ICS 11.100.10
CCS C 30

中华人民共和国国家标准

GB/TXXXXX—XXXX

单分子基因测序 第 1 部分 术语

Single-molecule gene sequencing - Part 1: Terms

征求意见稿

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。

2024 年 7 月 30 日

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。

本文件是GB/T ****-****《单分子基因测序》的第1部分，GB/T ****-****《单分子基因测序》已发布了：

—第1部分：术语。

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（SAC/TC136）归口。

本文件起草单位：

本文件主要起草人：

引 言

GB/T ****–规定了单分子基因测序的术语，测序仪性能评价方法，样本处理规范及质量要求，数据质控要求、分析流程及格式，以及软件要求。拟由5部分组成：

—第1部分：术语，目的在于界定单分子基因测序领域的术语和定义，规定具有连续测序特征的单分子基因测序主要参数指标术语和技术相关术语；

—第2部分：测序仪性能评价方法，目的在于规定单分子基因测序仪的性能评价要求及试验方法；

—第3部分：样本处理规范及质量要求，目的在于界定适用于单分子基因测序的样本类型及样本处理技术要求，规定样本质量要求及评价方法；

—第4部分：数据质控要求、分析流程及格式，目的在于界定单分子基因测序数据质量的术语和定义、规定数据质量要求及评价方法，描述测序信号至碱基序列的分析流程及测序序列层级目录格式和数据整体格式；

—第5部分：软件，目的在于对单分子基因测序数据进行生物信息学分析的软件结构组成、通用要求、专用要求和检验方法，及测序软件的计算机配置要求、接口规范等。

单分子基因测序 第1部分：术语

1 范围

本文件界定了单分子基因测序领域的术语和定义。

本文件适用于单分子实时荧光测序法、单分子纳米孔链测序法、单分子纳米孔标签测序法等技术为主要技术原理的具有连续测序特征的单分子基因测序领域。

本文件中参数指标术语、技术相关术语不适用于：

——桑格（Sanger）测序为主要技术原理的第一代基因测序领域；

——半导体测序法、可逆末端终止测序法、联合探针锚定连接测序法、联合探针锚定聚合测序法、焦磷酸测序法等技术为主要技术原理的大规模平行高通量测序领域；

——利用多个分立步骤进行非连续测序的单分子基因测序技术；

——利用共标签进行短序列连接的单分子长片段基因测序技术。

2 规范性引用文件

本文件没有规范性引用文件。

3 一般术语

3.1

基因组 genome

一种生物体具有的所有遗传信息的总和。

[来源：GB/T 30989—2014，3.2]

3.2

基因 gene

位于染色体上编码一个特定功能产物(如蛋白质或RNA分子)的一段核苷酸序列，是遗传信息的基本单位。

[来源：GB/T 30989—2014，3.1]

3.3

核酸 nucleic acid

作为遗传信息载体或信息表达中充当媒介的大分子。

注：存在两种类型的核酸，脱氧核糖核酸（deoxyribonucleic acid，DNA）和核糖核酸（ribonucleic acid，RNA）。

[来源：ISO 17822:2020，3.32]

3.4

脱氧核糖核酸 deoxyribonucleic acid;DNA

以双链或单链形式存在的脱氧核糖核苷酸聚合物。

[来源：ISO 17822:2020，3.17]

3.5

核糖核酸 ribonucleic acid;RNA

以双链或单链的形式存在的核糖核苷酸聚合物。

[来源：ISO 17822:2020，3.42]

3.6

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”。

注：DNA 中的碱基主要有腺嘌呤(Adenine，A),鸟嘌呤(Guanine，G)、胞嘧啶(Cytosine，C)和胸腺嘧啶(Thymine，T)；RNA 中的碱基主要有腺嘌呤(A)鸟嘌呤(G)胞嘧啶(C)和尿嘧啶(Uracil，U)。

[来源：GB/T 30989—2014，3.16]

3.7

碱基序列 base sequence

测序片段中记录碱基排列的字符串。

注：序列中的碱基可用对应的大写或小写字母表示，未测定的碱基用字母 N 或 n 表示。

[来源：GB/T 35890—2018，3.6，有修改]

