(19)中华人民共和国国家知识产权局

(12)发明专利

(10)授权公告号 CN 105658812 B
(45)授权公告日 2020.11.20

(21)申请号 201480048119.4
(22)申请日 2014.06.30
(65)同一申请的已公布的文献号
申请公布号 CN 105658812 A
(43)申请公布日 2016.06.08
(30)优先权数据
61/841,878 2013.07.01 US
62/001,580 2014.05.21 US
(85)PCT国际申请进入国家阶段日
2016.02.29
(86)PCT国际申请的申请数据
PCT/US2014/044971 2014.06.30
(87)PCT国际申请的公布数据
WO2015/002908 EN 2015.01.08
(73)专利权人 适应生物技术公司
地址 美国华盛顿州
(72)发明人 T·阿斯布瑞 K·赫特沃尔德
C·科特瓦利瓦勒 M·法哈姆
M·穆尔黑德 L·翁
T·威特科普 J·郑
(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.
C12Q 1/6874(2018.01)
C12Q 1/6806(2018.01)
C12Q 1/6886(2018.01)
(56)对比文件
CN 102272327 A,2011.12.07
US 7955794 B2,2011.06.07
WO 2013155119 A1,2013.10.17
WO 2013155119 A1,2013.10.17
审查员 杨佳倩

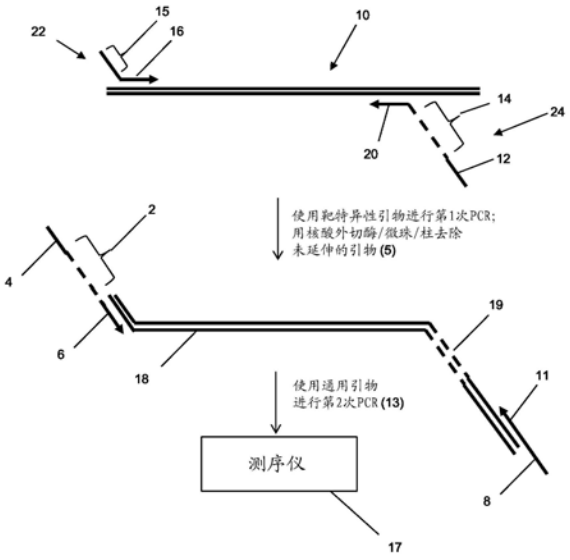
权利要求书3页 说明书47页
序列表3页 附图20页

(54)发明名称

用序列标签进行大规模生物分子分析

(57)摘要

本发明涉及通过对一个或多个序列标签进行多重扩增，并将所述序列标签附接至靶核酸和/或其拷贝，然后对所述扩增产物进行高通量测序，从而执对核酸群体进行基于序列的分析。在一些实施例中，本发明包括以下连续步骤：延伸引物、去除未延伸的引物和添加新的引物用于扩增(例如通过PCR)或用于额外的引物延伸。本发明的一些实施例涉及对接受癌症治疗的患者进行微小残留病灶(MRD)分析。将序列标签引入序列读段为测定克隆型提供了有效方法，同时为检测来自同一患者的其他样本的遗留污染或来自曾在同一实验室中曾检测的其他患者的样本的遗留污染提供了一种便利方法。



CN 105658812 B

1.一种由多个免疫细胞受体链产生克隆型谱的方法,所述方法包括以下步骤:

(a)在反应混合物中,使第一组引物在引物延伸条件下与来自B细胞和/或T细胞和/或无细胞DNA的重组核酸的样本混合,其中所述第一组引物中的每个引物均具有受体特异性部分,使得所述受体特异性部分在预定位置处与不同的重组核酸退火配对,然后延伸形成第一延伸产物,并且其中所述第一组引物中的每个引物均具有包含第一引物结合位点的5'非互补端;

(b)从所述反应混合物中去除所述第一组引物中的未延伸引物;

(c)在引物延伸条件下向所述反应混合物添加第二组引物,其中所述第二组引物中的每个引物均具有受体特异性部分,使得所述受体特异性部分在预定位置处与所述第一延伸产物退火配对并且具有包含第二引物结合位点的5'非互补端,所述第一组引物的引物包含位于所述受体特异性部分与所述第一引物结合位点之间的序列标签,并且其中所述第二组引物中的每个引物延伸形成第二延伸产物,使得每个第二延伸产物包含第一引物结合位点、第二引物结合位点、至少一个序列标签以及编码T细胞受体链或B细胞受体链的一部分的重组核酸;

(d)在所述反应混合物中进行聚合酶链反应以形成扩增子,所述聚合酶链反应使用特异于所述第一引物结合位点的正向引物和特异于所述第二引物结合位点的反向引物;以及

(e)对所述扩增子进行测序以形成多个T细胞受体链和/或B细胞受体链的克隆型谱。

2.根据权利要求1所述的方法,其中所述多个T细胞受体链包括TCRβ、TCRδ和TCRγ,并且其中所述第一组引物的引物和所述第二组引物的引物包括位于编码TCRβ和TCRδ的VDJ区的重组核酸区域旁侧的引物以及位于TCRγ的VJ区旁侧的引物。

3.根据权利要求1所述的方法,其中所述测序步骤包括:

产生各自具有错误率并且各自包含序列标签和重组核酸序列的序列读段;

比对具有类似序列标签的序列读段,以形成具有相同序列标签的序列读段组;

合并各组的序列读段以确定克隆型,其中每当所述序列读段组有至少95%的可能性是不同的时,则将所述序列读段组合并成不同的重组核酸序列;以及

由所述克隆型形成所述克隆型谱。

4.根据权利要求1所述的方法,所述方法还包括在形成所述第二延伸产物后从所述反应混合物中去除所述第二组引物中的未延伸引物的步骤。

5.根据权利要求1所述的方法,其中在解链所述第一延伸产物后,重复进行所述第一组引物的引物的退火和延伸。

6.根据权利要求1所述的方法,其中在解链所述第二延伸产物后,重复进行所述第二组引物的引物的退火和延伸。

7.根据权利要求1或4所述的方法,其中所述去除步骤包括核酸外切酶消化。

8.根据权利要求1或4所述的方法,其中所述去除步骤包括核酸外切酶I消化。

9.根据权利要求1所述的方法,其中所述多个B细胞受体链包括IgH和IgK,并且其中所述第一组引物的引物和所述第二组引物的引物包括位于编码IgH的VDJ区、IgH的DJ区和IgK的VJ区的重组核酸区域旁侧的引物。

10.一种用于确定样本中编码多个免疫受体链的重组核酸的克隆型谱的方法,所述方法包括以下步骤:

(a)将序列标签附接到来自包含T细胞和/或B细胞和/或无细胞DNA的个体样本的T细胞受体基因或免疫球蛋白基因的重组核酸分子,以形成标签-核酸缀合物,其中至少一个标签-核酸缀合物或所述标签-核酸缀合物的拷贝附接有不同的序列标签;

(b)扩增所述标签-核酸缀合物;

(c)对所述标签-核酸缀合物的样本测序以提供序列读段,所述序列读段各自具有错误率并且各自包含序列标签和重组核酸序列;

(d)比对具有类似序列标签的序列读段,以形成具有类似序列标签的序列读段组;

(e)合并各组的序列读段以确定克隆型,其中每当所述序列读段组有至少95%的可能性是不同的时,则将所述序列读段组合并成不同的重组核酸序列;以及

(f)通过确定所述克隆型的水平来确定所述样本的所述克隆型谱;

其中所述附接和扩增步骤包括:

在反应混合物中,使第一组引物在引物延伸条件下与所述样本混合,其中所述第一组引物中的每个引物均具有受体特异性部分,使得所述受体特异性部分在预定位置处与不同的重组核酸退火配对,然后延伸形成第一延伸产物,并且其中所述第一组引物中的每个引物均具有包含第一引物结合位点的5'非互补端;

从所述反应混合物中去除所述第一组引物中的未延伸引物;

在引物延伸条件下向所述反应混合物添加第二组引物,其中所述第二组引物中的每个引物均具有受体特异性部分,使得所述受体特异性部分在预定位置处与所述第一延伸产物退火配对并且具有包含第二引物结合位点的5'非互补端,所述第一组引物的引物包含位于所述受体特异性部分与所述第一引物结合位点之间的序列标签,并且其中所述第二组引物中的每个引物延伸形成第二延伸产物,使得每个第二延伸产物包含第一引物结合位点、第二引物结合位点、至少一个序列标签以及编码T细胞受体链或B细胞受体链的一部分的重组核酸;以及

在所述反应混合物中进行聚合酶链反应以形成扩增子,所述聚合酶链反应使用特异于所述第一引物结合位点的正向引物和特异于所述第二引物结合位点的反向引物。

11.根据权利要求10所述的方法,所述方法还包括在形成所述第二延伸产物后从所述反应混合物中去除所述第二组引物中的未延伸引物的步骤。

12.根据权利要求10所述的方法,其中在解链所述第一延伸产物后,重复进行所述第一组引物的引物的退火和延伸。

13.根据权利要求10所述的方法,其中在解链所述第二延伸产物后,重复进行所述第二组引物的引物的退火和延伸。

14.根据权利要求10或11所述的方法,其中所述去除步骤包括核酸外切酶消化。

15.根据权利要求10或11所述的方法,其中所述去除步骤包括核酸外切酶I消化。

16.根据权利要求10所述的方法,其中所述附接和扩增步骤将生成标签-核酸缀合物,其中至少一个重组核酸及其相应的拷贝附接有多个不同的序列标签。

17.根据权利要求16所述的方法,其中所述重组核酸包括编码TCRβ、TCRδ和TCRγ链的重组核酸,并且其中所述第一组引物的引物和所述第二组引物的引物包括位于编码TCRβ和TCRδ的VDJ区的重组核酸区域旁侧的引物以及位于TCRγ的VJ区旁侧的引物。

18.根据权利要求16所述的方法,其中所述重组核酸包括编码IgH和IgK的重组核酸,并

且其中所述第一组引物的引物和所述第二组引物的引物包括位于编码IgH的VDJ区、IgH的DJ区和IgK的VJ区的重组核酸区域旁侧的引物。

19.根据权利要求10所述的方法，其中所述合并步骤包括，每当所述序列读段组有至少99.9%的可能性是不同的时，则将所述序列读段组合并成不同的重组核酸序列。

20.根据权利要求1所述的方法，其中所述序列标签是镶嵌标签，其中所述镶嵌标签包含交替的恒定区与可变区。

21.根据权利要求1所述的方法，其中所述正向引物和/或所述反向引物包含含有样本标签的5’非互补端。

22.根据权利要求1所述的方法，其中所述第一组引物的引物的受体特异性部分对V区具有特异性。

23.根据权利要求22所述的方法，其中所述第一组引物的每个引物的受体特异性部分均与所述V区中的不同非重叠位置退火配对。

24.根据权利要求1所述的方法，其中所述第一延伸产物的预定位置是J区或C区。

25.根据权利要求24所述的方法，其中所述第二组引物的每个引物的受体特异性部分均与所述J区中的不同非重叠位置退火配对。

26.根据权利要求24所述的方法，其中所述第二组引物的每个引物的受体特异性部分均与所述J区中的单一引物结合位点退火配对。

27.根据权利要求10所述的方法，其中所述序列标签是镶嵌标签，其中所述镶嵌标签包含交替的恒定区与可变区。

28.根据权利要求10所述的方法，其中所述正向引物和/或所述反向引物包含含有样本标签的5’非互补端。

29.根据权利要求10所述的方法，其中所述第一组引物的引物的受体特异性部分对V区具有特异性。

30.根据权利要求29所述的方法，其中所述第一组引物的每个引物的受体特异性部分均与所述V区中的不同非重叠位置退火配对。

31.根据权利要求10所述的方法，其中所述第一延伸产物的预定位置是J区或C区。

32.根据权利要求31所述的方法，其中所述第二组引物的每个引物的受体特异性部分均与所述J区中的不同非重叠位置退火配对。

33.根据权利要求31所述的方法，其中所述第二组引物的每个引物的受体特异性部分均与所述J区中的单一引物结合位点退火配对。

用序列标签进行大规模生物分子分析

[0001] 交叉引用

[0002] 本专利申请要求于2013年7月1日提交的美国临时专利申请No.61/841,878以及于2014年5月21日提交的美国临时专利申请No.62/001,580的优先权,两者均以引用方式全文并入本文。

[0003] 序列表

[0004] 本申请包含序列表,该序列表以ASCII格式通过EFS-Web提交并且据此全文引入以供参考。所述ASCII副本创建于2014年6月27日,名称为848US00-SL-ST25.txt,文件大小为2千字节。没有添加新客体。

背景技术

[0005] 由于大规模DNA测序的速度和便利性提高,并且其每个碱基的成本降低,所以大规模DNA测序在诊断和预后中的应用迅速扩展,例如Ding et al,Nature,481(7382):506-510(2012)(Ding等人,《自然》,2012年,第481卷,第7382期,第506-510页);Chiu et al,Brit.Med.J.,342:c7401(2011)(Chiu等人,《英国医学期刊》,2011年,第342卷,第c7401页);Ku et al,Annals of Neurology,71(1):5-14(2012)(Ku等人,《神经学年鉴》,2012年,第71卷,第1期,第5-14页)等等。具体地讲,编码免疫分子(如T细胞或B细胞受体,或者它们的组分)的核酸谱包含大量有关生物体健康状态或疾病的信息,因而已提出将此类谱用作多种病症的诊断或预后指标,例如Faham and Willis,U.S.patents 8,236,503and 8,628927(Faham和Willis,美国专利8,236,503和8,628927);Freeman et al,Genome Research,19:1817-1824(2009)(Freeman等人,《基因组研究》,2009年,第19卷,第1817-1824页);Han et al,J.Immunol.,182(1001):42.6(2009)(Han等人,《免疫学杂志》,2009年,第182卷,第1001期,第42.6页);Boyd et al,Sci.Transl Med.,1(12):12ra23(2009)(Boyd等人,《科学转化医学》,2009年,第1卷,第12期,第12-23页);He et al,Oncotarget(March 8,2011)(He等人,《肿瘤标靶》,2011年3月8日)。

[0006] 举例来说,许多经治疗的癌症患者常常留有与癌症有关的微小残留病灶(MRD)。也就是说,即使可通过临床措施使患者的疾病响应治疗而得以完全缓解,但由于这样或那样的原因,一小部分癌细胞可能未被消灭而保留下来。对于患者的继续治疗而言,这种残留群体的类型和大小是重要的预后因子,例如Campana,Hematol.Oncol.Clin.North Am.,23(5):1083-1098(2009)(Campana,《北美血液学与肿瘤学临床》,2009年,第23卷,第5期,第1083-1098页);Buccisano et al,Blood,119(2):332-341(2012)(Buccisano等人,《血液》,2012年,第119卷,第2期,第332-341页)。因此,人们已开发了若干用于评估这种群体的技术,包括基于流式细胞术、原位杂交、细胞遗传学、核酸标记扩增等的技术,例如Buccisano et al,Current Opinion in Oncology,21:582-588(2009)(Buccisano等人,《肿瘤学新见》,2009年,第21卷,第582-588页);van Dongen et al,Leukemia,17(12):2257-2317(2003)(van Dongen等人,《白血病》,2003年,第17卷,第12期,第2257-2317页)等等。从T细胞和/或B细胞扩增编码免疫受体(即克隆型)区段的重组核酸,对于评估白血病和淋巴瘤

MRD特别有用，因为此类克隆型通常具有独特的序列，这些独特的序列可用作这些克隆型的相关癌细胞的分子标签。由于此类扩增是高度多重化的而难以建立，所以通常通过对编码单一受体链的核酸进行扩增和测序来部分地进行此类测定。随着多重扩增规模增大而遇到了若干问题，包括：由错误杂交引起的假性扩增的概率增加、形成引物二聚体、变化的扩增率导致序列表征偏倚等等，例如Elnifro et al,Clinical Microbiology Reviews,13(4):559-570(2000)(Elnifro等人，《临床微生物学评论》，2000年，第13卷，第4期，第559-570页)。此外，无论对于序列分析、样本跟踪、污染检测，还是诸如此类的工作，靶序列的相似性以及将序列标签掺入扩增序列均会加剧上述与大规模扩增有关的困难。这些困难阻碍了多免疫受体链的大规模单反应扩增的发展，而多免疫受体链的大规模单反应扩增将非常有利于减少测定微小病灶相关核酸序列所需的单独测定数。

[0007] 鉴于上文所述，如果存在更有效的方法用于评估在单一反应中选定的核酸，如癌基因的外显子或编码免疫受体链组的克隆型，那将会非常有利。

发明内容

[0008] 本发明涉及如下方法：在单一反应中对靶多核苷酸(如编码免疫受体链的重组核酸)群体进行大规模扩增，特别是通过聚合酶链反应(PCR)进行大规模扩增，随后采用大规模DNA测序对其进行鉴定。本发明包括将上述方法应用于监测癌症的微小残留病灶。在多个具体实施和应用中示例了本发明，其中一些汇总如下并贯穿于整个说明书中。

[0009] 在一些实施例中，本发明涉及生成核酸谱的方法，所述核酸编码所关注的生物分子(如免疫受体分子)群体。在一个方面，本发明的方法包括：将序列标签附接到样本中选定的核酸群体以形成标签-核酸缀合物；扩增该标签-核酸缀合物；以及对扩增的标签-核酸缀合物进行测序，以提供序列读段，每条序列读段均包含标签序列和核酸序列，针对这些序列生成核酸谱。在一些实施例中，序列标签的附接通过一个或多个连续的引物延伸和引物去除步骤实现，然后可通过通用的正向引物和反向引物无偏倚地进一步扩增所得产物。在一些实施例中，本发明涉及一些方法，用于检测和测定样本中由源于不同样本的物质所带来的污染(如遗留污染)。

[0010] 在一个实施例中，此类用于检测受到微小残留病灶监测的个体中污染的方法可包括以下步骤：(a)从个体获得组织样本；(b)将序列标签附接到癌基因分子或重组核酸形成标签-核酸缀合物，其中至少一个核酸或其拷贝所附接的序列标签不同，并且其中癌基因分子为该个体的癌症所特有的；(c)扩增该标签-核酸缀合物；(d)对该标签-核酸缀合物的样本进行测序，以提供序列读段，所述序列读段具有错误率并且包含标签序列和癌基因序列或重组核酸序列；(e)将标签序列与分别从其他组织样本确定的标签序列进行比较；以及(f)通过一条或多条标签序列与分别从其他组织样本确定的任意标签序列的同一性，确定是否存在污染和/或污染程度。

[0011] 在另一方面，本发明涉及上述的用于基于B细胞受体的至少两条链生成克隆型谱的方法，该方法包括在单一反应中扩增靶核酸，所述靶核酸编码B细胞受体的两条或更多条链。在另一方面，此类方法用于监测B细胞癌中的微小残留病灶。

[0012] 在另一方面，本发明涉及上述的基于T细胞受体的至少两条链生成克隆型谱的方法，该方法包括在单一反应中扩增靶核酸，所述靶核酸编码T细胞受体的两条或更多条链。

在另一方面，此类方法用于监测T细胞癌中的微小残留病灶。

[0013] 在多个示出的具体实施和应用中示例了本发明的上述这些特征方面以及其他方面，其中一些以图示出，并在下面的权利要求部分里进行了表征。然而，上述发明内容并非意图描述本发明的每个图示实施例或每种实施方式。

附图说明

[0014] 在所附权利要求中特别阐明了本发明的新颖特征。参考以下示出了示例性实施例的详细描述以及附图可更好地理解本发明的特征和优势，在这些示例性实施例中利用了本发明的原理，所述附图包括：

[0015] 图1A至图1C以图解方式示出了本发明的多种实施例。图1D示出了(使用或不使用序列标签)生成具有预定长度的重组核酸的模板的方法。

[0016] 图2A至图2G示出了为样本中基本上每条靶序列附接独特的序列标签的多种方法。

[0017] 图3A和图3B以图解方式示出了本发明的一个方面：用于从编码IgH链的核酸序列生成克隆型谱。

[0018] 图4A示出了序列标签在确定序列读段的克隆型序列中的用途。图4B示出了序列标签在多个不同序列标签附接到相同靶多核苷酸或其拷贝的实施例中的用途。

[0019] 图5A示出了序列空间中的克隆型以及紧密相关的克隆型之间距离的概念。

[0020] 图5B为流程图，示出了从仅因测序误差而有所不同的克隆型(其应被合并)中辨别出真正不同的克隆型的方法的一个实施例。

[0021] 图5C示出了一种数值函数的形式，该数值函数在一个实施例中用于确定是否合并相关的克隆型。

[0022] 图5D和图5E示出了序列树在合并序列读段的方法中的用途。

具体实施方式

[0023] 除非另外指明，否则本发明的实施可采用分子生物学(包括重组技术)、生物信息学、细胞生物学和生物化学的常规技术和描述，这些常规技术和描述在本领域的技术范围内。这类常规技术包括但不限于血细胞的采样和分析、核酸测序和分析等等。可参考下文中的例子获得合适技术的具体示例。然而，当然也可使用其他等效的常规方法。此类常规技术和描述可见于标准实验室手册，如Genome Analysis:A Laboratory Manual Series (Vols.I-IV)(《基因组分析：实验室手册系列》(I-IV卷))；PCR Primer:A Laboratory Manual(《PCR引物：实验室手册》)；以及Molecular Cloning:A Laboratory Manual(《分子克隆：实验室手册》)(均由冷泉港实验室出版社出版)等等。

[0024] 在一个方面，本发明涉及产生多条免疫受体链的克隆型谱的方法，所述方法通过对编码此类链的核酸进行大规模多重扩增，随后对扩增产物(或者说扩增子)进行高通量测序。在一些实施例中，本发明通过包括一些连续步骤而克服了多重扩增的常见缺陷，所述连续步骤包括：延伸引物、去除未延伸或未掺入的引物，以及添加新的引物用于(例如，通过PCR)扩增或用于另外的引物延伸。这些步骤也使得能够使用序列标签，否则不使用所述序列标签将促成非特异性扩增或假性扩增。在另一方面，在临床应用的实施例中采用了序列标签，特别是(例如，对来自正接受癌症治疗的患者的样本)进行微小残留病灶(MRD)分析。

掺入到序列读段中的序列标签提供了确定克隆型的有效手段，同时还提供了通过由前述测定检测序列标签是否存在来检测遗留污染的便利手段，所述序列标签可来自同一患者的样本，也可来自在同一实验室中测试的不同患者的样本。特别值得关注的是如下方法：其使用单一扩增反应为编码多条B细胞受体(BCR)链的重组核酸生成基于序列的克隆型谱，然后进行新一代的高通量测序。还特别值得关注的是如下方法：其使用单一扩增反应为编码多条T细胞受体(TCR)链的重组核酸生成基于序列的克隆型谱，然后进行新一代的高通量测序。本发明方法也可应用于其他所关注的核酸组(包括，例如，癌基因的外显子组)的其他大规模扩增和测序。在这些方面，利用序列标签既可监测遗留污染，又可基于易错测序方法更灵敏地确定靶多核苷酸的核苷酸序列。另外，在这些方面，一组序列标签(如下文中更充分讨论的)通常比样本中靶多核苷酸的数量大得多，并且附接于靶多核苷酸的序列标签之间的序列差异足够大，从而能有效地避免一个标签序列因测序误差而转变成另一个标签序列。

[0025] 本发明的一个实施例如图1A所示。在反应混合物中，来自第一组(第一组的每个引物具有受体特异性部分16和包含第一引物结合位点的5’非互补部分15)的引物22(在靶多核苷酸10解链后)与靶多核苷酸10的一端退火配对，而来自第二组(第二组的每个引物具有受体特异性部分20和5’非互补部分，该5’非互补部分包含序列标签14和第二引物结合位点12)的引物24与靶多核苷酸10的另一端退火配对。在一些实施例中，如下文所示，引物22的非互补部分15也可包含序列标签。在一些情况下，两个更短的序列标签可能比多样性等同的单个更长的序列标签更有利。因此，例如，两个8聚随机核苷酸序列标签可能比单个16聚随机核苷酸序列标签更不易造成假性启动、引物二聚体等等。靶多核苷酸10通常为来自T细胞或B细胞的体细胞重组核酸，其编码T细胞受体(TCR)或B细胞受体的链或者链的部分(例如，IgH链或IgK链的部分)。因此，在一些实施例中，引物22和24的受体特异性部分可分别对V区序列和J区序列具有特异性，或者，在其他实施例中，反之亦然。

[0026] 在一些实施例中，靶多核苷酸10可包含所需序列谱的核酸的复杂混合物，所述核酸的复杂混合物包括但不限于：编码免疫受体分子的部分的重组核酸；微生物群落的16S rDNA；编码在工业或医疗上具有重要性的蛋白质(诸如酶)的基因、与特定疾病(如癌症、传染病等)有关的人类或动物的基因和/或外显子的宏基因组扩增。在涉及编码免疫受体的重组核酸的实施例中，通常V区、D区或J区的至少部分存在于第一组引物和第二组引物的两个结合位置之间。在一些实施例中，在第一组引物和第二组引物的两个结合位置之间存在IgH的VDJ重排、IgH的DJ重排、IgK的VJ重排、IgL的VJ重排、TCRβ的VDJ重排、TCRβ的DJ重排、TCRα的VJ重排、TCRγ的VJ重排、TCRδ的VDJ重排或TCRδ的VD重排的至少一部分。在一些实施例中，在第一组引物和第二组引物的两个结合位置之间存在IgH的VDJ重排、IgH的DJ重排、IgK的VJ重排或IgL的VJ重排的至少一部分。在一些实施例中，在第一组引物和第二组引物的两个结合位置之间存在TCRβ的VDJ重排、TCRβ的DJ重排、TCRα的VJ重排、TCRγ的VJ重排、TCRδ的VDJ重排或TCRδ的VD重排的至少一部分。在其他实施例中，在第一组引物和第二组引物的两个结合位置之间存在IgH的VDJ重排、IgH的DJ重排和IgK的VJ重排的至少一部分。并且，在其他实施例中，在第一组引物和第二组引物的两个结合位置之间存在TCRβ的VDJ重排、TCRγ的VJ重排和TCRδ的VDJ重排或TCRδ的VD重排的至少一部分。在一些实施例中，VDJ重排的至少一部分包含能够满足其鉴定需求的完整的D或NDN部分以及V区段和J区段的一部分。在一些实施例中，VDJ重排的至少一部分包含至少50个核苷酸的区段，该50个核苷酸的区段包

含完整的D或NDN部分以及V区段和J区段的一部分。在一些实施例中，VDJ重排的至少一部分包含至少70个核苷酸的区段，该70个核苷酸的区段包含完整的D或NDN部分以及V区段和J区段的一部分。

[0027] 在一些实施例中，第一组包括一个或多个引物，每个引物对J区段或C区段具有特异性。来自此类第一组的引物与其靶序列退火配对，然后延伸，之后去除第一组中未延伸的引物。来自第二组的各自对V区段具有特异性的引物与其靶序列退火配对，然后延伸。在其他实施例中，第一组包括各自对V区段具有特异性的引物，此类第一组的引物与其靶序列退火配对，然后延伸，之后去除第一组中未延伸的引物；第二组中各自对J区段或C区段具有特异性的引物与其靶序列退火配对，然后延伸。在这两种实施例的替代形式中，第一组和第二组可各自包括多个引物，并且每个引物可对不同的免疫受体区段具有特异性。

[0028] 回到图1A，在一些实施例中，通过在可供选择的实施例中进行1-10或2-10或3-10或4-10或5-10个循环的解链、退火和延伸，然后使用常规技术将未延伸的引物从反应混合物中去除，来延伸第一组和第二组的引物5。在其他实施例中，通过在可供选择的实施例中进行2-5或3-5或4-5个循环的解链、退火和延伸，然后使用常规技术将未延伸的引物从反应混合物中去除，来延伸第一组和第二组的引物5。在另一个实施例中，通过进行两个循环的解链、退火和延伸来延伸第一组和第二组的引物。举例来说，可利用核酸外切酶酶切、与磁珠上的互补序列杂交、尺寸排阻色谱法、市售离心柱(如Qiagen QIAquick PCR纯化试剂盒(Qiagen QIAquick PCR Purification Kit))等去除未延伸的引物。在一个实施例中，通过(例如)使用核酸外切酶I酶切来去除未延伸的或未掺入的引物。双链DNA 18为延伸的产物5，其在各端具有通用的第一和第二引物结合位点，(在一些实施例中)可将具有互补序列6和11的正向引物和反向引物添加到所述结合位点上，用于稍后通过桥式PCR生成簇。在一些实施例中，双链DNA还具有序列标签19，并且正向或反向引物可包含样本标签2，用于鉴定或跟踪DNA 18，或者将其与样本或患者相关联。在一些实施例中，序列标签19基本上对于样本中每个不同的重组核酸都是独特的。正如将在下文中更充分说明的，序列标签19可用于将序列读段合并成克隆型，还可用于检测和跟踪样本污染。正向引物和反向引物还可包含引物结合位点4和8，用于针对某些测序方案在(例如)基因组分析仪17(Genome Analyzer)(圣地亚哥亿明达公司(Illumina,San Diego))上进行(13)桥式PCR。在用包含序列标签的引物进行了不止一次延伸的其他实施例中，样本中每个不同的重组核酸可具有拷贝，所述拷贝附接有不同的序列标签；因此，例如，如果根据图1A所示实施例对靶多核苷酸进行四个独立的解链、退火和延伸循环，并且如果样本包含重组核酸S_1，那么在用通用引物完成扩增(13)时，S_1的拷贝将最多具有四个不同的序列标签。因此，S_1的序列读段将最多具有四个不同的序列标签。正如将在下文中更充分说明的，在此类实施例中，可在由通用序列标签定义的各子集内通过比对序列标签与合并序列读段相结合的方式来确定克隆型。

[0029] 在另一个实施例中，在用通用引物进行PCR之前进行了至少两次延伸和两个去除未掺入引物的步骤。如图1B所示，引物101与靶多核苷酸100(如编码免疫受体链的重组核酸)的一端退火配对，然后(例如)用DNA聚合酶延伸。引物101可各自包括受体特异性部分103和5’非互补部分105，5’非互补部分105又包含序列标签104和第一引物结合位点102。如上所述进行延伸并去除未掺入的引物130后，向反应混合物中的第一延伸产物109添加了：(a)引物125，其中每个引物包含受体特异性部分106和5’非互补部分115(该部分包含引物

结合位点);以及(b)引物127,其包含对第一引物结合位点102和5'非互补部分117具有特异性的部分108。引物125和127与其引物结合位点退火配对后,使这些引物延伸(107)形成第二延伸产物118,然后去除未延伸的引物。向第二延伸产物118添加通用的正向引物112和反向引物110并进行PCR(111),然后对所得扩增子进行测序(120)。如上文结合图1A的实施例所述,每当在包含序列标签的引物(如101)存在的情况下进行不止一次延伸步骤时,同一靶多核苷酸100的拷贝就可被多个不同的序列标签标记。

[0030] 图1C示出了明确显示出V区、D区和J区的另一个实施例。在引物退火条件下的反应混合物中,向编码免疫受体(如TCR)的重组核酸1200添加引物1212(即对V区1226具有特异性的第一组引物)。第一组1212的每个引物均包含受体特异性部分和5'非互补部分,5'非互补部分又任选地包含序列标签和第一引物结合位点(如图1B所示的102、103和104)。第一组引物1212与重组核酸1200的V区1226退火配对,然后第一组引物1212穿过D区1224延伸(1202)到至少J区1222,并任选地延伸到C区1220,形成包含任选序列标签1228和第一引物结合位点1230的第一延伸产物1216。去除第一组1212的未延伸引物后,在退火条件下向反应混合物中添加第二组引物1240,以使这些引物与其相应的靶J区1222退火配对,然后使这些引物延伸(1204)形成第二延伸产物1232,所述第二延伸产物的每一个均包含序列标签1236(任选)和第二引物结合位点1234。第二延伸产物1232可包含单一序列标签,所述单一序列标签位于(例如)如序列标签1228所示的V区1226附近,或者如序列标签1236所示的J区1222附近;或者第二延伸产物1232可同时包含位于这两个位置的两个序列标签。在一个实施例中,第二延伸产物1232包含邻近V区1226的单一序列标签1228。在另一个实施例中,第二延伸产物1232包含邻近J区1222的单一序列标签1236。在一些实施例中,序列标签1228和/或1236为下文所述的镶嵌标签。去除第二组1240的未延伸引物后,添加分别对第一引物结合位点1230和第二引物结合位点1234具有特异性的通用正向引物和反向引物,然后进行PCR(1206)。对所得扩增子的样本进行测序(1208),生成用于构建克隆型和克隆型谱的序列读段。

