**向抗病毒产品部门提交下一代测序数据**

**行业指南**

技术规范文件

如对本技术规范文件有任何疑问，请发送电子邮件至cder-edata@fda.hhs.gov联系CDER

**美国卫生与公众服务部**

**美国食品药品监督管理局**

**药品评价与研究中心（CDER）**

**2019年7月**

**技术规范**

**修订历史**

|  |  |  |
| --- | --- | --- |
| **时间** | **版本** | **修订版本汇总** |
| 2019年7月 | 1.0 | 初始版本 |
|  |  |  |
|  |  |  |
|  |  |  |

**目录**

[1.0 引言 1](#_Toc91865457)

[2.0 可接受的NGS平台 2](#_Toc91865458)

[3.0 提交至该部门的信息概要 2](#_Toc91865459)

[3.1 NGS方案和数据分析方法 2](#_Toc91865460)

[3.2 Fastq格式的原始NGS数据 2](#_Toc91865461)

[3.3 频率表 2](#_Toc91865462)

[3.4 汇总表 3](#_Toc91865463)

[3.5 其他分析 - 就绪数据集（如适用） 3](#_Toc91865464)

[4.0 NGS方案 3](#_Toc91865465)

[4.1 通用方案设计要素 3](#_Toc91865466)

[4.2 样品制备 4](#_Toc91865467)

[5.0 NGS数据分析方法描述和结果报告 4](#_Toc91865468)

[5.1 每次序列运行的统计汇总数据 4](#_Toc91865469)

[5.2 对序列条形码处理方式的描述 5](#_Toc91865470)

[5.3 重叠群和映射报告 5](#_Toc91865471)

[5.3.1 序列读段映射到参考序列和变异检测 5](#_Toc91865472)

[5.3.2 报告读段映射和变异检测的结果 5](#_Toc91865473)

[5.3.3 重叠群的新基因组组装和变异检测（如适用） 6](#_Toc91865474)

[5.3.4 报告新的基因组组装和变异检测的结果 6](#_Toc91865475)

[6.0 NGS文件类型和提交程序 7](#_Toc91865476)

[6.1 通过移动硬盘提交的文件 7](#_Toc91865477)

[6.1.1 移动硬盘中的内容 7](#_Toc91865478)

[6.1.2 标记fastq序列文件 7](#_Toc91865479)

[6.2 通过CDER电子文件网关提交的文件 8](#_Toc91865480)

[7.0 频率表示例 8](#_Toc91865481)

**向抗病毒产品部门提交下一代测序数据**

行业指南
技术规范文件1

|  |
| --- |
| 本指南代表美国食品药品监督管理局（FDA或本机构）目前关于该主题的思考。其不会为任何人创造或赋予任何权利，也不会对FDA或公众产生约束。如果替代方法满足适用的法律法规的要求，则可以使用该方法。如需讨论替代方法，请联系标题页负责实施本指南文件的FDA工作人员或办公室。 |

**1.0** **引言**

本技术规范文件的目的是提供FDA抗病毒产品部门（以下简称为“该部门”）当前对提交下一代核苷酸序列分析程序和数据以支持抗病毒药品开发的耐药性评估的看法。

该部门对与开发的抗病毒药品相关的所有耐药性数据进行了独立分析，旨在确保耐药性的出现在新批准的抗病毒药品的标签中予以详细描述和解释。提供准确的耐药性信息对于保护公众健康至关重要，可防止出现新的耐药和交叉耐药病毒变异体，这些变异体有可能感染他人，并导致无法通过获批的药品予以控制的重大疾病暴发。此外，耐药性信息为负责监督抗病毒药品的使用的医疗服务专业人员提供了重要指导，包含在该部门批准的药品信息中。此外，该部门还可要求提供宿主靶向基因的核苷酸序列分析数据，以进行多态性分析，从而确定基于不同群体的等位基因是否对疗效有影响。

