(19)中华人民共和国国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 107437002 A
(43)申请公布日 2017.12.05

(21)申请号 201710293318.7

(22)申请日 2017.04.28

(71)申请人 首度生物科技(苏州)有限公司
地址 215123 江苏省苏州市苏州工业园区星湖街218号生物纳米园B8楼5层
申请人 苏州首度基因科技有限责任公司

(72)发明人 闫成海 唐元华 徐健

(74)专利代理机构 北京恒泰铭睿知识产权代理有限公司 11642
代理人 胡艳

(51)Int.Cl.
G06F 19/22(2011.01)

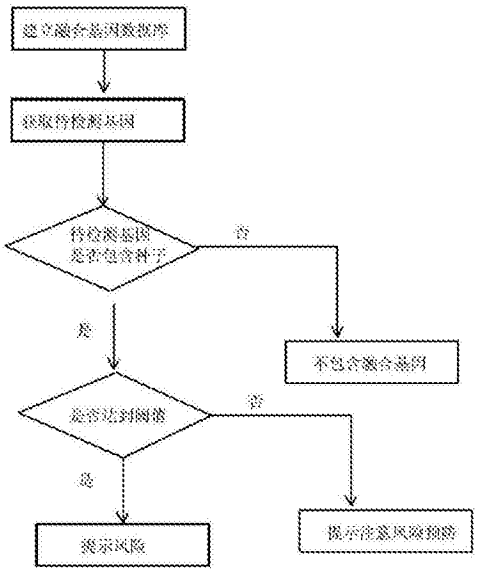
权利要求书1页 说明书4页 附图2页

(54)发明名称

一种快速检测融合基因的方法

(57)摘要

本发明提供一种快速检测融合基因的方法，包括以下步骤：A.建立融合基因数据库：将已知的融合基因断裂点以及断裂点前面的N个序列和断裂点后面的M个序列组成种子，种子的集合形成融合基因数据库；B.获取待检测基因；D.将种子与待检测基因的序列数据比对，确定待检测基因的序列数据是否包含种子信息；E.当包含待检测基因包含种子信息时，则认为待检测基因内含有融合基因；否则认为不包含融合基因。本发明弃用了常规的融合基因寻找方法，通过建立新的融合基因数据库，采用执果索因的方式，不需要与人类基因组进行比较，避免了既耗时又可能产生错误的比对基因组步骤，使得此种方法检测速度可提高几十倍，很少的内存需求下完成分析，并且防止了比较错误引起的误判。



CN 107437002 A

1.一种快速检测融合基因的方法,其特征在于,包括以下步骤:

A.建立融合基因数据库:将已知的融合基因断裂点以及断裂点前面的N个序列和断裂点后面的M个序列组成种子,种子的集合形成融合基因数据库;

B.获取待检测基因:通过基因检测装置,获取待检测基因的序列数据;

D.将种子与待检测基因的序列数据比对,确定待检测基因的序列数据是否包含种子信息;

E.当包含待检测基因包含种子信息时,则认为待检测基因内含有融合基因;否则认为不包含融合基因。

2.根据权利要求1所述的快速检测融合基因的方法,其特征在于,在所述步骤B之后还包括步骤C.将序列数据建立数据库索引。

3.根据权利要求1所述的快速检测融合基因的方法,其特征在于,在所述步骤D之后还包括步骤F.当含有融合基因时,判断待检测基因包含融合基因的含量,当融合基因的含量大于一定阈值时,提示存在风险。

4.根据权利要求1所述的快速检测融合基因的方法,其特征在于,所述步骤B中的基因检测装置为二代高通量测序平台或三代测序平台或基因芯片。

5.根据权利要求1所述的快速检测融合基因的方法,其特征在于,所述步骤A中,N≥5或M≥5。

6.根据权利要求1所述的快速检测融合基因的方法,其特征在于,所述步骤D中种子与序列数据采用局部比对。

一种快速检测融合基因的方法

技术领域

[0001] 本发明涉及生物信息技术领域，尤其是一种快速检测融合基因的方法。

背景技术

[0002] 融合基因是指两个基因的全部或者部分序列相互融合为一个全新的基因的过程，其有可能是染色体易位、中间缺失或染色体倒置所致的结果，通常具有致瘤性。1973年，芝加哥大学的Janet Rowley确认了费城染色体的形成机制来自于染色体易位，并在白血病中发现第一个融合基因。随后，在众多实体瘤如尤文肉瘤、滑膜肉瘤、前列腺癌、肺癌、乳腺癌、卵巢癌等中相继发现了融合基因的存在。据相关研究报道，90%以上的慢性粒细胞白血病(CML)会出现BCR-ABL融合基因，此基因产生一种新的mRNA，编码的蛋白为P210，P210会使细胞失去对周围环境的反应性，并抑制细胞凋亡的发生。因此，BCR-ABL融合基因也可以作为慢性粒细胞白血病的生物标记，来判别是否罹患慢性粒细胞白血病。常见的基因融合原理如图1所示，第一个基因从第二个序列断开，第二个基因从第二个序列断开，第一个基因的前段和第二个基因的后段组合形成新的基因。

