(19)中华人民共和国国家知识产权局

(12)发明专利

(10)授权公告号 CN 106575322 B
(45)授权公告日 2019.06.18

(21)申请号 201580042592.6
(22)申请日 2015.06.26
(65)同一申请的已公布的文献号
申请公布号 CN 106575322 A
(43)申请公布日 2017.04.19
(30)优先权数据
62/017,589 2014.06.26 US
(85)PCT国际申请进入国家阶段日
2017.02.08
(86)PCT国际申请的申请数据
PCT/US2015/038175 2015.06.26
(87)PCT国际申请的公布数据
WO2015/200891 EN 2015.12.30
(73)专利权人 10X基因组学有限公司
地址 美国加利福尼亚州
(72)发明人 M·史诺-莱文 I·麦克卡伦
(74)专利代理机构 深圳市百瑞专利商标事务所（普通合伙）44240
代理人 金辉
(51)Int.Cl.
G16B 30/20(2019.01)
(56)对比文件
WO 0218412 A1,2002.03.07,
CN 1358868 A,2002.07.17,
CN 101812530 A,2010.08.25,
CN 103667491 A,2014.03.26,
CN 101910399 A,2010.12.08,
CN 102365367 A,2012.02.29,
US 7566813 B2,2009.07.28,
审查员 葛晓倩

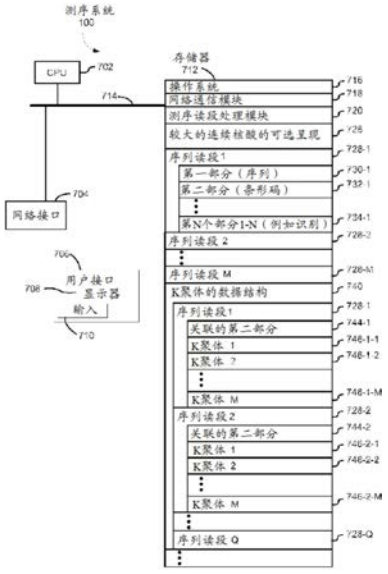
权利要求书7页 说明书28页 附图30页

(54)发明名称

核酸序列装配的方法和系统

(57)摘要

提供了用于分析遗传序列数据的方法、过程、且特别是计算机执行的过程和计算机程序产品。所述过程和产品用于将较短的核酸序列数据装配成较长连接的并且优选连续的遗传构建体，包括大的重叠群、染色体和整个基因组。



CN 106575322 B

1.一种装配核酸序列读段的测序方法,所述测序方法包括:

在计算机系统上,所述计算机系统具有一个或多个处理器,以及存储用于由所述一个或多个处理器执行的一个或多个程序的存储器:

获得源自较大的连续核酸的多个序列读段,其中源自所述较大的连续核酸的共同片段的两个或更多个序列读段包括共同条形码序列,

鉴定所述多个序列读段中包含重叠序列和共同条形码序列两者的序列读段的第一子集;并且

比对序列读段的所述第一子集以提供连续的线性核酸序列。

2.根据权利要求1所述的方法,所述方法还包括用序列读段的多个不同子集重复所述鉴定和比对步骤,以提供多个连续的线性核酸序列。

3.根据权利要求2所述的方法,所述方法还包括在所述较大的连续核酸内的序列背景中将所述多个不同的连续线性核酸序列排序。

4.根据权利要求3所述的方法,其中所述排序包括相对于参考序列定位所述多个不同的连续线性核酸序列。

5.根据权利要求3所述的方法,其中所述排序包括:

鉴定一个或多个序列读段,所述序列读段包含与第一连续线性核酸序列共同的条形码序列,还包括与第二连续线性核酸序列的重叠序列;并且

将所述第一和第二连续线性核酸鉴定为结构连接的。

6.一种将核酸序列读段装配成较大的连续序列的方法,所述方法包括:

在计算机系统上,所述计算机系统具有一个或多个处理器,以及存储用于由所述一个或多个处理器执行的一个或多个程序的存储器:

获得源自较大的连续核酸的多个序列读段,

从所述多个序列读段中的重叠序列读段集合中鉴定第一子序列;

基于在相邻序列上与所述第一子序列共同的条形码序列的存在,将所述第一子序列延伸至一个或多个相邻或重叠序列;并且

提供包含所述第一子序列和所述一个或多个相邻序列的线性核酸序列。

7.一种测序方法,所述测序方法包括在计算机系统上,所述计算机系统具有一个或多个处理器,以及存储用于由所述一个或多个处理器执行的一个或多个程序的存储器:

(A)获得多个序列读段,其中

所述多个序列读段包括序列读段的多个集合,

在序列读段集合中的每个相应的序列读段包含(i)对应于较大的连续核酸的子集的第一部分和(ii)形成标识符的共同第二部分,所述标识符不依赖于所述较大的连续核酸的序列,并且鉴定多个分区中形成所述相应的序列读段的分区,并且

在所述多个分区中的分区中形成所述序列读段的多个集合中的序列读段的每个相应集合,并且每个分区包含所述较大的连续核酸的一个或多个片段,所述较大的连续核酸用作所述分区中的每个相应序列读段的模板;

(B)为在所述多个序列读段中的每个序列读段创建相应的k聚体集合,其中

所述k聚体集合共同包含多个k聚体,

保留所述多个k聚体中的每个k聚体的所述序列读段的所述标识符,

k小于所述多个序列读段中的所述序列读段的平均长度,并且

每个相应的k聚体集合包括所述对应序列读段的第一部分的长度k的可能k聚体的至少80%;

(C)对于所述多个k聚体中的每个相应的k聚体,追踪含有所述相应k聚体的所述多个序列读段中的每个序列读段的身份和含有所述序列读段的所述序列读段集合的所述标识符;

(D)将所述多个k聚体绘制为包括通过多个有向弧连接的多个节点的图,其中,

每个节点包含长度k的多个k聚体中具有k-1重叠的k聚体的不间断集合,

每个弧在所述多个节点中将源节点连接到目标节点,

源节点的最终k聚体与目标节点的初始k聚体具有k-1重叠,并且

第一源节点具有在所述多个节点中的第一目标节点和第二目标节点二者的第一有向弧;并且

(E)测定是否将所述第一源节点与所述第一目标节点或所述第二目标节点合并,以便得到更可能代表所述较大的连续核酸的一部分的重叠群序列,其中所述重叠群序列包含(i)所述第一源节点和(ii)所述第一目标节点和所述第二目标节点中的一个,其中所述测定至少使用所述第一源节点、所述第一目标节点和所述第二目标节点中的k聚体的序列读段的标识符。

8.根据权利要求7所述的测序方法,其中,

所述第一源节点和所述第一目标节点是所述图中包括除所述第一源节点和所述第一目标节点之外的一个或多个额外节点的第一路径的一部分,

所述第一源节点和所述第二目标节点是所述图中包括除所述第一源节点和所述第二目标节点之外的一个或多个额外节点的第二路径的一部分,

所述测定(E)包括通过相对于所述第二路径的第一部分的节点的k聚体和所述第二路径的第二部分的节点的k聚体之间共享的标识符的数目评估所述第一路径的第一部分的节点的k聚体和所述第一路径的第二部分的节点的k聚体之间共享的标识符的数目测定所述第一路径是否比所述第二路径更可能代表所述较大的连续核酸。

9.根据权利要求7所述的测序方法,其中

所述第一源节点和所述第一目标节点是所述图中包括除所述第一源节点和所述第一目标节点之外的一个或多个额外节点的第一路径的一部分,

所述第一源节点和所述第二目标节点是所述图中包括除所述第一源节点和所述第二目标节点之外的一个或多个额外节点的第二路径的一部分,

当所述第一路径具有比所述第二路径更高的平均覆盖时,所述测定(E)相对于所述第二路径向上加权所述第一路径。

10.根据权利要求7所述的测序方法,其中

所述第一源节点和所述第一目标节点是所述图中包括除所述第一源节点和所述第一目标节点之外的一个或多个额外节点的第一路径的一部分,

所述第一源节点和所述第二目标节点是所述图中包括除所述第一源节点和所述第二目标节点之外的一个或多个额外节点的第二路径的一部分,

当所述第一路径代表比所述第二路径更长的所述较大的连续核酸序列的连续部分时,测定(E)相对于所述第二路径向上加权所述第一路径。

11.根据权利要求8-10中任一项所述的测序方法，其中所述第一源节点中的第一k聚体存在于所述多个序列读段的子多个中，并且在所述子多个序列读段中的每个序列读段的身份通过所述测定(E)以测定所述第一路径是否比所述第二路径更可能代表所述较大的连续核酸序列。

12.根据权利要求7所述的测序方法，其中

所述多个分区中的分区包含具有所述共同的第二部分的至少1000个分子，并且

所述至少1000个分子中的每个分子还包含与所述较大的连续核酸的至少一部分互补的引物序列。

13.根据权利要求7所述的测序方法，其中

所述多个分区中的分区包括具有所述共同的第二部分的至少1000个分子，并且

所述至少1000个分子中的每个分子还包含引物位点和与所述较大的连续核酸的一部分互补的半随机N聚体引发序列。

14.根据权利要求7所述的测序方法，其中所述多个分区中的分区中的所述较大的连续核酸的所述一个或多个片段的长度大于50千碱基。

15.根据权利要求7所述的测序方法，其中所述多个分区中的分区中的所述较大的连续核酸的所述一个或多个片段的长度在20千碱基和200千碱基之间。

16.根据权利要求7所述的测序方法，其中所述多个分区中的分区中的所述较大的连续核酸的所述一个或多个片段由所述较大的连续核酸的在1个和500个之间的不同片段组成。

17.根据权利要求7所述的测序方法，其中所述多个分区中的分区中的所述较大的连续核酸的所述一个或多个片段由所述较大的连续核酸的在5个和100个之间的片段组成。

18.根据权利要求7所述的测序方法，其中从小于5纳克的核酸或核糖核酸获得所述多个序列读段。

19.根据权利要求7所述的测序方法，其中在所述序列读段集合中的每个相应序列读段的所述第二部分中的所述标识符编码选自以下的共同值：集合{1,…,1024}、集合{1,…,4096}、集合{1,…,16384}、集合{1,…,65536}、集合{1,…,262144}、集合{1,…,1048576}、集合{1,…,4194304}、集合{1,…,16777216}、集合{1,…,67108864}、或集合{1,…,$1x10^{12}$}。

20.根据权利要求7所述的测序方法，其中所述标识符是N聚体，并且N是选自集合{4,…,20}的整数。

21.根据权利要求7所述的测序方法，其中所述多个序列读段的平均序列读段长度在40个碱基和200个碱基之间。

22.根据权利要求7所述的测序方法，其中所述多个序列读段的平均序列读段长度在60个碱基和140个碱基之间。

23.根据权利要求7所述的测序方法，其中

所述多个序列读段共同为所述较大的连续核酸提供至少15X覆盖，

所述多个k聚体中超过10％的所述k聚体来自所述多个序列读段中的超过一个序列读段，并且

保留对于由超过一个序列读段代表的每个k聚体的每个序列读段的所述标识符。

24.根据权利要求7所述的测序方法，其中

所述多个序列读段共同为所述较大的连续核酸提供至少25X覆盖，

所述多个k聚体的超过30％来自在所述多个序列读段中的超过一个序列读段，并且

保留对于由超过一个源序列代表的每个k聚体的每个序列读段的所述标识符。

25.根据权利要求7所述的测序方法，其中所述多个序列读段中的所述序列读段编码在75个碱基和125个碱基之间的所述较大的连续核酸，并且所述k值是在5和73之间的奇数整数。

26.根据权利要求7所述的测序方法，其中所述多个序列读段中的序列读段集合包括超过100个序列读段，并且所述超过100个序列读段的每个序列读段包括相同的共同第二部分。

27.根据权利要求7所述的测序方法，其中所述较大的连续核酸是染色体。

28.根据权利要求7所述的测序方法，其中所述较大的连续核酸的长度大于4000万个碱基对。

29.根据权利要求7所述的测序方法，其中创建所述相应的k聚体集合包括根据预定的k聚体长度散列所述多个序列读段中的每个序列读段，从而为所述多个序列读段中的每个序列读段创建相应的k聚体集合。

30.一种计算系统，所述计算系统包含：

一个或多个处理器；

存储器，所述存储器存储要由所述一个或多个处理器执行的一个或多个程序；

所述一个或多个程序包括指令，用于：

(A)获得多个序列读段，其中

所述多个序列读段包括序列读段的多个集合，

在序列读段集合中的每个相应的序列读段包含(i)对应于较大的连续核酸的子集的独特的第一部分和(ii)形成标识符的共同第二部分，所述标识符不依赖于所述较大的连续核酸的序列，并且鉴定多个分区中形成所述相应的序列读段的分区，并且

在所述多个分区中的分区中形成所述序列读段的多个集合中的序列读段的每个相应集合，并且每个分区包含所述较大的连续核酸的一个或多个片段，所述较大的连续核酸用作所述分区中的每个相应序列读段的模板；

(B)为在所述多个序列读段中的每个序列读段创建相应的k聚体集合，其中

所述k聚体集合共同包含多个k聚体，

保留所述多个k聚体中的每个k聚体的所述序列读段的所述标识符，

k小于所述多个序列读段中的所述序列读段的平均长度，并且

每个相应的k聚体集合包括所述对应序列读段的第一部分的长度k的可能k聚体的至少80％；

(C)对于所述多个k聚体中的每个相应的k聚体，追踪含有所述相应k聚体的所述多个序列读段中的每个序列读段的身份和含有所述序列读段的所述序列读段集合的所述标识符；

(D)将所述多个k聚体绘制为包括通过多个有向弧连接的多个节点的图，其中，

每个节点包含长度k的多个k聚体中具有k-1重叠的k聚体的不间断集合，

每个弧在所述多个节点中将源节点连接到目标节点，

源节点的最终k聚体与目标节点的初始k聚体具有k-1重叠，并且

第一源节点具有在所述多个节点中的第一目标节点和第二目标节点二者的第一有向弧;并且

(E)测定是否将所述第一源节点与所述第一目标节点或所述第二目标节点合并,以便得到更可能代表所述较大的连续核酸的一部分的重叠群序列,其中所述重叠群序列包含(i)所述第一源节点和(ii)所述第一目标节点和所述第二目标节点中的一个,其中所述测定至少使用所述第一源节点、所述第一目标节点和所述第二目标节点中的k聚体的序列读段的标识符。

31.一种测序方法,所述测序方法包括:

在计算机系统上,所述计算机系统具有一个或多个处理器,以及存储用于由所述一个或多个处理器执行的一个或多个程序的存储器:

(A)获得多个序列读段,其中

所述多个序列读段包括序列读段的多个集合,

在序列读段集合中的每个相应的序列读段包含(i)对应于较大的连续核酸的子集的独特的第一部分和(ii)形成标识符的共同第二部分,所述标识符不依赖于所述较大的连续核酸的序列,并且鉴定多个分区中形成所述相应的序列读段的分区,

在所述多个分区中的分区中形成所述序列读段的多个集合中的序列读段的每个相应集合,并且每个分区包含所述较大的连续核酸的一个或多个片段,所述较大的连续核酸用作所述分区中的每个相应序列读段的模板;

(B)为在所述多个序列读段中的每个序列读段创建相应的k聚体集合,其中

所述k聚体集合共同包含多个k聚体,

保留所述多个k聚体中的每个k聚体的所述序列读段的所述标识符,

k小于所述多个序列读段中的所述序列读段的平均长度,并且

每个相应的k聚体集合包括所述对应序列读段的所述第一部分的可能k聚体的至少80%;

(C)对于所述多个k聚体中的每个相应的k聚体,追踪含有所述相应k聚体的所述多个序列读段中的每个序列读段的身份;

(D)鉴定第一路径,其中所述第一路径包含在所述多个k聚体中具有k-1重叠的第一k聚体集合;

(E)鉴定第二路径,其中所述第二路径包含在所述多个k聚体中具有k-1重叠的第二k聚体集合,并且其中所述第一路径与所述第二路径相交,由此形成分支区段的集合,其包含:所述第一路径的左部分,所述第一路径的右部分,所述第二路径的左部分和所述第二路径的右部分;并且

(F)使用与所述分支区段的集合的k聚体有关的标识符,验证所述分支区段的连接性。

32.根据权利要求31所述的测序方法,其中使用(F)包括评估在所述分支区段的集合中的每个可能的分支区段对之间共享的标识符的数目。

33.根据权利要求31或32所述的测序方法,其中

所述多个分区中的分区包括具有所述共同的第二部分的至少1000个分子,并且

所述至少1000个分子中的每个分子包括与所述较大的连续核酸的至少一部分互补的引物序列。

34.根据权利要求31所述的测序方法，其中在所述序列读段集合中的每个相应序列读段的所述第二部分中的所述标识符编码选自集合{1,…,1024}、集合{1,…,4096}、集合{1,…,16384}、集合{1,…,65536}、集合{1,…,262144}、集合{1,…,1048576}、集合{1,…,4194304}、集合{1,…,16777216}、集合{1,…,67108864}、或集合{1,…,$1x10^{12}$}的共同值。

35.根据权利要求31所述的测序方法，其中所述标识符是N聚体，其中N是选自集合{4,…,20}的整数。

36.根据权利要求31所述的测序方法，其中所述多个序列读段的平均序列读段长度在40个碱基和200个碱基之间。

37.根据权利要求31所述的测序方法，其中所述多个序列读段的平均序列读段长度在60个碱基和140个碱基之间。

38.根据权利要求31所述的测序方法，其中

所述多个序列读段共同为所述较大的连续核酸提供至少15X覆盖，

所述多个k聚体中超过10％的k聚体来自所述多个序列读段中的超过一个源序列读段，并且

保留对于由超过一个源序列读段代表的每个k聚体的每个此类源序列读段的所述标识符。

39.根据权利要求31所述的测序方法，其中

所述多个序列读段共同为所述较大的连续核酸提供至少25X覆盖，

所述多个k聚体的超过30％来自所述多个序列读段中的超过一个源序列读段，并且

保留对于由超过一个源序列读段代表的每个k聚体的每个此类源序列读段的所述标识符。

40.根据权利要求31所述的测序方法，其中所述多个序列读段中的所述序列读段编码在75个碱基和125个碱基之间的所述较大的连续核酸，并且k是5-73的奇数整数。

41.根据权利要求31所述的测序方法，其中所述多个序列读段中的序列读段集合包括超过100个序列读段，每个具有相同的共同第二部分。

42.根据权利要求31所述的测序方法，其中所述较大的连续核酸是染色体。

43.根据权利要求31所述的测序方法，其中所述较大的连续核酸的长度大于4000万个碱基对。

44.一种计算系统，所述计算系统包含：

一个或多个处理器；

存储器，所述存储器存储要由所述一个或多个处理器执行的一个或多个程序；

所述一个或多个程序包含指令，用于：

(A)获得多个序列读段，其中

所述多个序列读段包括序列读段的多个集合，

在序列读段集合中的每个相应的序列读段包含(i)对应于较大的连续核酸的子集的独特的第一部分和(ii)形成标识符的共同第二部分，所述标识符不依赖于所述较大的连续核酸的序列，并且鉴定多个分区中形成所述相应的序列读段的分区，

在所述多个分区中的分区中形成所述序列读段的多个集合中的序列读段的每个相应集合，并且每个分区包含所述较大的连续核酸的一个或多个片段，所述较大的连续核酸用

作所述分区中的每个相应序列读段的模板；

(B)为在所述多个序列读段中的每个序列读段创建相应的k聚体集合，其中

所述k聚体集合共同包含多个k聚体，

保留所述多个k聚体中的每个k聚体的所述序列读段的所述标识符，

k小于所述多个序列读段中的所述序列读段的平均长度，并且

每个相应的k聚体集合包括所述对应序列读段的所述第一部分的可能k聚体的至少80％；

(C)对于所述多个k聚体中的每个相应的k聚体，追踪含有所述相应k聚体的所述多个序列读段中的每个序列读段的身份；

(D)鉴定第一路径，其中所述第一路径包含在所述多个k聚体中具有k-1重叠的第一k聚体集合；

(E)鉴定第二路径，其中所述第二路径包含在所述多个k聚体中具有k-1重叠的第二k聚体集合，并且其中所述第一路径与所述第二路径相交，由此形成分支区段的集合，其包含：所述第一路径的左部分，所述第一路径的右部分，所述第二路径的左部分和所述第二路径的右部分；并且

(F)使用与所述分支区段的集合的k聚体有关的标识符，验证所述分支区段的连接性。

核酸序列装配的方法和系统

[0001] 相关申请的交叉引用

[0002] 本申请要求2014年6月26日提交的名称为“Processes and Systems for Nucleic Acids Sequence Assembly”的美国专利申请号62/017,589的优先权,其通过引用的方式并入本文。

技术领域

[0003] 本公开涉及用于组装来自多种不同测序系统(包括例如短读和长读序列系统)的序列读数的基因序列组装方法。

背景技术

[0004] 与高通量下一代测序技术相关的重要挑战之一在于将相对短的序列读段装配成较长的连续序列。基因组序列装配过程通常类推为具有切割成小段的小说,然后必须将其重新装配成完整的小说。在序列装配中,这通常通过将要装配的整个序列的重叠子区段拼接在一起来实现。如将理解的,当各段为较大且不太模糊时,该装配过程变得较容易。例如,根据完整的段落,页或章节比从单个句子,句子片段或随机词或词的部分装配小说更容易。同样,使用核酸测序,单独的序列读段越短,将多个读段装配成较长的连续序列变得越困难。

[0005] 虽然下一代测序技术能够产生大量的序列数据,例如在单次运行中产生1兆兆碱基(terabase)序列数据,但是它们困扰于仅产生短的序列读段长度的困难。具体地,这些测序技术通常获得100个连续碱基或更少(或当作为配对末端测序来测序时,高达约200个碱基)的读段中的序列数据。这些读段然后必须装配成更长的连续序列。尽管某些测序技术提供长度为800、1000、5000或甚至20,000个碱基的更长序列读段,但这通常以系统通量为代价而来到,每次运行仅产生数百兆碱基序列数据。

[0006] 已经采用了许多过程和算法将相对短的序列读段装配成更长的连续序列。通常,这些短序列读段通常与许多其它短序列读段重叠,以在整个较长序列的不同部分上提供冗余覆盖(称为“覆盖”或“倍数覆盖”)。在高水平,来自多个读段的重叠序列用于将较长序列信息拼凑在一起。在许多情况下,例如来自相同物种的现有已知参考序列是可获得的。这些参考序列可以用作将较短序列读段定位到较大基因组或基因组片段上的支架上。然后在支架上层叠多倍序列覆盖以提供相对高的置信序列。

[0007] 然而,在许多情况下,参考序列可能不可用,需要被测序的基因组或基因组片段的从头装配。此外,在许多情况下,此类参考序列是没有帮助的,因为测序的目标实现它以检测靶序列中的突变。例如,在靶序列来自癌症的情况下,它是输入以确定靶序列中存在哪些突变,插入和缺失作为诊断癌症阶段或癌症的一些其它属性的方式。因此,利用参考序列在测序努力中并不总是有用的。

[0008] 因此,本文描述了用于将序列读段装配成更大基因组或基因组片段序列,特别是将序列信息从头装配成更大序列背景的改进的方法,过程和系统,所述序列背景包括例如

全染色体或甚至全基因组背景。

发明内容

[0009] 本文中描述了用于装配来自多种不同测序系统的序列读段(包括例如短读段和长读段序列系统)的改进的遗传序列装配方法。本文中所述的方法可用于从头装配过程和再测序装配过程或针对已知参考序列的装配过程两者。

[0010] 本文所述的方法,过程和系统采用具有相关联的条形码序列的序列读段,以在一个或多个不同的装配步骤帮助整体装配过程。

[0011] 在某些方面,提供了将核酸序列读段装配成较大连续序列的方法。所述方法包括在包括第一数据结构的计算机执行的系统中鉴定包含重叠序列和共同条形码序列二者的序列读段的第一子集,所述第一数据结构包含源自较大的连续核酸的多个序列读段。源自较大的连续核酸的共同片段的序列读段包括共同的条形码序列。比对序列读段的第一子集以提供连续的线性核酸序列。

[0012] 还提供了将核酸序列读段装配成较大的连续序列的方法。此类方法在包括第一数据结构的计算机实施系统中从重叠序列读段集合中鉴定第一子序列,所述第一数据结构包含源自较大的连续核酸的多个序列读段。然后,基于与第一子序列共同的相邻序列上条形码序列的存在,将第一子序列延伸至一个或多个相邻或重叠的序列。然后提供包含第一子序列和一个或多个相邻序列的线性核酸序列。

[0013] 本公开的一个方面提供了一种测序方法,所述测序方法包括在计算机系统上执行获得多个序列读段的方法,所述计算机系统具有一个或多个处理器,以及存储用于由所述一个或多个处理器执行的一个或多个程序的存储器。多个序列读段包含多个序列读段的集合。在序列读段集合中的每个相应序列读段包括(i)对应于较大的连续核酸的子集的第一部分和(ii)形成标识符的共同第二部分,所述标识符不依赖于所述较大的连续核酸的序列,并且鉴定多个分区中的分区,其中形成所述相应的序列读段。在多个分区中的分区中形成序列读段的多个集合中的序列读段的每个相应集合,并且每个分区包括较大的连续核酸的一个或多个片段(例如,2个或更多个,3个或更多个,10个或更多个),所述较大的连续核酸用作分区中每个相应序列读段的模板。

[0014] 在该方法中,对于在多个序列读段中的每个序列读段创建相应的k聚体集合。k聚体集合共同包含多个k聚体。保留多个k聚体中的每个k聚体的序列读段的标识符。在许多情况下,对于k聚体中的至少一些,存在多个此类测序读段标识符。k值小于多个序列读段中的序列读段的平均长度。每个相应的k聚体集合包括对应序列读段的第一部分的长度k的可能k聚体中的一些(例如,至少80%)。

[0015] 在该方法中,对于多个k聚体中的每个相应k聚体,追踪含有相应k聚体的多个序列读段中的每个序列读段的身份和含有序列读段的序列读段集合的标识符。

[0016] 在该方法中,将多个k聚体的全部或一部分(例如,至少1%,至少5%,至少50%)绘制为包括通过多个有向弧连接的多个节点的图。每个节点包含长度k的多个k聚体中具有k-1重叠的k聚体的不间断集合。每个弧将源节点连接到多个节点中的目标节点。源节点的最终k聚体与目标节点的初始k聚体具有k-1重叠。第一源节点具有在所述多个节点中的第一目标节点和第二目标节点二者的第一有向弧。

[0017] 在该方法中，测定是否将源节点与第一目标节点或第二目标节点合并，以便得到更可能代表较大的连续核酸的一部分的重叠群序列。重叠群序列包含(i)源节点和(ii)第一目标节点和第二目标节点中的一个。所述测定至少使用所述第一源节点、所述第一目标节点和所述第二目标节点中的k聚体的序列读段的标识符。

[0018] 另一方面提供了计算系统，所述计算系统包括一个或多个处理器，存储要由所述一个或多个处理器执行的一个或多个程序的存储器，所述一个或多个程序包括用于执行上述方法的指令。

[0019] 本公开的另一方面提供了测序方法，所述测序方法包括在计算机系统上获得多个序列读段，所述计算机系统具有一个或多个处理器以及存储用于由所述一个或多个处理器执行的一个或多个程序的存储器中。多个序列读段包括多个序列读段集合。在序列读段集合中的每个相应序列读段包括(i)对应于较大的连续核酸的子集的独特第一部分和(ii)形成标识符的共同第二部分，所述标识符不依赖于所述较大的连续核酸的序列，并且鉴定多个分区中的分区，其中形成所述相应的序列读段。在所述多个分区中的分区中形成所述序列读段的多个集合中的序列读段的每个相应集合。每个此类分区包括较大的连续核酸的一个或多个片段，所述较大的连续核酸用作分区中每个相应序列读段的模板。