下一代测序（NGS）是一种申请方在进行基于序列的耐药性分析时经常采用的新兴技术。由于该技术可生成临床样品中所有RNA或DNA的核苷酸序列，NGS增加了耐药性分析过程的复杂性。与提供病毒群平均序列的Sanger法测序核苷酸序列分析相比，NGS可提供病毒群中单个病毒的核苷酸序列信息，每个样品通常提供数百万或数十亿个短序列。数据的复杂性对审查员分析和验证序列信息构成挑战，特别是由于目前没有标准化的生物信息学分析方法可用于分析这些大型数据集。

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

1 本技术规范文件由美国食品药品监督管理局药品评价与研究中心抗病毒产品部门编制。可随时提交对本指南的意见。将意见提交至编号为FDA-2017-D-6821的备案文件中（可登录以下网址获取：https://www.regulations.gov/docket?D=FDA-2017-D-6821）（参见备案文件中有关提交意见的说明）。

FDA指导性文件，包括本指南在内，不具有法律强制责任。相反，指南表明了该机构目前关于该主题的思考，除非引用具体的法规或法律要求，否则只应视为建议。在本机构指南中使用词语“应”是指建议或推荐进行某一事项，并非强制要求。

**2.0** **可接受的NGS平台**

申请方提交测序平台的适当详细信息、样品制备方案、fastq格式的原始NGS数据以及用于分析数据的方法后，该部门将接受大多数标准NGS平台生成的核苷酸测序数据。FDA建议申请方在提交测序数据之前尽早与该部门沟通并提供上述详细信息。此外，FDA建议申请方在任何正式提交之前提交一个模拟NGS数据集，以确保数据格式和过程适当，并且可接受。准备NGS提交时，申请方应将以下章节中的信息考虑在内。

**3.0** **提交至该部门的信息概要**

**3.1** **NGS方案和数据分析方法**

申请方应提交一份详细描述样品处理和NGS分析程序的方案。有关所述的具体方法，请参见第4.0节“NGS方案”和第5.0节“NGS数据分析方法描述和结果报告”。

**3.2** **Fastq格式的原始NGS数据**

申请方应提供每次序列运行中fastq格式的所有原始NGS数据。组合读取映射可选用.fas、.ace、.sam或.bam格式提交。申请方应提供适当的参考序列（经部门同意）以及用于任何参考映射的登录号。对于对基线样品或目的基因的参考映射，申请方应提供基线或参考共识序列，并说明该序列是如何推导出来的。对于新的组装基因组组装，申请方应提供所有fastq格式的大于200个核苷酸的重叠群。

**3.3** **频率表**

申请方应提供频率表，报告频率大于或等于1%时与基线不同的所有氨基酸置换。关于频率表的示例，请参见第7.0节“频率表示例”。

**3.4** **汇总表**

申请方应将以下几点考虑在内：

a. 对于每项研究，以分析就绪病毒学耐药性数据格式提供用于基于群体的测序的数据，显示基线和失效序列，并在氨基酸与参考序列相同时使用空白序列。FDA建议申请方咨询该部门，确定耐药性数据集的敏感性水平（表中填充了以特定频率（即15%）存在的氨基酸置换），或提供所选敏感性水平的合理性说明。

b. 提供表格，列示每项研究或每个研究亚组（即按基因型、按剂量）主要置换的详细汇总。

**3.5** **其他分析 - 就绪数据集（如适用）**

FDA建议申请方在提交与NGS数据耐药性分析相关的其他分析 - 就绪数据集之前咨询该部门。

**4.0** **NGS方案**

**4.1** **通用方案设计要素**

申请方应在NGS方案中纳入以下通用方案设计要素：

a. 对所分析受试者、研究时间点和样本矩阵的描述。

b. 对所用NGS平台的描述，包括所有相关的性能特征。

c. 所分析的靶基因区域名称和大小。

d. 一般分析策略（例如识别相对于原型菌株的变化，比较同一受试者不同时间点的序列）。

e. 所尝试的覆盖水平。FDA建议覆盖目标为大于5,000次读取。但是，本机构认识到，一些病毒载量较低的样品可能不会在该覆盖水平下形成组装体。申请方应标识未达到该覆盖水平的样品。

