ICS 11.100.10
CCS C30

中华人民共和国国家标准

GB/T XXXX-XXXX
a

病原微生物基因组参考数据库通用要求

General requirements for construction of pathogen genome database

（申报稿）

b

c

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

目　次

前　言

本文件按照GB/T 1.1—2009给出的规则起草。

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这些专利的责任。

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

病原宏基因组测序（mNGS）技术因其检测时间短、分辨率高，能识别罕见、新发病原体引发的感染或者混合感染等优势，已经越来越广泛地用于临床疑难感染的辅助诊断。然而，由于目前尚缺乏高质量临床微生物基因组参考数据库等问题，部分程度上制约了该技术临床应用的进一步发展。

制定病原微生物基因组参考数据库建设标准，能够为相关产品的研发提供指导，为监管提供支撑。

病原微生物基因组参考数据库

1 范围

本文件规定了病原微生物基因组参考数据库的要求。本文件适用于基于病原微生物基因组和宏基因组高通量测序方法，对人体样品中病原微生物进行分析的参考数据库。

本文件适用的测序原理包括单分子纳米孔链测序、可逆末端终止测序、联合探针锚定聚合测序。

2 规范性引用文件

下列文件对于本文件的应用是必不可少的。凡是注日期的引用文件，仅所注日期的版本适用于本文件。凡是不注日期的引用文件，其最新版本（包括所有的修改单）适用于本文件。

《宏基因组数据采集的通用标准》

3 术语和定义

下列术语、缩略语和定义适用于本文件。

ISO和IEC的术语数据库可以通过下述网址访问：

——ISO：http://www.iso.org/obp；

——IEC：http://www.electropedia.org/

3.1 下一代测序（next-generation sequencing，NGS）

下一代测序（next-generation sequencing, NGS），也称为二代测序，是通过对短核苷酸序列的平行测序，读取每条序列（read）的碱基信息，用以确定整个基因组或DNA或RNA中目标区域核苷酸顺序的方法。

3.2 宏基因组高通量测序（metagenomic next-generation sequencing, mNGS）

病原微生物宏基因组也称元基因组高通量测序是利用NGS技术，检测来自临床患者标本中的微生物基因组的所有核酸序列。并通过生物信息学分析，判断和明确标本中的微生物存在与否及其组成与患者感染或疑似感染关系的方法。

3.3 湿实验

指mNGS检测中在实验室进行的生物标本处理、核酸提取、文库构建和测序的过程。

3.4.干实验

利用计算机和生物信息学技术进行的数据质控，数据分析和结果处理的实验过程。

3.5 低质量序列

mNGS 检测中测序下机数据中存在的不符合数据质控要求的序列，通常包括接头序列、低质量的末端序列、含 N 碱基数量过多的序列、过短序列等。

3.6 全基因组相似性（ANI）

指在核酸水平两两基因组之间所有直系同源蛋白编码基因的相似性，用于计算基因组之间的进化距离。通常在同一物种（species）层面的全基因组相似性≥95%。

3.7. 参考菌株基因组

指基因组平均相似度 ANI≥95%的同一物种基因组序列中序列长度最长、序列完整性最好的一株菌的基因组序列。

3.8 代表性基因组

从基因组相似性的角度，挑选能够代表某一个基因组聚类的代表性序列。

3.9 比对

指将测序所得序列与参考基因组序列进行匹配的过程。

3.10 覆盖深度

针对单个物种基因组，测序序列中比对上该基因组的总碱基数除以基因组被覆盖序列的基数的数值。

3.11 序列相似度

两条核酸序列对应位置上相似或/和相同残基的数目占总核酸序列长度的百分数。

3.12 覆盖度

测序标本检出的病原体核酸序列覆盖对应病原体参考全长基因组（属 / 种水平）的比例（覆盖度＝覆盖区域长度 / 参考基因组序列总长度）[2]。

3.13 均一性

基因组均一性是评价测序序列是否随机覆盖到参考基因组的指标，通常以物理熵值（Entropy）衡量，反映比对过程中匹配的数据是真实匹配还是同源性或是非特异性干扰[3]。

