(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 108368542 B
(45)授权公告日 2022.04.08

(21)申请号 201680074565.1
(22)申请日 2016.10.18
(65)同一申请的已公布的文献号
申请公布号 CN 108368542 A
(43)申请公布日 2018.08.03
(30)优先权数据
62/243,591 2015.10.19 US
62/243,576 2015.10.19 US
62/255,953 2015.11.16 US
62/294,198 2016.02.11 US
(85)PCT国际申请进入国家阶段日
2018.06.19
(86)PCT国际申请的申请数据
PCT/US2016/057557 2016.10.18
(87)PCT国际申请的公布数据
WO2017/070123 EN 2017.04.27
(73)专利权人 多弗泰尔基因组学有限责任公司
地址 美国加利福尼亚州
(72)发明人 理查德·E·格林 保罗·哈特利 克里斯多佛·特罗尔 闵艾艾
(74)专利代理机构 北京安信方达知识产权代理有限公司 11262
代理人 刘晓杰
(51)Int.Cl.
C12Q 1/6806(2018.01)
C12Q 1/6809(2018.01)
C12Q 1/6869(2018.01)
(56)对比文件
US 2006252061 A1,2006.11.09
US 8076070 B2,2011.12.13
WO 2012106546 A2,2012.08.09
WO 2014121091 A1,2014.08.07
Tatjana Schutze 等.A calibrated diversity assay for nucleic acid libraries using DiStRO—a Diversity Standard of Random Oligonucleotides.《Nucleic Acids Research》.2009,第38卷(第4期),第1-5页. (续)
审查员 贾星航

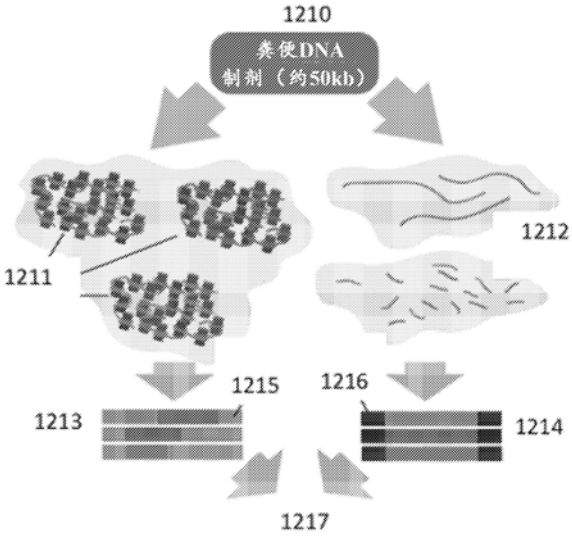
权利要求书2页 说明书114页
序列表6页 附图37页

(54)发明名称

用于基因组组装、单元型定相以及独立于靶标的核酸检测的方法

(57)摘要

本公开内容提供了用于组装真核或原核生物体的基因组的方法。本公开内容提供了用于单元型定相和宏基因组组装的方法。本公开内容提供了用于完成这些任务的流畅作业法,使得中间体无需被亲和标记物标记来促进与固体表面的结合。本公开内容还提供了用于从头生成异质宏基因组样品或从多个个体获得的样品中未知生物体的支架信息、连接信息和基因组信息的方法和组合物。所述方法的实施可允许对异质样品中的未培养或未鉴定的生物体的整个基因组进行从头测序,或确定从多个个体获得的包含核酸的样品中的核酸分子的连接信息。



CN 108368542 B

[转续页]

[接上页]

(56)对比文件

Andrew Adey 等.In vitro, long-range sequence information for de novo genome assembly via transposase contiguity.《Genome Research》.2014,第24卷第2041-2049页.

Ting Xie 等.De Novo Plant Genome Assembly Based on Chromatin Interactions: A Case Study of Arabidopsis thaliana.《Molecular Plant》.2014,第8卷第489-492页.

1.一种测定包含至少两种物种的异质样品中的核酸分子多样性的方法,该方法包括:

a)获得包含多个核酸分子的核酸样品,所述多个核酸分子经稳定剂处理,其中:(i)所述多个核酸分子中的至少第一核酸分子包含独立于第一共同磷酸二酯骨架而被保持在一起的第一核酸区段和第二核酸区段,其中在所述第一核酸区段与所述第二核酸区段之间裂解所述第一共同磷酸二酯骨架,以及(ii)所述多个核酸分子中的至少第二核酸分子包含独立于第二共同磷酸二酯骨架而被保持在一起的第三核酸区段和第四核酸区段,其中在所述第三核酸区段与所述第四核酸区段之间裂解所述第二共同磷酸二酯骨架;

b)用第一标记标记所述第一核酸区段并用第二标记标记所述第二核酸区段,使得所述第一核酸区段和所述第二核酸区段可被识别为由所述多个核酸分子中的所述第一核酸分子产生,并且用第三标记标记所述第三核酸区段并用第四标记标记所述第四核酸区段,使得所述第三核酸区段和所述第四核酸区段可被识别为由所述多个核酸分子中的所述第二核酸分子产生;

c)通过获得多个序列读取来对所述第一核酸区段的至少一个可识别部分及所述第一标记、所述第二核酸区段的可识别部分及所述第二标记、所述第三核酸区段的可识别部分及所述第三标记以及所述第四核酸区段的可识别部分及所述第四标记进行测序;

d)至少构建包含来自所述第一核酸区段和所述第二核酸区段的第一序列支架以及来自所述第三核酸区段和所述第四核酸区段的第二序列;使得所述多个核酸分子的多个区段的序列被分配给所述第一序列支架或所述第二序列支架中的至少一个;以及

e)计算所构架的多个序列支架的数目,其中当确定标记的核酸区段来自相同的核酸分子时,将标记的核酸分子的序列分配给共同的支架;

其中计算的序列支架的数目与所述异质样品的物种多样性对应。

2.根据权利要求1所述的方法,其中标记所述第一核酸区段和所述第二核酸区段包括将第一寡核苷酸添加至所述第一核酸区段以及将第二寡核苷酸添加至所述第二核酸区段,其中所述第一寡核苷酸和所述第二寡核苷酸包含共同的序列。

3.根据权利要求2所述的方法,其中将具有所述共同的寡核苷酸序列的核酸区段的序列分配给共同的支架。

4.根据权利要求3所述的方法,其包括将所述第一核酸区段的所述可识别部分映射至叠连群数据集,并且将所述叠连群数据集的任何匹配的叠连群纳入所述共同的支架。

5.根据权利要求4所述的方法,其中同时生成所述叠连群数据集。

6.根据权利要求4所述的方法,其中从数据库获得所述叠连群数据集。

7.根据权利要求1所述的方法,其中标记所述第一核酸区段和所述第二核酸区段包括连接所述第一核酸区段与所述第二核酸区段,并且其中将所述第一核酸区段和所述第二核酸区段分配给共同的支架。

8.根据权利要求7所述的方法,其包括将所述第一核酸区段的所述可识别部分映射至叠连群数据集,并且将所述叠连群数据集的任何匹配的叠连群纳入所述共同的支架。

9.根据权利要求8所述的方法,其中同时生成所述叠连群数据集。

10.根据权利要求8所述的方法,其中从数据库获得所述叠连群数据集。

11.根据权利要求1所述的方法,其中所述异质样品包含多个等位基因变体。

12.根据权利要求11所述的方法,其中等位基因变体的数目大于支架的数目。

13.根据权利要求11所述的方法，其中等位基因变体的数目等于生成的支架的数目。

14.根据权利要求1-13中任一项所述的方法，其中在所述获得所述核酸样品之后，裂解所述磷酸二酯骨架。

15.根据权利要求1-13中任一项所述的方法，其中使所述稳定剂包含交联剂。

16.根据权利要求1-13中任一项所述的方法，其中所述核酸样品为FFPE样品。

17.根据权利要求1-13中任一项所述的方法，其包括使所述核酸样品与逆转录酶接触。

18.根据权利要求1所述的方法，其包括在核酸序列数据库中搜索至少一个所述支架。

19.根据权利要求18所述的方法，其包括如果唯一映射至所述支架的核酸序列在所述数据库中不存在，则将所述支架分类为未知的。

20.根据权利要求1-13中任一项所述的方法，其中所述异质样品包含映射至共同物种的至少两个个体的核酸。

21.根据权利要求1-13中任一项所述的方法，其中所述异质样品包含映射至共同物种的至少三个个体的核酸。

22.根据权利要求1-13中任一项所述的方法，其中所述异质样品包含映射至至少三个物种的核酸。

23.根据权利要求1-13中任一项所述的方法，其中所述异质样品包含映射至至少四个物种的核酸。

24.根据权利要求1-13中任一项所述的方法，其中在不参考外源序列信息的情况下，多个序列读取组装成至少两个核酸支架。

25.根据权利要求1-13中任一项所述的方法，其中在不参考外源序列信息的情况下，多个序列读取组装成至少三个核酸支架。

26.根据权利要求1-13中任一项所述的方法，其中所述多个序列读取组装成至少两个核酸支架，使得至少50％的第一基因组和至少50％的第二基因组被表示在所述至少两个核酸支架中。

27.根据权利要求1-13中任一项所述的方法，其中所述多个序列读取组装成至少两个核酸支架，使得至少60％的第一基因组和至少60％的第二基因组被表示在所述至少两个核酸支架中。

28.根据权利要求1-13中任一项所述的方法，其中所述多个序列读取组装成至少两个核酸支架，使得至少70％的第一基因组和至少70％的第二基因组被表示在所述至少两个核酸支架中。

29.根据权利要求1-13中任一项所述的方法，其中所述多个序列读取组装成至少两个核酸支架，使得至少80％的第一基因组和至少80％的第二基因组被表示在所述至少两个核酸支架中。

30.根据权利要求1-13中任一项所述的方法，其中所述方法包括使用SPRI珠子。

31.根据权利要求1-13中任一项所述的方法，其中所述核酸样品包含不超过5微克的DNA。

用于基因组组装、单元型定相以及独立于靶标的核酸检测的方法

[0001] 交叉引用

[0002] 本申请要求于2015年10月19日提交的美国临时专利申请号62/243,576(通过引用以其全文并入本文)、于2015年10月19日提交的美国临时申请号62/243,591(通过引用以其全文并入本文)、于2015年11月16日提交的美国临时申请号62/255,953(通过引用以其全文并入本文)以及于2016年2月11日提交的美国临时专利申请号62/294,198(通过引用以其全文并入本文)的权益。

[0003] 关于联邦政府资助研究的声明

[0004] 本发明是根据国家人类基因组研究所的编号为5R44HG008719-02的合同在美国政府的支持下做出的。

背景技术

[0005] 在理论上和实践中,仍然难以产生高质量、高度连续的基因组序列。高通量测序允许对定居于具有生物医学、生态学或生物化学意义的众多环境中的生物体进行遗传分析。对通常含有难以培养的微生物的环境样品进行鸟枪法测序可揭示给定环境中的生物体内存在的基因和生物化学途径。小心过滤并分析这些数据还可以揭示数据中的读取之间系统发育关系的信号。然而,高质量从头组装这些高度复杂的数据集通常被认为是棘手的。

发明内容

[0006] 下一代测序(NGS)数据一直以来的缺点是由于读取长度短和插入片段大小相对较小而不能跨越基因组的大重复区域。这一缺陷显著地影响从头组装。因为基因组重排的性质和位置是不确定的,所以由长的重复区域隔开的叠连群不能被连接或重新测序。另外,因为变体不能确信地在长距离上与单元型相关联,所以定相信息是不能确定的。本公开内容通过采用合适的输入DNA生成跨越数百个千碱基至多达兆碱基的数量级的基因组距离的极长范围读取对(XLRP),可同时解决所有这些问题。这样的数据对于克服基因组(包括着丝粒)中的大重复区域所呈现的大量障碍可能是无价的;能实现成本效益好的从头组装;以及产生对于个性化医学而言完整性和准确度足够的重新测序数据。

[0007] 重要的是使用重构的染色质在距离非常远但分子连接的DNA区段之间形成缔合。本公开内容使得能够将远离的区段聚集在一起并通过染色质构象共价连接,从而物理连接DNA分子的先前远离的部分。后续的处理可允许确定缔合区段的序列,从而产生读取对,该读取对在基因组上的间隔延伸至输入DNA分子的全长。因为这些读取对来源于相同的分子,所以这些对还含有相位信息。

[0008] 健康和健身的许多方面受胃肠道中、皮肤上以及其他位置中丰富的微生物群落的影响。本文描述了揭示这样的微生物群落的全基因组复杂性的简单又有效的方法。这些技术可允许快速、准确且定量地测定在位置如人体(例如,肠道)和发现微生物群落的其他部位中存在的全部遗传组成成分。

[0009] 这样的技术包括体外邻近连接方法，例如，用于粪便宏基因组学应用。这些技术可提供用于从头宏基因组组装的有力且有效的方法，该方法将允许在诸如单个基因座分子计数或统计推断的方法以外进行研究和生物医学分析。

[0010] 本公开内容的技术可提供用于准确组装复杂的宏基因组群体的全部主要组分的单一、集成的工作流程。这些技术可使得能够全面理解微生物组(例如，肠道微生物组)影响人类、其他动物、植物、其他生命形式和环境的健康和疾病的方式。

[0011] 本文公开的技术可提供对样品如人类粪便样品中存在的微生物的多样性的有效捕获和表示。本文还公开了利用这些技术生成的丰富数据类型的宏基因组组装的计算方法。这样的计算方法可实现高度连续的支架化(scaffolding)和张量去卷积(strain deconvolution)。本公开内容的技术可提供强健的(robust)、防误操作的实验室方案和软件产品，所述实验室方案和软件产品可允许在数天内从小样品(例如，粪便样品)生成动态微生物环境(例如，人类肠道)的全面视图。

[0012] 在一些实施方案中，本公开内容提供了可采用远远少于先前所需要的数据产生高质量组装的方法。例如，本文公开的方法提供了来自仅两个通道的Illumina HiSeq数据的基因组组装。

[0013] 在其他实施方案中，本公开内容提供了可使用长距离读取对方法生成染色体水平定相的方法。例如，本文公开的方法可以以至少99％或更大的准确度对该个体的90％或更多的杂合单核苷酸多态性(SNP)进行定相。该准确度与通过实质上更昂贵且费力的方法所产生的定相相当。

[0014] 在一些实例中，可产生多达兆碱基规模的基因组DNA的片段的方法可与本文公开的方法一起使用。可生成长DNA片段来证实本发明方法生成跨越那些提取物提供的最长片段的读取对的能力。在一些情况下，可提取长度超过150kbp的DNA片段并使用该DNA片段生成XLRP文库。

[0015] 本公开内容提供了用于极大地促进和改善从头基因组组装的方法。本文公开的方法利用用于数据分析的方法，该方法允许来自一个或多个受试者的基因组的迅速且便宜的从头组装。本公开内容提出，本文公开的方法可用于多种应用，包括单元型定相和宏基因组分析。

[0016] 在某些实施方案中，本公开内容提供了用于基因组组装的方法，该方法包括以下步骤：生成多个叠连群；由通过探测染色体、染色质或重构的染色质的物理布局产生的数据生成多个读取对；将所述多个读取对映射或组装至所述多个叠连群；使用读取-映射或组装数据构建叠连群的邻接矩阵；以及分析所述邻接矩阵以确定穿过叠连群的路径，该路径代表叠连群相对于基因组的顺序和/或方向。在一些实施方案中，本公开内容提出，至少约90％的读取对通过获取每个读取到叠连群边缘的距离的函数来加权，以合并关于哪些读取对指示短范围接触和哪些读取对指示较长范围接触的信息。在其他实施方案中，可重新缩放所述邻接矩阵以降低一些叠连群上大量接触的权重，所述接触表示基因组的混杂区域，如针对调节染色质的支架化相互作用的一种或多种物质如转录抑制因子CTCF的保守结合位点。在其他实施方案中，本公开内容提供了用于人类受试者的基因组组装的方法，从而由人类受试者的DNA生成多个叠连群，并从而通过分析人类受试者的染色体、染色质或由该受试者的裸DNA制成的重构染色质生成多个读取对。

[0017] 在本文的一些实施方案中，一个益处是减少了分离被标记以便提供相位信息的复合体所需的步骤数。在现有技术的许多技术中，复合体包含标记的核酸或标记的缔合部分如蛋白质或纳米颗粒，例如生物素标记的，以促进复合体与采用例如亲和素或链霉亲和素标记的固体表面的结合。在本公开内容的一些方法和组合物中，固体表面包被有直接结合复合体或通过溶剂介导结合复合体的部分，使得该复合体不需要经配体修饰来促进与该固体表面的结合。本文考虑到许多部分，如亲水性部分、疏水性部分、带正电荷的部分、带负电荷的部分、PEG、多胺、氨基部分、聚羧酸部分或其他部分或部分的组合。在一些情况下，所述表面是SPRI表面，如直接地或通过溶剂与缔合部分-核酸复合体结合的SPRI表面。

[0018] 本公开内容提出，可通过使用鸟枪测序法生成多个叠连群，所述鸟枪测序法包括：使受试者的DNA的长序列段片段化为不确定大小的随机片段；使用高通量测序法对所述片段进行测序以生成多个测序读取；以及组装所述测序读取以形成多个叠连群。

[0019] 在某些实施方案中，本公开内容提出，可使用基于染色质捕获的技术，通过探测染色体、染色质或重构的染色质的物理布局生成多个读取对。在一些实施方案中，基于染色质捕获的技术包括使染色体、染色质或重构的染色质与固定剂如甲醛交联，以形成DNA-蛋白质交联；用一种或多种核酸酶(例如，限制酶)切割交联的DNA-蛋白质以生成包含粘性末端的多个DNA-蛋白质复合体；用含有一种或多种标志物如生物素的核苷酸补平该粘性末端，以产生钝性末端，随后将该钝性末端连接在一起；使所述多个DNA-蛋白质复合体片段化为片段；通过使用所述一种或多种标志物拉下(pull down)含有接头的片段；以及使用高通量测序法对含有接头的片段进行测序，以生成多个读取对。在一些实施方案中，用于本文公开的方法的多个读取对是由通过探测重构的染色质的物理布局所产生的数据生成的。

[0020] 在一些实施方案中，本公开内容提供了用于生成标记的序列的方法，该方法包括：使DNA分子与缔合分子结合；切割结合的DNA-蛋白质以生成包含区段末端的多个DNA-蛋白质复合体；使所述区段末端与标记(tag)连接；以及使用高通量测序法对含有接头的片段进行测序，以生成多个读取对。考虑许多结合DNA的缔合分子，包括狭义的(sensu stricto)染色质组分，如组蛋白，还包括更泛泛地定义的染色质组分，如DNA结合蛋白质、转录因子、核蛋白质、转座子，或非多肽DNA结合缔合分子，如具有包含DNA-亲和分子的表面的纳米颗粒。在一些情况下，使所述标记与区段末端连接，例如，使用连接酶或使用利用标记分子负载的转座酶。在一些情况下，包含共同标记的区段末端被分配给共同的起源分子，该共同的起源分子通常指示相位。在一些实施方案中，用于本文公开的方法的多个读取对是由通过探测重构的染色质的物理布局所产生的数据生成的。

[0021] 在各个实施方案中，本公开内容提出，可通过探测从培养的细胞或原代组织分离的染色体或染色质的物理布局来确定多个读取对。在其他实施方案中，可通过探测通过使从一个或多个受试者的样品获得的裸DNA与分离的组蛋白复合而形成的重构染色质的物理布局，来确定所述多个读取对。

[0022] 本公开内容提供了用于确定单元型定相的方法，该方法包括识别多个读取对中具有杂合性的一个或多个位点的步骤，其中等位基因变体的定相数据可通过识别包含杂合位点对的读取对来确定。

[0023] 在各个实施方案中，本公开内容提供了用于高通量细菌基因组组装的方法，该方法包括使用经修改的基于染色质捕获的方法通过探测多个微生物染色体的物理布局生成

多个读取对的步骤，所述经修改的基于染色质捕获的方法包括以下修改的步骤：从环境中收集微生物；添加固定剂，如甲醛，以在每个微生物细胞内形成交联，并且其中映射至不同叠连群的读取对指示哪些叠连群来自相同的物种。

[0024] 在一些实施方案中，本公开内容提供了用于基因组组装的方法，该方法包括：(a)生成多个叠连群；(b)由通过探测染色体、染色质或重构的染色质的物理布局生成的数据确定多个读取对；(c)将所述多个读取对映射至所述多个叠连群；(d)使用读取-映射数据构建叠连群的邻接矩阵；以及(e)分析所述邻接矩阵以确定穿过所述叠连群的路径，该路径代表叠连群相对于所述基因组的顺序和/或方向。

[0025] 本公开内容提供了使用基于染色质捕获的技术，通过探测染色体、染色质或重构的染色质的物理布局生成多个读取对的方法。在一些实施方案中，基于染色质捕获的技术包括(a)使染色体、染色质或重构的染色质与固定剂交联，以形成DNA-蛋白质交联；(b)用一种或多种核酸酶(例如，限制酶)切割交联的DNA-蛋白质以生成包含粘性末端的多个DNA-蛋白质复合体；(c)用含有一种或多种标志物的核苷酸补平所述粘性末端，以产生钝性末端，随后将该钝性末端连接在一起；(d)将所述多个DNA-蛋白质复合体剪切成片段；(e)通过使用一种或多种标志物拉下含有接头的片段；以及(f)使用高通量测序法对含有接头的片段进行测序，以生成多个读取对。

[0026] 在某些实施方案中，通过探测从培养的细胞或原代组织分离的染色体或染色质的物理布局来确定多个读取对。在其他实施方案中，通过探测通过使从一个或多个受试者的样品获得的裸DNA与分离的组蛋白复合而形成的重构染色质的物理布局，来确定所述多个读取对。

[0027] 在一些实施方案中，至少约60％、约70％、约80％、约90％、约95％或约99％或更多的所述多个读取对通过获取读取到叠连群边缘的距离的函数来加权，以合并与较长接触相比更高概率的较短接触。在一些实施方案中，重新缩放所述邻接矩阵以降低一些叠连群上大量接触的权重，所述接触表示基因组的混杂区域。

[0028] 在某些实施方案中，基因组的混杂区域包括针对调节染色质的支架化相互作用的一种或多种物质的一个或多个保守结合位点。在一些实例中，所述物质为转录抑制因子CTCF。

[0029] 在一些实施方案中，本文公开的方法提供了人类受试者的基因组组装，从而由人类受试者的DNA生成多个叠连群，并从而通过分析人类受试者的染色体、染色质或由该受试者的裸DNA制成的重构染色质生成多个读取对。

[0030] 在其他实施方案中，本公开内容提供了用于确定单元型定相的方法，该方法包括识别多个读取对中具有杂合性的一个或多个位点，其中等位基因变体的定相数据可通过识别包含杂合位点对的读取对来确定。

[0031] 在其他实施方案中，本公开内容提供了用于宏基因组组装的方法，其中使用经修改的基于染色质捕获的方法通过探测多个微生物染色体的物理布局生成多个读取对，所述经修改的基于染色质捕获的方法包括：从环境中收集微生物；以及添加固定剂，以在每个微生物细胞内形成交联，并且其中映射至不同叠连群的读取对指示哪些叠连群来自相同的物种。在一些实例中，所述固定剂为甲醛。

[0032] 在一些实施方案中，本公开内容提供了组装来源于DNA分子的多个叠连群的方法，

该方法包括由该DNA分子生成多个读取对以及使用所述读取对组装所述叠连群，其中至少1％的读取对跨越该DNA分子上大于50kB的距离并且所述读取对在14天内生成。在一些实施方案中，至少10％的读取对跨越该DNA分子上大于50kB的距离。在一些实施方案中，至少1％的读取对跨越该DNA分子上大于100kB的距离。在一些情况下，所述读取对在7天内生成。

[0033] 在其他实施方案中，本公开内容提供了组装来源于单个DNA分子的多个叠连群的方法，该方法包括在体外由该单个DNA分子生成多个读取对以及使用所述读取对组装所述叠连群，其中至少1％的读取对跨越该单个DNA分子上大于30kB的距离。在一些实施方案中，至少10％的读取对跨越该单个DNA分子上大于30kB的距离。在其他实施方案中，至少1％的读取对跨越该单个DNA分子上大于50kB的距离。

[0034] 在其他实施方案中，本公开内容提供了单元型定相的方法，该方法包括由单个DNA分子生成多个读取对以及使用所述读取对组装该DNA分子的多个叠连群，其中至少1％的读取对跨越该单个DNA分子上大于50kB的距离，并且以大于70％的准确度进行所述单元型定相。在一些实施方案中，至少10％的读取对跨越该单个DNA分子上大于50kB的距离。在其他实施方案中，其中至少1％的读取对跨越该单个DNA分子上大于100kB的距离。在一些实施方案中，以大于90％的准确度进行所述单元型定相。

[0035] 本公开内容提供了单元型定相的方法，该方法包括在体外由单个DNA分子生成多个读取对，以及使用所述读取对组装该DNA分子的多个叠连群，其中至少1％的读取对跨越该单个DNA分子上大于30kB的距离，并且以大于70％的准确度进行所述单元型定相。在一些实施方案中，至少10％的读取对跨越该单个DNA分子上大于30kB的距离。在其他实施方案中，至少1％的读取对跨越该单个DNA分子上大于50kB的距离。在其他实施方案中，以大于90％的准确度进行所述单元型定相。在一些实施方案中，以大于70％的准确度进行所述单元型定相。

[0036] 在一些实施方案中，本公开内容提供了由第一DNA分子生成第一读取对的方法，该方法包括：(a)在体外使所述第一DNA分子与多个缔合分子结合，其中所述第一DNA分子包含第一DNA区段和第二DNA区段；(b)标记所述第一DNA区段和所述第二DNA区段，从而形成至少一个标记的DNA区段；以及(c)对所述标记的DNA区段或所述标记的DNA区段的至少一个可识别部分，如与标记相邻的部分或在与标记的末端相对的末端的部分进行测序，从而获得标记的序列，其中所述多个缔合分子在步骤(a)和(b)之前和过程中不采用亲和标记物进行共价修饰。

[0037] 在某些实施方案中，本公开内容提供了由第一DNA分子生成标记的序列的方法，该方法包括：(a)在体外使所述第一DNA分子与多个缔合分子交联结合；(b)将所述第一DNA分子固定在固体支持体上；(c)切断所述第一DNA分子以生成第一DNA区段和第二DNA区段；(d)标记所述第一DNA区段和所述第二DNA区段，从而形成至少一个标记的DNA区段；以及对所述标记的DNA区段或所述标记的DNA区段的至少一个可识别部分，如与标记相邻的部分或在与标记的末端相对的末端的部分进行测序，或对所述标记的DNA区段的每一个末端的可识别部分进行测序，从而获得所述标记的序列，其中所述第一DNA分子直接结合至所述固体支持体。在一些实例中，所述固体支持体包括在没有用任何亲和标记物(例如，生物素、链霉亲和素、亲和素、多组氨酸、地高辛、EDTA或其衍生物)进一步修饰的情况下与DNA结合的聚合物珠子(例如，SPRI珠子)。

[0038] 在一些实施方案中，多个缔合分子，如来自重构的染色质的缔合分子，与所述第一DNA分子交联。在一些实例中，缔合分子包含氨基酸。在一些情况下，缔合分子是肽或蛋白质。在某些实例中，缔合分子是组蛋白。在一些情况下，组蛋白来自与所述第一DNA分子不同的来源。在各个实例中，缔合分子是转座酶。在一些情况下，第一DNA分子与缔合分子非共价结合。在其他情况下，第一DNA分子与缔合分子共价结合。在某些实例中，第一DNA分子与缔合分子交联。在某些实施方案中，第一DNA分子与固定剂交联。在一些实例中，所述固定剂为甲醛。在各个实施方案中，所述方法包括将所述多个缔合分子固定在固体支持体上。在一些情况下，固体支持体是珠子。在一些实例中，所述珠子包含聚合物。在一些实例中，所述聚合物是聚苯乙烯。在某些实例中，所述聚合物是聚乙二醇(PEG)。在某些实例中，所述珠子是磁珠。在一些实例中，所述珠子是固相可逆固定化(SPRI)珠子。在某些情况下，固体支持体包括表面，其中所述表面包含多个羧基基团。在各种情况下，固体支持体不与任何多肽(例如，链霉亲和素)共价连接。在一些情况下，缔合分子在固定至固体支持体之前不与亲和标记物(例如，生物素)共价连接。

[0039] 在一些实施方案中，通过切断所述第一DNA分子生成所述第一DNA区段和所述第二DNA区段。在一些情况下，在所述第一DNA分子与所述多个缔合分子结合之后切断所述第一DNA分子。在某些情况下，使用限制酶(例如，MboI)切断第一DNA分子。在一些情况下，使用转座酶(例如，Tn5)切断第一DNA分子。在其他情况下，使用物理方法(例如，声处理、机械剪切)切断第一DNA分子。在某些实施方案中，用亲和标记物修饰第一DNA和第二DNA区段。在一些实例中，亲和标记物可包括生物素，该生物素可用链霉亲和素珠子、亲和素珠子或其衍生物进行捕获。在某些实例中，亲和标记物是生物素修饰的核苷三磷酸(dNTP)。在一些实例中，亲和标记物是生物素修饰的脱氧核糖胞嘧啶三磷酸(dCTP)。在一些实例中，亲和标记物是生物素修饰的脱氧核糖胞嘧啶三磷酸(dGTP)。在一些实例中，亲和标记物是生物素修饰的脱氧核糖胞嘧啶三磷酸(dATP)。在一些实例中，亲和标记物是生物素修饰的脱氧核糖胞嘧啶三磷酸(dUTP)。在某些情况下，第一DNA区段用第一标记在至少第一端进行标记，而第二DNA区段用第二标记在至少第二端进行标记。在某些实例中，第一标记和第二标记相同。在各个实例中，使用转座酶(例如，Tn5)标记第一DNA区段和第二DNA区段。在一些情况下，用第二DNA区段标记第一DNA区段，并用第一DNA区段标记第二DNA区段。例如，可将第一DNA区段与第二DNA区段连接。在一些实例中，使用连接酶将第一DNA区段与第二DNA区段连接。在一些情况下，在步骤(c)中测序之前切断连接的DNA区段。在某些实例中，使用限制酶(例如，ExoIII)切断连接的DNA区段。在其他情况下，使用物理方法(例如，声处理、机械剪切)切断连接的DNA区段。

[0040] 在一些实施方案中，在第一DNA区段与第二DNA区段连接之前洗涤第一DNA区段少于约10次。在一些实施方案中，在第一DNA区段与第二DNA区段连接之前洗涤第一DNA区段少于约6次。在一些实施方案中，所述方法包括将所述连接的DNA区段与测序衔接子连接。

[0041] 在某些实施方案中，所述方法包括使用所述标记的序列组装多个叠连群。在一些实施方案中，将第一和第二DNA区段中的每一个均与至少一个亲和标记物连接，并使用亲和标记物捕获连接的DNA区段。在各个实施方案中，所述方法包括使用所述标记的序列对第一DNA区段和第二DNA区段进行定相。在一些情况下，通过连接第一DNA区段与第二DNA区段来完成‘标记’，从而生成读取对区段。

[0042] 在一些实施方案中，所述方法包括：(a)向至少第二DNA分子提供多个缔合分子(如来自重构的染色质)；(b)使所述缔合分子与所述第二DNA分子交联，从而在体外形成第二复合体；(c)切断所述第二复合体，从而生成第三DNA区段和第四区段；(d)连接所述第三DNA区段与第四DNA区段，从而形成第二连接的DNA区段；以及(e)对所述第二连接的DNA区段进行测序，从而获得第二读取对。在一些实例中，将来自DNA分子的少于40％的DNA区段与来自任何其他DNA分子的DNA区段连接。在一些实例中，将来自DNA分子的少于20％的DNA区段与来自任何其他DNA分子的DNA区段连接。

[0043] 在一些实施方案中，本公开内容提供了由包含预定序列的第一DNA分子生成第一读取对的方法，该方法包括：(a)向所述第一DNA分子提供一个或多个DNA结合分子，其中所述一个或多个DNA结合分子与所述预定序列结合；(b)使所述第一DNA分子在体外交联，其中所述第一DNA分子包含第一DNA区段和第二DNA区段；(c)连接所述第一DNA区段与所述第二DNA区段，从而形成第一连接的DNA区段；以及(d)对所述第一连接的DNA区段进行测序，从而获得所述第一读取对；其中所述预定序列在所述读取对中出现的概率受DNA结合分子与预定序列的结合的影响。

[0044] 在一些实施方案中，所述DNA结合分子是可与预定序列杂交的核酸。在一些实例中，该核酸是RNA。在其他实例中，该核酸是DNA。在其他实施方案中，所述DNA结合分子是小分子。在一些实例中，该小分子以小于100μM的结合亲和力与预定序列结合。在一些实例中，该小分子以小于1μM的结合亲和力与预定序列结合。在一些实施方案中，将DNA结合分子固定在表面或固体支持体上。

[0045] 在一些实施方案中，预定序列在读取对中出现的概率减小。在其他实施方案中，预定序列在读取对中出现的概率增加。

[0046] 本公开内容提供了用于由多个DNA分子生成多个标记的序列的方法，该方法包括：(a)在体外使所述多个DNA分子与多个缔合分子结合；(b)切断每一个DNA分子以生成至少多个DNA区段；(c)标记所述DNA区段的至少一部分以形成多个标记的DNA区段；以及(d)对所述标记的DNA区段或所述标记的DNA区段的至少一个可识别部分，如与标记相邻的部分或在与标记的末端相对的末端的部分进行测序，以获得多个标记的序列；其中所述多个缔合分子在步骤(a)和(b)之前和过程中不采用亲和标记物进行共价修饰。在一些情况下，将来自DNA分子的少于40％的DNA区段与来自任何其他DNA分子的DNA区段连接。在一些情况下，将来自DNA分子的少于20％的DNA区段与来自任何其他DNA分子的DNA区段连接。

[0047] 在一些实施方案中，所述缔合分子包含通过肽键连接的氨基酸。在某些实施方案中，所述缔合分子是多肽或蛋白质。在一些实例中，所述缔合分子是组蛋白。在一些实例中，该组蛋白来自与DNA分子不同的来源。例如，该组蛋白可从非人类生物体分离，而DNA分子可从人类分离。在多个实例中，所述缔合分子是转座酶(例如，Tn5)。在一些情况下，第一DNA分子与缔合分子非共价结合。在其他情况下，第一DNA分子与缔合分子共价结合。在某些实例中，第一DNA分子与缔合分子交联。在一些实例中，该DNA分子与固定剂交联。例如，该固定剂可以是甲醛。在一些情况下，所述方法包括将所述多个缔合分子固定在多个固体支持体上。在某些情况下，该固体支持体是珠子。在一些实例中，该珠子包含聚合物。在一些实例中，所述聚合物是聚苯乙烯。在某些实例中，所述聚合物是聚乙二醇(PEG)。在某些实例中，所述珠子是磁珠。在一些实例中，所述珠子是SPRI珠子。在多个实例中，所述固体支持体包括表面，

其中所述表面包含多个羧基基团。在多种情况下，所述固体支持体不与任何多肽(例如，链霉亲和素)共价连接。在一些情况下，所述缔合分子在固定至固体支持体之前不与亲和标记物(例如，生物素)共价连接。

[0048] 在一些实施方案中，在所述第一DNA分子与所述多个缔合分子结合之后切断所述第一DNA分子。在一些情况下，使用限制酶(例如，MboII)切断第一DNA分子。在某些情况下，使用转座酶(例如，Tn5)切断第一DNA分子。在某些实施方案中，用亲和标记物修饰DNA区段的部分。在一些情况下，该亲和标记物包括生物素。在一些实例中，该亲和标记物是生物素修饰的核苷三磷酸(dNTP)。在一些实例中，该生物素修饰的核苷三磷酸(dNTP)是生物素修饰的脱氧核糖胞嘧啶三磷酸(dCTP)。在一些情况下，DNA区段的一部分用第一标记至少在第一端进行标记。在一些实例中，使用转座酶标记DNA区段。在各种情况下，通过连接所述DNA区段中的每一个与至少一个其他DNA区段来标记DNA区段的一部分。在一些实例中，使用连接酶连接DNA区段的所述部分与其他DNA区段。在某些情况下，在步骤(c)之前切断连接的DNA区段。在多种情况下，使用物理方法(例如，声处理、机械剪切)切断连接的DNA区段。在一些实施方案中，所述方法包括将所述连接的DNA区段与测序衔接子连接。

[0049] 在一些情况下，在连接DNA区段以形成连接的DNA区段之前洗涤DNA区段少于约10次。在某些情况下，在连接DNA区段以形成连接的DNA区段之前洗涤DNA区段少于约6次。在各种情况下，所述方法包括使用标记的区段组装DNA分子的多个叠连群。在一些情况下，所述方法包括使用标记的区段对DNA区段进行定相。

[0050] 本公开内容提供了包含多个读取对的体外文库，所述多个读取对各自包含至少第一序列元件和第二序列元件，其中所述第一和第二序列元件来源于单个DNA分子，并且其中至少1％的读取对包含在单个DNA分子上相隔至少50kB的第一和第二序列元件。在一些实施方案中，至少10％的读取对包含在单个DNA分子上相隔至少50kB的第一和第二序列元件。在其他实施方案中，至少1％的读取对包含在单个DNA分子上相隔至少100kB的第一和第二序列元件。在一些实施方案中，少于20％的读取对包含一个或多个预定序列。在一些实施方案中，少于10％的读取对包含一个或多个预定序列。在一些实施方案中，少于5％的读取对包含一个或多个预定序列。

[0051] 在一些实施方案中，通过可与预定序列杂交的一个或多个核酸来确定预定序列。在一些实例中，所述一个或多个核酸是RNA。在其他实例中，所述一个或多个核酸是DNA。在一些实例中，将所述一个或多个核酸固定至表面或固体支持体。

[0052] 在一些实施方案中，通过一个或多个小分子确定预定序列。在一些实例中，所述一个或多个小分子以小于100μM的结合亲和力与预定序列结合。在一些实例中，所述一个或多个小分子以小于1μM的结合亲和力与预定序列结合。

[0053] 本公开内容提供了包含DNA片段和多个缔合分子(如来自重构的染色质)的组合物，其中：(a)使所述缔合分子与所述DNA片段以体外复合体的形式交联；以及(b)将该体外复合体固定在固体支持体上。

[0054] 本公开内容提供了包含DNA片段、多个缔合分子和DNA结合分子的组合物，其中：(a)使所述DNA结合分子与所述DNA片段的预定序列结合；以及(b)使所述缔合分子与所述DNA片段交联。在一些情况下，DNA结合分子是可与预定序列杂交的核酸。在一些实例中，该核酸是RNA。在其他实例中，该核酸是DNA。在一些实例中，将所述核酸固定至表面或固体支

持体。在其他实施方案中，DNA结合分子是小分子。在一些实例中，小分子以小于100μM的结合亲和力与预定序列结合。在其他实例中，小分子以小于1μM的结合亲和力与预定序列结合。

[0055] 本公开内容提供了包含与DNA片段以体外复合体形式结合的多个缔合分子的组合物，其中所述体外复合体固定在固体支持体上，并且其中所述固体支持体不与任何多肽共价连接。在一些情况下，固体支持体不与链霉亲和素共价连接。在一些情况下，固体支持体是珠子。在一些实例中，该珠子包含聚合物。在一些实例中，该聚合物是聚苯乙烯。在某些实例中，该聚合物是聚乙二醇(PEG)。在某些实例中，该珠子是磁珠。在一些实例中，该珠子是固相可逆固定化(SPRI)珠子。在某些情况下，固体支持体包括表面，其中所述表面包含多个羧基基团。在多种情况下，固体支持体不与任何多肽(例如，链霉亲和素)共价连接。

[0056] 在一些实例中，缔合分子包含通过肽键结合的氨基酸。在一些实例中，缔合分子是肽或蛋白质。在某些实例中，缔合分子是组蛋白。在一些情况下，该组蛋白来自与所述第一DNA分子不同的来源。在某些实例中，缔合分子是转座酶。在一些情况下，第一DNA分子与缔合分子非共价结合。在其他情况下，第一DNA分子与缔合分子非共价结合。在一些实例中，第一DNA分子与缔合分子交联。在某些实施方案中，第一DNA分子与固定剂交联。在一些实例中，所述固定剂为甲醛。

[0057] 在某些实施方案中，用亲和标记物修饰DNA片段。在一些实例中，该亲和标记物可包括生物素，该生物素可用链霉亲和素珠子、亲和素珠子或其衍生物进行捕获。在某些实例中，该亲和标记物是生物素修饰的核苷三磷酸(dNTP)。在一些实例中，该亲和标记物是生物素修饰的脱氧核糖胞嘧啶三磷酸(dCTP)。在一些情况下，在步骤(c)中测序之前进一步切断连接的DNA区段。在某些实例中，使用限制酶(例如，ExoIII)切断连接的DNA区段。在其他情况下，使用物理方法(例如，声处理、机械剪切)切断连接的DNA区段。

[0058] 本文公开的方法和组合物可用于将基因组信息组装成支架，直到并包括定相的全染色体。在一些情况下，本文生成的信息指导先前生成的序列信息组装成支架，直到并包括定相的全染色体。在一些情况下，本文的方法和组合物用于将从头生成的核酸信息组装成定相的支架，直到并包括全染色体。

[0059] 标记信息并非在所有情况下都与相位严格对应，而是提供关于相位信息的信息。通常提及本文的公开内容，在序列读取对上共同标记模式的存在指示所述读取1)来源于共同的分子，或2)偶然被共用。

[0060] 在大多数情况下，共同标记将不会偶然出现，并且因此可靠地推断大多数共同标记的序列，特别是独立地映射至共同叠连群的共同标记的序列，映射至所述叠连群的共同相位，即映射至二倍体生物体的相同单倍体分子。一起映射至单个或几个疑似相邻的叠连群以及共有标记序列的读取组很有可能在单个分子上同相位。例如，共有共同的标记序列但映射至疑似在单独染色体上的叠连群的读取组更有可能是偶然地获得了其共同的标记序列。然而，共有所述标记序列但映射至两个单独的叠连群或疑似染色体的序列集群的多个实例可能指示已发生易位，通过所述易位一个染色体的片段已附接至第二个染色体，使得所述读取由于易位而实际上在所述染色体上同相位。

[0061] 序列读取对中不相同的标记模式的存在指示这些序列并不是在临标记之前由共同的分子产生。然而，如果在单个样品中存在核酸分子的多个相同或重叠拷贝，则两组序列

读取可能在其标记模式中出现所述差异，这指示它们由所述样品中不同的分子产生，但尽管如此，仍映射至二倍体细胞中相同的同相位染色体。也就是说，标记模式信息指示序列是否由共同的分子产生，并且通常，标记模式信息与相位信息有关。然而，如上所讨论的，在不一致时，标记模式信息更恰当地指示起源的共同分子。在起源的分子和核酸相位确定显示出一些不一致的情况下，本领域技术人员能够解决这些不一致，使得一些相位信息仍可由通过本文的方法生成的标记模式信息来确定。

[0062] 本文公开了由第一DNA分子生成标记的序列的方法，该方法包括：(a)使所述第一DNA分子与多个缔合分子结合，以形成第一复合体，其中所述第一DNA分子包含第一DNA区段和第二DNA区段；(b)标记所述第一DNA区段和所述第二DNA区段，从而形成至少一个标记的DNA区段；(c)使所述复合体结合至具有直接结合所述复合体的组分的表面的固体支持体；以及(d)对所述标记的DNA区段的可识别部分，如与标记相邻的部分或在与标记的末端相对的末端的部分进行测序，从而获得所述标记的序列；其中所述多个缔合分子在步骤(a)和(b)之前或过程中不采用亲和标记物进行共价修饰。

[0063] 本文公开了由第一DNA分子生成标记的序列的方法，该方法包括：(a)使所述第一DNA分子与多个缔合分子结合；(b)将所述第一DNA分子固定在固体支持体上；(c)切断所述第一DNA分子以生成第一DNA区段和第二DNA区段；(d)标记所述第一DNA区段和所述第二DNA区段，从而形成至少一个标记的DNA区段；以及(e)对所述标记的DNA区段进行测序，从而获得所述标记的序列；其中所述第一DNA分子直接结合至所述固体支持体。

[0064] 本文公开了用于由多个DNA分子生成多个标记的序列的方法，该方法包括：(a)使所述多个DNA分子与多个缔合分子结合；(b)切断所述多个DNA分子以生成多个DNA区段；(c)标记所述DNA区段的至少一部分以形成多个标记的DNA区段；以及(d)对所述标记的DNA区段进行测序，以获得多个标记的序列；其中所述多个缔合分子在步骤(a)和(b)之前或过程中不采用亲和标记物进行共价修饰。

[0065] 本文公开了包含与DNA片段以体外复合体形式结合的多个缔合分子的组合物，其中所述体外复合体固定在固体支持体上，并且其中所述固体支持体不与任何多肽共价连接。

[0066] 本文公开了用于由多个DNA分子生成多个标记的序列的方法，该方法包括：(a)获得与多个缔合分子结合的多个DNA分子；(b)切断所述DNA分子以生成至少多个DNA区段；(c)标记所述DNA区段的至少一部分以形成多个标记的DNA区段；以及(d)对所述标记的DNA区段进行测序，以获得多个标记的序列；其中所述多个DNA分子的总量少于约5微克(μg)。

[0067] 本文公开了鉴定抗生素抗性基因的微生物宿主的方法，该方法包括：a)从患有显示微生物抗生素抗性的病况的个体获得稳定化的样品；b)处理所述稳定化的样品以裂解所述稳定化的样品中的双链DNA；c)标记暴露的DNA末端；d)连接标记的暴露的DNA末端以形成标记的成对末端；以及e)在标记的成对末端进行测序以生成成对序列；其中与抗生素抗性基因序列相邻的序列指示抗生素抗性基因的微生物宿主。

[0068] 本文公开了确定异质核酸样品的基因组连接信息的方法，该方法包括：(a)获得稳定化的异质核酸样品；(b)处理所述稳定化的样品以裂解所述稳定化的样品中的双链DNA；(c)标记暴露的DNA末端；(d)连接标记的暴露的DNA末端以形成标记的成对末端；(e)在标记的成对末端进行测序以生成多个成对序列读取；(f)将所述多个序列读取的成对序列读取

的每一半分配给共同的起源核酸分子。

[0069] 本文公开了用于宏基因组组装的方法，该方法包括：(a)从环境中收集微生物；(b)从所述微生物获得多个叠连群；(c)由通过探测重构的染色质的物理布局产生的数据生成多个读取对；以及(d)将所述多个读取对映射至所述多个叠连群，从而产生读取映射数据，其中映射至不同叠连群的读取对指示不同的叠连群来自共同的物种。

[0070] 本文公开了在宿主群体中检测病原体的方法，该方法包括：a)从疑似具有共同病原体的多个个体中的每一个获得稳定化的样品；b)处理所述稳定化的样品以裂解所述稳定化的样品中的双链DNA；c)使用第一条码标记来标记所述稳定化的样品的第一部分的暴露的DNA末端，并使用第二条码标记来标记所述稳定化的样品的第二部分的暴露的末端；d)在条码标记的末端进行测序以生成多个条码标记的序列读取；以及e)将所述多个序列读取的共同条码标记的序列读取分配给共同的起源生物体；其中对于疑似具有共同病原体的个体而言共同的起源生物体是所述病原体。

[0071] 本文公开了鉴定抗生素抗性基因的微生物宿主的方法，该方法包括：a)从患有显示微生物抗生素抗性的病况的个体获得稳定化的样品；b)处理所述稳定化的样品以裂解所述稳定化的样品中的双链DNA；c)使用第一条码标记来标记所述稳定化的样品的第一部分的暴露的DNA末端，并使用第二条码标记来标记所述稳定化的样品的第二部分的暴露的末端；d)在条码标记的末端进行测序以生成多个条码标记的序列读取；其中具有与抗生素抗性基因序列的条码标记相同的条码标记的序列指示抗生素抗性基因的微生物宿主。

[0072] 本文公开了确定异质核酸样品的基因组连接信息的方法，该方法包括：(a)获得稳定化的异质核酸样品；(b)处理所述稳定化的样品以裂解所述稳定化的样品中的双链DNA；(c)使用第一条码标记来标记所述稳定化的样品的第一部分的暴露的DNA末端，并使用第二条码标记来标记所述稳定化的样品的第二部分的暴露的末端；(d)在条码标记的末端进行测序以生成多个条码标记的序列读取；(e)将共同标记的序列读取分配给共同的起源核酸分子。

[0073] 本文公开了在宿主群体中检测病原体的方法，该方法包括：a)从多个受试者中的每一个获得稳定化的样品；b)处理所述稳定化的样品以裂解所述稳定化的样品中的双链DNA，从而生成暴露的DNA末端；c)标记所述暴露的DNA末端的至少一部分；d)连接所述暴露的DNA末端以形成标记的成对末端；e)对所述标记的成对末端的至少一个可识别部分进行测序以生成多个读取对；以及f)将读取对的每一半分配给共同的起源生物体；其中对于所述受试者而言共同的起源生物体被检测为所述病原体。

[0074] 本文公开了鉴定抗生素抗性基因的微生物宿主的方法，该方法包括：a)从患有显示微生物抗生素抗性的病况的受试者获得稳定化的样品；b)处理所述稳定化的样品以裂解所述稳定化的样品中的双链DNA，从而生成暴露的DNA末端；c)标记所述暴露的DNA末端的至少一部分；d)连接所述标记的暴露的DNA末端以形成标记的成对末端；以及e)对所述连接的成对末端的至少一个可识别部分进行测序以生成成对序列；其中与抗生素抗性基因序列相邻的成对序列指示抗生素抗性基因的微生物宿主。

[0075] 本文公开了确定异质核酸样品的基因组连接信息的方法，该方法包括：(a)使所述异质核酸样品稳定化；(b)处理所述稳定化的样品以裂解所述稳定化的样品中的双链DNA，从而生成暴露的DNA末端；(c)标记所述暴露的DNA末端的至少一部分；(d)连接所述标记的

暴露的DNA末端以形成标记的成对末端；(e)对所述标记的成对末端的至少一个可识别部分进行测序以生成多个读取对；(f)将读取对的每一半分配给共同的起源核酸分子。

[0076] 本文公开了用于宏基因组组装的方法，该方法包括：(a)从环境中收集微生物；(b)从所述微生物获得多个叠连群；(c)由通过探测重构的染色质的物理布局产生的数据生成多个读取对；以及(d)将所述多个读取对映射至所述多个叠连群，从而产生读取映射数据，其中映射至不同叠连群的读取对指示不同的叠连群来源于共同的个体。

[0077] 本文公开了用于检测细菌致病原的方法，该方法包括：(a)从所述细菌致病原获得多个叠连群；(b)由通过探测重构的染色质的物理布局产生的数据生成多个读取对；(c)将所述多个读取对映射至所述多个叠连群，从而产生读取映射数据；(d)使用所述读取映射数据排列所述叠连群以将所述叠连群组装成基因组组装；以及(e)使用所述基因组组装来确定所述细菌致病原的存在。

[0078] 本文公开了从生物体获得基因组序列信息的方法，该方法包括：(a)从所述生物体获得稳定化的样品；(b)处理所述稳定化的样品以裂解所述稳定化的样品中的双链DNA，从而生成暴露的DNA末端；(c)标记所述暴露的DNA末端的至少一部分，以生成标记的DNA区段；(d)对所述标记的DNA区段的至少一个可识别部分进行测序，从而获得标记的序列；以及(e)映射所述标记的序列以生成所述生物体的基因组序列信息，其中所述基因组序列信息覆盖所述生物体的基因组的至少75％。

[0079] 本文公开了分析样品的方法，该方法包括：(a)从多个生物体获得包含核酸的稳定化的样品；(b)处理所述稳定化的样品以裂解所述稳定化的样品中的双链DNA，从而产生暴露的DNA末端；(c)连接所述暴露的DNA末端以形成成对末端；(d)在所述成对末端进行测序以生成多个成对序列读取；以及(e)将所述多个序列读取的成对序列读取的每一半分配给共同的起源生物体。

[0080] 本文公开了测定异质样品的核酸分子多样性的方法，该方法包括：a)获得包含多样化的多个核酸的稳定化的核酸样品，所述多样化的多个核酸被稳定化，使得对于所述多个核酸中的至少一个成员，第一核酸区段和第二核酸区段独立于它们共同的磷酸二酯骨架而被保持在一起，其中在所述第一核酸区段与所述第二核酸区段之间裂解所述磷酸二酯骨架；b)标记所述第一核酸区段和所述第二核酸区段，使得在由所述多样化的多个核酸的共同核酸产生时，所述第一核酸区段和所述第二核酸区段是可识别的；c)对所述第一核酸区段的至少一个可识别部分及其标记以及所述第二核酸区段的可识别部分及其标记进行测序；d)将所述第一核酸区段和所述第二核酸区段分配给与所述标记对应的支架；e)使得所述多样化的多个核酸的多个区段被分配给至少一个支架；以及f)确定与生成多少支架对应的数目；其中生成的支架的数目与所述异质样品的核酸分子多样性对应。在一些方面，标记所述第一核酸区段和所述第二核酸区段包括将第一寡核苷酸添加至所述第一核酸区段以及将第二寡核苷酸添加至所述第二区段，所述第一寡核苷酸和所述第二寡核苷酸共有共同的序列。在一些方面，将具有所述共同的寡核苷酸序列的核酸区段分配给共同的支架。在一些方面，所述方法进一步包括将所述第一核酸区段的所述可识别部分映射至叠连群数据集，并且包括将所述叠连群数据集的叠连群任意匹配至所述共同的支架。在一些方面，同时生成所述叠连群数据集。在一些方面，从数据库获得所述叠连群数据集。在一些方面，标记所述第一核酸区段和所述第二核酸区段包括连接所述第一核酸区段与所述第二核酸区段，

并且其中将所述第一核酸区段和所述第二核酸区段分配给共同的支架。在一些方面，所述方法进一步包括将所述第一核酸区段的所述可识别部分映射至叠连群数据集，并且将所述叠连群数据集的任何匹配的叠连群纳入所述共同的支架。在一些方面，同时生成所述叠连群数据集。在一些方面，从数据库获得所述叠连群数据集。在一些方面，所述异质样品包含多个等位基因变体。在一些方面，等位基因变体的数目大于支架的数目。在一些方面，等位基因变体的数目等于生成的支架的数目。在一些方面，在所述获得稳定化的样品之后，裂解所述磷酸二酯骨架。在一些方面，使所述稳定化的样品与交联剂接触。在一些方面，所述稳定化的样品为FFPE样品。在一些方面，所述方法进一步包括使所述异质样品与逆转录酶接触。在一些方面，所述方法进一步包括在核酸序列数据库中搜索至少一个所述支架。在一些方面，所述方法进一步包括如果唯一映射至所述支架的核酸序列在所述数据库中不存在，则将所述支架分类为新型的。在一些方面，所述方法进一步包括当与样品条件相关的多个样品具有所述支架时并且在缺乏所述条件的多个样品缺乏所述样品的情况下，将所述支架分类为与样品条件对应。在一些方面，所述异质样品包含映射至共同物种的至少两个个体的核酸。在一些方面，所述异质样品包含映射至共同物种的至少三个个体的核酸。在一些方面，所述异质样品包含映射至至少两个物种的核酸。在一些方面，所述异质样品包含映射至至少三个物种的核酸。在一些方面，所述异质样品包含映射至至少四个物种的核酸。在一些方面，在不参考外源序列信息的情况下，序列读取组装成至少两个核酸支架。在一些方面，在不参考外源序列信息的情况下，序列读取组装成至少三个核酸支架。在一些方面，序列读取组装成至少两个核酸支架，使得至少50％的第一基因组和至少50％的第二基因组被表示在所述至少两个核酸支架中。在一些方面，序列读取组装成至少两个核酸支架，使得至少60％的第一基因组和至少60％的第二基因组被表示在所述至少两个核酸支架中。在一些方面，序列读取组装成至少两个核酸支架，使得至少70％的第一基因组和至少70％的第二基因组被表示在所述至少两个核酸支架中。在一些方面，序列读取组装成至少两个核酸支架，使得至少80％的第一基因组和至少80％的第二基因组被表示在所述至少两个核酸支架中。在一些方面，所述方法包括使用SPRI珠子。在一些方面，所述稳定化的样品包含不超过约5微克的DNA。

[0081] 援引并入

[0082] 本说明书中提到的所有出版物、专利和专利申请均通过引用并入本文，其程度如同特别地且单独地指出每一个单独的出版物、专利或专利申请均通过引用而并入。本说明书中提及的所有出版物、专利和专利申请均通过引用以其全文以及其中引用的任何参考文献并入本文。

附图说明

[0083] 本发明的新颖特征在本文所附的权利要求书中具体阐述。通过参考以下对利用本发明原理的说明性实施方案加以阐述的详细描述以及附图，将获得对本发明的特征和优点的更好理解，在这些附图中：

[0084] 图1呈现了使用高通量测序读取的基因组组装的图示。示出了待组装的基因组(顶部)。通常，基因组具有许多难以组装的重复序列。收集来自基因组的随机、高通量序列数据(中间)并将该数据组装成在所述基因组中独特的区域形式的“叠连群”(底部)。叠连群组装

通常在许多重复序列处停止。最终的输出是相对于彼此的顺序和方向未知的一组数以千计的叠连群。在该图中，将它们从最长至最短随意地进行编号。

[0085] 图2A-图2D示出了本公开内容的基于染色质捕获的方案：(A)显示了交联并处理DNA以产生生物素化的接头片段以进行测序；以及(B-D)提供了关于人chr14对于多种限制酶的接触映射数据。如图所示，大多数接触沿着染色体是局部的。

[0086] 图3A-图3C提供了本公开内容的使用染色质捕获序列数据来帮助基因组组装的方法：(A)示出了使用基于染色质捕获的方案交联并处理DNA；(B)显示了将读取对数据映射至由随机鸟枪法测序和组装生成的组装的叠连群；以及(C)示出了在过滤并加权之后，可构建概括所有叠连群间读取对数据的邻接矩阵。该矩阵可重新排序以指示正确的组装路径。如图所示，大多数读取对将在叠连群内映射。从中可以了解接触距离的分布(例如，参见图6)。映射至不同叠连群的读取对提供了关于在正确的基因组组装中哪些叠连群相邻的数据。

[0087] 图4示出了本公开内容的示例性方案：首先生成并制备DNA片段；然后进行体外染色质组装；随后将染色质/DNA复合体用甲醛固定并用SPRI珠子拉下；随后将所述复合体进行限制消化以生成粘性末端，然后在该粘性末端中装入生物素化dCTP并在内部装入硫酸化GTP；在钝性末端连接之后，使染色质/DNA复合体经历蛋白酶消化和剪切；之后将所述DNA片段用SPRI珠子拉下并用测序衔接子连接；最后，根据大小选择所述DNA片段并对该DNA片段进行测序。

[0088] 图5A-5B提供了基因组组装以及与基因组中的重复区域的比对中出现的分歧的图示。(A)由于读取对不能桥接重复区域引起连接不确定。(B)由于读取对不能跨越毗连重复引起区段的位置不确定。

[0089] 图6示出了来自人类XLRP文库的读取对之间的基因组距离的分布。显示了采用其他技术可得到的最大距离以供比较。

[0090] 图7示出了具有表征良好的单元型的样品NA12878的定相准确度。显示的距离是正在定相的SNP之间的距离。

[0091] 图8示出了根据本公开内容的各个实施方案的示例性计算机系统的各个构件。

[0092] 图9为示出可结合本公开内容的各个实施方案使用的示例性计算机系统的结构的框图。

[0093] 图10为示出可结合本公开内容的各个实施方案使用的示例性计算机网络的图。

[0094] 图11为示出可结合本公开内容的各个实施方案使用的另一个示例性计算机系统的结构的框图。

[0095] 图12A示出了用于邻近连接的程序的示例性示意图。

[0096] 图12B示出了用于宏基因组分析的样品制备的两种路线的示例性示意图。

[0097] 图12C示出了支架化技术的示例性示意图。

[0098] 图13A示出了根据本公开内容的方面，来自粪便DNA样品的DNA片段的大小分析。

[0099] 图13B示出了使用体外组装的染色质聚集体生成测序文库的方法。

[0100] 图14示出了根据本公开内容的方面，鸟枪法文库的插入片段大小分布。

[0101] 图15示出了来自使用映射至相同支架的体外组装染色质制备的文库的读取的大小分布。

[0102] 图16示出了来自针对鸟枪法测序制备的文库和来自使用体外组装染色质聚集体

制备的文库的命中的散点图。

[0103] 图17示出了按照叠连群长度的每个叠连群的鸟枪法命中/体外组装染色质命中的散点图。

[0104] 图18示出了TapeStation迹线，其显示粪便DNA制剂(蓝色，在x轴的100bp和15000bp处迅速上升接近y轴顶部)和天蓝色链霉菌(Streptomyces coelicolor)DNA(绿色，在15000bp处迅速上升至样品强度100)中的片段大小分布具有相似的长度。

[0105] 图19示出了对于每个水平的掺入(spiked-in)天蓝色链霉菌DNA，这些鸟枪法数据的覆盖倍数分布。

[0106] 图20示出了对于1％(红色，左侧)、5％(绿色，中间)和10％(蓝色，右侧)鸟枪法数据集，作为叠连群存在的天蓝色链霉菌基因组的总量。

[0107] 图21示出了映射至天蓝色链霉菌的已知基因组序列的来自邻近连接文库的读取对；x轴表示以千碱基为单位所跨越的距离，且y轴为在所有读取对中的累积分布。

[0108] 图22A描绘了已知的天蓝色链霉菌基因组(x轴)相对于在5％实验中如本文所述生成的三个支架的点阵图。

[0109] 图22B描绘了已知的天蓝色链霉菌基因组(x轴)相对于在10％实验中如本文所述生成的一个支架的点阵图。

[0110] 图23A描绘了来自粪便DNA制备试剂盒的DNA片段大小的图。

[0111] 图23B描绘了读取对的数目相对于所跨越的读取对距离的图。

[0112] 图24描绘了包含89％的8.67Mb天蓝色链霉菌基因组的单个支架。

[0113] 图25描绘了在掺入实验中，Chicago组装数据与鸟枪法数据中读取覆盖率之比的示例图。

[0114] 图26A描绘了掺入实验中针对支架的覆盖深度和GC含量的图。

[0115] 图26B描绘了对于每个支架的体外染色质组装连接性(作为所有连接相对于其第1至第4个最多连接的支架的分数)与支架对之间的GC+倍数覆盖空间的Euclidean距离的图。

[0116] 图27描绘了菌株变异对支架化性能的影响的图。

具体实施方式

[0117] 除非上下文另有明确说明，否则如本文和随附权利要求中所用的，单数形式“一个”、“一种”和“该”包括复数指代物。因此，例如，提及“叠连群”包括多个这样的叠连群，并且提及“探测染色体的物理布局”包括提及用于探测染色体的物理布局以及本领域技术人员已知的其等同项的一种或多种方法，等等。

[0118] 同样，使用“和”意指“和/或”，除非另有规定。类似地，“包含”、“包括”和“含有”可互换，并且并非意在限制。

[0119] 应进一步理解的是，在各个实施方案的描述中使用术语“包含”时，本领域技术人员将会理解在一些特定的情况下，实施方案可使用语言“基本上由...组成”或“由...组成”替代性地描述。

[0120] 除非另有规定，否则如本文用于描述数字的术语“约”是指包括该数字加上或减去该数字的10％的数值范围。

[0121] 如本文所用的术语“读取”、“序列读取”或“测序读取”是指在测序反应的单个反应

或运行中测定的DNA或RNA核酸的片段或区段的序列。

[0122] 如本文所用的术语“叠连群”是指DNA序列的连续区域。“叠连群”可通过本领域已知的任何数字方法来确定，例如通过比较重叠序列的测序读取和/或通过将测序读取与已知序列的数据库进行比较来识别哪些测序读取具有较高的连续概率。

[0123] 术语“多核苷酸”、“核苷酸”、“核酸”和“寡核苷酸”通常可互换使用。它们一般是指任意长度的核苷酸(脱氧核糖核苷酸或核糖核苷酸)的聚合形式，或其类似物。多核苷酸包含通过磷酸二酯键连接在其核糖骨架上的碱基单体。多核苷酸可具有任何三维结构，并且可执行已知或未知的任何功能。以下是多核苷酸的非限制性实例：基因或基因片段的编码或非编码区、基因间DNA、由连锁分析限定的多个基因座(单个基因座)、外显子、内含子、信使RNA(mRNA)、转移RNA、核糖体RNA、短干扰RNA(siRNA)、短发夹RNA(shRNA)、微RNA(miRNA)、小核仁RNA、核酶、互补DNA(cDNA)(其为mRNA的DNA表示，通常通过信使RNA(mRNA)的逆转录或通过扩增来获得)；通过合成或通过扩增产生的DNA分子、基因组DNA、重组多核苷酸、支链多核苷酸、质粒、载体、任何序列的分离的DNA、任何序列的分离的RNA、核酸探针和引物。多核苷酸可包含经修饰的核苷酸，如甲基化核苷酸和核苷酸类似物。如果存在的话，对核苷酸结构的修饰可在聚合物的组装之前或之后给予。通常，寡核苷酸仅包含几个碱基，而多核苷酸可包含任意数目但通常更长的碱基，而核酸可指任意长度的聚合物，直到并包括染色体或整个基因组的长度。另外，术语核酸经常被统一使用，使得核酸样品并不一定指单个核酸分子；而是可以指包含多个核酸分子的样品。术语核酸可涵盖双链或三链核酸以及单链分子。在双链或三链核酸中，核酸链不必是共同延伸的，例如，双链核酸不必沿着两条链的整个长度都是双链。术语核酸可涵盖它的任何化学修饰，如通过甲基化和/或通过加帽。核酸修饰可包括化学基团的添加，所述化学基团将附加电荷、极化性、氢键键合、静电作用和功能性并入单独核酸碱基或作为整体的核酸中。这类修饰可包括碱基修饰如2’-位置糖修饰、5-位置嘧啶修饰、8-位置嘌呤修饰、在胞嘧啶环外胺处的修饰、5-溴尿嘧啶的置换、骨架修饰、不寻常的碱基配对组合如异碱基异胞苷和异胍等。

[0124] 如本文所用的术语“受试者”可指任何真核或原核生物体。

[0125] 如本文所用的术语“裸DNA”可指基本不含复合DNA结合蛋白质的DNA。例如，它可指与少于约10％、约5％或约1％的在细胞核中发现的内源性蛋白，或少于约10％、约5％或约1％的通常在体内与核酸结合的内源性DNA结合蛋白质，或少于约10％、约5％或约1％的外源添加的核酸结合蛋白质或其他核酸结合部分如纳米颗粒复合的DNA。在一些情况下，裸DNA是指不与DNA结合蛋白质复合的DNA。

[0126] 术语“多肽”和“蛋白质”通常可互换使用并且一般是指氨基酸的聚合形式或通过多肽键结合的其类似物。多肽和蛋白质可以是任意长度的聚合物。多肽和蛋白质可具有任何三维结构，并且可执行已知或未知的任何功能。多肽和蛋白质可包含修饰，包括磷酸化、脂化、异戊烯化、硫酸化、羟基化、乙酰化、二硫键的形成等。在一些情况下，“蛋白质”是指具有已知功能或已知天然存在于生物系统中的多肽，但本领域中不总是遵守这一区别。

[0127] 如本文所用的，如果核酸通过一个结合部分或多个结合部分结合，使得核酸的单独区段独立于其共同的磷酸二酯骨架而保持在单个复合体中，则核酸被“稳定化”。复合体中稳定化的核酸独立于其磷酸二酯骨架而保持结合，使得采用限制内切核酸酶处理不会导致该复合体的分解，并且在该复合体不失去其完整性的情况下可获得内部双链DNA断裂。

[0128] 备选地或组合地，包含核酸和核酸结合部分的核酸复合体通过增强其结合或使得其抵抗降解或溶解的处理而被“稳定化”。使复合体稳定化的实例包括采用固定剂如甲醛或补骨脂素处理该复合体，或采用UV光处理，以诱导核酸与结合部分之间的交联或结合部分之间的交联，使得该复合体或多个复合体抵抗例如在限制内切核酸酶处理或用于诱导核酸剪切的处理之后的降解或溶解。

[0129] 如本文所用的术语“支架”一般是指被长度已知而序列未知的缺口隔开，或被长度未知的缺口隔开而已知存在于单个分子上的叠连群，或指通过测序读取的配偶对彼此连接的有序且定向的叠连群组。在叠连群被已知长度的缺口隔开的情况下，可通过多种方法测定所述缺口的序列，该方法包括PCR扩增，随后进行测序(对于较小的缺口)，以及细菌人工染色体(BAC)克隆方法，随后进行测序(对于较大的缺口)。

[0130] 如本文所用的术语“稳定化的样品”是指经由分子间相互作用关于缔合分子稳定化的核酸，使得所述核酸和缔合分子以抵抗分子操作如限制内切核酸酶处理、DNA剪切、核酸断裂的标记或连接的方式结合。本领域已知的核酸包括但不限于DNA和RNA及其衍生物。分子间相互作用可以是共价的或非共价的。共价结合的示例性方法包括但不限于交联技术、偶联反应或本领域普通技术人员已知的其他方法。非共价相互作用的示例性方法包括经由离子相互作用的结合、氢键键合、卤素键合、范德华力(例如，偶极相互作用)、π效应(例如，π-π相互作用、阳离子-π和阴离子-π相互作用、极性π相互作用等)、疏水作用以及本领域普通技术人员已知的其他非共价相互作用。缔合分子的实例包括但不限于染色体蛋白质(例如，组蛋白)、转座酶以及已知与核酸共价或非共价相互作用的任何纳米颗粒。

[0131] 如本文所用的术语“异质样品”是指包含各种群体的核酸(例如，DNA、RNA)、细胞、生物体或其他生物分子的生物样品。在许多情况下，核酸来源于超过一种生物体。例如，异质核酸样品可包含至少约1000、2000、3000、4000、5000、6000、7000、8000、9000、10,000、20,000、50,000、100,000、200,000、500,000、1,000,000、2,000,000、5,000,000、10,000,000或更多个DNA分子。另外，DNA分子中的每一个可包含至少一种或至少两种或超过两种生物体的全部或部分基因组，使得异质核酸样品可包含至少约1000、2000、3000、4000、5000、6000、7000、8000、9000、10,000、20,000、50,000、100,000、200,000、500,000、1,000,000、2,000,000、5,000,000、10,000,000个或更多个不同生物体的全部或部分基因组。异质样品的实例是从多种来源获得的样品，该来源包括但不限于受试者的血液、汗液、尿液、粪便或皮肤；或环境来源(例如，土壤、海水)；食物来源；废物处理场如垃圾场、下水道或公厕；或垃圾桶。

