(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
2012 年 3 月 22 日 (22.03.2012)

PCT

(10) 国际公布号
WO 2012/034251 A2

(51) 国际专利分类号:
C12Q 1/68 (2006.01)

(21) 国际申请号: PCT/CN2010/001409

(22) 国际申请日: 2010 年 9 月 14 日 (14.09.2010)

(25) 申请语言: 中文

(26) 公布语言: 中文

(71) 申请人 (对除美国外的所有指定国): **深圳华大基因科技有限公司 (BGI SHENZHEN CO., LIMITED)** [CN/CN]; 中国广东省深圳市盐田区北山路 146 号北山工业区综合楼 11F-3, Guangdong 518083 (CN)。

(72) 发明人; 及

(75) 发明人/申请人 (仅对美国): **罗锐邦 (LUO, Ruibang)** [CN/CN]; 中国广东省深圳市盐田区北山路 146 号北山区综合楼, Guangdong 518083 (CN)。 **邵浩靖 (SHAO, Haojing)** [CN/CN]; 中国广东省深圳市盐田区北山路 146 号北山区综合楼, Guangdong 518083 (CN)。 **林浩翔 (LIN, Haoxiang)** [CN/CN]; 中国广东省深圳市盐田区北山路 146 号北山区综合楼, Guangdong 518083 (CN)。

(74) 代理人: **中国国际贸易促进委员会专利商标事务所 (CCPIT PATENT AND TRADEMARK LAW OFFICE)**; 中国北京市西城区阜成门外大街 2 号万通新世界广场 8 层, Beijing 100037 (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, 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, IS, JP, KE, KG, KM, KN, KP, KR, KZ, LA, LC, LK, LR, LS, LT, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PE, PG, PH, PL, PT, RO, RS, RU, 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, SD, SL, SZ, TZ, UG, ZM, ZW), 欧亚 (AM, AZ, BY, KG, KZ, MD, 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, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG)。

本国际公布:

— 包括按条约第 17 条(2)(a)所作的宣布，不包括摘要，发明名称未经国际检索单位审查。

WO 2012/034251 A2

(54) **Title**: METHODS AND SYSTEMS FOR DETECTING GENOMIC STRUCTURE VARIATIONS

(54) **发明名称** : 一种基因组结构性变异检测方法和系统

(57) **Abstract**:

(57) **摘要**:

一种基因组结构性变异检测方法和系统

技术领域

本发明涉及生物信息学技术领域，尤其涉及一种基因组结构性变异（Structure Variation，SV）检测方法和系统。

背景技术

结构性变异在基因组中有重要的地位，结构性变异可能导致个体基因编码改变和功能改变。随着人类基因组计划和国际单体型图计划的顺利完成，生物学家通过遗传连锁或关联分析已经定位了大量与人类疾病相关的基因组候选区域。但是，识别这些区域中的致病基因或突变需要对这些区域进行重新测序。现有的全基因组重测序分析技术成本较高，而且通过全基因组重测序分析技术得到的信息对于部分研究和个体医疗指导来说包含大量冗余信息。为了提高获得有效信息的效率，将现有基因分析技术集中在高价值的基因研究区域对于科学研究和医疗指导具有重大意义。传统的基于 PCR（Polymerase Chain Reaction，聚合酶链反应）来对候选区域进行测序的方法由于耗时耗力已经无法满足研究者的要求，同时基于基因芯片的 SNP（Single Nucleotide Polymorphism，单核苷酸多态性）分型技术又无法找出基因组上的稀有变异。

随着新一代高通量测序技术的出现以及测序成本的降低，如 Solexa 测序技术，迫切需要一种可以对基因组上感兴趣的区域进行测序从而可以识别该区域上各种突变的技术。

发明内容

本公开的一个方面要解决的一个技术问题是提供一种基因组

结构性变异检测方法，准确性更高。

本公开的一个方面提供一种基因组结构性变异检测方法，包括:

组装步骤，将测序序列组装成骨架序列（scaffold）;

比对步骤，将骨架序列对参考基因组进行全局两两比对，获得含有变异信息的比对结果;

提取步骤，从含有变异信息的比对结果中提取变异信息。

根据本公开的一个方面，在组装步骤之前，还包括:

优化步骤，将测序序列通过比对参考基因组进行优化处理获得优化的测序序列;

组装步骤包括: 将优化的测序序列组装成骨架序列。

根据本公开的一个方面，在提取步骤之后，还包括:

验证步骤，对提取的变异信息进行验证以去除未通过验证的变异信息。

根据本公开的一个方面，验证步骤包括:

对于变异信息中长度大于等于 50bp 的变异，判断重复性是否小于 10%，如果是，则构建变异序列，将测序序列比对上变异序列，如果变异序列的深度符合逻辑理论分布，则通过验证，否则未通过验证，去除变异; 如果重复性大于等于 10%，则判断变异位点延伸序列是否无重复性，如果是，则构建变异序列，把测序序列比对上变异序列，延伸序列比对深度特征符合逻辑理论分布则通过验证，否则去除;

对于变异信息中长度小于 50bp 的变异，构建变异序列，通过短序列比对工具对测序序列和变异序列进行间隙比对，如果比对结果符合逻辑理论比对结果，则通过验证，否则未通过验证，去除变异。

根据本公开的一个方面，提取步骤还包括:

对含有变异信息的比对结果进行如下处理:

过滤或重新运行异常结果; 和/或

过滤逻辑错误结果; 和/或

去除常见结果不完整。

根据本公开的一个方面，优化步骤包括:

通过短序列比对工具将测序序列比对参考基因组获得比对序列;

优化步骤还包括:

通过短序列比对工具去除重复测序序列;

和/或

将比对上参考基因组的所有错误比对碱基置换成与参考基因组一致的碱基;

和/或

去除比对序列中平均质量低于预定值的测序序列。

根据本公开的一个方面，组装步骤包括:

将测序序列切成 N-mer 后构建德布鲁恩图;

根据德布鲁恩图输出重叠群（contig）和杂合序列;

运用测序得到的双端关系根据重叠群构建骨架序列;

对骨架序列进行补缺口得出最后的骨架序列。

通过本公开实施例的方法，对全基因组测序结果进行组装获得骨架序列，和参考基因组进行对比，得出与参考基因组无关的个人特有基因组，准确性高。

本公开的另一个方面要解决的一个技术问题是提供一种基因组结构性变异检测系统，准确性更高。

本公开的一个方面提供一种基因组结构性变异检测系统，包括:

组装装置，用于将测序序列组装成骨架序列（scaffold）;

比对装置，用于将骨架序列对参考基因组进行全局两两比对，获得含有变异信息的比对结果；

提取装置，用于从含有变异信息的比对结果中提取变异信息。

根据本公开的一个方面，该系统还包括：

优化装置，用于将测序序列通过比对参考基因组进行优化处理获得优化的测序序列；

组装装置用于将优化的测序序列组装成骨架序列。

根据本公开的一个方面，该系统还包括：

验证装置，用于对提取的变异信息进行验证，去除未通过验证的变异信息。

根据本公开的一个方面，验证装置对于变异信息中长度大于等于 **50bp** 的变异，判断重复性是否小于 **10%**，如果是，则构建变异序列，将测序序列比对上变异序列，如果变异序列的深度符合逻辑理论分布，则通过验证，否则未通过验证，去除变异；如果重复性大于等于 **10%**，则判断变异位点延伸序列是否无重复性，如果是，则构建变异序列，把测序序列比对上变异序列，延伸序列比对深度特征符合逻辑理论分布则通过验证，否则去除；对于变异信息中长度小于 **50bp** 的变异，构建变异序列，通过短序列比对工具对测序序列和变异序列进行间隙比对，如果比对结果符合逻辑理论比对结果，则通过验证，否则未通过验证，去除变异。

