

(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
2017 年 8 月 31 日 (31.08.2017)



(10) 国际公布号
WO 2017/143585 A1

- (51) 国际专利分类号:
C12Q 1/68 (2006.01) G06F 19/20 (2011.01)
- (21) 国际申请号: PCT/CN2016/074665
- (22) 国际申请日: 2016 年 2 月 26 日 (26.02.2016)
- (25) 申请语言: 中文
- (26) 公布语言: 中文
- (71) 申请人: 深圳华大基因研究院 (BGI SHENZHEN) [CN/CN]; 中国广东省深圳市盐田区北山工业区综合楼, Guangdong 518083 (CN)。
- (72) 发明人: 谢寅龙 (XIE, Yinlong); 中国广东省深圳市盐田区北山工业区综合楼, Guangdong 518083 (CN)。 黄伟华 (HUANG, Weihua); 中国广东省深圳市盐田区北山工业区综合楼, Guangdong 518083 (CN)。 李净净 (LI, Jingjing); 中国广东省深圳市盐田区北山工业区综合楼, Guangdong 518083 (CN)。 郭瑞东 (GUO, Ruidong); 中国广东省深圳市盐田区北山工业区综合楼, Guangdong 518083 (CN)。 唐静波 (TANG, Jingbo); 中国广东省深圳市盐田区北山工业区综合楼, Guangdong 518083 (CN)。 邓超 (DENG, Chao); 中国广东省深圳市盐田区北山工业区综合楼, 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熙源商业楼 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, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, JP, KE, KG, KN, KP, KR, 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)。

[见续页]

(54) Title: METHOD AND APPARATUS FOR ASSEMBLING SEPARATED LONG FRAGMENT SEQUENCES

(54) 发明名称: 对分隔长片段序列进行组装的方法和装置

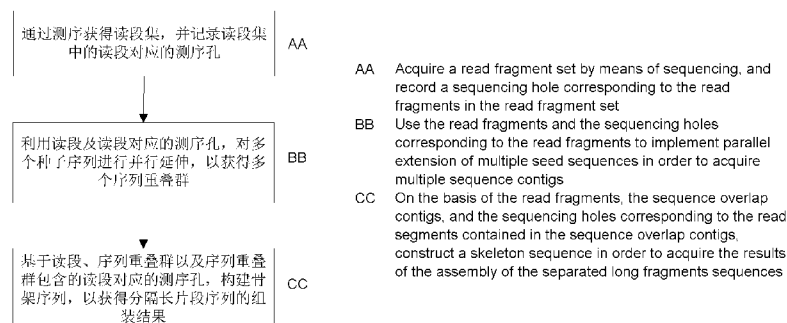


图 1

(57) Abstract: A method, apparatus, and system for assembling separated long fragment sequences. A method for assembling separated long fragment sequences, comprising: (a) acquiring a read fragment set by means of sequencing, and recording a sequencing hole corresponding to the read fragments in the read fragment set, the sequencing hole comprising at least one long fragment sequence; (b) using the read fragments and the sequencing holes corresponding to the read fragments to implement parallel extension of multiple seed sequences in order to acquire multiple sequence contigs, the multiple seed sequences being determined by means of known sequences; and (c) on the basis of the read fragments, the sequence overlap contigs, and the sequencing holes corresponding to the read segments contained in the sequence overlap contigs, constructing a skeleton sequence in order to acquire the results of the assembly of the separated long fragments sequences.

(57) 摘要: 一种对分隔长片段序列进行组装的方法、装置和系统。一种对分隔长片段序列进行组装的方法，包括：(a) 通过测序获得读段集，并记录所述读段集中的读段对应的测序孔，一个测序孔包含至少一条长片段序列；(b) 利用所述读段及所述读段对应的测序孔，对多个种子序列进行并行延伸，以获得多个序列重叠群，所述多个种子序列通过已知序列确定；(c) 基于所述读段、所述序列重叠群以及所述序列重叠群包含的读段对应的测序孔，构建骨架序列，以获得分隔长片段序列的组装结果。



WO 2017/143585 A1



本国际公布:

— 包括国际检索报告(条约第 21 条(3))。

对分隔长片段序列进行组装的方法和装置

技术领域

本发明涉及生物技术领域，具体地，本发明涉及对分隔长片段序列进行组装的方法和装置。

背景技术

从第一次人类基因组得到较完整的构建至今已有十数载，有了基因组图谱，各种针对人基因组重测序的生物信息学分析方法和软件如雨后春笋般涌现，对人类疾病、医学和健康研究的发展作出了重大贡献。

比较基因组学 (Comparative Genomics) 是一门对已知的基因和基因组结构进行比较以了解基因的功能、表达机理和物种进化的学科。其基本原则是两生物间相同的特性通常是由进化保守的 DNA 所编码的，那么两者间相关的 DNA 片段将是相同或相近的。比较基因组学方面的分析所需要的是基因组图谱 (基因组参考序列) 的存在和对比较对象的测序。变异检测则是比较基因组学中的重要内容，国际人类基因组单体型图计划 (The International HapMap Project) 和全基因组关联分析 (Genome-wide association study, GWAS) 正是基于对单核苷酸多态性 (single-nucleotide polymorphism, SNP) 这种类型的变异来进行相关研究的。

组装指的是将较短的碎片序列整合成较长序列的过程。受限于测序技术，基因组并不能通过对染色体从头到尾的测序读出其完整的序列内容，而往往将全基因组打碎成数十至数千长度的碱基片段，通过大规模并行测序来读出这些片段的内容，再使用组装来对这些零碎的片段进行分析整合，最后重新还原成相对完整的基因组序列。通过组装来识别变异是对组装技术的一种新的运用，而实际上组装主要的目的是构建基因组。当没有基因组参考序列时，基因组的构建组装则是一个从无到有的过程，显得尤为重要，这类组装也称为“从头组装” (denovo assembly)。人类基因组参考序列长度约为 3Gbp (3×10^9 bp)，这是单倍体的碱基对数量，而人类是二倍体生物，实际上应该有 $2 \times 3\text{Gbp} = 6\text{Gbp}$ 的基因组大小。一个人类个体的二倍体基因组来源于他的两个亲本各自贡献的一个单倍体，这两套单倍体间存在的各种差异使该个体在同源染色体上的一个或多个位点上有不同的等位基因存在，这种现象就是杂合现象。而且，人类参考序列是通过多个个体的数据构建而成，这导致它实际上是一个混杂嵌合了各种杂合的单倍体基因组。随着对基因组研究的深入，单倍体的人类参考序列越来越不能满足需求，构建单体型 (Haplotyping) 显得日益重要，基于单体型信息的基因组分析也不断出现。

单体型信息有助于深入解读基因型和表型之间的关系，杂合集完全相同的两个个体也会因单体型的不同产生不同的表型和疾病易感性，它对功能方面的研究(如等位基因的特异性表达)和疾病方面的研究(如地中海热，乳腺癌)有着重要意义。将多个杂合位点上的杂合信息分拨以确定单体型的作用称为定相 (Phasing)，是构建单体型的重要操作，而构建单体型的方法目前已有许多，主要分为以下五类：

- 1.对多个无亲缘关系个体的数据运用种群统计的方法
- 2.对家系数据运用孟德尔遗传的方法
- 3.直接利用测序序列信息的方法
- 4.通过实验操作的方法
- 5.通过物理分隔的方法

需要重点指出的，物理分隔的方法的核心是对断裂成 DNA 长片段的序列进行分隔，无论是通过 fosmid 质粒获取的方式，还是用多孔板直接物理分隔的方式。分隔后再进行测序所需的进一步打碎和扩增操作（构建文库），为了区分不同的分隔单元，分别对这些单元中的序列接上了不同的条码序列 (barcode)。这种方式通过分隔把全基因组分成了很多个子部份，当分离的子部份数量很多时，每个子部份则仅包含基因组上一小块区域的其中一个单倍体的内容。这使得全基因组水平上的杂合区域分别在这些小区域中以纯合的形态出现，这对单体型的构建有着极为重要的意义。

每个分隔单元有了自己独有的条码序列便可以在测序之后通过对条码序列的识别来找回属于各个分隔单元自己的读 (reads)。Fosmid 质粒分隔技术中称其分隔单元为 fosmid 质粒池 (fosmid pool)，每个 fosmid 质粒池中包含一条或多条长约 37Kbp 的长片段；而多孔板分隔技术中称其分隔单元为井 (well)，每个井中包含多条长片段，其长度在不同技术中各不相同。不管怎样，分隔的方法开创性地带来了新类型的信息，reads 总集合通过 barcode 分隔成了多群集合，这与全基因组鸟枪法 (Whole Genome Shotgun, WGS) 的测序手段相比，reads 不再随机地来源于全基因组上的任意位置，而是每群集合中的 reads 都来自于基因组上的一块共同的小区域内，这些小区域便是分隔时的长片段所覆盖的区域。虽然这些小区域依旧是来源于全基因组上的任意位置，但是每群集合中的 reads 却受到了约束，发生了聚集，这种聚集所新增的信息成为构建单体型的关键。

长片段读 (Long Fragment Reads, LFR) 技术大幅改善了文库构建的复杂程度，使得单体型的构建无论在时间还是成本上都得到了降低。

然而，目前的长片段读技术的组装仍有待改进。

发明内容

本发明旨在至少在一定程度上解决上述技术问题之一或至少提供一种有用的商业选择。

在本发明的第一方面，本发明提出了一种对分隔长片段序列进行组装的方法，包括：

(a) 通过测序获得读段集，并记录所述读段集中的读段对应的测序孔，一个测序孔包含至少一条长片段序列；

(b) 利用所述读段及所述读段对应的测序孔，对多个种子序列进行并行延伸，以获得多个序列重叠群，所述多个种子序列通过已知序列确定；

(c) 基于所述读段、所述序列重叠群以及所述序列重叠群包含的读段对应的测序孔，构建骨架序列，以获得分隔长片段序列的组装结果。

本领域技术人员可以理解，上述本发明一方面方法的全部或部分步骤可以通过程序来指令相关硬件完成，该程序可以存储于一计算机可读存储介质中，存储介质可以包括：只读存储器、随机存储器、磁盘或光盘等。

在本发明的第二方面，本发明提供一种对分隔长片段序列进行组装的系统，该系统包括：数据输入单元，用于输入数据；数据输出单元，用于输出数据；存储单元，用于存储数据，其中包括计算机可执行程序；处理器，与所述数据输入单元、所述数据输出单元和所述存储单元连接，用于执行所述计算机可执行程序，执行所述程序包括完成上述方法。

在本发明的第三方面，本发明提出了一种对分隔长片段序列进行组装的装置，包括：

输入模块，通过测序获得读段集，并记录所述读段集中的读段对应的测序孔，一个测序孔包含至少一条长片段序列；

延伸模块，利用所述读段及所述读段对应的测序孔，对多个种子序列进行并行延伸，以获得多个序列重叠群，所述多个种子序列通过已知序列确定；

骨架序列构建模块，基于所述读段、所述序列重叠群以及所述序列重叠群包含的读段对应的测序孔，构建骨架序列，以获得分隔长片段序列的组装结果。

上述本发明的方法和/或装置至少具有以下特性和优越性：

迭代延伸：种子序列的延伸是一个迭代过程，本次延伸的部分将作为下次迭代的延伸基座，使得种子序列能不断延长。

线性延伸：与其他基于图的组装算法不同，线性延伸方式使得组装时遇到的结构较为简单，逻辑关系相对清晰且容易分类处理，而图的算法会将所有重复区域的信息聚集于一处，其涉及的信息量更多且产生的结构更复杂。即图的解法是一次性解决，只要解掉一个重复区域，所有跟该处相关联的基因组区域（它们都具有相同或相近的重复序列）就会全

部同时解决，而线性方法一次仅解决当前区域的重复，对其他区域并不同时解决，当延伸至该重复在基因组的另一区域时则再解一次。解决了一个重复后会进行相关信息的记录，使得下一次解重复时遇到的信息更简单，这降低了额外的计算损耗。此外，线性延伸方式使各个种子序列的延伸相互独立，这种模式更易于计算并行化。

多路延伸：由于该算法的目标是构建多倍体基因组，对种子序列进行延伸时会利用测序孔（well）信息进行多个单倍体的多路联合延伸，每当遇到杂合区域时会结合前面已装成的杂合区状况即时进行定相的操作，即定相也是线性进行的，且贯穿整个组装装置的各个模块。为了节省资源消耗，纯合区的延伸依旧采用单路延伸，当遇到杂合区时再分为多路进行延伸，而杂合区延伸完毕后重新合并回单路继续延伸。

