(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 103476946 A
(43)申请公布日 2013.12.25

(21)申请号 201280005358.2

(22)申请日 2012.01.13

(30)优先权数据
61/432,915 2011.01.14 US

(85)PCT申请进入国家阶段日
2013.07.12

(86)PCT申请的申请数据
PCT/NL2012/050022 2012.01.13

(87)PCT申请的公布数据
WO2012/096579 EN 2012.07.19

(71)申请人 关键基因股份有限公司
地址 荷兰瓦赫宁恩

(72)发明人 M·J·T·范艾克

(74)专利代理机构 上海专利商标事务所有限公司 31100
代理人 陶家蓉

(51)Int.Cl.
C12Q 1/68(2006.01)

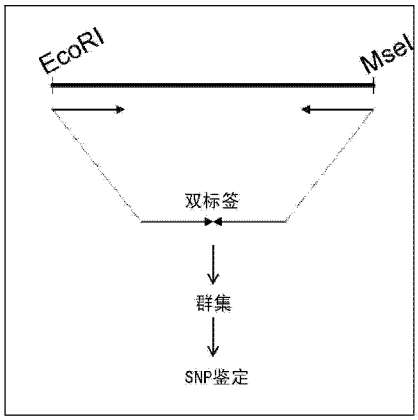
权利要求书2页 说明书10页 附图1页

(54)发明名称

基于配对末端随机序列的基因分型

(57)摘要

通过提供带标识物标签的限制性片段、使用配对高通量测序技术获得序列信息、组合该序列信息并鉴定该样品间多态性来同时发现、检测样品间多态性以及进行基因分型的方法。来自两个末端的序列信息组合能发现、检测高度重复性基因组中的多态性以及进行基因分型。



CN 103476946 A

1. 一种同时发现、检测一个或多个或大量样品中一种或多种多态性以及进行基因分型的方法，所述方法包括以下步骤：

(a) 从一个或多个或大量样品中提供 DNA;

