(19)国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 115651972 A
(43)申请公布日 2023.01.31

(21)申请号 202210652941.8
(22)申请日 2016.02.24
(30)优先权数据
62/119996 2015.02.24 US
62/146834 2015.04.13 US
(62)分案原申请数据
201680011692.7 2016.02.24
(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 M.施纳尔-莱文 M.贾罗什
(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 罗文锋 彭昶
(51)Int.Cl.
C12Q 1/6869(2018.01)
C12Q 1/6874(2018.01)

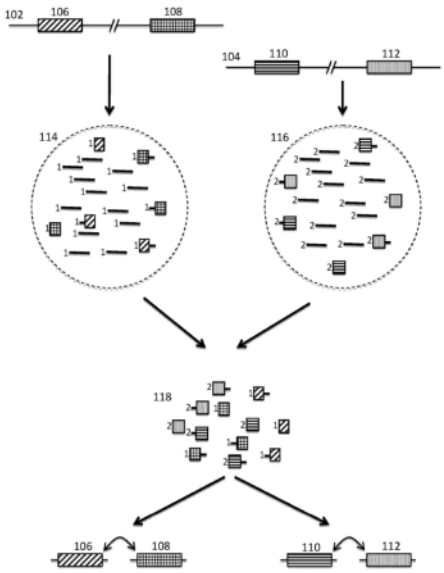
权利要求书2页 说明书29页 附图10页

(54)发明名称

用于靶向核酸序列覆盖的方法

(57)摘要

本发明涉及用于靶向核酸序列覆盖的方法。具体而言,本发明涉及用于分析来自基因组的靶向区域的序列信息的方法、组合物和系统。这种靶向区域可以包括基因组的不良表征、高度多态或与参考基因组序列偏离的区域。



CN 115651972 A

1.一种用于对基因组的一个或多个所选部分进行测序的方法,所述方法包括:

(a)提供起始基因组材料;

(b)将单独核酸分子从所述起始基因组材料分布到离散分区中,使得每个离散分区含有单独核酸分子;

(c)扩增所述离散分区中的至少一些所述单独核酸分子的所选部分以形成扩增子群;

(d)对所述扩增子群进行条形码编码以形成所述扩增子的多个带条形码的片段,其中给定离散分区中的片段各自包含共同条形码,从而将每个片段与其所来源于的所述单独核酸分子关联;

(e)从所述多个片段获得序列信息,从而对基因组的一个或多个所选部分进行测序。

2.一种从基因组样品的一个或多个不良表征的部分获得序列信息的方法,所述方法包括:

(a)在离散分区中提供所述基因组样品的单独第一核酸片段分子;

(b)将所述离散分区内的所述单独第一核酸片段分子片段化以从所述单独第一核酸片段分子产生多个第二片段;

(c)扩增不良表征的所述多个第二片段的所选区域以形成扩增子群;

(d)将共同条形码序列附接至每个所述离散分区内的所述扩增子,以使得每个所述扩增子可归属于其所被包含的所述离散分区;

(e)鉴定所述扩增子的序列,从而从所述基因组样品的一个或多个不良表征的部分获得序列信息。

3.一种从基因组样品的一个或多个部分获得序列信息,同时保持分子环境的方法,所述方法包括:

(a)提供起始基因组材料;

(b)将单独核酸分子从所述起始基因组材料分布到离散分区中,使得每个离散分区含有第一单独核酸分子;

(c)提供富集包含来自所述基因组的一个或多个所选部分的至少一部分的片段的群体;

(d)将共同条形码序列附接至每个离散分区内的所述片段,以使得每个所述片段可归属于其所被包含的所述离散分区;

(e)从所述片段获得序列信息,从而在保留分子环境的同时对所述基因组样品的一个或多个靶向部分进行测序。

4.一种从基因组样品的一个或多个部分获得序列信息,同时保持分子环境的方法,所述方法包括:

(a)提供起始基因组材料;

(b)将单独核酸分子从所述起始基因组材料分布到离散分区中,使得每个离散分区含有第一单独核酸分子;

(c)提供在所述离散分区中的至少一些内富集包含所述基因组的一个或多个所选部分的至少一部分的所述片段的序列的群体;

(d)将共同条形码序列附接至每个离散分区内的所述片段,以使得每个所述片段可归属于其所被包含的所述离散分区;

(e)分离含有包含所述基因组的一个或多个所选部分的至少一部分的片段的离散分区与不含有包含所述基因组的一个或多个所选部分的片段的离散分区；

(f)从包含所述基因组的一个或多个所选部分的至少一部分的所述片段获得序列信息，从而在保持分子环境的同时对所述基因组样品的一个或多个靶向部分进行测序。

5.一种从基因组样品的一个或多个部分获得序列信息，同时保持分子环境的方法，所述方法包括：

(a)提供基因组材料；

(b)将单独核酸分子从所述基因组材料分离，以形成分离的单独核酸分子；

(c)提供富集包含来自所述分离的单独核酸分子的所述基因组的一个或多个所选部分的至少一部分的片段的群体，其中所述片段中的至少多个片段可归属于其所来源于的所述单独核酸分子；

(d)从所述片段获得序列信息，从而在保留分子环境的同时对所述基因组样品的一个或多个靶向部分进行测序。

用于靶向核酸序列覆盖的方法

[0001] 本申请为分案申请，原申请的申请日为2016年2月24日，申请号为201680011692.7(PCT/US2016/019382)，发明名称为“用 于靶向核酸序列覆盖的方法”。

[0002] 相关申请的交叉引用

[0003] 本申请要求2015年4月13日提交的美国临时申请No.62/146,834 和2015年2月24日提交的美国临时申请No.62/119,996的权益，所 述专利在此出于所有目的以引用的方式整体并入。

背景技术

[0004] 尽管测序技术取得了显著的进步，但人基因组中的约5％至10％ 保持未装配、未作图，并且不良表征。参考组装序列(reference assembly) 通常将这些缺少区域注释为多兆碱基异染色质空位。基因组的此缺少 部分包括保持对使用通常所用的测序技术进行精确表征的抗性的结 构特征。对整个基因组进行从头测序在经济上是不可行的，因此仍然需要降低与基因组测序相关的成本，同时保留大规模基因组分析的益 处。

发明内容

[0005] 因此，本公开提供了用于提供对基因组的所选区域的靶向覆盖以 允许对那些所选区域进行从头序列装配，并且在一些方面允许将从头 覆盖与基因组的剩余区域的高通量和高精度的重新测序组合的方法、系统和组合物。

[0006] 在一些方面，本公开提供了一种对基因组的一个或多个所选部分 进行测序的方法，其中该方法包括以下步骤：(a)提供起始基因组材料；(b)将单独核酸分子从起始基因组材料分布到离散的分区(partition) 中，使得每个离散的分区含有单独核酸分子；(c)扩增该离散分区中的 至少一些单独核酸分子的所选部分以形成扩增子群；(d)对该扩增子群进行条形码编码以形成扩增子的多个带条形码的片段，其中给定离散 分区中的片段各自包含共同条形码，从而将每个片段与其所来源于的 单独核酸分子关联；(e)从多个片段获得序列信息，从而对基因组的一 个或多个所选部分进行测序。

[0007] 在其它实施方案中并且根据上述，该基因组的一个或多个所选部 分包含该基因组的高度多态性区域。在其它实施方案中，该基因组的 一个或多个所选部分的测序是从头测序。

[0008] 在其它实施方案中并且根据上述的任一项，扩增包括跨至少3.5 兆碱基对(Mb)的区域的PCR扩增。在其它实施方案中，扩增包括利 用跨至少3.0Mb的区域交错的多个引物对的PCR扩增。

[0009] 在一些实施方案中并且根据上述的任一项，测序反应是短读段、高精度的测序反应。在其它实施方案中，在获得步骤中产生的序列信 息保留了其来源的单独核酸的分子环境(molecular context)。

[0010] 在某些实施方案中并根据上述的任一项，在该获得步骤之前，通 过以下步骤从该多个片段进一步富集包含基因组的一个或多个所选 部分的至少一部分的片段：(i)将与该

基因组的一个或多个所选部分中 或附近的区域互补的探针与该片段杂交以形成探针-片段复合物;(ii) 将探针-片段复合物捕获到固体支撑物的表面。

[0011] 在一些实施方案中并且根据上述的任一项,离散分区内的扩增子 的带条形码的片段代表对基因组的一个或多个所选部分的约 100X-5000X的覆盖。在其它实施方案中,离散分区内的扩增子的带 条形码的片段代表对基因组的一个或多个所选部分的约200X-1000X 的覆盖。在其它实施方案中,离散分区内的扩增子的带条形码的片段 代表对基因组的一个或多个所选部分的至少1000X的覆盖。在其它 实施方案中,离散分区内的扩增子的带条形码的片段代表对基因组的 一个或多个所选部分的至少2000X或5000X的覆盖。

[0012] 在其它方面,本公开提供了一种从基因组样品的一个或多个不良 表征的部分获得序列信息的方法,其中该方法包括以下步骤:(a)在离 散分区中提供基因组样品的单独第一核酸片段分子;(b)将离散分区内 的单独第一核酸片段分子片段化以从该单独第一核酸片段分子产生 多个第二片段;(c)扩增不良表征的该多个第二片段的所选区域以形成扩增子群;(d)将共同条形码序列附接至每个离散分区内的扩增子,以 使得每个扩增子可归属于其所被包含的离散分区;(e)鉴定扩增子的序 列,从而从基因组样品的一个或多个不良表征的部分获得序列信息。

[0013] 在某些实施方案中,并且根据上述的任一项,扩增包括跨至少 3.5兆碱基对(Mb)的区域的PCR扩增。在其它实施方案中,扩增包括 利用跨至少3.0Mb的区域交错的多个引物对的PCR扩增。在其它实 施方案中,多个引物对含有尿嘧啶以防止引物序列扩增。

[0014] 在一些实施方案中,并且根据上述的任一项,鉴定步骤保留扩增 子序列的分子环境,使得鉴定进一步包括鉴定来源于相同的单独第一 核酸片段分子的扩增子。在其它实施方案中,该方法还包括基于该多 个第二片段的重叠序列来在推断的重叠群中连接两个或更多个单独 第一片段分子,其中该推断的重叠群包含至少10kb的长度N50。

[0015] 在一些实施方案中,并且根据上述的任一项,条形码序列还包括 附加的序列区段。在其它实施方案中,附加的序列区段包含选自由以 下项组成的组的一个或多个成员:引物、附接序列、无规n聚体寡核 苷酸、包含尿嘧啶核碱基的寡核苷酸。在其它实施方案中,条形码是 从至少70万条条形码的文库中选择的。

[0016] 在一些实施方案中,并且根据上述的任一项,每个离散分区内的 基因组样品包含来自单个细胞的基因组DNA。在其它实施方案中, 每个离散分区包含来自不同染色体的基因组DNA。

[0017] 在一些实施方案中,并且根据上述的任一项,离散分区包含乳液 中的液滴。

[0018] 在一些实施方案中,并且根据上述的任一项,离散分区内的带条 形码的扩增子代表对基因组的一个或多个不良表征的部分的约 1000X-5000X的覆盖。

[0019] 在其它方面,本申请提供了一种从基因组样品的一个或多个部分 获得序列信息,同时保留分子环境的方法,该方法包括以下步骤:(a) 提供起始基因组材料;(b)将单独核酸分子从起始基因组材料分布到离 散分区中,使得每个离散分区含有第一单独核酸分子;(c)提供富集包 含所述基因组的所述一个或多个所选部分的至少一部分的片段的群 体;(d)将共同条形码序列附接至每个离散分区内的片段,以使得每个 片段可归属于其所被包含的离散分区;(e)从该片段获得序列信息,从 而在保持分子环境的同时对基因组样品的一个或多个靶向部分进行 测序。

[0020] 在其它方面，本公开提供了一种从基因组样品的一个或多个部分 获得序列信息，同时保留分子环境的方法，该方法包括以下步骤：(a) 提供起始基因组材料；(b)将单独核酸分子从起始基因组材料分布到离 散分区中，使得每个离散分区含有第一单独核酸分子；(c)提供至少一 些离散分区内富集包含基因组的一个或多个所选部分的至少一部分 的片段的序列的的群体；(d)将共同条形码序列附接至每个离散分区内 的片段，以使得每个片段可归属于其所被包含的离散分区；(e)分离含 有包含基因组的一个或多个所选部分的至少一部分的片段的离散分 区与不含有包含基因组的该一个或多个所选部分的片段的离散分区； (f)从包含基因组的该一个或多个所选部分的至少一部分的片段获得 序列信息，从而在保留分子环境的同时对基因组样品的一个或多个靶 向部分进行测序。

[0021] 在其它实施方案中并根据上述的任一项，提供富集包含基因组的 一个或多个所选部分的至少一部分的片段的序列的群体的步骤包括 对包含基因组的一个或多个所选部分的至少一部分的片段进行定向 PCR扩增，以产生包含基因组的一个或多个所选部分的至少一部分的 扩增子群。在其它实施方案中，此提供步骤还包括将可检测标签附接 至扩增子，该可检测标记在一些实施方案中可以包括荧光分子。在其 它实施方案中，该分离含有包含基因组的一个或多个所选部分的至少 一部分的片段的离散分区与不含有包含基因组的一个或多个所选部 分的片段的离散分区的步骤包括分选发射来自可检测标签的信号的分区与没有此类信号的分区。

[0022] 在一些实施方案中并根据上述的任一项，在从片段获得序列信息 之前，将各离散分区混合并将各片段汇集在一起。在其它实施方案中， 从片段获得序列信息的步骤是以维持片段序列的分子环境的方式进 行的，使得鉴定进一步包括鉴定来源于相同的第一单独核酸分子的片 段。在其它实施方案中，此获得序列信息的步骤包括选自由以下项组 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 其它实施方案中，测序反应是短读段、高精度测序反应。

[0023] 在一些实施方案中并且根据上述的任一项，离散分区包含乳液中 的液滴。在其它实施方案中，离散分区内的带条形码的片段代表对基 因组的一个或多个所选部分的约100X-5000X的覆盖。在其它实施方 案中，离散分区内的带条形码的片段代表对基因组的一个或多个所选 部分的约200X-1000X的覆盖。在其它实施方案中，离散分区内的带 条形码的片段代表对基因组的一个或多个所选部分的至少1000X的 覆盖。在其它实施方案中，离散分区内的带条形码的片段代表对基因 组的一个或多个所选部分的至少2000X或5000X的覆盖。

[0024] 在一些方面并根据上述的任一项，本公开提供了从基因组样品的 一个或多个部分获得序列信息，同时保留分子环境的方法，该方法包 括以下步骤：(a)提供起始基因组材料；(b)从基因组材料中分离单独 核酸分子以形成分离的单独核酸分子；(c)提供富集包含来自分离的单 独核酸分子的基因组的一个或多个所选部分的至少一部分的片段的 群体。在某些实施方案中，该分离步骤是使用允许一个或多个核酸分 子与其它一个或多个核酸分子相对分离地被分选和加工的任何方法 完成的。在一些实施方案中，分离是物理分离成基底上的不同隔室或 分离成不同的分区。在其它实施方案中，至少多个片段可归属于它们所来源的单独核酸分子。此归属是使用允许将特定片段指定为来源于 特定单独核酸分子的任何方法获得的。在某些示例性实施方案中，此 归属是通过对片段进行条形码编码获得

的。在其它方面,从片段获得 序列信息,从而在保留分子环境的同时对基因组样品的一个或多个靶 向部分进行测序。

附图说明

[0025] 图1提供了使用常规方法对比本文所述的方法和系统进行的靶 向基因组区域鉴定和分析的示意图。

[0026] 图2提供了使用本文所述的方法和系统进行的靶基因组区域鉴 定和分析的示意图。

[0027] 图3示出了使用本文公开的方法和组合物执行测定以检测序列 信息的典型工作流程。

[0028] 图4提供了用于将核酸样品与珠粒合并以及将核酸和珠粒分隔 成离散液滴的方法的示意图。

[0029] 图5提供了用于对染色体核酸片段进行条形码编码和扩增的方 法的示意图。

[0030] 图6A和6B提供了对核酸片段进行条形码编码在将序列数据归 属于其起始来源核酸分子方面的用途的示意图。

[0031] 图7提供了本发明的一个实施方案的示意图。

[0032] 图8提供了本发明的一个实施方案的示意图。

[0033] 图9示出了来自比较使用模板进行的扩增反应与不含模板的 (NTC)那些扩增反应的实验的数据。

[0034] 图10示出了来自在退火温度范围内进行的扩增反应的数据。

具体实施方式

[0035] 除非另有说明,否则本发明的实践可以采用本领域技术范围内的 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、生物化学和免疫学的常规技术和描述。这种常规技术包括聚合物阵列 合成、杂交、连接、噬菌体展示以及使用标签检测杂交。可以参考下 文的实例具体说明合适的技术。然而,当然也可以使用其它等效的常 规程序。这些常规技术和描述可见于标准实验室手册中,诸如Genome Analysis:A Laboratory Manual Series(第I-IV卷),Using Antibodies:A Laboratory Manual,Cells:A Laboratory Manual,PCR Primer:A Laboratory Manual以及Molecular Cloning:A Laboratory Manual(均得 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 ofBiochemistry第3版,W.H.Freeman Pub.,New York,N.Y.以及Berg等人(2002)Biochemistry,第5版,W. H.Freeman Pub.,New York,N.Y.,该文献均出于所有目的以引用方式 整体并入本文。