根据本公开的一个方面，提取装置包括：

变异信息过滤单元，用于对含有变异信息的比对结果进行过滤或重新运行异常结果；和/或过滤逻辑错误结果；和/或去除常见结果不完整，输出过滤后的比对结果；

变异信息提取单元，用于从变异信息过滤单元输出的过滤后

的比对结果提取变异信息。

根据本公开的一个方面，优化装置包括:

对比单元，用于将测序序列比对参考基因组得到比对序列;

过滤单元，用于对比对序列进行过滤，去除比对结果中平均质量低于预定值的序列;

错误碱基置换单元，用于将比对上参考基因组的所有错误比对碱基置换成与参考基因组一致的碱基。

根据本公开的一个方面，组装装置包括:

图构建单元，用于将优化的测序序列切成 N-mer 后构建德布鲁恩图;

切割单元，用于对德布鲁恩图中的环状结构进行输出，切割该德布鲁恩图变成多条重叠群（contig）和杂合序列;

骨架构建单元，用于运用测序得到的双端关系根据多条重叠群构建骨架序列，对骨架序列进行补缺口得出最后的骨架序列。

本公开基因组结构性变异检测系统的实施例，通过组装装置对全基因组测序结果进行组装获得骨架序列，通过比对装置将骨架序列和参考基因组进行全局对比，得出与参考基因组无关的个人特有基因组，准确性高。

附图说明

图 1 示出本发明的基因组结构性变异检测方法的一个实施例的流程图;

图 2 示出本发明的基因组结构性变异检测方法的另一个实施例的流程图;

图 3 示出本发明的基因组结构性变异检测方法的又一个实施例的流程图;

图 4 示出本发明的基因组结构性变异检测系统的一个实施例

的结构图；

图 5 示出本发明的基因组结构性变异检测系统的另一个实施例的结构图；

图 6 示出本发明的基因组结构性变异检测系统的又一个实施例的结构图。

具体实施方式

下面参照附图对本发明进行更全面的描述，其中说明本发明的示例性实施例。

基于组装检测结构性变异的方法和系统是一种对基因组 DNA 序列信息进行一系列生物信息分析的方法和进行相关分析的工具，旨在解决基因组生物信息学分析方法和工具不完善的问题。

图 1 示出本发明的基因组结构性变异检测方法的一个实施例的流程图。

步骤 102，组装步骤。将测序序列组装成骨架序列（scaffold）。例如，把测序序列切成 N-mer 后构建德布鲁恩图，对德布鲁恩图中的部分环状结构进行输出，同时切割该德布鲁恩图变成多条重叠群（contig），和杂合序列；运用测序得到的双端关系对重叠群进行处理构建骨架序列。通过处理带缺口的骨架序列，对骨架序列用碱基“N”进行补缺口，得到最后的骨架序列。

步骤 104，比对步骤。将骨架序列对参考基因组进行全局两两比对，获得含有变异信息的比对结果。例如，对步骤 102 得出的组装结果使用长序列比对软件与参考基因组进行全局两两比对。长序列比对软件例如是 LASTZ，具体介绍可以见参考文献【Harris, R.S. Improved pairwise alignment of genomic DNA. PhD thesis, Pennsylvania State University （2007）】。

步骤 106，提取步骤，从含有变异信息的比对结果中提取变异信息。变异信息包括变异位点的位置，变异类型，变异的序列等信息。

在本发明的上述实施例中，对全基因组测序结果进行组装获得骨架序列，和参考基因组进行对比，得出与参考基因组无关的个人特有基因组，准确性高。

图 2 示出本发明的基因组结构性变异检测方法的另一个实施例的流程图。

如图 2 所示，步骤 202，优化步骤。将测序序列通过比对参考基因组后进行优化处理获得优化的测序序列。通过序列比对工具进行测序序列和参考基因组的比对获得比对序列，将比对序列进行优化处理，例如去重复、替换错误碱基和过滤后，转换成优化的测序序列。

例如，通过 BWA 软件进行测序序列和参考基因组的比对，BWA 具体参数采用“aln –e 0 –o 0”。该参数的含义为：“aln”是 BWA 的子功能，作用是比对；“-e”表示能进行间隙比对（gapped alignment）的间隙长度上限；“-o”表示间隙比对的间隙个数。BWA 是短序列对比软件，具体介绍可以参见参考文献【Heng Li, Richard Durbin. Fast and accurate short read alignment with Burrows-Wheeler transform. Nature Bioinformatics. Vol.25 no.14: 1754-1760（2009）】。

对比序列的去重复处理是指去除一些重复度高的序列区域。例如，一个序列区域为 ATCATCATCATCATC，包含多个 ATC，将会对比对造成影响，应当排除这样的序列区域。比对序列的替换错误碱基处理为把比对上参考基因组的所有错误比对碱基置换成跟参考基因组一致的碱基。比对序列的过滤处理为去除平均质量值低于预定值 X 的序列；例如，参数 X 根据测序的平

均质量值设定，质量值符合公式 Q=-10*lgPe，Pe 为出错概率，建议取值范围例如是[10-20]，对应平均错误率为[10%-1%]，默认选项是 15。通过对测序序列进行优化处理，可以提高下一步处理的精度。

步骤 204，组装步骤。将优化的测序序列组装成骨架序列。例如，采用华大基因研究院研发的软件 Soapdenovo 进行组装，具体组装参数是“-K 31”，其中，参数“－K”用于设定切 K-mer 的值。其中 Soapdenovo 软件的介绍可以参见参考文献:【Li, R. et al. De novo assembly of human genomes with massively parallel short read sequencing. Genome Res （2009）】。

步骤 206，比对步骤。将骨架序列与参考基因组进行全局两两比对，得出含有变异信息的比对结果。例如，用 LASTZ 把骨架序列对参考基因组进行全局两两比对，其中 LASTZ 的具体参数为：“ --strand=both --chain --ambiguousn –gapped --ydrop=20000 --gap=1000,1 --noentropy --format=axt”，对参考基因组建种子采用 12of19。参数定义可见 LASTZ 软件说明文档，“--strand=both”是指正负链都比对，“—chain”是指进行链接，“—ambiguousN”是指把 N 作为多种碱基型处理，“—gapped”是指进行间隙比对，“—ydrop=20000”是指间隙比对罚分的阀值为 20000，“--gap=1000,1”是指开一个间隙罚分 1000，延长一个间隙一个碱基罚分 1 分，“—noentropy”是指不引入熵对高精度结果进行过滤，“–format=axt”是指结果用 axt 格式输出。“12of19” 是种子的模式为 12of19。一个种子为参考序列中按软件设定规则选取的 19 个碱基长度的序列。目标序列能否比对上种子序列只考虑软件设定的种子中的 12 个碱基位置。如果种子区域比对上，比对将以种子区域为起点向两个方向延伸，直到比对完成，输出比对结果。

步骤 208，提取步骤。对包含变异信息的比对结果进行过滤，提取过滤后的比对结果中的变异信息。过滤包括:（1）过滤或重新运行异常结果，（2）过滤逻辑错误结果，和（3）常见结果不完整。（1）过滤或重新运行异常结果: 过滤 lastz 中运行不正常的结果，过滤 lastz 结果中注释的无意义部分，重新运行没有正常结尾标识符的 lastz 程序。（2）过滤逻辑错误结果: 这个包括一条组装序列比对上两条或以上染色体，一条染色体的同一个位置比对上两条或者以上的组装序列，从这些结果中挑选质量较好的保留之。（3）比对结果中常常有 N （ACGT 均有可能）与 -（比对间隙）成对出现，同时出现，应该去除这样的对。

