

基因测序分析方法、装置、存储介质和计算机设备

优先权信息

本申请请求 2021 年 06 月 23 日向中国国家知识产权局提交的、专利申请号为 202110698855.6
5 的专利申请的优先权和权益，并且通过参照将其全文并入此处。

技术领域

本发明涉及生物信息技术领域，尤其涉及一种基因测序分析方法、装置、存储介质和计算机设备。

10 背景技术

DNA 测序技术自发明以来就一直在推动分子生物学发展方面起着至关重要的作用。从早期 Frederick
Sanger 的手工测序，以及基于 Sanger 法开发的第一代自动化测序仪，到目前的新一代测序平台，这一领
域已经发生了巨大的变化。

新一代测序平台的测序过程主要是在每一轮测序循环中，标记不同荧光基团的 4 种核苷酸以及 DNA
15 聚合酶同时加入流通池通道中，按照碱基互补配对原则进行 DNA 链的延伸；接着采集荧光图像，碱基
特异的荧光标记揭示了这一轮中新加入的核苷酸是什么，也就获得了模板中这一位置的 DNA 序列；然
后，继续进行下一轮反应。这一过程重复多次后得到多个 DNA 序列，比如进行 50 轮测序循环后，将产
生 50 个碱基的 DNA 序列，且由于该方法获得的单条序列长度很短，因此叫做短序列。

由上述内容可见，新一代测序平台需要在生化和成像运行完后，再统一进行碱基识别，然后将识别
20 结果转化成其他形式进行存储；之后将存储后的文件传输到一级存储中进行数据质检，校验样本信息和
上机信息，再使用二级存储进行数据归档；该测序阶段耗时较长，传输效率较低，导致分析过程较慢。

发明内容

本发明的目的旨在至少能解决上述的技术缺陷之一，特别是现有技术中测序阶段耗时较长，传输效
25 率较低，导致分析过程较慢的技术缺陷。

在本发明的第一方面，本发明提供了一种基因测序分析方法，所述方法包括：获取切片数据，所述
切片数据为测序平台实时传入的，对测序文库进行至少一轮测序循环后所读取的短序列集合；将所述切
片数据输入内存中；调用所述内存中预先设置的封装程序，对所述切片数据进行数据处理及检测分析，
得到对应的分析结果；其中，所述封装程序为自定义程序。

30 根据本发明的实施例，上述一种基因测序分析方法还可以包括下列附加技术特征中的至少之一：

根据本发明的实施例，所述获取切片数据的步骤，包括：确定切片大小；根据所述切片大小对测序平台当前传入的、所有未切分的、单轮测序循环后读取的短序列进行切分，并将切分后的短序列集合作为切片数据。

- 根据本发明的实施例，所述确定切片大小的步骤，包括：获取当前CPU的空闲情况；依据所述当前CPU的空闲情况调整切片大小。

根据本发明的实施例，所述封装程序包括数据质控程序、数据处理程序和检测分析程序；所述调用所述内存中预先设置的封装程序，对所述切片数据进行数据处理及检测分析，得到对应的分析结果的步骤，包括：调用所述内存中预先设置的数据质控程序，对所述切片数据进行质量控制处理，得到质控结果；调用所述内存中预先设置的数据处理程序，对所述质控结果进行数据处理，得到处理结果；调用所述内存中预先设置的检测分析程序，对所述处理结果进行检测分析，得到最终的分析结果。

根据本发明的实施例，所述调用所述内存中预先设置的数据质控程序，对所述切片数据进行质量控制处理的步骤，包括：调用所述内存中预先设置的数据质控程序，计算所述切片数据中的短序列的位置出错率、统计所述切片数据中的碱基分布情况，和/或切除所述切片数据中的低质量碱基；其中，所述低质量碱基为碱基质量低于质量阈值的碱基。

- 根据本发明的实施例，所述调用所述内存中预先设置的数据处理程序，对所述质控结果进行数据处理的步骤，包括：调用所述内存中预先设置的数据处理程序，对所述质控结果中的短序列进行比对、排序、去重、重比对中的至少一种。

根据本发明的实施例，所述调用所述内存中预先设置的数据处理程序，对所述质控结果中的短序列进行比对、排序、去重、重比对中的至少一种的步骤，包括：调用所述内存中预先设置的数据处理程序，对所述质控结果中的短序列与参考基因之间进行比对，确定所述短序列在所述参考基因中的位置；基于所述短序列在所述参考基因中的位置，确定所述短序列的位置标签；确定所述内存中各预设的线程中，当前处于空闲状态的目标线程；参考所述短序列的位置标签，调用所述目标线程对所述短序列进行排序、去重、重比对中的至少一种。

- 根据本发明的实施例，所述调用所述内存中预先设置的检测分析程序，对所述处理结果进行检测分析的步骤，包括：调用所述内存中预先设置的检测分析程序，对所述处理结果进行变异检测、解读。

在本发明的第二方面，本发明还提供了一种基因测序分析装置，包括：数据获取模块，用于获取切片数据，所述切片数据为测序平台实时传入的，对测序文库进行至少一轮测序循环后所读取的短序列集合；切片传输模块，用于将所述切片数据输入内存中；处理分析模块，用于调用所述内存中预先设置的封装程序，对所述切片数据进行数据处理及检测分析，得到对应的分析结果；其中，所述封装程序为自定义程序。

在本发明的第三方面，本发明还提供了一种存储介质，所述存储介质中存储有计算机可读指令，所述计算机可读指令被一个或多个处理器执行时，使得一个或多个处理器执行如上述实施例中任一项所述基因测序分析方法的步骤。