[0132] 生物体的“部分基因组”可包含生物体的整个基因组的至少约10％、20％、30％、40％、50％、60％、70％、80％、90％、95％、99％或更多，或可包含含有整个基因组的至少约10％、20％、30％、40％、50％、60％、70％、80％、90％、95％、99％或更多序列信息的序列数据集。

[0133] 如本文所用的术语“重构的染色质”可指通过使分离的核蛋白与裸DNA复合而形成的形成染色质。

[0134] 如本文所用的术语“标记的序列”可指包含添加的序列的DNA序列，所述添加的序列可用于识别或关联所述序列以用于分析目的。例如，共有相同标记的一组标记的序列可分箱(bin)在一起。在一些实例中，在同一箱子中的标记的序列进一步被分配共同的相位或被分配给共同的起源分子。“标记”的示例性方法包括但不限于使用酶(例如，转座酶、连接

酶)引入标记,和/或将DNA区段彼此共价连接以获得读取对。标记的序列通过例如获得末端读取进行“测序”,其中一个末端读取包含标记序列,而另一个末端读取包含该标记已被添加到的区段的序列。在一些情况下,对整个标记、标记-区段接头和整个区段进行测序。然而,这对于有效标记和测序来说并非总是必需的。相反,在许多情况下,对标记端的可识别部分和区段端的可识别部分的测序足以实现对标记区段的“测序”,特别地但不是唯一地,当可获得叠连群信息如先前生成的或同时生成的叠连群信息时。类似地,在一些情况下,配对端标记序列通过获得末端读取进行“测序”,其中每个末端读取包含连接的区段的可识别序列。配对的端片段可进行完全测序,使得获得接合序列,但是这对于有效的配对端标记和测序来说并非总是必需的。因此,如本文所用,‘对标记的区段进行测序’或‘对配对的末端读取进行测序’不必包括获得所述连接的分子的完全的端对端序列。只要获得所述分子的任一端的可识别序列,使得获得被连接以形成连接的分子的核酸的同一性,就可将连接的片段称为已进行“测序”。在一些情况下,所述测序包括跨越连接接头的端对端测序。在一些情况下,所述测序包括由连接的分子的任一端生成读取。

[0135] 如本文所用的术语“读取对”或“读取-对”可指被连接以提供序列信息的两个或更多个元件。在一些情况下,读取对的数目可指可映射读取对的数目。在其他情况下,读取对的数目可指生成的读取对的总数目。

[0136] 如本文所用的术语“结合”或“缔合”或其派生词是指经由分子间相互作用使分子相对于另一分子稳定。分子间相互作用本质上可以是共价的或非共价的。共价结合的示例性方法包括但不限于交联技术、偶联反应或本领域普通技术人员已知的其他方法。非共价相互作用的示例性方法包括离子相互作用、氢键键合、卤素键合、范德华力(例如,偶极相互作用)、π效应(例如,π-π相互作用、阳离子-π和阴离子-π相互作用、极性π相互作用等)、疏水作用以及本领域普通技术人员已知的其他非共价相互作用。

[0137] 如本文所用的术语“固定”或“固定化”是指使分子或复合体对于对象稳定化。例如,当使DNA复合体对于固体支持体稳定化时,该DNA复合体被固定至固体支持体。在一些情况下,即使当经历多个洗涤步骤时,所述固定的DNA复合体仍将对于所述固体支持体保持稳定化。

[0138] 除非另有定义,否则本文使用的所有技术和科学术语具有如本发明所属领域的普通技术人员通常所理解的相同的含义。现在描述示例性的方法和材料,但与本文描述的任何方法和试剂类似或等效的方法和试剂也可在所公开的方法和组合物的实践中使用。

[0139] 本公开内容提供了用于生成极长范围读取对以及利用该数据推进所有前述追求的方法。在一些实施方案中,本公开内容提供了采用仅约3亿个读取对产生高度连续且准确的人类基因组组装的方法。在其他实施方案中,本公开内容提供了以99%或更大的准确度对人类基因组中90%或更多的杂合变体进行定相的方法。此外,本公开内容生成的读取对的范围可扩展至跨越大得多的基因组距离。由标准鸟枪法文库以及极长范围读取对文库产生所述组装。在其他实施方案中,本公开内容提供了能够利用这两个测序数据集的软件。采用单个长范围读取对文库产生定相的变体,所述读取从该文库映射至参考基因组,并随后用于将变体分配给个体的两个亲代染色体之一。最后,本公开内容提供了使用已知技术的甚至更大的DNA片段的提取,以生成特别长的读取。

[0140] 这些重复阻碍组装和比对过程的机理相当简单,并且最终是具有分歧的结果(图

5)。在大的重复区域的情况下,困难在于跨度。如果读取或读取对的长度不足以跨越重复区域,则不能确信地连接毗连该重复元件的区域。在重复元件较小的情况下,问题主要在于位置。当区域两侧为基因组中常见的两个重复元件时,确定该区域的确切位置由于在两侧的元件与其类别的所有其他元件的相似性而变得困难(即使不是不可能)。在这两种情况下,所述重复中区别信息的缺乏使得特定重复的识别具有挑战性,并因此使特定重复的位置具有挑战性。所需要的是用实验方法建立被重复区域包围或隔开的独特区段之间的连接。

[0141] 本公开内容的方法通过克服这些重复区域所形成的巨大障碍极大地推进了基因组学的领域,并且可因此实现基因组分析的许多领域中的重要进展。为了采用先前的技术进行从头组装,必须满足于将组装片段化成许多小支架,或调用大量时间和资源来产生大插入片段文库或使用其他方法生成更连续的组装。这类方法可包括获得很深的测序覆盖、构建BAC或fosmid文库、光映射,或者最可能包括这些技术及其他技术的一些组合。这种紧张的资源和时间需求使得这类方法对于大多数小的实验室都是遥不可及的,并且阻碍了对非模型生物体的研究。由于本文所述的方法可产生很长范围的读取对,因此可采用单次测序运行实现从头组装。这将会使组装成本按数量级削减,并将所需时间从数月或数年缩短至数周。在一些情况下,本文公开的方法允许在少于14天、少于13天、少于12天、少于11天、少于10天、少于9天、少于8天、少于7天、少于6天、少于5天、少于4天内或在任意两个前述规定的时间段之间的范围内生成多个读取对。例如,所述方法可允许在约10天至14天内生成多个读取对。构建针对甚至最具生态位的生物体的基因组将变成例行程序,系统发育分析将不会受限于缺乏比较,并且项目如Genome10k可能会实现。

[0142] 类似地,针对医疗目的的结构和定相分析也仍然具有挑战性。在癌症之间、患有相同类型的癌症的个体之间或甚至在同一肿瘤内存在令人惊讶的异质性。在低的每个样品成本下从结果性效果中找出原因需要很高的精度和通量。在个性化医学的领域中,基因组治疗的金标准之一是具有完全表征和定相的全部变体的经测序的基因组,包括大的和小的结构重排以及新型突变。为了实现该标准,先前的技术需要与从头组装所需的努力类似的努力,所述从头组装目前太过昂贵和费力而不能成为常规医疗程序。所公开的方法可在低成本下快速产生完全、准确的基因组,并且可因此产生在人类疾病的研究和治疗中许多高度寻求的能力。

[0143] 最后,将本文公开的方法应用于定相可将统计方法的便利与家族分析的准确性相结合,从而比单独使用任一种方法更节约金钱、劳动力和样品。从头变体定相是一种先前的技术无法实现(prohibitive)的非常可取的定相分析,它可使用本文公开的方法容易地执行。这尤其重要,因为绝大多数的人类变异是罕见的(低于5%次要等位基因频率)。定相信息对于种群遗传研究是有价值的,所述种群遗传研究因高度连接的单元型(被分配给单个染色体的变体的集合)的网络获得相对于未连接的基因型的显著优势。单元型信息可实现对群体大小的历史变化、迁移和亚群之间的更换的更高分辨率的研究,并且允许我们将具体的变体追溯回到特定的父母和祖父母。这反过来阐明了与疾病有关的变体的遗传传递,以及当在单个个体中聚在一起时变体之间的相互影响。本公开内容的方法可最终实现极长范围读取对(XLRP)文库的制备、测序和分析。

[0144] 在本公开内容的一些实施方案中,可提供来自受试者的组织或DNA样品,并且所述方法可返回组装的基因组、与判定的变体(包括大结构变体)的比对、定相的变体判定或任

何附加的分析。在其他实施方案中，本文公开的方法可直接为个体提供XLRP文库。

[0145] 在本公开内容的各个实施方案中，本文公开的方法可生成被长距离隔开的极长范围读取对。该距离的上限可通过收集大尺寸DNA样品的能力来改善。在一些情况下，读取对可跨越多达50、60、70、80、90、100、125、150、175、200、225、250、300、400、500、600、700、800、900、1000、1500、2000、2500、3000、4000、5000kbp或更远的基因组距离。在一些实例中，读取对可跨越多达500kbp的基因组距离。在其他实例中，读取对可跨越多达2000kbp的基因组距离。本文公开的方法可整合并基于分子生物学方面的标准技术进行构建，并且进一步地非常适合于提高效率、特异性和基因组覆盖率。在一些情况下，读取对可在少于约1、2、3、4、5、6、7、8、9、10、11、12、13、14、15、16、17、18、19、20、21、22、23、24、25、26、27、28、29、30、60或90天内生成。在一些实例中，读取对可在少于约14天内生成。在一些实例中，读取对可在少于约10天内生成。在一些情况下，本公开内容的方法可以在正确排序和/或定向所述多个叠连群方面以至少约50％、约60％、约70％、约80％、约90％、约95％、约99％或约100％的准确度提供大于约5％、约10％、约15％、约20％、约30％、约40％、约50％、约60％、约70％、约80％、约90％、约95％、约99％或约100％的读取对。例如，所述方法可在正确排序和/或定向所述多个叠连群方面提供约90％至100％的准确度。

[0146] 在其他实施方案中，本文公开的方法可与目前采用的测序技术一起使用。例如，所述方法可与经过良好测试的和/或广泛部署的测序仪器组合使用。在一些实施方案中，本文公开的方法可与来源于目前采用的测序技术的技术和方法一起使用。

[0147] 本公开内容的方法显著地简化了各种生物体的从头基因组组装。利用先前的技术，这类组装目前受限于经济配偶对文库的短插入片段。虽然在fosmid可访问的多达40-50kbp的基因组距离下有可能生成读取对，但这些读取对是昂贵的、麻烦的，并且太短而无法跨越最长的重复序列段，包括着丝粒内的那些重复序列段，所述重复序列段在人类中大小范围为300kbp至5Mbp。本文公开的方法可提供能够跨越长距离(例如，兆碱基或更长)的读取对，从而克服这些支架完整性挑战。因此，通过利用本公开内容的方法产生染色体水平的组装可以成为例行程序。用于组装的更多费力途径——目前花费研究实验室难以置信的大量时间和金钱，并且昂贵到无法接受(prohibiting expansive)的基因组目录——可能变得不必要，从而释放资源用于更有意义的分析。类似地，长范围定相信息的获取可向群体基因组、系统发育和疾病研究提供极大的附加力量。本文公开的方法实现了对大量个体的准确定相，从而扩展了我们在群体和深层时间水平下探测基因组的能力的宽度和深度。

[0148] 在个性化医学领域中，由本文公开的方法生成的XLRP读取对代表了朝着准确、低成本、定相以及快速产生的个人基因组的有意义的进展。当前方法在长距离下对变体定相的能力不足，从而阻碍了对复合杂合基因型的表型影响的表征。此外，对于基因组疾病的实质感兴趣的结构变体，由于与用于研究所述结构变体的读取和读取对插入片段相比的大尺寸，而难以采用当前技术准确地识别和表征。跨越几十个千碱基至兆碱基或更长碱基的读取对可帮助缓解这一困难，从而允许结构变异的高度平行和个性化的分析。

[0149] 高通量测序方面的技术进展正在驱动基础进化和生物医学研究。虽然全基因组测序和组装曾专属于大型基因组测序中心，然而目前商购可得的测序仪足够便宜，使得大多数研究型大学都拥有一台或数台这样的机器。目前，生成大量DNA序列数据是相对便宜的。然而，在理论和实践中，仍然难以采用当前技术产生高质量、高度连续的基因组序列。此外，

由于人们想要分析的大多数生物体(包括人类)是二倍体,因此每个个体都具有基因组的两个单倍体拷贝。在杂合性的位点(例如,其中母亲给予的等位基因不同于父亲给予的等位基因),很难知道哪些组的等位基因来自哪个父母(称为单元型定相)。这种信息可用于进行若干项进化和生物医学研究如疾病和特征相关研究。

[0150] 在各个实施方案中,本公开内容提供了用于基因组组装的方法,该方法将用于DNA制备的技术与用于高通量发现给定基因组内短期、中期和长期联系的配对端测序相结合。本公开内容进一步提供了使用这些联系帮助基因组组装的方法,以用于单元型定相和/或用于宏基因组研究。虽然本文提供的方法可用于测定受试者的基因组的组装,但是还应当理解的是,本文提供的方法也可用于测定该受试者的基因组的部分如染色体的组装,或该受试者的不同长度的染色质的组装。

[0151] 在一些实施方案中,本公开内容提供了本文公开的一种或多种方法,该方法包括由从受试者获得的靶DNA的测序片段生成多个叠连群的步骤。可通过采用一种或多种核酸酶(例如,限制酶)切割DNA、剪切DNA或其组合使靶DNA的长序列段片段化。可使用高通量测序法对得到的片段进行测序,以获得多个测序读取。可与本公开内容的方法一起使用的高通量测序法的实例包括但不限于Roche Diagnostics开发的454焦磷酸测序法、Illumina开发的“集群(clusters)”测序法、Life Technologies开发的SOLiD和Ion半导体测序法以及Complete Genomics开发的DNA纳米球测序法。随后可组装不同的测序读取的重叠端以形成叠连群。或者,可将片段化的靶DNA克隆至载体中。随后用该DNA载体转染细胞或生物体以形成文库。在复制所述转染的细胞或生物体后,分离载体并对该载体进行测序以生成多个测序读取。随后可组装不同的测序读取的重叠端以形成叠连群。

[0152] 如图1所示,基因组组装,特别是采用高通量测序技术可能是有问题的。通常,所述组装由数以千计或数以万计的短叠连群组成。这些叠连群的顺序和方向通常是未知的,从而限制了基因组组装的有用性。存在为这些支架排序和定向的技术,但是它们通常是昂贵的、劳动密集的,并且往往无法发现很长范围的相互作用。

[0153] 可通过任意数量的方法从受试者获得包含用于生成叠连群的靶DNA的样品,所述方法包括采集体液(例如,血液、尿液、血清、淋巴液、唾液、肛门和阴道分泌物、汗液和精液)、采集组织或收集细胞/生物体。获得的样品可由单一类型的细胞/生物体组成,或可由多种类型的细胞/生物体组成。可从该受试者的样品提取并制备DNA。例如,可使用已知的裂解缓冲液、声处理技术、电穿孔等处理该样品以裂解包含多核苷酸的细胞。可通过使用醇提法、铯梯度和/或柱色谱法进一步纯化靶DNA以去除污染物,如蛋白质。

[0154] 在本公开内容的其他实施方案中,提供了用于提取极高分子量DNA的方法。在一些情况下,可通过增加输入DNA的片段大小来改善来自XLRP文库的数据。在一些实例中,从细胞提取DNA的兆碱基大小的片段可在基因组中产生被兆碱基隔开的读取对。在一些情况下,产生的读取对可提供在大于约10kB、约50kB、约100kB、约200kB、约500kB、约1Mb、约2Mb、约5Mb、约10Mb或约100Mb的跨度上的序列信息。在一些实例中,读取对可提供在大于约500kB的跨度上的序列信息。在一些实例中,读取对可提供在大于约2Mb的跨度上的序列信息。在一些情况下,可通过非常温和的细胞裂解(Teague,B.等人.(2010)Proc.Nat.Acad.Sci.USA 107(24),10848-53)和琼脂糖块(Schwartz,D.C.,&Cantor,C.R.(1984)Cell,37(1),67-75)提取极高分子量DNA。在其他情况下,可纯化多达兆碱基长度的DNA分子的商购可得的机器

可用于提取极高分子量DNA。

[0155] 在各个实施方案中，本公开内容提供了一种或多种本文公开的方法，该方法包括探测活细胞内染色体的物理布局的步骤。通过测序探测染色体的物理布局的技术的实例包括“C”类技术，如染色体构象捕获(“3C”)、环状染色体构象捕获(“4C”)、副本染色体捕获(“5C”)以及其他基于染色质捕获的方法；以及基于ChIP的方法，如ChIP-loop、ChIP-PET。这些技术利用活细胞中染色质的固定来确定(cement)细胞核中的空间关系。产物的后续处理和测序允许研究人员恢复基因组区域之间邻近关联的矩阵。通过进一步的分析，这些关联可用于产生染色体的三维几何图，因为它们物理地排列在活细胞核中。这样的技术描述了活细胞中染色体的离散空间组织，并提供了染色体位点之间功能相互作用的准确视图。困扰这些功能研究的一个问题是非特异性相互作用的存在，在数据中存在的关联仅仅可归因于染色体邻近。在本公开内容中，通过本文提供的方法捕获这些非特异性染色体内相互作用，以提供对于组装有价值的信息。

[0156] 在一些实施方案中，染色体内相互作用与染色体连接性有关。在一些情况下，染色体内数据可帮助基因组组装。在一些情况下，在体外重构染色质。这可能是有利的，因为染色质——特别是组蛋白(染色质的主要蛋白质组分)——对于在用于通过测序检测染色质构象和结构的最常见的“C”类技术：3C、4C、5C和染色质捕获下进行固定是重要的。染色质关于序列是高度非特异性的，并且通常将在整个基因组上均匀地组装。在一些情况下，不使用染色质的物种的基因组可以在重构的染色质上进行组装，从而将本公开内容的范围扩大至所有生命领域。

[0157] 染色质构象捕获技术总结于图2。简而言之，在密切物理邻近的基因组区域之间产生交联。根据本文其他地方进一步详细描述或本领域已知的合适方法，可完成染色质内的蛋白质(如组蛋白)与DNA分子例如基因组DNA的交联。在一些情况下，两个或更多个核苷酸序列，或更严格地说，两个或更多个核酸区段可经由与一个或多个核苷酸序列结合的蛋白质进行交联。一种方法是将染色质暴露于紫外线照射(Gilmour等人，Proc.Nat’l.Acad.Sci.USA 81:4275-4279,1984)。还可以利用其他方法，如化学或物理(例如，光)交联进行多核苷酸区段的交联。合适的化学交联剂包括但不限于甲醛和补骨脂素(Solomon等人，Proc.NatL.Acad.Sci.USA 82:6470-6474,1985；Solomon等人，Cell 53:937-947,1988)。例如，可通过将2％甲醛添加至包含DNA分子和染色质蛋白质的混合物进行交联。可用于交联DNA的药剂的其他实例包括但不限于UV光、丝裂霉素C、氮芥、美法仑、1,3-丁二烯双环氧化物、顺二胺二氯铂(II)和环磷酰胺。适当地，所述交联剂将形成桥接相对较短距离如约2 Å的交联，从而选择可以逆转的密切相互作用。

[0158] 在一些实施方案中，DNA分子可在交联之前或之后进行免疫沉淀。在一些情况下，可使DNA分子片段化。片段可与结合配偶体如抗体接触，所述抗体特异性地识别并结合乙酰化组蛋白例如H3。这类抗体的实例包括但不限于可从Upstate Biotechnology，Lake Placid，N.Y.获得的抗乙酰化组蛋白H3。随后可从免疫沉淀物收集来自免疫沉淀物的多核苷酸。在使染色质片段化之前，乙酰化组蛋白可与相邻多核苷酸序列交联。

[0159] 在某些实施方案中，DNA分子与多个缔合分子结合，其中所述缔合分子不采用亲和标记物(例如，生物素、链霉亲和素、亲和素、多组氨酸、EDTA等)进行共价修饰。在一些情况下，缔合分子直接从生物体分离。在一些实例中，缔合分子包含氨基酸。在某些实例中，缔合

分子包括多肽或蛋白质。在一些实例中，缔合分子包括组蛋白。在各个实例中，缔合分子来自与DNA分子不同的来源。例如，DNA分子可与多个组蛋白交联，其中所述组蛋白不采用亲和标记物进行共价修饰。在其他情况下，缔合分子是转座酶。在一些实例中，第一DNA分子与缔合分子非共价结合。在其他实例中，第一DNA分子与缔合分子非共价结合。在一些情况下，第一DNA分子与缔合分子交联。在一些实例中，使用固定剂(例如，甲醛)使第一DNA分子与缔合分子交联。然而，在某些情况下，DNA分子包含可经亲和标记物修饰的DNA区段。在一些实例中，亲和标记物包括生物素。在某些实例中，亲和标记物是生物素修饰的核苷三磷酸(dNTP)。在一些实例中，亲和标记物是生物素修饰的脱氧核糖胞嘧啶三磷酸(dCTP)。在各种情况下，亲和标记物用于分离或纯化DNA区段。

[0160] 使用未经共价修饰的缔合分子减少了步骤数并且/或者提高了本公开内容中提供的方法的效率。在一些情况下，在连接DNA区段以形成连接的DNA区段之前洗涤DNA区段少于约20、18、16、15、14、13、12、11、10、9、8、7、6、5、4、3、2或1次。在某些情况下，在连接DNA区段以形成连接的DNA区段之前洗涤DNA区段少于约15、14、13、12、11、10、9、8、7、6或5次。在一些情况下，在连接DNA区段以形成连接的DNA区段之前洗涤DNA区段少于约12、11、10、9、8、7或6次。在一些实例中，在连接DNA区段以形成连接的DNA区段之前洗涤DNA区段少于约10次。在某些实例中，在连接DNA区段以形成连接的DNA区段之前洗涤DNA区段少于约8次。在一些实例中，在连接DNA区段以形成连接的DNA区段之前洗涤DNA区段少于约6次。

[0161] 在一些实施方案中，将结合的DNA分子固定在固体支持体上。在一些情况下，固体支持体是珠子。在一些实例中，珠子包含聚合物。在一些实例中，聚合物是聚苯乙烯。在其他实例中，聚合物是聚乙二醇(PEG)。在各个实例中，珠子是磁珠。在一些实例中，珠子是固相可逆固定化(SPRI)珠子。在其他情况下，固体支持体是阵列。在某些实例中，固体支持体不与亲和标记物(例如，生物素、链霉亲和素、亲和素、多组氨酸、EDTA或其衍生物)共价连接。在各个实例中，固体支持体不与任何多肽(例如，链霉亲和素、亲和素、多组氨酸标记或其衍生物)连接。

[0162] 在一些情况下，将固体支持体修饰为在不存在共价修饰的情况下结合缔合分子，而不是共价修饰缔合分子以通过结合至固体支持体的表面(如包被有用于结合共价附接至缔合分子的生物素的链霉亲和素的表面)促进其分离。在一些情况下，这是缔合分子与缔合分子表面的直接结合。或者，在一些情况下，通过溶剂中的至少一种组分来介导结合。在一些情况下，使用直接结合缔合分子的部分包被固体支持体。在一些情况下，使用直接结合核酸的部分包被固体表面。在各个实施方案中，合适的涂层包括多胺、带正电荷的部分、羧基基团和带负电荷的部分。

[0163] 在一些情况下，处理交联的DNA分子以分级分离或切断混合物中的多核苷酸。分级分离技术是本领域已知的，并且包括例如用于生成较小的基因组片段的剪切技术。可使用用于使染色质片段化的完善方法实现片段化，所述方法包括例如，声处理、剪切和/或使用核酸酶(例如，限制酶)或片段化酶(例如，dsDNA片段化酶(fragmentase))。限制酶可具有1、2、3、4、5或6个碱基长的限制位点。核酸酶可以是内切核酸酶、外切核酸酶或内切-外切核酸酶。核酸酶的实例包括但不限于脱氧核糖核酸酶I(DNase I)和微球菌酶(MNase)。限制酶的实例包括但不限于AatII、Acc65I、AccI、AciI、AclI、AcuI、AfeI、AflII、AflIII、AgeI、AhdI、AleI、AluI、AlwI、AlwNI、ApaI、ApaLI、ApeKI、ApoI、AscI、AseI、AsiSI、AvaI、AvaII、AvrII、

BaeGI、BaeI、BamHI、BanI、BanII、BbsI、BbvCI、BbvI、BccI、BceAI、BcgI、BciVI、BclI、BfaI、BfuAI、BfuCI、BglI、BglII、BlpI、BmgBI、BmrI、BmtI、BpmI、Bpu10I、BpuEI、BsaAI、BsaBI、BsaHI、BsaI、BsaJI、BsaWI、BsaXI、BscRI、BscYI、BsgI、BsiEI、BsiHKAI、BsiWI、BslI、BsmAI、BsmBI、BsmFI、BsmI、BsoBI、Bsp1286I、BspCNI、BspDI、BspEI、BspHI、BspMI、BspQI、BsrBI、BsrDI、BsrFI、BsrGI、BsrI、BssHII、BssKI、BssSI、BstAPI、BstBI、BstEII、BstNI、BstUI、BstXI、BstYI、BstZ17I、Bsu36I、BtgI、BtgZI、BtsCI、BtsI、Cac8I、ClaI、CspCI、CviAII、CviKI-1、CviQI、DdcI、DpnI、DpnII、DraI、DraIII、DrdI、EacI、EagI、EarI、EciI、Eco53kI、EcoNI、Eco0109I、EcoP15I、EcoRI、EcoRV、FatI、FauI、Fnu4HI、FokI、FseI、FspI、HaeII、HaeIII、HgaI、HhaI、HincII、HindIII、HinfI、HinPlI、HpaI、HpaII、HphI、Hpy166II、Hpy188I、Hpy188III、Hpy99I、HpyAV、HpyCH4III、HpyCH4IV、HpyCH4V、KasI、KpnI、MboI、MboII、MfeI、MluI、MlyI、MmeI、MnlI、MscI、MseI、MslI、MspAlI、MspI、MwoI、NaeI、NarI、Nb.BbvCI、Nb.BsmI、Nb.BsrDI、Nb.BtsI、NciI、NcoI、NdeI、NgoMIV、NheI、NlaIII、NlaIV、NmeAIII、NotI、NruI、NsiI、NspI、Nt.AlwI、Nt.BbvCI、Nt.BsmAI、Nt.BspQI、Nt.BstNBI、Nt.CviPII、PacI、PaeR7I、PciI、PflFI、PflMI、PhoI、PleI、PmeI、PmlI、PpuMI、PshAI、PsiI、PspGI、PspOMI、PspXI、PstI、PvuI、PvuII、RsaI、RsrII、SacI、SacII、SalI、SapI、Sau3AI、Sau96I、SbfI、ScaI、ScrFI、SexAI、SfaNI、SfcI、SfiI、SfoI、SgrAI、SmaI、SmlI、SnaBI、SpeI、SphI、SspI、StuI、StyD4I、StyI、SwaI、T、TaqαI、TfiI、TliI、TseI、Tsp45I、Tsp509I、TspMI、TspRI、Tth111I、XbaI、XcmI、XhoI、XmaI、XmnI和ZraI。得到的片段的大小可以变化。得到的片段还可在5’或3’端包含单链突出端。核酸酶可以是核酸引导的核酸酶。核酸引导的核酸酶可以是RNA引导的核酸酶，如来自核酸酶的Cas家族(例如，Cas9)，包括CAS I类I型、CAS I类III型、CAS I类IV型、CAS II类II型以及CAS II类V型，如Cas9、Cpf1、Cas3、Cas8a-c、Cas10、Cse1、Csy1、Csn2、Cas4、Csm2、Cm5和Csf1。

[0164] 在一些实施方案中，使用声处理技术可以获得约100至5000个核苷酸的片段。或者，可以获得约100至1000、约150至1000、约150至500、约200至500、或约200至400个核苷酸的片段。制备样品用于对交联的偶联序列区段进行测序。在一些情况下，例如，可通过连接分子内交联的两个序列区段产生多核苷酸的单个短序列段。可使用本文别处进一步详细描述的或本领域已知的任何合适的测序技术，如高通量测序法，从样品获得序列信息。例如，连接产物可经受配对端测序，从而从片段的每一端获得序列信息。序列区段对可以表示在获得的序列信息中，从而关联沿着所述多核苷酸将这两个序列区段隔开的线性距离内的单元型分析信息。

[0165] 通过染色质捕获生成的数据的一个特征是，大多数读取对在映射回基因组时被发现紧密地线性邻近。也就是说，大多数读取对被发现在基因组中彼此靠近。在得到的数据集中，正如在染色体占据不同区域的情况下所预期的，染色体内接触的概率平均比染色体间接触的概率高得多。此外，尽管相互作用的概率随着线性距离迅速衰减，但即使在相同染色体上被>200Mb隔开的基因座也比不同染色体上的基因座更有可能相互作用。在检测长范围染色体内，以及特别是染色体间接触时，这种短范围和中间范围染色体内接触的“背景”是将使用染色质捕获分析剔除的背景噪声。

[0166] 显著地，在真核生物中的染色质捕获实验已显示除了物种特异性和细胞类型特异性的染色质相互作用以外的两种典型的相互作用模式。一种模式，即距离依赖性衰减(DDD)

是作为基因组距离的函数的相互作用频率的衰减的一般趋势。第二种模式，即顺反比率(CTR)是与不同染色体上的基因座相比位于相同染色体上的基因座之间显著更高的相互作用频率，即使在被几十个兆碱基的序列隔开时也是如此。这些模式可以反映一般聚合物动力学，其中邻近的基因座具有随机相互作用的更高概率，以及特定的核组织特征，如染色体区域的形成，即间期染色体倾向于占据细胞核中不同容积，几乎没有相互混合(little mixing)的现象。尽管这两种模式的确切细节可在物种、细胞类型和细胞条件之间变化，但它们是普遍存在且突出的。这些模式是如此强大且一致，使得它们用于评估实验质量，且通常从所述数据进行归一化，以揭示详细的相互作用。-然而，在本文公开的方法中，基因组组装可利用基因组的三维结构。使典型的染色质捕获相互作用模式阻碍特定的成环相互作用的分析的特征，即所述模式的普遍性、强度和一致性，可用作用于估算叠连群的基因组位置的强有力的工具。

[0167] 在特定的实施中，染色体内读取对之间的物理距离的检测指示关于基因组组装的数据的若干有用的特征。首先，较短范围相互作用比较长范围相互作用更常见(例如，参见图6)。也就是说，与遥远的区域相比，读取对中的每个读取更有可能与实际基因组中附近的区域配对。第二，中间范围和长范围相互作用存在长尾(long tail)。也就是说，读取对携带关于千碱基(kB)或甚至兆碱基(Mb)距离下的染色体内排列的信息。例如，读取对可提供在大于约10kB、约50kB、约100kB、约200kB、约500kB、约1Mb、约2Mb、约5Mb、约10Mb或约100Mb的跨度上的序列信息。数据的这些特征简单地表明，在相同的染色体上邻近的基因组的区域更有可能在物理上紧密地邻近，这是由于它们通过DNA骨架彼此化学连接而预期的结果。据推测，基因组范围的染色质相互作用数据集，如通过染色质捕获生成的那些数据集，将提供关于沿着整个染色体的序列的分组和线性组织的长范围信息。

[0168] 尽管用于染色质捕获的实验方法简单明了且成本相对较低，但当前用于基因组组装和单元型分析的方案需要10^6-10^8个细胞，这是可能无法获得(特别是从某些人类患者样品获得)的相当大量的材料。相反，本文公开的方法包括允许针对采用来自细胞的明显较少的材料的基因型组装、单元型定相和宏基因组学的准确的预测结果的方法。例如，少于约0.1μg、约0.2μg、约0.3μg、约0.4μg、约0.5μg、约0.6μg、约0.7μg、约0.8μg、约0.9μg、约1.0μg、约1.2μg、约1.4μg、约1.6μg、约1.8μg、约2.0μg、约2.5μg、约3.0μg、约3.5μg、约4.0μg、约4.5μg、约5.0μg、约6.0μg、约7.0μg、约8.0μg、约9.0μg、约10μg、约15μg、约20μg、约30μg、约40μg、约50μg、约60μg、约70μg、约80μg、约90μg、约100μg、约150μg、约200μg、约300μg、约400μg、约500μg、约600μg、约700μg、约800μg、约900μg或约1000μg的DNA可与本文公开的方法一起使用。在一些实例中，可从少于约1,000,000、约500,000、约100,000、约50,000、约10,000、约5,000、约1,000、约5,000或约1,000、约500或约100个细胞中提取本文公开的方法中使用的DNA。

[0169] 在一些情况下，来自DNA分子的少于约80％、60％、50％、40％、30％、20％、15％、10％、9％、8％、7％、6％、5％、4％、3％、2％、1％、0.5％或0.1％的DNA区段与来自任何其他DNA分子的DNA区段连接。在某些情况下，来自DNA分子的少于50％、40％、30％、20％、15％、10％、9％、8％、7％、6％或5％的DNA区段与来自任何其他DNA分子的DNA区段连接。在一些情况下，来自DNA分子的少于40％、30％、20％、15％或10％的DNA区段与来自任何其他DNA分子的DNA区段连接。在一些实例中，来自DNA分子的少于40％的DNA区段与来自任何其他DNA分

子的DNA区段连接。在某些实例中，来自DNA分子的少于20％的DNA区段与来自任何其他DNA分子的DNA区段连接。在一些实例中，来自DNA分子的少于10％的DNA区段与来自任何其他DNA分子的DNA区段连接。

[0170] 普遍地，用于探测染色体的物理布局的程序，如基于染色质捕获的技术，利用在细胞/生物体内形成的染色质，如从培养的细胞或原代组织中分离的染色质。本公开内容不仅提供了采用从细胞/生物体中分离的染色质的此类技术的使用，而且提供了采用重构的染色质的此类技术的使用。重构的染色质在多种特征方面与细胞/生物体内形成的染色质不同。首先，对于许多样品，可通过使用多种非侵入性至侵入性的方法，如通过收集体液、擦拭口腔或直肠区域、采集上皮样品等，实现裸DNA样品的收集。第二，重构染色质基本上防止了染色体间和其他长范围相互作用的形成，所述相互作用生成用于基因组组装和单元型定相的人工制品。在一些情况下，根据本公开内容的方法和组合物，样品可具有少于约20％、15％、12％、11％、10％、9％、8％、7％、6％、5％、4％、3％、2％、1％、0.5％、0.4％、0.3％、0.2％、0.1％或更少的染色体间或分子间交联。在一些实例中，所述样品可具有少于约5％的染色体间或分子间交联。在一些实例中，所述样品可具有少于约3％的染色体间或分子间交联。在一些实例中，所述样品可具有少于约1％的染色体间或分子间交联。第三，可以调节能够交联的位点的频率，并因此可以调节多核苷酸内的分子内交联的频率。例如，DNA与组蛋白的比可以变化，使得核小体密度可以调节至所需的值。在一些情况下，核小体密度减小至生理学水平以下。因此，可以改变交联的分布以有利于较长范围相互作用。在一些实施方案中，可制备具有变化的交联密度的子样品以涵盖短范围和长范围缔合。例如，可以调节交联条件，使得至少约1％、约2％、约3％、约4％、约5％、约6％、约7％、约8％、约9％、约10％、约11％、约12％、约13％、约14％、约15％、约16％、约17％、约18％、约19％、约20％、约25％、约30％、约40％、约45％、约50％、约60％、约70％、约80％、约90％、约95％或约100％的交联发生在样品DNA分子上相隔至少约50kb、约60kb、约70kb、约80kb、约90kb、约100kb、约110kb、约120kb、约130kb、约140kb、约150kb、约160kb、约180kb、约200kb、约250kb、约300kb、约350kb、约400kb、约450kb或约500kb的DNA区段之间。

[0171] 在各个实施方案中，本公开内容提供了多种能够将多个读取对映射至多个叠连群的方法。存在若干个用于将读取映射至叠连群序列的公开可得的计算机程序。这些读取映射程序数据还提供了描述特定读取映射在基因组内的独特性的数据。从在叠连群内以高置信度独特映射的读取群，我们可以推断每个读取对中的读取之间的距离的分布。这些是图6中示出的数据。对于读取确信地映射至不同叠连群的读取对，该映射数据暗示正在讨论的这两个叠连群之间的连接。所述映射数据还暗示了与从以上所述的分析了解到的距离的分布成比例的这两个叠连群之间的距离。因此，读取映射至不同叠连群的每个读取对暗示正确的组装中的那两个叠连群之间的连接。从所有这类映射的读取对推断出的连接可总结于邻接矩阵中，其中每个叠连群通过行和列来表示。连接叠连群的读取对在表示读取对中的读取被映射至的叠连群的相应行和列中被标记为非零值。大多数读取对将映射在叠连群内，并且从该叠连群可了解读取对之间的距离的分布，并且可以使用映射至不同叠连群的读取对由该叠连群构建叠连群的邻接矩阵。

[0172] 在各个实施方案中，本公开内容提供了方法，该方法包括使用来自读取对数据的读取-映射数据构建叠连群的邻接矩阵。在一些实施方案中，邻接矩阵使用针对读取对的加

权方式，引入了短范围相互作用优于长范围相互作用的趋势(例如，参见图3)。跨越较短距离的读取对一般比跨越较长距离的读取对更常见。可使用映射至单个叠连群的读取对数据拟合描述特定距离的概率的函数以了解该分布。因此，映射至不同叠连群的读取对的一个重要特征是在该读取对映射的叠连群上的位置。对于都映射在叠连群的一端附近的读取对，所推断的这些叠连群之间的距离可以很短，并且因此连接的读取之间的距离很小。因为读取对之间较短距离比较长距离更常见，所以与映射为远离叠连群的边缘的读取相比，该结构提供了更强有力的证据，证明这两个叠连群相邻。因此，邻接矩阵中的连接通过读取到叠连群的边缘的距离来进一步加权。在一些实施方案中，所述邻接矩阵被重新缩放以降低一些叠连群上大量接触的权重，所述接触表示基因组的混杂区域。基因组的这些区域(通过具有高比例的映射至这些区域的读取而可识别)，根据先验更有可能含有可能误传组装的假读取映射。在其他实施方案中，可通过搜索调节染色质的支架化相互作用的一种或多种物质，如转录抑制因子CTCF、内分泌受体、黏结蛋白或共价修饰的组蛋白的一个或多个保守结合位点来指导这种缩放。

[0173] 在一些实施方案中，本公开内容提供了本文公开的一种或多种方法，该方法包括分析邻接矩阵以确定穿过叠连群的路径的步骤，该路径代表叠连群相对于基因组的顺序和/或方向。在其他实施方案中，可以选择穿过叠连群的路径，使得每个叠连群被恰好访问一次。在一些实施方案中，选择穿过叠连群的路径，使得穿过邻接矩阵的路径将被访问的边缘权重的总和最大化。以这种方式，为正确的组装推荐最可能的叠连群连接。在其他实施方案中，可以选择穿过叠连群的路径，使得每个叠连群被恰好访问一次，并且将邻接矩阵的边缘加权最大化。

[0174] 在二倍体基因组中，了解相同染色体上哪些等位基因变体连接往往是重要的。这被称为单元型定相。来自高通量序列数据的短读取很少允许直接观察哪些等位基因变体连接。在长距离下单元型定相的计算机推断可能是不可靠的。本公开内容提供了一种或多种方法，其允许使用读取对上的等位基因变体确定哪些等位基因变体连接。

[0175] 在各个实施方案中，本公开内容的方法和组合物能够使二倍体或多倍体基因组关于多个等位基因变体进行单元型定相。本文所述的方法因此可基于来自读取对和/或组装的叠连群的变体信息使用该变体信息确定连接的等位基因变体。等位基因变体的实例包括但不限于来自1000genomes、UK10K、HapMap和用于发现人类中的遗传变异的其他项目的已知的那些等位基因变体。通过拥有单元型定相数据，可更容易地揭示与特定基因有关的疾病，所述单元型定相数据是例如通过发现导致Charcot-Marie-Tooth神经病的SH3TC2的两个拷贝中未连接的、失活突变(Lupski JR,Reid JG,Gonzaga-Jauregui C等人.N.Engl.J.Med.362:1181-91,2010)以及导致高胆固醇血症9的ABCG5的两个拷贝中的未连接的、失活突变(Rios J,Stein E,Shendure J等人.Hum.Mol.Genet.19:4313-18,2010)所显示的。

[0176] 人类在平均1,000个位点中有1个位点杂合。在一些情况下，使用高通量测序法的单通道数据可生成至少约150,000,000个读取对。读取对可以为约100个碱基对长。从这些参数估计，来自人类样品的全部读取的十分之一覆盖杂合位点。因此，估计平均来自人类样品的全部读取对的百分之一覆盖一对杂合位点。因此，约1,500,000个读取对(150,000,000的百分之一)提供使用单通道的定相数据。由于在人类基因组中有约30亿个碱基，并且一千

个中有一个是杂合的，因此平均人类基因组中有约300万个杂合位点。在代表一对杂合位点的约1,500,000个读取对的情况下，使用典型的高通量测序机器，待使用单通道的高通量测序法定相的每个杂合位点的平均覆盖率为约(1X)。二倍体人类基因组因此可采用一个通道的高通量测序数据可靠地且完全地进行定相，从而关联来自使用本文公开的方法制备的样品的序列变体。在一些实例中，数据的通道可以是DNA序列读取数据的集合。在一些实例中，数据的通道可以是来自高通量测序仪器的单次运行的DNA序列读取数据的集合。

[0177] 由于人类基因组由两个同源染色体组组成，因此了解个体的真正基因组成需要描绘出遗传物质的母本和父本拷贝或单元型。获得个体中的单元型在若干方法中是有用的。首先，单元型用于在临床上预测器官移植中供体-宿主匹配的结果，并且越来越多地被用作检测疾病关联的手段。第二，在显示复合杂合性的基因中，单元型提供了关于两个有害变体是否位于同一等位基因上的信息，从而极大地影响了关于这些变体的遗传是否有害的预测。第三，个体群体的单元型提供了关于人口结构和人类进化史的信息。最后，最近描述的基因表达中广泛的等位基因失衡表明，等位基因之间的遗传或表观遗传差异可能导致定量的表达差异。对单元型结构的理解将描绘出导致等位基因失衡的变体的机制。

[0178] 在某些实施方案中，本文公开的方法包括如长范围连接和定相所需的，用于(例如，在体外或在体内)固定和捕获基因组的远离的区域之间的关联的技术。在一些情况下，所述方法包括构建XLRP文库并对该文库进行测序，以递送基因组上距离非常远的读取对。在一些情况下，相互作用主要由单个DNA片段内的随机关联产生。在一些实例中，可以推断出区段之间的基因组距离，因为在DNA分子中彼此接近的区段相互作用更频繁，且概率更高，而分子的远离部分之间的相互作用将不那么频繁。因此，连接两个基因座的对的数目与在输入DNA上所述对的邻近之间存在系统关系。本公开内容可以产生能够跨越提取物中最大的DNA片段的读取对，如图2所示。该文库的输入DNA的最大长度为150kbp，这是我们从测序数据中观察到的最长的有意义的读取对。这表明如果提供更大的输入DNA片段，本方法可以连接基因组上距离更远的基因座。通过应用特别适合于处理由本方法产生的数据类型的改进的组装软件工具，完整的基因组组装可能是可行的。

[0179] 通过使用本公开内容的方法和组合物产生的数据可以实现极高的定相准确度。与先前的方法相比，本文所述的方法可对更高比例的变体进行定相。可以在维持高水平的准确度的同时实现定相。该相位信息可以扩展至更长范围，例如大于约200kbp、约300kbp、约400kbp、约500kbp、约600kbp、约700kbp、约800kbp、约900kbp、约1Mbp、约2Mbp、约3Mbp、约4Mbp、约5Mbp或约10Mbp。在一些实施方案中，人类样品的大于90％的杂合SNP可以使用少于约2.5亿个读取或读取对(例如通过使用仅1个通道的Illumina HiSeq数据)，以大于99％的准确度进行定相。在其他情况下，人类样品的大于约40％、50％、60％、70％、80％、90％、95％或99％的杂合SNP可以使用少于约2.5亿或约5亿个读取或读取对(例如通过使用仅1或2个通道的Illumina HiSeq数据的)，以大于约70％、80％、90％、95％或99％的准确度进行定相。例如，人类样品的大于95％或99％的杂合SNP可以使用少于约2.5亿或约5亿个读取，以大于约95％或99％的准确度进行定相。在一些情况下，可以通过使读取长度增加至约200bp、250bp、300bp、350bp、400bp、450bp、500bp、600bp、800bp、1000bp、1500bp、2kbp、3kbp、4kbp、5kbp、10kbp、20kbp、50kbp或100kbp来捕获附加变体。

[0180] 在本公开内容的其他实施方案中，来自XLRP文库的数据可用于确认长范围读取对

的定相能力。如图6所示，那些结果的准确度与先前可用的最佳技术相当，但进一步扩展至显著更长的距离。用于特定测序方法的当前样品制备方案识别位于靶向限制位点的读取长度例如150bp内的变体，以进行定相。在一个实例中，在针对用于组装的基准样品NA12878构建的XLRP文库中，存在的1,703,909个杂合SNP中的44％以大于99％的准确度进行定相。在一些情况下，这一比例可以通过合理选择限制酶或采用不同的酶的组合扩展至几乎所有的变异位点。

[0181] 在一些实施方案中，本文所述的组合物和方法允许研究宏基因组，例如在人类的肠道中发现的那些宏基因组。因此，可以研究存在于给定生态环境中的一些或全部生物体的部分或全基因组序列。实例包括对全部肠道微生物、在皮肤的某些区域上发现的微生物以及生活在有毒废物处理场中的微生物的随机测序。这些环境中的微生物群体的组成可以使用本文所述的组合物和方法以及由微生物各自的基因组编码的相关生物化学的方面来确定。本文所述的方法可以实现复杂的生物环境的宏基因组研究，例如包含超过2、3、4、5、6、7、8、9、10、12、15、20、25、30、40、50、60、70、80、90、100、125、150、175、200、250、300、400、500、600、700、800、900、1000、5000、10000种或更多种生物体和/或生物体的变体的那些生物环境。

[0182] 可以使用本文所述的方法和系统实现癌症基因组测序所需的高准确度。当对癌症基因组进行测序时不准确的参考基因组可能带来碱基判定挑战。异质样品和小的起始材料，例如通过活检获得的样品引入了附加的挑战。此外，大规模的结构变体和/或杂合性丢失的检测对于癌症基因组测序，以及区分体细胞变体和碱基判定错误的能力来说往往至关重要。

[0183] 本文所述的系统和方法可由含有2、3、4、5、6、7、8、9、10、12、15、20个或更多个不同基因组的复杂样品生成准确的长序列。可以对正常、良性和/或肿瘤来源的混合样品进行分析(任选地不需要正常对照)。在一些实施方案中，利用仅100ng或甚至仅数百个基因组等同物的起始样品生成准确的长序列。本文所述的系统和方法可允许检测大规模结构变体和重排，可在跨越约1kbp、约2kbp、约5kbp、约10kbp、20kbp、约50kbp、约100kbp、约200kbp、约500kbp、约1Mbp、约2Mbp、约5Mbp、约10Mbp、约20Mbp、约50Mbp或约100Mbp或更多的核苷酸的长序列内获得定相的变体判定。例如，可在跨越约1Mbp或约2Mbp的长序列内获得相位变体判定。

[0184] 可将使用本文所述的方法和系统确定的单元型分配给计算资源，例如经由网络的计算资源，如云系统。如果需要的话，可以使用储存在计算资源中的有关信息来校正短变体判定。可以基于来自短变体判定的复合信息和储存在计算资源中的信息来检测结构变体。基因组的有问题部分，如区段重复、有结构变异倾向的区域、高度变异的医学上有关的MHC区域、着丝粒和端粒区域以及其他异染色质区域(包括但不限于具有重复区域、低序列准确度、高变异率、ALU重复、区段重复或本领域已知的任何其他有关的有问题部分的那些异染色质区域)，可进行重新组装以提高准确度。

[0185] 可将样品类型分配给本地或网络化的计算资源(诸如云)中的序列信息。在信息的来源已知的情况下，例如，当信息的来源来自癌症或正常组织时，可将来源以样品类型的一部分的形式分配给样品。其他样品类型实例通常包括但不限于组织类型、样品收集方法、感染的存在、感染的类型、处理方法、样品的大小等。在可获得完全或部分比较基因组序列，如正常基因组与癌症基因组的比较的情况下，可以确定样品数据与比较基因组序列之间的差异并任选地输出该差异。

[0186] 本文的方法可用于分析感兴趣的选择性基因组区域以及可与感兴趣的选择性区域相互作用的基因组区域的遗传信息。如本文公开的扩增方法可在遗传分析的领域已知的装置、试剂盒和方法中使用，例如但不限于美国专利号6,449,562、6,287,766、7,361,468、7,414,117、6,225,109和6,110,709中发现的那些装置、试剂盒和方法。在一些情况下，本公开内容的扩增方法可用于扩增靶核酸以进行DNA杂交研究，以确定多态性的存在或不存在。所述多态性或等位基因可与疾病或病况如遗传病有关。在其他情况下，多态性可与疾病或病况的易感性有关，例如，与成瘾、退行性和与年龄相关的病况、癌症等有关的多态性。在其他情况下，多态性可与有益的特征(如增加的冠状动脉健康，或对疾病如HIV或疟疾的抵抗力，或对退行性疾病如骨质疏松、阿尔茨海默病或痴呆的抵抗力)有关。

[0187] 本公开内容的组合物和方法可用于诊断、预后、治疗、患者分层、药物开发、治疗选择和筛选目的。本公开内容提供了以下优点，即使用本公开内容的方法可同时分析来自单个生物分子样品的许多不同靶分子。这允许例如对一个样品进行若干种诊断测试。

[0188] 本公开内容的组合物和方法可用于基因组学。本文所述的方法可快速提供答案，这对于该应用来说是非常理想的。本文所述的方法和组合物可用于发现生物标志物的过程，所述生物标志物可用于诊断或预后以及作为健康和疾病的指示物。本文所述的方法和组合物可用于筛选药物，例如药物开发、治疗的选择、疗效的确定，和/或确定药物开发的目标。在涉及药物的筛选测定中测试基因表达的能力是非常重要的，因为蛋白质是体内最终的基因产物。在一些实施方案中，本文所述的方法和组合物将同时测量蛋白质和基因表达，这将提供关于正在进行的特定筛选的大部分信息。

[0189] 本公开内容的组合物和方法可用于基因表达分析。本文所述的方法区分核苷酸序列。靶核苷酸序列之间的差异可以是例如，单个核酸碱基差异、核酸缺失、核酸插入或重排。还可以检测到涉及超过一个碱基的这样的序列差异。本公开内容的方法能够检测感染性疾病、遗传病和癌症。所述方法还用于环境监测、取证和食品科学。可对核酸进行的遗传分析的实例包括例如，SNP检测、STR检测、RNA表达分析、启动子甲基化、基因表达、病毒检测、病毒亚型分析和药物抗性。

[0190] 本发明方法可应用于获自或来源于患者的生物分子样品的分析，以确定样品中是否存在患病的细胞类型、疾病的阶段、患者的预后、患者对特定治疗作出反应的能力或对于患者的最佳治疗。本发明方法还可应用于鉴定特定疾病的生物标志物。

[0191] 在一些实施方案中，本文所述的方法用于病况的诊断。如本文所用的术语病况的“诊断(diagnose)”或“诊断(diagnosis)”可包括预测或诊断病况，从而确定病况的诱因、监测病况的治疗、诊断疾病的治疗反应，或对病况、病况进展或对病况的特定治疗的反应进行预后。例如，可根据本文所述的任何一种方法测定血液样品，以确定该样品中疾病或恶性细胞类型的标志物的存在和/或量，从而对疾病或癌症进行诊断或分期。

[0192] 在一些实施方案中，本文所述的方法和组合物用于病况的诊断和预后。

[0193] 许多免疫、增殖性和恶性疾病和病症特别适合于本文所述的方法。免疫疾病和病症包括变应性疾病和病症、免疫功能病症以及自身免疫疾病和病况。变应性疾病和病症包括但不限于变应性鼻炎、变应性结膜炎、变应性哮喘、特应性湿疹、特应性皮炎和食物变态反应。免疫缺陷包括但不限于重症联合免疫缺陷(SCID)、嗜酸细胞增多综合征、慢性肉芽肿病、白细胞粘附缺陷I和II、高IgE综合征、Chediak Higashi、嗜中性粒细胞增多症、嗜中性

粒细胞减少症、发育不全、丙种球蛋白缺乏症、高IgM综合征、DiGeorge/Velocardial面部综合征(DiGeorge/Velocardial-facial syndromes)和干扰素γ-TH1通路缺陷。自身免疫和免疫失调病症包括但不限于类风湿性关节炎、糖尿病、系统性红斑狼疮、格雷夫斯病(Graves'disease)、格雷夫斯眼病、克罗恩病(Crohn's disease)、多发性硬化、银屑病、系统性硬化、甲状腺肿和淋巴瘤性甲状腺肿(桥本甲状腺炎(Hashimoto's thyroiditis)、淋巴细胞性甲状腺肿)、脱发、自身免疫性心肌炎、硬化性苔藓、自身免疫性葡萄膜炎、艾迪生病(Addison's disease)、萎缩性胃炎、重症肌无力、特发性血小板减少性紫癜、溶血性贫血、原发性胆汁性肝硬化、韦格纳肉芽肿病(Wegener's granulomatosis)、结节性多动脉炎和炎性肠病、同种异体移植排斥以及来自对感染性微生物或环境抗原的变态反应的组织破坏。

[0194] 可通过本公开内容的方法评估的增殖性疾病和病症包括但不限于新生儿血管瘤病;继发性进行性多发性硬化;慢性进行性骨髓退行性疾病;神经节瘤病(ganglioneuromatosis);节细胞性神经瘤;瘢痕疙瘩形成;骨佩吉特病(Paget's Disease of the bone);纤维囊性疾病(例如,乳房或子宫的纤维囊性疾病);结节病;Peronies和Duputren纤维化,肝硬化,动脉粥样硬化和血管再狭窄。

[0195] 可通过本公开内容的方法评估的恶性疾病和病症包括恶性血液病和实体瘤。

[0196] 当样品为血液样品时,恶性血液病特别适合于本公开内容的方法,因为这样的恶性病涉及血源性细胞的变化。这样的恶性病包括非霍奇金淋巴瘤(non-Hodgkin's lymphoma)、霍奇金淋巴瘤、非B细胞淋巴瘤和其他淋巴瘤、急性或慢性白血病、红细胞增多症、血小板增多、多发性骨髓瘤、骨髓增生异常症、骨髓增生性疾病、骨髓纤维化、非典型免疫淋巴细胞增生和浆细胞病症。

[0197] 可通过本公开内容的方法评估的浆细胞病症包括多发性骨髓瘤、淀粉样变性和瓦尔登斯特伦巨球蛋白血症(Waldenstrom's macroglobulinemia)。

[0198] 实体瘤的实例包括但不限于结肠癌、乳腺癌、肺癌、前列腺癌、脑肿瘤、中枢神经系统肿瘤、膀胱肿瘤、黑色素瘤、肝癌、骨肉瘤和其他骨癌、睾丸和卵巢癌、头颈部肿瘤和宫颈肿瘤。

[0199] 还可以通过本公开内容的方法检测遗传病。这可以通过产前或产后筛查染色体和遗传变异或遗传病来进行。可检测的遗传病的实例包括:21羟化酶缺乏、囊性纤维化、脆性X综合征、特纳综合征(Turner Syndrome)、杜氏肌营养不良(Duchenne Muscular Dystrophy)、唐氏综合征或其他三体综合征、心脏病、单基因疾病、HLA分型、苯丙酮尿症、镰状细胞贫血、Tay-Sachs病、地中海贫血、克兰费尔特综合征(Klinefelter Syndrome)、亨廷顿病、自身免疫病、脂肪沉积、肥胖缺陷、血友病、先天性代谢缺陷和糖尿病。

[0200] 本文所述的方法可用于通过确定样品中细菌或病毒各自的标志物的存在和/或量来诊断病原体感染,例如通过细胞内细菌和病毒引起的感染。

[0201] 可通过本公开内容的方法检测各种感染性疾病。感染性疾病可能是由细菌、病毒、寄生虫和真菌致病原引起的。还可以使用本公开内容测定各种致病原对药物的抗性。

[0202] 可通过本公开内容检测的细菌致病原包括大肠杆菌(Escherichia coli)、沙门氏菌属(Salmonella)、志贺氏菌属(Shigella)、克雷伯氏菌属(KlESBiella)、假单胞菌属(Pseudomonas)、单核细胞增多性利斯特氏菌(Listeria monocytogenes)、结核分枝杆菌(Mycobacterium tuberculosis)、鸟胞内分枝杆菌(Mycobacterium aviumintracellulare)、

耶尔森氏菌属(Yersinia)、弗朗西丝氏菌属(Francisella)、巴斯德氏菌属(Pasteurella)、布鲁氏菌属(Brucella)、梭菌属(Clostridia)、百日咳博德特氏菌(Bordetellapertussis)、拟杆菌属(Bacteroides)、金黄色葡萄球菌(Staphylococcus aureus)、肺炎链球菌(Streptococcus pneumonia)、B-溶血性链球菌(B-Hemolytic strep.)、棒杆菌属(Corynebacteria)、军团菌属(Legionella)、枝原体属(Mycoplasma)、脲原体属(Ureaplasma)、衣原体属(Chlamydia)、淋病奈瑟氏球菌(Neisseria gonorrhea)、脑膜炎奈瑟氏菌(Neisseria meningitides)、流感嗜血杆菌(Hemophilus influenza)、粪肠球菌(Enterococcus faecalis)、普通变形杆菌(Proteus vulgaris)、奇异变形杆菌(Proteus mirabilis)、幽门螺杆菌(Helicobacter pylori)、梅毒螺旋体(Treponema palladium)、布氏疏螺旋体(Borrelia burgdorferi)、回归热疏螺旋体(Borrelia recurrentis)、立克次体病原体(Rickettsial pathogens)、诺卡氏菌属(Nocardia)和放线菌(Acitnomycetes)。

[0203] 可通过本公开内容检测的真菌致病原包括新型隐球菌(Cryptococcus neoformans)、皮炎芽生菌(Blastomyces dermatitidis)、荚膜组织胞浆菌(Histoplasma capsulatum)、粗球孢子菌(Coccidioides immitis)、巴西副球孢子菌(Paracoccidioides brasiliensis)、白假丝酵母(Candida albicans)、烟曲霉(Aspergillus fumigautus)、藻状菌纲(Phycomycetes)(根霉属(Rhizopus))、申克孢子丝菌(Sporothrix schenckii)、Chromomycosis和Maduromycosis。

[0204] 可通过本公开内容检测的病毒致病原包括人类免疫缺陷病毒、人类T细胞淋巴细胞营养病毒、肝炎病毒(例如,乙型肝炎病毒和丙型肝炎病毒)、EB病毒、巨细胞病毒、人类乳头瘤病毒、正粘病毒、副粘病毒、腺病毒、冠状病毒、弹状病毒、脊髓灰质炎病毒、披膜病毒、布尼亚病毒(bunya viruses)、沙粒病毒、风疹病毒和呼肠孤病毒。

[0205] 可通过本公开内容检测的寄生虫致病原包括恶性疟原虫(Plasmodium falciparum)、疟疾疟原虫(Plasmodium malaria)、间日疟原虫(Plasmodium vivax)、卵形疟原虫(Plasmodium ovale)、旋盘尾丝虫(Onchoverva volvulus)、利什曼原虫属(Leishmania)、锥虫属(Trypanosoma spp.)、血吸虫属(Schistosoma spp.)、溶组织内阿米巴(Entamoeba histolytica)、隐孢子虫(Cryptosporidum)、贾第虫属(Giardia spp.)、毛滴虫属(Trichimonas spp.)、结肠小袋虫(Balatidium coli)、班氏丝虫(Wuchereria bancrofti)、弓浆虫属(Toxoplasma spp.)、蠕形住肠蛲虫(Enterobius vermicularis)、人蛔虫(Ascaris lumbricoides)、毛首鞭形线虫(Trichuris trichiura)、Dracunculus medinesis、吸虫类(trematodes)、阔节裂头绦虫(Diphyllobothrium latum)、绦虫属(Taenia spp.)、卡氏肺孢子虫(Pneumocystis carinii)和Necator americanis。

[0206] 本公开内容还用于通过致病原检测药物抗性。例如,万古霉素抗性屎肠球菌(Enterococcus faecium)、甲氧西林抗性金黄色葡萄球菌、青霉素抗性肺炎链球菌、多种药物抗性结核分枝杆菌和AZT抗性人类免疫缺陷病毒均可采用本公开内容来鉴定。

[0207] 因此,使用本公开内容的组合物和方法检测的靶分子可以是患者标志物(如癌症标志物)或感染外来病原体的标志物,如细菌或病毒标志物。

[0208] 本公开内容的组合物和方法可用于鉴定和/或定量靶分子,所述靶分子的丰度指示生物状态或疾病状况,例如,作为疾病状态的结果而上调或下调的血液标志物。

[0209] 在一些实施方案中,本公开内容的方法和组合物可用于细胞因子表达。本文所述

方法的低灵敏度将有助于细胞因子(例如,作为疾病如癌症的状况、诊断或预后的生物标志物)的早期检测,以及亚临床状况的鉴定。

[0210] 靶多核苷酸所来源的不同样品可包含来自相同个体的多个样品、来自不同个体的样品或其组合。在一些实施方案中,样品包含来自单个个体的多个多核苷酸。在一些实施方案中,样品包含来自两个或更多个个体的多个多核苷酸。个体是可以产生靶多核苷酸的任何生物体或其部分,所述个体的非限制性实例包括植物、动物、真菌、原生生物、无核原生物、病毒、线粒体和叶绿体。可从受试者分离样品多核苷酸,如来源于受试者的细胞样品、组织样品或器官样品,包括例如,培养的细胞系、活检物、血液样品或含有细胞的流体样品。受试者可以为动物,包括但不限于动物如牛、猪、小鼠、大鼠、鸡、猫、狗等,并且通常为哺乳动物,如人。样品还可以通过人工获得,如通过化学合成。在一些实施方案中,样品包含DNA。在一些实施方案中,样品包含基因组DNA。在一些实施方案中,样品包含线粒体DNA、叶绿体DNA、质粒DNA、细菌人工染色体、酵母人工染色体、寡核苷酸标记或其组合。在一些实施方案中,样品包含通过使用引物和DNA聚合酶的任何合适的组合进行引物延伸反应而生成的DNA,所述引物延伸反应包括但不限于聚合酶链反应(PCR)、逆转录及其组合。在引物延伸反应的模板为RNA的情况下,逆转录的产物被称为互补DNA(cDNA)。用于引物延伸反应的引物可包括对一个或多个靶标具有特异性的序列、随机序列、部分随机序列及其组合。适合于引物延伸反应的反应条件是本领域已知的。通常,样品多核苷酸包含样品中存在的任何多核苷酸,其可能包括或可能不包括靶多核苷酸。

[0211] 在一些实施方案中,核酸模板分子(例如,DNA或RNA)从含有多种其他组分,如蛋白质、脂质和非模板核酸的生物样品中分离。可从获自动物、植物、细菌、真菌或任何其他细胞生物体的任何细胞材料获得核酸模板分子。用于本公开内容的生物样品包括病毒颗粒或制剂。核酸模板分子可直接从生物体或从获得自生物体的生物样品(例如从血液、尿液、脑脊液、精液、唾液、痰液、粪便和组织)获得。任何组织或体液样本均可用作用于本公开内容的核酸的来源。还可从培养的细胞,如原代细胞培养物或细胞系分离核酸模板分子。可用病毒或其他细胞内病原体感染从中获得模板核酸的细胞或组织。样品还可以是从生物样本提取的总RNA、cDNA文库、病毒或基因组DNA。样品还可以是来自非细胞来源的分离的DNA,例如来自冰箱的扩增/分离的DNA。

[0212] 用于提取和纯化核酸的方法是本领域公知的。例如,核酸可通过采用苯酚、苯酚/氯仿/异戊醇,或类似的制剂,包括TRIzol和TriReagent进行有机提取来纯化。提取技术的其他非限制性实例包括:(1)有机提取,然后进行乙醇沉淀,例如,在使用或不使用自动化核酸提取器例如,可从Applied Biosystems(Foster City,Calif.)获得的型号341DNA提取器的情况下,使用苯酚/氯仿有机试剂(Ausubel等人,1993);(2)固定相吸附法(美国专利号5,234,809;Walsh等人,1991);和(3)盐诱导的核酸沉淀法(Miller等人,(1988)),这样的沉淀法通常被称为“盐析”法。核酸分离和/或纯化的另一个实例包括使用核酸可特异性或非特异性结合的磁性粒子,然后使用磁铁分离珠子,洗涤珠子并从珠子洗脱核酸(参见,例如美国专利号5,705,628)。在一些实施方案中,以上的分离方法可在酶消化步骤之前,以帮助从样品中除去不需要的蛋白质,例如,采用蛋白酶K或其他类似的蛋白酶进行消化。参见,例如,美国专利号7,001,724。如果需要的话,可向裂解缓冲液中添加核糖核酸酶(RNase)抑制剂。对于某些细胞或样品类型,可能需要在方案中添加蛋白质变性/消化步骤。纯化方法可

针对分离DNA、RNA或二者。当DNA和RNA二者在提取程序过程中或之后被一起分离出时，可采用进一步的步骤通过使彼此分离而纯化其中的一种或二者。还可以生成提取的核酸的亚级分，例如，按大小、序列或其他物理或化学特征进行纯化。除了初始的核酸分离步骤以外，还可在本公开内容的方法中的任何步骤之后进行核酸的纯化，如用于去除过量或不需要的试剂、反应物或产物。

[0213] 可如2003年10月9日公开的美国专利申请公开号US2002/0190663A1中所述获得核酸模板分子。通常，可通过多种技术，如通过Maniatis等人，Molecular Cloning:A Laboratory Manual,Cold Spring Harbor,N.Y.,pp.280-281(1982)所述的那些技术从生物样品提取核酸。在一些情况下，首先可从生物样品提取核酸，然后在体外进行交联。在一些情况下，可进一步从核酸中去除天然缔合蛋白质(例如，组蛋白)。

[0214] 在其他实施方案中，本公开内容可容易地应用于任何高分子量双链DNA，包括例如从组织、细胞培养物、体液、动物组织、植物、细菌、真菌、病毒等分离的DNA。

[0215] 在一些实施方案中，多个独立样品中的每一个可独立地包含至少约1ng、2ng、5ng、10ng、20ng、30ng、40ng、50ng、75ng、100ng、150ng、200ng、250ng、300ng、400ng、500ng、1μg、1.5μg、2μg、5μg、10μg、20μg、50μg、100μg、200μg、500μg或1000μg或更多的核酸材料。在一些实施方案中，多个独立样品中的每一个可独立地包含少于约1ng、2ng、5ng、10ng、20ng、30ng、40ng、50ng、75ng、100ng、150ng、200ng、250ng、300ng、400ng、500ng、1μg、1.5μg、2μg、5μg、10μg、20μg、50μg、100μg、200μg、500μg或1000μg或更多的核酸。

[0216] 在一些实施方案中，使用商业试剂盒如可从Epicentre Biotechnologies(Madison,WI)获得的那些商业试剂盒，进行末端修复以生成钝性末端5’磷酸化核酸末端。

[0217] 衔接子寡核苷酸包括可与靶多核苷酸连接的具有序列(该序列的至少一部分是已知的)的任何寡核苷酸。衔接子寡核苷酸可包括DNA、RNA、核苷酸类似物、非典型核苷酸、标记的核苷酸、修饰的核苷酸或其组合。衔接子寡核苷酸可以是单链、双链或部分双链体。通常，部分双链体衔接子包含一个或多个单链区域和一个或多个双链区域。双链衔接子可包含两个彼此杂交的单独寡核苷酸(也称为“寡核苷酸双链体”)，并且杂交可留下由错配和/或不成对的核苷酸产生的一个或多个钝性末端、一个或多个3′突出端、一个或多个5′突出端、一个或多个凸起或这些的任何组合。在一些实施方案中，单链衔接子包含能够彼此杂交的两个或更多个序列。当两个这样的可杂交序列包含在单链衔接子中时，杂交产生发夹结构(发夹衔接子)。当衔接子的两个杂交区域通过非杂交区域彼此隔开时，产生“气泡(bubble)”结构。包含气泡结构的衔接子可由包含内部杂交的单个衔接子寡核苷酸组成，或可包含彼此杂交的两个或更多个衔接子寡核苷酸。内部序列杂交，如在衔接子中两个可杂交序列之间的杂交可在单链衔接子寡核苷酸中产生双链结构。不同种类的衔接子可组合使用，如发夹衔接子和双链衔接子或不同序列的衔接子。发夹衔接子中的可杂交序列可以包括或可以不包括寡核苷酸的一端或两端。当两端均不包括在可杂交序列中时，两端为“游离的”或“突出的”。当在衔接子中仅一端与另一序列可杂交时，另一端形成突出端，如3′突出端或5′突出端。当5′-末端核苷酸和3′-末端核苷酸被包括在可杂交序列中使得5′-末端核苷酸和3′-末端核苷酸互补并彼此杂交时，所述端被称为“钝性的”。不同的衔接子可以在顺序反应中或同时与靶多核苷酸连接。例如，可在同一反应中添加第一和第二衔接子。可在与靶多核苷酸组合之前操作衔接子。例如，可以添加或去除末端磷酸。