[0036] 应当注意,如本文和随附权利要求书中所使用,单数形式“一种”、“一个”和“该”包括复数指示物,除非上下文另有明确规定。因此,例 如,提及“一种聚合酶”是指一种试剂或此类试剂的混合物,并且提及 “该方法”包括对本领域的技术人员已知的等效步骤和方法的提及,等 等。

[0037] 除非另有定义,否则本文所使用的所有技术和科学术语具有与本 发明所属领域的普通技术人员通常理解的相同的含义。本文提及的所 有出版物出于描述和公开在出版物中描述并且可以结合当前描述的 发明使用的装置、组合物、制剂和方法的目的而以引用方式并入本文。

[0038] 在提供值范围的情况下,应理解,除非上下文另有明确规定,否 则每个居间值至该下限单位的十分之一在该范围的上限和下限之间, 并且任何其它所述范围或该所述范围中的居间值被涵盖在本发明内。这些较小范围的上限和下限可以独立地包括在该较小范围内,也被涵 盖在本发明内,经受所述范围中任何特别排除的极限。在所述范围包 括上限和下限中的一个或两个的情况下,排除了这些所包括的极限中 的一者或两者的范围也被包括在本发明中。

[0039] 在下面的描述中,阐述了许多具体细节来提供对本发明的更透彻 理解。然而,对于本领域的技术人员将显而易见的是,本发明可以在 没有这些具体细节中的一个或多个的情况下实践。在其它情况下,未 描述本领域的技术人员熟知的众所周知的特征和程序,以避免模糊本 发明。

[0040] 如本文所用,术语“包括/包含”旨在表示组合物和方法包括引用的 要素,但不排除其它要素。“基本上由.....组成”当用于限定组合物和 方法时,应表示排除其它要素对该组合物或方法具有任何显著重要 性。“由.....组成”应表示排除所要求保护的组合物和基本方法步骤中 的其它成分的元素超过痕量。由这些过渡术语中的各者限定的实施方案都在本发明的范围内。因此,该方法和组合物意图可以包括附加步 骤和组分,其(包括/包含)或者替代地包括不重要的步骤和组合物(基 本上由.....组成),或者替代地仅意图指所述方法步骤或组合物(由..... 组成)。

[0041] 所有的数字标记,例如pH、温度、时间、浓度和分子量,包括 范围在内,是以0.1的增量变化(+)或(-)的近似值。应当理解的是,虽 然并不总是明确地叙述为所有数字标记之前是术语“约”。术语“约”除 了“X”的较小增量(诸如“X+0.1”或“X-0.1”)之外,还包括确切值“X”。 还应当理解的是,尽管并不总是明确叙述,但是本文所述的试剂仅是 示例性的,并且其等效物是本领域中已知的。

[0042] I.综述

[0043] 本公开提供了可用于表征遗传物质的方法、组合物和系统。具体 地,本文所描述的方法、组合物和系统提供了对基因组的所选部分的 增加和冗余的覆盖,使得可以从基因组的那些所选部分获得附加的冗 余序列信息。在特定情况中,该附加序列信息提供了足够的信息以允 许对基因组的那些所选部分进行从头测序。

[0044] 通常,本文所述的方法、组合物和系统提供对基因组的所选区域 的遗传表征。这种遗传表征具有足够的深度以允许对基因组的所选区 域进行从头测序。这种从头测序对于基因组的不良表征、高度多态和 /或与参考序列偏离的区域特别有用。应当理解,人基因组的显著百 分比(根据例如Altemose等人,PLOS Computational Biology,2014年 5月15日,第10卷,第5期为至少5%至10%)保持未装配、未作图 并且不良表征。参考组装序列通常将这些缺少区域注释为多兆碱基异 染色质空位,这些缺少区域主要存在于着丝粒附近并且在近端着丝染 色体的短臂上。基因组的此缺少部分包括保持对使用通常所用的测序技术进行精确表征的抗性的结构特征。对精确表征具有抗性的示例性 区域包括具有近似

同源假基因的区域(例如SMN1/2CYP2D6)、在整 个基因组中具有基本上重复的序列的区域，包括但不限于转座子(例 如SINE、LINE)，特别是参考序列用作不良引导的具有巨大差异的区 域(诸如编码人白细胞抗原(HLA)复合物的基因的区域)。本文所述的 方法、组合物和系统将所关注区域的选择性扩增与维持分子环境的能 力结合在一起，从而允许对通常不良表征的基因组区域进行从头测 序，以及任选地提供这些区域在较大基因组中的长距离分子环境。

[0045] 在特定情况下，本文描述的方法包括在测序之前选择性扩增基因 组的所选区域的步骤。通常使用本领域中已知的方法(包括但不限于 PCR扩增)进行的此扩增提供了对基因组的所选区域的至少1X、10X、 20X、50X、100X、200X、500X、1000X、1500X、2000X、5000X 或10000X的覆盖，从而提供一定量的核酸以允许对那些所选区域进 行从头测序。在其它实施方案中，该扩增提供对基因组的所选区域的 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的覆盖。

[0046] 该扩增通常是通过与基因组的所选区域内或附近的序列互补的 引物的延伸进行的。在一些情况下，使用被设计成跨所关注区域拼接 (tile)的引物文库——换句话说，该引物文库被设计成扩增沿着基因组 的所选区域在特定距离处的区域。在一些情况下，选择性扩增利用与 沿着基因组的所选区域的每10个、15个、20个、25个、50个、100 个、200个、250个、500个、750个、1000个或10000个碱基互补 的引物。在其它实例中，该拼接的引物文库被设计成捕获各距离的混 合物——此混合物可以是各距离的随机混合或被智能地设计成使得 所选区域的特定部分或百分比被使用不同的引物对扩增。

[0047] 通常，本文所描述的方法和系统通过提供对基因组的所选区域的 序列的确定来实现靶向基因组测序，并且此测序信息是使用具有短读 段测序技术的极低测序错误率和高通量优点的方法获得的。

[0048] 核酸测序通常是以保留序列读段的分子环境或序列读段的部分 的分子环境的方式进行的。这意味着多个序列读段或序列读段的多个 部分可归属于单一的来源核酸分子。“可归属于”意指序列读段能够被 鉴定为落入其特定来源核酸分子的线性碱基序列内——换句话说，如 果片段1和2是从来源核酸分子A产生的，那么测序以使得来自片 段1、2、3和4的序列读段保留其分子环境并且容易确定片段1和2 来源于来源分子A的方式进行。

[0049] 虽然这种单一核酸单分子可以是各种长度中的任何一种，但是在 优选方面，其将是相对较长的分子，以允许保留长距离分子环境。具 体地，单一来源分子优选基本上长于典型的短读段序列长度，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，并且在一些情况下高达1兆碱基或 更长。

[0050] 通常，如图1所示，本文所述的方法和系统可以用于表征核酸， 尤其是来自基因组的所选区域的核酸，同时保留分子环境。如图所示， 示出了两个离散的单独核酸102和104，其各自具有多个所关注区域， 例如核酸102中的区域106和108，以及核酸104中的区域110和112。每个核酸中的所关注区域在相同核酸分子内连接(例如，来源于相同 核酸分子)，

但是在一些情况下,这些区域可以彼此相对分离,例如 间隔超过1kb、间隔超过5kb、间隔超过10kb、间隔超过20kb、间隔 超过30kb、间隔超过40kb、间隔超过50kb,并且在一些情况下间隔 多达100kb或更大。所关注区域通常是基因组的离散和分开的部分 ——在一些情况下,此类区域是不良表征的区域。所关注区域也可以 表示单独基因、基因群、外显子。如图所示,分离各个核酸102和 104。如图1所示,每个核酸分别被分离成其自身的分区114和116; 然而,应当理解,本文所描述的方法不限于使用此类分区,并且可以 使用任何分离核酸分子的方法,然后可以根据本文所公开的任何方法 来进一步处理那些分离的核酸分子。如本文中其它地方所述,图1中 的分区114和116在许多情况下是油包水乳液中的含水液滴。在每个 液滴内,每个片段的各部分以保留那些片段的原始分子环境的方式被 拷贝,例如为来源于相同的分子。可以使用任何允许将片段归属于其 所来源于的原始核酸分子的方法来保存这种分子环境。如图1所示, 一种实现此的方法是通过在每个拷贝的片段中包含条形码序列,例如 如图所示的条形码序列"1"或"2",该条形码序列代表来源片段被分隔成的液滴。对于全基因组序列分析应用,可以简单地汇集所有拷贝的 片段及其相关条形码,以便对来自各个来源核酸102和104的全长序 列信息(full range sequence information)进行测序和重新装配。然而, 在许多情况下,更期望仅分析整个基因组的特定靶向部分,以便提供 对基因组的科学相关部分的更大重视,并将执行对基因组的不太相关 或不相关部分的测序的时间和代价最小化。有助于保留分子环境的其 它测序方法包括单分子测序方法,诸如可从Pacific Biosciences获得 的SMRT测序,以及由例如OxfordNanopore描述的纳米孔测序以及 可得自Illumina,Inc的Truseq SLR方法。

[0051] 根据上述,除了条形码编码步骤之外,可以存在一个或多个选择 性扩增步骤,使得如果核酸102或104含有所关注的所选基因组区域, 则来自那些区域的扩增子将形成各个分区114和116中的片段的较大 百分比。这种扩增步骤将通常在根据本文所述的方法附接条形码之前 或同时进行,尽管在一些实施方案中,该扩增步骤也可以在条形码附 接之后发生。

[0052] 因为文库118中的汇集片段保留了它们原始分子环境,例如,通 过保留条形码信息,它们可以被重新装配成具有嵌入的(有时是长距 离的)连接信息的原始分子环境,例如,具有在所关注的各装配区域 106:108和110:112之间的推断连接。举例来说,可以鉴定基因组的 两个不同靶向部分之间的直接分子连接,例如两个或更多个外显子, 并且此直接分子连接可用于鉴定结构变化和其它基因组特征。对于其 中利用选择性扩增来增加含有基因组的所选区域的部分的核酸片段 的量的情况,则鉴定分子环境的能力还提供了一种对基因组的那些所 选区域(通常在允许对那些区域进行从头装配的覆盖深度处)进行测序 的方式。

[0053] 在某些情况下,本文描述的测序方法包括对所选区域的深覆盖与 跨基因组的较长范围的较低程度连接读段的组合。应当理解,从头和 重新测序的这种组合提供了对整个基因组和/或基因组的大部分进行 测序的有效方式。通过本文所述的选择性扩增方法对不良表征的和/ 或高度多态性区域进行靶向覆盖提供了那些区域的从头序列装配所 必需的覆盖水平的核酸材料量,而在该基因组的其它区域上的连锁基 因组测序允许通过提供关于通过保留其分子环境而连接在一起的离 散区域的序列信息来对基因组的其余部分进行高通量分析。本文所述 的方法和组合物独特地适于允许从头和连锁读段测序的组合,这是

因 为相同的测序平台和测序文库可用于这两种类型的覆盖。根据本文所 述的方法测序的核酸和/或核酸片段的群体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——覆盖用于 从头测序的所关注区域的核酸的比例高于覆盖基因组其它区域的核 酸的比例,这是由于本文进一步详细描述的靶向扩增方法。由于本文 描述的方法允许在装配期间保留相位信息,所以这种方法进一步适用 于单倍型的从头装配。

[0054] 除了提供从基因组的所选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,包括但不限于单 倍型定相、结构变异的鉴定以及鉴定拷贝数变异,如在美国专利申请 No.14/752,589和14/752,602中所描述的,该专利申请出于所有目的 并且具体地是关于针对基因组材料表征的所有书面描述、图式和工作 实施例而全文以引用方式并入本文。

[0055] 根据本申请中描述的方法和系统的核酸加工和测序方法还进一 步详细描述于美国专利申请No.14/316,383;No.14/316,398;No. 14/316,416;No.14/316,431;No.14/316,447;以及No.14/316,463中, 该专利申请出于所有目的并且具体地关于针对加工核酸和基因组材 料的测序和其它表征的所有书面描述、图式和工作实施例而以引用方 式整体并入本文。

[0056] 通常,本发明的方法包括如图2所示的步骤,其提供了在本文中 进一步详细论述的本发明的方法的示意性综述。应当理解,图2中概 述的方法是可以根据需要并如本文所述改变或修改的示例性实施方 案。

[0057] 如图2所示,本文所述的方法在大多数实例中将包括其中将含有 所关注的靶向区域的样品核酸例如分离成分区的步骤(201)。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靶向富集,以产 生其中大部分将含有来自所选基因组区域的序列的片段群体(202)。然 后,通常通过将对其被包含于其中的分区具有特异性的片段进行条形 码编码,来以保留片段的原始分子环境的方式将这些片段进一步片段 化或拷贝(203),尽管可以使用归属各片段的原始分子环境的任何其它 方法。在一些实例中,每个分区可以包含多于一个核酸,并且在一些 情况下将含有数百个核酸分子——在多个核酸位于分区的情况下,基 因组的任何特定基因座将通常在条形码编码之前用单个单独核酸表 示。可以使用本领域中已知的任何方法来产生步骤203的条形码片段 ——在一些实例中,寡核苷酸是不同分区内的样品。这种寡核苷酸可 以包括旨在随机引发样品的多个不同区域的无规序列,或者它们可以 包含靶向用于引发样品靶向区域上游的特异性引物序列。在其它实例中,这些寡核苷酸还含有条形码序列,使得复制过程也对原始样品核 酸的所得复制片段进行条形码编码。此类条形码可以使用本领域中已 知的任何方法来添加,包括在扩增各个核酸分子的区段的扩增方法期 间添加条形码序列,以及使用转座子将条形码插入原始的各个核酸分 子内,包括诸如在Amini等人,Nature Genetics 46:1343-1349 (2014)(2014年10月29日提前在线发布)中描述的那些的方法。用于 在对样品进行扩增和条形码编码的过程中使用这些条形码寡核苷酸 的特别好的方法详细描述于美国专利申请No.USSN 14/316,383; 14/316,398;14/316,416;14/316,431;14/316,447;以及14/316,463 中,该专利申请出于所有目的并且具体地关于针对加工核酸和基因组 材料的测序和其它表征的所有书面描述、图式和工作实施例而全文以 引用方式并入本文。也包含在分区中的延伸反应试剂,例如DNA聚 合酶、核苷三磷酸、辅因子(例如Mg^{2+}或Mn^{2+}等),然后使用样品作 为模板来

延伸引物序列,以产生引物退火至的模板链的互补片段,并 且互补片段包括寡核苷酸及其相关条形码序列。将多个引物退火和延 伸到样品的不同部分可以产生该样品的大量重叠互补片段,每个互补 片段具有其自身的条形码序列来指示产生其的分区。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核苷酸引发的模板,以 产生互补物的互补物,该互补物的互补物同样包含条形码序列。在其 它实例中,这种复制过程被配置为使得当第一互补物被复制时,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进一步重复的拷贝的基础的能 力。

[0058] 返回到图2中例示的方法,一旦分区特异性的条形码附接至拷贝 的片段,则将带条形码片段汇集(204)。然后对汇集的片段进行测序 (205),并且该片段的序列归属于其来源分子环境(206),使得所关注 的靶向区域既被鉴定,还被与所述来源分子环境相关联。本文描述的 方法和系统的一个优点是,在富集所靶向的基因组区域的片段之前将 分区-或样品-特异性的条形码附接到拷贝的片段保留了那些靶向区域 的原始分子环境,从而允许它们归属于它们的原始分区以及因此它们 的来源样品核酸分子。

[0059] 除了上述工作流程之外,可以使用包括基于芯片和基于溶液的捕 获方法两者的方法来进一步富集、分隔或分离所靶向的基因组区域, 即“沉淀(pulled down)”,以供进一步分析,特别是测序。此类方法利 用与所关注的基因组区域或所关注的基因组区域附近或邻近的区域 互补的探针。例如,在杂交(或基于芯片的)捕获中,含有捕获探针(通 常是单链寡核苷酸)的微阵列被固定在表面上,该微阵列具有一起覆 盖所关注区域的序列。基因组DNA被片段化并且可以进一步经受加 工(诸如末端修复),以产生平末端和/或添加附加特征如通用引发序 列。这些片段被与微阵列上的探针杂交。将未杂交的片段洗去,并将 所需的片段洗脱或以其它方式在表面上加工以供测序或其它分析,因 此保留在该表面上的片段群体富集含有所关注的靶向区域(例如,包 含与捕获探针中所包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来进一步扩增富集的片段群体。用于 这种靶沉淀富集方法的示例性方法描述于提交于2015年10月29日 的美国专利申请No.14/927,297中,该专利申请由此出于所有目的并 且具体是关于涉及与靶向沉淀富集方法和测序方法的所有教导,包括 所有书面描述、图式和实施例而以引用方式整体并入。

[0060] 如上所述,本文所描述的方法和系统提供了较长核酸的短序列读 段的单独分子环境。这种单独分子环境可以通过允许将较短序列读段 归属于来源单独核酸的任何方法或组合物来提供。如本文所用,单独 分子环境是指超过特定序列读段的序列环境,例如与不包括在序列读 段本身内的相邻或近侧序列的关系,并且因此将通常使得它们不被全 部或部分地包含在短序列读段中,例如约150个碱基、或约300个碱 基以供成对读段的读段。在特别优选的方面,所述方法和系统为短序 列读段提供长距离的序列环境。这种长距离环境包括给定的序列读段 与彼此之间相距长于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或者甚至长于100kb或更长的距离内的序列读 段的关系或关联。应当理解,通过提供长距离的单独分子环境,还可 以导出在该单独分子环境内的变体的定相信息,例如,特定长分子上 的变体将按照定义通常为定相的。

