



(45)授权公告日 2018.11.02

W02014/130388 EN 2014.08.28

1. 一种方法,其包括

a) 混合样品和一组标记多核苷酸,其中所述样品包括不同长度和/或不同序列的核酸混合物,其中所述标记多核苷酸单独地包括不变碱基和随机序列部分,其中所述标记多核苷酸包含被配置为自身杂交成为双链片段的回文序列,并且其中所述双链片段包含限制性位点,并且其中所述混合是在使得所述标记多核苷酸与所述核酸结合以形成单独地被所述随机序列标记的核酸的条件下进行;

b) 用滚环扩增(RCA)将单独地被随机序列标记的核酸扩增成同多联体的混合物,其中所述同多联体包括重复的来自样品的核酸和重复的标记多核苷酸;

c) 断裂所述同多联体,提供了同多联体片段;以及

d) 对同多联体片段进行测序。

2. 权利要求1的方法,其还包括在步骤c)之后且在步骤d)之前将所述同多联体片段与切割所述标记多核苷酸上不变碱基内的位点的限制性核酸酶混合,提供了经切割的同多联体片段。

3. 权利要求1的方法,其还包括以下步骤:鉴定所述同多联体片段内的单独地被所述随机序列标记的核酸,通过分离包含相同随机序列的单独标记的核酸进行鉴定,以及重新构建样品中的核酸序列。

4. 权利要求1的方法,其中所述限制性位点是不常见的限制性位点。

5. 一种方法,其包括:

a) 提供包括标记部分和靶标部分的双链核酸片段,其中所述标记部分包含被配置为自身杂交成为双链片段的回文序列,其中所述双链片段包含限制性位点,其中所述标记部分包含不变碱基片段和随机序列,其中所述不变碱基包含第一引物位点和限制性位点;

b) 将所述双链片段与针对所述限制性位点的限制性酶混合,提供了经切割的片段;

c) 将所述经切割的片段在使所述经切割的片段形成环状片段的条件下与酶混合;

d) 使所述环状片段在随机点断裂,提供了经剪切的片段;

e) 将衔接体连接到所述双链核酸的末端,其中所述衔接体包含第二引物位点,从而产生衔接体-核酸缀合物;

f) 使用针对第一和第二引物位点的引物来扩增所述衔接体核酸缀合物,其中所述第一引物在5'端包含第一捕获序列,所述第二引物在5'端包含第二捕获序列,以提供捕获靶标标记的缀合物;以及

g) 对所述捕获靶标标记的缀合物进行测序。

6. 权利要求5的方法,其中所述随机序列位于所述第一引物位点和所述靶标部分之间。

7. 权利要求5的方法,其中所述第一引物位点位于所述随机序列和所述靶标部分之间。

8. 权利要求5的方法,其中所述限制性位点位于所述随机序列和所述第一引物位点之间。

9. 权利要求5的方法,其中所述随机序列位于所述限制性位点和所述第一引物位点之间。

10. 权利要求5的方法,其中所述核酸片段包含两个随机序列,其中所述两个随机序列是相同的序列并且所述限制性位点在所述两个随机序列之间。

11. 一种方法,其包括:

a) 将样品和一组标记多核苷酸混合,其中所述样品包括不同长度和/或不同序列的核酸的混合物,其中所述标记多核苷酸单独地包括不变碱基和随机序列,其中所述标记多核苷酸包含被配置为自身杂交成为双链片段的回文序列,其中所述双链片段包含限制性位点,

其中所述随机序列在所述双链片段内,并且

其中所述混合是在使所述标记多核苷酸与所述核酸结合以形成单独地被随机序列标记的核酸的条件下进行;

b) 用滚环扩增(RCA)将单独地被随机序列标记的核酸混合物扩增成同多联体的混合物,其中所述同多联体包括重复的来自样品的核酸和重复的标记多核苷酸;

c) 使所述同多联体断裂,从而产生同多联体片段;

d) 将所述同多联体片段与限制性核酸酶混合,所述酶切割与所述标记多核苷酸上不变碱基相关的位点,提供了经切割的同多联体片段;以及

e) 对所述经切割的同多联体片段进行测序。

12. 一种组合物,其包括多核苷酸混合物,所述多核苷酸各自单独地包含不变碱基、随机序列、多聚-T部分、PCR引物序列和限制性位点,其中所述多核苷酸各自包含被设置为自身杂交成为双链片段的回文序列,其中所述双链片段包含限制性位点。

13. 权利要求12的组合物,其中所述多聚-T位于3'端附近,并且所述随机序列位于所述多聚-T和限制性位点之间。

14. 权利要求12的组合物,其中所述随机序列在所述双链片段内。

15. 权利要求13的组合物,其中所述多聚-T位于3'端附近,并且第二多聚-T位于5'端附近。

16. 权利要求12-13及15中任一项的组合物,其中所述限制性位点是不常见的限制性位点。

17. 权利要求12-13及15中任一项的组合物,其中所述随机序列包含散布着不变碱基的随机碱基位点或序列。

18. 一种包含多核苷酸的混合物的组合物,所述多核苷酸各自单独地包含不变碱基、随机序列、重复相同的随机序列、多聚-T部分,以及位于随机序列和重复相同的随机序列之间的限制性位点,其中所述多核苷酸各自包含被设置为自身杂交成为双链片段的回文序列,其中所述双链片段包含限制性位点。

19. 一种产生核酸的方法,其包括:

a) 将引物和复制试剂与起始发夹多核苷酸混合,所述多核苷酸包含3'多聚-T、不变碱基、随机序列和环序列,其中所述发夹多核苷酸包含被配置为自身杂交成为双链片段的回文序列,其中所述引物退火到所述环序列上,以形成部分双链和部分单链的核酸;以及

b) 将所述部分双链和部分单链的核酸与多聚-A引物和复制试剂混合,以形成完全双链的核酸。

20. 权利要求19的方法,其还包括切割所述多聚-A引物以提供具有多聚-T尾的双链核酸的步骤。

21. 权利要求20的方法,其还包括使所述双链核酸变性以形成具有25多聚-T尾的发夹核酸和所述起始发夹多核苷酸的步骤。

22. 权利要求19-21中任一项的方法,其中所述起始发夹多核苷酸被缀合至固体载体上。
23. 一种包括权利要求12-17中任一项的组合物的试剂盒。

混合物及其相关组合物中的核酸的测序方法

[0001] 相关申请的交叉引用

[0002] 本申请要求于2013年2月20日提交的美国临时申请号61/766,841的优先权,所述临时申请以引用的形式全文纳入本文。

背景技术

[0003] 单个基因可经常在不同的细胞或分化阶段产生新蛋白,包括在生物的生命周期中通常不会遇到的细胞(例如,癌细胞、培养物中的细胞、发育的神经解剖异常结构中的细胞)。不同的蛋白来自指定所述表达细胞中的蛋白的信使RNA(mRNA)的转录激活和转录后RNA加工的差异模式。

[0004] 细胞中存在的mRNA“转录物”的群体在本文被称为“转录组”。转录组测序的当前最先进的技术是“RNA-Seq”。参见Nature Methods(2008) 5,621-628。在该方法中,将分离自组织或细胞培养物的mRNA逆转录成互补的DNA(cDNA),对该cDNA进行加工并扩增以产生短片段的文库,对所述短片段进行测序。通过使cDNA片段的序列重叠并将其与基因组中的序列比对无法分析(profile)细胞中的mRNA。实际上,使用复杂的统计学算法来组装最可能的mRNA的群体,所述算法的有效性是目前研究的活跃主题。RNA-Seq确实提供关于组织特异性“外显子组”的信息,包括信使RNA中保留的基因组序列,包括指定蛋白编码结构域的片段。

[0005] RNA-Seq方法不保留关于序列变体的某些信息,主要是因为单个mRNA转录物通常包含若干可变区,其通常由远超过测序仪cDNA读段长度(read length)的距离隔开。因此,可变区的哪些组合存在于同一mRNA转录物上是不清楚的。

[0006] 以编码具有分隔开1500个核苷酸的如下两个“任选的”结构域的蛋白的基因为例:氨基端附近的钙结合结构域(C)和羧基端的钙调蛋白结合结构域(M)。该基因的转录物可被选择性地剪接,以在最终的mRNA中保留两个结构域(CM),仅保留一个结构域(cM或Cm)或都不保留(cm)。所表达的蛋白可具有四种非常不同的生理学行为,这取决于存在哪个结构域。如果RNA-Seq实验揭示出两个结构域的两种变化,则一种完全无法推论出哪些转录物真正存在于最初的mRNA池中,数据支持任一下转录物组:{CM,cm}、{cM,Cm}、{CM,cm,cM,Cm}等。这是因为连接结构域C和M的长区域在所有转录变体中包含相同的序列。

[0007] 如前面的说明书所证明的,大规模cDNA测序的挑战本质上与高等物种的基因生物学有关。关于哪些信息将在给定的细胞中或细胞分化阶段中表达的不确定性与哪些来自高度平行的cDNA测序的短读段(read)可被指定给具体的转录物的不确定性相匹配。因此,需要在基因组和蛋白质组之间的生物化学通道(conduit)中捕获更多的信息。

[0008] Fu等人报道了分子索引,其使得能够进行定量的靶向RNA测序并揭示出标准文库制备中的低效率。Proc Natl Acad Sci USA.2014,111(5):1891-6。

[0009] 某些方法已被描述为可能提供大规模转录组测序。这些方法在其应用上受到限制。Zamore等人在题目为“通过以RNA为模板的DNA连接/测序来推断外显子连通性(Deducing Exon Connectivity by RNA-Templated DNA Ligation/Sequencing)”的PCT公开文本W02011/049955中提供了某些测序方法,包括其中将RNA退火至与每一种均含有随机

条形码的已知选择性剪接点互补的低聚物的方法。然后进行连接和随后的测序。这种方法受到限制,因为它需要事先了解外显子连接点,并且无法对每一种mRNA的全长进行测序。

[0010] 平行标记测序(PTS)也是分子条形码方法。参见Meyer et al., Nature Protocols, 2007, 2, 267-278。这种方法依赖于通过连接和链置换将包含序列标签和限制性位点的样品特异性条形码衔接体连接到平末端修复的DNA样品上。通过使用所述标签序列来追踪每一种DNA序列的样品源。

[0011] Parameswaran et al., Nucleic Acids Res., 2007, 35(19):e130公开了这样一种方法,其通过组合提高条形码多样性以使得能够对来自样品源的文库进行混合测序。仅使用样品特异性标签。单个的转录物是不可区分的,或无法完全测序的。

[0012] Craig et al., Nat Methods., 2008, 5(10):887-893记载了使用简并索引DNA序列条形码在Illumina基因组分析仪上对人类基因组的靶区域进行多重测序的方法,所述DNA序列条形码在测序前连接至片段化的DNA。

[0013] Halbritter等人报道了使用基于多重条形码阵列的PCR扩增和下一代测序在患有肾消耗病相关的纤毛疾病患者中进行高通量突变分析。参见J Med Genet. 2012, 49:756-767。

[0014] Sharon等人报道了人类转录组的单分子长读段测量。Nat Biotechnol, 2013, 31:1009-14。

[0015] 本文所引用的参考文献并非是对现有技术的承认。

发明内容

[0016] 本公开内容涉及获得存在于异质混合物中的单一多核苷酸的全长(端到端)序列。还涉及使得能够进行此类分析的专用试剂的设计、合成和制备方法。在某些实施方案中,本公开内容涉及对高等多细胞生物的细胞或组织的转录组中的mRNA进行完全测序和定量。所公开的方法使得能够对全长mRNA进行有效、经济的测序,所述全长mRNA确定高等多细胞生物的细胞或组织的分子表型。在某些实施方案中,本公开内容涉及包括试剂的商业试剂盒,以及进行此类分析的应用方法。

[0017] 在某些实施方案中,本公开内容涉及包括以下步骤的方法:a)混合样品和一组标记多核苷酸,其中所述样品包括不同长度和/或不同序列的核酸的混合物,其中所述标记多核苷酸单独地包括重叠序列和随机序列部分,并且其中所述混合在使得所述标记多核苷酸与所述核酸结合以形成单独地被随机序列标记的核酸的条件下进行;b)将单独地被随机序列标记的核酸混合物复制成同聚物的混合物,其中所述同聚物包括重复的核酸和重复的序列标签;c)断裂同聚物,例如,通过酶促片段化、加热、剪切、超声或暴露于一种或多种限制性酶,提供了同聚物片段;以及d)对同聚物片段进行测序。所述同聚物片段的长度通常少于1000、2000或5000个核苷酸碱基。在某些实施方案中,断裂同聚物是随机进行的。在某些实施方案中,使用切割标记多核苷酸的重叠序列内的位点的限制性核酸酶或等同试剂来进行断裂同聚物,提供了经切割的同聚物片段。

[0018] 在某些实施方案中,所述方法还包括以下步骤:将同聚物片段与切割标记多核苷酸上的重叠序列内的位点的限制性核酸酶混合,提供了在一端具有标记序列并且在另一端具有靶核酸的随机内部断裂点的经切割的同聚物片段。

[0019] 在某些实施方案中,所述方法还包括对经切割的同聚物片段进行测序的步骤。

[0020] 在某些实施方案中,所述方法还包括以下步骤:鉴定同聚物片段内的标记序列,分离随机序列部分内的相同序列,以及根据靶核酸的相关随机内部序列重新构建样品中的核酸序列。

[0021] 在某些实施方案中,所述标记多核苷酸包含被配置为自身杂交成为双链片段的回文序列,其中所述双链片段包含限制性位点。通常,所述限制性位点为不常见的限制性位点。

[0022] 在某些实施方案中,本公开内容涉及试剂盒,其包括本文公开的标记多核苷酸和任选的逆转录酶,其他病毒逆转录酶,或从单链RNA生成双链核酸的其他来源的任何同等的酶,核苷酸,以及本文公开的其他试剂。在某些实施方案中,所述试剂盒包括详细说明本文公开的方法的说明书。

[0023] 在某些实施方案中,独特的标签与核酸样品中的每一种核酸相关。在一些实施方案中,所述独特的标签包含源分子标识符(source molecule identifier, SMID)。在某些实施方案中,单个核酸被复制成长的串联同聚物,其在每个连接点/亚单位重复处包含独特的鉴定序列。在步骤c)的某些实施方案中,同聚物被随机片段化,并且在一些实施方案中,同聚物被选择性地切割以提供包含SMID的片段。在某些实施方案中,通常以平行冗余的方式对SMID和随机断裂序列一起进行测序,并根据分子来源将序列分离。在一些实施方案中,通过使读段重叠来分析这些序列以提供全长mRNA源分子的共有序列。在某些实施方案中,通过与从已建立的基因数据库中选出的基因序列进行比对来分析所述序列,并与已建立的基因数据库中报道的外显子边界进行比较。在某些实施方案中,通过与从转录物数据库检索到的序列进行比对来分析序列。在某些实施方案中,通过与从cDNA数据库检索到的非编码cDNA序列进行比对来分析所述序列。在一些实施方案中,本公开内容提供了对样品中的mRNA(包括来自被激活用于进行转录的每一种基因的序列变体)进行定量的方法。

[0024] 在某些实施方案中,本公开内容涉及对溶液中具有不同长度和组成的多种异质多核苷酸一起进行全长测序的方法。通常,所述异质多核苷酸为RNA,例如mRNA或微小RNA。通常,所述mRNA是成熟且加帽的mRNA。在某些实施方案中,所述异质多核苷酸是微生物基因组或病毒基因组。

[0025] 在某些实施方案中,本公开内容涉及包括以下步骤的方法:a)提供通常少于1000、2000或5000个核苷酸的双链核酸片段,其包括标记部分和靶标部分,其中所述标记部分包含重叠序列片段和变化序列片段,并且其中所述重叠序列包含第一引物位点和限制性位点;b)将所述双链片段与针对限制性位点的限制性酶混合,提供了经切割的片段;c)将所述经切割的片段在使得所述经切割的片段形成环状片段的条件下与酶混合;d)使所述环状片段在随机点断裂,提供了经剪切的片段;e)将衔接体连接到所述双链核酸的末端,其中所述衔接体包含第二引物位点,提供了衔接体核酸缀合物;f)使用针对第一和第二引物位点的引物来扩增所述衔接体核酸缀合物,其中所述第一引物在5'端包含第一捕获序列,所述第二引物在5'端包含第二捕获序列,以提供捕获靶标标记的缀合物;以及g)对所述捕获靶标标记的缀合物进行测序。

[0026] 在某些实施方案中,所述变化序列片段在所述第一引物位点和所述靶标部分之间。在某些实施方案中,所述第一引物位点在所述变化序列片段和所述靶标部分之间。在某

些实施方案中,所述限制性位点在所述变化序列片段和所述第一引物位点之间。在某些实施方案中,所述变化序列片段在所述限制性位点和所述第一引物位点之间。在某些实施方案中,所述核酸片段包含两个变化序列片段,其中所述变化片段是相同的序列并且所述限制性位点在所述相同的序列之间。

[0027] 在某些实施方案中,本公开内容涉及包括以下步骤的方法:a)将样品和一组标记多核苷酸混合,其中所述样品包括具有不同长度和/或不同序列的核酸的混合物,其中所述标记多核苷酸单独地包括重叠序列和随机序列部分,并且其中所述混合在使得所述标记多核苷酸与所述核酸结合以形成单独地被随机序列标记的核酸的条件下进行;b)将单独地被随机序列标记的核酸混合物复制成同聚物的混合物,其中所述同聚物包括重复的核酸和重复的序列标签;c)使所述同聚物在随机点断裂,提供了同聚物片段;d)将所述同聚物片段与限制性核酸酶混合,所述酶切割与所述标记多核苷酸上的重叠序列相关的位点,提供了经切割的同聚物片段;以及e)对所述经切割的同聚物片段进行测序。

[0028] 在某些实施方案中,所述方法还包括以下步骤:鉴定同聚物片段内的标记序列,分离随机序列部分内的相同序列,以及重新构建所述样品中的核酸序列。在其他实施方案中,所述标记多核苷酸包含被配置为自身杂交成为双链片段的回文序列,其中所述双链片段包含限制性位点。通常,所述限制性位点为不常见的限制性位点。所述标记多核苷酸可共价或非共价地结合所述核酸。

[0029] 在某些实施方案中,本公开内容涉及包括以下步骤的方法:a)将样品和一组标记多核苷酸混合,其中所述样品包括具有不同长度和/或不同序列的核酸的混合物,其中所述标记多核苷酸单独地包括重叠序列和随机序列部分,并且其中所述标记多核苷酸包含被配置为自身杂交成为双链片段的回文序列,其中所述双链片段包含限制性位点,其中所述随机序列部分在所述双链片段内,并且其中所述混合在使得所述标记多核苷酸与所述核酸结合以形成单独地被随机序列标记的核酸的条件下进行;b)将单独地被随机序列标记的核酸混合物复制成同聚物的混合物,其中所述同聚物包括重复的核酸和重复的序列标签,以产生同聚物片段;c)将所述同聚物片段与限制性核酸酶混合,所述酶切割与所述标记多核苷酸上的重叠序列相关的位点,提供了经切割的同聚物片段;以及d)对所述经切割的同聚物片段进行测序。

[0030] 在某些实施方案中,本公开内容涉及包括以下步骤的方法:a)将具有不同大小和/或序列的多于3、4、5、10、100或1000种独特的mRNA与包含具有多于4、5、6、7、8、9或10个核苷酸的多聚-T尾的发夹多核苷酸在使得形成发夹mRNA缀合物的条件下混合,其中所述发夹多核苷酸包含在所述发夹的双链部分内的变化序列片段和在所述发夹的双链部分内的限制性位点;b)将所述发夹mRNA缀合物与复制试剂在使得形成cDNA互补物的条件下混合;c)使cDNA互补物环化;d)通过与引物和复制试剂混合来扩增所述cDNA互补物,形成了具有独特的mRNA序列和独特的序列片段的双链同聚物;e)将所述双链同聚物与针对所述发夹多核苷酸序列中的限制性位点的限制性酶混合,形成了片段,或与在发夹多核苷酸序列中具有切割位点的序列特异性化学试剂混合,形成了片段;以及f)对所述片段进行测序。

[0031] 在某些实施方案中,本公开内容涉及包括以下步骤的方法:a)将具有不同大小和/或序列的多于3、4、5、10、100或1000种独特的环状mRNA与包含多于4、5、6、7、8、9或10个核苷酸的多聚-T尾的发夹多核苷酸在使得形成发夹mRNA缀合物的条件下混合,其中所述发夹多

核苷酸包含在所述发夹的双链部分内的变化序列片段和在所述发夹的双链部分内的限制性位点缀合物;b)将所述发夹mRNA缀合物与复制试剂在使得形成环状cDNA互补物的条件下混合;c)通过与引物和复制试剂混合来扩增环状cDNA互补物,形成了具有独特的mRNA序列和独特的序列片段的双链同聚物;e)将所述双链同聚物与针对所述发夹多核苷酸序列中的限制性位点的限制性酶混合,形成了片段,或与在所述发夹多核苷酸序列中具有切割位点的序列特异性化学试剂混合,形成了片段;以及f)对所述片段进行测序。

[0032] 在某些实施方案中,本文公开的方法还包括对独特的序列片段进行分组以重新构建所述mRNA序列以及在计算机上记录所述序列的步骤。

[0033] 在某些实施方案中,本公开内容涉及扩增样品中多种mRNA的方法,其包括:a)将多种标记多核苷酸与包括多种mRNA的样品在使得所述标记多核苷酸与所述mRNA杂交的条件下混合,形成了mRNA标记试剂核酸,其中所述标记试剂多核苷酸包含多聚-T序列、基本上不同的随机序列的序列可识别区,以及限制性位点;b)将所述mRNA标记试剂核酸与逆转录酶在使得形成互补的标记核酸的条件下混合;c)使所述互补的标记核酸环化,提供了标记的环状互补核酸;以及e)扩增所述环状互补的标记核酸,提供根据mRNA源分子进行标记的经扩增的互补核酸。通常,所述限制性位点是不常见的限制性位点。

[0034] 在某些实施方案中,环化互补的标记核酸提供了环状的单链互补标记核酸。在某些实施方案中,环化所述单链互补标记核酸包括将所述单链互补标记核酸与无法连接双链核酸的连接酶混合。在某些实施方案中,扩增所述环状的互补标记核酸包括生成所述互补标记核酸的重复序列。在某些实施方案中,生成互补标记核酸的重复序列包括将所述环状的单链互补标记核酸与聚合酶和引物混合,其中所述聚合酶在核酸合成过程中从模板上置换双链核酸。在某些实施方案中,所述引物是随机序列,或与标记多核苷酸上的区域杂交,或与来自靶基因或来自靶多基因家族的成员或来自多个多基因家族的成员的序列杂交。在某些实施方案中,所述引物是选自五聚物、六聚物、七聚物及其组合的随机序列。

[0035] 在某些实施方案中,本公开内容涉及包括以下步骤的方法:使经扩增的互补标记核酸断裂成包含序列可识别区的片段。

[0036] 在某些实施方案中,断裂所述经扩增的互补标记核酸包括通过所述标记试剂序列中预定位点中的物理破坏和/或化学破坏而进行的随机和/或特异性断裂。在某些实施方案中,所述片段为少于2000、1000或500个核苷酸。在某些实施方案中,本文公开的方法包括以下步骤:对所述经扩增的互补标记核酸片段进行测序,将所述经测序的片段存储在计算机中,分析序列可识别区和重叠独特随机序列以产生样品中单独的mRNA序列;以及鉴定样品中单独的mRNA序列的模式。所述模式可与所述样品的表型相关。在某些实施方案中,所述样品包括疾病细胞,例如癌细胞。

[0037] 在某些实施方案中,本公开内容涉及标记试剂,例如I型、II-ps₁型、II-ps₂型、II-pa₁型、II-pa₂型和II-t型。在某些实施方案中,所述标记试剂包含随机序列的第一序列可识别区和随机序列的第二序列可识别区以及多聚-T片段。在一些实施方案中,所述第二序列可识别区是所述第一序列可识别区的反向互补物。通常,所述标记试剂还包含形成限制性位点序列的回文序列,其中所述回文序列位于所述随机序列的第一序列可识别区和所述随机序列的第二序列可识别区之间。在一些实施方案中,所述标记试剂包含环形(loop)序列,其具有多于5或10个核苷酸的引物位点序列。在一些实施方案中,所述标记试剂多核苷酸包

含所述多聚-T序列和所述随机序列的第一序列可识别区之间的引物位点。

[0038] 在某些实施方案中,本公开内容涉及包括多核苷酸的混合物的组合物,所述多核苷酸各自单独地包含重叠序列、随机序列部分、多于5、10或15个核苷酸的多聚-T部分以及限制性位点。在一些实施方案中,所述多聚-T位于3'端,并且所述随机序列部分在所述多聚-T和所述限制性位点之间。在一些实施方案中,多核苷酸包含被配置为自身杂交成为双链片段的回文序列,其中所述双链片段包含限制性位点。在一些实施方案中,所述随机序列部分在所述双链片段内。在一些实施方案中,所述多聚-T位于3'端,第二多聚-T位于5'端。通常,所述限制性位点为不常见的限制性位点。所述随机序列部分通常包含散布着(intersperse with)重叠序列的随机碱基位点或序列。

[0039] 在某些实施方案中,本公开内容涉及包括多核苷酸的混合物的组合物,所述多核苷酸各自单独地包含重叠序列、随机序列部分、重复相同随机序列的第二部分、多于5、10或15个核苷酸的多聚-T部分,以及位于所述随机序列部分和所述重复相同随机序列的第二部分之间的限制性位点。

[0040] 在某些实施方案中,本公开内容涉及包括多核苷酸混合物的组合物,其中单独的核苷酸的一部分序列包含基本上重叠的序列,并且单独的核苷酸的一部分序列包含基本上非重叠的序列,其中单独的核苷酸包含多于5、10、15或20个核苷酸的具有胸腺嘧啶或尿嘧啶碱基的重复核苷酸序列,并且其中所述基本上重叠的序列包含不常见的限制性位点。通常,单独的核苷酸包含多于50、100或150个核苷酸碱基。通常,单独的核苷酸包含少于500、1000、2000、5000或10000个核苷酸碱基。在一些实施方案中,所述多核苷酸混合物还包括具有重叠和非重叠序列部分的回文序列核苷酸。通常,所述回文序列产生形成发夹的结构,其中多于10、20、50个碱基对序列彼此杂交。

[0041] 在某些实施方案中,本公开内容涉及包括多核苷酸混合物的组合物,其中单独的核苷酸的一部分序列包含基本上重叠的序列,并且单独的核苷酸的一部分序列包含基本上非重叠的序列,其中单独的核苷酸包含多于10、15或20个核苷酸的具有腺嘌呤碱基的重复核苷酸序列,并且其中在单独的核苷酸中的重叠序列的一部分是回文序列。在一些实施方案中,在单独的核苷酸中非重叠序列的一部分是回文序列。在一些实施方案中,在单独的核苷酸中重叠序列的一部分不是反向互补物。

[0042] 在某些实施方案中,本公开内容涉及本文所公开的环形引物, RNA第二链引物, PCR引物, 衔接体, 单链、截短的且经分离的核酸。

[0043] 在某些实施方案中,本公开内容涉及用于本文公开的试剂的液相或固相合成的方法。在某些实施方案中,本公开内容涉及使用本文公开的试剂并结合大规模平行基因组测序技术的方法,包括分析转录组的统计学结构和组成,分离通过大规模平行测序鉴定的分子变体。

[0044] 在某些实施方案中,本公开内容涉及分离靶核酸的方法,其包括:a) 提供片段化的双链核酸,其包括标记部分和靶标部分,其中所述标记部分包含重叠序列片段和变化序列片段,其中所述重叠序列包含第一引物位点和第二引物位点,其中所述变化序列片段位于所述第一和第二引物位点之间,其中所述第一引物位点和第二引物位点是所述核酸的相反链上的相同序列;b) 将衔接体连接到所述双链核酸的末端,提供了衔接体核酸缀合物,其中所述衔接体包含第三引物位点;c) 在使得第一引物与第一和第二引物位点杂交并且第三引

物与第二引物位点杂交的条件下,通过混合所述衔接体核酸缀合物、第一引物、第二引物和复制试剂来扩增靶标部分,提供了所述引物位点之间的经分离的核酸,其包含变化序列和靶标部分。

[0045] 在某些实施方案中,所述核酸包含所述重叠序列内的限制性位点,并且所述方法还包括以下步骤:将所述核酸与针对所述限制性位点的限制性酶混合,提供了经切割的核酸,然后将衔接体连接到所述经切割的双链核酸上。在一些实施方案中,所述限制性位点位于所述变化序列和所述引物位点之间,在一些实施方案中,所述限制性位点位于两个相同的变化序列之间。

[0046] 在某些实施方案中,本公开内容涉及扩增样品中的多种mRNA的方法,其包括:a)在使得标记试剂多核苷酸与mRNA杂交的条件下,混合多种标记多核苷酸和包含多种mRNA的样品,形成了mRNA标记试剂核酸,其中所述标签包含回文序列,形成了限制性位点序列、随机序列的第一序列可识别区、随机序列的第二序列可识别区和一个末端附近的多聚-T片段,其中所述第二序列可识别区是所述第一序列可识别区的反向互补物,并且其中所述回文序列位于随机序列的第一和第二序列可识别区之间;b)在使得形成互补的经标记的核酸的条件下,将mRNA标记试剂核酸和逆转录酶混合;c)从mRNA中分离互补的经标记的核酸,提供了单链互补-接合物(joiner)核酸;d)使所述单链互补的经标记核酸环化,提供了环形互补的经标记核酸;和e)扩增所述环形互补的经标记核酸,提供了经扩增的互补-接合物核酸;f)通过与限制性酶混合使经扩增的互补的经标记核酸断裂成包含序列可识别区的片段。

