(19)国家知识产权局

(12)发明专利

(10)授权公告号 CN 111108218 B
(45)授权公告日 2024.06.25

(21)申请号 201880060283.5
(22)申请日 2018.09.20
(65)同一申请的已公布的文献号
申请公布号 CN 111108218 A
(43)申请公布日 2020.05.05
(30)优先权数据
62/560,745 2017.09.20 US
(85)PCT国际申请进入国家阶段日
2020.03.17
(86)PCT国际申请的申请数据
PCT/US2018/051872 2018.09.20
(87)PCT国际申请的公布数据
WO2019/060494 EN 2019.03.28
(73)专利权人 生命科技股份有限公司
地址 美国加利福尼亚州
(72)发明人 R·哥蒂姆卡拉 C-Z·白
D·布林扎 J·沙格曼
V·巴盖伊
(74)专利代理机构 上海专利商标事务所有限公司 31100
专利代理师 周全 陈洁
(51)Int.Cl.
C12Q 1/6853(2018.01)
G16B 30/10(2019.01)
(56)对比文件
Miguel Alcaide等.Targeted error-suppressed quantification of circulating tumor DNA using semi-degenerate barcoded adapters and biotinylated baits.《Scientific Reports》.2017,第7卷第13页第2-5段,第14页第1-3段,第16页第1-4段,图1.
Fredrik Lysholm等.FAAST: Flow-space Assisted Alignment Search Tool.《BMC Bioinformatics》.2011,第12卷摘要,第 5 页左栏第3段,表 1.
审查员 谢庆宁

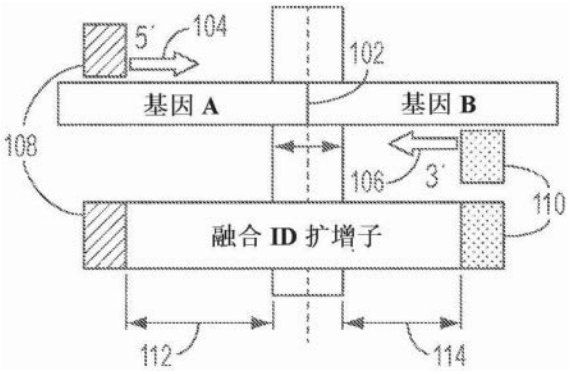
权利要求书3页 说明书21页 附图6页

(54)发明名称

使用压缩的分子标记的核酸序列数据检测融合的方法

(57)摘要

一种用于压缩核酸序列数据的方法,其中每个序列读数与分子标签序列相关,其中一部分序列读数比对对应于映射到靶向融合参考序列的序列读数,所述方法包括基于对应于序列读数家族的流动空间信号测量值确定每个序列读数家族的共有序列读数,确定每个序列读数家族的共有序列比对,其中所述共有序列比对的一部分对应于与所述靶向融合参考序列比对的共有序列读数,产生包含共有压缩数据的压缩数据结构,所述共有压缩数据包括每个家族的所述共有序列读数和所述共有序列比对,以及使用来自所述压缩数据结构的所述共有序列读数和所述共有序列比对,来检测融合。



CN 111108218 B

1.一种包括指令的非暂时性机器可读存储介质,当由处理器执行时,所述指令引起所述处理器执行一种使用压缩分子标记的核酸序列数据检测基因融合的方法,所述方法包括:

接收多个核酸序列读数和多个序列读数家族的多个序列比对,其中每个序列读数与分子标签序列连接,所述分子标签序列鉴定由核酸样品中的特定多核苷酸分子产生的序列读数家族,每个家族具有多个序列读数,其中所述序列比对的一部分对应于映射到靶向融合参考序列的序列读数,其中所述多个核酸序列读数通过在引物的存在下扩增所述核酸样品以产生多个扩增子来生成,所述引物包括5'引物和3'引物,所述5'引物和3'引物位于与基因融合相关的断点区域侧翼,其中所述5'引物附加有前缀标签并且所述3'引物附加有后缀标签,所述前缀标签和所述后缀标签包含用于所述核酸样品中的多核苷酸分子的独特分子标签,并且对所述多个扩增子进行测序以生成所述多个核酸序列读数,其中通过将所述核酸序列读数映射到参考序列以产生所述多个序列比对,所述参考序列包括所述靶向融合参考序列;

基于对应于所述家族的所述序列读数的流动空间信号测量值的向量,确定每个序列读数家族的共有序列读数,包括:

计算所述流动空间信号测量值的向量的算术平均值以形成所述家族的共有流动空间信号测量值的向量;

计算所述流动空间信号测量值的向量的标准偏差以形成所述家族的标准偏差向量;

基于所述家族的所述共有流动空间信号测量值的向量确定所述家族的共有序列读数;

确定每个序列读数家族的共有序列比对,包括:选择与所述序列读数家族相对应的具有最高映射质量的序列比对;以及将对应于具有所述最高映射质量的序列比对的序列读数与所述家族的共有序列读数进行比较,其中所述共有序列比对的一部分对应于与所述靶向融合参考序列比对的所述共有序列读数;

产生包含共有压缩数据的压缩数据结构,每个家族的所述共有压缩数据包括所述共有序列读数、所述共有序列比对、所述共有流动空间信号测量值的向量和所述标准偏差向量,所述压缩数据结构的数据量小于原始的多个序列读数和多个序列比对数据的数据量;以及

使用来自所述压缩数据结构的所述共有序列读数和所述共有序列比对,来检测基因融合,所述检测基因融合还包括:基于共有序列读数与靶向融合参考序列的共有序列比对的特征,鉴定合格的共有序列读数;以及确定对应于与靶向融合参考序列比对的所述合格的共有序列读数的家族的数目是否大于或等于最小分子计数阈值。

2.根据权利要求1所述的非暂时性机器可读存储介质,其中所述序列读数由双向测序产生,其中正向共有序列读数和反向共有序列读数在单独的家族中,包括:正向家族,其中所述正向家族中的每个序列读数与第一前缀标签和第一后缀标签连接;和反向家族,其中所述反向家族中的每个序列读书与第二前缀标签和第二后缀标签连接,所述方法还包括当所述第二前缀标签和第二后缀标签的反向互补序列匹配所述第一前缀标签和所述第一后缀标签时,组合所述正向家族和所述反向家族,以形成具有所述压缩数据结构的一个共有序列读数的组合家族。

3.根据权利要求1所述的非暂时性机器可读存储介质,其中所述特征包括同源性特征、映射质量特征和跨断点特征。

4.根据权利要求1所述的非暂时性机器可读存储介质,其中所述鉴定合格的共有序列读数还包括确定与所述靶向融合参考序列比对的所述共有序列读数是否跨越所述靶向融合参考序列的融合断点。

5.根据权利要求1所述的非暂时性机器可读存储介质,其中所述鉴定合格的共有序列读数还包括确定所述共有序列读数分别与所述靶向融合参考序列的第一和第二配偶体序列的第一和第二同源性水平是否大于或等于最小同源性阈值,其中所述配偶体序列是靶向融合参考序列在融合端点一侧的部分。

6.根据权利要求1所述的非暂时性机器可读存储介质,其中所述鉴定合格的共有序列读数还包括确定分别在靶向融合参考序列的第一和第二配偶体序列内的所述共有序列读数的第一和第二映射质量值是否大于或等于映射质量阈值。

7.根据权利要求6所述的非暂时性机器可读存储介质,其中鉴定合格的共有序列读数还包括通过计算与所述配偶体序列匹配的所述共有序列读数中的匹配碱基的数目和与所述配偶体序列重叠的所述共有序列读数中的重叠碱基的数目的比率,来确定映射质量值。

8.根据权利要求1所述的非暂时性机器可读存储介质,其中所述检测基因融合还包括确定读数计数是否大于或等于最小读数计数阈值,其中所述读数计数为对应于与所述靶向融合参考序列比对的所述合格的共有序列读数的所述家族的序列读数的数目的总和。

9.根据权利要求1所述的非暂时性机器可读存储介质,其中所述序列比对的第二部分对应于映射到对照基因参考序列的序列读数,其中所述共有压缩数据还包括对应于所述对照基因参考序列的共有序列读数和共有序列比对。

10.根据权利要求9所述的非暂时性机器可读存储介质,还包括当家族计数大于最小分子计数阈值并且读数计数大于读数计数阈值时,确定对应于所述对照基因参考序列的过程控制靶标的存在,其中所述家族计数为对应于与所述对照基因参考序列比对的所述共有序列读数的家族的数目,并且所述读数计数为所述对应家族的序列读数数目的总和。

11.根据权利要求1所述的非暂时性机器可读存储介质,其中所述基因融合包括基因间融合,并且所述靶向融合参考序列包括用于在融合断点处融合两个基因的参考序列。

12.根据权利要求1所述的非暂时性机器可读存储介质,其中所述基因融合包括基因内融合,并且所述靶向融合参考序列包括用于在相同基因内的融合断点处融合两个外显子的参考序列。

13.根据权利要求12所述的非暂时性机器可读存储介质,其中所述共有序列比对的第二部分对应于与所述相同基因的一个或多个野生型参考序列比对的共有序列读数。

14.根据权利要求13所述的非暂时性机器可读存储介质,其中所述检测基因融合还包括计算所述基因内融合的读数计数与平均读数计数的比率,所述平均读数计数对应于与所述相同基因的所述野生型参考序列比对的所述共有序列读数。

15.根据权利要求13所述的非暂时性机器可读存储介质,其中所述检测基因融合还包括计算所述基因内融合的读数计数与读数计数总和的比率,所述读数计数总和对应于与所述野生型参考序列比对的所述共有序列读数和与所述相同基因的所述靶向融合参考序列比对的所述共有序列读数。

16.根据权利要求1所述的非暂时性机器可读存储介质,其中所述共有序列读数部分地映射到所述靶向融合参考序列,其中检测基因融合还包括基于部分映射的共有序列读数检

测非靶向融合。

17.一种使用压缩分子标记的核酸序列数据检测基因融合的系统,所述系统包括:

机器可读存储器;以及

与所述存储器通信的处理器,其中所述处理器被配置成执行机器可读指令,当由所述处理器执行时,所述机器可读指令引起所述系统执行一种方法,所述方法包括:

接收多个核酸序列读数和多个序列读数家族的多个序列比对,其中每个序列读数与分子标签序列连接,所述分子标签序列鉴定由核酸样品中的特定多核苷酸分子产生的序列读数家族,每个家族具有多个序列读数,其中所述序列比对的一部分对应于映射到靶向融合参考序列的序列读数,其中所述多个核酸序列读数通过在引物的存在下扩增所述核酸样品以产生多个扩增子来生成,所述引物包括5'引物和3'引物,所述5'引物和3'引物位于与基因融合相关的断点区域侧翼,其中所述5'引物附加有前缀标签并且所述3'引物附加有后缀标签,所述前缀标签和所述后缀标签包含用于所述核酸样品中的多核苷酸分子的独特分子标签,并且对所述多个扩增子进行测序以生成所述多个核酸序列读数,其中通过将所述核酸序列读数映射到参考序列以产生所述多个序列比对,所述参考序列包括所述靶向融合参考序列;

基于对应于所述家族的所述序列读数的流动空间信号测量值的向量,确定每个序列读数家族的共有序列读数,包括:

计算所述流动空间信号测量值的向量的算术平均值以形成所述家族的共有流动空间信号测量值的向量;

计算所述流动空间信号测量值的向量的标准偏差以形成所述家族的标准偏差向量;

基于所述家族的所述共有流动空间信号测量值的向量确定所述家族的共有序列读数;

确定每个序列读数家族的共有序列比对,包括:选择与所述序列读数家族相对应的具有最高映射质量的序列比对;以及将对应于具有所述最高映射质量的序列比对的序列读数与所述家族的共有序列读数进行比较,其中所述共有序列比对的一部分对应于与所述靶向融合参考序列比对的所述共有序列读数;

产生包含共有压缩数据的压缩数据结构,每个家族的所述共有压缩数据包括所述共有序列读数、所述共有序列比对、所述共有流动空间信号测量值的向量和所述标准偏差向量,所述压缩数据结构的数据量小于原始的多个序列读数和多个序列比对数据的数据量;以及

使用来自所述压缩数据结构的所述共有序列读数和所述共有序列比对,来检测基因融合,所述检测基因融合还包括:基于共有序列读数与靶向融合参考序列的共有序列比对的特征,鉴定合格的共有序列读数;以及确定对应于与靶向融合参考序列比对的所述合格的共有序列读数的家族的数目是否大于或等于最小分子计数阈值。

使用压缩的分子标记的核酸序列数据检测融合的方法

[0001] 交叉引用

[0002] 根据美国法典35U.S.C.§119(e),本申请要求于2017年9月20日提交的美国临时申请号62/560,745的权益。前述申请的全部内容通过引用并入本文。