[0061] 通过提供较长距离的单独分子环境,本发明的方法和系统还提供 长得多的推断分子环境(在本文中也被称为“长虚拟单分子读段”)。如 本文所述的序列环境可以包括跨

全基因组序列的不同(通常在千碱基 量级)范围作图或提供各片段的关联。这些方法包括将短序列读段作 图到单独较长分子或连接分子的重叠群,以及对大部分较长单独分子(例如具有单独分子的连续确定序列)进行长距离测序,其中这种确定 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或者甚至长于100kb。与序列环境一样,短序列至较长核 酸(例如,单独长核酸分子或连接的核酸分子或重叠群的集合)的归属 可以包括以下两者:将短序列针对较长核酸段进行作图以提供高水平 序列环境,以及通过这些较长的核酸从短序列提供装配的序列。

[0062] 此外,虽然可以利用与长距离单独分子相关联的长距离序列环 境,但是具有这种长距离序列环境还允许推断甚至更长距离的序列环 境。作为一个实例,通过提供上述长距离分子环境,可以在来自不同 来源分子的长序列中鉴定重叠变体部分,例如定相变体、易位序列等, 从而允许推断这些分子之间的连接。这种推断的连接或分子环境在本 文中被称为“推断重叠群”。在一些情况下,当在定相序列的情形中论 述时,推断重叠群可以代表通常定相的序列,例如,在借助于重叠定 相变体的情况下,可以推断出实质上长度大于单独来源分子的定相重 叠群。这些定相重叠群在本文中被称为“相位区块”。

[0063] 通过用较长的单分子读段(例如,上文论述的“长虚拟单分子读 段”)开始,可以推导出比使用短读段测序技术或其它用于定相测序的 方法可获得的更长的推断重叠群或相位区块。参见例如所公开的美国 专利申请No.2013-0157870。具体地,通过使用本文描述的方法和系 统,可以获得N50为至少约10kb、至少约20kb、至少约50kb的推 断重叠群或相位区块长度(其中大于所述N50数值的多个区块长度的 总和是所有区块长度的总和的50%)。在更优选的方面,获得N50为 至少约100kb、至少约150kb、至少约200kb、以及在许多情况下至 少约250kb、至少约300kb、至少约350kb、至少约400kb、以及在一 些情况下至少约500kb或更高的推断重叠群或相位区块长度。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的最大相位区块长度。

[0064] 在一个方面,并结合本文中上文和下文所述的任何捕获方法,本 文所述的方法和系统提供用于分离样品核酸以供根据本文所述的任 何方法进一步加工。这种分离可以是允许核酸以与其所分离的其它核 酸相对隔离的方式经受进一步加工和反应的任何形式。分离可以关于 各自与所有其它核酸分离的单一核酸,或者将单一核酸分离成两个或 多个核酸的组,然后将该组与其它核酸组分离。在一些示例性实施方 案中,这种分离包括将样品核酸或其片段隔室化、沉积或分隔成离散 的隔室或分区(在本文中可互换地称为分区),其中每个分区保持将其 自身内含物与其它分区的内含物分离。独特标识符或其它归属手段 (在一些实例中,条形码)可以先前、随后或同时传递到分离的核酸, 以便允许稍后将特征(例如,核酸序列信息)归属至信息所来源于的样 品核酸。在其中将核酸分离成隔室或分区的某些示例性实施方案中, 标识符可以被包含在特定隔室中或被引入至特定隔室,并且具体地至 可以初始沉积到各分区内的连续样品核酸的相对较长段。

[0065] 在本文描述的方法中利用的样品核酸通常表示待分析的整个样 品的多个重叠部分,例如整个染色体、外显子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、扩增子或各 种不同所关注核酸中的任何一种。样品核酸通常经分隔为使得核酸以 连续核酸分子的相对较长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或者甚至长于100kb，这允许上述较长距离的分子环境。

[0066] 样品核酸还通常被以由此给定的分区具有极低的概率包含基因 组基因座的两个重叠片段的水平分隔。这通常通过在分隔过程期间以 低输入量和/或浓度提供样品核酸来实现。因此，在优选情况下，给 定的分区可以包括起始样品核酸的多个长但不重叠的片段。然后将不 同分区中的样品核酸与独特标识符相关联，其中对于任何给定的分 区，其中所包含的核酸具有相同的独特标识符，而不同分区可以包括 不同的独特标识符。此外，由于分隔步骤将样品组分分配成非常小体 积的分区或液滴，因此应当理解，为了实现如上所述的期望分配，不 需要如在较高体积过程中所需的例如在试管或多孔板的孔中对样品 进行大量稀释。此外，由于本文描述的系统采用这种高水平的条形码 多样性，所以可以如上所述在较高数目的基因组等同物之间分配不同 的条形码。具体地，先前描述的多孔板方法(参见例如美国公开申请 No.2013-0079231和No.2013-0157870)通常仅使用一百到几百种不同 的条形码序列操作，并且采用对其样品的有限稀释过程，以便能够将 条形码归属于不同的细胞/核酸。因此，它们通常使用远低于100个 细胞操作，这通常会提供1:10，并且一定高于1:100级别的基因组:(条 形码类型)比率。另一方面，本文描述的系统由于高水平的条形码多 样性，例如超过10,000、100,000、500,000等的不同条形码类型，而 可以1:50或更小、1:100或更小、1:1000或更小或甚至更小级别比率 的基因组:(条形码类型)比率操作，同时还允许加载更高数量的基因组 (例如，大于100个基因组/测定、大于500个基因组/测定、1000个基 因组/测定，或甚至更多的级别)，同时仍然提供每个基因组的远远改 善的条形码多样性。

[0067] 通常，在分隔步骤之前，将样品与可释放地附接至珠粒上的一组 寡核苷酸标签合并。用于对核酸进行条形码编码的方法是本领域中已 知并且在本文中描述的。在一些实例中，如在Amini等人，2014，Nature Genetics，提前在线发布(Advance Online Publication)中所述地利用方 法，该文献出于所有目的并且具体地是关于涉及将条形码或其它寡核 苷酸标记附接至核酸的所有教导以引用方式整体并入本文中。在其它 实例中，寡核苷酸可以包含至少第一和第二区域。第一区域可以是条 形码区域，其当在给定分区内的寡核苷酸之间时可以是基本上相同的 条形码序列，而当在不同分区之间时可以并且在大多数情况下是不同 的条形码序列。第二区域可以是可用于引发分区内的样品内的核酸的 N聚体(无规N聚体或设计成靶向特定序列的N聚体)。在某些情况下， 在N聚体被设计为靶向特定序列的情况下，该N聚体可被设计为靶 向特定染色体(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 特定基因或遗传区域，诸如与疾病或病症(例如癌症)相关的基因或区 域。在分区内，可以使用第二N聚体进行扩增反应，以沿着核酸的 长度在不同位置处引发核酸样品。作为该扩增的结果，每个分区可以 含有附接至相同或接近相同的条形码并且可以表示每个分区中的核 酸的更小重叠片段的核酸扩增产物。条形码可以用作标记来表示来源 于相同分区，以及因此潜在地还来源于相同的核酸链的一组核酸。在 扩增后，可以将核酸汇集、测序，以及使用测序算法进行比对。因为 较短的序列读段可以借助于其相关条形码序列来比对和归属于单一 较长的样品核酸片段，所以该序列上被鉴定的所有变体可以归属于单 一

起始片段和单一起始染色体。此外，通过在多个长片段间比对多个 共定位变体，可以进一步表征染色体贡献。因此，然后可以得出关于 特定遗传变体的定相的结论，诸如可以跨基因组序列的长距离进行分 析——例如跨基因组的不良表征区域的各段鉴定序列信息。此类信息 也可用于鉴定单倍型，单倍型通常是驻留在相同核酸链或不同核酸链 上的规定的遗传变异体集合。拷贝数变异也可以以这种方式鉴定。

[0068] 所描述的方法和系统提供了优于当前核酸测序技术及其相关样 品制备方法的显著优点。总体样品制备和测序方法被预先设置为主要 用于鉴定和表征样品中的主要成分，而非被设计用来鉴定和表征少量 成分(例如由一个染色体贡献的遗传物质，来自基因组的不良表征或 高度多态区域的遗传物质，或来自一个或多个细胞的物质，或在血流 中循环的片段化的肿瘤细胞DNA分子，它们构成了提取样品中总 DNA的较小百分比)。本文描述的方法包括增加来自这些少数成分的 遗传物质的选择性扩增方法，并且保留此遗传物质的分子环境的能力 进一步提供了这些成分的遗传表征。所描述的方法和系统还提供了检 测较大样品中存在的群体的显著优点。因此，它们尤其可用于评估单 倍型和拷贝数变异——本文公开的方法也可用于提供由于在样品制 备期间引入的偏差而在核酸靶群体中不良表征或不良呈现的基因组 区域上的序列信息。

[0069] 本文所公开的条形码编码技术的使用赋予了为给定的一组遗传 标记提供单独分子环境的独特能力，即将给定的一组遗传标记(与单 个标记相反)归属于单独的样品核酸分子，以及通过变体协调装配，来为多个样品核酸分子之间和/或特定染色体提供更宽广或甚至更长 距离的推断单独分子环境。这些遗传标记可以包括特异性遗传基因 座，例如变体(诸如SNP)，或者它们可以包括短序列。此外，条形码 的使用赋予了促进对从样品提取的总核酸群体中的少量成分和主要 成分之间进行区分的能力的附加优点，例如检测和表征血流中循环的 肿瘤DNA，并且还减少或消除了任选扩增步骤中的扩增偏差。此外，微流体形式的实施赋予了使用极小样品体积和低输入量的DNA的能 力，以及快速处理大量样品分区(液滴)以促进全基因组标记的能力。

[0070] 如前所述，本文描述的方法和系统的优点在于，它们可以通过使 用普遍存在可用的短读段测序技术来实现期望的结果。这些技术具有 易于获得和广泛分散在研究群落内，具有良好表征和高效的方案和试 剂体系的优点。这些短读段测序技术包括可从例如Illumina, inc.(GAIIx、NextSeq、MiSeq、HiSeq、X10)、Thermo-Fisher的Ion Torrent 分公司(Ion Proton和Ion PGM)购得的那些技术、焦磷酸测序方法，以 及其它技术。

[0071] 特别的优点在于本文所描述的方法和系统利用这些短读段测序 技术，并且由此具有其相关的低错误率和高通量。具体地，本文所描 述的方法和系统实现了如上所述期望的单独分子读段长度或环境，但 是在排除配对延伸的情况下所具有的单独测序读段短于1000bp、短 于500bp、短于300bp、短于200bp、短于150bp甚至更短；并且这 种单独分子读段长度的测序误差率小于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％。

[0072] II.工作流程综述

[0073] 本公开中描述的方法和系统提供用于将核酸分离成不同的组或 不同的区域，使得分离的核酸可以与一种或多种其它核酸相对隔离的 形式经受进一步加工和/或反应。在某些示例性情况下，这种分离可 以包括将单独样品(例如，核酸)沉积或分隔成离散分区，

其中每个分 区保持使其自身的内含物与其它分区中的内含物分离。如本文所用，分区是指可以包括各种不同形式的容器或器皿，例如孔、试管、微型 孔或纳米孔、通孔等。然而，在优选的方面，分区可在流体流内流动。这些器皿可以由例如具有包围内部流体中心或核心的外部阻挡层的 微胶囊或微囊泡构成，或者它们可以是能够夹带和/或保持其基质内 的材料的多孔基质。然而，在优选的方面，这些分区可以包括含水流 体在非水连续相(例如油相)内的液滴。各种不同的器皿描述于例如提 交于2013年8月13日的美国专利申请No.13/966,150中。类似地， 在非水或油连续相中形成稳定液滴的乳液体系详细描述于例如所公开的美国专利申请No.2010-0105112中。在某些情况下，微流通道网 络特别适用于产生如本文所述的分区。这种微流装置的实例包括在提 交于2015年4月9日的美国专利申请No.14/682,952中详细描述的那 些，该专利申请的全部公开内容出于所有目的整体并入本文。替代的 机制也可用于分隔单独的细胞，包括多孔膜，细胞的含水混合物穿过 多孔膜而被挤压成非水性流体。这种系统通常可购自例如Nanomi, Inc.。

[0074] 在利用乳液中的液滴的方法中，将样品材料(例如核酸)分隔成离 散分区通常可以通过将水性的含样品流流入到其中还流入分隔流体 的非水流(例如氟化油)的接合处，以便在分隔流体的流动流内形成含 水液滴，其中这些液滴包含样品材料。如下所述，分区，例如液滴， 通常还包括共分隔的条形码寡核苷酸。可以通过控制系统的各种不同 参数来调节任何特定分区内的样品材料的相对量，所述参数包括例如 含水流中的样品浓度、含水流的流速和/或非含水流等。本文中所描 述的分区通常具有非常小的体积。例如，在基于液滴的分区的情况下， 液滴的总体积可以小于1000pL、小于900pL、小于800pL、小于700pL、小于600pL、小于500pL、小于500pL、小于400pL、小于300pL、 小于200pL、小于100pL、小于50pL、小于20pL、小于10pL，或甚 至小于1pL。在使用珠粒共分隔的情况下，应当理解，分区内的样品 流体体积可以小于上述体积的90％、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，低反应体积分区的使用特别有利于使用非常 少量的起始试剂(例如输入核酸)进行反应。使用低输入核酸来分析样 品的方法和系统呈现于美国专利申请No.14/752,589和14/752,602中， 该专利申请的全部公开内容由此以引用方式整体并入。

[0075] 一旦样品被引入其相应的分区内，根据本文所述的方法和系统， 分区内的样品核酸通常经受选择性扩增，使得该基因组中用于进行靶 向覆盖以允许从头测序的所关注区域与该基因组的其它区域相比以 更高比例存在(尽管，如应理解的，基因组的那些其它区域也可以被 扩增，但是被扩增至较小的程度，这是因为它们不是被关注用于从头 覆盖的)。在某些实施方案中，所关注的基因组区域经扩增以提供对 基因组的那些所选区域的至少1X、2X、5X、10X、20X、30X、40X 或50X的覆盖。在其它实施方案中，分区内的所有核酸被扩增，但 是所选基因组区域是以靶向方式被扩增，以使得从那些所选的基因组 区域产生的扩增子比从该基因组的其它部分所产生的扩增子多至少 1-5倍、2-10倍、3-15倍、4-20倍、5-25倍、6-30倍、7-35倍、8-40 倍、9-45倍，或10-50倍。

[0076] 在选择性扩增基因组的所选区域的同时或之后，各分区内的核酸 (或其片段)具有独特的标识符，使得在表征那些核酸后，它们可以被 归属为来源于它们的相应的来源。因此，样品核酸通常与独特标识符 共分隔。在一些示例性实施方案中，这种独特标识符是条形码序列。为了清楚起见，本文的大部分论述针对包括条形码序列的标识符，但 是应当

理解，也可以根据本文所述的方法来使用可用于保留序列读段 的分子环境的任何独特标识符。在一些优选方面，独特标识符是以包 含可附接至核酸样品的核酸条形码序列的寡核苷酸形式提供的。将该 寡核苷酸分隔成使得当在给定分区中的寡核苷酸之间时，其中所包含 的核酸条形码序列是相同的，而当在不同分区之间时，寡核苷酸可以 并且优选具有不同的条形码序列。在优选方面，仅一个核酸条形码序 列将与给定的分区相关联，尽管在一些情况下可以存在两个或更多个 不同的条形码序列。

[0077] 核酸条形码序列将通常在寡核苷酸序列内包含6至约20个或更 多个核苷酸。这些核苷酸可以是完全连接的，即在相邻核苷酸的单个 段中，或者它们可以被分离成由一个或多个核苷酸分离的两个或更多 个分开的子序列。通常，分离的子序列的长度通常可以为约4至约16个核苷酸。

[0078] 共分隔的寡核苷酸通常还包含可用于加工分隔的核酸的其它功 能序列。这些序列包括例如用于扩增来自分区内的单独核酸的基因组 DNA的靶向或无规/通用扩增引物序列，同时附接相关条形码序列、测序引物、例如用于鉴定序列存在或用于沉淀带条形码的核酸的杂交 或探测序列，或许多其它潜在功能序列中的任何一种。此外，寡核苷 酸与相关条形码和其它功能序列的共分隔，以及样品材料描述于例如 美国专利申请No.14/316,383;14/316,398;14/316,416;14/316,431; 14/316,447;以及14/316,463中，该专利申请的全部公开内容由此以 引用方式整体并入。

[0079] 简言之，在一个示例性方法中，提供珠粒，每个珠粒可以包括可 释放地附接到珠粒上的大量上述寡核苷酸，其中附接至特定珠粒的所 有寡核苷酸可以包含相同的核酸条形码序列，但是其中大量不同的条 形码序列可跨所使用的珠粒群体呈现。通常，珠粒群体可以提供可包 括至少1000条不同条形码序列、至少10,000条不同条形码序列、至 少100,000条不同条形码序列或在一些情况下至少1,000,000条不同 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 有附接的大量寡核苷酸分子。具体地，包括单个珠粒上的条形码序列 的寡核苷酸分子数目可以是至少约10,000个寡核苷酸、至少100,000个寡核苷酸分子、至少1,000,000个寡核苷酸分子、至少100,000,000 个寡核苷酸分子，并且在一些情况下至少10亿个寡核苷酸分子。

