

(12) 按照专利合作条约所公布的国际申请

(19) 世界知识产权组织
国际局

(43) 国际公布日
2020 年 6 月 11 日 (11.06.2020)



(10) 国际公布号
WO 2020/113569 A1

(51) 国际专利分类号:
C40B 50/06 (2006.01) *C12Q 1/68* (2018.01)

(21) 国际申请号: PCT/CN2018/119857

(22) 国际申请日: 2018 年 12 月 7 日 (07.12.2018)

(25) 申请语言: 中文

(26) 公布语言: 中文

(71) 申请人: 深圳华大生命科学研究院 (BGI SHENZHEN) [CN/CN]; 中国广东省深圳市盐田区盐田街道北山工业区综合楼, Guangdong 518083 (CN)。

(72) 发明人: 何琳(HE, Lin); 中国广东省深圳市盐田区盐田街道北山工业区综合楼, Guangdong 518083 (CN)。 廖莎(LIAO, Sha); 中国广东省深圳市盐田区盐田街道北山工业区综合楼, Guangdong 518083 (CN)。 徐崇钧(XU, Chongjun); 中国广东省深圳市盐田区盐田街道北山工业区综合楼, Guangdong 518083 (CN)。 章文蔚(ZHANG, Wenwei); 中国广东省深圳市盐田区盐田街道北山工业区综合楼, Guangdong 518083 (CN)。 陈奥(CHEN, Ao); 中国广东省深圳市盐田区盐田街道北山工业区综合楼, Guangdong 518083 (CN)。

(74) 代理人: 北京清亦华知识产权代理事务所 (普通合伙) (TSINGYIHUA INTELLECTUAL PROPERTY LLC); 中国北京市海淀区清华园清华大学照澜院商业楼301室, Beijing 100084 (CN)。

(81) 指定国(除另有指明, 要求每一种可提供的国家保护): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW。

(84) 指定国(除另有指明, 要求每一种可提供的地区保护): ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), 欧亚 (AM, AZ, BY, KG, KZ, RU, TJ, TM), 欧洲 (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG)。

本国际公布:

— 包括国际检索报告(条约第21条(3))。

(54) Title: METHOD FOR SEQUENCING LONG FRAGMENT NUCLEIC ACID

(54) 发明名称: 长片段核酸测序的方法

(57) Abstract: The invention discloses a method for sequencing long fragment nucleic acid. Nucleic acid molecules comprising a long inserted fragment, a first sequencing joint and a second sequencing joint are used to establish a sequencing library, and the nucleic acid sequence of a long inserted fragment can be obtained by means of segmented sequencing.

(57) 摘要: 本发明公开一种长片段核酸测序的方法。利用含有长插入片段以及第一测序接头和第二测序接头的核酸分子, 构建测序文库, 分段测序, 获得长插入片段的核酸序列。



WO 2020/113569 A1

长片段核酸测序的方法

优先权信息

无。

5

技术领域

本发明涉及基因测序领域，具体涉及一种长片段核酸测序的方法。

背景技术

10 1977 年，桑格发明了双脱氧末端终止测序法，成为第一代测序技术的代表。2001 年，依托第一代测序技术，完成了人类基因组草图。桑格测序法具有实验操作简单、结果直观准确和实验周期短等特点，在对检测结果时效性要求较高的临床基因突变检测以及基因分型等领域有着广泛的应用。然而，桑格测序法通量小、成本高，限制了其在大规模基因测序中的应用。

15 为克服桑格测序法的缺点，第二代测序技术应运而生。与第一代测序技术相比，第二代测序技术具有通量大、成本低、自动化程度高等优点，适合于大规模测序。目前已开发的第二代测序技术主要涉及边连接边测序（sequencing by ligation, SBL）技术和边合成边测序（sequencing by synthesis, SBS）技术。这些测序技术的典型实例包括 Roche 454 测序法、Applied Biosystems 公司开发的 SOLiD 测序法、Complete Genomics 自主开发的联合探针锚定连接法（cPAL）和华大基因开发的联合探针锚定合成法（cPAS）、Illumina 公司和
20 Solexa technology 公司合作开发的 Illumina 测序法等等。测序检测方法主要有电化学法、光信号检测法等，其中，较为主流的检测方法为光信号检测。第二代测序技术由于硬件要求，仪器普遍较为庞大，不利于携带及搬运。而第三代测序技术虽然能够克服第二代测序技术仪器庞大的缺点，但是其测序错误率较高，限制了其大规模推广。

25 由此，对于基因测序方法，尤其是长片段核酸的测定还需要进一步改进。

发明内容

本发明旨在至少在一定程度上解决相关技术中的技术问题之一，提供了一种成本低，且能够实现长片段核酸测序的方法。

30 本发明的发明人在研究过程中发现：在基因测序领域，长序列读取有诸多优势例如有

利于转录组的研究，可以直接获得全长转录组。但读长短一直是二代测序的软肋，对于高度杂合的基因组、高度重复序列、高 GC 的区域、拷贝数变异、大的结构变异等问题，二代测序很难解决。二代测序 Illumina HiSeq 2500 测序平台是基于 DNA 单分子簇的边合成边测序技术以及其专有的可逆终止化学反应技术。基于该平台的测序需首先在 DNA 片段两端加上序列已知的通用接头（P7 和 P5 接头）构建文库，再将文库加载到测序芯片 Flowcell 上，Flowcell 是具有多个 lane（泳道）的玻璃板。每个 lane 可以测一个样本或者多样本的混合物，且随机布满了能够与文库两端接头分别互补配对或一致的寡核苷酸（oligos）。测序时将制备好的基因组 DNA 的随机片段附着到光学透明的玻璃表面（即 Flow Cell），这些 DNA 片段经过延伸和桥式扩增后，在 Flow Cell 上形成了数以亿计 Cluster，每个 Cluster 是具有数千份相同模板 DNA 的单分子簇。然后利用带荧光基团的四种特殊脱氧核糖核苷酸，通过可逆性终止的 SBS（边合成边测序）技术对待测的模板 DNA 进行测序：即在碱基延伸过程中，每个循环反应只能延伸一个正确互补的碱基，根据四种不同的荧光信号确认碱基种类，保证最终的核酸序列质量，经过多个循环后，完整读取核酸序列。这种新方法确保了高精确度和真实的一个碱基接一个碱基的测序，为同聚物和重复序列的测序提供了一个很好的解决方案。

基于此，发明人创造性的通过分段测序的方法，对含有长片段插入片段的核酸序列进行分段测定。例如可以在每段只测定大约 300bp 左右的碱基序列，然后通过序列拼接达到长片段序列测定的目的。

为达到上述目的，本发明的技术方案通过如下步骤实现的：

根据本发明的第一方面，本发明提供了一种用于确定待测核酸分子预定区域核酸序列的方法，所述待测核酸分子包括插入片段，所述插入片段的长度大于测序仪的读长，所述方法包括：使所述插入片段的两端分别带有第一测序接头和第二测序接头，所述第一测序接头与所述插入片段的一端相连，所述第二测序接头与所述插入片段的另一端相连；利用第一 dNTP 混合物和测序引物，对所述待测核酸分子进行第一延伸处理，所述测序引物与所述第一测序接头和所述第二测序接头之一匹配，以便将所述测序引物延伸至所述预定区域的上游；利用第二 dNTP 混合物对所述第一延伸处理的产物进行第二延伸处理，并在所述第二延伸过程中，确定所述预定区域的核酸序列；其中，所述第一 dNTP 混合物包括 dATP、dTTP、dGTP 和 dCTP，所述 dATP、dTTP、dGTP 和 dCTP 的至少一种携带聚合酶反应阻断基团；所述第二 dNTP 混合物包括 dATP、dTTP、dGTP 和 dCTP，所述 dATP、dTTP、dGTP 和 dCTP 均携带聚合酶反应阻断基团，并且所述 dATP、dTTP、dGTP 和 dCTP 的至少之一携带可检

测基团。所述“测序仪的读长”指的是测序仪能够准确读取的核酸序列的最大长度。插入片段的长度大于测序仪的读长，是指插入片段的长度至少为 400bp，至少为 500bp，或者至少为 600bp。

根据本发明的实施例，以上用于确定待测核酸分子预定区域核酸序列的方法可以进一步包括如下技术特征：

在本发明的一些实施例中，所述第一延伸处理包括：（1）将所述待测核酸分子与所述测序引物进行退火反应，以便形成起始双链体，所述起始双链体由所述待测核酸分子和所述测序引物构成；（2）以所述起始双链体中的测序引物为第一起始生长核酸链，在聚合酶的催化作用下，将所述第一 dNTP 混合物中的至少一个并入所述起始生长核酸链的 3'端，以便在所述起始生长核酸链的 3'端延伸新的碱基，形成第一产物双链体；（3）将所述第一产物双链体进行切割处理，以便去除所述第一产物双链体上的聚合酶反应阻断基团；（4）重复步骤（2）~（3）直至所述测序引物延伸至所述预定区域的上游预定位置。

在本发明的一些实施例中，所述第一 dNTP 混合物中所述 dATP、dTTP、dGTP 和 dCTP 全部携带聚合酶反应阻断基团。dATP、dTTP、dGTP 和 dCTP 上全部携带聚合酶反应阻断基团，可以实现每延伸一个碱基切割一次，保证将碱基延伸到预定位置。

在本发明的一些实施例中，所述第一 dNTP 混合物中所述 dATP、dTTP、dGTP 和 dCTP 的仅一种携带聚合酶反应阻断基团。

在本发明的一些实施例中，所述第一 dNTP 混合物中所述 dATP、dTTP、dGTP 和 dCTP 的至多三种携带聚合酶反应阻断基团。dATP、dTTP、dGTP 和 dCTP 的至多三种携带聚合酶反应阻断基团，可以在保证延伸准确性的同时，兼顾延伸的效率。

在本发明的一些实施例中，所述第一 dNTP 混合物中所述 dATP、dTTP、dGTP 和 dCTP 均不携带可检测基团。

在本发明的一些实施例中，在步骤（4）中，所述预定位置位于所述预定区域的上游 1~100bp 范围内。

在本发明的一些实施例中，所述第一延长处理产物的长度为至少 200bp，优选 200~600bp。

在本发明的一些实施例中，在步骤（4）中，所述预定位置与所述预定区域下游末端的距离不超过 400bp，优选不超过 300bp。

在本发明的一些实施例中，所述第二延伸处理包括：（a）以所述第一延伸处理的产物为起始点，在聚合酶的催化作用下，将所述第二 dNTP 混合物中的一个并入所述第一延伸产物

的 3'端,以便在所述第一延伸产物的 3'端延伸新的碱基,形成第二产物双链体,并通过对所述第二产物双链体的可检测基团进行检测,以便确定所述新的碱基的类型;(b)将所述第二产物双链体进行切割处理,以便去除所述第二产物双链体上的聚合酶反应阻断基团和所述可检测基团;(c)重复步骤(a)~(b)直至确定所述预定区域的至少一部分碱基类型。

