(19)国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 114807307 A
(43)申请公布日 2022.07.29

(21)申请号 202210420031.7
(22)申请日 2015.10.29
(30)优先权数据
62/072164 2014.10.29 US
(62)分案原申请数据
201580059008.8 2015.10.29
(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.贾罗什 M.施纳尔-莱文 S.萨克索诺夫 B.欣森 郑新颖
(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 彭昶
(51)Int.Cl.
C12Q 1/6806(2018.01)
C12Q 1/6874(2018.01)

权利要求书1页 说明书33页 附图8页

(54)发明名称

用于靶核酸测序的方法和组合物

(57)摘要

本发明涉及用于捕获和分析包含在基因组的靶区域中的序列信息的方法、组合物和系统。这样的靶区域可包括外显子组、部分外显子组、内含子、外显子区和内含子区的组合、基因、基因的组以及可能感兴趣的全基因组的任何其它子集。

CN 114807307 A

1.一种用于对基因组的一个或多个选定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)将来自所述起始基因组材料的个体核酸分子分配到离散分区中,使得每个离散分区包含第一个体核酸分子;

(c)将所述离散分区中的所述个体核酸分子片段化以形成多个片段,其中每个所述片段还包含条形码,并且其中给定离散分区内的片段各自包含共同条形码,由此将每个片段与其所源自的个体核酸分子相关联;

(d)提供富集包含所述基因组的所述一个或多个选定部分的至少一部分的片段的群体;

(e)从所述群体获得序列信息,从而对基因组的一个或多个选定部分进行测序。

2.根据权利要求1所述的方法,其中所述提供步骤(d)包括:

(i)将与所述基因组的所述一个或多个选定部分中或附近区域互补的探针与所述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持物的表面;

从而使所述群体富集包含所述基因组的所述一个或多个选定部分的至少一部分的片段。

3.根据权利要求2所述的方法,其中所述固体支持物包含珠。

4.根据权利要求2至3所述的方法,其中所述探针包含结合部分,并且所述表面包含捕获部分,并且其中所述探针-片段复合物通过所述结合部分与所述捕获部分之间的反应而捕获在所述表面上。

5.根据权利要求4所述的方法,其中所述捕获部分包含链霉亲和素,而所述结合部分包含生物素。

6.根据权利要求4所述的方法,其中所述捕获部分包含链霉亲和素磁珠,而所述结合部分包含生物素化RNA文库饵。

7.根据权利要求4所述的方法,其中所述捕获部分针对选自以下的成员:全外显子组或部分外显子组捕获、组捕获、靶向外显子捕获、锚定外显子组捕获和瓦片式覆盖基因组区域捕获。

8.根据权利要求1至7所述的方法,其中在所述获得步骤(e)之前,扩增所述片段以形成扩增产物。

9.根据权利要求8所述的方法,其中所述扩增产物能够形成部分或完整的发夹结构。

10.根据权利要求1至9所述的方法,其中所述获得步骤(e)包括选自以下的测序反应:短读取长度测序反应和长读取长度测序反应。

用于靶核酸测序的方法和组合物

[0001] 相关申请的交叉引用

[0002] 本申请要求于2014年10月29日提交的美国临时申请号62/072,164的权益,其明确地通过引用整体并入本文用于所有目的。

[0003] 发明背景

[0004] 精确并快速地对基因组进行测序的能力正在彻底改变生物学和医学。复杂基因组的研究,特别是对人类疾病的遗传基础的探索涉及大规模的遗传分析。在全基因组水平的这样的遗传分析不仅在金钱上而且在时间和劳资方面花费高。这些成本随着涉及不同个体DNA样品分析的方案而增加。基因组中与疾病发展相关的多态性区的测序(和重测序)将极大地有助于理解例如癌症的疾病和治疗剂研发,并且将有助于应对鉴定与药物反应的变化性相关的基因和功能多态性的药物基因组学挑战。在给定基因型与特定疾病之间可建立关联之前,需要对大到足以产生统计学显著数据的群体进行大量遗传标记的筛选。

[0005] 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的一种方法是对基因组的靶区域进行高通量、高精确度测序。广泛使用的方法捕获基因组的整个蛋白质编码区(外显子组)的大部分,外显子组占人类基因组的约1%;并且已经成为临床和基础研究中的常规技术。外显子组测序提供了优于全基因组测序的优点:它显著不那么贵、功能解释更易理解、分析显著更快、使得非常深入的测序负担得起,并且得到更易于管理的数据集。需要用于富集感兴趣的目标区域的方法、系统和组合物用于高精确度和高通量测序和遗传分析。

[0006] 发明概述

[0007] 因此,本发明提供了用于获得基因组靶区域的序列信息的方法、系统和组合物。

[0008] 在一些方面,本公开提供了一种用于对基因组的一个或多个选定部分进行测序的方法,所述方法大体上包括以下步骤:(a)提供起始基因组材料;(b)将来自所述起始基因组材料的个体核酸分子分配到离散分区中,使得每个离散分区包含第一个体核酸分子;(c)将所述离散分区中的所述个体核酸分子片段化以形成多个片段,其中每个片段还包含条形码,并且其中给定离散分区内的片段各自包括共同条形码,由此将每个片段与其所源自的个体核酸分子相关联;(d)提供富集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群体;(e)从所述群体获得序列信息,从而对基因组的一个或多个选定部分进行测序。

[0009] 在另一些实施方案中并且根据上述,提供富集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群体包括以下步骤:(i)使与所述基因组的一个或多个选定部分中或附近的区域互补的探针与所述片段杂交以形成探针-片段复合物;和(ii)将探针-片段复合物捕获到固体支持物的表面;从而使所述群体富集包含基因组的一个或多个选定部分的至少一部分的片段。在另一些实施方案中,所述固体支持物包括珠。在另一些实施方案中,所述探针包括结合部分,而所述表面包括捕获部分,并且所述探针-片段复合物通过结合部分与捕获部分之间的反应而捕获在表面上。在另一些实例中,所述捕获部分包括链霉亲和素,而所述结合部分包括生物素。在另一些实例中,所述捕获部分包括链霉亲和素磁珠,而所述结合部分包括生物素化RNA文库饵。

[0010] 在一些实施方案中并且根据任意上述，本发明的方法包括使用捕获部分，其涉及全外显子组或部分外显子组捕获、组捕获(panel capture)、靶向外显子捕获、锚定外显子组捕获或瓦片式覆盖(tiled)基因组区域捕获。

[0011] 在另一些实施方案中并且根据任意上述，本文公开的方法包括获得步骤，所述获得步骤包括测序反应。在另一些实施方案中，测序反应是短读取长度测序反应或长读取长度测序反应。在另一些实例中，测序反应提供起始基因组材料的小于90％、小于75％或小于50％的序列信息。

[0012] 在另一些实施方案中，本文所述的方法还包括基于分离的片段的重叠序列在推断的重叠群中连接两个或更多个个体核酸分子，其中所述推断的重叠群包含至少10kb、20kb、40kb、50kb、100kb或200kb的长度N50。

[0013] 在另一些实施例中并且根据任意上述，本文公开的方法还包括基于分离的片段的序列内的重叠定相变体(phased variants)在相区块(phase block)中连接两个或更多个个体核酸分子，其中相区块包含至少10kb、至少20kb、至少40kb、至少50kb、至少100kb或至少200kb的长度N50。

[0014] 在另一些实施方案中并且根据任意上述，本文公开的方法提供来自一起覆盖外显子组的基因组的选定部分的序列信息。在另一些实施方案中，离散分区中的个体核酸分子包括来自单细胞的基因组DNA。在另一些实施方案中，离散分区各自包括来自不同染色体的基因组DNA。

[0015] 在另一些方面，本公开提供了一种从基因组样品的一个或多个靶部分获得序列信息的方法。这样的方法包括但不限于以下步骤：(a)在离散分区中提供基因组样品的个体第一核酸片段分子；(b)将所述离散分区内的所述个体第一核酸片段分子片段化以由所述个体第一核酸片段分子中的每一个产生多个第二片段；(c)将共同条形码序列附接到离散分区内的所述多个第二片段，使得所述多个第二片段中的每一个可归属于其中包含它们的离散分区；(d)将针对所述基因组样品的一个或多个靶部分的探针文库应用于所述第二片段；(e)进行测序反应以鉴定与探针文库杂交的多个第二片段的序列，从而从基因组样品的一个或多个靶部分获得序列信息。在另一些实施方案中，将探针文库连接到结合部分，并且在进行步骤(e)之前，通过结合部分与捕获部分之间的反应，在包含捕获部分的表面上捕获第二片段。在另一些实施方案中并且在进行步骤(e)之前，第二片段在第二片段被捕获在表面上之前或之后被扩增。在另一些实施方案中，结合部分包含生物素，而捕获部分包含链霉亲和素。在另一些实施方案中，测序反应是短读取、高精确度测序反应。在另一些实施方案中，扩增第二片段以使得所得扩增产物能够形成部分或完整的发夹结构。

[0016] 在另一些方面并且根据任意上述，本公开提供了用于从基因组样品的一个或多个靶部分获得序列信息同时保留分子背景的方法。这样的方法包括以下步骤：(a)提供起始基因组材料；(b)将来自所述起始基因组材料的个体核酸分子分配到离散分区中，使得每个离散分区包含第一个体核酸分子；(c)将所述离散分区中的所述第一个体核酸分子片段化以形成多个片段；(d)提供富集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群体；(e)从所述群体获得序列信息，由此对基因组样品的一个或多个靶部分进行测序同时保留分子背景。在另一些实施方案中，在获得步骤(e)之前，用条形码标记多个片段以将每个片段与其形成于其中的离散分区相关联。在另一些实施方案中，将步骤(b)中的个体核酸分

子分配以使得保持每个第一个体核酸分子的分子背景。

[0017] 在一些方面,本公开提供了从基因组样品的一个或多个靶部分获得序列信息的方法。这样的方法包括但不限于以下步骤:(a)提供基因组样品的个体核酸分子;(b)将所述个体核酸分子片段化以形成多个片段,其中每个片段还包含条形码,并且其中来自同一个体核酸分子的片段具有共同条形码,由此将每个片段与其所源自的核酸分子相关联;(c)使所述多个片段富集含有所述基因组样品的一个或多个靶部分的片段;和(d)进行测序反应以鉴定富集多个片段的序列,从而从所述基因组样品的一个或多个靶部分获得序列信息。在另一些实施方案中,富集步骤包括应用针对基因组样品的一个或多个靶部分的探针文库。在另一些实施方案中,探针文库连接到结合部分,并且在进行步骤之前,通过结合部分与捕获部分之间的反应捕获片段。在一些示例性实施方案中,结合部分与捕获部分之间的反应将片段固定在表面上。

[0018] 附图简述

[0019] 图1提供了使用常规方法相对于本文所述的方法和系统鉴定和分析靶基因组区的示意图。

[0020] 图2A和图2B提供了使用本文所述的方法和系统鉴定和分析靶基因组区的示意图。

[0021] 图3示出了使用本文公开的方法和组合物进行检测序列信息的测定的典型工作流程。

[0022] 图4提供了用于将核酸样品与珠组合并将核酸和珠分区到离散液滴中的方法的示意图。

[0023] 图5提供了染色体核酸片段的加条形码和扩增方法的示意图。

[0024] 图6提供了在将序列数据归属于单个染色体时使用染色体核酸片段的加条形码的示意图。

[0025] 图7示出了本发明方法的一个一般实施方案。

[0026] 图8示出了本发明方法的一个一般实施方案。

[0027] 发明详述

[0028] 除非另有说明,否则本发明的实践可以采用在本领域技术人员能力范围内的有机化学、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。这样的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。通过参考下文的实例可以得到合适技术的具体说明。然而,当然也可以使用其它等同的常规程序。这样的常规技术和描述可见于标准实验室手册中,例如GenomeAnalysis:A LaboratoryManualSeries(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 ofBiochemistry第3版,W.H.Freeman Pub.,NewYork,N.Y.和Berg等(2002) Biochemistry,第5版,W.H.Freeman Pub.,New York,N.Y.,所有这些文献均通过引用整体并入本文用于所有目的。

[0029] 注意,除非上下文另有明确说明,否则如本文和所附权利要求书中所使用的,单数

形式“一个/种(a/an)”和“所述/该(the)”包括复数个指代物。因此,例如,提及“聚合酶”是指一种试剂或这样的试剂的混合物,提及“所述方法”包括提及本领域技术人员已知的等同步骤和方法等。

[0030] 除非另有定义,否则本文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通过引用并入本文,用于描述和披露在出版物中描述的和可能与目前描述的发明结合使用的装置、组合物、制剂和方法的目的。

[0031] 在提供值的范围时,应当理解,除非上下文另有明确说明,否则在该范围的上限和下限之间的每个中间值至下限单位的十分之一,以及在所述范围内的任何其它所述值中间值均包括在本发明内。可独立地包括在较小范围内的这些较小范围的上限和下限也包括在本发明内,服从于所述范围中的任何特别排除的限制。当所述范围包括一个或两个限制时,排除那些所包括的限制的范围的范围也包括在本发明中。

[0032] 在下面的描述中,阐述了许多具体细节以提供对本发明的更透彻的理解。然而,对本领域技术人员显而易见的是,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施。在其它情况下,没有描述本领域技术人员公知的众所周知的特征和程序,以避免模糊本发明。

[0033] 如本文所用,术语“包含”旨在表示组合物和方法包括所列举的要素,但不排除其它要素。当用于定义组合物和方法时,“基本上由……组成”意味着排除对组合物或方法具有任何本质意义的其它要素。“由……组成”是指排除对于所要求保护的组合物的其它成分和实质性方法步骤的超过痕量要素。由这些过渡术语中的每一个限定的实施方案在本发明的范围内。因此,意在方法和组合物可以包括另外的步骤和组分(包括)或可选地包括不重要的步骤和组合物(基本上由……组成),或者仅仅意指所述的方法步骤或组合物(由……组成)。

[0034] 所有数字表示,例如pH、温度、时间、浓度和分子量,包括范围,是以0.1的增量变化(+)或(-)的近似值。应当理解,虽然不总是明确地指出,但是所有数字指示前置有术语“约”。除了例如“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”。还应当理解,虽然不总是明确说明,但是本文所述的试剂仅是示例性的,并且其等同物在本领域中是已知的。

[0035] I.综述

[0036] 本公开提供了可用于表征遗传物质的方法、组合物和系统。特别地,本文所述的方法、组合物和系统提供了基因组靶区域的遗传表征,包括但不限于特定染色体、染色体区、所有外显子(外显子组)、外显子组的部分、特异性基因、激酶的组(例如,激酶组(kinome)或其它靶基因的组)、内含子区、基因组的瓦片式覆盖部分或基因组的任何其它所选部分。

[0037] 一般而言,本文所述的方法和系统通过提供长个体核酸分子的序列测定和/或鉴定由长延伸序列间隔开的两个序列区段之间的直接分子连接来完成靶基因组测序,其允许鉴定和使用长程序列信息,但是该测序信息是使用具有短读取测序技术的极低测序错误率和高通量的优点的方法获得的。本文所述的方法和系统将长核酸分子分成较小的片段,其可以使用高通量、更高精确度的短读取测序技术进行测序,并且以允许源自较小片段的序列信息保留原始长程分子序列背景,即允许较短读取序列归属于原始较长个体核酸分子的

方式完成分区段。通过将读取序列归属于原始较长核酸分子，可以获得通常不能单独从短读取序列获得的较长核酸序列的大量表征信息。该长程分子背景不仅通过测序过程保留，而且通过本文所述的靶向测序方法中使用的靶向富集过程来保留，没有其它测序方法显示出这种能力。

[0038] 一般来说，来自较小片段的序列信息将通过使用标记程序保留原始长程分子序列背景，所述标记程序包括添加如本文所述和本领域已知的条形码。在一些特定实例中，源自同一原始较长个体核酸分子的片段将用共同条形码标记，使得来自那些片段的任何后来的读取序列均可归属于该原始较长个体核酸分子。这样的条形码可以使用本领域已知的任何方法添加，包括在扩增个体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个体核酸分子，包括例如Amini等，Nature Genetics 46:1343-1349(2014)(在2014年10月29日提前在线出版)中描述的那些方法，该文献通过引用整体并入本文用于所有目的，特别是涉及使用转座子添加衔接子及其它寡核苷酸的所有教导。一旦使用这样的方法标记核酸，可以采用本文所述的方法富集得到的经标记片段，使得片段群体表示基因组的靶区域。因此，来自该群体的读取序列允许对基因组的选择区域进行靶向测序，并且那些读取序列也可以归属于原始核酸分子，从而保留原始长程分子序列背景。可以采用本领域已知的和本文描述的任何测序方法和平台获得读取序列。

[0039] 除了提供从基因组的靶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，包括但不限于单元型定相(haplotypephasing)、结构变化的鉴定和鉴定拷贝数变化，如共同未决的申请USSN 14/752,589和14/752,602中描述的(两者均于2015年6月26日提交)，这些专利通过引用整体并入本文用于所有目的，特别是用于针对基因组材料的表征的所有书面描述、附图和工作实例。

[0040] 根据本申请中描述的方法和系统对核酸进行处理和测序的方法还在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，这些专利通过引用整体并入本文用于所有目的，并且特别是用于涉及处理核酸和测序以及基因组材料的其它表征的所有书面描述、附图和工作实例。

