(19)国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 115369161 A
(43)申请公布日 2022.11.22

(21)申请号 202211197364.4

(22)申请日 2016.12.02

(30)优先权数据
62/263532 2015.12.04 US

(62)分案原申请数据
201680070638.X 2016.12.02

(71)申请人 10X 基因组学有限公司
地址 美国加利福尼亚州

(72)发明人 郑新颖 S.萨克索诺夫
M.施纳尔-莱文 K.内斯
R.巴拉瓦伊

(74)专利代理机构 中国专利代理(香港)有限公司 72001
专利代理师 初明明

(51)Int.Cl.
C12Q 1/6869(2018.01)
C12Q 1/6886(2018.01)

权利要求书1页 说明书31页 附图7页

(54)发明名称

用于核酸分析的方法和组合物

(57)摘要

本发明涉及用于分析序列信息同时保留所述序列信息的结构背景和分子背景的方法、组合物和系统。

CN 115369161 A

1.一种分析从福尔马林固定石蜡包埋(FFPE)的组织样品获得的核酸同时保留空间背景的方法,所述方法包括:

(a) 将从FFPE组织样品获得的核酸分配到多个孔中,其中在所述FFPE组织样品中彼此空间接近的核酸被引入到同一孔中;

(b) 将分配的核酸条形码化以形成多个条形码化的核酸,其中在给定的离散孔内的条形码化的核酸各自包含共同的分配特异性的条形码序列,使得条形码序列鉴定来自给定孔的核酸;

(c) 从所述多个条形码化的核酸获得序列信息,其中来自多个条形码化的核酸的序列信息包含分配特异性的条形码序列的序列信息;以及

(d) 将所述多个条形码化的核酸归属到空间接近的区域,其中在FFPE组织样品中源自空间接近的区域的条形码化的核酸包含相同的分配特异性的条形码序列。

2.如权利要求1所述的方法,其中所述条形码化包括用包含条形码序列的引物扩增。

3.如权利要求1所述的方法,其中在分配(a)步骤中分配到同一孔中的至少两个核酸包含不同的序列。

4.如权利要求1所述的方法,进一步包括在(a)之前,使所述FFPE组织样品成像的在先步骤。

5.如权利要求1所述的方法,其中所述FFPE组织样品是癌症组织样品。

6.如权利要求1所述的方法,其中在获得序列信息之前,将不同孔中的条形码化的核酸合并。

7.如权利要求1所述的方法,其中所述序列信息进一步包括涉及从所述FFPE组织样品获得的核酸的信息。

8.如权利要求1所述的方法,进一步包括在(a)之前,从所述FFPE组织样品释放核酸的在先步骤。

9.如权利要求1所述的方法,其中从所述FFPE组织样品获得核酸包含事先应用到样品的核酸标签。

10.如权利要求1所述的方法,其中获得序列信息的步骤包括所述多个条形码化的核酸的高通量测序。

用于核酸分析的方法和组合物

[0001] 相关申请的交叉引用

[0002] 本申请要求2015年12月4日提交的美国临时申请No.62/263,532的权益，该申请出于所有目的通过引用整体并入本文。

[0003] 发明背景

[0004] 多核苷酸测序越来越多地用于诸如肿瘤的遗传筛选和基因分型的医学应用中。许多多核苷酸测序方法依赖于原始样品的样品加工技术，包括多核苷酸的随机片段化。这些加工技术可以提供通量和效率方面的优势，但自这些加工过的样品获得的所得序列信息可能缺乏关于特定序列在含有那些序列的原始核酸分子的更宽线性(二维)序列内的位置的重要背景信息。原始样品的三维空间内的结构背景也因许多样品加工和测序技术而丢失。因此需要保留所鉴定的核酸序列的结构背景和分子背景的测序技术。

[0005] 发明概述

[0006] 因此，本发明提供用于提供保留原始核酸分子的分子背景和结构背景的序列信息的方法、系统和组合物。

[0007] 在一些方面，本公开内容提供分析核酸同时维持结构背景的方法。所述方法包括以下步骤：(a)提供含有核酸的样品，其中所述核酸包含三维结构；(b)将所述样品的部分分离到离散分区，使得所述核酸三维结构的部分也被分离到离散分区；(c)自所述核酸获得序列信息，由此分析核酸，同时维持结构背景。

[0008] 在一些实施方案中，来自获得步骤(c)的序列信息包括鉴定彼此空间接近的核酸。

[0009] 在进一步的实施方案中，来自获得步骤(c)的序列信息包括鉴定彼此空间接近的核酸。

[0010] 在进一步的实施方案中，获得步骤(c)提供关于基因组基因座之间的染色体内和/或染色体间相互作用的信息。

[0011] 在又进一步的实施方案中，获得步骤(c)提供关于染色体构象的信息。

[0012] 在进一步的实施方案中，在分离步骤(b)之前，加工至少一些三维结构以连接在三维结构内彼此接近的核酸的不同部分。

[0013] 在任何实施方案中，核酸在分离步骤(b)之前未从样品中分离。

[0014] 在任何实施方案中，在获得步骤(c)之前，对离散分区内的核酸进行条形编码以形成多个条形码化片段，其中给定离散分区内的片段各自包含共同的条形码，使得条形码鉴定来自给定分区的核酸。

[0015] 在进一步的实施方案中，获得步骤(c)包括选自由以下组成的组的测序反应：短读段长度测序反应和长读段长度测序反应。

[0016] 在一些方面，本公开内容提供分析核酸同时维持结构背景的方法，所述方法包括以下步骤：(a)在样品内形成连接的核酸，使得空间上相邻的核酸区段连接；(b)加工连接的核酸以产生多种连接产物，其中所述连接产物含有空间上相邻的核酸区段的部分；(c)将所述多种连接产物沉积到离散分区中；(d)对所述离散分区内的连接产物进行条形编码以形成多个条形码化片段，其中给定离散分区内的片段各自包含共同的条形码，从而将每个片

段与得到其的连接核酸相关联；(e)从所述多个条形码化片段获得序列信息，从而分析来自所述样品的核酸，同时维持结构背景。

[0017] 在一些方面，本公开内容提供分析核酸同时维持结构背景的方法，所述方法包括以下步骤：(a)在样品内形成连接的核酸，使得空间上相邻的核酸区段连接；(b)将连接的核酸沉积到离散分区中；(c)加工所述连接的核酸以产生多种连接产物，其中所述连接产物含有空间上相邻的核酸区段的部分；(d)对所述离散分区内的连接产物进行条形编码以形成多个条形码化片段，其中给定离散分区内的片段各自包含共同的条形码，从而将每个片段与得到其的连接核酸相关联；(e)从所述多个条形码化片段获得序列信息，从而分析来自所述样品的核酸，同时维持结构背景。

[0018] 在一些方面，本公开内容提供分析核酸同时维持结构背景的方法，所述方法包括以下步骤：(a)使样品内的核酸交联以形成交联的核酸，其中所述交联在空间上相邻的核酸片段之间形成共价键；(b)将所述交联的核酸沉积到离散分区中；(c)加工所述交联的核酸以产生多种连接产物，其中所述连接产物含有所述空间上相邻的核酸区段的部分；(d)自所述多种连接产物获得序列信息，从而分析来自所述样品的核酸，同时维持结构背景。

[0019] 在任何实施方案中，所述样品是福尔马林固定的石蜡样品。

[0020] 在任何实施方案中，所述离散分区包含珠粒。在进一步的实施方案中，所述珠粒为凝胶珠粒。

[0021] 在任何实施方案中，所述样品包括肿瘤样品。

[0022] 在任何实施方案中，所述样品包括肿瘤和正常细胞的混合物。

[0023] 在任何实施方案中，所述样品包括核基质。

[0024] 在任何实施方案中，所述核酸包括RNA。

[0025] 在任何实施方案中，样品中核酸的量小于5ng/ml、10ng/ml、15ng/ml、20ng/ml、25ng/ml、30ng/ml、35ng/ml、40ng/ml、45ng/ml或50ng/ml。

[0026] 在一些方面，本发明提供一种分析核酸同时维持结构背景的方法，其中所述方法包括以下步骤：(a)提供含有核酸的样品；(b)将标签文库施加给所述样品，使得所述样品的不同地理区域接收不同的标签或不同浓度的标签；(c)将所述样品的部分分离到离散分区，使得标签文库的部分和核酸的部分也被分离到离散分区；(d)自所述核酸获得序列信息；以及(e)鉴定所述离散分区中的标签或标签的浓度，由此分析核酸同时维持结构背景。

[0027] 附图简述

[0028] 图1提供根据本文所述的方法的分子背景和结构背景的示意图。

[0029] 图2提供本文所述的方法的示意图。

[0030] 图3示出使用本文公开的方法和组合物进行测定以检测序列信息的典型工作流程。

[0031] 图4提供将核酸样品与珠粒组合并将核酸和珠粒分配到离散液滴中的方法的示意图。

[0032] 图5提供用于对染色体核酸片段进行条形编码和扩增的方法的示意图。

[0033] 图6提供对核酸片段进行条形编码在将序列数据归属到它们的原始源核酸分子中的用途的示意图。

[0034] 图7提供示例性样品制备方法的示意图。

[0035] 发明详述

[0036] 除非另外指示,否则可以采用本领域的技术之内的有机化学、聚合物技术、分子生物学(包括重组技术)、细胞生物学、生物化学以及免疫学的常规技术和描述来实践本发明。所述常规技术包括聚合物阵列合成、杂交、连接、噬菌体展示和使用标记检测杂交。合适技术的具体说明可以参考下文的实例。然而,当然还可以使用其它等效的常规程序。所述常规技术和描述可以在诸如以下的标准实验室手册中找到:Genome Analysis:A Laboratory Manual Series(第I-IV卷),Using Antibodies:A Laboratory Manual,Cells:A Laboratory Manual,PCR Primer:A Laboratory Manual,and Molecular Cloning:A Laboratory Manual(均来自Cold Spring Harbor Laboratory Press),Stryer,L.(1995) Biochemistry(第4版)Freeman,New York,Gait,"Oligonucleotide Synthesis:A Practical Approach"1984,IRL Press,London,Nelson和Cox(2000),Lehninger, Principles of Biochemistry第3版,W.H.FreemanPub.,New York,N.Y.以及Berg等人,(2002)Biochemistry,第5版,W.H.FreemanPub.,NewYork,N.Y.,所有这些出于所有目的通过引用整体并入本文。

[0037] 应注意,除非上下文明确地另外规定,否则在本文和附加的权利要求书中使用的单数形式"一个"、"一种"和"所述"包括多个指示物。因此,例如,对"一种聚合酶"的引用是指一种试剂或所述试剂的混合物,并且对"所述方法"的引用包括对本领域技术人员已知的等价步骤和方法的引用等等。

[0038] 除非另外定义,否则本文所用的所有技术和科学术语都具有与本发明所属领域中的普通技术人员通常所理解相同的含义。出于描述和公开描述于公布中并且可能与目前所描述的发明结合使用的装置、组合物、制剂以及方法学的目的,本文所提及的所有公布均以引用的方式并入本文。

[0039] 如果提供一定范围的值,那么应理解除非上下文另外清楚地规定,否则本发明内包涵所述范围的上限与下限之间的各插入值(至下限单位的十分之一)和在所陈述的范围中的任何其它陈述值或插入值。本发明内还包涵可以独立地包括于这些较小范围中的所述较小范围的上限和下限,从属于所陈述的范围中的任何具体排除的限值。当所述范围包括一个或两个限值时,排除那些所包括的限值之一或两者的范围也包括在本发明中。

[0040] 在以下描述中,阐述大量具体细节以便提供本发明的更详尽的理解。然而,对于本领域技术人员将是清楚的:可以在无一种或多种这些具体细节的情况下实施本发明。在其它情况下,并未描述本领域技术人员所熟知的熟知特征和程序,以便避免混淆本发明。

[0041] 如本文所用,术语"包含"意图是指组合物和方法包括所述要素,但不排除其它要素。"基本上由……组成"在用于定义组合物和方法时应意指排除对组合物或方法具有任何重要意义的其它要素。"由……组成"应意指排除超过要求保护的组合物和实质方法步骤的其它成分的微量要素。由这些过渡术语的各者定义的实施方案在本发明的范围内。因此,意图是所述方法和组合物可以包括额外的步骤和组分(包含)或者可选地包括不重要的步骤和组成(基本上由……组成)或者可选地意图仅包括所述的方法步骤或组成(由……组成)。

[0042] 所有数字标号如pH、温度、时间、浓度和分子量(包括范围)都是近似值,其以0.1的增量变化(+)或(-)。尽管不总是明确地陈述,但是应当理解所有数字标号的前面都加了术语"约"。除了诸如"X+0.1"或"X-0.1"的"X"的小增量之外,术语"约"还包括精确值"X"。尽管

不总是明确地陈述，但是还应当理解本文描述的试剂仅是示例性的，并且所述试剂的等效物是本领域已知的。

[0043] I.概述

[0044] 本公开内容提供用于表征基因材料的方法、组合物和系统。一般来讲，本文所述的方法、组合物和系统提供分析样品的组分同时保留关于那些组分原样在样品中的结构以及分子背景的信息的方法。尽管这里的大部分论述都是关于核酸的分析，但应理解的是，本文讨论的方法和系统可以适用于样品的其它组分，这些组分包括蛋白质和其它分子。

[0045] 脱氧核糖核酸(DNA)是一种线性分子，因此经常用线性维度来描述和评估基因组。然而，染色体不是刚性的，并且两个基因组基因座之间的空间距离不一定总是与它们沿着基因组的线性序列的距离相对应。在三维空间中，由许多兆碱基分隔的区域可以直接相邻。从调控的角度来看，理解基因组基因座之间的长范围相互作用可能是有用的。例如，基因增强子、沉默子和绝缘子元件可能会在广阔的基因组距离上起作用。保留序列读段的结构背景和分子背景的能力提供了理解此类长范围相互作用的能力。

[0046] 如本文所用的“保留结构背景”意指多个序列读段或序列读段的多个部分可归属到样品内那些序列读段的原始三维相对位置。换句话说，序列读段可以与样品内关于该样品中的相邻核酸(以及在一些情况下，相关蛋白质)的相对位置相关联。该空间信息可通过本文讨论的方法获得，即使那些相邻核酸物理上并未位于单个原始核酸分子的线性序列内。参考图1中的示意图：在样品(101)中，序列(104)和(105)位于两个不同的原始核酸分子(分别地，(102)和(103))的线性序列内，但是在样品内空间上彼此接近地定位。本文所述的方法和组合物提供保留关于序列读段的结构背景的信息的能力，并且因此使来自序列(104)和(105)的读段归属到它们在原始样品内在得到那些序列读段的原始核酸分子(102)和(103)上的相对空间接近度。

[0047] 本文讨论的方法和组合物还提供保留分子背景的序列信息。如本文所用的“保留分子背景”意指多个序列读段或序列读段的多个部分可归属到核酸的单个原始分子。尽管该单个核酸分子可以具有各种长度中的任何长度，但在优选的方面，它将是相对长的分子，从而允许保存长范围分子背景。特定地讲，单个原始分子优选比典型的短读段序列长度基本上更长，例如长于200个碱基，并且通常为至少1000个碱基或更长、5000个碱基或更长、10,000个碱基或更长、20,000个碱基或更长、30,000个碱基或更长、40,000个碱基或更长、50,000个碱基或更长、60,000个碱基或更长、70,000个碱基或更长，80,000个碱基或更长，90,000个碱基或更长，或100,000个碱基或更长，并且在一些情况下最多1兆碱基或更长。

[0048] 通常，本文所述的方法包括分析核酸，同时维持结构背景和分子背景。此类分析包括其中提供含有核酸的样品的方法，其中所述核酸含有三维结构。将样品的部分分离到离散分区，使得核酸三维结构的部分也被分离到离散分区--彼此在空间上接近的核酸序列将倾向于被分离到相同分区，因此即使当后来获得的序列读段来自原本不在相同的单独原始核酸分子上的序列时，也保留此空间接近度的三维信息。再次参考图1：如果含有核酸分子102和103和106的样品101被分离到离散分区，使得样品的子集被配置到不同的离散分区中，则由于核酸分子106与102和103之间的物理距离，相较于核酸分子106，核酸分子102和103更可能彼此将被置于相同分区。因此，相同离散分区内的核酸分子是在原始样品中彼此空间接近的分子。从离散分区内的核酸获得的序列信息因此提供一种例如通过核酸测序进

行的分析核酸的方式,并将那些序列读段归属回原始核酸分子的结构背景。

[0049] 在进一步的实例中,结构背景(在本文中也被称为“地理背景”)可以通过使用标签(诸如条形码寡聚核苷酸)来编码样品的地理位置来维持。在一些情况下,这可以包括将编码条形码化序列(诸如,mRNA序列)的集合的病毒文库注射到样品中。条形码通过活动过程(activeprocesses)或通过扩散穿过样品。当样品随后根据本文所述且本领域已知的方法进一步加工时,可以将条形码与结构位置相关以鉴定样品内的来自相同地理位置的核酸序列。在条形码通过活动过程分布在样品中的实例中,具有相同条形码的序列可以在地理上连接和/或通过相同的过程连接。如将理解的,使用标签来编码结构背景的该系统可以单独使用或与利用离散分区的本文所述的方法组合使用以进一步保留结构背景和分子背景。在使用用于编码空间位置的标签和用于鉴定分离到相同离散分区的分子的条形码的实例中,样品本质上被标记或“双条形码化”,其中一组条形码用于鉴定空间位置,而一组条形码是分区特异性的。在所述实例中,两组条形码都可以用来提供信息以保留从样品产生的序列读段的结构背景和分子背景。

