(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 109072298 B
(45)授权公告日 2022.04.08

(21)申请号 201780025354.3

(22)申请日 2017.02.23

(65)同一申请的已公布的文献号
申请公布号 CN 109072298 A

(43)申请公布日 2018.12.21

(30)优先权数据
62/298,906 2016.02.23 US
62/298,966 2016.02.23 US
62/305,957 2016.03.09 US

(85)PCT国际申请进入国家阶段日
2018.10.23

(86)PCT国际申请的申请数据
PCT/US2017/019099 2017.02.23

(87)PCT国际申请的公布数据
WO2017/147279 EN 2017.08.31

(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 小理查德·E·格林
丹尼尔·S·罗赫萨尔
保罗·哈特利 马可·布兰切特

(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 贺淑东

(51)Int.Cl.
C12Q 1/6869(2018.01)
C40B 30/04(2006.01)
C40B 40/08(2006.01)

(56)对比文件
US 2015363550 A1,2015.12.17
WO 2015089243 A1,2015.06.18
WO 2016019360 A1,2016.02.04
Nicholas H. Putnam 等.Chromosome-scale shotgun assembly using an in vitro method for long-range linkage.《Genome Research》.2016,第26卷第342页摘要、第346-349页方法.
Jon-Matthew Belton 等.Hi-C: A comprehensive technique to capture the conformation of genomes.《Methods》.2012,第58卷第268页摘要、第267-272页方法.
Siddarth Selvaraj 等.Whole-genome haplotype reconstruction using proximity-ligation and shotgun sequencing.《nature biotechnology》.2013,第31卷(第12期),第1111-1118页.

审查员 贾星航

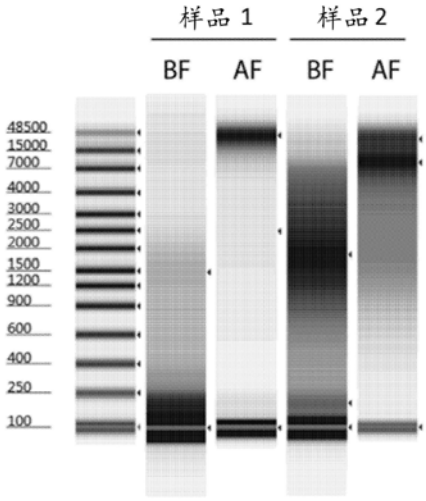
权利要求书2页 说明书50页 附图10页

(54)发明名称

用于基因组组装和单倍体型定相的定相读取集的生成

(57)摘要

本文公开了通过核酸分子的分段和重排以保留单个分子的相位信息或物理连锁信息的方式促进序列数据(如基因组序列数据)的准确定相的方法、组合物和系统。这通过独立于它们的磷酸二酯骨架结合分子、切割所述分子、连接以及通过长读取测序技术对所述分子进行测序以恢复跨越至少一个区段的区段序列信息来不同地实现。



CN 109072298 B

1.一种从第一DNA分子生成长距离相位信息的方法,该方法包括:

(a)提供具有第一区段、第二区段和第三区段的第一DNA分子,其中所述第一区段和所述第二区段在所述第一DNA分子上不相邻,并且其中所述第三区段与所述第二区段或所述第一区段在所述第一DNA分子上不相邻,使得所述第一区段、所述第二区段和所述第三区段独立于所述第一DNA分子的共同的磷酸二酯骨架而与DNA结合部分结合;

(b)切割所述第一DNA分子,使得所述第一区段、所述第二区段和所述第三区段不由共同的磷酸二酯骨架连接;

(c)通过磷酸二酯键将所述第一区段与所述第二区段附接并通过磷酸二酯键将所述第三区段与所述第二区段附接以形成重新组装的第一DNA分子;以及

(d)对单个测序读取中包含所述第一区段与所述第二区段之间的接头以及所述第二区段和所述第三区段之间的接头的所述重新组装的第一DNA分子的至少4kb的连续序列进行测序,

其中第一区段序列、第二区段序列和第三区段序列代表来自所述第一DNA分子的长距离相位信息。

2.根据权利要求1所述的方法,其中所述DNA结合部分包含多个DNA结合分子。

3.根据权利要求2所述的方法,其中使所述第一DNA分子与多个DNA结合分子接触包括与DNA结合蛋白的群体接触。

4.根据权利要求3所述的方法,其中所述DNA结合蛋白的群体包含核蛋白。

5.根据权利要求3所述的方法,其中所述DNA结合蛋白的群体包含核小体。

6.根据权利要求3所述的方法,其中所述DNA结合蛋白的群体包含组蛋白。

7.根据权利要求2所述的方法,其中使所述第一DNA分子与多个DNA结合部分接触包括与DNA结合纳米颗粒的群体接触。

8.根据权利要求1-7中任一项所述的方法,其包括使所述第一DNA分子与交联剂接触。

9.根据权利要求8所述的方法,其中所述交联剂为甲醛。

10.根据权利要求1所述的方法,其中所述DNA结合部分与包含多个DNA结合部分的表面结合。

11.根据权利要求1所述的方法,其中所述DNA结合部分与包含珠子的固体框架结合。

12.根据权利要求1所述的方法,其中切割所述第一DNA分子包括与限制性内切核酸酶接触。

13.根据权利要求1所述的方法,其中切割所述第一DNA分子包括与非特异性内切核酸酶接触。

14.根据权利要求1所述的方法,其中切割所述第一DNA分子包括与标签化酶接触。

15.根据权利要求1所述的方法,其中切割所述第一DNA分子包括与转座酶接触。

16.根据权利要求1所述的方法,其中切割所述第一DNA分子包括剪切所述第一DNA分子。

17.根据权利要求1所述的方法,其包括将标签添加至所述第一区段、所述第二区段或者所述第三区段的至少一个暴露的末端。

18.根据权利要求17所述的方法,其中所述标签包含标记的碱基。

19.根据权利要求17所述的方法,其中所述标签包含甲基化的碱基。

20.根据权利要求17所述的方法，其中所述标签包含生物素化的碱基。

21.根据权利要求17所述的方法，其中所述标签包含尿苷。

22.根据权利要求17所述的方法，其中所述标签包含非规范碱基。

23.根据权利要求17所述的方法，其中所述标签生成平端的暴露的末端。

24.根据权利要求1所述的方法，其包括将至少一个碱基添加至第一区段粘性末端的嵌入链。

25.根据权利要求1所述的方法，其中所述第一区段包含第一区段粘性末端，并且所述方法还包括添加包含与所述第一区段粘性末端退火的突出端的连接体寡核苷酸。

26.根据权利要求25所述的方法，其中所述连接体寡核苷酸包含与所述第一区段粘性末端退火的突出端和与第二区段粘性末端退火的突出端。

27.根据权利要求25所述的方法，其中所述连接体寡核苷酸不包含两个5’磷酸部分。

28.根据权利要求1所述的方法，其中附接包括连接。

29.根据权利要求1所述的方法，其中附接包括DNA单链切口修复。

30.根据权利要求1所述的方法，其中在切割所述第一DNA分子之前，所述第一区段和所述第二区段在所述第一DNA分子上相隔至少10kb。

31.根据权利要求1所述的方法，其中在切割所述第一DNA分子之前，所述第一区段和所述第二区段在所述第一DNA分子上相隔至少15kb。

32.根据权利要求1所述的方法，其中在切割所述第一DNA分子之前，所述第一区段和所述第二区段在所述第一DNA分子上相隔至少30kb。

33.根据权利要求1所述的方法，其中在切割所述第一DNA分子之前，所述第一区段和所述第二区段在所述第一DNA分子上相隔至少50kb。

34.根据权利要求1所述的方法，其中在切割所述第一DNA分子之前，所述第一区段和所述第二区段在所述第一DNA分子上相隔至少100kb。

35.根据权利要求1所述的方法，其中所述测序包括单分子长读取测序。

36.根据权利要求35所述的方法，其中所述长读取测序包括至少5kb的读取。

37.根据权利要求35所述的方法，其中所述长读取测序包括至少10kb的读取。

38.根据权利要求1所述的方法，其中所述第一重新组装的DNA分子包含在所述第一DNA分子的一端连接5’端与3’端的发夹部分。

39.根据权利要求1所述的方法，其包括对所述第一DNA分子的第二重新组装形式进行测序。

40.根据权利要求1所述的方法，其中所述第一区段和所述第二区段各自为至少500bp。

41.根据权利要求2所述的方法，其中所述第一区段、所述第二区段和所述第三区段各自为至少500bp。

用于基因组组装和单倍体型定相的定相读取集的生成

[0001] 交叉引用

[0002] 本申请要求于2016年2月23日提交的美国临时申请号62/298,906的权益，该临时申请通过引用以其全文明确地并入于此，并且本申请还要求于2016年2月23日提交的美国临时申请号62/298,966的权益，该临时申请通过引用以其全文明确地并入于此，并且本申请还要求于2016年3月9日提交的美国临时申请号62/305,957的权益，该临时申请通过引用以其全文明确地并入于此。

背景技术

[0003] 在理论上和实践中，仍然难以确定复杂DNA样品(如具有二倍体或多倍体基因组的DNA样品，或包含大量重复序列或相同序列的DNA样品)的单倍体型相位信息。由于感兴趣的基因座被高度重复的区域或长段的相同序列隔开而导致困难，使得读取信息的标准组装不足以将相位信息分配给基因座处的等位基因。

发明内容

[0004] 本文公开了与通过分段重排的核酸分子(如染色体)的生成和测序(如长读取测序)对核酸序列数据准确定相有关的方法、组合物和系统。

[0005] 本文公开了从第一DNA分子生成长距离相位信息的方法，该方法包括：a)提供具有第一区段和第二区段的第一DNA分子，其中所述第一区段和所述第二区段在所述第一DNA分子上不相邻；b)使所述第一DNA分子与DNA结合部分接触，使得所述第一区段和所述第二区段独立于所述第一DNA分子的共同的磷酸二酯骨架而与所述DNA结合部分结合；c)切割所述第一DNA分子，使得所述第一区段和所述第二区段不由共同的磷酸二酯骨架连接；d)通过磷酸二酯键将所述第一区段与所述第二区段附接以形成重新组装的第一DNA分子；以及e)对单个测序读取中包含所述第一区段与所述第二区段之间的接头的所述重新组装的第一DNA分子的至少4kb的连续序列进行测序，其中第一区段序列和第二区段序列代表来自第一DNA分子的长距离相位信息。在一些方面，所述DNA结合部分包含多个DNA结合分子，如DNA结合蛋白。在一些方面，所述DNA结合蛋白的群体广泛地包含核蛋白、核小体，或在一些情况下，更具体地包含组蛋白。在一些方面，使所述第一DNA分子与多个DNA结合部分接触包括与DNA结合纳米颗粒的群体接触。通常，所述第一DNA分子具有在所述第一DNA分子上与所述第一区段或所述第二区段不相邻的第三区段，其中进行(b)中的所述接触，使得所述第三区段独立于所述第一DNA分子的共同的磷酸二酯骨架而与所述DNA结合部分结合，其中进行(c)中的所述切割，使得所述第三区段不通过共同的磷酸二酯骨架与所述第一区段和所述第二区段连接，其中所述附接包括通过磷酸二酯键将所述第三区段与所述第二区段附接以形成所述重新组装的第一DNA分子，并且其中(e)中测序的连续序列包含单个测序读取中所述第二区段与所述第三区段之间的接头。该方法通常包括使所述第一DNA分子与交联剂如甲醛接触。在一些方面，所述DNA结合部分与包含多个DNA结合部分的表面结合。在一些方面，所述DNA结合部分与包含珠子的固体框架结合。在一些方面，切割所述第一DNA分子包括与限制

性内切核酸酶如非特异性内切核酸酶、标签化酶(tagmentation enzyme)或转座酶接触。在一些方面,切割所述第一DNA分子包括剪切所述第一分子。任选地,该方法包括将标签添加至至少一个暴露的末端。示例性的标签包含标记的碱基、甲基化的碱基、生物素化的碱基、尿苷或任何其他非规范碱基。在一些方面,所述标签生成平端的暴露的末端。在一些方面,该方法包括将至少一个碱基添加至第一区段粘性末端的嵌入链。在一些方面,该方法包括添加包含与所述第一区段粘性末端退火的突出端的连接体寡核苷酸。在一些方面,所述连接体寡核苷酸包含与所述第一区段粘性末端退火的突出端和与第二区段粘性末端退火的突出端。在一些方面,所述连接体寡核苷酸不包含两个5'磷酸部分。在一些方面,附接包括连接。在一些方面,附接包括DNA单链切口修复。在一些方面,在切割所述第一DNA分子之前,所述第一区段和所述第二区段在所述第一DNA分子上相隔至少10kb。在一些方面,在切割所述第一DNA分子之前,所述第一区段和所述第二区段在所述第一DNA分子上相隔至少15kb。在一些方面,在切割所述第一DNA分子之前,所述第一区段和所述第二区段在所述第一DNA分子上相隔至少30kb。在一些方面,在切割所述第一DNA分子之前,所述第一区段和所述第二区段在所述第一DNA分子上相隔至少50kb。在一些方面,在切割所述第一DNA分子之前,所述第一区段和所述第二区段在所述第一DNA分子上相隔至少100kb。在一些方面,所述测序包括单分子长读取测序。在一些方面,所述长读取测序包括至少5kb的读取。在一些方面,所述长读取测序包括至少10kb的读取。在一些方面,所述第一重新组装的DNA分子包含在所述第一DNA分子的一端连接5'端与3'端的发夹部分。在一些方面,该方法包括对所述第一DNA分子的第二重新组装形式进行测序。在一些方面,所述第一区段和所述第二区段各自为至少500bp。在一些方面,所述第一区段、所述第二区段和所述第三区段各自为至少500bp。

[0006] 本文公开了基因组组装的方法,该方法包括:a)获得与结构复合的第一DNA分子;b)切割所述第一DNA分子以形成第一暴露的末端和第二暴露的末端,其中在所述切割之前,所述第一暴露的末端和所述第二暴露的末端在所述分子上不相邻;c)切割所述第一DNA分子以形成第三暴露的末端和第四暴露的末端,其中在所述切割之前,所述第三暴露的末端和所述第四暴露的末端在所述分子上不相邻;d)将所述第一暴露的末端和所述第二暴露的末端附接以形成第一接头;e)将所述第三暴露的末端和所述第四暴露的末端附接以形成第二接头;f)在单个测序读取中跨越所述第一接头和所述第二接头进行测序;g)将所述第一接头的第一侧上的序列映射至所述多个叠连群的第一叠连群;h)将所述第一接头的第二侧上的序列映射至所述多个叠连群的第二叠连群;i)将所述第二接头的第一侧上的序列映射至所述多个叠连群的第二叠连群;j)将所述第二接头的第二侧上的序列映射至所述多个叠连群的第三叠连群;以及k)将所述第一叠连群、所述第二叠连群和所述第三叠连群分配给基因组组装的共同相位。在一些方面,所述多个叠连群由鸟枪序列数据生成。在一些方面,所述多个叠连群由单分子长读取数据生成。在一些方面,所述单分子长读取数据包含所述多个叠连群。在一些方面,所述多个叠连群通过在所述第一接头和所述第二接头进行测序而同时获得。在一些方面,在所述标志物寡核苷酸进行测序包括对至少10kb进行测序。在一些方面,所述结构包含与所述第一DNA分子结合以形成重构的染色质的DNA结合部分的群体。在一些方面,使所述重构的染色质与交联剂接触。在一些方面,所述交联剂包含甲醛。在一些方面,所述DNA结合部分的群体包含组蛋白。在一些方面,所述DNA结合部分的群体包含纳米颗粒。在一些方面,所述结构包含天然染色质。在一些方面,在切割所述第一DNA分子之

前，所述第一暴露的末端和所述第二暴露的末端在所述第一DNA分子上相隔至少10kb。在一些方面，在切割所述第一DNA分子之前，所述第一暴露的末端和所述第二暴露的末端在所述第一DNA分子上相隔至少15kb。在一些方面，在切割所述第一DNA分子之前，所述第一暴露的末端和所述第二暴露的末端在所述第一DNA分子上相隔至少30kb。在一些方面，在切割所述第一DNA分子之前，所述第一暴露的末端和所述第二暴露的末端在所述第一DNA分子上相隔至少50kb。在一些方面，在切割所述第一DNA分子之前，所述第一暴露的末端和所述第二暴露的末端在所述第一DNA分子上相隔至少100kb。在一些方面，该方法包括对所述第一DNA分子的第二拷贝进行测序。

[0007] 本文公开了至少5kb的重排核酸分子，其包含：a)第一区段；b)第二区段；以及c)第三区段；所述第一区段和所述第二区段在第一接头处连接；并且所述第二区段和所述第三区段在第二接头处连接；其中所述第一区段、所述第二区段和所述第三区段在未重排核酸分子中以相隔至少10kb的相位存在，并且其中至少70％的所述重排核酸分子映射至所述共同的未重排核酸分子。在一些方面，所述第一区段、所述第二区段和所述第三区段包含来自基因组的共同核酸分子的单独的基因组核酸序列。在一些方面，所述第一区段、所述第二区段和所述第三区段以重排核酸中重排的顺序存在于基因组中的共同分子中。在一些方面，所述核酸分子的长度为至少30kb。在一些方面，所述核酸在双链末端包含发夹环，使得所述分子包含含有30kb反向重复的单链。在一些方面，所述核酸为双链环状分子。在一些方面，至少80％的所述重排核酸分子映射至所述共同的未重排核酸分子。在一些方面，至少85％的所述重排核酸分子映射至所述共同的未重排核酸分子。在一些方面，至少90％的所述重排核酸分子映射至所述共同的未重排核酸分子。在一些方面，至少95％的所述重排核酸分子映射至所述共同的未重排核酸分子。在一些方面，至少99％的所述重排核酸分子映射至所述共同的未重排核酸分子。在一些方面，所述重排核酸分子的至少80％的区段映射至所述共同的未重排核酸分子。在一些方面，所述重排核酸分子的至少85％的区段映射至所述共同的未重排核酸分子。在一些方面，所述重排核酸分子的至少90％的区段映射至所述共同的未重排核酸分子。在一些方面，所述重排核酸分子的至少95％的区段映射至所述共同的未重排核酸分子。在一些方面，所述重排核酸分子的至少99％的区段映射至所述共同的未重排核酸分子。在一些方面，通过本文公开的任何方法的步骤生成所述重排核酸。

[0008] 本文公开了生成样品核酸分子的定相序列的方法，该方法包括：a)从所述样品核酸分子生成如本文公开的第一重排核酸分子；b)从所述样品核酸分子生成如本文公开的第二重排核酸分子；以及c)对所述第一重排核酸分子和所述第二重排核酸分子进行测序；其中所述第一重排核酸分子和所述第二重排核酸分子是独立生成的。

[0009] 本文公开了生成样品核酸分子的定相序列的方法，该方法包括：a)对来自所述样品核酸分子的如本文公开的第一重排核酸分子进行测序；b)对来自所述样品核酸分子的如本文公开的第二重排核酸分子进行测序；其中所述第一重排核酸分子和所述第二重排核酸分子是独立生成的；以及c)组装如本文公开的所述第一重排核酸分子和如本文公开的所述第二重排核酸分子的序列，使得组装的序列为样品核酸分子的未重排的定相序列。在一些方面，对第一重排核酸分子进行测序包括生成至少1kb的序列读取。在一些方面，对第一重排核酸分子进行测序包括生成至少2kb的序列读取。在一些方面，对第一重排核酸分子进行测序包括生成至少5kb的序列读取。在一些方面，该方法包括将至少70％的所述第一重排分

子分配给单个基因组分子的共同相位。在一些方面，该方法包括将至少70％的所述第二重排分子分配给单个基因组分子的共同相位。在一些方面，该方法包括将至少80％的所述第一重排分子分配给单个基因组分子的共同相位。在一些方面，该方法包括将至少80％的所述第二重排分子分配给单个基因组分子的共同相位。在一些方面，该方法包括将至少90％的所述第一重排分子分配给单个基因组分子的共同相位。在一些方面，该方法包括将至少90％的所述第二重排分子分配给单个基因组分子的共同相位。在一些方面，该方法包括将至少95％的所述第一重排分子分配给单个基因组分子的共同相位。在一些方面，该方法包括将至少95％的所述第二重排分子分配给单个基因组分子的共同相位。

[0010] 本文公开了对长读取序列数据进行定相的方法，该方法包括：a)从本文公开的任何核酸样品中获得序列数据；b)从如本文公开的任何重排核酸中获得长读取序列数据；c)将来自所述重排核酸的长读取序列数据映射至来自所述核酸样品的序列数据；以及d)将映射至来自所述重排核酸的长读取序列数据的来自所述核酸样品的序列数据分配给共同相位。

[0011] 本文公开了通过DNA测序技术向由核酸样品生成的核酸数据集提供相位信息的方法，该方法包括：a)获得具有相隔大于所述DNA测序技术的读取长度的距离的第一区段和第二区段的所述核酸样品的核酸；b)对所述核酸进行改组，使得所述第一区段和所述第二区段相隔小于所述DNA测序技术的读取长度的距离；c)使用所述DNA测序技术对所述改组的核酸进行测序，使得所述第一区段和所述第二区段出现在所述DNA测序技术的单个读取中；以及d)将包含第一区段序列的数据集的序列读取和包含第二区段序列的数据集的序列读取分配给共同相位。在一些方面，所述DNA测序技术生成具有至少10kb的读取长度的读取。在一些方面，改组包括进行本文公开的任何方法的步骤。在一些方面，所述第一区段和所述第二区段由标记区段末端的连接体寡核苷酸隔开。

[0012] 本文公开了包含从如本文公开的多个分子获得的序列信息的核酸序列数据库，其中从至少一个分析中排除对应于少于70％的区段映射至共同的支架的分子的序列信息。

[0013] 本文公开了包含从如本文公开的多个分子获得的序列信息的核酸序列数据库，其中从至少一个分析中排除对应于少于70％的序列信息映射至共同的支架的分子的序列信息。

[0014] 本文公开了对长读取序列数据进行定相的方法，该方法包括：a)从本文公开的任何核酸样品中获得序列数据；b)从本文公开的任何重排核酸的重排核酸中获得长读取序列数据；c)将所述重排核酸的第一区段、第二区段和第三区段映射至来自所述核酸样品的序列数据以获得核酸样品序列数据；以及d)当至少两个区段映射至共同的支架时，将所述区段的序列变异分配给共同的相位。在一些方面，所述第一区段包含相对于来自所述核酸样品的序列数据的单核苷酸多态性。在一些方面，所述第一区段包含相对于来自所述核酸样品的序列数据的插入。在一些方面，所述第一区段包含相对于来自所述核酸样品的序列数据的缺失。在一些方面，该方法包括将映射至第一共同的支架的第一组区段分配给所述第一共同的支架的共同相位，以及将映射至第二共同的支架的第二组区段分配给所述第二共同的支架的共同相位。

[0015] 本文公开了核酸样品的核酸序列文库，所述核酸序列文库包含具有平均读取长度的核酸序列读取的群体，至少一个所述读取包含第一核酸区段的至少500个碱基和第二核

酸区段的至少500个碱基，其中发现所述第一核酸区段和所述第二核酸区段在所述核酸样品的共同分子上同相位相隔大于所述平均读取长度的距离。在一些方面，发现所述第一核酸区段和所述第二核酸区段同相位相隔大于10kb的距离。在一些方面，发现所述第一核酸区段和所述第二核酸区段同相位相隔大于20kb的距离。在一些方面，发现所述第一核酸区段和所述第二核酸区段同相位相隔大于50kb的距离。在一些方面，发现所述第一核酸区段和所述第二核酸区段同相位相隔大于100kb的距离。在一些方面，至少一个所述读取包含至少1kb的第一核酸区段。在一些方面，至少一个所述读取包含至少5kb的第一核酸区段。在一些方面，至少一个所述读取包含至少10kb的第一核酸区段。在一些方面，至少一个所述读取包含至少20kb的第一核酸区段。在一些方面，至少一个所述读取包含至少50kb的第一核酸区段。在一些方面，核酸序列文库包含至少80％的所述核酸样品。在一些方面，核酸序列文库包含至少85％的所述核酸样品。在一些方面，核酸序列文库包含至少90％的所述核酸样品。在一些方面，核酸序列文库包含至少95％的所述核酸样品。在一些方面，核酸序列文库包含至少99％的所述核酸样品。在一些方面，核酸序列文库包含至少99.9％的所述核酸样品。

[0016] 本文公开了核酸样品的核酸序列文库，所述核酸序列文库包含平均长度为至少1kb的核酸序列读取的群体，所述读取独立地包含来自所述核酸样品的两个单独的同相位区域的至少300个碱基的序列，所述两个单独的同相位区域在所述核酸样品中相隔大于10kb的距离。在一些方面，所述读取独立地包含来自所述核酸样品的两个单独的同相位区域的至少500个碱基的序列。在一些方面，所述读取独立地包含来自所述核酸样品的两个单独的同相位区域的至少1kb的序列。在一些方面，所述读取独立地包含来自所述核酸样品的两个单独的同相位区域的至少2kb的序列。在一些方面，所述读取独立地包含来自所述核酸样品的两个单独的同相位区域的至少5kb的序列。在一些方面，所述读取独立地包含来自所述核酸样品的两个单独的同相位区域的至少10kb的序列。在一些方面，所述两个单独的同相位区域在所述核酸样品中相隔大于20kb的距离。在一些方面，所述两个单独的同相位区域在所述核酸样品中相隔大于30kb的距离。在一些方面，在至少1％的读取中所述两个单独的同相位区域在所述核酸样品中相隔大于50kb的距离。在一些方面，在至少1％的读取中所述两个单独的同相位区域在所述核酸样品中相隔大于100kb的距离。在一些方面，核酸序列文库包含至少80％的所述核酸样品。在一些方面，核酸序列文库包含至少85％的所述核酸样品。在一些方面，核酸序列文库包含至少90％的所述核酸样品。在一些方面，核酸序列文库包含至少95％的所述核酸样品。在一些方面，核酸序列文库包含至少99％的所述核酸样品。在一些方面，核酸序列文库包含至少99.9％的所述核酸样品。

[0017] 本文公开了由核酸样品生成的核酸文库，其中所述核酸样品中的至少80％的核酸序列在所述核酸文库中表示；并且所述核酸样品的同相位序列区段进行重排，使得在单个序列读取中读取所述核酸样品的至少一个远端定位的同相位区段对；使得对所述文库的测序同时生成跨越至少80％的核酸样品的叠连群信息，以及足以对所述叠连群信息进行排序和定向以生成所述核酸样品的定相序列的相位信息。在一些方面，所述核酸样品的至少90％的核酸序列在所述核酸文库中表示。在一些方面，所述核酸样品的至少95％的核酸序列在所述核酸文库中表示。在一些方面，所述核酸样品的至少99％的核酸序列在所述核酸文库中表示。在一些方面，所述核酸样品的所述80％的核酸序列获自不多于100,000个文库

组分。在一些方面，所述核酸样品的所述80％的核酸序列获自不多于10,000个文库组分。在一些方面，所述核酸样品的所述80％的核酸序列获自不多于1,000个文库组分。在一些方面，所述核酸样品的所述80％的核酸序列获自不多于500个文库组分。在一些方面，所述样品为基因组样品。在一些方面，所述样品为真核基因组样品。在一些方面，所述样品为植物基因组样品。在一些方面，所述样品为动物基因组样品。在一些方面，所述样品为哺乳动物基因组样品。在一些方面，所述样品为单细胞真核基因组样品。在一些方面，所述样品为人基因组样品。在一些方面，所述核酸文库未进行条码化以保留相位信息。在一些方面，所述文库的读取包含来自第一区域的至少1kb的序列和来自第二区域的至少100个碱基的序列，该第二区域与所述第一区域同相位并且在样品中与第一区域相隔大于50kb。