全局信息判定：一个算法若只利用当前判定下相关的信息，而不考虑其后的全局范围内的选择支，则容易出现错误或者无法达到全局最优解（如求单源最短路径中的 Dijkstra 贪婪算法）。对于基因组组装，贪婪算法往往因错误的选择而导致组装错误，所以本方法并不采用贪婪算法来解决延伸时的冲突，而用全局信息进行相关判定。由于 LFR 的长度（~100Kbp）较长，在路径选择时考虑了该位置所能提供的全部信息（Kmer、reads、mate-pair 及 well 信息），其中 well 信息可使得其判定涉及的范围在 100Kbp 左右，若大型重复没有大量存在，该算法是基本符合全局性的，而当延伸遇到复杂情况时，本算法会严格选择终止延伸而不是取较高值进行延伸，避免使用贪婪的判定模式。需要特别提出的是，由于延伸时考虑了该位置上所有的 well 信息，这使得该算法与常规的分级组装算法相比具有更强的全局性，可以大幅降低测序深度的要求，即每个 well 并不要求有特别高的测序深度，这节省了大量的资源，无论是成本还是时间。

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 是本发明的实施例中的对分隔长片段序列进行组装的方法的流程图。

图 2 是本发明的实施例中的对分隔长片段序列进行组装的方法的流程图。

图 3 是本发明的实施例中的对分隔长片段序列进行组装的装置的结构示意图。

图 4 是本发明的实施例中的对分隔长片段序列进行组装的装置的结构示意图。

图 5 是本发明的实施例中的种子序列迭代延伸过程的示意图。

图 6 是本发明的实施例中的识别重复区域的示意图。

图 7 是本发明的实施例中的处理具有相距较大的两个重复区域的大型双重复序列的冲突的示意图。

图 8 是本发明的实施例中的处理具有相距较小的两个重复区域的大型双重复序列的冲突的示意图。

图 9 是本发明的实施例中的处理串联重复序列的冲突的示意图。

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，其中自始至终相同或类似的标号表示相同或类似的元件或具有相同或类似功能的元件。下面通过参考附图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。

在本发明的第一方面，本发明提出了一种对分隔长片段序列进行组装的方法，参见图 1，该方法包括：

(a) 通过测序获得读段集，并记录所述读段集中的读段对应的测序孔，一个测序孔包含至少一条长片段序列；

(b) 利用所述读段及所述读段对应的测序孔，对多个种子序列进行并行延伸，以获得多个序列重叠群，所述多个种子序列通过已知序列确定；

(c) 基于所述读段、所述序列重叠群以及所述序列重叠群包含的读段对应的测序孔，构建骨架序列，以获得分隔长片段序列的组装结果。

根据本发明的实施例，所述种子序列是基于基因组参考序列按照下列步骤获得的：

在所述基因组参考序列中，按照 N 进行打断；以及

将经过打断的参考序列按照预定长度进行截断，以便获得所述种子序列。

对预定长度的大小不作特别限制，一般的，预定长度为不小于插入片段的大小，使来自同一插入片段的读段能够定位到同一种子序列上。根据本发明的实施例，测序为双末端测序，所述预定长度至少为所述测序中测序文库插入片段长度的两倍，利于成对读段定位至同一种子序列上。根据本发明的实施例，所述读段集包括成对读段，所述种子序列是基于所述读段按照下列步骤获得的：(1) 将所述读段滑动切割成多个 Kmer，构建 Kmer 对所述读段的索引 RKI，用于通过 Kmer 访问对应的读段；(2) 从所述读段集中提取一对不具有高频 Kmer 的成对读段；(3) 利用所述索引 RKI，分别确定 (2) 中的一对成对读段中的两个读段的 Kmer 对应的所有读段，获得第一读段群和第二读段群；(4) 分别确定 (3) 的第一读段群和第二读段群对应的测序孔，获得第一测序孔集和第二测序孔集；(5) 确定 (4)

中的第一测序孔集和第二测序孔集的交集，若所述交集的大小与碱基的有效测序孔的数量期望值无显著差异，则确定（2）中的成对读段为所述种子序列，所述碱基的有效测序孔的数量期望值由样本的核酸量决定。所称的高频是相对于平均频率来说的，某个 Kmer 的出现次数高于 Kmer 出现的平均数，即可认为其为高频 Kmer。在更多情况中，发明人根据需将出现次数远高于 Kmer 出现的平均数的 Kmer 定义为高频 Kmer，所称的“远高于”为出现次数为 Kmer 出现的平均次数的至少 10 倍、20 倍、30 倍、40 倍、50 倍或者 60 倍。Kmer 频率期望与测序深度、读段长度及 Kmer 大小相关，根据本发明的一个实施例，对于 $K=19$ ，即 19mer，100bp 读段长度，100X 测序深度，Kmer 期望频率为 $100(\text{测序深度}) * \{[100(\text{读长}) - 19(\text{Kmer 大小}) + 1] / 100(\text{读长})\} = 82$ ，在平均频率为 82 时，设置频率为 3000 至 5000 范围的 Kmer 为高频 Kmer。

所称的无显著差异，可以是无统计意义上的差异，也可以是一般所说的无明显差异。根据本发明的实施例，（5）中：若所述交集的大小在碱基的有效测序孔的数量期望值的一半到两倍之间，则确定（2）中的成对读段为所述种子序列。

上述种子序列的生成方式，前者属于只依据已知序列来构建，例如依据参考序列、测序获得的读段、组装片段等，后者属于结合已知序列、RKI 以及测序孔的信息来构建，后者可应用于完全没有参考序列的物种的从头组装。

根据本发明的实施例，（b）包括：（i）将所述读段滑动切割成多个 Kmer，构建 Kmer 对所述读段的索引 RKI，用于通过 Kmer 访问对应的读段；（ii）基于所述读段及其对应的索引 RKI，对所述多个种子序列进行并行延伸，以获得所述多个序列重叠群。

根据本发明的实施例，所述 RKI 是通过下列步骤获得的：对所述读段 reads 进行滑动切割成多个 Kmer；构建以 Kmer 为键值的哈希，所述哈希构成所述 RKI，并且所述哈希记载所述 Kmer 的频率、所属读段、所述 Kmer 在所述读段上的位置和方向。

根据本发明的实施例，通过重复下列步骤对所述种子序列进行延伸：选择适于延伸的种子序列；将所述读段定位至所述种子序列，以获得延伸序列；将定位在所述延伸序列末端的读段进行逐碱基一致性化处理；以及如果一致性化处理失败，则进行杂合识别、定相处理和/或对重复序列进行解析。

根据本发明的实施例，通过下列步骤选择适于延伸的种子序列：将种子序列滑动切割成 Kmer；通过所述 RKI 获取所述 Kmer 对应的读段；将所述对应的读段与所述种子序列进行比对；基于所述对应的读段对应的测序孔，确定测序孔对所述种子序列的覆盖状况；以及基于所述覆盖状况，确定适于延伸的种子序列。

根据本发明的实施例，通过下列步骤将所述读段定位至所述种子序列：将种子序列滑动切割成 Kmer；通过所述 RKI 获取所述 Kmer 对应的读段；将所述 Kmer 对应的读段定位至所述种子序列，并逐个碱基进行比对。

根据本发明的实施例，在一致性化处理过程中，如果延伸的位点的有效测序孔集合被该位点不同的碱基型平均分配，则判断存在杂合。

根据本发明的实施例，在确定存在杂合之后，将所述延伸序列分为多条分别进行延伸。

根据本发明的实施例，所述读段集包含多对成对读段，一对成对读段中的两个读段之间的距离为 L ，在一致性化处理过程中，如果成对读段中的定位在延伸方向下游的读段与对应的读段之间的距离为非 L ，则确定该定位在延伸方向下游的读段所定位的位置为重复序列的起点。

需要说明的是，本发明中涉及的数值大多具有统计意义，因此，如无特殊说明，任意以精确方式表达的数值均可代表一个范围，例如包含该数值正负 10% 的区间；或者该数值总体呈正态分布，以精确方式表达的数值包含该数值正负标准差的区间。所称一对成对读段中的两个读段之间的距离 L ，对应插入片段的长度。一般的，在试验阶段，构建的测序文库大小是一定的，即插入片段的大小是一个定值，理论上，双末端测序后获得的成对读段之间的外端距离为该定值，而在实际测序后获得的数据中，成对读段之间的距离呈正态分布。在该示例的重复序列判定与解析的过程中，发明人设置该 L 为试验阶段的插入片段的大小，本领域技术人员可以理解，设置该 L 为任一试验阶段插入片段大小的正负标准差之间的数值或者为试验阶段插入片段大小的正负标准差区间，利用本发明的解重复序列的方法，同样可以判定和解析延伸过程中的重复序列。

根据本发明的实施例，所述重复序列的终点为成对读段中的与对应的读段之间的距离为 L 的定位在延伸方向下游的读段所定位的位置，或者为延伸时的冲突位点。

根据本发明的实施例，通过下列步骤判断所述重复序列是否为串联重复序列：对定位至所述重复序列的读段进行滑动切割以获得 Kmer；判断每条所述读段进行滑动切割所得到的 Kmer 是否存在重复 Kmer，如果不存在所述重复 Kmer，则判断不存在串联重复序列，如果存在所述重复 Kmer，则判断存在串联重复序列。

根据本发明的实施例，若所述重复序列为串联重复序列，通过下列步骤对所述串联重复序列进行解析：确定所述串联重复序列的长度；将包含所述串联重复序列终点的读段用终点对齐的方式进行位置调整，以确定冲突位置上的碱基。

根据本发明的实施例，若所述重复序列非为串联重复序列，且存在以下情况，则通过

以下判断所述重复序列为大型双重复序列：所述重复序列的长度大于 L ，或者成对读段中定位在所述重复序列下游的读段对应的读段也定位在该重复序列上。

根据本发明的实施例，若所述重复序列为大型双重复序列，通过以下对所述大型双重复序列进行解析：比较所述大型双重复序列中的上游重复序列对应的有效测序孔的数量与下游重复序列对应的有效测序孔的数量的差值与碱基的有效测序孔的数量期望值的差异，依据差异的大小，解决所述大型双重复序列上的冲突。例如，若大型双重复序列中的两重复序列距离相对较远，处在上游的冲突碱基所含有的有效测序孔（EW）会明显比下游多，定义数量差值大于期望 EW 数量的一半时为明显，可以通过这种比较方式确定碱基来解决冲突；若上游与下游的 EW 数量差值小于等于期望 EW 数量的一半，可以利用离重复区起点一个插入片段长度范围内的错误下游臂所对应的上游臂来构建辅助重叠群（HC），比较 HC 上的 EW 集合与上游冲突碱基的 EW 集合比它与下游冲突碱基的差异。

根据本发明的实施例，若所述重复序列非为大型双重复序列，且存在以下情况，则判断所述重复序列为小型双重复序列：所述重复序列的长度小于 L 。

根据本发明的实施例，通过以下至少之一对所述小型双重复序列进行解析：（p）利用支持各碱基的成对读段之间的距离的均值作为冲突位点的期望位置，通过比较支持两种冲突碱基的成对读段之间的距离与所述均值的接近程度，以确定该冲突位点上的碱基；（k）利用成对读段中定位至所述延伸序列标准差范围中的非重复序列的最上游的读段对应的读段来构建辅助重叠群，利用成对读段中定位至所述辅助重叠群下游的读段对应的读段来确定冲突位点上的碱基。

根据本发明的实施例，若无法解析所述小型双重复序列，即无法解决该冲突，则终止对种子序列的延伸。

参见图 2，根据本发明的实施例，（c）包括：（iii）基于所述读段，建立所述序列重叠群之间的合并连接关系；（iv）基于所述序列重叠群和所述序列重叠群包含的读段对应的测序孔，构建初级骨架序列；（v）在对所述多个序列重叠群之间的合并连接关系与所述初级骨架序列进行相互检验之后，通过合并所述序列重叠群以构建所述骨架序列，获得分隔长片段序列的组装结果。

上述本发明任一实施例的方法具有以下四个特性和优越性中的至少之一：

迭代延伸：种子序列的延伸是一个迭代过程，本次延伸的部分将作为下次迭代的延伸基座，使得种子序列能不断延长。

线性延伸：与其他基于图的组装算法不同，线性延伸方式使得组装时遇到的结构较为

简单，逻辑关系相对清晰且容易分类处理，而图的算法会将所有重复区域的信息聚集于一处，其涉及的信息量更多且产生的结构更复杂。即图的解法是一次性解决，只要解掉一个重复区域，所有跟该处相关联的基因组区域（它们都具有相同或相近的重复序列）就会全部同时解决，而线性方法一次仅解决当前区域的重复，对其他区域并不同时解决，当延伸至该重复在基因组的另一区域时则再解一次。解决了一个重复后会进行相关信息的记录，使得下一次解重复时遇到的信息更简单，这降低了额外的计算损耗。此外，线性延伸方式使各个种子序列的延伸相互独立，这种模式更易于计算并行化。