(b) 用至少一种限制性内切酶消化所述 DNA 来降低所述样品 DNA 的复杂性以产生限制性片段；

(c) 将至少一个标识物标签提供给所述样品限制性片段以产生带标签的限制性片段；

(d) 对至少部分的所述带标签限制性片段进行配对末端测序；

(e) 鉴定所述样品间的多态性。

2. 如权利要求 1 所述的方法，其特征在于，所述配对末端测序读数的第一序列读数和第二序列读数组合成双标签，优选在计算机中进行。

3. 如权利要求 1 或 2 所述的方法，其特征在于，所述第一或第二序列读数之一在组合成双标签前反向互补。

4. 如权利要求 1-3 所述的方法，其特征在于，所述标识物标签通过以下方式提供：

- 将带标签衔接子连接到所述限制性片段上以产生连接有带标签衔接子的限制性片段；

或者

- 用至少一个带标签引物扩增连接有衔接子的限制性片段，所述引物与至少部分衔接子互补，以产生连接有带标签衔接子的限制性片段。

5. 如权利要求 1-4 所述的方法，其特征在于，所述序列根据标识物标签分配到所述样品中。

6. 如权利要求 1-5 所述的方法，其特征在于，在样品间比较所述分配的序列来鉴定样品间序列的多态性。

7. 如权利要求 1-6 所述的方法，其特征在于，在样品间比较所述双标签。

8. 如权利要求 1-7 所述的方法，其特征在于，根据所述鉴定的多态性对所述样品进行基因分型。

9. 如权利要求 1-8 所述的方法，其特征在于，所述复杂性降低包括用两种或多种限制性内切酶消化所述样品 DNA 以产生限制性片段。

10. 如权利要求 1-9 所述的方法，其特征在于，在所述限制性片段的一个末端或两个末端连接衔接子以提供连接有衔接子的片段。

11. 如权利要求 9 或 10 所述的方法，其特征在于，所述用不同限制性酶获得的限制性片段的各末端上连接不同的衔接子。

12. 如权利要求 10 或 11 所述的方法，其特征在于，所述复杂性降低还包括用至少一种引物扩增连接有衔接子的片段，所述引物至少与部分所述衔接子互补。

13. 如权利要求 12 所述的方法，其特征在于，所述引物还与限制性内切酶识别序列剩余部分的至少一部分互补。

14. 如权利要求 13 所述的方法，其特征在于，所述引物还包含位于引物 3’端的一种或多种随机选择性核苷酸。

15. 如权利要求 13 所述的方法，其特征在于，所述引物包含位于引物 3’端的一种或多种相同随机选择性核苷酸以用于一种或多种样品。

16. 如前述权利要求中任一项所述的方法，其特征在于，所述测序基于高通量测序。

17. 如权利要求 13 所述的方法，其特征在于，所述高通量测序基于焦磷酸测序，优选在固体载体上。

16. 如权利要求 13 所述的方法，其特征在于，所述高通量测序基于连接测序或纳米孔测序。

基于配对末端随机序列的基因分型

[0001] 发明背景

[0002] 目前使用的大部分标记物发现和基因分型技术主要依赖于两种不同的系统，一种是最初发现的SNP，另一种是之后对大量个体进行的基因分型。这激发本申请人开发一种基于序列的同步标记物发现和检测技术，称为基于随机序列的基因分型（rSBG）。该技术引入了 Illumina GAII 的高通量测序能力以及AFLP® (EP534858) 的基因组复杂性减低能力。其示例描述在本申请人的 WO2007073165 中。与各种通常经靶向的其它基因分型技术（即，先选择待检测 SNP，并用特异性检测探针靶向）相反，rSBG 是随机方法。原则上，当 AFLP 模板中含有特异性序列时可鉴定品系间存在的所有 SNP（通常在使用严格的挖掘过滤（mining filter）后）。其中一个问题是当分析样品来自含有相对大部分重复性序列的基因组时，例如辣椒，鉴定品系间多态性会由于重复性序列的存在而变得更难。

发明内容

[0003] 本发明人发现可改进在多种样品中评分和基因分型的多态性数目，特别是在使用来自视作高度重复性（即包含许多重复）的基因组样品，使用高通量测序方法来测序限制性片段的两个末端时。通过采用称为配对末端测序的方法，从相同限制性片段中获得两组序列数据（即序列读数），各来自限制性片段的一端。通过组合这些数据组，原来由于例如源自重复片段从而不能相互区分的来自限制性片段的序列数据（序列读数）现在变得可区分了。原因是来自位于上百或甚至上千个核苷酸之外限制性片段另一端的序列读数（通常依赖于所用的限制性酶或片段化方法），能产生独特的组合序列读数（参见图 1）。这也能发现、检测来自高度重复性基因组样品的多态性以及进行基因分型。因此本发明方法相比现有技术方法可以最广泛形式应用于更大范围的样品，因为其成功包括了高重复性样品。本发明人发现用该配对末端方法相比单独分析片段各端读数可发现更多 SNP 并进行基因分型，即，由于在 SNP 的同步发现和基因分型中使用配对末端测序而获得协同作用。

[0004] 附图简要说明

[0005] 图 1. 基于配对末端随机序列的基因分型示意图。从限制性片段各端形成双标签（ditag）以实现最大的重复序列分离从而用于 SNP 鉴定。

[0006] 定义

[0007] 在以下说明和实施例中，使用了一些术语。为了提供对说明书和权利要求的清楚和一致的理解，包括给予所述术语的范围，提供了下列定义。除非另外定义，本文中使用的所有技术和科学术语具有本发明所属领域普通技术人员通常所理解的同样含义。所有出版物、专利申请、专利和其他文献的公开内容都通过引用全文纳入本文。

[0008] 本领域技术人员清楚了解本发明方法实施所使用的常规技术。本领域技术人员熟知分子生物学、生物化学、计算化学、细胞培养、重组 DNA、生物信息学、基因组学、测序和相关的领域中的常规技术实践，并在例如以下参考文献中描述：Sambrook 等，《分子克隆．实验室手册》(Molecular Cloning：A Laboratory Manual)，第二版，纽约冷泉港的冷泉港出版社 (Cold Spring Harbor Laboratory Press)，1989；Ausubel 等，《新编分子生

物学实验指南》(Current Protocols in Molecular Biology),纽约的约翰威利父子公司(John Wiley&Sons),1987 及定期更新;和《酶学方法》系列(the series Methods in Enzymology),圣地亚哥的学术出版社(Academic Press)。

[0009] 本文所用的单数形式“一个”、“一种”和“该”包括复数指代形式,除非文中另有明确说明。例如,分离“一个”DNA 分子的方法包括分离多个分子(例如,十、百、千、万、十万、百万或更多的分子)。

[0010] 多态性:多态性指一个群体中存在两种或多种核苷酸序列变体。多态性可以包括一个或多个碱基改变,插入,重复,或缺失。多态性包括例如简单序列重复(SSR)和单核苷酸多态性(SNP),其为 DNA 序列变异,在单个核苷酸:腺嘌呤 (A)、胸腺嘧啶 (T)、胞嘧啶 (C) 或鸟嘌呤 (G) 改变时发生。通常群中至少 1%发生变异被视为 SNP。SNP 占(例如)所有人类遗传变异的 90%,人类基因组中每 100 至 300 个碱基即出现。每三个 SNP 中有两个是腺嘌呤 (T) 取代胞嘧啶 (C)。例如人类和植物的 DNA 序列变异会影响其如何处理疾病、细菌、病毒、化学试剂、药物等。