3.8

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以及碱基修饰信息。

[来源：YY/T 1723—2020，3.1，有修改]

3.9

文库 library

测序文库 sequencing library

具有特定的大小范围，通常包含接头和/或用于测序引物结合的特定序列、序列捕获和/或识别特定区域的标识，作为测序模板的DNA、cDNA或RNA的核酸片段。

[来源：ISO 20397-1:2022，有修改]

3.10

碱基识别 base calling

测序过程中从电信号、光信号或其他由于测序反应而产生的信号转换成碱基序列信息的过程。

3.11

碱基识别质量 quality of base calling

衡量碱基正确识别的概率。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:

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

式中:

Q——碱基识别质量；

P——碱基识别错误率。

[来源：GB/T 30989—2014，3.29]

3.12

单分子基因测序 single-molecule gene sequencing

在单分子水平上对核酸分子进行连续碱基序列测定。

注1：通常基于光学或电学等信号转换成碱基信息的单分子长片段测序方式，区别于依赖核酸模板扩增实现碱基序列测定的桑格测序、大规模平行高通量测序等以及基于共标签进行短序列连接的单分子测序方式。

注2：单分子基因测序中，根据不同的测序平台和建库方式，可实现万级或兆级连续碱基直接测量。

3.13

直接测序 direct sequencing

不经过任何扩增与转化处理构建文库，直接读取原始模板链碱基或修饰碱基所产生的测序信号的测序方式。

注：为单分子基因测序的技术特征。

3.14

直接RNA测序 direct RNA sequencing

待测RNA分子经文库构建后，直接读取原始RNA模板链碱基或修饰碱基所产生的测序信号的测序方式。

3.15

实时测序 real-time sequencing

在单个核酸分子的连续测序反应发生时同步进行碱基识别和序列输出的测序方式。

注：为单分子基因测序的技术特征。

3.16

表观修饰直接测序 epigenetic modification direct sequencing

待测核酸分子上带有的表观遗传修饰，可不经过任何化学、生物等方法转化处理而被直接测定的测序方式。

注：表观遗传修饰包括5mC（5-甲基胞嘧啶）、5hmC（5-羟甲基胞嘧啶）、6mA（6-甲基腺嘌呤）等。

3.17

单分子测序文库 single-molecule sequencing library

为单分子基因测序准备的核酸分子特殊结构，通常通过核酸工具酶或其他方法将待测核酸分子与同单分子基因测序平台适配的接头偶联后获得。

注：根据文库的拓扑结构可分为单链DNA环状文库、双链DNA环状文库、双链线性DNA文库等。

3.18

单分子一致性序列 single-molecule consensus sequence

在单分子基因测序中，通过整合目标区域/目标片段的多重拷贝、重复读段或互补链进行互相校正后得到的单条高置信度碱基序列。

4 参数指标术语

4.1

测序通量 throughput of sequencing

单次运行可获得序列信息的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。

注：通量指标需标注完整单位，以明确通量描述的是片段数或碱基数。

[来源：GB/T 30989—2014，3.21]

4.2

单芯片测序通量 throughput of sequencing per flow cell

单次运行过程中，单张测序芯片可获得序列信息的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。

4.3

单位时间测序通量 throughput of sequencing per unit time

单次运行过程中，单位时间内可获得序列信息的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。

注：单位时间（per unit time），通常为小时或分钟。

4.4

测序读长 read length of sequencing

单次运行可读取的质量合格的序列片段长度,通常以碱基数表示。

注：单分子基因测序评价指标包含最长读长、平均读长、读长N50等。

[来源：YY/T 1723—2020，3.4]

4.5

最长读长 maximum read length

单次运行获得的质量合格的最长序列片段的长度，以碱基数表示。

4.6

平均读长 average read length

单次运行获得的质量合格的序列的碱基总数与片段数相除得到的长度，以碱基数表示。

4.7

读长 N50 read length N50

将单次运行获得的质量合格的序列片段由长至短进行排序并依次相加，当相加的碱基数刚好达到或超过总碱基数一半时加上的最后一条片段的长度，以碱基数表示。

4.8

中位数读长 median read length

将单次运行获得的质量合格的序列片段按长度由长至短进行排序和累积计数，当累积数量刚好达到或超过总片段数的一半时计数的最后一条片段的长度，以碱基数表示。

4.9

测序准确度 accuracy of sequencing

单次运行获得的质量合格的序列片段，其原始序列或经处理后的单分子一致性序列与参考序列的一致程度。

测序准确度的计算方法可用式（2）表示：

$$Accuracy = Matches/(Matches + Substitutions + Insertions + Deletions) \quad \cdots\cdots(2)$$