[0020] 在该方法中，对于在多个序列读段中每个序列读段创建相应的k聚体集合。k聚体集合共同包含多个k聚体。保留多个k聚体中的每个k聚体的序列读段的标识符。k值小于多个序列读段中的序列读段的平均长度。每个相应的k聚体集合包括对应序列读段的第一部分的可能k聚体中的至少一些(例如，至少80％)。

[0021] 在该方法中，对于多个k聚体中的每个相应k聚体，追踪在含有相应k聚体的多个序列读段中每个序列读段的身份。鉴定第一路径。第一路径包括在多个k聚体中具有k-1重叠的第一k聚体集合。鉴定第二路径。第二路径包括在多个k聚体中具有k-1重叠的第二k聚体集合。第一路径与第二路径相交(例如，如图15所示)，从而形成分支区段的集合，其包含：第一路径的左部分，第一路径的右部分，第二路径的左部分和第二路径的右部分。与分支区段的集合的k聚体物相关联的标识符用于验证分支区段的连接性。在一些实施方案中，这包括评估在分支区段的集合中的每个可能的分支区段对之间共享的标识符的数目。

[0022] 在一些实施方案中，多个分区中的分区包含具有共同第二部分的至少1000个分子，并且至少1000个分子中的每个分子包括与较大的连续核酸的至少一部分互补的引物序列。

[0023] 在一些实施方案中，在序列读段集合中每个相应序列读段的第二部分中的标识符编码选自集合{1,…,1024}、集合{1,…,4096}、集合{1,…,16384}、集合{1,…,65536}、集合{1,…,262144}、集合{1,…,1048576}、集合{1,…,4194304}、集合{1,…,16777216}、集合{1,…,67108864}、或集合{1,…,$1x10^{12}$}的共同值。

[0024] 在一些实施方案中，标识符是N聚体，其中N是选自集合{4,...,20}的整数。

[0025] 在一些实施方案中，多个序列读段的平均序列读段长度在40个碱基和200个碱基之间或在60个碱基和140个碱基之间。

[0026] 在一些实施方案中，所述多个序列读段共同为较大的连续核酸提供至少15X覆盖，多个k聚体中超过10％的k聚体来自多个序列读段中的超过一个源序列读段，并且保留对于由超过一个源序列读段代表的每个k聚体的每个此类源序列读段的标识符。

[0027] 在一些实施方案中，多个序列读段共同为较大的连续核酸提供至少25X覆盖，超过30％的多个k聚体来自多个序列读段中的超过一个源序列读段，并且保留对于由超过一个源序列读段代表的每个k聚体的每个此类源序列读段的标识符。

[0028] 在一些实施方案中，多个序列读段中的序列读段编码在75个碱基和125个碱基之间的较大的连续核酸，并且k是在5和73之间的奇数整数。

[0029] 在一些实施方案中，多个序列读段中的序列读段集合包括超过100个序列读段，每个具有相同的共同第二部分。

[0030] 在一些实施方案中，较大的连续核酸是染色体和/或长度大于4000万碱基对。

附图说明

[0031] 图1A提供了重叠序列读段的比对的示意图。

[0032] 图1B示出了基于序列读段的样品de Bruijn初始装配图，及主路径鉴定(primary path identification)。

[0033] 图2示出了根据一些实施方案的总体装配过程中连接主路径的高水平连接图。

[0034] 图3示意性地示出了根据一些实施方案的条形码辅助装配过程。

[0035] 图4示意性地示出了根据一些实施方案的用于基因组装配的过程。

[0036] 图5示意性地示出了根据一些实施方案的装配过程的各方面。

[0037] 图6A，6B，6C和6D示出了使用本文中所述的条形码化测序方法从大肠杆菌序列数据的从头装配。

[0038] 图7是示出根据一些实施方案的计算设备的示例框图。

[0039] 图8示出了根据一些实施方案，较大的连续核酸，较大的连续核酸的不同片段和片段的序列读段之间的关系。

[0040] 图9示出了根据一些实施方案的长度3的64个可能k聚体的h(W)可能值。

[0041] 图10示出根据一些实施方案的从序列GATCCATCTT产生的长度3的可能的k聚体。

[0042] 图11A，11B，11C，11D和11E共同示出了根据一些实施方案的用于测序的过程流。

[0043] 图12A，12B，12C，12D，12E，12F，12G和12H共同示出了根据一些实施方案的用于获得多个序列读段的方法，其中多个序列读段包括多个序列读段集合，其中序列读段集合中的每个相应序列包括(i)对应于较大的连续核酸的子集的第一部分和(ii)形成标识符的共同第二部分，所述标识符不依赖于所述较大的连续核酸的序列，并且鉴定多个分区中的分区，其中形成所述相应的序列读段。

[0044] 图13示出了根据一些实施方案的用于获得多个序列读段并将它们合并在一起的方法。

[0045] 图14示出了根据一些实施方案的de Bruijn图。

[0046] 图15示出了根据一些实施方案的另一个de Bruijn图。

[0047] 贯穿本公开，相同的参考数字指代对应的部分。

具体实施方式

[0048] 本公开一般提供用于分析遗传序列数据的方法，过程，以及特别是计算机执行的过程和非瞬态计算机程序产品。本文中所述的过程和产品特别适用于将较短的核酸序列数

据装配成较长连接的并且优选连续的遗传构建体，包括大的重叠群，染色体和整个基因组。

[0049] 将参考以下附图和详细描述来理解本发明和各种具体方面和实施方案。在下面的一些附图和详细描述中，根据在逻辑处理设备(诸如计算机系统)上操作的系统的重要的独立实施方案来描述本发明。这不应被认为限制本发明，本发明使用本文中提供的教导可以应用于一起工作的任何数目的逻辑处理器，无论整合到单机计算机系统，仪器系统，或整合到实验室或诊断设备中或与其功能通信的其它信息使能设备或逻辑组件。为了清楚起见，本讨论根据具体实例提及设备、方法和概念。然而，本发明及其方面可以应用于多种类型的设备和系统。因此，除了所附权利要求书中提供之外，意图本发明不受限制。

[0050] 此外，在本领域中公知的是，诸如本文中描述的逻辑系统和方法可以以模块方式包括多种不同的组件和不同的功能。本发明的不同实施方案可以包括元件和功能的不同混合物，并且可以将各种功能分组为各种元件的部分。为了清楚起见，根据包括许多不同的创新组件和创新组件和已知组件的创新组合的系统来描述本发明。不应推断将本发明限制为包含本说明书中任何说明性实施方案中列出的所有创新组件的组合。如将从本文中的教导中理解的，在计算机或其它逻辑处理系统或电路上执行的本发明的功能方面可以使用任何适当的执行环境或编程语言(诸如C，C++，Cobol，Pascal，Java，Java脚本，HTML，XML，dHTML，汇编或机器代码编程，RTL，Python等)执行或实现。本文中引用的所有参考文献，出版物，专利和专利申请通过引用方式整体并入本文用于所有目的。

[0051] I.概述

[0052] 一般来说，将核酸序列信息装配成更长的连续序列区段通常涉及鉴定不同读段中的重叠序列，并使用那些重叠序列将这些序列彼此定位。因为基因组序列非常大，所以重叠序列的鉴定仅允许推断两个序列之间的连接。通过在整个序列的给定区域中具有较大数目的重叠序列，例如较高的覆盖，这种推断得到实质性加强。

[0053] 在一些实施方案中，当相对于参考序列定位序列时，可进一步加强此推断。特别地，对于许多生物体，已经阐明了参考序列，例如一个个体的基因组的完整或接近完整的序列或来自多个个体的生物体基因组的共同序列，其提供了该生物体的基因组序列的一般框架。然后，通常可以将重叠序列的组相对于DNA正被测序的给定生物体的参考序列定位，以便对所鉴定的序列连接提供较大的置信度，以及将不同的序列区段置于较大的基因组背景中。然而，在本公开的一些实施方案中，不使用参考序列。

[0054] 在装配较大的总体序列中，序列中具有不充足信息(例如，覆盖不足，缺乏与参考序列的明确匹配等)的区域可能导致总体推断序列的断裂。这些断裂导致许多单独的连续序列区段(或重叠群)，这在它们最终被连接成单个连续序列前可需要实质性额外处理。虽然这些挑战在参考序列可用时在测序过程中显著，但在从头序列装配中甚至更显著。根据定义，从头装配没有来自参考序列的序列连接的次级确认水平的益处。相反，这些序列必须从下向上装配，例如从头装配。

[0055] 在实例过程中，从多种不同测序过程中的任一种获得短序列读段的聚集。任选地，将序列读段存储在适当的数据结构中用于进一步评估。计算机算法分析这些序列以鉴定不同序列读段的子集中的共同序列区段(k聚体)。k聚体子序列的鉴定通常基于期望长度的k聚体序列的冗余或覆盖水平，其通常可取决于被测序的基因组的复杂性，测序覆盖的深度和变化的误差类型的耐受。在许多情况下，可以选择长度为约7至约31个碱基，或长度为约

15至约25个碱基的任何长度的k聚体长度。例如，在聚集序列数据库中给定k聚体仅表示少于3、4、5、10、20或更多次的情况下，可以认为它被覆盖得不足以有意义使用，这取决于正在分析的序列等。

[0056] 然后，将包括经鉴定的k聚体序列的序列片段沿着序列之间的重叠序列进行比对。然后，可以在这些重叠序列中鉴定另外的一个或多个k聚体序列以编译另外的序列片段的聚集，以使用从该k聚体延伸到重叠k聚体的额外装配算法在初始装配图中延伸连续序列，以产生完整的重叠群序列，例如使用DeBruijn图汇编程序，例如Velvet，Euler等。特别地，具有第一k聚体序列的读段的第一集合可以用作重复比对练习的锚，其中使用第一集合的一个或多个成员共同的第二k聚体序列来延伸推断的序列，即，建立短序列的第一集合和短序列的第二集合之间的连接。只要能够明确延伸该序列，就重复这点。

[0057] 然而，在基因组样品中，k聚体序列可以是总体序列的多个不同部分共同的，使得后续k聚体序列的鉴定可以牵涉两个或更多个不同序列读段集合作为给测序列的延伸，其中那两个序列实际上不相关，例如，那些序列中的一个或多个不是序列的正确延伸。这导致序列之间的推断连接中的分支，例如，两个或更多个序列中的一个可以在该分支点处构成连续序列。最大无分支推断序列通常称为“主路径”。为短序列读段数据集鉴定多个主路径。然后，不同的主路径可以通过其它过程彼此连接，例如，使用配对的末端序列数据，覆盖的长序列读段数据等。

[0058] 图1A和1B提供了上述从头装配过程的示意图。如所示，与在较低覆盖下代表的其它k聚体相比，根据以足够高的覆盖水平代表的包括的k聚体序列(例如k聚体104、106和108)比对序列片段102，例如，比对120和122k聚体中的k聚体。例如，在一些情况下，代表至少3-4X的覆盖的k聚体将用于比对。虽然图示为长度仅为8个碱基的k聚体，但是这是为了便于说明而完成，并且实际的k聚体长度可以变化并且可以更长，例如10聚体或更长，20聚体或更长，30聚体或更长，40聚体或更长，50聚体或更长，60聚体或更长，70聚体或更长，80聚体或更长，90聚体或更长，100聚体或更长。

[0059] 鉴定一系列重叠的k聚体序列，例如k聚体104和106，沿着聚集的重叠片段步行并装配较长的连续序列。如图1B所示，其中给定步骤导致两个或更多个重叠的k聚体，例如k聚体110和112，其在连续序列中创建分支点114。图1B的元件116示出了连续序列中的另一个断点。两个分支点之间的无分支序列的延伸被定义为主路径118。尽管被示为相对短的主路径，但是应当理解，通常产生较长的主路径，例如包含500个碱基，1kb，2kb，5kb或更长的长度。

[0060] 然后，从数据中鉴定种子主路径。种子主路径(也称为种子路径)通常代表在基因组或其它大样本核酸(例如，接近1的n)内更可能具有低拷贝数的主路径。此拷贝数可以从主路径中的序列的长度或从主路径的总体读段覆盖或两者来推断。然后，通过与种子路径的明确重叠，或通过使用其它连接信息，例如长读序列数据或配对末端序列读段，可以从每个种子路径鉴定相邻或连接的主路径，所述配对末端序列读段可以桥接两个主路径，例如，一个在一个主路径中表示，而另一个末端在不同的主路径中表示。两个主路径之间的连接可以例如通过对中间序列进行测序来明确确定，或者它可以是推断的结构连接，其中推断两个主路径的相对位置和连接，例如从配对的末端序列读段。

[0061] 这在图2中示出，其示意性地示出了多个单独的主路径202-210，其中结构连接(示

为连接212)确定为在主路径之间。如所示,在一些情况下,连接不可确定。

[0062] 然后,编译最长的装配的连续序列区段(在图2中显示为重叠群216和218)。通常,然后使用长的基因组整理过程,例如使用费力的Sanger测序或其它长读测序方法,例如来自Pacific Biosciences的SMRT®测序,以闭合总体序列中的剩余缺口,例如在重叠群206和208之间的缺口。

[0063] 从上述实例清楚的是,对于给定的基因组,仅仅装配合理的完成的基因组或合理数目的重叠群可以涉及多种不同的测序过程,例如高通量,短读段鸟枪法测序,配对末端长片段测序和/或低通量长读段DNA测序过程。这些过程中的每一个都是昂贵的且耗时的。

[0064] 然而,如本文所述,改进的测序和装配过程允许从更少或甚至单个更有效的测序过程获得基本上更完全的装配。特别地,根据本文所述的方法,使用例如短读段测序技术可测序的结构连接的寡核苷酸提供有共同的独特标识符,例如条形码,其可以在装配过程中用于推断不同序列区段和/或不同的主路径之间的连接。

[0065] II.连接读段测序

[0066] 结构连接的短序列和所得的序列读段的条形码化(barcoding)可以通过多种方法实施。示例性方法详细描述于例如2014年2月7日提交的美国临时专利申请号61/940,318,2014年5月9日提交的美国临时专利申请号61/991,018,2014年6月26日提交的美国专利申请号14/316,383以及2014年2月7日提交的美国专利申请号14/175,935,其全部公开内容通过引用的方式整体并入本文。

[0067] 简言之,从一个或多个大核酸序列,例如基因组,染色体或其片段制备具有并入的条形码序列的测序文库。将该大的样品核酸片段成通常长度等级为约1至约100kb或更长的第一片段。然后,将这些长的第一片段分隔成单独的反应体积。虽然这些反应体积可以是多种不同类型的反应容器中的任何一种,例如管,孔,微孔,纳米孔等,但是在优选方面,反应体积(也称为分区)由乳液,例如油包水乳液中的液滴构成。

[0068] 可以实施该分隔以便确保在任何分区中最多只含有单个第一片段。然而,此类过程往往具有减少的通量和较高的成本,因为分区被低效地使用,其中大多数分区将保持为空。因此,在优选的方面中,在单个分区中含有多个第一片段的条件下实施分隔,但是其中存在在给定分区内可能存在多个第一片段的高概率,则此类片段将是结构上独特的,例如,代表较大样品核酸的结构上离散且不同的区域。通常,这通过提供下述浓度的第一片段来实现,所述浓度基于样品核酸的大小,第一片段的大小和反应体积,将以统计法将结构独特的第一片段分配到每个单独的反应体积中。特别地,此类分隔将导致少于10%的分区具有结构上连接的分开的第一片段(重叠,不重叠但紧邻或在0.1至5kb内相邻),给定分区中的少于5%的第一片段在结构上连接,给定分区中少于1%的第一片段在结构上连接,并且在一些情况下,给定分区中少于0.5%,0.2%或甚至少于0.1%的第一片段在结构上连接。

[0069] 与第一片段共同分隔,例如在同一液滴内,是包含给定分区内的共同条形码序列的寡核苷酸的集合。寡核苷酸的集合取自不同条形码序列的广泛且多样的文库,使得不同的分区具有其中包含的不同条形码序列。在优选的方面,将包含条形码序列的寡核苷酸提供到可释放地附着到珠粒或微胶囊的各个区室中,其中附着到给定珠粒的寡核苷酸内包括的条形码序列都是基本相同的。此外,每个珠粒通常包含超过100,000,超过1,000,000,超过10,000,000和在许多情况下超过50,000,000个单独的寡核苷酸分子。寡核苷酸可以取自

包括大于1000,大于10,000,大于100,000,大于500,000,大于100万,或甚至大于200万个不同条形码序列的条形码库。

[0070] 在优选的方面,这些寡核苷酸还包括功能性序列,例如扩增引物序列(其可以靶向特定序列,通用引发序列或随机引发序列)或引物退火位点,测序引物序列或引物退火位点等。在优选的情况下,通用或随机引物序列被包括作为寡核苷酸内的功能性序列中的至少一个。这些和其它功能序列描述于例如2014年2月7日提交的美国临时专利申请61/940,318,2014年5月9日提交的美国临时专利申请61/991,018,2014年6月26日提交的美国专利申请14/316,383以及2014年2月7日提交的美国专利申请号14/175,935,其全部公开内容通过引用方式整体并入本文。

[0071] 然后在反应体积内产生构成第一片段的重叠片段的第二可测序的较短片段的群体,其中第二短片段包括共同的条形码序列区段。在一个示例性过程中,这通过将包含条形码序列的寡核苷酸从珠粒释放到分区中来进行。虽然寡核苷酸的释放可以使用从珠粒释放寡核苷酸的多种机制,例如热裂解,化学裂解或光诱导裂解来实施,但是在许多情况下,释放将通过将在共分隔步骤期间将化学切割剂导入分区中实施,其将切割寡核苷酸和珠粒之间的连接,或降解珠粒,使得寡核苷酸被释放,或两者。

[0072] 还与第一片段和寡核苷酸共分隔的是用于实施第一片段的不同部分的扩增的试剂,包括例如DNA聚合酶,核苷三磷酸,二价金属离子,例如Mn^{2+}或Mg^{2+}。然后使用寡核苷酸(例如包括通用或随机引物序列)作为该复制的引物复制给定分区中的第一片段的部分,使得引发并复制第一片段的不同部分。在一些情况下,这些第一复制片段可以另外由寡核苷酸引发并复制。作为这些复制的结果,创建给定分区中的第一片段的重叠片段的集合,其中每个片段包括该分区共同的条形码序列。

[0073] 在不同的分区中,以相似的方式从不同的第一片段产生不同的第二片段,不同的条形码序列附加到那些第二片段。然后在不同分区的集合中创建实质测序文库,其中在相同分区中创建的并且因此可以源自相同第一片段的该文库的元件将包括相同的条形码序列。

[0074] 然后合并代表测序文库的不同分区,并对其中包含的片段进行测序以鉴定短的第二片段序列及其附加的条形码序列。因为源自给定第一片段的第二片段将包含共同条形码序列,所以所述条形码序列在给定读段或读段集合中的存在提供可在整个装配过程中应用的额外连接推断。

[0075] III.应用于装配

[0076] A.概述

[0077] 如前所述,在将较短序列读段装配成较长连续的连接序列时使用许多不同的参数。这些包括在序列读段之间和之中重叠序列区段的存在,所述重叠序列区段提供了序列读段代表较大序列背景的重叠区段的推断。提供较长的连接的较短序列的能力以及提供来自给定核酸的较大序列片段之间的连接推断的能力,例如为较大分子提供较少重叠群或甚至单个统一序列的能力将在一定程度上取决于整个靶核酸的测序覆盖的水平和一致性,其中具有较低测序覆盖水平的区域可能导致重叠群之间的间缺口。因此,具有较多的低覆盖面积的测序过程通常导致较不连续的序列数据,例如较大数目的重叠群。虽然一些覆盖缺陷可以通过暴力解决,例如,将基因组测序到大得多的深度,但是就测序过程成本而言,该

解决方案具有潜在的大成本。然而，在许多情况下，覆盖缺陷是系统性的，例如部分源自序列本身。

[0078] 相比之下，如本文所述，条形码化的测序文库的装配可以在短序列读段水平以及在更高水平的装配中提供增强的连接信息，而不一定需要覆盖的显著增加或辅助附加测序过程的使用。

[0079] 特别地，在典型的短序列读段装配过程中，包括重叠序列区段的短序列读段彼此对齐以提供较大的连续序列背景或重叠群。然后可以将这些重叠群与其它重叠群重新进行进一步装配，或者通过相对于参考序列进行定位。然而，因为这些序列读段在数百万到数十亿碱基的基因组背景中极短，例如长度为100-200个碱基，所以任何给定的读段独特地定位到基因组内的单个基因座的潜力是相对低的，并且这可能在装配过程中引入显著的不确定性。通过对大量重叠片段进行测序以在给定基因座提供更大的测序覆盖度，可以减轻这种不确定性。然而，基因组的低覆盖区域仍将在任何装配过程中提供困难，通常在重叠群之间留下缺口。虽然较低的覆盖区域可以通过本文别处描述的其它方法(例如长读段排序)桥接，但是此类方法增加了方法的成本和低效率。

[0080] 然而，在本文所述的方法的背景中，为短读序列提供了条形码序列，其提供了两个或更多个序列读段之间的结构关系的额外推断，而不需要那么多的测序覆盖。具体地，可以基于与其它序列的重叠序列，以及基于其包括共同的条形码序列两者，将比对的短读序列分配给比对的重叠群。使用相关联的共同条形码序列的附加结构连接信息导致鉴定长得多的重叠群。

[0081] 同样，两个重叠群之间的桥接可以另外通过条形码序列的存在来辅助。例如，在相邻鉴定的重叠群包括共享重叠条形码序列但是以其它方式不重叠或不充分重叠以排列在一起的序列读段的情况下，可以推断那两个重叠群之间的结构连接。简言之，包含在两个序列读段中提供了合理强烈的推断，即两个读段是从在单个分区内创建的片段产生的。当与另外的连接信息组合时(例如，序列读段与两个或更多个重叠群的比对)，或在两个条形码跨越相同的两个重叠群的情况下，其给出那两个重叠群的连接的显著指示。

[0082] 就示例性计算机执行的过程而言，在数据结构内提供包括相关联的条形码序列的短读段序列的集合。基于那些序列内的重叠序列，并且还基于这些序列与相同条形码序列的关联，彼此比对序列的子集。不包括相同条形码序列的可比对序列或包含相同条形码序列但不可比对的序列是不包括在内的，并且可以相对于序列的不同子集进行比对。基于比对的、通常条形码化的序列的集合，产生线性连续序列或重叠群。这使用多个不同的序列子集重复，以提供多个不同的重叠群。如将理解的，重叠群可以比基于单独的序列比对将实现的重叠群长，或者可以提供有更高的置信水平，尽管基于更低的测序覆盖。

[0083] 然后，进一步处理所产生的重叠群以提供相对于彼此的相对取向和基因组背景，以建立总体基因组(或其它基因组组分)序列的支架。在一些情况下，这可以通过将离散重叠群相对于参考序列定位来实现。然而，在许多情况下，相对于参考的定位可能不是明确的，例如基于在基因组内的多个重复序列区域的存在，或者因为不存在合适的参考序列。

[0084] 在这种情况下，条形码序列可以另外用于提供关于重叠群之间的连接信息。例如，如上所述，在一些情况下，与第一序列子集比对但不共享共同的第一条形码序列(例如它们包括第二条形码序列)的序列不会用于产生第一重叠群。然而，此类序列可以在创建第二重

叠群时进行比对。同样地，包括第一条形码但与第二重叠群序列比对的序列可能尚未用于产生第二重叠群。然而，基于数据结构中存在的此类序列，可以确定第一和第二重叠群之间的连接。在所述方法的背景中，可以基于如与不同重叠群比对的序列间的共享条形码的阈值水平的存在，和/或基于与第一重叠群比对但包括与第二重叠群相关联的条形码的序列的阈值水平的存在鉴定重叠群连接。如本文所使用的，此类阈值水平可以包括单个序列或2个或更多个，3个或更多个，4个或更多个，5个或更多个，10个或更多个或20个或更多个序列。

[0085] B.应用于从头的装配过程

[0086] 上述的条形码化测序文库在从头装配中具有特别的优点。具体地，如上所述，在两个序列读段中存在共同的条形码可以提供两个序列之间的结构连接的推断，因为存在着这些序列在相同分区内，并且从同一第一片段创建的概率增加。虽然这可能不是决定性的，因为相同的条形码序列可以存在于其它分区中和结构上不相关的第一片段的片段中，但是当与结构关系的其它指示符(例如，重叠序列)组合时，其变成对结构连接的强有力的推断。作为实例，并且参考上述装配过程，图3示意性地示出了条形码序列对装配过程的潜在影响。

[0087] 在一个方面，序列读段内包括的条形码可以用于提供初始装配中的路径之间和之中的结构关系，以提供序列的高水平支架化(scaffolding)，例如提供高置信度主路径支架，其中鉴定并且定位主路径，并且优选地，高置信度主路径的相对顺序，并且可以相对于所述支架定位最终基因组装配。如将理解的，高置信度主路径将通常指主路径，所述主路径基于其长度和复杂度，将预期在基因组内具有重复或多重重复的低概率，从而允许更高的置信度水平，即事实上，连接到该主路径的序列或主路径实际上在基因组中连接，这与连接到具有重复序列的主路径相反。在优选的方面，高置信度主路径是将在相关基因组中具有预测的拷贝数接近1，例如小于2或等于约1的主要途径。

[0088] 简言之，如上所述，将从初始装配图中识别主路径集合，包括高置信度主路径(本文中也称为“单拷贝数”，“低拷贝数”或“CN1”路径)。包括在给定主路径内的每个序列读段将包括相关联的条形码序列，也如上所述。然后，可以为每个主路径产生与每个主路径相关联的条形码的列表。使用相关联的条形码序列，可以使用共享共同条形码序列的那些主路径(CN1和其它方式两者)在相邻的CN1主路径之间绘制初始序列路径。然后，将该路径延伸到下一个相邻的CN1路径，提供CN1路径相对于彼此排序。如将理解的，在许多情况下，重复的条形码序列可以存在于总体测序过程中，并且附着到较大核酸(例如染色体片段或基因组片段)的结构不同的部分。然而，由于装配过程关注于低拷贝数或CN1路径，利用高度不同的条形码文库(例如，大于1000、5000、10,000、100,000或甚至100万个不同的条形码序列)，并且在寻址非常大的遗传样品，将预期相同条形码序列附着到源自基因组或基因组片段的结构不同的(例如不是结构上连接的)部分的相似或重叠序列的概率将极低，例如在许多情况下，小于0.0001％。

[0089] 在CN1路径之间重复该连接过程，直到建立了所有CN1路径的顺序。然后，CN1路径可以用作支架，用于完成路径之间的序列的装配。这通常通过使用包括CN1路径共同的条形码的序列读段以类似的方式实现，以在相邻的CN1路径之间创建局部序列装配图。

[0090] 特别地，使用包括与给定主路径相同的条形码序列的序列读段以与如上述装配初始装配图相同的方式将支架中两个相邻主路径之间的局部序列装配图缝合在一起。该装配