[0018] 本文公开了配置核酸分子以用于在测序装置上测序的方法，其中所述核酸分子包含至少100kb的序列，并且其中所述至少100kb的序列包含相隔大于所述测序装置的读取长度的长度的第一区段和第二区段，所述方法包括改变所述核酸分子的第一区段相对于第二区段的相对位置，使得所述第一区段和所述第二区段相隔小于所述测序装置的读取长度；其中所述第一区段和所述第二区段的相位信息得以保持；并且其中不多于10％的核酸分子缺失。在一些方面，该方法包括生成跨越所述第一区段和所述第二区段的至少一部分的读取。在一些方面，该方法包括将所述第一区段和所述第二区段分配给所述核酸分子的序列的共同相位。在一些方面，不多于5％的核酸分子缺失。在一些方面，不多于1％的核酸分子缺失。在一些方面，在配置之前，所述第一区段和所述第二区段在核酸分子中相隔至少10kb。在一些方面，在配置之前，所述第一区段和所述第二区段在核酸分子中相隔至少50kb。在一些方面，在所述配置之后，所述第一区段和所述第二区段由接头标志物隔开。在一些方面，该方法包括将茎环附接在所述核酸的末端，从而将所述分子转化为单链。在一些方面，该方法包括将所述核酸分子环化。在一些方面，该方法包括将所述核酸分子附接至DNA聚合酶。在一些方面，该方法包括结合所述核酸分子，使得所述第一区段和所述第二区段独立于磷酸二酯骨架而保持在一起；在至少两个位置处切割所述第一区段与所述第二区段之间的磷酸二酯骨架；以及将所述第一区段与所述第二区段重新附接，使得所述第一区段和所述第二区段相隔小于所述测序装置的读取长度。在一些方面，所述切割和所述重新附接不会导致来自所述核酸分子的序列信息丢失。

[0019] 本文公开了从第一核酸分子生成长距离相位信息的方法，该方法包括：a)提供包含具有第一区段、第二区段和第三区段的第一核酸分子的样品，其中所述第一区段、所述第二区段和所述第三区段在所述第一核酸分子上均不相邻，其中使所述第一核酸分子与框架接触，使得所述第一区段、所述第二区段和所述第三区段独立于所述第一核酸分子的共同的磷酸二酯骨架而与所述框架结合；b)切割所述第一核酸分子，使得所述第一区段、所述第二区段和所述第三区段不由共同的磷酸二酯骨架连接；c)将所述第一区段与所述第二区段连接，并将所述第二区段与所述第三区段连接；以及d)对包含所述第一区段、所述第二区段和所述第三区段的所述第一核酸分子的第一部分进行测序，从而生成第一区段序列信息、第二区段序列信息和第三区段序列信息，其中所述第一区段序列信息、所述第二区段序列信息和所述第三区段序列信息提供关于所述第一核酸分子的长距离相位信息。在一些方面，所述框架包含重构的染色质。在一些方面，所述框架包含天然染色质。在一些方面，用限制酶进行所述切割。在一些方面，用片段化酶进行所述切割。在一些方面，该方法包括，在所

述测序之前，从样品中去除包含至多两个区段的第一核酸分子的第二部分。在一些方面，该方法包括使用所述第一区段序列信息、所述第二区段序列信息和所述第三区段序列信息组装所述第一核酸分子的序列。

[0020] 本文公开了对核酸分子进行测序的方法，该方法包括：a)获得包含共有共同的磷酸二酯骨架的第一区段、第二区段和第三区段的第一核酸分子，其中所述第一区段、第二区段和第三区段在所述第一核酸分子上均不相邻；b)对所述核酸分子进行分区，使得所述第一区段、第二区段和第三区段独立于它们共同的磷酸二酯骨架而相关联；c)切割所述核酸分子以生成片段，使得不存在连接所述第一区段、第二区段和第三区段的连续磷酸二酯骨架；d)连接所述片段，使得所述第一区段、第二区段和第三区段在共有共同的磷酸二酯骨架的重排核酸分子上是连续的；以及e)对所述重排核酸分子的至少一部分进行测序，使得在单个读取中对所述重排核酸分子的至少5,000个碱基进行测序。在一些方面，分区包括使所述核酸分子与结合部分接触，使得所述第一区段、第二区段和第三区段独立于它们共同的磷酸二酯骨架而结合在共同的复合体中。在一些方面，使所述核酸分子与多个DNA结合分子接触包括与DNA结合蛋白的群体接触。在一些方面，所述DNA结合蛋白的群体包含核蛋白。在一些方面，所述DNA结合蛋白的群体包含核小体。在一些方面，所述DNA结合蛋白的群体包含组蛋白。在一些方面，使所述核酸分子与多个DNA结合部分接触包括与DNA结合纳米颗粒的群体接触。在一些方面，切割所述核酸分子包括与限制性内切核酸酶接触。在一些方面，切割所述核酸分子包括与非特异性内切核酸酶接触。在一些方面，切割所述核酸分子包括与标签化酶接触。在一些方面，切割所述核酸分子包括与转座酶接触。在一些方面，切割所述核酸分子包括剪切所述第一分子。在一些方面，分区包括将所述核酸分子与样品的其他核酸分子分离。在一些方面，分区包括稀释所述核酸样品。在一些方面，分区包括将所述核酸分子分配至乳剂的微滴中。

[0021] 本文公开了代表生物体的基因组的基因组相位信息的核酸分子，所述核酸分子包含映射至单个基因组分子的至少20kb的核酸序列信息，其中所述序列信息包含相对于其在基因组分子中的位置重排的区段，并且其中至少70％的唯一地映射至所述生物体的基因组的序列信息映射至单个基因组分子。在一些方面，所述核酸分子包含至少20个区段。在一些方面，所述区段在所述生物体的基因组中不相邻。

[0022] 本文公开了包含至少100个至少20kb的核酸分子组分的核酸文库，其中组分包含生物体的基因组的重排区段；其中来自文库组分的至少70％的唯一映射区段映射至共同的基因组分子；并且其中组分不与核酸结合部分结合。

[0023] 本文公开了包含对应于至少100个至少20kb的核酸分子组分的序列的核酸数据集，其中组分包含生物体的基因组的至少5个重排区段，并且其中从下游分析中排除少于70％的所述重排区段映射至共同的支架的组分。

[0024] 本文公开了包含对应于至少100个至少20kb的核酸分子组分的序列的核酸数据集，其中组分包含生物体的基因组的至少5个重排区段，并且其中从下游分析中排除少于70％的所述序列唯一地映射至共同的支架的组分。

附图说明

[0025] 本发明的新颖特征在本文所附的权利要求书以及本文的发明内容和具体实施方

式中具体阐述。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：

[0026] 图1描绘了具有许多游离末端的消化的重构染色质聚集体，其具有与所有其他游离末端杂交相容的单链突出端。

[0027] 图2描绘了图1的消化的重构染色质聚集体，其具有单个碱基补平，使得每个单链突出端不相容以进行重新退火和重新连接。

[0028] 图3描绘了图2的部分补平的消化的重构染色质聚集体，其用与重构聚集体的修饰游离末端相容的标点寡核苷酸进行连接。

[0029] 图4描绘了由图3的连接产物产生并随后从DNA结合蛋白中释放的间断的(punctuated)DNA分子。每个基因组区段由通过其已知序列可识别的标点寡核苷酸勾画。基因组区段均代表该起始重构染色质聚集体中输入分子的一些区域。因此，该组中的读取是单倍体型定相的，并且可以用于组装或单元型相位重构。

[0030] 图5描绘了Chicago对的多联体生成。在顶部的图中，通过将消化的重构染色质聚集体的生物素化末端连接在一起(如图1中的末端，如果它们在连接之后得以生物素化和切割)生成Chicago读取对。这些分子在链霉亲和素包被的珠子上被捕获。然后，添加扩增衔接子。从链霉亲和素珠子上清液中大量扩增并收集所有分子。最后，将这些分子大量连接在一起以生成可使用长读取测序技术进行读取的长分子。嵌入的读取对可通过扩增衔接子进行识别。

[0031] 图6描绘了使间断分子如图4中描绘的分子或图5中生成的长分子条码化。首先，完成由条码和标点寡核苷酸的反向互补体组成的条码化寡核苷酸的递送。然后，延伸这些条码化寡核苷酸，使得产物含有条码、标点序列和一些基因组序列。

[0032] 图7描绘了在连接步骤('BF')之前和连接步骤('AF')之后两个样品的凝胶电泳分析，其证明成功连接以形成长的重排分子。

[0033] 图8呈现了从重排的基因组文库获得的数据。

[0034] 图9A描绘了由分隔成10kb区间(bin)的读取跨越的距离的频率分布。

[0035] 图9B描绘了由分隔成1kb区间的读取所跨越的距离的频率分布。

[0036] 图10描绘了用于实现本公开内容的计算机系统。

具体实施方式

[0037] 本文公开了使用长读取或短读取测序技术生成用于包括基因组组装和单倍体型定相的应用的读取集(包括定相读取集)的方法。可以将核酸分子结合(例如，在染色质结构中)、切割以暴露内部末端、在接头处重新附接至其他暴露的末端、从结合中释放出来并进行测序。该技术可产生包含多个序列区段的核酸分子。核酸分子内的多个序列区段可具有在相对于它们的天然或起始位置和方向进行重排时保留的相位信息。可以确信地认为接头的任一侧上的序列区段来自样品核酸分子的相同相位。

[0038] 核酸分子(包括高分子量DNA)可以结合或固定在至少一个核酸结合部分上。例如，组装成体外染色质聚集体并用甲醛处理进行固定的DNA与本文中的方法一致。核酸结合或固定方法包括但不限于体外或重构的染色质组装、天然染色质、DNA结合蛋白聚集体、纳米颗粒、DNA结合珠子或使用DNA结合物质包被的珠子、聚合物、合成的DNA结合分子或其他固

体或基本上固体的亲和分子。在一些情况下,所述珠子是固相可逆固定化(SPRI)珠子(例如,具有带负电荷的羧基基团的珠子如Beckman-Coulter Agencourt AMPure XP珠子)。

[0039] 可以保持与核酸结合部分(如本文所述的核酸结合部分)结合的核酸,使得具有在核酸分子上相隔大于测序装置上的读取距离的距离(例如,10kb、50kb、100kb或更大)的第一区段和第二区段的核酸分子独立于它们共同的磷酸二酯键而结合在一起。在切割这类结合的核酸分子后,第一区段和第二区段的暴露的末端可以彼此连接。在一些情况下,核酸分子以使得在固体表面上结合的核酸分子之间几乎没有或没有重叠的浓度结合,使得切割的分子的暴露的内部末端在切割之前可能仅重新连接或重新附接至在共同的核酸来源上同相位的其他区段的暴露的末端。因此,可以将DNA分子切割,并且经切割的暴露的内部末端可以例如随机地重新连接,而不丢失相位信息。

[0040] 可以通过任何数目的酶促和非酶促方法之一切割结合的核酸分子以暴露内部末端。例如,可以使用限制酶(如留下单链突出端的限制性内切核酸酶)消化核酸分子。例如,MboI消化适用于此目的,但也考虑了其他限制性内切核酸酶。可在例如大多数分子生物学产品目录中获得限制性内切核酸酶的列表。用于核酸切割的其他非限制性技术包括使用转座酶、标签化酶复合体、拓扑异构酶、非特异性内切核酸酶、DNA修复酶、RNA引导的核酸酶、片段化酶或备选的酶。例如,转座酶可以与未连接的左边界和右边界组合使用,以在核酸中产生序列非依赖性断裂,该断裂通过转座酶递送的寡核苷酸序列的附接进行标记。物理方法也可用于生成切割,包括机械方法(例如,声处理、剪切)、热方法(例如,温度变化)或电磁方法(例如,辐射,如UV辐射)。

[0041] 在该阶段固定核酸可以使切割的核酸分子片段保持紧密的物理邻近,从而保留初始分子的相位信息。图1中示意性地示出了从一个核酸结合部分所得到的示例性的染色质聚集体。所述固定的益处(例如固定至染色质聚集体上)是共同的核酸分子的单独的区域可以独立于它们的磷酸二酯骨架保持在一起,使得它们的相位信息在切割磷酸二酯骨架时不会丢失。该益处还可通过在切割之前附接核酸分子的备选支架来实现。

[0042] 任选地,对单链“粘性”末端突出端进行修饰以防止重新退火和重新连接。例如,将粘性末端部分补平,如通过添加一个核苷酸和聚合酶(图2)。以这种方式,不能补平整个单链末端,但可对末端进行修饰以防止与先前互补末端重新连接。在MboI消化(其留下5’GATC 5’突出端)的实例中,仅添加了鸟苷核苷酸三磷酸。这导致仅第一互补碱基(“C”)的“G”补平并导致5’GAT突出端。该步骤使得游离粘性末端不相容以便彼此重新连接,但保留粘性末端用于下游应用。或者,通过完全补平突出端、用平末端生成酶进行的限制性消化、用单链DNA外切核酸酶处理或非特异性切割来生成平末端。在一些情况下,转座酶用于将具有平末端或粘性末端的衔接子末端附接至DNA分子的暴露的内部末端。

[0043] 任选地,引入“标点(punctuation)寡核苷酸”(图3)。该标点寡核苷酸标记切割/重新连接位点。一些标点寡核苷酸在两端具有单链突出端,该单链突出端与在暴露的核酸样品内部末端上生成的部分补平的突出端相容。标点寡核苷酸的实例如下所示。在一些情况下,对具有单链突出端的双链寡核苷酸进行修饰,如通过在其5’端处的5’磷酸去除,使得它在连接过程中不能形成多联体。或者,使用钝性标点寡核苷酸,或不使用不同的标点寡核苷酸标记切割位点。在一些系统中,如当使用转座酶时,通过添加转座体边界序列,并随后将边界序列彼此连接或连接至标点寡核苷酸来完成加标点。示例性标点寡核苷酸如下所示。

然而，备选的标点寡核苷酸与本文中的公开内容一致，其在序列、长度、突出端的存在或序列、或修饰如5’去磷酸化方面不同。

[0044] 5’ATCACGCGC 3’

[0045] 3’TGCGCGCTA 5’

[0046] 在一些情况下，标点寡核苷酸的双链区域将有所不同。标点寡核苷酸的相关特征是其突出端的序列，从而允许与核酸样品连接，但任选地进行修饰，从而排除自动连接或多联体形成。通常优选的是，标点寡核苷酸包含不发生或不太可能发生在靶核酸分子中的序列，使得其在下游序列反应中容易被鉴定。任选地将标点寡核苷酸条码化，例如用已知的条码序列或用随机生成的唯一标识符序列。唯一标识符序列可被设计为使得核酸分子中或样品中的多个接头不太可能用相同的唯一标识符进行条码化。

[0047] 切割的末端可以直接或通过寡核苷酸(例如，标点寡核苷酸)彼此附接，例如使用连接酶或类似的酶。可以进行连接，使得固定的高分子量核酸分子的游离单链末端直接连接或与标点寡核苷酸连接(图3)。因为标点寡核苷酸(如果使用的话)可以具有两个可连接的末端，所以该连接可以有效地将高分子量核酸分子的区域链接在一起。也可以采用导致在两个暴露的末端之间附加间断序列或分子的备选方法，也可以采用直接连接两个暴露的末端而没有间断的方法。

[0048] 然后可以从核酸结合部分释放出核酸。在体外染色质聚集体的情况下，这可以通过将交联逆转、或消化蛋白质组分、或将交联逆转和消化蛋白质组分二者来实现。合适的方法是用蛋白酶K处理复合体，但也考虑了许多备选方案。对于其他结合技术，可以采用合适的方法，如切断连接体分子或降解基底。

[0049] 由这些技术产生的核酸分子可具有多种相关特征。核酸分子内的序列区段可以相对于它们的天然或起始位置和方向进行重排，但保留了相位信息。因此，接头的任一侧上的序列区段可以确信地分配给共同样品分子的共同相位。因此，通过这些技术，可以使在分子上彼此远离的区段聚集在一起或邻近，使得在单个分子测序装置的单次运行中对每个区段的部分或整体进行测序，从而允许确定的相位分配。或者，在一些情况下，最初相邻的区段可以与所得核酸中的区段分隔开。在一些情况下，可以重新连接核酸分子，使得至少约50％、55％、60％、65％、70％、75％、80％、85％、90％、91％、92％、93％、94％、95％、96％、97％、98％、99％、99.9％、99.99％、99.999％或100％的重新连接在切割之前位于在共同的核酸来源上同相位的区段之间。

[0050] 在一些情况下，所得分子的另一个相关特征是大部分或全部原始分子序列在最终的间断或重排分子中得以保留，尽管可能是重排的。例如，在一些情况下，在产生所得一个或多个分子时，丢失了不超过1％、2％、3％、4％、5％、10％、15％或20％的原始分子。因此，除了用作相位确定剂之外，所得分子还保留了很大比例的原始分子序列，使得所得分子任选地用于同时生成序列信息，如在从头测序中有用的叠连群信息或作为先前生成的叠连群信息的独立验证。

[0051] 一些所得分子的文库的另一个特征是切割接头(不是所得分子群的多个成员共有的)。也就是说，相同起始核酸分子的不同拷贝可以在末端具有不同的接头和重排模式。可以用非特异性切割分子或通过限制性内切核酸酶选择或消化参数的变化生成随机切割接头。

[0052] 具有分子特异性切割位点的结果是，在一些情况下，标点寡核苷酸被任选地从该过程中排除，其导致‘标点分子’重新改组并重新连接至无不良效果。通过比对三个或更多个重新改组的分子的区段，可以观察到切割位点很容易通过它们在文库的大多数其他成员中的缺失来鉴定。也就是说，当将三个或更多个重新改组的分子局部比对时，可以发现区段对于所有分子是共同的，但区段的边缘可以在分子之间变化。通过注意区段局部序列相似性在何处结束，可以在‘未间断的’重排核酸分子中映射切割接头。

[0053] 可以对得到的核酸分子(参见，例如，图4)进行测序，例如在长读取测序仪上。得到的序列读取含有在来自原始输入分子的核酸序列与(如果使用它们)标点寡核苷酸的序列之间交替的区段。这些读取可以由计算机进行处理以使用标点寡核苷酸序列从每个读取中分离序列数据，或者以其他方式进行处理来鉴定接头。每个读取内的序列区段可以是来自单个输入高分子量DNA分子的区段。原始核酸分子可包含基因组序列或其部分，如染色体。区段读取组在原始核酸分子中可以是不连续的，但揭示长范围单倍体型定相数据。这些数据可用于输入基因组中的从头基因组组装和定相杂合位置。接头之间的序列表示来源核酸样品中的连续核酸序列，而跨越接头的序列指示在核酸样品中同相位但可能在排列的支架中远离相邻区段的核酸区段。

[0054] 可以通过多种方法鉴定接头。如果使用标点寡核苷酸，可以在含有标点寡核苷酸序列的读取中鉴定接头。或者，可以通过与核酸分子的第二序列来源(并且优选地，第三序列来源)比较来鉴定接头，如先前生成的叠连群序列数据集或具有独立衍生的接头的第二独立生成的DNA链分子。例如，当比对序列时，与特定位置比对的质量或置信度可以指示一个区段结束而另一个区段开始的位置。如果使用限制酶生成切割，则可以评估含有限制酶识别位点的序列可能含有接头。注意，并非每个限制酶识别位点均可以含有接头，例如，当核酸与支持体结合时，一些限制酶识别位点可能尚未被酶物理接近。统计信息也可用于鉴定接头；例如，可以预测接头之间的长度区段具有某个平均值或遵循某种分布。

[0055] 本文中的操作的益处是它们可以保留分子相位信息，同时使分子的非相邻区域邻近，使得它们以适合于在单个读取(如长读取)中测序的距离包括在单个核酸分子中。因此，使在起始样品中相隔大于单个长读取操作的距离(例如10kb、15kb、20kb、30kb、50kb、100kb或更大)的区域局部邻近，使得它们处于由长范围测序反应的单个读取所覆盖的距离内。因此，在相位保留的重排分子中，在单个反应中读取对于原始样品中单个读取而言相隔超过测序技术范围的区域。

[0056] 可以对得到的重排分子进行测序，并将它们的序列信息映射至独立或同时生成的序列读取或叠连群信息，或映射至已知的参考基因组序列(例如，人类基因组的已知序列)。推测在得到的重排分子读取上相邻的区段是同相位的。因此，当将这些区段映射至不同的叠连群或长范围序列读取时，将读取分配给序列组装中的共同分子的共同相位。

[0057] 或者，如果同时对多个独立生成的所得重排分子进行测序，则任选地仅从这些分子生成定相样品数据，使得由接头隔开的区段序列被推断为同相位，而未由接头隔开的序列被推断为代表样品本身中连续的核酸段，并且可用于例如从头序列确定以及用于相位确定。然而，另外地或作为备选，仍然可以将同时测序的多个独立生成的所得重排分子与独立生成的支架或叠连群信息进行比较。

[0058] 本文提供的方法和组合物可以保留长范围相位信息，特别是对于相隔大于测序技

术中读取的长度(例如,10kb、20kb、50kb、100kb、500kb或更长)的分子区段,同时在区段相邻或足够接近以被单个读取覆盖的情况下在重排的或经常'间断的'分子中提供这样的非相邻区段。

[0059] 在一些情况下,将得到的重排分子与天然分子组合以用于测序。如果使用,可以通过缺少标点序列来信息地识别和利用天然分子。使用短读取或长读取技术对天然分子进行测序,并且由通过重排的分子或文库的测序生成的相位信息和区段序列信息来指导它们的组装。

[0060] 核酸提取

[0061] 用于提取和纯化适用于与本文公开内容一起使用的核酸的方法是本领域公知的。例如,通过采用苯酚、苯酚/氯仿/异戊醇,或类似制剂(包括TRIzol和TriReagent)进行有机提取来纯化核酸。提取技术的其他非限制性实例包括:(1)有机提取,随后进行乙醇沉淀,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);以及(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特异性或非特异性结合的磁性颗粒,随后使用磁铁分离珠子,并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,可在上述分离方法之前进行酶消化步骤,以帮助从样品中消除不需要的蛋白质,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RNA酶抑制剂。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。当在提取程序期间或之后将DNA和RNA二者一起分离时,可采用进一步的步骤来将其中一者或两者与另一者分离地进行纯化。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。

[0062] 可如例如在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,通过多种技术,如Maniatis等人,Molecular Cloning:A Laboratory Manual,Cold Spring Harbor,N.Y.,第280-281页(1982,其通过引用以其全文并入本文)所述的那些技术从生物样品中提取核酸。在一些情况下,首先可从生物样品中提取核酸,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(例如组蛋白)。在一些实施方案中,本公开内容易于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、真菌或病毒中分离的DNA。

[0063] 在一些实施方案中,从含有多种其他组分(如蛋白质、脂质和非模板核酸)的生物样品中分离核酸模板分子(例如,DNA或RNA)。核酸模板分子可以从获自动物、植物、细菌、真菌或任何其他细胞生物体或病毒的任何细胞材料获得,或者可以是人工合成的。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可以是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

品也可以是从生物样本提取的总RNA、cDNA文库、病毒或基因组DNA。样品还可以包含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。

[0064] 可将核酸分子(包括高分子量DNA)结合或固定在核酸结合部分上。例如，组装成体外染色质聚集体并用甲醛处理进行固定的DNA与本文中的方法一致。核酸结合或固定方法包括但不限于体外或重构的染色质组装、天然染色质、DNA结合蛋白聚集体、纳米颗粒、DNA结合珠子或使用DNA结合物质包被的珠子、聚合物、合成的DNA结合分子或其他固体或基本上固体的亲和分子。在一些情况下，珠子是固相可逆固定化(SPRI)珠子(例如，具有带负电荷的羧基基团的珠子如Beckman-Coulter Agencourt AMPure XP珠子)。

[0065] 可以保持核酸，如与核酸结合部分(如本文所述的那些核酸结合部分)结合的核酸，使得具有在核酸分子上相隔大于测序装置上的读取距离的距离(例如，10kb、50kb、100kb或更大)的第一区段和第二区段的核酸分子独立于它们共同的磷酸二酯键而结合在一起。在切割这类结合的核酸分子后，第一区段和第二区段的暴露的末端可以彼此连接。在一些情况下，核酸分子以使得在固体表面上结合的核酸分子之间几乎没有或没有重叠的浓度结合，使得切割的分子的暴露的内部末端在切割之前可能仅重新连接或重新附接至在共同的核酸来源上同相位的其他区段的暴露的末端。因此，可以将DNA分子切割，并且经切割的暴露的内部末端可以例如随机地重新连接，而不丢失相位信息。在一些情况下，核酸分子可以重新连接，使得至少约50％、55％、60％、65％、70％、75％、80％、85％、90％、91％、92％、93％、94％、95％、96％、97％、98％、99％、99.9％、99.99％、99.999％或100％的重新连接在切割之前位于在共同的核酸来源上同相位的区段之间。

[0066] 在一些情况下，通过可用于结合的表面积的量来控制表面上结合的核酸的表面密度。例如，选择用于结合核酸的珠子的大小可以影响或控制核酸之间的距离，或结合的核酸的平均表面密度。较大的珠子表面可导致结合的核酸之间的距离更大。这可导致核酸或核酸复合体之间的分子间连接事件的速率降低。所使用的珠子的直径可以为约100纳米(nm)、200nm、300nm、400nm、500nm、600nm、700nm、800nm、900nm、1微米(μm)、1.1μm、1.2μm、1.3μm、1.4μm、1.5μm、1.6μm、1.7μm、1.8μm、1.9μm、2μm、3μm、4μm、5μm、6μm、7μm、8μm、9μm、10μm、11μm、12μm、13μm、14μm、15μm、16μm、17μm、18μm、19μm、20μm、21μm、22μm、23μm、24μm、25μm、26μm、27μm、28μm、29μm、30μm、31μm、32μm、33μm、34μm、35μm、36μm、37μm、38μm、39μm、40μm、41μm、42μm、43μm、44μm、45μm、46μm、47μm、48μm、49μm、50μm、55μm、60μm、65μm、70μm、75μm、80μm、85μm、90μm、95μm、100μm、200μm、300μm、400μm、500μm、600μm、700μm、800μm、900μm或1毫米(mm)。所使用的珠子的直径可以为至少约100纳米(nm)、200nm、300nm、400nm、500nm、600nm、700nm、800nm、900nm、1微米(μm)、1.1μm、1.2μm、1.3μm、1.4μm、1.5μm、1.6μm、1.7μm、1.8μm、1.9μm、2μm、3μm、4μm、5μm、6μm、7μm、8μm、9μm、10μm、11μm、12μm、13μm、14μm、15μm、16μm、17μm、18μm、19μm、20μm、21μm、22μm、23μm、24μm、25μm、26μm、27μm、28μm、29μm、30μm、31μm、32μm、33μm、34μm、35μm、36μm、37μm、38μm、39μm、40μm、41μm、42μm、43μm、44μm、45μm、46μm、47μm、48μm、49μm、50μm、55μm、60μm、65μm、70μm、75μm、80μm、85μm、90μm、95μm、100μm、200μm、300μm、400μm、500μm、600μm、700μm、800μm、900μm或1毫米(mm)。所使用的珠子的直径可以为至多约100纳米(nm)、200nm、300nm、400nm、500nm、600nm、700nm、800nm、900nm、1微米(μm)、1.1μm、1.2μm、1.3μm、1.4μm、1.5μm、1.6μm、1.7μm、1.8μm、1.9μm、2μm、3μm、4μm、5μm、6μm、7μm、8μm、9μm、10μm、11μm、12μm、13μm、14μm、15μm、16μm、17μm、18μm、19μm、20μm、21μm、22μm、23μm、24μ

m、25μm、26μm、27μm、28μm、29μm、30μm、31μm、32μm、33μm、34μm、35μm、36μm、37μm、38μm、39μm、40μm、41μm、42μm、43μm、44μm、45μm、46μm、47μm、48μm、49μm、50μm、55μm、60μm、65μm、70μm、75μm、80μm、85μm、90μm、95μm、100μm、200μm、300μm、400μm、500μm、600μm、700μm、800μm、900μm或1毫米(mm)。

[0067] 核酸结合部分复合体的形成

[0068] 核酸可以与核酸结合部分结合以在核酸分子的切割之后保留相位信息。许多核酸结合部分形成与本文公开内容一致的支架。适合于本文公开内容的一些支架在多个点结合核酸,使得在核酸分子的切割和重新连接时不会丢失相位信息。

[0069] 在一些情况下,核酸结合部分是或包含一类蛋白质,如形成染色质的组蛋白。染色质可以是重构的染色质或天然染色质。在一些情况下,核酸结合部分分布在固体支持体,如微阵列、载玻片、芯片、微孔、柱、管、颗粒或珠子上。在一些实例中,固体支持体包被有链霉亲和素和/或亲和素。在其他实例中,固体支持体包被有抗体。此外,固体支持体可另外地或备选地包含玻璃、金属、陶瓷或聚合物材料。在一些实施方案中,固体支持体为核酸微阵列(例如,DNA微阵列)。在其他实施方案中,固体支持体可以为顺磁珠。

