(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 5 月 31 日 (31.05.2012)

WIPO | PCT

(10) 国际公布号
WO 2012/068919 A1

(51) 国际专利分类号:
C40B 50/06 (2006.01) *C12N 15/11* (2006.01)
C40B 40/08 (2006.01) *C12Q 1/68* (2006.01)
C40B 20/04 (2006.01)

(21) 国际申请号: PCT/CN2011/079971

(22) 国际申请日: 2011 年 9 月 21 日 (21.09.2011)

(25) 申请语言: 中文

(26) 公布语言: 中文

(30) 优先权:
201010555192.4 2010 年 11 月 23 日 (23.11.2010) CN

(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)。 **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; 及

(75) 发明人/申请人 (仅对美国): **杜野 (DU, Ye)** [CN/CN]; 中国广东省深圳市盐田区北山工业区综合楼, Guangdong 518083 (CN)。 **赵美茹 (ZHAO, Meiru)** [CN/CN]; 中国广东省深圳市盐田区北山工业区综合楼, Guangdong 518083 (CN)。 **陈颖 (CHEN, Ying)** [CN/CN]; 中国广东省深圳市盐田区北山工业区综合楼, Guangdong 518083 (CN)。 **武靖华 (WU, Jinghua)** [CN/CN]; 中国广东省深圳市盐田区北山工业区综合楼, Guangdong 518083 (CN)。 **田埂 (TIAN, Geng)** [CN/CN]; 中国广东省深圳市盐田区北山工业区综合楼, Guangdong 518083 (CN)。 **王俊 (WANG, Jun)** [CN/CN]; 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, 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, QA, RO, RS, RU, RW, 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, RS, 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:
— 包括国际检索报告(条约第 21 条(3))。

(54) Title: DNA LIBRARY AND PREPARATION METHOD THEREOF, AND METHOD AND DEVICE FOR DETECTING SNPS

(54) 发明名称 : DNA 文库及其制备方法、以及检测 SNPs 的方法和装置

(57) Abstract: Provided are a DNA library and preparation method thereof, a method for determining DNA sequence information, a device and a kit for detecting SNPs, and a method for genotyping. The method for preparing the DNA library comprises the following steps: digesting the genomic DNA of a sample with a restriction endonuclease to obtain enzyme-digested products, wherein the restriction endonuclease comprises at least one selected from *MBo* II and *Tsp* 45I; separating the enzyme-digested products to obtain DNA fragments of 100 bp to 1,000 bp in length; subjecting the DNA fragments to terminal repair to obtain terminally repaired DNA fragments; adding the base A to the termini of the terminally repaired DNA fragments to obtain DNA fragments with terminal base A; and linking the DNA fragments with terminal base A to linkers for sequencing to obtain the DNA library.

(57) 摘要:

提供了 DNA 文库及其制备方法、确定 DNA 序列信息的方法、检测 SNPs 的装置和试剂盒、以及基因分型方法。其中，所述制备 DNA 文库的方法包括如下步骤：使用限制性内切酶对样本基因组 DNA 进行酶切，以便获得酶切产物，其中所述限制性内切酶包括选自 *Mbo* II 和 *Tsp* 45I 中的至少一种；将所述酶切产物进行分离，以便获得长度为 100 bp－1,000 bp 的 DNA 片段；将所述 DNA 片段进行末端修复，以便获得经过末端修复的 DNA 片段；在所述经过末端修复的 DNA 片段的末端添加碱基 A，以便获得具有末端碱基 A 的 DNA 片段；以及将所述具有末端碱基 A 的 DNA 片段与测序接头连接，以便获得所述 DNA 文库。

WO 2012/068919 A1

DNA 文库及其制备方法、以及检测 SNPs 的方法和装置

优先权信息

本申请请求 2010 年 11 月 23 日向中国国家知识产权局提交的、专利申请号为 201010555192.4 的专利申请的优先权和权益，并且通过参照将其全文并入此处。

技术领域

本发明涉及分子生物学领域，具体地，本发明提供了 DNA 文库及其制备方法、确定 DNA 序列信息的方法、检测 SNPs 的装置和试剂盒、以及基因分型方法。

背景技术

单核苷酸位点多态性（single nucleotide polymorphism，SNP）是指基因组上单个核苷酸的变异，它的数量巨大而且多态性丰富。SNP 在比较基因组学和进化基因组学的研究中被认为是最理想的遗传标记，同时在与疾病相关的遗传学和药理基因组学研究中也被作为有效的分子标记。而无论对于哪个应用领域，都需要对大量的样本中的 SNPs 位点进行检测并分型，虽然基因组的深度重测序是检测 SNPs 最为直接且有效的办法，但是目前基因组测序的成本较为昂贵，无法满足大规模样本测序的要求，所以，许多高通量的 SNPs 分型方法和商业平台得以大力发展（Chunming Ding and Shengnan Jin.(2009). High-Throughput Methods for SNP Genotyping. *Single Nucleotide Polymorphisms, Methods in Molecular Biology*. AA. Komar (eds), Humana Press. p578，通过参照将其全文并入本文）。

然而，目前对样本 SNPs 进行检测的方法，仍有待改进。

发明内容

本发明是基于发明人的下列发现而完成的：

目前，应用比较普遍的高通量 SNPs 分型平台主要是基于单碱基延伸技术的 Iillumina® BeadArray 平台和基于差异杂交方法的 Affymetrix SNP microarray，两者都是基于已有的 SNPs 位点信息，通过设计合成探针根据各自不同的原理对特定的 tagSNPs 进行检测的方法，同时可以针对不同的待关联分析的性状设计不同的待检 SNPs 组合，使检测设计更加灵活，特异性更高。但这些方法也有一定的局限性，比如探针要经过严格的筛选设计，而且并不是所有的 tagSNPs 都能满足这些设计的要求。同时芯片的合成要求比较高，一般实验室很难实现，而购买商业芯片又导致成本过高同时需要专门的扫描仪器和分析软件。此外，一个重要的限制因素是该方法中探针的设计必须建立在已知 SNPs 数据库的基础上，无法发现未知的 SNPs 位点（Chunming Ding and Shengnan Jin.(2009). High-Throughput Methods for SNP Genotyping. *Single Nucleotide Polymorphisms, Methods in Molecular Biology*. AA. Komar (eds), Humana Press. p578，通过参照将其全文并入本文）。

此外，将限制性内切酶与下一代测序技术（NGS）结合，用于基因组范围内特异位点的多态性检测【Nathan A, Baird, Paul D, *et al.* (2008). Rapid SNP Discovery and Genetic Mapping Using Sequenced RAD Markers. *PLoS ONE*, 3(10):3376.; Michael A. Gore, *et al.*(2009). A First-Generation Haplotype Map of Maize. *Science*, 326:1115，通过参照将其全文并入本文）】的几个现有方法，或者依赖于对一种限制性酶切位点关联的 DNA（RAD）标签的测序（Nathan A, Baird, Paul D, *et al.* (2008) Rapid SNP Discovery and Genetic Mapping Using Sequenced RAD Markers. *PLoS ONE*, 3(10):3376.; S á nchez CC, Smith TPL, Wiedmann RT, *et al.* (2009). Single nucleotide polymorphism discovery in rainbow trout by deep sequencing of a reduced representation library. *BMC Genomics*, 10:559.，通过参照将其全文并入本文），或者至少要使用一种稀有位点的限制性内切酶，且酶切建库过程都比较繁琐，而其中有些方法因为要经过 PCR 扩增过程更容易将序列的偏差引入并放大。

本发明旨在解决现有技术问题的至少之一。为此，本发明的一个方面，提出了一种制备 DNA 文库的方法，其可以制备用于检测 SNPs 的 DNA 文库。根据本发明的实施例，其包括以下步骤：使用限制性内切酶，对样本基因组 DNA 进行酶切，以便获得酶切产物，其中所述限制性内切酶包括选自 *Mbo* II 和 *Tsp* 45I 的至少一种；将所述酶切产物进行分离，以便获得长度为 100 bp－1,000 bp 的 DNA 片段；将所述 DNA 片段进行末端修复，以便获得经过末端修复的 DNA 片段；在所述经过末端修复的 DNA 片段的末端添加碱基 A，以便获得具有末端碱基 A 的 DNA 片段；以及将所述具有末端碱基 A 的 DNA 片段与测序接头连接，以便获得所述 DNA 文库。利用根据本发明实施例的构建 DNA 文库的方法，能够有效地构建样本的 DNA 文库，从而可以通过对 DNA 文库进行测序，获得样品 DNA 的序列信息，最终通过对 DNA 序列信息进行 SNPs 数据分析就可以获得样本 DNA 的 SNPs 信息。另外，本发明人发现，上述方法过程简单，极易操作，操作流程易标准化，且成本较低。除此之外，发明人还惊奇地发现，当针对相同的样品，基于上述方法，采用候选的不同的限制性内切酶构建 DNA 文库时，所得到的测序数据结果的稳定性和可重复性非常好。

进一步，本发明还提供了一种 DNA 文库，其是由根据本发明实施例的制备 DNA 文库的方法所获得的。

根据本发明的又一方面，本发明还提供了一种确定DNA序列信息的方法。根据本发明的实施例，其包括下列步骤：根据本发明实施例的制备DNA文库的方法构建所述样本基因组DNA的DNA文库；以及对所述DNA文库进行测序，以便获得所述DNA序列信息。基于该方法，能够有效地获得DNA文库中DNA样品的序列信息，从而能够对DNA序列信息进行SNPs数据分析，以获得样本DNA的SNPs信息。另外，发明人惊奇地发现，利用根据本发明实施例的方法确定DNA样品序列信息，能够有效地减少数据产出偏向性的问题，并能够降低成本。

根据本发明的又一方面，本发明还提供了一种用于检测SNPs的装置，根据本发明的实施例，其包括如下单元：DNA文库制备单元，所述DNA文库制备单元用于制备DNA文库；测序单元，所述测序单元与所述DNA文库制备单元相连，用于对所述DNA文库进行测序，以便获得DNA序列信息；以及SNPs数据分析单元，所述SNPs数据分析单元与所述测序单元相连，用于对所述DNA序列信息进行SNPs数据分析，以便获得SNPs信息。利用该装置，能够方便地对样本进行SNPs检测，并能获得准确的SNPs信息，而且可以应用于大规模数量的样本的SNPs检测。