多路延伸：由于该算法的目标是构建多倍体基因组，对种子序列进行延伸时会利用 well 信息进行多个单倍体的多路联合延伸，每当遇到杂合区域时会结合前面已装成的杂合区状况即时进行定相的操作，即定相也是线性进行的，且贯穿整个组装系统的各个模块（如重叠群合并，骨架序列构建）。为了节省资源消耗，纯合区的延伸依旧采用单路延伸，当遇到杂合区时再分为多路进行延伸，而杂合区延伸完毕后重新合并回单路继续延伸。

全局信息判定：一个算法若只利用当前判定下相关的信息，而不考虑其后的全局范围内的选择支，则容易出现错误或者无法达到全局最优解（如求单源最短路径中的 Dijkstra 贪婪算法）。对于基因组组装，贪婪算法往往因错误的选择而导致组装错误，所以本算法并不采用贪婪算法来解决延伸时的冲突，而用全局信息进行相关判定。由于 LFR 的长度（~100Kbp）较长，在路径选择时考虑了该位置所能提供的全部信息（Kmer、read、mate-pair 及 well 信息），其中 well 信息可使得其判定涉及的范围在 100Kbp 左右，若大型重复不大量存在，该算法是基本符合全局性的，而当延伸遇到复杂情况时，本算法会严格选择终止延伸而不是取较高值进行延伸，避免使用贪婪的判定模式。需要特别提出的是，由于延伸时考虑了该位置上所有的 well 信息，这使得该算法与常规的分级组装算法相比具有更强的全局性，可以大幅降低测序深度的要求，即每个 well 并不要求有特别高的测序深度，这节省了大量的资源，无论是成本还是时间。

本领域技术人员可以理解，上述本发明一方面方法的全部或部分步骤可以通过程序来指令相关硬件完成，该程序可以存储于一计算机可读存储介质中，存储介质可以包括：只读存储器、随机存储器、磁盘或光盘等。

将上述过程编写成可执行程序，执行程序包括：首先，读入 seeds 及 reads，并把 reads 构建成 Kmer 对 reads 的索引（Read Kmer Index，RKI），以便比对时迅速通过 Kmer 来访问目标 reads。然后把 seeds 尽可能地延长，由于 seed 与 seed 间的延伸相互独立，这个步骤可以通过并行运算进行加速。延伸后的 seed 就是序列重叠群（contig），此时通过 well 的

信息开始预构建骨架序列(scaffold),仅建立 contig 间的前后关系,并不立即构建出 scaffold。同样,通过已使用 reads 和成对 reads 信息来建立 contig 间的合并连接关系,也不立即合并这些 contig。接着将 contig 间合并的关系与预构建 scaffold 中的关系互相检验,简化这些关系并解决冲突,再进行 contig 的合并及 scaffold 的构建,最后输出组装结果。

在又一方面,本发明提出了一种对分隔长片段序列进行组装的装置,该装置用以实施上述本发明任一实施例的方法。参见图 3,该装置包括:输入模块,通过测序获得读段集,并记录所述读段集中的读段对应的测序孔,一个测序孔包含至少一条长片段序列;延伸模块,利用所述读段及所述读段对应的测序孔,对多个种子序列进行并行延伸,以获得多个序列重叠群,所述多个种子序列通过已知序列确定;骨架序列构建模块,基于所述读段、所述序列重叠群以及所述序列重叠群包含的读段对应的测序孔,构建骨架序列,以获得分隔长片段序列的组装结果。

本领域技术人员能够理解,上述本发明任一实施例中的处理步骤或采用的手段,可以利用该装置的相应功能模块或者包含的子模块来实现。上述对本发明的方法的技术特征和优点的描述,同样适用本发明的这一装置。

根据本发明的实施例,所述种子序列是基于基因组参考序列按照下列步骤获得的:在所述基因组参考序列中,按照 N 进行打断;以及将经过打断的参考序列按照预定长度进行截断,以便获得所述种子序列。

根据本发明的实施例,所述预定长度不小于所述测序中测序文库插入片段的长度。

根据本发明的实施例,所述读段集包括成对读段,所述种子序列是基于所述读段按照下列步骤获得的:(1)将所述读段滑动切割成多个 Kmer,构建 Kmer 对所述读段的索引 RKI,用于通过 Kmer 访问对应的读段;(2)从所述读段集中提取一对不具有高频 Kmer 的成对读段;(3)利用所述索引 RKI,分别确定(2)中的一对成对读段中的两个读段的 Kmer 对应的所有读段,获得第一读段群和第二读段群;(4)分别确定(3)的第一读段群和第二读段群对应的测序孔,获得第一测序孔集和第二测序孔集;(5)确定(4)中的第一测序孔集和第二测序孔集的交集,若所述交集的大小与碱基的有效测序孔的数量期望值无显著差异,则确定(2)中的成对读段为所述种子序列。所述碱基的有效测序孔的数量期望值由样本的核酸量决定。

根据本发明的实施例,上述实施例的(5)中:若所述交集的大小在碱基的有效测序孔的数量期望值的一半到两倍之间,则确定(2)中的成对读段为所述种子序列。

根据本发明的实施例,所述还包括索引构建模块,与所述延伸模块连接,用于将所述

读段滑动切割成多个 Kmer，构建 Kmer 对所述读段的索引 RKI，用于通过 Kmer 访问对应的读段；接着利用所述延伸模块实现以下：基于所述读段及其对应的索引 RKI，对所述多个种子序列进行并行延伸，以获得所述多个序列重叠群。

根据本发明的实施例，所述 RKI 是通过下列步骤获得的：对所述读段进行滑动切割成多个 Kmer；构建以 Kmer 为键值的哈希，所述哈希构成所述 RKI，并且所述哈希记载所述 Kmer 的频率、所属读段、所述 Kmer 在所述读段上的位置和方向。

根据本发明的实施例，通过重复下列步骤对所述种子序列进行延伸：选择适于延伸的种子序列；将所述读段定位至所述种子序列，以获得延伸序列；将定位在所述延伸序列末端的读段进行逐碱基一致性化处理；以及如果一致性化处理失败，则进行杂合识别、定相处理和/或对重复序列进行解析。

根据本发明的实施例，通过下列步骤选择适于延伸的种子序列：将种子序列滑动切割成 Kmer；通过所述 RKI 获取所述 Kmer 对应的读段；将所述对应的读段与所述种子序列进行比对；基于所述对应的读段对应的测序孔，确定测序孔对所述种子序列的覆盖状况；以及基于所述覆盖状况，确定适于延伸的种子序列。

根据本发明的实施例，通过下列步骤将所述读段定位至所述种子序列：将种子序列滑动切割成 Kmer；通过所述 RKI 获取所述 Kmer 对应的读段；将所述 Kmer 对应的读段定位至所述种子序列，并逐个碱基进行比对。

根据本发明的实施例，在一致性化处理过程中，如果延伸的位点的有效测序孔集合被该位点不同的碱基型平均分配，则判断存在杂合。

根据本发明的实施例，在确定存在杂合之后，将所述延伸序列分为多条分别进行延伸。

根据本发明的实施例，所述读段集包含多对成对读段，一对成对读段中的两个读段之间的距离为 L ，在一致性化处理过程中，如果成对读段中的定位在延伸方向下游的读段与对应的读段之间的距离为非 L ，则确定该定位在延伸方向下游的读段所定位的位置为重复序列的起点。

根据本发明的实施例，所述重复序列的终点为成对读段中的与对应的读段之间的距离为 L 的定位在延伸方向下游的读段所定位的位置，或者为延伸时的冲突位点。

根据本发明的实施例，通过下列步骤判断所述重复序列是否为串联重复序列：对定位至所述重复序列的读段进行滑动切割以获得 Kmer；判断每条所述读段进行滑动切割所得到的 Kmer 是否存在重复 Kmer，如果不存在所述重复 Kmer，则判断不存在串联重复序列，如果存在所述重复 Kmer，则判断存在串联重复序列。

根据本发明的实施例，若所述重复序列为串联重复序列，通过下列步骤对所述串联重复序列进行解析：确定所述串联重复序列的长度；将包含所述串联重复序列终点的读段用终点对齐的方式进行位置调整。

根据本发明的实施例，若所述重复序列非为串联重复序列，且存在以下情况，则判断所述重复序列为大型双重复序列：所述重复序列的长度大于 L ，或者成对读段中的定位在所述重复序列下游的读段对应的读段也定位在该重复序列上。

根据本发明的实施例，通过以下对所述大型双重复序列进行解析：比较所述大型双重复序列中的上游重复序列对应的有效测序孔的数量与下游重复序列对应的有效测序孔的数量的差值与碱基的有效测序孔的数量期望值的差异，依据差异的大小，解决所述大型双重复序列上的冲突。

根据本发明的实施例，若所述重复序列非为大型双重复序列，且存在以下情况，则判断所述重复序列为小型双重复序列：所述重复序列的长度小于 L 。

根据本发明的实施例，若所述重复序列为小型双重复序列，通过以下至少之一对所述小型双重复序列进行解析：(p) 利用支持各碱基的成对读段之间的距离的均值作为冲突位点的期望位置，通过比较支持两种冲突碱基的成对读段之间的距离与所述均值的接近程度，以确定该冲突位点上的碱基；(k) 利用成对读段中定位至所述延伸序列标准差范围中的非重复序列的最上游的读段对应的读段来构建辅助重叠群，利用成对读段中定位至所述辅助重叠群下游的读段对应的读段，通过增加这些读段数据以更新距离均值，再同样通过比较支持两种冲突碱基的成对读段之间的距离与所述均值的接近程度，以确定冲突位点上的碱基。

根据本发明的实施例，若无法解析所述小型双重复序列，则终止对种子序列的延伸。

参见图 4，根据本发明的实施例，所述骨架序列构建模块包括初级骨架序列构建模块，用于基于所述读段，建立所述序列重叠群之间的合并连接关系；还包括合并连接关系建立模块，用于基于所述序列重叠群和所述序列重叠群包含的读段对应的测序孔，构建初级骨架序列；还包括组装模块，用于在对所述多个序列重叠群之间的合并连接关系与所述初级骨架序列进行相互检验之后，通过合并所述序列重叠群以构建所述骨架序列，获得分隔长片段序列的组装结果。

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一定

指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何一个或多个实施例或示例中以合适的方式结合。

为了方便理解，下面对根据本发明优选实施例的组装方法进行逐步详细分析。

程序首先读入 seed 数据及 reads 数据用于后续的步骤，本算法为了达到高效计算的目的，将延伸过程中不断需要参与比对的 reads 数据以二进制形式完全存入内存中直到整个程序终止。而测序中提供的质量值信息（如 Fastq 格式中的质量值）并不在组装过程中记录和使用，这些信息被设计在组装前的数据预处理步骤进行去除或纠正质量值异常碱基及 reads 的工作，即读入的序列被认为在测序质量值上都是正常的，组装时无视不同质量值碱基间的差异。

读入 reads 并记录其所属 well 的信息后，滑动切割这些 reads 成 Kmer 并构建 Reads Kmer Index (RKI)，这是以 Kmer 为键 (key) 的哈希 (hash)，记录了 Kmer 的频率、所属的 reads 集合及该 Kmer 在这些 reads 上的位置、方向等信息。哈希的特性是查找速度极快，时间复杂度仅为 $O(1)$ ，所以该索引的核心作用是迅速地通过 Kmer 来访问其对应的 reads 集合，以确定哪些 reads 要被定位到 seed 上来进行细化的比对。

这种 RKI 结构由于需要记录大量 reads 的信息，而进一步地，由于一条 reads 包含多个 Kmer，这条 reads 将会在多个 Kmer 的记录项中出现，这使得内存的消耗更为巨大，对于人基因组来说，100X 测序量下内存的耗用约为 3TB，比起传统的组装技术消耗 1TB 以下的内存，该算法所需要的内存较多。数据结构仍可以进一步优化，这可以降低组装的成本及资源瓶颈。可以进行优化的方式如将超高频的 Kmer 不予或限制存储其相关 reads 的信息以减少内存开销，因为这些 Kmer 对组装的作用很小，将其用于 reads 比对定位也会浪费较多的时间，效率极低。