3.14 组装

通过序列比对和序列拼接等算法，将短片段的测序序列建构成为较长的连续序列的过程。

3.15 物种注释

指通过生物信息分析，鉴定测序序列来自某种特定的微生物物种，从而注释标本中所包含的微生物物种种类。

3.16 物种谱

对特定标本中所包含的微生物物种组成及其对应相对丰度的描绘。

3.17 人体定植微生物（colonizing microorganisms）

在人体特定部位定居和不断生长、繁殖后代的微生物，主要是细菌、真菌、病毒。一般情况下，这类微生物不会引起人体疾病。人体标本定植微生物分类见背景微生物库构建。

3.18 元数据 metadata

定义和描述其他数据的数据。

[来源： ISO/IEC 11179-1:2015，3.2.16]

3.19 永久标识符 persistent identifier

唯一标识符，通过提供对数字对象的物理位置或当前所有权的访问，确保对数字对象的永久访问。

示例 https://doi.org/10.1016/S0140-6736（20）30251-8

注：用于互联网数字化对象，永久标识符不仅具有永久性，也就有可操作性（例如 URI）。

[来源：ISO 24619：2011,3.2.4,条目注释 1]

4 基因组数据库数据来源要求

4.1 数据库来源

选取具有权威性的公共数据来源作为建立数据库的重要来源，尽量优先选取数据库中经过人工校验的高质量数据集。数据来源的高质量数据集应尽量满足物种分类正确，基因组完整度高，污染度小，且冗余度低，具有相对完整的元数据描述信息。

注：权威性的公共数据来源，如NCBI数据库中的RefSeq数据集，ARGOS数据库等。

4.2 物种类型

微生物数据库物种类型完整性包括数据库所包含物种类型检测范围的完整性。数据库中的物种类型范围要尽可能覆盖已知病原体（包括罕见病原体），选择全面的病原微生物数据，包括但不限于细菌、真菌、病毒、寄生虫等。

4.3 公共数据库

在整合公共数据的同时，进行病原微生物基因组测序，补充常见但目前缺失的病原体基因组数据。

4.4 物种分类

对基因组序列集去冗余，在非冗余序列集中，要考虑物种间的特异性和物种内的保守性。数据库中

的序列宜用代表菌的全基因组序列作为物种代表基因组构建分类集合。

5 数据质量控制的要求

5.1 参考菌株基因组

优先选择高质量的微生物基因组序列作为参考菌株基因组，应考虑基于物种的命名规范、基因组的组装完整性、基因组长度、含N碱基比例、污染片段等质量控制要求。

5.2 参考菌株分类

比对数据库的准确性是指要尽可能保证数据库中的参考菌株分类准确、参考基因组序列无污染片段等。如需要过滤掉质粒序列、信息标注错误的序列、染色体组装不完整和分类错误的菌株、基因组重叠群片段（contig）长度小于100bp及低重复的基因组序列，并且去除人源污染序列及真菌、寄生虫基因组中细菌污染序列。

5.3 物种特异性

对于高同源物种，还要考虑数据库中物种间参考菌株的相似性对于物种检出灵敏度的影响，以提高对高同源物种检测的特异性。

5.4 物种分类命名

应确保物种分类名称/学名的准确性，方法是将名称的数据字段与公开可用的分类数据库链接起来，或根据相关的国际命名规范，重新检查已确认的有效分类名称的来源。

6 元数据内容的要求

6.1 数据字段

病原微生物基因组数据应包括所有必需数据字段，包括附件A.物种名称，微生物类型，基因组序列数据等。应根据附件A中规定的具体情况选择全部或部分可选数据字段。

6.2 物种命名

应将微生物物种的每个名称与公开可用的命名数据库链接起来，以便数据检索和访问。应将每个序列（例如基因组、元基因组、cDNA、RNA和蛋白质序列）链接到公共数据库。

7 数据库建设流程要求

7.1 数据来源

首先从各数据来源，包括公共数据来源及自有数据来源，对数据的元数据信息，包括序列编号，时间，分离源等信息以及对应的序列文件进行汇聚，建立一个待处理的中间数据库。

7.2 中间数据评估

对中间数据库的数据进行数据质量评估及过滤，包括元数据的完整性和准确性，以及测序数据的完整度、污染度等指标，选取符合数据质量要求的数据进行后续处理。

7.3 数据纯度

对通过质量评估后的数据，进一步进行污染序列过滤处理，去除污染序列。

7.4 分类信息评估

对数据进行分类信息评估与确认，存在于公共数据库中的数据，可能由于提交人的失误，或受限于提交时可进行比对的参考数据较少，存在分类错误。因此，对于通过质量控制的序列，应当进行平均核苷酸相似度（Average Nucleotide Identity，ANI）比较，或者构建进化树，确定每一个挑选的基因组具有正确的分类地位，剔除分类错误基因组。