[0041] 一般来说，如图1所示，本文所述的方法和系统可用于表征核酸。特别地，如所示，示出了两个离散的单个核酸102和104，每个具有数个目标区域，例如核酸102中的区域106和108，以及核酸104中的区域110和112。每个核酸的目标区域在同一核酸分子内相连，但是可以彼此相对分开，例如，相距超过1kb、相距超过5kb、相距超过10kb、相距超过20kb、相距超过30kb、相距超过40kb、相距超过50kb，且在一些情况下，相距达100kb。所述区域可以指单个基因、基因组、外显子或简单地离散和分离的基因组部分。仅为了便于讨论，图1中所示的区域将被称为外显子106、108、110和112。如所示，每个核酸102和104分别被分离成其自身的分区114和116。如本文别处所述，在许多情况下，这些分区是油包水乳液中的含水液滴。在每个液滴内，每个片段的部分以保留那些片段的原始分子背景的方式拷贝，例如，源自同一分子。如所示，这是通过在每个拷贝的片段中包含条形码序列(例如，如所示的条形码序列“1”或“2”，其代表原始片段被分成的液滴)来实现的。对于全基因组序列分析应用，可以简单地合并所有拷贝的片段及其相关条形码，以便对每个原始核酸102和104的全范围序列信息进行排序和重组。然而，在许多情况下，更希望仅分析整个基因组的特定靶部分，例如外显子组、特异性基因等，以便更集中于基因组的科学相关部分，并且使对基因组不太

相关或不相关的部分进行测序的时间和花费最小化。

[0042] 根据本文所述的方法，可以将靶标富集步骤应用于带条形码序列片段文库，以“下拉”与所需靶标相关的序列。这些可以包括外显子靶向下拉、基因的组特异性靶向下拉等。允许进行基因组的特异性靶区域的富集分离的大量靶向下拉试剂盒是市售的，例如Agilent SureSelect外显子下拉试剂盒等。如图1所示，靶向富集的应用得到富集的带条形码序列文库118。此外，因为文库118内的下拉片段保留其原始分子背景，例如通过保留条形码信息，它们可以重新组装入具有嵌入的长程连接信息的其原始分子背景中，例如，具有每个组装的目标区域106:108和110:112之间的推断的连接。例如，可以鉴定基因组的两个不同靶部分(例如，两个或更多个外显子)之间的直接分子连接，并且直接分子连接可用于鉴定结构变化和其它基因组特征，以及鉴定关于两个或更多个外显子的相信息，例如，提供定相外显子，包括潜在的全定相外显子组或基因组的其它定相靶部分。

[0043] 通常，本发明的方法包括如图7所示的步骤，其提供了本文进一步详细讨论的本发明方法的示意性概述。如将理解的，图9中概述的方法是可以根据需要和如本文所述改变或修改的示例性实施方案。

[0044] 如图7所示，本文所述的方法在大多数实例中将包括其中包含目标靶区的样品核酸分区的步骤(701)。通常，每个分区将包括来自特定基因座的单一个体核酸分子，然后通过以保留片段的原始分子背景的方式将其片段化或拷贝(702)，通常通过对它们所含分区具有特异性的区段加条形码。在一些实例中，每个分区可以包括多于一个核酸，并且在一些情况下包含数百个核酸分子-在多个核酸在分区内的情况下，基因组的任何特定基因座通常在加条形码之前由单一单个核酸表示。可使用本领域已知的任何方法产生步骤702的带条形码片段-在一些实例中，寡核苷酸是不同分区内的样品。这样的寡核苷酸可包含意在随机引发样品的许多不同区域的随机序列，或者它们可以包含靶向引发样品靶区域的上游的特异性引物序列。在另一些实例中，这些寡核苷酸还包含条形码序列，使得复制过程还使原始样品核酸的所得复制片段带条形码。在均于2014年6月26日提交的美国专利申请号14/316,383、14/316,398、14/316,416、14/316,431、14/316,447、14/316,463中详细描述了在放大和条形码样品中使用这些条形码寡核苷酸的特别优选的方法，这些专利申请各自通过引用并入本文用于所有目的。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核苷酸及其相关条形码序列。多个引物对样品的不同部分的退火和延伸可以导致样品的重叠互补片段的大合并，每个具有其自身的条形码序列，其指示其在其中产生的分区。在一些情况下，这些互补片段本身可用作由分区中存在的寡核苷酸引发的模板，以产生补体的补体，所述补体又包括带条形码序列。在另一些实例中，配置该复制过程，使得当第一互补序列复制时，其在其末端或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，其降低分子作为用于生成进一步的迭代拷贝的基础的能力。

[0045] 返回到图7中例示的方法，一旦分区专用条形码附接到拷贝的片段，则合并条形码片段(703)。然后可以应用靶富集技术(704)以“下拉”目标靶区域。然后对这些目标靶区域进行测序(705)并且将所述片段的序列归属于它们的原始分子背景(706)，使得目标靶区域被鉴别并且也与所述原始分子背景相关联。本文所述和图7中说明的所述方法和系统的独

特特征是在靶向富集步骤(704)之前将条形码附接到片段(702)。本文所述的方法和系统的优点是在使靶基因组区域富集片段之前将分区或样品专用条形码附接至所述拷贝的片段保留了那些靶区域的原始分子背景,从而使得它们可以归属于其原始分区,并因而归属于其原始样品核酸。

[0046] 通常,使用包括基于芯片的方法和基于溶液的捕获方法两者的方法,富集、分离或分隔(即“下拉”)靶基因组区域用于进一步分析,特别是测序。这样的方法利用与目标基因组区或与目标基因组区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)的微阵列固定在表面上,所述捕获探针具有一起覆盖目标区域的序列。基因组DNA是片段化的,并且可以进一步经历加工,例如末端修复以产生平端和/或添加额外的特征,例如通用引发序列。将这些片段与微阵列上的探针杂交。洗去未杂交的片段并洗脱出期望的片段或者在表面上以其它方式处理以用于测序或其它分析,因而保留在表面上的片段群富集了含有目标靶区域(例如,包含与捕获探针中所含的序列互补的序列)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。

[0047] 靶基因组区域捕获的其方法包括基于溶液的方法,其中使基因组DNA片段与寡核苷酸探针杂交。寡核苷酸探针通常称为“饵”。这些饵通常附接至捕获分子,包括但不限于生物素分子。饵与基因组的靶区域(或邻近或接近目标靶区域的区域)互补,使得在应用于基因组DNA片段时,饵与所述片段杂交,然后捕获分子(例如,生物素)用于选择性地拉下目标靶区域(例如,用磁性链霉亲和素珠),从而使所得的片段群富集含有目标靶区域的片段。

[0048] 在需要覆盖全外显子组的靶区域的实例中,使用一起覆盖全外显子组的饵文库来捕获那些靶序列。在这样的实例中,捕获方案可以包括本领域已知的任何方法,包括但不限于由Roche/NimbleGen、Illumina和Agilent生产的任何外显子组捕获方案和试剂盒。

[0049] 用于本文所述的方法和系统中的靶基因组区域的捕获不限于全外显子组,并且可以包括部分外显子组、基因、基因的组、内含子以及内含子与外显子组合中的任一种或组合。捕获这些不同类型的靶区域的程序遵循使用饵来下拉包含目标靶区域的片段的一般方法。饵的设计,特别是与目标靶区域或接近目标靶区域杂交的饵的寡核苷酸探针部分,将部分地取决于要捕获的靶区域的类型。

[0050] 在其中仅需要部分外显子组用于进一步分析的实例中,饵可被设计成捕获外显子组的那部分。在某些实例中,所需外显子组部分的特定身份是已知的,并且饵文库包含与那些识别部分或接近或邻近那些部分的区域互补的寡核苷酸。这样的实例可进一步包括但不限于捕获特定基因和/或基因的组,或已知与特定表型(例如,病症或疾病)相关的外显子组的鉴定部分。在一些实例中,可能需要外显子组或全基因组的某一部分(包括内含子区和外显子区两者)用于进一步分析,但是待捕获的基因组部分的特定序列是未知的。在这样的实施方案中,所使用的饵可以是针对全基因组的文库的子集,并且该子集可以随机选择或通过任何种类的智能设计选择,其中饵文库选择为或者富集与基因组或外显子组的靶子区段互补的探针。

[0051] 对于本文所述的任何方法,可以使用包含与靶区域的全部或一部分互补的寡核苷酸探针的饵捕获靶区域,或者寡核苷酸探针可以与靶区域附近或邻近靶区域的另一区域互补,例如,内含子区域。例如,如图2A中示意性说明的,基因组序列201包含外显子区202和

203。可以通过将饵引导至附近的一个或多个内含子序列(例如,捕获外显子区202时内含子区204和/或205以及捕获外显子区203时内含子区206)来捕获那些外显子区。换句话说,可以通过使用与内含子区204和/或205和206互补的饵来捕获包含外显子区202或203的片段群体。如图2A所示,用作附近外显子区的内含子饵的内含子区可以邻近目标外显子区-即,在内含子区与靶外显子区之间没有间隙。在另一些实例中,用于捕获附近外显子区的内含子区可能足够接近,使得两个区可能在同一片段中,但在外显子区与内含子区之间存在一个或多个核苷酸的缺口(例如,图2A中的202和205)。

[0052] 在一些实例中,采用瓦片式覆盖方法而非将饵设计为靶向基因组的特定区域。在这样的方法中,并非靶向特定外显子区或内含子区,而是将饵设计成与特定范围或距离处的基因组部分互补。例如,饵文库可以设计为沿基因组每5千碱基(kb)覆盖序列,使得将该饵文库应用于片段化基因组样品将仅捕获基因组的某个子集—即,包含在含有与饵互补的序列的片段中的那些区域。如应理解的,可以基于参考序列(例如,人基因组参考序列)设计饵。在另一些实例中,饵的瓦片式覆盖文库设计成捕获基因组的每1、2、5、10、15、20、25、50、100、200、250、500、750、1000或10000千碱基区域。在另一些实例中,饵的瓦片式覆盖文库设计成捕获距离的混合物-该混合物可以是距离的随机混合物,或者智能地设计为使得基因组的特定部分或百分比被捕获。如应理解的,这样的瓦片式覆盖捕获方法将捕获基因组的内含子区和外显子区以进行进一步分析,例如测序。本文所述的任何瓦片式覆盖或其它内含子饵诱方法提供了一种连接来自由长的间插内含子区域较大间距地分隔的外显子的序列信息的方式。

[0053] 在另一些实例中,本文所述的瓦片式覆盖或其它捕获方法将捕获约5%、10%、15%、20%、30%、40%、50%、60%、70%、80%、90%、95%的全基因组。在另一些实例中,本文所述的捕获方法捕获约1-10%、5-20%、10-30%、15-40%、20-50%、25-60%、30-70%、35-80%、40-90%或45-95%的全基因组。

[0054] 在一些实例中,样品制备方法,包括片段化、扩增、分区和以其它方式加工基因组DNA的方法,可导致基因组某些区域的偏差或较低覆盖。通过改变用于捕获基因组靶区域的饵的浓度或基因组位置,可以在本文公开的方法和系统中补偿这样的偏差或降低的覆盖。在一些实例中,可以知晓含有高GC含量或其它结构变化的基因组的某些区域将导致低覆盖-在这种情况下,可以改变饵文库以增加针对这些低覆盖区域的饵浓度-换句话说,可以“掺入”所使用的饵群体以确保在待测序的最终片段群体中获得足够数量的含有那些低覆盖区中的基因组的靶区域的片段。饵的这样的掺入可以在市售的全外显子组试剂盒中进行,使得向现成的外显子组捕获试剂盒中添加针对较低覆盖区域的常规饵文库。另外,可以将饵设计成靶向基因组中非常接近目标区域的区域,但是具有更有利的覆盖,如本文进一步详细讨论的,其实施方案在图2中示意性地示出。

[0055] 在另一些实例中,本发明方法中使用的饵文库是满足如本文进一步描述的一个或多个特征的知情设计的产物。这样的知情设计包括其中饵文库针对信息性单核苷酸多态性(SNP)的情况。本文使用的术语“信息性SNP”是指杂合的SNP。在一些实例中,饵文库设计为含有多个针对基因组样品中含信息性SNP的区域的探针。本文所用的“针对”意指探针含有与包含SNP的序列互补的序列。在另一些实例中,饵文库被设计为包含针对与外显子和内含子的边界相距预定距离的SNP的探针。在基因组的靶区域包括缺少或含有非常少的SNP的区

域的情况下，饵文库包括在预定距离处跨这样的区域瓦片式覆盖和/或与下一最近内含子或外显子内的第一信息性SNP杂交的探针。

[0056] 本文所述的方法和系统的优点在于，捕获的靶区域在捕获之前以这样的方式进行处理，使得即使在捕获靶区域的步骤和进行测序分析之后，仍保留那些靶区域的原始分子背景。如本文进一步详细讨论的，将特定靶区域归属于其原始分子背景(其可以包括它们所源自的原始染色体或染色体区域和/或特定靶区域在全基因组中相对于彼此的位置))的能力提供了一种从使用传统测序技术原本作图差或覆盖差的基因组区域获得序列信息的方式。

[0057] 例如，一些基因具有长的内含子，其太长以至于不能使用通常可用的测序技术，特别是使用具有比长读取技术更优异的精确度的短读取技术实现跨越。然而，在本文所述的方法和系统中，通常通过图1所示并在本文进一步详细描述的标记程序保留目标区域的分子背景。因此，可以在基因组的延伸区域上形成连接。例如，如图2B中示意性说明的，核酸分子207含有被长内含子区域(208)中断的两个外显子(阴影条)。通常使用的测序技术将不能跨越内含子的距离以提供关于两个外显子之间的关系的信息。在本文所述的方法中，个体核酸分子207分配到其自身的离散分区209中，然后片段化，使得不同的片段含有外显子和内含子的不同部分。因为那些片段中的每一个被标记，使得从片段获得的任何序列信息然后可归属于其在其中产生的离散分区，因此每个片段也可归属于其源自的个体核酸分子207。一般来说，并且如本文进一步详细描述的，在片段化和标记之后，来自不同分区的片段被组合在一起。然后可以使用靶向捕获方法来使经历进一步分析(例如测序)的片段群体富集含有目标靶区域的片段。在图2B所示的实例中，所使用的饵将富集该片段群体以仅捕获含有两个外显子之一的一部分的片段，但是不会捕获外显子与内含子之外的区域(例如209和210)。因此，经历测序的最终片段群体将富集含有两个目标外显子的片段。然后可以使用短读取、高精确度测序技术来鉴定该富集的片段群体的序列，并且因为每个片段被标记并且因此可归属于其原始分子背景，即其原始个体核酸分子，短读取序列可提供跨越长的插入内含子长度的信息，以提供有关两个外显子之间的关系的信息。

[0058] 如上所述，本文所述的方法和系统为较长核酸的短读取序列提供了个体分子背景。如本文所使用的，个体分子背景是指特定读取序列之外的序列背景，例如与相邻或近端序列的关系，其不包括在读取序列本身内，因此通常将使得它们不包括在短读取序列的全部或部分内，例如约150个碱基的读取或对于配对读取而言约300个碱基。在一些特别优选的方面，所述方法和系统为短读取序列提供长程序列背景。这样的长程背景包括给定读取序列与彼此相距长于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或甚至长于100kb或更长的读取序列的关系或连接。如应理解的，通过提供长程个体分子背景，还可以得到该个体分子背景中的变体的定相信息，例如，特定长分子上的变体通常通过定义定相。

[0059] 通过提供更长程的个体分子背景，本发明的方法和系统还提供了远远更长的推断的分子背景(在本文中也称为“长虚拟单分子读取”)。如本文所述的序列背景可以包括作图或提供跨越全基因组序列的不同(通常以千碱基尺度)范围的片段连接。这些方法包括将短读取序列作图到单个较长分子或连接的分子的重叠群，以及较长单个分子的大部分的长程测序，例如具有单个分子的连续确定序列，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，将短序列归属于较长核酸(例如，单个长核酸分子或连接的核酸分子或重叠群的集合两者)可以包括将短序列对较长核酸延伸序列作图，以提供高水平序列背景，以及通过这些较长的核酸从短序列提供装配序列。

[0060] 此外，在可以利用与长单个分子相关的长程序列背景时，具有这样的长程序列背景也使得可以推断更长程序列背景。作为一个实例，通过提供上述长程分子背景，可以鉴定来自不同原始分子的长序列中的重叠变体部分，例如，定相变体、易位序列等，从而允许那些分子之间的推断的连接。这样的推断的连接或分子背景在本文中称为“推断的重叠群”。在一些于定相序列的上下文中讨论时的一些情况下，推断的重叠群可以表示通常的定相序列，例如，其中由于重叠的定相变体，可以推断出比单个原始分子实质上长得多的定相重叠群。这些定相重叠群在本文中称为“相区块”。

[0061] 通过以较长的单分子读取(例如，上述“长虚拟单分子读取”)开始，可以得到比使用短读取测序技术或其它定相测序方法原本可获得的相比更长的推断的重叠群或相区块。参见例如公开的美国专利申请号2013-0157870中所公开的。特别地，使用本文描述的方法和系统，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区块长度(其中大于所述N50数值的区块长度总和为所有区块长度总和的50％)。在一些更优选的方面，获得了N50为至少约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的，并且在一些情况下至少约500kb或更大的推断的重叠群或相区块长度。在另一些情况下，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的最大相区块长度。