[0080] 当对珠粒施加特定的刺激后，寡核苷酸可以从珠粒上释放出来。 在一些情况下，该刺激可以是光刺激，例如通过对可释放寡核苷酸的 光不稳定连接进行切割。在一些情况下，可以使用热刺激，其中珠粒 环境温度的升高可以导致连接的切割或寡核苷酸从珠粒的其它释放。 在一些情况下，可以使用化学刺激，其切割寡核苷酸与珠粒的连接， 或可以以其它方式导致寡核苷酸从珠粒释放。

[0081] 根据本文所述的方法和系统，可将包括附接的寡核苷酸的珠粒与 单独样品共分隔，使得在单独分区内含有单个珠粒和单个样品。在一 些情况下，在期望单个珠粒分区的情况下，可能期望控制流体的相对 流速，使得平均来说，分区含有每分区少于一个珠粒，以便确保被占 据的那些分区主要是被单独占据的。同样，可能希望控制流速以提供 较高百分比的分区来被占据，例如仅允许小百分比的未占用分区。在 优选方面，控制流动和通道架构以确保期望数目的单独占据的分区， 小于未占据分区的某个水平并且小于多次占据的分区的某个水平。

[0082] 图3示出了用于对样品核酸进行条形码编码以及随后测序，特别 是用于拷贝数变

异或单倍型测定的一种特定示例方法。首先,可以从 来源获得包含核酸的样品(300),并且还可以获得一组带条形码的珠粒 (310)。珠粒优选连接至含有一个或多个条形码序列的寡核苷酸,以及 引物,诸如无规N聚体或其它引物。优选地,条形码序列可以例如 通过以下方式从带条形码的珠粒释放:通过切割条形码和珠粒之间的 连接,或通过降解下方的珠粒以释放条形码,或两者的组合。例如, 在某些优选方面,带条形码的珠粒可以用试剂(诸如还原剂)降解或溶 解以释放条形码序列。在此实例中,将包含核酸305、带条形码的珠 粒315和任选的其它试剂(例如还原剂)320的少量样品合并以及进行 分隔。举例来说,这种分隔可以涉及将组分引入到液滴产生系统,例 如微流装置325中。借助于微流装置325,可以形成油包水型乳液330, 其中该乳液含有含样品核酸305、还原剂320和带条形码的珠粒315的含水液滴。还原剂可以溶解或降解带条形码的珠粒,从而从液滴 335中的珠粒释放具有条形码和无规N聚体的寡核苷酸。无规N聚 体可以随后引发样品核酸的不同区域,从而导致扩增后样品的扩增拷 贝,其中每个拷贝用条形码序列340标记。优选地,每个液滴包含一组含有相同条形码序列和不同无规N聚体序列的寡核苷酸。随后, 将乳液破坏(345),并且可以经由例如扩增方法350(例如PCR)来添加 附加序列(例如,有助于特定测序方法的序列、附加条形码等)。然后 可以执行测序(355),并应用算法来解读测序数据(360)。测序算法通 常能够例如执行条形码分析以比对测度读段和/或鉴定特定序列读段 所归属于的样品。此外,并且如本文中所述,这些算法还可以用于将 拷贝的序列归属至其来源分子环境。

[0083] 应当理解,在使用条形码序列340进行标记之前或同时,可以根 据本文所述的任何方法扩增样品以提供对基因组的所选区域的靶向 覆盖。这种靶向覆盖通常导致与来自基因组的其它区域的扩增子相 比,代表含有基因组的那些所选区域的分区中的核酸(或其部分)的序 列的扩增子群体较大。因此,相较于来自基因组的其它区域的分区, 在来自基因组的所选区域的分区内将存在更大数量的含条形码序列 340的扩增拷贝。

[0084] 在一些实施方案中并根据上述的任一项,相较于用于将条形码序 列附接至片段的方案,使用不同扩增方案利于含有基因组的所选区域 的部分的片段的扩增。在一个非限制性实例中,使用靶向PCR引物 的选择性扩增是在标准PCR扩增热循环条件下进行的,而用于附接 条形码的扩增是在温度急剧下降,之后是升高温度缓慢增加的情况下 进行的,以允许引发和延伸无规N聚体。

[0085] 如上所述,虽然单个占用可能是最期望的状态,但是应当理解, 通常可能存在多次占用的分区或未占用的分区。在图4中示意性地示 出了用于共分隔样品和包含条形码寡核苷酸的珠粒的微流通道结构 的实例。如图所示,通道区段402、404、406、408和410提供为在 通道接合部412处流体连通。包含各个样品414的含水流穿过通道区 段402流向通道接合部412。如本文别处所述,这些样品可以在分隔 过程之前悬浮在水性流体中。

[0086] 同时,包含条形码运载珠粒416的含水流穿过通道区段404流向 通道接合部412。将非水分隔流体从侧通道406和408中的每一个引 入通道接合部412内,并且合并的流流入出口通道410内。在通道接 合部412内,来自通道区段402和404的两个合并的含水流被合并, 并被分隔分成液滴418,液滴418包含被共分隔的样品414和珠粒 416。如前所述,通过控制在通道接合部412处合并的流体的每一者 的流动特征以及控制通道接合部的几何形状,可以优化组合和分隔以 实现珠粒、样品或两者在产生的分区418内的期望占用水平。

[0087] 应当理解,许多其它试剂可以与样品和珠粒一起共分隔,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、 核苷三磷酸、或NTP类似物)、引物序列和附加辅因子(诸如在这种反 应中使用的二价金属离子)、连接反应试剂(诸如连接酶酶和连接序 列)、染料、标签或其它标记试剂。引物序列可以包括无规引物序列 或针对扩增基因组的所选区域的靶向PCR引物,或它们的组合。

[0088] 一旦被共分隔,设置在珠粒上的寡核苷酸可用于对经分隔的样品 进行条形码编码和扩增。用于在对样品进行扩增和条形码编码的过程 中使用这些条形码寡核苷酸的特别好的方法详细描述于美国专利申 请No.14/316,383;14/316,398;14/316,416;14/316,431;14/316,447; 以及14/316,463中,该专利申请的全部公开内容由此以引用方式整体并入。简而言之,在一个方面,寡核苷酸存在于与样品共分隔的珠粒 上,并被从其珠粒释放到含有样品的分区中。该寡核苷酸通常在其 5'末端包括引物序列以及条形序列。引物序列可以是无规的或结构化 的。无规引物序列通常用于随机引发样品的多个不同区域。结构化的 引物序列可以包括一系列不同的结构,包括靶向用于引发样品的特定 靶向区域上游的经限定序列,以及具有某种部分限定结构的引物,包 括但不限于含有某一百分比的特定碱基的引物(诸如某一百分比的 GC N聚体),含有部分或完全简并序列的引物和/或含有根据本文任 何描述的部分无规和部分结构化的序列的引物。应当理解,上述类型 的无规和结构化引物中的任何一种或多种可以以任何组合包含在寡 核苷酸中。

[0089] 一旦被释放,寡核苷酸的引物部分可以退火到样品的互补区域。 延伸反应试剂,例如DNA聚合酶、核苷三磷酸、辅因子(例如Mg2+ 或Mn2+等)也与样品和珠粒一起共分隔,然后使用样品作为模板来延 伸引物序列,以产生模板链的互补片段,使引物退火至模板链,其中 互补片段包括寡核苷酸及其相关条形码序列。将多个引物退火和延伸 到样品的不同部分可以产生该样品的大量重叠互补片段,每个互补片 段具有其自身的条形码序列来指示产生其的分区。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核苷酸引发的模板,以产 生互补物的互补物,该互补物的互补物同样包含条形码序列。在一些 情况下,这种复制过程被配置为使得当第一互补物被复制时,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进一步重复的拷贝的基础的能力。此情况的一个实例的示意图示出于图5中。

[0090] 如图所示,包含条形码序列的寡核苷酸与样品核酸504一起共分 隔在例如乳液中的液滴502中。如本文其它地方所述,寡核苷酸508 可以提供在与样品核酸504共分隔的珠粒506上,所述寡核苷酸优选 可从珠粒506释放,如小图A所示。寡核苷酸508除了一个或多个 功能序列(例如序列510、514和516)之外还包含条形码序列512。例 如,寡核苷酸508被示出为包含条形码序列512,以及可用作给定测 序系统的附接或固定序列的序列510,例如用于在Illumina Hiseq或 Miseq系统的流动池中附接的P5序列。如图所示,寡核苷酸还包含 引物序列516,该引物序列516可以包含用于引发样品核酸504的各 部分的复制的无规或靶向N聚体。寡核苷酸508中还包含序列514, 序列514可以提供测序引发区域,诸如"读段1"或R1引发区域,该 区域用于引发通过测序系统中的合成反应进行的聚合酶介导的模板定向测序。在许多情况下,条形码序列512、固定化序列510和R1 序列514可以是附接于给定珠粒上的所有寡核苷酸共有的。对于无规 N聚体引物,引物序列516可以改变,或者对于某些靶向应用,引物 序列516可以是在给定珠粒上的寡核苷酸共有的。

[0091] 基于引物序列516的存在,寡核苷酸能够如小图B所示引发样品 核酸,这允许使用

聚合酶酶和同样与珠粒506和样品核酸504共分隔 的其它延伸试剂来延伸寡核苷酸508和508a。如小图C中所示，在 寡核苷酸的延伸之后，无规N聚体引物将退火到样品核酸504的多个不同区域；产生核酸的多个重叠互补物或片段，例如片段518和 520。虽然包含与样品核酸的各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样品核酸504的片段，具有附接的条形码序列。应当理解，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”。尽管如上所述，然而，术 语“片段”涵盖来源核酸序列的一部分的任何表示，例如模板或样品核 酸，包括通过提供模板序列的各部分的其它机制产生的那些，所述机 制为诸如通过例如酶促、化学或机械片段化对给定序列分子进行的实 际片段化。然而，在优选方面，模板或样品核酸序列的片段将表示基 本序列(underlying sequence)的复制部分或其互补物。

[0092] 条形码核酸片段可以随后经受表征，例如通过序列分析，或者它 们可以在该过程中进一步扩增，如小图D所示。例如，附加寡核苷 酸，如同样从珠粒506释放的寡核苷酸508b，可以引发片段518和 520。具体地，此外基于寡核苷酸508b中的无规N聚体引物516b(其 在许多情况下将与给定分区(例如引物序列516)中的其它无规N聚体 不同)的存在，寡核苷酸退火至片段518，并且经延伸而产生片段518 的至少一部分的互补物526，该互补物526包含序列528，序列528 包含样品核酸序列的一部分的复制物。寡核苷酸508b继续延伸，直 到其已经复制经过片段518的寡核苷酸部分508。如本文其它地方所 述，并且如小图D所示，寡核苷酸可以被配置为促使使用聚合酶的 复制在所需点处停止，例如在复制经过包含在片段518内的寡核苷酸 508的序列516和514之后。如本文所述，这可以通过不同的方法来实现，包括例如并入不能被所用聚合酶酶处理的不同核苷酸和/或核 苷酸类似物。例如，这可以包括在序列区域512内包含含有尿嘧啶的 核苷酸，以阻碍非尿嘧啶耐受型聚合酶，从而停止对该区域的复制。 因此，产生了在一个端部处包括全长寡核苷酸508b的片段526，其包括条形码序列512、附接序列510、R1引物区514和无规N聚体序 列516b。在该序列的另一端将包括互补物516’至第一寡核苷酸508 的无规N聚体，以及R1序列的全部或部分的互补物(示出为序列 514’)。然后，R1序列514及其互补物514’能够一起杂交而形成部分 发夹结构528。应当理解，因为不同寡核苷酸中的无规N聚体不同， 所以预计这些序列及其互补物不会参与发夹形成，例如，序列516’ 是无规N聚体516的互补物，预计其不会与无规N聚体序列516b互 补。对于其它应用如靶向引物不会是这种情况，其中在给定分区内的 寡核苷酸中的N聚体将是共有的。通过形成这些部分发夹结构，其 允许从进一步复制物中移除样品序列的第一级复制物，例如防止对拷 贝的重复复制。部分发夹结构还为所产生的片段(例如片段526)的后 续加工提供了有用的结构。

[0093] 然后，可以汇集来自多个不同分区的所有片段，以用于如本文所 述在高通量测序仪上进行测序。因为每个片段是按照其来源分区编码 的，所以该片段的序列可以基于条形码的存在而被归属回其来源。这 在图6A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和源自不同染色体602或核 酸链的核酸606如上所述各自与其自身的一组条形码寡核苷酸一起 分隔。

[0094] 在每个分区内，然后加工各个核酸604和606以单独提供第一片 段的第二片段重叠集合，例如第二片段集合608和610。该加工还提 供给第二片段对于来源于特定第一片段的各个第二片段为相同的条 形码序列。如图所示，第二片段集合608的条形码序列由“1”表

示，而片段集合610的条形码序列由“2”表示。可以使用多样性条形码文 库来对大量不同的片段集合进行区别地条形码编码。然而，没有必要 使用不同条形码序列来对来自不同第一片段的每个第二片段集合进 行条形码编码。事实上，在许多情况下，可以将多个不同的第一片段 同时加工成包括相同的条形码序列。不同的条形码文库在本文其它地 方详细描述。

[0095] 例如来自片段集合608和610的带条形码的片段，可以随后被汇 集以供测序使用，例如通过可从Illumina或Thermo Fisher,Inc.的Ion Torrent分公司购得的合成技术进行测序。一旦被测序，来自汇集的片 段612的序列读段可以至少部分基于所包含的条形码，并且可选地且 优选地部分基于片段本身的序列，而被归属至它们相应的片段集合， 例如，如聚合读段614和616所示。然后每个片段集合的归属序列读 段经装配以提供每个样品片段的经装配序列，例如序列618和620， 该经装配序列继而可以被进一步归属回它们相应的来源染色体或来 源核酸分子(600和602)。用于装配基因组序列的方法和系统描述于例 如提交于2015年6月26日的美国专利申请No.14/752,773中，该专 利申请的全部公开内容由此以引用方式整体并入。

[0096] 在一些实施方案中并且如图6B所示，使用含有片段集合608或 610的分区包含的是引物集合613。在其它实施方案中，引物集合613 针对基因组的所选区域，使得在提供条形码序列(608的条形码为“1” 和610的条形码为“2”)之前、同时或之后，片段集合608和610被扩 增为使得基因组的所选区域被以另一程度覆盖到该基因组的其它区 域上。在图6B所示的示例性实施方案中，片段集合608含有来自引 物集合613所针对的基因组的所选区域的序列，而片段集合610不含 有来自该基因组的那些所选区域的序列。因此，相较于来自集合610 的片段的覆盖，将存在增加的来自集合608的片段的覆盖(例如，更 多拷贝)。因此，所汇集的片段612含有带条形码的片段，该带条形 码的片段含有已经用靶向方式扩增的片段，从而允许相较于来自片段 集合610(带条形码“2”的片段)，更大比例的序列读段来自片段集合 608(带条形码“1”的片段)。此外，由于条形码，来自集合608的较大 比例的序列读段，可以如所汇集集合612中的其余片段一样，被归属 回它们相应的原始来源核酸分子600和602(如图6A所示)。

[0097] III.应用于核酸测序的方法和系统

[0098] 本文所描述的方法、组合物和系统特别适用于核酸测序技术。这 种测序技术可以包括本领域中已知的任何技术，包括短读段和长读段 测序技术。在某些方面，本文所描述的方法、组合物和系统用于短读 段、高精度的测序技术。

[0099] 本文所描述的方法、组合物和系统允许对基因组的不良表征、高 度多态和/或与参考序列偏离的区域进行遗传表征。具体地，本文所 描述的方法、组合物和系统提供了对基因组的所选部分的增加和冗余 的覆盖，使得可以从基因组的那些所选部分获得附加的冗余序列信 息。在特定情况中，该附加序列信息(例如对基因组的靶向区域的增 加覆盖)提供了足够的信息来允许对基因组的那些所选部分进行从头 测序。这种从头测序对于基因组的不良表征、高度多态和/或与参考 序列偏离的区域特别有用。应当理解，人基因组的显著百分比(根据 例如Altemose等人，PLOS Computational Biology，2014年5月15 日，第10卷，第5期为至少5％至10％)保持未装配、未作图并且不 良表征。参考组装序列通常将这些缺少区域注释为多兆碱基异染色质 空位，这些缺少区域主要存在于着丝粒附近并且在

近端着丝染色体的 短臂上。基因组的此缺少部分包括保持对使用通常所用的测序技术进行精确表征的抗性的结构特征。对精确表征具有抗性的附加示例性区 域包括但不限于具有近似同源假基因的区域(例如SMN1/2Cyp2d6)、在整个基因组中具有基本上重复的序列的区域,包括但不限于转座子 (例如SINE、LINE),以及与用作不良引导的参考序列具有巨大差异 的区域(诸如编码人白细胞抗原(HLA)复合物的基因的区域)。本文所 述的方法、组合物和系统将所关注区域的选择性扩增与维持分子环境 的能力结合在一起,从而允许对通常不良表征的基因组区域进行从头 测序。

