

基于 FPGA 的重测序分析方法和装置

5 技术领域

本公开涉及生物信息分析技术领域，尤其涉及一种基于 FPGA 的重测序分析方法和装置。

背景技术

10 随着第二代基因测序技术的出现，基因测序的成本迅速下降，以及提升了测序数据的可用性。对基因组中的变异进行检测和关联性研究，能够促进医疗领域的发展，但是，人的基因组由三十多亿碱基对构成，数量较为庞大，因此，在对人的基因组进行测序分析时，需要强大的计算能力、存储容量和网络带宽。

相关技术中，重测序分析包括测序序列对比以及变异检测，其中，测序序列对比采用  
15 动态种子延伸的策略进行对比，当初始种子长度获取的比对位置数量超过固定值时，增加种子长度去比对，以获得更少的比对位置，而后采用史密斯-沃特曼算法进行比对，并使用多线程进行变异检测。

然而，上述方式中，由于步骤之间为串行执行，因此存在执行效率较低，分析速率较慢的问题。

20

发明内容

本公开提出一种基于 FPGA 的重测序分析方法和装置，以实现节省程序运行时间，节省计算成本，提升重测序效率，用于解决现有技术中执行效率较低，分析速率较慢的技术问题。

25 本公开实施例一方面实施例提出了一种基于 FPGA 的重测序分析方法，包括：  
接收基因组重测序数据；

将所述重测序数据作为 FPGA 的输入，根据所述 FPGA 的输出确定重测序过程中的比  
对结果，同时，对所述比对结果进行排序去重处理；

对排序去重处理后的比对结果进行碱基质量值的校正；以及

30 根据校正后的比对结果进行变异结果的检测。

本公开实施例的基于 FPGA 的重测序分析方法，通过接收基因组重测序数据，而后将重测序数据作为 FPGA 的输入，根据 FPGA 的输出确定重测序过程中的比对结果，同时，

对比对结果进行排序去重处理，接着，对排序去重处理后的比对结果进行碱基质量值的校正，最后，根据校正后的比对结果进行变异结果的检测。本公开中，对比过程、排序过程和去重处理过程可以同步执行，由此，可以节省程序运行时间，节省计算成本，提升重测序效率。

- 5       本公开实施例又一方面实施例提出了一种基于 FPGA 的重测序分析装置，包括：  
接收模块，用于接收基因组重测序数据；  
处理模块，用于将所述重测序数据作为 FPGA 的输入，根据所述 FPGA 的输出确定重测序过程中的比对结果，同时，对所述比对结果进行排序去重处理；  
校正模块，用于对排序去重处理后的比对结果进行碱基质量值的校正；以及  
10       检测模块，用于根据校正后的比对结果进行变异结果的检测。

本公开实施例的基于 FPGA 的重测序分析装置，通过接收基因组重测序数据，而后将重测序数据作为 FPGA 的输入，根据 FPGA 的输出确定重测序过程中的比对结果，同时，对比对结果进行排序去重处理，接着，对排序去重处理后的比对结果进行碱基质量值的校正，最后，根据校正后的比对结果进行变异结果的检测。本公开中，对比过程、排序过程  
15       和去重处理过程可以同步执行，由此，可以节省程序运行时间，节省计算成本，提升重测序效率。

本公开实施例又一方面实施例提出了一种计算机设备，包括：存储器、处理器及存储在存储器上并可在处理器上运行的计算机程序，所述处理器执行所述程序时，实现如本公开前述实施例所述的基于 FPGA 的重测序分析方法。

- 20       本公开实施例又一方面实施例提出了一种计算机可读存储介质，所述计算机可读存储介质存储有计算机可读指令，所述计算机可读指令用于使计算机执行如本公开前述实施例所述的基于 FPGA 的重测序分析方法。

本公开附加的方面和优点将在下面的描述中部分给出，部分将从下面的描述中变得明显，或通过本公开的实践了解到。

25

## 附图说明

为了更清楚地说明本公开实施例中的技术方案，下面将对实施例中所需要使用的附图作简单地介绍，显而易见地，下面描述中的附图是本公开的一些实施例，对于本领域普通技术人员来讲，在不付出创造性劳动的前提下，还可以根据这些附图获得其他的附图。

- 30       图 1 为本公开实施例一所提供的基于 FPGA 的重测序分析方法的流程示意图；  
图 2 为本公开实施例二所提供的基于 FPGA 的重测序分析方法的流程示意图；  
图 3 为本公开实施例的基于 FPGA 的重测序分析系统的结构示意图一；

- 图 4 为本公开实施例中使用乒乓缓存技术进行加速的结构示意图；
- 图 5 为本公开实施例的基于 FPGA 的重测序分析系统的结构示意图二；
- 图 6 为本公开实施例中的运行时序示意图；
- 图 7 为本公开实施例的基于 FPGA 的重测序分析系统的结构示意图三；
- 5 图 8 为本公开实施例三所提供的基于 FPGA 的重测序分析方法的流程示意图；
- 图 9 为本公开实施例四所提供的基于 FPGA 的重测序分析装置的结构示意图；
- 图 10 为本公开实施例五所提供的基于 FPGA 的重测序分析装置的结构示意图。

