(19)国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 119889444 A
(43)申请公布日 2025.04.25

(21)申请号 202411721596.4

(22)申请日 2020.05.22

(30)优先权数据
62/852,965 2019.05.24 US

(62)分案原申请数据
202080003658.1 2020.05.22

(71)申请人 伊鲁米那股份有限公司
地址 美国加利福尼亚州

(72)发明人 迈克尔·吕勒

(74)专利代理机构 广州三环专利商标代理有限公司 44202
专利代理师 郝传鑫

(51)Int.Cl.
G16B 30/10(2019.01)
G16B 20/30(2019.01)

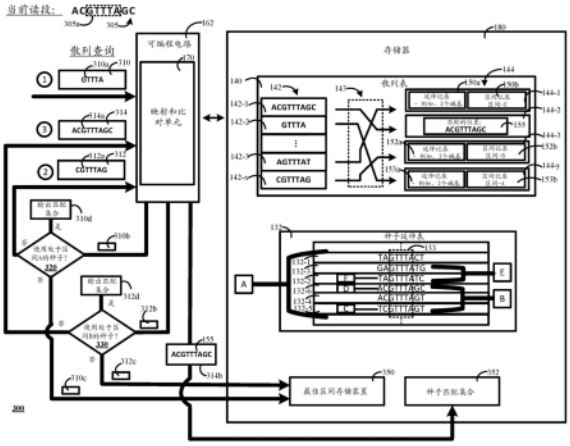
权利要求书6页 说明书36页 附图8页

(54)发明名称

用于散列表基因组映射的灵活种子延伸

(57)摘要

本发明公开了用于生成和使用被配置为改进读段映射的散列表的方法、系统和装置,包括计算机程序,该方法、系统和装置包括:从参考序列获得K个核苷酸的第一种子,生成具有节点的种子延伸树,其中该节点中的每个节点对应于(i)延伸种子,该延伸种子是第一种子的延伸并具有K*的核苷酸长度,和(ii)种子延伸表中的一个或多个位置,该一个或多个位置包括描述与该延伸种子匹配的参考序列位置的数据,并且对于每个节点:在该散列表的对应于延伸种子的索引键的位置存储区间信息,其中该区间信息引用该种子延伸表中的一个或多个位置,该一个或多个位置包括与和该节点相关联的该延伸种子匹配的参考序列位置。



CN 119889444 A

1.一种使用散列表来改进样品读段到参考序列的映射的方法,所述方法包括:

由映射和比对单元执行对散列表的查询,所述查询包括第一种子,其中所述第一种子包括从所述样品读段中的特定读段获得的核苷酸的子集;

由所述映射和比对单元获得对所执行的查询的响应,所述响应包括由所述散列表的被确定为响应于所述查询的位置存储的信息;

由所述映射和比对单元确定对所执行的查询的所述响应是否包括(i)延伸记录和(ii)区间记录,其中所述区间记录识别存储在延伸表中的与所述第一种子匹配的参考序列位置的连续集合;

基于由所述映射和比对单元确定对所执行的查询的所述响应包括(i)延伸记录和(ii)区间记录:

由所述映射和比对单元确定是否要存取所述延伸表以获得所述延伸表中的由所述区间记录引用的一个或多个匹配的参考序列位置;并且

基于确定要存取所述延伸表:由所述映射和比对单元存取所述延伸表以获得所述延伸表中的由所述区间记录引用的所述一个或多个匹配的参考序列位置;并且由所述映射和比对单元将所述一个或多个匹配的参考序列位置添加到种子匹配集合。

2.如权利要求1所述的方法,所述方法还包括:

基于确定不存取所述延伸表:

由所述映射和比对单元确定是否要将描述所述区间记录的第一信息作为描述候选最佳区间的信息存储在存储器设备中;

由所述映射和比对单元使用所述延伸记录生成第一延伸种子,所述第一延伸种子是所述第一种子的延伸;

由所述映射和比对单元生成包括所述第一延伸种子的后续散列查询;并且

由所述映射和比对单元执行对所述散列表的后续散列查询。

3.如权利要求2所述的方法,所述方法还包括:

由所述映射和比对单元确定对所执行的查询的所述响应包括一个或多个匹配的参考序列位置;并且

基于由所述映射和比对单元确定对所执行的查询的所述响应包括一个或多个匹配的参考序列位置:由所述映射和比对单元将所述一个或多个匹配的参考序列位置添加到种子匹配集合。

4.如权利要求2所述的方法,其中,由所述映射和比对单元确定是否要将描述所述区间记录的所述第一信息作为描述候选最佳区间的信息存储在存储器设备中包括:

由所述映射和比对单元确定不存在将区间记录描述为用于所述特定读段的候选最佳区间的先前信息;并且

由所述映射和比对单元将描述所述区间记录的所述第一信息作为描述候选最佳区间的信息存储在所述存储器设备中。

5.如权利要求2所述的方法,所述方法还包括:

由所述映射和比对单元获得对所述后续执行的查询的响应,所述响应包括由所述散列表的被确定为响应于所述查询的位置存储的信息;

由所述映射和比对单元确定对所述后续执行的查询的所述响应是否包括:(i)第二延

伸记录,(ii)第二区间记录,和(iii)一个或多个匹配的参考序列位置;

基于由所述映射和比对单元确定对所述后续执行的查询的所述响应包括(i)所述第二延伸记录和(ii)所述第二区间记录:

由所述映射和比对单元确定是否要存取延伸表以获得所述延伸表中的由所述第二区间记录引用的一个或多个匹配的参考序列位置;

基于确定不存取所述延伸表:使用一个或多个启发式规则并且由所述映射和比对单元确定描述所述第二区间记录的第二信息或描述所述候选最佳区间的所述第一信息是否将用作所述候选最佳区间;由所述映射和比对单元使用所述第二延伸记录生成第二延伸种子,所述第二延伸种子是所述第一延伸种子的延伸;由所述映射和比对单元生成包括所述第二延伸种子的第三散列查询;并且由所述映射和比对单元执行对所述散列表的包括所述第二延伸种子的第三查询。

6.如权利要求5所述的方法,其中,使用一个或多个启发式规则并且由所述映射和比对单元确定描述所述第二区间记录的所述第二信息或描述所述候选最佳区间的所述第一信息是否将用作所述最佳区间包括:

基于多个因素来选择描述所述第二区间记录的所述第二信息或描述所述候选最佳区间记录的所述第一信息,所述多个因素包括:(i)由所述区间记录和所述第二区间记录中的每一者返回的匹配的参考序列位置的数量,(ii)参考序列位置的预先确定的阈值级别,或(iii)到达存储所述区间记录和所述第二区间记录的所述散列位置的相应种子的每个种子长度。

7.一种使用散列表来改进样品读段到参考序列的映射的系统,所述系统包括一个或多个计算机及存储指令的一个或多个存储设备,

所述指令在由所述一个或多个计算机执行时,可操作以使得所述一个或多个计算机执行操作,所述操作包括:

由映射和比对单元执行对散列表的查询,所述查询包括第一种子,其中所述第一种子包括从所述样品读段中的特定读段获得的核苷酸的子集;

由所述映射和比对单元获得对所执行的查询的响应,所述响应包括由所述散列表的被确定为响应于所述查询的位置存储的信息;

由所述映射和比对单元确定对所执行的查询的所述响应是否包括(i)延伸记录和(ii)区间记录,其中所述区间记录识别存储在延伸表中的与所述第一种子匹配的参考序列位置的连续集合;

基于由所述映射和比对单元确定对所执行的查询的所述响应包括(i)延伸记录和(ii)区间记录:

由所述映射和比对单元确定是否要存取所述延伸表以获得所述延伸表中的由所述区间记录引用的一个或多个匹配的参考序列位置;并且

基于确定要存取所述延伸表:由所述映射和比对单元存取所述延伸表以获得所述延伸表中的由所述区间记录引用的所述一个或多个匹配的参考序列位置;并且由所述映射和比对单元将所述一个或多个匹配的参考序列位置添加到种子匹配集合。

8.如权利要求7所述的系统,所述操作还包括:

基于确定不存取所述延伸表:

由所述映射和比对单元确定是否要将描述所述区间记录的第一信息作为描述候选最佳区间的信息存储在存储器设备中;

由所述映射和比对单元使用所述延伸记录生成第一延伸种子,所述第一延伸种子是所述第一种子的延伸;

由所述映射和比对单元生成包括所述第一延伸种子的后续散列查询;并且

由所述映射和比对单元执行对所述散列表的后续散列查询。

9.如权利要求8所述的系统,所述操作还包括:

由所述映射和比对单元确定对所执行的查询的所述响应包括一个或多个匹配的参考序列位置;并且

基于由所述映射和比对单元确定对所执行的查询的所述响应包括一个或多个匹配的参考序列位置:由所述映射和比对单元将所述一个或多个匹配的参考序列位置添加到种子匹配集合。

10.如权利要求8所述的系统,其中,由所述映射和比对单元确定是否要将描述所述区间记录的所述第一信息作为描述候选最佳区间的信息存储在存储器设备中包括:

由所述映射和比对单元确定不存在将区间记录描述为用于所述特定读段的候选最佳区间的先前信息;并且

由所述映射和比对单元将描述所述区间记录的所述第一信息作为描述候选最佳区间的信息存储在所述存储器设备中。

11.如权利要求8所述的系统,所述操作还包括:

由所述映射和比对单元获得对所述后续执行的查询的响应,所述响应包括由所述散列表的被确定为响应于所述查询的位置存储的信息;

由所述映射和比对单元确定对所述后续执行的查询的所述响应是否包括:(i)第二延伸记录,(ii)第二区间记录,和(iii)一个或多个匹配的参考序列位置;

基于由所述映射和比对单元确定对所述后续执行的查询的所述响应包括(i)所述第二延伸记录和(ii)所述第二区间记录:

由所述映射和比对单元确定是否要存取延伸表以获得所述延伸表中的由所述第二区间记录引用的一个或多个匹配的参考序列位置;

基于确定不存取所述延伸表:使用一个或多个启发式规则并且由所述映射和比对单元确定描述所述第二区间记录的第二信息或描述所述候选最佳区间的所述第一信息是否将用作所述候选最佳区间;由所述映射和比对单元使用所述第二延伸记录生成第二延伸种子,所述第二延伸种子是所述第一延伸种子的延伸;由所述映射和比对单元生成包括所述第二延伸种子的第三散列查询;并且由所述映射和比对单元执行对所述散列表的包括所述第二延伸种子的第三查询。

12.如权利要求11所述的系统,其中,使用一个或多个启发式规则并且由所述映射和比对单元确定描述所述第二区间记录的所述第二信息或描述所述候选最佳区间的所述第一信息是否将用作所述最佳区间包括:

基于多个因素来选择描述所述第二区间记录的所述第二信息或描述所述候选最佳区间记录的所述第一信息,所述多个因素包括:(i)由所述区间记录和所述第二区间记录中的每一者返回的匹配的参考序列位置的数量,(ii)参考序列位置的预先确定的阈值级别,或

(iii)到达存储所述区间记录和所述第二区间记录的所述散列位置的相应种子的每个种子长度。

13.如权利要求7所述的系统,其中所述区间记录引用所述种子延伸表中的一个或多个位置,所述一个或多个位置包括描述与所述查询的所述第一种子匹配的参考序列位置的数据。

14.如权利要求13所述的系统,其中所述种子延伸表中的包括描述与所述查询的所述第一种子匹配的参考序列位置的数据的所述一个或多个位置包括:

延伸表中的与所述查询的所述第一种子匹配的参考序列位置的连续区间。

15.一种存储软件的非暂态计算机可读介质,所述软件包括能够由一个或多个计算机执行的指令,所述指令在经历此类执行时,使得所述一个或多个计算机执行操作,所述操作包括:

由映射和比对单元执行对散列表的查询,所述查询包括第一种子,其中所述第一种子包括从所述样品读段中的特定读段获得的核苷酸的子集;

由所述映射和比对单元获得对所执行的查询的响应,所述响应包括由所述散列表的被确定为响应于所述查询的位置存储的信息;

由所述映射和比对单元确定对所执行的查询的所述响应是否包括(i)延伸记录和(ii)区间记录,其中所述区间记录识别存储在延伸表中的与所述第一种子匹配的参考序列位置的连续集合;

基于由所述映射和比对单元确定对所执行的查询的所述响应包括(i)延伸记录和(ii)区间记录:

由所述映射和比对单元确定是否要存取延伸表以获得所述延伸表中的由所述区间记录引用的一个或多个匹配的参考序列位置;

基于确定要存取所述延伸表:由所述映射和比对单元存取所述延伸表以获得所述延伸表中的由所述区间记录引用的所述一个或多个匹配的参考序列位置;并且由所述映射和比对单元将所述一个或多个匹配的参考序列位置添加到种子匹配集合。

16.如权利要求15所述的计算机可读介质,所述操作还包括:

基于确定不存取所述延伸表:

由所述映射和比对单元确定是否要将描述所述区间记录的第一信息作为描述候选最佳区间的信息存储在存储器设备中;

由所述映射和比对单元使用所述延伸记录生成第一延伸种子,所述第一延伸种子是所述第一种子的延伸;

由所述映射和比对单元生成包括所述第一延伸种子的后续散列查询;并且

由所述映射和比对单元执行对所述散列表的后续散列查询。

17.如权利要求16所述的计算机可读介质,所述操作还包括:

由所述映射和比对单元确定对所执行的查询的所述响应包括一个或多个匹配的参考序列位置;并且

基于由所述映射和比对单元确定对所执行的查询的所述响应包括一个或多个匹配的参考序列位置:由所述映射和比对单元将所述一个或多个匹配的参考序列位置添加到种子匹配集合。

18.如权利要求16所述的计算机可读介质,其中,由所述映射和比对单元确定是否要将描述所述区间记录的所述第一信息作为描述候选最佳区间的信息存储在存储器设备中包括:

由所述映射和比对单元确定不存在将区间记录描述为用于所述特定读段的候选最佳区间的先前信息;并且

由所述映射和比对单元将描述所述区间记录的所述第一信息作为描述候选最佳区间的信息存储在所述存储器设备中。

19.如权利要求16所述的计算机可读介质,所述操作还包括:

由所述映射和比对单元获得对所述后续执行的查询的响应,所述响应包括由所述散列表的被确定为响应于所述查询的位置存储的信息;

由所述映射和比对单元确定对所述后续执行的查询的所述响应是否包括:(i)第二延伸记录,(ii)第二区间记录,或(iii)一个或多个匹配的参考序列位置;

基于由所述映射和比对单元确定对所述后续执行的查询的所述响应包括(i)所述第二延伸记录和(ii)所述第二区间记录:

由所述映射和比对单元确定是否要存取延伸表以获得所述延伸表中的由所述第二区间记录引用的一个或多个匹配的参考序列位置;

基于确定不存取所述延伸表:使用一个或多个启发式规则并且由所述映射和比对单元确定描述所述第二区间记录的第二信息或描述所述候选最佳区间的所述第一信息是否将用作所述候选最佳区间;由所述映射和比对单元使用所述第二延伸记录生成第二延伸种子,所述第二延伸种子是所述第一延伸种子的延伸;由所述映射和比对单元生成包括所述第二延伸种子的第三散列查询;并且由所述映射和比对单元执行对所述散列表的包括所述第二延伸种子的第三查询。

20.如权利要求19所述的计算机可读介质,其中,使用一个或多个启发式规则并且由所述映射和比对单元确定描述所述第二区间记录的所述第二信息或描述所述候选最佳区间的所述第一信息是否将用作所述最佳区间包括:

基于多个因素来选择描述所述第二区间记录的所述第二信息或描述所述候选最佳区间记录的所述第一信息,所述多个因素包括:(i)由所述区间记录和所述第二区间记录中的每一者返回的匹配的参考序列位置的数量,(ii)参考序列位置的预先确定的阈值级别,或(iii)到达存储所述区间记录和所述第二区间记录的所述散列位置的相应种子的每个种子长度。

21.一种使用散列表来改进样品读段到参考序列的映射的集成电路,所述集成电路包括多个硬件逻辑门,所述多个硬件逻辑门被物理配置成一个或多个硬件数字逻辑电路,所述硬件数字逻辑电路实现映射和比对单元的功能,所述集成电路包括:

执行对散列表的查询的一个或多个硬件逻辑电路,所述查询包括第一种子,其中所述第一种子包括从所述样品读段中的特定读段获得的核苷酸的子集;

获得对所执行的查询的响应的一个或多个硬件逻辑电路,所述响应包括由所述散列表的被确定为响应于所述查询的位置存储的信息;

确定对所执行的查询的所述响应是否包括(i)延伸记录和(ii)区间记录的一个或多个硬件逻辑电路,其中所述区间记录识别存储在延伸表中的与所述第一种子匹配的参考序列

位置的连续集合;

基于确定对所执行的查询的所述响应包括(i)延伸记录和(ii)区间记录,确定是否要存取所述延伸表以获得所述延伸表中的由所述区间记录引用的一个或多个匹配的参考序列位置的一个或多个硬件逻辑电路;以及

基于确定要存取所述延伸表执行下述操作的一个或多个硬件逻辑电路:存取所述延伸表以获得所述延伸表中的由所述区间记录引用的所述一个或多个匹配的参考序列位置;并且将所述一个或多个匹配的参考序列位置添加到种子匹配集合。

22.如权利要求21所述的集成电路,其中所述集成电路为现场可编程门阵列(FPGA)。

用于散列表基因组映射的灵活种子延伸

[0001] 本申请是名称为“用于散列表基因组映射的灵活种子延伸”、申请号为202080003658.1的发明专利申请的分案申请。

[0002] 相关专利申请的交叉引用

[0003] 本申请要求2019年5月24日提交的美国临时专利申请62/852,965的权益,该临时专利申请全文以引用方式并入本文。

背景技术

[0004] 核酸测序仪是被配置为使核酸测序过程自动化的仪器。核酸测序是确定核酸序列中核苷酸的顺序的过程。核酸可包括脱氧核糖核酸(DNA)或核糖核酸(RNA)。

[0005] 核酸测序仪被配置为接收核酸样品并生成表示核酸样品中的核苷酸的顺序的输出数据,该输出数据被称为一个或多个“读段”(read)。DNA样品中的核苷酸可包括一种或多种核苷酸碱基,这些核苷酸碱基包括形成任何组合的鸟嘌呤(G)、胞嘧啶(C)、腺嘌呤(A)和胸腺嘧啶(T)。RNA样品中的核苷酸可包括一种或多种碱基,这些碱基包括形成任何组合的G、C、A和尿嘧啶(U)。

[0006] 可使用映射和比对引擎来将由DNA测序仪生成的读段映射到参考基因组的已知核苷酸序列。将读段映射到参考基因组的核苷酸序列可由使用散列表索引的映射和比对引擎来实现。

发明内容

[0007] 本公开描述了散列表索引的构建和使用,该散列表索引有利于灵活的种子延伸以改进基因组映射和比对系统的性能。具体地讲,本公开可用于以如下方式执行灵活的种子延伸:(i)减少计算资源和功率的消耗,以及(ii)解决本文所述的常规种子延伸方法的问题。为了实现这些优点,本公开特别提供了可存储在散列表位置中的“区间记录”。

[0008] 本公开的各方面使得映射和比对单元能够单独使用区间记录,或结合一个或多个延伸记录使用区间记录,以减少要由所述映射和比对单元通过种子延伸处理的匹配位置的数量,同时还在确定使用动态种子延伸识别的匹配参考位置是否准确时,或者在一些情况下,确定是否应发生使用一个或多个延伸记录的种子延伸时,提供所述映射和比对单元的灵活性。这产生了一种映射和比对单元,它比利用常规种子延伸技术的其他映射和比对单元使用更少的功率和更少的计算资源,同时还更加准确。

[0009] 在一个方面,本公开提供了生成用于将样品读段映射到参考的散列表的方法。在一个方面,所述方法可包括以下动作:由计算机系统从参考序列获得第一核苷酸种子,其中所述第一种子具有K个核苷酸的长度;由所述计算机系统确定所述第一种子与超过预先确定数量的参考序列位置匹配;基于确定所述第一种子与超过预先确定数量的参考序列位置匹配,由所述计算机系统生成具有多个节点的种子延伸树,其中所述多个节点中的每个节点对应于(i)延伸种子,所述延伸种子是所述第一种子的延伸并且具有K*的核苷酸长度,其中K*为大于K的一个或多个核苷酸,和(ii)种子延伸表中的一个或多个位置,所述一个或多

个位置包括描述与所述延伸种子匹配的参考序列位置的数据,并且对于所述多个节点中的每个节点:由所述计算机系统在所述散列表的对应于所述延伸种子的索引键的位置存储区间信息,其中所述区间信息引用所述种子延伸表中的一个或多个位置,所述一个或多个位置包括描述与和所述节点相关联的所述延伸种子匹配的参考序列位置的数据。

[0010] 其他方面包括执行如本文所公开的方法的动作的对应系统、装置和计算机程序,如由在计算机可读存储设备上编码的指令所定义的。

[0011] 这些和其他方面可任选地包括以下特征中的一个或多个特征。例如,在一些具体实施中,每个所述匹配的参考序列位置包括所述第一种子的K个核苷酸。

[0012] 在一些具体实施中,所述方法还可包括:由所述计算机系统从所述参考序列获得与所述第一种子不同的第二核苷酸种子;由所述计算机系统确定所述第二种子不与超过所述预先确定数量的参考序列位置匹配;以及基于由所述计算机系统确定所述第二种子不与超过所述预先确定数量的参考序列位置匹配:由所述计算机系统获得描述与所述第二种子匹配的每个所述参考序列位置的数据,以及由所述计算机系统在所述散列表的对应于所述第二种子的索引键的第二位置存储描述与所述第二种子匹配的所述参考序列位置的所述数据。

[0013] 在一些具体实施中,所述种子延伸表中的包括描述与所述延伸种子匹配的参考序列位置的数据的所述一个或多个位置可包括所述种子延伸表中的包括描述与所述延伸种子匹配的参考序列位置的数据的位置的连续区间。

[0014] 在一些具体实施中,所述种子延伸表中的包括描述与和所述节点相关联的所述延伸种子匹配的参考序列位置的数据的所述一个或多个位置可包括与和所述节点相关联的所述延伸种子匹配的参考序列位置的延伸表中的连续区间。

[0015] 在一些具体实施中,由计算机系统从参考序列获得第一核苷酸种子,所述第一种子表示具有K个核苷酸的核苷酸长度的核苷酸序列,所述获得可包括:由所述计算机系统确定参考序列内的种子存取窗口的位置;以及由所述计算机系统获得所述参考序列的由所述种子存取窗口识别的子集。