[0011] 基因分型指测定种类中个体遗传变异的方法。生物基因型是其遗传密码内携带的遗传指令(inherited instruction)。并非所有具有相同基因型的生物表观或行为方式相同,因为表观和行为由环境和发育条件修饰。同样,并非所有看上去很像的生物肯定具有相同的基因型。单核苷酸多态性(SNP)是最常见的遗传变异类型并定义为在高于 1%群中发现的特异性基因座的单碱基差异。在基因组编码以及非编码区域都发现 SNP,其可能导致不同表型,例如,在编码区发现时,具有患病或耐受疾病的能力。因此,SNP 常用作特定疾病或一些表型的标记物。当发现于非编码区时,SNP 用作进化基因组研究的标记物。SNP 涉及不同长度的核苷酸“插入缺失标记(InDel)”或插入和缺失。第三种遗传变异类型是拷贝数变异(CNV),其源于不同基因组中具有不同拷贝数的 DNA 片段。编码基因拷贝数变化的情况中,所述变化会导致对疾病的易感性或抵抗性。一些表型还是剂量敏感性,拷贝数引起种类成员间的不同差异。对于 SNP 和 CNV 基因分型,存在许多测定个体间基因型的方法。选择方法通常取决于通量需要,其随着基因分型的个体数量和各个体所测的基因型数量而变化。所选方法还取决于各个体或样品中可得的样品材料量。

[0012] 基因型是细胞、器官、或个体(即个体的特异性等位基因组成)通常参照考虑中的特异性性状和特性的遗传组成。

[0013] 表型是可观察的生物性状和特性,例如其形态、发育、生化或生理特性、物候学、行为和行为产物。表型来自环境因素的基因表达和影响以及两者间的相互作用。虽然表型是生物表现出的总体可观察特性,术语表型组有时指特性集合,其同步研究称为表型组学。

[0014] 表型分型是测定生物表型。

[0015] 限制性内切核酸酶:限制性内切核酸酶或限制性酶是一种酶,其识别双链 DNA 分子中的特定核苷酸序列(靶位点),并能在每个靶位点处或附近切开两条链,得到钝端或交错末端。

[0016] 限制性片段:用限制性内切核酸酶消化产生的 DNA 分子被称作限制性片段。用特定限制性内切核酸酶消化任何给定基因组(或核酸,不论其来源),得到一组离散的限制性片段。限制性内切酶切割产生的 DNA 片段还可用于各种技术。

[0017] 加标签:术语加标签指在核酸样品中加入标签,从而能区分其与第二或更多核酸

样品。

[0018] 标识物或标识物标签：一种短序列，其可加到衔接子或引物上，或包括在其序列中，或用作标记以提供独特的识别物。这种序列标识物（标签）可以例如是长度变化（通常为 4-16bp）但确定的独特碱基序列。标识物或标识物组合可用于鉴定特异性核酸样品或连接或结合 DNA 产物，例如源自样品的该样品片段或 PCR 产物。例如 4bp 标签能够产生 4^4=256 个不同标签。使用该标识物，可在进一步加工后测定样品来源。在合并源于不同核酸样品的加工产物的情况下，通常可用不同标识物鉴定不同核酸样品。标识物优选彼此至少有两个碱基对的差异，优选不含两个相同的连续碱基，以防止误读。标识物功能有时可与其它功能联合，例如衔接子或引物。

[0019] 标签限制性片段：提供有标识物标签的限制性片段。

[0020] 连接有衔接子的限制性片段：被衔接子封端的限制性片段。

[0021] 衔接子：短双链 DNA 分子，具有有限数量的碱基对，例如长约 10-30 对碱基，设计成可与限制性片段的末端连接。衔接子通常由两种合成寡核苷酸组成，其具有彼此部分互补的核苷酸序列。当在合适条件下于溶液中混合两种合成寡核苷酸时，它们会彼此退火，形成双链结构。退火后，衔接子分子的一端被设计成与限制性片段末端相容，可与其连接；衔接子另一端可设计成不能连接，但不需如此（双连接衔接子）。

[0022] 连接：连接酶催化的酶反应，其中两个双链 DNA 分子彼此共价连接，称作连接。一般来说，两条 DNA 链共价连接，但也可通过化学或酶修饰链末端之一防止两条链之一连接。该情况下，共价连接将仅发生在两条 DNA 链的一条上。

[0023] 引物：一般，术语引物指能够引发 DNA 合成的 DNA 链。DNA 聚合酶不能在没有引物的情况下从头合成 DNA：在一个反应中只能延伸一条已存在的 DNA 链，其中互补链被用作模板，指导装配的核苷酸的顺序。我们将在聚合酶链式反应 (PCR) 中使用的合成寡核苷酸分子称作引物。

[0024] 合成寡核苷酸：可化学合成的具有优选 10-50 个碱基的单链 DNA 分子被称作合成寡核苷酸。一般这些合成 DNA 分子设计成具有独特或所需的核苷酸序列，虽然也可能合成具有相关序列的分子（其在核苷酸序列内的特定位点具有不同核苷酸组成）的家族。术语合成寡核苷酸用于指具有经设计或所需核苷酸序列的 DNA 分子。