具体实施方式

- 10 下面详细描述本公开的实施例，所述实施例的示例在附图中示出，其中自始至终相同或类似的标号表示相同或类似的元件或具有相同或类似功能的元件。下面通过参考附图描述的实施例是示例性的，旨在用于解释本公开，而不能理解为对本公开的限制。

下面参考附图描述本公开实施例的基于 FPGA 的重测序分析方法和装置。在具体描述本发明实施例之前，为了便于理解，首先对常用技术词进行介绍：

- 15 **FPGA**，现场可编程逻辑门阵列（Field—Programmable Gate Array），是一种可配置的集成电路，含有可编程布线网络，将逻辑阵列模块、嵌入式存储器模块以及数字信号处理器（Digital Signal Processor，简称 DSP）模块连接在一起，而中央处理器（Central Processing Unit，简称 CPU）和图形处理器（Graphics Processing Unit，简称 GPU）处理程序指令的数据通路和拓扑则是固定不变的。并且，FPGA 资源可以被配置并连接起来，建立定制指令流水线来处理数据，这种体系结构可以非常高效地解决很多类型的计算问题。此外，成熟开发并大规模使用的应用也可以依托 FPGA 为速度更快但更不灵活的专用集成电路（Application Specific Integrated Circuit，简称 ASIC）的设计做基础和验证。
- 20

- 重测序，是指对已知基因组序列的物种进行不同个体的基因组测序，并在此基础上对个体或群体进行差异性分析。针对基因组测序的个体，通过变异检测，可以确定大量的单核苷酸多态性位点，插入缺失位点。
- 25

图 1 为本公开实施例一所提供的基于 FPGA 的重测序分析方法的流程示意图。

本公开实施例以该基于 FPGA 的重测序分析方法被配置于基于 FPGA 的重测序分析装置中来举例说明，该基于 FPGA 的重测序分析装置可以被配置于任一计算机设备中，以使该计算机设备可以执行基于 FPGA 的重测序分析功能。

- 30 如图 1 所示，该基于 FPGA 的重测序分析方法包括以下步骤：  
步骤 101，接收基因组重测序数据。

本公开实施例中，基因组重测序数据包括：基因组重测序基础数据以及描述基础数据

的参数信息。

可选地，可以预先将基因组重测序数据写入预设文件中，在对基因组进行重测序时，可以读取该预设文件，根据该预设文件，可以获取基因组重测序数据。

进一步地，一般预设文件是以压缩文件进行保存的，例如以 gz 等格式进行保存的，因此，在读取该预设文件前，还可以对该预设文件进行解压缩处理。

步骤 102，将重测序数据作为 FPGA 的输入，根据 FPGA 的输出确定重测序过程中的比对结果，同时，对比对结果进行排序去重处理。

本公开实施例中，可以将基因组重测序数据，对比定位到参考基因组上，获取相应的比对位置等信息，而后将获取的信息传输到 FPGA，由 FPGA 进行比对结果的计算，得到对比结果。作为一种可能的实现方式，为了提升比对效率，可以根据 FPGA 的并行处理特性，将基因组重测序数据切分为多个固定长度的基因组重测序子数据，而后同时将多个固定长度的基因组重测序子数据分别与参考基因组进行比对，得到多个比对子结果，最后综合多个比对子结果，可以得到比对结果。

同时，在得到对比结果时，可以将比对结果按照参考基因组上的对比位置进行排序，并对相同对比位置的重测序数据进行去重处理。原因为，重复的重测序数据会影响测序深度等对比评估，最终导致基因突变检测的假阳性，因此，可以对相同对比位置的重测序数据进行去重处理。

可选地，可以调用生物信息学软件和比对脚本，对重测序数据进行比对，确定重测序过程中的比对结果，而后再根据生物信息学软件和排序去重脚本，对比对结果进行排序去重处理，确定排序去重处理后的比对结果。其中，生物信息学软件包括：BuildAlignerIndex、Alignment、SortMarkDup、BuildBQSRIndex、BQSR 和 HaplotypeCaller 等等。

例如，可以基于 FPGA 加速的高性能 DNA 序列对比软件，对重测序数据进程比对，该对比软件可以支持 33 至 160bp 双端短读对比。而后，基于多线程加速的高性能排序、去重软件，对比对结果进行排序去重处理，该排序、去重软件支持管道输入输出，以提升排序去重的效率。

本公开实施例中，对比过程、排序过程和去重处理过程可以同步执行，由此，可以节省程序运行时间，节省计算成本，提升重测序效率。

步骤 103，对排序去重处理后的比对结果进行碱基质量值的校正。

本公开实施例中，为了使得对比结果中的碱基质量值接近真实，且降低与参考基因组之间错配的概率，可以对排序去重处理后的比对结果进行碱基质量值的校正，即将碱基质量值进行数据内部的校准。

可选地，可以调用生物信息学软件和碱基质量值校正脚本，对排序去重处理后的比对结果进行碱基质量值的校正，得到校正后的比对结果。例如，可以基于多线程加速的高性能碱基质量值校正软件，对排序去重处理后的比对结果进行碱基质量值的校正，该校正软件可以支持管道输入输出，以提升校正的效率。

5 步骤 104，根据校正后的比对结果进行变异结果的检测。

本公开实施例中，变异结果的检测包括：单核苷酸多态性位点检测，插入缺失位点检测。可以根据校正后的比对结果及相关信息，进行变异结果的检测。

可选地，可以调用生物信息学软件和变异检测脚本，对校正后的比对结果进行变异结果的检测，确定检测结果。例如，可以基于 FPGA 加速的高性能变异检测软件，对校正后的  
10 的比对结果进行变异结果的检测，得到变异检测结果，以提升变异检测的效率。