步骤 210，验证步骤。对提取的变异信息进行验证以去除未通过验证的变异信息。可以通过各种计算方法对候选的变异信息进行验证去除未通过验证的变异信息。例如，通过深度和序列切割方法进行验证。对于长度大于等于 50bp 的变异，首先构建变异序列，然后把测序序列比对上变异序列，若变异序列的深度符合逻辑理论分布则通过验证，否则去除；对于长度少于 50bp 的变异，首先构建变异序列，然后用序列比对软件例如 BWA 对测序短序列和变异序列进行带间隙比对，比对参数为 “-e 50 –o 1 –i 5”，若比对结果符合逻辑理论比对结果则通过验证，否则去除。最后合并两者得出最终结果。上述变异序列的深度符合逻辑理论分布是指，如果目标序列跟参考序列一致，应该在该区域的各个点的深度会有比较高的值，而且每个点的深度都比较接近，反之，则值比较低。

需要指出，优化步骤和验证步骤作为本发明实施例的可选步骤，可以包含其中一个或者两个。

在上述实施例中，通过对测序序列进行优化处理，可以提高下一步处理的精度。对全基因组候选结构性变异集合进行多种方

法进行验证，去除未通过验证的变异信息，从而使得变异信息的假阳性低。通过实验表明，本发明实施例的方法可以得出假阳性10%以下的结构性变异集合。

图 3 示出本发明的基因组结构性变异检测方法的又一个实施例的流程图。

如图 3 所示，在步骤 301，BWA 比对。通过 BWA 软件进行测序序列和参考基因组的比对，获得对比序列。

在步骤 302，BWA 去重复。通过 BWA 软件去除重复度高的序列。

在步骤 303，把错误比对碱基置换成参考序列碱基和根据质量值过滤。把比对上参考基因组的所有错误比对碱基置换成跟参考基因组一致的碱基，去除平均质量值低于预定值 X 的序列。

在步骤 304，生成拼接的德布鲁恩图。

在步骤 305，根据德布鲁恩图输出重叠群和杂合序列。

在步骤 306，获得重叠群或杂合序列。后续步骤 307 至步骤 309 分别对重叠群和杂合序列进行处理。

在步骤 307，切分参考序列和拼接结果序列，该处结果序列指重叠群和杂合序列。

在步骤 308，拆分成多份的两两比对。将参考序列和结果序列拆分成多份，然后分别用一个来自参考序列的拆分过的小序列跟来自结果序列的小序列比对，直到所有小序列比对完。

在步骤 309，纠比对错误，去逻辑错误，输出变异信息。

在步骤 310，获取变异信息。

在步骤 311，判断变异长度是否大于等于 50bp 碱基对，如果是，则继续步骤 312，否则，继续步骤 317

在步骤 312，计算序列重复性。比较该序列的某个区域的信息与重复序列库中的信息，判断是否一致；若一致就判断该序列

区域为重复序列区域。也可能整条序列都为重复序列区域。通过计算重复序列区域的长度跟整条序列的比例，就能算出序列重复性。

在步骤 313，判断重复性是否少于 10%，如果是，则继续步骤 316，否则，继续步骤 314。

在步骤 314，判断变异位点延伸序列是否无重复性，如果是，则根继续步骤 315。

在步骤 315，得出变异序列，与参考序列进行比对。根据延伸序列比对深度特征得出验证序列，输出变异结果。

在步骤 316，得出变异序列，与参考序列进行比对。如果变异序列正确，该变异序列的比对深度会比较高，且比较平均。根据深度比得出验证变异，输出变异结果。

在步骤 317，得出变异序列。

在步骤 318，获得带间隙的单端或双端 BWA 比对结果。测序序列分两种，一种是单端（single-end），一种是双端（pair-end），BWA 比对的时候不同种类使用的方法不一样。具体可以参见：http://bio-bwa.sourceforge.net/bwa.shtml。

在步骤 319，提取变异位点附近序列。每个变异位点会有位置信息，在参考序列中找到这个位置，把这个位置的前后一定长度的序列截取下来跟这个变异位点的变异序列连接起来，变成一个新的序列。

在步骤 320，带间隙的 BWA 比对。 BWA 比对时使用-o 1 参数，允许目标序列与参考序列比对时存在间隙，或不存在间隙。

在步骤 322，根据比对结果的间隙情况和深度分布得出验证变异，输出变异结果。

下面介绍本发明方法的多个应用例。

应用例一，人类外显子捕捉测序。

以国际人类基因组单体型图计划个体 NA12156 外显子测序为例（样品号：NA12156；下载地址 ftp://ftp.ncbi.nlm.nih.gov/sra/static/SRX005/SRX005923）。原始数据，共 11346285 条短序列。

将人类外显子 NA12156 的测序结果用基础软件 BWA 工具和过滤程序软件对测序结果基于参考基因组进行过滤和优化；将过滤优化得出的序列用 soapdenovo 的进行组装；将组装结果使用基础软件 LASTZ 软件与参考基因组进行两两比对，比对结果用提取结构性变异信息软件进行过滤及去除异常结果，最后采用验证结构性变异软件通过深度和序列切割方法进行验证。对于长度大于等于 50bp 的变异，判断重复性是否少于 10%，如果是，则构建变异序列，把测序序列比对上变异序列，若变异序列的深度符合逻辑理论分布，则通过验证，否则去除；如果重复性大于等于 10%，则判断变异位点延伸序列是否无重复性，如果是，构建变异序列，然后把测序序列比对上变异序列，延伸序列比对深度特征符合逻辑理论分布则通过验证，否则去除；对于长度少于 50bp 的变异，构建变异序列，然后用 BWA 进行带间隙比对，比对参数为-e 50 -o 1 -i 5，若比对结果符合逻辑理论比对结果则通过验证，否则去除。最后合并两者得出最终结果。具体步骤如下：

第一步，优化步骤

对短序列进行优化处理（对比、去重复、替换、过滤）后，得到 9303954 条短序列。

第二步，组装步骤

对优化的短序列进行组装，组装结果基因组大小为 218030396bp，有 3941732 条组装序列，组装序列最长为

9042bp，N50 为 298bp 和 N90 为 122bp。

第三步，比对步骤

组装序列与参考基因组比对结果含有 64696911 对比对结果。

第四步，提取步骤

候选 SV 结果有 37014 个，大于 50bp 有 5695，少于 50bp 有 31253 个。

第五步，验证步骤

被验证的基因组变异结果有 3294 个，其中在捕捉区域的有 425 个。其中前 9 个 SV 如下表 1 所示：

组装序列 ID	变异类型	组装序列 ID 起始位置	组装序列 ID 终止位置	参考基因组染色体号	染色体开始	染色体终止	变异基因型
471724	Deletion	128	128	chr15	76846380	76846381	G
554286	Deletion	31	31	chr5	121358248	121358249	A
557038	Deletion	29	29	chr10	50393704	50393706	AA
573910	Deletion	183	183	chr2	79990300	79990303	AGC
574886	Deletion	120	120	chr6	31431989	31431990	C
576104	Insertion	697	700	chr2	73528736	73528736	CTC
596944	Insertion	178	179	chr8	145987638	145987638	T
662384	Deletion	40	40	chr8	55126183	55126184	T
729432	Insertion	117	118	chr17	6291667	6291667	C