根据本发明的再一方面，本发明还提供了一种用于检测 SNPs 的试剂盒，根据本发明的实施例，该试剂盒包括：限制性内切酶，所述限制性内切酶包括选自 *Mbo* II 和 *Tsp* 45I 的至少一种。由此，利用该试剂盒，能够方便地检测样本的 SNPs。

根据本发明的再一方面，本发明还提供了一种基因分型方法，根据本发明的实施例，其包括：提供样本基因组；根据本发明实施例的构建DNA文库的方法，制备样本基因组的DNA文库；对所述DNA文库进行测序，以便获得所述DNA序列信息；对所述DNA序列信息进行SNPs数据分析，以便获得所述样本的SNPs信息；以及基于所述SNPs信息对所述样本进行基因分型。利用上述方法，通过构建符合SNPs检测要求的样本DNA文库，以及对DNA文库进行测序获得DNA样品的序列信息，然后对DNA序列信息进行SNPs数据分析，就能够准确有效地获得样本DNA的SNPs信息，从而基于获得的样本的SNPs信息，结合该物种已有的基因型信息，就能够有效地对样本进行基因分型。另外，本发明人发现，该基因分型方法过程简单，操作容易，且成本很低。

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：显示了根据本发明实施例的 SNPs 检测方法的流程；

图 2：显示了根据本发明实施例的构建 DNA 文库的方法构建 DNA 文库时，基因组 DNA 被酶切后的电泳检测结果。

图 3：显示了根据本发明实施例的构建 DNA 文库的方法构建 DNA 文库时，基因组 DNA 分别被四种酶组合酶切后的 Agilent® Bioanalyzer 2100 检测结果。

图 4：显示了根据本发明实施例的构建 DNA 文库的方法，使用 *Tsp* 45I 构建的 DY 文库的插入片段范围的统计曲线。

图 5：显示了根据本发明实施例的构建 DNA 文库的方法，使用 *Tsp* 45I 构建的 YH 文库的插入片段范围的统计曲线。

图 6：显示了根据本发明实施例的构建 DNA 文库的方法，使用 *Tsp* 45I 构建的 DY 文库的测序数据深度的统计曲线。

图 7：显示了根据本发明实施例的构建 DNA 文库的方法，使用 *Tsp* 45I 构建的 YH 文库的测序数据深度的统计曲线。

图 8：显示了根据本发明实施例的构建 DNA 文库的方法，使用 *Tsp* 45I 分别构建的 DY 文库和 YH 文库间的目标区域覆盖深度一致性的比较图。

图 9：显示了根据本发明实施例的构建 DNA 文库的方法，两次构建的 YH 文库间的目标区域覆盖深度一致性的比较图。

图 10 显示了根据本发明一个实施例的用于检测 SNPs 的装置的示意图。

发明详细描述

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，仅用于解释本发明，而不能理解为对本发明的限制。

需要说明的是，术语“第一”、“第二”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一个或者更多个该特征。进一步地，在本发明的描述中，除非另有说明，“多个”的含义是两个或两个以上。

DNA 文库及其构建并测序的方法

根据本发明的一个方面，本发明提出了一种制备 DNA 文库的方法，其可以制备用于检测 SNPs 的 DNA 文库。具体地，根据本发明的实施例，参考图 1，该方法包括以下步骤：

首先，使用限制性内切酶，对样本基因组 DNA 进行酶切，以便获得酶切产物，其中限制性内切酶包括选自 *Mbo* II 和 *Tsp* 45I 的至少一种。根据本发明的实施例，限制性内切酶进一步包括选自 *Hin*d III 和 *Bcc* I 的至少一种。根据本发明的实施例，DNA 样品的来源并不受特别限制。根据本发明的具体示例，样本基因组 DNA 可以是来源于目前已有全基因组序列数据的任何物种（例如 http://www.ncbi.nlm.nih.gov/sites/genome 所列物种），具体地，样本基因组 DNA 可以取自该物种的个体、单个细胞或某个组织。优选地，根据本发明的实施例，样本基因组 DNA 为人的基因组 DNA。另外，根据本发明的实施例，基因组 DNA 的提取方法不受特别限制。本领域技术人员可以理解，基因组 DNA 的提取可以根据物种和样本的不同而选取不同的方法，具体地，可以按照本领域已知的方法完成（包括使用商品化的试剂盒），比如植物组织或微生物可以使用标准的 CTAB 法提取，人类血液基因组 DNA 可以使用 QIAamp® DNA Mini Kit(QIAGEN)完成等。根据本发明的实施例的构建 DNA 文库的方法，要求得到的基因组 DNA 应尽量保持完整，即要减少因人为断裂而产生过多的小 DNA 片段，一般认为，经琼脂糖凝胶电泳检测达到 23K 以上的标准为合格，同时要求 DNA 纯度尽量高，以避免影响酶切。

此外，本申请的发明人发现，构建 DNA 文库必须选择至少一种限制性内切酶对基

因组 DNA 进行酶切，所用限制性内切酶依赖于所研究的物种不同而略有不同，其中较常用的识别序列为 5 或 6 碱基的 II 型限制性内切酶，此外切割位点在识别位点以外的 IIs 型限制性内切酶也可以使用。一般来讲，所用限制性内切酶应该为 1-2 种，因为使用过多的限制性内切酶较难在一管反应体系中完成，不仅会增加操作的复杂性，而且容易导致酶切不完全或星号活性的出现。而目前，有许多商品化的限制性内切酶的可供选择，比如 NEB（NEW ENGLANG BioLabs）公司、TaKaRa 公司等等。为此，本申请的发明人进行了大量的筛选工作，并且选定了优选的根据本发明实施例的限制性内切酶，其为选自下列的至少一组：（1）*Mbo* II；（2）*Tsp* 45I；（3）*Mbo* II 和 *Hind* III；以及（4）*Mbo* II 和 *Bcc* I。

其次，将酶切产物进行分离，以便获得长度为 100 bp－1,000 bp 的 DNA 片段。根据本发明的实施例，对酶切产物进行分离回收的方法不受限制，可以按照本领域所熟知的方法进行。根据本发明的具体示例，可以使用合适浓度的琼脂糖凝胶电泳分离酶切产物。具体地，根据本发明的实施例，采用 2%的琼脂糖凝胶电泳分离酶切产物，接着切取目标长度范围内（100 bp－1,000 bp）的凝胶，然后可以利用商品化的凝胶回收试剂盒（例如 MinElute® PCR Purification Kit（QIAGEN））回收目标长度范围内的 DNA 片段。根据本发明的实施例，DNA 片段的目标长度范围为 100 bp－1,000 bp，进一步，DNA 片段的长度为 200 bp－700 bp。

接下来，将 DNA 片段进行末端修复，以便获得经过末端修复的 DNA 片段，以及在经过末端修复的 DNA 片段的末端添加碱基 A，以便获得具有末端碱基 A 的 DNA 片段。根据本发明的实施例，对 DNA 片段进行末端修复和加“A”反应使用标准化的流程，其具体过程如下：在一个反应体系中加入上述 DNA 片段、10mM dNTP、T4 DNA Polymerase、Klenow Fragment、T4 Polynucleotide Kinase 以及 T4 DNA ligase buffer (with 10mM ATP)，20℃温育 30 分钟，然后回收 DNA 片段，接着在另一反应体系中加入补平的 DNA、dATP、Klenow Fragment、Klenow (3’ -5’ exo$^-$)，于 37℃反应 30 分钟即可。

最后，将具有末端碱基 A 的 DNA 片段与测序接头连接，以便获得 DNA 文库。根据本发明的实施例，测序接头的选择不受特别限制，可以依据所使用的测序技术方法（高通量测序平台）而选择不同的接头。根据本发明的实施例，采用 Iillumina®公司的边合成边测序原理方法，因此测序接头就选择了相应的 Iillumina®接头，该接头序列包含与测序所用 flow cell 上具有的寡核苷酸的序列互补的序列，由此可以将文库片段连接到 flow cell 上，从而得以继续接下来的的测序流程。根据本发明的实施例，测序接头不需要包含扩增引物结合位点（因为根据本发明实施例的构建 DNA 文库的方法，不涉及 PCR 扩增），但需要带有测序引物的结合位点，进一步地，为了将来源于不同样本制备的 DNA 文库在测序后区分开来，8 bp 的标签序列以及标签测序引物序列也可被带入到一侧接头中，这样可以方便将不同文库直接混合后上机测序，得以应用于大规模数量样本的建库测序。另外，根据本发明的实施例，文库构建结束后，可以利用 Agilent® Bioanalyzer 2100 检测文库片段分布情况以及利用 Q-PCR 对文库进行定量。

利用根据本发明实施例的构建 DNA 文库的方法，能够有效地构建样本的 DNA 文库，通过对 DNA 文库测序获得的样品 DNA 的序列信息进行 SNPs 数据分析，就可以准确地获得样本 DNA 的 SNPs 信息，从而可以利用样本的 SNPs 信息进行许多相关科学研究。另外，本发明人发现，上述方法过程简单，极易操作，操作流程易标准化，且成本较低。除此之外，发明人还惊奇地发现，当针对相同的样品，基于上述方法，采用候选的不同的限制性内切酶构建 DNA 文库时，所得到的测序数据结果的稳定性和可重复性非常好；而针对相同样本进行多次平行建库时，测序数据结果稳定，表明根据本发明实施例的构建 DNA 文库的方法平行性及可重复性好。