[0031] 图1D示出了生成具有限定长度的模板并将一个或两个序列标签附接到其上的方法。图1D的实施例将信使RNA(mRNA)示作起始物,但该方法可适用于DNA或RNA样本中的任一者。对于包含VDJ区的mRNA 1300,一个或多个对C区1308具有特异性的引物1312("C引物")与mRNA 1300退火配对。通常只使用单个C引物。作为另外一种选择,可使用一个或多个(具有相似结构)对J区具有特异性的引物。C引物1312包含靶特异性区段1313、序列标签区段1314和通用引物结合位点1315。聚合酶阻断剂1310也与靶mRNA 1300退火配对,所述聚合酶阻断剂可为对V区1302具有特异性的寡核苷酸。在一些实施例中,只要用于延伸引物1312的聚合酶没有链置换活性或5'→3'外切酶活性,并且寡核苷酸不可延伸(例如,其具有3'-双脱氧核苷酸),则阻断剂1310可为天然寡核苷酸。通常,阻断剂1310为具有增强的结合活性和核酸酶抗性的寡核苷酸类似物,如反义化合物。在一些实施例中,阻断剂1310可为锁核酸(LNA)或肽核酸(PNA)或桥核酸(BNA),这些阻断剂公开于以下参考文献中:Wengel et al,U.S.patents 6,794,499;7,572,582(Wengel等人,美国专利6,794,499、7,572,582);Vester et al,Biochemistry,43(42):13233-13241(2004)(Vester等人,《生物化学》,2004年,第43卷,第42期,第13233-13241页)等等;以及Kazuyuki et al,Chem.Comm.,3765-3767(2007)(Kazuyuki等人,《化学通讯》,2007年,第3765-3767页);Nielson et al,

Chem.Soc.Rev.,26:73-78(1997)(Nielson等人,《化学会评论》,1997年,第26卷,第73-78页)等等。对阻断剂1310的序列进行选择,使得引物1312的延伸在V区1302上的预定位置停止。在一些实施例中,阻断剂1310经设计使得在延伸步骤中只复制足够的V区1302,从而可从复制的序列中识别出V区。在一些实施例中,由于只要能使聚合酶停止前进,可选择允许一些错配发生的共有序列,所以没必要获得对各个V区均具有特异性的阻断剂1310。阻断剂1310的长度可根据所使用的寡核苷酸或寡核苷酸类似物的种类而大幅变化。在一些实施例中,阻断剂1310的长度在10至25个单体的范围内。在一些实施例中,阻断剂1310可与不同V区序列上的不同位置退火配对。

[0032] 回到图1D,引物1312延伸至阻断剂1310处,形成具有预定长度的cDNA拷贝,该cDNA拷贝为靶1300的VDJ区的一部分。在一些实施例中,对预定长度(或者换句话讲,阻断剂1310的结合位点)进行选择,使得该方法所用测序技术测得的一个或多个序列读段涵盖所需的VDJ区部分。完成延伸后,采用常规技术对RNA模板1300进行酶切(1325)(例如,用核糖核酸酶(RNase)如RNase H和/或RNase A进行酶切),以得到单链cDNA 1326。使用常规方案中的末端脱氧核苷酸转移酶(TdT)为该cDNA添加3'单核苷酸尾,如聚C尾。对于有尾的cDNA 1331,衔接子1336具有与cDNA 1331的单核苷酸尾互补的突出端,该衔接子在这之后延伸,产生双链DNA 1340,可通过(例如)PCR(1337)对该双链DNA进行扩增,并对所得扩增子进行测序(1338)。

[0033] 可使用引物组扩增经过高频突变的重组核酸(如编码IgH的核酸),所述引物组包括与同一重组核酸上的不同引物结合位点结合的引物;也就是说,此类组可包括与一个或多个不重叠的引物结合位点结合的引物,所述一个或多个结合位点在编码受体链的同一重组核酸上。此类组可包括第一组引物和第二组引物中的一者或两者。在一些实施例中,用第一组引物和第二组引物扩增经受了高频突变的重组核酸,其中两组中的至少一组包括对多个不重叠引物结合位点具有特异性的引物,例如,一组可包括用于各个不同V区段的多个引物,每个引物对不同V区段上的不同非重叠引物结合具有特异性。适用于扩增发生高频突变的重组核酸的一个实施例如图3A-3B所示,其中引物嵌套组用于确保样本中每个重组核酸在(例如)体细胞高频突变、克隆进化等条件下扩增。在退火条件下,重组核酸(如编码IgH分子的重组核酸)在反应混合物中与第一嵌套组引物302结合,在此例中所述第一嵌套组引物包括引物群组304、306和308,这些引物群组对重组核酸300的V区316上的不同位点具有特异性。在该实施例中,第一嵌套组包括多个引物群组,每个引物群组对V区的不同位点或位置具有特异性,其中一个群组中的不同成员对V区中该位点的不同变体具有特异性。在一些实施例中,所述多个群组在2至4个范围内;在其他实施例中,所述多个为2或3。在一些实施例中,第一嵌套组302的各个引物可具有独特的序列标签314和5'非互补尾中的第一引物结合位点312。第一嵌套组引物302与其靶重组核酸退火配对,并延伸穿过D区318以及J区320的至少一部分,形成第一扩增子323,该扩增子包含与三个引物子集304、306和308分别对应的三种组分330、332和334。第一扩增子323的各成员均包含序列标签324和引物结合位点326。

[0034] 移除(322)未延伸的引物后,在退火条件下将第二嵌套组引物340添加到反应混合物。如图3A所示,第二嵌套组引物340包括引物子集336和338,这些引物子集在第一扩增子323成员的J区320上的不同非重叠位置处退火配对。在一些实施例中,第二嵌套组引物可只

包含单个引物群组。第二嵌套组引物340延伸形成第二延伸产物360，该第二延伸产物包括子集350、352和354，每个子集又包括两个与引物336和338对应的进一步子集(亚子集)。在一些实施例中，第二嵌套组引物340包含只对单一引物结合位点具有特异性的引物，而第一嵌套组引物302包含对至少两个非重叠引物结合位点具有特异性的引物。去除未延伸的引物(342)后，可添加通用的正向引物和反向引物进行PCR(356)，然后可对所得扩增子的样本进行测序(358)。在各种实施例中，第一嵌套组和第二嵌套组的引物均可包含序列标签339；第一嵌套组的引物可包含序列标签而第二嵌套组的引物可不包含序列标签；第二嵌套组的引物可包含序列标签而第一嵌套组的引物可不包含序列标签。在一些实施例中，首先延伸第一嵌套组的引物，接着去除或破坏未延伸的引物，然后对第二嵌套组的引物进行退火和延伸(如图3A-3B所示)。在其他实施例中，将退火、延伸和去除步骤的顺序颠倒；也就是说，首先延伸第二嵌套组的引物，接着去除或破坏未延伸的引物，然后对第一嵌套组的引物进行退火和延伸。

[0035] 在上述方法的一些实施例中，可执行不止一个延伸步骤(322或342)，以便(例如)将序列标签附接到样本中更大比例的靶多核苷酸上。在此类实施例中，可将不止一个不同的序列标签附接到靶多核苷酸和/或其拷贝上。也就是说，可将多个不同的序列标签附接到靶多核苷酸及其来自扩增反应(如PCR)的子代上；因此，原始靶多核苷酸的拷贝可标记有不止一个序列标签。正如将在下文中更充分说明的，此类多个序列标签在跟踪遗留污染和允许更灵敏地确定靶多核苷酸序列方面仍然有用。

[0036] 上述实施例中的一些可通过以下步骤进行。例如，从多种或多个T细胞受体链生成克隆型谱的方法可包括如下步骤：(a)在反应混合物中，使第一组引物在引物延伸条件下与来自T细胞的重组核酸样本混合，其中第一组的每个引物均具有受体特异性部分，该受体特异性部分具有一定长度，使得其在靶重组核酸上的预定位置或位点处与不同重组核酸退火配对，然后延伸形成第一延伸产物，并且其中第一组的每个引物均具有包含第一引物结合位点的5'非互补端；(b)从反应混合物中去除第一组的未延伸引物；(c)在引物延伸条件下向反应混合物添加第二组引物，其中第二组的每个引物均具有受体特异性部分，使得该受体特异性部分在预定位置或位点与第一延伸产物退火配对，并具有包含第二引物结合位点的5'非互补端，第一组引物和/或第二组引物包含分别位于受体特异性部分和第一或第二引物结合位点之间的序列标签，并且其中第二组的每个引物延伸形成第二延伸产物，从而每个第二延伸产物均包含第一引物结合位点、第二引物结合位点、至少一个序列标签以及下述中的任一者：(i)T细胞受体链的Vβ区段的一部分和Jβ区段的一部分，(ii)T细胞受体链的Vγ区段的一部分和Jδ区段的一部分，或者(iii)T细胞受体链的Vγ区段的一部分和Jγ区段的一部分；(d)在反应混合物中进行聚合酶链反应，形成扩增子，该聚合酶链反应使用对第一引物结合位点具有特异性的正向引物和对第二引物结合位点具有特异性的反向引物；以及(e)对扩增子的核酸进行测序，形成多种T细胞受体链的克隆型谱。如本文所用，反应混合物中的“引物延伸条件”包括让基本上所有引物结合位点呈单链状态的条件。在一些实施例中，通过解链双链靶核酸来获得此类条件，双链靶核酸解链使得引物结合位点呈单链形式，从而使引物能够与这些引物结合位点退火配对，形成聚合酶延伸的底物。

[0037] 正如在下文引用的参考文献中举例说明的，第一组和第二组引物所结合的预定位置或位点可通过多重核酸扩增领域的普通技术人员所知的常规方法(如多重PCR)来确定。

例如,就靶多核苷酸为编码免疫受体分子的重组核酸而言,Faham和Willis(上文有引用)、Van Dongen et al,Leukemia,17:2257-2317(2003)(van Dongen等人,《白血病》,2003年,第17卷,第2257-2317页)以及类似参考文献为选择用于多重扩增此类靶多核苷酸的引物结合位点提供了指导。在一些实施例中,选择此类预定位置或位点所依赖的几个因素包括:(i)这些位置或位点对扩增效率的影响(理想的情况是,扩增子中不同拷贝的频率忠实地反映样本中靶多核苷酸的频率),(ii)这些位置或位点对扩增子中拷贝长度的影响符合所采用的DNA测序化学的要求,(iii)所选引物是否跨越具有所需多样性的重组核酸的一部分,如VDJ区等。关于这方面,本发明在某种程度上涉及这样的理解和认识:由于一组序列是读数而非模拟信号,引物与不同靶多核苷酸的交叉反应不影响本发明方法的结果(相比于,例如,在完全以PCR扩增的模拟读数、谱型等为基础的方法中)。

[0038] 在一些实施例中,测序步骤包括以下步骤:(i)提供多个序列读段,每个序列读段均具有错误率,并且每个序列读段均包含核苷酸序列和标签序列;以及(ii)比对具有类似标签序列的序列读段群组,然后在群组内根据序列读段进行碱基判读,确定核苷酸序列。随后,可如下文所述的那样将此类群组级核苷酸序列合并为同一克隆型或不同克隆型。在一些实施例中,在PCR步骤中,对第一组和第二组引物的受体特异性部分的长度进行选择,使得扩增子中不同重组核酸的相对含量与样本中重组核酸的相对含量基本上一致。对引物进行此类选择时,位于引物各自的靶多核苷酸上的引物结合位点可能有各不相同的位置和长度。在一些实施例中,序列标签选自序列标签组,该组序列标签比样本中不同的靶多核苷酸的数量大得多,使得样本中基本上每个不同的靶多核苷酸及其拷贝都将具有不同的序列标签(例如,依照Brenner,U.S.patent 7,537,897(Brenner,美国专利7,537,897)中所描述的“取样标记(labeling by sampling)”法)。在一些实施例中,此类组中序列标签的数量为样本中靶多核苷酸群体大小的至少100倍。另外,在一些基本上每个原始靶多核苷酸及其拷贝都标记有同一独特序列标签的实施例中,测序步骤包括:生成扩增子的核酸的序列读段,以及比对具有相同序列标签的序列读段,以确定与样本中相同克隆型对应的序列读段。另外,在一些实施例中,比对步骤还包括:通过确定具有相同序列标签的序列读段各个核苷酸位置上的多数核苷酸,来确定各个克隆型的核苷酸序列。另外,在一些实施例中,可通过使用具有3’→5’单链核酸外切酶活性的核酸酶(所述核酸酶可由,例如,大肠杆菌(E.coli)核酸外切酶I提供,大肠杆菌核酸外切酶I可通过加热方便地灭活)酶切反应混合物中的单链核酸,来进行去除未延伸引物的步骤。在另外的实施例中,上述方法可用于生成克隆型谱,以诊断和/或监测癌症患者(如骨髓瘤、淋巴瘤或白血病患者)的微小残留病灶。在上述方法步骤之后,可用以下附加步骤进行此类诊断和/或监测:从克隆型谱中确定与癌症有关的一种或多种患者特异性克隆型是否存在以及/或者其水平。该实施例的方法可还包括以下步骤:确定一个或多个序列标签中每一个的序列,然后将这些序列与之前所确定克隆型谱的序列标签的序列进行比较,以确定污染序列是否存在以及/或者其水平。在一些实施例中,此类比较步骤包括将更多序列标签之一的序列与克隆型数据库的序列标签进行比较,所述克隆型数据库包含来自至少一个非患者个体的克隆型。

[0039] 在又另一个实施例中,在一次反应中对编码β、δ和γT细胞受体组分的多个重组核酸进行扩增的方法可包括以下步骤:(a)在反应混合物中,使第一组引物在引物延伸条件下与来自T细胞的重组核酸样本混合,其中重组核酸中的每一个均在第一端包含T细胞受体J

β、Jδ或Jγ区段的至少一部分，并且其中第一组的每个引物均具有受体特异性部分，该受体特异性部分具有一定长度，使其与不同重组核酸的第一端退火配对，然后延伸形成第一延伸产物，并且其中第一组的每个引物均具有5’非互补端，该5’非互补端按3’→5’顺序包含序列标签和第一引物结合位点，该序列标签不同于第一组的基本上每个引物；(b)从反应混合物中去除第一组的未延伸引物；(c)在引物延伸条件下向反应混合物添加第二组引物，第二组的每个引物均具有受体特异性部分，该受体特异性部分具有一定长度而能与第一延伸产物退火配对，然后延伸形成第二延伸产物，其中每个第二延伸产物均包含T细胞受体Vβ、Vδ或Vγ区段的至少一部分，并且其中第二组的每个引物均具有包含第二引物结合位点的5’非互补端；以及(d)在反应混合物中进行聚合酶链反应，形成扩增子，该聚合酶链反应使用对第一引物结合位点具有特异性的正向引物和对第二引物结合位点具有特异性的反向引物。上述方法可还包括对扩增子序列样本进行测序的步骤。通常此类样本为“代表性样本”，因为其大到足以使不同克隆型以与存在于生物材料原始样本中大致相同的频率存在于该样本中。在一些实施例中，测序步骤包括：提供多个序列读段，每个序列读段均具有错误率，并且每个序列读段均包含核苷酸序列和标签序列，然后比对具有类似标签序列的序列读段，确定与相同克隆型对应的序列读段。如下文中更充分描述的，每当将多个序列标签附接到原始靶多核苷酸或其拷贝上时，可在进一步的合并步骤中处理此类序列读段。

[0040] 在另一个实施例中，从多种T细胞受体链生成克隆型谱的方法可包括如下步骤：(a)在反应混合物中，使第一组引物在引物延伸条件下与来自T细胞的重组核酸样本混合，其中第一组的每个引物均具有受体特异性部分，该受体特异性部分具有一定长度，使得其在预定位置处与不同重组核酸退火配对，然后延伸形成第一延伸产物，并且其中第一组的每个引物均具有包含第一引物结合位点的5’非互补端；(b)从反应混合物中去除第一组的未延伸引物；(c)向反应混合物添加第二组引物，其中第二组的每个引物均具有一定长度的受体特异性部分，该受体特异性部分在预定位置处对第一延伸产物具有特异性，并具有包含第二引物结合位点的5’非互补端，第一组引物和/或第二组引物包含分别位于受体特异性部分和第一或第二引物结合位点之间的序列标签；(d)进行第一聚合酶链反应，形成第一扩增子，该第一聚合酶链反应使用对第一引物结合位点具有特异性的正向引物以及第二组引物，其中第一扩增子的每条核苷酸序列包含第一引物结合位点、第二引物结合位点、至少一个序列标签，并包含T细胞受体链的Vβ区段的一部分和Jβ区段的一部分，或者T细胞受体链的V区段的一部分和Jδ区段的一部分，或者T细胞受体链的Vγ区段的一部分和Jγ区段的一部分，并且其中对第一组和第二组引物的受体特异性部分的长度进行选择，使得扩增子中不同重组核酸的相对含量与样本中不同重组核酸的相对含量基本上一致；(e)添加对第二引物结合位点具有特异性的反向引物；(f)在反应混合物中进行第二聚合酶链反应，形成第二扩增子，该聚合酶链反应使用对第一引物结合位点具有特异性的正向引物和对第二引物结合位点具有特异性的反向引物；(g)对第二扩增子的核酸进行测序，形成多种T细胞受体链的克隆型谱。在一些实施例中，测序步骤包括：提供多个序列读段，每个序列读段均具有错误率，并且每个序列读段均包含核苷酸序列和标签序列，然后比对具有类似标签序列的序列读段，确定与相同克隆型对应的序列读段。在靶多核苷酸和/或其拷贝标记有不止一个序列标签的另外的实施例中，可如下文中更充分描述的那样在进一步的合并步骤中处理序列读段。

[0041] 又如，从多种B细胞受体链生成克隆型谱的方法可通过如下步骤进行：(a)在反应混合物中，使第一嵌套组引物在引物延伸条件下与来自B细胞的重组核酸样本混合，该第一嵌套组包括一个或多个引物群组，其中各群组的每个引物均具有受体特异性部分，该受体特异性部分具有一定长度，使得来自不同群组的每个引物的受体特异性部分与不同的重组核酸在预定位点退火配对，该预定位点不与第一嵌套组的任何其他引物的预定位点重叠，并且其中各群组的每个引物均具有包含第一引物结合位点的5’非互补端；(b)延伸第一嵌套组的引物，形成第一延伸产物；(c)从反应混合物中去除第一嵌套组的未延伸引物；(d)在引物延伸条件下向反应混合物添加第二嵌套组引物，该第二嵌套组包括一个或多个引物群组，其中各群组的每个引物均具有受体特异性部分，该受体特异性部分具有一定长度，使得来自不同群组的每个引物的受体特异性部分与第一延伸产物在预定位点退火配对，该预定位点不与第二嵌套组的任何其他引物的预定位点重叠，并且其中各群组的每个引物均具有包含第二引物结合位点的5’非互补端，并且其中第一嵌套组引物和/或第二嵌套组引物包含分别位于其受体特异性部分和其第一或第二引物结合位点之间的序列标签；(e)延伸第二嵌套组的引物，形成第二延伸产物，使得每个第二延伸产物均包含第一引物结合位点、第二引物结合位点、至少一个序列标签以及下述中的任一者：(i)B细胞受体重链的V区段的一部分和J区段的一部分，或(ii)B细胞受体κ轻链的V区段的一部分和J区段的一部分；(f)在反应混合物中进行聚合酶链反应，形成扩增子，该聚合酶链反应使用对第一引物结合位点具有特异性的正向引物和对第二引物结合位点具有特异性的反向引物；以及(g)对扩增子的核酸进行测序，形成多种B细胞受体链的克隆型谱。

[0042] 在一些实施例中，可在步骤(b)和/或(e)中对引物进行不止一个循环的退火和延伸(在解链延伸产物后)，在这种情况下，样本中原始重组核酸的拷贝可标记有一个或多个序列标签。在这些实施例中，测序步骤(g)可包括如下文所述的进一步的比对和合并步骤，用于确定克隆型和克隆型谱。在一些实施例中，例如，在步骤(b)和(e)中仅进行了单次延伸的实施例中，测序步骤包括：提供多个序列读段，每个序列读段均具有错误率，并且每个序列读段均包含核苷酸序列和标签序列，然后比对具有类似标签序列的序列读段，确定与相同克隆型对应的序列读段。同上，在一些实施例中，在PCR中对第一组和第二组引物的受体特异性部分的位置和长度进行选择，使得扩增子中不同重组核酸的相对含量与样本中不同重组核酸的相对含量基本上一致。

[0043] 在一些实施例中，在引物延伸步骤中将序列标签附接到靶多核苷酸或其拷贝上，其中基本上每个不同的靶多核苷酸及其拷贝都标记有相同的序列标签。在其他实施例中，样本的靶多核苷酸或其拷贝可标记有不止一个不同的序列标签。正如将在下文中进一步说明的，在一些实施例中，可在包含序列标签的引物(第一组引物或第二组引物)存在时进行多次延伸或多个PCR循环，这可使不同的序列标签附接到同一靶多核苷酸和/或其拷贝上。

[0044] 克隆型分析中的序列标签

[0045] 在一个方面，本发明涉及一种方法，用于获得和分析来自免疫分子组库如T细胞受体(TCR)或B细胞受体(BCR)或其确定片段的序列数据，以快速有效地确定克隆型谱。序列数据通常包括从用于分析免疫分子的DNA测序仪获得的大量序列读段，即碱基判读的序列和相关质量得分。构建克隆型谱的关键挑战在于，一些序列读段包含非生物来源(如提取步骤、测序化学、扩增化学等)的错误，需要从这些序列读段中快速准确地辨别出具有真正差

异的序列读段。本发明的一个方面包括将独特的序列标签附接到样本中的每个靶多核苷酸(例如重组核酸)上,以帮助确定此类缀合物的序列读段是否来源于同一原始靶多核苷酸。根据本发明的一个方面,将序列标签附接至体细胞重组核酸分子上,形成标签-分子缀合物,其中,此类缀合物的每个重组核酸均具有独特的序列标签。通常在将核酸分子从包含T细胞和/或B细胞和/或无细胞DNA的样本中提取出来后再进行这种附接。优选的是,如通过常规的序列距离测量法(如汉明距离法(Hamming distance)等)所测定的,此类独特序列标签彼此间的差异尽可能大。通过使标签-分子缀合物中序列标签之间的距离最大化,即使测序和扩增错误率很高,但比起与不同缀合物中任何其他标签序列的祖先标签序列的距离,一种缀合物的序列标签始终离其祖先标签序列近得多。例如,如果采用16聚序列标签,并且在一组克隆型上的每个此类标签与该组克隆型上所有其他的序列标签之间的汉明距离为至少50%(即八个核苷酸),那么将需要至少八个测序或扩增错误才会将一种此类标签转变成另一种,进而造成序列标签的错读(以及根据错误的序列标签对克隆型的序列读段进行不正确的分组)。在一个实施例中,对序列标签进行选择,使得标签附接到重组核酸分子上形成标签-分子缀合物后,标签-分子缀合物上标签之间汉明距离的数值为此类序列标签总长度的至少25%(也就是说,每个序列标签与所有其他的此类标签在序列上有至少25%的核苷酸不同);在另一个实施例中,此类序列标签之间汉明距离的数值为此类序列标签总长度的至少50%。

[0046] 在一个方面,本发明通过以下步骤实现:(a)从个体获得包含T细胞和/或B细胞和/或无细胞DNA的样本;(b)将序列标签附接到样本中T细胞受体基因或免疫球蛋白基因的重组核酸分子上,形成标签-分子缀合物,其中标签-分子缀合物的基本上每个分子均具有独特的序列标签;(c)扩增该标签-分子缀合物;(d)对标签-分子缀合物进行测序;以及(e)比对类似序列标签的序列读段,确定与样本中相同重组核酸对应的序列读段。采用常规技术获得包含B细胞或T细胞的样本。在附接序列标签的步骤中,优选的是,序列标签不仅独特,而且彼此完全不同,即使出现大量的测序或扩增错误,将一种序列标签转变成另一种的可能性也将接近于零。对于大部分测序技术而言,附接序列标签后必须扩增标签-分子缀合物;然而,每当采用单分子测序技术时,扩增步骤是任选的。单分子测序技术包括但不限于单分子实时(SMRT)测序、纳米孔测序等,例如美国专利7,313,308、8,153,375、7,907,800、7,960,116、8,137,569;Manrao et al,Nature Biotechnology,4(8):2685-2693(2012)(Manrao等人,《自然生物技术》,2012年,第4卷,第8期,第2685-2693页)等等。

[0047] 在另一方面,本发明包括通过对独特序列标签进行计数来测定样本中淋巴细胞数的方法。即使没有序列标签,TCRβ或IgH基因的克隆型,尤其是包含V(D)J区的那些,也会为淋巴细胞及其克隆提供独特的标记。每当从基因组DNA获得重组核酸后,可在测序后对独特的克隆型进行计数,由所得克隆型的数量来估计样本中的淋巴细胞计数。每当存在与相同克隆型相关的相同淋巴细胞的显著克隆群体时(或者当重组的核酸是从其个体序列的量可反映或依赖于表达率及细胞数的样本mRNA获得时),该方法失效。序列标签的使用克服了该缺陷,并且特别适用于对患有多种淋巴疾病(如淋巴瘤或白血病)的患者进行淋巴细胞计数。根据本发明的一个方面,无论是否存在大的显性克隆,例如伴随白血病存在,序列标签均可用于获得样本中淋巴细胞的绝对计数。可用以下步骤实施此类方法:(a)从个体获得包含淋巴细胞的样本;(b)将序列标签附接到淋巴细胞的T细胞受体基因或免疫球蛋白基因的

重组核酸分子上，形成标签-分子缀合物，其中标签-分子缀合物的基本上每个分子均具有独特的序列标签；(c)扩增该标签-分子缀合物；(d)对标签-分子缀合物进行测序；以及(e)对不同序列标签的数目进行计数，确定样本中淋巴细胞的数目。在一些实施例中，重组核酸分子来自基因组DNA。

[0048] 在本发明的一个实施例中，通过取样标记将序列标签附接到样本的重组核酸分子上，例如Brenner et al,U.S.patent 5,846,719(Brenner等人，美国专利5,846,719)；Brenner et al,U.S.patent 7,537,897(Brenner等人，美国专利7,537,897)；Macevicz,International patent publication WO 2005/111242(Macevicz，国际专利公布WO 2005/111242)等中所公开的，这些专利以引用并入本文。在取样标记中，将待标记(或者独特地添加标签)的多核苷酸群体用于(通过附接、连接等)对大得多的群体的序列标签进行采样。也就是说，如果多核苷酸群体具有K个成员(包括同一多核苷酸的重复)，并且序列标签的群体具有N个成员，则N>>K。在一个实施例中，本发明使用的序列标签的群体大小为样本中克隆型群体大小的至少10倍；在另一个实施例中，本发明使用的序列标签的群体大小为样本中克隆型群体大小的至少100倍；并且在另一个实施例中，本发明使用的序列标签的群体大小为样本中克隆型群体大小的至少1000倍。在其他实施例中，对序列标签的群体大小进行选择，使得每当将此类克隆型与此类序列标签群体在(例如)附接反应(如连接反应)、扩增反应等中结合时，样本中基本上每个克隆型都将具有独特的序列标签。在一些实施例中，基本上每个克隆型意指至少90％的此类克隆型将具有独特的序列标签；在其他实施例中，基本上每个克隆型意指至少99％的此类克隆型将具有独特的序列标签；在其他实施例中，基本上每个克隆型意指至少99.9％的此类克隆型将具有独特的序列标签。在许多组织样本或活体组织切片中，T细胞或B细胞的数目可多达或为约1百万个细胞；因此，在本发明的一些使用此类样本的实施例中，在取样标记中使用的独特序列标签的数目为至少10^8，或者在其他实施例中为至少10^9。

[0049] 在多达1百万个克隆型被取样标记的此类实施例中，可通过在合成反应的各个添加步骤中使全部四种核苷酸前体的混合物发生反应来进行组合合成，由此有效地产生大序列标签组，例如Church,U.S.patent 5,149,625(Church，美国专利5,149,625)中所公开的，该专利以引用方式并入。结果得到具有“$N_1N_2...N_k$”结构的序列标签组，其中每个N_i＝A、C、G或T，且k为标签中核苷酸的数目。这种组合合成产生的序列标签组中的序列标签数为4^k。因此，k为至少14或者k在约14至18范围内的此类序列标签组适于通过取样标记来将序列标签附接到10^6个成员的分子群体上。具有上述结构的序列标签组包括许多序列，所述序列可在实施本发明方法时造成困难或错误。例如，上述组合合成的序列标签组包括许多具有均聚物区段的成员标签，因此一些测序方法(如合成测序方法)在均聚物区段超过一定长度时难以准确地进行测定。因此，本发明包括组合合成的序列标签，其具有对特定的方法步骤(如测序)有效的结构。例如，可将四种天然的核苷酸分成不相交子集，在组合合成中交替地使用这些不相交子集，由此产生对合成测序化学反应有效的数个序列标签结构，从而防止均聚物区段超过给定的长度。例如，使z为A或C，并使x为G或T，以产生如下结构的序列标签：

[0050] $[(z)_1(z)_2...(z)_i][(x)_1(x)_2...(x)_j]...$

[0051] 其中，i和j可相同也可不同，其经选择以限制任何均聚物区段的大小。在一个实施例中，i和j在1至6的范围内。在此类实施例中，序列标签的长度可在12至36个核苷酸范围

内;并且在其他实施例中,此类序列标签的长度可在12至24个核苷酸范围内。在其他实施例中,可使用其他的核苷酸配对,例如z为A或T,而x为G或C;或者z为A或G,而x为T或C。或者,使z''为四种天然核苷酸中任意三种的组合,并使x'为不是z'的任何核苷酸(例如,z'为A、C或G,而x'为T)。这样产生了如下序列标签结构:

[0052] $[(z')_1(z')_2...(z')_i]x'[(z')_1(z')_2...(z')_i]x'...$

[0053] 其中,i如上所述进行选择,并且x'的出现作为终止任何非期望均聚物的标点。

[0054] 其他序列标签

[0055] 本发明使用了用独特的序列标签标记核酸(如基因组DNA的片段),而后进行扩增和测序的方法,所述标签可包括"镶嵌标签"。此类序列标签可用于识别扩增和测序错误。镶嵌标签将测序和扩增的人工痕迹减至最小,这些人工痕迹由不当的退火、启动、发夹结构形成等导致,可能伴随着现有技术的完全随机序列标签而出现。在一个方面,镶嵌标签为包含交替的恒定区与可变区的序列标签,其中每个恒定区位于镶嵌标签中且包含预定的核苷酸序列,而每个可变区位于镶嵌标签中且包含预定数目的随机选取的核苷酸。举例说明,22聚镶嵌标签(SEQ ID NO:1)可具有以下形式:

[0056] 核苷酸位置:

[0057]

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22
N	N	N	b	b	b	b	b	N	b	b	N	N	N	b	b	b	N	N	b	N	N
1			2				3	4		5			6		7		8		9		

[0058] 区域位置

[0059] 恒定区和可变区共九个,其中区域1(第1-3位核苷酸)、区域3(第9位核苷酸)、区域5(第12-14位核苷酸)、区域7(第18-19位核苷酸)和区域9(第21-22位核苷酸)为可变区(双下划线所示核苷酸),区域2(第4-8位核苷酸)、区域4(第10-11位核苷酸)、区域6(第15-17位核苷酸)和区域8(第20位核苷酸)为恒定区。N表示从A、C、G或T中随机选取的核苷酸;因此,本例的镶嵌标签数目为4^{11}=4,194,304种标签,b表示所示位置处的预定核苷酸。在一些实施例中,选择b的序列"***bbbb*bb***bbb**b**",使其在组成样本的生物体基因组中出现完全匹配的可能性最小。

[0060] 在一个方面,对于本发明方法的具体实施例中的镶嵌标签,位置相同的所有恒定区具有相同的长度,且位置相同的所有可变区具有相同的长度。这允许用传统的化学反应和仪器通过部分组合合成法来合成镶嵌标签。

[0061] 在一个方面,镶嵌标签包含10至100个核苷酸,或者12至80个核苷酸,或者15至60个核苷酸。在一些实施例中,镶嵌标签包含至少八个具有随机选取核苷酸的核苷酸位置;在其他实施例中,每当镶嵌标签的长度为至少15个核苷酸时,其包含至少12个具有随机选取核苷酸的核苷酸位置。在另一方面,镶嵌标签内可变区的长度不可超过七个核苷酸。

[0062] 在另一方面,可在以下步骤中使用镶嵌标签:(i)从样本中的核酸制备DNA模板;(ii)取样标记DNA模板,形成多重标签-模板缀合物,其中,标签-模板缀合物的基本上每个DNA模板均具有独特的镶嵌标签,所述镶嵌标签包含交替的恒定区和可变区,每个恒定区位

