JJF

中华人民共和国国家计量技术规范

JJF ××××－202×

高通量基因测序仪校准规范

Calibration Specification for High-Throughput Gene Sequencer

（征求意见稿）

202×－××－××发布　　202×－××－××实施

国家市场监督管理总局发布

高通量基因测序仪校准规范

JJF ××××－202×

Calibration Specification for High-Throughput Gene Sequencer

目录

引　　言

JJF 1071《国家计量校准规范编写规则》、 JJF 1001《通用计量术语及定义》和 JJF 1059.1《测量不确定度评定与表示》 共同构成支撑本规范制定工作的基础性系列规范。校准方法及计量特性等主要参考了 YY/T 1723-2020《高通量基因测序仪》、GB/T 30989-2014 《高通量基因测序技术规程》、GB/T 40226-2021《环境微生物宏基因组检测 高通量测序法》。

本规范为首次发布。

高通量基因测序仪校准规范（一）

1 范围

本规范适用于高通量基因测序仪（单次测序长度≤1000bp）的校准，不适用于桑格（Sanger）测序技术为原理的基因测序仪和纳米孔技术为主要原理的单分子测序仪的校准。

2 引用文件

本规范引用了下列文件：

JJF 1001—2011 通用计量术语及定义

JJF 1059.1 测量不确定度评定与表示

YY/T 1723-2020 高通量基因测序仪

GB/T 30989-2014 高通量基因测序技术规程

GB/T 40226-2021《环境微生物宏基因组检测 高通量测序法》

凡是注日期的引用文件，仅注日期的版本适用于本规范；凡是不注日期的引用文件，其最新版本（包括所有的修改单）适用于本规范。

3 术语和计量单位

YY/T1723-2020、GB/T 30989-2014、GB/T 40226-2021 界定的以及下列术语和定义适用于本规范。

3.1 高通量基因测序 high-throughput gene sequencing

将数万个以上的反应单元，随机或规则分布在固相载体表面上并行化运行，通过反应产生的光学或电学信号，同时监测每个测序反应的反应过程，来获得每个测序模板上的序列信息，从而得到大量乃至海量的序列信息。

3.2 高通量基因测序仪 high-throughput gene sequencer

具有明显区别于 Sanger 测序的基因测序仪，其特点主要表现在不必预先明确目的片

段的引物区序列、基于片段化的 DNA、依赖于独立反应体系进行克隆扩增、能一次进行对几十万到几百万条核酸序列（DNA）分子并行序列测定和读长一般较短等技术特征为标志。

[来源：YY/T 1723-2020，术语和定义 3.7]

3.3 **测序通量** throughput of gene sequencing

单次测序可获得序列信息的基因片段数量或可测定的脱氧核糖核酸和核糖核酸（以碱基表示）数量。

[来源：GB/T 30989-2014，术语和定义 3.2]

3.4 **碱基** base

一类含氮原子的有机杂环化合物，是组成嘌呤和嘧啶的主要成分，是拼出遗传密码的“字母”。

[来源：GB/T 30989-2014，术语和定义 3.16]

3.5 **文库** library

通过生物来源的、人工合成的或克隆技术等所得的一个重建分子群，如基因组文库、互补 DNA 文库、噬菌体展示肽文库等。

[来源：GB/T 30989-2014，术语和定义 3.5]

3.6 **标签文库** tag library

将 DNA 样品随机打断后，在其两端连接带特定序列的接头，在接头之间的序列称标签，所有不同的标签的集合即为标签文库。

[来源：GB/T 30989-2014，术语和定义 3.6]

3.7 **测序读长** read length of gene sequencing

单次运行可以读取的质量合格的序列片段长度，通常以碱基数量表示。

[来源：YY/T 1723-2020，术语和定义 3.4]

3.8 **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。

[来源：GB/T 30989-2014，术语和定义 3.26]

3.9 **碱基识别质量** quality of base calling

衡量碱基被正确识别的概率。通常以数字值直接表示。

碱基识别质量与碱基识别错误率之间的关系可用式（1）表示：

$$Q=-10 \lg P \tag{1}$$

式中：

Q——碱基识别质量；

P——碱基识别错误率。

[来源：GB/T 30989-2014，术语和定义 3.29]

3.10 Q20

测序数据中，碱基识别质量值为 20 时表示碱基识别准确率为 99%。

[来源：GB/T 40226-2021，术语和定义 3.5]