本公开实施例的基于 FPGA 的重测序分析，通过接收基因组重测序数据，而后将重测序数据作为 FPGA 的输入，根据 FPGA 的输出确定重测序过程中的比对结果，同时，对比对结果进行排序去重处理，接着，对排序去重处理后的比对结果进行碱基质量值的校正，最后，根据校正后的比对结果进行变异结果的检测。本公开中，对比过程、排序过程和去  
15 重处理过程可以同步执行，由此，可以节省程序运行时间，节省计算成本，提升重测序效率。

作为一种可能的实现方式，在接收到基因组重测序数据后，还可以对基因组重测序数据，基于预设长度序列进行哈希运算，并将进行哈希运算后的基因组重测序数据作为 FPGA  
的输入，由此，可以快速定位到基因组的比对位置，减少内存随机访问，提升对比效率。

20 为了清楚说明上述实施例，本公开提供了另一种基于 FPGA 的重测序分析方法。

图 2 为本公开实施例二所提供的基于 FPGA 的重测序分析方法的流程示意图。

如图 2 所示，该基于 FPGA 的重测序分析方法可以包括以下步骤：

步骤 201，接收基因组重测序数据，其中，基因组重测序数据包括：基因组重测序基础数据，以及描述基础数据的参数信息。

25 例如，参见图 3，图 3 为本公开实施例的基于 FPGA 的重测序分析系统的结构示意图一。其中，重测序分析系统包括：任务调度单元 1、分析单元 2 和结果生成单元 3。任务调度单元 1 用于接收基因组重测序基础数据，以及描述基础数据的参数信息。

步骤 202，根据基础数据，以及描述基础数据的参数信息，确定用于输入比对结果的第一任务，以及确定用于进行排序去重处理的第二任务。

30 本公开实施例中，为了提升重测序效率，从而节省程序运行时间，节省计算成本，对比过程、排序过程和去重处理过程可以同步执行。因此，当接收到基因组重测序数据后，可以根据基础数据，以及描述基础数据的参数信息，同时确定用于输入比对结果的第一任

务，以及确定用于进行排序去重处理的第二任务。

例如，参见图 3，任务调度单元 1 可以根据基础数据，以及描述基础数据的参数信息，同时确定用于输入比对结果的第一任务，以及确定用于进行排序去重处理的第二任务，而后可以将第一任务和第二任务发送至分析单元 2，并流程化调度分析单元 2。例如可以通过生物信息学软件和系统调度脚本，实现流程化调度分析单元 2。

步骤 203，从 FPGA 中分别为第一任务分配对应的线程和从 CPU 中为第二任务分配对应的线程。

本公开实施例中，在确定第一任务和第二任务后，可以从 FPGA 中为第一任务分配对应的线程和从 CPU 中为第二任务分配对应的线程。

步骤 204，启动，并且，并行执行第一任务和第二任务对应的线程，经由各对应的线程调用重测序数据。

本公开实施例中，在确定第一任务和第二任务对应的线程后，可以同时启动并执行第一任务和第二任务对应的线程，经由各对应的线程调用重测序数据，以进行对比过程、排序过程和去重处理过程。

步骤 205，经由第一任务对应的线程的输出结果确定重测序过程中的比对结果，同时，经由第二任务对应的线程的输出结果对比对结果进行排序去重处理。

本公开实施例中，经由第一任务对应的线程调用重测序数据后，可以根据第一任务对应的线程的输出结果，确定重测序过程中的比对结果。作为一种可能的实现方式，可以执行第一任务对应的线程，以确定目标比对位置，而后根据目标比对位置触发进行基因组重测序基础数据中重复碱基对的比对，最后将重复碱基对的比对结果作为重测序过程中的比对结果。例如，参见图 3，分析单元 2 可以通过流程化调用生物信息学软件和对比脚本，并通过对软件加速以及调用 FPGA 中第一任务对应的线程，对基因组重测序数据进行分析，确定对比结果。

本公开实施例中，第二任务对应的线程可以调用对比结果，对比对结果进行排序去重处理。例如，参见图 3，分析单元 2 可以通过流程化调用生物信息学软件和排序去重脚本，并通过对软件加速以及调用 ~~FPGA~~CPU 中第二任务对应的线程，对比对结果进行排序去重处理。

步骤 206，对排序去重处理后的比对结果进行碱基质量值的校正。

步骤 207，根据校正后的比对结果进行变异结果的检测。

例如，参见图 3，可以通过分析单元 2 流程化调用生物信息学软件和碱基质量值校正脚本，对排序去重处理后的比对结果进行碱基质量值的校正。以及，通过分析单元 2 流程化

调用生物信息学软件和变异检测脚本，根据校正后的比对结果进行变异结果的检测，得到变异检测结果。

进一步地，参见图 3，结果生成单元 3，用于在得到变异检测结果后，根据任务调度单元 1 传递的描述基础数据的参数信息，将分析单元 2 得到的变异检测结果发送至指定的位置。