于镶嵌标签中并包含长度为1至10个核苷酸的预定序列，而每个可变区位于镶嵌标签中且长度为1至10个随机选取的核苷酸，使得位置相同的恒定区具有相同的长度，且位置相同的可变区具有相同的长度；(iii)扩增所述多重标签-模板缀合物；(iv)针对每个扩增的标签-模板缀合物生成多个序列读段；以及(v)确定具有相同镶嵌标签的每组多个序列读段在各个核苷酸位置处的共有核苷酸，由此确定每个核酸的核苷酸序列。在另一方面，可在以下步骤中使用镶嵌标签：(a)从样本中的核酸制备单链DNA模板；(b)取样标记所述单链DNA模板，形成标签-模板缀合物，其中，标签-模板缀合物的基本上每个单链DNA模板均具有独特的序列标签(即镶嵌标签)，所述独特的序列标签的长度为至少15个核苷酸，并具有以下形式：

[0063] $[(N_1N_2...N_{Kj})(b_1b_2...b_{Lj})]_M$

[0064] 其中，每个N_i(i＝1,2,...K_j)为随机选自A、C、G和T的核苷酸；当每个j小于或等于M(即区域$N_1N_2...N_{Kj}$为可变区)时，K_j为1至10范围内的整数；每个b_i(i＝1,2,...L_j)为核苷酸；当每个j小于或等于M时，L_j为1至10范围内的整数；如此，每个序列标签(i)对于每个j具有相同的K_j，并且(ii)对于每个j具有相同的序列$b_1b_2...b_{Lj}$(即区域$b_1b_2...b_{Lj}$为恒定区)；且M为大于或等于2的整数；(c)扩增所述标签-模板缀合物；(d)针对每个扩增的标签-模板缀合物生成多个序列读段；并且(e)确定具有相同序列标签的每组多个序列读段在各个核苷酸位置处的共有核苷酸，由此确定每个核酸的核苷酸序列。在一些实施例中，所述多个序列读段为至少10^4个；在其他实施例中，所述多个序列读段为至少10^5个；在其他实施例中，所述多个序列读段为至少10^6个。在一些实施例中，上述序列标签的总长度在15至80个核苷酸的范围内。

[0065] 附接序列标签

[0066] 多种不同的附接反应可用于将独特的标签附接到样本中基本上每个克隆型上(除了上文示出的那些之外)。许多用于捕获样本核酸子集以(例如)降低微阵列或基因组测序技术中样本复杂性的技术可与本发明中用于将序列标签附接至重组核酸的常规修改形式一起使用。捕获不同靶核酸组以供后续操作(包括附接序列标签、测序等)的示例性技术包括以下内容：Willis et al,U.S.patent 7,700,323(Willis等人，美国专利7,700,323)；Jones et al,U.S.patent publication 2005/0142577(Jones等人，美国专利公布2005/0142577)；Gullberg et al,U.S.patent publication 2005/0037356(Gullberg等人，美国专利公布2005/0037356)；Porreca et al,Nature Methods,4(11):931-936(2007)(Porreca等人，《自然方法》，2007年，第4卷，第11期，第931-936页)；Turner et al,Nature Methods,6(5):315-316(2009)(Turner等人，《自然方法》，2009年，第6卷，第5期，第315-316页)；Church,U.S.patent 5,149,625(Church，美国专利5,149,625)；Macevicz,U.S.patent 8,137,936(Macevicz，美国专利8,137,936)等等。

[0067] 在一个实施例中，将包含重组核酸分子(该重组核酸分子又包含克隆型序列)的样本与序列标签的群体或文库组合，使得两个分子群体的成员能够随机组合，并且(例如，以共价键)关联或连接，由此完成此类附接。例如，此类随机组合可发生在双分子反应中，其中包含标签的引物与靶核酸退火配对并延伸，或者其中包含标签的衔接子连接至靶核酸的末端。在一些实施例中，附接标签的方法可部分地取决于DNA测序的方法。例如，在产生相对较长的准确序列读段的测序方法(如454测序)中，可使用常规技术(如5'-RACE)由包含重组核酸的mRNA制成cDNA文库，如Freeman et al,Genome Research,19:1817-1824(2009)

(Freeman等人,《基因组研究》,2009年,第19卷,第1817-1823页)中所公开的,然后可将包含序列标签的衔接子连接至一个或两个末端,由此附接序列标签。在其他实施例中,当使用测序方法(如"亿明达"测序或"离子激流"测序)来产生相对较短和易错的序列读段时,可能需要采取进一步的措施,以使得用于测序的扩增子具有该技术生成的序列读段所涵盖的长度。在此类标签附接反应中,克隆型序列包括线性单链或双链多核苷酸,并且序列标签由试剂如扩增引物(如PCR引物)、连接衔接子、可环化探针、质粒等携带。若干能够携带序列标签群体的此类试剂在如下文献中公开:Macevicz,U.S.patent 8,137,936(Macevicz,美国专利8,137,936);Faham et al,U.S.patent 7,862,999(Faham等人,美国专利7,862,999);Landegren et al,U.S.patent 8,053,188(Landegren等人,美国专利8,053,188);Unrau and Deugau,Gene,145:163-169(1994)(Unrau和Deugau,《基因》,1994年,第145卷,第163-169页);Church,U.S.patent 5,149,625(Church,美国专利5,149,625)等,这些文献以引用方式并入本文。

[0068] 图2A和2B示出了包括PCR的附接反应,其中将序列标签群体($T_1,T_2,T_3...T_j,T_{j+i}...T_k,T_{k+1}...T_{n-1},T_n$)掺入引物2100中。序列标签群体的大小比重组核酸分子2102的大小大得多。通过在PCR的第一个循环中使引物与核酸分子退火配对,并用DNA聚合酶延伸引物,来将序列标签附接到重组核酸分子上。该图描绘了重组核酸分子在(例如)V区2108中经由其通用引物结合区2104与引物随机退火配对,由此对整个序列标签群体的一小部分进行选取(或者说,取样)的方式。由于引物(以及由此序列标签)与重组核酸序列分子随机组合,所以相同序列标签附接到不同核酸分子上的可能性较小;然而,如果序列标签群体如本文所述的那样大,那么这种可能性将微乎其微,从而基本上每个重组核酸分子将附接有独特的序列标签。正向和反向引物对的其他引物2106与C区2110退火配对,以使得扩增子2112在多个退火、延伸和解链循环后形成,从而将独特的序列标签附接到V(D)J区,所述V(D)J区包括群体2102的克隆型。也就是说,扩增子2112包括来自附接反应的标签-分子缀合物。

[0069] 图2C和2D示出了将一对序列标签附接到样本中每一个或者基本上每一个重组核酸上的方法。如图2A和2B所示的方法中,带有序列标签($T_1,T_2,T_3...T_j,T_{j+i}...T_k,T_{k+1}...T_{n-1},T_n$)的引物2200被用作下游引物,此外,以带有序列标签($T_m,T_{m+1},T_{m+2}...T_q,T_{q+1},T_{q+2},...T_r,T_{r+1},T_{r+2},...T_s,T_{s+1},T_{s+2},...$)的引物2206取代通用引物2106被用作上游引物。如下游引物组一样,与重组核酸分子2202的数目相比,由上游引物2206携带的不同序列标签数目可较大,以使得在扩增后基本上每个重组核酸2202都将具有独特的标签。在一些实施例中,引物2206和2200中的每组序列标签不需要与图2A和2B所示实施例中的序列标签组一样大。由于每个重组核酸均由一对序列标签独特地标记,所以对于标记单个重组核酸的一对序列标签而言,与不同重组核酸共享该序列标签对中的一个序列标签不会减损其实质的独特性。因此,在图2C和2D的实施例中,每个引物组2200和2206的序列标签可不及引物组2100的序列标签多样化。例如,如果采用随机序列标签,并且引物2100包含16聚序列标签,那么引物2200和2206可各自包含8聚序列标签以得到相同的总体序列标签多样性。另外,图2C和2D的实施例与图2A和2B的实施例运作方式相似。通过在PCR的第一个循环中使引物与核酸分子退火配对,并用DNA聚合酶延伸引物,来将序列标签附接到重组核酸分子上。同上,图2C描绘了重组核酸分子分别在(例如)V区2208和C区2210中经由其通用引物结合区2204和2205与引物随机退火配对,由此对整个序列标签对群体的一小部分进行选取(或者说,取

样)的方式。由于引物(以及由此序列标签)与重组核酸序列分子随机组合,相同序列标签对附接到不同核酸分子上的可能性较小;然而,如果序列标签群体如本文所述的那样大,那么这种可能性将微乎其微,从而基本上每个重组核酸分子将附接有独特的序列标签对。扩增子2212在多个退火、延伸和解链循环后形成,从而将独特的序列标签对附接到V(D)J区,所述V(D)J区包括群体2202的克隆型。

[0070] 也就是说,扩增子2212包括来自附接反应的标签-分子缀合物。在一些实施例中,可通过(例如)Porreca等人(上文有引用)、Willis等人(上文有引用)或类似参考文献所公开的技术的常规修改形式,将可环化探针用于捕获序列标签,并将序列标签附接到所需的重组核酸上。如图2E和2F所示,提供的可环化探针2302包含以下元件:上游靶结合区段2304,具有5'-磷酸化末端2305的下游靶结合区段2306;序列标签2310;第二通用引物结合位点2314;任选的切割位点2308;以及第一通用引物结合位点2312。在退火条件下,可环化探针2302于反应混合物中与包含靶多核苷酸2300的样本结合,所述靶多核苷酸可为(例如)cDNA的第一链或第二链,所述cDNA通过常规技术由mRNA制备而成。如同所示,靶多核苷酸包含编码IgH或TCRβ链的重组核酸的V区、NDN区、J区和C区。在一些实施例中,分别对上游靶结合区段2304和下游靶结合区段2306的序列进行选择,使得其跨越靶多核苷酸的VDJ区的一部分。对上游靶结合区段2304和下游靶结合区段2306退火后,可环化探针2302和靶多核苷酸2300在反应混合物中形成复合物2330。在存在DNA聚合酶和dNTP的情况下,上游靶结合区段2304延伸(2340)至下游靶结合区段2306,复制(从而捕获)靶多核苷酸的VDJ区的一部分。当存在连接酶活性时,延伸的上游靶结合区段连接至下游靶结合区段2306,由此形成封闭的单链DNA环2342。然后,可任选地用核酸外切酶处理(2344)反应混合物,以去除未反应的探针和靶多核苷酸。在一些实施例中,通过在切割位点2308处(切割位点可为,例如,稀有切割核酸内切酶识别位点)进行切割,或将RNA单体插入探针并用RNase H或类似的酶切割,使单链环2342线性化,然后可通过引物2350和2352扩增线性化探针2348的VDJ标签插入序列。引物2350和2352可包含非互补区,该非互补区用于添加元件,以便随后能够进行DNA测序(2354)。或者,单链环可用于生成纳米球模板,以供直接测序,例如Drmanac et al,Science,327(5961):78-81(2010)(Drmanac等人,《科学》,2010年,第327卷,第5961期,第78-81页);美国专利8,445,196等。

[0071] 图2G示出了将序列标签附接到编码免疫受体分子的重组核酸上的另一个实施例。实施此实施例的指导可见于Faham and Zheng,U.S.patent 7,208,295(Faham和Zheng,美国专利7,208,295),该专利以引用方式并入本文。在退火条件下,重组核酸2450于反应混合物中与探针2454和衔接子2456结合。探针2454包含受体特异性部分2455和衔接子特异性部分2457。举例来说,探针2454可包括探针混合物,其中不同探针具有受体特异性部分,该受体特异性部分对不同J区具有特异性,或者在其他实施例中,该受体特异性部分对不同V区具有特异性。经5'磷酸化的衔接子2456包含位于其5'端的探针特异性部分2458、序列标签2460和第一引物结合位点2462。对探针2454的受体特异性部分2455和衔接子特异性部分2457以及探针特异性部分2458的位置、序列和长度进行选择,使得它们彼此杂交,形成结构体2452。结构体2452形成后,从重组核酸2450上切下单链部分2461,并将重组核酸2450的自由3'末端连接至衔接子2456的5'磷酸化末端,以形成第一延伸产物2459,然后去除(2474)探针2454。可如Faham和Zheng的专利中所述的那样,用单链核酸酶实现对2461的切割。在一

个实施例中，通过在(例如)用dUTP取代dTTP的PCR中用尿嘧啶替代胸苷来合成探针2454，并通过使用尿嘧啶DNA糖基化酶(UDG)处理来去除尿嘧啶，例如以引用方式并入的Faham et al,U.S.patent 7,208,295(Faham等人，美国专利7,208,295)中所述。UDG处理切割探针2454中的尿嘧啶，以产生片段2455。释放探针后，去除(2476)衔接子和翼，将正向引物2466和反向引物2468添加至延伸产物2464，并进行PCR(2470)，然后对所得扩增子的样本进行测序(2472)。

[0072] 在与图2G所示实施例类似的实施例中，类似的探针和衔接子可用于将序列标签附接到靶多核苷酸的预定位点上，其中瓣状核酸内切酶(如FEN-1)用于切割与2461对应的单链部分。在此实施例中，除了使用不同的核酸酶之外，探针和衔接子序列的极性还是相反的；也就是说，瓣状核酸内切酶的底物要求，与2454对应的衔接子的3'末端与靶序列2450退火配对，并且与2452对应的单链部分为靶序列的5'端。切割和去除探针序列后，剩下的步骤基本上相同。在检测分析中使用瓣状核酸内切酶的指导可见于以下参考文献:Lyamichev et al,Nature Biotechnology,17:292-296(1999)Lyamichev等人，《自然生物技术》，1999年，第17卷，第292-296页)；Eis et al,Nature Biotechnology,19:673-676(2001)(Eis等人，《自然生物技术》，2001年，第19卷，第673-676页)；以及类似的参考文献。

[0073] 在一些实施例中，重组核酸编码免疫受体分子链，所述免疫受体分子链通常形成免疫组库，该免疫组库可包括非常大的多核苷酸组(例如>1000，而更常见的是大于10,000，并且更加常见的是100,000至1,000,000，或者更大)，这些多核苷酸非常相似，其长度可小于500个核苷酸，或者在其他实施例中小于400个核苷酸，又或者在其他实施例中小于300个核苷酸。在本发明的一个方面，发明人认识和理解到，这些特性使得完全不同的序列标签能够用于有效地比较高度相似的克隆型的序列读段，以确定其是否来源于同一原始序列。

[0074] 样本

[0075] 术语“样本”指大量生物材料，在一些实施例中，这些生物材料来自患者，并且包含细胞和/或无细胞DNA；也就是说，该术语可与术语“标本”或“组织样本”交换使用。术语“样本”有时也分别用在获得(例如)重组核酸的更大组或量的子集或部分的统计学意义上；具体地讲，术语“样本”的统计学用法也可被理解为意指“代表性样本”，那样，这类样本被理解为反映或者近似于(例如)组织中不同核酸的相对频率。本领域的技术人员能够从上下文中辨别该术语的正确用法。

[0076] 可从免疫细胞或液体(如血液)的样本中获得克隆型谱，所述样本包含编码免疫受体链的无细胞核酸。举例来说，免疫细胞可包括T细胞和/或B细胞。T细胞(T淋巴细胞)包括(例如)表达T细胞受体的细胞。T细胞包括辅助性T细胞(效应T细胞或Th细胞)、细胞毒性T细胞(CTL)、记忆T细胞和调节性T细胞。在一个方面，T细胞样本包含至少1,000个T细胞；而更典型的是，样本包含至少10,000个T细胞，且更典型的是包含至少100,000个T细胞。在另一方面，样本包含的T细胞数目在1000至1,000,000个细胞范围内。免疫细胞的样本也可包含B细胞。B细胞包括(例如)浆B细胞、记忆B细胞、B1细胞、B2细胞、边缘区B细胞和滤泡性B细胞。B细胞可表达免疫球蛋白(抗体、B细胞受体)。同上，在一个方面，B细胞样本包含至少1,000个B细胞；而更典型的是，样本包含至少10,000个B细胞，且更典型的是包含至少100,000个B细胞。在另一方面，样本包含的B细胞数目在1000至1,000,000个B细胞范围内。

[0077] 本发明方法中使用的样本可来自多种组织，包括(例如)肿瘤组织、血液和血浆、淋

巴液、大脑和脊髓周围的脑脊液、骨关节周围的滑液等。在一个实施例中，所述样本为血样。该血样可为约0.1、0.2、0.3、0.4、0.5、0.6、0.7、0.8、0.9、1.0、1.5、2.0、2.5、3.0、3.5、4.0、4.5或5.0mL。所述样本可为肿瘤活体组织切片。该活体组织切片可来自(例如)大脑、肝脏、肺、心脏、结肠、肾或骨髓的肿瘤。本领域的技术人员使用的任何活检技术均可用于从受试者分离得到样本。例如，活检可为开放式活检，其中要使用全身麻醉。所述活检可为封闭式活检，其中造成的切口比开放式活检中更小。所述活检可为组织芯活检或切取活检，其中要移除组织的一部分。活检可为切除活检，其中要尝试移除整个病灶。活检可为细针抽吸活检，其中要用针移除组织或液体样本。

[0078] 在一些实施例中，就诊断样本而言，本发明方法的克隆型谱出自肿瘤或外周血，或者就用于监测残留病灶的样本而言，所述克隆型谱出自外周血。通过诊断样本确定与疾病(如，淋巴疾病或骨髓增生性疾病)相关的一种或多种克隆型。通常，与淋巴疾病或骨髓增生性疾病相关的一种或多种克隆型为克隆型谱中出现频率最高的那些。在一些情况下可能存在单一相关克隆型，而在另一些情况下可能存在多种与淋巴疾病或骨髓增生性疾病相关的克隆型。可从受此类疾病感染的任何组织中取得肿瘤样本，所述组织包括淋巴结或淋巴系统以外的其他组织。如上所述，用于监测残留病灶的克隆型谱可出自从外周血提取的核酸样本。样本的核酸可来自B细胞，所述B细胞来自外周血中包含细胞的部分(如血浆)，或者来自外周血中不含细胞的部分(如血清)。在一个实施例中，外周血样本包含至少1,000个B细胞；而更典型的是，此样本包含至少10,000个B细胞，且更典型的是包含至少100,000个B细胞。在另一方面，样本包含的B细胞数目在1000至1,000,000个B细胞范围内。在一些实施例中，样本中的细胞数目限制了测量的灵敏度。也就是说，使用更大的外周血样本，会使检测残留病灶的灵敏度更高。例如，在包含1,000个B细胞的样本中，无论在测序分析这些细胞的DNA时获得了多少测序读段，可检测克隆型的最低频率都为1/1000或0.001。样本的核酸可来自T细胞，所述T细胞来自外周血中包含细胞的部分(如血浆)，或者来自外周血中不含细胞的部分(如血清)。在一个实施例中，外周血样本包含至少1,000个T细胞；而更典型的是，此样本包含至少10,000个T细胞，且更典型的是包含至少100,000个T细胞。在另一方面，样本包含的T细胞数目在1000至1,000,000个T细胞范围内。在一些实施例中，样本中的细胞数目限制了测量的灵敏度。也就是说，使用更大的外周血样本，会使检测残留病灶的灵敏度更高。例如，在包含1,000个T细胞的样本中，无论在测序分析这些细胞的DNA时获得了多少测序读段，可检测克隆型的最低频率都为1/1000或0.001。

[0079] 用于本发明的样本可包括DNA(如基因组DNA)或RNA(如信使RNA)。所述核酸可为(例如，提取自循环系统的)无细胞DNA或RNA，Vlassov et al，Curr.Mol.Med.，10:142-165(2010)(Vlassov等人，《当今分子医学》，2010年，第10卷，第142-165页)；Swarup et al，FEBS Lett.，581:795-799(2007)(Swarup等人，《欧洲生物化学学会联盟通讯》，2007年，第581卷，第795-799页)。在所提供的本发明方法中，来自受试者的可供分析的RNA或DNA的量包括，例如，在一些应用(如，使用其他细胞选择标准如形态指标的校准试验)中低至单个细胞，以及高至1000万个细胞甚或更多，所述1000万甚或更多个细胞转换为DNA的量在6pg至60μg范围内，而转换为RNA的量在1pg至10μg范围内。在一些实施例中，核酸样本为6pg至60μg的DNA样本。在其他实施例中，核酸样本为来自100μL至10mL外周血的DNA样本；在其他实施例中，核酸样本为来自100μL至10mL外周血的无细胞部分的DNA样本。

[0080] 在一些实施例中，淋巴细胞或无细胞核酸的样本足够大，使得基本上每个具有不同克隆型的B细胞或T细胞都出现在其中，从而形成克隆型“组库”。在一个实施例中，为了获得每个不同克隆型的基本表征，取用的样本以99％的概率包含出现频率为0.001％或更高的群体中的每个克隆型。在另一个实施例中，取用的样本以99％的概率包含出现频率为0.0001％或更高的群体中的每个克隆型。并且在另一个实施例中，取用的样本以99％的概率包含出现频率为0.00001％或更高的群体中的每个克隆型。在一个实施例中，B细胞或T细胞的样本包含至少五十万个细胞，并且在另一个实施例中，这种样本包含至少一百万个细胞。

[0081] 可使用如下常规技术从外周血获得核酸样本，如Innis et al,editors,PCR Protocols(Academic Press,1990)(Innis等人编辑，《PCR方案》，学术出版社，1990年)等。例如，可使用如下常规技术从血样中分离出白细胞，如RosetteSep试剂盒(RosetteSep kit)(加拿大温哥华干细胞技术有限公司(Stem Cell Technologies,Vancouver,Canada))。血样的体积范围可为100μL至10mL；在一个方面，血样体积在100μL至2mL范围内。然后，可使用适用于本发明方法的如下常规技术从此类血样中提取DNA和/或RNA，如DNeasy血液和组织试剂盒(DNeasy Blood&Tissue Kit)(加利福尼亚州瓦伦西亚市凯杰公司(Qiagen,Valencia,CA))。任选地，可使用如下常规技术进一步分离白细胞的子集(如淋巴细胞)，如荧光激活细胞分选(FACS)(加利福尼亚州圣何塞市的BD公司(Becton Dickinson,San Jose,CA))、磁激活细胞分选(MACS)(加利福尼亚州奥本市的美天旎生物技术公司(Miltenyi Biotec,Auburn,CA))等。例如，记忆B细胞可经由表面标记CD19和CD27分离。

[0082] 还可使用如下常规技术从外周血样本提取无细胞DNA，如Lo et al,U.S.patent 6,258,540(Lo等人，美国专利6,258,540)；Huang et al,Methods Mol.Biol.,444:203-208(2008)(Huang等人，《分子生物学方法》，2008年，第444卷，第203-208页)等，这些文献以引用方式并入本文。以非限制性实例来说，可将外周血收集在EDTA管中，然后可通过离心将其分级分离成血浆、白细胞和红细胞组分。可根据制造商的方案，使用QIAamp血液DNA小量提取试剂盒(QIAamp DNA Blood Mini Kit)(加利福尼亚州瓦伦西亚市凯杰公司(Qiagen,Valencia,CA))或类似试剂盒从(如0.5至2.0mL)的无细胞血浆部分提取DNA。

[0083] 在一个方面，用于生成克隆型谱的淋巴细胞样本足够大，使得基本上每个具有不同克隆型的T细胞或B细胞都出现在其中。在一个实施例中，取用的样本以99％的概率包含出现频率为0.001％或更高的群体中的每个克隆型。在另一个实施例中，取用的样本以99％的概率包含出现频率为0.0001％或更高的群体中的每个克隆型。在另一个实施例中，取用的样本以99％的概率包含出现频率为0.00001％或更高的群体中的每个克隆型。在其他实施例中，取用的样本以95％的概率包含出现频率为0.001％或更高的群体中的每个克隆型。在另一个实施例中，取用的样本以95％的概率包含出现频率为0.0001％或更高的群体中的每个克隆型。在另一个实施例中，取用的样本以95％的概率包含出现频率为0.00001％或更高的群体中的每个克隆型。在又另一个实施例中，B细胞或T细胞的样本包含至少五十万个细胞，并且在另一个实施例中，这种样本包含至少一百万个细胞。

[0084] 每当获取样本的材料来源不足时，例如临床研究样本等，可通过非偏倚技术扩增来自材料的DNA，所述非偏倚技术如全基因组扩增(WGA)、多重置换扩增(MDA)或类似的技术，例如Hawkins et al,Curr.Opin.Biotech.,13:65-67(2002)(Hawkins等人，《生物技术

新见》,2002年,第13卷,第65-67页);Dean et al,Genome Research,11:1095-1099(2001)(Dean等人,《基因组研究》,2001年,第11卷,第1095-1099页);Wang et al,Nucleic Acids Research,32:e76(2004)(Wang等人,《核酸研究》,2004年,第32卷,第e76页);Hosono et al,Genome Research,13:954-964(2003)(Hosono等人,《基因组研究》,2003年,第13卷,第954-964页)等。

[0085] 血样受到特别关注,并且可使用如下常规技术获得,如Innis et al,editors,PCR Protocols(Academic Press,1990)(Innis等人编辑,《PCR方案》,学术出版社,1990年)等。例如,可使用如下常规技术从血样中分离出白细胞,如RosetteSep试剂盒(RosetteSep kit)(加拿大温哥华干细胞技术有限公司(Stem Cell Technologies,Vancouver,Canada))。血样的体积范围可为100μL至10mL;在一个方面,血样体积在100μL至2mL范围内。然后,可使用适用于本发明方法的如下常规技术从此类血样中提取DNA和/或RNA,如DNeasy血液和组织试剂盒(DNeasy Blood&Tissue Kit)(加利福尼亚州瓦伦西亚市凯杰公司(Qiagen,Valencia,CA))。任选地,可使用如下常规技术进一步分离白细胞的子集(如淋巴细胞),如荧光激活细胞分选(FACS)(加利福尼亚州圣何塞市的BD公司(Becton Dickinson,San Jose,CA))、磁激活细胞分选(MACS)(加利福尼亚州奥本市的美天旎生物技术公司(Miltenyi Biotec,Auburn,CA))等。

[0086] 由于每个个体的适应性免疫细胞的DNA以及其相关RNA转录物中存在识别重组,因此可以所提供的本发明的方法对RNA或DNA测序。来自T细胞或B细胞的重组序列被称为克隆型,所述重组序列编码T细胞受体或免疫球蛋白分子的链,或者它们的一部分。所述DNA或RNA可对应于来自编码抗体的T细胞受体(TCR)基因或免疫球蛋白(Ig)基因的序列。例如,所述DNA和RNA可对应于编码TCR中α、β、γ或δ链的序列。在大部分T细胞中,TCR为包含α链和β链的异源二聚体。TCRα链经VJ重组生成,而β链受体经V(D)J重组生成。人类中的TCRβ链有48个V区段、2个D区段和13个J区段。若干碱基可缺失,其他的(称作N核苷酸和P核苷酸)被添加到两个连接处中的每一个上。在少数T细胞中,TCR由γ和δ链组成。TCRγ链经VJ重组生成,而TCRδ链经V(D)J重组生成(Kenneth Murphy,Paul Travers,and Mark Walport,Janeway's Immunology 7th edition,Garland Science,2007(Kenneth Murphy、Paul Travers和Mark Walport,《Janeway免疫学》第七版,加兰科学出版社,2007年),该文献全文以引用方式并入本文)。

[0087] 在本发明方法中分析的DNA和RNA可与一些序列相对应,所述序列编码具有恒定区(α、δ、ε、γ或μ)的重链免疫球蛋白(IgH)或者具有恒定区λ或K的轻链免疫球蛋白(IgK或IgL)。每个抗体具有两条相同的轻链和两条相同的重链。每条链由恒定(C)区和可变区构成。对于重链而言,可变区由可变(V)区段、多样性(D)区段和连接(J)区段构成。基因组中存在若干不同序列对这些区段中的每种类型进行编码。特定的VDJ重组事件发生在B细胞的发育期间,将该细胞标记为产生特定的重链。轻链的多样性以类似的方式产生,不过没有D区,所以只有VJ重组。体细胞突变常常发生在重组位点附近,导致若干核苷酸的添加或缺失,进一步提高了B细胞所产生的重链和轻链的多样性。然后,不同的重链和轻链造成了B细胞所产生抗体可能的多样性。重链和轻链的可变区有助于形成抗原识别(或结合)区或位点。产生这种多样性的方式还有体细胞高频突变过程,所述体细胞高频突变可在对一些表位产生特异性应答后进行。

[0088] 如上所述，根据本发明，可选择引物来生成扩增子，所述扩增子包含重组核酸的来自淋巴细胞或来自组织(如血液)的无细胞核酸的部分。这些部分在本文中可称为“体细胞重排区”。体细胞重排区可包含来自发育中的或已充分发育的淋巴细胞的核酸，在发育中的淋巴细胞中，免疫基因的重排尚未完成，从而未形成具有(例如)完整V(D)J区的分子。示例性的不完整体细胞重排区包括不完整的IgH分子(如，只包含D-J区的分子)、不完整的TCR分子(如，只包含D-J区的分子)以及非活性IgK(例如，包含Kde-V区)。

[0089] 核酸群体的扩增

[0090] 在一些实施例中，可根据常规的多重聚合酶链反应(PCR)选择第一组引物和第二组引物的引物序列。例如，对选择引物以及对编码各种免疫受体链的核酸进行多重PCR的指导见于以下参考文献，这些参考文献以引用方式并入:Faham and Willis,U.S.patents 8,236,503and 8,628,927(Faham和Willis，美国专利8,236,503和8,628,927)；Morley,U.S.patent 5,296,351(Morley，美国专利5,296,351)；Gorski,U.S.patent 5,837,447(Gorski，美国专利5,837,447)；Dau,U.S.patent 6,087,096(Dau，美国专利6,087,096)；Van Dongen et al,U.S.patent publication 2006/0234234(Van Dongen等人，美国专利公布2006/0234234)；European patent publication EP 1544308B1(欧洲专利公布EP 1544308B1)；Van Dongen et al,Leukemia,17:2257-2317(2003)(Van Dongen等人，《白血病》，2003年，第17卷，第2257-2317页)等。对多重PCR的指导可见于Henegariu et al,BioTechniques,23:504-511(1997)(Henegariu等人，《生物技术》，1997年，第23卷，第504-511页)，以及类似参考文献。在一些实施例中，可对引物进行选择，使得终产物中扩增序列的频率基本上与起始反应混合物中序列的频率相同。此类引物选择可包括对引物长度、引物结合位点和引物浓度的选择。如上所述，根据所选的生成序列读段和附接序列标签的方法，多重化程度可能有很大差别。

[0091] 在一些实施例中，扩增靶核酸的步骤包括对靶核酸进行线性扩增，所述线性扩增可(例如)通过重复使一组引物(例如，第一组“上游”或“正向”引物)退火、延伸引物、使延伸链从模板解链的循环来进行，使得延伸链的量按与循环数呈线性函数的关系扩增。换句话讲，扩增步骤包括通过重复延伸一组引物来复制靶多核苷酸(即靶多核苷酸的至少一条链)。在一些实施例中，可在完成这种朝一个方向的单次或重复延伸后进行去除未延伸引物的步骤，并在其他方向上对另一组引物(例如，第二组“下游”或“反向”引物)进行单次或重复延伸。

[0092] 根据在分析中扩增的免疫受体链核酸的数目和种类，第一引物组和第二引物组中的引物数目可能有很大差别。在一些实施例中，可使用各种链的共有引物。在其他实施例中，可为待扩增的每个不同靶多核苷酸设计特异性引物。通常，第一引物组和第二引物组都各自包含多个引物。在一些实施例中，第一引物组或第二引物组中的多个引物为至少50个引物；在其他实施例中，第一引物组或第二引物组中的多个引物为至少100个引物；在其他实施例中，第一引物组或第二引物组中的多个引物为至少150个引物；在其他实施例中，第一引物组或第二引物组中的多个引物为至少200个引物；在其他实施例中，第一引物组或第二引物组中的多个引物为至少250个引物。第一组中的引物数目与第二组中的引物数目可相同，也可不同。

[0093] 在一些实施例中，对第一组和第二组的引物进行选择，使得克隆型的长度为至少

30个核苷酸;在其他实施例中,对第一组和第二组的引物进行选择,使得克隆型的长度在30至500个核苷酸的范围内;在其他实施例中,对第一组和第二组的引物进行选择,使得克隆型的长度在30至400个核苷酸的范围内;在其他实施例中,对第一组和第二组的引物进行选择,使得克隆型的长度在30至300个核苷酸的范围内;在其他实施例中,对第一组和第二组的引物进行选择,使得克隆型的长度在30至200个核苷酸的范围内。