3.11 Q30

测序数据中，碱基识别质量值为 30 时表示碱基识别准确率为 99.9%。

[来源：GB/T 40226-2021，术语和定义 3.6]

3.12**碱基识别错误率** inaccuracy of base calling

单次测序错误碱基数占碱基总数的比例。

[来源：GB/T 30989-2014，术语和定义 3.28]

3.13**测序覆盖率** coverage rate of sequencing

待测样本的核苷酸序列检测结果覆盖于参考序列上的比例（测序覆盖率=覆盖区域长度÷参考序列总长度）。

[来源：GB/T 30989-2014，术语和定义 3.30]

3.14**测序深度** depth of coverage

待测样本中某个指定的核苷酸被检测的次数。

[来源：GB/T 30989-2014，术语和定义 3.31]

3.15**测序平均深度** average depth of coverage

测序深度的平均值。

[来源：GB/T 30989-2014]

3.16**比对率** mapping rate

比对到参考序列上的 reads 数目除以总测序数据的 reads 数目。

3.17**序列相对丰度** sequence relative abundance

某一指定序列在总样本中所占的相对比例。

基因相对丰度可用式（2）表示：

$$N_{\text{相对丰度}} = \frac{n_i}{\sum_{i=1}^{n} n_i} \times 100\% \tag{2}$$

式中：

n_i——比对到 i 序列的 reads 数与 i 序列的长度的比；

N——序列相对丰度。

注：改写 GB/T 40226-2021，术语和定义 3.2。

3.18**序列数据库** sequence database

分子生物信息数据库中最基本的数据库，包括核酸和蛋白质两类，以核苷酸碱基顺序或氨基酸残基顺序为基本内容，并附有注释信息。

4 概述

高通量测序（High-Throughput Sequencing），也称为大规模并行测序（Massively Parallel Sequencing, MPS）或下一代测序（Next Generation Sequencing，NGS），有效解决一代测序（Sanger Sequencing）成本高、通量低、对人力需求高等问题，可一次对几百万到几十亿条核酸分子进行序列测定。目前高通量测序的平台主要有基于桥式 PCR

扩增结合边合成边测序、乳液 PCR 扩增结合半导体合成测序、DNA 纳米球滚环扩增结合联合探针锚定聚合技术等技术原理的测序仪。

高通量测序仪主要由主机、基因测序仪控制软件组成，其中主机包括主体架构、操作系统主机、光学系统、XYZ 平台、芯片平台、气液系统、电子控制系统、试剂存储系统、电源系统、显示系统等组成。详见图 1.

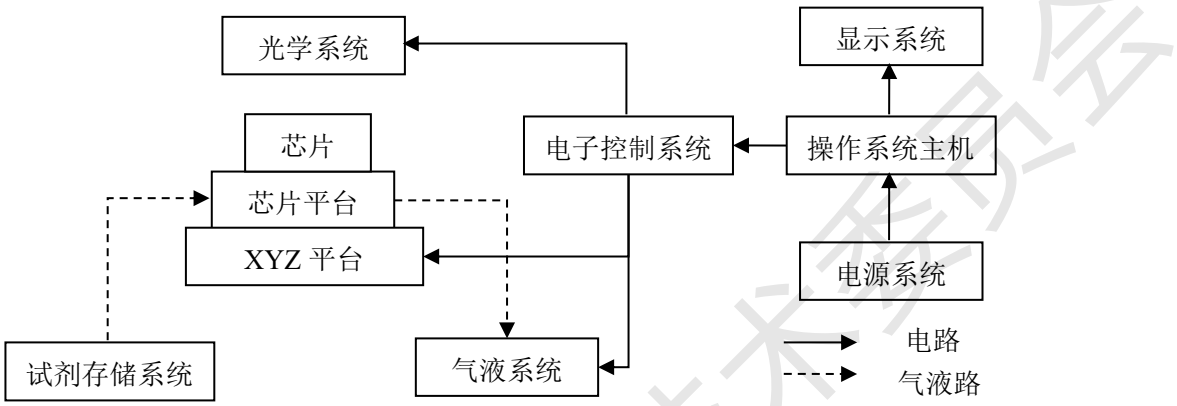

图 1　高通量基因测序仪结构示意图

5　计量特性

高通量测序仪的计量特性，如表 1 所示。

表 1 高通量测序仪的校准项目

序号	计量特性	计量性能指标	
1	读长总数重复性（Reads）	≤15%	
2	GC 含量占比偏差（GC%）	±10%	
3	碱基识别质量百分比	Q20	≥90%
		Q30	≥85%
4	平均碱基错误率	≤0.05%	
5	比对率*	≥99.00%	
6	比对率重复性*	≤5%	
7	序列覆盖率重复性*	⩽15%	
8	测序一致序列准确率*	≥99.00%	
9	序列相对丰度偏差*	±5%	