f. 对用于识别、过滤或处理测序错误的方法的描述。

**4.2** **样品制备**

获取可靠测序结果的关键是模板的制备，因此测序的样品代表所分析的群体。应描述以下信息：

a. 从样品中提取核酸的方法。

b. 从污染的背景核酸中纯化病毒序列的方法。

c. 浓缩病毒核酸的方法。包括每个样品的逆转录聚合酶链式反应（RT-PCR）（病毒RNA）或PCR（病毒DNA）反应的估计目标基因拷贝数输入。

d. 使二级结构变性的方法。

e. 产生双链DNA（dsDNA）的方法。包括对引物的描述。

f. 用于测序的dsDNA纯化方法。

g. NGS文库制备方法。

h. 添加用于多路复用的条形码的方法（如适用）。

**注：**许多已在科学文献中发表的NGS方案利用特异性（扩增子）或非特异性（随机引物）PCR扩增阶段来增加用于测序的dsDNA的浓度。尽管该方法将扩增主要基因型，但不太可能检测到随时间推移对耐药性突变的增加起重要作用的微小变异。因此，任何在NGS之前执行PCR扩增步骤的方案均应提供证据，证明扩增代表目标群体，并且NGS数据中仍存在微小变异。建议采用校正PCR重采样偏倚和（RT-）PCR及测序误差的方法，如互补DNA条形码技术。

**5.0** **NGS数据分析方法描述和结果报告**

提交的序列数据必须包括对用于分析测序数据集的分析流水线以及原始序列信息的全面描述，以便该部门能够对数据进行独立分析。应在这些报告中提供以下章节中的信息。

**5.1** **每次序列运行的统计汇总数据**

报告应包括每次序列运行测序的读取总数、读取的质量评分、读取的平均长度、对为减少平台特定错误的影响而执行的任何错误评估的描述，以及用于清除潜在序列错误或删除不满足任何质量或长度阈值读取的修剪参数相关信息。

**5.2** **对序列条形码处理方式的描述**

报告应包括对用于将条形码序列添加到相关序列的方法的描述、用于在序列运行后隔离读取的条形码标签的序列、用于按条形码将读取分类到bin文件中的程序、对序列中的条形码序列修剪方式的描述，以及对为消除潜在交叉条形码污染而采取的任何预防措施的描述。

**5.3** **重叠群和映射报告**

有两种方法可用于数据的NGS分析以支持抗病毒药品的开发：（1）将短序列读取映射到参考序列，或（2）对短序列读取进行新的组装基因组组装，以拼接重叠群的序列。

**5.3.1 序列读段映射到参考序列和变异检测**

申请方应考虑将以下几点纳入有关序列读取映射到参考序列和变异检测的报告中：

a. 识别用于映射的参考序列并提供核苷酸和氨基酸序列以及登录号。

b. 识别程序并描述用于将读取映射到参考序列的算法，并提供所用参数设置列表以及每种参数设置的依据。报告应提供有关算法的具体信息，包括但不限于：允许的错配百分比、允许的插入和删除（indel）数目以及同一性和相似性截止值。

c. 描述用于调用变体的算法并提供所用的参数。此外，描述如何区分单核苷酸多态性（SNP）和插入删除与序列错误。

**5.3.2 报告读段映射和变异检测的结果**

申请方应考虑将以下几点纳入包含读取映射和变异检测的结果的报告中：

a. 提供每个读段映射所达到的覆盖范围的概要，并纳入覆盖图。

b. 提供每个读段映射中存在的序列数。

c. 提供所有读段映射的频率表，显示主要SNP和插入删除。

d. 以氨基酸格式提供每个临床试验的NGS结果的组合频率表，包括以下内容：每个位点的氨基酸、每个位点的变异体频率，以及发生变异的每个位点的覆盖率（参见第7.0节“频率表示例”中的示例）。