表 1

应用例二，人类外显子捕捉测序。

该应用例以结肠癌癌变细胞的外显子测序为例（样品号：

yvf090508）。原始数据共 105972839 条短序列（测序序列）。

第一步，优化步骤

对短序列进行优化处理（对比、去重复、替换、过滤）后，共 69549590 条短序列。

第二步，组装步骤

对优化的短序列进行组装，组装结果基因组大小为 118938172bp，有 253868 条组装序列，组装序列最长为 16885bp，N50 为 793bp 和 N90 为 170bp。

第三步，比对步骤

组装序列与参考基因组比对结果含有 11882543 对比对结果。

第四步，提取步骤

候选 SV 结果有 57433 个，大于 50bp 有 12056，少于 50bp 有 45377 个。

第五步，验证步骤

被验证的 SV 有 9377 个，其中在捕捉区域的有 91 个，其中前 13 个 SV 如下表 2 所示：

组装序列 ID	变异类型	组装序列 ID 起始位置	组装序列 ID 终止位置	参考基因组染色体号	染色体开始	染色体终止	变异基因型
1811143	Insertion	36	38	chr7	65479979	65479979	AT
1833167	Insertion	261	264	chr3	126129396	126129396	AAG
1837575	Deletion	142	142	chr17	71800160	71800163	TGA
1848441	Insertion	17	20	chr3	46476289	46476289	CTT
1850771	Insertion	338	341	chr21	46546414	46546414	TGG

1852777	Deletion	343	343	chr7	15692332	15692335	TGG
1874031	Insertion	83	86	chr16	88444381	88444381	GAG
1874671	Deletion	410	410	chr7	143288092	143288093	G
1881421	Insertion	368	371	chr17	15284250	15284250	CTT
1883581	Deletion	215	215	chr6	160480888	160480896	TGGTAAGT
1887101	Deletion	146	146	chr19	1776928	1776931	CTC
1891753	Deletion	139	139	chr1	52078652	52078655	TCT
1896823	Deletion	363	363	chr19	59367581	59367584	CCC

表 2

应用例三，微生物测序。

该应用例以一株副溶血弧菌为例（样本号：VIBydvD10poolingIAAPEI-9-1）。原始数据共 5631982 条短序列。

第一步，优化步骤

对短序列进行优化处理（对比、去重复、替换、过滤）后，共 5213412 条短序列。

第二步，组装步骤

组装结果基因组大小为 5056512 bp，有 684 条组装序列，组装序列最长为 94989bp，N50 为 23988 bp 和 N90 为 5603bp。

第三步，比对步骤

组装序列与参考基因组比对结果含有 1442 对比对结果。

第四步，提取步骤

候选 SV 结果有 725 个，大于 50bp 有 196，少于 50bp 有 529 个。

第五步，验证步骤

被验证的 SV 有 180 个，其中前 19 个如表 3 所示：

组装序列 ID	变异类型	组装序列 ID 起始位置	组装序列 ID 终止位置	参考基因组染色体号	染色体开始	染色体终止	变异基因型
S001_4988	Deletion	623	623	Vibrio_parahaemolyticus_RIMD_1	281201	281202	T
S001_4998	Deletion	164	164	Vibrio_parahaemolyticus_RIMD_1	1336020	1336024	ATGT
S001_5000	Insertion	231	234	Vibrio_parahaemolyticus_RIMD_1	949303	949303	AAA
S001_5030	Deletion	536	536	Vibrio_parahaemolyticus_RIMD_1	1090795	1090796	T
S001_5176	Insertion	2626	2627	Vibrio_parahaemolyticus_RIMD_1	1499322	1499322	C
S001_5188	Deletion	2335	2335	Vibrio_parahaemolyticus_RIMD_1	723095	723096	A
S001_5240	Deletion	98	98	Vibrio_parahaemolyticus_RIMD_1	680139	680140	A
S001_5260	Deletion	1853	1853	Vibrio_parahaemolyticus_RIMD_1	676815	676816	A
S001_5348	Insertion	855	856	Vibrio_parahaemolyticus_RIMD_1	1335062	1335062	G
S001_5360	Insertion	5675	5676	Vibrio_parahaemolyticus_RIMD_1	341442	341442	T

S001_5364	Deletion	35	35	Vibrio_parahaemolyticus_RIMD_1	962113	962114	T
S001_5384	Insertion	5462	5463	Vibrio_parahaemolyticus_RIMD_1	312520	312520	A
S001_5388	Deletion	6105	6105	Vibrio_parahaemolyticus_RIMD_1	667732	667733	A
S001_5398	Deletion	6585	6585	Vibrio_parahaemolyticus_RIMD_1	996693	996694	A
S001_5408	Deletion	1482	1482	Vibrio_parahaemolyticus_RIMD_1	128682	128683	T
S001_5426	Insertion	5406	5407	Vibrio_parahaemolyticus_RIMD_1	71294	71294	T
S001_5436	Deletion	8239	8239	Vibrio_parahaemolyticus_RIMD_1	50680	50684	ACAT
S001_5436	Deletion	8185	8185	Vibrio_parahaemolyticus_RIMD_1	50738	50739	A
S001_5436	Deletion	49	49	Vibrio_parahaemolyticus_RIMD_1	58875	58877	AT

表 3

图 4 示出本发明的基因组结构性变异检测系统的一个实施例的结构图。如图 4 所示，该实施例的结构性变异检测系统 400 包括组装装置 41、比对装置 42 和提取装置 43。其中，组装装置 41 将测序序列组装成骨架序列（scaffold），输出骨架序列；比对装置 42 将组装装置 41 输出的骨架序列对参考基因组进行全局两两比对获得含有变异信息的比对结果；提取装置 43 从含有变异信息的比对结果中提取变异信息。

在上述实施例中，通过组装装置对全基因组测序结果进行组装获得骨架序列，通过比对装置将骨架序列和参考基因组进行全局对比，得出与参考基因组无关的个人特有基因组，准确性高。

图 5 示出本发明的基因组结构性变异检测系统的另一个实施例的结构图。和图 4 相比，该实施例的结构性变异检测系统 400 还可选地包括优化装置 50 和验证装置 54。优化装置 50 将测序序列通过比对参考基因组进行优化处理获得优化的测序序列，将优化的测序序列发送给组装装置 41。组装装置 41 将优化的测序序列组装成骨架序列（scaffold）。例如，优化装置 50 通过短序列比对软件将测序序列和参考基因组进行比对，获得比对序列，然后对比对序列进行去重复、替换、过滤等优化处理，获得优化的测序序列。

验证装置 54 对提取的变异信息进行验证，去除未通过验证的变异信息。验证装置 54 可以通过各种计算方法对候选的变异信息进行验证去除未通过验证的变异信息，例如，通过深度和序列切割方法进行验证。根据本发明的一个实施例，验证装置对于变异信息中长度大于等于 50bp 的变异，判断重复性是否小于 10%，如果是，则构建变异序列，将测序序列比对上变异序列，如果变异序列的深度符合逻辑理论分布，则通过验证，否则未通过验证，去除变异；如果重复性大于等于 10%，则判断变异位点延伸序列是否无重复性，如果是，则构建变异序列，把测序序列比对上变异序列，延伸序列比对深度特征符合逻辑理论分布则通过验证，否则去除；对于变异信息中长度小于 50bp 的变异，构建变异序列，通过短序列比对工具对测序序列和变异序列进行间隙比对，如果比对结果符合逻辑理论比对结果，则通过验证，否则未通过验证，去除变异。

在上述实施例中，通过优化装置对测序序列进行优化处理，

可以提高下一步处理的精度。通过验证装置对全基因组候选结构性变异集合进行多种方法进行验证，去除未通过验证的变异信息，从而使得变异信息的假阳性低。通过实验表明，本发明实施例的方法可以得出假阳性10%以下的结构性变异集合。