[0016] 在一些具体实施中,所述方法还可包括:由所述计算机系统将所述种子延伸窗口沿着所述参考序列向前调节K个核苷酸,以从所述参考序列识别具有K个核苷酸的核苷酸长度的第二核苷酸种子;由所述计算机系统从所述参考序列获得所述第二种子;由所述计算机系统确定所述第二种子与超过预先确定数量的参考序列位置匹配;基于确定所述第二种子与超过预先确定数量的参考序列位置匹配,由所述计算机系统生成具有多个第二节点的第二种子延伸树,其中所述多个第二节点中的每个第二节点对应于(i)第二延伸种子,所述第二延伸种子是所述第二种子的延伸并且具有K*的核苷酸长度,其中K*是大于K的一个或多个核苷酸,和(ii)第二种子延伸表中的第二一个或多个位置,所述第二一个或多个位置包括描述与所述第二延伸种子匹配的参考序列位置的数据,并且对于所述多个第二节点中的每个第二节点:由所述计算机系统在所述散列表的对应于所述第二延伸种子的索引键的位置存储第二区间信息,其中所述第二区间信息引用所述第二种子延伸表中的一个或多个位置,所述一个或多个位置包括描述与和所述第二节点相关联的所述第二延伸种子匹配的参考序列位置的数据。

[0017] 在一些具体实施中,所述方法还可包括:对于所述多个节点中的每个节点:由所述

计算机系统确定所述种子延伸树的所述节点是否为叶节点,并且基于由所述计算机系统确定所述延伸树的所述节点不是叶节点,由所述计算机系统在所述散列表的对应于所述延伸种子的所述索引键的所述位置存储延伸记录。

[0018] 在一些具体实施中,所述延伸记录包括一个或多个指令,所述一个或多个指令在由所述计算机系统执行时,使得所述计算机系统将一个或多个附加的核苷酸添加到与所述延伸记录相关联的种子。

[0019] 在一些具体实施中,所述方法还可包括:基于由所述计算机系统确定所述延伸树的所述节点是叶节点,由所述计算机系统确定不在所述散列表的对应于所述延伸种子的所述索引键的所述位置存储延伸记录。

[0020] 在一些具体实施中,所述方法还可包括由所述计算机系统生成所述种子延伸表。在此类具体实施中,生成所述种子延伸表可包括:由所述计算机系统识别与所述第一种子匹配的所述参考序列的每个种子,以及由所述计算机系统将识别所识别的种子的数据存储在所述种子延伸表中。

[0021] 在一些具体实施中,所述方法还可包括由所述计算机系统对所述种子延伸表中的所识别的种子进行排序。

[0022] 在一些具体实施中,所述方法还可包括:由所述计算机系统生成散列表安装包,所述散列表安装包包括指令,所述指令在由接收所述散列表安装包的一个或多个计算机处理时,使得所述一个或多个计算机在能够由可编程逻辑电路访问的存储器中安装所述散列表。

[0023] 在一些具体实施中,所述散列表安装包可包括所述种子延伸表,并且所述散列表安装包可包括用于指示(i)所述可编程逻辑电路或(ii)另一个计算机将所述种子延伸表存储在能够由所述可编程逻辑电路访问的存储器设备中的指令。

[0024] 在一些具体实施中,由所述计算机系统向另一个计算机提供所述散列表安装包。

[0025] 在一些具体实施中,所述另一个计算机可包括(i)被配置为与所述可编程逻辑电路通信的计算机或(ii)所述可编程逻辑电路本身。

[0026] 在一些具体实施中,所述计算机系统可包括多个计算机。

[0027] 在另一方面,本公开提供了使用散列表来改进样品读段到参考序列的映射的方法。在一个方面,所述方法可包括:由映射和比对单元执行对散列表的查询,所述查询包括第一种子,其中所述第一种子包括从所述样品读段中的特定读段获得的核苷酸的子集;由所述映射和比对单元获得对所执行的查询的响应,所述响应包括由所述散列表的被确定为响应于所述查询的位置所存储的信息;由所述映射和比对单元确定对所执行的查询的所述响应是否包括(i)延伸记录,(ii)区间记录,或(iii)一个或多个匹配的参考序列位置;基于由所述映射和比对单元确定对所执行的查询的所述响应包括(i)延伸记录和(ii)区间记录:由所述映射和比对单元确定是否要存取延伸表以获得所述延伸表中的由所述区间记录引用的一个或多个匹配的参考序列位置;基于确定不存取所述延伸表:由所述映射和比对单元确定是否要将描述所述区间记录的所述第一信息作为描述候选最佳区间的信息存储在存储器设备中;由所述映射和比对单元使用所述延伸记录生成第一延伸种子,所述第一延伸种子是所述第一种子的延伸;由所述映射和比对单元生成包括所述第一延伸种子的后续散列查询;以及由所述映射和比对单元执行对所述散列表的所述后续散列查询。

[0028] 其他型式包括执行由在计算机可读存储设备上编码的指令所定义的方法的动作的对应系统、装置和计算机程序。

[0029] 本公开的这些和其他方面可任选地包括以下特征中的一个或多个特征。例如,在一些具体实施中,所述方法还可包括:基于确定要存取所述延伸表:由所述映射和比对单元存取所述延伸表以获得所述延伸表中的由所述区间记录引用的所述一个或多个匹配的参考序列位置;以及由所述映射和比对单元将所述一个或多个匹配的参考序列位置添加到种子匹配集合。

[0030] 在一些具体实施中,所述方法还可包括:由所述映射和比对单元确定对所执行的查询的所述响应包括一个或多个匹配的参考序列位置;以及基于由所述映射和比对单元确定对所执行的查询的所述响应包括一个或多个匹配的参考序列位置:由所述映射和比对单元将所述一个或多个匹配的参考序列位置添加到种子匹配集合。

[0031] 在一些具体实施中,由所述映射和比对单元确定是否要将描述所述区间记录的所述第一信息作为描述候选最佳区间的信息存储在存储器设备中可包括:由所述映射和比对单元确定不存在将区间记录描述为用于所述特定读段的候选最佳区间的先前信息;以及由所述映射和比对单元将描述所述区间记录的所述第一信息作为描述候选最佳区间的信息存储在所述存储器设备中。

[0032] 在一些具体实施中,所述方法还可包括:由所述映射和比对单元获得对所述后续执行的查询的响应,所述响应包括由所述散列表的被确定为响应于所述查询的位置存储的信息;由所述映射和比对单元确定对所述后续执行的查询的所述响应是否包括(i)第二延伸记录,(ii)第二区间记录,或(iii)一个或多个匹配的参考序列位置;基于由所述映射和比对单元确定对所述后续执行的查询的所述响应包括(i)所述第二延伸记录或(ii)所述第二区间记录:由所述映射和比对单元确定是否要存取延伸表以获得所述延伸表中的由所述第二区间记录引用的一个或多个匹配的参考序列位置;基于确定不存取所述延伸表:使用一个或多个启发式规则并且由所述映射和比对单元确定描述所述第二区间记录的所述第二信息或描述所述候选最佳区间的所述第一信息是否将用作所述候选最佳区间;由所述映射和比对单元使用所述第二延伸记录生成第二延伸种子,所述第二延伸种子是所述第一延伸种子的延伸;由所述映射和比对单元生成包括所述第二延伸种子的第三散列查询;以及由所述映射和比对单元执行对所述散列表的包括所述第二延伸种子的所述第三查询。

[0033] 在一些具体实施中,使用一个或多个启发式规则并且由所述映射和比对单元确定描述所述第二区间记录的所述第二信息或描述所述候选最佳区间的所述第一信息是否将用作所述最佳区间可包括:基于多个因素来选择描述所述第二区间记录的所述第二信息或描述所述候选最佳区间记录的所述第一信息,所述多个因素包括(i)由所述区间记录和所述第二区间记录中的每一者返回的匹配的参考序列位置的数量,(ii)参考序列位置的预先确定的阈值级别,或(iii)到达存储所述区间记录和所述第二区间记录的所述散列位置的相应种子的每个种子长度。

[0034] 在一些具体实施中,所述区间记录引用所述种子延伸表中的多个位置,所述多个位置包括描述与所述查询的所述第一种子匹配的参考序列位置的数据。

[0035] 在一些具体实施中,所述种子延伸表中的包括描述与所述查询的所述第一种子匹配的参考序列位置的数据的所述多个位置可包括延伸表中的与所述查询的所述第一种子

匹配的参考序列位置的连续区间。

[0036] 除非另有定义,否则本文所用的所有技术和科学术语的含义与本发明所属领域的普通技术人员通常理解的含义相同。虽然与本文所述的方法和材料类似或等同的方法和材料也可用于本发明的实践或测试,但合适的方法和材料如下所述。本文提及的所有出版物、专利申请、专利和其他参考文献均全文以引用方式并入本文。如发生矛盾,以本说明书及其所包括的定义为准。此外,所述材料、方法和示例仅为例示性的,并非旨在进行限制。

[0037] 下面参考附图和权利要求书在详细描述中更详细地讨论本公开的这些和其他方面。

附图说明

[0038] 图1是用于生成散列表索引的系统的关系图,该散列表索引有利于散列表基因组映射的灵活种子延伸。

[0039] 图2是用于生成散列表索引的过程的流程图,该散列表索引有利于散列表基因组映射的灵活种子延伸。

[0040] 图3是用于执行散列表基因组映射的运行时灵活种子延伸的运行时系统的关系图。

[0041] 图4是用于执行散列表基因组映射的运行时灵活种子延伸的过程的流程图。

[0042] 图5是用于对读段的每个种子执行散列表基因组映射的迭代运行时灵活种子延伸的过程的流程图。

[0043] 图6是可用于实现本文所述的与散列表基因组映射的灵活种子延伸相关的系统的系统部件的示意图。

[0044] 图7是显示数据的条形图的图示,这些数据以使用如本文所述的灵活种子延伸方法的系统中未映射读段百分比的形式来表示测试结果,并与不使用灵活种子延伸方法的系统的测试结果进行比较。

[0045] 图8是显示数据的折线图的图示,这些数据以使用如本文所公开的灵活种子延伸方法的系统中读段映射准确率的形式来表示测试结果,并与不使用灵活种子延伸方法的系统的测试结果进行比较。

具体实施方式

[0046] 本公开描述了散列表索引的构建和使用,该散列表索引有利于灵活的种子延伸以改进基因组映射和比对系统的性能。如本文所用,术语“种子”是指存在于核酸读段(“读段”)或核酸参考序列(“参考序列”)中的有序核苷酸的子集。以举例的方式,读段的短种子可具有例如21个碱基或核苷酸,这些碱基或核苷酸从例如由核酸测序仪(“测序仪”)基于输入到测序仪中的生物样品而生成的150个碱基或核苷酸的读段提取而来。此类短种子可与参考序列的数百、数千、数十万或甚至更多位置匹配。参考序列的种子可包括来自参考序列的有序核苷酸的子集,该参考序列表示参考序列位置。对与读段的特定短种子匹配的如此大量的参考序列位置的识别可出于多种原因而发生,包括重复序列诸如“...ATATAT...”的出现,这种重复可出现在参考序列中的许多位置。另选地或除此之外,可出现如此大量的匹配参考序列位置,因为基因组序列的许多近拷贝可出现在参考序列中。

[0047] 这些与特定短种子匹配的大量参考序列位置可引起使用常规散列表索引的常规映射和比对单元上的资源紧张,因为映射和比对引擎可能被迫处理大量匹配项。对与特定短种子匹配的大量参考序列位置的此类过量处理导致不必要的计算资源消耗,包括占用大量处理资源、存储器资源,以及浪费用于为处理资源、存储器设备和冷却单元供电的功率,冷却单元用于冷却处理资源和存储器资源,或它们的任何组合。

[0048] 已采用常规方法来解决可因识别和处理与短种子匹配的大量参考序列位置而产生的问题。例如,采用了使用存储在散列表位置中的延伸记录来迭代地延伸短种子的常规方法。此类方法包括例如美国专利10,083,276所述的方法,该专利以引用方式并入本文,该专利的方法可返回“延伸记录”,该延伸记录存储在散列表的对应于散列查询的种子的位置。延伸记录可用于通过在接收的散列查询中对称地增加种子的长度来创建延伸种子,种子长度的增加是通过在种子的每一端添加一个或多个碱基或核苷酸来实现的。常规系统然后可使用包括延伸种子的另一个散列查询来再次查询散列表。具有延伸种子的该另一个散列查询可能对应于识别与延伸种子匹配的较少参考位置的散列位置,因为延伸种子较长。该迭代过程可以继续,直到(i)所得的匹配集合充分收缩到包括小于阈值数量的与延伸种子匹配的参考序列位置,(ii)匹配集合变为空,(iii)达到最大种子延伸,或(iv)无法进行下一次延伸,因为该延伸将移动超出作为该短种子的基础的读段的边缘。关键的是,在常规系统中,只有当该迭代过程以上述方式(i)终止时,映射和比对单元才能够获得匹配的参考位置的非空集合,而当该迭代过程以上述方式(ii)、(iii)或(iv)中的任何一者终止时则不能够。

[0049] 这些常规方法可有助于减少与短种子匹配的参考序列位置的数量。然而,这些常规方法存在三个明显的问题。

[0050] 首先,常规方法可能易受“未映射读段问题”的影响。当常规种子延伸方法针对延伸种子返回零匹配项时,会发生未映射读段问题。如果延伸种子结合了变异诸如SNP,或者如果延伸种子超出该延伸种子所对应的读段的边缘,则会发生此类零匹配结果集合。如果在使用常规方法时针对读段的每个种子位置都发生了这种情况,则该读段可能未映射。

[0051] 第二,常规方法可能易受“高置信度误映射”问题的影响。当延伸种子包括变异诸如SNP但仍与一个或多个参考位置匹配时,发生此类高置信度误映射问题。此类映射的特征可在于高置信度分数,诸如高MAPQ分数,即使延伸种子被错误地映射也如此。如果在使用常规方法时针对读段的每个种子位置都发生了这种情况,则该读段可能误映射,有可能具有高置信度。对于此类映射,可能会丢失相反的证据。与低置信度误映射相比,高置信度误映射可对映射器总体准确率造成更大损害。MAPQ分数可包括对已映射的读段被误置的概率进行量化的质量分数。

[0052] 第三,常规方法可能易受“固定最大匹配数问题”的影响。一般来讲,为种子延伸所构建的散列表使用最大匹配参数M,诸如M=16。该参数确保种子延伸树中没有叶节点超过参数M。然而,一些应用程序可受益于使用不同的最大匹配参数M*,诸如M=64。常规的种子延伸方法将继续迭代地延伸种子,直到到达叶节点。因此,除非重建散列表以使得最大匹配参数M被设定为64,否则当达到M=64的匹配集合时,使用常规方法的应用程序不能停止种子的延伸。

[0053] 本公开的创新方面可用于以如下方式执行灵活的种子延伸:(i)减少如本文所述

的计算资源和功率的消耗,以及(ii)解决常规种子延伸方法的问题,诸如上述那些问题。为了实现这些优点,本公开特别提供了可存储在散列表位置中的“区间记录”。对于特定种子,区间记录识别存储在种子延伸表中的与该特定种子匹配的参考序列位置的连续集合。在执行识别特定种子的散列查询时,映射和比对单元可基于响应于该查询的散列位置的内容来确定是否(i)基于存储在该散列位置的种子延伸记录来延伸种子,(ii)将识别与该特定种子匹配的参考位置的区间记录存储在种子延伸表中,或(iii)存取种子延伸表中的由存储在该散列位置的区间记录识别的参考序列位置。在一些具体实施中,可执行这些操作的组合,诸如延伸种子并存储区间。

[0054] 因此,将区间记录与一个或多个延伸记录结合使用使得映射器和比对器能够减少将由映射和比对单元通过种子延伸处理的匹配位置的数量,同时还在确定使用动态种子延伸识别的匹配参考位置是否准确时,或者在一些情况下,确定是否应发生使用一个或多个延伸记录的种子延伸时,提供映射和比对单元的灵活性。这产生了一种映射和比对单元,它比利用常规种子延伸技术的其他映射和比对单元使用更少的功率和更少的计算资源,同时还更加准确。

[0055] 灵活种子延伸的散列表索引的生成

[0056] 图1是用于生成散列表索引的系统100的关系图,该散列表索引有利于散列表基因组映射的灵活种子延伸。系统100包括计算机110、存储器112和存储器130。虽然存储器112和130在图1中被描绘为独立的存储器设备,但本公开不需要限于此。相反,在一些具体实施中,存储器112和存储器130可为相同的存储器设备。例如,存储器112和存储器130简单地指单个存储器设备上的两个单独的存储位置。另选地,存储器112和存储器130可各自分别存储在单独的存储器设备中,诸如计算机110可访问的单独的硬盘。又如,存储器112可以是存储参考基因组库的基于云的服务器的存储器设备,并且存储器130可以是计算机110的本地存储器。因此,在图1中将存储器112和存储器130描绘为单独的存储器对存储器112、130本身或这些存储器的内容没有限制,并且不需要这些存储器必须利用本公开的任何特定具体实施来被组织或存储。

[0057] 计算机110可包括一个或多个计算机,每个计算机包括一个或多个处理单元,这些处理单元被配置为通过执行一个或多个软件指令来执行操作。该一个或多个处理单元可包括一个或多个中央处理单元(CPU)、一个或多个图形处理单元(GPU)或它们的任何组合。计算机110可被配置成经由直接连接诸如一条或多条总线、一条或多条USB电缆、一条或多条USB-C电缆等或它们的任何组合来直接与存储器112、存储器130或可编程电路162交互。另选地或除此之外,计算机110可被配置为经由一个或多个网络与存储器112、存储器130或可编程电路162交互。该一个或多个网络可包括有线以太网网络、无线网络、光网络、LAN、WAN、蜂窝网络、互联网或它们的任何组合。

[0058] 以举例的方式,一个具体实施可包括计算机110,该计算机被配置为(i)与存储在计算机110可访问的一个或多个本地存储器设备中的存储器112和存储器130交互以生成种子延伸表132和散列表140,以及(ii)使用一个或多个网络将所生成的种子延伸表132和散列表140传送到与另一个设备160集成的可编程电路162。该另一个设备160可包括核酸测序仪、基于云的服务器或任何其他计算机。在一些具体实施中,可编程电路162可使用扩展卡诸如PCI卡与该另一个设备集成。在此类具体实施中,可编程电路162可被容纳在PCI卡的逻

辑板上,该PCI卡使用测序仪、基于云的服务器或该另一个计算机的主板上的PCI端口插入到该主板中。

[0059] 可编程电路162可包括一个或多个可编程集成电路,诸如一个或多个现场可编程门阵列(FPGA)。现场可编程门阵列是一类集成电路,它们包括能够被动态配置为实现一个或多个处理模块,诸如基因组分析流水线的基因组分析模块(诸如映射和比对单元170),或处理模块(诸如散列表140)的一部分的多个硬件数字逻辑门、硬件数字逻辑电路等。FPGA可使用硬件描述语言(HDL)诸如超高速集成电路硬件描述语言(VHDL)、Verilog等来编程。FPGA是灵活的,因为先前已被编程为包括基因组分析流水线的一个或多个基因组分析模块或其部分的FPGA能够被动态地重新配置为包括对一个或多个基因组分析模块、其他不同基因组分析模块等的更新。

[0060] 可使用其他类型的集成电路来代替或辅助可编程电路162以实现本文所述的可编程电路162的功能。例如,一个或多个专用集成电路(ASIC)可用于实现可编程电路162的功能或功能的一部分。ASIC是包括在制造时配置的多个硬件数字逻辑门、多个数字逻辑电路等的定制集成电路。ASIC与本文所述FPGA的类似之处在于ASIC的硬件数字逻辑门或多个数字逻辑电路可使用硬件描述语言诸如VHDL、Verilog等来描述或设计。然后可将ASIC制造或印刷成包括HDL所描述的数字逻辑或数字电路。然而,一旦制造或印刷,ASIC就不能像FPGA那样动态地重新配置。虽然本文所述的示例描述了可编程或定制电路,但本公开不需要限于此。在一些具体实施中,例如,可使用其他类型的集成电路来实现被描述为由可编程电路162执行的功能。

[0061] 存储器112可存储一个或多个参考序列114。参考序列可包括(i)代表物种的完整参考基因组,(ii)代表物种的参考基因组的一部分,或(iii)代表多个物种的完整和/或部分参考基因组。参考序列包括碱基或核苷酸的有序列表。构成参考序列的碱基或核苷酸的有序列表可在存储器112中作为数字核酸序列数据库来组织。特定参考序列可由人、计算机或两者从特定物种的多个不同供体组装而成,作为该物种的代表。

[0062] 在一些具体实施中,特定参考序列可被组装为代表特定族群,其中特定族群是物种的具有特定核酸序列的子集,所述特定核酸序列可唯一地将该特定族群与物种内的其他族群分开。物种可包括任何物种,包括人类、非人类哺乳动物、爬行动物、鱼、昆虫、植物、细菌、病毒等。参考序列可从未灭绝物种(诸如人类)的样品或当前已灭绝物种(诸如恐龙或猛犸族群)生成。可使用从包含在已灭绝物种诸如恐龙的化石、冰冻或以其他方式保存的遗骸内的生物材料获得的样品来组装已灭绝物种的参考序列。可从(i)对得自已灭绝物种的化石遗骸的生物遗骸进行测序和(ii)对得自未灭绝物种的生物样品进行测序的组合来组装得到已灭绝物种的参考序列。完整的参考基因组可包括许多有序碱基或核苷酸。例如,人类参考基因组可包括多达30亿个有序碱基或核苷酸。

[0063] 计算机110被配置为生成有利于灵活种子延伸的散列表140。计算机110通过存取存储在存储器112中的参考序列114以及获得参考序列的种子114-1、114-2、114-3至114-n来开始生成散列表140,其中n为大于零的任何非零整数。在一些具体实施中,计算机110可使用种子存取窗口来识别和获得参考序列114的种子114-1、114-2、114-3至114-n。计算机110可将种子存取窗口初始化为具有种子长度K,其中K为要包括在每个种子中的碱基或核苷酸的数量,其中K为大于零的任何非零整数。计算机110可通过如下方式开始存取参考序

列114的种子:将长度为K的种子存取窗口定位在参考序列的起点,以涵盖参考序列中的第一组K个核苷酸,诸如种子“GTTTA”114-1。在该示例中,K等于5,但K不限于这种核苷酸长度。相反,K可等于任何大于零的非零整数,并且在一些具体实施中可等于例如7、10、12、15、18、20、21、25或更多个碱基或核苷酸。种子114-1、114-2和114-3至114-n仅仅是参考序列114的种子的示例,并且在这个示例中不需要对应于参考序列114的一组4个序列种子。

[0064] 为了生成散列表140,计算机110被配置为存取参考序列114的每个种子114-1、114-2和114-3至114-n,并对每个种子114-1、114-2和114-3至114-n执行一组操作。该组操作被设计为生成用于存储到散列表140的对应于散列表140的索引键142的散列位置144的信息。每个索引键142可对应于参考序列114的多个种子114-1、114-2、114-3至114-n中的相应种子,种子114-1、114-2、114-3至114-n中的每个种子的反向互补序列,多个种子114-1、114-2、114-3至114-n中的一个或多个种子的延伸种子,或每个延伸种子中的反向互补序列。索引种子142中的每个种子可使用散列函数143映射到散列位置144。

