(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 106021997 B
(45)授权公告日 2019.03.29

(21)申请号 201610329027.4

(22)申请日 2016.05.17

(65)同一申请的已公布的文献号
申请公布号 CN 106021997 A

(43)申请公布日 2016.10.12

(73)专利权人 杭州和壹基因科技有限公司
地址 310052 浙江省杭州市滨江区长河街道聚才路88号远方科技中心15楼

(72)发明人 詹东亮 王军一 郝美荣 何荣军 俞凯成 高金龙 蔡庆乐

(74)专利代理机构 杭州中成专利事务所有限公司 33212
代理人 唐银益

(51)Int.Cl.
G16B 20/00(2019.01)

(56)对比文件
CN 104164479 A,2014.11.26,
CN 104531848 A,2015.04.22,

审查员 卜婷婷

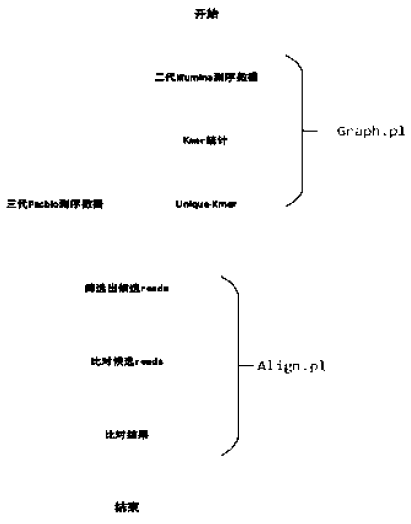
权利要求书1页 说明书5页 附图2页

(54)发明名称

一种三代PacBio测序数据的比对方法

(57)摘要

本发明提供一种有效降低重复序列造成的比对错误的三代PacBio测序数据的比对方法。它使用二代的Illumina数据建立k-mer模型，提取unique-kmer，在三代PacBio测序数据的比对中，使用这个unique-kmer来作为比对时使用的种子(seed)，能大大地降低重复序列的影响，提高比对的速度。



CN 106021997 B

1.一种三代PacBio测序数据的比对方法，其特征在于，它包括以下步骤：

(1)使用二代Illumina测序数据建立kmer模型，并从中提取出unique-kmer；

(2)使用unique-kmer把它作为比对的seed，与三代Pacbio测序数据进行比对，筛选出候选reads；

(3)对候选reads进行详细比对，包括以下步骤：

a．先对比对上的seed进行聚类，算出最可能的比对范围，方法如下：

建立坐标系，横坐标代表read1比对上的位置，纵坐标代表read2上比对上的位置，每个点代表两条read上共有的seed，将这些seed用斜率为1的直线进行聚类，将聚到最多点的直线作为比对上的区域；

b．再将比对范围进行小区域分割，对每一个分割区域，使用LCS算法计算相似度，再对整体进行打分，方法如下：

假设将比对范围分为n个区域，相似度大于0.8的区域有b个，这些小区域总体的相似碱基为c个，则区域相似度为b/n，碱基相似度为c/a，最后只保留这两个值都大于0.7的数据。

2.根据权利要求书1中所述的三代PacBio测序数据的比对方法，其特征在于，在所述步骤(1)中，使用jellyfish软件对二代Illumina测序数据进行k-mer统计，根据k-mer分布图获取二倍主峰以内的k-mer作为unique-kmer，并使用比特文件或GATB开源包，对所述unique-kmer进行存储。

3.根据权利要求书2中所述的三代PacBio测序数据的比对方法，其特征在于，对于$k \leq 17$，使用一个大小为2G的比特文件*.bit来存储，而对于$k > 17$的情况，把unique-kmer存入GATB开源包中的*.h5文件中。

4.根据权利要求书1中所述的三代PacBio测序数据的比对方法，其特征在于，在所述步骤(2)中，使用步骤(1)的unique-kmer，如果reads之间共有的unique-kmer计数超过3，就把这些reads筛选出来，作为候选reads。

一种三代PacBio测序数据的比对方法

技术领域

[0001] 本发明涉及生物信息技术领域,具体涉及DNA序列的比对方法,它使用二代的Illumina测序数据进行建模提取关键信息,并利用这些关键信息来辅助三代PacBio测序数据的比对。

背景技术

[0002] 三代PacBio的测序数据,单次测序的错误率约为15%,专门支持三代的比对软件并不多,目前使用最多的软件为以下两款:(1)blasr;(2)dalign。

[0003] 这两款都是非常优秀的三代比对软件,能支持PacBio的高错误率。由于基因组本身存在重复序列,它们拥有高度相似的序列。而这些比对软件,会将这些重复序列进行比对和输出,从而影响后续的生物学分析(比如组装,表达量分析等)。