图提供两个主路径之间的空间的序列，闭合序列缺口。通过利用与其中存在k聚体的每个序列相关联的条形码信息，除了k聚体中的重叠序列数据之外，可以明确地鉴定在两个主路径之间延伸的序列。这在图3中示意性显示。如所示，示出了一系列k聚体序列302-316，类似于图1A中的全局装配图所示意性示出的那些，除了其中每个k聚体与所掺入的条形码序列318，例如条形码序列“1”，条形码序列“2”或条形码序列“n”偶联外。与初始装配图一样，在一个或多个给定点处发散k聚体序列的存在可以产生序列内的分支点，例如分支点320和322。然而，通过部分地依靠与每个k-mer序列有关的条形码信息，可以有效地鉴定正确的序列路径，从而消除分支点。特别地，参考分支点322，虽然k聚体314和316都表示可能的下一个序列步骤，因为只有与k聚体314相关联的路径包括共同条形码“1”，它被鉴定为下一个序列步。

[0091] C.示例性实施方案

[0092] 图7是示出根据一些实现的测序系统100的框图。在一些实现中，设备100包括一个或多个处理单元CPU 702(也称为处理器)，一个或多个网络接口704，用户接口706，存储器712和一个或多个通信总线714，用于互连这些组件。通信总线714可选地包括互连和控制系统组件之间的通信的电路(有时称为芯片组)。存储器712通常包括高速随机存取存储器，诸如DRAM，SRAM，DDR RAM，ROM，EEPROM，闪存，CD-ROM，数字通用盘(DVD)或其它光学存储器，磁带盒，磁带，磁性盘存储或其它磁存储设备，其它随机存取固态存储器设备或可以用于存储所需信息的任何其它介质；并且可选地包括非易失性存储器，诸如一个或多个磁盘存储设备，光盘存储设备，闪存设备或其它非易失性固态存储设备。存储器712可选地包括远离CPU(一个或多个)702定位的一个或多个存储设备。存储器712或存储器712内的非易失性存储器设备包括非暂时性计算机可读存储介质。在一些实现中，存储器712或者非暂时性计算机可读存储介质存储以下程序，模块和数据结构或其子集：

[0093] • 操作系统716，其包括用于处理各种基本系统服务和用于执行依赖于硬件的任务的过程；

[0094] • 用于将设备700与其它设备或通信网络连接的网络通信模块(或指令)718；

[0095] • 用于处理序列读段的序列读段处理模块720；

[0096] • 从生物样品获得的较大的连续核酸726的任选表示；

[0097] • 多个序列读段728，在多个序列读段中每个相应序列读段包括对应于较大的连续核酸的子集的至少第一部分730和形成标识符的共同第二部分732，所述标识符不依赖于所述较大的连续核酸的序列，并且鉴定在其中形成相应的序列读段的多个分区中的分区；和

[0098] • 散列表(hash table)740，对于每个相应的序列读段728，包括相关联的第二部分744(条形码)和通过根据预定k聚体长度L的相应序列读段728的散列获得的k聚体

[0099] 在一些实现中，用户接口706包括用于用户与系统700和显示器708交互的输入设备(例如，键盘，鼠标，触摸板，追踪板和/或触摸屏)710。

[0100] 在一些实现中，一个或多个上面标识的元件存储在一个或多个前述存储器设备中，并且对应于用于执行上述功能的一组指令。以上鉴定的模块或程序(例如，指令集)不需要被实现为单独的软件程序，过程或模块，因此这些模块的各种子集可以在各种实现中组合或以其它方式重新布置。在一些实现中，存储器712可选地存储上面鉴定的模块和数据结

构的子集。此外，在一些实施方案中，存储器存储以上未描述的附加模块和数据结构。

[0101] 虽然图7示出了“序列系统700”，但是图7更旨在作为可能存在于计算机系统中的各种特征的功能描述，而不是作为这里描述的实现的结构示意图。实际上，并且如本领域普通技术人员所认识到的，可以组合单独示出的项，并且可以分离一些项。

[0102] 图8示出了较大的连续核酸802，较大的连续核酸的不同片段804和片段的序列读段728之间的关系。通常，将在1个和250个片段804之间，5个和500个片段804之间或10个和1000个片段804之间分隔成单独的分区。在任何情况下，将足够少的片段804分隔成同一分区，使得单个分区中的片段804具有任何可感知的重叠序列的机会是不可能的。进行每个片段804的序列读段728。在典型的实施方案中，序列读段728长度短(例如，小于1000个碱基)，使得它们可以在自动测序仪中测序。分区中的每个序列读段728包括形成标识符的共同第二部分732，所述标识符不依赖于较大的连续核酸802酸性核酸的序列，并且在多个分区中鉴定形成相应的序列读段的分区。

[0103] 在一些实施方案中，构建多个序列读段中的每个序列读段728的相应k聚体集合。在此类实施方案中，对于多个k聚体中的每个k聚体保留包括序列读段的序列读段集合的标识符(条形码732)。K聚体比序列读段短。在一些实施方案中，序列读段在50和250个碱基之间，而k聚体具有在7和51之间的k值(意味着其具有k的核酸长度)。每个相应的k聚体集合包括对应的序列读段的可能的k聚体的至少80％，至少85％，至少90％，至少95％。

[0104] 图11是示出根据一些实施方案对较大的连续核酸(例如，从生物样品获得的)(1102)进行测序的方法(1100)的流程图。该方法在计算机系统中实施，所述计算机系统具有一个或多个处理器和存储用于由一个或多个处理器(1104)执行的一个或多个程序的存储器。

[0105] 获得多个序列读段。根据所公开的系统和方法，获得多个序列读段728(1106)。多个序列读段包括多个序列读段集合。序列读段集合中的每个相应序列读段728包括对应于较大的连续核酸的子集的第一部分(130)和形成标识符的共同第二部分(732)，所述标识符不依赖于较大的连续核酸的序列并鉴定形成相应序列读段的多个分区中的分区。换言之，标识符不是从较大的连续核酸的测序数据的函数得到的，或者不是较大的连续核酸的测序数据的函数。

[0106] 在一些实施方案中，如图8所示，为了获得多个序列读段728，将较大的连续核酸802片段化以形成片段804，并且将这些片段区室化或分隔成离散的区室或分区(在本文中与分区可互换提及)。在一些实施方案中，获得超过10，超过100，超过1000，超过10,000，超过100,000，超过$1x10^6$，或超过$5x10^6$个序列读段集合，对应于超过10个，超过100，大于1000，大于10,000，大于100,000，大于$1x10^6$或大于$5x10^6$个分区。

[0107] 每个分区维持其自身的内容与其它分区的内容的分离。如本文所使用的，分区指可以包括多种不同形式的容器或器皿，例如孔，管，微孔或纳米孔，通孔等。然而，在优选的方面，分区在流体流内是可流动的。在一些实施方案中，这些容器由例如微胶囊或微囊组成，所述微胶囊或微囊具有围绕内部流体中心或核心的外部屏障，或具有能够在其基质内夹带和/或保留材料的多孔基质。然而，在优选的方面，这些分区包括在非水连续相(例如油相)内的水性流体的液滴。多种不同的合适容器描述于例如2013年8月13日提交的美国专利申请号13/966,150，其通过引用方式整体并入本文。同样地，在例如公开的美国专利申请

No.2010-0105112中详细描述了用于在非水相或油连续相中创建稳定液滴的乳液系统，其通过引用方式整体并入本文。在某些实施方案中，微流体通道网络特别适合于产生分区。此类微流体装置的实例包括在2014年4月4日提交的临时美国专利申请No.61/977,804中详细描述的那些，其全部内容通过引用方式整体并入本文用于所有目的。在单独室的分隔中也可以使用替代机制，包括多孔膜，通过所述多孔膜，将室的水性混合物挤入非水性流体中。此类系统通常可获自例如Nanomi，Inc。

[0108] 在乳液中的液滴的情况下，通常可以如下实现将片段802分隔成离散的分区：通过使含有水性样品的流流入其中也在流动分隔流体的非水性流(例如氟化油)的接合中，使得在流动流分隔流体内创建水性液滴，其中此类液滴包括样品材料。如下所述，分区，例如液滴，通常还包括共分区的条形码寡核苷酸。

[0109] 任何特定分区内的样品材料的相对量可以通过控制系统的多种不同参数来调节，包括例如水性流中的片段804的浓度，水性流和/或非水流的流速，等等。本文所述的分区通常特征在于具有小于1000pL，小于900pL，小于800pL，小于700pL，小于600pL，小于500pL，小于400pL，小于300pL，小于200pL，小于100pL，小于50pL，小于20pL，小于10pL或甚至小于1pL的总体体积。在与珠粒共分隔的情况下，应当理解，分区内的样品流体体积可以小于上述体积的90％，上述体积的小于80％，小于70％，小于60％，小于50％，小于40％，小于30％，小于20％，或甚至小于10％。在一些情况下，使用低反应体积分区在用少量起始试剂进行反应(例如，输入较大的连续核酸片段)中是特别有利的。用于分析具有低输入核酸的样品的方法和系统在2014年6月26日的美国临时专利申请No.62/017,580中提出，其全部内容通过引用方式并入本文。

[0110] 一旦将片段804导入它们各自的分区中，给分区内的片段804通常提供独特的标识符，使得在表征那些片段804时，可以将它们归因于已经从它们各自的分区衍生。在一些实施方案中，先前，随后或同时将此类独特标识符递送到保持区室化或分隔的片段804的分区，以允许稍后将特征(例如核酸序列信息)归因于包括在特定区室内的样本核酸，并且特别是可以最初沉积到分区中的连续样品核酸的相对长段。

[0111] 因此，片段804通常与独特的标识符732(例如，条形码序列)共分隔。在特别优选的方面，以包含核酸条形码序列732的寡核苷酸的形式提供独特的标识符732。将寡核苷酸732分隔，使得在给定分区中的寡核苷酸之间，其中含有的核酸条形码序列732是相同的，但是在不同的分区之间，寡核苷酸可以，并且优选地，具有不同的条形码序列。在优选的实施方案中，只有一个核酸条形码序列732与给定分区相关联，但在一些实施方案中，在给定分区中存在两个或更多个不同条形码序列。

[0112] 通常将较大的连续核酸802片段化和分隔，使得分区中较大的连续核酸802的片段804是连续核酸分子的长片段或段。如图8所示，这些片段804通常代表待分析的总体较大的连续核酸(例如整个染色体，外显子组或其它大基因组片段)的许多重叠片段。

[0113] 也通常将较大的连续核酸802片段化并在一定水平上分隔，由此给定分区具有包括开始较大的连续核酸802的两个重叠片段804的非常低的概率。这通常通过在分隔过程期间以低输入量和/或浓度提供较大的连续核酸802实现。因此，在优选的情况下，给定的分区包括开始较大的连续核酸的许多长的但不重叠的片段804。然后，将不同分区中的核酸片段与独特标识符732相关联，其中对于任何给定分区，其中含有的片段804具有相同的独特标

识符,但是其中不同分区包括不同的独特标识符。此外,由于分隔步骤将样品组分分配到非常小体积的分区或液滴中,应当理解,为了实现如上所述的期望分配,不需要对样品进行实质性稀释,如在较高体积的方法中,例如在管,或多孔板的孔中需要的。此外,因为本文中描述的系统采用如此高水平的条形码多样性,所以可以在更高数目的基因组等同物中分配多种多样的条形码,如上文提供。在一些实施方案中,使用超过10,000、100,000、500,000等不同条形码类型来实现大约1:50或更小,1:100或更小,1:1000或更小,或甚至更小比率的基因组:(条形码类型)比率,同时还允许加载更高数目的基因组(例如,大约每个测定大于100个基因组,每个测定大于500个基因组,每个测定1000个基因组或甚至更多,同时仍提供每个基因组的远远改善的条形码多样性。这里,每个此类基因组是较大的连续核酸的实例。

[0114] 参考图12A的图(B),通常通过将含有较大的连续核酸802的样品与一组寡核苷酸标签1202(含有条形码732)组合来进行上述分隔,所述寡核苷酸标签1202在分隔步骤前可释放地附着到珠粒。寡核苷酸1202可以至少包含引物区1210和条形码732区。在给定分区内的寡核苷酸1202之间,条形码区域732基本上是相同的条形码序列,但是在不同分区之间,条形码区域732在大多数情况下是不同的条形码序列。在一些实施方案中,引物区包含(或进一步包含)用于在分区内引发片段804的N聚体1216(随机N聚体或设计为靶向特定序列的N聚体)。在一些情况下,当N聚体设计为靶向特定核酸序列时,引物区1210设计为靶向特定染色体(例如,人染色体1、2、3、4、5、6、7、8、9、10、11、12、13、14、15、16、17、18、19、20、21、22、X、或Y)或染色体的区域,例如外显子组或其它靶向区域。在一些情况下,N聚体设计为靶向特定基因或遗传区域,如与疾病或病症(例如癌症)相关的基因或区域。在一些情况下,N聚体设计为靶向特定的结构变化。在分区内,使用引物序列1210/1216(例如N聚体)进行扩增反应,以在沿其长度的不同位置处引发片段804。作为扩增的结果,每个分区包含片段804的扩增产物,称为序列读段728,其包括相同或接近相同(共同)的条形码732,并且表示每个分区中的片段804的重叠的较小片段。因此,条形码732充当表示序列读段728集合源自相同分区,并因此潜在地也源自相同片段804的标志物。应当理解,在任何给定分区中通常存在有若干片段804。然而,在典型的实施方案中,在相同分区中的片段804通常不具有任何显著重叠,因此可能将扩增序列读段728定位于任何给定分区中的正确片段804。扩增后,将序列读段728合并,测序并使用测序算法比对。因为较短的序列读段728可以根据其相关的条形码序列732进行比对并且归因于较大的连续核酸802的单个长片段804,所以该序列上的所有鉴定的变体可以归因于单一起源片段802和较大的连续核酸的单一起源染色体。此外,通过在多个片段804间比对多个共定位变体,可以进一步表征染色体贡献。因此,可以得出关于特定遗传变体的定相的结论。此类信息可用于鉴定单倍型,其通常是驻留在相同核酸链或不同核酸链上的规定的遗传变体集合。此外,另外/或者,鉴定结构变体。

[0115] 在一些实施方案中,参考图12A,除条形码序列732和引物区序列1210/1216之外,共分隔的寡核苷酸1202还包含功能性序列。例如,在一些实施方案中,共分隔的寡核苷酸1202还包含可用于处理分隔的核酸的其它功能序列,如用于扩增分区内的片段804,同时附着相关联的条形码序列的靶向或随机/通用扩增引物序列,测序引物,杂交或探测序列,例如用于鉴测序列的存在,或用于下拉条形码化核酸,或许多其它潜在的功能序列中的任一种。参见例如关于寡核苷酸和相关条形码和其它功能序列的共分隔的公开内容,以及如在例如2014年2月7日提交的美国专利申请号61/940,318、2014年5月9日提交的61/991,018,

和2014年6月26日提交的美国专利申请号14/316,383以及2014年2月7日提交的美国专利申请号14/175,935中描述的样品材料,其全部公开内容通过引用方式整体并入本文。

[0116] 在一个示例性过程中,提供珠粒,每个珠粒包括可释放地附着到珠粒的大量上述寡核苷酸1202,其中附着到特定珠粒的所有寡核苷酸包括相同的核酸条码序列732,但是其中大量不同的条形码序列在所使用的珠粒群体上呈现。通常,珠粒群体提供包括至少1000个不同条形码序列,至少10,000个不同条形码序列,至少100,000个不同条形码序列或在一些情况下至少1,000,000个不同条形码序列的多样化条形码序列文库。另外,通常给每个珠粒提供附着的大量寡核苷酸分子1202。特别地,包括单个珠粒上的条形码序列732的寡核苷酸1202分子数目可以是至少约10,000个寡核苷酸,至少100,000个寡核苷酸分子,至少1,000,000个寡核苷酸分子,至少100,000,000个寡核苷酸分子,并且在一些情况下在至少10亿个寡核苷酸分子。

[0117] 在一些实施方案中,在对珠粒施加特定刺激时,寡核苷酸可从珠上释放。在一些情况下,刺激是光刺激,例如通过可释放寡核苷酸的光不稳定连接的切割。在一些情况下,使用热刺激,其中珠粒环境温度的升高导致连接的切割或寡核苷酸从珠粒的其它释放。在一些情况下,使用化学刺激,其切割寡核苷酸与珠粒的连接,或以其它方式导致寡核苷酸从珠粒释放。

[0118] 在一些实施方案中,将包括附着的寡核苷酸1202的珠粒与单独的样品共分隔,使得在单个分区内含有单个珠粒和单个样品。在一些情况下,在期望单个珠粒分区的情况下,可能期望控制流体的相对流速,使得平均来说,分区含有每个分区的少于一个珠粒,以便确保被占据的那些分区,主要是单独占据的。同样,在一些实施方案中,控制流速以提供较高百分比的分区被占据,例如仅允许小百分比的未占用分区。在优选的方面,控制流动和通道构造以确保期望数目的单独占据的分区,小于未占据分区的某个水平并且小于多次占据的分区的某个水平。

[0119] 2014年10月29日提交的名称为“Analysis of Nucleic Acid Sequences”的美国专利申请No.62/072,214(其通过引用的方式并入本文)的图3和描述图3的说明书的部分提供了用于根据本公开的一个实施方案的条形码化和随后测序较大的连续核酸(在参考文献中称为“样品核酸”)的一种方法的详细实例。如上所述,虽然单个珠粒占据可能是最期望的状态,但是应当理解,可能经常存在多次占据的分区或未占据的分区。2014年10月29日提交的名称为“Analysis of Nucleic Acid Sequences”的美国专利申请号62/072,214的图4(其通过引用方式并入本文)和描述图4的说明书的部分提供了根据本公开的一个实施方案的用于共分隔样品和包含条形码寡核苷酸的珠粒的微流体通道结构。

[0120] 一旦共分隔,布置在珠粒上的寡核苷酸1202用于条形码化和扩增分隔的样品。在2014年2月7日提交的美国专利申请号61/940,318,2014年5月9日提交的61/991,018号和2014年6月26日提交的美国专利申请号14/316,383(代理人案卷号43487-708.201)(其全部公开内容通过引用方式整体并入本文)中详细描述了在扩增和条形码化样本中使用这些条形码化寡核苷酸1202的一种过程。简言之,在一个方面,与样品共分隔的珠粒上存在的寡核苷酸从其珠粒释放到具有样品的分区中。寡核苷酸通常包括与条形码序列732一起在其5’端1216的引物序列。在一些实施方案中,该引物序列是意在随机引发样品的许多不同区域的随机寡核苷酸序列。在一些实施方案中,引物序列1216是在样品的特定靶向区域的上游

靶向引发的特异性引物序列。

[0121] 一旦释放，寡核苷酸1202的引物部分1216退火到分区中的片段804的互补区域。与片段804和珠粒共分隔的延伸反应试剂，例如DNA聚合酶，核苷三磷酸，辅因子(例如Mg^{2+}或Mn^{2+}等)使用片段804作为模板延伸引物序列，以产生与引物退火的片段804的链互补的序列，并且该互补序列包括寡核苷酸1202及其相关的条形码序列732。多个引物退火并延伸到分区中片段804的不同部分可以导致片段804的大量合并的重叠互补部分，每个拥有指示其被创建的分区的其自身的条形码序列732。在一些情况下，这些互补部分本身可用作由存在于分区中的寡核苷酸1202引发的模板，以产生再次包括条形码序列732的互补物的互补物。在一些情况下，该复制过程配置为使得当重复第一个互补物时，它在其末端处或附近产生两个互补序列，以允许形成发夹结构或部分发夹结构，其降低了分子作为产生进一步迭代拷贝的基础的能力。图12示出了这点的一个实例的示意图。

[0122] 如图12A所示，将包括条形码序列732的寡核苷酸1202与例如样品片段306一起共分隔在例如乳剂中的液滴1204中。在一些实施方案中，提供在珠粒1208上寡核苷酸302，所述珠粒1208与较大的连续核酸片段804共分隔。寡核苷酸1202优选为可从珠粒1208释放，如图12A的图(A)所示。如图12A的图(B)所示，寡核苷酸1202除了一个或多个功能序列(例如，序列1212、732、1210和1216)之外还包括条形码序列732。例如，寡核苷酸1202显示为进一步包括可发挥给定测序系统的附着或固定化序列功能的序列1212，例如用于ILLUMINA，HISEQ或MISEQ系统的流动池中附着的P5序列。换言之，在一些实施方案中，附着序列1212用于将寡核苷酸1202可逆地附着到珠粒1208。如图12A的图(B)所示，寡核苷酸1202还包括引物序列1210，其可以包括(或进一步包含)用于引发片段804的部分复制的随机或靶向N聚体1216(如上所述)。包括在图12A的图(B)的示例性寡核苷酸1202中，是可以提供测序引发区域，如“读段1”或R1引发区域的序列1210，其用于通过测序系统中的合成反应引发聚合酶介导的模板指导的测序。在许多情况下，条形码序列732和固定化序列1212对于附着到给定珠粒的所有寡核苷酸1202可以是共同的。对于随机N聚体引物，引物序列1210/1216可以变化，或者对于某些靶向应用，可以是给定珠粒上的寡核苷酸共同的。

[0123] 参考图12B，基于引物序列1216的存在，寡核苷酸1202a和1202b能够引发片段804，其允许寡核苷酸1202a和1202b的延伸，使用还与珠粒1208和片段804共分隔的聚合酶和其它延伸试剂进行。

[0124] 如图12C所示，在寡核苷酸(对于随机N聚体引物，所述寡核苷酸将退火到片段804的多个不同区域)延伸后；创建片段804的多个重叠互补，例如，序列读段728-1和728-2。因此，图12C示出了(A)获得多个序列读段，其中多个序列读段中的每个相应序列读段728包括对应于较大的连续核酸802的子集的第一部分730和形成标识符732的共同第二部分1250，所述标识符732不依赖于较大的连续核酸的序列，并且鉴定形成相应序列读段728(例如，条形码732)的多个分区中的分区。

[0125] 然后，可以对图12B的条形码化序列读段728进行表征，例如通过序列分析，或者可以在该过程中进一步扩增它们，如图12D所示。例如，也从珠粒1208释放的另外的寡核苷酸，例如寡核苷酸1202c，可以引发片段1202b。具体地，再次，基于寡核苷酸1202c中的随机N聚体引物1216的存在(其在许多情况下将不同于给定分区中的其它随机N聚体)，寡核苷酸与片段1202b退火，并且延伸以创建对片段1202b的至少部分的互补物728-3，所述部分包含较

大的连续核酸序列的一部分的重复。寡核苷酸1202b的延伸继续，直到它已经通过片段1202b的寡核苷酸部分730复制。如图12D所示，寡核苷酸可以配置成在所需点，例如在通过寡核苷酸1202b的序列1216和1210复制后，通过聚合酶迅速停止复制。如本文所述，这可以通过不同的方法实现，包括例如不能被所使用的聚合酶加工的不同核苷酸和/或核苷酸类似物的掺入。例如，这可以包括在序列区1210内纳入含有尿嘧啶的核苷酸，以防止非尿嘧啶耐受性聚合酶停止该区域的复制。因此，参考图12E，创建序列读段728-3，其包括在一端的全长寡核苷酸1202，包括条形码序列732，附着序列1212，引物区1210和随机N聚体序列1216。在序列的另一端包括与第一寡核苷酸1202的随机N聚体的互补物1216'，以及如序列1210'所示的引物序列的全部或一部分的互补物。然后，R1序列1210及其互补序列1210'能够杂交在一起以形成部分发夹结构1260。如应当理解，因为随机N聚体在不同寡核苷酸之间不同，所以预期这些序列及其互补序列不参与发夹形成，例如作为与随机N聚体1216互补的序列1216'，预期不会与随机N聚体序列1216b互补。对于其它应用，例如靶向引物，情况将不是这样，其中N聚体在给定分区内的寡核苷酸中是共同的。

[0126] 通过形成这些部分发夹结构，其允许从进一步复制中去除样品序列的第一水平重复，例如防止拷贝的迭代复制。部分发夹结构还提供可用于随后处理创建的片段(例如片段730-3)的结构。

[0127] 然后，可以将来自多个不同分区的所有序列读段728合并用于如本文中所述的高通量测序仪上的测序。因为每个序列读段728关于其起源分区编码，所以该序列读段728的序列可以基于条形码732的存在归因回其起源。

[0128] 这在图13中示意性示出。如在一个实例中所示，将片段804-1和片段804-2各自与其自身的条形码寡核苷酸1202集合一起分隔，如上所述。在每个分区内，然后分别处理每个片段(804-1和804-2)以提供片段804-1和804-2的重叠序列读段728，从而形成序列读段集合1302-1和1302-2的集合。该处理为序列读段728提供条形码序列732，其对于从特定第一片段804衍生的每个序列读段728是相同的。如所示，序列读段1302-1的集合由"1"表示，而序列读段1302-2的集合由"2"表示。条形码的多样化文库可以用于差异将序列读段的大量不同集合条形码化。然而，在给定分区中每个序列读段不必用不同的条形码序列条形码化。事实上，在许多情况下，可以在同一分区中同时处理多个不同的片段804以包括相同的条形码序列732。在本文别处详细描述了不同的条形码文库。

[0129] 然后，可以使用例如可从Illumina获得的合成序列技术或Thermo Fisher,Inc的Ion Torrent分割合并序列读段集合用于测序。一旦测序，序列读段728可以归因于它们各自的片段804集合，例如，如聚集读段1314和1316中所示，至少部分地基于所包括的条形码，并且可选地并且优选地部分地基于片段本身的序列。然后，装配每个片段集合的归因序列读段，以提供每个样品片段的装配序列，例如序列1318和1320，其又可以进一步归因回它们各自的起源片段804。用于装配基因组的方法和系统序列描述于例如于2014年6月26日提交的美国临时专利申请号62/017,589(代理人案卷号43487-729.101)中，其全部内容通过引用方式并入本文。

[0130] 参考图11A(1108)，在一些实施方案中，多个分区中的分区包含具有共同的第二部分732的至少10个分子，至少100个分子，至少1000个分子，至少$1x10^4$个分子，至少$1x10^5$个分子，至少$1x10^6$个分子，至少$1x10^7$个分子或至少$1x10^8$个分子。换言之，该分区包括至少10个

分子，至少100个分子，至少1000个分子，至少$1x10^4$分子，至少$1x10^5$分子，至少$1x10^6$分子，至少$1x10^7$分子或至少$1x10^8$分子1202(图12A)，并且每个此类分子1202还包含与较大的连续核酸802的至少一部分互补的引物序列1210。

[0131] 参考图11A(1110)，在一些实施方案中，多个分区中的分区包含具有共同的第二部分732的至少10个分子，至少100个分子，至少1000个分子，至少$1x10^4$个分子，至少$1x10^5$个分子，至少$1x10^6$个分子，至少$1x10^7$个分子或至少$1x10^8$个分子，并且每个此类分子1202还包括引物位点1210和半随机N聚体引发序列1216，其与较大的连续核酸(1110)的一部分互补。