发明内容

[0003] 结构变体,例如大规模缺失、插入、倒位、基因组重排、基因融合等,可以与各种遗传病症和癌症相关。结构变体通常可以导致对细胞的正常功能必需的蛋白质生产的显著破坏。例如,基因组重排和基因融合可以导致编码嵌合蛋白的mRNA,所述嵌合蛋白具有来自一种蛋白的第一部分和来自另一种蛋白的第二部分。通常,这些嵌合蛋白不再像第一蛋白或第二蛋白那样起作用,并且可导致规则性途径的破坏。在癌细胞中,被破坏的调节途径可能涉及细胞凋亡、细胞生长等的调节,并且由于基因融合,使得癌细胞能够不受抑制地生长。

[0004] 核酸序列的分子标记可用于鉴定源自相同多核苷酸分子(例如源自无细胞DNA(cfDNA)样品)的核酸序列读数,并基于其标签序列将其分类为家族。可以储存和处理使用各种技术、平台或科技从核酸样品得到的大量分子标记的核酸序列数据,用于检测融合。需要新的方法、系统和计算机可读介质,其压缩分子标记的核酸序列数据以便减少存储的存储器需求,并检测压缩的分子标记的核酸序列数据中的融合,包括从cfDNA样品所得到的。

[0005] 根据一种示例性实施方式,提供了一种用于压缩分子标记的核酸序列数据以供融合检测的方法,所述方法包括:(a)接收多个核酸序列读数和多个序列读数家族的多个序列比对,其中每个序列读数与分子标签序列相关,所述分子标签序列鉴定由核酸样品中的特定多核苷酸分子产生的序列读数家族,每个家族具有多个序列读数,其中所述序列比对的一部分对应于映射到靶向融合参考序列的序列读数;(b)基于对应于所述家族的所述序列读数的流动空间信号测量值,确定每个序列读数家族的共有序列读数;(c)确定每个序列读数家族的共有序列比对,其中所述共有序列比对的一部分对应于与所述靶向融合参考序列比对的所述共有序列读数;(d)产生包含共有压缩数据的压缩数据结构,所述共有压缩数据包括每个家族的所述共有序列读数和所述共有序列比对;以及(e)使用来自所述压缩数据结构的所述共有序列读数和所述共有序列比对,来检测融合。

[0006] 根据一种示例性实施方式,提供了一种包括指令的非暂时性机器可读存储介质,当由处理器执行时,所述指令引起所述处理器执行一种用于压缩分子标记的核酸序列数据以供融合检测的方法,所述方法包括:(a)接收多个核酸序列读数和多个序列读数家族的多个序列比对,其中每个序列读数与分子标签序列相关,所述分子标签序列鉴定由核酸样品中的特定多核苷酸分子产生的序列读数家族,每个家族具有多个序列读数,其中所述序列比对的一部分对应于映射到靶向融合参考序列的序列读数;(b)基于对应于所述家族的所述序列读数的流动空间信号测量值,确定每个序列读数家族的共有序列读数;(c)确定每个序列读数家族的共有序列比对,其中所述共有序列比对的一部分对应于与所述靶向融合参考序列比对的所述共有序列读数;(d)产生包含共有压缩数据的压缩数据结构,所述共有压缩数据包括每个家族的所述共有序列读数和所述共有序列比对;以及(e)使用来自所述压

缩数据结构的所述共有序列读数和所述共有序列比对,来检测融合。

[0007] 根据一种示例性实施方式,提供了一种用于压缩分子标记的核酸序列数据以供融合检测的系统,所述系统包括:机器可读存储器和与所述存储器通信的处理器,其中所述处理器被配置成执行机器可读指令,当由处理器执行时,所述指令引起所述系统执行一种方法,所述方法包括:(a)接收多个核酸序列读数和多个序列读数家族的多个序列比对,其中每个序列读数与分子标签序列相关,所述分子标签序列鉴定由核酸样品中的特定多核苷酸分子产生的序列读数家族,每个家族具有多个序列读数,其中所述序列比对的一部分对应于映射到靶向融合参考序列的序列读数;(b)基于对应于所述家族的所述序列读数的流动空间信号测量值,确定每个序列读数家族的共有序列读数;(c)确定每个序列读数家族的共有序列比对,其中所述共有序列比对的一部分对应于与所述靶向融合参考序列比对的所述共有序列读数;(d)产生包含共有压缩数据的压缩数据结构,所述共有压缩数据包括每个家族的所述共有序列读数和所述共有序列比对;以及(e)使用来自所述压缩数据结构的所述共有序列读数和所述共有序列比对,来检测融合。

附图说明

[0008] 本发明的新颖特征在所附权利要求中具体阐述。通过参考以下详细描述和附图,将获得对本发明的特征和优点的更好理解,以下详细描述阐述了利用本发明的原理的说明性实施方式,附图中:

[0009] 图1说明根据一种实施方式的用于检测融合的引物设计的示例,其中引物具有分子标签。

[0010] 图2为根据一种实施方式的用于产生用于检测融合的共有压缩数据的示例性方法的框图。

[0011] 图3为根据一种实施方式的用于流动空间共有管线的示例性方法的框图。

[0012] 图4显示可以从其进行碱基调用(base calls)的流动空间信号测量值的示例性表示。

[0013] 图5说明单个家族的流动空间信号测量值的示例性绘图。

[0014] 图6说明单个家族的共有流动空间信号测量值的示例性绘图。

[0015] 图7为根据一种实施方式的使用共有压缩数据进行融合调用(calling)操作的示例性方法的框图。

[0016] 图8为根据一种实施方式的用于核酸测序的示例性系统的框图。

具体实施方式

[0017] 根据本申请中体现的教导和原理,提供了新的方法、系统和非暂时性机器可读存储介质,以压缩分子标记的核酸序列数据,以形成与独特分子标签相关的核酸序列读数家族的共有压缩数据,并基于共有压缩数据检测融合。

[0018] 在各种实施方式中,DNA(脱氧核糖核酸)可以被称为由4种类型的核苷酸组成的核苷酸链;A(腺嘌呤)、T(胸腺嘧啶)、C(胞嘧啶)和G(鸟嘌呤),并且RNA(核糖核酸)由4种类型的核苷酸组成;A、U(尿嘧啶)、G和C。某些核苷酸对以互补方式(称为互补碱基配对)彼此特异性结合。也就是,腺嘌呤(A)与胸腺嘧啶(T)配对(然而,在RNA的情况下,腺嘌呤(A)与尿嘧

啶(U)配对,胞嘧啶(C)与鸟嘌呤(G)配对)。当第一核酸链与由与第一链中的核苷酸互补的核苷酸组成的第二核酸链结合时,这两条链结合以形成双链。在各种实施方式中,"核酸测序数据"、"核酸测序信息"、"核酸序列"、"基因组序列"、"遗传序列"或"片段序列"或"核酸测序读数"或"核酸序列读数"或"序列读数"表示指示DNA或RNA分子(例如全基因组、全转录组、外显子组、寡核苷酸、多核苷酸、片段等)中核苷酸碱基(例如腺嘌呤、鸟嘌呤、胞嘧啶和胸腺嘧啶/尿嘧啶)的顺序的任何信息或数据。

[0019] 在各种实施方案中,"多核苷酸"、"核酸"或"寡核苷酸"是指通过核苷间键连接的核苷(包括脱氧核糖核苷、核糖核苷或其类似物)的线性聚合物。通常,多核苷酸包含至少三个核苷。通常,多核苷酸包含至少三个核苷。通常,寡核苷酸的大小从几个单体单元(例如3-4个)到几百个单体单元。无论何时多核苷酸诸如寡核苷酸由字母序列表示,如"ATGCCTG",应当理解核苷酸从左到右为5'→3'顺序,并且除非另有说明,否则"A"表示脱氧腺苷,"C"表示脱氧胞苷,"G"表示脱氧鸟苷,以及"T"表示胸苷。字母A、C、G和T可用于表示碱基本身、核苷或包含碱基的核苷酸,如本领域的标准。

[0020] 如本文所用的术语"基因座"是指染色体或核酸分子上的特定位置。基因座的等位基因位于同源染色体上的相同位点。

[0021] 如本文所用,术语"衔接子"或"衔接子及其互补物"以及其衍生物是指可与本公开的核酸分子连接(ligated)的任何线性寡核苷酸。任选地,衔接子包括与样品内至少一个靶序列的3'端或5'端基本上不互补的核酸序列。在一些实施方式中,衔接子与样品中存在的任何靶序列的3'端或5'端基本上不互补。在一些实施方式中,衔接子包括与扩增的靶序列基本上不互补的任何单链或双链线性寡核苷酸。在一些实施方式中,衔接子与样品的至少一个、一些或所有核酸分子基本上不互补。在一些实施方式中,合适的衔接子长度为约10-100个核苷酸、约12-60个核苷酸和约15-50个核苷酸的长度。衔接子可以包括核苷酸和/或核酸的任何组合。在一些方面,衔接子可以在一个或多个位置包括一个或多个可切割的基团。在另一方面,衔接子可以包括与引物(例如通用引物)的至少一部分基本上同一或基本上互补的序列。在一些实施方式中,衔接子可以包括条形码或标签以帮助下游编目、鉴定或测序。在一些实施方式中,当与扩增的靶序列连接时,单链衔接子可充当扩增的底物,尤其是在聚合酶和dNTPs存在下在合适的温度和pH下。

[0022] 如本文所用,"DNA条形码"或"DNA标签序列"及其衍生物是指衔接子内的独特短(例如6-14个核苷酸)核酸序列,其可充当区分或分离样品中的多个扩增的靶序列的"键"。为了本公开的目的,DNA条形码或DNA标记序列可以并入衔接子的核苷酸序列中。

[0023] 在一些实施方式中,本公开提供了来自靶核酸分子群的多个靶特异性序列的扩增。在一些实施方式中,所述方法包括使一个或多个靶特异性引物对与靶序列杂交,延伸引物对的第一引物,使来自核酸分子群的延伸的第一引物产物变性,使引物对的第二引物与延伸的第一引物产物杂交,延伸第二引物以形成双链产物,以及将靶特异性引物对从双链产物消化掉,以产生多个扩增的靶序列。在一些实施方式中,消化包括从扩增的靶序列部分消化一个或多个靶特异性引物。在一些实施方式中,扩增的靶序列可与一个或多个衔接子连接。在一些实施方式中,衔接子可以包括一个或多个DNA条形码或标记序列。在一些实施方式中,一旦与衔接子连接,扩增的靶序列可经历切口平移反应和/或进一步扩增以产生衔接子连接的扩增的靶序列的文库。

[0024] 在一些实施方式中,本公开的方法包括选择性扩增含有多个核酸分子的样品中的靶序列,并将扩增的靶序列与至少一个衔接子和/或条形码连接(ligating)。用于分子生物学文库制备技术的衔接子和条形码为本领域技术人员公知。本文所用的衔接子和条形码的定义与本领域所用的术语一致。例如,条形码的使用允许每个多重反应的多个样品、来源、组织或核酸分子群的检测和分析。条形码化且扩增的靶序列含有独特的核酸序列,通常是短的6-15个核苷酸序列,其鉴定并区分一个扩增的核酸分子与另一个扩增的核酸分子,即使当两个核酸分子减去条形码均含有相同的核酸序列时。衔接子的使用允许以统一的方式扩增每个扩增的核酸分子,并帮助减少链偏差(strand bias)。衔接子可以包括通用衔接子或适当衔接子,这两者都可以在下游使用以执行一个或多个不同的功能。例如,通过本文公开的方法制备的扩增的靶序列可以与衔接子连接,所述衔接子可以在下游用作克隆扩增的平台。衔接子可以作为模板链,用于使用第二组引物的随后的扩增,因此允许衔接子连接的扩增的靶序列的通用扩增。在一些实施方式中,选择性扩增靶核酸以产生扩增子库可以进一步包括将一个或多个条形码和/或衔接子与扩增的靶序列连接。结合条形码的能力提高了样品通量,并且允许同时分析多个样品或材料源。

[0025] 在本申请中,"反应限制区域"通常是指其中可以限制反应的任何区域,并且包括例如"反应室"、"孔"和"微孔"(它们中的每一个可以互换使用)。反应限制区域可以包括例如其中固体底物的物理或化学属性可允许感兴趣的反应定位的区域,以及可特异性结合感兴趣的分析物的底物表面的离散区域(例如具有共价连接到该表面的寡核苷酸或抗体的离散区域)。反应限制区域可为中空的,或者具有定义明确的形状和体积,其可以被制造成底物。这些后一类型的反应限制区域在本文中称为微孔或反应室,并且可以使用任何合适的微制造技术制造。反应限制区域也可为例如在没有孔的底物上的基本上平坦的区域。