[0003] 目前融合基因的检测，多是基于高通量测序技术，首先进行转录组测序，获得全部转录本的序列信息；然后将这些序列回帖到人类基因组上，寻找可以比对上不同区域上的嵌合序列，对于双端测序，可以寻找横跨某一区域的双端序列，。最后根据嵌合序列比对到的基因，确定融合基因的名称。

[0004] 然而，由于现有测序技术具有一定的错误率，加之人类基因组的复杂性，现有融合基因的检测装置和方法并不能很好的完成检测目标。现有检测方法存在以下不足：

1.检测过程对短序列比对软件的依赖较高，比对结果的好坏对检测到融合基因的有较大影响；

2.检测时间相对较长、内存消耗较大。现有检测方法一般要花费数小时或者数天来才能检测到结果，且对计算内存消耗较大，一般的计算设备较难满足要求。

[0005] 3.检测结果的假阳性较高。由于测序错误、比对错误等原因，传统的检测方法会产生较多的假阳性结果，导致分析结果需要进一步验证才能最终确定真正的融合基因。

发明内容

[0006] 为了解决上述技术问题，本发明提供了一种快速检测融合基因的方法，其不但可以快速检测出融合基因，同时可以防止软件比对错误引起的误判。

[0007] 一种快速检测融合基因的方法，包括以下步骤：

A.建立融合基因数据库：将已知的融合基因断裂点以及断裂点前面的N个序列和断裂点后面的M个序列组成种子，种子的集合形成融合基因数据库；

B.获取待检测基因：通过基因检测装置，获取待检测基因的序列数据；

D.将种子与待检测基因的序列数据比对，确定待检测基因的序列数据是否包含种子信息；

E.当包含待检测基因包含种子信息时，则认为待检测基因内含有融合基因；否则认为不包含融合基因。

[0008] 进一步地，在所述步骤B之后还包括步骤C.将序列数据建立数据库索引。

[0009] 进一步地，在所述步骤D之后还包括步骤F.当含有融合基因时，判断待检测基因包含融合基因的含量，当融合基因的含量大于一定阈值时，提示存在风险。

[0010] 进一步地，所述步骤B中的基因检测装置为二代高通量测序平台或三代测序平台或基因芯片。

[0011] 进一步地，所述步骤A中，N≥5或M≥5。

[0012] 进一步地，所述步骤D中种子与序列数据采用局部比对。

[0013] 采用上述方法，本发明具有以下的技术效果：

1.由于本发明将将已知的融合基因断裂点以及断裂点前面的N个序列和断裂点后面的M个序列组成种子，种子的集合形成融合基因数据库，并将种子与待检测基因的序列数据进行比对，弃用了常规的融合基因寻找方法，通过建立新的融合基因数据库，采用执果索因的方式，不需要与人类基因组进行比较，避免了既耗时又可能产生错误的比对基因组步骤，使得此种方法检测速度可提高几十倍，很少的内存需求下完成分析，并且防止了比较错误引起的误判。

[0014] 2.当融合基因大于阈值时，提示存在风险，通过本发明的检测融合基因的方法，可以有效检测融合基因并且提示用户注意，提前做出预防。

[0015] 3.基因检测装置为二代高通量测序平台或三代测序平台或基因芯片，通过快速测序平台，使得检测融合基因的方法的检测速度进一步提高，防止测序的时间影响整个检测时间。

[0016] 4.N≥5或M≥5时，种子与待检测基因的序列数据比对速度较快，并且可以有效保障准确率，如果N<5，则可能引起误判；同样M<5，也有可能引起误判。

[0017] 5.种子与序列数据采用局部比对，采用局部比对的方式，忽略了不相关的基因数据，不但可以提高比对的准确度，还具有较高的敏感性与特异性。

附图说明

[0018] 图1是现有技术融合基因的原理。

[0019] 图2是本发明实施例1的流程图。

[0020] 图3是本发明实施例2的流程图。

具体实施例

[0021] 下面结合本发明实施例的附图对本发明实施例的技术方案进行解释和说明，但下述实施例仅为本发明的优选实施例，并非全部。基于实施方式中的实施例，本领域技术人员在没有做出创造性劳动的前提下所获得其他实施例，都属于本发明的保护范围。

[0022] 实施例一：

如图2所示，一种快速检测融合基因的方法，包括以下步骤：

A.建立融合基因数据库：融合基因是已知的，例如可以从已经发表的文献或者数据库中获取，为了实现快速检测，本实施例根据已知的融合基因形成融合基因的嵌合序列，该嵌

合序列即称为种子，具体方式为，将已知的融合基因断裂点以及断裂点前面的N个序列和断裂点后面的M个序列组成种子，种子的集合形成融合基因数据库；在本实施例中，M=5，N=5。

