**《微生物超低频突变测定 双重测序法》国家标准**

**编制说明**

（征求意见稿）

一、任务来源

本国家标准的制定任务列入国家标准化管理委员会《国家标准委关于下达2018年第二批国家标准制修订计划的通知》（国标委综合〔2018〕41号），项目编号“20181040-T-424”。该标准由中国标准化研究院提出并归口，由清华大学等单位承担该标准的制定任务。

二、标准制定目的和意义

随着二代测序技术的发展，单位测序成本越来越低，各领域仰赖其高通量的特性获取基因序列信息，结合不同分析手段，获得更多发现，突破科研瓶颈，并逐步将前沿技术带入社会大众的生活当中，拓展应用价值。

然而，对于单碱基差异便影响巨大的工作，包括但不限于疾病研究，二代测序拥有足够的通量却未能达到需求的测序精度，当待测文库的突变率低于一定数值时，便难以常规的测序方法准确检测。许多技术围绕此痛点展开，并取得良好的成果。此标准的编制旨在规范提高测序精度的工作方法，以更好地推动相关领域的发展。

（一）测序精度的提高

二代测序的在样本制备、测序、和分析步骤，会引入约10-2~10-4的突变误差，其可能来自于建库中PCR过程的保真性、人为操作或机器检测的偏差等，当待测DNA文库中的突变率小于10-4时，便无法以现有平台正确识别。互补标签的设计，有助于在不影响通量的情形下，排除引入的误差，还原真实的变异位点，提高测序的精度。

（二）适配于现有测序平台

双重测序法透过建库过程的设计和数据分析的模型建立来达到增加准确性的目的，完全适配于现有测序平台，无需过多的调整变动。

三、标准制定原则和依据

**（一）标准编制原则**

本标准参考了其他同类国家标准的内容，同时充分考虑我国超低频突变测定的使用经验和要求。

**（二）标准制订主要依据**

1、标准编写遵循GB1.1-2009《标准化工作导则第1部分：标准的结构和编写规则》的有关要求。

2、标准编写内容参考的相关标准，包括：

GB/T 6682-2008 分析实验室用水规格和试验方法

GB/T 30989-2014高通量基因测序技术规程

**四、标准**主要技术内容

**（一）标准适用范围说明**

本标准规定了双重测序法的建库操作、样品质量的控制及测序数据的处理模型；超低频突变检测的仪器设备及器具、主要试剂、操作步骤和结果分析。

**（二）内容提要**

其技术原理为通过DNA双链体的两条链进行独立标记和测序。在双螺旋的DNA片段的两条链，接上一段12nt随机且互补的双链核苷酸序列。后续透过文库制备及二代测序，从结果信息进行对比。因为DNA双链体的两股练是互补的，突变会同时存在互补家族的同一个位置上，排除掉制备过程中所造成的变异，正确侦得实际的突变位点。

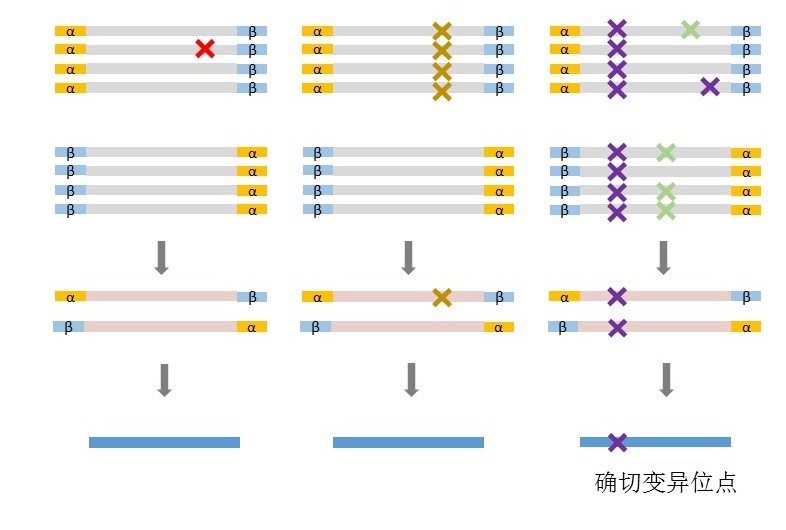


图 1、双重测序法作用原理图

**五、主要工作过程**

2016年接到编制任务，开展资料收集和分析工作。研究了GB/T 6682 《分析实验室用水规格和试验方法》、GB/T 30989-2014《高通量基因测序技术规程》。了解二代测序的相关要求，同时收集查阅了相关资料。为了使编制的标准能更好的应用于超低频突变的检测，在2016年10月提出了标准讨论提纲，确定了编制原则和主要内容。该标准于2017年10月在全国标准信息公共服务平台上进行公示，于2018年2月召开专家会，听取专家意见。