[0100] 在特定情况下,本文描述的方法包括在测序之前选择性扩增基因 组的所选区域的步骤。通常使用本领域中已知的方法(包括但不限于 PCR扩增)进行的此扩增提供了对基因组的所选区域的至少1X、2X、 3X、4X、5X、6X、7X、8X、9X、10X、11X、12X、13X、14X、15X、 16X、17X、18X、19X或20X的覆盖,从而提供一定量的核酸以允 许对那些所选区域进行从头测序。在其它实施方案中,该扩增提供对 基因组的所选区域的至少1X至30X、2X至25X、3X至20X、4X至 15X、或5X至10X的覆盖。

[0101] 该扩增通常是通过与基因组的所选区域内或附近的序列互补的 引物的延伸进行的。在一些情况下,使用被设计成跨所关注区域拼接 (tile)的引物文库——换句话说,该引物文库被设计成扩增沿着基因组 的所选区域在特定距离处的区域。在一些情况下,选择性扩增利用与 沿着基因组的所选区域的每10个、15个、20个、25个、50个、100 个、200个、250个、500个、750个、1000个或10000个碱基互补 的引物。在其它实例中,该拼接的引物文库被设计成捕获各距离的混 合物——此混合物可以是各距离的随机混合或被智能地设计成使得 所选区域的特定部分或百分比被使用不同的引物对扩增。在其它实施 方案中,引物对被设计为使得每对扩增基因组的所选部分的任何连续 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%。

[0102] 在某些实施方案中并且根据上述的任一项,扩增是跨至少3兆碱 基对长(Mb)的基因组区域进行的。在其它实施方案中,根据本文所述 的任何方法选择性扩增的基因组所选区域为至少3.5Mb、4Mb、 4.5Mb、5Mb、5.5Mb、6Mb、6.5Mb、7Mb、7.5Mb、8Mb、8.5Mb、 9Mb、9.5Mb或10Mb。在其它实施方案中,基因组的所选区域的长 度为约2-20Mb、3-18Mb、4-16Mb、5-14Mb、6-12Mb、或7-10Mb。 如上所述,可以使用与这些区域的末端处或末端附近的序列互补的单 个引物对来跨这些区域进行扩增。在其它实施方案中,扩增是使用跨 区域长度拼接的引物对文库进行的,使得沿着该区域的规则区段、无 规区段或不同区段距离的某种组合被扩增,具有根据上述描述的覆盖 程度。

[0103] 在一些实施方案中,用于选择性扩增基因组的所选区域的引物含 有尿嘧啶,以使得引物本身不被扩增。

[0104] 通常,本文所描述的方法和系统通过提供对基因组的所选区域的 序列的确定来实现靶向基因组测序,并且此测序信息通常是使用具有 短读段测序技术的极低测序错误率和高通量优点的方法获得的。如前 所述,本文描述的方法和系统的优点在于,它们可以通过使用普遍存 在可用的短读段测序技术来实现期望的结果。这些技术具有易于获得 和广泛分散在研究群落内,具有良好表征和高效的方案和试剂体系的 优点。这些短读段测序技术包括可从例如Illumina,inc.(GAIIx、NextSeq、MiSeq、HiSeq、X10)、Thermo-Fisher的Ion Torrent分公司 (Ion Proton和Ion PGM)购得的那些技术、焦磷酸测序方法,以及其它

技术。

[0105] 特别的优点在于本文所描述的方法和系统利用这些短读段测序 技术,并且由此具有其相关的低错误率。具体地,本文所描述的方法 和系统实现了如上所述期望的单独分子读段长度或环境,但是在排除 配对延伸的情况下所具有的单独测序读段短于1000bp、短于500bp、 短于300bp、短于200bp、短于150bp甚至更短;并且这种单独分子 读段长度的测序误差率小于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%。

[0106] 根据本申请中描述的方法和系统来对核酸加工和测序的方法还 进一步详细描述于USSN 14/316,383;14/316,398;14/316,416; 14/316,431;14/316,447;以及14/316,463中,该专利申请出于所有目 的并且具体地关于针对加工核酸和基因组材料的测序和其它表征的 所有书面描述、图式和工作实施例以引用方式整体并入本文。

[0107] 不管使用的测序平台如何,一般地并且根据本文所述的任何方 法,核酸测序通常是以保留序列读段的分子环境或序列读段的部分的 分子环境的方式进行的。这意味着多个序列读段或序列读段的多个部 分可归属于单一的来源核酸分子。“可归属于”意指序列读段能够被鉴 定为落入其特定来源核酸分子的线性碱基序列内——换句话说并且 参照图7,如果片段703、704、705和706是从来源核酸分子701和702产生的,那么测序以使得来自片段703、704、705和706的序列 读段保留其分子环境并且容易确定片段703和704来源于来源分子 701,而片段705和706来源于来源分子702的方式进行,即使将所 有片段汇集在一起进行测序反应也是这样。此外,测序通常被进行为 使得不仅来源分子被确定,而且每个片段沿着该线性分子的相对位置 也被确定——例如,可以确定片段703沿着来源核酸701的线性序列 在片段704的“上游”。通常,通过使用任何标识符或将一个或多个片 段与其它片段区分开的任何其它方法来保留分子环境。通常,这种标 识符用于已经被分成多组或分成单独实体的片段。在一些实例中,这 种分离是分离成离散分区,尽管应当理解可以使用任何其它分子分离 方法。在其它实例中,所用的标识符是条形码,并且线性位置通过条 形码编码以及来自重叠片段的序列读段的算法装配两者来确定。尽管 为了清楚起见,本文的大部分论述关于分离成分区和/或条形码编码, 但是应当理解,任何分离来源核酸分子的方法和任何鉴定或以其它方 式归属片段的方法都可用于本文所描述的方法和系统。

[0108] 应当理解,虽然该单一来源核酸分子可以是各种长度中的任何一 者,但是在优选方面,其应是相对较长的分子,以允许保留长距离分 子环境。具体地,该单一来源分子优选基本上长于典型的短读段序列 长度,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 为1兆碱基或更长。

[0109] 在某些情况下,本文描述的测序方法包括对所选区域的深覆盖与 跨基因组的较长范围的较低程度连接读段的组合。应当理解,从头和 重新测序的这种组合提供了对整个基因组和/或基因组的大部分进行 测序的有效方式。通过本文所述的选择性扩增方法对不良表征的和/ 或高度多态性区域进行靶向覆盖提供了从头序列装配所必需的核酸 材料量,而在该基因组的其它区域上的连锁基因组测序维持对基因组 的其余部分进行高通量

测序。本文所述的方法和组合物独特地适于允 许从头和连锁读段测序的这种组合，这是因为相同的测序平台可用于 这两种类型的覆盖。根据本文所述的方法测序的核酸和/或核酸片段 的群体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——覆盖用于从头测序的所关注区域的核酸的比 例高于覆盖基因组其它区域的核酸的比例，这是由于本文进一步详细 描述的靶向扩增方法。

[0110] 通常，如图1所示，本文所述的方法和系统可以用于表征核酸，尤其是来自基因组的所选区域的核酸，同时保留分子环境。如图所示，示出了两个离散的单独核酸102和104，其各自具有多个所关注区域，例如核酸102中的区域106和108，以及核酸104中的区域110和112。每个核酸中的所关注区域在相同核酸分子内连接(例如，来源于相同 核酸分子)，但是在一些情况下，这些区域可以彼此相对分离，例如 间隔超过1kb、间隔超过5kb、间隔超过10kb、间隔超过20kb、间隔 超过30kb、间隔超过40kb、间隔超过50kb，并且在一些情况下间隔 多达100kb。所关注区域通常是基因组的离散和分开的部分——在一 些情况下，此类区域是不良表征的区域。所关注区域也可以表示单独 基因、基因群、外显子。如图所示，将各个核酸102和104分别分离 到其自身的分区114和116。如本文中其它地方所述，这些分区在许 多情况下是油包水乳液中的含水液滴。在每个液滴内，每个片段的各 部分以保留那些片段的原始分子环境的方式被拷贝，例如为来源于相 同的分子。如图所示，这是通过在每个拷贝的片段中包含条形码序列 实现的，例如如图所示的条形码序列“1”或“2”，该条形码序列代表来 源片段被分隔成的液滴。对于全基因组序列分析应用，可以简单地汇 集所有拷贝的片段及其相关条形码，以便对来自各个来源核酸102和 104的全长序列信息进行测序和重新装配。然而，在许多情况下，更 期望仅分析整个基因组的特定靶向部分，以便提供对基因组的科学相 关部分的更大重视，并将执行对基因组的不太相关或不相关部分的测 序的时间和代价最小化。

[0111] 根据上述，除了条形码编码步骤之外，可以存在一个或多个选择 性扩增步骤，使得如果核酸102或104含有所关注的所选基因组区域，则来自那些区域的扩增子将形成各个分区114和116中的片段的较大 百分比。这种扩增步骤将通常在根据本文所述的方法附接条形码之前 或同时进行，尽管在一些实施方案中，该扩增步骤也可以在条形码附 接之后发生。

[0112] 因为文库118中的汇集片段保留了它们原始分子环境，例如，通 过保留条形码信息，它们可以被重新装配成具有嵌入的(有时是长距 离的)连接信息的原始分子环境，例如，具有在所关注的各装配区域 106:108和110:112之间的推断连接。举例来说，可以鉴定基因组的 两个不同靶向部分之间的直接分子连接，例如两个或更多个外显子，并且此直接分子连接可用于鉴定结构变化和其它基因组特征。对于其 中利用选择性扩增来增加含有基因组的所选区域的部分的核酸片段 的量的情况，则鉴定分子环境的能力还提供了一种对基因组的那些所 选区域(通常在允许对那些区域进行从头装配的深度处)进行测序的方 式。

[0113] 通常，本发明的方法包括如图2所示的步骤，其提供了在本文中 进一步详细论述的本发明的方法的示意性综述。应当理解，图2中概 述的方法是可以根据需要并如本文所述改变或修改的示例性实施方 案。

[0114] 如图2所示，本文所述的方法在大多数实例中将包括其中将含有 所关注的靶向区

域的样品核酸分隔的步骤(201)。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靶向富集,以产生其中大部 分将含有来自所选基因组区域的序列的片段群体(202)。然后,通常通 过将对其被包含于其中的分区具有特异性的片段进行条形码编码,来以保留片段的原始分子环境的方式将这些片段进一步片段化或拷贝(203)。在一些实例中,每个分区可以包含多于一个核酸,并且在一些 情况下将含有数百个核酸分子——在多个核酸位于分区的情况下,基 因组的任何特定基因座将通常在条形码编码之前用单个单独核酸表 示。可以使用本领域中已知的任何方法来产生步骤203的条形码片段 ——在一些实例中,寡核苷酸是不同分区内的样品。这种寡核苷酸可 以包括旨在随机引发样品的多个不同区域的无规序列,或者它们可以 包含靶向用于引发样品靶向区域上游的特异性引物序列。在其它实例 中,这些寡核苷酸还含有条形码序列,使得复制过程也对原始样品核 酸的所得复制片段进行条形码编码。用于在对样品进行扩增和条形码 编码的过程中使用这些条形码寡核苷酸的特别好的方法详细描述于 美国专利申请No.14/316,383;No.14/316,398;No.14/316,416;No. 14/316,431;No.14/316,447;以及No.14/316,463中,该专利申请出 于所有目的并且具体地关于针对加工核酸和基因组材料的测序和其 它表征的所有书面描述、图式和工作实施例而以引用方式整体并入本 文。也包含在分区中的延伸反应试剂,例如DNA聚合酶、核苷三磷 酸、辅因子(例如Mg^{2+}或Mn^{2+}等),然后使用样品作为模板来延伸引 物序列,以产生引物退火至的模板链的互补片段,并且互补片段包括 寡核苷酸及其相关条形码序列。将多个引物退火和延伸到样品的不同 部分可以产生该样品的大量重叠互补片段,每个互补片段具有其自身 的条形码序列来指示产生其的分区。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核苷酸引发的模板,以产生互补物的互 补物,该互补物的互补物同样包含条形码序列。在其它实例中,这种 复制过程被配置为使得当第一互补物被复制时,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进一步重复的拷贝的基础的能力。

[0115] 返回到图2中例示的方法,一旦分区特异性的条形码附接至拷贝 的片段,则将带条形码片段汇集(204)。然后对汇集的片段进行测序 (205),并且该片段的序列归属于其来源分子环境(206),使得所关注 的靶向区域既被鉴定,还被与所述来源分子环境相关联。本文描述的 方法和系统的一个优点是,在富集所靶向的基因组区域的片段之前将 分区-或样品-特异性的条形码附接到拷贝的片段保留了那些靶向区域 的原始分子环境,从而允许它们归属于它们的原始分区以及因此它们 的来源样品核酸。

[0116] 除了上述工作流程之外,可以使用包括基于芯片和基于溶液的捕 获方法两者的方法来进一步富集、分隔或分离所靶向的基因组区域, 即“沉淀(pulled down)”,以供进一步分析,特别是测序。此类方法利 用与所关注的基因组区域或所关注的基因组区域附近或邻近的区域 互补的探针。例如,在杂交(或基于芯片的)捕获中,含有捕获探针(通 常是单链寡核苷酸)的微阵列被固定在表面上,该微阵列具有一起覆 盖所关注区域的序列。基因组DNA被片段化并且可以进一步经受加 工(诸如末端修复),以产生平末端和/或添加附加特征如通用引发序 列。这些片段被与微阵列上的探针杂交。将未杂交的片段洗去,并将 所需的片段洗脱或以其它方式在表面上加工以供测序或其它分析,因 此保留在该表面上的片段群体富集含有所关注的靶向区域(例如,包 含与捕获探针中所包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来进一步扩增富集的片段群

体。用于 这种靶沉淀富集方法的示例性方法描述于提交于2015年10月29日 的USSN 14/927,297中,该专利申请由此出于所有目的并且具体是关 于涉及与靶向沉淀富集方法和测序方法的所有教导,包括所有书面描 述、图式和实施例而全文以引用方式并入本文。

[0117] 在一些方面,用于覆盖基因组的所选区域的方法包括其中将含有 来自那些所选区域的核酸分子和/或其片段的离散分区本身分选以供 进一步加工的方法。应当理解,离散分区的这种分选可以与本文所述 对所关注基因组区域的其它选择性扩增和/或靶向沉淀方法的任何组 合,特别是与上述工作流程的步骤的任何组合。

[0118] 一般而言,这种分离分区的分选方法包括其中将包含基因组的一 个或多个所选部分的至少一部分的分区与不含来自该基因组的那些 部分的任何序列的分区分离的步骤。这些方法包括以下步骤:提供在 包含来自基因组的一个或多个所选部分的至少一部分的序列的离散 分区内富集包含该基因组的那些部分的片段的序列的群体。这种富集 通常是通过对离散分区内包含基因组的一个或多个所选部分的至少 一部分的片段使用定向PCR扩增以产生群体来实现的。因此,这种 定向PCR扩增产生了包含基因组的一个或多个所选部分的至少一部 分的扩增子。在某些实施方案中,这些扩增子附接至可检测标签,在 一些非限制性实施方案中可检测标签可以包括荧光分子。通常,这种 附接发生而使得仅从包含基因组的一个或多个所选部分的片段产生 的那些扩增子附接至可检测标签。在一些实施方案中,可检测标签的 附接在基因组的一个或多个所选部分的选择性扩增期间发生。在其它 实施方案中,这种可检测标签可以包括但不限于荧光标签、电化学标 签、磁珠以及纳米颗粒。可检测标签的此附接可以使用本领域中已知 的方法来实现。在其它实施方案中,含有包含基因组的一个或多个所 选部分的至少一部分的片段的离散分区基于从附接到那些分区内的 扩增子的可检测标签发射的信号而被分选。

[0119] 在其它实施方案中,将含有基因组的所选部分的离散分区从不含 有这种序列的那些分区分选的步骤包括以下步骤:(a)提供起始基因组 材料;(b)将单独核酸分子从起始基因组材料分布到离散分区中,使得 每个离散分区含有第一单独核酸分子;(c)提供至少一些离散分区内富 集包含基因组的一个或多个所选部分的至少一部分的片段的序列的的群体;(d)将共同条形码序列附接至每个离散分区内的片段,以使得 每个片段可归属于其所被包含的离散分区;(e)分离含有包含基因组的 一个或多个所选部分的至少一部分的片段的离散分区与不含有包含 基因组的该一个或多个所选部分的片段的离散分区;(f)从包含基因组 的该一个或多个所选部分的至少一部分的片段获得序列信息,从而在 保留分子环境的同时对基因组样品的一个或多个靶向部分进行测序。

[0120] 在其它实施方案中并根据上述的任一项,在从片段获得序列信息 之前,将各离散分区合并并将各片段汇集在一起。在其它实施方案中, 从片段获得序列信息的步骤是以维持片段序列的分子环境的方式进 行的,使得鉴定进一步包括鉴定来源于相同的第一单独核酸分子的片 段。在其它实施方案中,此获得序列信息的步骤包括选自由以下项组 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 其它实施方案中,测序反应是短读段、高精度测序反应。

[0121] 在其它实施方案中并根据上述的任一项,离散分区包含乳液中的 液滴。在其它实施方案中,离散分区内的带条形码的片段代表对基因 组的一个或多个所选部分的约1X-10X的覆盖。在其它实施方案中, 离散分区内的带条形码的片段代表对基因组的一个或多

个所选部分 的约2X-5X的覆盖。在其它实施方案中，离散分区内的带条形码的 片段代表对基因组的一个或多个所选部分的至少1X的覆盖。在其它 实施方案中，离散分区内的带条形码的片段代表对基因组的一个或多 个所选部分的至少2X或5X的覆盖。