[0050] 在一些实例中,从核酸获得的序列信息提供关于基因组基因座之间染色体内和/或染色体间相互作用的信息。在进一步的实例中,序列信息包括关于染色体构象的信息。

[0051] 在进一步的实例中,在分离到离散分区之前,可以加工样品中的核酸以连接其三维结构的不同区域,使得在那些三维结构内彼此接近的序列区域彼此附接。这样,将样品分离到离散分区将会将这些连接的区域分离到相同分区,由此进一步确保保留来自那些核酸的任何序列读段的结构背景。

[0052] 在一些情况下,核酸的连接可以使用本领域已知的用于使空间上接近的分子交联的任何方法来完成。所述交联剂可以包括而不限于烷化剂、顺铂、一氧化二氮、补骨脂素、醛、丙烯醛、乙二醛、四氧化锇、碳二亚胺、氯化汞、锌盐、苦味酸、重铬酸钾、乙醇、甲醇、丙酮、乙酸等等。在具体实例中,使用设计用于分析基因组的三维结构的方案连接核酸,诸如例如Dekker等人,“Capturing chromosome conformation”Science 295:1306-1311(2002)和Berkum等人,J.Vis.Exp.(39),e1869,doi:10.3791/1869(2010)描述的“Hi-C”方案,其各自出于所有目的通过引用整体地并且特别是对于涉及连接核酸分子的所有教导并入本文。所述方案通常包括通过使样品交联使得紧密空间接近的基因组基因座连接来产生分子文库。在进一步的实施方案中,将交联之间的居间DNA环消化掉,然后将序列内区域反交联以添加到文库。消化和反交联步骤可以在将样品分到离散分区的步骤之前发生,或者可以在分离步骤之后在分区内发生。

[0053] 在更进一步的实例中,核酸可以经历标记或条形编码步骤,该步骤为分区内的所有核酸提供共同的条形码。如将理解的,该条形编码可以在有或没有上文讨论的核酸连接/交联步骤的情况下发生。使用本文公开的条形编码技术赋予为基因组区域提供单独的结构背景和分子背景的独特能力-即通过将某些序列读段归属到单独的样品核酸分子并通过变体配位组装实现,以在多个样品核酸分子之中和/或对特定的染色体提供更宽或甚至更长范围的推断背景。如本文所用,术语“基因组区域”或“区域”是指任何限定长度的基因组和/或染色体。例如,基因组区域可以指多于一个染色体之间的关联(即,例如相互作用)。基因组区域也可以涵盖完整的染色体或部分染色体。此外,基因组区域可以包括染色体上的特定核酸序列(即,例如开放阅读框和/或调控基因)或基因间非编码区域。

[0054] 使用条形编码赋予促进区分从样品中提取的总核酸群体的少数组分和多数组分例如用于检测和表征血流中的循环肿瘤DNA的能力的另外优势，并且还降低或消除在任选的扩增步骤期间的扩增偏差。此外，以微流体形式实施赋予在极小样品量和低DNA输入量下工作的能力以及快速加工大量样品分区(液滴)以促进全基因组标记的能力。

[0055] 除了提供从基因组的整个或选定区域获得序列信息的能力之外，本文所述的方法和系统还可以提供基因组材料的其它表征，这些表征包括但不限于单倍型定相、结构变异和拷贝数变异的鉴定，如USSN14/316,383、14/316,398、14/316,416、14/316,431、14/316,447和14/316,463中所述，其出于所有目的且特别是对于涉及基因组材料表征的所有书面描述、附图和工作实施例通过引用整体并入本文。

[0056] 通常，本发明的方法包括如图2说明的步骤，其提供本文中进一步详细讨论的本发明的方法的示意性图。如将理解的，图2中概述的方法是可以根据需要并如本文所述进行改变或修改的示例性实施方案。如图2所示，本文所述的方法可以包括任选的步骤201，其中加工样品核酸以连接空间上彼此接近的核酸。经过或不经过此初步加工步骤(201)，本文所述的方法在大多数实例中将包括其中分配含有核酸的样品的步骤(202)。通常，含有来自所关注的基因组区域的核酸的每个分区将经历产生含有条形码的片段的过程(203)。然后，在测序(205)之前，可以将那些片段汇集(204)。来自(205)的序列读段可以归属到通常归因于分区特异性条形码(203)的原始结构背景和分子背景(206)。每个分区在一些实例中可以包括多于一个核酸，并且在一些情况下将含有数百个核酸分子。可以使用本领域已知的任何方法产生步骤203的条形码化片段-在一些实例中，寡聚核苷酸与样品一起包括在不同分区内。所述寡聚核苷酸可以包含旨在随机引发样品的多个不同区域的随机序列，或者它们可以包含被靶向以引发样品的靶向区域上游的特异性引物序列。在进一步的实例中，这些寡聚核苷酸还含有条形码序列，使得复制过程还条形编码原始样品核酸的所得复制片段。在扩增和条形编码样品中使用这些条形码寡聚核苷酸的特别优雅的方法在USSN 14/316,383;14/316,398;14/316,416;14/316,431;14/316,447;和14/316,463中详细描述，其各自出于所有目的通过引用整体地且特别是对于涉及条形编码和扩增寡聚核苷酸的所有教导并入本文。也包含在分区中的延伸反应试剂如DNA聚合酶、三磷酸核苷、辅因子(例如，Mg^{2+}或Mn^{2+}等)然后使用样品作为模板延伸引物序列以产生与引物退火的模板链的互补片段，并且所述互补片段包括寡聚核苷酸及其相关的条形码序列。将多个引物退火并延伸至样品的不同部分可以产生样品的重叠互补片段的大型汇集物，其各自具有指示产生其的分区的其自己的条形码序列。在一些情况下，这些互补片段本身可以用作由分区中存在的寡聚核苷酸引发的模板以产生补体的补体，所述补体再次包括条形码序列。在进一步的实例中，构造该复制过程，使得当复制第一补体时，其在其末端处或其末端附近产生两条互补序列，以允许形成发夹结构或部分发夹结构，这降低了分子成为产生另外迭代拷贝的基础的能力。本文所述的方法和系统的优势在于将分区-或样品-特异性条形码附接到拷贝的片段保存了测序片段的原始分子背景，从而使得它们归属到它们的原始分区并因此归属到它们的原始样品核酸分子。

[0057] 通常，将样品与在分配步骤之前可释放地附接至珠粒的寡聚核苷酸标签集组合。用于条形编码核酸的方法在本领域中已知并且在本文中描述。在一些实例中，利用如Amini等人，2014,Nature Genetics,Advance Online Publication)中所述的方法，其出于所有

目的通过引用整体地且特别是对于涉及附接条形码或其它寡聚核苷酸标签到核酸的所有教导并入本文。根据本申请中描述的方法和系统加工核酸并对核酸测序的方法也在USSN 14/316,383;14/316,398;14/316,416;14/316,431;14/316,447;和14/316,463中进一步详细描述,其出于所有目的通过引用整体地且特别是对于涉及加工核酸和对基因组材料进行测序和其它表征的所有书面描述、附图和工作实施例并入本文。

[0058] 除了上述工作流程之外,使用包括基于芯片的捕获方法和基于溶液的捕获方法的方法,可以将靶向基因组区域富集、分离(isolate)或分离(separate),即“下拉”,以便进一步分析,特别是测序。此类方法利用与所关注的基因组区域或与在所关注的基因组区域附近或与其相邻的区域互补的探针。例如,在杂交(或基于芯片)捕获中,将含有捕获探针(通常是单链寡聚核苷酸)的微阵列固定到表面上,所述捕获探针具有一起覆盖所关注的区域的序列。将基因组DNA片段化,并且可以进一步进行加工,诸如末端修复,以产生平端和/或添加额外特征物如通用引发序列。这些片段与微阵列上的探针杂交。将未杂交的片段洗掉并将所需片段洗脱或以其它方式在表面上加工以便测序或其它分析,并且因此富集表面上残留的片段群体的含有所关注的靶向区域(例如,包含与捕获探针中所含的那些序列互补的序列的区域)的片段。可以使用本领域已知的任何扩增技术进一步扩增富集的片段群体。用于所述靶向下拉富集方法的示例性方法在2015年10月29日提交的USSN 14/927,297中描述,其出于所有目的通过引用整体地并且特别是对于包括所有书面描述、附图和实施例的涉及靶向下拉富集方法和测序方法的所有教导并入本文。靶向基因组区域的群体可以在上述下拉方法之前通过使用增加那些靶向区域的覆盖度的方法进一步富集。可以例如使用靶向扩增方法来实现所述增加的覆盖度,上述靶向扩增方法包括例如在2015年2月24日提交的USSN62/119,996中描述的那些方法,其出于所有目的通过引用整体地并且特别是对于涉及核酸分子的靶向覆盖的所有教导并入本文。

[0059] 在具体情况下,本文所述的方法包括在测序之前选择性扩增基因组的选定区域的步骤。通常使用本领域已知的方法(包括但不限于PCR扩增)进行的该扩增提供基因组的选定区域的至少1X、10X、20X、50X、100X、200X、500X、1000X、1500X、2000X、5000X或10000X覆盖度,由此提供一定量的核酸以允许那些选定区域的从头测序。在进一步的实施方案中,扩增提供基因组的选定区域的至少1X-20X、50X-100X、200X-1000X、1500X-5000X、5000X-10,000X、1000X-10000X、1500X-9000X、2000X-8000X、2500X-7000X、3000X-6500X、3500X-6000X、4000X-5500X覆盖度。

[0060] 通常通过延伸与基因组的选定区域内或附近的序列互补的引物进行扩增。在一些情况下,使用设计成跨所关注的区域平铺的引物文库-换句话说,引物文库设计成沿基因组的选定区域以特定的距离扩增区域。在一些情况下,选择性扩增利用与沿着基因组的选定区域的每10、15、20、25、50、100、200、250、500、750、1000或10000个碱基互补的引物。在更进一步的实例中,引物的平铺文库设计成捕获距离的混合,所述混合可以是距离的随机混合或经智能设计使得选定区域的特定部分或百分比由不同引物对扩增。例如在2015年4月13日提交的USSN 62/146,834中提供了根据本文所述的方法使用的基因组的靶向覆盖的另外信息,其出于所有目的通过引用整体地且特别是对于涉及基因组的靶向覆盖的所有教导并入本文。

[0061] 通常,本文所述的方法和系统提供用于诸如测序的分析的核酸。使用具有极低测

序错误率和短读段测序技术的高通量的优势的方法获得测序信息。如上所述，核酸的测序通常以保存序列读段或序列读段部分的结构背景和分子背景的方式进行。这意味着多个序列读段或序列读段的多个部分可以归属到相对于原始样品中其它核酸的空间位置(结构背景)以及沿着核酸的单个原始分子的线性序列的序列读段的位置(分子背景)。尽管该单个核酸分子可以具有各种长度中的任何长度，但在优选的方面，它将是相对长的分子，从而允许保存长范围分子背景。特定地讲，单个原始分子优选比典型的短读段序列长度基本上更长，例如长于200个碱基，并且通常为至少1000个碱基或更长、5000个碱基或更长、10,000个碱基或更长、20,000个碱基或更长、30,000个碱基或更长、40,000个碱基或更长、50,000个碱基或更长、60,000个碱基或更长、70,000个碱基或更长，80,000个碱基或更长，90,000个碱基或更长，或100,000个碱基或更长，并且在一些情况下最多1兆碱基或更长。

[0062] 如上所指出，本文所述的方法和系统为更长核酸的短序列读段提供单独的分子背景。如本文所用，单独分子背景是指超出特定序列读段的序列背景，例如与不包括在序列读段本身内的相邻序列或近侧序列有关，并且因此通常将使得它们并不全部或部分地包括在用于配对读段的短序列读段如约150个碱基或约300个碱基的读段中。在特别优选的方面，所述方法和系统为短序列读段提供长范围序列背景。所述长范围背景包括给定序列读段与如下序列读段的关系或键联，所述序列读段彼此之间的距离长于1kb、长于5kb、长于10kb、长于15kb、长于20kb、长于30kb、长于40kb、长于50kb、长于60kb、长于70kb、长于80kb、长于90kb或甚至长于100kb或更长。如将理解的，通过提供长范围单独分子背景，还可以得出单独分子背景内的变体的相位信息，例如，特定长分子的变体根据定义通常将是分相位的。

[0063] 通过提供更长范围的单独分子背景，本发明的方法和系统还提供长得多的推断分子背景(在本文中也称为“长虚拟单分子读段”)。如本文所述的序列背景可以包括作图或提供跨完整基因组序列的不同(通常在千碱基规模)范围的片段的键联。这些方法包括将短序列读段作图到单独的更长分子或连接分子的重叠群，以及对例如具有单独分子的连续确定序列的更长单独分子的大部分进行长范围测序，其中此类确定的序列长于1kb、长于5kb、长于10kb、长于15kb、长于20kb、长于30kb、长于40kb、长于50kb、长于60kb、长于70kb、长于80kb、长于90kb或甚至长于100kb。与序列背景一样，短序列归属到较长核酸(例如，单独的长核酸分子或连接的核酸分子或重叠群的集合)可以包括针对较长核酸段将短序列作图以提供高水平的序列背景以及通过这些较长核酸提供来自短序列的组装序列。

[0064] 此外，虽然可以利用与长单独分子相关联的长范围序列背景，但是具有所述长范围序列背景还允许推断甚至更长范围的序列背景。举一个实例，通过提供上述长范围分子背景，可以鉴定在来自不同原始分子的长序列之中的重叠变体部分，例如相位变体、易位序列等，从而允许那些分子之间的推断键联。此类推断键联或分子背景在本文中称为“推断的重叠群”。在一些情况下，当在相位序列的上下文中讨论时，推断的重叠群可以代表通常的相位序列，例如，其中通过重叠相位变体，可以推断比单独原始分子基本上更长的相位重叠群。这些相位重叠群在本文中称为“相块”。

[0065] 通过用较长单分子读段(例如，上文讨论的“长虚拟单分子读段”)开始，可以得出比原本可以使用短读段测序技术或其它相位测序方法得到的重叠群或相块长的推断重叠群或相块。参见例如公布的美国专利申请No.2013-0157870。特别地讲，使用本文所述的方法和系统，可以获得具有至少约10kb、至少约20kb、至少约50kb的N50的推断的重叠群或相

块长度(其中大于所述N50数值的块长之和是所有块长之和的50%)。在更优选的方面,得到具有至少约100kb、至少约150kb、至少约200kb,并且在许多情况下,至少约250kb、至少约300kb、至少约350kb、至少约400kb,并且在一些情况下,至少约500kb或更长的N50的推断的重叠群或相块长度。在其它情况下,可以获得超过200kb、超过300kb、超过400kb、超过500kb、超过1Mb或甚至超过2Mb的最大相块长度。

[0066] 在一方面,并且结合上文和随后本文描述的任何方法,本文所述的方法和系统提供样品核酸或其片段的区室化、沉积或分区到离散的隔室或分区(在本文中可互换地称为分区),其中每个分区维持其自己的内含物与其它分区的内含物的分离。可以将独特的标识符如条形码预先、随后或同时递送至容纳区室化或分区样品核酸的分区,以允许将特征如核酸序列信息随后归属到包括在特定区室内的样品核酸,且特别是可以原始沉积到分区中的连续样品核酸的相对长的段。该随后归属进一步允许归属到原始样品中那些样品核酸的原始结构背景,因为在原始样品的三个维度内彼此靠近的核酸将更可能沉积到相同的分区。因此,序列读段归属到分区(以及那些分区内所含的核酸)不仅提供关于沿着得到所述序列读段的原始核酸分子的线性位置的分子背景,而且还提供鉴定在原始样品的三维背景中彼此紧密空间接近的核酸的序列读段的结构背景。

[0067] 本文所述的方法中利用的样品核酸通常代表待分析的总体样品的多个重叠部分,例如整个染色体、外显子组或其它大基因组部分。这些样品核酸可以包括全基因组、单独染色体、外显子组、扩增子或任何各种不同的所关注核酸。通常分配样品核酸使得核酸以连续核酸分子的相对长的片段或段存在于分区中。通常,样品核酸的这些片段可以长于1kb、长于5kb、长于10kb、长于15kb、长于20kb、长于30kb、长于40kb、长于50kb、长于60kb、长于70kb、长于80kb、长于90kb或甚至长于100kb,这容许上述较长范围的结构背景和分子背景。

[0068] 样品核酸通常也以一定的水平分配,由此给定的分区具有包括基因组基因座的两个重叠片段的极低概率。这通常通过在分配过程期间以低输入量和/或浓度提供样品核酸来完成。结果,在优选的情况下,给定的分区可以包括起始样品核酸的许多长但不重叠的片段。不同分区中的样品核酸然后与独特标识符相关联,其中对于任何给定的分区,其中所含的核酸具有相同的独特标识符,但是其中不同的分区可以包括不同的独特标识符。此外,由于分配步骤将样品组分配置到体积非常小的分区或液滴中,所以应理解,为了实现如上所述的期望配置,无需进行样品的大量稀释,如将在更高容量过程中如在管或多孔板的各孔中所需要的。另外,由于本文所述的系统采用如此高水平的条形码多样性,因此可以如上文所提供在较高数量的基因组等同物之中配置多样的条形码。特定地讲,先前描述的多孔板方法(参见例如美国公开的申请No.2013-0079231和2013-0157870)通常仅用一百至几百种不同的条形码序列操作,并且采用其样品的有限稀释过程以便能够将条形码归属到不同的细胞/核酸。因此,它们通常将用远少于100个细胞操作,这通常将提供近似1:10且肯定远高于1:100的基因组:(条形码类型)比率。另一方面,本文所述的系统由于高水平的条形码多样性如超过10,000、100,000、500,000等多样条形码类型而可以以近似地1:50或更低、1:100或更低、1:1000或更低、或甚至更小比率的基因组:(条形码类型)比率操作,同时还允许载入更高数量的基因组(例如,近似地每次测定大于100个基因组,每次测定大于500个基因组,每次测定1000个基因组,或甚至更多),同时仍提供每个基因组远远提高的条形码多样性。