本公开实施例中，通过调用生物信息学软件和个性化分析脚本，并通过对软件加速以及调用 FPGA 中的资源，对基因组重测序数据进行分析，可以提升基因组重测序效率。

作为一种可能的实现方式，可以采用乒乓缓存技术对目标比对位置进行缓存处理，例如，参见图 4，图 4 为本公开实施例中采用乒乓缓存技术进行加速的结构示意图。对于 FPGA 全局内存的数据，采用乒乓缓存技术进行缓存处理。原因为，CPU 端对 FPGA 全局内存的读/写操作，可以与 FPGA 内核模块运算同步进行，提升运行效率。

如果使用传统的“写全局内存→运行内核→读全局内存”串行运行方式，FPGA 的运算模块有一段时间将处于等待 I/O 状态，导致运行效率较低，因此，采用乒乓缓存技术对 FPGA 全局内存的数据，比如目标比对位置进行缓存处理，可以使得 FPGA 的运算模块始终处于高效的工作状态，并且可以使得单个任务的多个子任务流水化执行，实现任务的并行处理。例如，参见图 4，在时间片 C3，任务 T1 从输出缓存 1 的读回、任务 T2 使用缓存 2 数据的运算和任务 T3 对输入缓存 1 的写入可以同时进行。

作为一种示例，参见图 5，图 5 为本公开实施例的基于 FPGA 的重测序分析系统的结构示意图二。该重测序分析系统包括：比对模块 4、排序去重模块 5、碱基质量值校正模块 6 以及变异检测模块 7。

其中，比对模块 4，用于将接收的基因组重测序数据对比到参考基因组，而后将比对后的数据传输至 FPGA 进行计算，得到比对结果，比对结果可以以 SAM 格式的文件进行存储。

排序去重模块 5，用于将对模块 4 输出的数据按照比对位置进行排序，而后标记并过滤重复的数据，排序去重结果可以以 BAM 格式的文件进行存储。

碱基质量值校正模块 6，用于将排序去重模块 5 输出的数据的碱基质量值进行校正，校正结果可以以 BAM 格式的文件进行存储。

变异检测模块 7，用于通过调用生物信息学软件和分析脚本，对碱基质量值校正模块 6 输出的数据，进行变异检测，包含单核苷酸多态性位点检测，插入缺失位点检测，检测结果可以 VCF 或 gVCF 格式的文件进行存储。

作为一种示例，参见图 6，图 6 为本公开实施例中的运行时序示意图。图 6 中，Fastq 输入文件，通过调整相邻步骤间数据的计算模式，并增加相应的输入、输出缓冲区，将整个计算流程优化成流水线结构。

具体地，对比模块在程序开始时运行，将输入文件中的基因组重测数据比对到参考基因组，对比结果输出到排序去重模块。

排序去重模块，与比对模块同步运行，将对模块输出数据按照比对位置进行排序，而后标记并过滤重复的数据。

- 5        碱基质量值校正模块，在比对模块完成后运行，运行时与排序去重模块同步运行，用于将排序去重后的数据的碱基质量值进行校正。

变异检测模块，用于通过调用生物信息学软件和分析脚本，对碱基质量值校正模块输出的数据，进行变异检测，包含单核苷酸多态性位点检测，插入缺失位点检测。

- 10       本公开实施例中，通过流程化调用生物信息学软件和分析脚本，加速基因组重测序分析过程，提高了基因组重测序效率。

作为一种示例，参见图 7，图 7 为本公开实施例的基于 FPGA 的重测序分析系统的结构示意图三。

- 15       CPU 端读取输入文件（Fastq），将需要进行计算的数据存放在内存中，通过总线 and 接口标准，将内存中的数据传输至 FPGA 的全局内存，在 FPGA 中进行输入数据的处理、比对和结果记录 and 输出，结果存放在全局内存中，之后，FPGA 再通过总线 and 接口标准将处理后的数据传输至 CPU 端内存进行后续处理，输出 SAM 格式的文件。

本公开实施例中，采用 CPU 和 FPGA 异构的方式，实现基因组重测数据的高速比对，通过 CPU 多线程技术可以实现数据的并行化预处理和后续处理，FPGA 采用多个功能模块以通道连接的方式，实现数据的细粒度流水操作。

- 20       作为一种可能的实现方式，在 FPGA 中，I/O 读写操作可以和压缩/解压缩运算剥离，利用循环缓存和多线程，可以实现压缩/解压缩运算并行计算，提升基因组重测序效率。

作为一种示例，参见图 8，图 8 为本公开实施例三所提供的基于 FPGA 的重测序分析方法的流程图。

如图 8 所示，该基于 FPGA 的重测序分析方法可以包括以下步骤：

- 25       300，任务开始。

301，程序调用，参数设置及初始化。具体地，可以将所有的参数初始化后，开始执行任务调度，接收基因组重测序基础数据，以及描述基础数据的参数信息。

302，判断是否构建比对索引，若是，则执行步骤 316。

303，判断是否构建碱基质量值校正索引，若是，执行步骤 313。

- 30       304 或 317，判断是否比对，若是，则执行步骤 318。

305，判断是否排序去重，若是，则执行 319。

306，判断是否对碱基质量值进行校正，若是，执行步骤 320。

307，判断是否比对排序去重，若是，执行步骤 321。

308, 判断是否拷贝比对结果, 若是, 执行步骤 322。

309 或 314, 判断是否对碱基质量值校正, 若是, 执行步骤 315。

310, 判断是否拷贝碱基质量值校正结果, 若是, 执行步骤 323。

311, 判断是否进行变异检测, 若是, 执行步骤 324。

5 312, 等待所有进程执行完成。

313, 构建碱基质量值校正索引。

315, 调用碱基质量值校正。

316, 构建比对索引, 用于将基因组重测序数据快速比对到参考基因组。

318, 调用比对, 在 CPU 端将基因组重测序数据对比到参考基因组上, 获取相应的比  
10 对位置等信息, 并将获取的信息传输至 FPGA 进行比对计算, FPGA 将比对结果在发送至 CPU 端。