[0094] 示例性的PCR扩增方案可见于van Dongen et al,Leukemia,17:2257-2317(2003)(van Dongen等人,2003年,第17卷,第2257-2317页)或van Dongen et al,U.S.patent publication 2006/0234234(van Dongen等人,美国专利公布2006/0234234),这些参考文献以引用方式并入。简而言之,一个示例性方案如下:反应缓冲液:ABI缓冲液II(ABI Buffer II)或ABI Gold缓冲液(ABI Gold Buffer)(加利福尼亚州圣地亚哥市美国生命技术公司(Life Technologies,San Diego,CA));50μL最终反应体积;100ng样本DNA;引物各10pmol(按下文所述的那样进行调整以平衡扩增);终浓度为200μM的dNTP;终浓度为1.5mM的MgCl2(根据靶序列和聚合酶进行优化);Taq聚合酶(1-2U/管);循环条件:95℃下预激活7min;60℃下退火;循环时间:30s变性;30s退火;30s延伸。可在本发明方法中用于扩增的聚合酶可商购获得,包括(例如)Taq聚合酶、AccuPrime聚合酶(AccuPrime polymerase)或Pfu。可根据优选保真性还是效率来选择要使用的聚合酶。

[0095] 可在最初步骤中使用实时PCR、picogreen染色(picogreen staining)、纳流控电泳(如LabChip)或UV吸收测量来估计样本中可扩增材料的功能量。

[0096] 在一个方面,进行本发明所述的多重扩增,以使起始群体中的序列相对量与扩增群体(或者说扩增子)中的序列相对量基本相同。也就是说,多重扩增在样本群体成员序列中扩增偏倚最低的情况下进行。在一个实施例中,如果扩增子中各相对量为起始样本中其值的五倍以内,则此类相对量基本相同。在另一个实施例中,如果扩增子中各相对量为起始样本中其值的两倍以内,则此类相对量基本相同。如下文中更充分讨论的,可使用常规技术检测和校正PCR中的扩增偏倚,从而可为预定的组库选择一组PCR引物,该PCR引物组会提供任何样本的非偏倚扩增。

[0097] 在一些实施例中,可通过进行两阶段扩增(例如上文引用的Faham和Willis中所述)来避免扩增偏倚,其中,在第一阶段(或者说初级阶段)完成少量扩增循环(例如2-5或2-10或2-15个循环),该阶段使用的引物具有与靶序列不互补的尾。所述尾包含添加到初级扩增子序列末端的引物结合位点,使得这些位点可用于只使用单一正向引物和单一反向引物的第二阶段扩增中,从而消除扩增偏倚的主要成因。在第二阶段扩增起始前,将第一阶段的未延伸引物从反应混合物中去除,或者将其灭活。在一些实施例中,初级PCR的循环数将少到(如2-10)足以将由不同引物引起的差异扩增减到最小。然后用一对引物完成次级扩增,其消除了差异扩增的来源。在一些实施例中,将初级PCR的一小部分(如1%)的反应体积直接加入次级PCR反应混合物中。在一些实施例中,分配给第一阶段扩增和第二阶段扩增的循环总共为至少三十五个。

[0098] 在一些实施例中,可将内标与样本重组核酸混合,并在同一反应中扩增。内标为序列和浓度均已知的核酸。例如,它们可能是天然核酸的克隆拷贝,所述天然核酸编码免疫受体链的部分;或者,它们可能是合成的核酸。在一些实施例中,对内标的长度和碱基组成进行选择以代表扩增中的特定免疫受体链。通过监测扩增后内标的相对浓度变化,可检测扩

增偏倚，并且可确定非偏倚扩增的条件。例如，可改变引物长度、位置和浓度来将扩增产物的偏倚减至最小。在一些实施例中，多种内标被用于反应中；在一些实施例中，2至50种不同的内标被用于反应中；在其他实施例中，2至25种不同的内标被用于反应中；并且在一些实施例中，2至10种不同的内标被用于反应中。在一些实施例中，通过测量扩增产物中不同靶核苷酸(例如，全部克隆型或所选克隆型或内标)的序列的相对频率来确定扩增偏倚。在其他实施例中，可通过对所选核酸(如两种或更多内标)进行实时定量PCR来确定扩增偏倚的有无或水平。内标还可用于对原始样本中不同克隆型的数目进行定量。用于此类分子计数的技术是众所周知的，如Brenner et al,U.S.patent 7,537,897(Brenner等人，美国专利7,537,897)，其以引用方式并入本文。

[0099] 生成序列读段

[0100] 本发明的方法可使用任何用于对核酸测序的高通量技术。优选地，这种技术能够以经济有效的方式产生大量的序列数据，从这些数据中可确定至少1000种克隆型，优选地，从这些数据中可确定至少10,000至1,000,000种克隆型。DNA测序技术包括：经典的双脱氧测序反应(Sanger法，该法使用标记的终止子或引物并在板或毛细管中进行凝胶分离)、使用可逆终止的标记核苷酸进行的边合成边测序、焦磷酸测序、454测序、等位基因特异性杂交至标记寡核苷酸探针文库、使用等位基因特异性杂交至标记克隆文库然后连接而进行的边合成边测序、在聚合步骤中掺入标记核苷酸的实时监测、聚合酶克隆测序和SOLiD测序。分离分子的测序最近已通过使用聚合酶和连接酶进行连续或单次延伸反应，以及通过使用探针文库进行单次或连续的差异杂交得到证实。人们已对许多平行克隆序列进行了这些反应，包括在目前超过1亿个平行序列的商业应用中的实证。因此这些测序方法可以用于研究T细胞受体(TCR)和/或B细胞受体(BCR)组库。在本发明的一个方面，采用了高通量测序方法，该方法包括在对各分子进行平行测序的固体表面上将各分子在空间上隔开的步骤。此类固体表面可包括无孔表面(诸如Solexa测序中使用的无孔表面，例如Bentley et al,Nature,456:53-59(2008)(Bentley等人，《自然》，2008年，第456卷，第53-59页)，或完整基因组测序中使用的无孔表面，例如Drmanac et al,Science,327:78-81(2010)(Drmanac等人，《科学》，2010年，第327卷，第78-81页))、可包括微珠或粒子结合模板的孔阵列(所述模板诸如454测序中使用的模板，例如Margulies et al,Nature,437:376-380(2005)(Margulies等人，《自然》，2005年，第437卷，第376-380页)，或离子激流测序中使用的模板(美国专利公布2010/0137143或2010/0304982))、微机械薄膜(诸如SMRT测序中使用的微机械薄膜，例如Eid et al,Science,323:133-138(2009)(Eid等人，《科学》，2009年，第323卷，第133-138页))，或微珠阵列(如SOLiD测序或聚合酶克隆测序中使用的微珠阵列，例如Kim et al,Science,316:1481-1414(2007)(Kim等人，《科学》，2007年，第316卷，第1481-1414页))。在另一方面，此类方法包括在固体表面上将各分子在空间上隔开之前或之后，扩增已被隔开的分子。前扩增可包括基于乳液的扩增(诸如乳液PCR)或滚环扩增。

[0101] 特别值得关注的是使用可逆终止子进行边合成边测序的方法，诸如基于Solexa的测序，在该测序方法中，在固体表面上将各模板分子在空间上隔开，之后将所述分子通过桥式PCR平行地扩增，形成单独的克隆群体或簇，然后测序，诸如在Bentley等人(上文有引用)和制造商说明书(例如，TruSeq™样本制备试剂盒和数据表(TruSeq™Sample Preparation Kit and Data Sheet)，2010年得自加利福尼亚州圣地亚哥市的亿明达公司(Illumina,

Inc.,San Diego,CA))中所描述的;并且在以下参考文献中也有所描述:美国专利6,090,592、6,300,070、7,115,400和EP0972081B1;这些参考文献以引用方式并入。在一个实施例中,在固体表面上排布并扩增的各分子形成的簇密度为至少10^5簇/cm^2;或密度为至少5×10^5簇/cm^2;或密度为至少10^6簇/cm^2。基于Solexa的测序还能够从簇中的同一靶序列(或模板)生成两个序列读段,所述两个序列读段分别来自靶序列的相对两端。在一些实施例中,可将这种序列读段对结合起来,并且在随后的分析中作为单个序列读段处理,或者可分别处理这种序列读段对,但应考虑到其来自同一簇。有时,来自同一模板的序列读段对被称为"配对",而从一个模板的两端进行测序的过程被称为"双向"测序。在一些实施例中,使用可逆终止的标记核苷酸进行的边合成边测序的步骤包括:为模板的每个簇或克隆群体生成单个序列读段,以及为模板的每个簇或克隆群体生成多个序列读段(包括但不限于配对)。在另外的实施例中,当为模板的每个簇或克隆群体生成多个序列读段时,可将此类多个序列读段结合起来,形成用于后续分析(诸如合并步骤)的单个有效的序列读段。

[0102] 在一个方面,使用以下步骤获得个体样本的基于序列的克隆型谱:(a)从个体的T细胞和/或B细胞获得核酸样本;(b)将源自这种核酸样本的各分子在空间上隔开,所述各分子包含由样本中的核酸产生的至少一个模板,该模板包括体细胞重排区或该区的一部分,各分子中的每一个均能够产生至少一个序列读段;(c)对所述在空间上隔开的各分子进行测序;以及(d)确定核酸样本中核酸分子的不同序列的丰度,生成克隆型谱。在一个实施例中,每个体细胞重排区包含V区和J区。在另一个实施例中,测序步骤包括为每个确定的克隆型生成多个序列读段。在其他实施例中,测序步骤包括将来自多个序列读段的信息或数据结合起来,形成每个克隆型。在一些实施例中,这种结合步骤可通过合并序列读段来进行,如Faham和Willis在美国专利8,628,927(该专利据此以引用方式并入以用于本教导内容)中所述,或通过使用序列标签来进行,如Faham等人在美国专利公布2013/0236895A1(该专利公布据此以引用方式并入以用于本教导内容)中所述。在另一个实施例中,测序步骤包括对在空间上隔开的各分子中的每一个进行双向测序,产生至少一个正向序列读段和至少一个反向序列读段。

[0103] 就后一个实施例进一步来说,所述至少一个正向序列读段和至少一个反向序列读段具有重叠区域,如此一来,通过这种序列读段之间的反向互补关系可确定该重叠区域的碱基。在又另一个实施例中,每个体细胞重排区包括V区和J区,测序步骤还包括由各核酸分子的一个或多个正向序列读段和至少一个反向序列读段确定各核酸分子中每一个的序列,所述反向序列读段起始于J区中的某个位置,并在其相关的V区方向上延伸。在另一个实施例中,各分子包含选自以下分子的核酸:完整IgH分子、不完整IgH分子、完整IgK分子、完整IgK非活性分子、TCRβ分子、TCRγ分子、完整TCRδ分子和不完整TCRδ分子。在另一个实施例中,测序步骤包括生成具有单调递减的质量得分的序列读段。在另一个实施例中,上述方法包括以下步骤:(a)从个体的T细胞和/或B细胞获得核酸样本;(b)将源自这种核酸样本的各分子在空间上隔开,所述各分子包括模板的嵌套组,每个模板由样本中的核酸产生并且包括体细胞重排区或该区的一部分,每个嵌套组能够产生多个序列读段,每个序列读段在同一方向上延伸并且起始于核酸上产生该嵌套组的不同位置;(c)对所述在空间上隔开的各分子进行测序;以及(d)确定核酸样本中核酸分子的不同序列的丰度,生成克隆型谱。在一个实施例中,测序步骤包括为每个嵌套组生成多个序列读段。在另一个实施例中,每个体细

胞重排区包含V区和J区，所述多个序列读段中的每一个起始于V区中的不同位置，并在其相关的J区方向上延伸。

[0104] 在一个方面，对于来自个体的每个样本，本发明方法中使用的测序技术每次运行产生至少1000个克隆型的序列；在另一方面，这种技术每次运行产生至少10,000个克隆型的序列；在另一方面，这种技术每次运行产生至少100,000个克隆型的序列；在另一方面，这种技术每次运行产生至少500,000个克隆型的序列；在另一方面，这种技术每次运行产生至少1,000,000个克隆型的序列。在另一方面，对于每个个体样本，这种技术每次运行产生介于100,000个至1,000,000个之间的克隆型的序列。在上述每个方面，每次运行的每个克隆型由至少10个序列读段确定。

[0105] 所提供的本发明方法中使用的测序技术每次读数可产生约30bp、约40bp、约50bp、约60bp、约70bp、约80bp、约90bp、约100bp、约110bp、约120bp、约150bp、约200bp、约250bp、约300bp、约350bp、约400bp、约450bp、约500bp、约550bp、或约600bp。

[0106] 由序列数据确定克隆型

[0107] 在本发明的一些实施例中，使用序列标签来确定克隆型，在其他实施例中，将序列标签与序列读段合并步骤结合使用来确定克隆型。在将单个独特的序列标签附接至基本上每个不同的靶多核苷酸的实施例中，可直接使用序列标签确定克隆型。在这样的实施例中，首先基于序列读段的序列标签对序列读段进行分组来确定样本的克隆型。可通过常规的序列比对方法来完成这种分组。选择比对方法的指南可见于Batzoglou,Briefings in Bioinformatics,6:6-22(2005)(Batzoglou,《生物信息学简讯》,2005年,第6卷,第6-22页)，该参考文献以引用的方式并入。在将序列读段对应于独特序列标签分组之后，可分析相关克隆型的序列，从而确定来自样本的克隆型的序列。图4A示出了确定与独特序列标签相关的克隆型的序列(SEQ ID NO:2)的示例性比对及方法。在这个例子中，将十一个序列读段通过各自的序列标签4302进行比对，比对后，对序列读段(示为1,2,3,4,...n)的克隆型部分4304每个位置处的核苷酸进行比较。例如，位置6(4306)处的核苷酸为t、t、g、t、t、t、t、t、t、c、t；也就是说，九个碱基判读为t，一个为“g”(4308)(SEQ ID NO:3)，还有一个为“c”(4310)(SEQ ID NO:4)。在一个实施例中，某一位置的克隆型序列的正确碱基判读为占多数的碱基的类型。在位置6(4306)的例子中，碱基判读是“t”，因为t是该位置大多数序列读段中的核苷酸。在其他实施例中，可考虑其他因素来确定克隆型序列的正确碱基判读，诸如所述序列读段的碱基判读的质量得分、相邻碱基的种类，等等。如上所述确定了克隆型后，即可组合出克隆型谱，其包括样本中每个不同克隆型的丰度或频率。

[0108] 在一些实施例中，可使用含有序列标签的引物进行不止一个延伸步骤，从而增加样本中在扩增前标有序列标签的靶多核苷酸的比例。在此类实施例中，在存在含有序列标签的引物的情况下进行所述不止一个延伸步骤，会使靶多核苷酸和/或其拷贝被多个不同的序列标签所标记。不同的序列标签的多少取决于在存在含有序列标签的引物的情况下进行的延伸步骤数目、扩增反应的效率、正向引物和反向引物是否都具有序列标签，等等。在一些此类实施例中，不同的序列标签的数目在2至15的范围内、或在2至10的范围内、或2至5的范围内。在一些此类实施例中，扩增后，样本中每个靶多核苷酸的拷贝可分为多个群组或子集，其中每个群组或子集的成员标记有相同的序列标签，而每个不同群组或子集的成员标记有不同的序列标签；也就是说，同一组的成员具有相同的序列标签，而不同组的成员具

有不同的序列标签。换句话讲，扩增后，在一些实施例中，样本中靶多核苷酸的每个拷贝会被标记上两个不同的序列标签之一；或者在其他实施例中，样本中靶多核苷酸的每个拷贝会被标记上三个不同的序列标签之一；或者在其他实施例中，样本中靶多核苷酸的拷贝会被标记上四个不同的序列标签之一；以此类推。在这些实施例中，可通过在序列标签比对后进行合并步骤确定克隆型，所述合并步骤基于共同起源为真的可能性将一个群组内的序列读段视为源于相同的亲本序列，所述可能性随错误率、相对频率等变化。图4B示出了来自这种实施例的序列读段。在一个方法中，首先通过通用序列标签4402将序列读段分组，在图中示为三个群组4420、4422和4424。在一些实施例中，分析每个群组内的序列4404，以确定该组的共有序列；例如，如上所述，在每个核苷酸位置处碱基可称为多数碱基或最高频率碱基等等。然后该群组的共有序列可彼此合并，从而确定克隆型。

[0109] 在一些实施例中，可在分析样本中几乎所有核酸群体的方法中实施本发明的上述方面。这种方法可以包括以下步骤：(a)获得包含核酸群体的样本；(b)将序列标签附接到该群体的核酸，以形成标签-核酸缀合物，其中该群体的至少一个核酸或其拷贝附接有不同的序列标签；(c)扩增标签-核酸缀合物；(d)对标签-核酸缀合物进行测序以产生序列读段，所述序列读段具有错误率并且包含核酸序列和标签序列；(e)比对具有类似标签序列的序列读段以形成具有相同序列标签的序列读段群组；(f)合并各群组的序列读段以确定所述核酸的序列，其中每当所述序列读段群组有至少95％的可能性不同时，将序列读段群组合并为不同的序列；以及(g)通过确定序列的水平来确定该群体的序列谱。当应用这种方法分析重组核酸群体时，可通过以下步骤实施这种方法：(a)从个体获得包含T细胞和/或B细胞和/或无细胞DNA的样本；(b)将序列标签附接到来自样本的T细胞受体基因或免疫球蛋白基因的重组核酸分子，以形成标签-核酸缀合物，其中来自样本的至少一个重组核酸或其拷贝附接有不同的序列标签；(c)扩增标签-核酸缀合物；(d)对标签-核酸缀合物的样本进行测序以提供序列读段，所述序列读段具有错误率并且包含标签序列和重组核酸序列；(e)比对具有类似标签序列的序列读段以形成具有相同序列标签的序列读段群组；(f)合并各群组的序列读段以确定克隆型，其中每当所述序列读段群组有至少95％的可能性不同时，将序列读段群组合并为不同的序列；以及(g)通过确定克隆型的水平来确定样本的克隆型谱。

[0110] 在上述实施例和本文所公开的其他实施例中，对标签核酸缀合物进行测序的步骤包括对来自扩增子的标签核酸缀合物样本进行测序。通常，这种样本为代表性样本，其中，原始样本(即组织样本、血样等等)中靶多核苷酸的相对频率在来自扩增反应产物的标签-核酸缀合物样本中得到保持。在一些分析编码免疫受体分子的重组核酸群体的实施例中，标签-核酸缀合物样本包含至少10^4个标签-核酸缀合物；在其他实施例中，此类样本包含至少10^5个标签-核酸缀合物；在其他实施例中，此类样本包含至少10^6个标签-核酸缀合物；在其他实施例中，此类样本包含至少10^7个标签-核酸缀合物。

[0111] 合并序列读段

[0112] 在多个序列标签附接至原始重组核酸或其拷贝的实施例中，可进行合并序列读段(或各群组的共有序列读段)的步骤，来确定克隆型。如果测序技术没有错误，那么将给定样本的一组序列读段缩减成一组不同的克隆型并记录每个克隆型的序列读段数目将会是轻而易举。然而，在存在测序错误的情况下，每个真正的克隆型都被具有不同数量的其序列相关错误的序列读段“云”所围绕。随着测序错误“云”与序列空间中克隆型的距离增加，“云”

的密度下降。有多种算法可用于将序列读段转换成克隆型。在一个方面，合并序列读段(即，归并已确定具有一个或多个测序错误的候选克隆型)至少取决于三个因素：对于每个进行比较的克隆型所获得的序列数、序列中不相同的碱基数、以及在序列不一致位置处的测序质量得分。在一些实施例中，可根据预期错误率和错误的二项分布构建并评估似然比。例如，有两个克隆型，其中一个具有150个序列读段，而另一个具有2个序列读段，在测序质量差的区域这两个克隆型之间存在一个差异，因为二者很可能是由测序错误产生的，所以可将二者合并起来。另一方面，有两个克隆型，其中一个具有100个序列读段，而另一个具有50个序列读段，二者之间存在两个差异，因为考虑到这两个克隆型不太可能是由测序错误产生的，所以不将二者合并起来。在一些实施例中，可使用下述算法由序列读段来确定克隆型。图5A中例示出了这些概念的其中一些。在合并步骤的一些实施例中，序列读段首先被转换成候选克隆型。这种转换取决于所采用的测序平台。对于产生高质量得分的长序列读段的平台，可将序列读段或它的一部分直接作为候选克隆型。对于产生低质量得分的较短序列读段的平台，则需要进行一些比对和分组步骤，才能将一组相关的序列读段转换成候选克隆型。例如，在一些实施例中，对于基于Solexa的平台，候选克隆型是由如上所述来自多个簇(如10个或更多个簇)的成对序列读段的集合产生的。

[0113] 如图5A所示，可在序列空间中绘制候选克隆型的频率，图中为了便于说明，将此类序列空间简化为一维(水平轴)。垂直轴表示每个候选克隆型的频率大小，或者说log(读段计数)，或一些类似量度。在图中，候选克隆型由各种符号530来表示。根据本发明的一个实施例，是否合并两个候选克隆型，取决于它们各自的频率或读段计数(如上所述)、二者之间的碱基差异数(差异越多，越不可能合并)，以及在它们各自序列相异位置处的碱基质量得分(质量得分越高，越不可能合并)。候选克隆型可按照其相应频率的顺序来考虑。图5A示出的候选克隆型1(532)、候选克隆型7(534)和候选克隆型11(536)是具有最高的三个频率的三个候选克隆型。与每个这种候选克隆型相关的是其他在序列上相近、但频率较小的候选克隆型，诸如(i)对于候选克隆型1(532)，候选克隆型2(538)和锥形(540)所包围的候选克隆型3、4、5和6与之相关；(ii)对于候选克隆型7(534)，锥形(542)所包围的候选克隆型8、9和10与之相关；(iii)对于候选克隆型11，锥形(544)所包围的候选克隆型12与之相关。锥形代表可能性边界，在该边界内，较低频率的候选克隆型会与较高频率的候选克隆型1、7或11之一合并。这种可能性边界是附近的候选克隆型(1附近是3、4、5和6；7附近是8、9和10；11附近是12)的频率与其在序列空间中与相应较高频率候选克隆型的距离的函数。候选克隆型2(538)在锥形(540)之外；因此，不会将它与候选克隆型1(532)合并。同样，(合并的)可能性边界示为锥形是因为具有较高频率的候选克隆型比具有较低频率的候选克隆型更有可能是真正不同的克隆型，而较低频率处的多个差异比起较高频率处的多个差异更有可能是错误。

[0114] 可使用二项分布和单个碱基错误概率的简单模型对每个候选克隆型周围的序列读段云进行建模。可由映射V区段和J区段或由克隆型查找算法本身，通过自洽性和收敛性推断出后一个错误模型。构建模型是为了得到具有读段计数C2和E个错误(相对于序列X)的给定“云”序列Y在零模型下为具有理想读段计数C1的真实克隆型序列X的一部分的概率，所述零模型即X是序列空间的该区域内唯一的真实克隆型。根据参数C1、C2和E作出是否将序列Y合并为克隆型X的决定。对于任何给定的C1和E，预先计算最大值C2，以便确定是否合并

序列Y。选择C2的最大值，使得在Y是克隆型X的一部分的零假设下，在将具有错误E的所有可能的序列Y整合到序列X附近后，不能合并Y的概率小于某个P值。P值可控制算法的运行状态并使得合并或多或少地得到允许。

[0115] 如果由于序列Y的读段计数高于合并为克隆型X的阈值C2而使序列Y不能合并为克隆型X，则序列Y成为用于形成另一个克隆型(诸如图5A中的候选克隆型2(538))的候选序列。实施这种原则的算法还将确保“更靠近”序列Y(被认为独立于X)的任何其他序列Y2、Y3等不会聚集于X。“靠近”的概念既包括关于Y和X的错误计数，还包括X和Y的绝对读取计数，即，它以与上述用于围绕克隆型X的错误序列云的模型相同的方式进行模拟。这样，如果“云”序列正好“靠近”一个以上克隆型，则可以正确地归属于它们正确的克隆型。因此，参见图5A，如果候选克隆型2被认为真正不同于候选克隆型1(532)，那么专用程序或子算法将提供规则来确定应合并为候选克隆型1(532)和2(538)、候选克隆型4和5以及1和2之间中的哪一者(如果任一者均可)。

[0116] 在一个实施例中，算法从具有最高读段计数的序列X开始，以从上至下的方式进行。此序列形成第一克隆型。如果邻近序列的计数低于预先计算的阈值(参见上文)，则将其合并到此克隆型中，如果邻近序列的读数高于阈值或“更接近”另一个未合并的序列，则置之不理。在搜索最大错误计数内的所有邻近序列后，即停止将读段合并到克隆型X中的过程。解读其读段以及所有已经合并到其中的读段，并将其从可用于产生其他克隆型的读段列表中去除。然后转向下一个具有最高读段计数的序列。如上所述，将邻近的读段合并到该克隆型中，并且此过程一直持续到不再有读段计数高于给定阈值的序列为止，例如直到计数大于1的所有序列都已用来形成克隆型。

[0117] 如上所述，在上述算法的另一个实施例中，可增加进一步的检验，用于确定是否将候选序列Y合并到现有克隆型X中，该检验考虑了相关序列读段的质量得分。在序列Y和X不同的情况下，确定序列Y的平均质量得分(取所有具有序列Y的读段的平均值)。如果平均得分高于预定值，那么该差异更有可能说明该序列为真正不同的克隆型，不应被合并；如果平均得分低于该预定值，那么序列Y更可能是由测序错误产生的，因此应该合并到X中。

[0118] 上述用于合并候选克隆型的算法的成功实施取决于是否能从一些输入序列X中有效地找到错误数量小于E(即，小于一些序列距离量度)的所有序列。此问题可使用序列树来解决。这种序列树的实施方式具有一些不寻常的特征，因为树的节点不限于候选克隆型的DNA序列的单一碱基，如图5D所示。这些节点可具有任意长的序列，以便更有效地使用计算机存储器。

[0119] 将给定样本的所有读段置于序列树中。每个叶节点指向与它相关联的读段。通过从叶节点到根节点逆向穿过序列树，得到候选克隆型的独特序列。将第一个序列放在具有一个根节点和一个叶节点的简易树中，该简易树包含读段的全部序列。然后逐一添加序列。对于每个添加的序列而言，在读数和现有树之间的共同序列的最后一点处形成新的分支，或者如果该树已包含此序列，则将该读段添加到现有的叶节点上。将所有读段放入树中后，可轻松使用该树实现以下目的：1)找出最高读段计数：通过读段计数来分选叶节点，可找出具有最多读段的叶节点(即，序列)，接着找出具有较低数量的读段的叶节点；2)找出相邻叶：对于任何序列，可搜索树中相对于该序列而言错误小于X的所有路径。路径始于根部，并分支为多个单独的路径继续沿着序列树行进。记录每条路径沿着序列树延伸时的当前错误

计数。当错误计数超过允许的最大值时，终止给定路径。这样，该树的大部分会尽可能早地得到修正。这种方法可有效地从任何给定序列中找到错误在X内的所有路径(即，所有叶)。

[0120] 图5B的流程图更详细地示出了上述概念的特征。通过对从T细胞或B细胞的样本中提取的重组核酸进行测序获得序列数据，并从该序列数据中获得一组候选克隆型。在一个方面，候选克隆型各自包含NDN区以及部分V区和J区。将这些序列置于数据结构(550)中，该数据结构可以是序列树。在一个实施例中，作为生成一组候选克隆型的一部分，还可针对已知V区和已知J区构建序列树(未在图5B中示出)。然后可通过序列树将组成候选克隆型的序列读段映射到这些已知的序列或与这些已知的序列进行比对，以有效确定候选克隆型的最可能已知的V序列和J序列。转到图5B，生成候选克隆型后，数据结构(诸如序列树)被构建为能够用于将真正的克隆型与包含实验误差或测量误差(诸如测序错误)的候选克隆型区分开来。从数据结构(例如，序列树)中选择当前候选克隆型中具有最高出现频率的候选克隆型($HFCC_k$)(552)；换句话讲，$HFCC_k$是循环k中具有最高拷贝数或最高读数计数的候选克隆型。接着，识别相邻的较低频率的候选克隆型(LFCC)(554)；也就是说，识别距离在D_k内的候选克隆型。在本发明的一个方面，使用序列树进行此识别，这将允许对相对较短(<300bp)的序列进行有效序列比较。

[0121] 在一个实施例中，使用例如由Gusfield公开的动态规划(如上文所提及)进行比较或序列比对。在另一个实施例中，这种动态规划为带状动态规划，其中不考虑与所选HFCC相差超过预定距离的序列，通过这种方式加快计算速度。可根据许多不同的标准或特性来比较候选$HFCC_k$和$LFCC_j$。在一个方面，如上所述，根据至少两个特性来比较候选克隆型：(i)频率或读段计数和(ii)序列差异。在另一个方面，如上所述，根据至少三个特性来比较候选克隆型：(i)频率或读段计数、(ii)序列差异和(iii)出现差异时碱基的质量得分或度量。在一个实施例中，序列差异包括碱基置换；在另一个实施例中，序列差异包括碱基置换、缺失和插入。后一实施例尤其适用于通过不使用终止子的边合成边测序方法(诸如454测序仪和离子激流测序仪)来生成序列数据。这种测序方法通过信号振幅来区分不同大小的均聚物延伸；因此，这种方法中的碱基读出程序容易发生插入和缺失错误，这是因为相差一个核苷酸的均聚物在信号水平上的差异程度随均聚物大小的增加而急剧下降(也就是说，2-mer很容易与3-mer区分开，但8-mer与9-mer却很难区分)。在一个方面，可使用函数(本文称为“合并似然函数”)来比较HFCC和LFCC，诸如判定框(558)中所示的$P(HFCC_k,LFCC_j,D,Q)$函数，具体取决于如上所述的数量(i)至(iii)。这种函数可采用许多不同的形式，但P值一般随(i)、(ii)和(iii)的变化发生如下变化：P值优选地随HFCC的频率以及HFCC频率与LFCC频率的比值单调上升，使得HFCC频率与LFCC频率的比值越高，将要合并到HFCC的似然LFCC就越高。同样，P值优选地随HFCC和LFCC的序列差异程度单调下降，使得HFCC和LFCC之间的差异(例如，由将一个序列变为另一个序列的置换、插入或缺失的最低数量测得)越大，将要合并到HFCC的似然LFCC越低。最终，P值优选地随HFCC和LFCC出现序列差异的位置的质量得分的增加而单调下降，使得质量得分越高，将要合并到HFCC的似然LFCC越低。

[0122] 当HFCC和LFCC的序列差异出现在不止一个位置时，可通过多种不同方法合并不同位置处的质量得分。在一个实施例中，每当存在多个这样的差异时，以平均值表示多个质量得分，该平均值可以是未加权平均数，也可以是加权平均数。图5C示出了示例性函数P，用于计算给定序列差异的不同质量值(曲线a至e)。如图5C所示，每当HFCC处于约200个读段计数

的水平(570)时，如果通过曲线(a)确定质量得分，那么读段计数小于约50的任何LFCC(572)将被合并到HFCC。函数P的自变量D是序列$HFCC_k$和$LFCC_j$之间的距离的量度，并且其值可随着分析的进行而在循环间变化。(符号“k”指示具有下标“k”的常数的值可取决于计算循环k。)在一个实施例中，$D=D_k$，使得其值为循环数的函数。在另一个实施例中，D＝D(HFCC频率)，使得其值为HFCC频率的函数，与循环数无关。例如，随着HFCC频率的降低，则待比较的候选项之间的距离减小。在一个实施例中，D为$HFCC_k$与$LFCC_j$之间的汉明距离(Hamming distance)；然而，也可使用其他距离量度。在一个实施例中，D_k为k的非递增函数；并且在另一个实施例中，D_k为k的递减函数。在一些实施例中，随着循环数的增加或者随着HFCC频率的下降而降低D的量值是有利的，这是因为随着计算的进行候选克隆型的频率越来越低，其中这些候选克隆型中的大部分为单例对象(singleton)，使得序列距离(而不是频率差异)成为主要比较。通过随着计算的进行减小D，减少了与远离的低频率候选克隆型的无效比较，从而加快计算速度。函数P可以是复杂的表达式，具体取决于所考虑因素的数量。图5C示出针对P的一个实施例所计算的值，所述值与在针对不同质量得分给定HFCC读段计数的情况下合并LFCC的读段计数阈值相关，如上所述。曲线“a”至“e”表示不同质量得分的关系(其中曲线“a”对应于最高的质量得分)。