以上关于 RKI 的数据结构设计与一些 DBG 算法中的设计存在明显的不同，它们通过从文件中读取 reads 来构建出 Kmer 间关系的图，并不需要把 reads 存储于内存中，且 Kmer 对应的 reads 关系也无需详细记录，因为基因组的表示已从一堆杂乱的 reads 转化为一群有逻辑关联的 Kmer，数据进行了大量的整理与压缩，之后基因组的构建仅需对 Kmer 图进行操作即可，无需直接操作 reads。而本算法并不构建一幅整体的图，在 seed 延伸（基因组的构建）过程中需要不断地通过 Kmer 来对 reads 进行随机访问以获取延伸（组装）素材。

最后，关于实验方面的相关参数也需要读入，如文库构建中成对 reads 的插入片段大小，分离 LFR 的 well 数量，以及投入的细胞数，目标基因组的倍体数（通过光学手段或信息学分析获得）。但其中插入片段大小及投入细胞数是可以由种子序列初始化时训练统计得

到的，而目标基因组的倍体数也是可以通过杂合识别计算得到。以上获取信息的不同方式都会有一定误差，实际组装时应结合分析及运用。

seed 在被选取后，在开始延伸之前会先进行初始化及合法性的判定。初始化获得 seed 上的相关信息，这些信息用于之后的合法性判定及 seed 延伸素材。而合法性判定用于丢弃那些在已延伸区域上的 seed 及不适于延伸的 seed。

首先通过滑动切割 seed 成 Kmer，通过 RKI 获取这些 Kmer 所对应的 reads 后再将这些 reads 精细比对上 seed，并利用这些 reads 对应的 well 信息来确定 well 对这个 seed 的覆盖状况。

基本完全覆盖整个 seed 的 well 被定义为“有效的 well (Effective Well, EW)”，其他覆盖尚不足的 well 将作为候选被记录下来并在之后的延伸中进行更新。在分隔长片段序列组装方式中，基因组上每个位点都会被多条 LFR 所覆盖，其覆盖的数量与细胞数及倍体数相关，随着延伸的前进，覆盖延伸区域的 LFR 集合将发生过渡性变化，新的 LFR 不断替换掉旧的 LFR，就如接力赛一样。例如，如果用 10 个二倍体细胞建 LFR，每个位点上的 well 数量期望则为 $10 \text{ 个细胞} \times \text{二倍体} \times \text{DNA 双链} = 40$ 个，若 LFR 长度为 100Kbp，那么相邻 LFR 的平均间隔为 $100\text{Kbp}/40 = 2.5\text{Kbp}$ ，即在 2.5Kbp 范围内 well 的数量期望值为 40，若参与计算的范围扩大，则每增加 2.5Kbp，well 的数量就会增加一个。

在 seed 延伸过程中，EW 的集合是通过其对已延伸（或初始化）区块的 well 覆盖度来确认的。当一条 LFR 刚开始覆盖到正在延伸的 seed 时，其对该 seed 的覆盖度会随着 seed 的延伸不断上升直至完全覆盖，这种基本完全覆盖 seed 的 LFR 所对应的 well 就可以成为辅助延伸用的 EW；而随着 seed 继续往前延伸，这条 LFR 将逐渐退出 seed 所延伸的区域，其覆盖度会不断下降直至为 0，当 LFR 对 seed 覆盖明显不足时，其所对应的 well 将不再作为 EW 用以辅助延伸。这就产生了延伸时 EW 集合的过渡性变化。

这个随着延伸不断更新的 EW 集合将指导组装时 reads 的过滤，可以从整个基因组众多的 reads 中分离出只属于当前 seed 附近区域的 reads，使得一些在全基因组范围上复杂的区域退化为 LFR 长度范围上的简单区域而变得易于组装，这就是分隔长片段序列组装方式的核心所在。

这种在延伸时同时考虑多个 well 的集合的组装方式与传统的分开考虑单个 well 的二级组装方式（即对每个 well 单独进行组装后再合并这些组装结果来进行二次组装）的明显差异之处正是对 EW 的运用。这种方法可使组装时需要的深度信息从单个 well 扩展到联合整个 EW 集合的累积深度，降低了单个 well 所需的测序深度，仅需较低的覆盖度即可。以 10

个二倍体细胞（40 份拷贝）为例，以往的二级组装方式需要 $40 \times 50X = 2,000X$ 的测序深度，其中 50X 为常规基因组组装所需的测序深度；而本算法仅需 $40 \times 2.5X = 100X$ 的测序深度，其中 2.5X 为 EW 在容忍部分区域没有被测序到的情况下被识别出来所需的测序深度，而延伸时联合整个 EW 集合的累积深度 100X，即 2.5X 用于识别 EW 而累积 100X 用于延伸。这种方式极大地降低了测序成本，并节约了时间。

此外，EW 的过渡性变化也可以用于确定 contig 间的位置关系，构建出极长的骨架序列。

根据本发明的实施例，seed 初始化及延伸时都需要将 reads 比对到 seed 上以提供素材来获取信息（如 well 和 reads 的成对信息）及延伸序列，而把所有 reads 都尝试去比对是不切实际的，这会消耗大量的时间，所以利用 RKI 来初筛 reads，命中了 seed 上的 Kmer 的 reads 才能进行精细的比对，方法如下：

1. 滑动切割 seed 序列以获取 Kmer，并通过 RKI 来获得这些 Kmer 所对应的 reads。seed 初始化时对全 seed 序列进行切割，而在延伸过程中仅对 contig 末端的序列进行切割。基于效率方面的考虑，极高频的 Kmer（一个 Kmer 对应极多的 reads）不会用来获取 reads。特殊地，对于延伸过程中的 reads 获取，若发现末端全都是极高频 Kmer 时，仅取最末端的一个 Kmer 用于 reads 定位。

2. 通过 Kmer 在 seed 和 reads 上的位置记录把 reads 定位到 seed 上，并逐碱基进行比对，对测序中的替换型错误有一定的容忍，但暂无开 gap 的模块，无法容忍 indel 类型的错误。

3. reads 过滤的目的是为了去除由于重复导致错误定位的 reads。本算法首先通过 EW 信息滤掉因全基因组级别的重复而错误定位的 reads，再通过 mate-pair 信息滤掉因 LFR 内部的简单重复而错误定位的右臂 reads（左臂 reads 不进行过滤）。需要注意的是，在重复区内的左臂 reads 不会参与其对应右臂 reads 的合法性判断，因为其本身也可能定位错误，而复杂类型的重复将交由重复解决模块处理。

4. 为了获得延伸的序列，需要将定位在 contig 末端的 reads 进行逐碱基的一致性化。与 reads 比对类似，一致性化过程中仅容忍替换型测序错误，暂无 indel 类型的容忍。由于 reads 的 mate-pair 过滤只能作用于右臂 reads，这使得右臂 reads 不会产生定位错误，与左臂 reads 有着明显优势。基于这个特征，当过滤后的右臂 reads 数量充足时，本模块仅使用这些 reads 进行一致性化，此时若能合为一条序列，本次延伸将完成，接下来进入相关信息更新的模块，若无法合为一条序列，即一致性化时发生冲突，则需联合该区域的左臂 reads

进行杂合或重复的识别与解决。而当右臂 reads 数量不足时，也需联合左臂 reads 进行一致性化。

5. 一致性化的逐碱基位点合并过程中，若发现有超过一种碱基在同一位置非低频地出现（低频主要是测序错误造成），则会造成冲突，判定该位点的一致性化失败，这主要是由杂合和重复造成。与重复相比，二倍体基因组的杂合识别较为简单，其特性是冲突的碱基有且仅有两种，对半平分 EW 集合，且各自支持的 EW 集基本没有交集，所以一致性化失败后会先进行杂合的识别。

与杂合相比，测序错误就不符合这种状况，测序错误的主要特征是低频与随机。因为绝大多数的测序技术中，测序错误只占总测序量的极少数，一致性化时只会发现极少量的差异，即使是在测序深度有较大波动的状况下，测序错误造成的差异依旧只占极少的成分，测序错误的绝对数量虽然会随着测序深度波动一同变化，但比例并不改变，也就是测序错误率不变。另一方面，其随机性的表现则在于测序错误在不同 well 中并没有偏差，各个 well 中都会有相同概率的测序错误出现，即使测序错误本身在测错模式上可能有一定的偏向性，但各个 well 中的状况也都是一样的。这意味着测序错误的碱基在 well 上面并没有集中性，一致性化冲突时不同碱基各自支持 EW 集合间存在交集，无法被区分开。需要留意的是，当整体测序深度低下时测序错误的特征将变得不明显，这种时候极易出现误判，所以，测序深度不能过低。

当确认为杂合后，contig 将被分为两条来进行双路延伸，分离时会结合前一个杂合区域的 EW 状况进行定相。原理上来自于同一个单倍体上的两个杂合区域应具有相近的 EW 集合，相反，来自不同单倍体上的两个杂合区域的 EW 集合间应没有交集。通过这种方式就可以进行临近杂合位点的定相操作，特殊地，若前一个杂合区域距离过远（超过 LFR 长度），此时即使是同一单倍体上的两杂合区域的 EW 也不会有交集，定相将被判定为失败，并建立一个新的定相区块，而 contig 本身依旧是连续的，并不同时打断。

种子序列迭代延伸过程参见图 5。以上方法与另外一些定相的算法不同，仅考虑了两临近位点的 EW 关系，并没有联合多个位点一同考虑，且定相的方式也是线性的，逻辑较为简洁，定相的算法仍存在改进的空间。当然，本算法中的定相方法也有着一定的优势，与一些只考虑单位点杂合的方法不同，本算法对所有杂合类型都会进行定相，这使得定相的长度得到提升。

和其他基于组装来识别结构变异的方法一样，本算法也能识别出大型的结构变异型杂合，这是本算法比起其他处理分隔长片段序列来分离单体型方法的优势所在。更进一步，

由于插入-删除型杂合和大型结构变异型杂合在本算法中的定相方式上并没有差异，这使得本算法比其他仅考虑单一替换型杂合位点的定相方法能构建出更完整更长的单体型，还能获得那些重测序方法所不具有的功能，尤其是能对大型结构变异进行定相，这对下游的分析有着重大意义。

重复序列（Repetitive Sequence）是基因组中不同位置出现的相同或相似的序列。它在基因组中大量出现，如人类基因组各种类型的重复序列的总长占将近半个基因组的大小。重复序列一直以来都是影响组装质量的重要问题，能否正确地解决重复也是各种组装算法最为关心并不断尝试新策略以求突破的问题。重复的解决自然也是本算法中的关键模块，主要处理 LFR 内的复杂重复，主要包括邻近的小型双重复，大型双重复，以及串联重复。以下对这些重复的解决方法进行描述。

不同的重复类型将交由不同的模块进行解决，而重复类型的识别需要相关信息的辅助，重复区域识别就是识别重复类型的重要信息之一，主要目的是用于识别不能作过滤用的左臂 reads，即 mate-pair reads 都在重复区内，这是因为重复区的特性之一是 reads 会被错误地定位到此处，若使用了错误定位的 reads 用于延伸，往往会导致错误的延伸。

如图 6 所示，重复区域识别主要包括起点和终点的识别。当延伸时发现延伸区中的右臂 reads 所对应的左臂不在离上游一个插入片段长度的位置附近时，可判定出重复区域的起点。

重复区的终点主要分为两类，对于简单重复（仅通过 mate-pair 关系来过滤错误定位的 reads 就可解决），重复区的终点是丢失左臂的右臂 reads 不再出现时的位置；而对于复杂重复，延伸时的冲突点即为终点。

此外，重复片段中出现的差异区域将被认为是非重复区，即相似的重复片段会被严格分割为多个子重复片段来处理。而这种情况下差异点后的区域性质识别则有所不同，重复区起点的右臂 reads 所对应的左臂在上一个重复区中而并不是在非重复区内，发生了成对 reads 定位错误的状况。此时检查延伸点前一个插入片段长度位置附近的左臂 reads，若它们对应的右臂都参与了延伸，则说明当前延伸的区域是重复区，若发现有一部分并未参与延伸，则说明已经延伸入非重复区。

由于 reads 的 mate-pair 过滤只能作用于右臂 reads，延伸时右臂 reads 和左臂 reads 的地位并不相同，然而反过来看，若存在一区域能通过右臂 reads 来检验左臂 reads，那左臂 reads 就可以发挥解决重复的作用。这正是辅助重叠群（Helper Contig，HC）的概念。

