(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 102791880 A
(43)申请公布日 2012.11.21

(21)申请号 201080056670.5
(22)申请日 2010.12.17
(30)优先权数据
61/287,927 2009.12.18 US
2004019 2009.12.24 NL
(85)PCT申请进入国家阶段日
2012.06.14
(86)PCT申请的申请数据
PCT/NL2010/050860 2010.12.17
(87)PCT申请的公布数据
WO2011/074964 EN 2011.06.23
(71)申请人 凯津公司
地址 荷兰瓦赫宁恩
(72)发明人 J·斯图尔曼 F·亚尔钦
(74)专利代理机构 北京戈程知识产权代理有限公司 11314
代理人 程伟 韩文华
(51)Int.Cl.
C12Q 1/68(2006.01)

权利要求书 1 页 说明书 14 页 附图 1 页

(54)发明名称

改进的大突变分析

(57)摘要

本发明涉及鉴别和任选分离在生物体中表达的并与所述生物体特定表型(特征的性状)因果上相关的核酸序列的新策略。与本领域已知的方法相反,本发明的方法有可能高效地富集、鉴别、分离和/或克隆生物体例如(作物)植物的基因,这些植物没有或只有有限的基因组相关的信息是可获得的。

CN 102791880 A

1. 一种方法，其用于鉴别和任选分离与生物体特征相关的表达的核酸序列，其特征在于方法包括以下步骤：

a）提供具有所述特征的性状 A 的所述生物体的至少两个成员，其中具有性状 A 的所述成员源自，优选独立地源自，具有性状 B 的所述生物体的同基因成员，其中性状 A 和性状 B 是不同的；

b）从具有性状 A 的步骤 a）的每个成员中获得 cDNA，并确定至少部分，优选每个获自具有性状 A 的成员的个体 cDNA 的核苷酸序列；

c）通过与具有性状 A 的其它成员的相应 cDNA 比较，在具有性状 A 的成员的每个个体 cDNA 中确定单核苷酸多态性位置；

d）鉴别个体 cDNA，其中具有性状 A 的成员中的至少 2 个，优选所有具有性状 A 的成员，含有至少一个单核苷酸多态性，以及优选其中对于所述至少两个，优选所有具有性状 A 的成员，这些单核苷酸多态性位于 cDNA 的不同位置；

e）任选，验证与生物体的特征有关的步骤 d）中鉴别的每个 cDNA。

2. 根据权利要求 1 的方法，其中具有性状 A 的所述生物体的成员是通过具有所述性状 B 的所述生物体的成员致突变而获得的，其中具有性状 B 的所述成员在所述致突变处理之前是同基因的。

3. 根据权利要求 2 的方法，其中致突变的方法诱导点突变。

4. 根据权利要求 2-3 任一项的方法，其中致突变是通过非生物诱变剂的使用进行的。

5. 根据前述权利要求中的任一项的方法，其中所述生物体是植物，优选选自如下的作物植物：番茄、胡椒、茄子、莴苣、胡萝卜、洋葱、韭菜、菊苣、萝卜、欧芹、菠菜、甜瓜、黄瓜。

6. 根据前述权利要求中的任一项的方法，其中所述生物体是植物并且其中在步骤（a）之前建立编码隐性性状 A 的等位基因和编码显性性状 B 的第二等位基因的杂合的 F1 杂交群，并且其中所述 F1 杂交群经致突变，且其中在所述致突变之后从所述 F1 群中选择具有性状 A 的成员。

改进的大突变分析

技术领域

[0001] 克隆负责仅通过表型界定的生物性状的基因(所谓的正向基因克隆)已成为遗传学和生物技术中长期的挑战。特别是在(作物)植物领域中,快速而经济的正向基因克隆方法是急需的。目前的要求是找到提高正向基因分离速度降低其成本,特别是扩展可以进行基因克隆的物种范围的方法。在具有含有大量重复 DNA 的大或无特征(uncharacterized)的基因组的物种中,基因克隆尤其繁重。这种物种是莴苣、胡椒、洋葱、韭菜、豆、小麦、黑麦、大麦以及许多其它的。

背景技术

[0002] 正向基因克隆的目的是鉴别仅从表型获知,而没有或仅非常有限的分子信息或没有序列信息可获得的那些基因。正向遗传学的起点可以是例如天然存在的表型变体或人工诱导的突变体。

[0003] 本质上,已经认识了两组正向基因克隆的方法:绘图(mapping)策略和标签(tagging)策略。这些策略是互补的,各自都存在其固有局限性。

[0004] 在绘图策略中,通过标记辅助的减数分裂重组绘图将表型精确定点在染色体的最小片段上。基于绘图的克隆是非常繁杂的程序,特别可用在一小群模式物种上,对于该物种可获得其广泛的基因组信息,优选全基因组 DNA 序列。

[0005] 理论上,基于绘图的克隆对于通过有性繁殖(cycle)而复制的任何生物体是普遍适用程序(Peters 等人,(2003)Trends in Plant Science Vol.8No.10pp 484-491)。然而在实践中,存在基本的生物学限制。最重要的是,基于绘图的克隆关键取决于在感兴趣基因区域内减数分裂重组的良好频率。其次,这在特征很少的大基因组例如在植物如莴苣、胡椒、洋葱和许多其它基因组中变得越来越困难,因为在其中重复 DNA 构成了基因组的大部分,阻碍了 DNA 标记的清晰绘图以及阻碍了跨越遗传间隔的大的连续 DNA 序列的拼接。

[0006] 在经典的标签策略中,良好特征的生物插入诱变剂(转座子或 T-DNA 插入)已被用作载体来克隆基因。将可转座元件插入基因中会导致功能丧失或获得、表达模式的变化、或对基因功能可以没有影响,这取决于插入片段是否占据例如基因的编码或非编码区。负责任何新产生的表型的基因经界定位于插入元件的两侧,并且因此通过沿着基因组 DNA 两侧片段找回插入序列而进行克隆。

[0007] 通过插入突变的帮助进行加标签理论上是克隆基因的优选方法,因为这是快速的,并独立于减数分裂重组,而且不要求事先的基因组资源比如遗传图或大量的序列信息(见 Maes 等人,Trends Plant Sci.1999Mar;4(3):90-96)。通过加标签克隆突变的基因不受重复基因组序列或基因组大小的阻碍。

[0008] 加标签方法在少数模式生物体中已经成功,在这些生物体中天然基因标签(转座子)体系是已知的,或大量个体可用随机整合 DNA (T-DNA)进行转化(见例如 Settles 等人,BMC Genomics 2007,8:116,用玉米)。除了这一小群物种之外,标签不能或很难应用,由于缺乏已知的插入元件,或者由于群大小的逻辑限制。逻辑上,为了寻找一个或几个特异突变

体，标签要求几千个个体的群，因为每基因组插入序列的数量很低（1-200）。

发明内容

[0009] 本发明的目标之一在于提供一种新方法及其用途，其允许高效鉴别参与特定表型显现的核酸。该方法可以应用于任何可人工杂交和操作的（植物）物种，而无需该物种的广泛的减数分裂重组绘图和预先存在的基因组 DNA 序列知识。本发明的其他目标将明显地体现在本发明的说明书、实施方式和权利要求中。