[0047] 在某些实施方案中,本文公开的方法包括将标记物缀合至所述片段的末端,提供了标记物结合的经扩增的互补的经标记核酸片段。在一些实施方案中,所述标记物是生物素。在某些实施方案中,所述方法还包括以下步骤:使所述片段环化和片段化;通过将所述标记物结合的经扩增的互补的经标记核酸片段与结合所述标记物的底物混合来纯化所述片段;以及释放所述经扩增的互补的经标记核酸片段并且测序所述片段。

[0048] 在某些实施方案中,本公开内容涉及产生核酸的方法,其包括:a)将引物和复制试剂与起始发夹多核苷酸混合,所述多核苷酸包含3'多聚-T、重叠序列、随机序列部分和环序列(loop),其中所述引物与所述环序列结合,以形成部分双链和部分单链的核酸;以及b)将所述部分双链和部分单链的核酸与多聚-A引物和复制试剂混合,以形成完全双链的核酸。在某些实施方案中,所述方法还包括切割所述多聚-A引物以提供具有多聚-T尾的双链核酸的步骤。在某些实施方案中,所述方法还包括使所述双链核酸变性以形成具有多聚-T尾的发夹核酸和所述起始发夹多核苷酸的步骤。所述起始发夹多核苷酸通常缀合至固体载体。

[0049] 在某些实施方案中,本公开内容涉及通过上述方法制备的固体载体,其包括如本文所述的发夹多核苷酸。

[0050] 在某些实施方案中,本公开内容涉及产生多核苷酸的方法,其包括:a)混合除了环形序列以外基本上是双链核酸的模板多核苷酸、针对环形序列引物位点的引物、以及聚合酶,其中所述模板多核苷酸包含环形序列引物位点和所述双链核酸内的第二引物位点,其中所述聚合酶在核酸合成过程中从所述模板置换双链核酸以形成部分双链和部分单链的核酸;b)将所述部分双链和部分单链的核酸与针对所述第二引物位点的引物以及聚合酶混合,其中所述聚合酶在核酸合成过程中从所述模板置换双链核酸以形成双链核酸;以及c)加热所述双链核酸以变性、释放出多核苷酸并重新形成所述模板多核苷酸。在某些实施方

案中,将所述模板多核苷酸缀合至固体载体;所述模板多核苷酸在一个末端附近所述多聚-A片段相邻;所述模板多核苷酸包含与用于第二链合成的引物位点相邻的可识别随机序列的第一区域;并且所述模板包含含有与环形序列相邻的限制性位点的回文序列。在某些实施方案中,所述回文序列与可识别随机序列的第一区域相邻。

[0051] 在某些实施方案中,本公开内容涉及分析细胞或组织的分子表型,分析疾病细胞或组织,以及建立转录组数据库。

[0052] 在某些实施方案中,本文公开的方法包括以这样的方式加工经标记的寡核苷酸以进行扩增,然后以这样的方式加工每个片段拷贝以使初始的标签与所产生的内部片段一起被复制,以及对重新构建初始的异质溶液中寡核苷酸的序列和相对数量所需的相关信息进行分析计算机恢复。

附图说明

[0053] 图1A示意性地说明了标记物块 (Marker-Block) 的嵌套结构,其包括SMID、5'和3'封套(wrapper)序列和包括5'尾、3'连接元件(例如,多聚-T尾)和环形结构的侧翼区。在本实施方案中,标记物块包含被保留在最终测序文库中并且可用于根据批次、来源和链来对读段进行分类的那些序列。标记物块可包含多种序列元件,例如,用于标记试剂链(TRS)合成中的引物序列和退火位点,或PCR引物,簇合成和测序元件,用于本文所示的改造(adaptation)中的测序方案的限制性酶切位点。

[0054] 图1B表示I型、II-ps型、II-pa型和II-t型标签,示出了5'尾、标记物块、环和3'连接元件。在本文所述的实施例,3'连接元件是3'单链寡聚-dT (T₂₂或T₂₂V)。可变残基(V=A、G或C)从与约300-400个碱基的多聚-A尾非常接近的编码信息的3'末端碱基引发合成。标签由四种实例代表。单SMID单链I型标记物在引发cDNA合成和环化反应方面是高度有效的;这些标记物是原型分子,其中在文库制备过程中产生的限制性切割片段中仅有一半将被末端标记。II-ps型标签引入SMID的两个拷贝,并且两个酶切片段都将被末端标记。这种试剂并不直接鉴定源分子模板的义(sense),其可仅通过参考基因序列来评估。II-pa型标签类似于II-ps型标签,但是也允许通过计算机来区分模板源链的义,因此允许鉴定多聚-A非编码RNA。II-t型标签具有独特的性质,即SMID序列相对于所述源链的义来取向,不需要计算机分析步骤。这些标签还允许特定的SMID鉴定的cDNA可以通过PCR直接扩增,用于克隆和表达。

[0055] 图1C表示标记试剂的确认。说明了II-pa型和II-ps型的两种一般构象。

[0056] 图1D说明了标记物块的自身互补性;I型和II-t型通常缺少二级结构(参见上述B)。II-pa型和II-ps型标签的互补性导致当逆转录物被转换为双链cDNA时引入独特的分子标识符(SMID)的两个相同拷贝,使片段标记的效率加倍。对这些区域自身退火的偏好允许合成后对TRS的有效纯化;此外,该偏好阻断了文库形成过程中在两个末端被标记的片段(例如,完整的cDNA)的扩增。

[0057] 图1E表示具有特定DNA序列的I型和II-pa型的标记试剂。在II-pa型中,所示的标记物块内部是用于在配对测序中使用具有PCR 1.0的Illumina衔接体、捕获、簇合成、测序和限制性酶序列的序列。示出了标记元件,包括SMID、有义链区分元件和限制性位点。

[0058] 图2A表示单尾的II-ps₁型和II-pa₁型标记试剂的II型标记试剂的合成。此处所述

的反应可在溶液中进行,不涉及固相,或使用固相步骤。单尾或双尾II-p型形式的合成具有共同的前体和第一步骤;II-ps和II-pa形式的区别源自标记物块的具体内容(参见图1C)。

(1) HP是TRS随后在其上合成的发夹构象模板。使用DNA聚合酶Phi 29,通过延伸前体寡聚-J来生成HP。寡聚-J前体具有中等稳定性的末端夹板(cclamp),其导致分子向自身折叠以引发其延伸,产生标记物的独特SMID元件的分子内互补物。该反应通常是定量的(参见图7A)。

(2) 从HP合成3Q。此第二反应也由Phi 29DNA聚合酶介导。环形引物退火至HP的开放环以引发HP的5'部分的一个拷贝(1Q),释放3'端作为单链DNA以允许进行反应(3)。环形引物通常在5'末端是非磷酸化的,但是也有例外。1Q-HP复合物被指定为3Q。(3) TRS-HP复合物。本反应与反应(2)同时进行,由相同的酶介导。2SP RNA保护引物退火至3Q的3'端以保护末端免受Phi 29的核酸外切酶活性并引发RNA嵌合的TRS-HP复合物的合成,从3Q释放出游离的1Q。在Phi 29失活之后,RNA酶H移除RNA部分。该反应通常是定量的。(4) 通过TRS-HP的凝胶纯化或其他纯化除去1Q、反应引物和酶。可通过Phi 29的3'→5'核酸外切酶活性除去1Q和过量的反应引物。(在1Q作为能够引发cDNA合成的污染物被保留的情况下,由于缺少5'磷酸基所获得的产物将无法在随后的反应中环化。)瞬时变性不可逆地分离了单链HP和TRS DNA。由于强的内部互补性,自身退火从数量上控制了所述两条链的重新结合,阻止复合物的重新形成。作为等摩尔副产物,HP并不会导致不利后果,它在文库制备中对于cDNA引发或环化都是惰性的。反应(2)和(3)通常是定量的(参见图7B)。II-pa型/HP₁或II-ps型/HP产物试剂是纯的并且直接用于合成经标记的cDNA。

[0059] 图2B说明了双尾II-ps₂型和II-pa₂型标记试剂的制备。这些反应通常是当一些或全部步骤使用固体底物进行时实施。在商业上,制备具有共价延伸的寡聚-J前体,所述共价延伸可连接至生物素基团或可以是用于共价连接至葡聚糖凝胶、玻璃或其他固体底物(星号)的化学接头。此处,示出了使用生物素化的形式,反应中的第一步骤在溶液中进行。(1) 如上述图2A(1,2)进行HP延伸和3Q合成。(2) 使用2sP引发TRS-HP'合成,所述2sP仅退火至3Q的3'单链末端中封套序列的5'区。2sP的非互补部分被示为寡聚-dT,但是可以是变化的。按照上述图2A(3)置换1Q。有意地不阻止Phi 29的3'核酸外切酶活性,最后消除HP的3'单链部分,然后合成新的3'尾作为2sP的非互补部分的拷贝,以生成修饰的模板,将其指定为HP'。(3) TRS-HP'复合物被吸附到链霉亲和素小珠上;通过冲洗除去1Q、引物和酶。瞬时变性后洗脱纯的TRS以将所述链分开;在洗脱过程中可保持温和变性条件以阻止与HP'的互补叉状尾(forked tail)退火。反应通常是定量的(参见图7C)。(4) 因为HP'是在固相上重新生成,其可用于进一步的循环以合成TRS链。如果HP'共价连接至固相,该反应效果最佳。不再需要使用环引物来打开HP'结构以引发TRS合成,并且不合成1Q。与HP'的暴露的3'叉互补的经修饰的2sP(未示出)引发TRS合成,冲洗除去残余的2sP和酶。这使得重复循环从相同的模板进行的TRS合成。图例地示出了固相合成的循环。

[0060] 图2C表示II-t型标签的合成。

[0061] 图3A表示用于文库制备的方法的一个实施方案。用标记试剂引发cDNA,然后除去RNA并用单链RNA/DNA连接酶进行环化反应。用核酸外切酶I除去线性残留物。

[0062] 图3B表示区分加帽和未加帽的mRNA。对于一些应用而言,需要区分成熟的Gppp加帽的mRNA和具有末端5'磷酸基或5'羟基的不成熟形式。(1) 标记。(a) 加帽形式:用碱性磷酸酶处理总多聚-A mRNA以除去5'磷酸基末端,留下未加帽的分子作为5'羟基末端。随后用烟

草酸磷酸酶除去Gppp帽释放了可用单链RNA/DNA连接酶进行环化的5'磷酸基形式。(磷酸化形式——在环化前省略(1)中使用的两个酶步骤,仅在信使群体中捕获作为环状RNA形式的内源5'磷酸基形式。)(b)完全未加帽形式。在环化之前用多核苷酸激酶处理多聚-A mRNA并省略(1)的烟草酸磷酸酶步骤,使5'羟基形式磷酸化。在随后用单链RNA/DNA连接酶处理过程中留下加帽的分子不进行修饰,导致仅有未加帽的mRNA群体被环化。

[0063] 图3C表示cDNA合成。低于化学计量的($<<1:10$) II型试剂(例如,II-ps₂型或II-pa₂型)双尾变体退火到之前环化的信使RNA和残余的线性RNA,然后进行逆转录。cDNA合成之后,通过T4DNA连接酶进行有效的cDNA分子内连接以形成环状cDNA。环状而非线性mRNA为酶作用提供了增强速率的“夹板(splint)”；此外,从退火位点的3'多聚-A信使的“突出”通常阻止环状的分子内或分子间DNA连接。使用RNA酶H除去RNA。使用核酸外切酶I除去线性单链cDNA(未示出)。随后以与从总多聚-A mRNA生成的产物(3A)同样的方式加工所得的环化单链cDNA。

[0064] 图4表示滚环扩增。每种经标记的环化单链cDNA退火到多个引物(例如,但不限于硫代磷酸基随机六聚物)；使用DNA聚合酶Phi 29催化第二链合成。因为这种高度进行性的酶环绕模板并与自身末端或另一种酶的第二链相遇,所述第二链被置换为长同聚物；对这条链的随后引发相应地产生了双链产物,其通常在所述逆反应的多个引发位点处分支化。也会产生线性同聚物。从双链同聚物DNA的II-ps型和II-pa型标签生成SMID的第二拷贝。通过标记试剂的插入拷贝来分离每种cDNA的串联拷贝,在这种情况下包含对称排列的独特SMID的双拷贝；通过具有不常见的限制性酶切位点的环衍生片段彼此分开。在下一个步骤之前使用单链核酸酶(SI或绿豆核酸酶)可使同聚物去分支化。

[0065] 图5表示RCA扩增的cDNA同聚物的加工。A)片段化。在去分支化之后,通过超声、酶促片段化、水剪切或同等的物理过程使串联的(concatenated)同聚物片段化成具有平均靶cDNA大小的数量级的任意长度。B)限制性切割。使用不常见的限制性酶(或其他序列特异性切割试剂)在标记物块之间的环衍生连接元件(connector)中的位点处切割随机片段。具有经标记的接合点的片段将产生两条切割链,其具有通过酶促片段化产生的在一个末端和另一个末端包含独特的SMID的标记物块,或通过超声产生的来自cDNA内部序列的随机末端。通常,使用任一种下一代鸟枪测序平台的方案,使用这些片段产生双端或配对文库。此处,通过对用于Illumina High Seq或Illumina MiSeq设备的配对方案的改造来说明这一点。C) Illumina配对方案中的末端标记片段的结局(fate)。Illumina配对方案广泛用于基因组测序,其适用于显著扩大待测序的末端的分离。这适用于cDNA测序,其中末端标记物在每个cDNA的长度(平均为1.7kb,最长为15kb)上必须与随机读段相关。此处示出了标准Illumina配对方案的初始步骤中末端标记的片段的结局。缺少标记物块的片段将经历这些步骤,以待随后被消除。(1)在初始剪切(超声或酶促片段化)和限制性消化之后,对DNA片段进行末端修复、末端生物素化和环形连接。用核酸外切酶I和II除去线性残余物。(2)通过雾化作用破坏环化的cDNA,然后通过凝胶纯化300–500bp的片段。通过链霉亲和素小珠捕获生物素化的接合片段,并通过冲洗除去未接合的片段。

[0066] 图6A示意性地说明了含有标记物的片段的选择性扩增。A)在标准方案中,被捕获的接合点具有A尾并连接至允许进行PCR扩增的Illumina叉状衔接体(PCR引物1.0;PCR引物2.0),所述PCR扩增提高每一种片段的丰度同时引入独特的双末端,其除了包含PCR引物位

点之外还包含捕获、簇合成、A和B型限制性位点序列以及测序引物。

[0067] 通过将PCR引物1.0的退火位点引入标记物序列来修改这一步骤。代替专用的叉状衔接体,经修饰的衔接体使仅与PCR引物2.0互补的片段连接至所述片段的3'端。因此,(a)仅扩增了链霉亲和素捕获的包含标记物的接合序列;(b)在阶段I读取开始时对标记物序列(包含SMID)进行测序,扩展到具有酶促片段化或超声随机断裂点的标记物-接合点中。阶段II报告了从通过雾化作用产生的第二随机断裂位点开始的读段序列。因此,选择性地报告与包含SMID的标记物结合的两个内部序列,鉴定其初始来源分子。标记试剂的区别在于其报告源分子链的义的能力。此处所述的图形元件描述了使用I型、II-p型和II-t型标记物的扩增反应。

[0068] 图6B.a) I型标记物块在SMID的5'和3'侧具有两种类型的限制性位点。b)在cDNA合成和扩增后,对cDNA进行酶促片段化或通过超声进行随机片段化;象征性地示出剪接位点的位置。c)示出了样品片段。d)在酶促片段化或超声之后,将片段分成等分试样,使用一种或其他限制性酶切割并重新组合(5B的修改)。e)对片段进行末端修复、生物素化、环化,雾化并在链霉亲和素小珠上捕获接合片段。然后使其形成A尾,并使3'端连接到仅具有PCR引物2.0的退火位点的经修饰的衔接体上。f)使用引物1.0和引物2.0进行扩增,所述引物1.0与标记物上的互补序列相互作用,所述引物2.0退火至随机断裂点处的3'衔接体。扩增的结果是,在最终的配对文库中仅表示了在一端具有标记物且在另一端具有随机断裂点的片段。g)配对文库的首次序列读取产生标记物(封套和相关的SMID序列)和源分子链的义。

[0069] 图6C.以相同的方式加工II-pa型和II-ps型标记的cDNA;示出了II-pa型加工。a) II型标记物块在重复的SMID之间的环中具有不常见限制性位点的两个拷贝,以及PCR引物1.0及其互补物的两个序列。b)合成cDNA,使其扩增和片段化(酶促片段化或超声);示意性地示出随机断裂位点。c)示意性地示出片段。d)通过限制性酶切割除去SMID之间的片段。对片段进行末端修复、生物素化、连接成环、雾化并在链霉亲和素小珠上捕获接合的片段。环化使由酶促片段化或超声产生的随机或非随机断裂位点位于SMID或SMID互补物的附近。e)对捕获的生物素化的接合雾化片段进行末端修复,使其形成A尾并将其3'端连接到仅与PCR引物2.0互补的经修饰的衔接体。f)使用PCR引物1.0或该引物的修饰形式以及使用PCR引物2.0进行PCR,所述PCR引物1.0或该引物的修饰形式与其在标记物块中的互补位点相互作用,所述PCR引物2.0与随机断裂位点相关的衔接体相互作用。g)该扩增的结果是,(a)仅携带标记物的片段包含配对文库;(b)在初始读取中选择性报告标记物序列,然后是来自初始片段化(酶促片段化或超声)的内部断裂序列;(c)配对序列来自雾化作用产生的第二随机断裂位点。基本上,来自配对文库的全部读段对(read pair)都根据衍生出它们的源分子进行索引。对于II-pa型标记的分子而言,不对称标记物(环形标记)表明了源分子链的义;对于II-ps型标记的分子而言,无法获得这个信息。

[0070] 图6D.a-f)对于II-p型标记物,产生了文库,但是说明了SMID取向的结局。g)在初次读取中揭示了标记物序列(或其互补物),然后是对应于酶促片段化或超声断裂位点的接合点;配对序列揭示出雾化作用断裂位点,如以上所述。可将每个读段对报告为经测序的读段或其互补物。II-t型标记物具有有用的属性——所述标记试剂标记物(通过封套序列和校对碱基(check base)来区分)的互补物独特地与源分子链的序列相关。直接分离cDNA构建体:II型试剂产生这样的cDNA,其同时在3'端和5'端的侧接相同的标记物序列;因此,通

过标记物(例如, SMID) 指导的PCR或亚克隆, 可直接从由酶促片段化或超声样品所保存的等分试样中获得来自特定源分子的全长cDNA。

[0071] 图7A示出了HP延伸的结果(4%琼脂糖凝胶)。用DNA聚合酶Phi 29延伸寡聚-J前体。所述前体条带浅且弥散, 原因是(a) 夹板(clamped)形式和延伸形式之间的构象平衡; 和(b) 单链DNA与溴化乙锭的较弱结合。相比之下, 延伸后, 由于更刚性的双链构象, HP明显成为具有较高表观分子量的明亮条带。

[0072] 图7B示出了II-ps₁型TRS的合成的结果(4%琼脂糖凝胶)。合成、凝胶纯化和链分离之后, 在游离溶液(非固相)中生成的这种物质是分离的TRS和HP的等摩尔混合物。在RCA之前在引发cDNA合成或环化方面, HP是惰性的: (环化后, 通过环化后的核酸外切酶I和II处理除去HP)。转化通常是定量的。

[0073] 图7C示出了II-ps₂合成的结果(4%琼脂糖凝胶)。在HP-TRS合成之后, 吸附于链霉亲和素小珠和冲洗除去1Q、试剂引物和酶; 瞬时加热和变性从固相上释放纯TRS, 生物素化的HP被保留在小珠上。该反应通常是定量的。

[0074] 图7D示出了由I型、II-ps₁型和II-ps₂型标记试剂引发的cDNA合成的结果(4%琼脂糖凝胶)。使用有限量(0.25pmol)的每一种标记试剂对来自人胚胎肾(HEK-293)细胞的多聚-A mRNA进行逆转录。大cDNA聚集在凝胶的上部。每个cDNA条带对应于足够数量的分子(约150,000,000,000), 以大规模地报告存在于细胞中的mRNA群。

[0075] 图7E示出了来自I型、II-ps₁型和II-ps₂型标记试剂的环化cDNA的RCA结果(4%琼脂糖凝胶)。使用硫代磷酸基随机六聚物和Phi 29DNA聚合酶对(D)中生成的cDNA的小等分试样进行RCA, 并且使小等分试样在凝胶上电泳。较大的多联体(concatemer)被捕获在凝胶的上部。这些反应是高度可重复的; 每个RCA反应可足以制备多个文库。

[0076] 图7F示出了去分支化RCA产物的酶促片段化或超声处理和限制性酶切割的结果(1%琼脂糖凝胶)。首先对同聚物进行酶促片段化或超声处理以生成平均约4kbp的片段, 使其在凝胶上电泳。在用不常见的限制性酶AlwNI切割以除去重复标记物之间的片段之后, 注意到分子量改变。所述图表说明了酶处理之前和之后的凝胶扫描。实线预测了100%切割的大小分布; 小偏移量(offset)反映了这一事实: 即数学建模无法估计缺少内部标签的片段。

[0077] 图7H示出了使用PCR扩增接合片段所生成的配对文库的重复凝胶和凝胶扫描结果(Agilent分析凝胶)。在这种情况下, 从II-ps₁型标签内的引物位点和连接至随机断裂位点的末端衔接体中的末端PCR 2.0位点引发PCR。使用High Seq设备的15个实验反应室之一, 用该文库进行的测序运行产生174,000,000个读段对。

[0078] 图7I示出了由上述末端标记的配对文库的High Seq测序生成的末端标记的读段对的结果。(1)展示了来自随机选取的末端标记序列的配对读段的实例; 突出显示标记物的5'和3'封套和读段I中的校对碱基。在下方示出了对应的读段II序列。以红色表示校对碱基中的测序错误。

[0079] 这些数据表明, 所述方案的步骤实现了其所需的目的, 特别是: (a) 标记试剂合成成功产生了预期序列的标记物和用于配对文库的多样性; (b) cDNA合成从组织样品(培养的人胚胎肾(HEK 293)细胞)中标记了单独的mRNA——30000个读段对的测量鉴定出细胞中表达的4000多个基因; (c) 所使用的双标记的II-ps₁型试剂的切割是有效的; (d) 所述反应将具有内部cDNA序列的源分子的独特SMID集合到一起; (e) 对Illumina配对方案的修改产生

末端标记的配对文库,帮助鉴定标记物序列;(f) 配对序列高效率和高保真地与已知基因产物而非基因间DNA序列匹配;(g) 可根据使用独特SMID鉴定的源分子,对数据进行分类。(2) 来自Blast搜索实施例I的配对匹配。在此实施例中,人乙醛脱氢酶基因的mRNA被标记。因为从通常位于雾化片段中间的标记物片段内扩增被测序的片段,cDNA序列的163个碱基(63个来自读段I,100个来自读段II)通常在使用300-400bp雾化片段生成的文库中重叠。通过突出显示查询序列来指示这种重叠。(3) 来自Blast搜索实施例II的配对匹配。在该第二个实施例中,人大脑my047基因的mRNA被标记,观察到同等程度的重叠。尽管应用PCR扩增中使用的稍大的雾化产物可增加可使用的序列,需要一些重叠以控制长测序循环结束时识别碱基保真度的下降(此处是100bp)。

[0080] 图8A图解说明了将序列读段组装成重叠的重叠群(contig),以及可在没有缺口的情况下构建被两个重叠群覆盖的模板,这是能保证的包含三个串联cDNA序列的模板的最小值。

[0081] 图8B是对所计算的参数的读段数量相对于覆盖度作图的图形表示。这示出了总体行为。对于一个转录物长度(T),一个转录物重复值(r)和一个读段长度(L)而言,每条线段连接每个总序列(包括重复)的k(1、2、3、5和10)重叠群的5个数值的点。三个重复水平(r)将图分成三组:(r=1)、(r=2)和(r=3)。这显示出如果转录物以串联方式被拷贝但被固定数量的重叠群覆盖,则单个转录物的覆盖度如何增大。更大的转录物位于上部,因为需要更多的读段来覆盖它们。

[0082] 图8C是对数据的一个子集作图以说明读段长度对捕获全部转录物序列的总覆盖度的影响的图形表示。红色曲线表示9个转录物长度值对5个重叠群值作图,每个重叠群值处于固定的读段长度(50bp)和重复数(3)。蓝色曲线同样如此,但是具有更长的读段长度(150bp)。为保证覆盖转录物的相同水平,与150bp读段相比,更长的转录物对50bp读段需要多读取3.5至4倍碱基。

[0083] 图8D是关于读段、转录物长度、覆盖度和读段数的表格。这个表示对于读段长度的特定覆盖深度的读段数、每个读段长度的重叠群数、重复水平和转录物长度之间的关系的关系的简易型表格,允许估计在实验中生成的配对读段数与使用特定的覆盖度深度可完全测序的给定大小的cDNA数量之间的关系。

[0084] 图9表示经修饰的引物的实施方案。通过在文库制备的PCR步骤中使用针对标签(接合物)的引物,可提高包含SMID的配对序列的产生。这进一步保证SMID序列将位于基本上每个文库片段的末端,以使得读段对之一将在标签序列中产生SMID。因此,在标准配对或双端方案中,用经修饰的引物取代Illumina试剂的A引物。

[0085] 图10表示可来自于假定基因的选择性转录组(alternative transcriptome),所述假定基因经受由于使用选择性启动子(alternative promoter)以及RNA链终止和多腺苷酸化的选择性共有位点(alternative consensus site)导致的5'和3'UTR中的变异,以及由于选择性RNA剪接导致的UTR和ORF中的变异。符号A-E表示与选择性启动子相关的可变3'UTR片段;x-z表示与选择性终止和多腺苷酸化位点相关的不同3'UTR片段;绿色符号代表选择性剪接的外显子,其保留或缺失是通过括号内变体矩阵中的1或0来表示。虚线表示假定的ORF。

[0086] 图11A和11B示出了对从测序运行返回的数据进行计算机分析的流程图。G+T代表

基因组和转录物数据库;PTMP表示部分标记的配对序列(mate pair),其可由具有断裂或原本经部分测序的SMID或有少量PCR或测序错误的SMID的文库片段产生,根据SMID序列及其与特定基因转录物的相关性,对部分标记的配对序列的鉴定被证明超过了特定的可能性阈值(即, $p < 10^{-9}$)。

[0087] 图12表示在原型实验中从标记的cDNA同聚物中测序并通过Illumina HiSeq 2000 配对测序(Nature.2008 456 (7218):53-59)分析的全长信使RNA,所述标记的cDNA同聚物是从人胚胎肾细胞系制备。A图表示使用配对读段拼接成的乳酸脱氢酶A基因的整个序列,每个读段都标记有通过逆转录纳入初始cDNA分子中的SMID的复制序列;表示了经RefSeq注释的外显子边界;通过检查确认末端序列。B图说明了读段的连接。细水平线连接片段和与染色体不相交部分匹配的单一读段。在许多情况下,插入的基因组序列是已被剪接掉的内含子,并且两个侧翼部分实际上是读段中的连续序列。在其他情况下,经连接的序列实际上在读段内被分开;这些来自于配对文库制备中被集合到一起的远距离的cDNA序列。图C说明了所述两种连接。图D表示,在推导的信使RNA上(其缺少外显子2)相同的读段的比对,表明全部分子内剪接点的覆盖度。颜色始终是关于在被测序的配对序列中报告的信使RNA的义。

具体实施方式

[0088] 本文所述的方法克服了RNASeq的某些局限。RNA-Seq以及全基因组测序或“外显子组测序”策略预测或解释细胞和组织表型的局限示于图10中所描绘的两个选择性转录组的示意图中,所述选择性转录组可产自单个假定的断裂基因。显然,即使对单一的选择性剪接片段(此处两个转录组中是相同的)的表达频率进行定量,外显子测序仍无法区分表达完全不同的蛋白的情况——不同的结构背景,其中致病突变可影响功能。