[0070] 在一些情况下,DNA样品与多个缔合分子交联。在各种情况下,缔合分子包含氨基酸。在许多情况下,缔合分子包含肽或蛋白质。在其他情况下,缔合分子包含组蛋白。在其他情况下,缔合分子包含纳米颗粒。在一些情况下,纳米颗粒为铂基纳米颗粒。在其他情况下,纳米颗粒为DNA嵌入剂或其任何衍生物。在其他情况下,纳米颗粒为双嵌入剂或其任何衍生物。在某些情况下,缔合分子来自与第一DNA分子不同的来源。交联可以作为本文公开的方案的一部分进行,或者可以先前已经进行。例如,可以采用本公开内容的技术处理和分析先前固定的样品(例如,福尔马林固定石蜡包埋(FFPE))样品。

[0071] 形成结构的核酸结合部分的实例是重构的染色质。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,在一些情况下从分离的裸DNA中生成重构的染色质。对于许多样品,通过使用多种非侵入性至侵入性的方法中的任一种,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。这些方法通常比分离天然染色质更容易、更快且更便宜。

[0072] 第二,重构染色质大幅减少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倍体型定相的人工制品。在一些情况下,根据本公开内容的方法和组合物,样品具有少于约30%、29%、28%、27%、26%、25%、24%、23%、22%、21%、20%、19%、18%、17%、16%、15%、13%、14%、12%、11%、10%、9%、8%、7%、6%、5%、4%、3%、2%、1%、0.5%、0.4%、0.3%、0.2%、0.1%、0.01%、0.001%或更少的染色体间或分子间交联。在一些实例中,所述样品具有少于约30%的染色体间或分子间交联。在一些实例中,所述样品具有少于约25%的染色体间或分子间交联。在一些实例中,所述样品具有少于约20%的染色体间或分子间交联。在一些实例中,所述样品具有少于约15%的染色体间或分子间交联。在一些实例中,所述样品具有少于约10%的染色体间或分子间交联。在一些实例中,所述样品具有少于约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。在其他实例中,所述样品可具有少于约1%的染色体间或分子间交联。由于染色体间相互作用表示非同相位的分子部分之间的相互作用,因此它们的减少或消除有益于本公开内容的一些目标,即,定相的核酸信息的有效、快速组

装。

[0073] 第三，调节能够交联的位点的频率，并因此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。

[0074] 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％的交联以便连接在样品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。

[0075] 核酸结合部分支架如重构的染色质的重要益处是其独立于组成核酸的磷酸二酯键而保留了其组成核酸的物理连锁信息。因此，通过重构的染色质保持在一起的核酸任选地进行交联以维持稳定性，即使它们的磷酸二酯键断裂也将保持它们的邻近，如在内部标记中可能发生的那样。由于重构的染色质，即使进行切割，片段也将保持邻近，从而在内部标记过程中保留相位信息或物理连锁信息。因此，当重新连接暴露的末端时，这些末端将连接至来源于共同分子的共同相位的区段。

[0076] 重构的染色质组装

[0077] 在一些情况下，通过将重构的染色质组装到核酸样品上来完成核酸组装到核酸结合部分上以在核酸分子的切割和重排过程中保留相位信息。如本文所用的重构染色质被广泛使用，范围从将天然染色质组分重新组装到核酸上至核酸与非生物颗粒结合。

[0078] 参考传统意义上的重构染色质，核芯组蛋白和DNA组装成核小体是由伴侣蛋白和相关的组装因子介导的。几乎所有这些因子均为核芯组蛋白结合蛋白。一些组蛋白伴侣蛋白，如核小体组装蛋白-1(NAP-1)，显示出与组蛋白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起重要作用。

[0079] 通常，已经开发了两种体外方法用于重构或组装染色质，但考虑了这些方法的变化。一组方法涉及不依赖ATP的组装，而第二组方法为ATP依赖性的。

[0080] 用于重构染色质的不依赖ATP的方法涉及DNA和核芯组蛋白加上蛋白质如NAP-1或盐以充当组蛋白伴侣蛋白。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。

[0081] 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)是两种用于使核小体的延伸有序的阵列在体外生成染色质的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)。

[0082] 还考虑了备选的组装方法，例如不依赖组蛋白来构成重构染色质的方法。可以向核酸添加任何DNA结合部分以形成广泛定义的某些类型的重构染色质。

[0083] 在一些实施方案中，考虑了非天然染色质类似物。本文考虑了纳米颗粒，如具有促进核酸结合的正涂覆的外表面、或可激活用于与核酸交联的表面，或者促进核酸结合的正涂覆的外表面和可激活用于与核酸交联的表面二者的纳米颗粒。在一些实施方案中，纳米颗粒包含硅。

[0084] 在一些情况下，本文公开的方法与同纳米颗粒相关的DNA一起使用。在一些情况下，所述纳米颗粒带正电荷。例如，所述纳米颗粒涂覆有胺基团和/或含胺的分子。所述DNA和纳米颗粒聚集并凝聚，类似于天然或重构的染色质。此外，诱导纳米颗粒结合的DNA来以模拟生物核小体的有序阵列(即染色质)的方式聚集。基于纳米颗粒的方法可以更便宜、更快地组装，提供比使用重构染色质更好的回收率，并且/或者允许降低DNA输入要求。

[0085] 可以改变许多因素来影响凝聚的程度和形式，该凝聚包括溶液中纳米颗粒的浓度、纳米颗粒与DNA的比例，以及所用纳米颗粒的大小。在一些情况下，将所述纳米颗粒以大于约1ng/mL、2ng/mL、3ng/mL、4ng/mL、5ng/mL、6ng/mL、7ng/mL、8ng/mL、9ng/mL、10ng/mL、15ng/mL、20ng/mL、25ng/mL、30ng/mL、40ng/mL、50ng/mL、60ng/mL、70ng/mL、80ng/mL、90ng/mL、100ng/mL、120ng/mL、140ng/mL、160ng/mL、180ng/mL、200ng/mL、250ng/mL、300ng/mL、400ng/mL、500ng/mL、600ng/mL、700ng/mL、800ng/mL、900ng/mL、1μg/mL、2μg/mL、3μg/mL、4μg/mL、5μg/mL、6μg/mL、7μg/mL、8μg/mL、9μg/mL、10μg/mL、15μg/mL、20μg/mL、25μg/mL、30μg/mL、40μg/mL、50μg/mL、60μg/mL、70μg/mL、80μg/mL、90μg/mL、100μg/mL、120μg/mL、140μg/mL、160μg/mL、180μg/mL、200μg/mL、250μg/mL、300μg/mL、400μg/mL、500μg/mL、600μg/mL、700μg/mL、800μg/mL、900μg/mL、1mg/mL、2mg/mL、3mg/mL、4mg/mL、5mg/mL、6mg/mL、7mg/mL、8mg/mL、9mg/mL、10mg/mL、15mg/mL、20mg/mL、25mg/mL、30mg/mL、40mg/mL、50mg/mL、60mg/mL、70mg/mL、80mg/mL、90mg/mL或100mg/mL的浓度添加至所述DNA。在一些情况下，将所述纳米颗粒以小于约1ng/mL、2ng/mL、3ng/mL、4ng/mL、5ng/mL、6ng/mL、7ng/mL、8ng/mL、9ng/mL、10ng/mL、15ng/mL、20ng/mL、25ng/mL、30ng/mL、40ng/mL、50ng/mL、60ng/mL、70ng/mL、80ng/mL、90ng/mL、100ng/mL、120ng/mL、140ng/mL、160ng/mL、180ng/mL、200ng/mL、250ng/mL、300ng/mL、400ng/mL、500ng/mL、600ng/mL、700ng/mL、800ng/mL、900ng/mL、1μg/mL、2μg/mL、3μg/mL、4μg/mL、5μg/mL、6μg/mL、7μg/mL、8μg/mL、9μg/mL、10μg/mL、15μg/mL、20μg/mL、25μg/mL、30μg/mL、40μg/mL、50μg/mL、60μg/mL、70μg/mL、80μg/mL、90μg/mL、100μg/mL、120μg/mL、140μg/mL、160μg/mL、180μg/mL、200μg/mL、250μg/mL、300μg/mL、400μg/mL、500μg/mL、600μg/mL、700μg/mL、800μg/mL、900μg/mL、1mg/mL、2mg/mL、3mg/mL、4mg/mL、5mg/mL、6mg/mL、7mg/mL、8mg/mL、9mg/mL、10mg/mL、15mg/mL、20mg/mL、25mg/mL、30mg/mL、40mg/mL、50mg/mL、60mg/mL、70mg/mL、80mg/mL、90mg/mL或100mg/mL的浓度添加至所述DNA。在一些情况下，将所述纳米颗粒以大于约1:10000、1:5000、1:2000、1:1000、1:500、1:200、1:100、1:50、1:20、1:10、1:5、1:2、1:1、2:1、5:1、10:1、20:1、50:1、100:1、200:1、500:1、1000:1、2000:1、5000:1或10000:1的重量-重量(w/w)比添加至所述DNA。在一些情况下，将所述纳米颗粒以小于约1:10000、1:5000、1:2000、1:1000、1:500、1:200、1:100、1:50、1:20、1:10、1:5、1:2、1:1、2:1、5:1、10:1、20:1、50:1、100:1、200:1、500:1、1000:1、2000:1、5000:1或10000:1的重量-重量(w/w)比添加至所述DNA。在一些情况下，所述纳米颗粒具有大于约1nm、1nm、2nm、

3nm、4nm、5nm、6nm、7nm、8nm、9nm、10nm、15nm、20nm、25nm、30nm、40nm、50nm、60nm、70nm、80nm、90nm、100nm、120nm、140nm、160nm、180nm、200nm、250nm、300nm、400nm、500nm、600nm、700nm、800nm、900nm、1μm、2μm、3μm、4μm、5μm、6μm、7μm、8μm、9μm、10μm、15μm、20μm、25μm、30μm、40μm、50μm、60μm、70μm、80μm、90μm或100μm的直径。在一些情况下，所述纳米颗粒具有小于约1nm、1nm、2nm、3nm、4nm、5nm、6nm、7nm、8nm、9nm、10nm、15nm、20nm、25nm、30nm、40nm、50nm、60nm、70nm、80nm、90nm、100nm、120nm、140nm、160nm、180nm、200nm、250nm、300nm、400nm、500nm、600nm、700nm、800nm、900nm、1μm、2μm、3μm、4μm、5μm、6μm、7μm、8μm、9μm、10μm、15μm、20μm、25μm、30μm、40μm、50μm、60μm、70μm、80μm、90μm或100μm的直径。

[0086] 此外，可以通过施加磁场(在顺磁性纳米颗粒的情况下)或通过共价附接(例如通过与聚赖氨酸涂覆的基底交联)将所述纳米颗粒固定在固体基底(例如珠子、载玻片或管壁)上。所述纳米颗粒的固定化可提高连接效率，从而相对于不期望的产物(噪音)增加期望的产物(信号)的数目。

[0087] 任选地使重构的染色质与交联剂如甲醛接触，以进一步使DNA-染色质复合体稳定。

[0088] 核酸切割

[0089] 可以处理结合的核酸以暴露内部双链末端。可以用限制酶(如限制性内切核酸酶)进行切割。备选的切割方法也与本文的公开内容一致。例如，转座酶任选地与未连接的左边界和右边界寡核酸(oligonucleic acid)分子组合使用，以便在核酸中产生不依赖序列的断裂，该断裂通过转座酶递送的寡核酸分子的附接进行标记。在一些情况下，将寡核酸分子合成为包含标点相容的突出端，或彼此相容，使得寡核酸分子彼此连接并用作标点分子。这种类型的备选方法的益处在于切割是不依赖于序列的，并因此即使两个核酸分子的序列局部相同，也更可能从核酸的一个拷贝变化到另一个拷贝。

[0090] 在一些情况下，所述暴露的核酸末端理想地是粘性末端，例如作为与限制性内切核酸酶接触的结果。在一些情况下，使用限制性内切核酸酶切割可预测的突出端，随后与核酸末端(如标点寡核苷酸)连接，该核酸末端包含与DNA片段上可预测的突出端互补的突出端。在一些实施方案中，部分地补平限制性内切核酸酶生成的突出端的5’和/或3’端。在一些情况下，采用单个核苷酸补平突出端。

[0091] 在一些情况下，具有突出端的DNA片段可与一个或多个核酸(如具有互补突出端的标点寡核苷酸、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，使用不依赖于模板的聚合酶将单个腺嘌呤添加至末端修复的DNA片段的3’端，随后与一个或多个标点寡核苷酸连接，每个标点寡核苷酸在3’端处具有胸腺嘧啶。在一些实施方案中，核酸如寡核苷酸或多核苷酸与平末端双链DNA分子连接，该平末端双链DNA分子已经通过用一个或多个核苷酸延伸3’端并随后进行5’磷酸化而得以修饰。在一些情况下，在含有镁的合适的缓冲液中存在一种或多种dNTP的情况下，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，具有平末端的靶多核苷酸与一个或多个包含平末端的衔接子连接。可以例如在含有ATP和镁的合适缓冲液中采用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。

[0092] 标点寡核苷酸

[0093] 在一些情况下，标点寡核苷酸可用于连接暴露的切割的末端。标点寡核苷酸包括可以与靶多核苷酸连接以桥接经历相位保留重排的样品分子的两个切割的内部末端的任何寡核苷酸。标点寡核苷酸可包括DNA、RNA、核苷酸类似物、非规范核苷酸、标记的核苷酸、修饰的核苷酸或其组合。在许多实例中，双链标点寡核苷酸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平末端、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。在一些情况下，不同的标点寡核苷酸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标点寡核苷酸。或者，标点寡核苷酸群体在一些情况下是均匀的。

[0094] 可在与靶多核苷酸组合之前操作标点寡核苷酸。例如，可去除末端磷酸。这样的修饰排除了标点寡核苷酸相对于彼此的位置(而不是相对于样品分子的切割的内部末端的位置)。

[0095] 标点寡核苷酸含有多种序列元件中的一种或多种，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、多个不同标点寡核苷酸或不同标点寡核苷酸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、一个或多个随机或近似随机序列及其组合。在一些实例中，两个或更多个序列元件彼此不相邻(例如，由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充当测序引物退火序列。在某些情况下，序列元件位于3’端或其附近、5’端或其附近或标点寡核苷酸的内部。

[0096] 在备选的实施方案中，标点寡核苷酸包含最小的碱基互补体以维持双链分子的完整性，从而使其在测序反应中占据的序列信息的量最小化，或者标点寡核苷酸包含用于连接的最佳数目的碱基，或标点寡核苷酸长度是任意确定的。

[0097] 在一些实施方案中，标点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。在某些情况下，互补突出端的长度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，互补突出端的长度为约1、2、3、4、5或6个核苷酸。在一些实施方案中，标点寡核苷酸突出端与通过限制性内切核酸酶消化或其他DNA切割方法产生的靶多核苷酸突出端互补。

[0098] 标点寡核苷酸可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，标点寡核苷酸的长度为约、小于约或大于约4、5、6、7、8、9、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，标点寡核苷酸的长度为5至15个核苷酸。在其他实例中，标点寡核苷酸的长度为约20至约40个核苷酸。

[0099] 优选地，对标点寡核苷酸进行修饰，例如通过5’磷酸切除(经由小牛碱性磷酸酶处理，或在没有这些部分的情况下通过合成从头进行)，使得它们彼此不连接形成多聚体。3’OH(羟基)部分能够与切割的核酸上的5’磷酸连接，从而支持与第一或第二核酸区段的连接。

[0100] 衔接子寡核苷酸

[0101] 衔接子包括具有可与靶多核苷酸连接的序列的任何寡核苷酸。在各种实例中，衔

接子寡核苷酸包含DNA、RNA、核苷酸类似物、非规范核苷酸、标记的核苷酸、修饰的核苷酸或其组合。在一些情况下，衔接子寡核苷酸可以是单链、双链或部分双链体。通常，部分双链体衔接子寡核苷酸包含一个或多个单链区域和一个或多个双链区域。双链衔接子寡核苷酸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平末端、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，单链衔接子寡核苷酸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)。当衔接子寡核苷酸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寡核苷酸由包含内部杂交的单个衔接子寡核苷酸组成，或包含彼此杂交的两个或更多个衔接子寡核苷酸。在一些情况下，内部序列杂交，如在衔接子寡核苷酸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。在一些实例中，不同种类的衔接子寡核苷酸组合使用，如发夹衔接子和双链衔接子或不同序列的衔接子。在某些情况下，发夹衔接子中的可杂交序列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”。在一些情况下，不同的衔接子寡核苷酸在顺序反应中或同时与靶多核苷酸连接。例如，在同一反应中添加第一和第二衔接子寡核苷酸。在一些实例中，在与靶多核苷酸组合之前操作衔接子寡核苷酸。例如，可以添加或去除末端磷酸。

[0102] 衔接子寡核苷酸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。在许多实例中，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充当测序引物退火序列。序列元件位于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。互补突出端可另外地或备选地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寡核苷酸库中。在一些实施方案中，衔接子寡核苷酸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，衔接子寡核苷酸突出端由腺嘌呤或胸腺嘧啶组成。

[0103] 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，衔接子寡核苷酸的长度为约、小于约或大于约4、5、6、7、8、9、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，衔接子寡核苷酸的长度为5至15个核苷酸。在其他实例中，衔接子寡核苷酸的长度为约20至约40个核苷酸。

[0104] 优选地，对衔接子寡核苷酸进行修饰，例如通过5’磷酸切除(经由小牛碱性磷酸酶处理，或在没有这些部分的情况下通过合成从头进行)，使得它们彼此不连接形成多聚体。3’OH(羟基)部分能够与切割的核酸上的5’磷酸连接，从而支持与第一或第二核酸区段的连接。

[0105] 确定核酸样品的相位信息

[0106] 为了确定核酸样品的相位信息，首先例如通过本文讨论的提取方法获取核酸。在许多情况下，然后将核酸附接至固体表面，以在核酸分子切割后保留相位信息。优选地，将核酸分子在体外与核酸结合蛋白组装以生成重构的染色质，尽管其他合适的固体表面包括核酸结合蛋白聚集体、纳米颗粒、核酸结合珠子或使用核酸结合物质包被的珠子、聚合物、合成的核酸结合分子，或其他固体或基本上固体的亲和分子。还可以获得已经附接至固体表面的核酸样品，如在天然染色质的情况下。可以获得已经固定的天然染色质，如以福尔马林固定石蜡包埋(FFPE)或类似保存的样品的形式。

[0107] 在附接至核酸结合部分后，可以切割结合的核酸分子。用任何合适的核酸切割实体(包括任何数目的酶促和非酶促方法)进行切割。优选地，用限制性内切核酸酶、片段化酶或转座酶进行DNA切割。备选地或另外地，用其他限制酶、拓扑异构酶、非特异性内切核酸酶、核酸修复酶、RNA引导的核酸酶或备选的酶实现核酸切割。也可使用物理方法来生成切割，包括机械方法(例如，声处理、剪切)、热方法(例如，温度变化)或电磁方法(例如，辐射，如UV辐射)。核酸切割产生游离核酸末端，或具有“粘性”突出端或平末端，其取决于所用的切割方法。当生成粘性突出端末端时，任选地部分补平粘性末端以防止重新连接。或者，完全补平突出端以产生平末端。

[0108] 在许多情况下，用任选地进行标记的dNTP部分地或完全地补平突出端末端。在这样的情况下，dNTP可以是生物素化的、硫酸化的、与荧光团附接的、去磷酸化的或任何其他数目的核苷酸修饰。核苷酸修饰还可包括表观遗传修饰，如甲基化(例如，5-mC、5-hmC、5-fC、5-caC、4-mC、6-mA、8-oxoG、8-oxoA)。可从测序过程中可检测的标记物或修饰(如通过纳米孔测序可检测的表观遗传修饰)来选择标记物或修饰；以这种方式，可在测序过程中检测

连接接头的位置。这些标记物或修饰也可靶向用于结合或富集;例如,靶向甲基胞嘧啶的抗体可用于捕获、靶向、结合或标记用甲基胞嘧啶补平的平末端。非天然核苷酸、非规范或修饰的核苷酸和核酸类似物也可用于标记平末端补平的位置。非规范或修饰的核苷酸可包括假尿苷(Ψ)、二氢尿苷(D)、肌苷(I)、7-甲基鸟苷(m7G)、黄嘌呤、次黄嘌呤、嘌呤、2,6-二氨基嘌呤和6,8-二氨基嘌呤。核酸类似物可包括肽核酸(PNA)、吗啉代核酸和锁核酸(LNA)、乙二醇核酸(GNA)和苏糖核酸(TNA)。在一些情况下,用未标记的dNTP(如不含生物素的dNTP)补平突出端。在一些情况下,如通过用转座子切割,生成不需要补平的平末端。这些游离的平末端在转座酶插入两个未连接的标点寡核苷酸时生成。然而,标点寡核苷酸被合成为具有所需的粘性或平末端。也可对与样品核酸相关的蛋白质(如组蛋白)进行修饰。例如,可将组蛋白乙酰化(例如,在赖氨酸残基处)和/或甲基化(例如,在赖氨酸和精氨酸残基处)。

[0109] 接下来,虽然切割的核酸分子仍然与固体表面结合,但游离核酸末端连接在一起。在一些情况下,通过在游离末端之间连接或与单独的实体(如寡核苷酸)连接发生连接。在一些情况下,所述寡核苷酸为标点寡核苷酸。在这样的情况下,标点分子末端与切割的核酸分子的游离末端相容。在许多情况下,将标点分子去磷酸化以防止寡核苷酸的连环化。在大多数情况下,标点分子在每个末端上连接至切割的核酸分子的游离核酸末端。在许多情况下,该连接步骤导致切割的核酸分子的重排,使得起始核酸分子中最初彼此不相邻的两个游离末端现在以成对末端连接。

[0110] 在连接切割的核酸分子的游离末端后,使用任何数目的标准酶促和非酶促方法从核酸结合部分释放重排的核酸样品。例如,在体外重构的染色质的情况下,通过核酸结合蛋白的变性或降解释放重排的核酸分子。在其他实例中,交联是逆转的。在其他实例中,逆转或阻断亲和力相互作用。与输入的核酸分子相比,重排释放的核酸分子。在使用标点分子的情况下,由于穿插在整个重排的核酸分子中的标点寡核苷酸,得到的重排分子被称为间断分子。在这些情况下,位于标点侧翼的核酸区段构成成对末端。

[0111] 在本文公开的方法的切割和连接步骤期间,由于在整个这些过程中核酸分子与固体表面结合,因此相位信息得以维持。这可以在不依赖于来自其他标志物的信息(如单核苷酸多态性(SNP))的情况下实现相位信息的分析。使用本文公开的方法和组合物,在一些情况下,对核酸分子内的两个核酸区段进行重排,使得它们比它们在原始核酸分子上更接近。在许多实例中,起始核酸样品中两个核酸区段的原始分隔距离大于标准测序技术的平均读取长度。例如,输入核酸样品中两个核酸区段之间的起始分隔距离为约10kb、12.5kb、15kb、17.5kb、20kb、25kb、30kb、35kb、40kb、45kb、50kb、60kb、70kb、80kb、90kb、100kb、125kb、150kb、200kb、300kb、400kb、500kb、600kb、700kb、800kb、900kb、1Mb或更大。在优选的实例中,两个重排的DNA区段之间的分隔距离小于标准测序技术的平均读取长度。例如,在重排的DNA分子内分隔两个重排的DNA区段的距离小于约50kb、40kb、30kb、25kb、20kb、17kb、15kb、14kb、13kb、12kb、11kb、10kb、9kb、8kb、7kb、6kb、5kb或更小。在优选的情况下,该分隔距离小于长读取测序仪的平均读取长度的距离。在这些情况下,当重排的DNA样品从核酸结合部分释放并进行测序时,确定相位信息并生成足以生成从头序列支架的序列信息。

[0112] 将重排核酸分子条码化

[0113] 在一些实例中,在测序之前进一步处理本文所述的释放的重排核酸分子。例如,可将包含在重排核酸分子内的核酸区段条码化。条码化可允许更容易地对序列读取进行分

组。例如，条码可用于鉴定源自相同重排核酸分子的序列。条码还可用于唯一地标识单个接头。例如，每个接头可以用唯一的(例如，随机生成的)条码进行标记，该条码可以唯一地鉴定接头。多个条码可以一起使用，如鉴定源自相同重排核酸分子的序列的第一条码和唯一地鉴定单个接头的第二条码。

[0114] 可通过许多技术来实现条码化。在一些情况下，条码可作为序列包含在标点寡核苷酸中。在其他情况下，可以使释放的重排核酸分子与包含至少两个区段(一个区段含有条码，并且第二区段含有与标点序列互补的序列)的寡核苷酸接触。在与标点序列退火后，用聚合酶延伸条码化的寡核苷酸，以从相同的间断核酸分子产生条码化分子。由于间断核酸分子是输入核酸分子的重排形式(其中相位信息被保留)，因此所生成的条码化分子也来自相同的输入核酸分子。这些条码化分子包含条码序列、标点互补序列和基因组序列。

[0115] 对于具有或不具有标点的重排核酸分子，可以通过其他方式对分子进行条码化。例如，可使重排的核酸分子与条码化的寡核苷酸接触，该条码化的寡核苷酸可进行延伸以掺入来自重排的核酸分子的序列。条码可以与标点序列、限制酶识别位点、感兴趣的位点(例如，感兴趣的基因组区域)或随机位点(例如，通过条码寡核苷酸上的随机n-聚体序列)杂交。可以使用适当浓度和/或与样品中其他重排的核酸分子的分离(例如，空间或时间分离)使重排的核酸分子与条码接触，使得随后不给予多个重排的核酸分子相同的条码序列。例如，可以将包含重排的核酸分子的溶液稀释至这样的浓度，使得只有一个重排的核酸分子与具有给定条码序列的条码或条码组接触。条码可在游离溶液中、在流体分区(例如，液滴或孔)中或在阵列上(例如，在特定阵列点处)与重排的核酸分子接触。

[0116] 可以例如在短读取测序仪上对条码化核酸分子(例如，延伸产物)进行测序，并且通过将具有相同条码的序列读取分组成共同相位来确定相位信息。或者，在测序之前，可以将条码化产物连接在一起，例如通过大量连接，以生成经测序(例如使用长读取测序技术)的长分子。在这些情况下，嵌入的读取对可通过扩增衔接子和标点序列进行识别。从读取对的条码序列获得进一步的相位信息。

[0117] 用成对末端确定相位信息

[0118] 本文进一步提供了用于从成对末端确定相位信息的方法和组合物。可通过公开的任何方法或在所提供的实施例中进一步说明的方法生成成对末端。例如，在核酸分子与固体表面结合并随后进行切割的情况下，在重新连接游离末端后，重新连接的核酸区段从固相附接的核酸分子释放，例如，通过限制性消化。该释放导致多个成对末端。在一些情况下，将成对末端与扩增衔接子连接、扩增，并用短距离技术进行测序。在这些情况下，来自多个不同核酸结合部分结合的核酸分子的成对末端位于测序的样品内。然而，可以确信地得出结论，对于成对末端接头的任一侧，接头相邻序列来源于共同的分子的共同相位。在成对末端与标点寡核苷酸连接的情况下，测序读取中的成对末端接头由标点寡核苷酸序列鉴定。在其他情况下，成对末端通过修饰的核苷酸进行连接，其可以基于所用的修饰的核苷酸的序列进行鉴定。

[0119] 或者，在释放成对末端后，可将游离的成对末端与扩增衔接子连接并进行扩增。在这些情况下，然后将多个成对末端大量连接在一起以生成使用长读取测序技术读取的长分子。在其他实例中，释放的成对末端彼此大量连接而没有插入的扩增步骤。在任一情况下，嵌入的读取对可通过与连接序列(如标点序列或修饰的核苷酸)相邻的天然DNA序列进行识

别。在长序列装置上读取连接的成对末端，并获得多个接头的序列信息。由于成对末端来源于多个不同的核酸结合部分结合的DNA分子，因此发现跨越两个个体成对末端的序列，如位于扩增衔接子序列侧翼的序列，映射至多个不同的DNA分子。然而，可以确信地得出结论，对于成对末端接头的任一侧，接头相邻序列来源于共同的分子的共同相位。例如，在成对末端来源于间隔分子的情况下，将位于标点序列侧翼的序列确信地分配给共同的DNA分子。在优选的情况下，因为使用本文公开的方法和组合物连接单个成对末端，所以可在单个读取中对多个成对末端进行测序。