图 6 示出本发明的基因组结构性变异检测系统的又一个实施例的结构图。如图 6 所示，在该实施例的结构性变异检测系统600 中，优化装置 50 包括对比单元 501、过滤单元 502 和错误碱基置换单元 503。组装装置 41 包括图构建单元 411、切割单元 412 和骨架构建单元 413。提取装置 43 包括变异信息过滤单元 431 和变异信息提取单元 432。

其中，对比单元 501 将测序序列比对参考基因组得到比对序列；过滤单元 502 用于对比对序列进行过滤，去除比对队列中平均质量低于预定值的序列；错误碱基置换单元 503 将比对上参考基因组的所有错误比对碱基置换成与参考基因组一致的碱基。图构建单元 411 将优化的测序序列切成 N-mer 后构建德布鲁恩图；切割单元 412 对德布鲁恩图中的部分环状结构进行输出，切割该德布鲁恩图变成多条重叠群（contig）；骨架构建单元 413 运用测序得到的双端关系构建骨架序列，对骨架序列进行补缺口得出最后的骨架序列。变异信息过滤单元 431 对含有变异信息的比对结果进行过滤或重新运行异常结果；和/或过滤逻辑错误结果；和/或去除常见结果不完整，输出过滤后的比对结果；变异信息提取单元 432 从变异信息过滤单元输出的过滤后的比对结果提取变异信息。

对于图 4 至图 6 中各个装置或单元的功能，可以参考上文中关于本发明方法的实施例中对应部分的说明，为简洁起见，在此不再详述。

本领域的技术人员应当理解，对于图 4 至图 6 中的各个装

置，可以通过单独的计算处理设备实现，或者将其集成为一个独立的设备实现。在图 4 至图 6 中用框示出以说明它们的功能。这些功能块可以用硬件、软件、固件、中间件、微代码、硬件描述语音或者它们的任意组合来实现。举例来说，一个或者两个功能块都可以利用运行在微处理器、数字信号处理器（DSP）或任何其他适当计算设备上的代码实现。代码可以表示过程、功能、子程序、程序、例行程序、子例行程序、模块或者指令、数据结构或程序语句的任意组合。代码可以位于计算机可读介质中。计算机可读介质可以包括一个或者多个存储设备，例如，包括 RAM 存储器、闪存存储器、ROM 存储器、EPROM 存储器、EEPROM 存储器、寄存器、硬盘、移动硬盘、CD-ROM 或本领域公知的其他任何形式的存储介质。计算机可读介质还可以包括编码数据信号的载波。

本领域技术人员将意识到硬件、固件和软件配置在这些情况下的可替换性，以及如何最好地实现每个特定应用地该功能。

在本发明的上述实施例中，对全基因组测序结果进行组装获得骨架序列，和参考基因组进行对比，得出与参考基因组无关的个人特有基因组，准确性高。实验数据表明，本发明实施例的方法在基因组大小为 1M-3G 之间均可表现出极佳的准确性。此外，通过对全基因组测序组装结果进行分析得出候选结构性变异集合，使得结果更加全面。该候选结构性变异集合，可以进行下一步分析。本发明对全基因组候选结构性变异集合进行多种其他方法进行验证，得出假阳性 10%以下的结构性变异集合，阳性低。

本发明的描述是为了示例和描述起见而给出的，而并不是无遗漏的或者将本发明限于所公开的形式。很多修改和变化对于本领域的普通技术人员而言是显然的。选择和描述实施例是为了更

好说明本发明的原理和实际应用，并且使本领域的普通技术人员能够理解本发明从而设计适于特定用途的带有各种修改的各种实施例。

权 利 要 求

1. 一种基因组结构性变异检测方法，其特征在于，包括:

组装步骤，将测序序列组装成骨架序列（scaffold）;

比对步骤，将所述骨架序列对参考基因组进行全局两两比对，获得含有变异信息的比对结果;

提取步骤，从所述含有变异信息的比对结果中提取变异信息。

2. 根据权利要求 1 所述的基因组结构性变异检测方法，其特征在于，在所述组装步骤之前，还包括:

优化步骤，将测序序列通过比对参考基因组进行优化处理获得优化的测序序列;

所述组装步骤包括:

将所述优化的测序序列组装成骨架序列。

3. 根据权利要求 1 所述的基因组结构性变异检测方法，其特征在于，在所述提取步骤之后，还包括:

验证步骤，对所述提取的变异信息进行验证以去除未通过验证的变异信息。

4. 根据权利要求 3 所述的基因组结构性变异检测方法，其特征在于，所述验证步骤包括:

对于所述变异信息中长度大于等于 50bp 的变异，判断重复性是否小于 10%，如果是，则构建变异序列，将所述测序序列比对上所述变异序列，如果所述变异序列的深度符合逻辑理论分布，则通过验证，否则未通过验证，去除所述变异；如果重复性大于等于 10%，则判断变异位点延伸序列是否无重复性，如果是，则构建变异序列，把所述测序序列比对上所述变异序列，延伸序列比对深度特征符合逻辑理论分布则通过验证，否则去除;

对于所述变异信息中长度小于 50bp 的变异，构建变异序列，通

过短序列比对工具对所述测序序列和所述变异序列进行间隙比对，如果比对结果符合逻辑理论比对结果，则通过验证，否则未通过验证，去除所述变异。

5. 根据权利要求 1 所述的基因组结构性变异检测方法，其特征在于，所述提取步骤还包括:

对所述含有变异信息的比对结果进行如下处理:

过滤或重新运行异常结果；和/或

过滤逻辑错误结果；和/或

去除常见结果不完整。

6. 根据权利要求 2 所述的基因组结构性变异检测方法，其特征在于， 所述优化步骤包括:

通过短序列比对工具将测序序列比对参考基因组获得比对序列；

所述优化步骤还包括:

通过短序列比对工具去除重复测序序列；

和/或

将比对上参考基因组的所有错误比对碱基置换成与参考基因组一致的碱基；

和/或

去除所述比对序列中平均质量低于预定值的测序序列。

7. 根据权利要求 1 所述的基因组结构性变异检测方法，其特征在于，将所述组装步骤包括:

将所述测序序列切成 N-mer 后构建德布鲁恩图；

根据所述德布鲁恩图输出重叠群（contig）和杂合序列；

运用测序得到的双端关系根据重叠群构建骨架序列；

对骨架序列进行补缺口得出最后的骨架序列。

8. 一种基因组结构性变异检测系统，其特征在于，包括:

组装装置，用于将测序序列组装成骨架序列（scaffold）；

比对装置，用于将所述骨架序列对参考基因组进行全局两两比对，获得含有变异信息的比对结果；

提取装置，用于从所述含有变异信息的比对结果中提取变异信息。

9. 根据权利要求 8 所述的基因组结构性变异检测系统，其特征在于，还包括：

优化装置，用于将测序序列通过比对参考基因组进行优化处理获得优化的测序序列；

所述组装装置用于将所述优化的测序序列组装成骨架序列。

10. 根据权利要求 8 所述的基因组结构性变异检测系统，其特征在于，还包括：

验证装置，用于对所述提取的变异信息进行验证，去除未通过验证的变异信息。

11. 根据权利要求 10 所述的基因组结构性变异检测系统，其特征在于，所述验证装置对于所述变异信息中长度大于等于 50bp 的变异，判断重复性是否小于 10%，如果是，则构建变异序列，将所述测序序列比对上所述变异序列，如果所述变异序列的深度符合逻辑理论分布，则通过验证，否则未通过验证，去除所述变异；如果重复性大于等于 10%，则判断变异位点延伸序列是否无重复性，如果是，则构建变异序列，把所述测序序列比对上所述变异序列，延伸序列比对深度特征符合逻辑理论分布则通过验证，否则去除；对于所述变异信息中长度小于 50bp 的变异，构建变异序列，通过短序列比对工具对所述测序序列和所述变异序列进行间隙比对，如果比对结果符合逻辑理论比对结果，则通过验证，否则未通过验证，去除所述变异。