5 在本发明的一些实施例中,所述可检测基团为荧光基团或磷光基团。

在本发明的一些实施例中,所述 dATP、dTTP、dGTP 和 dCTP 均携带不同的所述可检测基团。

根据本发明的第二方面,本发明提供了一种用于本发明第一方面任一实施例所述的测序方法的 dNTP 混合物。

10 根据本发明的第三方面,本发明提供了一种用于核酸测序的核酸分子,包括:插入片段,所述插入片段的长度为至少 500bp,优选至少 800bp,例如 1000~1200bp;第一测序接头,所述第一测序接头与所述插入片段的一端相连;第二测序接头,所述第二测序接头与所述插入片段的另一端相连。利用本申请提供的用于核酸测序的核酸分子,可以实现对于长片段插入片段的序列测定,从而可以保证测序的准确率和快捷性。

15 根据本发明的实施例,所述用于核酸测序的核酸分子进一步包括:标签序列,所述标签序列设置在所述第一测序接头和所述第二测序接头中。

根据本发明的第四方面,本发明提供了一种测序文库,所述测序文库中含有本发明第一方面所述的核酸分子。由本发明第一方面所述的核酸分子所构成的测序文库可以实现对于核酸分子中插入片段的核酸序列进行测定,或者对于核酸分子中插入片段的某一段预定区域

20 的核酸序列进行测定。

根据本发明的第五方面,本发明提供了一种测序芯片,所述测序芯片上携带本发明第一方面所述的核酸分子。测序芯片上携带有用于核酸测序的核酸分子,可以直接利用测序芯片实现对于核酸分子的序列测定,从而可以方便快捷的实现测序。

根据本发明的第六方面,本发明提供了一种用于确定待测核酸分子的核酸序列的方法,所述待测核酸分子的插入片段为至少 700bp,所述待测核酸分子的插入片段分为第一区域、第二区域和第三区域,所述第一区域、第二区域和第三区域依次相连,所述第一区域、第二区域和第三区域的片段大小大致相同,所述方法包括:(1')根据本发明第一方面任一实施例所述的方法,分别确定所述待测核酸分子第一区域、第二区域、第三区域的核酸序列;

25 (2')基于步骤(1')的结果,确定所述核酸分子的核酸序列。

30 “所述第一区域、第二区域和第三区域的片段大致相同”是指:第一区域的片段大小、

第二区域的片段大小和第三区域的片段大小不要求绝对的完全相同，根据插入片段的大小尽量平均即可。例如每个区域的片段大小可以相差 10bp, 20bp, 30bp, 40bp, 50bp, 甚至是 80bp, 100bp 等等。当然，第一区域、第二区域和第三区域的片段大小不全相同，例如第一区域的大小为 200bp, 第二区域的大小为 300bp, 第三区域的大小为 300bp 也可以利用测序平台实现一次测序。本文中的“一次测序”或者“一轮测序”是指在一个混合体系和一个环境下所进行的一次操作。当进行完第一次测序或者第一轮测序后，进行第二次测序或者第二轮测序，是指进行了另一次操作，这次操作所对应的混合体系和环境与第一次测序或者第一轮测序并不必然相同或者不同。所述第一区域、第二区域和第三区域的片段大致相同，就尽可能保证每次测序或者每轮测序的效率，不至于某一区域片段过大或者过小时，对于测序资源的浪费。

例如，当所述待测核酸分子的插入片段为 700bp 时，可以将插入片段大致平均分成三个区域，也可以适当进行调整，例如第一区域为 250bp 左右，第二区域为 250bp 左右，第三区域为 200bp 左右。然后分别确定第一区域、第二区域和第三区域的核酸序列，通过序列拼接获得插入片段的测序序列。

在本发明的一些实施例中，以上用于确定待测核酸分子的核酸序列的方法可以进一步包括如下技术特征：

在本发明的一些实施例中，所述待测核酸分子的插入片段为大约 1000bp，所述第一区域的起始位置距离所述插入片段的上游末端至少 600bp，例如为 660bp，所述第二区域的起始位置距离所述插入片段的上游末端至少 300bp，例如为 330bp，所述第三区域的起始位置为所述插入片段的上游末端。“插入片段为大约 1000bp”是指并非绝对要求插入片段的大小为 1000bp，可以在 1000bp 大小左右浮动，例如可以比 1000bp 大或者小 10bp, 20bp, 30bp, 40bp, 50bp, 甚至是 80bp, 100bp 等等。

在本发明的一些实施例中，步骤（1'）进一步包括：（a）利用阻抑性引物和部分所述待测核酸分子结合，并根据本发明第一方面任一实施例所述的方法确定所述待测核酸分子第一区域的核酸序列，所述阻抑性引物与所述测序引物的核酸序列相同，所述阻抑性引物的 3' 端带有阻抑基团；（b）切除部分所述阻抑性引物的阻抑基团，并根据本发明第一方面任一实施例所述的方法确定所述待测核酸分子第二区域的核酸序列；（c）切除部分所述阻抑性引物的阻抑基团，并根据本发明第一方面任一实施例所述的方法确定所述待测核酸分子第三区域的核酸序列。

在本发明的一些实施例中，利用核酸内切酶切除所述阻抑性引物的阻抑基团。

在本发明的一些实施例中，步骤（1'）进一步包括：（a）根据本发明第一方面任一实施例所述的方法确定所述待测核酸分子第一区域的核酸序列，并利用碱性溶液对所述第二延伸处理的产物进行解链、洗脱，获得所述待测核酸分子；（b）根据本发明第一方面任一实施例所述的方法确定所述待测核酸分子第二区域的核酸序列，并利用碱性溶液对所述第二延伸处理的产物进行解链、洗脱，获得所述待测核酸分子；（c）根据本发明第一方面任一实施例所述的方法确定所述待测核酸分子第三区域的核酸序列。

在本发明的一些实施例中，所述碱性溶液为 NaOH 溶液。

在本发明的一些实施例中，在确定所述待测核酸分子的所述第三区域的核酸序列时，不进行所述第一延伸处理。

根据本发明的第七方面，本发明提供了一种用于确定待测核酸分子的核酸序列的方法，所述待测核酸分子的插入片段为至少 700bp，所述方法包括：

（A）将所述待测核酸分子的插入片段划分为多个区域，所述多个区域的每一个均不超过 400bp，优选不超过 300bp，根据本发明第四方面任一实施例所述的方法，分别确定所述待测核酸分子多个区域每一个的核酸序列；

（B）基于步骤（A）的结果，确定所述核酸分子的核酸序列。

本发明所取得的有益效果为：利用本发明提供的测序方法，可以在低成本测序的基础上，实现对于长片段或者长片段中预定区域的核酸序列进行测定，而且精确度高。

本发明的附加方面和优点将在下面的描述中部分给出，部分将从下面的描述中变得明显，或通过本发明的实践了解到。

附图说明

本发明的上述和/或附加的方面和优点从结合下面附图对实施例的描述中将变得明显和容易理解，其中：

图 1 为根据本发明的一个实施例提供的用于核酸测序的核酸分子的示意图。

图 2 为根据本发明的一个实施例提供的用于核酸测序的核酸分子的示意图。

图 3 为根据本发明的一个实施例提供的用于确定待测核酸分子的核酸序列的方法中所述待测核酸分子的区域测定示意图。

图 4 为根据本发明的一个实施例提供的构建携带有插入片段的测序芯片的示意图。

图 5 为根据本发明的一个实施例提供的利用三轮测序实现长片段插入片段的测定的示意图。

图 6 为根据本发明的一个实施例提供的利用三轮测序实现长片段插入片段的测定的示意图。

图 7 为根据本发明的一个实施例提供的利用三轮测序实现长片段插入片段的测定的示意图。

5 图 8 为根据本发明的一个实施例提供的测序质量比 (Q30) 分布示意图。

图 9 为根据本发明的一个实施例提供的测序质量比 (Q30) 分布示意图。

具体实施方式

下面详细描述本发明的实施例，所述实施例的示例在附图中示出。下面通过参考附图
10 描述的实施例是示例性的，旨在用于解释本发明，而不能理解为对本发明的限制。在本发明的描述中，当表示核酸分子“相连”“连接”时，指的是两个核酸分子通过 3' ,5' 磷酸二酯键进行连接。

用于核酸测序的核酸分子

本发明提供了一种用于核酸测序的核酸分子，包括：插入片段、第一测序接头以及第二
15 测序接头，其中插入片段的长度为至少 500bp，任选至少为 800bp，例如可以为 1000~1200bp；第一测序接头与插入片段的一端相连；第二测序接头与插入片段的另一端相连。

根据本发明的一个实施例，用于核酸测序的核酸分子如图 1 所示，第一测序接头与插入
片段的 5' 端相连，第二测序接头与插入片段的 3' 端相连。根据本发明的实施例，第一测
序接头和第二测序接头的位置也可以互换，即第一测序接头可以与插入片段的 3' 端相连，
20 第二测序接头与插入片段的 5' 端相连。

用于核酸测序的核酸分子还可以包括标签序列，例如 barcode 序列，标签序列设置在第
一测序接头中或者第二测序接头中。如图 2 所示，标签序列设置在第一测序接头中。标签
序列用来区分不同的生物样本。在利用 Illumina 测序平台进行测序的过程中，核酸分子中的标签序
列和插入序列分别进行测定。因此可以将标签序列设置在第一测序接头或者第二测序接头中，远离插入
25 片段。利用测序引物对标签序列和插入片段分别进行测定。

基于如上用于核酸测序的核酸分子，本发明还提供了一种测序文库，所述测序文库含有
上述用于核酸测序的核酸分子。由上述用于核酸测序的核酸分子所构成的测序文库可以实
现对于核酸分子中插入片段的核酸序列进行测定，或者实现对于核酸分子中插入片段的预
定区域的核酸序列进行测定。例如可以借助于 Illumina HiSeq 2500 测序平台实现对于核酸分
30 子中插入片段的序列的测定，其中，第一测序接头和第二测序接头可以分别含有通用接头

P5 接头和 P7 接头。P5 接头和 P7 接头作为 Illumina HiSeq 2500 测序平台的锚定序列，可以将插入片段锚定到测序芯片上。

以 Illumina HiSeq2000 或者 Illumina HiSeq2500 测序平台为例，它们是基于边合成边测序技术进行测序的平台。通过利用单分子阵列实现在测序芯片（flow cell）上进行桥式 PCR 反应，并利用可逆阻断技术可以实现每次只合成一个碱基，并标记荧光基团，再利用相应的激光激发荧光基团，捕获激发光，从而读取碱基信息。类似地，可应用可逆阻断技术的测序平台都可以按照本申请的方法对长片段核酸进行文库构建和测序，例如也可以采用 Illumina 的 cBot 测序平台。当然，本领域技术人员可以根据不同的测序平台，选择相应的通用接头，来满足建库的要求。