[0132] 参考图11A(1112)，在一些实施方案中，多个分区中的分区中的较大的连续核酸802的一个或多个片段中的每个片段804的长度大于1千碱基，长度大于5千碱基，长度大于50千碱基，长度大于100千碱基，长度大于200千碱基，长度大于250千碱基，长度大于300千碱基，或长度大于350千碱基。通常，分隔成分区的较大的连续核酸802的片段804长于1kb，长于5kb，长于10kb，长于15kb，长于20kb，长于30kb，长于40kb，长于50kb，长于60kb，长于70kb，长于80kb，长于90kb或甚至长于100kb。在一些实施方案中，多个分区中的分区中的一个或多个片段具有至少约10kb，至少约20kb，至少约50kb，至少约100kb，至少约150kb，至少约200kb，至少约250kb，至少约300kb，至少约350kb，至少约400kb，至少约500kb，超过1Mb，或甚至超过2Mb的N50(其中大于所述N50数的片段804长度的总和是所有片段804长度的总和的50％)。在一些实施方案中，跨越多个分区的每个分区中的一个或多个片段具有至少约10kb，至少约20kb，至少约50kb，至少约100kb，至少约150kb，至少约200kb，至少约250kb，至少约300kb，至少约350kb，至少约400kb至少约500kb，超过1Mb，或甚至超过2Mb的总体N50(其中大于所述N50数的片段804长度的总和是所有片段804长度的总和的50％)。在一些实施方案中，多个分区中的分区中的较大的连续核酸802的一个或多个片段804的长度(1114)在20千碱基至200千碱基之间。在一些实施方案中，多个分区中的分区中的较大的连续核酸的一个或多个片段长度(1116)在20千碱基和200千碱基之间。

[0133] 参考图11B(1118)，在一些实施方案中，多个分区中的分区中较大的连续核酸的一个或多个片段804由较大的连续核酸(1118)的1个和500个之间的不同片段，10个和1000个之间的不同片段，100个和1000个之间的不同片段，25个和150个之间的不同片段，5个和500个之间的不同片段，25个和400个之间的不同片段，100个和10,000个之间的不同片段，50个和250个之间的不同片段或2个和125个之间的不同片段组成。在一些实施方案中，多个分区中的分区中的较大的连续核酸的一个或多个片段804包含较大的连续核酸(1118)的超过2个不同片段804，超过10个不同片段804，超过20个不同片段804，超过50个不同的片段804，超过100个不同片段804，超过200个不同片段804，超过300个不同片段804，超过400个不同片段804或超过500个不同片段804。在具体实施方案中，多个分区中的分区中的较大的连续核酸802的一个或多个片段804由较大的连续核酸802(1120)的5个和100个之间的片段804组成。

[0134] 参考图11B(1122)，在一些实施方案中，多个序列读段自小于5纳克的核酸或核糖核酸获得。也就是说，在一些实施方案中，多个分区所需的核酸804的总量有利地小于5纳克。在根据本公开的不同实施方案中，多个序列读段自小于500纳克的核酸或核糖核酸，小于250纳克的核酸或核糖核酸，小于50纳克的核酸或核糖核酸，小于5纳克的核酸或核糖核酸，或小于2纳克的核酸或核糖核酸获得。

[0135] 参考图11B(1124),在一些实施方案中,序列读段集合中的每个相应序列读段728的第二部分732中的标识符编码选自以下的共同数值:集合{1,…,1024}、集合{1,…,4096}、集合{1,…,16384}、集合{1,…,65536}、集合{1,…,262144}、集合{1,…,1048576}、集合{1,…,4194304}、集合{1,…,16777216}、集合{1,…,67108864}、或集合{1,…,$1x10^{12}$}。例如,考虑其中标识符由一组五个核苷酸位置表示的情况。在这种情况下,每个核苷酸位置贡献四种可能性(A,T,C或G),在考虑所有五个位置时产生4x4x4x4x4=1024种可能性。因此,五个核苷酸位置形成集合{1,...,1024}的基础。因此,当条形码序列是5聚体时,每个测序读段728的第二部分732编码从集合{1,...,1024}中选择的独特预定值。同样,当条形码序列由一组六个核苷酸位置表示时,六个核苷酸位置共同贡献4x4x4x4x4x4=4096种可能性。因此,六个核苷酸位置形成集合{1,...,4096}的基础。换言之,当条形码序列是6聚体时,每个测序读段128的第二部分732对从集合{1,...,4096}中选择的独特预定值进行编码。在一些实施方案中,标识符是N聚体,并且N是从集合{4,...,20}(1126)中选择的整数。

[0136] 在一些实施方案中,多个测序读段中的测序读段728的标识符732定位于测序读段内的寡核苷酸的连续集合。在一个此类示例性实施方案中,寡核苷酸的连续集合是N聚体,其中N是选自集合{4,...,20}(214)的整数。换言之,在一些实施方案中,例如图12A的图(B)中的条形码732是寡核苷酸标签1202内的核苷酸位置的连续集合(例如,4个连续核苷酸位置,5个连续核苷酸位置,6个连续核苷酸位置,7个连续核苷酸位置,8个连续核苷酸位置,9个连续核苷酸位置,10个连续核苷酸位置,11个连续核苷酸位置,12个连续核苷酸位置,13个连续核苷酸位置,14个连续核苷酸位置,15个连续核苷酸位置,16个连续核苷酸位置,17个连续核苷酸位置,18个连续核苷酸位置,19个连续核苷酸位置或20个连续核苷酸位置),所述寡核苷酸标签1202在较大的连续核酸转录时最终变成第二部分732。

[0137] 相比之下,在一些实施方案中,多个测序读段中的测序读段的第二部分中的标识符定位于测序读段内的寡核苷酸的不连续的集合。在一个此类示例性实施例中,核苷酸的预定的非连续集合共同由N个核苷酸组成,其中N是集合{4,...,20}中的整数。作为实例,在一些实施方案中,参考图12A的图(B),条形码序列732包含寡核苷酸标签1202中第一位置处的连续核苷酸位置的第一集合和寡核苷酸标签120中第二位置处的连续核苷酸位置的第二集合,其通过间隔物与连续核苷酸位置的第一集合移位。在一个具体实例中,条形码序列732包含$(X1)_nY_z(X2)_m$,其中X1是n个连续核苷酸位置,Y是z个连续核苷酸位置的恒定预定集合,并且X2是m个连续核苷酸位置。在该实例中,由调用该示例性条形码的模式产生的测序读段728的第二部分中的标识符定位于寡核苷酸的不连续集合,即$(X1)_n$和$(X2)_m$。这只是在本公开的范围内的条形码序列的非连续格式的许多实例中的一个。

[0138] 核酸条形码序列732通常在寡核苷酸序列内包括6至约20个或更多个核苷酸。在一些实施方案中,这些核苷酸是完全连续的,例如在相邻核苷酸的单一段中。在可选的实施方案中,它们被分离成由一个或多个核苷酸分开的两个或更多个单独的子序列。通常,分开的子序列以约4至约16个间插核苷酸分开。

[0139] 参考图11B(1128和1130),在一些实施方案中,多个序列读段的平均序列读段长度在40个碱基和200个碱基(1128)之间,40个碱基和200个碱基之间,60个碱基和140个碱基之间,或30个碱基和130个碱基之间。在一些实施方案中,序列读段具有与ILLUMINA,HISEQ,

MISEQ或相关系统的流动池相容的序列长度。

[0140] 在一些实施方案中，所述多个序列读段在聚集体中具有至少约10kb，至少约20kb或至少约50kb的N50(其中大于所述N50数的序列读段长度的总和是所有序列读段长度之和的50％)。在一些方面，获得具有至少约100kb，至少约150kb，至少约200kb，和在一些情况下，至少约250kb，至少约300kb，至少约350kb，至少约400kb，并且在一些情况下，或至少约500kb或更多的N50的序列读段长度。在其它情况下，根据本公开获得超过200kb，超过300kb，超过400kb，超过500kb，超过1Mb或甚至超过2Mb的最大序列读段长度。

[0141] 参考图11C(1128和1130)，在一些实施方案中，多个序列读段中的序列读段集合包括超过100个序列读段，并且超过100个序列读段中的每个序列读段728包括相同的共同第二部分(1132)。也就是说，多个分区中的给定分区产生100个序列读段，并且每个此类序列读段728具有相同的标识符732。更一般地，在一些实施方案中，多个序列读段中的序列读段集合包括超过10个序列读段，超过200个序列读段，超过500个序列读段，超过1000个序列读段，超过2500个序列读段或超过5000个序列读段，并且每个此类序列读段728包括相同的共同第二部分(相同的标识符732)。

[0142] 参考图11C(1134,1136)，在一些实施方案中，较大的连续核酸802是染色体(1134)。在一些实施方案中，较大的连续核酸802是多染色体生物体如人的基因组。在一些实施方案中，较大的连续核酸802包括完整基因组，单独的染色体，外显子组，扩增子或多种不同目的核酸中的任一种。在一些实施方案中，较大的连续核酸802在长度上大于4000万碱基对(1136)。在一些实施方案中，较大的连续核酸802长度大于100,000个碱基对，长度大于100万个碱基对，长度大于500万个碱基对，长度大于1000万个碱基对，长度大于2000万碱基对，长度大于3000万碱基对，或长度大于5000万碱基对。

[0143] 创建k聚体集合。参考图11C(1138)，一旦获得多个序列读段，就为多个序列读段中的每个序列读段728创建相应的k聚体集合。总之，k聚体集合共同包含多个k聚体。也就是说，将来自多个序列读段的k聚体的集合合并以形成多个k聚体。重要的是，保留多个k聚体中的每个k聚体746的源序列读段728的标识符732。例如，在本公开的一些实施方案中使用对于所观察的每个相应k聚体746，列出含有相应k聚体的序列读段728的条形码(第二部分)732的数据结构。通常，追踪与含有相应k聚体746的序列读段728相关联的条形码732的任何技术或数据结构在本公开的范围内。

[0144] k值必然小于多个序列读段中的序列读段728的平均长度。在一些实施方案中，k具有11、13、15、17、19、21、23、25、27、29、31、33、35、37、39、41、43、45、47、49、51、或43的值。在一些实施方案中，k是在6和100之间的偶数整数。在一些实施方案中，k是在5和99之间的奇数整数。在一些实施方案中，每个相应的k聚体集合包括对应的序列读段728(1138)的长度k的可能k聚体的至少80％。例如，参考图10，如果序列读段具有序列GATCCATCTT并且k设置为3，则对应的k聚体集合包括8种可能的k聚体的至少80％，例如8种可能的k聚体的至少6种。

[0145] 在一些实施方案中，每个相应的k聚体集合包括对应序列读段728(1138)的长度k的可能的k聚体的至少60％，至少70％，至少80％，至少90％，至少95％或全部。在一些实施方案中，相应的k聚体集合包括对应序列读段728(1138)的第一部分的长度k的可能的k聚体的至少60％，至少70％，至少80％，至少90％，至少95％或全部。在一些实施方案中，k聚体的两个或更多个集合包括对应序列读段728(1138)的长度k的至少60％，至少70％，至少80％，

至少90％，至少95％或全部。

[0146] 参考图11C(1140)，在一些实施方案中，多个序列读段共同为较大的连续核酸802提供至少15X或至少25X的覆盖，多个k聚体中的k聚体746的超过10％或超过30％来自多个序列读段中的超过一个序列读段728，并且保留对于由超过一个序列读段表示的每个k聚体746的每个此类序列读段728的标识符732。

[0147] 在一些实施方案中，多个k聚体中超过10％，超过20％，超过30％，超过40％，超过50％，超过60％，超过70％或超过80％的k聚体来自多个序列读段中的超过一个序列读段728。在超过一个序列读段728中发现相应k聚体746的情况下，每个序列读段728的条形码(第二部分)732与相应k聚体746相关联。

[0148] 在一些实施方案中，多个序列读段共同为较大的连续核酸802提供至少5X，10X，15X，20X，25X或至少30X的覆盖。

[0149] 参考图11C(1140)，在一些实施方案中，多个序列读段中的序列读段728编码在75个碱基和125个碱基之间的较大的连续核酸802，并且k值是在5和73之间的奇数整数(1142)。在一些实施方案中，多个序列读段中的序列读段728编码在80个碱基和120个碱基之间的较大的连续核酸802，并且k值是在11和51之间的奇数整数。在一些实施方案中，多个序列读段中的序列读段728编码在90个碱基和110个碱基之间的较大的连续核酸802，并且k值是在13和45之间的奇数整数。

[0150] 参考图11C(1144)，在一些实施方案中，如下创建相应的k聚体集合，通过根据预定的k聚体长度k对多个序列读段中的每个序列读段728散列第一部分730，从而创建多个序列读段中的每个序列读段728的相应k聚体集合。从序列读段产生k聚体的一种此类方法包括在覆盖在序列读段728的第一部分730上的核苷酸长度k的滑动窗口的每个位置处计算散列函数，以便形成与序列读段的第一部分730对应的k聚体的集合。在该方法中，使用称为散列函数的简单函数对长度为k的每个可能的k聚体进行散列。核酸序列的一个可能的散列函数h(W)根据下式从序列读段728的第一部分730定位形式$W=w_1w_1\cdots w_k$的k聚体：

[0151] $(W)=f(w_k)+f(w_{k-1})\times 4+\cdots+f(w_1)\times 4^{k-1}$

[0152] 其中对于w_i＝A、C、G、T，$f(w_i)$＝0、1、2、3。其它散列函数是可能的。通常，可以使用将序列$W=w_1w_1\cdots w_k$的k聚体从序列读段分配到可以用于重建k聚体的实际核酸序列的独特数目的任何函数。图9显示了长度为3的64种可能的k聚体的h(W)的可能值。因此，在一些实施方案中，每个k聚体746的值h(W)存储在散列表740或某种其它形式的数据结构中。在一些实施方案中，如图7中所示，对于每个序列读段728，保留序列读段728的分区的相关联的第二部分732(例如，条形码)。此外，在一些实施方案中，k聚体740的数据结构还包括每个k聚体746在对应的序列读段728中的位置。在典型的实施方案中，将序列读段728的第一部分730中的每个可能k聚体转化为k聚体746。例如，在k聚体长度为3并且$W=w_1w_1\cdots w_k$(例如，序列读段728的第二部分732)为GATCCATCTT的实例中，将产生图10的k聚体，并且将产生这些k聚体中的每一个的散列值，例如h(W)值。

[0153] 追踪k-聚体。参考图11D(1146)，对于多个k聚体中的每个相应的k聚体746，追踪多个序列读段中的每个序列读段728的身份，所述序列读段728含有相应的k聚体746和含有序列读段的序列读段集合的标识符732。例如，参考图7，实现这点的一种方式是追踪从序列读段728产生的所有k聚体746以及条形码732。可选择的方式是可能的。例如，可以实现追踪含

有相应的k聚体746的所有序列读段728的数据结构或数据结构的集合，以及这些序列读段728的条形码732。此数据结构或数据结构的集合将进一步追踪含有所有其它相应的k聚体746的序列读段728以及这些序列读段728的条形码732。

[0154] 绘制k聚体。参考图11D(1148)，将多个k聚体作为包括通过多个有向弧连接的多个节点的图绘制。每个节点包括长度k的多个k聚体中具有k-1重叠的k聚体的不间断集合。每个弧将源节点连接到多个节点中的目标节点。源节点的最终k聚体具有与目标节点的初始k聚体的k-1重叠。第一源节点具有含多个节点(1148)中的第一目标节点和第二目标节点两者的第一定向弧。

[0155] 图14示出了一个此类图1400，其在本领域中被称为de Bruijn图。参见Zerbino和Birney，“Velvet:Algorithms for de novo short read assembly using de Bruijn graphs”，Genome Reach 2008,18:821-829，其通过引用方式整体并入本文。每个节点1402表示一系列重叠的k聚体。在图14的情况下，k＝5。节点1402中的相邻k聚体重叠k-1个核苷酸。由k聚体含有的边缘信息是其最后(例如，最后3’)核苷酸。那些最终核苷酸的序列称为节点1402b的序列，或s(N)。在图14中，并且在可选实施方案中，每个节点1402附着到双节点1402b，所述双节点1402b表示反向互补物k聚体的反向系列。这确保了来自相对链的读段之间的重叠得到考虑。注意，附着到节点及其双节点的序列不需要是彼此的反向互补物。节点1402及其双节点1402b的并集被称为“块”。节点1402由有向“弧”1404连接。在所述情况下，弧的源节点的最后k聚体与其第一个目标节点重叠。由于块的对称性，如果弧从节点1402-1到节点1402-2，则对称弧从节点1402b-2到1402b-1。在这些节点1402和弧1404上，读段定位为遍历图形的“路径”。从一个路径提取核苷酸序列取自第一节点的初始k聚体和该路径中所有节点的序列。

[0156] 图14示出了第一源节点具有第一有向弧的情况，所述有向弧具有多个节点(1148)中的第一目标节点和第二目标节点两者。具体地，节点1402-2是具有到节点1402-3的第一有向弧1404-2和到节点1402-4的第二有向弧1404-3的源节点。这是因为存在有在节点1402-2的最后k聚体(GATTG)与节点1402-3的第一k聚体(ATTGA)之间的k-1重叠，即，ATTG的重叠，以及节点1402-2的最后k聚体(GATTG)和节点1402-4的第一k聚体(ATTGC)间的重叠，即ATTG的重叠。因此，在图1400中存在两条路径，包括1402-1→1404-1→1402-2→1404-2→1402-3的路径以及包括1402-1→1404-1→1402-2→1404-3→1402-4的路径。所公开的方法可用于确定哪个路径更可能代表较大核酸802的对应序列。

[0157] 测定合并哪些节点。参考图11D(1150)，进行关于是否将源节点与第一目标节点或第二目标节点合并的测定，以便导出重叠群序列，该重叠群序列包括(i)源节点和(ii)所述第一目标节点和所述第二目标节点中的一个。测定和选择更可能代表较大的连续核酸802的对应部分的重叠群序列。例如，参考图14，测定是否合并节点1402-2(源节点)和1402-3(第一目标节点)或合并1402-2(源节点)和1402-4(第二目标节点)。该测定1150有利地至少使用第一源节点、第一目标节点和第二目标节点中的k聚体740的序列读段728的标识符732来鉴定更可能代表较大的连续核酸802的对应部分的重叠群。

[0158] 在一些实施方案中，至少部分地通过评估与两个或更多个路径中的每个k聚体相关联的条形码732来完成步骤1150。例如，在图1400中，评估路径1402-1→1404-1→1402-2→1404-2→1402-3中的k聚体的条形码732和路径1402-1→1404-1→1402-2→1404-3→

1402-4中的k聚体的条形码732。由于源核酸802通常包括靶核酸的多个拷贝(例如,相同靶染色体或待测序的其它核酸的多个拷贝),因此对于节点中的每个k聚体通常存在多个条形码732。然而,正确的路径,即代表在较大的连续核酸802中发现的真实序列的路径将具有比不表示在较大的连续核酸802中发现的真实序列的路径更多的共同条形码。使用条形码来鉴定正确路径的实际方法取决于多个变量,因此可以变化。

[0159] 在一些实施方案中,通过施加截留,例如需要至少30%条形码来选择更可能的路径,即包含(i)起点节点和(ii)第一目标节点和第二目标节点中的一个的更好的重叠群序列,所述条形码代表源节点中的任何k聚体和两个节点之间共同的选择的目标节点中的任何k聚体。

[0160] 在一些实施方案中,通过施加截留来选择更可能的路径,即包含(i)起点节点和(ii)第一目标节点和第二目标节点中的一个的更好的重叠群序列,例如,需要至少10%的条形码,至少20%的条形码,至少30%的条形码,至少40%的条形码,至少50%的条形码,至少60%的条形码,所述条形码代表源节点中的任何k聚体和两个节点之间共同的选择的目标节点中的任何k聚体。

[0161] 在一些实施方案中,通过鉴定具有最高百分比的共享条形码的节点对,选择更可能的路径,即包括(i)源节点和(ii)第一目标节点和第二目标节点中的一个的更好的重叠群序列,。

[0162] 参考图11D(1152),在一些实施方案中,第一源节点和第一目标节点是图中包括除源节点和第一目标节点之外的一个或多个额外节点的第一路径的一部分。第一源节点和第二目标节点是图中包括除源节点和第二目标节点之外的一个或多个额外节点的第二路径的一部分。这在图14中被示为路径1402-1→1404-1→1402-2→1404-2→1402-3和路径1402-1→1404-1→1402-2→1404-3→1402-4,其中节点1402在这两个实例中是额外节点。在一些此类实施方案中,1150的测定包括测定通过相对于第二途径的第一部分和第二路径的第二部分之间共享的标识符的数目评估第一路径的第一部分和第一路径的第二部分之间共享的标识符728的数目测定第一路径是否比第二路径更可能代表所述较大的连续核酸(1152)。在一些实施方案中,这是在相对基础上完成的。例如,在与路径中的任何k聚体相关联的标识符746的总数中,在路径的第一部分和路径的第二部分之间共享它们的何种百分比。例如,参考图15,考虑其中路径1包括节点1502、1504、1506、1508、1510,并且路径2包括节点1512、1514、1506、1516和1518的实例。可以将路径1划分为由节点1502、1504和1506组成的第一部分以及由节点1508和1510组成的第二部分。测定与节点1502、1504和1506的k个节点中的任一个相关联的标识符732。测定与节点1508和1510的k个节点中的任一个相关联的标识符732。这对于第二路径重复。也就是说,将路径2划分为由节点1512、1514和1506组成的第一部分和由节点1516和1518组成的第二部分。与节点1512、1514和1506的k个节点中的任一个相关联的标识符732。测定与节点1516和1518的k聚体中的任一个相关联的标识符732。将第一路径的第一和第二部分的标识符732之间存在重叠的程度与第二路径的第一和第二部分的标识符732之间存在重叠的程度进行比较。例如,考虑在第一路径的第一和第二部分二者中找到在第一路径的第一或第二部分中找到的标识符总数的30%的情况。此外,在第二路径的第一和第二部分二者中找到的标识符总数的45%在第二路径的第一和第二部分二者中都找到。在这种情况下,第二路径将被认为更可能代表较大的连续核酸802的对

应部分。可以使用用于比较路径中k聚体的条形码732的其它技术，例如贝叶斯分析或Pearson卡方检验。实例发现是使用条形码732来破坏路径中的模糊性。

[0163] 参考图11E(1154)，在一些实施方案中，第一源节点和第一目标节点是图中包括除源节点和第一目标节点之外的一个或多个额外节点的第一路径的一部分。第一源节点和第二目标节点是图中包括除源节点和第二目标节点之外的一个或多个额外节点的第二路径的一部分。当第一路径具有比第二路径更高的平均覆盖时，测定相对于第二路径向上加权第一路径(1154)。再次参考图15，示出并再次考虑路径1(1502、1504、1506、1508、1510)和路径2(节点1512、1514、1506、1516和1518)，如果路径1的节点的k聚体相对于路径2的节点的k聚体存在更高的覆盖，则这可以在图11的过程1150或1152不能基于单独的序列标识符鉴定明确获胜者的情况下用作连接中断器。

[0164] 参考图11E(1156)，在一些实施方案中，第一源节点和第一目标节点是图中包括除源节点和第一目标节点之外的一个或多个额外节点的第一路径的一部分。第一源节点和第二目标节点是图中包括除源节点和第二目标节点之外的一个或多个额外节点的第二路径的一部分。当第一路径代表比第二路径更长的较大的连续核酸序列的连续部分时，确定相对于第二路径向上加权第一路径(1156)。在一些实施方案中，在图11的过程1150、1152或1154不能基于单独的序列标识符鉴定明确的获胜者的情况下，过程1156用作连接中断器。在一些实施方案中，过程1156与图11的过程1150、1152和/或1154结合使用，以使用多个信息源来识别获胜者。

[0165] 参考图11E(1158)，在一些实施方案中，第一节点中的第一k聚体746存在于多个序列读段728的子多个中，并且子多个序列读段中的每个序列读段728的身份对于第一k聚体728得到保留，并使用测定1150以确定第一路径是否比第二路径更可能代表较大的连续核酸序列(1158)。这在许多情况下出现，其中分隔成片段804的较大的连续核酸802的样品实际上构成靶核酸的多个拷贝(例如，相同靶染色体或待测序的其它核酸的多个拷贝)。这导致节点中的每个k聚体746的多个条形码732。因此，过程1150、1152、1154、1156和/或1158中的任一个能够考虑这点。

[0166] 本公开的另一方面提供了测序方法，其包括在具有一个或多个处理器以及存储用于由所述一个或多个处理器执行的一个或多个程序的存储器的计算机系统上获得多个序列读段，如上文关于图11A的1106公开。多个序列读段包括序列读段的多个集合。在序列读段的集合中每个相应序列读段包括(i)对应于较大的连续核酸的子集的独特第一部分和(ii)形成条形码的共同第二部分，所述条形码不依赖于较大的连续核酸的序列，并识别形成相应序列读段的多个分区中的分区。在多个分区中的分区中形成序列读段的多个集合中的序列读段的每个相应集合。每个此类分区包括较大的连续核酸的一个或多个片段，其用作分区中每个相应序列读段的模板。

[0167] 在该方法中，针对在多个序列读段中每个序列读段创建k聚体的相应集合，如上文关于图11C的元件1138所讨论的。k聚体集合共同包含多个k聚体。保留多个k聚体中的每个k聚体746的序列读段728的标识符732。k值小于多个序列读段中的序列读段的平均长度。每个相应的k聚体集合包括对应序列读段的第一部分的可能k聚体中的至少一些(例如，至少80％)。

[0168] 在该方法中，对于多个k聚体中的每个相应k聚体，追踪包含相应k聚体的多个序列

读段中的每个序列读段的身份。鉴定第一路径。第一路径包括在多个k聚体中具有k-1重叠的k聚体的第一集合(例如,图15的节点1502、1504、1506、1508和1510)。鉴定第二路径。第二路径包括在多个k聚体中具有k-1重叠的k聚体的第二集合(例如,图15的节点1512、1514、1506、1516和1518)。第一路径与第二路径相交(例如,如图15中所示),从而形成分支区段的集合,包括:第一路径的左部分(节点1502和1504),第一路径的右部分(节点1508和1510),第二路径的左部分(节点1512和154)和第二路径的右部分(1516和1518),其中用二分节点1506完成的事情的选择(例如,不使用它,将它置于路径中,或将它置于每个路径的一部分中)是应用依赖的。与分支区段的集合的k聚体物相关联的标识符用于验证分支区段的连接性。在一些实施方案中,这包括评估在分支区段的集合中的每个可能的分支区段对之间共享的标识符的数目。在某些情况下,这可能导致重新路由路径。例如,在评估图15的图时,四个可能的路径是可能的,如下表所示。如表所示,路径1520到1522(节点1502、1504、1506、1508和1510)和路径1524到1526(1512、1514、1506、1516和1518)验证为正确的,因为在相对的基础上,它们各自比路径1520到1526和1524到1522具有更多重叠的条形码。

[0169]

左路径分支区段	左路径分支区段	左分支区段中的序列标识符732的数目	右分支区段中的序列标识符732的数目	重叠左和右分支的序列标识符732的数目
1520	1522	100	110	65
1520	1526	100	105	3
1524	1522	90	110	8
1524	1526	90	105	50

[0170] 在一些实施方案中,使用诸如路径长度和路径覆盖的附加信息来验证这些连接。应当理解,任何形式的统计分析可以应用于信息诸如上表中的信息以验证路径。

[0171] III.示例性装配过程

[0172] 以下是如上所述的总体从头装配过程的实例,在下文阐明,并且在图4所示的过程流程图中示意性示出。如所示,在测序和读段产生之后,将短读段序列在数据结构中聚集,并在步骤402进行预装配处理。在这些读段中,与测序片段的功能性组分(例如衔接子序列,读段引物序列等)相关的序列数据从序列读段中去除。可以通过鉴定在有义读段和配偶或反义读段的前两个k聚体中复制的那些来消除重复读段。当初始k聚体在重复读段和它们的互补物中为共同时,除了一个以外的所有序列可以使用例如FastUniq算法或其它类似算法作为重复来去除。在一些情况下,重复读段可以通过预装配过程,其中那些重复中的一个或多个包括测序错误。因为这前几个k聚体发生在错误较少的测序运行的开始时,所以预期这发生的概率将较低。除了重复读段外,低复杂性读段,例如单个碱基构成读段的90%或更多的读段,或完全或基本上由2、3或4个碱基重复单元组成的读段。