12. 根据权利要求 8 所述的基因组结构性变异检测系统，其特征在于，所述提取装置包括：

变异信息过滤单元，用于对所述含有变异信息的比对结果进行过滤或重新运行异常结果；和/或过滤逻辑错误结果；和/或去除常见结果不完整，输出过滤后的比对结果；

变异信息提取单元，用于从所述变异信息过滤单元输出的过滤后的比对结果提取变异信息。

13. 根据权利要求 9 所述的基因组结构性变异检测系统，其特征在于，所述优化装置包括:

对比单元，用于将所述测序序列比对参考基因组得到比对序列；

过滤单元，用于对所述比对序列进行过滤，去除所述比对结果中平均质量低于预定值的序列；

错误碱基置换单元，用于将比对上参考基因组的所有错误比对碱基置换成与参考基因组一致的碱基。

14. 根据权利要求 8 所述的基因组结构性变异检测系统，其特征在于，所述组装装置包括:

图构建单元，用于将所述优化的测序序列切成 **N-mer** 后构建德布鲁恩图；

切割单元，用于对所述德布鲁恩图中的环状结构进行输出，切割该德布鲁恩图变成多条重叠群（**contig**）和杂合序列；

骨架构建单元，用于运用测序得到的双端关系根据多条重叠群构建骨架序列，对骨架序列进行补缺口得出最后的骨架序列。

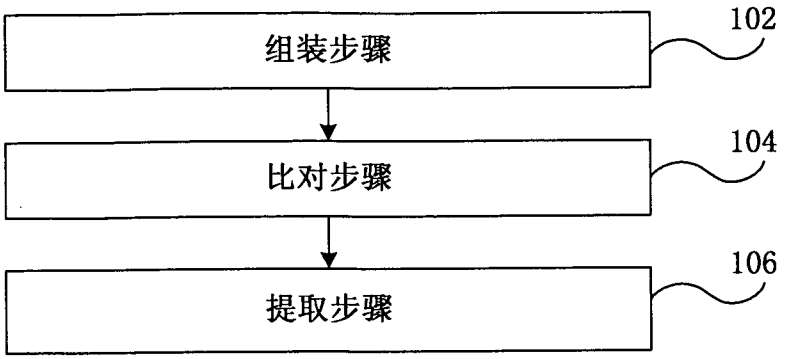
组装步骤
102
比对步骤
104
提取步骤
106

FIG.1

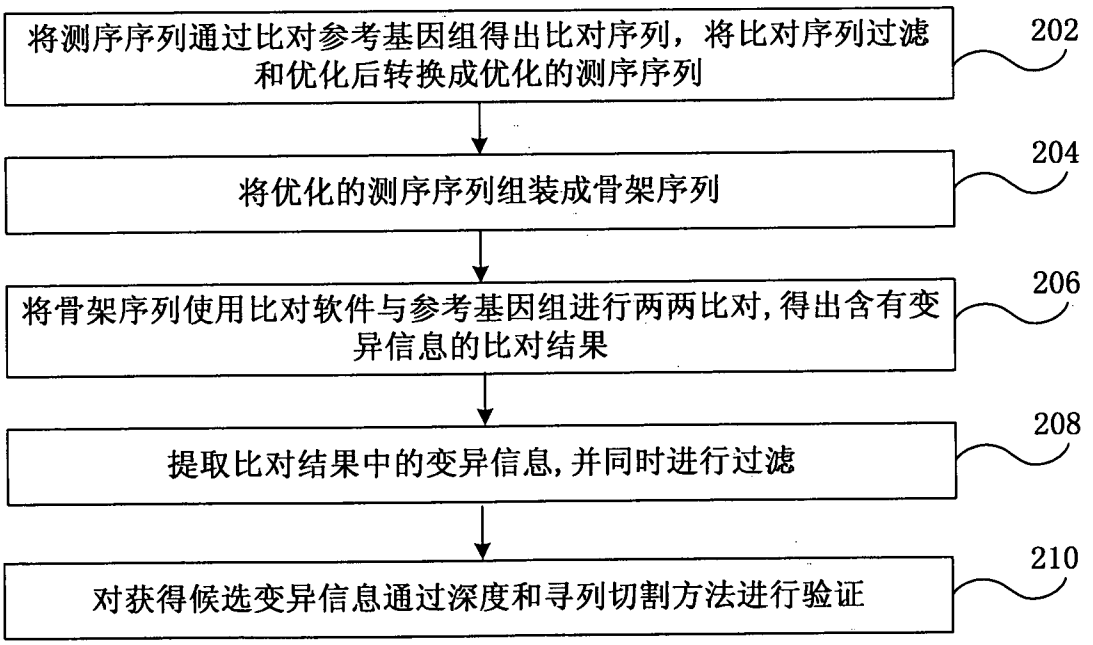
将测序序列通过比对参考基因组得出比对序列，将比对序列过滤和优化后转换成优化的测序序列
202
将优化的测序序列组装成骨架序列
204
将骨架序列使用比对软件与参考基因组进行两两比对，得出含有变异信息的比对结果
206
提取比对结果中的变异信息，并同时进行过滤
208
对获得候选变异信息通过深度和寻列切割方法进行验证
210