[0218] 衔接子可含有多种序列元件中的一个或多个，包括但不限于一个或多个扩增引物退火序列或其互补体、一个或多个测序引物退火序列或其互补体、一个或多个条码序列、多个不同衔接子或不同衔接子的亚组之间共有的一个或多个共同的序列、一个或多个限制酶识别位点、与一个或多个靶多核苷酸突出端互补的一个或多个突出端、一个或多个探针结合位点(例如，用于附接至测序平台，如用于大规模平行测序的流通池，如Illumina,Inc.开发的)、一个或多个随机或近似随机序列(例如，在一个或多个位置从两个或更多个不同核苷酸的组随机选择的一个或多个核苷酸，其中在一个或多个位置选择的不同核苷酸中的每一个被表示在包含随机序列的衔接子库中)及其组合。两个或更多个序列元件可彼此不相邻(例如，被一个或多个核苷酸隔开)、彼此相邻、部分重叠或完全重叠。例如，扩增引物退火序列也可充当测序引物退火序列。序列元件可位于3′端或其附近、5′端或其附近或衔接子寡核苷酸的内部。当衔接子寡核苷酸能够形成二级结构如发夹时，序列元件可部分或完全位于二级结构的外部、部分或完全位于二级结构的内部，或位于参与二级结构的序列之间。例如，当衔接子寡核苷酸包含发夹结构时，序列元件可部分或完全位于可杂交序列(“茎”)的内部或外部，包括在可杂交序列之间的序列(“环”)中。在一些实施方案中，具有不同条码序列的多个第一衔接子寡核苷酸中的第一衔接子寡核苷酸包含所述多个第一衔接子寡核苷酸中的所有第一衔接子寡核苷酸之间共同的序列元件。在一些实施方案中，所有第二衔接子寡核苷酸包含所有第二衔接子寡核苷酸之间共同的序列元件，该序列元件不同于所述第一衔接子寡核苷酸共有的共同序列元件。序列元件的差异可以是任何差异，使得例如，由于序列长度的变化、一个或多个核苷酸的缺失或插入或在一个或多个核苷酸位置的核苷酸组合物的变化(如碱基变化或碱基修饰)，不同衔接子的至少一部分不完全对齐。在一些实施方案中，衔接子寡核苷酸包含与一个或多个靶多核苷酸互补的5′突出端、3′突出端或二者。互补突出端的长度可以为一个或多个核苷酸，包括但不限于长度为1、2、3、4、5、6、7、8、9、10、11、12、13、14、15个或更多个核苷酸。例如，互补突出端的长度可以为约1、2、3、4、5或6个核苷酸。互补突出端可包含固定的序列。互补突出端可包含一个或多个核苷酸的随机序列，使得在一个或多个位置从两个或更多个不同核苷酸的组随机选择一个或多个核苷酸，其中在一个或多个位置选择的不同核苷酸中的每一个被表示在包含随机序列的具有互补突出端的衔接子库中。在一些实施方案中，衔接子突出端与通过限制性内切核酸酶消化产生的靶多核苷酸突出端互补。在一些实施方案中，衔接子突出端由腺嘌呤或胸腺嘧啶组成。

[0219] 衔接子寡核苷酸可以具有任何合适的长度，至少足以容纳它们所包含的一个或多个序列元件。在一些实施方案中，衔接子的长度为约、小于约或大于约10、15、20、25、30、35、40、45、50、55、60、65、70、75、80、90、100、200个或更多个核苷酸。在一些实例中，衔接子的长度可以为约10至约50个核苷酸。在一些实例中，衔接子的长度可以为约20至约40个核苷酸。

[0220] 如本文所用的术语“条码”是指允许鉴定条码所关联的多核苷酸的某种特征的已知核酸序列。在一些实施方案中，待鉴定的多核苷酸的特征是多核苷酸所来源的样品。在一些实施方案中，条码的长度可以为至少3、4、5、6、7、8、9、10、11、12、13、14、15个或更多个核苷酸。例如，条码的长度可以为至少10、11、12、13、14或15个核苷酸。在一些实施方案中，条码的长度可以短于10、9、8、7、6、5或4个核苷酸。例如，条码的长度可以短于10个核苷酸。在一些实施方案中，与一些多核苷酸关联的条码具有不同于与其他多核苷酸关联的条码的长度。通常，条码具有足够的长度，并且包含足够不同的序列以允许基于与它们关联的条码来

鉴定样品。在一些实施方案中，在条码序列中的一个或多个核苷酸的突变、插入或缺失，如1、2、3、4、5、6、7、8、9、10个或更多个核苷酸的突变、插入或缺失后，可以准确鉴定条码和与其关联的样品来源。在一些实例中，1、2或3个核苷酸可以突变、插入和/或缺失。在一些实施方案中，多个条码中的每个条码与多个至少两个核苷酸位置，如至少2、3、4、5、6、7、8、9、10个或更多个位置中的每一个其他的条码不同。在一些实例中，每个条码可与至少2、3、4或5个位置中的每一个其他的条码不同。在一些实施方案中，第一位点和第二位点包含多个条码序列中的至少一个。在一些实施方案中，从第一衔接子寡核苷酸的条码中独立地选择第二位点的条码。在一些实施方案中，具有条码的第一位点和第二位点配对，使得所述对的序列包含相同或不同的一个或多个条码。在一些实施方案中，本公开内容的方法进一步包括基于靶多核苷酸连接的条码序列鉴定该靶多核苷酸所来源的样品。通常，条码可包含核酸序列，该核酸序列当与靶多核苷酸连接时充当该靶多核苷酸所来源的样品的标识符。

[0221] 在真核生物中，基因组DNA被包装成染色质，以构成细胞核内的染色体。染色质的基本结构单元是核小体，该核小体由包裹在组蛋白八聚体周围的146个DNA碱基对(bp)组成。组蛋白八聚体由两个拷贝组成，每一个拷贝的核芯为组蛋白H2A-H2B二聚体和H3-H4二聚体。核小体沿着DNA规则地隔开，在DNA中其通常被称为“串上的珠子”。

[0222] 核芯组蛋白和DNA组装成核小体是由伴侣蛋白质和相关的组装因子介导的。几乎所有这些因子都是是核芯组蛋白结合蛋白质。一些组蛋白伴侣蛋白质，如核小体组装蛋白质-1(NAP-1)，表现出与组蛋白H3和H4结合的偏好。还观察到新合成的组蛋白被乙酰化并随后在组装成染色质之后进行脱乙酰化。因此，介导组蛋白乙酰化或脱乙酰化的因子在染色质组装过程中发挥重要作用。

[0223] 通常，已开发了两种体外方法用于重构或组装染色质。一种方法为不依赖ATP的，而第二种方法为ATP依赖性的。用于重构染色质的不依赖ATP的方法包括使DNA和核芯组蛋白加上蛋白质如NAP-1或盐充当组蛋白伴侣蛋白质。该方法导致组蛋白在DNA上的随机排列，该随机排列没有准确地模拟细胞中的天然核芯核小体颗粒。这些颗粒通常被称为单核小体，因为它们不是规则排序、延伸的核小体阵列，并且所用的DNA序列通常不长于250bp(Kundu,T.K.等人,Mol.Cell 6:551-561,2000)。为了在更长长度的DNA序列上生成有序的核小体的延伸阵列，必须通过ATP依赖性方法组装染色质。

[0224] 周期性核小体阵列的ATP依赖性组装，与天然染色质中所见的类似，需要DNA序列、核芯组蛋白颗粒、伴侣蛋白质以及利用ATP的染色质组装因子。ACF(利用ATP的染色质组装和重塑因子)或RSF(重塑和间距因子)是用于使核小体的延伸有序的阵列在体外生成染色质的两种广泛研究的组装因子(Fyodorov,D.V.和Kadonaga,J.T.Method Enzymol.371:499-515,2003;Kundu,T.K.等人.Mol.Cell 6:551-561,2000)。

[0225] 在特定的实施方案中，本公开内容的方法可容易地应用于任何类型的片段化的双链DNA，包括但不限于例如，从血浆、血清和/或尿液分离的游离DNA;来自细胞和/或组织的凋亡DNA;在体外酶促片段化的DNA(例如，通过脱氧核糖核酸酶I和/或限制性内切核酸酶);和/或通过机械力(水力剪切、声处理、雾化等)片段化的DNA。

[0226] 可使从生物样品获得的核酸片段化以产生用于分析的合适的片段。可使用多种机械、化学和/或酶法，将模板核酸片段化或剪切成所需长度。可经由声处理，例如Covaris法，短暂暴露于脱氧核糖核酸酶(DNase)，或使用一种或多种限制酶的混合物或转座酶或切口

酶(nicking enzyme)对DNA进行随机剪切。可通过短暂暴露于核糖核酸酶、热加镁或通过剪切使RNA片段化。RNA可转化为cDNA。如果采用片段化，则RNA可在片段化之前或之后转化为cDNA。在一些实施方案中，通过声处理使来自生物样品的核酸片段化。在其他实施方案中，通过水力剪切仪器使核酸片段化。通常，单独的核酸模板分子可为约2kb碱基至约40kb。在各个实施方案中，核酸可为约6kb-10kb片段。核酸分子可以是单链、双链或具有单链区域的双链(例如，茎和环结构)。

[0227] 在一些实施方案中，交联的DNA分子可经历大小选择步骤。可对小于或大于特定大小的交联的DNA分子进行核酸的大小选择。大小选择可进一步受到交联的频率和/或片段化方法(例如受到选择常见的或罕见的切割限制酶)的影响。在一些实施方案中，可制备组合物，其包括使在约1kb至5Mb、约5kb至5Mb、约5kB至2Mb、约10kb至2Mb、约10kb至1Mb、约20kb至1Mb、约20kb至500kb、约50kb至500kb、约50kb至200kb、约60kb至200kb、约60kb至150kb、约80kb至150kb、约80kb至120kb或约100kb至120kb范围内，或由这些值中的任何值所界定的任何范围内(例如，约150kb至1Mb)的DNA分子交联。

[0228] 在一些实施方案中，使样品多核苷酸片段化成具有一个或多个特定大小范围的片段化的DNA分子的群体。在一些实施方案中，可由至少约1、约2、约5、约10、约20、约50、约100、约200、约500、约1000、约2000、约5000、约10,000、约20,000、约50,000、约100,000、约200,000、约500,000、约1,000,000、约2,000,000、约5,000,000、约10,000,000个或更多个起始DNA的基因组等同物生成片段。可通过本领域已知的方法实现片段化，包括化学、酶和机械片段化。在一些实施方案中，片段具有约10至约10,000、约20,000、约30,000、约40,000、约50,000、约60,000、约70,000、约80,000、约90,000、约100,000、约150,000、约200,000、约300,000、约400,000、约500,000、约600,000、约700,000、约800,000、约900,000、约1,000,000、约2,000,000、约5,000,000、约10,000,000个或更多个核苷酸的平均长度。在一些实施方案中，片段具有约1kb至约10Mb的平均长度。在一些实施方案中，片段具有约1kb至5Mb、约5kb至5Mb、约5kB至2Mb、约10kb至2Mb、约10kb至1Mb、约20kb至1Mb、约20kb至500kb、约50kb至500kb、约50kb至200kb、约60kb至200kb、约60kb至150kb、约80kb至150kb、约80kb至120kb或约100kb至120kb，或由这些值中的任何值所界定的任何范围(例如，约60至120kb)的平均长度。在一些实施方案中，片段具有少于约10Mb、少于约5Mb、少于约1Mb、少于约500kb、少于约200kb、少于约100kb或少于约50kb的平均长度。在其他实施方案中，片段具有大于约5kb、大于约10kb、大于约50kb、大于约100kb、大于约200kb、大于约500kb、大于约1Mb、大于约5Mb或大于约10Mb的平均长度。在一些实施方案中，机械地实现片段化，其包括使样品DNA分子进行声波的声处理。在一些实施方案中，片段化包括采用一种或多种酶在适合于该一种或多种酶生成双链核酸断裂的条件下处理样品DNA分子。用于生成DNA片段的酶的实例包括序列特异性和非序列特异性核酸酶。核酸酶的非限制性实例包括脱氧核糖核酸酶I、片段化酶、限制性内切核酸酶、其变体及其组合。例如，采用脱氧核糖核酸酶I进行的消化可在Mg^{++}的不存在和Mn^{++}的存在下诱导DNA中的随机双链断裂。在一些实施方案中，片段化包括采用一种或多种限制性内切核酸酶处理样品DNA分子。片段化可产生具有5′突出端、3′突出端、钝性末端或其组合的片段。在一些实施方案中，如当片段化包括使用一种或多种限制性内切核酸酶时，样品DNA分子的裂解保留了具有可预测序列的突出端。在一些实施方案中，所述方法包括经由标准方法如柱纯化或从琼脂糖凝胶分离，对片段进行大小选择的步骤。

[0229] 在一些实施方案中，片段化的DNA的5′和/或3′端核苷酸序列在连接之前不进行修饰。例如，可使用通过限制性内切核酸酶的片段化来保留可预测的突出端，然后与包含与DNA片段上的可预测突出端互补的突出端的核酸末端连接。在另一个实例中，可在保留了可预测的钝性末端的通过酶的裂解之后，使钝性末端DNA片段与包含钝性末端的核酸如衔接子、寡核苷酸或多核苷酸连接。在一些实施方案中，在与衔接子连接之前，将片段化的DNA分子进行钝性末端抛光(或“末端修复”)，以产生具有钝性末端的DNA片段。钝性末端抛光步骤可通过与合适的酶(如具有3′至5′外切核酸酶活性和5′至3′聚合酶活性的DNA聚合酶，例如T4聚合酶)一起温育来实现。在一些实施方案中，可在末端修复之后添加1、2、3、4、5、6、7、8、9、10、11、12、13、14、15、16、17、18、19、20个或更多个核苷酸，如一个或多个腺嘌呤、一个或多个胸腺嘧啶、一个或多个鸟嘌呤或一个或多个胞嘧啶，以产生突出端。例如，可在末端修复之后添加1、2、3、4、5或6个核苷酸。具有突出端的DNA片段可与具有互补突出端的一个或多个核酸如寡核苷酸、衔接子寡核苷酸或多核苷酸连接，如在连接反应中。例如，可使用模板独立的聚合酶将单个腺嘌呤添加至末端修复的DNA片段的3′端，然后与各自在3′端具有胸腺嘧啶的一个或多个衔接子连接。在一些实施方案中，核酸如寡核苷酸或多核苷酸可与钝性末端双链DNA分子连接，所述钝性末端双链DNA分子通过采用一个或多个核苷酸延伸3′端，然后进行5′磷酸化进行修饰。在一些情况下，可在可含有镁的合适的缓冲液中在一个或多个dNTP的存在下，采用聚合酶如Klenow聚合酶或本文提供的任何合适的聚合酶，或通过使用末端脱氧核苷酸转移酶进行3′端的延伸。在一些实施方案中，具有钝性末端的靶多核苷酸与包含钝性末端的一个或多个衔接子连接。可在含有ATP和镁的合适的缓冲液中，例如采用T4多核苷酸激酶进行DNA片段分子的5′端的磷酸化。可任选地处理片段化的DNA分子以使5′端或3′端去磷酸化，例如，通过使用本领域已知的酶，如磷酸酶。

[0230] 关于两个多核苷酸如衔接子寡核苷酸和靶多核苷酸，如本文所用的术语“连接(connecting)”、“连接(joining)”和“连接(ligation)”是指两个单独的DNA区段的共价附接，以产生具有连续骨架的单个较大多核苷酸。用于连接两个DNA区段的方法是本领域已知的，并且包括但不限于酶法和非酶法(例如，化学法)。非酶促的连接反应的实例包括通过引用并入本文的美国专利号5,780,613和5,476,930中描述的非酶促连接技术。在一些实施方案中，衔接子寡核苷酸通过连接酶，例如DNA连接酶或RNA连接酶与靶多核苷酸连接。各自具有特征化反应条件的多种连接酶是本领域已知的，并且包括但不限于NAD^+-依赖性连接酶，包括tRNA连接酶、Taq DNA连接酶、丝状栖热菌(Thermus filiformis)DNA连接酶、大肠杆菌DNA连接酶、Tth DNA连接酶、水管致黑栖热菌(Thermus scotoductus)DNA连接酶(I和II)、热稳定的连接酶、Ampligase热稳定的DNA连接酶、VanC型连接酶、9°N DNA连接酶、Tsp DNA连接酶和通过生物勘探发现的新型连接酶；ATP依赖性连接酶，包括T4 RNA连接酶、T4 DNA连接酶、T3 DNA连接酶、T7 DNA连接酶、Pfu DNA连接酶、DNA连接酶1、DNA连接酶III、DNA连接酶IV和通过生物勘探发现的新型连接酶；以及它们的野生型、突变的同种型和基因工程化变体。

[0231] 连接可以在具有可杂交序列的DNA区段如互补突出端之间。连接还可以在两个钝性末端之间。通常，在连接反应中使用5′磷酸。可通过靶多核苷酸、衔接子寡核苷酸或二者提供5′磷酸。可根据需要在待连接的DNA区段中添加或去除5′磷酸。用于添加或去除5′磷酸的方法是本领域已知的，并且包括但不限于酶法和化学法。用于添加和/或去除5′磷酸的酶

包括激酶、磷酸酶和聚合酶。在一些实施方案中，在连接反应中连接的两端(例如，衔接子端和靶多核苷酸端)均提供5′磷酸，使得在连接两端时生成两个共价键。在一些实施方案中，在连接反应中连接的两端中的仅一端(例如，衔接子端和靶多核苷酸端中的仅一个)提供5′磷酸，使得在连接两端时仅生成一个共价键。

[0232] 在一些实施方案中，在靶多核苷酸的一端或两端的仅一条链与衔接子寡核苷酸连接。在一些实施方案中，在靶多核苷酸的一端或两端的两条链均与衔接子寡核苷酸连接。在一些实施方案中，在连接之前去除3′磷酸。在一些实施方案中，向靶多核苷酸的两端添加衔接子寡核苷酸，其中在每一端的一条或两条链与一个或多个衔接子寡核苷酸连接。当在两端的两条链均与衔接子寡核苷酸连接时，在连接之后可进行裂解反应，该裂解反应保留了可充当用于延伸相应3′端的模板的5′突出端，所述3′端可以包括或可以不包括来源于衔接子寡核苷酸的一个或多个核苷酸。在一些实施方案中，靶多核苷酸在一端与第一衔接子寡核苷酸连接而在另一端与第二衔接子寡核苷酸连接。在一些实施方案中，靶多核苷酸的两端与单个衔接子寡核苷酸的相对端连接。在一些实施方案中，靶多核苷酸以及与其连接的衔接子寡核苷酸包含钝性末端。在一些实施方案中，针对每个样品使用包含至少一个条码序列的不同的第一衔接子寡核苷酸，可对每个样品进行单独的连接反应，使得没有条码序列与超过一个样品的靶多核苷酸连接。具有与其连接的衔接子寡核苷酸的DNA区段或靶多核苷酸被认为通过连接的衔接子进行“标记”。

[0233] 在一些情况下，可在约0.1ng/μL、约0.2ng/μL、约0.3ng/μL、约0.4ng/μL、约0.5ng/μL、约0.6ng/μL、约0.7ng/μL、约0.8ng/μL、约0.9ng/μL、约1.0ng/μL、约1.2ng/μL、约1.4ng/μL、约1.6ng/μL、约1.8ng/μL、约2.0ng/μL、约2.5ng/μL、约3.0ng/μL、约3.5ng/μL、约4.0ng/μL、约4.5ng/μL、约5.0ng/μL、约6.0ng/μL、约7.0ng/μL、约8.0ng/μL、约9.0ng/μL、约10ng/μL、约15ng/μL、约20ng/μL、约30ng/μL、约40ng/μL、约50ng/μL、约60ng/μL、约70ng/μL、约80ng/μL、约90ng/μL、约100ng/μL、约150ng/μL、约200ng/μL、约300ng/μL、约400ng/μL、约500ng/μL、约600ng/μL、约800ng/μL或约1000ng/μL的DNA区段或靶多核苷酸浓度下进行连接反应。例如，可在约100ng/μL、约150ng/μL、约200ng/μL、约300ng/μL、约400ng/μL或约500ng/μL的DNA区段或靶多核苷酸浓度下进行连接。

[0234] 在一些情况下，可在约0.1至1000ng/μL、约1至1000ng/μL、约1至800ng/μL、约10至800ng/μL、约10至600ng/μL、约100至600ng/μL或约100至500ng/μL的DNA区段或靶多核苷酸浓度下进行连接反应。

[0235] 在一些情况下，连接反应可进行多于约5分钟、约10分钟、约20分钟、约30分钟、约40分钟、约50分钟、约60分钟、约90分钟、约2小时、约3小时、约4小时、约5小时、约6小时、约8小时、约10小时、约12小时、约18小时、约24小时、约36小时、约48小时或约96小时。在其他情况下，连接反应可进行少于约5分钟、约10分钟、约20分钟、约30分钟、约40分钟、约50分钟、约60分钟、约90分钟、约2小时、约3小时、约4小时、约5小时、约6小时、约8小时、约10小时、约12小时、约18小时、约24小时、约36小时、约48小时或约96小时。例如，连接反应可进行约30分钟至约90分钟。在一些实施方案中，衔接子与靶多核苷酸的连接产生连接产物多核苷酸，该连接产物多核苷酸具有包含来源于衔接子的核苷酸序列的3′突出端。

[0236] 在一些实施方案中，在使至少一个衔接子寡核苷酸与靶多核苷酸连接之后，使用一个或多个连接的衔接子寡核苷酸作为模板延伸一个或多个靶多核苷酸的3′端。例如，包

含仅与靶多核苷酸的5′端连接的两个杂交寡核苷酸的衔接子允许在未连接链的置换的同时或之后，使用衔接子的连接链作为模板延伸所述靶标的未连接的3’端。包含两个杂交寡核苷酸的衔接子的两条链可与靶多核苷酸连接，使得连接产物具有5′突出端，并且互补的3′端可以使用5′突出端作为模板进行延伸。作为进一步的实例，发夹衔接子寡核苷酸可与靶多核苷酸的5′端连接。在一些实施方案中，被延伸的靶多核苷酸的3′端包含来自衔接子寡核苷酸的一个或多个核苷酸。对于在两端均与衔接子连接的靶多核苷酸，可对具有5′突出端的双链靶多核苷酸的两个3′端进行延伸。这种3′端延伸或“补平”反应生成了与模板杂交的衔接子寡核苷酸模板的互补序列或“互补体”，从而补平5′突出端以产生双链序列区域。在双链靶多核苷酸的两端均具有通过互补链3′端的延伸而被补平的5′突出端的情况下，该产物完全是双链的。可通过本领域已知的任何合适的聚合酶如DNA聚合酶进行延伸，所述聚合酶中有许多是商购可得的。DNA聚合酶可包括DNA依赖性的DNA聚合酶活性、RNA依赖性的DNA聚合酶活性、或DNA依赖性的和RNA依赖性的DNA聚合酶活性。DNA聚合酶可以是热稳定或非热稳定的。DNA聚合酶的实例包括但不限于Taq聚合酶、Tth聚合酶、Tli聚合酶、Pfu聚合酶、Pfutubo聚合酶、Pyrobest聚合酶、Pwo聚合酶、KOD聚合酶、Bst聚合酶、Sac聚合酶、Sso聚合酶、Poc聚合酶、Pab聚合酶、Mth聚合酶、Pho聚合酶、ES4聚合酶、VENT聚合酶、DEEPVENT聚合酶、EX-Taq聚合酶、LA-Taq聚合酶、Expand聚合酶、Platinum Taq聚合酶、Hi-Fi聚合酶、Tbr聚合酶、Tfl聚合酶、Tru聚合酶、Tac聚合酶、Tne聚合酶、Tma聚合酶、Tih聚合酶、Tfi聚合酶、Klenow片段，及其变体、修饰的产物和衍生物。3′端延伸可在从独立的样品汇集靶多核苷酸之前或之后进行。

[0237] 在某些实施方案中，本公开内容提供了用于富集靶核酸和分析该靶核酸的方法。在一些情况下，用于富集的方法为基于溶液的形式。在一些情况下，靶核酸可采用标记试剂进行标记。在其他情况下，靶核酸可与采用标记试剂标记的一个或多个缔合分子交联。标记试剂的实例包括但不限于生物素、多组氨酸标记物和化学标记物(例如，点击化学法中使用的炔和叠氮化物衍生物)。此外，可捕获标记的靶核酸，从而通过使用捕获剂进行富集。捕获剂可以是链霉亲和素和/或亲和素、抗体、化学部分(例如炔、叠氮化物)，以及本领域已知的用于亲和纯化的任何生物、化学、物理或酶试剂。

[0238] 在一些情况下，固定化或非固定化核酸探针可用于捕获靶核酸。例如，靶核酸可通过与固体支持体上或溶液中的探针杂交而从样品中富集。在一些实例中，样品可以是基因组样品。在一些实例中，探针可以是扩增子。扩增子可包含预定序列。此外，可洗涤和/或从探针洗脱杂交的靶核酸。靶核酸可以是DNA、RNA、cDNA或mRNA分子。

[0239] 在一些情况下，富集方法可包括使包含靶核酸的样品与探针接触以及使靶核酸与固体支持体结合。在一些情况下，可使用化学、物理或酶法使样品片段化以产生靶核酸。在一些情况下，探针可与靶核酸特异性杂交。在一些情况下，靶核酸的平均大小可为约50至5000、约50至2000、约100至2000、约100至1000、约200至1000、约200至800、或约300至800、约300至600或约400至600个核苷酸残基。靶核酸可进一步与样品中未结合的核酸分离。可洗涤和/或洗脱固体支持体以提供富集的靶核酸。在一些实例中，富集步骤可重复约1、2、3、4、5、6、7、8、9或10次。例如，富集步骤可重复约1、2或3次。

[0240] 在一些情况下，富集方法可包括提供探针来源的扩增子，其中用于扩增的探针附接至固体支持体。固体支持体可包含支持体固定化核酸探针以捕获来自样品的特异性靶核酸。

探针来源的扩增子可与靶核酸杂交。在与探针扩增子杂交之后，可通过捕获(例如，经由捕获剂如生物素、抗体等)以及洗涤和/或洗脱来自捕获的探针的杂交的靶核酸来富集样品中的靶核酸(图4)。可以使用例如PCR方法进一步扩增靶核酸序列，以产生富集的PCR产物的扩增库。

[0241] 在一些情况下，固体支持体可以是微阵列、载玻片、芯片、微孔、柱子、管、颗粒或珠子。在一些实例中，固体支持体可包被有链霉亲和素和/或亲和素。在其他实例中，固体支持体可包被有抗体。此外，固体支持体可包括玻璃、金属、陶瓷或聚合物材料。在一些实施方案中，固体支持体可以是核酸微阵列(例如，DNA微阵列)。在其他实施方案中，固体支持体可以是顺磁珠。

[0242] 在一些情况下，富集方法可包括用二级限制酶消化、自连接(例如自循环)和用原始限制酶重新消化。在特定的实例中，只有连接产物将被线性化且可用于衔接子连接和测序。在其他情况下，连接接头序列本身可以用于使用与该接头序列互补的诱饵-探针的基于杂交的富集。

[0243] 在特定的实施方案中，本公开内容提供了用于扩增富集的DNA的方法。在一些情况下，富集的DNA为读取对。可通过本公开内容的方法获得读取对。

[0244] 在一些实施方案中，使用一个或多个扩增和/或复制步骤制备待测序的文库。可使用本领域已知的任何扩增方法。可使用的扩增技术的实例包括但不限于定量PCR、定量荧光PCR(QF-PCR)、多重荧光PCR(MF-PCR)、实时PCR(RTPCR)、单细胞PCR、限制性片段长度多态性PCR(PCR-RFLP)、PCK-RFLPIRT-PCR-IRFLP、热启动PCR、嵌套式PCR、原位聚合酶克隆PCR、原位滚环扩增(RCA)、桥式PCR、连接介导的PCR、Qb复制酶扩增、反向PCR、picotiter PCR和乳液PCR。其他合适的扩增方法包括连接酶链反应(LCR)、转录扩增、自持续序列复制、靶多核苷酸序列的选择性扩增、共有序列引物聚合酶链反应(CP-PCR)、任意引物聚合酶链反应(AP-PCR)、简并寡核苷酸引物PCR(DOP-PCR)和基于核酸的序列扩增(NABSA)。本文可使用的其他扩增方法包括美国专利号5,242,794;5,494,810;4,988,617;和6,582,938中描述的那些方法。

[0245] 在特定的实施方案中，在DNA分子被分配给单独的分区后使用PCR扩增DNA分子。在一些情况下，使用扩增衔接子内的一个或多个特定的引发序列进行PCR扩增。在分配给单独的分区之前或之后，扩增衔接子可与片段化的DNA分子连接。包含在两端具有合适的引发序列的扩增衔接子的多核苷酸可以以指数方式进行PCR扩增。具有仅一个合适的引发序列的多核苷酸由于例如包含引发序列的扩增衔接子的不完全连接效率，可能仅经历线性扩增。此外，如果没有连接包含合适的引发序列的衔接子，则可以从扩增例如PCR扩增中将多核苷酸一起去除。在一些实施方案中，PCR循环数在10-30之间变化，但可低至9、8、7、6、5、4、3、2或更少，或多达40、45、50、55、60或更多。作为结果，在PCR扩增之后，与可线性扩增或不可扩增的片段相比，携带具有合适的引发序列的扩增衔接子的可以以指数方式扩增的片段可以以更高(1000倍或更多)的浓度存在。与全基因组扩增技术(如采用随机化引物的扩增或使用phi29聚合酶的多重置换扩增)相比，PCR的益处包括但不限于更均匀的相对序列覆盖——因为在每个循环中每个片段可以被拷贝至多一次以及因为扩增受热循环程序的控制，形成嵌合分子的比率大幅低于例如MDA(Lasken等人，2007，BMC Biotechnology)——因为嵌合分子通过在组装图中呈现非生物序列而对于准确的序列组装构成重大挑战(这可导致较高的错误组装率或高度模糊且片段化的组装)，与使用具有特异性序列的特异性引发位点相比降低的序列特异性偏差(可能由MDA中常用的随机化引物的结合产生)，最终扩增

的DNA产物的量的较高再现性(这可通过选择PCR循环数来控制),以及与本领域中已知的常见的全基因组扩增技术相比在采用PCR中常用的聚合酶的复制中较高的保真度。

[0246] 在一些实施方案中,补平反应在使用第一引物和第二引物的一个或多个靶多核苷酸的扩增之后或作为所述扩增的一部分进行,其中所述第一引物包含与一个或多个第一衔接子寡核苷酸的互补体的至少一部分可杂交的序列,并且此外其中所述第二引物包含与一个或多个第二衔接子寡核苷酸的互补体的至少一部分可杂交的序列。第一和第二引物中的每一个可以具有任何合适的长度,如约、少于约或多于约10、15、20、25、30、35、40、45、50、55、60、65、70、75、80、90、100个或更多个核苷酸,其任何部分或全部可与相应的靶序列(例如,约、少于约或多于约5、10、15、20、25、30、35、40、45、50个或更多个核苷酸)互补。例如,约10至50个核苷酸可与相应的靶序列互补。

[0247] 在一些情况下,扩增衔接子在文库生成过程中使用。扩增衔接子是共有部分反向互补性的寡聚体对,使得寡聚体对可退火形成具有双链部分和单链部分的分子。通过使用扩增衔接子,能够将分开的退火靶标与文库分子的每一端连接。因为扩增衔接子的单链部分包含非反向互补的序列,所以可获得仅与扩增衔接子的单链臂的其中一个或另一个,或另一个的反向互补体退火的引物。因此,扩增衔接子允许将第一不同的引物结合位点添加至文库分子的第一端,以及将第二不同的引物结合位点添加至文库分子的第二端。

[0248] 适合于生成扩增衔接子的寡核苷酸如下所示(*为硫代磷酸酯键)。寡核苷酸被列为P5/P7对,其中每个P7寡核苷酸被合成为与紧接地在它前面的P5寡核苷酸一起作用。对于每一对,在P5寡核苷酸的硫代磷酸酯键之前的最后十个核苷酸碱基与第二寡核苷酸的/5Phos/之后的前十个碱基反向互补。

[0249] SEQ ID NO位置序列(5’到3’)

[0250] 1 P5_full ACACTCTTTCCCTACACGACGCTCTTCCGATG*T

[0251] 2 P7_rev /5Phos/CATCGGAAGAGCACACGTCTGAACTCCAGTCA*/3ddC/

[0252] 3 P5_full ACACTCTTTCCCTACACGACGCTCTTCCGACC*T

[0253] 4 P7_rev /5Phos/GGTCGGAAGAGCACACGTCTGAACTCCAGTCA*/3ddC/

[0254] 5 P5_full ACACTCTTTCCCTACACGACGCTCTACCGATC*T

[0255] 6 P7_rev /5Phos/GATCGGTAGAGCACACGTCTGAACTCCAGTCA*/3ddC/

[0256] 7 P5_full ACACTCTTTCCCTACACGACGCTATTCCGATC*T

[0257] 8 P7_rev /5Phos/GATCGGAATAGCACACGTCTGAACTCCAGTCA*/3ddC/

[0258] 9 P5_full ACACTCTTTCCCTACACGACGCTCTTCGGATC*T

[0259] 10 P7_rev /5Phos/GATCCGAAGAGCACACGTCTGAACTCCAGTCA*/3ddC/

[0260] 11 P5_full ACACTCTTTCCCTACACGACCCTCTTCCGATC*T

[0261] 12 P7_rev /5Phos/GATCGGAAGAGGACACGTCTGAACTCCAGTCA*/3ddC/

[0262] 13 P5_full ACACTCTTTCCCTACACGACGCACTTCCGATC*T

[0263] 14 P7_rev /5Phos/GATCGGAAGTGCACACGTCTGAACTCCAGTCA*/3ddC/

[0264] 15 P5_full ACACTCTTTCCCTACACGACGCTCTTCCGATC*T

[0265] 16 P7_rev /5Phos/GATCGGAAGAGCACACGTCTGAACTCCAGTCA*/3ddC/

[0266] “扩增”是指增加靶序列的拷贝数的任何过程。在一些情况下,复制反应可产生多核苷酸的仅单个互补拷贝/复制物。用于靶多核苷酸的引物引导的扩增的方法是本领域已

知的，并且包括但不限于基于聚合酶链反应(PCR)的方法。对通过PCR扩增靶序列有利的条件是本领域已知的，可在所述方法的多个步骤中优化，并依赖于反应中元件的特征，如靶标类型、靶标浓度、待扩增的序列长度、靶标和/或一个或多个引物的序列、引物长度、引物浓度、使用的聚合酶、反应体积、一个或多个元件与一个或多个其他元件的比等等，所述元件特征中的一些或全部可以改变。通常，PCR包括以下步骤：使待扩增的靶标变性(如果是双链的话)、使一个或多个引物与靶标杂交以及通过DNA聚合酶延伸引物，其中所述步骤被重复(或“循环”)以扩增靶序列。该过程中的步骤可以针对各种结果进行优化，如以提高产率、减少假产物的形成和/或增加或降低引物退火的特异性。优化方法在本领域中是公知的，并且包括对扩增反应中的元件的类型或量的调整和/或对该过程中给定步骤的条件(如特定步骤中的温度、特定步骤的持续时间和/或循环数)的调整。

[0267] 在一些实施方案中，扩增反应可包含至少约5、10、15、20、25、30、35、40、50、60、70、80、90、100、150、200个或更多个循环。在一些实例中，扩增反应可包含至少约20、25、30、35或40个循环。在一些实施方案中，扩增反应包含不多于约5、10、15、20、25、35、40、50、60、70、80、90、100、150、200个或更多个循环。循环可含有任何数目的步骤，如1、2、3、4、5、6、7、8、9、10个或更多个步骤。步骤可以包括适合于达到给定步骤的目的的任何温度或温度梯度，所述给定步骤的目的包括但不限于3′端延伸(例如衔接子补平)、引物退火、引物延伸和链变性。步骤可具有任何持续时间，包括但不限于约、少于约或多于约1、5、10、15、20、25、30、35、40、45、50、55、60、70、80、90、100、120、180、240、300、360、420、480、540、600、1200、1800秒或更长，包括无限期直到手动中断。包含不同步骤的任何数目的循环可以以任何顺序进行组合。在一些实施方案中，包含不同步骤的不同循环进行组合，使得组合中的总循环数为约、少于约或多于约5、10、15、20、25、30、35、40、50、60、70、80、90、100、150、200个或更多个循环。在一些实施方案中，扩增在补平反应之后进行。

[0268] 在一些实施方案中，可对至少约1、2、3、4、5、6、7、8、9、10、12、14、16、18、20、25、30、40、50、100、200、300、400、500、600、800、1000ng的靶DNA分子进行扩增反应。在其他实施方案中，可对少于约1、2、3、4、5、6、7、8、9、10、12、14、16、18、20、25、30、40、50、100、200、300、400、500、600、800、1000ng的靶DNA分子进行扩增反应。

[0269] 扩增可在从独立的样品汇集靶多核苷酸之前或之后进行。

[0270] 本公开内容的方法包括测定样品中存在的可扩增的核酸的量。可使用任何已知的方法定量可扩增的核酸，并且示例性方法为聚合酶链反应(PCR)，特别是定量聚合酶链反应(qPCR)。qPCR是基于聚合酶链反应的技术，并且用于扩增且同时定量靶向的核酸分子。qPCR允许检测和定量(当相对于DNA输入或另外的归一化基因进行归一化时，以绝对拷贝数或相对量的形式)DNA样品中的特异性序列。该过程遵循聚合酶链反应的一般原理，具有附加的特征，即在每个扩增循环后扩增的DNA随着它在反应中的积累进行实时定量。QPCR描述于例如Kurnit等人(美国专利号6,033,854)、Wang等人(美国专利号5,567,583和5,348,853)、Ma等人(The Journal of American Science,2(3),2006)、Heid等人(Genome Research 986-994,1996)、Sambrook和Russell(Quantitative PCR,Cold Spring Harbor Protocols,2006)以及Higuchi(美国专利号6,171,785和5,994,056)中。这些文献的内容通过引用以其全文并入本文。

[0271] 其他的定量方法包括使用插入双链DNA的荧光染料，以及当与互补DNA杂交时发荧

光的修饰的DNA寡核苷酸探针。这些方法可广泛使用,但也特别适用于作为实例进一步详细描述的实时PCR。在第一种方法中,在PCR中DNA结合染料与全部双链(ds)DNA结合,从而产生染料的荧光。因此,PCR过程中DNA产物的增加导致荧光强度的增加,并且在每个循环中测量荧光强度,从而允许对DNA浓度进行定量。与标准PCR反应类似地准备所述反应,其中加入荧光(ds)DNA染料。反应在热循环仪中运行,并且在每个循环之后,用检测器测量荧光水平;所述染料仅在与(ds)DNA(即,PCR产物)结合时发荧光。参照标准稀释,可以测定PCR中的(ds)DNA浓度。类似于其他实时PCR方法,所获得的值不具有与之相关联的绝对单位。测量的DNA/RNA样品与标准稀释的比较给出样品相对于该标准的分数或比率,从而允许不同组织或实验条件之间的相对比较。为了确保靶基因的定量和/或表达的准确性,可以关于稳定表达的基因进行归一化。类似地,拷贝数未知的基因可相对于已知拷贝数的基因进行归一化。

[0272] 第二种方法使用基于序列特异性RNA或DNA的探针仅对含有探针序列的DNA进行定量;因此,报道分子探针的使用显著增加了特异性,并且即使在一些非特异性DNA扩增的存在下也允许定量。这允许多路化,即,通过使用具有不同颜色标记物的特定探针在同一反应中测定多个基因,条件是所有基因以相似的效率进行扩增。

[0273] 这种方法通常采用基于DNA的探针进行,该探针的一端具有荧光报道分子(例如6-羧基荧光素),而在相对端具有荧光猝灭剂(例如,6-羧基-四甲基罗丹明)。报道分子与猝灭剂的密切接近妨碍其荧光的检测。通过聚合酶(例如,Taq聚合酶)的5'到3'外切核酸酶活性的探针分解破坏了报道分子-猝灭剂接近,从而允许荧光的未猝灭发射(可被检测到)。在每个PCR循环中,由报道分子探针靶向的产物的增加导致荧光的成比例增加,这是由于探针的分解和报道分子的释放。与标准PCR反应类似地准备所述反应,并加入报道分子探针。当反应开始时,在PCR的退火阶段,探针和引物均与DNA靶标退火。新的DNA链的聚合从引物启动,并且一旦聚合酶到达探针,其5'-3'-外切核酸酶便使探针降解,将荧光报道分子与猝灭剂物理隔开,从而导致荧光的增加。在实时PCR热循环仪中检测并测量荧光,并使用对应于产物的指数式增长的荧光的几何增长来确定每个反应中的循环阈值。

[0274] 通过在对数尺度上绘制荧光与循环数的图(因此以指数方式增加的量将得到一条直线)来测定在反应的指数阶段中存在的DNA的相对浓度。确定大于背景的荧光的检测阈值。来自样品的荧光跨越阈值时的循环被称为循环阈值,C_t。因为在指数阶段DNA的量在每个循环都加倍,所以可计算DNA的相对量,例如,C_t比另一个样品早3个循环的样品具有比另一个样品多$2^3=8$倍的模板。随后通过将结果与由已知量的核酸的系列稀释(例如,未稀释的、1:4、1:16、1:64)的实时PCR产生的标准曲线进行比较,来确定核酸(例如,RNA或DNA)的量。

[0275] 在某些实施方案中,qPCR反应涉及双荧光团方法,该方法利用荧光共振能量转移(FRET),例如LIGHTCYCLER杂交探针,其中两个寡核苷酸探针与扩增子退火(例如参见美国专利号6,174,670)。寡核苷酸被设计为在头到尾的方向上杂交,其中荧光团以与有效能量转移兼容的距离隔开。被结构化为当与核酸结合或并入延伸产物中时发出信号的标记的寡核苷酸的其他实例包括:SCORPIONS探针(例如,Whitcombe等人,Nature Biotechnology 17:804-807,1999,和美国专利号6,326,145)、Sunrise(或AMPLIFLOUR)引物(例如,Nazarenko等人,Nuc.Acids Res.25:2516-2521,1997,和美国专利号6,117,635)以及LUX引物和MOLECULAR BEACONS探针(例如,Tyagi等人,Nature Biotechnology14:303-308,1996,

和美国专利号5,989,823)。

[0276] 在其他实施方案中,qPCR反应使用荧光Taqman方法和能够实时测量荧光的仪器(例如,ABI Prism 7700序列检测器)。Taqman反应使用采用两种不同荧光染料标记的杂交探针。一种染料是报道分子染料(6-羧基荧光素),另一种是猝灭染料(6-羧基-四甲基罗丹明)。当探针完整时,发生荧光能量转移,并且报道分子染料荧光发射被猝灭染料吸收。在PCR循环的延伸阶段中,荧光杂交探针通过DNA聚合酶的5'-3'溶核活性被裂解。在探针裂解时,报道分子染料发射不再有效地转移至猝灭染料,从而导致报道分子染料荧光发射光谱的增加。任何核酸定量方法,包括实时方法或单点检测方法,均可用于定量样品中的核酸的量。可通过若干种不同的方法(例如,染色、与标记的探针杂交;并入生物素化的引物,然后进行亲和素-酶缀合物检测;将32P标记的脱氧核苷酸三磷酸,如dCTP或dATP并入扩增的区段),以及用于核酸定量的本领域已知的任何其他合适的检测方法进行检测。定量可包括或可不包括扩增步骤。

[0277] 在一些实施方案中,本公开内容提供了用于鉴定或定量连接的DNA区段的标记物。在一些情况下,可标记连接的DNA区段以有助于下游应用,如阵列杂交。例如,可使用随机引发或切口平移标记连接的DNA区段。

[0278] 可以使用各种标记物(例如报道分子)来标记本文所述的核苷酸序列,包括但不限于在扩增步骤中。合适的标记物包括放射性核素、酶、荧光剂、化学发光剂或显色剂以及配体、辅因子、抑制剂、磁性粒子等。这类标记物的实例包括在美国专利号3,817,837;美国专利号3,850,752;美国专利号3,939,350;美国专利号3,996,345;美国专利号4,277,437;美国专利号4,275,149和美国专利号4,366,241中,这些专利通过引用以其全文并入本文。

[0279] 附加的标记物包括但不限于β-半乳糖苷酶、转化酶、绿色荧光蛋白、荧光素酶、氯霉素、乙酰转移酶、β-葡萄糖醛酸酶、外切葡聚糖酶和葡糖淀粉酶。还可以使用荧光标记物,以及采用特定化学性质特别合成的荧光试剂。可使用各种测量荧光的方法。例如,一些荧光标记物显示出激发或发射光谱的变化,一些显示出共振能量转移,其中一个荧光报道分子失去荧光,而第二个荧光报道分子的荧光增加,一些显示出荧光的损失(猝灭)或出现,而一些报告旋转运动。

[0280] 此外,为了获得足够的标记材料,可以汇集多个扩增,而不是增加每个反应的扩增循环数。或者,可将标记的核苷酸并入最后的扩增反应循环,例如,30个PCR循环(无标记物)+10个PCR循环(加上标记物)。

[0281] 在特定的实施方案中,本公开内容提供了可附接至连接的DNA区段的探针。如本文所用的术语"探针"是指这样一种分子(例如,寡核苷酸,无论是在纯化的限制性消化物中天然存在的,还是通过合成、重组或通过PCR扩增产生的),该分子能够与另一个感兴趣的分子(例如,另一个寡核苷酸)杂交。当探针为寡核苷酸时,其可能是单链或双链的。探针用于特定靶标(例如,基因序列)的检测、鉴定和分离。在一些情况下,探针可以与标记物关联,使得探针可在任何检测系统中被检测到,该检测系统包括但不限于酶(例如,ELISA,以及基于酶的组织化学分析)、荧光、放射性和发光系统。

[0282] 关于阵列和微阵列,术语"探针"用于指固定于阵列的任何可杂交材料,该阵列用于检测已经与探针杂交的核苷酸序列。在一些情况下,探针可以为约10bp至500bp、约10bp至250bp、约20bp至250bp、约20bp至200bp、约25bp至200bp、约25bp至100bp、约30bp至100bp

或约30bp至80bp。在一些情况下，探针的长度可以大于约10bp、约20bp、约30bp、约40bp、约50bp、约60bp、约70bp、约80bp、约90bp、约100bp、约150bp、约200bp、约250bp、约300bp、约400bp或约500bp。例如，探针的长度可以为约20至约50bp。可在WO95/11995、EP 717,113和WO97/29212中找到探针设计的实例和基本原理。

[0283] 在一些情况下，可以设计一个或多个探针，使得它们可与通过限制酶消化的位点密切杂交。例如，探针可以在限制酶识别位点的约10bp、约20bp、约30bp、约40bp、约50bp、约60bp、约70bp、约80bp、约90bp、约100bp、约150bp、约200bp、约250bp、约300bp、约400bp或约500bp内。

[0284] 在其他情况下，单个独特的探针可被设计为在通过限制酶消化的位点的每一侧约10bp、约20bp、约30bp、约40bp、约50bp、约60bp、约70bp、约80bp、约90bp、约100bp、约150bp、约200bp、约250bp、约300bp、约400bp或约500bp内。可以设计探针，使得它们可在通过限制酶消化的位点的任一侧杂交。例如，可使用在主要限制酶识别位点的每一侧的单个探针。

[0285] 在一些情况下，可在限制酶识别位点中的每一侧设计2、3、4、5、6、7、8个或更多个探针，这些探针可随后用于研究相同的连接事件。例如，可在限制酶识别位点的每一侧设计2或3个探针。在一些实例中，在每个主要限制酶识别位点的多个(例如，2、3、4、5、6、7或8个或更多个)探针的使用可用于使从单独探针获得假阴性结果的问题最小化。

[0286] 如本文所用的术语“探针集”是指可与基因组中的主要限制酶的一个或多个主要限制酶识别位点杂交的探针的组或集合。

[0287] 在一些情况下，探针集可在序列上与核酸序列互补，所述核酸序列与基因组DNA中的限制酶的一个或多个主要限制酶识别位点相邻。例如，探针集可在序列上与约10bp至500bp、约10bp至250bp、约20bp至250bp、约20bp至200bp、约25bp至200bp、约25bp至100bp、约30bp至100bp或约30bp至80bp的核苷酸互补，所述核苷酸与基因组DNA中的一个或多个限制酶识别位点相邻。探针集可在序列上与限制酶识别位点的一(例如，任一)侧或两侧互补。因此，探针可在序列上与核酸序列互补，所述核酸序列与基因组DNA中的一个或多个主要限制酶识别位点的每一侧相邻。此外，探针集可在序列上与核酸序列互补，所述核酸序列小于来自基因组DNA中的一个或多个主要限制酶识别位点的约10bp、约20bp、约30bp、约40bp、约50bp、约60bp、约70bp、约80bp、约90bp、约100bp、约150bp、约200bp、约250bp、约300bp、约400bp或约500bp。

[0288] 在一些情况下，两个或更多个探针可被设计为能够与序列杂交，所述序列与基因组DNA中的一个或多个限制酶识别位点相邻。探针可重叠或部分重叠。

[0289] 探针、探针的阵列或探针集可固定在支持体上。支持体(例如固体支持体)可由多种材料(如玻璃、二氧化硅、塑料、尼龙或硝化纤维素)制成。支持体优选刚性的且具有平坦表面。支持体可具有约1至10,000,000个分辨位点。例如，支持体可具有约10至10,000,000、约10至5,000,000、约100至5,000,000、约100至4,000,000、约1000至4,000,000、约1000至3,000,000、约10,000至3,000,000、约10,000至2,000,000、约100,000至2,000,000或约100,000至1,000,000个分辨位点。分辨位点的密度可以为平方厘米内至少约10、约100、约1000、约10,000、约100,000或约1,000,000个分辨位点。在一些情况下，每个分辨位点可被>95％的单一类型的寡核苷酸占据。在其他情况下，每个分辨位点可被汇集的探针混合物或探针集占据。在一些情况下，一些分辨位点被汇集的探针混合物或探针集占据，并且其他分

辨位点被>95％的单一类型的寡核苷酸占据。

[0290] 在一些情况下，阵列上针对给定核苷酸序列的探针的数目可以大大超过有待于与这种阵列杂交的DNA样品。例如，相对于输入样品中DNA的量，阵列可具有约10、约100、约1000、约10,000、约100,000、约1,000,000、约10,000,000或约100,000,000倍的探针数。

[0291] 在一些情况下，阵列可具有约10、约100、约1000、约10,000、约100,000、约1,000,000、约10,000,000、约100,000,000或约1,000,000,000个探针。

[0292] 探针的阵列或探针集可以以逐步的方式在支持体上合成，或可以以预合成的形式进行附接。一种合成方法为VLSIPS™(如美国专利号5,143,854和EP 476,014中所描述的)，这需要使用光来指导在高密度、小型化阵列中的寡核苷酸探针的合成。在美国专利号5,571,639和美国专利号5,593,839中描述了用于减少合成循环数的掩模设计的算法。阵列还可以通过将单体经由机械限制的流道输送至支持体的单元来合成，如EP 624,059中所述。阵列还可通过使用喷墨式打印机将试剂点样到支持体上来合成(参见，例如，EP 728,520)。

[0293] 在一些实施方案中，本公开内容提供了用于使连接的DNA区段杂交至阵列上的方法。“基底”或“阵列”是一种有意创建的核酸集合，所述核酸可通过合成或生物合成来制备，并针对多种不同的形式的生物活性进行筛选(例如，可溶性分子的文库；和拴系至树脂珠子的寡核苷酸的文库、二氧化硅芯片或其他固体支持体)。此外，术语“阵列”包括核酸的文库，所述核酸的文库可通过将基本任何长度的核酸(例如，长度为1至约1000个核苷酸单体)点样到基底上进行制备。

[0294] 阵列技术以及各种相关技术和应用一般描述于许多教科书和文献中。例如，这些教科书和文献包括Lemieux等人，1998,Molecular Breeding4,277-289;Schena和Davis,Parallel Analysis with Biological Chips.在PCR Methods Manual(M.Innis,D.Gelfand,J.Sninsky编著)中;Schena和Davis,1999,Genes,Genomes and Chips.在DNA Microarrays:A Practical Approach(M.Schena编著),Oxford University Press,Oxford,UK,1999中;The Chipping Forecast(Nature Genetics special issue;1999年1月增刊);Mark Schena(编著),Microarray Biochip Technology,(Eaton Publishing Company);Cortes,2000,The Scientist 14[17]:25;Gwynn和Page,Microarray analysis:the next revolution in molecular biology,Science,1999年8月6日;以及Eakins和Chu,1999,Trends in Biotechnology,17,217-218。

[0295] 通常，通过空间隔开文库的成员，任何文库都可以以有序的方式排列成阵列。适合排列的文库的实例包括核酸文库(包括DNA、cDNA、寡核苷酸等文库)、肽、多肽和蛋白质文库，以及包含任何分子的文库，如配体文库等。

[0296] 文库可固定或固定化在固相(例如，固体基底)上，以限制成员的扩散和混合。在一些情况下，可以制备DNA结合配体的文库。特别地，文库可固定至基本平坦的固相，包括膜和无孔基底，如塑料和玻璃。此外，文库可以以便于索引(即，参考或访问特定成员)的这样一种方式进行排列。在一些实例中，文库的成员可作为网格形成中的点来应用。常见的测定系统可适用于该目的。例如，阵列可以固定在微板的表面上，所述微板在孔中具有多个成员，或者在每个孔中具有单个成员。此外，固体基底可以是膜，如硝化纤维素或尼龙膜(例如，印迹实验中使用的膜)。备选的基底包括基于玻璃或二氧化硅的基底。因此，可通过本领域已知的任何合适的方法使文库固定化，例如，通过电荷相互作用，或通过化学偶合至孔的壁或

底部或膜的表面。可以使用其他的排列和固定方法，例如移液、滴触、压电方式、喷墨和气泡喷射技术、静电应用等。在硅基芯片的情况下，可以利用光刻法来排列文库并将文库固定在芯片上。

[0297] 可通过“点样”至固体基底上来排列文库；这可手动进行或通过利用机器人放置成员来进行。通常，阵列可描述为宏阵列或微阵列，差异在于点的大小。宏阵列可含有大小为约300微米或更大的点，并且可容易地通过现有的凝胶和印迹扫描仪成像。微阵列中的点大小可为小于200微米的直径，并且这些阵列通常含有数千个点。因此，微阵列可能需要专业的机器人和成像设备，所述机器人和成像设备可能需要定制。仪器装置一般描述在Cortese,2000,The Scientist 14[11]:26的综述中。

[0298] 本领域已描述了用于产生DNA分子的固定化文库的技术。一般来说，大多数现有技术方法描述了如何合成单链核酸分子文库，例如使用掩蔽技术在固体基底上的不同离散位置建立序列的各种排列。美国专利号5,837,832描述了基于超大规模集成技术产生固定至硅基底的DNA阵列的改进方法。特别地，美国专利号5,837,832描述了在基底上的空间限定位置处合成特定探针集的称为“平铺(tiling)”的策略，其可用于产生本公开内容的固定化DNA文库。美国专利号5,837,832还提供了关于也可能使用的早期技术的参考。在其他情况下，也可使用光沉积化学构建阵列。

[0299] 还可以以将每个不同的文库成员(例如，独特的肽序列)放置在阵列中离散的、预定的位置的方式，在表面上合成肽(或模拟肽)的阵列。每个文库成员的同一性通过它在阵列中的空间位置来确定。确定阵列中预定分子(例如，靶标或探针)与反应性文库成员之间发生结合相互作用的位置，从而基于空间位置鉴定反应性文库成员的序列。这些方法描述于美国专利号5,143,854;WO90/15070和WO92/10092;Fodor等人.(1991)Science,251:767;Dower和Fodor(1991)Ann.Rep.Med.Chem.,26:271中。

[0300] 为了帮助检测，可使用标记物(如上所讨论的)——如任何容易检测的报道分子，例如，荧光、生物发光、发出磷光、放射性的报道分子等。在本文的其他地方讨论了这样的报道分子、其检测、与靶标/探针的偶合。探针和靶标的标记还公开于Shalon等人,1996,Genome Res6(7):639-45中。

[0301] 一些商购可得的微阵列形式的实例列于下表1中(还参见Marshall和Hodgson,1998,Nature Biotechnology,16(1),27-31)。

[0302] 表1

[0303]

目前可获得的杂交微阵列形式的实例

公司	产品名称	排列方法	杂交步骤	读出
Affymetrix, Inc., Santa Clara, California	GeneChip®	原位（在芯片上）光刻合成约 20-25-聚体寡核苷酸至被切成 1.25 cm^2 或 5.25 cm^2 芯片的硅片上	用样品 cDNA 或反义 RNA 的标记的 30-40 个核苷酸的片段探测 10,000-260,000 个寡核苷酸特征	荧光
Brax, Cambridge, UK		短合成寡核苷酸，在芯片外合成	用标记的核酸探测“通用芯片”上的 1000 个寡核苷酸	质谱分析法
Gene Logic, Inc., Columbia, Maryland	READS ™			
Genometrix Inc., The	Universal Arrays ™			

	Woodlands, Texas			
	GENSET, Paris, France			
[0304]	Hyseq Inc., Sunnyvale, California	HyChip ™	将 500-2000 nt DNA 样品打印在 0.6 cm^2（HyGnostics）或约 18 cm^2（Gene Discovery）膜上	用 8,000 个 7-聚体寡核苷酸（HyGnostics）探测 64 个样品 cDNA 点，或用 300 个 7-聚体寡核苷酸（Gene Discovery）探测＜＝55,000 个样品 cDNA 点	放射性同位素
			制成的 5-聚体寡核苷酸以 1.15 cm^2 阵列打印到玻璃（HyChip）上	用 10 kb 样品 cDNA、标记的 5-聚体寡核苷酸和连接酶探测通用的 1024 个寡核苷酸点	荧光

[0305] 表1-续

目前可获得的杂交微阵列形式的实例

公司	产品名称	排列方法	杂交步骤	读出
[0306] Incyte Pharmaceuticals, Inc., Palo Alto, California	GEM	压电打印以将 PCR 片段点样以及在芯片上合成寡核苷酸	用标记的 RNA 探测＜=1000（最终为 10,000）个寡核	荧光和放射性同位素

			苷酸/PCR 片段点	
Molecular Dynamics, Inc., Sunnyvale, California	Storm® FluorImager®	通过笔将 500-5000 nt cDNA 打印到载玻片上的约 10 cm^2	用 200-400 nt 标记的样品 cDNA 探测约 10,000 个 cDNA 点	荧光
Nanogen, San Diego, California	半导体微芯片（Semiconductor Microchip）	预先制成约 20-聚体的寡核苷酸，捕获到被切成＜=1 cm^2 芯片的硅片的电活性点上	使 25、64、400（以及最终 10,000）个寡核苷酸点极化以增强与 200-400 nt 标记的样品 cDNA 的杂交	荧光
Protogene Laboratories, Palo Alto, California		经由打印至表面张力阵列，通过芯片上合成将 40-50-聚体寡核苷酸合成到 9 cm^2 的玻璃芯片上	用 200-400 nt 标记的样品核酸探测＜=8,000 个寡核苷酸点	荧光
Sequenom, Hamburg, Germany 和 San Diego, California	MassArray SpectroChip	胶印阵列；约 20-25-聚体寡核苷酸	通过激光解吸和质谱分析法询问每个 SpectroChip 的 250 个位置	质谱分析法
Synteni, Inc., Fremont, California	UniGEM ™	通过覆盖在约 4 cm^2 玻璃芯片上打印 500-5,000 nt cDNA	用 200-400 nt 标记的样品 cDNA 探测＜=10,000 个	荧光

[0307]

[0308]

			cDNA 点	
Nimblegen Systems Inc., Madison	智人全基因组 60-聚体微阵列	每个基因 38,000 个转录物（具有 5 个探针），17.4 mm × 13 mm		5 微米扫描平台
The German Cancer Institute, Heidelberg, Germany		使用 f-moc 或 t-moc 化学通过芯片上合成探针来合成原型 PNA 宏芯片（macrochip）	在 8×12 cm 芯片上约 1,000 个点	荧光/质谱分析法

[0309] 为了从基于阵列的测定生成数据，可以检测信号来表示探针与核苷酸序列之间杂交的存在或不存在。此外，还可以使用直接和间接的标记技术。例如，直接标记将荧光染料直接并入与阵列相关探针杂交的核苷酸序列中（例如，在标记的核苷酸或PCR引物的存在下，通过酶合成将染料并入核苷酸序列中）。直接标记方案可例如通过使用具有类似的化学结构和特征的荧光染料家族来产生强烈的杂交信号，并且可简单地实施。在包括直接标记核酸的情况下，可使用花青或alexa类似物进行多氟比较阵列分析。在其他实施方案中，可使用间接标记方案在核酸与微阵列探针杂交之前或之后将表位并入核酸中。一个或多个染色程序和试剂可用于标记杂交的复合体（例如，与表位结合的荧光分子，从而通过染料分子与杂交种类的表位缀合来提供荧光信号）。

[0310] 在各个实施方案中，本文描述的或本领域中已知的合适的测序方法将用于从样品内的核酸分子获得序列信息。可通过本领域公知的经典桑格测序法来实现测序。也可以使用高通量系统实现测序，其中一些高通量系统允许在经测序的核苷酸并入增长的链中之后立即或在并入之时检测经测序的核苷酸，即实时或基本实时地检测序列。在一些情况下，高通量测序每小时生成至少1,000、至少5,000、至少10,000、至少20,000、至少30,000、至少40,000、至少50,000、至少100,000或至少500,000个序列读取；其中所述测序读取的每个读取可以为至少约50、约60、约70、约80、约90、约100、约120、约150、约180、约210、约240、约270、约300、约350、约400、约450、约500、约600、约700、约800、约900或约1000个碱基。

[0311] 在一些实施方案中，高通量测序包括使用可通过Illumina的基因组分析仪IIX、MiSeq个人测序仪或HiSeq系统，如使用HiSeq 2500、HiSeq1500、HiSeq 2000或HiSeq 1000机器的那些HiSeq系统获得的技术。这些机器使用基于可逆终止子的合成化学测序。这些机器可在八天内进行2000亿个或更多的DNA读取。较小的系统可用于在3、2、1天或更短的时间内的运行。

[0312] 在一些实施方案中，高通量测序包括使用可通过ABI Solid系统获得的技术。该遗传分析平台实现了与珠子连接的克隆扩增的DNA片段的大规模平行测序。该测序方法基于与染料标记的寡核苷酸的顺序连接。

[0313] 下一代测序可包括离子半导体测序（例如，使用来自Life Technologies（Ion

Torrent)的技术)。离子半导体测序可以利用这样一个事实,即当核苷酸并入DNA链中时,离子可被释放。为了进行离子半导体测序,可形成微型机械化孔的高密度阵列。每个孔可容纳单个DNA模板。孔的下面可以是离子敏感层,而离子敏感层下面可以是离子传感器。当向DNA添加核苷酸时,H+可被释放,其可以被测量为pH的变化。H+离子可以被转换成电压并由半导体传感器记录。一个接一个的核苷酸可顺序地涌入阵列芯片。可不需要扫描、光或摄像机。在一些情况下,使用IONPROTON™测序仪对核酸进行测序。在一些情况下,使用IONPGM™测序仪。Ion Torrent Personal Genome Machine(PGM)。PGM可在两小时内进行1000万个读取。

[0314] 在一些实施方案中,高通量测序包括使用可通过Helicos BioSciences Corporation(Cambridge,Massachusetts)获得的技术,如单分子合成测序(SMSS)方法。SMSS是独特的,因为它允许在长达24小时内对整个人类基因组进行测序。最终,SMSS部分地描述于美国公开申请号20060024711;20060024678;20060012793;20060012784;和20050100932中。

[0315] 在一些实施方案中,高通量测序包括使用可通过454Lifesciences,Inc.(Branford,Connecticut)获得的技术,如包括光纤板的PicoTiterPlate装置,所述光纤板传输将通过仪器中的CCD摄像机记录的由测序反应生成的化学发光信号。光纤的这种使用允许在4.5小时内检测最少2000万个碱基对。

[0316] 使用珠子扩增并随后进行光纤检测的方法描述于Marguiles,M等人."Genome sequencing in microfabricated high-density picolitre reactors",Nature,doi:10.1038/nature03959;以及美国公开申请号20020012930;20030068629;20030100102;20030148344;20040248161;20050079510、20050124022;以及20060078909中。

[0317] 在一些实施方案中,使用克隆单分子阵列(Solexa,Inc.)进行高通量测序,或利用可逆终止子化学进行合成测序(SBS)。这些技术部分地描述于美国专利号6,969,488;6,897,023;6,833,246;6,787,308;和美国公开申请号20040106110;20030064398;20030022207;以及Constans,A.,The Scientist 2003,17(13):36中。

[0318] 下一代测序技术可包括Pacific Biosciences的实时(SMRT™)技术。在SMRT中,四种DNA碱基中的每一种均可附接至四种不同荧光染料中的一种。这些染料可以被磷酸连接。可以用单分子的模板单链DNA将单个DNA聚合酶固定在零模式波导(ZMW)的底部。ZMW可以是限制结构,该限制结构使得能够观察单个核苷酸以荧光核苷酸作为背景通过DNA聚合酶的并入,所述单个核苷酸可在ZMW之外迅速扩散(在微秒内)。将核苷酸并入增长的链中可能需要几毫秒。在这个时间内,荧光标记物可被激发并产生荧光信号,并且荧光标记物可被裂解。可如下阐明ZMW。来自激发光束的衰减光可以穿透每个ZMW的下部20-30nm。可以创建检测限为20仄升(zepto liters)(10″升)的显微镜。微小的检测体积可以在降低背景噪声方面提供1000倍的改善。染料的相应荧光的检测可以指示哪个碱基被并入。该过程可以重复。

[0319] 在一些情况下,下一代测序为纳米孔测序(参见,例如,Soni GV和Meller A.(2007)Clin Chem 53:1996-2001)。纳米孔可以是小孔,直径为约一纳米的量级。将纳米孔浸入导电流体中并在导电流体上施加电势可导致由于通过纳米孔的离子传导而产生轻微电流。流动的电流量可对纳米孔的大小敏感。当DNA分子穿过纳米孔时,DNA分子上的每个核苷酸可能在不同程度上堵塞纳米孔。因此,当DNA分子穿过纳米孔时,穿过纳米孔的电流的变化可以代表DNA序列的读数。纳米孔测序技术可以来自Oxford Nanopore Technologies,

例如GridlON系统。单个纳米孔可穿过微孔的顶部插入聚合物膜中。每个微孔可具有用于个体感测的电极。微孔可制成阵列芯片，其中每个芯片具有100,000个或更多个微孔(例如，大于200,000、300,000、400,000、500,000、600,000、700,000、800,000、900,000或1,000,000个)。可使用仪器(或节点)分析芯片。可实时分析数据。可同时操作一个或多个仪器。纳米孔可以是蛋白质纳米孔，例如，蛋白质α-溶血素、七聚体蛋白质孔。纳米孔可以是制备的固态纳米孔，例如，在合成膜(例如，SiN_x或SiO_2)中形成的纳米大小的孔。纳米孔可以是混合孔(例如，蛋白质孔集成到固态膜中)。纳米孔可以是具有集成传感器(例如，隧道电极探测器、电容探测器或基于石墨烯的纳米间隙或边缘状态探测器(参见例如，Garaj等人.(2010) Nature vol.67,doi:10.1038/nature09379))的纳米孔。纳米孔可被功能化用于分析特定类型的分子(例如，DNA、RNA或蛋白质)。纳米孔测序可包括“链测序”，其中完整的DNA聚合物可穿过蛋白质纳米孔，其中随着DNA使孔易位进行实时测序。酶可以分离双链DNA的链，并通过纳米孔供给链。DNA可在一端具有发夹，并且系统可读取两条链。在一些情况下，纳米孔测序为“外切核酸酶测序”，其中单独的核苷酸可通过持续的外切核酸酶从DNA链裂解，并且核苷酸可穿过蛋白质纳米孔。核苷酸可与孔中的分子(例如，环糊精)瞬时结合。电流的特征性中断可用于鉴定碱基。

[0320] 可以使用来自GENIA的纳米孔测序技术。可将工程化蛋白质孔嵌入脂质双层膜中。“主动控制”技术可用于实现有效的纳米孔-膜组装以及对DNA通过通道的移动的控制。在一些情况下，纳米孔测序技术来自NABsys。基因组DNA可以片段化成平均长度约100kb的链。100kb片段可制成单链，并随后与6-聚体探针杂交。具有探针的基因组片段可被驱动通过纳米孔，这可产生电流与时间描记图。电流描记图可提供每个基因组片段上探针的位置。基因组片段可排成一行以产生基因组的探针示意图。该过程可针对探针的文库平行进行。可以生成针对每个探针的基因组长度探针示意图。可采用被称为“移动窗口杂交测序(moving window Sequencing By Hybridization,mwSBH)”的方法校正误差。在一些情况下，纳米孔测序技术来自IBM/Roche。电子束可用于在微芯片中形成纳米孔大小的开口。可利用电场通过纳米孔将DNA拉出或穿出。纳米孔中的DNA晶体管装置可包含交替的纳米大小的金属和电介质层。DNA骨架中的离散电荷可以被DNA纳米孔内的电场捕获。关闭和打开栅压可以允许读取DNA序列。

[0321] 下一代测序可包括DNA纳米球测序(如例如通过Complete Genomics进行的;参见例如，Drmanac等人.(2010)Science 327:78-81)。可对DNA进行分离、片段化和大小选择。例如，可将DNA片段化(例如，通过声处理)成约500bp的平均长度。衔接子(Adl)可附接至片段的末端。衔接子可用于与锚形体杂交以用于测序反应。每一端结合衔接子的DNA可进行PCR扩增。可对衔接子序列进行修饰，使得互补的单链末端彼此结合，从而形成环状DNA。可将DNA甲基化以保护其免受后续步骤中使用的IIS类限制酶的裂解。衔接子(例如，右侧衔接子)可具有限制性识别位点，并且该限制性识别位点可以保持非甲基化。衔接子中的非甲基化限制性识别位点可以由限制酶(例如，Acul)识别，并且DNA可以由Acul裂解成13bp到右侧衔接子的右侧，以形成线性双链DNA。第二轮右侧和左侧衔接子(Ad2)可连接至线性DNA的任一端上，并且与两个衔接子结合的全部DNA均可进行PCR扩增(例如，通过PCR)。可以修饰Ad2序列以使它们彼此结合并形成环状DNA。可将DNA甲基化，但限制酶识别位点可以在左侧Adl衔接子上保持非甲基化。可以应用限制酶(例如，Acul)，并且可以将DNA裂解成13bp到Adl的

左侧以形成线性DNA片段。第三轮右侧和左侧衔接子(Ad3)可连接至线性DNA的右侧和左侧，并且得到的片段可进行PCR扩增。可以修饰衔接子，使得它们可以彼此结合并形成环状DNA。可添加III类限制酶(例如，EcoP15)；EcoP15可将DNA裂解为26bp到Ad3的左侧，以及裂解为26bp到Ad2的右侧。该裂解可去除DNA的大区段，并使DNA再次线性化。可以将第四轮左侧和右侧衔接子(Ad4)连接至DNA，可以对DNA进行扩增(例如，通过PCR)，并进行修饰，使它们彼此结合并形成完整的环状DNA模板。