e. 提供汇总表，显示相关群体的主要变化。

**5.3.3 重叠群的新组装基因组组装和变异检测（如适用）**

申请方应考虑将以下几点纳入有关重叠群的新组装基因组组装和变异检测的报告中：

a. 提供重叠群统计数据，显示通过读段生成的重叠群数目。

b. 识别程序并描述用于对短序列读段进行新的组装基因组组装，以拼接重叠群序列的算法，并提供所用参数设置列表以及每种参数设置的依据。报告应提供有关算法的具体信息，包括但不限于：允许的错配百分比、允许的插入和删除数目以及同一性和相似性截止值。

c. 描述用于比较重叠群和调用变体的算法并提供所用的参数。此外，描述如何区分SNP和插入删除与序列错误。

d. 描述如何通过标识用于识别与已知序列相似性的程序对重叠群进行注释，识别查询的数据库，提供所用的检索标准，定义同一性百分比的截止值，并提供如何识别和报告污染物或其他病毒的详细信息。

**5.3.4 报告新的组装基因组组装和变异检测的结果**

申请方应考虑将以下几点纳入有关新的组装基因组组装和变异检测的报告中：

a. 提供每个重叠群的覆盖范围概要。

b. 提供每个重叠群中的序列数和未形成重叠群的读段数。

c. 描述注释期间检测到的污染物或其他病毒，并报告每种污染物的百分比。

d. 提供所有主要重叠群的频率表，显示主要SNP和插入删除。

e. 以氨基酸格式提供每个临床试验的NGS结果的组合频率表，包括以下内容：每个位点的氨基酸、每个位点的变异体频率，以及发生变异的每个位点的覆盖率（参见第7.0节“频率表示例”中的示例）。

f. 提供汇总表，显示相关群体的主要变化。

g. 建议对所有未映射到参考序列的读段进行新的组装基因组组装并注释重叠群，以评估样品的潜在交叉污染。

**6.0** **NGS文件类型和提交程序**

**6.1** **通过移动硬盘提交的文件**

fastq格式的原始NGS数据和频率表应按照技术规范文件《利用eCTD规范传输电子提交材料》（2019年4月）中概述的指导原则，通过安全的移动硬盘发送给该部门。2

**6.1.1 移动硬盘中的内容**

请注意，只有原始NGS数据、频率表和目录在硬盘上提交。其他文件，如扩展名为“.exe”的文件，可能导致提交被拒绝。

如果硬盘有密码保护（此时不需要或不推荐），申请方应提前咨询该部门，以确保将密码提供给药品评价与研究中心（CDER）档案室的相关人员。

**6.1.2 标记fastq序列文件**

所有NGS数据均应以fastq格式提供，其中包括核苷酸序列和质量信息。这应包括每例受试者在不同时间点的NGS序列运行，包括基线和接近观察到治疗失败的时间点。申请方应提供标有受试者唯一标识符和时间点的fastq文件。例如，对于研究ABC123、受试者0001和3个样品（基线、第16周和随访第4周），文件应标记如下：

a. ABC123-0001.baseline.fastq

b. ABC123-0001.WEEK16.fastq

c. ABC123-0001.WEEK4FU.fastq

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

2 可登录以下网址获取：https://www.fda.gov/media/76812/download。另请参见电子监管申请和审查网页（网址：https://www.fda.gov/Drugs/DevelopmentApprovalProcess/FormsSubmissionRequirements/ElectronicSubmissions/d efault.htm）。