319, 调用排序去重。

320, 调用碱基质量值校正, 使得碱基质量值能够接近真实, 且降低与参考基因组之间的错配率。

15 321, 调用比对排序去重, 利用比对位置等信息过滤比对结果中的重复基因序列, 重复基因序列主要是由于测序 PCR 产生的, 会影响测序深度等对比评估, 最终导致基因突变检测的假阳性, 比对结果一般以 SAM/BAM 的格式文件进行存储。

322, 调用拷贝对比结果, 可以在后台进行调用。

323, 调用拷贝碱基质量值校正结果, 可以在后台进行调用。

20 324, 调用变异检测。

325, 任务结束。

需要说明的是, 在硬件上, 还可以使用 FPGA 加速卡实现提升重测序效率, 或者, 使用 GPU 实现提升重测序效率, 并且, 本公开的基于 FPGA 的重测序分析系统由多个模块构成, 对任一模块方法的改进、变形、替代, 例如, 使用 bwt 算法替换对比, 皆属于本公开  
25 的保护范围内, 因此, 无论采用硬件加速 bwt 算法的比对, 还是直接采用 bwt 算法替代对比, 皆可实现本公开所实现的技术效果。

作为一种示例:

使用如下已公开数据进行测试:

数据网页路径如下: <http://gigadb.org/dataset/100274>

30 数据路径:

[ftp://parrot.genomics.cn/gigadb/pub/10.5524/100001\\_101000/100274/CL100008589\\_L02\\_r  
ead\\_1.fq.gz](ftp://parrot.genomics.cn/gigadb/pub/10.5524/100001_101000/100274/CL100008589_L02_read_1.fq.gz)

[ftp://parrot.genomics.cn/gigadb/pub/10.5524/100001\\_101000/100274/CL100008589\\_L02\\_r](ftp://parrot.genomics.cn/gigadb/pub/10.5524/100001_101000/100274/CL100008589_L02_r)

ead\_2.fq.gz

相关技术中流程如下 BWA+GATK:

1、比对排序去重

比对和排序:

```
5 ./BIN/bwa-0.7.13/bwa mem -t 72 -M -Y \
    -R "@RG\tID:PE100\tPL:Illumina\tPU:PE100\tLB:NA12878\tSM:PE100\tCN:NIST" \
    ./REF_hs37d5/hs37d5/hs37d5.fa \
    ./read_1.fq.gz ./read_2.fq.gz \
    | ./BIN/samtools-1.3/bin/samtools view -buhS \
10 -t ./REF_hs37d5/hs37d5/hs37d5.fa.fai - \
    | ./BIN/samtools-1.3/bin/samtools sort -m 1000000000 \
    -T ./PE100.sort \
    -o ./PE100.1.sorted.bam -
```

去重:

```
15 ./BIN/jdk1.8.0_131/bin/java -jar ./BIN/picard.jar MarkDuplicates \
    I=./PE100.1.sorted.bam \
    O=./PE100.2.rmdup.bam \
    M=./PE100.2.rmdup.mat \
    REMOVE_DUPLICATES=false \
```

```
20 TMP_DIR=./tmp
```

建索引 index:

```
./BIN/samtools-1.3/bin/samtools index \
    ./PE100.2.rmdup.bam
```

2、碱基质量值校正

```
25 ./BIN/jdk1.8.0_131/bin/java
    -jar ./GATK_stat/lib/GenomeAnalysisTK.3.7-0/GenomeAnalysisTK.jar \
    --disable_auto_index_creation_and_locking_when_reading_rods \
    -T BaseRecalibrator \
    -nct 72 \
30 -R REF_hs37d5/hs37d5/hs37d5.fa \
    -I ./PE100.2.rmdup.bam \
    -o ./PE100.4.bqsr.grp \
    -knownSites ./REF_hs37d5/GATK.b37/Mills_and_1000G_gold_standard.indels.b37.vcf
```

```

\
    -knownSites ./REF_hs37d5/GATK.b37/1000G_phase1.indels.b37.vcf\
    -knownSites ./REF_hs37d5/dbsnp_GRCh37_b149/dbsnp-149.vcf
PrintReads:
5  ./BIN/jdk1.8.0_131/bin/java
    -jar ./GATK_stat/lib/GenomeAnalysisTK.3.7-0/GenomeAnalysisTK.jar \
        --disable_auto_index_creation_and_locking_when_reading_rods \
        -T PrintReads \
        -nct 72 \
10  -R REF_hs37d5/hs37d5/hs37d5.fa \
        -I ./PE100.2.rmdup.bam \
        -BQSR ./PE100.4.bqsr.grp \
        -o ./PE100.4.bqsr.bam \
        -baqGOP 30
15  建索引 index:
    ./BIN/samtools-1.3/bin/samtools index \
        ./PE100.4.bqsr.bam
3、变异检测
    ./BIN/jdk1.8.0_131/bin/java
20  -jar ./GATK_stat/lib/GenomeAnalysisTK.3.7-0/GenomeAnalysisTK.jar \
        --disable_auto_index_creation_and_locking_when_reading_rods \
        -R REF_hs37d5/hs37d5/hs37d5.fa \
        -T HaplotypeCaller \
        -nct 72 \
25  -I ./PE100.4.bqsr.bam \
        --dbsnp ./REF_hs37d5/dbsnp_GRCh37_b149/dbsnp-149.vcf \
        -stand_call_conf 10.0 \
        -o ./raw.snp_indel.vcf
    本公开流程如下所示:
30  MegaBOLT -type allwithindex -list ./my.list -scalafile ./my.scala
    -ref ./REF_hs37d5/hs37d5/hs37d5.fa -alignerindex ./REF_hs37d5/hs37d5/hs37d5.fa
    -bqsrindex ./REF_hs37d5/hs37d5/bqsrRef
    -vcf ./REF_hs37d5/dbsnp_GRCh37_b149/dbsnp-149.vcf