[0173] 然后,在步骤404对读段进行错误校正算法,因为序列读段中的未校正错误导致片段化的装配并增加计算复杂性。不同的校正算法通常是可用的(参见例如Tahir等人,“Review of Genome Sequence Short Read Error Correction Algorithms”,Am.J.BioinformaticsRes.2013;3(1)1-9,其通过引用的方式在此并入本文),并且包括例如QUAKE和单机ALLPATHS-LG错误校正算法。

[0174] 在预装配处理和错误校正之后,使用例如k=20或k=24的k聚体大小,在步骤406在所有聚集读段间产生k聚体的频率谱。假设典型的基因组概貌,在没有显著偏差的情况下

测序，预期将是对于k聚体频率的高斯分布，其中峰的中心指示基因组的平均k聚体覆盖。用于计算频率谱的k聚体的大小通常可以根据正在分析的基因组的复杂性，测序覆盖的预期水平等而变化。

[0175] 然后，使用参考图1B描述的k聚体装配过程，在步骤408从错误校正的读段中装配初始装配图。在一些情况下，可以使用更长的k聚体序列，例如，对于150个碱基的序列读段，可以使用40个或更多，50个或更多个，或60个或更多个碱基的k聚体。从图中，将未分支路径瓦解为单个主路径。

[0176] 代表例如1、2或3k聚体的相对短的分支的悬挂末端通常在序列中是普遍的，并且可以由多个来源产生，包括来自未修正的读段。这些短路径容易鉴定，并且可以在鉴定主路径之前或之后从装配图中移除。在一些情况下，可能出现短分支，但可以从错误校正的读段中的缺失序列导出，这可以由低覆盖率导致。

[0177] 然后，在步骤410，与错误校正的读段偶联的条形码序列与它们各自的主路径相关联，产生与每个主路径相关联的条形码的列表。对于每个主路径，估计基因组内的所述主路径的拷贝数。如所述，给定主路径的拷贝数通常取决于基因组的复杂性和主路径的长度和复杂性。拷贝数通常可以从触碰主路径的读段的数目估计，并将其与从在步骤406计算的k聚体谱导出的平均k聚体覆盖进行比较。短主路径通常可以被忽略，由于确定它们拷贝数更加困难，因为在那些主路径上触碰的读段的大得多的比例可能仅部分地与它们重叠。由此，可以在步骤412鉴定低拷贝数主路径，例如，具有接近1的拷贝数，也被称为CN1路径，如上所述。

[0178] 然后，借助于相关的条形码序列将CN1主路径装配在支架中。特别地，在步骤414，选择与条形码的最大数目相关联的CN1主路径作为种子路径。在步骤416，然后将种子路径扩展到含有共同条形码序列的所有其它重叠主路径，考虑每个主路径仅一次，并且忽略具有低于期望水平的条形码频率(例如，小于种子路径中的条形码频率的一半)的主路径。图5更详细示出了该过程。特别地，在步骤502鉴定第一种子路径。如上面对步骤416所述，将所述第一种子路径扩展十次。特别地，在步骤504，通过主路径和附加CN1路径继续扩展，直到到达给定分支的末端。然后，将路径修整回到在步骤506中路径行进通过的最后的CN1路径，使得装配的序列将始终在CN1路径上开始末端，并且围绕种子路径锚定。

[0179] 然后，通过移除所有非CN1路径，然后合并在一起以形成单个支架，来缩减通过CN1路径的路径，所述支架在步骤508对围绕第一种子路径的CN1路径定义顺序。然后，指定这些CN1路径为已使用。

[0180] 然后，在步骤512使用CN1路径从剩余的CN1路径(其表示最高的剩余条形码频率作为下一个种子路径)以及针对剩余的CN1路径集合的支架重复该过程。这将继续，直到指定所有CN1路径为已使用。然后进一步合并缩短主路径的集合以提供在步骤514给出基因组中所有CN1路径的相对顺序的最终路径或支架。这形成为最终装配提供基础的锚支架。

[0181] 一旦装配了CN1路径支架，就在步骤420进行局部装配以填充CN1路径之间的支架缺口。对于支架中的每个缺口，选择相邻的主路径，并且鉴定与它们相关联的条形码。包括这些条形码的读段组然后与相邻的CN1路径结合使用，以与用于装配全局主路径图的方式相同的方式装配局部主路径装配图。从每个CN1路径，鉴定连接到另一个主路径的所有路径。在许多情况下，将仅鉴定单个路径，从而允许使用该路径作为缺口的补丁。在两个相邻

CN1主路径之间鉴定出超过一个路径的情况下，具有最强读段支持的路径用于将相邻CN1主路径拼接在一起并且闭合该缺口。在读段支持不允许在两个路径之间进行选择的情况下，这两个路径可以包括在图中，或者可以允许缺口保持打开。

[0182] 从完全装配中，创建了整个基因组的线性图，理想地，其不包括剩余缺口或双倍路径。然而，在许多情况下，线性图可以包括双重路径或剩余缺口的一些位置，但是序列的总体顺序是已知的。对于剩余的缺口或双重路径，可以进一步检查未校正的序列数据或丢弃的序列数据以确定是否任何缺口或双倍可以由早期从处理中移除的那些序列告知。或者，可以使用其它消除二义性技术，诸如使用拷贝数信息来通知模糊装配，或使用读段对信息来评估多个解决方案对于模糊装配的相对可能性。

[0183] 在某些方面，本文中提供的方法是计算机执行的方法，其中该方法的至少一个或多个步骤由计算机程序执行。在一些实施方案中，本文中提供的方法在存储在计算机可读介质(诸如标准计算机的硬盘驱动器)上的计算机程序中实现。例如，用于从复制序列读段中确定至少一个共同序列的计算机程序可以包括以下中的一个或多个：用于提供或接收序列读段的代码，用于鉴定序列读段之间的序列重叠区域的代码，用于比对所述序列读段以产生布局，重叠群或支架的代码，用于共同序列确定的代码，用于在计算机监视器上转换或显示所述装配的代码，用于应用本文中所述的各种算法的代码，以及包括所述代码的计算机可读存储介质。

[0184] 在一些实施方案中，从复制序列集合确定至少一个装配的系统(例如，数据处理系统)包括处理器，可操作地耦合到处理器以用于存储存储器的计算机可读介质，其中所述存储器具有用于由所述处理器执行的指令，所述指令包括以下的一个或多个：用于接收序列读段的输入的指令，用于所述序列读段之间的重叠检测的指令，比对所述序列读段以产生布局，重叠群或支架的指令，应用共同序列算法以产生至少一个共同序列(例如，“最佳”共同序列和任选地一个或多个另外的共同序列)的指令，计算/存储与该方法的各个步骤相关的信息的指令，以及记录该方法的结果的指令。

[0185] 在某些实施方案中，本文中描述的过程的各个步骤利用信息和/或程序并产生存储在计算机可读介质(例如，硬盘驱动器，辅助存储器，外部存储器，服务器，数据库，便携式存储器设备等)上的结果。例如，可以存储在非暂时性计算机可读介质上的方法所使用的信息和产生的结果包括但不限于输入序列读段信息，成对重叠的集合，新产生的共同序列，质量信息，技术信息和同源或参考序列信息。

[0186] 在一些方面，本发明包括用于确定至少一个序列装配的制品，其包括含有一个或多个程序的机器可读介质，所述程序在被执行时实施如本文所述的本发明的步骤。

[0187] 如本领域技术人员从本文提供的教导将理解的，本发明可以在硬件和/或软件中实现。在本发明的一些实施方案中，本发明的不同方面可以在客户端逻辑或服务器侧逻辑中实现。如本领域中将理解的，本发明或其组件可以体现在包含逻辑指令和/或数据的固定媒体程序组件中，当被加载到适当配置的计算设备中时，使得该设备根据本发明执行。如本领域中将理解的，可以将含有逻辑指令的固定介质递送到固定介质上的观看者，用于物理加载到观看者的计算机中，或者含有逻辑指令的固定介质可以驻留在远程服务器上，观看者通过通信介质访问所述远程服务器以便下载程序组件，或者根据本文阐述的过程在一个或读个远程位置，用户使用上述硬件和/或软件来将要处理的数据上传到远程服务器中。

[0188] 通常与本文中描述的过程结合使用的信息设备(或数字设备)包括常规逻辑系统，诸如常规计算机系统，包括个体或个人计算机，联网计算机，计算机主机等。与这些过程的交互可以不限于此类常规计算系统，而是可以另外或备选包括用于与远程数据应用交互的任何信息装置，并且可以包括诸如数字使能的电视，蜂窝电话/智能电话，个人数字助手，平板电脑，可穿戴电子设备，例如Google眼镜，智能手表等装置。用于在上述类型的设备中的任一种上执行本文所述的过程的编程可以驻留在设备本身的存储器(主存储器或辅助存储器)上，或者可以通过例如设备上的API和适当的通信网络(例如LAN，WAN，Wi-Fi，蜂窝，蓝牙或其它通信系统)从其它位置和设备，例如计算机访问。

[0189] 大肠杆菌基因组(菌株K12_DH10B)用于模拟本文中所述的从头装配过程。将基因组的60个物理拷贝剪切至平均大小为50千碱基(标准偏差5千碱基)，并随机分成5000个分区，每个分区含有独特的条形码。在每个分区内模拟创建条形码测序文库的过程。这继之以利用150bp配对末端读段模拟该文库的测序至平均深度30X。测序过程的模拟结合了源自实际测序数据的经验误差率模型。图6示出了上述建模数据的从头装配图。如所示，基于单独的短读段序列(例如没有条形码)，与60kb的N50重叠群相比，装配产生750kb的N50重叠群。

[0190] 虽然为了清楚和理解的目的已经相当详细地描述了前述发明，但是本领域技术人员通过阅读本公开内容将清楚，在不脱离本发明的真实范围的情况下，可以进行形式和细节上的各种改变。例如，上述所有技术和装置可以以各种组合使用。例如，可以使用如所述的阵列孔尺寸测定方法来实施颗粒递送。在本申请中引用的所有出版物，专利、专利申请和/或其它文献通过引用的方式以其整体并入本文用于所有目的，程度如同每个单独的出版物，专利、专利申请和/或其它文献单独和分开指明为所有目的通过引用方式并入一样。

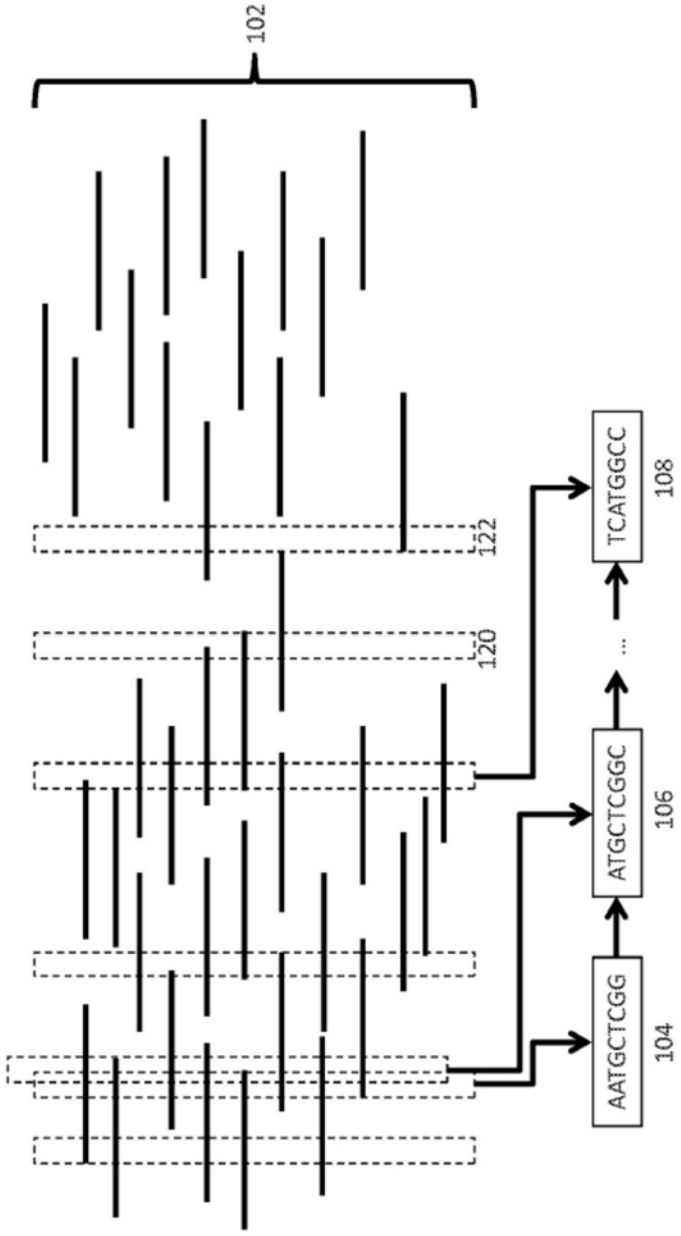


图1A

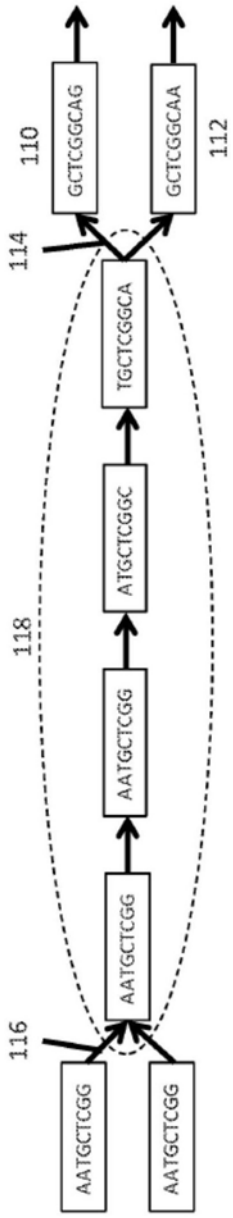

110
GCTCGGCAG
112
GCTCGGCAA
114
TGCTCGGCA
ATGCTCGGC
AATGCTCGG
AATGCTCGG
118
116
AATGCTCGG
AATGCTCGG

图1B

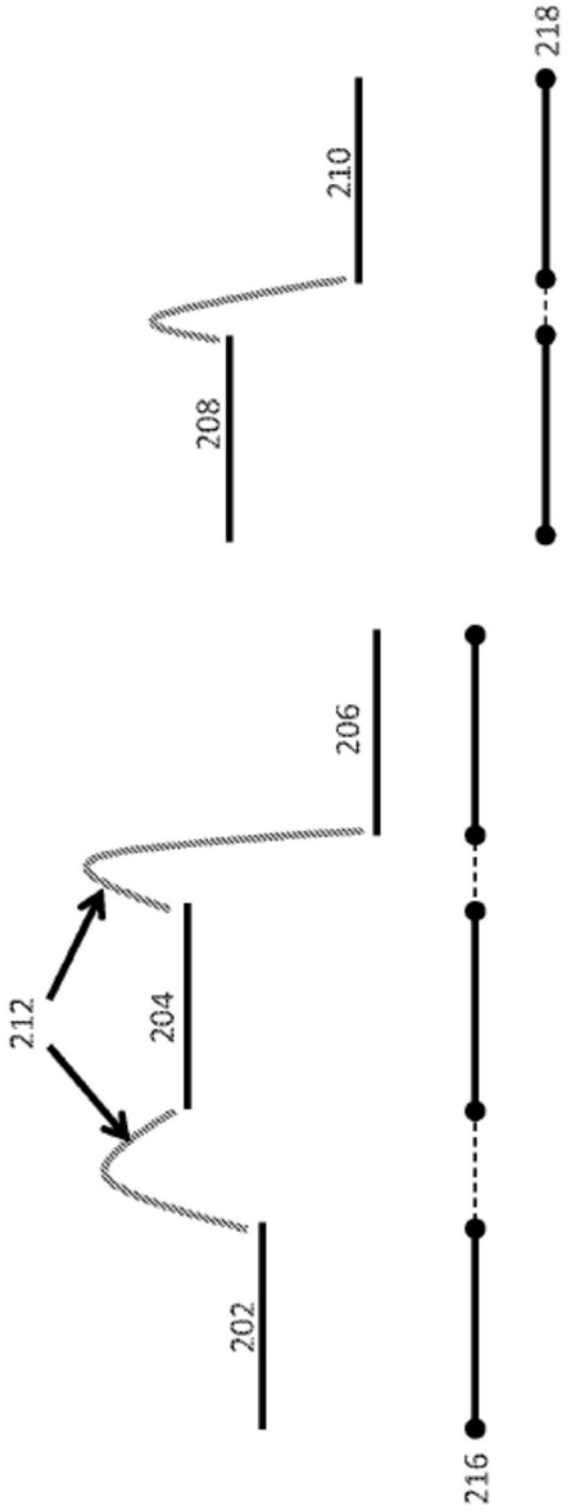


图2

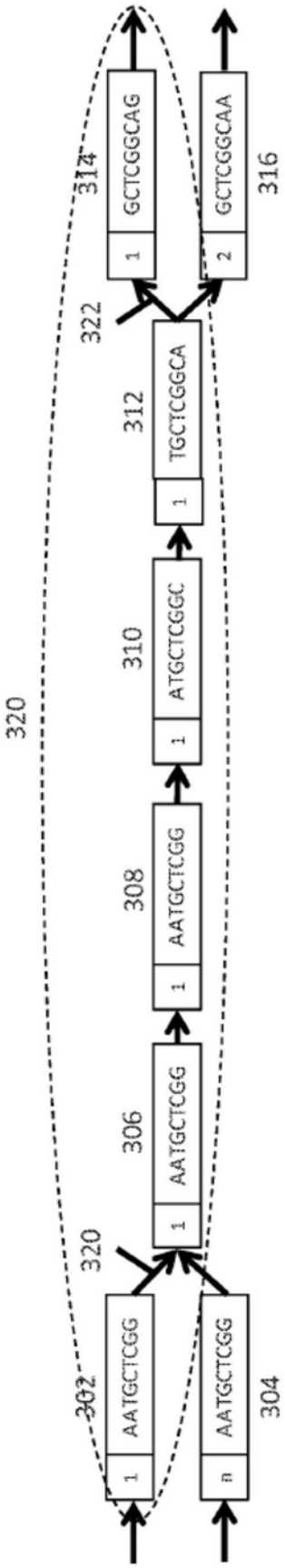

320
302
AATGCTCGG
1
304
AATGCTCGG
n
320
306
AATGCTCGG
1
308
AATGCTCGG
1
310
ATGCTCGGC
1
312
TGCTCGGCA
1
322
314
GCTCGGCAG
1
316
GCTCGGCAA
2

图3

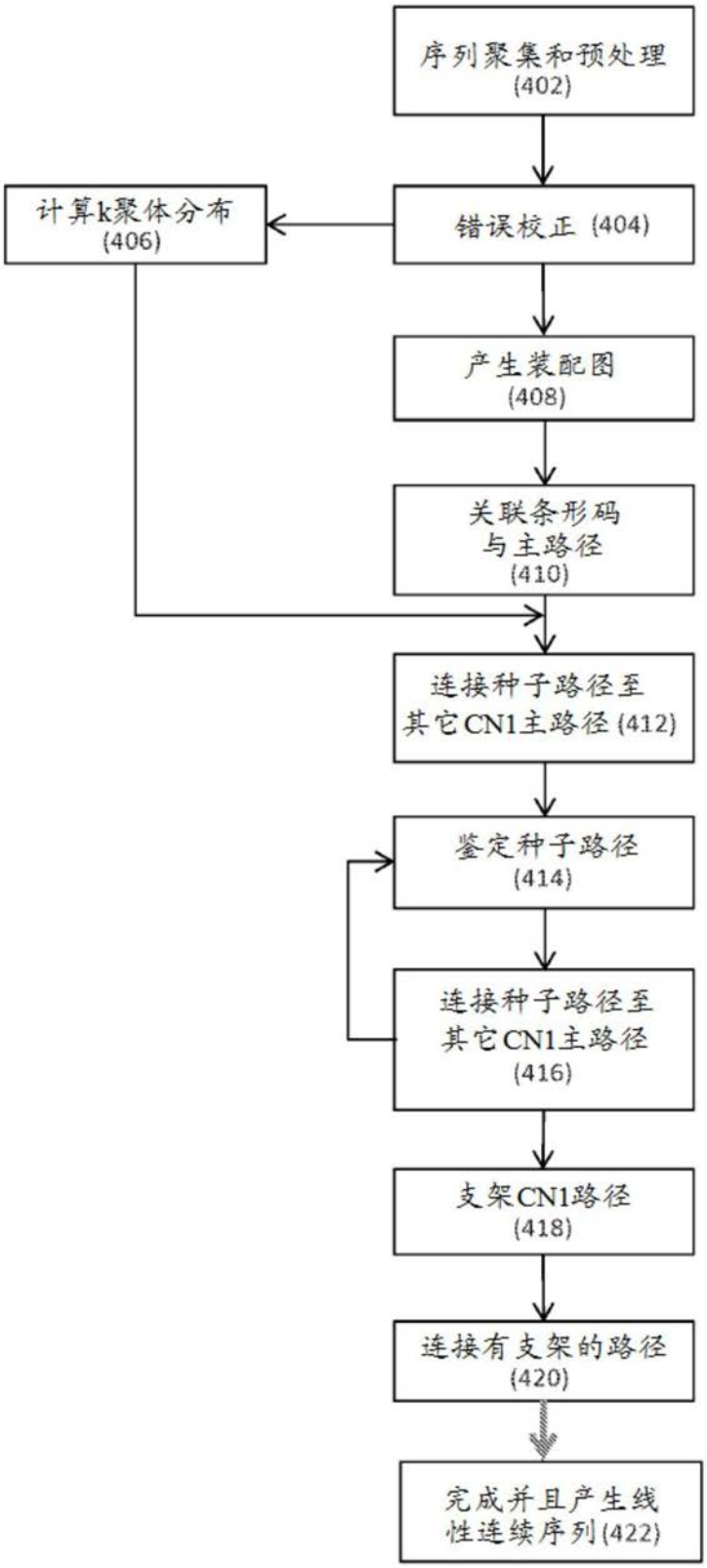

序列聚集和预处理
(402)
计算k聚体分布
(406)
错误校正 (404)
产生装配图
(408)
关联条形码
与主路径
(410)
连接种子路径至
其它CN1主路径 (412)
鉴定种子路径
(414)
连接种子路径至
其它CN1主路径
(416)
支架CN1路径
(418)
连接有支架的路径
(420)
完成并且产生线
性连续序列(422)

图4

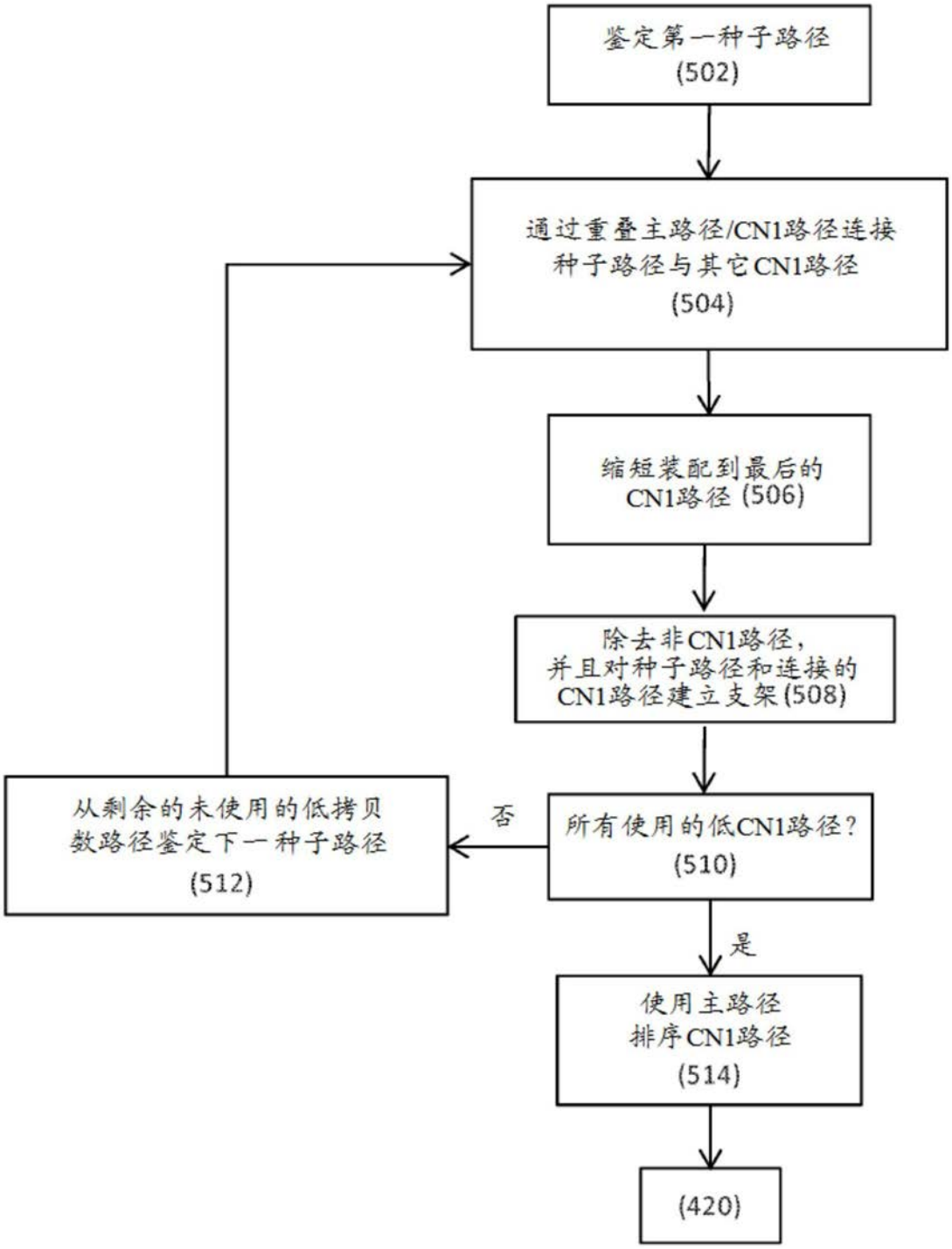

鉴定第一种子路径
(502)
通过重叠主路径/CN1路径连接
种子路径与其它CN1路径
(504)
缩短装配到最后的
CN1路径 (506)
除去非CN1路径，
并且对种子路径和连接的
CN1路径建立支架(508)
所有使用的低CN1路径？
(510)
否
从剩余的未使用的低拷贝
数路径鉴定下一种子路径
(512)
是
使用主路径
排序CN1路径
(514)
(420)