进一步，根据本发明的实施例，本发明提供了一种构建 DNA 文库的方法，其包括：

1）使用至少一种限制性内切酶，对样本基因组 DNA 进行酶切，得到酶切产物；

2）将酶切产物进行分离，得到长度为 100 bp－10,00 bp 的 DNA 片段；以及

3）将步骤 2）中得到的 DNA 片段进行末端修复；

优选地，还包括下述步骤：

4）将步骤 3）中得到的 DNA 片段的末端添加碱基 A；

优选地，还包括下述步骤：

5）将步骤 4）中得到的 DNA 片段连接测序接头。

根据本发明的一些具体示例，上述根据本发明实施例的构建 DNA 标签文库的方法的步骤 1）：所用样本基因组 DNA 可以是来源于目前已有全基因组序列数据的任何物种（例如 http://www.ncbi.nlm.nih.gov/sites/genome 所列物种），基因组 DNA 可以取自该物种的个体、单个细胞或某个组织。优选地，为人的基因组 DNA。对本领域技术人员而言，基因组 DNA 的提取方法根据物种和样本的不同，可以按照本领域已知的方法完成（包括使用商品化的试剂盒），比如植物组织或微生物可以使用标准的 CTAB 法提取，人类血液基因组 DNA 可以使用 QIAamp® DNA Mini Kit(QIAGEN)完成等。得到的基因组 DNA 应尽量保持完整，减少因人为断裂而产生过多的小 DNA 片段，一般经琼脂糖凝胶电泳检测达到 23K 以上的标准视为合格，同时 DNA 纯度尽量高以避免影响酶切过程的因素存在。另外，选择至少一种限制性内切酶对基因组 DNA 进行酶切，所用限制性内切酶依赖于所研究的物种不同而略有不同，其中较常用的识别序列为 5 或 6 碱基的 II 型限制性内切酶，此外切割位点在识别位点以外的 II s 型限制性内切酶也可以使用。一般来讲，所用限制性内切酶应该为 1-2 种，因为使用过多的限制性内切酶较难在一管反应体系中完成，不仅会增加操作的复杂性，而且容易导致酶切不完全或星号活性的出现。目前，有许多商品化的限制性内切酶的可供选择，比如 NEB（NEW ENGLANG BioLabs）公司、TaKaRa 公司等等，反应条件以限制性内切酶提供说明书为准，以保证达到优选的酶切效果。优选地，所述酶切为完全酶切。根据本发明的实施例，以人类基因组为主要研究对象，分别设计了不同的酶切组合，其中优选酶切组合如表 1 中所示。其中，限制性内切酶名称以 NEB 公司公布为准。

根据本发明的一些具体示例，上述根据本发明实施例的构建 DNA 标签文库的方法的步骤 2）：按照本领域所熟知的方法进行酶切后基因组片段的回收，例如使用合适浓度的琼脂糖凝胶电泳分离酶切 DNA 片段。一般地，对于回收 1 kb 以下范围内的 DNA 片段，2%的琼脂糖凝胶是比较合适的选择，电泳结束后切取目标长度范围内的凝胶。然后可以使用商品化的凝胶回收试剂盒（例如 MinElute® PCR Purification Kit（QIAGEN））， 回收目标长度范围内的 DNA 片段。另外，限制性内切酶将人类基因组切割成基本相同的长度分布（例如 100 bp－10,00 bp），该范围内分布的片段是为了得到一部分基因组，并且一个库中片段长度相差过大会影响最后测序数据的质量，并且会导致很大的增加成本。根据本发明的实施例，得到的 DNA 片段的长度为 100 bp－1,000 bp，进一步，根据本发明的实施例，DNA 片段的长度为 200 bp－700 bp。为了有效地得到该长度范围的 DNA 片段，本发明人进行了大量的研究和不懈的努力，发现此方法在步骤 1）中的限制性内切酶优选地为选自根据本发明实施例的下面的（1）－（4）中的至少一组（如表 1 所示）：（1）*Mbo* II；（2）*Tsp* 45I（3）*Mbo* II 和 *Hin*d III；以及（4）*Mbo* II 和 *Bcc* I。

根据本发明的一些具体示例，上述根据本发明实施例的构建 DNA 标签文库的方法的步骤 3）和 4）：回收后的酶切 DNA 片段使用标准化的流程进行末端修复和加“A”反应，具体过程如下：在一个反应体系中加入回收的 DNA、10mM dNTP、T4 DNA Polymerase、Klenow Fragment、T4 Polynucleotide Kinase 以及 T4 DNA ligase buffer (with 10mM ATP)在 20℃温育 30 分钟后，回收片段，在另一反应体系中加入补平的 DNA、dATP、Klenow Fragment、Klenow (3’ -5’ exo$^-$)于 37℃反应 30 分钟。

根据本发明的一些具体示例，上述根据本发明实施例的构建 DNA 标签文库的方法的步骤 5）：接头与限制性片段的连接，接头的选择会因所使用的测序技术方法（高通量测序平台）的不同而有所不同。在本发明实施例 2 中所用为 illumina®公司的边合成边测序原理方法，所以，illumina®接头序列包含与测序所用 flow cell 上连接寡核苷酸互补的序列以便于将文库片段连接到 flow cell 上。由于本发明并不使用 PCR 扩增的方法，所以，所加接头不需要包含扩增引物结合位点，但需要带有测序引物的结合位点，为了将来源于不同样本制备的 DNA 文库在测序后区分开来，8 bp 的 Index 标签序列以及 index 标签测序引物序列也可被带入到一侧接头中，这样可以方便将不同文库直接混合后上机测序。文库构建结束后，需经 Agilent® Bioanalyzer 2100 检测文库片段分布情况以及经过 Q-PCR 对文库进行定量。

利用根据本发明实施例的构建 DNA 文库的方法，能够有效地构建样本的 DNA 文库，对 DNA 文库测序后，能准确地获得样品 DNA 的序列信息，通过对样品 DNA 的序列信息进行 SNPs 数据分析，就可以获得样本 DNA 的 SNPs 信息，从而可以成功地应用于许多下游的相关科学研究。另外，本发明人发现，上述方法过程简单，操作流程可标准化，则操作方便，而且成本较低。此外，发明人惊奇地发现，当针对相同的样品，基于上述方法，采用不同的限制性内切酶构建 DNA 文库时，测序数据结果稳定，可重复性好；而针对相同样本多次平行建库时，测序数据结果稳定性好，则表明根据本发明实施例的构建 DNA 文库的方法平行性及可重复性好。

根据本发明的又一方面，本发明还提供了一种 DNA 文库，其是根据本发明的构建 DNA 文库的方法所构建的。该 DNA 文库可以有效地应用于高通量测序技术例如 Solexa 技术，从而可以通过获得样本 DNA 的序列信息，进而对其进行进行 SNPs 数据分析，从而可以获得样本 DNA 的 SNPs 信息，以为应用于下游的相关科学研究做好准备。

根据本发明的再一方面，本发明还提供了一种确定 DNA 序列信息的方法，其是通过对根据本发明实施例的构建 DNA 文库的方法构建的 DNA 文库进行测序而实现的。根据本发明的具体示例，其包括下列步骤：根据本发明实施例的制备 DNA 文库的方法构建样本基因组 DNA 的 DNA 文库；以及对 DNA 文库进行测序，以便获得 DNA 序列信息。进一步地，根据本发明的实施例，还包括对 DNA 序列信息进行 SNPs 数据分析的步骤，以便获得所述 DNA 的 SNPs 信息。根据本发明的实施例，利用选自 GS 测序平台、GA 测序平台、HiSeq2000™ 测序平台、以及 SOLiD™ 测序平台对所述 DNA 文库进行测序。基于该方法，能够有效地获得 DNA 文库中 DNA 样品的序列信息，从而能够对 DNA 序列信息进行 SNPs 数据分析，以获得样本 DNA 的 SNPs 信息，进而可以依据得到的样本的 SNPs 信息，对各样本进行基因分型等科学研究。另外，发明人惊奇地发现，利用根据本发明实施例的方法确定 DNA 样品序列信息，能够有效地减少数据产出偏向性的问题，而且此方法可操作性和平行性好，针对大规模样本的测序时还能够有效地简化流程，并降低测序成本。

检测SNPs的装置、试剂盒以及基因分型方法

根据本发明的又一方面，本发明还提供了一种用于检测 SNPs 的装置。参考图 10，根据本发明的实施例，该用于检测 SNPs 的装置 1000 包括：DNA 文库制备单元 100、测序单元 200 以及 SNPs 数据分析单元 300。根据本发明的实施例，DNA 文库制备单元 100 用于制备 DNA 文库，例如可以采用适于前面所述的文库构建方法的任意装置作为 DNA 文库制备单元 100。测序单元 200 与 DNA 文库制备单元 100 相连，可以从 DNA 文库制备单元 100 接收所制备的 DNA 文库，并对所接收的 DNA 文库进行测序，从而可以获得样本的 DNA 序列信息。SNPs 数据分析单元 300 与测序单元 200 相连，可以从测序单元 200 接收所获得的样本的 DNA 序列信息，并且能够进一步对 DNA 序列信息进行 SNPs 数据分析，从而获得 SNPs 信息。本领域技术人员能够理解的是，可以采用本领域中已知的任何适于进行上述操作的装置作为上述各个单元的组成部件。另外，

这里所使用的术语“相连”应作广义理解，可以是直接相连，也可以通过中间媒介间接相连，对于本领域的普通技术人员而言，可以根据具体情况理解上述术语的具体含义。

利用根据本发明实施例的上述装置，能够方便地对样本进行 SNPs 检测，并能获得准确的 SNPs 信息。另外，本发明人发现，根据本发明的实施例的检测 SNPs 的装置能够应用于大规模数量的样本的 SNPs 检测，从而简化测序流程，节省测序时间及成本，并且获得的 SNPs 信息较多而准确，此用途只需要在 DNA 文库制备单元中于 DNA 文库加入 Index 标签，并将来自于多个样本的 DNA 文库进行混合测序即可实现。

根据本发明的再一方面，本发明还提供了一种用于检测 SNPs 的试剂盒，根据本发明的实施例，该试剂盒包括：限制性内切酶，所述限制性内切酶包括选自 *Mbo* II 和 *Tsp* 45I 的至少一种。由此，利用该试剂盒，能够方便地检测样本的 SNPs。

根据本发明的再一方面，本发明还提供了一种基因分型方法，根据本发明的实施例，其包括：首先，提供样本基因组；接下来，根据本发明实施例的构建DNA文库的方法，制备样本基因组的DNA文库；对DNA文库进行测序，以便获得DNA序列信息；对DNA序列信息进行SNPs数据分析，以便获得所述样本的SNPs信息；以及基于SNPs信息对样本进行基因分型。利用上述方法，通过构建符合SNPs检测要求的高质量的样本DNA文库，基于对高质量的DNA文库的测序可以获得DNA样品的序列信息，然后基于对DNA序列信息进行SNPs数据分析获得的准确有效的SNPs信息，再结合已有的基因型信息，就能够有效地对样本进行基因分型。另外，本发明人发现，该基因分型方法过程简单，操作容易，能够同时应用于大规模样本，且成本很低。

需要说明的是，根据本发明实施例的确定 DNA 样品序列信息的方法是本申请的发明人经过艰苦的创造性劳动和优化工作才完成的。

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（例如参考 J.萨姆布鲁克等著，黄培堂等译的《分子克隆实验指南》，第三版，科学出版社）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，例如可以采购自 Illumina 公司。