7.5 物种分类命名

对微生物的命名进行校正及统一。微生物命名需要参考国际权威的微生物命名数据库，细菌可以参考原核生物名称列表（List of Prokaryotic names with Standing in Nomenclature，LPSN），真菌可以参考真菌命名（Fungal Names），病毒可以参考国际病毒分类委员会（International Committee on Taxonomy of Viruses， ICTV）等数据库。然而，无论哪类数据库，微生物的命名都会随时更新曾经分类错误的数据，包括分类地位的改变或者命名的改变。

7.6 数据代表性

选取代表性基因组，可选择不同血清型、基因型的代表基因组。对于病毒数据，由于序列较短，且病毒分型较为复杂，在选取参考基因组时，应尽量全面地纳入代表性数据，甚至应包括部分不完整但能提供分型特征基因序列，从而实现更加准确的鉴定。对于近缘物种，选取代表性基因组无法代表种内所有的基因特征，或无法实现与近缘种的区分，可以考虑通过构建进化树或基于基因组相似性进行聚类分

析，选取不同进化分支或聚类的代表性基因组；或者可以构建种内共性及特异性基因的泛基因组集，通过比对的方式来代替单个代表性基因组的选择。

8 数据库性能验证和评价要求

8.1 数据产生

采用序列模拟技术生成数字参考品，构建数据库建设评估工具；通过统计对比和标准流程分析，建立多样性、基因组质量、序列一致性与同源序列等质量指标与分析性能的映射关系，评估数据库性能。

8.2 抗干扰评价

数据库的性能评价应当从罕见物种的准确检出、低丰度物种的准确检出、近源微生物同源干扰、污染序列的检出等角度评价数据库的复杂度、序列一致性、完整度、抗干扰性能、灵敏度、特异性、抗干扰性能等。

9 数据库更新要求

9.1 更新条件

数据库应保持定期更新，更新的内容包括与公共数据来源保持同步更新，持续新增代表性序列。及时更新数据库中已有记录的内容，如参考序列的物种分类地位发生变化时，应及时更新。及时去除经验证后出现错误的数据，如分类状态错误或者其他信息存疑数据。

9.2 更新频率

数据库更新的频率应考虑数据内容的变化，对于数据内容变化较为频繁的数据库，及时提高更新频率，对于数据内容变化较少的数据库，保持适当更新频率，保持数据库稳定性。

9.3 更新验证

每次更新数据库后，及时验证数据库各项性能指标。

10 应用数据库评价现有产品

利用病原基因组参考数据库结合数字参考品和标准化生物信息学分析流程，对申报产品中的干实验部分进行性能评价。

11 数据库通用程序和原则

11.1 数据库纠错

应采用程序来检测当前数据库中新数据和预先存在的数据中的错误，以提高质量和一致性。

11.2 文件格式规范

文件的日期和时间格式应符合ISO 8601和ISO 20387，第7.1.3条。

注：日期可以表示为YYYY-MM-DD(例如2018-04-25)，时间可以表示为hh: mm: ss(例如04:26:55)。

11.3 数据交换和集成

应按照数据交换和集成的相关标准采用标准格式分发微生物相关（元）数据，以确保数据共享的可信赖性和及时性，以及提供人类和机器可读格式的数据的标准化协议，以促进与各种生物信息学工具之间的互操作性。

11.4 同物种数据接口

用于同一物种的所有名称的清单，例如本地名称、同义词、通用名称、误用名称和拼写错误，应记录为元数据，并且用户可以轻松访问。

11.5 数据唯一标识

应确保微生物的（元）数据可通过适当通信协议采用唯一标识符进行检索，并确保（元）数据包括对其他相关（元）数据的合格引用。

附 录 A
（规范性）
推荐的数据字段

1 物种名称和分类

a) 名称：

数据类型：字符串或链接

说明：当前科学物种包含属名和种名，或者至少分类名称显示最低的分类范畴，可以确定目前，按照 4.2.2 中指定的要求，也可以包含一个链接到授权分类数据库供参考。作为复合种的微生物材料例外。