[0120] 测序方法

[0121] 与输入的DNA样品相比，本文公开的方法和组合物可用于生成包含重排区段的长DNA分子。这些分子是使用任何数目的测序技术的序列。优选地，使用标准长读取测序技术对长分子进行测序。另外地或备选地，可以如本文所公开的那样修饰生成的长分子以使它们与短读取测序技术相容。

[0122] 示例性的长读取测序技术包括但不限于纳米孔测序技术和其他长读取测序技术如Pacific Biosciences单分子实时(SMRT)测序。纳米孔测序技术包括但不限于Oxford纳米孔测序技术(例如，GridION、MinION)和Genia测序技术。

[0123] 序列读取长度可以为至少约100bp、200bp、300bp、400bp、500bp、600bp、700bp、800bp、900bp、1kb、2kb、3kb、4kb、5kb、6kb、7kb、8kb、9kb、10kb、20kb、30kb、40kb、50kb、60kb、70kb、80kb、90kb、100kb、200kb、300kb、400kb、500kb、600kb、700kb、800kb、900kb、1Mb、2Mb、3Mb、4Mb、5Mb、6Mb、7Mb、8Mb、9Mb或10Mb。序列读取长度可以为约100bp、200bp、300bp、400bp、500bp、600bp、700bp、800bp、900bp、1kb、2kb、3kb、4kb、5kb、6kb、7kb、8kb、9kb、10kb、20kb、30kb、40kb、50kb、60kb、70kb、80kb、90kb、100kb、200kb、300kb、400kb、500kb、600kb、700kb、800kb、900kb、1Mb、2Mb、3Mb、4Mb、5Mb、6Mb、7Mb、8Mb、9Mb或10Mb。在一些情况下，序列读取长度为至少约5kb。在一些情况下，序列读取长度为约5kb。

[0124] 在一些实例中，使用本文公开的方法和组合物生成的长的重排DNA分子在一端上与测序衔接子连接。在优选的实例中，测序衔接子为发夹衔接子，产生具有反向重复的自退火单链分子。在这些情况下，通过测序酶馈送分子，并获得反向重复序列每一侧的全长序列。在大多数情况下，得到的序列读取对应于DNA分子，如具有多个重排区段(每个区段传递相位信息)的间断DNA分子的2x覆盖度。在有利的情况下，生成足够的序列以独立地生成核酸样品的从头支架。

[0125] 或者，将使用本文公开的方法和组合物生成的长的重排DNA分子切割以形成所需长度的双链分子的群体。在这些情况下，这些分子在每一端上均与单链衔接子连接。结果是双链DNA模板的两端均覆盖有发夹环。通过连续测序技术对环状分子进行测序。含有长双链区段的分子的连续长读取测序导致每个分子的单个连续读取。含有短双链区段的分子的连续测序导致分子的多个读取，这些读取单独使用或与连续的长读取序列信息一起使用以确认分子的共有序列。在大多数情况下，鉴定由标点寡核苷酸标记的基因组区段边界，并且得出结论，与标点边界相邻的序列是同相位的。在优选的情况下，生成足够的序列以独立地生成核酸样品的从头支架。

[0126] 在一些情况下，基于长度选择重排的核酸分子以用于测序。基于长度的选择可用于选择含有更多重排区段的重排核酸分子，使得仅含有少量重排区段的较短重排核酸分子

未进行测序或以较少数目进行测序。含有更多重排区段的重排核酸分子可提供比含有较少重排区段的那些分子更多的定相信息。可以选择含有至少1、2、3、4、5、6、7、8、9、10个或更多个重排区段的重排核酸分子。例如，可以选择重排的核酸分子的长度为至少100bp、200bp、300bp、400bp、500bp、600bp、700bp、800bp、900bp、1kb、2kb、3kb、4kb、5kb、6kb、7kb、8kb、9kb、10kb、20kb、30kb、40kb、50kb、60kb、70kb、80kb、90kb、100kb、200kb、300kb、400kb、500kb、600kb、700kb、800kb、900kb、1Mb、2Mb、3Mb、4Mb、5Mb、6Mb、7Mb、8Mb、9Mb、10Mb或更长。基于长度的选择可以是严格的排除，排除100％的低于所选长度的重排核酸分子。或者，基于长度的选择可以是对更长分子的富集，去除了至少99.999％、99.99％、99.9％、99％、98％、97％、96％、95％、94％、93％、92％、91％、90％、85％、80％、75％、70％、65％、60％、55％、50％、45％、40％、35％、30％、25％、20％、15％、10％、5％、4％、3％、2％或1％的低于所选长度的重排核酸分子。核酸的长度选择可以通过多种技术进行，包括但不限于电泳(例如，凝胶或毛细管)、过滤、珠子结合(例如，SPRI珠粒大小选择)和基于流动的方法。

[0127] 定相序列组装

[0128] 在优选的实施方案中，使用本文所述的方法和组合物生成的测序数据来生成定相的从头序列组装。

[0129] 在一些实例中，如本文所公开的，生成多个重排的(和任选地间断的)DNA分子，并随后使用长读取测序技术进行测序。比较来自多个重排的(和任选地间断的)DNA分子的序列，并且在许多情况下，使用第一重排的(和任选地间断的)分子来确定其组成区段的相位信息，同时使用与第二(和另外的)重排的(和任选地间断的)DNA分子的未重排的(和任选地间断的)区域进行比较来对第一间断的分子的区段进行排序。相互地重复该过程，确定多个重排分子中的每一个中的大多数区段的相位和顺序信息。在优选的情况下，得到的组装序列是在重排发生之前输入DNA分子的定相序列，并且代表核酸样品的从头、定相的组装。

[0130] 或者，使用长读取测序技术对如使用本文公开的方法和组合物生成的重排DNA分子进行测序，并且平行地，使用标准短读取鸟枪测序技术对输入DNA进行测序。在这些情况下，将来自样品的鸟枪序列映射至由重排的DNA分子生成的长读取数据，并且/或者将来自重排的分子的定相基因组序列读取映射至由同时生成的短读取测序获得的测序数据。在一些情况下，一些短读取映射至长读取生成的序列。在这样的情况下，这种重叠允许将短序列读取分配给与由重排的DNA分子长序列读取生成的基因组序列相同的相位。

[0131] 可丢弃与生成定相序列组装无关的信息。在一个实例中，生成如本文所讨论的重排的DNA分子并对其进行测序。发现重排的DNA分子包含映射至染色体A的区段和映射至染色体B的区段。在一些情况下，可丢弃或不使用映射至染色体B的区段的序列读取信息，并且仅使用映射至染色体A的区段来生成定相序列信息。在其他情况下，可使用映射至染色体A的区段的序列读取信息来生成关于染色体A的定相序列信息，同时可使用映射至染色体B的区段的序列读取信息来生成关于染色体B的定相序列信息，但仍未使用或丢弃了关于染色体A区段与染色体B区段之间的接头的信息。

[0132] 可以操作样品以减少或去除染色体间邻近性或接头信息。例如，如本文所述，在重排和测序之前可将细胞样品在有丝分裂中冷冻，从而破坏细胞中染色体的常见三维结构。这可以减少或消除染色体间连接。在另一个实例中，可以在分析之前去除组蛋白翻译后修饰。

[0133] 核酸序列文库

[0134] 本文还公开了用于生成核酸序列文库的方法和组合物。对重排的分子进行测序，并分析序列读取。对于给定的读取，可观察到序列区段并将其解析为多个重排的区段。如果采用标点寡核苷酸，则可观察到通过标点元件局部不间断的序列区段。假定序列区段中的序列信息是同相位的，并且在局部正确地排序和定向。推断接头的任一侧上的区段在共同的样品核酸分子上彼此同相位，但不一定在重排的核酸分子上相对于彼此正确地排序和定向。重排的益处是彼此远离的区段有时会邻近，使得它们在共同的读取中进行读取并确信地分配给共同的相位，即使在样品分子中它们相隔相同的、难以将序列定相的远距离。另一个益处是区段序列本身包含大部分、基本上全部或全部原始样品序列，使得除了相位信息之外，在一些情况下，还确定叠连群信息足以在一些情况下进行从头序列组装。该从头序列任选地用于生成新的支架或叠连群组，或增加先前或独立生成的叠连群或支架序列组。

[0135] 重排的分子，如在测序文库中，可包含至少2、3、4、5、6、7、8、9、10个或更多个区段，其中该区段不与原始输入核酸分子(例如，输入基因组DNA)上的其他区段相邻。在一些情况下，给定重排分子上的至少约50％、55％、60％、65％、70％、75％、80％、85％、90％、95％、96％、97％、98％、99％、99.9％、99.99％、99.999％或100％的区段映射至共同的支架。在一些情况下，平均而言，在重排分子群体(如测序文库)上，给定重排分子上的至少约50％、55％、60％、65％、70％、75％、80％、85％、90％、95％、96％、97％、98％、99％、99.9％、99.99％、99.999％或100％的区段映射至共同的支架。

[0136] 区段的长度可以为约100bp、200bp、300bp、400bp、500bp、600bp、700bp、800bp、900bp、1kb、1.1kb、1.2kb、1.3kb、1.4kb、1.5kb、2.0kb、2.5kb、3.0kb、3.5kb、4.0kb、4.5kb、5.0kb、5.5kb、6.0kb、6.5kb、7.0kb、7.5kb、8.0kb、8.5kb、9.0kb、9.5kb、10.0kb或更长。区段的长度可以为至少约100bp、200bp、300bp、400bp、500bp、600bp、700bp、800bp、900bp、1kb、1.1kb、1.2kb、1.3kb、1.4kb、1.5kb、2.0kb、2.5kb、3.0kb、3.5kb、4.0kb、4.5kb、5.0kb、5.5kb、6.0kb、6.5kb、7.0kb、7.5kb、8.0kb、8.5kb、9.0kb、9.5kb、10.0kb或更长。区段的长度可以为至多约100bp、200bp、300bp、400bp、500bp、600bp、700bp、800bp、900bp、1kb、1.1kb、1.2kb、1.3kb、1.4kb、1.5kb、2.0kb、2.5kb、3.0kb、3.5kb、4.0kb、4.5kb、5.0kb、5.5kb、6.0kb、6.5kb、7.0kb、7.5kb、8.0kb、8.5kb、9.0kb、9.5kb、10.0kb或更长。

[0137] 重排的分子可具有至少2、3、4、5、6、7、8、9、10个或更多个长度为至少约100bp、200bp、300bp、400bp、500bp、600bp、700bp、800bp、900bp、1kb、1.1kb、1.2kb、1.3kb、1.4kb、1.5kb、2.0kb、2.5kb、3.0kb、3.5kb、4.0kb、4.5kb、5.0kb、5.5kb、6.0kb、6.5kb、7.0kb、7.5kb、8.0kb、8.5kb、9.0kb、9.5kb、10.0kb或更长的区段。在一些情况下，重排的分子具有至少3个长度为至少500bp的区段。在一些情况下，重排的分子具有至少4个长度为至少500bp的区段。在一些情况下，重排的分子具有至少5个长度为至少500bp的区段。在一些情况下，重排的分子具有至少6个长度为至少500bp的区段。

[0138] 当在重排的分子中的所有区段上相加时，重排的分子可包含来自一个原始核酸分子(例如，来自一个染色体)的至少100bp、200bp、300bp、400bp、500bp、600bp、700bp、800bp、900bp、1kb、1.1kb、1.2kb、1.3kb、1.4kb、1.5kb、2.0kb、2.5kb、3.0kb、3.5kb、4.0kb、4.5kb、5.0kb、5.5kb、6.0kb、6.5kb、7.0kb、7.5kb、8.0kb、8.5kb、9.0kb、9.5kb、10.0kb。在一些情况下，当在重排的分子中的所有区段上相加时，重排的分子包含来自一个原始核酸分子(例

如,来自一个染色体)的至少1000bp。在一些情况下,当在重排的分子中的所有区段上相加时,重排的分子包含来自一个原始核酸分子(例如,来自一个染色体)的至少2000bp。在一些情况下,当在重排的分子中的所有区段上相加时,重排的分子包含来自一个原始核酸分子(例如,来自一个染色体)的至少3000bp。在一些情况下,当在重排的分子中的所有区段上相加时,重排的分子包含来自一个原始核酸分子(例如,来自一个染色体)的至少4000bp。在一些情况下,当在重排的分子中的所有区段上相加时,重排的分子包含来自一个原始核酸分子(例如,来自一个染色体)的至少5000bp。

[0139] 在一些情况下,可以采用强制唯一映射来进行映射。在一些情况下,小于约50%、45%、40%、35%、30%、25%、20%、15%、10%、5%、4%、3%、2%、1%、0.1%、0.01%或0.001%的区段模糊地映射(例如,映射至多个位置)。

[0140] 测序文库可包含至少约10、100、1000、10,000、100,000、100万、110万、120万、130万、140万、150万、160万、170万、180万、190万、200万、300万、400万、500万、600万、700万、800万、900万、1000万、2000万、3000万、4000万、5000万、6000万、7000万、8000万、9000万、1亿、2亿、3亿、4亿、5亿、6亿、7亿、8亿、9亿、10亿、20亿、30亿、40亿、50亿、60亿、70亿、80亿、90亿、100亿、1000亿、2000亿、3000亿、4000亿、5000亿、6000亿、7000亿、8000亿、9000亿或1万亿个重排的分子。

[0141] 测序文库中的重排的分子可包含必需的衔接子、标记物或用于测序的其他组分,如特定识别序列、杂交序列、发夹(例如,用于SMRTbell)、标签(例如,NanoTag)、标记物、染料或条码。

[0142] 在一些情况下,如本文所公开生成多个重排的DNA分子,并随后使用长读取测序技术进行测序。对每个重排的分子进行测序,并分析序列读取。在优选的实例中,对于序列反应而言,序列读取平均至少约5kb或至少约10kb。在其他实例中,序列读取平均至少约5kb、6kb、7kb、8kb、9kb、10kb、11kb、12kb、13kb、14kb、15kb、16kb、17kb、18kb、19kb、20kb、21kb、22kb、25kb、30kb、35kb、40kb或更大。在有利的实例中,鉴定包含第一区段的至少500个碱基和第二区段的500个碱基的序列读取,其中所述第一区段和第二区段在原始样品输入核酸上不相邻。区段可通过标点寡核苷酸序列进行连接。在其他实例中,序列读取包含第一DNA区段的至少约100个碱基、200个碱基、300个碱基、400个碱基、500个碱基、600个碱基、700个碱基、800个碱基、900个碱基、1000个碱基或更多个碱基和第二DNA区段的至少约100个碱基、200个碱基、300个碱基、400个碱基、500个碱基、600个碱基、700个碱基、800个碱基、900个碱基、1000个碱基或更多个碱基。在一些实例中,将所述第一和第二区段序列映射至支架基因组,并且发现其映射至相隔至少100kb的叠连群。在其他实例中,分隔距离为至少约8kb、9kb、10kb、12.5kb、15kb、17.5kb、20kb、25kb、30kb、35kb、40kb、45kb、50kb、60kb、70kb、80kb、90kb、100kb、125kb、150kb、200kb、300kb、400kb、500kb、600kb、700kb、800kb、900kb、1Mb或更大。在大多数情况下,第一叠连群和第二叠连群各自包含单个杂合位置,该杂合位置的相位未在支架中确定。在优选的实例中,第一叠连群的杂合位置被长读取的第一区段跨越,并且第二叠连群的杂合位置被长读取的第二区段跨越。在这样的情况下,这些读取各自跨越它们的叠连群的各自的杂合区域,并且读取区段的序列指示第一叠连群的第一等位基因和第二叠连群的第一等位基因是同相位的。如果在单个长序列读取中检测到来自第一核酸区段和第二核酸区段的序列,则确定第一核酸区段和第二核酸区段包含在输入DNA样

品中的相同DNA分子上。在这些优选的实施方案中，通过本文公开的方法和组合物生成的核酸序列文库提供了在基因组支架上彼此远离的叠连群的相位信息。

[0143] 或者，如本文所述生成多个成对末端分子，并随后使用长读取测序技术进行测序。在一些情况下，文库的平均读取长度被确定为约1kb。在其他情况下，文库的平均读取长度为约100bp、200bp、300bp、400bp、500bp、600bp、700bp、800bp、900bp、1kb、1.1kb、1.2kb、1.3kb、1.4kb、1.5kb、2.0kb、2.5kb、3.0kb、3.5kb、4.0kb、4.5kb、5.0kb、5.5kb、6.0kb、6.5kb、7.0kb、7.5kb、8.0kb、8.5kb、9.0kb、9.5kb、10.0kb或更长。在许多实例中，成对末端分子包含第一DNA区段和第二DNA区段，该第一DNA区段和第二DNA区段在输入DNA样品内是同相位的，并且相隔大于10kb的距离。在一些实例中，两个这样的DNA区段之间的分隔距离大于约5kb、6kb、7kb、8kb、9kb、10kb、11kb、12kb、13kb、14kb、15kb、20kb、23kb、25kb、30kb、32kb、35kb、40kb、50kb、60kb、75kb、100kb、200kb、300kb、400kb、500kb、750kb、1Mb或更大。在大多数情况下，序列读取由成对末端分子生成，其中一些序列读取包含来自第一核酸区段的至少300个碱基的序列和来自第二核酸区段的至少300个碱基的序列。在其他实例中，序列读取包含第一DNA区段的至少约50个碱基、100个碱基、150个碱基、200个碱基、250个碱基、300个碱基、350个碱基、400个碱基、450个碱基、500个碱基、550个碱基、600个碱基、650个碱基、700个碱基、750个碱基、800个碱基或更多个碱基和第二DNA区段的至少约50个碱基、100个碱基、150个碱基、200个碱基、250个碱基、300个碱基、350个碱基、400个碱基、450个碱基、500个碱基、550个碱基、600个碱基、650个碱基、700个碱基、750个碱基、800个碱基或更多个碱基。如果在单个序列读取中检测到来自第一核酸区段和第二核酸区段的序列，则可以确定第一核酸区段和第二核酸区段在输入DNA样品中的同一DNA分子上是同相位的。在这样的情况下，生成的序列文库产生DNA区段的相位信息，这些DNA区段在核酸样品中相隔大于用于对它们进行测序的测序技术的读取长度。

[0144] 或者，如本文所公开的，从重排的DNA核酸序列文库生成多个序列读取。在一些情况下，如本文所公开和如所提供的实施例中所述，所述文库传递相位信息，使得在区段接头的任一侧上的区段被确定为在单个分子上是同相位的。在一些实例中，生成的序列读取代表输入DNA样品的至少80％的核酸序列。在其他实例中，生成的序列读取代表输入DNA样品的至少约45％、50％、55％、60％、65％、70％、75％、80％、85％、90％、95％或100％的核酸序列。在优选的实例中，序列读取用于生成跨越至少80％的输入DNA样品的从头叠连群信息。在其他实例中，序列读取用于生成跨越至少约45％、50％、55％、60％、65％、70％、75％、80％、85％、90％、95％或100％的输入DNA样品的从头叠连群信息。在大多数情况下，序列读取用于确定相位信息，其随后任选地用于将叠连群相对于彼此进行排序和定向，以生成输入DNA样品的定相序列组装。在优选的实施方案中，由重排的DNA分子生成的核酸序列文库传递相位信息，并且优选地还包括包含总核酸序列的大部分的序列信息，使得同时生成从头序列组装。

[0145] 可以进行重排分子的文库的测序以实现至少约1X、2X、3X、4X、5X、6X、7X、8X、9X、10X、11X、12X、13X、14X、15X、16X、17X、18X、19X、20X、21X、22X、23X、24X、25X、26X、27X、28X、29X、30X、31X、32X、33X、34X、35X、336X、37X、38X、39X、40X、41X、42X、43X、44X、45X、46X、47X、48X、49X、50X、55X、60X、65X、70X、75X、80X、85X、90X、95X、100X或更多倍的测序覆盖度。

[0146] 保留的DNA分子定相

[0147] 此外，本文公开了用于对核酸序列进行定相和从头组装的方法和组合物，在优选的实施方案中，所述核酸序列包含几乎整个输入核酸分子。

[0148] 本公开内容的技术可用于对多种标志物进行定相，该标志物包括但不限于单核苷酸多态性(SNP)、插入或缺失(INDEL)和结构变体(SV)。例如，两个或更多个区段在重排的DNA分子上一起存在可用于推断区段的序列是同相位的。这可以允许定相而不依赖于先前已知的标志物定相。在一些情况下，对SNP进行定相。在一些情况下，对INDEL进行定相。在一些情况下，对SV进行定相。可以参考一种或多种标志物来确认定相。在一些情况下，参考SNP确认定相。在一些情况下，不参考SNP确认定相。在一些情况下，参考INDEL确认定相。在一些情况下，不参考INDEL确认定相。在一些情况下，参考SV确认定相。在一些情况下，不参考SV确认定相。在一些实例中，使用本领域已知的标准方法提取高分子量(HMW)核酸样品。在大多数情况下，这些HMW核酸样品包含至少一些长度为至少100kb的核酸分子。一个或多个的100kb核酸分子包含相隔大于标准测序技术的平均读取长度的距离的第一核酸区段和第二核酸区段。在其他实例中，核酸样品包含至少一些核酸分子，该核酸分子的长度为至少约30kb、40kb、50kb、60kb、70kb、80kb、90kb、100kb、110kb、120kb、130kb、140kb、150kb或更长，其中一个或多个核酸分子包含相隔大于标准测序技术(如本文所述的技术)的平均读取长度的距离的至少第一核酸区段和第二核酸区段。

[0149] 为了确定相位信息是这样的实例，需要在单个测序读取内检测第一核酸区段和第二核酸区段。因此，必须改变第一核酸区段和第二核酸区段的相对位置，使得第一DNA区段和第二DNA区段相隔小于标准测序技术的平均读取长度的距离。为了生成所需的相位信息，这种重排不应导致相位信息的丢失。在优选的实例中，这种重排是通过本文公开的方法和组合物以及如所提供的实施例中所述实现的。在有利的实例中，在相位保持重排期间，不多于10％的起始核酸分子缺失。也就是说，仅通过使插入序列缺失就不会使第一区段和第二区段邻近。相反，这些区段相对于彼此进行重排而不会使大部分插入序列缺失。在其他实例中，不多于约2％、5％、7％、10％、12％、13％、14％、15％、20％、23％、25％、30％、35％、40％、50％、55％、60％、70％、80％、90％或95％的起始核酸分子缺失。因为，在有利的实例中，保留了几乎整个输入核酸分子，所以在测序后，使用所生成的序列读取对从头生成的叠连群进行组装、排序和定向，使得几乎整个输入核酸分子得以测序、组装并定相。

[0150] 应用

[0151] 本公开内容的技术可用于多种遗传学和基因组学应用，包括但不限于生成从头序列组装(包括定相序列组装)、将读取映射至支架(包括具有定相信息)、确定定相信息以及鉴定结构变体。

[0152] 本文公开的技术可用于许多领域，包括(作为非限制性实例)，法医学、农业、环境研究、可再生能源、流行病学或疾病爆发应对和物种保存。

[0153] 本公开内容的技术可用于诊断疾病状态，如癌症。本公开内容的技术可用于临床重要区域的定相、结构变体的分析、假基因的解析(例如，STRC)、癌症中可药用结构变体的靶向板和其他应用。例如，在线性相距甚远的基因组区域之间或在单独的染色体上，过多的邻近连接事件可以指示像癌症这样的疾病。

[0154] 可以使用本公开内容的技术分析来自患病或怀疑患病的组织的天然染色质。可以分析在这样的组织样品内基因组的三维结构，例如通过分析来自组织体积内不同位置的几

个样品。

[0155] 在一些情况下，如对于从头基因组组装，可以从这些数据中去除生物或病理信号。例如，在添加在邻近连接之前锁定在三维结构中的固定剂之前，可以用引起有丝分裂阻滞的试剂或破坏异染色质或基因组结构的其他区域特征的试剂处理细胞。在这样的情况下，得到的数据可能缺乏诊断效用，但对于基因组组装可能是最有用的。

[0156] 如本文所公开生成的分子和文库用于许多应用，如与基因组组装和叠连群或其他序列信息定相相关的应用，如用于将序列信息分配给来自二倍体生物体的基因组组装的特定起源分子或姐妹染色单体。

[0157] 对分子进行测序，并且将连续区段鉴定为映射至共同的叠连群或支架的连续碱基。将区段之间的接头鉴定为碱基停止映射至共同的叠连群或支架的连续碱基的区域。在一些情况下，丢弃映射至基因组的多个区域的核酸序列(如重复序列)。或者，特别是如果重复序列的一端或两端映射至共同的支架并且重复序列末端处的唯一映射序列的序列位置之间的差异与支架中包含的重复区域一致，则将重复区域分配给具有其相邻唯一序列的共同区段。

[0158] 在优选的实施方案中，将如本文公开的分子或文库组分的相邻区段分配给基因组的共同分子的共同相位。也就是说，将区段映射的叠连群分配给共同相位，并且将单核苷酸多态性、插入、缺失、颠换、易位或由一个或两个区段指示的其他核酸特征分配给共同的分子。

[0159] 通常，所有或大多数区段映射至共同的支架或叠连群，使得它们在文库的单个分子上的共存指示单核苷酸多态性、插入、缺失、颠换、易位或由一个或两个区段指示的其他核酸特征被分配给共同的分子。在一些情况下，至少50％、55％、60％、65％、70％、75％、80％、85％、90％、91％、92％、93％、94％、95％、96％、97％、98％、99％或多于99％的区段，或至少50％、55％、60％、65％、70％、75％、80％、85％、90％、91％、92％、93％、94％、95％、96％、97％、98％、99％或多于99％的区段序列映射至共同的支架。

[0160] 在一些情况下，富集分子生成以确保或增加区段连接的可能性，从而反映物理连锁信息或相位信息或者使得连接区段由共同的物理起源分子产生是有益的。许多方法均能实现这一目标。

[0161] 如本文所讨论的，在一些情况下，通过切割和重新连接分离的核酸分子生成文库，该核酸分子上已经组装了染色质或其他核酸结合部分。通过分离分子，例如通过将它们与核酸结合蛋白或其他天然染色质成分分离，允许单个分子彼此分离。通过结合分离的核酸分子使得分离的区段独立于它们共同的磷酸二酯骨架而保持在一起，在切割和降级过程中保留了起源分子的区段共有的相位信息，使得切割的区段可能与来源于两个区段共有的起源分子的第二区段重新连接。通过任何数目的方法增加该频率，例如通过在染色质组装之前稀释分子或通过将核酸分子以低于在该区段处形成单独分子可能连接的密度附接至共同表面上的不同位置。当使用珠子如SPRI珠子来锚定分子以用于消化和组装时，选择具有更大表面积的珠子，或添加更多珠子以增加可用于结合的总表面积，在一些情况下降低分子间连接事件的机会。

[0162] 或者，在一些情况下，采取步骤以减少由天然染色质结合的核酸分子之间的分子间相互作用，如当使用固定剂处理细胞时发生。此类步骤的实例包括在细胞周期中的一点

处主动靶向细胞，使得分子间相互作用可能最小化。在一些情况下，这通过在有丝分裂中冷冻或固定细胞来实现，以在染色体不太可能组装成可能导致分子间连接事件的亚核结构时选择性地接近该细胞的核酸。备选地或组合地，处理细胞、细胞核或来自细胞的分离的染色质，以去除组蛋白翻译后修饰，从而去除三维映射信息并同时提高在文库生成中来自单个分子的区段彼此相互连接的机会以用于测序/定相信息。

[0163] 除了在重排的文库形成中减少分子间连接事件的生物化学或“湿实验室”方法之外，还可使用计算方法来减少分子间连接事件对相位确定的影响。例如，在一些情况下，通过评估连接的重排分子中唯一映射区段的映射分布来筛选单个分子。排除包含映射至阈值水平以上的可能不同分子的区段的分子。也就是说，在一些情况下，从进一步分析中排除包含区段的分子的序列信息，该区段以小于50％、55％、60％、65％、70％、75％、80％、85％、90％、95％或最高99％或更高的百分比唯一映射至共同的支架。在示例性情况下，该阈值为70％或约70％、或80％或约80％、或90％或约90％。在这些情况下，从分析中排除包含映射至除第一共同的支架之外的其他地方的一定百分比的区段的分子序列。

[0164] 类似地，在一些情况下，从进一步分析中排除包含聚集唯一映射序列的分子的序列信息，该唯一映射序列以小于50％、55％、60％、65％、70％、75％、80％、85％、90％、95％或最高99％或更高的百分比映射至共同的支架。在示例性情况下，该阈值为70％或约70％、或80％或约80％、或90％或约90％。在这些情况下，从分析中排除包含映射至除第一共同的支架之外的其他地方的一定百分比的唯一映射序列的分子序列。