[0062] 在一个方面，并结合上文和本文后面所述的任何捕获方法，本文所述的方法和系统提供了将样品核酸或其片段区室化、沉积或分区到离散的隔室或分区中(在本文中可互换地称为分区)，其中每个分区保持其自身的内容物与其它分区的内容物相分离。可以预先、随后或同时将独特的标识符(例如，条形码)递送到保持分区室的或分区的样品核酸的分区，以允许稍后将特征(例如，核酸序列信息)归属于包括在特定区室内的样品核酸，并且特别是可以原始沉积到分区中的连续样品核酸的相对较长延伸序列。

[0063] 本文所述方法中使用的样品核酸通常代表待分析的总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、单个染色体、外显子组、扩增子，或多种不同的目标核酸中的任一种。通常对样品核酸进行分区，使得核酸以连续核酸分子的相对较长片段或延伸序列存在于分区中。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，这允许上述更长程分子背景。

[0064] 样品核酸通常也以一定水平进行分区，由此给定分区具有非常低的包括起始样品核酸的两个重叠片段的可能性。这通常通过在分区过程期间以低输入量和/或浓度提供样品核酸来实现。因此，在一些优选的情况下，给定的分区可以包括起始样品核酸的多个长的但不重叠的片段。然后将不同分区中的样品核酸与独特标识符相关联，其中对于任何给定分区，其中所含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，由于分区步骤将样品组分分配到非常小体积的分区或液滴中，应当理解，为了

实现如上所述的期望分配，不需要进行样品的实质性稀释，如在更高体积的方法，例如在管中或在多孔板的孔中所需的。此外，因为本文描述的系统采用如此高水平的条形码多样性，所以可以在更高数量的基因组等同物之间分配不同的条形码，如上所述。特别地，前述多孔板方法(参见例如美国公开申请号2013-0079231和2013-0157870)通常仅用一百至几百个不同的条形码序列操作，并且采用其样品的有限稀释方法以便能够将条形码归属于不同的细胞/核酸。因此，它们通常以远远少于100个细胞操作，这通常提供约1:10的基因组:(条形码类型)比率，当然远高于1:100。另一方面，由于高水平的条形码多样性，例如超过10,000、100,000、500,000等多种条形码类型，本文所述的系统可以按约1:50或更小、1:100或更小、1:1000或更小或甚至更小的基因组:(条码类型)比率操作，同时还允许加载更高数量的基因组(例如，每个测定约大于100个基因组、每个测定大于500个基因组、每个测定1000个基因组或甚至更多)，同时仍然提供每个基因组的远远改善的条形码多样性。

[0065] 通常，在分区步骤之前，将样品与一组可释放地附接到珠上的寡核苷酸标签组合。然后，可以采用本领域已知的和本文描述的方法使该组合导致样品中核酸的加条形码。在一些实例中，采用扩增方法对所得扩增产物添加条形码，在一些实例中，所述扩增产物含有其所源自的全部原始核酸分子的较小区段(片段)。在一些实例中，如Amini等，Nature Genetics 46:1343-1349(2014)(在2014年10月29日提前在线出版)(其通过引用整体并入本文用于所有目的)并且特别是涉及将条形码或其它寡核苷酸标签附接至核酸的所有教导中所述，采用使用转座子的方法。在另一些实例中，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(例如recA)以沿着双链样品核酸产生间隙-然后可以将条形码插入那些间隙中。

[0066] 在其中采用扩增来标记核酸片段的实例中，寡核苷酸标签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核苷酸之间可以是基本上相同的条形码序列，但是在不同分区之间，并且在大多数情况下为不同的条形码序列。第二区域可以是可用于引发分区内的样品内的核酸的N聚物(随机N聚物或设计为靶向特定序列的N聚物)。在一些情况下，当N聚物设计为靶向特定序列时，可以将其设计成靶向特定染色体(例如，染色体1、13、18或21)或染色体的区域，例如外显子组或其它靶区域。在一些情况下，可以将N聚物设计成靶向特定基因或遗传区域，例如与疾病或病症(例如，癌症)相关的基因或区域。在分区内，可以使用第二N聚物进行扩增反应，以在沿着核酸长度的不同位置引发核酸样品。作为扩增的结果，每个分区可以含有附接到相同或接近相同的条形码的核酸的扩增产物，并且可以代表每个分区中的核酸的重叠的较小片段。条形码可以用作表示来源于相同分区的一组核酸的标志物，并且因此潜在地也源自相同的核酸链。扩增后，可以使用测序算法合并、测序和比对核酸。因为较短的读取序列可以根据其相关的条形码序列进行比对并且归属于样品核酸的单一长片段，所述序列上的所有鉴定的变体均可归属于单一原始片段和单一原始染色体。此外，通过在多个长片段上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，如可以跨越长程基因组序列进行分析-例如，跨基因组的表征不足的区域的延伸序列的序列信息的鉴定。这样的信息也可用于鉴定单倍型，其通常是驻留在相同核酸链或不同核酸链上的指定组的遗传变体。也可以以这种方式识别拷贝数变化。

[0067] 所述方法和系统提供了相对于当前核酸测序技术及其相关的样品制备方法的显

著优点。整体样品制备和测序方法倾向于主要鉴定和表征样品中的大多数成分，并且不被设计为鉴定和表征少数成分(其在所提取的样品中占总DNA的小百分比)，例如由一条染色体贡献的遗传物质、或由一个或几个细胞贡献的遗传物质、或在血流中循环的片段化肿瘤细胞DNA分子。所述方法和系统还提供了用于检测较大样品中存在的群体的显著优点。因此，它们对于评估单倍型和拷贝数变化特别有用-本文公开的方法还可用于提供由于在样品制备期间引入的偏差而在核酸靶标群体中表征不良或表现不佳的基因组区域的序列信息。

[0068] 本文公开的加条形码技术的使用针对给定组的遗传标志物赋予提供个体分子背景的独特能力，即将给定组的遗传标志物(与单个标记相反)归属于单个样品核酸分子，并且通过变体配位装配以在多个样品核酸分子中提供更宽或甚至更长程推断的个体分子背景和/或归属于特定染色体。这些遗传标志物可以包括特定遗传基因座，例如变体，例如SNP，或者它们可以包括短序列。此外，加条形码的使用赋予了促进区分从样品提取的总核酸群体的少数成分和主要成分的能力的另外的优点，例如，用于检测和表征血流中的循环肿瘤DNA，并且还在任选的扩增步骤期间减少或消除扩增偏差。此外，以微流体形式实施赋予了与极小的样品体积和低输入量的DNA一起工作的能力，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。

[0069] 如前所述，本文描述的方法和系统的优点是它们可以通过使用普遍可用的短读取测序技术来实现期望的结果。这样的技术具有如下优点：容易获得并且广泛分散在研究界、具有良好表征和高度有效的方案和试剂系统。这些短读取测序技术包括可从例如Illumina，inc.(GXII、NextSeq、MiSeq、HiSeq、X10)、Thermo-Fisher的Ion Torrent分离(IonProton和Ion PGM)、焦磷酸测序方法以及其它获得的那些。

[0070] 特定优点在于本文所述的方法和系统利用这些短读取测序技术，并且利用它们相关的低错误率进行。特别地，本文所述的方法和系统获得了如上所述的期望的个体分子读取长度或背景，但是除配对延伸外具有小于1000bp、小于500bp、小于300bp、小于200bp、小于150bp或甚至更短的单个读取序列；以及这样的单个分子读取长度的测序错误率小于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％。

[0071] II.工作流程综述

[0072] 在一个示例性方面，本公开中描述的方法和系统提供了将个体样品(例如，核酸)沉积或分区到离散分区中，其中每个分区保持其自身内容物与其它分区中的内容物分离。如本文所使用的，分区指可包括多种不同形式的容器或器皿，例如孔、管、微孔或纳米孔、通孔等。然而，在一些优选方面，分区在流体流内是可流动的。这些器皿可以由例如具有围绕内部流体中心或芯的外部屏障的微胶囊或微泡构成，或者它们可以是能够在其基质内夹带和/或保留材料的多孔基质。然而，在优选的方面，这些分区可以包含在非水连续相(例如，油相)内的含水流体的液滴。在例如2013年8月13日提交的美国专利申请号13/966,150中描述了各种不同的器皿。同样，用于在非水连续相或油连续相中产生稳定液滴的乳液系统详细描述于例如公开的美国专利申请号2010-0105112中。在某些情况下，微流体通道网络特别适于产生如本文所述的分区。这样的微流体装置的实例包括在2015年4月9日提交的美国专利申请号14/682,952中详细描述的那些，其全部公开内容通过引用整体并入本文用于所

有目的。在单个细胞的分区中也可以使用替代机制，包括多孔膜，将细胞的含水混合物通过多孔膜挤出成不含水流体。这样的系统通常可得自例如Nanomi，Inc。

[0073] 在乳液中的液滴的情况下，将样品材料(例如，核酸)分区到离散分区中通常可以通过使含水的含样品流入不含水分区流体流(例如，氟化油)也流入的接合部，使得在流动的流分区流体内产生含水液滴，其中这样的液滴包括样品材料。如下文所述，分区(例如，液滴)通常还包括共分区的条形码寡核苷酸。可通过控制系统的各种不同参数来调整任何特定分区内样品材料的相对量，所述参数包括例如含水流中样品的浓度、含水流的流速和/或不含水的流等。本文所述的分区的特征通常在于具有极小体积。例如，在基于液滴的分区的情况下，液滴可以具有如下总体积：小于1000pL、小于900pL、小于800pL、小于700pL、小于600pL、小于500pL、小于400pL、小于300pL、小于200pL、小于100pL、小于50pL、小于20pL、小于10pL或甚至小于1pL。在与珠共分区的情况下，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10％的上述体积。在一些情况下，在用非常少量的起始试剂(例如，输入核酸)进行反应时使用低反应体积分区是特别有利的。用于分析具有低输入核酸的样品的方法和系统示于2015年6月26日提交的美国专利申请号14/752,602中，其全部公开内容通过引用整体并入本文。

[0074] 一旦根据本文所述的方法和系统将样品引入其各自的分区中，分区内的样品核酸通常提供有独特的标识符，使得在表征那些核酸时，它们可归属于其各自来源。因此，样品核酸通常与独特标识符(例如，条形码序列)共分区。在一些特别优选的方面，独特标识符以包含可附接至那些样品的核酸条形码序列的寡核苷酸形式提供。将寡核苷酸分区以使得在给定分区中的寡核苷酸之间，其中所含核酸条形码序列是相同的，但是在不同分区之间，寡核苷酸可以且优选具有不同的条形码序列。在一些优选方面，只有一种核酸条形码序列将与给定的分区相关联，但在一些情况下，可能存在两种或更多种不同的条形码序列。

[0075] 核酸条形码序列通常将在寡核苷酸序列内包括6至约20个或更多个核苷酸。这些核苷酸可以是完全邻接的，即在相邻核苷酸的单个延伸序列中，或者它们可以分成由一个或多个核苷酸分隔开的两个或更多个单独的子序列。通常，分隔开的子序列的长度通常可以为约4至约16个核苷酸。

[0076] 共分区的寡核苷酸还通常包含可用于加工分区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，以扩增来自分区内各个核酸的基因组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，或用于拉下带条形码核酸)或众多种其它潜在功能序列中的任一种。此外，寡核苷酸和相关条形码及其它功能序列以及样品材料的共分区描述于例如均于2014年6月26日提交的美国专利申请号US 14/316,383、US 14/316,398、US 14/316,416、US 14/316,431、14/316,447、14/316,463以及2014年2月7日提交的美国专利申请号14/175,935中，这些专利的全部公开内容通过引用整体并入本文。

[0077] 简而言之，在一个示例性方法中，提供珠，每个珠可包括大量可释放地附接到珠上的上述寡核苷酸，其中附接到特定珠的所有寡核苷酸可包括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以呈现于所用珠群中。通常，珠群可提供可包括至少1000个不同条形码序列、至少10,000个不同条形码序列、至少100,000个不同条形码序列或在一些

情况下至少1,000,000个不同条形码序列的不同条形码序列文库。另外,每个珠通常可以提供有附接的大量寡核苷酸分子。特别地,在个体珠上包括条形码序列的寡核苷酸分子的数目可以是至少约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。

[0078] 在对珠施加特定刺激时,寡核苷酸可以从珠释放。在一些情况下,刺激可以是光刺激,例如通过可释放寡核苷酸的光不稳定键的切割。在一些情况下,可以使用热刺激,其中珠环境温度的升高可导致连接的切割或形成珠的寡核苷酸的其它释放。在一些情况下,可以使用化学刺激以切割寡核苷酸与珠的连接,或者可以导致寡核苷酸从珠的释放。

[0079] 根据本文所述的方法和系统,包括附接的寡核苷酸的珠可以与个体样品共分区,使得单个珠和单个样品包含在个体分区内。在一些情况下,在期望单个珠分隔区时,可能期望控制流体的相对流速,使得平均来说,所述分区在每个分区包含少于一个珠粒,以确保被占据的那些分区主要是单独占用。同样,可能希望控制流速以使较高百分比的分区被占用,例如仅允许小百分比的未占用分区。在一些优选方面,控制流和通道架构以确保所需数量的单独占用的分区、小于一定水平的未占用分区和小于一定水平的多次占用的分区。

[0080] 图3示出了用于对样品核酸进行加条形码和随后测序,特别是用于拷贝数变化或单倍型测定的一种特定示例性方法。首先,可以从源获得包含核酸的样品,300,并且还可以获得一组带条形码珠,310。优选地将珠与含有一个或多个带条形码序列的寡核苷酸以及引物(例如,随机N聚物或其它引物)相连。优选地,条形码序列可以从带条形码珠释放,例如通过切割条形码与珠之间的连接或通过降解下面的珠以释放条形码,或两者的组合。例如,在某些优选方面,带条形码珠可以通过试剂(例如,还原剂)降解或溶解以释放带条形码序列。在该实例中,将少量的包含核酸305、带条形码珠315和任选的其它试剂(例如,还原剂320)的样品组合并进行分区。例如,这样的分区可以涉及将组分引入液滴生成系统,例如微流体装置,325。借助于微流体装置325,可以形成油包水乳液330,其中该乳液包含含有样品核酸305、还原剂320和带条形码珠315的含水液滴。还原剂可以溶解或降解带条形码珠,从而从液滴中的珠释放具有条形码和随机N聚物的寡核苷酸,335。然后,随机N聚物可引发样品核酸的不同区域,从而在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,每个液滴含有一组包含相同条形码序列和不同的随机N聚物序列的寡核苷酸。随后,破坏乳液,345,并且可通过例如扩增方法,350(例如,PCR)添加另外的序列(例如,特别有助于测序方法的序列、另外的条形码等)。然后可以进行测序,355,并且应用算法来解译测序数据,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取和/或识别特定读取序列所属样品。此外,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于其原始分子背景。

[0081] 如上所述,虽然单个珠占据可以是最期望的状态,但是应当理解,通常可存在多次占用分区或未占用分区。图4中示意性地示出了用于共分区包含条形码的寡核苷酸的样品和珠的微流体通道结构的实例。如图所示,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。包含个体样品414的含水流流经通道区段402流向通道接合部412。如本文别处所述,这些样品可以在分区过程之前悬浮在含水流体中。

[0082] 同时,包含携带条形码的珠416的含水流流经通道区段404流向通道接合部412。不

含水分区流体从通道406和408中的每一侧引入通道接合部412，并且合并流流入通道出口410。来自通道区段402和404的两个组合的含水流在通道接合部412内合并，并分区入包括共分区样品414和珠416的液滴418。如前所述，通过控制在通道接合部412处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，可以优化组合和分区以在所产生的分区418内实现珠、样品或两者的期望的占用水平。

[0083] 如将理解的，许多其它试剂可以与样品和珠一起共分区，包括例如化学刺激试剂、核酸延伸试剂、转录试剂和/或扩增试剂，例如聚合酶、逆转录酶、核苷三磷酸或NTP类似物、引物序列和另外的辅因子，例如用于这样的反应中的二价金属离子、连接反应剂例如连接酶和附接序列、染料、标记或其它标记试剂。

[0084] 一旦共分区，设置于珠上的寡核苷酸可用于对分区样品进行加条形码和扩增。使用这些条形码寡核苷酸对样品进行扩增和加条形码的优良方法详细描述于均于2014年6月26日提交的美国专利申请号14/316,383、14/316,398、14/316,416、14/316,431、14/316,447、14/316,463中，这些专利的全部公开内容通过引用整体并入本文。简而言之，在一个方面，寡核苷酸存在于珠上，其与样品共分区并从其珠释放到与样品的分区中。寡核苷酸通常包括与条形码序列一起在其5'端的引物序列。该引物序列可以是随机的或结构化的。随机引物序列通常意在随机引发样品的多个不同区域。结构化引物序列可包括一系列不同结构，包括靶向样品的特定靶区域上游的限定序列以及具有某种部分限定结构的引物，包括但不限于含有一定百分比的特定碱基的引物(例如，一定百分比的GC N聚物)、含有部分或完全简并序列的引物和/或含有根据本文任意描述部分随机结构化和部分结构化的序列的引物。如应理解的，任意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含在寡核苷酸中。

[0085] 一旦释放，寡核苷酸的引物部分即可退火至样品的互补区。延伸反应试剂，例如DNA聚合酶、三磷酸核苷、辅因子(例如Mg2+或Mn2+等)、其也与样品和珠共分区，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核苷酸及其相关的条形码序列。多个引物对样品的不同部分的退火和延伸可导致样品的重叠互补片段的大合并，每个具有指示其所产生自的分区的其自身条形码序列。在一些情况下，这些互补片段本身可用作由分区中存在的寡核苷酸引发的模板，以产生补体的补体，所述补体又包括条形码序列。在一些情况下，配置该复制过程，使得当复制第一互补序列时，其在其末端或其附近产生两个互补序列，以允许形成发夹结构或部分发夹结构，其降低分子作为生成进一步迭代拷贝的基础的能力。图5中示出了其一个实例的示意图。