*数据相关参数依据所用标准物质要求进行分析计算。
注：1、以上参考指标不适用于合格判定，仅供参考。

2、测序数据量至少占被校准测序仪最大通量的1%或满足相关标准物质最低分析数据量。

6 校准条件

6.1 环境条件

校准前需将高通量测序仪开机预热30min，在下述适用的环境条件下，确保仪器达到正常工作状态。

6.1.1 环境温度：(10～30) ℃。

6.1.2 相对湿度：≤80%，无冷凝水。

6.1.3 室内应具备良好的防尘措施，高通量测序仪应远离振动、电磁干扰。

如有特殊条件保证仪器正常工作，参考仪器设备安装要求。

6.2 校准用标准物质、试剂

6.2.1 标准物质：校准时应采用经计量行政部门批准发布的片段化DNA序列有证标准物质。至少包含两种以上明确量值信息的参考序列，且每种参考序列长度≥5 kbp。参考序列相对丰度的扩展不确定度≤15% (k=2)。标准物质的选择原则参照附录A。

6.2.2 试剂：配套高通量测序仪相关试剂盒。

6.2.3 配制校准用标准物质所需一级水（18.2 MΩ）。

6.3 校准用设备及分析软件

6.3.1 移液器：规格为 10 μL、100 μL、1000 μL，经计量检定合格。

6.3.2 分析软件：推荐使用FastQC、Fastp、Trimmomatic、BWA、Bowtie、Bamstats、Samtools、IGV、SAS、SPSS等标准统计分析软件。

7 校准项目和校准方法

将高通量测序仪开机预热 30min，确保仪器达到正常工作状态后，使用标准物质按照待校准测序仪上机流程分别上机测序3次，将下机数据（FASTQ文件）使用分析软件（FastQC或Fastp等）处理，处理后与参考序列进行比对并开展统计分析，按照下面的具体指标开展校准。

7.1 读长总数重复性（Reads）

使用标准物质测序获得总 Reads 值，根据公式（3）计算测序总 Reads 的重复性。

$$RSD_{\text{reads}} = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}} \times \frac{1}{\bar{x}} \times 100\% \tag{3}$$

式中：

RSD_{reads}——测序总 Reads 的相对标准偏差，以百分数表示；

x_i——第 i 次测序结果的总 Reads 值；

$\bar{x}$——3 次测序结果总 Reads 的平均值；

n——测量次数。

7.2 GC 含量百分比偏差

使用标准物质测序获得 GC 的碱基数和总的碱基数，根据公式（4）计算测序 GC 含量百分比。测序 3 次，分别获得 GC 含量百分比，根据（5）和（6）计算 GC 含量百分比偏差。