```

表 1，流程运行时间对比

| 数据集      | 碱基数 (Gbp) | 流程       | 运行时间(小时) |
|----------|-----------|----------|----------|
| 双端 100bp | 114       | MegaBOLT | 2.4      |
|          |           | BWA+GATK | 59.0     |

表 2，变异检测结果评估对比

| 数据集      | 流程       | SNPs      |        |       |        |        | INDELs  |        |        |        |        |
|----------|----------|-----------|--------|-------|--------|--------|---------|--------|--------|--------|--------|
|          |          | 真阳性       | 假阳性    | 假阴性   | 准确性    | 敏感性    | 真阳性     | 假阳性    | 假阴性    | 准确性    | 敏感性    |
| 双端 100bp | MegaBOLT | 3,201,059 | 15,401 | 9,198 | 99.52% | 99.71% | 461,465 | 27,966 | 19,801 | 94.29% | 95.89% |
|          | BWA+GATK | 3,201,095 | 6,987  | 9,162 | 99.78% | 99.71% | 462,450 | 17,681 | 18,817 | 96.32% | 96.09% |

从以上结果可以看出，如表 1 所示，本公开在运行时间上比相关技术中流程加速约 20 倍；如表 2 所示，本公开相比相关技术中流程，在 SNP 的准确性和敏感性上基本保持一致，INDEL 的准确性上会比相关技术中流程差一些，敏感性基本保持一致，总体结果与相关技术中流程相当。

为了实现上述实施例，本公开还提出一种基于 FPGA 的重测序分析装置。

图 9 为本公开实施例四所提供的基于 FPGA 的重测序分析装置的结构示意图。

如图 9 所示，该基于 FPGA 的重测序分析装置 100 包括：接收模块 110、处理模块 120、校正模块 130，以及检测模块 140。

其中，接收模块 110，用于接收基因组重测序数据。

处理模块 120，用于将重测序数据作为 FPGA 的输入，根据 FPGA 的输出确定重测序过程中的比对结果，同时，对比对结果进行排序去重处理。

校正模块 130，用于对排序去重处理后的比对结果进行碱基质量值的校正。

检测模块 140，用于根据校正后的比对结果进行变异结果的检测。

进一步地，在本公开实施例的一种可能的实现方式中，参见图 10，在图 9 所示实施例的基础上，该基于 FPGA 的重测序分析装置 100 还可以包括：

运算模块 150，用于在在接收基因组重测序数据之后，对基因组重测序数据，基于预设长度序列进行哈希运算。

相应的，处理模块 120，具体用于：将进行哈希运算后的基因组重测序数据作为 FPGA 的输入。

作为一种可能的实现方式，基因组重测序数据包括：基因组重测序基础数据，以及描述基础数据的参数信息。

确定模块 160, 用于在接收基因组重测序数据之后, 根据基础数据, 以及描述基础数据的参数信息, 确定用于输入比对结果的第一任务, 以及确定用于进行排序去重处理的第二任务。

5 分配模块 170, 用于从 FPGA 中为第一任务分配对应的线程和从 CPU 中为第二任务分配对应的线程。

处理模块 120, 包括:

启动执行单元 121, 用于启动, 并且, 并行执行第一任务和第二任务对应的线程, 经由各对应的线程调用重测序数据。

10 确定处理单元 122, 用于经由第一任务对应的线程的输出结果确定重测序过程中的比对结果, 同时, 经由第二任务对应的线程的输出结果对比对结果进行排序去重处理。

作为一种可能的实现方式, 确定处理单元 122, 具体用于: 执行第一任务对应的线程, 以确定目标比对位置; 根据目标比对位置触发进行基因组重测序基础数据中重复碱基对的比对; 将重复碱基对的比对结果作为重测序过程中的比对结果。