[0026] 多个限定的空间或反应限制区域可以以阵列排列,并且每个限定的空间或反应限制区域可以与至少一个传感器电通信,以允许检测或测量一个或多个可检测或可测量的参数或特征。该阵列在此被称为传感器阵列。传感器可以将反应副产物的存在、浓度或量的变化(或反应物的离子特性的变化)转换成输出信号,该输出信号可以被电子地记录为例如电压水平或电流水平的变化,该电压水平或电流水平的变化进而可以被处理以提取关于化学反应或期望的缔合事件(例如核苷酸掺入事件)的信息。传感器可以包括至少一个化学敏感场效应晶体管("chemFET"),其可被配置成产生与化学反应的性质或其附近感兴趣的目标靶分析物相关的至少一个输出信号。这些性质可以包括反应物、产物或副产物的浓度(或浓度变化),或物理性质的值(或该值的变化),例如离子浓度。例如,对于限定的空间或反应限制区域的pH的初始测量或查询可以表示为电信号或电压,其可以被数字化(例如,转换为电信号或电压的数字表示)。这些测量值和表示中的任何一个可以被认为是原始数据或原始信号。

[0027] 在各种实施方式中,短语"碱基空间"是指核苷酸序列的表示。短语"流动空间"是指特定核苷酸流动的掺入事件或非掺入事件的表示。例如,流动空间可为代表该特定核苷酸流动的核苷酸掺入事件(例如一,"1")或非掺入事件(例如零,"0")的一系列值。具有非掺入事件的核苷酸流动可被称为空流动,并且具有核苷酸掺入事件的核苷酸流动可被称为正流动。应当理解零和一是非掺入事件和核苷酸掺入事件的方便表示;然而,任何其他符号或指定可以替代地用于表示和/或鉴定这些事件和非事件。特别地,当多个核苷酸在给定位置

处掺入时,诸如对于均聚物拉伸(homopolymer stretch),该值可与核苷酸掺入事件的数目成比例,并因此与均聚物拉伸的长度成比例。

[0028] 图1说明用于检测融合的引物设计的示例,其中5'引物和3'引物各自具有分子标签。为了检测基因间融合,在每个基因的已知断点的可扩增范围内设计引物。例如,5'引物106被设计在基因A的断点102的左侧,3'引物被设计在基因B的断点102的右侧。每个引物在各自断点102的多个碱基或范围112和114内,使得所得扩增子包括在融合断点102处融合的基因A和基因B的部分。范围112和114与扩增子大小相关。取决于扩增子大小,引物可以被设计在距断点102的一定范围112和114处。在一些实施方式中,融合设计范围112和114可指示可用于引物设计的序列的最大量。靶向特异性融合断点102的引物104和106位于断点102的侧翼,以产生融合扩增子。图1中的融合ID扩增子代表可以鉴定靶向融合的存在的理想融合扩增子。相同的策略可用于被设计成检测基因内融合事件的测定。例如,在图1中,基因A为外显子A,基因B为外显子B。基因内事件可以包括例如外显子跳跃、非规范的和野生型转录物的事件。为了鉴定单个多核苷酸分子,分子标签108和110分别被附加到5'引物104和3'引物106上,包括附加到5'引物104的前缀标签108和附加到3'引物106的后缀标签110。用独特的分子标签标记单个多核苷酸分子,在PCR反应中扩增并测序,产生融合扩增子。用于给定靶向融合的融合扩增子可以包括用于融合ID扩增子的序列和用于5'端的前缀标签108和用于3'端的后缀标签110。PCR扩增和测序可以产生多个融合扩增子,从而当靶向融合存在时,每个原始标记的多核苷酸分子产生多个序列读数。独特的分子标签用于鉴定源自相同多核苷酸分子的序列读数,并将它们分类成具有相同标签序列的家族。

[0029] 家族或分子家族是指具有相同独特分子标签的序列读数组。家族大小是家族中序列读数的数目。功能家族是具有大于最小家族大小的成员数量的家族。最小家族大小可为任何整数值。例如,最小家族大小可为3或更大。

[0030] 图2是根据一种实施方式的产生用于检测融合的共有压缩数据的示例性方法的框图。流动空间信号测量值可以通过核酸测序装置提供给处理器。在一些实施方式中,每个流动空间信号测量值表示响应于传感器阵列的微孔中的样品核酸掺入或未掺入流动的核苷酸而测量的信号振幅或强度。对于掺入事件,信号振幅取决于在一个流处掺入的碱基的数目。对于均聚物,信号振幅随着均聚物长度的增加而增加。处理器可以应用碱基调用者202,以通过分析流动空间信号测量值来产生序列读数的碱基调用。

[0031] 图4显示可以从其进行碱基调用的流动空间信号测量值的示例性表示。在该实施例中,x轴显示流动指数和在流动序列中流动的核苷酸。图中的条显示来自传感器阵列中微孔的特定位置的每个流的流动空间信号测量值的振幅。流动空间信号测量值可为原始采集数据或已经处理的数据,例如通过缩放、背景过滤、归一化、信号衰减校正和/或相位误差或效应校正等。可以通过分析任何合适的信号特征(例如,信号振幅或强度)来进行碱基调用。用于与本教导一起使用的传感器阵列、信号处理和碱基调用的结构和/或设计可以包括2013年4月11日的美国专利申请公开号2013/0090860中描述的一个或多个特征,其全部内容通过引用并入本文。

[0032] 一旦确定了序列读数的碱基序列,就可以将序列读数提供给映射器204,例如,在未映射的BAM文件中。在一些实施方式中,映射器204将序列读数与定制融合参考序列和对照基因参考序列比对,以确定比对的序列读数和相关的映射质量参数。定制融合参考可以

包括用于靶向融合的嵌合序列。对照基因参考序列可以包括管家基因的RNA转录序列。管家基因是维持基本细胞功能所必需的,并且在正常和病理条件下在生物体的细胞中表达。靶向融合参考序列和对照基因参考序列可以使用FASTA文件格式或其他合适的文件格式在文件中提供。用于与本教导一起使用的比对序列读数的方法可以包括在2012年8月2日公开的美国专利申请公开号2012/0197623中描述的一种或多种特征,其全部内容通过引用并入本文。可以将比对的序列读数提供给流动空间共有管线206,例如,在映射的BAM文件中。

[0033] BAM文件格式结构在2014年9月12日的"Sequence Alignment/Map Format Specification(序列比对/映射格式规范)"(https://github.com/samtools/hts-speces)中描述,本文中称为"BAM规范"。如本文所述,"BAM文件"是指与BAM格式兼容的文件。如本文所述,"未映射的"BAM文件是指不包含比对的序列读数信息或映射质量参数的BAM文件,并且"映射的"BAM文件是指含有比对的序列读数信息和映射质量参数的BAM文件。如本文所述,"共有"BAM文件是指含有共有压缩数据的BAM文件。

[0034] 在一些实施方式中,具有分子标记的序列读数的读数结构可以包括从5'端开始的文库键、条形码序列、条形码衔接子、前缀分子标签、序列模板、后缀分子标签和P1衔接子。碱基调用可以包括从序列读数的剩余部分修剪文库键、条形码序列和条形码衔接子,并将它们存储在BAM文件格式的读数组标题@RG的键序列(KS)标签字段中。碱基调用可以包括从序列读数中修剪P1衔接子,并将其存储在BAM标题的注释行@CO中。

[0035] 在一些实施方式中,碱基调用者202可以被配置成检测标签结构并从序列读数中修剪标签。经修剪的标签可以存储在定制标签ZT(例如,用于前缀标签)和YT(例如,用于后缀标签)的字段中的BAM读数组标题(@RG)中。由于读数组标题与模板的序列读数数据相关,可以维持标签与家族组的关联的完整性。随后的与参考序列的映射或比对可应用于没有前缀标签或后缀标签的模板序列。这降低了标签的一部分错误地映射到参考序列的可能性。

[0036] 在一些实施方式中,标签序列可以包括随机碱基的子集和已知碱基的子集。标签修剪方法可能需要序列读数的标签部分中的碱基序列与已知碱基匹配。标签修剪方法可以选择具有等于标签的已知长度的多个碱基的碱基串。在一些实施方式中,标签修剪方法可检测和校正标签中的测序错误,例如插入和缺失。校正标签中的测序错误可提供更准确的家族鉴定。

[0037] 在一些实施方式中,映射的BAM文件可以存储多个序列读数、多个流动空间信号测量值的向量和对应于序列读数的多个序列比对。映射的BAM文件可以将流动空间信号测量值的向量存储在定制标签字段ZM中。映射的BAM文件可以将模型参数存储在定制标签字段ZP中。映射的BAM文件可以将与序列读数相关的分子标签序列存储在BAM读数组标题中,如上所述。映射的BAM文件可以存储在存储器中,并提供给流动空间共有管线206。在一些实施方式中,可使用其他文件格式来存储多个序列读数、多个流动空间信号测量值的向量、多个序列比对和对应于序列读数的分子标签序列。

[0038] 图3是根据一种实施方式的流动空间共有管线206的示例性方法的框图。分组操作302可以使用分子标签序列信息,来鉴定序列读数家族和对应的流动空间信号测量值。分组操作302可比较与序列读数相关的分子标签序列,并应用分组阈值。例如,分组阈值的标准可要求序列读数组的成员的所有标签序列具有100%标签序列同一性。通过满足分组阈值的标准,被确定为共享共同标签序列的序列读数和对应的流动空间信号测量值被分组到给

定家族中,其中所述共同标签序列对于该家族是独特的。每个家族将具有多个成员,其是在家族中分组的序列读数的数目。在一些实施方式中,不具有至少最小数量的成员的家族将不被进一步处理,并且可被从存储器中移除。用于与本教导一起使用的基于分子标签序列将序列读数分组的方法可以包括在2016年12月15日公开的美国专利申请公开号2016/0362748中描述的一个或多个特征,其全部内容通过引用并入本文。

[0039] 在一些实施方式中,流动空间共有压缩器304可以基于每个分组的家族的流动空间信号测量值来确定共有压缩数据,如下:

[0040] A.计算每个分组的家族的流动空间信号测量值的向量的算术平均值,以形成每个家族的共有流动空间信号测量值的向量。

[0041] B.计算每个家族的流动空间信号测量值的向量的标准偏差,以形成每个家族的标准偏差的向量。

[0042] 在一些实施方式中,流动空间共有压缩器304可以接收对应于流动空间信号测量值的每个向量的至少一个模型参数。流动空间共有压缩器304可以计算该家族的模型参数的算术平均值,以形成该家族的至少一个共有模型参数。模型参数可用于碱基调用,如下所述。在一些实施方式中,模型参数可以包括用于流动空间信号测量值的每个向量的不完全扩展(IE)参数和结转(carry forward)(CF)参数。流动空间共有压缩器304可计算每个家族的IE参数的算术平均值和CF参数的算术平均值,以形成每个家族的共有IE参数和共有CF参数。

[0043] 在一些实施方式中,碱基调用者202可以被应用于每个家族的共有流动空间信号测量值的向量,以产生相应家族的共有碱基序列。共有碱基序列在本文中也称为共有序列读数。共有模型参数可用于应用碱基调用模型。例如,可以将每个家族的共有不完全扩展(IE)参数和共有结转(CF)参数提供给碱基调用者202。碱基调用可以包括在2013年4月11日公开的美国专利申请公开号2013/0090860和在2012年5月3日公开的美国专利申请公开号2012/0109598中描述的一个或多个特征,其全部内容通过引用并入本文。共有碱基序列的共有序列比对可以通过将共有碱基序列与具有最高映射质量的家族中的序列读数进行比较来确定。如果共有碱基序列与具有最高映射质量的序列读数相匹配,则选择相应的序列比对作为共有序列比对。如果共有碱基序列与具有最高映射质量的家族中的序列读数不匹配,则映射器204可以将共有碱基序列与靶向融合参考序列和对照基因参考序列进行比对,以确定共有序列比对。用于比对共有序列读数的方法可以包括在2012年8月2日公开的美国专利申请公开号2012/0197623中描述的一个或个种特征,其全部内容通过引用并入本文。在一些实施方式中,平均约1%的共有序列读数可能需要通过映射器204进行重新比对。

[0044] 在一些实施方式中,处理器可以将每个家族的共有压缩数据存储在存储器中的压缩数据结构中。共有压缩数据可以包括共有序列读数、共有序列比对、共有流动空间信号测量值的向量、标准偏差的向量和每个家族的成员的数目。共有压缩数据可以进一步包括用于每个家族的共有模型参数组。如果该家族已被分成亚家族,则共有压缩数据可以进一步包括共有序列读数、共有序列比对、共有流动空间信号测量值的向量、标准偏差的向量和每个亚家族的成员的数目。在一些实施方式中,压缩数据结构可以与BAM文件格式兼容,以产生映射的共有BAM文件。BAM规范允许用户定义定制标签字段。例如,可以为用于存储一些共有压缩数据的BAM文件定义定制标签字段,如表1所示。

[0045] 表1。

[0046]

BAM定制标签字段	数据
ZM	共有流动空间信号测量值
ZP	共有模型参数
ZS	流动空间信号测量值的标准偏差
ZR	家族或亚家族中序列读数或成员的数目

[0047] 每个家族的原始序列读数、流动空间信号测量值的原始向量和原始模型参数不包括在共有压缩数据中,并且可以从存储器中移除。在一些实施方式中,压缩数据结构可以使用与BAM文件格式不同的格式协议,包括定制文件格式。