图5

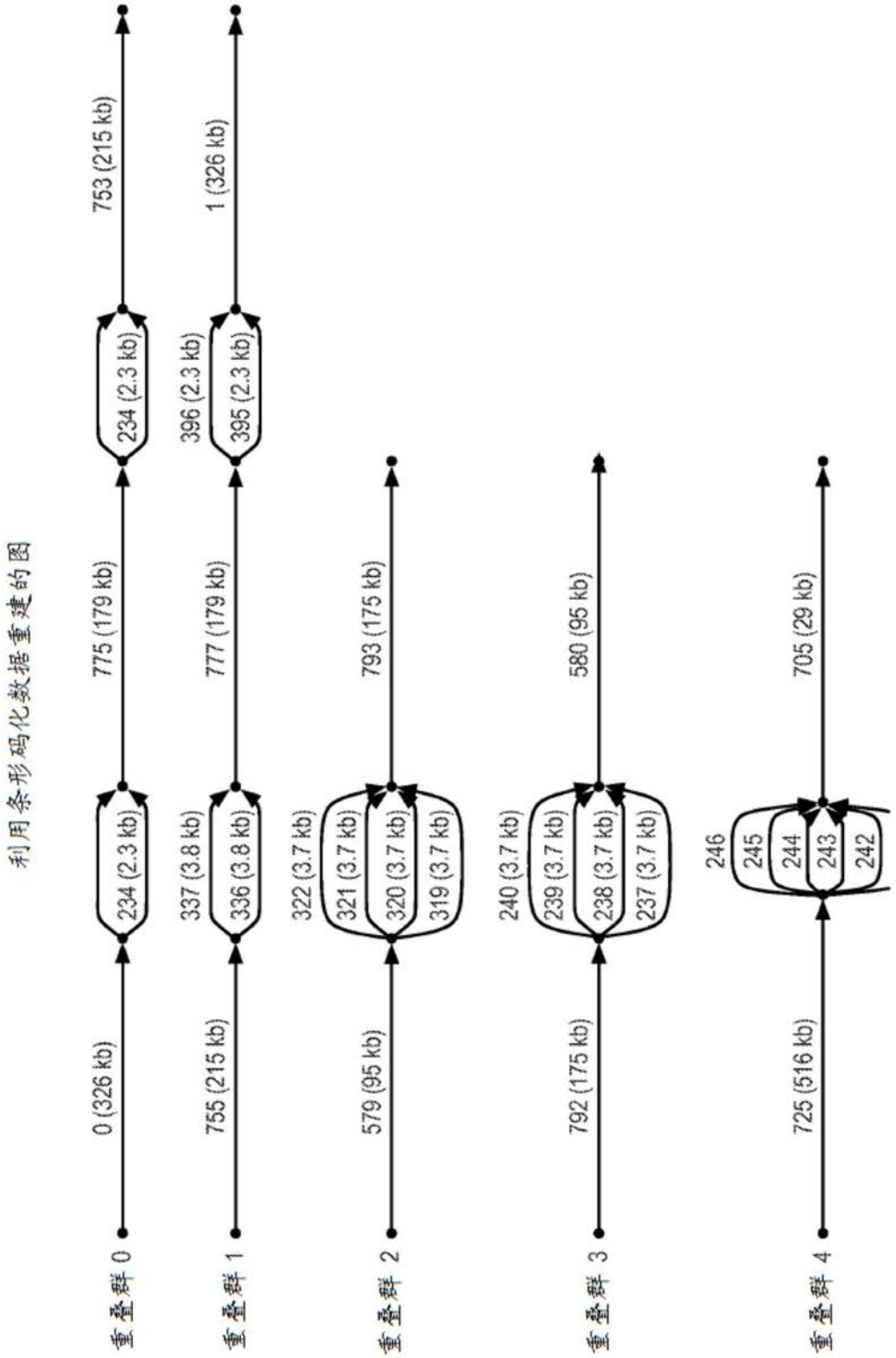

利用条形码化数据重建的图
重叠群 0
重叠群 1
重叠群 2
重叠群 3
重叠群 4
0 (326 kb)
755 (215 kb)
579 (95 kb)
792 (175 kb)
725 (516 kb)
234 (2.3 kb)
336 (3.8 kb)
337 (3.8 kb)
322 (3.7 kb)
321 (3.7 kb)
320 (3.7 kb)
319 (3.7 kb)
240 (3.7 kb)
239 (3.7 kb)
238 (3.7 kb)
237 (3.7 kb)
246
245
244
243
242
775 (179 kb)
777 (179 kb)
793 (175 kb)
580 (95 kb)
705 (29 kb)
234 (2.3 kb)
396 (2.3 kb)
395 (2.3 kb)
753 (215 kb)
1 (326 kb)

图6A

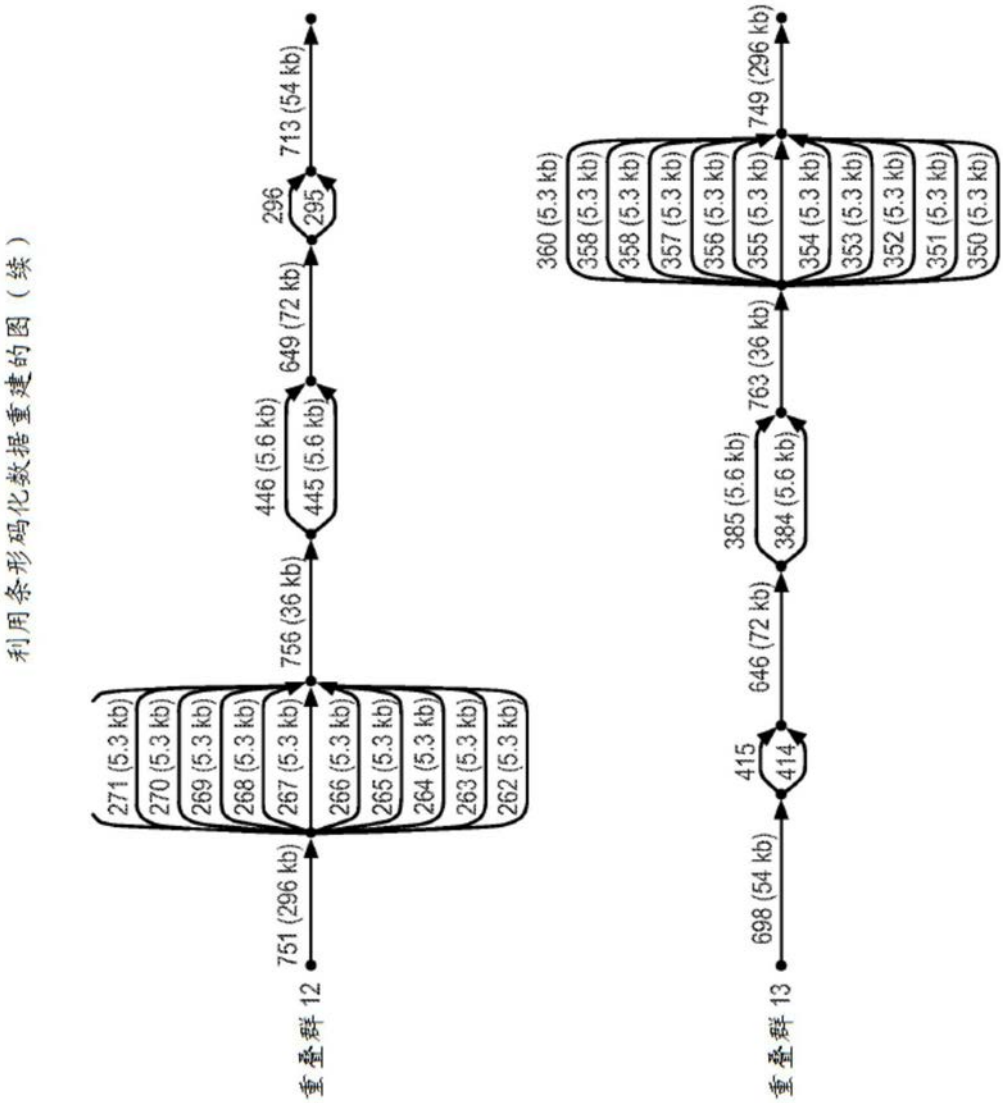

利用条形码化数据重建的图（续）
重叠群 12
751 (296 kb)
271 (5.3 kb)
270 (5.3 kb)
269 (5.3 kb)
268 (5.3 kb)
267 (5.3 kb)
266 (5.3 kb)
265 (5.3 kb)
264 (5.3 kb)
263 (5.3 kb)
262 (5.3 kb)
756 (36 kb)
446 (5.6 kb)
445 (5.6 kb)
649 (72 kb)
296
295
713 (54 kb)
重叠群 13
698 (54 kb)
415
414
646 (72 kb)
385 (5.6 kb)
384 (5.6 kb)
763 (36 kb)
360 (5.3 kb)
358 (5.3 kb)
358 (5.3 kb)
357 (5.3 kb)
356 (5.3 kb)
355 (5.3 kb)
354 (5.3 kb)
353 (5.3 kb)
352 (5.3 kb)
351 (5.3 kb)
350 (5.3 kb)
749 (296 kb)

图6B

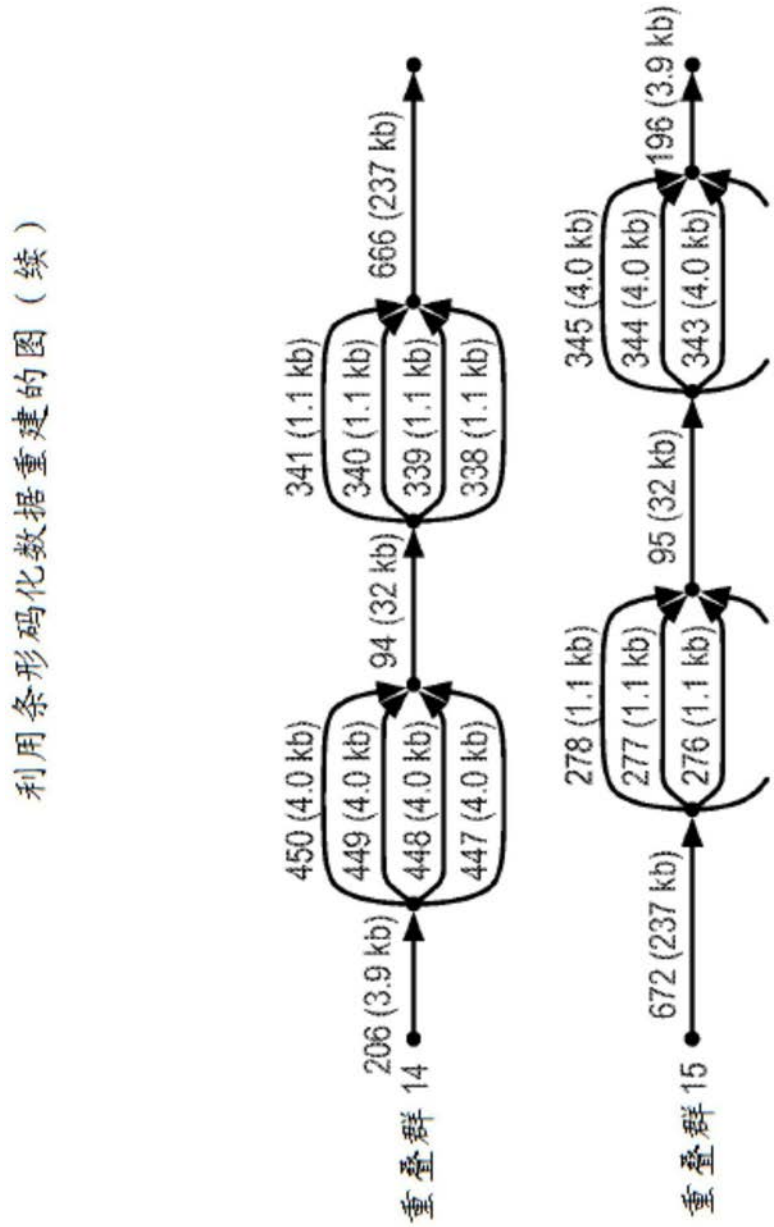

利用条形码化数据重建的图（续）
重叠群 14
206 (3.9 kb)
450 (4.0 kb)
449 (4.0 kb)
448 (4.0 kb)
447 (4.0 kb)
94 (32 kb)
341 (1.1 kb)
340 (1.1 kb)
339 (1.1 kb)
338 (1.1 kb)
666 (237 kb)
重叠群 15
672 (237 kb)
278 (1.1 kb)
277 (1.1 kb)
276 (1.1 kb)
95 (32 kb)
345 (4.0 kb)
344 (4.0 kb)
343 (4.0 kb)
196 (3.9 kb)

图6C

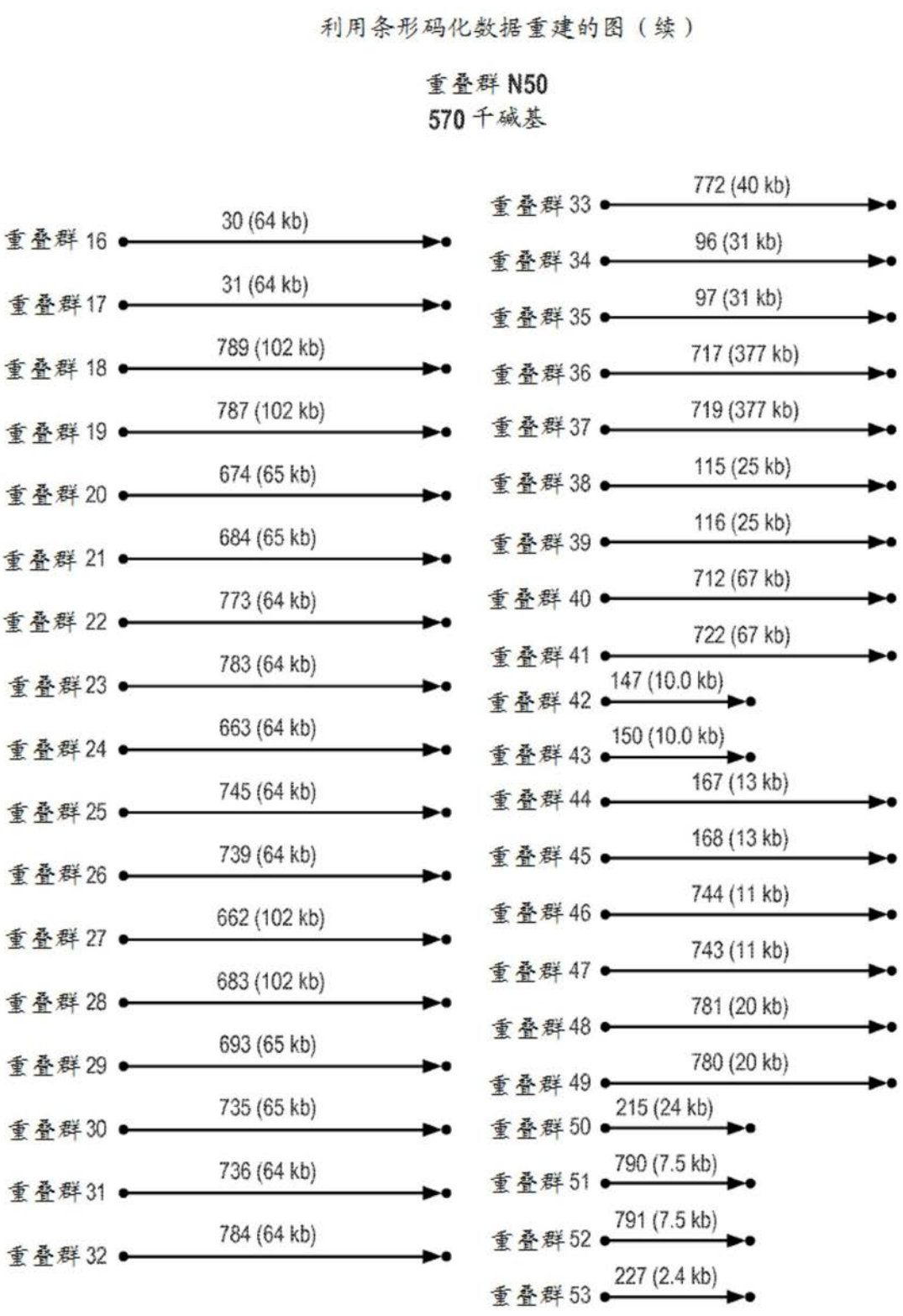

利用条形码化数据重建的图（续）
重叠群 N50
570 千碱基
重叠群 16 30 (64 kb)
重叠群 17 31 (64 kb)
重叠群 18 789 (102 kb)
重叠群 19 787 (102 kb)
重叠群 20 674 (65 kb)
重叠群 21 684 (65 kb)
重叠群 22 773 (64 kb)
重叠群 23 783 (64 kb)
重叠群 24 663 (64 kb)
重叠群 25 745 (64 kb)
重叠群 26 739 (64 kb)
重叠群 27 662 (102 kb)
重叠群 28 683 (102 kb)
重叠群 29 693 (65 kb)
重叠群 30 735 (65 kb)
重叠群 31 736 (64 kb)
重叠群 32 784 (64 kb)
重叠群 33 772 (40 kb)
重叠群 34 96 (31 kb)
重叠群 35 97 (31 kb)
重叠群 36 717 (377 kb)
重叠群 37 719 (377 kb)
重叠群 38 115 (25 kb)
重叠群 39 116 (25 kb)
重叠群 40 712 (67 kb)
重叠群 41 722 (67 kb)
重叠群 42 147 (10.0 kb)
重叠群 43 150 (10.0 kb)
重叠群 44 167 (13 kb)
重叠群 45 168 (13 kb)
重叠群 46 744 (11 kb)
重叠群 47 743 (11 kb)
重叠群 48 781 (20 kb)
重叠群 49 780 (20 kb)
重叠群 50 215 (24 kb)
重叠群 51 790 (7.5 kb)
重叠群 52 791 (7.5 kb)
重叠群 53 227 (2.4 kb)

图6D

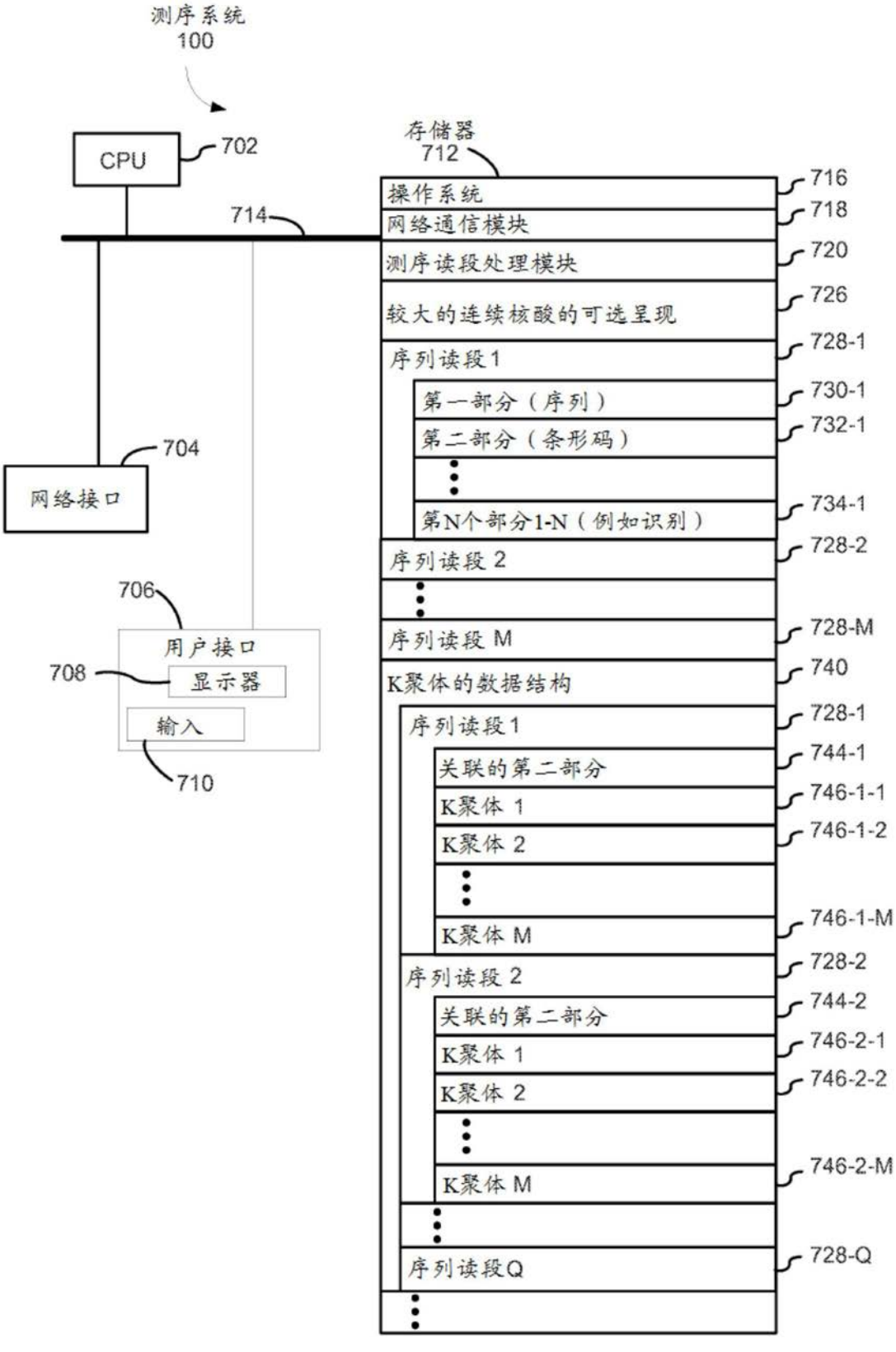

测序系统
100
CPU
702
存储器
712
操作系统
716
网络通信模块
718
714
测序读段处理模块
720
较大的连续核酸的可选呈现
726
序列读段1
728-1
第一部分（序列）
730-1
第二部分（条形码）
732-1
第N个部分1-N（例如识别）
734-1
704
网络接口
序列读段 2
728-2
706
序列读段 M
728-M
用户接口
708
显示器
K聚体的数据结构
740
输入
序列读段1
728-1
关联的第二部分
744-1
K聚体 1
746-1-1
K聚体 2
746-1-2
K聚体 M
746-1-M
710
序列读段 2
728-2
关联的第二部分
744-2
K聚体 1
746-2-1
K聚体 2
746-2-2
K聚体 M
746-2-M
序列读段Q
728-Q

图7

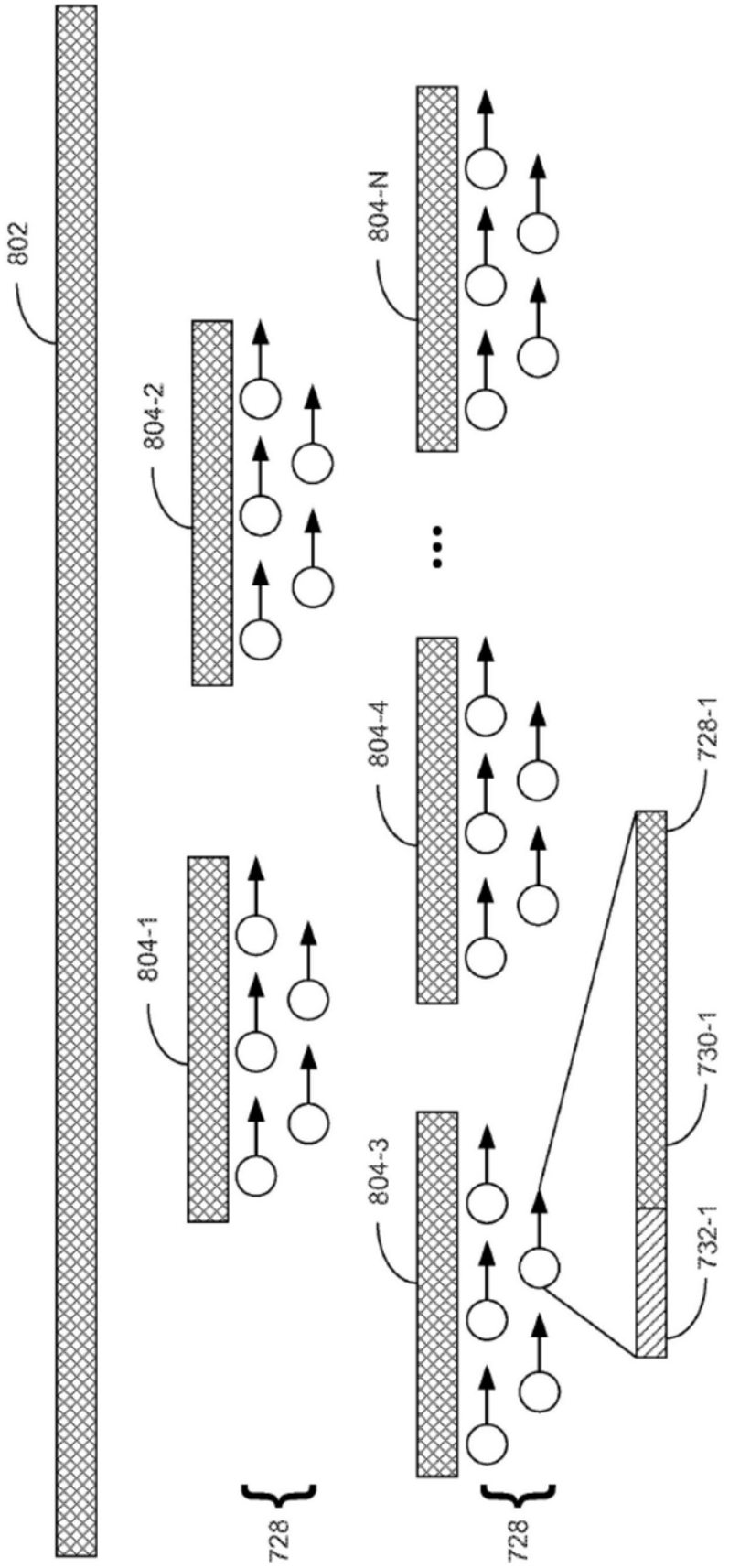


图8

3聚体

h(W)	W	h(W)	W	h(W)	W	h(W)	W
0	AAA	16	CAA	32	GAA	48	TAA
1	AAC	17	CAC	33	GAC	49	TAC
2	AAG	18	CAG	34	GAG	50	TAG
3	AAT	19	CAT	35	GAT	51	TAT
4	ACA	20	CCA	36	GCA	52	TCA
5	ACC	21	CCC	37	GCC	53	TCC
6	ACG	22	CCG	38	GCG	54	TCG
7	ACT	23	CCT	39	GCT	55	TCT
8	AGA	24	CGA	40	GGA	56	TGA
9	AGC	25	CGC	41	GGC	57	TGC
10	AGG	26	CGG	42	GGG	58	TGG
11	AGT	27	CGT	43	GGT	59	TGT
12	ATA	28	CTA	44	GTA	60	TTA
13	ATC	29	CTC	45	GTC	61	TTC
14	ATG	30	CTG	46	GTG	62	TTG
15	ATT	31	CTT	47	GTT	63	TTT