发明内容

[0004] 本发明的目的是解决以上提出的问题,提供一种有效降低重复序列造成的比对错误的三代PacBio测序数据的比对方法。它使用二代的Illumina数据建立kmer模型,提取unique-kmer,在三代PacBio测序数据的比对中,使用这个unique-kmer来作为比对时使用的种子(seed),能大大地降低重复序列的影响,提高比对的速度。

[0005] 本发明是通过以下技术方案实现的:

[0006] 本发明是一种三代PacBio测序数据的比对方法,它包括以下步骤:

[0007] (1)使用Illumina测序数据建立kmer模型,从中提取unique-kmer;

[0008] (2)使用unique-mer作为比对的seed进行候选reads筛选

[0009] (3)对候选reads进行详细比对。

[0010] 作为优化,使用jellyfish软件对二代Illumina测序数据进行k-mer统计,根据k-mer分布图获取二倍主峰以内的k-mer作为unique-kmer,并使用比特文件或GATB开源包,对所述unique-kmer进行存储。

[0011] 作为优化,对于k≤17,使用一个大小为2G的比特文件(*.bit)来存储,而对于k>17的情况,把unique-kmer存入GATB开源包中的(*.h5)文件中。

[0012] 作为优化,在步骤(2)中,使用步骤(1)的unique-kmer,如果reads之间共有的unique-kmer计数超过3,就把这些reads筛选出来,作为候选reads。

[0013] 作为优化,所述步骤(3)包括以下步骤:

[0014] a.先对比对上的seed进行聚类,算出最可能的比对范围,方法如下:

[0015] 建立坐标系,横坐标代表read1比对上的位置,纵坐标代表read2上比对上的位置,每个点代表两条read上共有的seed,将这些seed用斜率为1的直线进行聚类,将聚到最多点的直线作为比对上的区域;

[0016] b.再将比对范围进行小区域分割,对每一个分割区域,使用LCS算法计算相似度,再对整体进行打分,方法如下:

[0017] 假设将比对范围分为n个区域，相似度大于0.8的区域有b个，这些小区域总体的相似碱基为c个，则区域相似度为b/n，碱基相似度为c/a，最后只保留这两个值都大于0.7的数据。

[0018] 本发明的有益效果如下：

[0019] 1、使用二代Illumina测序数据提取unique-kmer，提高比对的准确率和速度。

[0020] 在基因组中，存在许多重复序列，有些短重复序列甚至出现成百上千次，从而会影响比对的准确度，增加比对的时间。为了提高比对的准确度，降低比对时间，我们提取在contig中只出现一次的k-mer，作为unique-kmer。因为二代Illumina测序数据的质量非常高，在测序深度足够随机的情况下(一般情况为～40x)，使用Jellyfish软件对二代Illumina测序数据进行kmer统计，可以得到k-mer的分布图(图1)。将峰值2倍内区域的k-mer作为unique-kmer。对于k<＝17，使用一个大小为2G的比特文件(*.bit文件)来存储，而对于k>17的情况，使用GATB(开源框架)，把unique-kmer存入文件(*.h5文件)。其中所使用的二代Illumina测序数据质量较高，Jellyfish软件具有多线程运行，速度快，内存消耗小的优点，保证了整个方法具有较高的数据处理质量，以及明显的处理速度优势；

[0021] 2、使用unique-kmer作为比对的seed进行候选reads筛选，节约比对时间，提高比对速度。

[0022] 因为unique-kmer在概率和理论上，在单倍体的基因组中，只会出现一次，从而能避免重复序列造成的影响。另一方面，由于避免了重复序列的影响，找到的候选reads准确度非常高，节约了很多比对时间，大大提高了比对速度。

[0023] 3、对候选的reads进行详细比对，节约了内存和比对时间，提高比对速度。

[0024] 很多比对软件的比对方法，都使用了最长公共子序列(LCS)的算法，直接对整体区域进行LCS计算，对于大于100k的比对区域则非常浪费内存和时间。本方法也是使用这个算法，但是做了两方面的改进：(1)事先对seed的比对关系进行聚类，算出最优的比对范围；(2)分区域进行比对。从而节约了内存和比对时间，提高比对速度。

附图说明

[0025] 图1：kmer分布图

[0026] 将所有的数据打断成长度为k的片断(称为k-mer)，横坐标为在k-mer的频数，纵坐标为该频数k-mer的种类，将峰值2倍内区域的k-mer作为unique-kmer。