[0048] 图5显示单个家族的流动空间信号测量值的示例性绘图。流动指数指示流序列中的第j个流。归一化振幅指示流动空间信号测量值的值。绘图符号的类型对应于特定流处的核苷酸。该流动空间信号测量值的绘图对应于与共同分子标签相关的序列读数的单个家族。在每个流处的流动空间信号测量值的值被聚集在相似的值附近。流动指数对应于流动空间信号测量值的向量中的元素指数。在该绘图中表示的流动空间信号测量值可以被输入到流动空间共有压缩器304。

[0049] 图6说明单个家族的共有流动空间信号测量值的示例性绘图。该绘图显示由对图5所示的流动空间信号测量值的共有计算得到的共有流动空间信号测量值。绘图符号指示算术平均值,其是该家族的共有流动空间测量值的向量的元素。条表示标准偏差,其是该家族的标准偏差的向量的元素。

[0050] 对于双向测序,第一家族可被指定用于正向序列读数,并且第二家族可被指定用于反向序列读数。正向读数的前缀和后缀标签可为反向读数的前缀和后缀标签的反向互补物,如表2的示例所示。

[0051] 表2

[0052]

读数方向	前缀标签	后缀标签	标签	序列读数的数目
正向读数	ACT	GGT	ACTGGT	10
反向读数	ACC	AGT	ACCAGT	10
反向读数标签的反向互补物	GGT	ACT	ACTGGT	20

[0053] 在一些实施方式中,家族可被分成亚家族,导致具有相同分子标签的每个家族有一个以上共有序列读数。可形成用于流同步的亚家族,使得每个亚家族具有同步的流动空间信号测量值,用于确定共有流动空间测量值的向量。当家族内的序列读数存在变化时,家族可被分成亚家族,使得对于每个亚家族产生共有序列读数。用于与本教导一起使用的分子标记的核酸序列数据的流动空间共有压缩的方法可以包括在2018年5月15日提交的美国专利申请号15/979,804中描述的一个或多个特征,其通过引用整体并入本文。

[0054] 回到图2,在一些实施方式中,在融合分析之前,可以将第二级压缩器208应用于共有压缩数据。第二级压缩器208可将具有相同分子标签的亚家族组合成包括一个共有序列读数的单个家族。在双向测序读数的一些实施方式中,第二级压缩器208可以如下组合正向和反向序列读数的家族:

[0055] 1.确定反向读数的前缀标签和后缀标签的反向互补,以形成反向互补物标签,

[0056] 2.将反向互补物标签与正向读数标签匹配,

[0057] 3.将正向读数家族与匹配标签和反向读数家族组合成包括一个共有序列读数的一个家族。

[0058] 参考表2,由组合家族表示的读数的数目是正向和反向读数家族中序列读数数目的总和。映射的共有BAM文件可以被修改以包括组合家族信息并去除亚家族信息。总和值可以进入映射的共有BAM文件的ZR字段中。第二级压缩器208提供组合家族的单个共有序列读数。通过消除每个组合家族的一个亚家族的共有序列读数,第二级压缩器208提供额外的数据压缩。在第二级压缩之后,共有压缩数据可以被提供给融合调用者210。

[0059] 图7为使用共有压缩数据进行融合调用操作的示例性方法的框图。在一些实施方式中,在步骤701处,处理器可通过以下来验证并提供靶向融合的计数:

[0060] a.通过分析共有序列读数与靶向融合参考序列的比对的特征,鉴定有资格被计数用于靶向融合检测的共有序列读数。

[0061] b.对于合格的序列读数,计数对应于合格的共有序列读数的家族的数目,所述合格的共有序列读数与每个靶向融合参考序列比对,以给出家族计数。对于双向测序,计数每条链的家族的数目,以给出每条链的家族计数。

[0062] c.将最小分子计数阈值应用于每个靶向融合的家族计数。例如,最小分子计数阈值可以被设定为大于或等于3的值。对于双向测序,将每条链的最小分子计数阈值应用于每条链的家族计数。例如,每条链的最小分子计数阈值可以被设定为大于或等于1的值。

[0063] d.将家族的序列读数的数目求和,这些家族的序列读数对应于与每个靶向融合参考序列比对的合格的共有序列读数,以给出每个靶向融合的读数计数。每个家族的序列读数的数目可以包括在映射的共有BAM文件中。

[0064] e.将最小读数计数阈值应用于每个靶向融合的读数计数。例如,最小读数计数阈值可以被设定为大于或等于21的值。

[0065] f.确定与过程控制相关的样品质量控制(QC)参数。过程控制或表达控制指示共有序列读数与对照基因参考序列的比对结果,所述对照基因参考序列由映射器204确定并且包括在映射的共有BAM文件中。在一些实施方式中,样品QC参数可以包括验证所需的最小数量的过程控制。例如,当使用2个对照基因时,将过程控制的最小数量设定为1,意味着对于2个对照基因中的至少1个,应当存在与对照基因参考序列比对的共有序列读数。例如,对于使用6个对照基因的双向测序,将过程控制的最小数量设定为3,意味着对于6个对照基因中的至少3个,应当存在与对照基因参考序列比对的共有序列读数。如果对于与对照基因参考序列比对的共有序列读数,家族计数大于最小分子计数阈值并且读数计数大于或等于读数计数阈值,则过程控制目标可以被称为存在。在一些实施方式中,样品QC参数可以包括融合组(fusion panel)的映射序列读数的最小总和。例如,映射的序列读数的最小总和的值可以为20000。在一些实施方式中,样品QC参数可以包括样品中所有序列读数的最小平均序列读数长度。例如,最小平均读数长度的值可以为50。

[0066] g.确定关于靶向融合的存在的调用或无调用决定。如果对于靶向融合断点满足最小分子计数阈值、最小读数计数阈值和样品QC参数,则可作出存在靶向融合的调用决定。如果靶向融合断点不满足最小分子计数阈值和最小读数计数阈值中的一个或多个,则可作出不存在的决定。在一些实施方式中,如果样品具有少于最小总映射序列读数,则对组中的所有靶向融合给予无调用决定。在一些实施方式中,如果计算来自所有序列的平均序列读数

长度,并且如果平均值小于最小平均序列读数长度,则对组中的所有靶向融合给予无调用决定。在一些实施方式中,如果检测到的过程控制的数目小于过程控制阈值的最小数目,则对组中的所有靶向融合给予无调用决定。

[0067] 在一些实施方式中,鉴定合格的共有序列读数的步骤(以上步骤a)包括分析共有序列读数,以确保每个比对的共有序列读数在其被计数用于融合检测之前具有来自涉及基因融合的两个基因配偶体的代表。共有序列读数与靶向融合参考序列的比对的特征可以包括同源性特征、映射质量特征和断点扫描特征。处理器如下分析每个比对的共有序列读数的这些特征:

[0068] i.确定共有序列读数是否跨越靶向融合参考序列上的融合断点。

[0069] ii.确定共有序列读数与各配偶体序列的同源性水平,以给出第一和第二同源性水平。同源性为共有序列读数和靶序列之间的重叠中的碱基的数目。配偶体序列为靶向融合参考序列在融合断点一侧的部分。每个靶向融合具有第一和第二配偶体序列,在靶向融合参考序列中的断点的每一侧上各一个。

[0070] iii.将第一和第二同源性水平与最小同源性阈值进行比较。最小同源性阈值可以基于所需的灵敏度和严格性来选择。例如,最小同源性阈值可以针对高灵敏度/低严格性设定为60%,针对默认值设定为70%,针对高严格性设定为80%。也可以使用最小同源性阈值的其他值。这些值可为50%至100%。

[0071] iv.确定在每个配偶体序列内的比对的共有序列读数的映射质量值,以产生第一和第二映射质量值。可通过计算与配偶体序列匹配的比对的共有序列读数中的匹配碱基的数目与和配偶体序列重叠的比对的共有序列读数中的重叠碱基的数目的比率,来确定映射质量值。

[0072] v.将第一映射质量值和第二映射质量值与映射质量阈值进行比较。例如,映射质量阈值的值可为66.6%。也可以使用其他映射质量阈值的值。这些值可为50%至100%。

[0073] vi.当共有序列读数满足跨越融合断点、具有至少为最小同源性阈值的第一和第二同源性水平以及具有至少为映射质量阈值的第一和第二映射质量值的标准时,其为上述步骤b的合格的共有序列读数。在一些实施方式中,当共有序列读数不满足这些标准时,可将其提供给图7中的步骤702以鉴定非靶向融合的存在,或可将其过滤掉。

[0074] 在一些实施方式中,不满足上述标准的共有序列读数可部分地映射到靶向融合参考序列。部分映射的共有序列读数可具有映射部分和未映射部分。映射部分可以靠近共有序列读数的开始,而未映射部分靠近共有序列读数的末端,或者未映射部分可以靠近共有序列读数的开始,而映射部分可以靠近共有序列读数的末端。

[0075] 在一些实施方式中,在步骤702,处理器如下分析部分映射的共有序列读数,以鉴定非靶向融合组合的任何证据:

[0076] a)将部分映射的共有序列读数分成映射部分和未映射部分,使得部分映射的共有序列读数产生两个读数片段。

[0077] b)将读数片段独立地与融合参考序列比对。例如,部分映射的读数的第一片段将映射到融合参考序列内的第一基因座,并且部分映射的读数的第二片段将映射到融合参考序列内的第二基因座。基因座可为参考序列上读数片段的映射位置。例如,两个读数片段可与两个不同的融合参考序列比对。例如,两个读数片段可与相同的融合参考序列比对。

[0078] c)确定第一片段是否与相应融合参考序列的5'端比对,并且第二片段是否与相应融合参考序列的3'端比对。

[0079] d)确定每个配偶体序列内的比对的读数片段的映射质量值是否大于或等于映射质量阈值。映射质量值在上述步骤iv和步骤v中描述。例如,映射质量阈值的值可为66.6%。

[0080] e)确定第一和第二比对的读数片段的同源性水平。同源性水平描述于上述步骤ii中。将同源性水平求和,以给出两个读数片段的组合同源性水平。

[0081] f)确定所述组合同源性水平是否大于或等于组合同源性阈值。例如,组合同源性阈值可为150%。例如,允许第一片段的同源性水平为100%,第二片段的同源性水平为50%。例如,允许第一片段的同源性水平为90%,第二片段的同源性水平为70%。例如,第一片段的同源性水平为80%且第二片段的同源性水平为60%是不够的,因为总和小于150%。

[0082] g)对于满足组合同源性水平阈值和映射质量阈值的比对的读数片段,计数对应于产生与相应融合参考序列比对的读数片段的共有读数序列的家族的数目,以给出家族计数。

[0083] h)将最小分子计数阈值应用于家族计数。例如,最小分子计数阈值可以被设定为大于或等于3的值。对于双向测序,将每条链的最小分子计数阈值应用于每条链的家族计数。例如,每条链的最小分子计数阈值可以被设定为大于或等于1的值。

[0084] i)将家族的序列读数的数目求和,这些家族的序列读数对应于产生与相应融合参考序列比对的读数片段的共有序列读数,以给出读数计数。

[0085] j)将最小读数计数阈值应用于每个靶向融合的读数计数。例如,最小读数计数阈值可以被设定为大于或等于21的值。

[0086] 用于与本教导一起使用的确定部分映射的序列读数中的融合的方法可以包括在2016年12月15日公开的美国专利申请公开号2016/0362748中描述的一个或多个特征,其全部内容通过引用结合到本文中。

[0087] 在一些实施方式中,从步骤701和步骤702产生的家族计数可被聚集,以形成总家族计数。最小家族计数阈值可应用于总家族计数。在一些实施方式中,可以将从步骤701和步骤702产生的读数计数聚集,以形成总读数计数。最小读数计数阈值可以应用于总读数计数。

[0088] 关于图1、图2、图3和图7描述的方法可应用于基因间融合事件和基因内融合事件。对于基因间融合事件,靶向融合参考序列包含用于在融合断点处融合两个基因的参考序列。对于基因内融合事件,靶向融合参考序列包含用于在来自相同基因的融合断点处融合两个外显子的参考序列。

[0089] 基因内融合事件可被设计成检测外显子缺失、非规范和野生型转录物。外显子缺失靶的示例包括EGFR-EGFR.E1E8.Delpositive(通常称为EGFRvIII缺失)和MET-MET.M13M15(MET外显子14跳跃测定)。基因内融合事件被称为RNAExonVariants(RNA外显子变体)。在一些实施方式中,对于具有基因内融合靶的每个基因,加入至少一个野生型靶,所述野生型靶被设计为扩增该基因的规范转录物。定制融合参考和对照基因可以包括野生型靶标的参考序列。共有序列读数的一部分可以与野生型参考序列比对。野生型靶称为野生型RNAExonVariants(RNA外显子变体)。可以例如以BED文件格式为野生型靶提供附加注释。