**六、方法验证及结果**

**（一）接头合成验证**

双重测序法的接头合成完成之后，以放射性同位素及PAGE验证结果如图 2，以确保后续建库测序能够顺利以双重测序法进行。



图 2、(**a**) 14％（wt / vol）聚丙烯酰胺凝膠图。

各泳道分别对应《超低频突变侦测 双重测序法》中8.1各步骤。

泳道1：退火后接头；泳道2：延伸后接头；泳道3：酶切后接头；泳道4：最终的接头。

(**b**)各步骤接头示意图。数字和字母标记分别对应**a**中各泳道及各条带。

**（二）建库过程的样品验证**

在进行PCR前，须先对样品进行长度定量，连接反应可能导致多峰值的产生，如图 3中的双峰是很常见的，峰的分布可能会有多种状况，此将不影响后续的测序准确度。

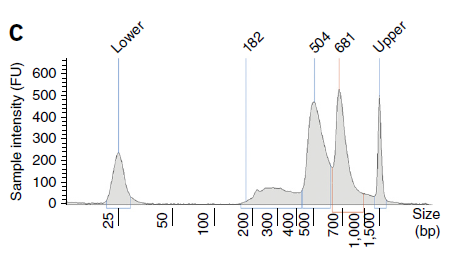


图 3、PCR反应前对样品进行长度定量，数据由Agilent TapeStation 2200 测得。

另外，对于DNA样品和双重测序法接头的添加比例请务必参考《超低频突变频率检测方法：双重测序》中规范，若浓度失衡将导致长度分布混乱，所测得数据质量低，而不可用于后续的分析。图 4为接头浓度过高，导致接头本身互相连接之情况。

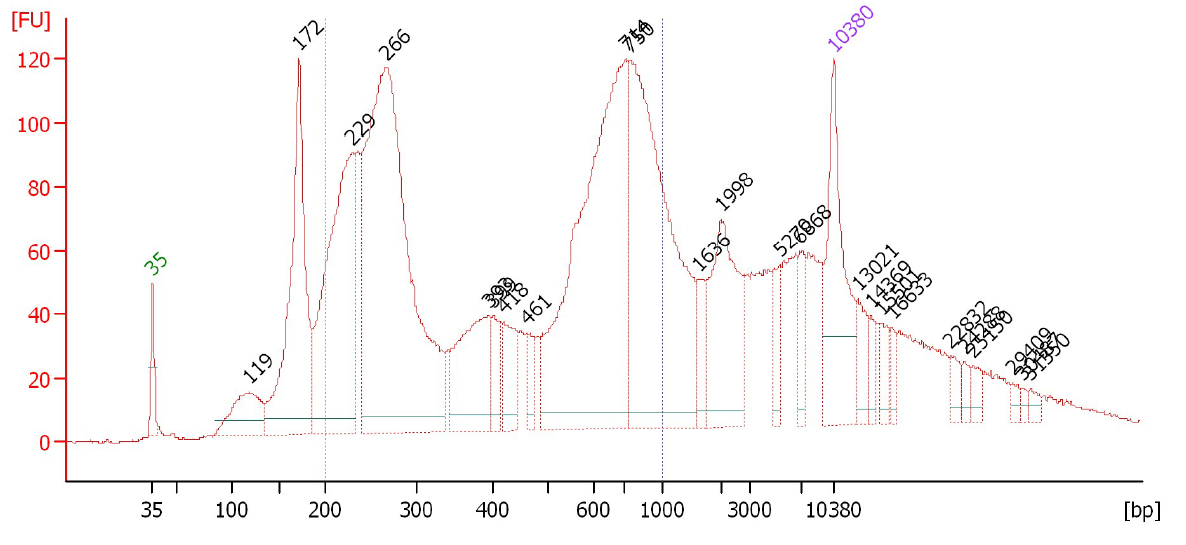


图 4、双重测序法接头与待测片段浓度不当导致长度分布混乱。数据由Bioanalyzer

2100测得。

样品正式上机测序前，对DNA片段进行长度定量，推测操作过程中PCR轮次是否需要增减，以确保后续数据质量，下图数据对应《超低频突变频率检测方法：双重测序》中8.3之步骤结果。若使用单一酶切片段，则确认无多余峰值产生，如图 7所示。