在本发明的第四方面，本发明还提供了一种计算机设备，所述计算机设备中存储有计算机可读指令，
5 所述计算机可读指令被一个或多个处理器执行时，使得一个或多个处理器执行如上述实施例中任一项所述基因测序分析方法的步骤。

从以上技术方案可以看出，本发明实施例具有以下优点：

本发明提供的基因测序分析方法、装置、存储介质和计算机设备，首先将测序平台实时传入的短序列进行分切片处理后得到对应的切片处理，然后将切片数据输入内存中，以便调用内存中预先加载的封装程序对切片数据进行检测分析，得到对应的分析结果；该过程无需等待测序平台完全测序完毕后将整个测序结果传输至对应平台进行处理分析，而是实时获取测序平台测序后的切片数据，对切片数据进行流式测序及分析，因而本申请能够从整体上加快测序与分析的进程，且分析时的数据为切片数据，该切片数据相对于整个测序结果而言，传输速度更快，耗时更短。
10

另外，本申请的分析过程由内存中预先设置的封装程序进行，无需将测序数据传输至其他平台进行处理，有效降低了I/O拥堵阻塞和内存消耗，进一步提高机器运行效率的同时，也延长了机器的使用寿命。
15

附图说明

为了更清楚地说明本发明实施例或现有技术中的技术方案，下面将对实施例或现有技术描述中需要使用的附图作简单地介绍，显而易见地，下面描述中的附图仅仅是本发明的一些实施例，对于本领域普通技术人员来讲，在不付出创造性劳动性的前提下，还可以根据这些附图获得其它的附图。

20 图1为本发明实施例提供的一种基因测序分析方法的流程示意图；

图2为本发明实施例提供的流式分析的过程示意图；

图3为本发明实施例提供的基因测序分析过程示意图；

图4为本发明实施例提供的基因测序分析过程中的测序过程和时间的关系示意图；

图5为本发明实施例提供的一种基因测序分析装置的结构示意图；

25 图6为本发明实施例提供的一种计算机设备的内部结构示意图。

具体实施方式

下面将结合本发明实施例中的附图，对本发明实施例中的技术方案进行清楚、完整地描述，显然，所描述的实施例仅仅是本发明一部分实施例，而不是全部的实施例。基于本发明中的实施例，本领域普通技术人员在没有做出创造性劳动前提下所获得的所有其他实施例，都属于本发明保护的范围。

测序，简单来说就是将DNA化学信号转变为计算机可处理的数字信号。从早期Frederick Sanger的手工测序，以及基于Sanger法开发的第一代自动化测序仪，到目前的新一代测序平台，这一领域已经发生了巨大的变化，且测序技术的每一次变革和突破，都对基因组学研究，疾病医疗研究，药物研发，育种等领域产生巨大的推动作用。

5 目前，新一代测序平台的测序过程主要是在每一轮测序循环中，标记不同荧光基团的4种核苷酸以及DNA聚合酶同时加入流通池通道中，按照碱基互补配对原则进行DNA链的延伸；接着采集荧光图像，碱基特异的荧光标记揭示了这一轮中新加入的核苷酸是什么，也就获得了模板中这一位置的DNA序列；然后，继续进行下一轮反应。这一过程重复多次后得到多个DNA序列，比如进行50轮测序循环后，将产生50个碱基的DNA序列，且由于该方法获得的单条序列长度很短，因此叫做短序列。

10 由上述内容可见，新一代测序平台需要在生化和成像运行完后，再统一进行碱基识别，然后将识别结果转化成其他形式进行存储；之后将存储后的文件传输到一级存储中进行数据质控，校验样本信息和上机信息，再使用二级存储进行数据归档；该测序阶段耗时较长，传输效率较低，导致分析过程较慢。

因此，本发明的目的在于解决现有技术中测序阶段耗时较长，传输效率较低，导致分析过程较慢的技术问题，并具体提出以下技术方案：

15 示意性地，如图1所示，图1为本发明实施例提供的一种基因测序分析方法的流程示意图；本发明提供了一种基因测序分析方法，具体包括如下步骤：

S110：获取切片数据。

本步骤中，在进行生信分析之前，需要获取测序平台传入的切片数据，以便根据该切片数据进行相应的生信分析。

20 可以理解的是，本申请考虑到后续平台的分析效率以及内存的占用情况等，需要对测序平台传入的实时数据进行分切片处理，且切片大小可随服务器的空闲程度自适应调整。因此，本申请获取的切片数据为测序平台实时传入的、对测序文库进行至少一轮测序循环后所读取的短序列集合。

需要说明的是，这里的切片大小可随服务器的空闲程度进行自适应调整的情况主要包括：若服务器不空闲时，可将切片大小调整至能够切分一个测序循环后读取的短序列；若服务器较为空闲，则可以将切片大小调整至能够切分多个测序循环后读取的短序列。另外，首次进行切分时，可等待多轮测序循环，25 如等待10轮测序循环后切分为一个切片。

举例来说，本申请可以使用二代测序平台来获取切片数据，二代测序（NGS）和Sanger测序相同，都是在每一个测序周期中，利用计算机检测DNA聚合酶催化荧光标记的dNTP结合到DNA模板时产生的荧光信号。但与Sanger单位时间检测单片段不同的是，NGS能同时检测成千上万的孔道的信号，因此大30 大大提高了效率。

可以理解的是，这里的二代测序平台包括但不限于Illumina (Solexa) sequencing、Roche 454 sequencing、Ion torrent: Proton / PGM sequencing、SOLiD sequencing。