$$GC\% = \frac{n_{\text{GC}}}{N_{\text{碱基}}} \times 100\% \tag{4}$$

$$\Delta GC\% = GC\%_s - \overline{GC\%} \tag{5}$$

$$\overline{GC\%} = \frac{1}{n}\sum_{i=1}^{n} GC\%_i \tag{6}$$

式中：

n_{GC}——测序碱基是 G 和 C 的碱基数；

$N_{\text{碱基}}$——测序获得总碱基数；

$GC\%$——测序碱基是 G 和 C 的碱基数占比，以百分数表示；

$\Delta GC\%$——GC 含量百分比偏差；

$\overline{GC\%}$——n 次测序 GC 含量百分比平均值；

$GC\%_i$——第 i 次 GC 含量占比百分数；

$GC\%_s$——GC 含量标准值。

7.3 Q20

使用标准物质测序获得碱基识别准确率为 99%的碱基数量和总的碱基数，根据公式（7）计算测序 Q20 的结果。

$$Q20=\frac{n_{\text{Q20碱基}}}{N_{\text{总碱基}}}\times 100\% \tag{7}$$

式中：

Q20——测序碱基识别准确率为 99%的碱基含量占比，以百分数表示；

$n_{\text{Q20 碱基}}$——碱基识别准确率为 99%的碱基数；

$N_{\text{总碱基}}$——测序获得总碱基数。

7.4 Q30

使用标准物质测序获得碱基识别准确率为 99.9%的碱基数量和总的碱基数，根据公式（5）计算测序 Q30 的结果。

$$Q30=\frac{n_{\text{Q30碱基}}}{N_{\text{总碱基}}}\times 100\% \tag{8}$$

式中：

Q30——测序碱基识别准确率为 99.9%的碱基含量占比，以百分数表示；

$n_{\text{Q30 碱基}}$——碱基识别准确率为 99.9%的碱基数；

$N_{\text{总碱基}}$——测序获得总碱基数。

7.5 平均碱基错误率

使用标准物质测序，获得每个碱基识别质量值，根据公式（9）和（10）计算平均碱基错误率。

$$P=10^{-\frac{\overline{Q}}{10}} \tag{9}$$

$$\overline{Q}=\frac{1}{n}\sum_{i=1}^{n}Q_i \tag{10}$$

式中：

P——碱基识别错误率；

Q——碱基识别质量；

$\overline{Q}$——所有碱基识别质量的平均值；

Q_i——第 *i* 个碱基识别质量值。

7.6 比对率重复性

使用标准物质测序获得的 FASTQ 数据，经比对后进行统计分析（Bamstats、

Samtools），获得比对率，根据公式（11）计算测序比对率的重复性。

$$RSD_{比对率}=\sqrt{\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1}}\times\frac{1}{\bar{x}}\times100\% \tag{11}$$

式中：

$RSD_{比对率}$——测序比对率的重复性；

x_i——第 i 次测序结果的比对率值；

$\bar{x}$——3 次测序结果比对率的平均值；

n——测量次数。

7.7 序列覆盖率重复性

使用标准物质测序获得的 FASTQ 数据，经比对后进行统计分析（Bamstats、Samtools），获得覆盖率。根据公式（12）别计算测序覆盖率。

$$RSD_{覆盖率}=\sqrt{\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1}}\times\frac{1}{\bar{x}}\times100\% \tag{12}$$

式中：

$RSD_{覆盖率}$——序列覆盖率的重复性；

x_i——第 i 次测序结果的覆盖率值；

$\bar{x}$——3 次测序结果覆盖率的平均值；

n——测量次数。

7.8 测序一致序列准确率

使用标准物质测序获得的 FASTQ 数据，经比对后进行统计分析（Bamstats、Samtools）。根据公式（13）计算测序一致性序列准确率。

$$AC=\left(1-\frac{MS}{MB}\right)\times100\% \tag{13}$$

式中：

AC——测序一致序列准确率；

MS——错配碱基数；

MB——指定目标区域基因组碱基数。

7.9 序列相对丰度偏差

使用标准物质测序获得的 FASTQ 数据，经比对后进行统计分析（Bamstats、Samtools），根据公式（9）计算参考基因相对丰度。

$$N_{相对丰度}=\frac{n_i}{\sum_{i=1}^{n} n_i}\times 100\% \tag{9}$$

$$\Delta N = N_s - \overline{N} \tag{5}$$

式中：

ΔN——目标序列相对丰度偏差；

$\overline{N}$——n 次测序目标序列相对丰度平均值；

N_s——目标序列相对丰度标准值；

$N_{相对丰度}$——基因相对丰度；

n_i——目标基因含量

8 校准结果表达

校准记录应尽可能详尽地记载测量数据和计算结果，推荐的校准记录格式见附录 B。各测量结果的测量不确定度应按 JJF 1059.1 的要求评定，不确定度评定示例见附录 C。经校准的高通量测序仪应出具校准证书，校准证书应符合 JJF 1071—2010 中 5.12 的要求。

9 复校时间间隔

由于复校时间间隔的长短是由高通量测序仪的使用情况、使用者、仪器本身质量等诸因素所决定的，因此，送校单位可根据实际使用情况自主决定复校时间间隔，建议不超过 1 年。