[0086] 如图所示，包括条形码序列的寡核苷酸与例如样品核酸504一起共分区在例如乳液中的液滴502中。如本文别处所述，寡核苷酸508可以在与样品核酸504共分区的珠506上提供，所述寡核苷酸优选可从珠506中释放，如图A所示。除了一个或多个功能序列(例如序列510、514和516)之外，寡核苷酸508还包括条形码序列512。例如，寡核苷酸508显示为包含条形码序列512以及可用作给定测序系统的附接或固定化序列的序列510，例如用于在Illumina Hiseq或Miseq系统的流动池内附接的P5序列。如所示，寡核苷酸还包括引物序列516，其可包括用于引发样品核酸504的部分复制的随机或靶向N聚物。寡核苷酸508内还包括序列514，序列514可提供测序引发区，例如“read1”或R1引发区，其用于通过测序系统中的合成反应引发聚合酶介导的模板指导的测序。在许多情况下，条形码序列512、固定化序

列510和R1序列514对于附接至给定珠的所有寡核苷酸可以是共同的。对于随机N聚物引物，引物序列516可以变化，或者可以为用于某些靶向应用的给定珠子上的寡核苷酸所共有。

[0087] 基于引物序列516的存在，寡核苷酸能够引发样品核酸，如图B所示，从而允许使用与珠506和样品核酸504共分区的聚合酶和其它延伸试剂将寡核苷酸508和508a延伸。如图C所示，在寡核苷酸延伸之后，随机N聚物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序列或片段，例如片段518和520。虽然包括与样品核酸部分(例如，序列522和524)互补的序列部分，但这些构建体在本文中通常是指包含样品核酸504的片段，具有附接的条形码序列。如应理解的，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，尽管有上述，术语“片段”涵盖原始核酸序列(例如，模板或样品核酸)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产生的那些，例如给定分子序列的实际片段化，例如通过酶促、化学或机械片段化。然而，在一些优选方面，模板或样品核酸序列的片段将表示下面序列或其互补序列的复制部分。

[0088] 然后可以对带条形码核酸片段进行表征，例如通过序列分析，或者它们可以在该过程中进一步扩增，如图D所示。例如，同样从珠306释放的另外的寡核苷酸(例如，寡核苷酸508b)可引发片段518和520。特别地，再次基于寡核苷酸508b中随机N聚物引物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聚物，例如引物序列516)，寡核苷酸与片段518退火，并且延伸以产生与片段518的至少一部分的补体526，片段518包括序列528，其包含样品核酸序列的一部分的副本。寡核苷酸508b的延伸继续，直到它已经通过片段518的寡核苷酸部分508复制。如本文别处所述，并且如图D中所示，寡核苷酸可以被配置成促使聚合酶在所需点的复制中停止，例如在通过包含在片段518内的寡核苷酸508的序列516和514复制之后。如本文所述，这可以通过不同的方法完成，包括例如不同核苷酸和/或核苷酸类似物的掺入不能被所使用的聚合酶加工。例如，这可以包括在序列区512内包含含有尿嘧啶的核苷酸，以防止非尿嘧啶耐受性聚合酶停止该区域的复制。结果，产生了在一端包括全长寡核苷酸508b的片段526，包括条形码序列512、附接序列510、R1引物区514和随机N聚物序列516b。在序列的另一端将包括与第一寡核苷酸508的随机N聚物的补体516’以及显示为序列514’的R1序列的全部或一部分的互补物。然后，R1序列514及其互补序列514’能够杂交在一起以形成部分发夹结构528。如应理解的，因为不同寡核苷酸之间的随机N聚物不同，预期这些序列及其互补序列不参与发夹形成，例如预期作为随机N聚物516的补体的互补序列516’将不会与随机N聚物序列516b互补。对于其它应用，例如靶向引物，情况将不是这样，其中N聚物在给定分区内的寡核苷酸中是常见的。

[0089] 通过形成这些部分发夹结构，使得可以从进一步复制中去除样品序列的第一级副本，例如防止拷贝的迭代拷贝。部分发夹结构还提供用于随后处理所产生的片段(例如，片段526)的有用结构。

[0090] 然后可以按照本文所述将来自多个不同分区的所有片段合并用于在高通量测序仪上测序。因为每个片段根据其原始分区编码，所以该片段的序列可以基于条形码的存在归属于其来源。示意性说明示于图6中。如一个实例所示，源自第一来源600(例如个别染色体、核酸链等)的核酸604和来自不同染色体602的核酸606或核酸链各自与其自身的条形码寡核苷酸组一起分区，如上所述。

[0091] 在每个分区内，然后处理每个核酸604和606以分别提供第一片段的第二片段的重

叠组，例如第二片段组608和610。该处理还提供具有条形码序列的第二片段，其对于源自特定第一片段的每个第二片段而言是相同的。如所示，第二片段组608的条形码序列由“1”表示，而片段组610的条形码序列由“2”表示。可以使用多样化的条形码文库来差异地为大量不同的片段组加条形码。然而，不必需用不同的条形码对来自不同的第一片段的每个第二片段组进行条形码编码。实际上，在许多情况下，可以同时处理多个不同的第一片段以包括相同的条形码序列。多样的条形码文库在本文别处详细描述。

[0092] 然后可以使用例如来自Illumina或Ion Torrent division ofThermo Fisher，Inc.的合成技术的序列合并例如来自片段组608和610的条形码片段用于测序。一旦测序，读取序列612可归属于它们各自的片段集合，例如，如聚合读取614和616中所示，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地基于片段本身的序列。然后将每个片段组的所归属的读取序列装配，以提供每个样品片段的装配序列，例如序列618和620，其进而可进一步归属于它们各自的原始染色体(600和602)。用于组装基因组序列的方法和系统描述于例如2015年6月26日提交的美国专利申请号14/752,773中，其全部公开内容通过引用整体并入本文。

[0093] III.施用方法和系统进行靶向测序

[0094] 在一个方面，本文所述的系统和方法用于从基因组的靶区域获得序列信息。

[0095] 基因组的“靶向”区域(及其任何语法等同物)是指全基因组或通过本文所述的一种或多种方法鉴定为目标和/或所选的基因组的任何一个或多个区域。通过本文描述的方法和系统测序的基因组的靶区域包括但不限于内含子、外显子、基因间区或其任意组合。在某些实例中，本文所述的方法和系统提供关于全外显子组、外显子的一部分、一个或多个选定基因(包括选定的基因的组)、一个或多个内含子以及内含子序列和外显子序列的组合的序列信息。

[0096] 基因组的靶区域还可以包括基因组的某些部分或百分比而不是由序列鉴定的区域。在某些实施方案中，根据本文所述的方法捕获和分析的基因组的靶区域包括位于基因组的每1、2、5、10、15、20、25、50、100、200、250、500、750、1000或10000kb处的基因组部分。在另一些实施方案中，基因组的靶区域包含全基因组的5％、10％、15％、20％、30％、40％、50％、60％、70％、80％、90％、95％。在另一些实施方案中，靶区域包含全基因组的1-10％、5-20％、10-30％、15-40％、20-50％、25-60％、30-70％、35-80％、40-90％或45-95％。

[0097] 通常，捕获基因组的靶区域以用于本领域已知和本文描述的任何测序方法。本文所用的“捕获的”是指用于富集核酸和/或核酸片段群体使得所得群体包含与非目标的基因组区域相比增加百分比的靶区域的任何方法和系统。在另一些实施方案中，富集群体含有至少50％、55％、60％、70％、75％、80％、85％、90％、95％、97％、98％、99％或100％的包含靶区域的核酸和/或核酸片段。

[0098] 捕获方法通常包括基于芯片的方法，其中通过与表面上的捕获分子杂交或其它缔合来捕获靶区域；以及基于溶液的方法，其中使与靶区域(或接近靶区域的区域)互补的寡核苷酸探针(饵)与基因组片段文库杂交。在本文公开的捕获方法中使用的探针通常附接到捕获分子，例如生物素，其可用于“下拉”探针和与其杂交的片段—这些下拉方法包括能将与含有目标靶区域的核酸或核酸片段杂交的饵与不含目标区域的片段分开的任何方法。在其中探针是生物素化的实施方案中，使用磁性链霉亲和素珠来选择性地下拉和富集具有结

合的靶区域的饵。

[0099] 在另一些方面，使用饵文库，其涵盖期望用于进一步研究的所有靶区域。在全外显子组分析的情况下，这样的饵文库因此包括一起覆盖完整外显子组的寡核苷酸探针。在某些实施方案中，仅需要部分外显子组用于进一步分析。在这样的实施方案中，饵被设计成靶向外显子组的子集。这样的设计可以使用本领域已知的方法和算法来实现，并且通常基于参考序列，例如人类基因组。

[0100] 在一些实例中，根据本文所述的方法和系统加工和测序的靶基因组区域是全部或部分外显子组。使用本领域已知的任何方法可捕获这些全部或部分外显子组，这些方法包括但不限于任何Roche/NimbleGen外显子组方案(包括NimbleGen 2.1M人外显子组阵列和NimbleGen SeqCap EZ外显子组文库)、任何Agilent SureSelect产品、任何Illumina外显子组捕获产品(包括TruSeq和Nextera外显子组产品)，以及本领域已知的任何其它产品、方法、系统和方案。

[0101] 在另一些实施方案中，当目标靶区域包含外显子组的全部或部分时，用于捕获那些靶区域的饵可以设计为与那些外显子序列互补。在另一些实施方案中，饵不与外显子序列本身互补，而是与外显子序列附近的序列或两个外显子之间的内含子序列互补。这样的设计在本文中也称为“锚定外显子捕获”或“内含子饵诱”，如本文所述，其意指通过使用与目标外显子组的一个或多个部分接近或临近的一个或多个内含子序列互补的饵来捕获外显子组的所述一个或多个部分的方法。例如，如图2中示意性说明，基因组序列201包含外显子区域202和203。可以通过利用针对附近的一个或多个内含子序列(例如，用于捕获外显子区域202的内含子区域204和/或205和用于捕获外显子区域203的内含子区域206)的饵来捕获那些外显子区域。换句话说，包含外显子区域202或203的片段群体将通过使用与内含子区域204和/或205和206互补的饵被捕获。在一些实施方案中，使用内含子饵诱通过稀疏地饵诱较长的内含子来桥接由长内含子区域分隔开的外显子。在这样的实施方案中，饵不必靶向接近目标外显子区域的内含子区域，而是将饵设计为靶向分隔开特定距离(或距离组)的区域，或设计成在内含子区域瓦片式覆盖特定数目的碱基或许多碱基的组合。下面进一步详细描述这样的实施方案。

[0102] 在一些实施方案中，用于本发明的锚定外显子捕获/内含子饵诱技术的内含子区域与待捕获的外显子区域相邻。在另一些实施方案中，内含子区域与待捕获的外显子区域分隔开约1-50、2-45、3-40、4-35、5-30、6-25、7-20、8-15、9-10、2-20、3-15、4-10、5-30、10-40、15-50、20-75、25-100个核苷酸。在另一些实施方案中，内含子区域与待捕获的外显子区域分隔开约5、10、20、30、40、50、60、70、80、90、100、125、150、175、200、300、400或500个核苷酸。在另一些实施方案中，特别是对于其中内含子区域的稀疏饵诱有用的情况(例如，跨大内含子距离的相连外显子区域的相变体检测或鉴定)，内含子区域与待捕获外显子区域分隔开约数千碱基的距离，例如1-20、2-18、3-16、4-14、5-12、6-10千碱基。由于保留了富集的寡核苷酸群的原始分子背景，这种内含子区的稀疏饵诱允许由长内含子分隔开的外显子区之间的序列信息的连接。

[0103] 在另一些方面，不是将饵设计为靶向基因组的特定区域，而是使用瓦片式覆盖方法。在这种方法中，并非靶向特定外显子或内含子区域，而是将饵设计为与特定范围或距离处的基因组部分互补。例如，可以将饵文库设计为与沿着基因组每5千碱基(kb)定位的序列

杂交，使得将该饵文库应用于片段化的基因组样品将仅捕获基因组的某个子集—即那些包含在含有与饵互补的序列的片段中的区域。如应理解的，可以基于参考序列(例如，人基因组参考序列)设计饵。在另一些实施方案中，将饵的瓦片式覆盖文库设计成捕获基因组的每1、2、5、10、15、20、25、50、100、200、250、500、750、1000或10000千碱基的区域。在一些实例中，该瓦片式覆盖方法具有稀疏捕获内含子区域的效果，因此提供了连接由长内含子区域分隔开的外显子区域的序列信息的方式，因为通过内含子区域的稀疏捕获而捕获的那些外显子区域的原始分子背景得以保留。

[0104] 在另一些实施方案中，将饵设计成以随机或组合的方式瓦片式覆盖基因组—例如，可以使用瓦片式覆盖文库的混合物，其中一些文库每1kb捕获区域，而混合物中的其它文库每100kb捕获区域。在另一些实施方案中，将瓦片式覆盖文库设计成使得在基因组内的位置范围内的饵靶标—例如，饵可以靶向基因组的每1-10、2-5、5-200、10-175、15-150、20-125、30-100、40-75、50-60kb的区域。在另一些实例中，本文所述的瓦片式覆盖或其它捕获方法将捕获全基因组的约5％、10％、15％、20％、30％、40％、50％、60％、70％、80％、90％、95％。如应理解的，这种瓦片式覆盖捕获方法将捕获基因组的内含子区和外显子区两者以进行进一步分析，例如测序。

[0105] 在另一些实施方案中并根据本文所述的任何方法，本发明方法中使用的饵文库是满足如本文进一步描述的一个或多个特征的知情设计的产物。这样的知情设计包括其中饵文库针对信息性单核苷酸多态性(SNP)的情况。如上文所讨论的，本文使用的术语“信息性SNP”是指杂合的SNP。在一些实例中，饵文库设计为含有信息性SNP。本文所用的“针对”意指探针含有与基因组序列的那些区域互补的序列。知情饵设计能够通过允许具有完全覆盖的靶向富集而同时减少所需探针数量(从而降低成本和简化工作流程)而优化靶向测序方法。

[0106] 通常，对于利用知情饵设计的方法，将饵文库设计为包括针对基因组靶区域中的特定序列的饵，此基于在那些区域中存在或不存在信息性SNP和/或那些信息性SNP的位置。图8中提供了用于知情饵设计的一般考虑的示例性说明。基因组801的区域可以包括外显子(802和803)。在一些实例中，信息性SNP 804将位于外显子(802)与相邻内含子之间的边界。在这样的情况下，可以将饵文库设计为包括针对在离边界指定距离处的一个或多个核苷酸(805)的探针。在其中在外显子和相邻内含子(806)之间的边界处没有信息性SNP的另一些实例中，饵库可以设计为包括针对在该边界附近的内含子中的一个或多个位置(807和808)的探针。那些位置将优选地包括信息性SNP，但是也可以根据需要包括其它SNP和/或其它序列。在其中外显子803在外显子内部包含信息性SNP 809但在边界处没有信息性SNP的另一些实例中，可以将饵库设计为包括针对相邻内含子的几个位置810、811和812的探针，这些位置包括信息性和非信息性SNP(以及根据需要的任何其它序列)的混合物。

[0107] 在一些方面，使用一个或多个输入特征来设计探针饵文库，该探针饵文库基于那些输入特征以及各个区域中的图谱质量而针对沿着基因组的移动位置。该设计通常基于信息性SNP之间的间隔，而不是基于内含子和外显子的位置。然而，如应理解的，本文提供的关于基于内含子和外显子位置的饵设计的任何描述也可以与基于信息性SNP的知情饵设计方法组合使用。在知情饵设计中使用的输入特征包括但不限于下列以及下列的任何组合：外显子、内含子、基因间区、信息性SNP的位置，以及重复序列区域(例如富含GC的区域)、着丝粒和样品核酸的长度。

[0108] 为了便于讨论，下文以不同的可能实施方案描述了知情设计探针文库的不同特征。如应理解的，本文讨论的任意探针文库，无论是使用任何知情设计要素还是任何上述讨论的其它类型的设计，都可以单独使用或以任何组合使用。基于目标靶基因组区域以及样品输入和那些目标区域的作图质量来选择使用的设计要素。

[0109] 在一些实施方案中，将探针饵文库设计为包括针对在给定样品中含有信息性SNP的可能性高的的区域的探针。这样的靶标可包括单独的碱基(信息性SNP本身)或靠近或邻近信息性SNP的一个或多个碱基。在另一些实施方案中，探针饵的靶标可以直接邻近信息性SNP，或者以距信息性SNP约1-200、10-190、20-180、30-170、40-160、50-150、60-140、70-130、80-120、90-100个碱基的距离分隔开。

[0110] 在另一些实施方案中，探针饵文库包括针对与核酸分子的平均长度相关的特定密度的区域的探针。例如，可以将探针设计为包括靶序列密度比探针杂交的核酸分子/片段的平均长度稠密x倍的探针，其中x可以是但不限于1、5、10、20、50、75、100、125、150或200。相对于核酸的长度增加探针靶标的密度增加了跨相同物理分子上的基因座连接探针的能力。这样的方法还可以提高连接的区域将包括信息性SNP的可能性，因而进一步提高了探针饵文库附接到基因组的靶区域的能力。