[0065] 计算机110可采用以下方法来识别并存取多个种子114-1、114-2和114-3至114-n中的每个种子:通过在每个相应种子被存取并用于执行该组操作之后将种子存取窗口在参考序列114中推进K个位置。下文更详细地描述了对每个相应种子114-1、114-2和114-3至114-n执行的该组操作。该组操作可包括使用所生成的信息来填充散列表140。另选地,对散列表140的填充可在对每个种子的该组操作结束之后发生。

[0066] 计算机110对参考序列114的每个种子114-1、114-2和114-3至114-n执行的该组操作的第一步是计算机110获得由种子存取窗口识别的种子。在图1的示例中,假定参考序列114的由种子存取窗口识别的种子为“GTTTA”114-1。

[0067] 计算机110可确定所获得的种子“GTTTA”114-1是否与超过预先确定数量的参考序列114位置匹配。匹配的参考序列位置可包括参考序列114的包含种子114-1的子集。参考序列114的该子集可包括一组有序排序的核苷酸,该组核苷酸大于或等于所获得的种子中的K个核苷酸。在一些具体实施中,预先确定数量的匹配的参考序列位置可包括一个匹配的参考序列位置。然而,在其他具体实施中,可将该预先确定数量设置为多于一个匹配的参考序列位置。

[0068] 如果计算机110确定种子114-1与小于或等于预先确定数量的参考序列位置匹配,则计算机可将与种子114-1匹配的参考位置填充到种子“GTTTA”114-1所到达的散列位置144。如果种子诸如种子114-1与散列函数143映射到散列位置的散列键142匹配,则散列位置144可被种子114-1“到达”。另选地,如果计算机110确定预先确定数量的匹配参考序列位置大于预先确定数量的参考序列位置,则计算机110可为种子114-1生成种子延伸树。在图1的示例中,计算机110确定种子“GTTTA”114-1与超过预先确定数量的参考序列位置匹配。因此,计算机110为种子114-1生成种子延伸树120。

[0069] 计算机110可从根节点120开始、逐个节点地为种子114-1生成种子延伸树120。可生成种子延伸树120,使得由叶节点识别的一组匹配的参考位置不超过预先确定的匹配限值,除非不可能进行进一步的种子延伸。种子延伸树120的每个节点可包括种子和种子延伸表132中的连续地址的区间。在一些具体实施中,种子延伸表132包括与种子(诸如,由计算机110使用种子存取窗口获得的种子114-1)匹配的131-1至131-6参考序列位置的中心字典序有序列表。中心字典排序可包括例如建立符号位置的优先级顺序,然后从中心符号开始

向左和向右交替。另选地,中心字典排序可包括例如建立符号位置的优先级顺序,然后从中心符号开始向右和向左交替。不过,可使用其他变型。

[0070] 在图1的示例中,基于种子114-1“GTTTA”对种子延伸表132进行中心字典排序133。该示例假设左侧开始顺序和正常字母表核苷酸顺序(即,A、C、G、T)以实现图1所示的中心字典排序顺序。计算机110可为被确定为具有超过预先确定的阈值数量的匹配参考序列位置的每个种子114-1、114-2、114-3、114-n生成种子延伸表,诸如种子延伸表132。在一些具体实施中,对于每个合格的种子,可在计算机110使用种子存取窗口存取特定种子之后并且在为该种子生成种子延伸树120之前,为该特定种子生成种子延伸表132。

[0071] 上文对种子延伸树120的节点的描述指示每个节点的地址的区间是连续的。然而,本公开不需要限于此。相反,节点的地址的区间可以是非连续的。例如,特定具体实施可使用区间来描述种子延伸表的一个或多个连续位置的多个不同集合,或存储在一个或多个存储器设备中的其他数据结构,其中一个或多个连续位置的每个连续集合相对于彼此不连续。也就是说,在每个相应集合之间可存在连续性的中断。

[0072] 每个合格种子的种子延伸表可存储在存储器130中。这可产生n个种子延伸表,即,对于参考序列114的n个种子中的每个都产生一个种子延伸表。另选地,种子延伸表的数量可小于n,诸如如果仅为具有超过预先确定的阈值数量的匹配参考序列位置的种子生成和存储种子延伸表的话。在生成每个种子延伸表之后,可将种子延伸表中的每个的集合132A提供给容纳可编程电路162的设备160,并存储在可编程电路162可访问的存储器180中。存储器180可包括DRAM存储器、SRAM存储器、NAND存储器等。在一些具体实施中,种子延伸表的集合132A可作为单独的种子延伸表提供给容纳可编程电路162的设备160。在其他具体实施中,种子延伸表的集合132A可作为单个主种子延伸表提供,该主种子延伸表由每个种子的相应种子延伸表中的每个的级联构成。种子延伸表的集合132A可以任何数量的格式提供。在一些具体实施中,种子延伸表的集合132A可由计算机110压缩以减小提供给设备160的种子延伸表文件的大小,然后由设备160、可编程电路162等解压缩以便存储在存储器180中。

[0073] 计算机110可生成种子延伸树120的根节点121以包括种子“GTTTA”121a和区间A 121b。区间A 121b识别种子延伸表132中存储与根节点121所表示的种子“GTTTA”121a匹配的参考序列位置的位置的连续区间。在该示例中,区间A识别种子延伸表132的跨越131-1至132-6并且包括“TAGTTTACT”、“TAGTTTATC”、“GAGTTTATG”、“ACGTTTAGT”、“TCGTTTAGT”和“ACGTTTAGC”的位置。计算机110可通过存取种子延伸表132以确定种子延伸表132的具有与节点诸如节点121的特定种子匹配的参考序列位置的位置,来为节点121的该种子确定一个或多个适当区间。

[0074] 在一些具体实施中,可使用种子延伸表132中的区间的起始位置地址和种子延伸表132中的区间的结束位置地址来描述用于节点诸如节点121的特定种子的区间121b。在其他具体实施中,可使用种子延伸表132中的区间的起始位置地址和与起始位置地址的偏移来描述用于节点诸如节点121的特定种子的区间121b。在此类具体实施中,以后可使用区间的起始地址和结束地址或者区间的起始地址和偏移来计算区间。然而,本公开不需要限于此。相反,应当理解,可使用以任何适当方式结构化或非结构化的任何形式的信息在散列表位置144中表示区间记录。例如,在一些具体实施中,可使用固定大小和格式的一个记录来实现区间记录。在其他具体实施中,可通过在不同大小的多种格式中进行选择来实现区间

记录,包括记录计数等,以优化散列表140所消耗的存储空间、启用散列表140的可压缩性、提高散列查询相对于其他区间记录格式的效率等。

[0075] 计算机110可通过延伸在根节点中识别的用于种子121a的碱基或核苷酸的数量来继续生成种子延伸树120。例如,计算机110可将根节点的种子长度从5个碱基或核苷酸延伸到7个碱基或核苷酸,并识别种子延伸表中具有7个匹配碱基或核苷酸的参考序列位置的最大子集。在图1的示例中,计算机110可确定具有7个匹配核苷酸的参考序列位置的最大子集是“CGTTTAG”。区间B识别种子延伸表132中存储与种子“CGTTTAG”匹配的参考序列位置的位置的连续区间。在该示例中,区间B识别种子延伸表132的跨越132-4至132-6并且包括“ACGTTTAGT”、“TCGTTTAGT”和“ACGTTTAGC”的位置。计算机110可使用利用种子延伸表132确定的信息来生成节点122。例如,计算机110可生成节点122以包括种子“CGTTTAG”122a和区间B 122b。

[0076] 参考图1的示例,计算机110可通过确定种子延伸表中是否存在具有7个匹配碱基或核苷酸的其他参考序列位置来继续生成种子延伸树120。如果在种子延伸表132中存在具有7个匹配碱基或核苷酸的其他参考序列位置,则计算机110使用具有7个匹配碱基或核苷酸的参考序列位置的第二大集合来生成种子延伸树的下一个节点。在图1的示例中,计算机110可确定具有7个匹配核苷酸的参考序列位置的第二大子集为“AGTTTAT”。区间B识别种子延伸表132中存储与种子“AGTTTAT”匹配的参考序列位置的位置的连续区间。在该示例中,区间E识别种子延伸表132的跨越132-2至132-3并且包括“TAGTTTATC”和“GAGTTTATG”的位置。计算机110可使用利用种子延伸表132确定的信息来生成节点123。例如,计算机110可生成节点123以包括种子“AGTTTAT”123a和区间E 123b。

[0077] 参考图1的示例,计算机110可通过确定种子延伸表中是否存在具有7个匹配碱基或核苷酸的其他参考序列位置来继续生成种子延伸树120。如果种子延伸表中的其他参考序列位置被识别为具有7个匹配碱基或核苷酸,则计算机110可使用具有7个匹配碱基或核苷酸的参考序列位置的第二大集合来生成种子延伸表120的新节点,如上所述。然而,在图1的该示例中,种子延伸表132中不存在具有7个匹配碱基或核苷酸的其他参考序列位置。因此,计算机110可确定将核苷酸的碱基数量从7延伸至9并继续分析种子延伸表132中的参考序列位置。

[0078] 参考图1的示例,计算机110可识别具有9个匹配核苷酸的参考序列位置的最大子集。在该示例中,存在具有9个匹配核苷酸的参考序列位置的多个子集。在此类情况下,计算机110可确定为具有9个匹配的参考序列核苷酸的参考序列位置的每个集合创建种子延伸树的节点。在一些具体实施中,计算机110可随机确定种子延伸树节点的创建顺序。在其他具体实施中,计算机110可基于后续延伸树节点的中心字典顺序来开始生成这些节点。

[0079] 无论它们的创建顺序如何,计算机110都可通过为具有9个匹配核苷酸的参考序列位置的每个子集生成种子延伸表的节点来继续。以举例的方式,计算机110可生成种子延伸树120的节点124以包括延伸的短种子“TCGTTTAGT”124a和区间C 124b。区间C 124b识别种子延伸表132中存储与短种子“TCGTTTAGT”124a匹配的参考序列位置的位置的连续区间。在该示例中,区间C识别种子延伸表132的跨越132-5并且包括“TCGTTTAGT”的位置。计算机110可通过存取种子延伸表132以确定种子延伸表132的具有与节点诸如节点124的特定短种子匹配的参考序列位置的位置,来为节点124的该短种子确定适当的区间。

[0080] 参考图1的示例,计算机110可通过为具有9个匹配核苷酸的参考序列位置的每个子集生成种子延伸表的节点来继续。以举例的方式,计算机110可生成种子延伸树120的节点125以包括延伸的短种子“ACGTTTAGC”125a和区间D 125b。区间D 125b识别种子延伸表132中存储与短种子“ACGTTTAGC”125a匹配的参考序列位置的位置的连续区间。在该示例中,区间D识别种子延伸表132的跨越132-6并且包括“ACGTTTAGC”的位置。计算机110可通过存取种子延伸表132以确定种子延伸表132的具有与节点诸如节点125的特定短种子匹配的参考序列位置的位置,来为节点125的该短种子确定适当的区间。

[0081] 参考图1的示例,计算机110可通过为具有9个匹配核苷酸的参考序列位置的每个子集生成种子延伸表的节点来继续。以举例的方式,计算机110可生成种子延伸树120的节点126以包括延伸的短种子“TAGTTTATC”126a和区间F 126b。区间F 126b识别种子延伸表132中存储与短种子“TAGTTTATC”126a匹配的参考序列位置的位置的连续区间。在该示例中,区间F识别种子延伸表132的跨越132-2并且包括“TAGTTTATC”的位置。计算机110可通过存取种子延伸表132以确定种子延伸表132的具有与节点诸如节点126的特定短种子匹配的参考序列位置的位置,来为节点126的该短种子确定适当的区间。

[0082] 本公开描述了以从匹配碱基的最大集合到匹配碱基的最小集合的特定有序顺序来构建种子延伸表的示例。然而,本公开不需要限于使用以这种方式构建的种子延伸树。相反,可使用用于构建种子延伸表的任何过程,只要该种子延伸表构建过程的结果产生种子延伸表即可。例如,可按照从匹配碱基的最小集合到匹配碱基的最大集合,或者根本不按特定顺序来生成种子延伸树。在一些具体实施中,系统100可生成并使用先前生成的种子延伸表,而不需要系统100构建种子延伸表。

[0083] 计算机110可使用所生成的种子延伸树120来填充散列表140的散列位置144,这些散列位置由对应于特定散列索引键142的特定种子输入到达。以举例的方式,计算机110可确定节点121是否为叶节点。基于确定节点121不是叶节点,计算机110可使用根节点121来填充散列位置144-y,其中y为任何非零整数。使用根节点121来填充散列位置144-y可包括将区间记录153b存储到种子121a所到达的散列表位置144-y中。区间记录153b识别节点121的区间121b。散列表140可包括每个种子114-1、114-2、114-3至114-n的散列表索引键142,每个种子114-1、114-2、114-3至114-n的反向互补序列,或它们的组合。可使用散列函数143将每个散列表索引键142映射到一个或多个散列位置144。每个散列位置144可使用一个或多个存储桶来实现,其中存储桶对应于存储器设备的一组一个或多个存储位置。存储器设备的该一个或多个存储位置中的每个可以是连续或非连续的存储器位置。

[0084] 图1的示例仅示出散列表140的一部分,该部分具有对应于种子121、122、123和125中的前向种子的键142。然而,本公开不需要限于此。例如,在一些具体实施中,可使用散列函数143来将种子散列,使得任何种子的反向互补核苷酸序列产生与原始前向种子相同的散列。核苷酸序列的反向互补序列可通过将原始核苷酸序列的顺序颠倒并且用T交换A、用A交换T、用G交换C以及用C交换G来确定。以举例的方式,用于原始前向种子GTTTA 121A的散列键142可具有与用于种子GTTTA的反向互补序列(即TAAAC)的散列键142相同的散列。在此类具体实施中,当匹配参考序列位置存储在散列位置144或作为种子延伸表132中的条目时,可使用例如反向互补(RC)标记来注释它们的序列取向。然而,在其他具体实施中,种子的反向互补序列可导致不同的散列,并且不需要在存储在散列表140的散列位置144或种子

延伸表132中的匹配参考序列位置中注释取向。

[0085] 填充散列表140的位置144还可包括确定是否要将延伸记录填充到散列位置144中。确定是否应将延伸记录填充到散列位置144中可包括确定种子延伸树120的用于填充散列位置的节点是否为叶节点。如果确定该节点是叶节点,则计算机110可确定不将延伸记录存储在与该节点相关联的种子所到达的散列位置。另选地,如果确定该节点不是叶节点,则计算机110可生成延伸记录并将所生成的延伸记录存储在散列表位置144。参考图1的示例,计算机110可确定或先前已经确定节点121不是叶节点。在此类情况下,计算机110可生成延伸记录153a并将其存储在种子121a所到达的散列表位置144-y。因此,散列位置144-y可包括延伸记录153a和区间记录153b。

[0086] 延伸记录可包括一个或多个指令,这些指令当由计算机诸如执行软件指令的中央处理单元(CPU)或图形处理单元(GPU)或可编程电路162执行时,可使得该CPU、GPU或可编程电路162延伸散列查询中使用的种子,该种子到达散列位置,该散列位置通过一个或多个核苷酸来存储延伸记录。在一些具体实施中,可生成延伸记录,使得延伸记录指示计算机在种子的每一端对称地延伸种子。因此,以举例的方式,可生成延伸记录以指示计算机诸如CPU、GPU或可编程电路162将种子延伸两个核苷酸、四个核苷酸、六个核苷酸等。在此类具体实施中,可通过使种子在种子的每个相应端上延伸一个核苷酸,在种子的每个相应端上延伸两个核苷酸,在种子的每个相应端上延伸三个核苷酸等来实现种子的对称延伸。在图1的示例中,延伸记录153a被配置为将初始种子121a对称地延伸2个碱基。计算机110可基于多种因素来确定用于包括在延伸记录中的延伸长度,所述因素包括(i)与种子匹配的参考序列位置的数量,(ii)所需的运行时种子延伸迭代数量,(iii)为每个迭代搜寻的匹配参考序列位置的数量等。下文中参考图3更详细地描述使用散列表140的运行时灵活种子延伸。

[0087] 核苷酸的种子通常已被描述为由有序核苷酸的连续集合构成。类似地,延伸记录被描述为以对称或不对称、但仍然连续的方式来将有序核苷酸的连续集合有序地延伸一个或多个附加的核苷酸。然而,本公开不限于使用有序核苷酸的连续集合。相反,读段或参考序列的种子可以是来自读段或参考序列的种子的非连续图案。类似地,延伸记录可包括指令,该指令在由CPU、GPU或可编程电路162处理时,使得该CPU、GPU或可编程电路162将初始种子延伸为结合不连续的附近碱基或核苷酸。在此类具体实施中,可以考虑使用不连续种子的方式在种子延伸表132中将每个根节点种子的匹配参考序列位置进行字典排序。

[0088] 计算机110可继续将信息填充到种子延伸树120的每个剩余节点122、123、124、125、126的散列位置144中。以举例的方式,计算机110可确定节点122是否为叶节点。基于确定节点122不是叶节点,计算机110可使用节点122来填充散列位置144-3。使用节点122来填充散列位置144-3可包括将区间记录152b存储到种子122a所到达的散列表位置144-3中。区间记录152b识别节点122的区间122b。计算机110可确定或先前已确定节点122不是叶节点,并且生成延伸记录152a以便存储在散列位置144-3。在图1的示例中,延伸记录152A包括将种子122a对称地延伸2个碱基或核苷酸的指令。如果例如响应于对种子122a的查询而未存取区间B,则可在运行时执行延伸记录152a的这些指令。

[0089] 然而,本公开不限于此,因为还生成其他延伸记录扫描,其指示CPU、GPU或可编程电路162将种子延伸不同的附加核苷酸长度(例如,2个、4个、6个、8个等)或以不同方式来延伸(例如,使用1、3、5等附加核苷酸长度非对称地延伸)。尽管图1的示例示出了散列位置

144-3中的单个延伸记录,但本公开不限于此。相反,在一些具体实施中,多个延伸记录可存储在单个散列位置144-3。例如,计算机110还可在散列位置144-3存储一个或多个附加的延伸记录,这些附加的延伸记录被配置为将初始种子122a延伸四个碱基。在此类具体实施中,CPU、GPU或可编程电路162在运行时可首先尝试将初始种子122A延伸四个碱基。如果此类种子延伸失败,因为在运行时对散列表140的后续查询没有产生匹配的参考序列位置,则CPU、GPU或可编程电路可获得另一个延伸记录152A,该延伸记录包括将初始碱基仅延伸2个碱基的指令。这可增大返回匹配的参考序列位置的可能性。

[0090] 计算机110可继续将信息填充到种子延伸树120的每个节点123、124、125、126的散列位置144中。以举例的方式,该计算机可确定节点123是否为叶节点。基于确定节点123不是叶节点,计算机110可使用节点123来填充散列位置144-1。填充散列位置144-1可包括将区间记录150B存储到种子123a所到达的散列表位置144-1中。区间记录150b识别节点123的区间123b。计算机110可确定或先前已确定节点123不是叶节点,并且生成延伸记录150a以便存储在散列位置144-1。在该示例中,延伸记录150a包括将种子123a对称地延伸2个碱基或核苷酸的指令。如果例如响应于对种子123a的查询而未存取区间E,则可在运行时执行延伸记录150a的这些指令。

[0091] 计算机110可继续将信息填充到种子延伸树120的每个节点124、125、126的散列位置144中。以举例的方式,计算机110可确定节点125是否为叶节点。基于确定节点125是叶节点,计算机110可确定通过将由区间D 125b识别的与种子“ACGTTTAGC”匹配的匹配参考序列位置155存储到散列位置144-2中来填充散列表140。另选地,在其他具体实施中,计算机110可确定将识别区间D 125b的区间记录存储到散列位置144-2中。在一些具体实施中,可由计算机110基于在用于叶节点的散列表位置144中存储每个匹配的参考序列位置是否是对存储器资源的最佳使用来作出此类确定。因此,如果确定在用于叶节点的散列表位置144中存储匹配的参考序列位置不满足对存储器资源的预先确定的阈值使用,则计算机110可将匹配的参考序列位置存储到种子延伸树的叶节点的种子所到达的散列位置中。否则,如果超过该存储器资源使用阈值,则计算机110可在种子延伸树的叶节点的种子所到达的散列位置144中存储区间记录。计算机110可确定或先前已确定节点125是叶节点,并且不生成用于存储在散列位置144-2的延伸记录。因此,在该示例中,在运行时不会发生种子“ACGTTTAGC”的进一步延伸。

[0092] 如上所述,散列位置144-2可仅存储与种子125a匹配并且对应于散列键142-1的匹配参考序列位置。这是因为,在该示例中,种子125a是不能延伸的叶节点125种子。然而,在没有延伸记录或区间记录中的一者或两者的情况下填充参考序列位置不限于叶节点的种子所到达的散列位置144。相反,在其他情况下,计算机110可确定在没有延伸记录或区间记录中的一者或两者的情况下利用匹配参考序列位置来填充散列位置144。例如,在一些具体实施中,如果计算机110确定用于特定种子的种子延伸表132仅识别小于阈值数量的匹配参考序列位置的匹配参考序列位置的区间,则计算机110可在没有延伸记录或区间记录中的一者或两者的情况下利用匹配参考序列位置来填充该特定种子所到达的散列位置144。

[0093] 其他类型的信息可存储在散列表140的散列位置144中。例如,计算机110可接收将一个或多个“停止”(STOP)记录插入散列表140的散列位置144中的指令。此类“停止”散列记录可使得存储区间记录或一组一个或多个匹配参考位置的特定散列位置140返回(i)区间

记录或(ii)该组一个或多个匹配参考位置,而无需将用于到达该散列位置的种子进一步延伸。在其他具体实施中,计算机110可接收将“停止”记录插入已包括延伸记录的散列位置中的指令。在此类具体实施中,当CPU、GPU或可编程电路162遇到“停止”记录时,该CPU、GPU或可编程电路162可有条件地确定是(i)忽略延伸记录并返回(i)区间记录或(ii)与用于到达具有“停止”记录的散列位置的种子匹配的一组一个或多个匹配参考位置,还是(ii)执行延伸记录所描述的种子延伸。在一些具体实施中,可基于一个或多个因素进行有条件确定,所述因素诸如(i)区间记录或(ii)该组一个或多个匹配参考序列位置所识别的匹配的参考序列的数量。因此,响应于相应输入种子将一个或多个“停止”记录插入特定散列位置的插入可用作设计工具来避免固定最大错配问题,而无需重建散列表诸如散列表140。

