



(12)发明专利



(10)授权公告号 CN 105492625 B

(45)授权公告日 2020.04.07

(21)申请号 201480021755.8

(22)申请日 2014.04.17

(65)同一申请的已公布的文献号
申请公布号 CN 105492625 A

(43)申请公布日 2016.04.13

(30)优先权数据
61/812876 2013.04.17 US
61/813001 2013.04.17 US

(85)PCT国际申请进入国家阶段日
2015.10.16

(86)PCT国际申请的申请数据
PCT/US2014/034476 2014.04.17

(87)PCT国际申请的公布数据
W02014/172529 EN 2014.10.23

(73)专利权人 先锋国际良种公司
地址 美国依阿华州

(72)发明人 M.贝蒂 K.R.哈耶斯 J.L.霍夫曼
H.林 G.M.扎斯特罗-哈耶斯

(74)专利代理机构 中国专利代理(香港)有限公
司 72001
代理人 唐华东 黄希贵

(51)Int.Cl.

C12Q 1/6869(2018.01)

(56)对比文件

CN 1434873 A,2003.08.06,
Jennifer Monson-Miller等.reference
genome-independent assessment of mutation
density using restriction enzyme-phased
sequencing.《BMC genomics》.2012,摘要、图1、
图5、第12页左栏倒数第3段-右栏倒数第2段、第
13页右栏第3段.

Dorian Fabian Urbanski等.genome-wide
LOREl retrotransposon mutagenesis and
high-throughput insertion detection in
Lotus japonicus.《the plant journal》.2011,
摘要、第732页最后一段-第733页左栏倒数第2
段、第734-735页图1、第738页左栏第2-3段、第
739页左栏第4段-右栏第3段、第739-740页跨页
段.

审查员 罗洋

权利要求书2页 说明书16页
序列表2页 附图11页

(54)发明名称

用于在基因组中表征DNA序列组成的方法

(57)摘要

本文公开了针对转基因事件进行高通量分析的方法。所述方法使用经剪切的基因组DNA的文库,其连接到专门的接头并且合并以用于序列分析以及与已知基因组和插入序列的比较。所述方法可用于检测表征插入位点、转基因完整性,以及转基因拷贝数。

1. 一种用于在样品植物的基因组中表征靶序列的方法,所述方法包括:
 - a) 分离和纯化基因组DNA的样品;
 - b) 将所述基因组DNA剪切成片段以创建文库,其中所述经剪切的基因组DNA片段在200个碱基对长度至2.5kb长度的范围内;
 - c) 将所述文库片段连接到具有条形码的接头序列;
 - d) 设计构建体特异性PCR引物,所述构建体特异性PCR引物由第一PCR引物对和第二PCR引物对组成,所述第一PCR引物对由插入序列特异性引物和接头特异性引物组成,所述第二PCR引物对由接头特异性引物和嵌套的PCR引物组成;其中所述第一和第二PCR引物对被设计用于交替链上的每200个碱基对,或用于单链上的每400个碱基对;
 - e) 使用d) 所述的第一和第二PCR引物对扩增所述DNA,从而富集DNA序列的文库;
 - f) 以相等摩尔比将富集的文库合并到样品池中;
 - g) 对所述样品池进行测序以获得读数;
 - h) 将所述读数过滤并且与对照植物的基因组序列以及与所关注靶序列比对,其中,对步骤(g)中获得的所述读数进行处理以移除任何接头序列信息,选择得自步骤(g)的前60%最丰富的读数以用于与对照植物的所述基因组序列以及与所述所关注靶序列的比对;并且将与对照植物的基因组对齐的任何读数从进一步的分析中排除;
 - i) 选择与所述所关注靶序列对齐的读数;
 - j) 从所述所选择的读数确定连接序列,其中所述连接序列表示DNA的包含接点的区段,并且是通过比对读数和识别其中读数不再与靶序列对齐的接点来识别的,所述接点是指DNA的两个不同片段接合在一起的点;以及
 - k) 使用步骤j) 所述的连接序列在所述样品植物的所述基因组中表征所述所关注靶序列的完整性。
2. 根据权利要求1所述的方法,其中所述经剪切的基因组DNA片段在200个碱基对长度至1kb长度的范围内。
3. 根据权利要求1所述的方法,其中所述经剪切的基因组DNA片段的长度为400个碱基对。
4. 根据权利要求1所述的方法,其中所述第一和第二PCR引物对是嵌套的。
5. 根据权利要求1所述的方法,其中所述第一和第二PCR引物对是重叠的。
6. 根据权利要求1所述的方法,其中分析所述富集的文库的PCR伪迹。
7. 根据权利要求1所述的方法,其中步骤(g)中的测序产生至少100万个读数。
8. 根据权利要求1所述的方法,其中步骤(g)中获得的所述读数为100bp配对末端读数。
9. 根据权利要求1所述的方法,其中在步骤(j)中确定所述连接序列包括比对对应于连接序列的至少两个读数以及识别共有连接序列。
10. 根据权利要求9所述的方法,其中所述共有连接序列包含:
与所述对照植物的所述基因组序列对齐的30个核苷酸,以及
与所述所关注靶序列对齐的20个核苷酸。
11. 根据权利要求1所述的方法,其中所述连接序列延伸以提供100至3000个核苷酸的连接重叠群。
12. 根据权利要求11所述的方法,其中通过在所述连接序列的每个末端处比对重叠读

数来延伸所述连接序列以便识别位于所述连接序列的每个末端旁侧的核苷酸序列。

13. 根据权利要求11所述的方法, 其中使用SSAKE来延伸所述连接序列。

14. 根据权利要求1所述的方法, 其中所述连接序列与对照植物的所述基因组序列以及与所述所关注靶序列比对。

15. 根据权利要求1所述的方法, 还包括在植物育种程序中使用所述样品植物。

用于在基因组中表征DNA序列组成的方法

技术领域

[0001] 本发明涉及植物生物技术领域。更具体地讲,本发明涉及确定植物基因组内的DNA序列组成的方法。

[0002] 对通过EFS-WEB作为文本文件提交的

[0003] 序列列表的引用

[0004] 遵照美国信息交换标准码(American Standard Code for Information Interchange(ASCII)),通过EFS-Web将序列列表的正式文本作为文本文件与本发明说明书同时提交,文件名为431978seqlist.txt,创建日期为2013年4月17日,文件大小为2Kb。通过EFS-Web提交的序列列表是说明书的一部分,藉此将其全文以引用的方式并入本文。

背景技术

[0005] 当含有所关注的靶序列的质粒转化到植物中时,需要进行测试来确认转化已发生并且评估转化的质量。例如,当在已用相同构建体转化的多个植物中选择时,所选择的植物应具有完整的所关注靶序列,该完整靶序列没有重排、插入、缺失、或外源侧翼序列。

[0006] 历史上,Southern印迹方法已用于确认质粒构建体的转化并且识别潜在的重排、多拷贝、或部分事件。Southern印迹实验可为耗时的、提供低分辨率、具有高成本,并且需要多次手动检查。此外,Southern印迹方法既不能识别靶序列整合位点,也不能识别整合位点的用于设计事件特异性PCR实验的侧翼序列。

[0007] 侧翼序列分析(FSA)已成功地用于识别转基因整合位点并且用于获得插入位点的侧翼序列。然而,由于FSA仅针对有限的边界区域,因此FSA不检测潜在的部分片段、重排、或超出目标边界区域的截短。此外,由FSA使用的搜索算法可识别由质粒的参考序列中的任何误差造成的假阳性。

[0008] 因此,一直需要迅速、低成本的方法来有效地表征插入植物基因组中的靶序列的位置、数量和完整性。

发明内容

[0009] 基于测序的Southern(SbS)是整合的高通量序列和生物信息学分析流水线,其针对大规模事件选择和推进决策而评估和表征转化事件。SbS实施一系列过滤策略来确保检测的精确度和灵敏度。通过用富集含有所关注的靶序列的片段的shotgun文库开始,SbS可迅速过滤出内源读数并且识别连接序列。连接序列被识别和延伸之后,接点被定位到植物基因组和靶序列构建体以确定插入序列的位置、数量和完整性。SbS可检测小的部分片段并且容忍质粒的参考序列中的误差。

附图说明

[0010] 图1汇总了SbS数据分析流水线。

[0011] 图2示出了来自识别染色体2上的所关注靶序列的单个插入的SbS流水线的输出数

据。

[0012] 图3a和图3b示出与靶序列片段的插入相邻的靶序列的单个插入。靶序列插入与靶序列片段插入之间的接点被识别为构建体：构建体插入，而靶序列与植物基因组或者靶序列片段与植物基因组之间的接点被识别为构建体：基因组。

[0013] 图4a和图4b示出复杂插入事件，其中连接序列在染色体6和染色体9上被检测到。靶序列在染色体9上的插入被复制并且处于相反方向。

[0014] 图5识别靶序列的截短插入，证据是其不存在与农杆菌 (*Agrobacterium*) 构建体上的靶DNA的一部分对准的读数。

[0015] 图6识别农杆菌骨架的一部分插入到植物的基因组中。农杆菌骨架的插入可通过读数 (黑盒) 与骨架的一部分的对准来识别。

[0016] 图7a和图7b描述了合成连接序列的方法。第一个表示出在合成步骤之前预测的推定接点。合成脚本根据相同取向中的每个接点的30_20聚体对所有接点支撑读数进行分组。对于足够接近的两个接点 (默认距离2bp)，如果30_20聚体在偏移距离后相同，则两个接点冷凝成一个。如图所示，两个接点为11708和11709。在被合成之后，具有更多独特支撑读数的接点 (接点11708) 从接点11709接管支撑读数。粗体的核苷酸表示单核苷酸多态性 (SNP) 并且列出的序列通过算法的分裂和合成特征来移除。

具体实施方式

[0017] 借助前面的描述中给出的教导内容，本发明所属领域的技术人员将会想到本文所述的发明的许多修改形式和其他实施例。因此，应当理解，本发明不限于所公开的特定实施例，并旨在将修改形式和其他实施例包括在所附权利要求的范围内。虽然采用特定术语，但所述术语仅在一般性和描述性意义上使用而并非用于限制目的。

[0018] 对于商业产品批准的遗传修饰的作物的表征目前需要转基因DNA插入序列的详细分子表征和转基因基因座的完整性。此外，分子分析是产品开发过程中事件选择和推进决策的重要组成部分。

[0019] 已知外来基因在植物中的表达受它们在植物基因组中的位置影响，这可能归因于染色质结构 (例如，异染色质) 或转录调节元件 (例如，增强子) 接近整合位点的接近度 (Weising et al. (1988) *Ann. Rev. Genet.* 22:421-477 (Weising等人, 1988年, 《遗传学年评》，第22卷, 第421-477页))。与此同时，转基因在基因组中不同位置处的存在将以不同方式影响植物的总体表型。此外，插入转基因的拷贝数可影响植物的表型。

[0020] 出于这个原因，经常需要筛选大量事件，以识别出以引入的所关注基因的最佳表达为特征的事件。例如，已经在植物及在其他生物体中观察到，在事件之间可能存在引入基因的表达水平的巨大差异。在表达的空间或时间模式方面也可以存在差异，例如，转基因在多种植物组织中相对表达方面的差异，这可能不对应于从存在于引入的基因构建体中的转录调节元件所预期的模式。

[0021] 因此，常见的是产生数百至数千个不同事件，并从这些事件中筛选出具有所需的转基因表达水平和用于商业目的模式的单一事件。具有所需的转基因表达水平或模式的事件可用于将转基因通过有性远交或其他常规育种方法渐渗至其他遗传背景中。此类杂交的子代维持原始转化体的转基因表达特征。使用该杂交育种策略以在充分适应于局部生长条

件的众多品种中确保可靠的基因表达。

[0022] 通常,该分子分析依赖于Southern印迹以确定横跨任何插入DNA的PCR产物的基因座和拷贝数以及靶标序列来完成表征过程。Southern印迹的缺点包括:低通量,每个样品成本高,未知的序列组成和位置,以及检测到的DNA片段缺乏完整性。

[0023] 最近,通过生物信息学的下一代(NextGen)测序和连接序列分析已在成本和时间上均优于Southern印迹分析。

[0024] 本发明涉及靶序列的扩增或捕获,合并扩增或捕获的序列以及通过DNA测序对合并样品的表征。DNA序列数据被汇集并且与参考序列相比。转基因插入的表征在植物、动物、和微生物物种,人类疾病诊断、靶序列的单个或多个拷贝的基因组位置、以及纯度测试中是有用的。

[0025] 本发明进一步涉及植物的基因组中的所关注靶序列的生物信息学分析和表征的方法。当与本发明的新型扩增和捕获方法结合时,该表征被称为基于测序的Southern(SbS)。如本文所用,术语“基于测序的Southern”是指连续的一系列步骤,所述步骤捕获和扩增DNA,合并样品,以及分析序列数据来表征体内结构。