下面以二代测序平台为例，进一步说明如何获取切片数据。示意性地，本申请中的二代测序平台在对测序文库进行测序前，需要构建测序文库，该测序文库指的是DNA片段在两头接上特定的DNA接头序列后组成的DNA混合物。例如，使用超声波等将基因组DNA打断，然后把打断DNA片段的两头用酶补平，并用连接酶把接头序列连接上去，连接好的一段DNA混合物，就称为一个“文库”(library)。

而使用二代测序平台对测序文库进行测序前，还可以对测序文库进行PCR扩增，PCR扩增的作用是为了将待测序溶液中微弱的DNA片段复制几倍甚至几十倍，以便增大它们在待测序溶液中分布的密度，使得取样时能够被获取到。

正式测序时，可以加入中性溶液，并在中性溶液中加入测序dNTP引物。由于dNTP的3'端被叠氮基团堵住，因此，一轮测序循环只能延长1个碱基，一个测序循环结束之后，加入特定化学试剂将叠氮基团和荧光标记基团切掉，使3'端的羟基暴露出来，就可以把碱基序列读取出来。

本申请中，将测序过程中每个测序循环的生化和成像部分都实时传入内存进行碱基识别(Basecall)，得到对应的短序列，并在满足一个切片大小时输出切片数据存入内存中。

举例来说，原有的二代测序平台进行测序后输出为fastq文件，fastq文件是一种包含质量值的序列文件，其中的q为quality，一般用来存储原始测序数据，扩展名一般为fastq或者fq。下面是fastq文件常见的序列格式：

```
@DJB775P1:248:D0MDGACXX:7:1202:12362:49613
TGCTTACTCTGCGTTGATACCACTGCTTAGATCGGAAGAGCACACGTCTGAA
+
JJJJJJJJJJHIIHHGHFFFFFCEEEEEEDBD?DDDDDDDBDDABDDCA
@DJB775P1:248:D0MDGACXX:7:1202:12782:49716
CTCTGCGTTGATACCACTGCTTACTCTGCGTTGATACCACTGCTTAGATCGG
+
IIIIIIIIIIHHHHHHHHFFFFFEECCCCBCECCCCCCCCCCCCCCCC
```

由上述序列格式可知，fastq文件中每四行成为一个独立的单元，我们称之为read，即短序列。第一行以‘@’开头，DJB775P1是这条短序列的名字，第一行的字符串是根据测序时的状态信息转换过来的，中间不会有空格，它是每一条短序列的唯一标识符；第二行是测序read的序列，由A，C，G，T和N这五种字母构成，这也是我们真正关心的DNA序列，N代表的是测序时那些无法被识别出来的碱基；第三行以‘+’开头，在旧版的FASTQ文件中会直接重复第一行的信息，但现在一般什么也不加（节省存储空间）；

第四行是测序read的质量值，这个和第二行的碱基信息一样重要，它描述的是每个测序碱基的可靠程度，用ASCII码表示。

而本申请为了提高分析效率，将测序和分析过程分切片处理，采用流式结构依次处理，实现边测序，边编码，边传输，边分析。首先，将测序过程中的每个测序循环的生化 and 成像部分获得的结果传入内存

5 进行basecall（碱基识别）过程的图像识别处理，从而得到多个短序列，然后根据服务器的空闲程度确定切片大小，即依据服务器空闲程度对上述第二行中的碱基进行切分，服务器若空闲，则切分多个碱基作为一个切片数据，若不空闲，则切分单个碱基作为一个切片数据，切分后得到的切片数据保存至内存中。

接着，在读取内存中的切片数据进行分析时，可先去掉接头（位于第一个切片初始，一般是6-8个碱基，接头数据可用于区分不同的样本，方便之后输出），然后将剩余切片数据比对到人类基因组上，每次新的切片数据的每个碱基都使read定位到人类基因组的范围不断缩小，比对位置更准确。

10

另外，本申请在进行测序与分析时，所选用的平台可以是将测序与分析进行合成的机器，也可以是将现有测序仪和分析仪通过网络连接起来，实现数据传输和运行分析，在此不做限制。

S120：将切片数据输入内存中。

本步骤中，通过步骤S110获取切片数据后，可输入至合成的机器中的内存，也可以输入至与测序仪

15 连接的分析仪的内存中，以便通过内存中的封装程序来对切片数据进行数据处理及检测分析。

可以理解的是，现有的测序仪对测序文库进行测序后，得到的测序结果一般输出为fastq文件，并存储到硬盘里；做比对分析时，从硬盘中读取fastq文件，再比对到人类基因组（reference.fa）上；而本申请获取到的切片数据，则不需要输出到硬盘中，而是直接放到内存里，即可进行比对处理，该方式减少了写入硬盘和从硬盘中读取的过程，降低了I/O消耗。

20 另外，需要说明的是，这里的内存可以是DDR内存。由于基因数据较大，服务器内存有限，常扩展为DDR内存，DDR内存存在时钟信号上升沿与下降沿各传输一次数据，这使得DDR内存的数据传输速度为传统SDRAM的两倍。并且，由于仅多采用了下降缘信号，因此并不会造成能耗增加。至于定址与控制信号则与传统SDRAM相同，仅在时钟上升缘传输。