HC 是一条解决复杂重复时使用的 contig，仅作辅助使用，并不作为正式的 contig 出现

在组装结果中。其本质是进一步利用 reads 的 mate-pair 信息，并期望 HC 能跨过当前的重复区，出现在下游的非重复区上，利用这个非重复区来帮助解决重复。若 HC 未能成功跨过重复区，则一般无法发挥效果。目前对 HC 的应用对象主要分为以下两类：

(1) HC case1-邻近的小型双重复：

由于 mate-pair reads 的插入片段长度具有一定标准差 (standard deviation, SD)，若存在两个邻近的比 reads 长度-Kmer+1 略长的小型重复，则延伸时会因这个 SD 偏差造成 mate-pair 信息的混乱，一致性化时将发生冲突，冲突的两种碱基都有 EW 和 mate-pair reads 的支持。

对于这种问题，本算法首先利用支持各碱基的 mate-pair 信息计算出的均值作为期望位置，距离近的将被认为是应该延伸的碱基，若这两个位置距离过近，实际状况较模糊，则需要构建 HC 辅助进行识别。

首先将 contig 末端 SD 长度范围内不在重复区中的最左端的左臂 reads 所对应的右臂作为构建 HC 的起始，对这条 reads 采用类似于 seed 的方式进行延伸。延伸到足够的长度时就可以通过定位在这个 HC 上的右臂 reads 所对应的左臂来分别计算冲突碱基的位置。实质上，这种情况下使用 HC 仅提升了计算距离的可靠性，若 SD 较大，则仍可能产生错误。

(2) HC case2-大型双重复：

当重复区域长度大于插入片段长度时，左臂 reads 将无法用于 reads 的过滤，实质上，左臂只要在重复区内就无法用于过滤重复区内的右臂，无论两臂之间的区域是重复抑或是非重复。

大型的重复由于长度较长，well 信息较容易发生差异，所以依照不同状况可以利用 well 信息来进行解决，部分情况则需要结合 HC 来解决，主要分为两类：

两重复序列距离较远，这种情况下虽然两冲突碱基都有 EW 支持，但处在上游的碱基所含有的 EW 会明显比下游的多。此时可以通过这种比较来解决冲突。于此相反，对于尚未识别为 EW 的其他 well，可发现支持下游的会较上游的多，这也可以用来辅助冲突的解决，如图 7 所示。若以上信息依旧比较模糊，可以利用离重复区起点一个插入片段长度范围内的错误右臂所对应的左臂来构建 HC，并期望 HC 上的 EW 集合与上游冲突碱基的 EW 集合比它与下游冲突碱基的差异更小。

两重复序列距离较近，这种情况下 well 的信息将不会发生明显变化而无法用于冲突的解决，但此时可能存在一些简单情况使得 HC 能够解决冲突，参见图 8a 和 8b：

a) 当冲突位点与重复区终点距离小于插入片段长度时，可发现支持上游碱基的左臂

reads 所对应的右臂能被定位到 HC 上 (HC 的构建方式与上一类大型重复相同, 也是利用离重复区起点一个插入片段长度范围内的错误右臂所对应的左臂来构建), 如图 8a 所示;

b) 当冲突位点与下游重复的起点间距离小于插入片段长度时, 可发现支持上游碱基的左臂 reads 所对应的右臂能在重复区内找到, 如图 8b 所示。

定义由重复单元的多个拷贝连续相接组成的重复为串联重复 (Tandem Repeat, TR), 在串联重复区中最短的重复片段为串联重复单元 (Tandem Repeat Unit, TRU)。一个串联重复单元具有多种不同的相位 (如 ACT, CTA, TAC), 其种类数目与其长度相等。

与常规重复造成的冲突相同, 串联重复区末端序列及重复区内的差异序列与重复区内的序列造成了冲突, 而与常规重复不同的是一个串联重复区相当于多块常规重复区连接而成。这使得这种重复的解决方法与其他常规重复的方法既有相同点, 又有不同之处。

如图 9 所示, 串联重复会使 reads 的定位发生偏移, 这往往造成一致性化时因重复终点或重复差异的 reads 定位错误而发生冲突。产生这种问题的核心实质是用于定位 reads 的 Kmer 是重复 Kmer。这种情况下错误定位的 reads 的头端都会在串联重复区起点之后 (因为起点前的非重复区会使错误定位的 reads 在精细比对过程中失败), 而 reads 因错误定位的偏移又小于插入片段长度的 SD, 即 mate-pair 无法滤掉这些定位错误的 reads, 进一步说, 如果串联重复的长度大于插入片段长度的 SD, 那么其重复区中的 reads 定位将发生以 SD 为单元的周期性的定位向延伸上游错误聚集的状况, 错误定位的 reads 距离偏差不会大于插入片段长度的 SD, 且都聚集在每个 SD 单元的头端, 这意味着在没遇到冲突之前串联重复区内每段 SD 长度单元都会被压缩变短, 下一 SD 单元的 reads 不断地被向前挪用。

解决 TR 的首要前提是正确识别, 在本算法中, TR 是通过发现串联重复单元来确认的。而由于大于插入片段长度的 TR 也符合 HC case2 的激活条件, 本算法将 TR 的激活判定置于 HC case2 的激活判定之前。

TRU 主要通过发现 Kmer 被周期性地使用来识别。对于小于 reads 长度-Kmer+1 的 TRU (即在一条 reads 上 TRU 出现两次或以上), 可以把冲突位点上的 reads 滑动切割成 Kmer 并查找一条 reads 中是否存在多次出现的 Kmer, 若大部分 reads 都存在这种状况, 且这些 reads 重复的 Kmer 一致或在 TRU 相位上一致, 则可以判断当前冲突是由串联重复造成, 解串联重复模块将被激活。而对于大于 reads 长度-Kmer+1 的 TRU, 则扫描重复区起点至冲突点的范围内的 Kmer, 若发现呈固定周期出现的 Kmer, 则可判断出冲突由串联重复造成。

识别出 TRU 的内容后, 可将冲突位点上的 reads 分为四类: 1) 在 TR 区内的 reads, 它们完全被 TRU 覆盖; 2) 包含 TR 起点的 reads, 这些 reads 仅在尾端发现 TRU; 3) 包含

TR 终点的 reads, 这些 reads 仅在头端发现 TRU; 4) 同时包含 TR 起点和终点的 reads, 这种情况仅在 TR 小于 reads 长度时才会存在, 且此时不会发现类型 1 的完全被 TRU 覆盖的 reads。

对于小于 reads 长度的 TR, 将包含 TR 终点的 reads 用终点对齐的方式来调整位置后即可消除冲突。而对于大于 reads 长度的 TR 则只能跨过并填入“N”来解决。其中对于大于 reads 长度并小于 reads 插入片段长度的 TR, 将有在非重复区内的左臂支持且包含 TR 终点的右臂 reads 按终点对齐并进行一致性化后就可依照这些 reads 的插入片段长度来估算距离并填入预估量数目的“N”; 对于大于插入片段长度的 TR, 可发现包含 TR 终点的右臂 reads 都没有在非重复区内的左臂 reads 支持, 此时直接对这些 reads 一致性化即可, 仅填入一个“N”作为标记。需要注意的是, TR 区内存在的细小差异(如出现个别碱基替换、插入和删除)会与 TR 终点混淆, 对识别出的包含 TR 终点的 reads 进行一致性化时就会发生冲突。此时可以采用 DBG 的方式对这些冲突的 reads 进行组装以自然构建出分离的 TR 区内的差异序列及 TR 终点序列, 再利用这些组装所得的序列上所定位的右臂 reads 对应的在前面已延伸的非重复区上的左臂 reads 的位置来计算组装所得序列的位置, 这样距离延伸点最远的一段序列就是正确的 TR 终点。

根据本发明的实施例, 由于 seed 的延伸是并行的, 为避免同一区域被多个 seed 重复延伸, 需要把已参与延伸的 reads 标记为“已使用”, 当其他 seed 初始化或延伸时发现这种 reads 时将停止延伸, 之后交由 contig 合并模块进行连接。需要注意的是, mate-pair 都在重复区内的 reads 是无法明确定位的, 为防止错误的延伸停止, 重复性质的 reads 将被标为“重复”, 并不会被用作冗余延伸的判定。

根据本发明的实施例, 当一个 well 中的 reads 被定位到 contig 上时, 其有可能是因重复或测序错误导致的错误定位, 所以本算法要求 well 对已延伸的 seed 区有相对足够的覆盖度时才能被认为是 EW, 同时覆盖度不足的 EW 将被丢弃。基于效率的考虑, 每达到一定的延伸长度才会进行 EW 集的更新, 且仅考察这个长度范围内的覆盖状况。

在所有 contig 都延伸完毕之后将对 contig 进行合并, 同时也会进行 contig 间定相的操作。可以合并的 contig 分为两种情况:

因延伸时发现含有“已使用”reads 而停止的 contig。此时直接通过这些“已使用”reads 在两条 contig 上的位置来计算两条 contig 的合并点。

因重复、局部测序深度过低等原因导致延伸停止的 contig。此时考察 contig 末端非重复区上的外向 reads(下游延伸时为左臂 reads, 上游延伸时为右臂 reads)的 mate-pair reads

是否被定位于其他 contig 上，若成功发现，则通过插入片段长度来估算两 contig 间的距离并填入预估量数目的“N”。

因为本算法中骨架序列构建的基本单元也是延伸终止而尚未合并的 contig，所以实际处理时重叠群合并步骤先建立出 contig 间的关系（交叠或是存在间隔地相接），再结合构建骨架序列步骤中建立的 contig 间关系进行互相检验，简化这些关系并解决冲突，再实质地进行 contig 的合并操作。

随着延伸的不断进行，可以发现 well/LFR 具有过渡性变化的特征，这种特征可以用于确定 contig 之间的位置关系，本算法最后对所有的 contig 进行了一次骨架序列构建（scaffolding）的操作。实际上，这种思维已在前面的大型重复的解决方法中有所体现。

本算法中，定义 contig 间主要通过 well 的信息来确定前后关系所形成的序列为骨架序列（scaffold），这与传统定义上通过 reads 的 mate-pair 信息来构建的骨架序列有着一定的区别。本算法中的 scaffold 仅表述 contig 的前后关系，但 contig 间的具体距离并无法确认，它们之间仅用一个“N”进行相连。而传统意义上的 scaffold 不仅通过 reads 的 mate-pair 来确定 contig 间的前后关系，还能通过成对 reads 分别定位在两 contig 上的位置及它们的插入片段长度来计算出 contig 间的距离，这与 contig 合并中的方法是一致的。无论是那种 scaffolding 方式都涉及 contig 的顺序排布，这是一个最优线性排布问题（Optimal Linear Arrangement, OLA），具有 NP-Hard 的性质。另外需要注意的是 scaffold 中的 contig 也是有自身方向的（四种脱氧核苷酸 A、C、T、G 按照以 3', 5' 磷酸二酯键相连形成的多聚脱氧核苷酸称为 DNA 碱基序列，是组装中 contig 和 scaffold 的表示形式，脱氧核苷酸的连接具有严格的方向性，是前一脱氧核苷酸的 3'-OH 与下一位脱氧核苷酸的 5'-位磷酸间形成 3',5' 磷酸二酯键，构成一个没有分支的线性 DNA 大分子，DNA 的表述中定义 5' 端向 3' 端为“+”向，3' 端向 5' 端为“-”向），位置正确而方向错误的 contig 也会造成大型组装错误，表现为子序列翻转型严重错误。

一般的，一个 well 会在多条临近的 contig 中出现，通过这个 well 就可以获得一组临近的 contig 集合。首先在 contig 所属的 well 中找到以该 contig 作为起点的 well，定义这个 contig 为这个 well 的起点 contig。然后找出包含这个 well 的其他 contig，并分别计算出它们所含的 well 集合与起点 contig 的 well 集合的交集，显然，与起点 contig 的 well 集合交集越大的 contig 离这个起点 contig 越近，由此可以确定一组 contig 的前后位置关系继而构建出一条 scaffold。

对于 contig 在 scaffold 中的方向性判断，需要将其头端和尾端的 well 集合分开考虑，

各自计算与附近 contig 头端或尾端的 well 集合的近似度。而当 contig 较短时，Well 的过渡性变化并不明显，此时可能会出现位置不明的情况，所以当完成所有 contig 的初步 scaffolding 之后，需要结合 contig 合并的结果进行检验和纠正，并合并两种方法所获得的关系，最终输出 scaffold 序列。

尽管已经示出和描述了本发明的实施例，本领域的普通技术人员可以理解：在不脱离本发明的原理和宗旨的情况下可以对这些实施例进行多种变化、修改、替换和变型，本发明的范围由权利要求及其等同物限定。