基于如上用于核酸测序的核酸分子，本发明还提供了一种测序芯片，所述测序芯片上携带上述用于核酸测序的核酸分子。将上述用于核酸测序的核酸分子通过与测序芯片上的接头互补序列配对，使得测序芯片上携带有上述用于核酸测序的核酸分子。然后可以借助于测序平台实现对于核酸分子的测序。

用于确定待测核酸分子预定区域核酸序列的方法

根据本发明的另一方面，对于上述用于核酸测序的核酸分子，可以借助于如下方法实现对于核酸分子预定区域核酸序列的测定：

（1）利用第一 dNTP 混合物和测序引物，对所述待测核酸分子进行第一延伸处理，所述测序引物与所述第一测序接头和所述第二测序接头之一匹配，以便将所述测序引物延伸至所述预定区域的上游，其中，所述第一 dNTP 混合物包括 dATP、dTTP、dGTP 和 dCTP，所述 dATP、dTTP、dGTP 和 dCTP 的至少一种携带聚合酶反应阻断基团；

（2）利用第二 dNTP 混合物对所述第一延伸处理的产物进行第二延伸处理，并在所述第二延伸过程中，确定所述预定区域的核酸序列，其中所述第二 dNTP 混合物包括 dATP、dTTP、dGTP 和 dCTP，所述 dATP、dTTP、dGTP 和 dCTP 均携带聚合酶反应阻断基团，并且所述 dATP、dTTP、dGTP 和 dCTP 的至少之一携带可检测基团。

第一 dNTP 混合物中包括 dATP、dTTP、dGTP 和 dCTP，其中 dATP、dTTP、dGTP 和 dCTP 的至少一种携带有聚合酶反应阻断基团，将第一 dNTP 混合物和测序引物于待测核酸分子进行混合，测序引物能够与待测核酸分子的第一测序接头或者第二测序接头匹配，然后在聚合酶以及 dNTP 混合物的作用下，以待测核酸分子为模板，沿着测序引物进行聚合，直到将测序引物延伸至待测核酸分子的预定区域的上游。其中聚合酶反应阻断基团意指可以阻断相应的脱氧核糖核苷酸发生聚合反应的基团，例如可以是叠氮基团、磷酸酯、氢乙

烯或者基于二硫键的阻断基团。例如可以在 dGTP 上携带有聚合酶反应阻断基团，当 dGTP 发生聚合反应之后，需要对 dGTP 上携带的聚合酶反应阻断基团进行切除，然后继续进行后续聚合反应。通过切除反应发生的次数，能够确定将测序引物延伸至待测核酸分子的预定区域的上游。

5 根据本发明的另一种具体实施方式，所述第一 dNTP 混合物中所述 dATP、dTTP、dGTP 和 dCTP 中仅一种携带聚合酶反应阻断基团。由此，可以快速将测序引物延伸至待测核酸分子的预定区域的上游。

根据本发明的另一种具体实施方式，所述第一 dNTP 混合物中所述 dATP、dTTP、dGTP 和 dCTP 均携带聚合酶反应阻断基团。

10 根据本发明的一种具体实施方式，第一 dNTP 混合物中 dATP、dTTP、dGTP 和 dCTP 的至多三种携带聚合酶反应阻断基团。由此可以更精确将测序引物延伸至待测核酸分子的预定区域的上游，阻断效果更好。

其中，第一 dNTP 混合物中 dATP、dTTP、dGTP 和 dCTP 均不携带可检测基团。

根据本发明的具体实施例，所述第一延伸处理包括如下方法：（1）将所述待测核酸分子
15 与所述测序引物进行退火反应，以便形成起始双链体，所述起始双链体由所述待测核酸分子和所述测序引物构成；（2）以所述起始双链体中的测序引物为第一起始生长核酸链，在聚合酶的催化作用下，将所述第一 dNTP 混合物中的至少一个并入所述起始生长核酸链的 3'端，以便在所述起始生长核酸链的 3'端延伸新的碱基，形成第一产物双链体；（3）将所述第一产物双链体进行切割处理，以便去除所述第一产物双链体上的聚合酶反应阻断基团；
20 （4）重复步骤（2）~（3）直至所述测序引物延伸至所述预定区域的上游预定位置。由此，可以实现将测序引物延伸至所述预定区域的上游预定位置。

根据本发明的一种具体实施方式，所述预定位置可以位于预定区域的上游 1~100bp 范围内，例如可以是 1~80bp 范围内，也可以是 1~50bp、1~30bp、1~20bp 范围内。

根据本发明的另一种具体实施方式，所述预定位置与预定区域的下游末端的距离不超过
25 400bp，优选不超过 300bp 或者不超过 200bp。经过第一延伸处理，将测序引物延伸至预定区域的上游预定位置，预定位置与预定区域的下游末端的距离不超过 400bp，优选不超过 300bp 或者不超过 200bp，这样可以借助于测序平台，一次性实现对于预定区域的序列测定。

对于经过第一延伸处理的产物，即将测序引物延伸至预定区域的上游后，将第二 dNTP 混合物与第一延伸处理的产物混合进行第二延伸处理。其中，第二 dNTP 混合物包括 dATP、
30 dTTP、dGTP 和 dCTP，所述 dATP、dTTP、dGTP 和 dCTP 均携带聚合酶反应阻断基团，并

且所述 dATP、dTTP、dGTP 和 dCTP 的至少之一携带可检测基团。在延伸过程中，利用可检测基团以及聚合酶反应阻断基团可以实现对于每种碱基类型的精确测定。

根据本发明的一种具体实施方式，所述第二延伸处理包括：(a) 以所述第一延伸处理的产物为起始点，在聚合酶的催化作用下，将所述第二 dNTP 混合物中的一个并入所述第一延伸产物的 3'端，以便在所述第一延伸产物的 3'端延伸新的碱基，形成第二产物双链体，并通过对所述第二产物双链体的可检测基团进行检测，以便确定所述新的碱基的类型；(b) 将所述第二产物双链体进行切割处理，以便去除所述第二产物双链体上的聚合酶反应阻断基团和所述可检测基团；(c) 重复步骤 (a) ~ (b) 直至确定所述预定区域的至少一部分碱基类型。

根据本发明的一些实施例，所述可检测基团可以为荧光基团或者磷光基团。所述 dATP、dTTP、dGTP 和 dCTP 上分别均携带不同的所述可检测基团。所述荧光基团包括但不限于 AF532、IF700、Cy5 或者 ROX，所述磷光基团包括但不限于 zinc sulfide、strontium aluminate 或者 calcium sulfide。

基于以上用于确定待测核酸分子预定区域核酸序列的方法，可以实现对于待测核酸分子预定区域核酸序列的精确测定。同时，借助于该方法，还可以同时实现对于长片段插入片段的核酸序列的精确测定。例如可以实现对于长度为 700bp 以上，800bp 以上，1000bp 以上的核酸分子的序列测定。这种对于长片段核酸分子的序列测定，可以根据长片段核酸分子的长度，分段实现长片段核酸分子不同区域的序列的测定，即利用分段测序实现对于长片段核酸分子的序列测定。例如，当长片段核酸分子的长度在 1000bp 左右时，可以在第一轮测序中只测定待测核酸分子最末端（即 3' 端）的 330bp 左右的核酸序列，然后在第二轮测定中只测定待测核酸分子中间 330bp 左右的核酸序列，在第三轮测序中只测定待测核酸分子最前端（即 5' 端）330bp 左右的核酸序列，最后通过序列拼接获得长片段插入片段的核酸序列的精确测定。

为此，根据本发明的又一个方面，本发明提供了一种用于确定待测核酸分子的核酸序列的方法，所述待测核酸分子的插入片段为至少 700bp，所述插入片段分为第一区域、第二区域和第三区域，所述第一区域、第二区域和第三区域的片段大小大致相同，所述方法包括：

(1') 根据上述用于确定待测核酸分子预定区域核酸序列的方法，分别确定所述待测核酸分子第一区域、第二区域、第三区域的核酸序列；(2') 基于步骤 (1') 的结果，确定所述核酸分子的核酸序列。当插入片段的大小大约为 1000bp 时，如图 3 所示，所述第一区域的起始位置距离所述插入片段的上游末端至少 600bp，例如可以为 660bp，所述第二区域的起

始位置距离所述插入片段的上游末端至少 300bp，例如可以为 330bp，所述第三区域的起始位置为所述插入片段的上游末端。由此，通过三次测定，可以实现对于长片段的插入片段的序列测定。

以上通过对于待测核酸分子的不同区域的核酸序列进行测定，通过分段测定可以实现含有长片段插入片段的核酸分子的测定。当然在对于第三区域的核酸序列测定时，不再序列进行第一延伸处理。

在本发明的至少一些实施方式中，步骤（1'）进一步包括：（a）利用阻抑性引物和部分所述待测核酸分子结合，并根据上述用于确定待测核酸分子预定区域核酸序列的方法确定所述待测核酸分子第一区域的核酸序列，所述阻抑性引物与所述测序引物的核酸序列相同，所述阻抑性引物的 3' 端带有阻抑基团；（b）切除部分所述阻抑性引物的阻抑基团，并根据上述用于确定待测核酸分子预定区域核酸序列的方法确定所述待测核酸分子第二区域的核酸序列；（c）切除部分所述阻抑性引物的阻抑基团，并根据上述用于确定待测核酸分子预定区域核酸序列的方法确定所述待测核酸分子第三区域的核酸序列。

通过调整阻抑性引物和正常测序引物的比例，使得阻抑性引物和部分待测核酸分子结合，从而在第一轮测序时实现对于第一区域的核酸序列的测序，然后通过切除部分阻抑基团，在第二轮测序时测得与这些被切除阻抑基团互补结合的核酸分子上的第二区域的序列；依次类推，实现核酸分子的不同区域的序列测定。

在至少一些实施方式中，在完成第一区域和第二区域的核酸序列的测定后，分别利用 ddNTP 封端，再进行下一区域的核酸序列的测定。在对待测核酸分子每一区域的核酸序列完成测定后，可以采用 ddNTP 封端，以免在下一轮对于另一区域进行测序时，会对下一轮的测序产生信号干扰。具体而言，以一轮测序为例，例如在对待测核酸分子的第一区域进行测序时，按照上述用于确定待测核酸分子预定区域核酸序列的方法对待测核酸分子的第一区域进行测序。当利用第二 dNTP 混合物对第一延伸处理的产物进行第二延伸处理，并完后最后一个 dNTP 的聚合酶反应阻断基团的切除，并确定相应碱基后，利用 ddNTP 封端。然后再开始下一轮的测序。

在具体实施时，可以利用混合引物体系以及 Primer Walking（PW）技术实现不同区域的测序，即实现分段测序。混合引物体系为含有标准测序引物及阻抑性引物的混合物，该技术能确保实现在首轮测序中只针对 flowcell 上 DNA 簇中的约 1% DNA 链进行基因测序。其中，能够与混合物中的阻抑性引物互补结合的 DNA 链不发生引物的延伸，待到第二轮或者第三轮测序时，可以切割阻抑性引物上的阻抑基团后可以再做延伸。阻抑性引物与标准测