[0123] 回到图5B，如果$P<P_k$，则不将$LFCC_j$与$HFCC_k$合并，并且选择另一个LFCC(560)。如果$P>P_k$，则将$LFCC_j$与$HFCC_k$合并(562)，在这种情况下选择另一个LFCC(566)，除非没有留下待评估的其他LFCC(564)。如果不再有待评估的LFCC(564)，则从数据结构(例如，序列树)中去除当前的$HFCC_k$(包括合并到$HFCC_k$的所有LFCC)(568)。这样的去除在图5D-5E的简易序列树(590)中示出。在所述图中，序列树(590)中的路径(592)(用虚线指示)对应于HFCC(596)，所述HFCC与LFCC(598)合并。在合并之后，从序列树(590)去除路径(592)在阴影区域(599)中的区段，以得到图5E所示的缩减序列树(597)，所述缩减序列树用于在随后的计算中找出相邻的LFCC(554)。在所述去除后，如果满足停止标准(570)，则完成克隆型的确定。在一个实施例中，停止标准(570)为是否已处理最后的非单例对象候选克隆型(552)。在另一个实施例中，停止标准(570)为所选HFCC的频率或读段计数是否低于与单个淋巴细胞对应的HFCC频率或读段计数。在本发明的方法的一个方面，扩增步骤可导致样本中的每个淋巴细胞由相同克隆型的多个拷贝表示；因此，在一个实施例中，每当HFCC具有的读段计数数量低于与单个淋巴细胞对应的数量时，则停止计算。在一些实施例中，这样的读段计数(或候选克隆型拷贝)的数量为至少10；在另一个实施例中，此数量为至少20；在另一个实施例中，此数量为至少30；在另一个实施例中，此数量为至少40。如果未满足停止标准，则选择下一个HFCC(572)。在图5B的流程图中总结的分析步骤可以使用任何合适的编程语言如C、C++、Java、C#、Fortran、Pascal等执行。

[0124] 根据本发明的一个方面，上述确定克隆型和/或克隆型谱(clonotype profile)的方法包括以下步骤：(a)根据通过高通量核酸测序获得的序列读段来形成重组免疫分子的数据结构；(b)每当任何较低频率的候选克隆型的频率低于预定频率值并且最高频率的候选克隆型与所述较低频率的候选克隆型之间的序列差异低于预定差异值时，将最高频率的候选克隆型与此类较低频率的候选克隆型合并以形成一种克隆型；(c)从所述数据结构中去除合并的候选克隆型；以及(d)重复步骤(b)和(c)，直到形成克隆型谱为止。在一个实施例中，数据结构为序列树。

[0125] 根据本发明的另一个方面，可通过以下步骤来进行上述确定克隆型的方法：(a)通过各自具有V区、NDN区和J区的重组免疫分子的组库提供一组序列读段，其中对于每个这样的分子，至少一个序列读段涵盖这类分子的NDN区的至少一部分；(b)由涵盖NDN区的至少一部分的序列读段形成具有表示候选克隆型的叶的序列树，每一叶及其对应的候选克隆型具有一定频率；(c)每当任何较低频率的候选克隆型的频率低于预定频率值并且最高频率的候选克隆型与所述较低频率的候选克隆型之间的序列差异低于预定差异值时，将最高频率的候选克隆型与此类较低频率的候选克隆型合并以形成具有最高频率的候选克隆型的序列的一种克隆型；(d)从所述序列树中去除对应于合并的候选克隆型的叶；以及(e)重复步骤(c)和(d)，直到较低频率的候选克隆型的最高频率低于预定的停止值。在一个实施例中，所述形成步骤还包括选择最高频率的候选克隆型，并识别与所选最高频率候选克隆型的序列差异小于预定差异值的所有所述较低频率的候选克隆型，以形成合并子集。因此，在此类实施例中，可以限制合并操作必须比较的LFCC的总数(仅考虑在预定差异值内的LFCC)。此类值为取决于具体应用的过程输入，例如组库大小、使用了多少计算时间，等等。如上所述，用于决定是否将HFCC与LFCC合并的函数可具有多种形式。在一个一般方面，对于合并步骤，此类函数可具有以下特性：其取决于HFCC的频率、LFCC的频率、HFCC与LFCC之间的序列差异(所述序列差异可表示为常规的字符串差异量度，诸如汉明距离)，以及HFCC和LFCC发生差异的一个或多个核苷酸位置的质量得分；以使得所述函数(i)随着HFCC频率与LFCC频率的比值的增大而单调上升，(ii)随着HFCC与LFCC之间的序列差异的增大而单调下降，以及(iii)随着一个或多个核苷酸位置的质量得分的增加而单调下降。也就是说，关于特性(iii)，越确定HFCC和LFCC是不同的(例如，因为碱基判读中存在高置信水平)，则越不可能合并HFCC和LFCC。

[0126] 在一些实施例中，对合并似然函数进行选择，使得每当序列读段有至少95％的可能性是不同的时，则将此类序列读段合并为不同的克隆型(或靶多核苷酸，诸如重组核酸)；在其他实施例中，对合并似然函数进行选择，使得每当序列读段有至少99％的可能性是不同的时，则将此类序列读段合并为不同的克隆型；在其他实施例中，对合并似然函数进行选择，使得每当序列读段有至少99.9％的可能性是不同的时，则将此类序列读段合并为不同的克隆型。如上所述，在一些实施例中，合并似然函数取决于所用的测序化学反应的错误率、所比较的序列读段中不一致核苷酸的数量以及所比较的序列读段的相对频率；在另一个实施例中，合并似然函数取决于所用的测序化学反应的错误率、所比较的序列读段中不一致核苷酸的数量、所比较的序列读段的相对频率以及不一致核苷酸的质量得分。在上文中，预定频率值和预定差异值的选择是取决于具体应用的设计选择。影响此类选择的因素可包括生物学详细信息、实施速度等等。

[0127] 监测应用

[0128] 在一个方面，本发明涉及用于通过确定样本中表征疾病或与疾病相关的核酸的存在、缺失和/或水平来监测微小残留病灶的方法。在一些实施例中，此类核酸为体细胞重组核酸，或者克隆型，所述核酸与癌前病症或癌症病症(诸如，淋巴或骨髓增生性疾病)相关，并且其可用于监测疾病或病症的状态。此类核酸，尤其是克隆型，可用于在治疗后监测癌症的微小残留病灶，其中此类监测的结果是决定继续治疗、停止治疗还是以其他方式改进治疗的关键因素。对于许多恶性淋巴和骨髓肿瘤，在治疗前获取诊断性的组织样本，诸如外周

血样本或骨髓样本，利用所述样本生成克隆型谱(“诊断性克隆型谱”)。对于淋巴或骨髓增生性疾病，在诊断性样本之前往往不知道哪些免疫受体链与疾病或病症的淋巴或骨髓克隆相关。因此，在现行方法中，必须对编码不同候选免疫受体链的不同重组核酸进行多个单独的扩增和测序，以便鉴定与患者的疾病或病症相关的克隆型。通过从此类扩增和测序工作产生的克隆型谱鉴定一种或多种与疾病相关的克隆型(即，“相关克隆型”)。通常，选取克隆型谱中具有最高频率的克隆型作为相关克隆型。在本发明的一个方面，通过对编码多个不同免疫受体链的重组核酸的若干部分进行单一反应的方式提供较大规模的多重扩增，大大减少了鉴定相关克隆型所需的单独扩增和测序运行的次数。在一些实施例中，所述“多个”在2至4个单独的免疫受体链的范围内；并且在其他实施例中，所述“多个”在2至3个单独的免疫受体链的范围内。更具体地讲，在一些实施例中，在BCR链中，通过单个多重反应来扩增以下序列：编码包括VDJ区域的至少一部分的IgH、包括DJ区域的至少一部分的IgH、以及IgK的重组核酸；并且在其他实施例中，在TCR链中，通过单个多重反应来扩增以下序列：TCRβ、TCRδ和TCRγ。

[0129] 在治疗之后，并且优选地在实现癌症的完全缓解之后，周期性地评估这种相关克隆型或核酸的存在、缺失或频率，以便基于治疗后的克隆型谱或核酸谱中相关核酸或克隆型(或所涉及的克隆型)的存在或频率增高来确定是缓解得到保持还是肿瘤复发。也就是说，在治疗之后，基于相关克隆型或特征性核酸的存在、缺失或频率来评估癌症的微小残留病灶。如上所述，当这种相关克隆型常见于或对应于缺乏足够的多样性的重排受体区段(使得非癌细胞可共享所述克隆型)时，这种克隆型在治疗后克隆型谱中的存在可产生复发的假阳性指示。

[0130] 本发明的方法适用于监测任何增生性疾病，在所述增生性疾病中，编码免疫受体或其一部分的重排核酸可用作疾病所涉及的细胞的标志物。在一个方面，本发明的方法适用于淋巴和骨髓增生性疾病。在另一方面，本发明的方法适用于淋巴瘤和白血病。在另一方面，本发明的方法适用于监测滤泡性淋巴瘤、慢性淋巴细胞白血病(CLL)、急性淋巴细胞白血病(ALL)、慢性骨髓性白血病(CML)、急性骨髓性白血病(AML)、霍奇金氏淋巴瘤和非霍奇金氏淋巴瘤、多发性骨髓瘤(MM)、意义未明的单克隆丙种球蛋白病(MGUS)、套细胞淋巴瘤(MCL)、弥漫性大B细胞淋巴瘤(DLBCL)、骨髓增生异常综合征(MDS)、T细胞淋巴瘤等等的MRD(微小残留病灶)。在具体实施例中，本发明的方法尤其适用于监测ALL、MM或DLBCL的MRD(微小残留病灶)。

[0131] 在一些实施例中，患者样本如血液或骨髓经过诊断性分析，以鉴定多个免疫受体链中的哪个免疫受体链可包含由疾病的克隆产生的克隆型(即，相关克隆型)。一旦确定相关克隆型的免疫受体链，就可针对该特定的免疫受体链进行后续监测分析。例如，在一些实施例中，诊断性分析可在同一反应中生成多个BCR链(如IgH(VDJ)、IgH(DJ)和IgK)的基于序列的克隆型谱。如果相关克隆型为IgH(VDJ)链，则后续监测分析可仅生成IgH(VDJ)的克隆型谱。在一些实施例中，诊断样本中的测序深度可与监测样本中的测序深度不同。“测序深度”是指所分析的用于构建克隆型谱的序列读段的总数量。对于癌症如白血病和淋巴瘤，因为在治疗之前先对患者样本进行了诊断性分析，所以样本中相关克隆型的频率或水平通常较高并且很容易鉴定。例如，频率超过预定水平的任何克隆型可定义为相关克隆型。在使用其他患者指标的情况下此预定水平可以不同；而预定水平通常可在2％至5％的范围内；或

者在一些实施例中为5％。因此，在一些实施例中，所进行的测序的深度为能可靠检测出以1％或2％或更高的频率存在的克隆型所需的测序深度。在一些实施例中，诊断样本的测序深度产生至少10,000个序列读段；或者在其他实施例中，产生至少100,000个序列读段；在又其他实施例中，诊断样本的测序深度产生至少10^6个序列读段。在一些实施例中，监测样本的测序深度为至少100,000个序列读段；在其他实施例中，监测样本的测序深度为至少10^6个序列读段。

[0132] 在一些实施例中，可通过根据从患者连续获得的样本(或组织样本)产生克隆型谱来监测患者所患的淋巴增生性疾病，如白血病或者淋巴瘤。可如上文所述生成此克隆型谱。在一些实施例中，这种监测可通过以下步骤实施：(a)从个体获取包含T细胞和/或B细胞和/或无细胞DNA的样本；(b)将序列标签附接到来自样本的T细胞受体基因或免疫球蛋白基因的重组核酸分子，以形成标签-核酸缀合物，其中至少一个重组核酸或其拷贝附接有不同的序列标签；(c)扩增标签-核酸缀合物；(d)对标签-核酸缀合物的样本进行测序以提供序列读段，所述序列读段各自具有错误率并且各自包含标签序列和重组核酸序列；(e)比对具有类似标签序列的序列读段以形成具有相同序列标签的序列读段组；(f)合并所述组中的序列读段以确定克隆型，其中每当所述序列读段组有至少95％的可能性是不同的时，则将序列读段组合并成不同的重组核酸序列；(g)通过确定克隆型的水平来确定样本的克隆型谱；以及(h)确定克隆型谱中相关克隆型的水平。在一些实施例中，可在监测患者的过程中重复步骤(a)至(h)，以确定相关克隆型的水平是否证明疾病的复发。在一些实施例中，附接和扩增的步骤可包括下列步骤：(a)在反应混合物中，使第一组引物在引物延伸条件下与来自表达免疫受体的免疫细胞和/或无细胞DNA的重组核酸样本混合，其中第一组的每个引物均具有受体特异性部分，使得该受体特异性部分在预定位置处与不同重组核酸退火配对，然后延伸形成第一延伸产物，并且其中第一组中的每个引物均具有包含第一引物结合位点的5’非互补端；(b)从反应混合物中去除第一组的未延伸引物；以及(c)在引物延伸条件下向反应混合物添加第二组引物，其中第二组中的每个引物均具有受体特异性部分，使得该受体特异性部分在预定位置处与第一延伸产物退火配对，并具有包含第二引物结合位点的5’非互补端，第一组的引物和/或第二组的引物分别包含位于受体特异性部分与第一或第二引物结合位点之间的序列标签，并且其中第二组的每个引物延伸形成第二延伸产物，使得每个第二延伸产物均包含第一引物结合位点、第二引物结合位点、至少一个序列标签以及编码免疫细胞受体链的一部分的重组核酸。在一些实施例中，合并重组核酸的步骤包括：每当不同重组核酸的序列读段有至少99％的可能性是不同的时；并且在其他实施例中有至少99.9％的可能性是不同的时，则合并此类序列读段。

[0133] 本发明的方法还适用于监测患者体内的癌症微小残留病灶，包括非淋巴癌或非骨髓癌，所述微小残留病灶具有例如在所选癌基因组中的突变鉴定模式。此类突变模式，也就是含有此类突变的基因的存在、缺失和/或水平可指示疾病复发的可能性。在一些实施例中，用于此类监测的靶多核苷酸可为外显子、部分外显子、所选内含子和/或基因表达控制区，例如多种基因的启动子(在本文中称为“癌基因分子”)。可使用常规技术从组织样本中分离癌基因分子，诸如外显子捕获技术，例如TruSeq™外显子组富集试剂盒(加利福尼亚州圣地亚哥市的亿明达公司(Illumina,San Diego,CA))；Frampton et al,Nature Biotechnology,31(11):1023-1031(2013)(Frampton等人，《自然－生物技术》，第31卷第11

期第1023-1031页,2013年);等等。获得此类癌基因分子之后,附接序列标签以形成标签-核酸缀合物,根据本发明扩增标签-核酸缀合物并对其测序。

[0134] 最近的癌基因组测序研究表明,在不同的癌症之间、患有相同癌症的不同患者之间、同一肿瘤的各细胞之间以及同一患者的不同转移部位的各细胞之间存在显著的突变模式不一致;然而,在同一患者体内,异质癌细胞通常从共同的祖细胞进化而来,因此它们共享突变并且癌细胞之间的进化关系可以随时间推移通过一系列测量来认识,例如Vogelstein et al,Science,339:1546-1558(2013)(Vogelstein等人,《科学》,第339卷第1546-1558页,2013年);Ding et al,Nature,481(7382):506-510(2012)(Ding等人,《自然》,第481卷第7382期第506-510页,2012年);等等;因此,与在诊断样本中测得的癌症相关的突变模式为检测相同癌症的复发或所述癌症的克隆性进化版本提供了手段。

[0135] 癌基因分子可选自多种基因,包括但不限于表I中的基因。

[0136] 表I

[0137] 示例性癌基因

[0138]

ABL1	AKT1	ALK	APC	ATM
BRAF	CDH1	CSF1R	CTNNB1	EGFR
ERBB2	ERBB4	FBXW7	FGFR1	FGFR2
FGFR3	FLT3	GNA11	GNAC	GNAS
HNF1A	HRAS	IDH1	JAK2	JAK3
KDR	KIT	KRAS	MET	MLH1
MPL	NOTCH 1	NPM1	NRAS	PGGFRA
PIK3CA	PTEN	PTPN11	RB1	RET
SMAD4	SMO	SRC	STK	TP53
VHL				

[0139] 在一些实施例中,上述监测癌症的微小残留病灶的方法可包括以下步骤:(a)从个体获取组织样本;(b)将序列标签附接到样本中多个癌基因分子中的每一个,以形成标签-核酸缀合物,其中至少一个核酸或其拷贝附接有不同的序列标签,并且其中癌基因分子为该个体的癌症所特有的;(c)扩增标签-核酸缀合物;(d)对标签-核酸缀合物的样本测序,以提供具有错误率并且包含标签序列和癌基因序列的序列读段;(e)比对具有类似标签序列的序列读段以形成具有相同序列标签的序列读段组;(f)合并组中的癌基因序列以确定癌基因分子的序列,其中每当所述癌基因序列组有至少95%的可能性是不同的时,则将序列读段组合并成不同的癌基因分子;以及(g)在癌基因分子谱中检测所述个体的癌症所特有的癌基因分子的存在、缺失和/或水平。在一些实施例中,合并癌基因序列的步骤包括:每当不同癌基因分子的序列读段有至少99%的可能性是不同的时;并且在其他实施例中有至少99.9%的可能性是不同的时,则合并此类序列读段。

[0140] 使用序列标签检测遗留污染

[0141] 遗留污染对于包括核酸扩增在内的技术而言是一个重要的问题,例如Borst et al,Eur.J.Clin.Microbiol.Infect.Dis.,23(4):289-299(2004)(Borst等人,《欧洲临床微生物和感染性疾病杂志》,第23卷第4期第289-299页,2004年);Aslanzadeh,Ann.Clin.Lab.Sci.,34(4):389-396(2004)(Aslanzadeh,《临床和实验室科学年报》,第34卷第4期第389-396页,2004年);等等。当在分析样本时无意扩增了痕量的样本外来核酸并且其对测量结果产生影响时,出现此类污染。在更糟糕的情况下,患者的医学样本中的遗留

污染可导致对分析结果的假阳性判读。外来核酸可来自与具体患者无关的来源;例如,其可来自另一患者的样本。或者,外来核酸可来自与患者相关的来源;例如,其可来自过去在同一实验室中处理的同一患者的不同样本或者来自对过去在同一实验室中处理的同一患者的不同样本的分析反应。

[0142] 当测量高度复杂的相关核酸群体(诸如编码T细胞受体或免疫球蛋白等免疫分子的重组核酸群体)时,临床环境中的遗留污染尤其具有挑战性。所述挑战的存在是因为难以确定序列读段或克隆型是否为目标样本的真正多样性的构成部分或者序列读段或克隆型是否来源于外来核酸,诸如在同一实验室的相同类型的分析中所处理的另一患者的样本或同一患者的在先样本。在本发明的一个方面,此类遗留污染可使用序列标签来检测,这样不仅可根据序列读段确定克隆型,还可确定序列标签是源自当前样本还是源自另一样本。这是通过以下步骤实现的:保存关于从每个患者样本确定的序列标签的记录,然后每当进行后续测量时,将当前测量的序列标签与先前测量的序列标签进行比较。与克隆型相关的这些序列标签记录被便利地保存为大容量存储装置上的电子记录,这是由于存在大量来自每次测量的标签,而且这样便于搜索电子记录和使用常规算法对电子记录进行比较。如果存在匹配,那么在测量中所采用的序列标签群体足够大的前提下,最可能的原因是遗留污染。克隆型群体的大小与用于上文所述的取样标记的序列标签群体的相同示例性比率适用于检测遗留污染。在一个实施例中,此比率为100:1或更大。

[0143] 可使用多种搜索方法或算法来进行比较所测克隆型与数据库克隆型的步骤。许多常规序列比对和搜索算法为可公开获得的并且已经在下列以引用方式并入的参考文献中有所描述:Mount,Bioinformatics Sequence and Genome Analysis,Second Edition(Cold Spring Harbor Press,2004)(Mount,《生物信息学:序列与基因组分析》,第二版,冷泉港实验室出版社,2004年);Batzoglou,Briefings in Bioinformatics,6:6-22(2005)(Batzoglou,《生物信息学简报》,第6卷第6-22页,2005年);Altschul et al,J.Mol.Biol.,215(3):403-410(1990)(Altschul等人,《分子生物学杂志》,第215卷第3期第403-410页,1990年);Needleman and Wunsch,J.Mol.Biol.,48:443-453(1970)(Needleman和Wunsch,《分子生物学杂志》,第48卷第443-453页,1970年);Smith and Waterman,Advances in Applied Mathematics,2:482-489(1981)(Smith和Waterman,《应用数学进展》,第2卷第482-489页,1981年);等等。

[0144] 在一些实施例中,上述用于检测和测定样本中由源于不同样本的物质所带来的污染(如遗留污染)的方法可包括以下步骤:(a)从个体获得组织样本;(b)将序列标签附接到癌基因分子或重组核酸,以形成标签-核酸缀合物,其中至少一个核酸或其拷贝附接有不同的序列标签,并且其中癌基因分子为该个体的癌症所特有的;(c)扩增标签-核酸缀合物;(d)对该标签-核酸缀合物的样本测序以提供序列读段,所述序列读段各自具有错误率并且各自包含标签序列和癌基因序列或重组核酸序列;(e)将标签序列与单独测定的来自其他组织样本的标签序列进行比较;以及(f)通过一条或多条标签序列与从单独测定的来自其他组织样本的任何标签序列的一致性,确定是否存在污染和/或污染程度。一旦在分析中确定标签序列,则可将这些标签序列与根据对其他患者的分析所记录的标签序列数据库中的标签序列进行比较。这样的比较步骤可在分析时进行,或者此类步骤可回顾性地实施,例如在分析后的某一时段。在一个实施例中,将序列标签附接到来自患有淋巴增生性疾病如淋

巴癌的个体的组织样本(诸如,血液或骨髓)中的重组核酸。在另一个实施例中,将序列标签附接到癌基因分子,如上所述。

[0145] 在其中监测重组核酸以防组织样本交叉污染的其他实施例中,可实施如下附接和扩增步骤:(a)在反应混合物中,使第一组引物在引物延伸条件下与来自T细胞和/或无细胞DNA的重组核酸样本混合,其中第一组的每个引物均具有受体特异性部分,使得该受体特异性部分在预定位置处与不同重组核酸退火配对,然后延伸形成第一延伸产物,并且其中第一组中的每个引物均具有包含第一引物结合位点的5'非互补端;(b)从反应混合物中去除第一组的未延伸引物;(c)在引物延伸条件下向反应混合物添加第二组引物,其中第二组中的每个引物均具有受体特异性部分,使得该受体特异性部分在预定位置处与第一延伸产物退火配对,并具有包含第二引物结合位点的5'非互补端,第一组的引物和/或第二组的引物分别包含位于受体特异性部分与第一或第二引物结合位点之间的序列标签,并且其中第二组的每个引物延伸形成第二延伸产物,使得每个第二延伸产物均包含第一引物结合位点、第二引物结合位点、至少一个序列标签以及编码免疫受体链的一部分的重组核酸;以及(d)在反应混合物中进行聚合酶链反应以形成扩增子,所述聚合酶链反应使用特异于第一引物结合位点的正向引物和特异于第二引物结合位点的反向引物。

[0146] 试剂盒

[0147] 本发明包括用于执行本发明方法的多种试剂盒。在一些实施例中,试剂盒包括(a)一组正向引物和一组反向引物,这些引物用于在多重PCR中扩增编码多个免疫受体链的重组核酸,其中正向引物和/或反向引物各自具有靶特异性部分、序列标签和共同的引物结合位点;以及(b)引物去除元件,用于在至少第一次延伸后去除引物组中未被掺入的引物(即,未延伸的引物)。在一些实施例中,试剂盒还包括特异于共同的引物结合位点的通用引物。在一些实施例中,试剂盒还包括关于在本发明方法中使用试剂盒组分的书面说明。在一些实施例中,试剂盒还包括专用于扩增编码IgH(VDJ)、IgH(DJ)和IgK的重组核酸的正向和反向引物。在一些实施例中,试剂盒还包括专用于扩增编码TCRβ、TCRδ和TCRγ的重组核酸的正向和反向引物。在一些实施例中,试剂盒还包括内标,所述内标包含具有靶重组核酸的代表性长度和组成的多个核酸,其中内标以已知浓度提供。在一些实施例中,试剂盒包括单链核酸外切酶作为引物去除元件,诸如大肠杆菌核酸外切酶I。在一些实施例中,试剂盒包括能够按大小分选双链DNA的离心柱作为引物去除元件。

[0148] 虽然已结合若干具体示例性实施例描述了本发明,但是本领域的技术人员将认识到,在不脱离本发明的实质和范围的情况下可对本发明作出多种改变。除了上文论述的那些之外,本发明还可应用于多种传感器的具体实施和其他主题。

[0149] 定义

[0150] 除非本文另外明确定义,否则本文使用的核酸化学、生物化学、遗传学和分子生物学的术语和符号遵循本领域标准论文和教科书中的那些,例如Kornberg and Baker,DNA Replication,Second Edition(W.H.Freeman,New York,1992)(Kornberg和Baker,《DNA的复制》,第二版(W.H.弗里曼出版社,纽约,1992年));Lehninger,Biochemistry,Second Edition(Worth Publishers,New York,1975)(Lehninger,《生物化学》,第二版(沃思出版社,纽约,1975年));Strachan and Read,Human Molecular Genetics,Second Edition(Wiley-Liss,New York,1999)(Strachan和Read,《人类分子遗传学》,第二版(威立-利斯出

版公司,纽约,1999年));Abbas et al,Cellular and Molecular Immunology,6th edition (Saunders,2007)(Abbas等人,《细胞与分子免疫学》,第6版(桑德斯出版社,2007年))。

[0151] “比对”是指基于某一序列距离量度来比较测试序列(诸如,序列读段)与一个或多个参考序列以确定哪个参考序列或参考序列的哪一部分最接近的方法。比对核苷酸序列的示例性方法为史密斯-沃特曼(Smith Waterman)算法。距离量度可包括汉明距离、莱温斯坦距离(Levenshtein distance)等等。距离量度可包括与所比较序列的核苷酸的质量值有关的部分。

[0152] “扩增子”是指多核苷酸扩增反应的产物,即多核苷酸的克隆群体,其可为单链或双链的,所述多核苷酸是从一个或多个起始序列复制的。所述一个或多个起始序列可以是同一序列的一个或多个拷贝,或者它们可以是不同序列的混合物。扩增子可通过多种扩增反应产生,所述扩增反应的产物包括一个或多个起始核酸或靶核酸的复制物。在一个方面,产生扩增子的扩增反应是“模板驱动的”,因为反应物(无论是核苷酸还是寡核苷酸)的碱基配对在产生反应产物所需的模板多核苷酸中具有互补物。在一个方面,模板驱动的反应是使用核酸聚合酶进行的引物延伸或使用核酸连接酶进行的寡核苷酸连接。此类反应包括但不限于聚合酶链反应(PCR)、线性聚合酶反应、基于核酸序列的扩增(NASBA)、滚环扩增等,如在下列以引用方式并入本文的参考文献中所公开:Mullis等人,美国专利4,683,195、4,965,188、4,683,202、4,800,159(PCR);Gelfand等人,美国专利5,210,015(使用“taqman”探针的实时PCR);Wittwer等人,美国专利6,174,670;Kacian等人,美国专利5,399,491(“NASBA”);Lizardi,美国专利5,854,033;Aono等人,日本专利公布JP 4-262799(滚环扩增);等等。在一个方面,本发明的扩增子通过PCR产生。如果提供可以随着扩增反应的进行而测量反应产物的检测化反应,则扩增反应可以是“实时”扩增,例如,如下所述的“实时PCR”或如Leone et al,Nucleic Acids Research,26:2150-2155(1998)(Leone等人,《核酸研究》,第26卷第2150-2155页,1998年)和类似参考文献中所描述的“实时NASBA”。如本文所用,术语“扩增”是指进行扩增反应。“反应混合物”是指包含进行反应的所有必需反应物的溶液,其可包含但不限于在反应过程中将pH维持在选定水平的缓冲剂、盐、辅因子、清除剂等。

[0153] 如本文所用,“克隆性(clonality)”是指对谱系的克隆型之间克隆型丰度的分布偏斜至单个或几个克隆型的程度的度量。大致上,克隆性是克隆型多样性的相反度量。从描述物种-丰度关系的生态学可以获得许多度量或统计学数据,所述物种-丰度关系可用于根据本发明的克隆性度量,例如Pielou,An Introduction to Mathematical Ecology,(Wiley-Interscience,1969)(Pielou,《数理生态学简介》(威立国际科学出版社,1969年))中的第17章和第18章。在一个方面,本发明使用的克隆性度量是克隆型谱(即检测到的不同克隆型的数目和它们的丰度)的函数,使得在测定克隆型谱后,可以从克隆型谱计算克隆性以得到单个数字。一种克隆性度量是Simpson度量,其只是两个随机绘制的克隆型将会相同的概率。其他克隆性度量包括在Pielou(如上文所引用)中公开的基于信息的度量和McIntosh多样性指数。

[0154] “克隆型”是指编码免疫受体或其一部分的淋巴细胞的重组核酸。更具体地讲,“克隆型”是指通常从T细胞或B细胞提取但是也可来自于无细胞来源的重组核酸,其编码T细胞受体(TCR)或B细胞受体(BCR)或其一部分。在各种实施例中,克隆型可编码IgH的VDJ重排、

IgH的DJ重排、IgK的VJ重排、IgL的VJ重排、TCRβ的VDJ重排、TCRβ的DJ重排、TCRα的VJ重排、TCRγ的VJ重排、TCRδ的VDJ重排、TCRδ的VD重排、Kde-V重排等等的全部或一部分。克隆型还可编码涉及免疫受体基因(诸如Bcl1-IgH或Bcl1-IgH)的易位断点区。在一个方面,克隆型具有足够长的序列以表征或反映其所衍生自的免疫分子的多样性;因此,克隆型的长度可以变化很大。在一些实施例中,克隆型的长度在25至400个核苷酸的范围内;在其他实施例中,克隆型的长度在25至200个核苷酸的范围内。

[0155] “克隆型谱”是指衍生自淋巴细胞群体的不同克隆型的列表以及它们的相对丰度,其中例如相对丰度可表示为在给定群体中的频率(也就是说,基于0和1之间的数值)。通常,从组织样本获得淋巴细胞群体。术语“克隆型谱”涉及如诸如以下参考文献中所述的免疫“谱系”的免疫学概念,但是比此概念更广义:Arstila et al,Science,286:958-961(1999)(Arstila等人,《科学》,第286卷第958-961页,1999年);Yassai et al,Immunogenetics,61:493-502(2009)(Yassai等人,《免疫遗传学》,第61卷第493-502页,2009年);Kedzierska et al,Mol.Immunol.,45(3):607-618(2008)(Kedzierska等人,《分子免疫学》,第45卷第3期第607-618页,2008年);等等。术语“克隆型谱”包括编码免疫受体的重排核酸的大量列表和丰度,所述重排核酸可源自淋巴细胞的所选子集(例如,组织浸润的淋巴细胞、免疫表型的子集等),或者所述重排核酸可编码与完整的免疫受体相比具有较低多样性的免疫受体的部分。在一些实施例中,克隆型谱可包含至少10^3个不同的克隆型;在其他实施例中,克隆型谱可包含至少10^4个不同的克隆型;在其他实施例中,克隆型谱可包含至少10^5个不同的克隆型;在其他实施例中,克隆型谱可包含至少10^6个不同的克隆型。在此类实施例中,此类克隆型谱还可包含不同的克隆型中每一个的丰度或相对频率。在一个方面,克隆型谱是一组不同的重组核苷酸序列(以及它们的丰度),所述重组核苷酸序列在个体的淋巴细胞群体中分别编码T细胞受体(TCR)或B细胞受体(BCR)或其片段,其中,该组中的核苷酸序列具有与所述群体中基本上全部淋巴细胞不同的淋巴细胞或它们的克隆亚群一一对应性。在一个方面,对限定克隆型的核酸区段进行选择,使得它们的多样性(即,所述组中不同核酸序列的数量)大到足以使个体中基本上每个T细胞或B细胞或它们的克隆均携带此谱系的独特核酸序列。也就是说,样本的每个不同的克隆优选具有不同的克隆型。在本发明的其他方面,对应于谱系的淋巴细胞群体可以是循环的B细胞,或者可以是循环的T细胞,或者可以是上述群体中任一种的亚群,包括但不限于CD4+T细胞或CD8+T细胞,或者由细胞表面标记限定的其它亚群等。可基于一种或更多种细胞表面标记、大小、形态等,通过从特定组织(例如,骨髓或淋巴结等)采集样本,或通过从样本(例如,外周血)分选或富集细胞而获得此类亚群。在又其他方面,对应于谱系的淋巴细胞群体可来源于疾病组织(例如,肿瘤组织、感染的组织等)。在一个实施例中,包含人TCRβ链或其片段的克隆型谱包含在0.1×10^6至1.8×10^6的范围内,或在0.5×10^6至1.5×10^6的范围内,或在0.8×10^6至1.2×10^6的范围内的多个不同的核苷酸序列。在另一个实施例中,包含人IgH链或其片段的克隆型谱包含在0.1×10^6至1.8×10^6的范围内,或在0.5×10^6至1.5×10^6的范围内,或在0.8×10^6至1.2×106的范围内的多个不同的核苷酸序列。在具体实施例中,本发明的克隆型谱包含一组编码IgH链的V(D)J区的基本上全部区段的核苷酸序列。在一个方面,如本文所用的“基本上全部”是指每个区段具有0.001%或更高的相对丰度;或者在另一方面,如本文所用的“基本上全部”是指每个区段具有0.0001%或更高的相对丰度。在另一个具体实施例中,本发明的克隆型谱包含一