权利要求书

1、一种对分隔长片段序列进行组装的方法，其特征在于，包括：

(a) 通过测序获得读段集，并记录所述读段集中的读段对应的测序孔，一个测序孔包含至少一条长片段序列；

(b) 利用所述读段及所述读段对应的测序孔，对多个种子序列进行并行延伸，以获得多个序列重叠群，所述多个种子序列通过已知序列确定；

(c) 基于所述读段、所述序列重叠群以及所述序列重叠群包含的读段对应的测序孔，构建骨架序列，以获得分隔长片段序列的组装结果。

2、根据权利要求 1 所述的方法，其特征在于，所述种子序列是基于基因组参考序列按照下列步骤获得的：

在所述基因组参考序列中，按照 N 进行打断；以及

将经过打断的参考序列按照预定长度进行截断，以便获得所述种子序列。

3、根据权利要求 2 所述的方法，其特征在于，所述预定长度不小于所述测序中测序文库插入片段的长度。

4、根据权利要求 1 所述的方法，其特征在于，所述读段集包括成对读段，所述种子序列是基于所述读段按照下列步骤获得的：

(1) 将所述读段滑动切割成多个 Kmer，构建 Kmer 对所述读段的索引 RKI，用于通过 Kmer 访问对应的读段；

(2) 从所述读段集中提取一对不具有高频 Kmer 的成对读段；

(3) 利用所述索引 RKI，分别确定 (2) 中的一对成对读段中的两个读段的 Kmer 对应的所有读段，获得第一读段群和第二读段群；

(4) 分别确定 (3) 的第一读段群和第二读段群对应的测序孔，获得第一测序孔集和第二测序孔集；

(5) 确定 (4) 中的第一测序孔集和第二测序孔集的交集，若所述交集的大小与碱基的有效测序孔的数量期望值无显著差异，则确定 (2) 中的成对读段为所述种子序列。

5、根据权利要求 4 所述的方法，其特征在于，(5) 中：

若所述交集的大小在碱基的有效测序孔的数量期望值的一半到两倍之间，则确定 (2) 中的成对读段为所述种子序列。

6、根据权利要求 1 所述的方法，其特征在于，(b) 包括：

(i) 将所述读段滑动切割成多个 Kmer, 构建 Kmer 对所述读段的索引 RKI, 用于通过 Kmer 访问对应的读段;

(ii) 基于所述读段及其对应的索引 RKI, 对所述多个种子序列进行并行延伸, 以获得所述多个序列重叠群。

7、根据权利要求 6 所述的方法, 其特征在于, 所述 RKI 是通过下列步骤获得的:

对所述读段 reads 进行滑动切割成多个 Kmer;

构建以 Kmer 为键值的哈希, 所述哈希构成所述 RKI, 并且所述哈希记载所述 Kmer 的频率、所属读段、所述 Kmer 在所述读段上的位置和方向。

8、根据权利要求 6 所述的方法, 其特征在于, 通过重复下列步骤对所述种子序列进行延伸:

选择适于延伸的种子序列;

将所述读段定位至所述种子序列, 以获得延伸序列;

将定位在所述延伸序列末端的读段进行逐碱基一致性化处理; 以及

如果一致性化处理失败, 则进行杂合识别、定相处理和/或对重复序列进行解析。

9、根据权利要求 8 所述的方法, 其特征在于, 通过下列步骤选择适于延伸的种子序列:

将种子序列滑动切割成 Kmer;

通过所述 RKI 获取所述 Kmer 对应的读段;

将所述对应的读段与所述种子序列进行比对;

基于所述对应的读段对应的测序孔, 确定测序孔对所述种子序列的覆盖状况; 以及

基于所述覆盖状况, 确定适于延伸的种子序列。

10、根据权利要求 8 所述的方法, 其特征在于, 通过下列步骤将所述读段定位至所述种子序列:

将种子序列滑动切割成 Kmer;

通过所述 RKI 获取所述 Kmer 对应的读段;

将所述 Kmer 对应的读段定位至所述种子序列, 并逐个碱基进行比对。

11、根据权利要求 8 所述的方法, 其特征在于, 在一致性化处理过程中, 如果延伸的位点的有效测序孔集合被该位点不同的碱基型平均分配, 则判断存在杂合。

12、根据权利要求 11 所述的方法, 其特征在于, 在确定存在杂合之后, 将所述延伸序列分为多条分别进行延伸。

13、根据权利要求 8 所述的方法, 其特征在于, 所述读段集包含多对成对读段, 一对

成对读段中的两个读段之间的距离为 L ,

在一致性化处理过程中, 如果成对读段中的定位在延伸方向下游的读段与对应的读段之间的距离为非 L , 则确定该定位在延伸方向下游的读段所定位的位置为重复序列的起点。

14、根据权利要求 13 所述的方法, 其特征在于, 所述重复序列的终点为成对读段中的定位在延伸方向下游的读段所定位的位置, 该定位在延伸方向下游的读段与对应的读段之间的距离为 L ,

或者为延伸时的冲突位点。

15、根据权利要求 14 所述的方法, 其特征在于, 通过下列步骤判断所述重复序列是否为串联重复序列:

对定位至所述重复序列的读段进行滑动切割以获得 $Kmer$;

判断每条所述读段进行滑动切割所得到的 $Kmer$ 是否存在重复 $Kmer$, 如果不存在所述重复 $Kmer$, 则判断不存在串联重复序列, 如果存在所述重复 $Kmer$, 则判断存在串联重复序列。

16、根据权利要求 15 所述的方法, 其特征在于, 若所述重复序列为串联重复序列, 通过下列步骤对所述串联重复序列进行解析:

(m) 确定所述串联重复序列的长度;

(n) 将包含所述串联重复序列终点的读段用终点对齐的方式进行位置调整。

17、根据权利要求 15 所述的方法, 其特征在于, 若所述重复序列非为串联重复序列, 且存在以下情况, 则通过以下判断所述重复序列为大型双重复序列:

所述重复序列的长度大于 L , 或者成对读段中的定位在所述重复序列下游的读段对应的读段也定位在该重复序列上。

18、根据权利要求 17 所述的方法, 其特征在于, 若所述重复序列为大型双重复序列, 通过以下对所述大型双重复序列进行解析:

比较所述大型双重复序列中的上游重复序列对应的有效测序孔的数量与下游重复序列对应的有效测序孔的数量的差值与碱基的有效测序孔的数量期望值的差异, 依据差异的大小, 解决所述大型双重复序列上的冲突。

19、根据权利要求 17 所述的方法, 其特征在于, 若所述重复序列非为大型双重复序列, 且存在以下情况, 则判断所述重复序列为小型双重复序列:

所述重复序列的长度小于 L 。

20、根据权利要求 19 所述的方法, 其特征在于, 通过以下至少之一对所述小型双重复

序列进行解析：

(p) 利用支持各碱基的成对读段之间的距离的均值作为冲突位点的期望位置，通过比较支持两种冲突碱基的成对读段之间的距离与所述均值的接近程度，以确定该冲突位点上的碱基；

(k) 利用成对读段中定位至所述延伸序列标准差范围中的非重复序列的最上游的读段对应的读段来构建辅助重叠群，利用成对读段中定位至所述辅助重叠群下游的读段对应的读段来确定冲突位点上的碱基。

21、根据权利要求 20 所述的方法，其特征在于，若无法解析所述小型双重复序列，则终止对种子序列的延伸。

22、根据权利要求 1 所述的方法，其特征在于，(c) 包括：

(iii) 基于所述读段，建立所述序列重叠群之间的合并连接关系；

(iv) 基于所述序列重叠群和所述序列重叠群包含的读段对应的测序孔，构建初级骨架序列；

(v) 在对所述多个序列重叠群之间的合并连接关系与所述初级骨架序列进行相互检验之后，通过合并所述序列重叠群以构建所述骨架序列，获得分隔长片段序列的组装结果。

23、一种对分隔长片段序列进行组装的装置，其特征在于，包括：

输入模块，通过测序获得读段集，并记录所述读段集中的读段对应的测序孔，一个测序孔包含至少一条长片段序列；

延伸模块，利用所述读段及所述读段对应的测序孔，对多个种子序列进行并行延伸，以获得多个序列重叠群，所述多个种子序列通过已知序列确定；

骨架序列构建模块，基于所述读段、所述序列重叠群以及所述序列重叠群包含的读段对应的测序孔，构建骨架序列，以获得分隔长片段序列的组装结果。

24、根据权利要求 23 所述的装置，其特征在于，所述种子序列是基于基因组参考序列按照下列步骤获得的：

在所述基因组参考序列中，按照 N 进行打断；以及

将经过打断的参考序列按照预定长度进行截断，以便获得所述种子序列。

25、根据权利要求 24 所述的装置，其特征在于，所述预定长度不小于所述测序中测序文库插入片段的长度。

26、根据权利要求 23 所述的装置，其特征在于，所述读段集包括成对读段，所述种子序列是基于所述读段按照下列步骤获得的：

(1) 将所述读段滑动切割成多个 Kmer, 构建 Kmer 对所述读段的索引 RKI, 用于通过 Kmer 访问对应的读段;

(2) 从所述读段集中提取一对不具有高频 Kmer 的成对读段;

(3) 利用所述索引 RKI, 分别确定 (2) 中的一对成对读段中的两个读段的 Kmer 对应的所有读段, 获得第一读段群和第二读段群;

(4) 分别确定 (3) 的第一读段群和第二读段群对应的测序孔, 获得第一测序孔集和第二测序孔集;

(5) 确定 (4) 中的第一测序孔集和第二测序孔集的交集, 若所述交集的大小与碱基的有效测序孔的数量期望值无显著差异, 则确定 (2) 中的成对读段为所述种子序列。

27、根据权利要求 26 所述的装置, 其特征在于, (5) 中:

若所述交集的大小在碱基的有效测序孔的数量期望值的一半到两倍之间, 则确定 (2) 中的成对读段为所述种子序列。

28、根据权利要求 23 所述的装置, 其特征在于, 在所述延伸模块中包括进行:

(i) 将所述读段滑动切割成多个 Kmer, 构建 Kmer 对所述读段的索引 RKI, 用于通过 Kmer 访问对应的读段;

(ii) 基于所述读段及其对应的索引 RKI, 对所述多个种子序列进行并行延伸, 以获得所述多个序列重叠群。

29、根据权利要求 28 所述的装置, 其特征在于, 所述 RKI 是通过下列步骤获得的:

对所述读段进行滑动切割成多个 Kmer;

构建以 Kmer 为键值的哈希, 所述哈希构成所述 RKI, 并且所述哈希记载所述 Kmer 的频率、所属读段、所述 Kmer 在所述读段上的位置和方向。

30、根据权利要求 28 所述的装置, 其特征在于, 通过重复下列步骤对所述种子序列进行延伸:

选择适于延伸的种子序列;

将所述读段定位至所述种子序列, 以获得延伸序列;

将定位在所述延伸序列末端的读段进行逐碱基一致性化处理; 以及

如果一致性化处理失败, 则进行杂合识别、定相处理和/或对重复序列进行解析。

31、根据权利要求 30 所述的装置, 其特征在于, 通过下列步骤选择适于延伸的种子序列:

将种子序列滑动切割成 Kmer;

通过所述 RKI 获取所述 Kmer 对应的读段；

将所述对应的读段与所述种子序列进行比对；

基于所述对应的读段对应的测序孔，确定测序孔对所述种子序列的覆盖状况；以及

基于所述覆盖状况，确定适于延伸的种子序列。

32、根据权利要求 30 所述的装置，其特征在于，通过下列步骤将所述读段定位至所述种子序列：

将种子序列滑动切割成 Kmer；

通过所述 RKI 获取所述 Kmer 对应的读段；

将所述 Kmer 对应的读段定位至所述种子序列，并逐个碱基进行比对。

33、根据权利要求 30 所述的装置，其特征在于，在一致性化处理过程中，如果延伸的位点的有效测序孔集合被该位点不同的碱基型平均分配，则判断存在杂合。

34、根据权利要求 33 所述的装置，其特征在于，在确定存在杂合之后，将所述延伸序列分为多条分别进行延伸。

35、根据权利要求 30 所述的装置，其特征在于，所述读段集包含多对成对读段，一对成对读段中的两个读段之间的距离为 L，

在一致性化处理过程中，如果成对读段中的定位在延伸方向下游的读段与对应的读段之间的距离为非 L，则确定该定位在延伸方向下游的读段所定位的位置为重复序列的起点。

36、根据权利要求 35 所述的装置，其特征在于，所述重复序列的终点为成对读段中的与对应的读段之间的距离为 L 的定位在延伸方向下游的读段所定位的位置，或者为延伸时的冲突位点。