实施例 1：优选的限制性内切酶或者酶组合的确定

按照下表 1 中的酶或酶组合的识别序列，通过已知的酶切识别位点信息，以 hg18 基因组序列为参考序列，以酶切位点为分界将基因组按长度范围分类，最终选取 200bp-700bp 范围的片段作为待测的文库集合。对本领域技术人员而言，hg18 基因组序列数据可以从已知的数据库下载，例如从 http://genome.ucsc.edu/上下载。

按照 Iillumina® HiSeq2000™ PE91 index 测序参数过滤产生数据。由于在实际测序中使用 PE91 循环数测序，所以，将以上文库集合中每个片段两端 91bp 的碱基作为目标区域，以按照 PE91 长度测序参数将在选定范围内的片段包含酶切位点两端的 91bp 作为目标区域，统计目标区域覆盖 dbSNP v128 数据库（http://www.ncbi.nlm.nih.gov/projects/SNP/）中 SNP 位点数目，以及该数目所占 dbSNP v128 中总数的比例。

由于所用参考序列为国际上公用，特别是不涉及实际实验中会产生的其它因素的干扰（比如 DNA 的不可避免的断裂，酶切的不完全等），因此得到的结果是最理想状态下的结果，也就是最优化的结果。

表 1:人类基因组限制性内切酶酶切建库的优选酶组合

回收片段范围	酶或酶的组合	可检测到 SNP 数目	dbSNP v128 覆盖度
200bp-700bp	*Mbo* II	3338421	26.90%

	Tsp 45I	1579936	12.73%
	Mbo II 和 *Hin*d III	3597897	28.99%
	Mbo II 和 *Bcc* I	4835970	38.97%

与上面的检验方法类似，本发明人还检验了大量其它的酶或酶的组合，计算得到的 dbSNP v128 覆盖度一般都在 10%以下，部分酶或酶的组合的检验结果如下表 2 所示：

表 2：检验过的部分其它酶和酶的组合

回收片段范围	酶或酶的组合	可检测到 SNP 数目	dbSNP v128 覆盖度
200bp-700bp	*Bcc* I	751883	6.56%
	Bgl II	76415	0.67%
	*Bam*H I 和 *Bcc*I	893795	7.80%
	*Hin*d III 和 *Bgl*II	277079	2.42%

从表 2 可见，表 2 中的酶或酶的组合的可检测到的 SNP 数目和 dbSNP v128 覆盖度都远低于上述表 1 中所列的酶或酶的组合。

因此，表 1 中的酶或酶的组合是优选的方案。

实施例 2：炎黄一号 DNA 文库的测序

针对于人类基因组，如详细技术方法表 1 中所表述的优选酶切组合，选取其中回收片段在 200 bp－700 bp 范围内的四种优选酶切组合进行酶切建库，通过数据分析并与表 1 所示的结果相比。具体操作如下：

人类基因组 DNA 提取自炎黄一号（YH1）的血液细胞，提取使用 QIAamp® DNA Mini Kit（QIAGEN）完成，操作完全按照说明书进行。最后基因组 DNA 溶解于 EB 缓冲液中，经 NonoDrop® ND-1000 以 A260 处吸光值进行定量后，取 5 微克进行酶切。限制性内切酶全部购买自 NEB 公司，缓冲液随酶提供，共进行四种酶切组合。

每个酶切反应体系中基因组 DNA 都为 5 微克，限制性内切酶用量为 20U(NEB 定义单位)，每个反应中因酶组合的不同而选用最适合的缓冲液以及反应条件，详细见下面的表 3。

表 3：酶切体系

酶组合	*Mbo* II	*Tsp* 45I	*Mbo* II +*Hin*d III	*Mbo* II +*Bcc* I
缓冲液	NEBuffer4	NEBuffer1 + BSA	NEBuffer2	NEBuffer1 +BSA
反应条件	37℃、1hr	65℃、1hr	37℃、1hr	37℃、1hr

以上反应缓冲液都是 10×母液，最后以超纯水将反应体系补平至 100 微升，按照最适反应条件进行。

酶切后的基因组 DNA 经 2%琼脂糖凝胶电泳（TAE 缓冲系统）分离后（图 2），手工切取 200 bp－700 bp 长度范围内的片段经 QIAquick® Gel Extraction Kit（QIAGEN）凝胶回收，将溶于 30 微升超纯水中。

末端修复反应按照如下体系进行：

T4 DNA ligase buffer with 10mM ATP	10 微升
dNTPs	4 微升
T4 DNA Polymerase	5 微升
Klenow Fragment	1 微升
T4 Polynucleotide Kinase	5 微升
DNA	30 微升

ddH_2O	up to 100 微升

20℃反应 30 分钟后，使用 MinElute® PCR Purification Kit(QIAGEN)回收补平的 DNA 片段。样品最后溶于 32 微升的 EB 缓冲液中。

加“A”反应按照以下体系完成:

Klenow buffer	5 微升
dATP	10 微升
Klenow (3’ -5’ exo-)	3 微升
DNA	32 微升

37℃温育 30 分钟后，经 MinElute® PCR Purification Kit(QIAGEN)纯化并溶于 35 微升的 EB 中。

接头的连接反应如下:

10x T4 DNA Ligation buffer	5 微升
PCR-free Adapter oligo mix	5 微升
T4 DNA Ligase	5 微升
加“A”后的样品 DNA	35 微升

连接反应于 16℃连接过夜。其中接头为 Iillumina®公司 PCR-free index 接头，四个库分别带有唯一的 8 bp index 标签序列，将构建好的文库经 Agilent® Bioanalyzer 2100 检测片段分布范围（图 3，A－D）。从图 3 可见，文库切割的片段范围为 200 bp－700 bp，在连接接头以后片段长度增加约 120 bp 左右，由图 3 可以看出四个文库片段范围基本符合要求，而且文库质量符合测序要求。将其中使用 *Tsp* 45I 酶构建的文库命名为 YH 文库（YH 文库 trial 1）。

再经过 Q-PCR 方法对四个文库进行定量，并以此为标准将除 *Mbo* II+*Bcc* I 文库外的其他三个文库进行 1:1 等量混合，而 *Mbo* II+*Bcc* I 文库则为其它文库上样量的两倍，将该混合文库使用 flow cell 一个 lane 的测序量进行上机测序。测序使用 Iillumina®公司的 HiSeq2000™ 测序系统完成，操作完全按照相应的操作指导进行。

数据分析主要按照 jun wang *et al.*, Nature(2008)（J Wang, *et al.*, (2008).The diploid genome sequence of an Asian individual. *Nature*,456:60.）中描述的方法操作，由于双向测序，所以通过设定成对测序读长的方向及间隔距离参数（50 bp－2000 bp）对原始数据进行过滤，满足条件的测序读长以成对进行比对，不满足的则以单独的测序读长进行比对，比对方法可以使用 SOAP v2.20 将测序读长比对到参考序列 hg18 上，比对过程允许有两个碱基的错配，计算所有测序读长可以比对到参考序列上的比例。最后再检测这些可以比对上的读长有多少比例可以落在不同酶切组合结果（表 1 所示）的目标区域上，以及目标区域的覆盖度和覆盖深度等数据，结果如表 4 所示。

表 4: 数据分析结果

建库用酶或组合	*Mbo* II	*Tsp* 45I	*Mbo* II-*Hin*d III	*Mbo* II-*Bcc* I
测序总读数	20406253	16964596	19182040	35838376
收获数据量（Mb）	3673	3054	3453	6451
可比对到基因组的碱基数（比例）	2863709280 (78.0%)	2137535730 (70.0%)	2707424190 (78.4%)	5241764970 (81.3%)

可比对到目标区域的碱基数（比例）	1867134613 (65.2%)	1232551200 (57.7%)	1717052782 (63.4%)	3873866058 (73.9%)
目标区域的覆盖度	81.10%	89.00%	72.60%	87.70%
目标区域平均覆盖深度	3.13	4.67	2.675	4.643

由最终数据结果可以看出，选用的 4 个酶切组合最后结果基本一致，除去测序上样量加倍的 *Mbo* II-*Bcc* I 组合，其余三个测序文库都产生 3Gb－4Gb 的数据量，而这些序列有 70%－80%可以比对到基因组中，而这其中又有 57%－73%的数据可以比对在目标区域，最后与表 1 所示结果相比，72%－90%的目标区域被测序所覆盖，且平均的覆盖深度为 3×－5×，由此可见，该方法使用较好的酶切组合可以得到约 90%的目标区域，而且与表 1 所示的结果相比，应用不同的酶切组合的一致性较好。

实施例 3：使用 *Tsp* 45I 酶切建库的 SNPs 检测和基因分型

为了检测对于不同样本间的平行性，以及实际的 SNPs 位点检测情况，本实施例中除了使用炎黄一号（标注为 YH）基因组外，选用了另一个健康男性（标注为 DY）基因组进行平行实验。按照与实施例 2 中类似的方法，用 *Tsp* 45I 酶分别构建两个 DNA 文库：YH 文库（YH 文库 trial2）和 DY 文库。

SNP 的检测使用 SOAPsnp 程序，按照 Q20. mean quality of best allele>20.copy number≤1.1 的过滤参数进行过滤，最后统计实际得到的 SNPs 数目，以及这些位点占 dbSNP 数据库的比例。同时，根据炎黄一号全基因组已有的 SNP 位点信息（Ruiqiang Li *et al.*,(2010).SNP detection for massively parallel whole-genome resequencing. Genome Research,19:1124），选取以 *Tsp* 45I 酶切建库的目标区域范围内的 SNP 位点信息，与本实施例中鉴定的 SNP 位点相比较，计算实际检测到的 SNPs 位点占已有结果的比例。

具体地，将使用 *Tsp* 45I 独立构建两个文库的测序数据与 hg18 基因组序列为参考进行比对，使用这些可以正确比对到参考基因组的测序序列，统计了插入片段的长度分布，结果显示，无论使用 DY 基因组（图 4）还是 YH 基因组（图 5）构建的文库，插入片段都正常分布在 200bp－700bp 之间，这与最初的实验设计和操作是一致的，而且两个文库间，在该片段长度范围（X 坐标）内测序数据分布比例（Y 坐标）也比较一致。此外，统计了两个文库测序数据的分布情况，其中 DY 文库（图 6）平均的测序深度为 11×左右，而 YH 文库（图 7）平均测序深度达到 20×，而且二者的深度分布基本近似于泊松分布，而 DY 文库由于测序最后得到的数据量较 YH 文库要小，所以其测序深度较低。