[0025] 扩增：术语扩增通常用于指体外合成双链 DNA 分子，一般使用 PCR。应注意存在其它扩增方法，其可在本发明中使用而不违背其要旨。

[0026] 扩增子：多核苷酸扩增反应的产物，即从一个或多个起始序列复制的多核苷酸群。扩增子可通过各种扩增反应产生，包括但不限于聚合酶链反应（PCR）、线性聚合酶反应、基于核酸序列的扩增、滚环扩增等反应。

[0027] 复杂性降低：术语复杂性降低用于指一种方法，其中通过产生样品亚组降低了核酸样品，例如基因组 DNA 的复杂性。该亚组可代表全（即复杂）样品，优选是可重复亚组。可重复在此表示当用相同的方法降低同一样品的复杂性时，可获得相同或至少相当的亚组。用于复杂性减低的方法可以是任何本领域已知的复杂性降低的方法。复杂性降低的方法示例包括例如AFLP® (Keygene N.V., 荷兰；参见例如 EP0534858)，Dong 所述的方法（参见例如 WO03/012118, WO00/24939)，索引连接 (Unrau 等，1994，基因(Gene)，145:163-169) 等。用于本发明的复杂性降低法方法的共同之处是它们都是可重复的。可重复意味着当以

相同方式降低相同样品的复杂性时，获得相同样品亚组，与之截然相反的是更随机的复杂性降低法，例如显微切割或使用 mRNA(cDNA)，其代表了所选组织中一部分转录的基因组，其可重复性取决于组织选择和分离时间等。

[0028] 选择性碱基，选择性核苷酸，随机选择性核苷酸：位于引物的 3’端，选择性碱基随机选自 A、C、T 或 G（或 U，视情况而定）。通过用选择性碱基延伸引物，随后的扩增仅仅得到连接有衔接子的限制性片段的可重复亚组，即仅用携带选择性碱基的引物能扩增出的片段。加到引物 3’端的选择性核苷酸数量可在 1-10 间不等。通常 1-4 足够。两种引物（PCR 中）都可含有不同数量的选择性碱基。使用各加入的选择性碱基情况下，亚组使扩增的连接有衔接子的限制性片段量降低了约 4 倍。该类型复杂性降低视为随机降低，因为其不需要或考虑任何之前序列信息，仅仅基于选择性核苷酸。通常，用于 AFLP 技术（EP534858）的选择性碱基的数目以 +N+M 表示，其中一个引物携带 N 个选择性核苷酸，另一个引物携带 M 个选择性核苷酸。因此，Eco/Mse+1/+2AFLP 简写形式代表用 EcoRI 和 MseI 消化起始 DNA，连接合适的衔接子并扩增，一种引物针对 EcoRI 限制性位点，其携带 1 个选择性碱基，另一种引物针对 MseI 限制性位点，携带 2 个选择性核苷酸。用于 AFLP 的在 3’末端携带至少一种选择性核苷酸的引物也称为 AFLP- 引物。在 3’末端不携带选择性核苷酸，实际上与衔接子和其余限制性位点互补的引物有时被称作 AFLP+0 引物。术语选择性核苷酸也用于目标序列的核苷酸，其毗邻衔接子部分并通过使用选择性引物鉴定，因此该核苷酸被如此称呼。

[0029] 测序：术语测序指测定核酸样品，例如 DNA 或 RNA 中的核苷酸顺序（碱基序列）。可用许多技术例如桑格测序和高通量测序技术（也称作下一代测序技术），例如罗氏应用科学公司（Roche Applied Science）的 GS FLX 平台，和亿明达公司（Illumina）的基因组分析仪（Genome Analyzer），它们都基于焦磷酸测序。还存在其它平台。

[0030] 高通量测序或下一代测序是能产生大量读数的测序技术，通常是上千（即上万或上十万）或百万等级的序列读数，而不是一次数百个。高通量测序区别于且不同于常规桑格或毛细管测序。一般地，测序产物是通常本身具有相对较短读数（约 600-30bp）的测序产物。该方法的示例为基于焦磷酸测序的方法，描述于 WO03/004690、WO03/054142、WO2004/069849、WO2004/070005、WO2004/070007 和 WO2005/003375，Seo 等，(2004)Proc. Natl.Acad.Sci.USA101:5488-93。该技术通常还包括广泛且精确的数据储存和用于读数装配的加工工作流程等。可用的高通量测序需要许多基因组分析的传统工作流程和方法重新设计成容纳目前产生的数据类型和质量。