[0069] 在进一步的实例中，与分成离散分区的样品部分一起包括的寡聚核苷酸可以包含至少第一区域和第二区域。第一区域可以是条形码区域，其在给定分区内的寡聚核苷酸之间可以是基本相同的条形码序列，但是在不同分区之间可以是并且在大多数情况下是不同的条形码序列。第二区域可以是可以用于在分区内引发在样品内的核酸的N-聚体(随机N-聚体或设计成靶向特定序列的N-聚体)。在一些情况下，在N-聚体设计成靶向特定序列的情况下，其可以被设计成靶向特定的染色体(例如，染色体1、13、18或21)或染色体的区域，例如外显子或其它靶向区域。在一些情况下，N-聚体可以被设计成靶向特定基因或基因区域，诸如与疾病或病症(例如癌症)相关联的基因或区域。在分区内，扩增反应可以使用第二N-聚体进行以沿着核酸的长度在不同地方引发核酸样品。作为扩增的结果，每个分区可以含有附接到相同或近乎相同的条形码并且可以代表每个分区中核酸的重叠的较小片段的核酸的扩增产物。条形码可以充当预示起源于相同分区且因此也潜在地起源于核酸的相同链的核酸集的标志物。扩增后，可以将核酸汇集、测序并使用测序算法比对。由于较短的序列读段可以凭借其相关条形码序列进行比对并归属到样品核酸的单个长片段，因此该序列上的所有鉴定的变体可以归属到单个原始片段和单个原始染色体。另外，通过比对跨多个长片段的多个共定位变体，可以进一步表征该染色体贡献。因此，然后可以绘制关于特定基因变体的定相的结论，如可以跨长范围基因组序列进行分析-例如，跨基因组的表征欠佳的区域的段的序列信息的鉴定。所述信息也可以用于鉴定单元型，其通常是驻留在相同核酸链或不同核酸链上的一组特定的基因变体。拷贝数变化也可以以此方式鉴定。

[0070] 所描述的方法和系统提供优于当前的核酸测序技术及其相关的样品制备方法的显著优势。全体样品制备和测序方法倾向于主要鉴定和表征样品中的多数组分，并且不旨在鉴定和表征构成所提取样品中总DNA的小百分比的少数组分，例如来自基因组的表征欠佳或高度多态性的区域的由一种染色体贡献的基因材料、或来自一个或几个细胞的材料、或在血流中循环的片段化肿瘤细胞DNA分子。本文所述的方法包括增加来自这些少数组分的基因材料的选择性扩增方法，并且保留该基因材料的分子背景的能力进一步提供这些组分的基因表征。所描述的方法和系统还为检测存在于较大样品内的群体提供显著的优势。因此，它们对于评估单体型和拷贝数变化特别有用-本文公开的方法还可用于提供基因组区域的序列信息，所述区域表征欠佳或者由于在样品制备期间引入的偏差而在核酸靶标群体中代表欠佳。

[0071] 使用本文公开的条形编码技术赋予为给定基因标志物集提供单独的分子背景的独特能力，即将给定基因标志物集(与单个标志物相反)归属到单独的样品核酸分子，并且通过变体配位组装，以在多个样品核酸分子之中和/或对特定的染色体提供更宽或甚至更长范围的推断单独结构背景和分子背景。这些基因标志物可以包括特定的遗传基因座，例如变体，诸如SNP，或者它们可以包括短序列。此外，使用条形编码赋予促进区分从样品中提取的总核酸群体的少数组分和多数组分例如用于检测和表征血流中的循环肿瘤DNA的能力的附加优势，并且还降低或消除在任选的扩增步骤期间的扩增偏差。此外，以微流体形式实施赋予在极小样品量和低DNA输入量下工作的能力以及快速加工大量样品分区(液滴)以促进全基因组标记的能力。

[0072] 如前所述，本文所述的方法和系统的优势在于它们可以通过使用无所不在的短读段测序技术来获得期望的结果。所述技术具有易于获得并且在研究界用充分表征且高效的

方案和试剂系统广泛分散的优势。这些短读段测序技术包括可得自例如Illumina,Inc.(GAIIx、NextSeq、MiSeq、HiSeq、X10)、Thermo-Fisher的Ion Torrent部门(Ion Proton和IonPGM)的那些,焦磷酸测序方法以及其它。

[0073] 特别有利的是,本文所述的方法和系统利用这些短读段测序技术并且在其相关的低错误率和高通量下进行。特别地讲,本文所述的方法和系统获得如上所述的期望单独分子读段长度或背景,但所述期望单独分子读段长度或背景具有短于1000bp、短于500bp、短于300bp、短于200bp、短于150bp或甚至更短的单独测序读段(排除配偶对延伸);并且对于所述单独分子读段长度具有小于5%、小于1%、小于0.5%、小于0.1%、小于0.05%、小于0.01%、小于0.005%或甚至小于0.001%的测序错误率。

[0074] II.工作流程概述

[0075] 在一个示例性方面,本公开内容中描述的方法和系统提供将样品沉积或分配到离散分区,其中每个分区维持其自己的内含物与其它分区中的内含物的分离。如本文进一步详细讨论的,样品可以包括来源于患者的样品,诸如细胞或组织样品,其可以含有核酸,并且在某些情况下还可以含有相关蛋白质。在具体方面,用于本文所述方法的样品包括福尔马林固定石蜡包埋(FFPE)细胞和组织样品等等,以及其中样品降解风险高的任何其它样品类型。

[0076] 如本文所用,分区是指可以包括各种不同形式的器皿或容器,例如孔、管、微米或纳米孔、通孔等。然而,在优选的方面,分区可在流体物流内流动。这些容器可以由例如具有围绕内部流体中心或核心的外部屏障的微胶囊或微囊泡组成,或者它们可以是能够将材料夹带和/或保留在其基质内的多孔基质。然而,在优选的方面,这些分区可以包括在非水性连续相如油相内的水性流体液滴。例如在2013年8月13日提交的美国专利申请No.13/966,150中描述了各种不同的容器。同样,用于在非水性或油连续相中产生稳定液滴的乳液体系在例如公开的美国专利申请No.2010-0105112中详细描述。在某些情况下,微流体通道网络特别适合如本文所述产生分区。所述微流体装置的实例包括2014年4月4日提交的临时美国专利申请No.61/977,804中详细描述的那些,其全部公开内容出于所有目的通过引用整体地并入本文。替代机制也可以用于分配单独的细胞,包括多孔膜,细胞的水性混合物经所述多孔膜挤出到非水性流体中。所述系统通常可以从例如Nanomi,Inc.得到。

[0077] 在乳液中的液滴的情况下,将样品材料分配到离散分区中通常可以通过使水性含样品的物流流进接合部,分配流体的非水性物流如氟化油也流入所述接合部中,使得在流动物流分配流体内产生水性液滴,其中所述液滴包括样品材料。如下所述,分区如液滴通常还包括共分区的条形码寡聚核苷酸。任何特定分区内样品材料的相对量可以通过控制系统的多种不同参数来调节,所述参数包括例如水性物流中样品的浓度、水性物流和/或非水性物流的流速等等。本文描述的分区通常通过极小的体积表征。例如,在基于液滴的分区的情况下,液滴可以具有小于1000pL、小于900pL、小于800pL、小于700pL、小于600pL、小于500pL、小于小于400pL、小于300pL、小于200pL、小于100pL、小于50pL、小于20pL、小于10pL或甚至小于1pL的总体积。在与珠粒共分配的情况下,如将理解的,分区内的样品流体体积可以小于上述体积的90%、小于80%、小于70%、小于60%、小于50%、小于40%、小于30%、小于20%或甚至小于上述体积的10%。在一些情况下,使用低反应体积分区在用非常少量的起始试剂如输入核酸进行反应中特别有利。用于用低输入核酸分析样品的方法和系统在

2014年6月26日提交的美国临时专利申请No.62/017,580(代理人案号43487-727.101)中提出,其全部公开内容通过引用整体并入本文。

[0078] 在涉及经受降解和/或含有低浓度的所关注组分的样品的情况下,可以在分配之前或在分区内进一步加工样品以进一步释放核酸和/或任何相关蛋白质以便进一步分析。例如,通常使用本领域已知的方法提取FFPE样品中所含的核酸。为了分离更长的核酸分子,还可以通过添加有机催化剂来加工所述样品以除去甲醛加合物(参见例如Karmakar等人,(2015),Nature Chemistry,DOI:10.1038/NCHEM.2307,其通过引用整体地且特别是对于涉及FFPE样品的处理和加工的所有教导并入本文)。

[0079] 一旦将样品引入其各自的分区中,分区内的样品核酸可以经受扩增以增加用于后续应用(诸如本文所述且本领域已知的测序方法)的核酸量。在某些实施方案中,该扩增用针对基因组序列的不同部分的引物文库进行,使得所得扩增产物代表来自原始核酸分子的子部分的序列。在关注选择基因组区域的实施方案中,该扩增可以包括一轮或多轮的选择性扩增,使得与基因组的其它区域相比,对于靶向覆盖所关注的基因组的区域以更高的比例存在(尽管,如将理解的,也可以扩增基因组的那些其它区域,但扩增程度较小,因为它们对于从头覆盖而言不受关注)。在某些实施方案中,扩增提供基因组的全部或选定区域的至少1X、2X、5X、10X、20X、30X、40X或50X覆盖度。在进一步的实施方案中,扩增分区内的所有核酸,但以靶向方式扩增选定的基因组区域,使得与基因组的其它部分相比,至少1-5、2-10、3-15、4-20、5-25、6-30、7-35、8-40、9-45或10-50倍以上的扩增子是由这些选定的基因组区域产生。

[0080] 与上述扩增同时或在其之后,分区内的核酸(或其片段)提供有独特的标识符,使得在表征那些核酸时,它们可以归属为自其各自起源得到。因此,样品核酸通常与独特标识符(例如,条形码序列)共分配。在特别优选的方面,独特标识符以包含可以附接到那些样品的核酸条形码序列的寡聚核苷酸的形式提供。将寡聚核苷酸分区,使得在给定分区中的寡聚核苷酸之间,其中所含的核酸条形码序列是相同的,但是在不同分区之间,寡聚核苷酸可以并且优选具有不同的条形码序列。在示例性方面,只有一个核酸条形码序列将与给定分区相关联,尽管在一些情况下,可能存在两个或更多个不同的条形码序列。

[0081] 核酸条形码序列通常将在寡聚核苷酸的序列内包括6至约20个或更多个核苷酸。这些核苷酸可以是完全连续的,即在单段相邻核苷酸中,或者它们可以分离成由一个或多个核苷酸分离的两个或多个分离的子序列。通常,分离的子序列长度通常可以是约4至约16个核苷酸。

[0082] 共分配的寡聚核苷酸通常还包含可用于加工分配的核酸的其它功能序列。这些序列包括例如靶向或随机/通用扩增引物序列,其用于扩增来自分区内单独核酸的基因组DNA,同时附接相关的条形码序列,测序引物,杂交或探测序列,例如用于鉴定序列的存在或用于下拉条形码化的核酸,或许多其它潜在的功能序列中的任一种。此外,寡聚核苷酸和相关条形码及其它功能序列连同样品材料的共分配在例如在USSN 14/175,935;14/316,383;14/316,398;14/316,416;14/316,431;14/316,447;和14/316,463中描述,其出于所有目的通过引用整体地且特别是对于涉及加工核酸和对基因组材料进行测序和其它表征的所有书面描述、附图和工作实施例并入本文。

[0083] 简而言之,在一种示例性方法中,提供珠粒,每个珠粒都可以包括大量的可以释放

地附接到珠粒的上述寡聚核苷酸，其中附接到特定珠粒的所有寡聚核苷酸可以包括相同的核酸条形码序列，但是其中大量的多样化条形码序列的数量可以跨使用的珠粒群体代表。通常，珠粒群体可以提供可以包括至少1000个不同的条形码序列、至少10,000个不同的条形码序列、至少100,000个不同的条形码序列，或者在一些情况下，至少1,000,000个不同的条形码序列的多样化条形码序列文库。另外，每个珠粒通常可以提供有大量附接的寡聚核苷酸分子。特定地讲，包括单独珠粒上的条形码序列的寡聚核苷酸分子的数量可以是至少约10,000个寡聚核苷酸、至少100,000个寡聚核苷酸分子、至少1,000,000个寡聚核苷酸分子、至少100,000,000个寡聚核苷酸分子，并且在一些情况下至少10亿个寡聚核苷酸分子。

[0084] 在对珠粒施加特定刺激时，寡聚核苷酸可以从珠粒释放。在一些情况下，所述刺激可以是光刺激，例如通过裂解光不稳定的键，可以释放寡聚核苷酸。在一些情况下，可以使用热刺激，其中珠粒环境的温度升高可以导致键裂解或寡聚核苷酸从珠粒的其它释放。在一些情况下，可以使用化学刺激来裂解寡聚核苷酸与珠粒的键，或以其它方式可以引起寡聚核苷酸从珠粒释放。

[0085] 根据本文所述的方法和系统，包括附接的寡聚核苷酸的珠粒可以与单独样品共分配，使得单个珠粒和单个样品包含在单独分区内。在一些情况下，在期望单珠粒分区的情况下，可能需要控制流体的相对流速，使得平均地讲，分区含有小于1个珠粒/分区，以确保那些被占用的分区主要被单一地占用。同样，可能希望控制流速以提供更高百分比的分区被占用，例如允许仅有小百分比的未占分区。在优选的方面，控制流动和通道结构以确保有期望数量的单一占用分区，所述数量小于某一水平的未占用分区并且小于某一水平的多重占用分区。

[0086] 图3图示用于对样品核酸进行条形编码和随后测序的一种特定示例性方法。首先，可以从来源获得包含核酸的样品，300，并且还可以获得条形码化珠粒集，310。珠粒优选连接至含有一个或多个条形码序列的寡聚核苷酸以及引物如随机N-聚体或其它引物。优选地，条形码序列可例如通过裂解在条形码和珠粒之间的键或通过降解下面的珠粒以释放条形码或两者的组合而从条形码化珠粒释放。例如，在某些优选的方面，条形码化珠粒可以被试剂如还原剂降解或溶解以释放条形码序列。在该实例中，将包含核酸305、条形码化珠粒315和任选地其它试剂如还原剂320的少量样品组合并进行分配。举例来说，所述分配可以包括将组分引入液滴产生系统，诸如微流体装置325。在微流体装置325的帮助下，可以形成油包水乳液330，其中乳液含有包含样品核酸305、还原剂320和条形码珠粒315的水性液滴。还原剂可以溶解或降解条形码化珠粒，从而在液滴内从珠粒释放具有条形码的寡聚核苷酸和随机N-聚体，335。随机N-聚体然后可以引发样品核酸的不同区域，扩增后产生样品的扩增拷贝，其中每个拷贝用条形码序列标记340。优选地，每个液滴含有一组含有相同条形码序列和不同随机N-聚体序列的寡聚核苷酸。随后，使乳液破乳345并且可以经由例如扩增方法添加额外序列(例如，有助于特定测序方法的序列、额外条形码等)350(例如，PCR)。然后可以进行测序355，并且应用算法来解释测序数据360。测序算法通常能够例如进行条形码的分析以比对测序读段和/或鉴定特定序列读段所属的样品。此外，并且如本文所述，这些算法还可以进一步用于将拷贝的序列归属到它们的原始分子背景。

[0087] 如将理解的，在用条形码序列标记340之前或与其同时，可以根据本文所述的任何方法扩增样品以提供对全基因组或基因组的选定区域的覆盖。对于期望靶向覆盖的实施方

案，靶向扩增通常导致与来自基因组的其它区域的扩增子相比，在含有基因组的那些选定区域的分区中代表核酸(或其部分)的序列的更大扩增子群体。结果，与基因组的其它区域相比，在来自基因组的选定区域的分区内将存在更大量的含有条形码序列的扩增拷贝340。在期望全基因组扩增的实施方案中，扩增可以使用设计成使扩增偏差最小化并提供跨整个基因组的稳健水平覆盖的引物文库进行。

[0088] 如上所指出，尽管单占用可能是最期望的状态，但是应该理解通常可能存在多重占用分区或未占用分区。用于共分配包含条形码寡聚核苷酸的样品和珠粒的微流体通道结构的实例在图4中示意性地示出。如所示，在通道接合部412处提供流体连通的通道区段402、404、406、408和410。包含单独样品414的水性物流经通道区段402流向通道接合部412。如本文其它地方所述，可以在分配过程之前将这些样品悬浮在水性流体内。

[0089] 同时，包含载有条形码的珠粒416的水性物流经通道区段404流向通道接合部412。将非水性分配流体从每个侧通道406和408引入通道接合部412，并且组合的物流流入出口通道410。在通道接合部412内，将来自通道区段402和404的两个组合的水性物流组合，并且分配到包括共分配的样品414和珠粒416的液滴418。如先前所指出，通过控制在通道接合部412组合的每种流体的流动特性，以及控制通道接合部的几何形状，可以优化组合和分配以实现珠粒、样品或两者在产生的分区418内的期望占用水平。

