



(12) 发明专利申请

(10) 申请公布号 CN 104685116 A

(43) 申请公布日 2015. 06. 03

(21) 申请号 201380040527. 0

代理人 余颖

(22) 申请日 2013. 06. 24

(51) Int. Cl.

(30) 优先权数据

C40B 50/06(2006. 01)

61/664, 118 2012. 06. 25 US

61/731, 627 2012. 11. 30 US

(85) PCT国际申请进入国家阶段日

2015. 01. 30

(86) PCT国际申请的申请数据

PCT/US2013/047370 2013. 06. 24

(87) PCT国际申请的公布数据

W02014/004393 EN 2014. 01. 03

(71) 申请人 GEN9 股份有限公司

地址 美国马萨诸塞州

(72) 发明人 M·E·赫德森 L·A·昆

D·辛德勒 S·阿彻 I·萨奥尔默

(74) 专利代理机构 上海专利商标事务所有限公

司 31100

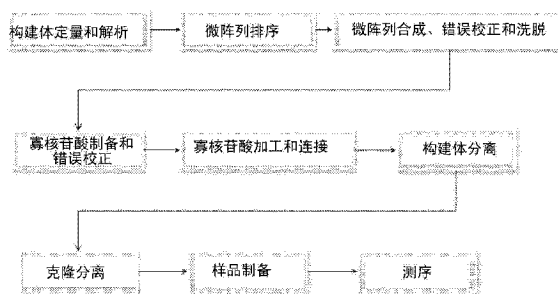
权利要求书2页 说明书22页
序列表2页 附图5页

(54) 发明名称

用于核酸组装和高通量测序的方法

(57) 摘要

本发明的一些方面的方法和设备涉及高保真多核苷酸的合成。具体而言,本发明的各方面涉及同时进行酶促去除扩增序列和将经加工的寡核苷酸连接成核酸组装体。根据一些实施方式,该方法包括提供多个寡核苷酸的步骤,其中各寡核苷酸包含 (i) 与目标核酸序列的不同部分相同的内部序列, (ii) 该内部序列 5' 端之侧的 5' 序列和在该内部序列 3' 端之侧的 3' 侧接序列,各所述侧接序列包含引物对的引物识别位点和限制性酶识别位点。



1. 一种生成有预定序列的目标核酸的方法,所述方法包括:
 - a) 提供第一混合物,所述第一混合物包含
 - (i) 第一寡核苷酸池,所述第一寡核苷酸池包含:包含与所述目标核酸 5' 端相同序列的第一组多个寡核苷酸、包含与所述目标核酸 3' 端相同序列的第二组多个寡核苷酸和包含与目标核酸序列的另外部分相同序列的多寡核苷酸组,所述寡核苷酸各自具有对应于下一个寡核苷酸中序列区域的重叠序列区域,所述第一混合物中的寡核苷酸合起来包含所述目标核酸序列;
 - (ii) 限制性酶,以及
 - b) 使所述第一混合物接触连接酶,从而生成所述目标核酸。
2. 如权利要求 1 所述的方法,所述方法还包括对所述目标核酸进行测序验证。
3. 如权利要求 1 所述的方法,所述方法还包括在步骤 (a) 前提供多个构建寡核苷酸,所述构建寡核苷酸各自包含 (i) 与目标核酸序列的不同部分相同的内部序列, (ii) 所述内部序列 5' 端和 3' 端之侧的 5' 和 3' 侧接序列,所述侧接序列各自包含引物对的引物识别位点和限制性酶识别位点。
4. 如权利要求 3 所述的方法,所述方法还包括扩增所述多个构建寡核苷酸。
5. 如权利要求 3 所述的方法,所述方法还包括对所述多个经扩增的寡核苷酸进行错误去除。
6. 如权利要求 5 所述的方法,其中使所述多个经扩增的寡核苷酸接触错配结合剂,所述错配结合剂选择性结合并剪切包含错配的双链寡核苷酸。
7. 如权利要求 4 所述的方法,其中在适于促进消化和连接的条件将所述限制性酶和所述连接酶加入单个经扩增寡核苷酸池中,从而生成包含组装的所述目标核酸序列和所述侧接序列的混合物。
8. 如权利要求 1 所述的方法,所述侧接序列各自包含共有的引物识别位点。
9. 如权利要求 1 所述的方法,所述限制性酶是 IIS 型限制性酶。
10. 如权利要求 7 所述的方法,其中使用 IIS 型限制性酶进行的消化产生多个粘性末端双链寡核苷酸,将所述多个粘性末端双链寡核苷酸以独特的线性排列方式进行连接。
11. 如权利要求 1 所述的方法,所述方法还包括使用能够识别位于第一寡核苷酸 5' 端和第二寡核苷酸 3' 端的引物识别位点的引物对扩增所述目标核酸。
12. 如权利要求 1 所述的方法,所述方法还包括对所述目标核酸进行测序以确认其序列准确性。
13. 如权利要求 12 所述的方法,其中通过高通量测序进行测序步骤。
14. 如权利要求 1 所述的方法,所述方法还包括从核酸序列池中分离至少一种具有预定序列的目标核酸。
15. 如权利要求 1 所述的方法,所述方法还包括加工所述目标核酸。
16. 如权利要求 1 所述的方法,所述方法还包括
 - c) 提供第二混合物,所述第二混合物包含
 - (i) 第二寡核苷酸池,所述第二寡核苷酸池包含:包含与所述目标核酸 5' 端相同序列的第一组多个寡核苷酸、包含与所述目标核酸 3' 端相同序列的第二组多个寡核苷酸和包含与目标核酸序列的另外部分相同的序列的多寡核苷酸组,所述寡核苷酸各自具有对应于下