[0111] 在其中(在群体水平)在给定目标区域中信息性SNP的概率不高的情况下，探针靶标的密度也可以增加。在这样的区域中，例如本文所述的那些的瓦片式覆盖方法可用于沿着该区域以周期性间隔引导探针。在某些实施方案中，间隔的密度可以是差异性的，使得缺少信息性SNP的这些区域中的探针间隔的密度比含有信息性SNP区域中的探针间距短1、2、5、10、25、50倍。

[0112] 在另一些实施方案中，将探针饵文库设计为仅考虑基因(包括外显子和内含子)内的信息性SNP分布。该设计方法针对在关键位置捕获足够数量的杂合SNP以从基因的一端连接/定相到另一端。这样的设计方法包括针对组合外显子信息性SNP与一个或多个非外显子SNP的靶标组的饵，使得基因中信息性SNP之间的距离低于上述间隔密度。

[0113] 这样的知情设计方法允许不仅检测基因组的一般靶区域，而且允许基因组结构变化(例如易位和基因融合)的检测和定相。通过确保任何单个基因可以定相，因此可以使用本文所述的方法对绝大多数基因融合事件进行检测和定相。

[0114] 在某些实施方案中并且根据任意上述，将饵文库设计为以约1kb至约2Mb的距离靶向探针。在另一些实施方案中，距离为约1-50、5-45、10-40、15-35、20-30、10-50kb。

[0115] 在另一些实施方案中，被探针饵靶向的核酸片段为约2kb至约250Mb。在另一些实施方案中，片段为约10-1000、20-900、30-800、40-700、50-600、60-500、70-400、80-300、90-200、100-150、50-500、25-300kb。

[0116] 在一些实施方案中，设计探针饵文库，使得约60-95％的探针与含有信息性SNP的序列杂交。在另一些实施方案中，将探针饵文库设计为使得探针文库中约65％-85％、70％-80％、60-90％、80-90％、90-95％、95％-99％的探针与信息性SNP杂交。在另一些实施方案中，将探针文库中至少65％、75％、85％、90％、91％、92％、93％、94％、95％、96％、97％的探针与信息性SNP杂交。如应理解的，待设计为“与信息性SNP杂交”的探针意指这样的探针与包括信息性SNP的序列区域杂交。

[0117] 在另一些实施方案中，将探针饵文库设计成包括针对位于基因组样品靶部分中的

外显子和内含子两者内的信息性SNP的多个探针。

[0118] 在另一些实施方案中，将所述文库设计为使得文库中的大部分探针与间隔开约1-15、5-10、3-6kb的信息性SNP杂交。在另一些实施方案中，将探针文库中的大多数探针进一步设计为与间隔开约1、3、5、10、20、30、50kb的信息性SNP杂交。

[0119] 在另一些实施方案中，设计探针文库内的多个探针，使得对于其中在外显子与内含子之间的边界的5-300、10-50、20-100、30-150或40-200kb内不存在信息性SNP的基因组样品的靶部分，将所述多个探针设计为在来自那些边界的内含子内的信息性SNP处杂交。

[0120] 在另一些实施方案中，设计探针文库内的多个探针，使得对于如下所述基因组样品的靶部分：其中在外显子内存在第一信息性SNP并且所述第一信息性SNP位于与相邻内含子边界相距5-300、10-50、20-100、30-150或40-200kb，并且在相邻内含子内存在第二信息性SNP且所述第二信息性SNP位于与边界相距10-50kb的位置，所述多个探针设计为与第一信息性SNP与第二信息性SNP之间的基因组样品的区域杂交；

[0121] 在另一些实施方案中，设计探针文库中的多个探针，使得对于至少5-300、10-50、20-100、30-150或40-200kb不含信息性SNP的基因组样品的靶部分，将所述多个探针设计为每0.5、1、3或5kb与所述基因组样品的那些靶部分杂交。在另一些实施方案中，将所述多个探针设计成沿着基因组样品的那些靶部分每0.1、0.5、1、1.5、3、5、10、15、20、30、35、40、45、50kb杂交。

[0122] 在另一些实施方案中，设计探针文库内的多个探针，使得对于其中在外显子与内含子之间的5-300、10-50、20-100、30-150或40-200kb的边界内不存在信息性SNP的基因组样品的靶部分，所述多个探针设计成与同外显子-内含子边界的下一最接近的信息性SNP杂交。

[0123] 在另一些实施方案中，探针文库包含设计为以提供跨条形码的连接信息的密度侧接外显子的基因组样品区域杂交的探针。

[0124] 在另一些实施方案中，由探针文库表示的覆盖范围与离散分区中基因组样品的个体核酸片段分子的长度分布成反比，使得含有较高比例的较长个体核酸片段分子的方法使用具有较小覆盖范围的探针文库。

[0125] 在另一些实施方案中，优化探针文库以覆盖基因组样品的靶部分。在另一些实施方案中，对于高地图质量的区域，特别是对于包含信息性SNP的区域，覆盖的密度可以更低，并且对于低地图质量的区域，密度可以进一步更高，以确保跨目标区域提供连接信息。

[0126] 在另一些实施方案中，探针文库具有由基因组样品的一个或多个靶部分的特征而获知的特征，使得对于具有高图谱质量的靶部分，探针文库包括在外显子和内含子的1kb-1Mb边界内与信息性SNP杂交的探针。在这种情况下，探针文库可以进一步包括在外显子和内含子边界的10-500、20-450、30-400、40-350、50-300、60-250、70-200、80-150、90-100kb内与信息性SNP杂交。

[0127] 在另一些实施方案中，探针文库具有由基因组样品的一个或多个靶部分的特征而获知的特征，使得对于其中带条形码片段的长度分布具有高比例片段的靶部分，大于约100、150、200、250kb，探针文库包含与分隔至少50kb的信息性SNP杂交的探针。在这种情况下，探针文库可以进一步包括分隔至少5、10、15、20、25、30、35、40、45、50、75、100、125、150、175、200kb的信息性SNP杂交的探针。

[0128] 在另一些实施方案中，探针文库具有由基因组样品的一个或多个靶部分的特征而获知的特征，使得对于具有低图谱质量的靶部分，探针文库包括在1kb的外显子-内含子边界内与信息性SNP杂交的探针。在这种情况下，探针文库可以进一步包括在外显子-内含子边界的2、5、10、15、20、25、30、35、40、45、50、75、100、125、150、175、200kb内与信息性SNP杂交的探针。在这种情况下，文库将进一步包括与外显子内、内含子内或两者中的信息性SNP杂交的探针。

[0129] 在另一些实施方案中，探针文库具有由基因组样品的一个或多个靶部分的特征而获知的特征，使得对于包含基因间区的靶部分，探针文库包含与间隔开至少1、2、5、10、15、20、25、30、35、40、45、50、75、100kb信息性SNP杂交的探针。

[0130] 在本文所述的捕获方法中使用的饵可以是可用于富集片段群体用于含有基因组靶区域的片段的任何大小或结构。如上所述，通常本发明中使用的饵包含附接到捕获分子(例如，生物素)的寡核苷酸探针。寡核苷酸探针可以与目标靶区域内的序列互补，或者它们可以与靶区域外部的区域互补但足够接近该靶区域，使得“锚定”区域和靶区域两者在同一片段内，使得饵能够通过与该附近区域(例如，侧接内含子)杂交而下拉靶区域。

[0131] 附接到饵的捕获分子可以是可用于从群体中的其它片段分离饵及其杂交伴侣的任何捕获分子。通常，本文所用的饵附接至生物素，然后包含链霉亲和素(包括但不限于磁性链霉亲和素珠)的固体支持物可用于捕获饵及其所杂交的片段。其它捕获分子对可包括但不限于生物素/中性抗生物素蛋白、抗原/抗体或互补寡核苷酸序列。

[0132] 在另一些实施方案中，饵的寡核苷酸探针部分可以是适合于与靶区域或接近靶区域的区域杂交的任何长度。在一些实施方案中，根据本文所述方法使用的饵的寡核苷酸探针部分—即与基因组的靶区域杂交或与靶区域附近的区域杂交的部分—通常具有约10至约150个核苷酸长度(例如，35个核苷酸、50个核苷酸、100个核苷酸)，并且被选择为与目标靶序列特异性杂交。在另一些实施方案中，寡核苷酸探针部分包含长度为约5-10、10-50、20-100、30-90、40-80、50-70个核苷酸的长度。如应理解的，本文所述的任何寡核苷酸探针部分均可包含RNA、DNA、非天然核苷酸例如PNA、LNA等或其任意组合。

[0133] 本文所述的方法和系统的优点在于，捕获的靶区域在捕获之前以这样的方式进行处理，使得即使在捕获靶区域的步骤和进行测序分析之后，仍保留那些靶区域的原始分子背景。将特定靶区域归属于其原始分子背景(其可以包括它们所源自的原始染色体或染色体区域和/或特定靶区域在全基因组中相对于彼此的位置)的能力提供了一种从使用传统测序技术原本作图差或覆盖差的基因组区域获得序列信息的方式。

[0134] 例如，一些基因具有长的内含子，其太长以至于不能使用通常可用的测序技术，特别是使用短读取技术实现跨越。短读取技术通常是优选的测序技术，因为它们具有比长读取技术更优异的精确度。然而，通常使用的短读取技术不能跨越基因组的长区域，因此在基因组中由于例如长串联重复序列长度，高GC含量和含有长内含子的外显子的结构特征而难以表征的区域中可能无法使用这些常规技术获得信息。然而，在本文所述的方法和系统中，通常通过图1所示并在本文进一步详细描述的标记程序保留目标区域的分子背景。因此，可以在基因组的延伸区域上形成连接。例如，如图2B中示意性说明，核酸分子207包含具有长内含子区(208)的两个外显子(阴影条)。在本文所述的方法中，个体核酸分子207分布到其自身的离散分区211中，然后片段化，使得不同的片段含有外显子和内含子的不同部分。因

为那些片段中的每一个被标记，使得从片段获得的任何序列信息然后可归属于其在其中产生的离散分区，因此每个片段也可归属于其源自的个体核酸分子207。

[0135] 一般来说，并且如本文进一步详细描述的，在片段化和标记之后，来自不同分区的片段被组合在一起。然后可以使用靶向捕获方法来使经历进一步分析(例如测序)的片段群体富集含有目标靶区域的片段。在图2B所示的实例中，所使用的饵将富集片段群体以仅捕获含有外显子的一部分的那些片段，但不会捕获外显子与内含子之外的区域(例如209和210)。因此，经历测序的最终片段群体将富集含有外显子部分的片段，即使那些外显子被长内含子区域分隔开。然后可以使用短读取、高精确度测序技术来鉴定该富集的片段群体的序列，并且因为每个片段被标记并且因此可归属于其原始分子背景，即其原始个体核酸分子，短读取序列可以拼接在一起以提供关于外显子之间的关系的信息。在一些实施方案中，用于捕获含有一个或多个外显子的全部或部分的片段的饵与所述一个或多个外显子本身的一个或多个部分互补。在其它实施方案中，饵与插入内含子的一个或多个部分或与外显子区域的3'或5'侧上的外显子相邻或接近的序列互补(这样的饵在本文中也称为“内含子饵”)。在另一些实施方案中，用于捕获含有全部或部分外显子的片段的饵包括与外显子本身互补的饵和内含子饵。

[0136] 保留为测序捕获的靶区域的分子背景的能力还提供允许对基因组的不良表征区域进行测序的优点。如将理解的，人类基因组的显著百分比(至少5-10％，根据例如Altemose等，PLOS Computational Biology，2014年5月15日，第10卷，第5期)保持未装配、未作图和表征不良。参考装配通常将这些缺失区域注释为多兆碱基异染色空位，主要发现在着丝粒附近和近端染色体的短臂上。基因组的这种缺失部分包括使用通常使用的测序技术保持抗精确表征的结构特征。通过提供跨基因组的扩展区域连接信息的能力，本文所述的方法提供了允许对这些表征不良的区域进行测序的方式。

[0137] 在一些实例中，样品制备方法，包括片段化、扩增、分区和以其它方式加工基因组DNA的方法，可导致基因组某些区域的偏差或较低覆盖。通过改变用于捕获基因组靶区域的饵的浓度，可以在本文公开的方法和系统中补偿这样的偏差或降低的覆盖。例如，在一些情况下，可以知晓基因组的某些区域在片段文库被加工后将具有低覆盖，例如可导致基因组中的某些区相对于其它区存在偏差的含有高GC含量或其它结构变化的区域。在这种情况下，可以改变饵文库以增加针对这些低覆盖区域的饵浓度-换句话说，可以“掺入”所使用的饵群体以确保在待测序的最终片段群体中获得足够数量的含有那些低覆盖区中的基因组的靶区域的片段。在一些实施方案中，饵的这样的掺入可以通过常规文库设计进行。在另一些实施方案中，饵的掺入可以在市售的全外显子组试剂盒中进行，使得向现成的外显子组捕获试剂盒中添加针对较低覆盖区域的常规饵文库。

[0138] 本文所述的方法和系统的优点在于，捕获的靶区域在捕获之前以这样的方式进行处理，使得即使在捕获靶区域的步骤和进行测序分析之后，仍保留那些靶区域的原始分子背景。如本文进一步详细讨论的，将特定靶区域归属于其原始分子背景(其可以包括它们所源自的原始染色体或染色体区域和/或特定靶区域在全基因组中相对于彼此的位置))的能力提供了一种从使用传统测序技术原本作图差或覆盖差的基因组区域获得序列信息的方式。

[0139] 例如，一些基因具有长的内含子，其太长以至于不能使用通常可用的测序技术，特

别是使用具有比长读取技术更优异的精确度的短读取技术实现跨越。然而，在本文所述的方法和系统中，通常通过图1所示并在本文进一步详细描述的标记程序保留靶区域的分子背景。因此，可以在基因组的延伸区域上形成连接。例如，如图2B中示意性说明的，核酸分子207含有被长内含子区域中断的外显子(阴影条)。通常使用的测序技术将不能跨越内含子的距离以提供关于两个外显子之间的关系的信息。在本文所述的方法中，个体核酸分子207分配到其自身的离散分区209中，然后片段化，使得不同的片段含有外显子和内含子的不同部分。因为那些片段中的每一个被标记，使得从片段获得的任何序列信息然后可归属于其在其中产生的离散分区，因此每个片段也可归属于其源自的个体核酸分子207。一般来说，并且如本文进一步详细描述的，在片段化和标记之后，来自不同分区的片段被组合在一起。然后可以使用靶向捕获方法来使经历进一步分析(例如测序)的片段群体富集含有目标靶区域的片段。在图2B所示的实例中，所使用的饵将富集片段群体以仅捕获含有外显子之一的一部分的片段，但是不会捕获外显子之外的区域(例如209和210)。因此，经历测序的最终片段群体将富集含有目标外显子的片段。然后可以使用短读取、高精确度测序技术来鉴定该富集的片段群体的序列，并且因为每个片段被标记并且因此可归属于其原始分子背景，即其原始个体核酸分子，短读取序列可拼接在一起以跨越插入内含子的长度(在一些实例中，其长度可为约1、2、5、10或更多千碱基)，从而提供关于两个外显子的连接序列信息。

[0140] 如上所述，本文所述的方法和系统为较长核酸的短读取序列提供了个体分子背景。如本文所使用的，个体分子背景是指特定读取序列之外的序列背景，例如与相邻或近端序列的关系，其不包括在读取序列本身内，因此通常将使得它们不包括在短读取序列的全部或部分内，例如约150个碱基的读取或对于配对读取而言约300个碱基。在一些特别优选的方面，所述方法和系统为短读取序列提供长程序列背景。这样的长程背景包括给定读取序列与彼此相距长于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或甚至长于100kb或更长的读取序列的关系或连接。通过提供长程个体分子背景，本发明的方法和系统还提供了远远更长的推断的分子背景。如本文所述的序列背景可以包括较低分辨率背景，例如从将短读取序列作图到单个较长分子或连接的分子的重叠群以及较高分辨率序列背景，例如来自大部分更长的个体分子的大部分长程测序，例如具有个体分子的连续确定序列，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，将短序列归属于较长核酸(例如，单个长核酸分子或连接的核酸分子或重叠群的集合两者)可以包括将短序列对较长核酸延伸序列作图，以提供高水平序列背景，以及通过这些较长的核酸从短序列提供装配序列。

[0141] IV.样品

[0142] 如将理解的，本文所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靶向序列信息。这样的基因组材料可以从取自患者的样品获得。在本文讨论的方法和系统中使用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、寡核苷酸、循环无细胞核酸、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)、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、无细胞的胎儿DNA(cffDNA)、

mRNA、rRNA、tRNA、nRNA、siRNA、snRNA、snoRNA、scaRNA、microRNA、dsRNA、病毒RNA等。总之，所使用的样品可以根据具体的加工需要而变化。