[0026] 本文所用的冠词“一个”和“一种”是指一个(种)或不止一个(种)(即,指至少一个(种))该冠词的语法对象。举例来说,“一个要素”是指一个或多个要素。

[0027] 在整个公开中,各种专门术语将被使用并且在本文中如下定义:

[0028] DNA“构建体”是被连接在一起的提供一个或多个表达盒的DNA分子的装配物。DNA构建体可以是能够在细菌细胞中自我复制并且含有可用于引入DNA分子的多个核酸内切酶限制性位点的质粒,所述DNA分子提供功能性遗传元件即启动子、内含子、前导序列、编码序列、3'终止区及其他;或DNA构建体可以是DNA分子的线性装配物,诸如表达盒。包含在DNA构建体内的表达盒包括提供信使RNA转录的必需的遗传元件。可以将表达盒设计成在原核细胞或真核细胞中表达。将本发明实施例的表达盒设计成在植物细胞中表达。

[0029] “转基因”是已经通过转化程序引入基因组中的基因。植物基因组中已经插入重组DNA的位点可以称作“插入位点”或“靶位点”。

[0030] 当重组DNA通过传统杂交引入到植物中时,其侧翼区一般不改变。如本文所用,“侧翼区”或“侧翼序列”是指至少20bp、优选地至少50bp和至多5000bp的序列,其紧邻原始外来插入DNA分子的上游或下游并与该DNA分子保持连续。

[0031] 如本文所用,“插入DNA”是指在用来转化植物材料的表达盒内部的异源DNA,而“侧翼DNA”可由生物体诸如植物中天然存在的基因组DNA或经转化方法引入的相对原始插入DNA分子为外源的外来(异源)DNA(例如,与转化事件相关的片段)组成。

[0032] 应当理解,如本文所用的术语“转基因的”包括任何细胞、细胞系、愈伤组织、组织、植物部分、或植物,其基因型已经因异源核酸的存在而改变,包括最初如此改变的那些转基因以及通过有性杂交或无性繁殖从初始转基因产生的那些转基因。如本文所用,术语“转基因”不涵盖通过常规植物育种方法或通过自然发生的事件(如随机异花受精、非重组病毒感染、非重组细菌转化、非重组转座或自发突变)进行的基因组(染色体或染色体外)变更。

[0033] 通过以下方式产生转基因“事件”:用异源DNA构建体(包括包含所关注转基因的核酸表达盒)转化植物细胞,由于所述转基因插入到所述植物的基因组中而再生出一批植物,以及选择以插入到特定基因组位置中为特征的特定植物。事件通过转基因的表达在表型上

得以表征。在遗传水平上,事件是植物的基因构成的一部分。术语“事件”还指通过转化体和包含异源DNA的另一个品种之间有性远交所产生的子代。甚至与轮回亲本反复回交后,插入的DNA和来自转化亲本的侧翼DNA存在于杂交子代中相同的染色体位置处。术语“事件”也指来自原始转化体的包含所插入DNA和与插入DNA紧邻的侧翼序列的DNA,其中预期所述的DNA转移至子代,所述子代因包括所述插入DNA的一个亲本系(例如,原始转化体和因自交产生的子代)与不含所述插入DNA的亲本系的有性杂交而接受包含所关注转基因的插入DNA。

[0034] “转化”是指核酸片段转移至宿主生物体的基因组中,从而得到遗传上稳定的遗传性。含有转化的核酸片段的宿主生物体称作“转基因”生物体。植物转化方法的例子包括农杆菌介导的转化法(De Blaere et al. (1987) *Meth. Enzymol.* 143:277 (De Blaere等人, 1987年,《酶学方法》,第143卷,第277页))和粒子加速或“基因枪”转化技术(Klein et al. (1987) *Nature (London)* 327:7073 (Klein等人, 1987年,《自然(伦敦)》,第327卷,第70-73页);美国专利No. 4,945,050,所述文献以引用方式并入本文)。

[0035] 转化体在一片异源插入DNA和基因组DNA或两(2)片基因组DNA或两(2)片异源DNA之间含有独特接点。“接点”是两个(2)特定DNA片段接合的点。例如,插入DNA接合侧翼DNA之处存在的接点。接点也存在于其中两个(2)DNA片段以改自于天然生物体中存在的方式接合在一起的转化生物体中。“接点DNA”是指包含接点的DNA。

[0036] “PCR”或“聚合酶链反应”是用于扩增特定DNA片段的技术(参见美国专利No. 4,683,195和No. 4,800,159;其以引用的方式并入本文)。在PCR方案中,可设计寡核苷酸引物用于PCR反应,以从提取自任何所关注生物体的cDNA或基因组DNA扩增(或“捕获”)相应的DNA序列。设计PCR引物和PCR克隆的方法是本领域熟知的,并且公开于以下文献中: Sambrook et al. (1989) *Molecular Cloning: A Laboratory Manual* (2d ed., Cold Spring Harbor Laboratory Press, Plainview, New York) (Sambrook等人, 1989年,《分子克隆: 实验室手册》,第2版,纽约普莱恩维尤冷泉港实验室出版社)。另外参见Innis et al., eds. (1990) *PCR Protocols: A Guide to Methods and Applications* (Academic Press, New York) (Innis等人编辑, 1990年,《PCR方案: 方法和应用指南》,纽约学术出版社); Innis and Gelfand, eds. (1995) *PCR Strategies* (Academic Press, New York) (Innis和Gelfand编辑, 1995年,《PCR策略》,纽约学术出版社);以及Innis and Gelfand, eds. (1999) *PCR Methods Manual* (Academic Press, New York) (Innis和Gelfand编辑, 1999年,《PCR方法手册》,纽约学术出版社)。已知的PCR方法包括但不限于利用成对引物、巢式引物、单一特异性引物、简并引物、基因特异性引物、载体特异性引物、部分错配引物等的方法。

[0037] “探针”是与常规的可检测标记或报道分子(例如放射性同位素、配体、化学发光试剂、或酶)连接的分离的核酸。这样的探针与来自包括来自事件的DNA的样品的靶核酸的链互补。根据本发明的探针不仅包含脱氧核糖核酸或核糖核酸,而且包含与靶DNA序列特异性结合并且可以用来检测这种靶DNA序列存在的聚酰胺及其他探针材料。

[0038] “引物”是分离的核酸,其通过核酸杂交与互补性靶DNA链退火以形成引物和靶DNA链之间的杂交体,随后由聚合酶(例如,DNA聚合酶)沿靶DNA链延伸。本发明的引物对是指它们用于扩增靶核酸序列的用途,例如,通过PCR或其他常规的核酸扩增方法

[0039] 探针和引物要有足够的核苷酸长度,以在操作员确定的杂交条件或反应条件下与靶DNA序列特异性地相结合。该长度可以是任何长度,其是在选择的检测方法中有用的足够

长度。通常,使用11个或更多个核苷酸的长度、18个或更多个核苷酸的长度、和22个或更多个核苷酸的长度。此类探针和引物在高严格性杂交条件下与靶序列特异性杂交。根据本发明实施例的探针和引物可以与靶序列具有完全连续核苷酸DNA序列相似性,但是可以通过常规方法设计与靶DNA序列不同并保持与靶DNA序列杂交的能力的探针。探针可以作为引物使用,但是通常设计成与靶DNA或RNA结合并且不通常用于扩增过程中。

[0040] 特异性引物可以用来扩增整合片段以产生可以作为识别生物学样品中事件的“特异性探针”使用的扩增子。当探针与生物学样品的核酸在允许探针与所述样品结合的条件下杂交时,这种结合可以被检测到并因此允许指示事件的存在。已经在本领域中描述了对结合的探针的这种识别。在本发明的一个实施例中,特异性探针是这样的序列,所述序列在优化条件下与事件的所需区域特异性杂交并且还包含与所述区域邻接的外来DNA的部分。特异性探针可以包含与事件的特定区域至少80%、80%至85%、85%至90%、90%至95%、和95%至100%相同(或互补)的序列。

[0041] “所关注靶序列”可为天然或非天然的、整合或部分整合到植物基因组中的任何核苷酸序列。在某些实施例中,所关注靶序列为异源序列。涉及多核苷酸序列的“异源”为起源于外来物种或人工源的序列,或者,如果起源于相同物种的话,则为通过蓄意的人为干预从其天然形式在组成和/或基因座方面进行实质性修饰得到的序列。所关注靶序列可包括但不限于:转基因、天然性状、或天然或诱导的突变。

[0042] 所关注靶序列可反映出作物开发的参与者的商业市场和利益。所关注作物和市场改变,并且随着发展中国家打开世界市场,将出现新的作物和技术。另外,随着我们对农学性状和特性诸如产量和杂种优势的理解的增加,对用于转化的靶序列的选择将会相应变化。所关注靶序列的大体类别包括例如涉及信息的那些靶序列(诸如锌指)、涉及通信的那些靶序列(诸如激酶)和涉及看家的那些靶序列(诸如热休克蛋白)。靶序列的更具体类别例如包括编码对农艺学、昆虫抗性、病害抗性、除草剂抗性、不育性、籽粒特性和商业产品重要的性状的多核苷酸。一般而言,所关注靶序列包括涉及油脂、淀粉、碳水化合物或营养物质代谢的那些以及影响仁尺寸、蔗糖载量等的那些。

[0043] 除了使用传统的育种方法之外,还可通过遗传方式变更农艺上重要的性状诸如油脂含量、淀粉含量和蛋白质含量。修饰包括增加油酸、饱和或不饱和油的含量、增加赖氨酸和硫的水平、提供必需氨基酸以及修饰淀粉。美国专利No.5,703,049、No.5,885,801、No.5,885,802和No.5,990,389中描述了Hordothionin蛋白修饰,将这些专利以引用的方式并入本文。另一个例子是美国专利No.5,850,016中所描述的由大豆2S白蛋白编码的富赖氨酸和/或富硫种子蛋白,和Williamson et al.(1987) Eur. J. Biochem. 165:99-106 (Williamson等人,1987年,《欧洲生物化学杂志》,第165卷,第99-106页)中所述的来自大麦的胰凝乳蛋白酶抑制剂,将所述专利和文献的公开内容以引用方式并入本文。

[0044] 可通过定点诱变产生编码序列的衍生物来增加预选的氨基酸在所编码的多肽中的水平。例如,编码大麦高赖氨酸多肽(BHL)的基因源自大麦胰凝乳蛋白酶抑制剂(1996年11月1日提交的美国专利申请序列号08/740,682和W0 98/20133,将它们的公开内容以引用方式并入本文)。其他蛋白质包括富含甲硫氨酸的植物蛋白,诸如来自葵花籽(Lilley et al.(1989) Proceedings of the World Congress on Vegetable Protein Utilization in Human Foods and Animal Feedstuffs, ed. Applewhite (American Oil Chemists

Society, Champaign, Illinois), pp. 497-502 (Lilley等人, 1989年, 植物蛋白在人类食品和动物饲料中的利用的世界大会会议录, Applewhite编辑 (美国油类化学家学会, 伊利诺伊州香巴尼市), 第497-502页); 其以引用方式并入本文); 玉米 (Pedersen et al. 1986) J. Biol. Chem. 261:6279 (Pedersen等人, 1986年, 《生物化学杂志》, 第261卷, 第6279页); Kirihara et al. (1988) Gene 71:359 (Kirihara等人, 1988年, 《基因》, 第71卷, 第359页), 这两篇文献均以引用方式并入本文); 和水稻 (Musumura et al. (1989) Plant Mol. Biol. 12:123 (Musumura等人, 1989年, 《植物分子生物学》, 第12卷, 第123页), 其以引用方式并入本文)。其他农艺上重要的靶序列编码胶乳、Floury 2、生长因子、种子贮藏因子和转录因子。

[0045] 昆虫抗性靶序列可以编码针对严重影响产量的害虫的抗性, 所述害虫诸如根虫、切根虫、欧洲玉蜀黍螟等。这种多核苷酸包括例如苏云金杆菌 (*Bacillus thuringiensis*) 毒蛋白基因 (美国专利No. 5,366,892; No. 5,747,450; No. 5,736,514; No. 5,723,756; No. 5,593,881; 和 Geiser et al. (1986) Gene 48:109 (Geiser等人, 1986年, 《基因》, 第48卷, 第109页)); 等等。

[0046] 编码病害抗性性状的靶序列包括解毒基因, 诸如抗伏马毒素 (美国专利No. 5,792,931); 无毒性 (avr) 和病害抗性 (R) 多核苷酸 (Jones et al. (1994) Science 266:789 (Jones等人, 1994年, 《科学》, 第266卷, 第789页); Martin et al. (1993) Science 262:1432 (Martin等人, 1993年, 《科学》, 第262卷, 第1432页); 和 Mindrinis et al. (1994) Cell 78:1089 (Mindrinis等人, 1994年, 《细胞》, 第78卷, 第1089页)); 等等。