S130：调用内存中的封装程序对切片数据进行数据处理及检测分析，得到分析结果。

25 本步骤中，在对切片数据进行处理时，需要调用内存中预先设置的封装程序，通过封装程序来对切片数据进行数据处理及检测分析，并得到对应的分析结果。

需要说明的是，这里的封装程序指的是将多个程序封装为一个程序的软件，且该封装程序为自定义程序，可以根据用户需求封装多个不同的程序为一个程序，来对切片数据进行数据处理和检测分析。

并且，这里的封装程序是在已有的国际公认的程序基础上做了优化，增加了内存衔接，增加了多线程

30 程，还增加了对切片数据的处理功能，当需要使用该封装程序进行数据处理时，可触发放置在磁盘中的

对应的封装程序，并将该封装程序加载到内存中成为任务进程，从而有效降低I/O消耗，减少读写时间，提高分析效率。

举例来说，在对切片数据进行检测分析前，需要对其进行数据处理，该数据处理过程可通过一个程序或多个程序来实现，如数据质控程序、数据处理程序等，然后再连接一个检测分析程序来对数据处理

5 后的结果进行检测分析，使得最终的分析结果能够更加准确。

示意性地，如图2所示，图2为本发明实施例提供的流式分析的过程示意图；图2中，测序仪可使用生物和成像系统对测序文库进行生化成像，并将生化成像后的图像以切片传输的方式传输至中控服务器中，通过中控服务器传输至分析仪器中进行数据处理和检测分析。其中数据质控过程为通过数据质控服务器中的服务器内存获取生产信息系统中的样本和上机信息，以及中控服务器发送的切片数据，然后对该切片数据

10 片数据进行数据质控后进行数据归档并交付，最后将分析结果进行存储，从而实现流式分析过程。

由图2可见，本申请使用的流式分析过程，通过对测序平台传入的短序列集合进行切片，并将一部分分析时间隐藏在原有的测序时间内，且数据处理过程与检测分析过程通过内存衔接，能够有效降低I/O消耗，减少读写时间的同时，提高分析效率。

上述实施例提供的基因测序分析方法，首先将测序平台实时传入的短序列进行分切片处理后得到对应的切片处理，然后将切片数据输入内存中，以便调用内存中预先加载的封装程序对切片数据进行检测分析，得到对应的分析结果；该过程无需等待测序平台完全测序完毕后将整个测序结果传输至对应平台进行处理分析，而是实时获取测序平台测序后的切片数据，对切片数据进行流式测序及分析，因而本申请能够从整体上加快测序与分析的进程，且分析时的数据为切片数据，该切片数据相对于整个测序结果而言，传输速度更快，耗时更短。

15

20 另外，本申请的分析过程由内存中预先设置的封装程序进行，无需将测序数据传输至其他平台进行处理，有效降低了I/O拥堵阻塞和内存消耗，进一步提高机器运行效率的同时，也延长了机器的使用寿命。

下面将通过多个实施例来对本申请的基因测序分析方法进行进一步说明。如下所述，下述实施例中，主要是对如何获取切片数据的过程进行展开描述。

25 在一个实施例中，步骤S110中获取切片数据的步骤，可以包括：

S111：确定切片大小。

S112：根据所述切片大小对测序平台当前传入的、所有未切分的、单轮测序循环后读取的短序列进行切分，并将切分后的短序列集合作为切片数据。

本实施例中，将测序过程中每个测序循环的生化成像部分都实时传入内存进行碱基识别，得到对应的短序列后，需要先确定切片大小，并将满足一个切片大小的短序列结合输出为切片数据，将切片数据存入内存中。

5 举例来说，每 $m+n$ 个cycle（测序循环）为一个切片数据，其中 m 是第一个切片， $m \geq 10$ ， n 指代之后的所有切片， $n \geq 1$ 且 $n \leq 20$ ， $m+n$ 可以根据分析速度自适应调节切片大小。

示意性地，如图3、图4所示，图3为本发明实施例提供的基因测序分析过程示意图，图4为本发明实施例提供的基因测序分析过程中的测序过程和时间的关系示意图；由图3和图4可知，本申请通过实时获取测序平台测序后的切片数据，并对切片数据进行流式测序及分析，从而能够在整体上加快测序与分析的进程，之后完成变异检测或比对到RNA病原库等，再进行后续的注释，结束分析流程，完成交付。

10

上述实施例对如何获取切片数据的过程进行展开描述，下面将对上述实施例中确定切片大小的步骤进行进一步说明。

在一个实施例中，步骤S111中确定切片大小的步骤，可以包括：

S1111：获取当前CPU的空闲情况。

15 S1112：依据所述当前CPU的空闲情况调整切片大小。

本实施例中，在获取切片数据进行检测分析时，该切片数据为测序平台实时传入的、对测序文库中的基因片段进行至少一轮测序循环后所读取的短序列集合。

其中，切片大小可随分析速度进行自适应调整，而分析速度则可以通过获取当前CPU的空闲情况进行分析后得到，当前CPU的空闲情况则与当前内存的占用情况相关。如当前内存占用较多，则表示不空闲，此时内存的分析速度较慢，可将切片大小调整至能够切分多个测序循环后读取的短序列集合；若当前内存占用较少，则表示空闲，此时内存的分析速度较快，则可以将切片大小调整至能够切分一个测序循环后读取的短序列集合。

25 举例来说，每 $m+n$ 个cycle（测序循环）为一个切片数据， m 之后的切片是自动控制的，即从第二个切片开始到最后一个切片结束，切片 n 的大小是根据机器的空闲情况决定的，若较为空闲，则每个cycle传递一次，此时 $n=1$ ；若较忙，则等 $n=20$ 的时候再传输。

