



(12)发明专利申请

(10)申请公布号 CN 108473984 A

(43)申请公布日 2018.08.31

(21)申请号 201680077294.5

(22)申请日 2016.11.04

(30)优先权数据

62/250,900 2015.11.04 US

(85)PCT国际申请进入国家阶段日

2018.06.29

(86)PCT国际申请的申请数据

PCT/US2016/060576 2016.11.04

(87)PCT国际申请的公布数据

W02017/079593 EN 2017.05.11

(71)申请人 阿特雷卡公司

地址 美国加利福尼亚州

(72)发明人 G·威西

(74)专利代理机构 北京市金杜律师事务所

11256

代理人 陈文平

(51)Int.Cl.

C12N 15/10(2006.01)

C12Q 1/6806(2018.01)

权利要求书4页 说明书26页

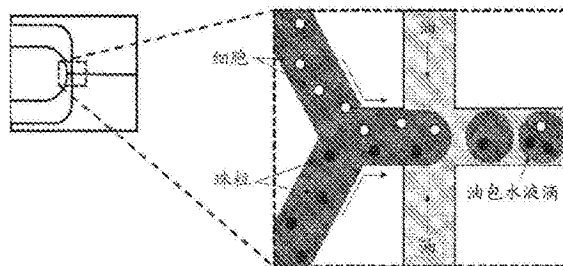
序列表3页 附图13页

(54)发明名称

用于分析与单个细胞相关联的核酸的核酸条形码的组合组

(57)摘要

本文提供鉴定核酸样品来源的方法。所述方法包括形成反应混合物,所述反应混合物包含核酸样品,所述核酸样品包含来自单个细胞的核酸分子和条形码组,将所述条形码组并入所述样品的核酸分子中,以及鉴定并入所述单个细胞的核酸分子中的所述条形码组,从而鉴定所述核酸样品的来源。



1. 一种鉴定核酸样品来源的方法,其包括以下步骤:

(a) 形成包含核酸样品的反应混合物,所述核酸样品包含来自单个细胞的核酸分子和条形码组;

(b) 将所述条形码组并入所述样品的核酸分子中;以及

(c) 鉴定并入所述单个细胞的核酸分子中的所述条形码组,从而鉴定所述核酸样品的来源。

2. 根据权利要求1所述的方法,其中形成所述反应混合物包括使所述单个细胞与所述条形码组接触并裂解所述细胞以提供来自所述单个细胞的核酸样品。

3. 根据权利要求1所述的方法,其中形成所述反应混合物包括提供在悬浮缓冲液中的细胞和使包含所述细胞的悬浮缓冲液与包含所述条形码组的条形编码缓冲液在一定条件下接触以形成包含所述单个细胞和所述条形码组的反应混合物。

4. 根据权利要求1至3中任一项所述的方法,其中所述条形码组包含各自包含独特的5'条形码的一种或多种5'核酸构建体;和各自包含独特的3'条形码的一种或多种3'核酸构建体。

5. 根据权利要求4所述的方法,其中形成所述反应混合物包括提供在悬浮缓冲液中的细胞,以及使所述悬浮缓冲液与包含所述5'和3'核酸构建体的条形编码缓冲液在一定条件下接触以形成包含所述单个细胞和至少一种5'核酸构建体以及至少一种3'核酸构建体的反应混合物。

6. 根据权利要求4所述的方法,其中形成所述反应混合物包括提供在悬浮缓冲液中的细胞,以及使所述悬浮缓冲液与包含所述5'核酸构建体的第一条形编码缓冲液和包含所述3'核酸构建体的第二条形编码缓冲液在一定条件下接触以形成包含所述单个细胞和至少一种5'核酸构建体以及至少一种3'核酸构建体的反应混合物。

7. 根据权利要求4至6中任一项所述的方法,其中所述构建体还包含独特的分子标识符(UMI)序列。

8. 根据权利要求4至7中任一项所述的方法,其中每种5'核酸构建体以附接到固体表面的多个拷贝提供。

9. 根据权利要求4至8中任一项所述的方法,其中每种3'核酸构建体以附接到固体表面的多个拷贝提供。

10. 根据权利要求8或9所述的方法,其中所述固体表面是珠粒、微珠、微粒、微球、纳米粒子、纳米珠粒或水凝胶。

11. 根据权利要求4至10中任一项所述的方法,其中所述反应混合物包含来自所述单个细胞的核酸样品、包含所述5'核酸构建体的多个拷贝的至少一个珠粒和包含所述3'核酸构建体的多个拷贝的至少一个珠粒。

12. 根据权利要求4至11中任一项所述的方法,其中每个5'和3'核酸构建体还包含可光裂解的连接子或限制性酶位点。

13. 根据权利要求4至11中任一项所述的方法,其中每种5'核酸构建体还包含可光裂解的连接子,并且每种3'核酸构建体还包含限制性酶位点。

14. 根据权利要求4至11中任一项所述的方法,其中每种5'核酸构建体还包含限制性酶位点,并且每种3'核酸构建体还包含可光裂解的连接子。

15. 根据权利要求13或14所述的方法, 其中所述方法还包括使所述反应混合物与限制性酶接触并将所述反应混合物暴露于紫外光以从所述固体表面释放所述5'和3'核酸构建体。

16. 根据权利要求12所述的方法, 其中所述5'和3'核酸构建体还包含限制性酶位点, 并且所述方法还包括使所述反应混合物与限制性酶接触以从所述固体表面释放所述5'和3'核酸构建体。

17. 根据权利要求12所述的方法, 其中所述5'和3'核酸构建体还包含可光裂解的连接子, 并且所述方法还包括将所述反应混合物暴露于紫外光以从所述固体表面释放所述5'和3'核酸构建体。

18. 根据权利要求4至17中任一项所述的方法, 其中所述5'核酸构建体是包含启动子的双链核酸构建体, 并且其中所述5'核酸构建体用作模板以生成所述5'核酸构建体的多个拷贝。

19. 根据权利要求1至18中任一项所述的方法, 其中所述条形码组包含1、2、3、4、5、6、7或8种5'核酸构建体和1、2、3、4、5、6、7或8种3'核酸构建体。

20. 根据权利要求1至19中任一项所述的方法, 其中所述核酸样品中5'和3'条形码的组合鉴定所述核酸样品的来源。

21. 根据权利要求1至20中任一项所述的方法, 其中所述3'核酸构建体还包含结合序列。

22. 根据权利要求21所述的方法, 其中所述结合序列包含poly-T序列。

23. 根据权利要求22所述的方法, 其中所述poly-T序列结合所述样品的核酸分子上的poly-A序列。

24. 根据权利要求1至23中任一项所述的方法, 其中所述3'核酸构建体上的条形码通过逆转录并入。

25. 根据权利要求24所述的方法, 其中所述逆转录生成包含所述3'条形码序列的核酸分子。

26. 根据权利要求25所述的方法, 其中所述逆转录酶将2至5个半胱氨酸残基添加至所述样品中的所述核酸分子的末端。

27. 根据权利要求26所述的方法, 其中所述2至5个半胱氨酸残基提供与所述5'核酸构建体上的结合序列互补的结合序列。

28. 根据权利要求27所述的方法, 其中所述5'核酸构建体上的结合序列包含poly-G序列。

29. 根据权利要求28所述的方法, 其中所述5'核酸构建体上的结合序列结合所述样品的核酸分子上的poly-C序列。

30. 根据权利要求1至29中任一项所述的方法, 其中所述5'核酸构建体使用逆转录酶并入所述核酸样品中。

31. 根据权利要求24至30中任一项所述的方法, 其中所述逆转录酶为M-MLV酶。

32. 根据权利要求31所述的方法, 其中所述逆转录酶为MMLV H-逆转录酶。

33. 根据权利要求1至32中任一项所述的方法, 其中进一步扩增所述核酸样品以生成各自包含5'条形码和3'条形码的多个核酸序列的多个拷贝。

34. 根据权利要求1至33中任一项所述的方法,其中所述核酸分子包含mRNA。
35. 根据权利要求34所述的方法,其中所述mRNA包含至少1、3、10、30、100、300、1,000、3,000、10,000、30,000、100,000、300,000或1,000,000个mRNA。
36. 根据权利要求1至35中任一项所述的方法,其中所述核酸样品中的条形码通过测序鉴定。
37. 根据权利要求1至3中任一项所述的方法,其中所述条形码组包含各自包含独特的5'条形码的至少两种5'核酸构建体。
38. 根据权利要求1至3中任一项所述的方法,其中所述条形码组包含各自包含独特的3'条形码的至少两种3'核酸构建体。
39. 根据权利要求1所述的方法,其中所述核酸分子中的一种或多种包含两种或更多种独特的条形码。
40. 根据权利要求39所述的方法,其中所述独特的条形码位于所述核酸分子的5'端、3'端或5'端和3'端两者。
41. 根据权利要求1至40中任一项所述的方法,其中所述方法包括形成多种反应混合物。
42. 根据权利要求41所述的方法,其中两种或更多种独特的条形码存在于30%、40%、50%、60%、70%、80%、90%或100%的所述反应混合物中。
43. 根据权利要求41所述的方法,其中形成所述多种反应混合物包括以大于0.5、0.6、0.7、0.8、0.9、1.0、2.0、3.0、4.0或5.0的 λ 值将所述条形码组添加到所述多种反应混合物中。
44. 根据权利要求1至40中任一项所述的方法,其中所述反应混合物为液滴。
45. 一种鉴定核酸样品来源的方法,其包括以下步骤:
- (a) 提供各自包含独特的3'条形码序列、结合序列和条形码关联序列的3'核酸构建体组;
 - (b) 提供包含来自单个细胞的核酸分子的核酸样品;
 - (c) 使所述核酸样品与所述3'核酸构建体组接触;
 - (d) 分离所述3'条形码序列和结合序列与所述条形码关联序列;
 - (e) 将来自所述3'核酸构建体的3'条形码序列并入所述核酸分子中;以及
 - (f) 使用所述条形码关联序列鉴定所述3'条形码序列组,由此鉴定来自所述单个细胞的所述核酸样品的来源。
46. 根据权利要求45所述的方法,其中所述3'核酸构建体组包含2、3、4、5、6、7或8种3'核酸构建体。
47. 根据权利要求45或46所述的方法,其中所述3'核酸构建体还包含独特的分子标识符(UMI)序列。
48. 根据权利要求45至47中任一项所述的方法,其中所述3'核酸构建体能够形成发夹。
49. 根据权利要求45至48中任一项所述的方法,其中所述3'条形码序列和结合序列通过限制性酶序列或可光裂解的连接子与所述条形码关联区段分开。
50. 根据权利要求45至49中任一项所述的方法,其中所述3'核酸构建体的多个拷贝附接到固体表面。

51. 根据权利要求50所述的方法, 其中所述固体表面是珠粒、微珠、微粒、微球、纳米粒子、纳米珠粒或水凝胶。

52. 根据权利要求51所述的方法, 其中限制性酶从所述固体表面释放所述3'条形码序列, 并将所述结合序列与所述条形码关联区段分离。

53. 根据权利要求51所述的方法, 其中紫外光从所述固体表面释放所述3'条形码序列, 并将所述结合序列与所述条形码关联区段分离。

54. 根据权利要求45至53中任一项所述的方法, 其中所述3'核酸构建体上的结合序列包含poly-T序列。

55. 根据权利要求54所述的方法, 其中所述3'核酸构建体的结合序列结合所述样品的核酸分子上的所述poly-A序列。

56. 根据权利要求45至55中任一项所述的方法, 其中所述3'核酸构建体上的条形码序列通过逆转录并入。

57. 根据权利要求56所述的方法, 其中所述逆转录生成包含所述3'条形码序列的核酸分子。

58. 根据权利要求45至57中任一项所述的方法, 其中所述条形码关联区段包含引发序列。

59. 根据权利要求58所述的方法, 其中所述引发序列为回文序列。

60. 根据权利要求58所述的方法, 其中一个条形码关联区段的引发序列能够结合所述核酸样品中任何其他条形码关联区段的引发序列。

61. 根据权利要求59所述的方法, 其中彼此接触的所述条形码关联区段通过DNA聚合酶延伸以将条形码关联区段两者并入单个核酸中。

62. 根据权利要求45至61中任一项所述的方法, 其中所述方法还包括鉴定并入所述核酸分子中的所述条形码序列。

63. 根据权利要求62所述的方法, 其中所述核酸样品中的条形码序列通过测序鉴定。

64. 根据权利要求62所述的方法, 其中所述条形码关联区段通过测序鉴定。

用于分析与单个细胞相关联的核酸的核酸条形码的组合组

[0001] 相关申请的交叉引用

[0002] 本申请要求2015年11月4日提交的美国临时申请62,250,900的优先权,所述临时申请以引用的方式整体并入本文。

[0003] 发明背景

[0004] 细胞群体的基因组和转录组特征谱分析是一个强烈且日益增长的关注领域,其在生物技术和医学领域中具有相当重要的意义。核酸测序技术的最新进展使特征谱分析研究的范围远远超过了常规方法的实际通过量。针对单细胞基因组或转录组特征谱分析的早期努力涉及FACS分选或甚至将个别细胞人工挑选到单独的反应小瓶如多孔微量滴定板的孔中,其中可以应用所选择的基因组或转录组分析方法。在这种格式下,可以个别地操作并控制单独的反应。例如,确保每个孔仅含有单个细胞是比较简单的。这是重要的,因为多个细胞的共隔离会导致这些细胞的基因组或转录组特征谱彼此卷积,并且解开个别特征谱将是困难的或不可能的。这些格式的缺点在于时间密集、试剂密集并且劳动密集,并且由于这些原因而仅允许相对低通过量的过程。更近期的方法趋向于将细胞和所需的化学和分子试剂在诸如微米或纳米孔或液滴的纳升或皮升尺寸的容器中隔离。这些格式允许在时间、成本和通过量方面的巨大改善,其缺点是控制细胞和其他反应组分精确装载到容器中显著更具挑战性。组分通常随机分配到它们的容器,在这种情况下,组分在许多容器之中的分布遵循泊松分布。例如,在细胞的情况下,如果系统要求至少95%的细胞被单隔离(换句话说,仅容忍5%的具有细胞的容器含有两个或更多个细胞),则条件必须设定为每个容器平均接收0.1个细胞。在泊松项中,这是0.1的预期值,即“ λ ”。在这种情况下,90.48%的容器将不接收细胞,9.05%的容器将仅接收单个细胞,0.45%的容器将接收两个细胞,0.02%的容器将接收三个细胞,而较低的百分比将接收四个、五个、六个等的细胞。在这种情况下接收一个细胞的容器之中,95.1%的容器含有恰好一个细胞。这种系统的低效率是显而易见的,因为所有反应容器中超过90%缺乏细胞,因此是非生产性的。当涉及这些泊松分布中的两个或多个的交叉时,低效率就会复杂化。基因组和转录组特征谱分析的常用方法涉及将细胞与核酸条形码一起隔离,所述核酸条形码然后附加到源自该细胞的基因和/或转录物上。条形码稍后用于将每个基因或转录物明确地分配到来其来源细胞,从而包含该细胞的基因组或转录组特征谱。这种方案的解疑不仅取决于细胞被单个隔离,而且也取决于相同的容器仅接收单个条形码。条形码可以是单个核酸分子,但更常见的是通过诸如聚苯乙烯珠粒的粒子递送,其表面用许多拷贝的相同核酸条形码官能化。在任一种情况下,独特的条形码随机分配到反应容器将遵循与以前相同的装载统计学,即由预期值(λ)确定的泊松分布。仅其中恰好一个细胞与恰好一个条形码共隔离的容器将产生明确的基因组或转录组特征谱。在下游数据分析时,具有特定条形码的多细胞隔离将表现为单个基因组或转录组特征谱,所述基因组或转录组特征谱是共隔离的细胞的个别特征谱的联合。共隔离多个细胞与多个条形码的情况产生更加复杂的模糊性。这种模糊性对涉及单细胞基因组或转录组特征谱分析的所有应用都有不利的影响。例如,如果应用涉及通过其转录组特征谱鉴定的特定克隆的定量,则多个条形码与该克隆类型的细胞共隔离的情况将在下游分析中给出该克隆类型的多个细

胞的外观,因此人为地夸大了样品中细胞总群的表观百分比。对于要求至少95%的接收一个细胞和一个条形码的容器接收恰好一个细胞和恰好一个条形码的这种设计的特征谱分析系统,存在给出期望结果的细胞和条形码 λ 组合的连续体(应注意,如果容器未能接收其中的至少一个,其是非生产性的,因此不予考虑)。一种代表性组合是对于细胞和条形码二者来讲 λ 为0.05。在这种情况下,一个容器平均将接收0.05个细胞和0.05个条形码。在细胞和条形码各自之中,95.12%的容器将接收零个,4.76%的容器将接收一个,0.12%的容器将接收两个,而较低百分比的容器将接收三个、四个、五个等。在接收至少一个细胞的容器之中,97.52%的容器将接收恰好一个细胞,并且对于条形码也相似。重叠这两个百分比给出 $97.52\% \times 97.52\% = 95.1\%$ 的细胞-条形码共隔离事件为恰好一个细胞和恰好一个条形码。然而,为了达到这个标准,大于99.76%的容器是无价值的,因为它们不会接收细胞,不会接收条形码,或者两者都没有。换句话说,少于0.24%的容器将具有生成细胞特征谱的机会。此外,由于少于5%的细胞与条形码共隔离,超过95%的细胞因分析而损失。这极大地减小了系统的通过量,并且当待分析的关注细胞是稀有细胞时尤其不利,因为这导致95%的细胞因分析而损失。这些皮升和纳升级方法可以在通过量(即,处理的细胞的数量)和样品覆盖率(即,处理的原始样品中细胞总数的百分比)方面实现有意义的改善,即,在不牺牲条形码方案的保真度的情况下可以增加接收条形码的容器的百分比。