[0322] 滚环复制(例如，使用Phi 29DNA聚合酶)可用于扩增DNA的小片段。四个衔接子序列可以包含可杂交的回文序列，并且单链可以折叠到其自身上以形成DNA纳米球(DNB™)，所述DNA纳米球的平均直径可为约200-300纳米。DNA纳米球可附接(例如，通过吸附)至微阵列(测序流动池)。流动池可以是包被有二氧化硅、钛和六甲基二硅氮烷(HMDS)以及光阻材料的硅片。测序可以通过连接荧光探针与DNA的非链式测序进行。询问位置的荧光的颜色可以通过高分辨率摄像机来可视化。可以确定衔接子序列之间的核苷酸序列的同一性。

[0323] 在一些实施方案中，可利用AnyDot.芯片(Genovoxx,Germany)进行高通量测序。特别地，AnyDot.芯片允许核苷酸荧光信号检测增强10x-50x。AnyDot.芯片及其使用方法部分地描述于国际公布申请号WO 02088382、WO 03020968、WO 03031947、WO 2005044836、PCT/EP 05/05657、PCT/EP 05/05655；以及德国专利申请号DE 101 49 786、DE 102 14 395、DE 103 56 837、DE 10 2004 009 704、DE 10 2004 025 696、DE 10 2004 025 746、DE 10 2004 025 694、DE 10 2004 025 695、DE 10 2004 025 744、DE 10 2004 025 745和DE 10 2005 012 301中。

[0324] 其他高通量测序系统包括公开于以下文献中的那些高通量测序系统：Venter,J.等人.Science 2001年2月16日；Adams,M.等人.Science2000年3月24日；和M.J.Levene等人.Science 299:682-686,2003年1月；以及美国公开申请号20030044781和2006/0078937。总的来说，这样的系统涉及对通过经由聚合反应暂时添加碱基而具有多个碱基(在核酸分子上测量的)的靶核酸分子进行测序，即，实时地追踪待测序的模板核酸分子上核酸聚合酶的活性。然后，可通过由在添加碱基的序列的每一步中核酸聚合酶的催化活性鉴定哪个碱基正在并入靶核酸的增长的互补链中，来推导序列。靶核酸分子复合体上的聚合酶提供于适合于沿着靶核酸分子移动的位置中，并且使寡核苷酸引物在活性位点处延伸。多个标记类型的核苷酸类似物提供于活性位点附近，其中每个可区别类型的核苷酸类似物与靶核酸序列中的不同的核苷酸互补。增长的核酸链通过使用聚合酶进行延伸，以在活性位点处将核苷酸类似物添加至核酸链，其中被添加的核苷酸类似物与活性位点处的靶核酸的核苷酸互补。鉴定出作为聚合步骤的结果而被添加至寡核苷酸引物的核苷酸类似物。重复提供标记的核苷酸类似物、使增长的核酸链聚合以及鉴定添加的核苷酸类似物的步骤，使得核酸链进一步延伸并确定靶核酸的序列。

[0325] 本公开内容提供了单元型定相的方法，该方法包括由单个DNA分子生成多个读取对以及使用该读取对组装该DNA分子的多个叠连群，其中至少1％的读取对跨越单个DNA分子上大于50kB的距离，并且以大于70％的准确度进行单元型定相。在一些实施方案中，至少10％的读取对跨越单个DNA分子上大于50kB的距离。在其他实施方案中，其中至少1％的读取对跨越单个DNA分子上大于100kB的距离。在一些实施方案中，以大于90％的准确度进行单元型定相。

[0326] 在进一步的方面，本公开内容提供了单元型定相的方法，该方法包括由单个DNA分子(例如，在体外)生成多个读取对以及使用该读取对组装该DNA分子的多个叠连群，其中至少1％的读取对跨越单个DNA分子上大于30kB的距离，并且以大于70％的准确度进行单元型定相。在一些实施方案中，至少10％的读取对跨越单个DNA分子上大于30kB的距离。在其他实施方案中，至少1％的读取对跨越单个DNA分子上大于50kB的距离。在其他实施方案中，以大于90％的准确度进行单元型定相。在一些实施方案中，以大于70％的准确度进行单元型定相。

[0327] 在特定的实施方案中，本公开内容进一步提供了包含本公开内容的一个或多个组分的试剂盒。所述试剂盒可用于对本领域技术人员来说显而易见的任何应用，包括上述那些应用。所述试剂盒可包含例如多个缔合分子、固定剂、内切核酸酶(例如，限制性内切核酸酶)、连接酶和/或其组合。在一些情况下，缔合分子可以是蛋白质，包括例如组蛋白。在一些情况下，固定剂可以是甲醛或任何其他DNA交联剂。

[0328] 在一些情况下，试剂盒包含多个珠子。珠子可以是顺磁性的和/或包被有捕获剂。例如，珠子可包被有链霉亲和素和/或抗体。

[0329] 在一些情况下，试剂盒可包含衔接子寡核苷酸和/或测序引物。此外，试剂盒可包含能够使用衔接子寡核苷酸和/或测序引物扩增读取对的装置。

[0330] 在一些情况下，试剂盒还可包含其他试剂，包括但不限于裂解缓冲液、连接试剂(例如，dNTP、聚合酶、多核苷酸激酶和/或连接酶缓冲液等)和PCR试剂(例如，dNTP、聚合酶和/或PCR缓冲液等)。

[0331] 试剂盒还可包含针对使用试剂盒的组分和/或针对生成读取对的说明书。

[0332] 本公开内容的技术与其他技术如其他染色质组装程序相比，可提供许多优点。优点包括但不限于减少输入DNA量的要求、缩短完成方案的总时间、缩短完成方案的操作时间(hands-on time)、提高DNA回收率、去除昂贵和/或耗时的步骤、更容易自动化、更容易放大和更高的吞吐量。

[0333] 本文公开的技术可要求少量的输入DNA。例如，所需的输入DNA可少于约5微克(μg)、少于约4.5μg、少于约4μg、少于约3.5μg、少于约3μg、少于约2.5μg、少于约2μg、少于约1.5μg、少于约1μg、少于约900纳克(ng)、少于约800ng、少于约700ng、少于约600ng、少于约500ng、少于约400ng、少于约300ng、少于约200ng或少于约100ng。在一些情况下，所需的输入DNA少于约500ng。

[0334] 由样品制备测序文库的总经过时间(即，“时钟时间”)可以很短。例如，由样品制备测序文库(例如，染色质组装文库)的总时间可以少于约5.5天、少于约5天、少于约4.5天、少于约4天、少于约3.5天、少于约3天、少于约2.5天、少于约2天、少于约1.5天、少于约1天或少于约0.5天。在一些情况下，制备测序文库的总时间少于约2天。

[0335] 用户(例如，科学家或技术员)制备测序文库所需的有效时间(即，“操作时间”)的量可以很短。例如，操作时间的量可以少于约8小时、少于约7小时、少于约6小时、少于约5小时、少于约4小时、少于约3小时、少于约2小时或少于约1小时。在一些情况下，制备测序文库的操作时间的量少于约4小时。

[0336] 可使用本文公开的技术提高例如在交联逆转步骤之后回收的DNA的量。例如，交联逆转步骤之后的DNA回收率可以为至少5％、至少10％、至少15％、至少20％、至少25％、至少

30％、至少35％、至少40％、至少45％、至少50％、至少55％、至少60％、至少65％、至少70％、至少75％、至少80％、至少85％、至少90％或至少95％。在一些情况下，交联逆转步骤之后的DNA回收率为至少30％到至少50％。

[0337] 可使用本公开内容的技术避免某些步骤，包括昂贵或耗时的步骤。例如，可在不需要透析的情况下制备测序文库。可在不需要染色质生物素化的情况下制备测序文库。可在不需要染色质下拉的情况下制备测序文库。可在不需要生物素珠子占据步骤的情况下制备测序文库。可在不需要特定消化如ExoIII消化的情况下制备测序文库。还可以减少所需的染色质的量。例如，与先前染色质组装文库制备相比，所需的染色质的量可减少至少2倍、至少3倍、至少4倍、至少5倍、至少6倍、至少7倍、至少8倍、至少9倍或至少10倍。所需的染色质的量可少于约5个单位、少于约4.5个单位、少于约4个单位、少于约3.5个单位、少于约3个单位、少于约2.5个单位、少于约2个单位、少于约1.5个单位、少于约1个单位、少于约0.9个单位、少于约0.8个单位、少于约0.7个单位、少于约0.6个单位、少于约0.5个单位、少于约0.4个单位、少于约0.3个单位、少于约0.2个单位或少于约0.1个单位。1个单位的染色质相当于组装成染色质的1微克(μg)的DNA。

[0338] 图8中所示的计算机系统500可被理解为可从介质511和/或网络端口505读取指令的逻辑装置，该装置可任选地连接到具有固定介质512的服务器509。该系统(诸如图8中所示)可包含CPU 501、磁盘驱动器503、可选的输入设备如键盘515和/或鼠标516以及可选的监视器507。通过指示的通信介质可实现与本地或远程位置处的服务器的数据通信。通信介质可包括传送和/或接收数据的任何工具。例如，通信介质可以是网络连接、无线连接或因特网连接。这样的连接可提供通过万维网的通信。可以设想，与本发明相关的数据可通过这样的网络或如图8所示的由一方522接收和/或检查的连接进行传送。

[0339] 图9是示出可结合本公开内容的示例性实施方案使用的计算机系统100的第一示例性结构的框图。如图9中所示，示例性计算机系统可包含用于处理指令的处理器102。处理器的非限制性实例包括：Intel Xeon™处理器、AMD Opteron™处理器、Samsung 32-bit RISC ARM1176JZ(F)-S v1.0™处理器、ARM Cortex-A8 Samsung S5PC100™处理器、ARM Cortex-A8 Apple A4™处理器、Marvell PXA 930™处理器或功能相当的处理器。多个执行线程可用于并行处理。在一些实施方案中，无论是在单个计算机系统中、集群中，还是通过包括多个计算机、蜂窝电话和/或个人数据助理设备的网络的跨系统分布，也均可使用多个处理器或具有多个核的处理器。

[0340] 如图9中所示，高速缓冲存储器104可连接或并入到处理器102中，以为处理器102最近或频繁使用的指令或数据提供高速存储器。处理器102通过处理器总线108与北桥106连接。北桥106通过存储器总线112与随机存取存储器(RAM)110连接，并且管理处理器102对RAM110的访问。北桥106还通过芯片组总线116与南桥114连接。南桥114进而与外围总线118连接。外围总线可为例如PCI、PCI-X、PCI Express或其他外围总线。北桥和南桥通常被称为处理器芯片组，并且管理处理器、RAM和外围总线118上的外围组件之间的数据传送。在一些替代的架构中，北桥的功能可被并入处理器中而不使用单独的北桥芯片。

[0341] 在一些实施方案中，系统100可包含附接到外围总线118的加速器卡122。加速器可包括现场可编程门阵列(FPGA)或用于加速某些处理的其他硬件。例如，可使用加速器用于自适应数据重组或评价在扩展集处理中使用的代数表达式。

[0342] 软件和数据存储在外部存储124中，并且可被加载到RAM 110和/或缓存104以供处理器使用。系统100包括用于管理系统资源的操作系统；操作系统的非限制性实例包括：Linux、Windows™、MACOS™、BlackBerry OS™、iOS™和其他功能相当的操作系统，以及用于根据本公开内容的示例性实施方案管理数据存储和优化的运行在操作系统之上的应用软件。

[0343] 在该实例中，系统100还包括与外围总线连接的网络接口卡(NIC)120和121，用于向外部存储提供网络接口，诸如网络附加存储(NAS)和可用于分布式并行处理的其他计算机系统。

[0344] 图10为示出具有多个计算机系统202a和202b、多个蜂窝电话和个人数据助理202c以及网络附加存储(NAS)204a和204b的网络200的示意图。在示例性实施方案中，系统202a、202b和202c可管理数据存储并优化对网络附加存储(NAS)204a和204b中存储的数据的数据访问。将数学模型用于数据，并且使用在计算机系统202a和202b以及蜂窝电话和个人数据助理系统202c的分布式并行处理进行评价。计算机系统202a和202b以及蜂窝电话和个人数据助理系统202c还可提供并行处理，用于存储在网络附加存储(NAS)204a和204b中的数据的自适应数据重组。图10仅示出了示例，并且可结合本公开内容的多种实施方案使用多种其他计算机架构和系统。例如，可使用刀片服务器提供并行处理。处理器刀片可通过背板进行连接以提供并行处理。存储还可通过单独的网络接口与背板或网络附加存储(NAS)连接。

[0345] 在一些示例性实施方案中，处理器可维持单独的存储器空间，并通过网络接口、背板或其他连接器传送数据用于通过其他处理器进行并行处理。在其他实施方案中，一些或全部处理器可使用共享虚拟地址存储器空间。

[0346] 图11为根据示例性实施方案使用共享虚拟地址存储器空间的多处理器计算机系统300的框图。该系统包含可访问共享存储器子系统304的多个处理器302a-f。该系统在存储器子系统304中包含多个可编程硬件存储器算法处理器(MAP)306a-f。每个MAP 306a-f可包含存储器308a-f和一个或多个现场可编程门阵列(FPGA)310a-f。MAP提供可配置的功能单元，并且可向FPGA 310a-f提供特定算法或算法部分用于与相应处理器紧密配合进行处理。例如，在示例性实施方案中，MAP可用于评估关于数据模型的代数表达式并执行自适应数据重组。在该实例中，为达到这些目的，每个MAP均可被所有处理器全局访问。在一种配置中，每个MAP可使用直接存储器访问(DMA)来访问相关联的存储器308a-f，从而使其独立于相应的微处理器302a-f且与该微处理器异步地执行任务。在这种配置中，MAP可将结果直接提供给另一个MAP用于流水操作和算法的并行执行。

[0347] 上述计算机架构和系统仅是示例性的，并且可结合示例性实施方案使用多种其他计算机、蜂窝电话和个人数据助理架构和系统，包括使用通用处理器、协同处理器、FPGA和其他可编程逻辑设备、片上系统(SOC)、专用集成电路(ASIC)和其他处理和逻辑元件的任意组合的系统。在一些实施方案中，计算机系统的全部或部分以软件或硬件可实现。任何种类的数据存储介质均可与示例性实施方案结合使用，包括随机存取存储器、硬盘驱动器、闪速存储器、磁带驱动器、磁盘阵列、网络附加存储(NAS)以及其他局部或分布式数据存储设备和系统。

[0348] 在示例性实施方案中，可使用在上述任一种或其他计算机架构和系统上执行的软件模块来实现计算机系统。在其他实施方案中，该系统的功能部分地或完全地在固件、可编程逻辑设备如图11所示的现场可编程门阵列(FPGA)、片上系统(SOC)、专用集成电路(ASIC)

或其他处理和逻辑元件中可实现。例如，可以通过使用硬件加速卡如图9所示的加速卡122，采用硬件加速来实现集处理器和优化器。

[0349] 宏基因组学和复杂样品

[0350] 生物或生物医学样品、生态或环境样品和食品样品的微生物含量常常通过依赖于培养的方法来鉴定或定量。由于许多微生物是不可培养的，或者不适合在实验室中培养，因此大量的微生物生物多样性可能被基于培养的方法所忽略。对数以千计的生物体进行平行测序的鸟枪宏基因组测序方法可以使研究人员对给定的复杂样品中存在的大多数生物体中的大多数基因进行全面采样。这种方法可以实现可能难以分析的细菌多样性的评估和不可培养的微生物的研究。然而，不被支持的鸟枪测序法生成包含短读序列的大量读取，所述短读序列在没有参考序列或没有从头组装序列所需的长范围连接信息的一些来源的情况下可能难以组装。短读鸟枪法数据的生物信息学分析(例如，ConStrains)可能只需要鸟枪法数据；然而，输出由通过序列特征分箱但未组装的叠连群组成，并且最近的水平转移区段可能被错误地分箱。单分子长读测序(例如，Pacific Biosciences和Oxford Nanopore Technologies MinION)提供长范围组装的可能性；然而，它们可能提供低丰度基因组的低覆盖率，并且每个组装碱基的成本相对较高。16S RNA扩增可用于对群落16S RNA进行深度采样；然而，该技术仅提供粗分类信息，而不分析菌株差异、病原体类型等。合成长读取(例如，Moleculo，10X)可提供叠连群的真正支架化；然而，样品制备可能是复杂的且不标准化的，每个样品的成本可能更高，并且在Moleculo研究中报告了高水平的污染。体内邻近连接可以提供长范围支架化，并可以将额外的基因组元件(例如，质粒)与宿主一起放置；然而，体内邻近连接需要完整的细胞，并且由于基因组不均匀的压紧或与DNA结合蛋白质的缔合，可能导致邻近数据中的群落组分的不均匀呈现。

[0351] 微生物群落通常由丰度极不均匀的数十个、数百个或数千个可识别的操作分类单位(OTU)组成，其每一个都具有不同量的菌株变异。进一步结合该问题，微生物常常通过各种配偶交换手段来交换遗传物质，并且遗传物质的这些区段可被并入它们的宿主的染色体中，从而导致细菌群落内猖獗的水平基因转移。因此，微生物基因组通常以广泛存在的基因的核心基因组和特定菌株中可能存在或可能不存在的其他基因的核心基因组来描述。描述复杂的微生物群落如人类肠道微生物组的组成基因组和动态是重要且困难的挑战。

[0352] 由于从头宏基因组组装的困难，已经开发并广泛采用了几种较简单的方法来询问和描述所述微生物群落的组分。例如，16S RNA扩增和测序是评估群落组成的常见方法。虽然可以在比较框架中使用这种方法描述微生物群落在各种刺激或处理前后的动态，但它提供了实际群落组成的非常狭窄的视野(因为并没有了解到关于其16S区域以外的实际基因组的信息)。分箱方法也已被证明可用于将鸟枪读取或由鸟枪读取组装的叠连群分类。这些方法用于将分离的基因组片段临时分配给OTU。然而，它们本质上是假设生成器，并且无法对这些片段进行排序和定向，或者将片段分配给OTU内的菌株。重要的是，这些方法不适合鉴定水平转移序列，因为它们检测原点的OTU而不是当前的连接。从这个角度来看，这些基于k-聚体发生、测序深度和其他特征的分箱方法是理解分离的宏基因组组分的权宜之计，因为迄今为止高度连续的组装在可靠、快速且经济合理的方法中是不可能的。

[0353] 本文公开的技术提供了相对于现有技术的几个关键优点。首先，我们的“Chicago”文库可以提供广泛的基因组连接信息，并且可以快速且可靠地制备。如本文所述，该方案可

解决来源于宏基因组群落的DNA的特殊特征。可生成测序文库以准备用于在少于两天内测序。此外，因为这些文库可以在完全体外方案中生成，所以可能不需要培养任何物质。那么原则上这些技术可以组装任何可被回收DNA的微生物组群落成员。第三，该方法比从头组装和支架化的其他方法更简单、更快速且更全面。

[0354] 本文公开了用于宏基因组样品中的生物体的遗传分析的方法和工具，如不能在实验室环境中培养的微生物和存在于各种环境中的微生物。本公开内容提供了来自包含连接性数据的复杂宏基因组数据集的读取数据的从头基因组组装的方法。本文公开的方法和组合物生成支架化数据，该数据统一地且完全地表示宏基因组样品中的复合种类。

[0355] 图12A示出了用于邻近连接的程序的示意图。DNA 1201如高分子量DNA与组蛋白1202一起温育，并随后交联1203(例如，用甲醛)以形成染色质聚集体1204。这将DNA分子锁进支架中，以供进一步的操作和分析。该DNA随后进行消化1205，并用标志物如生物素将经消化的末端补平1206。标记的末端随后彼此随机连接1207，然后例如通过蛋白质消化将连接的聚集体释放1208。标志物可随后用于选择含有连接接头的DNA分子1209，如通过链霉亲和素-生物素结合。这些分子可随后进行测序，并且每个读取对中的读取来源于源分子的两个不同的区域，所述区域被至多输入DNA大小的某一插入片段距离分隔开。

[0356] 图12B示出了用于宏基因组分析的样品制备的两条线路，它们可以单独地使用或一起使用。将单个DNA制剂1210(例如，来自粪便样品)输入该过程中。在粪便样品的情况下，收集的DNA可以为约50千碱基片段，如来自使用Qiagen粪便DNA试剂盒的制剂。可由该DNA制备体外染色质组装1211(例如，“Chicago”)和鸟枪1212文库制剂。染色质组装文库1213和鸟枪法文库1214可使用彼此不同的条码1215和1216。这两种文库可随后汇集以用于测序1217。通过使用这样的方案，单个DNA制剂可充当以下两个测序文库的输入：鸟枪和体外染色质组装。需要少于1μg的输入DNA来生成两个文库，并且这些文库可以在测序过程中单独地条码化以进行汇集。随后可将这些数据首先组装成叠连群，然后使用来自体外染色质组装文库的长范围连接信息进行支架化。仅这些数据就可以生成许多大于一兆碱基的支架，从而能够比目前可实现的更为全面地了解微生物基因组结构和动力学。从样品到高度连续组装的处理时间可以在一个星期内。

[0357] 图12C示出了本公开内容的程序可采用的支架化技术的示例性示意图。体外染色质组装读取对可用于生成叠连群的生成树(未示出)，以确定正确的组装中哪些叠连群(彩色箭头)彼此邻近。然后，在本地窗口(例如，1220)中，可以针对体外染色质组装数据对所有可能的排序和方向进行测试。如图12C所示，在绿色叠连群1221的两个可能的方向中，体外染色质组装对1222将跨越短距离(顶部)或更远的距离(底部)。可针对每个文库训练的体外染色质组装距离的模型比较每一种方向的可能性。在邻近连接过程中，可以通过两个区段沿着DNA的线性聚合物相隔的距离的缓慢递减函数来描述连接这两个区段的概率。因此，回收跨越短、中和长距离的全部来自相同的单个文库的对。可以通过递减幂律函数很好地模拟特定距离的概率。也就是说，越来越不可能观察到跨越越来越大的距离的读取对。本文公开的组装技术(例如，“HiRise”)可以利用数据的这一方面来准确地将叠连群排序并定向到支架中。

[0358] 本发明方法的一些实施方案包括体外组装的染色质聚集体的邻近连接和测序，所述染色质聚集体包括宏基因组DNA样品，或来自直接从样品例如生物医学或生物样品、生态

或环境样品或食物样品获得的未培养的微生物的DNA样品。在兼容的实施方案中，将核酸组装成复合体，结合，裂解以暴露内部双链断裂，进行标记以促进断裂接头的分离，以及进行重新连接以生成进行测序的成对末端序列。在一些这样的成对末端序列中，推断出成对末端读取的两端映射至共同的核酸分子，即使成对读取的序列映射至不同的叠连群。

[0359] 在类似优选的实施方案中，使用标识符如核酸条码标记结合的复合体的暴露末端，使得复合体被标记或条码化，从而推断出标记相邻序列很可能来自单个核酸。另外，通常条码化序列可以映射至多个叠连群，但是叠连群随后被推断出映射至共同的核酸分子。

[0360] 在类似优选的实施方案中，通过添加除组蛋白以外的核酸结合蛋白质，如核蛋白、转座酶、转录因子、拓扑异构酶、特异性或非特异性双链DNA结合蛋白质或其他合适的蛋白质，来组装复合体。备选地或组合地，使用纳米颗粒而不是组蛋白或其他核酸结合蛋白质来组装复合体。

[0361] 在类似优选的实施方案中，依靠天然存在的复合体来保留核酸复合体的连接信息。在一些这样的情况下，分离核酸以保留天然组装的复合体，或者在处理或分离之前用稳定剂如固定剂进行处理。

[0362] 在任何组装或分离的复合体中，在一些情况下，可以依靠交联来稳定核酸复合体的形成，而在备选的情况下，核酸结合部分相互作用足以在没有交联的情况下保持复合体完整性。

[0363] 本文的方法和组合物可单独地或与独立获得或生成的序列数据如鸟枪法测序数据组合，生成异质核酸样品中的基因组、染色体或独立核酸分子的基因组信息的组装。可组装基因组，以代表可培养或不可培养的生物体，如在广泛的宏基因组群落如人类口腔或肠道微生物组中的丰富或稀有的生物体，且包括不适合培养生长的生物体。生物体也可以是具有来自其他个体的混合群或群体的遗传物质的样品(如含有来自多个不同人类个体的细胞或核酸的样品)中的个体。在一些情况下，通过使用广泛可用的高通量测序技术，本公开内容的方法提供了高通量、无培养的基因组组装的快速且简单的方法。

[0364] 独立于靶标的微生物检测的应用

[0365] 生物或生物医学样品、生态或环境样品、工业微生物样品和食品样品的微生物含量常常通过依赖于培养的方法来鉴定或定量。培养微生物可以依赖于各种因素，包括但不限于pH、温度、湿度和营养物质。为未知的或先前未培养的生物体确定培养条件往往是耗时且困难的过程。

[0366] 许多微生物目前不能在实验室中培养。基于培养的方法忽视了大量的微生物的生物多样性。本公开内容的方法和组合物可应用于宏基因组样品中生物体的遗传分析，如不能在实验室环境中培养且存在于众多环境中的微生物或病毒。宏基因组样品的非限制性实例包括生物样品，其包括组织、尿、汗、唾液、痰液和粪便；空气和大气；来自诸如池塘、湖泊、海、洋等水体的水样品；生态样品如土壤和灰尘；以及食品。各种宏基因组样品中的微生物含量的分析可用于包括但不限于医学、法医学、环境监测和食品科学的应用。

[0367] 在从受试者，例如哺乳动物受试者如人类或其他动物获得的生物或生物医学样品中，鉴定出包含一组微生物的个体微生物或“微生物标签”或“微生物指纹”。在一些方面，这样的信息用于医学应用或目的。在一些方面，鉴定包括确定存在或不存在微生物属或种，或具有先前未鉴定的或不常见的基因突变如可赋予细菌菌株抗生素抗性的突变的微生物属

或种。在一些方面，鉴定包括测定来自一个或多个微生物种或一个或多个微生物属的微生物DNA的水平。在一些情况下，微生物标签或指纹表明特定属或种的微生物DNA的水平与样品中来自不同属或种的微生物DNA的水平相比升高或显著较高。在一些方面，样品的微生物标签或指纹表明来自特定属或种的微生物DNA的水平与该样品中来自其他属或种的微生物DNA的水平相比降低或显著较低。在一些方面，通过定量样品中存在的各种类型的微生物(例如，不同的属或种)的微生物DNA的水平来确定样品的微生物标签或指纹。在一些方面，确定样品中存在的微生物的各种属或种的微生物DNA水平，并与对照样品或标准物的微生物DNA水平进行比较。

[0368] 在一些方面，疑似患有医学病况的受试者中微生物属或种的存在被确信地诊断为患有由该微生物属或种引起的医学病况。在一些情况下，如果微生物属或种疑似可传播给其他个体，例如通过接触或接近，则该信息用于将个体与其他个体隔离。在一些情况下，关于样品中存在的微生物或微生物种的信息用于确定特定医学治疗是否消除受试者中的微生物并治疗例如细菌感染。

[0369] 在一些方面，如果样品中特定属或种的微生物DNA水平降低或显著低于对照样品或标准物，则从中获得样品的受试者被诊断为患有疾病，例如癌症(例如，乳腺癌)。在一些方面，确定样品中存在的微生物的各种属或种的微生物DNA的水平，并与该样品中存在的其他各种属或种进行比较。在一些方面，如果样品中特定属或种的微生物DNA的水平降低或显著低于该样品中检测到的其他微生物属或种的微生物DNA，则从中获得样品的受试者很有可能患有疾病，例如癌症。

[0370] 在环境或生态样品，例如空气样品、水样品和土壤或灰尘样品中，鉴定出包含一组微生物的个体微生物或“微生物标签”或“微生物指纹”。在一些方面，使用环境或生态样品中微生物的鉴定和微生物多样性的分析来改进用于监测污染物对生态系统的影响和净化污染环境的策略。对微生物群落如何处理污染物的增进的了解改善了对污染场所从污染中恢复的潜力的评估，并增加了生物强化或生物刺激的可能性。这样的信息提供了对环境群落的功能生态学的有价值的见解。在一些情况下，微生物分析也更广泛地应用于鉴定空气、特定水体以及土壤和灰尘样品中存在的物种。例如，这可以用来建立入侵物种和濒危物种的范围，并追踪季节性种群。

[0371] 对环境或生态样品中的微生物群落的鉴定和分析还可用于农业应用。微生物群体执行植物生长所需的各种各样的生态系统功能，包括固定大气氮、营养物循环、抑制疾病以及螯合铁和其他金属。这样的信息可用于，例如，改善作物和牲畜中的病害检测以及对增强的农业实践的适应，所述增强的农业实践通过利用微生物与植物之间的关系来提高作物健康。

[0372] 在一些实施方案中，在工业微生物样品中，例如在用于产生各种生物活性化学品如精细化学品、农业化学品和药物的微生物群落中，鉴定出包含一组微生物的个体微生物或“微生物标签”或“微生物指纹”。微生物群落产生大量的生物活性化学品。

[0373] 基于序列分析的微生物检测和鉴定还可用于食品安全、食品确证和欺诈检测。例如，在宏基因组样品中的微生物检测和鉴定允许检测和鉴定疑似腐败或污染的食物中的不可培养的和先前未知的病原体，包括细菌、病毒和寄生虫。估计美国约80％的食源性疾病病例是由非特定的病原体引起的，所述病原体包括尚未公认为引起食源性疾病的已知病原

体、已知存在于食物中但未证实致病性的物质以及未知的病原体，所以对整个群体的微生物分析可以提供减少食源性疾病的机会。随着对食品全球供应的认识不断增加以及对获取食品如海鲜和贝类的可持续实践的认识不断增加，微生物检测可用于评估食品的真伪，例如确定声称来自世界上特定区域的鱼类是否真正来自世界的这个区域。

[0374] 连接测定在异质样品中的应用

[0375] 本文方法的应用还涉及到异质样品中的已知或未知分子的连接测定。本文还考虑了除新型生物体检测以外的与异质样品中连接信息测定有关的应用。在一些实施方案中，测定异质核酸样品中的核酸如染色体的连接信息。获得来自多个个体的包含DNA的样品，如来自犯罪现场、小便池或厕所、战场、水槽或垃圾废物场的样品。核酸信息例如经由鸟枪法测序来获得，并测定连接信息。通常，个体的独特基因组信息不是通过单个基因座来鉴定，而是通过基因座的组合来鉴定，如单核苷酸多态性(SNP)、插入或缺失(in/del)或点突变或等位基因，它们共同表示独特或基本独特的遗传性状组合。在许多情况下，单独的性状不足以鉴定特定的个体。然而，使用连接信息，如通过实施本文方法可获得的连接信息，人们不仅鉴定异质样品中存在的聚集等位基因，如采用本领域可获得的鸟枪法或备选的高通量测序方法鉴定，而且还确定所述样品中的特定分子中存在的特定等位基因组合。因此，人们不是简单地确定样品中的特定等位基因，而是确定将等位基因组合映射至特定个体所需的染色体上这些等位基因的组合，对于所述特定个体，其基因组信息可通过先前获得的基因组序列得到，或通过可从亲属获得的序列信息得到。连接信息在以下情况下也具有价值，其中已知基因存在于异质样品中，但其基因组环境未知。例如，在一些情况下，已知个体具有抵抗抗生素治疗的有害感染。鸟枪法测序很有可能鉴定出抗生素抗性基因。然而，通过实施本文的方法，获得了关于抗生素抗性基因的基因组环境的有价值的信息。因此，通过不仅鉴定抗生素抗性基因而且鉴定该基因所处的生物体的基因组，人们能够根据生物体基因组信息的其余部分鉴定靶向抗生素抗性基因宿主的备选治疗。例如，靶向在抗性微生物中缺乏或易受第二抗生素影响的代谢途径，使得抗性微生物被清除，尽管在第一选择时其对抗生素具有抗性。或者，使用关于患者中抗生素抗性基因的宿主的更完整基因组信息，确定该抗性基因是否来自“野生型”微生物，或者它是否可能来源于从实验室“逃出”或故意释放的微生物的实验室菌株。

[0376] 样品

[0377] 检测其中微生物的样品可以是包含微生物群体或异质核酸群体的任何样品。实例包括来自人类受试者或动物受试者的生物或生物医学样品；环境和生态样品，包括但不限于土壤和水样品，如来自池塘、湖泊、海、洋等的水样品；或疑似变质或污染的食品。

[0378] 可从生物受试者获得生物样品。受试者可以指任何动物(例如，哺乳动物)，包括但不限于人类、非人灵长类、啮齿动物、狗、猫、猪、鱼等。样品可以从任何受试者、个体或生物来源获得，其包括例如人类或非人类动物，包括哺乳动物和非哺乳动物、脊椎动物和无脊椎动物。样品可以包括受感染或污染的组织样品，例如包括皮肤、心脏、肺、肾、乳房、胰、肝、肌肉、平滑肌、膀胱、胆囊、结肠、肠、脑、前列腺、食道和甲状腺的组织样品。样品可包括受感染或污染的生物样品，例如血液、尿、脑脊液、精液、唾液、痰液和粪便。

[0379] 在一些情况下，异质样品包括来源于至少两个个体的核酸，如从两个或更多个个体使用的小便池或厕所获得的样品，或者从来自至少两个个体的血液或组织合并在一起的

场所如战场或犯罪现场获得的样品。通过实施本文公开的方法，确定样品的连接信息。

[0380] 可选择获得样品的方法用于合适的样品类型和期望的应用。例如，组织样品可以在手术过程中通过活检或切除获得；血液可以通过静脉穿刺获得；而唾液、痰液和粪便可以由个体自行提供于容器中。

[0381] 在一些方面，粪便样品来源于动物，如哺乳动物(例如，非人灵长类、马、牛、犬、猫、猪和人)。粪便样品可具有任何合适的重量。粪便样品可以为至少50g、60g、70g、80g、90g、100g、110g、120g、130g、140g、150g或更多。粪便样品可含有水。在一些方面，粪便样品含有至少60％、65％、70％、75％、80％、85％或90％或更多的水。在一些方面，储存粪便样品。粪便样品可以在2-8℃下储存几天(例如，3-5天)，或在-20℃下或更低的温度下储存更长的时间段(例如，超过5天)。在一些方面，粪便样品可由个体或受试者提供。在一些方面，粪便样品可从粪便存放的地方收集。在一些方面，粪便样品可包含在预定时间段从单个个体收集的多个样品。在一段时间内于多个时间点收集的粪便样品可用于例如在治疗感染的过程中监测个体粪便中的生物多样性。在一些方面，粪便样品包含来自若干个体的样品，例如疑似感染同一病原体或患有相同疾病的若干个体。

[0382] 在一些情况下，样品包括包含微生物群体或群落的环境样品或生态样品。环境样品的非限制性实例包括大气或空气样品、土壤或灰尘样品和水样品。可分析空气样品以确定空气中的微生物组成，例如，疑似具有被认为是健康威胁的微生物例如引起疾病的病毒的地区中的空气。在一些方面，对空气样品的微生物组成的了解可用于监测环境的变化。

[0383] 可针对包括但不限于公共安全和环境监测的目的分析水样品。可分析例如来自饮用水供应库的水样品，以确定饮用水供应中的微生物多样性以及对人类健康的潜在影响。可以分析水样品以确定由于大气的局部温度和气体组成的变化而对微生物环境的影响。可以在一年中的不同时间对水样品，例如来自池塘、湖泊、海、洋或其他水体的水样品进行取样。在一些方面，在一年中的不同时间获取多个样品。可以在距离水体表面不同的深度收集水样品。例如，可以在水体表面或在距离水体表面至少1米(例如至少2、3、4、5、6、7、8、9米或更深)处收集水样品。在一些方面，可以从水体的底部收集水样品。

[0384] 可对土壤和灰尘样品取样以研究微生物多样性。土壤样品可以提供关于土壤和水中病毒和细菌运动的信息，并可用于生物修复，其中可应用基因工程来开发能够降解有害污染物的土壤微生物。土壤微生物群落可以具有数千种不同的生物体，它们包含大量的遗传信息，例如在一克土壤中估计有2,000到18,000个不同的基因组。可以在距离表面不同的深度收集土壤样品。在一些方面，在表面收集土壤。在一些方面，在表面以下至少1in(例如，至少2、3、4、5、6、7、8、9或10in或更深)处收集土壤。在一些方面，在表面以下1-10in(例如2-9in、3-8in、4-7in或5-6in)的深度收集土壤。可以在一年中的不同时间收集土壤样品。在一些方面，在特定季节如冬季、春季、夏季或秋季收集土壤样品。在一些方面，在特定月份收集土壤样品。在一些方面，在包括但不限于龙卷风、飓风或雷暴的环境现象之后收集土壤样品。在一些情况下，在一段时间内收集多个土壤样品以允许在时间过程中监测微生物多样性。可从不同的生态系统如农业生态系统、森林生态系统和来自不同地理区域的生态系统收集土壤样品。

[0385] 食品样品可以是任何疑似污染、腐败、引起人类疾病的或疑似具有感兴趣的微生物或核酸的食品。可小规模产生食品样品，如在单个商店中。可以工业规模生产食品样品，

如在大型食品制造或食品加工厂中。食品样品的实例非限制性地包括动物产品，包括生的或煮熟的海鲜、贝类、生的或煮熟的鸡蛋、未煮熟的肉类(包括牛肉、猪肉和家禽肉)、未经巴氏杀菌的牛奶、未经巴氏杀菌的软奶酪、生热狗和熟食肉类；植物产品，包括新鲜农产品和沙拉；水果产品，如新鲜农产品和果汁；以及加工和/或制备的食品，如自制罐头食品、大量生产的罐头食品和三明治。在一些方面，用于分析的食品样品，例如疑似污染或变质的食品样品，可在室温例如20℃至25℃下储存。在一些方面，食品样品在低于室温的温度，如低于20℃、18℃、16℃、14℃、12℃、10℃、8℃、6℃、4℃、2℃、0℃、-10℃、-20℃、-40℃、-60℃或-80℃或更低的温度下储存。在一些方面，食品样品在高于室温的温度，如高于26℃、28℃、30℃、32℃、34℃、36℃、38℃、40℃或50℃或更高的温度下储存。在一些方面，食品样品在未知温度下储存。食品样品可储存特定的时间段，例如1天、1周、1个月或1年。在一些情况下，食品样品储存至少1天、1周、1个月、6个月、1年、2年或更长时间。食品样品可能是易腐烂的并且具有有限的保质期。在制造厂中生产的食品样品可从特定的生产批次或生产阶段获得。食品样品可从不同社区的不同商店和不同的制造厂获得。

[0386] 核酸分子

[0387] 可从含有多种其他组分，如蛋白质、脂质和非模板核酸的宏基因组样品分离核酸分子(例如，DNA或RNA)。可从获自动物、植物、细菌、真菌或任何其他细胞生物体的任何细胞材料获得核酸分子。用于本公开内容的生物样品还包括病毒颗粒或制剂。核酸分子可直接从生物体获得，或从获自生物体的生物样品，例如从血液、尿、脑脊液、精液、唾液、痰液、粪便和组织获得。核酸分子可直接从获自生物体的生态或环境样品，例如从空气样品、水样品和土壤样品获得。核酸模板可直接从疑似变质或污染的食品样品，例如肉类样品、农产品样品、水果样品、生食样品、加工食品样品、冷冻样品等获得。

[0388] 使用多种方法提取并纯化核酸。在一些情况下，通过采用苯酚、苯酚/氯仿/异戊醇，或类似的制剂，包括TRIzol和TriReagent进行有机萃取来纯化核酸。萃取技术的其他非限制性实例包括：(1)有机萃取，然后进行乙醇沉淀，例如，在使用或不使用自动化核酸提取器，例如，可从Applied Biosystems(Foster City,Calif.)获得的341型DNA提取器的情况下，使用苯酚/氯仿有机试剂(Ausubel等人，1993)；(2)固定相吸附法(美国专利号5,234,809;Walsh等人，1991)；和(3)盐诱导的核酸沉淀法(Miller等人，1988)，这类沉淀法通常被称为“盐析”法。核酸分离和/或纯化可包括核酸可特异性或非特异性结合的磁性粒子的使用，然后使用磁体分离珠子，洗涤珠子并从珠子上洗脱核酸(参见，例如美国专利号5,705,628)。以上分离方法之前可以是酶消化步骤，以帮助从样品中除去不需要的蛋白质，例如，采用蛋白酶K或其他酶如蛋白酶进行消化。参见，例如，美国专利号7,001,724。如果需要的话，可向裂解缓冲液中添加RNA酶抑制剂。对于某些细胞或样品类型，可在方案中增加蛋白质变性/消化步骤。纯化方法可针对分离DNA、RNA或二者。当DNA和RNA二者在提取程序过程中或之后被共同分离出时，可采用进一步的步骤分开纯化其中一种或二者。可以生成提取的核酸的亚级分，例如，通过根据大小、序列或其他物理或化学特征的纯化。除了初始的核酸分离步骤以外，还可在本公开内容的方法中的任何步骤之后进行核酸的纯化，如用于去除过量或不需要的试剂、反应物或产物。在一些情况下，例如当考虑检测RNA编码的基因组时，用逆转录酶处理核酸样品，使得核酸样品中的RNA分子充当用于合成互补DNA分子的模板。在一些情况下，这样的处理促进核酸样品的下游分析。

[0389] 在一些情况下，如2003年10月9日公开的美国专利申请公开号US2002/0190663 A1中所述获得核酸模板分子。在一些情况下，通过各种技术，如通过Maniatis等人，Molecular Cloning:A Laboratory Manual,Cold Spring Harbor,N.Y.,pp.280-281(1982)所述以及公知的实验室资源的最新资料中的那些技术从生物样品中提取核酸分子。可以首先从生物样品中提取核酸，然后在体外进行交联。可进一步从核酸中去除天然缔合蛋白质(例如，组蛋白)。

[0390] 本文公开的方法可应用于任何高分子量双链DNA，包括例如从组织、细胞培养物、体液、动物组织、植物、细菌、真菌、病毒等分离的DNA。

[0391] 多个单独样品中的每一个可独立地包含至少1ng、2ng、5ng、10ng、20ng、30ng、40ng、50ng、75ng、100ng、150ng、200ng、250ng、300ng、400ng、500ng、1μg、1.5μg、2μg、5μg、10μg、20μg、50μg、100μg、200μg、500μg或1000μg或更多的核酸物质。在一些情况下，多个单独样品中的每一个可独立地包含少于约1ng、2ng、5ng、10ng、20ng、30ng、40ng、50ng、75ng、100ng、150ng、200ng、250ng、300ng、400ng、500ng、1μg、1.5μg、2μg、5μg、10μg、20μg、50μg、100μg、200μg、500μg、1000μg或更多的核酸。

[0392] 可使用多种核酸定量方法。核酸定量方法的非限制性实例包括分光光度分析，以及测量与核酸结合并在结合时选择性地发荧光的染料例如溴化乙锭的荧光强度。

[0393] 核酸复合体

[0394] 在一些情况下，包含来自一个或多个宏基因组或异质样品的DNA的核酸与缔合分子或核酸结合部分结合形成核酸复合体。在一些情况下，核酸复合体包含与多个缔合分子或部分结合的核酸，如多肽；非蛋白质的有机分子；以及纳米颗粒。在一些情况下，结合剂在多个接触点与单独的核酸结合，使得在这些接触点的区段独立于其共同的磷酸二酯骨架而被保持在一起。

[0395] 在一些情况下，结合核酸包括在核酸分子的区段之间形成连接，例如共价连接。可在核酸分子的远隔区段之间形成连接。在一些情况下，结合核酸形成核酸复合体包括使核酸与缔合分子或部分(本文也称为核酸结合分子或部分)交联。在一些情况下，缔合分子包含氨基酸，包括但不限于肽和蛋白质，如DNA结合蛋白质。示例性DNA结合蛋白质包括天然染色质组分，如组蛋白，例如组蛋白2A、2B、3A、3B、4A和4B。在一些情况下，多个核酸结合部分包含重构染色质或体外组装染色质。染色质可由长度约为150kbp的DNA分子重构。在一些情况下，染色质由长度为至少50、100、125、150、200、250kbp或更长的DNA分子重构。在一些情况下，结合蛋白质包含转录因子或转座酶。非蛋白质有机分子也与本公开内容相容，如鱼精蛋白、精胺、亚精胺或其他带正电荷的分子。在一些情况下，缔合分子包含纳米颗粒，如具有带正电荷的表面的纳米颗粒。许多纳米颗粒组合物与本公开内容相容。在一些方面，纳米颗粒包含硅，如包被有阳性涂层的硅，以结合带负电荷的核酸。在一些情况下，纳米颗粒为基于铂的纳米颗粒。纳米颗粒可以是磁性的，它可促进交联的序列区段的分离。

[0396] 核酸通过与本公开内容一致的多种方法与缔合分子结合。在一些情况下，核酸与缔合分子交联。交联的方法包括紫外线照射、化学和物理(例如，光学)交联。化学交联剂的非限制性实例包括甲醛和补骨脂素(Solomon等人，Proc.Natl.Acad.Sci.USA 82:6470-6474,1985；Solomon等人，Cell 53:937-947,1988)。在一些情况下，交联通过向包含核酸分子和染色质蛋白质的混合物中添加包含约2％甲醛的溶液来进行。可用于交联DNA的试剂的其他非限制性实

例包括但不限于丝裂霉素C、氮芥、美法仑、1,3-丁二烯双环氧化物、顺铂(II)和环磷酰胺。在一些情况下,交联剂形成连接相对较短距离的交联——如约2 Å、3 Å、4 Å或5 Å。

[0397] 在一些情况下,核酸复合体,例如与体外组装的染色质结合的核酸(本文称为染色质聚集体)附接至固体支持体,包括但不限于珠子,例如磁珠。

[0398] 在一些实施方案中,核酸复合体存在于样品中,而不是在提取之后或在提取的同时进行组装。通常,在这样的情况中,核酸复合体包含天然核小体或与样品的核酸复合的其他天然核酸结合分子。

[0399] 天然或后续生成的核酸复合体在一些情况下是独立稳定的。在一些情况下,通过采用交联剂处理将天然或后续生成的核酸复合体稳定化。

[0400] 染色质重构

[0401] 重构的染色质作为结合部分通过许多方法来实现。本文所设想的重构染色质被广泛地用于涵盖大量结合部分与裸核酸的结合。结合部分包括组蛋白和核小体,但在对重构染色质的一些解释中还包括其他核蛋白质,如转录因子、转座子或其他DNA,或其他核酸结合蛋白质、精胺或亚精胺,或其他非多肽核酸结合部分,纳米颗粒如有机或无机纳米颗粒核酸结合剂。

[0402] 在一些情况下,使用重构染色质将天然染色质组分或天然染色质组分的同源物重新组装到裸核酸上,如组蛋白或核小体重新组装到天然核酸上。

[0403] 重构染色质的两种方法包括(1)组蛋白向DNA上的不依赖于ATP的随机沉积,和(2)周期性核小体的ATP依赖性组装。本公开内容考虑以上任一种方法与本文公开的一种或多种方法一起使用。两种生成染色质的方法的实例可见于Lusser等人("Strategies for the reconstitution of chromatin,"Nature Methods(2004),1(1):19-26),其通过引用以其全文并入本文。

[0404] 本文考虑了其他重构染色质的方法,它们被严格定义为向裸核酸添加核小体或组蛋白,或更宽泛地定义为向裸核酸添加任何部分,并且染色质的组成和其重构方法都不应被认为在一些实施方案中是限制性的。在一些情况下,'染色质重构'指的不是天然染色质的生成,而是新核酸复合体的生成,如包含通过与纳米颗粒结合而稳定的核酸的复合体,该纳米颗粒例如是具有包含促进核酸结合或核酸结合和交联的部分的表面的纳米颗粒。

[0405] 或者,在一些情况下,不进行重构,而是依靠天然核酸复合体来稳定核酸以用于下游分析。通常,这样的核酸复合体包含天然组蛋白,但考虑了包含其他核蛋白质、DNA结合蛋白质、转座酶、拓扑异构酶或其他DNA结合蛋白质的复合体。

[0406] 裂解核酸分子

[0407] 可裂解核酸分子,如来自核酸复合体中宏基因组样品的结合核酸分子,以暴露内部核酸末端并产生双链断裂。在一些情况下,裂解核酸分子,如核酸复合体中的核酸分子,以暴露核酸末端并形成未在其磷酸二酯骨架上物理连接的至少两个片段或区段。各种方法可用于裂解内部核酸末端和/或生成来源于核酸的片段,包括但不限于机械、化学和酶法,如剪切、声处理、非特异性内切核酸酶处理或特异性内切核酸酶处理。备选方法涉及酶切,如采用拓扑异构酶、碱基修复酶、转座酶如Tn5或磷酸二酯骨架切口酶的酶切。

[0408] 在一些情况下,通过消化裂解核酸。消化可包括与限制性内切核酸酶接触。可以根据已知的基因组序列信息来选择限制性内切核酸酶,以适应消化产生的游离核酸末端的平

均数目。限制性内切核酸酶可在被称为限制位点的特异性识别核苷酸序列处或附近切割。在消化过程中可以使用在整个基因组中具有较高相对丰度的限制位点的限制性内切核酸酶，以产生与具有相对较低相对丰度的限制位点的限制内切核酸酶相比更大数目的暴露核酸末端，因为更多的限制位点可产生更多的切割位点。在一些情况下，使用具有非特异性限制位点或多于一个限制位点的限制性内切核酸酶。非特异性限制位点的非限制性实例是CCTNN。碱基A、C、G和T是指DNA链的四种核苷酸碱基——腺嘌呤、胞嘧啶、鸟嘌呤和胸腺嘧啶。碱基N代表四种DNA碱基中的任何一种——A、C、G和T。具有相应限制位点的酶并非识别特定序列以用于切割，而是可识别多于一个序列以用于切割。例如，被识别的前五个碱基可以是CCTAA、CCTAT、CCTAG、CCTAC、CCTTA、CCTTT、CCTTG、CCTTC、CCTCA、CCTCT、CCTCG、CCTCC、CCTGA、CCTGT、CCTGG或CCTGC(16种可能性)。在一些情况下，具有非特异性限制位点的酶的使用与具有特异性限制位点的酶相比产生了更多的切割位点。限制性内切核酸酶可以具有至少4、5、6、7、8个碱基对或更长的限制识别序列。用于消化核酸复合体的限制酶可切割单链和/或双链核酸。限制性内切核酸酶可产生单链断裂或双链断裂。限制性内切核酸酶切割可产生钝性末端、3’突出端或5’突出端。3’突出端的长度可以为至少1、2、3、4、5、6、7、8或9个碱基或更长。5’突出端的长度可以为至少1、2、3、4、5、6、7、8或9个碱基或更长。限制酶的实例包括但不限于AatII、Acc65I、AccI、AciI、AclI、AcuI、AfeI、AflII、AflIII、AgeI、AhdI、AleI、AluI、AlwI、AlwNI、ApaI、ApaLI、ApeKI、ApoI、AscI、AseI、AsiSI、AvaI、AvaII、AvrII、BaeGI、BaeI、BamHI、BanI、BanII、BbsI、BbvCI、BbvI、BccI、BceAI、BcgI、BciVI、BclI、BfaI、BfuAI、BfuCI、BglI、BglII、BlpI、BmgBI、BmrI、BmtI、BpmI、Bpu10I、BpuEI、BsaAI、BsaBI、BsaHI、BsaI、BsaJI、BsaWI、BsaXI、BscRI、BscYI、BsgI、BsiEI、BsiHKAI、BsiWI、BslI、BsmAI、BsmBI、BsmFI、BsmI、BsoBI、Bsp1286I、BspCNI、BspDI、BspEI、BspHI、BspMI、BspQI、BsrBI、BsrDI、BsrFI、BsrGI、BsrI、BssHII、BssKI、BssSI、BstAPI、BstBI、BstEII、BstNI、BstUI、BstXI、BstYI、BstZ17I、Bsu36I、BtgI、BtgZI、BtsCI、BtsI、Cac8I、ClaI、CspCI、CviAII、CviKI-1、CviQI、DdeI、DpnI、DpnII、DraI、DraIII、DrdI、EacI、EagI、EarI、EciI、Eco53kI、EcoNI、Eco0109I、EcoP15I、EcoRI、EcoRV、FatI、FauI、Fnu4HI、FokI、FseI、FspI、HaeII、HaeIII、HgaI、HhaI、HincII、HindIII、HinfI、HinP1I、HpaI、HpaII、HphI、Hpy166II、Hpy188I、Hpy188III、Hpy99I、HpyAV、HpyCH4III、HpyCH4IV、HpyCH4V、KasI、KpnI、MboI、MboII、MfeI、MluI、MlyI、MmeI、MnlI、MscI、MseI、MslI、MspA1I、MspI、MwoI、NaeI、NarI、Nb.BbvCI、Nb.BsmI、Nb.BsrDI、Nb.BtsI、NciI、NcoI、NdeI、NgoMIV、NheI、NlaIII、NlaIV、NmeAIII、NotI、NruI、NsiI、NspI、Nt.AlwI、Nt.BbvCI、Nt.BsmAI、Nt.BspQI、Nt.BstNBI、Nt.CviPII、PacI、PaeR7I、PciI、PflFI、PflMI、PhoI、PleI、PmeI、PmlI、PpuMI、PshAI、PsiI、PspGI、PspOMI、PspXI、PstI、PvuI、PvuII、RsaI、RsrII、SacI、SacII、SalI、SapI、Sau3AI、Sau96I、SbfI、ScaI、ScrFI、SexAI、SfaNI、SfcI、SfiI、SfoI、SgrAI、SmaI、SmlI、SnaBI、SpeI、SphI、SspI、StuI、StyD4I、StyI、SwaI、T、Taqα I、TfiI、TliI、TseI、Tsp45I、Tsp509I、TspMI、TspRI、Tth111I、XbaI、XcmI、XhoI、XmaI、XmnI和ZraI。

[0409] 连接

[0410] 裂解的核酸分子可通过邻近连接使用多种方法连接。裂解的核酸分子的连接可通过酶促和非酶促方案实现。非酶促连接反应的实例可包括美国专利号5,780,613和5,476,930中描述的非酶促连接技术，其每一个均通过引用以其全文并入本文。酶促连接反应可包

括使用连接酶。连接酶的非限制性实例是ATP依赖性双链多核苷酸连接酶、NAD+依赖性DNA或RNA连接酶和单链多核苷酸连接酶。连接酶的非限制性实例是大肠杆菌DNA连接酶、丝状栖热菌DNA连接酶、Tth DNA连接酶、水管致黑栖热菌DNA连接酶(I和II)、T3 DNA连接酶、T4 DNA连接酶、T4 RNA连接酶、T7 DNA连接酶、Taq连接酶、Ampligase(Epicentre® Technologies Corp.)、VanC型连接酶、9°N DNA连接酶、Tsp DNA连接酶、DNA连接酶I、DNA连接酶III、DNA连接酶IV、Sso7-T3 DNA连接酶、Sso7-T4 DNA连接酶、Sso7-T7 DNA连接酶、Sso7-Taq DNA连接酶、Sso7-大肠杆菌DNA连接酶、Sso7-Ampligase DNA连接酶以及热稳定的连接酶。连接酶可以是野生型、突变同种型和遗传工程变体。连接反应可含有缓冲液组分、小分子连接增强剂和其他反应组分。

[0411] 测序

[0412] 本文描述的或本领域中已知的合适的测序方法可用于从核酸分子获得序列信息。测序可通过经典的Sanger测序法来实现。测序还可使用高通量下一代测序系统来实现。下一代测序方法的非限制性实例包括单分子实时测序、离子半导体测序、焦磷酸测序、合成测序、连接测序和链终止。

[0413] 微生物

[0414] 在此检测的微生物可以是细菌、病毒、真菌、霉菌或任何其他微生物或其组合。

[0415] 在一些方面,在生物医学样品,例如生物流体或固体样品(包括但不限于唾液、血液和粪便)中检测到的微生物是与医学病况有关的至少一种细菌物种。临床上相关的细菌的非限制性实例包括橙黄弗拉托菌(Acetobacter aurantius)、鲍氏不动杆菌(Acinetobacter baumannii)、衣氏放线菌(Actinomyces israelii)、放射形土壤杆菌(Agrobacterium radiobacter)、根癌土壤杆菌(Agrobacterium tumefaciens)、无形体吞噬细胞(Anaplasma phagocytophilum)、茎瘤固氮根瘤菌(Azorhizobium caulinodans)、维涅兰德固氮菌(Azotobacter vinelandii)、炭疽芽孢杆菌(Bacillus anthracis)、短芽孢杆菌(Bacillus brevis)、蜡状芽孢杆菌(Bacillus cereus)、梭形气芽孢杆菌(Bacillus fusiformis)、地衣芽孢杆菌(Bacillus licheniformis)、巨大芽孢杆菌(Bacillus megaterium)、蕈状芽孢杆菌(Bacillus mycoides)、嗜热脂肪芽孢杆菌(Bacillus stearothermophilus)、枯草芽孢杆菌(Bacillus subtilis)、脆弱类杆菌(Bacteroides fragilis)、牙龈红棕色单胞菌(Bacteroides gingivalis)、产黑素拟杆菌(Bacteroides melaninogenicus)(现在称为产黑素普雷沃氏菌(Prevotella melaninogenica))、汉氏巴尔通氏体(Bartonella henselae)、五日热巴尔通氏体(Bartonella quintana)、支气管炎博德特氏菌(Bordetella bronchiseptica)、百日咳博德特氏菌(Bordetella pertussis)、布氏疏螺旋体(Borrelia burgdorferi)、流产布鲁氏菌(Brucella abortus)、马尔他布鲁氏菌(Brucella melitensis)、猪布鲁氏菌(Brucella suis)、鼻疽伯克霍尔德氏菌(Burkholderia mallei)、类鼻疽伯克霍尔德氏菌(Burkholderia pseudomallei)、洋葱伯克霍尔德氏菌(Burkholderia cepacia)、肉芽肿鞘杆菌(Calymmatobacterium granulomatis)、曲形弯曲杆菌(Campylobacter coli)、胚胎弯曲杆菌(Campylobacter fetus)、空肠弯曲杆菌(Campylobacter jejuni)、幽门螺杆菌(Campylobacter pylori)、砂眼衣原体(Chlamydia trachomatis)、肺炎嗜衣原体(Chlamydophila pneumoniae)(先前称为肺炎衣原体(Chlamydia pneumoniae))、鹦鹉热嗜衣原体(Chlamydophila psittaci)(先

前称为鹦鹉热衣原体(Chlamydia psittaci))、肉毒梭菌(Clostridium botulinum)、艰难梭菌(Clostridium difficile)、产气荚膜梭菌(Clostridium perfringens)(先前称为韦氏梭菌(Clostridium welchii))、破伤风梭菌(Clostridium tetani)、白喉棒杆菌(Corynebacterium diphtheriae)、梭形棒杆菌(Corynebacterium fusiforme)、伯氏考克斯氏体(Coxiella burnetii)、恰菲埃里希氏体(Ehrlichia chaffeensis)、阴沟肠杆菌(Enterobacter cloacae)、鸟肠球菌(Enterococcus avium)、耐久肠球菌(Enterococcus durans)、粪肠球菌(Enterococcus faecalis)、屎肠球菌(Enterococcus faecium)、鹑鸡肠球菌(Enterococcus galllinarum)、病臭肠球菌(Enterococcus maloratus)、大肠杆菌(Escherichia coli)、土拉热弗朗西丝氏菌(Francisella tularensis)、具核梭杆菌(Fusobacterium nucleatum)、阴道加德纳氏菌(Gardnerella vaginalis)、杜氏嗜血菌(Haemophilus ducreyi)、流感嗜血菌(Haemophilus influenzae)、副流感嗜血菌(Haemophilus parainfluenzae)、百日咳嗜血菌(Haemophilus pertussis)、阴道嗜血菌(Haemophilus vaginalis)、幽门螺杆菌(Helicobacter pylori)、肺炎克雷伯氏菌(Klebsiella pneumoniae)、嗜酸乳杆菌(Lactobacillus acidophilus)、布氏乳杆菌(Lactobacillus bulgaricus)、干酪乳杆菌(Lactobacillus casei)、乳酸乳球菌(Lactococcus lactis)、侵肺军团菌(Legionella pneumophila)、单核细胞增生利斯特氏菌(Listeria monocytogenes)、扭脱甲烷杆菌(Methanobacterium extroquens)、多形微杆菌(Microbacterium multiforme)、藤黄微球菌(Micrococcus luteus)、粘膜炎莫拉氏菌(Moraxella catarrhalis)、鸟分枝杆菌(Mycobacterium avium)、牛分枝杆菌(Mycobacterium bovis)、白喉分枝杆菌(Mycobacterium diphtheriae)、胞内分枝杆菌(Mycobacterium intracellulare)、麻风分枝杆菌(Mycobacterium leprae)、鼠麻风分枝杆菌(Mycobacterium lepraemurium)、草分枝杆菌(Mycobacterium phlei)、耻垢分枝杆菌(Mycobacterium smegmatis)、结核分枝杆菌(Mycobacterium tuberculosis)、发酵枝原体(Mycoplasma fermentans)、生殖道枝原体(Mycoplasma genitalium)、人型枝原体(Mycoplasma hominis)、穿透枝原体(Mycoplasma penetrans)、肺炎枝原体(Mycoplasma pneumoniae)、淋病奈瑟氏球菌(Neisseria gonorrhoeae)、脑膜炎奈瑟氏球菌(Neisseria meningitidis)、出血败血性巴斯德氏菌(Pasteurella multocida)、土拉热巴斯德氏菌(Pasteurella tularensis)、消化链球菌属(Peptostreptococcus)、牙龈红棕色单胞菌(Porphyromonas gingivalis)、产黑素普雷沃氏菌(Prevotella melaninogenica)(先前称为产黑素拟杆菌(Bacteroides melaninogenicus))、铜绿假单胞菌(Pseudomonas aeruginosa)、放射根瘤菌(Rhizobium radiobacter)、普氏立克次氏体(Rickettsia prowazekii)、鹦鹉热宫川氏体(Rickettsia psittaci)、五日热立克次氏体(Rickettsia quintana)、立氏立克次氏体(Rickettsia rickettsii)、砂眼立克次氏体(Rickettsia trachomae)、汉氏巴尔通氏体(Rochalimaea henselae)、五日热巴尔通氏体(Rochalimaea quintana)、龋齿罗氏菌(Rothia dentocariosa)、肠炎沙门氏菌(Salmonella enteritidis)、伤寒沙门氏菌(Salmonella typhi)、鼠伤寒沙门氏菌(Salmonella typhimurium)、粘质沙雷氏菌(Serratia marcescens)、痢疾志贺氏菌(Shigella dysenteriae)、金黄色葡萄球菌(Staphylococcus aureus)、表皮葡萄球菌(Staphylococcus epidermidis)、嗜麦芽糖寡养单胞菌(Stenotrophomonas

maltophilia)、无乳链球菌(Streptococcus agalactiae)、鸟肠球菌(Streptococcus avium)、牛链球菌(Streptococcus bovis)、大鼠链球菌(Streptococcus cricetus)、屎链球菌(Streptococcus faceium)、粪链球菌(Streptococcus faecalis)、野生链球菌(Streptococcus ferus)、鸡链球菌(Streptococcus gallinarum)、乳链球菌(Streptococcus lactis)、温和链球菌(Streptococcus mitior)、缓症链球菌(Streptococcus mitis)、变异链球菌(Streptococcus mutans)、口腔链球菌(Streptococcus oralis)、肺炎链球菌(Streptococcus pneumoniae)、酿脓链球菌(Streptococcus pyogenes)、鼠链球菌(Streptococcus rattus)、唾液链球菌(Streptococcus salivarius)、血链球菌(Streptococcus sanguis)、表兄链球菌(Streptococcus sobrinus)、苍白密螺旋体(Treponema pallidum)、齿垢密螺旋体(Treponema denticola)、霍乱弧菌(Vibrio cholerae)、逗号弧菌(Vibrio comma)、副溶血弧菌(Vibrio parahaemolyticus)、创伤弧菌(Vibrio vulnificus)、沃尔巴克氏体属(Wolbachia)、小肠结肠炎耶尔森氏菌(Yersinia enterocolitica)、鼠疫耶尔森氏菌(Yersinia pestis)和假结核耶尔森氏菌(Yersinia pseudotuberculosis)。

[0416] 在一些方面,在生物医学样品,例如生物流体或固体样品(包括但不限于唾液、血液和粪便)中检测到的微生物是与医学病况有关的至少一种病毒。在一些方面,病毒为DNA病毒。在一些方面,病毒为RNA病毒。人类病毒感染可具有人畜共患病或野生或家养动物起源。几种人畜共患病病毒经由与动物接触直接地,或经由暴露于受感染动物的尿或粪便或吸血节肢动物叮咬而间接地传染给人类。如果病毒能够在其新的人类宿主中适应并复制,则可能发生人-人传播。在一些方面,在生物医学样品中检测到的微生物是具有人畜共患病起源的病毒。

[0417] 在一些方面,在生物医学样品,例如生物流体或固体样品(包括但不限于唾液、血液和粪便)中检测到的微生物是与医学病况有关的至少一种真菌。临床上相关的真菌属的非限制性实例包括曲霉属(Aspergillus)、蛙粪霉属(Basidiobolus)、芽酵母属(Blastomyces)、假丝酵母属(Candida)、金孢子菌属(Chrysosporium)、球孢子菌属(Coccidioides)、耳霉属(Conidiobolus)、隐球酵母属(Cryptococcus)、表皮癣菌属(Epidermophyton)、组织胞浆菌属(Histoplasma)、小孢子菌属(Microsporum)、肺囊虫属(Pneumocystis)、孢子丝菌属(Sporothrix)和毛癣菌属(Trichophyton)。

[0418] 在一些方面,在食品样品,如疑似引起疾病的食品样品中检测到的微生物可以是致病细菌、病毒或寄生虫。可引起疾病的致病细菌、病毒或寄生虫的非限制性实例包括沙门氏菌属的种,如肠沙门氏菌和乍得沙门氏菌(S.bongori);弯曲杆菌属的种,如空肠弯曲杆菌、大肠弯曲杆菌(C.coli)和胚胎弯曲杆菌(C.fetus);耶尔森氏菌属的种,如小肠结肠炎耶尔森氏菌和假结核耶尔森氏菌;志贺氏菌属的种,如索氏志贺氏菌(S.sonnei)、鲍氏志贺氏菌(S.boydii)、弗氏志贺氏菌(S.flexneri)和痢疾志贺氏菌;弧菌属的种,如副溶血弧菌、霍乱弧菌血清型01和0139、霍乱弧菌血清型非-01和非-0139、创伤弧菌;考克斯氏体属的种,如伯氏考克斯氏体;分枝杆菌属的种,如牛分枝杆菌,它是牛肺结核的病原体,但也可传染人类;布鲁氏菌属的种,如马尔他布鲁氏菌(B.melitensis)、流产布鲁氏菌(B.abortus)、猪布鲁氏菌(B.suis)、木鼠布鲁氏菌(B.neotomae)、狗布鲁氏菌(B.canis)和羊布鲁氏菌(B.ovis);阪崎肠杆菌属的种(阪崎肠杆菌);气单胞菌属的种,如嗜水气单胞菌

(A.hydrophila)；邻单胞菌属的种，如类志贺邻单胞菌(P.shigelloides)；弗朗西丝氏菌属的种，如土拉热弗朗西丝氏菌；梭菌属的种，如产气荚膜梭菌和肉毒梭菌；葡萄球菌属的种，如金黄色葡萄球菌；芽孢杆菌属的种，如蜡状芽孢杆菌；利斯特氏菌属的种，如单核细胞增生利斯特氏菌；链球菌属的种，如酿脓链球菌A群(S.pyogenes of Group A)；诺如病毒(NoV、GI、GII、GIII、GIV和GV群)；甲型肝炎病毒(HAV，I-VI基因型)；戊型肝炎病毒(HEV)；呼肠孤病毒科病毒(Reoviridae viruse)，如轮状病毒；星状病毒科病毒(Astroviridae virus)，如星形病毒；钙毒病毒科病毒(Calciviridae virus)，如Sapo病毒；腺病毒科病毒(Adenoviridae virus)，如肠腺病毒；细小病毒科病毒(Parvoviridae virus)，如细小病毒；以及小核糖核酸病毒科病毒(Picornarviridae virus)，如爱知病毒。

[0419] 本文公开的方法的益处在于其有助于检测样品中未知身份的微生物或病原体，并将所述未知微生物或病原体的序列信息单独地或与附加序列信息(如通过鸟枪法测序或其他手段同时生成的附加序列信息)结合组装成部分或完全组装的基因组。因此，本文公开的方法不限于检测以上刚刚列出的一种或多种生物体；相反，通过本文公开的方法，人们能够鉴定和确定以上列表中的未知病原体，或不在以上列表中的生物体，或序列信息不可获得的生物体或科学上未知的生物体的实质性部分或全部基因组信息。

[0420] 本文公开的方法适用于许多异质核酸样品，如肠道微生物群落的探索性调查；患病个体或群体如患有未知原因的流行病的群体中的病原体检测；异质核酸样品中具有已知个体的连接信息特征的核酸的存在的检测；或与显示抗生素抗性感染的个体中的抗生素抗性有关的一种或多种微生物的检测。许多这些实施方案的一个共同的方面是它们受益于长范围连接信息的生成，如适合于将鸟枪法序列信息组装成叠连群、支架或部分或完整基因组序列的长范围连接信息的生成。鸟枪法或其他高通量序列信息与以上列出的至少一些问题相关，但是从本文公开的方法的实践结果获得了实质性的益处，以将鸟枪法序列组装成更大的定相核酸组装体，直到并包含部分、基本完整或完整的基因组。因此，使用本文公开的方法提供的益处大大超过了本领域已知的对异质样品单独实施鸟枪法测序。

[0421] 除了摄入污染和/或变质食品后由直接细菌感染引起的疾病外，微生物还可产生引起疾病的毒素，如肠毒素。在一些方面，在食品样品中检测到的微生物可以产生毒素，如肠毒素，它是一种靶向肠道的蛋白质外毒素；以及真菌毒素，它是一种由真菌界的生物体(俗称霉菌)产生的有毒次级代谢物。