[0031] 本文所用的“配对末端测序”是基于高通量测序的方法，特别是基于亿明达公司和罗氏公司目前销售的平台。亿明达公司发布了一种硬件模块（PE 模块），其可安装在现有测序仪中作为升级形式，能对模板两端测序，从而产生配对末端读数。配对末端测序可通过在载体上对待测序 DNA 分子链重新取向来实现，在该载体中进行测序，例如 Lakdawalla 所述的“Next generation sequencing:towards personalized medicine（下一代测序：走向个体化医疗）”，Michael Janitz 编，2008，威利（Wiley）部分 2.4。这类配对末端测序通常用于更小的片段（高至约 1000bp）。配对末端测序的另一变体有时称为伴侣配对测序，其中测序衔接子连接于 DNA 片段，该连接的 DNA 用识别序列包含在衔接子中的 II 亚类限制酶消化，自身环化，II 亚类酶消化，得到配对末端测序。这特别有助于分析较大片段（约 >1000bp）。也参见 Wei 等“下一代测序：走向个体化医疗”，Michael Janitz 编，2008，威利部分 13.2，

图 13.1。

[0032] II 亚类限制性内切酶是识别序列远离限制性位点的内切酶。换言之，II 亚类限制性内切酶在识别序列外部的一侧切割。其示例为 NmeAIII(GCCGAG(21/19) 和 FokI、AlwI、MmeI。存在在识别为序列外部两侧切割的 II 亚类酶。

[0033] 对齐和比对：术语“对齐”和“比对”表示根据相同或相似核苷酸的短或长延伸段的存在比较两种或更多核苷酸序列。比对核苷酸序列的数种方法为本领域已知。

[0034] 本文所用的“收集”指将多个样品(或人工染色体或克隆或基因组亚组或可重复的复杂性降低基因组)组合到库中。收集可以是将许多单独样品简单合并成一个样品（例如 100 个样品合并成 10 个库，每个含有 10 个样品），也可以使用更精细的收集策略。库中样品的分布优选使得每个样品存在于至少两个或多个库中。优选每库含有 10-10000 个，优选 100-1000，更优选 250-750 个样品。观察到每个库的样品数可广泛变化，该变化与例如研究的基因组大小或样品数目有关。通常，库或亚库的最大尺寸由独特鉴定一个库中某一样品的能力决定，例如通过一组标识物。用本领域熟知的收集策略产生库。本领域技术人员能够根据基因组大小、样品数目等因素选择最佳收集策略。得到的收集策略将视环境而定，其例子如平板收集，N- 维收集例如 3D 收集、6D 收集或复杂收集。为了便于处理大量库，库本身可以组合成超级库（即，超级库是样品池的库）或分成亚库。收集策略及其去卷积的其它例子（即通过检测一个或多个库或亚库中样品的已知相关标志（即标记或标识物）的存在来正确鉴定文库中每一个样品）如 US6975943 或 Klein 等，Genome Research,(2000),10,798-807 所述。收集策略优选文库中每个样品的分布使得对于每个样品有独特的库组合。其结果是某个（亚）库的组合独特鉴定一个样品。

[0035] 群聚：术语“群聚”意味着在相同或相似核苷酸的短或长延伸段的存在下，比较两个或多个核苷酸序列，并基于相同或类似序列的短（或更长）延伸段将具有某一最小水平序列同源性的序列分到一组。

[0036] 发明详述

[0037] 第一方面，本发明涉及同时发现、检测一个或多个或大量样品中的一种或多种多态性以及进行基因分型，包括以下步骤：

[0038] (a) 从一个或多个或大量样品中提供 DNA;

[0039] (b) 用至少一种限制性内切酶消化该 DNA 来降低样品 DNA 复杂性以产生限制性片段；

[0040] (c) 将标识物标签提供给样品限制性片段以产生带标签的限制性片段；

[0041] (d) 对至少部分的带标签限制性片段进行配对末端测序；

[0042] (e) 鉴定样品间的多态性。

[0043] 复杂性降低可仅仅基于用一种或多种限制性内切酶消化来自样品的 DNA。在某些实施方式中，可使用两种或多种限制性酶。对于限制性片段，可连接衔接子。该衔接子可连接到限制性片段的一端或两端，它们可相同或不同。用两种或多种不同的限制性酶限制DNA 从而获得限制性片段时，可使用不同的衔接子。复杂性降低还可通过扩增限制性片段来实现，例如使用针对衔接子或其部分的引物。用于扩增的引物还可包含与限制性酶识别序列剩余部分互补的一部分。在某些实施方式中，可使用既定技术例如AFLP® (EP534858)，其中在至少一个引物的 3’端添加 1-10 个随机选择性核苷酸以提供可重复的片段亚组。其它