组编码TCRβ链的V(D)J区的基本上全部区段的核苷酸序列。在另一个实施例中，本发明的克隆型谱包含一组长度在25至200个核苷酸范围内并具有TCRβ链的V区、D区和J区的区段的核苷酸序列。在另一个实施例中，本发明的克隆型谱包含一组长度在25至200个核苷酸范围内并具有IgH链的V区、D区和J区的区段的核苷酸序列。在另一个实施例中，本发明的克隆型谱包含数量基本上相当于表达不同IgH链的淋巴细胞的数量的多个不同核苷酸序列。在另一个实施例中，本发明的克隆型谱包含数量基本上相当于表达不同TCRβ链的淋巴细胞的数量的多个不同核苷酸序列。在又另一个实施例中，“基本上相当于”是指克隆型谱将有99％的概率包含这样的核苷酸序列：该核苷酸序列编码由个体的淋巴细胞群体中的每一个以0.001％或更大的频率携带或表达的IgH或TCRβ或其一部分。在又另一个实施例中，“基本上相当于”是指核苷酸序列组库将有99％的概率包含这样的核苷酸序列：该核苷酸序列编码由每个淋巴细胞以0.0001％或更大的频率携带或表达的IgH或TCRβ或其一部分。在一些实施例中，克隆型谱得自包含105至107个淋巴细胞的样本。这种数量的淋巴细胞可从1-10mL的外周血样本获得。

[0156] “互补决定区”(CDR)是指免疫球蛋白(即，抗体)或T细胞受体的区域，在该区域中分子与抗原的构象互补，从而决定分子对特异性抗原的特异性以及与其的接触。T细胞受体和免疫球蛋白各自具有三个CDR：CDR1和CDR2存在于可变(V)结构域中，而CDR3包含部分V结构域、全部的多变结构域(D)(仅重链)和连接结构域(J)以及部分恒定结构域(C)。

[0157] “克隆型数据库”是指为了简化和加快搜索、对比和检索而格式化和排列的克隆型的集合。在一些实施例中，克隆型数据库包括编码免疫受体的相同区域或区段的克隆型的集合。在一些实施例中，克隆型数据库包含来自多个个体的克隆型谱的克隆型。在一些实施例中，克隆型数据库包含来自至少10个个体的至少10^4个克隆型的克隆型谱中的克隆型。在一些实施例中，克隆型数据库包含至少10^6个克隆型，或至少10^8个克隆型，或至少10^9个克隆型，或至少10^{10}个克隆型。克隆型数据库可以是包含克隆型的公共数据库，例如IMGT数据库(www.imgt.org)，如Nucleic Acids Research,31:307-310(2003)(《核酸研究》，第31卷第307-310页，2003年)中所述。克隆型数据库可为FASTA格式，并且可使用BLAST算法(如Altschul et al,J.Mol.Biol.,215(3):403-410(1990)(Altschul等人，《分子生物学杂志》，第215卷第3期第403-410页，1990年))或类似算法来搜索或对比克隆型数据库条目。

[0158] “合并”是指通过确定此类序列差异是由于实验或测量误差，而不是由于真的生物差异而将具有序列差异的两个候选克隆型当作同一克隆型处理。在一个方面，将更高频率候选克隆型的序列与更低频率候选克隆型的序列相比较，如果满足预定标准，则将更低频率候选克隆型的数目加到更高频率克隆型的数目中，然后忽略更低频率的候选克隆型。也就是说，将与更低频率候选克隆型相关的读段计数加到更高频率候选克隆型的读段计数中，并且更高频率候选克隆型和更低频率候选克隆型被当作同一克隆型处理；也就是说，在这两种候选克隆型之间所观察到的差异被确定为是因为错误(例如，测序错误、扩增错误等等)造成的。在一些实施例中，预定标准是取决于诸如所比较的候选克隆型的相对频率、候选克隆型发生差异的位置的数目、所述位置的质量得分等因素的似然函数。

[0159] “互补决定区”(CDR)是指免疫球蛋白(即，抗体)或T细胞受体的区域，在该区域中分子与抗原的构象互补，从而决定分子对特异性抗原的特异性以及与其的接触。T细胞受体和免疫球蛋白各自具有三个CDR：CDR1和CDR2存在于可变(V)结构域中，而CDR3包含部分V结

构域、全部的多变结构域(D)(仅重链)和连接结构域(J)以及部分恒定结构域(C)。

[0160] 如本文所用的“污染”是指在一个个体的组织样本中存在来自另一个个体的核酸。在一个方面,“污染”是指存在并非源自患者的核酸,这会影响对该患者的克隆型谱的解读。

[0161] “遗传鉴定”是指个体与来自所述个体的一个或多个基因座的遗传标记的一组值(或状态)之间的唯一对应性。

[0162] “遗传标记”是指基因座处的DNA的多态性区段,其可用于鉴定个体。遗传标记可通过其序列或通过其相邻或旁侧序列来鉴定。通常,在群体的不同个体中,遗传标记可具有多个序列或值。示例性遗传标记包括但不限于短串联重复序列(STR)、单核苷酸多态性(SNP)等等。DNA的多态性区段可以是基因组DNA,也可以是逆转录RNA。在一个实施例中,多态性区段为基因组DNA。在一个实施例中,通过使用常规技术扩增和测序来鉴定用于本发明的遗传标记。在另一个实施例中,在产生克隆型谱的过程中,将遗传标记与免疫分子一起扩增和测序。

[0163] “内标”是指在同一反应中以与一种或多种靶多核苷酸相同的方式进行处理的核酸序列,用于对样本中的靶多核苷酸进行绝对或相对定量。在一个方面,所述反应为扩增反应,诸如PCR。内标可为内源或外源的。也就是说,内标可以天然存在于样本中,也可以在反应前被添加至样本中。在一个方面,可将一种或多种外源的内标序列以预定浓度添加至反应混合物中以提供校准样,扩增的序列可以与所述校准样对比以确定样本中其相应的靶多核苷酸的量。对于本领域的普通技术人员而言,对外源内标的数量、序列、长度和其他特征的选择属于常规设计选择。内源内标在本文中也称为“参考序列”,是样本天然存在的序列,其对应于表现出恒定的和不依赖于细胞周期的转录水平的最低调控基因,例如Selvey et al,Mol.Cell Probes,15:307-311(2001)(Selvey等人,《分子细胞探针》,第15卷第307-311页,2001年)。示例性内标包括但不限于来自以下基因的序列:GAPDH、β_2-微球蛋白、18S核糖体RNA和β-肌动蛋白。

[0164] “试剂盒”是指用于递送用于实施本发明方法的材料或试剂的任何递送系统。在本发明方法的背景中,此类递送系统包括允许将反应试剂(例如,置于适当容器中的引物、酶、内标等)和/或支持材料(例如,缓冲液、关于如何进行分析的书面说明等)从一个位置到另一个位置储存、转运或递送的系统。例如,试剂盒包括装有相关反应试剂和/或支持材料的一个或多个封装件(例如,盒子)。这些内容物可一起或单独地被递送至目标受体。例如,第一容器可包含用于分析的酶,而第二容器包含引物。

[0165] “微小残留病灶”是指治疗后残余的癌细胞。该术语常与淋巴瘤和白血病的治疗结合使用。

[0166] “淋巴或骨髓增生性疾病”是指任何异常的增生性疾病,在所述疾病中编码一个或多个重排免疫受体的一种或多种核苷酸序列可被用作监测此类疾病的标记。“淋巴或骨髓肿瘤”是指淋巴细胞或骨髓细胞的恶性或非恶性的异常增殖。淋巴癌是一种恶性淋巴肿瘤。骨髓癌是一种恶性骨髓肿瘤。淋巴肿瘤和骨髓肿瘤是由淋巴或骨髓增生性疾病导致的,或与淋巴或骨髓增生性疾病相关,并且包括但不限于滤泡性淋巴瘤、慢性淋巴细胞白血病(CLL)、急性淋巴细胞白血病(ALL)、慢性骨髓性白血病(CML)、急性骨髓性白血病(AML)、霍奇金氏淋巴瘤和非霍奇金氏淋巴瘤、多发性骨髓瘤(MM)、意义未明的单克隆丙种球蛋白病(MGUS)、套细胞淋巴瘤(MCL)、弥漫性大B细胞淋巴瘤(DLBCL)、骨髓增生异常综合征(MDS)、T

细胞淋巴瘤等等，例如Jaffe et al,Blood,112:4384-4399(2008)(Jaffe等人，《血液》，第112卷第4384-4399页，2008年)；Swerdlow et al,WHO Classification of Tumours of Haematopoietic and Lymphoid Tissues(e.4th)(IARC Press,2008)(Swerdlow等人，WHO造血和淋巴组织肿瘤分类(第4版)(国际癌症研究机构(IARC)新闻稿，2008年))。如本文所用，“B细胞癌”是指涉及B细胞或者从B细胞发育的细胞(例如，浆细胞)的淋巴或骨髓肿瘤。同样地，“T细胞癌”是指涉及T细胞或者从T细胞发育的细胞的淋巴或骨髓肿瘤。

[0167] 关于参考序列和另一序列(“比较序列”)的比较中使用的“同源性百分比”、“同一性百分比”或类似术语是指在两个序列之间的最佳比对中，比较序列在相当于所示百分比的多个亚单元位置处与参考序列相同，对于多核苷酸比较而言亚单元是核苷酸，或者对于多肽比较而言亚单元是氨基酸。如本文所用，所比较序列的“最佳比对”是使亚单元间的匹配最大化并且使构建比对时所用空位的数目最小化的比对。同一性百分比可以用商购获得的算法实施方式诸如Needleman and Wunsch,J.Mol.Biol.,48:443-453(1970)(Needleman和Wunsch，《分子生物学杂志》，第48卷第443-453页，1970年)所述的算法(威斯康辛序列分析包的“GAP”程序，威斯康辛大学麦迪逊分校的遗传学计算机小组(“GAP”program of Wisconsin Sequence Analysis Package,Genetics Computer Group,Madison,WI))或类似算法来确定。本领域中用于构建比对和计算同一性百分比或其他相似性度量的其他软件包包括基于Smith and Waterman,Advances in Applied Mathematics,2:482-489(1981)(Smith和Waterman，《应用数学进展》，第2卷第482-489页，1981年)的算法的“BestFit”程序(威斯康辛序列分析包，威斯康辛大学麦迪逊分校的遗传学计算机小组(Wisconsin Sequence Analysis Package,Genetics Computer Group,Madison,WI))。换句话讲，例如，为了获得具有与参考核苷酸序列具有至少95％同一性的核苷酸序列的多核苷酸，参考序列中至多5％的核苷酸可缺失或被置换为另一种核苷酸，或者可以将占参考序列中核苷酸总数至多5％的核苷酸插入参考序列中。

[0168] “聚合酶链反应”或“PCR”是指用于通过DNA互补链的同时引物延伸而体外扩增特定DNA序列的反应。换句话讲，PCR是用于制备旁侧为引物结合位点的靶核酸的多个拷贝或复制物的反应，这种反应包括以下步骤的一次或多次重复：(i)使靶核酸变性，(ii)将引物与引物结合位点退火配对，以及(iii)在核苷三磷酸的存在下，利用核酸聚合酶延伸引物。如本文所用，术语“正向引物”和“上游引物”可互换使用，并且术语“反向引物”和“下游引物”可互换使用。还如本文所用，如果双链靶多核苷酸表现为其有义链为5’→3’从左到右取向，则正向引物将结合至左侧的反义链并且向右侧延伸，并且反向引物将结合到右侧的有义链并且向左侧延伸。通常，使用针对热循环仪中的各个步骤优化的不同温度来循环所述反应。具体温度、每个步骤的持续时间和步骤之间的改变速率取决于本领域的普通技术人员熟知的许多因素，如由以下参考文献所例示：McPherson et al,editors,PCR:A Practical Approach and PCR2:A Practical Approach(IRL Press,Oxford,1991and 1995,respectively)(McPherson等人编辑，《PCR:实用方法》和《PCR2:实用方法》(牛津大学出版社的IRL新闻稿，分别于1991年和1995年出版))。例如，在使用Taq DNA聚合酶的常规PCR中，双链的靶核酸可在>90℃的温度下变性，引物在50-75℃范围内的温度下退火，并且引物在72-78℃范围内的温度下延伸。术语“PCR”涵盖反应的派生形式，包括但不限于RT-PCR、实时PCR、巢式PCR、定量PCR、多重PCR等。反应体积在几百纳升(如200nL)到几百μL(如

200μL)的范围内。“逆转录PCR”或“RT-PCR”是指之前先将靶RNA转化成互补的单链DNA的逆转录反应的PCR,互补的单链DNA然后被扩增,例如Tecott等人,美国专利5,168,038中所述,该专利以引用方式并入本文。“实时PCR”是指随着反应进行而监测其反应产物即扩增子的量的PCR。实时PCR有多种形式,区别主要在于用于监测反应产物的检测化学反应,例如Gelfand等人,美国专利5,210,015(“taqman”);Wittwer等人,美国专利6,174,670和6,569,627(嵌入染料);Tyagi等人,美国专利5,925,517(分子信标);这些专利以引用方式并入本文。用于实时PCR的检测化学反应在Mackay et al,Nucleic Acids Research,30:1292-1305(2002)(Mackay等人,《核酸研究》,第30卷第1292-1305页,2002年)中有所描述,该文献也以引用方式并入本文。“巢式PCR”指两阶段PCR,其中第一PCR的扩增子成为第二PCR的样本,第二PCR使用一组新的引物,该新引物组中的至少一个结合至第一扩增子的内部位置。如本文所用,关于巢式扩增反应的“初始引物”是指用于生成第一扩增子的引物,并且“第二引物”是指用于生成第二或巢式扩增子的一条或多条引物。“多重PCR”是指其中多个靶序列(或单个靶序列和一个或多个参考序列)在同一反应混合物中同时进行的PCR,例如Bernard et al,Anal.Biochem.,273:221-228(1999)(Bernard等人,《分析生物化学》,第273卷第221-228页,1999年)(双色实时PCR)。通常,对所扩增的每个序列采用不同的引物组。“定量PCR”是指被设计用于测量样本或标本中一种或多种特定靶序列的丰度的PCR。定量PCR既包括此类靶序列的绝对定量,还包括其相对定量。使用一个或多个参考序列或内标进行定量测量,所述参考序列或内标可单独地或与靶序列一起分析。参考序列可以对于样本或标本是内源的或外源的,并且在后一种情况下可包含一个或多个竞争模板。典型的内源参考序列包括以下基因的转录物的区段:β-肌动蛋白、GAPDH、β_2-微球蛋白、核糖体RNA等等。用于定量PCR的技术为本领域的普通技术人员所熟知,如在以下以引用方式并入的参考文献中所例示:Freeman et al,Biotechniques,26:112-126(1999)(Freeman等人,《生物技术》,第26卷第112-126页,1999年);Becker-Andre et al,Nucleic Acids Research,17:9437-9447(1989)(Becker-Andre等人,《核酸研究》,第17卷第9437-9447页,1989年);Zimmerman et al,Biotechniques,21:268-279(1996)(Zimmerman等人,《生物技术》,第21卷第268-279页,1996年);Diviacco et al,Gene,122:3013-3020(1992)(Diviacco等人,《基因》,第122卷第3013-3020页,1992年);Becker-Andre et al,Nucleic Acids Research,17:9437-9446(1989)(Becker-Andre等人,《核酸研究》,第17卷第9437-9446页,1989年);等等。

[0169] “引物”是指天然的或合成的寡核苷酸,其在与多核苷酸模板形成双链体后,能够充当核酸合成的起始点并从其3’端沿着模板延伸以形成延伸的双链体。引物的延伸通常使用核酸聚合酶诸如DNA或RNA聚合酶进行。在延伸过程中添加的核苷酸的序列由模板多核苷酸的序列决定。通常,使用DNA聚合酶延伸引物。引物的长度通常在14至40个核苷酸的范围内,或在18至36个核苷酸的范围内。引物用于多种核酸扩增反应中,例如,使用单个引物的线性扩增反应,或采用两个或更多个引物的聚合酶链式反应。关于选择用于具体应用的引物的长度和序列的指导为本领域的普通技术人员所熟知,如由以引用方式并入的以下参考文献所证实:Dieffenbach,editor,PCR Primer:A Laboratory Manual,2nd Edition(Cold Spring Harbor Press,New York,2003)(Dieffenbach编辑,《PCR引物:实验室手册》,第2版(冷泉港实验室出版社,纽约,2003年))。

[0170] “质量得分”是指对将碱基正确分配到特定序列位置的概率的量度。对于特定情

况，诸如，针对由不同的测序化学反应、检测系统、碱基判读算法等读出的碱基来计算质量得分的多种方法是本领域的普通技术人员熟知的。一般来讲，质量得分值与正确的碱基判读的概率单调相关。例如，质量得分或Q为10可表示碱基有90％的可能性被正确判读，Q为20可表示碱基有99％的可能性被正确判读，等等。对于一些测序平台，尤其是使用边合成边测序化学反应的那些测序平台，平均质量得分作为序列读段长度的函数而下降，使得序列读段开始处的质量得分高于序列读段结束处的质量得分，这种下降是由于诸如不完全延伸、推进延伸(carry forward extension)、模板损耗、聚合酶损耗、加帽失败、脱保护失败等现象引起的。

[0171] “序列读段”是指由通过测序技术生成的序列或数据流而确定的核苷酸序列，这种确定例如借助于与该技术相关的碱基判读软件(如来自DNA测序平台供应商的碱基判读软件)实现。序列读段通常包括序列中各核苷酸的质量得分。通常，序列读段通过例如用DNA聚合酶或DNA连接酶沿着模板核酸延伸引物来获得。通过记录与此类延伸相关的信号诸如光学信号、化学信号(例如，pH变化)或电信号生成数据。将这种初始数据转换成序列读段。

[0172] “序列标签”(或“标签”)或“条码”是指附接到多核苷酸或模板分子的寡核苷酸，并被用于鉴定和/或追踪一个反应或一系列反应中的多核苷酸或模板。每个序列标签具有在本文中有时被称作“标签序列”的核苷酸序列。序列标签可附接到多核苷酸或模板的3’-端或5’-端，也可以插入此类多核苷酸或模板的内部以形成线性或环状缀合物，所述线性或环状缀合物在本文中有时被称为“加标签的多核苷酸”或“加标签的模板”或“标签-多核苷酸缀合物”、“标签-分子缀合物”等。序列标签可在大小和组成上有很大差异；以下以引用方式并入本文的参考文献提供了关于选择适用于具体实施例的序列标签组的指导：Brenner，美国专利5,635,400；Brenner和Macevicz，美国专利7,537,897；Brenner et al，Proc.Natl.Acad.Sci.，97:1665-1670(2000)(Brenner等人，《美国国家科学院院刊》，第97卷第1665-1670页，2000年)；Church等人，欧洲专利公布0 303 459；Shoemaker et al，Nature Genetics，14:450-456(1996)(Shoemaker等人，《自然遗传学》，第14卷第450-456页，1996年)；Morris等人，欧洲专利公布0799897A1；Wallace，美国专利5,981,179；等等。对特定标签长度和/或组成的选择可取决于若干因素，包括但不限于：用于解码标签的测序技术；明确鉴定一组靶多核苷酸所需的可区分标签的数目；为了确保可靠的鉴定例如避免交叉杂交或源于测序错误的错误鉴定，一组标签必须得如何不同；等等。在一些实施例中，序列标签的长度可各自在6至100个核苷酸的范围内，或10至100个核苷酸的范围内，或12至50个核苷酸的范围内，或12至25个核苷酸的范围内。在一些实施例中，使用序列标签组，其中一组中的每个序列标签具有与同一组的所有其他标签的序列相差至少四个碱基的独特核苷酸序列；在其他实施例中，使用序列标签组，其中一组中的每个标签的序列与同一组的所有其他标签的序列相差至少五个碱基；在另外的其他实施例中，使用序列标签组，其中一组中的每个标签的序列与同一组的所有其他标签的序列在其核苷酸上相差至少10％；或者在其他实施例中，它们的核苷酸相差至少25％；或者在其他实施例中，它们的核苷酸相差至少50％。

序列表

<110> 赛昆塔有限责任公司
Asbury, Thomas
Hervold, Kieran
Kotwaliwale, Chitra
Faham, Malek
Moorhead, Martin
Weng, Li
Wittkop, Tobias
Zheng, Jianbiao

<120> 用序列标签进行大规模生物分子分析
<130> 848US00 (37623-739.601)
<150> 61/841878
<151> 2013-07-01
[0173]
<150> 62/001580
<151> 2014-05-21
<160> 6
<170> PatentIn version 3.5

<210> 1
<211> 24
<212> DNA
<213> 人工序列
<220>
<223> 引物
<400> 1
agttctggct aacctgtaga gcca
24
<210> 2
<211> 24

<212> DNA

<213> 人工序列

<220>

<223> 引物

<400> 2

agttcgggct aacctgtcga gcca

24

<210> 3

<211> 24

<212> DNA

<213> 人工序列

<220>

<223> 引物

<400> 3

agttccggct aacctgtcga gcca

24

[0174]

<210> 4

<211> 22

<212> DNA

<213> 人工序列

<220>

<223> 引物

<220>

<221> misc_feature

<222> (1)..(22)

<223> n 是 a, c, g, 或 t

<400> 4

nnnnnnnnnn nnnnnnnnnn nn

22

<210> 5

<211> 12

<212> DNA

<213> 人工序列

<220>

<223> 引物

<400> 5

gtatttttt ct

12

<210> 6

[0175] <211> 13

<212> DNA

<213> 人工序列

<220>

<223> 引物

<400> 6

ttcaggggggg gct

13

序列表

<110> 赛昆塔有限责任公司
Asbury, Thomas
Hervold, Kieran
Kotwaliwale, Chitra
Faham, Malek
Moorhead, Martin
Weng, Li
Wittkop, Tobias
Zheng, Jianbiao

<120> 用序列标签进行大规模生物分子分析

<130> 848US00 (37623-739.601)

<150> 61/841878
<151> 2013-07-01

<150> 62/001580
<151> 2014-05-21

[0001]

<160> 6

<170> PatentIn version 3.5

<210> 1
<211> 24
<212> DNA
<213> 人工序列

<220>
<223> 引物

<400> 1
agttctggct aacctgtaga gcca 24

<210> 2
<211> 24
<212> DNA
<213> 人工序列

<220>

<223> 引物

<400> 2
agttcgggct aacctgtcga gcca 24

<210> 3
<211> 24
<212> DNA
<213> 人工序列

<220>
<223> 引物

<400> 3
agttccggct aacctgtcga gcca 24

<210> 4
<211> 22
<212> DNA
[0002] <213> 人工序列

<220>
<223> 引物

<220>
<221> misc_feature
<222> (1)..(22)
<223> n is a, c, g, or t

<400> 4
nnnnnnnnnn nnnnnnnnnn nn 22

<210> 5
<211> 12
<212> DNA
<213> 人工序列

<220>
<223> 引物

<400> 5
gtatttttt ct 12

[0003]

<210> 6
<211> 13
<212> DNA
<213> 人工序列

<220>
<223> 引物

<400> 6
ttcaggggggg gct 13

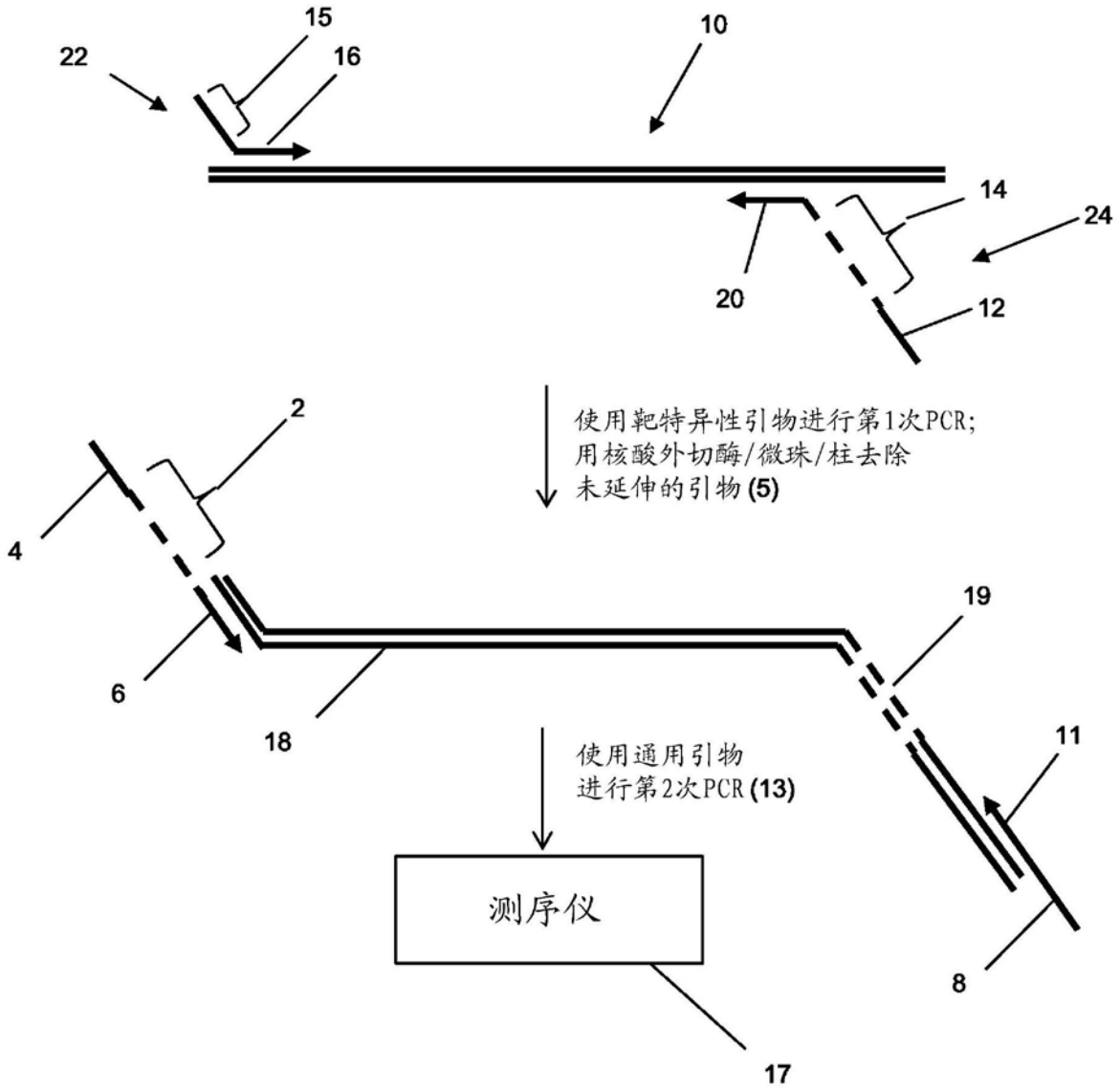

15
22
16
10
14
24
20
12
使用靶特异性引物进行第1次PCR;
用核酸外切酶/微珠/柱去除
未延伸的引物 (5)
2
4
19
6
18
11
使用通用引物
进行第2次PCR (13)
测序仪
8
17

图1A

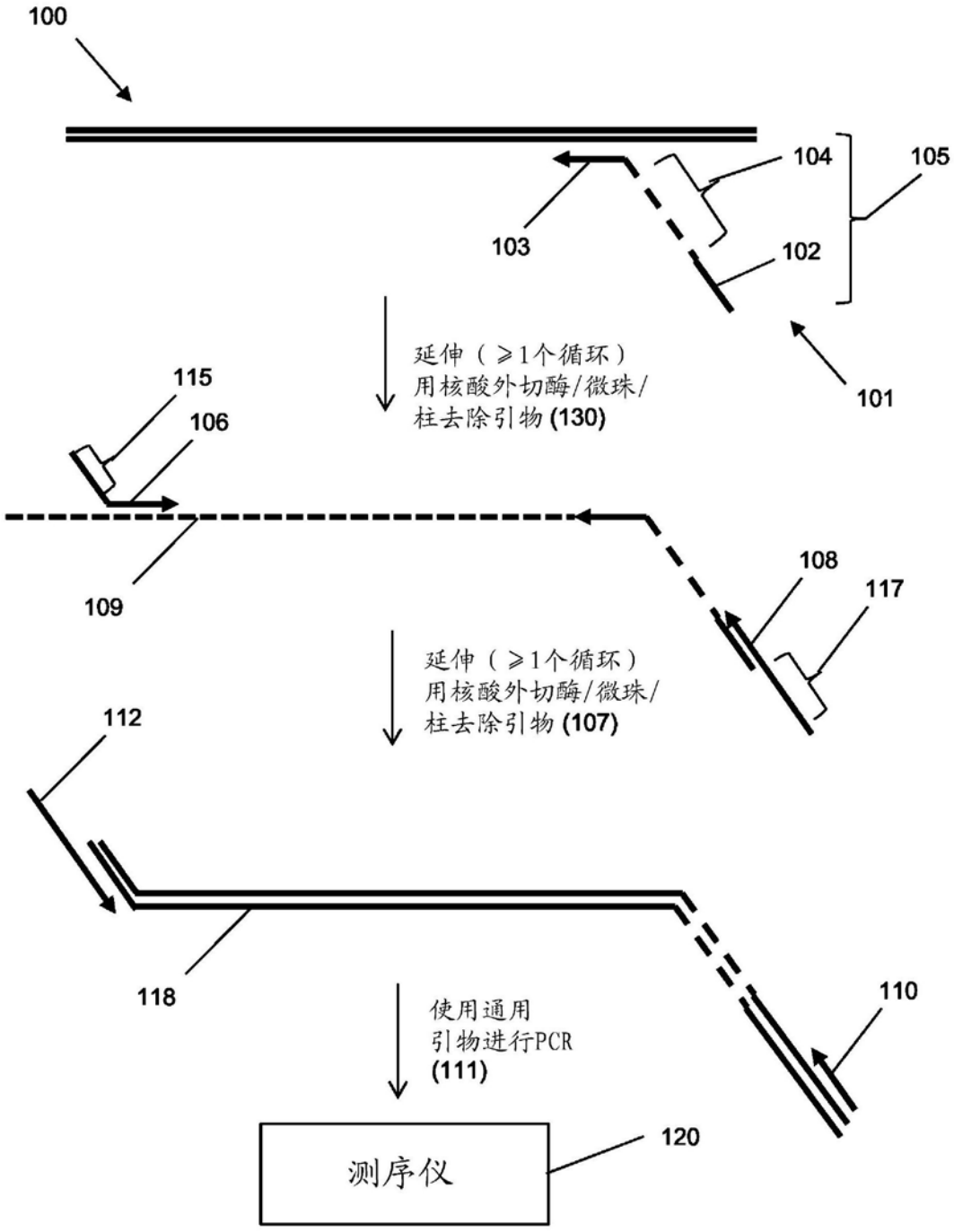

100
104
105
103
102
延伸（≥1个循环）
用核酸外切酶/微珠/
柱去除引物 (130)
101
115
106
109
108
117
延伸（≥1个循环）
用核酸外切酶/微珠/
柱去除引物 (107)
112
118
110
使用通用
引物进行PCR
(111)
测序仪
120