[0010] 定义

[0011] 在以下说明书和实施例中使用了几个术语。为了在说明书和权利要求中提供明确且一致的理解，包括这些术语给出的范围，提供以下定义。除非此处另有定义，否则所有使用的技术和科学术语具有发明所属领域普通技术人员所通常理解的含义。所有发表、专利申请、专利和其它参考文献的公开通过引用其全部并入此处。

[0012] 等位基因：在同源染色体上可以具有相同位置并负责可替换的性状的至少两个可替换形式基因的其中一个。非限制性实例是花中的花簇颜色基因——单个基因可控制花瓣的颜色，但是可能存在基因的几个不同版本或等位基因。一个版本可形成红色花瓣，而另一可形成白色花瓣。个体花的最终颜色取决于其拥有基因的哪两个等位基因以及两者如何相互作用。

[0013] 特征：与生物体的表型质量相关。特征可使自身显现成不同性状。例如，生物体可以是以花朵颜色作为特征的植物，红色或白色花朵是该特征的性状 A 和 B。在本发明内，特征（或性状）可以是任意的，只要具有特征第一性状的生物体成员可以在表型上区别于具有特征第二性状的生物体成员。这不仅局限于那些可通过观测生物体直接观察到的区别，也包括那些通过进一步分析生物体而呈现出的特征 / 性状，例如基于对特定环境（如对特定的疾病）的抗性的分析，或基于对这些生物体中特定代谢产物的存在的分析。

[0014] 本发明上下文中“基因的表达”或“表达的核酸”涉及这样的过程，其中与合适的调控区特别是启动子可操作连接的 DNA 区被转录成生物活性的 RNA，即，该 RNA 能够被翻译成生物活性蛋白质或肽或其自身就有活性。

[0015] “基因”是含有与适宜转录调控区（如启动子）可操作连接的区域（转录区）的 DNA 序列，在细胞中该区域被转录成 RNA 分子（如：mRNA）。因而基因可含有几个可操作连接的序列，例如启动子、含有例如参与翻译起始的序列的 5’非翻译前导序列（也称为 5’UTR，其对应着翻译起始密码子的转录 mRNA 序列上游）、（蛋白）编码区（cDNA 或基因组 DNA）和含有如转录终止位点和多腺苷酸化位点的 3’非翻译序列（也称为 3’非翻译区，或 3’UTR）。

[0016] 同基因的表示遗传上相同。同基因的群内的个体细胞通常是具有等价的遗传构成的单一祖先的子代。本发明中，除了由于自然变异所造成的，或在本发明优选实施方式中，由致突变处理所造成的任何点突变之外，“同基因的”被解释为在 cDNA 水平上个体成员 95-99% 甚至 100% 相同。实际上术语“同基因”是公知的也是本领域技术人员可以理解的。

[0017] 致突变或致突变处理：此处该术语涉及导致在核酸、基因、或基因组中引入变化的处理，例如导致引入点突变和 / 或插入或删除长达 10 个连续核苷酸。

[0018] 核酸：本发明的核酸可包括嘧啶和嘌呤碱基（分别优选胞嘧啶、胸腺嘧啶和尿嘧啶，以及腺嘌呤和鸟嘌呤）的任何聚合物或寡聚物（见 Albert L.Lehninger,Principles of

Biochemistry, 在 793-800 (Worth Pub. 1982),出于所有目的其通过引用全部并入此处)。核酸可以是 DNA,包括 cDNA,或 RNA,或其混合物,其可以是永久或暂时存在的单链或双链形式,包括同源双链、异源双链(heteroduplex)和杂交状态。

[0019] 测序:该术语测序涉及确定核酸样品如 DNA 或 RNA 中核苷酸顺序(碱基序列)。

[0020] 性状:生物学中,性状涉及生物体的个体成员与相同生物体的(任何)其他个体成员相比时任何表型上可区别的特征。在本发明的上下文中,性状可以是遗传的,即通过生物体的遗传信息方式被传代至生物体的下一世代。性状的非限定性实例包括,例如,可以视觉观察的性状或可以通过分析观察的性状(例如,对于代谢物的存在或水平)。

[0021] "相同特征的性状"或"所述特征的性状":一个特征所存在(或呈现)的一组至少两种性状的任一个。例如,对于特征"花的颜色"的情况,表型上的显现可包括蓝色、红色、白色等。在以上实例中,蓝色、红色和白色都是相同特征的不同性状。

[0022] 发明详述

[0023] 通过提供所附权利要求描述的方法令人惊奇地解决了以上所述的目的。

[0024] 更具体地,提供了鉴别和任选分离与生物体特征相关的表达的核酸序列的方法,其特征在于该方法包括以下步骤

[0025] a)提供具有所述特征的性状 A 的所述生物体的至少两个成员,其中具有性状 A 的所述成员源自,优选独立地源自,具有性状 B 的所述生物体的同基因成员,其中性状 A 和性状 B 不同;

[0026] b)从具有性状 A 的步骤 a)的各成员中获得 cDNA,并确定至少部分,优选获自具有性状 A 的成员的每个个体 cDNA 的核苷酸序列;

[0027] c)通过与具有性状 A 的其它成员的相应 cDNA 的比较,在具有性状 A 的成员的每个个体 cDNA 中确定单核苷酸多态性位置;

[0028] d)鉴别个体 cDNA,其中具有性状 A 的成员中的至少 2 个,优选所有具有性状 A 的成员,含有至少一个单核苷酸多态性,以及优选其中对于所述至少两个优选所有具有性状 A 的成员,这些单核苷酸多态性位于 cDNA 的不同位置;

[0029] e)任选,验证与生物体的特征有关的步骤 d)中鉴别的每个 cDNA。

[0030] 任选并优选,方法可以包括鉴别来自 d)中鉴别的每个 cDNA 的个体 cDNA 的步骤(f),其中至少两个,优选所有具有性状 A 的成员具有至少一个非同义的单核苷酸多态性。

[0031] 然后,在一个优选实施方式中,方法可包括鉴别来自前述步骤(f)中鉴别的每个 cDNA 的个体 cDNA 的步骤(g),其中至少两个,优选所有具有性状 A 的成员恰好具有一个非同义的单核苷酸多态性。

[0032] 并且,方法可包括列出进行的步骤(d)、(f)和/或(g)中鉴别的所有个体 cDNA、根据所遵循的进行的步骤(d)、(f)和/或(g)的数目将个体 cDNA 排名(依从性最大的排名最高)的步骤(h)。

[0033] 方法还可包括,优选,鉴别表达的核酸序列(鉴别的 cDNA 源自表达的核酸序列),以及任选,克隆包括表达的核酸序列的基因。

[0034] 本发明是基于这样的实现:目前正向克隆策略的上述问题可通过本发明的方法解决,以及通过使用可用于任何生物体的非生物诱变剂来克服基于生物诱变剂的加标签方法的局限性,其可以在每基因组中产生几千个易于检测的突变体从而尽可能地消除逻辑群限