的复杂性降低技术也可用，只要其可重复。这里的可重复指相同样品进行两次复杂性降低时获得相同亚组以及两个基本相同的样品间获得相同亚组。

[0044] 产生带标签限制性片段的标识物标签可以许多方式提供。标识物标签可通过以下方式提供：

[0045] -与限制性片段连接待标签衔接子以产生连接有带标签衔接子的限制性片段；

[0046] 或者

[0047] 用至少一个带标签引物扩增连接衔接子的限制性片段，该引物与至少部分衔接子互补，以产生连接带标签衔接子的限制性片段。

[0048] 该衔接子可仅由标识物标签组成，或该衔接子可含有其它官能团，例如能选择(部分)带标签限制性片段，例如以降低样品的复杂性，例如在阵列上。

[0049] 标志物标签也可在衔接子连接、扩增或复杂性降低之前或之后的单独步骤中添加，只要每个样品提供独特标签，该标签将限制性片段与其来源样品相关联。

[0050] 测序步骤优选使用高通量测序进行，使用包括伴侣配对测序在内的末端配对测序。

[0051] 在本发明的一个优选实施方式中，测定限制性片段的部分序列。优选测定限制性片段的两端序列，优选同时检测，即在同一测序运行中。用于该序列测定的方案通常指定GA//和罗氏平台用作配对末端测序，包括伴侣配对测序，如本文他处所定义。

[0052] 使用配对末端测序，通常包括伴侣配对测序，获得限制性片段两端的序列信息。来自限制性片段两端的序列信息(第一读数和第二读数，包括标识物)可合并，产生所谓的“双标签(ditag)”。双标签包含第一和第二读数的组合信息，优选可使用标识物标签与样品相联。该标识物标签优选与第一读数相关(或包含于其中)。可用计算机产生双标签。在一个优选实施方式中，读数之一，优选第二读数，在产生双标签前反向互补。这里的反向互补指读取的序列是反向的(例如，N1N2N3N4N5N6变为N6N5N4N3N2N1)。因此，详细的双标签为：

[0053] ID-读数1-读数2(反向互补)：IDIDIDIDM1M2M3M4M5M6N6N5N4N3N2N1

[0054] 还可参见描述该概念的图1。可从重复序列中获得双标签一部分，但另一部分来自基因组序列的另一部分，因此增加产生两部分独特组合的可能性。这能鉴定其它情况下不可能鉴定的序列间多态性。现有技术允许从片段两端获得150个核苷酸，产生300个参考性核苷酸。这显著提高了每个样品中独特的组合片段数目，因此提高待鉴定的多态性数目。可在允许配对末端(包括伴侣配对测序)的其它测序平台上实施相同的技术概念。

[0055] 高通量测序优选基于以下方式的测序：合成测序，焦磷酸测序(固体载体上)例如亿明达公司提供的平台(Ga//、Hiseq、MiSeq)或罗氏GS FLX，通常称为下一代测序。也可使用称为下下代测序的技术。其示例是基于连接测序、杂交测序、纳米孔测序(牛津纳米孔技术或NABsys(US20100096268,US20100078325,US20090099786))或太平洋生物科学公司(Pacific Biosciences)和离子激流公司(Ion torrent)公司提供的那些(Nature475,348-352页)。

[0056] 获得序列信息后，根据标识物标签将序列分配到每个样品。通过群集(或比对)序列，可鉴定序列间从而鉴定样品间的多态性。这使得在多个样品中同时鉴定SNP、检测SNP并测定基因型。可用本领域常规技术进行群集或比对。

[0057] 出于比较目的比对序列的方法为本领域熟知。各种程序和比对算法如下所

述：Smith 和 Waterman(1981)Adv.Appl.Math.2:482;Needleman 和 Wunsch(1970)J.MoI.Biol.48:443;Pearson 和 Lipman(1988)Proc.Natl.Acad.Sci.USA85:2444;Higgins 和 Sharp(1988)Gene73:237-244;Higgins 和 Sharp(1989)CABIOS5:151-153;Corpet 等，(1988)Nucl.Acids Res.16:10881-90;Huang 等，(1992)Computer Appl.inthe Biosci.8:155-65; 以及 Pearson 等，(1994)Meth.MoI.Biol.24:307-31，其通过引用纳入本文。Altschul 等，(1994)Nature Genet.6:119-29（通过引用纳入本文）提出了序列比对方法和同源性计算的详细考量。

[0058] NCBI 基本局部比对搜索工具（Basic Local Alignment Search Tool）（BLAST）（Altschul 等，1990J Mol Biol.5;215(3):403-10）来自各种来源，包括国家生物信息中心（NCBI，马里兰州贝塞斯达）和因特网，用于和序列分析程序 blastp、blastn、blastx、tblastn 和 tblastx 联用。其可在 http://www.ncbi.nlm.nih.gov/BLAST/ 获得。怎样使用该程序检测序列相同性的说明可在 http://www.ncbi.nlm.nih.gov/BLAST/blast help.html 获得。

[0059] 通常对就衔接子 / 引物和 / 或标识物修整的序列数据进行比对，即仅使用源自核酸样品的片段的序列数据。通常，获得的序列数据用于鉴定片段来源（即来自哪个样品），来自衔接子和 / 或标识物的序列从数据中移出并对该修整组进行比对。