[0422] 本公开内容的益处是，它能够使人们在不依赖于先前或甚至同时生成的有待于组装的一个或多个基因组的序列信息的情况下，获得异质样品的长范围基因组邻接信息。表示样品中生物体的基因组或染色体的支架使用通常标记的读取进行组装，该读取例如是共有共同的寡核苷酸标记的读取或彼此连接或以其他方式融合的成对末端读取，从而指示通常标记的序列信息来自于共同的基因组或染色体分子。

[0423] 因此，在不依赖于先前生成的叠连群或其他序列读取信息的情况下生成支架信息。从头支架信息有很多益处。例如，即使不能获得先前的序列信息，序列读取也可以被分配给共同的支架，使得在不依赖于先前的测序工作的情况下对全新的基因组进行支架化。当异质样品包含未知的、未培养的或不可培养的生物体时，这种益处尤其有用。依赖于未靶向的序列读取生成的测序项目可生成未被分配给任何已知叠连群序列的序列读取的集合，而很少有或没有与未知生物体(从中获得序列读取)的数目或身份有关的信息。例如，它们

可以代表单个个体、在基因组序列中具有高度异质性或杂合性的共同物种的个体群体、密切相关物种的复合体或不同物种的复合体。仅仅依赖于序列读取信息，人们将不能容易地区分上述情况。

[0424] 然而，通过使用本文公开的方法或组合物，人们能够将例如包含共同基因型或基因组的克隆复制物的样品，与包含单个物种的代表性异质群体的样品，与包含不同物种的松散相关生物体的样品，或这些情况的组合区分开来。通过依靠序列相似性来组装叠连群而不是独立地生成支架信息，区分杂合性与测序错误是具有挑战性的。即使假设没有发生实质性测序错误，估算从中获得密切相关基因组信息的基因型的数目也是具有挑战性的。例如，人们不能将包含单个物种的两个广泛不同代表(在许多不同的基因座处相对于彼此是杂合的)的样品与包含广泛多样性的密切相关基因型(在一个或仅几个基因座处，每一个基因型不同于其他基因型)的样品区分开来。通过单独使用序列读取信息，这两种情况都表现为具有大量等位基因多样性的单个叠连群组装体。然而，通过使用本文公开的方法和组合物，人们能够可靠地确定哪些等位基因映射至共同的支架，即使等位基因被均匀或未知序列的相当大的区域分隔开。

[0425] 在一些情况下，当研究包含病毒群体，如基于DNA基因组的病毒群体或逆转录病毒或其他基于RNA的病毒群体的异质样品(经由RNA基因组的逆转录，或者备选地或组合地，组装样品中的RNA上的复合体)时，本文生成的数据的这种益处尤其有用。由于病毒群体通常是相当异质的，了解群体内异质性的分布(在几个高度不同的群体中或在大量密切相关的群体中)在选择治疗靶标方面以及在追踪被研究的异质样品中病毒的起源方面具有特定益处。

[0426] 这并不是说本文公开的组合物和方法与叠连群信息或同时生成的序列读取不相容。相反，通过使用本文的方法和组合物生成的支架化信息特别适合于改进的叠连群组装或叠连群排列成支架。实际上，在本公开内容的一些实施方案中，同时生成的序列读取信息被组装成叠连群。使用传统的测序方法，如下一代测序方法并行生成序列读取信息。备选地或组合地，成对读取或寡核苷酸标记的读取信息用作序列信息本身，以“传统地”使用对齐的重叠序列生成叠连群。该信息进一步用于根据通过本文公开的组合物和方法生成的支架化信息来相对于彼此定位叠连群。

[0427] 还通过下面的编号实施方案阐明本公开内容的实施方案。

[0428] 编号实施方案1包括由第一DNA分子生成标记的序列的方法，该方法包括：(a)使所述第一DNA分子与多个缔合分子结合，以形成第一复合体，其中所述第一DNA分子包含第一DNA区段和第二DNA区段；(b)标记所述第一DNA区段和所述第二DNA区段，从而形成至少一个标记的DNA区段；(c)使所述复合体结合至具有直接结合所述复合体的组分的表面的固体支持体；以及(d)对所述标记的DNA区段的可识别部分，如与标记相邻的部分或在与标记的末端相对的末端的部分进行测序，从而获得所述标记的序列；其中所述多个缔合分子在步骤(a)和(b)之前或过程中不采用亲和标记物进行共价修饰。编号实施方案2包括编号实施方案1所述的方法，其中所述缔合分子包含通过肽键结合的氨基酸。编号实施方案3包括编号实施方案1-2中任一实施方案所述的方法，其中所述缔合分子包括多肽或蛋白质。编号实施方案4包括编号实施方案1-3中任一实施方案所述的方法，其中所述缔合分子包括组蛋白。编号实施方案5包括编号实施方案1-3中任一实施方案所述的方法，其中所述组蛋白来自与

所述第一DNA分子不同的来源。编号实施方案6包括编号实施方案1-3中任一实施方案所述的方法，其中所述缔合分子包括转座酶。编号实施方案7包括编号实施方案1-6中任一实施方案所述的方法，其中所述第一DNA分子与至少一个所述缔合分子非共价结合。编号实施方案8包括编号实施方案1-7中任一实施方案所述的方法，其中所述第一DNA分子与至少一个所述缔合分子共价结合。编号实施方案9包括编号实施方案1-8中任一实施方案所述的方法，其中所述第一DNA分子与至少一个所述缔合分子交联。编号实施方案10包括编号实施方案1-9中任一实施方案所述的方法，其中所述第一DNA分子使用固定剂进行交联。编号实施方案11包括编号实施方案1-10中任一实施方案所述的方法，其中所述固定剂包括甲醛。编号实施方案12包括编号实施方案1-11中任一实施方案所述的方法，其包括将所述多个缔合分子固定在固体支持体上。编号实施方案13包括编号实施方案1-12中任一实施方案所述的方法，其中所述固体支持体包括珠子。编号实施方案14包括编号实施方案1-13中任一实施方案所述的方法，其中所述珠子包含聚合物。编号实施方案15包括编号实施方案1-14中任一实施方案所述的方法，其中所述聚合物为聚苯乙烯或聚乙二醇(PEG)。编号实施方案16包括编号实施方案1-13中任一实施方案所述的方法，其中所述珠子为磁珠。编号实施方案17包括编号实施方案1-13中任一实施方案所述的方法，其中所述珠子为固相可逆固定化(SPRI)珠子。编号实施方案18包括编号实施方案1-13中任一实施方案所述的方法，其中所述固体支持体包括表面，并且其中所述表面包含多个羧基基团。编号实施方案19包括编号实施方案1-12中任一实施方案所述的方法，其中所述固体支持体不与任何多肽共价连接。编号实施方案20包括编号实施方案1-12中任一实施方案所述的方法，其中所述缔合分子在固定至所述固体支持体之前不与生物素共价连接。编号实施方案21包括编号实施方案1-20中任一实施方案所述的方法，其中通过切断所述第一DNA分子生成所述第一DNA区段和所述第二DNA区段。编号实施方案22包括编号实施方案1-21中任一实施方案所述的方法，其中在所述第一DNA分子与所述多个缔合分子结合之后切断所述第一DNA分子。编号实施方案23包括编号实施方案1-21中任一实施方案所述的方法，其中使用核酸酶切断所述第一DNA分子。编号实施方案24包括编号实施方案1-23中任一实施方案所述的方法，其中使用亲和标记物修饰所述第一DNA区段和所述第二DNA区段。编号实施方案25包括编号实施方案1-24中任一实施方案所述的方法，其中所述亲和标记物包含生物素。编号实施方案26包括编号实施方案1-25中任一实施方案所述的方法，其中所述亲和标记物为生物素修饰的核苷三磷酸(dNTP)。编号实施方案27包括编号实施方案1-26中任一实施方案所述的方法，其中所述亲和标记物为生物素修饰的脱氧核糖胞嘧啶三磷酸(dCTP)。编号实施方案28包括编号实施方案1-27中任一实施方案所述的方法，其中所述第一DNA区段用第一标记在至少第一端进行标记，而第二DNA区段用第二标记在至少第二端进行标记。编号实施方案29包括编号实施方案1-28中任一实施方案所述的方法，其中所述第一标记和所述第二标记相同。编号实施方案30包括编号实施方案1-28中任一实施方案所述的方法，其中使用转座酶标记所述第一DNA区段和所述第二DNA区段。编号实施方案31包括编号实施方案1-30中任一实施方案所述的方法，其中通过将所述第一DNA区段与所述第二DNA区段连接，使所述第一DNA区段用所述第二DNA区段进行标记，且使所述第二DNA区段用所述第一DNA区段进行标记。编号实施方案32包括编号实施方案1-31中任一实施方案所述的方法，其中使用连接酶将所述第一DNA区段与所述第二DNA区段连接。编号实施方案33包括编号实施方案1-32中任一实施方案所述

的方法，其中在步骤(c)之前切断所述连接的DNA区段。编号实施方案34包括编号实施方案1-24中任一实施方案所述的方法，其中使用物理方法切断所述连接的DNA区段。编号实施方案35包括编号实施方案1-34中任一实施方案所述的方法，其包括将所述连接的DNA区段与测序衔接子连接。编号实施方案36包括编号实施方案1-35中任一实施方案所述的方法，其中在所述第一DNA区段与所述第二DNA区段连接之前洗涤所述第一DNA区段少于10次。编号实施方案37包括编号实施方案1-36中任一实施方案所述的方法，其中在所述第一DNA区段与所述第二DNA区段连接之前洗涤所述第一DNA区段少于6次。编号实施方案38包括编号实施方案1至37中任一实施方案所述的方法，其包括使用所述标记的序列组装所述第一DNA分子的多个叠连群。

[0429] 编号实施方案39包括编号实施方案1至37中任一实施方案所述的方法，其包括使用所述标记的序列对所述第一DNA区段和所述第二DNA区段进行定相。编号实施方案40包括编号实施方案1至39中任一实施方案所述的方法，其中所述方法在不超过两天内完成。编号实施方案41包括编号实施方案1-40中任一实施方案所述的方法，其中所述第一DNA分子的所述结合在体外进行。编号实施方案42包括编号实施方案1-41中任一实施方案所述的方法，其中所述第一DNA分子的所述结合在体内进行。编号实施方案43包括编号实施方案1-42中任一实施方案所述的方法，其中所述方法在不超过2天内完成。编号实施方案44包括编号实施方案1-43中任一实施方案所述的方法，其中步骤(a)-(d)所需的操作时间的量不多于6小时。编号实施方案45包括编号实施方案1-44中任一实施方案所述的方法，其中所述第一DNA分子直接结合至所述固体支持体。编号实施方案46包括编号实施方案1-45中任一实施方案所述的方法，其中在步骤(a)-(d)之间不进行透析。

[0430] 编号实施方案47包括由第一DNA分子生成标记的序列的方法，该方法包括：(a)使所述第一DNA分子与多个缔合分子结合；(b)将所述第一DNA分子固定在固体支持体上；(c)切断所述第一DNA分子以生成第一DNA区段和第二DNA区段；(d)标记所述第一DNA区段和所述第二DNA区段，从而形成至少一个标记的DNA区段；以及(e)对所述标记的DNA区段进行测序，从而获得所述标记的序列；其中所述第一DNA分子直接结合至所述固体支持体。编号实施方案48包括编号实施方案47所述的方法，其中所述缔合分子包含氨基酸。编号实施方案49包括编号实施方案47-48中任一实施方案所述的方法，其中所述缔合分子包括多肽或蛋白质。编号实施方案50包括编号实施方案47-49中任一实施方案所述的方法，其中所述缔合分子包括组蛋白。编号实施方案51包括编号实施方案47-49中任一实施方案所述的方法，其中所述组蛋白来自与所述第一DNA分子不同的来源。编号实施方案52包括编号实施方案47-51中任一实施方案所述的方法，其中所述缔合分子包括转座酶。编号实施方案53包括编号实施方案47-52中任一实施方案所述的方法，其中所述第一DNA分子与所述缔合分子非共价结合。编号实施方案54包括编号实施方案47-53中任一实施方案所述的方法，其中所述第一DNA分子与所述缔合分子共价结合。编号实施方案55包括编号实施方案47-54中任一实施方案所述的方法，其中所述第一DNA分子与所述缔合分子交联。编号实施方案56包括编号实施方案47-55中任一实施方案所述的方法，其中所述第一DNA分子使用固定剂进行交联。编号实施方案57包括编号实施方案47-56中任一实施方案所述的方法，其中所述固定剂为甲醛。编号实施方案58包括编号实施方案47-57中任一实施方案所述的方法，其中所述固体支持体包括珠子。编号实施方案59包括编号实施方案47-58中任一实施方案所述的方法，其中所

述珠子包含聚合物。编号实施方案60包括编号实施方案47-59中任一实施方案所述的方法，其中所述聚合物包括聚苯乙烯或聚乙二醇(PEG)。编号实施方案61包括编号实施方案47-58中任一实施方案所述的方法，其中所述珠子为磁珠。编号实施方案62包括编号实施方案47-58中任一实施方案所述的方法，其中所述珠子为SPRI珠子。编号实施方案63包括编号实施方案47-62中任一实施方案所述的方法，其中所述固体支持体包括表面，并且其中所述表面包含多个羧基基团。编号实施方案64包括编号实施方案47-63中任一实施方案所述的方法，其中所述固体支持体不与任何多肽共价连接。编号实施方案65包括编号实施方案47-64中任一实施方案所述的方法，其中所述缔合分子在固定至所述固体支持体之前不与生物素共价连接。编号实施方案66包括编号实施方案47-65中任一实施方案所述的方法，其中在所述第一DNA分子与所述多个缔合分子中的至少一个结合之后切断所述第一DNA分子。编号实施方案67包括编号实施方案47-66中任一实施方案所述的方法，其中使用核酸酶切断所述第一DNA分子。编号实施方案68包括编号实施方案47-67中任一实施方案所述的方法，其中使用亲和标记物修饰所述第一DNA区段和所述第二DNA区段。编号实施方案69包括编号实施方案47-68中任一实施方案所述的方法，其中所述亲和标记物包含生物素。编号实施方案70包括编号实施方案47-69中任一实施方案所述的方法，其中所述亲和标记物为生物素修饰的核苷三磷酸(dNTP)。编号实施方案71包括编号实施方案47-70中任一实施方案所述的方法，其中所述亲和标记物为生物素修饰的脱氧核糖胞嘧啶三磷酸(dCTP)。编号实施方案72包括编号实施方案47-71中任一实施方案所述的方法，其中所述第一DNA区段用第一标记在至少第一端进行标记，而第二DNA区段用第二标记在至少第二端进行标记。编号实施方案73包括编号实施方案47-72中任一实施方案所述的方法，其中所述第一标记和所述第二标记相同。编号实施方案74包括编号实施方案47-72中任一实施方案所述的方法，其中使用转座酶标记所述第一DNA区段和所述第二DNA区段。编号实施方案75包括编号实施方案47-74中任一实施方案所述的方法，其中通过将所述第一DNA区段与所述第二DNA区段连接，使所述第一DNA区段用所述第二DNA区段进行标记，且使所述第二DNA区段用所述第一DNA区段进行标记。编号实施方案76包括编号实施方案47-75中任一实施方案所述的方法，其中使用连接酶将所述第一DNA区段与所述第二DNA区段连接。编号实施方案77包括编号实施方案47-76中任一实施方案所述的方法，其中使用物理方法切断所述连接的DNA区段。编号实施方案78包括编号实施方案47-77中任一实施方案所述的方法，其包括将所述连接的DNA区段与测序衔接子连接。编号实施方案79包括编号实施方案47-78中任一实施方案所述的方法，其中在所述第一DNA区段与所述第二DNA区段连接之前洗涤所述第一DNA区段少于10次。编号实施方案80包括编号实施方案47-79中任一实施方案所述的方法，其中在所述第一DNA区段与所述第二DNA区段连接之前洗涤所述第一DNA区段少于6次。编号实施方案81包括编号实施方案47至80中任一实施方案所述的方法，其包括使用所述标记的序列组装所述第一DNA分子的多个叠连群。编号实施方案82包括编号实施方案47至80中任一实施方案所述的方法，其包括使用所述标记的序列对所述第一DNA区段和所述第二DNA区段进行定相。编号实施方案83包括编号实施方案47-82中任一实施方案所述的方法，其中所述标记的序列包含读取对。编号实施方案84包括编号实施方案47至83中任一实施方案所述的方法，其中所述方法在不超过2天内完成。编号实施方案85包括编号实施方案47-84中任一实施方案所述的方法，其中所述第一DNA分子的所述结合在体外进行。编号实施方案86包括编号实施方案47-85中任一

实施方案所述的方法，其中所述第一DNA分子的所述结合在体内进行。编号实施方案87包括编号实施方案47-86中任一实施方案所述的方法，其中步骤(a)-(d)所需的操作时间的量不多于6小时。编号实施方案88包括编号实施方案47-87中任一实施方案所述的方法，其中在步骤(a)-(d)之间不进行透析。

[0431] 编号实施方案89包括用于由多个DNA分子生成多个标记的序列的方法，该方法包括：(a)使所述多个DNA分子与多个缔合分子结合；(b)切断所述多个DNA分子以生成多个DNA区段；(c)标记所述DNA区段的至少一部分以形成多个标记的DNA区段；以及(d)对所述标记的DNA区段进行测序，以获得多个标记的序列；其中所述多个缔合分子在步骤(a)和(b)之前或过程中不采用亲和标记物进行共价修饰。编号实施方案90包括编号实施方案89所述的方法，其中在步骤(b)之前，来自所述DNA分子的少于40％的DNA区段与不具有共同磷酸二酯键的其他DNA区段连接。编号实施方案91包括编号实施方案89-90中任一实施方案所述的方法，其中在步骤(b)之前，来自所述DNA分子的少于20％的DNA区段与不具有共同磷酸二酯键的其他DNA区段连接。编号实施方案92包括编号实施方案89-91中任一实施方案所述的方法，其中所述缔合分子包含氨基酸。编号实施方案93包括编号实施方案89-92中任一实施方案所述的方法，其中所述缔合分子包括多肽或蛋白质。编号实施方案94包括编号实施方案89-93中任一实施方案所述的方法，其中所述缔合分子包括组蛋白。编号实施方案95包括编号实施方案89-94中任一实施方案所述的方法，其中所述组蛋白来自与所述DNA分子不同的来源。编号实施方案96包括编号实施方案89-95中任一实施方案所述的方法，其中所述缔合分子包括转座酶。编号实施方案97包括编号实施方案89-96中任一实施方案所述的方法，其中所述DNA分子与所述缔合分子非共价结合。编号实施方案98包括编号实施方案89-97中任一实施方案所述的方法，其中所述DNA分子与所述缔合分子共价结合。编号实施方案99包括编号实施方案89-98中任一实施方案所述的方法，其中所述DNA分子与所述缔合分子交联。编号实施方案100包括编号实施方案89-99中任一实施方案所述的方法，其中所述DNA分子使用固定剂进行交联。编号实施方案101包括编号实施方案89-100中任一实施方案所述的方法，其中所述固定剂为甲醛。编号实施方案102包括编号实施方案89-101中任一实施方案所述的方法，其包括将所述多个缔合分子固定在多个固体支持体上。编号实施方案103包括编号实施方案89-102中任一实施方案所述的方法，其中所述固体支持体为珠子。编号实施方案104包括编号实施方案89-103中任一实施方案所述的方法，其中所述珠子包含聚合物。编号实施方案105包括编号实施方案89-104中任一实施方案所述的方法，其中所述聚合物包括聚苯乙烯或聚乙二醇(PEG)。编号实施方案106包括编号实施方案89-103中任一实施方案所述的方法，其中所述珠子包括磁珠。编号实施方案107包括编号实施方案89-103中任一实施方案所述的方法，其中所述珠子包括SPRI珠子。编号实施方案108包括编号实施方案89-102中任一实施方案所述的方法，其中所述固体支持体包括表面，并且其中所述表面包含多个羧基基团。编号实施方案109包括编号实施方案89-102中任一实施方案所述的方法，其中所述固体支持体不与任何多肽共价连接。编号实施方案110包括编号实施方案89-109中任一实施方案所述的方法，其中所述缔合分子在固定至所述固体支持体之前不与生物素共价连接。编号实施方案111包括编号实施方案89-110中任一实施方案所述的方法，其中使用亲和标记物修饰所述DNA区段的所述部分。编号实施方案112包括编号实施方案89-111中任一实施方案所述的方法，其中所述亲和标记物包含生物素。编号实施方案113包括编号实

施方案89-112中任一实施方案所述的方法，其中所述亲和标记物为生物素修饰的核苷三磷酸(dNTP)。编号实施方案114包括编号实施方案89-113中任一实施方案所述的方法，其中所述生物素修饰的核苷三磷酸(dNTP)为生物素修饰的脱氧核糖胞嘧啶三磷酸(dCTP)。编号实施方案115包括编号实施方案89-114中任一实施方案所述的方法，其中使用第一标记在至少第一端标记所述DNA区段的一部分。编号实施方案116包括编号实施方案89-115中任一实施方案所述的方法，其中使用转座酶标记所述DNA区段。编号实施方案117包括编号实施方案89-116中任一实施方案所述的方法，其中通过连接所述DNA区段与至少一个其他DNA区段来标记所述DNA区段的一部分。编号实施方案118包括编号实施方案89-117中任一实施方案所述的方法，其中使用连接酶将DNA区段的所述部分与所述其他DNA区段连接。编号实施方案119包括编号实施方案89-118中任一实施方案所述的方法，其中使用核酸酶切断所述DNA分子。编号实施方案120包括编号实施方案89-119中任一实施方案所述的方法，其中在步骤(c)之前切断所述连接的DNA区段。编号实施方案121包括编号实施方案89-120中任一实施方案所述的方法，其中使用物理方法切断所述连接的DNA区段。编号实施方案122包括编号实施方案89-121中任一实施方案所述的方法，其包括将所述连接的DNA区段与测序衔接子连接。编号实施方案123包括编号实施方案89-122中任一实施方案所述的方法，其中在将所述DNA区段连接以形成所述连接的DNA区段之前，将所述DNA区段洗涤少于10次。编号实施方案124包括编号实施方案89-123中任一实施方案所述的方法，其中在将所述DNA区段连接以形成所述连接的DNA区段之前，将所述DNA区段洗涤少于6次。编号实施方案125包括编号实施方案89至124中任一实施方案所述的方法，其包括使用所述读取对组装所述DNA分子的多个叠连群。编号实施方案126包括编号实施方案89至124中任一实施方案所述的方法，其包括使用所述读取对对所述DNA区段进行定相。编号实施方案127包括编号实施方案89至126中任一实施方案所述的方法，其中所述方法在不超过2天内完成。编号实施方案128包括编号实施方案89-127中任一实施方案所述的方法，其中步骤(a)-(d)所需的操作时间的量不多于6小时。编号实施方案129包括编号实施方案89-128中任一实施方案所述的方法，其中在步骤(a)-(d)之间不进行透析。编号实施方案130包括编号实施方案89-129中任一实施方案所述的方法，其中所述方法在少于2天内完成。编号实施方案131包括编号实施方案89-130中任一实施方案所述的方法，其中所述多个DNA分子不多于约5微克。编号实施方案132包括编号实施方案89-131中任一实施方案所述的方法，其中所述多个DNA分子的所述结合在体外进行。编号实施方案133包括编号实施方案89-132中任一实施方案所述的方法，其中所述多个DNA分子的所述结合在体内进行。

[0432] 编号实施方案134包括包含与DNA片段以体外复合体形式结合的多个缔合分子的组合物，其中所述体外复合体固定在固体支持体上，并且其中所述固体支持体不与任何多肽共价连接。编号实施方案135包括编号实施方案89-134中任一实施方案所述的组合物，其中所述固体支持体不与链霉亲和素共价连接。编号实施方案136包括编号实施方案89-134中任一实施方案所述的组合物，其中所述固体支持体包括珠子。编号实施方案137包括编号实施方案89-136中任一实施方案所述的组合物，其中所述珠子包含聚合物。编号实施方案138包括编号实施方案89-137中任一实施方案所述的组合物，其中所述聚合物包括聚苯乙烯或聚乙二醇(PEG)。编号实施方案139包括编号实施方案89-134中任一实施方案所述的组合物，其中所述珠子为SPRI珠子。编号实施方案140包括编号实施方案89-134中任一实施方

案所述的组合物，其中所述固体支持体包被有多个羧基基团。编号实施方案141包括编号实施方案89-134中任一实施方案所述的组合物，其中所述固体支持体不与任何多肽共价连接。编号实施方案142包括编号实施方案89-134中任一实施方案所述的组合物，其中所述缔合分子包含氨基酸。编号实施方案143包括编号实施方案89-134中任一实施方案所述的组合物，其中所述缔合分子包括多肽或蛋白质。编号实施方案144包括编号实施方案89-143中任一实施方案所述的组合物，其中所述缔合分子包括组蛋白。编号实施方案145包括编号实施方案89-144中任一实施方案所述的组合物，其中所述组蛋白来自与所述DNA分子不同的来源。编号实施方案146包括编号实施方案89-134中任一实施方案所述的组合物，其中所述缔合分子包括转座酶。编号实施方案147包括编号实施方案89-134中任一实施方案所述的组合物，其中所述第一DNA分子与所述缔合分子非共价结合。编号实施方案148包括编号实施方案89-134中任一实施方案所述的组合物，其中所述第一DNA分子与所述缔合分子共价结合。编号实施方案149包括编号实施方案89-148中任一实施方案所述的组合物，其中所述第一DNA分子与所述缔合分子交联。编号实施方案150包括编号实施方案89-134中任一实施方案所述的组合物，其中采用固定剂使所述缔合分子与所述DNA片段交联。编号实施方案151包括编号实施方案89-150中任一实施方案所述的组合物，其中所述固定剂为甲醛。编号实施方案152包括编号实施方案89-134中任一实施方案所述的组合物，其中采用亲和标记物修饰所述DNA片段。编号实施方案153包括编号实施方案89-152中任一实施方案所述的组合物，其中所述亲和标记物包含生物素。编号实施方案154包括编号实施方案89-153中任一实施方案所述的组合物，其中所述亲和标记物为生物素修饰的核苷三磷酸(dNTP)。编号实施方案155包括编号实施方案89-154中任一实施方案所述的组合物，其中所述生物素修饰的核苷三磷酸(dNTP)为生物素修饰的脱氧核糖胞嘧啶三磷酸(dCTP)。编号实施方案156包括编号实施方案89-155中任一实施方案所述的方法，其中所述多个DNA分子的所述结合在体外进行。编号实施方案157包括编号实施方案89-156中任一实施方案所述的方法，其中所述多个DNA分子的所述结合在体内进行。

[0433] 编号实施方案158包括用于由多个DNA分子生成多个标记的序列的方法，该方法包括：(a)获得与多个缔合分子结合的多个DNA分子；(b)切断所述DNA分子以生成至少多个DNA区段；(c)标记所述DNA区段的至少一部分以形成多个标记的DNA区段；以及(d)对所述标记的DNA区段进行测序，以获得多个标记的序列；其中所述多个DNA分子的总量少于约5微克(μg)。编号实施方案159包括用于由多个DNA分子生成多个标记的序列的方法，该方法包括：(a)获得与多个缔合分子结合的多个DNA分子；(b)切断所述DNA分子以生成至少多个DNA区段；(c)标记所述DNA区段的至少一部分以形成多个标记的DNA区段；以及(d)对所述标记的DNA区段进行测序，以获得多个标记的序列；其中在步骤(a)与步骤(d)之间不进行透析。编号实施方案160包括用于由多个DNA分子生成多个标记的序列的方法，该方法包括：(a)获得与多个缔合分子结合的多个DNA分子；(b)切断所述DNA分子以生成至少多个DNA区段；(c)标记所述DNA区段的至少一部分以形成多个标记的DNA区段；以及(d)对所述标记的DNA区段进行测序，以获得多个标记的序列；其中步骤(a)-(d)所需的操作时间的量少于6小时。编号实施方案161包括编号实施方案158、159或160中任一实施方案所述的方法，其中来自所述DNA分子的少于40％的DNA区段与来自任何其他DNA分子的DNA区段连接。编号实施方案162包括编号实施方案158-161中任一实施方案所述的方法，其中来自所述DNA分子的少于20％的

DNA区段与来自任何其他DNA分子的DNA区段连接。编号实施方案163包括编号实施方案158-162中任一实施方案所述的方法，其中所述缔合分子包含氨基酸。编号实施方案164包括编号实施方案158-162中任一实施方案所述的方法，其中所述缔合分子为多肽或蛋白质。编号实施方案165包括编号实施方案158-164中任一实施方案所述的方法，其中所述缔合分子为组蛋白。编号实施方案166包括编号实施方案158-165中任一实施方案所述的方法，其中所述组蛋白来自与所述DNA分子不同的来源。编号实施方案167包括编号实施方案158-166中任一实施方案所述的方法，其中所述缔合分子为转座酶。编号实施方案168包括编号实施方案158-167中任一实施方案所述的方法，其中所述DNA分子与所述缔合分子非共价结合。编号实施方案169包括编号实施方案158-168中任一实施方案所述的方法，其中所述DNA分子与所述缔合分子共价结合。编号实施方案170包括编号实施方案158-169中任一实施方案所述的方法，其中所述DNA分子与所述缔合分子交联。编号实施方案171包括编号实施方案158-170中任一实施方案所述的方法，其中所述DNA分子使用固定剂进行交联。编号实施方案172包括编号实施方案158-171中任一实施方案所述的方法，其中所述DNA分子使用甲醛进行交联。编号实施方案173包括编号实施方案158-172中任一实施方案所述的方法，其包括将所述多个缔合分子固定在多个固体支持体上。编号实施方案174包括编号实施方案158-173中任一实施方案所述的方法，其中所述固体支持体为珠子。编号实施方案175包括编号实施方案158-174中任一实施方案所述的方法，其中所述珠子包含聚合物。编号实施方案176包括编号实施方案158-175中任一实施方案所述的方法，其中所述聚合物为聚苯乙烯或聚乙二醇(PEG)。编号实施方案177包括编号实施方案158-176中任一实施方案所述的方法，其中所述珠子为磁珠。编号实施方案178包括编号实施方案158-177中任一实施方案所述的方法，其中所述珠子为SPRI珠子。编号实施方案179包括编号实施方案158-178中任一实施方案所述的方法，其中所述固体支持体包括表面，并且其中所述表面包含多个羧基基团。编号实施方案180包括编号实施方案158-179中任一实施方案所述的方法，其中所述固体支持体不与任何多肽共价连接。编号实施方案181包括编号实施方案158-180中任一实施方案所述的方法，其中所述缔合分子在固定至所述固体支持体之前不与生物素共价连接。编号实施方案182包括编号实施方案158-181中任一实施方案所述的方法，其中采用亲和标记物修饰所述DNA区段的所述部分。编号实施方案183包括编号实施方案158-182中任一实施方案所述的方法，其中所述亲和标记物包含生物素。编号实施方案184包括编号实施方案158-183中任一实施方案所述的方法，其中所述亲和标记物为生物素修饰的核苷三磷酸(dNTP)。编号实施方案185包括编号实施方案158-184中任一实施方案所述的方法，其中所述生物素修饰的核苷三磷酸(dNTP)为生物素修饰的脱氧核糖胞嘧啶三磷酸(dCTP)。编号实施方案186包括编号实施方案158-185中任一实施方案所述的方法，其中采用第一标记在至少第一端标记所述DNA区段的一部分。编号实施方案187包括编号实施方案158-186中任一实施方案所述的方法，其中使用转座酶标记所述DNA区段。编号实施方案188包括编号实施方案158-187中任一实施方案所述的方法，其中通过连接每一个所述DNA区段与至少一个其他DNA区段来标记所述DNA区段的一部分。编号实施方案189包括编号实施方案158-188中任一实施方案所述的方法，其中使用连接酶将DNA区段的所述部分与所述其他DNA区段连接。编号实施方案190包括编号实施方案158-189中任一实施方案所述的方法，其中使用核酸酶切断所述DNA分子。编号实施方案191包括编号实施方案158-190中任一实施方案所述的方

法，其中在步骤(c)之前切断所述连接的DNA区段。编号实施方案192包括编号实施方案158-191中任一实施方案所述的方法，其中使用物理方法切断所述连接的DNA区段。编号实施方案193包括编号实施方案158-192中任一实施方案所述的方法，其包括将所述连接的DNA区段与测序衔接子连接。编号实施方案194包括编号实施方案158-193中任一实施方案所述的方法，其中在将所述DNA区段连接以形成所述连接的DNA区段之前，将所述DNA区段洗涤少于约10次。编号实施方案195包括编号实施方案158-194中任一实施方案所述的方法，其中在将所述DNA区段连接以形成所述连接的DNA区段之前，将所述DNA区段洗涤少于约6次。编号实施方案196包括编号实施方案158-195中任一实施方案所述的方法，其包括使用所述读取对组装所述DNA分子的多个叠连群。编号实施方案197包括编号实施方案158-196中任一实施方案所述的方法，其包括使用所述读取对对所述DNA区段进行定相。编号实施方案198包括编号实施方案158-197中任一实施方案所述的方法，其中所述方法在不超过2天内完成。编号实施方案199包括编号实施方案158-198中任一实施方案所述的方法，其中步骤(a)中的所述获得包括使所述多个DNA分子与所述多个缔合分子结合。编号实施方案200包括编号实施方案158-199中任一实施方案所述的方法，其中步骤(a)中的所述获得包括收集与所述多个缔合分子结合的所述多个DNA分子。编号实施方案201包括编号实施方案158-200中任一实施方案所述的方法，其中所述多个DNA分子的总量不多于4μg。编号实施方案202包括编号实施方案158-201中任一实施方案所述的方法，其中所述多个DNA分子的总量不多于3μg。编号实施方案203包括编号实施方案158-202中任一实施方案所述的方法，其中所述多个DNA分子的总量不多于2μg。编号实施方案204包括编号实施方案158-203中任一实施方案所述的方法，其中步骤(a)-(d)所需的操作时间的量不多于5小时。编号实施方案205包括编号实施方案158-204中任一实施方案所述的方法，其中步骤(a)-(d)所需的操作时间的量不多于4小时。编号实施方案206包括编号实施方案158-205中任一实施方案所述的方法，其中在步骤(a)-(d)之间不进行透析。编号实施方案207包括编号实施方案158-206中任一实施方案所述的方法，其中所述方法在少于2天内完成。编号实施方案208包括编号实施方案158-207中任一实施方案所述的方法，其中所述多个DNA分子的所述结合在体外进行。编号实施方案209包括编号实施方案158-208中任一实施方案所述的方法，其中所述多个DNA分子的所述结合在体内进行。

[0434] 编号实施方案210包括在宿主群体中检测病原体的方法，该方法包括：a)从疑似具有共同病原体的多个个体中的每一个获得稳定化的样品；b)处理所述稳定化的样品以裂解所述稳定化的样品中的双链DNA；c)标记暴露的DNA末端；d)连接标记的暴露的DNA末端以形成标记的成对末端；e)在标记的成对末端进行测序以生成多个成对序列读取；f)将所述多个序列读取的成对序列读取的每一半分配给共同的起源生物体；其中对于疑似具有共同病原体的个体而言共同的起源生物体是所述病原体。编号实施方案211包括编号实施方案210所述的方法，其中所述起源生物体的序列读取映射至已知病原体。编号实施方案212包括编号实施方案210-211中任一实施方案所述的方法，其中所述起源生物体的序列读取在序列数据库搜索时鉴定已知病原体。编号实施方案213包括编号实施方案210-212中任一实施方案所述的方法，其中所述起源生物体的序列读取不存在于多个成对序列读取中，所述多个成对序列读取从获自非疑似具有共同病原体的多个个体中的每一个的稳定化的样品获得。编号实施方案214包括编号实施方案210-213中任一实施方案所述的方法，其中所述起源生

物体的序列读取鉴定在序列数据库中没有表示出的生物体。编号实施方案215包括编号实施方案210-214中任一实施方案所述的方法，其中所述稳定化的样品已进行交联。编号实施方案216包括编号实施方案210-215中任一实施方案所述的方法，其中所述稳定化的样品已与甲醛接触。编号实施方案217包括编号实施方案210-215中任一实施方案所述的方法，其中所述稳定化的样品已与补骨脂素接触。编号实施方案218包括编号实施方案210-215中任一实施方案所述的方法，其中所述稳定化的样品已暴露于UV辐射。编号实施方案219包括编号实施方案210-218中任一实施方案所述的方法，其中所述样品已与DNA结合部分接触。编号实施方案220包括编号实施方案210-219中任一实施方案所述的方法，其中所述DNA结合部分包括组蛋白。编号实施方案221包括编号实施方案210-220中任一实施方案所述的方法，其中处理所述稳定化的样品以裂解双链DNA包括使所述样品与限制性内切核酸酶接触。编号实施方案222包括编号实施方案210-221中任一实施方案所述的方法，其中处理所述稳定化的样品以裂解双链DNA包括对所述样品进行声处理。编号实施方案223包括编号实施方案210-22中任一实施方案所述的方法，其中标记暴露的DNA末端包括向暴露的DNA末端添加生物素部分。编号实施方案224包括编号实施方案210-223中任一实施方案所述的方法，其中所述样品来源于血液、汗液、尿液或粪便。编号实施方案225包括编号实施方案210-224中任一实施方案所述的方法，其中所述方法在不超过2天内完成。编号实施方案226包括编号实施方案210-225中任一实施方案所述的方法，其中完成所述方法所需的操作时间的量不多于6小时。编号实施方案227包括编号实施方案210-226中任一实施方案所述的方法，其中所述方法包括使用SPRI珠子。编号实施方案228包括编号实施方案210-227中任一实施方案所述的方法，其中所述稳定化的样品包含不超过约5微克的DNA。

[0435] 编号实施方案229包括鉴定抗生素抗性基因的微生物宿主的方法，该方法包括：a)从患有显示微生物抗生素抗性的病况的个体获得稳定化的样品；b)处理所述稳定化的样品以裂解所述稳定化的样品中的双链DNA；c)标记暴露的DNA末端；d)连接标记的暴露的DNA末端以形成标记的成对末端；以及e)在标记的成对末端进行测序以生成成对序列；其中与抗生素抗性基因序列相邻的序列指示抗生素抗性基因的微生物宿主。编号实施方案230包括编号实施方案229所述的方法，其中所述稳定化的样品已进行交联。编号实施方案231包括编号实施方案229-230中任一实施方案所述的方法，其中所述稳定化的样品已与甲醛接触。编号实施方案232包括编号实施方案229-230中任一实施方案所述的方法，其中所述稳定化的样品已与补骨脂素接触。编号实施方案233包括编号实施方案229-230中任一实施方案所述的方法，其中所述稳定化的样品已暴露于UV辐射。编号实施方案234包括编号实施方案229-233中任一实施方案所述的方法，其中所述样品已与DNA结合部分接触。

[0436] 编号实施方案235包括编号实施方案229-234中任一实施方案所述的方法，其中所述DNA结合部分包括组蛋白。编号实施方案236包括编号实施方案229-235中任一实施方案所述的方法，其中处理所述稳定化的样品以裂解双链DNA包括使所述样品与限制性内切核酸酶接触。编号实施方案237包括编号实施方案229-236中任一实施方案所述的方法，其中处理所述稳定化的样品以裂解双链DNA包括对所述样品进行声处理。编号实施方案238包括编号实施方案229-237中任一实施方案所述的方法，其中标记暴露的DNA末端包括向暴露的DNA末端添加生物素部分。编号实施方案239包括编号实施方案229-238中任一实施方案所述的方法，其包括在DNA数据库中搜索所述成对序列。编号实施方案240包括编号实施方案

229-239中任一实施方案所述的方法，其中所述方法在不超过2天内完成。编号实施方案241包括编号实施方案229-240中任一实施方案所述的方法，其中完成所述方法所需的操作时间的量不多于6小时。编号实施方案242包括编号实施方案229-241中任一实施方案所述的方法，其中所述方法包括使用SPRI珠子。编号实施方案243包括编号实施方案229-242中任一实施方案所述的方法，其中所述稳定化的样品包含不超过约5微克的DNA。

[0437] 编号实施方案244包括确定异质核酸样品的基因组连接信息的方法，该方法包括：(a)获得稳定化的异质核酸样品；(b)处理所述稳定化的样品以裂解所述稳定化的样品中的双链DNA；(c)标记暴露的DNA末端；(d)连接标记的暴露的DNA末端以形成标记的成对末端；(e)在标记的成对末端进行测序以生成多个成对序列读取；(f)将所述多个序列读取的成对序列读取的每一半分配给共同的起源核酸分子。编号实施方案245包括编号实施方案244所述的方法，其中从血液、汗液、尿液或粪便获得所述异质核酸样品。编号实施方案246包括编号实施方案244-245中任一实施方案所述的方法，其中所述稳定化的样品已进行交联。编号实施方案247包括编号实施方案244-246中任一实施方案所述的方法，其中所述稳定化的样品已与甲醛接触。编号实施方案248包括编号实施方案244-246中任一实施方案所述的方法，其中所述稳定化的样品已与补骨脂素接触。编号实施方案249包括编号实施方案244-246中任一实施方案所述的方法，其中所述稳定化的样品已暴露于UV辐射。编号实施方案250包括编号实施方案244-249中任一实施方案所述的方法，其中所述样品已与DNA结合部分接触。编号实施方案251包括编号实施方案244-250中任一实施方案所述的方法，其中所述DNA结合部分包括组蛋白。编号实施方案252包括编号实施方案244-251中任一实施方案所述的方法，其中处理所述稳定化的样品以裂解双链DNA包括使所述样品与限制性内切核酸酶接触。编号实施方案253包括编号实施方案244-252中任一实施方案所述的方法，其中处理所述稳定化的样品以裂解双链DNA包括对所述样品进行声处理。编号实施方案254包括编号实施方案244-253中任一实施方案所述的方法，其中标记暴露的DNA末端包括向暴露的DNA末端添加生物素部分。编号实施方案255包括编号实施方案244-254中任一实施方案所述的方法，其包括在DNA数据库中搜索成对序列。编号实施方案256包括编号实施方案244-255中任一实施方案所述的方法，其中所述共同的起源核酸分子映射至单个个体。编号实施方案257包括编号实施方案244-256中任一实施方案所述的方法，其中所述共同的起源核酸分子鉴定群体的亚组。编号实施方案258包括编号实施方案244-257中任一实施方案所述的方法，其中所述方法在不超过2天内完成。编号实施方案259包括编号实施方案244-258中任一实施方案所述的方法，其中完成所述方法所需的操作时间的量不多于6小时。编号实施方案260包括编号实施方案244-259中任一实施方案所述的方法，其中所述方法包括使用SPRI珠子。编号实施方案261包括编号实施方案244-260中任一实施方案所述的方法，其中所述稳定化的样品包含不超过约5微克的DNA。

[0438] 编号实施方案262包括用于宏基因组组装的方法，该方法包括：(a)从环境中收集微生物；(b)从所述微生物获得多个叠连群；(c)由通过探测重构的染色质的物理布局产生的数据生成多个读取对；以及(d)将所述多个读取对映射至所述多个叠连群，从而产生读取映射数据，其中映射至不同叠连群的读取对指示不同的叠连群来自共同的物种。编号实施方案263包括编号实施方案262中任一实施方案所述的方法，其中从人类肠道收集所述微生物。编号实施方案264包括用于检测细菌致病原的方法，该方法包括：(a)从所述细菌致病原

获得多个叠连群;(b)由通过探测重构的染色质的物理布局产生的数据生成多个读取对;(c)将所述多个读取对映射至所述多个叠连群,从而产生读取映射数据;(d)使用所述读取映射数据排列所述叠连群以将所述叠连群组装成基因组组装;以及(e)使用所述基因组组装来确定所述细菌致病原的存在。

[0439] 编号实施方案265包括在宿主群体中检测病原体的方法,该方法包括:a)从疑似具有共同病原体的多个个体中的每一个获得稳定化的样品;b)处理所述稳定化的样品以裂解所述稳定化的样品中的双链DNA;c)使用第一条码标记来标记所述稳定化的样品的第一部分的暴露的DNA末端,并使用第二条码标记来标记所述稳定化的样品的第二部分的暴露的末端;d)在条码标记的末端进行测序以生成多个条码标记的序列读取;以及e)将所述多个序列读取的共同条码标记的序列读取分配给共同的起源生物体;其中对于疑似具有共同病原体的个体而言共同的起源生物体是所述病原体。编号实施方案266包括编号实施方案265所述的方法,其中所述起源生物体的序列读取映射至已知病原体。编号实施方案267包括编号实施方案265-266中任一实施方案所述的方法,其中所述起源生物体的序列读取在序列数据库搜索时鉴定已知病原体。编号实施方案268包括编号实施方案265-267中任一实施方案所述的方法,其中所述起源生物体的序列读取不存在于多个成对序列读取中,所述多个成对序列读取从获自非疑似具有共同病原体的多个个体中的每一个的稳定化的样品获得。编号实施方案269包括编号实施方案265-268中任一实施方案所述的方法,其中所述起源生物体的序列读取鉴定在序列数据库中没有表示出的生物体。编号实施方案270包括编号实施方案265-269中任一实施方案所述的方法,其中所述稳定化的样品已进行交联。编号实施方案271包括编号实施方案265-270中任一实施方案所述的方法,其中所述稳定化的样品已与甲醛接触。编号实施方案272包括编号实施方案265-271中任一实施方案所述的方法,其中所述稳定化的样品已与补骨脂素接触。编号实施方案273包括编号实施方案265-272中任一实施方案所述的方法,其中所述稳定化的样品已暴露于UV辐射。编号实施方案274包括编号实施方案265-273中任一实施方案所述的方法,其中所述样品已与DNA结合部分接触。编号实施方案275包括编号实施方案265-274中任一实施方案所述的方法,其中所述DNA结合部分包括组蛋白。编号实施方案276包括编号实施方案265-275中任一实施方案所述的方法,其中处理所述稳定化的样品以裂解双链DNA包括使所述样品与限制性内切核酸酶接触。编号实施方案277包括编号实施方案265-276中任一实施方案所述的方法,其中处理所述稳定化的样品以裂解双链DNA包括对所述样品进行声处理。编号实施方案278包括编号实施方案265-277中任一实施方案所述的方法,其中标记暴露的DNA末端包括向暴露的DNA末端添加生物素部分。编号实施方案279包括编号实施方案265-278中任一实施方案所述的方法,其中所述样品来源于血液、汗液、尿液或粪便。编号实施方案280包括编号实施方案265-279中任一实施方案所述的方法,其中所述方法在不超过2天内完成。编号实施方案281包括编号实施方案265-280中任一实施方案所述的方法,其中完成所述方法所需的操作时间的量不多于6小时。编号实施方案282包括编号实施方案265-281中任一实施方案所述的方法,其中所述方法包括使用SPRI珠子。编号实施方案283包括编号实施方案265-282中任一实施方案所述的方法,其中所述稳定化的样品包含不超过约5微克的DNA。

[0440] 编号实施方案284包括鉴定抗生素抗性基因的微生物宿主的方法,该方法包括:a)从患有显示微生物抗生素抗性的病况的个体获得稳定化的样品;b)处理所述稳定化的样品

以裂解所述稳定化的样品中的双链DNA;c)使用第一条码标记来标记所述稳定化的样品的第一部分的暴露的DNA末端,并使用第二条码标记来标记所述稳定化的样品的第二部分的暴露的末端;d)在条码标记的末端进行测序以生成多个条码标记的序列读取;其中具有与抗生素抗性基因序列的条码标记相同的条码标记的序列指示抗生素抗性基因的微生物宿主。编号实施方案285包括编号实施方案284所述的方法,其中所述稳定化的样品已进行交联。编号实施方案286包括编号实施方案284-285中任一实施方案所述的方法,其中所述稳定化的样品已与甲醛接触。编号实施方案287包括编号实施方案284-285中任一实施方案所述的方法,其中所述稳定化的样品已与补骨脂素接触。编号实施方案288包括编号实施方案284-285中任一实施方案所述的方法,其中所述稳定化的样品已暴露于UV辐射。编号实施方案289包括编号实施方案284-288中任一实施方案所述的方法,其中所述样品已与DNA结合部分接触。编号实施方案290包括编号实施方案284-289中任一实施方案所述的方法,其中所述DNA结合部分包括组蛋白。编号实施方案291包括编号实施方案284-290中任一实施方案所述的方法,其中处理所述稳定化的样品以裂解双链DNA包括使所述样品与限制性内切核酸酶接触。编号实施方案292包括编号实施方案284-291中任一实施方案所述的方法,其中处理所述稳定化的样品以裂解双链DNA包括对所述样品进行声处理。编号实施方案293包括编号实施方案284-292中任一实施方案所述的方法,其中标记暴露的DNA末端包括向暴露的DNA末端添加生物素部分。编号实施方案294包括编号实施方案284-293中任一实施方案所述的方法,其包括在DNA数据库中搜索成对序列。编号实施方案295包括编号实施方案284-294中任一实施方案所述的方法,其中所述方法在不超过2天内完成。编号实施方案296包括编号实施方案284-295中任一实施方案所述的方法,其中完成所述方法所需的操作时间的量不多于6小时。编号实施方案297包括编号实施方案284-296中任一实施方案所述的方法,其中所述方法包括使用SPRI珠子。编号实施方案298包括编号实施方案284-297中任一实施方案所述的方法,其中所述稳定化的样品包含不超过约5微克的DNA。

[0441] 编号实施方案299包括确定异质核酸样品的基因组连接信息的方法,该方法包括:(a)获得稳定化的异质核酸样品;(b)处理所述稳定化的样品以裂解所述稳定化的样品中的双链DNA;(c)使用第一条码标记来标记所述稳定化的样品的第一部分的暴露的DNA末端,并使用第二条码标记来标记所述稳定化的样品的第二部分的暴露的末端;(d)在条码标记的末端进行测序以生成多个条码标记的序列读取;(e)将共同标记的序列读取分配给共同的起源核酸分子。编号实施方案300包括编号实施方案299所述的方法,其中从血液、汗液、尿液或粪便获得所述异质核酸样品。编号实施方案301包括编号实施方案299-300中任一实施方案所述的方法,其中所述稳定化的样品已进行交联。编号实施方案302包括编号实施方案299-301中任一实施方案所述的方法,其中所述稳定化的样品已与甲醛接触。编号实施方案303包括编号实施方案299-301中任一实施方案所述的方法,其中所述稳定化的样品已与补骨脂素接触。编号实施方案304包括编号实施方案299-301中任一实施方案所述的方法,其中所述稳定化的样品已暴露于UV辐射。编号实施方案305包括编号实施方案299-304中任一实施方案所述的方法,其中所述样品已与DNA结合部分接触。编号实施方案306包括编号实施方案299-305中任一实施方案所述的方法,其中所述DNA结合部分包括组蛋白。编号实施方案307包括编号实施方案299-306中任一实施方案所述的方法,其中处理所述稳定化的样品以裂解双链DNA包括使所述样品与核酸酶接触。编号实施方案308包括编号实施方案299-

307中任一实施方案所述的方法，其中所述核酸酶为限制性内切核酸酶。编号实施方案309包括编号实施方案299-308中任一实施方案所述的方法，其中处理所述稳定化的样品以裂解双链DNA包括对所述样品进行声处理。编号实施方案310包括编号实施方案299-309中任一实施方案所述的方法，其中标记暴露的DNA末端包括向暴露的DNA末端添加生物素部分。编号实施方案311包括编号实施方案299-310中任一实施方案所述的方法，其包括在DNA数据库中搜索成对序列。编号实施方案312包括编号实施方案299-311中任一实施方案所述的方法，其中所述共同的起源核酸分子映射至单个个体。编号实施方案313包括编号实施方案299-312中任一实施方案所述的方法，其中所述共同的起源核酸分子鉴定群体的亚组。编号实施方案314包括编号实施方案299-313中任一实施方案所述的方法，其中所述异质样品包含映射至共同物种的至少两个个体的核酸。编号实施方案315包括编号实施方案299-314中任一实施方案所述的方法，其中所述异质样品包含映射至共同物种的至少三个个体的核酸。编号实施方案316包括编号实施方案299-315中任一实施方案所述的方法，其中所述异质样品包含映射至至少两个物种的核酸。编号实施方案317包括编号实施方案299-316中任一实施方案所述的方法，其中所述异质样品包含映射至至少三个物种的核酸。编号实施方案318包括编号实施方案299-317中任一实施方案所述的方法，其中所述异质样品包含映射至至少四个物种的核酸。编号实施方案319包括编号实施方案299-318中任一实施方案所述的方法，其中在不参考外源序列信息的情况下序列读取组装成至少两个核酸支架。编号实施方案320包括编号实施方案299-319中任一实施方案所述的方法，其中在不参考外源序列信息的情况下序列读取组装成至少三个核酸支架。编号实施方案321包括编号实施方案299-320中任一实施方案所述的方法，其中所述序列读取组装成至少两个核酸支架，使得至少50％的第一基因组和至少50％的第二基因组被表示在所述至少两个核酸支架中。编号实施方案322包括编号实施方案299-321中任一实施方案所述的方法，其中所述序列读取组装成至少两个核酸支架，使得至少60％的第一基因组和至少60％的第二基因组被表示在所述至少两个核酸支架中。编号实施方案323包括编号实施方案299-322中任一实施方案所述的方法，其中所述序列读取组装成至少两个核酸支架，使得至少70％的第一基因组和至少70％的第二基因组被表示在所述至少两个核酸支架中。编号实施方案324包括编号实施方案299-323中任一实施方案所述的方法，其中所述序列读取组装成至少两个核酸支架，使得至少80％的第一基因组和至少80％的第二基因组被表示在所述至少两个核酸支架中。编号实施方案325包括编号实施方案299-324中任一实施方案所述的方法，其中所述方法在不超过2天内完成。编号实施方案326包括编号实施方案299-325中任一实施方案所述的方法，其中完成所述方法所需的操作时间的量不多于6小时。编号实施方案327包括编号实施方案299-326中任一实施方案所述的方法，其中所述方法包括使用SPRI珠子。编号实施方案328包括编号实施方案299-327中任一实施方案所述的方法，其中所述稳定化的样品包含不超过约5微克的DNA。

[0442] 编号实施方案329包括在宿主群体中检测病原体的方法，该方法包括：a)从多个受试者中的每一个获得稳定化的样品；b)处理所述稳定化的样品以裂解所述稳定化的样品中的双链DNA，从而生成暴露的DNA末端；c)标记所述暴露的DNA末端的至少一部分；d)连接所述暴露的DNA末端以形成标记的成对末端；e)对所述标记的成对末端的至少一个可识别部分进行测序以生成多个读取对；以及f)将读取对的每一半分配给共同的起源生物体；其中

对于所述受试者而言共同的起源生物体被检测为所述病原体。编号实施方案330包括编号实施方案329所述的方法，其中所述起源生物体的读取对映射至已知病原体。编号实施方案331包括编号实施方案329-330中任一实施方案所述的方法，其中所述起源生物体的读取对在序列数据库搜索时鉴定已知病原体。编号实施方案332包括编号实施方案329-331中任一实施方案所述的方法，其中所述起源生物体的读取对不存在于多个读取对中，所述多个读取对从获自不具有共同病原体的多个受试者中的每一个的稳定化的样品获得。编号实施方案333包括编号实施方案329-332中任一实施方案所述的方法，其中所述起源生物体的读取对鉴定在序列数据库中没有表示出的生物体。编号实施方案334包括编号实施方案329-333中任一实施方案所述的方法，其中所述稳定化的样品已进行交联。编号实施方案335包括编号实施方案329-334中任一实施方案所述的方法，其中所述稳定化的样品已与甲醛接触。编号实施方案336包括编号实施方案329-334中任一实施方案所述的方法，其中所述稳定化的样品已与补骨脂素接触。编号实施方案337包括编号实施方案329-334中任一实施方案所述的方法，其中所述稳定化的样品已暴露于UV辐射。编号实施方案338包括编号实施方案329-337中任一实施方案所述的方法，其中通过使样品与DNA结合部分接触获得所述稳定化的样品。编号实施方案339包括编号实施方案329-338中任一实施方案所述的方法，其中所述DNA结合部分包括组蛋白。编号实施方案340包括编号实施方案329-339中任一实施方案所述的方法，其中处理所述稳定化的样品以裂解双链DNA包括使所述稳定化的样品与限制性内切核酸酶接触。编号实施方案341包括编号实施方案329-340中任一实施方案所述的方法，其中处理所述稳定化的样品以裂解双链DNA包括对所述稳定化的样品进行声处理。编号实施方案342包括编号实施方案329-341中任一实施方案所述的方法，其中标记暴露的DNA末端包括向暴露的DNA末端添加生物素部分。编号实施方案343包括编号实施方案329-342中任一实施方案所述的方法，其中所述稳定化的样品来源于血液、汗液、尿液或粪便。编号实施方案344包括编号实施方案329-343中任一实施方案所述的方法，其中所述方法在不超过2天内完成。编号实施方案345包括编号实施方案329-344中任一实施方案所述的方法，其中完成所述方法所需的操作时间的量不多于6小时。编号实施方案346包括编号实施方案329-345中任一实施方案所述的方法，其中所述方法包括使用SPRI珠子。编号实施方案347包括编号实施方案329-346中任一实施方案所述的方法，其中所述稳定化的样品包含不超过约5微克的DNA。

[0443] 编号实施方案348包括鉴定抗生素抗性基因的微生物宿主的方法，该方法包括：a)从患有显示微生物抗生素抗性的病况的受试者获得稳定化的样品；b)处理所述稳定化的样品以裂解所述稳定化的样品中的双链DNA，从而生成暴露的DNA末端；c)标记所述暴露的DNA末端的至少一部分；d)连接所述标记的暴露的DNA末端以形成标记的成对末端；以及e)对所述连接的成对末端的至少一个可识别部分进行测序以生成成对序列；其中与抗生素抗性基因序列相邻的成对序列指示抗生素抗性基因的微生物宿主。编号实施方案349包括编号实施方案348所述的方法，其中所述稳定化的样品已进行交联。编号实施方案350包括编号实施方案348-349中任一实施方案所述的方法，其中所述稳定化的样品已与甲醛接触。编号实施方案351包括编号实施方案348-349中任一实施方案所述的方法，其中所述稳定化的样品已与补骨脂素接触。编号实施方案352包括编号实施方案348-349中任一实施方案所述的方法，其中所述稳定化的样品已暴露于UV辐射。编号实施方案353包括编号实施方案348-352

中任一实施方案所述的方法，其中所述样品已与DNA结合部分接触。编号实施方案354包括编号实施方案348-353中任一实施方案所述的方法，其中所述DNA结合部分包括组蛋白。编号实施方案355包括编号实施方案348-354中任一实施方案所述的方法，其中处理所述稳定化的样品以裂解双链DNA包括使所述样品与限制性内切核酸酶接触。编号实施方案356包括编号实施方案348-355中任一实施方案所述的方法，其中处理所述稳定化的样品以裂解双链DNA包括对所述样品进行声处理。编号实施方案357包括编号实施方案348-356中任一实施方案所述的方法，其中标记暴露的DNA末端包括向暴露的DNA末端添加生物素部分。编号实施方案358包括编号实施方案348-357中任一实施方案所述的方法，其包括在DNA数据库中搜索所述成对序列。编号实施方案359包括编号实施方案348-358中任一实施方案所述的方法，其中所述方法在不超过2天内完成。编号实施方案360包括编号实施方案348-359中任一实施方案所述的方法，其中完成所述方法所需的操作时间的量不多于6小时。编号实施方案361包括编号实施方案348-360中任一实施方案所述的方法，其中所述方法包括使用SPRI珠子。编号实施方案362包括编号实施方案348-361中任一实施方案所述的方法，其中所述稳定化的样品包含不超过约5微克的DNA。

[0444] 编号实施方案363包括确定异质核酸样品的基因组连接信息的方法，该方法包括：(a)使所述异质核酸样品稳定化；(b)处理所述稳定化的样品以裂解所述稳定化的样品中的双链DNA，从而生成暴露的DNA末端；(c)标记所述暴露的DNA末端的至少一部分；(d)连接所述标记的暴露的DNA末端以形成标记的成对末端；(e)对所述标记的成对末端的至少一个可识别部分进行测序以生成多个读取对；(f)将读取对的每一半分配给共同的起源核酸分子。编号实施方案364包括编号实施方案363所述的方法，其中从血液、汗液、尿液或粪便获得所述异质核酸样品。编号实施方案365包括编号实施方案363-364中任一实施方案所述的方法，其中所述稳定化的样品已进行交联。编号实施方案366包括编号实施方案363-365中任一实施方案所述的方法，其中所述稳定化的样品已与甲醛接触。编号实施方案367包括编号实施方案363-365中任一实施方案所述的方法，其中所述稳定化的样品已与补骨脂素接触。编号实施方案368包括编号实施方案363-365中任一实施方案所述的方法，其中所述稳定化的样品已暴露于UV辐射。编号实施方案369包括编号实施方案363-368中任一实施方案所述的方法，其中所述样品已与DNA结合部分接触。编号实施方案370包括编号实施方案363-369中任一实施方案所述的方法，其中所述DNA结合部分包括组蛋白。编号实施方案371包括编号实施方案363-370中任一实施方案所述的方法，其中处理所述稳定化的样品以裂解双链DNA包括使所述样品与限制性内切核酸酶接触。编号实施方案372包括编号实施方案363-371中任一实施方案所述的方法，其中处理所述稳定化的样品以裂解双链DNA包括对所述样品进行声处理。编号实施方案373包括编号实施方案363-372中任一实施方案所述的方法，其中标记暴露的DNA末端包括向暴露的DNA末端添加生物素部分。编号实施方案374包括编号实施方案363-373中任一实施方案所述的方法，其中在DNA数据库中搜索成对序列。编号实施方案375包括编号实施方案363-374中任一实施方案所述的方法，其中所述共同的起源核酸分子映射至单个个体。编号实施方案376包括编号实施方案363-375中任一实施方案所述的方法，其中所述共同的起源核酸分子鉴定群体的亚组。编号实施方案377包括编号实施方案363-376中任一实施方案所述的方法，其中所述异质样品包含映射至共同物种的至少两个个体的核酸。编号实施方案378包括编号实施方案363-377中任一实施方案所述的方

法，其中所述异质样品包含映射至共同物种的至少三个个体的核酸。编号实施方案379包括编号实施方案363-378中任一实施方案所述的方法，其中所述异质样品包含映射至至少两个物种的核酸。编号实施方案380包括编号实施方案363-379中任一实施方案所述的方法，其中所述异质样品包含映射至至少三个物种的核酸。编号实施方案381包括编号实施方案363-380中任一实施方案所述的方法，其中所述异质样品包含映射至至少四个物种的核酸。编号实施方案382包括编号实施方案363-381中任一实施方案所述的方法，其中在不参考外源序列信息的情况下序列读取组装成至少两个核酸支架。编号实施方案383包括编号实施方案363-382中任一实施方案所述的方法，其中在不参考外源序列信息的情况下序列读取组装成至少三个核酸支架。编号实施方案384包括编号实施方案363-383中任一实施方案所述的方法，其中所述序列读取组装成至少两个核酸支架，使得至少50％的第一基因组和至少50％的第二基因组被表示在所述至少两个核酸支架中。编号实施方案385包括编号实施方案363-384中任一实施方案所述的方法，其中所述序列读取组装成至少两个核酸支架，使得至少60％的第一基因组和至少60％的第二基因组被表示在所述至少两个核酸支架中。编号实施方案386包括编号实施方案363-385中任一实施方案所述的方法，其中所述序列读取组装成至少两个核酸支架，使得至少70％的第一基因组和至少70％的第二基因组被表示在所述至少两个核酸支架中。编号实施方案387包括编号实施方案363-386中任一实施方案所述的方法，其中所述序列读取组装成至少两个核酸支架，使得至少80％的第一基因组和至少80％的第二基因组被表示在所述至少两个核酸支架中。编号实施方案388包括编号实施方案363-387中任一实施方案所述的方法，其中所述方法在不超过2天内完成。编号实施方案389包括编号实施方案363-388中任一实施方案所述的方法，其中完成所述方法所需的操作时间的量不多于6小时。编号实施方案390包括编号实施方案363-389中任一实施方案所述的方法，其中所述方法包括使用SPRI珠子。编号实施方案391包括编号实施方案363-390中任一实施方案所述的方法，其中所述稳定化的样品包含不超过约5微克的DNA。

[0445] 编号实施方案392包括用于宏基因组组装的方法，该方法包括：(a)从环境中收集微生物；(b)从所述微生物获得多个叠连群；(c)由通过探测重构的染色质的物理布局产生的数据生成多个读取对；以及(d)将所述多个读取对映射至所述多个叠连群，从而产生读取映射数据，其中映射至不同叠连群的读取对指示不同的叠连群来源于共同的个体。编号实施方案393包括编号实施方案392中任一实施方案所述的方法，其中从人类肠道收集所述微生物。编号实施方案394包括编号实施方案392所述的方法，其中从人类皮肤收集所述微生物。编号实施方案395包括编号实施方案392-394中任一实施方案所述的方法，其中从有毒废弃物收集所述微生物。编号实施方案396包括编号实施方案392-395中任一实施方案所述的方法，其中从分解的木材或纤维素收集所述微生物。编号实施方案397包括编号实施方案392-396中任一实施方案所述的方法，其中从水生环境收集所述微生物。编号实施方案398包括编号实施方案392-397中任一实施方案所述的方法，其中从海底收集所述微生物。编号实施方案399包括编号实施方案392-398中任一实施方案所述的方法，其中从陆地环境收集所述微生物。编号实施方案400包括编号实施方案392-399中任一实施方案所述的方法，其中从生物环境收集所述微生物。编号实施方案401包括编号实施方案392-400中任一实施方案所述的方法，其中所述异质样品包含映射至共同物种的至少两个个体的核酸。编号实施方案402包括编号实施方案392-401中任一实施方案所述的方法，其中所述异质样品包含映

射至共同物种的至少三个个体的核酸。编号实施方案403包括编号实施方案392-402中任一实施方案所述的方法，其中所述异质样品包含映射至至少两个物种的核酸。编号实施方案404包括编号实施方案392-403中任一实施方案所述的方法，其中所述异质样品包含映射至至少三个物种的核酸。编号实施方案405包括编号实施方案392-404中任一实施方案所述的方法，其中所述异质样品包含映射至至少四个物种的核酸。编号实施方案406包括编号实施方案392-405中任一实施方案所述的方法，其中在不参考外源序列信息的情况下序列读取组装成至少两个核酸支架。编号实施方案407包括编号实施方案392-406中任一实施方案所述的方法，其中在不参考外源序列信息的情况下序列读取组装成至少三个核酸支架。编号实施方案408包括编号实施方案392-407中任一实施方案所述的方法，其中所述序列读取组装成至少两个核酸支架，使得至少50％的第一基因组和至少50％的第二基因组被表示在所述至少两个核酸支架中。编号实施方案409包括编号实施方案392-408中任一实施方案所述的方法，其中所述序列读取组装成至少两个核酸支架，使得至少60％的第一基因组和至少60％的第二基因组被表示在所述至少两个核酸支架中。编号实施方案410包括编号实施方案392-409中任一实施方案所述的方法，其中所述序列读取组装成至少两个核酸支架，使得至少70％的第一基因组和至少70％的第二基因组被表示在所述至少两个核酸支架中。编号实施方案411包括编号实施方案392-410中任一实施方案所述的方法，其中所述序列读取组装成至少两个核酸支架，使得至少80％的第一基因组和至少80％的第二基因组被表示在所述至少两个核酸支架中。编号实施方案412包括编号实施方案392-411中任一实施方案所述的方法，其中所述方法包括使用SPRI珠子。编号实施方案413包括编号实施方案392-412中任一实施方案所述的方法，其中所述稳定化的样品包含不超过约5微克的DNA。

[0446] 编号实施方案414包括用于检测细菌致病原的方法，该方法包括：(a)从所述细菌致病原获得多个叠连群；(b)由通过探测重构的染色质的物理布局产生的数据生成多个读取对；(c)将所述多个读取对映射至所述多个叠连群，从而产生读取映射数据；(d)使用所述读取映射数据排列所述叠连群以将所述叠连群组装成基因组组装；以及(e)使用所述基因组组装来确定所述细菌致病原的存在。

[0447] 编号实施方案415包括从生物体获得基因组序列信息的方法，该方法包括：(a)从所述生物体获得稳定化的样品；(b)处理所述稳定化的样品以裂解所述稳定化的样品中的双链DNA，从而生成暴露的DNA末端；(c)标记所述暴露的DNA末端的至少一部分，以生成标记的DNA区段；(d)对所述标记的DNA区段的至少一个可识别部分进行测序，从而获得标记的序列；以及(e)映射所述标记的序列以生成所述生物体的基因组序列信息，其中所述基因组序列信息覆盖所述生物体的基因组的至少75％。编号实施方案416包括编号实施方案415所述的方法，其中所述异质样品包含映射至共同物种的至少两个个体的核酸。编号实施方案417包括编号实施方案415-416中任一实施方案所述的方法，其中所述异质样品包含映射至共同物种的至少三个个体的核酸。编号实施方案418包括编号实施方案415-417中任一实施方案所述的方法，其中所述异质样品包含映射至至少两个物种的核酸。编号实施方案419包括编号实施方案415-418中任一实施方案所述的方法，其中所述异质样品包含映射至至少三个物种的核酸。编号实施方案420包括编号实施方案415-419中任一实施方案所述的方法，其中所述异质样品包含映射至至少四个物种的核酸。编号实施方案421包括编号实施方案415-420中任一实施方案所述的方法，其中在不参考外源序列信息的情况下序列读取组装