[0005] 对于使用核酸条形码的单克隆“包”或“组”的基因组和转录组特征谱分析方法,诸如条形码核酸官能化珠粒、装载有核酸条形码的水凝胶或相似的条形码递送方法,核酸条形码包几乎总是表现出一定程度的多克隆性。多克隆性的程度取决于用于产生条形码包的方法。例如,一种方法使用具有通用引发序列的DNA寡聚物官能化并且用个别条形码分子乳化的珠粒,所述条形码分子然后通过PCR扩增到珠粒的表面上。在该方案中,乳液中的每个液滴可以含有几个全部接收相同条形码序列的珠粒,但是在理想的情况下,任何给定的液滴将仅接收一个条形码核酸。如果两个或多个条形码核酸进入单个液滴,则在该液滴中生成的任何珠粒都将是多克隆的。根据可以通过改变条形码浓度来调谐的泊松分布将条形码核酸分子分配给液滴。这里又是一个折衷。较低浓度的条形码将生成具有较高程度的单克隆性的珠粒,而具有将生成较少的总体条形码化珠粒的下行成本。更高的浓度将增加条形码化珠粒产量,其中更大比率的珠粒将表现出更高的多克隆性。多克隆性成为这种方法的设计的一部分。包装核酸条形码的另一方法涉及组合合成。在384孔板中执行,每个孔含有通用DNA寡聚引物官能化的大量粒子,在这种情况下,所述粒子为水凝胶。每个孔还含有独特的初始条形码序列,所述条形码序列通过PCR并入该孔中的每个水凝胶中。然后将水凝胶汇集在一起并重新分布到另一384孔板中,其中的每个孔都含有不同的二次条形码序列,所述二次条形码序列再次通过PCR并入每个水凝胶中。通过这种方法得到的水凝胶文库具有约150K条形码的多样性,并且理论上应该表现出高度的单克隆性。然而,实际上,已经报道该方法产生不可忽略程度(约8%)的多克隆性(Klein, A. Cell 161 (2015))。这种多克隆性可能由于PCR步骤中在PCR板的孔之间水凝胶或条形码核酸的污染或者原始条形码核酸文库-初始或二次条形码序列或两者-的污染引起。无论是通过设计还是通过污染,条形码包文库的任何多克隆性都会混淆目的在于为每个细胞分配单个条形码的基因组或转录组特征谱分析系统。条形码编码方案将受益于任何能够处理条形码包文库中的一定程度的多克隆性的能力。

[0006] 已经采取了一些方法的实例来克服或避免由核酸条形码的基于泊松的分布引起的低效率。已经报道了使用水凝胶粒子的转录组特征谱分析方法,所述水凝胶粒子装载有个别条形码并且引入液滴装置的水流中,紧接与细胞(其根据泊松分布随机装载,其中 λ 为0.1-0.2)一起封装在液滴中。由于水凝胶具有很好的顺从性,它们可以在狭窄的入口内填充成单行,并以高度有序的行列引入,其可以与液滴形成的速率同步。结果,几乎所有的液滴都接收具有独特条形码的许多拷贝的一个水凝胶粒子(Abate, R. Lab on a Chip 9 (2009))。然而,虽然这样一个经过仔细调谐的实验可以在高度优化的环境中并通过专家的手执行,然而这种方法是否具有足够的可扩展性和稳健的广泛用途还有待观察。此外,该方案对多克隆条形码包不稳健。其他方法涉及将多个关注的基因组或转录组目标一起并入单个连续扩增子中,如重叠延伸PCR的情况。这些方法依靠定制的通用引物组来物理连接关注的目标,而不是为每个目标附加条形码。由于引物组是通用的,并且许多拷贝引入每个容器中,所以将目标分配给其来源细胞的机制并不面对泊松驱动分布的缺点。然而,这些方法在可以自每个细胞靶向的潜在基因或转录物的数量方面受到严重限制。大多数(如果不是全部的话)现有方法目前仅限于两种(DeKosky, B. J., Nat Biotechnol 31 (2) (2013))。

[0007] 发明简述

[0008] 本文提供鉴定核酸样品来源的方法。所述方法包括形成反应混合物,所述反应混合物包含核酸样品,所述核酸样品包含来自单个细胞的核酸分子和条形码组;将所述条形码组并入所述样品的核酸分子中;以及鉴定并入所述单个细胞的核酸分子中的所述条形码组,从而鉴定所述核酸样品的来源。还提供了鉴定核酸样品来源的方法,所述方法包括提供5'和3'核酸构建体组,每种构建体分别包含独特的5'或3'条形码序列;提供包含来自单个细胞的核酸分子的核酸样品;使所述核酸样品与所述组的5'和3'核酸构建体接触;将来自所述5'和3'核酸构建体的5'和3'条形码序列分别并入所述核酸分子中;以及鉴定所述组的5'和3'条形码序列,从而鉴定来自所述单个细胞的核酸样品的来源。

[0009] 附图描述

[0010] 图1是具有以用于液滴形成的流动聚焦构造的油输入通道和以下水性输入通道的微流体液滴芯片的示意图:(1)在悬浮缓冲液中的细胞和(2)在条形码化混合物中的条形码化珠粒。悬浮缓冲液和条形码化混合物位于不同的微流体通道中,这些微流体通道会聚以按比率合并它们的组分,以构成液滴中所需的最终混合物。如泊松分布所述,将条形码化珠粒和细胞装载到水滴中。每个液滴的珠粒的平均值(λ)和每个液滴的细胞的平均值(λ)是这些组分在其输入流中的浓度的函数。液滴是发生条形码化反应的反应容器。

[0011] 图2是示出两个珠粒组(3'和5')的示例性组成的示意图。每个3'珠粒上的条形码核酸包含用于从珠粒上裂解条形码序列的限制性位点(RS)、用于下游分析的独特分子识别序列(UMI)、条形码序列(BC3)和poly-T序列(TTT)。每个5'珠粒上的条形码核酸包含限制性位点(RS)、条形码序列(BC5)和三个鸟嘌呤残基。

[0012] 图3是示出在液滴内裂解细胞并且释放mRNA转录物并且将其可用于条形码化反应的示意图。限制性酶(RE)在限制性位点从它们的珠粒裂解3'条形码和5'条形码两者。当3'条形码的poly-T序列退火成mRNA转录物的3'端的poly-A尾部时,引发逆转录。从该3'条形码引物开始,逆转录酶(RT)酶以5'至3'方向逆转录以生成包括3'条形码和mRNA序列两者的单个cDNA扩增子。达到转录物的末端后,RT酶(M-MLV酶)增加为5'条形码的三个鸟嘌呤残基

提供退火位点的数个胞嘧啶残基。一旦5'条形码退火,RT酶就“模板切换”并将5'条形码序列并入同一cDNA中。

[0013] 图4是示出两个3'条形码化珠粒和三个5'条形码化珠粒与细胞共封装的示意图。每个3'条形码与每个5'条形码配对,反之亦然,形成一个有内聚力的独特条形码组,所述条形码组鉴定所有转录体来源于该细胞。

[0014] 图5是示出当过早终止逆转录时,cDNA将不能并入5'条形码的示意图。3'条形码文库可以设计成使得(就条形码总数而言)任何一个条形码将与多个细胞共封装变得极不可能。在这种情况下,非全长条形码化cDNA可以以高置信度分配给适当的特征谱。

[0015] 图6是一个示意图,其示出尽管相同的两个3'条形码序列共封装在两个不同的液滴中,但是由于一个液滴中额外的3'条形码,这两个条形码组仍是可区分的。由于该第三3'BC不与另一液滴的两个5'BC配对,因此可以明确地绘制分开两个条形码组的线。

[0016] 图7是示出使用一个示例性条形码组的示意图。在此,5'条形码珠粒组用双链DNA寡聚物官能化,其杂交(非共价键合)链包含T7启动子序列(“T7”)、条形码序列和三个胞嘧啶残基。T7启动子不是通过限制性酶从珠粒表面上裂解,而是允许T7RNA聚合酶结合并转录来自珠粒表面上的DNA寡聚物模板的RNA条形码拷贝。这些RNA条形码拷贝将在3'端以三个鸟嘌呤残基终止,这将使得它们能够退火到全长3'-条形码化cDNA的3'端。RT酶模板-切换至5'条形码RNA寡聚物模板,并将5'条形码并入cDNA中。

[0017] 图8是示出用单链条形码寡核苷酸表面官能化的每个珠粒的示意图。寡聚物形成双链区段含有限制性位点的发夹结构。限制性酶裂解两条链并释放转录物条形编码区段和环状区段。环状区段包括与转录物条形编码区段上的条形码相同或简单相关联的条形码,以及“条形码关联区段”(BAS)。BAS允许多个环状区段在DNA聚合酶反应中彼此引发(prime off),所述聚合酶反应将两个环状区段的两个条形码在单个扩增子中物理连接在一起。为了实现这一点,BAS可以是回文式的,或者可以由离散的组(例如,两个互补序列)组成,这两者都合并每个珠粒上。

[0018] 图9是示出具有选择性退火至mRNA的poly-A尾部的poly-T尾部的转录物-条形编码区段的示意图。转录物序列被逆转录到也携带条形码的单个cDNA。

[0019] 图10是示意图,其示出当两个条形码化珠粒与细胞共封装时,来自该细胞的每个转录物可以由这两个条形码(BC1a或BC2a)中的任一个条形编码。同一液滴还将含有与这两个转录物条形码中的每一个对应的环状区段的许多拷贝。

[0020] 图11是示意图,其示出环状区段可以设计成使得与单个环状区段在环状结构中保持自退火相比,优选经由BAS在环状区段之间交叉退火。可以包括DNA聚合酶以延伸每个环状区段以并入另一条形码的互补序列。

[0021] 图12是示出与样品相关联的条形码的示意图。5'条形码(5'BC)或3'条形码(3'BC)可以如下与样品相关联:(a)在核酸的每个末端上的一者与样品相关联,或(b)在核酸的5'端上的两者与样品相关联,或(c)在核酸的3'端上的两者与样品相关联,或者(d-f)在核酸内或在与样品相关联的两个核酸共价连接的情况下在内部与样品相关联。可以通过已知方法将条形码添加到样品中,所述方法包括但不限于连接或扩增技术,诸如聚合酶链式反应(PCR)。

[0022] 图13是示出与样品相关联的条形码组的分析示意图。条形码化核酸可以表示为

二分图,其中圆形顶点(5'条形码)和矩形顶点(3'条形码)通过边(代表核酸)连接。在反应容器中,5'和3'条形码的所有组合都可以与核酸相关联,因此所有圆形顶点都应该通过边连接到所有矩形顶点,形成最大聚类。使用最大聚类将正确地关联样品内的条形码和核酸。

[0023] 图14是示出与样品解析条形码“碰撞”相关联的多个条形码组的示意图。(顶部)如果使用单个条形码组,并且不同的反应容器碰巧含有相同的条形码珠粒,则会发生来自不同样品的核酸的错误关联。(中间)如果使用多个条形码组,则两个反应容器中的相同条形码(5'条形码)可以通过它们与其他条形码组(这里是3'条形码)的关联来解疑。共有的5'条形码将属于两个聚类,但自5'条形码顶点的边(代表核酸)将独特地仅属于1个聚类。(底部)如果在两个反应容器中有相同的5'条形码和3'条形码,则边将属于两个聚类,并且不知道核酸与哪一样品相关联。这可以从分析中检测和丢弃(用虚线表示)。

[0024] 图15是示出在序列读取对中观察到的条形码对的频率(y轴,分数)相对于观察到的次数(x轴,计数)的图。超过一半的条形码对恰好观察到一次,与代表可以直接滤出的实验噪声(诸如测序错误)的这些一致。很少条形码对被观察到超过10次。在50处的箱包括观察到50次或更多次的所有条形码对。

具体实施方案

[0025] 不是试图将单个独特的条形码引入反应容器中,本申请涉及在生产性反应的大多数或全部情况下有意引入形成鉴定所述容器中所有基因组或转录组目标的来源细胞的可区分的条形码组的多个条形码。这可以以几种不同方式中的任一种实现,并且关键的特征是当将多个条形码引入单个容器中时,除了对关注的基因组或转录组目标进行条形编码之外,还存在一种机制,所述容器中的所有条形码可以通过所述机制容器与容器中的所有其他条形码直接或间接相关联。当将多个而不是单个条形码分配到个别容器时,可以应用条形码的随机分布,而不要求大多数容器是空的,因此是非生产性的。所提供的方法还降低了应用两个或更多个泊松分布的交叉的低效率。

[0026] 为了证明这种条形编码方案在通过量和覆盖率方面的益处,考虑至少95%的接收一个细胞和一个条形码的容器恰好接收一个细胞和一个条形码的情况。在条形码方案中,相应的要求是大于95%的接收一个细胞的容器接收恰好一个细胞,因为接收的条形码的数量现在与实际情况无关。该标志物达到平均值为0.1个细胞/容器,在这种情况下,95.1%的接收至少一个细胞的容器将接收恰好一个细胞。可以基于条形码经济性、下游分析复杂性和样品覆盖率以及其他因素之间的期望平衡来选择每个容器的平均条形码数量。对于每个容器的三个条形码的平均值,相应的泊松分布表明95.02%的容器将接收至少一个条形码。然而,在单条形码方法中,少于0.24%的容器接收至少一个细胞和一个条形码,因此有机会多产地参与基因组或转录组特征谱分析系统,在这种类似的多条形码方法中,9.04%的容器接收至少一个细胞和一个条形码。这是潜在生产性容器数量的近四十倍增加,并且可以预期所得通过量(在特征谱分析的细胞数量方面)相应地增加。重要的是,这不仅反映特征谱分析的细胞的数量,而且反映特征谱分析的细胞的百分比。然而在单条形码方法中,仅4.88%的容器接收至少一个条形码,因此相应低百分比的细胞被特征谱分析,在平均值为三个条形码/容器的多珠粒方法中,95.02%的容器接收至少一个条形码。这可以使样品覆盖率增加近20倍,并且如果选择每个容器的平均条形码数量更高,则可以实现甚

至更大的增益。样品覆盖率的这种改善是极其有益的,并且对于寻求相对稀少的细胞克隆的特征谱分析的应用尤其如此。

[0027] 多条形码方法的另一益处是该方案对于多克隆性是稳健的,这是旨在生成多组单克隆条形码的多数或所有方法的常见结果。而在单一条形码方法中,含有两个或更多个不同条形码的单个组会混淆该条形码组所分配给的细胞的特征谱,而将给定容器中的所有条形码一起关联到一个综合条形码组的特征谱分析方案能够处理多克隆条形码组。它还可以实现珠粒条形码工艺的收率和/或成本方面的显著改善。使用这里描述的基于乳液PCR的方法作为实例,如果可容许的阈值使得至少95%的条形码珠粒必须是单克隆的,则乳化过程期间的条形码拷贝 λ 必须设定为0.1,结果是少于10%的珠粒将接收一个条形码,而剩余的珠粒被浪费掉。如果系统设计成可以容许更大比例的多克隆珠粒-99%的珠粒必须是单编码、双编码或三重编码的(在这种情况下,约70%是单编码的)-条形码拷贝 λ 可以设定为0.7,结果是大于50%的珠粒接收至少一个条形码。这引起产量增加超过五倍。

[0028] 表1. 作为emPCR珠粒条形码反应中每个液滴的条形码的平均拷贝数(λ)以及相对于 $\lambda=0.1$ 的情况的条形码珠粒收率的相应倍数增加的函数的珠粒的克隆性百分比。

0	克隆性				通过量成 倍增加相 对于 0.1 λ
	单	双	三	四	
0.1	95.1%	4.8%	0.2%	0.0%	NA
0.2	90.3%	9.0%	0.6%	0.0%	1.9
0.3	85.7%	12.9%	1.3%	0.1%	2.72
0.4	81.3%	16.3%	2.2%	0.2%	3.46
0.5	77.1%	19.3%	3.2%	0.4%	4.13
0.6	73.0%	21.9%	4.4%	0.7%	4.74
0.7	69.1%	24.2%	5.6%	1.0%	5.29

[0030] 通过使用可区分的条形码组而不是单个条形码来鉴定个别基因组或转录组特征谱,该方案还可以鉴定任何一个条形码进入多个反应容器的情况。单个条形码序列与另外两个不同组的关联将允许鉴定并通过单独的分析机制分类或丢弃这类情况。

[0031] 可以使用图论,特别是使用二分图并搜索最大聚类来完成与样品关联的条形码和核酸的正确分配。二分图是其顶点(或节点)可以被分成两个不相交组的图,其中每条边将第一组中的顶点连接到第二组中的顶点。对于我们的目的,这两组顶点表示两个条形码组(5'和3'条形码组),且连接顶点的边表示具有该5'条形码的核酸以及并入其中的特定3'条形码(图13)。在反应容器中,5'和3'条形码的所有组合都可以与核酸相关联,因此所有圆形顶点应该通过边连接到所有矩形顶点,形成最大聚类。聚类是无向图的顶点的子组,使得聚类中的每两个不同的顶点相邻。最大聚类是不能通过包含一个相邻顶点来扩展的聚类,这意味着它不是一个较大聚类的子组。使用计算机程序以信息方式解析最大聚类将正确地将样品内的条形码和核酸相关联(参见图13和14)。图14显示使用多个条形码组的额外优点。可以检测并解析任何一个条形码进入多个反应容器(或条形码“碰撞”)的实例。这在使用仅一个条形码时是不可能的。

[0032] 功能组内的多个条形码的组合固有地增加了代码空间的更大多样性。例如,在每

个cDNA拷贝在每个末端(5'和3')接收一个条形码并且每个组(5'和3')中的编码空间是100的本发明的一个可能的实施方案中,系统的总编码空间是 $100 \times 100 = 10,000$ 个可能的条形码组合。这可以允许生成更小的条形码珠粒文库,以便以更高的条形码编码保真度对更大数量的细胞进行条形码编码。

[0033] 为了方便,条形码组被称为5'条形码和3'条形码。可以使用多个条形码组将核酸分配到适当的反应容器,并因此鉴定核酸样品的来源。条形码组可以在5'端和3'端与核酸相关联,或在5'端两者相关联,或在3'端两者相关联,或甚至在核酸样品内部相关联(参见图12)。可以使用本领域已知的方法将条形码添加到核酸中,所述方法诸如但不限于扩增(诸如聚合酶链式反应(PCR)、连接和转座反应(请参阅Green等Molecular Cloning, A Laboratory Manual, 第4版, Cold Spring Harbor Press (2012))。只要至少两个或更多个条形码组与核酸样品相关联,就可以将核酸分配到具有本文提及的所有优点的适当反应容器(并且因此是样品),包括在用条形码珠粒装载反应容器和解吸条形码“碰撞”方面冲击泊松统计学,其中任何一个条形码进入多个反应容器中。

[0034] 在单个液滴中包含多个条形码化珠粒也可能有益于基因或转录物条形码编码反应经其进行的分子生物学。由于条形码核酸分子是条形码编码反应中的反应物并且任何反应的效率都与其反应物的供应直接相关,所以在每个反应中包含多个条形码将引起条形码化产物的收率增加。