[0060] 本发明的一个示例中，用两种限制性酶 EcoRI 和 MseI 消化基因组 DNA 样品，将衔接子连接到片段上。可采用 AFLP 复杂性降低（取决于基因组复杂度）。最后，得到的片段适于 GAII 测序并用配对末端形式测序（每个方向 76 个核苷酸）。使用针对标签定义的生物信息学方法和基因型鉴定（genotype calling），分析所得数据从而鉴定样品间的多态性。详细结果描述在实施例中。

[0061] 该技术的附加值从以下几方面体现：

[0062] 基于高通量测序限制性片段，通过使用制作物理图谱所用的相同限制性酶，经测序的标签和所得基因分型可容易地与物理图谱相关联。

[0063] 通过采用配对末端测序（即，对限制性片段两端测序，即各片段的 EcoRI 和 MseI 端）随后只比对独特的 EcoRI 和 MseI 标签组合，最大程度地在重复区域中进行 SNP 鉴定及基因分型。

[0064] 通过 AFLP 采用稳健的复杂性降低能收集大量样品。因此在某些实施方式中，在测序前将复杂性降低的样品收集在库中。

[0065] 优选基于总基因组 DNA 的技术。

[0066] 本申请通篇中，各种参考文献以括号引用从而更全面描述本发明涉及的技术状态。本说明书引用的所有专利和参考文献均通过引用全文纳入本文。

[0067] 显然上述说明和附图意在说明本发明的一些实施方式，而并非限制该保护范围。从本公开入手，更多的实施方式对于本领域技术人员是显而易见的并在本发明的保护范围和实质内容中，其为现有技术和本专利公开内容的明显组合。下文中本发明将通过非限制性实施例进一步说明。

实施例

[0068] 本项目目的是产生在基于随机序列的基因分型（rSBG）情况下分析配对末端序列

数据的策略。使用配对末端(双标签)相比单末端策略对拟南芥(Arabidopsis)分析数据，评价并比较其性能。为了这些目的，用来自亿明达 GAII NGS 平台的序列数据通过从头装配策略产生参考序列。随后，用亿明达读数对参考序列作图。然后就 SNP 的存在检测作图结果。

[0069] 拟南芥数据组遗传材料由两个亲本、两个 F1 个体和 28 个来自回交(BC)群的后代组成。

[0070] 用配对末端读数建立构建物，称为双标签，其中所述读数组合成单个“读数”。双标签长度是每对读数中各读数的长度总和。此外，建立双标签前反向互补读取读数 2，从而使双标签就参考(基因组)序列作图。因此，双标签的最终结构是 :ID 标签 - 读数 1- 读数 2(反向互补)。双标签建立在任何质量控制步骤实施前，修改质量控制方法以用过滤配对末端序列数据中各读文件所用的相同标准来过滤双标签。

[0071] ID 标签存在于配对末端序列数据的读数 1 以及读数 2 序列中。

[0072] 针对各拟南芥样品产生 EcoRI/MseI 文库并用亿明达 GAII 测序。针对双标签以及来自配对末端序列数据的读数 1 和读数 2 文件实施质量控制方法。

[0073] 应用于拟南芥序列数据的质量控制过滤所得的概括统计示于表 1。

[0074] 表 1 —拟南芥中亿明达 GAII 序列数据的描述统计。

[0075]

	双标签	读数 1	读数 2
读取起始数	19,622,319	19,622,319	19,622,319
无 ID 标签的读数	594,273	594,273	594,273
无 EcoRI 限制性酶模式的读数	3,136,557	3,136,557	无
无 MseI 限制性酶模式的读数 §	1,495,914	无	4,390,806
含有均聚物延伸的读数	18,298	39,849	17,766
与叶绿体 / 线粒体数据库显著配对的读数	3,438,320	3,704,597	3,399,068
含有未确定核苷酸的读数	25,632	32,177	23,621
低质量读数	32,452	54,647	138,345
过滤后的最终读取数目	10,880,838	12,060,184	11,058,405
通过 QC 的含有 ID 标签的读数%	55.5	61.5	56.4

[0076] §. 双标签质量控制中，在双标签最后评价 MseI 模式的存在。

[0077] 该测序通道中产生的总读数为 19,622,319。总共 97%的读数在开始时具有 ID 标签，这表明由于读数不匹配任何样品而仅去除了小部分序列。采用所有过滤标准后，数据集中读数范围保持为 10.9M (双标签)-12.1M (读数 1)。从数据集中去除读数的主要原因是

缺失预期的限制性酶基序(EcoRI 或 MseI),以及显著命中叶绿体 / 线粒体数据库的读数。

[0078] 采用 CAP3 组装(Huang 等, Genome Res. 1999Sep;9(9):868-77),用独特的读数完成拟南芥中双标签与单末端的比较,来评价双标签或单末端数据的分析性能。单末端分析单独用于配对末端序列数据的各读文件,通过分析各读文件获得的数目总和用来测定最终结果。该评价的概括性结果示于表 3。

[0079] 表 3 —拟南芥序列数据(用 CAP3 实施的组装以及独特读数)中比较组装策略的概括性结果(双标签相比单末端)。

[0080]