序引物的核酸序列相同，均能与 DNA 链互补结合，区别在于：阻抑性引物的 3' 末端带有阻抑基团，阻止后续的延伸反应。叠氮基团、磷酸酯、氢乙烯、或者基于二硫键的基团等只要可以阻止延伸反应，均可以作为阻抑基团。当然由于 dNTP 上携带有聚合酶反应阻断基团（叠氮基团），因此，叠氮基团不能作为阻抑性引物的阻抑基团，否则后续在切割 dNTP 上携带的聚合酶反应阻断基团时也会解除阻抑性引物的阻抑性。

以插入片段为 1000bp 左右为例，在第一轮测序的 primer walking (PW) 阶段，投放的 dNTP 的混合物主要有以下构成 dATP; dTTP; dCTP 以及携带有聚合酶反应阻断基团（叠氮基团）的 dGTP。聚合反应后都针对 dGTP 上携带的叠氮基团做切除反应以保证后续聚合反应的进行。在进行 N 个 cycle 的聚合/切除反应后，新合成链的长度将达到~660bp。以此为节点，投放带有荧光基团的 dNTP 混合物进行常规单端测序以获得插入片段末端 340bp 的基因序列。

下一轮测序开始前，针对部分阻抑性引物切除其阻抑基团。可以利用核酸内切酶 Endo IV 酶切除阻抑基团。在后续的 PW 阶段将聚合/切除反应的 cycle 数（约）减半，使新合成链的长度约达到 330bp。以此为节点，投放带有荧光基团的 dNTP 混合物进行常规单端测序以获得 DNA 插入片段中端约 330bp 的基因序列。

在第三轮测序开始前，切除剩余阻抑性引物所携带的阻抑基团。后投放带有荧光基团的 dNTP 混合物针对同一 DNA 簇中剩余 DNA 链进行常规单端测序以获得 DNA 插入片段前端约 330bp 的基因序列。最后将三轮测序得到的 DNA 序列进行分析并拼接成完整的 1000bp DNA 插入片段。

在至少一些实施方式中，步骤（1'）进一步包括：（a）根据上述用于确定待测核酸分子预定区域核酸序列的方法确定所述待测核酸分子第一区域的核酸序列，并利用碱性溶液对所述第二延伸处理的产物进行解链、洗脱，获得所述待测核酸分子；（b）根据上述用于确定待测核酸分子预定区域核酸序列的方法确定所述待测核酸分子第二区域的核酸序列，并利用碱性溶液对所述第二延伸处理的产物进行解链、洗脱，获得所述待测核酸分子；（c）根据上述用于确定待测核酸分子预定区域核酸序列的方法确定所述待测核酸分子第三区域的核酸序列。

在具体实施时，可以利用 Primer walking (PW) 技术以及新合成链洗脱技术实现上述对于不同区域的片段进行测序（即分段测序）。

以 1000bp 的插入片段为例，在第一轮测序中先将足量的标准测序引物与待测 DNA 链上的引物结合位点杂交互补，再投放由 dATP; dTTP; dCTP 以及阻抑性 dGTP 组成的 dNTP

混合物。聚合反应后都针对 dGTP 上携带的阻抑集团做切除反应以保证后续聚合反应的进行。在进行 N 个 cycle 的聚合/切除反应后（即 PW），新合成链的长度将达到~660bp。以此为节点，投放带有荧光集团的 dNTP 混合物进行常规单端测序以获得插入片段末端 340bp 的基因序列。第一轮测序完成后使用 NaOH 将双链解链，并将新合成链从芯片上洗脱。第二轮测序开始前同样先将足量的标准测序引物与待测 DNA 链上的引物结合位点杂交互补，再投放由 dATP; dTTP; dCTP 以及阻抑性 dGTP 组成的 dNTP 混合物。聚合反应后都针对 dGTP 上携带的阻抑集团做切除反应以保证后续聚合反应的进行。在进行 N 个 cycle 的聚合/切除反应后（即 PW），新合成链的长度将达到~330bp。以此为节点，投放带有荧光集团的 dNTP 混合物进行常规单端测序以获得插入片段中端 330bp 的基因序列。第二轮测序完成后使用 NaOH 将双链解链，并将新合成链从芯片上洗脱。

第三轮测序开始前将标准测序引物与待测 DNA 链上的引物结合位点杂交互补，后投放带有荧光集团的 dNTP 混合物针对同一 DNA 簇中剩余 DNA 链进行常规单端测序以获得 DNA 插入片段前端约 330bp 的基因序列。最后将三轮测序得到的 DNA 序列进行分析并拼接成完整的 1000bp DNA 插入片段。

利用 NaOH 碱性溶液将双链解链，并洗脱。相比较于通过传统退火的方式来解链，可以避免由于双链太长，退火温度过高，影响双链核酸分子中化学键的稳定性。同时也可以避免由于退火温度过高对于测序平台的破坏。

为此，根据本发明的另一个方面，本发明还提供了一种用于确定待测核酸分子的核酸序列的方法，所述待测核酸分子的插入片段为至少 700bp，所述方法包括：（A）将所述待测核酸分子的插入片段划分为多个区域，所述多个区域的每一个均不超过 400bp，优选不超过 300bp，根据上述用于确定待测核酸分子预定区域核酸序列的方法，分别确定所述待测核酸分子多个区域每一个的核酸序列；（B）基于步骤（A）的结果，确定所述核酸分子的核酸序列。在至少一种实施方式中，所述待测核酸分子的插入片段的长度至多 1500bp。当待测核酸分子的插入片段的长度过长，可能会影响测序的准确度。合适片段大小的插入片段在 1500bp 以下。在本发明的至少一些实施方式中，在完成每一区域的核酸序列的测定后，分别利用 ddNTP 封端，再进行下一区域的核酸序列的测定。从而避免不同区域核酸序列测定时，对于检测信号的干扰。

下面将结合实施例对本发明的方案进行解释。本领域技术人员将会理解，下面的实施例仅用于说明本发明，而不应视为限定本发明的范围。实施例中未注明具体技术或条件的，按照本领域内的文献所描述的技术或条件或者按照产品说明书进行。所用试剂或仪器未注

明生产厂商者，均为可以通过市购获得的常规产品。

实施例 1

利用 Illumina HiSeq 2500 测序平台，根据本发明提供的基因测序方法，通过分段测序使基因单端测序达到 1000bp 及以上的读长。

5 如图 4 所示，在长约 1000bp 的单链 DNA 片段的一端分别添加 barcode 序列、测序引物序列以及通用接头 P7 或 P5（作为接头序列），而 DNA 插入片段的另一端则直接添加 P5 或 P7（作为另一接头序列）以制备文库。测序芯片 Flowcell 上随机分布了这两种不同的寡核苷酸序列，分别与 P5 互补（即 P5'）；与 P7 一致（即 P7）。待测 DNA 片段通过自身携带的 P5 与 flowcell 上的 P5' 序列杂交互补，以待测 DNA 片段为模板进行互补链的延伸，互补链的两端分别为 P5' 和 P7'。接下来将模板链切断并洗脱只保留互补链，而互补链的 P7' 10 则与 Flowcell 上的 P7 杂交互补，进行链的合成即桥式 PCR。接下来合成的双链被解链，再分别与 Flowcell 上的接头杂交互补，重复进行延伸、解链、杂交，如此重复若干循环。桥式 PCR 完成后，使用 NaOH 将双链解链，并利用甲酰胺基嘧啶糖苷酶对 8-氧鸟嘌呤糖苷的选择性切断作用，选择性地将 P5' 与链的连接切断，留下与 Flowcell 上 P7 连接的链。同时 15 利用 ddNTPs，将游离的 3' 端被阻断，防止不必要的 DNA 延伸。至此芯片 flowcell 上已新合成若干相对独立的 DNA 簇（DNA cluster），每一个 DNA 簇则由携带同一 DNA 插入片段的上千拷贝 DNA 链组成。

然后按照图 5 所示，对长 DNA 文库片段进行分段测序，即在第一轮测序中只测序待检测 DNA 插入片段最末端的 330bp，在随后的第二轮测序中只测序待检测 DNA 插入片段的中端 330bp，在第三轮测序中只测序待检测 DNA 插入片段的前端 330bp，最后通过序列拼 20 接间接达到长读长的实验目的。

上述分段测序可以利用混合引物体系以及 Primer walking（PW）技术实现，如图 6 和图 7 所示。

其中，混合引物体系为标准测序引物及阻抑性引物的混合物，该项杂交技术能确保实现 25 在首轮测序中只针对 flowcell 上 DNA 簇中的约 1% DNA 链进行基因测序。其中，能够与混合物中的阻抑性引物互补结合的 DNA 链不发生引物的延伸，待到第二轮或者第三轮测序时，可以切割阻抑性引物上的阻抑基团后可以再做延伸。阻抑性引物与标准测序引物的核酸序列相同，均能与 DNA 链互补结合，区别在于：阻抑性引物的 3' 末端带有磷酸化修饰，可以阻止后续的延伸反应。

30 在第一轮测序的 primer walking（PW）阶段，投放的 dNTP 的混合物主要有以下构成

dATP; dTTP; dCTP 以及携带有聚合酶反应阻断基团（叠氮基团）的 dGTP。聚合反应后都针对 dGTP 上携带的叠氮基团做切除反应以保证后续聚合反应的进行。在进行 N 个 cycle 的聚合/切除反应后，新合成链的长度将达到~660bp。以此为节点，投放带有荧光基团的 dNTP 混合物进行常规单端测序以获得插入片段末端 340bp 的基因序列。

- 5 下一轮测序开始前，针对部分阻抑性引物切除其阻抑基团。可以利用核酸内切酶 Endo IV 酶切除阻抑基团。在后续的 PW 阶段将聚合/切除反应的 cycle 数（约）减半，使新合成链的长度约达到 330bp。以此为节点，投放带有荧光基团的 dNTP 混合物进行常规单端测序以获得 DNA 插入片段中端约 330bp 的基因序列。

- 10 在第三轮测序开始前，切除剩余阻抑性引物所携带的阻抑基团。后投放带有荧光基团的 dNTP 混合物针对同一 DNA 簇中剩余 DNA 链进行常规单端测序以获得 DNA 插入片段前端约 330bp 的基因序列。

最后将三轮测序得到的 DNA 序列进行分析并拼接成完整的 1000bp DNA 插入片段。

下面通过对大肠杆菌序列进行测定和分析来验证上述方法的精确性，其中具体实施材料和实验步骤以及实验结果如下所示：

15 1. 实验材料

- 1). 大肠杆菌单链 DNA 模板（约 1000 碱基长度）
- 2). AmpliSeq for Illumina 文库制备、标签接头及配套产品
- 3). HiSeq SR 簇生成试剂盒 v4-cBot
- 4). HiSeq SBS 试剂盒 V4