[0047] 除草剂抗性性状可包括编码对用于抑制乙酰乳酸合酶 (ALS) 的作用的除草剂 (特别是磺酰脲类除草剂) 的抗性的靶序列 (例如, 含有导致这种抗性的突变, 特别是S4和/或Hra突变的乙酰乳酸合酶 (ALS) 多核苷酸)、编码对用于抑制谷氨酰胺合酶的作用的除草剂 (诸如草胺膦或basta) 的抗性的靶序列 (例如bar基因); 草甘膦 (例如, EPSPS基因和GAT基因; 参见例如美国专利公布No. 20040082770和WO 03/092360); 或者本领域已知的其他此类多核苷酸。bar基因编码对除草剂basta的抗性, nptII基因编码针对抗生素卡那霉素和遗传霉素的抗性, ALS基因突变体编码针对除草剂氯磺隆的抗性。

[0048] 靶序列还包含不育基因。不育基因可为物理去雄提供另选方案。以这种方式使用的多核苷酸的例子包括雄性组织偏好的基因和具有雄性不育表型的基因 (诸如QM), 在美国专利No. 5,583,210中有描述。其他靶序列包括激酶和编码对雄性或者雌性配子体发育有毒的化合物的那些。

[0049] 谷粒的质量反映在诸如以下的性状: 饱和及不饱和油的水平和类型、必需氨基酸的质量和数量、纤维素的水平。在玉米中, 经修饰的hordothionin蛋白在美国专利No. 5,703,049、No. 5,885,801、No. 5,885,802和No. 5,990,389中有所描述。

[0050] 还可在靶序列上编码商业性状, 所述靶序列可增加例如用于乙醇生产的淀粉, 或提供蛋白的表达。经转化的植物的另一个重要商业用途是生产聚合物和生物塑料, 诸如美国专利No. 5,602,321中所述。诸如 β -酮基硫解酶、PHB酶 (聚羟基丁酸合酶) 和乙酰乙酰基-CoA还原酶之类的靶序列 (参见 Schubert et al. (1988) J. Bacteriol. 170:5837-5847 (Schubert等人, 1988年, 《细菌学杂志》, 第170卷, 第5837-5847页)) 有利于聚羟基链烷酸酯 (PHA) 的表达。

[0051] 本文所述的方法可用于表征在任何所关注植物的基因组中的靶序列的数量、位置和完整性。为了获得适用于生物信息学流水线中的分析的序列读数,必须对特异于靶序列和相应连接序列的DNA的多个片段进行测序。可制备对应于所关注序列和/或内源基因组DNA的DNA片段以用于通过适用于对短片段进行高通量测序的任何方法来进行测序。在一些实施例中,使用shotgun克隆策略来制备DNA片段以产生用于高通量双脱氧核苷酸测序或下一代测序的模板。

[0052] 为了制备用于测序的DNA shotgun文库,来自样品植物的基因组DNA必须被收集和分离。如本文所用,术语“样品植物”是指具有所关注序列的任何植物。在一些实施例中,所关注序列对样品植物是异源的。因此,本文所公开的方法可用于检测样品植物中所关注序列的存在。一般来讲,植物基因组DNA的分离导致获得不含脂质、蛋白质和其他细胞碎片的纯化植物DNA。优选的植物DNA分离方法包括:裂解、加热、醇沉淀、盐沉淀、有机萃取、固相萃取、硅胶膜萃取、CsCl梯度纯化,以及它们的任何组合。在一些实施例中,基因组DNA可通过Stacey和Isaac(1994 In *Methods in Molecular Biology* Vol.28, pp.9-15, Ed.P.G.Isaac, Humana Press, Totowa, N.J. (1994年,载于《分子生物学方法》,第28卷,第9-15页, P.G.Isaac编辑,胡马纳出版社,新泽西州托托瓦))所述的CTAB(鲸蜡基三乙基溴化铵, Sigma H5882)方法,佐治亚州诺克斯市的奥美嘉生物技术公司(Omega Biotek (norcross, GA))的EZNA植物96试剂盒,或以DNeasy试剂盒(加利福尼亚州瓦伦西亚市凯杰公司(Qiagen, Valencia, Calif.))市售的硅胶膜技术与样品植物分离。在分离之后,来自样品植物的基因组DNA被剪切以提供适用于文库构建的基因组DNA的多个片段。植物DNA的剪切可利用经酶促的超声处理,通过加热、或任何其他适用于产生配合shotgun文库构建的基因组DNA片段的方法来实现。在一个实施例中,基因组DNA通过超声处理来剪切。

[0053] 在一些实施例中,来自样品植物的基因组DNA通过超声处理来剪切,对其进行末端修复,加A尾,并且连接到接头序列。可使用本领域中适用于高通量测序的任何方法来对片段末端进行修复和加A尾。在某些实施例中,接头连接到剪切基因组DNA的末端以使样品能够在杂交和测序阶段合并。在连接到接头序列之后,基因组DNA片段可通过PCR扩增。例如,具有接头序列的DNA片段可通过5、6、7、8、9、10、12、15个或任何数目的PCR循环来扩增以产生所需数量的扩增基因组片段以用于测序。

[0054] 在某些实施例中,序列捕获可用于富集包含所关注靶序列的区域的的那些的基因组DNA片段的池。如本文所用,术语“富集”或“富含”是指增加特定组的基因组DNA片段的浓度。例如,在对于具有所关注靶序列的区域的DNA片段进行富集之后,片段文库将比富集处理之前包含更高比例的具有所关注靶序列的区域的片段。如本文所用,术语“靶序列的区域”是指对应于靶序列中的任何多核苷酸。在某些实施例中,靶序列的区域包含对应于所关注靶序列中的至少2个、至少5个、至少10个、至少15个、至少20个、至少25个、至少30个、至少35个、至少40个、至少45个、至少50个、至少60个、至少70个、至少80个、至少90个、至少100个、至少125个、至少150个、至少175个、至少200个、至少225个、至少250个、至少275个、至少300个、至少350个、或至少400个连续核苷酸。

[0055] “序列捕获”是指使用特异于所关注多核苷酸的探针从DNA文库选择特定DNA片段的方法。在一些实施例中,探针被设计为表示靶序列的所有核苷酸的独特序列。任何数目的探针都可用于本文所公开的方法中,而且探针的数目将随靶序列的长度而改变。例如,至少

2个、至少25个、至少50个、或约100个、约200个、约300个、约400个、约500个、约750个、约1000个、约1500个、约2000个、约3000个、约4000个、约5000个、约10,000个、约50,000个、约100,000个、约250,000个、约500,000个、约750,000个、或约1,000,000个独特的探针可被设计成覆盖所关注靶序列的整个长度。在一些实施例中,探针被修饰成包含促进后续捕获以及探针/DNA片段复合物的纯化的特征。例如,在一些实施例中,探针被修饰成具有稍后可用链亲和素珠捕获的生物素标记。序列还可使用具有附接到表面的探针的微阵列载玻片来捕获和富集。

[0056] 基于序列捕获的方法可用于富集DNA片段文库,该文库针对包含所关注靶序列的区域的那些片段。序列捕获方法可根据在位于www.nimblegen.com的Roche NimbleGen网站可获得的标准方案来执行。简而言之,DNA shotgun文库在混合物中利用杂交缓冲液变性,所述杂交缓冲液为诸如SeqCap EZ显色试剂、和对应于在文库构建中使用的任何接头序列的阻断寡核苷酸。在变性之后,混合物与生物素酰化探针文库混合并且被温育以允许探针文库与基因组片段文库杂交。在杂交之后,混合物与链亲和素珠混合并且随后洗涤和洗脱以提供与所关注靶序列具有同源性的结合DNA片段。在一个实施例中,洗涤和洗脱的文库可通过经历3、4、5、6、7、8、9、10、12、15个或任何数目的扩增循环(例如,PCR)来扩增以提供足够数量的DNA以用于测序或又一轮的序列捕获。因此,在一些实施例中,为了针对具有所关注靶序列的区域的片段进一步富集DNA片段文库,可使用多轮序列捕获。例如,可使用与所关注靶序列具有同源性的探针来执行2轮、3轮或4轮序列捕获。

[0057] 在最后一轮的序列捕获之后,可将DNA片段文库合并、扩增并且在制备中纯化以用于高通量测序。例如,DNA片段文库可通过5、6、7、8、9、10、12、15、17、20、25个或任何数目的扩增循环(例如,PCR)来扩增以提供足够数量的DNA以用于测序。可通过本领域中已知的任何方法(例如,Qiagen Qiaquick柱)对DNA片段文库进行纯化。

[0058] 在对于具有所关注靶序列的区域的那些DNA片段进行选择之后,可使用本领域中已知的任何方法对DNA片段文库进行测序。在一些实施例中,DNA片段文库可通过下一代测序来进行测序。短语“下一代测序”或NGS是指与传统的基于Sanger和基于毛细管电泳的方法相比具有增加通量的测序技术,该测序技术例如具有一次产生成百上千的相对较小的序列读数的能力。下一代测序技术的一些例子包括但不限于通过合成测序、通过连接测序,以及通过杂交测序。在特定实施例中,使用Illumina MiSeq或HiSeq2500系统对DNA片段文库进行测序。

[0059] DNA片段文库的测序将导致收集对应于具有所关注靶序列的区域的所选择DNA片段的单独序列。如本文所用,术语“读数”是指在测序之后获得的DNA片段的序列。在一些实施例中,测序从DNA序列文库产生约500,000、约100万、约150万、约200万、约250万、约300万、或约500万读数。在某些实施例中,读数是配对末端读数,其中DNA片段从分子的两端测序。

[0060] 在一个实施例中,该方法使用构建体特异性PCR引物和下一代(NextGen)测序技术来表征转基因事件。通过该方法生成的序列读数可用于,但不限于:识别插入位点、转基因完整性,以及转基因拷贝数。

[0061] 基因组DNA可使用本领域中已知的任何技术来分离和纯化。对于转基因植物分析,该方法的一个益处在于,可通过叶圆片、单叶、或叶部分、或其他允许植物继续正常生长和

发育的样品来获得足够的基因组DNA。然后可将分离的基因组DNA纯化并且对于质量进行分析,并且使用本领域中已知的任何方法来量化,诸如使用Fragment Analyzer™(爱荷华州埃姆斯市Advanced Analytical公司(Advanced Analytical, Ames, IA))。

[0062] 然后使用本领域中已知若干技术中的任一项将基因组DNA剪切成片段,使用本领域中已知的任何方案将其剪切成方便的片段大小,诸如使用Covaris E210™(马萨诸塞州沃本市Covaris公司(Covaris Inc, Woburn MA))。片段大小可在约50个碱基对至约2.5kb长度的范围内,包括但不限于75、100、150、200、250、300、350、400、450、500、550、600、650、700、750、800、850、900、950、或1kb的长度;或约1.25、1.5、1.75、2.0、或2.5或者更大kb的长度或这些范围内的任何长度。然后对剪切片段进行末端修复,加A尾,并通过本领域中已知的任何方案连接到专门的接头,诸如Kapa Biosystems™(马萨诸塞州沃本市(Woburn MA))所提供的方案。

[0063] 接头设计有九十六个独特的六碱基对区段,称为“条形码”、也称为“标签”、“多路复用标识符”、“索引”或“索引”序列。这些条形码用作唯一标识符并且有助于序列分析。同时,具有附接条形码的DNA片段形成片段文库,该片段文库可通过利用构建体特异性和接头特异性PCR引物的PCR扩增来富集。

[0064] 可通过沿着构建体和接头序列的长度设计构建体特异性PCR引物来调整和测定该方法的灵敏度和特异性。两种引物被设计用于交替链上的每200个碱基对,或单链上的400个碱基对。引物可根据所需分辨率嵌套或重叠。

[0065] 首轮PCR使用靶向构建体序列的第一引物(可为嵌套对),以及作为反向引物的接头特异性引物,从而锚固每个所得扩增子的一个末端。第二轮PCR将接头特异性引物与巢式PCR引物配对。巢式PCR引物可包括与将用于测序和分析的测序平台相关的序列标签。例如, Illumina P5™序列用于在Illumina™测序系统上使用。

[0066] 在PCR之后,可通过本领域中已知的任何方案(诸如Ampure Beads™(马萨诸塞州丹弗斯市贝克曼基因组学公司(Beckman Genomics, Danvers, MA))对片段文库进行纯化并且对于PCR伪影进行分析。文库以相等摩尔比合并并且稀释成优选浓度以用于测序。在另一个实施例中,文库可以以将实现适当测序数据的任何方式合并。

[0067] 在进一步的实施例中,该方法使用所关注转基因构建体的生物素酰化探针文库,其作为收集被分析并且简并成表示收集内的所有碱基的一组独特序列。DNA探针文库被设计为使得构建体池内的几乎所有碱基将在本文所述的富集步骤期间被靶向。与放置在玻璃载玻片或板微阵列上相反,探针文库被保持在溶液中。