进一步的数据统计分析结果如下表 5 所示，其中，两个文库上机后分别得到了 4.5Gb 和 7.8Gb 的测序原始数据，这其中分别有 76.8%和 84.6%分别可以比对到 hg18 参考基因组上，在正确比对上这部分的数据中，分别有 80.9%和 78.5%是正确位于目标区域的，而统计目标区域被至少一个测序数据所覆盖的比例，两个文库中分别为 91.9%和 95.2%。由该数据结果可以看出，使用该限制性内切酶建库的方法，可以稳定得到 90%以上的目标区域，而且测序数据的比对率都在正常范围内。

表 5: 初步数据分析结果

基因组文库	DY 文库	YH 文库
总测序收获数据量(Mb)	4540	7885
比对到参考基因组的数据量 Mb(比例)	3482(76.84%)	6664(84.62%)
可比对到目标区域 Mb(比例)	2818（80.93%）	5228（78.47%）
至少被覆盖一次的目标区域 Mb(比例)	242 (91.91%)	251 (95.24%)

为了进一步比较该建库方法的平行性，以目标区域中不同碱基的覆盖深度为参考，分别选取使用 *Tsp* 45I 构建的三个文库进行了两两间的比较，分别比较了 YH 文库和 DY 文库（图 8）和两次构建的 YH 样品文库（图 9，“YH 文库 trial 1”表示实施例 2 中构建的文库，“YH 文库 trial2”表示实施例 3 中构建的文库）的平行性，其中 X 轴和 Y 轴分别对应不同的样品或不同实验批次（如图 8 和 9 中标注），其坐标是按照不同的覆盖深度由小到大分为相应的区间等级，由 1 至 10 表示由低到高的覆盖深度。Z 轴表示的是位于该深度区间的碱基数目，由图 8 和 9 中可以看出，无论是使用不同的样品还是不同的批次，建库的平行性都较好，大部分碱基在两个库中被覆盖的深度也基本一致。

同时，分析了相互比较的文库之间目标区域被共同覆盖的情况显示，两次构建的三个文库一致性较好，其中有 3%的目标区域在相互比较的两个文库中都没有测序数据覆盖，而被覆盖的目标区域有 90%是一致的，此外大约 7%的目标区域仅在一个库中被覆盖，说明，该方法建库的平行性在 93%以上。

由于第二次构建的 YH 文库，平均测序深度达到了 20×，所以，使用此次数据进行了 SNP 检测，使用 SOAPsnp 软件，以 Q20. mean quality of best allele>20.copy number ≤1.1 为过滤参数，以 hg18 为参考基因组序列，一共得到了 264K 的 SNPs 位点信息，通过与已发表的 YH 基因组 SNPs 位点信息比较，应该有 294K 的 SNPs 位点位于 *Tsp* 45I 酶切后测序的目标区域内，而本次实验得到的 SNPs 位点中有 219K（74.6%）为一致的，其中假阳性有 44K（17%），假阴性为 74K（25%），通过分析确定，假阳性中有 28K（65%）位点虽然在已报道的 YH 基因组中并未检测到，但是在 dbSNP 数据库中是被收录的，说明这部分可能是在 YH 的参考 SNP 数据集中因某种原因被过滤掉，而在本实验中被正确的检测出来，所以，除去这部分原因，假阳性率也可以控制在合理范围内。而假阴性部分有约 21K（28%）是由于 SNP 位于限制性内切酶的识别位点内，最后导致了酶无法识别和切割而丢掉了该目标区域片段及 SNP 位点信息，而另外大部分则是因为测序深度不够或者该位点测序质量值不高导致的，这部分与本方法无关，可以在后续实验中通过提高测序量来进一步优化。

为了进一步验证该方法得到 SNP 位点的准确性，将此次得到的数据与使用目前主流的基因分型芯片（Iillumina 1M BeadChip）对 YH 基因组的分型信息比较，在芯片上涵盖的约 1M 的 SNPs 位点有 100K 位于本方法的目标区域内，而使用本方法覆盖了约 98K（90%），在共同覆盖的部分，其中对于纯合位点的一致率达到 99%以上，而杂合位点的一致率为 92%，准确率和覆盖度都较好。

由以上结果可以看出，通过根据本发明实施例的构建 DNA 文库、确定 DNA 序列信息及检测 SNP 位点的方法可以有效地得到预先模拟（表 1）90%以上的目标区域片段，并成功且准确检测该区域内大部分的 SNPs 位点信息，这些 SNP 信息可以用于后续的基因分型或者 GWAS 研究中。

工业实用性

本发明的 DNA 文库及其制备方法、确定 DNA 序列信息的方法、检测 SNPs 的装置

和试剂盒、以及基因分型方法，能够应用于 DNA 测序，进而应用于 SNPs 检测以及基因分型，并且能够有效地提高测序平台，例如 Solexa 测序平台的测序通量。

尽管本发明的具体实施方式已经得到详细的描述，本领域技术人员将会理解。根据已经公开的所有教导，可以对那些细节进行各种修改和替换，这些改变均在本发明的保护范围之内。本发明的全部范围由所附权利要求及其任何等同物给出。

在本说明书的描述中，参考术语“一个实施例”、“一些实施例”、“示意性实施例”、“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的一个或多个实施例或示例中以合适的方式结合。

权利要求书

1. 一种制备 DNA 文库的方法，包括如下步骤：

使用限制性内切酶，对样本基因组 DNA 进行酶切，以便获得酶切产物，其中所述限制性内切酶包括选自 *Mbo* II 和 *Tsp* 45I 的至少一种；

将所述酶切产物进行分离，以便获得长度为 100 bp－1,000 bp 的 DNA 片段；

将所述 DNA 片段进行末端修复，以便获得经过末端修复的 DNA 片段；

在所述经过末端修复的 DNA 片段的末端添加碱基 A，以便获得具有末端碱基 A 的 DNA 片段；以及

将所述具有末端碱基 A 的 DNA 片段与测序接头连接，以便获得所述 DNA 文库。

2、根据权利要求 1 所述的方法，其中所述限制性内切酶进一步包括选自 *Hind* III 和 *Bcc* I 的至少一种。

3、根据权利要求 1 所述的方法，其中所述限制性内切酶为选自下列的至少一组：

（1）*Mbo* II；

（2）*Tsp* 45I；

（3）*Mbo* II 和 *Hind* III；以及

（4）*Mbo* II 和 *Bcc* I。

4、根据权利要求 1 所述的方法，其中，通过琼脂糖凝胶电泳和切胶回收将所述酶切产物进行分离。

5、根据权利要求 1 所述的方法，其中所述 DNA 片段的长度为 200 bp－700 bp。

6、一种 DNA 文库，其是根据权利要求 1 至 5 任一项所述的方法构建的。

7. 一种确定 DNA 序列信息的方法，其特征在于包括以下步骤：

根据权利要求 1-5 任一项所述的方法构建所述 DNA 的 DNA 文库；以及

对所述 DNA 文库进行测序，以便获得所述 DNA 序列信息。

8、根据权利要求 7 所述的方法，其特征在于，利用选自 Illumina、Roche 454 以及 SOLiD 测序平台对所述 DNA 文库进行测序。

9. 根据权利要求 7 所述的方法，其特征在于，进一步包括对所述 DNA 序列信息进行 SNPs 数据分析的步骤，以便获得所述 DNA 的 SNPs 信息。

10. 一种用于检测SNPs的装置，包括如下单元：

DNA文库制备单元，所述DNA文库制备单元用于制备DNA文库；

测序单元，所述测序单元与所述DNA文库制备单元相连，用于对所述DNA文库进行测序，以便获得DNA序列信息；以及

SNPs数据分析单元，所述SNPs数据分析单元与所述测序单元相连，用于对所述DNA序列信息进行SNPs数据分析，以便获得SNPs信息。

11. 一种用于检测SNPs的试剂盒，其包括：

限制性内切酶，所述限制性内切酶包括选自*Mbo* II和*Tsp* 45I的至少一种。

12、根据权利要求 11 所述的试剂盒，其中所述限制性内切酶进一步包括选自 *Hind* III 和 *Bcc* I 的至少一种。

13、根据权利要求 11 所述的试剂盒，其中所述限制性内切酶为选自下列的至少一组：

（1）*Mbo* II；

（2）*Tsp* 45I；

（3）*Mbo* II 和 *Hind* III；以及

（4）*Mbo* II和*Bcc* I。

14. 一种基因分型方法，其包括：

提供样本基因组；

根据权利要求1-5所述的方法，制备所述样本基因组的DNA文库；
对所述DNA文库进行测序，以便获得所述DNA序列信息；
对所述DNA序列信息进行SNPs数据分析，以便获得所述样本的SNPs信息；以及基于所述SNPs信息对所述样本进行基因分型。

5

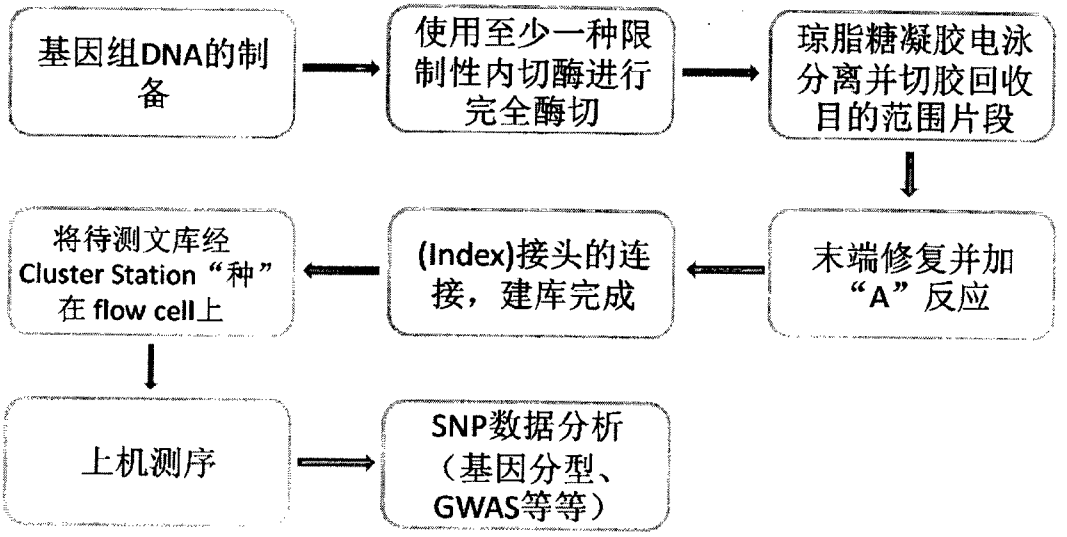

基因组DNA的制备
使用至少一种限制性内切酶进行完全酶切
琼脂糖凝胶电泳分离并切胶回收目的范围片段
末端修复并加“A”反应
(Index)接头的连接，建库完成
将待测文库经Cluster Station“种”在flow cell上
上机测序
SNP数据分析（基因分型、GWAS等等）