图5、理想的定量后结果，数据由Agilent TapeStation 2200 测得。

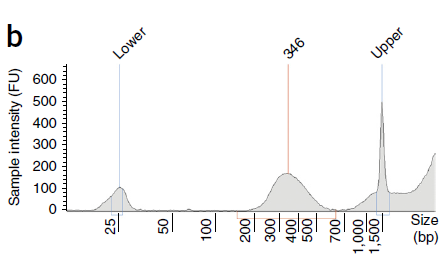
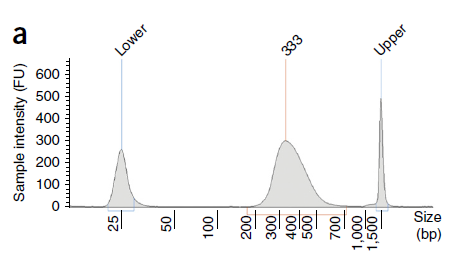


图 6、PCR轮次过度造成较多大片段的产生，需调整PCR轮次，数据由Agilent TapeStation 2200 测得。

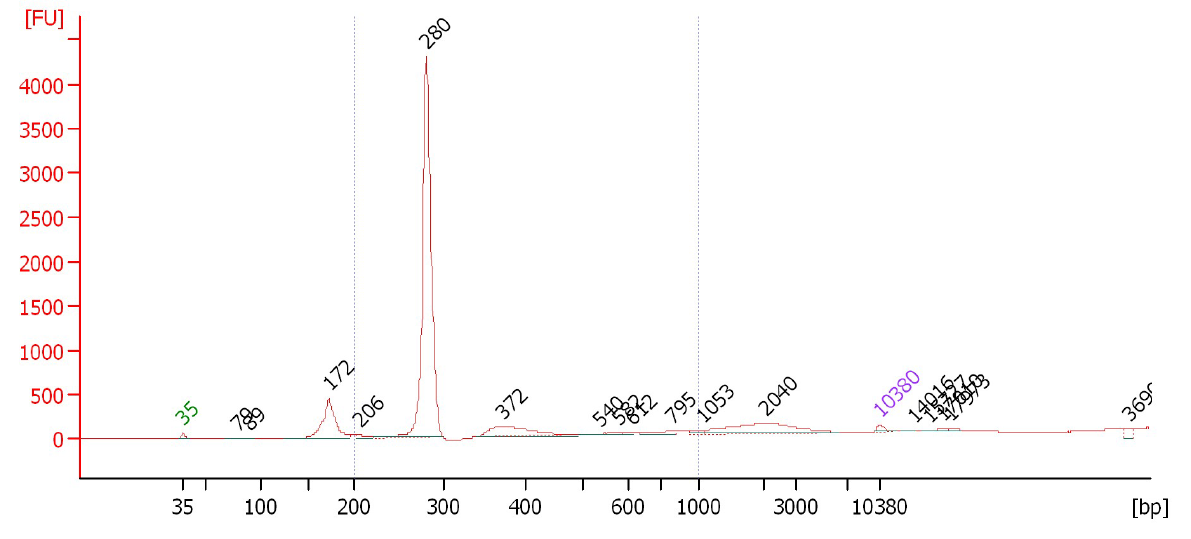


图 7、酶切片段长度116 bp经连接和PCR后的长度定量结果，数据由Bioanalyzer

2100测得。

**（三）测序结果分析**

获得测序的原始数据(raw data)后，需评估数据质量，若为Illumina平台注意Q20代表被识别错误机率为1%，即正确率是99%，其值不可小于90；Q30代表被识别错误机率为0.1%，即正确率是99.9%，其值不可小于80。

表 1为图 7之样品经Illumina平台测序后，分析得到的质量数据；序列准确性如表 2。

表 1、图 7样品经Illumina平台测序分析质量结果

|  |  |  |
| --- | --- | --- |
|  | Read 1 | Read 2 |
| Q20 | 94.74 | 92.52 |
| Q30 | 92.2 | 88.99 |

表 2、图 7样品经Illumina平台测序结果(样品为单一序列片段，无突变库混合)

|  |  |
| --- | --- |
| Mutation Site |  |
| 0 | 99.88% |
| 1 | 0.12% |
| 2 | 0.0004% |

分析测序数据得到标签家族(tag family)的大小分布，亦即为各组标签的读数(reads)。若所得结果偏低，如图 9所示，代表太少的读数共用一组标签(tag)，难以对比找出突变位点；若偏高如图 10，则不能有效产生更多双链同源序列(DSCs)，浪费了测序的数据容量。而图 6样品分析标签家族大小的整体分布结果如图 10所示。