[0090] 在一些实施方式中,RNA外显子变体检测外显子缺失和替代转录物的分析可以包

括计算和报告以下度量：

[0091] I.基因内融合RNA外显子变体V_i与野生型RNA外显子变体的比率：

[0092] 比率＝(V_i的读数计数)/(相同基因的所有野生型RNA外显子变体的平均读数计数)。

[0093] 可以通过将家族的序列读数的数目相加,这些家族的序列读数对应于与靶向融合参考序列比对的共有序列读数的一部分,来确定基因内融合的V_i的读数计数。可通过将家族的序列读数的数目相加,这些家族的序列读数对应于与野生型参考序列比对的共有序列读数的一部分,来确定野生型RNA外显子变体的读数计数。

[0094] II.基因内融合RNA外显子变体V_i在相同基因内的归一化计数：

[0095] 范数计数＝(V_i的读数计数)/(相同基因的所有RNA外显子变体的读数计数的总和)。

[0096] 可以通过将对应于野生型参考序列的读数计数和对应于相同基因的靶向融合参考序列的读数计数相加,来确定所有RNA外显子变体的读数计数的总和。可通过将家族的序列读数的数目相加,这些家族的序列读数对应于与相同基因的野生型参考序列比对的共有序列读数的一部分,来确定对应于野生型参考序列的读数计数。可通过将家族的读数计数相加,这些家族的读数对应于与相同基因的靶向融合参考序列比对的共有序列读数,来确定对应于靶融合参考序列的读数计数。

[0097] 在步骤703,处理器可以核对并格式化在步骤701中确定的靶向融合得到的总计数,可以核对并格式化在步骤703中产生的非靶向融合和RNA外显子变体的另外的度量以在文件中存储。在步骤704,处理器可产生融合结果的可视化,诸如每个融合/对照基因的归一化计数的概要表和动态热图,以呈现给用户。在表3中给出了结果的示例。

[0098] 表3

[0099]

变体(外显子)	摩尔计数	读数计数	检测	QC测试	类型	与野生型的比率	基因内的范数计数
EML4-ALK.E6aA20.AB374361	16	202	存在	通过	融合		
EML4-ALK.E6bA20.AB374362	5	56	存在	通过	融合		
SLC34A2-ROS1.S4R32.COSF1197	340	5011	存在	通过	融合		
SLC34A2-ROS1.S4R34.COSF1198	28	464	存在	通过	融合		
CCDC6-RET.C1R12.COSF1271	36	569	存在	通过	融合		
TBP.ENCTRL.E3E4	328	5724	存在	通过	过程控制		
HMBS.ENCTRL.E8E9	606	9008	存在	通过	过程控制		
MET.E6E7.WT	2830	43648	存在	通过	RNA 外显子变体	1.152864	0.53447
MET.E11E12.WT	2906	32073	存在	通过	RNA 外显子变体	0.847136	0.392734
MET-MET.M13M15	432	5945	存在	通过	RNA 外显子变体	0.157024	0.072797

[0100] 在结果表中还可以包括其他信息,例如基因座信息、注释信息、分析名称、样品名称和条形码ID。

[0101] 表4显示通过本文所述的方法实现的双向序列读数数据的所得压缩。

[0102] 表4

[0103]

变体(外显子)	原始读数计数	共有压缩	第2次压缩和过滤
AFM.E12E13.NM_001133.2	95	19	2
APOB.E7E8.NM_000384.2	3157	483	107
CHMP2A.E4E5.NM_014453.2	505656	85252	15094
FBXW2.E5E6.NM_012164.3	583591	72010	10671
GUSB.E7E8.NM_000181.3	500218	76878	9779
HMBS.E3E4.NM_000190.3	108978	13847	2050
ITGB7.E9E10.NM_000889.1	24197	3482	602
LMNA.E1E2.NM_005572.3	1259880	268496	44624
LRP1.E41E42.NM_002332.2	2078078	305540	42782
MET.E11E12.NM_000245.2	138973	21579	2647
MET.E6E7.NM_000245.2	121732	17652	2618
MRPL13.E1E2.NM_014078.5	546165	70526	9997
MTTP.E12E13.NM_000253.2	1291	85	25
MYC.E1E2.NM_002467.4	329437	63108	8104
PSMB2.E3E4.NM_001199779.1	1087559	159101	20576

PUM1.E3E4.NM_001020658.1	825501	116543	19967
SNRPD3.E1E2.NM_004175.3	718619	121620	13362
TBP.E3E4.NM_001172085.1	174953	16809	3069
TRIM27.E7E8.NM_006510.4	140781	19916	2637
VCP.E7E8.NM_007126.3	1744231	323054	33616

[0104] 表5显示通过本文所述的方法实现的单向序列读数数据的所得压缩。

[0105] 表5

[0106]

变体(外显子)	原始读数计数	共有压缩	第2次压缩和过滤
CCDC6-RET.C1R12.COSF1271	521	144	35
CUX1-RET.C10R12	156	128	0
EML4-ALK.E15A20.COSF413.1	136	102	0
EML4-ALK.E18A20.COSF487.1	9	6	0
EML4-ALK.E6A19.COSF1296.2	2	2	0
EML4-ALK.E6aA20.AB374361	79	16	10
EML4-ALK.E6bA20.AB374362	84	16	7
EZR-ROS1.E10R34.COSF1267	38	31	0
GOPC-ROS1.G4R36.COSF1188	50	39	0
GOPC-ROS1.G8R35.COSF1139	1363	1130	0
HIP1-ALK.H28A20.1	20	17	0
HMBS.ENCTRL.E8E9	5085	710	482
KIF5B-ALK.K17A20.COSF1257	157	134	0
KIF5B-RET.K16R12.COSF1230	1947	1505	2
KIF5B-RET.K22R12.COSF1253	123	96	0
KIF5B-RET.K24R8.COSF1236	1	1	0
KLC1-ALK.K9A20.COSF1276	4	3	0
LRIG3-ROS1.L16R35.COSF1269.1	47	35	0
MET-MET.M13M15	5209	2058	429
MET.E11E12.WT	19028	3639	2255
MET.E6E7.WT	17573	2934	1676
SLC34A2-ROS1.S4R32.COSF1196	4357	965	561
SLC34A2-ROS1.S4R34.COSF1198	474	88	42
TBP.ENCTRL.E3E4	3762	644	340
TPR-ALK.T15A20	1	1	0

[0107] 在表4和表5中,原始读数计数给出与变体(外显子)列中所示的基因座比对的序列读数的原始数目。共有压缩列给出在通过流动空间共有管线206压缩后,比对的共有序列读数的数目或家族的数目。更多的众多原始序列读数被共有压缩数据的共有序列读数替换。结果是,实现了数据量的显著减少。第2次压缩和过滤列显示比对的共有序列读数的数目,或家族的数目,在应用第二级压缩器208并过滤出具有不符合融合检测计数的共有序列读数的那些之后的剩余。共有序列读数数目的进一步减少导致数据量的进一步减少。整体压

缩导致原始序列读数数据的数据量显著减少。压缩率直接与存储量的减少有关,该存储量是存储来自存储原始序列读数数据所需的存储量的共有序列读数数据所需的。

[0108] 在一些实施方式中,本文描述的方法可以至少部分地使用分布式、集群(clustered)、远程或云计算资源来执行或实施。压缩序列读数数据以提供共有压缩数据,提供了将数据传输到分布式、集群、远程或云计算资源中的处理器的优点。由于数据量减少,所以计算资源之间的跨数据传送接口的传输所需的带宽和/或时间减少。例如,映射的共有BAM文件可以从本地计算资源传送到云计算资源,以用于融合检测操作。映射的共有BAM文件的大小将显著小于原始映射的BAM文件的大小。映射的共有BAM文件的较小的大小将减少通过跨数据传送接口传输到云计算资源所需的带宽和/或时间。

[0109] 根据示例性实施方式,提供了一种用于压缩分子标记的核酸序列数据以供融合检测的方法,所述方法包括:(a)接收多个核酸序列读数和多个序列读数家族的多个序列比对,其中每个序列读数与分子标签序列相关,所述分子标签序列鉴定由核酸样品中的特定多核苷酸分子产生的序列读数家族,每个家族具有多个序列读数,其中所述序列比对的一部分对应于映射到靶向融合参考序列的序列读数;(b)基于对应于所述家族的所述序列读数的流动空间信号测量值,确定每个序列读数家族的共有序列读数;(c)确定每个序列读数家族的共有序列比对,其中所述共有序列比对的一部分对应于与所述靶向融合参考序列比对的所述共有序列读数;(d)产生包含共有压缩数据的压缩数据结构,所述共有压缩数据包括每个家族的共有序列读数和共有序列比对;以及(e)使用来自所述压缩数据结构的所述共有序列读数和所述共有序列比对,来检测融合。对于双向测序,其中正向共有序列读数和反向共有序列读数在单独的家族中,其中正向家族与第一前缀标签和第一后缀标签相关,反向家族与第二前缀标签和第二后缀标签相关,所述方法还包括当第二前缀标签和第二后缀标签的反向互补物与第一前缀标签和第一后缀标签匹配时,组合正向家族和反向家族,以形成具有压缩数据结构的一个共有序列读数的组合家族。检测融合的步骤可以进一步包括基于共有序列读数与靶向融合参考序列的共有序列比对的特征,鉴定合格的共有序列读数。所述特征可以包括同源性特征、映射质量特征和跨断点的特征。鉴定合格的共有序列读数的步骤可以进一步包括,确定与靶向融合参考序列比对的共有序列读数是否跨越靶向融合参考序列的融合断点。鉴定合格的共有序列读数的步骤可以进一步包括,确定共有序列读数分别与靶向融合参考序列的第一和第二配偶体序列的第一和第二同源性水平是否大于或等于最小同源性阈值。鉴定合格的共有序列读数的步骤可以进一步包括,确定分别在靶向融合参考序列的第一和第二配偶体序列内的共有序列读数的第一和第二映射质量值是否大于或等于映射质量阈值。鉴定合格的共有序列读数的步骤可以进一步包括通过计算与配偶体序列匹配的共有序列读数中的匹配碱基的数目和与配偶体序列重叠的共有序列读数中的重叠碱基的数目的比率,来确定映射质量值。鉴定合格的共有序列读数的步骤可以进一步包括,确定对应于与靶向融合参考序列比对的合格的共有序列读数的家族的数目是否大于或等于最小分子计数阈值。鉴定合格的共有序列读数的步骤可以进一步包括,确定读数计数是否大于或等于最小读数计数阈值,其中读数计数为家族的序列读数数目的总和,这些家族的序列读数对应于与靶向融合参考序列比对的合格的共有序列读数。对应于序列读数的序列比对的一部分可以映射到对照基因参考序列,其中共有压缩数据还包括对应于对照基因参考序列的共有序列读数和共有序列比对。所述方法可以进一步包括当家族

计数大于最小分子计数阈值并且读数计数大于读数计数阈值时，确定对应于对照基因参考序列的过程控制靶标的存在，其中家族计数为对应于与对照基因参考序列比对的共有序列读数的家族的数目，并且读数计数为对应家族的序列读数数目的总和。融合可以包括基因间融合，并且靶向融合参考序列可以包括用于在融合断点处融合两个基因的参考序列。融合可以包括基因内融合，并且靶向融合参考序列可以包括用于在相同基因内的融合断点处融合两个外显子的参考序列。共有序列比对的一部分可以对应于与相同基因的一个或多个野生型参考序列比对的共有序列读数。对于基因内融合，检测融合的步骤还可以进一步包括，计算基因内融合的读数计数与平均读数计数的比率，该平均读数计数对应于与相同基因的野生型参考序列比对的共有序列读数。对于基因内融合，用于检测融合的步骤还可以进一步包括计算基因内融合的读数计数与读数计数总和的比率，所述读数计数总和对应于与野生型参考序列比对的共有序列读数和与相同基因的靶向融合参考序列比对的共有序列读数。共有序列读数的一部分可以部分地映射到靶向融合参考序列。检测融合的步骤可以进一步包括基于部分映射的共有序列读数检测非靶向融合。