如果对样品进行重新测序，则第一次和第二次序列运行区分如下：

a. ABC123-0001.baseline-1.fastq

b. ABC123-0001.baseline-2.fastq

**6.2** **通过CDER电子文件网关提交的文件**

应通过CDER电子文件网关提交NGS方案、分析就绪耐药性数据集和任何其他支持信息。

**7.0** **频率表示例**

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **STUDYID** | **USUBJID** | **NGSPL** | **VISIT** | **AAPOS** | **AAREF** | **AASUB** | **TCOV** | **VCOV** | **AAFREQ** |
| ABC123-999 | 0123 | Illumina | BL | 81 | R | K | 4317 | 156 | 0.036 |
| ABC123-999 | 0123 | Illumina | BL | 98 | K | R | 2841 | 99 | 0.035 |
| ABC123-999 | 0123 | Illumina | 第2天 | 98 | K | R | 9487 | 366 | 0.039 |
| ABC123-999 | 0123 | Illumina | 第3天 | 98 | K | R | 9474 | 378 | 0.040 |
| ABC123-999 | 0123 | Illumina | BL | 120 | R | Q | 4310 | 200 | 0.046 |
| ABC123-999 | 0123 | Illumina | 第2天 | 120 | R | Q | 12722 | 470 | 0.037 |
| ABC123-999 | 0123 | Illumina | 第3天 | 120 | R | Q | 12466 | 489 | 0.039 |
| ABC123-999 | 0123 | Illumina | BL | 147 | I | V | 3456 | 742 | 0.215 |
| ABC123-999 | 0123 | Illumina | 第2天 | 147 | I | V | 13456 | 2709 | 0.201 |
| ABC123-999 | 0123 | Illumina | 第3天 | 147 | I | V | 13297 | 1934 | 0.145 |
| ABC123-999 | 0123 | Illumina | BL | 150 | A | V | 3107 | 43 | 0.014 |
| ABC123-999 | 0123 | Illumina | 第2天 | 150 | A | T | 13116 | 167 | 0.013 |
| ABC123-999 | 0123 | Illumina | BL | 154 | K | R | 2987 | 124 | 0.042 |
| ABC123-999 | 0123 | Illumina | 第2天 | 154 | K | R | 13434 | 1350 | 0.101 |
| ABC123-999 | 0123 | Illumina | 第3天 | 154 | K | R | 13077 | 1206 | 0.092 |
| ABC123-999 | 0123 | Illumina | 第3天 | 155 | R | K | 12459 | 9837 | 0.781 |
| ABC123-999 | 0123 | Illumina | 第3天 | 156 | P | S | 13385 | 172 | 0.013 |
| ABC123-999 | 0123 | Illumina | BL | 186 | V | I | 6155 | 129 | 0.021 |
| ABC123-999 | 0123 | Illumina | 第2天 | 186 | V | I | 17698 | 269 | 0.015 |
| ABC123-999 | 0123 | Illumina | 第3天 | 186 | V | I | 16474 | 460 | 0.028 |
| ABC123-999 | 0123 | Illumina | BL | 206 | K | H | 9698 | 165 | 0.017 |
| ABC123-999 | 0123 | Illumina | 第2天 | 206 | K | R | 24601 | 292 | 0.012 |
| ABC123-999 | 0123 | Illumina | 第3天 | 210 | S | N | 23001 | 255 | 0.011 |
| ABC123-999 | 0123 | Illumina | 第3天 | 254 | H | R | 25145 | 290 | 0.012 |

**STUDYID** =研究方案编号；**USUBJID** =受试者唯一ID；**NGSPL** = 用于测序的下一代测序平台；**VISIT** =期间采集样品的研究访视；**AAPOS** =氨基酸在靶基因中的位置；**AAREF** = 参考序列中该位置的氨基酸；**AASUB** =通过测序检测到的氨基酸置换；**TCOV** =核苷酸位点的总覆盖率；**VCOV** =变体核苷酸位点的总覆盖率；**AAFREQ** =检测到的置换频率；**BL** =基线。