[0165] 备选地或组合地，进一步处理包含唯一映射至多于一个支架的区段的分子序列，以最小化对相位结论的影响而不丢失序列信息，如SNP数据、插入数据、缺失数据、倒位数据或可在测序的区段中捕获的其他基因组重排信息。例如，对于包含唯一映射至两个支架(主要或排他)的区段的分子序列，映射至第一支架的区段被分配给该支架的共同相位，而映射至第二支架的区段被分配给第二支架的共同相位。也就是说，映射至第一共同的支架的区段被分配给该支架上的共同相位，而映射至第二共同的支架的区段被确定为提供第二支架的共同相位信息的信息，但是映射(如唯一映射)至第一支架的区段未被确定为提供关于映射至第二支架的区段的相位信息的信息。

[0166] 或者，在一些情况下，获得多个独立的分子序列，该分子序列包含唯一映射至第一支架的第一区段群和唯一映射至第二支架的第二区段群。在这些情况下，任选地推断第一支架和第二支架实际上在核酸样品中同相位，例如由于在分析中的样品基因组中的易位。

[0167] 这些方法允许选择性富集分子序列，该分子序列可能提供关于重排的文库和重排的文库的序列数据所源自的潜在分子的相位的信息。

[0168] 在一些情况下，组合使用文库生成和序列分析以获得序列信息和相位信息。在一些这样的情况下，对连接接头进行标记，例如使用与长读取测序技术相容并且在这种技术的读取中容易鉴定的经修饰的核苷酸碱基。本文提供了实例。

[0169] 使用这样的接头标志物，能够独立于区段序列以高置信度鉴定区段接头。因此，文库构建中的序列重排容易地与样品中发生的“重排事件”区分开，并且反映了样品核酸序列或结构。此类事件包括例如插入、缺失、倒位、颠换或易位。当此类事件未经接头标志物如修饰的核酸标记时，观察区段中的此类事件指示该事件反映了潜在的样品序列。

[0170] 备选地或组合地，可以依赖于文库覆盖的深度来提供关于分子结构的某种程度的

置信度。也就是说，在对多个独立生成的文库组分进行测序时，可以发现共享共同的重排特征的多个独立生成的区段。如果这样的特征包含多个独立衍生的文库组分中的共同“重排事件”，则可以推断它们指示的“重排事件”反映了潜在的样品序列而不是文库生成过程的产物。

[0171] 很多种文库组分与本文的公开内容一致。文库组分优选平均长于普遍的长读取测序技术的单个读取，使得测序技术在对文库测序中最有效地使用。然而，这不是绝对要求，并且包含、主要包含或由小于长范围测序运行的长度的成分组成的文库与本文的公开内容一致。

[0172] 本文公开的文库可以在文库中表示的总样品的分数、平均或中位重排分子大小、区段大小和每分子的区段数方面变化。在许多实施方案中，文库被配置成使得单个长读取跨越文库的分子成分的三个区段的至少一部分。在许多实施方案中，文库被配置成使得同相位但分散在整个基因组样品中的区段得以重新配置，以便它们相邻或以其他方式包含在单个长范围序列读取中，从而促进将它们分配给共同分子的共同相位。

[0173] 计算机系统及其操作的改进

[0174] 在一些情况下，通过存储在服务器1001的电子存储位置上(例如在存储器1010或电子存储单元1015上)的机器(或计算机处理器)可执行代码(或软件)来实现如本文所述的方法。在使用期间，该代码可以由处理器1005执行。在一些情况下，可以从存储单元1015检索该代码并将其存储在存储器1010上以备处理器1005访问。在一些情况下，可以排除电子存储单元115，并且机器可执行指令存储在存储器1010上。或者，该代码可以在第二计算机系统1040上执行。

[0175] 本文提供的系统和方法的各方面，如服务器1001，可体现在编程中。本技术的多个方面可以被认为是“产品”或“制品”，其通常为某种类型的机器可读介质上携带或体现的机器(或处理器)可执行代码和/或相关数据的形式。机器可执行代码可存储在诸如存储器(例如，只读存储器、随机存取存储器、闪速存储器)的电子存储单元或硬盘上。“存储”型介质可包括计算机的任何或全部有形存储器、处理器等，或其相关模块，如各种半导体存储器、磁带驱动器、磁盘驱动器等，其可在任何时候为软件编程提供非暂时性存储。该软件的全部或部分有时可以通过因特网或各种其他电信网络进行通信。这样的通信，例如，可使软件能够从一个计算机或处理器加载到另一个计算机或处理器，例如，从管理服务器或主机加载到应用服务器的计算机平台中。因此，可承载软件元素的另一类型的介质包括光波、电波和电磁波，如跨本地设备之间的物理接口、通过有线和光学陆线网络以及通过各种空中链路而使用的。携带这类波的物理元件，如有线或无线链路、光学链路等，也可以被认为是承载软件的介质。如本文所用的，除非受限于非暂时性有形“存储”介质，否则诸如计算机或机器“可读介质”的术语可以指参与向处理器提供指令以供执行的任何介质。

[0176] 因此，机器可读介质如计算机可执行代码可以采用许多形式，包括但不限于有形存储介质、载波介质或物理传输介质。非易失性存储介质可包括例如光盘或磁盘，如任何计算机中的任何存储设备等，例如可用于实现该系统。有形传输介质可以包括：同轴电缆、铜线和光纤(包括电线，该电线包括计算机系统内的总线)。载波传输介质可采用电信号或电磁信号或者声波或光波的形式，如在射频(RF)和红外(IR)数据通信过程中生成的那些电信号或电磁信号或者声波或者光波。因此，计算机可读介质的常见形式包括例如：软盘、柔性

盘、硬盘、磁带、任何其他磁性介质、CD-ROM、DVD、DVD-ROM、任何其他光学介质、穿孔卡片纸带、任何其他具有孔洞图案的物理存储介质、RAM、ROM、PROM和EPROM、FLASH-EPROM、任何其他存储器芯片或匣盒、传送数据或指令的载波、传送这样的载波的电缆或链路，或者计算机可从中读取编程代码和/或数据的任何其他介质。这些计算机可读介质形式中的许多形式可以参与将一个或多个指令的一个或多个序列传送至处理器以供执行。

[0177] 计算机系统可用于实现本文所述方法的一个或多个步骤，包括例如样品收集、样品处理、序列生成和序列分析。

[0178] 客户端-服务器和/或关系数据库架构可用于本文所述的任何方法中。通常，客户端-服务器架构为网络架构，其中网络上的每个计算机或进程是客户端或服务器。服务器计算机可以是专用于管理磁盘驱动器(文件服务器)、打印机(打印服务器)或网络流量(网络服务器)的功能强大的计算机。客户端计算机可包括用户在其上运行应用的PC(个人计算机)或工作站，以及如本文所公开的示例性输出设备。客户端计算机可依赖于服务器计算机来获取资源，如文件、设备以及甚至处理能力。服务器计算机处理所有数据库功能。客户端计算机可具有处理前端数据管理和接收来自用户的数据输入的软件。

[0179] 在执行计算之后，处理器可将输出(如来自计算)返回至例如输入设备或存储单元、相同或不同计算机系统的另一存储单元或输出设备。来自处理器的输出可以通过数据显示器显示，该数据显示器例如显示屏(例如，监视器或数字设备上的屏幕)、打印输出、数据信号(例如，数据包)、图形用户界面(例如，网页)、警报(例如，闪光或声音)，或上述的任何组合。在一个实施方案中，输出通过网络(例如，无线网络)传送至输出设备。用户可以使用输出设备来接收来自数据处理计算机系统的输出。在用户接收到输出之后，用户可以确定行动过程，或者可以执行行动过程，如当用户是医务人员时的医疗处理。在一些实施方案中，输出设备是与输入设备相同的设备。示例性输出设备包括但不限于电话、无线电话、移动电话、PDA、闪存驱动器、光源、声音发生器、传真机、计算机、计算机显示器、打印机、iPod和网页。用户站可以与打印机或显示监视器通信，以输出由服务器处理的信息。这样的显示器、输出设备和用户站可用于向受试者或其护理人员提供警报。

[0180] 与本公开内容有关的数据可以通过网络或连接进行传送以供接收者接收和/或查看。接收者可以是但不限于报告所涉及的受试者；或其护理人员，例如，医疗保健提供者、经理、其他医疗保健专业人员或其他看护人；执行和/或订购基因分型分析的个人或实体；遗传咨询师。接收者也可以是用于存储这类报告的本地系统或远程系统(例如服务器或“云计算”架构的其他系统)。在一个实施方案中，计算机可读介质包括适用于传送生物样品分析结果的介质。

[0181] 如本文所公开的数据集和序列文库与核酸序列信息(如通过杂合二倍体真核基因组的测序获得的核酸序列信息)的基于计算机的相位分配一致。分析此类数据的计算机可以将读取分配至支架中，在一些情况下生成包含样品基因组的整个“端到端”染色体图谱的图谱。然而，当所述杂合序列相隔大于测序技术的读取长度时，大多数方法不能将杂合序列分配至共同相位。因此，使用大多数基于计算机的基因组组装方法，不能使杂合基因座准确地映射至共同相位。

[0182] 本文公开的方法、数据库和系统允许将杂合序列信息分配至共同相位，即使当杂合基因座相隔大于单个长读取生成的序列距离时。因此，本文公开的方法、数据库和系统提

供了与基因组测序和基因组序列组装相关的计算机系统的性能改进。例如，本公开内容的技术可以允许提高计算速度，从而减少计算时间或计算负担。这些技术还可以允许减少存储器需求，包括暂时性存储器和非暂时性数据存储需求。在一些情况下，本公开内容的技术可以使得能够计算先前不可计算的计算。

[0183] 参考以下编号实施方案进一步补充详细描述。1.一种从第一DNA分子生成长距离相位信息的方法，该方法包括：a)提供具有第一区段和第二区段的第一DNA分子，其中所述第一区段和所述第二区段在所述第一DNA分子上不相邻；b)使所述第一DNA分子与DNA结合部分接触，使得所述第一区段和所述第二区段独立于所述第一DNA分子的共同的磷酸二酯骨架而与所述DNA结合部分结合；c)切割所述第一DNA分子，使得所述第一区段和所述第二区段不由共同的磷酸二酯骨架连接；d)通过磷酸二酯键将所述第一区段与所述第二区段附接以形成重新组装的第一DNA分子；以及e)对单个测序读取中包含所述第一区段与所述第二区段之间的接头的所述重新组装的第一DNA分子的至少4kb的连续序列进行测序，其中第一区段序列和第二区段序列代表来自第一DNA分子的长距离相位信息。2.根据编号实施方案1所述的方法，其中所述DNA结合部分包含多个DNA结合分子。3.根据编号实施方案1-2中任一实施方案所述的方法，其中使所述第一DNA分子与多个DNA结合分子接触包括与DNA结合蛋白的群体接触。4.根据编号实施方案1-3中任一实施方案所述的方法，其中所述DNA结合蛋白的群体包含核蛋白。5.根据编号实施方案1-4中任一实施方案所述的方法，其中所述DNA结合蛋白的群体包含核小体。6.根据编号实施方案1-5中任一实施方案所述的方法，其中所述DNA结合蛋白的群体包含组蛋白。7.根据编号实施方案1-6中任一实施方案所述的方法，其中使所述第一DNA分子与多个DNA结合部分接触包括与DNA结合纳米颗粒的群体接触。8.根据编号实施方案1-7中任一实施方案所述的方法，其中所述第一DNA分子具有在所述第一DNA分子上与所述第一区段或所述第二区段不相邻的第三区段，其中进行(b)中的所述接触，使得所述第三区段独立于所述第一DNA分子的共同的磷酸二酯骨架而与所述DNA结合部分结合，其中进行(c)中的所述切割，使得所述第三区段不通过共同的磷酸二酯骨架与所述第一区段和所述第二区段连接，其中所述附接包括通过磷酸二酯键将所述第三区段与所述第二区段附接以形成所述重新组装的第一DNA分子，并且其中(e)中测序的连续序列包含单个测序读取中所述第二区段与所述第三区段之间的接头。9.根据编号实施方案1-9中任一实施方案所述的方法，其包括使所述第一DNA分子与交联剂接触。10.根据编号实施方案1-9中任一实施方案所述的方法，其包括使所述第一DNA分子与交联剂接触。11.根据编号实施方案1-10中任一实施方案所述的方法，其中所述交联剂为甲醛。12.根据编号实施方案1-11中任一实施方案所述的方法，其中所述交联剂为甲醛。13.根据编号实施方案1-12中任一实施方案所述的方法，其中所述DNA结合部分与包含多个DNA结合部分的表面结合。14.根据编号实施方案1-13中任一实施方案所述的方法，其中所述DNA结合部分与包含珠子的固体框架结合。15.根据编号实施方案1-14中任一实施方案所述的方法，其中切割所述第一DNA分子包括与限制性内切核酸酶接触。16.根据编号实施方案1-15中任一实施方案所述的方法，其中切割所述第一DNA分子包括与非特异性内切核酸酶接触。17.根据编号实施方案1-16中任一实施方案所述的方法，其中切割所述第一DNA分子包括与标签化酶接触。18.根据编号实施方案1-17中任一实施方案所述的方法，其中切割所述第一DNA分子包括与转座酶接触。19.根据编号实施方案1-18中任一实施方案所述的方法，其中切割所述第一DNA分子包括剪

切所述第一分子。20.根据编号实施方案1-19中任一实施方案所述的方法，其包括将标签添加至至少一个暴露的末端。21.根据编号实施方案1-20中任一实施方案所述的方法，其中所述标签包含标记的碱基。22.根据编号实施方案1-21中任一实施方案所述的方法，其中所述标签包含甲基化的碱基。23.根据编号实施方案1-22中任一实施方案所述的方法，其中所述标签包含生物素化的碱基。24.根据编号实施方案1-23中任一实施方案所述的方法，其中所述标签包含尿苷。25.根据编号实施方案1-24中任一实施方案所述的方法，其中所述标签包含非规范碱基。26.根据编号实施方案1-25中任一实施方案所述的方法，其中所述标签生成平端的暴露的末端。27.根据编号实施方案1-26中任一实施方案所述的方法，其包括将至少一个碱基添加至第一区段粘性末端的嵌入链。28.根据编号实施方案1-27中任一实施方案所述的方法，其包括添加包含与所述第一区段粘性末端退火的突出端的连接体寡核苷酸。29.根据编号实施方案1-28中任一实施方案所述的方法，其中所述连接体寡核苷酸包含与所述第一区段粘性末端退火的突出端和与第二区段粘性末端退火的突出端。30.根据编号实施方案1-29中任一实施方案所述的方法，其中所述连接体寡核苷酸不包含两个5'磷酸部分。31.根据编号实施方案1-30中任一实施方案所述的方法，其中附接包括连接。32.根据编号实施方案1-31中任一实施方案所述的方法，其中附接包括DNA单链切口修复。33.根据编号实施方案1-32中任一实施方案所述的方法，其中在切割所述第一DNA分子之前，所述第一区段和所述第二区段在所述第一DNA分子上相隔至少10kb。34.根据编号实施方案1-33中任一实施方案所述的方法，其中在切割所述第一DNA分子之前，所述第一区段和所述第二区段在所述第一DNA分子上相隔至少15kb。35.根据编号实施方案1-34中任一实施方案所述的方法，其中在切割所述第一DNA分子之前，所述第一区段和所述第二区段在所述第一DNA分子上相隔至少30kb。36.根据编号实施方案1-35中任一实施方案所述的方法，其中在切割所述第一DNA分子之前，所述第一区段和所述第二区段在所述第一DNA分子上相隔至少50kb。37.根据编号实施方案1-36中任一实施方案所述的方法，其中在切割所述第一DNA分子之前，所述第一区段和所述第二区段在所述第一DNA分子上相隔至少100kb。38.根据编号实施方案1-37中任一实施方案所述的方法，其中所述测序包括单分子长读取测序。39.根据编号实施方案1-38中任一实施方案所述的方法，其中所述长读取测序包括至少5kb的读取。40.根据编号实施方案1-39中任一实施方案所述的方法，其中所述长读取测序包括至少10kb的读取。41.根据编号实施方案1-40中任一实施方案所述的方法，其中所述第一重新组装的DNA分子包含在所述第一DNA分子的一端连接5'端与3'端的发夹部分。42.根据编号实施方案1-41中任一实施方案所述的方法，其包括对所述第一DNA分子的第二重新组装形式进行测序。43.根据编号实施方案1-42中任一实施方案所述的方法，其中所述第一区段和所述第二区段各自为至少500bp。44.根据编号实施方案1-43中任一实施方案所述的方法，其中所述第一区段、所述第二区段和所述第三区段各自为至少500bp。45.一种基因组组装的方法，该方法包括：a)获得与结构复合的第一DNA分子；b)切割所述第一DNA分子以形成第一暴露的末端和第二暴露的末端，其中在所述切割之前，所述第一暴露的末端和所述第二暴露的末端在所述分子上不相邻；c)切割所述第一DNA分子以形成第三暴露的末端和第四暴露的末端，其中在所述切割之前，所述第三暴露的末端和所述第四暴露的末端在所述分子上不相邻；d)将所述第一暴露的末端和所述第二暴露的末端附接以形成第一接头；e)将所述第三暴露的末端和所述第四暴露的末端附接以形成第二接头；f)在单个测序读取中跨越所述第一接

头和所述第二接头进行测序;g)将所述第一接头的第一侧上的序列映射至所述多个叠连群的第一叠连群;h)将所述第一接头的第二侧上的序列映射至所述多个叠连群的第二叠连群;i)将所述第二接头的第一侧上的序列映射至所述多个叠连群的第二叠连群;j)将所述第二接头的第二侧上的序列映射至所述多个叠连群的第三叠连群;以及k)将所述第一叠连群、所述第二叠连群和所述第三叠连群分配给基因组组装的共同相位。46.根据编号实施方案45所述的方法,其中所述多个叠连群由鸟枪序列数据生成。47.根据编号实施方案45-46中任一实施方案所述的方法,其中所述多个叠连群由单分子长读取数据生成。48.根据编号实施方案45-47中任一实施方案所述的方法,其中所述单分子长读取数据包含所述多个叠连群。49.根据编号实施方案45-48中任一实施方案所述的方法,其中所述多个叠连群通过在所述第一接头和所述第二接头进行测序而同时获得。50.根据编号实施方案45-49中任一实施方案所述的方法,其中在所述标志物寡核苷酸进行测序包括对至少10kb进行测序。51.根据编号实施方案45-50中任一实施方案所述的方法,其中所述结构包含与所述第一DNA分子结合以形成重构的染色质的DNA结合部分的群体。52.根据编号实施方案45-51中任一实施方案所述的方法,其中使所述重构的染色质与交联剂接触。53.根据编号实施方案45-52中任一实施方案所述的方法,其中所述交联剂包含甲醛。54.根据编号实施方案45-53中任一实施方案所述的方法,其中所述DNA结合部分的群体包含组蛋白。55.根据编号实施方案45-54中任一实施方案所述的方法,其中所述DNA结合部分的群体包含纳米颗粒。56.根据编号实施方案45-55中任一实施方案所述的方法,其中所述结构包含天然染色质。57.根据编号实施方案45-56中任一实施方案所述的方法,其中在切割所述第一DNA分子之前,所述第一暴露的末端和所述第二暴露的末端在所述第一DNA分子上相隔至少10kb。58.根据编号实施方案45-57中任一实施方案所述的方法,其中在切割所述第一DNA分子之前,所述第一暴露的末端和所述第二暴露的末端在所述第一DNA分子上相隔至少15kb。59.根据编号实施方案45-58中任一实施方案所述的方法,其中在切割所述第一DNA分子之前,所述第一暴露的末端和所述第二暴露的末端在所述第一DNA分子上相隔至少30kb。60.根据编号实施方案45-59中任一实施方案所述的方法,其中在切割所述第一DNA分子之前,所述第一暴露的末端和所述第二暴露的末端在所述第一DNA分子上相隔至少50kb。61.根据编号实施方案45-60中任一实施方案所述的方法,其中在切割所述第一DNA分子之前,所述第一暴露的末端和所述第二暴露的末端在所述第一DNA分子上相隔至少100kb。62.根据编号实施方案45-61中任一实施方案所述的方法,其包括对所述第一DNA分子的第二拷贝进行测序。63.一种至少5kb的重排核酸分子,其包含:a)第一区段;b)第二区段;以及c)第三区段;d)所述第一区段和所述第二区段在第一接头处连接;并且e)所述第二区段和所述第三区段在第二接头处连接;其中所述第一区段、所述第二区段和所述第三区段在未重排核酸分子中以相隔至少10kb的相位存在,并且其中至少70%的所述重排核酸分子映射至所述共同的未重排核酸分子。64.根据编号实施方案63所述的重排核酸,其中所述第一区段、所述第二区段和所述第三区段包含来自基因组的共同核酸分子的单独的基因组核酸序列。65.根据编号实施方案63-64中任一实施方案所述的重排核酸,其中所述第一区段、所述第二区段和所述第三区段以重排核酸中重排的顺序存在于基因组中的共同分子中。66.根据编号实施方案63-65中任一实施方案所述的重排核酸,其中所述核酸分子的长度为至少30kb。67.根据编号实施方案63-66中任一实施方案所述的重排核酸,其中所述核酸在双链末端包含发夹环,使得所述分

子包含含有30kb反向重复的单链。68.根据编号实施方案63-67中任一实施方案所述的重排核酸，其中所述核酸为双链环状分子。69.根据编号实施方案63-68中任一实施方案所述的重排核酸，其中至少80％的所述重排核酸分子映射至所述共同的未重排核酸分子。70.根据编号实施方案63-69中任一实施方案所述的重排核酸，其中至少85％的所述重排核酸分子映射至所述共同的未重排核酸分子。71.根据编号实施方案63-70中任一实施方案所述的重排核酸，其中至少90％的所述重排核酸分子映射至所述共同的未重排核酸分子。72.根据编号实施方案63-71中任一实施方案所述的重排核酸，其中至少95％的所述重排核酸分子映射至所述共同的未重排核酸分子。73.根据编号实施方案63-72中任一实施方案所述的重排核酸，其中至少99％的所述重排核酸分子映射至所述共同的未重排核酸分子。74.根据编号实施方案63-73中任一实施方案所述的重排核酸，其中所述重排核酸分子的至少80％的区段映射至所述共同的未重排核酸分子。75.根据编号实施方案63-74中任一实施方案所述的重排核酸，其中所述重排核酸分子的至少85％的区段映射至所述共同的未重排核酸分子。76.根据编号实施方案63-75中任一实施方案所述的重排核酸，其中所述重排核酸分子的至少90％的区段映射至所述共同的未重排核酸分子。77.根据编号实施方案63-76中任一实施方案所述的重排核酸，其中所述重排核酸分子的至少95％的区段映射至所述共同的未重排核酸分子。78.根据编号实施方案63-77中任一实施方案所述的重排核酸，其中所述重排核酸分子的至少99％的区段映射至所述共同的未重排核酸分子。79.根据编号实施方案63-78中任一实施方案所述的重排核酸，其中所述重排核酸通过编号实施方案1-62中任何一个或多个实施方案所述的方法的步骤生成。80.一种生成样品核酸分子的定相序列的方法，该方法包括：a)从所述样品核酸分子生成编号实施方案63-78中任一实施方案的第一重排核酸分子；b)从所述样品核酸分子生成编号实施方案63-78中任一实施方案的第二重排核酸分子；以及c)对所述第一重排核酸分子和所述第二重排核酸分子进行测序；其中所述第一重排核酸分子和所述第二重排核酸分子是独立生成的。81.一种生成样品核酸分子的定相序列的方法，该方法包括：a)对来自所述样品核酸分子的编号实施方案63-78中任一实施方案的第一重排核酸分子进行测序；b)对来自所述样品核酸分子的编号实施方案63-78中任一实施方案的第二重排核酸分子进行测序；其中所述第一重排核酸分子和所述第二重排核酸分子是独立生成的；以及c)组装编号实施方案63-78中任一实施方案的第一重排核酸分子和编号实施方案63-78中任一实施方案的第二重排核酸分子的序列，使得组装的序列为样品核酸分子的未重排的定相序列。82.根据编号实施方案80-81中任一实施方案所述的方法，其中对第一重排核酸分子进行测序包括生成至少1kb的序列读取。83.根据编号实施方案80-82中任一实施方案所述的方法，其中对第一重排核酸分子进行测序包括生成至少2kb的序列读取。84.根据编号实施方案80-83中任一实施方案所述的方法，其中对第一重排核酸分子进行测序包括生成至少5kb的序列读取。85.根据编号实施方案80-84中任一实施方案所述的方法，其包括将至少70％的所述第一重排分子分配给单个基因组分子的共同相位。86.根据编号实施方案80-85中任一实施方案所述的方法，其包括将至少70％的所述第二重排分子分配给单个基因组分子的共同相位。87.根据编号实施方案80-86中任一实施方案所述的方法，其包括将至少80％的所述第一重排分子分配给单个基因组分子的共同相位。88.根据编号实施方案80-87中任一实施方案所述的方法，其包括将至少80％的所述第二重排分子分配给单个基因组分子的共同相位。89.根据编号实施方案80-88中任一实施方

案所述的方法，其包括将至少90％的所述第一重排分子分配给单个基因组分子的共同相位。90.根据编号实施方案80-89中任一实施方案所述的方法，其包括将至少90％的所述第二重排分子分配给单个基因组分子的共同相位。91.根据编号实施方案80-90中任一实施方案所述的方法，其包括将至少95％的所述第一重排分子分配给单个基因组分子的共同相位。92.根据编号实施方案80-91中任一实施方案所述的方法，其包括将至少95％的所述第二重排分子分配给单个基因组分子的共同相位。93.一种对长读取序列数据进行定相的方法，该方法包括：a)从编号实施方案63-78中任一实施方案的核酸样品中获得序列数据；b)从编号实施方案63-78中任一实施方案的重排核酸中获得长读取序列数据；c)将来自编号实施方案63-78中任一实施方案的重排核酸的长读取序列数据映射至来自所述核酸样品的序列数据；以及d)将映射至来自编号实施方案63-78中任一实施方案的重排核酸的长读取序列数据的来自所述核酸样品的序列数据分配给共同相位。94.一种通过DNA测序技术向由核酸样品生成的核酸数据集提供相位信息的方法，该方法包括：a)获得具有相隔大于所述DNA测序技术的读取长度的距离的第一区段和第二区段的所述核酸样品的核酸；b)对所述核酸进行改组，使得所述第一区段和所述第二区段相隔小于所述DNA测序技术的读取长度的距离；c)使用所述DNA测序技术对所述改组的核酸进行测序，使得所述第一区段和所述第二区段出现在所述DNA测序技术的单个读取中；以及d)将包含第一区段序列的数据集的序列读取和包含第二区段序列的数据集的序列读取分配给共同相位。95.根据编号实施方案94所述的方法，其中所述DNA测序技术生成具有至少10kb的读取长度的读取。96.根据编号实施方案94-94中任一实施方案所述的方法，其中改组包括进行编号实施方案1-62中任一实施方案的步骤。97.根据编号实施方案94-94中任一实施方案所述的方法，其中所述第一区段和所述第二区段由标记区段末端的连接体寡核苷酸隔开。98.一种核酸序列数据库，其包含从编号实施方案63-78中任一实施方案的多个分子获得的序列信息，其中从至少一个分析中排除对应于少于70％的区段映射至共同的支架的分子的序列信息。99.一种核酸序列数据库，其包含从编号实施方案63-78中任一实施方案的多个分子获得的序列信息，其中从至少一个分析中排除对应于少于70％的序列信息映射至共同的支架的分子的序列信息。100.一种对长读取序列数据进行定相的方法，该方法包括：a)从编号实施方案63-78中任一实施方案的核酸样品中获得序列数据；b)从编号实施方案63-78中任一实施方案的重排核酸中获得长读取序列数据；c)将编号实施方案63-78中任一实施方案的重排核酸的第一区段、第二区段和第三区段映射至来自所述核酸样品的序列数据以获得核酸样品序列数据；以及d)当至少两个区段映射至共同的支架时，将所述区段的序列变异分配给共同的相位。101.根据编号实施方案100所述的方法，其中所述第一区段包含相对于来自所述核酸样品的序列数据的单核苷酸多态性。102.根据编号实施方案100-101中任一实施方案所述的方法，其中所述第一区段包含相对于来自所述核酸样品的序列数据的插入。103.根据编号实施方案100-102中任一实施方案所述的方法，其中所述第一区段包含相对于来自所述核酸样品的序列数据的缺失。104.根据编号实施方案100-103中任一实施方案所述的方法，其包括将映射至第一共同的支架的第一组区段分配给所述第一共同的支架的共同相位，以及将映射至第二共同的支架的第二组区段分配给所述第二共同的支架的共同相位。105.一种核酸样品的核酸序列文库，所述核酸序列文库包含具有平均读取长度的核酸序列读取的群体，至少一个所述读取包含第一核酸区段的至少500个碱基和第二核酸区段的至少500个碱