制，结合上对获自显示表型(性状 A)的(至少两个成员的)突变体集合(collection)的转录物组(cDNA)进行测序，并在此基础上在所述突变体集合的各成员中鉴别携带至少一个单核苷酸多态性(优选至少一个非同义的单核苷酸多态性)的任何基因(但不一定在所述基因 / cDNA 中的相同位置)。此外，突变体集合可获自自然变异。

[0035] 除了通过用例如诱变剂处理所诱导的突变以外，还可使用发生在其它同基因的群中的自发突变体。换言之，如果在具有性状 B 的同基因的群中，发现至少两个成员具有性状 A (如由于自发突变)，可以用如此处所公开的方法鉴别负责特征和性状的基因。

[0036] 认识到，用化学诱导的突变如点突变可以在任何生物体的基因中加标签(并从而可显现区别表型)。然而主要问题在于如何在全基因组中检测这些小而随机的突变(标签)，因为这些突变的质量不足以进行任何形式的特定 PCR 扩增。而且，化学诱导的突变以高频率随机引入到基因组中。因此，与原始物种的基因相比，许多基因可能同时含有突变，从而使得正确鉴别负责观察到的表型或显示特定性状的基因复杂化。

[0037] 本发明人已认识到，解决方案可以来自在待克隆的基因座上创建独立且区别的突变体等位基因的集合(但是尚且未知)，并将来自这种集合的所有 cDNA 进行多次再测序。当与这种集合中的(所有)cDNA 序列比较时，突变体池中的一个 cDNA 将显示出一致的突变。这极可能是形成突变表型基础的基因。

[0038] 换言之，以及例如，可以用例如在基因组中诱导点突变的诱变剂，处理显现特定性状 B (例如红色的花)的同基因的群(的种子)。由于用所述诱变剂处理的结果，随机突变引入至各处理的成员的基因组中。通常，处理之后大多经处理的成员将仍然显现特定性状 B。在那些成员中，用诱变剂处理引入的突变不改变性状。换言之，在这些成员中没有涉及性状的特征的基因经突变而改变性状(例如变成性状 A，例如，黄色的花)。然而，在一些成员中，诱变剂可以这样的方式诱导构成观察到的性状基础的基因的突变，即在那些成员中原始性状 B 不再显现，但是性状 A 显现。

[0039] 由于诱变剂的随机特征，如果至少两个成员，优选更多成员被观察到具有性状 A，那么极有可能由诱变剂诱导的并负责从性状 B 到性状 A 的改变的突变尽管是位于相同基因，但不在所述基因的相同位置。例如，在具有性状 A 的第一成员中，突变可能位于核苷酸 R，在具有性状 A 的第二成员中，突变可能位于核苷酸 S，在具有性状 A 的第三、第四和第五成员中，突变可能分别位于位置 T、U 和 V (基因中 R、S、T、U 和 V 位于不同的位置)。换言之，当互相比较时，这些成员均显示出在负责的基因不同位置的突变。与此同时，尽管显示性状 A 的成员还携带遍及基因上其它位置的突变，在现在显示性状 A 的每个成员中另一个表达基因被修饰的可能几乎为零。

[0040] 现在通过相互比较获自具有性状 A 的成员的 cDNA 时，可鉴别出在具有性状 A 的各成员中携带至少一个核苷酸的修饰的 cDNA，该修饰位于至少两个所述成员中所述 cDNA 中的不同位置。所述 cDNA 负责观察到的相同特征的性状。

[0041] 在本发明的方法中，池由另外的同基因遗传背景中的(诱导的)等位基因系列组成。本发明的方法检测位于待克隆的基因中的 SNP。

[0042] 本方法适用于可以人工杂交和致突变的所有物种，包括这些众所周知的难以操作的作物如胡椒和洋葱，且本发明不依赖于遗传图的存在并将在任何大小和复杂性的基因组中起作用。

[0043] 根据本发明的方法，因此现在已可能在大量的基因中鉴别那些为可能参与或负责观察到的性状的候选者的基因。在实践中，当采用本发明的方法时，可能的候选者的组可以快速且可靠地限定为，例如 10、5、3、2 或甚至 1 个候选基因。使用根据本发明的方法，负责或参与特定性状的显现的基因的快速鉴别，现在已经成为可能。

[0044] 更详细地，在本发明的方法中，比较具有特定表型(或具有特定性状 A)的生物体的至少两个个体成员。至少两个成员源自具有(不同的)性状 B 的同基因生物体。

[0045] 对于本发明重要的是，具有性状 A 或性状 B 的生物体的成员是(性状 A 的情况时)源自(如通过诱变剂处理或通过自发突变)或是(性状 B 的情况时)所述生物体的同基因成员，换言之，生物体具有相同的遗传背景(例如源自相同近交系)。除了由于自然变异或者，在本发明的优选实施方式中，由于致突变处理引入遗传物质中的变化外，具有性状 A 或性状 B 的个体成员是同基因的，即具有相同的遗传背景。

[0046] 具有性状 A 的成员和具有性状 B 的成员之间遗传物质中的这些差异中，包括观察到的表型。由于特征中观察的变化(从性状 A 至性状 B)，具有性状 A 的成员必须进行这样的修饰从而它们表达至少一个不同于具有性状 B 的生物体的成员所表达核酸的核酸，并且核酸与生物体的特征有关，性状 A 和性状 B 是特征的表型形式。

[0047] 优选，具有性状 A 的至少两个成员独立地源自具有性状 B 的所述生物体的同基因成员。

[0048] 术语“独立地源自”表示具有性状 A 的至少两个成员是例如由于诱变剂处理而彼此独立地获得所述性状 A 的成员。如果生物体是植物，由于例如种子(植物生长自种子)的诱变剂处理，至少两个成员可能已各自获得性状 A。在这个实施例中，性状引入独立于显现性状 A 的任何其它成员的第一成员。换言之，例如在最后的 3、4、5、6、7、8、9 或 10 代期间，优选具有性状 A 的成员不共享已显现所述性状 A 的共同前辈(predecessor)。优选所有成员都独立地源自具有性状 B 的所述生物体的同基因成员。

[0049] 如上所述，由于两个不同性状的存在所显现的，具有性状 A 的成员表达至少一个不同于具有性状 B 的其它成员所表达核酸的核酸，并且该核酸与生物体的特征有关，性状 A 和性状 B 是该特征的表型形式。

[0050] 为了检测这种带标签的(通过，例如，点突变)核酸，在本发明方法中，获得来自具有性状 A 的步骤 a)的各成员的 cDNA，优选总 cDNA，以及确定来自具有性状 A 的所述成员的个体 cDNA，优选每个 cDNA 的核苷酸序列。换言之，具有性状 A 的所有成员的转录物组(优选总转录物组)，与具有独立获得性状 A 的所有其它成员的转录物组(优选总转录物组)比较。优选比较所选组织中所有表达的基因的完整序列。