[0094] 计算机110可继续将信息迭代地填充到种子延伸树120的每个剩余节点诸如节点124、126的散列位置144中。用于这些叶节点中的每个的条目可以上文相对于节点125所述的方式填充,因为节点124、126和节点125一样是叶节点。

[0095] 此外,计算机110可继续将上文在图1的示例中参考种子“GTTTA”114-1迭代地描述的过程应用于参考序列114的每个种子。例如,一旦已如上所述处理了种子“GTTTA”114-1,计算机110就可将种子存取窗口推进到参考序列中的下一个后续种子,存取该种子,然后对参考序列的n个种子中的每个迭代地执行上文参考种子“GTTTA”114-1所述的过程。这些过程可包括:获得由种子存取窗口识别的种子,确定该种子是否具有超过预先确定数量的匹配参考序列位置,如果存在超过预先确定数量的匹配参考序列位置,则生成种子延伸树,然后使用由该种子延伸树的节点识别的种子和区间来填充散列表144。在一些具体实施中,计算机110还可迭代地执行上文参考种子“GTTTA”114-1所述的相同过程,以将参考序列114的n个种子中的每个反向互补序列。每个参考种子和每个反向互补序列的这些迭代过程最终可得到散列表140,该散列表具有x个索引条目和y个散列位置,其中对于特定参考序列诸如人类基因组,x和y各自为数亿甚至数十亿。

[0096] 在一些具体实施中,散列表140可由计算机诸如计算机110用来执行运行时灵活种子延伸,具体方法是使用一个或多个CPU、GPU或它们的组合对软件中的散列表140执行散列查询,从而执行软件指令,该软件指令在被执行时,使得所述一个或多个CPU、GPU或它们的组合执行参考图3和图4所述的过程。在其他具体实施中,计算机110可生成散列表安装包,该散列表安装包包括用于在另一个计算机上安装散列表140和种子延伸表的集合132A的软件指令。例如,散列表安装包可包括软件指令,该软件指令在被执行时,执行图2的过程200所述的操作。计算机110可向该另一个计算机提供包括该软件指令的散列表安装包。该另一个计算机可接收散列表安装包并安装散列表140和种子延伸表的集合132A。然后该另一个计算机可执行运行时灵活种子延伸,具体方法是使用一个或多个CPU、GPU或它们的组合对软件中的散列表140执行散列查询,从而执行软件指令,该软件指令在被执行时,使得所述CPU、GPU或它们的组合执行参考图3和图4所述的过程。

[0097] 然而,在一些具体实施中,计算机110可生成散列表安装包146,该散列表安装包包括硬件编程语言指令,该指令可将可编程电路162配置为在硬件数字逻辑电路中实现映射和比对单元170。该硬件编程语言指令可以是文件的形式,诸如二进制比特流文件。该二进制比特流文件可在其被包括在散列表安装包146中之前通过编译描述将由可编程电路162实现的电路的硬件编程语言代码诸如VHDL、Verilog等来生成。当由可编程电路162处理时,

散列表安装包的硬件编程语言指令可使得可编程逻辑电路对可编程逻辑电路的可动态配置的逻辑电路进行编程,以通过使用参考图3和图4所述的过程对硬件中的散列表140执行散列查询来实现灵活种子延伸。散列表安装包146还可包括种子延伸表的集合132A和用于将种子延伸表的集合132A安装在可编程电路162可访问的存储器180中的指令。散列表安装包146还可包括散列表140和用于将散列表140安装在可编程电路162可访问的存储器180中的指令。可编程电路160可被编程为将散列表140用作映射/比对单元170的一部分来执行短种子到参考序列的映射,如本文参考图3更详细地讨论的。计算机110可使用一个或多个网络、直接连接诸如一条或多条总线、USB电缆、USB-C电缆或它们的任何组合将散列表安装包提供给设备160,诸如台式计算机、膝上型计算机、平板电脑、智能电话、基于云的服务器、测序仪或容纳可编程电路160的其他设备。设备160可接收散列表安装包并且对可编程电路162进行编程,以使用散列表安装包的硬件编程语言指令在可编程电路162的硬件逻辑门中实现映射和比对单元170。

[0098] 因此,相应的散列表安装包146可用于以多种不同的方式管理散列表140和种子延伸表的安装、使用甚至移除。例如,在一些具体实施中,散列表140和种子延伸表的集合132A可各自作为文件存储在硬盘或其他存储介质上,然后各自可在运行时存取之前被加载到包括用于实现运行时灵活种子延伸的一个或多个部件或模块的公共存储器诸如DRAM中,如相对于本文参考图3和图4所述的过程所述。然而,在其他具体实施中,散列表140或种子延伸表的集合132A可一起存储或单独存储,各自作为存储器设备中的一个或多个不同连续部分或存储器设备的非连续部分。同样,在运行时映射或其他操作中,散列表140或种子延伸表的集合132A可被压缩或不被压缩,存储在公共或单独的存储介质和/或存储器上,或者被高速缓存或不被高速缓存,只要存在用于使运行时映射和比对单元170存取散列表140和种子延伸表的集合132A两者的所选部分的某种路径和方法。在其他具体实施中,散列表140可完全在可编程电路162的硬件逻辑电路中实现,并且种子延伸表的集合132A可存储在可编程逻辑电路162可访问的存储器180(诸如DRAM存储器单元)中。在其他具体实施中,散列表140可存储在可编程逻辑电路162可访问的存储器180(诸如DRAM存储器单元)中,并且种子延伸表的集合132A可完全在可编程电路162的硬件逻辑电路中实现。

[0099] 在一些具体实施中,计算机110还可生成包括散列表和种子延伸构建器的安装包,如本文所述。计算机110可通过网络向另一个计算机提供该安装包。该安装包可用于在另一个计算机或一个不同的计算机上安装该散列表和种子延伸构建器,以使得接收和安装该散列表和种子延伸构建器的一方能够从其自身所选择的参考序列并利用其自身所选择的设置来构建其自身的散列表和种子延伸表。因此,该散列表和种子延伸构建器安装包的接收方可在任何时间从其所选择的参考来构建其自己的散列表,将该散列表存储在磁盘上,将该散列表加载到可编程电路162可访问的存储器180中,以及使用可编程电路162执行映射和比对。

[0100] 图2是用于生成散列表索引的过程的流程图200,该散列表索引有利于散列表基因组映射的灵活种子延伸。一般来讲,过程200可包括:由计算机系统从参考序列获得特定核苷酸种子,该特定种子表示具有K个核苷酸的核苷酸长度的核苷酸序列(210);由该计算机系统确定该特定种子与超过预先确定数量的参考序列位置匹配(220);基于确定该特定种子与超过预先确定数量的参考序列位置匹配,由该计算机系统生成具有多个节点的种子延

伸树,其中该多个节点中的每个节点对应于(i)延伸种子,该延伸种子是该特定种子的延伸并且具有K*的核苷酸长度,其中K*为大于K的一个或多个核苷酸,和(ii)种子延伸表中的多个位置,该多个位置包括描述与该延伸种子匹配的参考序列位置的数据(230),并且对于该多个节点中的每个节点:由该计算机系统在散列表的对应于该延伸种子的索引键的位置存储区间信息,其中该区间信息引用种子延伸表中的多个位置,该多个位置包括描述与和该节点相关联的延伸种子匹配的参考序列位置的数据(240),从而生成该散列表。过程200将在下文中更详细地描述为由计算机系统诸如计算机110执行。

[0101] 更详细地,计算机系统可通过以下操作来开始执行过程200:由计算机系统从参考序列获得特定核苷酸种子,该特定种子表示具有K个核苷酸的核苷酸长度的核苷酸序列(210)。在一些具体实施中,获得特定种子可包括由计算机系统确定参考序列内的种子存取窗口的位置。然后,该计算机系统可获得参考序列的由种子存取窗口识别的子集。该计算机系统可包括一个或多个计算机。

[0102] 该计算机系统可通过以下操作来继续执行过程200:由该计算机系统确定该特定种子是否与超过预先确定数量的参考序列位置匹配(220)。如果该计算机系统确定该特定种子不与超过预先确定数量的参考序列位置匹配,则该计算机系统可确定不为该特定种子生成种子延伸树。相反,该计算机系统可获得描述与第二种子匹配的每个参考序列位置的数据。然后,该计算机系统可在散列表的对应于该特定种子的索引键的第二位置存储描述与该特定种子匹配的参考序列位置的数据。

[0103] 另选地,如果该计算机系统确定该特定种子与超过预先确定数量的参考序列位置匹配,则该计算机系统可生成具有多个节点的种子延伸树(230)。该多个节点中的每个节点可包括表示以下项的数据:(i)延伸种子,该延伸种子是该特定种子的延伸并且具有K*的核苷酸长度,其中K*是大于K的一个或多个核苷酸,和(ii)种子延伸表中的多个位置,该多个位置包括描述与延伸种子匹配的参考序列位置的数据。在一些具体实施中,该多个位置可包括延伸表中与和节点相关联的延伸种子匹配的参考序列位置的连续区间。

[0104] 该计算机系统可通过以下操作来继续执行过程200:对于种子延伸树的每个节点,将区间信息存储在散列表的散列位置中。在一些具体实施中,对于种子延伸树的每个节点,该计算机系统可将区间信息存储在散列表的对应于延伸种子的索引键的散列位置处,从而生成散列表(240)。区间信息可包括对多个种子延伸位置的引用,该多个种子延伸位置包括描述与和节点相关联的延伸种子匹配的参考序列位置的数据。在一些具体实施中,由区间信息描述的多个种子延伸表位置可包括种子延伸表中的包括描述与延伸种子匹配的参考序列位置的数据的位置的连续区间。

[0105] 使用散列表基因组映射的运行时灵活种子延伸

[0106] 图3是用于执行散列表基因组映射的运行时灵活种子延伸的运行时系统300的关系图。运行时系统300包括可编程逻辑电路162、映射和比对单元170、散列表140、存储器18和存储在存储器180中的多个种子延伸表,诸如种子延伸表132。尽管图3的示例描述了使用可编程逻辑单元162的硬件逻辑电路在硬件中实现的映射和比对单元170以及散列表140,但本公开不限于此。相反,映射和比对单元170可以是使用由存取存储在存储器单元中的散列表140的一个或多个CPU、GPU或它们的组合执行的软件指令来实现的软件应用程序。

[0107] 由系统300执行的用于散列表基因组映射的运行时灵活种子延伸可首先由映射和

比对单元170存取当前读段305。当前读段305可由已对生物样品进行初步分析的核酸测序仪生成。初步分析可包括由核酸测序仪接收生物样品诸如血液样品、组织样品或痰,以及生成输出数据诸如一个或多个读段305,该读段表示所接收的生物样品中的核酸序列中的核苷酸顺序。在一些具体实施中,生物样品可包括DNA样品,并且核酸测序仪可包括DNA测序仪。在此类具体实施中,由核酸测序仪生成的读段305中的有序核苷酸的顺序可包括形成任何组合的鸟嘌呤(G)、胞嘧啶(C)、腺嘌呤(A)和胸腺嘧啶(T)中的一者或多者。在其他具体实施中,核酸测序仪可包括RNA测序仪,并且生物样品可包括RNA样品。在此类具体实施中,由核酸测序仪生成的读段中的有序核苷酸的顺序可包括形成任何组合的G、C、A和尿嘧啶(U)中的一者或多者。因此,虽然图3的示例描述了对由DNA测序仪基于DNA样品生成的由G、C、A和T构成的读段的处理,但本公开不限于此。相反,其他具体实施可处理由RNA测序仪基于RNA样品生成的由C、G、A和U构成的读段。

[0108] 一般来讲,映射和比对单元170可被配置成不知道其接收、映射和比对的读段的类型。例如,在一些具体实施中,相同的二进制代码可用于表示“T”和“U”。由映射和比对单元170接收的读段可包括DNA、cDNA和/或RNA,并且参考可以是DNA、cDNA和/或RNA。在此类具体实施中,读段碱基T和/或U可共享单个二进制代码,使得读段T和/或U与参考T和/或U匹配。

[0109] 在一些具体实施中,核酸测序仪可包括下一代测序仪(NGS),该下一代测序仪被配置为通过使用大规模并行测序技术以实现超高通量、可延伸性和速度的方式生成给定样品的序列读段诸如读段305。NGS使得能够对整个基因组进行快速测序,能够扩大到深度测序的靶区域中,利用RNA测序(RNA-Seq)发现新型RNA变异和剪接位点,或定量mRNA用于基因表达分析,表观遗传因子诸如全基因组DNA甲基化和DNA蛋白相互作用的分析,癌症样品测序以研究罕见体细胞变异和肿瘤亚克隆,以及研究人类或环境中的微生物多样性。

[0110] 由核酸测序仪生成的序列读段诸如读段305可由辅助分析单元诸如映射和比对单元170存取和处理。在一些具体实施中,辅助分析单元诸如映射和比对单元170可使用可编程电路162诸如现场可编程门阵列(FPGA)或专用集成电路(ASIC)在硬件诸如数字逻辑电路中实现。在其他具体实施中,辅助分析单元诸如映射和比对单元170可使用一个或多个CPU、GPU或两者的组合来实现,以实现映射和比对单元170的功能。虽然在一些具体实施中可在可编程电路162的硬件逻辑电路中实现散列表140,诸如在使用可编程电路162实现映射和比对单元170的情况下,但本公开不限于此。相反,散列表140可存储在存储器设备中并在需要时由以下部件存取:(i)CPU、GPU或两者的组合,它们执行实现映射和比对单元170的功能的软件指令,或(ii)已在硬件数字逻辑电路中实现的映射和比对单元170。

[0111] 在一些具体实施中,可编程电路162可与生成了读段305的核酸测序仪集成。在此类具体实施中,例如,可编程电路162可被容纳在扩展卡诸如外围部件互连(PCI)扩展卡中并安装到核酸测序仪中。在其他具体实施中,例如,可编程电路162中的每个可以是不同于核酸测序仪的另一个计算机的一部分,并且使用以太网电缆、USB电缆、USB-C电缆等直接连接到核酸测序仪。在其他具体实施中,例如,可编程电路162可集成到基于云的服务器中,该基于云的服务器能够由生成了读段305的核酸测序仪使用一个或多个有线或无线网络诸如局域网(LAN)、广域网(WAN)、蜂窝网络、互联网或它们的组合远程访问。

[0112] 映射和比对单元170可接收包括初始种子“GTTTA”310a的第一散列查询310。在一些具体实施中,散列查询可仅由用作散列表140的输入的样品读段诸如当前读段305的种子

构成。在其他具体实施中,可将附加数据、元数据等添加到样品读段的种子以将样品读段转换成可用于搜索散列表140的格式。

[0113] 在图3的示例中,从当前读段“ACGTTTAGC”305的使用种子存取窗口305a识别的第一部分获得散列查询310中包括的初始种子“GTTTA”310a。映射和比对单元170可使用散列表140执行散列查询310,以使用散列函数143将初始短种子310a映射到散列位置144。在图3的示例中,散列查询310的执行确定种子“GTTTA”310a与散列索引键“GTTTA”142-2匹配,该散列索引键由散列函数143映射到散列位置144-y。

[0114] 映射和比对单元170可使用散列表140生成对散列查询310的响应310b。响应310b可包括散列查询310的种子310a所到达的散列位置144-y的内容。映射和比对单元170评估对散列查询310的响应310b,并确定该内容是否包括一组匹配的参考序列位置、延伸记录、区间记录或它们的组合。如果响应310b仅包括一组匹配的参考序列位置而不具有延伸记录或区间记录,则映射和比对单元170可将该组匹配的参考序列位置连同将匹配的参考位置与所接收的查询的种子相关联的元数据一起存储在种子匹配集合352中。另选地,如果映射和比对单元170确定该响应包括区间记录、延伸记录或两者,则映射和比对单元170必须确定320是使用由区间记录识别的匹配参考种子还是继续将查询的种子延伸。

[0115] 在图3的示例中,对响应310b的评估表明(i)该响应不包括一组匹配的参考序列位置,并且(ii)响应310b包括延伸记录153a和区间记录153b。基于响应310b,映射和比对单元170可确定320是否要存取由区间记录153b识别的匹配参考位置。在一些具体实施中,如果对散列查询310的响应诸如310b包括延伸记录153a,则映射和比对单元170将不存取由区间记录诸如区间记录153b识别的匹配参考序列位置。

[0116] 然而,在其他具体实施中,映射和比对单元170可被配置为在使用延伸记录153b延伸种子310a之前评估由区间记录153b识别的匹配参考序列位置的数量。在此类具体实施中,如果匹配参考序列位置的数量低于预先确定的阈值,则映射和比对单元170可输出310d处于由区间记录识别的区间A的匹配参考序列位置。输出匹配参考序列位置可包括由映射和比对单元170存取以种子延伸表132的区间A存储在存储器180中的匹配参考序列位置,以及将所存取的匹配参考序列位置存储在种子匹配集合存储装置352中。一旦所存取的匹配参考序列位置被存储在种子匹配集合352中,图3所述的过程就可以结束,而无需进一步延伸种子310a。然后,可将种子存取窗口305a沿着当前读段305向前调节一个或多个核苷酸。一旦调节了种子存取窗口305a,图3所述的过程就可再次开始并迭代地继续,直到查询了整个当前读段305。另一方面,在该另选的具体实施中,如果确定匹配参考序列位置的数量不低于预先确定的阈值,则可使用延伸记录152a来延伸种子310a。

[0117] 回到图3的示例,映射和比对单元170未将上述阈值应用于由区间记录153b识别的匹配项。相反,映射和比对单元170确定320不使用由区间记录153b识别的匹配参考序列位置,因为输出310b包括延伸记录153a。因此,在这种情况下,映射和比对单元170确定要延伸种子310a。

[0118] 在继续基于延伸种子执行后续查询之前,映射和比对单元可将描述区间A 310c的信息存储在“最佳区间”存储装置350中。区间A可被认为是用于种子310a的种子延伸表132中匹配参考序列位置的“最佳区间”,因为在该过程中此时尚未识别和评估其他区间。然而,在图3所述过程的后续迭代中,可启发式评估所识别的每个后续区间,以确定该区间是否好

于存储在用于初始种子310a或其延伸种子的最佳区间存储装置中的现有区间。将描述区间A 310c的信息存储在最佳区间存储装置340中使得在初始种子310a的延伸导致映射失败(诸如未读映射问题或高置信度映射问题)的情况下,区间A的匹配参考序列位置能够被重新访问。描述区间A 310c的信息可包括描述与初始种子匹配的参考序列位置的连续列表的起始位置和结束位置的数据。在一些具体实施中,描述区间A 310c的信息还可包括识别由区间A所识别的参考序列位置与之匹配的种子的数据。

[0119] 映射和比对单元170执行的灵活种子延伸可通过映射和比对单元170生成第一延伸种子312a而继续,该第一延伸种子是使用延伸记录153a对初始种子310a的延伸。在图3的示例中,延伸记录153a可包括指示映射和比对单元170将初始种子310a对称地延伸两个碱基或核苷酸的一个或多个指令。在图3的示例中,将初始种子“GTTTA”310a对称地延伸两个碱基或核苷酸得到了读段305的延伸种子“CGTTTAG”312a。在一些具体实施中,用于延伸初始种子310a的附加核苷酸“C”和“G”可从读段305的在由种子存取窗口305a识别的初始种子310a的相对侧上的下一个核苷酸获得。

[0120] 虽然在种子存取窗口的每侧上存在附加的种子以有利于该种子延伸,但在其他具体实施中,诸如当种子存取窗口在读段305的起点时,延伸可能导致初始种子的延伸超出读段305的边界。在此类具体实施中,种子延伸可能失败,并且使用散列表140将初始种子映射到匹配参考序列位置的过程可终止而无需针对从初始种子开始的查询循环将任何匹配参考序列位置添加到种子匹配集合352。然而,在此类具体实施中,可将种子存取窗口305a沿着读段305向前调节一个或多个核苷酸,并且可获得读段305的由经调节的种子存取窗口识别的下一个种子,以用作使用散列表140的新查询循环的散列查询的初始种子。对下一个种子以及其后的每个种子执行该新查询循环直到读段305的每个种子都已被处理为止可用于更新最佳区间存储装置350,将一组或多组匹配参考序列位置存储在种子匹配集合存储装置352中或这两项操作,可评估这些匹配参考序列位置以识别针对读段305的最佳匹配参考序列位置集合,如参考图5所述(即使种子延伸失败),从而解决常规方法中可能存在的未映射读段问题。

[0121] 当种子存取窗口305a朝向读段305的相对端推进时,可能由于类似原因而发生类似的种子延伸失败。本公开类似地通过如参考图5所述的从读段的散列查询的先前迭代中评估最佳区间存储装置250、种子匹配集合353或两者来解决这些种子延伸失败问题。

[0122] 回到图3的示例,映射和比对单元170可生成包括第一延伸种子312a的后续散列查询312。映射和比对单元170可从散列查询312获得第一延伸种子312a,并使用散列表来通过使用散列函数143将第一延伸短种子312a映射到散列位置144。在一些具体实施中,使用第一延伸种子312a生成散列查询312可包括将第一延伸种子312a作为用于使用散列表140进行种子映射的输入提供给映射和比对单元170,而无需生成查询。在图3的示例中,散列查询312的执行确定种子“CGTTTAG”312a与散列索引键“CGTTTAG”142-x匹配,该散列索引键由散列函数143映射到散列位置144-3。

[0123] 映射和比对单元170可使用散列表140生成对散列查询312的响应312b。响应312b可包括散列查询312的种子312a所到达的散列位置144-3的内容。映射和比对单元170可评估对散列查询312的响应312b,并确定响应312b(i)不包括一组匹配的参考序列位置并且(ii)包括延伸记录152a和区间记录152b。基于响应312b,映射和比对单元170可确定330是

否要存取由区间记录152b识别的匹配参考位置。在一些具体实施中,如果对散列查询312的响应诸如312b包括延伸记录152a,则映射和比对单元170将不存取由区间记录诸如区间记录152b识别的匹配参考序列位置。

[0124] 然而,在其他具体实施中,映射和比对单元170可被配置为在使用延伸记录152b延伸种子312a之前评估由区间记录152b识别的匹配参考序列位置的数量。在此类具体实施中,如果由区间记录152b识别的匹配参考序列位置的数量低于预先确定的阈值,则映射和比对单元170可以由区间记录152b识别的区间B输出312d匹配参考序列位置。输出匹配参考序列位置可包括由映射和比对单元170存取以种子延伸表132的区间B存储在存储器180中的匹配参考序列位置,以及将所存取的匹配参考序列位置存储在种子匹配集合存储装置352中。一旦所存取的匹配参考序列位置被存储在种子匹配集合存储装置352中,图3所述的过程就可以结束,而无需进一步延伸种子312a。然后,可将种子存取窗口305a沿着当前读段305调节一个或多个核苷酸。一旦调节了种子存取窗口305a,图3所述的过程就可再次开始并迭代地继续,直到查询了整个当前读段305。另一方面,在该另选的具体实施中,如果确定匹配参考序列位置的数量不低于预先确定的阈值,则可使用延伸记录152a来延伸种子312a。