附录 A 标准物质的选择原则

标准物质的选择原则

A.1 标准物质的选择

A1.1 采用来源稳定、溯源信息完备和参考序列信息已知的核酸样本；

A1.2 选用的标准物质 GC 占比应该在 40%-60%之间，为含量均一的片段化 DNA，且不具有影响校准性能评估的特征序列；

A1.3 应至少包含组装成两种以上不同的参考序列，每种参考序列长度≥5 kbp；

A1.4 选用的标准物质涵盖的所有片段 DNA 均需要有明确量值和序列信息。

A.2 参考序列相对丰度

A.2.1 目标序列相对丰度范围为 40%-60%，相对丰度的相对扩展不确定度 U≤15% (k=2)。

附录 B 校准原始记录参考格式

校准原始记录参考格式

客户名称	
器具名称	
型号/规格	
出厂编号	
生产厂商	
客户地址	
校准日期	
校准环境条件及地点	
地点	
温度	
湿度	
其他	

基础信息	
读长类型	
设备最大通量	
测序试剂型号	
数据量	

一、读长总数（reads）重复性

	测定值			平均值	重复性
	1	2	3		
reads					

二、GC 含量占比偏差

	测定值			平均值	相对偏差	重复性
	1	2	3			
GC%						

三、碱基识别质量

	测定值			平均值	参考指标
	1	2	3		
Q20					
Q30					
平均碱基错误率					

四、比对率重复性

	测定值			平均值	重复性	参考指标
	1	2	3			
比对率						

五、序列覆盖率重复性

	测定值			平均值	重复性
	1	2	3		
覆盖率					

六、测序一致序列准确率

	测定值			平均值	重复性	参考指标
	1	2	3			
准确率						

七、序列相对丰度偏差

	测定值			平均值	相对偏差	重复性
	1	2	3			
序列相对丰度						

附录 C 校准证书(内页）参考格式

校准证书(内页）参考格式

共　页，第　页

序号	校准项目	校准结果
1	读长总数重复性（Reads）	
2	GC 含量占比偏差（GC%）	
3	碱基识别质量百分比	
4	平均碱基错误率	
5	比对率*	
6	比对率重复性*	
7	序列覆盖率重复性*	
8	测序一致序列准确率*	
9	序列相对丰度偏差*	，不确定度：　　（k=2）

校准员：_____　　核验员：_____

附录 D 序列相对丰度偏差测量结果的不确定度评定示例

D.1 测量模型

相对丰度偏差可由公式(D.1)给出：

$$\Delta c = \bar{c} - c_s \tag{D.1}$$

式中：

Δc ——相对丰度偏差，单位为%；

$\bar{c}$ ——3 次相对丰度测定结果的算术平均值，单位为%；

c_s ——标准物质的相对丰度标准值，单位为%。

D.2 不确定度来源

不确定度来源包括：

a）高通量测序仪测量重复性引入的标准不确定度 $u(\bar{c})$；

b）标准物质引入的标准不确定度 $u(c_s)$

D.3 标准不确定度分量的评定

D.3.1 测量重复性引入的标准不确定度分量 $u(\bar{c})$

实际校准时在重复性条件下连续测量 3 次，以 3 次测量的算术平均值作为结果，计算其实验标准偏差 s 和标准不确定度分量 $u(\bar{c})$ 。其中，

$$s = \frac{C_{\max} - C_{\min}}{C_n} \tag{D.2}$$

式中：

s ——实验标准偏差，单位为%；

$C_{\max}$——3 次测定结果的最大值，单位为%；

$C_{\min}$——3 次测定结果的最小值，单位为%。

C_n ——极差系数，n=3 时，C_n=1.69。

$$u(\bar{c}) = \frac{s}{\sqrt{3}} \tag{D.3}$$

结果见表 D.1

表 D.1 测量重复性引入的标准不确定度分量数据

C_i			$\overline{C}$	s	$u(\bar{c})$
1	2	3			
47.77%	50.95%	51.36%	51.70%	2.12%	1.23%

D.3.2 标准物质引入的标准不确定度分量 $u(c_s)$

由标准物质引入的不确定度分量 $u(c_s)$可以根据标准物质证书提供的扩展不确定度 $U(c_s)$和包含因子 k 根据公式(C.5)计算：

$$u(c_s)=\frac{U(c_s)}{k} \tag{C.5}$$

式中：

$u(c_s)$——标准物质引入的标准不确定度；

$U(c_s)$——标准物质证书提供的扩展不确定度；

k——标准物质证书提供的包含因子。

得 $u(c_s)$=2.60%

D.4 **合成标准不确定度** u_c

由公式（C.6）可得合成标准不确定度：

$$u_c = \sqrt{u(\bar{c})^2 + u(c_s)^2} \quad (C.6)$$

D.5 扩展不确定度U

取包含因子 k=2，则扩展不确定度U =2u_c

结果见表 C.2

表 C.1 合成标准不确定度和扩展不确定度

项目	u_c	U	k
基因相对丰度偏差	2.87%	5.8%	2