[0089] 技术人员能够确认倾向于在RNA-Seq和一般外显子测序策略中丢失的信息:考虑一系列选择性RNA,其来自于复杂组织中的假定基因 g ,具有衍生自转录激活的选择性5'位点和3'共有多聚腺苷酸化位点的多个UTR,并且UTR和ORF片段都经过选择性的RNA剪接(或RNA编辑)。假设综合测量中的变体以染色体上5'至3'的顺序包括 n 个可变片段(广泛定义为“外显子”)的线性组合, $X = (e_1, e_2, \dots, e_n)$ 。通过 $S_i = (a_{i1}e_1 + a_{i2}e_2 + \dots + a_{ij}e_j + \dots + a_{in}e_n)$ 来指定 i^{th} 信使变体的序列 S_i ,其中 a_{ij} 是保留系数,当外显子 e_j 缺失或保留时, a_{ij} 分别为0或1,或更简单地,

$$[0090] \quad S_i = a_{ig} \cdot X_g$$

[0091] 其中 a_{ig} 是基因 g 的 i^{th} 信使的保留系数的矢量, X_g 是指 g 的保留外显子的集合。

[0092] 认识到即使是对剪接变体的大量测量也可能是不全面的,例如,未经检查的细胞类型可向 X_g 引入其他变化,这个术语被基因边界内的基因组序列取代,写成不同的方式(符号为 Ψ_g)。 Ψ_g 表示类似 X_g 的矩阵,但是包括来自基因边界内所有有序碱基的所有可想到的相邻片段,所述基因边界可位于剪接受体和供体位点之间。因此,

$$[0093] \quad S_i = a_{ig} \cdot \Psi_g$$

[0094] 这一关系从形式上将来自 g 的基因组序列(具体为 Ψ_g)的信使密码中的信息与产生于表达所述变体的细胞中经高度调控的RNA加工的细胞指令(a_{ig})区分开(有时被称为“相互作用组(interactome)”)。这两种信息贡献指定蛋白质的结构和分子特性;它们必须经过变异和自然选择。编码蛋白的基因可在多个方向上同时进化以产生介导不同生物过程的蛋白

质。

[0095] 转录组 τ 有时被视为相对基因表达水平的分布： $\tau = (p_1, \dots, p_g, \dots, p_N)$ ，其中 N 是基因的数量； $p_g = n_g / \sum_{g=1}^N n_g$ ； n_g 是基因 g 的所有变体的信使分子的总数。

[0096] 相应地，每个基因有其自身的转录组， $\tau_g = (q_{g1}, \dots, q_{gk})$ ，其中 k 是基因 g 的信使变体的数量； $q_{gi} = n_{gi} / \sum_{i=1}^k n_{gi}$ ； n_{gi} 是变体 i 的信使分子的数量。将真正的基因序列元件吸收到 τ_g 中重新构建了作为变体序列的加权库存 (weighted inventory) 的转录组，例如，通常需要的信息形式。在以下公式中，

$$[0097] \quad \tau = \sum_{g=1}^N p_g \left(\sum_{i=1}^{k_g} q_{gi} \cdot (a_{ig} \cdot \psi_g) \right) = \sum_{g=1}^N p_g \cdot \tau_g$$

[0098] 最佳实施的RNASeq捕获从 ψ_g 和一些单个剪接点而不是更长的距离连接 (range linkage) 表达的全部信息。RNASeq通常使用复杂的统计算法以预测最可能表达的变体，但是无法通过试验捕获 a_{ig} ；它是计算密集型的不确定策略，而直接测序提供了“地面实况 (ground truth)”数据集，在任何情况下必须针对该数据集验证RNASeq算法。类似地，所述外显子组的基因组测序通过设计省略了指定 a_{ig} 的信息，因此无法明确确定哪些蛋白 (或功能性RNA) 可在细胞或组织中表达。特别设计此处公开的技术，以捕获分子内连接信息和需要用于描述转录组的相对变体分布。

[0099] 如之前的说明书中所说明的，大规模cDNA测序的挑战在本质上与高等物种基因的生物学及其与经典的顺反子基因模型的差异有关。可以容易推断出所述差异的规模：即，所述细菌大肠杆菌 (*Escherichia coli*)，一种约为线粒体大小、具有有限的表型多样性的单细胞生物，具有 (随着菌株变化) 约4700个顺反子蛋白质和功能性RNA编码基因。相反，仅有稍多于4倍的“断裂基因”的人受精卵产生100万亿个细胞——每个细胞都比细菌复杂得多——以形成组织、器官和器官系统 (包括人大脑的1000亿神经元和100万亿突触)。顺反子和“断裂”基因的信息不对等是明显的。

[0100] 在本公开内容的一个典型的实施方案中，在逆转录时用独特的标记试剂标记每个mRNA，所述标记试剂包含源分子标识符 (SMID) 的一个或多个拷贝。逆转录之后，全长、经标记的cDNA经历一系列步骤，所述步骤产生重叠片段文库，其中每个片段具有初始SMID的一个拷贝。与精选的 (curated) 基因序列的比对能够鉴定新的外显子并确认之前注释的外显子。

[0101] 可使用几种已知的鸟枪测序平台的任何一种对该文库进行测序，以产生特定源分子的标记物和来自该分子的随机序列。然后将序列对分成特异性针对每个源cDNA分子的组 (“区段 (bin)”)，然后组装出该cDNA的全长、端到端的序列。

[0102] 通常，在样品中每个分子的预设的覆盖度水平 (例如，5x、10x、25x等) 下，所述方法允许进行完全测序：由更高的覆盖度深度带来的精确度使得能够检测单碱基突变、单核苷酸多态性 (SNP) 或RNA编辑位点。独特的标识符的分布允许重新构建初始mRNA群的结构。产生完整的序列，包括5'非编码区 (UTR)、开放阅读框 (ORF) 和3'UTR，包括沉默突变或错义突变。这可使得转录激活控制和序列元件的转录后拼接相互关联。所述方案的中间步骤保留了初始cDNA池的全长样品，其可用于立即克隆通过测序检测到的任何特定的源mRNA的拷

贝。可通过异质表达来表征这类克隆,或使用这类克隆来制备分子探针。可进行文库构建以同时分析来自多个组织来源的多个转录组。因此,可在同一实验中加工来自同一患者的多个肿瘤的mRNA,或来自不同发育阶段或疾病进展的mRNA,测序后对数据进行分类(多路复用)。

[0103] 除了对mRNA进行测序,以及所产生的基础和临床研究应用,本文所述技术具有其他可能的用途:对在患者疾病进展过程中进化的逆转录病毒的异质基因组进行测序;监测动物储存库中人和动物病毒元件的重组,这是流感病毒变异产生的基础;在进行预防性抗生素治疗的动物的微生物种群中,或在进行改变土壤生态学的杀虫剂方案的农业作物中出现的微生物中,分析与药物抗性或毒素产生相关的基因交换动力学。

[0104] 一个设想的应用在于个体化医疗的新兴领域。某些治疗产品以不同方式影响人。个体化医疗寻求使用个体患者的遗传信息以预测或解释疾病的发生以及选择或优化治疗策略。尽管这一方法取得了显著的成功,但在大多数病例中,个体化医学的目标仍未达到。即使在最好的情况下,更多了解细胞或组织的分子表型将明显对结果产生益处,这通过若干实施例来说明。

[0105] 在乳腺癌治疗中,对肿瘤样本中HER2/neu——一种参与信号转导途径或细胞增殖的酪氨酸激酶——的组织学检查可指向使用曲妥珠单抗(trastuzumab)(赫赛汀(Herceptin)),一种针对HER2/neu胞外结构域的重组人单克隆抗体。成功的治疗下调作为细胞增殖的直接原因的HER2/neu表达,并可引发对癌细胞的免疫杀伤。不幸的是,在大多数患者中出现对曲妥珠单抗的原发性和获得性抗性;对心脏组织的破坏性脱靶效应使其无法用于原本适合的有冠心病史的患者中。此外,在三名阴性乳腺癌患者中,三个细胞标记物HER2/neu、ER(雌激素受体)和PR(孕激素受体)都不表达:这些癌症通常更具侵入性,通常在年轻女性或非裔美国女性中被诊断出,并且缺少良好的诊断标记物或治疗药物靶标。全面的cDNA分析将可能提供对疾病机理的认识、新的诊断标记物和可能的治疗药物靶标。

[0106] 在某些实施方案中,本公开内容涉及癌症基因组测序方法。在某些实施方案中,本公开内容涉及比较特定基因组的方法,所述方法通过平行分析肿瘤和正常对照细胞的mRNA以检测导致癌症表型的突变蛋白、激活的基因级联和其他标记物,所述癌症表型例如在由于随机积累的信使突变导致的变异背景下,指定细胞谱系为“黑色素瘤”的那些特征。鉴定独特的癌症蛋白将提供药物或免疫抑制治疗的靶标。

[0107] 在某些实施方案中,本公开内容涉及免疫系统工程的方法。在引人注目的试点研究中,三名慢性淋巴细胞贫血症患者的免疫系统被重新改造以攻击在正常和恶性免疫β细胞上表达的标记蛋白CD 19。N Engl J Med (2011);365:725-733。在两个病例中,这导致消除了疾病征象的自身免疫攻击;在第三个病例中,注意到显著的改善。不幸的是,在针对实体瘤癌症的类似研究中,因为平均的人类细胞表达25-30%的蛋白编码基因互补物的产物,严重的(在若干病例中是致命的)脱靶反应使与肿瘤共享表面标记物的其他重要器官受到损伤。了解更多的关于在重要组织中表达的特定蛋白质变体可为靶标选择带来必要的改进。

[0108] 在某些实施方案中,本公开内容涉及选择性药物治疗的方法。对受体和酶的选择性调节已有相当多的关注。参见Journal of Clinical Oncology, (2007) 1 25,5815-5824。最多的关注在于表征存在于靶组织中的共激活因子和共抑制因子的相对水平。相比较而言

很少考虑这些受体的约24个剪接变体的组织特异性表达,或孕激素和雌激素受体中类似的变异解释这些差异的可能性。使用本文公开的实施方案进行mRNA分析能够捕获与全部这些可能性相关的信息。

[0109] 类似地,激酶是多种病症的有效药理靶标;特别是丝氨酸/苏氨酸激酶似乎是多种癌症形式的根源。最近,518个基因被分类为人蛋白激酶基因超家族的成员。参见Science, 2002, 298 (5600):1912-34。这些基因中的许多基因在健康和疾病中,在不同的组织中发挥不同的生理作用。这些基因可全部或部分药理学上可识别的可能性能够使药物替代物改进,以鉴定具有最小脱靶效应的化合物。

[0110] 在某些实施方案中,本公开内容涉及评估进行性疾病的方法。对与多种进行性疾病相关的细胞改变的全面分析正在广泛进行中。候选的病症包括阿尔茨海默病、肌萎缩侧索硬化症(ALS)、帕金森病、多发性硬化症和充血性心力衰竭。已检测多个基因,其中的突变似乎改变线粒体、突触功能、选择性RNA剪接的广泛调控,以及通常由泛素系统介导的蛋白翻转。参见Nature (2011) 477, 211-215。即使在鉴定的突变不存在时,这些元素中的一些似乎也发生了错误调控;表型分析可提供对基因激活的改变的更广泛的指示,同时提供与疾病进展相关的突变和/或剪接变异的存货单。

[0111] 在某些实施方案中,本公开内容涉及评估传染性疾病的方法。转录组分析能够在加速的时间进程中揭示出传染性病原体的证据。

[0112] 对免疫系统进行mRNA分析,特别是研究对传染性病原体的成熟、成功应答中不成比例地上调的免疫球蛋白,能够允许直接克隆人免疫球蛋白,其类似于可被表达为生物工程疫苗的单克隆抗体。例如,晚期狂犬病的进展速度太快以致免疫系统无法跟上,其通常有较高的发病率。这项技术能够扩大单克隆抗体的商业开发,其通常由于需要使小鼠免疫球蛋白人源化以克服其自身固有的免疫原性而受到阻碍。

[0113] 临床研究的一个经济上重要的领域关注小干扰RNA(或siRNA)的用途。这些小干扰RNA可作为治疗试剂被引入,以选择性降解或阻止特定mRNA的翻译。尽管在基因特异性靶向作用中取得了巨大进展且在临床治疗中取得了一些成功,一个令人困惑的问题是,无法评估靶标特异性。不想要的基因抑制通常是副作用。在siRNA表达之前和之后在靶组织中分析mRNA的全部阵列的能力成为推动基于siRNA的治疗的有用工具。

[0114] 在某些实施方案中,本文公开的方法允许在个体研究者付出最小代价的情况下,在单个实验中对大量cDNA进行端到端测序并定量(每次实验 10^4 、 10^5 、 10^6 、 10^8 cDNA)。

[0115] 一个特别强有力的应用在于mRNA的全长深度测序,通过在高等多细胞物种的复杂、断裂基因的表达过程中激活转录和转录后RNA加工而产生所述mRNA。这捕获了在基因组和蛋白质组之间的生物化学通道中产生的信息。这一应用支持了以下假设:所表达的蛋白或功能性RNA比被激活用于转录的基因更能指定分子表型。

[0116] 转录组学

[0117] 而在某些实施方案中,本公开内容可用于对任何的长聚合物(如染色体DNA)进行测序和从头组装,在这一部分中描述的是对转录组学的应用。该实施方案允许对来自细胞或组织转录组中混合信使群的mRNA变体进行全长测序和相对丰度的定量。

[0118] 在该实施方案中,本公开内容绕过克隆每一种全长mRNA的cDNA复制子的常规步骤,相反其允许对来自cDNA异质混合物的最多达十亿个cDNA片段进行随机、大规模平行测

序之后,根据序列信息的分子来源进行分离(“虚拟克隆”),所述cDNA异质混合物包括区别仅在于在大于任何单个序列读段的长度上分布的共有序列元件的连接方式的形式。

[0119] 本公开内容提供了试剂、步骤和方法的组合,其使得可使用多种不同的DNA测序技术平台对来自细胞或组织的mRNA(“转录组”)的情况进行测序和定量。

[0120] 在某些实施方案中,所述方法通常包括以下步骤:

[0121] 1) 将独特的标识符序列“标签”连接至混合物中的各多核苷酸;

[0122] 2) 复制经标记的多核苷酸,通常(但不一定)为串联的经标记的同聚物;

[0123] 3) 剪切(例如,通过物理方法)经标记的复制产物以使cDNA复制子在随机点断裂;

[0124] 4) 在鉴定标签内确定的位点进行酶法切割,以对每一种酶法切割产物一端上的标识符定位;

[0125] 5) 对每个经标记的片段进行测序,以捕获标识符标签和来自随机剪切点的相关序列;

[0126] 6) 根据源分子将标记的序列对分开,所述源分子被鉴定用于单分子序列组装,用于记录相同序列的多核苷酸,以及用于重新构建初始mRNA群的统计结构。

[0127] 所述“标签”通常具有两个元件:(1)独特的SMID,其包含具有足够复杂度的随机序列的序列可识别区以确保大多数多核苷酸不可能接收到相同的标识符;(2)使SMID被清楚地鉴定的非变异序列区(“SMID封套”)。具体实施方案中的标签设计的其他方面在具体方法中列出。

[0128] 在以下实施方案中,提供了使这项技术成为能够的SMID试剂的设计、合成、应用和使用方法。通过末端连接可直接将标识符标签纳入mRNA;在某些实施方案中,所述标记试剂具有多聚-T尾;用来从mRNA的多聚-A尾引发逆转录,所述标记试剂将标签纳入所得的cDNA的5'端。当添加过剩时,该试剂将捕获悬液中的mRNA;当以低于化学计量算的量添加时,该试剂捕获无偏随机样品。

[0129] 可通过PCR复制标记的多核苷酸,例如,有限效率的最小方案。优选的方法是,通过酶法消化从cDNA/mRNA异源双链体除去mRNA,然后使用能够作用于单链DNA的RNA连接酶的合适形式使经标记的单链cDNA环化。使用核酸外切酶I除去任何残留的线性形式。使用T4RNA连接酶(参见Epicentre;Promega Corp.)的变体可将单链cDNA连接成共价环。已将源于嗜热细菌的这种酶的一些形式进行进一步修饰以用于在升高的温度下的最佳使用(例如,CircLigase:Epicentre)。这种酶有效地环化单链RNA或DNA,具有低水平的线性或环化的分子间连接产物,并且其似乎不依赖于寡核苷酸序列。尽管所述酶无法连接双链DNA,但是发现其在60℃孵育温度下能够将游离的单链3'端连接至折叠成双链构象的单链分子的5'端。

[0130] 将环化的标记的cDNA分子等分成由待进行的测序规模确定的量;理想地,在预先确定的深度(例如,5x、10x、25x等)下样品中的每个单个分子都将被序列读段覆盖。使用高度进行性、链置换phi 29 DNA聚合酶,对等分的材料(环化的标记的cDNA分子)进行滚环扩增(RCA)。使用随机序列引物可非选择性地引发RCA;或使用针对标签非变异区的多核苷酸;或使用针对各个基因的特异性序列的多核苷酸;或多基因家族的旁系同源成员;或针对直系同源基因或多基因家族的共有序列。每个RCA产物组成了单一cDNA的延伸的同多联体,每个拷贝通过包含SMID标签的标记物拷贝与下一个拷贝相连。根据引物的选择,RCA可产生连

续的单链或分支的(或“超分支的”)双链产物。使用S1或绿豆核酸酶可使分支产物去分支化。

[0131] 同聚物的片段化通常在两步中进行。第一步,通过酶促片段化、超声作用、水剪切或等同的物理方法,将去分支的同聚物片段化为序列所需的平均mRNA长度的大小(通常为2-6kbp),或取决于靶mRNA群的长度的其他大小。第二步,使用具有不常见的识别序列的限制性酶切割随机的物理片段化的产物,为此一个或多个共有位点已被设计到SMID试剂中。这些步骤产生片段,每个片段在一端具有由随机片段化而暴露的内部序列,在另一端具有SMID。然后可使用若干大规模平行测序平台的任何一个来对每个SMID标记的片段的双端进行测序。对Illumina High Seq的配对(mate-pair)方案的修改在本文被描述为对标记的cDNA片段进行有效测序的方法实用性的一个经证实的实例。

[0132] 在某些实施方案中,本公开内容涉及源分子鉴定标签的设计、合成和应用,以及其合成和应用的试剂盒,以及本文公开的使用方法。用于mRNA测序和定量的标记试剂是通过本文所述的I型和II型DNA多核苷酸接合物/引物来举例说明。这些标记试剂通常是具有暴露的单链3'多聚-dT延伸的单链DNA分子,所述3'多聚-dT延伸能够与mRNA的多聚-A片段退火并引发标记的单链cDNA的逆转录病毒的逆转录酶合成。

[0133] I型标记试剂具有SMID和其他功能序列,并且表现出缺少二级结构。通过商业合成可从设计序列获得这些I型标记试剂而不需要进一步修饰。II型标签包含SMID的重复拷贝。所述标记试剂可包含SMID的两个拷贝,或可被设计以使得其在被拷贝成双链DNA时单个SMID被复制。

[0134] II型试剂及其商业合成的前体具有二级结构,所述二级结构为:(A)用在从商业制备的前体进行的标记试剂合成中;以及(B)在合成用于大规模平行测序的富集的SMID末端标记的cDNA片段文库中发挥有用的功能。以下所述的是II型标记试剂设计,其区别在于二级结构和应用。它们还包含更多的功能结构域,所述功能结构域在试剂合成或标记的cDNA测序文库的最终制备中提供多功能性。以下所述的其他试剂包括用于文库制备的特定应用的II型合成的前体多核苷酸和引物。

[0135] II型标记试剂合成利用被设计到化学合成前体中的二级结构以直接制成复杂的标记试剂结构。或者,使用二级结构以制成合成最终标记试剂所依赖的复杂模板,并且允许作为即用型试剂而有效纯化。这种策略允许使用单一的合成酶,高效的合成和纯化这些纯化形式的复杂且可变的试剂。

[0136] 在某些实施方案中,本公开内容涉及包括从附着于固体底物的模板合成所需要的标记试剂变化的替代方案的方法。

[0137] 在某些实施方案中,分析测序标记文库所产生的数据通常包括以下步骤:

[0138] 1. SMID检测——通过侧翼序列元件(“封套”)或通过在每条文库链一端的均匀配置,或通过这两种方法来定位鉴定随机化序列。

[0139] 2. 读段分类(虚拟克隆)——根据SMID将测序读段分成单个“区段”。区段是存储相关序列数据的计算机存储器中的地址块。将每个包含SMID的读段及其配对读段指定到该SMID区段。每个区段代表初始样品中的单个源分子(即,单一的完整mRNA分子)并且该区段中的每条序列可追踪到该单一分子。这是在测序前从混合物克隆物理cDNA的测序后信息等价物(因此是“虚拟克隆”)。

[0140] 3. 在使用SMID标签鉴定由其中标签序列被从经记录的读段中去除的相关读段所代表的链之后进行序列修剪,仅留下来自源分子的信息。

[0141] 4. 组装源分子序列——以最大重叠比对的方式排列每个区段中的经修剪的读段以产生最少数量的重叠群,每个重叠群具有最长的长度。在足够的覆盖度下,每个区段产生包含源分子的端到端序列的单一重叠群。因为每个区段包含来自一条相对短的cDNA源链的序列读段,避免了妨碍大规模(例如,基因组)组装的复杂问题。转录组分析的主要问题——将序列读段指定到单个转录物——得到解决。使用已知的从头组装软件可完成整个组装过程。

[0142] 5. 将每个组装的序列指示到其来源基因(或是反式剪接中可能的基因)——可使用已知软件来更新每个基因的外显子/内含子组织的综合数据(curation)。

[0143] 6. 鉴定来源链——来自标签取向的信息或对来源基因的参照表明,相关的cDNA序列是否对应于“正义”(蛋白编码)mRNA序列或其反义互补物。这能够区分mRNA序列与多聚-A标记的非编码反义序列,所述非编码反义序列在基因表达中可发挥调控作用。

[0144] 7. mRNA特征的总体重新构建——对所有表达的内源和外源(在病原体感染的情况下)基因的稳态表达的相对水平,以及每个基因的每个序列变体的相对表达水平进行定量。这些数据提供了关于序列变异的联系的相关信息;例如,特定剪接或RNA编辑变体与特定的选择性启动子序列的相关性;特定的编码结构域的协同连接,其可反映控制蛋白质功能机制的相互作用蛋白结构域,等等。重新构建可包括可一起被测序(多重测序)的来自多个组织的信使特征的比较结构,如:在来自癌症患者多个肿瘤和未受影响的非恶性对照组织的样品中;在发育和分化的不同阶段取样的组织中;在疾病进展过程中取样的组织中。

[0145] 试剂

[0146] 在本公开内容的某些实施方案中,涉及以下术语。

[0147] “标签”是指连接至目的多核苷酸(例如,靶序列)的多核苷酸,其使所得的缀合物可被复制并通过鉴定其连接物或复制物的一部分(或全部)来区分。标签通常具有多个元件,其有助于制备和分析测序文库。

[0148] “标记试剂”是指用来将独特的标签引入多核苷酸样品的多核苷酸试剂。在某些上下文中,所述“标记试剂”是指一组多核苷酸,所述多核苷酸的一些部分包含基本上重叠的序列片段和基本上不重叠的序列片段,即,其中不重叠序列片段中的重叠序列群在统计学上是非常少的。一旦标签缀合至多核苷酸,它通常被复制——因此,标签生成具有相同序列的多个拷贝。

[0149] “标记物”是指与内部cDNA序列一起被复制的标记试剂部分。标记物根据初始悬液中的单个分子来鉴定每个经测序的片段,所述片段来源于所述悬液。有时,整个标记物原则上可用于最终的数据组装软件。标记物可包含两部分。一部分是“SMID”或“源分子标识符”,其是指随机序列的序列可识别区,例如,分布在不变的“校对”碱基之间的一系列随机碱基:所述随机碱基在数量上足以产生标记试剂的高度多样性,例如,通常但不限于,多于 10^9 条独特的SMID序列。第二,“SMID封套”包含不变碱基(重叠序列),其位于SMID的一侧或双侧。校对碱基和封套都有助于在大量文库序列中检测标签。如图1所示,标记物块包含具有其他相邻序列的标记物序列(例如,SMID和侧翼封套序列),所述其他相邻序列可包括,例如,位于标记物一侧的PCR引物序列或其互补物和位于标记物另一侧的一种或多种限制性核酸内

切酶的识别序列。因此,SMID、标记物和标记物块包含标签的嵌套序列元件。

[0150] 对于本文描述更多的II型标签,“插入的环”或“环结构域”可包含一个或多个引物或其互补物的结合位点。一个或多个限制性核酸内切酶识别和切割序列也可存在于插入的环中。

[0151] “夹板(clamp)”是指位于标记试剂前体的插入的环侧翼的退火序列(参见图2A,“退火”),其允许延伸所述前体3'端的自身引发,从而产生标记物块的分子内互补拷贝(包含SMID)。3'尾结构域通常是用来从多腺苷酸化mRNA引发cDNA合成的多聚-dT或变体(例如,dT22dV)。任选的5'尾结构域通常是多聚-dT或其他序列的延伸,其通常不与标记试剂的任何其他部分杂交。所述5'尾为通过RNA连接酶进行完整cDNA的有效环化提供了单链5'末端,而不需要行部分变性的高温。试剂具有的一对5'和3'尾使其可用于从环化的RNA复制和环化cDNA,如在用于测序Gppp加帽的mRNA的方法的实施方案中所示。(试剂的特征可在于,具有一个(仅3')或两个尾(5'和3'),其用标签名称(例如,II-ps₁型或II-ps₂型)中作为最后元素的下标表示。)

[0152] “批代码(batch-code)”是指保留在标记物中的任何变异,其可用于根据核酸(例如,mRNA、源分子)区分文库片段。多个批代码可使得能够同时分析若干来源的mRNA群,如在组织的发育系列中,在疾病进展的不同阶段的组织中,或在来自不同物种的组织的基因产物之间的比较中,等等。

[0153] 标签类型

[0154] 包括不同类型的标签:I型(单标记物)、II-ps型(两个回文的、对称的标记物)、II-pa型(两个回文的、不对称的标记物)和II-t型(两个串联的标记物,非回文的)。

[0155] I型标签通常具有一个标记物,位于标记物的5'侧翼的一个限制性酶的不常见的共有结合和切割位点的一个或多个拷贝,和位于标记物的3'侧翼的第二个酶的第二不常见的共有结合和切割位点的一个或多个拷贝。尾可包含3'多聚-dT以引发cDNA合成(从信使RNA的多聚A延伸或多腺苷酸化的非编码RNA),或可促进环形连接的5'序列。

[0156] I型标记试剂的实例在图1E中、示意性地显示为举例说明了这类标记试剂的特性的DNA序列。所述试剂是单链DNA多核苷酸,其通常(但不排除例外的可能性)缺少二级结构。在本实施方案中,所述试剂具有包含寡聚-dT片段的3'尾,所述寡聚-dT片段有但不限于22个碱基,其能够与成熟mRNA的多聚-A尾退火并通过RNA依赖的逆转录酶引发cDNA合成。每个标记分子的SMID代表总共但不限于23个碱基,包括以被校对碱基分隔开的6个三联体形式分布的18个随机碱基。在图1E的实例中,所述标记物块包含复制的不常见的限制性酶共有序列,一种类型的序列对位于标记物的3'侧翼,第二种类型的序列对位于标记物的5'侧翼。因此,SMID封套包含5'和3'各自的限制性切割位点之间的间隔,以及SMID的边界。

[0157] II型标签是指最少具有三个连续结构域的单链多核苷酸,所述三个连续结构域依次为标记物块、插入的环、标记物块;在任一端可包含任选的尾结构域,因此II型标签结构域的结构通常为:

[0158] 5'-[尾]-[标记物块]-[插入的环]-[标记物块]-[尾]-3'

[0159] 其中尾表明所述结构域是任选的。具有一个或两个尾的标签通过下标来表示,例如,II-pa₁型和II-pa₂型分别是单尾或双尾形式的实例。

[0160] 单尾II型标签有效地从多聚-A mRNA引发cDNA合成,并且在提高的温度下使用

RNA/DNA单链连接酶的热稳定形式使其环化。双尾标签更适用于通过一般的RNA/DNA单链连接酶环化,不需要升高温度以除去二级结构;它们可用来区分成熟的(加帽的)多聚-A mRNA与不成熟的5'磷酸化或5'羟基形式。在此处所述的实施方案中,3'连接元件是cDNA合成的单链寡聚-dT引物,而5'单链尾可以是寡聚-dT或根据其应用的替代序列。

[0161] 一个例外是,II型标签缺少可形成稳定双链体的自身互补区域,所述稳定双链体用于合成中的酶反应或随后的标签反应用途。唯一的例外是标记物块结构域,其包含标记物序列和其他功能序列。

[0162] 在II-p型标签中,所述第二标记物块在大范围的结构域上是第一标记物块的碱基互补物,以使所述标签多核苷酸将折叠成由标记物块结构域的互补碱基之间的碱基配对形成的双链体。在此标记物块双链体构象中,其他结构域可保持大部分而不具有二级结构。

[0163] 在II-ps型标签中,所述两个标记物块结构域在全长上互补。II-ps型标记试剂的实例示意性地示于图1D中。在该实例中,寡聚-dT的延伸的3'单链片段形成能够与mRNA的多聚-A尾退火以引发cDNA合成的3'尾。II-ps具有精确互补的两个标记物块序列,从而在溶液中形成连续的双链体:当被复制成双链cDNA时,所述两个标记物块将生成两个相同的SMID拷贝,每个拷贝在两条链中以相同的5'到3'方向排列。

[0164] 在II-pa型标签中,一个或两个标记物块可包含与另一个标记物块不互补的内部序列片段。当II-pa型标签的两个标记物块结构域杂交时,单链的一段长度将从双链体的一个或两个臂伸出,在重新结合成双链体之前形成不配对的环。标记物块内的不匹配区域的目的是,使两个标记物在随后的序列分析过程中可区分。这使单个读段中的标记物序列能够鉴定初始源分子链的义。对于标记物的两个相同的拷贝以相反的义附加于源链上的II-ps型标签而言,这是不可能的。