	双标签	单末端读数 1	单末端读数 2	单末端组合
读取数	689,512	597,878	784,782	1,382,660
毗连群数	31,891	38,236	43,448	81,684
具有 SNP 的毗连群	4,371	1,956	1,774	3,730
SNP 数§	**8,760**	3,338	2,838	**6,176**
每毗连群的 SNP 数	2.0	1.7	1.6	1.7
SNP 数¶ (90%基因分型率)	**2,634**	1,385	997	**2,382**
基因分型数 (90%基因分型率)	**78,174**	41,063	29,533	**70,596**
SNP 数† (80%基因分型率)	**3,346**	1,791	1,352	**3,143**
基因分型数 (80%基因分型率)	**96,779**	51,645	38,808	**90,453**

[0081] § 以严格的默认设置鉴定的 SNP 数,¶90%基因分型率中至少 28 个个体进行基因分型,†80%基因分型率中至少 25 个个体进行基因分型。

[0082] 两个基因分型率阈值处的 SNP 和基因分型数目高于作为双标签分析的数据。提高的性能使得在 90%和 80%基因分型率分别额外鉴定了 11%和 7%的 SNP 和基因型。随后,利用拟南芥数据可用的回交群结构测定各 SNP 数据集的 A、B 和 H 基因型数目。由于群体的回交特性,应该只观察到一个纯合子基因型,因此 B 基因分型数目是基因型鉴定中总错误率的良好指示。此外,A 和 H 基因型的频率应该约为 50%,这些频率的大偏差也标志着基因分型鉴定的问题。表 4 显示了拟南芥数据中基因型检查的结果。

[0083] 表 4 —拟南芥数据中基因型检查的结果

[0084]

	双标签	读数 1	读数 2
SNP 数†	2,334	1,101	1,024
基因型数	58,452	27,647	25,078
A 基因型	25,108	11,911	10,841
% A	43.0	43.1	43.2
B 基因型	447	252	156
% B	0.8	0.9	0.6
H 基因型	32,897	15,484	14,081
% H	56.3	56.0	56.1

[0085] 亲本是交替等位基因的纯合子时，仅对 SNP 检测基因型鉴定准确性。这些结果确定基因分型准确性很高，因为就所有测试的策略而言 B 基因型频率小于 1％。另外，它也揭示了所测的三个分析策略之间基因分型准确性没有实质性差异，因为各基因型种类的频率在所有测试策略中极相似。

[0086] 这些结果确定双标签分析产生较高数目的 SNP 和基因型，但不影响 SNP 鉴定和基因分型的准确性。

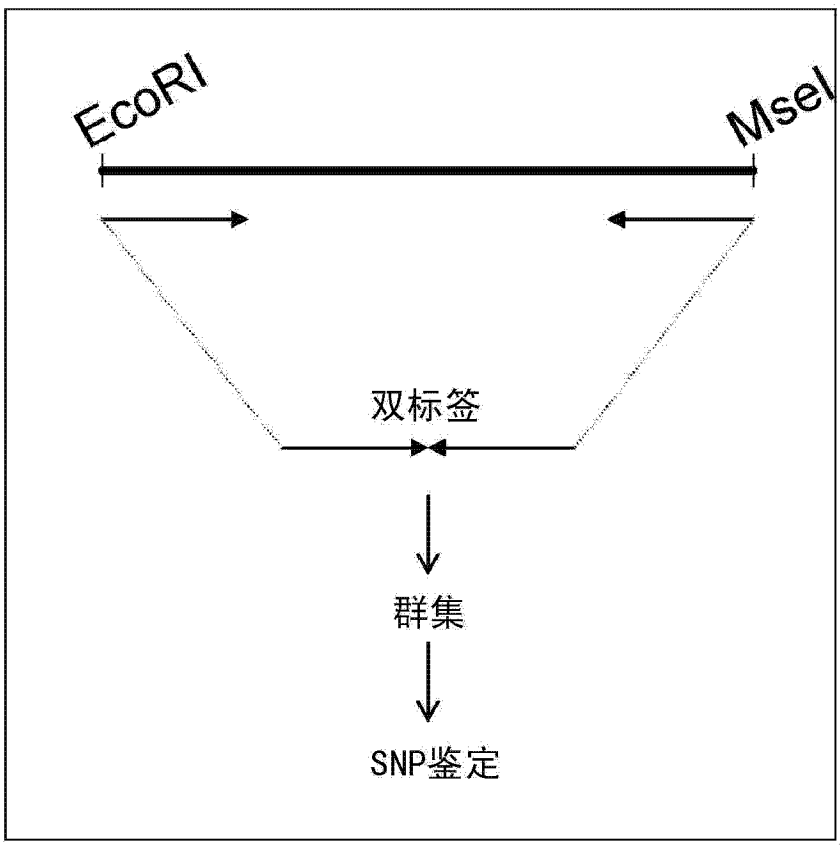


图 1