[0125] 回到图3的示例,映射和比对单元170未将上述阈值应用于由区间记录152b识别的匹配项。相反,映射和比对单元170确定330不使用由区间记录152b识别的匹配参考序列位置,因为输出312b包括延伸记录152a。因此,映射和比对单元170确定要延伸种子312a。

[0126] 在继续基于延伸种子执行后续查询之前,映射和比对单元可确定是否要将把区间B 312c描述为“最佳区间”的信息存储在最佳区间存储装置350中。确定是否要存储把区间B 312c描述为“最佳区间”的信息包括启发式地确定区间B是否是比当前存储在最佳区间存储装置352中的用于第一延伸种子的先前迭代的区间更好的区间,在该示例中该先前区间是区间A。在一个具体实施中,可通过评估为每个区间返回的目标命中项的数量来确定多个区间中的最佳区间。在此类具体实施中,可根据multipart规则来选择“最佳”区间。以举例的方式,映射和比对单元170可将第一优先级分配给包含至少预先确定数量的匹配参考序列位置的区间,该区间可通过阈值诸如intvl-target-hits(32)个匹配项来指代。然而,如果每个区间具有少于intvl-target-hits(32)个匹配项,则把具有最多匹配项的区间作为最佳区间来存储。此外,映射和比对单元170可将第二优先级分配给与较长延伸种子相关联的区间,因为此类区间可为优选的。此外,如果映射和比对单元170确定至少一个区间具有至少intvl-target-hits(32)个匹配项,则在满足至少intvl-target-hits(32)个匹配项的所有区间中,基于与最长延伸种子相关联的区间来选择最佳区间。虽然此处的示例引用了具有32个匹配项的阈值intvl-target-hits(32),但本公开不需要限于此。相反,可以将阈值intvl-target-hits()发送到任何数量的匹配参考序列位置,以实现该multi-part启发式规则。

[0127] 在图3的示例中,最佳区间存储装置350中的先前作为最佳区间存储的区间A识别六个匹配的参考序列位置132-1至132-6,并且区间B识别三个匹配的参考序列位置132-4至132-6。应用十个匹配项的示例性intvl-target-hits(10)阈值,映射和比对单元170可应用multi-part启发式规则并确定这两个区间都不满足intvl-target-hits(10)阈值。因此,根据multi-part启发式规则,映射和比对单元170可选择区间A作为最佳区间,因为在区间A和

区间B两者中,区间A具有最多匹配项,即六个。基于该示例性multi-part启发式规则的应用,可丢弃描述区间B 321c的信息,并且保持将区间A作为最佳区间来存储。然而,在应用不同的启发式规则(该规则不需要是multi-part启发式规则)的其他示例中,可选择区间B作为最佳区间并将其存储在最佳区间存储装置350中来替换区间A。此类结果可最终留给特定设计配置,诸如intvl-target-hits()阈值的设置、一个或多个启发式规则的设计等。

[0128] 在图3的示例中,前述启发式规则用于将先前存储在最佳区间存储装置350中的区间A和包括在对查询312的响应312b中的区间B进行比较。然而,本公开不需要限于此。例如,在一些具体实施中,对散列查询的响应可包括存储在散列查询的特定种子所到达的散列位置144的多个区间记录。在此类具体实施中,映射和比对单元170可应用上述启发式规则来确定应存取该多个区间记录中的哪一个。同样,映射和比对单元170还可使用此类启发式规则从查询响应中返回的每个区间记录中确定要存储在最佳区间存储装置450中的最佳区间。又如,映射和比对单元170还可使用此类启发式规则,从查询响应中返回的每个区间记录和先前存储在最佳区间存储装置350中的用于返回多个区间的查询的种子的先前迭代的另一个区间中确定要存储在最佳区间存储装置450中的最佳区间。

[0129] 在一些具体实施中,系统300可有利于在最佳区间存储装置350中存储多于一个最佳区间。例如,在一些具体实施中,可记录最多2个最佳区间。在一些具体实施中,可记录最多N个最佳区间。在此类具体实施中,如果存储了N>1个最佳区间,则用于确定哪些区间被保留的标准可涉及评估候选区间、其相关联的延伸种子或这两者之间的关系,诸如要求该N个最佳区间与在读段内彼此不重叠的延伸种子相关联。

[0130] 映射和比对单元170执行的灵活种子延伸可通过映射和比对单元170生成第二延伸种子314a而继续,该第二延伸种子是使用延伸记录152a对第一延伸种子312a的延伸。在图3的示例中,延伸记录152a可包括指示映射和比对单元170将第一延伸种子312a对称地延伸两个碱基或核苷酸的一个或多个指令。在图3的示例中,将第一延伸种子“CGTTTAG”312a对称地延伸两个碱基或核苷酸得到了读段305的第二延伸种子“ACGTTTAGC”314a。在一些具体实施中,用于延伸第一延伸种子312a的附加核苷酸“A”和“C”可从读段305的在第一延伸种子“CGTTTAG”312a的相对侧上的下一个核苷酸获得。

[0131] 回到图3的示例,映射和比对单元170可生成包括第二延伸种子314a的后续散列查询314。映射和比对单元170可从散列查询314获得第二延伸种子314a,并使用散列表来通过使用散列函数143将第二延伸短种子314a映射到散列位置144。在一些具体实施中,使用第二延伸种子314a生成散列查询314可包括将第二延伸种子314a作为用于使用散列表140进行种子映射的输入提供给映射和比对单元170,而无需生成查询。在图3的示例中,散列查询314的执行确定种子“ACGTTTAGC”314a与散列索引键“ACGTTTAGC”142-1匹配,该散列索引键由散列函数143映射到散列位置144-2。

[0132] 映射和比对单元170可使用散列表140生成对散列查询314的响应314b。响应314b可包括散列查询314的第二延伸种子314a所到达的散列位置144-2的内容。映射和比对单元170评估对散列查询314的响应314b,并确定响应314b(i)包括一组匹配的参考序列位置155,(ii)不包括延伸记录,并且(iii)不包括区间记录。基于响应314b,映射和比对单元170可确定应将匹配参考序列位置155存储在种子匹配集合存储装置352中。

[0133] 由于响应314b不包括延伸记录,因此针对读段305的种子“GTTTA”310a的运行时灵

活种子延伸过程结束。种子存取窗口305a可继续沿着读段305前进一个或多个核苷酸,直到对读段305的每个相应种子执行了参考图3所述的每个过程。该过程也相对于图4的流程图进行描述。如上所述,当种子存取窗口305a朝向读段305的末端延伸时,尝试将输入到映射和比对单元170的种子进行延伸可能会失败,从而产生潜在的未映射读段问题。然而,本公开可使用存储在最佳区间存储装置中的一个或多个区间、存储在种子匹配集合352中的一个或多个读段或前述两者的组合来识别用于读段305的一组匹配的参考序列位置,如相对于至少图5所述。

[0134] 图4是用于执行散列表基因组映射的运行时灵活种子延伸的过程400的流程图。过程400将在下文中解释为由一个或多个计算机组成的计算机系统执行。该一个或多个计算机可包括例如映射和比对单元170。出于本公开的目的,该一个或多个计算机可包括被配置为获得和执行软件指令以实现由软件指令描述的特定编程功能的CPU或GPU。另选地或除此之外,该一个或多个计算机可包括可编程电路,该可编程电路已被配置为使得该可编程电路的硬件数字逻辑电路已被配置为在硬件中实现特定编程功能。

[0135] 该计算机系统可通过执行对散列表405的查询来开始执行过程400。该查询可包括核苷酸种子。核苷酸种子可包括从读段获得的核苷酸子集。读段可包括由核酸测序仪基于输入到该核酸测序仪中的生物样品而生成的一组核苷酸。生物样品可包括例如血液样品、组织样品、痰等。

[0136] 以举例的方式,由核酸测序仪基于生物样品生成的读段可包括一系列核苷酸,诸如“ACGTTTAGC”。该示例包括9个核苷酸组成的读段。然而,9个核苷酸组成的读段仅用作示例。代替限制在9个核苷酸,如本公开所述的读段可具有任何核苷酸长度,包括但不限于5个碱基或核苷酸、10个碱基或核苷酸、12个碱基或核苷酸、15个碱基或核苷酸、18个碱基或核苷酸、21个碱基或核苷酸、25个碱基或核苷酸、35个碱基或核苷酸、50个碱基或核苷酸、100个碱基或核苷酸、150个碱基或核苷酸、1,000个碱基或核苷酸、1,000,000个碱基或核苷酸,甚至更多个碱基或核苷酸。查询的种子可包括读段的一部分,诸如“GTTTA”。在过程400的第一迭代期间从用于第一散列查询的读段获得的种子可具有任何长度K,其中K小于该读段中碱基或核苷酸的数量。在一些具体实施中,K可显著小于读段核苷酸长度,诸如读段长度的1/100、读段长度的1/10或读段长度的1/5等。

[0137] 计算机系统可通过获得种子并将种子与散列表的散列键进行比较来执行包括该种子的查询。散列键可对应于每个参考序列种子、每个参考序列种子的反向互补序列、参考序列的每个延伸种子,以及参考序列的每个延伸种子的反向互补序列。参考序列可包括例如物种诸如人类或其他动物的参考基因组或其部分。当计算机系统识别出与查询的种子匹配的散列键时,计算机系统可使用散列函数将该散列键映射到一个或多个散列位置。在本公开的一些方面,一个或多个散列位置可存储(i)延伸记录,(ii)区间记录,或(iii)一个或多个参考序列位置。计算机系统可生成对查询的响应,该响应包括查询的种子所到达的一个或多个散列位置的内容。

[0138] 计算机系统可通过以下操作来继续执行过程400:获得对所执行的查询的响应,该响应包括由散列表的被确定为由该查询到达的一个或多个位置所存储的信息(410)。如果确定查询的种子与使用散列函数映射到散列表的该一个或多个位置的散列键匹配,则确定该查询到达该一个或多个位置。

[0139] 计算机系统可通过以下操作来继续执行过程400:确定对所执行的查询的响应是否包括(i)延伸记录,(ii)区间记录,或(iii)一个或多个匹配的参考序列位置(415)。由计算机系统确定对所执行的查询的响应是否包括(i)延伸记录(ii)区间记录或(iii)一个或多个匹配的参考序列位置(415)可包括解析所接收的响应并分析所解析的响应数据。计算机系统可基于所解析的数据来确定所解析的数据是否表示(i)延伸记录,(ii)区间记录或(iii)一个或多个匹配的参考序列位置。在其他具体实施中,对所执行的查询的响应可包括表明响应是否包括(i)延伸记录,(ii)区间记录或(iii)一个或多个匹配的参考序列位置的一个或多个数据标记。

[0140] 在一些情况下,计算机系统可通过以下操作来继续执行过程400:在阶段415处,确定响应不包括延伸记录、区间记录或一个或多个匹配的参考序列位置。如果计算机系统确定响应不包括(i)延伸记录,(ii)区间记录或(iii)一个或多个匹配的参考序列位置,则该过程在阶段420处结束,而不向用于该查询的种子的种子匹配集合添加任何匹配的参考序列位置。以举例的方式,如果种子是延伸种子并且存在种子延伸错误,则对包括该种子的查询的获得的响应可不包括(i)延伸记录,(ii)区间记录,或(iii)一个或多个匹配的参考序列位置。例如,如果计算机系统尝试将种子延伸超出用来获得种子的读段的末端,则可能存在这种种子延伸错误。

[0141] 另选地,在其他情况下,计算机系统可通过以下操作来继续执行过程400:在阶段415处,确定对所执行的查询的响应包括(i)延伸记录,(ii)区间记录或(iii)两者。在此类情况下,计算机系统可通过以下操作来继续执行过程400:确定是否要存取延伸表以获得该延伸表中的由区间记录引用的一个或多个匹配的参考序列位置(430)。

[0142] 在一些情况下,计算机系统可通过以下操作来继续执行过程400:确定要存取种子延伸表以获得该延伸表中的由区间记录引用的一个或多个匹配的参考序列位置。例如,在一些具体实施中,计算机系统可被配置为:在匹配的参考序列位置的数量低于预先确定的阈值的情况下,存取种子延伸表以获得由区间记录识别的一个或多个匹配的参考序列位置。另选地或除此之外,计算机系统可被配置为:在对所执行的查询的响应还包括存储在散列查询的种子所到达的散列位置的“停止”记录的情况下,存取种子延伸表以获得由区间记录识别的一个或多个匹配的参考序列位置。“停止”记录可指示计算机系统不执行查询中的种子的任何另外的种子延伸,并且优先存取由区间记录识别的一个或多个匹配的参考序列位置,诸如如果匹配的参考序列位置的数量降至预先确定的阈值以下。

[0143] 在此类情况下,在计算机系统在阶段430处确定要存取种子延伸表的情况下,计算机系统可通过以下操作来继续执行过程400:存取种子延伸表以获得种子延伸表中的该一个或多个参考序列位置(450)。计算机系统可通过使用区间记录来识别要从种子延伸表获得的一个或多个匹配参考序列位置的特定集合。区间记录可包括引用种子延伸表中的多个位置的信息,该多个位置包括描述与查询的种子匹配的参考序列位置的数据。在一些具体实施中,引用该多个位置的信息可包括延伸表中与查询的延伸种子匹配的参考序列位置的连续区间。另选地,在其他具体实施中,引用该多个位置的信息可包括延伸表中与查询的种子匹配的参考序列位置的一个或多个非连续区间。

[0144] 在此类情况下,计算机系统可从种子延伸表获得使用区间记录识别的一个或多个匹配的参考序列位置。可将所获得的一个或多个参考序列位置添加到种子匹配集合(455)。

在一些具体实施中,将一个或多个匹配的参考序列位置添加到种子匹配集合可包括获得表示该一个或多个匹配参考序列位置的数据并将其存储在存储器设备的被分配用于种子匹配集合存储的位置。在其他具体实施中,将一个或多个匹配的参考序列位置添加到种子匹配集合可包括存储引用存储一个或多个参考序列位置的种子延伸表的区间的数据,诸如指针。因此,种子匹配集合可以是存储一组已识别并获得的匹配参考序列位置的存储位置。另选地,种子匹配集合可包括存储对一个或多个匹配参考序列位置的引用的一个或多个存储位置。在将由区间记录识别的一个或多个匹配的参考序列位置添加到种子匹配之后,计算机系统可在460处结束过程400的该实例。

[0145] 在其他情况下,在计算机系统确定(415)响应包括至少(i)延伸记录,(ii)区间记录或(iii)或两者之后,计算机系统可确定(430)不存取种子延伸表来获得一个或多个匹配的参考序列位置。计算机系统关于不存取种子延伸表来获得一个或多个匹配的参考序列位置的确定可基于多种因素。以举例的方式,在一些具体实施中,如果响应返回延伸记录,则计算机系统可确定不存取种子延伸表来获得由区间记录识别的匹配参考序列位置。这种确定可以是优选的,因为延伸的种子可能产生比由区间记录识别的匹配参考序列位置集合更小的匹配参考序列位置集合。

[0146] 又如,在其他具体实施中,如果计算机系统确定匹配参考序列位置的数量超过预先确定的匹配参考序列位置的阈值数量,则计算机系统可确定不存取种子延伸表来获得由区间记录识别的匹配参考序列位置。类似地,在此类具体实施中,当由区间识别的匹配参考序列位置超过匹配阈值时,计算机系统可确定不存取种子延伸表。

[0147] 当计算机系统确定430不存取种子延伸表时,计算机系统可通过以下操作来继续执行过程400:确定465所获得的响应是否包括区间记录和延伸记录。如果计算机系统确定465所获得的响应包括区间记录和延伸记录,则计算机系统可确定435是否将包括在对所执行查询的响应中的区间记录或描述区间记录的信息作为候选最佳区间来存储。在过程400的针对具有尚未延伸的初始种子的查询的首次迭代期间,计算机系统可确定将区间记录或描述区间记录的信息作为候选最佳区间存储在存储器设备的最佳区间存储装置中。因为在过程400的针对具有尚未延伸的初始种子的查询的初始迭代期间遇到此类区间记录,所以在对一个或多个后续延伸种子的其他查询的响应中没有遇到其他区间记录。因此,在对具有尚未延伸的初始种子的查询的响应中返回的第一区间必须是“最佳区间”,因为尚未识别出其他区间用来进行比较。

[0148] 然而,对于过程400中的在已接收到对具有延伸种子的查询的响应之后的后续交互,计算机系统可从对具有延伸种子的查询的响应获得第二区间记录。在此类情况下,计算机系统可启发式地确定是否应将第二区间记录用于替换最佳区间存储装置中先前存储的候选最佳区间。关于是保持先前存储的候选最佳区间还是将候选最佳区间替换为第二区间或描述该区间的信息的确定可通过应用一个或多个启发式规则来进行,如参考图3的示例所述。在一些具体实施中,该启发式规则可包括一个或多个multi-part启发式规则。

[0149] 虽然本公开的一些具体实施可涉及对照先前存储的候选最佳区间迭代地评估每个后续返回的区间记录以确定应针对查询的种子所依据的当前读段来存储的单个最佳区间,但本公开不需要限于此。相反,在一些具体实施中,可将所有区间保存在区间存储装置中并随后评估以用于补充种子匹配集合。

[0150] 计算机系统可通过生成延伸种子(440)来继续执行过程400。可基于在对查询的响应中返回的延伸记录中包括的指令来生成延伸种子。以举例的方式,延伸记录可包括一个或多个指令,这些指令当由计算机诸如执行软件指令的中央处理单元(CPU)或图形处理单元(GPU)或可编程电路162执行时,可使得该CPU、GPU或可编程电路延伸散列查询中使用的种子,该种子到达散列位置,该散列位置通过一个或多个核苷酸来存储延伸记录。在一些具体实施中,可生成延伸记录,使得延伸记录指示计算机在种子的每一端对称地延伸种子。因此,以举例的方式,可生成延伸记录以指示计算机诸如CPU、GPU或可编程电路162将种子延伸两个核苷酸、四个核苷酸、六个核苷酸等。在此类具体实施中,可通过使种子在种子的每个相应端上延伸一个核苷酸,在种子的每个相应端上延伸两个核苷酸,在种子的每个相应端上延伸三个核苷酸等来实现种子的对称延伸。然而,本公开不应限于种子的对称延伸。相反,本公开还设想了种子的非对称延伸。

[0151] 计算机系统可通过生成445包括延伸种子的散列查询来继续执行过程400。然后,计算机系统可通过在阶段405处利用延伸查询执行查询来执行过程400的另一次迭代,然后继续执行过程400,直到(a)该过程在427或460处通过将一个或多个匹配的参考序列位置添加到种子匹配集合而结束,该过程在475处在确定是否要将区间记录作为候选最佳区间来存储之后结束,或者(c)该过程在阶段420处由于一个或多个错误而结束,所述错误诸如种子延伸错误,这种错误导致未接收到对所执行的查询的包括(i)延伸记录,(ii)区间记录或(iii)一个或多个匹配的参考序列位置的响应的查询。

[0152] 另选地,如果在阶段465处计算机系统确定所获得的响应不包括区间记录和延伸记录两者,则计算机系统可通过确定所获得的响应是否包括延伸记录来继续执行过程400。

[0153] 如果计算机系统确定所获得的响应包括延伸记录,则计算机系统可通过以下操作来继续执行过程400:在阶段440处生成延伸种子,生成包括延伸种子的散列查询445,以及通过在阶段405处利用延伸查询执行查询来执行过程400的另一次迭代。然后,计算机系统可继续执行过程400,直到(a)该过程在427、420、460、475处结束。

[0154] 另一方面,如果计算机系统确定所获得的响应不包括延伸记录,则计算机系统可通过以下操作来继续执行过程400:在阶段470处确定是否要将区间记录或描述区间记录的信息作为候选最佳区间来存储。计算机系统可使用相对于在阶段435处确定是否要将区间记录作为候选最佳区间来存储所述的相同过程来在阶段470处确定是否要将区间记录作为候选最佳区间来存储。无论计算机系统在阶段470处是否确定将区间记录作为候选最佳区间来存储,过程400都在阶段475处结束。

[0155] 可实现过程400的至少一个变型,其中计算机系统转为在阶段470处确定所获得的响应是否包括区间记录。在此类情况下,从逻辑上讲是这样:如果计算机系统确定所获得的响应包括区间记录,则计算机系统可在阶段470处继续执行该过程。另选地,如果计算机系统确定所获得的响应不包括区间记录,则该过程通过在阶段440处生成延伸种子而继续。可类似地实现过程400的过程流程的其他变型,并且这些变型落入本公开的实质和范围内。

[0156] 图5是用于对读段的每个种子执行散列表基因组映射的迭代运行时灵活种子延伸的过程500的流程图。一般来讲,过程500可包括:获得由核酸测序仪生成的读段(505);确定种子存取窗口的位置,其中该种子存取窗口识别读段的种子(510);生成散列查询,该散列查询包括由种子存取窗口识别的种子(515);以及开始执行图4所述的过程400,具体是通过

在阶段410处执行生成的散列查询并继续过程400的迭代执行直到过程400结束(520);确定读段是否包括另一个种子(525);以及如果确定读段包括另一个种子(525),则调节种子存取窗口以识别该另一个种子(530);以及执行阶段515以生成使用该另一个种子的散列查询(515)。

[0157] 过程500可以继续执行阶段515、520、525和530的处理循环,直到在阶段525处确定在阶段505处获得的读段不包括要使用过程400映射和比对的另一个种子为止。在此类情况下,可确定是否使用最佳区间来补充读段的当前种子匹配集合(535)。如果在阶段535处确定要使用最佳区间来补充当前种子匹配集合,则过程500可通过以下操作来继续:在阶段540处通过处理最佳区间(540),使用从使用最佳区间识别的种子延伸表的一部分获得的一个或多个匹配参考序列位置补充当前种子匹配集(545),以及确定550是否存在准备好使用过程500进行映射和比对的另一个读段。如果不存在准备好用于映射和比对的另一个读段,则过程500在阶段555处结束。另选地,如果存在准备好使用过程500进行映射和比对的另一个读段,则过程500通过在阶段505处获得准备好进行映射和比对的另一个读段来继续。过程500然后可继续迭代地执行过程500,直到在阶段550处确定不存在准备好使用过程500进行映射和比对的另一个读段为止。

[0158] 过程500将在下文更详细地描述为由一个或多个计算机组成的计算机系统执行。该一个或多个计算机可包括例如映射和比对单元170。出于本公开的目的,该一个或多个计算机可包括被配置为获得和执行软件指令以实现由软件指令描述的特定编程功能的CPU或GPU。另选地或除此之外,该一个或多个计算机可包括可编程电路,该可编程电路已被配置为使得该可编程电路的硬件数字逻辑电路已被配置为在硬件中实现特定编程功能。