一个寡核苷酸中序列区域的重叠序列区域,所述第二混合物中的寡核苷酸合起来包含第二目标核酸;

(ii) 限制性酶,以及

d) 使所述第二混合物接触连接酶,从而生成第二目标核酸。

17. 如权利要求 16 所述的方法,所述方法还包括组装至少两个目标核酸。

18. 如权利要求 17 所述的方法,所述组装通过分级组装进行。

19. 如权利要求 16 所述的方法,所述第一池中的所述第二组多个寡核苷酸包含限制性内切酶的限制性内切酶识别位点且所述第二池中的所述第一组多个寡核苷酸包含所述限制性内切酶的限制性内切酶识别位点。

20. 如权利要求 19 所述的方法,其中对所述至少两个目标核酸进行限制性内切酶消化和连接,从而形成长的的目标核酸构建体。

21. 如权利要求 20 所述的方法,所述长的目标核酸构建体的长度为至少约 10 千碱基。

22. 如权利要求 20 所述的方法,所述长的目标核酸构建体的长度为至少约 100 千碱基。

23. 一种生成有预定序列的目标核酸的方法,所述方法包括:

a) 提供多个寡核苷酸,所述寡核苷酸各自包含 (i) 与目标核酸序列的不同部分相同的内部序列, (ii) 所述内部序列 5' 端和 3' 端之侧的 5' 和 3' 侧接序列,所述侧接序列各自包含引物对的引物识别位点和限制性内切酶的限制性酶识别位点;

b) 使用所述引物对扩增至少一个亚组的所述寡核苷酸,从而生成多个经扩增的寡核苷酸;

c) 可选地对所述多个经扩增的寡核苷酸进行错误去除;

d) 提供具有所述限制性内切酶的限制性酶识别位点的环形载体;以及

c) 使所述多个经扩增的寡核苷酸和所述环形载体在单个池中接触所述限制性酶和连接酶,从而在所述载体中组装所述目标核酸,所述限制性酶能够识别所述限制性酶识别位点。

24. 如权利要求 23 所述的方法,所述方法还包括将所述载体转化入宿主细胞。

25. 一种用于组装具有预定序列的目标核酸的组合物,所述组合物包含:

a) 寡核苷酸池,所述寡核苷酸池包含:包含与所述目标核酸 5' 端相同序列的第一组多个寡核苷酸、包含与所述目标核酸 3' 端相同序列的第二组多个寡核苷酸和包含与目标核酸序列的另外部分相同序列的一组或多组多个寡核苷酸,所述寡核苷酸各自具有对应于下一个寡核苷酸中序列区域的重叠序列区域,所述池中的寡核苷酸合起来包含所述目标核酸;

b) 包含引物对的引物识别位点和限制性内切酶识别位点的多个共有序列;

c) 限制性内切酶;以及

d) 连接酶。

26. 如权利要求 25 所述的组合物,所述寡核苷酸是经扩增的。

27. 如权利要求 25 所述的组合物,所述寡核苷酸是经错误校正的。

28. 如权利要求 25 所述的组合物,还包含线性化的载体,所述载体包含与第一寡核苷酸相容的 5' 端和与第二寡核苷酸相容的 3' 端。

29. 如权利要求 25 所述的组合物,所述限制性内切酶是 IIS 型限制性内切酶。

用于核酸组装和高通量测序的方法

[0001] 相关申请

[0002] 本申请要求 2012 年 6 月 25 日提交的美国临时专利申请号 61/664,118 和 2012 年 11 月 30 日提交的美国临时专利申请号 61/731,627 的权益和优先权,上述各申请通过引用全文纳入本文。