- 20 5). 阻抑性引物与标准测序引物序列一致，但在 3' 生长端做磷酸化修饰，并由上海生工合成提供。其中阻抑性引物序列为 5' -GCGTAGCGCGGAACCATGGTTCCGCGCTACGGCTCACAGAACGACATGGCTACGATC CGACTT-3' P。

- 6). 核酸内切酶 4，购买自 NEB。

25 2. 实验步骤

1). 参照 AmpliSeq for Illumina 制备文库、添加标签接头试剂盒说明书，制备用于 Illumina 测序仪的扩增子文库。

2). 将上一步骤制备好的文库参照 HiSeq SR 簇生成试剂盒 v4-cBot 说明书，loading 到 flow cell 上并进行等温扩增生成每簇具有大约 1000 个拷贝的克隆簇，以供在支持 HiSeq 的系统
30 上进行测序。

3).将混合引物体系即标准测序引物及阻抑性引物的混合物与 flowcell 上 DNA 链上的引物结合位点进行互补杂交。

4).测序仪启动：参照 Illumina 基因测序仪使用说明书，启动测序仪对应的内置控制程序（单端测序）。

5 5).仪器清洗：用清洗芯片进行 Illumina 基因测序仪的流道清洗。具体请参照仪器使用说明书。

6).PW 试剂装载：将 HiSeq SBS 试剂盒 V4 中的荧光 dNTP 混合物替换成非荧光 dATP+dTTP+dCT+带有叠氮阻抑集团的 dGTP 混合物，后将带有 PW 试剂的试剂盒放入测序仪冰箱。

10 7). 试剂预载：参照 Illumina 基因测序仪使用说明书进行试剂预载。

8). 安装待测芯片（上述步骤 2 制备好的待测芯片）开始 PW 并运行 167 轮（660bp）聚合/切除反应。

9). PW 完成以后，取出试剂槽。

10).测序试剂装载：将 HiSeq SBS 试剂盒 V4 放入测序仪冰箱。

15 11). 试剂预载：参照基因测序仪 Illumina 使用说明书进行试剂预载。

12). 首轮测序：进行分段测序流程中的针对 DNA 插入片段末端 330bp 的测序。在首轮测序结束后对新合成的 DNA 链采用 ddNTP 混合物进行末端阻断。

13). 恢复部分阻抑性测序引物聚合能力：将核酸内切酶 IV（购买自 NEB）与商品配套缓冲液按照 1:100 稀释后通入芯片中，在 37 摄氏度下孵育 5 分钟。采用核酸内切酶 IV 对部分阻抑性引物 3' 生长端修饰的磷酸集团做切除以恢复其生长聚合能力。14).按顺序重复上述步骤 6)，7）并开始第二轮 PW 运行 84 轮（约 340bp）聚合/切除反应。

15).按顺序重复上述步骤 9）-11）并进行分段测序流程中的针对 DNA 插入片段中端 330bp 的第二轮测序。在此轮测序结束后对新合成的 DNA 链再次采用 ddNTP 混合物进行末端阻断。

25 16). 恢复剩余阻抑性测序引物聚合能力：采用核酸内切酶 4 对剩余阻抑性引物 3' 生长端修饰的磷酸集团做切除以恢复其生长聚合能力。

17).按顺序重复上述步骤 10）-11）并进行分段测序流程中的针对 DNA 插入片段最前端 330bp 的最终轮测序。

18).参照 Illumina 基因测序仪使用说明书清洗仪器。

30 19).将三轮测序得到的 DNA 序列进行分析并拼接成完整的 1000bp DNA 插入片段。

3.实验结果

对比现有单端测序技术（即利用Illumina测序平台对大肠杆菌中长度为400bp的插入片段进行测序），该400bp的插入片段对应于本发明分段测序的第一段序列和部分中段序列。比对结果如下：

表1 测序结果

	单 端 测 序 400bp	分段测序第一轮 测序（330bp）	分段测序第二 轮测序（330bp）	分段测序第三 轮（340bp）
Q30%	78.57	91.5	91.8	92.6
ESR%	87.84	88.37	89.02	89.75
Mapping Rate%	88.9	95.3	96.1	96.3
Avg Error Rate%	1.96	0.19	0.16	0.16

其中 Q30 的结果图如图 8 所示。由表 1 和图 8 可以看出，分段测序技术在各项重要测序质量指标的表现，包括 Q30；ESR；mapping rate 以及平均错误率（Avg Error Rate）均明显优于前者且超过目前对测序质量的要求，同时满足了对长序列读取的需求。其中，表 1 中 Q30（Quality Proportion Distribution）代表所测碱基质量，即利用 basecall 测序软件结果错误率低于 0.001（准确性高于 99.9%）的 base 占有比率。ESR（effective spot rate）则是为了排除没有 loading 上或者发光不正常的 DNB，保证结果的准确性，对每个 DNB 的每个碱基计算质量值，通过质量值计算决定 reads 是否会因为质量较低而被过滤。ESR 值越大，代表质量合格的 reads 数越多。Mapping rate 代表了能够准确比对到参考序列（reference 序列）上的 reads 占 total reads 的比率。Mapping rate 值越大，代表能够准确比对到参考序列上的 reads 数占比越大。平均错误率（Avg Error Rate%）代表错误匹配的比率，为整条 lane 各位点错误匹配的平均值。平均错误率值越小，测序效果越好。

借助于 Illumina 测序平台，对于同样大小例如，400bp 的插入片段进行测序，利用本发明所提供的方法所获得的测序结果更加精确。不受理论的限制，猜测原因可能有：1）聚合所用的 dNTPs 是不一样的，常规单端测序每一轮的聚合都是采用的带有修饰的阻抑性 dNTPs，而新方法则是 3/4 的 dNTPs 是天然的（native）的，聚合酶针对带修饰的和天然的 dNTPs 的聚合效率和准确性可能是不一样的；2）碱基错配是会随着引物延伸而累积的，引物延伸重复的次数越少碱基错配可能性的概率越低。新方法在引物延伸这一步骤预估每一轮可聚合四个碱基，是常规单端测序方法的四倍。

实施例 2

实施例 2 提供了利用 Primer walking (PW) 技术以及新合成链洗脱技术实现上述分段测序, 并通过对大肠杆菌序列进行测定与分析, 进行了验证。

在第一轮测序中先将足量的标准测序引物与待测 DNA 链上的引物结合位点杂交互补, 再投放由 dATP; dTTP; dCTP 以及阻抑性 dGTP 组成的 dNTP 混合物。聚合反应后都针对 dGTP 上携带的阻抑集团做切除反应以保证后续聚合反应的进行。在进行 N 个 cycle 的聚合/切除反应后 (即 PW), 新合成链的长度将达到~660bp。以此为节点, 投放带有荧光集团的 dNTP 混合物进行常规单端测序以获得插入片段末端 340bp 的基因序列。第一轮测序完成后使用 NaOH 将双链解链, 并将新合成链从芯片上洗脱。

第二轮测序开始前同样先将足量的标准测序引物与待测 DNA 链上的引物结合位点杂交互补, 再投放由 dATP; dTTP; dCTP 以及阻抑性 dGTP 组成的 dNTP 混合物。聚合反应后都针对 dGTP 上携带的阻抑集团做切除反应以保证后续聚合反应的进行。在进行 N 个 cycle 的聚合/切除反应后 (即 PW), 新合成链的长度将达到~330bp。以此为节点, 投放带有荧光集团的 dNTP 混合物进行常规单端测序以获得插入片段中端 330bp 的基因序列。第二轮测序完成后使用 NaOH 将双链解链, 并将新合成链从芯片上洗脱。

第三轮测序开始前将标准测序引物与待测 DNA 链上的引物结合位点杂交互补, 后投放带有荧光集团的 dNTP 混合物针对同一 DNA 簇中剩余 DNA 链进行常规单端测序以获得 DNA 插入片段前端约 330bp 的基因序列。最后将三轮测序得到的 DNA 序列进行分析并拼接成完整的 1000bp DNA 插入片段。

具体实验材料和实施步骤以及实验结果如下:

1. 实验材料

- 1). 大肠杆菌单链 DNA 模板 (约 1000 碱基长度)
- 2). AmpliSeq for Illumina 文库制备、标签接头及配套产品
- 3). HiSeq SR 簇生成试剂盒 v4-cBot
- 4). HiSeq SBS 试剂盒 V4

2. 实验步骤

1). 参照 AmpliSeq for Illumina 制备文库、添加标签接头试剂盒说明书, 制备用于 Illumina 测序仪的扩增子文库。

2). 将上一步骤制备好的文库参照 HiSeq SR 簇生成试剂盒 v4-cBot 说明书, loading 到 flow cell 上并进行等温扩增生成每簇具有大约 1000 个拷贝的克隆簇, 以供在支持 HiSeq 的系统

上进行测序。

3).将过量标准测序引物与 flowcell 上 DNA 链上的引物结合位点进行互补杂交。

4).测序仪启动：参照 Illumina 基因测序仪使用说明书，启动测序仪对应的内置控制程序（单端测序）。

5 5).仪器清洗：用清洗芯片进行 Illumina 基因测序仪的流道清洗。具体请参照仪器使用说明书。

6).PW 试剂装载：将 HiSeq SBS 试剂盒 V4 中的荧光 dNTP 混合物替换成非荧光 dATP+dTTP+dCT+带有叠氮阻抑集团的 dGTP 混合物，后将带有 PW 试剂的试剂盒放入测序仪冰箱。

10 7).试剂预载：参照 Illumina 基因测序仪使用说明书进行试剂预载。

8).安装待测芯片（上述步骤 2）制备好的待测芯片）开始 PW 并运行 167 轮（660bp）聚合/切除反应。

9).PW 完成以后，取出试剂槽。

10).测序试剂装载：将 HiSeq SBS 试剂盒 V4 放入测序仪冰箱。

15 11).试剂预载：参照基因测序仪 Illumina 使用说明书进行试剂预载。

12).首轮测序：进行分段测序流程中的针对 DNA 插入片段末端 330bp 的测序。在首轮测序结束后对新合成的 DNA 链采用 ddNTP 混合物进行末端阻断。

13).采用 0.5M NaOH 孵育 10 分钟使双链 DNA 变性后将上述测序步骤中的新合成链洗脱。

20 14).按顺序重复上述步骤 6, 7 并开始第二轮 PW 运行 84 轮（约 340bp）聚合/切除反应。

15).按顺序重复上述步骤 9）-11）并进行分段测序流程中的针对 DNA 插入片段中端 330bp 的第二轮测序。在此轮测序结束后对新合成的 DNA 链再次采用 ddNTP 混合物进行末端阻断。