可以理解的是，根据本申请中的比对原理，切片数据被输送到到比对过程中与人类基因组进行比对，并确定切片数据的比对的范围，若切片数据越大，则比对范围缩小的越精确。

上述实施例中对确定切片大小的步骤进行进一步说明，下面将通过实施例对封装程序以及如何调用封装程序进行数据处理及检测分析的过程进行展开描述。

30

在一个实施例中，所述封装程序可以包括数据质控程序、数据处理程序和检测分析程序。

步骤S130中调用所述内存中预先设置的封装程序，对所述切片数据进行数据处理及检测分析，得到对应的分析结果的步骤，可以包括：

5 S131：调用所述内存中预先设置的数据质控程序，对所述切片数据进行质量控制处理，得到质控结果。

S132：调用所述内存中预先设置的数据处理程序，对所述质控结果进行数据处理，得到处理结果。

S133：调用所述内存中预先设置的检测分析程序，对所述处理结果进行检测分析，得到最终的分析结果。

10 本实施例中，在对切片数据进行处理时，可以调用内存中预先设置的封装程序，通过封装程序来对切片数据进行数据处理及检测分析，并得到对应的分析结果。

而内存中预先设置的封装程序可以是自定义程序，该自定义程序可由多个程序封装为一个程序。举例来说，该封装程序可以包括数据质控程序、数据处理程序和检测分析程序。

其中，数据质控程序主要用于对切片数据进行质量控制处理，并得到质控结果；数据处理程序主要用于对质控结果进行数据处理，并得到处理结果；而检测分析程序则是用于对处理结果进行检测分析，
15 从而得到最终的分析结果。

上述实施例中对封装程序以及如何调用封装程序进行数据处理及检测分析的过程进行展开描述，下面将对如何调用数据质控程序进行质量控制处理过程进行详细说明。

20 在一个实施例中，步骤S131中调用所述内存中预先设置的数据质控程序，对所述切片数据进行质量控制处理的步骤，可以包括：

调用所述内存中预先设置的数据质控程序，计算所述切片数据中的短序列的位置出错率、统计所述切片数据中的碱基分布情况，和/或切除所述切片数据中的低质量碱基；其中，所述低质量碱基为碱基质量低于质量阈值的碱基。

25 本实施例中，在对切片数据进行质量控制处理时，可调用封装程序中的数据质控程序来进行质量控制处理。

具体地，数据质控程序主要是对切片数据进行质量控制处理，该质量控制处理过程可以包括计算切片数据中的短序列的位置出错率、统计切片数据中的碱基分布情况、切除切片数据中的低质量碱基，以及其他适用于短序列的质量控制处理过程，具体选择哪一种或多种质量控制处理过程，且质量控制处理过程使用何种工具，可依据实际情况进行设置。

例如，当需要切除切片数据中的低质量碱基时，可以通过滑动一定长度的窗口，并计算窗口内的碱基平均质量，如果过低，就直接往后全部切除。

上述实施例中对如何调用数据质控程序进行质量控制处理过程进行详细说明，下面将通过实施例来

5 说明如何调用数据处理程序进行数据处理过程。

在一个实施例中，步骤S132中调用所述内存中预先设置的数据处理程序，对所述质控结果进行数据处理的步骤，可以包括：

调用所述内存中预先设置的数据处理程序，对所述质控结果中的短序列进行比对、排序、去重、重比对的至少一种。

10 本实施例中，在对切片数据进行质量控制处理后，可调用内存中预先设置的数据处理程序，来对质量控制处理后的质控结果中的短序列进行比对、排序、去重和/或重比对。

具体地，虽然切片数据均来自于有序的基因组，但经过DNA建库和测序之后，切片数据中不同短序列之间的前后顺序关系已经全部丢失，因此，切片数据中相邻的两条短序列之间没有任何位置关系，他们都是随机来自于原本基因组中的某个位置的短序列。

15 因此，本申请在对切片数据进行质量控制处理后，可调用内存中预先设置的数据处理程序，将质控结果对应的切片数据中的短序列一个个地与该物种的参考基因进行比对，找到每一条短序列在参考基因上的位置，然后按顺序排列好，这个过程就称为切片数据的比对过程。

需要说明的是，这里的参考基因指的是该物种的基因组序列，即为已经组装成的完整基因组序列，常将该完整基因组序列作为该物种的标准参照物。

20 完成短序列比对后，即可进行排序流程，该排序流程的作用主要是将之前比对后的短序列按照位置先后顺序从小到大进行排序，以使多个短序列之间形成连续的位置关系。

而在排序流程后，还可以对切片数据中的短序列进行去重操作，该去重操作主要是去除切片数据中重复的短序列，这样方便后续检测分析。

25 最后，短序列的重比对过程，主要是将上述对比过程中发现的潜在序列插入或序列删除的区域进行重新校正，以使后续检测分析的结果更加准确。

可以理解的是，上述比对、排序、去重、重比对的数据处理过程，可依据实际情况有选择地进行设置，且前后顺序也可以人为设置，在此不做限制。

30 例如，正常的分析流程是比对、排序、去重、重比对、变异检测、注释；比对时的输入文件是fastq文件，输出文件是bam文件，之后一直到重比对输出都是bam文件，而变异检测的输出是vcf文件，注释时根据变异检测的结果注释的；而使用本申请的基因测序分析方法时，若使用二代测序平台进行测序，