[0023] B.获取待检测基因：通过illumina，获取待检测基因的序列数据；获取序列数据的方法在本领域中比较常见，在本实施中不做详细阐述。当然，本实施例的基因序列数据的获取并不限于使用illumina，也可以是454，Life Technologies等二代高通量测序平台，或者来自三代测序平台，如Pacbio等主流或者其他高通量测序平台产生的测序数据。此外，也可以是基因芯片测序产生的序列数据。

[0024] D.将种子与待检测基因的序列数据比对，确定待检测基因的序列数据是否包含种子信息。

[0025] E.当包含待检测基因包含种子信息时，则认为待检测基因内含有融合基因；否则认为不包含融合基因。

[0026] F.当含有融合基因时，判断待检测基因包含融合基因的含量，当融合基因的含量大于一定阈值时，提示存在风险；否则提示注意风险预防。

[0027] 由于本发明将将已知的融合基因断裂点以及断裂点前面的N个序列和断裂点后面的M个序列组成种子，种子的集合形成融合基因数据库，并将种子与待检测基因的序列数据进行比对，弃用了常规的融合基因寻找方法，通过建立新的融合基因数据库，采用执果索因的方式，不需要与人类基因组进行比较，避免了既耗时又可能产生错误的比对基因组步骤，使得此种方法检测速度可提高几十倍，很少的内存需求下完成分析，并且防止了比较错误引起的误判。

[0028] 可以理解，本实施例的M或N并不限于5，也可以是6、7、8等，通常来说，N≥5或M≥5时，种子与待检测基因的序列数据比对速度较快，并且可以有效保障准确率，如果N<5，则可能引起误判；同样M<5，也有可能引起误判。

[0029] 实施例二：

本实施例与实施例一的区别在于，该方法还包括建立数据库索引和比对方式不同。

[0030] 如图3所示，

一种快速检测融合基因的方法，包括以下步骤：

A.建立融合基因数据库：将已知的融合基因断裂点以及断裂点前面的N个序列和断裂点后面的M个序列组成种子，种子的集合形成融合基因数据库；

B.获取待检测基因：通过基因检测装置，获取待检测基因的序列数据；

C.将序列数据建立数据库索引。

[0031] D.通过blast软件种子与待检测基因的序列数据进行局部比对，确定待检测基因的序列数据是否包含种子信息。

[0032] E.当包含待检测基因包含种子信息时，则认为待检测基因内含有融合基因；否则认为不包含融合基因。

[0033] 通过对测序样本数据进行索引构建，将局部比对算法灵活的运用到融合基因的查找中，忽略了不想关的基因数据，使得检测速度更快，并且具有较高的精确度，对于融合基因的检测具有较高的敏感性与特异性。

[0034] 可以理解，本实施例的比较软件并不限于blast，也可以是其他以局部比对算法为核心思想开发的软件工具。

[0035] 以上所述者，仅为本发明的较佳实施例而已，并非用来限定本发明的实施范围，即凡依本发明所作的均等变化与修饰，皆为本发明权利要求范围所涵盖，这里不再一一举例。

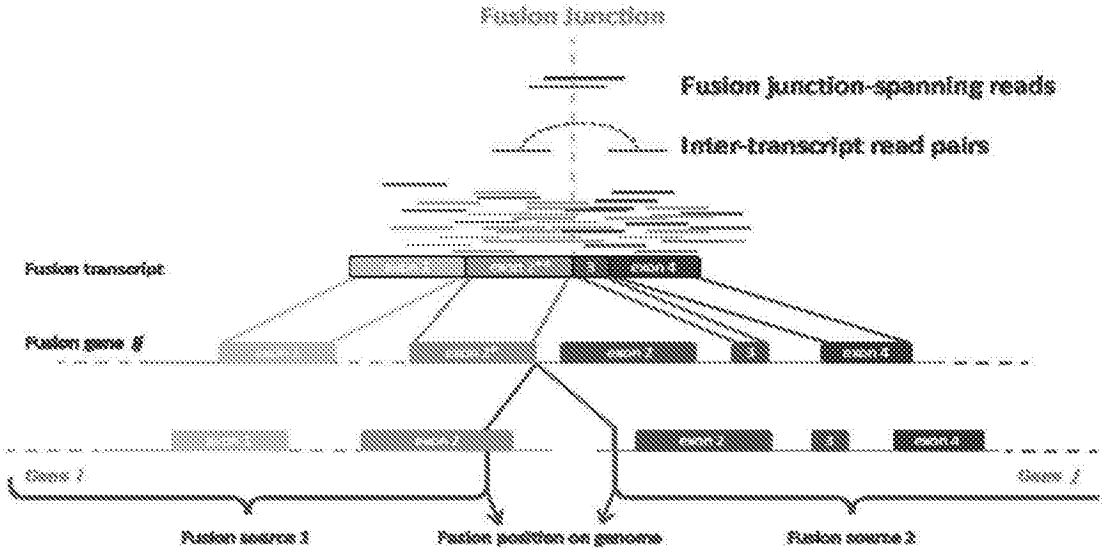

Fusion junction
Fusion junction-spanning reads
Inter-transcript read pairs
Fusion transcript
Fusion gene B
Fusion gene A
Fusion source A
Fusion position on genome
Fusion source B