[0110] 根据示例性实施方式，提供了一种包括指令的非暂时性机器可读存储介质，当由处理器执行时，所述指令引起所述处理器执行一种用于压缩分子标记的核酸序列数据以供融合检测的方法，所述方法包括：(a)接收多个核酸序列读数和多个序列读数家族的多个序列比对，其中每个序列读数与分子标签序列相关，所述分子标签序列鉴定由核酸样品中的特定多核苷酸分子产生的序列读数家族，每个家族具有多个序列读数，其中所述序列比对的一部分对应于映射到靶向融合参考序列的序列读数；(b)基于对应于所述家族的所述序列读数的流动空间信号测量值，确定每个序列读数家族的共有序列读数；(c)确定每个序列读数家族的共有序列比对，其中所述共有序列比对的一部分对应于与所述靶向融合参考序列比对的所述共有序列读数；(d)产生包含共有压缩数据的压缩数据结构，所述共有压缩数据包括每个家族的共有序列读数和共有序列比对；以及(e)使用来自所述压缩数据结构的所述共有序列读数和所述共有序列比对，来检测融合。对于双向测序，其中正向共有序列读数和反向共有序列读数在单独的家族中，其中正向家族与第一前缀标签和第一后缀标签相关，反向家族与第二前缀标签和第二后缀标签相关，所述方法还包括当第二前缀标签和第二后缀标签的反向互补物与第一前缀标签和第一后缀标签匹配时，组合正向家族和反向家族，以形成具有压缩数据结构的一个共有序列读数的组合家族。检测融合的步骤可以进一步包括基于共有序列读数与靶向融合参考序列的共有序列比对的特征，鉴定合格的共有序列读数。所述特征可以包括同源性特征、映射质量特征和跨断点的特征。鉴定合格的共有序列读数的步骤可以进一步包括，确定与靶向融合参考序列比对的共有序列读数是否跨越靶向融合参考序列的融合断点。鉴定合格的共有序列读数的步骤可以进一步包括，确定共有序列读数分别与靶向融合参考序列的第一和第二配偶体序列的第一和第二同源性水平是否大于或等于最小同源性阈值。鉴定合格的共有序列读数的步骤可以进一步包括，确定分别在靶向融合参考序列的第一和第二配偶体序列内的共有序列读数的第一和第二映射质量值是否大于或等于映射质量阈值。鉴定合格的共有序列读数的步骤可以进一步包括通过计算与配偶体序列匹配的共有序列读数中的匹配碱基的数目和与配偶体序列重叠的共有序列读数中的重叠碱基的数目的比率，来确定映射质量值。鉴定合格的共有序列读数的步骤可以进一步包括，确定对应于与靶向融合参考序列比对的合格的共有序列读数的家族的

数目是否大于或等于最小分子计数阈值。鉴定合格的共有序列读数的步骤可以进一步包括,确定读数计数是否大于或等于最小读数计数阈值,其中读数计数为家族的序列读数数目的总和,这些家族的序列读数对应于与靶向融合参考序列比对的合格的共有序列读数。对应于序列读数的序列比对的一部分可以映射到对照基因参考序列,其中共有压缩数据还包括对应于对照基因参考序列的共有序列读数和共有序列比对。所述方法可以进一步包括当家族计数大于最小分子计数阈值并且读数计数大于读数计数阈值时,确定对应于对照基因参考序列的过程控制靶标的存在,其中家族计数为对应于与对照基因参考序列比对的共有序列读数的家族的数目,并且读数计数为对应家族的序列读数数目的总和。融合可以包括基因间融合,并且靶向融合参考序列可以包括用于在融合断点处融合两个基因的参考序列。融合可以包括基因内融合,并且靶向融合参考序列可以包括用于在相同基因内的融合断点处融合两个外显子的参考序列。共有序列比对的一部分可以对应于与相同基因的一个或多个野生型参考序列比对的共有序列读数。对于基因内融合,检测融合的步骤还可以进一步包括计算基因内融合的读数计数与平均读数计数的比率,该平均读数计数对应于与相同基因的野生型参考序列比对的共有序列读数。对于基因内融合,用于检测融合的步骤还可以进一步包括计算基因内融合的读数计数与读数计数总和的比率,所述读数计数总和对应于与野生型参考序列比对的共有序列读数和与相同基因的靶向融合参考序列比对的共有序列读数。共有序列读数的一部分可以部分地映射到靶向融合参考序列。检测融合的步骤可以进一步包括基于部分映射的共有序列读数检测非靶向融合。

[0111] 根据示例性实施方式,提供了一种用于压缩分子标记的核酸序列数据以供融合检测的系统,所述系统包括:机器可读存储器和与所述存储器通信的处理器,其中所述处理器被配置成执行机器可读指令,当由处理器执行时,所述指令引起所述系统执行一种方法,所述方法包括:(a)接收多个核酸序列读数和多个序列读数家族的多个序列比对,其中每个序列读数与分子标签序列相关,所述分子标签序列鉴定由核酸样品中的特定多核苷酸分子产生的序列读数家族,每个家族具有多个序列读数,其中所述序列比对的一部分对应于映射到靶向融合参考序列的序列读数;(b)基于对应于所述家族的所述序列读数的流动空间信号测量值,确定每个序列读数家族的共有序列读数;(c)确定每个序列读数家族的共有序列比对,其中所述共有序列比对的一部分对应于与所述靶向融合参考序列比对的所述共有序列读数;(d)产生包含共有压缩数据的压缩数据结构,所述共有压缩数据包括每个家族的共有序列读数和共有序列比对;以及(e)使用来自所述压缩数据结构的所述共有序列读数和所述共有序列比对,来检测融合。对于双向测序,其中正向共有序列读数和反向共有序列读数在单独的家族中,其中正向家族与第一前缀标签和第一后缀标签相关,反向家族与第二前缀标签和第二后缀标签相关,所述方法还包括当第二前缀标签和第二后缀标签的反向互补物与第一前缀标签和第一后缀标签时匹配时,组合正向家族和反向家族,以形成具有压缩数据结构的一个共有序列读数的组合家族。检测融合的步骤可以进一步包括基于共有序列读数与靶向融合参考序列的共有序列比对的特征,鉴定合格的共有序列读数。所述特征可以包括同源性特征、映射质量特征和跨断点的特征。鉴定合格的共有序列读数的步骤可以进一步包括,确定与靶向融合参考序列比对的共有序列读数是否跨越靶向融合参考序列的融合断点。鉴定合格的共有序列读数的步骤可以进一步包括,确定共有序列读数分别与靶向融合参考序列的第一和第二配偶体序列的第一和第二同源性水平是否大于或等于最

小同源性阈值。鉴定合格的共有序列读数的步骤可以进一步包括,确定分别在靶向融合参考序列的第一和第二配偶体序列内的共有序列读数的第一和第二映射质量值是否大于或等于映射质量阈值。鉴定合格的共有序列读数的步骤可以进一步包括,通过计算与配偶体序列匹配的共有序列读数中的匹配碱基的数目和与配偶体序列重叠的共有序列读数中的重叠碱基的数目的比率,来确定映射质量值。鉴定合格的共有序列读数的步骤可以进一步包括,确定对应于与靶向融合参考序列比对的合格的共有序列读数的家族的数目是否大于或等于最小分子计数阈值。鉴定合格的共有序列读数的步骤可以进一步包括,确定读数计数是否大于或等于最小读数计数阈值,其中读数计数为家族的序列读数数目的总和,这些家族的序列读数对应于与靶向融合参考序列比对的合格的共有序列读数。对应于序列读数的序列比对的一部分可以映射到对照基因参考序列,其中共有压缩数据还包括对应于对照基因参考序列的共有序列读数和共有序列比对。所述方法可以进一步包括当家族计数大于最小分子计数阈值并且读数计数大于读数计数阈值时,确定对应于对照基因参考序列的过程控制靶标的存在,其中家族计数为对应于与对照基因参考序列比对的共有序列读数的家族的数目,并且读数计数为对应家族的序列读数数目的总和。融合可以包括基因间融合,并且靶向融合参考序列可以包括用于在融合断点处融合两个基因的参考序列。融合可以包括基因内融合,并且靶向融合参考序列可以包括用于在相同基因内的融合断点处融合两个外显子的参考序列。共有序列比对的一部分可以对应于与相同基因的一个或多个野生型参考序列比对的共有序列读数。对于基因内融合,检测融合的步骤还可以进一步包括,计算基因内融合的读数计数与平均读数计数的比率,该平均读数计数对应于与相同基因的野生型参考序列比对的共有序列读数。对于基因内融合,用于检测融合的步骤还可以进一步包括计算基因内融合的读数计数与读数计数总和的比率,所述读数计数总和对应于与野生型参考序列比对的共有序列读数和与相同基因的靶向融合参考序列比对的共有序列读数。共有序列读数的一部分可以部分地映射到靶向融合参考序列。检测融合的步骤可以进一步包括基于部分映射的共有序列读数检测非靶向融合。

[0112] 核酸序列数据可以使用各种技术、平台或科技产生,包括但不限于:毛细管电泳、微阵列、基于连接的系统、基于聚合酶的系统、基于杂交的系统、直接或间接核苷酸鉴定系统、焦磷酸测序、基于离子或pH的检测系统、基于电子签名的系统等。

[0113] 核酸测序平台(如核酸测序仪)的各种实施方式可以包括如图10的框图中所显示的组件。根据各种实施方式,测序仪器1200可以包括流体递送和控制单元1202、样品处理单元1204、信号检测单元1206以及数据采集、分析和控制单元1208。用于下一代测序的仪器、试剂、文库和方法的各种实施方式描述于美国专利申请公开号2009/0127589和2009/0026082中。仪器1200的各种实施方式可提供自动测序,其可用于并行地(例如基本上同时地)从多个序列收集序列信息。

[0114] 在各种实施方式中,流体递送和控制单元1202可以包括试剂递送系统。试剂递送系统可以包括用于储存各种试剂的试剂储器。试剂可以包括基于RNA的引物、正向/反向DNA引物、用于连接测序(ligation sequencing)的寡核苷酸混合物、用于边合成边测序的核苷酸混合物、任选的ECC寡核苷酸混合物、缓冲液、洗涤试剂、阻断试剂、剥离试剂等。另外,试剂递送系统可以包括移液系统或连续流动系统,其将样品处理单元与试剂储器连接。

[0115] 在各种实施方式中,样品处理单元1204可以包括样品室,诸如流动池、底物、微阵

列、多孔托盘等。样品处理单元1204可以包括多个线路、多个通道、多个井或基本上同时处理多个样品集的其他装置。另外,样品处理单元可以包括多个样品室,以使得能够同时处理多个运行。在特定实施方式中,系统可以在基本上同时处理另一个样品室的同时对一个样品室执行信号检测。另外,样品处理单元可以包括用于移动或操纵样品室的自动化系统。

[0116] 在各种实施方式中,信号检测单元1206可以包括成像或检测传感器。例如,成像或检测传感器可以包括CCD,CMOS,离子或化学传感器,例如覆盖CMOS或FET的离子敏感层,电流或电压检测器等。信号检测单元1206可以包括激发系统,以使诸如荧光染料的探针发射信号。激发系统可以包括照明源,例如弧光灯、激光器、发光二极管(LED)等。在特定实施方式中,信号检测单元1206可包含用于将光从照明源传输到样品或从样品传输到成像或检测传感器的光学器件。或者,信号检测单元1206可提供用于检测的基于电子或非光子的方法,且因此不包括照明源。在各种实施方式中,当在测序反应期间产生可检测信号或物质时,可发生基于电子的信号检测。例如,信号可以通过释放的副产物或一部分,例如释放的离子,如氢离子,与离子或化学敏感层相互作用而产生。在其他实施方式中,可检测信号作为例如用于焦磷酸测序(参见,例如,美国专利申请公开号2009/0325145)酶级联的结果而产生,其中焦磷酸盐通过聚合酶的碱基掺入而产生,所述聚合酶在腺苷5'-磷酸硫酸的存在下进一步与ATP硫酸化酶反应以产生ATP,其中产生的ATP可在萤光素酶介导的反应中消耗以产生化学发光信号。在另一个实例中,当核酸通过纳米孔时,可以检测电流的变化,而不需要照明源。

[0117] 在各种实施方式中,数据采集分析和控制单元1208可以监测各种系统参数。系统参数可以包括仪器1200的各个部分的温度,诸如样品处理单元或试剂储器,各种试剂的体积,各种系统子部件的状态,诸如操纵器、步进电机、泵等,或其任何组合。

[0118] 本领域技术人员将理解,仪器1200的各种实施方式可用于实践各种测序方法,该方法包括基于连接的方法、合成测序、单分子方法、纳米孔测序和其他测序技术。

[0119] 在各种实施方式中,测序仪器1200可以确定核酸的序列,例如多核苷酸或寡核苷酸。核酸可以包括DNA或RNA,并且可为单链的,例如ssDNA和RNA,或者为双链的,例如dsDNA或RNA/cDNA对。在各种实施方式中,核酸可以包括或衍生自片段文库、配对物文库、ChIP片段等。在特定实施方式中,测序仪器1200可从单个核酸分子或从一组实质上相同的核酸分子得到序列信息。

[0120] 在各种实施方式中,测序仪器1200可以各种不同的输出数据文件类型/格式输出核酸测序读数数据,包括但不限于:*.fasta、*.csfasta、*seq.txt、*qseq.txt、*.fastq、*.sff、*prb.txt、*.sms、*srs和/或*.qv。

[0121] 根据各种示例性实施方式,可以使用适当配置的和/或编程的硬件和/或软件元件来执行或实施上述讨论的教导和/或示例性实施方式中的任何一个或多个的一个或多个特征。确定实施方式是否使用硬件和/或软件元件来实现可以基于任何数量的因素,诸如期望的计算速率、功率水平、耐热性、处理周期预算、输入数据速率、输出数据速率、存储器资源、数据总线速度等,以及其他设计或性能约束。