成至少两个核酸支架。编号实施方案422包括编号实施方案415-421中任一实施方案所述的方法，其中在不参考外源序列信息的情况下序列读取组装成至少三个核酸支架。编号实施方案423包括编号实施方案415-422中任一实施方案所述的方法，其中所述序列读取组装成至少两个核酸支架，使得至少50％的第一基因组和至少50％的第二基因组被表示在所述至少两个核酸支架中。编号实施方案424包括编号实施方案415-423中任一实施方案所述的方法，其中所述序列读取组装成至少两个核酸支架，使得至少60％的第一基因组和至少60％的第二基因组被表示在所述至少两个核酸支架中。编号实施方案425包括编号实施方案415-424中任一实施方案所述的方法，其中所述序列读取组装成至少两个核酸支架，使得至少70％的第一基因组和至少70％的第二基因组被表示在所述至少两个核酸支架中。编号实施方案426包括编号实施方案415-425中任一实施方案所述的方法，其中所述序列读取组装成至少两个核酸支架，使得至少80％的第一基因组和至少80％的第二基因组被表示在所述至少两个核酸支架中。编号实施方案427包括编号实施方案415-426中任一实施方案所述的方法，其中从异质样品收集所述生物体。编号实施方案428包括编号实施方案415-427中任一实施方案所述的方法，其中所述异质样品包含至少1000种生物体，每一种包含不同的基因组。编号实施方案429包括编号实施方案415-428中任一实施方案所述的方法，其中通过使来自所述生物体的DNA与DNA结合部分接触获得所述稳定化的样品。编号实施方案430包括编号实施方案415-429中任一实施方案所述的方法，其中所述DNA结合部分为组蛋白。编号实施方案431包括编号实施方案415-429中任一实施方案所述的方法，其中所述DNA结合部分为纳米颗粒。编号实施方案432包括编号实施方案415-429中任一实施方案所述的方法，其中所述DNA结合部分为转座酶。编号实施方案433包括编号实施方案415-432中任一实施方案所述的方法，其中使用转座酶标记所述暴露的DNA末端。编号实施方案434包括编号实施方案415-433中任一实施方案所述的方法，其中通过连接所述暴露的DNA末端与另一个暴露的DNA末端来标记暴露的DNA末端的所述部分。编号实施方案435包括编号实施方案415-434中任一实施方案所述的方法，其中使用连接酶将暴露的DNA末端的所述部分与所述其他暴露的DNA末端连接。编号实施方案436包括编号实施方案415-435中任一实施方案所述的方法，其中在不使用从所述基因组获得的附加叠连群序列的情况下生成所述基因组序列信息。编号实施方案437包括编号实施方案415-436中任一实施方案所述的方法，其中所述方法包括使用SPRI珠子。编号实施方案438包括编号实施方案415-437中任一实施方案所述的方法，其中所述稳定化的样品包含不超过约5微克的DNA。

[0448] 编号实施方案439包括分析样品的方法，该方法包括：(a)从多个生物体获得包含核酸的稳定化的样品；(b)处理所述稳定化的样品以裂解所述稳定化的样品中的双链DNA，从而产生暴露的DNA末端；(c)连接所述暴露的DNA末端以形成成对末端；(d)在所述成对末端进行测序以生成多个成对序列读取；以及(e)将所述多个序列读取的成对序列读取的每一半分配给共同的起源生物体。编号实施方案440包括编号实施方案439所述的方法，其进一步包括在所述连接之前标记所述暴露的DNA末端。编号实施方案441包括编号实施方案439-440中任一实施方案所述的方法，其中起源生物体的序列读取鉴定在序列数据库中没有表示出的生物体。编号实施方案442包括编号实施方案439-441中任一实施方案所述的方法，其进一步包括将所述序列读取组装成在序列数据库中没有表示出的基因序列。编号实施方案443包括编号实施方案439-442中任一实施方案所述的方法，其进一步包括根据所述

分配生成所述样品的标签。编号实施方案444包括编号实施方案439-443中任一实施方案所述的方法，其中所述标签指示所述样品的微生物环境。编号实施方案445包括编号实施方案439-444中任一实施方案所述的方法，其进一步包括根据所述分配鉴定一种或多种个体生物体的存在。编号实施方案446包括编号实施方案439-445中任一实施方案所述的方法，其中所述一种或多种个体生物体为人。编号实施方案447包括编号实施方案439-446中任一实施方案所述的方法，其中所述稳定化的样品已进行交联。编号实施方案448包括编号实施方案439-447中任一实施方案所述的方法，其中所述稳定化的样品已与甲醛接触。编号实施方案449包括编号实施方案439-447中任一实施方案所述的方法，其中所述稳定化的样品已与补骨脂素接触。编号实施方案450包括编号实施方案439-447中任一实施方案所述的方法，其中所述稳定化的样品已暴露于UV辐射。编号实施方案451包括编号实施方案439-450中任一实施方案所述的方法，其中所述样品已与DNA结合部分接触。编号实施方案452包括编号实施方案439-451中任一实施方案所述的方法，其中所述DNA结合部分包括组蛋白。编号实施方案453包括编号实施方案439-452中任一实施方案所述的方法，其中所述处理所述稳定化的样品以裂解双链DNA包括使所述样品与核酸酶接触。编号实施方案454包括编号实施方案439-453中任一实施方案所述的方法，其中所述核酸酶为内切核酸酶。编号实施方案455包括编号实施方案439-454中任一实施方案所述的方法，其中所述内切核酸酶为限制性内切核酸酶。编号实施方案456包括编号实施方案439-455中任一实施方案所述的方法，其中所述核酸酶为核酸引导的核酸酶。编号实施方案457包括编号实施方案439-456中任一实施方案所述的方法，其中所述异质样品包含映射至共同物种的至少两个个体的核酸。编号实施方案458包括编号实施方案439-457中任一实施方案所述的方法，其中所述异质样品包含映射至共同物种的至少三个个体的核酸。编号实施方案459包括编号实施方案439-458中任一实施方案所述的方法，其中所述异质样品包含映射至至少两个物种的核酸。编号实施方案460包括编号实施方案439-459中任一实施方案所述的方法，其中所述异质样品包含映射至至少三个物种的核酸。编号实施方案461包括编号实施方案439-460中任一实施方案所述的方法，其中所述异质样品包含映射至至少四个物种的核酸。编号实施方案462包括编号实施方案439-461中任一实施方案所述的方法，其中在不参考外源序列信息的情况下序列读取组装成至少两个核酸支架。编号实施方案463包括编号实施方案439-462中任一实施方案所述的方法，其中在不参考外源序列信息的情况下序列读取组装成至少三个核酸支架。编号实施方案464包括编号实施方案439-463中任一实施方案所述的方法，其中所述序列读取组装成至少两个核酸支架，使得至少50％的第一基因组和至少50％的第二基因组被表示在所述至少两个核酸支架中。编号实施方案465包括编号实施方案439-464中任一实施方案所述的方法，其中所述序列读取组装成至少两个核酸支架，使得至少60％的第一基因组和至少60％的第二基因组被表示在所述至少两个核酸支架中。编号实施方案466包括编号实施方案439-465中任一实施方案所述的方法，其中所述序列读取组装成至少两个核酸支架，使得至少70％的第一基因组和至少70％的第二基因组被表示在所述至少两个核酸支架中。编号实施方案467包括编号实施方案439-466中任一实施方案所述的方法，其中所述序列读取组装成至少两个核酸支架，使得至少80％的第一基因组和至少80％的第二基因组被表示在所述至少两个核酸支架中。编号实施方案468包括编号实施方案439-467中任一实施方案所述的方法，其中所述处理所述稳定化的样品以裂解双链DNA包括对所述样品进行声处理。编

号实施方案469包括编号实施方案439-468中任一实施方案所述的方法，其中所述标记暴露的DNA末端包括向暴露的DNA末端添加生物素部分。编号实施方案470包括编号实施方案439-469中任一实施方案所述的方法，其中所述方法包括使用SPRI珠子。编号实施方案471包括编号实施方案439-470中任一实施方案所述的方法，其中所述稳定化的样品包含不超过约5微克的DNA。

[0449] 实施例

[0450] 给出以下实施例的目的是为了说明本发明的各种实施方案，而不意味着以任何方式限制本发明。本发明实施例以及本文所述的方法目前是优选实施方案的代表，是示例性的，而并非旨在限制本发明的范围。本领域技术人员将会想到包含在由权利要求范围所限定的本发明精神内的其中的变化以及其他用途。

[0451] 实施例1.体外生成染色质的方法

[0452] 特别关注两种重构染色质的方法：一种方法是不依赖于ATP将组蛋白随机沉积到DNA上，而另一种方法使用周期性核小体的ATP依赖性组装。本公开内容允许将以上任一种方法与本文所公开的一种或多种方法一起使用。这两种生成染色质的方法的实例均可以在Lusser等人("Strategies for the reconstitution of chromatin,"Nature Methods (2004),1(1):19-26)中找到，其通过引用以其全文并入本文，包括其中引用的参考文献。

[0453] 使用来自受试者的包含基因组核酸的样品来制备核酸文库，并随后对该文库进行测序。作为实例，从人类样品收集基因组核酸。使用来自人类受试者的50kb样品作为阳性对照。通常，同时制备多个样品以生成多个文库。在一些情况下，一次制备4个样品和50kb人类对照。在一些情况下，一次制备9个样品和50kb人类对照。在一些情况下，制备12、15、20个或更多个样品。

[0454] 反应参数如下：将来自Active Motif染色质组装试剂盒的一组组分在冰上的硅化管中混合。在一些情况下，制备总反应体积的1.25倍的混合物。通常，将约2.1μl的h-Nap-1添加到约2.7μl的核心组蛋白(Core Histones)和约15μl的高盐缓冲液(High Salt Buffer)中以生成溶液A。将溶液A的组分混合并在冰上温育约15分钟。通过在冰上混合制备10X ATP再生体系(ATP Regeneration System)的混合物。简言之，将约15μl的10X ATPRegen缓冲液添加到约0.45μl的肌酸激酶中以生成溶液B，并在冰上混合。

[0455] 在冰上温育溶液A后，将约96.45μl的低盐缓冲液(Low Salt Buffer)添加到约3.75μl的溶液B中，再添加到约15μl的10X ATP再生体系中以生成溶液B。将溶液B混合，并将约135μl的该混合液分配至约1.5μg的DNA以生成溶液C。将水添加到溶液4以得到约150μl的最终体积。将溶液C混合并在27℃下温育过夜。在一些实例中，将溶液C混合并在27℃下温育至多、至少或约12小时、约14小时、约18小时、约20小时或约24小时。在其他实例中，将溶液C混合并在27℃下温育1天、2天、3天、4天、5天、6天、7天、8天、9天、10天或更多天。

[0456] 在27℃下温育过夜后，收集大约10μl的溶液C并转移至硅化管中。保存收集的溶液C以用于测试染色质组装的效率。通常，在MboI消化过程中通过MNase消化实现该测试。

[0457] 实施例2.缓冲液和溶液

[0458] 此处所述的缓冲液和溶液可以通过以下参数来制备：

[0459] SPRI重构缓冲液：SPRI重构缓冲液通常如下制备：将9g的PEG 8000粉末添加到约10ml的1M NaCl中。添加一定量的水以使混合物达到50ml。通常，PEG 8000粉末的工作浓度

约为18％且NaCl约为1M。

[0460] 洗涤缓冲液:洗涤缓冲液通常如下制备:将约500μl的1M Tris-Cl pH8.0添加到约500μl 5M NaCl中。添加一定量的水以使混合物达到50ml。在一些情况下,Tris-Cl pH8.0的工作浓度约为10mM且NaCl约为100mM。

[0461] LWB:LWB通常如下制备:将约500μl的1M Tris-Cl pH8.0添加到约12.5ml 4M LiCl、约100μl 0.5M EDTA和约200μl 10％吐温20中。添加一定量的水以使混合物达到50ml。在某些情况下,Tris-Cl pH8.0的工作浓度为10mM,LiCl为1M,EDTA为1mM且吐温20为0.05％。

[0462] NWB:NWB通常如下制备:将约500μl的1M Tris-Cl pH8.0添加到约10ml的56M NaCl、约100μl的0.5M EDTA和约200μl的10％吐温20中。添加一定量的水以使混合物达到50ml。在各种情况下,Tris-Cl pH8.0的工作浓度为10mM,NaCl的工作浓度为1M,EDTA为1mM且吐温20为0.05％。

[0463] 实施例3.基于染色质捕获来捕获读取对的方法

[0464] 将来自人类受试者的基因组片段化成具有500kb大小的伪叠连群。使用基于染色质捕获的方法,通过探测活细胞内染色体的物理布局生成多个读取对。可以使用任何数目的基于染色质捕获的方法生成读取对,包括Lieberman-Aiden等人("Comprehensive mapping of long range interactions reveals folding principles of the human genome,"Science(2009),326(5950):289-293)中提出的方法,其以全文并入本文,包括其中引用的参考文献。

[0465] 在各种情况下,染色质组装体用甲醛交联。通常,将约4.05μl的约37％的甲醛添加到温育的溶液C中,该混合物在室温下温育约15分钟,随后添加约8.1μl的2.5M甘氨酸以生成溶液D。将溶液D混合并在冰上温育约10分钟。

[0466] 在甲醛交联后,将包含交联染色质的溶液D添加到约330μl的在约18％的PEG 8000/1M NaCl中重建的GE SPRI珠子中,混合并静置以供温育。去除上清液。用约400μl 1X 10mM Tris/50mM NaCl洗涤珠子至少两次。去除上清液并使珠子干燥。在一个实例中,使珠子风干。

[0467] 接下来,制备用于酶消化的溶液。向约175μl水中添加约20μl的10X NEB CutSmart缓冲液和约5μl的NEB MboI并混合以生成溶液E。将大约200μl的溶液E添加到干燥的珠子中,并在37℃下温育约60分钟。在一些实例中,温育在37℃下进行至多、至少或约30分钟、约60分钟、约90分钟、约120分钟、约180分钟或约240分钟。在某些实例中,温育在4℃下进行至多、至少或约1小时、约2小时、约6小时、约12小时、约14小时、约16小时或约24小时。在各种实例中,温育在4℃下进行至多、至少或约1小时、约2小时、约6小时、约12小时,在4℃下进行至多、至少或约1天、约2天、约5天或约10天。

[0468] 酶消化后,处理温育后的珠子以用于缓冲液更换。简言之,将磁体放在包含溶液E和珠子的混合物上,并弃去上清液。用约400μl的1X 10mM Tris/50mM NaCl将沉淀物洗涤至少两次。在一个实例中,使沉淀物/洗涤的珠子风干。

[0469] 制备溶液以用于末端补平,并将生物素添加到珠子。简言之,将约160μl水添加到约20μl的10X NEB缓冲液#2、约1μl的10mM dATP、约1μl的10mM dTTP、约1μl 10mM dGTP、约8μl 10mM生物素-dCTP和约2.5μl NEB Klenow 5U/μl中以生成溶液F。将大约200μl的溶液F

添加到珠子，然后使其在25℃下温育约40分钟。在一个实例中，包含溶液F和珠子的混合物在25℃下温育至多、至少或约30分钟、约60分钟、约120分钟或约180分钟。

[0470] 然后通过缓冲液更换处理珠子。将磁体添加到溶液F和珠子的混合物中，并弃去上清液。用约400μl 1X 10mM Tris/50mM NaCl将沉淀物洗涤至少两次。在一个实例中，使沉淀物/洗涤的珠子风干。

[0471] 然后处理样品以进行聚集体内DNA末端连接。简言之，将约870μl水添加到约100μl的10X T4连接酶缓冲液、约50μl Thermo BSA20mg/ml、约25μl的10％Triton X-100和约0.5μl的NEB T4 DNA连接酶400U/μl中以生成溶液G。然后在约200μl的溶液G中添加经洗涤的珠子，并在16℃下采用设定为约1000RPM的搅拌(Thermo Block摇床)温育过夜。在一个实例中，将经洗涤的珠子和溶液G温育至多、至少或约12小时、约14小时、约16小时、约20小时、约24小时或约48小时。

[0472] 然后处理经温育的珠子用于缓冲液更换。将磁体添加到溶液G和珠子的混合物中，并弃去上清液。随后用约400μl 10mM Tris/50mM NaCl将沉淀物/珠子洗涤至少两次。在一个实例中，使沉淀物/珠子风干。

[0473] 通过用逆向交联处理释放交联组装体中的DNA。制备混合物以用于交联逆转。例如，将约172μl的水添加到约10μl 1M Tris pH8.0、约10μl 20％SDS、约0.5μl 0.1M $CaCl_2$和约5μl NEB蛋白酶K 20mg/ml中以生成溶液I。在一个实例中，该溶液中的每种组分的最终浓度如下：约50mM的Tris pH8.0、约1％的20％SDS、约0.25mM的$CaCl_2$和约0.5mg/ml的NEB蛋白酶K。将大约200μl的溶液I添加到包含交联DNA的珠子，并将混合物在约55℃下温育约15分钟，随后在约68℃下温育约45分钟。

[0474] 将交联的逆转溶液提供给磁珠，并将溶液转移至干净的1.5ml管。向交联的逆转溶液添加约400μl的Normal SPRI珠子，并将该混合物在室温下温育约5分钟。接下来，将磁体添加到该混合物中并弃去上清液。用约400μl的80％乙醇将沉淀物/珠子洗涤至少两次。弃去上清液，并使沉淀物/珠子风干约10-15分钟。最后，用约100μl TE将珠子重悬并温育约2分钟。在Qubit上检测来自交联逆转的DNA的量，并且与起点相比，预期DNA具有至少约30％至约75％的回收率。在一个实例中，从交联逆转中回收超过75％的DNA。

[0475] 为了定量DNA的量和DNA交联逆转的效率，在TapeStation上分析DNA。将约2μl的基因组DNA样品缓冲液分配于8联管PCR条中。简言之，将约2μl的基因组DNA分子量标准物添加到第一管中。将约2μl的Chicago DNA添加到后面的管中。随后封闭管并在TapeStation漩涡振荡器中进行涡旋。然后将基因组DNA带加载到机器中进行分析。

[0476] 使约200ng的DNA经受片段化。将200ng DNA添加到100μl溶液中。将含有DNA的溶液在冰上冷却至少10分钟。将BioRuptor设定在4℃，并将含有DNA的溶液置于BioRuptor上，运行15秒开/90秒关的7个循环。

[0477] 在TapeStation中分析片段化的DNA。将约1μl的片段化DNA在约4μl的TE中稀释，并将2μl的混合物加载到使用高灵敏度D1000芯片的tapestation上。预期有以约350nt为中心的宽分布。

[0478] 然后处理片段化的DNA以用于末端修复。通过将约67.8μl的水添加到约20μl的10X NEB T4连接酶缓冲液、约3.2μl的dNTP 25mM、约1μl的Klenow大片段5U/μl、约3μl的T4 DNA Pol 5U/μl(thermo)和约5μl的T4 PNK 10U/μl(thermo)中制备100μl溶液以生成溶液J。将

约100μl的溶液J添加到具有片段Chicago DNA的管中，并在20℃下温育约20分钟以修复片段化末端。

[0479] 收集约100μl的C1珠子并置于磁体上。取出上清液并丢弃。用约400μl的1X TWB将沉淀物/珠子洗涤至少两次。取出上清液并丢弃。然后将沉淀物/珠子重悬于约200μl的2X NTB中。接下来，将约200μl的末端修复反应液添加到珠子，并将混合物在室温下温育一段时间，使管旋转。将磁体置于溶液上并弃去上清液。用约400μl LWB将沉淀物/珠子洗涤至少1次，随后用约400μl NWB洗涤至少两次，接着用约400μl的10mM Tris/50mM NaCl洗涤至少两次。

[0480] 实施例4.基于染色质捕获方法生成读取对的方法

[0481] 然后将沉淀物/珠子与衔接子连接。通过将约77.5μl的水添加到约20μl的5X Quick连接酶、约1μl的P5/P7衔接子和约2.5μl的NEB T4 DNA连接酶400U/μl中制备衔接子连接溶液。将沉淀物/珠子重悬于约100μl的衔接子连接溶液中。然后将混合物在25℃下温育约30分钟。将磁体置于溶液上，并弃去上清液。用约400μl 10mM Tris/50mM NaCl将沉淀物/珠子洗涤至少两次，随后用约400μl TE洗涤至少两次。

[0482] 通过将约85.25μl的水添加到约10μl的10X Thermo Pol、约1μl的25mM dNTP和约3.75μl的NEB BST Pol 8U/μl中制备用于衔接子补平的溶液。将珠子重悬于约100μl的衔接子补平溶液中并在37℃下温育约20分钟。将磁体添加到混合物中并弃去上清液。用约400μl的10mM Tris/50mM NaCl将沉淀物/珠子洗涤至少两次。

[0483] 通过将约48μl的水与约2μl ISA引物(10mM)和约50μl的2X KAPA MIX混合制备用于索引PCR的溶液。将沉淀物/珠子重悬于约98μl的索引PCR溶液中。向8联管的每一个管中添加约2μl的索引引物。然后盖上管并送去进行采用以下参数的PCR扩增:PCR混合物扩增13个循环，每个循环包括以下步骤:98℃下温育3分钟、98℃下变性20秒、65℃下退火30秒、72℃下延伸30秒、72℃下继续延伸1分钟，最后在12℃下保持直到下一步。在一个实例中，PCR产物在12℃下保持至多、至少或约1小时、2小时、5小时、10小时、15小时、20小时或24小时。在一个实例中，PCR产物在4℃、-20℃、-80℃下、液氮中以玻璃态储存，或在室温下干燥储存。

[0484] 为了纯化扩增的DNA或PCR产物，在新的干净管中合并至少两个PCR反应并放置于磁体上。将溶液转移至干净的1.5ml管中并添加约200μl的Normal SPRI珠子。将含有珠子的混合物在室温下温育约5分钟。将磁体添加到混合物中，并弃去上清液。用约400μl 80％乙醇将沉淀物/珠子洗涤至少两次。弃去上清液。使沉淀物/珠子风干约10-15分钟。然后将沉淀物/珠子重悬于约20μl TE中并温育约2分钟。对重悬的DNA进行定量，例如在宽范围Qubit上。通常，预期浓度为约60ng/μl。

[0485] 分析了索引PCR的DNA产物。首先，通过在约4.5μl的TE中添加约0.5μl的PCR DNA，以1:10稀释DNA。将大约2μl的混合物加载到使用高灵敏度D1000芯片的tapestation上。在某些情况下，预期有以约550nt为中心的宽分布。在一些实例中，按大小选择索引PCR的DNA产物。简言之，用TE使PCR DNA样品达到约30μl(例如添加约18μl的TE)。将约10μl的1.5％DF Pippin Prep样品缓冲液添加到该混合物中。根据制造商手册准备Pippin Prep仪器。将大约40μl的制备的混合物添加到筒匣中。根据在TapeStation分析中观察到的分布中心周围约300nt的宽范围选择DNA大小。通常，DNA的大小约为400-700nt。然后通过使用Qubit高灵

敏度分析对DNA进行定量，预期回收率约为5-10ng/μl。随后通过在4.5TE中添加约0.5μl以1:10稀释DNA。将约2μl的混合物加载到在tapestation上的高灵敏度D1000带上。然后将浓度记录到JIRA中。通常，浓度以pg/μl和摩尔为单位进行记录。

[0486] 在一些情况下，使用酶消化测试染色质组装体的量。一个实例为MNase消化。通常，使用的参数如下所列：通过首先用水将MNase50U/μl稀释至1:10，以1:1000稀释MNase溶液。例如，将约1μl的MNase 50U/μl添加到9μl的水中。通过将1μl的1:10MNase添加到99μl的水中来将稀释的MNase进一步稀释至1:1000。

[0487] 通常通过将约480μl水添加到约5μl 10mM Tris-Cl pH8.0、约5μl 1mM CaCl和约1μl MNase 5mU中，在溶液(例如500μl混合物)中制备MNase消化混合物。一般而言，每种组分的原料浓度为约1MTris-Cl pH8.0、0.1M CaCl和50mU/μl MNase。

[0488] 通过将约362.5μl的水添加到约100μl的10mM EDTA、约25μl的1％SDS和约12.5μl的0.5mg/ml蛋白酶K中来制备终止缓冲液，例如500μl的溶液。在某些情况下，混合物中的每种组分的原料浓度为约0.5M EDTA、约20％SDS和约20mg/ml蛋白酶K。

[0489] 通过MNase消化测试染色质组装体的质量。一般而言，将约45μl的MNase消化混合物分配于1.5ml Eppendorf管中。反应在37℃下预温约2分钟。向每个管中添加大约5μl的组装染色质，并且在添加下一个样品之前温育约15秒。约5分钟后，向样品中添加约50μl的终止缓冲液，从第一管开始，管之间等待约15秒，使得每个样品通常消化约5分钟。随后将样品在37℃下温育约30分钟。在将样品转移至MiniElute Reaction Cleanup柱之前，向温育的样品中添加约300μl的Qiagen Buffer ERC。以下是通常建议的制造程序。通常，将柱子离心约1分钟，并弃去流过液。向每个柱子上添加约700μl的缓冲液PE，随后离心约1分钟，并弃去流过液。通常将柱子离心另外30秒或1分钟以洗脱残留的PE缓冲液。向每个柱子上添加约10μl的EB缓冲液，并通常温育约1分钟。将柱子离心以收集纯化的DNA。为了测试MNase消化的效率，在TapeStation上运行约2μl洗脱的DNA。

[0490] 实施例5.使用读取对的基因组组装

[0491] 将读取对映射至所有伪叠连群，并且使用映射至两个单独的伪叠连群的那些读取对在映射数据的基础上构建邻接矩阵。至少约50％、约60％、约70％、约80％、约90％、约95％或约99％的读取对通过获得读取到伪叠连群边缘的距离的函数来加权，以便在数学上引入根据经验已知的、与较长接触相比更高概率的较短接触。随后，对于每个伪叠连群，分析邻接矩阵以通过找到单个最邻近的伪叠连群来确定穿过伪叠连群的路径，该最邻近的伪叠连群通过具有最高的权重总和来确定。通过执行这些方法，发现所有伪叠连群中的97％以上鉴别出它们的正确邻居。可以进行附加的实验来测试较短叠连群和替代加权对路径查找方案的影响。

[0492] 或者，使用染色质捕获数据的基因组组装可包括利用染色质捕获数据集中的基因组接近信号进行从头基因组组装的超长支架化的计算方法。可以与本文公开的方法一起使用的这类计算方法的实例包括Burton等人(Nature Biotechnology31:1119-1125(2013))的连接相邻染色质方法；和Kaplan等人(Nature Biotechnology 31:1143-47(2013))的DNA三角测量法，这些参考文献以全文以及其中引用的任何参考文献并入本文。此外，应当理解，这些计算方法可以组合使用，包括与本文提供的其他基因组组装方法组合使用。

[0493] 例如，基于Burton等人的连接相邻染色质方法包括以下步骤：(a)使叠连群聚类为

染色体组，(b)对一个或多个染色体组内的叠连群进行排序，并随后(c)将相对方向分配给单个叠连群，该方法可以与本文公开的方法一起使用。对于步骤(a)，使用分层聚类将叠连群放入组中。建立图，其中每个节点最初代表一个叠连群，并且节点之间的每个边缘具有等于连接这两个叠连群的染色质捕获读取对的数目的权重。使用具有平均连接度量的分层凝聚聚类将叠连群合并在一起，应用该分层凝聚聚类直至组的数目减少到不同染色体的预期数目(仅对具有多于一个叠连群的组进行计数)。不对重复叠连群(所述叠连群与其他叠连群的平均连接密度(通过限制性片段位点的数目进行归一化)大于平均连接密度的两倍)和具有太少限制性片段位点的叠连群进行聚类。然而，在聚类之后，如果这些叠连群与一个组的平均连接密度大于其与任何其他组的平均连接密度的四倍，则将这些叠连群中的每一个分配给该组。对于步骤(b)，像在聚类步骤中那样建立图，但是节点之间的边缘权重等于叠连群之间染色质捕获连接数的倒数，其通过每个叠连群的限制性片段位点的数目进行归一化。该图中不包括短叠连群。针对该图计算最小生成树。找到该树中的最长路径，即“树干”。然后对生成树进行修饰，以便以试探性地保持低的总边缘权重的方式通过向树干添加与树干相邻的叠连群来延长该树干。在为每个组找到延长的树干后，将其如下转换为完整的排序。从生成树中移除树干，留下一组含有不在树干中的所有叠连群的“分支”。将这些分支重新插入树干，首先插入最长的分支，其中选择插入位点，以便使排序中相邻叠连群之间的连接数目最大化。没有重新插入短片段;因此，许多聚类的小叠连群被排除在最终的组装体之外。对于步骤(c)，通过考虑每个叠连群上染色质捕获连接比对的确切位置来确定每个叠连群在其排序中的方向。假设对于x≥约100Kb而言，染色质捕获连接使基因组距离为x的两个读取连接的可能性大致为1/x。建立加权的、定向的非循环图(WDAG)，其代表以给定顺序使叠连群定向的所有可能的方式。WDAG中的每个边缘对应于在其四种可能的组合方向之一上的一对相邻叠连群，并且边缘权重被设定为观察到两个叠连群之间的染色质捕获连接距离的集合的对数似然，其中假设它们以给定的方向紧密相邻。对于每个叠连群，如下计算其方向的质量评分。发现所观察到的在当前方向上的该叠连群与其邻居之间的染色质捕获连接的集合的对数似然。随后翻转叠连群并再次计算对数似然。由于计算方向的方式，第一对数似然保证会更高。对数似然之间的差异被视为质量评分。

[0494] 类似于Kaplan等人的备选DNA三角测量法也可在本文公开的方法中使用，以由叠连群和读取对组装基因组。DNA三角测量基于使用高通量体内全基因组染色质相互作用数据来推断基因组位置。对于DNA三角测量法，首先通过将基因组划分成100-kb的箱元(每个箱元代表大的虚拟叠连群)并计算每个放置的叠连群与每个染色体的平均相互作用频率来对CTR模式进行定量。为了评估长距离内的定位，省略了叠连群与其每侧1mb的侧翼叠连群的相互作用数据。平均相互作用频率强有力地区分了染色体内和染色体间相互作用，并且高度预测叠连群属于哪个染色体。接下来，训练简单的多类模型(朴素贝叶斯分类器)，以基于每个叠连群与每个染色体的平均相互作用频率来预测每个叠连群的染色体。基因组的组装部分用于拟合描述染色质捕获相互作用频率与基因组距离之间关系的概率性单参数指数衰减模型(DDD模式)。在每一轮中，从染色体上去除叠连群以及每侧1Mb的侧翼区域。然后基于相互作用谱和衰变模型估计每个叠连群最可能的位置。预测误差被量化为预测位置与实际位置之间的距离的绝对值。

[0495] 通过将DNA三角测量法与长插入片段文库组合，可进一步提高每个叠连群的可预

测性。通过了解每个叠连群的染色体分配和大致位置可显著降低长插入片段支架化的计算复杂性，因为每个叠连群只需要与其附近的叠连群配对，从而分辨不明确的叠连群连接，以及减少位于染色体远处区域处或不同染色体上的叠连群被错误连接的组装错误。

[0496] 实施例6.单元型定相方法

[0497] 由于通过本文公开的方法生成的读取对一般来源于染色体内接触，所以含有杂合性位点的任何读取对也将携带关于其定相的信息。利用该信息，可以快速且准确地进行短距离、中距离以及甚至长距离(兆碱基)的可靠定相。设计用于定相来自1000个基因组三元组(母亲/父亲/后代基因组的集合)之一的数据的实验已可靠地推断定相。此外，使用类似于Selvaraj等人(Nature Biotechnology 31:1111-1118(2013))的邻近连接的单元型重构也可与本文公开的单元型定相方法一起使用。

[0498] 例如，使用基于邻近连接的方法的单元型重构也可以在本文公开的方法中用于基因组定相。使用基于邻近连接的方法的单元型重构将邻近连接和DNA测序与单元型组装的概率算法相结合。首先，使用染色体捕获方案如染色质捕获方案进行邻近连接测序。这些方法可以捕获来自在三维空间中成环的两个远离的基因组基因座的DNA片段。对所得DNA文库进行鸟枪法DNA测序后，成对末端测序读取具有在几百个碱基对到数千万个碱基对范围内的“插入片段大小”。因此，在染色质捕获实验中生成的短DNA片段可以产生小的单元型区块，长片段最终可以将这些小区块连接在一起。在足够的测序覆盖度的情况下，这种方法能够连接不连续区块中的变体，并且将每一个这样的区块组装成单个单元型。然后将该数据与概率算法相结合以用于单元型组装。概率算法利用节点对应于杂合变体且边缘对应于可连接变体的重叠序列片段的图。该图可能含有由测序错误或反式相互作用造成的假边缘。随后使用最大切割算法来预测与由输入测序读取集合提供的单元型信息具有最大一致性的简约解决方案。由于邻近连接生成比常规基因组测序或配对测序更大的图，因此修改计算时间和迭代次数，使得可以以合理的速度和高精确度预测单元型。所得数据随后可用于使用Beagle软件和来自基因组计划的测序数据来指导局部定相，以生成具有高分辨率和准确度的跨越染色体的单元型。

[0499] 实施例7.宏基因组组装方法

[0500] 从环境中收集微生物并用固定剂如甲醛固定，以在微生物细胞内形成交联。通过使用高通量测序生成来自微生物的多个叠连群。通过使用基于染色质捕获的技术生成多个读取对。映射至不同叠连群的读取对指示哪些叠连群来自同一物种。

[0501] 实施例8.产生极长范围读取对(XLRP)的方法

[0502] 使用商购可得的试剂盒，将DNA提取至最高达150kbp的片段大小。使用来自Active Motif的商业试剂盒在体外将DNA组装成重构的染色质结构。用甲醛固定染色质，并将其固定在SPRI珠子上。用限制酶消化DNA片段并温育过夜。得到的粘性末端用α-硫代-dGTP和生物素化的dCTP补平以生成钝性末端。用T4连接酶连接钝性末端。用蛋白酶消化重构的染色质以回收连接的DNA。从珠子提取DNA，进行剪切并用dNTP修复末端。通过用SPRI珠子下拉来纯化片段。在一些情况下，连接衔接子并对片段进行PCR扩增以用于高通量测序。

[0503] 实施例9.产生高质量人类基因组组装体的方法

[0504] 在了解了通过本公开内容可生成跨越相当大的基因组距离的读取对的情况下，可对于该信息在基因组组装中的利用进行测试。本公开内容可显著改善从头组装体可能与染

色体长度的支架的连接。可对于使用本公开内容能产生多么完整的组装体和将会需要多少数据进行评估。为了评估本方法产生对组装有价值的数据的效率，可建立标准的Illumina鸟枪法文库和XLRP文库并进行测序。在一种情况下，使用来自标准鸟枪法文库和XLRP文库各1个Illumina HiSeq通道的数据。对由每种方法生成的数据进行测试并与各种现有的组装器(assembler)进行比较。任选地，还编写了新的组装器以特别地适应由本公开内容产生的独特数据。任选地，使用良好表征的人类样品提供参考，以比较由本方法产生的组装体，从而评估其准确度和完整性。利用先前分析中获得的知识，创建组装器以提高XLRP和鸟枪法数据的效率和有效利用率。使用本文所述的方法生成2002年12月小鼠基因组草图的质量的或更好的基因组组装体。

[0505] 可用于该分析的一个样品为NA12878。使用被设计为使DNA片段长度最大化的多种公开技术来提取来自样品细胞的DNA。分别建立了标准的Illumina TruSeq鸟枪法文库和XLRP文库。对于每个文库而言获得2x150 bp序列的单个HiSeq通道，每个文库可产生大约1.5亿个读取对。使用全基因组组装的算法将鸟枪法数据组装成叠连群。这类算法的实例包括：如Chapman等人(PLOS ONE 6(8):e2350(2011))所述的Meraculous或如Simpson等人(Genome research 22(3):549-56(2012))所述的SGA。将XLRP文库读取与通过初始组装产生的叠连群进行比对。该比对用于进一步连接叠连群。一旦确定连接叠连群的XLRP文库的有效性，则延伸Meraculous组装体以将鸟枪法和XLRP文库二者同时集成到单个组装过程中。Meraculous为组装器提供了坚实的基础。任选地，创建一体化组装器以适应本公开内容的具体需求。将通过本公开内容组装的人类基因组与任何已知序列进行比较，从而评估基因组组装体的质量。

[0506] 实施例10.用于以高准确度由小数据集对人类样品的杂合SNP进行定相的方法

[0507] 在一个实验中，对测试人类样品数据集中的大约44％的杂合变体进行定相。捕获了位于限制位点的一个读取长度距离内的所有或几乎所有定相变体。通过使用计算机模拟分析，可通过使用更长的读取长度和使用一种或多种用于消化的限制酶组合来捕获更多的定相变体。使用具有不同限制位点的限制酶的组合增加参与每个读取对的两个限制位点之一的范围内的基因组的比例(并且由此增加了杂合位点的比例)。计算机模拟分析显示，本公开内容的方法可以使用两种限制酶的不同组合对超过95％的已知杂合位置进行定相。附加的酶和更长的读取长度进一步增加了被观察和定相的杂合位点的分数，直至完全覆盖和定相。

[0508] 计算了两种限制酶的不同组合可达到的杂合位点覆盖度。在读取邻近的杂合位点方面，用所述方案对前三种组合进行测试。对于这些组合中的每一种，均生成XLRP文库并进行测序。将所得读取与人类参考基因组进行比对，并与样品的已知单元型进行比较以确定该方案的准确度。仅使用Illumina HiSeq数据的1个通道，以99％或更高的精确度对高达90％或更多的人类样品的杂合SNP进行定相。另外，通过将读取长度增加到300bp来捕获其他变体。可观察到的限制位点周围的读取面积实质上加倍。进行另外的限制酶组合，从而增加覆盖度和准确度。

[0509] 实施例11.高分子量DNA的提取和效果：

[0510] 用商购可得的试剂盒提取高达150kbp的DNA。图7显示可由长达提取的DNA的最大片段长度的捕获读取对生成XLRP文库。因此，可预期本文公开的方法能够由甚至更长的DNA段生成读取对。有许多完善的高分子量DNA回收方法，并且这些方法可与本文公开的方法或

方案一起使用。使用提取方法产生DNA的大片段长度，由这些片段创建XLRP文库，并且可以评估产生的读取对。例如，可以通过以下方法提取大分子量DNA：(1)根据Teague等人(Proc.Nat.Acad.Sci.USA 107(24):10848-53(2010))或Zhou等人(PLOS Genetics,5(11):e1000711(2009))的细胞的温和裂解；和(2)根据Wing等人(The Plant Journal:for Cell and Molecular Biology,4(5):893-8(1993))的琼脂糖凝胶塞，这些参考文献以全文(包括其中引用的任何参考文献)并入本文，或者通过使用来自Boreal Genomics的Aurora系统。这些方法能够生成超出下一代测序常规要求的长DNA片段；然而，本领域已知的任何其他合适的方法可以代替实现类似的结果。Aurora系统提供了出色的结果，并且可以从组织或其他制剂中分离和浓缩DNA，直至并超过兆碱基长度。使用这些方法中的每一种从单个GM12878细胞培养物开始制备DNA提取物，以控制样品水平上可能的差异。可以根据Herschleb等人(Nature Protocols2(3):677-84(2007))，通过脉冲场凝胶电泳评估片段的大小分布。使用前述方法，可以提取极大的DNA段并用于建立XLRP文库。随后对该XLRP文库进行测序和比对。通过比较读取对之间的基因组距离与从凝胶中观察到的片段大小来分析所得的读取数据。

[0511] 实施例12.减少来自不期望的基因组区域的读取对

[0512] 通过体外转录产生与不期望的基因组区域互补的RNA，并在交联之前将其添加到重构的染色质中。当补充的RNA与一个或多个不期望的基因组区域结合时，RNA结合降低了这些区域处的交联效率。由此减少了交联复合物中来自这些区域的DNA的丰度。将重构的染色质固定化，并如上所述使用。在一些情况下，RNA被设计为靶向基因组中的重复区域。

[0513] 实施例13.增加来自期望的染色质区域的读取对

[0514] 以双链形式产生来自期望的染色质区域的DNA以供基因组装或单元型分析。来自不期望的区域的DNA的呈现相应地减少。来自期望的染色质区域的双链DNA通过以多千碱基间隔在这样的区域上平铺的引物生成。在该方法的其他实施中，平铺间隔是变化的，从而以期望的复制效率来处理不同大小的期望区域。任选地通过DNA解链，使跨越所期望的区域的引物结合位点与引物接触。使用平铺的引物合成DNA的新链。例如通过用单链DNA特异性内切核酸酶靶向这些区域，减少或消除不期望的区域。可任选地扩增剩余的期望的区域。对制备的样品进行如本文别处所述的测序文库制备方法。在一些实施中，由每个这样的期望的染色质区域生成跨越可达每个期望的染色质区域的长度的距离的读取对。

[0515] 实施例14.Rapid Chicago文库制备方案

[0516] 该方案仅进行两天，并产生用于确定核酸样品中的连续性信息的高质量文库。

[0517] 在第1天，进行以下步骤。

[0518] 染色质组装。在冰上解冻Active Motif试剂盒组分。同时，Qubit(Broad Range)对1μl待组装的gDNA进行定量；包括用于准确度的大小标准。在移液前加热特别高分子量/粘性的样品，以确保均匀重悬。

[0519] 在硅化管中，在冰上将以下Active Motif染色质组装试剂盒组分按顺序混合在一起(制备0.25X额外的主混合物)：

[0520] h-NAP-1 0.7μl

[0521] HeLa核心组蛋白 0.9μl

[0522] 高盐缓冲液 5μl

[0523] 在冰上温育15分钟。

[0524] 同时,通过在冰上混合以下组分来制备10X ATP再生体系:

[0525] 10X ATP再生体系 5μl

[0526] 肌酸激酶 0.15μl

[0527] 在冰上温育后,将以下组分按顺序添加到组蛋白混合物中:

[0528] 低盐缓冲液 32.15μl

[0529] ACF 1.25μl

[0530] 10x ATP再生体系 5μl

[0531] 将45μl的主混合物分配至:

[0532] DNA 0.5μg

[0533] H_2O DNA+H_2O的最终体积为5μl

[0534] 在27℃下温育1hr。

[0535] 在一些情况下,待添加ACF/10x ATP再生体系的组蛋白混合物中的DNA浓度应为至少100ng/μl。然而,该方法成功地组装了染色质,该方法通过在10μl的DNA样品上方添加45μl的主混合物,使用低至50ng/μl的DNA得到了成功的Chicago文库。总体积增加10%不会影响组装染色质的总体质量。

[0536] 任选地,将5μl保存到硅化管中以供通过MNase消化(在如下的DpnII消化过程中)测试染色质组装体。

[0537] 甲醛交联。添加1.35μl的37%甲醛管(在室温下,白盖2ml管)。轻弹混合物并旋转。在室温(RT)下温育15分钟。添加2.7μl的2.5M甘氨酸管(在室温下,绿盖2ml管)。在冰上温育10分钟。

[0538] 使染色质结合至SPRI珠子。添加100μl的SPRI珠子;通过移液约10次进行混合。在室温下温育5分钟。在磁体上使管澄清5分钟,随后弃去上清液(SN)。用250μl洗涤缓冲液(10mM Tris/50mM NaCl)洗涤2次。

[0539] 可以在这些温育过程中制备消化主混合物(如下)。

[0540] DpnII消化。在结合至SPRI珠子之前,在冰上解冻一管DpnII消化混合物(在-30℃下,紫盖2ml管)。去除洗液后,用50ul的DpnII消化混合物使珠子重悬。弃去混合物的剩余物。在37℃下,在混匀仪(thermomixeer)中以>1000rpm的转速消化1小时。

[0541] 缓冲液更换。将样品放置在磁体上以分离上清液,并弃去上清液。用250μl洗涤缓冲液洗涤1次。

[0542] 可在这些温育过程中制备主混合物(如下)。

[0543] 末端补平。在Dpn II消化结束前15分钟,在冰上解冻一管末端补平混合物(在-30℃下,绿盖2ml管)。去除洗液后,用50ul的末端补平混合物使珠子重悬。弃去混合物的剩余物。

[0544] 在25℃下,在热混仪(thermomixer)中以>1000rpm温育30分钟。

[0545] 缓冲液更换。将样品放置在磁体上以分离上清液,并弃去上清液。用250μl洗涤缓冲液洗涤1次。

[0546] 可在这些温育过程中制备主要混合物(如下)。

[0547] 聚集体内DNA末端连接。末端补平反应结束前30分钟,在冰上解冻一管聚集体内连

接混合物(在-30℃下,假底3ml管)。去除洗液后,用250ul的聚集体内混合物使珠子重悬。弃去混合物的剩余物。

[0548] 在16℃下,在热混仪中以>1000rpm温育至少1小时。

[0549] 末端核苷酸交换。在聚集体内连接反应结束前5分钟,在冰上解冻一管末端核苷酸交换混合物(在-30℃下,黄盖2ml管)。向反应中直接添加5ul的末端核苷酸交换混合物。弃去混合物的剩余物。

[0550] 在16℃下,在热混仪中以>1000rpm温育15分钟。

[0551] 缓冲液更换。

[0552] 将样品放置在在磁体上以分离上清液,并弃去上清液。用250μl洗涤缓冲液洗涤1次。

[0553] 可在这些温育过程中制备主混合物(如下)。

[0554] 交联逆转。

[0555] 在末端核苷酸交换反应结束前5分钟,向一个满的(full)交联逆转缓冲液管(在室温下,红盖2ml管)中添加11μl的NEB蛋白酶K(在-30℃下,20mg/ml)。去除上清液后,用50ul的交联逆转/蛋白酶K混合物使珠子重悬。弃去混合物的剩余物。

[0556] 在55℃下,在热混仪中以>1000rpm温育15分钟。

[0557] 在68℃下,在热混仪中以>1000rpm温育45分钟。

[0558] 在SPRI上纯化DNA。将交联逆转反应液置于磁体上以分离上清液。将上清液转移至干净的1.5ml管。添加100μl的SPRI珠子;通过移液约10次进行混合。在室温下温育5分钟。将样品放在磁体上5分钟,然后取出并弃去上清液。

[0559] 用250μl新鲜制备的80%EtOH洗涤3次。风干5分钟,注意不要过度干燥。用78μl TE使珠子重悬,等待2min。置于磁体上,将75μl的上清液转移至Bioruptor 0.65ml管中。用Qubit HS定量1ul DNA;预期回收率为输入的30%-75%。

[0560] 在第2天,进行以下步骤。

[0561] 片段化。将Bioruptor冷却至4℃。在冰上冷却DNA最少10min。涡旋,旋转样品。将管放入Bioruptor转盘中,注意不要溅出DNA。运行15秒开/90秒关的4次循环。从转盘上取下。涡旋,旋转管。运行15秒开/90秒关的3次循环。从转盘上取下。涡旋,旋转管。

[0562] 在TapeStation上分析Chicago DNA。将2ul的片段化DNA加载到使用高灵敏度D1000带的TapeStation上。预期有以约350nt为中心的宽分布。

[0563] 末端修复。将55.5μl的片段化DNA转移至含有以下NEBNext Ultra试剂的PCR管(绿盖):末端制备酶混合物3.0μl,末端修复反应缓冲液6.5μl。使用NEB-END方案在PCR机器中温育:20℃下30分钟,65℃下30分钟,4℃下保持。

[0564] 衔接子连接。将以下NEBNext Ultra试剂(红盖)添加到反应中:Blunt/TA连接酶主混合物15μl,连接增强剂1.0μl,自制Y-衔接子15μM 2.5μl。

[0565] 使用NEB-Ligate方案在PCR机器中温育:20℃下15min。

[0566] 连接事件的捕获。为每个Chicago反应制备25μl的C1珠子的主混合物。将样品置于磁体上以分离上清液,并弃去上清液。用250μl的1X TWB(参见缓冲液配方页)洗涤两次。使珠子重悬于85μl乘以Chicago反应数的2X NTB中。将85μl的在2X NTB中的珠子分配到一组干净的1.5μl管中。将85μl末端修复反应转移至珠子。在LabQuake旋转器上于室温下温育

30min。

[0567] 将样品放在磁体上以分离上清液，并弃去上清液。用250μl LWB洗涤1次。用250μl NWB洗涤2次。用250μl洗涤缓冲液洗涤2次。

[0568] 索引PCR。使珠子在49μl的如下混合物(主混合物+0.25％Rx)中重悬：H_2O 23μl；IS4引物(10uM)1.0μl；2X KAPA MIX 25μl。

[0569] 转移至PCR条管。向每个管中添加1μl的10μM索引引物；确保记录每个样品的索引ID。

[0570] 用这些步骤扩增13次循环：在98℃下3min；在98℃下20sec；在65℃下30sec；在72℃下30sec；从步骤2开始重复12次以上；在72℃下1min；在12℃下保持。

[0571] 在SPRI上纯化扩增的DNA。将样品放在磁体上以分离上清液。将上清液转移至干净的1.5ml管。添加100μl的SPRI珠子；通过移液约10次进行混合。在室温下温育5min。将样品置于磁体上以分离上清液5min；弃去上清液。用250μl新鲜制成的80％EtOH洗涤2次。风干5min，注意不要过度干燥。用32μl TE使珠子重悬，等待2min。在磁体上浓缩。将洗脱的DNA转移至新的1.5ml管。在宽范围Qubit上定量DNA；预期浓度约30ng/ul。

[0572] 在TapeStation上分析索引化的PCR DNA。通过将0.5ul的纯化的PCR添加到4.5μl的TE中以1:10稀释。在TapeStation高灵敏度D1000带上加载2μl。预期有以约550nt为中心的宽分布。

[0573] 在Pippin Prep上对索引PCR DNA进行大小选择。添加10μl的1.5％DF Pippin Prep样品缓冲液(标记K)。根据制造商的方案准备仪器和凝胶。使用在TapeStation分析中观察到的分布中心周围300nt的宽范围窗口进行大小选择；通常为400-700nt。使用Qubit高灵敏度定量DNA；回收率应为约5-10ng/ul。

[0574] 在TapeStation上分析大小选择的DNA。通过添加1μl至4μlTE中以1:5稀释。在TapeStation高灵敏度D1000带上加载2ul。将浓度(以pg/ul和摩尔为单位)记录到JIRA中。

[0575] 实施例15

[0576] 根据Chicago文库的生成，进行微球菌核酸酶(MNase)消化以测试染色质组装。

[0577] 主混合物制备。在室温下制备消化和终止主混合物。如下将MNase稀释至1:1000：在H_2O中进行1:10稀释(1μl的MNase 50U/μl+9μl的H_2O)；在H_2O中进行1:1000稀释(1μl的1:10稀释液+99μl的H_2O)；通过将1μl的MNase 1:1000添加到一管MNase消化缓冲液(在室温下，黄盖管)中制备MNase消化混合物；通过将11μl的NEB蛋白酶K 20mg/ml添加到一个满的终止缓冲液管(在室温下，蓝盖管)中制备终止缓冲液混合物。

[0578] MNase消化。在37℃下预温MNase消化混合物2min。每管添加45μl至5μl组装的染色质，每个样品之间等待30sec。在第一次样品添加时启动计时器，并按顺序保持样品。5min后，从第一管开始，添加50μl的终止缓冲液混合物。再次，每个管之间等待30sec，使得每个样品恰好消化5min。在37℃下温育另外30min。

[0579] 使用Qiagen MinElute试剂盒纯化：添加300ul的Qiagen缓冲液ERC，混匀；转移至MinElute Reaction Cleanup柱；离心1min，弃去流过液；添加700μl的缓冲液PE(确保已经添加乙醇)；离心1min，弃去流过液；离心1min以确保没有留下PE缓冲液；将柱子转移至1.5ml管中；添加10ul的EB缓冲液，等待1min；离心1min以回收DNA。

[0580] 在HS DNA 1000 TapeStation带上运行2μl的MNase消化的样品。

[0581] 实施例16

[0582] 通过退火制备扩增衔接子。如下实现15μM部分双链扩增衔接子的制备。在1.5ml管中将以下物质混合在一起:37.5μl的在TE+50mM NaCl中的200μM P5_full_A(寡核苷酸#111);37.5μl的在TE+50mM NaCl中的200μM P7_Y_Rev(寡核苷酸#132);420μl的TE;5μl的NaCl 5M。等分到热循环仪中的两个PCR管中,运行退火程序:95℃2min;以0.1℃/sec缓慢降至25℃。

[0583] 适用于扩增衔接子的寡核苷酸如下所示(*是硫代磷酸酯键)

[0584] EQ ID NO位置序列(5'至3')

[0585] 1 P5_full ACACTCTTTCCCTACACGACGCTCTTCCGATG*T

[0586] 2 P7_rev /5Phos/CATCGGAAGAGCACACGTCTGAACTCCAGTCA*/3ddC/

[0587] 3 P5_full ACACTCTTTCCCTACACGACGCTCTTCCGACC*T

[0588] 4 P7_rev /5Phos/GGTCGGAAGAGCACACGTCTGAACTCCAGTCA*/3ddC/

[0589] 5 P5_full ACACTCTTTCCCTACACGACGCTCTACCGATC*T

[0590] 6 P7_rev /5Phos/GATCGGTAGAGCACACGTCTGAACTCCAGTCA*/3ddC/

[0591] 7 P5_full ACACTCTTTCCCTACACGACGCTATTCCGATC*T

[0592] 8 P7_rev /5Phos/GATCGGAATAGCACACGTCTGAACTCCAGTCA*/3ddC/

[0593] 9 P5_full ACACTCTTTCCCTACACGACGCTCTTCGGATC*T

[0594] 10 P7_rev /5Phos/GATCCGAAGAGCACACGTCTGAACTCCAGTCA*/3ddC/

[0595] 11 P5_full ACACTCTTTCCCTACACGACCCTCTTCCGATC*T

[0596] 12 P7_rev /5Phos/GATCGGAAGAGGACACGTCTGAACTCCAGTCA*/3ddC/

[0597] 13 P5_full ACACTCTTTCCCTACACGACGCACTTCCGATC*T

[0598] 14 P7_rev /5Phos/GATCGGAAGTGCACACGTCTGAACTCCAGTCA*/3ddC/

[0599] 15 P5_full ACACTCTTTCCCTACACGACGCTCTTCCGATC*T

[0600] 16 P7_rev /5Phos/GATCGGAAGAGCACACGTCTGAACTCCAGTCA*/3ddC/

[0601] 实施例17

[0602] 制备SPRI珠子。量入50ml管中:PEG-8000粉末9g。

[0603] 然后添加:

[0604]

原料浓度	最终浓度
1M Tris-Cl pH 8.0	500 μl 10mM
0.5M EDTA	100 μl 1mM
NaCl	1M
H_2O	至约 48 mL

[0605] 摇动以溶解PEG。然后添加吐温并轻轻混合:10%吐温20 250μl 0.05%。

[0606] 同时,使Sera-Mag珠子重悬。将1ml转移至1.5ml管中。在磁体上使管澄清,随后弃去上清液(SN)。用1ml TE将珠子洗涤4次。重悬于1ml TE中。全部转移至PEG溶液并通过倒置进行混合。用H_2O使该溶液达到50ml。在4℃下储存。用50bp序列梯(例如,GeneRuler或Hyperladder)以不同比率对每批进行校准。

[0607] 实施例18.使用由来源于粪便样品中核酸的体外组装的染色质聚集体生成的序列读取进行人粪便宏基因组组装

[0608] 用MoBio Powerfecal试剂盒制备用于粪便宏基因组组装的DNA。根据试剂盒中提供的DNA分离方案制备粪便子样品(在单个时间点从单个个体收集的样品的子样品)。制备四个约250mg的子样品。每个样品的DNA产量如下:(1)4.28μg;(2)7.28μg;(3)6.48μg;和(4)5.56μg。

[0609] 选择样品(2)用于进一步处理,因为它在四个子样品中具有最高DNA产量。使用TapeStation(Agilent)分析样品(2)中的DNA片段的大小。如图13A所示,该样品的片段大小中值约为22kb,并且不存在小片段。为宏基因组组装制备了两个文库——第一个文库使用体外组装的染色质聚集体和邻近连接制备,而第二个文库被制备用于鸟枪法测序。

[0610] 如图13B所示使用来自样品(2)的500ng的DNA和体外组装的染色质制备第一个文库。染色质在来自样品(2)的裸DNA 1301上进行体外重构1302。然后将染色质用甲醛固定以形成如1303所示的染色质聚集体。用限制酶消化固定的染色质以生成如1304所示的游离粘性末端。如1305所示,用生物素化(圆圈)和硫醇化(正方形)的核苷酸补平游离末端。如1306所示,连接游离钝性末端(用星号表示连接)。使交联逆转并去除染色质缔合的蛋白质以产生如1307所示的文库片段。在MiSeq(Illumina,2x75bp)上对文库进行测序。生成了5,026,934个读取对。

[0611] 制备第二个文库以用于鸟枪法测序。第二个文库是使用文库制备试剂盒由2μg样品(2)制备的无TrueSeq PCR的文库。在MiSeq(Illumina,2x150bp)上对鸟枪法文库进行测序。在使用Omega(重叠图宏基因组组装器,Haider等人Bioinformatics(2014)doi:10.1093/bioinformatics/btu39)生成宏基因组组装体之前,使用SeqPrep对读取进行修剪和合并。有15,758,635个读取对,并且1,810,877个读取对合并为单个读取。

[0612] 如图14所示将鸟枪法读取映射至组装体以评估插入片段长度分布和覆盖度。在图14中,x轴表示以bp为单位的插入片段长度,而y轴表示读取对的数目。合并的读取对显示为虚线,而未合并的读取对显示为实线。

[0613] 来自用体外染色质聚集体制备的文库的读取被映射至组装体以评估插入片段长度分布。将819,566个读取对映射至相同的支架。图15中示出了在映射位置之间的插入片段分布。在图15中,x轴表示以kb为单位的插入片段大小,而y轴表示读取对的数目。相同链读取对以短虚线显示。还显示了两个读取对类别——“内部(innies)”用长虚线表示,而“外部(outties)”用实线表示。在读取对中,有1,358,770个映射至不同的支架。剩余的读取对没有映射或没有唯一地映射。

[0614] 图16和图17示出了使用两种文库制备方法的命中覆盖率的比较。图16示出了来自针对鸟枪法测序制备的文库与使用体外组装染色质聚集体制备的文库(“Chicago”)的命中散点图。图17示出了按照叠连群长度的每个叠连群的鸟枪法命中/体外组装染色质命中(“Chicago”)的散点图。用HiRise软件分析读取,该HiRise软件使用似然模型来构建支架并且也破坏似乎不正确的输入支架。与Omega输出中的15.7kb相比,最终的支架N50约为53.4kb。

[0615] 实施例19.人群中未知病原体的检测和测序

[0616] 使用来自粪便样品的读取数据的从头基因组组装体来鉴定受试者中的未知病原

体。随着国际上健康的改善，发现没有已知原因或病原体来源的疾病的暴发越来越普遍。因为病原体难以分离或培养，所以分离该病原体的工作通常耗时且具有挑战性。

[0617] 从疑似患有未知疾病的患者或确诊患者收集粪便标本和/或尿液标本。用粪便DNA提取方法，如MetaHIT(人类肠道宏基因组学)方法或HMP(人类微生物组计划)方法，粪便DNA提取试剂盒，如来自MO BIO的MoBio Powerfecal试剂盒、来自Qiagen的QIAmp DNAStool Mini试剂盒或来自Zymo Research的ZR Fecal DNA MiniPrep试剂盒，制备用于粪便宏基因组组装的DNA。用DNA提取方法或DNA提取试剂盒，如来自Qiagen的QIAamp DNA Micro试剂盒；来自Intron Biotechnology的i-genomic Urine DNA Extraction Mini试剂盒；来自Zymo Research的ZR Urine DNA分离试剂盒；来自Norgen Biotek的Norgen RNA/DNA/蛋白质纯化试剂盒；以及来自Abcam的Abcam Urine分离试剂盒，提取尿中的DNA。

[0618] 用体外组装的染色质聚集体和500ng来自粪便DNA样品的DNA或尿DNA制备文库。在体外在来自粪便或尿样品的裸DNA上重构染色质，并用甲醛固定染色质和DNA以形成染色质聚集体。用限制酶消化固定的染色质以生成游离的粘性末端。用生物素化和硫醇化的核苷酸补平游离末端，随后连接游离的钝性末端。逆转交联并去除染色质缔合的蛋白质以产生文库片段。对文库进行测序并组装读取对。

[0619] 然后，使用来自粪便样品的读取数据的从头基因组组装体来鉴定与受试者群体中的有病或患病个体相对应的核酸分子。将核酸信息组装成基因组大小的叠连群，以便把序列信息分组为染色体或基因组大小的单元。

[0620] 在分析中不再强调与健康个体中可能存在的生物体相对应的基因组。在分析中，也不再强调与在显示疾病症状的个体中可能更为丰富的生物体相对应的基因组。

[0621] 鉴定与先前未表征的生物体相对应的基因组。对基因组进行分析以确定其中编码的代谢途径，并设计培养方案以促进具有基因组的微生物的独立于宿主的培养。继续进行代谢途径的分析，以鉴定选择性地阻断微生物复制的潜在的药物靶标。根据在此生成的基因组信息，在生成的微生物培养物上测试药物靶标，并显示该药物靶标阻断复制。将这些药物施用给表现出爆发症状的个体，并且证明该药物治疗能减轻症状。

[0622] 实施例20.使用鸟枪法测序对人群中未知病原体的检测和测序

[0623] 使用来自粪便样品的读取数据的从头鸟枪法测序来鉴定受试者中未知病原体的基因组序列。如以上实施例中所述分离核酸，并仅进行鸟枪法测序。

[0624] 鉴定与已知和未知微生物相对应的测序读取。确定在患有疾病的个体中存在一种或多种未知生物体。然而，不能确定代谢途径信息，并且鸟枪法序列信息不提供关于微生物可能如何培养或哪些药物可用于阻止微生物在人类宿主中生长或增殖的见解。从所述结果无法得出治疗方案的建议。

[0625] 实施例21.患者中的抗生素抗性基因的检测

[0626] 一名患者具有对抗生素治疗有抗性的感染。从该患者获得粪便样品，并从该样品中提取核酸。

[0627] 对该核酸进行鸟枪法序列分析，并生成许多序列读取。一些单独的序列读取足够长，从而允许它们可靠地映射至假定的宿主生物体。一些读取映射至假定的抗生素抗性基因座，并且怀疑编码提供抗生素抗性的基因产物的核酸是否存在于患者中。

[0628] 所述序列信息不足以确定哪些抗生素抗性基因座映射至哪些宿主微生物。

[0629] 实施例22.患者中的抗生素抗性基因宿主的检测

[0630] 一名患者具有对多重抗生素治疗有抗性的感染。从该患者获得粪便样品，并从该样品中提取核酸。

[0631] 对该核酸进行鸟枪法序列分析，并生成许多序列读取。一些单独的序列读取足够长，从而允许它们可靠地映射至假定的宿主生物体。一些读取映射至假定的抗生素抗性基因座，并且怀疑编码提供抗生素抗性的基因产物的核酸是否存在于患者中。

[0632] 对所述核酸进行如本文公开的分析。确定连接信息，以确定相对于抗生素抗性基因的由共同核酸分子产生的核酸序列。将鸟枪法序列信息组装成与微生物基因组相对应的叠连群。

[0633] 确定了多个抗生素抗性基因映射至单个微生物宿主。根据对组装的微生物基因组中存在和不存在的代谢途径的分析，还确定了抗生素抗性基因的微生物宿主很可能易受先前未施用的抗生素的影响。

[0634] 向所述患者施用先前未施用的抗生素，并且感染症状得到减轻。

[0635] 实施例23.患者中的抗生素抗性基因宿主的检测

[0636] 一名患者具有对连续施用的多种抗生素的治疗有抗性的感染。从该患者获得粪便样品，并从该样品中提取核酸。

[0637] 对该核酸进行鸟枪法序列分析，并生成许多序列读取。一些单独的序列读取足够长，从而允许它们可靠地映射至假定的宿主生物体。一些读取映射至假定的抗生素抗性基因座，并且怀疑编码提供抗生素抗性的基因产物的核酸是否存在于患者中。

[0638] 对所述核酸进行如本文公开的分析。确定连接信息，以确定相对于抗生素抗性基因的由共同核酸分子产生的核酸序列。将鸟枪法序列信息组装成与微生物基因组相对应的叠连群。

[0639] 确定了多个抗生素抗性基因映射至多个微生物宿主，以及没有微生物宿主拥有多于一个微生物抗性基因。

[0640] 向所述患者施用先前施用的抗生素治疗，但抗生素是平行施用的而不是连续施用的。也就是说，同时施用先前发现在一次施用时无效的抗生素，并且感染症状得到减轻。

[0641] 实施例24.异质样品中的个体序列的检测

[0642] 寻找感兴趣的个体。从个体的父母提供的核酸样品合理地推断该个体的基因组信息。确定了在该个体中预期的SNP(单核苷酸多态性)模式。在给定染色体上的SNP模式包括许多在单独时常见但合起来不太可能在任何单个个体中组合出现的SNP。

[0643] 怀疑所述个体存在于某个位置。调查该位置，并从该位置获得异质DNA样品。对DNA进行鸟枪法测序，并确定大量读取。鉴定了预期存在于感兴趣的个体的基因组中的每个SNP。然而，无法得到SNP之间的连接信息，并且研究者不能确定检测到的SNP是否来自单个个体或与单个核酸分子相对应。

[0644] 实施例25.异质样品中的个体基因组标签的检测

[0645] 如以上实施例24所述寻找感兴趣的个体。对DNA进行鸟枪法测序，并确定大量读取。鉴定了预期存在于感兴趣的个体的基因组中的每个SNP。

[0646] 对从该位置获得的异质DNA的第二样品进行如本文公开的分析。鉴定跨越感兴趣的SNP的序列读取，并将所述序列读取以及其他共有共同标记信息的读取映射至特定的核

酸分子。确定SNP的相位信息，并且确定具有对于感兴趣的个体预测的SNP模式的个体最近处于所调查的位置。

[0647] 同时，根据来源于从该位置获得的异质DNA样品的鸟枪法和连接信息来确定该位置处其他个体的SNP模式。

[0648] 实施例26.新型生物体测定

[0649] 选择已知具有感兴趣的肠道生物群系的白蚁进行测序。已知白蚁缺乏编码木材降解所必需的酶的基因。怀疑白蚁的肠道中是否具有单独地或组合地编码代谢纤维素所必需的酶的一种或多种微生物。

[0650] 从白蚁群体获得核酸并对其进行鸟枪法测序。获得指示代谢纤维素的能力的分离的读取。然而，序列读取不能组装成高等级支架，以便鉴定在白蚁肠道中栖息的生物体的数目或身份。

[0651] 实施例27.新型生物体发现

[0652] 选择已知具有感兴趣的肠道生物群系的白蚁进行测序。已知白蚁缺乏编码木材降解所必需的酶的基因。怀疑白蚁的肠道中是否具有单独地或组合地编码代谢纤维素所必需的酶的一种或多种微生物。

[0653] 从白蚁群体获得核酸，并如以上实施例16中所述进行鸟枪法测序，同时使用本文公开的方法对相同核酸的第二样品进行分析。鸟枪法序列读取被映射至与许多不同生物体的基本完整基因组相对应的不同集群，包括厌氧细菌和新型蜂窝状物种。

[0654] 对由此生成的基因组的分析表明，至少一些基因组缺少生物体在需氧条件下培养所必需的生物合成途径，或者在不存在由肠道微生物群落其他成员产生的复杂代谢物组合的情况下培养所需的生物合成途径。因此，确定先前未知且不太可能使用标准方法培养的生物体的基因组。

[0655] 实施例28.粪便宏基因组组装中的掺入实验

[0656] 来自复杂宏基因组群落的基因组的从头组装呈现出特殊的挑战。不同于典型的单个生物体的从头组装项目，输入DNA来源于多达数百或数千或更多的野生丰度不同的不相关生物体。此外，单独的物种可以在不同的菌株中表现出小的或大的等位基因变异。我们描述了全基因组宏基因组组装的新方法，该方法利用通过邻近连接可获得的长范围接触信息。我们进行了一组对照实验，其中我们添加了来自基因组被良好表征的细菌物种天蓝色链霉菌、但在粪便样品中不存在的DNA。我们制备了两个文库：标准的短插入鸟枪法文库和邻近连接文库和并对两者均进行了测序。利用这些数据，我们证明了可以生成天蓝色链霉菌的已知基因组的完整组装体。因此，使用这种方法可以准确地重构来自复杂宏基因组样品的微生物的基因组。

[0657] DNA收集：使用MoBio PowerFecal收集试剂盒，根据方案，我们从250mg粪便样品中收集了2微克DNA。我们从ATCC订购了天蓝色链霉菌的基因组DNA制剂。为了模拟PowerFecal纯化后的DNA片段的大小分布，我们使天蓝色链霉菌DNA通过在PowerFecal试剂盒中提供的离心柱。如图18所示，在TapeStation迹线中，粪便DNA制剂(1801，蓝色，在x轴的100bp和15000bp处迅速上升接近y轴顶部)和天蓝色链霉菌DNA(1802，绿色，在15000bp处迅速上升至样品强度为100)中的片段大小分布具有类似的长度。x轴显示大小(bp)，标记从左至右为100、250、400、600、900、1200、1500、2000、2500、3000、4000、7000、15000和48500。y轴显示以

荧光单位(FU)计的样品强度。

[0658] 测序文库的制备:我们制备了三种粪便DNA混合物,其中以总量的1%、5%和10%添加天蓝色链霉菌。这是为了粗略估计当基因组包含总宏基因组样品的1%、5%和10%时正确组装基因组的困难。对于每种混合物,我们如前所述(Putnam等人,Genome Research,2016)使用体外重构的染色质制备了Illumina鸟枪法文库和邻近连接文库。随后我们在Illumina MySeq测序仪上对这些文库进行测序。

[0659] 鸟枪法读取的分析和叠连群组装:通过比对鸟枪法读取与天蓝色链霉菌的已知基因组序列(GenBank ID:NC_003888.3),我们估计了鸟枪法数据中天蓝色链霉菌基因组的覆盖度。图19示出了对于每个水平的掺入天蓝色链霉菌DNA,这些鸟枪法数据中的覆盖倍数分布。x轴显示覆盖倍数,而y轴显示天蓝色链霉菌上的位置数目。如图所示,1%掺入(最左峰)实验中的基因组覆盖度倍数(13倍中值)不够高,以至于不能支持通常需要至少30倍基因组覆盖度的精确的叠连群组装。另一方面,5%(中间峰)和10%(最右峰)掺入实验对于叠连群组装不太可能受覆盖度限制。

[0660] 我们使用Omega(Haider等人,2014Bioinformatics)组装每个数据集的叠连群。然后,我们将这些叠连群映射至已知的天蓝色链霉菌基因组序列,以评估这些数据中组装体的完整性和片段化。图20示出了对于1%(红色,左侧)、5%(绿色,中间)和10%(蓝色,右侧)鸟枪法数据集,作为叠连群存在的天蓝色链霉菌基因组的总量。每一个数据集周围的外部黑圆圈与天蓝色链霉菌的总基因组大小成比例。正如预期的那样,1%掺入实验无法将大量基因组组装成叠连群,而5%和10%实验将大部分基因组组装成叠连群。每个实验的叠连群的总数在表2中给出。

[0661] 表2.叠连群的总数。

[0662]