图1

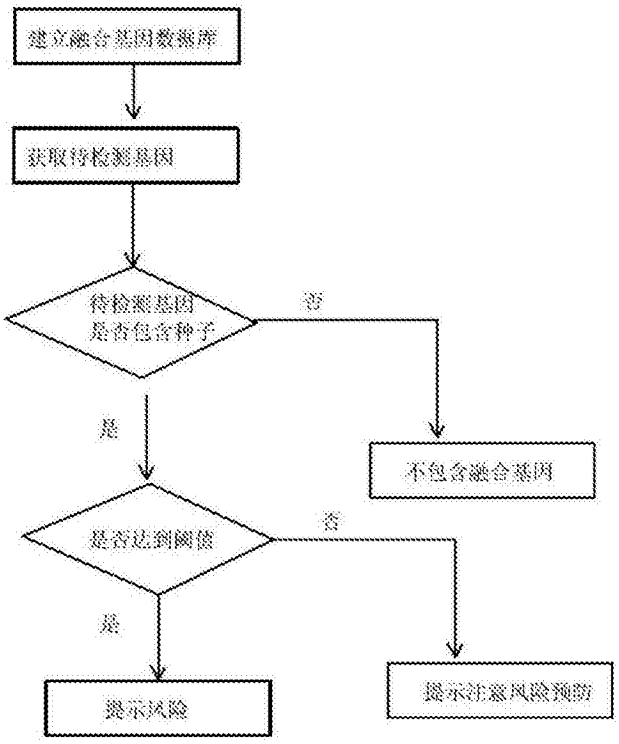

建立融合基因数据库
获取待检测基因
待检测基因
是否包含种子
否
是
不包含融合基因
是否达到阈值
否
是
提示风险
提示注意风险预防

图2

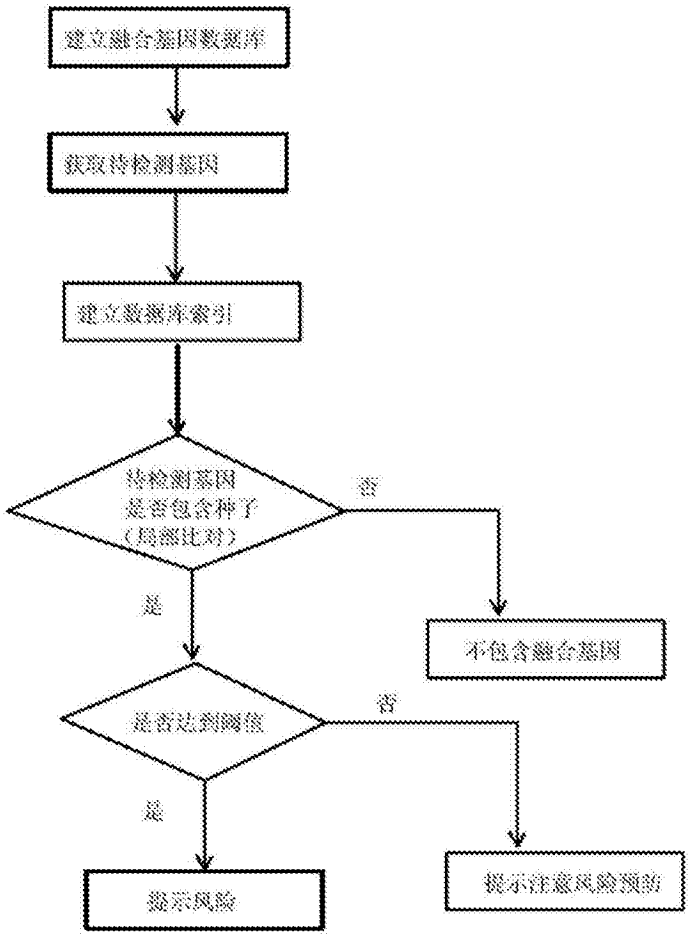

建立融合基因数据库
获取待检测基因
建立数据库索引
待检测基因
是否包含种子
（局部比对）
否
是
不包含融合基因
是否达到阈值
否
是
提示风险
提示注意风险预防

图3