[0090] 如将理解的，许多其它试剂可以与样品和珠粒一起共分配，所述试剂包括例如化学刺激、核酸延伸、转录和/或扩增试剂，诸如聚合酶，逆转录酶，三磷酸核苷或NTP类似物，引物序列和额外辅因子，例如用于所述反应的二价金属离子，连接反应试剂，诸如连接酶和连接序列，染料，标记或其它标记试剂。引物序列可以包括针对扩增基因组的选定区域的随机引物序列或靶向PCR引物或其组合。

[0091] 一旦共分配，置于珠粒上的寡聚核苷酸则可以用于条形编码和扩增分配的样品。在扩增和条形编码样品中使用这些条形码寡聚核苷酸的特别优雅的方法在USSN 14/175,935;14/316,383;14/316,398;14/316,416;14/316,431;14/316,447;和14/316,463中详细描述，其全部公开内容通过引用整体并入本文。简而言之，一方面，珠粒上存在的寡聚核苷酸与样品共分配并自其珠粒释放到具有样品的分区中。寡聚核苷酸通常包括与条形码序列一起的在其5'端的引物序列。引物序列可以是随机的或结构化的。随机引物序列通常旨在随机引发样品的许多不同区域。结构化引物序列可以包括一系列的不同结构，这些结构包括靶向以引发样品的特异性靶向区域上游的限定序列以及具有某种部分限定的结构的引物，所述引物包括但不限于含有一定百分比的特异性碱基(例如，一定百分比的GC N-聚体)的引物、含有部分或完全简并序列的引物和/或含有根据本文中的任何描述部分随机和部分结构化的序列的引物。如将理解的，任何一种或多种上述类型的随机和结构化引物可以以任何组合包括在寡聚核苷酸中。

[0092] 一旦释放，寡聚核苷酸的引物部分可以退火至样品的互补区域。然后与样品和珠粒共分配的延伸反应试剂如DNA聚合酶、三磷酸核苷、辅因子(例如，Mg2+或Mn2+等)然后使用样品作为模板延伸引物序列以产生与引物退火的模板链互补的片段，其中互补片段包括寡聚核苷酸及其相关的条形码序列。将多个引物退火并延伸至样品的不同部分可以产生样品的重叠互补片段的大型汇集物，其每自具有指示产生其的分区的其自己的条形码序列。在一些情况下，这些互补片段本身可以用作由分区中存在的寡聚核苷酸引发的模板以产生

补体的补体，所述补体再次包括条形码序列。在一些情况下，构造该复制过程，使得当复制第一补体时，其在其末端处或其末端附近产生两个互补序列，以允许形成发夹结构或部分发夹结构，这降低了分子成为产生另外迭代拷贝的基础的能力。图5中示出其一个实例的示意图。

[0093] 如该图所示，包括条形码序列的寡聚核苷酸与样品核酸504一起在例如乳液中的液滴502中共分配。如本文别处所指出，寡聚核苷酸508可以提供在与样品核酸504共分配的珠粒506上，如版面A所示，所述寡聚核苷酸优选可以从珠粒506释放。除了一个或多个功能序列如序列510、514和516之外，寡聚核苷酸508还包括条形码序列512。例如，寡聚核苷酸508显示为包含条形码序列512以及序列510，序列510可以充当给定测序系统的附接或固定序列，例如用于在IlluminaHiseq或Miseq系统的流动细胞中附接的P5序列。如所示，寡聚核苷酸还包括引物序列516，其可以包括用于引发样品核酸504的部分复制的随机或靶向N-聚体。寡聚核苷酸508内还包括序列514，其可以提供测序引发区域，诸如“读段1”或R1引发区域，所述测序引发区域用于在测序系统中通过合成反应引发聚合酶介导的模板定向测序。在许多情况下，条形码序列512、固定序列510和R1序列514对于附接到给定珠粒的所有寡聚核苷酸可以是共有的。引物序列516对于随机N-聚体引物可以不同，或者对于某些靶向应用，对于给定珠粒上的寡聚核苷酸可以是共有的。

[0094] 基于引物序列516的存在，寡聚核苷酸能够如版面B所示引发样品核酸，其允许使用聚合酶和也与珠粒506和样品核酸504共分配的其它延伸试剂延伸寡聚核苷酸508和508a。如版面C所示，在寡聚核苷酸延伸后，对于随机N-聚体引物，将退火至样品核酸504的多个不同区域；产生核酸的多个重叠补体或片段，例如片段518和520。虽然包括与样品核酸的部分互补的序列部分，例如序列522和524，但是这些构建体在本文中通常被称为包含具有附接的条形码序列的样品核酸504的片段。如应理解，如上所述的模板序列的复制部分在本文中通常被称为该模板序列的“片段”。然而，尽管如此，术语“片段”涵盖例如模板或样品核酸的原始核酸序列的一部分的任何表示，包括通过提供模板序列的部分的其它机制如例如通过酶、化学或机械片段化实现的给定序列分子的实际片段化产生的那些。然而，在优选的方面，模板或样品核酸序列的片段将表示潜在序列或其补体的复制部分。

[0095] 条形码化核酸片段然后可以例如通过序列分析进行表征，或者它们可以在该过程中进一步扩增，如版面D中所示。例如，例如寡聚核苷酸508b的额外寡聚核苷酸也从珠粒506释放并且可以引发片段518和520。特定地讲，再次，基于寡聚核苷酸508b中的随机N-聚体引物516b的存在(其在许多情况下将不同于给定分区中的其它随机N-聚体，例如引物序列516)，寡聚核苷酸用片段518退火，并延伸以产生包含序列528的片段518的至少一部分的补体526，其包含样品核酸序列的一部分的复本。寡聚核苷酸508b继续延伸，直到它已经通过片段518的寡聚核苷酸部分508复制。如本文别处所指出，并且如版面D中图示，寡聚核苷酸可以构造成在期望的点处例如在通过包括在片段518内的寡聚核苷酸508的序列516和514复制之后通过聚合酶促使复制终止。如本文所述，这可以通过不同的方法完成，所述方法包括例如掺入不能被所使用的聚合酶加工的不同核苷酸和/或核苷酸类似物。例如，这可以包括在序列区域512内包括含尿嘧啶的核苷酸以防止非尿嘧啶耐受性聚合酶停止该区域的复制。结果，产生在一端包括全长寡聚核苷酸508b的片段526，其包括条形码序列512、附接序列510、R1引物区域514和随机N-聚体序列516b。在序列的另一端将包括第一寡聚核苷酸508

的随机N-聚体的补体516'以及R1序列的全部或一部分的补体，如序列514'所示。然后R1序列514及其补体514'能够一起杂交以形成部分发夹结构528。如将理解的，因为随机N聚体在不同的寡聚核苷酸中不同，所以将预期这些序列及其补体不参与发夹形成，例如，将预期作为随机N-聚体516的补体的序列516'不与随机N-聚体序列516b互补。对于其它应用，例如靶向引物，情况并非如此，其中N-聚体在给定分区内的寡聚核苷酸之中是共有的。通过形成这些部分发夹结构，它允许从进一步复制中除去样品序列的第一级重复体，例如防止拷贝的迭代拷贝。部分发夹结构还为例如片段526的所产生的片段的后续加工提供有用的结构。

[0096] 然后可以汇集来自多个不同分区的所有片段，以在如本文所述的高通量测序仪上进行测序。因为每个片段都按其原始分区进行编码，所以该片段的序列可以基于条形码的存在归属回其原点。这在图6中示意性地示出。如在一个实例中所示，源自第一来源600(例如，单独染色体、核酸链等)的核酸604和来源于不同染色体602或核酸链的核酸606各自与如上所述其自己的条形码寡聚核苷酸集一起分区。

[0097] 在每个分区内，然后加工每个核酸604和606以单独地提供第一片段的第二片段的重叠集，例如第二片段集608和610。该加工还提供具有条形码序列的第二片段，所述条形码序列对于来源于特定第一片段的每个第二片段是相同的。如所示，第二片段集608的条形码序列由“1”表示，而片段集610的条形码序列由“2”表示。可以使用多样化条形码文库来区别地条形编码大量的不同片段集。然而，来自不同第一片段的每一第二片段集不必用不同的条形码序列条形编码。实际上，在许多情况下，可以同时加工多个不同的第一片段以包括相同的条形码序列。本文中别处详细地描述了多样化条形码文库。

[0098] 然后可以汇集例如来自片段集608和610的条形码化片段以便使用例如通过得自Illumina或Thermo Fisher,Inc.的Ion Torrent部门的合成技术等等得到的序列测序。一旦被测序，来自汇集片段612的序列读段可以归属到它们各自的片段集，例如如聚合读段614和616所示，至少部分地基于所包括的条形码，并且任选地且优选地部分基于根据片段本身的序列。此外，序列读段可以归属到自其得到那些读段的核酸关于原始样品内紧密空间接近的其它核酸分子的相对位置的结构背景。然后组装每个片段集的所属序列读段以提供每个样品片段的组装序列，例如序列618和620，所述序列618和620又可以进一步归属回它们各自的原始染色体或源核酸分子(600和602)。用于组装基因组序列的方法和系统在例如2015年6月26日提交的美国专利申请No.14/752,773中描述，其全部公开内容通过引用整体地并且特别用于涉及基因组序列组装的所有教导并入本文。

[0099] III.保留结构背景的方法和组合物

[0100] 本公开内容提供用于表征基因材料的方法、组合物和系统。一般来讲，本文所述的方法、组合物和系统提供分析样品的组分同时保留关于那些组分原样在样品中的结构以及分子背景的信息的方法。换句话说，本文的描述通常涉及样品中核酸的空间检测，包括使用本领域已知的方法已经固定或将要固定的组织样品，诸如福尔马林固定的石蜡包埋样品。如将理解的，本部分中描述的任何方法都可以与上面在标题为“概述”和“工作流程概述”的部分中描述的任何方法以及在本说明书的后续部分中描述的核酸测序方法组合。

[0101] 通常，本文公开的方法涉及确定和/或分析样品中的核酸，包括样品的基因组，特别是全基因组。本文所述的方法提供定量或定性分析样品中核酸序列(包括基因组序列)的分布、位置或表达的能力，其中保留样品内的空间背景。本文公开的方法提供优于地理编码

样品中的核酸的常规方法的优势，因为在高通量加工方法中保留关于结构背景的信息，而不需要在加工用于序列读段的样品之前鉴定特定分子靶标(诸如特异性基因或其它核酸序列)。此外，需要少量的核酸，这在诸如FFPE样品的样品中是特别有利的，在所述样品中特别是DNA的输入核酸常常被片段化或以低浓度存在。

[0102] 尽管这里的大部分论述都是关于核酸的分析，但应理解的是，本文讨论的方法和系统可以适用于样品的其它组分，这些组分包括蛋白质和其它分子。

[0103] 如上讨论，维持结构背景(在本文中也被称为维持地理背景和编码地理)意味着使用允许获得多个序列读段或多个序列读段部分的方法，所述序列读段或序列读段部分可归属到样品内的那些序列读段的原始三维相对位置。换句话说，序列读段可以与样品内相对于该样品中的相邻核酸(以及在一些情况下，相关蛋白质)的相对位置相关联。该空间信息可以得到，即使那些相邻核酸物理上并未位于单个原始核酸分子的线性序列内。

[0104] 一般来讲，本文所述的方法包括提供含有核酸的样品的分析，其中所述核酸含有三维结构。将样品的部分分离到离散分区，使得核酸三维结构的部分也被分离到离散分区-彼此在空间上接近的核酸序列将倾向于被分离到相同分区，因此即使当后来获得的序列读段来自原本不在相同的单独原始核酸分子上的序列时，也保留该空间接近度的三维信息。参考图1：如果含有核酸分子102和103和106的样品101被分离到离散分区，使得样品的子集被配置到不同的离散分区中，则由于核酸分子106与102和103之间的物理距离，与核酸分子106不同，核酸分子102和103更可能彼此将被置于相同分区。因此，相同离散分区内的核酸分子是在原始样品中彼此空间接近的分子。从离散分区内的核酸获得的序列信息因此提供一种分析核酸的方式，例如通过核酸测序分析并将那些序列读段归属回原始核酸分子的结构背景。

[0105] 在一些实例中，将标签文库施加给样品以进行样品的空间或地理编码。在某些实施方案中，标签是寡聚核苷酸标签(其可以包括“寡聚核苷酸条形码”和“DNA条形码”)，但是如将理解的，可以使用能够添加到样品中的任何类型的标签，包括但不限于粒子、珠粒、染料、分子倒置探针(MIP)等等。标签文库可以通过简单扩散或通过活动过程，诸如组织培养或细胞培养样品中的细胞过程施加给样品。细胞转运过程包括但不限于渗透，通过细胞转运蛋白的参与促进扩散，被动转运和通过细胞转运蛋白的参与和来自诸如ATP的分子的能量输入的主动转运。通常，施加标签使得样品内的不同空间/地理位置接收不同的标签和/或不同浓度的标签。样品的任何进一步加工和样品内核酸的分析都可以通过鉴定标签而归属到特定的空间背景。例如，参考图1，向样品101中添加标签文库将产生与核酸106不同的对不同部分或浓度的标签文库具有空间接近度的核酸102和103。根据本文所述的工作流程对样品进行的任何进一步加工然后都将产生与相同部分/浓度的标签相关联的核酸102和103，并且因此这些标签的鉴定将指示核酸102和103在原始样品101中彼此空间接近。具有不同部分/浓度的标签的核酸106的鉴定将显示核酸106在原始样品中处于与核酸102和103不同的空间位置处。

[0106] 在进一步的实例中，还采用分区特异性条形码，使得获得的任何序列读段都可归属回原始核酸分子所处的分区。如上文讨论，使序列读段与特定分区相关联鉴定在原始样品的地理位置中彼此空间接近的核酸分子。诸如图2中所绘的那些的工作流程的进一步使用也提供关于序列读段的分子背景的信息，使得单独序列读段可以归属到它们所源自的单

独核酸分子。

[0107] 为了能够标记样品，可以使用本领域已知的任何方法加工样品以允许施加诸如寡聚核苷酸标签或其它标签的外源性分子。例如，在使用FFPE样品的实施方案中，标签可以通过加热样品以允许标签嵌入样品中而施加到样品，然后可以将样品冷却并根据本文所述的任何方法进一步加工，所述方法包括分到离散分区和进一步分析以鉴定样品中的核酸序列以及也与那些序列读段紧密空间接近的标签，从而保留那些序列读段的结构背景。其它样品加工方法包括除去细胞外基质和/或其它结构障碍同时保留分子和蛋白质要素的组织加工方法。这些方法在一些非限制性实例中包括CLARITY方法以及其它组织清除和标记方法的使用，包括例如在以下文献中描述的那些：Tomer等人，第9卷，第7期，2014，Nature Protocols；Kebschull等人，Neuron，第91卷，第5刊，2016年9月7日，第975-987页；Chung，K.等人，Structural and molecular interrogation ofintactbiological systems.Nature497,332-337(2013)；Susaki,E.A.等人，Whole-brain imaging with single-cell resolution using chemical cocktails and computational analysis.Cell 157,726-739(2014)；和Lee等人，ACT-PRESTO:Rapid and consistenttissue clearing and labeling method for 3-dimensional(3D)imaging，Scientific Reports，2016/01/11/在线；第6卷，第18631页，其各自出于所有目的通过引用整体地且特别是对于涉及加工用于结构和分子询问方法的样品的任何教导并入本文中。

[0108] 在某些实施方案中，本文所述的方法与成像技术组合使用以鉴定样品，特别是固定在载玻片上的样品如FFPE样品内标签的空间位置。所述成像技术可以允许序列读段与载玻片上的特定位置相关，这允许与可以用那些样品进行的其它病理学/成像研究相关。例如，可以使用成像技术来提供病理的初步鉴定。本文所述的进一步提供序列读段同时维持结构背景的测序技术可以与所述成像分析相组合以将序列读段与结构背景相关以证实或提供关于病理的初步鉴定的另外信息。此外，成像技术可以与具有光学性质的标签组合使用，使得特定标签与成像样品的特定区域相关联。与那些鉴定的标签相关的序列读段然后可以通过它们与这些标签的位置而与成像样品的区域进一步相关。然而，应该理解，本文所述的方法独立于任何这样的成像技术，并且保留结构上下文的能力不依赖于使用成像技术来确定样品中的核酸的空间信息。

[0109] 在一个示例性方面，在样品中产生寡聚核苷酸的梯度以提供可以通过测序通过后续加工解码的坐标系统。这样的梯度将允许用寡聚核苷酸或寡聚核苷酸浓度标记样品中的细胞和/或核酸，其可以映射到原始样品内的物理位置。该坐标系统可以通过使寡聚核苷酸文库扩散到样品中和/或通过将寡聚核苷酸注射到样品的特定区域中来开发。当使用扩散时，扩散动力学的标准计算将提供寡聚核苷酸标签的浓度与其在原始样品中的空间位置之间的相关性。因此，用该浓度的寡聚核苷酸标签鉴定的任何其它核酸又可以与样品的特定地理区域相关。

[0110] 在进一步的示例性实施方案中，所述方法包括用于分析核酸，同时维持结构背景的过程，其中将标签文库施加给样品，使得样品的不同地理区域接收不同标签。然后将现在含有其原始核酸以及添加的标签的样品部分分离到离散分区，使得样品内在地理位置上彼此接近的标签文库的部分和核酸部分该样品最终在同一离散分区中。测序过程，诸如本文详细描述的过程，用于提供离散分区中核酸的序列读段。标签也可以在那些测序过程之前，