[0051] 对此，cDNA 优选总 cDNA 是从具有性状 A 的至少两个成员中获得的。可以通过技术人员已知的任何适宜方法完成 cDNA 优选总 cDNA 的制备。可购买许多商业可获得的 cDNA 合成试剂盒，例如从 ABgene、Ambion、Applied Biosystems、BioChain、Bio-Rad、Clontech、GE Healthcare、GeneChoice、Invitrogen、Novagen、Qiagen、Roche Applied Science、Stratagene 等。这种方法例如在 Sambrook 等人(Sambrook,J.、Fritsch,E.F. 和 Maniatis,T.，在 Molecular Cloning:A Laboratory Manual.Cold Spring Harbor Laboratory Press,NY,Vol.1,2,3(1989))中有描述。技术人员还将理解，例如通过正态化来降低体现为个体 cDNA 形式的差异，cDNA 可进一步进行处理(Evrogen 等)。

[0052] 如所示，本发明的方法要求获得、测序并相互比较具有性状 A 的至少两个成员的 cDNA 优选总 cDNA。如上所解释的，不需要和获自具有性状 B 的成员的 cDNA 进行比较。更优选地，本发明方法中提供 / 使用至少 3、4、5、6、7 或 8 个生物体的成员，并且各具有性状 A。

[0053] 如上所解释的，本发明是基于这样的实现 ：可通过比较独立地获得表型性状（性状 A）的生物体的（优选全）转录物组的序列而鉴别与特定特征（或性状）相关的遗传差异。因为在优选实施方式中，具有性状 A 的成员是通过对具有性状 B 的成员进行随机的（化学）致突变处理而产生的，这种处理的生物体的遗传物质将含有许多且随机的突变，其中大部分并不与所观察的性状相关。比较具有性状 A 的一个成员的 cDNA（优选总 cDNA）和具有性状 B 的一个成员的 cDNA（优选总 cDNA）将因此允许鉴别与所观察表型相关的核酸。然而本发明人认识到当具有性状 A 的至少两个成员相互比较时，并各自优选具有独立获得的所述性状 A，有可能鉴别负责（或有关）特征（或性状）的核酸 ：

[0054] 具有性状 A 的成员（独立地）源自相同的同基因的来源。由于同基因的来源的突变处理，变化被随机引入遗传物质中。在第一成员中诱导的变化将不同于在第二成员中诱导的变化。然而，如果所述第一成员和所述第二成员，由于致突变处理，现在都表达出表型，即具有性状 A，则极可能是在两个成员中，相同 cDNA 将表现出单核苷酸多态性（SNP）。由于突变处理的随机过程，这种 SNP 不太可能位于这种 cDNA 内的相同位置。所述 cDNA 现在可鉴别为源自表达核酸的 cDNA，而该表达核酸与所研究的生物体的特定性状或特征相关，因为在具有性状 A 的至少两个成员中，这种非相关的 cDNA 将表现这种 SNP 的变化几乎为零。

[0055] 换言之，本发明是基于这样的实现 ：在来自如一个基因（性状 A）中的 5 个等位基因突变的“突变池”的所有 cDNA 序列中，当相互比较时，将只有一个个体 cDNA 一直显示出序列中的差异。

[0056] 为了比较具有性状 A 的各成员的 cDNA 优选总 cDNA，具有性状 A 的成员的 cDNA 优选总 cDNA 需要进行测序。cDNA 的测序可通过技术人员已知的任何适宜方法完成。然而，尤其非常优选 Margulies M. 等人（http://www.454.com，Genome sequencing in microfabricated high-density picolitre reactors. Nature 437:376-80, 2005）中描述的这些方法，其允许快速并高效地对全转录物组（所有 cDNA）进行测序。例如，在优选实施方式中，所获得的 cDNA 片段的核苷酸序列可通过高通量测序方法确定，例如 WO 03/004690、WO 03/054142、WO 2004/069849、WO2004/070005、WO 2004/070007 和 WO 2005/003375、Seo 等人（2004）Proc. Natl. Acad. Sci. USA 101:5488-93 以及 Helicos、Solexa、US Genomics、etcetera 的技术中所公开的那些，其都通过参考引入此处。

[0057] 在本发明所述方法的下一步中，将所获得的具有性状 A 的成员的 cDNA 优选总 cDNA 序列与具有性状 A 的其它成员的 cDNA 优选总 cDNA 比较，从而建立 cDNA 中单核苷酸多态性频率。可以通过技术人员已知的任何适宜方法完成这种确定，例如在所附实施例中展示的那些。

[0058] 简言之，例如 cDNA 片段核苷酸序列的比对可用于收集源自相同转录基因（但是来自具有性状 A 的不同成员）的核苷酸序列，并比较这些核苷酸序列。核苷酸序列是否源自相同的转录基因，可以基于序列之间的同源性来建立。为了本发明的目的，假定他们在至少 30、优选至少 50、更优选至少 90、还更优选至少 100、150、200 个核苷酸长度中至少百分之 95、96、97、98、99、100 是同源的，则核苷酸序列是源自相同的转录基因。

[0059] 在下一步，基于获得的数据，可以鉴别具有提高的 SNP 频率的具有性状 A 的成员的 cDNA 集合中的 cDNA，优选总 cDNA。在本发明方法中，鉴别个体 cDNA，其中具有性状 A 的至少两个成员含有至少一个单核苷酸多态性，以及优选其中这些单核苷酸多态性位于 cDNA 中的不同位置。优选所分析的具有性状 A 的各成员在所述 cDNA 中优选在 cDNA 中不同位置含有至少一个单核苷酸多态性。例如，如果分析具有性状 A 的四个植物，可鉴别在四个植物的（相同，即源自各植物的相同基因）cDNA 中对于植物 1、2、3 和 4 分别在位置 25、34、56 和 101 上含有 SNP 的 cDNA。或者，对于一些植物，而非另一些植物，SNP 可能在相同位置上。例如，对于植物 1、2、3 和 4，SNP 可能分别在位置 25、25、56 和 101，或甚至 25、25、101 和 101。

[0060] 在本发明的方法中，在具有性状 A 的成员中测序的所有 cDNA 中，至少一个特异性 cDNA 将在至少两个成员中，但优选具有性状 A 的各成员中，携带至少一个遗传上（优选非同义的）单核苷酸多态性（SNP）。

[0061] 换言之，这些特异性 cDNA 将在经分析的具有性状 A 的各成员中包括至少一个 SNP（不一定在相应 cDNA 中的相同位置）。这种 cDNA 可能是源自参与或负责所观察到的性状 A 的基因的 cDNA 的候选者，并且这一信息可随后通过技术人员已知的方法用于鉴别表达的核酸序列并克隆相应的基因。

[0062] 任选且优选，在本发明方法的下一步骤中，对来自之前步骤中鉴别的每个 cDNA 的个体 cDNA 进行鉴别，其中具有性状 A 的至少两个优选所有成员具有至少一个非同义的单核苷酸多态性。

[0063] 技术人员将了解，之前的步骤中鉴别的 cDNA 中发现的至少一个 SNP，必须极有可能与编码的蛋白质中氨基酸的变化有关，以便参与或负责或因果上与观察到的表型（性状 A）相关。因此，鉴别至少一个非同义的存在（导致密码子的变化，导致编码的氨基酸的变化）进一步改善本发明的方法，并允许甚至更可靠地鉴别参与观察到的性状 A 的基因。

[0064] 的确，技术人员理解本发明上下文中的术语“非同义的”，表示由于三联体碱基中的至少一个核苷酸的变化，发生密码子变化，其中与原始密码子相比，变化的密码子编码不同的氨基酸。