37、根据权利要求 36 所述的装置，其特征在于，通过下列步骤判断所述重复序列是否为串联重复序列：

对定位至所述重复序列的读段进行滑动切割以获得 Kmer；

判断每条所述读段进行滑动切割所得到的 Kmer 是否存在重复 Kmer，如果不存在所述重复 Kmer，则判断不存在串联重复序列，如果存在所述重复 Kmer，则判断存在串联重复序列。

38、根据权利要求 37 所述的装置，其特征在于，若所述重复序列为串联重复序列，通过下列步骤对所述串联重复序列进行解析：

(m) 确定所述串联重复序列的长度；

(n) 将包含所述串联重复序列终点的读段用终点对齐的方式进行位置调整。

39、根据权利要求 37 所述的装置，其特征在于，若所述重复序列非为串联重复序列，且存在以下情况，则判断所述重复序列为大型双重复序列：

所述重复序列的长度大于 L ，或者成对读段中的定位在所述重复序列下游的读段对应的读段也定位在该重复序列上。

40、根据权利要求 39 所述的装置，其特征在于，若所述重复序列为大型双重复序列，通过以下对所述大型双重复序列进行解析：

比较所述大型双重复序列中的上游重复序列对应的有效测序孔的数量与下游重复序列对应的有效测序孔的数量的差值与碱基的有效测序孔的数量期望值的差异，依据差异的大小，解决所述大型双重复序列上的冲突。

41、根据权利要求 39 所述的装置，其特征在于，若所述重复序列非为大型双重复序列，且存在以下情况，则判断所述重复序列为小型双重复序列：

所述重复序列的长度小于 L 。

42、根据权利要求 41 所述的装置，其特征在于，若所述重复序列为小型双重复序列，通过以下至少之一对所述小型双重复序列进行解析：

(p) 利用支持各碱基的成对读段之间的距离的均值作为冲突位点的期望位置，通过比较支持两种冲突碱基的成对读段之间的距离与所述均值的接近程度，以确定该冲突位点上的碱基；

(k) 利用成对读段中定位至所述延伸序列标准差范围中的非重复序列的最上游的读段对应的读段来构建辅助重叠群，利用成对读段中定位至所述辅助重叠群下游的读段对应的读段来确定冲突位点上的碱基。

43、根据权利要求 42 所述的装置，其特征在于，若无法解析所述小型双重复序列，则终止对种子序列的延伸。

44、根据权利要求 1 所述的方法，其特征在于，在所述骨架序列构建模块中进行：

(iii) 基于所述读段，建立所述序列重叠群之间的合并连接关系；

(iv) 基于所述序列重叠群和所述序列重叠群包含的读段对应的测序孔，构建初级骨架序列；

(v) 在对所述多个序列重叠群之间的合并连接关系与所述初级骨架序列进行相互检验之后，通过合并所述序列重叠群以构建所述骨架序列，获得分隔长片段序列的组装结果。

45. 一种计算机可读介质，其特征在于，用于存储计算机可执行程序，执行所述程序包括完成权利要求 1-22 任一方法。

46. 一种对分隔长片段序列进行组装的系统，其特征在于，包括：

数据输入单元，用于输入数据；

数据输出单元，用于输出数据；

存储单元，用于存储数据，其中包括计算机可执行程序；

处理器，与所述数据输入单元、所述数据输出单元和所述存储单元连接，用于执行所述计算机可执行程序，执行所述程序包括完成权利要求 1-22 任一方法。

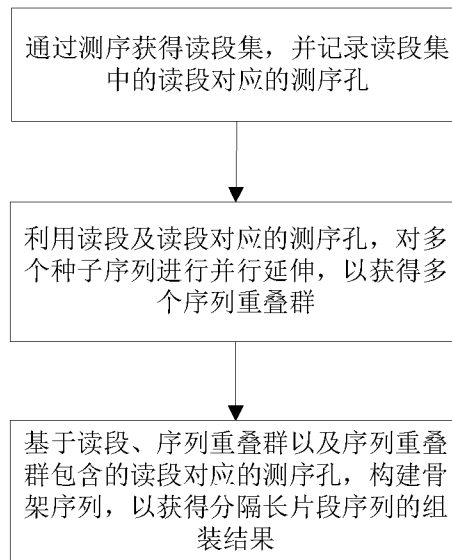


图 1

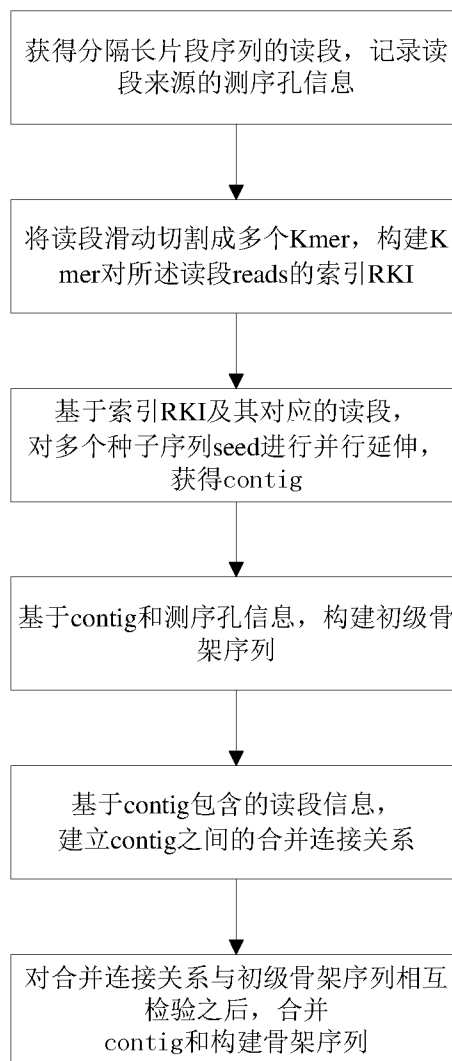


图 2

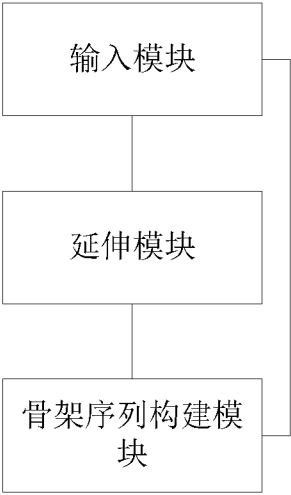


图 3

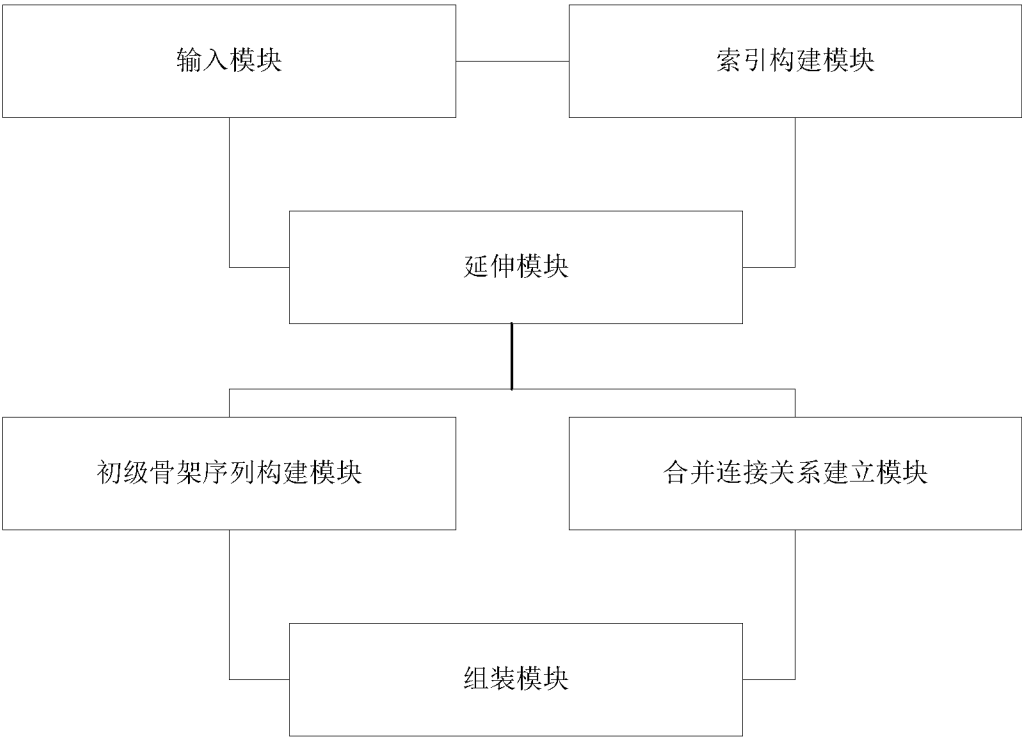


图 4

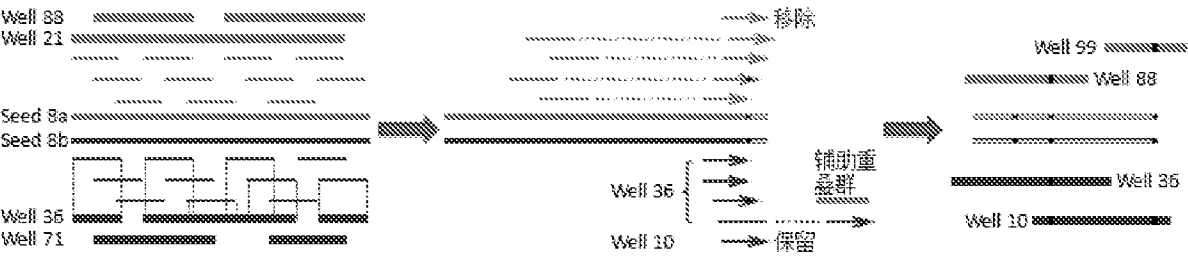


图 5

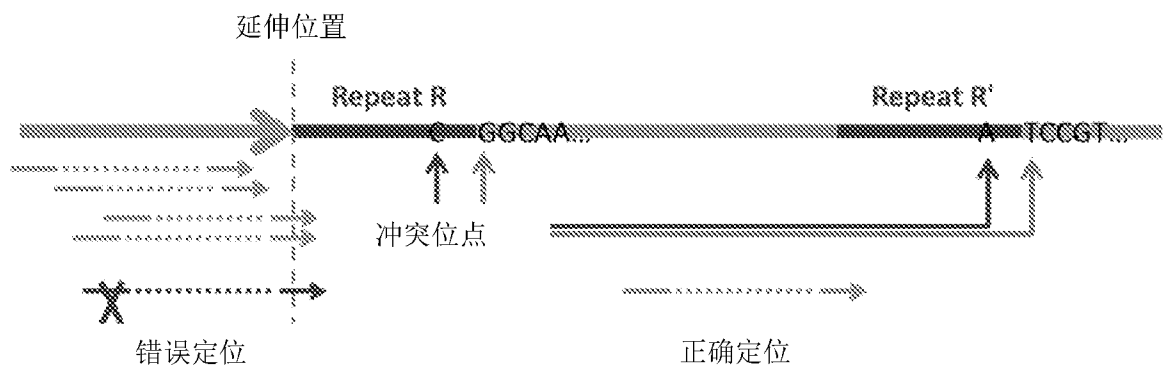


图 6

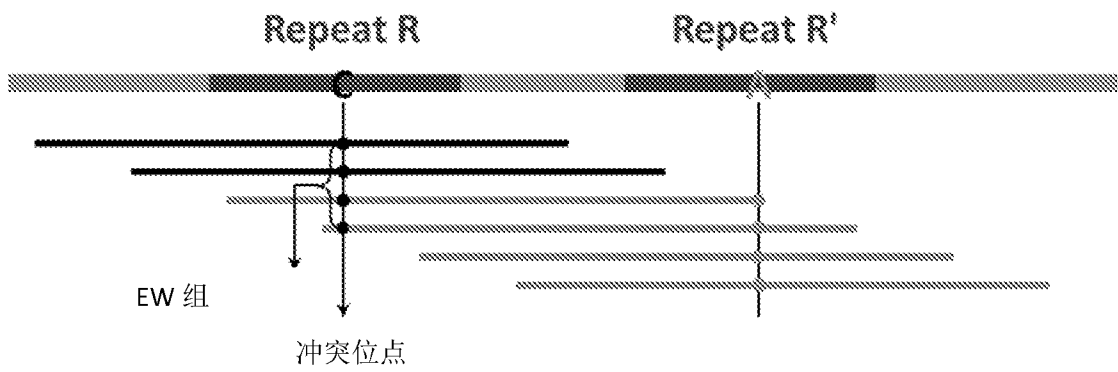
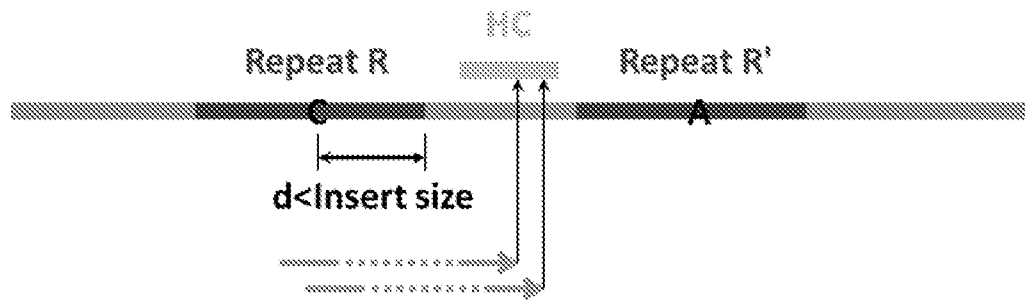