实验	天蓝色链霉菌的叠连群总数	OMEGA 组装中的总叠连群
1%	297	24,333
5%	2,647	26,567
10%	1,524	25,347

[0663] 这些结果对于从宏基因组从头组装的一些方法是典型的:大多数组成基因组可以被组装成小叠连群。在典型的情况下,人们不会知道例如10%掺入实验中1,524个叠连群全部来自天蓝色链霉菌。

[0664] 邻近连接文库中的连接信息的评估:为了确定邻近连接文库是否含有可用于正确地对这些叠连群进行支架化的信息,我们将来自这些文库的读取对映射至已知的天蓝色链霉菌基因组序列。参见图21,其显示了每个读取对所跨越的距离,其中x轴显示以千碱基(kb)为单位的跨越的距离,而y轴为所有读取对上的累积分布。对于邻近连接文库而言典型的是,读取对所跨越的距离覆盖了超过用来生成文库的输入DNA片段的大小的所有距离。这表明即使对于这些细菌DNA制剂,体外邻近连接文库制剂也起作用,并含有可用于基因组支架化和组装的信息。

[0665] 基因组支架化:我们使用了邻近连接文库数据来对所有叠连群进行支架化。然后,

我们通过鉴定与5％和10％实验中的天蓝色链霉菌相对应的基因组支架来评估支架化准确性和完整性，其中存在代表大多数天蓝色链霉菌基因组的叠连群。注意，在1％实验中，天蓝色链霉菌的支架化在对本实验选择的参数下是不可能的，因为叠连群覆盖度太小而不能进行支架化。备选的参数可产生单独的结果。还要注意，在这些实验中的任何一个中生成更多的鸟枪法数据可能会增加存在的所有基因组(包括天蓝色链霉菌)的叠连群覆盖度。

[0666] 图22A和图22B中示出了在5％和10％实验中代表天蓝色链霉菌的支架。图22A描绘了已知的链霉菌基因组(x轴)相对于在5％实验中如本文所述生成的三个支架的点阵图。在5％实验中，天蓝色链霉菌存在于3个大支架中，不同于在采用邻近连接数据进行支架化之前的2,647个叠连群。图22B描绘了已知的天蓝色链霉菌基因组(x轴)相对于在10％实验中如本文所述生成的一个支架的点阵图。在10％实验中，天蓝色链霉菌基因组存在于1个大支架中。

[0667] 实施例29:人类粪便DNA

[0668] 进行一系列实验以评估上述从头宏基因组测序和组装的方法。由人类粪便DNA提取物生成鸟枪法和“Chicago”体外邻近连接文库，并进行“HiRise”从头叠连群组装和支架化。设计这些概念验证实验以确定:(1)如何快速且可靠地从粪便样品中提取高分子量DNA;(2)如何利用Chicago实验室方案由从粪便样品中回收的DNA生成体外染色质邻近连接文库，所述DNA主要来自原核生物体;(3)Chicago数据是否可用来有效地对来自同一DNA制剂的宏基因组叠连群进行支架化;(4)已知的基因组(其DNA被掺入宏基因组样品中，从而以相同的方式进行处理)是否可以可靠地组装;以及(5)用什么方式可使HyRise基因组组装策略适合于宏基因组组装的特殊挑战。

[0669] 针对粪便DNA的DNA提取，对几种商购可得的试剂盒进行了测试。Qiagen粪便DNA试剂盒一致地产生30-40千碱基的DNA，这是所有测试试剂盒中最长的，几乎没有较短的片段(见图23A，其中用于从健康供体收集DNA的Qiagen Fecal prep试剂盒产生的DNA片段大小显示为单一模式分布，其中大多数片段为30至40kb)。在组装之后(下文所述)，通过将读取映射至几个最大的组装支架并测量邻近连接事件之间的推断距离的分布来评估邻近连接文库(参见图23B，其中在组装和支架化之后，来自该文库的Chicago对(实验2，以虚线示出)被映射至支架)。在典型的Chicago文库中，读取对可跨越长达输入DNA的大小的距离。该分析可以是路线中“Chicago”文库的标准质量控制程序的一部分，并且可提供邻近连接产物在标准Chicago文库中分布的有效评估。注意，该分析可能需要读取可被映射至的基因组组装体。对于该分析，HiRise的宏基因组形式用于对这些数据进行支架化，如下文所述对宏基因组数据进行修改。根据该分析，可以证明，Chicago程序对粪便样品中的至少一部分DNA的表现如预期的那样。

[0670] 还测试了当原核生物体的基因组是混合物的已知组分时准确组装以低丰度存在的该基因组的能力。在该实验中，使用来自天蓝色链霉菌的DNA，其完整基因组是已知的。将来自天蓝色链霉菌的DNA添加至粪便DNA制剂，使得该DNA为总DNA质量的1％。重要的是，通过使输入天蓝色链霉菌DNA穿过在粪便制剂中使用的Qiagen柱，使输入天蓝色链霉菌DNA断裂为与粪便DNA相当的大小。在该实验中，回收7.68Mb的单个支架，其包含89％的8.67Mb天蓝色链霉菌基因组。该单个支架(参见图24)缺乏任何大的与已知基因组相比的结构差异。该天蓝色链霉菌基因组在x轴上，而在此生成的支架在y轴上。因为新的支架与参考序列不

在同一起点开始,所以点阵图被卷绕。注意到组装体没有错误连接并且几乎完整。“丢失的”区段为单个区域,其本身几乎被完整组装为另一个大支架,并且这两个支架提供几乎完整的天蓝色链霉菌的组装体。从该分析可以看出,该组装策略可准确地对已知基因组进行支架化,即使在它是整个群落的次要组分时——在该测试例中为1%。

[0671] 鉴于掺入物的正确且几乎完整的组装,接下来评估在支架化前后组装体的邻接。对于叠连群组装步骤,使用了Meraculous组装器的形式,其被修改以允许如在宏基因组数据中预期的宽覆盖范围。还成功地使用了其他宏基因组组装器(未示出)。随后使用宏基因组形式的HiRise(meta-HiRise)对叠连群进行支架化,其放宽了关于以标准HiRise方法制备的支架之间的覆盖均匀性的假设。

[0672] 对于该分析,采用了被称为宏基因组群落N50(MGC N50)的度量,它通过(1)将支架从最大到最小进行排序,并(2)将鸟枪法读取映射至所有支架来计算。MGC N50是所有鸟枪法读取的累积计数达到总数的50%时支架的大小。在鸟枪法读取代表每个OTU的群落丰度的统计数字的假设下,该度量描述了宏基因组组装体的总体邻接,因为其涉及存在于样品中的OTU的丰度。注意,如果少于50%的读取可以可靠地映射至组装体,那么MGC N50是未定义的。采用在此收集的数据,实现了1.5-25倍的MGC N50的提高。此外,在每个实验中生成了若干个多兆碱基支架。

[0673] 这些结果显示,用于有效生成长范围邻接信息的本文公开的体外染色质组装框架可应用于宏基因组背景。该程序可能需要约1微克的高分子量DNA。使用标准的商业粪便DNA制备试剂盒,可以可靠地从正常的粪便样品中提取所述量的高分子量DNA。该DNA适合于本文采用的体外染色质组装方法。生成的邻近连接文库可用于准确地将宏基因组样品中的基因组进行支架化,如采用天蓝色链霉菌的掺入阳性对照实验所示。

[0674] 实施例30:代表性偏差的最小化

[0675] 如本文所公开的,已证明Chicago方案可与作为输入的来自粪便样品的DNA一起使用。在此讨论了在该方案上扩展的示例性方法。

[0676] Chicago方案可依赖于采用特异性限制酶MboI对体外染色质聚集体的消化,该限制酶的切割位点为GATC。该方案可被修改为使用其他限制酶,如MboI的甲基化不敏感的同切点酶(例如,DpnII)。宏基因组群落成员的不同碱基组成可导致不均匀的切割,从而导致在组装文库中不均匀的表示。图25示出了在掺入实验中,Chicago组装数据与鸟枪法数据中读取覆盖度之比的示例图。如图25所示,支架的每个碱基对的鸟枪法覆盖度与样品中的丰度成比例。鸟枪法覆盖度与Chicago覆盖度之比在约一个数量级内变化。在许多情况下,甚至在该比值低时,产生大支架。对于大多数支架长度,该比值在十倍的范围内。注意,具有中等GC分数的支架具有中间水平的Chicago覆盖度,这与作为每个OTU的Chicago文库效率中的因素的碱基组成相一致。为了降低这一偏差,可以采用不同的策略。

[0677] 测试限制酶组合的使用:对于具有极高A/T含量的项目,可使用备选的限制酶,其限制位点比MboI具有更丰富的A/T(GATC)。宏基因组群落拥有具有各种各样G/C含量的基因组;因此单个限制酶对于产生所有群落OTU的有效Chicago文库生成可能并不理想。酶的组合可以用于采用各种粪便样品的Chicago文库制备。

[0678] 使不含限制酶的方案适应于宏基因组用途:不含限制酶的方案也可用于Chicago文库。这样的方法可采用以不依赖序列的方式切割DNA的核酸酶。例如,随后使用生物素化

的衔接子桥接钝性末端并标记连接的区域。

[0679] 实施例31:宏基因组组装软件平台

[0680] 使用两步法分析数据。首先,使用Meraculous的依情况(ad hoc)修改将成对末端片段鸟枪法数据组装成支架。这些组装的序列以及来自相同样品的Chicago数据用作HiRise的输入。对于这些实验,对Meraculous和HiRise均进行依情况修改以允许(1)改变代表不同物种的支架中的序列覆盖度(即,丰度),和(2)物种内的株间多态性。采用其他宏基因组组装器(例如,Omega和metaSpades)的实验在第一阶段不提供相比修改的Meraculous的实质性改进(未示出)。HiRise最初开发用于二倍体基因组组装,因此假设一致的Chicago和鸟枪法覆盖度。在支架化步骤中将该特征针对宏基因组进行了修改。通过这种组装方法,采用Chicago数据可获得显著的支架大小。这两个步骤也可以被集成用于改进的组装和相异菌株的单独组装。

[0681] 多态性区域的改进组装:在掺入对照实验中,最长的支架来自天蓝色链霉菌(1%掺入),尽管事实是许多其他OTU以更高丰度存在于粪便样品中。重要的是,我们注意到,(克隆)掺入对照在分类上不同于存在的其他OTU,因为所述对照没有菌株变异。因此,通过菌株变异进行检测和组装的有效方法可提高物种水平上的邻接。

[0682] 最初的Meraculous算法被设计用于组装二倍体基因组。在所述设置中,多态性表现为相同频率的两个等位基因变体,使得其总和为二倍体基因组的覆盖度的(一致)深度。这些等位基因变体可以容易地与测序错误区分开,所述测序错误以低水平出现(例如,Illumina数据中<1%)。相反,在宏基因组中,(1)单元型可根据菌株丰度以不同频率发生;(2)菌株的所有单元型的总深度代表物种的丰度,该丰度在物种之间(因此在支架之间)不同;以及(3)在非常丰富的物种中,即使低错误率也可以产生可容易混淆真实变体的经常性错误。

[0683] 因此,对于宏基因组,Meraculous可以适合于(1)允许不同频率的单元型(在deBuijn图中表现为叉),(2)允许深度为局部约束而不是全局限制,以及(3)相对于局部深度的过滤错误,而不是具有全局截断。可对开放源Meraculous代码作出这些改变,并根据经验采用由两个或更多个密切相关的菌株的掺入生成的测试数据进行验证。存在对Meraculous的这些调整的自我一致性的元件,因为局部深度(每个物种的丰度)可从数据中得知。这些方法可针对多种粪便样品进行测试,以确保我们的算法是稳健的。

[0684] 如图26A和图26B所示,初步组装表明,Chicago数据含有用于进一步支架化的剩余的未开发的信息。例如,当前的组装策略可生成许多具有相似GC含量和覆盖深度的未连接支架,相比于具有广泛不同的GC含量或深度的支架,所述未连接支架更可能代表来自相同物种的支架。以依情况方式对这些支架进行分组是最初分箱策略的基础,其可以被认为是进行进一步连接的假设。

[0685] 关于Chicago数据是否可以提供这些假设的独立的实验证实,进行进一步的研究。图26A和图26B示出了通过Chicago读取对高度连接的鸟枪法支架更有可能具有相似的GC含量和覆盖深度。图26A示出了掺入实验中所有支架的覆盖深度(y-轴)和GC含量(色标);支架的条纹位于很可能来自相同OTU的相似覆盖度和GC含量。图26B在x轴上示出了每个支架的Chicago连通性,其作为所有Chicago连接相对于第1至第4个连接最多的支架的分数,而y轴示出了在支架对之间的GC+倍数覆盖空间的Euclidean距离;与Chicago连接高度连接的支

架对倾向于具有相似的GC含量和覆盖倍数。与微生物隔离群的已知基因组的比较进一步支持了这些是由Chicago读取对支持的，但不是由当前HiRise算法制作的接头。多种方法可用于对此进行校正。首先，可以分析由HiRise给予这些未制作的接头的内部权重，并且可以采用改进的启发法，由掺入的真值或来自已知基因组的外部支持来指导。第二，可采用明确考虑GC含量和深度的启发法。

[0686] GC含量和深度是将支架划分成假定的连接组的方式。自最初的Tyson报告以来，已经开发出了更精细的方法，并且基于支架特征的不同统计特征(例如，四聚体频率)有多种方法来解决这个问题。也可以从Chicago数据中提取完整的连接信息。

[0687] 为了实现单独组装菌株的目标，可以采用能实现以下迭代方法的软件模块：

[0688] (1)将所有读取映射回初始Meraculous/HiRise组装体。BWA-MEM是一种通用的比对器(aligner)，它可以容易地比对高达3-4％相异的序列，如菌株变异所预期的；

[0689] (2)鉴定这些比对中的变异位置并“定相”，以提取单元型。包括GATK和HapCut在内的现有方法可适于与宏基因组一起使用，从而显著地预期超过两种单元型的可能性和不相等的频率。从鸟枪法序列鉴定单元型可受制于读取长度，因为定相需要读取/读取对映射至多个变体上；以及

[0690] (3)最后，随着在合适的多态性区域中鉴定出单元型，可鉴定与这些单元型相匹配的Chicago读取，并且Chicago对可用于产生菌株特异性支架化。根据菌株的(strain-aware)组装可以显著提高组装质量，因为不同的菌株往往显示出结构可变性；如果多个这样的菌株折叠成一个“共有”物种组装体，则支架化将终止于结构差异(参见图27)。图27示出了菌株变异对支架化性能的影响的图；示出了每个支架的长度与显示出菌株变异(备选碱基)证据的位点的分数，其中在顶部鉴定出最多变异的支架。

[0691] 虽然本公开内容的优选实施方案已经在本文中示出和描述，但对于本领域技术人员来说显而易见的是，这样的实施方案仅通过示例的方式提供。在不脱离本公开内容的情况下，本领域技术人员现在将会想到许多改变、变化和替换。应当理解，本文所描述的公开内容的实施方案的各种备选方案可用于实施本公开内容。旨在由以下权利要求限定本公开内容的范围，并旨在由此涵盖这些权利要求的范围内的方法和结构及其等同物。

<110> 多弗泰尔基因组学有限责任公司

<120> 用于基因组组装、单元型定相以及独立于靶标的核酸检测的方法

<130> 45269-713.601

<140> PCT/US2016/057557
<141> 2016-10-18

<150> 62/294,198
<151> 2016-02-11

<150> 62/255,953
<151> 2015-11-16

<150> 62/243,576
<151> 2015-10-19

<150> 62/243,591
<151> 2015-10-19

[0001]

<160> 18

<170> PatentIn version 3.5

<210> 1
<211> 33
<212> DNA
<213> 人工序列

<220>
<223> 人工序列的描述：合成
寡核苷酸

<400> 1
acactctttc cctacacgac gctcttccga tgt 33

<210> 2
<211> 32
<212> DNA
<213> 人工序列

<220>
<223> 人工序列的描述：合成
寡核苷酸

<400> 2
catcggaaga gcacacgtct gaactccagt ca 32

<210> 3
<211> 33
<212> DNA
<213> 人工序列

<220>
<223> 人工序列的描述：合成
寡核苷酸

<400> 3
acactctttc cctacacgac gctcttccga cct 33

<210> 4
[0002] <211> 32
<212> DNA
<213> 人工序列

<220>
<223> 人工序列的描述：合成
寡核苷酸

<400> 4
ggtcggaaga gcacacgtct gaactccagt ca 32

<210> 5
<211> 33
<212> DNA
<213> 人工序列

<220>
<223> 人工序列的描述：合成
寡核苷酸

<400> 5
acactctttc cctacacgac gctctaccga tct 33

<210> 6
<211> 32
<212> DNA
<213> 人工序列

<220>
<223> 人工序列的描述：合成
寡核苷酸

<400> 6
gatcggtaga gcacacgtct gaactccagt ca 32

<210> 7
<211> 33
<212> DNA
<213> 人工序列

<220>
<223> 人工序列的描述：合成
寡核苷酸

[0003]

<400> 7
acactctttc cctacacgac gctattccga tct 33

<210> 8
<211> 32
<212> DNA
<213> 人工序列

<220>
<223> 人工序列的描述：合成
寡核苷酸

<400> 8
gatcggaata gcacacgtct gaactccagt ca 32

<210> 9
<211> 33
<212> DNA
<213> 人工序列

<220>
<223> 人工序列的描述：合成
寡核苷酸

<400> 9
acactctttc cctacacgac gctcttcgga tct 33

<210> 10
<211> 32
<212> DNA
<213> 人工序列

<220>
<223> 人工序列的描述：合成
寡核苷酸

<400> 10
gatccgaaga gcacacgtct gaactccagt ca 32

[0004]

<210> 11
<211> 33
<212> DNA
<213> 人工序列

<220>
<223> 人工序列的描述：合成
寡核苷酸

<400> 11
acactctttc cctacacgac cctcttccga tct 33

<210> 12
<211> 32
<212> DNA
<213> 人工序列

<220>
<223> 人工序列的描述：合成
寡核苷酸

<400> 12

gatcggaaga ggacacgtct gaactccagt ca 32

<210> 13
<211> 33
<212> DNA
<213> 人工序列

<220>
<223> 人工序列的描述：合成
寡核苷酸

<400> 13
acactctttc cctacacgac gcacttccga tct 33

<210> 14
<211> 32
<212> DNA
<213> 人工序列

<220>
[0005] <223> 人工序列的描述：合成
寡核苷酸

<400> 14
gatcggaagt gcacacgtct gaactccagt ca 32

<210> 15
<211> 33
<212> DNA
<213> 人工序列

<220>
<223> 人工序列的描述：合成
寡核苷酸

<400> 15
acactctttc cctacacgac gctcttccga tct 33

<210> 16
<211> 32
<212> DNA

<213> 人工序列

<220>
<223> 人工序列的描述：合成
寡核苷酸

<400> 16
gatcggaaga gcacacgtct gaactccagt ca 32

<210> 17
<211> 10
<212> DNA
<213> 人工序列

<220>
<223> 人工序列的描述：合成
寡核苷酸

<400> 17
aagctagctt 10

[0006]

<210> 18
<211> 10
<212> DNA
<213> 人工序列

<220>
<223> 人工序列的描述：合成
寡核苷酸

<220>
<221> 修饰的碱基
<222> (1)..(3)
<223> a、c、t、g、未知的或其他

<220>
<221> 修饰的碱基
<222> (8)..(10)
<223> a、c、t、g、未知的或其他

<400> 18
nnngatcnnn 10

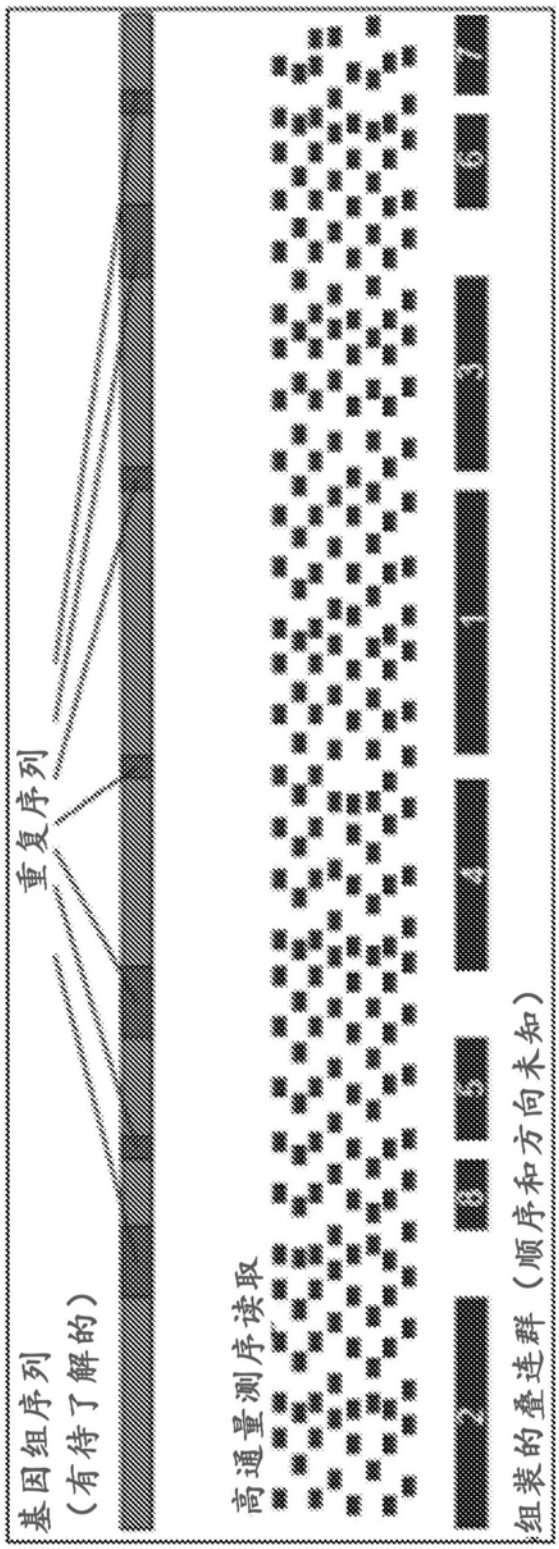


图1

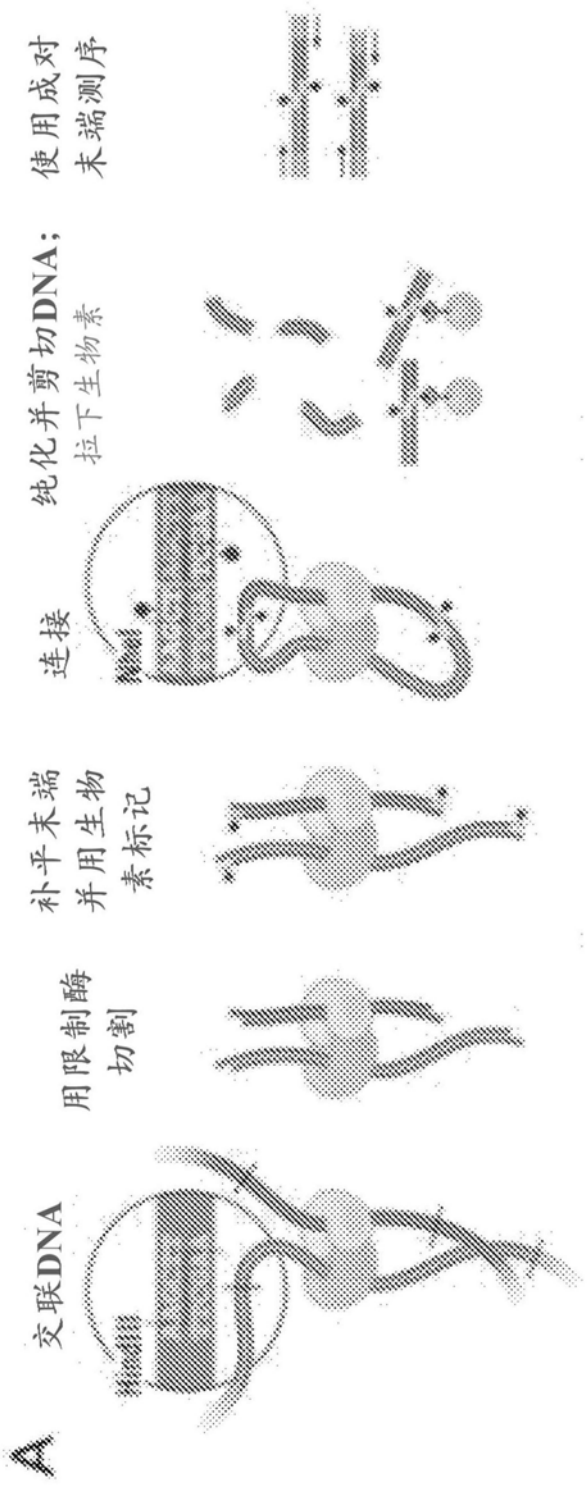


图2A

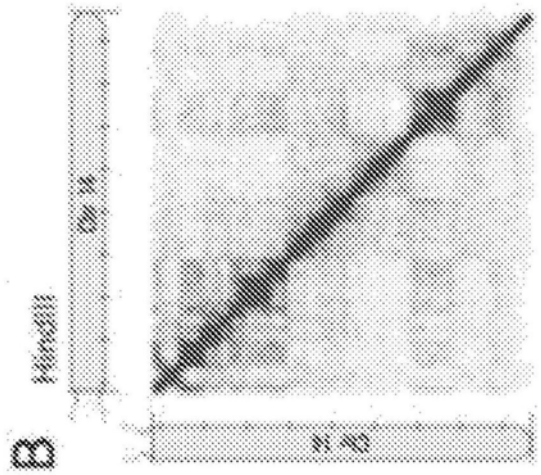


图2B

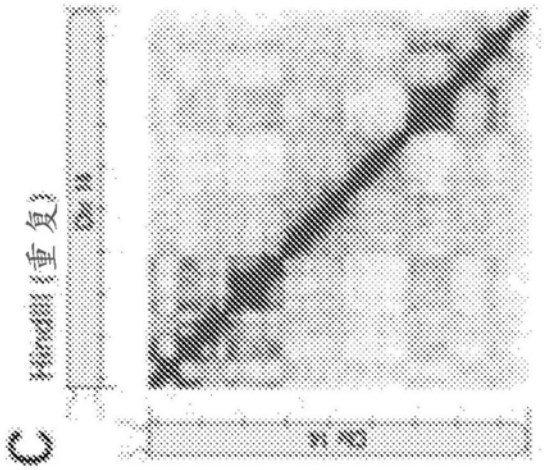


图2C

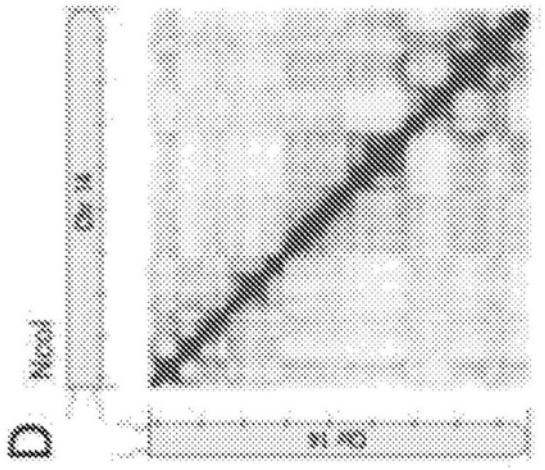


图2D

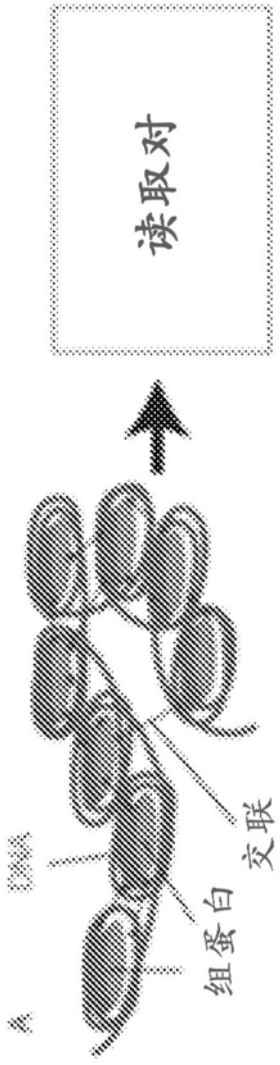


图3A

图3B

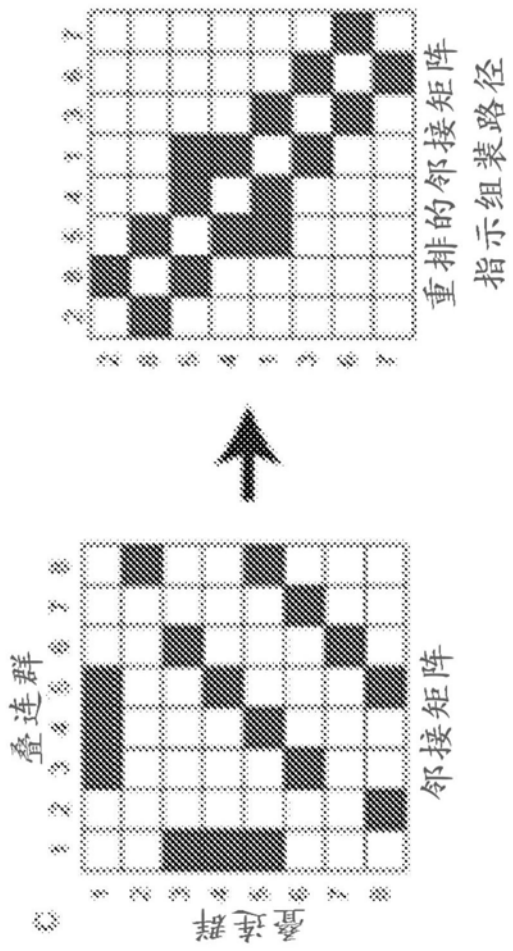


图3C

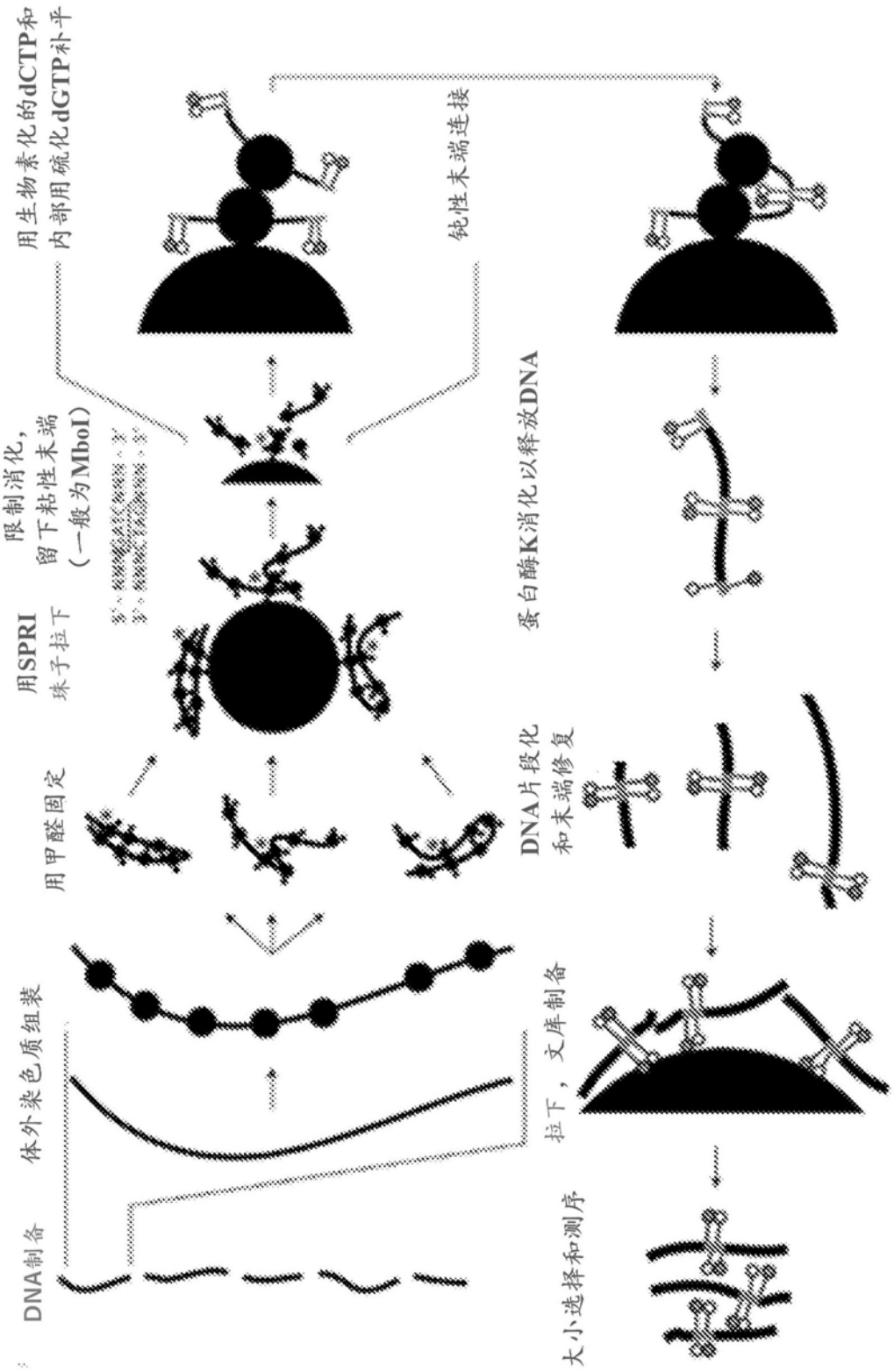

DNA制备
体外染色质组装
用甲醛固定
用SPRI
珠子拉下
限制消化，
留下粘性末端
（一般为MboI）
用生物素化的dCTP和
内部用硫化dGTP补平
钝性末端连接
蛋白酶K消化以释放DNA
DNA片段化
和末端修复
拉下，文库制备
大小选择和测序

图4

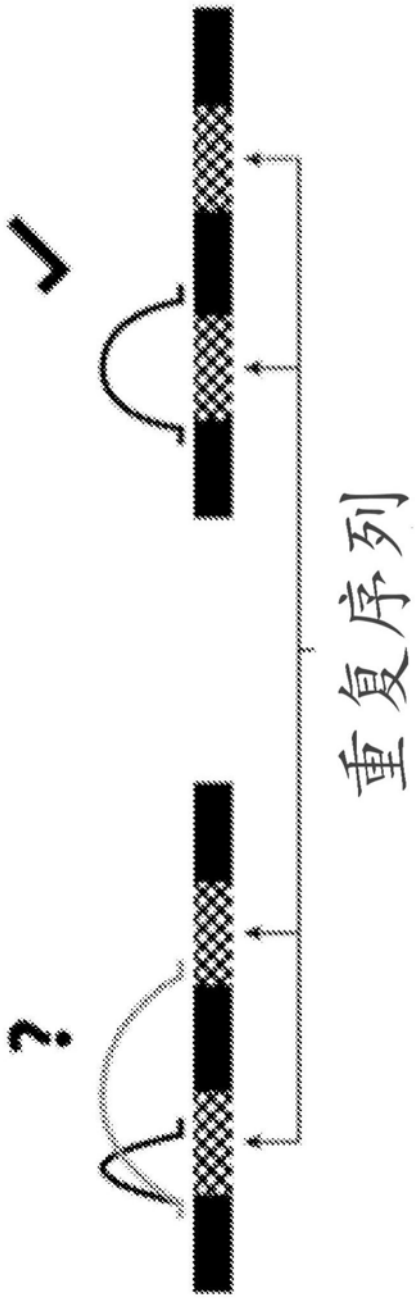


图5A

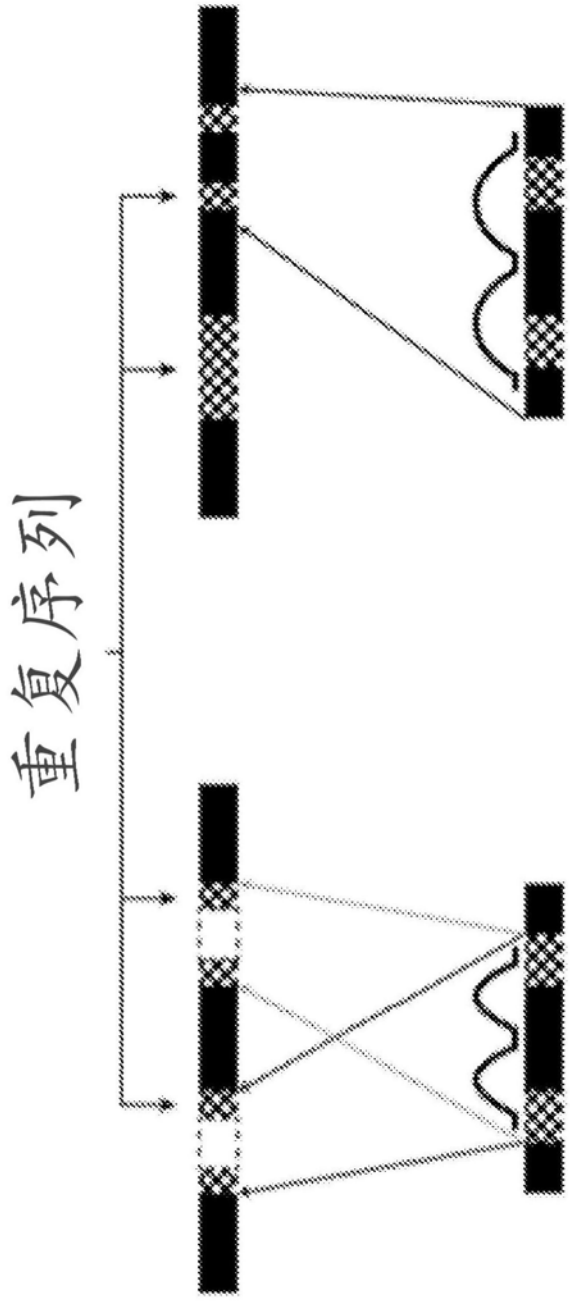


图5B

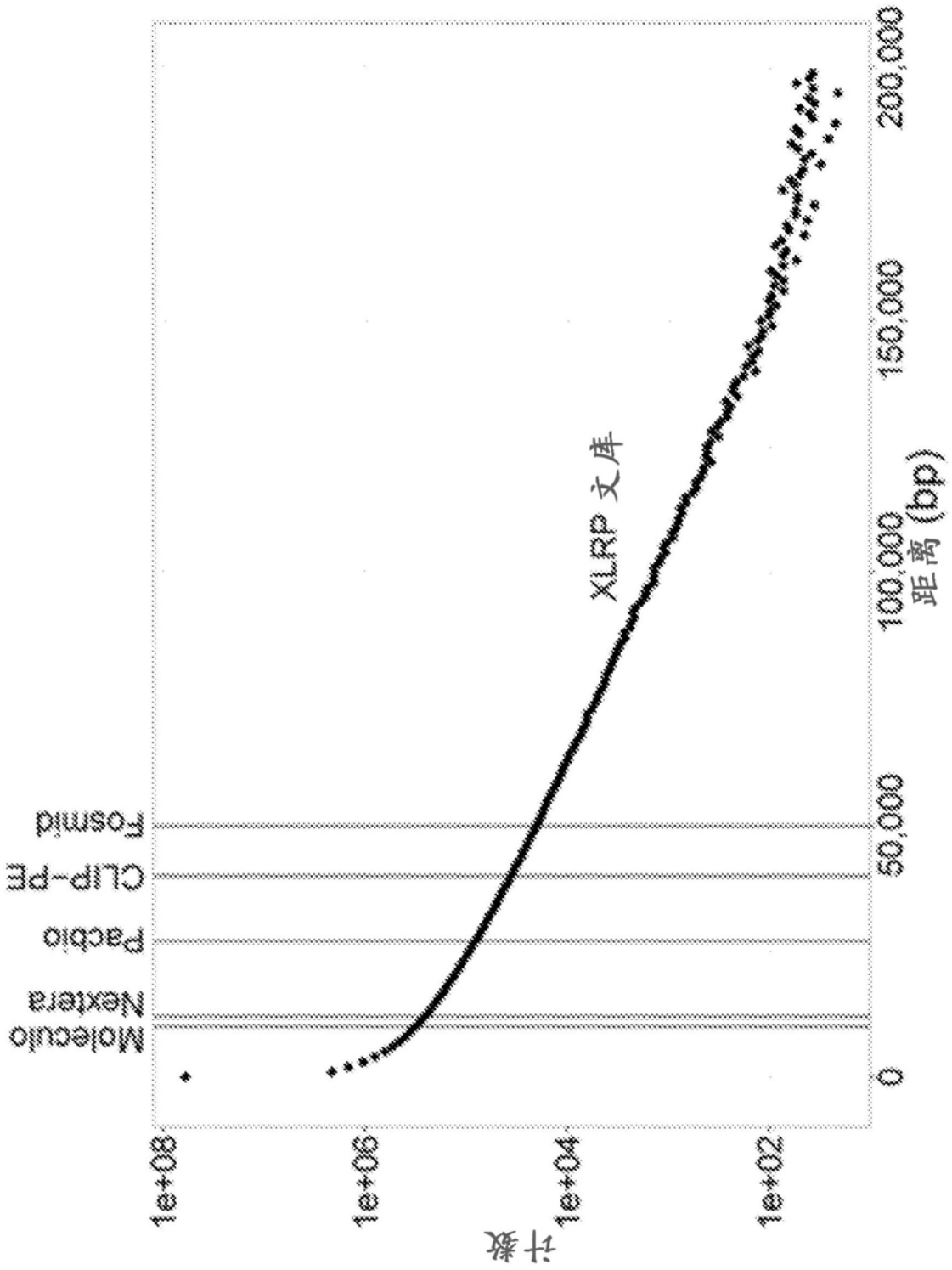

XLRP 文库
计数
1e+08
1e+06
1e+04
1e+02
0
50,000
100,000
150,000
200,000
距离 (bp)
Fosmid
CLIP-PE
Pacbio
Nextera
Molecula

图6

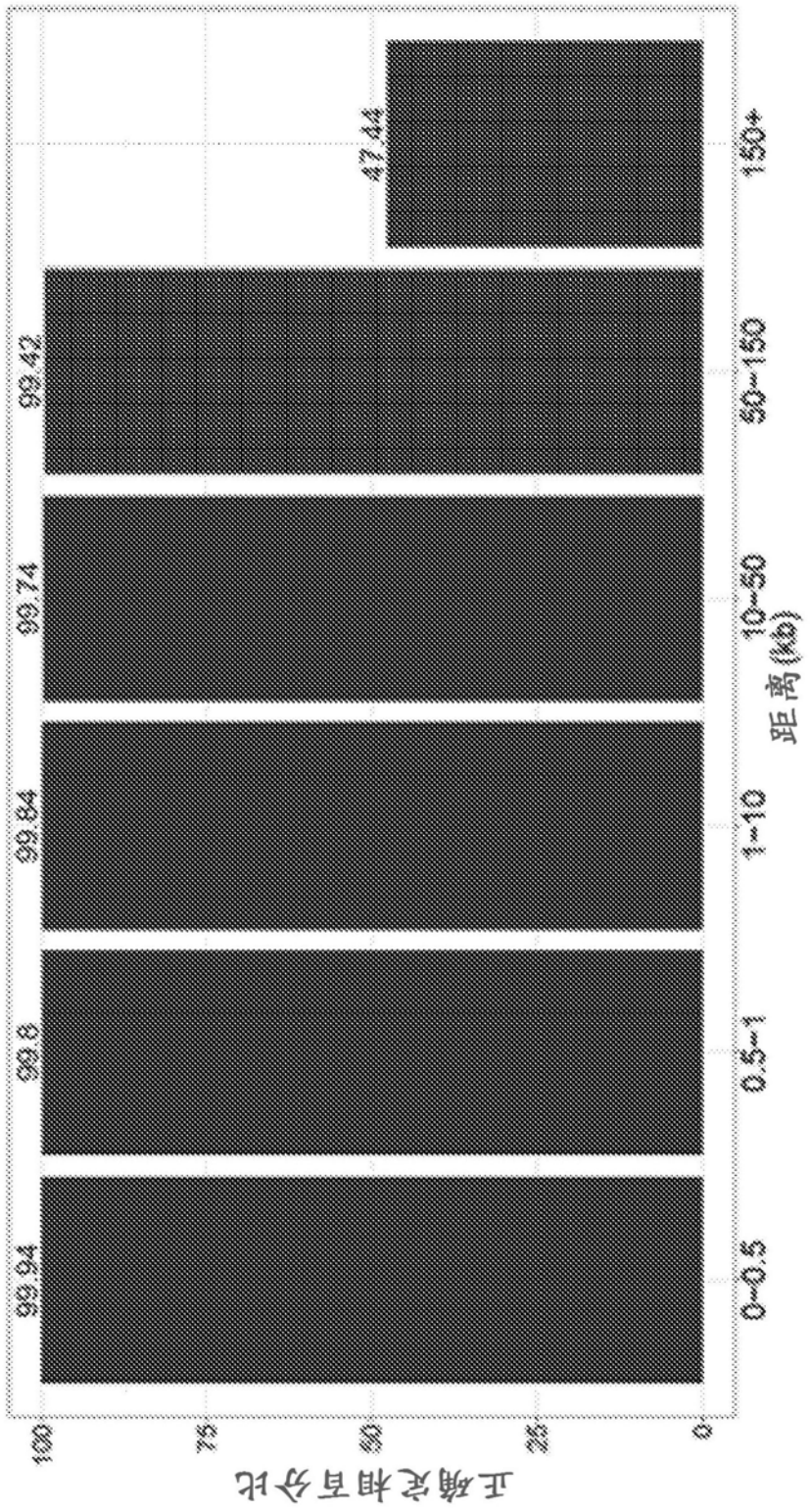

99.94
99.9
99.84
99.74
99.42
47.44
0-0.5
0.5-1
1-10
10-50
50-150
150+
距离(kb)
0
25
50
75
100
正确定相百分比

图7

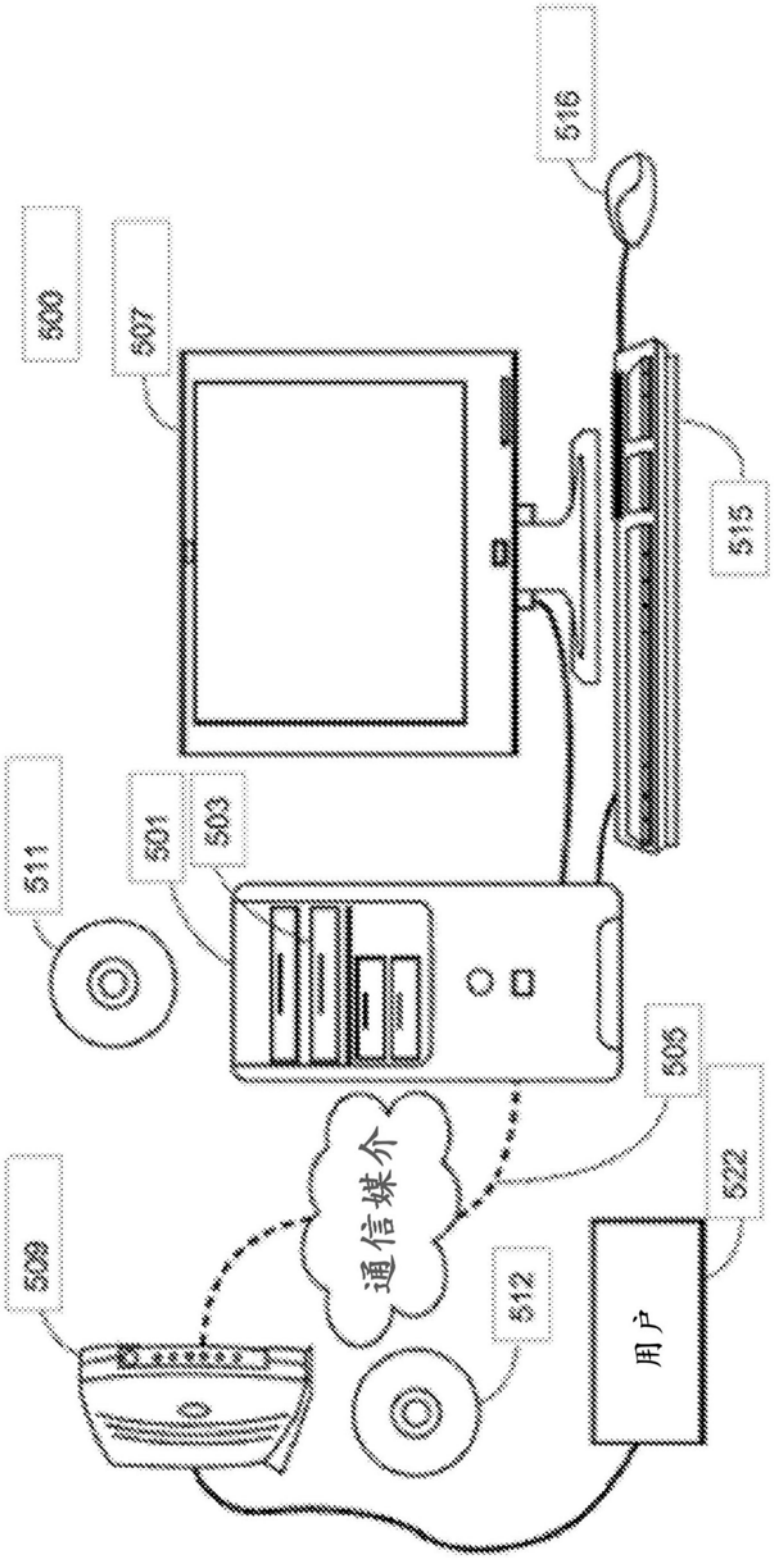


图8

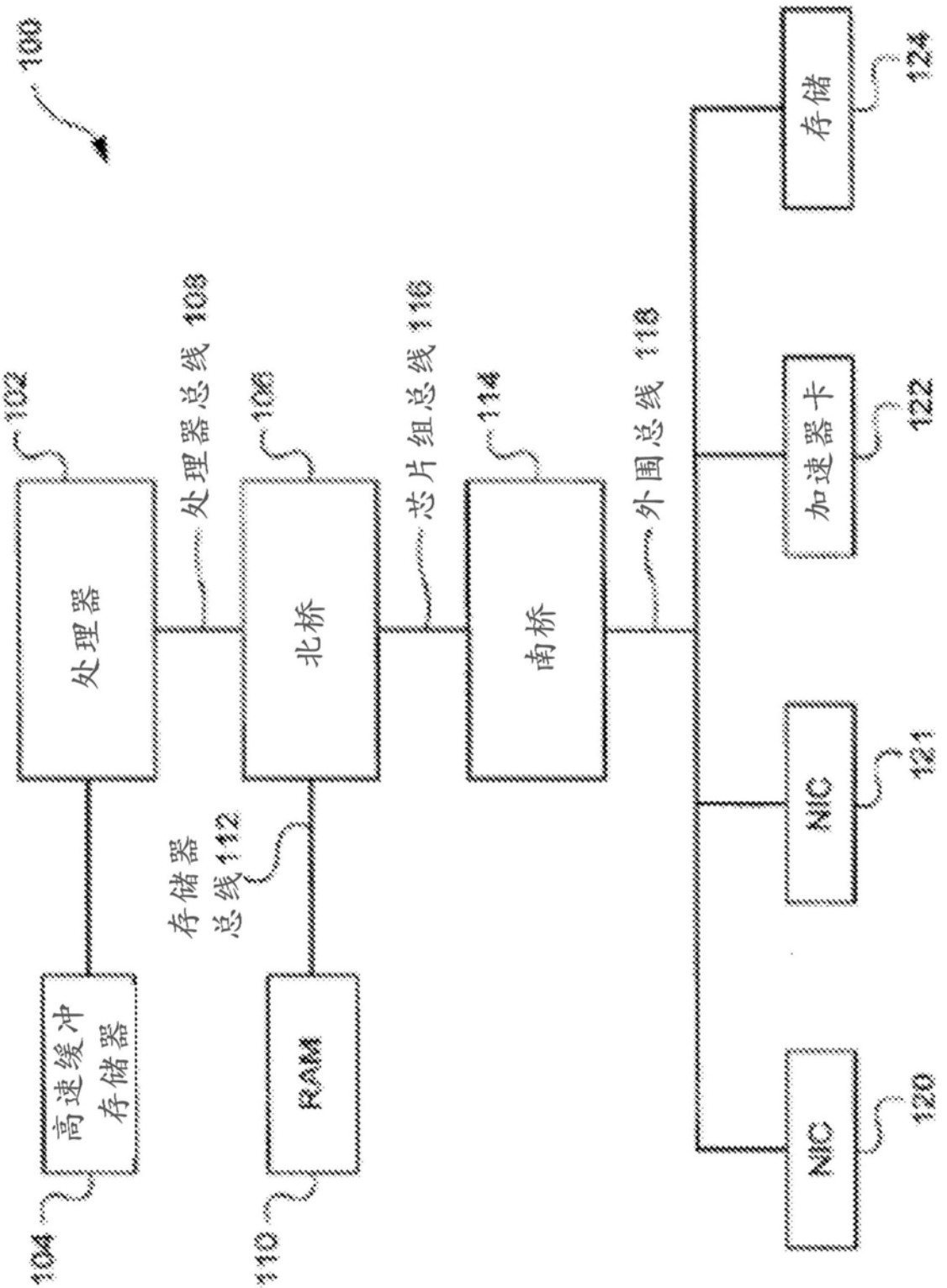

100
102
处理器
处理器总线 108
106
北桥
芯片组总线 116
114
南桥
外围总线 118
存储 124
加速器卡 122
NIC 121
NIC 120
存储器总线 112
RAM
110
高速缓冲存储器
104

图9

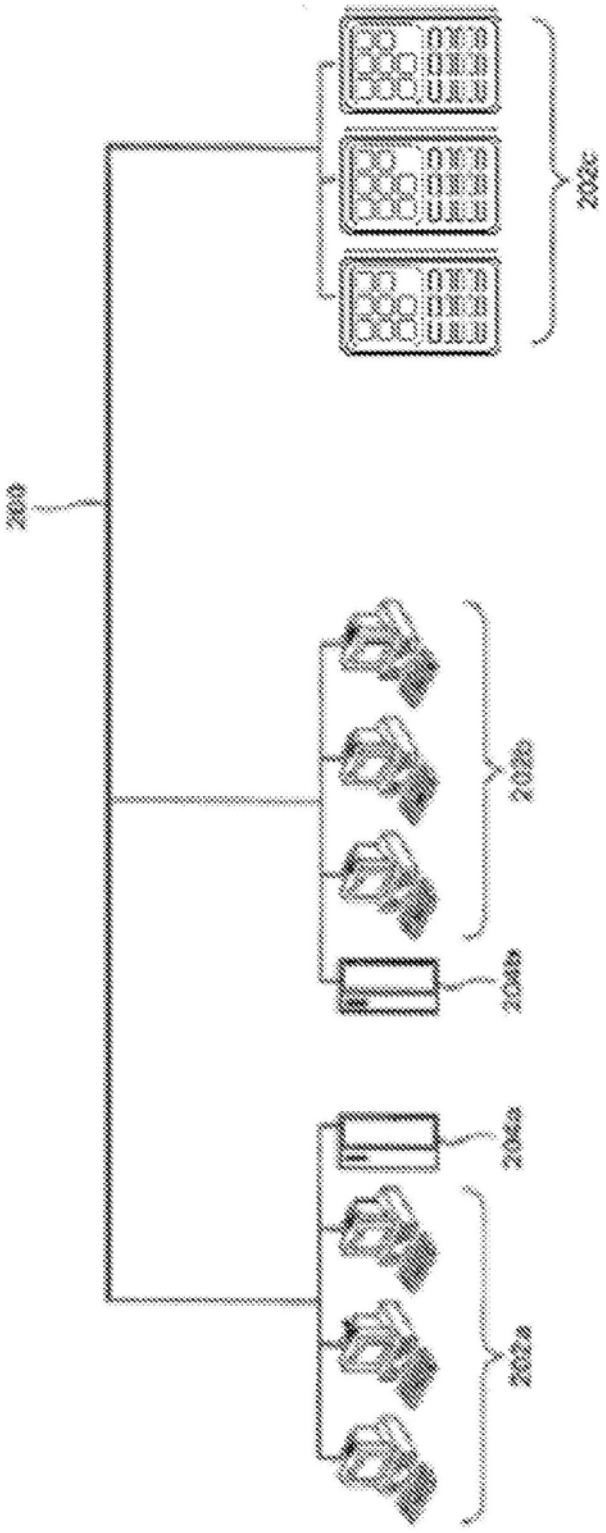


图10

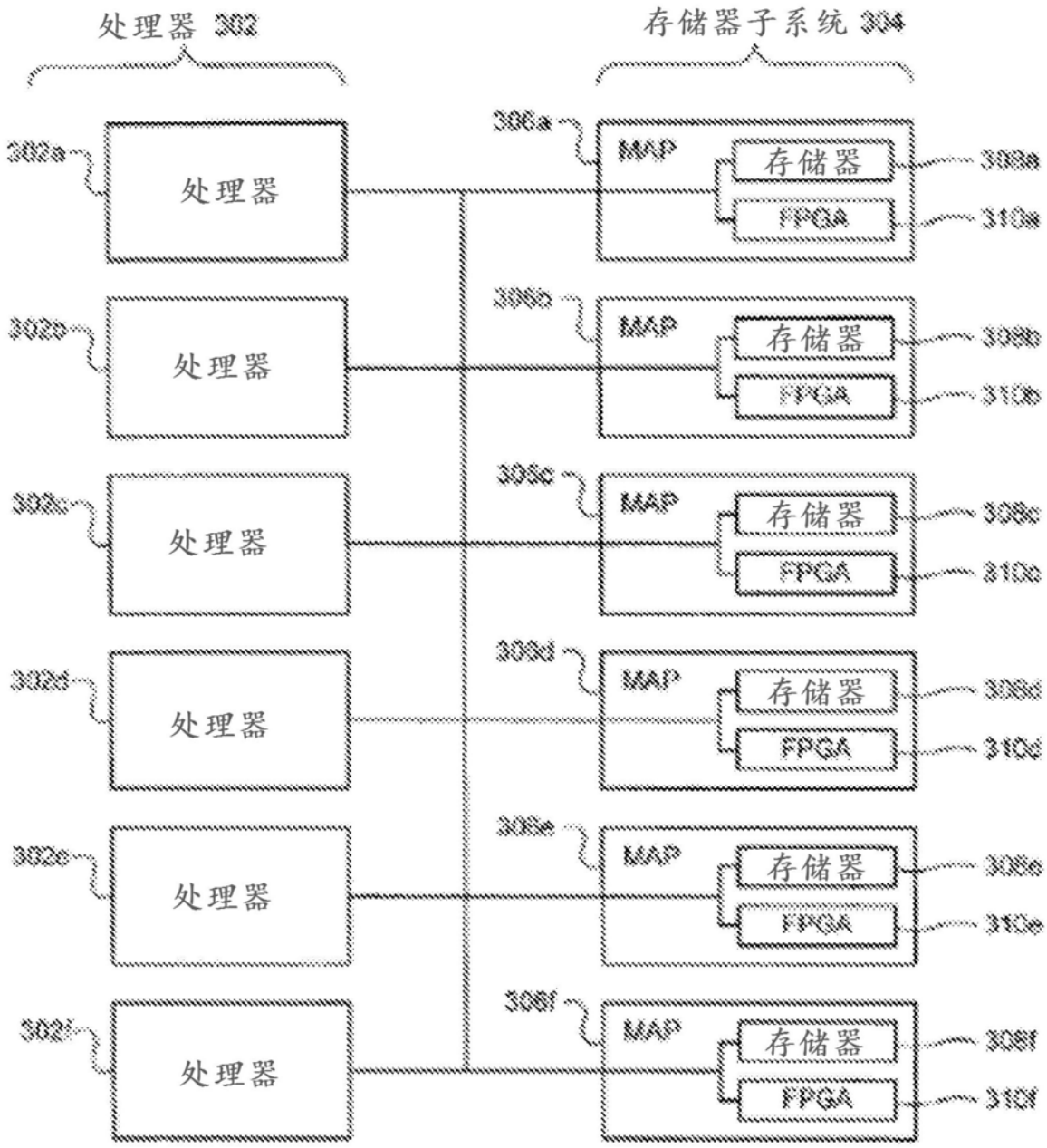

处理器 302
存储器子系统 304
302a
处理器
306a
MAP
存储器
FPGA
308a
310a
302b
处理器
306b
MAP
存储器
FPGA
308b
310b
302c
处理器
306c
MAP
存储器
FPGA
308c
310c
302d
处理器
306d
MAP
存储器
FPGA
308d
310d
302e
处理器
306e
MAP
存储器
FPGA
308e
310e
302f
处理器
306f
MAP
存储器
FPGA
308f
310f

图11

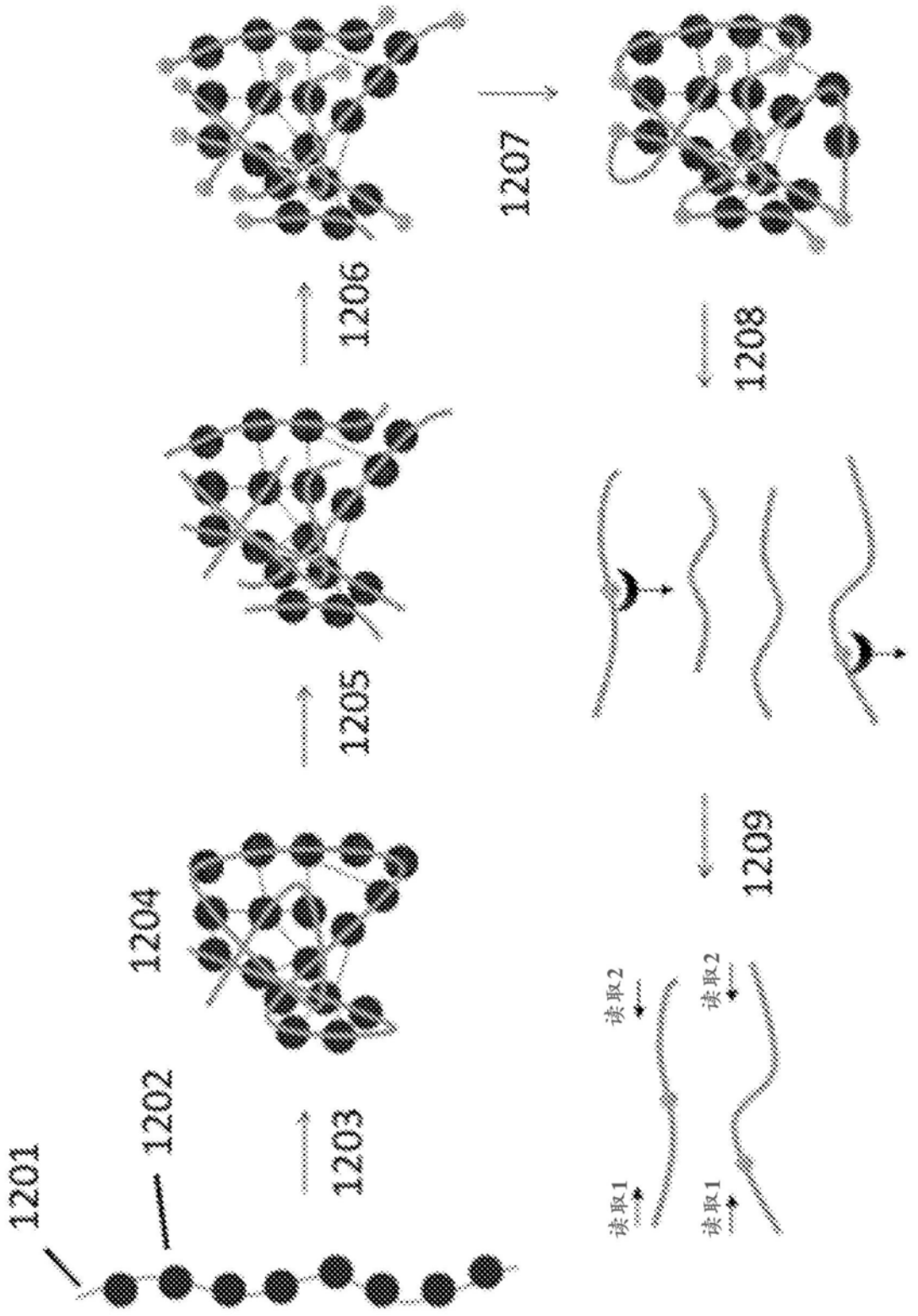


图12A

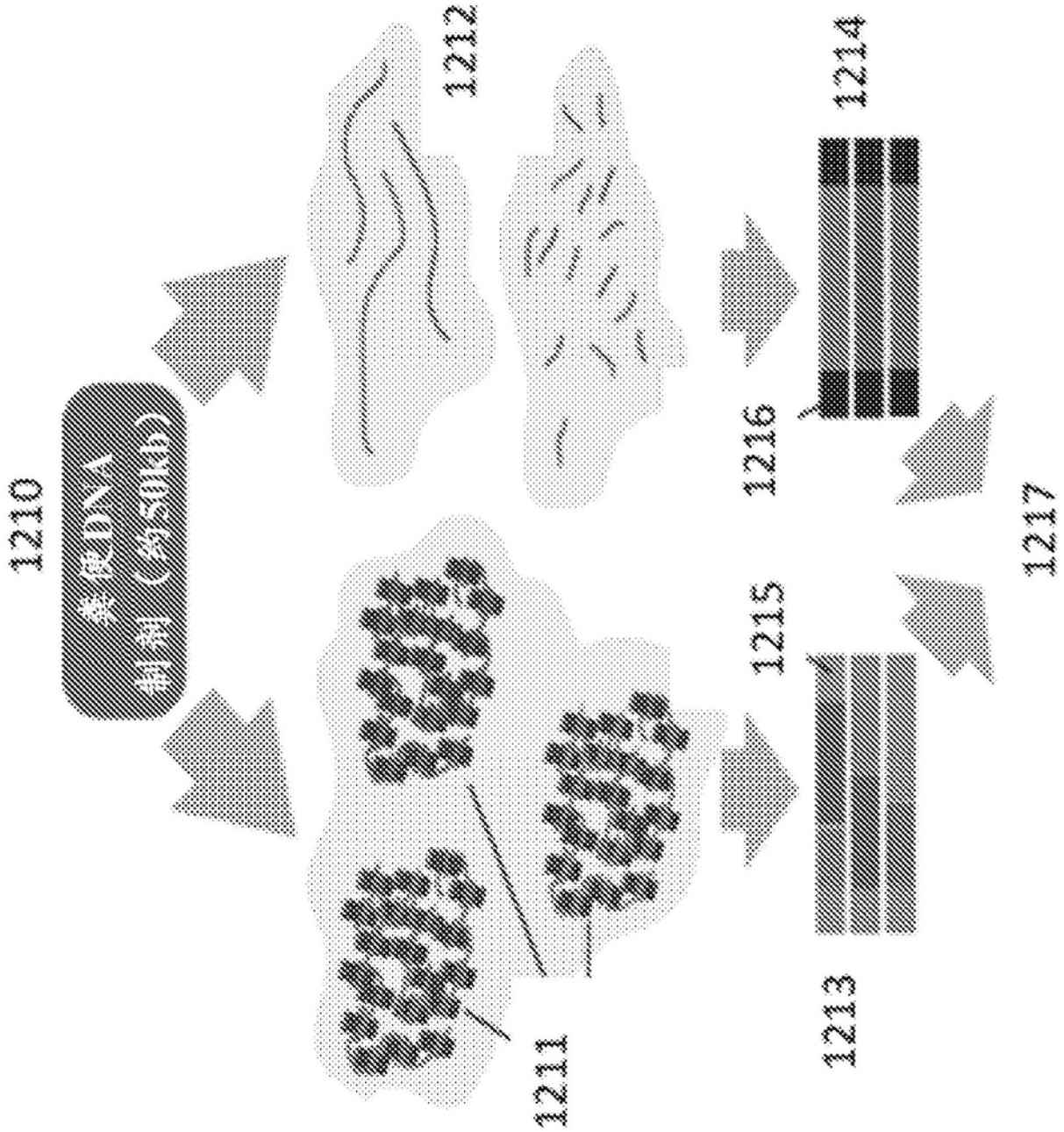


图12B

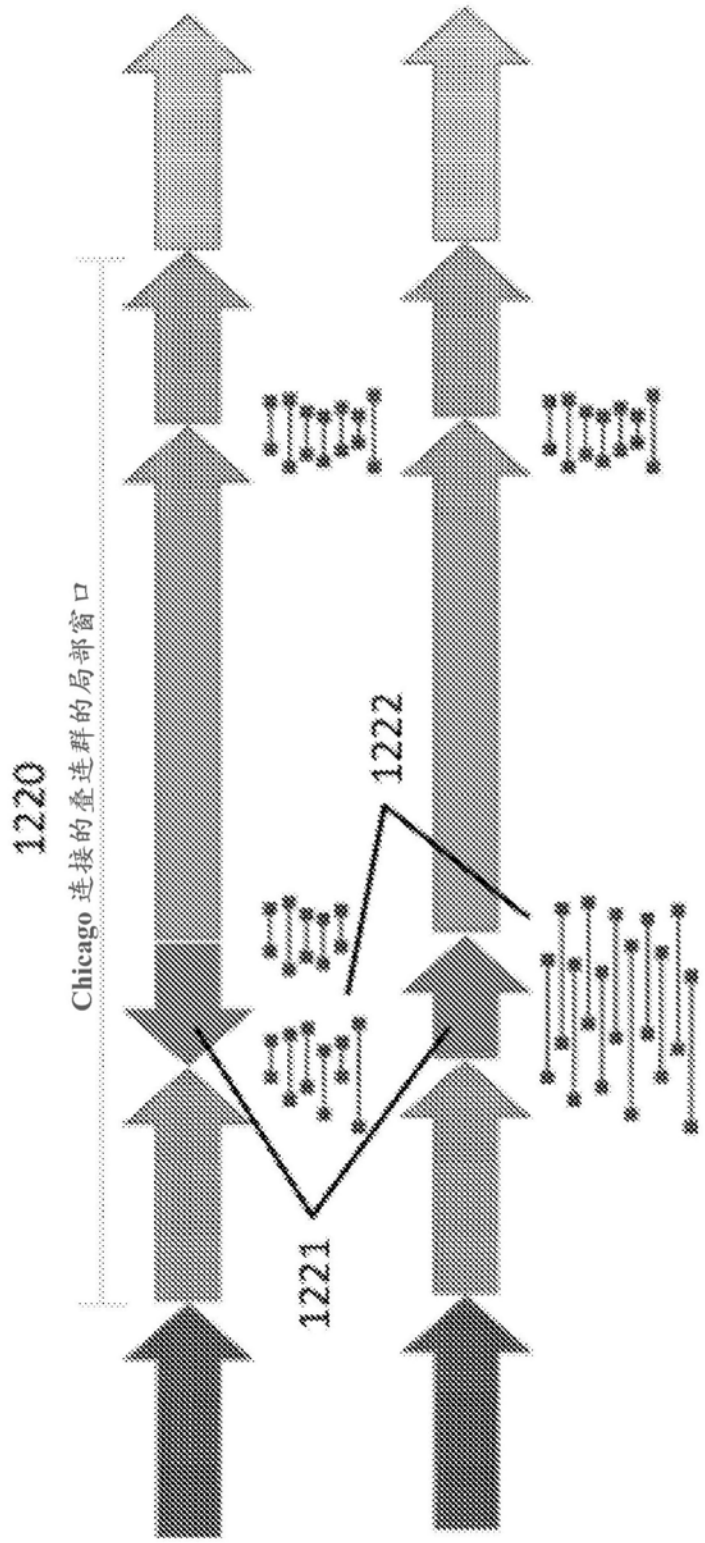


图12C

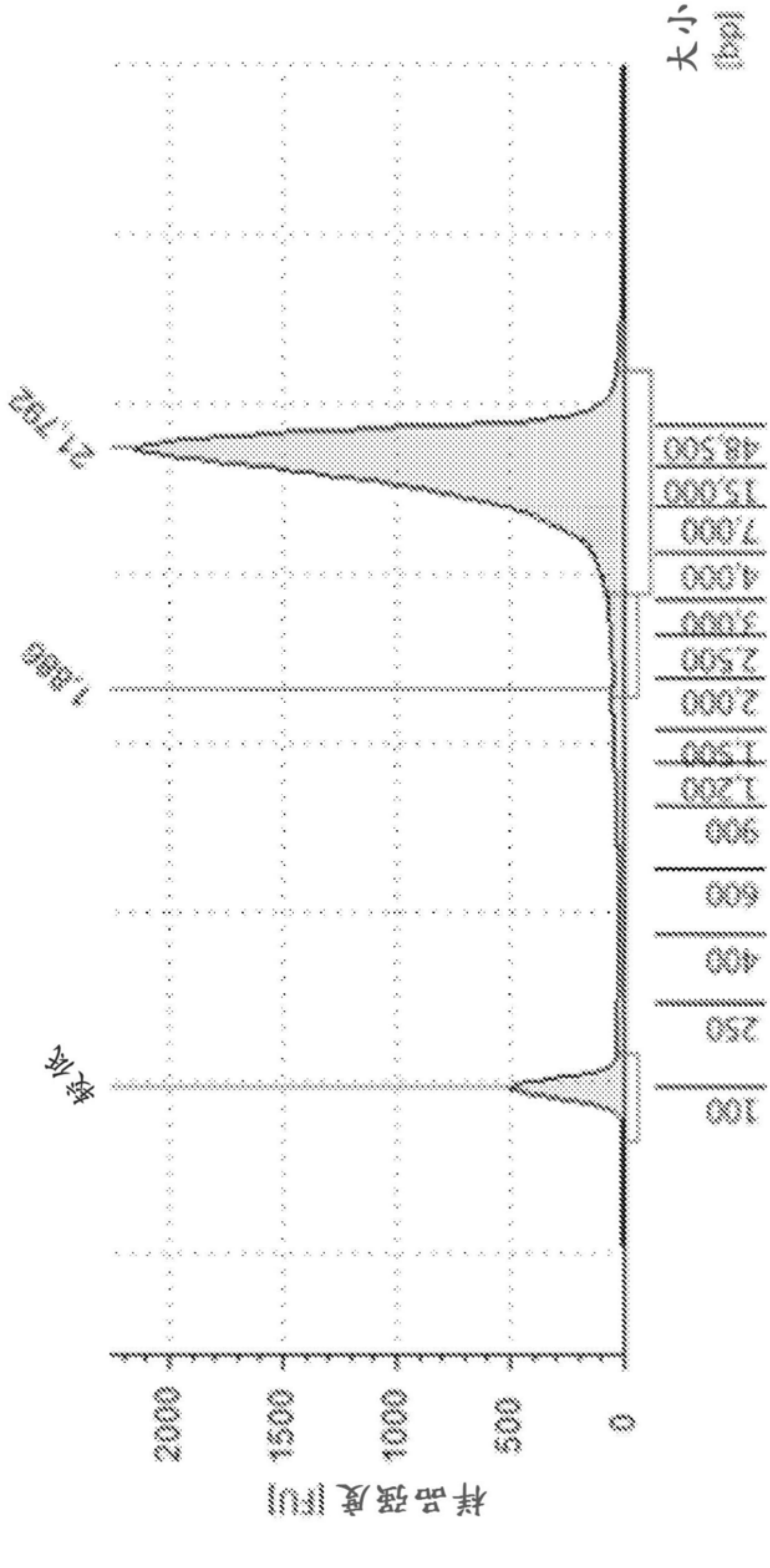

大小 [bp]
48,500
15,000
7,000
4,000
3,000
2,500
2,000
1,500
1,200
900
600
400
250
100
21,792
1,886
样品强度 [FU]
2000
1500
1000
500
0