[0035] 本文提供鉴定核酸样品来源的方法。所述方法包括形成反应混合物,所述反应混合物包含核酸样品,所述核酸样品包含来自单个细胞的核酸分子和条形码组;将所述条形码组并入所述样品的核酸分子中;以及鉴定并入所述单个细胞的核酸分子中的所述条形码组,从而鉴定所述核酸样品的来源。任选地,所述条形码组包含各自包含独特的5'条形码的至少两种5'核酸构建体。任选地,所述条形码组包含各自包含独特的3'条形码的至少两种3'核酸构建体。任选地,核酸分子中的一种或多种包含两种或更多种独特的条形码。任选地,独特的条形码位于核酸分子的5'端、3'端或5'端和3'端两者。任选地,所述方法包括形成多种反应混合物。任选地,两种或更多种独特的条形码存在于30%、40%、50%、60%、70%、80%、90%或100%的反应混合物中。任选地,形成多种反应混合物包括以大于0.5、0.6、0.7、0.8、0.9、1.0、2.0、3.0、4.0或5.0的 λ 值将所述条形码组添加到所述多种反应混合物中。

[0036] 还提供了鉴定核酸样品来源的方法,所述方法包括提供3'核酸构建体组,每种构建体包含独特的3'条形码序列;提供包含来自单个细胞的核酸分子的核酸样品;使所述核酸样品与所述3'核酸构建体组接触;将来自所述3'核酸构建体的3'条形码序列并入所述核酸分子中;以及鉴定所述3'条形码序列组,从而鉴定来自所述单个细胞的核酸样品的来源。

[0037] 如本文所用的术语,将序列“并入”多核苷酸是指通过磷酸二酯键将一系列核苷酸与多核苷酸的其余部分例如在多核苷酸的3'或5'端共价连接,其中核苷酸按所述序列规定的顺序连接。如果多核苷酸含有序列或其互补序列,则所述序列已被“并入”多核苷酸中,或者等同地多核苷酸“并入”所述序列。将序列并入多核苷酸可以酶促地(例如,通过连接或聚合)或使用化学合成(例如,通过亚磷酰胺化学)发生。

[0038] 术语“引物”是指能够选择性退火或杂交至目标核酸的多核苷酸(寡核苷酸)及其类似物。在合适的条件下,诸如在一种或多种适当的酶、辅因子、底物如核苷酸(dNTP)等存

在下,引物用作DNA合成的起始引物。引物允许合成与相应多核苷酸模板互补的序列、侧翼序列或来自引物的3'端的扩增产物。通常,引物的长度可以在约10至100个核苷酸之间。引物结合位点或通用引发序列是指引物能够结合以引发扩增的核酸序列。

[0039] 如本文所用,术语“扩增(amplify)”和“扩增(amplification)”是指全部或部分地酶促拷贝多核苷酸的序列,以便产生也含有所述序列或其互补序列的更多多核苷酸。被拷贝的序列被称为模板序列。扩增的实例包括通过RNA聚合酶进行的DNA模板化RNA合成、通过逆转录酶进行的RNA模板化第一链cDNA合成和使用热稳定DNA聚合酶进行的DNA模板化PCR扩增。扩增包括所有引物延伸反应。

[0040] 如本文所用,术语“等温”是指在恒定温度或温度范围下进行的反应,诸如酶反应。

[0041] 术语“相关联”在本文中用于指样品与源自或来源于该样品的DNA分子、RNA分子或其他多核苷酸之间的关系。该术语可以用于指在一个条形码序列与另一条形码序列之间或在一个条形码序列与一组其他条形码序列之间的关系。例如,如果条形码在相同条形码组中,则所述条形码与另一条形码相关联。或者,如果条形码序列是条形码序列组中的一个条形码,则所述条形码可以与所述条形码组相关联。如果多核苷酸是内源性多核苷酸,则其与样品相关联,即其在样品被选择时存在于样品中,或来源于内源多核苷酸。例如,细胞内源性的mRNA与所述细胞相关联。由这些mRNA的逆转录产生的cDNA和由cDNA的PCR扩增产生的DNA扩增子含有mRNA的序列并且也与细胞相关联。与样品相关联的多核苷酸不需要在样品中定位或合成,并且即使样品已经被破坏之后(例如,在细胞已被裂解之后)也认为其与样品相关联。可以使用分子条形码编码或其他技术来确定混合物中的哪些多核苷酸与特定样品相关联。

[0042] 如本文所用的术语,“反应体积”(或等同地“容器”或“隔室”)是可以容纳一定体积的液体如水溶液并保持其与其他这样的体积的液体或周围介质隔离(例如,分离)的空间。反应体积与其周围环境之间的隔离可以由反应体积周围的固体阻隔或由相分离产生。例如,悬浮在疏水性载体流体中的水性微流体微滴可以构成反应体积,因为水在载体流体中不混溶。因此,在载体流体中彼此分离的两个液滴保持隔离,并且溶解在一个液滴中的核酸或其他亲水性物质不能离开该液滴或转移到另一液滴。反应体积也可以通过例如烧瓶、烧杯、离心管和多孔板中的孔来限定。

[0043] 将条形码组“添加”到与例如来自单个细胞的样品相关联的RNA中包括形成含有条形码和单个细胞或来自单个细胞的RNA的反应混合物,使得RNA和条形码可以参与条形码反应。一旦添加,条形码则可以例如通过与RNA杂交而与一种或多种RNA直接反应,或者可以参与聚合反应或其中RNA分子用作模板的一系列反应(例如,逆转录或RT-PCR)。

[0044] 术语“一种或多种多核苷酸”是指核酸如DNA分子和RNA分子及其类似物(例如,使用核苷酸类似物或使用核酸化学产生的DNA或RNA)。根据需要,可以例如使用本领域公认的核酸化学合成制备多核苷酸,或者使用例如聚合酶酶促制备多核苷酸,并且如果需要,则可以对其进行修饰。典型的修饰包括甲基化、生物素化(biotinylation)和其他的本领域内已知的修饰。此外,多核苷酸可以是单链的或双链的,并且在需要时可以连接至可检测部分。在一些方面,多核苷酸可以包括例如包含DNA和RNA的杂合分子。

[0045] “G”、“C”、“A”、“T”和“U”各自通常分别代表含有鸟嘌呤、胞嘧啶、腺嘌呤、胸腺嘧啶核苷和尿嘧啶作为碱基的核苷酸。然而,应该理解,术语“核糖核苷酸”或“核苷酸”还可以指

修饰的核苷酸或替代置换部分。本领域技术人员很清楚,鸟嘌呤、胞嘧啶、腺嘌呤和尿嘧啶可以被其他部分置换而基本上不改变包含携带这种置换部分的核苷酸的寡核苷酸的碱基配对性质。例如但不限于,包含肌苷作为其碱基的核苷酸可以与含有腺嘌呤、胞嘧啶或尿嘧啶的核苷酸碱基配对。因此,含有尿嘧啶、鸟嘌呤或腺嘌呤的核苷酸可以在核苷酸序列中被含有例如肌苷的核苷酸置换。在另一实施例中,寡核苷酸中任何位置处的腺嘌呤和胞嘧啶可以分别被鸟嘌呤和尿嘧啶置换以与目标mRNA形成G-U摆动碱基配对。含有这种置换部分的序列适用于本文所述的组合物和方法。

[0046] 如本文所用,并且除非另有说明,否则如技术人员所理解,术语“互补”在用于描述与第二核苷酸序列有关的第一核苷酸序列时,是指包含第一核苷酸序列的多核苷酸在某些条件下与包含第二核苷酸序列的多核苷酸杂交并形成双链体结构的能力。这样的条件例如可以是严格条件,其中严格条件可以包括:400mM NaCl、40mM PIPES pH 6.4、1mM EDTA,50℃或70℃,12至16小时,然后洗涤。其他条件如生物体内可能遇到的生理学有关条件可以适用。根据杂交核苷酸的最终应用,技术人员将能够确定最适合两个序列的互补性的测试的一组条件。

[0047] 互补序列包括包含第一核苷酸序列的多核苷酸的区域与包含第二核苷酸序列的多核苷酸的区域在一个或两个核苷酸序列的长度或所述长度的一部分上碱基配对。这样的序列在本文中可以被称为彼此“互补”。然而,在第一序列被称为相对于本文中的第二序列“基本上互补”的情况下,这两个序列可以是互补的,或者它们可以在碱基配对的区域内包括一个或多个,但通常不超过约5、4、3或2个错配碱基对。对于具有错配碱基对的两个序列,只要这两个核苷酸序列通过碱基配对相互结合,就认为序列“基本上互补”。

[0048] 如本文所用,在实现关于其杂交能力的上述实施方案的范围内,“互补”序列还可以包括非沃森-克里克(non-Watson-Crick)碱基对和/或由非天然和修饰的核苷酸形成的碱基对或完全由其形成。这样的非沃森-克里克碱基对包括但不限于G:U摆动或Hoogsteen碱基配对。

[0049] 在两个或更多个核酸或多肽序列的上下文中,术语百分比“同一性”是指,如使用下文所述的序列比较算法(例如,BLASTP和BLASTN或技术人员可用的其他算法)之一或通过视觉检查所测量,在对于最大一致性进行比较和比对时两个或多个序列或子序列具有特定百分比的相同的核苷酸或氨基酸残基。取决于应用,百分比“同一性”可以存在于被比较的序列的区域上,例如在功能域上,或者可选地,存在于待比较的两个序列的全长上。

[0050] 对于序列比较,通常一个序列充当与测试序列进行比较的参考序列。当使用序列比较算法时,将测试序列和参考序列输入计算机中,必要时指定子序列坐标,并指定序列算法程序参数。序列比较算法随后基于指定的程序参数计算一个或多个测试序列相对于参考序列的百分比序列同一性。

[0051] 用于比较的序列的最佳比对可以例如通过以下方法进行:Smith和Waterman, Adv. Appl. Math. 2:482 (1981) 的局部同源性算法;Needleman和Wunsch, J. Mol. Biol. 48:443 (1970) 的同源性比对算法;Pearson和Lipman, Proc. Nat'l. Acad. Sci. USA 85:2444 (1988) 的对于相似性方法的探索;这些算法的计算机化实现(在Wisconsin Genetics Software Package中的GAP、BESTFIT、FASTA和TFASTA, Genetics Computer Group, 575 Science Dr., Madison, Wis.);或视觉检查(通常参见Ausubel等人,下文)。

[0052] 适合确定百分比序列同一性和序列相似性的算法的一个实例是BLAST算法,其在Altschul等人,J.Mol.Biol.215:403-410(1990)中描述。用于执行BLAST分析的软件可通过国家生物技术信息中心研究中心(National Center for Biotechnology Information)网站公开获得。BLAST算法参数W、T和X确定所述比对的灵敏度和速度。BLASTN程序(用于核苷酸序列)默认使用11的字长(W)、期望值(E)或10、M=5、N=-4以及两条链的比较。

[0053] 在一个或两个核苷酸序列的全长上,相同的序列包括包含第一核苷酸序列的多核苷酸与含有第二核苷酸序列的多核苷酸的100%同一性。这样的序列在本文中可以被称为相对于彼此“完全相同”。然而,在一些方面中,在第一序列在本文中被称为相对于第二序列“基本上相同”的情况下,这两个序列可以是完全互补的,或者在比对时它们可以具有一个或多个,但通常不超过约5个、4个、3个或2个错配的核苷酸。在一些方面,在第一序列在本文中被称为相对于第二序列“基本上相同”的情况下,这两个序列可以是完全互补的,或者它们可以是彼此至少约50%、60%、70%、80%、81%、82%、83%、84%、85%、86%、87%、88%、89%、90%、91%、92%、93%、94%、95%、96%、97%、98%或99%相同。为了确定本文所述的两种核苷酸序列的百分比同一性,可以使用上述BLASTN的默认设置。

[0054] 当第一序列在本文中被称为相对于第二序列的同一性“不同”时,在比对时这两个序列具有至少一个或多个错配的核苷酸。在一些方面,在比对时不同的序列可以具有2、3、4、5、6、7、8、9、10、11、12、13、14、15、16、17、18、19、20、21、22、23、24、25、26、27、28、29、30、31、32、33、34、35、36、37、38、39、40、41、42、43、44、45、46、47、48、49、50个或更多个错配核苷酸。在一些方面,不同的序列可以是彼此约1%、2%、3%、4%、5%、6%、7%、8%、9%、10%、11%、12%、13%、14%、15%、16%、17%、18%、19%、20%、21%、22%、23%、24%、25%、26%、27%、28%、29%、30%、31%、32%、33%、34%、35%、36%、37%、38%、39%、40%、41%、42%、43%、44%、45%、46%、47%、48%、49%、50%、60%、70%、80%、81%、82%、83%、84%、85%、86%、87%、88%、89%、90%、91%、92%、93%、94%、95%、96%、97%、98%、99%或小于100%相同。在一些方面,在第一序列在本文中被称为相对于第二序列“不同”的情况下,这两个序列可以具有基本上或完全相同的序列,但是基于序列内的不同修饰模式彼此不同。这样的修饰在本领域中通常是已知的,例如甲基化。

[0055] 在一些方面,多核苷酸可以存在于多核苷酸文库中。在一些方面,多核苷酸文库可以包括多个多核苷酸。在一些方面,所述多个多核苷酸中的每个多核苷酸可以来源于单个样品。在一些方面,单个样品是单个细胞。

[0056] 本文中使用常规符号来描述核苷酸序列:单链核苷酸序列的左端是5'-端;将双链核苷酸序列的左侧方向称为5'-方向。将核苷酸的5'至3'添加至新生RNA转录物的方向称为转录方向。将具有与mRNA相同的序列的DNA链称为“编码链”。

[0057] 术语“信使RNA”或“mRNA”是指不含内含子并可以翻译成多肽的RNA。

[0058] 术语“cDNA”是指以单链或双链形式与mRNA互补或相同的DNA。

[0059] 术语“扩增子”是指核酸扩增反应如RT-PCR的扩增产物。

[0060] 术语“杂交”是指与互补核酸的序列特异性非共价结合相互作用。杂交可以发生在核酸序列的全部或一部分上。本领域技术人员将认识到,核酸双链体或杂合体的稳定性可以由T_m确定。有关杂交条件的额外指导可以参见:Current Protocols in Molecular Biology, John Wiley&Sons, N.Y., 1989, 6.3.1-6.3.6和Sambrook等人, Molecular

Cloning, a Laboratory Manual, Cold Spring Harbor Laboratory Press, 1989, 第3卷。

[0061] 如本文所用,“区域”是指多核苷酸的核苷酸序列的连续部分。本文描述的区域实例包括鉴定区域、样品鉴定区域、条形码区域、结合区域、cDNA区域等等。在一些方面,多核苷酸可以包含一个或多个区域。在一些方面,多核苷酸可以包含小于2、2.3、4、5、6、7、8、9、10、11、12、13、14、15、16、17、18、19、20、21、22、23、24、25、26、27、28、29、30、31、32、33、34、35、36、37、38、39、40、41、42、43、44、45、46、47、48、49、50个或更多个区域。在一些方面,区域可以偶联。在一些方面,区域可以可操作地偶联。在一些方面,区域可以物理偶联。

[0062] 如本文所用,“条形码”或“条形码序列”是指可以与至少一个核苷酸序列偶联用于例如稍后鉴定所述至少一个核苷酸序列的任何独特的序列标记。

[0063] 如本文所用,“条形码组”或“条形码的组”是指可以与来自样品的核苷酸序列偶联的任何独特的序列组,其中核苷酸序列与所述组中的一个条形码序列偶联用于例如稍后鉴定核苷酸序列。

[0064] 术语“条形码构建体”、“条形码化构建体”和“条形码构建体分子”在本文中可互换地用以指包含独特条形码序列的寡核苷酸。

[0065] 术语“3'核酸构建体”或“3'核酸条形码构建体”是指包含条形码序列的寡核苷酸,所述条形码序列是“5'核酸条形码构建体”的3'。可以将3'核酸构建体并入核酸分子中,例如mRNA的3'端。术语“3'条形码”或“3'条形码序列”是指位于3'核酸构建体或3'核酸条形码构建体上的条形码。通常,3'核酸构建体含有与核酸分子的3'端的序列互补的序列,例如,3'核酸构建体包括与mRNA的poly-A序列互补的poly-T序列。典型地,3'构建体与mRNA序列的3'端结合并逆转录以产生在其5'端包含3'核酸构建体的cDNA分子。参见例如图2和图3。其还可以在5'端、3'端或内部并入核酸中(在将其并入核酸的3'端和另一核酸的5'端的情况下,形成一个更长的核酸)。参见例如图12。除了条形码序列之外,3'核酸条形码构建体可以包括通用引发序列或通用引发区域和结合位点。

[0066] 术语“5'核酸构建体”或“5'核酸条形码构建体”是指包含条形码序列的寡核苷酸,所述条形码序列是“3'核酸条形码构建体”的5'。可以将5'核酸构建体并入核酸分子中,例如mRNA的5'端或cDNA的3'端。术语“5'条形码”或“5'条形码序列”是指位于5'核酸构建体或5'核酸条形码构建体上的条形码。通常,5'核酸构建体含有与核酸分子的3'端的序列互补的序列,例如,5'核酸构建体包括与并入cRNA序列的3'端的poly-C序列互补的poly-G序列。可以将5'核酸构建体并入包含3'核酸构建体的cDNA分子的3'端。参见例如图2和图3。其还可以在5'端、3'端或内部并入核酸中(在将其并入核酸的3'端和另一核酸的5'端的情况下,形成一个更长的核酸)。参见例如图12。除了条形码序列之外,5'核酸条形码构建体可以包括通用引发序列或通用引发区域和结合位点。

[0067] 如本文所用,“条形码构建体珠粒”是指与一个或多个条形码构建体偶联的珠粒。

[0068] 如本文所用,“条形码编码混合物”是指包含一个或多个条形码构建体的组合物。所述混合物可以包含用于裂解细胞、逆转录、限制性酶消化和/或进行条形码编码反应的额外试剂。

[0069] 如本文所用,“条形码编码”或“条形码编码反应”是指将条形码序列或条形码序列的互补序列与核酸连接的反应。条形码构建体不一定与核酸共价连接,但条形码序列信息本身与核酸连接或并入核酸中。“条形码编码核酸”、“条形码编码细胞”、“条形码编码来自细胞的核

酸”、“条形编码来自反应容器的核酸”和“条形编码反应容器”可互换使用。

[0070] 如本文所用,“鉴定区域”是指可以与至少一个核苷酸序列偶联以例如稍后鉴定所述至少一个核苷酸序列的核苷酸序列标记(例如,独特的条形码序列)。在一些方面,将条形码组用作样品鉴定区域。