基,其中发现所述第一核酸区段和所述第二核酸区段在所述核酸样品的共同分子上同相位相隔大于所述平均读取长度的距离。106.根据编号实施方案105所述的核酸序列文库,其中发现所述第一核酸区段和所述第二核酸区段同相位相隔大于10kb的距离。107.根据编号实施方案105-106中任一实施方案所述的核酸序列文库,其中发现所述第一核酸区段和所述第二核酸区段同相位相隔大于20kb的距离。108.根据编号实施方案105-107中任一实施方案所述的核酸序列文库,其中发现所述第一核酸区段和所述第二核酸区段同相位相隔大于50kb的距离。109.根据编号实施方案105-108中任一实施方案所述的核酸序列文库,其中发现所述第一核酸区段和所述第二核酸区段同相位相隔大于100kb的距离。110.根据编号实施方案105-109中任一实施方案所述的核酸序列文库,其中至少一个所述读取包含至少1kb的第一核酸区段。111.根据编号实施方案105-110中任一实施方案所述的核酸序列文库,其中至少一个所述读取包含至少5kb的第一核酸区段。112.根据编号实施方案105-111中任一实施方案所述的核酸序列文库,其中至少一个所述读取包含至少10kb的第一核酸区段。113.根据编号实施方案105-112中任一实施方案所述的核酸序列文库,其中至少一个所述读取包含至少20kb的第一核酸区段。114.根据编号实施方案105-113中任一实施方案所述的核酸序列文库,其中至少一个所述读取包含至少50kb的第一核酸区段。115.根据编号实施方案105-114中任一实施方案所述的核酸序列文库,其中核酸序列文库包含至少80%的所述核酸样品。116.根据编号实施方案105-115中任一实施方案所述的核酸序列文库,其中核酸序列文库包含至少85%的所述核酸样品。117.根据编号实施方案105-116中任一实施方案所述的核酸序列文库,其中核酸序列文库包含至少90%的所述核酸样品。118.根据编号实施方案105-117中任一实施方案所述的核酸序列文库,其中核酸序列文库包含至少95%的所述核酸样品。119.根据编号实施方案105-118中任一实施方案所述的核酸序列文库,其中核酸序列文库包含至少99%的所述核酸样品。120.根据编号实施方案105-119中任一实施方案所述的核酸序列文库,其中核酸序列文库包含至少99.9%的所述核酸样品。121.一种核酸样品的核酸序列文库,所述核酸序列文库包含平均长度为至少1kb的核酸序列读取的群体,所述读取独立地包含来自所述核酸样品的两个单独的同相位区域的至少300个碱基的序列,所述两个单独的同相位区域在所述核酸样品中相隔大于10kb的距离。122.根据编号实施方案121所述的核酸序列文库,其中所述读取独立地包含来自所述核酸样品的两个单独的同相位区域的至少500个碱基的序列。123.根据编号实施方案121-122中任一实施方案所述的核酸序列文库,其中所述读取独立地包含来自所述核酸样品的两个单独的同相位区域的至少1kb的序列。124.根据编号实施方案121-123中任一实施方案所述的核酸序列文库,其中所述读取独立地包含来自所述核酸样品的两个单独的同相位区域的至少2kb的序列。125.根据编号实施方案121-124中任一实施方案所述的核酸序列文库,其中所述读取独立地包含来自所述核酸样品的两个单独的同相位区域的至少5kb的序列。126.根据编号实施方案121-125中任一实施方案所述的核酸序列文库,其中所述读取独立地包含来自所述核酸样品的两个单独的同相位区域的至少10kb的序列。127.根据编号实施方案121-126中任一实施方案所述的核酸序列文库,其中所述两个单独的同相位区域在所述核酸样品中相隔大于20kb的距离。128.根据编号实施方案121-127中任一实施方案所述的核酸序列文库,其中所述两个单独的同相位区域在所述核酸样品中相隔大于30kb的距离。129.根据编号实施方案121-128中任一实施方案所述的核酸序列文库,其中在至少1%的读

取中所述两个单独的同相位区域在所述核酸样品中相隔大于50kb的距离。130.根据编号实施方案121-129中任一实施方案所述的核酸序列文库，其中在至少1％的读取中所述两个单独的同相位区域在所述核酸样品中相隔大于100kb的距离。131.根据编号实施方案121-130中任一实施方案所述的核酸序列文库，其中核酸序列文库包含至少80％的所述核酸样品。132.根据编号实施方案121-131中任一实施方案所述的核酸序列文库，其中核酸序列文库包含至少85％的所述核酸样品。133.根据编号实施方案121-132中任一实施方案所述的核酸序列文库，其中核酸序列文库包含至少90％的所述核酸样品。134.根据编号实施方案121-133中任一实施方案所述的核酸序列文库，其中核酸序列文库包含至少95％的所述核酸样品。135.根据编号实施方案121-134中任一实施方案所述的核酸序列文库，其中核酸序列文库包含至少99％的所述核酸样品。136.根据编号实施方案121-135中任一实施方案所述的核酸序列文库，其中核酸序列文库包含至少99.9％的所述核酸样品。137.一种由核酸样品生成的核酸文库，其中所述核酸样品中的至少80％的核酸序列在所述核酸文库中表示；并且所述核酸样品的同相位序列区段进行重排，使得在单个序列读取中读取所述核酸样品的至少一个远端定位的同相位区段对；使得对所述文库的测序同时生成跨越至少80％的核酸样品的叠连群信息，以及足以对所述叠连群信息进行排序和定向以生成所述核酸样品的定相序列的相位信息。138.根据编号实施方案137所述的核酸文库，其中所述核酸样品的至少90％的核酸序列在所述核酸文库中表示。139.根据编号实施方案137-138中任一实施方案所述的核酸文库，其中所述核酸样品的至少95％的核酸序列在所述核酸文库中表示。140.根据编号实施方案137-139中任一实施方案所述的核酸文库，其中所述核酸样品的至少99％的核酸序列在所述核酸文库中表示。141.根据编号实施方案137-140中任一实施方案所述的核酸文库，其中所述核酸样品的所述80％的核酸序列获自不多于100,000个文库组分。142.根据编号实施方案137-141中任一实施方案所述的核酸文库，其中所述核酸样品的所述80％的核酸序列获自不多于10,000个文库组分。143.根据编号实施方案137-142中任一实施方案所述的核酸文库，其中所述核酸样品的所述80％的核酸序列获自不多于1,000个文库组分。144.根据编号实施方案137-143中任一实施方案所述的核酸文库，其中所述核酸样品的所述80％的核酸序列获自不多于500个文库组分。145.根据编号实施方案137-144中任一实施方案所述的核酸文库，其中所述样品为基因组样品。146.根据编号实施方案137-145中任一实施方案所述的核酸文库，其中所述样品为真核基因组样品。147.根据编号实施方案137-146中任一实施方案所述的核酸文库，其中所述样品为植物基因组样品。148.根据编号实施方案137-147中任一实施方案所述的核酸文库，其中所述样品为动物基因组样品。149.根据编号实施方案137-148中任一实施方案所述的核酸文库，其中所述样品为哺乳动物基因组样品。150.根据编号实施方案137-149中任一实施方案所述的核酸文库，其中所述样品为单细胞真核基因组样品。151.根据编号实施方案137-150中任一实施方案所述的核酸文库，其中所述样品为人基因组样品。152.根据编号实施方案137-151中任一实施方案所述的核酸文库，其中所述核酸文库未进行条码化以保留相位信息。153.根据编号实施方案137-152中任一实施方案所述的核酸文库，其中所述文库的读取包含来自第一区域的至少1kb的序列和来自第二区域的至少100个碱基的序列，该第二区域与所述第一区域同相位并且在样品中与第一区域相隔大于50kb。154.一种配置核酸分子以用于在测序装置上测序的方法，其中所述核酸分子包含至少100kb的序列，并且其中所述至少100kb的序列

包含相隔大于所述测序装置的读取长度的长度的第一区段和第二区段，所述方法包括改变所述核酸分子的第一区段相对于第二区段的相对位置，使得所述第一区段和所述区段相隔小于所述测序装置的读取长度；其中所述第一区段和所述第二区段的相位信息得以保持；并且其中不多于10％的核酸分子缺失。155.根据编号实施方案154所述的方法，其包括生成跨越所述第一区段和所述第二区段的至少一部分的读取。156.根据编号实施方案154-155中任一实施方案所述的方法，其包括将所述第一区段和所述第二区段分配给所述核酸分子的序列的共同相位。157.根据编号实施方案154-156中任一实施方案所述的方法，其中不多于5％的核酸分子缺失。158.根据编号实施方案154-157中任一实施方案所述的方法，其中不多于1％的核酸分子缺失。159.根据编号实施方案154-158中任一实施方案所述的方法，其中在配置之前，所述第一区段和所述第二区段在核酸分子中相隔至少10kb。160.根据编号实施方案154-159中任一实施方案所述的方法，其中在配置之前，所述第一区段和所述第二区段在核酸分子中相隔至少50kb。161.根据编号实施方案154-160中任一实施方案所述的方法，其中在所述配置之后，所述第一区段和所述第二区段由接头标志物隔开。162.根据编号实施方案154-161中任一实施方案所述的方法，其包括将茎环附接在所述核酸的末端，从而将所述分子转化为单链。163.根据编号实施方案154-162中任一实施方案所述的方法，其包括将所述核酸分子环化。164.根据编号实施方案154-163中任一实施方案所述的方法，其包括将所述核酸分子附接至DNA聚合酶。165.根据编号实施方案154-164中任一实施方案所述的方法，其包括结合所述核酸分子，使得所述第一区段和所述第二区段独立于磷酸二酯骨架而保持在一起；在至少两个位置处切割所述第一区段与所述第二区段之间的磷酸二酯骨架；以及将所述第一区段与所述第二区段重新附接，使得所述第一区段和所述第二区段相隔小于所述测序装置的读取长度。166.根据编号实施方案154-165中任一实施方案所述的方法，其中所述切割和所述重新附接不会导致来自所述核酸分子的序列信息丢失。167.一种从第一核酸分子生成长距离相位信息的方法，该方法包括：a)提供包含具有第一区段、第二区段和第三区段的第一核酸分子的样品，其中所述第一区段、所述第二区段和所述第三区段在所述第一核酸分子上均不相邻，其中使所述第一核酸分子与框架接触，使得所述第一区段、所述第二区段和所述第三区段独立于所述第一核酸分子的共同的磷酸二酯骨架而与所述框架结合；b)切割所述第一核酸分子，使得所述第一区段、所述第二区段和所述第三区段不由共同的磷酸二酯骨架连接；c)将所述第一区段与所述第二区段连接，并将所述第二区段与所述第三区段连接；以及d)对包含所述第一区段、所述第二区段和所述第三区段的所述第一核酸分子的第一部分进行测序，从而生成第一区段序列信息、第二区段序列信息和第三区段序列信息，其中所述第一区段序列信息、所述第二区段序列信息和所述第三区段序列信息提供关于所述第一核酸分子的长距离相位信息。168.根据编号实施方案167所述的方法，其中所述框架包含重构的染色质。169.根据编号实施方案167-168中任一实施方案所述的方法，其中所述框架包含天然染色质。170.根据编号实施方案167-169中任一实施方案所述的方法，其中用限制酶进行所述切割。171.根据编号实施方案167-170中任一实施方案所述的方法，其中用片段化酶进行所述切割。172.根据编号实施方案167-171中任一实施方案所述的方法，其进一步包括，在所述测序之前，从样品中去除包含至多两个区段的第一核酸分子的第二部分。173.根据编号实施方案167-172中任一实施方案所述的方法，其进一步包括使用所述第一区段序列信息、所述第二区段序列信息和所述第三区段

序列信息组装所述第一核酸分子的序列。174.一种对核酸分子进行测序的方法，该方法包括：获得包含共有共同的磷酸二酯骨架的第一区段、第二区段和第三区段的第一核酸分子，其中所述第一区段、第二区段和第三区段在所述第一核酸分子上均不相邻；对所述核酸分子进行分区，使得所述第一区段、第二区段和第三区段独立于它们共同的磷酸二酯骨架而相关联；切割所述核酸分子以生成片段，使得不存在连接所述第一区段、第二区段和第三区段的连续磷酸二酯骨架；连接所述片段，使得所述第一区段、第二区段和第三区段在共有共同的磷酸二酯骨架的重排核酸分子上是连续的；以及对所述重排核酸分子的至少一部分进行测序，使得在单个读取中对所述重排核酸分子的至少5,000个碱基进行测序。175.根据编号实施方案174所述的方法，其中分区包括使所述核酸分子与结合部分接触，使得所述第一区段、第二区段和第三区段独立于它们共同的磷酸二酯骨架而结合在共同的复合体中。176.根据编号实施方案174-175中任一实施方案所述的方法，其中使所述核酸分子与多个DNA结合分子接触包括与DNA结合蛋白的群体接触。177.根据编号实施方案174-176中任一实施方案所述的方法，其中所述DNA结合蛋白的群体包含核蛋白。178.根据编号实施方案174-177中任一实施方案所述的方法，其中所述DNA结合蛋白的群体包含核小体。179.根据编号实施方案174-178中任一实施方案所述的方法，其中所述DNA结合蛋白的群体包含组蛋白。180.根据编号实施方案174-179中任一实施方案所述的方法，其中使所述核酸分子与多个DNA结合部分接触包括与DNA结合纳米颗粒的群体接触。181.根据编号实施方案174-180中任一实施方案所述的方法，其中切割所述核酸分子包括与限制性内切核酸酶接触。182.根据编号实施方案174-181中任一实施方案所述的方法，其中切割所述核酸分子包括与非特异性内切核酸酶接触。183.根据编号实施方案174-182中任一实施方案所述的方法，其中切割所述核酸分子包括与标签化酶接触。184.根据编号实施方案174-183中任一实施方案所述的方法，其中切割所述核酸分子包括与转座酶接触。185.根据编号实施方案174-184中任一实施方案所述的方法，其中切割所述核酸分子包括剪切所述第一分子。186.根据编号实施方案174-185中任一实施方案所述的方法，其中分区包括将所述核酸分子与样品的其他核酸分子分离。187.根据编号实施方案174-186中任一实施方案所述的方法，其中分区包括稀释所述核酸样品。188.根据编号实施方案174-187中任一实施方案所述的方法，其中分区包括将所述核酸分子分配至乳剂的微滴中。189.一种代表生物体的基因组的基因组相位信息的核酸分子，所述核酸分子包含映射至单个基因组分子的至少20kb的核酸序列信息，其中所述序列信息包含相对于其在基因组分子中的位置重排的区段，并且其中至少70％的唯一地映射至所述生物体的基因组的序列信息映射至单个基因组分子。190.根据编号实施方案189所述的核酸分子，其中所述核酸分子包含至少20个区段。191.根据编号实施方案189-190中任一实施方案所述的核酸分子，其中所述区段在所述生物体的基因组中不相邻。192.一种包含至少100个至少20kb的核酸分子组分的核酸文库，其中组分包含生物体的基因组的重排区段；其中来自文库组分的至少70％的唯一映射区段映射至共同的基因组分子；并且其中组分不与核酸结合部分结合。193.一种核酸数据集，其包含对应于至少100个至少20kb的核酸分子组分的序列，其中组分包含生物体的基因组的至少5个重排区段，并且其中从下游分析中排除少于70％的所述重排区段映射至共同的支架的组分。194.一种核酸数据集，其包含对应于至少100个至少20kb的核酸分子组分的序列，其中组分包含生物体的基因组的至少5个重排区段，并且其中从下游分析中排除少于70％的所述序列唯一地映射

至共同的支架的组分。

[0184] 参考附图，可以看到本文所讨论的某些实施方案的说明。在图1中，可以看到在构建间断的、重排的相位保留核酸分子的过程中的中间体。单核酸分子已经与核酸结合部分结合，如与重构的染色质复合体结合，并与甲醛接触以交联该复合体。该复合体涉及单个核酸起始分子，其与核酸结合组分形成簇，统称为重构染色质，使得仅核酸分子的内环从簇突出。如图1所描绘的，使用限制性内切核酸酶MboI切割突出的环以生成粘性末端。

[0185] 在备选的实施方案中，核酸分子与珠子或表面结合，如SPRI包被的珠子或其他核酸结合剂包被的珠子。核酸样品在这样的条件下结合，使得每个珠子仅结合一个核酸分子，或者使得结合的核酸在切割后不可能交叉连接。而且，使用另一种限制性内切核酸酶、转座酶、标签化酶、非特异性内切核酸酶、拓扑异构酶或具有内切核酸酶活性的其他试剂交替完成切割。

[0186] 在图2中，可以看到使用核酸聚合酶和单个dGTP群体处理图1的切割的核酸复合体，以补平突出端的单个位置。补平步骤防止复合体的粘性末端在后续步骤中交叉退火和连接。在一些情况下，排除该步骤，并且允许复合体交叉连接而没有标点寡核苷酸。或者，生成平末端，或通过转座酶而不是限制性内切核酸酶的作用添加标签化衔接子。

[0187] 图3示出了在标点寡核苷酸与复合体的暴露末端退火并连接之后图1和图2的复合体。标点寡核苷酸被描绘为细实线而不是核酸碱基序列。任选地修饰标点寡核苷酸以防止连环化，例如通过去除5’磷酸基团。标点寡核苷酸任选地被设计成与图2中修饰的游离粘性末端相容。在其他实施方案中，切割的核酸末端可以直接彼此连接，而不插入标点寡核苷酸。

[0188] 图4描绘了在通过使用蛋白酶K处理逆转交联和从重构的染色质中释放之后释放的间断核酸分子。终产物间断核酸包含由标点寡核苷酸401隔开的区段400。该区段保留原始核酸分子的相位信息，但相对于起始分子是随机排序和定向的。基本上所有原始核酸分子的序列均存在于间断分子中，使得对间断分子的测序生成足以生成从头叠连群的序列信息。

[0189] 在使用长读取测序装置对间断核酸进行测序后，观察对应于未切割区段的序列段，导出其局部顺序和方向以及相位信息。还观察到跨越标点寡核苷酸序列的长序列读取区域。已知标点寡核苷酸的任一侧上的这些序列区段彼此同相位(并且与间断分子上的其他区段同相位)，但不太可能处于正确的顺序和方向。重排过程的益处是使样品分子上彼此远离的区段邻近，使得它们跨越单个读取。另一个益处是极大地保留了原始样品分子的序列信息，从而同时生成了从头叠连群信息。

[0190] 图5示出了本公开内容的备选实施方案。一系列短成对末端500(每个指示在该对中连接的序列是同相位的)经衔接子标记(例如，用扩增衔接子)501并连接以形成连接的成对末端多聚体502。它们唯一映射至的个体对或叠连群被确信地分配给共同的相位。除非在多联体组装中采取另外的措施，否则扩增衔接子任一侧上的读取对单元不能被推断为彼此具有顺序、方向或相位关系。

[0191] 图5的连接分子的益处是将多个成对末端读取组装成单个分子，该单个分子在单个或更少数目的长读取反应中进行测序，而不是在更多数目的短运行读取中进行测序。然而，因为单个成对末端的区段长度较短，所以起始样品的总体序列不可能保留在连接分子

中，从而使从头测序复杂化。

[0192] 图6示出了替代情形，其中使用间断核酸分子600来生成用于短读取测序的模板。使间断核酸分子与引物601的群体接触，所述引物与标点序列退火并且包含区间(bin)特异性寡核苷酸条码602。然后可以对引物进行延伸，例如，以掺入与间断核酸分子互补的序列603。通过这种方法，从条码信息中导出相位信息。益处是促进了短读取测序。

[0193] 图7示出了在连接步骤('BF')之前和连接步骤('AF')之后两个样品的凝胶电泳分析。最左边的泳道包含DNA梯，其大小从上到下为48500、15000、7000、4000、3000、2500、2000、1500、1200、900、600、400、250和100bp。从左侧开始的第二泳道和第三泳道分别在连接之前和之后包含样品1。从左侧开始的第四泳道和第五泳道分别在连接之前和之后包含样品2。与样品1和样品2二者连接的泳道显示在7000-48500bp范围内(远大于任一连接前泳道中的带)的DNA暗带。样品1包含约7纳克DNA/微升(ng/μL)，总共约200ng DNA，并且样品2包含约115ng/μL DNA，总共约3.4μg DNA。

[0194] 图8呈现了关于样品的测序信息的代表性信息。生成超过1,000,000个环状共有序列(CSS)读取，其具有300,000个未映射的读取(25%)。有1,500,000个映射区段(-q 1)和1,350,000个映射区段(-q 20)。对于具有1个映射区段的读取，n＝500,000；对于具有2个映射区段的读取，n＝175,000；对于具有3个映射区段的读取，n＝75,000；对于具有4个映射区段的读取，n＝30,000；对于5个映射区段的读取，n＝15,000；对于具有6个映射区段的读取，n＝7,000。表1示出了具有X个最大映射区段数的读取的克隆覆盖度。

[0195] 图9A和图9B示出了对于具有10kb区间(图9A)和1kb区间(图9B)的样品而言，由具有X个映射区段的读取所跨越的距离的频率分布。y轴显示PacBio CCS读取的数目(轴线从下到上：1、10、100、1000、10000)。x轴显示由读取所跨越的距离(轴线从左到右：图9A：0、200000、400000、600000、800000、1000000；图9B：0、20000、40000、60000、80000、100000)。显示了具有1个映射区段(901,911)、2个映射区段(902,912)、3个映射区段(903,913)、4个映射区段(904,914)和5个映射区段(905,915)的读取的频率分布。

[0196] 图10描绘了适于实现本文所述方法的示例性计算机系统1000。系统1000包括中央计算机服务器1001，其被编程为实现本文所述的示例性方法。服务器1001包括中央处理单元(CPU，也称为“处理器”)1005，其可以是单核处理器、多核处理器或用于并行处理的多个处理器。服务器1001还包括存储器1010(例如随机存取存储器、只读存储器、闪速存储器)；电子存储单元1015(例如硬盘)；用于与一个或多个其他系统通信的通信接口1020(例如网络适配器)；以及外围设备1025，其可包括高速缓冲存储器、其他存储器、数据存储和/或电子显示适配器。存储器1010、存储单元1015、接口1020和外围设备1025通过诸如主板的通信总线(实线)与处理器1005通信。存储单元1015可以是用于存储数据的数据存储单元。服务器1001借助于通信接口1020可操作地耦合至计算机网络(“网络”)1030。网络1030可以是因特网、内联网和/或外联网，与因特网、电信或数据网络通信的内联网和/或外联网。在一些情况下，网络1030借助于服务器1001可以实现对等网络，这可以使得耦合至服务器1001的设备能够起到客户端或服务器的作用。

[0197] 存储单元1015可以存储文件，如受试者报告、和/或与护理人员的通信、测序数据、关于个体的数据或与本发明相关的数据的任何方面。

[0198] 服务器可以通过网络1030与一个或多个远程计算机系统通信。该一个或多个远程

计算机系统可以是例如个人计算机、便携式计算机、平板型计算机、电话、智能电话或个人数字助理。

[0199] 在一些情况下，系统1000包括单个服务器1001。在其他情况下，系统包括通过内联网、外联网和/或因特网彼此通信的多个服务器。

[0200] 服务器1001可以适于存储来自受试者的测量数据、患者信息，例如多态性、突变、病史、家族史、人口统计数据和/或潜在相关性的其他信息。这样的信息可以存储在存储单元1015或服务器1001上，并且这样的数据可以通过网络传送。

[0201] 如本文所用，核酸区段在它们同相位时“邻近”并且可以至少部分地包括在单个读取中。

[0202] 实施例

[0203] 实施例1.一些长读取测序方法无法对二倍体DNA样品中的一些突变进行定相

[0204] 特定人类疾病的治疗取决于功能性基因产物的存在。在该基因产物存在下，治疗性分子被代谢以产生有效的代谢物。在缺乏该基因产物的情况下，治疗性分子累积并对患者有害。

[0205] 对患者基因组进行鸟枪法测序，并确定两个点突变映射至编码对于治疗功效所必需的基因产物的基因座。在组装的鸟枪支架中，两个点突变间隔30kb。两个点突变的相位信息是不可用的，因此从业者无法确定患者是否具有野生型等位基因和双突变等位基因，或者确定在备选方案中患者是否独立地具有两个单突变无效等位基因，一个位于基因座的5’端且第二个位于基因座的3’端。

[0206] 从患者提取DNA，并在长读取测序仪上对样品进行测序。单个长读取的平均极限为10-15kb。该读取证实患者对于第一突变和第二突变二者均是杂合的。然而，鉴于患者基因组中的突变间隔30kb，使用生成的序列信息不能获得相位信息。因此，从业者无法确定患者是否具有野生型等位基因和双重突变无效等位基因并因此适合使用治疗性分子进行治疗，或者确定患者是否具有两个单突变无效等位基因并因此无法代谢治疗性分子。患者被拒绝治疗并继续遭受该病况。

[0207] 本实施例证明与鸟枪读取组合使用的长范围测序方法不能准确地对突变进行定相，特别是当突变由长段的纯合DNA隔开时。此外，本实施例说明了未能准确地将相位信息分配给基因组序列对患者健康具有影响。

[0208] 实施例2.二倍体DNA样品中突变的成功定相

[0209] 使用本文公开的方法对来自实施例1的患者的DNA进行相位分析。

[0210] 从实施例1中描述的患者中提取DNA。生成间断的插入改组分子的文库，使得在对序列区段的相对位置进行重排时保留相位信息。

[0211] 将提取的DNA在体外组装成重构的染色质。用限制酶MboI切割重构的染色质。用单个碱基部分地补平得到的粘性末端，以防止限制酶生成的突出端的重新连接。将具有与消化的DNA样品的部分补平的突出端相容的5’端和3’端的标点寡核苷酸与DNA连接酶一起添加至DNA样品中。标点寡核苷酸缺少5’磷酸基团以避免寡核苷酸的连环化。该连接步骤导致DNA区段的重组，因为最初彼此不相邻的末端在连接后彼此相邻。由于DNA分子在该过程期间与交联的重构染色质支架结合，因此维持了相位信息。

[0212] 确定足够的序列信息，使得在不使用独立于相位确定的鸟枪序列步骤的情况下获

得完整的基因组信息。确定患者对于感兴趣的基因中的第一无效突变和第二无效突变是杂合的。

[0213] 此外，观察到文库分子，其中对包含两个突变的第一DNA区段和第二DNA区段进行重排而不丢失相位信息，使得小于15kb的序列将它们隔开。生成跨越重排区域的读取，并且发现该读取包含第一无效突变和第二无效突变。由于重排的DNA样品中的第一DNA区段和第二DNA区段相隔小于15kb，因此两个突变均能够在单个测序读取中检测到，从而导致定相信息。该定相信息用于确定患者携带双突变等位基因。观察到第二读取，其具有不同的接头点，但也具有跨越基因座的第一杂合区和第二杂合区的第一区段和第二区段。观察到重排分子中的第一区域和第二区域均编码野生型序列。

[0214] 对包含相位保留重排的另外的分子进行测序。发现另外的分子在相对于彼此的不同位置处具有标点插入物。重排的分子均不具有单个无效突变和单个野生型等位基因。相反，跨越两个杂合区的所有序列读取均包含两个基因座处的野生型等位基因或两个基因座处的无效突变。

[0215] 确定患者基因组包含双突变无效等位基因和野生型等位基因。结论是治疗可能是有效的。向患者施用治疗性分子，并通过治疗性分子的有益活性减轻患者的病况。

[0216] 本实施例说明了本文公开的方法和组合物允许从单个模板文库同时进行从头序列生成和定相。不需要单独的鸟枪测序文库和相位确定文库，因此大大降低了测序测定的成本。

[0217] 本实施例还说明了本文公开的方法和组合物允许准确地、冗余地对分子进行定相，即使分子大部分相同，并且杂合位置在使用的测序技术中由大于读取长度的两倍的同一性区域隔开。