[0065] 如果进行在先的鉴别 cDNA 的步骤，其中具有性状 A 的至少两个优选所有成员具有至少一个非同义单核苷酸多态性，任选，但优选，在下一步中，鉴别来自之前步骤中的每个 cDNA 的个体 cDNA，其中具有性状 A 的至少两个优选所有成员具有不超过一个非同义单核苷酸多态性。

[0066] 本发明人发现实际上这种 cDNA 极有可能是源自负责或因果上与观察到的表型（性状 A）相关的基因的候选者，从而进一步改善本发明的方法。

[0067] 本发明方法的进一步改善可以涉及鉴别来自任意在先步骤中鉴别的每个 cDNA 的个体 cDNA，其中具有性状 A 的至少两个优选所有成员具有非同义单核苷酸多态性，其在进化上保守的氨基酸位置上引起氨基酸变化，并优选这些新的氨基酸化学上区别于原始非突变的氨基酸（即芳香的相对于非芳香的，极性相对于非极性，在特定 pH 下负电的、正电的或中性的，亲水性正的或负的（亲水的或疏水的）等）。

[0068] 根据本发明方法所进行的以上公开的步骤，可以根据所符合的（所进行的）步骤 d-f 的数目对在进行的步骤（d）、（f）、（g）和 / 或（h）中鉴别的所有个体 cDNA 排名。在下一步骤中，符合所有步骤（d）、（f）、（g）和 / 或（h）的表达的核酸序列因而被排为最高，可以被

鉴别为最有可能参与、负责或因果上与观察到的表型 / 尤其(./esp)相关的候选者。

[0069] 基于以上,可以鉴别表达的核酸序列并可以克隆基因。

[0070] 在优选实施方式中,提供本发明所述的方法,其中具有性状 A 的所述生物体的成员是通过具有所述性状 B 的所述生物体的成员致突变而获得的,其中具有性状 B 的所述成员在所述致突变处理之前是同基因的。

[0071] 在本发明方法的一个实施方式中,具有区别特征性状 A 的至少两个成员可以是自然变异的结果,即由于在之前显现性状 B 的所述生物体中核酸中的自然或自发变化。在遗传信息中的这些突变是非刻意的,但揭示了生物体携带负责所观察的表型变化的遗传信息(例如,从性状 B 至性状 A)。

[0072] 然而优选地,本发明方法不依赖于遗传信息中的这种非刻意和不可控的变异,反而依赖于刻意致突变引起的具有特定性状的突变。技术人员可以理解如何使任何生物体的遗传信息发生突变,例如通过采用已知的诱变剂。由于采用这些诱变剂,突变可随机发生在生物体成员的基因组中。换言之,核酸例如基因可以用(化学地、生物地或通过辐射方式)诱导的突变(如点突变,如通过甲磺酸乙酯)进行加标签(与非突变核酸相比)。

[0073] “辐射致突变”的实例包括 X- 射线、γ - 射线、UV 光、或电离粒子。“生物致突变”的实例包括例如 WO0150847 所述的这种方法,例如使用重组酶如 recA。可用于本发明方法的“化学诱变剂”的实例包括应用特定化学品如甲磺酸乙酯(EMS)、硫酸二乙酯(DES)、N- 亚硝基 -N- 乙脲(ENU)、双环氧丁烷、2- 氨基嘌呤、5- 溴尿嘧啶、溴化乙啶、亚硝酸、亚硝基胍、羟胺、叠氮钠或甲醛。

[0074] 优选地,致突变方法诱导基因组的点突变、或插入、置换或高达 10 个连续核苷酸的删除。

[0075] 技术人员将理解在该实施方式上下文中,具有性状 A 的成员和具有性状 B 的成员是同基因的(因为它们源自相同的遗传背景),除所述致突变处理引入的突变之外。

[0076] 然而,与经典绘图和用插入突变的加标签策略相反,这种小而随机的突变难以检测,因为它们的质量不足以进行任何形式的特定 PCR 扩增。然而,使用以上策略存在几个好的理由在于致突变本质上可应用于任何感兴趣的物种、诱导突变体的谱比加标签途径更广、致突变通常更高效并且更易于获得第二位点突变。

[0077] 尽管可以采用多种适宜的致突变方法,优选地诱变剂不是选自转座子插入物或 T-DNA 插入物的生物诱变剂。优选地,所用的致突变方法在遗传物质中引入点突变。

[0078] 在另一实施方式中,提供了本发明的方法,并且其中生物体是植物,优选选自如下的作物植物 :番茄、胡椒、茄子、莴苣、胡萝卜、洋葱、韭菜、菊苣、萝卜、欧芹、菠菜、甜瓜、和黄瓜。

[0079] 可用于本发明的生物体可以是任何生物体,包括细菌、原核生物和真核生物。然而优选生物体是真核生物,特别是植物,更特别是作物植物。优选地,植物是属于以下重要作物植物的植物 :包括番茄、胡椒、茄子、莴苣、胡萝卜,但是本方法本质上可应用于所有其他植物。

[0080] 特别地,本发明允许核酸的鉴别和克隆,如来自作物植物的基因,这些作物植物中现有技术中可用的绘图和加标签技术变得不可能或很难采用。

[0081] 在另一实施方式中,提供了本发明的方法,其中生物体是植物,并且其中在步骤 a)

之前建立编码隐性性状 A 的等位基因和编码显性性状 B 的第二等位基因的杂合的 F1 杂交群，并且其中所述 F1 杂交群经致突变，且其中在所述致突变之后从所述 F1 群中选择具有性状 A 的成员。

[0082] 优选地，可通过例如在一个基因座选择新的等位基因的标准遗传方法构建等位基因系列。这种方法是由建立一个参照等位基因（其产生这个基因座之前识别的隐性表型，即性状 A）和一个显性等位基因（产生这个基因座性状 B）的杂合 F1 群组成的。当有害突变发生在等位基因 B 中时，F1 的致突变将显露新的产生性状 A 的等位基因。这种带有性状 A 的突变 F1 植物以一定的频率出现，而大部分 F1 群将显示性状 B。在一个实施例中，可以使携带基因座的显性等位基因（编码性状 B）的开红花的栽培品种的种子（或花粉，或任何其它组织）致突变，并使用这种材料与携带基因座的隐性参考等位基因（编码性状 A）的非致突变的开白花的栽培品种进行有性杂交。在获得的 F1 子代中，大部分个体将是红色的，除了少数白色个体携带显性 B 等位基因中新诱导的、独立的突变从而它们表达性状 A。在第二个实施例中，可直接使 F1 材料（如种子，或任何适宜的组织）致突变，并在生长于致突变材料的获得的个体植物中，对性状 A 的表达进行评分。实施例 1 和 2 的优势在于，使用位于基因座的 A 和 B 等位基因的 F1 材料杂合子，允许将新产生的性状 A 等位基因相对确定地指定至待克隆的基因座。没有 F1 方法，例如：原始致突变的材料自体受精所取得的 F2 群，理论上可能是开白花的突变体植物由不同基因（而不是待克隆的基因）突变的结果而产生。因此这些新产生的突变可能是非等位基因的，并将本发明上下文中混淆。