[0143] 包含核酸的任何物质可以是样品的来源。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、汗液、血清、精液、阴道液、胃液和消化液、脊髓液、胎盘液、腔液、眼液、血清、乳汁、淋巴液或其组合。所述物质可以是实体，例如生物组织。所述物质可以包括正常健康组织、患病组织或健康组合与患病组织的混合体。在一些情况下，所述物质可以包括肿瘤。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包括：纤维肉瘤、粘液肉瘤、脂肪肉瘤、软骨肉瘤、骨源性肉瘤、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、维尔姆斯氏瘤、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、神经胶质瘤、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、成视网膜细胞瘤或其组合。所述物质可以与各种类型的器官相关。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、食管、胃或其组合。在一些情况下，所述物质可包含多种细胞，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、从人类微生物组样品分离的细菌细胞等。在一些情况下，所述物质可包含细胞的内容物，例如，单细胞的内容物或多个细胞的内容物。用于分析个体细胞的方法和系统提供于例如2015年6月26日提交的美国专利申请号14/752,641中，该专利的全部公开内容在此通过引用整体并入，特别是涉及分析来自个体细胞的核酸的所有教导。

[0144] 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、鸟类、鼠、兔、昆虫、蛞蝓、微生物、细菌、寄生虫或鱼类。在一些情况下，受试者可以是患有、疑似患有疾病或病症或处于产生疾病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，受试者可以是处于怀有具有某些出生缺陷的婴儿的风险的孕妇。

[0145] 可以通过本领域已知的任何手段从受试者获得样品。例如，可以通过进入循环系统(例如，通过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、痰液、尿液、粪便等)、手术(例如，活检)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)或移液从受试者获得样品。

[0146] 虽然本文已经示出并描述了本发明的优选实施方案，但是对于本领域技术人员来说显而易见的是，这样的实施方案仅以示例的方式提供。在不脱离本发明的情况下，本领域技术人员将会想到许多变化、改变和替换。应当理解，在实施本发明时可以采用本文所述的本发明实施方案的各种替代方案。意图是以下权利要求限定本发明的范围，并且由此涵盖这些权利要求及其等同变化形式的范围内的方法和结构。

实施例

[0147] 实施例1:全外显子组捕获和测序:NA12878

[0148] 使用Blue PippinDNA大小筛分系统对来自NA12878人细胞系的基因组DNA进行基于大小的片段分离,以回收长度大于或等于约10kb的片段。然后使用微流体分区系统(参见例如2015年4月9日提交并通过引用整体并入本文用于所有目的的美国专利申请号14/682,952)将所选大小的样品核酸与条形码珠在氟化油连续相内的含水液滴中共分区,其中含水液滴还包括dNTP、热稳定的DNA聚合酶及用于在液滴内进行扩增的其它试剂以及用于从珠中释放条形码寡核苷酸的DTT。将其对于1ng的总输入DNA和2ng的总输入DNA两者重复。获得条形码珠作为呈现条形码多样性(具有超过700,000种不同条形码序列)的储备文库的子集。含有寡核苷酸的条形码包括其它序列组分并具有以下一般结构:

[0149] 珠-P5-BC-R1-Nmer

[0150] 其中P5和R1分别指Illumina附接序列和Read1引物序列,BC表示寡核苷酸的条形码部分,且Nmer表示用于引发模板核酸的随机10碱基N-mer引发序列。参见例如于2014年6月26日提交的美国专利申请号14/316,383,其全部公开内容通过引用整体并入本文用于所有目的。

[0151] 珠溶解后,将液滴热循环以容许条形码寡核苷酸针对每个小滴内的样品核酸的模板的引物延伸。除了上文所述的其它包括的序列之外,这还产生了包含代表原始分区的条形码序列的样品核酸的扩增拷贝片段。

[0152] 对拷贝片段进行条形码标记后,破坏包含扩增拷贝片段的液滴的乳液,并且通过另外的扩增步骤将另外的测序仪所需组分(例如,read2引物序列和P7附接序列)添加到拷贝片段,其将这些序列附接到拷贝片段的另一端。然后使用Agilent SureSelect外显子组捕获试剂盒对带条形码的DNA进行杂交捕获。

[0153] 下表提供了NA12878基因组的靶向统计学:

[0154]

样品	中值插入物大小	靶标上片段%	靶标上碱基%
版本1.A	258	81%	51%
版本1.B	224	81%	55%
版本1.C	165	81%	63%

[0155] 上面列出的三种不同版本代表在第二接头附接步骤之前带条形码片段的三种不同剪切长度。

[0156] 实施例2:全外显子组捕获和测序:NA19701和NA19661

[0157] 根据上文实施例1中所述的方法制备来自NA19701和NA19661细胞系的基因组DNA。来自这两种细胞系的数据(包括定相数据)提供在下表中:

[0158]

	NA19661	NA19701
相区块的N50	29,535	83,953
相区块的N90	8,595	25,684
相区块的平均值	5,968	21,128
相区块的中值	0	76.5
相区块的最长值	209,323	504,140
定相的基因的分率	0.719	0.841
完全定相的基因的分率	0.679	0.778
定相的snp的分率	0.869	0.832
带条形码和等位基固两者的snp的分率	0.328	0.351
基因中正确的snp的概率	0.906	0.927
基因中定相的snp的概率	0.807	0.889
短snp交换错误率	0.013	0.013
长snp交换错误率	0.012	0.013

[0159] 表3.定相量度。如图1所示,NA19701的片段长度比NA19661长得多,从而产生好得多的定相性能。

[0160] 本说明书在当前描述的技术的示例方面中提供了方法、系统和/或结构及其用途的完整描述。虽然上文已经以某种程度的特殊性或参考一个或多个个别方面描述了本技术的各个方面,但是本领域技术人员可以对所公开的方面进行多种改变,而不脱离本技术的精神或范围。由于可以在不脱离当前描述的技术的精神和范围的情况下做出许多方面,因此适当的范围在所附权利要求中。因此设想了其它方面。此外,应当理解,可以以任何顺序进行任何操作,除非另有明确要求或者特定顺序由权利要求语言固有地必需的。旨在包含在上述描述中和在附图中示出的所有内容均应被解释为仅仅是对特定方面的说明,而不限于所示的实施方案。除非上下文另有说明或明确阐述,否则本文提供的任何浓度值通常以混合物值或百分比的形式给出,而不考虑在加入混合物的特定组分时或之后发生的任何转化。在未明确并入本文的程度,本公开中提及的所有公开的参考文献和专利文献通过引用整体并入本文用于所有目的。在不脱离所附权利要求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的改变。

[0161] 本发明涉及以下实施方案:

[0162] 1.一种用于对基因组的一个或多个选定部分进行测序的方法,所述方法包括:

[0163] (a)提供起始基因组材料;

[0164] (b)将来自所述起始基因组材料的个体核酸分子分配到离散分区中,使得每个离散分区包含第一个体核酸分子;

[0165] (c)将所述离散分区中的所述个体核酸分子片段化以形成多个片段,其中每个所述片段还包含条形码,并且其中给定离散分区内的片段各自包含共同条形码,由此将每个片段与其所源自的个体核酸分子相关联;

[0166] (d)提供富集包含所述基因组的所述一个或多个选定部分的至少一部分的片段的群体;

[0167] (e)从所述群体获得序列信息,从而对基因组的一个或多个选定部分进行测序。

[0168] 2.根据实施方案1所述的方法,其中所述提供步骤(d)包括:

[0169] (i)将与所述基因组的所述一个或多个选定部分中或附近区域互补的探针与所述片段杂交以形成探针-片段复合物;

[0170] (ii)将探针-片段复合物捕获到固体支持物的表面;

[0171] 从而使所述群体富集包含所述基因组的所述一个或多个选定部分的至少一部分的片段。

[0172] 3.根据实施方案2所述的方法,其中所述固体支持物包含珠。

[0173] 4.根据实施方案2至3所述的方法,其中所述探针包含结合部分,并且所述表面包含捕获部分,并且其中所述探针-片段复合物通过所述结合部分与所述捕获部分之间的反应而捕获在所述表面上。

[0174] 5.根据实施方案4所述的方法,其中所述捕获部分包含链霉亲和素,而所述结合部分包含生物素。

[0175] 6.根据实施方案4所述的方法,其中所述捕获部分包含链霉亲和素磁珠,而所述结合部分包含生物素化RNA文库饵。

[0176] 7.根据实施方案4所述的方法,其中所述捕获部分针对选自以下的成员:全外显子组或部分外显子组捕获、组捕获、靶向外显子捕获、锚定外显子组捕获和瓦片式覆盖基因组区域捕获。

[0177] 8.根据实施方案1至7所述的方法,其中在所述获得步骤(e)之前,扩增所述片段以形成扩增产物。

[0178] 9.根据实施方案8所述的方法,其中所述扩增产物能够形成部分或完整的发夹结构。

[0179] 10.根据实施方案1至9所述的方法,其中所述获得步骤(e)包括选自以下的测序反应:短读取长度测序反应和长读取长度测序反应。

[0180] 11.根据实施方案10所述的方法,其中所述测序反应是短读取、高精确度测序反应。

[0181] 12.根据实施方案1至11所述的方法,其中所述获得步骤(e)提供关于所述起始基因组材料的小于90%的序列信息。

[0182] 13.根据实施方案1至11所述的方法,其中所述获得步骤(e)提供关于所述起始基因组材料的小于75%的序列信息。

[0183] 14.根据实施方案1至11所述的方法,其中所述获得步骤(e)提供关于所述起始基因组材料的小于50%的序列信息。

[0184] 15.根据实施方案1至14所述的方法,其中所述方法还包括基于所述分离的片段的重叠序列在推断的重叠群中连接两个或更多个所述个体核酸分子,其中所述推断的重叠群包含至少10kb的长度N50。

[0185] 16.根据实施方案15所述的方法,其中所述推断的重叠群包含至少20kb的长度N50。

[0186] 17.根据实施方案15所述的方法,其中所述推断的重叠群包含至少40kb的长度N50。

[0187] 18.根据实施方案15所述的方法,其中所述推断的重叠群包含至少50kb的长度

N50。

[0188] 19.根据实施方案15所述的方法，其中所述推断的重叠群包含至少100kb的长度N50。

[0189] 20.根据实施方案15所述的方法，其中所述推断的重叠群包含至少200kb的长度N50。

[0190] 21.根据实施方案1至14所述的方法，其中所述方法还包括基于所述分离的片段的序列内的重叠定相变体，在相区块中连接两个或更多个所述个体核酸分子，其中所述相区块包含至少10kb的长度N50。

[0191] 22.根据实施方案21所述的方法，其中所述相区块包含至少20kb的长度N50。

[0192] 23.根据实施方案21所述的方法，其中所述相区块包含至少40kb的长度N50。

[0193] 24.根据实施方案18所述的方法，其中所述相区块包含至少50kb的长度N50。

[0194] 25.根据实施方案21所述的方法，其中所述相区块包含至少100kb的长度N50。

[0195] 26.根据实施方案21所述的方法，其中所述相区块包含至少200kb的长度N50。

[0196] 27.根据实施方案1至26所述的方法，其中所述基因组的选定部分包含外显子组。

[0197] 28.根据实施方案1至26所述的方法，其中每个离散分区中的所述个体核酸分子包含来自单细胞的基因组DNA。

[0198] 29.根据实施方案1至28所述的方法，其中每个离散分区包含来自不同染色体的基因组DNA。

[0199] 30.根据实施方案1至29所述的方法，其中所述离散分区包含乳液中的液滴。

[0200] 31.根据实施方案1至30所述的方法，其中附接于所述片段的所述条形码来自至少700,000个条形码的文库。

[0201] 32.根据实施方案1至31所述的方法，其中所述条形码还包含另外的序列区段。

[0202] 33.根据实施方案32所述的方法，其中所述另外的序列区段包含选自以下的一个或多个成员：引物、附接序列、随机n聚寡核苷酸、包含尿嘧啶核碱基的寡核苷酸。

[0203] 34.一种由基因组样品的一个或多个靶部分获得序列信息的方法，所述方法包括：

[0204] (a)在离散分区中提供所述基因组样品的个体第一核酸片段分子；

[0205] (b)将所述离散分区内的所述个体第一核酸片段分子片段化以由所述个体第一核酸片段分子中的每一个产生多个第二片段；

[0206] (c)将共同条形码序列附接至离散分区内的所述多个第二片段，使得所述多个第二片段中的每一个均可归属于其中包含它们的所述离散分区；

[0207] (d)将针对所述基因组样品的所述一个或多个靶部分的探针文库应用于所述第二片段；

[0208] (e)进行测序反应以鉴定与所述探针文库杂交的所述多个第二片段的序列，从而由所述基因组样品的所述一个或多个靶部分获得序列信息。

[0209] 35.根据实施方案34所述的方法，其中将所述探针文库附接至结合部分，并且其中在所述进行步骤(e)之前，通过所述结合部分与所述捕获部分之间的反应将所述第二片段捕获在包含捕获部分的表面上。

[0210] 36.根据实施方案35所述的方法，其中在所述进行步骤(e)之前，在所述第二片段捕获在所述表面上之前或之后扩增所述第二片段。

[0211] 37.根据实施方案35至36所述的方法,其中所述结合部分包含生物素,而所述捕获部分包含链霉亲和素。

[0212] 38.根据实施方案34至37所述的方法,其中所述测序反应是选自以下的成员:短读取长度测序反应和长读取长度测序反应。

[0213] 39.根据实施方案34至38所述的方法,其中所述测序反应是短读取、高精确度测序反应。

[0214] 40.根据实施方案34至39所述的方法,其中所述方法还包括基于所述多个第二片段的重叠序列在推断的重叠群中连接两个或更多个所述个体片段分子,其中所述推断的重叠群包含至少10kb的长度N50。

[0215] 41.根据实施方案40所述的方法,其中所述推断的重叠群包含至少20kb的长度N50。

[0216] 42.根据实施方案40所述的方法,其中所述推断的重叠群包含至少40kb的长度N50。

[0217] 43.根据实施方案40所述的方法,其中所述推断的重叠群包含至少50kb的长度N50。

[0218] 44.根据实施方案40所述的方法,其中所述推断的重叠群包含至少100kb的长度N50。

[0219] 45.根据实施方案40所述的方法,其中所述推断的重叠群包含至少200kb的长度N50。

[0220] 46.根据实施方案34至39所述的方法,其中所述方法还包括基于所述多个第二片段的序列内的重叠定相变体,在相区块中连接所述多个个体核酸片段分子中的两个或更多个,其中所述相区块包含至少10kb的长度N50。

[0221] 47.根据实施方案46所述的方法,其中所述相区块包含至少20kb的长度N50。

[0222] 48.根据实施方案46所述的方法,其中所述相区块包含至少40kb的长度N50。

[0223] 49.根据实施方案46所述的方法,其中所述相区块包含至少50kb的长度N50。

[0224] 50.根据实施方案46所述的方法,其中所述相区块包含至少100kb的长度N50。

[0225] 51.根据实施方案46所述的方法,其中所述相区块包含至少200kb的长度N50。

[0226] 52.根据实施方案34至51所述的方法,其中所述基因组样品的所述靶部分包含外显子组。

[0227] 53.根据实施方案34至51所述的方法,其中每个离散分区中的所述基因组样品包含来自单细胞的基因组DNA。

[0228] 54.根据实施方案34至51所述的方法,其中每个离散分区包含来自不同染色体的基因组DNA。

[0229] 55.根据实施方案34至54所述的方法,其中所述离散分区包含乳液中的液滴。

[0230] 56.根据实施方案34至55所述的方法,其中附接至所述第二片段的所述条形码序列来自至少700,000个条形码的文库。

[0231] 57.根据实施方案34至56所述的方法,其中所述条形码还包含另外的序列区段。

[0232] 58.根据实施方案57所述的方法,其中所述另外的序列区段包含选自以下的一个或多个成员:引物、附接序列、随机n聚寡核苷酸、包含尿嘧啶核碱基的寡核苷酸。

[0233] 59.根据实施方案36至58所述的方法,其中扩增所述第二片段以使得所得扩增产物能够形成部分或完整的发夹结构。

[0234] 60.一种用于从基因组样品的一个或多个靶部分获得序列信息同时保留分子背景的方法,所述方法包括:

[0235] (a)提供起始基因组材料;

[0236] (b)将来自所述起始基因组材料的个体核酸分子分配到离散分区中,使得每个离散分区包含第一个体核酸分子;

[0237] (c)将所述离散分区中的所述第一个体核酸分子片段化以形成多个片段;

[0238] (d)提供富集包含所述基因组的一个或多个选定部分的至少一部分的片段的群体;

[0239] (e)从所述群体获得序列信息,由此对所述基因组样品的一个或多个靶部分进行测序同时保留分子背景。

[0240] 61.根据实施方案60所述的方法,其中在所述获得步骤(e)之前,用条形码标记所述多个片段,以将每个片段与其形成于其中的所述离散分区相关联。

[0241] 62.根据实施方案60所述的方法,其中将步骤(b)中的所述个体核酸分子分配以使得保持每个第一个体核酸分子的分子背景。

[0242] 63.一种从基因组样品的一个或多个靶部分获得序列信息的方法,所述方法包括:

[0243] (a)在离散分区中提供所述基因组样品的个体核酸分子;

[0244] (b)将所述离散分区中的所述个体核酸分子片段化以形成多个片段,其中每个所述片段还包含条形码,并且其中给定离散分区内的片段各自包含共同条形码,由此将每个片段与其所源自的个体核酸分子相关联;

[0245] (c)将针对所述基因组样品的所述一个或多个靶部分的探针文库应用于所述多个片段,其中将所述探针文库中的至少大多数探针设计为与信息性单核苷酸多态性(SNP)杂交;