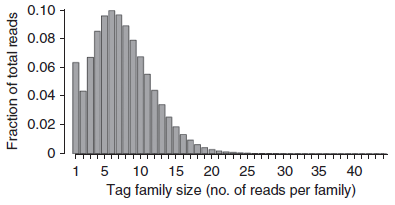


图 8、理想的标签家族(tag family)分布图。

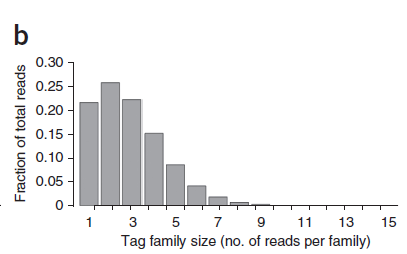


图 9、标签家族(tag family)的分布偏小，可能是进行PCR的样品太少所致。

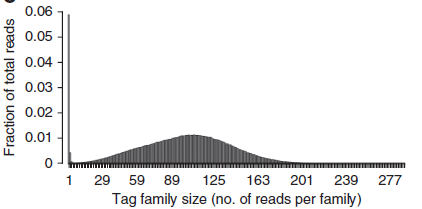


图 10、标签家族(tag family)的分布偏大，可能是进行PCR的样品太多所致。

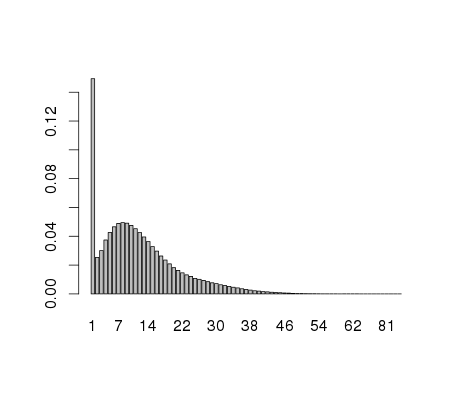


图 11、样品测序数据分析标签家族(tag family)大小的整体分布结果

在得到标签家族(tag family)的大小分布后，分析峰族(peak family)的大小。由于每个标签家族的读数不同，且存在一定的分布性，以峰族大小(peak family size)作为后续双链同源序列(DSCs)的数据指标。峰族大小(peak family size)定义为大于1的标签家族所拥有最高比例的读数。图 12和图 13以图 8数据，进行峰族大小分析。

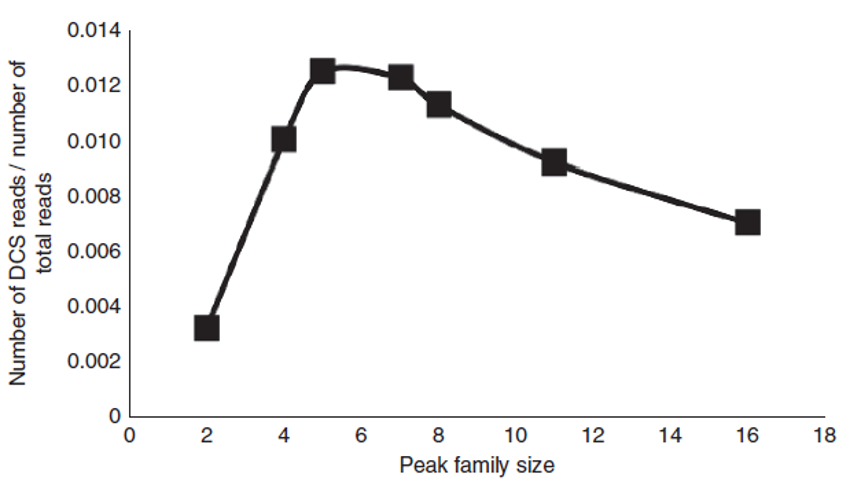


图 12、有效产生双链同源序列(DSCs)数据的峰族大小。在峰族大小为6时，能最高效的产生双链同源序列(DSCs)，此时每个双链同源序列对应的原始数据读数约为40。