[0159] 计算机系统可通过以下操作来开始执行过程500:获得表示由核酸测序仪生成的核酸读段(在本文中也称为“读段”)的数据(505)。在核酸测序仪生成读段之后,可由计算机系统将该读段作为来自核酸测序仪的输入而接收。另选地或除此之外,由核酸测序仪生成的读段可存储在计算机系统可访问的存储器设备中。然后,计算机系统500可通过存取该存储器以从存储器设备检索一个或多个读段来获得所存储的读段。以举例的方式,读段可包括一组核苷酸,诸如“ACGTTTAGC”。该示例包括9个核苷酸组成的读段。然而,9个核苷酸组成的读段仅用作示例。代替限制在9个核苷酸,如本公开所述的读段可具有任何核苷酸长度,包括但不限于5个碱基或核苷酸、10个碱基或核苷酸、12个碱基或核苷酸、15个碱基或核苷酸、18个碱基或核苷酸、21个碱基或核苷酸、25个碱基或核苷酸、35个碱基或核苷酸、50个碱基或核苷酸、100个碱基或核苷酸、150个碱基或核苷酸、1,000个碱基或核苷酸、1,000,000个碱基或核苷酸,甚至更多个碱基或核苷酸。

[0160] 计算机系统可通过确定种子存取窗口的位置(510)来继续执行过程500。种子存取窗口可用于识别由读段的核苷酸的子集构成的核苷酸种子。种子的示例是有序核苷酸的集合“GTTTA”,它是读段“ACGTTTAGC”的种子。虽然有序核苷酸的集合“GTTTA”表示读段“ACGTTTAGC”的连续种子的示例,但本公开不需要限于此。相反,在一些具体实施中,可使用本公开所述的系统和方法获得非连续种子并对其进行分析。例如,还可从读段“ACGTTTAGC”获得非连续种子诸如“G_T_A”,并且使用本文所述的系统和方法对其进行分析。在此类具体实施中,本公开的系统和方法可将由下划线“_”表示的跳过位置视为可与任何碱基或核苷酸匹配的通配符。

[0161] 种子存取窗口可被配置为具有小于读段长度的任何碱基或核苷酸长度。种子存取窗口可被配置为沿着有序读段向前或向后移动以识别读段的要用于处理的种子。如果要利用非连续种子,则可相应地配置种子存取窗口。以举例的方式,种子存取窗口可被配置为识别九个核苷酸的非连续种子,其中在核苷酸位置6和核苷酸位置8处插入通配符。

[0162] 计算机系统可通过生成包括由种子存取窗口识别的种子的散列查询(515)来继续执行过程500。在一些具体实施中,散列查询可仅由读段诸如“ACGTTTAGC”的种子“GTTTA”构成。在其他具体实施中,可将附加数据、元数据等添加到样品读段的种子以将该种子转换成可用于搜索散列表的格式。

[0163] 计算机系统可通过以下操作来继续执行过程500:执行图4所述的过程400(520)以将所生成的查询的种子映射和比对到一个或多个参考序列位置。计算机系统通过在阶段410处执行在阶段515处生成的散列查询来开始执行过程400。然后,计算机系统可继续迭代执行过程400,直到过程400在阶段420、427、460或475处终止为止,这时可能已经在阶段425或455处将匹配的参考序列位置添加到种子匹配集合。

[0164] 在过程400终止之后,计算机系统可确定在阶段505处获得的读段是否包括另一个种子(525)。在一些具体实施中,确定读段是否包括另一个种子包括考虑读段中所有可能的种子存取窗口位置。另选地,确定读段是否包括另一个种子可包括仅考虑所有可能的种子存取窗口位置的预先确定的子集,诸如仅偶数种子存取窗口位置或仅奇数种子存取窗口位置。因此,本公开不要求读段的每个种子都使用过程500来评估。相反,在一些具体实施中,计算机系统可在阶段505处确定读段的种子的预先确定的子集是否存在要使用过程500评估的另一个种子。

[0165] 如果计算机系统在阶段525处确定读段包括另一个种子,则计算机系统可调节种子存取窗口以识别该另一个种子(530),并且计算机系统可执行阶段515以生成使用经调节的种子存取窗口识别的该另一个种子的散列查询(515)。调节种子存取窗口可包括例如将种子存取窗口沿着在阶段505处获得的读段向前移动一个或多个碱基或核苷酸位置。计算机系统可继续执行阶段515、520、525和530的处理循环,直到计算机系统在阶段525处确定在阶段505处获得的读段不包括要使用过程400映射和比对的另一个种子为止。

[0166] 一旦计算机系统确定在阶段505处获得的读段不包括要映射和比对的另一个种子,计算机系统就可确定是否使用最佳区间来补充读段的当前种子匹配集合(535)。在一些情况下,如果计算机系统确定不应补充种子匹配集合,则计算机系统可确定(550)是否存在准备好使用过程500进行映射和比对的另一个读段。在此类情况下,如果计算系统确定存在准备好用于映射和比对的另一读段,则计算机系统可通过在阶段505处获得准备好用于映射和比对的另一个读段来继续执行过程500。然后计算机系统可继续迭代地执行过程500,直到在阶段550处确定不存在准备好使用过程500进行映射和比对的另一个读段为止。

[0167] 另选地,在其他情况下,计算机系统可确定应使用由最佳区间识别的一个或多个匹配参考序列位置来补充读段的当前种子匹配集合。计算机系统可通过将一个或多个启发式规则应用于(i)延伸种子的种子长度,该延伸种子的查询产生最佳区间,(ii)一个或多个匹配参考序列位置的种子长度,(iii)所生成的种子链的数量,或它们的组合,来确定应使用由最佳区间识别的一个或多个匹配参考序列位置来补充当前种子匹配集合。在一些具体实施中,启发式规则可指定一个或多个独立触发条件,这些条件如果被触发,则使得计算机

系统处理最佳区间。

[0168] 以举例的方式,可触发计算机系统对最佳区间进行处理的第一独立触发条件是确定其查询产生最佳区间的延伸种子的种子长度是否大于或等于intvl-seed-length(60)个碱基或核苷酸。在该示例中,阈值intvl-seed-length(60)是预先确定的阈值,该阈值可由计算机系统用于评估产生最佳区间的延伸种子的长度。在该示例中,产生供计算机系统检查的最佳区间的延伸种子的种子长度为六十个核苷酸。然而,本公开不需要限于此。相反,可将阈值intvl-seed-length()设定为任何核苷酸长度。如果计算机系统确定不满足intvl-seed-length()阈值,则计算机系统可评估其他触发条件来确定是否要处理最佳区间。

[0169] 又如,可触发计算机系统对最佳区间进行处理的第二独立触发条件是确定其查询产生最佳区间的延伸种子的种子长度是否大于由至少intvl-seed-longer(8)个碱基或核苷酸处理的最长匹配参考序列位置。在该示例中,阈值intvl-seed-longer(8)是可由计算机系统用来评估(i)其查询产生最佳区间的延伸种子的种子长度与(ii)最长匹配参考序列位置的比较的预先确定的阈值。在该示例中,如果计算机系统确定其查询产生最佳区间的延伸种子的种子长度为八个碱基或核苷酸,或多于、大于任何匹配种子,则触发对最佳区间的处理。

[0170] 又如,可触发计算机系统对最佳区间进行处理的第三独立触发条件是确定种子链的数量是否小于intvl-min-chains(8)。种子链可包括一组类似定位的参考序列位置匹配项。在该示例中,阈值intvl-min-chains(8)是可用于评估所生成的种子链的数量的预先确定的阈值。在该示例中,如果生成少于八个种子链,则触发对最佳区间的处理。

[0171] 虽然描述了用于触发对最佳区间的处理以补充种子匹配集合的三个独立触发条件的示例,但本公开不需要限于此。相反,其他触发条件可被构造成根据特定计算机系统可能需要的来触发对最佳区间的处理。

[0172] 如果计算机系统在阶段535处确定要补充种子匹配集合,因为例如已经满足用于处理最佳区间的触发条件的一个或多个阈值,则计算机系统可在阶段535处确定使用最佳区间来补充当前种子匹配集合。使用最佳区间来补充当前种子匹配集合可包括计算机系统处理最佳区间(540)。处理最佳区间可包括将一个或多个启发式规则应用于最佳区间以识别由最佳区间识别并存储在种子延伸表中的一个或多个匹配的参考序列位置。

[0173] 以举例的方式,如果由最佳区间识别的参考序列位置的数量小于或等于intvl-max-hits(64),则计算机系统可确定处理由最佳区间识别的一个或多个参考序列位置中的全部。在该示例中,如果计算机系统确定识别出的最佳区间小于或等于六十四个匹配的参考序列位置,则计算机系统可从使用最佳区间的种子延伸表获得由最佳区间识别的所有匹配的参考序列位置。另选地,如果计算机系统确定最佳区间识别出多于六十四个的匹配参考序列位置,则计算机系统可从由最佳区间识别的匹配参考序列位置集合中随机获得intvl-sample-hits(32)个匹配参考序列。

[0174] 随机获得阈值量的三十二个匹配参考序列位置可包括随机地或通过确定性伪随机选择,从使用最佳区间的种子延伸表获得阈值量的三十二个匹配参考序列位置。最佳区间可包括识别(i)种子延伸表的一个或多个停止位置和起始位置,(ii)一个或多个起始位置和一个或多个偏移,或它们的组合的数据。虽然描述了阈值的示例,诸如六十四个匹配参

考位置和32个随机采样命中项,但是本公开不需要限于此。相反,具有其他数值的其他阈值可用于实现本公开的优点。

[0175] 使用最佳区间获得的匹配参考序列位置可用于补充当前种子匹配集合545。这样使用最佳区间来补充种子匹配集合可解决诸如未映射读段问题或高置信度误映射问题之类的问题,这些问题可能分别导致没有匹配的参考序列位置被存储在种子匹配集合中或只有非常少的匹配参考序列位置被存储种子匹配集合中。匹配的参考序列位置可能已经或可从种子延伸表的由最佳区间识别的部分获得(540)。

[0176] 一旦已经补充了种子匹配集合,计算机系统就可确定是否存在准备好使用过程500进行映射和比对的另一个读段。如果存在准备好用于映射和比对的另一个读段,则计算机系统通过获得该另一个读段来继续执行过程500。另选地,如果不存在准备好用于映射和比对的另一个读段,则过程500可在555处结束。

[0177] 需注意,在参考过程500所述的示例中,评估最佳区间以确定最佳区间或其一部分是否可用于补充种子匹配集合。然而,并不要求在最佳区间存储装置中仅存储单个最佳区间。在一些具体实施中,计算机系统可有利于在最佳区间存储装置中存储多于一个最佳区间。例如,在一些具体实施中,可记录最多2个最佳区间。在一些具体实施中,可记录最多N个最佳区间。在此类具体实施中,如果存储了N>1个最佳区间,则用于确定哪些区间被保留的标准可涉及评估候选区间、其相关联的延伸种子或这两者之间的关系,诸如要求该N个最佳区间与在读段内彼此不重叠的延伸种子相关联。在一些具体实施中,计算机系统甚至可从多个不同的最佳区间中选择匹配的参考序列位置。可随机地、伪随机地或通过应用一个或多个启发法从多个不同的最佳区间中对匹配的参考序列位置进行这种选择。

[0178] 系统部件

[0179] 图6是可用于实现本文所述的与散列表基因组映射的灵活种子延伸相关的系统的系统部件的示意图。

[0180] 计算设备600旨在表示各种形式的数字计算机,诸如膝上型计算机、台式计算机、工作站、个人数字助理、服务器、刀片式服务器、大型机和其他适当的计算机。计算设备650旨在表示各种形式的移动设备,诸如个人数字助理、蜂窝电话、智能电话和其他类似的计算设备。另外,计算设备600或650可包括通用串行总线(USB)闪存驱动器。USB闪存驱动器可存储操作系统和其他应用程序。USB闪存驱动器可包括输入/输出部件,诸如可插入到另一个计算设备的USB端口中的无线发射器或USB连接器。此处所示的部件、它们的连接和关系以及它们的功能仅意在作为示例,并不意在限制本文档中描述和/或要求保护的本发明的具体实施。

[0181] 计算设备600包括处理器602、存储器604、存储设备608、连接到存储器604和高速扩展端口610的高速接口608,和连接到低速总线614和存储设备608的低速接口612。部件602、604、608、608、610和612中的每个部件使用各种总线互连,并且可安装在公共主板上或视情况以其他方式安装。处理器602可处理用于在计算设备600内执行的指令,包括存储在存储器604中或存储设备608上的指令,以在外部输入/输出设备(诸如耦接到高速接口608的显示器616)上显示GUI的图形信息。在其他具体实施中,可视情况将多个处理器和/或多个总线与多个存储器和多种类型的存储器一起使用。另外,可连接多个计算设备600,每个设备提供必要操作的一些部分,例如,作为服务器库、一组刀片式服务器或多处理器系统。

[0182] 存储器604将信息存储在计算设备600内。在一个具体实施中,存储器604是一个或多个易失性存储器单元。在另一个具体实施中,存储器604是一个或多个非易失性存储器单元。存储器604还可以是另一种形式的计算机可读介质,诸如磁盘或光盘。

[0183] 存储设备608能够为计算设备600提供海量存储。在一个具体实施中,存储设备608可以是或包含计算机可读介质,诸如软盘设备、硬盘设备、光盘设备或磁带设备、闪存存储器或其他类似的固态存储器设备,或设备阵列,包括存储区域网络中的设备或其他配置。计算机程序产品可在信息载体中有形地体现。计算机程序产品还可包含指令,该指令在被执行时,执行一种或多种方法,诸如上文所述的那些方法。信息载体是计算机可读介质或机器可读介质,诸如存储器604、存储设备608或处理器602上的存储器。

[0184] 高速控制器608管理计算设备600的带宽密集操作,而低速控制器612管理带宽较不密集操作。这种功能分配仅为示例。在一个具体实施中,高速控制器608例如通过图形处理器或加速器耦接到存储器604、显示器616,并且耦接到高速扩展端口610,该端口可接受各种扩展卡(未示出)。在该具体实施中,低速控制器612耦接到存储设备608和低速扩展端口614。低速扩展端口(该端口可包括各种通信端口,例如USB、蓝牙、以太网、无线以太网)可例如通过网络适配器耦接到一个或多个输入/输出设备,诸如键盘、指向设备、麦克风/扬声器对、扫描仪或联网设备诸如交换机或路由器。计算设备600可以多种不同的形式实现,如图所示。例如,该计算设备可被实现为标准服务器620,或者在一组此类服务器中被实现多次。它还可被实现为机架式服务器系统624的一部分。此外,该计算设备可在个人计算机诸如膝上型计算机622中实现。另选地,来自计算设备600的部件可与移动设备(未示出)诸如设备650中的其他部件组合。此类设备中的每个设备可包含计算设备600、650中的一个或多个设备,并且整个系统可由彼此通信的多个计算设备600、650组成。

[0185] 计算设备600可以多种不同的形式实现,如图所示。例如,该计算设备可被实现为标准服务器620,或者在一组此类服务器中被实现多次。它还可被实现为机架式服务器系统624的一部分。此外,该计算设备可在个人计算机诸如膝上型计算机622中实现。另选地,来自计算设备600的部件可与移动设备(未示出)诸如设备650中的其他部件组合。此类设备中的每个设备可包含计算设备600、650中的一个或多个设备,并且整个系统可由彼此通信的多个计算设备600、650组成。

[0186] 计算设备650包括处理器652、存储器664和输入/输出设备诸如显示器654、通信接口666和收发器668,以及其他部件。设备650还可设置有存储设备,诸如微驱动器或其他设备,以提供附加的存储装置。部件650、652、664、654、666和668中的每个部件使用各种总线互连,并且这些部件中的若干部件可安装在公共主板上或视情况以其他方式安装。

[0187] 处理器652可执行计算设备650内的指令,包括存储在存储器664中的指令。处理器可被实现为包括独立的多个模拟处理器和数字处理器的芯片的芯片组。另外,处理器可使用多种架构中的任一种架构来实现。例如,处理器610可以是CISC(复杂指令集计算机)处理器、RISC(精简指令集计算机)处理器或MISC(最小指令集计算机)处理器。处理器可提供例如设备650的其他部件的协调,诸如对用户接口的控制、由设备650运行的应用程序以及由设备650进行的无线通信。

[0188] 处理器652可通过耦接到显示器654的控制接口658和显示接口656与用户通信。显示器654可为例如TFT(薄膜晶体管液晶显示器)显示器或OLED(有机发光二极管)显示器或

其他适当的显示技术。显示接口656可包括用于驱动显示器654以向用户呈现图形和其他信息的适当电路。控制接口658可接收来自用户的命令并转换这些命令以提交给处理器652。此外,可提供与处理器652通信的外部接口662,以便实现设备650与其他设备的近距区域通信。外部接口662可例如在一些具体实施中提供有线通信,或在其他具体实施中提供无线通信,并且还可使用多个接口。

[0189] 存储器664将信息存储在计算设备650内。存储器664可被实现为一个或多个计算机可读介质、一个或多个易失性存储器单元,或一个或多个非易失性存储器单元中的一者或多者。还可提供扩展存储器674并通过扩展接口672将其连接到设备650,该扩展接口可包括例如SIMM(单列直插式存储器模块)卡接口。此类扩展存储器674可为设备650提供额外的存储空间,或者还可为设备650存储应用程序或其他信息。具体地,扩展存储器674可包括用于执行或补充上述过程的指令,并且还可包括安全信息。因此,例如,扩展存储器674可被提供作为用于设备650的安全模块,并且可被编程为具有实现设备650的安全使用的指令。此外,安全应用程序可经由SIMM卡连同附加信息一起提供,诸如将识别信息以不可破解的方式放置在SIMM卡上。

[0190] 存储器可包括例如闪存存储器和/或非易失性随机存取存储器(NVRAM)存储器,如下所述。在一个具体实施中,计算机程序产品在信息载体中有形地体现。计算机程序产品包含指令,该指令在被执行时,执行一种或多种方法,诸如上文所述的那些方法。信息载体是计算机可读介质或机器可读介质,诸如存储器664、扩展存储器674,或处理器652上的可通过例如收发器668或外部接口662接收的存储器。

[0191] 设备650可通过通信接口666进行无线通信,该通信接口在需要时可包括数字信号处理电路。通信接口666可以提供在各种模式或协议下的通信,诸如GSM语音呼叫、SMS、EMS或MMS消息传送、CDMA、TDMA、PDC、WCDMA、CDMA2000或GPRS等。此类通信可例如通过射频收发器668发生。此外,可发生近程通信,诸如使用蓝牙、Wi-Fi或其他此类收发器(未示出)。此外,GPS(全球定位系统)接收器模块670可向设备650提供附加的导航和位置相关的无线数据,该无线数据可由在设备650上运行的应用程序视情况使用。

[0192] 设备650还可使用音频编解码器660可听地通信,该音频编解码器可从用户接收口头信息并将其转换为可用的数字信息。音频编解码器660同样可诸如通过扬声器(例如在设备650的手持终端中)为用户生成可听声。此类声音可包括来自语音电话呼叫的声音,可包括录制的声音,例如语音消息、音乐文件等,并且还可包括由在设备650上操作的应用程序生成的声音。

[0193] 计算设备650可以多种不同的形式实现,如图所示。例如,该计算设备可被实现为蜂窝电话680。该计算设备还可被实现为智能电话682、个人数字助理或其他类似的移动设备的一部分。

[0194] 本文所述的系统和方法的各种具体实施可在数字电子电路、集成电路、特别设计的ASIC(专用集成电路)、计算机硬件、固件、软件和/或此类具体实施的组合中实现。这些各种具体实施可包括在一个或多个计算机程序中的具体实施,该一个或多个计算机程序能够在包括至少一个可编程处理器的可编程系统上执行和/或解释,该至少一个可编程处理器可以是专用或通用处理器,被耦接以从存储系统、至少一个输入设备和至少一个输出设备接收数据和指令以及将数据和指令发送到存储系统、至少一个输入设备和至少一个输出设

备。

[0195] 这些计算机程序(也称为程序、软件、软件应用程序或代码)包括用于可编程处理器的机器指令,并且可以高级程序化和/或面向对象的编程语言和/或以汇编语言/机器语言来实现。如本文所用,术语“机器可读介质”、“计算机可读介质”是指用于向可编程处理器提供机器指令和/或数据的任何计算机程序产品、装置和/或设备,例如磁盘、光盘、存储器、可编程逻辑设备(PLD),包括接收机器指令作为机器可读信号的机器可读介质。术语“机器可读信号”是指用于向可编程处理器提供机器指令和/或数据的任何信号。

[0196] 为了提供与用户的交互,本文所述的系统和技术可在计算机上实现,该计算机具有用于向用户显示信息的显示设备(例如CRT(阴极射线管)或LCD(液晶显示器)监视器),以及用户可用来向该计算机提供输入的键盘和指向设备(例如鼠标或轨迹球)。也可使用其他类型的设备来提供与用户的交互;例如,提供给用户的反馈可以是任何形式的感官反馈,例如视觉反馈、听觉反馈或触觉反馈;并且可以任何形式接收来自用户的输入,包括声音、语音或触觉输入。

[0197] 本文所述的系统和技术可在计算系统中实现,该计算系统包括后端部件(例如,作为数据服务器)或包括中间件部件(例如,应用程序服务器)或包括前端部件(例如,具有图形用户界面或Web浏览器的客户端计算机),用户可通过该计算系统与本文所述的系统和技术的具体实施进行交互,或者与此类后端部件、中间件部件或前端部件的任何组合进行交互。该系统的部件可通过数字数据通信的任何形式或介质(例如,通信网络)互连。通信网络的示例包括局域网(“LAN”)、广域网(“WAN”)和互联网。

[0198] 该计算系统可包括客户端和服务器。客户端和服务器通常彼此远离,并且通常通过通信网络进行交互。客户端和服务器的关系借助于在相应计算机上运行并彼此具有客户端-服务器关系的计算机程序而产生。

[0199] 实施例

[0200] 本公开在以下实施例中进一步描述,这些实施例不限制权利要求的范围。

[0201] 实施例1:在使用灵活种子延伸的系统与未使用灵活种子延伸的系统之间比较未映射读段的百分比

[0202] 在本实施例中,使用不同的核酸测序仪来对特定样品进行测序,这些核酸测序仪包括HiSeq® 2500测序仪、HiSeq® X测序仪和NovaSeq®测序仪。然后,使用DRAGEN™平台在利用本文所述的灵活种子延伸的情况下和不利用灵活种子延伸的情况下映射由每个相应测序仪产生的读段。一旦映射,计算机系统就确定由每个测序仪的每个映射操作产生的未映射读段的百分比。

[0203] DRAGEN™平台是在现场可编程门阵列(FPGA)的硬件电路中实现的映射和比对单元。如本文所述,DRAGEN™v7平台当前未利用灵活种子延伸,而DRAGEN™v8平台确实利用灵活种子延伸。虽然本文所用的DRAGEN™平台在FPGA中实现,但一般来讲,DRAGEN™平台也可在其他集成电路诸如专用集成电路(ASIC)中实现。