[0071] 如本文所用,“接头区域”或“接头分子”是指将第一核苷酸序列与第二核苷酸序列偶联的连接子。在一些方面,接头区域可以包括充当连接子的核苷酸序列的连续部分。

[0072] 术语“样品”可以包括取自受试者(例如,哺乳动物受试者、动物受试者、人受试者或非人受试者)的RNA、DNA、单个细胞或多个细胞或细胞片段或体液等分试样。本领域技术人员可以使用现在已知或后来发现的任何方式来选择样品,所述方式包括离心、静脉穿刺、抽血、排泄、擦拭、射精、按摩、活检、针抽吸、灌洗样品、刮擦、手术切口、激光捕获显微切片、梯度分离或干预或本领域已知的其他方式。本领域技术人员还可以使用已知与所关注的样品相关联的一种或多种标志物来选择样品。还可以使用本领域已知的方法如细胞分选和FACS来选择样品。

[0073] 本发明的方法可以用任何期望的样品来实施。在一些实施方案中,每种样品包含细胞,并且可以是例如单个细胞。细胞可以封闭在反应体积如微流液体滴中,并且如果需要,则可以裂解细胞以将RNA分子释放到反应体积中。为此目的,可以使细胞在任何方便的时间与裂解缓冲液接触。

[0074] 本文提供了鉴定核酸样品来源的方法。任选地,所述方法包括形成反应混合物,所述反应混合物包含核酸样品,所述核酸样品包含来自单个细胞的核酸分子和条形码组;将所述条形码组并入所述样品的核酸分子中;以及鉴定并入所述单个细胞的核酸分子中的所述条形码组,从而鉴定所述核酸样品的来源。任选地,形成所述反应混合物包括使所述单个细胞与所述条形码组接触并裂解所述细胞以提供来自单个细胞的核酸样品。任选地,形成所述反应混合物包括提供在悬浮缓冲液中的细胞和使包含所述细胞的悬浮缓冲液与包含所述条形码组的条形编码缓冲液在一定条件下接触以形成包含所述单个细胞和所述条形码组的反应混合物。任选地,所述反应混合物是液滴。

[0075] 如本文所提供,条形码组包含至少两个独特条形码的一个或多个拷贝。因此,例如,条形码组可以包含各自包含独特的5'条形码的5'核酸构建体中的一个或多个和各自包含独特的3'条形码的3'核酸构建体中的一个或多个。任选地,所述条形码组包含各自包含独特的3'条形码的至少两种3'核酸构建体。任选地,所述条形码组包含1、2、3、4、5、6、7或8种5'核酸构建体和1、2、3、4、5、6、7或8种3'核酸构建体。1、2、3、4、5、6、7或8种类型的核酸构建体中的每一种具有独特的条形码序列并且可以以多个拷贝提供。任选地,所述条形码组包含包含两个不同的5'条形码的两种不同的5'核酸构建体和包含两个不同的3'条形码的两种不同的3'核酸构建体。任选地,所述条形码组包含包含三个不同的5'条形码的三种不同的5'核酸构建体和包含三个不同的3'条形码的三种不同的3'核酸构建体。任选地,所述条形码组包含包含两个不同的3'条形码的两种不同的3'核酸构建体和包含三个不同的5'条形码的三种不同的5'核酸构建体。可以提供每种类型的核酸构建体的多个拷贝,相同类型的多个拷贝各自具有相同的条形码。因此,例如,条形码组可以包括1、2、3、4、5、6、7或8种不同类型的5'核酸构建体或3'核酸构建体的多个拷贝,不同类型各自具有独特的5'或3'条形码序列。换言之,条形码组可以包含第一、第二、第三、第四、第五、第六、第七或第八类型

的5'核酸构建体中的一种或多种的一个或多个拷贝以及第一、第二、第三、第四、第五、第六、第七或第八类型的3'核酸构建体中的一种或多种的一个或多个拷贝。任选地,条形码组包含第一、第二、第三、第四、第五、第六、第七或第八类型的3'核酸构建体中的一种或多种的一个或多个拷贝。构建体可以是RNA或DNA分子或RNA-DNA杂合体。例如,构建体可以包括共价连接到共同的寡核苷酸链中的DNA核苷酸的RNA核苷酸。构建体也可以是单链的或双链的。如果是双链的,构建体可以具有一个或多个平端或单链悬伸端。

[0076] 本文还提供了一种使用3'核酸构建体组鉴定核酸样品来源的方法。所述方法包括提供3'核酸构建体组,每种构建体包含独特的3'条形码序列、结合序列和条形码关联序列;提供包含来自单个细胞的核酸分子的核酸样品;使核酸样品与所述3'核酸构建体组接触;将3'条形码序列和结合序列与条形码关联序列分离;将来自3'核酸构建体的3'条形码序列并入核酸分子中;以及使用条形码关联序列鉴定所述3'条形码序列组,从而鉴定来自单个细胞的核酸样品的来源。任选地,所述3'核酸构建体组包含2、3、4、5、6、7或8种3'核酸构建体。任选地,将3'核酸构建体的多个拷贝附接至固体表面如珠粒、微珠、微粒、微球、纳米珠粒或水凝胶。任选地,每个珠粒包含相同的核酸构建体,并且可以在条形码组中提供多个珠粒。因此,任选地,每种固体表面(例如,珠粒)含有相同类型的3'核酸构建体的多个拷贝。因此,例如,条形码组可以包括第一表面和第二表面,所述第一表面包含含有独特3'条形码的第一3'核酸构建体的多个拷贝,所述第二表面包含含有与第一构建体上的条形码不同的独特3'条形码的第二3'核酸构建体的多个拷贝。

[0077] 任选地,形成反应混合物包括提供在悬浮缓冲液中的细胞,并使悬浮缓冲液与包含5'和/或3'核酸构建体的条形编码缓冲液在一定条件下接触以形成包含单个细胞的反应混合物。反应混合物包含至少两种构建体,每种构建体具有独特的条形码和单个细胞。任选地,反应混合物包含至少一种5'核酸构建体和至少一种3'核酸构建体。任选地,反应混合物包含至少两种3'核酸构建体。任选地,形成反应混合物包括提供在悬浮缓冲液中的细胞,以及使悬浮缓冲液与包含5'核酸构建体的第一条形编码缓冲液和包含3'核酸构建体的第二条形编码缓冲液在一定条件下接触以形成包含单个细胞和至少一种5'核酸构建体以及至少一种3'核酸构建体的反应混合物。任选地,反应混合物包含来自单个细胞的核酸样品、包含5'核酸构建体的多个拷贝的至少一个珠粒和包含3'核酸构建体的多个拷贝的至少一个珠粒。

[0078] 在裂解之前,可以将细胞有利地悬浮在包含渗透保护剂的细胞悬浮缓冲液中。渗透保护剂可以保护细胞免受渗透应力,并确保细胞生理在条形编码之前保持稳定或无干扰。在一些实施方案中,细胞与条形码构建体一起悬浮在细胞悬浮缓冲液中。在一些实施方案中,在与用于逆转录、PCR和/或裂解的试剂接触之前,将细胞悬浮在细胞悬浮缓冲液中。细胞悬浮缓冲液可以包含在任何反应体积中并且与本文所述的用于形成和组合水性反应体积的方法相容。

[0079] 在与用于逆转录、PCR和/或裂解的试剂接触之前,将细胞悬浮在细胞悬浮缓冲液中。细胞悬浮缓冲液可以包含在任何反应体积中并且与本文所述的用于形成和组合水性反应体积的方法相容。

[0080] 任选地,每种5'核酸构建体以附接到固体表面的多个拷贝提供。任选地,每种3'核酸构建体以附接到固体表面的多个拷贝提供。具有可用于本发明的方法和组合物中的固体

表面的固体载体的实例包括珠粒、色谱树脂、多孔板、微量离心管或具有固体表面的任何其他物体。任选地,固体表面是珠粒、微珠、微粒、微球、纳米粒子、纳米珠粒或水凝胶。因此,可以使用任何期望的机制或捕获化学如生物素-抗生物素蛋白、生物素-链霉亲和素或金-硫醇相互作用将条形码构建体结合到固体载体上。任选地,使条形码构建体附接的任何固体表面与水溶液接触,并且将条形码构建体释放到该溶液中。所述水溶液可以与和例如来自单个细胞的样品相关联的RNA分子处于相同的反应体积,向其中添加条形码构建体。另选地,可以将接触用于条形码构建体的固体表面的水溶液容纳在与目标RNA或细胞不同的反应体积中,并且可以在组合两个反应体积时将条形码构建体添加到这些RNA或细胞中以产生包含条形码组和单个细胞或来自单个细胞的RNA样品的反应混合物。

[0081] 提供用于将多核苷酸附接至固体载体的方法,其中所述多核苷酸含有条形码序列。多核苷酸可以是条形码构建体。任选地,所述方法包括产生反相乳液的亲水性隔室(即,水性液滴)。可以根据需要例如通过将水溶液混入疏水性载体流体中并任选地搅拌混合物而产生隔室。水溶液可以具有固体载体、寡核苷酸和悬浮于其中的试剂,使得每个隔室含有用于在形成隔室时将多核苷酸附接至固体载体的所有必需组分。在这些实施方案中,在将固体载体添加至隔室中之前,寡核苷酸经由捕获部分结合至固体载体的表面。该寡核苷酸在本文中称为“结合寡核苷酸”并且含有与条形码寡核苷酸的3'序列互补的3'序列。由此通过涉及结合的寡核苷酸和条形码寡核苷酸的聚合酶延伸反应在固体载体上形成多核苷酸,并且该反应在隔室内发生。

[0082] 任选地,当形成亲水性隔室时,条形码寡核苷酸以低浓度或极限浓度(例如,1个分子/隔室)存在。当使用具有随机化序列的条形码寡核苷酸的文库制备多个条形码构建体珠粒时,该浓度是方便的。如果假定每个条形码寡核苷酸具有不同的条形码序列,并且期望每个隔室中的固体载体仅具有一个条形码序列,则每个隔室可以存在一个条形码寡核苷酸(最多或平均)。一旦满足该条件,多个固体载体(例如,多个珠粒)可以存在于隔室中,或者结合的寡核苷酸的多个拷贝可以与每个固体载体结合,但是由隔室中的聚合酶延伸反应产生的所有多核苷酸将含有相同的条形码序列。

[0083] 用于本发明方法的固体载体包括珠粒,例如由金属和/或聚合物材料制成并且直径在约0.1微米至10微米范围内的球形珠粒。可以替代或此外使用具有其他特性的珠粒。固体载体可以用捕获部分官能化以将结合的寡核苷酸附接到表面。捕获部分的实例包括抗生物素蛋白、链霉亲和素、生物素、羧基基团、环氧基团、羟基基团、硫醇基团和金。一些捕获部分具有特异性和非共价结合的结合伴侣。例如,链霉亲和素采用生物素作为其结合伴侣。这样的捕获部分可以直接(例如,共价地)偶联至固体载体,并且结合伴侣可以偶联至结合的寡核苷酸,或者反之亦然,使得结合的寡核苷酸通过非共价相互作用与固体载体结合。其他捕获部分提供在结合的寡核苷酸和固体载体之间的直接共价连接。

[0084] 结合的寡核苷酸优选是在其5'端结合到固体载体的单链DNA分子。因此,结合的寡核苷酸的3'端在溶液中是游离的,并且当与条形码寡核苷酸杂交时,可以通过酶如DNA聚合酶延伸。延伸反应使用条形码寡核苷酸模板化,使得条形码序列并入与珠粒结合的DNA链中。如果需要,结合的寡核苷酸和/或条形码寡核苷酸则可以具有设计成使分子内二级结构最小化的序列。

[0085] 本文所述的用于将多核苷酸附接到固体载体上的任何方法可以用于制备用于条

形编码样品、细胞或RNA的一种或多种固体载体。附接到每种固体载体上的一种或多种多核苷酸包括条形码序列并且可以用作条形码构建体。本发明方法还可以用于制备条形码文库,所述条形码文库包括多种固体载体,其各自与条形码序列相关联。任何两种固体载体(例如,珠粒)可以具有彼此全部或部分地不同的条形码序列。任选地,条形码文库中的每种固体载体都与不同的条形码序列相关联。

[0086] 根据本发明方法制备的条形码构建体珠粒包括与条形码构建体结合的珠粒。所述珠粒可以与条形码构建体的多个拷贝结合,例如至少10、30、100、300、1,000、3,000、10,000、30,000、100,000、300,000或1,000,000个拷贝。任选地,与一个珠粒结合的条形码构建体的每个拷贝包含相同的条形码序列。本发明方法还允许制备包含多种条形码构建体珠粒的珠粒化条形码文库。文库中的每种珠粒可以与不同的条形码序列相关联,并且每种珠粒上的条形码构建体的拷贝可以包含相同的条形码序列。

[0087] 任选地,本发明方法可以用于通过物理捕获制备自或获自条形码构建体珠粒上的一个或多个样品(例如,细胞)的cDNA来制备多核苷酸文库。每种珠粒包括在3'端具有cDNA结合位点的条形码构建体。可以使珠粒与酶接触以使结合位点成为单链(例如,保持模板分子的末端的3'悬伸部在溶液中游离)。然后使珠粒与来自样品的一种或多种cDNA接触,使得cDNA通过结合位点结合模板分子的拷贝。在优选的实施方案中,结合位点包括一个或多个G核苷酸,例如poly-G道,并且与通过逆转录酶添加到cDNA的末端的非模板化poly-C道互补。

[0088] 可以根据需要使用多核苷酸文库中的珠粒,例如以对来自多个样品的cDNA进行测序或从不同样品中分离cDNA。在后一种情况下,可以使用离心或磁力将对应于不同样品的珠粒制粒,然后使用标准方法重新悬浮和分离。如果需要,则在将cDNA与珠粒上的模板分子结合之后,可以将模板分子酶促延伸,由此将cDNA序列并入结合珠粒的DNA双链体并将这些序列与条形码序列相关联。如果来自样品的cDNA分子的拷贝数与珠粒上条形码构建体的拷贝数相当,则可以将这些cDNA分子捕获在少量珠粒上(例如,至多约1、3、10、30、100、300或1000个珠粒/样品)。可以使用标准方法或如上所讨论逆转录来自样品的RNA以产生cDNA。

[0089] 可以使用各种方法释放5'和3'核酸构建体,在它们结合到固体表面时允许将包含条形码的构建体并入核酸样品的核酸中。例如,每种5'和3'核酸构建体可以包含可光裂解的连接子或限制性酶位点。任选地,每种5'核酸构建体还包含可光裂解的连接子,并且每种3'核酸构建体还包含限制性酶位点。任选地,每种5'核酸构建体还包含限制性酶位点,并且每种3'核酸构建体还包含可光裂解的连接子。任选地,所述方法还包括使反应混合物与限制性酶接触并将反应混合物暴露于紫外光以从固体表面释放5'和3'核酸构建体。任选地,5'和3'核酸构建体还包含限制性酶位点,并且所述方法还包括使反应混合物与限制性酶接触以从固体表面释放5'和3'核酸构建体。任选地,5'和3'核酸构建体还包含可光裂解的连接子,并且所述方法还包括将反应混合物暴露于紫外光以从固体表面释放5'和3'核酸构建体。

[0090] 如上所述,任选地,3'核酸构建体包含独特的3'条形码序列、结合序列和条形码关联序列。任选地,3'核酸构建体能够形成发夹。任选地,3'条形码序列和结合序列通过限制性酶序列或可光裂解的连接子与条形码关联区段分开。任选地,限制性酶从固体表面释放3'条形码序列,并将结合序列与条形码关联区段分离。任选地,紫外光从固体表面释放3'条形码序列,并将结合序列与条形码关联区段分离。任选地,3'核酸构建体上的结合序列包含

poly-T序列。任选地,3'核酸构建体的结合序列结合样品的核酸分子上的poly-A序列。任选地,3'核酸构建体上的条形码序列通过逆转录并入。任选地,逆转录生成包含3'条形码序列的核酸分子。任选地,条形码关联区段包括引发序列。任选地,所述引发序列是回文序列。任选地,一个条形码关联区段的引发序列能够结合核酸样品中任何其他条形码关联区段的引发序列。任选地,彼此接触的条形码关联区段通过DNA聚合酶延伸以将条形码关联区段两者并入单个核酸中。

[0091] 限制性酶和限制性酶位点是已知的,并且这些酶是可商购获得的。限制性酶消化是指用仅在核酸分子中特定位置发挥功能的酶催化裂解核酸分子。这样的酶的用途为本领域技术人员所理解。因此,5'和3'核酸构建体可以含有在暴露于适当的酶(例如,限制性内切核酸酶)时促进构建体裂解的内切核酸酶限制性位点。任选地,限制性酶选自由Nb.BsrDI、Nb.BtsI、Nt.BbvCI、Nt.BspQI、Nt.BsmAI、Nt.BstNBI、Nt.AlwI和Nt.BsmAI组成的组。任选地,附接到固体表面的5'或3'核酸构建体是双链的,并且构建体的仅一条链从固体表面释放。因此,任选地,5'和/或3'核酸构建体可以包含切口内切核酸酶限制性位点,并且可以使用一种或多种切口内切核酸酶释放所述链。切口内切核酸酶限制性位点可以是对选自Nb.BsrDI、Nb.BtsI、Nt.BbvCI、Nt.BspQI、Nt.BsmAI、Nt.BstNBI、Nt.AlwI和Nt.BsmAI的酶具有特异性的位点。任选地,可以包含与限制性内切酶一起的链置换DNA聚合酶。链置换聚合酶可以在切口内切核酸酶之前,与其同时或在其之后添加。链置换DNA聚合酶可以选自由Klenow exo-、Bst Large Fragment和Bst Large Fragment的工程化变体组成的组。

[0092] 可光裂解的连接子及其使用方法也是已知的。如本文所用,术语“可光裂解的连接子”是指附接或可操作地连接两个分子如两个核酸或一个核酸和一个固体表面的任何化学基团。任选地,连接子包含2-硝基苄基部分。参见,例如,美国专利5,643,722,其全部内容通过引用并入本文。可光裂解的连接子还包括但不限于 α -取代的2-硝基苄基部分[例如,1-(2-硝基苄基)乙基部分]、3,5-二甲氧基苄基部分、硫代异羟肟酸、7-硝基吡啶啉部分、9-苯基黄嘌呤酸基部分、苯偶因部分、羟基苯甲酰甲基部分和NHS-ASA部分。连接子可以通过暴露于适当波长的光如紫外光而被裂解,以从底物收获核酸分子(参见J.Olejnik和K.Rothschild,Methods Enzymol 291:135-154,1998)。