图 1

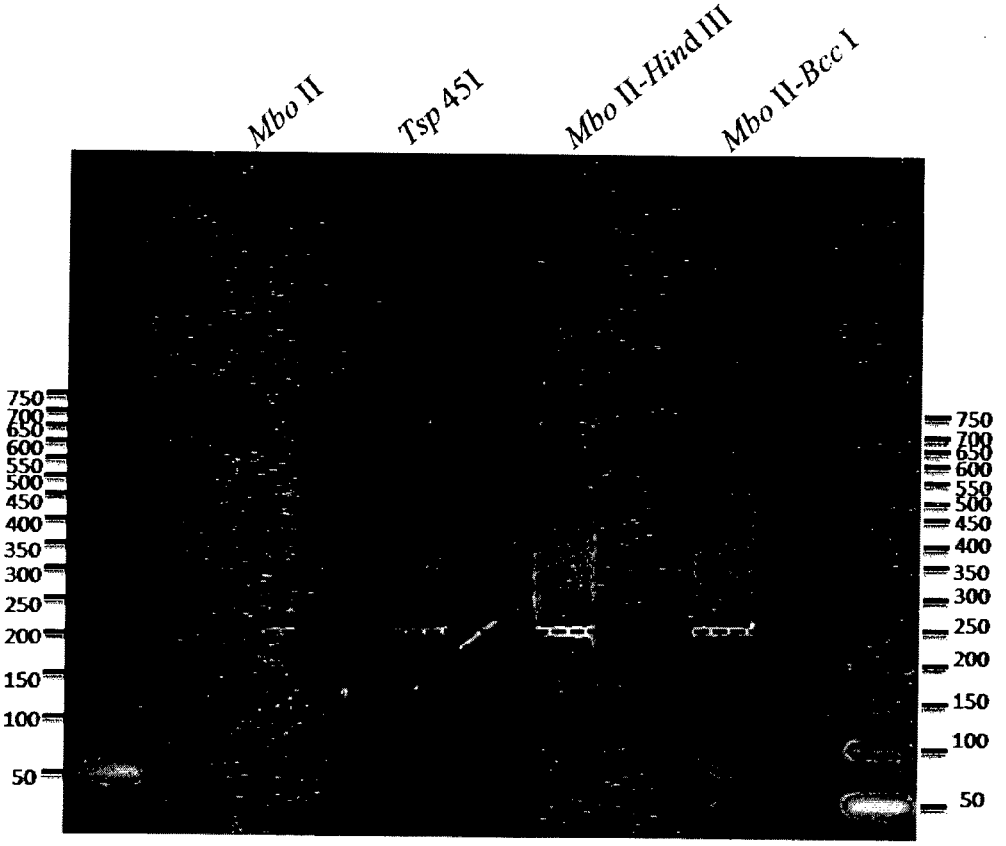
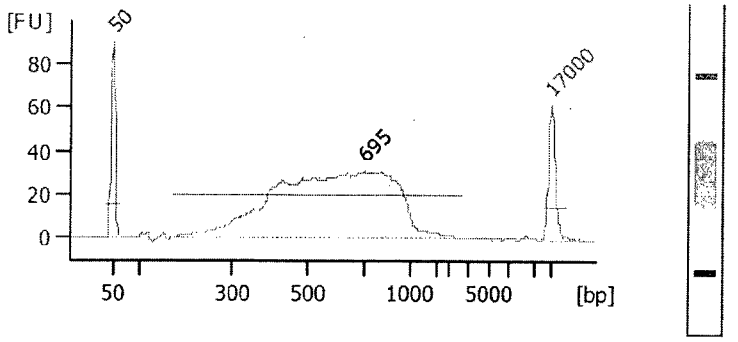
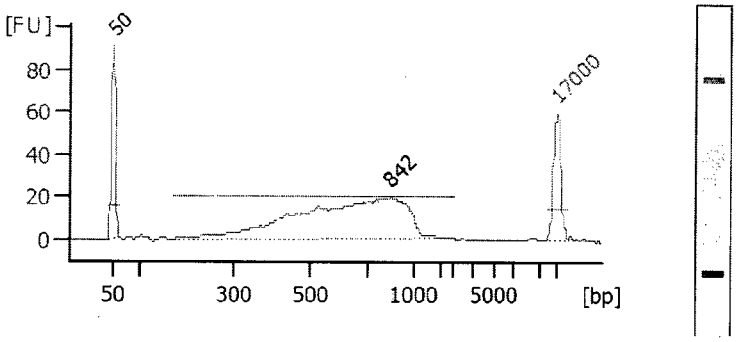

Mbo II
Tsp 45I
Mbo II-Hind III
Mbo II-Bcc I
750
700
650
600
550
500
450
400
350
300
250
200
150
100
50

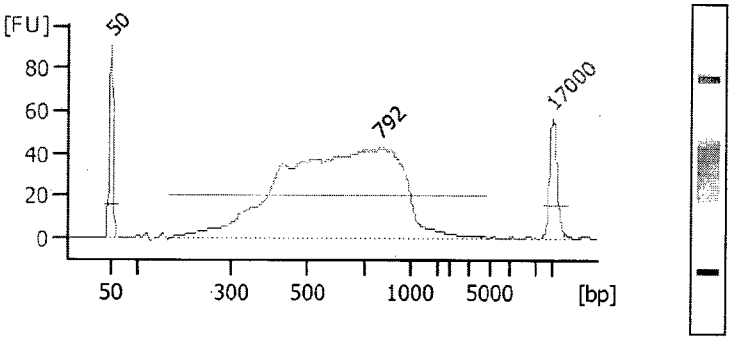
图 2


[FU]
80
60
40
20
0
50
695
17000
50
300
500
1000
5000
[bp]

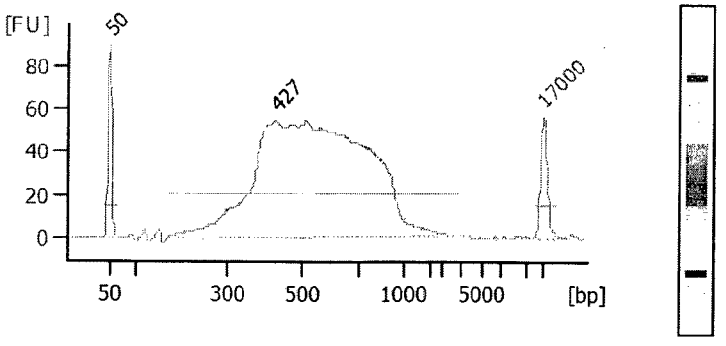
A


[FU]
80
60
40
20
0
50
842
17000
50
300
500
1000
5000
[bp]

B


[FU]
80
60
40
20
0
50
792
17000
50
300
500
1000
5000
[bp]

C


[FU]
80
60
40
20
0
50
427
17000
50
300
500
1000
5000
[bp]

D

图 3

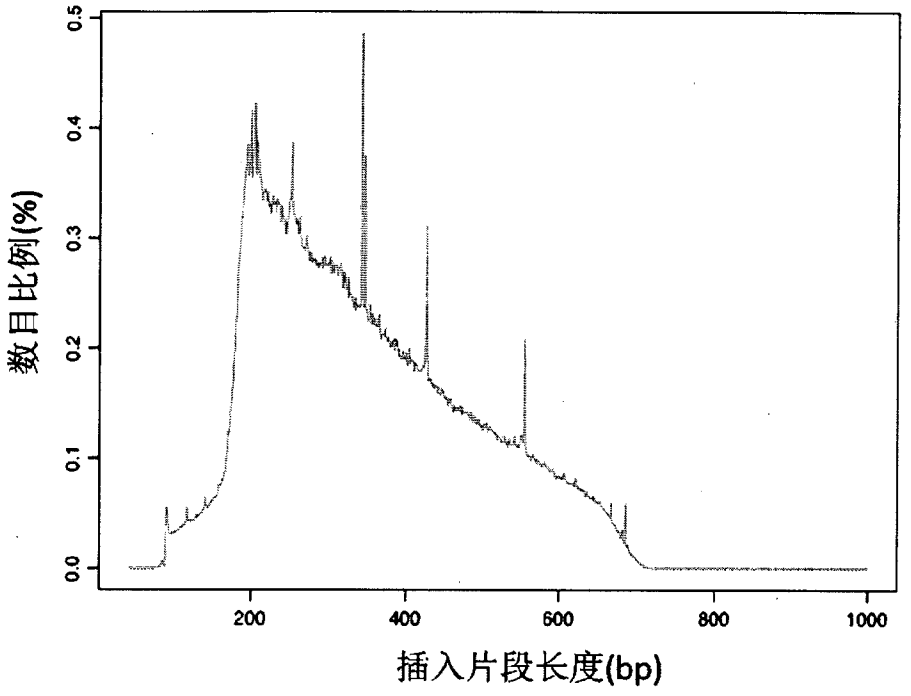

数目比例(%)
0.0
0.1
0.2
0.3
0.4
0.5
200
400
600
800
1000
插入片段长度(bp)