所有的二代fastq的下机数据分析流程都要做比对，之后有的做后续流程，有的不做；如果不从下机数据开始，也可以输入任意阶段的bam文件做后续分析处理。

下面将对本申请的比对过程进行进一步说明。举例来说，本申请在对切片数据进行比对时，可以将切片数据中的短序列的碱基与全基因组进行比对，找到每一条短序列在参考基因上的位置，待全部切片数据都定位到全基因组中的特定位置后，再整合匹配到的比对信息，进行相邻比对位置的筛选扩展，且允许空位存在（即没比对上的位置），以此获得更长的比对到全基因组的序列链。

进一步地，在上述比对过程中，可以使用综合全局和局部比对情况的动态规划算法，选取最优的比对结果信息，然后根据接头信息，将不同样本比对结果分别输出到内存。

例如，常用的BWA-MEM算法主要采用seed-and-extend策略。在seed阶段，BWA取read的碱基片段在reference上进行精确匹配，并选择满足一定匹配次数和长度要求的read片段作为seed，这个阶段算法的核心是基于FM-index的精确匹配；在extend阶段，BWA利用Smith-Waterman算法将seed在read和reference上向两边延伸比对（容忍gap），进而找到整个read在reference上符合条件的全局匹配。

上述实施例主要说明如何调用数据处理程序进行数据处理过程，下面将对数据处理过程的具体实施方式进一步说明。

在一个实施例中，所述调用所述内存中预先设置的数据处理程序，对所述质控结果中的短序列进行比对、排序、去重、重比对中的至少一种的步骤，可以包括：

A11：调用所述内存中预先设置的数据处理程序，对所述质控结果中的短序列与参考基因之间进行比对，确定所述短序列在所述参考基因中的位置。

A12：基于所述短序列在所述参考基因中的位置，确定所述短序列的位置标签。

A13：确定所述内存中各预设的线程中，当前处于空闲状态的目标线程。

A14：参考所述短序列的位置标签，调用所述目标线程对所述短序列进行排序、去重、重比对中的至少一种。

本实施例中，在对质控结果中的短序列进行数据处理时，可将比对后的短序列增加位置标签，然后再确定当前内存中各个预设的线程中，是否有处于空闲状态的线程，如果有的话，则将该线程作为目标线程，接着参考短序列的位置标签，调用该目标线程对短序列进行排序、去重和/或重比对，使得排序和去重更为高效。

举例来说，通过调用数据处理程序对质控结果中的短序列与参考基因进行比对，并基于该短序列在参考基因中的位置确定与短序列对应的位置标签后，可调用目标线程对该短序列进行排序、去重等操作。

比如在排序过程中，使用多线程进行处理时，为了提高多线程的处理效率，可参考该短序列的位置标签进行排序，以使得多线程排序后得到的排序结果准确无误，同时排序效率更高。

上述实施例中对数据处理过程的具体实施方式进行进一步说明，下面将对如何调用检测分析程序对

5 处理结果进行检测分析展开具体描述。

在一个实施例中，步骤S133中调用所述内存中预先设置的检测分析程序，对所述处理结果进行检测分析的步骤，可以包括：调用所述内存中预先设置的检测分析程序，对所述处理结果进行变异检测、解

本实施例中，进行基因测序分析的其中一个目标是获得样本准确的变异集合，而获得样本准确的变
10 异集合则需要通过变异检测、解读等来进行确定。目前，变异检测的内容包括有SNP、Indel、CNV、SV等。变异检测的过程可以使用应用贝叶斯推断的算法，也可以使用HaplotypeCaller算法。

例如，使用HaplotypeCaller算法进行变异检测时，会先推断群体的单倍体组合情况，计算各个组合的几率，然后根据这个信息再反推每个样本的基因型组合。

进一步地，本申请在对处理结果进行检测分析时，如进行变异检测，变异检测过程不需要等待比
15 对完成后再做，而是可以将全基因组中的不同染色体分成若干区域，当每个区域积累到一部分比对结果后，即可寻找高突变区进行检测；后续比对结果中增加不断验证现有高突变区的变异情况，包括错配/插入/缺失等；接着将高突变区的数据进行组装，获得精简的单倍型数据，再利用隐马尔可夫模型估计单倍型的最大似然估计，获区每个位点的分型结果，输出变异信息。

可以理解的是，为了提高整个生信分析的分析效率，本申请在对处理结果进行变异检测时，并不会
20 对全基因组范围内的所有位点都进行变异检测，而是选择高变区进行检测，而高变区的选择可依据每个位点出现突变的概率进行选择。例如，计算得到每个位点出现突变的概率后，可根据预定概率阈值来判断是否为高变区。

并且，将高突变区的数据进行组装时，若参考基因组对应的片段集合中存在重复，则会将短片段的
25 长度递增至没有重复或达到最大长度限制，且组装获得的边缘可以根据对比得到的短序列的数目来分配权重。

更进一步地，在获得变异检测结果后，还可以对该变异检测结果进行质控和过滤，以便进一步区分变异检测结果的好坏。

下面是对本申请实施例提供的基因测序分析装置进行描述，下文描述的基因测序分析装置与上文描述的基因测序分析方法可相互对应参照。

在一个实施例中，如图5所示，图5为本发明实施例提供的一种基因测序分析装置的结构示意图；本发明还提供了一种基因测序分析装置，包括数据获取模块210、切片传输模块220、处理分析模块230，具体包括如下：