[0122] 硬件元件的示例可以包括:处理器,微处理器,经由本地接口电路通信地耦合的一个或多个输入和/或一个或多个输出(I/O)设备(或外围设备),电路元件(例如,晶体管、电阻器、电容器、电感器等),集成电路,专用集成电路(ASIC),可编程逻辑设备(PLD),数字信

号处理器(DSP),现场可编程门阵列(FPGA),逻辑门,寄存器,半导体设备,芯片,微芯片,芯片组等。本地接口可以包括例如一个或多个总线或其他有线或无线连接、控制器、缓冲器(高速缓存)、驱动器、转发器和接收器等,以允许硬件组件之间的适当通信。处理器是用于执行软件、特别是存储在存储器中的软件的硬件设备。处理器可为任何定制的或商业上可得到的处理器、中央处理单元(CPU)、与计算机相关的若干处理器中的辅助处理器、基于半导体的微处理器(例如,以微芯片或芯片组的形式)、宏处理器或一般地用于执行软件指令的任何设备。处理器也可以表示分布式处理体系结构。I/O设备可以包括输入设备,例如键盘、鼠标、扫描仪、麦克风、触摸屏、用于各种医疗设备和/或实验室仪器的接口、条形码读取器、指示笔、激光阅读器、射频设备阅读器等。此外,I/O设备可以进一步包括输出设备,例如打印机、条形码打印机、显示器等等;最后,I/O设备还可以包括作为输入和输出进行通信的设备,例如调制器/解调器(调制解调器;用于访问另一设备、系统或网络)、射频(RF)或其他收发器、电话接口、桥接器、路由器等。

[0123] 软件的示例可以包括软件组件、程序、应用、计算机程序、应用程序、系统程序、机器程序、操作系统软件、中间件、固件、软件模块、例程、子程序、子例程、方法、工序、软件接口、应用程序接口(API)、指令集、计算代码、计算机代码、代码段、计算机代码段、字、值、符号或其任意组合。存储器中的软件可以包括一个或多个单独的程序,其可以包括用于实施逻辑功能的可执行指令的有序列表。存储器中的软件可以包括根据本教导的用于鉴定数据流的系统以及任何适当的定制的或商业上可得到的操作系统(O/S),其可以控制诸如系统的其他计算机程序的执行,并且提供调度、输入-输出控制、文件和数据管理、存储器管理、通信控制等。

[0124] 根据各种示例性实施方式,可以使用适当配置的和/或编程的非暂时性机器可读介质或物品来执行或实施上述讨论的教导和/或示例性实施方式中的任何一个或多个的一个或多个特征,所述非暂时性机器可读介质或物品可以存储指令或指令集,所述指令或指令集如果由机器执行,则可以引起机器执行根据示例性实施方式的方法和/或操作。这样的机器可以包括例如任何合适的处理平台、计算平台、计算设备、处理设备、计算系统、处理系统、计算机、处理器、科学或实验室仪器等,并且可以使用硬件和/或软件的任何合适的组合来实施。机器可读介质或物品可以包括例如任何适当类型的存储器单元、存储器设备、存储器物品、存储器介质、存储设备、存储物品、存储介质和/或存储单元,例如存储器、可移动或不可移动介质、可擦除或不可擦除介质、可写或可重写介质、数字或模拟介质、硬盘、软盘、只读存储器光盘(CD-ROM)、可记录光盘(CD-R)、可重写光盘(CD-RW)、光盘、磁介质、磁光介质、可移动存储卡或盘、各种类型的数字多功能盘(DVD)、磁带、盒式磁带等,包括适合用在计算机中的任何介质。存储器可以包括易失性(volatile)存储器元件(例如,随机存取存储器(RAM,诸如DRAM、SRAM、SDRAM等))和非易失性存储器元件(例如,ROM、EPROM、EEROM、闪存、硬盘驱动器、磁带、CDROM等)中的任何一个或组合。此外,存储器可以结合电子、磁、光和/或其他类型的存储介质。存储器可以具有分布式体系结构,其中各种组件处于彼此远离的位置,但是仍然由处理器访问。指令可以包括使用任何合适的高级、低级、面向对象、可视、编译和/或解释性程序语言实施的任何合适类型的代码,诸如源代码、编译代码、解释代码、可执行代码、静态代码、动态代码、加密代码等。

[0125] 根据各种示例性实施方式,可以至少部分地使用分布式、集群、远程或云计算资源

来执行或实施上述讨论的教导和/或示例性实施方式中的任何一个或多个的一个或多个特征。

[0126] 根据各种示例性实施方式,可以使用源程序、可执行程序(目标代码)、脚本或包括要执行的指令集的任何其他实体,来执行或实施上述讨论的教导和/或示例性实施方式中的任何一个或多个的一个或多个特征。当使用源程序时,该程序可以经由可以包括在存储器中或不包括在存储器中的编译器、汇编器、解释器等来翻译,以便与O/S结合适当地操作。指令可以使用(a)面向对象的编程语言,其具有数据和方法类,或(b)程序性编程语言,其具有例程、子程序和/或子例程,其可以包括例如C、C++、R、Pascal、Basic、Fortran、Cobol、Perl、Java和Ada。

[0127] 根据各种示例性实施方式,一个或多个上述讨论的示例性实施方式可以包括向用户接口设备、计算机可读存储介质、本地计算机系统或远程计算机系统传输、显示、存储、打印或输出与可能已经由这样的示例性实施方式产生、访问或使用的任何信息、信号、数据和/或中间或最终结果有关的信息。例如,这种传输的、显示的、存储的、打印的或输出的信息可以采取可搜索的和/或可过滤的运行和报告列表、图片、表格、图表、图形、电子表格、相关性、序列及其组合的形式。

[0128] 虽然在此已经显示和描述了本发明的优选实施方式,但是对于本领域技术人员来说显而易见的是,这些实施方式仅作为示例提供。在不背离本发明的情况下,本领域技术人员可以想到许多变化、改变和替换。应当理解,在实践本发明时可以采用本文所述的本发明实施方式的各种替代方案。下面的权利要求书旨在定义本发明的范围,并且这些权利要求的范围内的方法和结构及其等同物由此被覆盖。

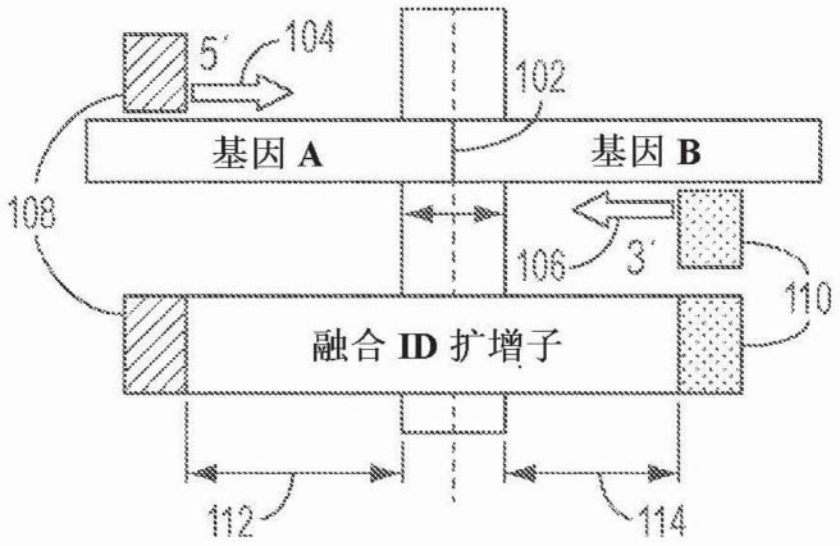

5'
104
102
基因 A
基因 B
108
106
3'
110
融合 ID 扩增子
112
114

图1

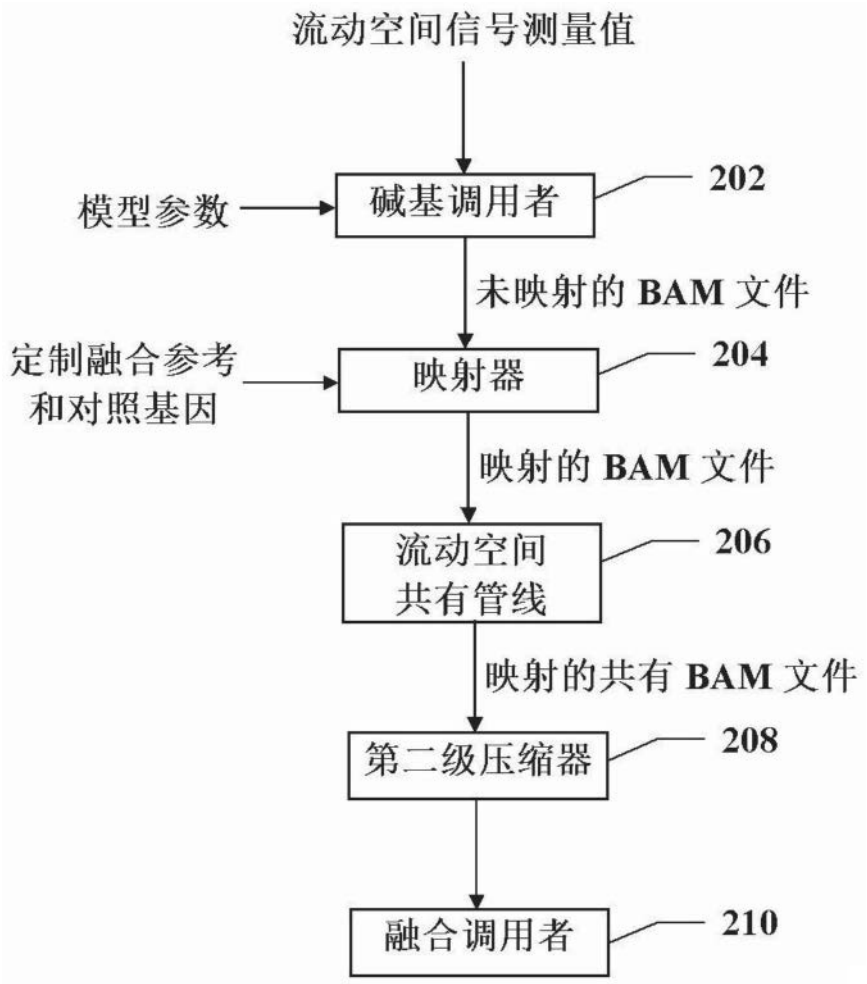

流动空间信号测量值
模型参数
碱基调用者
202
未映射的 BAM 文件
定制融合参考
和对照基因
映射器
204
映射的 BAM 文件
流动空间
共有管线
206
映射的共有 BAM 文件
第二级压缩器
208
融合调用者
210

图2

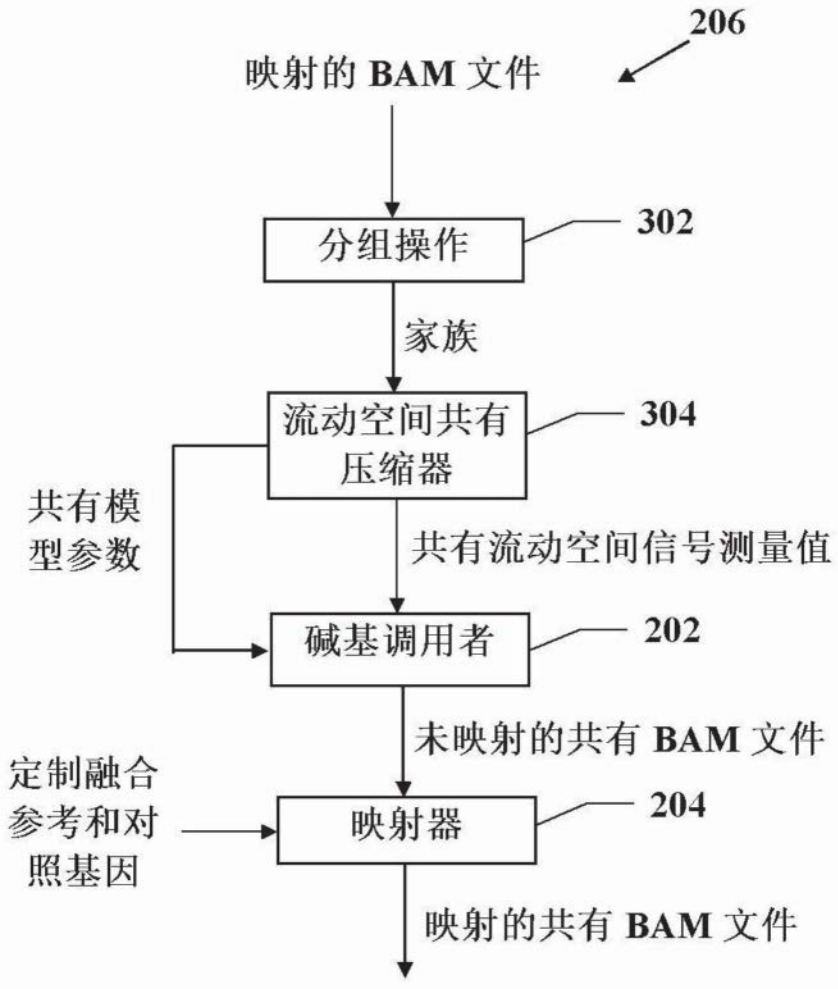

206
映射的 BAM 文件
分组操作
302
家族
流动空间共有压缩器
304
共有模型参数
共有流动空间信号测量值
碱基调用者
202
未映射的共有 BAM 文件
定制融合参考和对照基因
映射器
204
映射的共有 BAM 文件