FIG.2

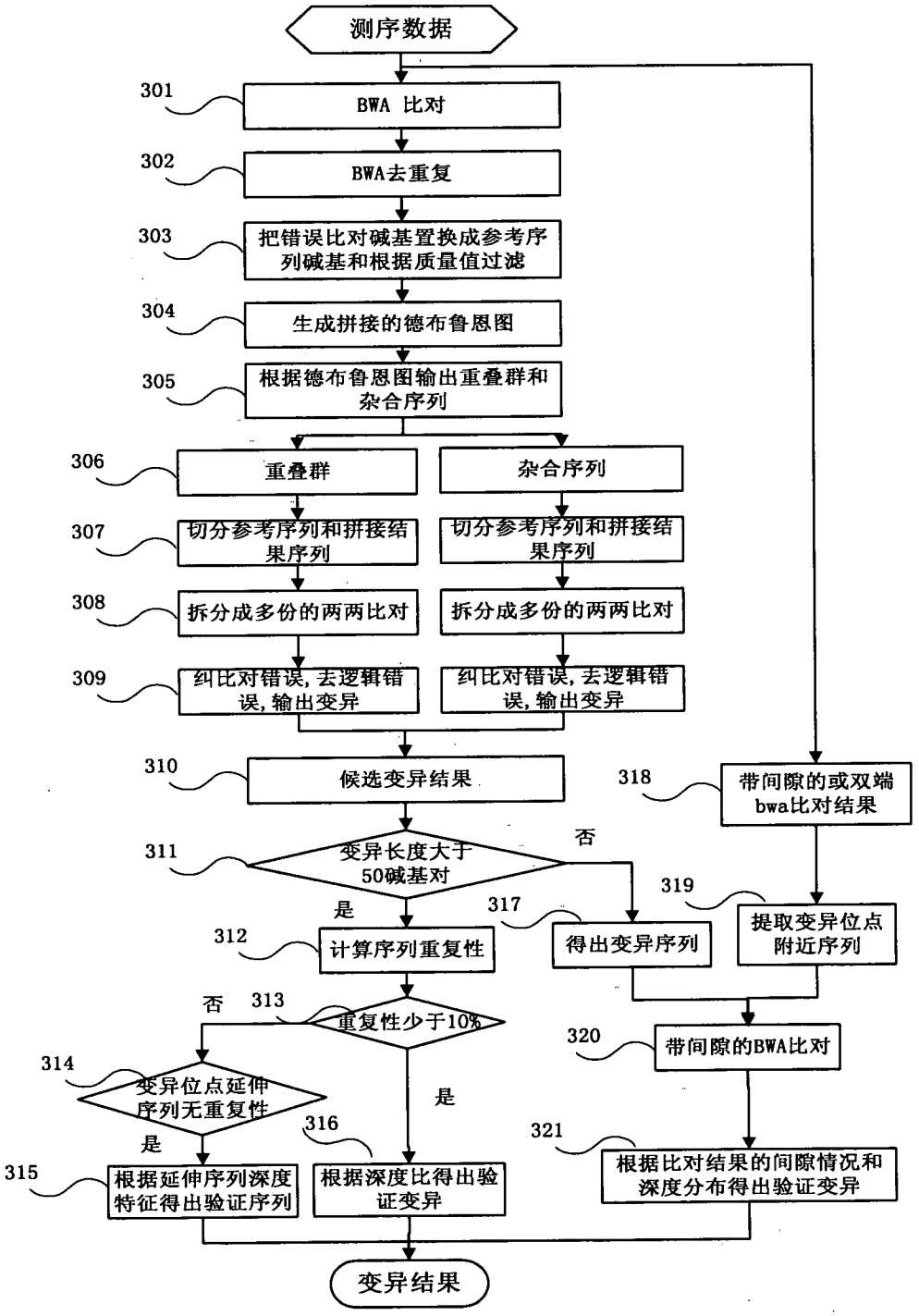
测序数据
301
BWA 比对
302
BWA去重复
303
把错误比对碱基置换成参考序列碱基和根据质量值过滤
304
生成拼接的德布鲁恩图
305
根据德布鲁恩图输出重叠群和杂合序列
306
重叠群
杂合序列
307
切分参考序列和拼接结果序列
切分参考序列和拼接结果序列
308
拆分成多份的两两比对
拆分成多份的两两比对
309
纠比对错误，去逻辑错误，输出变异
纠比对错误，去逻辑错误，输出变异
310
候选变异结果
318
带间隙的或双端bwa比对结果
311
变异长度大于50碱基对
否
是
317
得出变异序列
319
提取变异位点附近序列
312
计算序列重复性
否
313
重复性少于10%
320
带间隙的BWA比对
314
变异位点延伸序列无重复性
是
是
316
321
根据比对结果的间隙情况和深度分布得出验证变异
315
根据延伸序列深度特征得出验证序列
根据深度比得出验证变异
变异结果

FIG.3

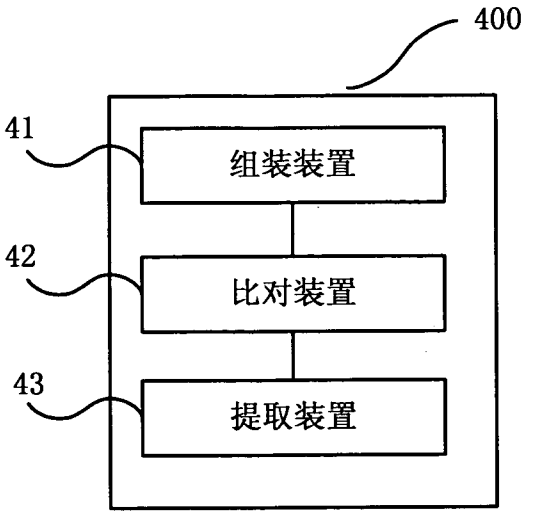
400
41
组装装置
42
比对装置
43
提取装置

FIG.4

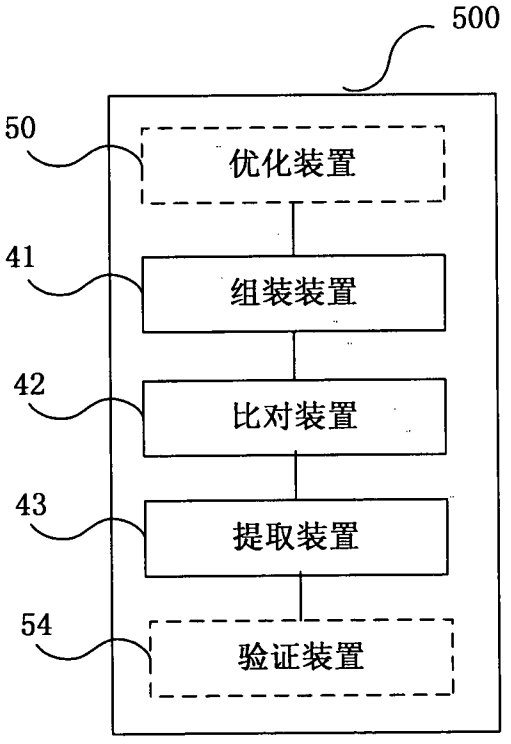
500
50
优化装置
41
组装装置
42
比对装置
43
提取装置
54
验证装置

FIG.5

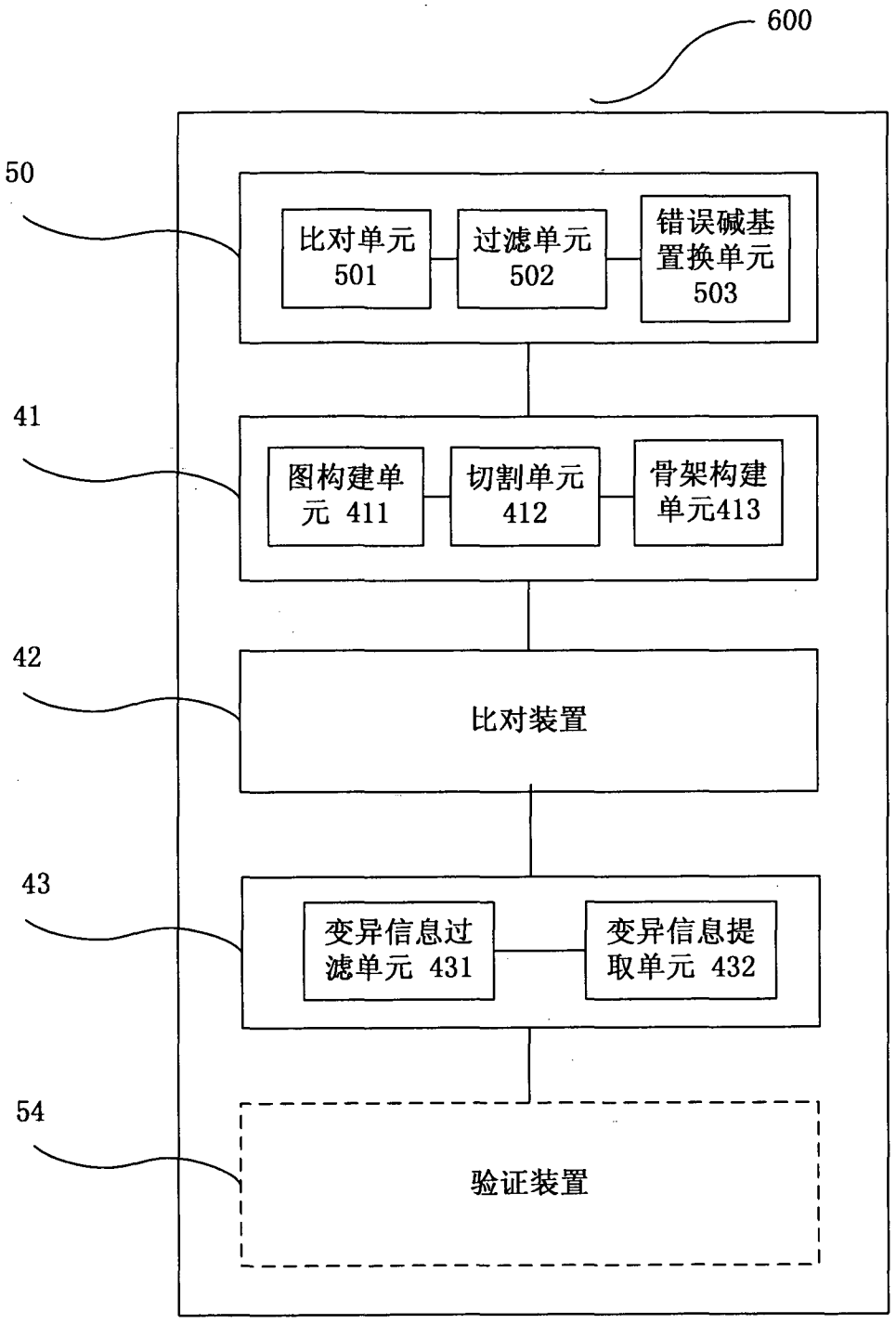
600
50
比对单元 501
过滤单元 502
错误碱基置换单元 503
41
图构建单元 411
切割单元 412
骨架构建单元413
42
比对装置
43
变异信息过滤单元 431
变异信息提取单元 432
54
验证装置