[0083] 通过本发明方法鉴别的核酸，之后可以被分离、克隆（基因）或引入宿主细胞、或用于例如植物培育程序。

[0084] 综上所述，本发明涉及鉴别和任选地分离在生物体中所表达的并与所述生物体的特定表型（特征的性状）相关的核酸序列的新策略。与本领域已知的方法相反，本发明的方法可以高效地鉴别、分离或克隆生物体例如（作物）植物的基因，因为这些植物没有或只有有限的基因组相关信息是可获得的。此外，现在可以利用化学诱导点突变作为通用、广泛地可用于建立致突变的群的方法，例如这比通过使用 T-DNA 或转座子系统生成高得多的突变频率，从而减少该方法所需要的生物体（例如植物）的量。而且，该方法不依赖于标记在基因组中的使用或存在，因为它直接检测感兴趣的基因的变化。

附图说明

[0085] 图 1 是本发明方法的（实施方式的）示意概况。概况中，并适用于此处公开的发明，显示了通过单个个体来描述同基因的群，其中单个个体是在基因（其为当纯合时可观察到的表型显现的基因）上携带突变（黑圈）的参考等位基因的杂合体，其是通过本发明方法所要鉴别的基因（图 1A，左图）。致突变处理之后，群将含有消除或补偿基因功能的新型等位基因（图 1A 右图，灰圈），从而参考等位基因的表型作用通过杂合性丢失变得可观察。在图 1A 中描述了五种独立的非同义等位基因。图 1B 显示这五个植物的等位基因组成，选作用于与杂合性丢失有关的表型显现的组。在图 1B 的左图中，这五个植物的基因型描述为 1-5，亲本或母本等位基因用词缀“a”或“b”表示，对于与所选表型显现无关的基因。可见植物和亲本中这种基因不含有有关的多态性。相反，因果上与所选表型有关的所选基因（图 1B，右图）含有一个非同义的、或遗传上非中性的等位基因，等位基因由五个同基因的植物共享，而且

对于亲本之一是排他的(这种情况中,亲本“a”)。此外,其在五个表型上选择的植物中的每个中含有新型且不同的非同义、或遗传上非中性的多态性。这种多态性模式必然是极罕见的,并因此可用于鉴别构成由所选表型所限定的性状的核酸序列(基因)。

[0086] 图 2 即为表 1。

[0087] 表 1

[0088] A) 三个独立矮牵牛花突变体的转录物组测序之后的组合结果(参考转录物组)的总结,具有通过杂合性丢失获得的红色花颜色。

[0089] B)根据本发明中的步骤,运用基于单基因(unigenes)中多态性模式的选择原则之后,剩余的单基因数目。还给出了选择之后的富集因素,表示为参考转录物组中单基因总数的百分比。

[0090] C)基于本发明的方法,剩余的基因的排名列表。最遵循所有所用原则的三个基因。经推理(a priori)构成红色花颜色的基因是花青素 -3-O- 葡糖基转移酶,与两个其它基因共同排为 1。

具体实施方式

[0091] 实施例

[0092] 矮牵牛花中花着色基因的从头克隆

[0093] 在单个花颜色基因中产生突变体

[0094] 为了证实本发明的所述方法,我们意图从头鉴别之前已知的杂交矮牵牛花的花颜色基因(RT)。由自交系 W5 (rt::dTph3 ;Stuurman 和 Kuhlemeier 2005, Plant J. 41(6):945-55) 和 M1 (RT, Snowden KC 和 Napoli CA(1998)Psl:a novel Spm-like transposable element from Petunia hybrida. Plant J. 14:43-54) 之间杂交产生开紫红色花的 F1 杂交矮牵牛花,其导致在基因 RT 上携带隐性等位基因的开紫红色花的杂合基因型,其编码 UDP 鼠李糖 :花青素 -3- 葡萄糖苷鼠李糖基转移酶。在所用遗传背景中,RT 的无效突变产生红色的花,这是花青素 -3- 葡萄糖苷积累的结果(Kroon 等人,1994, Plant J. 5(1):69-80 ;Brugliera 等人,1994, Plant J. 5(1):81-92)。用甲磺酸乙酯(EMS)致突变处理 F1 杂交种子,一些获得的植物在野生型 RT 等位基因上会携带新的突变,这造成了杂合性丢失并表达红色花冠颜色。

[0095] 从 EMS 处理的种子生长出 3600 个 F1 植物的群,可以目测记录所有个体植物的花的颜色。观察了七个开红色花的突变体。因为红色是隐性的,在原始致突变的群中表达这种颜色表示开红花的突变体在 RT 基因的野生型拷贝中携带了 EMS 诱导的点突变。这些新分离的突变必然是杂合的。

[0096] 为了在新鉴别的突变等位基因中从头鉴别构成红色的基因,在三个独立的突变体中,对发育时间点时的花冠檐(corolla limb)完整且正态化的转录物组进行测序,这正是花青素色素变得可见的发育时间点。使用 SMART cDNA 合成试剂盒(Clontech 公司,1290 Terra Bella Avenue, Mountain View, CA 94043, USA) 将 3 个突变体的每个总 RNA 转化成 ds-cDNA。使用 Trimmer 试剂盒(Evrogen 公司, Miklukho-Maklaya str, 16/10, 117997, Moscow, Russia)对双链(ds-)cDNA 进行正态化。来自各个体突变体植物的正态化的 cDNA 分别经加工用于在 GS-FLX 肽测序仪(454 Life Sciences)上进行测序。加

工涉及 cDNA 片段化为 ~700nt，并用 10bp 的多重识别符(multiplex identifier)(MID)加标签。随后将来自所有突变体的所有经加工的 cDNA 样本合并起来，在 GS-FLX 肽测序仪上进行两轮测序。我们获得平均 300 个核苷酸长度的总共 1′ 977′ 422 个读取。

[0097] 使用拼接程序 Newbler (Roche 454 软件发布 :2.3-PreRelease-9/14/2009)使用缺省运行拼接设置，将所有合并的序列数据经修剪去除接头序列和 MID 序列，随后拼接成重叠群的非冗余组。结果概况在表 1A 中给出。获得总共 39306 个独特的重叠群(平均大小为 955 个核苷酸，中值大小为 1025 个核苷酸)，这代表花冠中表达的单基因的充分补充。每碱基覆盖度是 24×，意味着核苷酸平均进行 24 次测序。这种补充因此称为“参考”，因为它代表随后用于本发明方法的一致参考转录物组序列。相似地，重叠群因此称为“单基因”。

[0098] 根据本发明方法的假设，单基因中，一个特异性单基因在用于生成转录物组序列的大多或各突变体中将携带遗传上非同义单核苷酸多态性(SNP)。在此处给出的特异实施例中，负责生成红色花颜色的基因应突变三次，构成 cDNA 池的个体突变体植物中各一次。因为随机 EMS 突变频率典型地为 ~1 次每 200,000 个碱基，如果单基因因果上与表型(在表型的基础上选择突变植物)无关，那么 3 个区别的核苷酸突变将最终发生在任何单个单基因中的机会几乎为零。符合这种突变原则并且其中碱基变化还是非同义的特定基因，这意味着引起该基因的蛋白产物的氨基酸变化，极有可能是造成红色花颜色的基因。在其余上下文中，碱基变化表示为单核苷酸突变(SNM)并且非同义单核苷酸突变表示为 nSNM。可以注意到同义碱基变化，即那些不会引起氨基酸变化的变化，有时可以是遗传上非中性的，如果它们如位于内含子外显子剪接位点。然而注意的是，本发明的方法对于相对罕见的情况，如至少一对或优选一组等位基因用于比较，是强大的，因为剪接错误将导致重大改变的 cDNA。