[0027] 图2：计算出比对范围示意图

[0028] 图上的每个点代表两条read上共有的seed，横坐标代表read1比对上的位置，纵坐标代表read2比对上的位置，将这些seed用斜率为1的直线进行聚类，选出聚类最多的直线，将这个区域作为比对上的范围。

[0029] 图3：本发明流程图

具体实施方式

[0030] 下面结合附图对本发明的实施例进行进一步详细说明：

[0031] 实施例：

[0032] (1)使用二代Illumina测序数据建立kmer模型，从中提取unique-kmer

[0033] 使用jellyfish软件对二代Illumina测序数据进行k-mer统计，将所有的数据打断成长度为k的片断(称为k-mer)，横坐标为在k-mer的频数，纵坐标为该频数k-mer的种类。根据k-mer分布图获取二倍主峰以内的k-mer作为unique-kmer，对于k≤17，使用一个大小为2G的比特文件(*.bit)来存储，而对于k＞17的情况，把unique-kmer存入GATB开源包中的(*.h5)文件中。其中，二代Illumina测序数据是指通过Illumina公司测序仪获得的二代测序数据。

[0034] 根据上述方法，编写如下程序，用来提取unique-kmer，具体操作命令使用说明如下：

[0035]

```
perl Graph.pl
    Name
        Graph.pl    --The De novo tool to build k-mer graph
     Usage
         Graph.pl    <command> [arguments]
             Command should be one of the following command.
             Arguments depend on specific command.
     Command List:
                    count    Record k-mer and related occurrence by jellyfish
                               Arguments:
                                 -i <FILE>          the file list to count kmer
                                 -k <INT>           the k-mer size to store occupied k-mer [17]

                    kmer     Record the unique k-mers into .h5 file, recommend for k>17.
                               Arguments:
                                 -i <FILE>          the kmer table file with format "ATGC 1"
                                 -m <INT>           the mininum occurrence of kmer [3]
                                 -x <INT>           the maxinum occurrence of kmer [-1]
                                 -k <INT>           the k-mer size to store occupied k-mer [17]

                    graph    Build graph by using GATB tookit(for Lordec)
                               Arguments:
                                 -i <FILE>          the kmer table file with format "ATGC 1"
                                 -m <INT>           the mininum occurrence of kmer [3]
                                 -k <INT>           the k-mer size

                    bit      Record k-mer into bitset, this method is for k<=17.
```

[0036]

```
                    Arguments:
                        -i <FILE>          the kmer table file with format "ATGC 1"
                        -m <INT>           the mininum occurrence of kmer [3]
                        -x <INT>           the maxinum occurrence of kmer [-1]
                        -k <INT>           the k-mer size to store occupied k-mer [17]

            pipe    Combine step(s) above
                    Arguments:
                        -i <FILE>          the file list
                        -m <INT>           the mininum occurrence of kmer [3]
                        -k <INT>           the k-mer size to store occupied k-mer [17]
                        -s <INTs>          the step to do
                                               1: count k-mer by jellyfish
                                               2: record unique k-mer into .h5 file
                                               3: record all k-mer into .h5 file
                                               4: record unique k-mer into bitset

                                -d                 the output directory
```

[0037] 具体案例实施操作如下：

[0038] 从二代的Illumina测序数据中，筛选大约40X的数据，把它写入一个叫fq.lst文件中：

[0039]

```
read1.fastq
read2.fastq
```

[0040] 然后运行程序，来获取unique-kmer:

[0041]

```
perl Graph.pl pipe -i fq.lst -m 2 -k 17 -s 1,4 -dir Kmer_17
```

[0042] 因为选取k＝17，将结果存入比特文件中：k17.bit

[0043] (2)使用unique-kmer与三代Pacbio测序数据进行比对，筛选候选reads

[0044] 使用这个unique-kmer来作为比对时使用的种子(seed)，如果reads间共有的unique-kmer超过3时，把它们作为候选reads。其中，三代Pacbio测序数据是指通过Pacbio公司测序仪获得的二代测序数据。

[0045] 根据上述方法，编写一个比对程序，来对三代Pacbio测序数据进行比对，具体操作命令使用说明如下：

[0046]

```
perl Align.pl
    Name
        Align.pl    --The Alignment Tool
```

[0047]

```
Usage
    perl  Align.pl  [arguments]  <reference.fa>  <query.fa>
                    Argument List:
                        -g <FILE>       the unique kmer(.h5 or .bit)
                        -k <INT>        the kmer size of the unique graph
                        -f <INT>        the min kmer num[1]
                        -u <INT>        the min unique kmer[0]

                        -s <INT>        split the reference file, in unit of M[100].
                        -s2<INT>        split the query file, in unit of M[1000].
                        -x <INT>        the scope to align[-1]
                        -j <INT>        the jump length to get kmer[1]
                        -n <INT>        max align number for query [20]
                        -t <INT>        thread number[4]
                        -m <INT>        align mode[1]
                                          1. align with LCS,for uncrrected reads
                                          2. align with kmer, for corrected reads
                        -c <FILT>       the config file for qsub [align.cfg]
                        -d <DIR>        the output directory [Align]
```