[0165] II-pa型标记试剂的实例在如图1E中,示意性地显示为DNA序列。寡聚-dT的延伸的3'单链片段(tract)形成能够与mRNA的多聚-A尾退火以引发cDNA合成的3'尾。它具有基本上互补的两个标记物块序列,从而在溶液中形成双链体:当被复制成双链cDNA时,所述两个标记物块将生成两个相同的SMID拷贝,每个拷贝在两条链中以相同的5'到3'方向排列。标记物块在标记物的3'侧翼包含不常见的限制性酶的共有序列。因此,当被复制成双链cDNA时,这些限制性酶的作用为切割每条链的3'至标记物序列的每个拷贝,从cDNA构建体删除“插入的环”。标记物块在标记物的5'侧具有引发位点,以在文库制备过程中允许进行SMID的PCR扩增。参见图1E中的“Seq”。所述标签在标记物块中还具有两个不配对的DNA片段,其位于SMID和限制性酶共有位点之间,所述限制性酶共有位点位于终止于插入的环内的标记物块的末端。参见图1E中的“标记物鉴别元件(marker discriminator)”。当被复制成双链cDNA并通过所述步骤被加工成最终的测序文库时,这些不同的序列将使与每种特定的SMID读段相关的初始源分子链能够被区分。

[0166] 在II-t型标签中,所述第二标记物块是第一标记物块的串联、精确的复制物,二者具有相同的义(所述两个拷贝被插入的环序列分隔开)。因此,标记物的两个拷贝以相同的义附加于源链上,因此可从测序仪输出的标记物序列推断源分子的义。

[0167] II-t型标记试剂是单链DNA多核苷酸,其通常没有双链体结构,具有被插入的环分隔开的两个相同的标记物。这种试剂提供与II-pa型标记试剂相同的益处,除了所述标记物总是表示在DNA合成中被复制的链的义。II-t型允许通过实验来测定所测序的链,而不需要

计算机操作。更重要的是,II-t型允许立即和直接克隆任何单一的cDNA。

[0168] 单尾或双尾标记试剂的合成方法:

[0169] 可通过连续固相合成(一次合成一个核苷酸)或通过连接单独生成的片段而获得I型标记试剂。可通过连接核苷酸混合物而生成随机碱基位点。

[0170] 可从商业多核苷酸前体合成II-p型(ps和pa)标记试剂,使用了生物合成酶phi 29 DNA聚合酶的特性以利用被设计到前体和反应中间产物中的二级结构来制成最终试剂。

[0171] 对于单尾II-ps型标签(II-ps₁型)的合成而言,可商业合成前体分子,所述前体分子以5'到3'的顺序含有如下四个结构域:

[0172] (1) 最终分子中所需的3'单链尾的互补物(例如但不限于:5'-WA₂₂,其中W是V的互补碱基)。

[0173] (2) 标记物块(5'-A-[B-SMID-C]-D 3'),其中"[B-SMID-C]"是标记物本身(在最终测序文库中被复制和保留的元件),A和D是标记物块的相邻的5'和3'部分。

[0174] (3) 插入的环;尽管这个环没有前体内的互补位点,其可包含与多核苷酸("环引物",LP)互补的序列,所述多核苷酸可用于在反应系列中引发第二链中间产物的合成。

[0175] (4) 标记物块3'至SMID(例如,C-D的部分或全部)部分的互补物:这可称为用于自身引发的分子内"夹板(clamp)"。

[0176] 所述前体在溶液中将折叠形成结构域(4)及其在结构域(2)中的互补物之间的双链体——即,5'-D'-C,标记物块中亚结构域C-D的反向互补物:这些元件一起构成了"夹板(clamp)"。参见图2A中的"寡聚-J前体"。

[0177] 这形成了茎-环构象,所述茎包含双链体,环是插入的环结构域(3)。在所述环的相对一侧,从所述茎的5'端延伸,是DNA单链,其包含以上所述的结构域(2)的未配对片段和尾结构域(1)。

[0178] 所述前体多核苷酸被称为"寡聚-J",因为折叠结构的一些形式类似于字母"J",单链环对应于该字母的碱基。参见图2A。

[0179] 合成"环引物"(LP)多核苷酸,其结合于寡聚-J的插入的环结构域(3)。LP结合序列可在标记物块结构域的任一侧偏移若干个碱基以允许未受阻碍的引物结合于单链环并有效引发聚合酶反应。一般而言,但不排除例外,LP的5'端不会被磷酸化以防止其参与(如果其作为痕量污染物被保留在最终的TRS制品中)cDNA合成和随后的环化反应。

[0180] 合成与寡聚-J的5'尾结构域相同的RNA多核苷酸,其被称为"RNA保护物/引物"(RPP)。在一些实施方案中,短DNA多核苷酸可被替换为RPP,被阻断以阻止DNA合成,与用于第二链合成的单个引物一起加入。

[0181] 通常,在合成的第一步中,双链体的3'端作为寡聚-J上的自身引物,使用DNA聚合酶使该自身引物延伸以复制SMID和相邻的标记物块以及尾结构域。参见图2A。这产生了平端的茎-环结构,此时所述茎包含的片段包括一条链上的尾和整个标记物块以及另一条链上的其精确互补物。仅有插入的环结构域是单链的。该结构被称为"发夹"(HP)。

[0182] 通常,在合成的第二步中,按照以下步骤在HP的3'端生成RPP的游离结合位点。"环引物"(LP)结合于HP单链环上的其互补序列。参见图2A。使用phi 29聚合酶延伸该引物,所述聚合酶复制模板最多至其5'端,在此过程中从所述茎置换链的3'端。

[0183] 产物链仍然与HP模板的5'半边形成双链体,但是经置换的模板3'半边(包含完整

的标记物块和尾结构域)此时是单链。该两条链的结构被称为“3Q”。单个短产物链被称为“1Q”。

[0184] 3Q双链体的结构:

[0185] 5'op-[D'-C'-SMID'-B'-A']-[尾]3'

[0186] 3'[尾]-[A'-[B'-SMID'-C']-D']-环-[D-C-SMID-B-A]-[尾]5'

[0187] 通常,在合成的第三步中,通过使RPP结合于3Q的游离3'端并使用phi 29聚合酶延伸来生成HP的完整拷贝,这可容易地由RNA多核苷酸引发。该合成复制模板链并从复合物置换1Q链。参见图2A。

[0188] 第二和第三步通常同时进行。当所述模板的3'端在环引发的反应中被置换时,它可被聚合酶的3'核酸外切酶活性降解。因此,同时加入引物LP和RPP以使得RPP在暴露时可与模板的3'端杂交,保护其免于降解。

[0189] 在合成的第四步中,然后使用RNA酶H除去产物的RNA引物部分。双链产物包含与HP模板链形成双链体的所需的TRS。

[0190] 可使用多种方法除去引物和蛋白质,例如,“辅助”链1Q——其此时代表TRS的截短拷贝的污染部分,并且分离出TRS和HP链。单链引物通常被Phi 29的3'→5'核酸外切酶活性降解。通过琼脂糖凝胶电泳可从1Q分离出HP-TRS复合物,同时除去反应2和3的残余的多核苷酸引物和任何蛋白质。通过常规方法(电洗脱、在离液盐中熔化琼脂、提取和膜纯化等)从凝胶切片回收纯化的HP-TRS。如果在从凝胶提取的过程中或之后使HP-TRS双链体瞬时变性,这两条链随后将不会重新退火。每个分子的占绝对优势的动力学途径是由于分子内互补性而自身发生折叠。所得的等摩尔HP和TRS溶液可直接用于标记mRNA。在与mRNA退火并引发cDNA合成方面,以及在所述方法的随后步骤(例如,单链DNA的环化)中,HP是惰性的。

[0191] 在第二个实施方案中,所述寡聚-J前体衍生出一个使其能够以不干扰对多核苷酸的酶活性的方式连接至固体载体的官能团。在该实施方案的一个实施例中,可用连接至生物素的化学延伸部分来修饰所述寡聚-J前体。在TRS-HP双链体合成之后,它将结合至已连接有链霉亲和素的琼脂糖或玻璃小珠或其他固体基质,但是污染的1Q或残余的多核苷酸引物和酶不发生结合。通过冲洗除去1Q、多核苷酸引物和任何蛋白质。然后处理所述小珠以使TRS-HP双链体瞬时变性,导致重新生成粘附于基质的HP。将TRS释放到溶液中,通过冲洗而洗脱,并使其处于适合备用于标记反应的形式。

[0192] 图2B中说明了双尾II-ps型标签(II-ps₂型)的合成。使用如之前实施方案中的环引物来启动1Q合成,在所述实施方案中,HP前体衍生出用于最终产物的物理分离的生物素化的延伸部分(或适合与固体基质相互作用的其他延伸部分)。在这种情况下,在与第二个反应同时进行的第三个反应中,可合成DNA多核苷酸(TRS引物)而不是RPP,所述DNA多核苷酸的3'端序列等同于寡聚-J中标记物块的一部分(不包括SMID,即结构域“A-B”)。该引物的5'尾不是HP的3'尾序列的互补物:(这个非互补部分可根据需要变化,并且可以是多聚-dT的5'延伸部分)。该引物以引物的5'尾和HP的3'尾都保持单链的方式结合于HP的新暴露的3'区域。该引物从其双链3'端的延伸产生具有非互补的5'和3'尾的TRS,并且置换1Q。在瞬时变性引起释放之后,这条链将发生自身折叠以形成在5'和3'端都有单链尾的茎-环结构。HP和TRS在温和变性条件下分离,所述变性条件阻止TRS的3'端和HP的游离5'端退火。

[0193] 以与II-ps₁型标签相同的方式合成单尾II-pa型(II-pa₁型)标签,但有以下例外:

第一标记物块的SMID封套的3'部分(在上图中的结构域C)与其在第二标记物块中的对应部分(C')不完全互补,形成了寡聚-J前体的夹板(clamp)。唯一的限制是,这两个封套部分必须在SMID附近互补,并且在该区域中的互补程度必须足以形成双链体,所述双链体足够稳定以允许HP合成的自身引发。

[0194] 双尾II-pa型(II-pa₂型)标签的合成与双尾II-ps₂型形式的完全相同,其开始于以上实施方案中所述的适合的寡聚-J前体。

[0195] 在II-pa型和II-ps型的固相合成中,标记试剂通过化学键共价连接至固体基质如玻璃上。

[0196] 在单尾形式的合成中,所述寡聚-J前体多核苷酸包含一个使其能够以不干扰对所述多核苷酸的酶活性的方式共价连接至固体载体(如玻璃)的官能团。以这种方式,反应中间产物可被冲洗掉并且以最少的操作或损耗引入新反应物。简单地通过加热以熔化HP-TRS双链体来回收最终产物(TRS)。一旦冷却,TRS和HP链发生自身折叠以形成发夹。在溶液中回收TRS,而HP仍结合于载体上。

[0197] 在双尾形式的合成中,在第一个循环中,如在之前的实施方案中那样用环引物来启动1Q合成。但是在与第二个反应同时进行的第三个反应中,可合成DNA多核苷酸(TRS引物)而不是RPP,所述DNA多核苷酸的3'端序列等同于寡聚-J中标记物块的一部分(不包括SMID,即结构域“A-B”)。该引物的5'尾不是HP的3'尾序列的互补物:(这个非互补部分可根据需要改变,并且可以是多聚-dT的5'延伸部分(stretch))。该引物仅结合于HP的新暴露的3'区域上,因此,所述引物的5'尾和HP的3'尾都保持单链。该引物从其双链3'端的延伸产生具有非互补的5'和3'尾的TRS,并且置换1Q。在瞬时变性引起释放之后,这条链将发生自身折叠以形成在5'和3'端都有单链尾的茎-环结构。

[0198] 在双尾标记试剂的固相合成中,如果广泛暴露于phi 29聚合酶的3'核酸外切酶活性,HP的游离3'端将被降解直至与TRS DNA引物形成双链体的点,然后phi 29聚合酶复制TRS引物的5'端以形成修饰的共价连接的HP。在变性条件下从HP-TRS复合物中除去TRS,然后经修饰的HP将采取具有非互补的5'和3'尾的叉状末端构象。于是可用仅针对HP的游离3'端的引物来启动TRS的进一步重复循环合成,而不需要提前合成1Q。引发反应,冲洗固相以除去试剂并在变性条件下回收TRS。在固相存在的情况下,通常无法回收TRS的二级结构,因为TRS可与修饰的HP的靠后游离3'端重新退火。

[0199] 因此,在固相上制备具有叉状尾且经修饰的HP可用于TRS多循环形成而不消耗HP(和消耗额外的寡聚-J前体的成本)。这样的固相合成模板可为标记试剂合成试剂盒提供基础;图2B中所示的重复步骤可使固相试剂被修饰以产生具有由研究者决定的不同5'尾的TRS。

[0200] II-t型试剂的合成如图2C中所说明。可商业合成前体分子,其以5'到3'顺序含有以下11个结构域(小括号中的数字对应于图2C中的元件):

[0201] (1) 包含最终分子中所需的3'尾的互补物(例如但不限于:5'-WA₂₂);

[0202] (2) (任选地) 包含PCR引物互补序列;

[0203] (3) 包含与阻断DNA合成的多核苷酸互补的序列;

[0204] (4) 包含与自身引发夹板(clamp)序列互补的序列(SP-2互补物);

[0205] (5) 包含SMID互补序列;

[0206] (6) 包含与自身引发夹板 (clamp) 序列互补的序列 (SP-1 互补物)；

[0207] (7) 包含与多核苷酸 (“LP-1”) 互补的序列。LP-1 结合位点从结构域 (5) 的 3' 端偏移若干个碱基以使环引物 LP-1 正确结合以及通过 phi 29 DNA 聚合酶使结合的引物有效延伸。结构域 (6) 还包含在文库制备中待使用的限制性核酸内切酶的不常见的识别序列 (RE-L (图中的冲程))。

[0208] (8) 包含夹板 (clamp) 序列 SP-2 互补物 (结构域 3 的重复)。

[0209] (9) 包含与多核苷酸 (“LP-2”) 互补的序列。LP-2 结合位点从结构域 (7) 的 3' 端偏移若干个碱基以使环引物 LP-2 正确结合以及通过 phi 29 DNA 聚合酶使结合的引物有效延伸。结构域 (8) 还包含在标签合成中待使用的限制性核酸内切酶的识别序列 (RE-T (图中的双冲程))。如果需要, LP-2 序列可从其 3' 端延伸, 超过结构域 (8) 并进入下一个结构域。序列 RE-T 必须延伸到 LP-2 的 3' 端, 但是应该从那个位置偏移许多碱基, 所偏移的碱基足以使限制性核酸内切酶切割末端有 LP-2 的双链底物的两条链。

[0210] (10) 包含 PCR 引物互补序列 (如果存在的话, 结构域 2 的重复)。

[0211] (11) 包含与结构域 (5) 互补的自身引发的夹板 (clamp) 序列 1 (SP-1)。

[0212] 合成两个“环引物”多核苷酸 LP-1 和 LP-2, 其具有与前体多核苷酸结构域 (6) 和 (8) 中的其各自的结合位点互补的序列。

[0213] 合成与结构域 (3) 互补的阻断 DNA 单链多核苷酸。前体多核苷酸形成茎-环构象, 所述茎包括结构域 (5) 和 (10) 之间的双链体 (SP-1 及其互补物)。所述环包含结构域 (6) 至 (9)。从所述茎的 5' 端延伸、位于所述环相对一侧的是包括结构域 (1) - (4) 的 DNA 单链。参见图 2C (A)、(B)。

[0214] 前体退火至阻断 DNA 多核苷酸, 然后使用 DNA 聚合酶 (Klenow) 延伸茎-环构象中的前体多核苷酸的 3' 端, 从 SP-1 引发, 复制 5' 单链以形成部分发夹 (HP-1, 图 2C (C))。发夹茎的臂包含与模板的结构域 (4) 互补的新合成的夹板 (clamp) 序列 SP-2, 但是没有延伸过阻断多核苷酸。

[0215] 引入环引物 LP-1 并使其与发夹的环杂交。使用 phi 29 聚合酶延伸复制发夹的 5' 臂, 置换 3' 臂作为单链片段 (图 2C (E)) 并置换阻断多核苷酸至溶液中。新合成的产物链仍然与 HP-1 模板的 5' 半边形成双链体, 但是被置换的模板 3' 部分此时是单链。这个双链结构被称为“3Q”。单个短产物链被称为“1Q” (参见图 2C (H))。

[0216] 在 HP-1 合成过程中生成的新的 SP-2 序列此时是未配对的, 并且与其在环中的互补物——结构域 (7) 杂交 (图 2C (F))。3' 端剩下的未杂交的核苷酸被 phi 29 聚合酶的 3'→5' 核酸外切酶活性连续地去除, 回到 SP-2 双链体 (图 2C (D))。杂交的 SP-2 此时充当引物, phi 29 聚合酶延伸 3' 端, 将 1Q 置换至溶液中 (图 2C (H))。这生成了更长的发夹 (HP-2), 其单链环包含来自初始的前体多核苷酸的结构域 (8) - (11) (图 2C (H))。

[0217] 引入第二环引物 LP-2 并使其与环中的互补序列 (结构域 (8)) 杂交。该引物的延伸复制 HP-2 的 5' 臂, 置换 3' 臂作为单链 (Figure 2C (F))。通过限制性核酸内切酶的消化在 RE-1 处切割该产物将双链部分与单链部分分离, 所述单链部分是所需的 TRS, 即 II-t 型试剂。

[0218] 如果所述前体多核苷酸被生物素化, 可通过与链霉亲和素小珠孵育并单独回收上清液中的 TRS, 从双链切割产物中分离 TRS。

[0219] 用于高度平行测序平台的试剂

[0220] 在Illumina High Seq设备的配对或双端测序模式中,对待纳入各文库的DNA片段进行平端处理(end-polished),加A-尾,并连接至具有若干功能元件的叉状衔接体;PCR位点,捕获序列,簇合成序列,共有切割位点和测序引物。

[0221] 如所说明的,在衔接体连接之后,PCR扩增导致独特的末端对连接至双链DNA的相反末端(例如,多核苷酸的A和B互补对)。尽管形成扩增文库的各衔接体修饰的DNA片段的相反末端具有A末端和B末端,链同样可能在每条链的5'或3'端具有B-多核苷酸之一,或在每条链的5'或3'端具有A-多核苷酸之一。

[0222] A和B指导的PCR引物(分别是PCR 1.0和PCR 2.0)扩增内部捕获的DNA片段。为了选择性产生在一端独有标记物并且在另一端独有cDNA内部序列的文库片段,可使用修饰的衔接体,修饰的PCR引物,或同时使用。

[0223] 在一个实施方案中,修饰的衔接体包括对应于标准叉状引物的B-叉的单链。一旦在第一轮PCR中被复制,这提供了能够与PCR 1.0相互作用的小片段;在所使用的退火温度下,从A末端引发DNA合成降低或受阻。

[0224] 在第二个实施方案中,修饰的衔接体包括对应于标准叉状引物的B-叉的单链和截短的A-叉的小互补部分。这提供了在退火温度下能够与PCR 1.0相互作用的小片段;在所使用的退火温度下,从A末端引发DNA合成降低或受阻。

[0225] 在一个实施方案中,合成A-叉特异性PCR 1.0的修饰形式,其具有(但不限于)在3'端突出的5个碱基。连接所述碱基的磷酸基键被修饰为硫代磷酸酯(phosphothioate)、磷酸甲酯或磷酸酰胺键;这降低了聚合酶的核酸外切酶活性消除突出端的能力。因此,从A-末端进行DNA合成,不管是从未修饰的叉状引物、具有截短的A-链的经修饰的衔接体、还是仅包括B链的修饰引物,都极大地降低。将经修饰的衔接体与经修饰的PCR 1.0结合极大地降低了从A-末端的DNA合成。

[0226] 在某些实施方案中,如此设计标记试剂使其包含经修饰的PCR 1.0引物的退火位点。由于这几个步骤,使用经修饰的A-引物(PCR 1.0,修饰的)进行PCR扩增仅能从标记物序列开始,通过SMID并到达相关的随机断裂位点。从B末端扩增是不变的。这些限制的结果是,每个阶段I序列开始于标记物序列,进行到cDNA中的随机内部断裂位点。每个配对的阶段II序列开始于cDNA内的第二随机断裂位点。在这些条件下,几乎全部测序读段对具有标记物,因此可用于构建源cDNA分子的序列。

[0227] 方法

[0228] 在某些实施方案中,本公开内容涉及在异质悬液中的标记多核苷酸以使其可区分性最大化。在多核苷酸的异质溶液中,单个分子仅在其序列不同的情况下可被区分。为了根据大规模平行短序列读段重新构建定量的群特征,首先修饰每个分子以最终根据其完整序列将其与其他分子区分开。

[0229] 在某些实施方案中,所述方法允许以这样的方式加工标记的寡核苷酸以进行扩增,然后以这样的方式加工每个片段拷贝以使得初始的标签与所产生的内部片段一起被复制。这允许对重新构建初始的异质溶液中所有寡核苷酸的序列和相对数量所需的相关信息,进行计算机恢复。

[0230] 在某些实施方案中,本文公开的方法能够返回细胞或组织中基本上全部信使RNA的序列或其统计学上有代表性的量,以及对其相对表达水平的评估。这些信息包括“转录

组”的子集。信使特征 T_m 涉及以下方程：

$$[0231] \quad r = \sum_{g=1}^N p_g \left(\sum_{i=1}^{k_g} q_{gi} \cdot S_i \right) = \sum_{g=1}^N p_g \cdot r_g$$

[0232] 信使RNA的权重分布，其中 S_i 代表特异性的信使序列； q_{gi} 代表基因 g 的 i^{th} 信息的相对丰度； p_g 代表每个 N 表达的基因的转录物的相对水平。这一信息提供了分析转录组的统计学结构的基础，以揭示出控制基因组和所表达的细胞分子表型之间的信息增量的复杂机制。

[0233] 在某些实施方案中，本公开内容涉及区分样品中基本上全部mRNA或其统计学上有代表性的量的方法。在某些实施方案中，本公开内容涉及重新构建mRNA特征的方法，其中分离多聚腺苷酸化mRNA并使其与低于化学计量的标记试剂（例如，II-pa1型试剂）悬液结合。由于3'单链的寡聚-dT延伸，标记试剂退火到mRNA的多聚-A尾。在该实施方案中，源分子的定量因此通常不依赖于所述方案的后期步骤的复制效率的任何差异。在有效生成全长cDNA的条件下，使用逆转录病毒的RNA依赖的DNA聚合酶（例如，莫洛尼鼠白血病病毒逆转录酶或其他来源的逆转录酶）启动从3'标签末端引发的通过逆转录进行的cDNA合成。

[0234] 在某些实施方案中，本公开内容涉及区分样品中所有Gppp加帽的mRNA的方法。成熟的mRNA在一端具有Gppp帽，在另一端具有多聚-A尾。不成熟形式具有5'磷酸基或5'羟基。在一个涉及仅捕获样品中加帽mRNA的特征的实施方案中，根据标准方案分离包括Gppp加帽形式的多腺苷酸化的mRNA；缺少末端帽但是具有5'末端磷酸的mRNA；缺少末端帽但是具有5'羟基的mRNA。使用碱性磷酸酶（AP）处理多聚-A mRNA以从未加帽的、5'磷酸基化的种类除去末端磷酸。使用烟草酸磷酸酶（TAP）处理样品以除去末端Gppp基团，仅在那些分子上留下5'末端磷酸。使用RNA连接酶使样品中具有5'末端磷酸的之前加帽的mRNA连接成环状RNA，而5'羟基形式仍保持线性。将混合的环化和线性mRNA与同时具有5'和3'单链多聚-dT片段的II型标记试剂（例如，II-pa2型）的悬液结合。退火到mRNA的多聚-A尾的标签引发使用逆转录病毒的逆转录酶的逆转录。经环化的RNA为使用T4DNA连接酶有效连接成环状cDNA提供了模板或“夹板（splint）”。与线性RNA分子形成双链体的线性cDNA未被有效连接，大部分仍保持线性形式。用RNA酶H处理样品以从mRNA/cDNA双链体除去RNA，并且使用RNA酶R以除去残留的非复制的线性RNA，然后用核酸内切酶I除去线性cDNA。剩余的环化cDNA反映了初始样品中成熟、加帽mRNA的特征；所述环化形式可被纳入本文所述的方法中的随后步骤。

[0235] 在某些实施方案中，本公开内容涉及区分未加帽但是具有末端5'磷酸的mRNA的方法。本实施方案与上述实施方案相同，除了Gppp帽被保持完整；使5'磷酸基形式进行环形连接，然后进行之前的实施例中的其余步骤。所述环化形式可被纳入本文所述的方法中的随后步骤。在一个替代实施方案中，使mRNA的5'羟基形式磷酸化，使RNA环化，保持Gppp帽完整。这将捕获“不成熟的形式”并可用于对比其与成熟形式的组成。

[0236] 在某些实施方案中，本文所述的方法包括使单独标记的cDNA多核苷酸环化的步骤。使用RNA酶H处理从逆转录产生的RNA/cDNA异源双链体，以除去RNA链，然后使RNA酶H热失活。根据标准方案使用RNA连接酶环化5'标记的单链cDNA，然后使所述连接酶失活，并使用核酸内切酶I除去剩余的线性cDNA链。经环化的cDNA的数量可根据所添加的以引发cDNA合成的引物的（限制性）化学计量来估计，或通过光谱学（spectroscopic）或其他方法来估

计,并被分成适于最终测序规模和所需的覆盖度深度的等分试样。

[0237] 在某些实施方案中,本文公开的方法包括将环化的cDNA复制成分支、线性、标记的同聚物的步骤。在一个实施方案中,使用高度进行性的DNA聚合酶phi 29(用随机寡聚物引发),对用于测序的经环化的cDNA的等分试样进行RCA。RCA的产物是双链DNA的长的、频繁分支的同聚物,每个同聚物包含串联的由源分子特异性标签的重复分隔开的单一cDNA重复。在一个替代的实施方案中,使用phi 29聚合酶(用退火到标签中独特序列上的反向寡聚物引发),对用于测序的cDNA的等分试样进行RCA。在一个替代的实施方案中,使用phi 29聚合酶(用与在mRNA剪接变体中可能是保守的所选择基因的序列互补的寡聚物引发),对用于测序的cDNA的等分试样进行RCA。在一个替代的实施方案中,使用phi 29聚合酶(用与多基因家族或超家族的旁系同源成员的通常保守序列互补的多核苷酸引发),对用于测序的cDNA的等分试样进行RCA。在一个替代的实施方案中,使用phi 29聚合酶(用与来自不同物种的直系同源基因的成员的序列互补的寡核苷酸引发),对用于测序的cDNA的等分试样进行RCA。

[0238] 在某些实施方案中,本文公开的方法包括使单个多核苷酸的随机内部片段与来自源分子标记试剂的区分标记物结合的步骤。随机内部片段被转化为可与鉴定标记物一起被测序的形式,所述鉴定标记物来自cDNA合成时引入的标记试剂。可通过用单链核酸酶(如S1核酸酶或绿豆核酸酶)切割单链分叉对长链多核苷酸进行去分支化。通过物理方法使去分支的长链多核苷酸片段化成研究者所指定的平均大小范围,所述物理方法包括但不限于超声作用、酶促片段化、水剪切或雾化作用。在一些实施方案中,最佳平均大小接近于平均mRNA的大小(约1.7kb),例如,约3kbp。在一些实施方案中,最佳大小范围小于平均mRNA的大小(例如,约500bp)。在一些实施方案中,最佳大小接近于待测序的最长mRNA的大小(例如,约30kbp)。在一些实施方案中,可保留经片段化的材料的等分试样用于特定mRNA(其随后在测序中鉴定为目的物)的随后克隆,用于功能性表达或其他研究。

[0239] 在某些实施方案中,本文公开的方法包括切割来自I型标记试剂的cDNA聚合物的片段的步骤。对于用I型标记试剂产生的并具有单一标记物的cDNA聚合物而言,在酶促去分支和物理片段化之后,使用针对共有序列的限制性内切酶单独地处理片段的等分试样,所述共有序列仅位于标记物的5'侧或仅位于标记物的3'侧。切割后,重新结合这些片段等分试样。

[0240] 在某些实施方案中,本文公开的方法包括切割来自II型标记试剂的cDNA聚合物的片段的步骤。在酶促去分支和物理片段化之后,使用限制性酶处理复制的cDNA多核苷酸,所述限制性酶具有之前被设计到标签中的不常见的共有序列。在端到端序列中的任何位置具有所述标签的片段将被切割以在每个切割片段的一端留下标记物并在另一端留下随机断裂位点。在该步骤中,在双功能标记试剂中分开两个标记物元件的环形序列通常(但并非在所有情况下)被切除。

[0241] 通常,这些步骤的净效应是,产生多个片段,其中许多片段一端具有来自单个cDNA序列内部的随机断裂位点,另一端具有鉴定初始的单个源分子的标记物的拷贝。通过多种大规模平行测序平台的任何一种(包括双端或配对测序方法)处理,这些片段将产生大量标记的读段或双端读段,根据组装成长全序列的标记物的SMID可区分所述读段或双端读段,反映了起始样品或样品中初始多核苷酸混合物中的那些。