[0122] 除了提供从基因组的所选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，包括但不限于单 倍型定相、结构变异的鉴定以及鉴定拷贝数变异，如在美国专利申请 No.14/752,589和14/752,602中所描述的，该专利申请出于所有目的 并且具体地是关于针对基因组材料表征的所有书面描述、图式和工作 实施例而全文以引用方式并入本文。

[0123] 如上所述，本文所描述的方法和系统提供了较长核酸的短序列读 段的单独分子环境。如本文所用，单独分子环境是指超过特定序列读 段的序列环境，例如与不包括在序列读段本身内的相邻或近侧序列的 关系，并且因此将通常使得它们不被全部或部分地包含在短序列读段 中，例如约150个碱基、或约300个碱基以供成对读段的读段。在特 别优选的方面，所述方法和系统为短序列读段提供长距离的序列环 境。这种长距离环境包括给定的序列读段与彼此之间相距长于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或者甚至 长于100kb或更长的距离内的序列读段的关系或关联。应当理解，通 过提供长距离的单独分子环境，还可以导出在该单独分子环境内的变 体的定相信息，例如，特定长分子上的变体将按照定义通常为定相的。

[0124] 通过提供较长距离的单独分子环境，本发明的方法和系统还提供 较长的推断分子环境(在本文中也被称为“长虚拟单分子读段”)。如本 文所述的序列环境可以包括在全基因组序列的不同(通常在千碱基量 级)范围上映射或提供各片段的关联。这些方法包括将短序列读段作 图到单独较长分子或连接分子的重叠群，以及对大部分较长单独分子(例如具有单独分子的连续确定序列)进行长距离测序，其中这种确定 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或者甚至长于100kb。与序列环境一样，短序列至较长核 酸(例如，单独长核酸分子或连接的核酸分子或重叠群的集合)的归属 可以包括以下两者：将短序列针对较长核酸段进行作图以提供高水平 序列环境，以及通过这些较长的核酸从短序列提供装配的序列。

[0125] 此外，虽然可以利用与长距离单独分子相关联的长距离序列环 境，但是具有这种长距离序列环境还允许推断甚至更长距离的序列环 境。作为一个实例，通过提供上述长距离分子环境，可以在来自不同 来源分子的长序列中鉴定重叠变体部分，例如定相变体、易位序列等， 从而允许推断这些分子之间的连接。这种推断的连接或分子环境在本 文中被称为“推断重叠群”。在一些情况下，当在定相序列的情形中论 述时，推断重叠群可以代表通常定相的序列，例如，在借助于重叠定 相变体的情况下，可以推断出实质上长度大于单独来源分子的定相重 叠群。这些定相重叠群在本文中被称为“相位区块”。

[0126] 通过用较长的单分子读段(例如，上文论述的“长虚拟单分子读 段”)开始，可以推导出比使用短读段测序技术或其它用于定相测序的 方法可获得的更长的推断重叠群或相位区块。参见例如所公开的美国 专利申请No.2013-0157870。具体地，通过使用本文描述的方法和系 统，可以获得N50为至少约10kb、至少约20kb、至少约50kb的推 断重叠群或相位

区块长度(其中大于所述N50数值的多个区块长度的 总和是所有区块长度的总和的50%)。在更优选的方面,获得N50为 至少约100kb、至少约150kb、至少约200kb、以及在许多情况下至 少约250kb、至少约300kb、至少约350kb、至少约400kb、以及在一 些情况下至少约500kb或更高的推断重叠群或相位区块长度。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的最大相位区块长度。

[0127] 在一个方面,并结合本文中上文和下文所述的任何捕获方法,本 文所述的方法和系统提供用于将样品核酸或其片段隔室化、沉积或分 隔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 保持将其自身内含物与其它分区的内含物分离。独特标识符(例如条 形码)可以先前、随后或同时传递到保持隔室化或分隔的样品核酸的 分区,以便允许稍后将特征(例如,核酸序列信息)归属至特定隔室内 所包含的样品核酸,特别是可以最初沉积到分区内的连续样品核酸的 相对较长段。

[0128] 在本文描述的方法中利用的样品核酸通常表示待分析的整个样 品的多个重叠部分,例如整个染色体、外显子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、扩增子或各 种不同所关注核酸中的任何一种。样品核酸通常经分隔为使得核酸以 连续核酸分子的相对较长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或者甚至长于100kb,这允许上述较长距离的分子环境。

[0129] 样品核酸还通常被以由此给定的分区具有极低的概率包含该起 始样品核酸的两个重叠片段的水平分隔。这通常通过在分隔过程期间 以低输入量和/或浓度提供样品核酸来实现。因此,在优选情况下, 给定的分区可以包括起始样品核酸的多个长但不重叠的片段。然后将 不同分区中的样品核酸与独特标识符相关联,其中对于任何给定的分 区,其中所包含的核酸具有相同的独特标识符,而不同分区可以包括 不同的独特标识符。此外,由于分隔步骤将样品组分分配成非常小体 积的分区或液滴,因此应当理解,为了实现如上所述的期望分配,不 需要如在较高体积过程中所需的例如在试管或多孔板的孔中对样品 进行大量稀释。此外,由于本文描述的系统采用这种高水平的条形码 多样性,所以可以如上所述在较高数目的基因组等同物之间分配不同 的条形码。具体地,先前描述的多孔板方法(参见例如美国公开申请 No.2013-0079231和No.2013-0157870)通常仅使用一百到几百种不同 的条形码序列操作,并且采用对其样品的有限稀释过程,以便能够将 条形码归属于不同的细胞/核酸。因此,它们通常使用远低于100个 细胞操作,这通常会提供1:10,并且一定高于1:100级别的基因组:(条 形码类型)比率。另一方面,本文描述的系统由于高水平的条形码多 样性,例如超过10,000、100,000、500,000、600,000、700,000等的 不同条形码类型,而可以1:50或更小、1:100或更小、1:1000或更小 或甚至更小比率级别的基因组:(条形码类型)比率操作,同时还允许加 载更高数量的基因组(例如,大于100个基因组/测定、大于500个基 因组/测定、1000个基因组/测定,或甚至更多的级别),同时仍然提供 每个基因组的远远改善的条形码多样性。

[0130] 通常,在分隔步骤之前,将样品与可释放地附接至珠粒上的一组 寡核苷酸标签合并。在一些实例中,使用扩增方法来向所得扩增产物 中添加条形码,该条形码在一些实例中含有其所来源于的完整来源核 酸分子的较小区段(片段)。在一些实例中,如在Amini等

人,Nature Genetics 46:1343-1349(2014)(2014年10月29日提前在线发布)中所 述地利用转座子的方法,该文献出于所有目的并且具体地是关于涉及 将条形码或其它寡核苷酸标记附接至核酸的所有教导以引用方式整 体并入本文中。在其它实例中,附接条形码的方法可以包括使用切口 酶或聚合酶和/或侵入性探针如recA来产生沿着双链样品核酸的空位 ——然后可以将条形码插入到那些空位中。

[0131] 在使用扩增来标记核酸片段的实例中,寡核苷酸标记可以包含至 少第一和第二区域。第一区域可以是条形码区域,其当在给定分区内 的寡核苷酸之间时可以是基本上相同的条形码序列,而当在不同分区 之间时可以并且在大多数情况下是不同的条形码序列。第二区域可以 是可用于引发分区内的样品内的核酸的N聚体(无规N聚体或设计成 靶向特定序列的N聚体)。在某些情况下,在N聚体被设计为靶向特 定序列的情况下,该N聚体可被设计为靶向特定染色体(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。如本文所论述的,N聚体还可以被设计为倾向于不良表征或具有高度 多态性或与参考序列偏离的基因组所选区域。在一些情况下,N聚体 可以被设计成靶向特定基因或遗传区域,诸如与疾病或病症(例如癌 症)相关的基因或区域。在分区内,可以使用第二N聚体进行扩增反 应,以沿着核酸的长度在不同位置处引发核酸样品。作为该扩增的结 果,每个分区可以含有附接至相同或接近相同的条形码并且可以表示 每个分区中的核酸的更小重叠片段的核酸扩增产物。条形码可以用作 标记来表示来源于相同分区,以及因此潜在地还来源于相同的核酸链 的一组核酸。在扩增后,可以将核酸汇集、测序,以及使用测序算法进行比对。因为较短的序列读段可以借助于其相关条形码序列来比对 和归属于单一较长的样品核酸片段,所以该序列上被鉴定的所有变体 可以归属于单一起始片段和单一起始染色体。此外,通过在多个长片 段间比对多个共定位变体,可以进一步表征染色体贡献。因此,然后 可以得出关于特定遗传变体的定相的结论,诸如可以跨基因组序列的 长距离进行分析——例如跨基因组的不良表征区域的各段鉴定序列 信息。此类信息也可用于鉴定单倍型,单倍型通常是驻留在相同核酸 链或不同核酸链上的规定的遗传变异体集合。拷贝数变异也可以以这 种方式鉴定。

[0132] 所描述的方法和系统提供了优于当前核酸测序技术及其相关样 品制备方法的显著优点。总体样品制备和测序方法被预先设置为主要 用于鉴定和表征样品中的主要成分,而非被设计用来鉴定和表征少量 成分(例如由一个染色体贡献的遗传物质,来自基因组的不良表征或 高度多态区域的遗传物质,或来自一个或多个细胞的物质,或在血流 中循环的片段化的肿瘤细胞DNA分子,它们构成了提取样品中总 DNA的较小百分比)。本文描述的方法包括增加来自这些少数成分的 遗传物质的选择性扩增方法,并且保留此遗传物质的分子环境的能力 进一步提供了这些成分的遗传表征。所描述的方法和系统还提供了检 测较大样品中存在的群体的显著优点。因此,它们尤其可用于评估单 倍型和拷贝数变异——本文公开的方法也可用于提供由于在样品制 备期间引入的偏差而在核酸靶群体中不良表征或不良呈现的基因组 区域上的序列信息。

[0133] 本文所公开的条形码编码技术的使用赋予了为给定的一组遗传 标记提供单独分子环境的独特能力,即将给定的一组遗传标记(与单 个标记相反)归属于单独的样品核酸分子,以及通过变体协调装配, 来为多个样品核酸分子之间和/或特定染色体提供更宽广或甚至更长 距离的推断单独分子环境。这些遗传标记可以包括特异性遗传基因 座,例如

变体(诸如SNP),或者它们可以包括短序列。此外,条形码 的使用赋予了促进对从样品提取的总核酸群体中的少量成分和主要 成分之间进行区分的能力的附加优点,例如检测和表征血流中循环的 肿瘤DNA,并且还减少或消除了任选扩增步骤中的扩增偏差。此外, 微流体形式的实施赋予了使用极小样品体积和低输入量的DNA的能 力,以及快速处理大量样品分区(液滴)以促进全基因组标记的能力。

[0134] 如上所述,本文所描述的方法和系统提供了较长核酸的短序列读 段的单独分子环境。如本文所用,单独分子环境是指超过特定序列读 段的序列环境,例如与不包括在序列读段本身内的相邻或近侧序列的 关系,并且因此将通常使得它们不被全部或部分地包含在短序列读段 中,例如约150个碱基、或约300个碱基以供成对读段的读段。在特 别优选的方面,所述方法和系统为短序列读段提供长距离的序列环 境。这种长距离环境包括给定的序列读段与彼此之间相距长于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或者甚至 长于100kb或更长的距离内的序列读段的关系或关联。通过提供较长 距离的单独分子环境,本发明的方法和系统还提供较长的推断分子环 境。如本文所述的序列环境可以包括较低分辨率的环境,例如来自将 短序列读段映射到单独较长分子或连接分子的重叠群,以及较高分辨 率的环境,例如来自对大部分较长单独分子(例如具有单独分子的连 续确定序列)进行的长距离测序,其中这种确定序列长于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或者甚至长 于100kb。与序列环境一样,短序列至较长核酸(例如,单独长核酸分 子或连接的核酸分子或重叠群的集合)的归属可以包括以下两者:将 短序列针对较长核酸段进行作图以提供高水平序列环境,以及通过这 些较长的核酸从短序列提供装配的序列。

[0135] IV.样品

[0136] 应当理解,本文所论述的方法和系统可用于从任何类型的基因组 材料获得靶序列信息。此类基因组材料可以从取自患者的样品获得。用于本文论述的方法和系统中的示例性样品和基因组材料的类型包 括但不限于多核苷酸、核酸、寡核苷酸、循环的无细胞核酸、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)、血浆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。总的来说,所使用的样品可以根据具体的加工需要来变化。

[0137] 包含核酸的任何物质可以是样品的来源。该物质可以是流体,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、阴道分泌物、胃液和消化液、脊髓液、胎水、腔 液、眼内液、血清、母乳、淋巴液,或它们的组合。该物质可以是固体,例如生物组织。该物质可以包括正常的健康组织、患病组织,或 健康组织和患病组织的混合物。在一些情况下,该物质可以包括肿瘤。 肿瘤可能是良性(非癌症)或恶性肿瘤(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、骨原性肉瘤、 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、肝癌、胆管癌、绒 毛膜癌、精原细胞瘤、胚胎性癌、维尔姆斯瘤、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、 胶质瘤、星形细胞瘤、成神经管细胞瘤、颅咽管瘤、室管膜瘤、松果 体瘤、成血管细胞瘤、听神经瘤、少突神经胶质瘤、脑膜瘤、黑素瘤、 成神经细胞瘤、成视网膜细胞瘤，或它们的组合。该物质可能与各种 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，或它们的组合。在一些情况下，该物质可以 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、 诱导性多能干细胞、胃肠细胞、血细胞、癌细胞、细菌细胞、从人微 生物样品中分离的细菌细胞等。在一些情况下，该物质可以包含细胞 的内容物，例如单个细胞的内容物或多个细胞的内容物。用于分析单 独细胞的方法和系统提供于例如提交于2015年6月26日的美国专利 申请No.14/752,641中，该专利申请的全部公开内容由此全文以引用 方式并入本文。

[0138] 样品可以从多个受试者获得。受试者可以是活的受试者或死亡的 受试者。受试者的实例可包括但不限于人、哺乳动物、非人哺乳动物、 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、艾维茵肉鸡(avines)、小鼠、兔、昆虫、蛞蝓、微生物、细菌、寄生虫或鱼类。在一些情况下，受试者可 能是患有、疑似患有或有患上某疾病或病症风险的患者。在一些情况 下，受试者可以是孕妇。在一些情况下，受试者可以是正常的健康孕 妇。在一些情况下，该受试者可以是有怀上具有某些出生缺陷的婴儿的风险的孕妇。

[0139] 可以通过本领域中已知的任何方式从受试者获得样品。例如，可 以通过以下方式从受试者获得样品：接取循环系统(例如，通过注射 器或其它装置静脉内或动脉内地)、收集分泌的生物样品(例如，唾液、 痰液、尿、粪便等)、手术地(例如，活体组织切片)获取生物样品(例 如手术中样品、手术后样品等)、擦拭(例如颊拭子、口咽拭子)，或移 液。

[0140] 虽然本文已经示出和描述了本发明的优选实施方案，但是对于本 领域的技术人员将显而易见的是，这些实施方案仅以举例方式提供。 在不脱离本发明的情况下，本领域的技术人员现将思及许多变化、改 变和替换。应当理解，本文所述的本发明的实施方案的各种替代方案 可用于实践本发明。以下权利要求旨在限定本发明的范围，并且落入 这些权利要求范围内的方法和结构及其等同物将由此被涵盖在本发 明的范围内。

[0141] 实施例

[0142] 实施例：对TP53基因的靶向覆盖

[0143] 进行靶向TP53基因的扩增反应。肿瘤蛋白p53，也被称为p53、细胞肿瘤抗原p53(UniProt名称)、磷蛋白p53、肿瘤抑制因子p53、抗原NY-CO-13或转化相关蛋白53(TRP53)，是由人TP53基因编码 的蛋白。p53蛋白在多细胞生物体中是至关重要的，其在多细胞生物体中调节细胞周期，并且因而用作肿瘤抑制因子，从而预防癌症。因 此，p53由于其通过防止基因组突变来保存稳定性的作用而被描述为 “基因组卫士”。因此，TP53被分类为肿瘤抑制基因。

[0144] 使用在多重反应中跨整个基因的总共96个引物进行对含有TP53 基因的基因组区域(其长度为约19149bp)的靶向扩增。将引物设计成 跨基因组的该区域以约400bp的间隔拼接。使用退火步骤的温度梯度 进行扩增反应达14个循环，并且输入量为约3ng DNA。本实

施例中 使用的热循环方案如下：

[0145]

初始预变性	98℃	30秒
18个循环	98℃	10秒
	30-55℃	15秒
	72℃	15秒
最终延伸	72℃	2分钟
保存	4C	

[0146] 这种类型的反应的示例性工作流程示出于图8中。应当理解，这 是根据本文所述的本发明方法的一个示例性实施方案，并且可以使用 已知方法进行改变或扩展。如图8所示，使用靶特异性引物(诸如示 出为802和803的那些)来扩增基因组的所选区域(在这种情况下为 TP53基因)。此外，具有条形码801的引物也被并入扩增子中，其可 以在如本文所述的某些实施方案中为后续序列读段(808)提供分子环 境。

[0147] 引物802和803在本实验中具有“尾”R1和R2，R1和R2使得所 得扩增子适于在特定平台(诸如Illumina平台)上进行测序。使用SI引 物(806)的扩增进一步提供了同样用于Illumina平台的样品指标。应当 理解，可以使用可用于其它测序平台的序列来代替R1和R2以及S1 引物。