之后或同时进行鉴定。序列读段与特定标签(或其中使用标签的浓度梯度的实施方案中的标签浓度)的相关性因此有助于提供序列读段的空间背景。如上讨论,其中用于空间编码的标签与分区特异性条形码结合使用的实施方案进一步提供序列读段的结构背景和分子背景。

[0111] IV.应用方法和系统进行核酸测序

[0112] 本文所述的方法、组合物和系统特别适用于核酸测序技术。所述测序技术可以包括本领域已知的任何技术,包括短读段和长读段测序技术。在某些方面,本文所述的方法、组合物和系统用于短读段高准确度测序技术。

[0113] 一般来讲,本文所述的方法和系统使用具有短读段测序技术的极低测序错误率和高通量的优势的方法完成基因组测序。如前所述,本文所述的方法和系统的优势在于它们可以通过使用无所不在的短读段测序技术来获得期望的结果。所述技术具有易于获得并且在研究界用充分表征且高效的方案和试剂系统广泛分散的优势。这些短读段测序技术包括可得自例如Illumina,Inc.(GAIIx、NextSeq、MiSeq、HiSeq、X10)、Thermo-Fisher的Ion Torrent部门(Ion Proton和Ion PGM)的那些,焦磷酸测序方法以及其它。

[0114] 特别有利的是,本文所述的方法和系统利用这些短读段测序技术并且在其相关的低错误率下进行。特别地讲,本文所述的方法和系统获得如上所述的期望单独分子读段长度或背景,但所述期望单独分子读段长度或背景具有短于1000bp、短于500bp、短于300bp、短于200bp、短于150bp或甚至更短的单独测序读段(排除配偶对延伸);并且对于所述单独分子读段长度具有小于5%、小于1%、小于0.5%、小于0.1%、小于0.05%、小于0.01%、小于0.005%或甚至小于0.001%的测序错误率。

[0115] 根据本申请中描述的方法和系统加工并测序核酸的方法也在USSN 14/316,383;14/316,398;14/316,416;14/316,431;14/316,447;和14/316,463中进一步详细描述,其出于所有目的通过引用整体地且特别是对于涉及加工核酸和对基因组材料进行测序和其它表征的所有书面描述、附图和工作实施例并入本文。

[0116] 在一些实施方案中,本文所述的用于获得序列信息同时保留结构背景和分子背景的方法和系统用于全基因组测序。在一些实施方案中,本文所述的方法用于对基因组的靶向区域进行测序。在进一步的实施方案中,本文所述的测序方法包括选定区域的深度覆盖与跨基因组的更长范围的较低水平连接读段的组合。如将理解的,从头测序和重新测序的这种组合提供一种测序整个基因组和/或大部分基因组的高效方法。表征欠佳和/或高度多态的区域的靶向覆盖进一步提供了从头序列组装所需的核酸物质的量,而基因组的其它区域上的连接基因组测序维持基因组的其余部分的高通量测序。本文所述的方法和组合物适合于允许从头测序和连接读段测序的这种组合,因为对于两种类型的覆盖可以使用相同的测序平台。根据本文所述的方法测序的核酸和/或核酸片段的群体可以含有来自用于从头测序的基因组区域和用于重新测序的基因组区域两者的序列。

[0117] 在具体情况下,本文所述的方法包括在测序之前扩增基因组的全部或选定区域的步骤。通常使用本领域已知的方法(包括但不限于PCR扩增)进行的这种扩增提供基因组的全部或选定区域的至少1X、2X、3X、4X、5X、6X、7X、8X、9X、10X、11X、12X、13X、14X、15X、16X、17X、18X、19X或20X覆盖度。在进一步的实施方案中,扩增提供基因组的全部或选定区域的至少1X-30X、2X-25X、3X-20X、4X-15X或5X-10X覆盖度。

[0118] 通常通过延伸与基因组选定区域内或其附近的序列互补的引物进行扩增以覆盖全基因组和/或基因组的选定靶向区域。在一些情况下，使用设计成跨所关注的基因组区域平铺的引物文库-换句话说，引物文库设计成沿基因组以特定的距离扩增区域，无论这是跨选定区域还是跨全基因组。在一些情况下，选择性扩增利用与沿着基因组的选定区域的每10、15、20、25、50、100、200、250、500、750、1000或10000个碱基互补的引物。在更进一步的实例中，引物的平铺文库设计成捕获距离的混合，所述混合可以是距离的随机混合或经智能设计使得选定区域的特定部分或百分比由不同引物对扩增。在进一步的实施方案中，引物对设计成使得每一对扩增基因组的选定部分的任何连续区域的约1-5％、2-10％、3-15％、4-20％、5-25％、6-30％、7-35％、8-40％、9-45％或10-50％。

[0119] 在某些实施方案中并且根据以上任何描述，扩增跨长度为至少3兆碱基对(Mb)的基因组区域发生。在进一步的实施方案中，基因组的选定区域根据本文所述的任何方法选择性地扩增，并且所述选定区域至少3.5、4、4.5、5、5.5、6、6.5、7、7.5、8、8.5、9、9.5或10Mb长。在又进一步的实施方案中，基因组的选定区域的长度为约2-20、3-18、4-16、5-14、6-12或7-10Mb。使用与这些区域的末端或末端附近的序列互补的单个引物对，扩增可以跨这些区域发生。在其它实施方案中，扩增是通过跨区域的长度平铺的引物对的文库进行，从而在根据上述的覆盖程度下扩增沿该区域的规则区段、随机区段或不同区段距离的一些组合。

[0120] 在一些实施方案中，用于选择性扩增基因组的选定区域的引物含有尿嘧啶，从而不扩增引物本身。

[0121] 与所用的测序平台无关，一般来讲且根据本文所述的任何方法，核酸的测序通常以保存序列读段或序列读段部分的结构背景和分子背景的方式进行。这意味着多个序列读段或多个序列读段部分可以归属到相对于其它核酸的原始样品内的相对空间位置(结构背景)和/或核酸的单个原始分子的线性序列内的位置(分子背景)。

[0122] 如将理解的，尽管核酸的单个原始分子可以具有各种长度中的任何长度，但在优选的方面，它将是相对长的分子，从而允许保存长范围分子背景。特定地讲，单个原始分子优选比典型的短读段序列长度基本上更长，例如长于200个碱基，并且通常为至少1000个碱基或更长、5000个碱基或更长、10,000个碱基或更长、20,000个碱基或更长、30,000个碱基或更长、40,000个碱基或更长、50,000个碱基或更长、60,000个碱基或更长、70,000个碱基或更长，80,000个碱基或更长，90,000个碱基或更长，或100,000个碱基或更长，并且在一些情况下1兆碱基或更长。

[0123] 通常，本发明的方法包括如图2图示的步骤，其提供本文中进一步详细讨论的本发明的方法的示意性图。如将理解的，图2中概述的方法是可以根据需要并如本文所述进行改变或修改的示例性实施方案。

[0124] 如图2所示，本文所述的方法在大多数实例中将包括分配样品的步骤(202)。在该分配步骤之前，可以存在任选的步骤(201)，其中连接样品中的核酸以附接彼此紧密空间接近的序列区域。一般来讲，含有来自所关注的基因组区域的核酸的每个分区将经历某种片段化过程，并且通常通过条码编码对包含它们的分区具有特异性的片段通常将保留片段的原始分子背景(203)。每个分区在一些实例中可以包括多于一个核酸，并且在一些情况下将含有数百个核酸分子-在多个核酸在一个分区内的情况下，在条形编码之前，基因组的任何特定基因座通常将由单个单独核酸代表。如上讨论，步骤203的条形码化片段可以使用本领

域已知的任何方法产生-在一些实例中，寡聚核苷酸为在不同分区内的样品。所述寡聚核苷酸可以包含旨在随机引发样品的多个不同区域的随机序列，或者它们可以包含被靶向以引发样品的靶向区域上游的特异性引物序列。在进一步的实例中，这些寡聚核苷酸还含有条形码序列，使得复制过程还条形编码原始样品核酸的所得复制片段。也包含在分区中的延伸反应试剂如DNA聚合酶、三磷酸核苷、辅因子(例如，Mg^{2+}或Mn^{2+}等)然后使用样品作为模板延伸引物序列以产生与引物退火的模板链的互补片段，并且所述互补片段包括寡聚核苷酸及其相关的条形码序列。将多个引物退火并延伸至样品的不同部分可以产生样品的重叠互补片段的大型汇集物，其各自具有指示产生其的分区的其自己的条形码序列。在一些情况下，这些互补片段本身可以用作由分区中存在的寡聚核苷酸引发的模板以产生补体的补体，所述补体再次包括条形码序列。在进一步的实例中，构造该复制过程，使得当复制第一补体时，其在其末端处或其末端附近产生两个互补序列，以允许形成发夹结构或部分发夹结构，这降低了分子成为产生另外迭代拷贝的基础的能力。

[0125] 回到图2中例示的方法，一旦分区特异性条形码附接到拷贝的片段，然后可以任选地汇集条形码化片段(204)。然后对汇集的片段进行测序(205)，并将片段的序列归属到它们的原始分子背景(206)，从而鉴定所关注的靶向区域并且还将其与原始分子背景连接。本文所述的方法和系统的优势在于在对于靶向基因组区域富集片段之前将分区或样品特异性条形码附接到拷贝的片段保存了那些靶向区域的原始分子背景，允许它们归属到它们的原始分区，并因此归属到它们的原始样品核酸分子。

[0126] 除了上述工作流程之外，使用包括基于芯片的捕获方法和基于溶液的捕获方法的方法，可以将靶向基因组区域进一步富集、分离(isolate)或分离(separate)，即“下拉”，以便进一步分析，特别是测序。此类方法利用与所关注的基因组区域或与在所关注的基因组区域附近或与其相邻的区域互补的探针。例如，在杂交(或基于芯片)捕获中，将含有捕获探针(通常是单链寡聚核苷酸)的微阵列固定到表面上，所述捕获探针具有一起覆盖所关注的区域的序列。将基因组DNA片段化，并且可以进一步进行加工，诸如末端修复，以产生平端和/或添加额外特征物如通用引发序列。这些片段与微阵列上的探针杂交。将未杂交的片段洗掉并将所需片段洗脱或以其它方式在表面上加工以便测序或其它分析，并且因此富集表面上残留的片段群体的含有所关注的靶向区域(例如，包含与捕获探针中所含的那些序列互补的序列的区域)的片段。富集的片段群体可以使用本领域已知的任何扩增技术进一步扩增。用于这样的靶向下拉富集方法的示例性方法在2014年10月29日提交的USSN 62/072,164中描述，其出于所有目的通过引用整体地并且特别是对于包括所有书面描述、附图和实施例的涉及靶向下拉富集方法和测序方法的所有教导并入本文。

[0127] 在一些实例中，不是全基因组测序，而是期望集中在基因组的选定区域上。本文所述的方法特别适用于这样的分析，因为即使当基因组亚集在原始样品的三维背景中处于大线性距离但潜在地极为接近时，靶向这些亚集的能力也是这些方法的有利特征。在一些方面，用于覆盖基因组的选定区域的方法包括其中将含有来自那些选定区域的核酸分子和/或其片段的离散分区本身分类以供进一步加工的方法。如将理解的，对离散分区的该分类可以以与本文所述的所关注的基因组区域的其它选择性扩增和/或靶向下拉方法的任何组合进行，特别是以与上述工作流程的步骤的任何组合进行。

[0128] 一般来讲，离散分区的分类方法包括以下步骤：其中将包含基因组的一个或多个

选定部分的至少一部分的分区与不包含来自基因组的那些部分的任何序列的分区分开。这些方法包括以下步骤:在含有来自基因组的一个或多个选定部分的序列的离散分区内提供富含包含基因组的那些部分的至少一部分的片段的序列的群体。所述富集通常通过在包括基因组的一个或多个选定部分的至少一部分的离散分区内使用的片段的定向PCR扩增来产生群体。该定向PCR扩增因此产生包含基因组的一个或多个选定部分的至少一部分的扩增子。在某些实施方案中,这些扩增子附接到可检测标记,所述可检测标记在一些非限制性实施方案中可以包括荧光分子。一般来讲,发生所述附接,使得只有那些由含有基因组的一个或多个选定部分的片段产生的扩增子附接到可检测标记上。在一些实施方案中,可检测标记的附接在基因组的一个或多个选定部分的选择性扩增期间发生。所述可检测标记在进一步的实施方案中可以包括但不限于荧光标记、电化学标记、磁性珠粒和纳米粒子。可检测标记的这种附接可以使用本领域已知的方法来完成。在又进一步的实施方案中,基于从连接于这些分区内的扩增子的可检测标记发射的信号对含有包含基因组的一个或多个选定部分的至少一部分的片段的离散分区进行分选。

[0129] 在进一步的实施方案中,分类含有基因组的选定部分的离散分区与不含所述序列的那些离散分区的步骤包括以下步骤:(a)提供起始基因组材料;(b)将来自起始基因组材料的单独核酸分子分布到离散分区中,使得每个离散分区含有第一单独核酸分子;(c)提供在离散分区的至少一些内富含包含基因组的一个或多个选定部分的至少一部分的片段的序列的群体;(d)将共同的条形码序列附接到每个离散分区内的片段上,使得每个片段归属到包含其的离散分区;(e)将含有包含基因组的一个或多个选定部分的至少一部分的片段的离散分区与不含包含基因组的一个或多个选定部分的片段的离散分区中分开;(f)自包含基因组的一个或多个选定部分的至少一部分的片段获得序列信息,从而对基因组样品的一个或多个靶向部分进行测序,同时保留分子背景。如将理解的,这样的方法的步骤(a)可以包括多于一个单独核酸分子。

[0130] 在进一步的实施方案中并且根据上述任何一个实施方案,在自片段获得序列信息之前,组合离散分区并将片段汇集在一起。在进一步的实施方案中,自片段获得序列信息的步骤以维持片段序列的结构背景和分子背景的方式进行,使得鉴定还包括鉴定来源于原始样品内紧密物理接近定位和/或在相同第一单独核酸分子上定位的核酸的片段。在更进一步的实施方案中,序列信息的该获得包括选自由以下组成的组的测序反应:短读段长度测序反应和长读段长度测序反应。在又进一步的实施方案中,测序反应是短读段高精确度测序反应。

[0131] 在更进一步的实施方案中并且根据上述任何一个实施方案,离散分区包含在乳液中的液体。在进一步的实施方案中,离散分区内的条形码化片段代表基因组的一个或多个选定部分的约1X-10X覆盖度。在更进一步的实施方案中,离散分区内的条形码化片段代表基因组的一个或多个选定部分的约2X-5X覆盖度。在又进一步的实施方案中,离散分区内的扩增子的条形码化片段代表基因组的一个或多个选定部分的至少1X覆盖度。在更进一步的实施方案中,离散分区内的条形码化片段代表基因组的一个或多个选定部分的至少2X或5X覆盖度。

[0132] 除了提供自基因组的选定区域获得序列信息的能力之外,本文所述的方法和系统还可以提供基因组材料的其它表征,这些表征包括但不限于单倍型定相、鉴定结构变异和

鉴定拷贝数变异的鉴定，如USSN 14/316,383、14/316,398、14/316,416、14/316,431、14/316,447和14/316,463中详细所述，其出于所有目的通过引用整体地并入本文其出于所有目的通过引用整体地且特别是对于涉及基因组材料表征的所有书面描述、附图和工作实施例并入本文。

[0133] 在一方面，并且结合上文和随后在此描述的任何方法，本文所述的方法和系统提供样品核酸或其片段的区室化、沉积或分区成离散的隔室或分区(在本文中可互换地称为分区)，其中每个分区维持其自己的内含物与其它分区的内含物的分离。可以将独特的标识符如条形码预先、随后或同时递送至容纳区室化或分区样品核酸的分区，以允许将特征如核酸序列信息随后归属到包括在特定区室内的样品核酸，且特别是可以原始沉积到分区中的连续样品核酸的相对长的段。

[0134] 本文所述的方法中利用的样品核酸通常代表待分析的总体样品的多个重叠部分，例如整个染色体、外显子组或其它大基因组部分。这些样品核酸可以包括全基因组、单独染色体、外显子组、扩增子或任何各种不同的所关注核酸。通常分配样品核酸使得核酸以连续核酸分子的相对长的片段或段存在于分区中。通常，样品核酸的这些片段可以长于1kb、长于5kb、长于10kb、长于15kb、长于20kb、长于30kb、长于40kb、长于50kb、长于60kb、长于70kb、长于80kb、长于90kb或甚至长于100kb，这容许上述较长范围的分子背景。

[0135] 样品核酸通常也以一定的水平分配，由此给定的分区具有包括起始样品核酸的两个重叠片段的很低的可能性。这通常通过在分配过程期间以低输入量和/或浓度提供样品核酸来完成。结果，在优选的情况下，给定的分区可以包括起始样品核酸的许多长但不重叠的片段。不同分区中的样品核酸然后与独特标识符相关联，其中对于任何给定的分区，其中所含的核酸具有相同的独特标识符，但是其中不同的分区可以包括不同的独特标识符。此外，由于分配步骤将样品组分配置到体积非常小的分区或液滴中，所以应理解，为了实现如上所述的期望配置，无需进行样品的大量稀释，如将在更高容量过程中如在管或多孔板的各孔中所需要的。另外，由于本文所述的系统采用如此高水平的条形码多样性，因此可以如上文所提供在较高数量的基因组等同物之中配置多样的条形码。特定地讲，先前描述的多孔板方法(参见例如美国公开的申请No.2013-0079231和2013-0157870)通常仅用一百至几百种不同的条形码序列操作，并且采用其样品的有限稀释过程以便能够将条形码归属到不同的细胞/核酸。因此，它们通常将用远少于100个细胞操作，这通常将提供近似1:10且肯定远高于1:100的基因组:(条形码类型)比率。另一方面，本文所述的系统由于高水平的条形码多样性如超过10,000、100,000、500,000、600,000、700,000等多样条形码类型而可以以近似地1:50或更低、1:100或更低、1:1000或更低、或甚至更小比率的基因组:(条形码类型)比率操作，同时还允许载入更高数量的基因组(例如，近似地每次测定大于100个基因组，每次测定大于500个基因组，每次测定1000个基因组，或甚至更多)，同时仍提供每个基因组远远提高的条形码多样性。