[0204] 具体地讲,HiSeq® 2500测序仪用于对“DNA_Nexus_hiseq2500”样品进行测序,HiSeq®X测序仪用于对“DNA_Nexus_hiseqX”样品进行测序,并且NovaSeq®测序仪用于对“DNA_Nexus_NovaSeq”样品、“NovaSeq_NA12878_rep1”样品、“NovaSeq_TruSeq-nano-550”

样品和“AWS_HG005_40x”样品进行测序。“AWS_HG005_40X”来自受试者HG005。所有其他样品来自受试者HG001。

[0205] 图7是显示数据的条形图700的图示,这些数据以使用如本文所述的灵活种子延伸方法的系统中未映射读段百分比的形式来表示测试结果,并与不使用灵活种子延伸方法的系统的测试结果进行比较。条形图700是测试结果710、720、730、740、750和760的图形表示,比较了对由不同Illumina,Inc.测序设备生成的基因组读段进行的映射操作的结果。

[0206] 在第一实施例中,测试结果710示出当HiSeq® 2500测序仪对“DNA_Nexus_hiseq2500”样品进行测序并且在映射期间利用如本文的一个或多个具体实施中所述的灵活种子延伸时出现的未映射读段的百分比710b显著小于当该HiSeq® 2500测序仪对该“DNA_Nexus_hiseq2500”样品进行测序而在映射期间在未利用如本文的一个或多个具体实施中所述的灵活种子延伸时出现的未映射读段的百分比710a。

[0207] 在第二实施例中,测试结果720示出当NovaSeq®测序仪对“DNA_Nexus_NovaSeq”样品进行测序并且在映射期间利用如本文的一个或多个具体实施中所述的灵活种子延伸时出现的未映射读段的百分比720b显著小于当该NovaSeq测序仪对该“DNA_Nexus_NovaSeq”样品进行测序而在映射期间未利用如本文的一个或多个具体实施中所述的灵活种子延伸时出现的未映射读段的百分比720a。

[0208] 在第三实施例中,测试结果730示出当HiSeq®X测序仪对“DNA_Nexus_hiseqX”样品进行测序并且在映射期间利用如本文的一个或多个具体实施中所述的灵活种子延伸时出现的未映射读段的百分比730b显著小于当该HiSeq®X测序仪对该“DNA_Nexus_hiseqX”样品进行测序而在映射期间在未利用如本文的一个或多个具体实施中所述的灵活种子延伸时出现的未映射读段的百分比730a。

[0209] 在第四实施例中,测试结果740示出当NovaSeq®测序仪对“NovaSeq_NA12878_rep1”样品进行测序并且在映射期间利用如本文的一个或多个具体实施中所述的灵活种子延伸时出现的未映射读段的百分比740b显著小于当该NovaSeq®测序仪对该“NovaSeq_NA12878_rep1”样品进行测序而在映射期间未利用如本文的一个或多个具体实施中所述的灵活种子延伸时出现的未映射读段的百分比740a。

[0210] 在第五实施例中,测试结果750示出当NovaSeq®测序仪对“NovaSeq_TruSeq-nano-550”样品进行测序并且在映射期间利用如本文的一个或多个具体实施中所述的灵活种子延伸时出现的未映射读段的百分比750b显著小于当该NovaSeq®测序仪对该“NovaSeq_TruSeq-nano-550”样品进行测序而在映射期间未利用如本文的一个或多个具体实施中所述的灵活种子延伸时出现的未映射读段的百分比750a。

[0211] 在第六实施例中,测试结果760示出当NovaSeq®测序仪对“AWS_HG005_40X”样品进行测序并且在映射期间利用如本文的一个或多个具体实施中所述的灵活种子延伸时出现的未映射读段的百分比760b显著小于当该NovaSeq®测序仪对该“AWS_HG005_40X”样品进行测序而在映射期间未利用如本文的一个或多个具体实施中所述的灵活种子延伸时出现的未映射读段的百分比760a。

[0212] 因此,与不生成或不使用本文所述的散列表的常规方法相比,使用本文所述的散

列表的灵活种子延伸的具体实施在减少未映射读段方面实现了显著的性能提升。

[0213] 实施例2:在使用灵活种子延伸的系统与未使用灵活种子延伸的系统之间比较读段映射准确率

[0214] 在该实施例2中,使用DRAGEN™平台将由核酸测序仪生成的读段映射到参考序列。每个DRAGEN™平台将同一组读段映射到同一个参考测序仪。一旦完成映射,计算机系统就为每个映射操作确定作为映射错误率的函数的读段映射准确率。

[0215] DRAGEN™平台是在现场可编程门阵列(FPGA)的硬件电路中实现的映射和比对单元。如本文所述,DRAGEN™v7平台当前未利用灵活种子延伸,而DRAGEN™v8平台和DRAGEN™v8 hi-effort平台确实利用灵活种子延伸。

[0216] 虽然本文所用的DRAGEN™平台在FPGA中实现,但一般来讲,DRAGEN™平台也可在其他集成电路诸如专用集成电路(ASIC)中实现。

[0217] DRAGEN™v8平台和DRAGEN™v8 hi-effort平台之间的差异在于启发法和其他参数的设置。DRAGEN™v8平台使用以下启发法:intvl-target-hits=32,intvl-max-hits=16和intvl-sample-hits=16。本文描述了这些启发法中的每个启发法。此外,DRAGEN™v8平台使用其他参数:max-hifreq-hits=16,rescue-hifreq=0和sw-extra-intvl=1。Max-hifreq-hits参数指示从种子延伸失败之前到达的匹配区间获得的随机样本匹配项的最大数量(例如,每次延伸失败一个样本,直到达到限值为止)。rescue-hifreq参数确定是否将昂贵的急救扫描操作用于仅通过来自匹配区间的随机样本发现的匹配项。急救扫描是用于搜索候选读段比对附近的可能的配对读段比对的方法。sw-extra-intvl参数确定将昂贵的Smith-Waterman比对用于通过存取最佳(“额外”)区间而找到的或通过对匹配区间随机采样而找到的匹配项的策略。在无间隙比对未被修剪时通常不使用Smith-Waterman,但在无间隙比对经修剪时可采用Smith-Waterman,具体取决于包括该设置的启发法。设置为“1”意指Smith-Waterman可用于来自额外/最佳匹配区间的候选项,所述额外/最佳匹配区间整体被存取而不是通过随机采样。设置为“2”意指Smith-Waterman还可用于来自对匹配区间的随机采样的候选项。设置为“0”将意指Smith-Waterman不应用于来自额外/最佳区间处理或来自对匹配区间的随机采样的候选项。

[0218] 另一方面,DRAGEN™v8 hi-effort平台使用以下启发法:intvl-target-hits=32,intvl-max-hits=64和intvl-sample-hits=48。此外,DRAGEN™v8 hi-effort平台使用其他参数:max-hifreq-hits=32,rescue-hifreq=0和sw-extra-intvl=2。因此,DRAGEN™v8 hi-effort平台具有比DRAGEN™v8平台更宽泛的一组启发法。

[0219] 图8是显示数据的折线图800的图示,这些数据以使用如本文所公开的灵活种子延伸方法的系统中读段映射准确率的形式来表示测试结果,并与不使用灵活种子延伸方法的系统的测试结果进行比较。具体地讲,折线图800使用接受者操作特性(“ROC”)曲线(或折线)形式的准确率曲线来示出当使用置信度度量对数据进行分层时假阳性和假阴性之间的权衡。在图8的图示中,更靠近折线图800的顶壁和左壁的曲线(或折线)意味着更好的读段映射准确率。

[0220] 描绘了曲线810,它表示在映射期间未使用如本文的一个或多个具体实施中所述的灵活种子延伸的DRAGEN™v7平台的读段映射准确率。描绘了曲线820,它表示在映射期间使用如本文的一个或多个具体实施中所述的灵活种子延伸的DRAGEN™v8平台的读段映射准

确率。曲线810和曲线820的比较显示曲线820比曲线810更靠近顶壁和左壁。因此,仅通过以某种容量实现如本文的一个或多个具体实施中所述的灵活种子延伸就实现了读段映射准确率的改进。

[0221] 图8还描绘了曲线830,它表示DRAGEN™v8 hi-effort具体实施。与DRAGEN™v8具体实施类似,DRAGEN™v8 hi-effort具体实施也在映射期间采用如本文所述的灵活种子延伸方法。然而,如上所述,由DRAGEN™v8 hi-effort具体实施采用的启发法比用于采用DRAGEN™v8具体实施(其性能由820曲线表示)的启发法更宽泛。为DRAGEN™的“hi-effort”v8版本分配参数(例如,sw-extra-intvl=2),该参数相对于DRAGEN™v8具体实施(例如,sw-extra-intvl=1)增加了在下游执行更多Smith-Waterman比对工作的意向。如图8所示,曲线830比曲线810和820两者更靠近顶壁和左壁,从而表现出DRAGEN™v8 hi-effort具体实现在读段映射准确率方面显著的性能增进。

[0222] 图8还示出了曲线840,该曲线表示由BWA-MEM软件映射工具实现的读段映射准确率。BWA-MEM软件映射工具使用参考基因组的Burrows-Wheeler变换(BWT)作为其索引。这种表示参考基因组的方法可固有地提供与灵活种子延伸所提供的相似的有益效果,诸如检索对应于任何匹配长度的匹配项完整集合的能力。如曲线830和曲线840所示,DRAGEN™v8“hi-effort”具体实施可实现与基于软件的BWA软件映射工具相同的读段映射准确率。因此,重要的是,DRAGEN™v8“hi-effort”能够实现与基于软件的BWA映射工具相当的读段映射准确率水平,因为DRAGEN™v8“hi-effort”还利用了DRAGEN™平台的其他益处,包括例如对映射种子的较少存储器存取。然而,在实现本文所述的基于硬件的灵活种子延伸之前,DRAGEN™平台能够实现与BWA软件映射工具实现的读段映射准确率相同的读段映射准确率。

[0223] 因此,与不生成或不使用本文所述的散列表的常规方法相比,使用本文所述的散列表的灵活种子延伸的具体实施在读段映射准确率方面实现了显著的性能提升。

[0224] 其他实施方案

[0225] 已经描述了多个实施方案。然而,应当理解,在不脱离本发明的实质和范围的情况下,可进行各种修改。此外,附图中所示的逻辑流程不需要所示的特定顺序或有序顺序来实现所需的结果。此外,可在所述流程中提供其他步骤,或者可消除步骤,并且可将其他部件添加到所述系统或从所述系统中移除。因此,其他实施方案也在以下权利要求书的范围内。

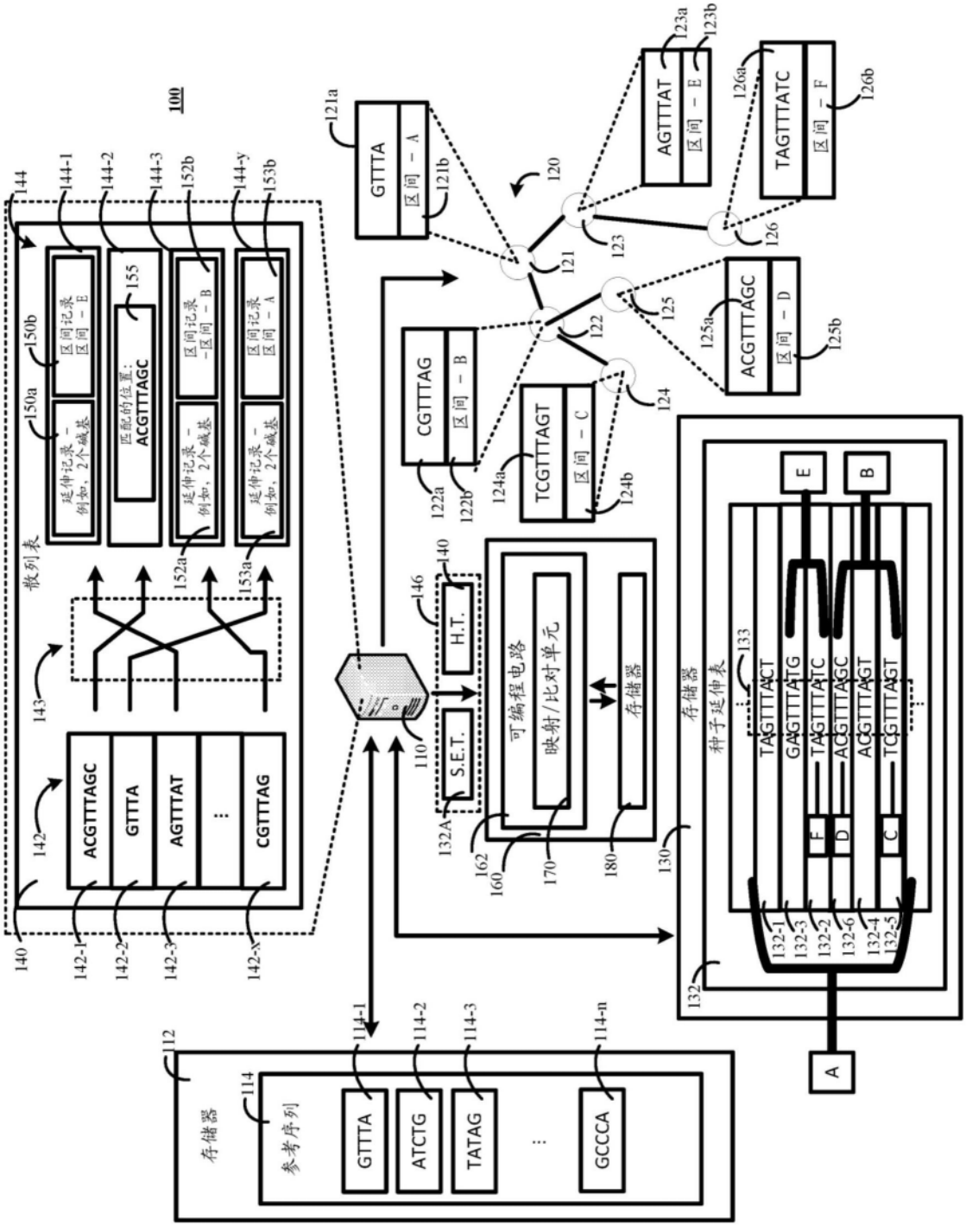

100
散列表
ACGTTTAGC
GTTTA
AGTTTAT
CGTTTAG
匹配的位置:
ACGTTTAGC
区间记录
区间 – E
区间记录
区间 – B
区间记录
区间 – A
延伸记录 –
例如，2个碱基
GTTTA
区间 – A
AGTTTAT
区间 – E
TAGTTTATC
区间 – F
CGTTTAG
区间 – B
TCGTTTAGT
区间 – C
ACGTTTAGC
区间 – D
S.E.T.
H.T.
可编程电路
映射/比对单元
存储器
存储器
种子延伸表
TAGTTTACT
GAGTTTATG
TAGTTTATC
ACGTTTAGC
ACGTTTAGT
TCGTTTAGT
E
B
F
D
C
A
存储器
参考序列
GTTTA
ATCTG
TATAG
GCCCA

图1

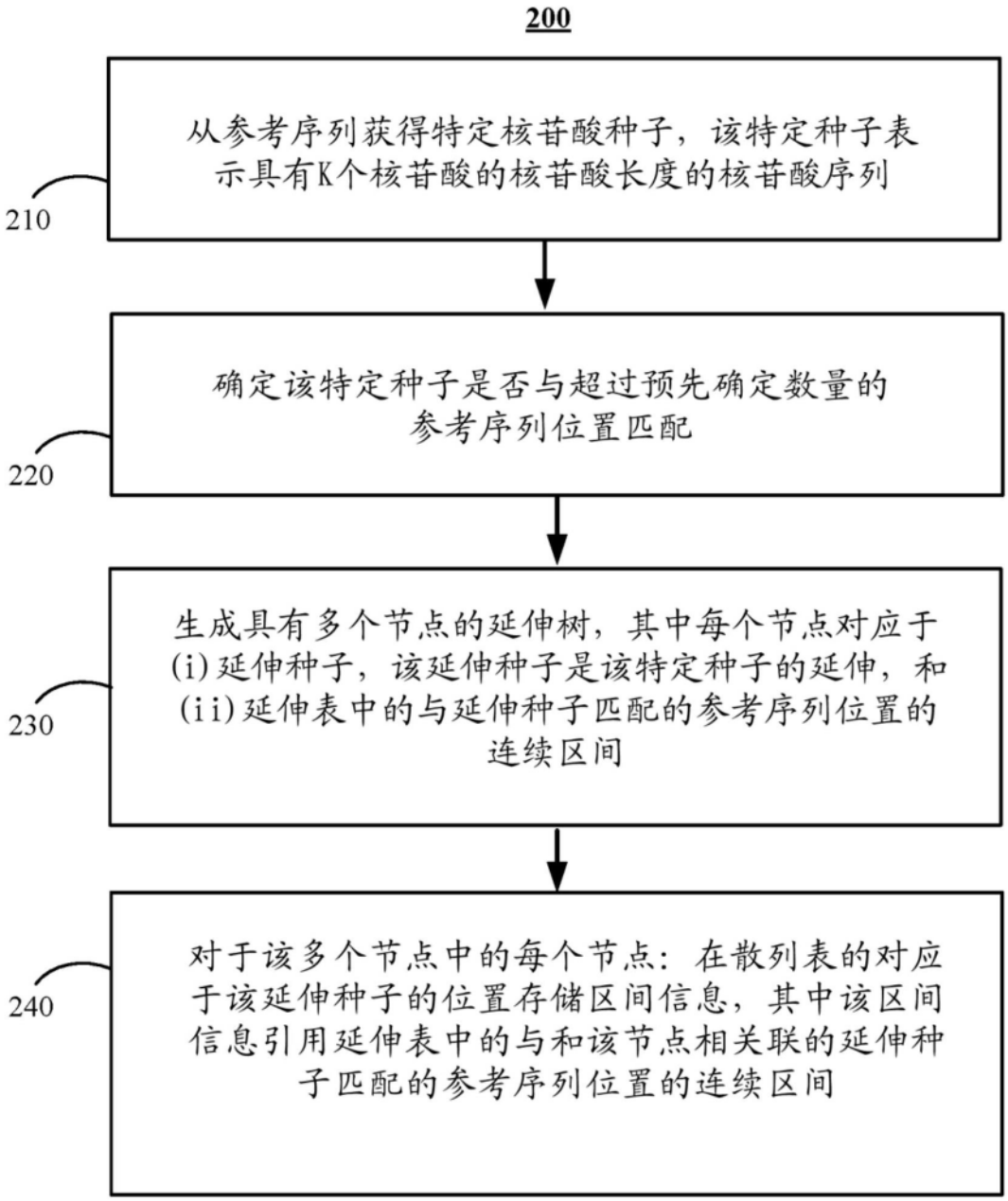

200
从参考序列获得特定核苷酸种子，该特定种子表示具有K个核苷酸的核苷酸长度的核苷酸序列
210
确定该特定种子是否与超过预先确定数量的参考序列位置匹配
220
生成具有多个节点的延伸树，其中每个节点对应于(i)延伸种子，该延伸种子是该特定种子的延伸，和(ii)延伸表中的与延伸种子匹配的参考序列位置的连续区间
230
对于该多个节点中的每个节点：在散列表的对应于该延伸种子的位置存储区间信息，其中该区间信息引用延伸表中的与和该节点相关联的延伸种子匹配的参考序列位置的连续区间
240

图2

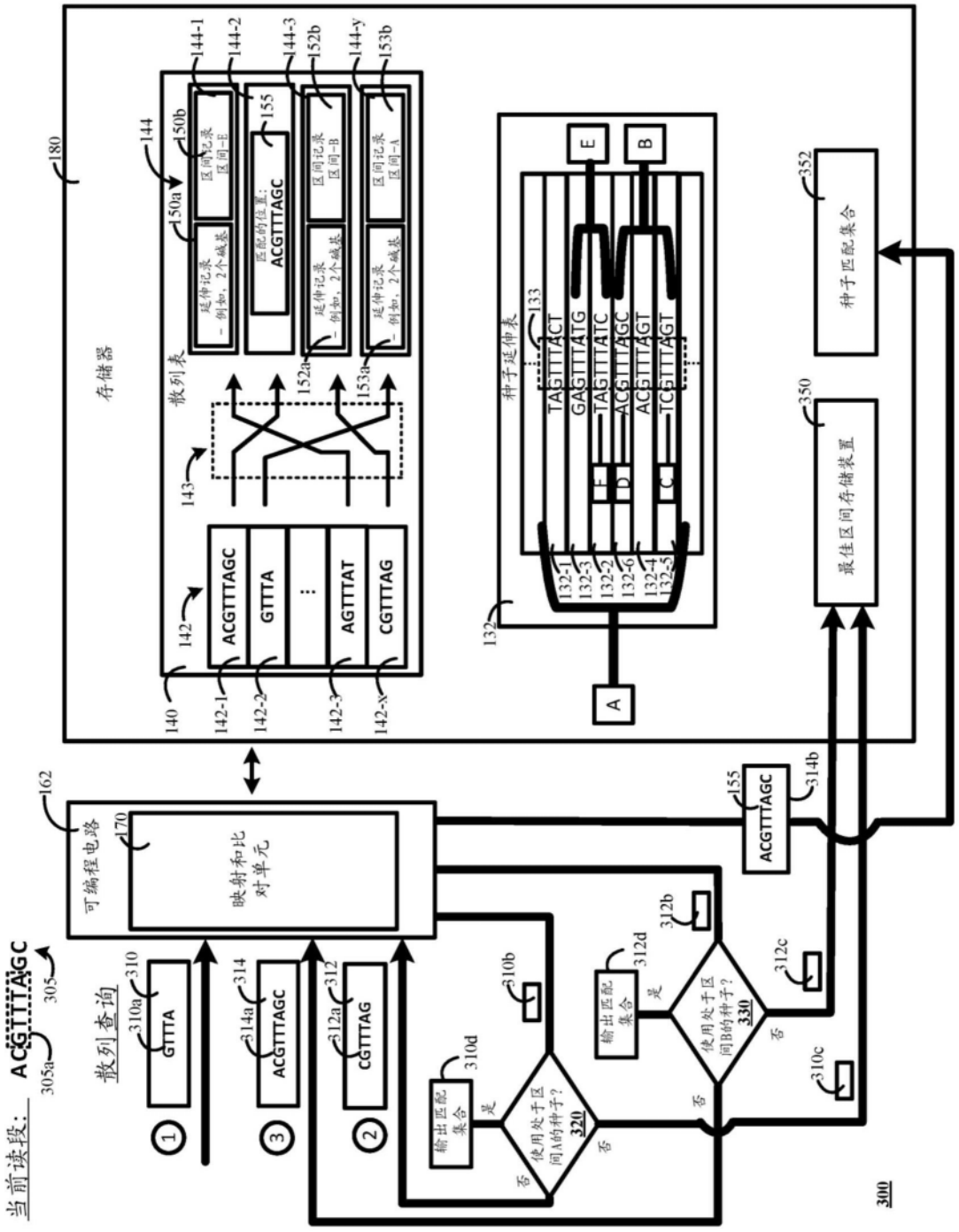

当前读段：
ACGTTTAGC
305a
305
散列查询
310a
GTTTA
314a
ACGTTTAGC
312a
CGTTTAG
可编程电路
170
映射和比对单元
162
存储器
180
散列表
144
150a
150b
延伸记录 - 例如，2个碱基
区间记录 区间-E
区间记录 区间-B
区间记录 区间-A
匹配的位置：
ACGTTTAGC
155
144-1
144-2
144-3
152b
144-y
153b
152a
153a
143
142
ACGTTTAGC
GTTTA
AGTTTAT
CGTTTAG
140
142-1
142-2
142-3
142-x
种子延伸表
133
TAGTTTACT
GAGTTTATG
TAGTTTATC
ACGTTTAGC
ACGTTTAGT
TCGTTTAGT
132
132-1
132-3
132-2
132-6
132-4
132-5
A
B
C
D
E
F
种子匹配集合
352
最佳区间存储装置
350
ACGTTTAGC
314b
310b
310c
310d
312b
312c
312d
输出匹配集合
使用处于区间A的种子？
320
使用处于区间B的种子？
330
是
否
300