图1B

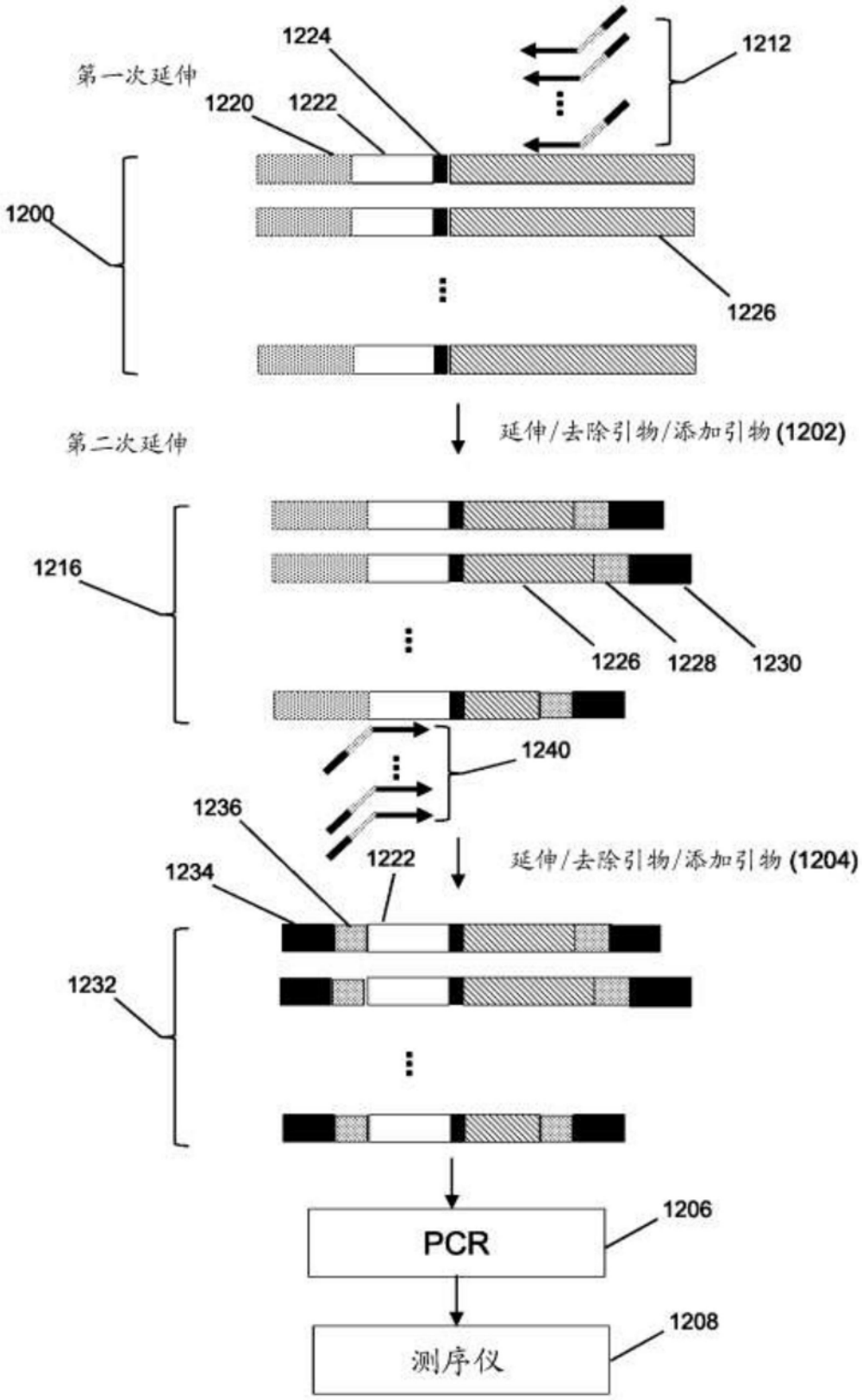

第一次延伸
1224
1212
1220
1222
1200
1226
延伸/去除引物/添加引物 (1202)
第二次延伸
1216
1226
1228
1230
1240
1236
1222
延伸/去除引物/添加引物 (1204)
1234
1232
PCR
1206
测序仪
1208

图1C

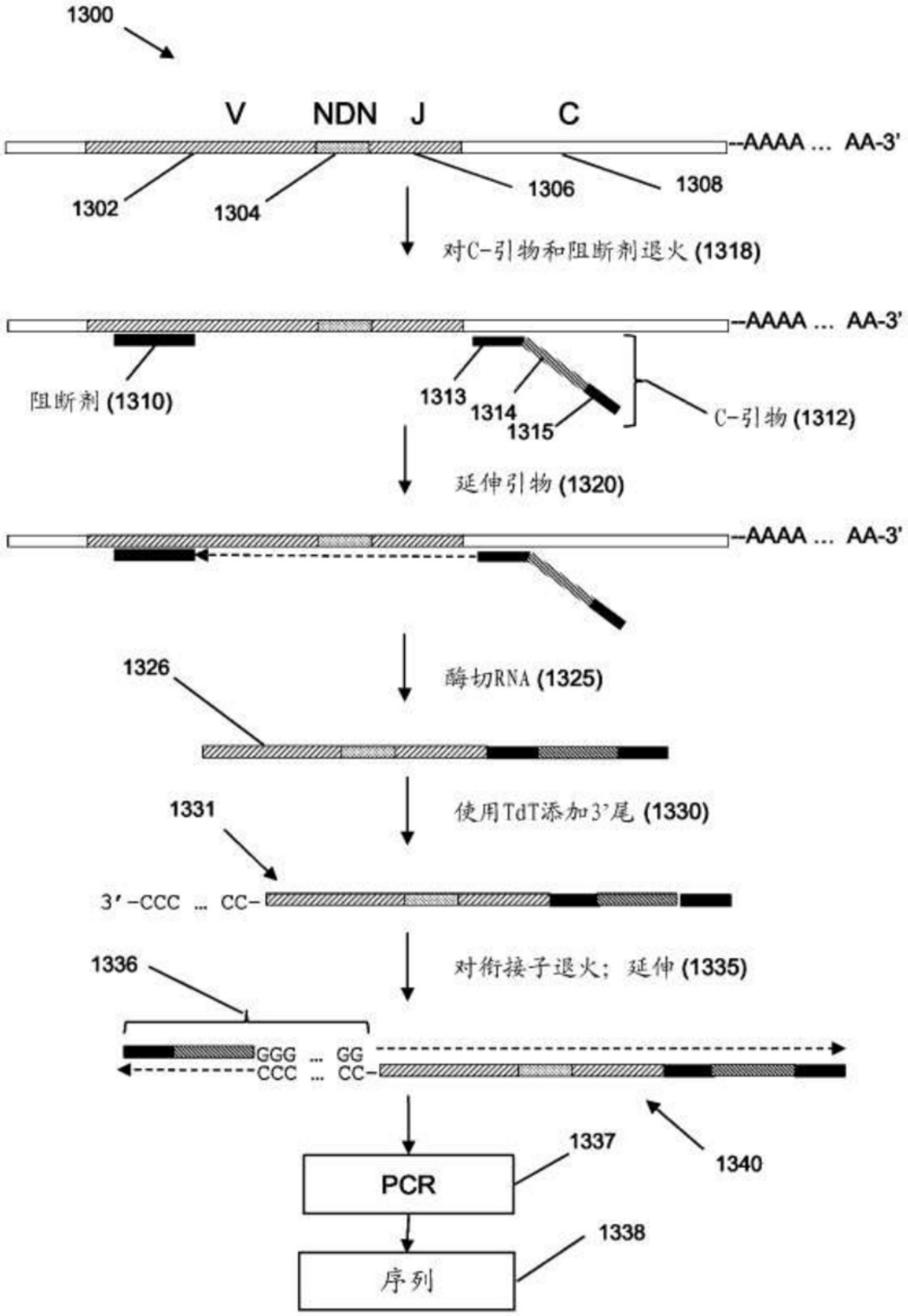

1300
V
NDN
J
C
--AAAA ... AA-3'
1302
1304
1306
1308
对C-引物和阻断剂退火 (1318)
--AAAA ... AA-3'
阻断剂 (1310)
1313
1314
1315
C-引物 (1312)
延伸引物 (1320)
--AAAA ... AA-3'
1326
酶切RNA (1325)
1331
使用TdT添加3'尾 (1330)
3'-CCC ... CC-
1336
对衔接子退火; 延伸 (1335)
GGG ... GG
CCC ... CC-
1337
1340
PCR
1338
序列

图1D

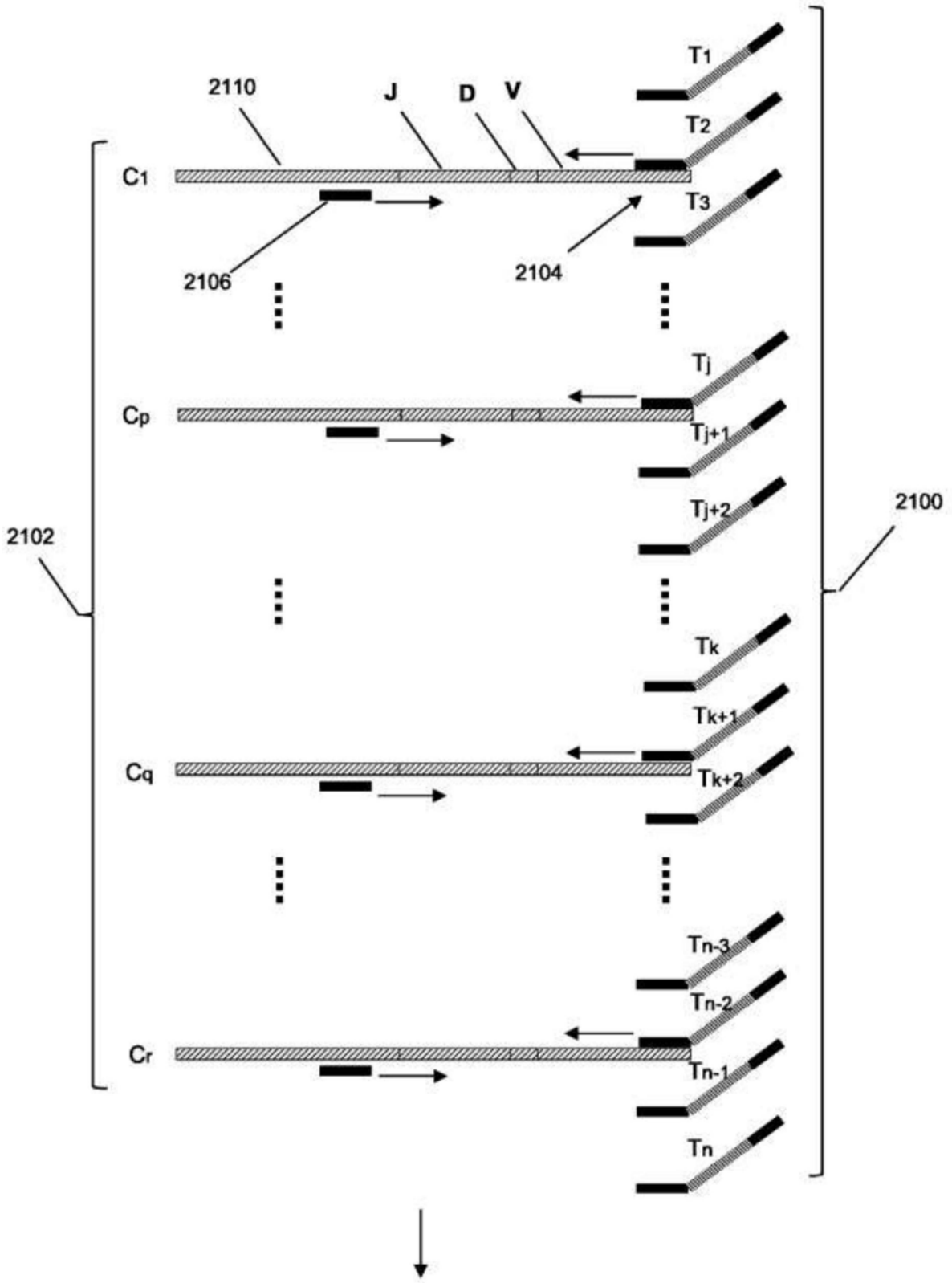

2110
J
D
V
C1
T1
T2
T3
2106
2104
Tj
Cp
Tj+1
Tj+2
2102
2100
Tk
Tk+1
Cq
Tk+2
Tn-3
Tn-2
Cr
Tn-1
Tn

图2A

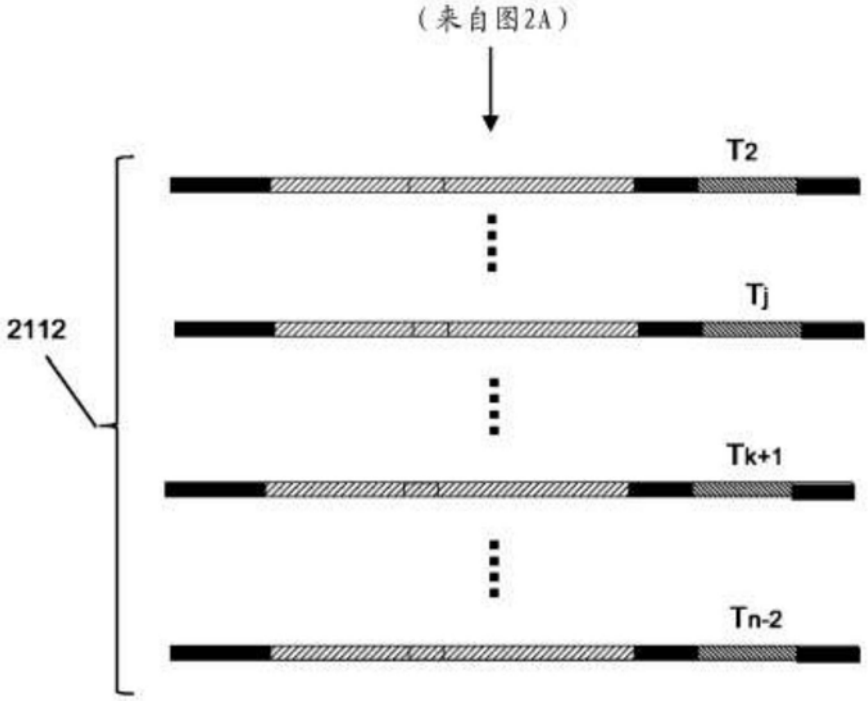


图2B

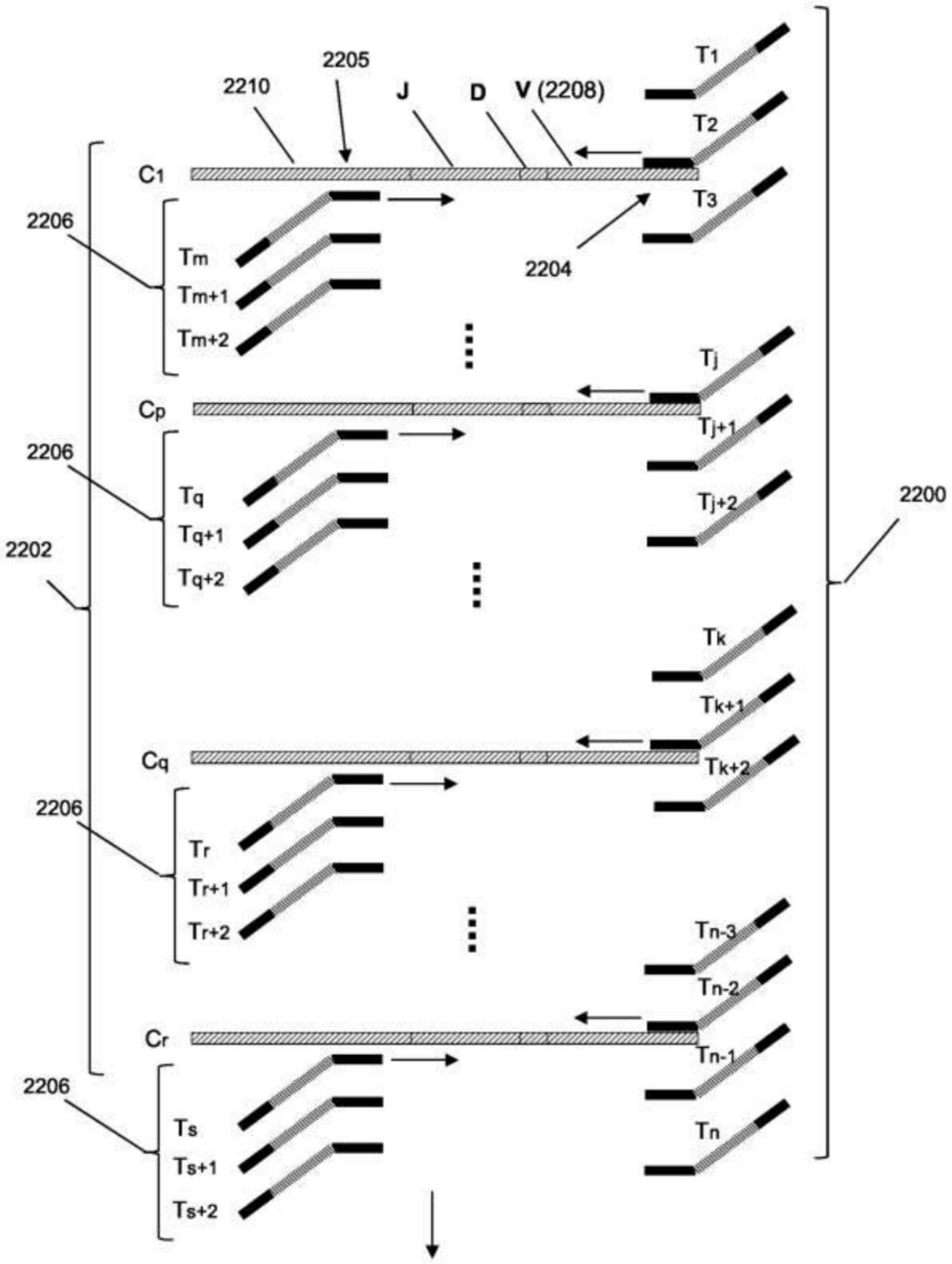

2205
2210
J
D
V (2208)
T1
T2
T3
C1
2206
Tm
Tm+1
Tm+2
2204
Tj
Cp
Tj+1
2206
Tq
Tq+1
Tq+2
Tj+2
2200
2202
Tk
Tk+1
Cq
Tk+2
2206
Tr
Tr+1
Tr+2
Tn-3
Tn-2
Cr
Tn-1
2206
Ts
Ts+1
Ts+2
Tn

图2C

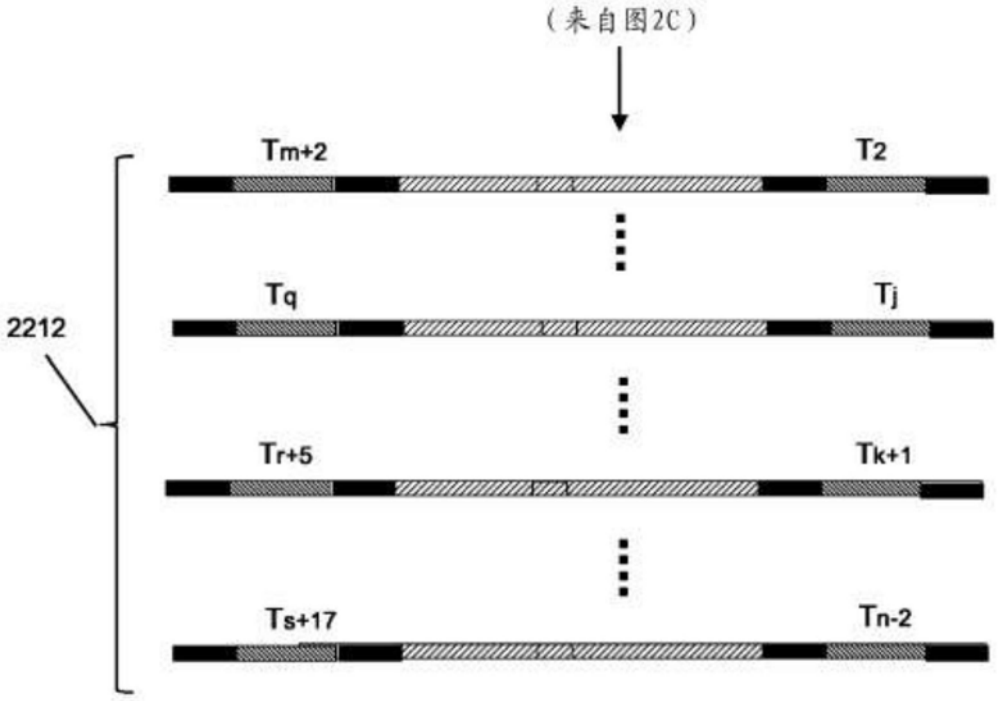


图2D

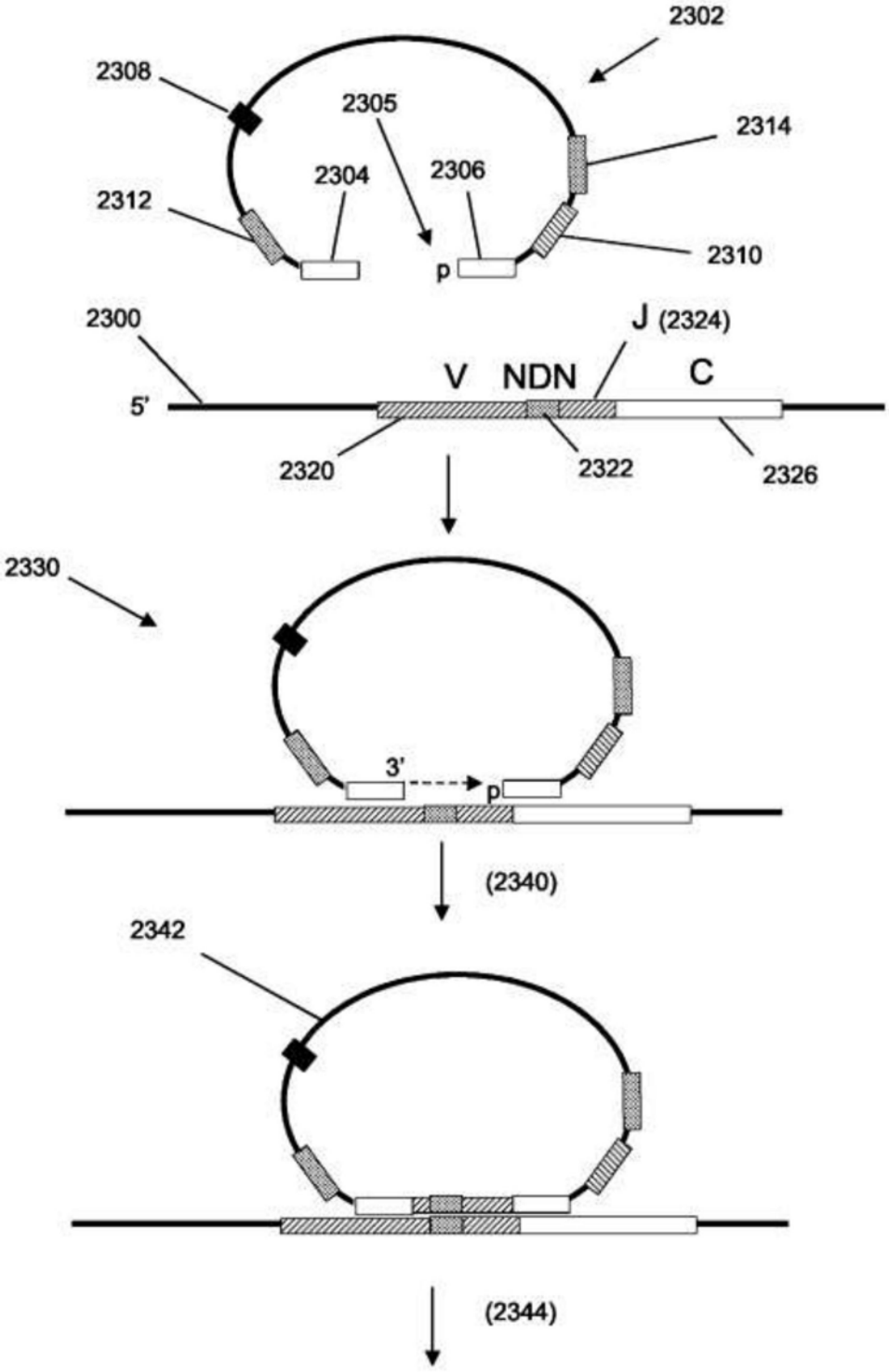

2302
2308
2305
2314
2304
2306
2312
2310
p
2300
J (2324)
V
NDN
C
5'
2320
2322
2326
2330
3'
p
(2340)
2342
(2344)

图2E

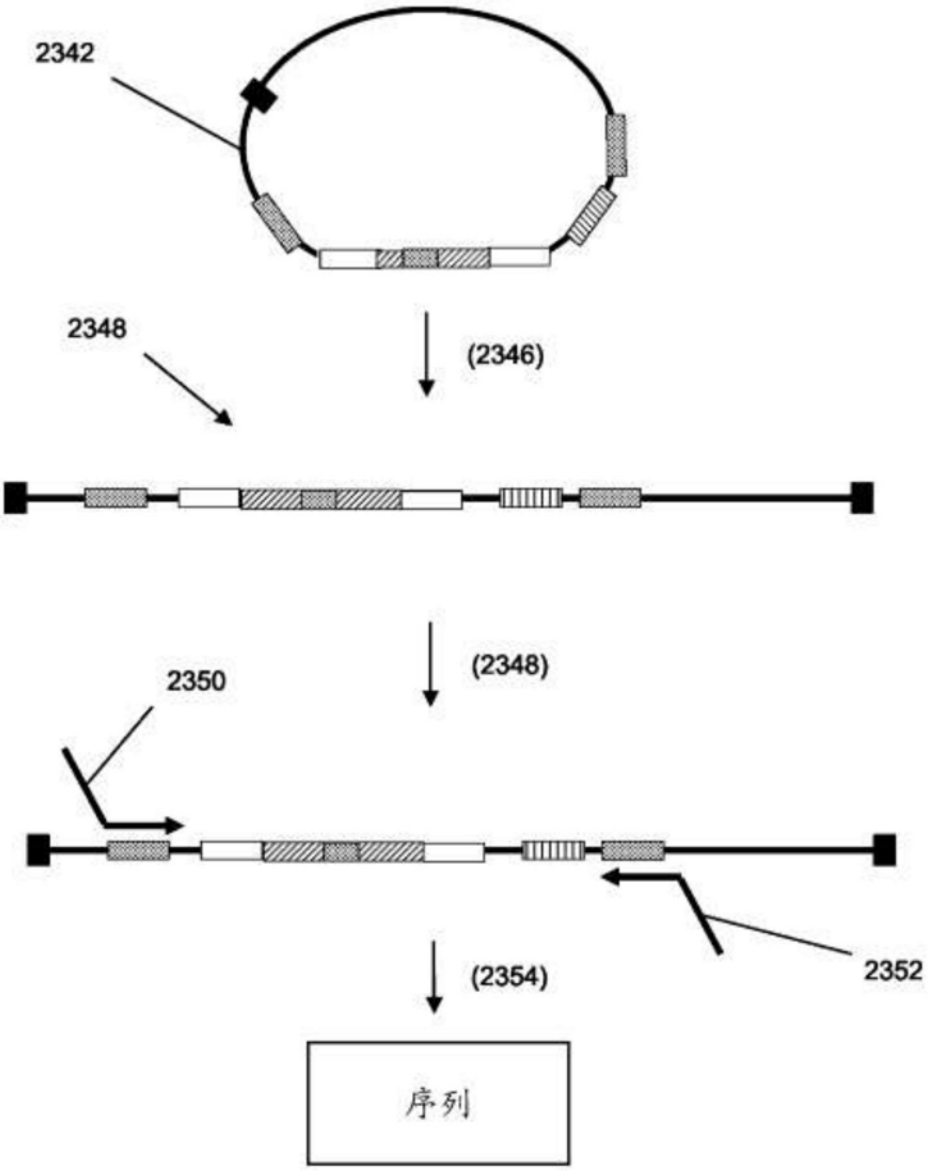


图2F

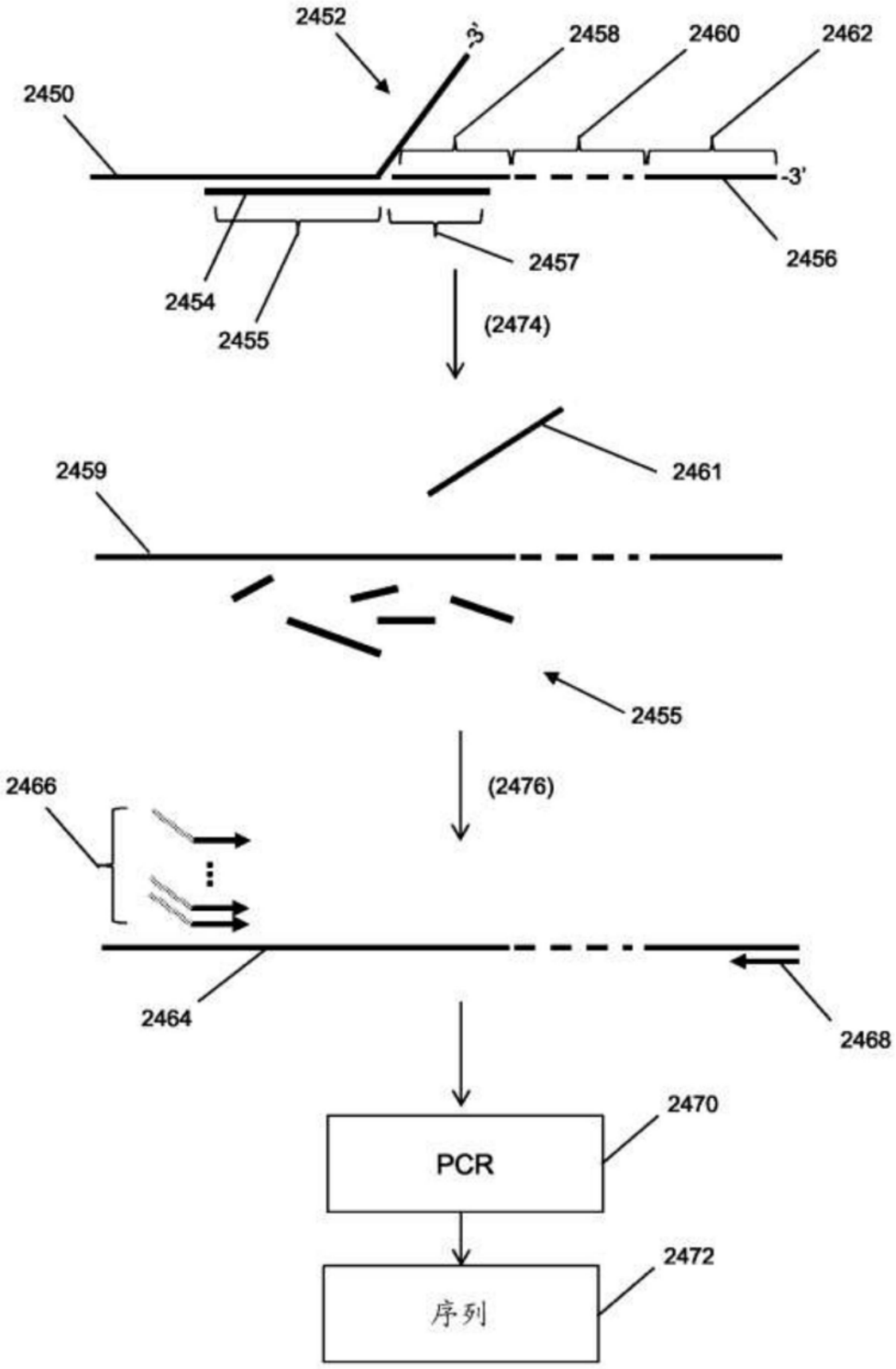

2452
5'
2458
2460
2462
2450
3'
2457
2456
2454
2455
(2474)
2459
2461
2455
(2476)
2466
2464
2468
2470
PCR
2472
序列