技术领域

[0003] 本文提供了涉及合成和组装有预定序列的高保真核酸和核酸库的方法和设备。更具体的,提供了用于多核苷酸合成、错误降低和 / 或高通量测序验证的方法和设备。

[0004] 发明背景

[0005] 使用重组 DNA 化学技术,从自然中复制和扩增 DNA 序列并随后分解成组成部分是常见的。作为组成部分,随后序列重组或重新组装成新的 DNA 序列。然而,对天然可得序列的依赖性严重限制了研究者探索的可能性。虽然现在可以从单个核苷直接合成短 DNA 序列,但通常直接构建多核苷酸的大片段或组件(即长于约 400 碱基对的多核苷酸序列)是不可行的。

[0006] 能通过微芯片上的大规模平行定制合成进行寡核苷酸合成(Zhou 等,(2004) *Nucleic Acids Res.* 32:5409 ;Fodor 等(1991) *Science* 251:767)。然而,当前微芯片的表面积非常小,并且因而仅能生成少量的寡核苷酸。在释放到溶液中时,寡核苷酸以每条序列皮摩尔或更低浓度存在,该浓度不足以有效驱动双分子活化反应。无法对目前用于组装少量变体核酸的方法以成本节约的方式进行放大以生成大量特定的变体。同样,仍然需要用于高保真基因组组装等的改良的方法和装置。

[0007] 此外,通常通过化学反应合成微芯片上的寡核苷酸。错误的化学反应导致寡核苷酸中随机的碱基错误。化学核酸合成中关键限制因素之一是错误率。化学合成的寡核苷酸的错误率(例如每 100 个碱基中 1 个的缺失和约每 400 个碱基中 1 个的错配和插入)超过了通过酶的方式复制现有核酸(例如 PCR) 所得的错误率。因此,迫切需要一种新的技术来以成本节约的方式生产高产量的高保真多核苷酸。

[0008] 概述

[0009] 本发明的各方面涉及制备和 / 或组装高保真聚合物的方法、系统和组合物。本发明还提供了进行核酸组装反应和组装核酸的设备和方法。本发明的一个目的是提供合成定制多核苷酸的可行、经济的方法。本发明的另一个目的是提供生产合成的多核苷酸的方法,该方法的错误率低于通过本领域已知方法制备的合成的多核苷酸。

[0010] 根据一些实施方式,本发明提供了一种用于生产具有预定序列的目标核酸的方法。在一些实施方式中,该方法包括提供多个寡核苷酸的步骤,其中各寡核苷酸包含 (i) 与目标核酸的序列的不同部分相同的内部序列,(ii) 内部序列 5' 端之侧的 5' 侧接序列和内部序列 3' 端之侧的 3' 侧接序列,各侧接序列包含引物对的引物识别位点和限制性酶识别位点。在一些实施方式中,该方法还包括使用引物对扩增至少一组寡核苷酸,从而生成多个经扩增的寡核苷酸。随后可使多个经扩增的寡核苷酸在单个池中接触限制性酶和连接酶,

其中限制性酶能够识别限制性酶识别位点,从而生成目标核酸。

[0011] 在一些实施方式中,该方法包括对组装的目标核酸进行测序确认。在一些实施方式中,经扩增的双链寡核苷酸可包含一个序列错误或错配。在一些实施方式中,该方法包括对多个经扩增的寡核苷酸进行错误去除。在一些实施方式中,可使多个经扩增的寡核苷酸接触错配结合剂。所述错配结合剂可选择性地结合包含错配的双链寡核苷酸,导致结合和剪切作用。在一些实施方式中,可使多个经扩增的寡核苷酸接触错配识别剂,例如化学品,如赖氨酸、哌啶等。

[0012] 在一些实施方式中,在适于促进消化和连接的条件下将限制性酶和连接酶加入单个经扩增的寡核苷酸的池中,从而生成包含组装的目标核酸序列和侧接区的混合物。在一些实施方式中,各侧接区包含共有的引物识别位点。在一些实施方式中,所述限制性酶是 IIS 型限制性酶。使用 IIS 型限制性酶进行消化可产生多个粘性末端的双链结构寡核苷酸且所述多个粘性末端的双链结构寡核苷酸可以特定的线性排列形式进行连接。