[0099] 在所有个体 454 序列读取中，通过将它们与上述参考花冠转录物组比，我们分析了花冠 454 序列数据组的 SNM 发生率。只有当读取与参考至少 98％一致时，454 读取的比对用于 SNM 分析。可以通过本领域技术人员已知的各种生物信息学方法进行比对。在本实施例中 SNM 被定义为 :

[0100] (1)在完整的 454 读取组中发生至少 3 次(3× 覆盖度)

[0101] (2)只发生于 3 个个体突变体植物中的一个

[0102] (3)发生于来自个体突变体植物中所有读取的 10％ <N<90％

[0103] 这三个原则证实 SNM 不是测序错误(1)，是独立的(2)，并且在杂合情况下发生(3)。技术人员将理解，根据池中突变体植物的数目、SNM 调取(calling)中 1 型和 2 型错误的水平的规定，这些原则可以制定得更严格或更宽松。此外，在这种实施例中，只有 G->A 或 C->T 的转换被考虑为是可能的 SNM，由于这些是由所用诱变剂(EMS)诱导的主要突变。技术人员将理解，绘图的读取的 98% 同一性阈值可以制定得更严格或更宽松，这种阈值可以影响发现的 SNM 的数目。

[0104] 作为下一个步骤，我们搜索其中存在至少三个 SNM 的单基因，池中三个突变体各一个，并通过上述指定的 SNM 调取原则(1、2、3)进行鉴别。因果上与开红花表型有关的任何单基因必须理论上遵循这种规则，假设测序覆盖度足以使所有 SNM 满足 SNM 调取的三个原则。发现 26 个单基因在编码区含有至少三个 SNM，各自源自不同的突变体植物。其中，20 个含有刚好含有三个 SNM，而 6 个含有 4-7 个 SNM。

[0105] 遵循所有上述 SNM 的形式遗传模式的 26 个单基因随后检验 nSNM 的发生率。首

先，通过找到最长的预测的开放阅读框（ORF），根据标准的将 DNA 三联体翻译成氨基酸的真核遗传密码子，将 26 个单基因的 DNA 序列转换成氨基酸序列。然后使用标准 BLASTp 程序，与蛋白质序列的 UniProt 数据库比较这些 ORF，以便检测同源蛋白质，从而确认预测的 ORF 是在正确的有生物学意义的翻译阅读框中。然后将各 26 参考单基因的蛋白质序列，与它们对应的含有 SNM 的突变体植物的蛋白质序列比较。如果氨基酸变化发生在包括 SNM 的密码子中，SNM 标记为 nSNM。根据定义，只有 nSNM 可以导致编码区内的表型变化。因此，任何不含至少三个 nSNM 的 26 个单基因以及含有至少一个 nSNM 的各突变体植物，作为引起开红花表型的可能候选者而被弃掉。此操作之后，9 个单基因仍然留在候选名单中（表 1B）。

[0106] 留下的 9 个单基因的组随后将根据形式遗传原则进行排序过程。对于存在三个以上的 nSNM 给予罚分，其中如刚好三个突变体植物用于分析的事实所示的，完美的模式是刚好三个。nSNM 越多，排名越低。本操作之后，3 个单基因共同排为第 1。9 个单基因及其排名的完整列表给于表 1C 中，包括和 UniProt 数据库的最佳 BLAST 匹配。可见 RT 基因，根据推理被认为引起开红花的表型并编码 UDP 鼠李糖 :花青素 -3- 葡萄糖苷鼠李糖基转移酶，排名 1。

[0107] 从这些数据可以得出结论，本发明的方法允许在 39306 个单基因的背景中一小组候选者之中鉴别引起突变体表型的单一基因。本领域技术人员将理解，可以在除矮牵牛花以外的任何物种中进行本发明的方法。本领域技术人员将理解，可以用每个池不同的突变体个体数目、绘图期间读取相对于参考转录物组不同的同一性百分比、每 SNM 覆盖度的不同严谨度（stringencies）、或搜索所采用的其它可衡量的设置进行方法。还明确的是可以单独或组合使用用于候选者排名的额外客观标准，其可包括 :

[0108] - 获自原始测序数据的可能的 SNM 的碱基调取质量分数

[0109] - 根据所用突变体植物的池组成以及杂合或纯合状态所遵循的形式遗传预测，在池中个体等位基因中 SNM 最佳分布的匹配度。

[0110] - 和野生型氨基酸相比，由于 nSNM 的氨基酸取代的化学相似性得分

[0111] 材料和方法

[0112] 1）制备 cDNA 用于 GS-FLX 肽测序

[0113] 为了制备适于分析的 cDNA 样本，按照如下步骤。

[0114] I. 根据制造商的说明书，使用 Qiagen RNeasy Plant Mini 试剂盒（Qiagen 公司，27220 Turnberry Lane Valencia, CA 91355）分离正处于着色开始阶段的各突变体植物的花瓣组织的总 RNA。

[0115] II. 根据制造商的说明书，根据用于文库构建操作的 SMART cDNA 合成（Clontech 公司，1290 Terra Bella Avenue, Mountain View, CA 94043, USA）进行 SMART 双链（ds-）cDNA 的合成。

[0116] III. 根据制造商的说明书，使用 TRIMMER cDNA Normalization 试剂盒（Evrogen 公司，Miklukho-Maklaya str, 16/10, 117997, Moscow, Russia）对 cDNA 正态化。

[0117] IV. 使用如下反应设置，通过 BAL31 核酸外切酶（New England Biolabs, 240 County Road, Ipswich, MA 01938-2723, USA）从 ds-DNA 去除 polyA 尾 :

[0118]

2×BAL31 反应缓冲液	25.0 μl
1 U/μl BAL-31	0.25 μl
cDNA (1 μg)	x μl
双蒸水	y μl
总共	50 μl

[0119] 允许根据制造商说明书进行 BAL31 核酸外切酶促反应，直至从 cDNA 的 5′ 和 3′ 端去除 ~50 个核苷酸，正如通过琼脂糖凝胶电泳和观察到 50+50=100 个碱基对的平均分子量大小的变化所判断的。然后根据制造商说明书使用 Qiagen PCR 纯化试剂盒（Qiagen 公司，27220Turnberry Lane Valencia, CA 91355）纯化处理的 cDNA 样本。