25 16).再次采用 0.5M NaOH 孵育 10 分钟使双链 DNA 变性后将第二轮测序步骤中的新合成链洗脱。

17).按顺序重复上述步骤 10）-11）并进行分段测序流程中的针对 DNA 插入片段最前端 330bp 的最终轮测序。

18).参照 Illumina 基因测序仪使用说明书清洗仪器。

19).将三轮测序得到的 DNA 序列进行分析并拼接成完整的 1000bp DNA 插入片段。

30 3.实验结果

对比现有单端测序技术（即利用 Illumina 测序平台对大肠杆菌中长度为 400bp 的插入片

段进行测序), 该 400bp 的插入片段对应于本发明分段测序的第一段序列和部分中段序列。
 比对结果如下:

表 2 测序结果

	单 端 测 序 400bp	分段测序第一轮 测序 330bp	分段测序第二轮 测序 330bp	分段测序第三 轮测序 340bp
Q30%	81.12	92.6	93.0	93.7
ESR%	88.56	89.77	90.63	90.98
Mapping rate%	89.33	95.7	96.4	96.9
Avg Error rate%	1.82	0.18	0.14	0.15

其中 Q30 的结果图如图 9 所示。表 2 中所示出的 Q30、ESR、Mapping rate 以及 Avg Error
 rate 等与表 1 示出的具有相同的含义。从表 2 和图 9 可以看出, 分段测序技术在各项重要测
 序质量指标的表现, 包括 Q30; ESR; mapping rate 以及平均错误率 (Avg Error Rate) 均明
 显优于前者且超过目前对测序质量的要求, 同时满足了对长序列读取的需求。

在本发明的描述中, 术语“第一”、“第二”等仅用于描述目的, 而不能理解为指示或
 暗示相对重要性或者隐含指明所指示的技术特征的数量。由此, 限定有“第一”、“第二”
 的特征可以明示或者隐含地包括至少一个该特征。在本发明的描述中, “多个”的含义是至
 少两个, 例如两个, 三个等, 除非另有明确具体的限定。

在本说明书的描述中, 参考术语“一个实施例”、“一些实施例”、“示例”、“具体示例”、
 或“一些示例”等的描述意指结合该实施例或示例描述的具体特征、结构、材料或者特点
 包含于本发明的至少一个实施例或示例中。在本说明书中, 对上述术语的示意性表述不必
 须针对的是相同的实施例或示例。而且, 描述的具体特征、结构、材料或者特点可以在任
 一个或多个实施例或示例中以合适的方式结合。此外, 在不相互矛盾的情况下, 本领域的
 技术人员可以将本说明书中描述的不同实施例或示例以及不同实施例或示例的特征进行结
 合和组合。

尽管上面已经示出和描述了本发明的实施例, 可以理解的是, 上述实施例是示例性的,
 不能理解为对本发明的限制, 本领域的普通技术人员在本发明的范围内可以对上述实施例
 进行变化、修改、替换和变型。

权利要求书

1、一种用于确定待测核酸分子预定区域核酸序列的方法，其特征在于，所述待测核酸分子包括插入片段，所述插入片段的长度大于测序仪的读长，所述方法包括：

5 使所述插入片段的两端分别带有第一测序接头和第二测序接头，所述第一测序接头与所述插入片段的一端相连，所述第二测序接头与所述插入片段的另一端相连；

利用第一 dNTP 混合物和测序引物，对所述待测核酸分子进行第一延伸处理，所述测序引物与所述第一测序接头和所述第二测序接头之一匹配，以便将所述测序引物延伸至所述预定区域的上游；

10 利用第二 dNTP 混合物对所述第一延伸处理的产物进行第二延伸处理，并在所述第二延伸过程中，确定所述预定区域的核酸序列，

其中，所述第一 dNTP 混合物包括 dATP、dTTP、dGTP 和 dCTP，所述 dATP、dTTP、dGTP 和 dCTP 的至少一种携带聚合酶反应阻断基团；

15 所述第二 dNTP 混合物包括 dATP、dTTP、dGTP 和 dCTP，所述 dATP、dTTP、dGTP 和 dCTP 均携带聚合酶反应阻断基团，并且所述 dATP、dTTP、dGTP 和 dCTP 的至少之一携带可检测基团。

2、根据权利要求 1 所述的方法，其特征在于，所述第一延伸处理包括：

(1) 将所述待测核酸分子与所述测序引物进行退火反应，以便形成起始双链体，所述起始双链体由所述待测核酸分子和所述测序引物构成；

20 (2) 以所述起始双链体中的引物为第一起始生长核酸链，在聚合酶的催化作用下，将所述第一 dNTP 混合物中的至少一个并入所述起始生长核酸链的 3'端，以便在所述起始生长核酸链的 3'端延伸新的碱基，形成第一产物双链体；

(3) 将所述第一产物双链体进行切割处理，以便去除所述第一产物双链体上的聚合酶反应阻断基团；

25 (4) 重复步骤 (2) ~ (3) 直至所述测序引物延伸至所述预定区域的上游预定位置。

3、根据权利要求 1 所述的方法，其特征在于，所述第一 dNTP 混合物中所述 dATP、dTTP、dGTP 和 dCTP 的全部携带聚合酶反应阻断基团。

4、根据权利要求 1 所述的方法，其特征在于，所述第一 dNTP 混合物中所述 dATP、dTTP、dGTP 和 dCTP 的仅一种携带聚合酶反应阻断基团。

30 5、根据权利要求 1 所述的方法，其特征在于，所述第一 dNTP 混合物中所述 dATP、dTTP、

dGTP 和 dCTP 的至多三种携带聚合酶反应阻断基团。

6、根据权利要求 1 所述的方法，其特征在于，所述第一 dNTP 混合物中所述 dATP、dTTP、dGTP 和 dCTP 均不携带可检测基团。

7、根据权利要求 1 所述的方法，其特征在于，在步骤（4）中，所述预定位置位于所述
5 预定区域的上游 1~100bp 范围内。

8、根据权利要求 1 所述的方法，其特征在于，所述第一延长处理产物的长度为至少 200bp。

9、根据权利要求 1 所述的方法，其特征在于，所述第一延长处理产物的长度为 200~600bp。

10 10、根据权利要求 1 所述的方法，其特征在于，在步骤（4）中，所述预定位置与所述预定区域下游末端的距离不超过 400bp。

11、根据权利要求 1 所述的方法，其特征在于，在步骤（4）中，所述预定位置与所述预定区域下游末端的距离不超过 300bp。

12、根据权利要求 1 所述的方法，其特征在于，所述第二延伸处理包括：

15 （a）以所述第一延伸处理的产物为起始点，在聚合酶的催化作用下，将所述第二 dNTP 混合物中的一个并入所述第一延伸产物的 3'端，以便在所述第一延伸产物的 3'端延伸新的碱基，形成第二产物双链体，并通过对所述第二产物双链体的可检测基团进行检测，以便确定所述新的碱基的类型；

（b）将所述第二产物双链体进行切割处理，以便去除所述第二产物双链体上的聚合酶
20 反应阻断基团和所述可检测基团；

（c）重复步骤（a）~（b）直至确定所述预定区域的至少一部分碱基类型。

13、根据权利要求 12 所述的方法，其特征在于，所述可检测基团为荧光基团或磷光基团。

14、根据权利要求 12 所述的方法，其特征在于，所述 dATP、dTTP、dGTP 和 dCTP 均
25 携带不同的所述可检测基团。

15、用于权利要求 1~14 中任一项所述的测序方法的 dNTP 混合物。

16、一种用于核酸测序的核酸分子，其特征在于，包括：

插入片段，所述插入片段的长度为至少 500bp，优选至少 800bp，例如 1000~1200bp；

第一测序接头，所述第一测序接头与所述插入片段的一端相连；

30 第二测序接头，所述第二测序接头与所述插入片段的另一端相连。

17、根据权利要求 16 所述的核酸分子，其特征在于，进一步包括：

标签序列，所述标签序列设置在所述第一测序接头中或者所述第二测序接头中。

18、一种测序文库，所述测序文库中含有权利要求 16 或 17 所述的核酸分子。

19、一种测序芯片，其特征在于，所述测序芯片上携带权利要求 16 或 17 所述的核酸分子。

20、一种用于确定待测核酸分子的核酸序列的方法，所述待测核酸分子的插入片段为至少 700bp，其特征在于，所述待测核酸分子的插入片段区分为第一区域、第二区域和第三区域，所述第一区域、第二区域和第三区域依次相连，所述第一区域、第二区域和第三区域的片段大小大致相同，所述方法包括：

（1'）根据权利要求 1~14 任一项所述的方法，分别确定所述待测核酸分子第一区域、第二区域、第三区域的核酸序列；

（2'）基于步骤（1'）的结果，确定所述核酸分子的核酸序列。

21、根据权利要求 20 所述的方法，其特征在于，所述待测核酸分子的插入片段为大约 1000bp，所述第一区域的起始位置距离所述插入片段的上游末端至少 600bp，例如为 660bp，所述第二区域的起始位置距离所述插入片段的上游末端至少 300bp，例如为 330bp，所述第三区域的起始位置为所述插入片段的上游末端。

22、根据权利要求 20 所述的方法，其特征在于，步骤（1'）进一步包括：

（a）利用阻抑性引物和部分所述待测核酸分子结合，并根据权利要求 1~14 任一项所述的方法确定所述待测核酸分子第一区域的核酸序列，所述阻抑性引物与所述测序引物的核酸序列相同，所述阻抑性引物的 3' 端带有阻抑基团；

（b）切除部分所述阻抑性引物的阻抑基团，并根据权利要求 1~14 任一项所述的方法确定所述待测核酸分子第二区域的核酸序列；

（c）切除部分所述阻抑性引物的阻抑基团，并根据权利要求 1~14 任一项所述的方法确定所述待测核酸分子第三区域的核酸序列。

23、根据权利要求 22 所述的方法，其特征在于，利用核酸内切酶切除所述阻抑性引物的阻抑基团。

24、根据权利要求 20 所述的方法，其特征在于，步骤（1'）进一步包括：

（a）根据权利要求 1~14 任一项所述的方法确定所述待测核酸分子第一区域的核酸序列，并利用碱性溶液对所述第二延伸处理的产物进行解链、洗脱，获得所述待测核酸分子；

（b）根据权利要求 1~14 任一项所述的方法确定所述待测核酸分子第二区域的核酸序

列，并利用碱性溶液对所述第二延伸处理的产物进行解链、洗脱，获得所述待测核酸分子；

(c) 根据权利要求 1~14 任一项所述的方法确定所述待测核酸分子第三区域的核酸序列。

25、根据权利要求 24 所述的方法，其特征在于，所述碱性溶液为 NaOH 溶液。

5 26、根据权利要求 22~25 中任一项所述的方法，其特征在于，在确定所述待测核酸分子的所述第三区域的核酸序列时，不进行所述第一延伸处理。

27、一种用于确定待测核酸分子的核酸序列的方法，所述待测核酸分子的插入片段为至少 700bp，其特征在于，所述方法包括：

10 (A) 将所述待测核酸分子的插入片段划分为多个区域，所述多个区域的每一个均不超过 400bp，优选不超过 300bp，根据权利要求 1~14 任一项所述的方法，分别确定所述待测核酸分子多个区域每一个的核酸序列；