[0242] 以下实施例描述了使用Illumina High Seq设备平台的双端或配对测序方案的实施方案。双端方案通常(但并非在所有情况下)限于提供来自经标记的片段末端的最大约800bp的内部序列,通常限制cDNA全长序列大小为约1.6kb,接近于mRNA的平均大小,将分析限制于约一半的mRNA群。相反,双端方案没有这样的限制,通常可应用于分析整个mRNA群。

[0243] 在某些实施方案中,本文公开的方法包括制备双端测序文库的步骤。使用本文所述的方法,可将双端测序的Illumina方案改造为适用于全长mRNA测序。在此处给出的实施例中,对如上所述制备的标记物标记的片段进行以下改造。

[0244] 对不大于约800bp的酶法切割、标记物标记的片段进行末端修复、加A-尾并将其连接至Illumina叉状衔接体。通过冲洗除去过量的衔接体之后,使用特异性针对叉状衔接体的引物通过PCR产生索引文库。如下所述对所得的文库进行测序,用于配对测序。

[0245] 在一端具有标记物序列并且另一端具有随机断裂序列的片段将会被捕获,在两端都具有标记物序列的片段和没有标记物序列的片段也会被捕获。在一些实施方案中,可使用修饰的衔接体和PCR引物以产生在一端具有标记物并且在另一端具有随机断裂序列的文库。这些修饰的步骤在以下配对测序中描述。这些方法通常产生落在cDNA的5'末端或3'末端的800bp以内的内部序列,因此无法捕获长度远大于约1.6kbp的cDNA的全长序列。

[0246] Illumina配对方案是一个修改的方法,其基本上延伸了DNA的大小,提供了所述DNA全长序列,因此是mRNA分析的常用方法。如上所述生成经酶法切割的标记物标记的片段,其可在以上所述的大小范围内。通过标准方案对所述片段进行末端修复,在每条链的5'端使其生物素化并使其环化。因此,生物素基团标记了环化反应的接合点。因此,在一端具有标记物并且在相对端具有随机断裂位点的片段中,环化产生了源分子鉴定SMID和cDNA中随机断裂位点的物理连接,并且该接合点共价连接于生物素残基。

[0247] 通过雾化作用再次对环化的、生物素化的cDNA进行片段化,产生一系列平均长度为300-500bp的片段。对这些片段进行末端修复并加A尾,并使其与标准Illumina叉状衔接体或如上所述的衔接体的定制修饰形式连接。将这些吸附到链霉亲和素小珠上,并且通过冲洗除去未接合的片段。使用Illumina PCR 1.0 (A) 或PCR 2.0 (B) 对与标准Illumina叉状衔接体连接的片段进行PCR。

[0248] PCR扩增的净效应是产生双链DNA片段,其在一端具有A-引物序列及其互补物,在另一端具有B引物序列及其互补物。将产生相同的cDNA片段的重复拷贝,其中A和B对会被连接至任一端。A和B末端包含用于使用PCR 1.0或PCR 2.0引物进行的PCR的引物,使单链DNA在测序室中退火到单链A和B特异性多核苷酸上的捕获序列;用于簇合成的引发序列;在配对测序方案中使用的A和B末端特异性试剂的切割位点;以及引物位点,籍此捕获多核苷酸的3'端在测序过程中引发DNA合成。在一个实施方案中,可使用经修饰的衔接体,其包括对应于标准叉状衔接体的B-链的T-尾DNA的单链。在一个实施方案中,可使用经修饰的衔接体,其包括对应于标准叉状衔接体的B-链的T-尾DNA的单链,退火到标准叉状衔接体A-链的短片段,但是缺少在PCR反应条件下允许A引物(PCR 1.0)退火的片段。在某些实施方案中,已添加了衔接体的雾化作用片段被吸附到链霉亲和素小珠上,通过冲洗除去不包含环化反应接合点的未被生物素化的DNA片段。使用PCR 1.0和PCR2.0引物对经生物素化、被吸附的片段进行PCR,将一端有A-引物对且另一端有B-引物对的双链DNA释放到溶液中。这组成了配对文库,所述文库获得了环化反应的接合点。

[0249] 在替代的实施方案中,从经修饰的衔接体之一获得具有末端片段的生物素化、被吸附的片段,通过PCR 2.0而非PCR 1.0使所述末端片段经受DNA合成的引发。对于这些片段而言,进行用PCR 1.0和PCR 2.0引发的PCR,其中一条链由退火到末端衔接体链的PCR 2.0引发,而相反方向的合成由结合于A-链序列的PCR 1.0引发,所述A链序列之前作为PCR引物位点被纳入来自标记试剂并保留在所述片段中的标记物块。

[0250] 在其他实施方案中,修改所述方案,以使得对于这些片段而言,可使用退火到末端衔接体链的PCR 2.0在一个方向上引发DNA合成,而相反方向上的合成用PCR 1.0(修饰的)引发,其中由于既不与末端衔接体互补也通常不与靶DNA序列末端互补的若干个碱基突出,使PCR 1.0(修饰的)不能从末端引发合成。

[0251] 在一个实施方案中,可使用经修饰的PCR 1.0引物和标准Illumina叉状引物衔接体。在一个实施方案中,可使用经修饰的PCR 1.0引物和修饰的单链B衔接体。在一个实施方案中,可使用修饰的PCR引物与包括常规B链和截短的A链的修饰的引物。

[0252] 在一些实施方案中,可了解的是,通过设计,在使用II-ps₁型、II-ps₂型、II-pa₁型、II-pa₂型而非I型或II-t型标记物形成的文库中,在5'和3'端都具有标记物的片段在最终文库中将不会被扩增,因为在雾化作用之前进行环化时,标记物块(而非插入的环)将再接合,并且由于内部互补性,PCR 1.0引物或PCR 1.0修饰的引物的退火位点将无法用于扩增反应。

[0253] 在一些实施方案中,可了解的是,在使用II-ps₁型、II-ps₂型、II-pa₁型、II-pa₂型而非I型或II-t型标记物形成的文库中,在序列的任何位置都没有标记物的片段将缺少PCR 1.0引物或PCR 1.0修饰的引物的退火位点,因此不会被扩增。

[0254] 经修改的标准方案的净结果是产生配对文库,其中每个片段优选地(>80%)在一端具有A-引物/测序对,并且在另一端具有B-引物/测序对,但是其中紧邻A-引物对的序列始终是标记物(包括源分子鉴定SMID),其与cDNA序列内的随机断裂序列直接连接。B-引物/测序对连接于通过雾化作用产生的相同cDNA序列中的第二随机断裂序列,通常对应于由平均大小的雾化作用片段(例如,300-500bp)分隔开的A-连接序列的下游区域。

[0255] 经修改的标准方案的净结果是使用High Seq设备中的单室通常会产生>100,000,000对的序列读段,基本上全部这些都会被源分子SMID可识别地标记。使用全部15个可用的室通常将产生>1,500,000,000对的序列读段,基本上所有这些都会被源分子SMID可识别地标记。因此,根据所需的覆盖度深度,单室可产生高达>1,000,000全长信使的组装序列,并且已有设备的组合室可产生>15,000,000全长信使的组装序列。

[0256] 某些方法使用和生长中的双链序列连接的荧光标记的核苷酸,其中使用化学官能团来控制聚合作用。使用相同的多核苷酸来增大固体表面的面积,并且经荧光标记的核苷酸表明添加了哪个碱基。所述方法还可拓展到其他方案,包括中等大小片段(>300bp)的全长测序。

[0257] 在Illumina High Seq设备的双端方法中,所述文库包括通常小于800bp的片段。所述文库由变性成单链的双链、矢量修饰的、平端DNA片段组成。这些被退火到捕获室(流动池)表面小区(tile)上的共价连接的单链多核苷酸(与3'端延伸B或A的互补)上。

[0258] 捕获多核苷酸引发与退火的单链DNA互补的链的合成,然后使产物变性,将(非共价)模板冲洗掉。然后被保留的链退火到与其游离3'端互补的附近捕获多核苷酸。从该捕获

多核苷酸延伸出第二条链,产生双链“桥”,其任一端仅与DNA双链体的5'端相连。

[0259] 使这些桥变性,并且将单链重新退火到新的捕获多核苷酸上,重复该过程,直到对于每个最初退火到室表面的DNA片段而言,扩增产生多核苷酸簇——其通过其A或B端连接至所述表面。这通常在每个流动池测序室中产生大量(例如,100000000-600000000)簇。

[0260] 测序通常分两个阶段进行。在阶段I,使用特异性针对两个连接元件之一的试剂来切割DNA桥的群,使非共价连接的链变性并将其冲洗掉。这留下仅在两个方向之一(例如,B共价的5'端)的单链DNA被从游离端进行测序,使用与上述叉状衔接体一起引入的A测序引物。

[0261] 通过以下方式进行测序:从A-捕获多核苷酸引发,从四种可区分的荧光三磷酸核苷酸衍生物的溶液中连续纳入适合的碱基;在光学记录每个新加入的碱基之后,荧光基团被水解,重复进行反应。以这种方式,可光学记录游离(3')端的最多达150个碱基。

[0262] 在阶段I之后,使在测序过程中产生的未束缚(untethered)的链变性并将其冲洗掉。然后,使保留的模板在其游离3'端重新退火到小区上的捕获多核苷酸(例如,A)。通过延伸所述捕获多核苷酸来合成相反方向的新链。然后在第二(例如,B)接头处切割所得的桥接多核苷酸的群,使所得的未束缚的链变性并将其冲洗掉,如所说明的。剩余的群代表游离3'("B")端,如之前所述进行测序,使用"B"多核苷酸引发以产生第一测序链相反末端的互补序列。在收集来自每个簇的阶段I和阶段I的光学记录的序列之后,一起报告每个簇的这些读段对。

[0263] 在配对测序中,使核酸片段化(例如,但不限于,通过超声作用、酶促片段化或水剪切)成片段,通常为数kb。因此,所得的序列读段捕获到最长达平均片段大小二倍的插入序列。在选择起始片段化大小范围时,将全长组装序列的最大尺寸设定为初始片段尺寸的约二倍。在其他方面,双端方案和配对方案中的设备测序步骤是相同的,差异仅在于文库制备。

[0264] 通过计算机组装序列(参见图11)。概括地说,根据独特的SMID标识符来分离读段对,所述SMID标识符指定衍生出经测序的cDNA的单个源分子。避免物理处理每个cDNA是允许大量产生序列数据;这捕获“虚拟克隆”的目的含义,其中被分离的仅是序列信息而不是物理cDNA,每一种都与悬液中的其他分离。

[0265] 鉴定单个源分子并因此计数。简单的统计学分析对初始样品中每种cDNA已被测序的可能性进行定量。鉴定和定量每个表达的基因的相对转录物不需要如微芯片研究那样提前了解要搜索哪些基因,并且同时捕获外源和内源(例如,病原体)基因产物。

[0266] 对来自每个单个基因的单个源分子按照来自每个基因的序列变体进行分类。类似地,对揭示出相对单倍型基因表达、基因表达的表观调控的SNP变异或反映体细胞突变的序列变异进行定量。这些数量共同提供了mRNA群的统计学结构。这提供了关于可与特定的启动子元件相关联的基因级联的相对转录激活的信息,以及关于与RNA翻转、翻译速率、RNA转运相关的序列元件的协同选择以及可反映相互作用以影响表达蛋白分子机制的结构域的序列元件的协同选择的信息,从而控制组成分子表型的生物化学特性。

[0267] 序列分析的步骤可如下所示。通过侧翼序列元件(“封套”)或通过在每个文库链的一端的均匀配置来定位鉴定SMID。这可使用已有的软件来完成。

[0268] 根据SMID将标记的读段对分类成不同“区段”。区段是存储相关序列数据的计算机

存储器中的地址块。将每个包含SMID的读段及其配对读段指定到该SMID区段。每个区段代表初始样品中的单个源分子(即,单一的完整mRNA分子)并且该区段中的每条序列可追踪到该单一分子。当在同一实验中对多个mRNA群(例如,不同组织)进行测序时,读段分类可首先按照样品源分离标记的对。

[0269] 在使用SMID标签鉴定由相关读段代表的链之后,从记录的读段去除标签序列,仅留下来自源分子的信息。每个读段对提供了两个内部读段——其由初始片段化和随后的雾化作用过程中随机断裂点产生;这两个内部读段分别是与SMID标识符相邻的片段和距离第二随机断裂点下游约为文库片段平均长度处的片段。每对的两个读段对应于互补链,因此在组装前必须被转换为相同的义。

[0270] 以最大重叠比对的形式排列每个区段内修剪的读段(转化为相同的义),以产生数量最少的重叠群,每个重叠群长度最大。在充分的覆盖度下,每个区段产生包含源分子端到端序列的单一重叠群。使用已有的从头组装软件(例如,Velvet)可实现整个组装过程。

[0271] 每个组装的序列以其源基因(或在反式剪接的情况下多个基因)为参考。可使用已有的软件来更新每个基因外显子/内含子结构的综合数据(例如,Spidey)。

[0272] 鉴定源链,不管来自标签定向信息(II-pa型或II-t型标记试剂)或通过参照源基因都表明相关的cDNA序列是否对应于“正义”(蛋白编码)mRNA序列还是其反义互补物。这一步骤能够区分mRNA序列与多聚-dA标记的非编码反义序列,所述非编码反义序列可在基因表达中发挥调控或其他意料不到的作用。

[0273] 通过存在于每个基因的信息中的独特的SMID的数量,对表达的内源或外源(其中存在病原体)基因的稳态表达的相对水平进行定量,以及类似地对每个基因的每个序列变体的相对表达水平进行定量。

[0274] 这些数据提供了关于序列变异连接的相关信息;例如,特定的剪接或RNA编辑变体与特定的选择性启动子序列的相关性;特定的编码结构域的连接,其可反映控制蛋白质功能机制的相互作用的蛋白结构域,等等。重新构建可包括可一起被测序(多重测序)的来自多个组织的信使特征的比较结构,如:在来自癌症患者多个肿瘤和未受影响的非恶性对照组织的样品中;于在发育和分化的不同阶段取样的组织中;于在疾病进展过程中取样的组织中。

[0275] 随后,可对获自mRNA特征的初次重新构建的信息进行更高级的分析,例如搜索体细胞突变或遗传突变;搜索上调基因或下调基因;搜索多基因表达的组织特征模式;搜索病原体基因表达等。

[0276] 在需要对在数据分析中鉴定的特定信使的全长构建体进行功能或其他分析的情况下,可使用SMID标识符特异性和基因特异性的PCR引物的结合物扩增任何特定源分子的全长cDNA,然后可进行亚克隆和确认性测序。

[0277] 单次测序运行的产率取决于设备平台和获得的序列读段的特征。此外,下一代测序平台的能力继续增大,目前的估计必须被认为是下限。每个cDNA的覆盖度水平取决于应用。因此,相对低深度的覆盖度可足以确定SMID种类并评估剪接变体中的外显子保留:因为高通量方法的固有误差频率,需要更高的覆盖度以高精度水平识别单个碱基变化。

[0278] 在所述方法的一些实施方案中,考虑的完整序列包括长度为T(例如,但不限于,500至10000bp)的转录物的1、2或3个串联重复的覆盖度r。考虑的序列读段可以是,但不限

于,长度L (25-200bp)。因此,“重叠群”是指完全被一组重叠读段覆盖的原始序列区;即,重叠群内的每个碱基在至少一个读段中被表示,并且覆盖组中的每个读段与相同组的另一读段共有至少一个碱基。“覆盖”一个序列是指所述序列的每个核苷酸都被包含在至少一个读段内。一个序列可被多于一个重叠群完全覆盖。在这种情况下,存在相邻的核苷酸对,其都被读段覆盖,但是从不起存在于相同的读段中。这类“拆分对”定义了两个重叠群之间的界限,因此覆盖一个序列的重叠群数量比拆分对的数量多一个。45个碱基的序列被定义2个重叠群的6或7个碱基的读段覆盖。示出了定义重叠群界限的拆分对。

[0279] 明确的转录物全长序列是其中全长序列的转录物重复中不存在拆分对的那个,因为在那种情况下无法排除可能已经丢失任何插入序列的可能性。没有重复的单一序列必须被不多于一个重叠群覆盖。串联重复可被两个重叠群覆盖,因为一个拷贝中的拆分对在另一个拷贝中是未拆分的。

[0280] 因此,如果序列包含r个转录物的串联重复,则获得全长转录物序列需要重叠群的数量 $k \leq r$ 。注意这是最小值。本发明人需要,通过共有最少数量(如3个)碱基的读段重叠,重叠群中的每个碱基可从同一重叠群内的每个其他碱基达到。这给予本发明人更大的信心,读段覆盖了真正相邻的序列。未对此建模,本发明人反而确定覆盖度超过了通过一些协定量从该模型获得的最小值。由两个重叠群覆盖的三个串联重复的序列使重叠的重叠群的数量比覆盖单拷贝所需的数量增加1.5倍,并且增大了重叠群会在其末端具有更大重叠的可能性。

[0281] 按照 $\text{covg} = NL/T$ 计算覆盖度,其中N是读段总数,L是读段长度,T是转录物长度。如果通过用k个重叠群覆盖r个串联重复来获得covg,那么单一转录物覆盖度是具有所述重复的完整序列的覆盖度的r倍。

[0282] $k = N \exp(-NL/rT)$

[0283] 求解N得出:

[0284] $N = -kA W_{-1}(-1/A)$

[0285] 其中 $A = rT/kL$,并且 W_{-1} 是朗伯W函数的实数的分支,在本发明的情况中其返回N的实数(即,不是复数)值(Adv,Comparative Mathematics,5,329-359,1996)。

[0286] 根据L、r、T和k的不同数值计算数据表格。该信息被最直观地绘制出来,示于图8B、8C中。所计算的参数的读段数量相对于覆盖度的图显示出总体行为。对于一个转录物长度(T),一个转录物重复值(r)和一个读段长度(L)而言,每条线段连接每个总序列(包括重复)的k(1、2、3、5和10)个重叠群的5个数值的点。三个重复水平(r)将图分成三组:(r=1)、(r=2)和(r=3)。这显示出如果转录物以串联方式被拷贝但被固定数量的重叠群覆盖,单个转录物的覆盖度如何增大。更大的转录物位于上部,因为需要更多的读段来覆盖它们。

[0287] 对数据的子集作图以说明读段长度对捕获全部cDNA序列所需的总覆盖度的影响。为保证覆盖cDNA的相同水平,更长的cDNA(例如,10kbp)对50bp读段比对150bp读段需要多读取3.5到4倍碱基。

[0288] 图8D中所示的缩略表格表明,3kbp cDNA所需的约10x深度需要约150bp的230次读取。读段为167bp时的覆盖度的更低深度可需要约100次读取。目前,Illumina High Seq平台的一个室的一次典型运行产生约2亿个读段对,每个读段序列为167bp,足以完全测序约200万cDNA,其为数均mRNA大小的约二倍。因此,使用该设备的全部15个室的能力足以对约

3000万平均大小的cDNA进行测序。

[0289] 假定一般细胞表达7000-8000蛋白编码基因的产物,对单个室而言,这允许每个基因约250个mRNA的动态范围,或在一次运行中使用全部室分析范围为3000-4000之间的mRNA。不考虑该范围只会随着设备性能改善而增大的事实,这似乎足以分析甚至中等复杂的组织的全部信使RNA。

[0290] 术语

[0291] 术语“多核苷酸”是指由两个或多个(优选多于三个,通常多于10个)脱氧核糖核苷酸或核糖核苷酸组成的分子。精确大小取决于许多因素,进而又取决于所述多核苷酸的最终功能或用途。可以通过任何方法生成多核苷酸,所述方法包括化学合成、DNA复制、逆转录或其组合。

[0292] 术语“核酸”是指核苷酸聚合物或多核苷酸,如上所述。使用所述术语来指定单个分子或分子的集合。核酸可以是单链或双链,并且可包含编码区和多种控制元件区域。

[0293] 术语“互补”和“互补性”是指通过碱基配对原则相关的多核苷酸(即,核苷酸序列)。例如,序列“A-G-T”与序列“T-C-A”互补。互补性可以是“部分的”,其中仅有一些核酸碱基根据碱基配对原则匹配。或者,核酸之间存在“完全”或“总体”互补性。核酸链之间的互补性程度对核酸链之间的杂交效率和强度具有显著影响。在扩增反应中,以及依赖于核酸之间结合的检测方法中,互补性程度尤其重要。

[0294] 术语“回文序列”是指无论从一条链的5' (5端引物)读到3' (5端引物)或从互补链的5'读到3'都是相同的核酸序列(DNA或RNA)——如果核苷酸序列与其反向互补物相同,则将其称为回文序列。回文核苷酸序列能够形成发夹。所述术语意欲包括存在基本互补性但可包含一些错配碱基对(例如,不破坏自身杂交或形成多个环的错配碱基对)的序列。

[0295] 限制性位点或限制性识别位点是核酸分子上包含特定核苷酸序列的位置,其被限制性酶(核酸酶)或其他有能力的分子切割。在本文公开的任何实施方案中,限制性位点可称为切割位点。所述位点通常是回文序列,特定的切割分子(例如,限制性酶)可切割其识别位点内或附近位置处的两个或多个核苷酸之间的序列。天然存在的限制性酶通常识别长度为4-6bp的序列。这些术语与限制性酶共有序列是同义的。考虑非天然存在的切割酶和分子。Chu和Orgel报道了单链DNA的非酶促序列特异性切割。参见PNAS,1985,82:963-967。还可参见Dervan,Science,1986,232:464-47;Dreyer&Dervan PNAS,1985,82(4):968-972;和美国专利号6555692和4795700。

[0296] “不常见的限制性位点”是指长度大于6、7或8bp的被切割分子或其他限制性酶切割的位点。通过突变或改造已有的酶或产生嵌合的限制性核酸酶,限制性修饰酶产生具有更长识别位点的限制性核酸内切酶。锌指蛋白通常用于具有非常合适的序列特异性的嵌合限制性酶。这些蛋白通常通过将 α 螺旋插入双螺旋的大沟内而结合于核酸。例如,可设计核酸酶,其通过使锌指蛋白融合至Fok I核酸内切酶的切割结构域而在优选的位点切割DNA。参见Kim et al.,Proc.Natl.Acad.Sci.USA 1996,93,1156-1160。

[0297] 术语“杂交”是指互补核酸的配对。杂交和杂交强度(即,核酸之间结合的强度)受到如核酸之间的互补程度、反应条件的严格度、形成的杂交体的 T_m 和核酸内的G:C比的因素的影响。在其结构中包含互补核酸配对的单个分子被称为“自身杂交”。

[0298] 术语“引物”是指在纯化的限制酶切消化中天然存在的或通过合成产生的多核苷

酸,当被置于诱导与核酸链互补的引物延伸产物合成的条件下(即,在核苷酸和诱导试剂如DNA聚合酶存在时,在适合的温度和pH下)时,其能够作为合成的起始点。为了扩增效率最大,引物优选为单链,但是也可选择为双链。如果是双链,在用于制备延伸产物前首先处理引物以分开两条链。引物必须足够长以在诱导试剂的存在下引发延伸产物的合成。引物的精确长度取决于许多因素,包括温度、引物来源和方法使用。

[0299] 术语“测序”是指可用于鉴定特定核酸中核苷酸顺序的任何方法。已知用于核酸测序的方法和设备,在某些实施方案中,测序方法不限于所使用的特定方法、设备或数据/质量过滤。Bokulich等人报道,质量过滤改进了由Illumina GAIIX、HiSeq和MiSeq设备产生的测序。参见Nature Methods,2013,10:57-59。

[0300] 术语“聚合酶链式反应”(PCR)是指K.B.Mullis在美国专利号4,683,195、4,683,202和4,965,188中记载的方法,其中描述了用于提高靶序列片段在混合物中浓度的方法。这个用于扩增靶序列的方法由以下步骤组成:将大大过量的两种多核苷酸引物引入到包含所需靶序列的DNA混合物中,然后在DNA聚合酶的存在下进行精确顺序的热循环。所述两个引物与双链靶序列中它们各自的链互补。为了进行扩增,使混合物变性,然后引物退火到其在靶分子内的互补序列上。退火之后,使用聚合酶延伸引物以形成新的互补链对。变性、引物退火和聚合酶延伸步骤可多次重复(即,变性、退火和延伸组成一个“循环”;可以有多个“循环”),以获得所需靶序列的高浓度扩增片段。所需靶序列的扩增片段的长度由引物彼此间的相对位置决定,因此,该长度是一个可控的参数。由于这个过程的重复方面,所述方法被称为“聚合酶链式反应”(以下称为PCR)。因为所需的靶序列扩增片段成为混合物中的主要序列(在浓度方面),它们被称为“PCR扩增的”。

[0301] 通过PCR,可将基因组DNA中特定的靶序列的单拷贝扩增到通过若干不同的方法可检测的水平(例如,使用标记探针杂交;纳入生物素化引物,然后进行抗生物素蛋白-酶缀合物检测;将³²P标记的三磷酸脱氧核苷酸如dCTP或dATP纳入到经扩增的片段中)。除了基因组DNA,可使用适合的引物分子组扩增任何多核苷酸或多核苷酸序列。特别地,通过PCR方法本身生成的扩增片段本身是用于随后PCR扩增的有效模板。

[0302] 术语“PCR产物”、“PCR片段”和“扩增产物”是指完成两轮或多轮变性、退火和延伸的PCR步骤之后所得的化合物混合物。这些术语包括已经扩增了一个或多个靶序列的一个或多个片段的情况。

[0303] 术语“扩增试剂”是指扩增所需的那些试剂(三磷酸脱氧核糖核苷酸、缓冲液、引物、核酸模板和扩增酶等)。通常,扩增试剂以及其他反应成分被放置并包含于一个反应容器(试管、微孔等)中。

[0304] 在某些实施方案中,将本文公开的方法与以下进一步描述的以及Bentley et al.,Nature,2008,456,53-59和Meyer et al.,Nature protocols,2008,3,267-278中所描述的双端、配对方法结合使用,所述文献以引用的方式纳入本文。

[0305] 某些方法使用和生长中的双链序列连接的荧光标记的核苷酸,其中使用化学官能团来控制聚合作用。使用相同的多核苷酸来增大固体表面的面积,经荧光标记的核苷酸表明添加了哪个碱基。所述方法还可拓展到其他方案,包括中等大小片段(>300bp)的全长测序。

[0306] 在双端方法中,使核酸断裂成通常小于800bp的片段(例如,通过但不限于,酶促片

段化、超声作用、水剪切、雾化作用)。对(双链)片段进行末端平端化,加A-尾,并将其连接至具有单链延伸的叉状衔接体,所述单链延伸导致PCR扩增以将不同的(A和B)双链延伸引入到每个片段的相反末端。通过PCR产生的这些末端包含功能性位点,其以后用于文库PCR、簇合成和引物指导的末端测序。使用每个DNA片段的正链和负链的两个方向的末端标签(A和B)生成PCR产物。

[0307] PCR扩增和凝胶纯化后,使双链的、矢量修饰的、平端DNA片段变性成单链。将它们退火到捕获室(流动池)的表面小区上的共价连接的多个单链多核苷酸(与3'端延伸B或A互补)上。

[0308] 捕获多核苷酸引发与退火的单链DNA互补的链的合成,然后使产物变性,将(非共价)模板冲洗掉。然后使保留的链退火到与其游离3'端互补的附近的捕获多核苷酸上。从该捕获多核苷酸延伸出第二条链,产生双链“桥”,其任一端仅与DNA双链体的5'端相连。

[0309] 使这些桥变性,单链重新退火到新的捕获多核苷酸,重复该过程,直到对于每个最初退火到室表面的DNA片段而言,扩增产生多核苷酸簇——其通过其A或B端连接至所述表面。这通常产生每个流动池的大量簇。

[0310] 测序通常分两个阶段进行。在阶段I,使用特异性针对两个连接元件之一的试剂来切割DNA桥的群,使非共价连接的链变性并将其冲洗掉。这留下仅在两个方向之一(例如,B共价的5'端)的单链DNA被从游离端进行测序,使用与上述叉状衔接体一起引入的A测序引物。

[0311] 通过以下方式进行测序:用A-寡核苷酸进行引发,从四种可区分的荧光三磷酸核苷酸衍生物的溶液中连续纳入适合的碱基;在光学记录每个新加入的碱基之后,荧光基团被水解,重复进行反应。以这种方式,可光学记录游离端的最多达150、250或更多个碱基。

[0312] 在阶段I之后,使在测序过程中产生的未束缚的链变性并将其冲洗掉。然后,使保留的模板的游离3'端重新退火到小区上的捕获寡核苷酸(例如,A)。通过延伸所述捕获寡核苷酸来合成相反方向的新链。然后在第二(例如,B)连接元件处切割所得的桥式寡核苷酸群,使所得的未束缚的链变性并将其冲洗掉,如所说明的。剩余的群代表游离3' (“B”)端,如之前所述进行测序,使用“B”多核苷酸引发以产生第一测序链相反末端的互补序列。一起报道每个簇的这些读段对。

[0313] 在配对测序中,使核酸片段化(例如,但不限于,通过酶促片段化、超声作用或水剪切)成片段,通常为数kb。对这些随机片段进行末端平端化,在其末端生物素化并通过酶连接进行环化;使用核酸外切酶I和II除去剩余的线性产物。