式中:

Accuracy——测序准确度；

Matches——比对正确碱基数；

Substitutions——替换错误碱基数；

Insertions——插入错误碱基数；

Deletions——缺失错误碱基数。

注：用于单条序列的评价。

4.10

平均准确度 average accuracy

单次运行获得的质量合格的所有序列片段，其原始序列或经处理后的单分子一致性序列经与参考序列比对后，所有序列片段与参考序列的一致程度。

注：用于所有序列的整体统计评价。

4.11

中位数准确度 median accuracy

将单次运行获得的质量合格的序列片段按测序准确度由高至低进行排序和累积计数，当累积数量刚好达到或超过总片段数的一半时计数的最后一条片段的准确度。

4.12

众数准确度 modal accuracy

单次运行获得的质量合格的各序列片段的测序准确度所做直方图中的最高峰对应的准确度。

4.13

一致性准确度 consensus accuracy

单次运行获得的质量合格的序列片段，经多序列比对校正处理后得到的一致性序列与参考序列的一致程度。

注1：同测序准确度的计算方法类似，用一致性序列中比对正确的碱基数与目标区域所有类型碱基的总数的比值来表示。

注2：用于一致性序列的评价。

4.14

单次测序准确度 single-pass accuracy of sequencing

测序获得的序列片段，其未经校正的原始序列与参考序列的一致程度。

4.15 单分子一致性测序准确度 single-molecule accuracy of sequencing

针对单个模板分子进行测序、碱基识别与校正（使用多重拷贝测序、循环一致性测序及双链测序方法时）后得到的序列与参考序列的一致程度。

注：对文库分子进行单次测序的平台，单分子测序准确度等同于单次测序准确度；对文库分子进行循环一致性测序的平台，单分子一致性测序准确度等同于循环一致性测序准确度；对文库分子进行双链测序的平台，单分子一致性测序准确度等同于分子内互补链一致性准确度。

5 技术相关术语

5.1

单分子实时荧光测序 single-molecule real-time fluorescent sequencing

单个核酸分子被固定在零模波导孔底部的单个聚合酶捕获后，聚合酶延伸引物链添加新碱基，并产生连续脉冲荧光信号，经碱基识别获得核酸分子的碱基序列与修饰信息的测序方式。

5.2

单分子纳米孔链测序 single-molecule nanopore strand sequencing

单个核酸分子被纳米孔捕获后，碱基单元在马达蛋白或其他方式的控制下逐一通过纳米孔，并产生连续的电信号变化，该时序电信号经碱基识别获得核酸分子的碱基序列及修饰信息的测序方式。

5.3

单分子纳米孔标签测序 single-molecule nanopore tag sequencing

单个核酸分子被连接于纳米孔的聚合酶捕获后，在延伸引物链过程中，标签标记的核苷酸与纳米孔相互作用（穿过或其他）产生特定电流信号，经碱基识别获得碱基序列及修饰信息的测序方式。