[0068] 基因组DNA与生物样品分离并且如上所述剪切和连接到接头。然后通过最多至八轮PCR扩增连接的片段。这些扩增文库可针对质量和PCR伪影进行评估,然后根据操作者偏好以24、48或96的组或其他组合成相等摩尔比,并且稀释成优选ng/μl的工作贮备液。

[0069] 扩增文库用杂交缓冲液、显色试剂和对应于接头序列的阻断寡核苷酸来变性。在变性之后,池混合有生物素酰化探针并且在47℃下温育16小时。在杂交之后,溶液结合到链亲和素珠并且洗涤。

[0070] 对洗涤和洗脱池进行PCR扩增最多至五个循环,再次纯化和扩增。最终文库池被量化和稀释以用于测序。

[0071] 该实施例的一个益处在于,来自若干不同事件和生物体的样品可立刻合并和分

析。该实施例可捕捉未被QT-PCR或Southern检测到的异常。

[0072] 从针对具有所关注靶序列的区域的基因组DNA片段选择的DNA片段文库进行测序获得的测序读数可以以高通量方式进行处理以便表征样品植物中的转化事件。在一些实施例中,对从测序获得的读数进行后处理以移除任何接头序列。例如,任何接头序列(诸如NEXTFlex接头序列)的序列可被搜索并且从包含最少3bp的序列的任何读数的末端移除。还可使用K聚体分析对读数的集合进行处理以移除低质量序列以使所选择读数的特异性和灵敏度最大化。通过内部K聚体分析流水线推动读数的集合,该内部K聚体分析流水线使用水母(参见bioinformatics.oxfordjournals.org/content/27/6/764的网站)以用于K聚体计数。k聚体(默认值:31聚体)计数小于或等于2的读数从下游分析中移除。还可对读数进行后处理以选择前60%最丰富的读数以用于在本文所公开的方法中进一步分析。作为另外一种选择,前40%、45%、50%、55%、65%、70%、75%、80%、85%、90%、95%或100%的读数可被选择以用于进一步分析。在具体实施例中,前60%读数是适当的选择以使对于合理特异性的接点检测的灵敏度最大化。前60%可适于约50Kb长的构建体(T-DNA大小为约20Kb)的100-200万的总读数计数。随着靶向读数深度改变,该60%值可改变。为了选择最佳数目的读数以用于进一步分析,应该考虑读数的背景、特异性、灵敏度和质量。

[0073] 为了从进一步分析中排除任何内源读数,读数可与对照植物的基因组比对。如本文所用,“对照植物”的基因组是指不具有所关注靶序列的相同或种系发生上相似的基因型的植物的基因组。与对照植物的基因组对齐的任何读数都被认为是“内源读数”并且从本文所公开的方法中的进一步分析排除。也就是说,对应于靶序列的至少一部分的读数被选择以用于SbS流水线中的进一步分析。

[0074] 将序列进行比对以作比较的方法是本领域公知的。因此,可以使用数学算法完成任意两个序列的比对。此类数学算法的非限制性例子是Myers and Miller(1988) CABIOS 4:11-17 (Myers和Miller,1988年,《计算机在生物科学中的应用》,第4卷,第11-17页)的算法;Smith et al. (1981) Adv. Appl. Math. 2:482 (Smith等人,1981年,《应用数学进展》,第2卷,第482页)的局部比对算法;Needleman and Wunsch(1970) J. Mol. Biol. 48:443-453 (Needleman和Wunsch,《分子生物学杂志》,第48卷,第443-453页)的全局比对算法;Pearson and Lipman(1988) Proc. Natl. Acad. Sci. 85:2444-2448 (Pearson和Lipman,1988年,《美国国家科学院院刊》,第85卷,第2444-2448页)的搜索局部比对方法;Karlin and Altschul (1990) Proc. Natl. Acad. Sci. USA 87:2264 (Karlin和Altschul,1990年,《美国国家科学院院刊》,第87卷,第2264页)的算法,其在Karlin and Altschul (1993) Proc. Natl. Acad. Sci. USA 90:5873-5877 (Karlin和Altschul,1993年,《美国国家科学院院刊》,第90卷,第5873-5877页)中进行了修改。

[0075] 可利用这些数学算法的计算机实现进行序列的比较以确定最佳比对。此类实施包括但不限于:PC/Gene程序(可得自加利福尼亚州山景城的Intelligenetics公司(Intelligenetics,Mountain View,California))中的CLUSTAL;GCG威斯康星遗传学软件包(Wisconsin Genetics Software Package)版本10(可得自美国加利福尼亚州圣地亚哥斯克兰顿路9685号的Accelrys公司(Accelrys Inc.,9685Scranton Road,San Diego,California,USA))中的ALIGN程序(版本2.0)和GAP、BESTFIT、BLAST、FASTA和TFASTA。使用这些程序的比对可以使用默认参数进行。以下文献对CLUSTAL程序进行了详细描述:

Higgins et al. (1988) *Gene* 73:237-244 (Higgins等人, 1988年, 《基因》, 第73卷, 第237-244页); Higgins et al. (1989) *CABIOS* 5:151-153 (Higgins等人, 《计算机在生物学中的应用》, 第5卷, 第151-153页); Corpet et al. (1988) *Nucleic Acids Res.* 16:10881-90 (Corpet等人, 1988年, 《核酸研究》, 第16卷, 第10881-10890页); Huang et al. (1992) *CABIOS* 8:155-65 (Huang等人, 1992年, 《计算机在生物学中的应用》, 第8卷, 第155-165页); 以及 Pearson et al. (1994) *Meth. Mol. Biol.* 24:307-331 (Pearson等人, 1994年, 《分子生物学方法》, 第24卷, 第307-331页)。ALIGN程序基于Myers和Miller (1988年) (出处同上) 的算法。当比较氨基酸序列时, ALIGN程序可使用PAM120加权残基表、空位长度罚分12和空位罚分4。Altschul et al (1990) *J. Mol. Biol.* 215:403 (Altschul等人, 1990年, 《分子生物学杂志》, 第215卷, 第403页) 的BLAST程序基于上文的Karlin和Altschul (1990年) 算法。为了出于比较目的获得带空位的比对结果, 可以如Altschul et al. (1997) *Nucleic Acids Res.* 25:3389 (Altschul等人, 1997年, 《核酸研究》, 第25卷, 第3389页) 中所描述采用空位BLAST (在BLAST 2.0中)。作为另外一种选择, PSI-BLAST (在BLAST 2.0中) 可以用来执行检测分子之间远源关系的迭代搜索。参见Altschul et al. (1997) (Altschul等人, 1997年), 出处同上。当采用BLAST、空位BLAST、PSI-BLAST时, 可以使用各个程序的默认参数 (例如, BLASTN用于核苷酸序列)。参见www.ncbi.nlm.nih.gov的网站。Bowtie2 (位于网站nature.com/nmeth/journal/v9/n4/full/nmeth.1923.html) 和BWA (位于网站ncbi.nlm.nih.gov/pubmed/19451168) 还可用于将由NGS产生的数百万短读数与基因组有效地比对。在另一个实施例中, 还可使用GSNAP (Thomas D. Wu, Serban Nacu “Fast and SNP-tolerant detection of complex variants and splicing in short reads. *Bioinformatics*. 2010 Apr 1; 26 (7): 873-81. Epub 2010 Feb 10 (Thomas D. Wu, Serban Nacu, “短读数中的复杂变体和剪接的快速和SNP耐受性检测”, 《生物信息学》, 2010年4月1日, 第26卷, 第7期, 第873-881页, 2010年2月10日电子出版))。

[0076] 用于比对的算法和参数可根据所选择植物的类型、被表征的靶序列的类型、以及用于将靶序列引入到样品植物中的转化方法来调整。

[0077] 然后将具有与所关注靶序列的一些对齐的读数彼此比对以识别连接序列。为了使读数具有与所关注靶序列的一些对齐, 读数必须具有与靶序列共同的约2、约4、约6、约8、约10、约12、约15、约20、或约30个核苷酸, 或者在约5、约10、约15、约20、约25、约30、约35、约40、约45、约50、约75、或约100个连续核苷酸长度的靶序列的片段上共享与所关注靶序列的至少约50%、约60%、约70%、约75%、约80%、约85%、约90%、约95%、或约100%同源性。如本文所用, “接点”是指其中DNA的两个不同片段接合在一起的点。例如, 接点可存在于插入DNA (例如, 质粒、靶序列等) 与样品植物的基因组DNA之间, 或存在于两个插入DNA序列之间。接点也存在于其中两个DNA片段以改自于天然植物中存在的方式接合在一起的转化植物中。“连接序列”是指DNA的包含接点的区段。连接序列可为约5、约7、约10、约12、约15、约17、约20、约25、约26、约27、约28、约29、约30、约31、约32、约33、约34、约35、约40、约50、约60、约75、约100、或约5-10、10-15、10-20、10-30、10-40、20-40、或25-35个核苷酸的长度。

[0078] 连接序列可通过比对读数和识别其中读数不再与靶序列对齐的接点来识别。由于比对算法的性质, 接点有时预测为彼此接近。当接点被预测为彼此距离在约5、约4、约3、约2、或约1个核苷酸内时, 接点可被合成。如本文所用, 术语“合成的”或“合成”是指将相似的

接点预测结合到单个共有连接序列中的方法。为了合成连接序列,比对围绕预测连接序列的每个读数的片段。在比对之后,具有更多独特支撑读数的接点被识别为接点。在一些实施例中,围绕预测连接序列的读数的片段由基因组DNA的三十(30)个核苷酸和靶序列的二十(20)个核苷酸组成。这还可为基因组DNA序列的约10、15、20、25、30、35、40、45、50、60、70、80、90至100个核苷酸以及靶序列的约10、15、20、25、30、35、40、45、50、60、70、80、90至100个核苷酸。如本文所用,包含基因组DNA的三十(30)个核苷酸和靶序列的二十(20)个核苷酸的DNA片段被称为“30_20聚体”。合成可用于移除具有测序误差的接点读数。

[0079] 在一些实施例中,连接序列被识别并且30_20聚体连接序列延伸以便促进基因组作图。连接序列可通过使用序列组装工具来延伸,所述序列组装工具诸如SSAKE (Warren R., et al. (2007) *Bioinformatics* 23(47):500-501 (Warren R等人,2007年,《生物信息学》,第23卷,第47期,第500-501页)),其以引用的方式并入本文。在某些实施例中,来自基因组文库测序的读数数据、或用于对照植物的现有基因组序列数据可被组装到连接序列的末端以便获得延伸的连接序列。如本文所用,“重叠群”是指延伸的连接序列。

[0080] 重叠群可被定位到对照植物的基因组并且定位到所关注靶序列以便识别到样品植物的基因组中的靶序列插入的数目、位置和完整性。靶序列的数目可通过识别靶序列与植物基因组之间的接点的连接序列的数目,连同两个靶序列之间的连接序列的数目来确定。例如,如果接点在靶序列与植物基因组之间的靶序列的每个末端上识别,则预期在植物基因组中仅存在一个拷贝的靶序列。然而,如果在两个插入序列之间检测到另外的连接序列,则预期存在多个靶序列。可通过分析与靶序列的读数比对来评价靶序列插入的完整性以识别所关注靶序列的插入、缺失、或重排。例如,靶序列内的SNP可被识别,并且可检测到农杆菌骨架污染。重叠群可被定位到对照植物的基因组以便识别靶序列插入的位置。

[0081] 用于在本文所公开的方法中使用的植物包括整株植物、植物器官、植物组织、种子和植物细胞及其子代。植物细胞包括但不限于来自种子、悬浮培养物、胚、分生组织区、愈伤组织、叶、根、苗、配子体、孢子体、花粉和小孢子的细胞。

[0082] 本文所用的“转基因植物”包括指涉在其基因组内包含异源多核苷酸的植物。一般来讲,异源多核苷酸稳定地整合在基因组内使得该多核苷酸得以传递到连续世代。异源多核苷酸可单独整合进基因组中,或者作为重组表达盒的一部分整合进基因组中。“转基因”在本文中用于包括其基因型已因异源核酸的存在而改变的任何细胞、细胞系、愈伤组织、组织、植物部分或植物,包括那些最初经如此改变的转基因植物以及通过有性杂交或无性繁殖从最初的转基因植物产生的那些。本文所用的术语“转基因”不涵盖通过常规植物育种方法或通过诸如随机异花受精、非重组病毒感染、非重组细菌转化、非重组转座或自发突变之类的自然发生事件导致的基因组(染色体基因组或染色体外基因组)的改变。