[0093] 任选地,所提供的构建体可以包括独特的分子标识符(UMI)序列。任选地,UMI序列含有随机化的核苷酸并独立于条形码序列并入条形码构建体中。因此,含有相同条形码序列的条形码构建体组可以含有不同的UMI序列。当将含有相同条形码序列但不同UMI序列的条形码构建体组添加到与一个样品相关的RNA中时,每个RNA序列可以在条形码编码期间与不同的UMI序列连接。可以产生具有UMI序列的条形码构建体珠粒,其中每个珠粒上的构建体含有相同的条形码序列和不同UMI序列的文库。

[0094] 除条形码序列之外,条形码构建体可以包括通用引发序列或通用引发区域和结合位点。因此,任选地,3'核酸构建体和5'核酸构建体通常包括结合序列。任选地,5'和3'核酸构建体包含引物结合位点或通用引发序列。任选地,3'核酸构建体具有包含poly-T序列的结合序列。poly-T序列结合样品的核酸分子上的poly-A序列。在3'核酸构建体与核酸分子结合后,3'核酸构建体上的条形码然后可以通过逆转录并入。逆转录生成包含3'条形码序列的核酸分子。因此,3'核酸构建体可以充当用于逆转录的引物。这里,构建体的结合位点是RNA结合位点(例如,mRNA结合位点),并含有与一个或多个RNA中的序列区域互补的序列

区域。任选地,结合位点与添加有条形码构建体的样品中的所有RNA共有的序列区域互补。例如,结合位点可以是poly-T道,其与真核mRNA的poly-A尾部互补(图3)。可选地或此外地,结合位点可以包括随机序列道。在将3'构建体添加至与样品相关联的RNA中时,可以发生逆转录并且可以合成cDNA的第一链,从而将3'条形码序列并入cDNA的第一链中。应该认识到,逆转录需要适当的条件,例如存在适当的缓冲液和逆转录酶以及适合条形码构建体退火至RNA的温度以及酶的活性。还将认识到,当引物的3'端与模板互补并且可以直接退火至模板时,涉及DNA引物和RNA模板的逆转录是最高效的。因此,可以设计3'核酸构建体,使得结合位点在RNA的3'端发生。

[0095] 如上所讨论,任选地,使用包含条形码、结合序列和条形码关联序列的3'核酸构建体组。任选地,3'核酸构建体包含引物结合位点或通用引发序列。每种构建体可以是形成发夹结构的单链DNA寡聚物。任选地,构建体是线性的。如果发夹用于构建体,则可以将限制性位点并入发夹的双链区段中,所述限制性位点导向两条链的限制。这将条形码和结合区域从可能还含有条形码的单独环状区段以及“条形码关联序列”(BAS)中裂解并分离。参见例如图8。线性寡聚物可以与多个限制性位点一起使用,这也引起从珠粒上切下两个或更多个区段。条形码区域包括条形码、poly-T区段和任选的UMI。该区段退火至由液滴内的细胞释放的mRNA转录物的poly-A尾部并引发该mRNA序列的逆转录,从而将条形码并入生成的cDNA中。参见例如图9。单独地,环状区段包含其自己的条形码,所述条形码与裂解其的条形码相同或相对应。当多个珠粒与细胞共封装时,两个转录物条形编码区段(BC1a和BC2a,图10)都将作为RT条形编码反应的引物,并且来自该细胞的转录组将合并两个条形码。单独地,还会存在携带其各自条形码的两个相应的环状区段(BC1b和BC2b,图10)。BAS上的结合区域将使BC1b和BC2b环状区段实现交叉退火,并且DNA聚合酶可以包含在反应混合物中以延伸每个区段,从而合并其他条形码的补体(图11)。通过这种方式,BAS'物理连接在单个扩增子中,然后可以在下游进行测序,以便将条形码作为一组相关联以将统一来自该细胞的转录组。BAS上的结合序列可以是回文序列,或者每个BAS可以包括与如在实施例中进一步描述的另一BAS上的序列互补的序列。

[0096] 当使用5'和3'核酸构建体时,一旦3'核酸构建体已经并入cDNA的第一链中,则任选地,逆转录酶将2至5个半胱氨酸残基加到样品中的核酸分子的末端。这2至5个半胱氨酸残基提供与5'核酸构建体上的结合序列互补的结合序列。任选地,5'核酸构建体上的结合序列包含poly-G序列。5'核酸构建体上的结合序列然后可以结合样品的核酸分子上的poly-C序列。使用逆转录酶将5'核酸构建体并入核酸样品中。参见例如图3。至少在W02012/148497中已经描述了该方法,该申请的全部内容通过引用并入本文。简言之,H-MMLV逆转录酶具有3'dC拖尾活性,并将非模板化dC加到第一链cDNA。如果还存在以至少1个G结尾的条形码化构建体,例如5'核酸构建体,则该构建体可以与第一链cDNA的3'dC碱基配对,并且逆转录酶经历模板切换并继续转录。逆转录酶因此通过磷酸二酯键将条形码序列共价添加至第一链cDNA的3'端。例如如图7所示,所得构建体是通常包含5'至3'序列的核酸分子:5'条形码-cDNA序列-3'条形码。任选地,进一步扩增核酸样品以生成各自包含5'条形码和3'条形码的多个核酸序列的多个拷贝。任选地,5'和3'核酸构建体包含用于扩增包含5'和3'条形码的核酸序列的通用引发位点。

[0097] 任选地,逆转录酶是M-MLV酶。任选地,逆转录酶是MMLV H-逆转录酶。

[0098] 如上所述,如果裂解试剂存在于液滴中以从细胞中释放RNA,并且如果存在逆转录酶、引物和其他适当的试剂,则可以发生逆转录。用于促进裂解和逆转录的酶和试剂可以例如通过将含有酶和试剂的液滴与含有条形码和细胞的液滴合并来一次性添加,或者可以分步添加。

[0099] 任选地,5'核酸构建体是包含启动子的双链核酸构建体,并且其中5'核酸构建体用作模板以生成5'核酸构建体的多个拷贝。任选地,启动子是RNA聚合酶启动子。参见例如图7。因此,任选地,5'核酸是具有以下序列的双链DNA(dsDNA)模板分子:5'-T7启动子-通用引发序列-条形码序列-结合序列-3'。T7启动子序列允许通过T7RNA聚合酶由模板合成5'核酸构建体。通用引发序列用于与下游使用的PCR引物互补。结合序列由1个或多个鸟嘌呤碱基(G)组成,并且允许条形码化构建体与第一链cDNA的3'端互补碱基配对。可以使用其他启动子序列,例如但不限于T3和SP6启动子序列,其允许分别通过T3和SP6RNA聚合酶合成RNA条形码化构建体。还可以使用不具有特异性启动子序列的其他RNA聚合酶,只要在大部分情况下合成了全长或接近全长的5'核酸构建体即可。

[0100] 提供的核酸样品来自单个细胞,并且与样品相关联的RNA可以包括mRNA。样品可以包含例如至少1、3、10、30、100、300、1,000、3,000、10,000、30,000、100,000、300,000或1,000,000个mRNA分子,其可以表示任何数目的基因、等位基因、阅读框或不同的序列。任选地,与样品相关联的RNA包括来自样品的所有mRNA、细胞的全部或部分转录组或来自细胞的总RNA。

[0101] 不受任何理论的约束,本发明方法对每个样品可以条形编码的RNA的数量没有限制。因此,每个样品生成的关注多核苷酸的数目可以是至少10、30、100、300、1,000、3,000、10,000、30,000、100,000、300,000或1,000,000个。每种关注的多核苷酸可以以多个拷贝存在。另外,可以在所述方法的一次执行中条形编码的细胞或样品的数量仅受限于用条形码序列的独特组合制备多组条形码的挑战(上文讨论)。在一些实施方案中,所述一个或多个样品包含至少10、30、100、300、1,000、3,000、10,000、30,000、100,000、300,000或1,000,000个细胞。样品(例如,每个样品是单个细胞)可以从相同的受试者或不同的受试者获得。例如,至少2、3、4、5、6、7、8、9、10、20、30、40、50、60、70、80、90或100个不同的受试者可以提供样品。

[0102] 核酸样品中5'和3'条形码的组合鉴定来自单个细胞的核酸样品的来源。任选地,通过测序鉴定核酸样品中的条形码。预期每个条形码序列与来自不同样品的组装重叠群相关联,因为条形码序列被重新用于独特的条形码组中。如本文所用的术语重叠群是指重叠序列数据(读数)。可以观察到相同的重叠群以使用条形码序列a、b和c。并且可以观察条形码序列a、b和d以与另一重叠群相关联。由此可以得出结论:a、b和c包含条形码组1,且a、b和d包含条形码组2。条形码组1与来自一个细胞的核酸样品相关联,而条形码组2与来自第二不同细胞的核酸样品相关联。由于组中的所有5'条形码都与一组中的所有3'条形码相关联,因此它们可以在图论中表示为最大二分聚类(参见图13)。最大二分聚类将使条形码组合核酸与单个细胞相关联。

[0103] 还提供了使用本文公开的方法产生的组合物。因此,本申请提供用于其产生的RNA和DNA构建体的组合物。还提供了条形码构建体珠粒文库、装载有RNA条形码构建体的乳液液滴文库、具有细胞的含有条形码构建体文库的乳液、条形码化cDNA文库和微流体微滴产

生装置等等。

[0104] 提供了包含条形码化珠粒群的组合物。所述组合物包含多个珠粒，每个珠粒具有附接到其表面的5'或3'核酸构建体的多个拷贝。因此，提供了具有多个珠粒的组合物，每个珠粒具有附接到其表面上的5'核酸构建体的多个拷贝。还提供了包含多个珠粒的组合物，每个珠粒具有附接到其表面上的3'核酸构建体的多个拷贝。任选地，所述组合物是条形编码混合物，其在适当地添加到包含细胞的悬浮缓冲液时产生包含至少一个包含5'核酸构建体的珠粒、至少一个包含3'核酸构建体的珠粒和单个细胞的反应混合物。优选地，所述反应混合物包含至少三种不同的核酸构建体，例如两种5'核酸构建体和一种3'核酸构建体或两种3'核酸构建体和一种5'核酸构建体。

[0105] 本文进一步公开了包含本文所述的提供的构建体的试剂盒。试剂盒可以包含与本文所述的条形码构建体偶联的多个固体载体。任选地，所述试剂盒包含一个或多个条形码构建体文库，每个文库包含具有相同条形码的多个条形码构建体。例如，提供了一种包含条形码珠粒的文库的试剂盒，所述条形码珠粒包含5'和3'核酸构建体，包括1、2、3、4、5、6、7、8种或更多种类型的5'和/或3'核酸构建体。任选地，试剂盒包含本文所述的条形编码混合物和/或细胞悬液缓冲液。试剂盒可以包括使用说明。试剂盒可以在合适的容器中包含本文公开的珠粒、构建体或文库，一种或多种对照物，以及本领域公知的各种缓冲液、试剂、酶和其他标准成分。因此，试剂盒还可以包含用于进行细胞裂解、限制性酶消化和条形编码反应的酶或任何其他试剂。

[0106] 容器可以包括在包括一个或多个孔的板上的至少一个孔。容器可以包括至少一个小瓶、试管、烧瓶、瓶子、注射器、液滴或其他容器装置，所提供的试剂、核酸构建体、珠粒、文库或细胞可以放置于其中，并且在一些情况下，适当地等分。在提供额外组分的情况下，试剂盒可以含有可以放置该组分的额外容器。试剂盒还可以包括用于容纳试剂、核酸构建体、珠粒、文库或细胞以及严格限制用于商业销售的任何其他试剂容器的装置。这样的容器可以包括注塑或吹塑的塑料容器，其中期望的小瓶保留在其中。容器可以包括具有使用说明和/或警告的标记。

[0107] 还提供了用于产生和运输反应体积的装置，例如在U.S.S.N.14/586,857中公开的那些，U.S.S.N.14/586,857的全文通过引用并入本文。这些体积可以以微流体尺度存在并且可以与载体流体相分离。可以由装置处理的反应体积的实例包括在反相乳液(即，水/油乳液)中的水性液滴。这些装置允许条形码构建体、样品(例如，细胞)和/或从这些样品获得的RNA单独地或一起封装在液滴中。所述装置还允许将试剂引入液滴中，从而可以酶促产生条形码构建体分子，并且可以对来自个别样品的RNA进行条形编码。

[0108] 装置通常包括三个微流体路径，每个微流路径偶联到压力源和流量传感器。微流体路径的压力源驱动流体通过所述路径，并且可以使用在压力源下游存在的流量传感器来测量通过该路径的流速。在一些实施方案中，第一路径和第二路径在第一连接点合并以形成组合路径。参见例如图1。

[0109] 如本文所述的装置可以由可购自IDEX Corporation(Lake Forest, Illinois, U.S.A.)的管道和流体组件并使用可购自Dolomite Microfluidics(Charlestown, Massachusetts, U.S.A.)的微流体液滴芯片组装。微流体液滴芯片的一些特征在美国专利7,268,167、7,375,140、7,717,615、7,772,287、8,741,192和8,883,864中描述，这些专利通

过引用并入本文。合适的压力源包括注射泵和压力泵。压力泵可购自Dolomite Microfluidics。可以独立地控制压力源。

[0110] 任选地,第一和第二微流体路径输送水溶液。每个路径可以包括注射口和阀(例如,四通阀)以在内嵌在路径中的注射口中引入溶液。任选地,容纳水性载体流体的贮存器安置在每个四通阀的上游。当载体流体向下游驱动时,水性载体流体可以在四通阀中与水溶液混合,或者将水溶液的塞子向下游推向第一接合点。任选地,流阻器安置在每个微流体路径中。

[0111] 一旦将水溶液引入第一或第二微流体路径中,其就可以经过样品环路,所述样品环路计量流向第一接合点的溶液流。可以根据需要例如流体阻力或使用沿样品环路安置的阀实现计量。任选地,一个样品环路与第一和第二微流体路径中的每一个相关联,并且样品环路与热冷却单元接触。可以包括热冷却单元以防止水溶液中的酶、核酸或其他生物组分的热变性,或为酶反应确立最佳温度。可以独立地或共同地控制与用于第一和第二微流体路径的样品环路接触的热冷却单元的部分。可以使用任何物质或设备作为热冷却装置,条件是它能使通过样品环路的水溶液的温度偏离环境温度。合适的热冷却装置的实例是帕耳帖(Peltier)装置和冰箱。

[0112] 任选地,通过第一微流体路径输送的水溶液含有细胞和条形码构建体珠粒。在一些实施方案中,通过第二微流体路径运输的水溶液含有用于细胞裂解的试剂和用于生成关注的多核苷酸的试剂(例如,用于生成条形码构建体分子的酶)。与每个微流体路径相关联的注射口、阀和/或样品环路可以构造或定制成容纳经过该路径的水溶液的内容物。例如,与第一微流体路径相关联的样品环路可以具有扩大的内径以容纳细胞和珠粒。应认识到在第一微流体路径和第二微流体路径之间存在用于分配细胞、珠粒和试剂的许多其他选择,使得所有这些组分在第一接合点组合。例如,细胞可以通过第一微流体路径运输,并且珠粒可以通过第二微流体路径运输。鉴于其携带的水溶液的内容物,可以根据需要构造每个路径。

[0113] 由组合路径的合并产生的含有液滴的流体路径构成样品路径。样品路径被递送到样品收集容器,该容器存在于第二接合点的下游。在样品收集容器中,液滴可以经受热循环。液滴还可以被打开并且可以收获条形码化核酸。

[0114] 在操作中,可以使用装置以将条形码构建体珠粒和细胞封装到水性微流体液滴中,使得每个液滴平均含有大约两个珠粒和一个细胞。每个液滴中珠粒和细胞的数量可以根据需要例如通过调整装载到装置中的溶液中的珠粒或细胞的浓度或通过调整三个微流体路径中的流速来调谐。包含在每个液滴中的试剂允许条形码构建体分子由液滴中的一个珠粒酶促产生。这些试剂还允许一个细胞被裂解并且来自细胞的RNA经历条形编码反应。因此,来自细胞的RNA可以在液滴内条形编码,并且当混合来自多个细胞的核酸时,来源于这些RNA(并且含有条形码序列)的核酸可以稍后追溯到一个细胞。

[0115] 公开了材料、组合物和组分,其可以用于所公开的方法和组合物的产物、可以与所述产物结合使用,可以用于制备所述产物;以及所公开的方法和组合物的产物。本文公开了这些材料及其他材料,并且应理解,若公开了这些材料的组合、子组、相互作用、群组等,而没有明确地具体提及各种个体和集体组合的每一种组合以及这些化合物的排列方式,则每种情况均具体地涵盖并描述在本文中。例如,如果公开并讨论了方法并且讨论了可以对包

括所讨论的方法的多个分子进行的多种修改,除非明确地相反指示,否则明确涵盖了所述方法的每一种组合和排列以及可能的修改。同样地,还具体地涵盖并公开了这些分子的任何子组或组合。这个概念适用于本公开的所有方面,包括但不限于使用所公开的组合物中的方法中的步骤。因此,如果有多种额外步骤可以进行,则应理解,这些额外步骤中的每一个可与所公开的方法的任何具体方法步骤或方法步骤的组合一起进行,并且具体地涵盖每种此类组合或组合的子组并应将其视为得到公开。

[0116] 本文引用的出版物和引用它们的材料特此通过引用整体明确地并入。

[0117] 已经描述了多个实施方案。然而,应理解,可以做出各种修改。因此,其他实施方案在下面的权利要求的范围内。

[0118] 实施例

[0119] 实施例1.用于分析与单个细胞相关联的核酸的核酸条形码的组。

[0120] 使用两组不同的条形码化珠粒文库(5'组和3'组)对每个细胞的转录组进行条形码编码。参见图2。每个3'条形码由限制性位点、独特的分子标识符、条形码序列和将特异性靶向mRNA转录物的poly-A尾部的poly-T序列组成(因此避免每个真核细胞中包含相当大部分的总RNA的核糖体和tRNA)。每个5'条形码包含可以与3'珠粒组上的限制性位点相同的限制性位点(RE)、最终的条形码序列的互补序列和三个鸟嘌呤残基。这两个珠粒组与关注的细胞一起封装在液滴中。可以使用液滴装置,诸如图1所示的液滴装置。