图9

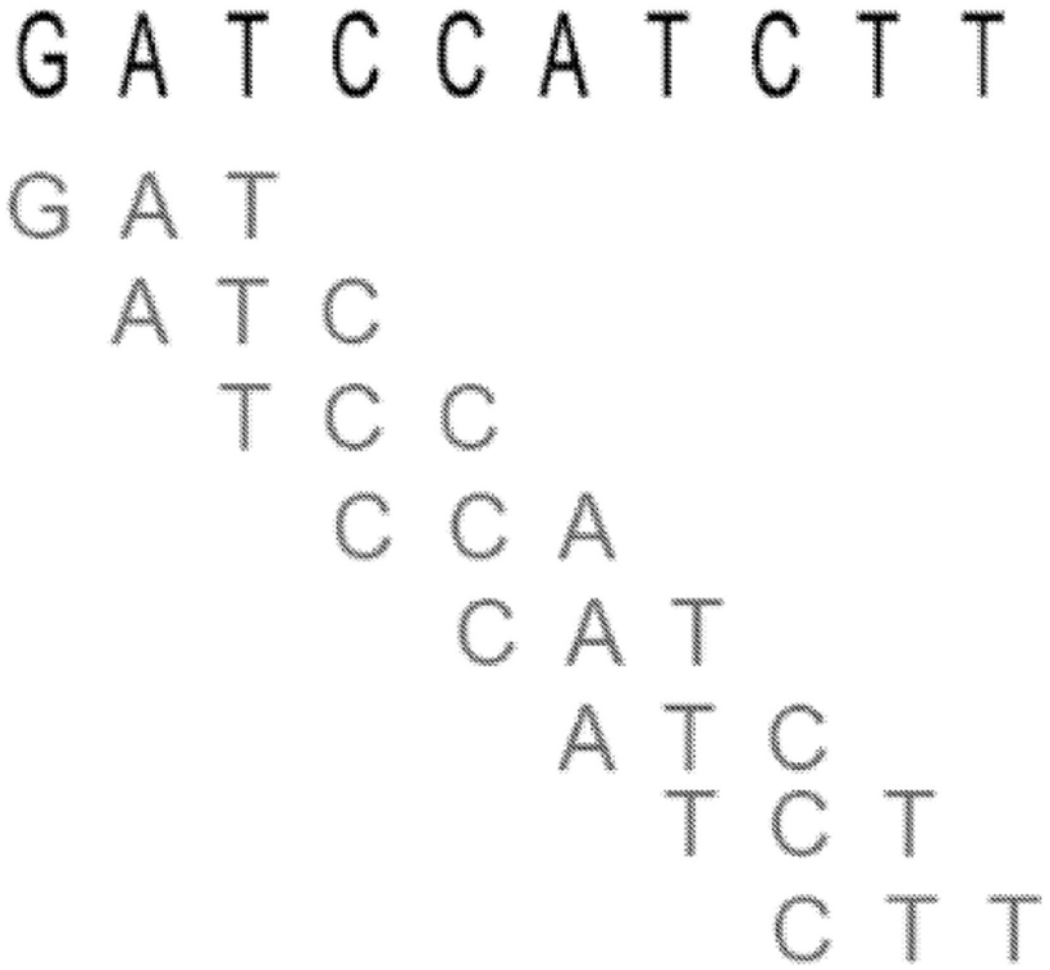


图10

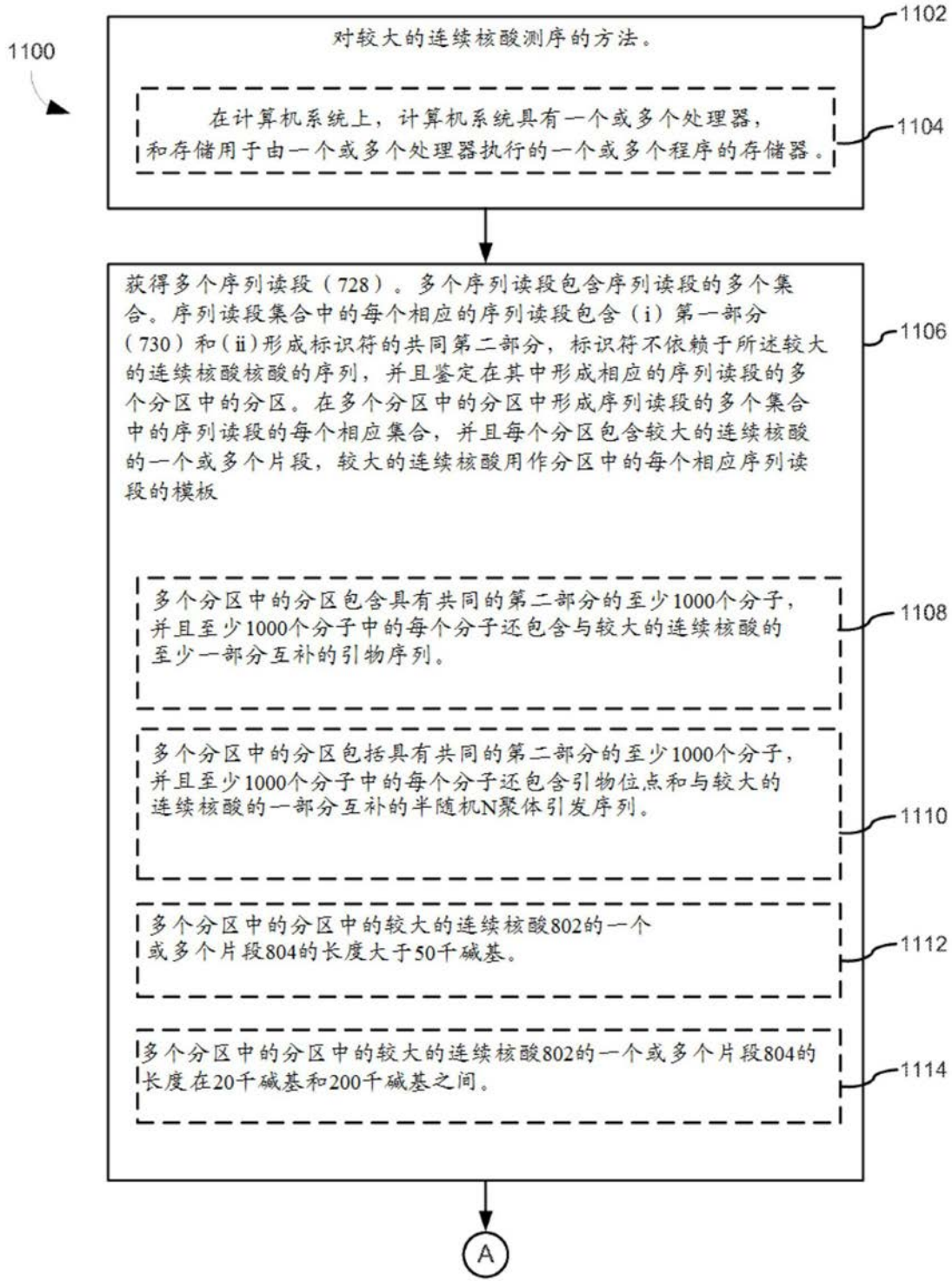

1100
对较大的连续核酸测序的方法。
1102
在计算机系统上，计算机系统具有一个或多个处理器，
和存储用于由一个或多个处理器执行的一个或多个程序的存储器。
1104
获得多个序列读段（728）。多个序列读段包含序列读段的多个集合。序列读段集合中的每个相应的序列读段包含（i）第一部分（730）和（ii）形成标识符的共同第二部分，标识符不依赖于所述较大的连续核酸核酸的序列，并且鉴定在其中形成相应的序列读段的多个分区中的分区。在多个分区中的分区中形成序列读段的多个集合中的序列读段的每个相应集合，并且每个分区包含较大的连续核酸的一个或多个片段，较大的连续核酸用作分区中的每个相应序列读段的模板
1106
多个分区中的分区包含具有共同的第二部分的至少1000个分子，并且至少1000个分子中的每个分子还包含与较大的连续核酸的至少一部分互补的引物序列。
1108
多个分区中的分区包括具有共同的第二部分的至少1000个分子，并且至少1000个分子中的每个分子还包含引物位点和与较大的连续核酸的一部分互补的半随机N聚体引发序列。
1110
多个分区中的分区中的较大的连续核酸802的一个或多个片段804的长度大于50千碱基。
1112
多个分区中的分区中的较大的连续核酸802的一个或多个片段804的长度在20千碱基和200千碱基之间。
1114
A

图11A

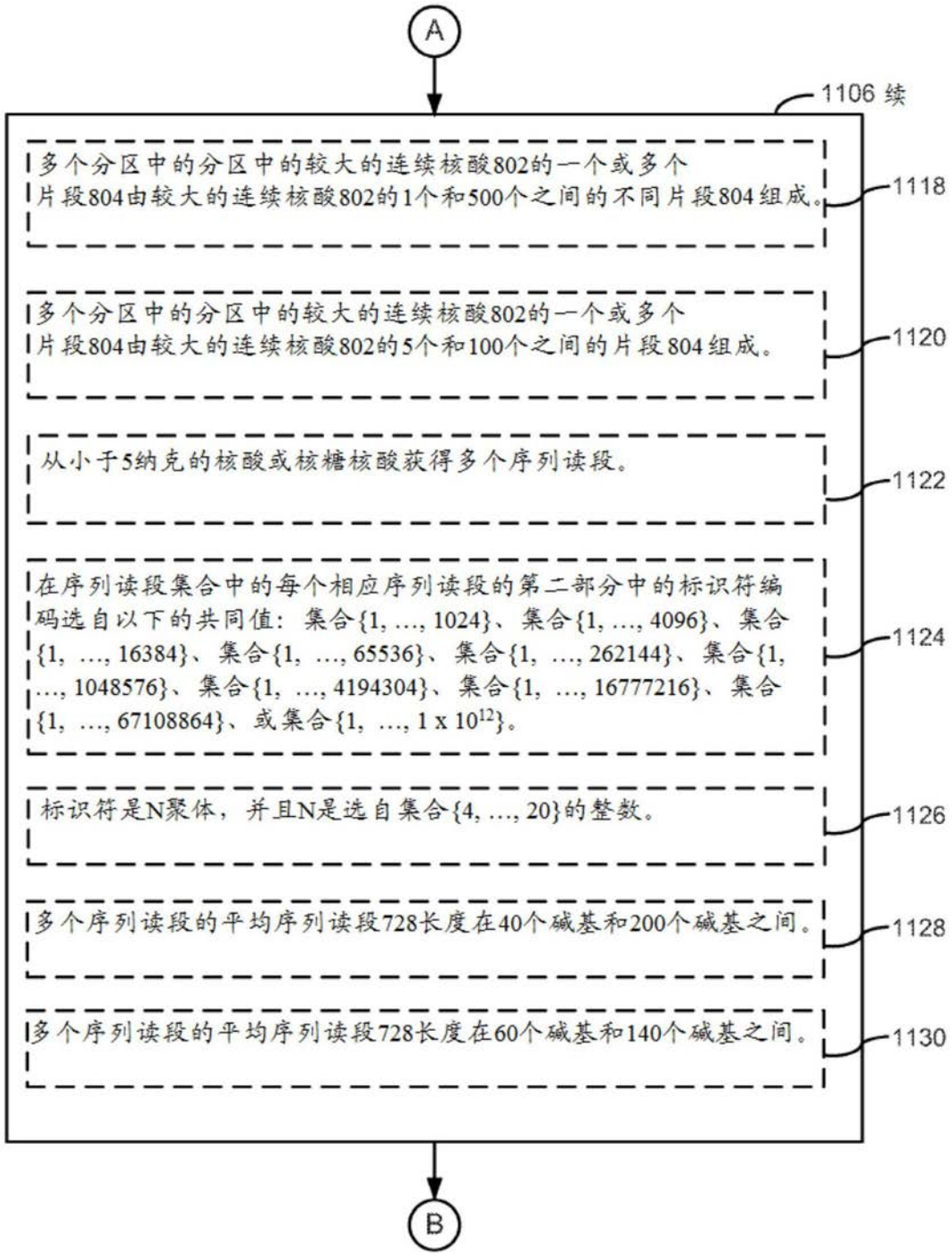

A
1106 续
多个分区中的分区中的较大的连续核酸802的一个或多个片段804由较大的连续核酸802的1个和500个之间的不同片段804组成。
1118
多个分区中的分区中的较大的连续核酸802的一个或多个片段804由较大的连续核酸802的5个和100个之间的片段804组成。
1120
从小于5纳克的核酸或核糖核酸获得多个序列读段。
1122
在序列读段集合中的每个相应序列读段的第二部分中的标识符编码选自以下的共同值：集合{1, ..., 1024}、集合{1, ..., 4096}、集合{1, ..., 16384}、集合{1, ..., 65536}、集合{1, ..., 262144}、集合{1, ..., 1048576}、集合{1, ..., 4194304}、集合{1, ..., 16777216}、集合{1, ..., 67108864}、或集合{1, ..., 1 x 10^12}。
1124
标识符是N聚体，并且N是选自集合{4, ..., 20}的整数。
1126
多个序列读段的平均序列读段728长度在40个碱基和200个碱基之间。
1128
多个序列读段的平均序列读段728长度在60个碱基和140个碱基之间。
1130
B

图11B

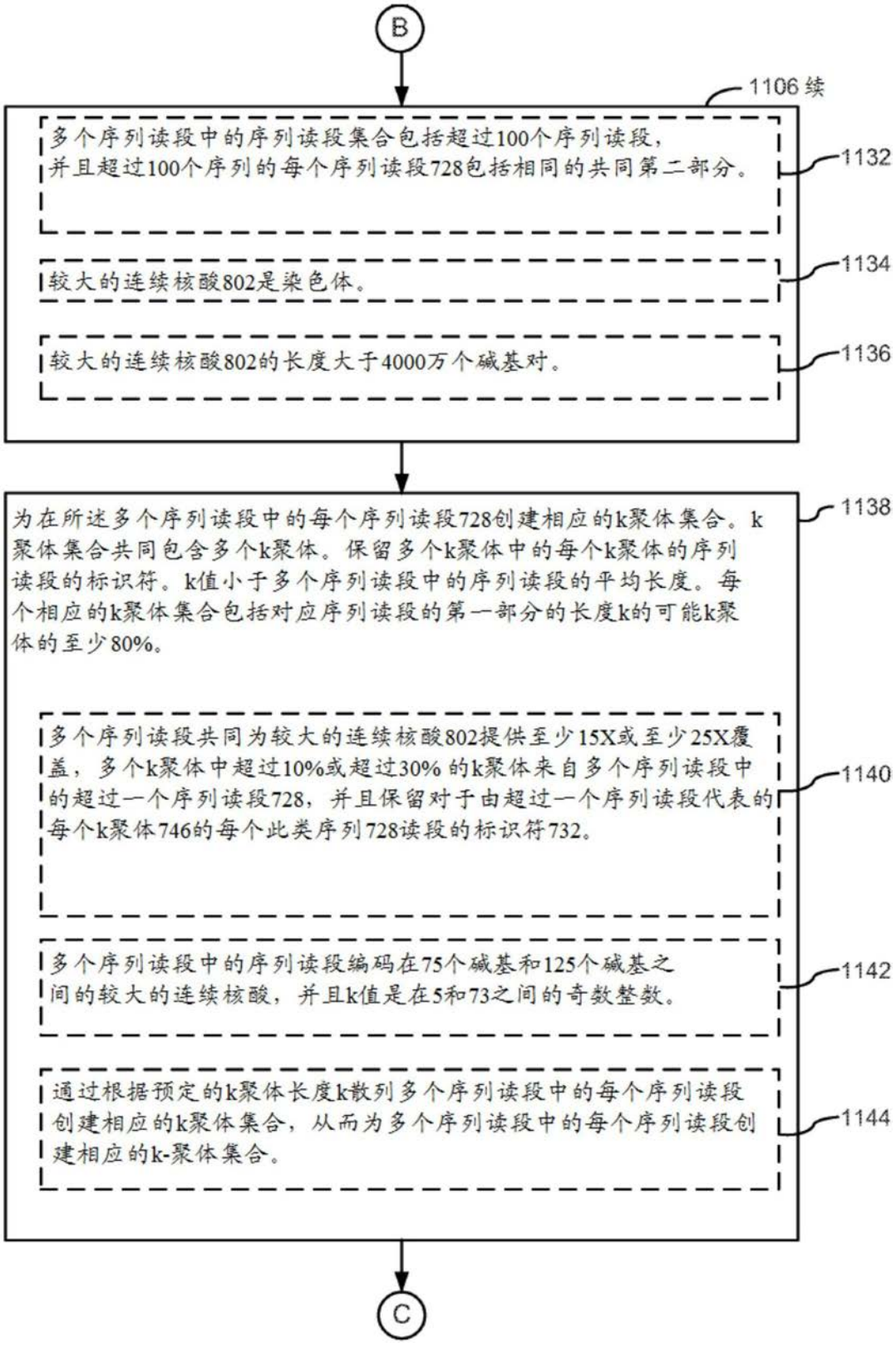

B
1106 续
多个序列读段中的序列读段集合包括超过100个序列读段，并且超过100个序列的每个序列读段728包括相同的共同第二部分。
1132
较大的连续核酸802是染色体。
1134
较大的连续核酸802的长度大于4000万个碱基对。
1136
为在所述多个序列读段中的每个序列读段728创建相应的k聚体集合。k聚体集合共同包含多个k聚体。保留多个k聚体中的每个k聚体的序列读段的标识符。k值小于多个序列读段中的序列读段的平均长度。每个相应的k聚体集合包括对应序列读段的第一部分的长度k的可能k聚体的至少80%。
1138
多个序列读段共同为较大的连续核酸802提供至少15X或至少25X覆盖，多个k聚体中超过10%或超过30%的k聚体来自多个序列读段中的超过一个序列读段728，并且保留对于由超过一个序列读段代表的每个k聚体746的每个此类序列728读段的标识符732。
1140
多个序列读段中的序列读段编码在75个碱基和125个碱基之间的较大的连续核酸，并且k值是在5和73之间的奇数整数。
1142
通过根据预定的k聚体长度k散列多个序列读段中的每个序列读段创建相应的k聚体集合，从而为多个序列读段中的每个序列读段创建相应的k-聚体集合。
1144
C

图11C

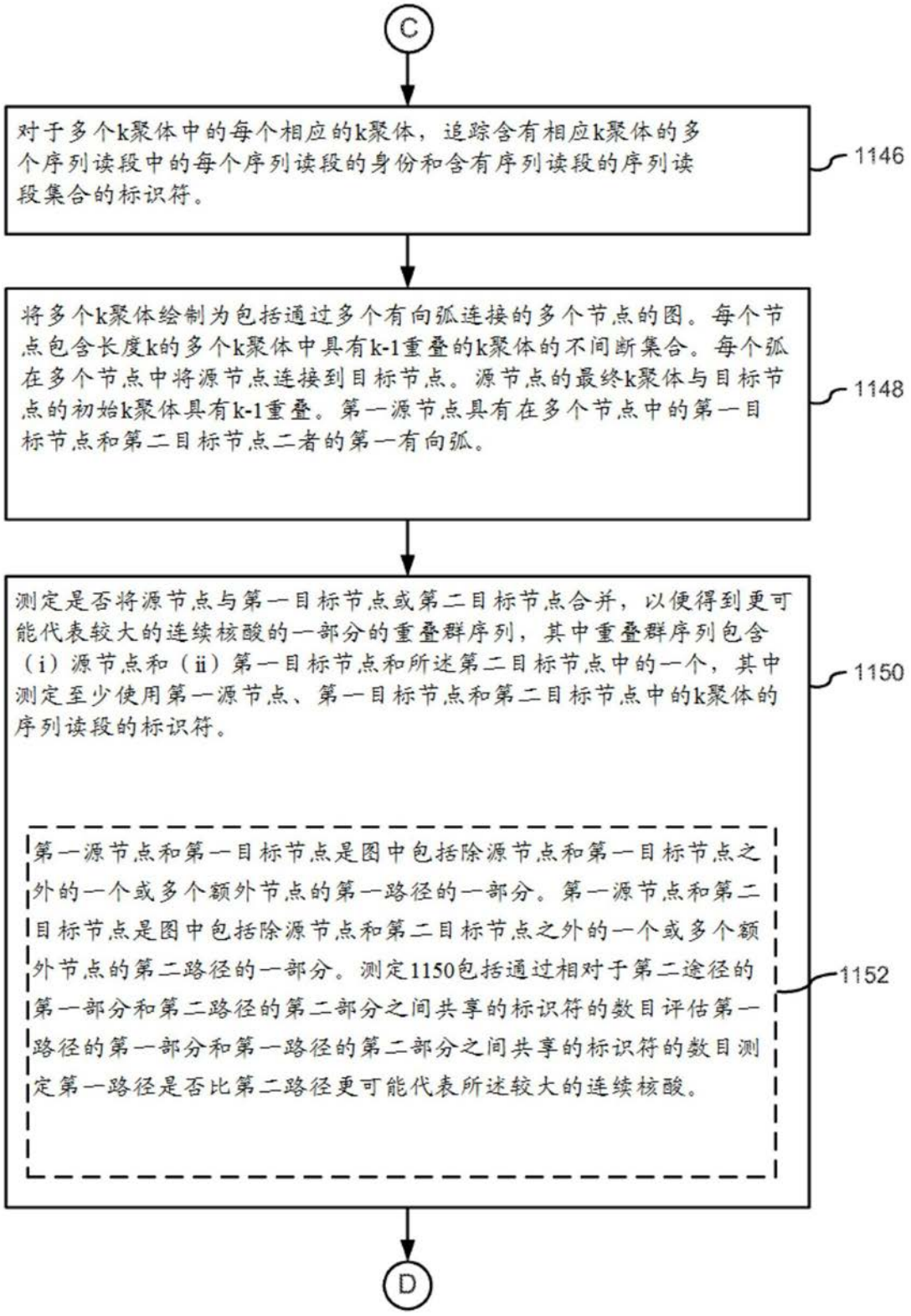

C
对于多个k聚体中的每个相应的k聚体，追踪含有相应k聚体的多个序列读段中的每个序列读段的身份和含有序列读段的序列读段集合的标识符。
1146
将多个k聚体绘制为包括通过多个有向弧连接的多个节点的图。每个节点包含长度k的多个k聚体中具有k-1重叠的k聚体的不间断集合。每个弧在多个节点中将源节点连接到目标节点。源节点的最终k聚体与目标节点的初始k聚体具有k-1重叠。第一源节点具有在多个节点中的第一目标节点和第二目标节点二者的第一有向弧。
1148
测定是否将源节点与第一目标节点或第二目标节点合并，以便得到更可能代表较大的连续核酸的一部分的重叠群序列，其中重叠群序列包含（i）源节点和（ii）第一目标节点和所述第二目标节点中的一个，其中测定至少使用第一源节点、第一目标节点和第二目标节点中的k聚体的序列读段的标识符。
1150
第一源节点和第一目标节点是图中包括除源节点和第一目标节点之外的一个或多个额外节点的第一路径的一部分。第一源节点和第二目标节点是图中包括除源节点和第二目标节点之外的一个或多个额外节点的第二路径的一部分。测定1150包括通过相对于第二途径的第一部分和第二路径的第二部分之间共享的标识符的数目评估第一路径的第一部分和第一路径的第二部分之间共享的标识符的数目测定第一路径是否比第二路径更可能代表所述较大的连续核酸。
1152
D

图11D

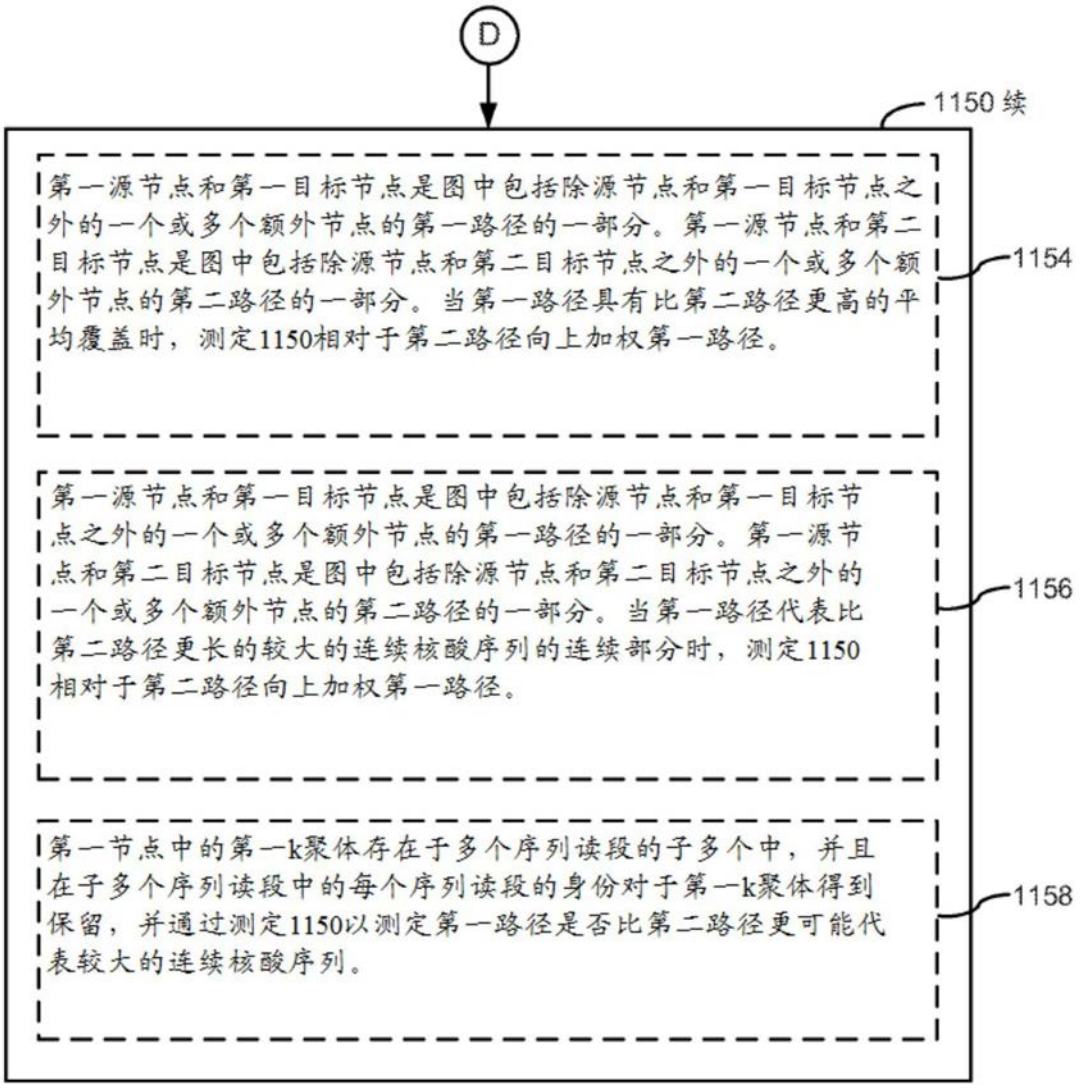

D
1150 续
第一源节点和第一目标节点是图中包括除源节点和第一目标节点之外的一个或多个额外节点的第一路径的一部分。第一源节点和第二目标节点是图中包括除源节点和第二目标节点之外的一个或多个额外节点的第二路径的一部分。当第一路径具有比第二路径更高的平均覆盖时，测定1150相对于第二路径向上加权第一路径。
1154
第一源节点和第一目标节点是图中包括除源节点和第一目标节点之外的一个或多个额外节点的第一路径的一部分。第一源节点和第二目标节点是图中包括除源节点和第二目标节点之外的一个或多个额外节点的第二路径的一部分。当第一路径代表比第二路径更长的较大的连续核酸序列的连续部分时，测定1150相对于第二路径向上加权第一路径。
1156
第一节点中的第一k聚体存在于多个序列读段的子多个中，并且在子多个序列读段中的每个序列读段的身份对于第一k聚体得到保留，并通过测定1150以测定第一路径是否比第二路径更可能代表较大的连续核酸序列。
1158