图4

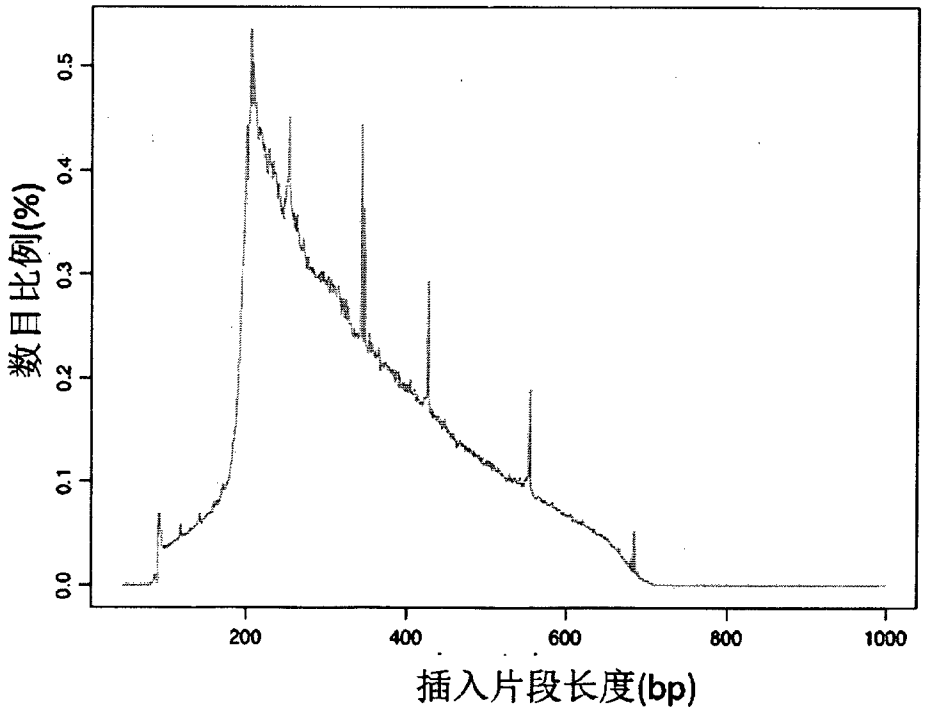

数目比例(%)
0.0
0.1
0.2
0.3
0.4
0.5
200
400
600
800
1000
插入片段长度(bp)

图5

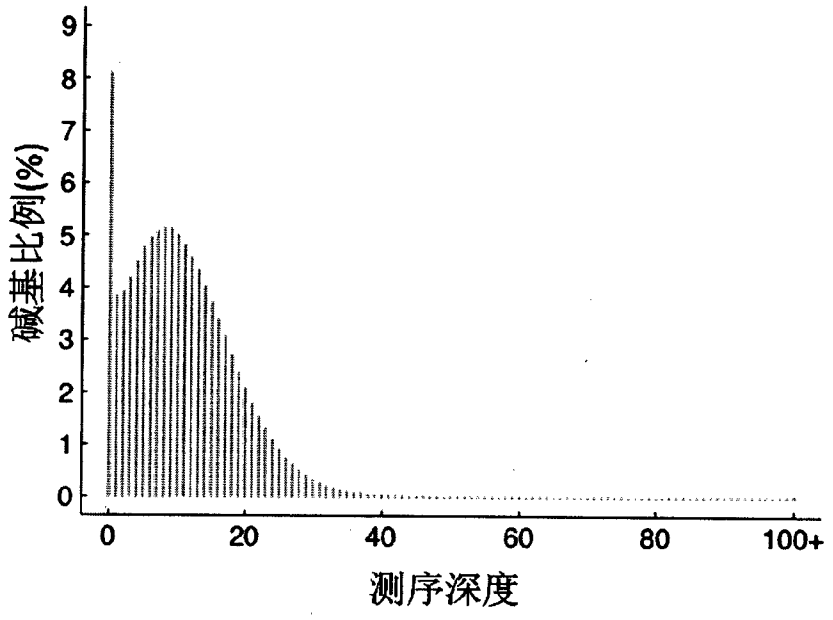

碱基比例(%)
9
8
7
6
5
4
3
2
1
0
0
20
40
60
80
100+
测序深度

图6

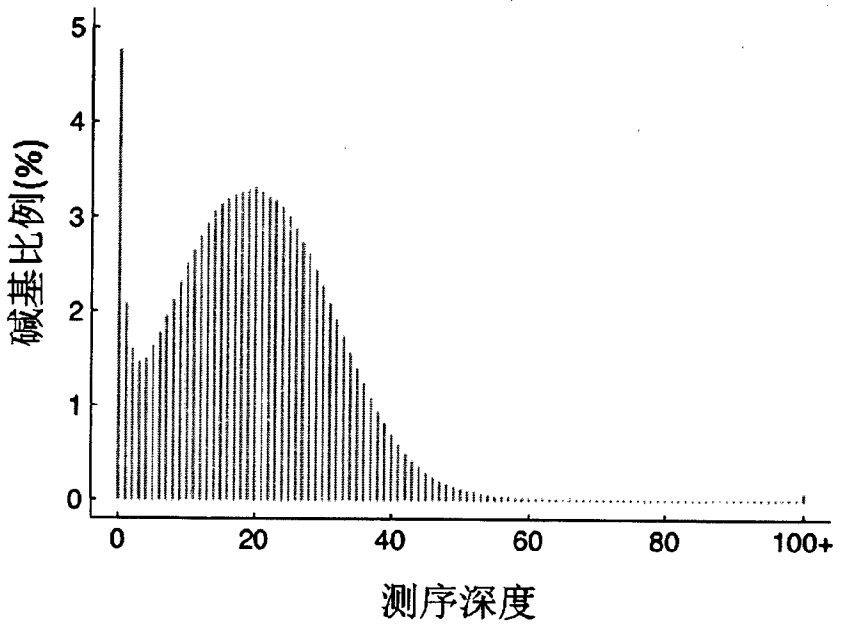

碱基比例(%)
5
4
3
2
1
0
0
20
40
60
80
100+
测序深度

图7

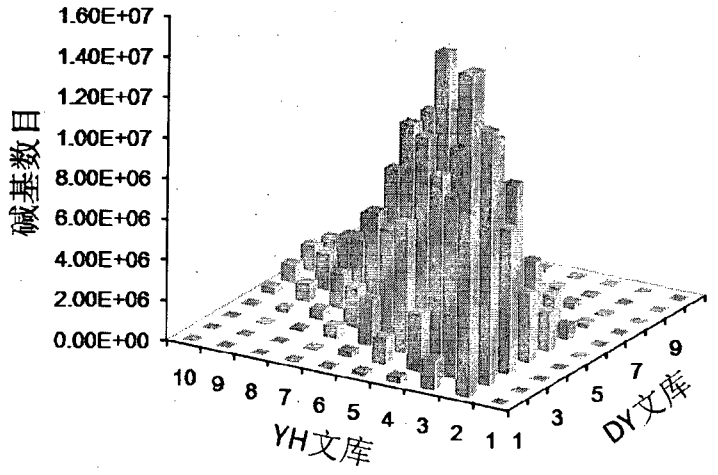

1.60E+07
1.40E+07
1.20E+07
1.00E+07
8.00E+06
6.00E+06
4.00E+06
2.00E+06
0.00E+00
碱基数目
10 9 8 7 6 5 4 3 2 1
YH文库
1 3 5 7 9
DY文库

图8

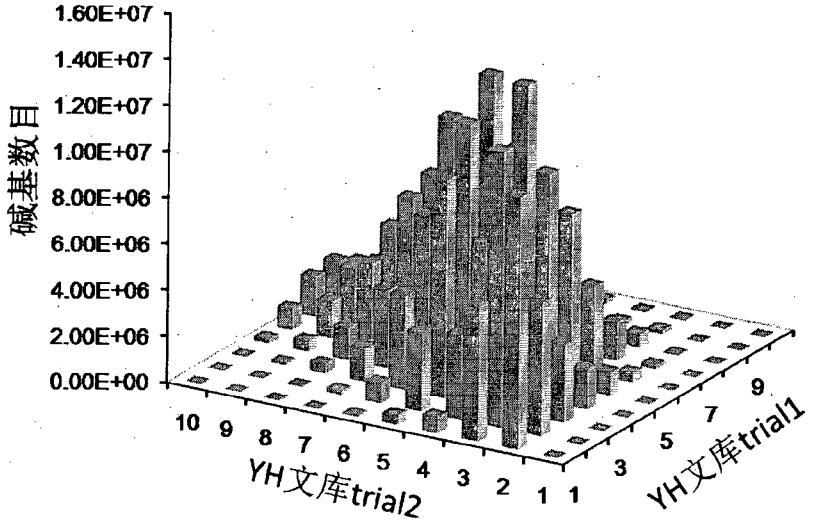

1.60E+07
1.40E+07
1.20E+07
1.00E+07
8.00E+06
6.00E+06
4.00E+06
2.00E+06
0.00E+00
碱基数目
10 9 8 7 6 5 4 3 2 1
YH文库trial2
1 3 5 7 9
YH文库trial1

图9

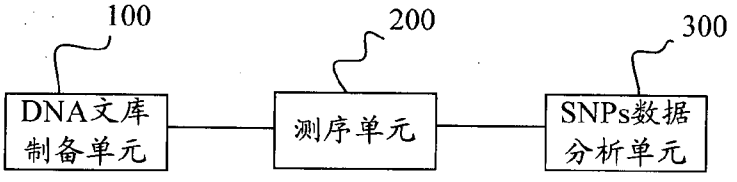

100
200
300
DNA文库
制备单元
测序单元
SNPs数据
分析单元

图10

INTERNATIONAL SEARCH REPORT

International application No.
PCT/CN2011/079971

A. CLASSIFICATION OF SUBJECT MATTER

See the extra sheet

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)

IPC: C40B, C12N, 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)
CPRSABS, CNABS, DWPI, SIPOABS, CNKI, BA, CA, PUBMED: Mbo II, Tsp45I, genomic DNA, single nucleotide polymorphism、single nucleotide polymorphisms、SNP, SNPs, Hind III, Bcc I, sequencing、restriction enzyme、restriction enzymes、DNA library、DNA libraries、kit、kits

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
PX	CN102061526A (SHENZHEN BGI TECHNOLOGY CO LTD), 18 May 2011 (18.05.2011), see the whole document	1-14
X	KR20090033307A (UNIV SEOUL NAT IND FOUND), 02 Apr. 2009 (02.04.2009), see abstract and claims 1-9	11,13
A	Nathan A, et al., Rapid SNP discovery and genetic mapping using sequenced RAD markers, PLoS ONE, 3(10), 30 Oct. 2008(30.10.2008), pe3376, see the whole document	1-14
A	WO2009126395A1 (TRANSGENOMIC INC),	1-14

☒ 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	Date of mailing of the international search report
23 Nov. 2011 (23.11.2011)	**15 Dec. 2011 (15.12.2011)**
Name and mailing address of the ISA 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 SONG, Zhigang Telephone No. (86-10) 62411078

Form PCT/ISA /210 (second sheet) (July 2009)

INTERNATIONAL SEARCH REPORT

International application No.
PCT/CN2011/079971

C (Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
	15 Oct. 2009 (15.10.2009), see the whole document	
A	CN101845489A (UNIV CHINA OCEAN), 29 Sep. 2010 (29.09.2010), see the whole document	1-14
A	CN101230490A (UNIV BEIJING FORESTRY), 30 July 2008 (30.07.2008), see the whole document	1-14
A	CN101343667A (HUANGHAI SEA AQUATIC INST CHINESE ACAD), 14 Jan. 2009 (14.01.2009), see the whole document	1-14
A	CN1341750A (UNIV HUBEI), 27 Mar. 2002 (27.03.2002), see the whole document	1-14
A	WO2010091111A1 (BIOHELIX CORP), 12 Aug. 2010 (12.08.2010), see claims 1-29	11-13

INTERNATIONAL SEARCH REPORT

International application No.
PCT/CN2011/079971

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.:
because they relate to subject matter not required to be searched by this Authority, namely:

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).