15 作为另一种可能的实现方式, 确定处理单元 122, 还用于: 采用乒乓缓存技术对目标比对位置进行缓存处理。

需要说明的是, 前述对基于 FPGA 的重测序分析方法实施例的解释说明也适用于该实施例的基于 FPGA 的重测序分析装置 100, 此处不再赘述。

20 本公开实施例的基于 FPGA 的重测序分析装置, 通过接收基因组重测序数据, 而后将重测序数据作为 FPGA 的输入, 根据 FPGA 的输出确定重测序过程中的比对结果, 同时, 对比对结果进行排序去重处理, 接着, 对排序去重处理后的比对结果进行碱基质量值的校正, 最后, 根据校正后的比对结果进行变异结果的检测。本公开中, 对比过程、排序过程和去重处理过程可以同步执行, 由此, 可以节省程序运行时间, 节省计算成本, 提升重测序效率。

25 为了实现上述实施例, 本公开还提出一种计算机设备, 包括: 存储器、处理器及存储在存储器上并可在处理器上运行的计算机程序, 处理器执行程序时, 实现如本公开前述实施例提出的基于 FPGA 的重测序分析方法。

为了实现上述实施例, 本公开还提出一种计算机可读存储介质, 计算机可读存储介质存储有计算机可读指令, 计算机可读指令用于使计算机执行本公开前述实施例提出的基于 FPGA 的重测序分析方法。

30 在本说明书的描述中, 参考术语“一个实施例”、“一些实施例”、“示例”、“具体示例”、或“一些示例”等的描述意指结合该实施例或示例描述的具体特征、结构、材料或者特点包含于本公开的至少一个实施例或示例中。在本说明书中, 对上述术语的示意性表述不必须

针对的是相同的实施例或示例。而且，描述的具体特征、结构、材料或者特点可以在任一个或多个实施例或示例中以合适的方式结合。此外，在不相互矛盾的情况下，本领域的技术人员可以将本说明书中描述的不同实施例或示例以及不同实施例或示例的特征进行结合和组合。

- 5       此外，术语“第一”、“第二”仅用于描述目的，而不能理解为指示或暗示相对重要性或者隐含指明所指示的技术特征的数量。由此，限定有“第一”、“第二”的特征可以明示或者隐含地包括至少一个该特征。在本公开的描述中，“多个”的含义是至少两个，例如两个，三个等，除非另有明确具体的限定。

- 10       流程图中或在此以其他方式描述的任何过程或方法描述可以被理解为，表示包括一个或多个用于实现定制逻辑功能或过程的步骤的可执行指令的代码的模块、片段或部分，并且本公开的优选实施方式的范围包括另外的实现，其中可以不按所示出或讨论的顺序，包括根据所涉及的功能按基本同时的方式或按相反的顺序，来执行功能，这应被本公开的实施例所属技术领域的技术人员所理解。

- 15       在流程图中表示或在此以其他方式描述的逻辑和/或步骤，例如，可以被认为用于实现逻辑功能的可执行指令的定序列表，可以具体实现在任何计算机可读介质中，以供指令执行系统、装置或设备（如基于计算机的系统、包括处理器的系统或其他可以从指令执行系统、装置或设备取指令并执行指令的系统）使用，或结合这些指令执行系统、装置或设备而使用。就本说明书而言，“计算机可读介质”可以是任何可以包含、存储、通信、传播或传输程序以供指令执行系统、装置或设备或结合这些指令执行系统、装置或设备而使用的装置。计算机可读介质的更具体的示例（非穷尽性列表）包括以下：具有一个或多个布线的电连接部（电子装置），便携式计算机盘盒（磁装置），随机存取存储器（RAM），只读存储器（ROM），可擦除可编程只读存储器（EPROM 或闪速存储器），光纤装置，以及便携式光盘只读存储器（CDROM）。另外，计算机可读介质甚至可以是可在其上打印所述程序的纸或其他合适的介质，因为可以例如通过对纸或其他介质进行光学扫描，接着进行编辑、解译或必要时以其他合适方式进行处理来以电子方式获得所述程序，然后将其存储在计算机存储器中。
- 20
- 25

- 30       应当理解，本公开的各部分可以用硬件、软件、固件或它们的组合来实现。在上述实施方式中，多个步骤或方法可以用存储在存储器中且由合适的指令执行系统执行的软件或固件来实现。如，如果用硬件来实现和在另一实施方式中一样，可用本领域公知的下列技术中的任一项或他们的组合来实现：具有用于对数据信号实现逻辑功能的逻辑门电路的离散逻辑电路，具有合适的组合逻辑门电路的专用集成电路，可编程门阵列（PGA），现场可编程门阵列（FPGA）等。

本技术领域的普通技术人员可以理解实现上述实施例方法携带的全部或部分步骤是可

以通过程序来指令相关的硬件完成，所述的程序可以存储于一种计算机可读存储介质中，该程序在执行时，包括方法实施例的步骤之一或其组合。

此外，在本公开各个实施例中的各功能单元可以集成在一个处理模块中，也可以是各个单元单独物理存在，也可以两个或两个以上单元集成在一个模块中。上述集成的模块既可以采用硬件的形式实现，也可以采用软件功能模块的形式实现。所述集成的模块如果以软件功能模块的形式实现并作为独立的产品销售或使用，也可以存储在一个计算机可读存储介质中。

上述提到的存储介质可以是只读存储器，磁盘或光盘等。尽管上面已经示出和描述了本公开的实施例，可以理解的是，上述实施例是示例性的，不能理解为对本公开的限制，本领域的普通技术人员在本公开的范围内可以对上述实施例进行变化、修改、替换和变型。