[0083] 本文所公开的方法可用于表征任何植物物种(包括但不限于单子叶植物和双子叶植物)中的所关注靶序列。所关注植物物种的例子包括但不限于玉米(*Zea mays*)、芸苔属物种(*Brassica* sp.) (例如甘蓝型油菜(*B. napus*)、芜菁(*B. rapa*)、芥菜(*B. juncea*)),特别是可用作种子油来源的那些芸苔属物种、苜蓿(*Medicago sativa*)、水稻(*Oryza sativa*)、黑麦(*Secale cereale*)、高粱(*Sorghum bicolor*、*Sorghum vulgare*)、粟(例如珍珠粟(*Pennisetum glaucum*)、黄米(*Panicum miliaceum*)、谷子(*Setaria italica*)、龙爪稷(*Eleusine coracana*))、向日葵(*Helianthus annuus*)、红花(*Carthamus tinctorius*)、小

麦 (*Triticum aestivum*)、大豆 (*Glycine max*)、烟草 (*Nicotiana tabacum*)、马铃薯 (*Solanum tuberosum*)、落花生 (*Arachis hypogaea*)、棉花 (海岛棉 (*Gossypium barbadense*)、陆地棉 (*Gossypium hirsutum*))、甘薯 (*Ipomoea batatas*)、木薯 (*Manihot esculenta*)、咖啡 (*Coffea* spp.)、椰子 (*Cocos nucifera*)、菠萝 (*Ananas comosus*)、柑橘 (*Citrus* spp.)、可可 (*Theobroma cacao*)、茶 (*Camellia sinensis*)、香蕉 (*Musa* spp.)、鳄梨 (*Persea americana*)、无花果 (*Ficus casica*)、番石榴 (*Psidium guajava*)、芒果 (*Mangifera indica*)、橄榄 (*Olea europaea*)、木瓜 (*Carica papaya*)、腰果 (*Anacardium occidentale*)、澳洲坚果 (*Macadamia integrifolia*)、杏树 (*Prunus amygdalus*)、糖用甜菜 (*Beta vulgaris*)、甘蔗 (*Saccharum* spp.)、燕麦、大麦、蔬菜类、观赏植物类和针叶树类。

[0084] 蔬菜包括番茄 (*Lycopersicon esculentum*)、莴苣 (例如 *Lactuca sativa*)、青豆 (*Phaseolus vulgaris*)、利马豆 (*Phaseolus limensis*)、豌豆 (*Lathyrus* spp.) 和黄瓜属 (*Cucumis*) 的成员诸如黄瓜 (*C. sativus*)、香瓜 (*C. cantalupensis*) 和甜瓜 (*C. melo*)。观赏植物包括杜鹃 (*Rhododendron* spp.)、八仙花 (*Macrophylla hydrangea*)、朱槿 (*Hibiscus rosasanensis*)、玫瑰 (*Rosa* spp.)、郁金香 (*Tulipa* spp.)、水仙 (*Narcissus* spp.)、矮牵牛 (*Petunia hybrida*)、康乃馨 (*Dianthus caryophyllus*)、一品红 (*Euphorbia pulcherrima*) 和菊花。

[0085] 可用于实施本发明的针叶树类包括例如松树类,如厚皮刺果松 (*Pinus taeda*)、沼泽松 (*Pinus elliotii*)、美国黄松 (*Pinus ponderosa*)、黑松 (*Pinus contorta*) 和辐射松 (*Pinus radiata*);黄-杉 (*Pseudotsuga menziesii*);西铁杉 (*Tsuga canadensis*);北美云杉 (*Picea glauca*);红杉 (*Sequoia sempervirens*);枞树 (true firs) 如银枞 (*Abies amabilis*) 和胶枞 (*Abies balsamea*);和雪松如西部红雪松 (*Thuja plicata*) 和阿拉斯加黄-雪松 (*Chamaecyparis nootkatensis*)。在具体的实施例中,本发明的植物是作物植物 (例如玉米、苜蓿、向日葵、芸苔、大豆、棉花、红花、花生、高粱、小麦、稷、烟草等)。在其他实施例中,玉米和大豆以及甘蔗植物是优选的,而在另外的实施例中玉米植物是优选的。

[0086] 其他所关注植物包括提供所关注种子的谷物植物、油料种子植物和豆科植物。所关注种子包括谷物种子,诸如玉米、小麦、大麦、水稻、高粱、黑麦等。油料种子植物包括棉花、大豆、红花、向日葵、芸苔、玉蜀黍、苜蓿、棕榈、椰子等。豆科植物包括豆类和豌豆。豆类包括瓜尔豆、槐豆、葫芦巴、大豆、四季豆、豇豆、绿豆、利马豆、蚕豆、小扁豆、鹰嘴豆等等。

[0087] 在一些实施例中,样品植物被选择以用于根据样品植物的基因组中的所关注靶序列的数目、位置、完整性、或它们的任意组合来开发和育种。样品植物可被选择以用于在育种程序中使用,所述育种程序诸如谱系育种、轮回选择、混合选择、或突变育种。

[0088] 本发明的实施例在以下实例中进一步限定。应当理解这些实例仅以示例的方式给出。通过以上讨论和这些实例,本领域技术人员可以确定本发明的基本特征,并在不背离本发明的精神和范围的情况下,可对本发明的实施例作出各种改变和修饰以使之适应各种用法和条件。因此,除了本文显示和描述的那些实施例之外,本发明实施例的各种修改形式将因前面的描述而对本领域技术人员是显而易见的。此类修改形式也旨在落入所附权利要求的范围内。

[0089] 本文列出的每个参考文献的公开内容以引用方式全文并入本文。

[0090] 实验

[0091] 实例1:平铺方法

[0092] 使用来自转基因构建体的序列信息,设计了构建体特异性连接介导的巢式PCR (LMN-平铺引物。通过巢式PCR引物设计来确定测定法灵敏度和特异性,其中两种引物被设计用于交替链上的每200个碱基对,或单链上的400个碱基对。

[0093] 在引物设计之后,使用EZNA Plate 96TM试剂盒(佐治亚州诺克斯市的Omega Biotek公司(Omega Biotek,Norcross,GA))从冻干叶圆片提取DNA。利用Fragment AnalyzerTM(爱荷华州埃姆斯市Advanced Analytical公司(Advanced Analytical,Ames,IA))评估纯化基因组DNA的质量和数量,并且随后利用Covaris E210TM(马萨诸塞州沃本市Covaris公司(Covaris Inc,Woburn MA))将其剪切成平均片段大小为1500个碱基对。对剪切的DNA进行修复,加A尾,并且根据Kapa BiosystemsTM(马萨诸塞州沃本市(Woburn MA))所提供的方案连接。连接的接头定制设计有九十六个独特的六碱基对条形码并且连接到Illumina P7TM序列以使Illumina能够在PCR之后测序。

[0094] 在连接之后,针对转基因序列通过两轮二十循环扩增富集片段文库。初次PCR使用巢式对的第一引物作为正向引物,以及接头特异性引物作为反向引物,从而锚固每个扩增子的一个末端。第二PCR将接头特异性引物与巢式PCR引物配对,其包括Illumina P5TM序列,从而整理用于IlluminaTM序列的片段。在用AmpureXPTM珠(马萨诸塞州丹弗斯市贝克曼基因组学公司(Beckman Genomics,Danvers,MA))纯化之后,在Fragment AnalyzerTM上分析片段文库,将其以相等摩尔比合并到九十六个样品池中并且稀释至2nM。在加利福尼亚州圣地亚哥亿明达公司(Illumina(San Diego,CA))的MiSeq或HiSeq 2500TM系统上对池进行测序,从而按照制造商方案产生每个样品100万至200万个100碱基对配对末端。

[0095] 所产生的序列用于识别插入位点、转基因完整性和转基因拷贝数。

[0096] 实例2:基于测序的Southern方法

[0097] 基于测序的Southern(SbS)应用采用基于序列捕获的方法来针对包含片段的构建体富集IlluminaTM测序文库。该方法中的第一步骤用于设计生物素酰化探针文库,该文库在先锋设计组批准之后通过Roche NimbleGenTM(威斯康星州麦迪逊(Madison,WI))合成。所关注转基因构建体作为收集被分析并且简并成表示收集内的所有碱基的一组独特序列。DNA探针文库被设计为使得构建体池内的几乎所有碱基在富集过程期间被靶向。

[0098] 在探针文库设计之后,通过标准分子操作产生下一代DNAShotgun文库以用于单个事件。简而言之,DNA通过佐治亚州诺克斯市的Omega Biotek公司(Omega Biotek,Norcross,GA)的EZNA Plant 96TM试剂盒从叶圆片分离。利用Fragment AnalyzerTM(爱荷华州埃姆斯市Advanced Analytical公司(Advanced Analytical,Ames,IA))评估纯化基因组DNA的质量和数量,并且随后利用Covaris E210TM(马萨诸塞州沃本市Covaris公司(Covaris Inc,Woburn MA))通过超声处理将其剪切成平均片段大小为400bp。对剪切的DNA进行修复,加A尾,并且根据Kapa BiosystemsTM(马萨诸塞州沃本市(Woburn MA))所提供的方案连接。

[0099] 包括九十六个独特六碱基对条形码的连接的德克萨斯州奥斯汀BI00 Scientific公司(BI00 Scientific(Austin,TX))的NEXTflexTM接头序列旁侧是IlluminaTM特定序列,以使样品能够在杂交和测序阶段合并。

[0100] 这些分子条形码(也称为标签、索引或多路复用标识符)是短的DNA序列,其出现在每个测序读数的末端(5'或3'),并且用于将读数连接到其文库源。要支持样品的有效合并,

需通过以下所述将索引条形码结合到Illumina文库构建方法:将其添加到Illumina的I5TM接头,并且在Illumina的I7TM接头中使用标准Illumina条形码。与Illumina的I7接头条形码(其中目前为24)相比,这提供了与独特条形码标识符一起在每个样品上运行2,304个样品的方式。

[0101] 连接片段文库根据NimbleGenTM捕获方案扩增八个循环。利用Advanced Analytical Fragment AnalyzeTM再次评估扩增文库的质量和数量,将其以相等摩尔比在24、48或96的组中合并并且稀释成5ng/ul的工作贮备液。

[0102] 根据NimbleGenTM方案,使用双捕获方法来实现序列富集以增加靶读数。利用杂交缓冲液、SeqCap EZ Developer ReagentTM、和对应于池中的接头序列的阻断寡核苷酸使上述DNA shotgun文库在混合物中变性。在变性之后,将混合物与生物素酰化寡核苷酸文库混合并且在四十七摄氏度下温育十六小时。在杂交之后,将混合物与链亲和素Dyanbeads M-270TM(纽约州格兰德岛生命技术公司(LifeTech,Grand Island,NY))混合。使用DynaMag-2TM(纽约州格兰德岛生命技术公司(LifeTech,Grand Island,NY)),结合的DNA片段根据NimbleGenTM捕获方案来洗涤。将洗涤和洗脱的文库池扩增五个循环,利用马里兰州日耳曼敦的凯杰公司(Qiagen(Germantown,MD))的QiaquickTM柱根据制造商说明书对其进行纯化,然后捕获,扩增十六个循环,并且使用上述方法第二次纯化。

[0103] 利用安捷伦(Agilent)带站对最终捕获文库池进行量化并且稀释成2nM以用于测序。在IlluminaTM(加利福尼亚州圣地亚哥(San Diego,CA))MiSeqTM或HiSeq 2500 systemTM上对池进行测序,从而产生每个样品100万至200万个100碱基对配对末端读数。

[0104] 所产生的序列用于识别插入位点、转基因完整性和转基因拷贝数。

[0105] 实例3:基于测序的Southern生物信息学流水线

[0106] SbS通过检测转化质粒与基因组DNA或不连续质粒DNA之间的嵌合连接序列来识别质粒插入的整合位点、拷贝数、完整性、骨架存在和重排。对从转基因植物产生的短序列进行后处理以移除低质量序列和测序误差并且修剪掉任何接头序列。然后将来自前60%丰富清洁序列的代表性序列与植物基因组比对。识别内源读数并且将其从下游接点识别步骤排除。

[0107] 然后通过比对剩余非内源性读数与质粒参考来识别质粒与植物基因组区段之间或不连续质粒区段之间的接点。然后根据包含与质粒比对的20个碱基和30个不连续碱基的连接序列的50个碱基来对接点进行合成。分裂和合成特征检查被识别为在相同区域中的连接序列。如果连接序列是相同的,但存在单个核苷酸差异(SNP、插入、缺失),则算法报道具有最多支持的独特30_20聚体,即对具有低支撑读数计数的接点进行过滤。(参见图7a和图7b)。还对存在于来自相同或接近基因型的非转基因对照物中的接点进行过滤。

[0108] 对于剩余接点中的每个,工具SSAKE(通过K聚体搜索和3' 读数扩展的短序列组装)用于使用所有清洁读数将嵌合接点读数延伸到较长重叠群中。每个接点可具有所产生的多个SSAKE重叠群。每个SSAKE重叠群基于30_20聚体在接点位置处分裂成一个近侧序列和一个远侧序列,其中近侧序列是指含有30_20聚体的20聚体的亚序列,而远侧序列是指含有30_20聚体的30聚体的亚序列。然后将每个接点的最长远侧序列定位到植物基因组以识别整合位点以及定位到质粒以识别任何重排。最长近侧序列被定位到质粒以进一步确认质粒的接点位置。然后根据基于分析结果的一组标准作出推进决策,所述分析结果为诸如拷贝