[0246] (d)进行测序反应以鉴定与所述探针文库杂交的所述多个片段的序列,从而从所述基因组样品的所述一个或多个靶部分获得序列信息。

[0247] 64.根据实施方案63所述的方法,其中将所述探针文库中的约80%至99%的探针设计为与信息性SNP杂交。

[0248] 65.根据实施方案63所述的方法,其中将所述探针文库中的约65%至85%的探针设计为与信息性SNP杂交。

[0249] 66.根据实施方案63所述的方法,其中将所述探针文库中的约70%至80%的探针设计为与信息性SNP杂交。

[0250] 67.根据实施方案63所述的方法,其中将所述探针文库中的至少65%的探针设计为与信息性SNP杂交。

[0251] 68.根据实施方案63所述的方法,其中将所述探针文库中的至少75%的探针设计为与信息性SNP杂交。

[0252] 69.根据实施方案63所述的方法,其中将所述探针文库中的至少85%的探针设计为与信息性SNP杂交。

[0253] 70.根据实施方案63所述的方法,其中将所述探针文库中的至少90%的探针设计

为与信息性SNP杂交。

[0254] 71.根据实施方案63至70所述的方法,其中所述信息性SNP位于所述基因组样品的所述靶部分中的外显子和内含子两者内。

[0255] 72.根据实施方案63至70所述的方法,其中将所述探针文库中的所述大多数探针进一步设计为与间隔开约1千碱基至约15千碱基(kb)的信息性SNP杂交。

[0256] 73.根据实施方案63至70所述的方法,其中将所述探针文库中的所述大多数探针进一步设计为与间隔开约5kb至约10kb的信息性SNP杂交。

[0257] 74.根据实施方案63至70所述的方法,其中将所述探针文库中的所述大多数探针进一步设计为与间隔开约3kb至约6kb的信息性SNP杂交。

[0258] 75.根据实施方案63至70所述的方法,其中将所述探针文库中的所述大多数探针进一步设计为与间隔开约1kb的信息性SNP杂交。

[0259] 76.根据实施方案63至70所述的方法,其中将所述探针文库中的所述大多数探针进一步设计为与间隔开约3kb的信息性SNP杂交。

[0260] 77.根据实施方案63至70所述的方法,其中将所述探针文库中的所述大多数探针进一步设计为与间隔开约10kb的信息性SNP杂交。

[0261] 78.根据实施方案63至77所述的方法,其中将所述探针文库内的多个探针进一步设计为以任意组合满足以下条件中的一个或多个:

[0262] (i)对于其中在外显子与内含子之间边界的10至50kb内没有信息性SNP的所述基因组样品的靶部分,将所述多个探针设计为在来自那些边界的内含子内的信息性SNP处杂交;

[0263] (ii)对于其中在外显子内存在第一信息性SNP且所述第一信息性SNP位于与相邻内含子的边界相距10至50kb处,并且在相邻内含子内存在第二信息性SNP且所述第二信息性SNP位于距所述边界10至50kb处的所述基因组样品的靶部分,将所述多个探针设计为与所述第一信息性SNP和所述第二信息性SNP之间的基因组样品的区域杂交;

[0264] (iii)对于至少10至50kb不包含信息性SNP的所述基因组样品的靶部分,将所述多个探针设计为每0.5、1、3或5kb与所述基因组样品的那些靶部分杂交;

[0265] (iv)对于其中在外显子与内含子之间边界的10至50kb内没有信息性SNP的所述基因组样品的靶部分,将所述多个探针设计为与同所述外显子-内含子边界下一最接近的信息性SNP杂交。

[0266] 79.根据实施方案63至78所述的方法,其中所述探针文库包含设计为以提供跨条形码的连接信息的密度与侧接外显子的所述基因组样品区域杂交的探针。

[0267] 80.根据实施方案63至79所述的方法,其中由所述探针文库表示的覆盖范围与所述离散分区中的所述基因组样品的所述个体核酸片段分子的长度分布成反比,使得包含较高比例的较长个体核酸片段分子的方法使用具有较小覆盖范围的探针文库。

[0268] 81.根据实施方案63至80所述的方法,其中所述探针文库针对所述基因组样品的所述靶部分的覆盖进行优化,并且其中所述基因组样品的所述靶部分包含高图谱质量区。

[0269] 82.根据实施方案63至80所述的方法,其中所述探针文库具有由基因组样品的所述一个或多个靶部分的特征而获知的特征,使得:

[0270] (i)对于具有高图谱质量的靶部分,所述探针文库包含与外显子和内含子的边界

的1kb至1兆碱基(Mb)内的信息性SNP杂交的探针；

[0271] (ii)对于其中所述带条形码片段的长度分布具有高比例的长于约250kb的片段的靶部分，所述探针文库包含与分隔至少50kb的信息性SNP杂交的探针；

[0272] (iii)对于具有低图谱质量的靶部分，所述探针文库包含与外显子-内含子边界1kb内的信息性SNP杂交的探针和与外显子内和内含子内的信息性SNP杂交的探针；

[0273] (iv)对于包含基因间区的靶部分，所述探针文库包含与以至少2kb的距离间隔开的信息性SNP杂交的探针。

[0274] 83.根据实施方案63至82所述的方法，其中步骤(d)中获得的所述序列信息包含关于以下中的一个或多个成员的信息：基因融合、拷贝数变化、插入和缺失。

[0275] 84.一种从基因组样品的一个或多个靶部分获得序列信息的方法，所述方法包括：

[0276] (a)提供所述基因组样品的个体核酸分子；

[0277] (b)将所述个体核酸分子片段化以形成多个片段，其中每个所述片段还包含条形码，并且其中来自同一个体核酸分子的片段包含共同条形码，由此将每个片段与其所源自的个体核酸分子相关联；

[0278] (c)使所述多个片段富集含有所述基因组样品的所述一个或多个靶部分的片段；

[0279] (d)进行测序反应以鉴定所述富集多个片段的序列，从而从所述基因组样品的所述一个或多个靶部分获得序列信息。

[0280] 85.根据实施方案84所述的方法，其中在所述片段化步骤(b)之前将条形码添加到所述个体核酸分子。

[0281] 86.根据实施方案85所述的方法，其中使用转座子将所述条形码添加到所述个体核酸分子。

[0282] 87.根据实施方案84所述的方法，其中在所述片段化的同时添加条形码。

[0283] 88.根据实施方案87所述的方法，其中所述片段化包括扩增步骤。

[0284] 89.根据实施方案84至88所述的方法，其中所述富集步骤(c)包括应用针对所述基因组样品的所述一个或多个靶部分的探针文库。

[0285] 90.根据实施方案89所述的方法，其中将所述探针文库附接到结合部分，并且其中在所述进行步骤(d)之前，通过所述结合部分与所述捕获部分之间的反应捕获所述片段。

[0286] 91.根据实施方案90所述的方法，其中所述结合部分与所述捕获部分之间的所述反应将所述片段固定在表面上。

[0287] 92.根据实施方案90至91所述的方法，其中所述结合部分包含生物素，而所述捕获部分包含链霉亲和素。

[0288] 93.根据实施方案84至92所述的方法，其中所述测序反应是选自以下的成员：短读取长度测序反应和长读取长度测序反应。

[0289] 94.根据实施方案84至92所述的方法，其中所述测序反应是短读取、高精确度测序反应。

[0290] 95.根据实施方案84至94所述的方法，其中扩增所述片段以使得所得扩增产物能够形成部分或完整的发夹结构。

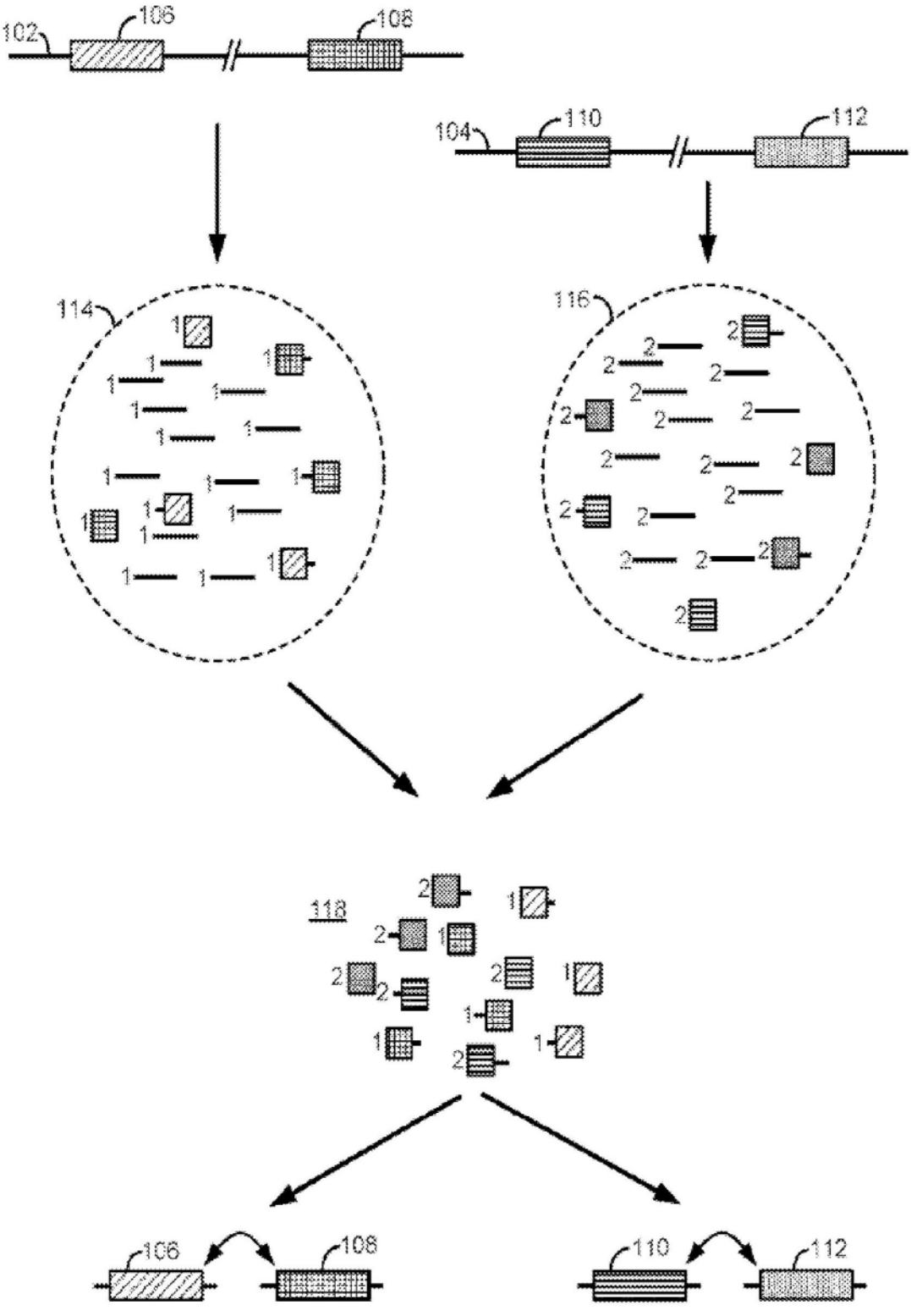

图1

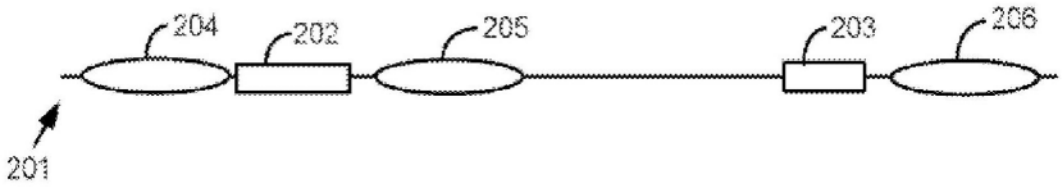

图2A

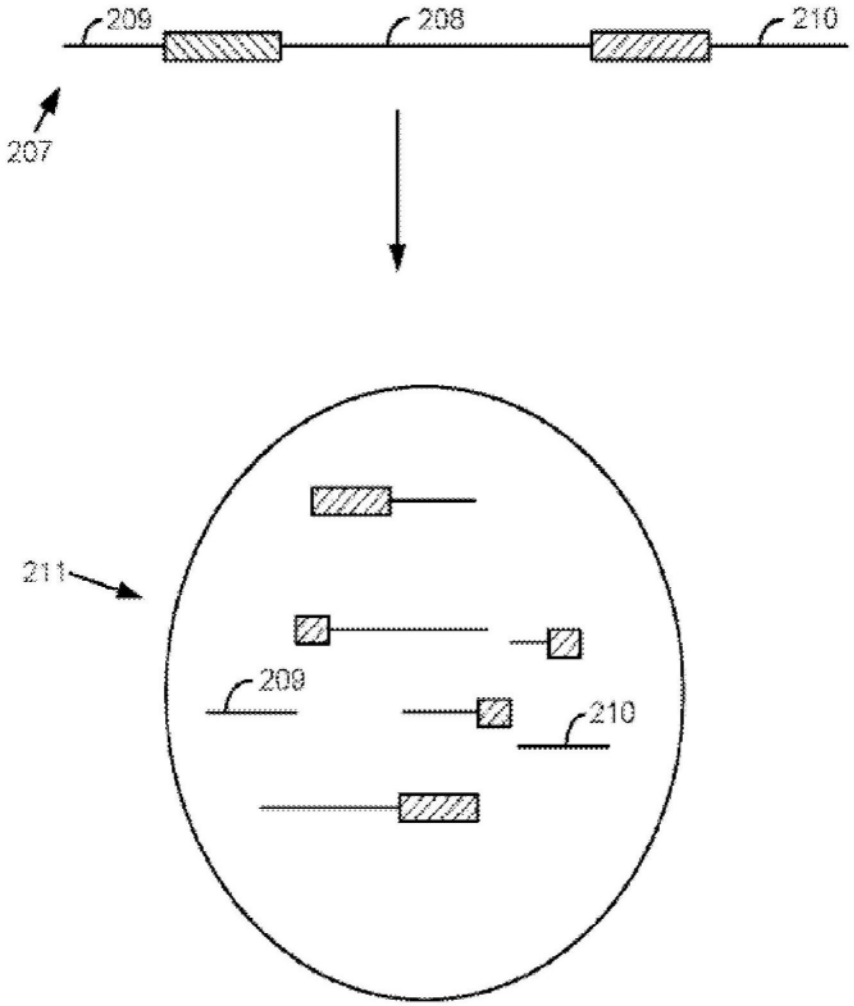

图2B

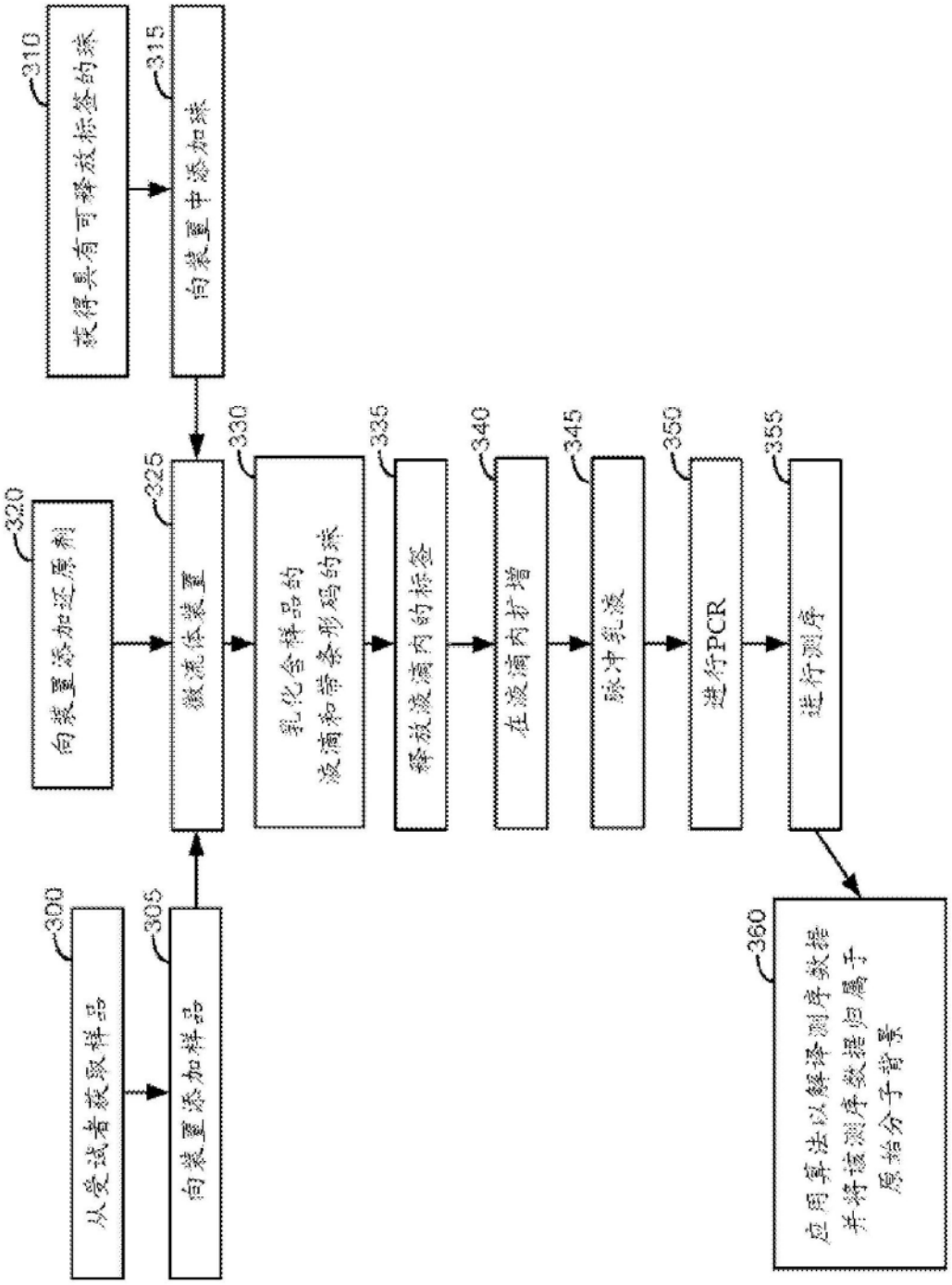
310
获得具有可释放标签的珠
315
向装置中添加珠
320
向装置添加还原剂
325
微流体装置
330
乳化含样品的液滴和带条形码的珠
335
释放液滴内的标签
340
在液滴内扩增
345
破坏乳液
350
进行PCR
355
进行测序
300
从受试者获取样品
305
向装置添加样品
360
应用算法以解译测序数据并将该测序数据归属于原始分子群集