 (B) 基于步骤 (A) 的结果，确定所述核酸分子的核酸序列。



图 1



图 2

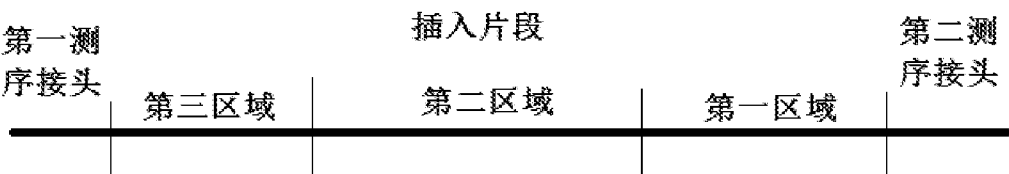


图 3

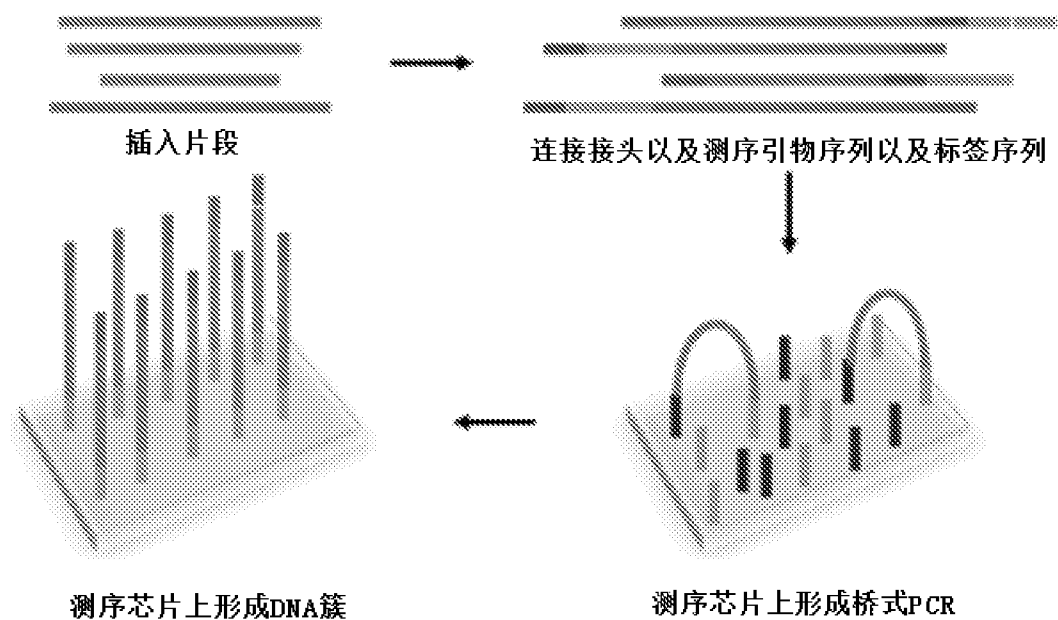


图 4

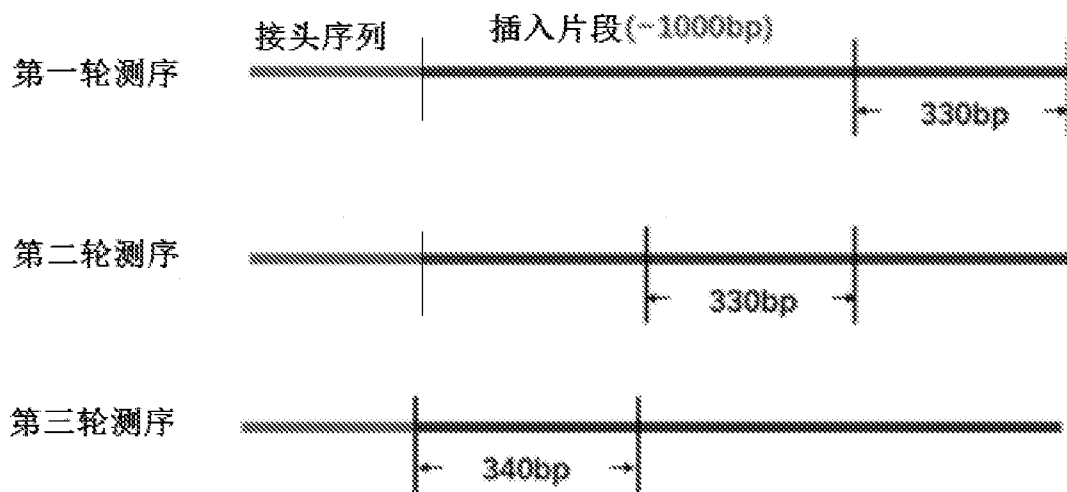


图 5

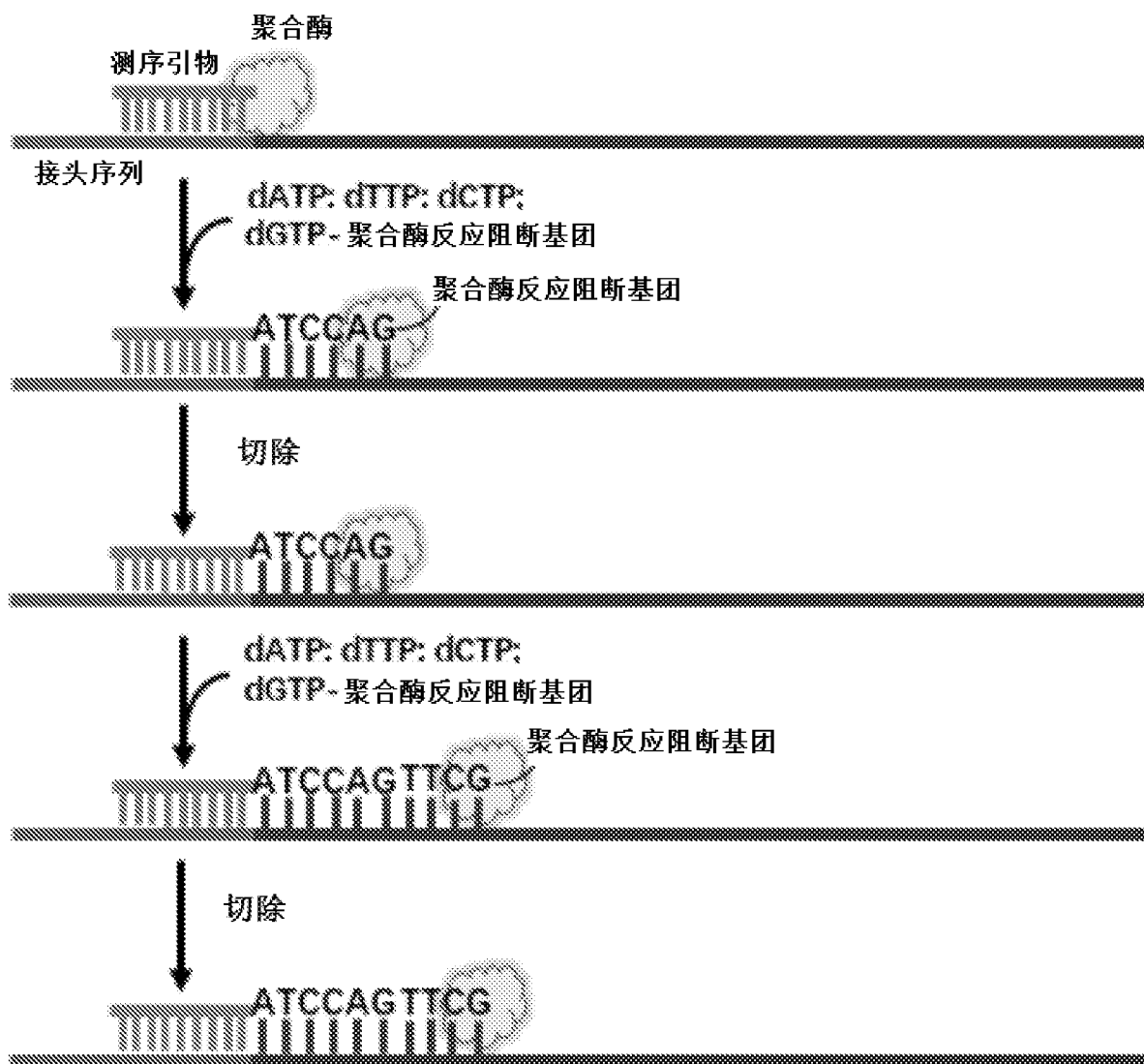


图 6

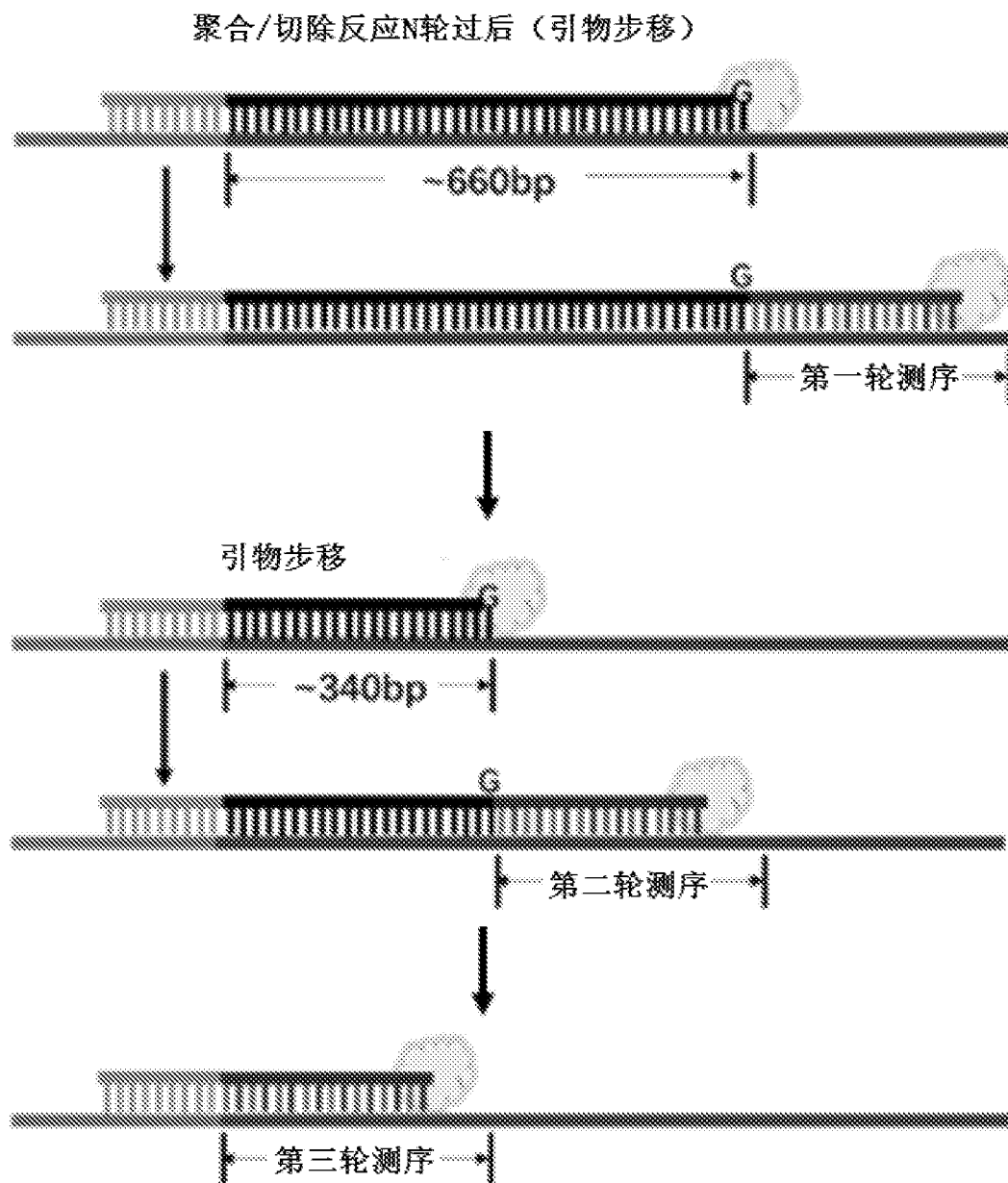


图 7

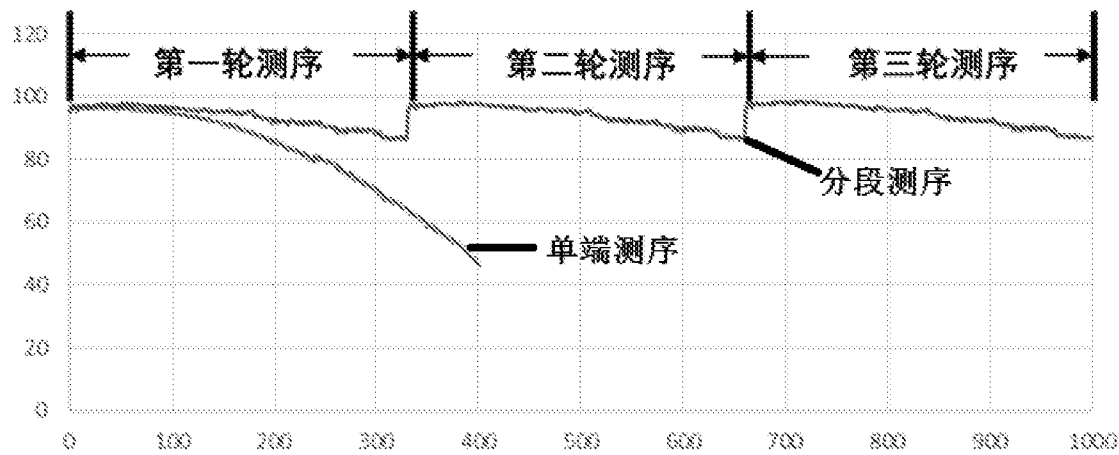


图 8

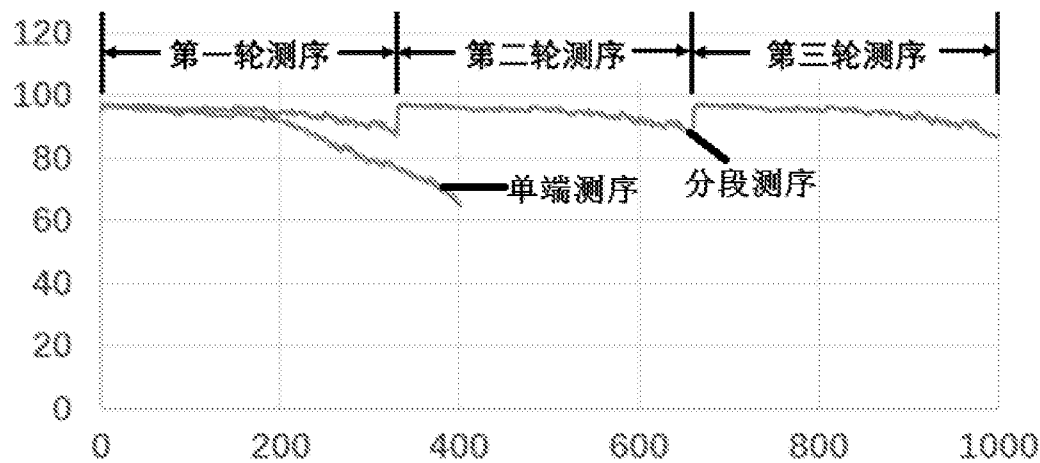


图 9

INTERNATIONAL SEARCH REPORT

International application No.

PCT/CN2018/119857

A. CLASSIFICATION OF SUBJECT MATTER

C40B 50/06(2006.01)i; C12Q 1/68(2018.01)i

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

C04B; C12Q

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

CNABS, CPRSABS, SIPOABS, DWPI, CNTXT, WOTXT, EPTXT, USTXT, CNKI, 百度搜索, BAIDU SCHOLAR SEARCH, WEB OF SCIENCE, PubMed: 接头, 引物, 插入, 测序, 序列, dNTP, insert+, sequencing linkers, primer

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	WO 0166802 A1 (GENETAG TECHNOLOGY INC. et al.) 13 September 2001 (2001-09-13) see entire document	1-27
A	CN 104404160 A (NANJING UNIVERSITY) 11 March 2015 (2015-03-11) see entire document	1-27
A	CN 105506063 A (BGI SHENZHEN CO., LTD.) 20 April 2016 (2016-04-20) see entire document	1-27
A	CN 105525357 A (BGI SHENZHEN CO., LTD.) 27 April 2016 (2016-04-27) see entire document	1-27
A	CN 105624272 A (BGI SHENZHEN CO., LTD.) 01 June 2016 (2016-06-01) see entire document	1-27



Further documents are listed in the continuation of Box C.



See patent family annex.

* Special categories of cited documents:

“A” document defining the general state of the art which is not considered to be of particular relevance

“E” earlier application or patent but published on or after the international filing date

“L” document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

“O” document referring to an oral disclosure, use, exhibition or other means

“P” document published prior to the international filing date but later than the priority date claimed

“T” later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

“X” document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

“Y” document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

“&” document member of the same patent family

Date of the actual completion of the international search

03 September 2019

Date of mailing of the international search report

12 September 2019