[0136] 通常，样品与在分配步骤之前可释放地附接至珠粒的寡聚核苷酸标签集组合。用于条形编码核酸的方法在本领域中已知并且在本文中描述。在一些实例中，利用如Amini等人，2014,Nature Genetics,Advance Online Publication)中所述的方法，其出于所有目的通过引用整体地且特别是对于涉及附接条形码或其它寡聚核苷酸标签到核酸的所有教导并入本文。在进一步的实例中，寡聚核苷酸可以包含至少第一区域和第二区域。第一区域

可以是条形码区域,其在给定分区内的寡聚核苷酸之间可以是基本相同的条形码序列,但是在不同分区之间可以是并且在大多数情况下是不同的条形码序列。第二区域可以是可以用于在分区内引发在样品内的核酸的N-聚体(随机N-聚体或设计成靶向特定序列的N-聚体)。在一些情况下,在N-聚体设计成靶向特定序列的情况下,其可以被设计成靶向特定的染色体(例如,染色体1、13、18或21)或染色体的区域,例如外显子或其它靶向区域。如本文所讨论,N-聚体也可以被设计用于基因组的倾向于表征欠佳或相对于参考序列高度多态或趋异的选定区域。在一些情况下,N-聚体可以被设计成靶向特定基因或基因区域,诸如与疾病或病症(例如癌症)相关联的基因或区域。在分区内,扩增反应可以使用第二N-聚体进行以沿着核酸的长度在不同地方引发核酸样品。作为扩增的结果,每个分区可以含有附接到相同或近乎相同的条形码并且可以代表每个分区中核酸的重叠的较小片段的核酸的扩增产物。条形码可以充当预示起源于相同分区且因此也潜在地起源于核酸的相同链的核酸集的标志物。扩增后,可以将核酸汇集、测序并使用测序算法比对。由于较短的序列读段可以凭借其相关条形码序列进行比对并归属到样品核酸的单个长片段,因此该序列上的所有鉴定的变体可以归属到单个原始片段和单个原始染色体。另外,通过比对跨多个长片段的多个共定位变体,可以进一步表征该染色体贡献。因此,然后可以绘制关于特定基因变体的定相的结论,如可以跨长范围基因组序列进行分析-例如,跨基因组的表征欠佳的区域的段的序列信息的鉴定。所述信息也可以用于鉴定单元型,其通常是驻留在相同核酸链或不同核酸链上的一组特定的基因变体。拷贝数变化也可以以此方式鉴定。

[0137] 所描述的方法和系统提供优于当前的核酸测序技术及其相关的样品制备方法的显著优势。全体样品制备和测序方法倾向于主要鉴定和表征样品中的多数组分,并且不旨在鉴定和表征构成所提取样品中总DNA的小百分比的少数组分,例如来自基因组的表征欠佳或高度多态性的区域的由一种染色体贡献的基因材料、或来自一个或几个细胞的材料、或在血流中循环的片段化肿瘤细胞DNA分子。本文所述的方法包括增加来自这些少数组分的基因材料的选择性扩增方法,并且保留该基因材料的分子背景的能力进一步提供这些组分的基因表征。所描述的方法和系统还为检测存在于较大样品内的群体提供显著的优势。如此,它们对于评估单体型和拷贝数变化特别有用-本文公开的方法也可用于提供在原始样品的三维空间内彼此空间接近定位的序列或得到那些序列的原始核酸分子的序列信息。

[0138] 使用本文公开的条形码技术赋予为基因组的序列和区域提供单独结构背景和分子背景的独特能力。基因组的所述区域可以包括给定基因标志物集,即将给定基因标志物集(与单个标志物相反)归属到单独的样品核酸分子,并且通过变体配位组装,以在多个样品核酸分子之中和/或对特定的染色体提供更宽或甚至更长范围的推断单独分子背景。这些基因标志物可以包括特定的遗传基因座,例如变体,诸如SNP,或者它们可以包括短序列。此外,使用条形编码赋予促进区分从样品中提取的总核酸群体的少数组分和多数组分例如用于检测和表征血流中的循环肿瘤DNA的能力的附加优势,并且还降低或消除在任选的扩增步骤期间的扩增偏差。此外,以微流体形式实施赋予在极小样品量和低输入量的DNA下工作的能力以及快速加工大量样品分区(液滴)以促进全基因组标记的能力。

[0139] 如上所指出,本文所述的方法和系统为较长核酸的短序列读段提供单独的结构背景和分子背景。如本文所用,结构背景是指原始样品内其原始核酸分子的三维空间内的序列的位置。如上所讨论,尽管基因组通常被认为是线性的,但染色体并不是刚性的,并且两

个基因组基因座之间的空间距离不一定与它们沿基因组的距离相关-沿线性序列的数兆碱基分离的基因组区域可以在三维空间中彼此直接接近。通过保留序列读段的原始空间接近度的信息，本文所述的方法和组合物提供一种将序列读段归属到长范围基因组相互作用的方式。

[0140] 类似地，用本文所述的方法可以实现单独分子背景的保留提供超出特定序列读段的序列背景，例如与不包括在序列读段本身内的相邻序列或近侧序列有关，并且因此通常将使得它们并不全部或部分地包括在用于配对读段的短序列读段如约150个碱基或约300个碱基的读段中。在特别优选的方面，所述方法和系统为短序列读段提供长范围序列背景。所述长范围背景包括给定序列读段与如下序列读段的关系或键联，所述序列读段彼此之间的距离长于1kb、长于5kb、长于10kb、长于15kb、长于20kb、长于30kb、长于40kb、长于50kb、长于60kb、长于70kb、长于80kb、长于90kb或甚至长于100kb或更长。通过提供更长范围的单独分子背景，本发明的方法和系统还提供长得多的推断分子背景。如本文所述的序列背景可以包括例如来自将短序列读段作图到单独的更长分子或连接分子的重叠群的较低分辨率的背景，以及例如来自例如具有单独分子的连续确定序列的更长单独分子的大部分的长范围测序的较高分辨率的序列背景，其中这样确定的序列长于1kb、长于5kb、长于10kb、长于15kb、长于20kb、长于30kb、长于40kb、长于50kb、长于60kb、长于70kb、长于80kb、长于90kb或甚至长于100kb。与序列背景一样，将短序列归属到较长核酸(例如，单独的长核酸分子或连接的核酸分子或重叠群的集合)可以包括针对较长核酸段将短序列作图以提供高水平的序列背景以及通过这些较长核酸提供来自短序列的组装序列。

[0141] 本文所述的方法、组合物和系统允许表征跨基因组的长范围相互作用以及表征样品内的相关蛋白和其它分子。像更高水平的蛋白质组织一样，DNA和染色质的弯曲和折叠以各种规模上产生功能重要的结构。在小规模的情况下，众所周知DNA常常缠绕在诸如组蛋白的蛋白质周围以产生称为核小体的结构。这些核小体包装成较大的“染色质纤维”，并且包装模式已经牵涉到受诸如转录的细胞过程影响。功能结构也以更大规模存在：由基因组的多兆碱基长的线性序列分离的区域可以在三维空间中直接相邻。基因组基因座之间的所述长范围相互作用可以在功能特性方面起作用：例如，基因增强子、沉默子和绝缘子元件都可以跨广泛的基因组距离起作用，并且其主要作用模式可以涉及与靶基因、非编码RNA和/或调控元件的直接物理关联。长范围相互作用不限于位于顺式，即沿相同染色体的元件，而是也可以在位于反式，即不同染色体上的基因组基因座之间发生。长范围相互作用的存在会使理解调控细胞过程的途径的努力变得复杂化，因为相互作用调控元件可以位于距靶基因的大基因组距离处，甚至位于另一染色体上。在癌基因和其它疾病相关基因的情况下，鉴定长范围基因调控子可能在鉴定负责疾病状态的基因组变体和引起疾病状态的过程方面具有重大用途。因此，根据本文所述的方法保留结构背景和分子背景的能力提供一种鉴定长范围基因组相互作用和表征任何相关蛋白质的方式。

[0142] 本文所述的方法特别可用于表征来自包括历史FFPE组织样品的FFPE组织样品的核酸。FFPE样品通常对核酸表征提出挑战，因为核酸经常被片段化或以其它方式降解，这会限制可以使用常规方法获得的信息量。本文描述的方法中保留的结构背景和分子背景信息为所述样品提供独特的机会，因为该背景信息可以提供甚至对于降解的样品的长范围基因组相互作用的表征，因为长范围信息可以通过短读段测序技术得到。FFPE核酸表征的应用

包括将来自一个或多个历史样品的序列与来自例如癌症患者的受试者的样品的序列进行比较以提供诊断或预后信息。例如,历史样品中的一种或多种分子标志物的状态可以与一种或多种治疗结果相关,并且治疗结果与一种或多种历史样品中的分子标志物状态的相关性可以用于预测例如癌症患者的受试者的治疗结果。这些预测可以作为确定是否向受试者推荐药物治疗选择的基础。

[0143] V.样品

[0144] 如将理解的,本文讨论的方法和系统可以用于从任何类型的基因组材料获得序列信息。所述基因组材料可以从取自患者的样品获得。用于本文讨论的方法和系统的基因组材料的示例性样品和类型包括但不限于多核苷酸、核酸、寡聚核苷酸、无循环细胞的核酸、循环肿瘤细胞(CTC)、核酸片段、核苷酸、DNA、RNA、肽多核苷酸、互补DNA(cDNA)、双链DNA(dsDNA)、单链DNA(ssDNA)、质粒DNA、粘粒DNA、染色体DNA、基因组DNA(gDNA)、病毒DNA、细菌DNA、mtDNA(线粒体DNA)、核糖体RNA、无细胞DNA、无细胞胎儿DNA(cffDNA)、mRNA、rRNA、tRNA、nRNA、siRNA、snRNA、snoRNA、scaRNA、微RNA、dsRNA、病毒RNA等等。总之,使用的样品可以根据特定的加工需要而变。

[0145] 在特定的方面,用于本发明的样品包括福尔马林固定的石蜡包埋(FFPE)细胞和组织样品等等,以及其中样品降解风险高的任何其它样品类型。其它类型的固定样品包括但不限于使用以下固定的样品:丙烯醛、乙二醛、四氧化锇、碳二亚胺、氯化汞、锌盐、苦味酸、重铬酸钾、乙醇、甲醇、丙酮和/或乙酸。

[0146] 在进一步的实施方案中,用于本文所述的方法和系统的样品包括核基质。“核基质”是指包含核酸和蛋白质的任何组合物。核酸可以被组织成染色体,其中蛋白质(即,例如组蛋白)可以与具有调控功能的染色体相关联。

[0147] 本文提供的方法和系统特别可用于核酸测序应用,其中起始核酸(例如,DNA、mRNA等)-或起始靶核酸-以少量存在,或者其中对于分析靶向的核酸样品内以总核酸的相对低比例存在。在一方面,本公开内容提供分析核酸的方法,其中输入核酸分子以小于50纳克(ng)的量存在。在进一步的实施方案中,核酸分子的输入量小于小于40ng。在一些实施方案中,所述量小于20ng。在一些实施方案中,所述量小于10ng。在一些实施方案中,所述量小于5ng。在一些实施方案中,所述量小于1ng。在一些实施方案中,所述量小于0.1ng。起始输入量为少量的用于分离和分析核酸的方法例如在2015年6月26日提交的USSN 14/752,602中进一步描述,其出于所有目的通过引用整体地且特别是对于涉及自其中存在少量核酸的样品得到的核酸的分离和表征的所有教导并入本文。

[0148] 如将理解的,可以在本文所述的方法期间的任何点使用本领域已知的方法加工样品。例如,可以在分配之前或在已经将样品分配到离散分区中之后加工样品。

[0149] 在某些实施方案中,加工样品以确保保留较长核酸链。在使用FFPE样品的实施方案中,可以对所述样品进行加工以除去甲醛加合物,从而提高核酸收率。此类加工方法可以在一个非限制性实例中包括使用水溶性有机催化剂以加速甲醛加合物自RNA和DNA碱基反转,如在Karmakar等人,(2015),Nature Chemistry,DOI:10.1038/NCHEM.2307中所述,其通过引用整体地且特别是对于涉及FFPE样品的处理和加工的所有教导并入本文。

[0150] 包含核酸的任何物质都可以是样品的来源。所述物质可以是流体,例如生物流体。流体物质可以包括但不限于血液、脐带血、唾液、尿液、汗液、血清、精液、阴道液、胃和消化

液、脊髓液、胎盘液、腔液、眼液、血清、母乳、淋巴液或其组合。所述物质可以是固体，例如生物组织。所述物质可以包括正常健康组织、患病组织或者健康组织和患病组织的混合物。在某些情况下，所述物质可以包含肿瘤。肿瘤可以是良性的(非癌症)或恶性的(癌症)。肿瘤的非限制性实例可以包括：纤维肉瘤、粘液肉瘤、脂肪肉瘤、软骨肉瘤、成骨肉瘤、脊索瘤、血管肉瘤、内皮肉瘤、淋巴管肉瘤、淋巴管内皮肉瘤、滑膜瘤、间皮瘤、尤因氏肉瘤、平滑肌肉瘤、横纹肌肉瘤、胃肠系统癌、结肠癌、胰腺癌、乳腺癌、泌尿生殖系统癌、卵巢癌、前列腺癌、鳞状细胞癌、基底细胞癌、腺癌、汗腺癌、皮脂腺癌、乳头状癌、乳头状腺癌、囊腺癌、髓样癌、支气管癌、肾细胞癌、胆管癌、胆管癌、绒毛膜癌、精原细胞瘤、胚胎癌、维尔姆斯瘤(Wilms' tumor)、宫颈癌、内分泌系统癌、睾丸肿瘤、肺癌、小细胞肺癌、非小细胞肺癌、膀胱癌、上皮癌、神经胶质瘤、星形细胞瘤、髓母细胞瘤、颅咽管瘤、室管膜瘤、松果体瘤、血管母细胞瘤、听神经瘤、少突胶质细胞瘤、脑膜瘤、黑素瘤、成神经细胞瘤、成视网膜细胞瘤或其组合。所述物质可以与各种类型的器官相关联。器官的非限制性实例可以包括脑、肝、肺、肾、前列腺、卵巢、脾、淋巴结(包括扁桃体)、甲状腺、胰腺、心脏、骨骼肌、肠、喉、食道、胃或其组合。在一些情况下，所述物质可以包括多种细胞，包括但不限于：真核细胞、原核细胞、真菌细胞、心脏细胞、肺细胞、肾细胞、肝细胞、胰腺细胞、生殖细胞、干细胞、诱导的多能干细胞、胃肠细胞、血细胞、癌细胞、细菌细胞、从人微生物组样品分离的细菌细胞等。在一些情况下，所述物质可以包含细胞内容物，诸如，例如单个细胞的内容物或多个细胞的内容物。在例如2015年6月26日提交的USSN 14/752,641中提供了用于分析单独细胞的方法和系统，其全部公开内容通过引用整体并入本文。

[0151] 样品可以从各种受试者获得。受试者可以是活受试者或死受试者。受试者的实例可以包括但不限于人类、哺乳动物、非人类哺乳动物、啮齿动物、两栖动物、爬行动物、犬科动物、猫科动物、牛科动物、马科动物、山羊、绵羊、母鸡、小鼠、兔、昆虫、鼻涕虫、微生物、细菌、寄生虫或鱼。在一些情况下，受试者可以是患有疾病或病症、怀疑患有疾病和病症或处于发展病症或病症的风险中的患者。在一些情况下，受试者可以是孕妇。在一些情况下，受试者可以是正常健康的孕妇。在一些情况下，受试者可以是可能处于怀有具有某些出生缺陷的婴儿的风险的孕妇。

[0152] 样品可以通过本领域已知的任何手段自受试者获得。例如，样品可以通过进入循环系统(例如，经由注射器或其它装置静脉内或动脉内)，收集分泌的生物样品(例如，唾液、痰尿、粪便等)、手术(例如，活组织检查)，采集生物样品(例如，手术中样品、手术后样品等)，擦拭(例如，口腔拭子、口咽拭子)或移液从受试者获得。

[0153] VI.实施方案

[0154] 在一些方面，本公开内容提供分析核酸同时维持结构背景的方法。此类方法包括以下步骤：(a)提供含有核酸的样品，其中所述核酸包含三维结构；(b)将所述样品的部分分离到离散分区，使得所述核酸三维结构的部分也被分离到所述离散分区；(c)自所述核酸获得序列信息，由此分析核酸，同时维持结构背景。