图3

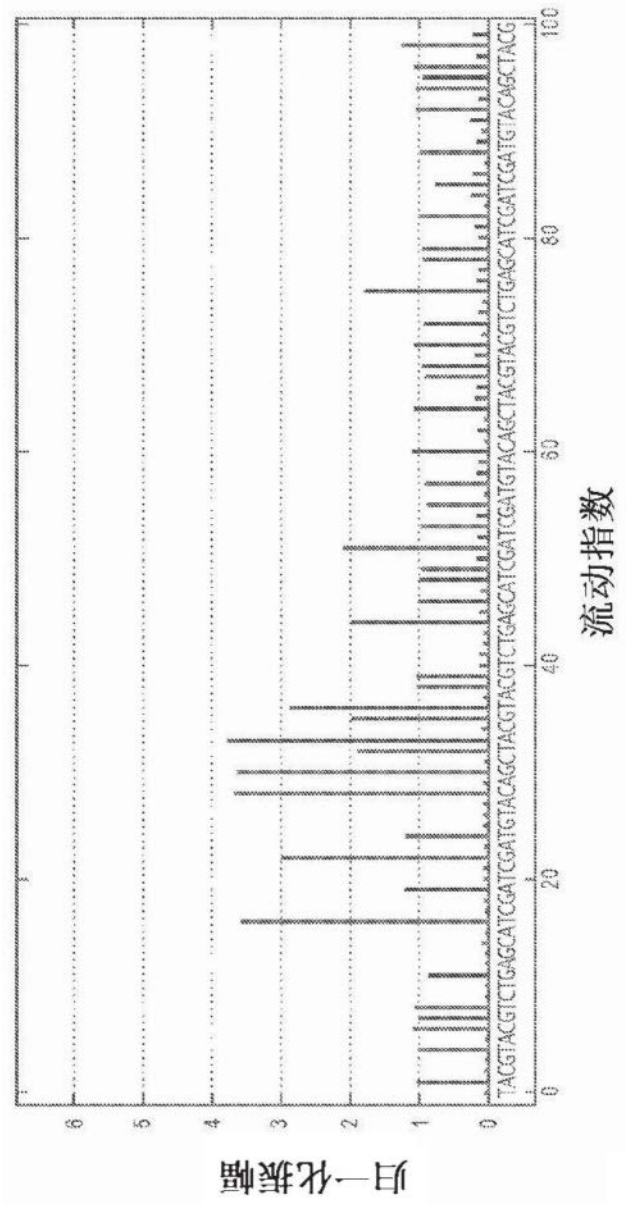

归一化振幅
流动指数
TACGTACGTCTGAGCATCGATCGATGTACAGCTACGTACGTCTGAGCATCGATCGATGTACAGCTACGTACGTCTGAGCATCGATCGATGTACAGCTACG
0
1
2
3
4
5
6
0
20
40
60
80
100

图4

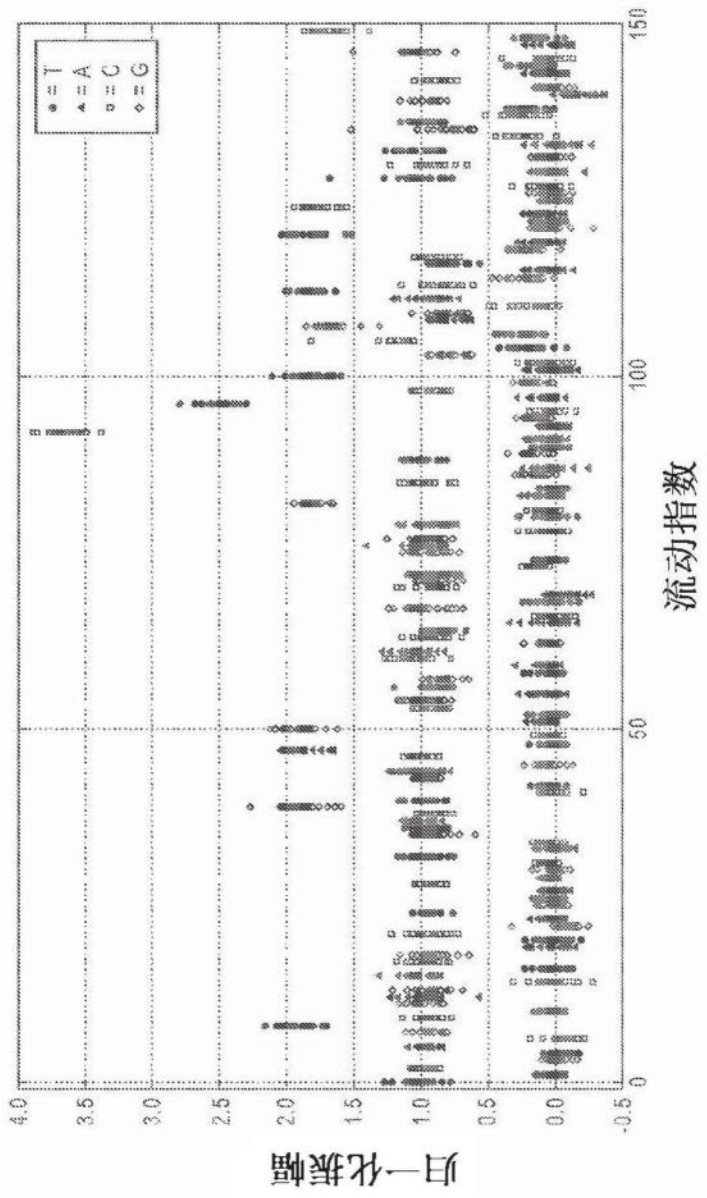


图5

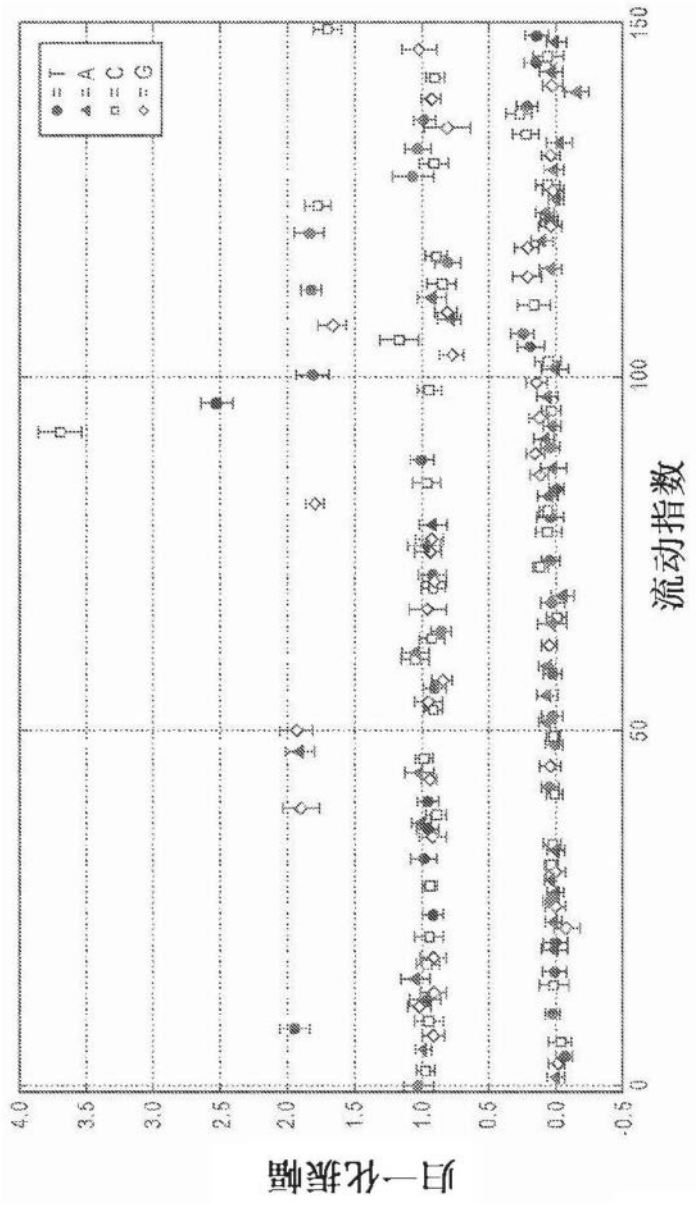


图6

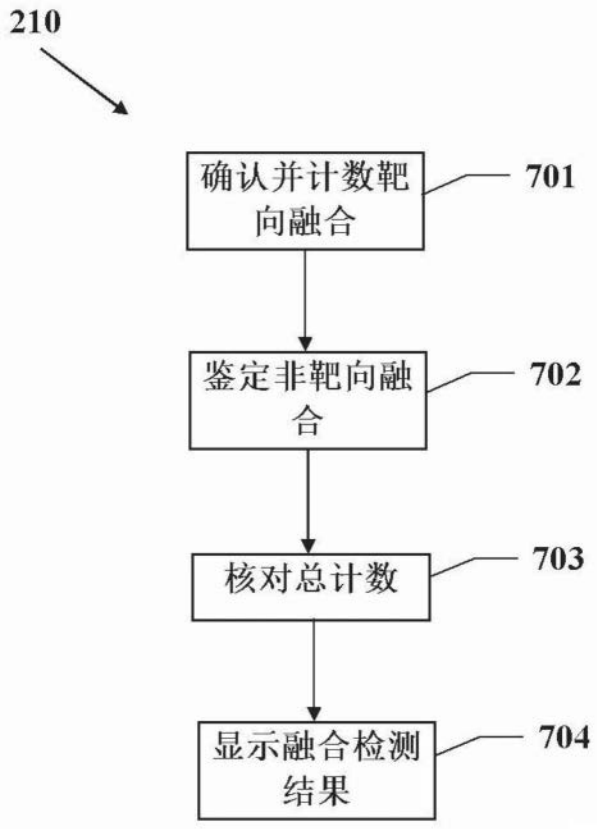

210
确认并计数靶向融合
701
鉴定非靶向融合
702
核对总计数
703
显示融合检测结果
704

图7

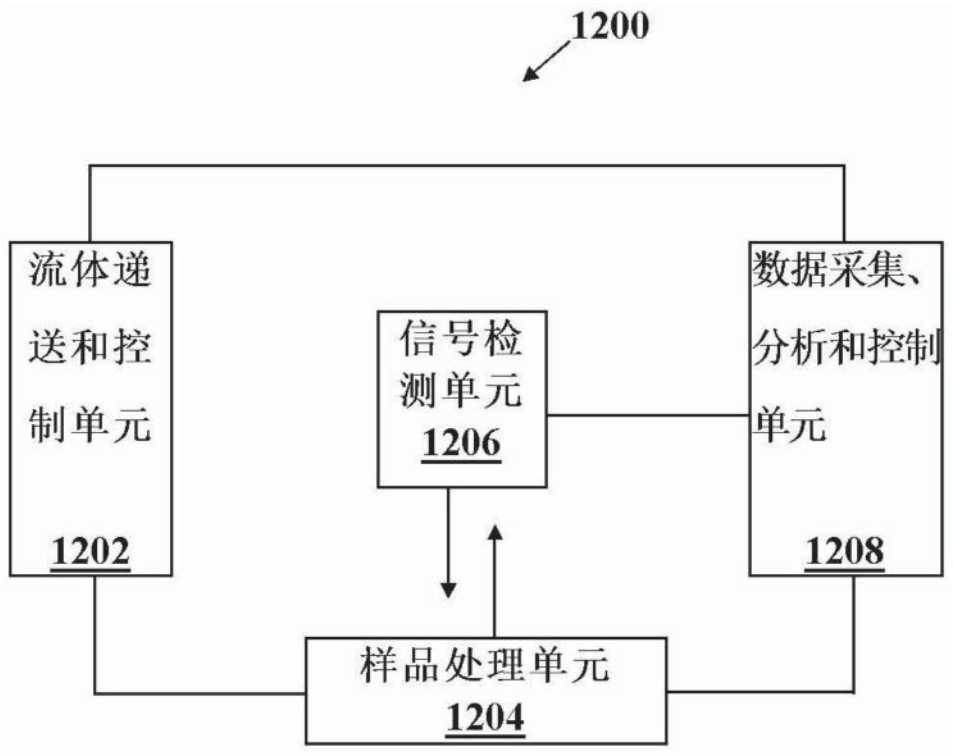

1200
流体递送和控制单元
1202
信号检测单元
1206
数据采集、分析和控制单元
1208
样品处理单元
1204

图8