[0314] 环化使剪切片段的两个生物素化末端连接在一起。环状核酸被随机断裂成长度通常为300-500bp的更短的线性片段。使所述具有生物素的短片段吸附到链霉亲和素小珠上,将未生物素化的片段冲洗掉并丢弃。对保留的片段进行末端平端化,加A-尾,将其连接到叉状衔接体(如上所述)并通过凝胶纯化进行大小选择。所得的片段组成随机分布的序列元件对的文库,在所述文库中,每个元件或元件对是通过核酸上的已知平均距离(第一剪切长度)与其他元件分开。根据之前部分中描述的方案对该配对文库进行测序。

[0315] 实施例

[0316] 实施例1:使用标记试剂对mRNA进行测序

[0317] 使用标准试剂盒分离来自细胞和组织的多聚-A mRNA,并且除去基因组DNA的残留

物是常规的 (DNA-Free™, LifeTechnology)。

[0318] 1. 从RNA逆转录出cDNA (莫洛尼鼠白血病病毒逆转录酶), 使用包含SMID的标记试剂引发; 用RNA酶H处理异源双链体。可以用其他病毒逆转录酶或任何能够进行RNA逆转录的其他来源的同等的酶来代替莫洛尼鼠白血病病毒逆转录酶。

[0319] 2. 使标记的单链cDNA环化 (T4 RNA, DNA连接酶 (CircLigase; Epicentre)); 使用核酸外切酶I除去剩余的线性cDNA。

[0320] 3. 将环化的cDNA悬液分成等分试样, 并对其进行滚环扩增 (RCA) (phi 29 DNA聚合酶) [待扩增的cDNA群可随着引物的选择而变化。]

[0321] 4. 任选地使用S1核酸酶或绿豆核酸酶使超分支 (hyperbranch) 的RCA cDNA同聚物去支链 (debranch); 转移至酶促片段化或超声缓冲液, 并且片段化成预先选择的平均大小 (例如, 2-4kb)。

[0322] 5. 使用限制性酶切割片段; 缓冲液交换。将该物质提交给商业基因组中心, 用于通过标准方法进行文库制备和测序。可提供引物/标记的特定修饰的PCR引物以代替标准试剂盒的A引物。

[0323] 实施例2: 非多聚腺苷酸化的RNA

[0324] Salzman, J. 等人在PloS One, 2012, vol 7, issue 2, e30733中报道了环状RNA是多种细胞类型中几百个人类基因的主要转录物同工型。这些RNA未被多聚腺苷酸化。这类RNA产物适于使用标记试剂通过这种技术进行测序, 所述标记试剂为低化学计量的, 具有随机3'末端序列以获得RNA的拷贝, 然后进行如本文所述的环化和加工。