[0218] 实施例3.一些长读取测序方法在富含转座子的作物DNA样品的定相中是不成功的

[0219] 估计大约90％的玉米基因组是转座因子，如转座子。由于一些转座子的重复性，对等位基因进行定相是困难的。为了产生具有提高的产量和改善的营养含量的玉米菌株，需要玉米双突变系。两种突变均是显性的，并且发现在染色体的相对末端上。将高产量玉米菌株与高类胡萝卜素水平玉米菌株杂交以产生杂合系，然后将其自交以生成分离的后代。

[0220] 观察到一些后代表现出提高的产量和增加的营养含量。本项目的下一步是将高产量和高营养含量菌株中的一种与表现出枯萎病抗性的菌株杂交。已知如果枯萎病抗性突变与高产量突变或改善的营养含量突变包含在相同DNA分子上，则其失去功效。为了使及时和昂贵的下游测序以及表型分型实验最小化，期望进行枯萎病抗性菌株与亲本菌株的杂交，所述亲本菌株在相同DNA分子上含有高产量和高营养含量突变。

[0221] 来自初始杂交的两个亲本系是近等基因系，进行繁殖以使其基因组的变异最小化。结果，发现极少标志物可用于促进相位确定。从数千株得到的幼苗中提取DNA进行测序，以确定哪些含有在相同DNA分子上同相位的产量和营养突变。因为产量基因和类胡萝卜素基因由重复的、高度保守的转座因子隔开，并且因为除了这些突变之外的品系之间几乎没有变异，所以短读取测序仪不能提供定相信息。因为在染色体的相对末端发现了产量基因突变和类胡萝卜素基因突变，所以通过长读取测序技术在单个长读取中不能检测到这两种突变。结果，数千株幼苗中的任一株是否在单个染色体上具有所需的高产量突变和高营养突变的组合尚不得知。确定该项目不能保持在预算范围内，并因此取消了该项目。

[0222] 实施例4.富含转座子的作物DNA样品的成功定相

[0223] 提取并修饰来自实施例3的玉米幼苗的DNA样品以生成区段改组的相位保留测序文库。在长读取测序仪上对得到的重排DNA分子进行测序。获得跨越产量突变基因座和营养突变基因座的单个序列读取，其由一个或多个标点寡核苷酸隔开。对于一些幼苗样品，观察到表明两个有益突变在单个分子上同相位的读取。选择经确认的同相位高产量和改善的营养含量菌株中的一种并将其与枯萎病抗性菌株杂交，以产生强健的玉米菌株，其将在发展中国家产生急需的增加的营养。

[0224] 本实施例证明了本文公开的方法和组合物如何用于确定具有多个重复元件的复杂基因组的相位信息。本技术允许甚至在复杂的基因组(如相关作物物种的基因组)中进行准确、快速的相位确定。

[0225] 实施例5.具有难以区分的相位的含突变核酸

[0226] 二倍体生物体含有遗传物质的每个染色体的两个拷贝。由至少30kb的相同序列隔开的两个突变存在于二倍体基因组的单个染色体上。在平均读取长度为15kb的长读取测序仪上对DNA样品进行测序。不可能确定两个突变是否包含在相同或不同的核酸分子上。

[0227] 实施例6.确定核酸样品的相位信息

[0228] 从实施例5的生物体中提取DNA。用DNA结合蛋白体外组装DNA以生成重构的染色质。将重构的染色质切割以产生粘性末端，该粘性末端被部分补平以防止重新连接。将具有与部分补平的粘性末端相容的末端的标点寡核苷酸与DNA连接酶一起添加至染色质样品中。在一些情况下，将标点寡核苷酸去磷酸化以避免寡核苷酸的多联体化(contatemerization)。与起始DNA样品相比，对重新连接的染色质样品的DNA区段进行重排，尽管由于分子通过加标点过程与染色质蛋白结合而维持了相位信息。在一些情况下，对基因组内的两个突变进行重排，使得它们相隔小于15kb。在这种情况下，分隔距离小于长读取测序仪的平均读取长度。当重排的DNA样品从染色质蛋白中释放并进行测序时，确定相位信息并生成足以生成从头序列支架的序列信息。

[0229] 实施例7.确定核酸样品-平端连接的相位信息

[0230] 从实施例5的生物体中提取DNA，并在体外用DNA结合蛋白重新组装以生成重构的染色质。将DNA进行切割以产生平末端。将具有平末端的标点寡核苷酸连接至切割的DNA样品的平末端。将标点寡核苷酸去磷酸化以避免寡核苷酸的多联体化。重排的DNA样品从染色质蛋白中释放并如实施例6中那样进行测序。当重排的DNA样品从染色质蛋白中释放并进行测序时，确定相位信息并生成足以生成从头序列支架的序列信息。

[0231] 实施例8.将标点分子-短读取条码化

[0232] 如实施例6-7中任一个所述生成包含标点寡核苷酸的DNA样品。在从DNA结合蛋白释放后，游离DNA样品(被称为间断DNA分子)与包含至少两个区段的寡核苷酸接触。一个区段包含条码，而第二区段包含与标点序列互补的序列。在与标点序列退火后，用聚合酶延伸条码化的寡核苷酸以产生来自相同DNA分子的条码化分子。这些条码化分子包含条码序列、标点互补序列和基因组序列。在短读取测序仪上对延伸产物进行测序，并通过将具有相同条码的序列读取分组至共同的相位来确定相位信息。

[0233] 实施例9.将标点分子-长读取条码化

[0234] 如实施例8中那样对DNA样品进行提取、间断和条码化。在延伸后，将条码化产物大

量连接在一起以生成长分子，该长分子使用长读取测序技术进行读取。嵌入的读取对可通过扩增衔接子和标点序列识别。从读取对的条码序列获得进一步的相位信息。

[0235] 实施例10.用转座子标点确定相位信息

[0236] 提取实施例5的DNA样品并在体外用DNA结合蛋白重新组装以生成重构的染色质。将与两个未连接的标点寡核苷酸结合的转座酶添加至DNA样品。转座酶切割暴露的DNA区段并将两个标点寡核苷酸插入DNA中。因为给定转座酶中的标点寡核苷酸是未连接的，所以插入导致两个游离DNA末端，每个末端被两个标点寡核苷酸之一终止，并且每个末端拴系至重构的染色质以保留相位信息。将DNA连接酶添加至样品中以将平DNA末端连接在一起，从而导致DNA区段的重排，尽管由于DNA分子在整个过程中与染色质蛋白结合而维持了相位信息。重排的DNA样品从染色质蛋白中释放并如实施例6那样进行测序以确定相位信息。

[0237] 实施例11.用转座子标点-短读取确定相位信息

[0238] 如实施例10中所述，提取DNA样品，将其在体外重新组装成重构的染色质，并用转座酶进行间断。在重新连接平末端后，通过限制性消化从蛋白质-DNA复合体中释放重新连接的DNA区段，导致多个成对末端，其随后与扩增衔接子连接。扩增后，用短距离技术对成对末端进行测序。可以确信地得出结论，对于间断接头的任一侧，标点相邻序列来源于共同分子的共同相位。

[0239] 实施例12.用转座子标点-长读取确定相位信息

[0240] 如实施例10中所述提取DNA样品，将其在体外重新组装成重构的染色质，并用转座酶进行间断。在重新连接平末端后，通过限制性消化从蛋白质-DNA复合体中释放重新连接的DNA区段，导致多个成对末端，其随后与扩增衔接子连接。扩增后，将多个成对末端大量连接在一起以生成长分子，该长分子使用长读取测序技术进行读取。嵌入的读取对可通过与转座酶标点序列相邻的天然DNA序列识别。在长序列装置上读取连接的间断接头，并获得多个接头的序列信息。发现接头映射至多个不同的染色体。然而，可以确信地得出结论，对于间断接头的任一侧，标点相邻序列来源于共同分子的共同相位。

[0241] 实施例13.Chicago对的多联体生成

[0242] 提取DNA样品，并将其在体外用DNA结合蛋白组装以生成重构的染色质。将DNA进行切割以产生粘性末端。用生物素化的核苷酸补平粘性末端，随后平接补平的末端以生成DNA区段对(被称为Chicago对)。这些重新改组的核酸从染色质蛋白中释放，将其进行切割，并分离链霉亲和素结合的连接接头。将扩增衔接子添加至Chicago对的游离端。扩增后，将Chicago对大量连接在一起以生成长分子，该长分子使用长读取测序技术进行读取。嵌入的读取对可通过扩增衔接子进行识别。在用于引入生物素化碱基的‘补平过程’中生成的序列重复也用于鉴定在相位序列中连接的接头。

[0243] 在长读取测序装置的单个读取中对连接的多联体进行测序。由于各个接头是连接的，所以能够在单个读取中对多个接头进行测序。

[0244] 实施例14.对发夹DNA分子进行定相

[0245] 将如在实施例6、7、9、10或12中的任一实施例中生成的长的、间断的DNA分子在一端上连接至发夹衔接子，产生了具有反向重复的自退火单链分子。通过测序酶馈送分子，并获得反向重复各侧的全长序列。得到的序列读取对应于具有多个重排区段(每个重排区段传递相位信息)的间断DNA分子的2x覆盖度。生成足够的序列以独立地生成核酸样品的从头

支架。

[0246] 实施例15.对环化的DNA分子进行定相

[0247] 对如在实施例6、7、9、10或12中的任一实施例中生成的长的、间断的DNA分子进行切割以形成所需长度的双链分子群。将这些分子在每一端上与单链衔接子连接。结果是双链DNA模板在两端处覆盖有发夹环。通过连续测序技术对环状分子进行测序。含有长双链区段的分子的连续长读取测序导致每个分子的单个连续读取。含有短双链区段的分子的连续测序导致分子的多个读取,其单独使用或与连续长读取序列信息一起使用以确认分子的共有序列。鉴定由标点寡核苷酸标记的基因组区段边界,并且得出结论,与标点边界相邻的序列是同相位的。生成足够的序列以独立地生成核酸样品的从头支架。

[0248] 实施例16.使用多个间断DNA分子的定相序列组装

[0249] 如实施例6、7、9、10或12中任一实施例所述生成多个间断DNA分子,并随后使用长读取测序技术进行测序。比较来自多个间断DNA分子的序列。观察到所述多个分子中的两个分子共有共同的序列,但是已经独立地衍生并具有不同的标点寡核苷酸。对于第一分子上的给定标点寡核苷酸,在每个标点寡核苷酸的任一侧上确定序列,并且得出结论,标点寡核苷酸的任一侧上的序列区段在共同分子上是同相位的。然而,同相位区段的相对位置是不清楚的。

[0250] 将第一间断DNS分子的一个区段与第二间断DNA分子的序列进行比较。发现第一分子的标点寡核苷酸附近的区段末端映射至第二间断DNA分子的区段的内部。将在第一间断DNA分子的标点寡核苷酸之外对齐的第二间断寡核苷酸的区段的序列映射至第一标点DNA分子,并鉴定远端区段。使用第二DNA分子区段作为指导,确定第一间断DNA分子的两个区段在原始核酸样品中彼此相邻定位。

[0251] 也就是说,使用第一间断分子来确定其组成区段的相位信息,而使用与第二(和另外的)间断DNA分子的未间断区域的比较来对第一间断分子的区段进行排序。相互重复该过程,确定多个标点寡核苷酸中的每一个中的大多数区段的相位和顺序信息。

[0252] 得到的组装序列是在重排发生之前输入DNA分子的定相序列,并且代表核酸样品的从头、定相组装。

[0253] 实施例17.使用长读取序列数据对短读取测序数据进行定相

[0254] 如实施例6、7、9、10或12中任一实施例所述生成间断DNA分子,并随后使用长读取测序技术进行测序。平行地,使用标准的短读取鸟枪测序技术对输入DNA进行测序。将来自样品的鸟枪序列映射至从重排的DNA分子生成的长读取数据。将来自间断分子的定相基因组序列读取映射至从同时生成的短读取测序获得的测序数据。一些短读取映射至长读取生成的序列。这种重叠允许将短序列读取分配给与从间断DNA分子长序列读取生成的基因组序列相同的相位。

[0255] 实施例18.核酸序列文库-长读取

[0256] 如实施例6、7、9、10或12中任一实施例所述生成多个间断DNA分子,并随后使用长读取测序技术进行测序。对每个间断分子进行测序,并分析序列读取。对于序列反应而言,序列读取平均为10kb。鉴定序列读取,该读取包含由标点寡核苷酸序列连接的第一区段的至少500个碱基和第二区段的500个碱基。将第一区段序列和第二区段序列映射至支架基因组,并发现第一区段序列和第二区段序列映射至间隔至少100kb的叠连群。

[0257] 第一叠连群和第二叠连群各自包含单个杂合位置，其相位未在支架中确定。第一叠连群的杂合位置由长读取的第一区段跨越，并且第二叠连群的杂合位置由长读取的第二区段的500个碱基跨越。

[0258] 所述读取各自跨越它们的叠连群的各自的杂合区。读取区段的序列表明第一叠连群的第一等位基因和第二叠连群的第一等位基因是同相位的。由于在单个长序列读取中检测到来自第一核酸区段和第二核酸区段的序列，因此确定第一核酸区段和第二核酸区段包含在输入DNA样品中的相同DNA分子上。

[0259] 本实施例证明了来自标点分子的长读取提供了在基因组支架上彼此远离的叠连群的相位信息。本实施例还表明，因为与标点寡核苷酸相邻的每个区段的大小足以促进精确映射，所以所述映射是以高置信度完成的，并且增加了跨越杂合位置的可能性。

[0260] 实施例19.核酸序列文库-短读取

[0261] 如实施例8或11中所述生成多个成对末端分子，并随后使用长读取测序技术进行测序。文库的平均读取长度被确定为1kb。成对末端分子包含第一DNA区段和第二DNA区段，其在输入DNA样品内是同相位的并相隔大于10kb的距离。由成对末端分子生成序列读取，其中一些序列读取包含来自第一核酸区段的至少300个碱基的序列和来自第二核酸区段的至少300个碱基的序列。由于在单个序列读取中检测到来自第一核酸区段和第二核酸区段的序列，因此确定第一核酸区段和第二核酸区段在输入DNA样品中的相同DNA分子上是同相位的。

[0262] 本实施例说明了使用如本文所教导的重排的间断分子，可以生成序列文库，该序列文库产生DNA区段的相位信息，该DNA区段在核酸样品中相隔大于用于对它们测序的测序技术的读取长度。

[0263] 实施例20.核酸序列文库-同时定相的DNA组装

[0264] 从间断DNA文库生成多个序列读取。该文库如实施例18或19中所述传递相位信息，使得标点事件的任一侧上的区段被确定为在单个分子上同相位。另外，生成的序列读取代表输入DNA样品的至少80％的核酸序列。序列读取用于生成跨越至少80％的输入DNA样品的从头叠连群信息。另外，序列读取用于确定相位信息，该相位信息随后用于将叠连群相对于彼此进行排序和定向，以生成输入DNA样品的定相序列组装。

[0265] 本实施例说明了间断DNA分子传递相位信息，并且在一些情况下还包括含有总核酸序列的实质部分的序列信息，使得同时生成从头序列组装。

[0266] 实施例21.DNA分子定相

[0267] 提取高分子量(HMW)DNA样品，该DNA样品包含至少一些长度为至少100kb的DNA分子。一种100kb DNA分子包含第一DNA区段和第二DNA区段，该第一DNA区段和第二DNA区段相隔大于标准测序技术的平均读取长度的距离。核酸样品为二倍体，但包含大的序列同一性区域，从而使相位确定复杂化。

[0268] 为了确信地确定相位，需要在单个测序读取内检测第一DNA区段和第二DNA区段。因此，必须改变第一DNA区段和第二DNA区段的相对位置，使得第一DNA区段和第二DNA区段相隔小于标准测序技术的平均读取长度的距离。这种重排不得导致相位信息的丢失。通过本文公开的方法和如实施例6、7或10中任一实施例所述的方法实现这种重排。在相位保持重排过程中，不多于10％的起始HMW DNA分子缺失。也就是说，仅通过缺失插入序列不会使

第一区段和第二区段邻近。相反，所述区段相对于彼此重排而不缺失大部分插入序列。由于几乎整个输入DNA分子得以保留，因此在测序后，使用生成的序列读取对从头生成的叠连群进行组装、排序和定向，使得对几乎整个输入DNA分子进行测序、组装和定相。

[0269] 实施例22.哺乳动物细胞培养物的分析

[0270] 使用本文所述的技术分析哺乳动物细胞培养物的样品。简言之，培养哺乳动物细胞的细胞培养物。使细胞进行交联，将交联猝灭，并且将细胞沉淀物在-20℃下储存。将细胞匀浆化并在裂解缓冲液中回收细胞核。使匀浆中的细胞核与SPRI珠子结合并使用DpnII限制酶进行消化。在没有生物素-11-dCTP的情况下补平末端并连接平末端。使交联逆转，将DNA回收并清理并准备用于测序。采用Pacific Biosciences SMRT长读取测序进行测序。在一些情况下，在测序之前，可以针对长度为至少约6kb的分子对DNA进行大小选择。

[0271] 测试两个样品以确保适当地发生连接。图7为指示在单独样品中成功连接的结果的代表。可以看到对于每个样品，连接导致向高得多的分子量核酸的转变。

[0272] 在图8中，可以看到这样的文库生成过程的结果。在超过1,000,000个环状共有序列(CSS)读取中，只有300,000个未映射。有1,500,000个映射区段(-q 1)和1,350,000个映射区段(-q 20)。对于具有1个映射区段的读取，n＝500,000;对于具有2个映射区段的读取，n＝175,000;对于具有3个映射区段的读取，n＝75,000;对于具有4个映射区段的读取，n＝30,000;对于5个映射区段的读取，n＝15,000;对于具有6个映射区段的读取，n＝7,000。这表明容易鉴定区段，并且对文库生成方案进行测序生成跨越多个重排区段的读取。

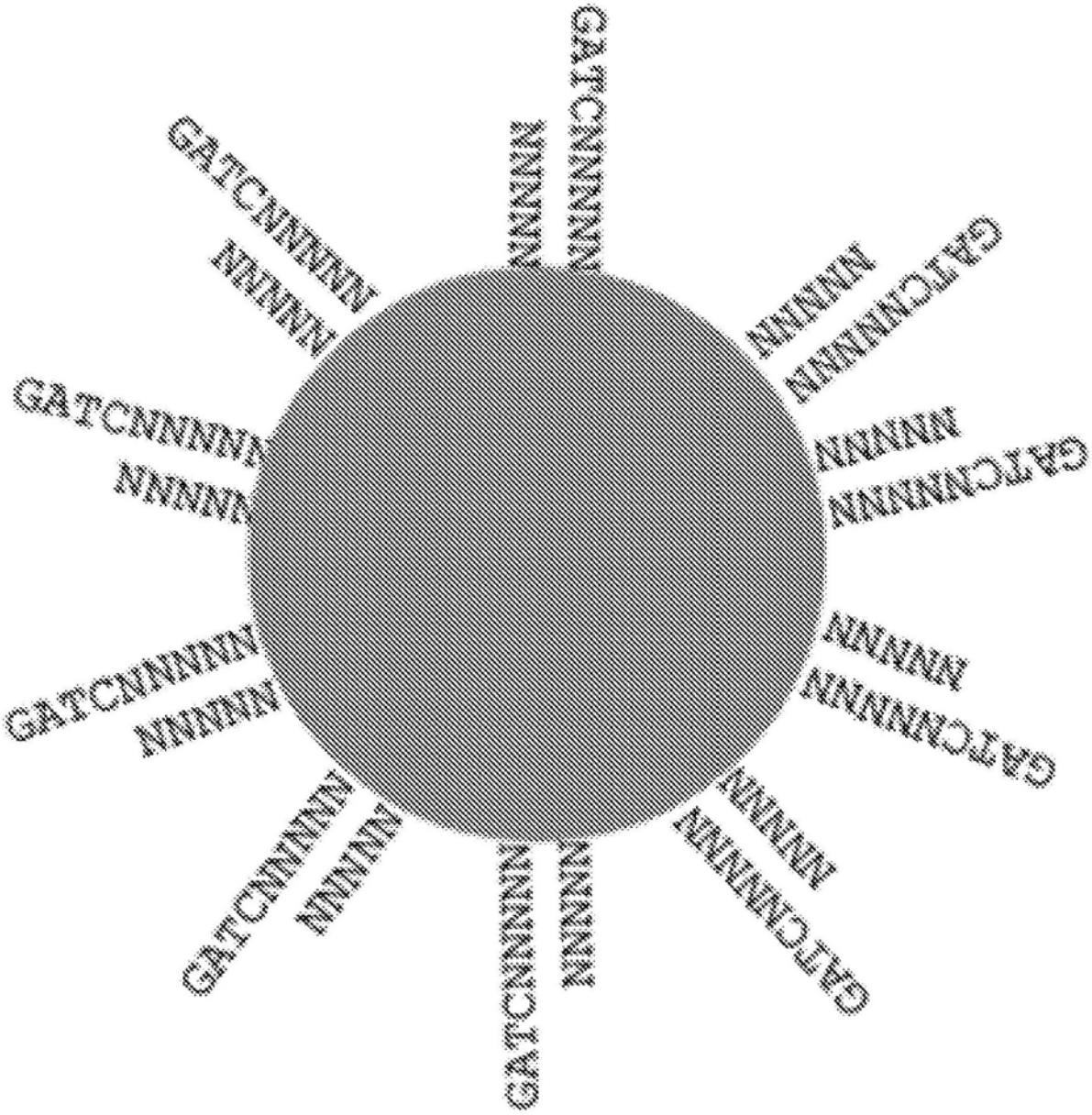
[0273] 表1示出了具有指定数目的映射区段的读取的克隆覆盖度。如其中所指出的，文库生成方案在总区段序列中产生大量的全基因组覆盖，同时产生了有价值的定相信息(如具有两个或更多个映射区段的克隆的数目所示)。由于许多基因组具有重复序列，因此唯一映射区段的数目是重排的文库组成分子中的区段的总数目的低估值。

[0274] 表1.具有X最大数目的映射区段的读取的近似克隆覆盖度。

[0275]

映射区段编号	1	2	3	4	5	6
碱基对	1,3000,000,000	18,000,000,000	9,000,000,000	4,000,000,000	1,500,000,000	500,000,000
基因组倍数(3.Gb)	0.4 X	6 X	3 X	1.2 X	0.5 X	0.5 X

[0276] 在图9A-9B中，可以看到针对样品的由具有X个映射区段的读取跨越的距离的频率分布，其被分类为10kb区间(图9A)和1kb区间(图9B)。该图中的数据再次证实了如下结论：本文公开的文库生成方案产生具有在可识别的接头处连接的多个唯一映射区段的读取，以提供基因组序列信息(通常包括多态性)和相位信息，从而可以相对于彼此对这些多态性进行定相，即使它们在样品基因组中以大于序列读取的长度的距离出现并且由不具有杂合性标志物的序列隔开。



消化的染色质聚集体

图1

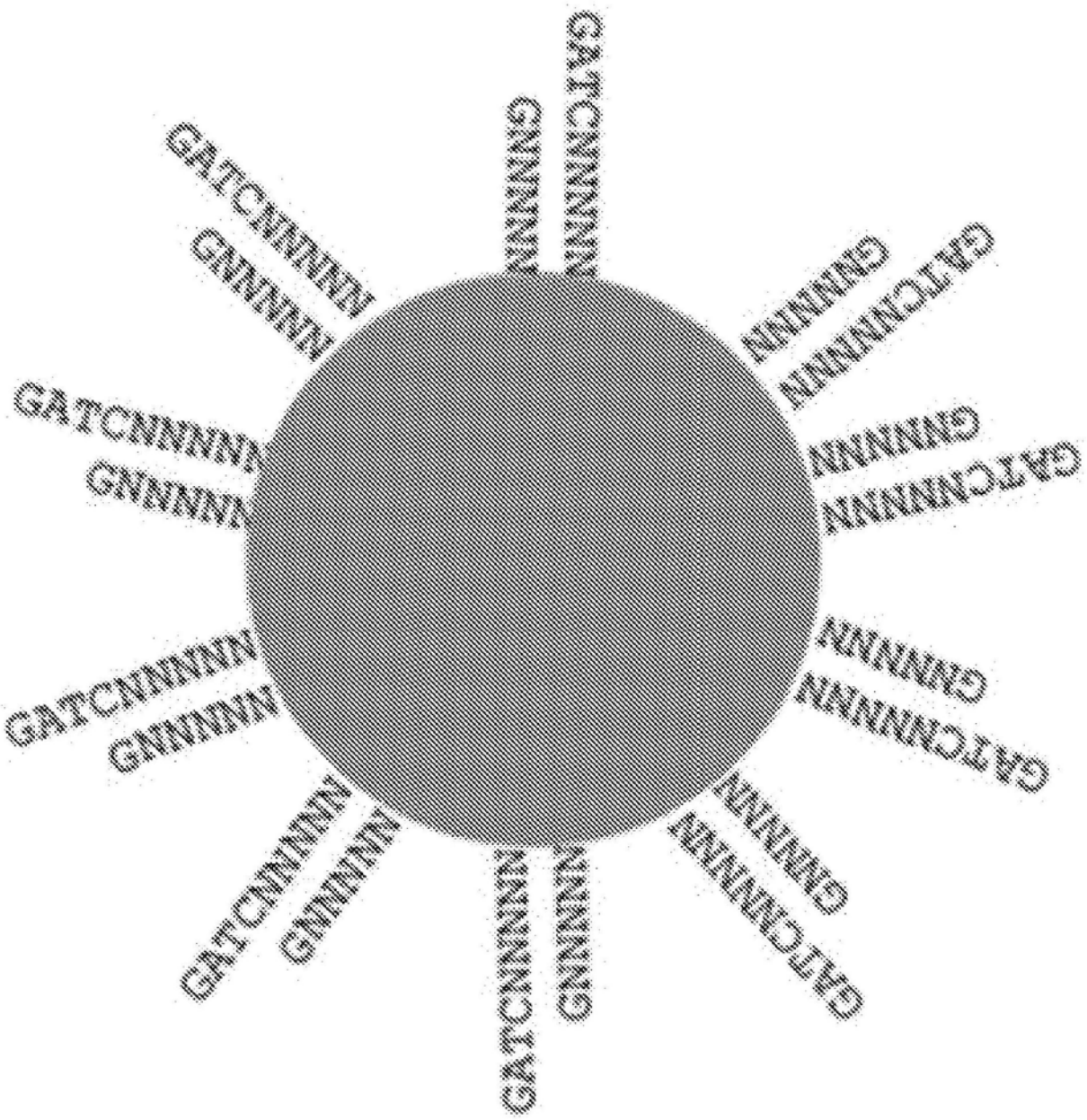

GATCNNNNN
GNNNNN
GATCNNNNN
GNNNNN
GATCNNNNN
GNNNNN
GATCNNNNN
GNNNNN
GATCNNNNN
GNNNNN
GATCNNNNN
GNNNNN
GATCNNNNN
GNNNNN
GATCNNNNN
GNNNNN
GATCNNNNN
GNNNNN
补平一个碱基以使突出端
不能重新退火和连接