[0148] 图9示出扩增反应是特异性的，因为无模板对照(NTC)显示为没 有产物。图10提供了被看作上述方案在一定温度范围内的结果的富 集倍数。

[0149] 本说明书提供了对当前描述的技术在示例性方面中的方法、系统 和/或结构及其用途的全面描述。虽然上文已经以某种程度的特殊性 或参考一个或多个单独方面对本技术的各个方面进行了描述，但是本 领域技术人员可以在不脱离本发明技术的精神或范围的情况下对所 公开的各方面进行多种改变。由于在不脱离目前描述的技术的精神和 范围的情况下可以进行许多方面，因此适当的范围取决于下文随附的 权利要求书。因此设想了其它方面。此外，应当理解，可以以任何顺 序执行任何操作，除非以其它方式明确地要求保护或权利要求语言固 有地需要特定的次序。包含在上述描述中并在附图中示出的所有内容 应旨在被解释为仅为对特定方面的说明，而不是限制所示出实施方 案。除非根据上下文可清楚或另有明确说明，本文所提供的任何浓度 值一般是就混合物值或百分比给定的，而与添加混合物的特定组分时 或之后发生的任何转化无关。对于本文尚未明确包含的内容，在本公 开中引用的所有公开的参考文献和专利文献出于所有目的而全文以 引用方式并入本文。在不脱离如以下权利要求书所限定的本技术的基 本要素的情况下，可以进行细节或结构方面的改变。

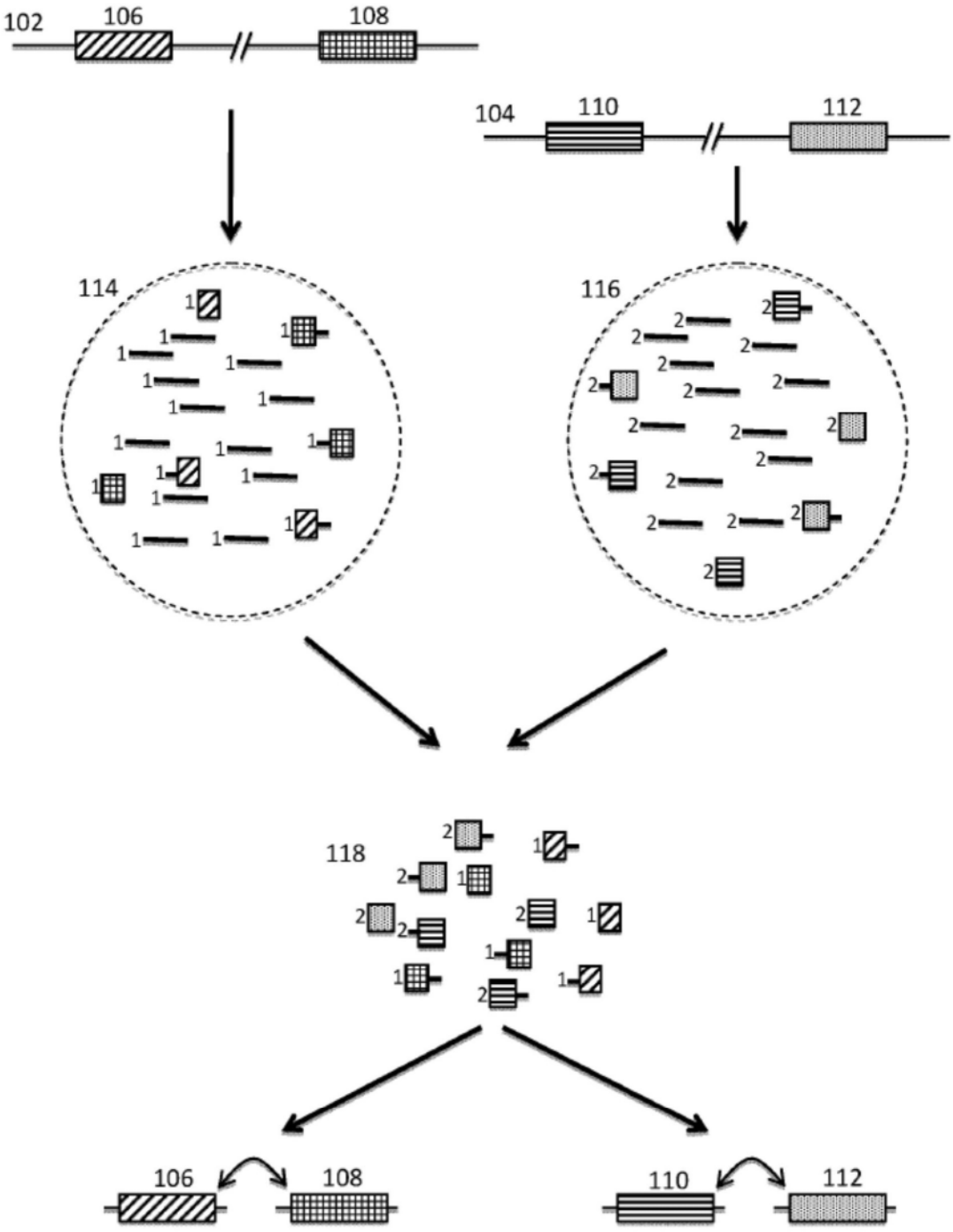


图1

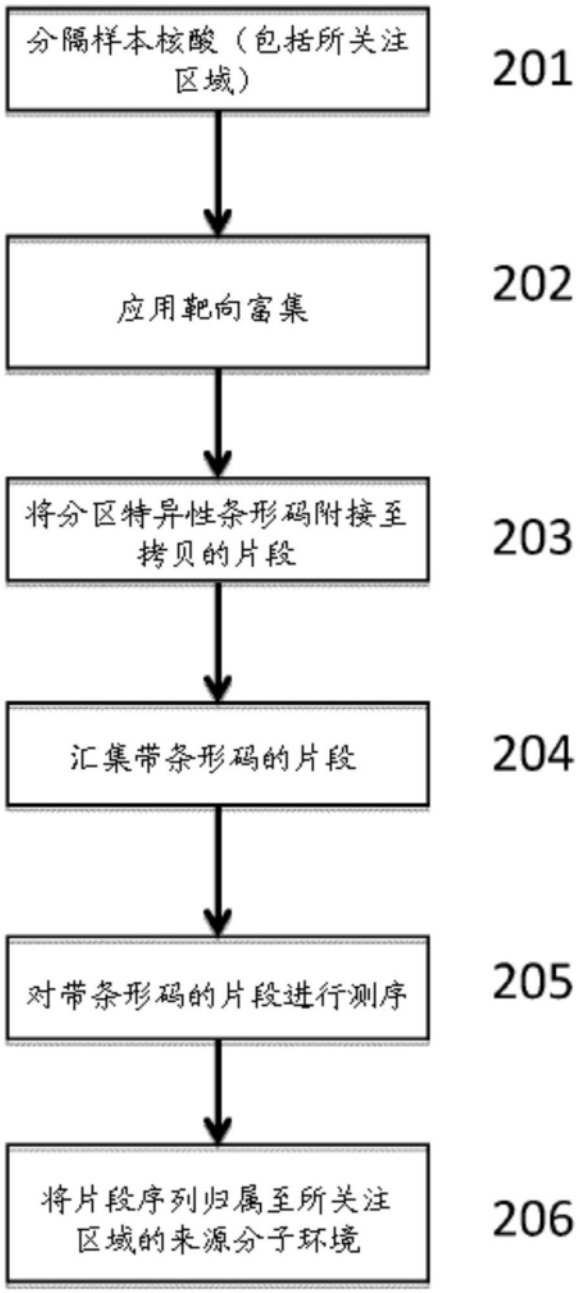

分隔样本核酸（包括所关注区域）
201
应用靶向富集
202
将分区特异性条形码附接至拷贝的片段
203
汇集带条形码的片段
204
对带条形码的片段进行测序
205
将片段序列归属至所关注区域的来源分子环境
206

图2

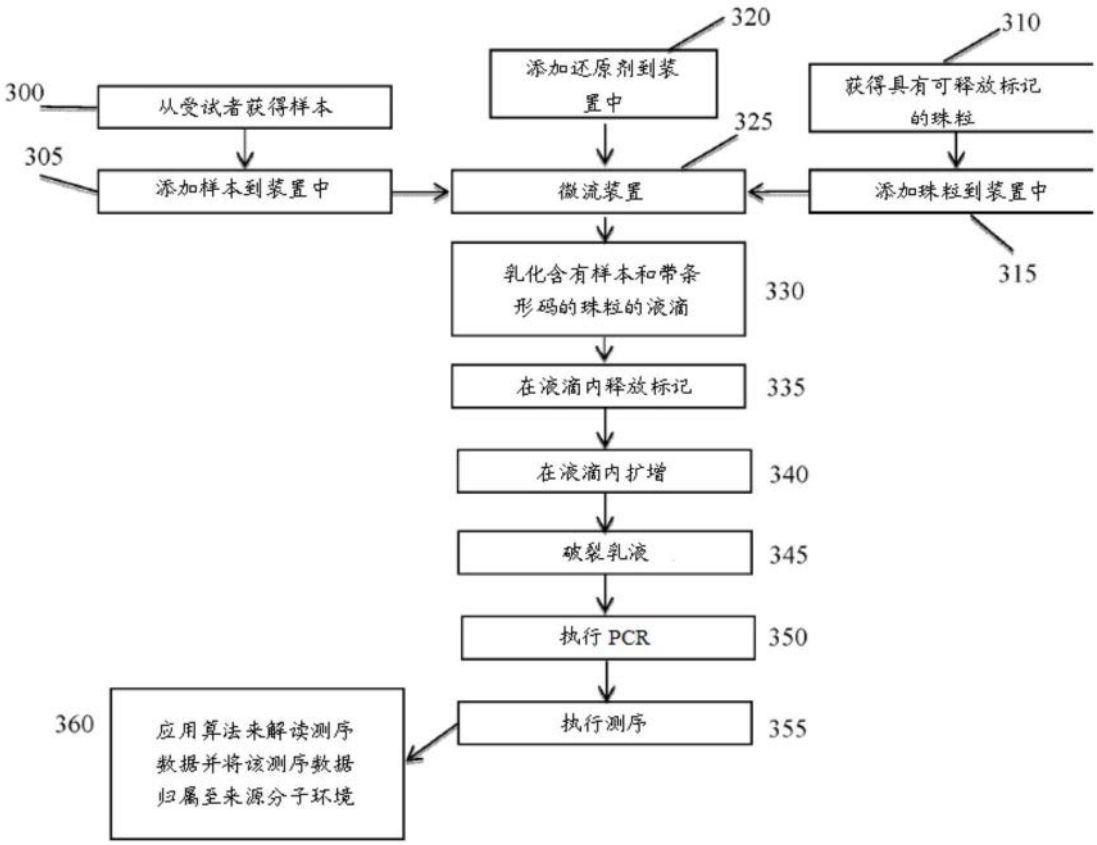

300
从受试者获得样本
305
添加样本到装置中
320
添加还原剂到装置中
325
微流装置
310
获得具有可释放标记的珠粒
添加珠粒到装置中
315
乳化含有样本和带条形码的珠粒的液滴
330
在液滴内释放标记
335
在液滴内扩增
340
破裂乳液
345
执行 PCR
350
执行测序
355
360
应用算法来解读测序数据并将该测序数据归属至来源分子环境

图3

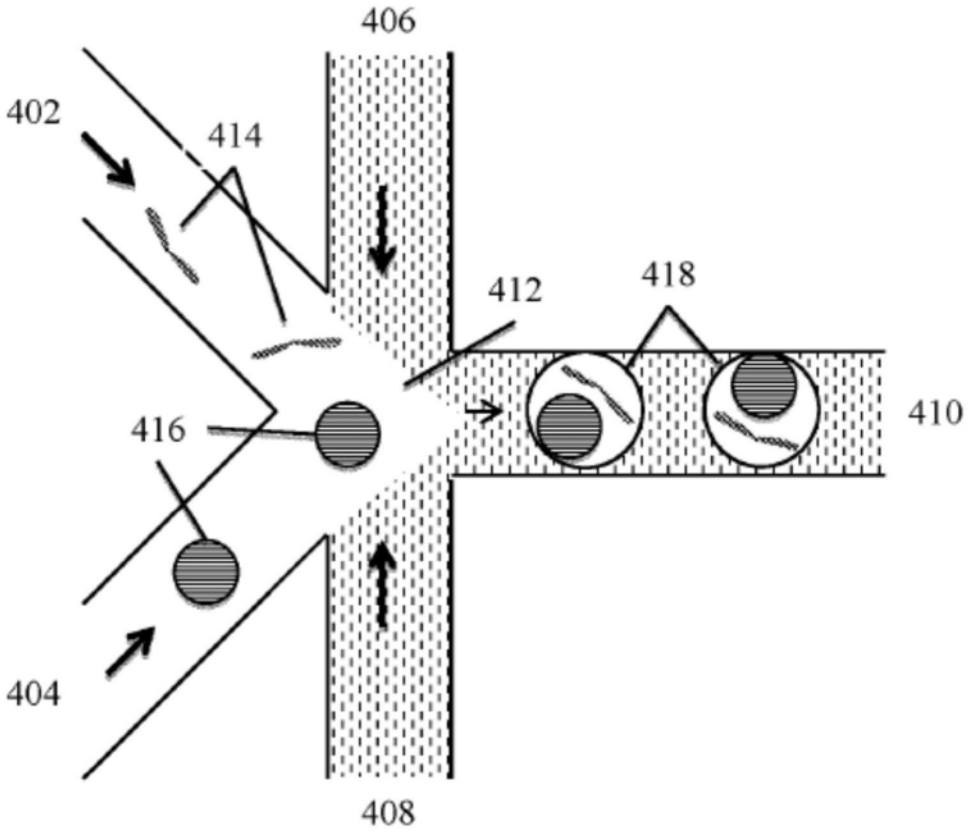


图4

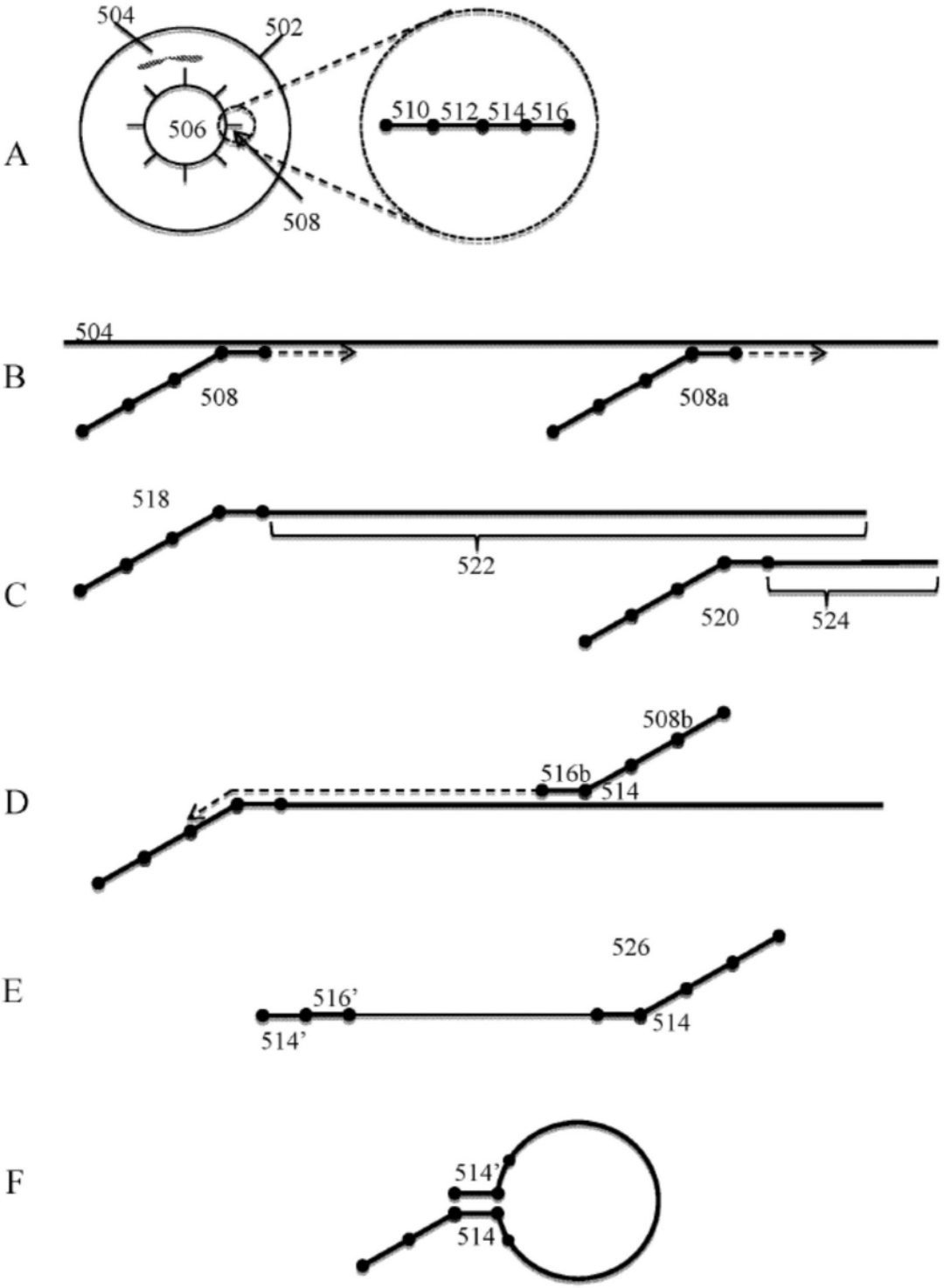

A
504
502
506
510 512 514 516
508
B
504
508
508a
C
518
522
520
524
D
508b
516b
514
E
526
516'
514'
514
F
514'
514