Box No. III Observations where unity of invention is lacking (Continuation of item 3 of first sheet)

This International Searching Authority found multiple inventions in this international application, as follows:

See extra sheet

1. ☐ As all required additional search fees were timely paid by the applicant, this international search report covers all searchable claims.

2. ☒ As all searchable claims could be searched without effort justifying additional fees, this Authority did not invite payment of additional fees.

3. ☐ As only some of the required additional search fees were timely paid by the applicant, this international search report covers only those claims for which fees were paid, specifically claims Nos.:

4. ☐ No required additional search fees were timely paid by the applicant. Consequently, this international search report is restricted to the invention first mentioned in the claims; it is covered by claims Nos.:

Remark on protest

☐ The additional search fees were accompanied by the applicant's protest and, where applicable, the payment of a protest fee.

☐ The additional search fees were accompanied by the applicant's protest but the applicable protest fee was not paid within the time limit specified in the invitation.

☐ No protest accompanied the payment of additional search fees.

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.
PCT/CN2011/079971

Patent Documents referred in the Report	Publication Date	Patent Family	Publication Date
CN102061526A	18-05-2011	None	
KR20090033307A	02-04-2009	None	
WO2009126395A1	15-10-2009	US2009068659A1	12-03-2009
		US7579155B2	25-08-2009
		EP2276860A1	26-01-2011
CN101845489A	29-09-2010	None	
CN101230490A	30-07-2008	None	
CN101343667A	14-01-2009	None	
CN1341750A	27-03-2002	CN1188526C	09-02-2005
WO2010091111A1	12-08-2010	None	

INTERNATIONAL SEARCH REPORT

International application No.

PCT/CN2011/079971

A. CLASSIFICATION OF SUBJECT MATTER

C40B 50/06 (2006.01) i

C40B 40/08 (2006.01) i

C40B 20/04 (2006.01) i

C12N 15/11 (2006.01) i

C12Q 1/68 (2006.01) i

Box No. III Observations where unity of invention is lacking

This International Searching Authority found three groups of inventions in this international application, as follows:

1. Claims: 1-9, 14

related to a method for constructing a DNA library, comprising performing enzyme cutting for sample genomic DNA using at least one restriction enzyme selected from Mbo II and Tsp 45I; a DNA libray constructed by the said method; a method for determining DNA sequence information by sequencing the DNA libray constructed by the said method; a method for genotyping sample genomic by sequencing the DNA libray constructed by the said method;

2. Claim: 10

related to a device for detecting SNPs, comprising the following units: DNA library constructing unit, wherein the said DNA library constructing unit is used to construct DNA library; sequence measuring unit, wherein the sequence measuring unit is connected with the said DNA library to sequence the said DNA library in order to obtain the DNA sequence information; and the SNPs data analyzing unit, wherein the said SNPs data analyzing unit is connected with the said sequence measuring unit to perform the data analyze for the said DNA sequence information to obtain the SNPs information;

3. claims: 11-13

related to a kit for detecting SNPs, comprising at least one restriction enzyme selected from Mbo II and Tsp 45I;

Groups 1-3 do not have the same or corresponding special technical feature, and do not linked by a single general inventive concept. Therefore, they do not meet the requirement of unity of invention in accordance with Rules 13.1, 13.2 and 13.3.

国际检索报告

国际申请号
PCT/CN2011/079971

A. 主题的分类

参见附加页

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)

IPC：C40B、C12N、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）)

CPRSABS、CNABS、DWPI、SIPOABS、CNKI、BA、CA、PUBMED：Mbo II、Tsp45I、基因组 DNA、单核苷酸多态性、SNP、SNPs、测序、内切酶、DNA 文库、、Hind III、Bcc I、试剂盒、genomic DNA、single nucleotide polymorphism、single nucleotide polymorphisms、sequencing、restriction enzyme、restriction enzymes、DNA library、DNA libraries、kit、kits

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
PX	CN102061526A（深圳华大基因科技有限公司），18.5 月 2011（18.05.2011），参见全文	1－14
X	KR20090033307A（UNIV SEOUL NAT IND FOUND），02.04 月 2009（02.04.2009），参见摘要和权利要求 1－9	11，13
A	Nathan A，等，Rapid SNP discovery and genetic mapping using sequenced RAD markers，PLoS ONE，3（10），30.10 月 2008（30.10.2008），pe3376，参见全文	1－14
A	WO2009126395A1（TRANSGENOMIC INC），15.10 月 2009（15.10.2009），参见全文	1－14
A	CN101845489A（中国海洋大学），	1－14

☒ 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	国际检索报告邮寄日期
23.11 月 2011（23.11.2011）	**15.12 月 2011 (15.12.2011)**

ISA/CN 的名称和邮寄地址：	受权官员
中华人民共和国国家知识产权局 中国北京市海淀区蓟门桥西土城路 6 号 100088 传真号：　(86-10)62019451	宋智刚 电话号码：(86-10) **62411078**

C(续). 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
	29.9 月 2010（29.10.2010），参见全文	
A	CN101230490A（北京林业大学）	1－14
	30.7 月 2008（30.07.2008），参见全文	
A	CN101343667A（中国水产科学研究院黄海水产研究所）， 14.1 月 2009（14.01.2009），参见全文	1－14
A	CN1341750A（湖北大学）， 27.3 月 2002（27.03.2002），参见全文	1－14
A	WO2010091111A1（BIOHELIX CORP）， 12.08 月 2010（12.08.2010），参见权利要求 1－29	11－13

国际检索报告

国际申请号
PCT/CN2011/079971

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：

因为它们涉及不要求本单位进行检索的主题，即：

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。

第III 栏　缺乏发明单一性的意见(续第 1 页第 3 项)

本国际检索单位在该国际申请中发现多项发明，即：

参见附加页

1. ☐ 由于申请人按时缴纳了被要求缴纳的全部附加检索费，本国际检索报告涉及全部可作检索的权利要求。

2. ☒ 由于无需付出有理由要求附加费的劳动即能对全部可检索的权利要求进行检索，本单位未通知缴纳任何附加费。

3. ☐ 由于申请人仅按时缴纳了部分被要求缴纳的附加检索费，本国际检索报告仅涉及已缴费的那些权利要求。

具体地说，是权利要求：

4. ☐ 申请人未按时缴纳被要求缴纳的附加检索费。因此，本国际检索报告仅涉及权利要求书中首先提及的发明；包含该发明的权利要求是：

关于异议的说明：

☐ 申请人缴纳了附加检索费，同时提交了异议书，适用时，缴纳了异议费。

☐ 申请人缴纳了附加检索费，同时提交了异议书，但未在通知书规定的时间期限内缴纳异议费。

☐ 缴纳附加检索费时未提交异议书。

国际检索报告

关于同族专利的信息

国际申请号
PCT/CN2011/079971

检索报告中引用的专利文件	公布日期	同族专利	公布日期
CN102061526A	18-05-2011	无	
KR20090033307A	02-04-2009	无	
WO2009126395A1	15-10-2009	US2009068659A1	12-03-2009
		US7579155B2	25-08-2009
		EP2276860A1	26-01-2011
CN101845489A	29-09-2010	无	
CN101230490A	30-07-2008	无	
CN101343667A	14-01-2009	无	
CN1341750A	27-03-2002	CN1188526C	09-02-2005
WO2010091111A1	12-08-2010	无	

A. 主题的分类

C40B 50/06 (2006.01) i

C40B 40/08 (2006.01) i

C40B 20/04 (2006.01) i

C12N 15/11 (2006.01) i

C12Q 1/68 (2006.01) i

第III栏　缺乏发明单一性的意见

本国际检索单位发现权利要求书中包括三项发明，如下所示：

发明 1：权利要求 1－9 和 14；

涉及一种制备 DNA 文库的方法，包括使用选自 Mbo II 和 Tsp 45I 中的至少一种限制性内切酶对样本基因组 DNA 进行酶切；一种由该方法所构建的 DNA 文库；一种通过对上述方法所构建的 DNA 文库进行测序来确定 DNA 序列信息的方法；通过对上述方法所构建的 DNA 文库进行测序，从而对样本基因组进行基因分型的方法；

发明 2：权利要求 10；

涉及一种用于检测 SNPs 的装置，包括如下单元：DNA 文库制备单元，所述 DNA 文库制备单元用于制备 DNA 文库；测序单元，所述测序单元与所述 DNA 文库相连，用于对所述 DNA 文库进行测序，以便获得 DNA 序列信息；以及 SNPs 数据分析单元，所述 SNPs 数据分析单元与所述测序单元相连，用于对所述 DNA 序列信息进行 SNPs 数据分析，以便获得 SNPs 信息；

发明 3：权利要求 11－13

涉及一种用于检测 SNPs 的试剂盒，其包括选自 Mbo II 和 Tsp 45I 中的至少一种限制性内切酶；

发明 1－3 之间没有相同或相应的特定技术特征，这些发明不能相互关联，从而不能形成一个总的发明构思。因此，不符合 PCT 实施细则 13.1、13.2 和 13.3 的规定。