图2

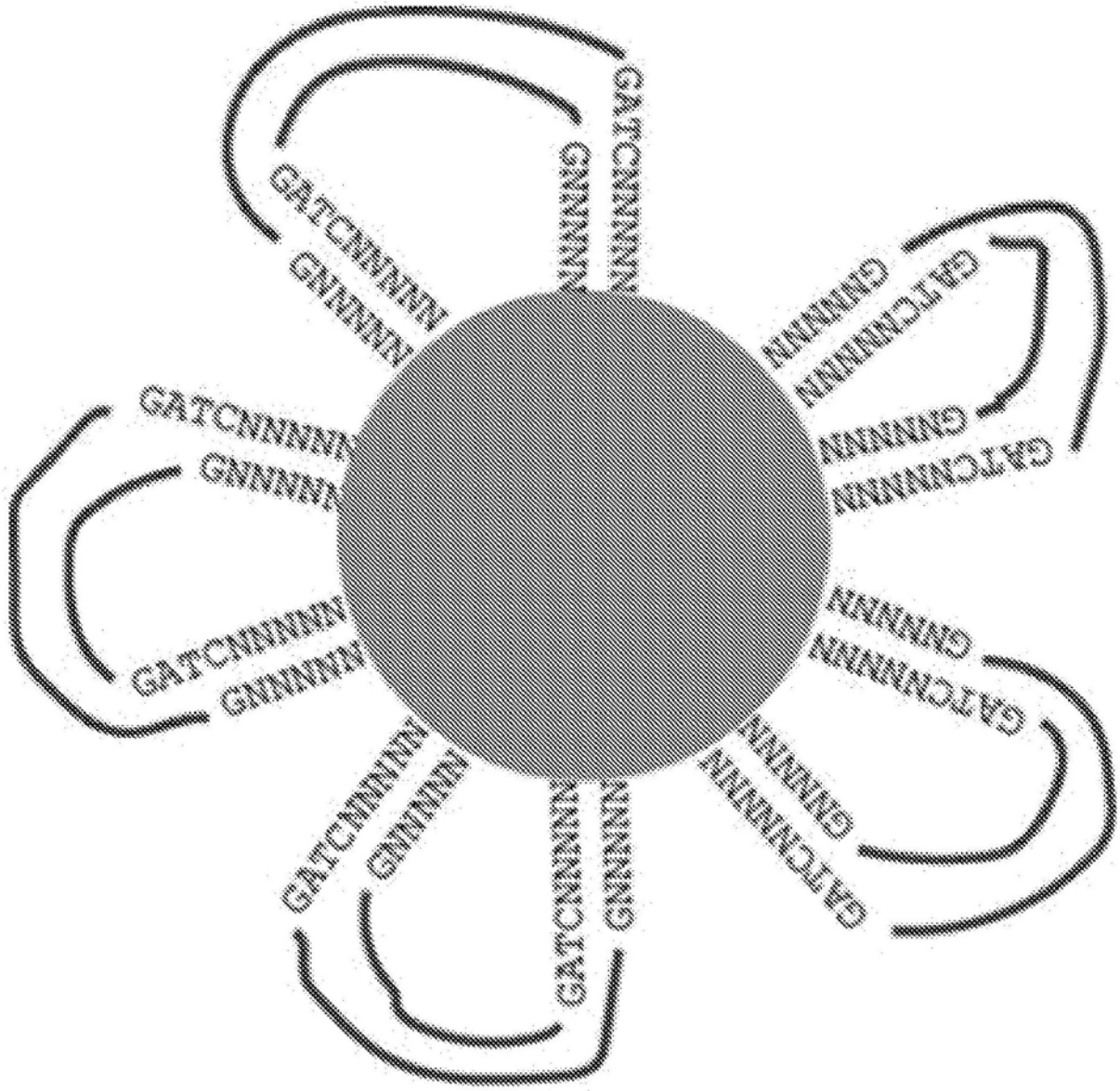

GATCNNNNN
GNNNNN
GATCNNNNN
GNNNNN
GATCNNNNN
GNNNNN
GATCNNNNN
GNNNNN
GATCNNNNN
GNNNNN
GATCNNNNN
GNNNNN
GATCNNNNN
GNNNNN
GATCNNNNN
GNNNNN
GATCNNNNN
GNNNNN
连接通过标点寡核苷酸接合游离末端

图3

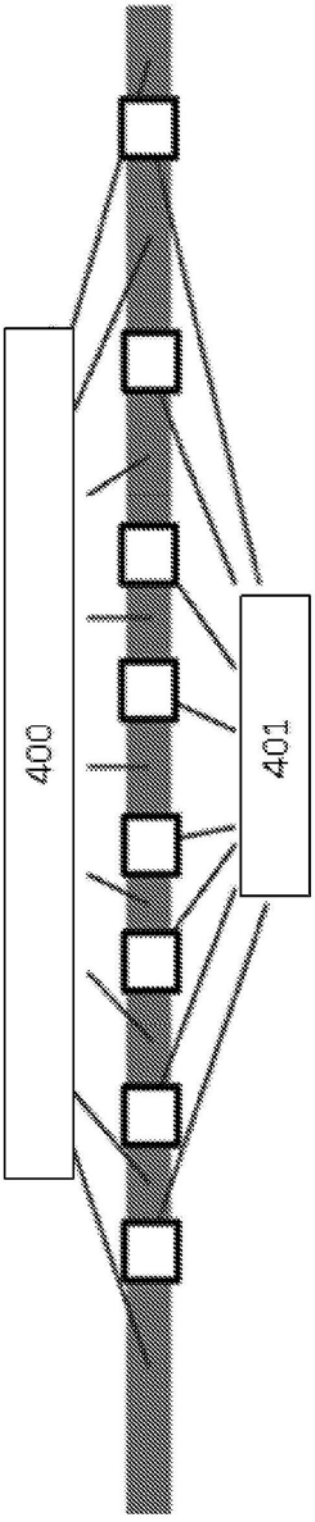


图4

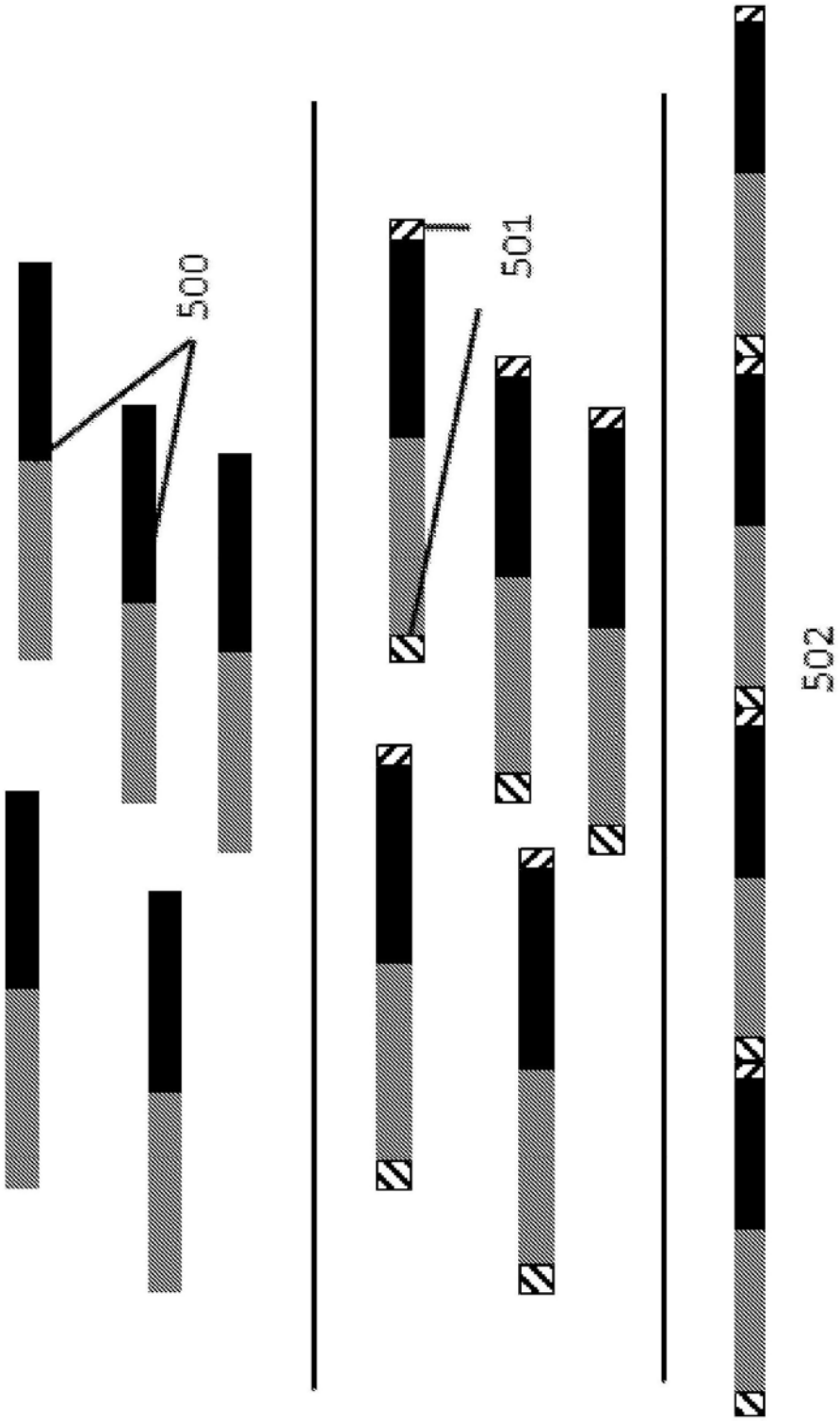


图5

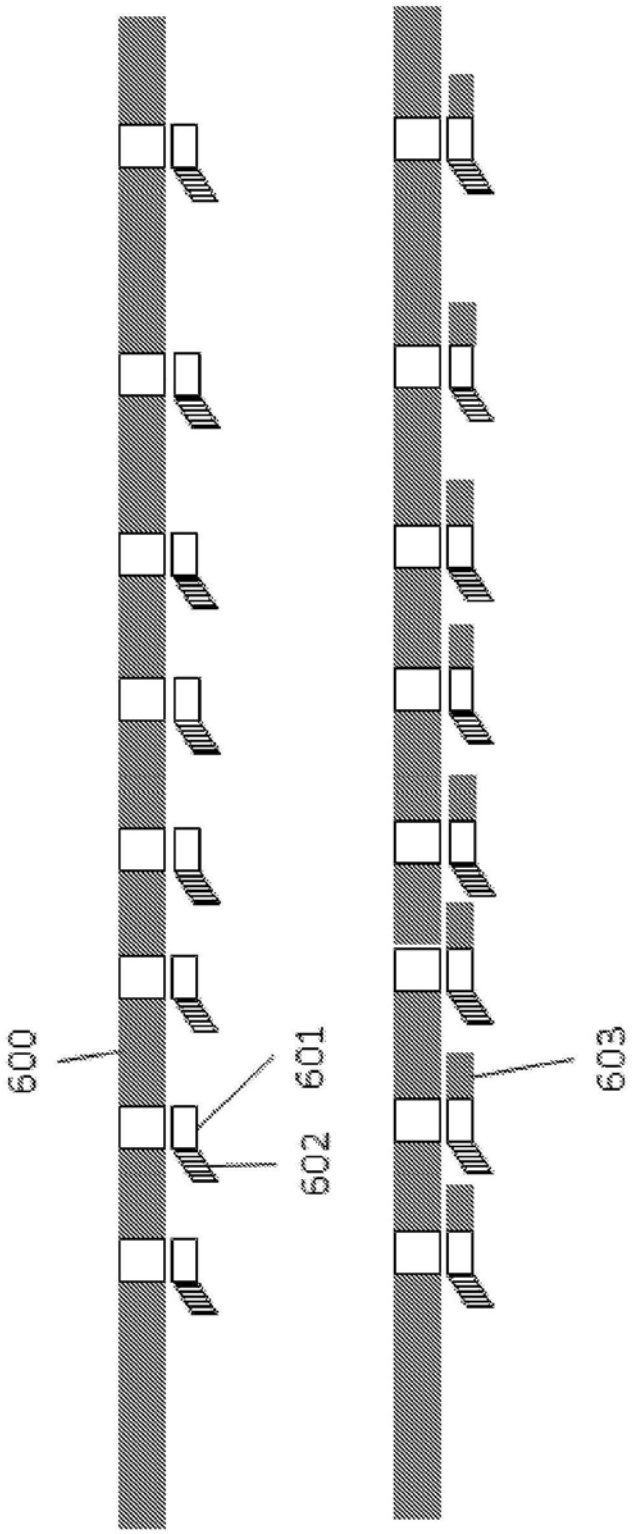


图6

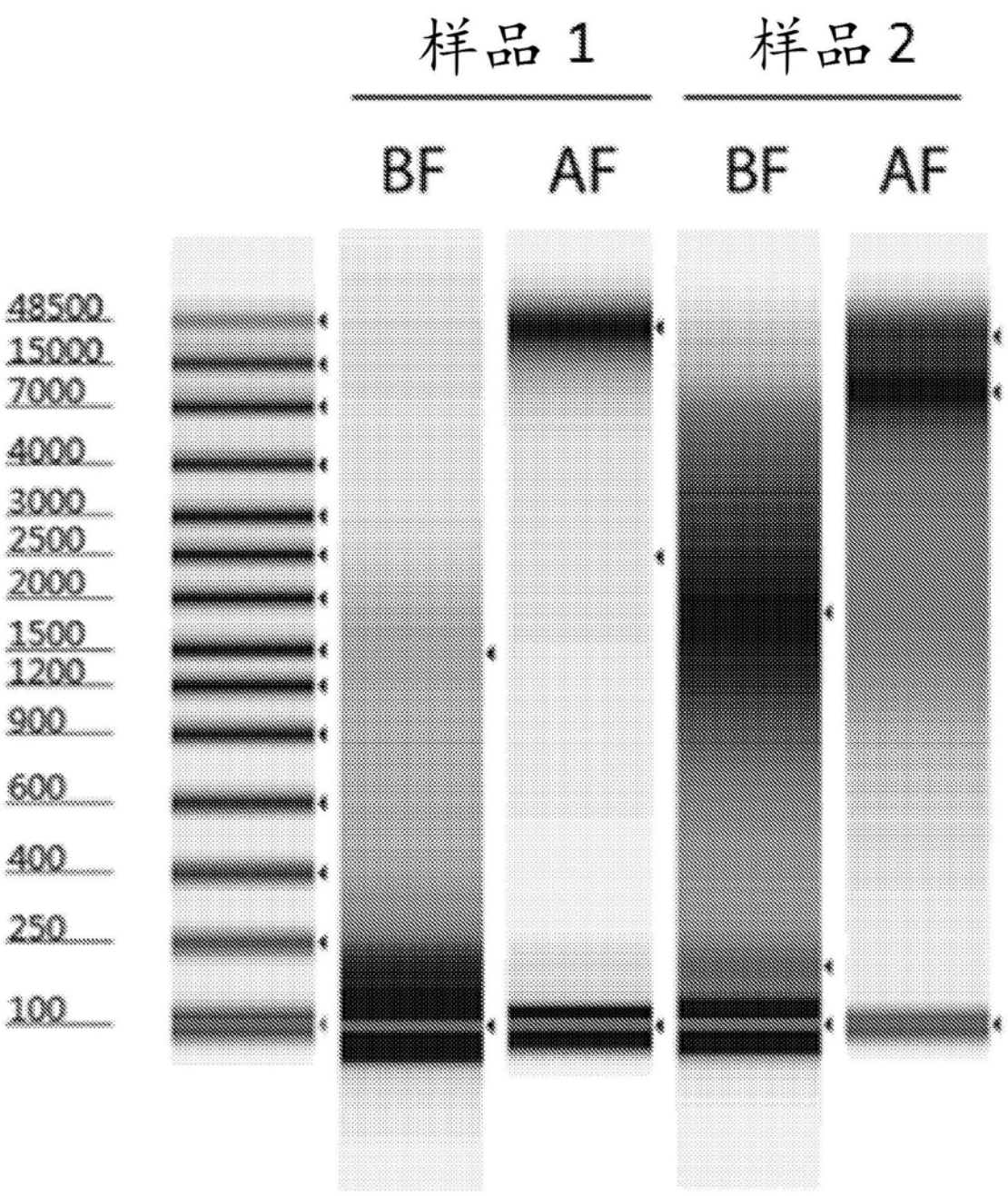

样品 1
样品 2
BF
AF
BF
AF
48500
15000
7000
4000
3000
2500
2000
1500
1200
900
600
400
250
100

图7

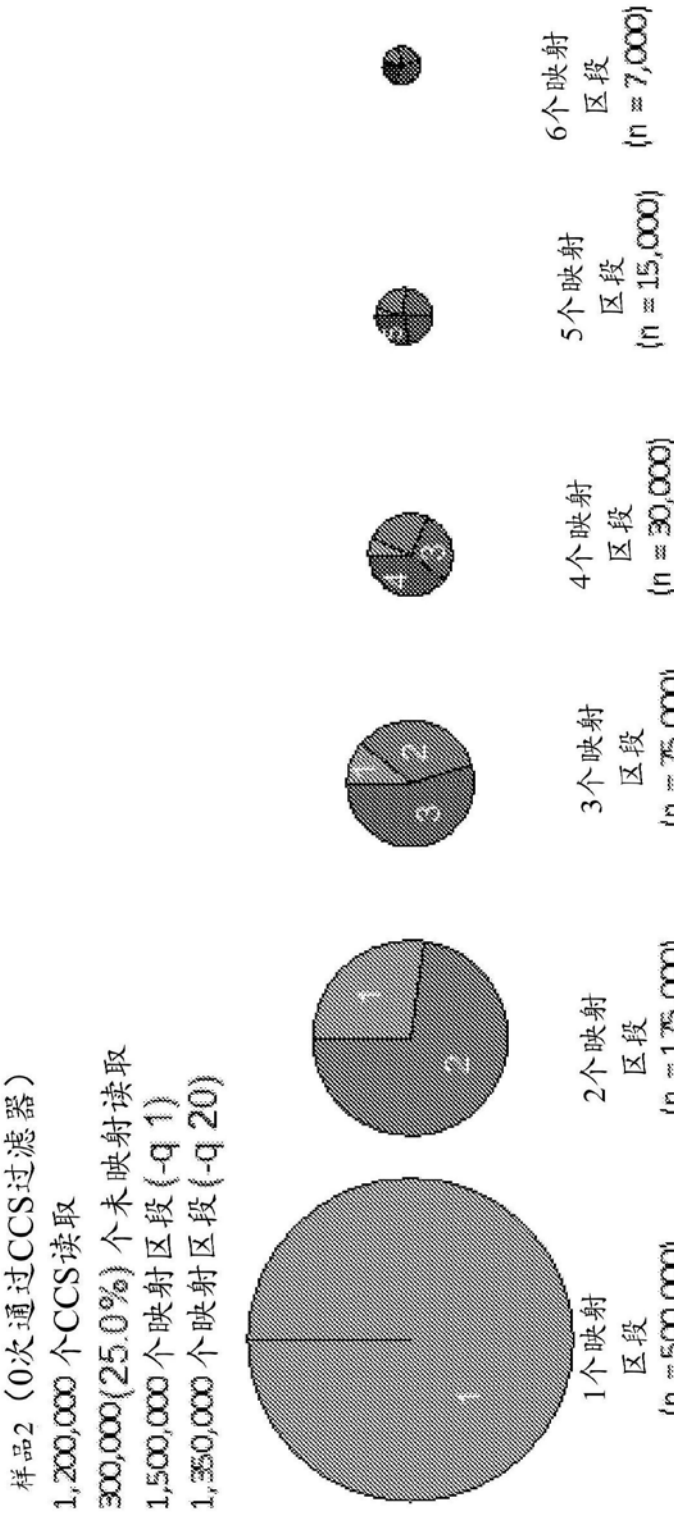

样品2（0次通过CCS过滤器）
1,200,000个CCS读取
300,000(25.0%)个未映射读取
1,500,000个映射区段(-q 1)
1,350,000个映射区段(-q 20)
1个映射区段 {n = 500,000}
2个映射区段 {n = 175,000}
3个映射区段 {n = 75,000}
4个映射区段 {n = 30,000}
5个映射区段 {n = 15,000}
6个映射区段 {n = 7,000}

图8

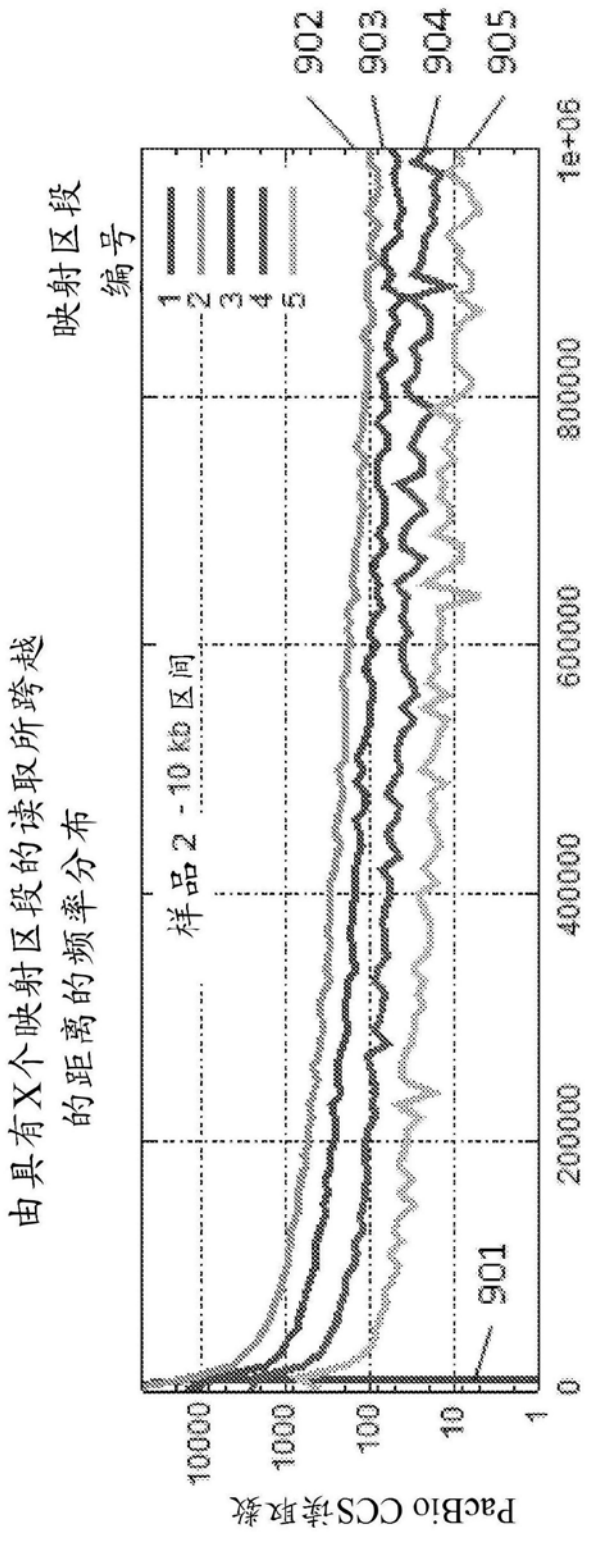

由具有X个映射区段的读取所跨越
的距离的频率分布
映射区段
编号
1
2
3
4
5
样品2 - 10 kb区间
PacBio CCS读取数
10000
1000
100
10
1
0
200000
400000
600000
800000
1e+06
901
902
903
904
905

图 9A

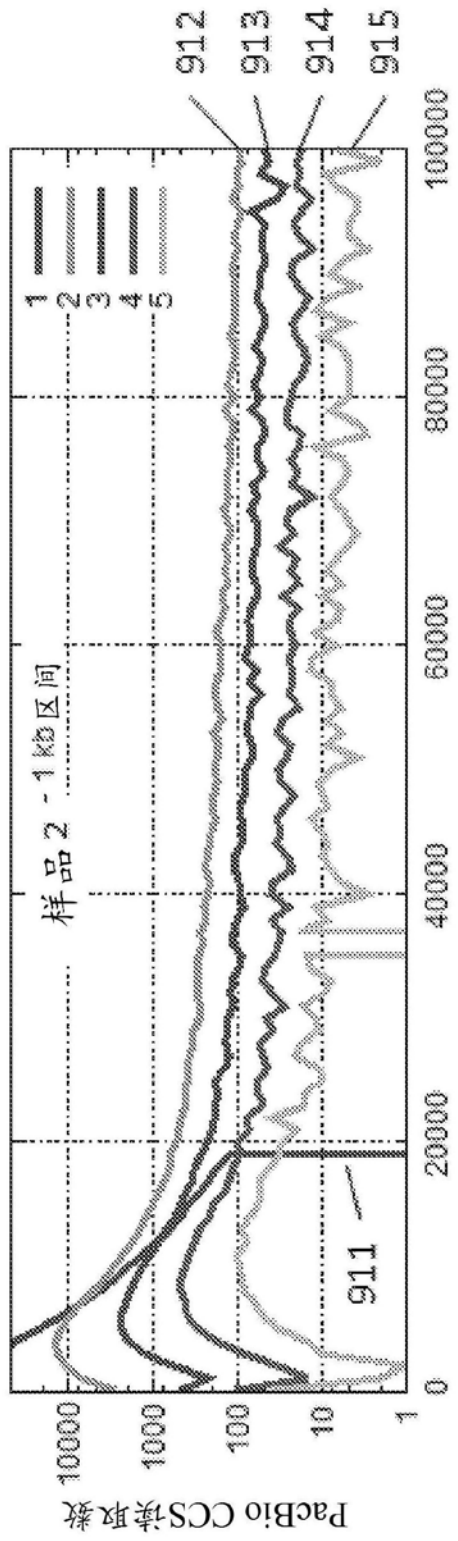

1
2
3
4
5
样品2 - 1 kb区间
PacBio CCS读取数
10000
1000
100
10
1
0
20000
40000
60000
80000
100000
911
912
913
914
915

图 9B

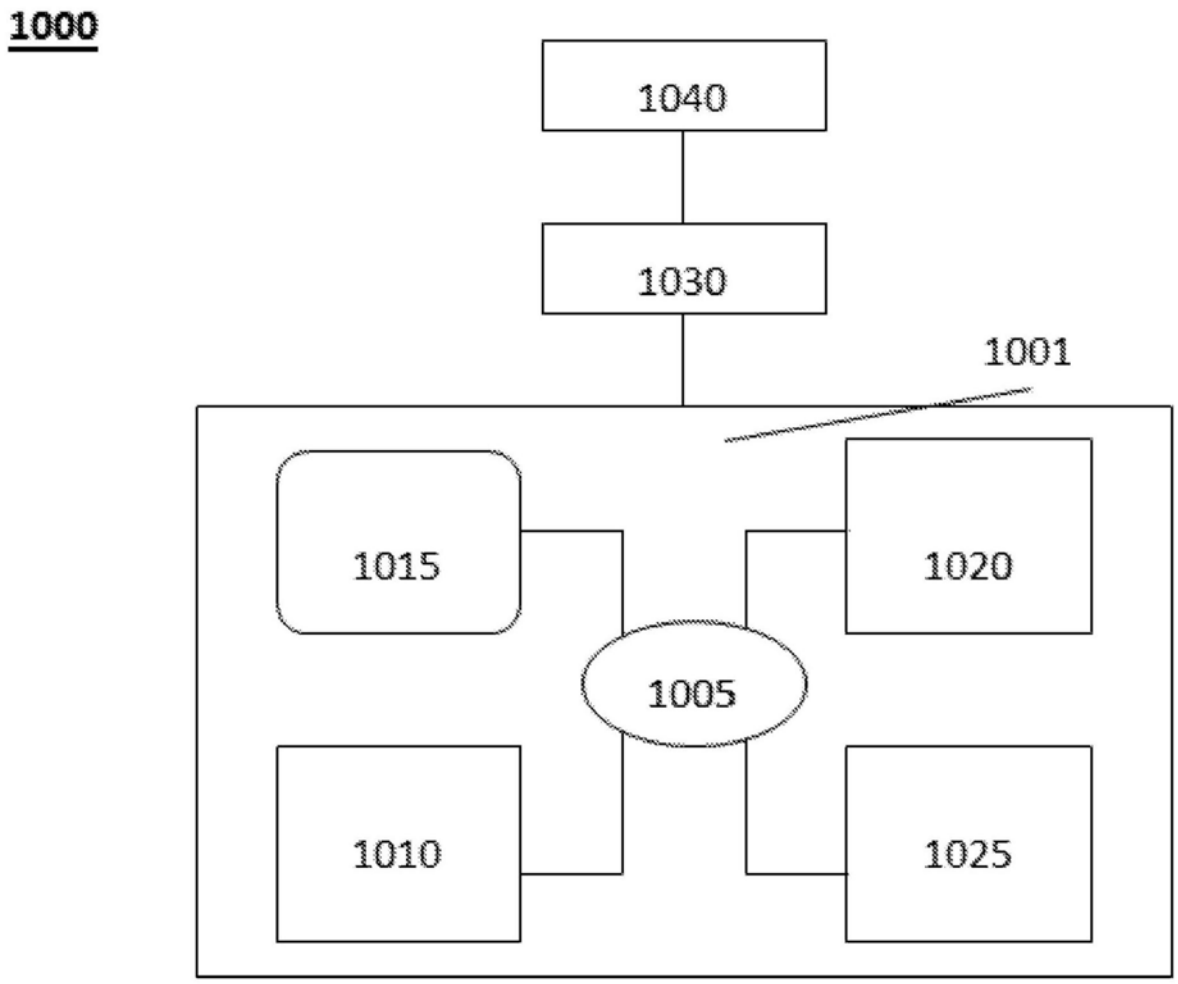


图10