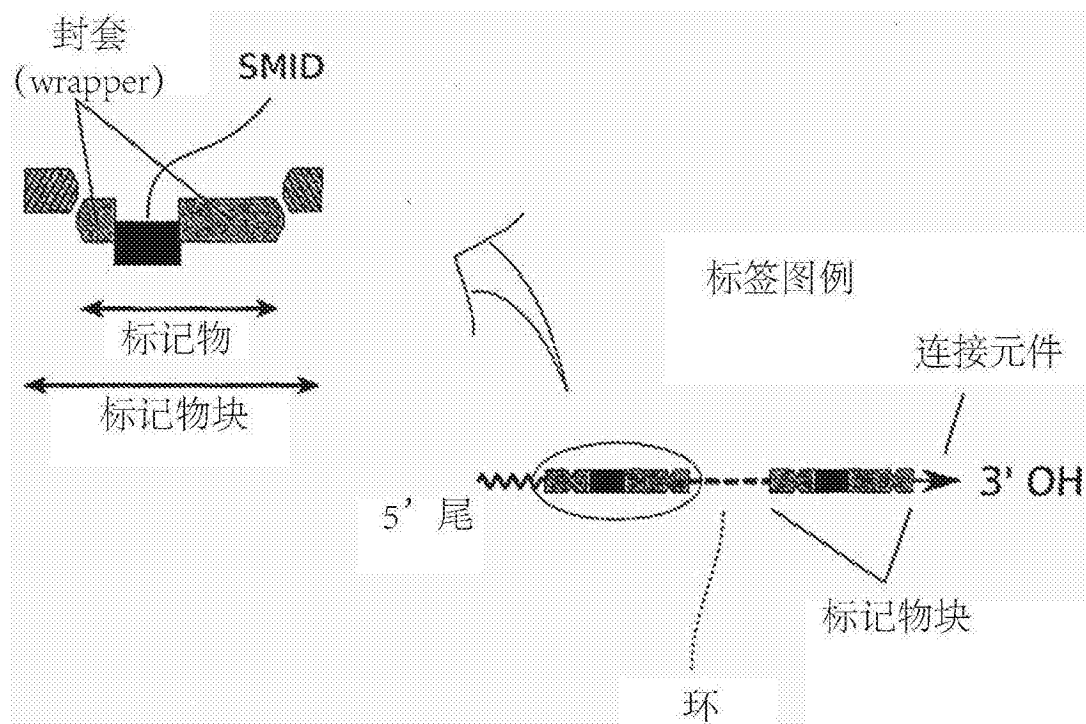


图1A

四种基本的标
签类型

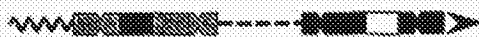
标签类型

标签

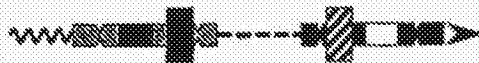
I



II-ps



II-pa



II-t

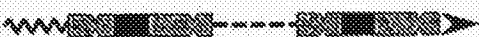


图1B

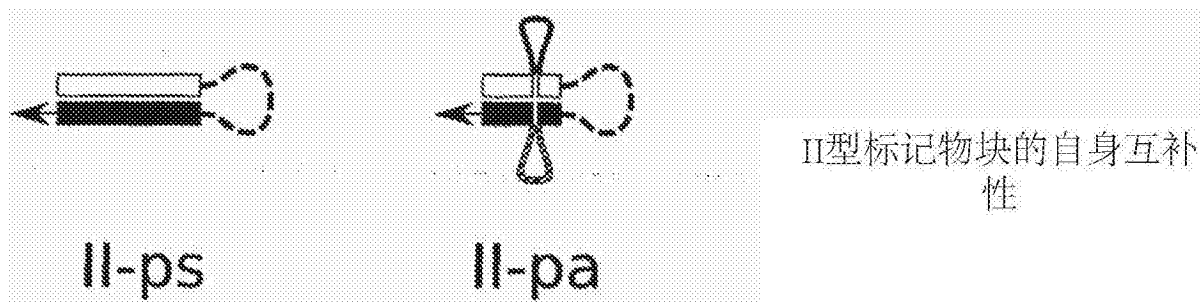


图1C

标签类型	标记试剂的构象	
	单尾	双尾
I		
II-ps		
II-pa		
II-t		

图1D

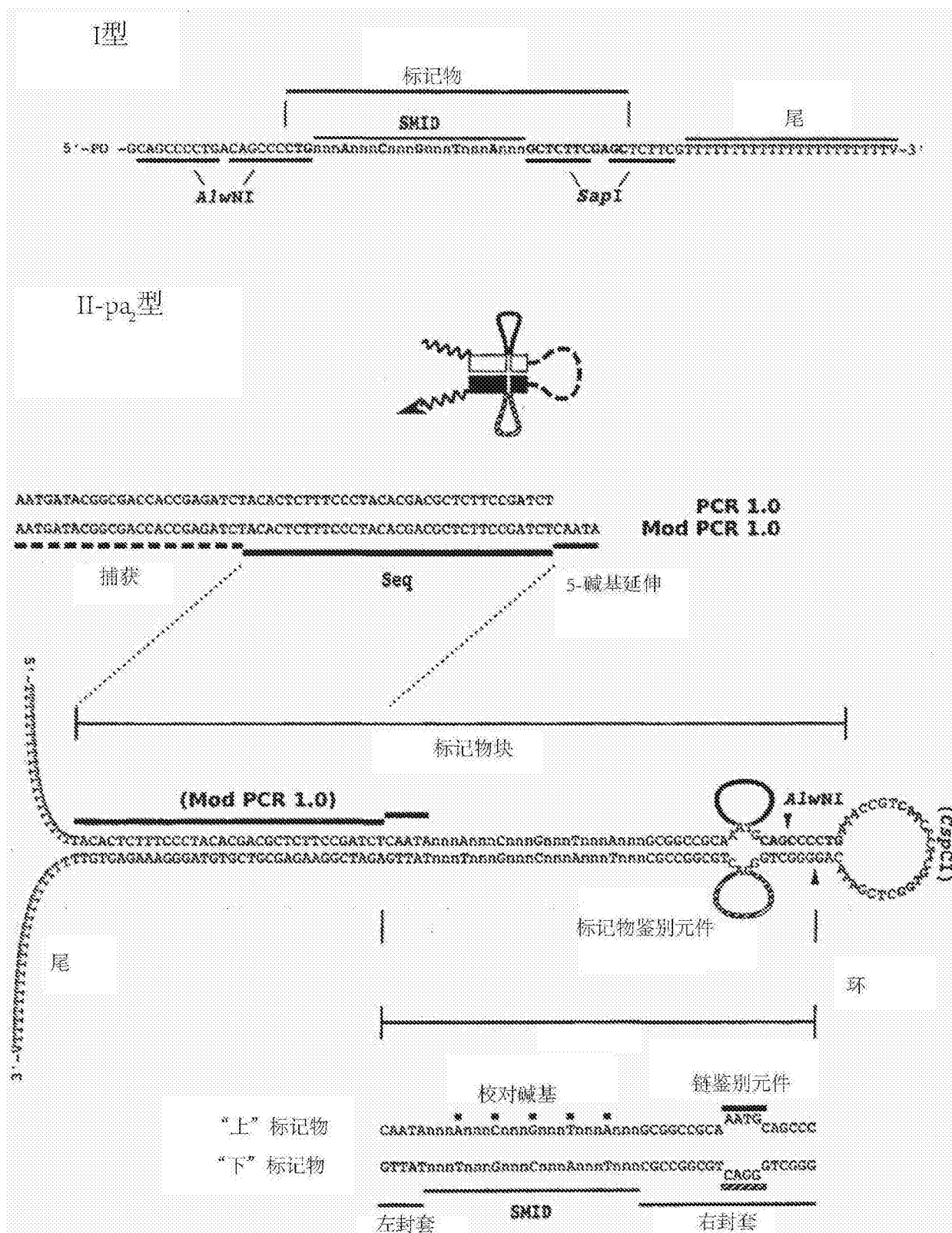


图1E

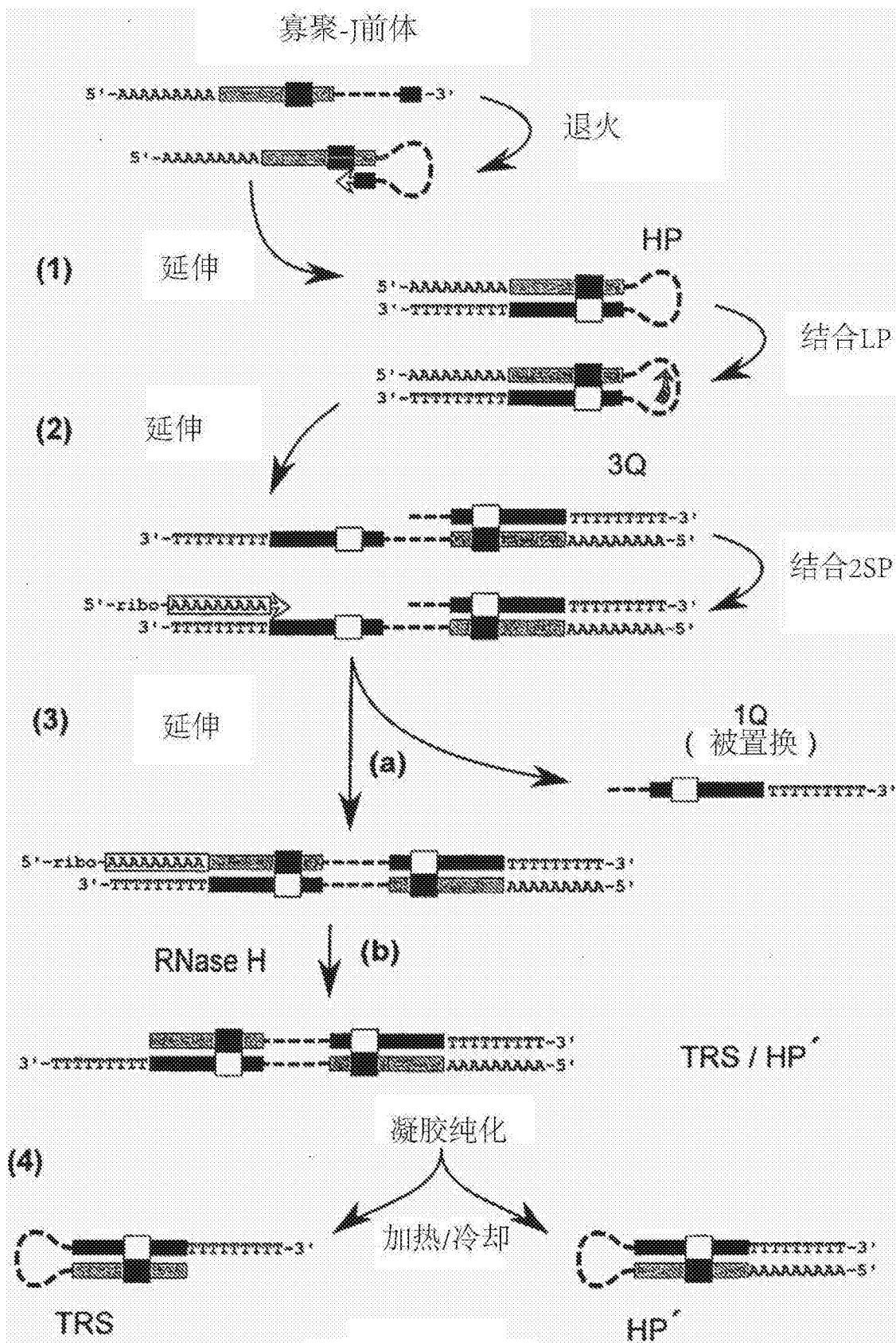


图2A

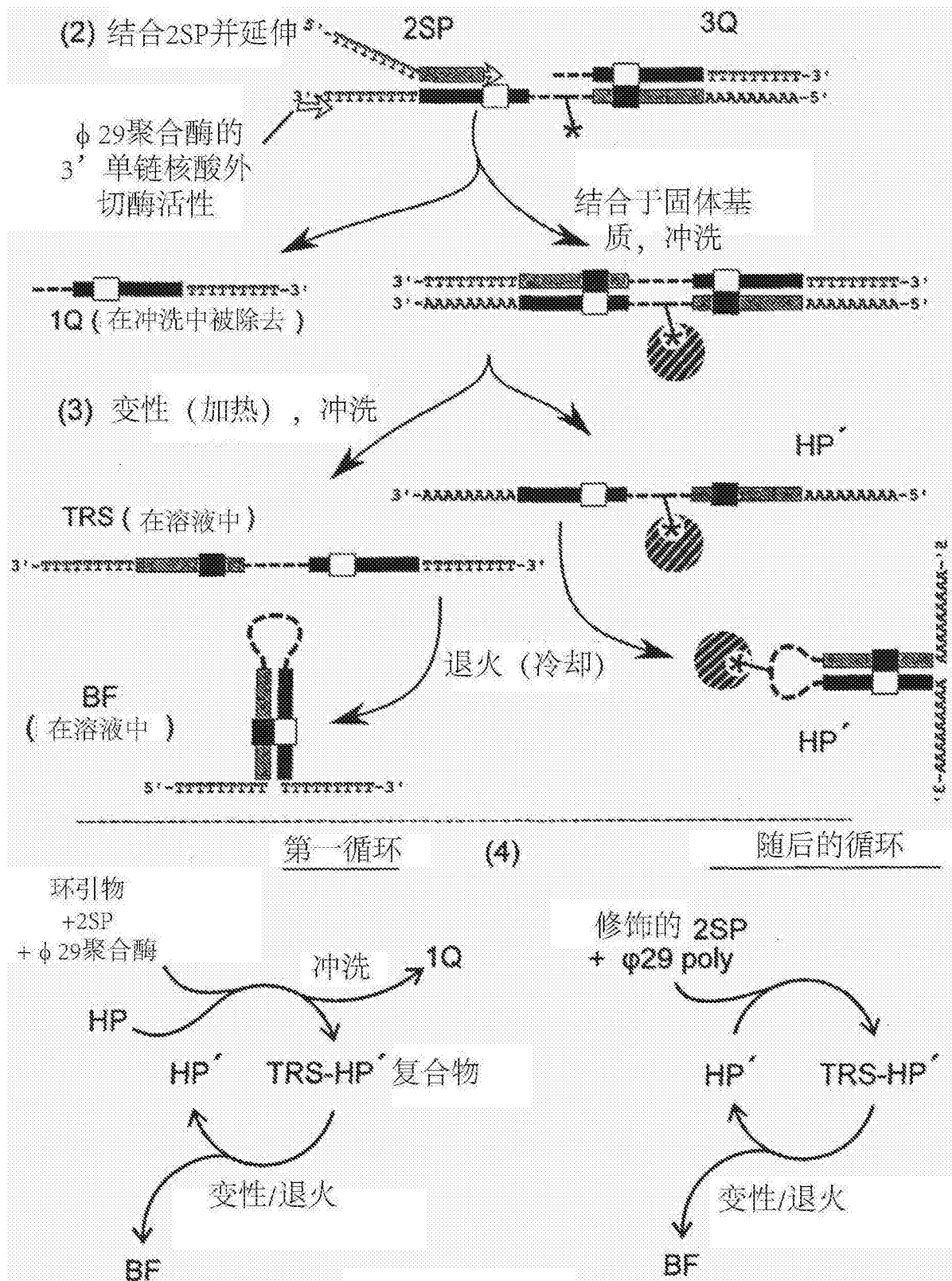


图2B

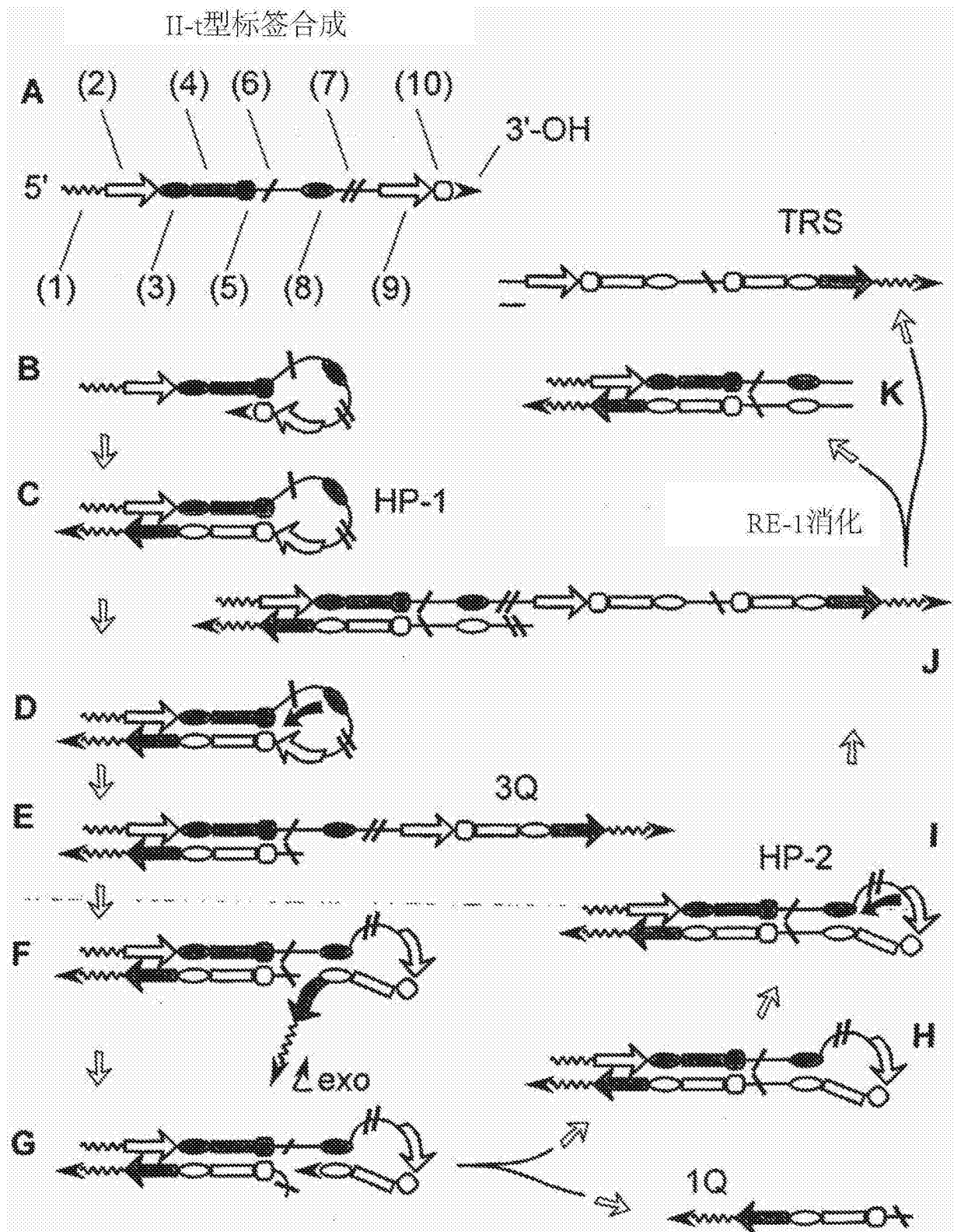


图2C

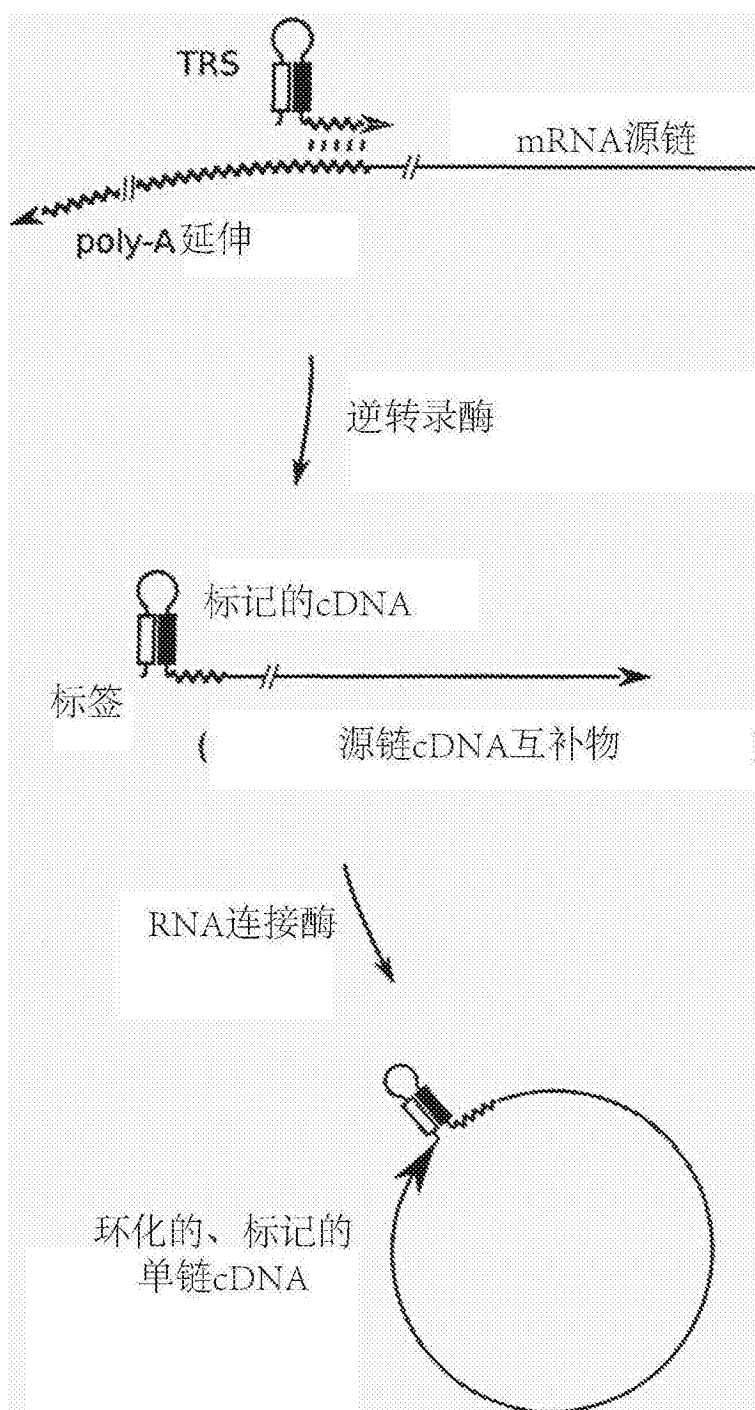


图3A

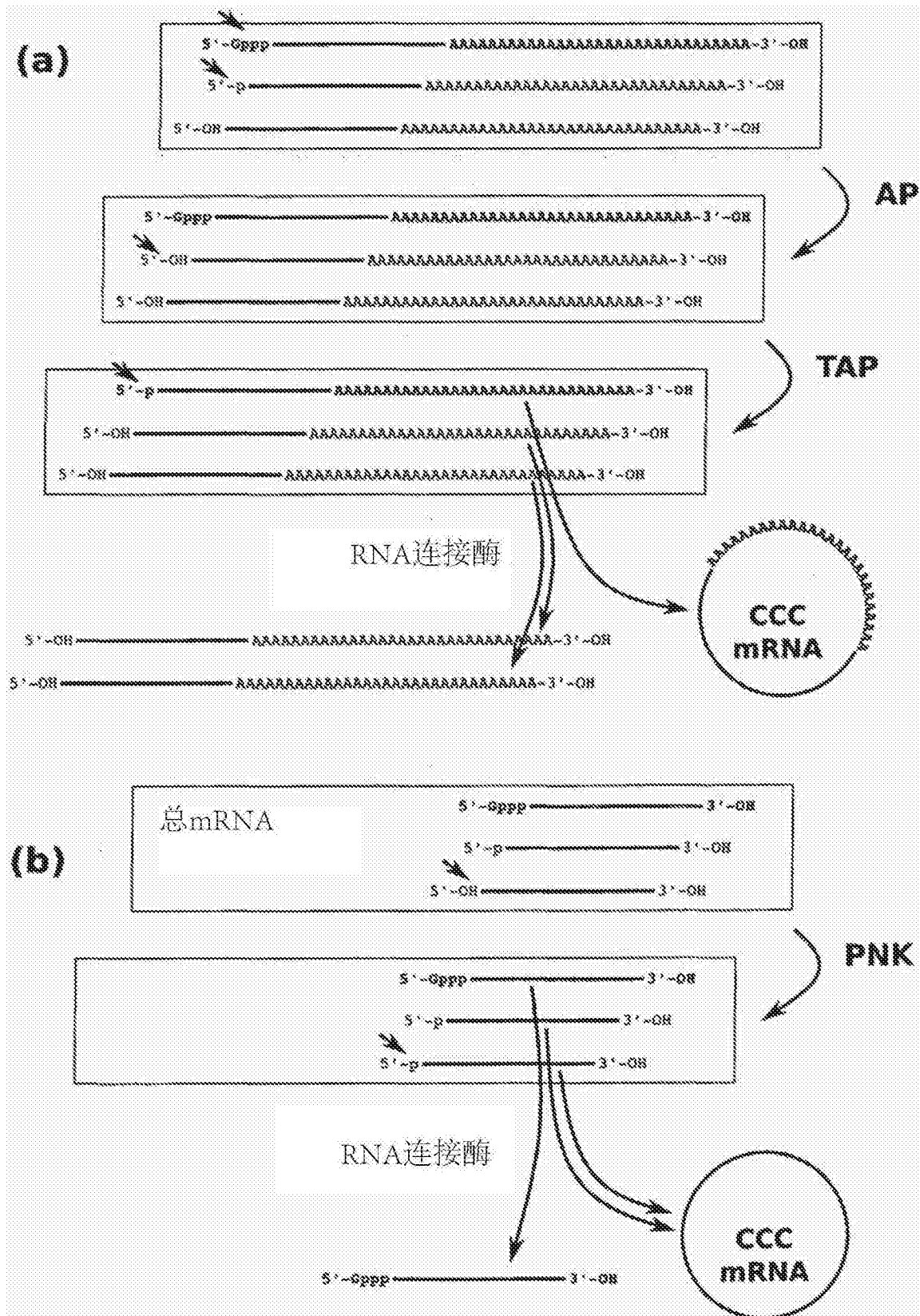


图3B

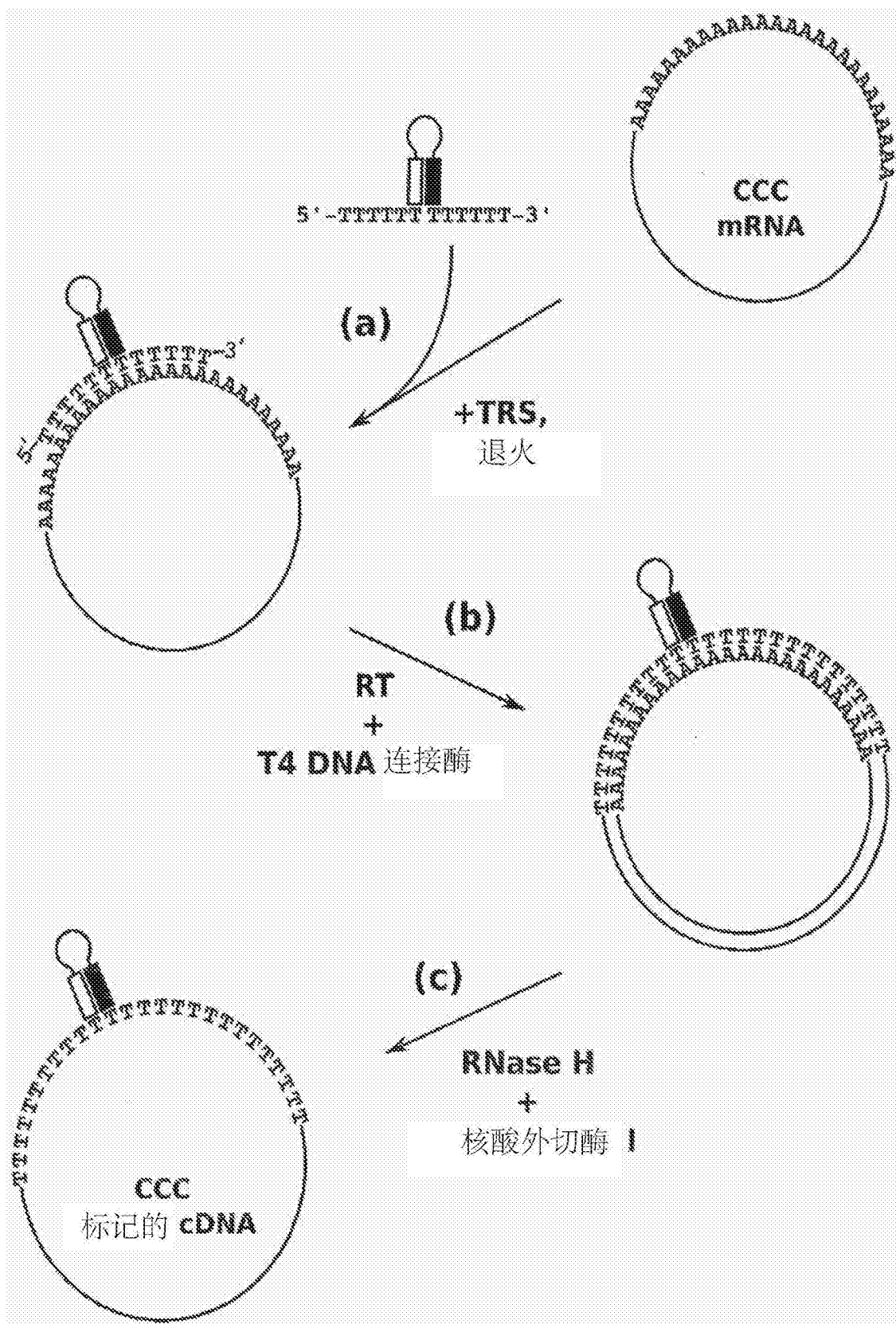


图3C

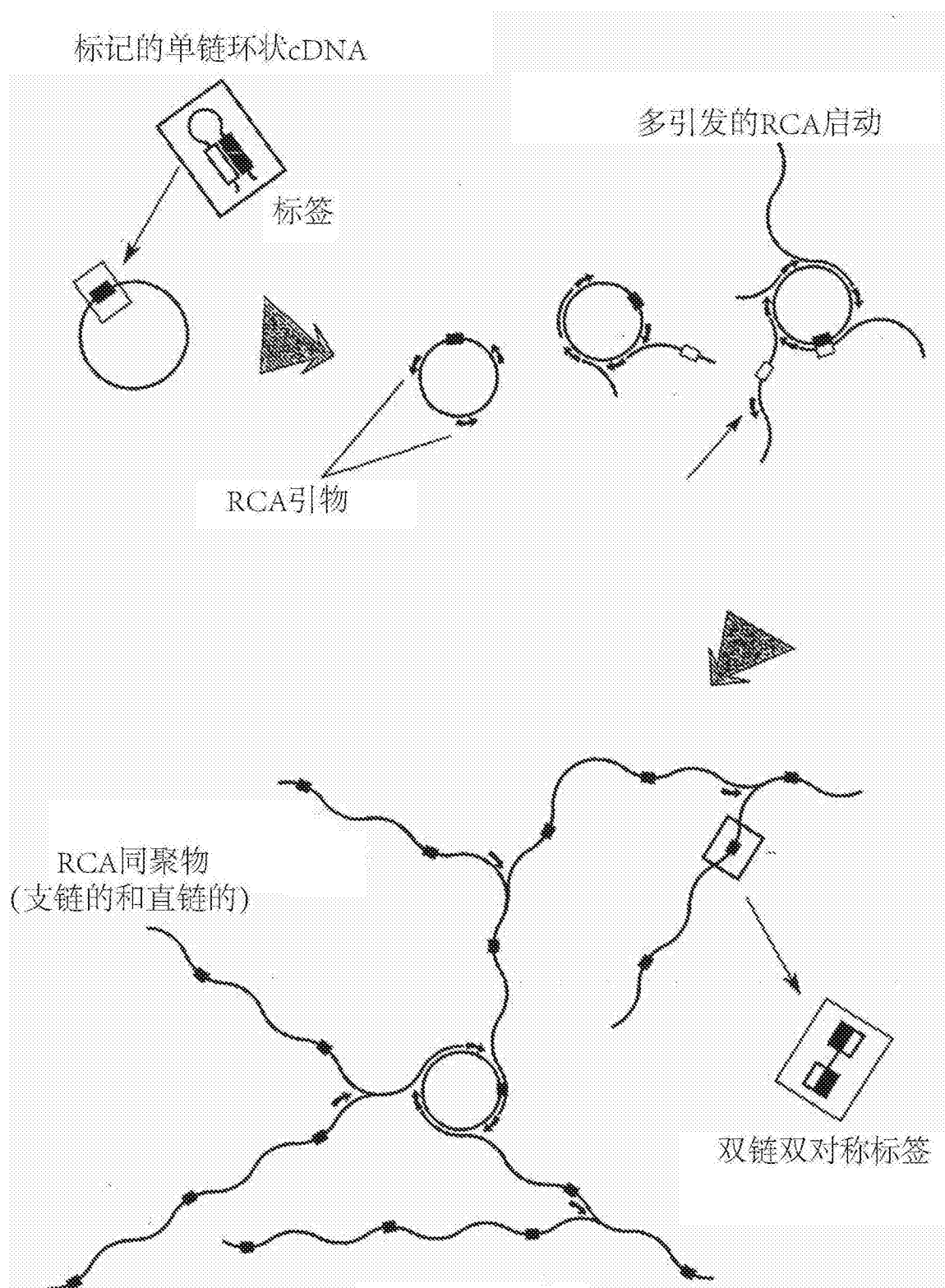


图4

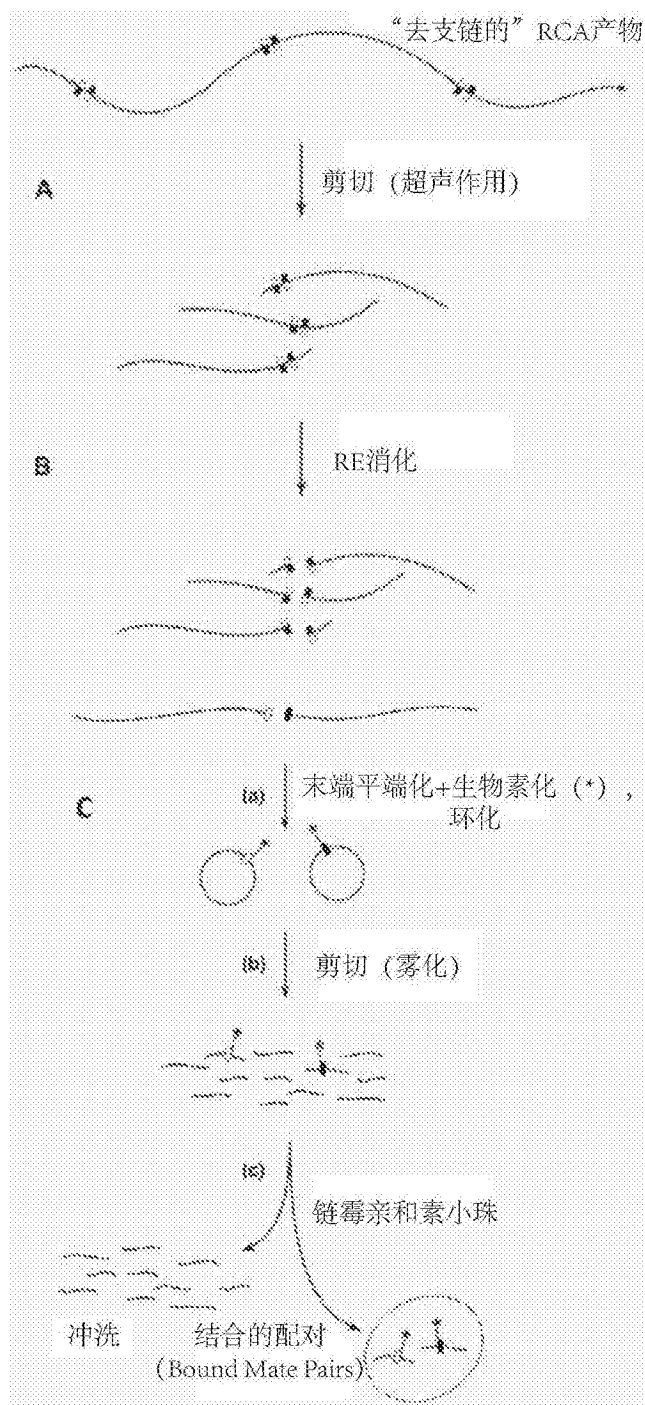


图5

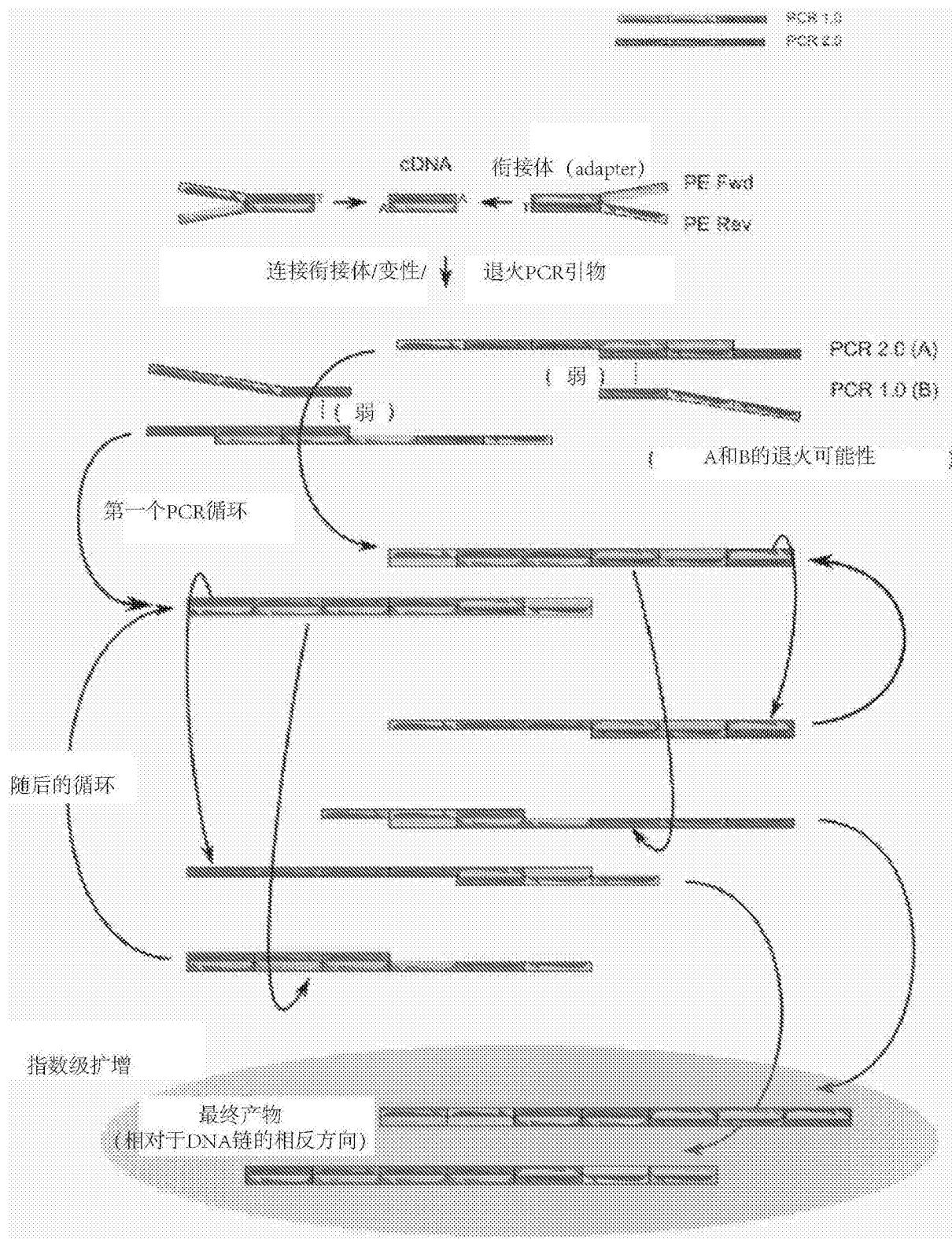


图6A

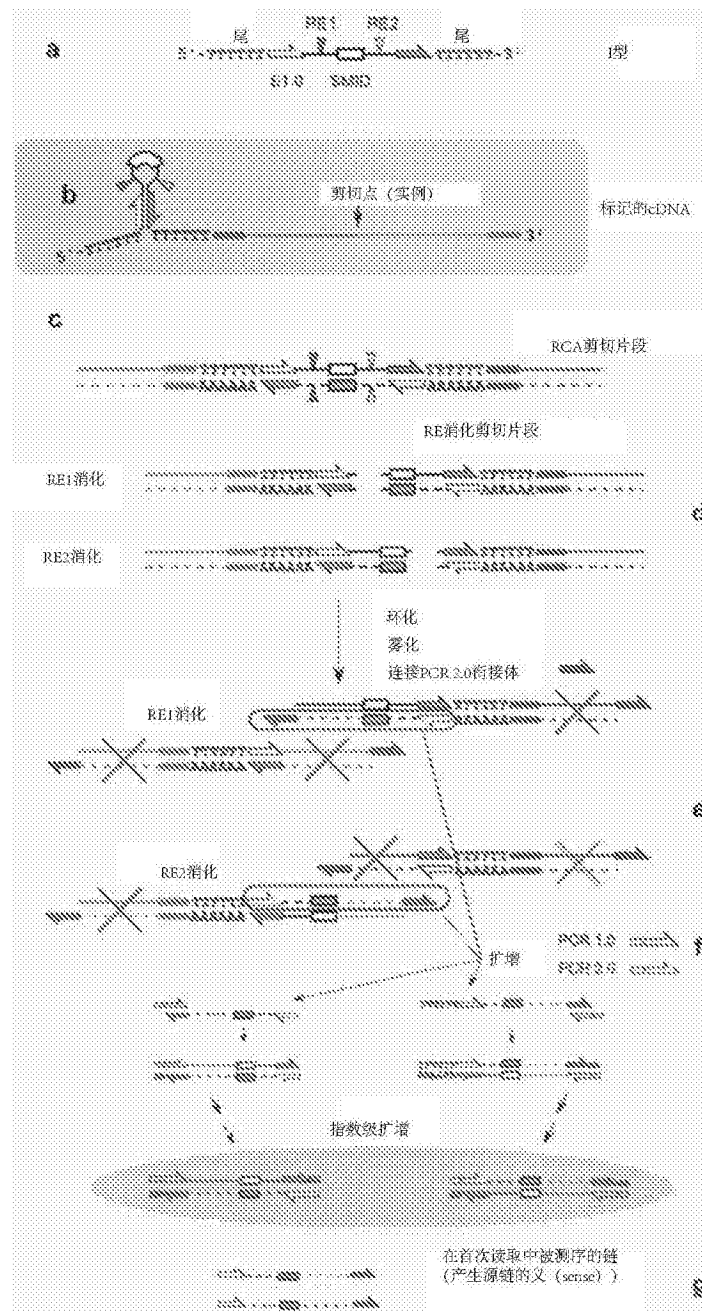


图6B

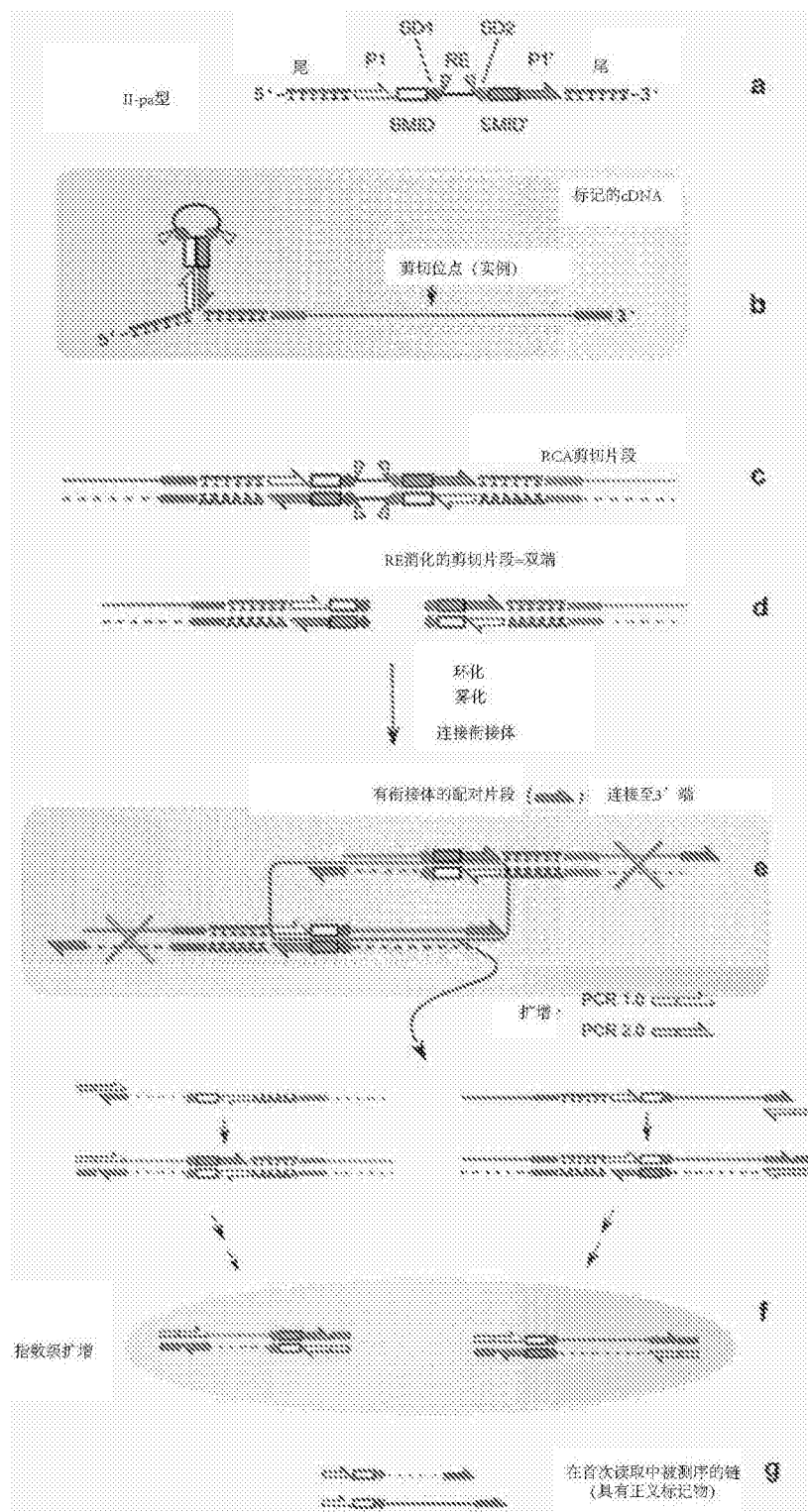


图6C

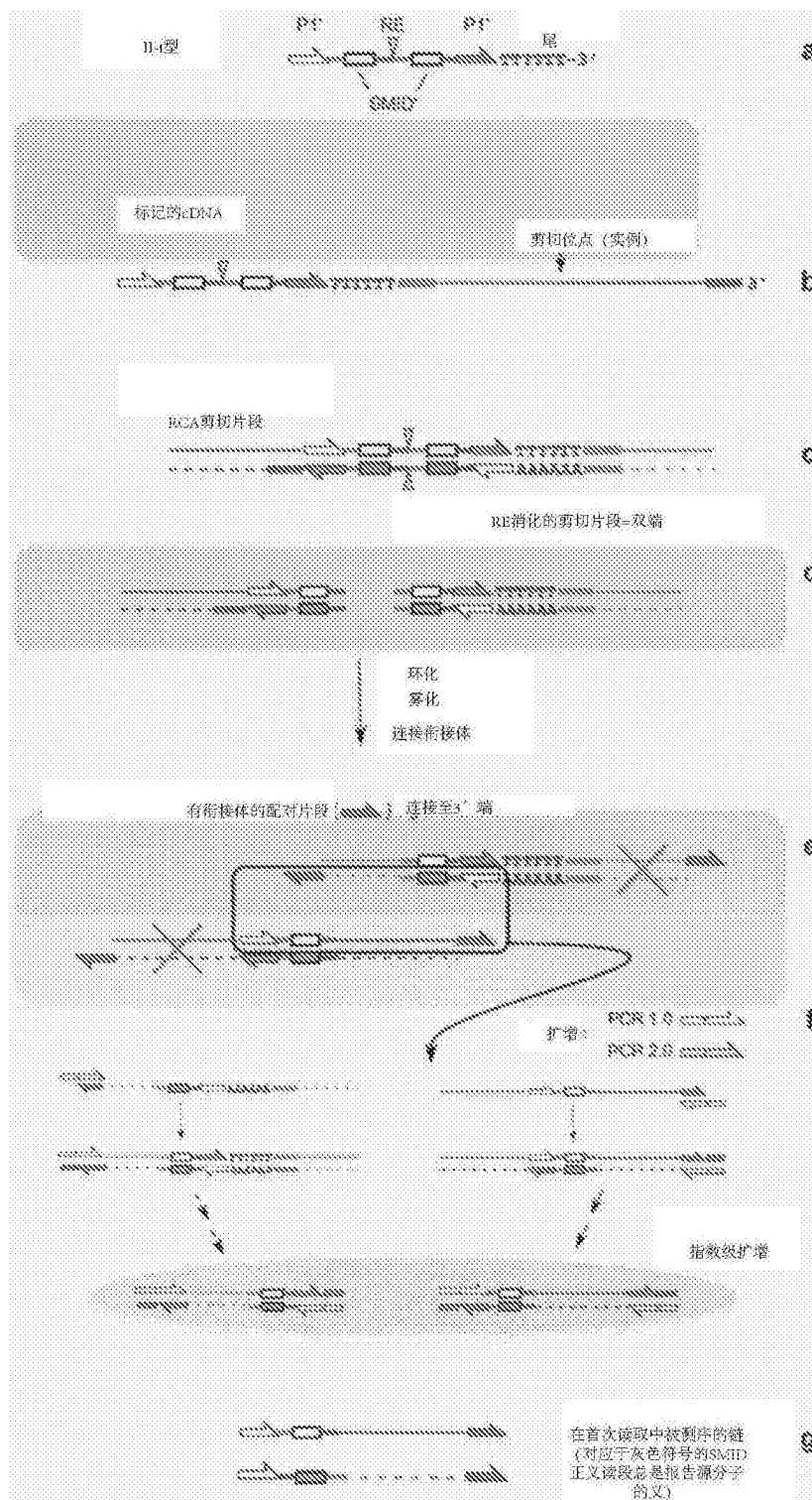


图6D

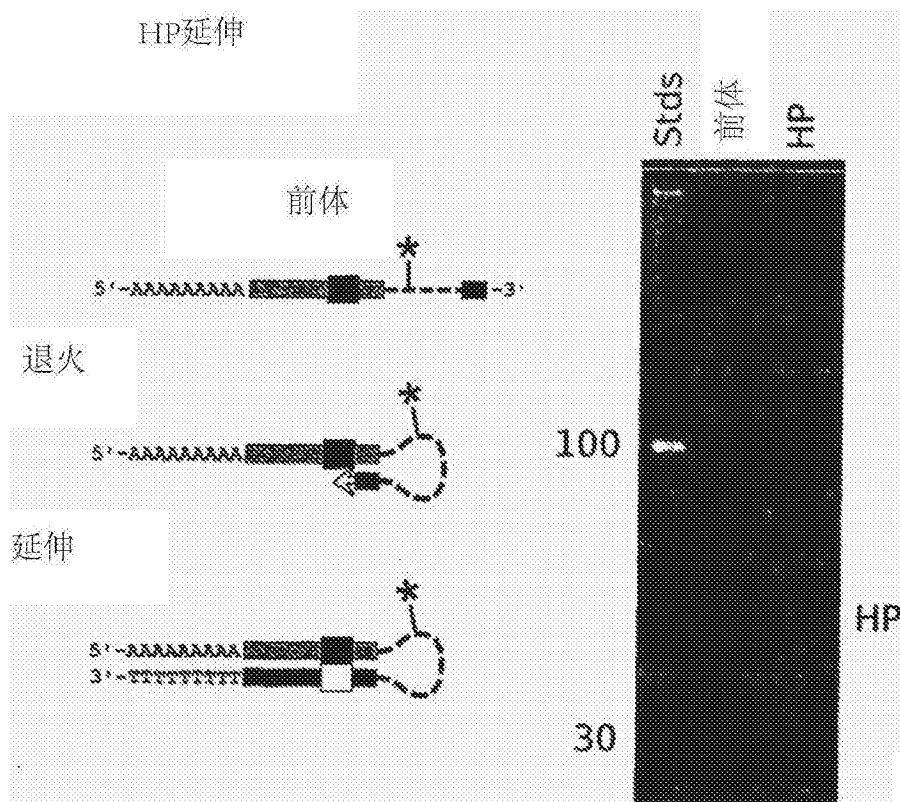


图7A

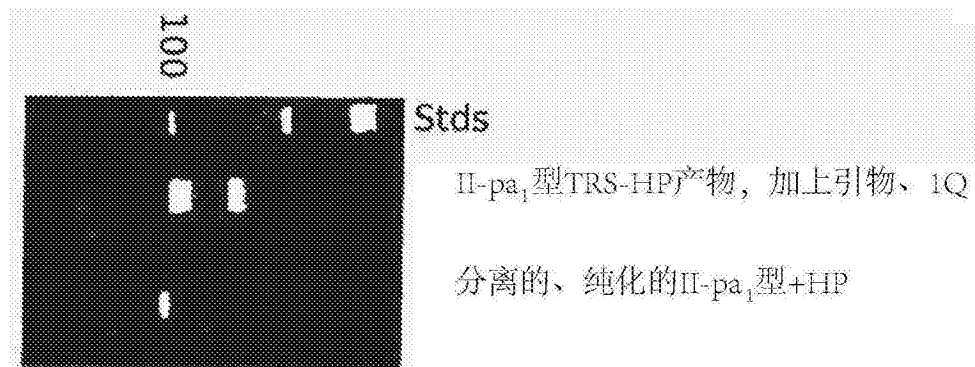