数据获取模块210，用于获取切片数据，所述切片数据为测序平台实时传入的，对测序文库进行至少
5 一轮测序循环后所读取的短序列集合。

切片传输模块220，用于将所述切片数据输入内存中。

处理分析模块230，用于调用所述内存中预先设置的封装程序，对所述切片数据进行数据处理及检测分析，得到对应的分析结果；其中，所述封装程序为自定义程序。

上述实施例提供的基因测序分析装置，首先将测序平台实时传入的短序列进行分切片处理后得到对
10 应的切片处理，然后将切片数据输入内存中，以便调用内存中预先加载的封装程序对切片数据进行检测分析，得到对应的分析结果；该过程无需等待测序平台完全测序完毕后将整个测序结果传输至对应平台进行处理分析，而是实时获取测序平台测序后的切片数据，对切片数据进行流式测序及分析，因而本申请能够从整体上加快测序与分析的进程，且分析时的数据为切片数据，该切片数据相对于整个测序结果而言，传输速度更快，耗时更短。

15 另外，本申请的分析过程由内存中预先设置的封装程序进行，无需将测序数据传输至其他平台进行处理，有效降低了I/O拥堵阻塞和内存消耗，进一步提高机器运行效率的同时，也延长了机器的使用寿命。

在一个实施例中，数据获取模块210可以包括：

确定模块211，用于确定切片大小。

20 切分模块212，用于根据所述切片大小对测序平台当前传入的、所有未切分的、单轮测序循环后读取的短序列进行切分，并将切分后的短序列集合作为切片数据。

在一个实施例中，确定模块211可以包括：

内存情况获取模块2111，用于获取当前内存的空闲情况。

25 切片调整模块2112，用于依据所述当前内存的空闲情况调整切片大小。

在一个实施例中，处理分析模块230中的封装程序可以包括数据质控程序、数据处理程序和检测分析程序。

处理分析模块230，用于调用所述内存中预先设置的封装程序，对所述切片数据进行数据处理及检测
30 分析，得到对应的分析结果的步骤，可以包括：

数据质控模块231，用于调用所述内存中预先设置的数据质控程序，对所述切片数据进行质量控制处理，得到质控结果。

数据处理模块232，用于调用所述内存中预先设置的数据处理程序，对所述质控结果进行数据处理，得到处理结果。

- 5 检测分析模块233，用于调用所述内存中预先设置的检测分析程序，对所述处理结果进行检测分析，得到最终的分析结果。

在一个实施例中，数据质控模块231可以包括：

- 10 第一处理模块2311，用于调用所述内存中预先设置的数据质控程序，计算所述切片数据中的短序列的位置出错率、统计所述切片数据中的碱基分布情况，和/或切除所述切片数据中的低质量碱基；其中，所述低质量碱基为碱基质量低于质量阈值的碱基。

在一个实施例中，数据处理模块232可以包括：

- 15 第二处理模块2321，用于调用所述内存中预先设置的数据处理程序，对所述质控结果中的短序列进行比对、排序、去重、重比对中的至少一种。

在一个实施例中，第二处理模块2321可以包括：

比对模块310，用于调用所述内存中预先设置的数据处理程序，对所述质控结果中的短序列与参考基因之间进行比对，确定所述短序列在所述参考基因中的位置。

- 20 确定标签模块311，用于基于所述短序列在所述参考基因中的位置，确定所述短序列的位置标签。

确定线程模块312，用于确定所述内存中各预设的线程中，当前处于空闲状态的目标线程。

多线程操作模块313，用于参考所述短序列的位置标签，调用所述目标线程对所述短序列进行排序、去重、重比对中的至少一种。

- 25 在一个实施例中，检测分析模块233可以包括：

第三处理模块2331，用于调用所述内存中预先设置的检测分析程序，对所述处理结果进行变异检测、解读。

在一个实施例中，本发明还提供了一种存储介质，所述存储介质中存储有计算机可读指令，所述计算机可读指令被一个或多个处理器执行时，使得一个或多个处理器执行如上述实施例中任一项所述基因测序分析方法的步骤。

- 5 在一个实施例中，本发明还提供了一种计算机设备，所述计算机设备中存储有计算机可读指令，所述计算机可读指令被一个或多个处理器执行时，使得一个或多个处理器执行如上述实施例中任一项所述基因测序分析方法的步骤。

示意性地，如图6所示，图6为本发明实施例提供的一种计算机设备的内部结构示意图，该计算机设备300可以被提供为一服务器。参照图6，计算机设备300包括处理组件302，其进一步包括一个或多个处
10 理器，以及由存储器301所代表的存储器资源，用于存储可由处理组件302的执行的指令，例如应用程序。存储器301中存储的应用程序可以包括一个或一个以上的每一个对应于一组指令的模块。此外，处理组件302被配置为执行指令，以执行上述任意实施例的基因测序分析方法。

计算机设备300还可以包括一个电源组件303被配置为执行计算机设备300的电源管理，一个有线或无线网络接口304被配置为将计算机设备300连接到网络，和一个输入输出（I/O）接口305。计算机设备300
15 可以操作基于存储在存储器301的操作系统，例如Windows Server™、Mac OS X™、Unix™、Linux™、Free BSD™或类似。

本领域技术人员可以理解，图6中示出的结构，仅仅是与本申请方案相关的部分结构的框图，并不构成对本申请方案所应用于其上的计算机设备的限定，具体的计算机设备可以包括比图中所示更多或更少的部件，或者组合某些部件，或者具有不同的部件布置。