图2G

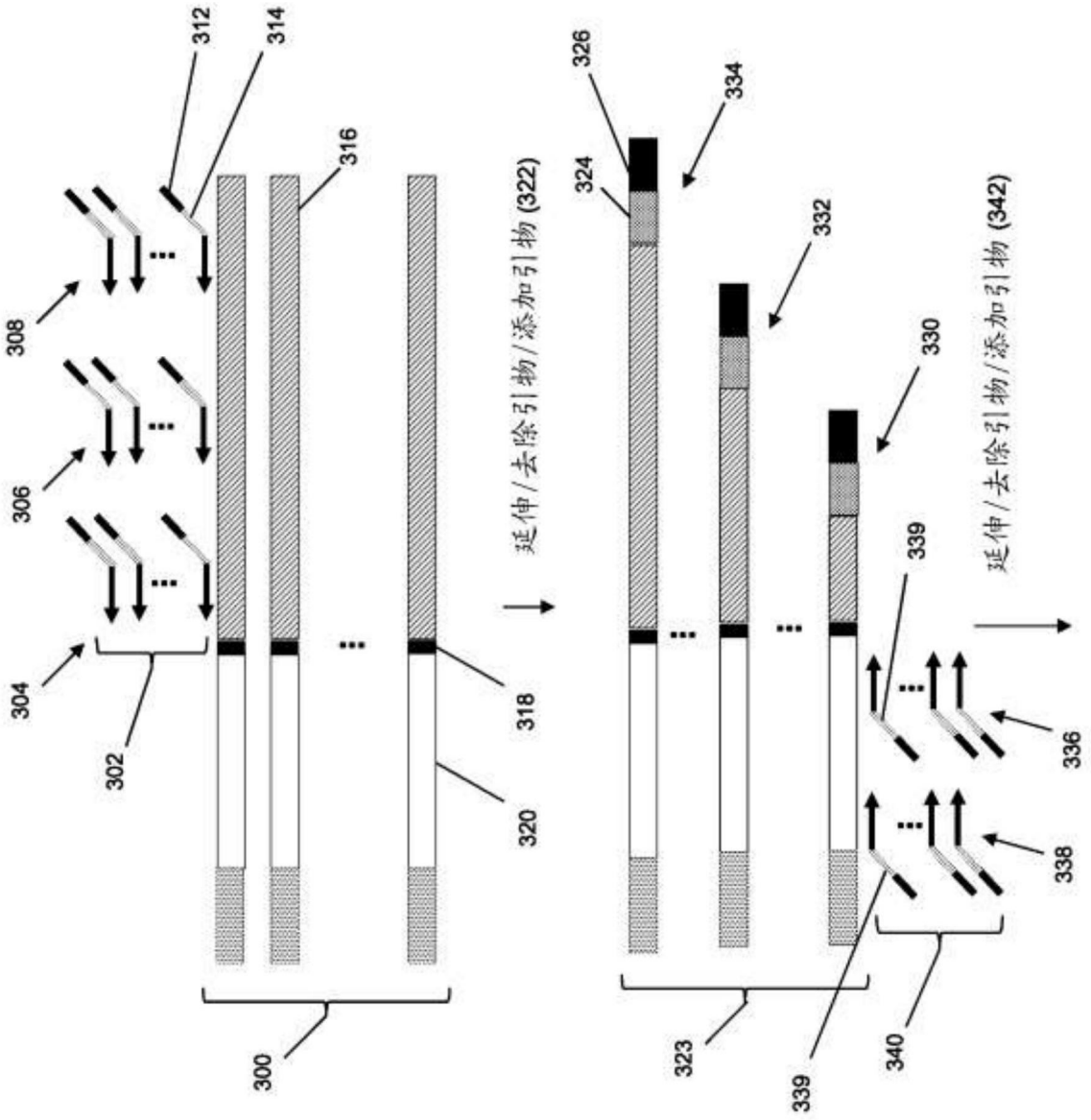

312
314
316
308
306
304
302
318
320
300
延伸/去除引物/添加引物 (322)
326
334
324
332
330
339
延伸/去除引物/添加引物 (342)
336
338
323
339
340

图3A

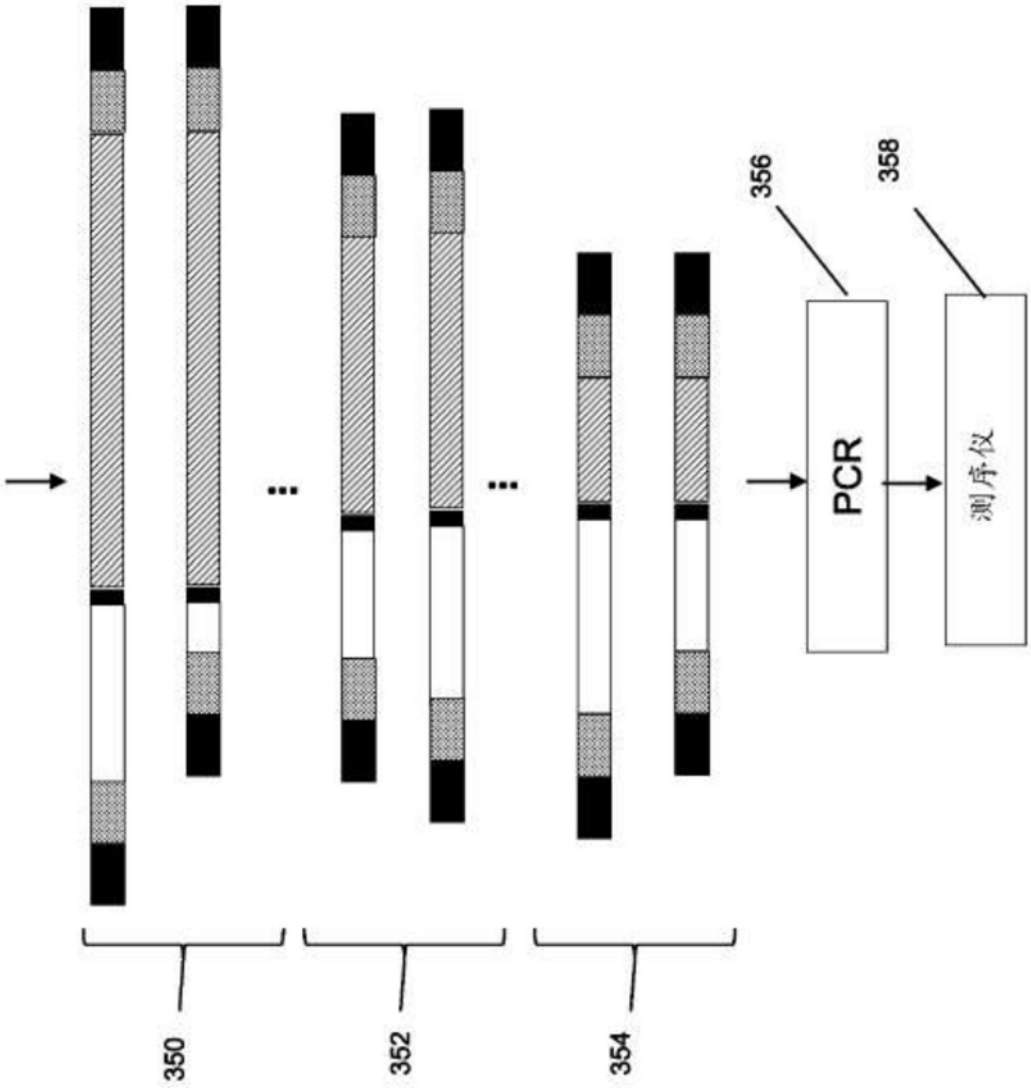


图3B

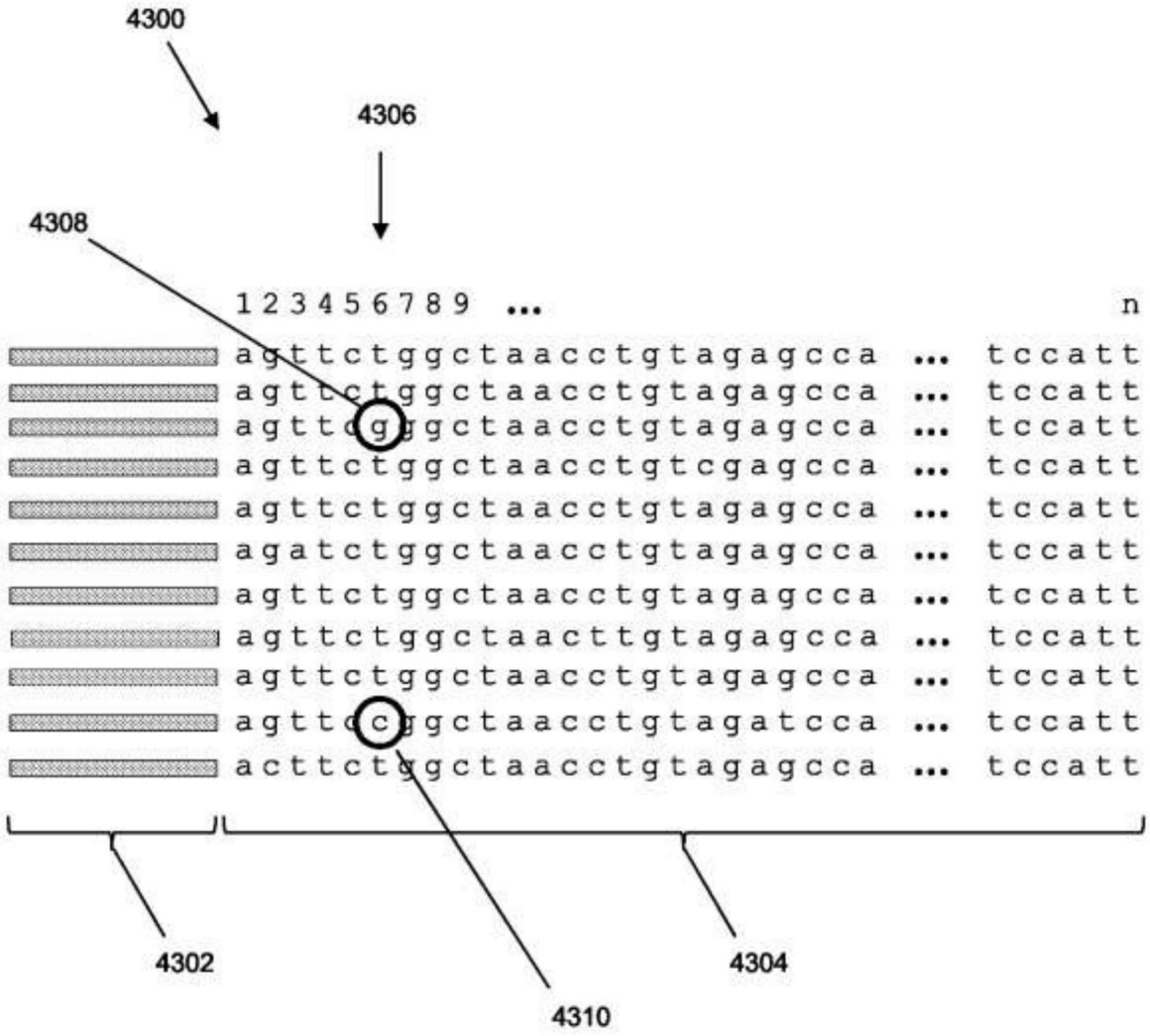

4300
4306
4308
1 2 3 4 5 6 7 8 9 ... n
agttctggctaacctgtagagcca ... tccatt
agttctggctaacctgtagagcca ... tccatt
agttctggctaacctgtagagcca ... tccatt
agttctggctaacctgtcgagcca ... tccatt
agttctggctaacctgtagagcca ... tccatt
agatctggctaacctgtagagcca ... tccatt
agttctggctaacctgtagagcca ... tccatt
agttctggctaacttgtagagcca ... tccatt
agttctggctaacctgtagagcca ... tccatt
agttctggctaacctgtagatcca ... tccatt
acttctggctaacctgtagagcca ... tccatt
4302
4304
4310

图4A

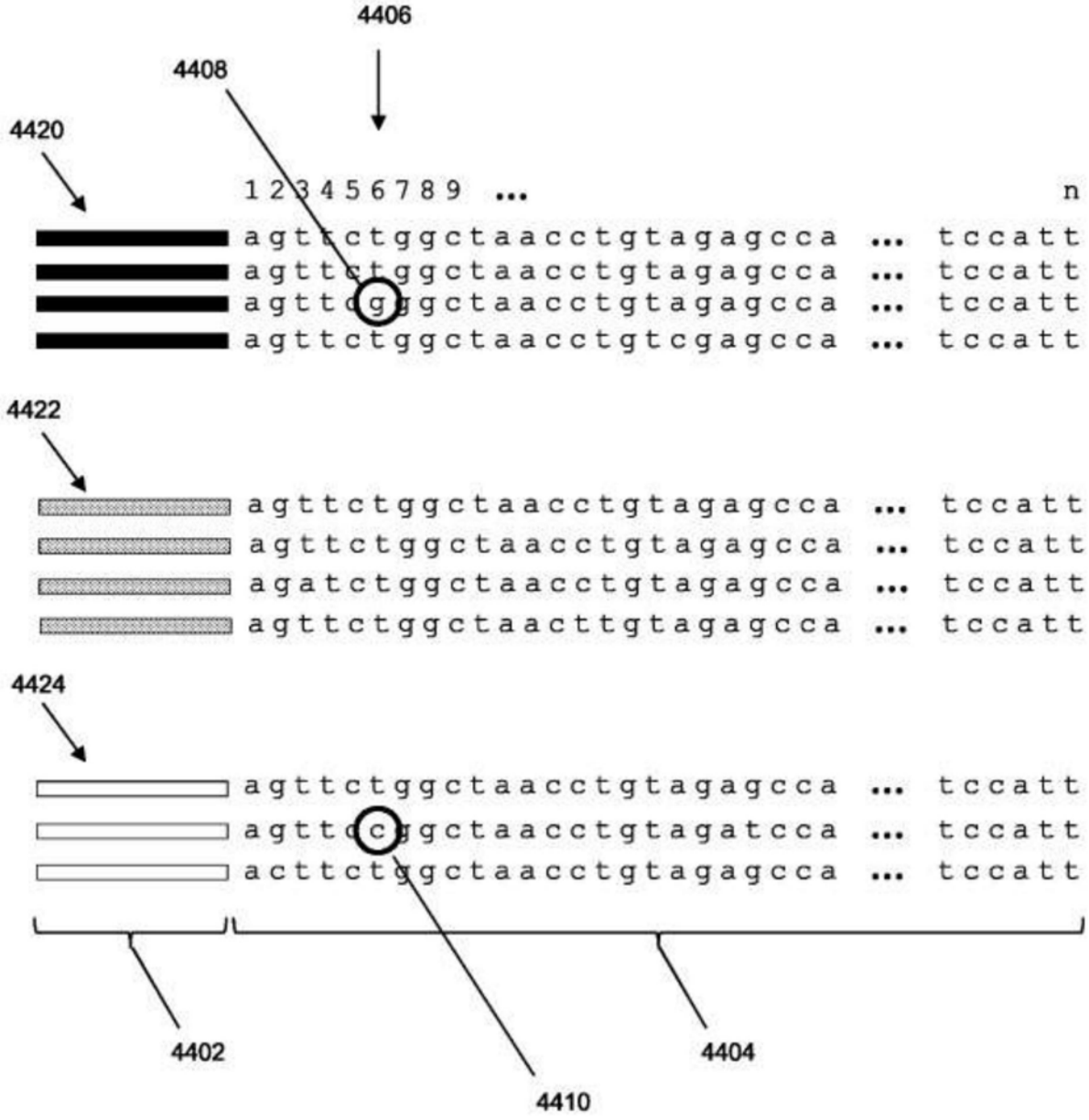

4406
4408
4420
1 2 3 4 5 6 7 8 9 ... n
agttctggctaacctgtagagcca ... tccatt
agttctggctaacctgtagagcca ... tccatt
agttcgggctaacctgtagagcca ... tccatt
agttctggctaacctgtcgagcca ... tccatt
4422
agttctggctaacctgtagagcca ... tccatt
agttctggctaacctgtagagcca ... tccatt
agatctggctaacctgtagagcca ... tccatt
agttctggctaacttgtagagcca ... tccatt
4424
agttctggctaacctgtagagcca ... tccatt
agttccggctaacctgtagatcca ... tccatt
acttctggctaacctgtagagcca ... tccatt
4402
4404
4410

图4B

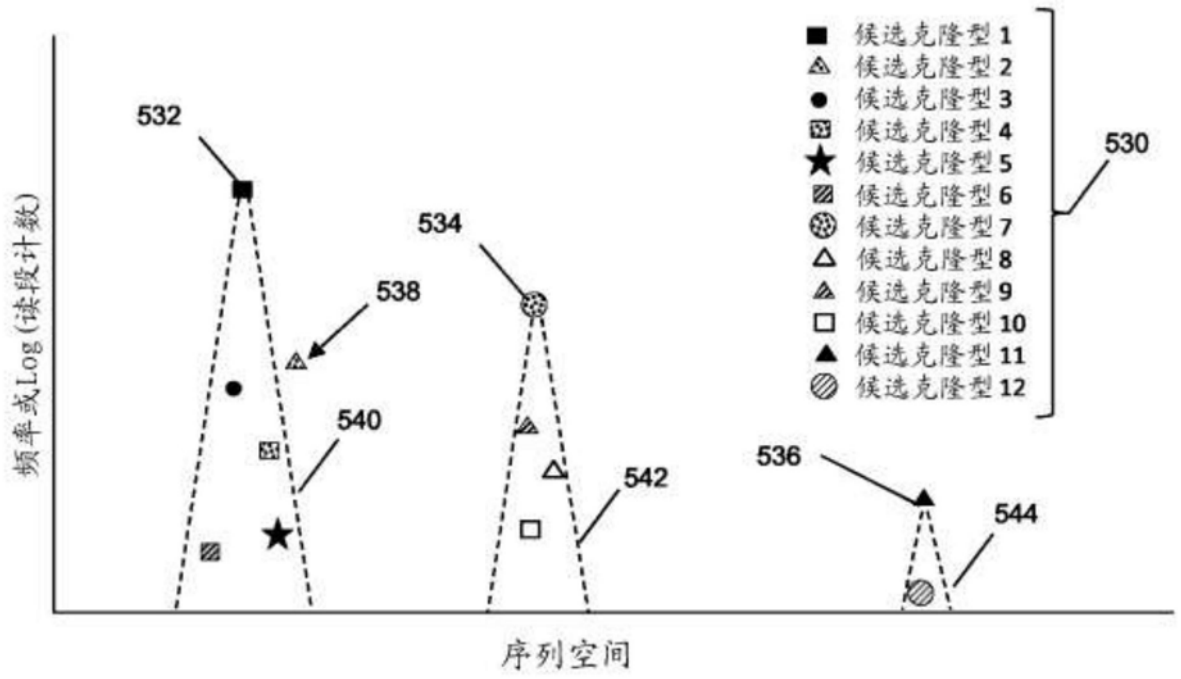

候选克隆型 1
候选克隆型 2
候选克隆型 3
候选克隆型 4
候选克隆型 5
候选克隆型 6
候选克隆型 7
候选克隆型 8
候选克隆型 9
候选克隆型 10
候选克隆型 11
候选克隆型 12
530
532
534
536
538
540
542
544
频率或Log(读段计数)
序列空间

图5A

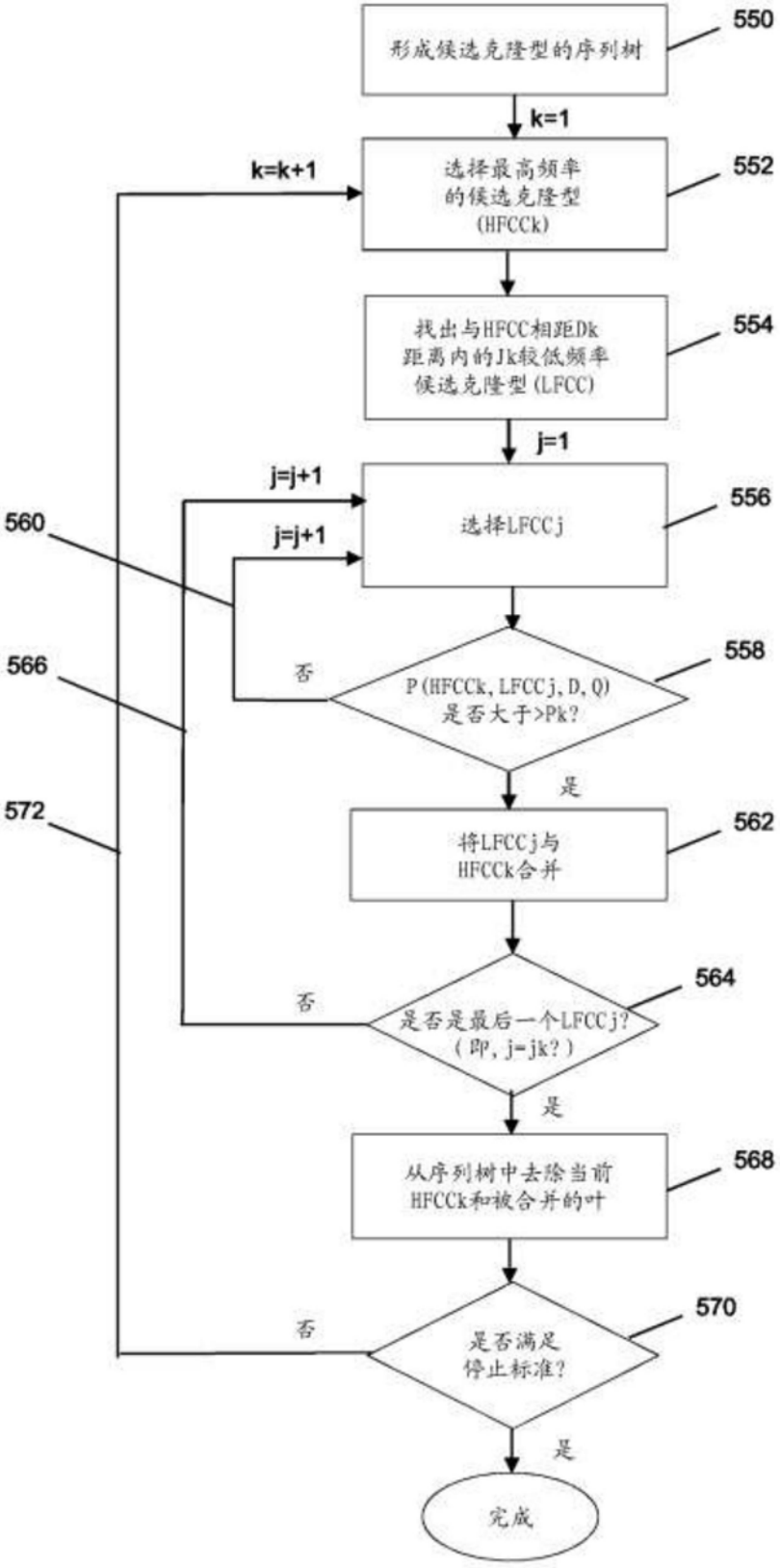

形成候选克隆型的序列树
550
k=1
k=k+1
选择最高频率
的候选克隆型
(HFCCk)
552
找出与HFCC相距Dk
距离内的Jk较低频率
候选克隆型(LFCC)
554
j=1
j=j+1
j=j+1
选择LFCCj
556
560
566
否
P(HFCCk, LFCCj, D, Q)
是否大于>Pk?
558
是
572
将LFCCj与
HFCCk合并
562
否
是否是最后一个LFCCj?
(即, j=jk?)
564
是
从序列树中去除当前
HFCCk和被合并的叶
568
否
是否满足
停止标准?
570
是
完成

图5B

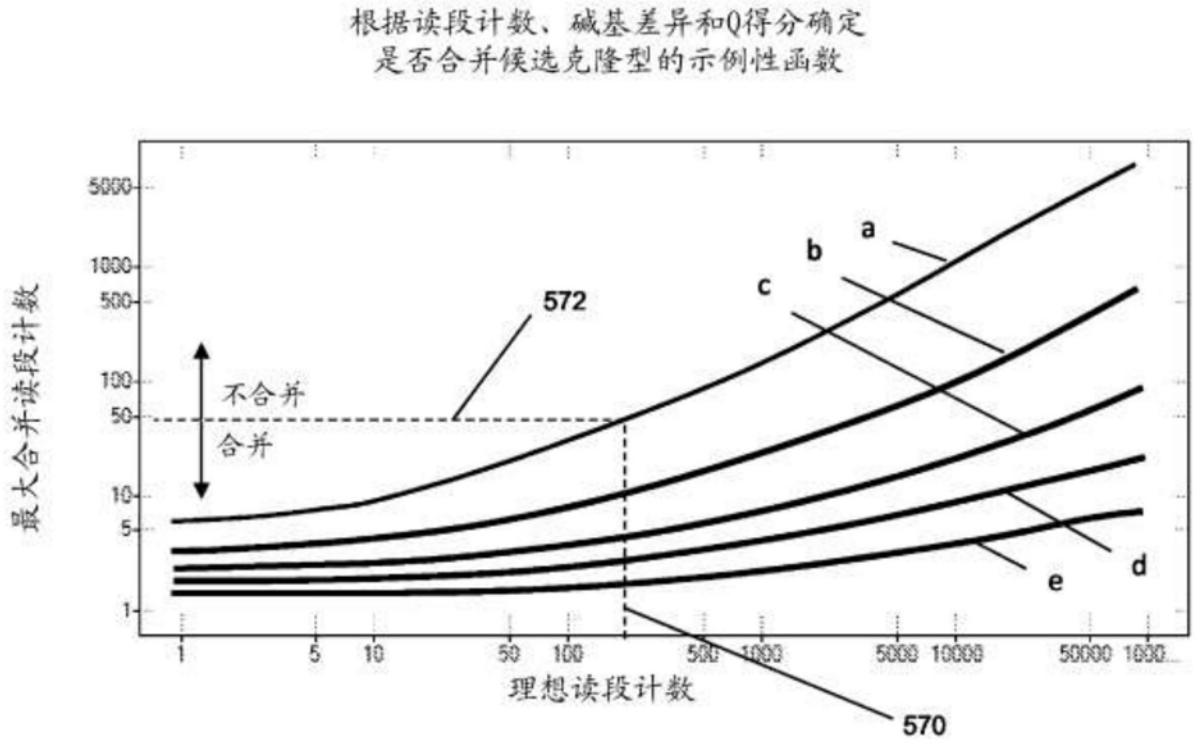

根据读段计数、碱基差异和Q得分确定
是否合并候选克隆型的示例性函数
最大合并读段计数
5000
1000
500
100
50
10
5
1
572
不合并
合并
a
b
c
d
e
1
5
10
50
100
500
1000
5000
10000
50000
理想读段计数
570

图5C

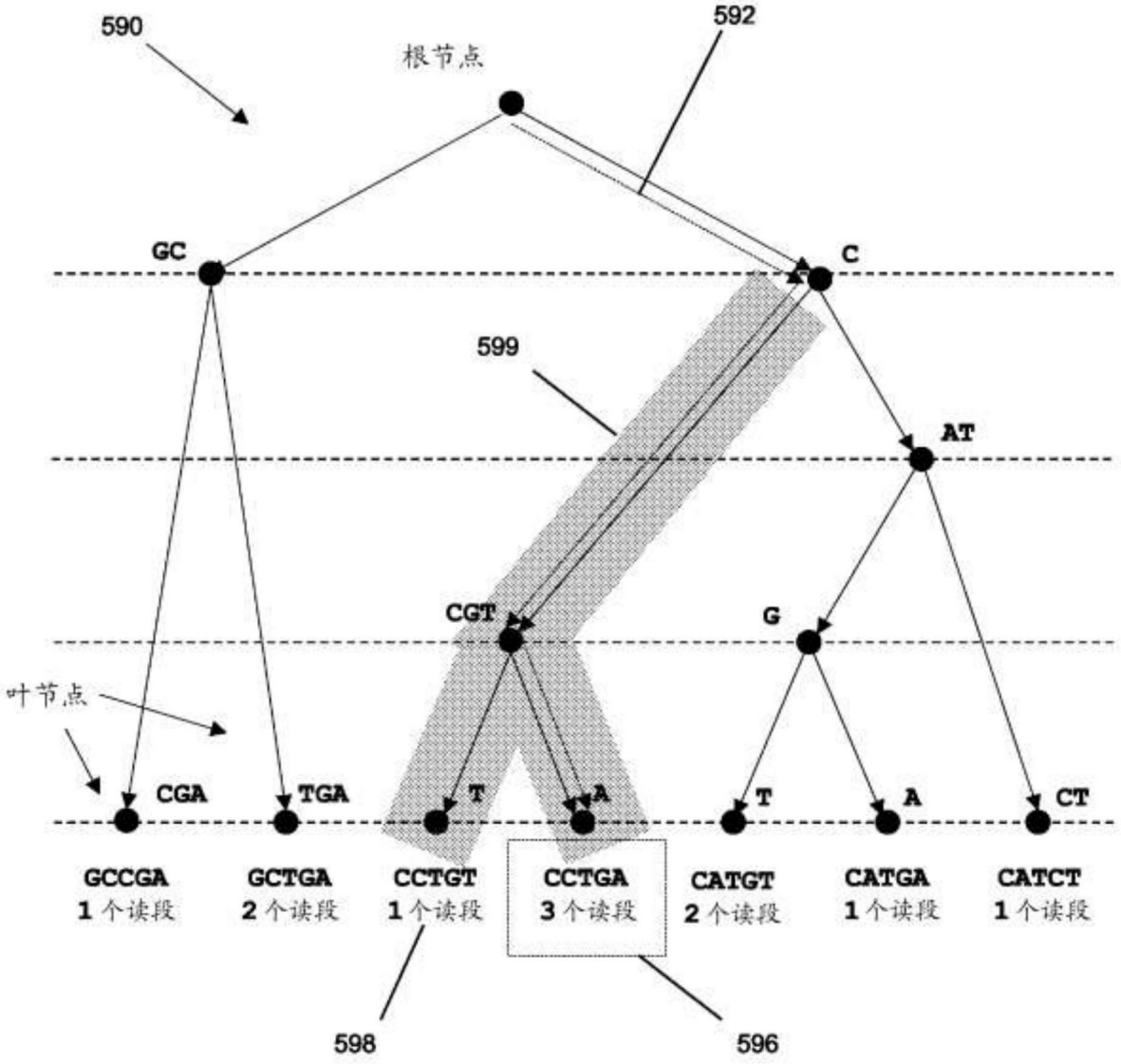

590
592
根节点
GC
C
599
AT
CGT
G
叶节点
CGA
TGA
T
A
T
A
CT
GCCGA
1 个读段
GCTGA
2 个读段
CCTGT
1 个读段
CCTGA
3 个读段
CATGT
2 个读段
CATGA
1 个读段
CATCT
1 个读段
598
596

图5D

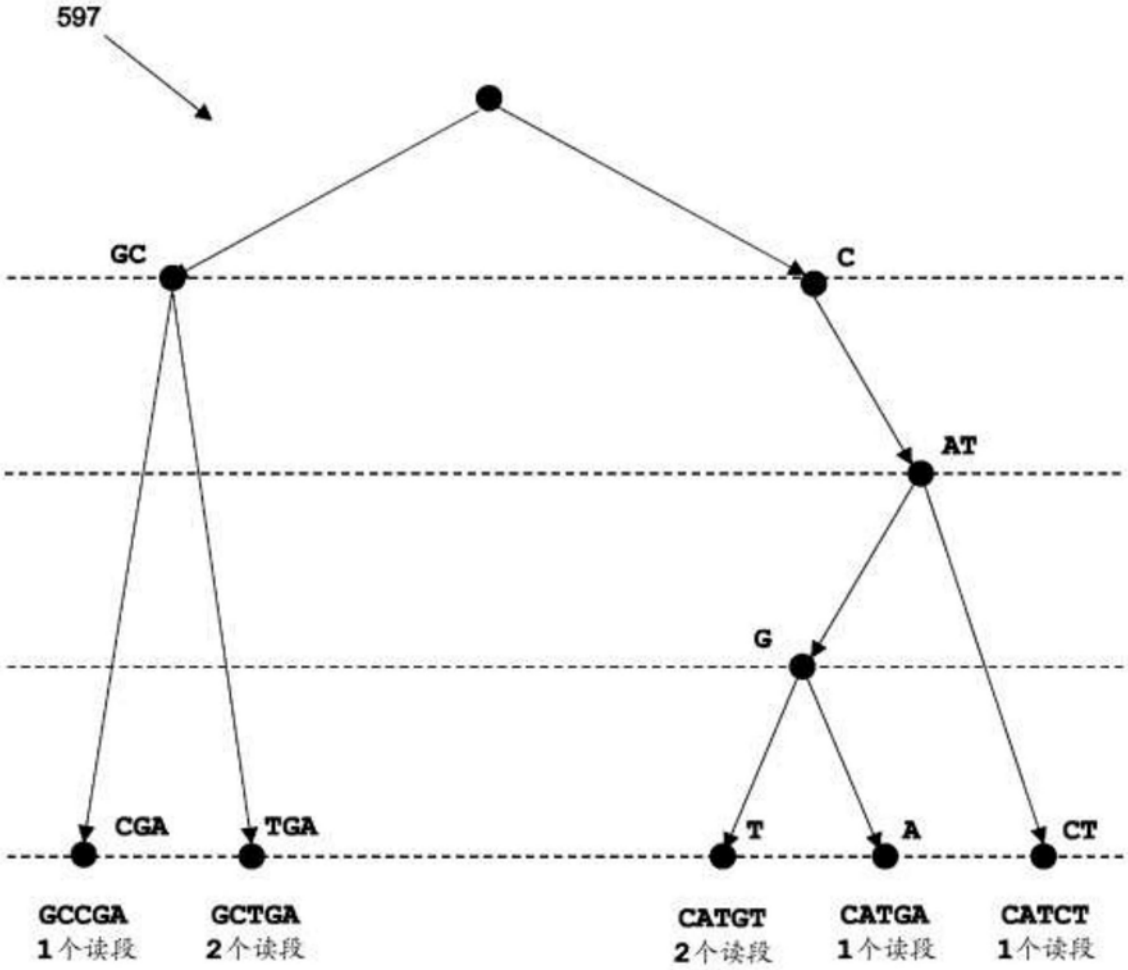

597
GC
C
AT
G
CGA
TGA
T
A
CT
GCCGA
1个读段
GCTGA
2个读段
CATGT
2个读段
CATGA
1个读段
CATCT
1个读段

图5E