## 权利要求书

1、一种基于 FPGA 的重测序分析方法，其特征在于，包括以下步骤：

接收基因组重测序数据；

5 将所述重测序数据作为 FPGA 的输入，根据所述 FPGA 的输出确定重测序过程中的比对结果，同时，对所述比对结果进行排序去重处理；

对排序去重处理后的比对结果进行碱基质量值的校正；以及

根据校正后的比对结果进行变异结果的检测。

2、如权利要求 1 所述的基于 FPGA 的重测序分析方法，其特征在于，在所述接收基因组重测序数据之后，还包括：

对所述基因组重测序数据，基于预设长度序列进行哈希运算；

所述将所述重测序数据作为 FPGA 的输入，包括：

将进行所述哈希运算后的基因组重测序数据作为所述 FPGA 的输入。

3、如权利要求 1 所述的基于 FPGA 的重测序分析方法，其特征在于，所述基因组重测序数据包括：基因组重测序基础数据，以及描述所述基础数据的参数信息，在所述接收基因组重测序数据之后，还包括：

根据所述基础数据，以及描述所述基础数据的参数信息，确定用于输入比对结果的第一任务，以及确定用于进行排序去重处理的第二任务；

从所述 FPGA 中为所述第一任务分配对应的线程和从 CPU 中为所述第二任务分配对应的线程。

4、如权利要求 3 所述的基于 FPGA 的重测序分析方法，其特征在于，所述将所述重测序数据作为 FPGA 的输入，根据所述 FPGA 的输出确定重测序过程中的比对结果，同时，对所述比对结果进行排序去重处理，包括：

启动，并且，并行执行所述第一任务和所述第二任务对应的线程，经由各所述对应的线程调用所述重测序数据；

经由所述第一任务对应的线程的输出结果确定重测序过程中的比对结果，同时，经由所述第二任务对应的线程的输出结果对所述比对结果进行排序去重处理。

5、如权利要求 4 所述的基于 FPGA 的重测序分析方法，其特征在于，所述经由所述第一任务对应的线程的输出结果确定重测序过程中的比对结果，包括：

30 执行所述第一任务对应的线程，以确定目标比对位置；

根据所述目标比对位置触发进行所述基因组重测序基础数据中重复碱基对的比对；

将所述重复碱基对的比对结果作为所述重测序过程中的比对结果。

6、如权利要求4所述的基于FPGA的重测序分析方法，其特征在于，在所述确定目标比对位置之后，还包括：

采用乒乓缓存技术对所述目标比对位置进行缓存处理。

5 7、一种基于FPGA的重测序分析装置，其特征在于，包括：

接收模块，用于接收基因组重测序数据；

处理模块，用于将所述重测序数据作为FPGA的输入，根据所述FPGA的输出确定重测序过程中的比对结果，同时，对所述比对结果进行排序去重处理；

校正模块，用于对排序去重处理后的比对结果进行碱基质量值的校正；以及

10 检测模块，用于根据校正后的比对结果进行变异结果的检测。

8、如权利要求7所述的基于FPGA的重测序分析装置，其特征在于，所述装置还包括：

运算模块，用于在在所述接收基因组重测序数据之后，对所述基因组重测序数据，基于预设长度序列进行哈希运算；

所述处理模块，具体用于：

15 将进行所述哈希运算后的基因组重测序数据作为所述FPGA的输入。

9、如权利要求7所述的基于FPGA的重测序分析装置，其特征在于，所述基因组重测序数据包括：基因组重测序基础数据，以及描述所述基础数据的参数信息，所述装置还包括：

20 确定模块，用于在所述接收基因组重测序数据之后，根据所述基础数据，以及描述所述基础数据的参数信息，确定用于输入比对结果的第一任务，以及确定用于进行排序去重处理的第二任务；

分配模块，用于从所述FPGA中为所述第一任务分配对应的线程和从CPU中为所述第二任务分配对应的线程。

25 10、如权利要求9所述的基于FPGA的重测序分析装置，其特征在于，所述处理模块，包括：

启动执行单元，用于启动，并且，并行执行所述第一任务和所述第二任务对应的线程，经由各所述对应的线程调用所述重测序数据；

30 确定处理单元，用于经由所述第一任务对应的线程的输出结果确定重测序过程中的比对结果，同时，经由所述第二任务对应的线程的输出结果对所述比对结果进行排序去重处理。

11、如权利要求10所述的基于FPGA的重测序分析装置，其特征在于，所述确定处理

单元，具体用于：

执行所述第一任务对应的线程，以确定目标比对位置；

根据所述目标比对位置触发进行所述基因组重测序基础数据中重复碱基对的比对；

将所述重复碱基对的比对结果作为所述重测序过程中的比对结果。

- 5        12、如权利要求 10 所述的基于 FPGA 的重测序分析装置，其特征在于，所述确定处理单元，还用于：

采用乒乓缓存技术对所述目标比对位置进行缓存处理。

- 13、一种计算机设备，其特征在于，包括：存储器、处理器及存储在存储器上并可在处理器上运行的计算机程序，所述处理器执行所述程序时，实现如权利要求 1-6 中任一所述的基于 FPGA 的重测序分析方法。
- 10

14、一种计算机可读存储介质，其特征在于，所述计算机可读存储介质存储有计算机可读指令，所述计算机可读指令用于使计算机执行权利要求 1-6 中任一所述的基于 FPGA 的重测序分析方法。