图3

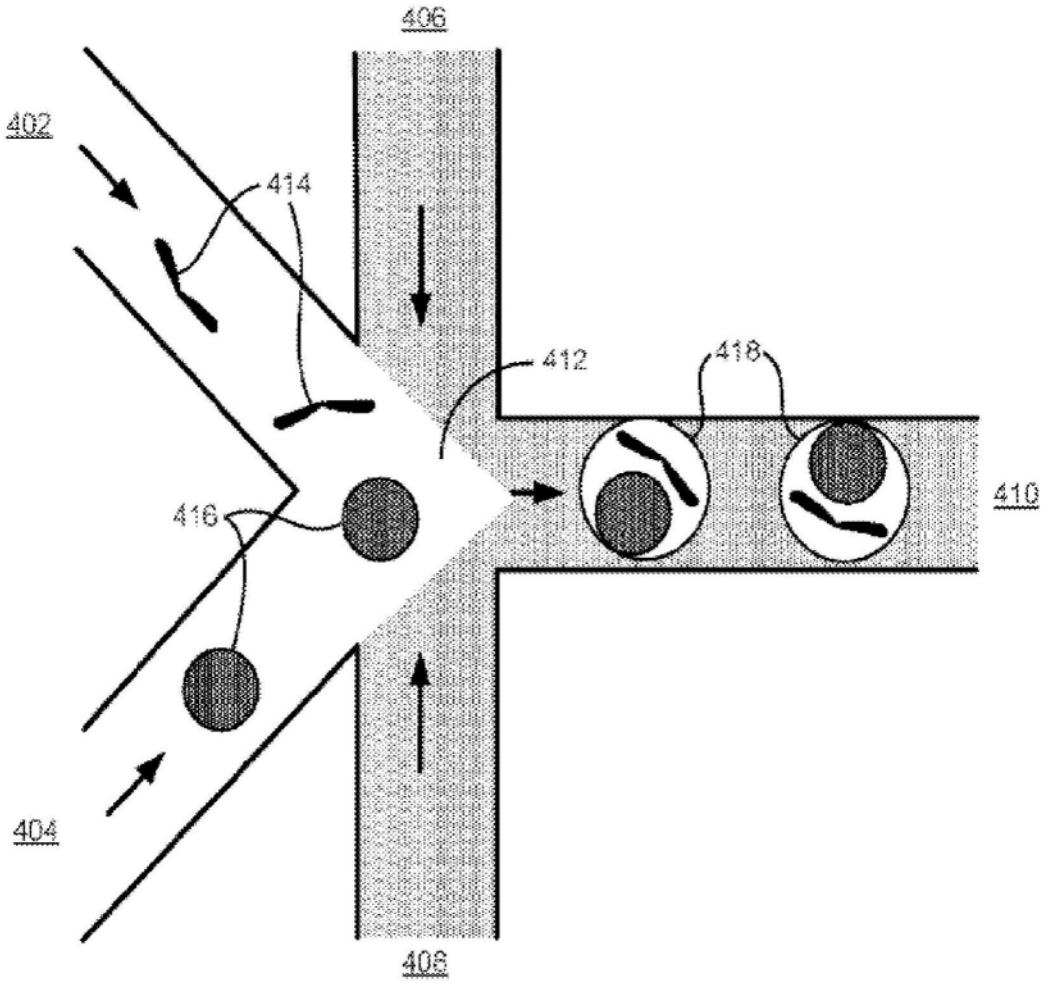

图4

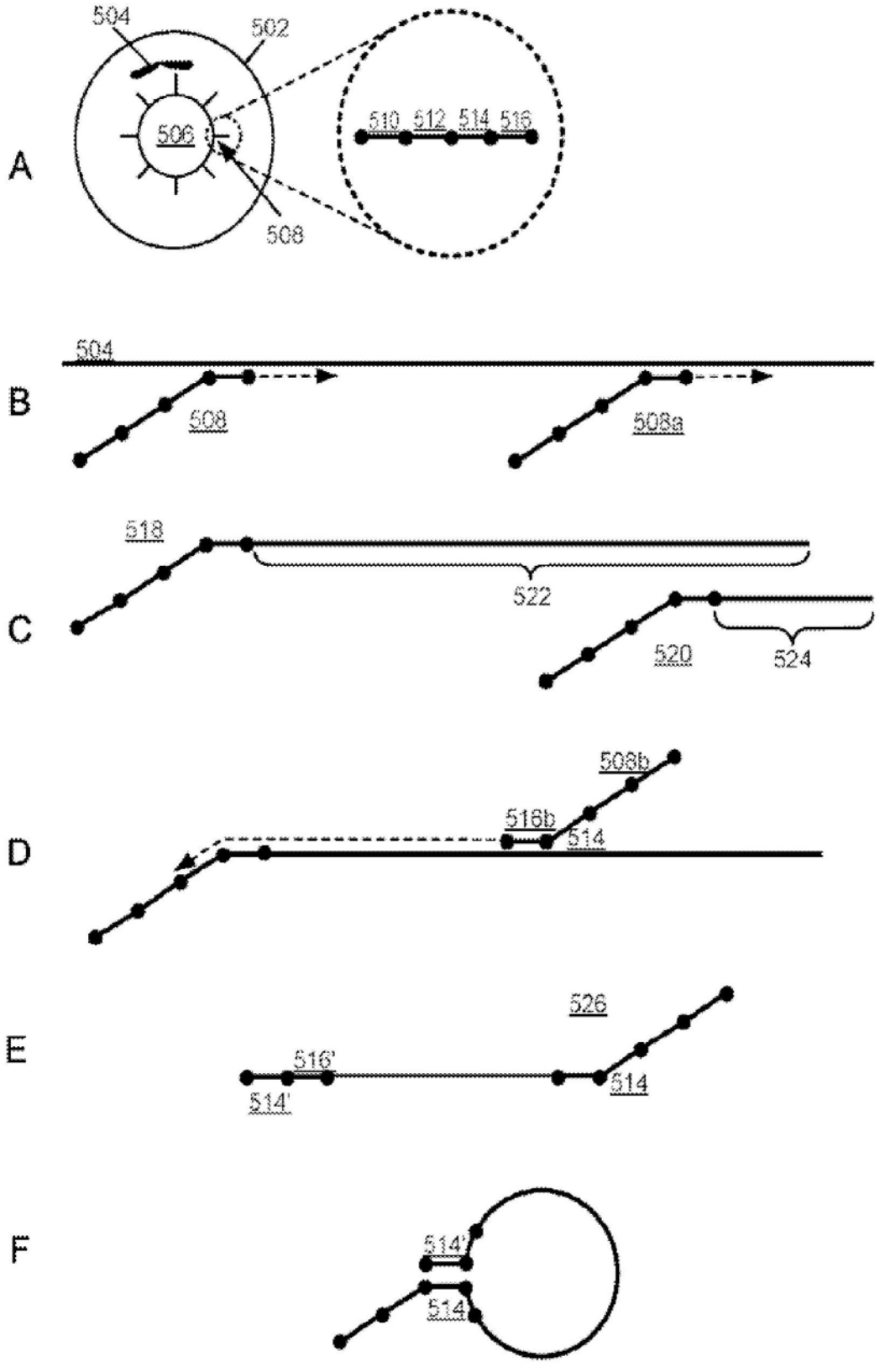
A
504
502
506
508
510 512 514 516
B
504
508
508a
C
518
522
520
524
D
508b
516b
514
E
526
516'
514'
514
F
514'
514

图5

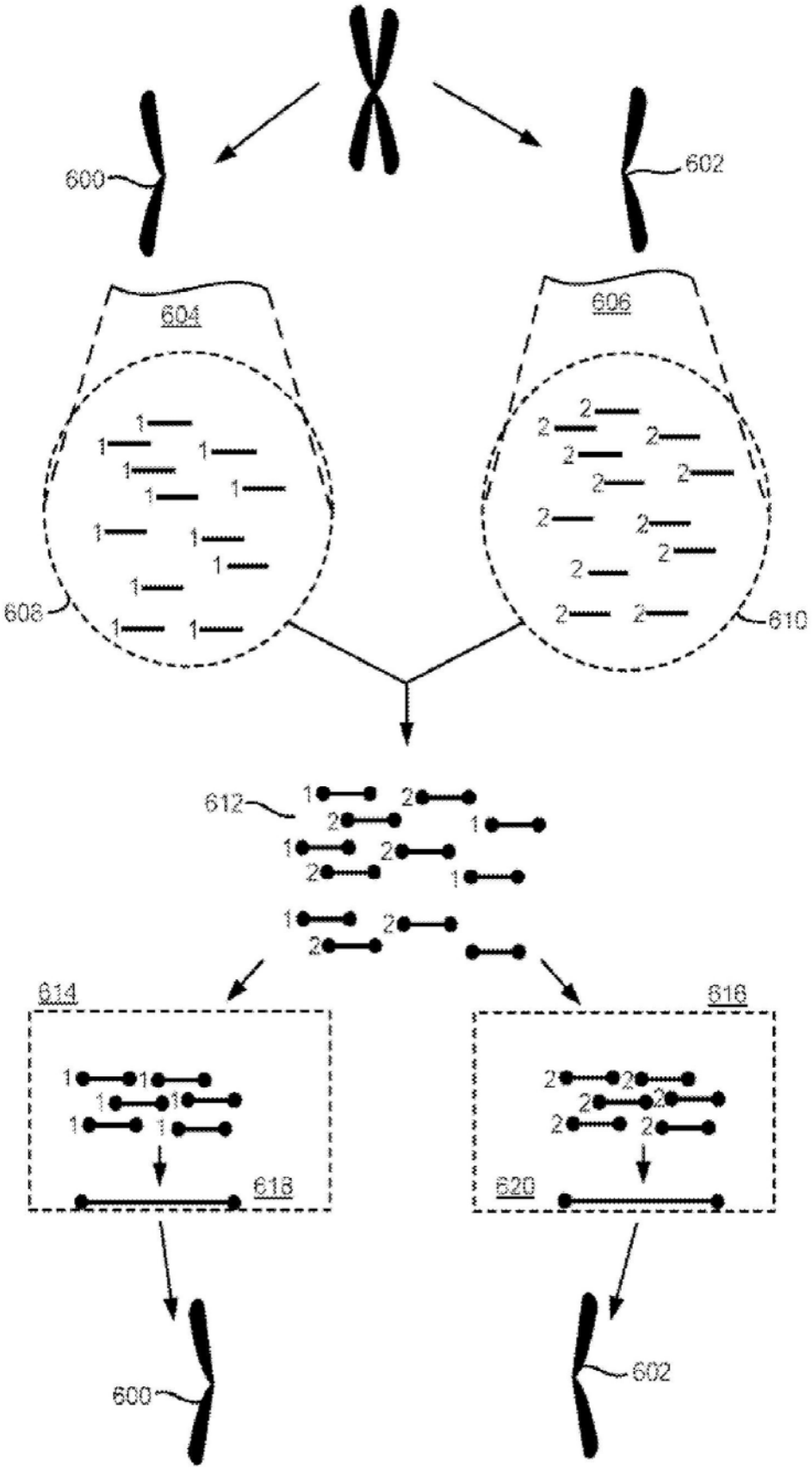

图6

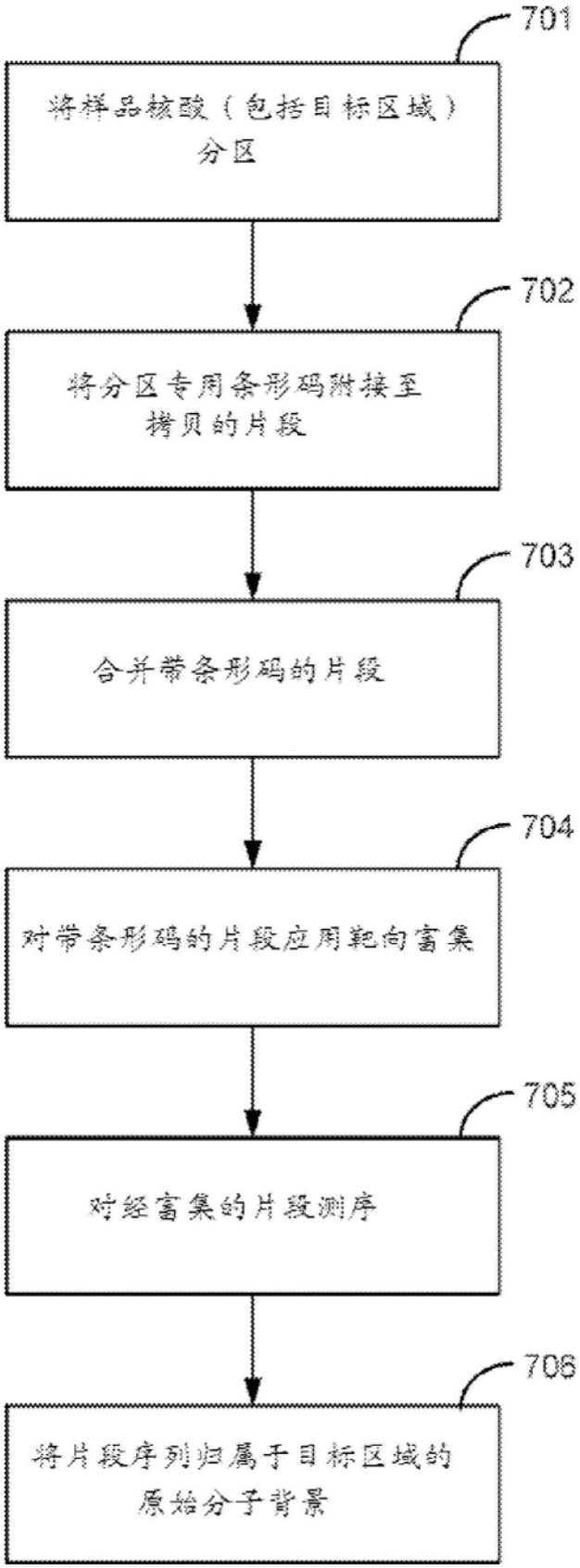
701
将样品核酸（包括目标区域）分区
702
将分区专用条形码附接至拷贝的片段
703
合并带条形码的片段
704
对带条形码的片段应用靶向富集
705
对经富集的片段测序
706
将片段序列归属于目标区域的原始分子背景

图7

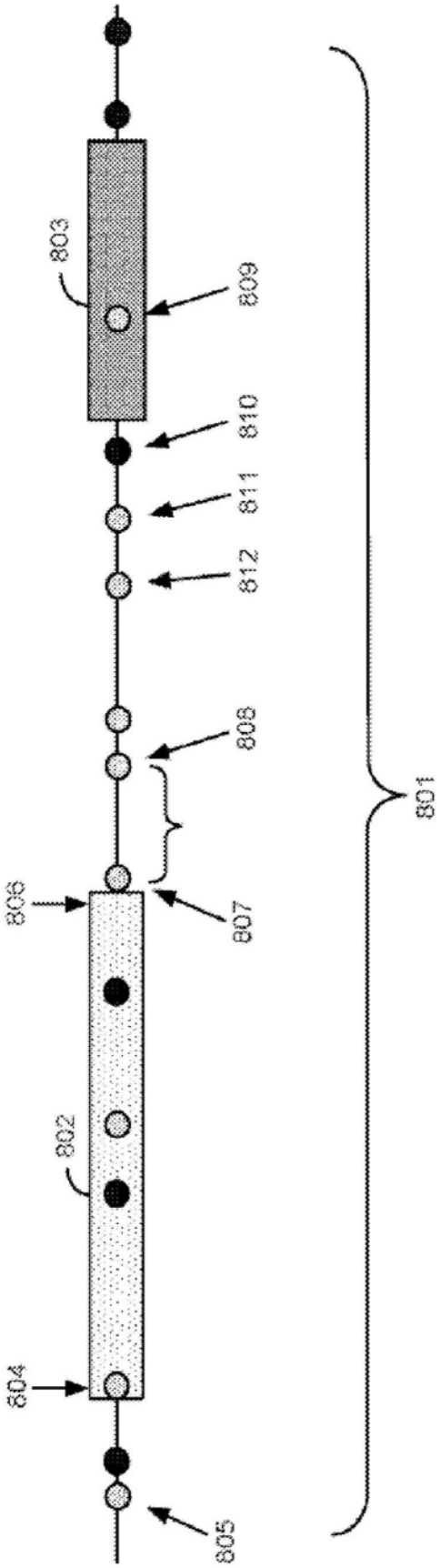

图8