Name and mailing address of the ISA/CN

State Intellectual Property Office of the P. R. China
No. 6, Xitucheng Road, Jimenqiao Haidian District, Beijing
100088
China

Facsimile No. (86-10)62019451

Authorized officer

Telephone No.

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.

PCT/CN2018/119857

Patent document cited in search report			Publication date (day/month/year)	Patent family member(s)			Publication date (day/month/year)
WO	0166802	A1	13 September 2001	None			
CN	104404160	A	11 March 2015	CN	104404160	B	09 October 2018
CN	105506063	A	20 April 2016	None			
CN	105525357	A	27 April 2016	HK	1223660	A1	04 August 2017
				CN	105525357	B	21 August 2018
CN	105624272	A	01 June 2016	None			

国际检索报告

国际申请号

PCT/CN2018/119857

A. 主题的分类

C40B 50/06 (2006.01) i; C12Q 1/68 (2018.01) i

按照国际专利分类 (IPC) 或者同时按照国家分类和 IPC 两种分类

B. 检索领域

检索的最低限度文献 (标明分类系统和分类号)

C04B; C12Q

包含在检索领域中的除最低限度文献以外的检索文献

在国际检索时查阅的电子数据库 (数据库的名称, 和使用的检索词 (如使用))

CNABS, CPRSABS, SIPOABS, DWPI, CNTXT, WOTXT, EPTXT, USTXT, CNKI, 百度学术搜索, WEB OF SCIENCE, PubMed: 接头, 引物, 插入, 测序, 序列, dNTP, insert+, sequencing linkers, primer

C. 相关文件

类 型*	引用文件, 必要时, 指明相关段落	相关的权利要求
A	WO 0166802 A1 (GENETAG TECHNOLOGY INC等) 2001年 9月 13日 (2001 - 09 - 13) 参考全文	1-27
A	CN 104404160 A (南京大学) 2015年 3月 11日 (2015 - 03 - 11) 参考全文	1-27
A	CN 105506063 A (深圳华大基因科技有限公司) 2016年 4月 20日 (2016 - 04 - 20) 参考全文	1-27
A	CN 105525357 A (深圳华大基因股份有限公司) 2016年 4月 27日 (2016 - 04 - 27) 参考全文	1-27
A	CN 105624272 A (深圳华大基因科技有限公司) 2016年 6月 1日 (2016 - 06 - 01) 参考全文	1-27

☐ 其余文件在C栏的续页中列出。☒ 见同族专利附件。

* 引用文件的具体类型:

“A” 认为不特别相关的表示了现有技术一般状态的文件

“E” 在国际申请日的当天或之后公布的在先申请或专利

“L” 可能对优先权要求构成怀疑的文件, 或为确定另一篇引用文件的公布日而引用的或者因其他特殊理由而引用的文件 (如具体说明的)

“O” 涉及口头公开、使用、展览或其他方式公开的文件

“P” 公布日先于国际申请日但迟于所要求的优先权日的文件

“T” 在申请日或优先权日之后公布, 与申请不相抵触, 但为了理解发明之理论或原理的在后文件

“X” 特别相关的文件, 单独考虑该文件, 认定要求保护的发明不是新颖的或不具有创造性

“Y” 特别相关的文件, 当该文件与另一篇或者多篇该类文件结合并且这种结合对于本领域技术人员为显而易见时, 要求保护的发明不具有创造性

“&” 同族专利的文件

国际检索实际完成的日期

2019年 9月 3日

国际检索报告邮寄日期

2019年 9月 12日

ISA/CN的名称和邮寄地址

中国国家知识产权局 (ISA/CN)

中国北京市海淀区蓟门桥西土城路6号 100088

传真号 (86-10)62019451

受权官员

赵彦豪

电话号码 62411043

国际检索报告
关于同族专利的信息

国际申请号

PCT/CN2018/119857

检索报告引用的专利文件			公布日 (年/月/日)	同族专利	公布日 (年/月/日)
WO	0166802	A1	2001年 9月 13日	无	
CN	104404160	A	2015年 3月 11日	CN 104404160	B 2018年 10月 9日
CN	105506063	A	2016年 4月 20日	无	
CN	105525357	A	2016年 4月 27日	HK 1223660	A1 2017年 8月 4日
				CN 105525357	B 2018年 8月 21日
CN	105624272	A	2016年 6月 1日	无	