[0048] 具体案例实施操作如下：

[0049] 使用两个三代Pacbio测序的数据文件，分别为read1.fa,read2.fa，另外还有一个二代Illumina测序数据提取的unique-kmer文件：k17.bit,运行以下命令来进行比对：

[0050]

```
perl Align.pl -g k17.bit -k 17 -f 5 -u 3 -n 3 read1.fa read2.fa
```

[0051] (3)对候选reads进行详细比对。

[0052] a.先对比对上的seed进行聚类，算出最可能的比对范围，方法如下：

[0053] 建立坐标系，横坐标代表read1比对上的位置，纵坐标代表read2上比对上的位置，每个点代表两条read上共有的seed，将这些seed用斜率为1的直线进行聚类，将聚到最多点的直线作为比对上的区域；

[0054] b.再将比对范围进行小区域分割(可以设定分割长度为100bp)，对每一个分割区域，使用LCS算法计算相似度，再对整体进行打分，方法如下：

[0055] 假设将比对范围分为n个区域，相似度大于0.8的区域有b个，这些小区域总体的相似碱基为c个，则区域相似度为b/n，碱基相似度为c/a，最后只保留这两个值都大于0.7的数据。

[0056] 以上所述的仅是本发明的优选实施方式，应当指出，对于本技术领域中的普通技术人员来说，在不脱离本发明核心技术特征的前提下，还可以做出若干改进和润饰，这些改进和润饰也应视为本发明的保护范围。