图5

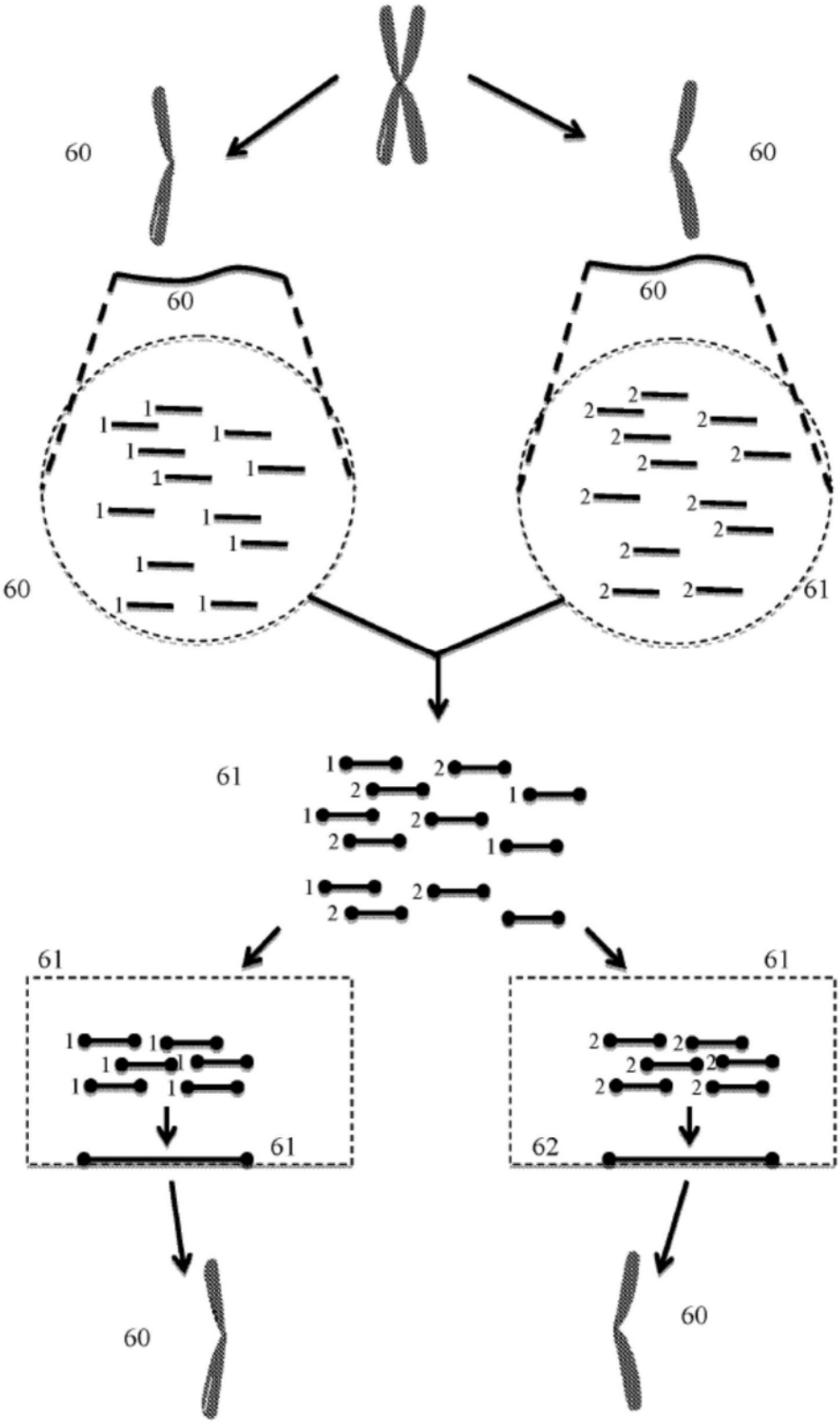


图6A

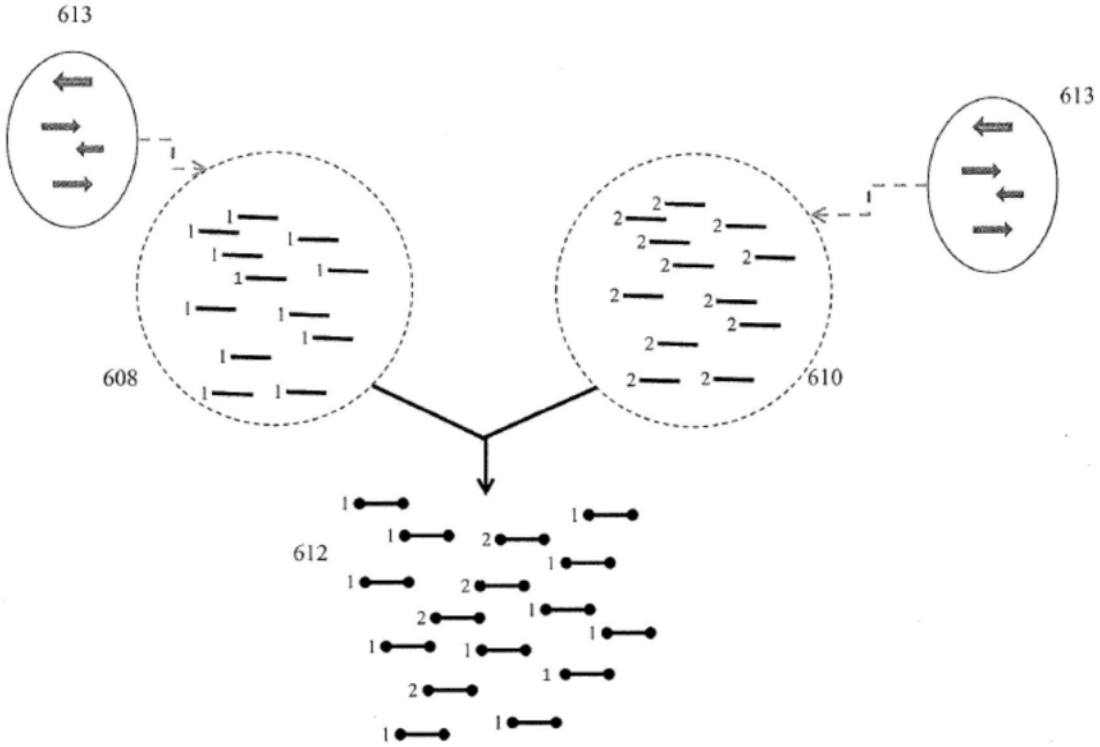


图6B

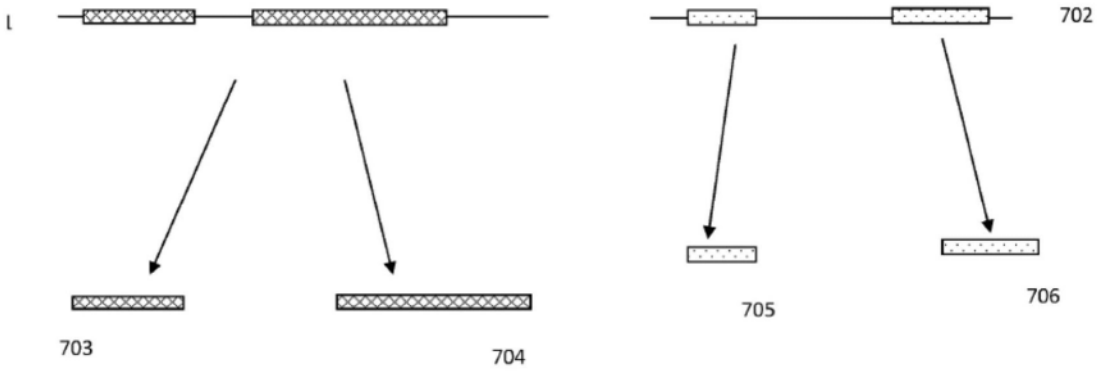


图7

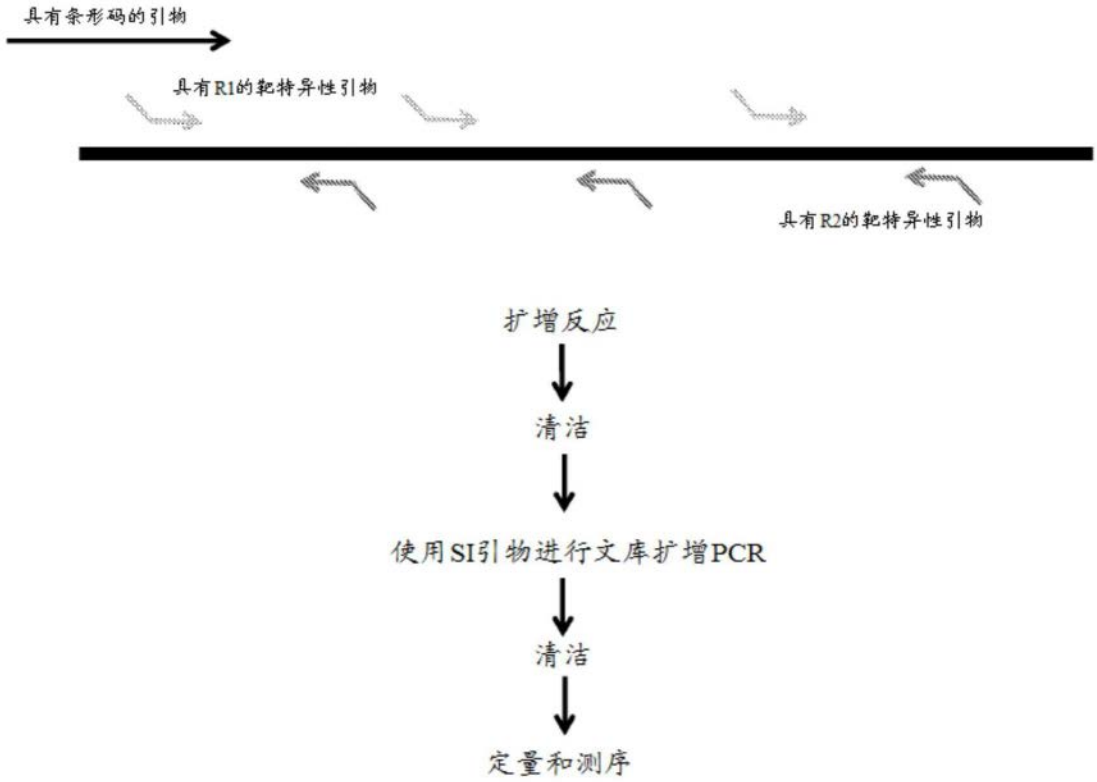

具有条形码的引物
具有R1的靶特异性引物
具有R2的靶特异性引物
扩增反应
清洁
使用SI引物进行文库扩增PCR
清洁
定量和测序

图8

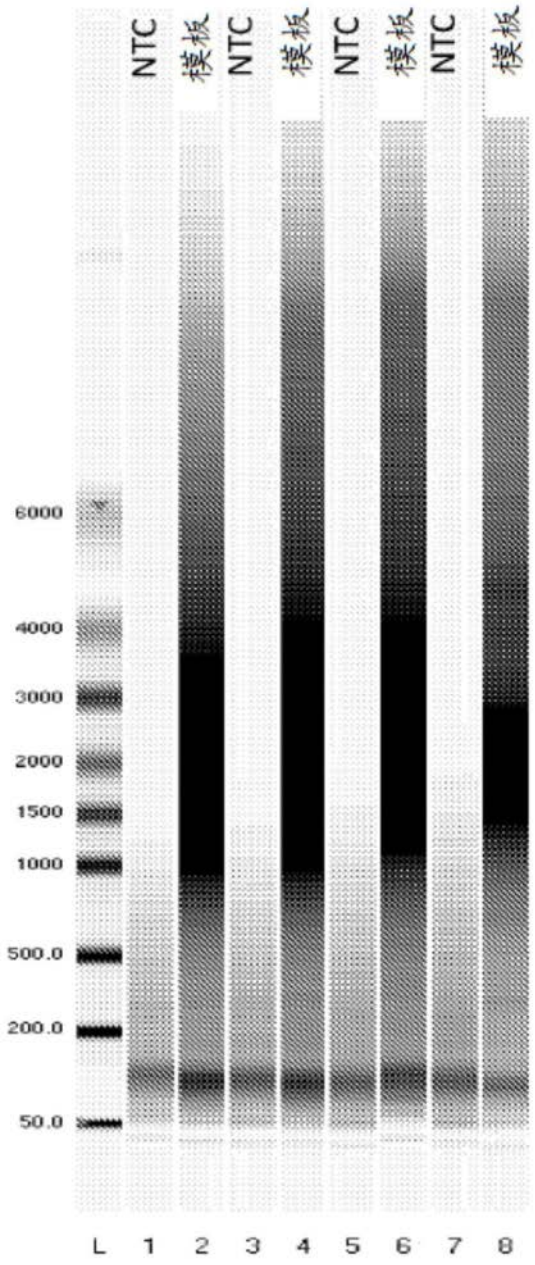

NTC
模板
NTC
模板
NTC
模板
NTC
模板
6000
4000
3000
2000
1500
1000
500.0
200.0
50.0
L 1 2 3 4 5 6 7 8

图9

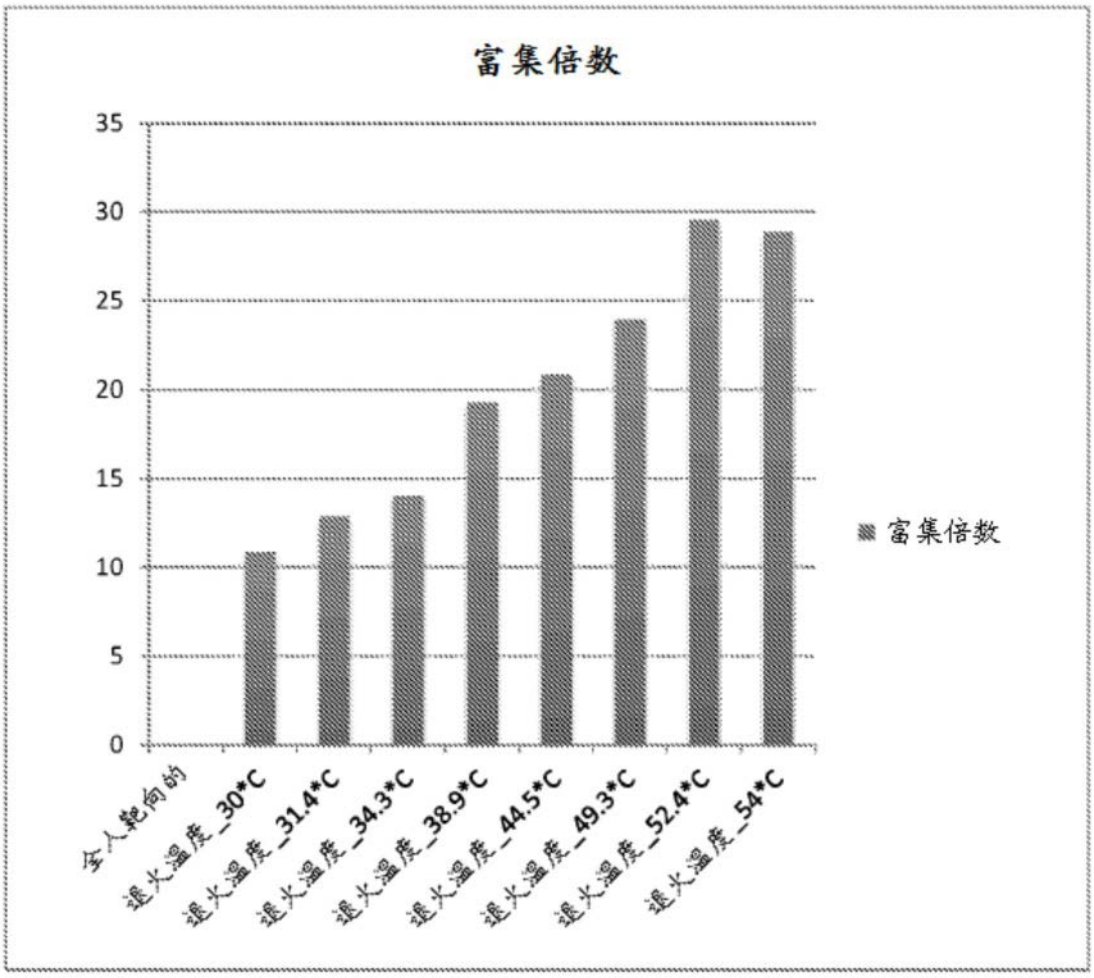

富集倍数
35
30
25
20
15
10
5
0
全人靶向的
退火温度_30*C
退火温度_31.4*C
退火温度_34.3*C
退火温度_38.9*C
退火温度_44.5*C
退火温度_49.3*C
退火温度_52.4*C
退火温度_54*C
富集倍数

图10