图7B

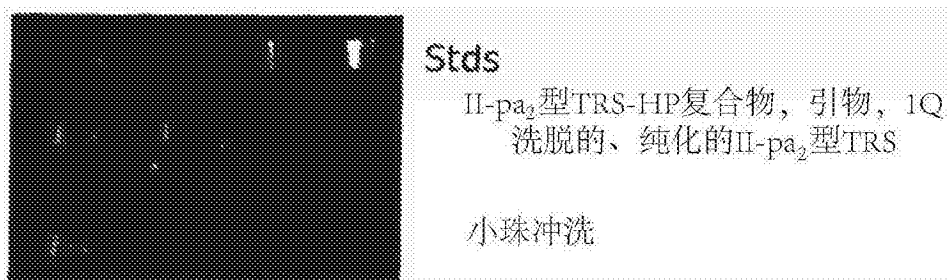


图7C

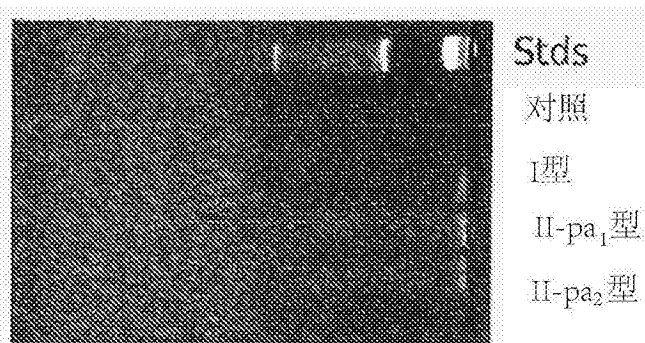


图7D

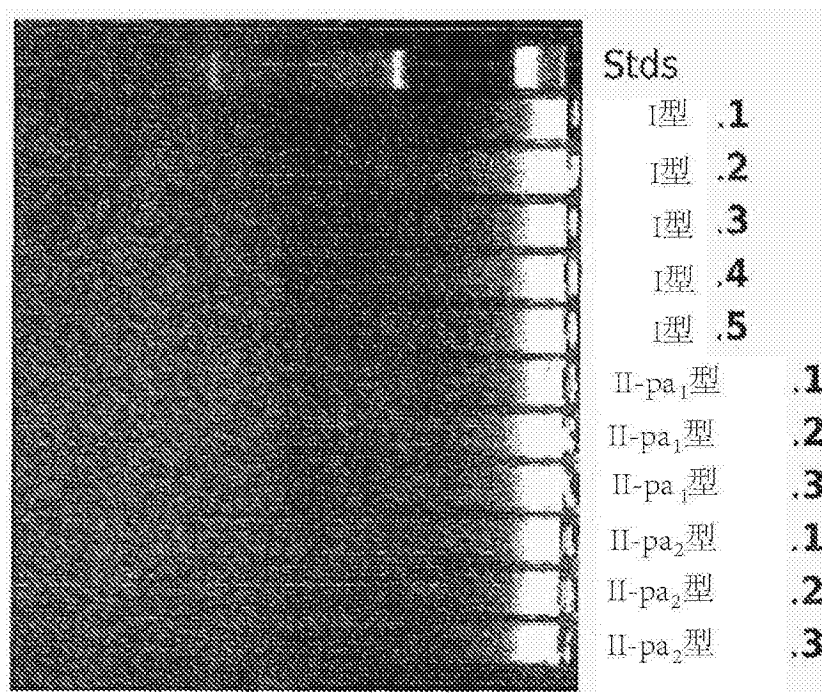


图7E

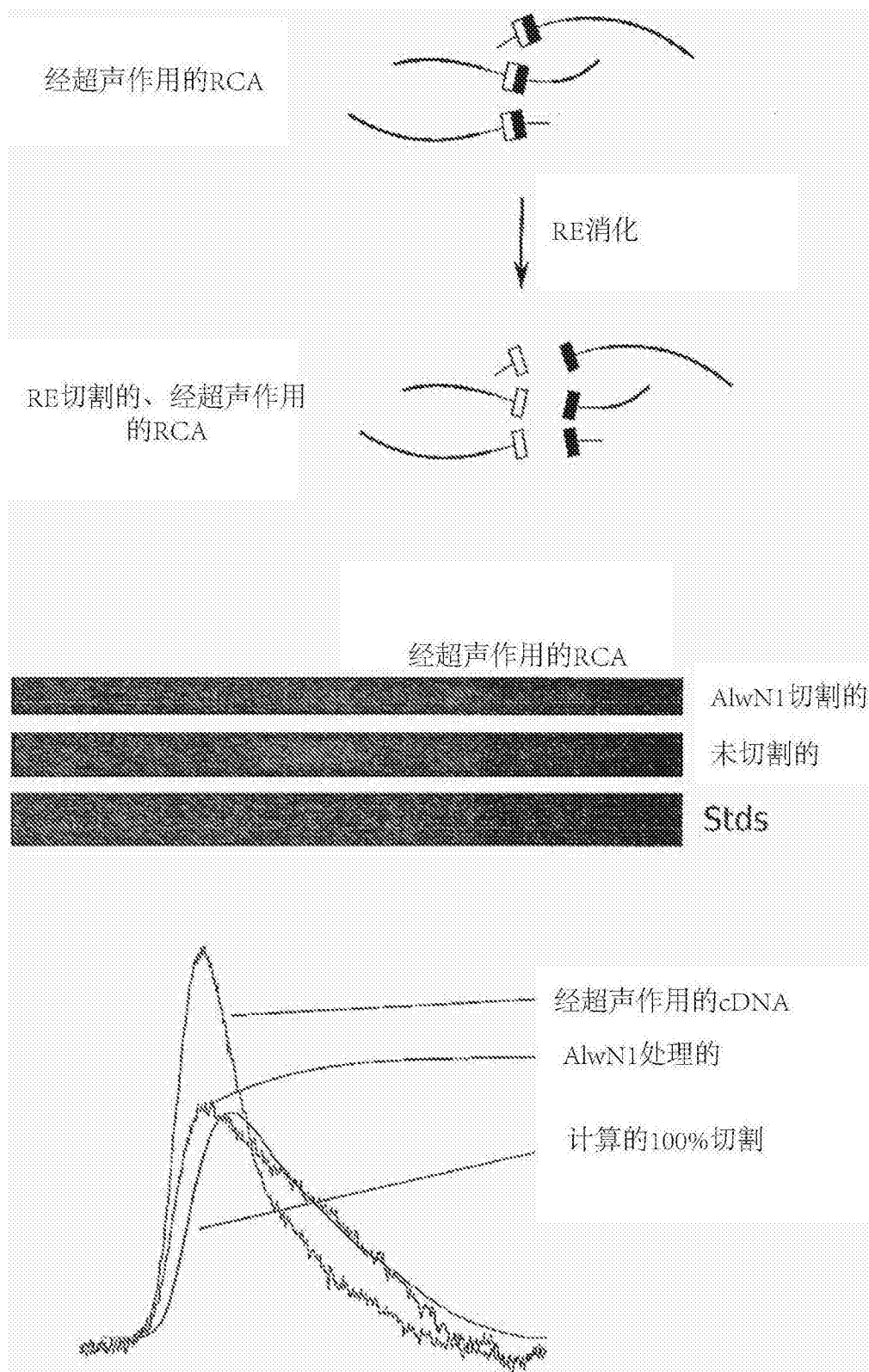


图7F

复制的、标签-引物扩增的配对文库

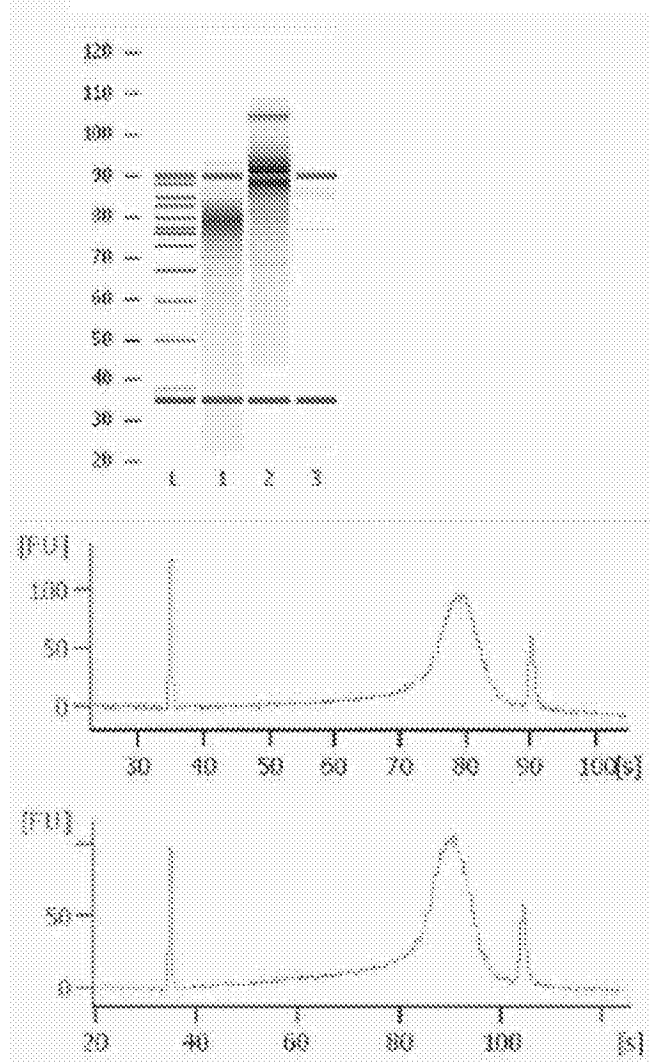


图7G

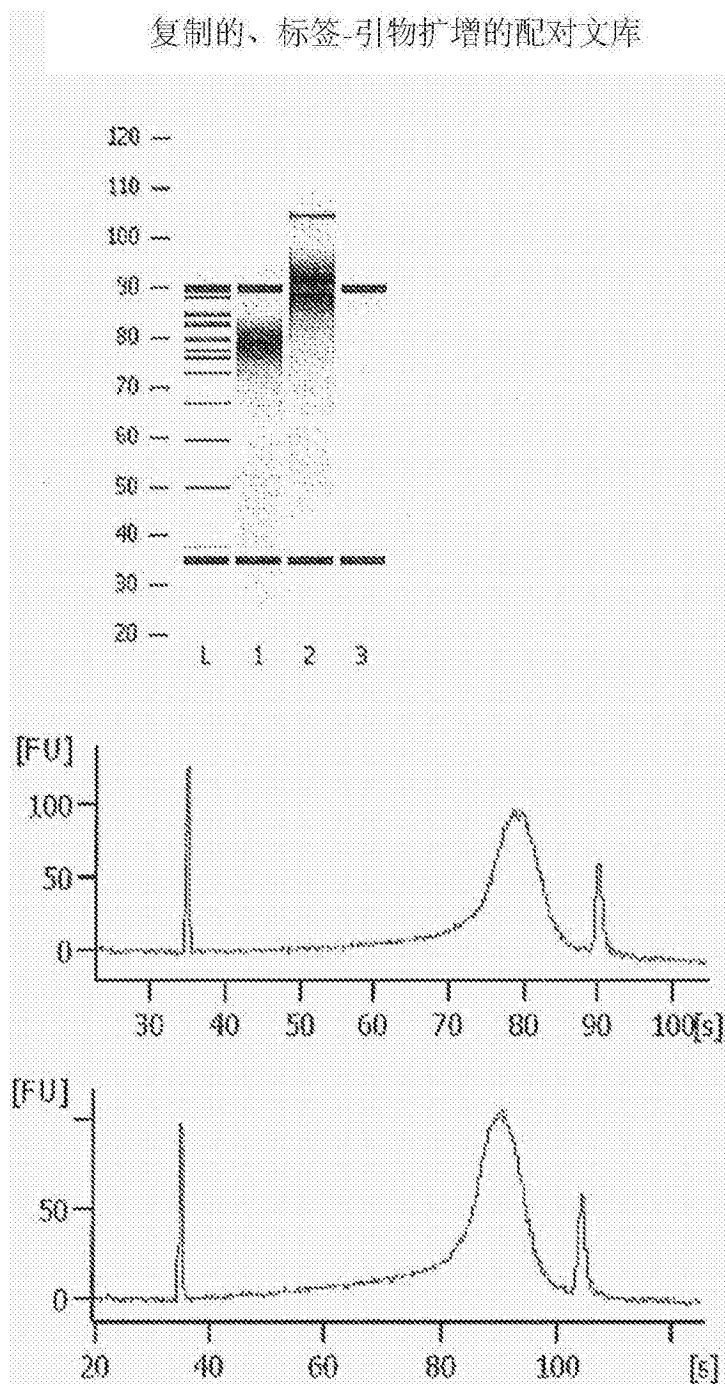


图7H



图7I

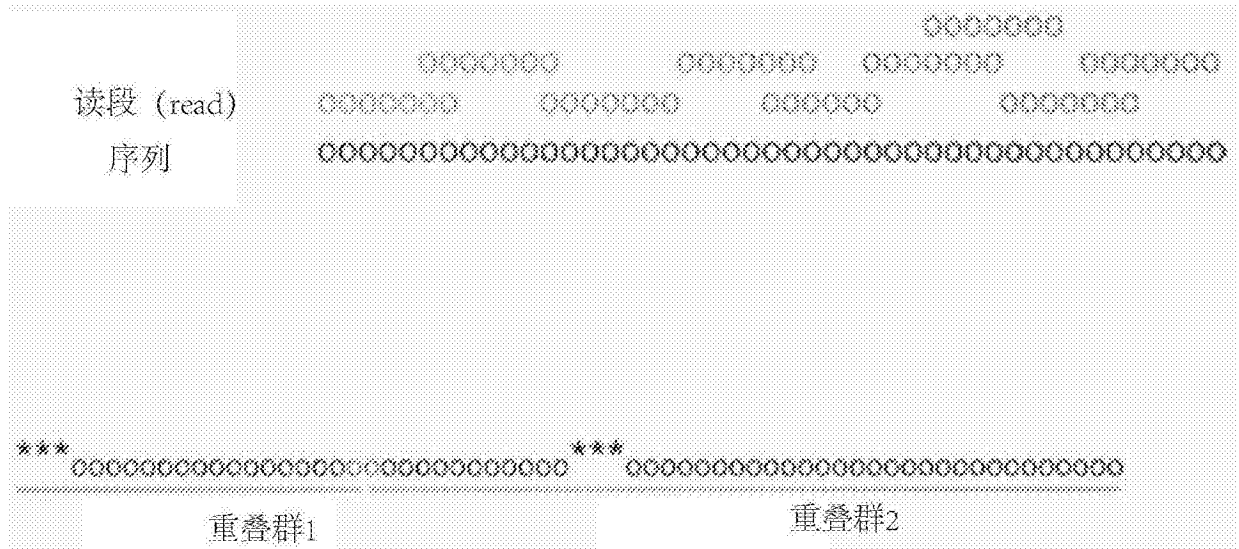


图8A

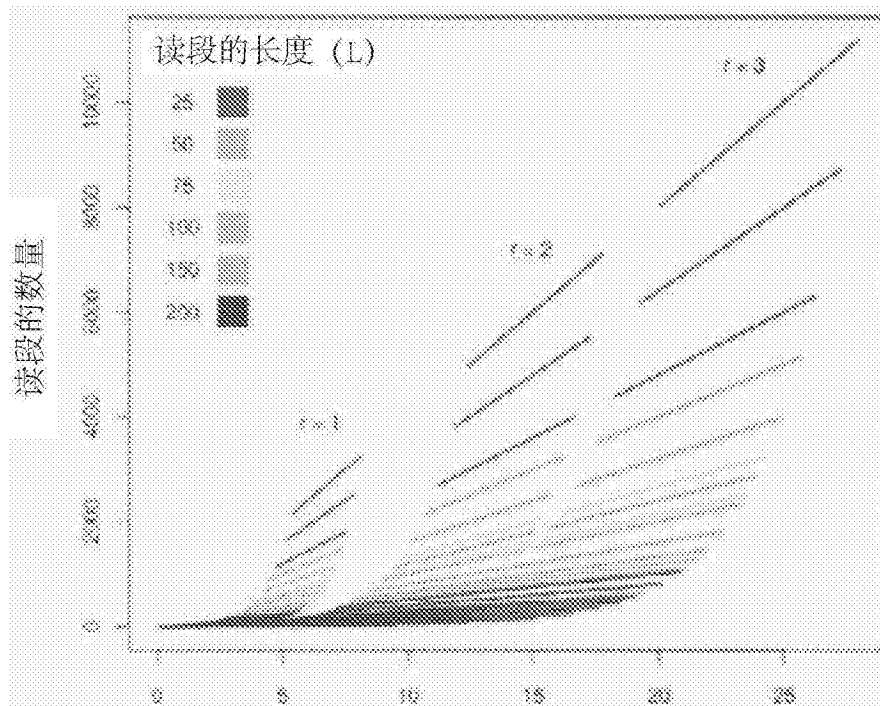


图8B

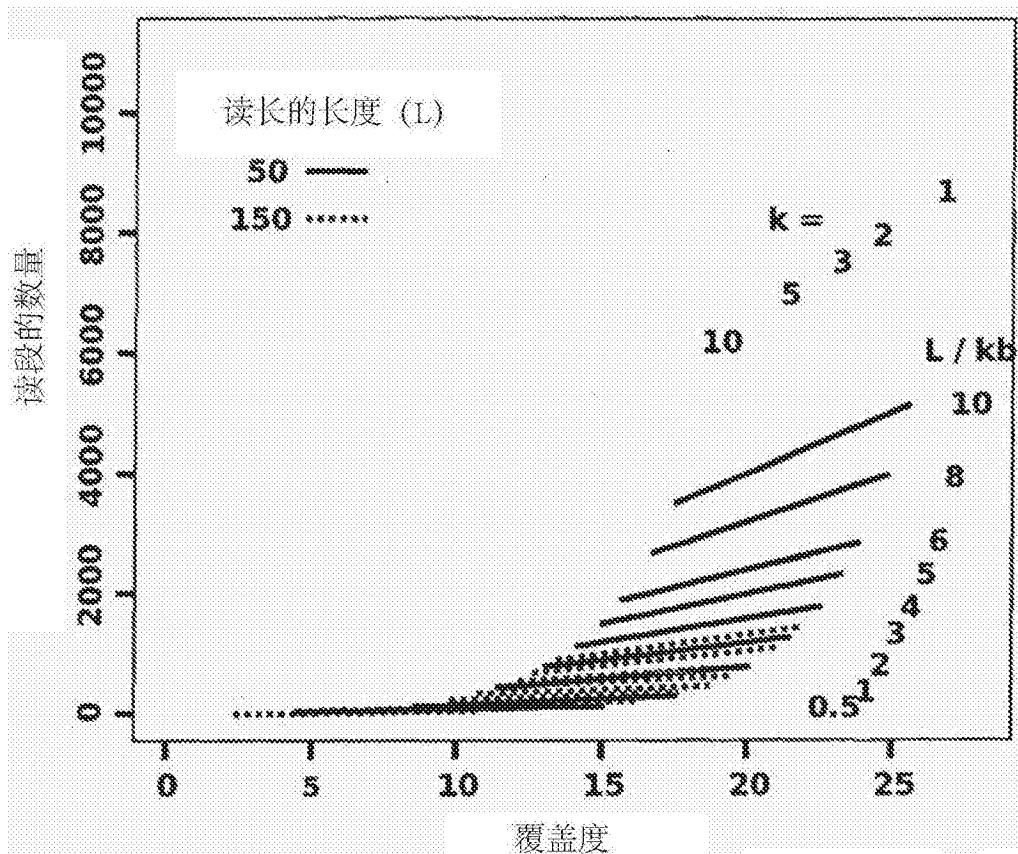


图8C

T	A(150)	A(200)	N(150)	N(200)	cvg(150)	cvg(200)
500	10	7.5	3.59	4.5	2.38	1.79
1,000	20	15	43.1	22.7	6.46	4.54
2,000	40	30	131	85	9.79	8.9
3,000	60	45	230	154	11.5	10.3
4,000	80	60	337	230	12.6	11.5
5,000	100	75	450	309	13.5	12.4
6,000	120	90	568	393	14.2	13.1
8,000	160	120	815	568	15.3	14.2
10,000	180	150	1074	752	16.1	15
对于所有的实例, $r=3$ $k=5$ $A = rT/kL$						

图8D

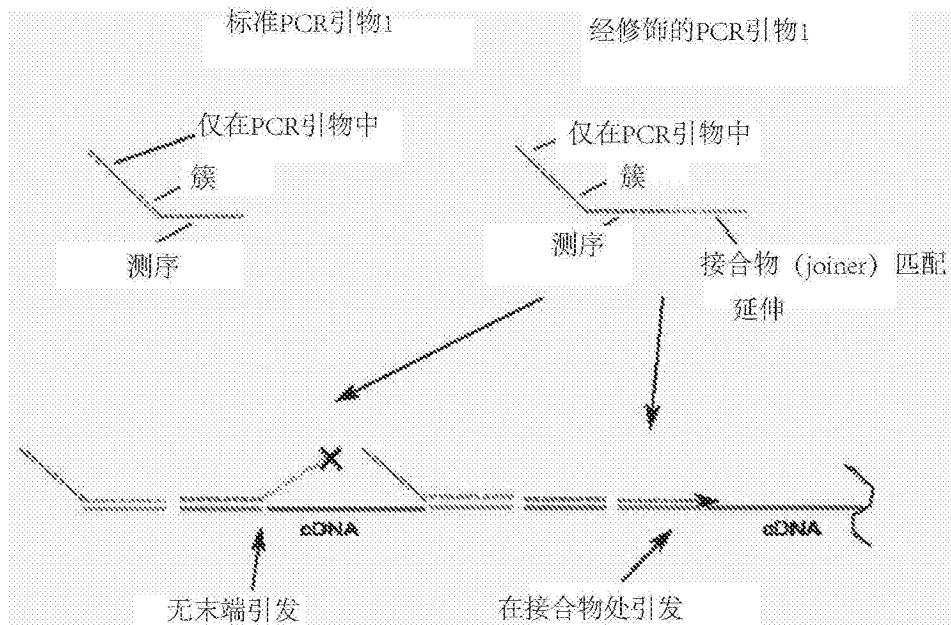


图9

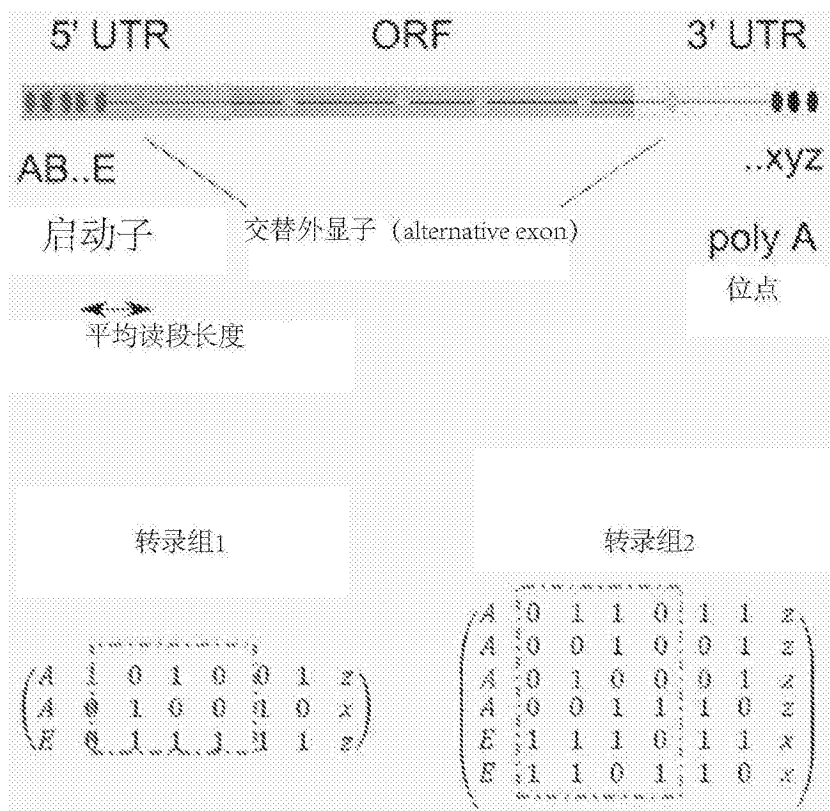


图10

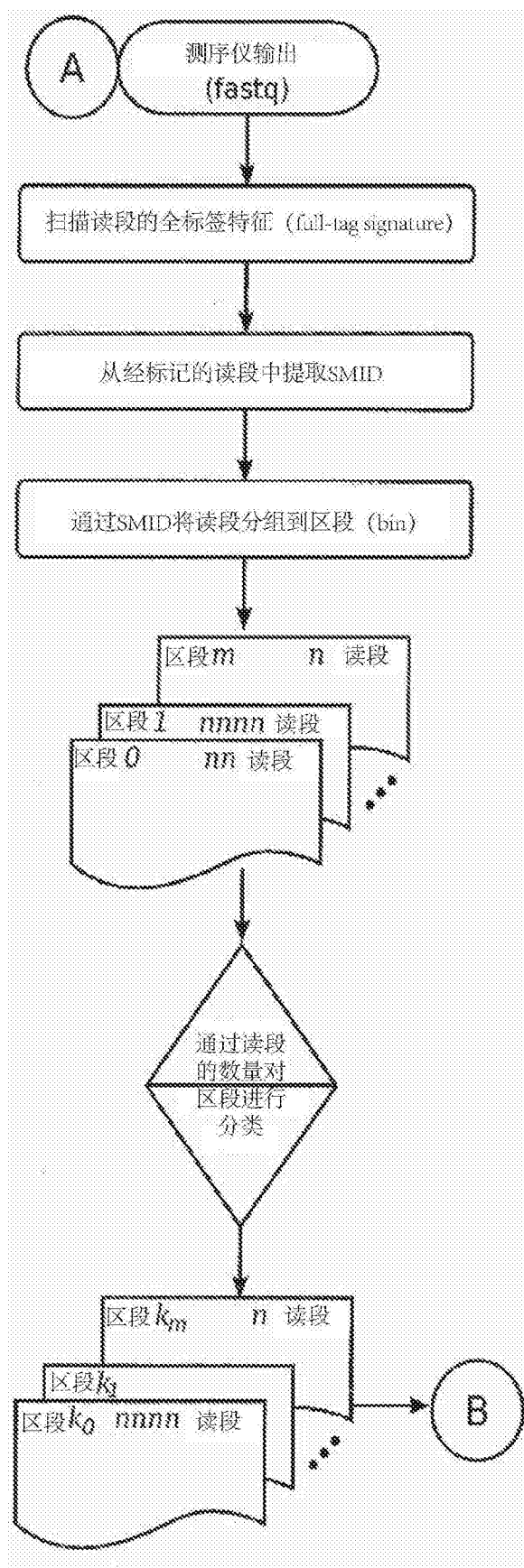


图11A

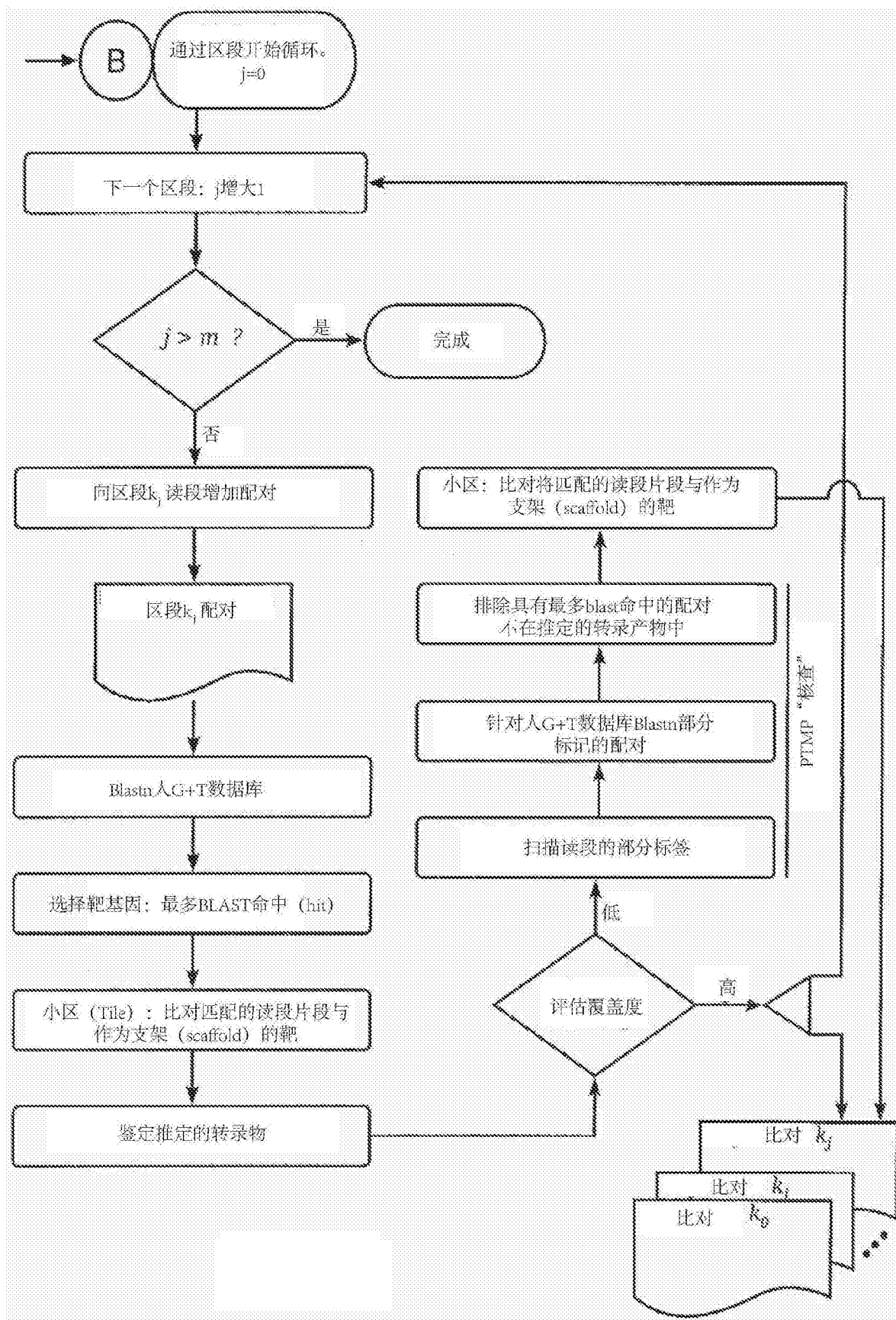


图11B