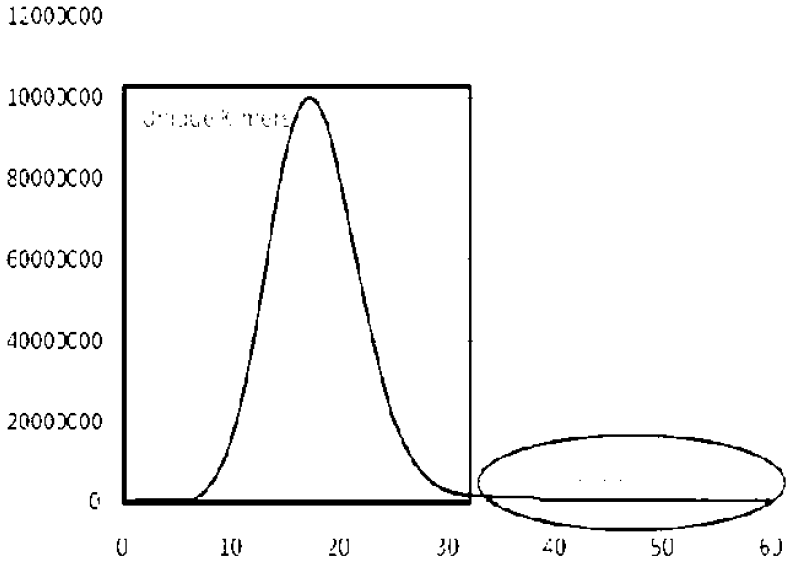

12000000
10000000
80000000
60000000
40000000
20000000
0
0
10
20
30
40
50
60

图1

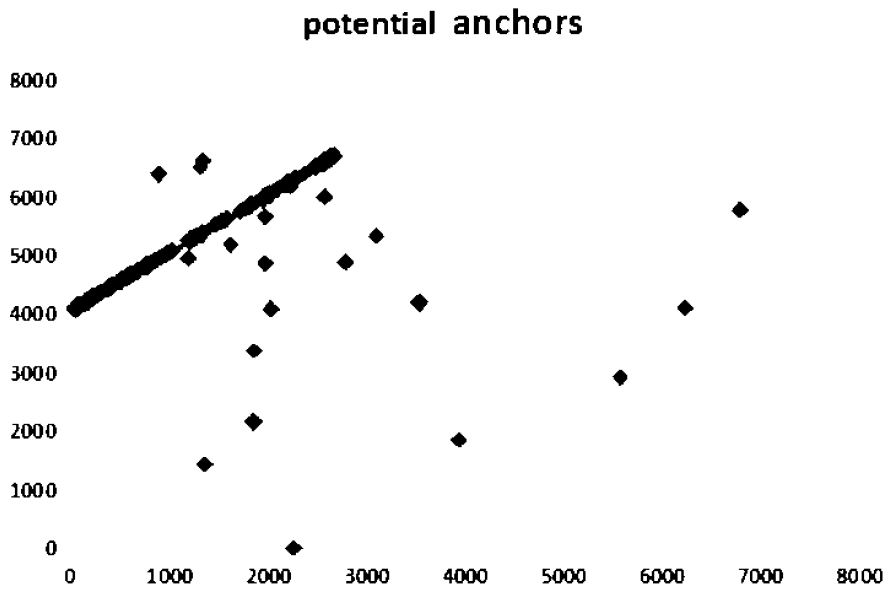

potential anchors
8000
7000
6000
5000
4000
3000
2000
1000
0
0
1000
2000
3000
4000
5000
6000
7000
8000

图2

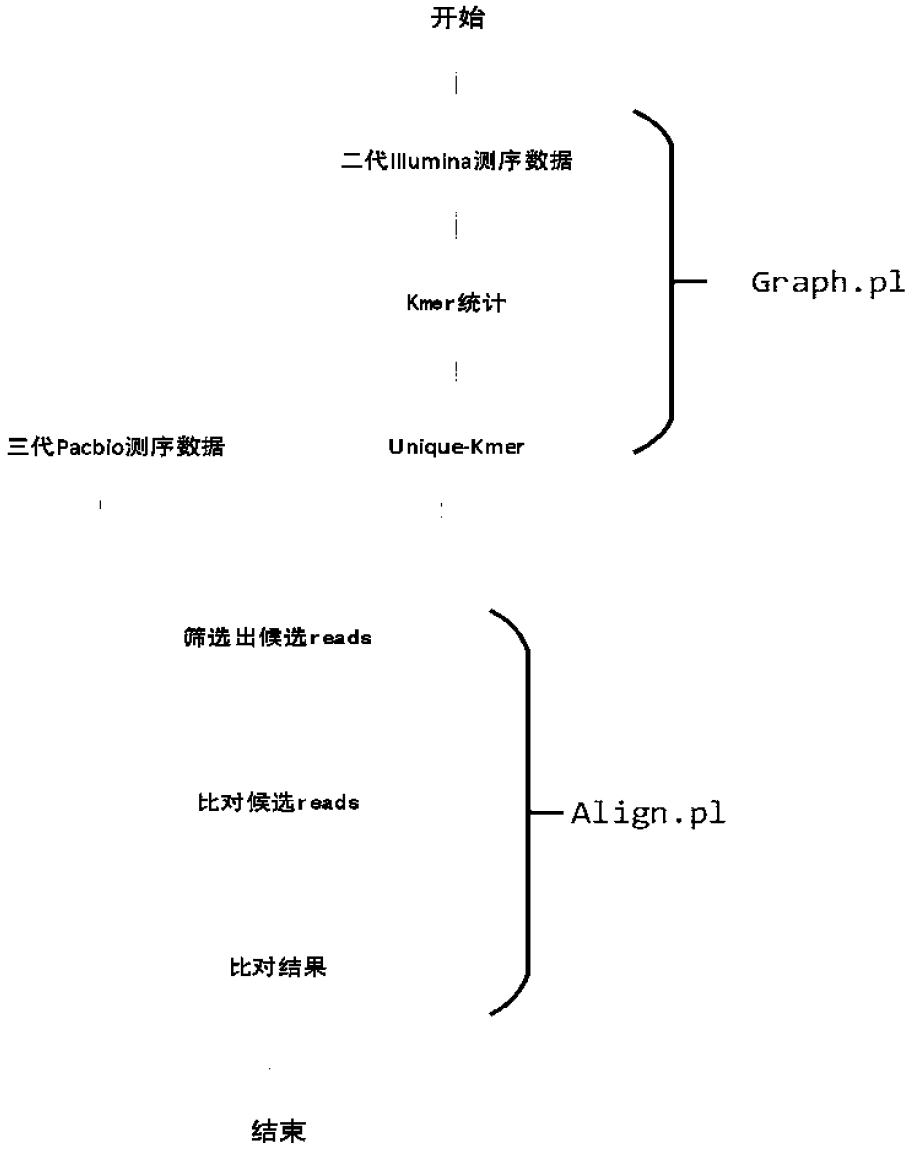

开始
二代Illumina测序数据
Kmer统计
Graph.pl
三代Pacbio测序数据
Unique-Kmer
筛选出候选reads
比对候选reads
Align.pl
比对结果
结束

图3