图11E

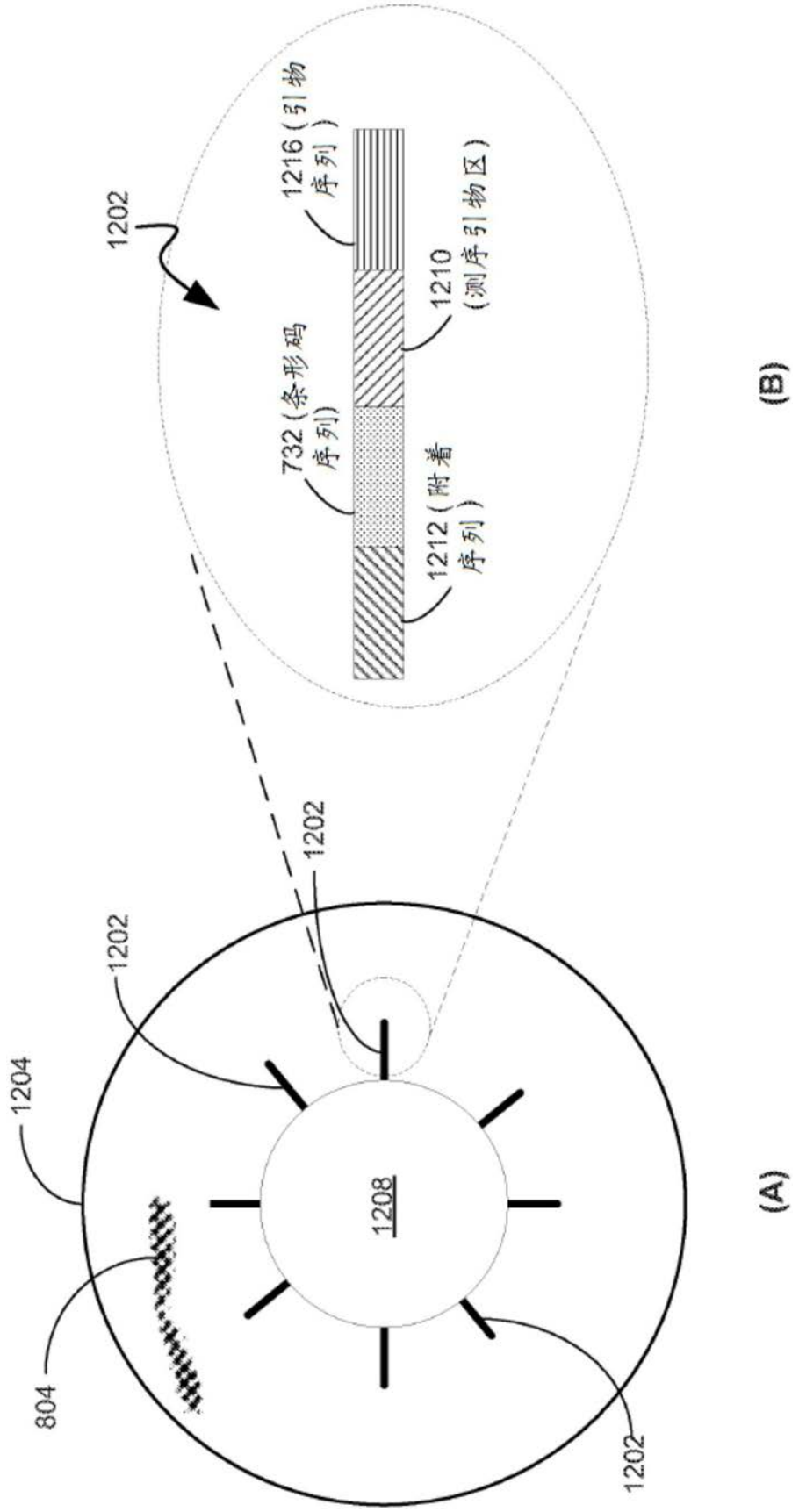

1202
1216（引物序列）
1210（测序引物区）
732（条形码序列）
1212（附着序列）
(B)
1202
1202
1204
1208
804
1202
(A)

图12A

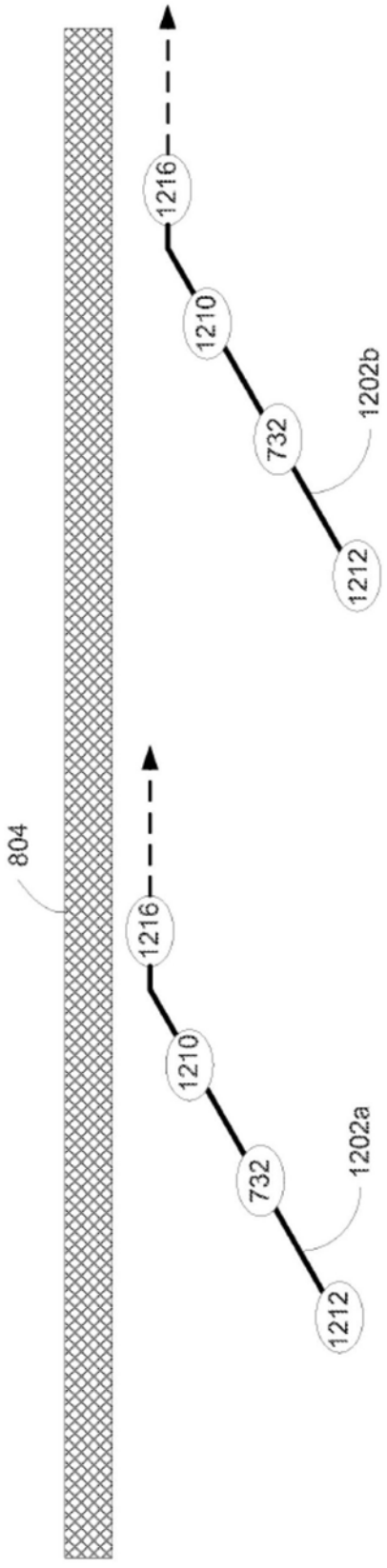


图12B

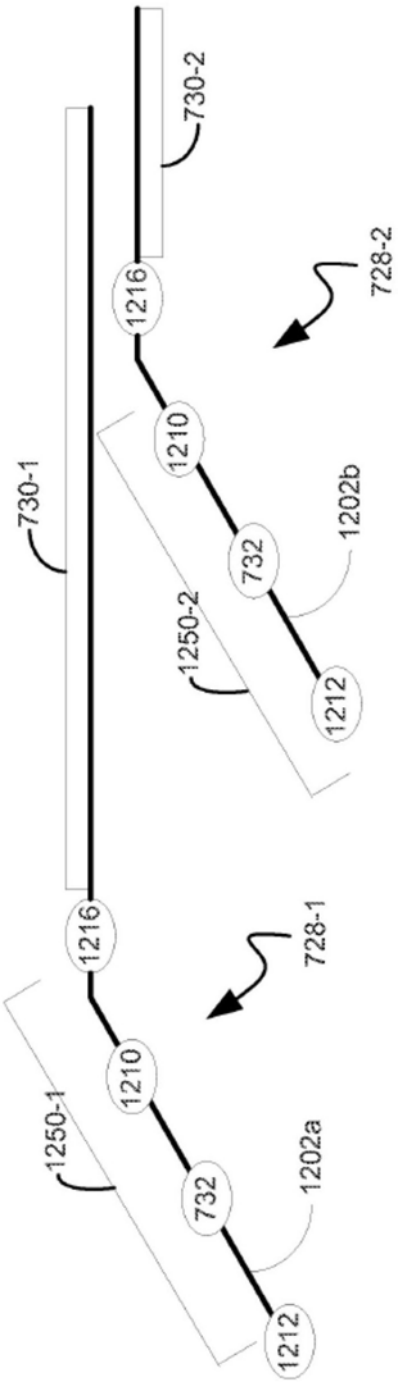

730-2
728-2
1216
1210
730-1
1202b
732
1250-2
1212
1216
728-1
1210
1250-1
1202a
732
1212

图12C

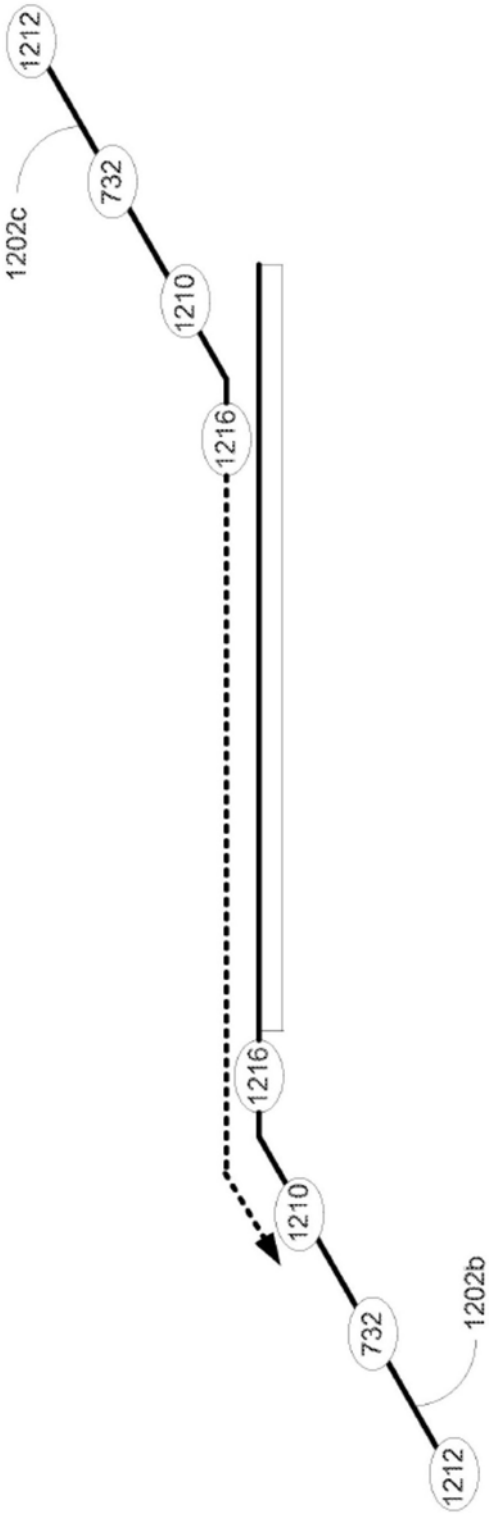


图12D

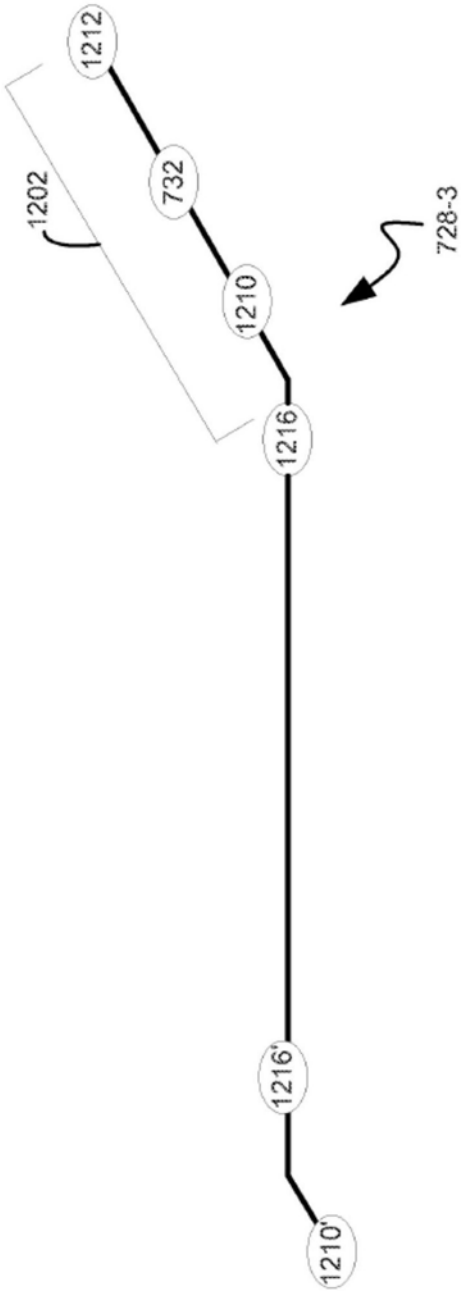


图12E

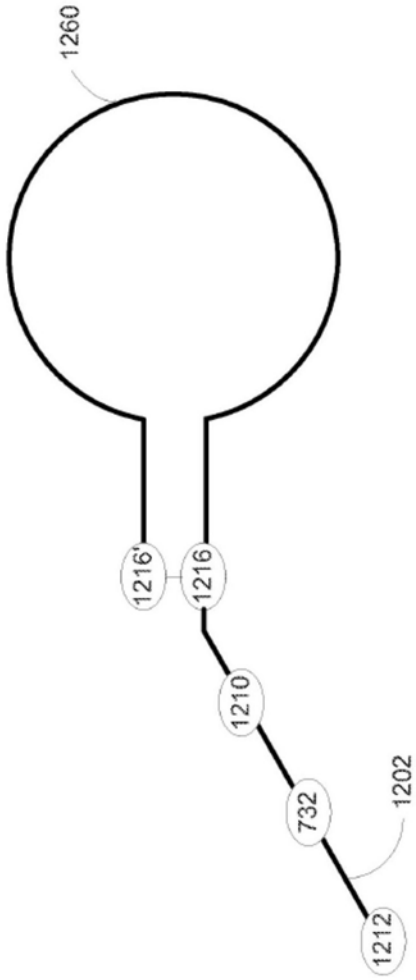


图12F

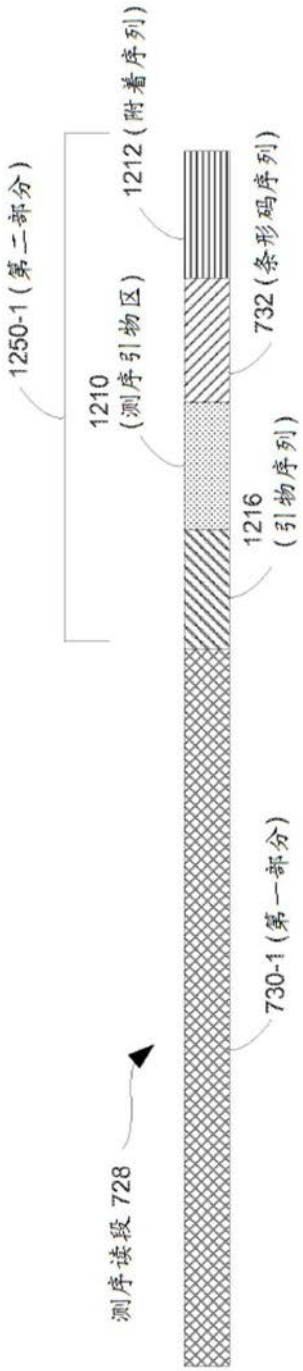

1250-1（第二部分）
1212（附着序列）
1210
（测序引物区）
732（余形码序列）
1216
（引物序列）
730-1（第一部分）
测序读段 728

图12G

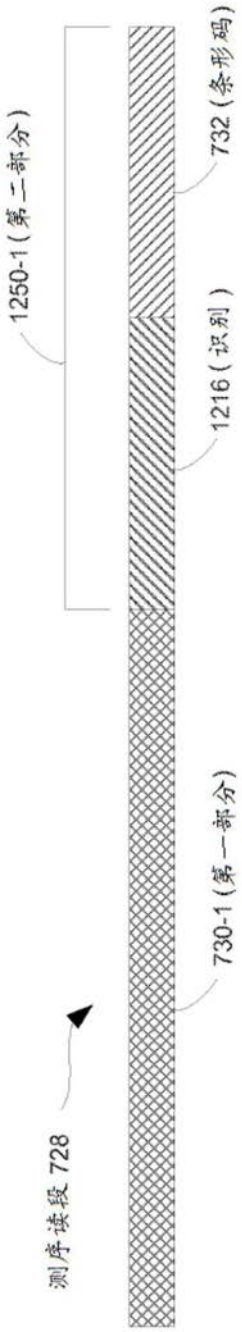


图12H

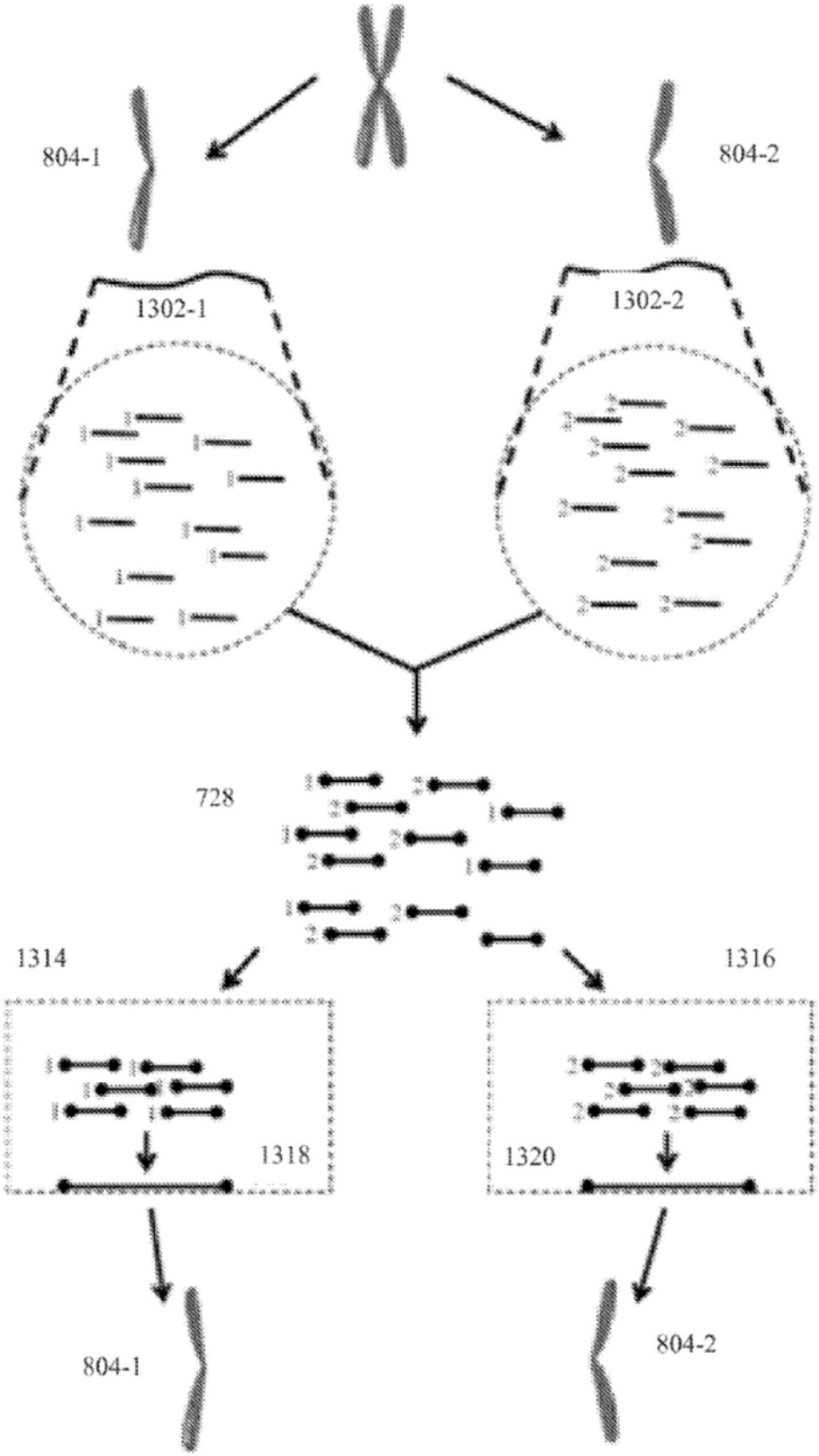


图13

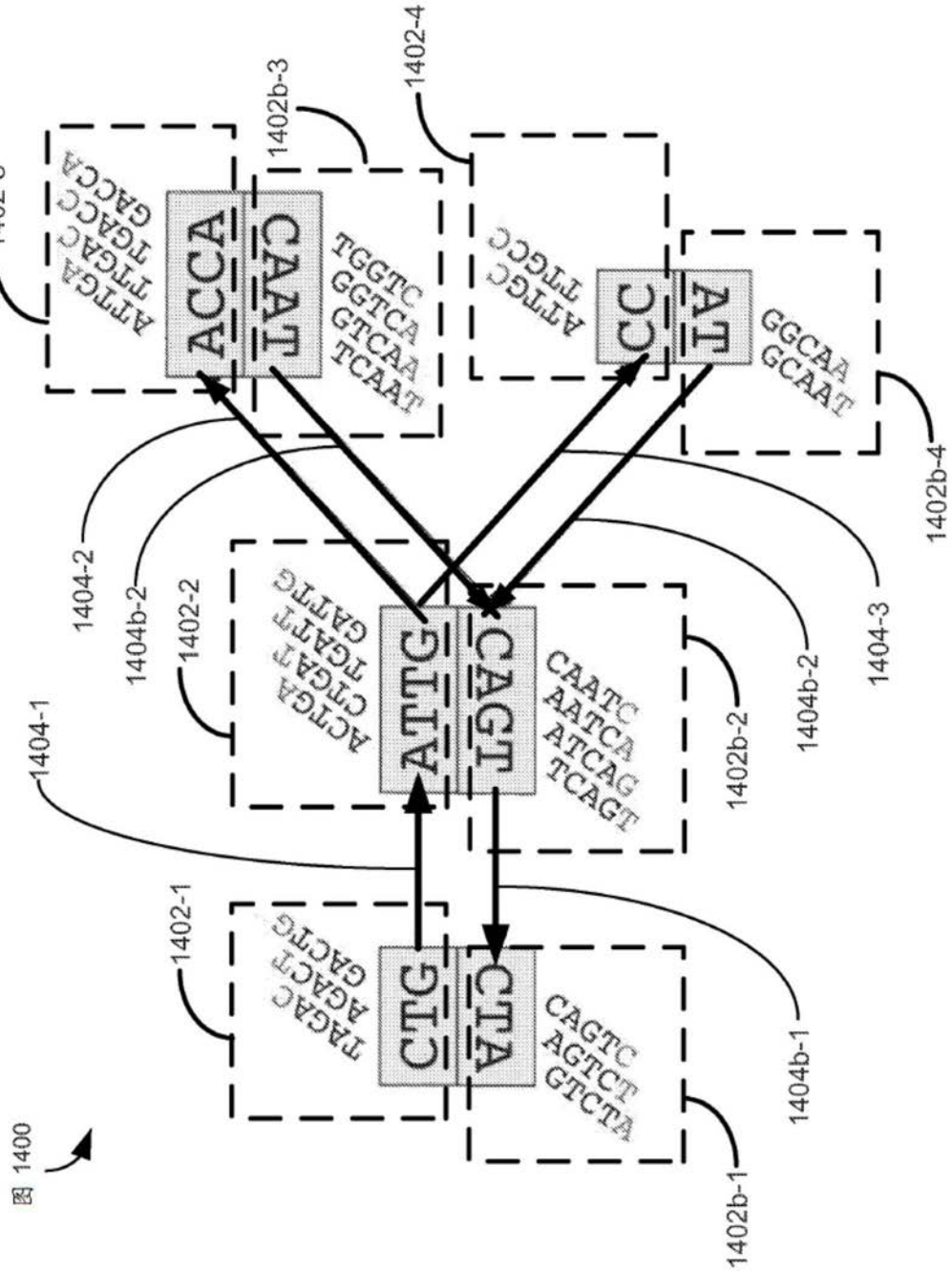

1402-4
1402b-3
1402-3
ACCA
CAAT
TGGTC
GGTCA
GTCAA
TCAAT
ATTGA
TTGAC
TGACC
GACCA
ATTGC
TTGCC
CC
AT
GGCAA
GCAAT
1402b-4
1404-2
1404b-2
1402-2
1404-3
1404b-2
1402b-2
1404-1
ACTGA
CTGAT
TGATT
GATTG
ATTG
CAGT
CAATC
AATCA
ATCAG
TCAGT
CTG
CTA
TAGAC
AGACT
GACTG
CAGTC
AGTCT
GTCTA
1402-1
1404b-1
1402b-1
图 1400

图14

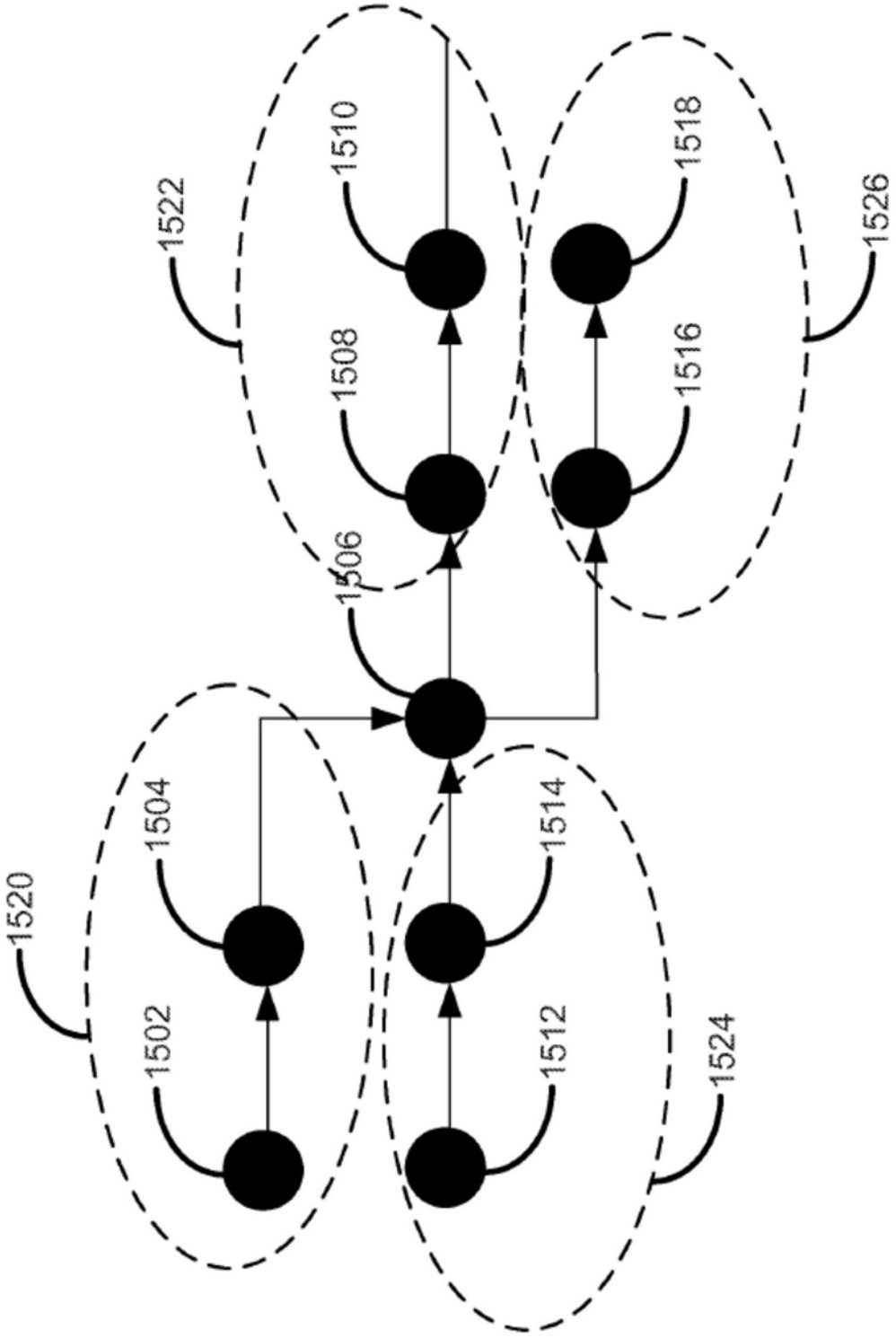

1522
1510
1518
1526
1508
1516
1506
1520
1504
1514
1502
1512
1524

图15