FIG.6

PATENT COOPERATION TREATY

PCT

DECLARATION OF NON－ESTABLISHMENT OF INTERNATIONAL SEARCH REPORT

(PCT Article 17(2)(a),Rules 13 *ter*.1(c) and 39)

Applicant's agent's reference	**IMPORTANT DECLARATION**	Date of mailing *(day/month/year)*
IEE100022PCT		30 Jun. 2011 (30.06.2011)
International application No.	International filing date *(day/month/year)*	(Earliest)Priority date*(day/month/year)*
PCT/CN2010/001409	14 Sep. 2010(14.09.2010)	
International Patent Classification (IPC) or both national classification and IPC		
C12Q 1/68 (2006.01) i		
Applicant		
BGI SHENZHEN CO.,LIMITED et al.		

This International Searching Authority hereby declares, according to Article 17(2)(a), that **no international search report will be established** on the international application for the reasons indicated below.

1. ☒ The subject matter of the international application relates to:
   - a. ☐ scientific theories
   - b. ☐ mathematical theories
   - c. ☐ plant varieties
   - d. ☐ animal varieties
   - e. ☐ essentially biological processes for the production of plants and animals, other than microbiological processes and the products of such processes
   - f. ☐ schemes, rules or methods of doing business
   - g. ☒ schemes, rules or methods of performing purely mental acts
   - h. ☐ schemes, rules or methods of playing games
   - i. ☐ methods for treatment of the human body by surgery or therapy
   - j. ☐ methods for treatment of the animal body by surgery or therapy
   - k. ☐ diagnostic methods practised on the human or animal body
   - l. ☐ mere presentations of information
   - m. ☐ computer programs for which this International Searching Authority is not equipped to search prior art

2. ☐ The failure of the following parts of the international application to comply with prescribed requirements prevents a meaningful search from being carried out:
   ☐ the description ☐ the claims ☐ the drawings

3. ☐ A meaningful search could not be carried out without the sequence listing; the applicant did not, within the prescribed time limit:
   - ☐ furnish a sequence listing on paper complying with the standard provided for in Annex C of the Administrative Instructions, and such listing was not available to the International Searching Authority in a form and manner acceptable to it.
   - ☐ furnish a sequence listing in electronic form complying with the standard provided for in Annex C of the Administrative Instructions, and such listing was not available to the International Searching Authority in a form and manner acceptable to it.
   - ☐ pay the required late furnishing fee for the furnishing of a sequence listing in response to an invitation under Rule 13ter.1(a) or (b).

4. Further comments:

Assembling scaffolds based on de Bruijn graph and the like is an algorithm defined by people, and the other steps are directed to people's inference and judgment , thus claims 1-7 relate to rules and methods for governing people's inference and judgment, and systems of claims 7-14 also relate to rules and methods for governing people's inference and judgment in substance, that is, all the claims are schemes, rules or methods of performing purely mental acts, no international search report will be established on the international application under Rule 39 .1 (iii) PCT.

Name and mailing address of the ISA/CN	Authorized officer
The State Intellectual Property Office, the P.R.China 6 Xitucheng Rd., Jimen Bridge, Haidian District, Beijing, China 100088 Facsimile No. 86-10-62019451	ZHAO,Yanhao Telephone No. (86-10)62411043

专 利 合 作 条 约
PCT

宣布不制定国际检索报告

(PCT 条约 17(2)(a)、细则 13 之三 1(c)和(d)和 39)

申请人或者代理人的档案号： **IEE100022PCT**	重 要 宣 布	发文日（日/月/年） **30.6 月 2011 (30.06.2011)**
国际申请号 **PCT/CN2010/001409**	国际申请日（日/月/年）： **14.9 月 2010(14.09.2010)**	(最早的) 优先权日（日/月/年）
国际专利分类（IPC）或者国家分类和 IPC 分类 C12Q 1/68 (2006.01) i		
申请人： **深圳华大基因科技有限公司 等**		

本国际检索单位根据条约 17(2)(a)宣布，对于国际申请**将不制定国际检索报告**，其理由如下：

1. ☒ 国际申请的主题涉及：
   - a. ☐ 科学理论。
   - b. ☐ 数学理论。
   - c. ☐ 植物品种
   - d. ☐ 动物品种。
   - e. ☐ 本质上是生物学方法的动植物生产方法，微生物方法以及用所述方法获得的产品除外。
   - f. ☐ 经营的方案、规则或方法。
   - g. ☒ 纯智力活动的方案、规则或方法。
   - h. ☐ 游戏的方案、规则或者方法。
   - i. ☐ 处理人体的外科手术或治疗方法。
   - j. ☐ 处理动物体的外科手术或治疗方法。
   - k. ☐ 在人体或动物体上实施的诊断方法。
   - l. ☐ 单纯的信息提供。
   - m. ☐ 本国际检索单位不具备检索现有技术的计算机程序。
2. ☐ 国际申请的下列部分不符合规定的要求，以致不能进行有意义的检索：
   ☐说明书　　☐权利要求书　　☐附图
3. ☐ 没有序列表,不能进行有意义的检索；申请人在规定的期限内：
   - ☐ 没有提交符合《行政规程》附录 C 规定标准的纸件形式的序列表，并且国际检索单位也未获得形式和方式可以接受的序列表。
   - ☐ 没有提交符合《行政规程》附录 C 规定标准的电子形式的序列表，并且国际检索单位也未获得形式和方式可以接受的序列表。
   - ☐ 在答复根据细则 13 之三.1(a)或(b)的通知提交序列表时，未缴纳后提交费用。
4. ☒ 补充意见：

由于根据德布鲁恩图等组装骨架序列只是一种人为规定的算法，而其他步骤也是人的推理、判断，所以权利要求 1-7 仅涉及指导人们进行推理、判断的规则和方法，而权利要求 7-14 所涉及的系统，其实质也是指导人们进行推理、判断的规则和方法，因此全部权利要求属于纯智力行为的范畴，属于 PCT 细则 39.1（iii）中所列的无需进行国际检索的情形，因此未进行国际检索。

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 **赵彦豪** 电话号码：(86-10) **62411043**