图3

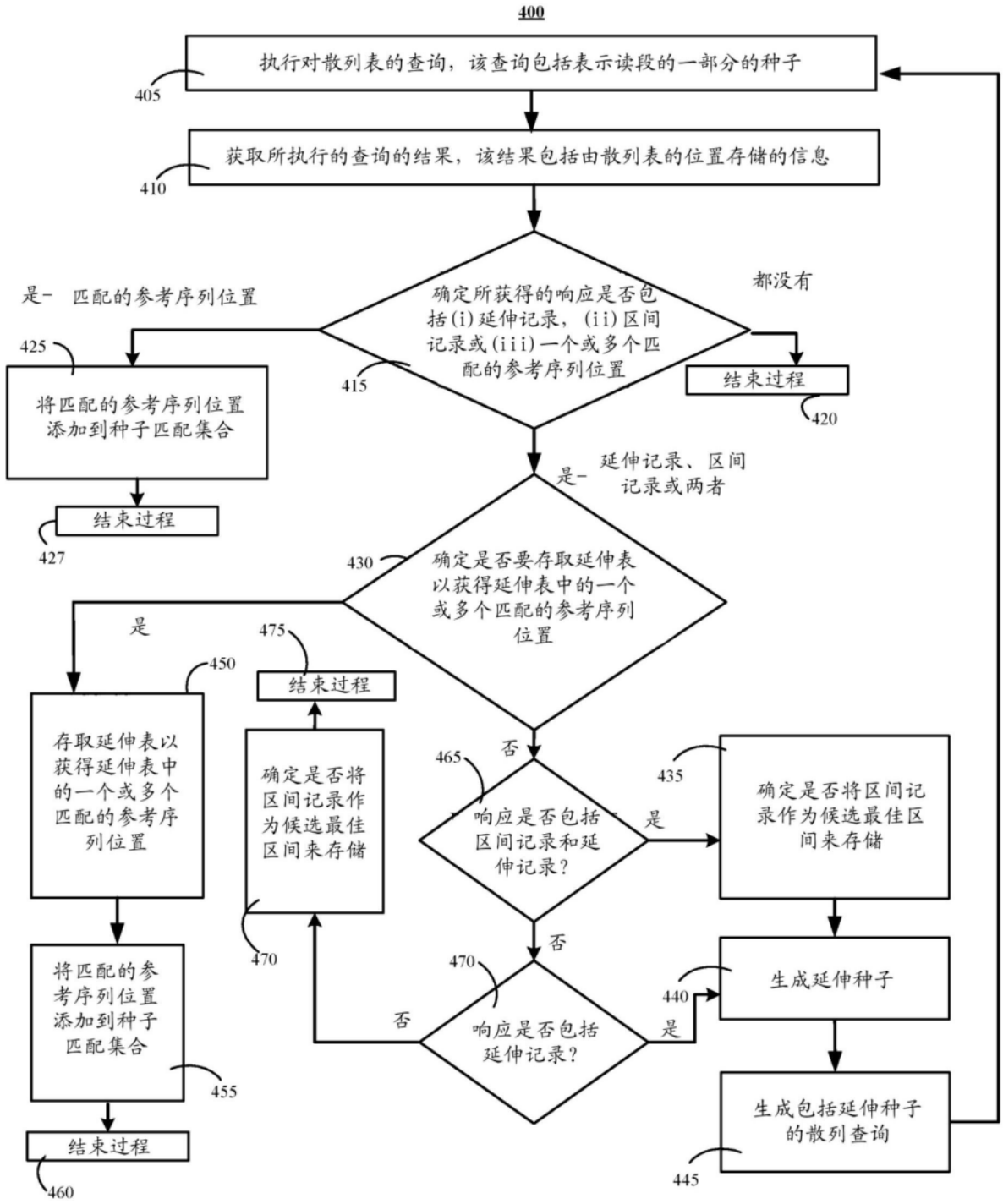

400
执行对散列表的查询，该查询包括表示读段的一部分的种子
405
获取所执行的查询的结果，该结果包括由散列表的位置存储的信息
410
确定所获得的响应是否包括(i)延伸记录，(ii)区间记录或(iii)一个或多个匹配的参考序列位置
415
是- 匹配的参考序列位置
都没有
将匹配的参考序列位置添加到种子匹配集合
425
结束过程
420
结束过程
427
是- 延伸记录、区间记录或两者
确定是否要存取延伸表以获得延伸表中的一个或多个匹配的参考序列位置
430
是
否
存取延伸表以获得延伸表中的一个或多个匹配的参考序列位置
450
结束过程
475
确定是否将区间记录作为候选最佳区间来存储
470
响应是否包括区间记录和延伸记录?
465
是
确定是否将区间记录作为候选最佳区间来存储
435
否
将匹配的参考序列位置添加到种子匹配集合
455
结束过程
460
响应是否包括延伸记录?
470
否
是
生成延伸种子
440
生成包括延伸种子的散列查询
445

图4

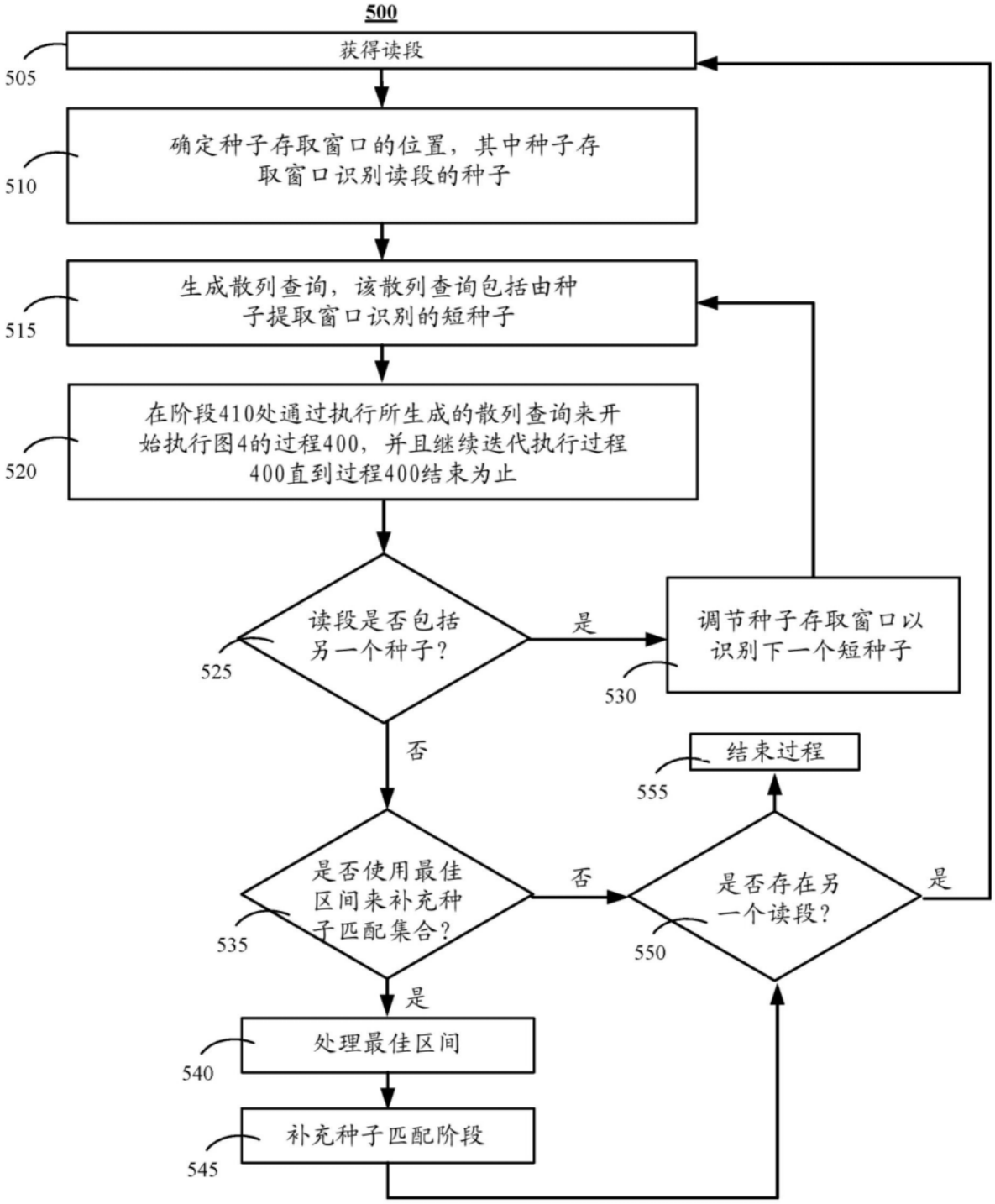

500
获得读段
505
确定种子存取窗口的位置，其中种子存取窗口识别读段的种子
510
生成散列查询，该散列查询包括由种子提取窗口识别的短种子
515
在阶段410处通过执行所生成的散列查询来开始执行图4的过程400，并且继续迭代执行过程400直到过程400结束为止
520
读段是否包括另一个种子？
525
是
调节种子存取窗口以识别下一个短种子
530
否
结束过程
555
是否使用最佳区间来补充种子匹配集合？
535
否
是否存在另一个读段？
550
是
是
处理最佳区间
540
补充种子匹配阶段
545

图5

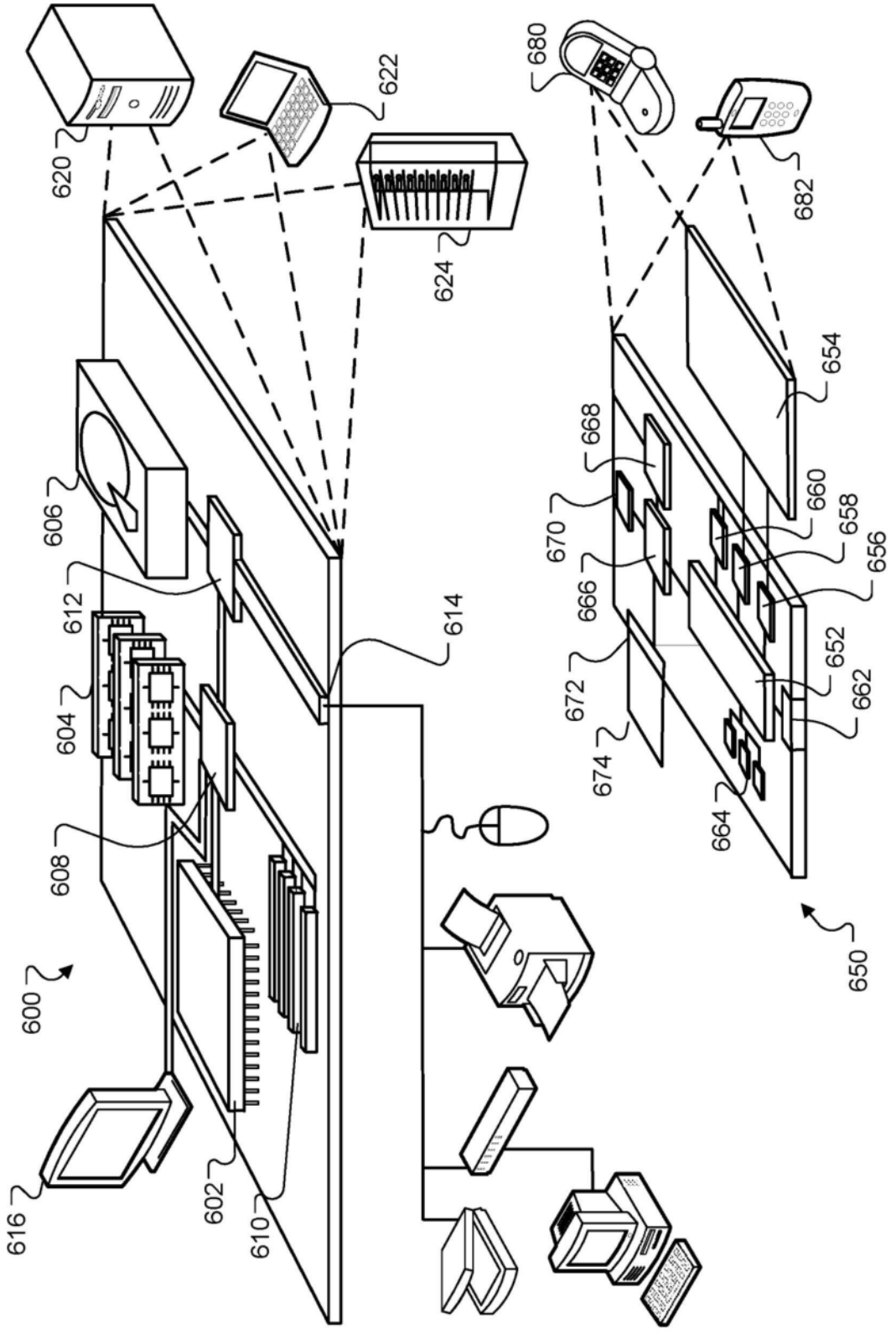

620
622
624
680
682
606
612
614
604
608
600
616
602
610
654
668
660
658
656
670
666
652
672
662
674
664
650

图6

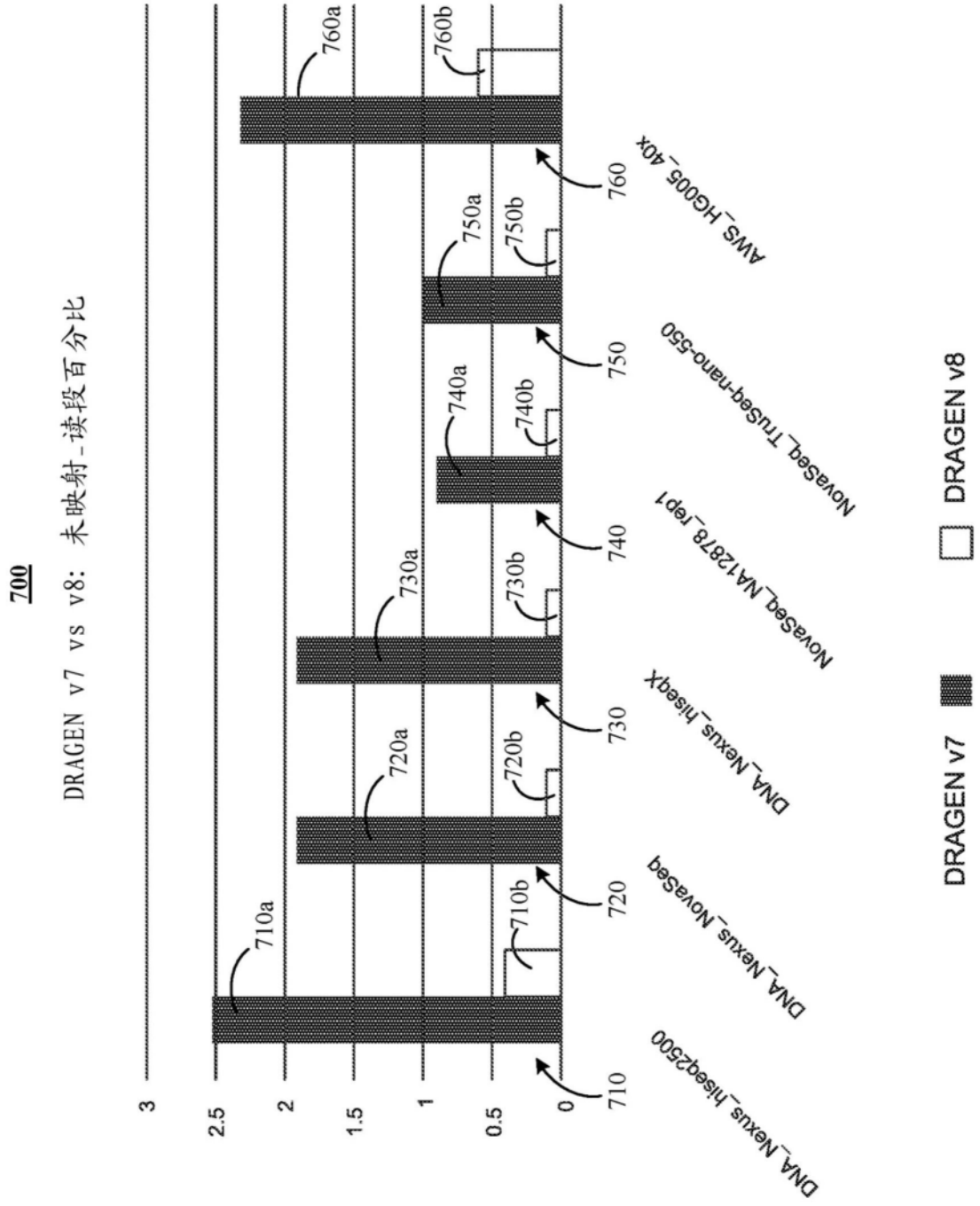

700
DRAGEN v7 vs v8: 未映射-读段百分比
3
2.5
2
1.5
1
0.5
0
710
720
730
740
750
760
710a
710b
720a
720b
730a
730b
740a
740b
750a
750b
760a
760b
DNA_Nexus_hiseq2500
DNA_Nexus_NovaSeq
DNA_Nexus_hiseqx
NovaSeq_NA12878_rep1
NovaSeq_TruSeq-nano-550
AWS_HG005_40x
DRAGEN v7
DRAGEN v8

图7

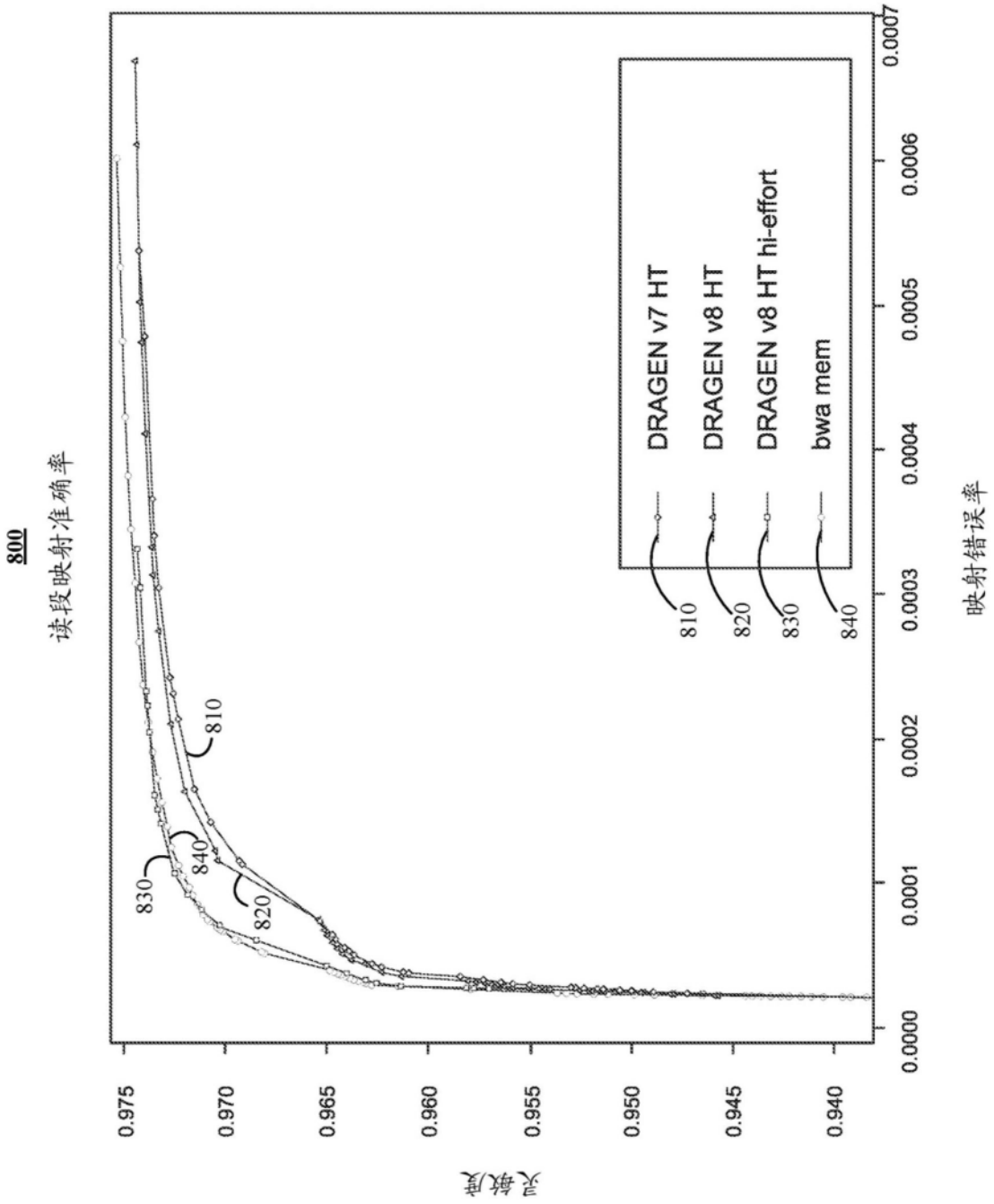

800
读段映射准确率
0.975
0.970
0.965
0.960
0.955
0.950
0.945
0.940
0.0000
0.0001
0.0002
0.0003
0.0004
0.0005
0.0006
0.0007
映射错误率
DRAGEN v7 HT
DRAGEN v8 HT
DRAGEN v8 HT hi-effort
bwa mem
810
820
830
840

图8