示例 1：红酵母属

示例 2：pFBAOT6

示例 3：芽胞杆菌属

1）亚种名称：

数据类型：字符串或链接

说明：包含亚种加词和（如果与种加词不同）作者。

2）亚种下名称：

数据类型：字符串或链接

说明：包含变种、名称、加词、作者和参考。不适用于所有菌株。

3）变种名称：

数据类型：字符串或链接

说明：变种名称，作者。

4）曾用名称：

数据类型：字符串或链接

说明：名称，作者。包含变型加词和（如果不同于种加词）。

5）特殊变型名称：

数据类型：字符串或链接

说明：名称，作者，包括专化型加词和作者。

6）误用名称：

数据类型：字符串或链接

说明：名称，作者。误用于当前物种的名称，也可能正确地应用于另一物种。

b) 命名史：

数据类型：文本或链接

说明：分类名，日期，修改的原因，以及谁做了修改。

c) 分类水平：

数据类型：受控词汇

说明：微生物分类的所有分类层次，包括界、门、纲、目、科、属、种、亚种和所有其他中间层。

d) 分类 ID：

数据类型：字符串或链接

说明：一组分类单元信息的标识符。

e) 血清型：

数据类型：字符串

说明：血清型名称和作者，这是一些医学上重要的物种所必须的信息。

2. 微生物类型：

数据类型：命令，受控词汇

说明：命令值可以是真菌、酵母、细菌、古生菌、微藻、蓝藻、原生动物、质粒、噬菌体、病毒、cDNA

示例：酵母

3. **生物安全与生物安保：**

数据类型：命令，字符串

说明：根据致病性等级及病原微生物危害分类(世卫组织标准和/或国家标准)的相关规定，对内部政策的注意事项和对菌毒种的分配、进口和出口的限制。

示例 1：二级生物安全防护实验室

示例 2：国家生物安全标准 3 级

示例 3：与植物免疫有关

示例 4：仅在生物危害层流箱内转移

4. 采集国家：

数据类型：受控词汇

说明：获得微生物资源的国家。

5. 访问权限：

数据类型：字符串

说明：有关谁可以访问资源或其安全状态指示的信息。访问权限可包括基于隐私、安全性或其他策略的访问或限制信息。

6. 分离基质：

数据类型：字符串

说明：分离出微生物样本的基质（例如土壤、水、血液、叶子等）。基于同源词库的宿主植物/动

物的名称。

示例 1：被污染的石油

示例 2：植物残留物

7. **生境：**

数据类型：字符串

说明：表示所发现物种的群落生境。空间区域或命名位置。该字段包含纬度、经度、海拔和深度的位置信息。物理因素，如湿度、温度范围、pH 值和环境光强度，以及生物因素，如食物的可用性和捕食者的存在与否。

示例 1：山毛榉林

示例 2：水域

1）**纬度：**

说明：以小数形式输入纬度地理坐标，不要以度数、分和秒的格式输入。

2）**经度：**

说明：以小数形式输入逻辑地理坐标，不要以度数、分和秒的格式输入。

3）**海拔：**

说明：海拔是地球表面高于海平面与采样位置之间的垂直距离。

4）**深度：**

说明：深度定义为局部表面以下的垂直距离，例如，对于泥沙或土壤样品，深度分别从泥沙或土壤表面测量。深度可以记录为地下样品的间隔。

8. **收集者：**

数据类型：字符串

说明：收集者的姓名。

9. **收集日期：**

数据类型：日期

说明：采集微生物资源的日期，并使用格式[年]-[月]-[日] （如适用）

示例 1：1988-09-21

示例 2：1989 年前

10. **基因型和遗传**

a) **基因型：**

数据类型：字符串

说明：菌毒株染色体标记的名称。特别推荐用于存在许多转基因菌毒株的菌毒株。

b) **基因组入藏号：**

数据类型：字符串和链接

说明：序列表模式，包括序列类型、登录号（例如 NCBl EMBL/DDBI 号码）以及到公开可用数据库的链接。

c) **基因组大小：**

数据类型：字符串

说明：以 kbp 为单位（如果完整，则报告总基因组大小，包括所有质粒和微染色体；如果不完整，则报告测序延伸的大小）

d) **基因组序列：**

数据类型：文件

说明：通常以 fasta 文件进行存储

参 考 文 献