[0121] 调节每个珠粒组(3'和5')的浓度,使得所需比率的液滴接收每组的至少一个珠粒。例如,对于每个液滴中每种类型的三个珠粒的平均值而言,超过90%的液滴将接收每种类型的至少一个珠粒。对于通过量和单个细胞封装之间的最佳平衡,类似地调谐细胞浓度。对于该实施例而言,平均值是0.1个细胞/液滴,在这种情况下,大于95%的接收一个细胞的液滴将接收恰好一个细胞。一个多产性液滴将接收一个细胞和至少一个来自5'和3'组中的每一个的珠粒。在成功封装事件之后,反应可以如图3所示进行。

[0122] mRNA被完全逆转录,并且在cDNA的每一端并入条形码序列。在每种类型恰好一个珠粒与细胞共封装的情况下,完全逆转录的mRNA转录物将表现出相同的5'和3'条形码对,包含对该细胞独特的条形码组。在一个或两个珠粒组中的多个与细胞共封装的情况下,预期每个5'/3'组合在所生成的cDNA文库中表示,并且这组5'/3'组合是对于该细胞独特的条形码组。每个cDNA将至少携带3'条形码。在逆转录不完全的情况下,5'条形码可能没有并入cDNA中。即使在仅并入3'条形码的情况下,只要该条形码与液滴中足够数量的5'条形码相关联以将其确立为该条形码组的一部分,不完整的cDNA仍可以潜在地被分配给下游分析中的其适当转录组。

[0123] 图4图示其中每个条形码组的多个条形码化珠粒与细胞一起共封装在单个液滴中的情况。由于每个细胞内大量的mRNA转录物,所有3'条形码将有机会通过上述条形编码方案与液滴中的所有5'条形码匹配。这些条形码形成一个独特的条形码组,所述条形码组用于鉴定源自细胞的所有转录物。

[0124] 在许多情况下,条形编码反应可以无法完成并且仅一部分转录物序列将被转录。虽然不完整,但无论如何这都是潜在有用的信息。图5图示,即使在未能并入5'条形码的非全长cDNA的情况下,这些cDNA仍可以被分配到正确的条形码组。因此,仅具有一个条形码的核酸可以以高置信度分配到正确的转录组(反之亦然)。对此的要求是3'条形码代码空间

(独特3'条形码的总数量)足够大,以致于任何一个3'条形码都不足以与多个细胞共封装。即使在出现这样的情况的情况下,也可以检测到这种重叠,并且可以安全地丢弃具有该特定共有的3'条形码的非全长cDNA。相比之下,在单个条形码系统的情况下,在分析期间将合并接收相同条形码的两个细胞的特征谱。

[0125] 图6表明,即使当相同的个别条形码有一些重叠时,所述条形码组也可以通过包含一个独特的条形码来区分。

[0126] 必须确定对于5'和3'组中的每一组的所需代码空间(独特条形码的数量)。条形码也必须具有足够的独特性,在存在测序错误的情况下,可以区分它们,即,在少量错误的碱基的情况下,一个条形码不应该被转换成另一个。确立条形码组将有助于解析测序错误可能会在条形码之间产生歧义的情况。

[0127] 一个替代实施方案类似于上述条形编码方案之处在于其涉及模板切换,但在这种情况下,5'条形码珠粒组具有不同的组成,如图7所示。这表明存在用于“模板切换”条形编码方案的多种方法,所述方案将条形码并入从mRNA转录物完全逆转录的cDNA的每个末端。

[0128] 实施例2.用于分析与单个细胞相关联的核酸的核酸条形码组-非模板切换。

[0129] 该实施例描述非模板切换方法。仅存在单组条形码化珠粒。每个珠粒具有形成发夹结构的单链DNA寡聚物。在发夹的双链区段中,并入限制性位点,其导向两条链的限制。这裂解转录物条形编码区段和还含有条形码以及“条形码关联序列”(BAS)的单独环状区段两者。参见图8。

[0130] 转录物条形码区段包括UMI、条形码和poly-T区段。该片段退火至由液滴内的细胞释放的mRNA转录物的poly-A尾部并引发该mRNA序列的逆转录,从而将条形码并入生成的cDNA中(图9)。

[0131] 单独地,环状区段包含其自己的条形码,所述条形码与裂解其的条形码相同或对应。环状区段还包含引发序列。BAS,这是一个回文序列。

[0132] 在多个珠粒与细胞共封装的情况下,两个转录物条形编码区段(BC1a和BC2a,图10)都将用作RT条形编码反应的引物,并且来自该细胞的转录组将并入两个条形码。单独地,还将存在携带其各自条形码的两个相应的环状区段(BC1b和BC2b,图10)。

[0133] 回文BAS将使BC1b和BC2b环状区段实现交叉退火,并且DNA聚合酶包括在反应混合物中以延伸每个区段以并入其他条形码的补体(图11)。通过这种方式,BAS在单个扩增子中物理连接,其然后可以在下游进行测序,以便将条形码关联为一个以统一来自该细胞的转录组。

[0134] 如果回文序列是不合需要的,则也可以设计一种系统,其中BAS包含两个互补序列,并且每个珠以大致相等的比例携带两个BAS序列。被描述为“环状”区段可以是线性的而不是环状的。这取决于使用的限制性酶。类似地,不是使用发夹,而线性寡核苷酸可以与多个限制性位点一起使用,这也导致从珠粒上切下两个或多个区段。这里再一次,“环状”区段可能不是环状的,而是线性的。

[0135] 实施例3.使用多次条形编码来鉴定哪些核酸与同一样品相关联,而无需事先知道条形码所处的容器

[0136] PBMC制备至条形码细胞。将冷冻保存的PBMC解冻并将细胞球粒在磷酸盐缓冲盐水(Corning,Manassas,VA)中洗涤三次并单独FACS分选到96孔板中或重新悬浮于包含25mM

NaCl和153mM甜菜碱的等渗细胞悬浮缓冲液(CSB)中。然后计数细胞并将其用CSB稀释至每孔平均1个细胞以便逆转录。

[0137] 寡聚物板制备。从Integrated DNA Technologies(Coralville,IA)订购条形码化的oligoDT寡聚物(3'条形码,在下文中称为“oDT BC”)和条形码化模板切换寡聚物(5'条形码,在下文中称为“wID BC”)。抵达后,将oDT BC和wID BC都稀释至10 μ M,并保持在-20℃下,直至使用。其具有以下序列:

[0138] 表2.寡聚物序列

名称	序列
[0139]	oDT BC GGC TCG GAG ATG TGT ATA AGA GAC AG [7-mer 条形码] NNN NNN NNT TTT TTT TTT TTT TTT V (SEQ ID NO:1)
	wID BC GGA AGA TAG GGA TAA CAG GGT AAT G [7-mer 条形码] ACG GG (SEQ ID NO:2)
	pltwll_LibPCR_Fw AAT GAT ACG GCG ACC ACC GAG ATC TAC ACT CGT CGG CAG CGG AAG ATA GGG ATA ACA GGG TAA TG (SEQ ID NO:3)
[0140]	pltwll_LibPCR_RV CAA GCA GAA GAC GGC ATA CGA GAT [指标序 列] GTC TCG TGG GCT CGG AGA TGT GTA TAA GAG A (SEQ ID NO:4)
	pltwell_PCR2_Fw AATGATACGGCGACCAC (SEQ ID NO:5)
	pltwell_PCR2_Rv CAAGCAGAAGACGGCATAC (SEQ ID NO:6)

[0141] 条形码设计成使得它们彼此间隔至少2个编辑距离,即,将其采用至少两个突变(置换、插入或缺失)以将一个条形码转换成另一个条形码。这大大降低了条形码分配错误,因为条形码需要至少2个测序错误才有由于这些错误而被错误分配为另一个条形码的机会。

[0142] 反转录和条形编码。制备96孔板,其含有经由随机泊松分布在 λ 为2下分配到每个孔中的oDT BC和wID BC。通过这样的分配,一些孔将不接收寡聚物,而其他孔将接收多个寡聚物,并且平均而言,每个孔将接收2个寡聚物。在R中写入随机泊松分布脚本以确定将哪些寡聚物放置于每个孔中,并且使用液体处理器(FeliX,Cybio,Germany)以将条形码分配到孔中。反应在每孔中包含6mM Mg²⁺、0.3%Tween20、RNase抑制剂、Maxima H Minus逆转录酶和牛血清白蛋白(BSA)的RT缓冲液中执行。在每个孔中,样品是通过FACS沉积的单个细胞。将反应混合物逆转录,并且使条形码化寡聚物在55℃下3分钟,在42℃下1小时并且在85℃下10分钟而并入cDNA中。然后将板在-20℃下冷冻过夜。

[0143] 样品汇集、文库制备和测序。将冷冻的板在冰上解冻以进行样品汇集。根据制造商的说明,将来自每个板的RT产物汇集成两半(96孔板的左半和右半)并使用Select-a-Size DNA Clean&Concentrator(Zymo Research,Irvine,CA)浓缩以除去短于300bp的物质。简言之,将500 μ L select-a-size DNA结合缓冲液添加到100 μ L的cDNA中,并通过轻轻移液来混合,然后以10,000g离心30秒。含有cDNA的柱用DNA洗涤缓冲液洗涤两次,并使用DNA洗脱缓冲液在15 μ L中洗脱。使用1 μ L的每种样品以使用Q5DNA聚合酶(New England Biosciences)

和0.2 μ M的引物pltw11_LibPCR_Fw和pltw11_LibPCR_RV进行PCR。使用以下PCR热循环条件：95℃变性5分钟；10个以下循环：98℃下30秒，60℃下30秒以及68℃下1分钟；并使用Eppendorf Mastercycler Nexus热循环仪 (Hauppauge, NY) 在68℃下最终延伸5分钟。如前所述，使用select-a-Size DNA Clean&Concentrator (Zymo Research, Irvine, CA) 再次将PCR产物浓缩以除去短于300bp的产物。然后使用0.2 μ M的引物pltw11_PCR2_Fw和pltw11_PCR2_Rv进行第二次PCR。使用Pippin Prep DNA大小选择系统 (Beverly, MA) 对大小为300bp至1200bp的产物进行大小选择并通过Illumina MiSeq系统 (San Diego, CA) 进行测序。

[0144] 分析和结果。在文库的MiSeq测序之后，使用Python脚本分析来自一组48个孔的所得读数对。由于根据泊松分布的条形码的随机分布，这些孔中有39个含有至少一个wID和至少一个oID。九(9)个孔含有零个wID条形码或零个oID条形码。使用读数的已知结构(5'-wID-ACGGG-mRNA-BdA16-N8-oID)，自反向读数鉴定oID BC，并且自正向读数鉴定wID BC。在下面的分析中没有使用已知的条形码序列，这表明该方法可以在不事先知道哪些条形码与哪个样品相关联的情况下确定核酸来源的样品，并且数据分析方法对由测序器引入的错误具有鲁棒性。

[0145] 由鉴定oID BC和wID BC两者的读数对产生一组“双条形码转录物读数”。该组由总共107,795个读数对组成，其包含6,964个不同的条形码对。图15中示出来自双重条形码化转录物的观察的条形码对的频率直方图。

[0146] 从寡聚物的合成预期低水平的交叉污染(参见IDT DNA Technologies, Reduced Barcode Contamination Using Oligonucleotides With TruGrade™ Processing, Decoded 2(4), 2012)。条形码寡聚物一起成组地合成。在一起合成N个条形码的情况下，如果每个递送的寡聚物含有每种其他N-1寡聚物的相等混合物，使得每种所需产物寡聚物仅包含1-p分数的生成的寡聚物，则可以估算将具有预期的条形码对相对由污染产生的对的读数的分数。具有预期对的读数的分数可以估算为 $(1-p) * (1-p)$ 。污染物的分数可以估算为 $(1 - (1-p) * (1-p))$ 。如果由于污染而存在(N-1)个不同的寡聚物条形码，则总共有N*N-1个污染物对。然后将污染物对的比例估算为 $f = (1-p*p) / \{p*p * (N*N-1)\}$ ，并且对于给定数量R的读数，每个污染物的读数对的数量是 $f*R$ 。

[0147] 一次订购24个板中的288个总条形码。在摩尔水平下保守地允许多达10%的寡聚物交叉污染，据估计具有少于40个不同读数对的条形码对可能是污染物。因此设定了阈值，要求在我们的分析中进一步考虑至少50个读数取对中观察到给定的条形码对。

[0148] 数据分配到孔和细胞。输出读数被分配给单个细胞和孔，而不用事先知道哪些条形码分布到每个孔。该分析的起点是在数据中由至少50个读数对支持的183对wID BC和oID BC的组。这些183对BC由共计94,712个读数对支持。这些对中的每一对在图中都被视为边(wID BC、oID BC)，其顶点是观察到的wID BC和oID BC，分配给边至少50个读数。表示我们的数据集的图是具有包含一组节点的wID BC和包含另一组节点的oID BC组成的二分图。注意到具有一个或多个wID BC和一个或多个oID BC的单元/孔应该产生一组边，这些边形成由数据中的所有边组成的图的完整二分图子图(也参见图13和14)。因此，通过开发一种将所有数据组的图分解为完整的二分图的算法，将图的边分配给细胞/孔。

[0149] 该算法的步骤如下：

[0150] 1. 计算图中所有连接的组分

- [0151] 2.完整的二分图(二聚类)的连接组分各自被分配到一个孔/细胞
- [0152] 3.其他连接的组分被分解成如下的一组二聚类:
- [0153] a.确定作为该组分的子图的所有二聚类(算法在下面解释)
- [0154] b.在3.a.中的图表组中,选择了最大的二聚类(未包括在任何其他二聚类中)
- [0155] c.应用“修剪”步骤以除去不可能存在于数据中的二聚类
- [0156] 上述算法的步骤3a.如下执行:
- [0157] 1.如下产生组O和I:
- [0158] a.对于连接的组分中的每个wID BC,确定通过边连接到它的一组oDT BC,并将其称为输出(wID BC),然后将其添加到O两个输出(wID BC)及其所有子组中。
- [0159] b.以类似于a.的方式,如下产生组I:对于连接的组分中的每个oDT BC,确定通过边连接到它的一组wID BC并将其称为进入(oDT BC),然后将其添加到I两个进入(oDT BC)及其所有子组。
- [0160] 2.现在,对于I中的每个组i和O中的每个组o,确定i中的wID BC和o中的oDT BC之间的所有边是否存在于数据中,并且如果是这样的情况,则将二聚类添加到作为该组分的子图的所有二聚类的组中
- [0161] 上述算法的步骤3c.如下执行:
- [0162] 1.在 $\lambda=2$ 时,仅小分数0.0166的细胞将具有给定类型的多于5个条形码。因此,具有超过5个wID BC或5个oDT BC的通过该算法构建的任何二聚类都被拒绝作为有效构建体并被滤出
- [0163] 2.具有单个wID BC和4个或更多个oDT BC(或相反)的聚类具有是由较小二聚类组成的模糊构建体的高度可能性。鉴于这种模糊性,这些二聚类也被过滤。
- [0164] 步骤3b.和3c.的启发法可以用一种最大似然方法来取代,这种方法可以将每个组分分解成它所包含的最可能的二分聚类,并且这有望进一步改善结果。可能性取决于参数 λ ,正如我们的启发式3c.所讨论的,并且也取决于构成组分的多个二聚类之间wID BC和oDT BC中的重叠量。
- [0165] 表3.分析结果。

[0166]

组分	组分大小	wBC 数量	odtB C 数量	修剪最大的二聚类的数量	完美重构的二聚类的数量	部分重构的二聚类的数量	相交 1 个孔的二聚类的数量	相交 2 个孔的二聚类的数量
1	48	25	23	12	6	3	0	3
2	20	12	8	3	3	0	0	0
3	15	8	7	4	2	1	0	1
4	11	4	7	2	2	0	0	0
5	10	2	8	2	1	0	0	1
6	8	2	6	1	1	0	0	0
7	6	2	4	1	1	0	0	0
8	5	2	3	1	1	0	0	0
9	4	3	1	1	0	0	1	0
10	4	2	2	1	1	0	0	0
11	4	2	2	2	1	0	1	0
12	4	3	1	1	1	0	0	0
13	2	1	1	1	1	0	0	0

[0167]

14	2	1	1	1	1	0	0	0
15	2	1	1	1	1	0	0	0

[0168] 表2示出来自贡献该数据组的39个孔的数据产生15种组分(注意48个孔中仅有39个孔具有wID BC和odT BC两者。9个孔仅具有wID条形码或仅具有odT条形码,并且因此不能产生条形码对)。给出了由分解算法产生的wID BC和odT BC的数量以及修剪的最大二聚类的数量。此外,表2给出了与预期聚类精确匹配的二聚类的数量,这些二聚类是预期二聚类的一个恰当子组,或者与一个或两个预期图相交。

[0169] 从15种连接的组分中产生34个完整的二分图。在这34个之中,23个与预期图完美匹配。另外的4个图是预期图的适当子组。因此,完美地产生 $23/34=67\%$ 的所得聚类,并且至少部分地重构 $27/34=79\%$,并没有错误的边缘分配给它们。总体上 $27/48=56\%$ 的实验中包括的孔成功地分配了相关联的条形码和mRNA。这些成功的重构是可能的,尽管由于在条形码寡聚物制造期间的交叉污染而存在虚假读数对的数据集,其可以在数据处理中被过滤出来。

[0170] 这种多条形编码方案大大优于单条形编码方案,其中目标是将单个条形码分配给每个细胞。在具有泊松装载的单条形编码方案中,使用 $\lambda=0.1$ 或更小将是自然选择(参见Macosko等人,Cell,161:5(2015),其使用0.06的 λ)。在 $\lambda=0.1$ 时,大约90.4%的细胞会不接收条形码,大约9.0%的细胞会接收恰好一个条形码,并且大约0.5%的细胞会接收多于一个条形码。在这里描述的实验中, $56\%/(100-90.4)\%$ 等于使用单条形编码方案成功分析的细胞数的5.8倍。

序列表

	<110>	Atreca, Inc.	
	<120>	用于分析与单个细胞相关联的核酸的核酸条形码的组合组	
	<130>	097519-1028178-001010PC	
	<150>	62,250,900	
	<151>	2015-11-04	
	<160>	6	
	<170>	PatentIn 3.5 版	
	<210>	1	
	<211>	51	
	<212>	DNA	
	<213>	人工序列	
	<220>		
[0001]	<223>	合成构建体	
	<220>		
	<221>	misc_feature	
	<222>	(27)..(34)	
	<223>	n 为 a、c、g 或 t。	
	<400>	1	
		ggctcggaga tgtgtataag agacagnnnn nnnntttttt ttttttttt v	51
	<210>	2	
	<211>	30	
	<212>	DNA	
	<213>	人工序列	
	<220>		
	<223>	合成构建体	
	<400>	2	
		ggaagatagg gataacaggg taatgacggg	30