- 20 最后，还需要说明的是，在本文中，诸如第一和第二等之类的关系术语仅仅用来将一个实体或者操作与另一个实体或操作区分开来，而不一定要求或者暗示这些实体或操作之间存在任何这种实际的关系或者顺序。而且，术语“包括”、“包含”或者其任何其他变体意在涵盖非排他性的包含，从而使得包括一系列要素的过程、方法、物品或者设备不仅包括那些要素，而且还包括没有明确列出的其他要素，或者是还包括为这种过程、方法、物品或者设备所固有的要素。在没有更多限制的情况下，由语句“包括一个……”限定的要素，并不排除在包括所述要素的过程、方法、物品或者设备中还存在另外的相同要素。
25 个……”限定的要素，并不排除在包括所述要素的过程、方法、物品或者设备中还存在另外的相同要素。

本说明书中各个实施例采用递进的方式描述，每个实施例重点说明的都是与其他实施例的不同之处，各个实施例之间可以根据需要进行组合，且相同相似部分互相参见即可。

- 对所公开的实施例的上述说明，使本领域专业技术人员能够实现或使用本申请。对这些实施例的多种修改对本领域的专业技术人员来说将是显而易见的，本文中所定义的一般原理可以在不脱离本申请的精神或范围的情况下，在其它实施例中实现。因此，本申请将不会被限制于本文所示的这些实施例，而
30 精神或范围的情况下，在其它实施例中实现。因此，本申请将不会被限制于本文所示的这些实施例，而

是要符合与本文所公开的原理和新颖特点相一致的最宽的范围。

权利要求书

- 1、一种基因测序分析方法，其特征在于，所述方法包括：
- 获取切片数据，所述切片数据为测序平台实时传入的，对测序文库进行至少一轮测序循环后所读取
- 5 的短序列集合；
- 将所述切片数据输入内存中；
- 调用所述内存中预先设置的封装程序，对所述切片数据进行数据处理及检测分析，得到对应的分析结果；其中，所述封装程序为自定义程序。
- 2、根据权利要求1所述的基因测序分析方法，其特征在于，所述获取切片数据的步骤，包括：
- 10 确定切片大小；
- 根据所述切片大小对测序平台当前传入的、所有未切分的、单轮测序循环后读取的短序列进行切分，并将切分后的短序列集合作为切片数据。
- 3、根据权利要求2所述的基因测序分析方法，其特征在于，所述确定切片大小的步骤，包括：
- 获取当前 CPU 的空闲情况；
- 15 依据所述当前 CPU 的空闲情况调整切片大小。
- 4、根据权利要求1所述的基因测序分析方法，其特征在于，所述封装程序包括数据质控程序、数据处理程序和检测分析程序；
- 所述调用所述内存中预先设置的封装程序，对所述切片数据进行数据处理及检测分析，得到对应的分析结果的步骤，包括：
- 20 调用所述内存中预先设置的数据质控程序，对所述切片数据进行质量控制处理，得到质控结果；
- 调用所述内存中预先设置的数据处理程序，对所述质控结果进行数据处理，得到处理结果；
- 调用所述内存中预先设置的检测分析程序，对所述处理结果进行检测分析，得到最终的分析结果。
- 5、根据权利要求4所述的基因测序分析方法，其特征在于，所述调用所述内存中预先设置的数据质控程序，对所述切片数据进行质量控制处理的步骤，包括：
- 25 调用所述内存中预先设置的数据质控程序，计算所述切片数据中的短序列的位置出错率、统计所述切片数据中的碱基分布情况，和/或切除所述切片数据中的低质量碱基；
- 其中，所述低质量碱基为碱基质量低于质量阈值的碱基。
- 6、根据权利要求4所述的基因测序分析方法，其特征在于，所述调用所述内存中预先设置的数据处理程序，对所述质控结果进行数据处理的步骤，包括：
- 30 调用所述内存中预先设置的数据处理程序，对所述质控结果中的短序列进行比对、排序、去重、重

比对、碱基质量值校正中的至少一种。

7、根据权利要求6所述的基因测序分析方法，其特征在于，所述调用所述内存中预先设置的数据处理程序，对所述质控结果中的短序列进行比对、排序、去重、重比对中的至少一种的步骤，包括：

调用所述内存中预先设置的数据处理程序，对所述质控结果中的短序列与参考基因之间进行比对，

5 确定所述短序列在所述参考基因中的位置；

基于所述短序列在所述参考基因中的位置，确定所述短序列的位置标签；

确定所述内存中各预设的线程中，当前处于空闲状态的目标线程；

参考所述短序列的位置标签，调用所述目标线程对所述短序列进行排序、去重、重比对中的至少一种。

10 8、根据权利要求4所述的基因测序分析方法，其特征在于，所述调用所述内存中预先设置的检测分析程序，对所述处理结果进行检测分析的步骤，包括：

调用所述内存中预先设置的检测分析程序，对所述处理结果进行变异检测、解读。

9、一种基因测序分析装置，其特征在于，包括：

数据获取模块，用于获取切片数据，所述切片数据为测序平台实时传入的，对测序文库进行至少一

15 轮测序循环后所读取的短序列集合；

切片传输模块，用于将所述切片数据输入内存中；

处理分析模块，用于调用所述内存中预先设置的封装程序，对所述切片数据进行数据处理及检测分析，得到对应的分析结果；其中，所述封装程序为自定义程序。

10、一种存储介质，其特征在于：所述存储介质中存储有计算机可读指令，所述计算机可读指令被一个或多个处理器执行时，使得一个或多个处理器执行如权利要求1至8中任一项所述基因测序分析方法的步骤。

11、一种计算机设备，其特征在于：所述计算机设备中存储有计算机可读指令，所述计算机可读指令被一个或多个处理器执行时，使得一个或多个处理器执行如权利要求1至8中任一项所述基因测序分析方法的步骤。

25