[0120] V. BAL31 处理的 cDNA 样本分成 15 个子样本，并且通过使用如下列出的一种限制性酶消化对每个样本进行片段化，括号中为它们的 DNA 识别序列：AccI(GTMKAC)、AluI(AGCT)、ApoI(RAATTY)、AvaI(CYCGRG)、AvaII(GGWCC)、BsaAI(YACGTR)、BsiEI(CGRYCG)、BstYI(RGATCY)、HaeII(RGCGCY)、Hpy99I(CGWCG)、MseI(TTAA)、NspI(RCATGY)、RsaI(GTAC)、SfcI(CTRYAG)、StyI(CGWWGG)。根据它们产生片段的能力选择这些限制性酶，使得片段的大小范围适合 GS-FLX 肽测序，并确保 cDNA 的良好覆盖度以及片段之间良好的重叠概率，来允许建立重叠群。根据酶制造商的说明书，允许反应结束并随后灭活。所有酶获自 New England Biolabs（240County Road, Ipswich, MA 01938-2723, USA）。完成消化之后，每单个突变体植物的所有 15 个限制性反应进行合并，并根据制造商的说明书用 Qiagen PCR 纯化试剂盒（Qiagen 公司，27220 Turnberry Lane Valencia, CA 91355）进行纯化。

[0121] VI 根据制造商的说明书，使用 GS-FLX 肽通用文库制备试剂盒（Roche Diagnostics Corporation, Roche Applied Science, P. O. Box50414, 9115 Hague Road, Indianapolis, IN 46250-0414, USA），将 3-5μg 来自各突变体植物的纯化的 cDNA 分别用于产生用于测序的 GS-FLX 肽文库。用不同的多重识别符（MID）制备个体文库。本工作所用的三个突变体植物的 MID 可区分的文库然后进行合并，并使用 GS-FLX 肽进行两轮测序。

[0122] 生物信息学分析

[0123] 原始 GS-FLX 肽读取的预加工

[0124] 读取经过预加工来去除 GS-FLX 肽通用文库制备试剂盒（Roche Diagnostics Corporation, Roche Applied Science, P. O. Box 50414, 9115Hague Road, Indianapolis, IN 46250-0414, USA）中提供的 MID 和接头序列。

[0125] 参考转录物组序列的拼接

[0126] 用 454 Life Sciences 的 Newbler 拼接软件（454 Life Sciences, 15Commercial Street, Branford, CT 06405, USA），将预加工的序列作为输入，使用 454 的缺省运行拼接设置，完成拼接。

[0127] 分析

[0128] 使用 98% 的最小同一性，将预加工的读取绘制于拼接的参考重叠群上。通过在单基因中的不同位置上具有至少三个 SNM 以及通过与不同的突变体植物个体的相关性，选择

候选单基因。从 3× 的最小覆盖度调取 SNM。SNM 的性质被迫使为 G->A 或 C->T。每 SNM 位置，只允许一个，并且是杂合的，突变体植物。如果在每个体植物中以≥ 90% 的情况出现相同的核苷酸，那么核苷酸位置被认为是纯合的。

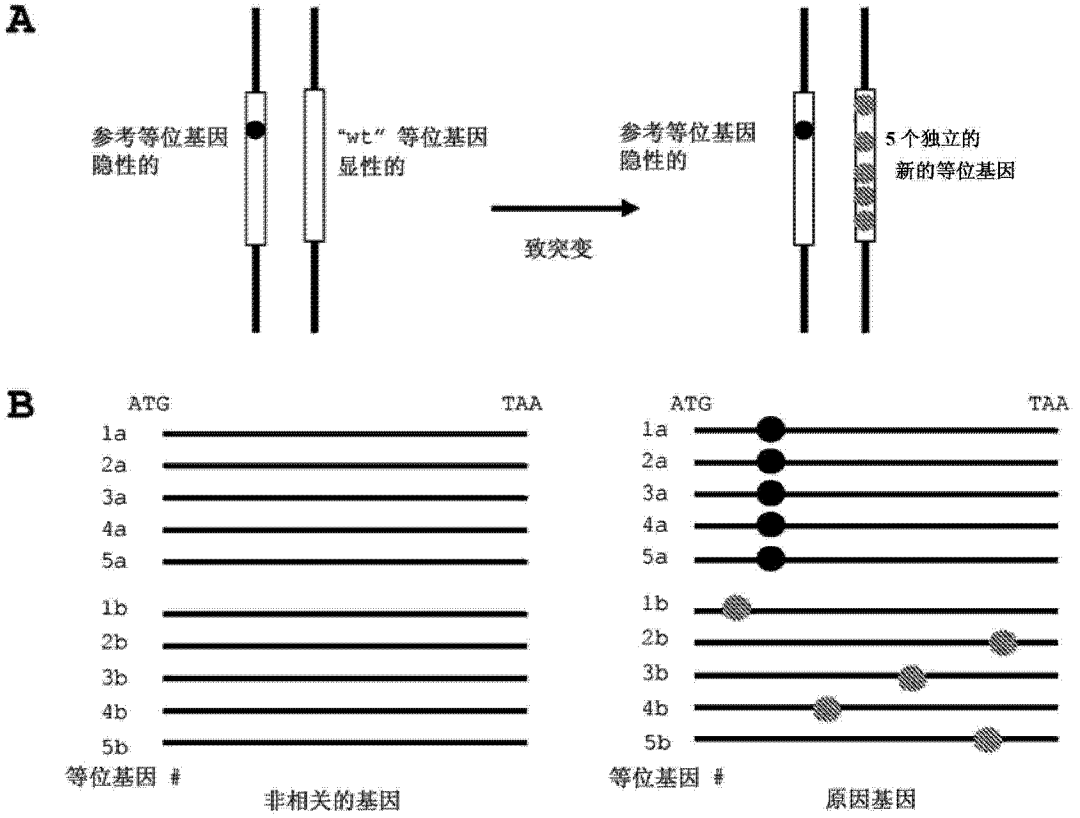

图 1

表 1

A

GS-FLX 钛测序	参考转录物组特征
GS-FLX 钛读取的 nr.	1977422
GS-FLX 钛碱基的 nr.	592944352
平均读取长度（nt）	299.86
重叠群总 nr.	39306
平均重叠群长度（nt）	955
重叠群长度中值（nt）	1025
总重叠群复杂性（nt）	24841943
重叠群中每个碱基平均覆盖度	23.87

B

标准 SNM 尺度 =3 个杂合的个体，≥3 不同的位置
与参考重叠群相比，98% 读取同一性，≥3X 覆盖度每个等位基因

步骤	尺度	# 单基因	% 参考
步骤 d	标准	9	0.023
步骤 e	在≥2 个个体中的 nSNM	9	0.023
步骤 f	>3 个不同 nSMN 的罚分	3	0.008

C

步骤 g	鉴别的 cDNA 同源物	acc.nr.	# nSNM
排名 1	花青素 -3-0 葡萄糖基转移酶	**>sp|Q43716**	3
排名 1	血管 ass. 膜蛋白	**>gi|18415696**	3
排名 1	无同源性	-	3
排名 2	RNA 聚合酶相关的	>gi|22331241	4
排名 3	RNA 结合蛋白	>gi|18651027(	4
排名 4	未知蛋白	>gi|15242587	5
排名 5	谷胱甘肽 S- 转移酶	>gi|15218517	5
排名 6	核糖体蛋白 S13	>gi|18411224	6
排名 7	母源胚胎捕获蛋白 59	>gi|15235514	7

图 2