	<210> 3	
	<211> 65	
	<212> DNA	
	<213> 人工序列	
	<220>	
	<223> 合成构建体	
	<400> 3	
	aatgatacgg cgaccaccga gatctacact cgtcggcagc ggaagatagg gataacaggg	60
	taatg	65
	<210> 4	
	<211> 55	
	<212> DNA	
	<213> 人工序列	
	<220>	
[0002]	<223> 合成构建体	
	<400> 4	
	caagcagaag aeggcatacg agatgtctcg tgggctcgga gatgtgtata agaga	55
	<210> 5	
	<211> 17	
	<212> DNA	
	<213> 人工序列	
	<220>	
	<223> 合成构建体	
	<400> 5	
	aatgatacgg cgaccac	17
	<210> 6	
	<211> 19	
	<212> DNA	
	<213> 人工序列	

	<220>	
	<223>	合成构建体
[0003]	<400>	6
	caagcagaag acggcatac	19

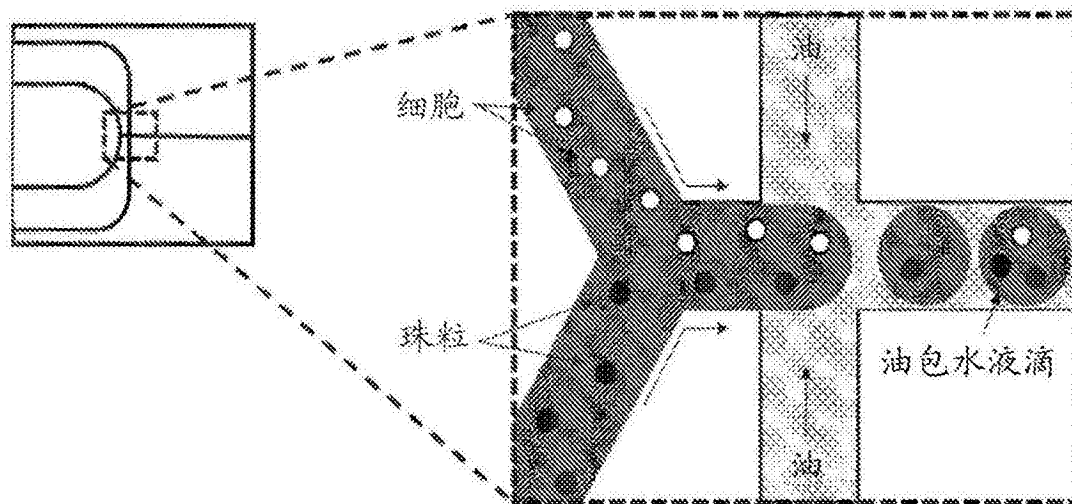


图1

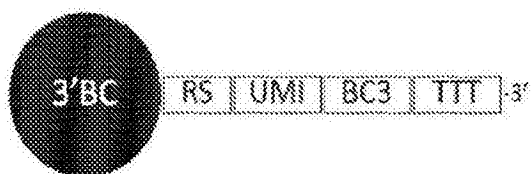


图2A

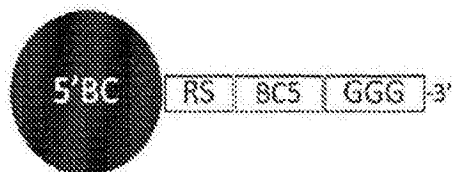


图2B

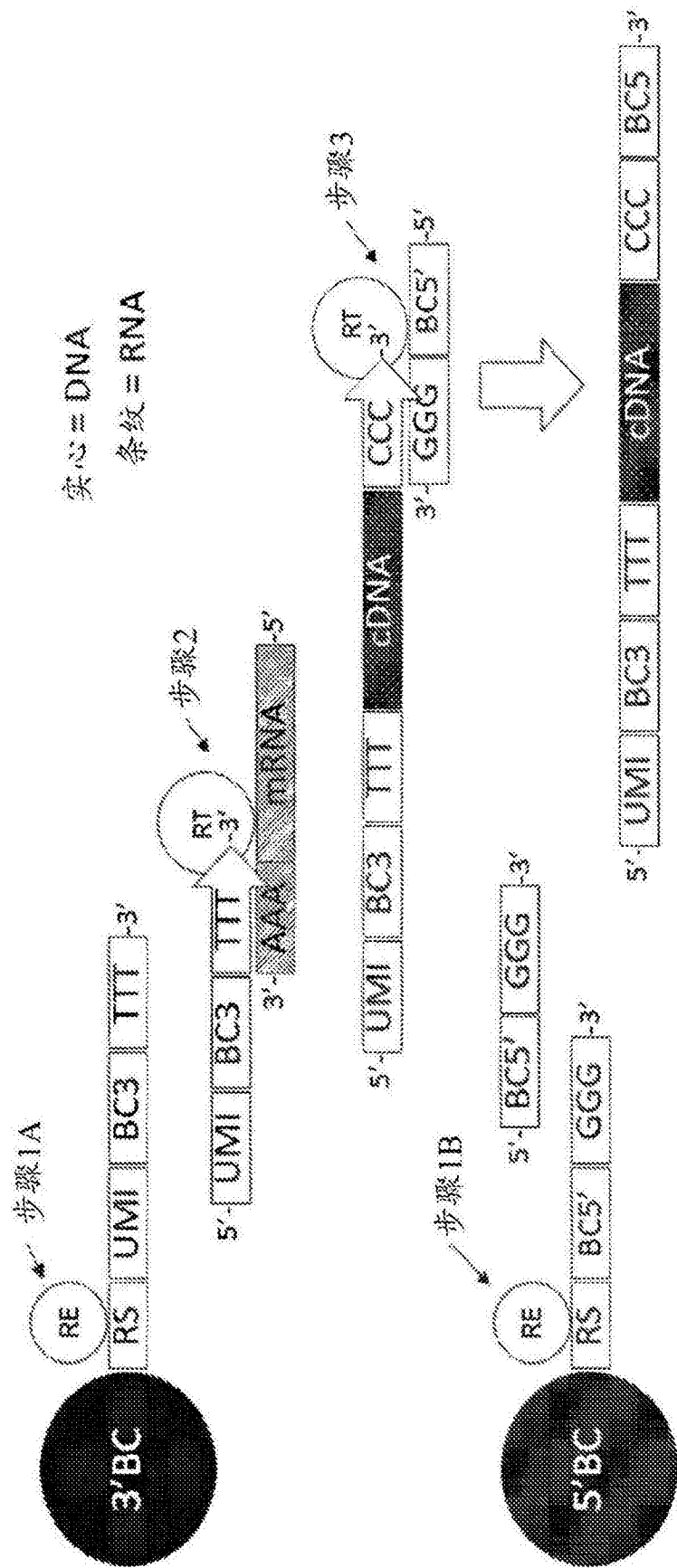


图3

