ICS 11.100.10
CCS C30

中华人民共和国国家标准

GB/T XXXXX—XXXX

高通量测序样品标签通用要求

General requirements of high-throughput sequencing sample index

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。

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/TC 136）归口。

本文件起草单位：TBD

本文件主要起草人：TBD

高通量测序样品标签通用要求

1 范围

本文件规定了高通量测序样品标签的设计，合成和质量评价的要求。

本文件适用于基于高通量测序法中使用的样品标签。

本文件不适用于单分子测序法中使用的样品标签。

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 35537-2017 高通量基因测序结果评价要求

GB/T 34797-2017 核酸引物探针质量技术要求

YY/T 1723—2020 高通量测序仪

3 术语和定义

下列术语和定义适用于本文件。

3.1

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。

[来源：YY/T 1723—2020，术语和定义 3.1]

3.2

高通量测序 high-throughput sequencing

以一次并行几十万到几百万条核酸分子序列测定和一般读长较短等为标志，适用于 DNA 的测序技术。

[来源:GB/T 35890-2018]

3.3

样本标签 sample index

测序片段的ID，保证一个序列编号对应一段序列片段或对应一个测序文库，具有唯一性。

3.4

GC含量 GC content

测序片段碱基中G（鸟嘌呤）和C（胞嘧啶）所占的百分比。

3.5

单端标签 single index

样本拆分使用的标签仅包含一段序列。

3.6

组合型双标签 combinatorial dual index

样本拆分使用的标签包含两段序列，单端序列在一组标签间可能存在重复。

3.7

唯一性双标签 unique dual index

样本拆分使用的标签包含两段序列，单端序列在一组标签间不会重复。

3.8

莱文斯坦距离 Levenshtein distance

两个字符串之间的莱文斯坦距离是将一个字符串更改为另一个字符串所需的最少编辑操作次数。莱文斯坦距离允许的编辑操作包括单个字符的 替换/插入/删除。可用于评价两个序列之间的相似程度。

4 要求

4.1 样本标签设计

4.1.1 序列相似度

在设计一组样本标签时，使用莱温斯坦距离对标签间相似性进行评估。单端标签，任意标签间的莱温斯坦距离不小于3；唯一性双标签，任意标签间的单端莱温斯坦距离不小于2，双端莱温斯坦距离之和不小于5；组合型双标签序，任意标签间至少一端莱温斯坦距离不小于3。

4.1.2 碱基平衡

在一组样本标签的每一个位置,同时存在A,T,C,G四种碱基,最优选的情况下每种碱基占比均为25%。单一碱基占比的为10%~50%。

4.1.3 连续相同碱基

样本标签的连续相同碱基长度不大于3。

4.1.4 GC 含量

样本标签的GC含量为25%~75%。

4.1.5 特殊序列

标签的序列与文库接头通用序列任一连续区域不相同，不反向互补。

4.2 样本标签合成

样本标签作为寡核苷酸序列生产时，应满足GB/T 34797-2017 核酸引物探针治疗技术要求。

4.3 样本标签质量

4.3.1 样本标签纯度

样本标签纯度应满足产品性能要求，应不小于 85%。

4.3.2 样本标签交叉污染比例

样本标签交叉污染比例应满足产品性能要求，单一交叉污染类型比例最高不超过0.1%,交叉污染总和不超过0.5%。

5 试验方法

5.1 样本标签设计

5.1.1 序列相似度

计算一组样本标签中，每两个标签间的莱文斯坦距离。唯一性双标签和组合型双标签的为两端序列莱文斯坦距离之和。结果应符合 4.1.1 的要求。

5.1.2 碱基平衡

计算一组样本标签中，每个位置的碱基占比。结果应符合4.1.2的要求。

5.1.3 连续相同碱基

计算每个样本标签中，最大的连续相同碱基值。结果应符合4.1.3要求。

5.1.4 GC 含量

计算每个样本标签的GC含量，唯一性双标签和组合型双标签单独计算两端序列的GC含量。结果应符合4.1.4要求。

5.1.5 特殊序列

核查样本标签中是否存在特定序列。结果应符合4.1.5要求。

5.2 实验流程

5.2.1 样本

实验选择具有唯一性的样本序列，可通过插入片段识别样本，插入片段应与测序读长匹配，不宜太长过太短。DNA样本的完整度、纯度、浓度及总量满足制造商规定的样本质量要求。样本具有稳定的来源，不同检验批次间实验结果稳定。

5.2.2 样本标签合成

样本标签作为寡核苷酸序列生产时，应满足4.2的要求。

5.2.3 文库制备

可采用靶向扩、接头链接等方式制备文库，样本文库序列与样本标签对应关系具有唯一性。文库经过质控后进行测序。

5.2.4 数据拆分

根据测序插入序列对样本数据进行拆分。并统计样本标签序列组成。

5.2.5 数据分析

统计每个样本的样本标签组成。测序总 reads 数（TR），与预期样本标签相同的序列为正确序列其数量（CR），与其他样本标签相同的序列为交叉污染序列（PR1，PR2，…，PRn）。根据式（1）计算样本标签纯度，根据式（2）计算各种类交叉污染比例，根据式（3）计算交叉污染交叉污染比总和。

$$纯度 = \frac{CR}{TR} \times 100\% \quad \cdots\cdots (1)$$

$$交叉污染n = \frac{PRn}{TR} \times 100\% \quad \cdots\cdots (2)$$

$$交叉污染总和 = 交叉污染1 + 交叉污染2 + \cdots + 交叉污染n \quad \cdots (3)$$

5.3 样本标签质量

5.3.1 样本标签纯度

按照5.2.5计算样本标签纯度。结果应满足4.3.1要求。

5.3.2 样本标签交叉污染比例

按照5.2.5计算单一交叉污染比例，并确定最大值；按照5.2.5计算所有交叉污染总和。结果应满足4.3.2要求。

5.3.3 实验方法稳定性

采用多批次多重复试验验证试验方法稳定性，应符合 4.3.1-4.3.2 的要求。

参 考 文 献

[1]GB/T 35537-2017 高通量基因测序结果评价要求
[2]GB/T 34797-2017 核酸引物探针质量技术要求
[3]YY/T 1723—2020 高通量测序仪