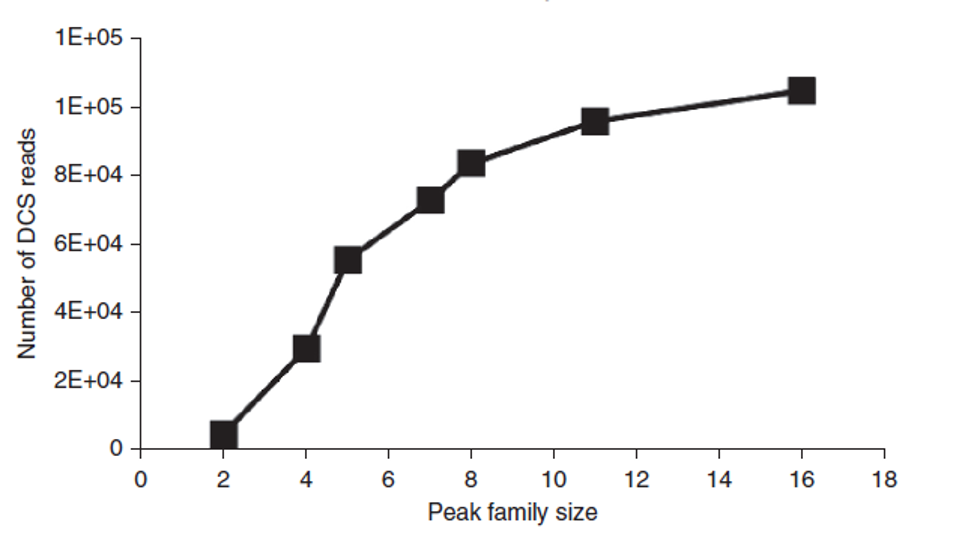


图 13、在峰族大小低于16之前，双链同源序列(DSCs)的数量持续增加；当峰族大小大于16之后，则双链同源数列的数量无明显增加。

**（三）突变率识别分析**

以带有LacZ α片段的M13mp2质粒制备突变库，对其做单位点的突变设计，按比例混合两者，后转入大肠杆菌进行蓝白筛选验证，并以Illumina HiSEq 2000同时进行常规测序及双重测序法。从图 14可看出当突变率低于10-2时，常规的测序方法开始出现很大的偏差，而双重测序法在突变率达10-4时，皆能准确识别突变。

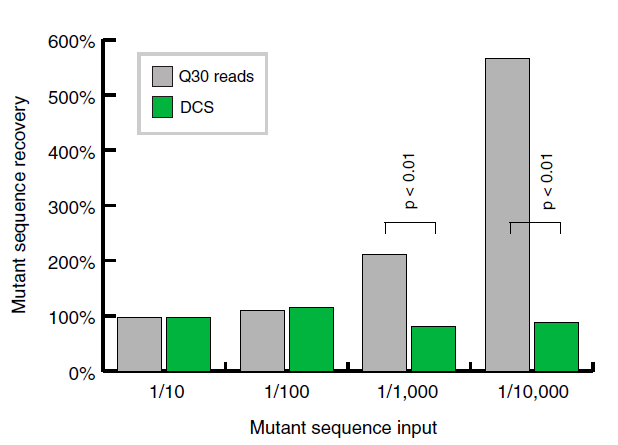


图 14、将已知的单点突变M13mp2与野生M13mp2按比例混合，分析常规测序及双重测序法的数据结果。测序深度约20,000x。

图 15为混合样品比例达时，对比常规方法及双重测序法经Illumina HiSeq 2000测序所得的突变率检测结果。常规方法所分析得到的突变率为，比样品库真正的突变率高了1000倍，暗示一般的测序方法有的突变识别是有误的。比对单链同源序列得到的突变率结果，代表99%的测序偏差已被修正，但所得结果仍比样品库的突变率高了10倍，有约90%的突变识别是不准确的。而双重测序法经双链同源序列分析后，所得基本与样品本身一致，分析结果为。

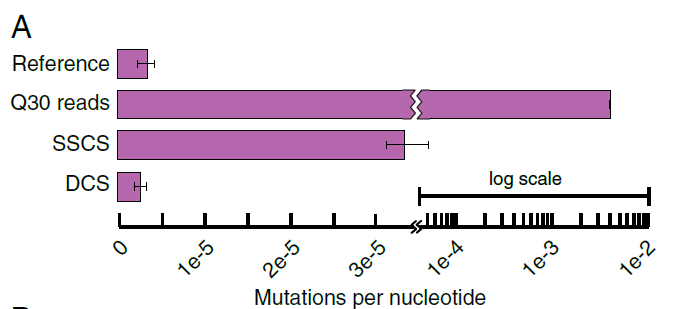


图 15、比较常规测序Q30数据、单链同源序列数据、双链同源序列数据的突变率分析平均结果。样品库的突变率为。

**七、与有关的现行法律、法规和强制性国家标准的关系**

本标准符合国家现行法律、法规、规章和强制性国家标准的要求。本标准的实施不涉及对现行标准的废止情况。

**八、标准属性的建议**

本标准为推荐性国家标准。

**九、贯彻国家标准的要求和措施建议**

标准发布实施后，建议由标准编制单位组织有关单位进行宣传贯彻。

**十、其他应予说明的事项**

无其他事项说明。