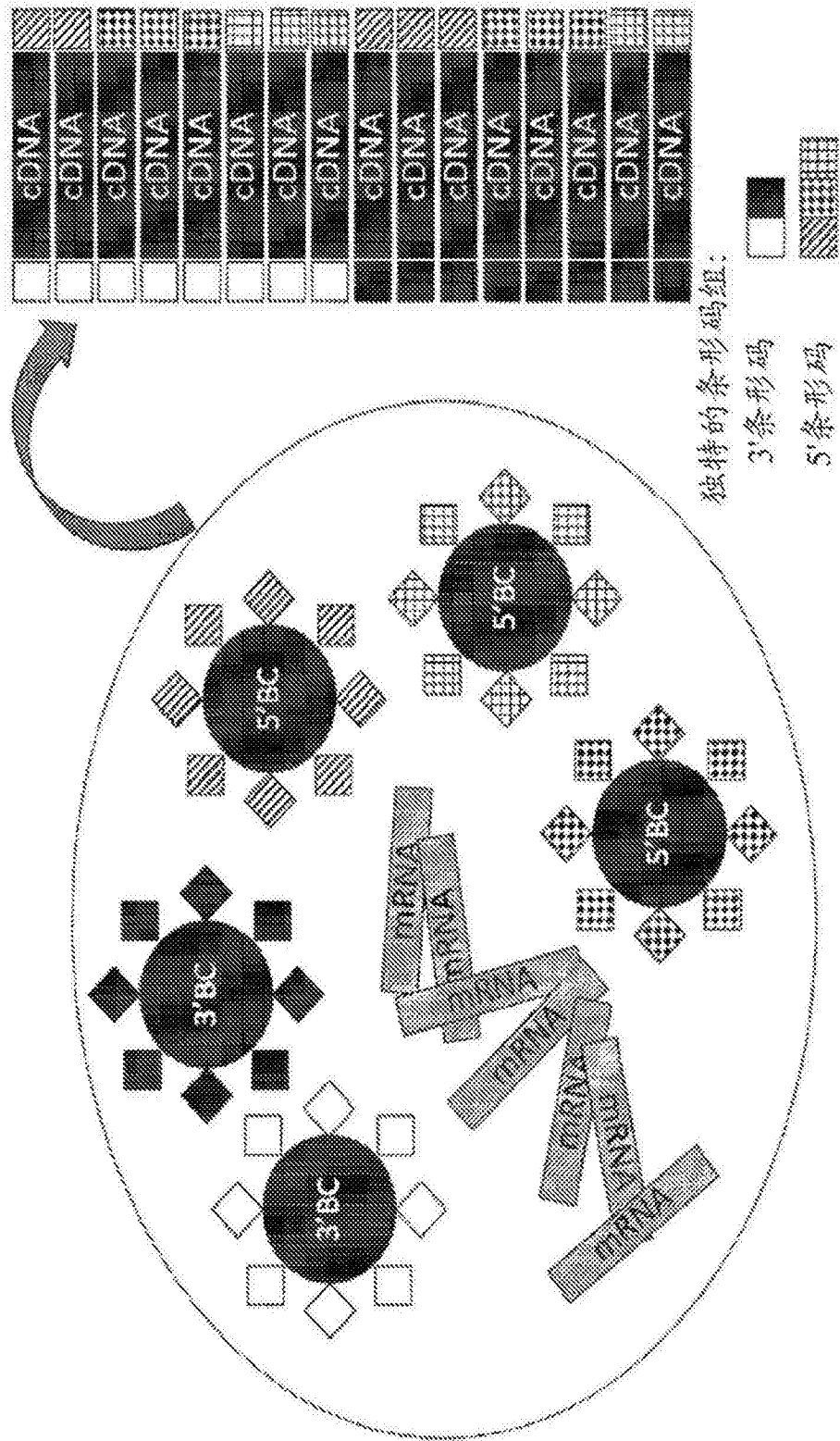


图4

如果3'BC编码空间足够多样，则非全长cDNA(缺乏5'BC)可以高置信度分配给正确的转录组：

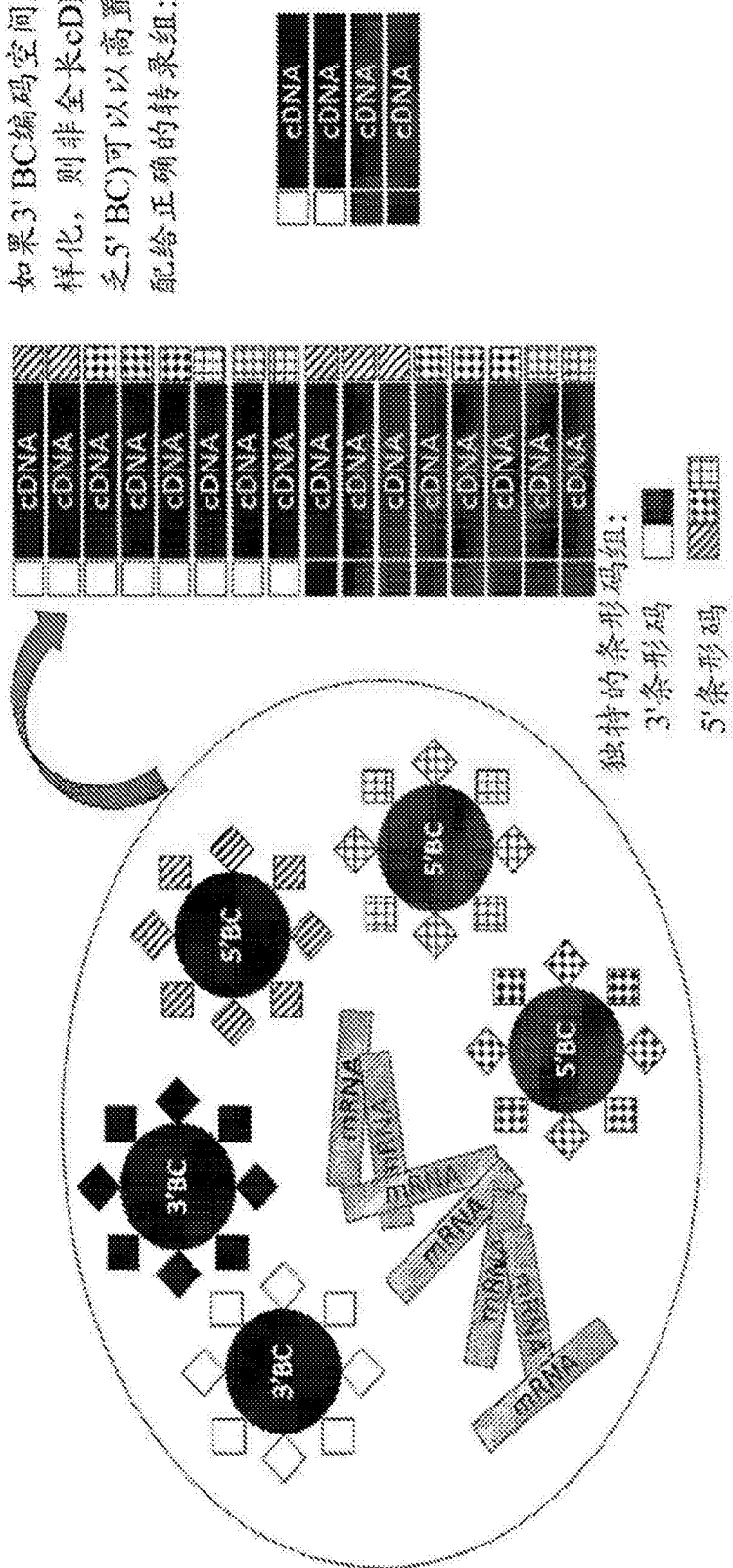


图5

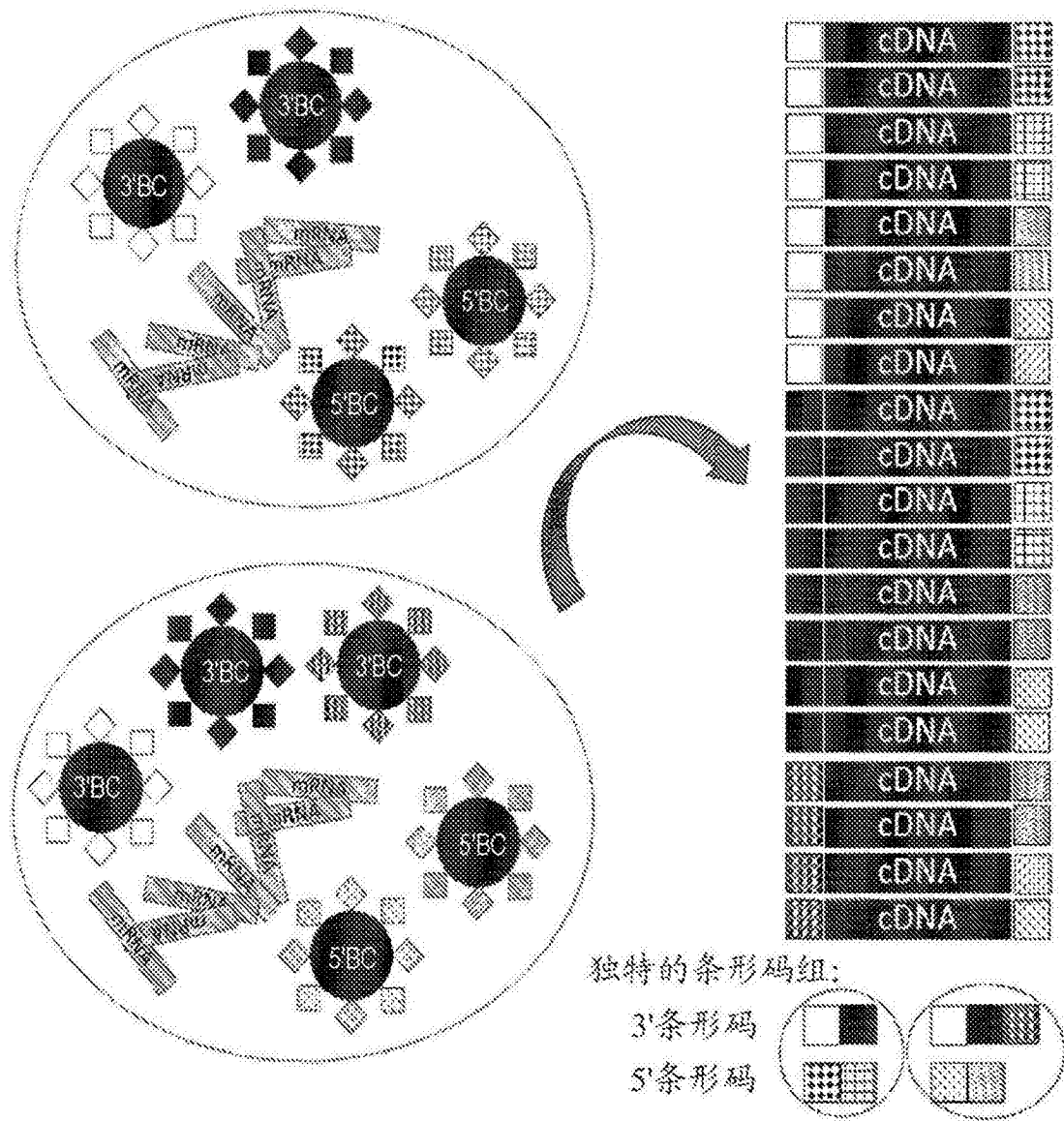


图6

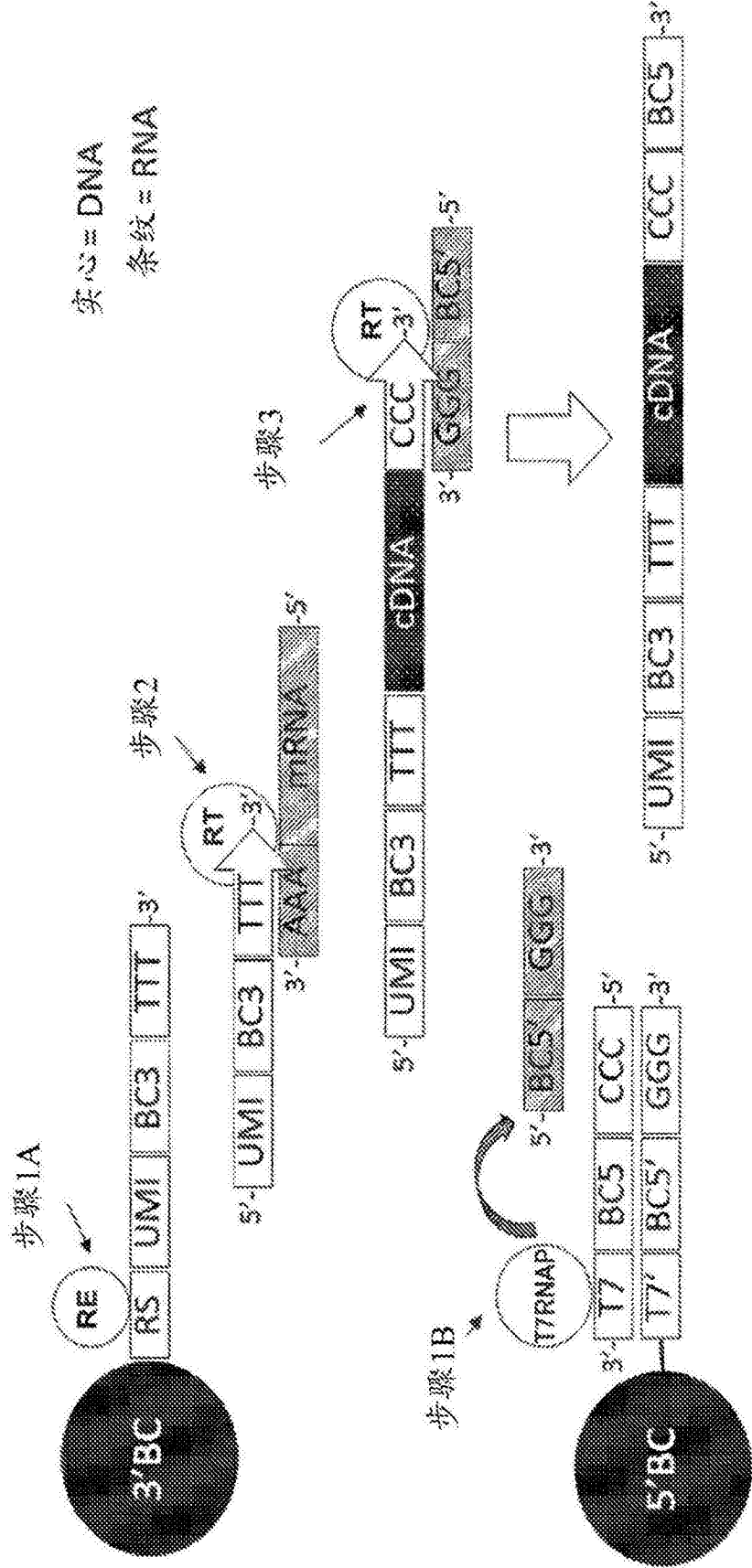


图7

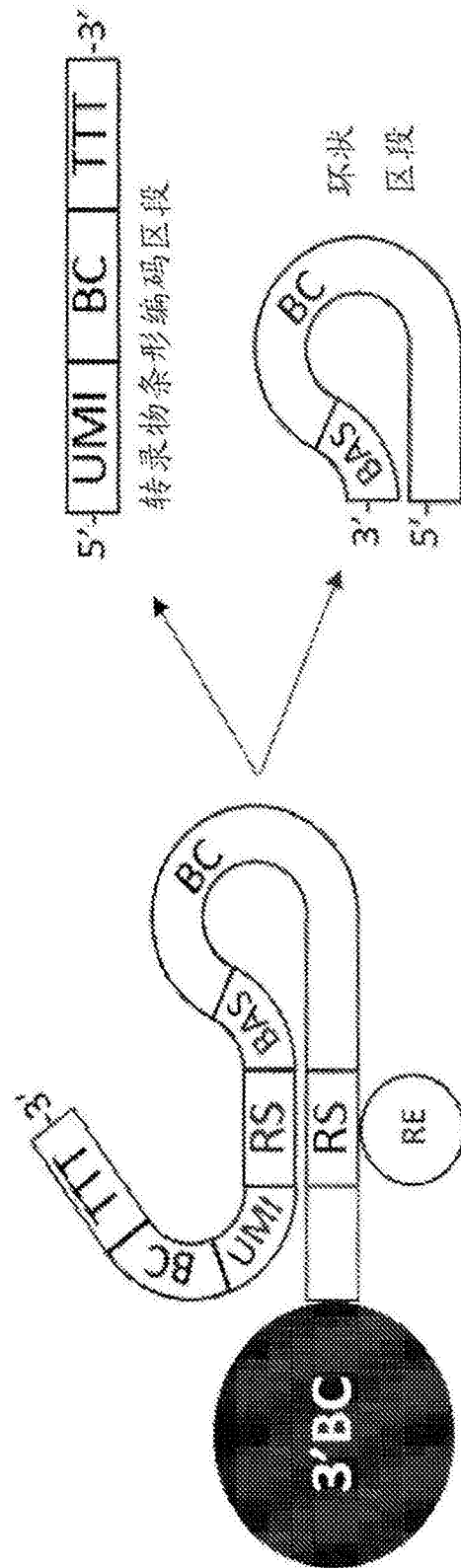


图8

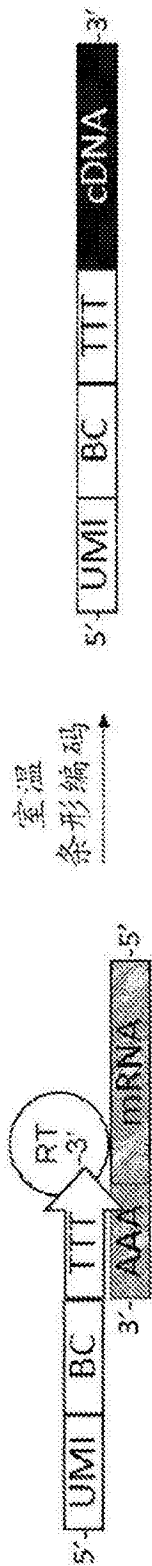


图9

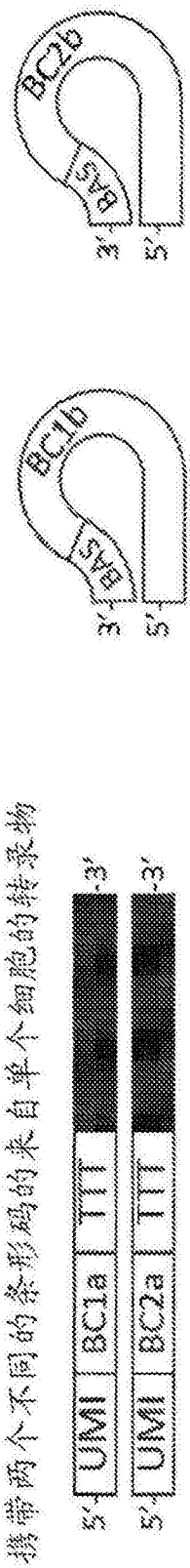


图10

BAC的退火，每个BC延伸
以并入其他的补体

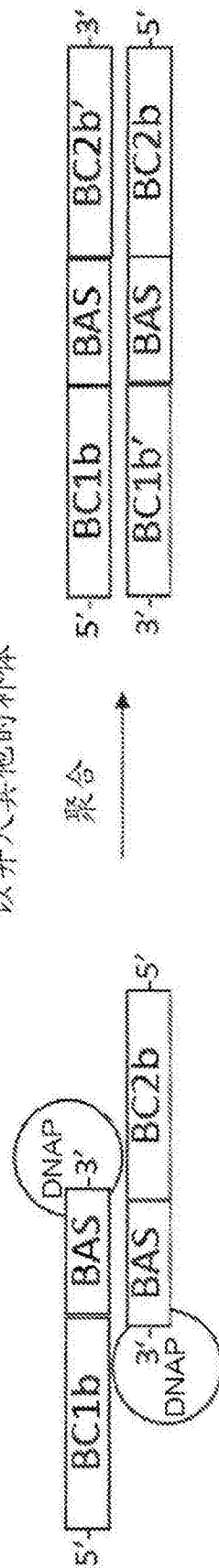


图11

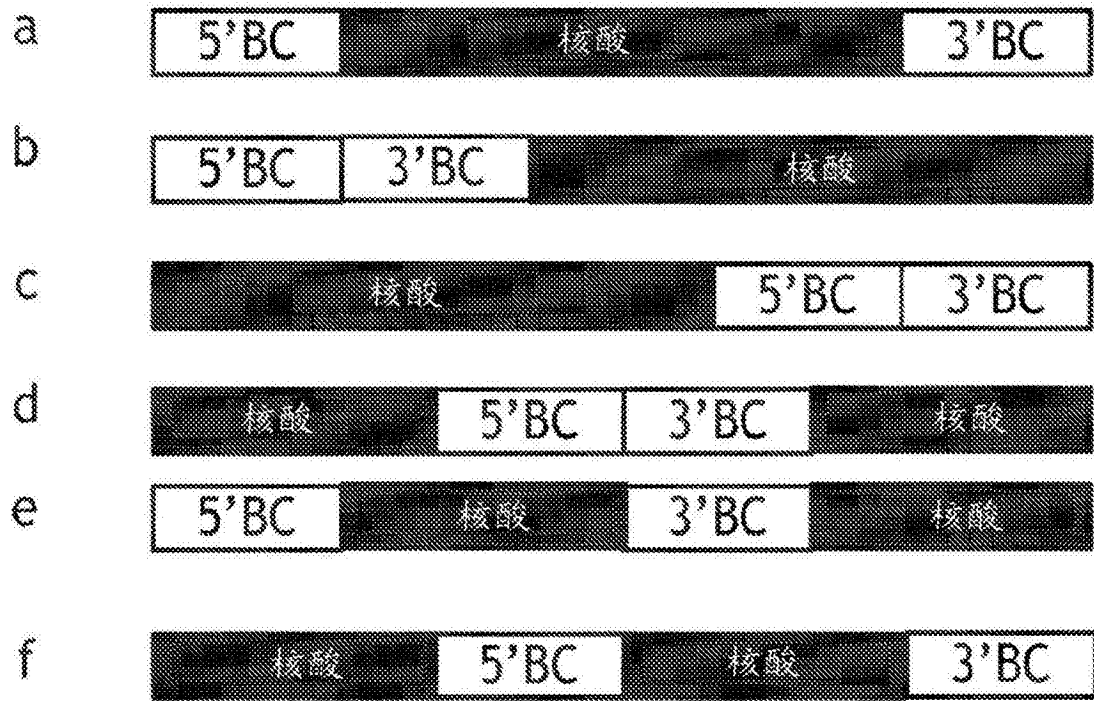
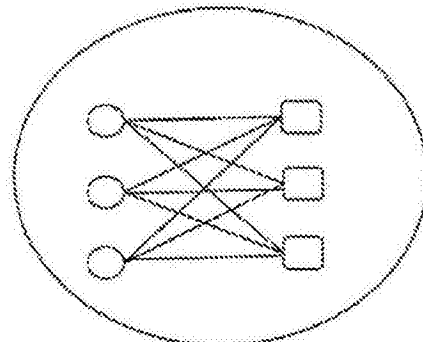


图12

与样品1相关联的条形码



与样品2相关联的条形码

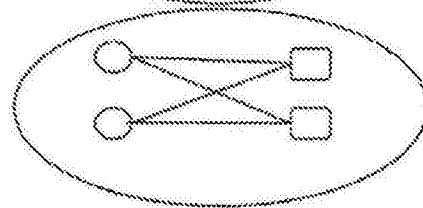


图13

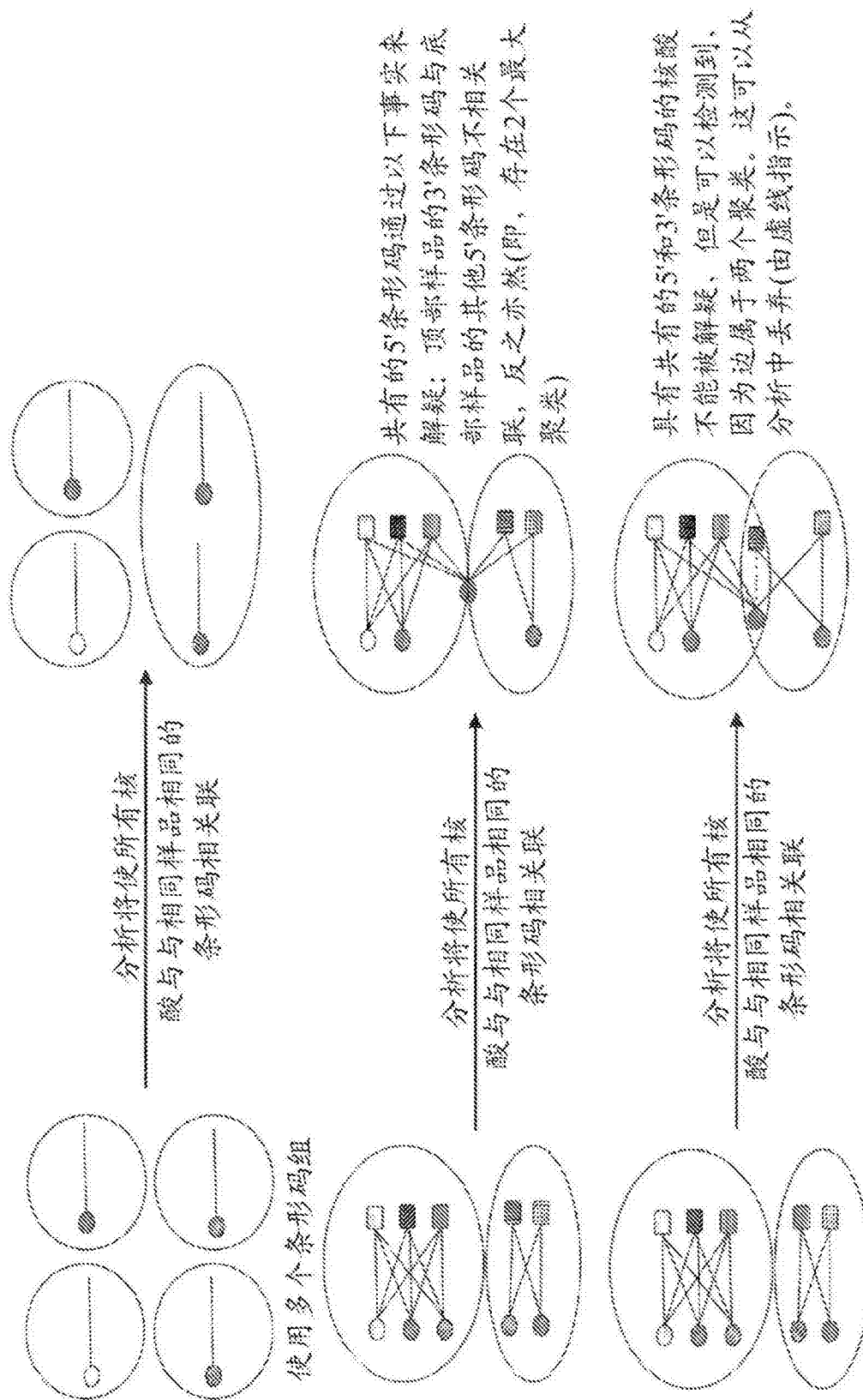


图14

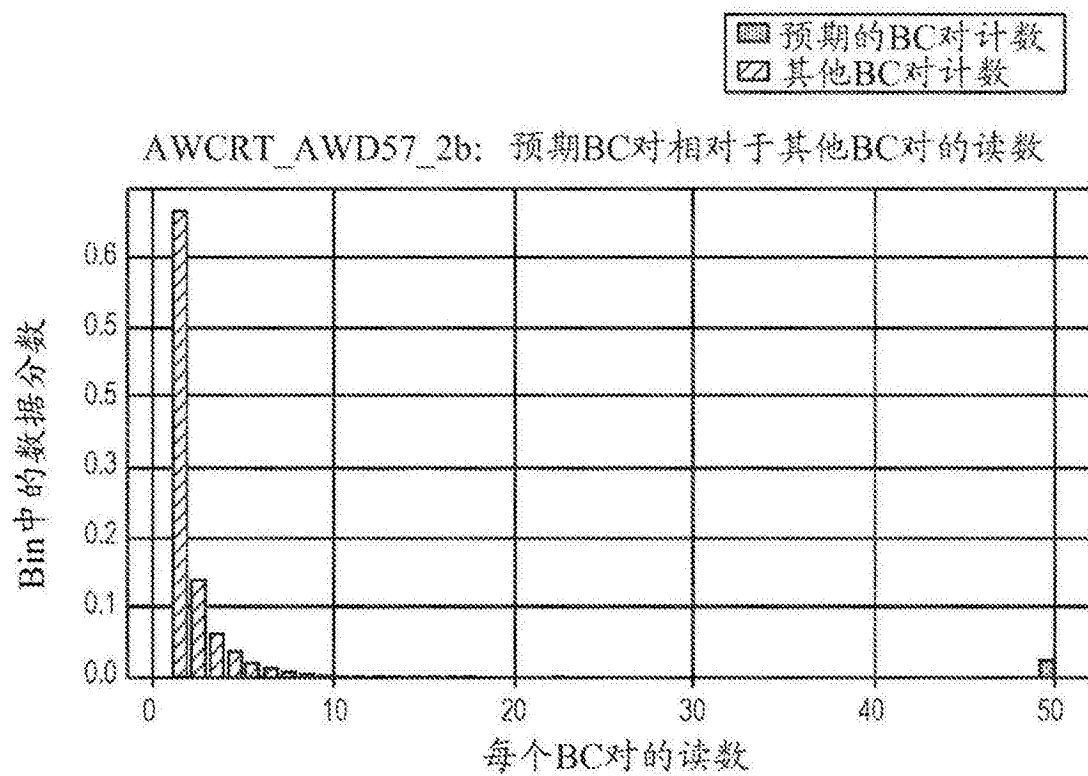


图15