数、完整性、骨干不存在/存在等。

[0109] SbS流水线很好地作用于富集质粒的序列和通过序列捕获方法产生的侧翼序列。其还可应用于转基因植物的整个基因组shotgun测序。

[0110] SbS是高通量流水线,其被开发以使对较差转化事件的推进最小化,所述较差转化事件将在下游产品开发阶段中浪费数百万美元。SbS的分辨率足够高以用于监管要求。

[0111] 本说明书中提到的所有公布和专利申请指示了本发明所属领域的技术人员的水平。所有公布和专利申请以引用方式并入本文,如同每个单独的公布或专利申请被具体地和独立地指出以引用方式并入本文一样。

[0112] 虽然为了理解清晰目的已经通过举例说明和实例方式较详细地描述了本发明,但显然可以在所附权利要求书范围内实施一些改变和修改。

序列表

	<110> 先锋良种国际公司	
	<120> 用于在基因组中表征DNA序列组成的方法	
	<130> 5917-W0-PCT	
	<150> 61/812, 876	
	<151> 2013-04-17	
	<150> 61/813, 001	
	<151> 2013-04-17	
	<160> 8	
	<170> FastSEQ for Windows 4.0版	
	<210> 1	
	<211> 50	
	<212> DNA	
	<213> 人工序列	
	<220>	
	<223> 推定连接序列	
	<400> 1	
	aagttgtaat ttaaattgaa tacagaaggg ctccggcccc ggcgccacca	50
	<210> 2	
	<211> 50	
	<212> DNA	
	<213> 人工序列	
	<220>	
	<223> 推定连接序列	
[0001]	<400> 2	
	aagttgtaat ttaaattgaa tagagaaggg ctccggcccc ggcgccacca	50
	<210> 3	
	<211> 50	
	<212> DNA	
	<213> 人工序列	
	<220>	
	<223> 推定连接序列	
	<400> 3	
	aagttgtaat ttaaattgaa tacagaagga ctccggcccc ggcgccacca	50
	<210> 4	
	<211> 50	
	<212> DNA	
	<213> 人工序列	
	<220>	
	<223> 推定连接序列	
	<400> 4	
	aagttgtaat ttaaattgaa tacagagggg ctccggcccc ggcgccacca	50
	<210> 5	
	<211> 50	
	<212> DNA	
	<213> 人工序列	
	<220>	
	<223> 推定连接序列	
	<400> 5	
	aagttgtaat ttaaattgaa tacagaaggg ctccggcccc ggcaccacca	50
	<210> 6	
	<211> 50	

	<212> DNA	
	<213> 人工序列	
	<220>	
	<223> 推定连接序列	
	<400> 6	
	aagttgtaat ttaagttgaa tacagaaggg ctccggcccc ggcgccacca	50
	<210> 7	
	<211> 50	
	<212> DNA	
	<213> 人工序列	
[0002]	<220>	
	<223> 推定连接序列	
	<400> 7	
	aagttgtaat ttaaattgaa tacagaaggg ctccggcccc ggcgccacca	50
	<210> 8	
	<211> 50	
	<212> DNA	
	<213> 人工序列	
	<220>	
	<223> 推定连接序列	
	<400> 8	
	agttgtaatt taaattgaat acagaagggc tccggccccg gcgccaccaa	50