图13A

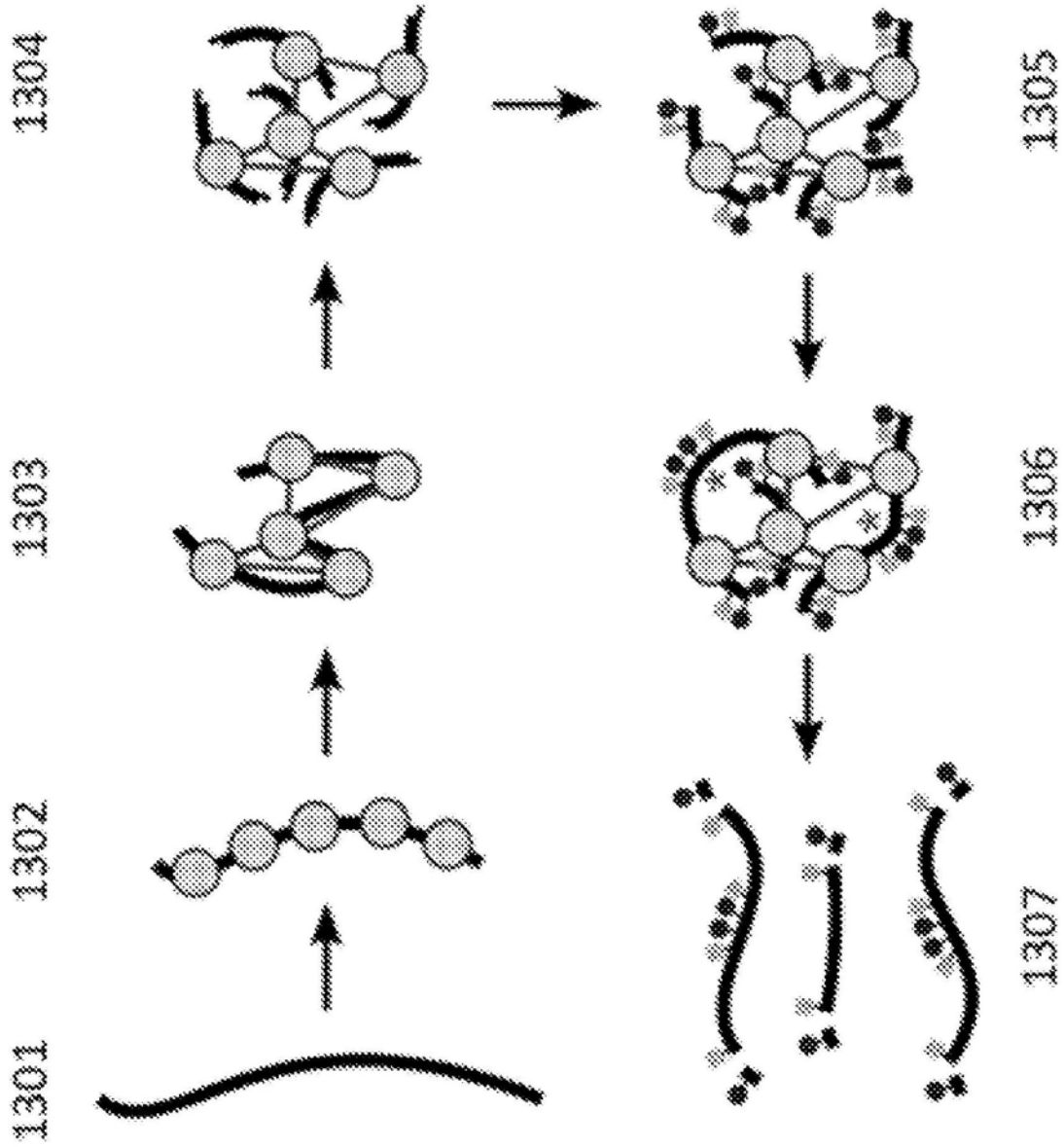


图13B

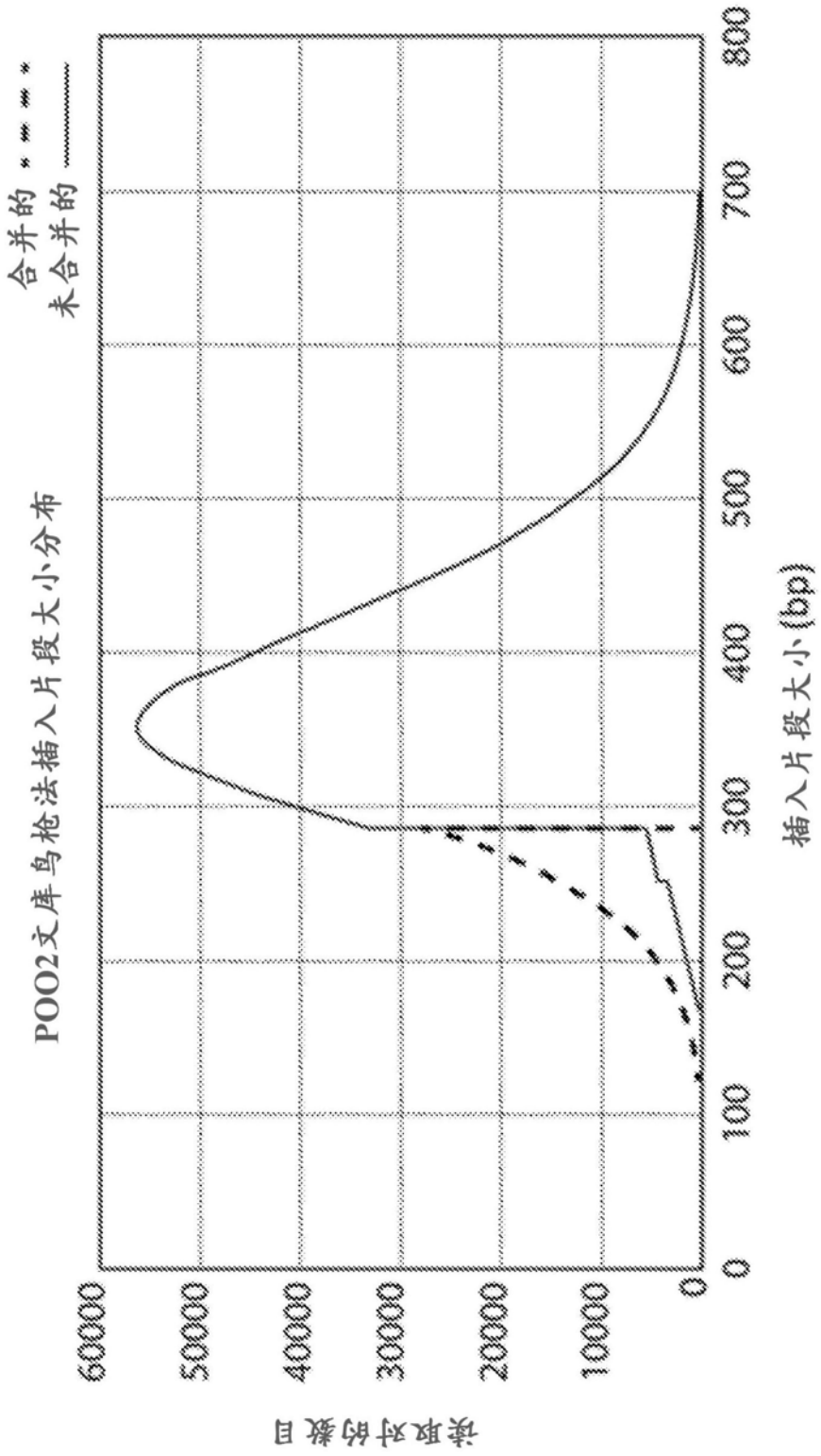

PO02文库鸟枪法插入片段大小分布
合并的
未合并的
读取对的数目
插入片段大小(bp)
60000
50000
40000
30000
20000
10000
0
0
100
200
300
400
500
600
700
800

图14

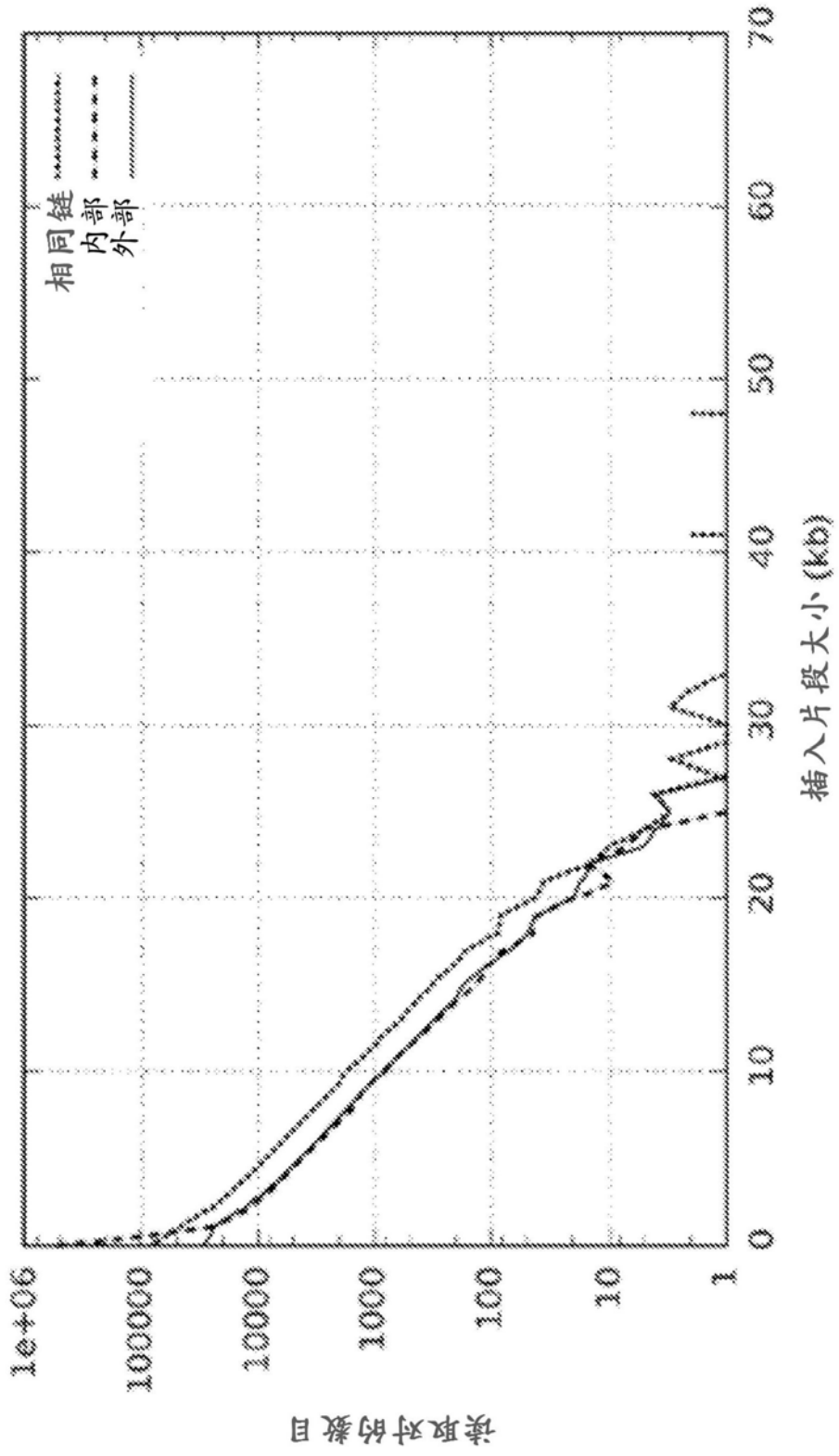

相同链
内部
外部
读取对的数目
插入片段大小 (kb)
1e+06
100000
10000
1000
100
10
1
0
10
20
30
40
50
60
70

图15

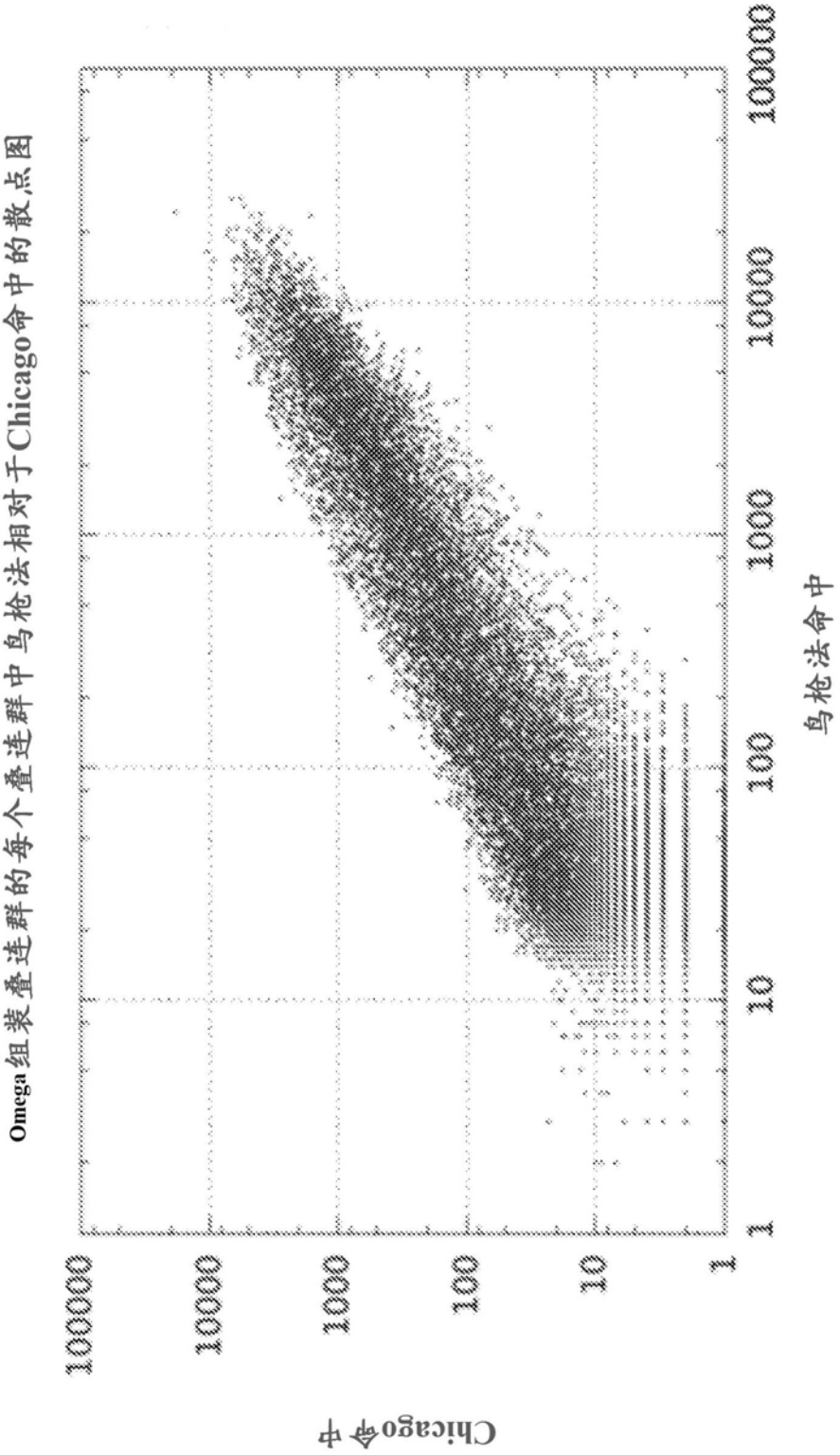

Omega组装叠连群的每个叠连群中鸟枪法相对于Chicago命中的散点图
Chicago命中
鸟枪法命中
100000
10000
1000
100
10
1
1
10
100
1000
10000
100000

图16

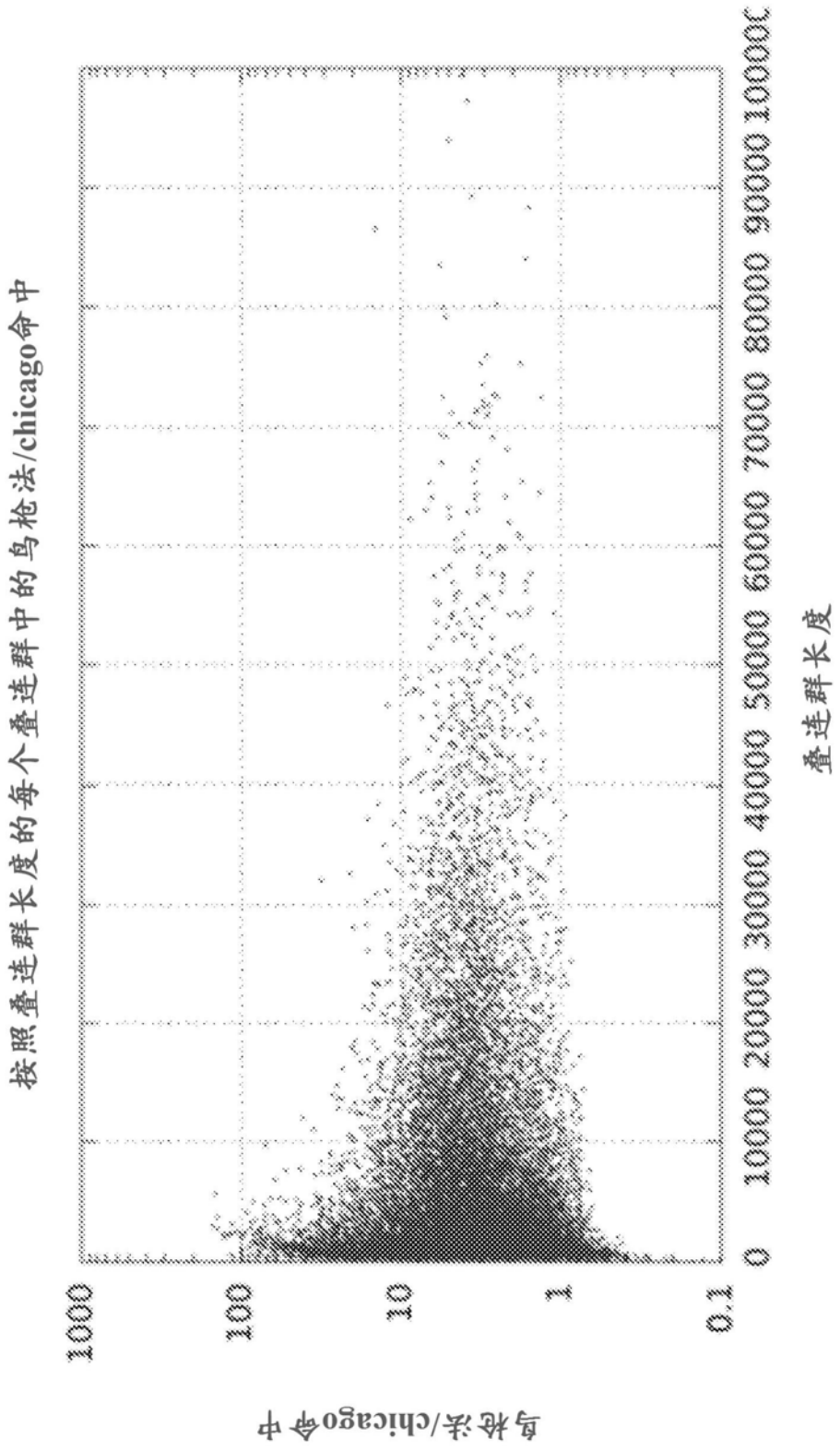

按照叠连群长度的每个叠连群中的鸟枪法/chicago命中
叠连群长度
鸟枪法/chicago命中
0
10000
20000
30000
40000
50000
60000
70000
80000
90000
100000
1000
100
10
1
0.1

图17

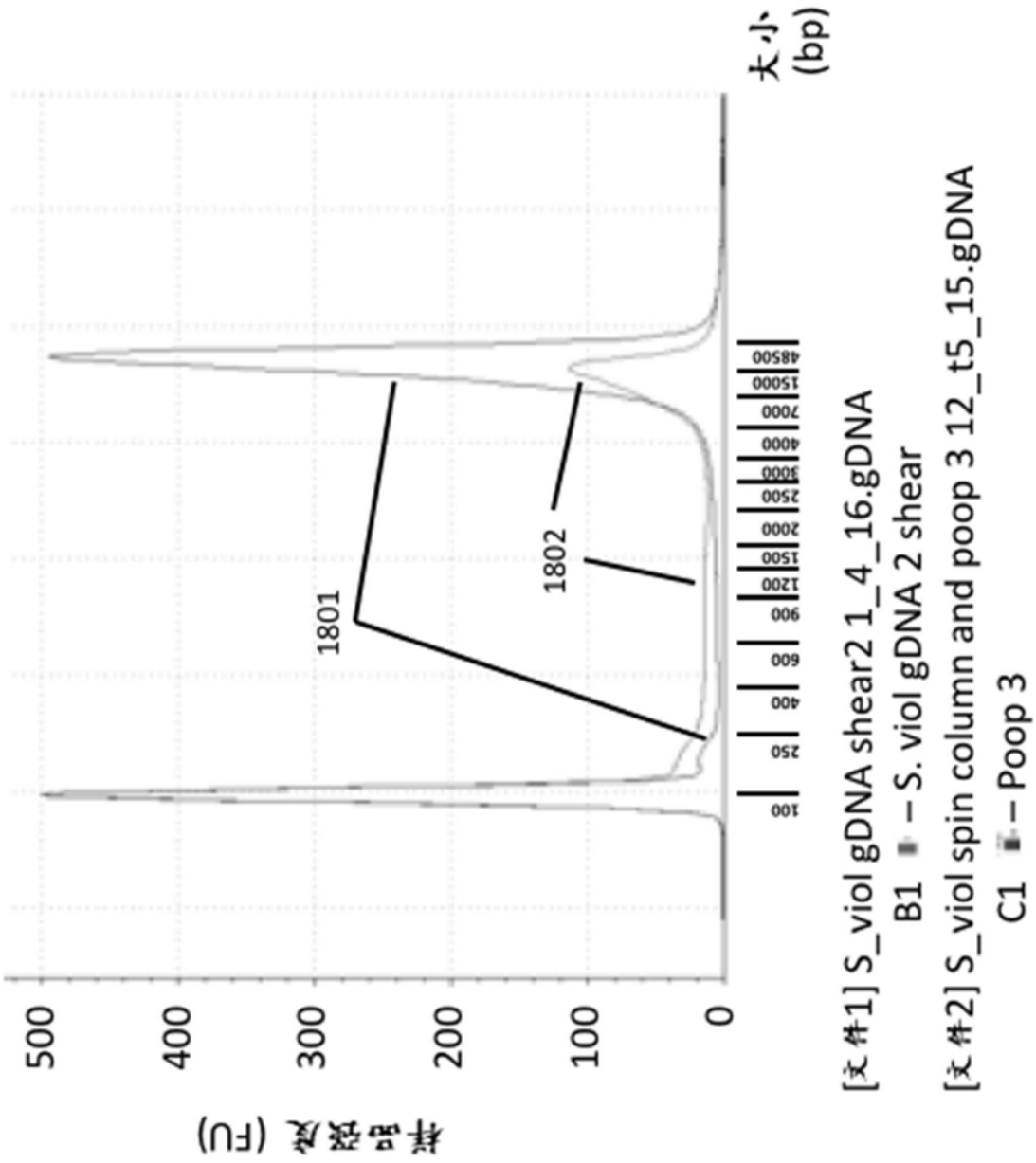

荧光强度 (FU)
500
400
300
200
100
0
大小 (bp)
100
250
400
600
900
1200
1500
2000
2500
3000
4000
7000
15000
48500
1801
1802
[文件1] S_viol gDNA shear2 1_4_16.gDNA
B1 – S. viol gDNA 2 shear
[文件2] S_viol spin column and poop 3 12_t5_15.gDNA
C1 – Poop 3

图18

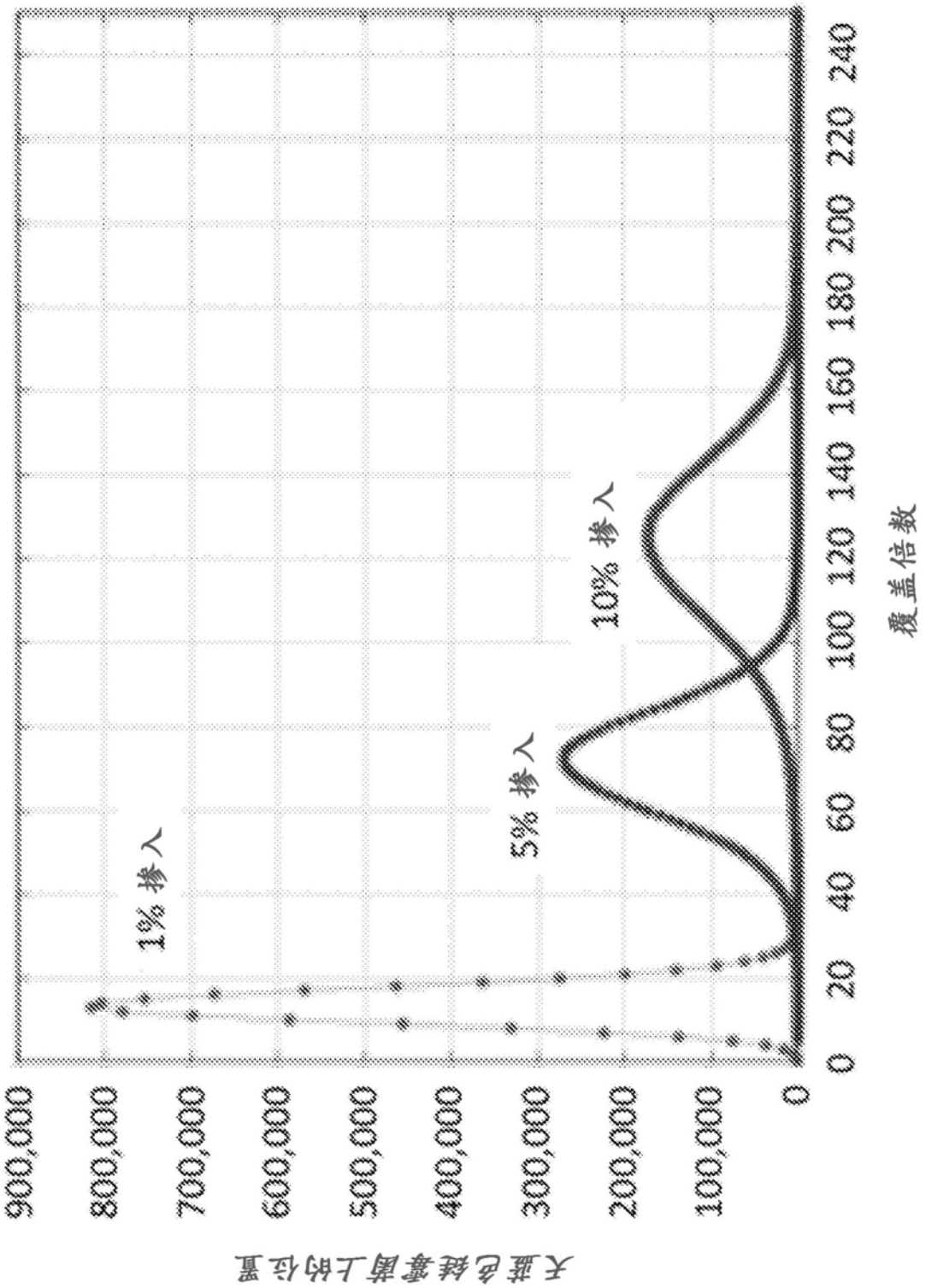

900,000
800,000
700,000
600,000
500,000
400,000
300,000
200,000
100,000
0
0 20 40 60 80 100 120 140 160 180 200 220 240
覆盖倍数
1% 掺入
5% 掺入
10% 掺入
无意色线赛道上的位置

图19

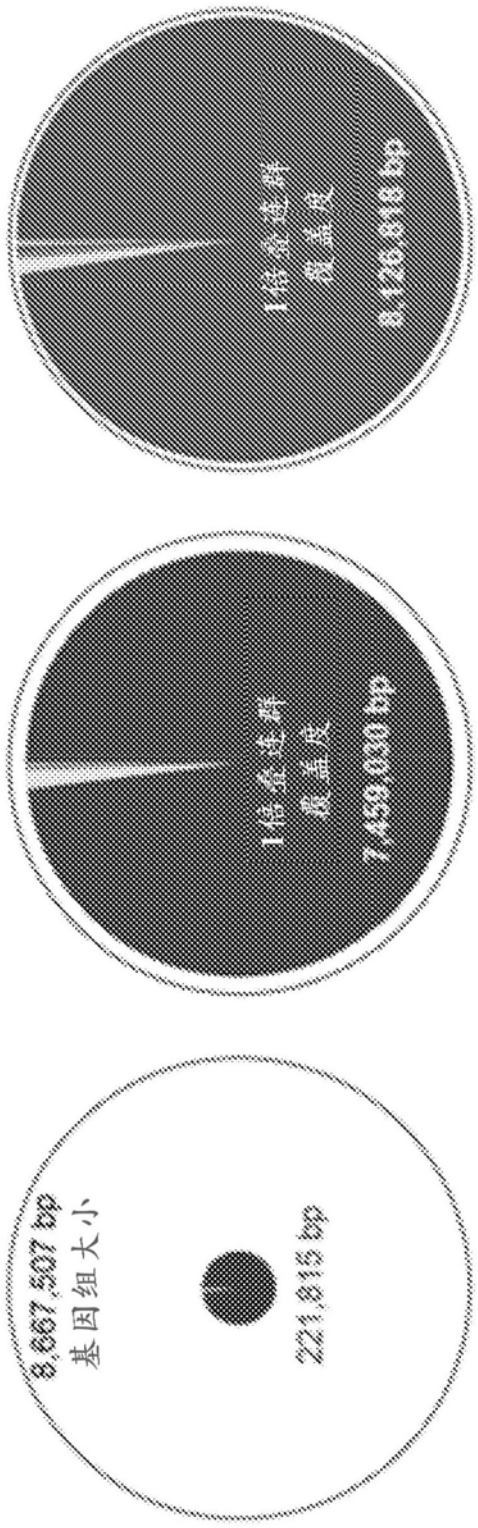

1倍重叠群
覆盖度
8,126,818 bp
1倍重叠群
覆盖度
7,459,030 bp
8,867,507 bp
基因组大小
221,815 bp

图20

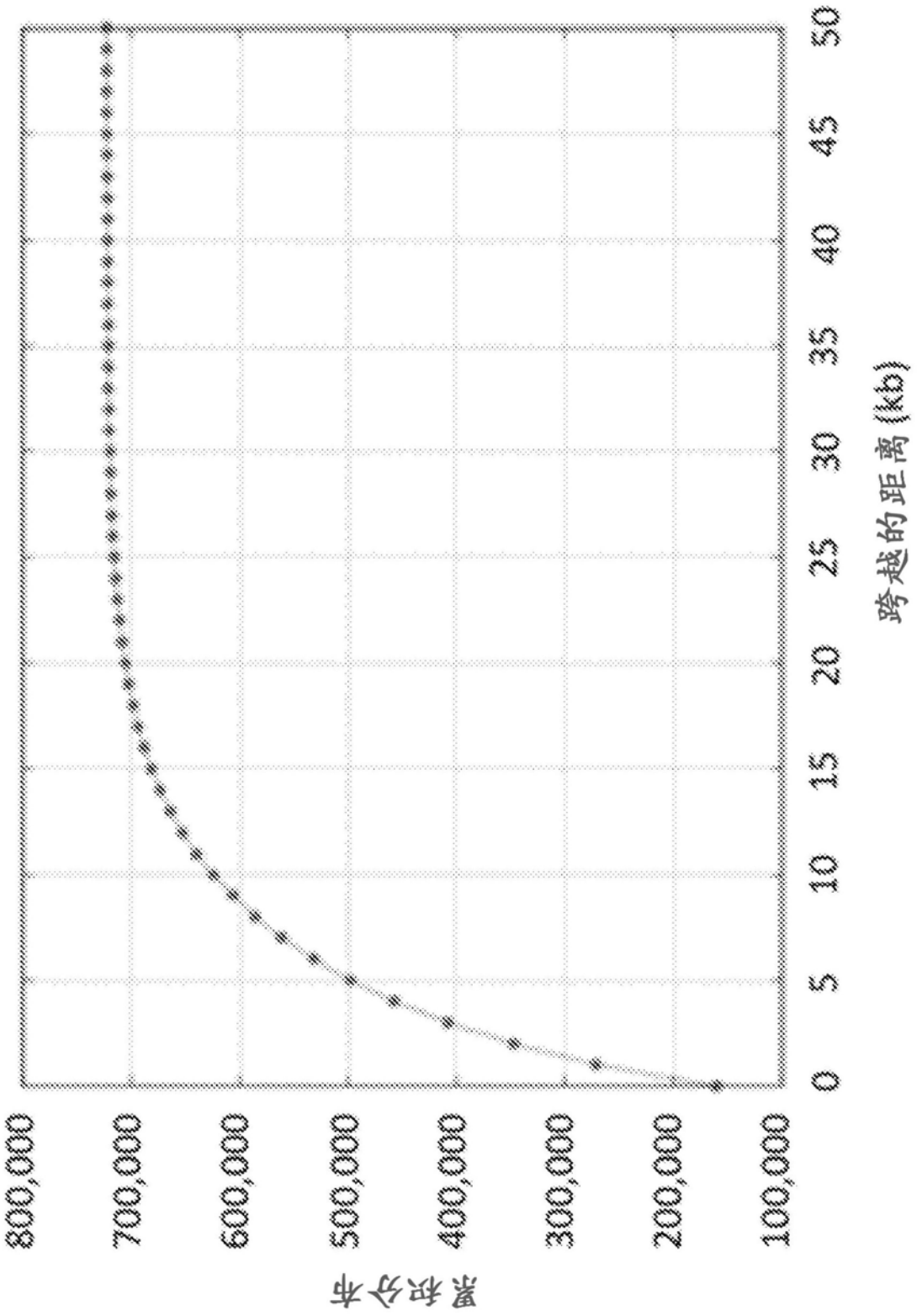

聚积分布
800,000
700,000
600,000
500,000
400,000
300,000
200,000
100,000
0
5
10
15
20
25
30
35
40
45
50
跨越的距离(kb)

图21

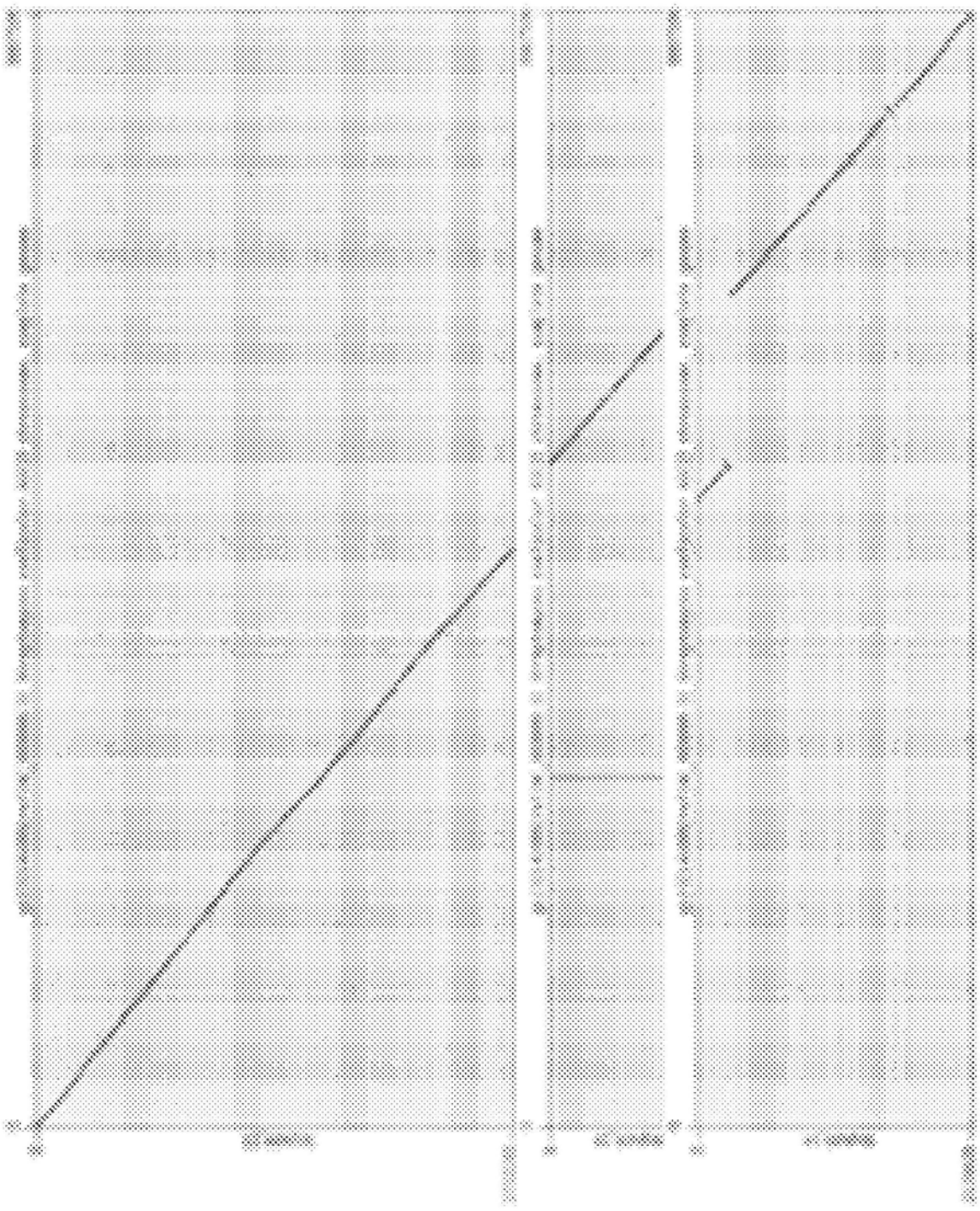


图22A

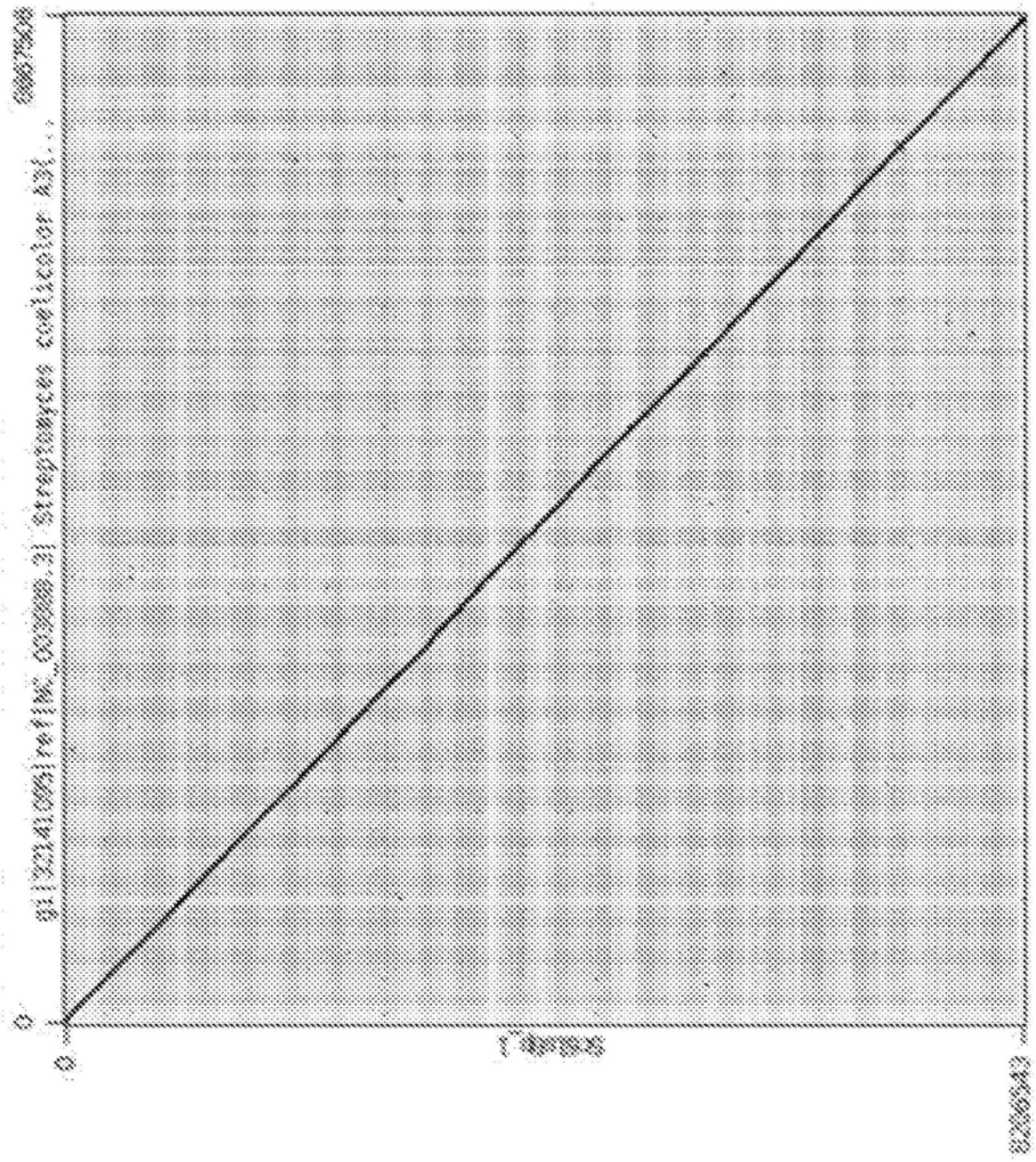


图22B

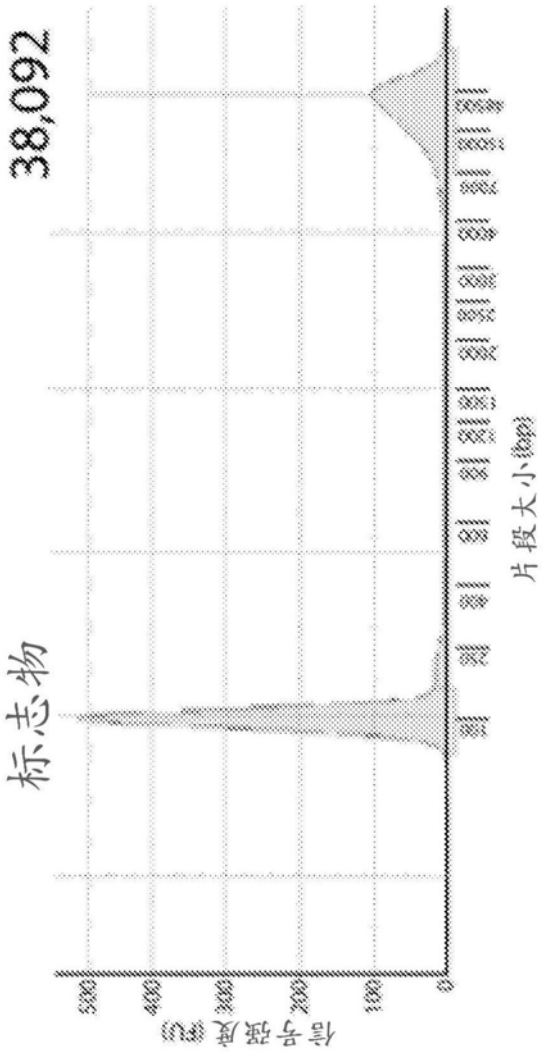


图23A

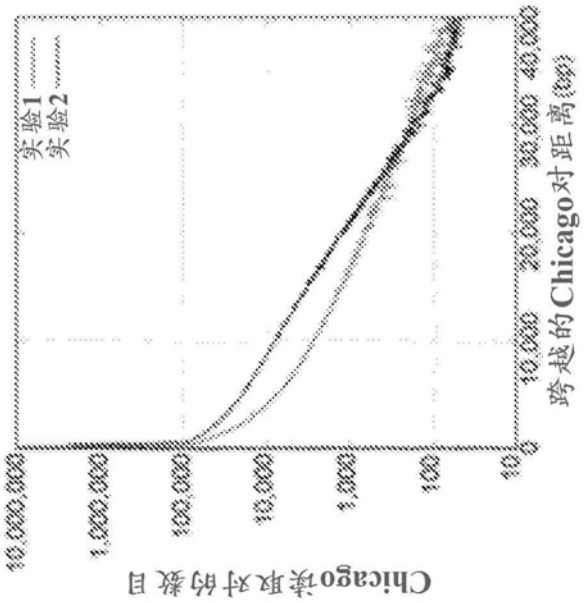

实验1
实验2
10,000,000
1,000,000
100,000
10,000
1,000
100
10
Chicago读取对的数目
10,000
20,000
30,000
40,000
跨越的Chicago对距离(m)

图23B

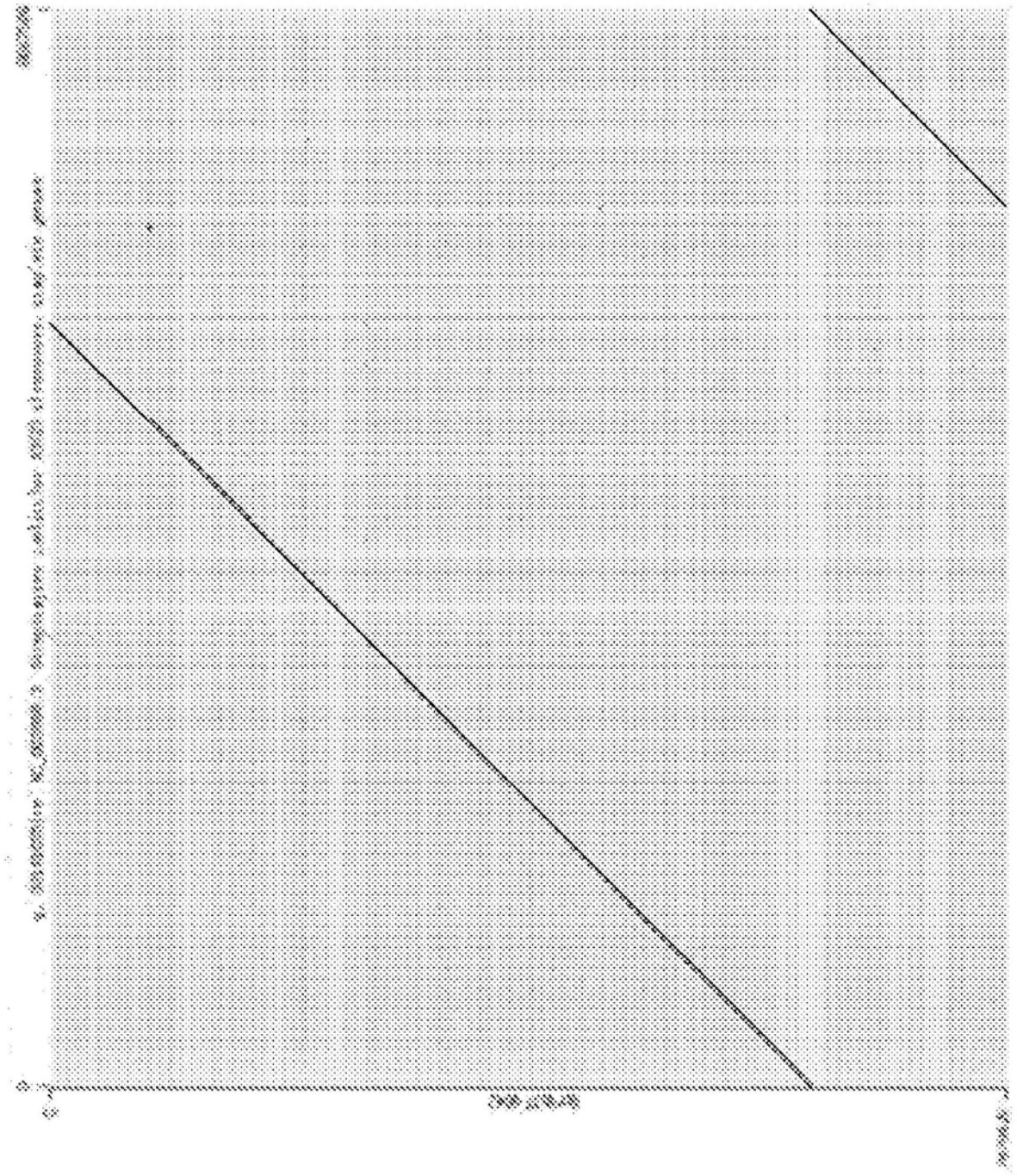


图24

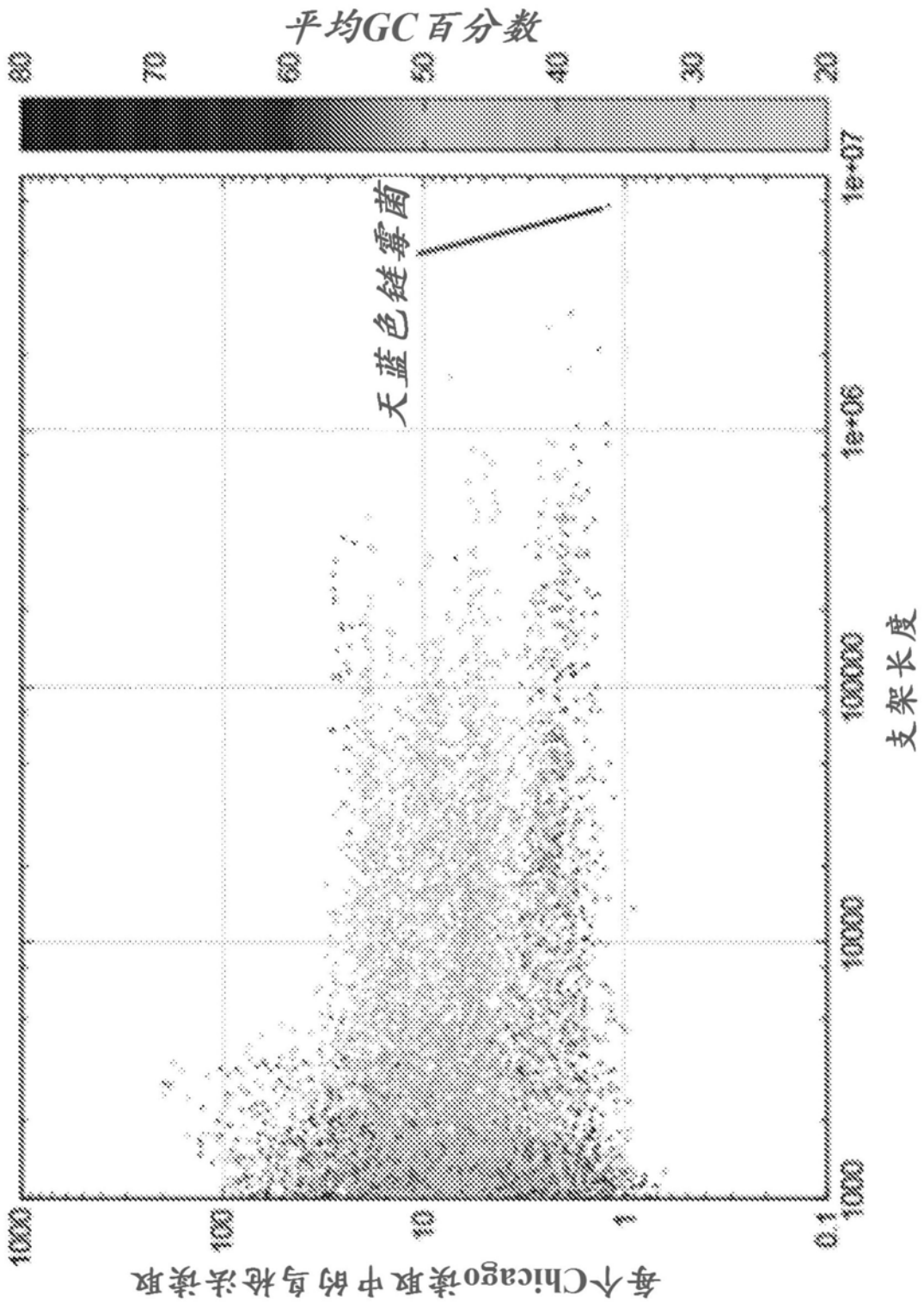

平均GC百分数
80
70
60
50
40
30
20
天蓝色链霉菌
1e+07
1e+06
100000
10000
1000
支架长度
每个Chicago读取中的鸟枪法读取
1000
100
10
1
0.1

图25

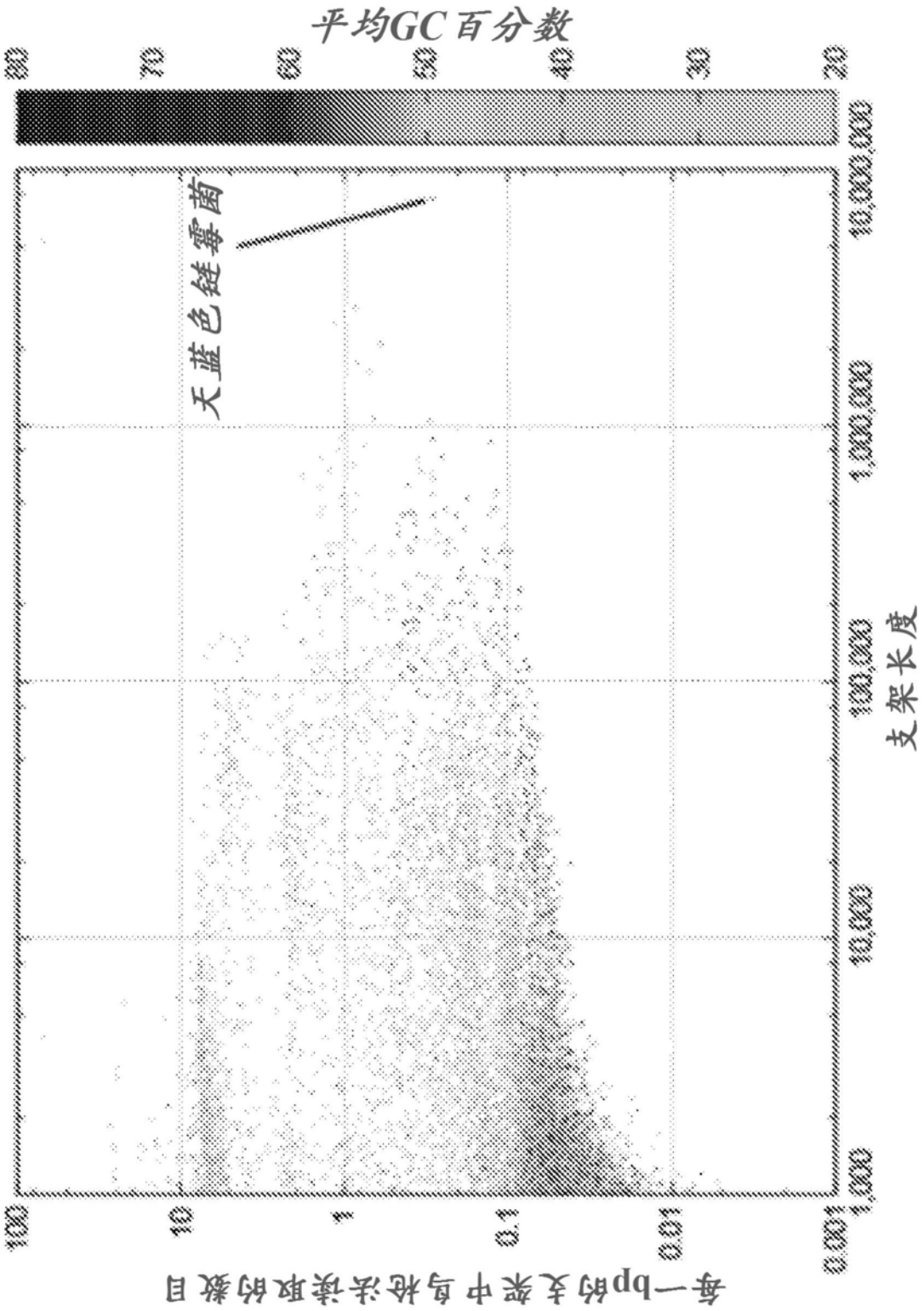

平均GC百分数
80
70
60
50
40
30
20
天蓝色链霉菌
10,000,000
1,000,000
100,000
10,000
1,000
支架长度
100
10
1
0.1
0.01
0.001
每一bp的支架中鸟枪法读取的数目

图26A

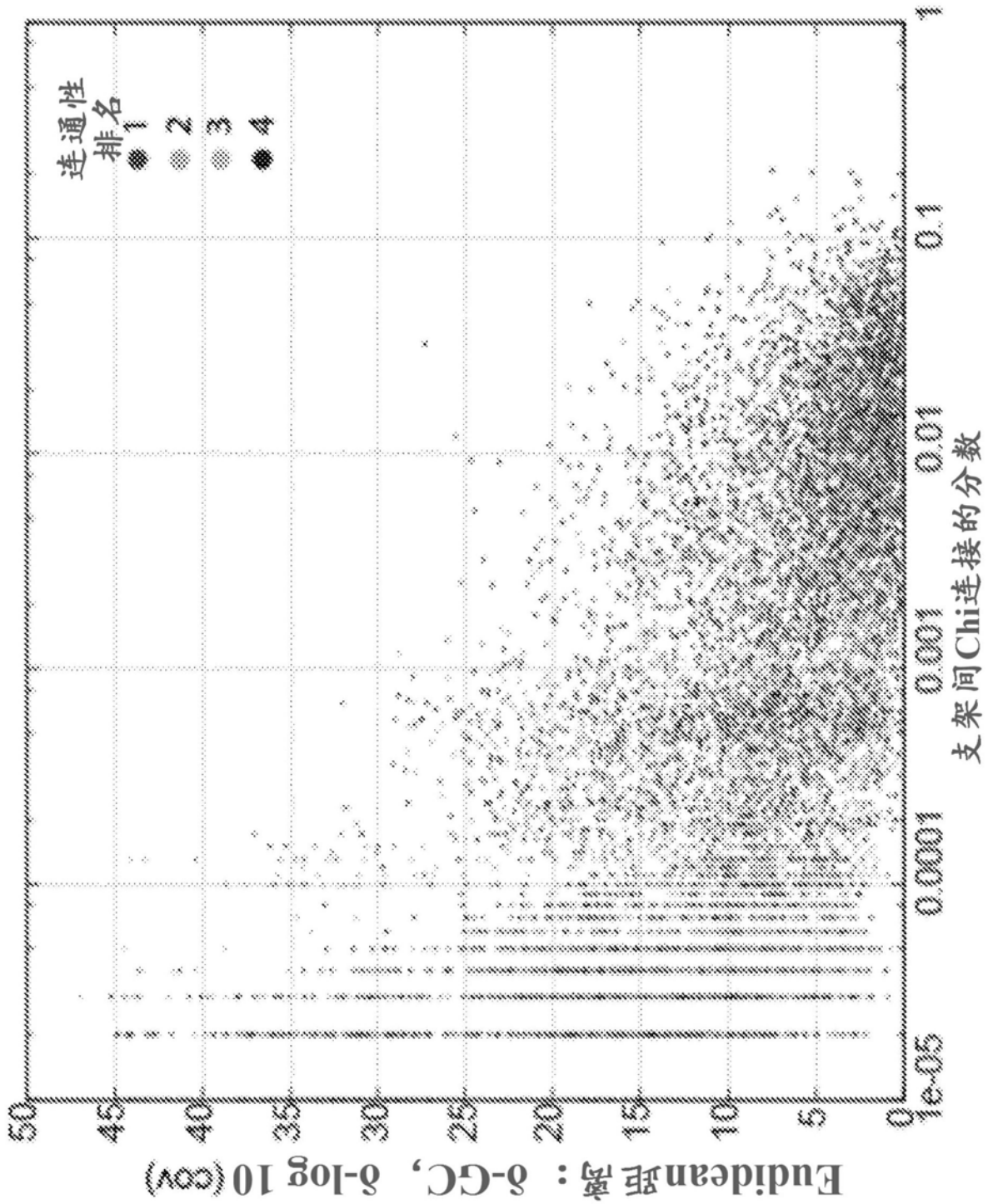

连通性
排名
1
2
3
4
Euclidean距离：δ-GC，δ-log 10(cov)
支架间Chi连接的分数
50
45
40
35
30
25
20
15
10
5
0
1e-05
0.0001
0.001
0.01
0.1
1

图26B

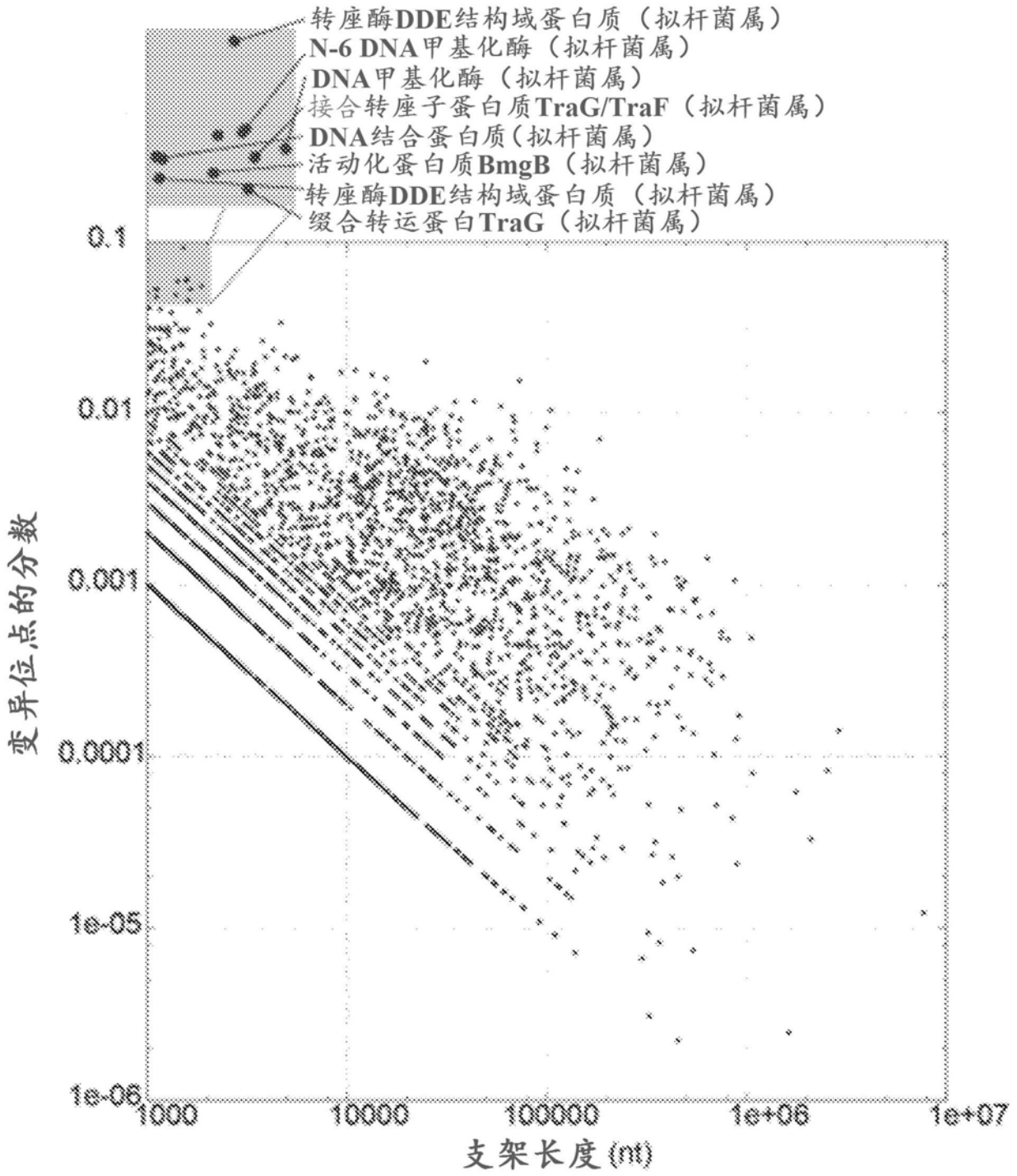

转座酶DDE结构域蛋白质（拟杆菌属）
N-6 DNA甲基化酶（拟杆菌属）
DNA甲基化酶（拟杆菌属）
接合转座子蛋白质TraG/TraF（拟杆菌属）
DNA结合蛋白质(拟杆菌属)
活动化蛋白质BmgB（拟杆菌属）
转座酶DDE结构域蛋白质（拟杆菌属）
缀合转运蛋白TraG（拟杆菌属）
0.1
0.01
0.001
0.0001
1e-05
1e-06
变异位点的分数
1000
10000
100000
1e+06
1e+07
支架长度(nt)

图27