5.4

单分子边合成边测序 single-molecule sequencing by synthesis

通过荧光标记或标签标记核苷酸或其衍生物被连续添加到新合成核酸链的过程，实现对单个核酸分子碱基序列识别的测序方式。

5.5

双链测序 duplex sequencing

对模板链及其互补链先后进行测序和碱基识别的测序方式。

注：对互补链进行测序的目的是与模板链的碱基序列相互进行校正，以提高测序准确度。

5.6

循环一致性测序 circular consensus sequencing

通过构建环状文库，对模板分子循环多次测定序列，产生子读序，将多个子读序对齐后进行比对校正的测序方式。

注：可应用于单分子实时荧光测序和单分子纳米孔链测序等测序方式。

5.7

子读序 subread

循环一致性测序中，模板分子的每一轮测序获得的一个测定序列。

注：完整的模板链或互补链测定一次即为一轮。例如，针对哑铃型文库，滚环一周完成模板链与互补链各测一次，此情况按两轮计算。

5.8

脉冲荧光信号 pulsed fluorescent signal

单分子实时荧光测序时，荧光染料标记的核苷酸进入检测区域并在聚合酶介导下发生连接反应延伸引物链，在此过程中染料分子被激发而发射的短时间持续的荧光信号。

5.9

零模波导孔 zero‐mode waveguide (ZMW)

一种直径为数十纳米孔道结构的纳米级别光子器件，用于检测在限定的极小体积内发出的光学信号。

注：由于其直径小于激发光波长而将激发范围限制在孔底部的聚合酶附近，从而实现对单个分子合成反应荧光信号的观测，可用于探测特定波长范围内的光信号，有效避免背景游离荧光分子的干扰。

5.10

哑铃型文库 dumbbell-shaped library

使用发夹结构序列将待测双链模板分子的两端进行连接而形成的哑铃形状的测序文库。

注：可用于多种单分子基因测序技术，包括单分子实时荧光测序与单分子纳米孔标签测序。

5.11

测序动力学 kinetics of sequencing

测序过程中，核苷酸在合成时由于其类型及所带修饰不同，产生的脉冲式信号在脉冲宽度与脉冲间持续时间方面存在差异性表现的现象，基于此现象可建模分析核酸分子上的修饰事件。

5.12

标签标记核苷酸 tagged nucleotide

用于单分子纳米孔标签测序的磷酸端带有特殊标签分子的修饰核苷酸。

注：不同核苷酸带有不同的标签分子。

5.13

纳米孔 nanopore

具有贯穿通道的纳米量级的孔道，是实现碱基序列编码为电信号的转换器。

注：通常分为生物蛋白孔道与固态孔道。

5.14

纳米孔测序芯片 nanopore sequencing flow cell/flow cell

集成纳米孔传感、信号检测及配套流体结构的完整测序单元。

5.15

马达蛋白 motor protein

单分子纳米孔链测序中，用于对待测核酸分子进行控速的蛋白分子。

注：通常为解旋酶，也可以是聚合酶或其他类型蛋白。

5.16

自适应性采样 Adaptive Sampling

无需前置文库富集步骤，允许测序仪在测序过程中实时选择或拒绝序列的靶向测序方法。

注：技术使用实时碱基识别，可根据 DNA 片段的初始序列组成，接受或拒绝其用于进一步测序，以达到特定的测序目标或提高测序数据的质量。

参 考 文 献

[1] GB/T 30989—2014 高通量基因测序技术规程
[2] GB/T 35890—2018 高通量测序数据序列格式规范
[3] YY/T 1723—2020 高通量基因测序仪
[4] ISO 17822:2020 In vitro diagnostic test systems — Nucleic acid amplification-based examination procedures for detection and identification of microbial pathogens — Laboratory quality practice guide
[5] ISO 20397-1:2022 Biotechnology — Massively parallel sequencing Part 1: Nucleic acid and library preparation
[6] Eid J., Fehr A., Gray J. et al. Real-time DNA sequencing from single polymerase molecules. Science 2009 Jan, 233 pp.133-138
[7] Wenger A.M., Peluso P., Rowell W.J. et al. Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. Nat. Biotechnol. 2019 Oct, 37 pp.1155-1162
[8] Wang Y.H., Zhao Y., Bollas A. et al. Nanopore sequencing technology, bioinformatics and applications. Nat. Biotechnol. 2021 Nov, 39 pp.1348-1365
[9] Hook P.W., Timp W. Beyond assembly: the increasing flexibility of single-molecule sequencing technology. Nat. Rev. Genet. 2023 Sep, 24 pp.627-641
[10] Fuller C.W., Kumar S., Porel M. et al. Real-time single-molecule electronic DNA sequencing by synthesis using polymer-tagged nucleotides on a nanopore array. Proc. Natl. Acad. Sci. U.S.A. 2016 May, 113 pp.5233-5238