图 7

a)



b)

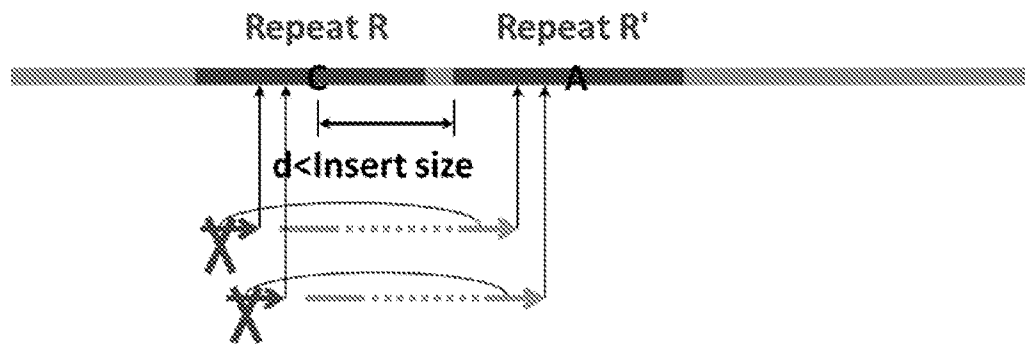


图 8

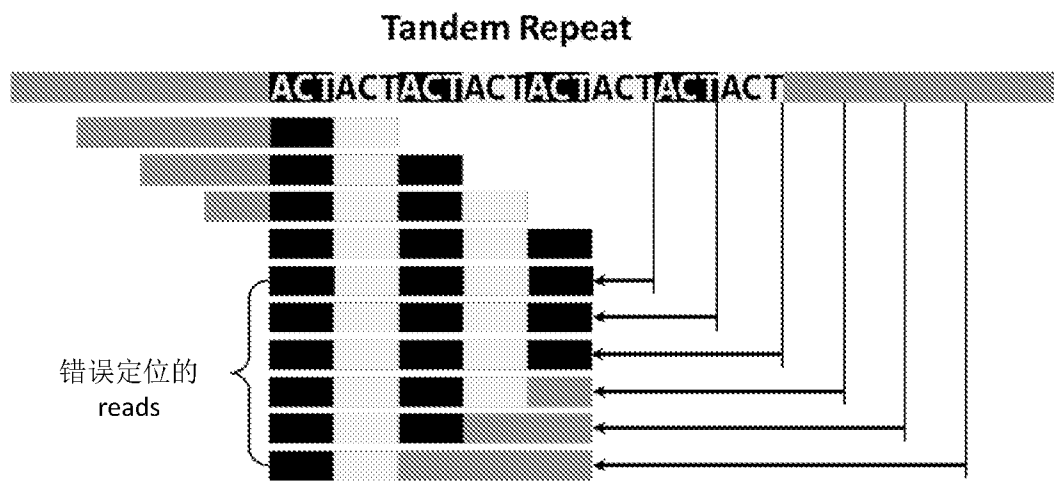


图 9

INTERNATIONAL SEARCH REPORT

International application No.

PCT/CN2016/074665

A. CLASSIFICATION OF SUBJECT MATTER

C12Q 1/68 (2006.01) i; G06F 19/20 (2011.01) 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)

C12Q; G06F

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; DWPI; LEXTXT; CATXT; USTXT; GBTXT; EPTXT; CNTXT; CNABS; WOTXT; CHTXT; SGTXT; ATTXT; TWTXT; ISI WEB OF KNOWLEDGE; SPRINGER LINK; GOOGLE; PUBMED; CNKI; BGI SHENZHEN, assembling, sequence, splicing and assembling, perforated plate, seed sequence, physical separation, fragment reads, well, long fragment reads, framework, separate, kmer, fragment, contig, haplotyping, plate, assembl+, LFR, pool, reads, reference sequence, splicing, read?, seed?, RKI, index, delay, scaffold, long, fosmid

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	CN 102206704 A (BGI SHENZHEN CO., LIMITED et al.), 05 October 2011 (05.10.2011), the whole document	1-44, 46
A	CN 102789553 A (CHINESE ACADEMY OF FISHERY SCIENCES), 21 November 2012 (21.11.2012), the whole document	1-44, 46
A	US 2015094961 A1 (COMPLETE GENOMICS, INC.), 02 April 2015 (02.04.2015), the whole document	1-44, 46
A	CN 104164479 A (BGI TECH SOLUTIONS CO., LTD.), 26 November 2014 (26.11.2014), the whole document	1-44, 46
A	CN 104017883 A (BGI TECH SOLUTIONS CO., LTD.), 03 September 2014 (03.09.2014), the whole document	1-44, 46
A	CN 105189308 A (COMPLETE GENOMICS INCORPORATED), 23 December 2015 (23.12.2015), the whole document	1-44, 46

☐ Further documents are listed in the continuation of Box C.☒ See patent family annex.

* Special categories of cited documents:	“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
“A” document defining the general state of the art which is not considered to be of particular relevance	“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
“E” earlier application or patent but published on or after the international filing date	“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
“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)	“&” document member of the same patent family
“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	

Date of the actual completion of the international search 14 November 2016 (14.11.2016)	Date of mailing of the international search report 25 November 2016 (25.11.2016)
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 ZHANG, Yanqing Telephone No.: (86-10) 010-62413836

INTERNATIONAL SEARCH REPORT

International application No.

PCT/CN2016/074665

Box No. II Observations where certain claims were found unsearchable (Continuation of item 2 of first sheet)

This international search report has not been established in respect of certain claims under Article 17(2)(a) for the following reasons:

1. ☒ Claims Nos.: 45
because they relate to subject matter not required to be searched by this Authority, namely:
[1] Claim 45 relates to a computer readable medium, characterized in that it is used for storing a computer-executable program, and it is substantially information record on a carrier, which falls within simple information expression of the cases set out in PCT Rule 39.1(iv), and thus an international search is not required.
2. ☐ Claims Nos.:
because they relate to parts of the international application that do not comply with the prescribed requirements to such an extent that no meaningful international search can be carried out, specifically:
3. ☐ Claims Nos.:
because they are dependent claims and are not drafted in accordance with the second and third sentences of Rule 6.4(a).

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.

PCT/CN2016/074665

Patent Documents referred in the Report	Publication Date	Patent Family	Publication Date
CN 102206704 A	05 October 2011	CN 102206704 B	20 November 2013
		HK 1162614 A1	11 April 2014
		WO 2012116658 A2	07 September 2012
		US 2013345095 A1	26 December 2013
		HK 1162614 A0	31 August 2012
CN 102789553 A	21 November 2012	CN 102789553 B	15 April 2015
US 2015094961 A1	02 April 2015	CN 105593683 A	18 May 2016
		WO 2015051006 A2	09 April 2015
		WO 2015051006 A3	28 May 2015
CN 104164479 A	26 November 2014	WO 2015149719 A1	08 October 2015
		HK 1201887 A0	11 September 2015
CN 104017883 A	03 September 2014	CN 104017883 B	18 November 2015
CN 105189308 A	23 December 2015	US 2016046985 A1	18 February 2016
		WO 2014145820 A3	27 November 2014
		WO 2014145820 A2	18 September 2014
		US 9328382 B2	03 May 2016
		CA 2902882 A1	18 September 2014
		US 2014323316 A1	30 October 2014
		AU 2014232515 A1	03 September 2015
		EP 2969847 A2	20 January 2016

A. 主题的分类 C12Q 1/68(2006.01)i; G06F 19/20(2011.01)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) C12Q; G06F 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;DWPI;LEXTXT;CATXT;USTXT;GBTXT;EPTXT;CNTXT;CNABS;WOTXT;CHTXT;SGTXT;ATTXT;TWTXT;ISI WEB OF KNOWLEDGE;SPRINGER LINK;GOOGLE;PUBMED;CNKI:深圳华大基因, 组装, 序列, 拼接组装, 多孔板, 种子序列, 物理分隔, fragment reads, well, 长片段读, 骨架, 分隔, kmer, fragment, contig, 重叠群, haplotyping, plate, assembl+, LFR, 读段, pool, reads, 种子, 序列, fosmid质粒, 参考序列, 拼接, read?, 单体型, seed?, RFL, 索引, 延, scaffold, long, fosmid		
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	CN 102206704 A (深圳华大基因科技有限公司等) 2011年 10月 5日 (2011 - 10 - 05) 全文	1-44, 46
A	CN 102789553 A (中国水产科学研究院) 2012年 11月 21日 (2012 - 11 - 21) 全文	1-44, 46
A	US 2015094961 A1 (COMPLETE GENOMICS, INC.) 2015年 4月 2日 (2015 - 04 - 02) 全文	1-44, 46
A	CN 104164479 A (深圳华大基因科技服务有限公司) 2014年 11月 26日 (2014 - 11 - 26) 全文	1-44, 46
A	CN 104017883 A (深圳华大基因科技服务有限公司) 2014年 9月 3日 (2014 - 09 - 03) 全文	1-44, 46
A	CN 105189308 A (考利达基因组股份有限公司) 2015年 12月 23日 (2015 - 12 - 23) 全文	1-44, 46
☐ 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 2016年 11月 14日		国际检索报告邮寄日期 2016年 11月 25日
ISA/CN的名称和邮寄地址 中华人民共和国国家知识产权局(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 张艳青 电话号码 (86-10)010-62413836

第II栏

某些权利要求被认为是不能检索的意见(续第1页第2项)

根据条约第17条(2)(a)，对某些权利要求未做国际检索报告的理由如下：

1. ☒ 权利要求： 45
因为它们涉及不要求本单位进行检索的主题，即：
[1] 权利要求45涉及一种计算机可读介质，其特征不在于用于存储计算机可执行程序，实质上是记录在载体上的信息，属于PCT细则39.1(5)所列的单纯的信息表达的情形，无需进行国际检索。
2. ☐ 权利要求：
因为它们涉及国际申请中不符合规定的要求的部分，以致不能进行任何有意义的国际检索， 具体地说：
3. ☐ 权利要求：
因为它们是从属权利要求，并且没有按照细则6.4(a)第2句和第3句的要求撰写。

国际检索报告
关于同族专利的信息

国际申请号

PCT/CN2016/074665

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)
CN	102206704	A	2011年 10月 5日	CN	102206704	B	2013年 11月 20日
				HK	1162614	A1	2014年 4月 11日
				WO	2012116658	A2	2012年 9月 7日
				US	2013345095	A1	2013年 12月 26日
				HK	1162614	A0	2012年 8月 31日
CN	102789553	A	2012年 11月 21日	CN	102789553	B	2015年 4月 15日
US	2015094961	A1	2015年 4月 2日	CN	105593683	A	2016年 5月 18日
				WO	2015051006	A2	2015年 4月 9日
				WO	2015051006	A3	2015年 5月 28日
CN	104164479	A	2014年 11月 26日	WO	2015149719	A1	2015年 10月 8日
				HK	1201887	A0	2015年 9月 11日
CN	104017883	A	2014年 9月 3日	CN	104017883	B	2015年 11月 18日
CN	105189308	A	2015年 12月 23日	US	2016046985	A1	2016年 2月 18日
				WO	2014145820	A3	2014年 11月 27日
				WO	2014145820	A2	2014年 9月 18日
				US	9328382	B2	2016年 5月 3日
				CA	2902882	A1	2014年 9月 18日
				US	2014323316	A1	2014年 10月 30日
				AU	2014232515	A1	2015年 9月 3日
				EP	2969847	A2	2016年 1月 20日