[0155] 在一些实施方案中，来自获得步骤(c)的序列信息包括鉴定彼此空间接近的核酸。

[0156] 在任何实施方案中，获得步骤(c)提供关于基因组基因座之间的染色体内和/或染色体间相互作用的信息。

[0157] 在任何实施方案中，获得步骤(c)提供关于染色体构象的信息。

[0158] 在任何实施方案中，在分离步骤(b)之前，加工所述三维结构中的至少一些以连接在三维结构内彼此接近的核酸的不同部分。

[0159] 在任何实施方案中，所述样品是福尔马林固定的石蜡样品。

[0160] 在任何实施方案中，核酸在分离步骤(b)之前未从样品中分离。

[0161] 在任何实施方案中，所述离散分区包含珠粒。

[0162] 在任何实施方案中，所述珠粒为凝胶珠粒。

[0163] 在任何实施方案中，在获得步骤(c)之前，对离散分区内的核酸进行条形编码以形成多个条形码化片段，其中给定离散分区内的片段各自包含共同的条形码，使得条形码鉴定来自给定分区的核酸。

[0164] 在任何实施方案中，获得步骤(c)包括选自由以下组成的组的测序反应：短读段长度测序反应和长读段长度测序反应。

[0165] 在任何实施方案中，所述样品包含肿瘤样品。

[0166] 在任何实施方案中，所述样品包括肿瘤和正常细胞的混合物。

[0167] 在任何实施方案中，所述样品包括核基质。

[0168] 在任何实施方案中，所述核酸包括RNA。

[0169] 在任何实施方案中，样品中核酸的量小于5ng/ml、10ng/ml、15ng/ml、20ng/ml、25ng/ml、30ng/ml、35ng/ml、40ng/ml、45ng/ml或50ng/ml。

[0170] 在一些方面，本公开内容提供分析核酸同时维持结构背景的方法，所述方法包括以下步骤：(a)在样品内形成连接的核酸，使得空间上相邻的核酸区段连接；(b)加工连接的核酸以产生多种连接产物，其中所述连接产物含有空间上相邻的核酸区段的部分；(c)将所述多种连接产物沉积到离散分区中；(d)对所述离散分区内的连接产物进行条形编码以形成多个条形码化片段，其中给定离散分区内的片段各自包含共同的条形码，从而将每个片段与得到其的连接核酸相关联；(e)从所述多个条形码化片段获得序列信息，从而分析来自所述样品的核酸，同时维持结构背景。

[0171] 在进一步的实施方案中，加工步骤(b)包括在有利于分子内连接的条件下的平端连接，使得空间上相邻的核酸区段在同一分子内连接。

[0172] 在任何实施方案中，有利于分子内连接的条件包含稀释样品以将核酸的浓度降低到10ng/μL以下。

[0173] 在任何实施方案中，核酸在步骤(a)之前未从样品中分离。

[0174] 在任何实施方案中，在步骤形成(a)之前，使核酸免疫沉淀，使得相关的DNA结合蛋白保持与核酸结合。

[0175] 在任何实施方案中，所述分区包含珠粒。

[0176] 在任何实施方案中，所述珠粒为凝胶珠粒。

[0177] 在任何实施方案中，所述样品包括肿瘤样品。

[0178] 在任何实施方案中，样品包含肿瘤细胞和正常细胞的混合物。

[0179] 在任何实施方案中，加工步骤包括在形成连接产物之后反转连接。

[0180] 在任何实施方案中，获得步骤(e)提供关于基因组基因座之间的染色体内和/或染色体间相互作用的信息。

[0181] 在任何实施方案中，获得步骤(e)提供关于染色体构象的信息。

[0182] 在任何实施方案中，染色体构象与疾病状态相关联。

[0183] 在任何实施方案中，加工步骤产生包含最初在样品中紧密空间接近的核酸的连接产物。

[0184] 在任何实施方案中，获得步骤(e)包括选自由以下组成的组的测序反应：短读段长度测序反应和长读段长度测序反应。

[0185] 在任何实施方案中，测序反应是短读段高精确度测序反应。

[0186] 在任何实施方案中，形成步骤(a)包括使样品中的核酸交联。

[0187] 在任何实施方案中，形成步骤(a)产生在空间上相邻的核酸区段之间的共价键。

[0188] 在一些方面，本公开内容提供分析核酸同时维持结构背景的方法，所述方法包括以下步骤：(a)在样品内形成连接的核酸，使得空间上相邻的核酸区段连接；(b)将连接的核酸沉积到离散分区中；(c)加工所述连接的核酸以产生多种连接产物，其中所述连接产物含有空间上相邻的核酸区段的部分；(d)对所述离散分区内的连接产物进行条形编码以形成多个条形码化片段，其中给定离散分区内的片段各自包含共同的条形码，从而将每个片段与得到其的连接核酸相关联；(e)从所述多个条形码化片段获得序列信息，从而分析来自所述样品的核酸，同时维持结构背景。

[0189] 在进一步的实施方案中，加工步骤(c)包括在有利于分子内连接的条件下的平端连接，使得空间上相邻的核酸区段在同一分子内连接。

[0190] 在任何实施方案中，所述样品是福尔马林固定的石蜡样品。

[0191] 在任何实施方案中，样品包括核基质。

[0192] 在任何实施方案中，所述核酸包括RNA。

[0193] 在任何实施方案中，核酸在步骤(a)之前未从样品中分离。

[0194] 在任何实施方案中，在形成步骤(a)之前，使核酸免疫沉淀，使得相关的DNA结合蛋白保持与核酸结合。

[0195] 在任何实施方案中，所述分区包含珠粒。

[0196] 在任何实施方案中，所述珠粒为凝胶珠粒。

[0197] 在任何实施方案中，所述样品包括肿瘤样品。

[0198] 在任何实施方案中，样品包括肿瘤和正常细胞的混合物。

[0199] 在任何实施方案中，加工步骤(c)产生包含最初在样品中紧密空间接近的核酸的连接产物。

[0200] 在任何实施方案中，获得步骤(e)提供关于基因组基因座之间的染色体内和/或染色体间相互作用的信息。

[0201] 在任何实施方案中，获得步骤(e)包括选自由以下组成的组的测序反应：短读段长度测序反应和长读段长度测序反应。

[0202] 在任何实施方案中，测序反应是短读段高精确度测序反应。

[0203] 在一些方面，本公开内容提供分析核酸同时维持结构背景的方法，所述方法包括以下步骤：(a)使样品内的核酸交联以形成交联的核酸，其中所述交联在空间上相邻的核酸片段之间形成共价键；(b)将所述交联的核酸沉积到离散分区中；(c)加工所述交联的核酸以产生多种连接产物，其中所述连接产物含有所述空间上相邻的核酸区段的部分；(d)自所述多种连接产物获得序列信息，从而分析来自所述样品的核酸，同时维持结构背景。

[0204] 在进一步的实施方案中，加工步骤(b)包括在有利于分子内连接的条件下的平端连接，使得空间上相邻的核酸区段在同一分子内连接。

[0205] 在任何实施方案中，所述样品是福尔马林固定的石蜡样品。

[0206] 在任何实施方案中，样品包括核基质。

[0207] 在任何实施方案中，所述核酸包括RNA。

[0208] 在任何实施方案中，核酸在交联步骤(a)之前未从样品中分离。

[0209] 在任何实施方案中，样品中核酸的量小于5ng/ml、10ng/ml、15ng/ml、20ng/ml、25ng/ml、30ng/ml、35ng/ml、40ng/ml、45ng/ml或50ng/ml。

[0210] 在任何实施方案中，在交联步骤(a)之前，使核酸免疫沉淀，使得相关的DNA结合蛋白保持与核酸结合。

[0211] 在任何实施方案中，在获得步骤(d)之前，连接产物与条形码相关联。

[0212] 在任何实施方案中，同一分区内的连接产物接收共同的条形码，使得条形码鉴定来自给定分区的连接产物。

[0213] 在任何实施方案中，获得步骤(d)包括选自由以下组成的组的测序反应：短读段长度测序反应和长读段长度测序反应。

[0214] 尽管本文已示出和描述了本发明的优选实施方案，但对于本领域技术人员来说将显而易见，此类实施方案仅作为实例提供。在不脱离本发明的情况下，本领域技术人员现将会想到众多的变化、改变和替代。应了解本文描述的本发明的实施方案的各种替代方案可用于实施本发明。意图在于，上文的权利要求书定义本发明的范围，并且因此可涵盖处于这些权利要求范围内的方法和结构以及其等效物。

实施例

[0215] 实施例1：样品制备

[0216] 修改样品制备方法以提供来自FFPE样品的长DNA分子。图7图示示例性工作流程，其中修改指示制备用于全基因组测序(WGS)和全外显子组测序(WES)的FFPE样品。例如，DNA提取后，在701处修改标准热循环方案以将98度变性步骤从每个循环结束时移至开始。此外，在每个周期结束时增加70度保持，持续2分钟。

[0217] 在循环后清理702和WES文库制备和靶标富集步骤704和705期间，使用超过正常方案的1.8X固相可逆固定化(Solid Phase Reversible Immobilisation，SPRI)珠粒。

[0218] 另一修改包括改变在剪切步骤703期间的条件，其中与具有50的峰值入射功率的标准超声波发生器相反，使用具有约450的峰值入射功率的超声波发生器。

[0219] 可以在某些情况下使用的另外修改是首先用有机催化剂加工FFPE样品以除去甲醛加合物，如例如在Karmakar等人，(2015)，NatureChemistry，DOI:10.1038/NCHEM.2307中描述。此类方案包括将30mMpH 7Tris缓冲液中的5mM有机催化剂添加到样品中以实现加合物反转。有效的有机催化剂包括但不限于水溶性双功能催化剂，诸如Karmakar等人描述的邻氨基苯甲酸盐和氨基苯磷酸盐催化剂。加合物的反转具有提高自样品产生核酸的收率的效应。

[0220] 实施例2：FFPE样品的条形编码

[0221] FFPE样品(其可以包括载玻片上的FFPE样品)可以用以空间明确的模式施加的DNA

条形码如DNA微阵列印刷中使用的那些进行标记。DNA条形码(以下称为条形码-1)为长的，使得它不会在后续步骤中扩散开，或共价地施加到FFPE样品。为了能够条形编码DNA以使其嵌入FFPE载玻片中，可将样品加热，然后添加条形码。条形码通常是条形码文库，使得在载玻片的不同部分提供不同的条形码。条形码也可以在载玻片的不同部分以不同的浓度添加以帮助地理编码-在那种情况下，条形码文库可以包含相同或不同的条形码。在添加条形码之后，然后将载玻片冷却，然后通常通过诸如使用激光显微切割、机械/声学手段等的方式切割而分成各部分。也可以使用荧光团或量子点(Qdots)代替条形码，然而，条形编码能够大量平行地随机包封样品部分，同时保留局部空间信息(例如，肿瘤相对于正常细胞)。

[0222] 然后可以将含有条形码的样品部分放入测序系统中，包括基于液滴的系统，诸如10X Genomics Chromium™系统，使得每个液滴包封单个条形码化部分。

[0223] 样品的脱蜡可以通过加热在液滴中进行。石蜡在水中不混溶，但可溶于某些油中，因此在加热片上液滴后，石蜡可以容易地从液滴中除去。二甲苯也可以用于液-液提取过程中，以对样品部分进行脱蜡并准备其核酸内容物用于进一步加工。

[0224] 其它步骤包括使脱石蜡样品的亚甲基桥解交联。对于该步骤，可以使用专门的化学手段来除去交联并且由此能够获得所含的核酸用于任何后续加工，包括本文讨论的核酸条形编码、扩增和文库制备步骤(参见例如图2)。请注意，空间条形编码DNA也被包封在液滴中。单独核酸的第二条形编码步骤将用于对核酸和条形码(用于空间编码样品)进行条形编码。然后可以将序列读段拼接在一起以提供信息，然后可以将这些信息与样品中的原始空间位置进行比较，并因此与病理数据关联。

[0225] 在该空间编码工作流程的替代版本中，解交联步骤首先在液滴内进行，然后将样品中的核酸(包括基因组DNA以及空间编码条形码)附接到粒子上或以其它方式自样品分离。然后将核酸重新包封并且进行本文所述方法中的条形编码和测序的工作流程，包括图2所绘的方法。

[0226] 本说明书在目前描述的技术的实施例方面提供方法、系统和/或结构以及其用途的完整描述。虽然该技术的多种方面已经如上在一定程度的具体性上或参考一个或多个独特的方面作了描述，但是本领域技术人员不脱离此技术的精神或范畴就能够对公开的方面做大量的改变。因为可以不脱离目前描述的技术的精神和范畴而产生许多方面，适当的范畴存在于下文附加的权利要求书中。因此涵盖其它方面。此外，应理解的是，除非另外明确地主张或主张语言本身需要特定的顺序，否则任何操作可以以任何顺序进行。意在以上描述所含和在附图中示出的所有事物解释为只说明特定方面并且不限于示出的实施方案。除非另外从上下文清晰可见或明确说明，否则本文中提供的任何浓度值通常就混合物值或百分数给出，而不考虑添加混合物的特定组分时或之后发生的任何转化。如果尚未明确地并入本文中，那么在本公开中涉及的所有公开的参考文献和专利文件处于所有目的均以全文引用方式并入本文。如在以下权利要求书中定义，可以进行不脱离本技术的基本要素的细节或结构上的改变。

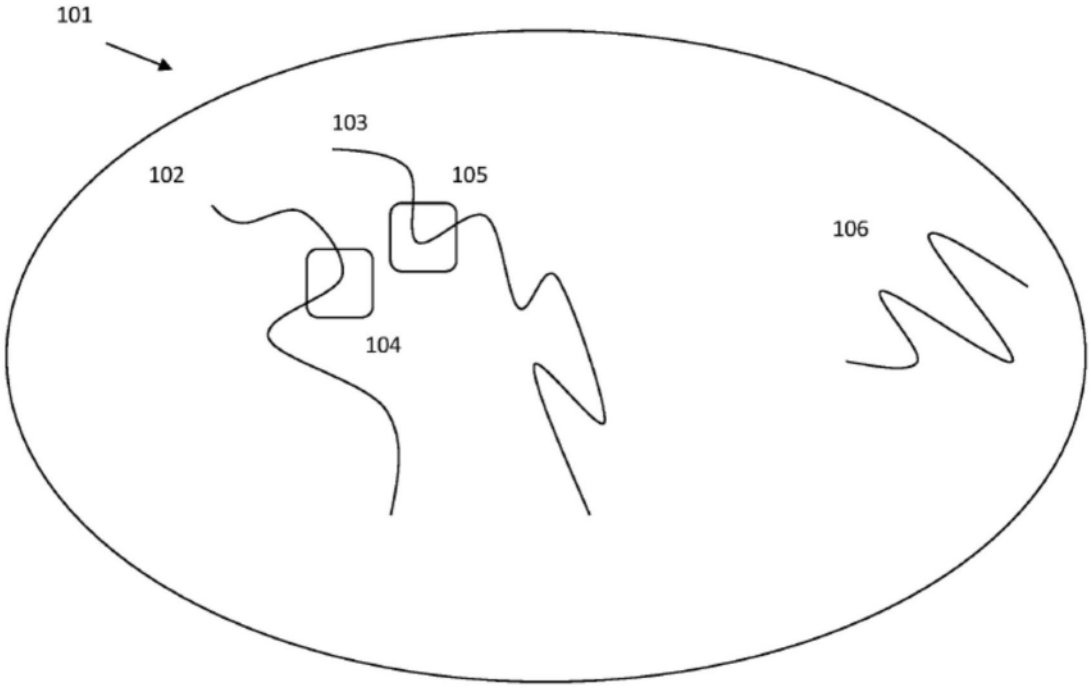

图1

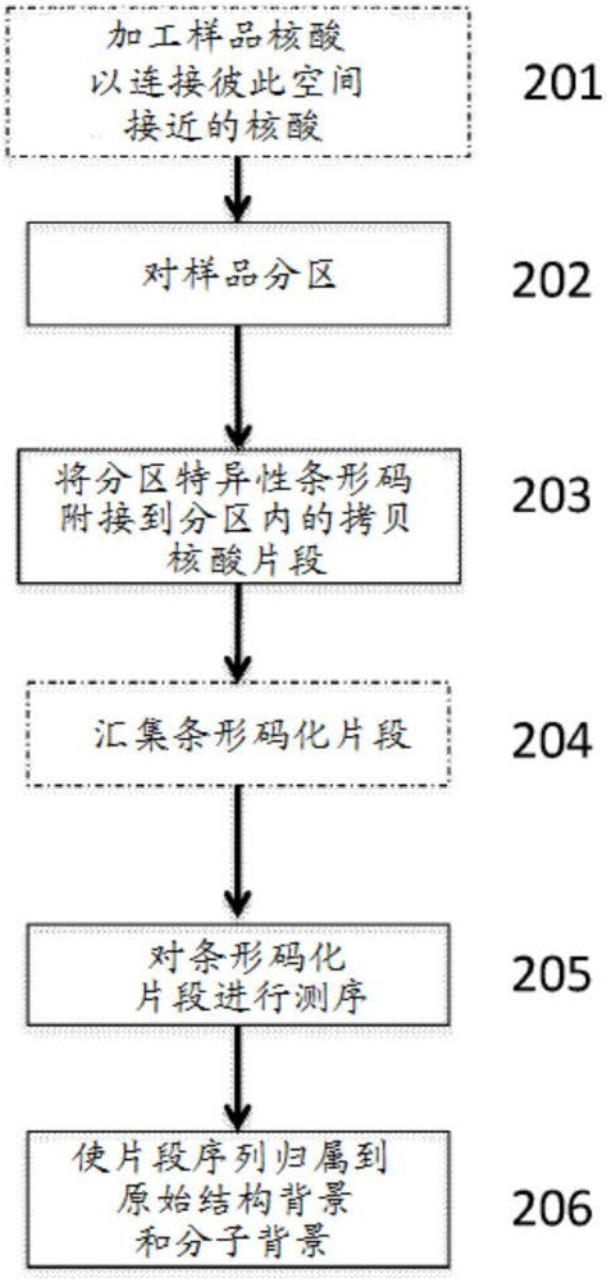
加工样品核酸
以连接彼此空间
接近的核酸
201
对样品分区
202
将分区特异性条形码
附接到分区内的拷贝
核酸片段
203
汇集条形码化片段
204
对条形码化
片段进行测序
205
使片段序列归属到
原始结构背景
和分子背景
206