图1

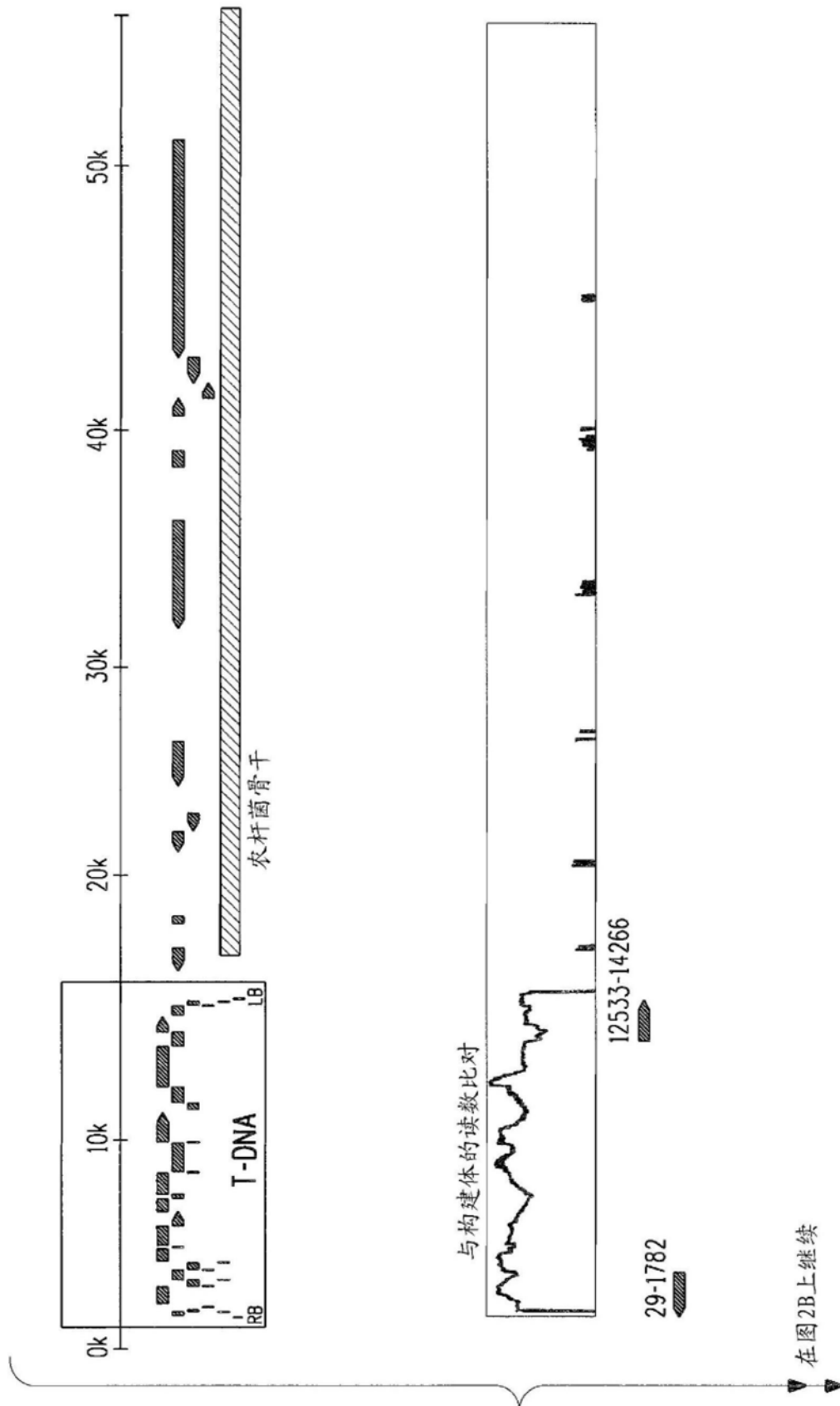


图2A

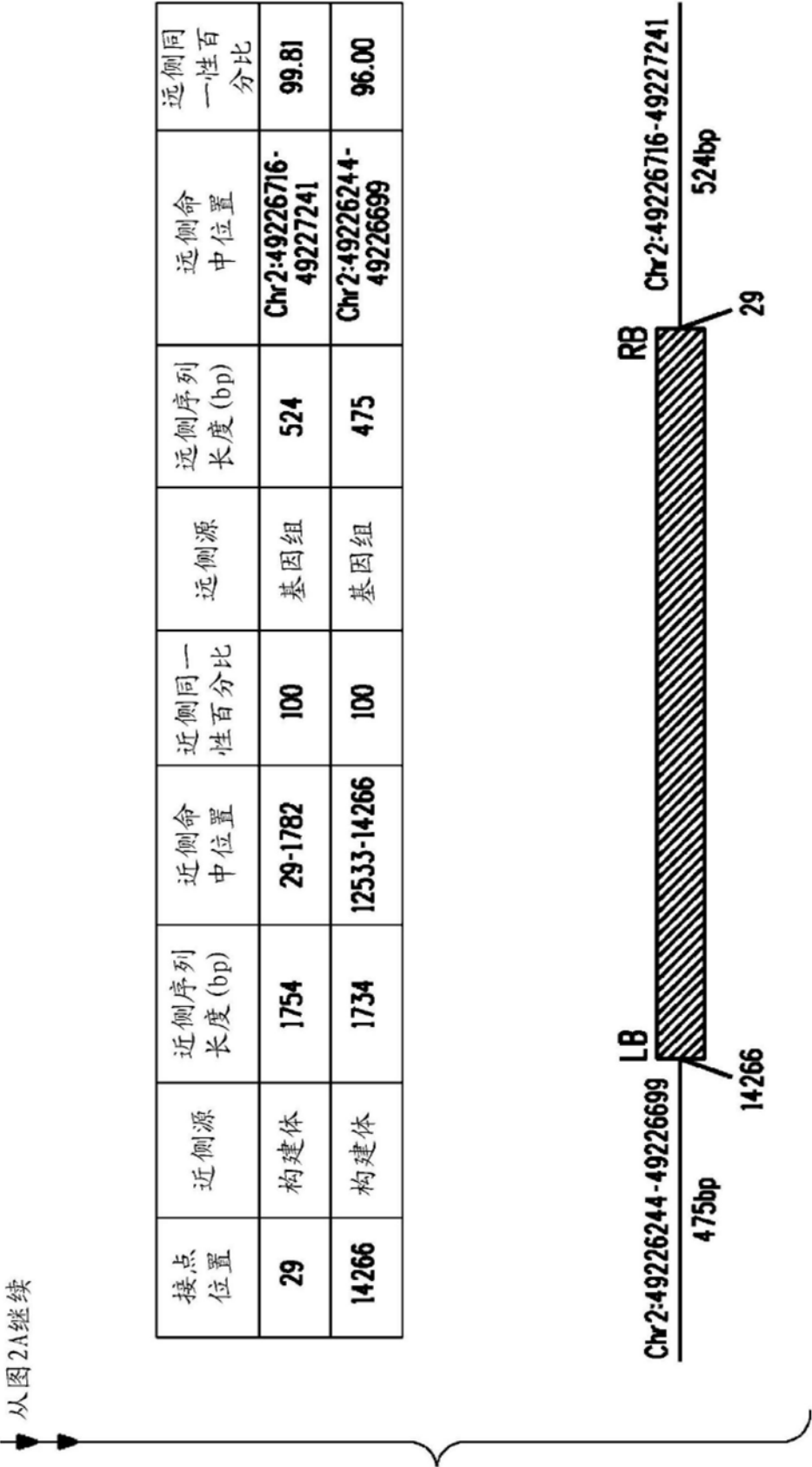


图2B

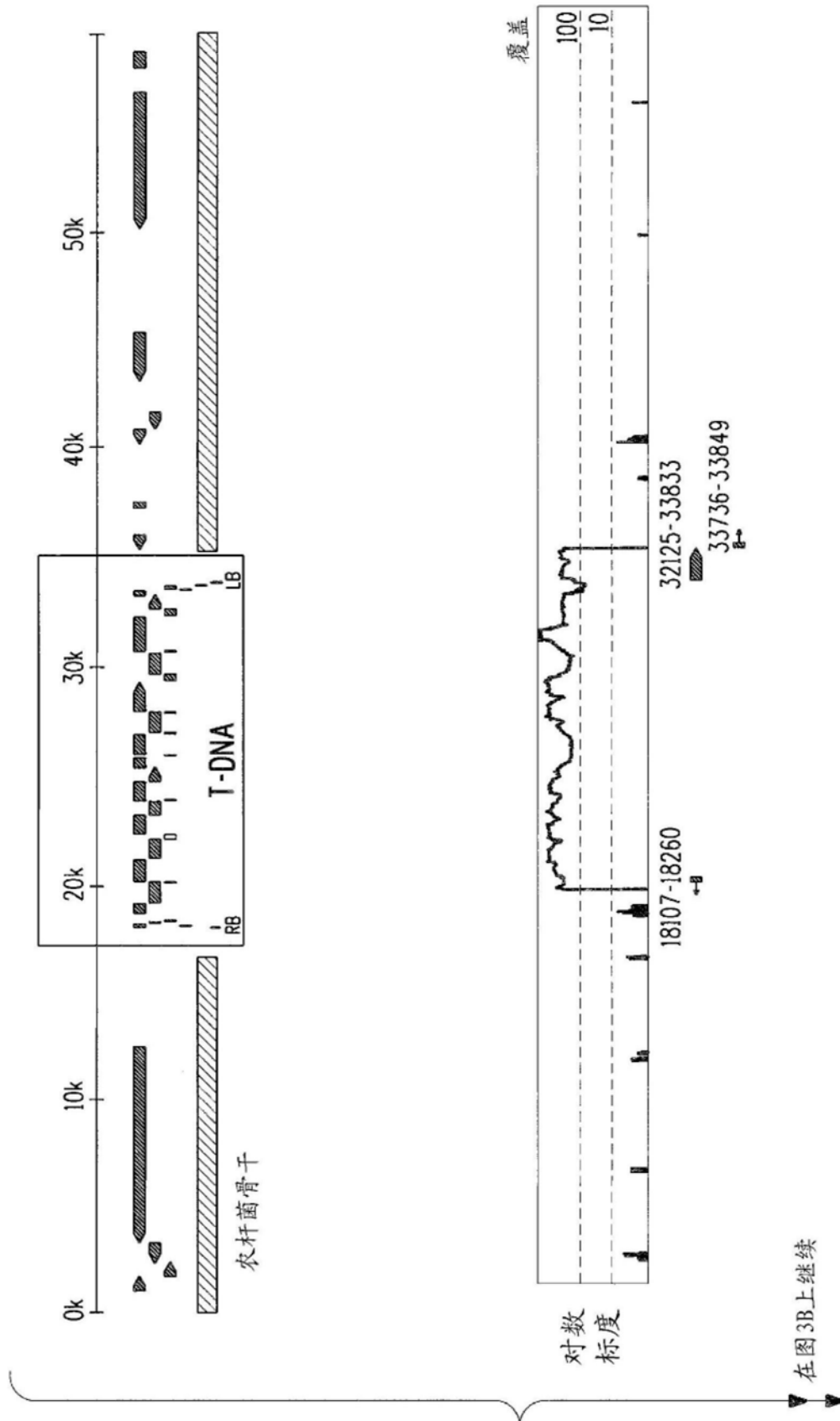


图3A

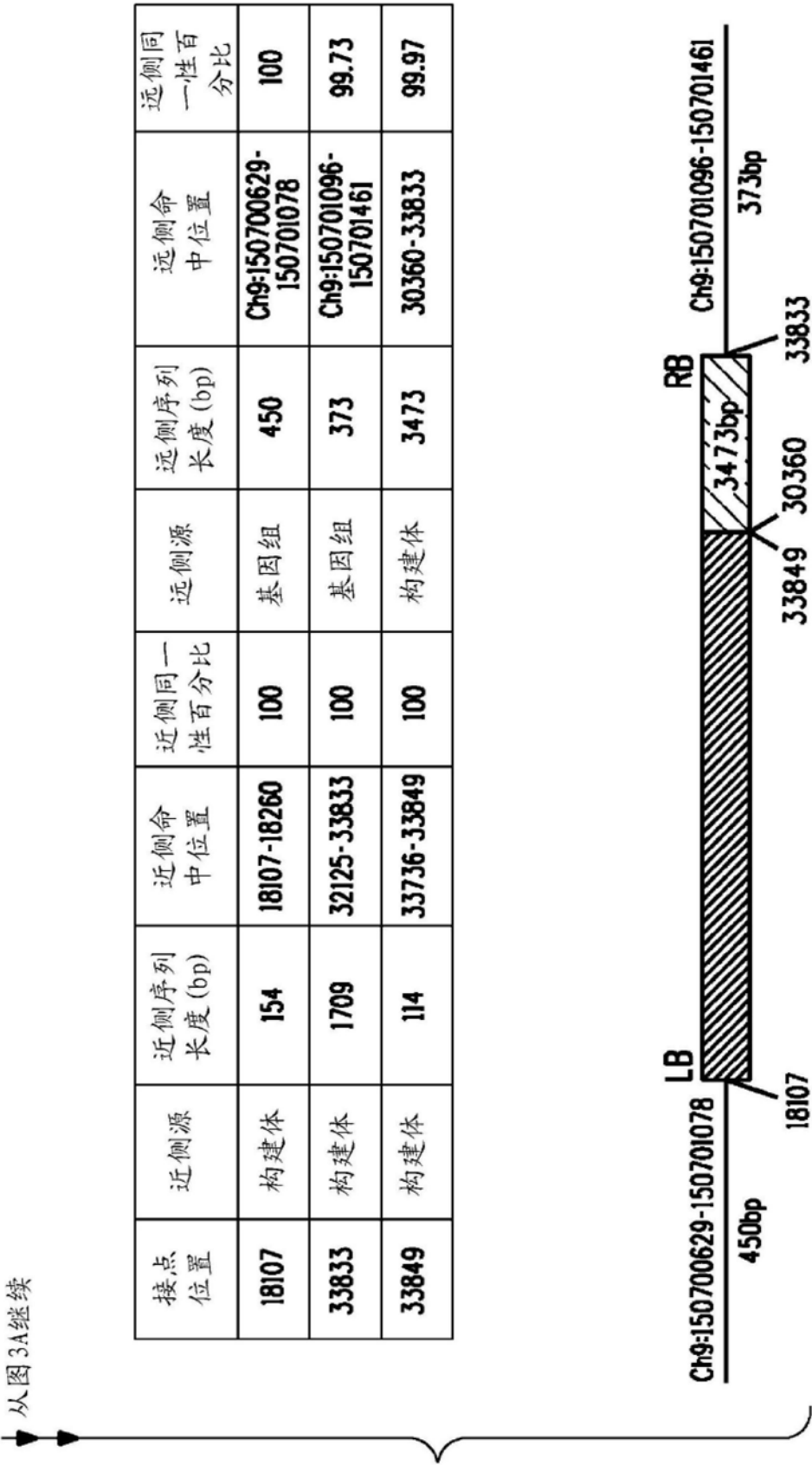


图3B

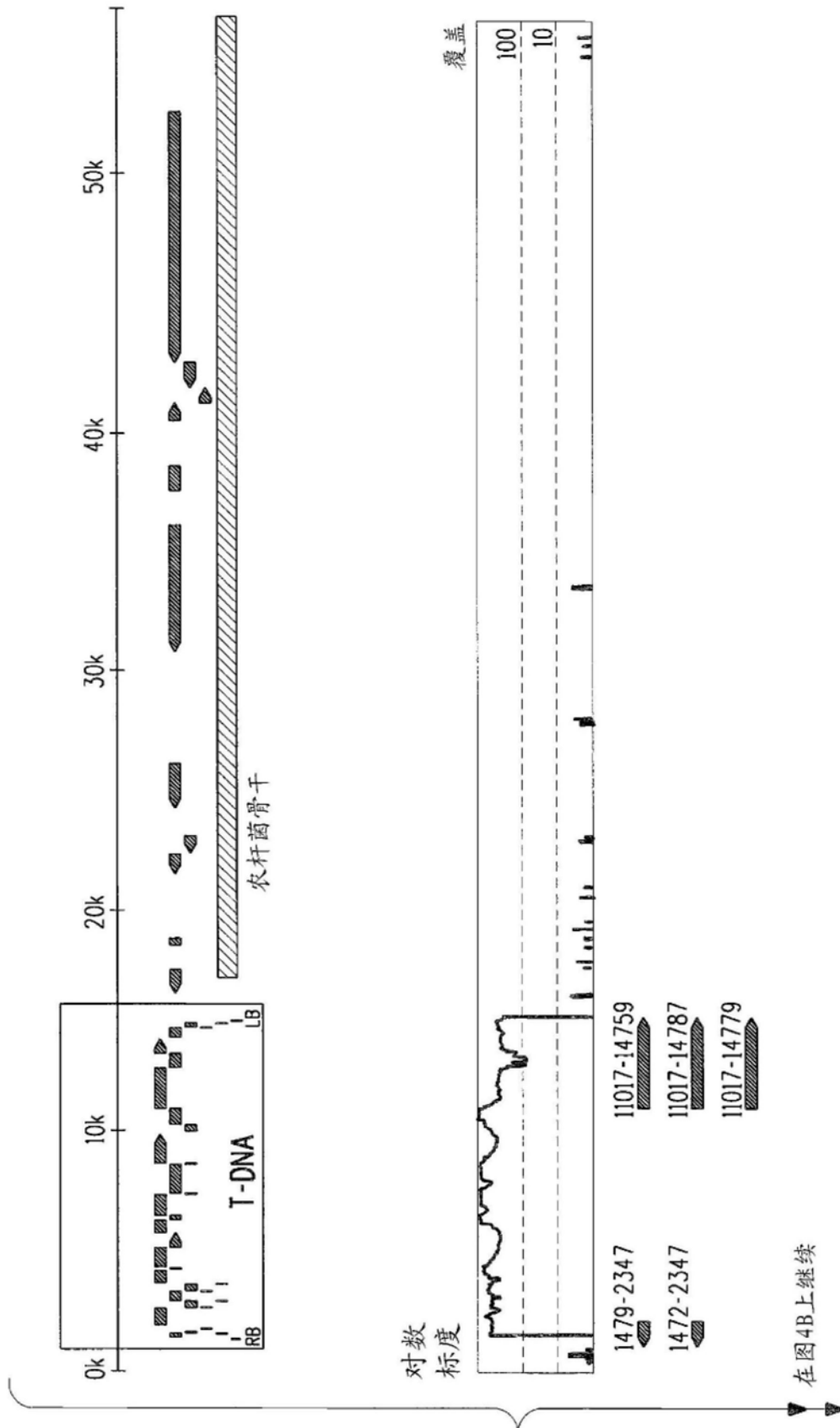


图4A

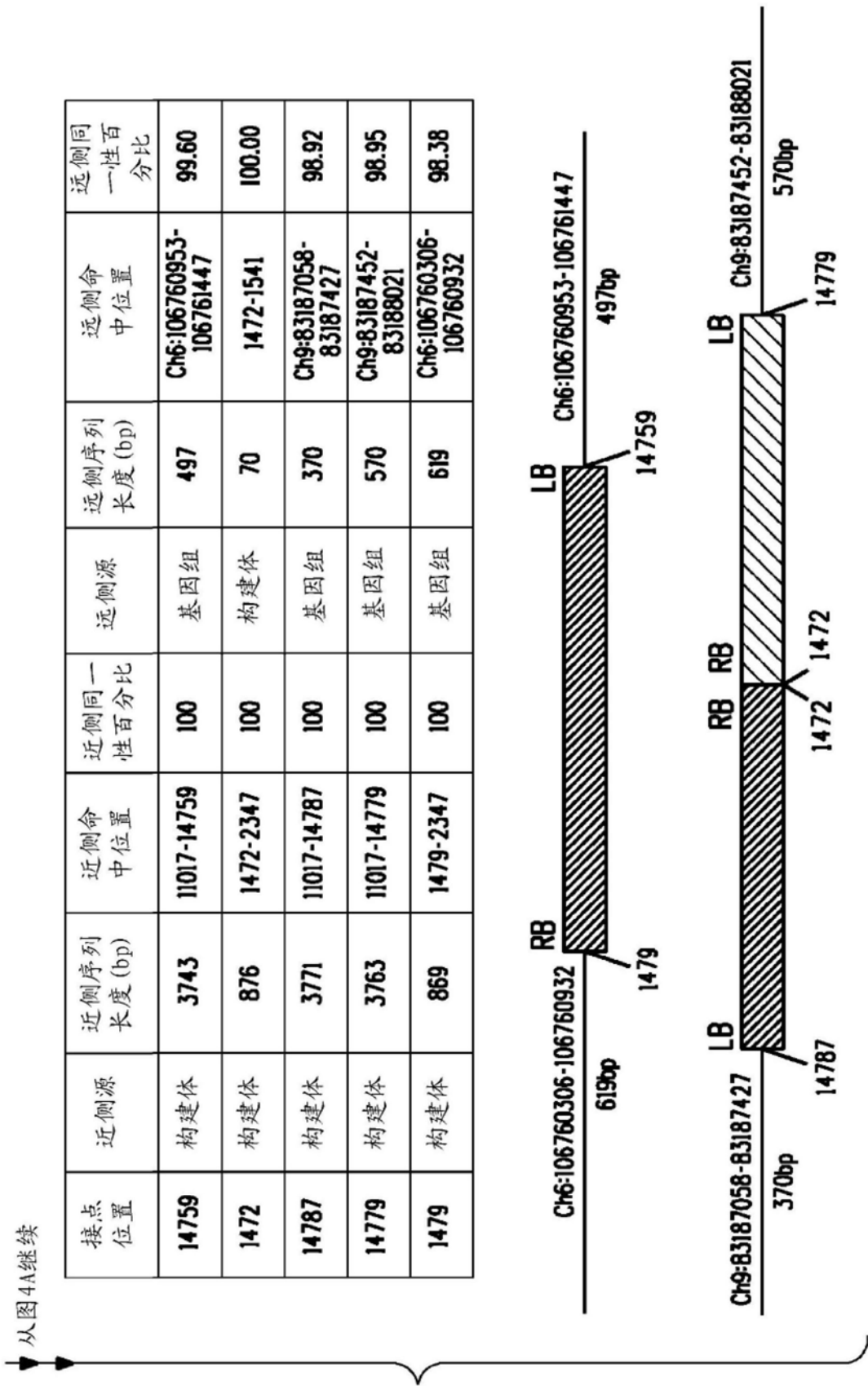


图4B

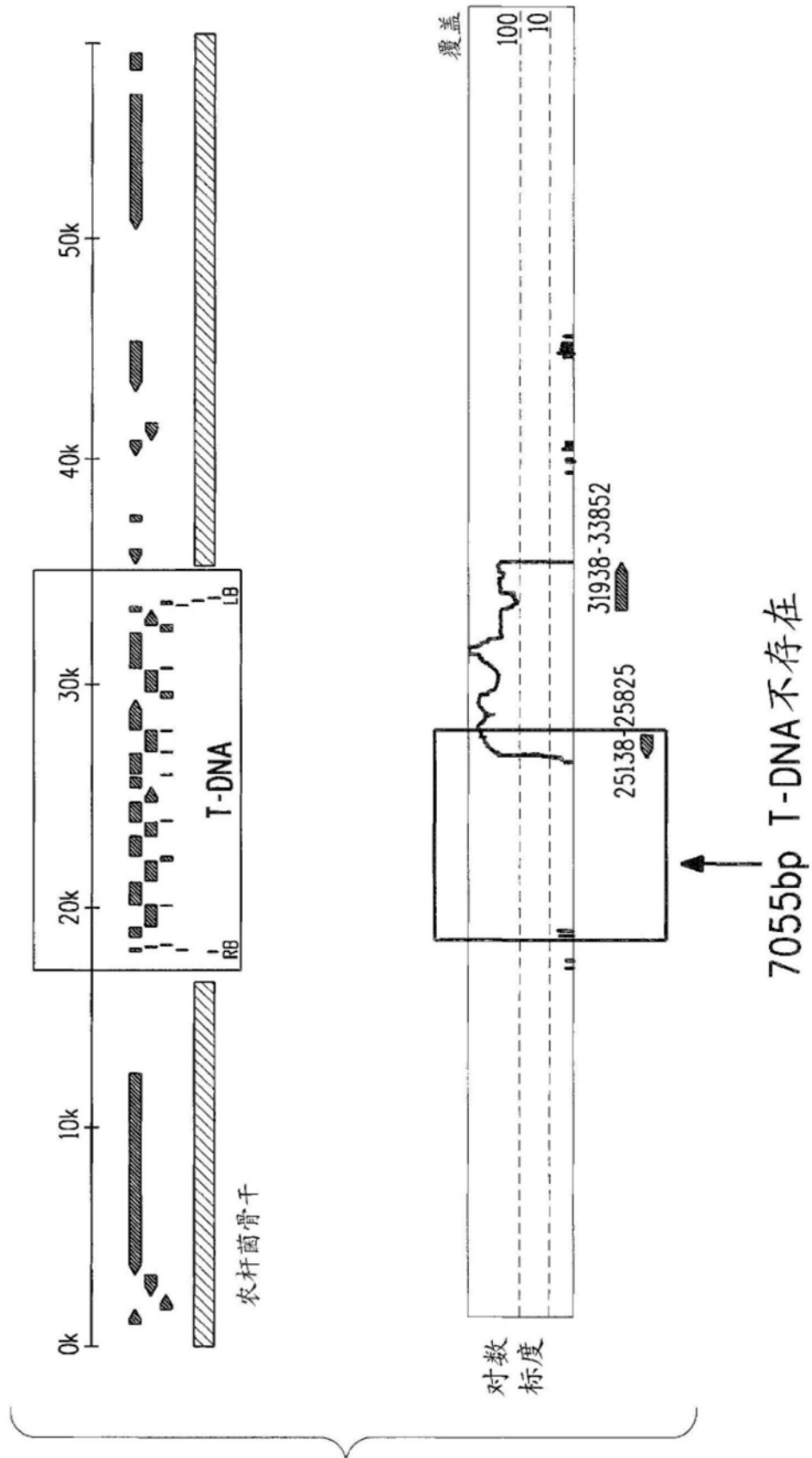


图5

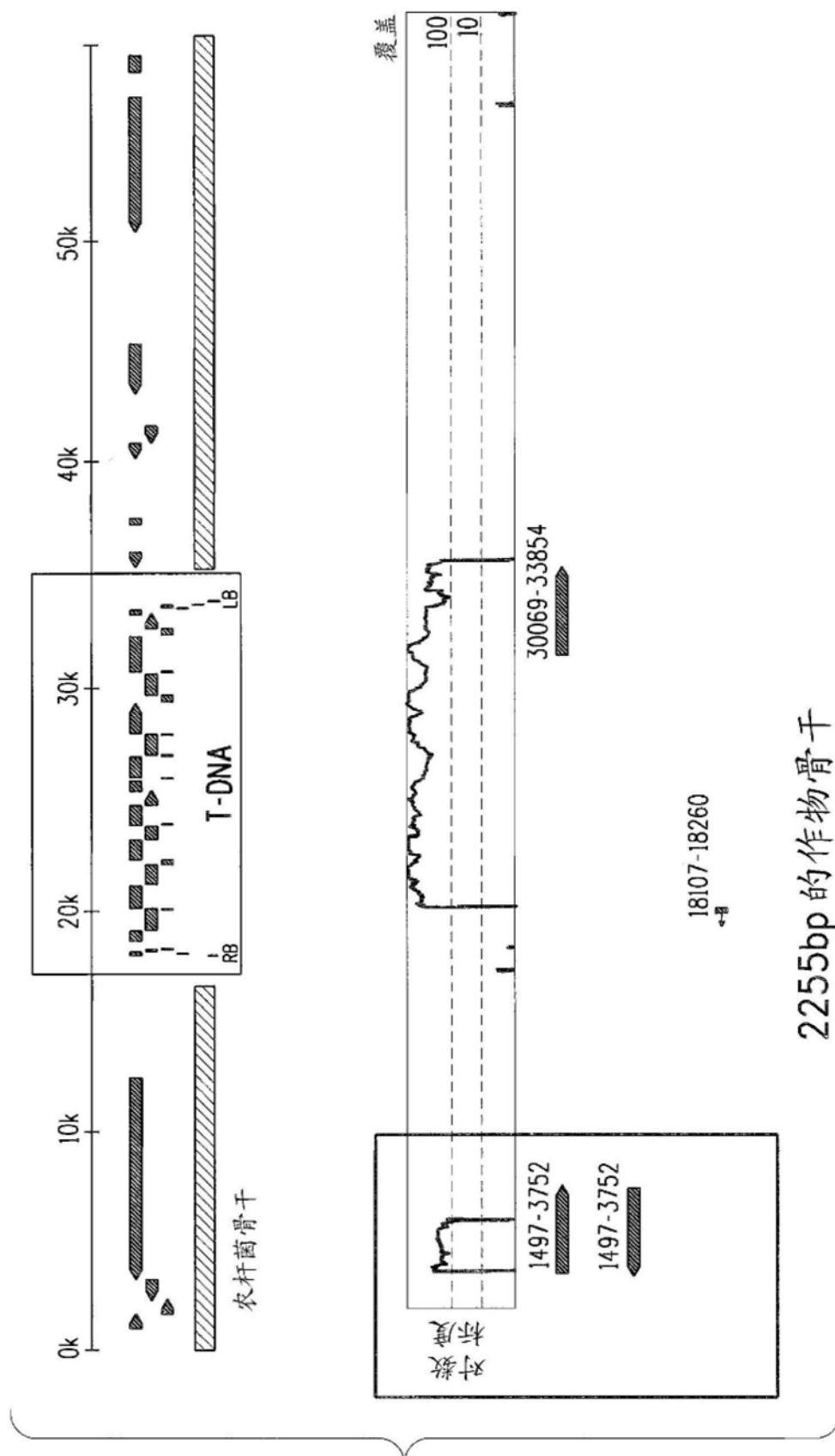


图6

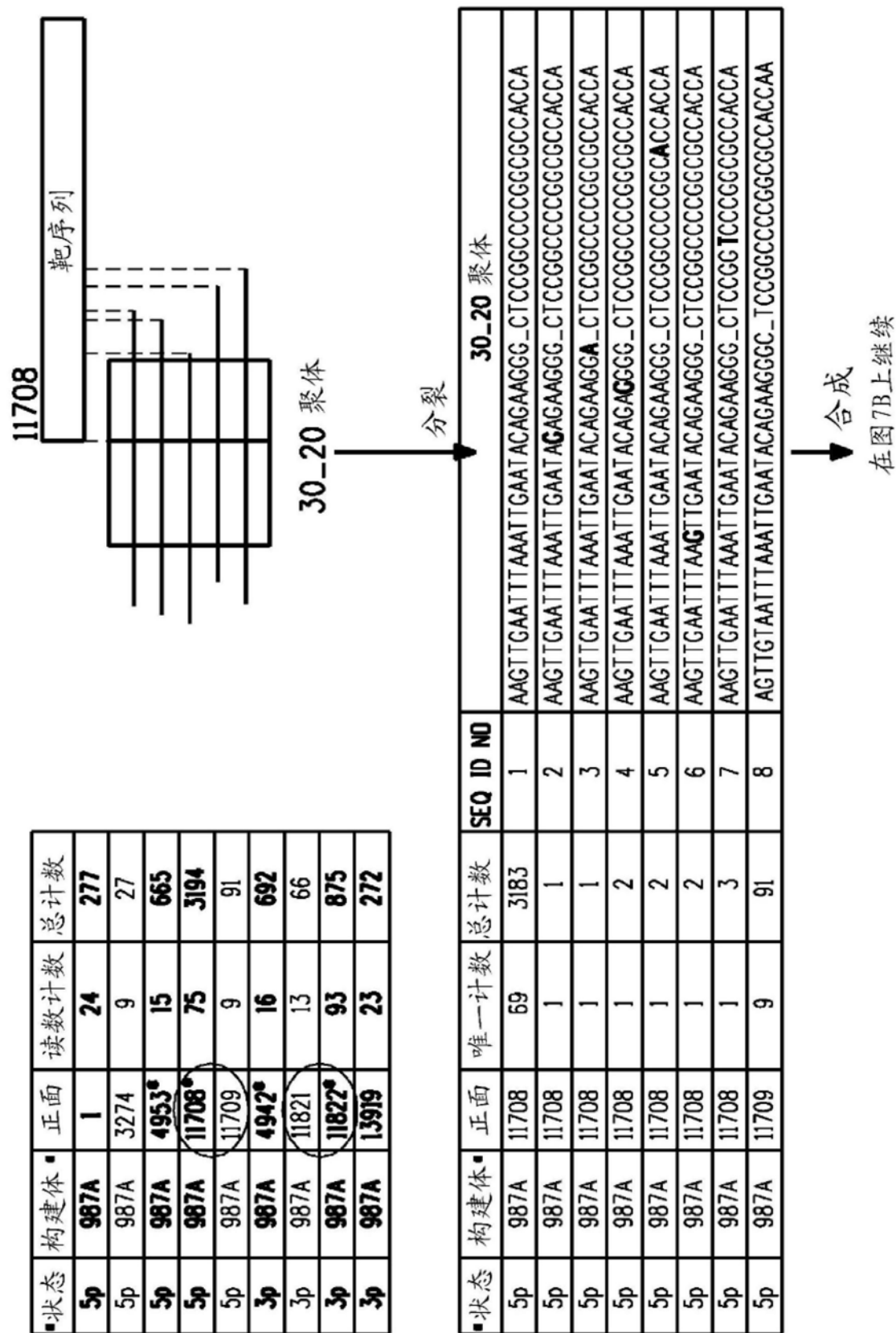


图7A

从图7A继续

合成

状态	构建体	正面	唯一计数	总计数	30_20 聚体
5p	987A	11708	69	3183	AAGTTGAATTTAAATTGAATACAGAGGG_CTCGGCCCCCGGGCCACCA
-5p	-987A	-11708	-1	-1	-AAGTTGAATTTAAATTGAATACAGAGGG_CTCGGCCCCCGGGCCACCA-
-5p	-987A	-11708	-1	-1	-AAGTTGAATTTAAATTGAATACAGAGGA_CTCGGCCCCCGGGCCACCA-
-5p	-987A	-11708	-1	-2	-AAGTTGAATTTAAATTGAATACAGAGGG_CTCGGCCCCCGGGCCACCA-
-5p	-987A	-11708	-1	-2	-AAGTTGAATTTAAATTGAATACAGAGGG_CTCGGCCCCCGGGCCACCA-
-5p	-987A	-11708	-1	-2	-AAGTTGAATTTAAAGTTGAATACAGAGGG_CTCGGCCCCCGGGCCACCA-
-5p	-987A	-11708	-1	-3	-AAGTTGAATTTAAATTGAATACAGAGGG_CTCGGGGTGGGGCCGGCCACCA-

图7B