图2

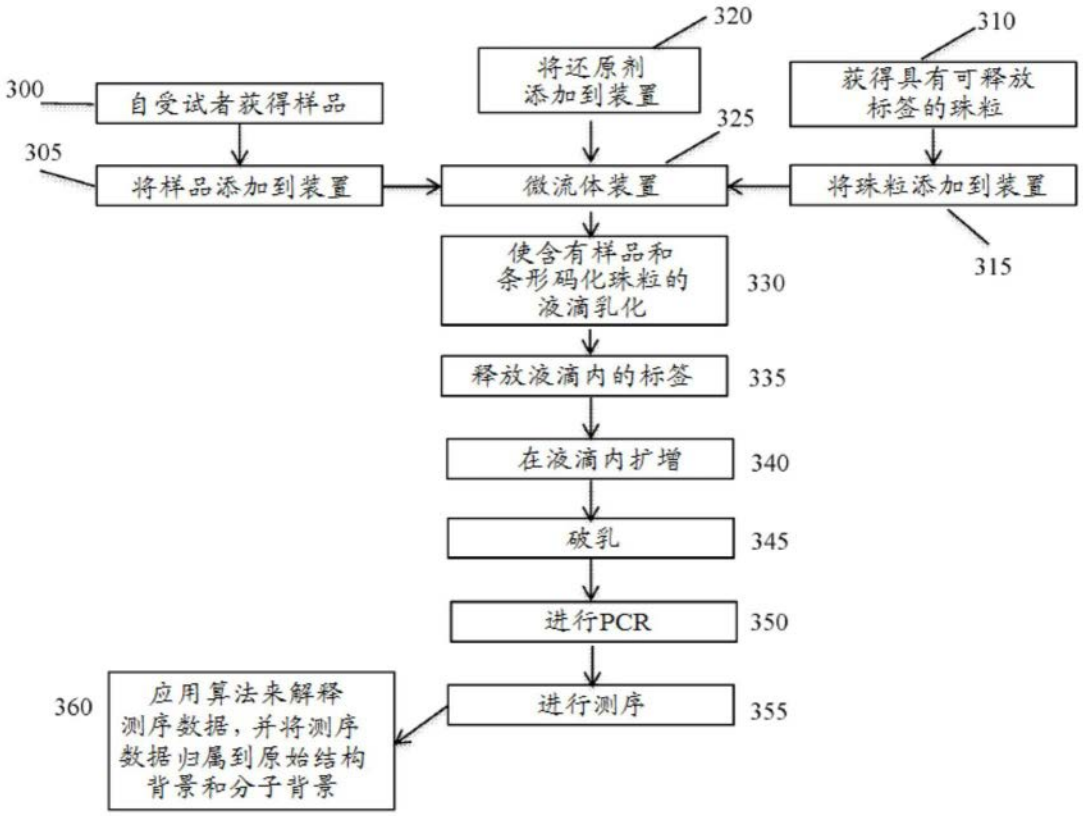
300
自受试者获得样品
305
将样品添加到装置
320
将还原剂
添加到装置
325
微流体装置
310
获得具有可释放
标签的珠粒
将珠粒添加到装置
315
使含有样品和
条形码化珠粒的
液滴乳化
330
释放液滴内的标签
335
在液滴内扩增
340
破乳
345
进行PCR
350
进行测序
355
360
应用算法来解释
测序数据，并将测序
数据归属到原始结构
背景和分子背景

图3

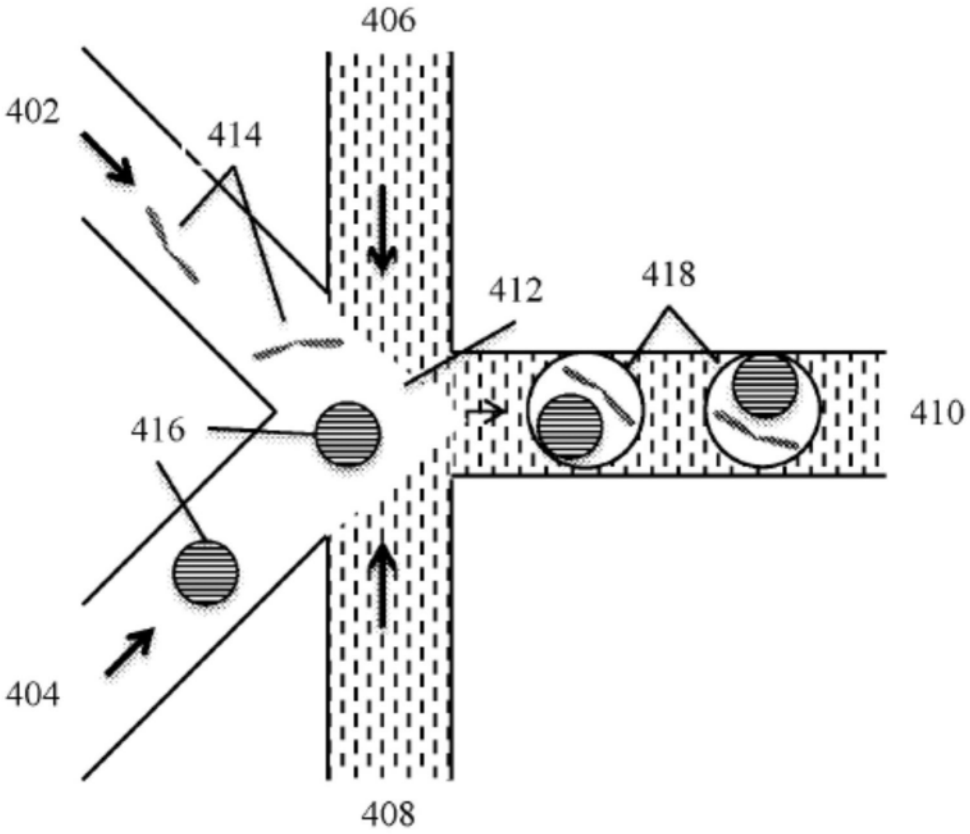

图4

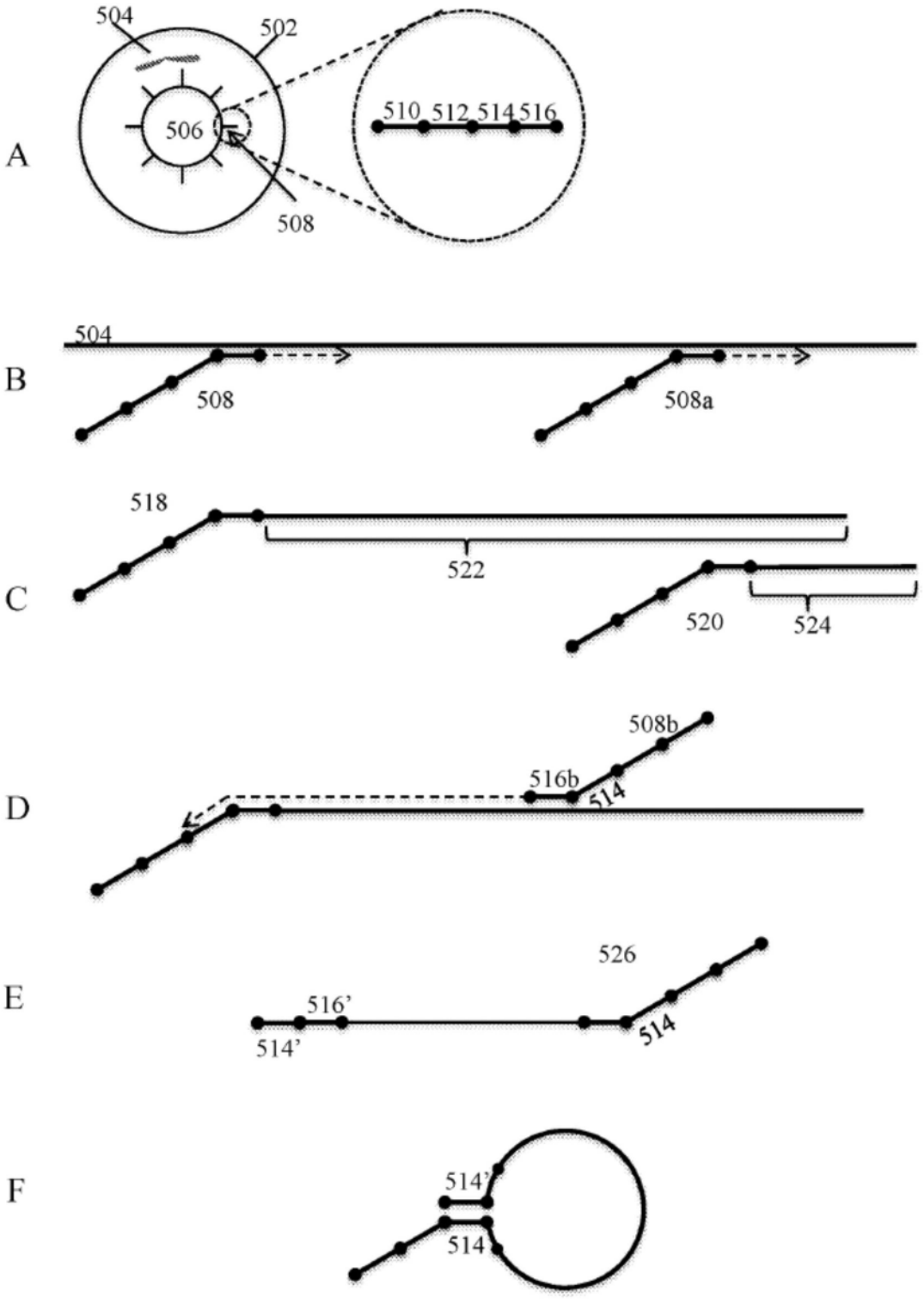
A
504
502
506
508
510 512 514 516
B
504
508
508a
C
518
522
520
524
D
508b
516b
514
E
526
516'
514'
514
F
514'
514

图5

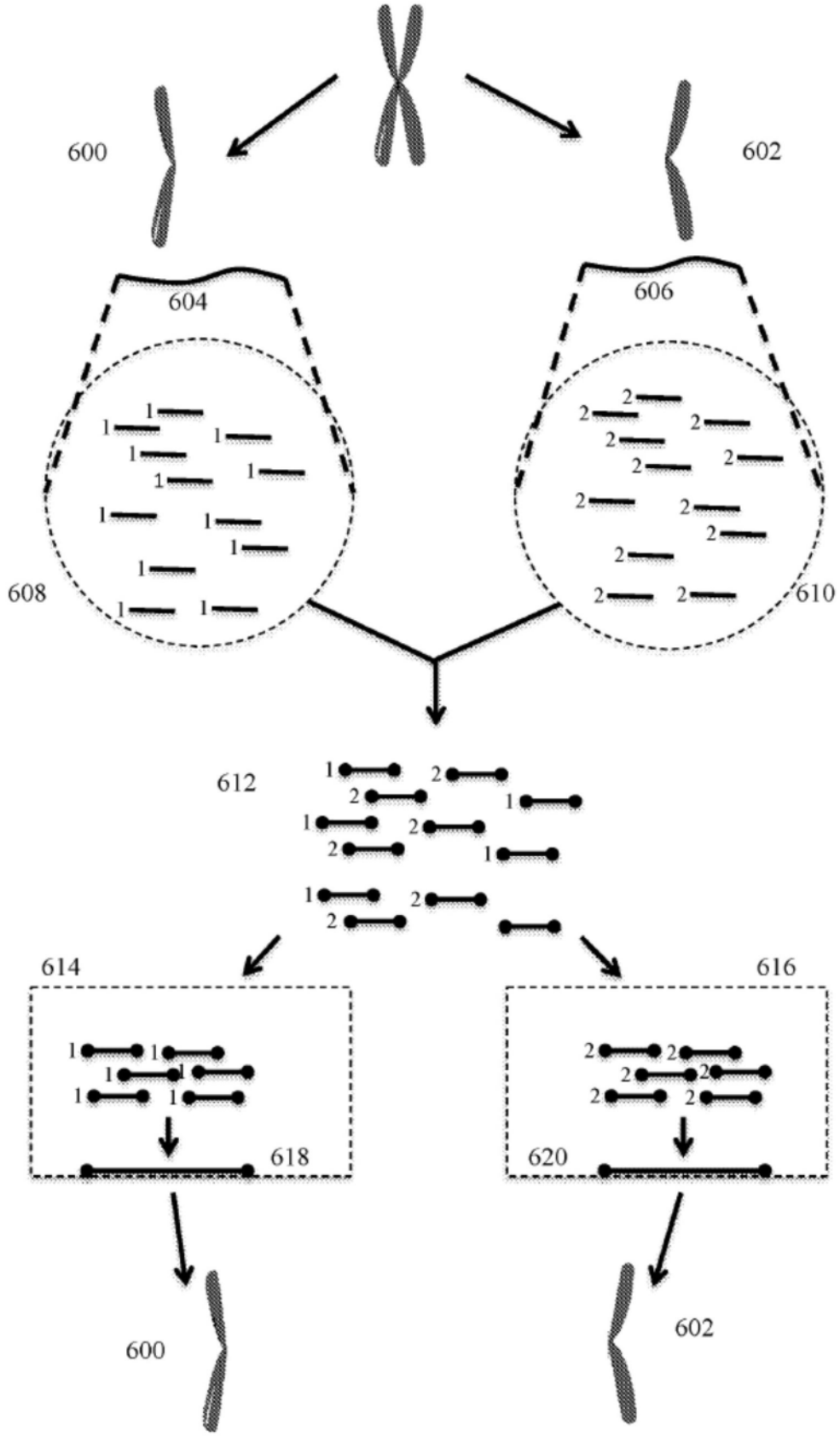

图6

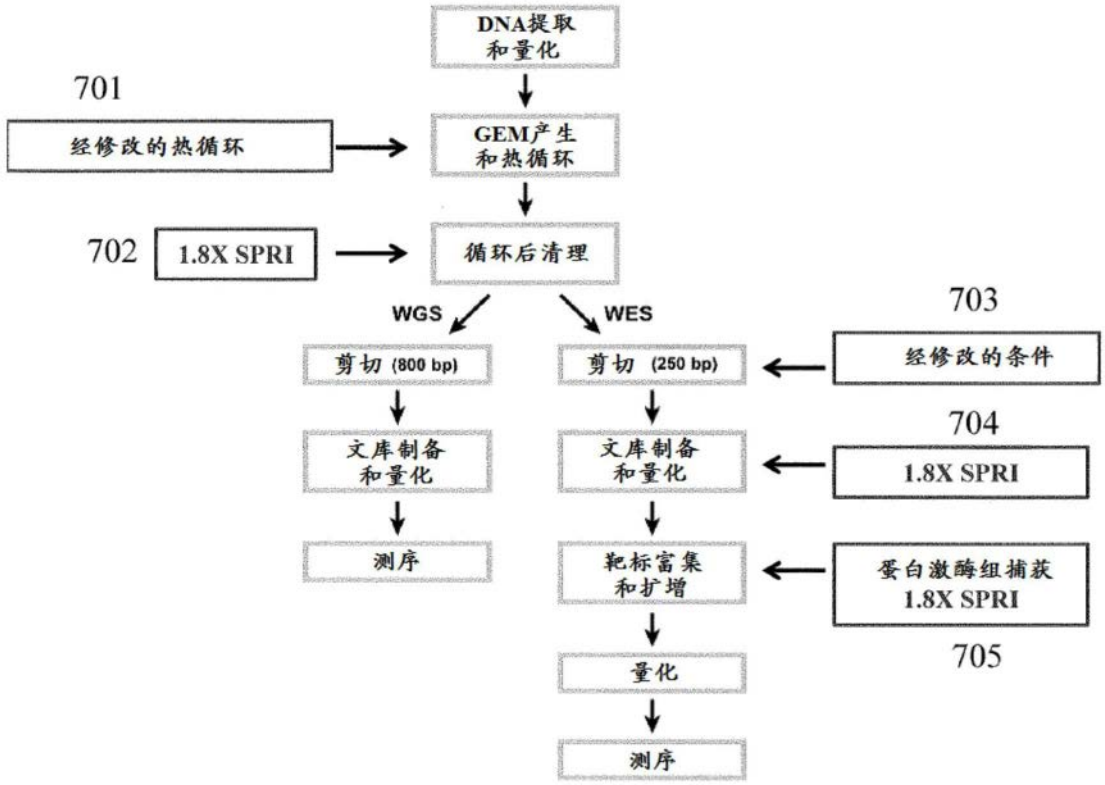
DNA提取
和量化
701
经修改的热循环
GEM产生
和热循环
702
1.8X SPRI
循环后清理
WGS
WES
703
剪切 (800 bp)
剪切 (250 bp)
经修改的条件
文库制备
和量化
704
文库制备
和量化
1.8X SPRI
测序
靶标富集
和扩增
蛋白激酶组捕获
1.8X SPRI
705
量化
测序

图7