[0013] 在一些实施方式中,该方法包括使用能够识别位于目标核酸的 5' 端和 3' 端的引物识别位点的引物对扩增目标核酸。在一些实施方式中,该方法包括对目标核酸进行测序以确认其序列准确性,例如通过高通量测序。在一些实施方式中,该方法包括从核酸序列池中分离至少一种具有预定序列的目标核酸。

[0014] 根据一些实施方式,本发明提供了一种进一步加工分离的核酸的方法。在一些实施方式中,该方法包括组装至少两种目标核酸。组装的步骤可以通过分级组装。在一些实施方式中,可对至少两种目标核酸进行限制性酶消化和连接,从而形成长的目標核酸构建体,例如长度至少约 10 千碱基或 100 千碱基。

[0015] 根据一些实施方式,本发明提供了一种用于生产载体中具有预定序列的目标核酸的方法。在一些实施方式中,提供了多个寡核苷酸,各寡核苷酸包含 (i) 与目标核酸的序列的不同部分相同的内部序列,(ii) 内部序列 5' 端之侧的 5' 侧接序列和内部序列 3' 端之侧的 3' 侧接序列,各侧接序列包含引物对的引物识别位点和限制性内切酶的限制性酶识别位点。在一些实施方式中,可使用引物对扩增至少一组寡核苷酸,从而生成多个经扩增的寡核苷酸。在一些实施方式中,可对多个经扩增的寡核苷酸进行错误去除和 / 或校正。在一些实施方式中,提供了具有限制性内切酶的限制性酶识别位点的环形载体。在一些实施方式中,可使多个经扩增的寡核苷酸和环形载体在单个池中接触限制性酶和连接酶,其中限制性酶能够识别限制性酶识别位点,从而在载体中组装目标核酸。在一些实施方式中,该方法还包括将载体转化至宿主细胞中并测序以验证目标核酸序列。

[0016] 根据一些实施方式,本发明提供了一种用于组装具有预定序列的目标核酸的组合物。在一些实施方式中,所述组合物包含多个寡核苷酸,其中各寡核苷酸包含 (i) 与目标核酸的序列的不同部分相同的内部序列,(ii) 内部序列 5' 端之侧的 5' 侧接序列和内部序列 3' 端之侧的 3' 侧接序列,各侧接序列包含引物对的引物识别位点和限制性内切酶的限制性酶识别位点。在一些实施方式中,所述组合物还包含限制性内切酶和 / 或连接酶。在一些实施方式中,所述组合物还包含载体,所述载体包含限制性内切酶的一对酶识别位点。在一些实施方式中,所述限制性内切酶是 IIS 型限制性内切酶。

[0017] 在一些实施方式中,对多个寡核苷酸进行扩增和 / 或错误校正。

[0018] 在本发明的一些方面中,生成具有预定序列的目标核酸的方法包括提供第一混合

物,其包含 (i) 限制性酶和 (ii) 第一寡核苷酸池,所述第一寡核苷酸池包括:包含与目标核酸 5' 端相同序列的第一寡核苷酸、包含与目标核酸 3' 端相同序列的第二寡核苷酸和包含与目标核酸序列另外部分相同序列的多寡核苷酸组,各寡核苷酸具有对应于下一寡核苷酸中序列区域的重叠序列区域,第一池中的寡核苷酸合起来包含目标核酸序列;并使混合物与连接酶接触,从而生成目标核酸。随后可对目标核酸进行测序验证。

[0019] 在一些实施方式中,本发明的方法包括提供一个构建寡核苷酸池并涉及不同阶段时寡核苷酸的扩增。术语“构建寡核苷酸”指可用于组装比构建寡核苷酸本身长的核酸分子的单链寡核苷酸。构建寡核苷酸可以是单链寡核苷酸或双链寡核苷酸。在一些实施方式中,构建寡核苷酸是合成的寡核苷酸并可以在基质上平行地合成。

[0020] 在一些实施方式中,该方法还包括在提供第一混合物前提供多个构建寡核苷酸的步骤,其中各构建寡核苷酸包含 (i) 与目标核酸的序列的不同部分相同的内部序列,(ii) 内部序列 5' 端之侧的 5' 侧接序列和内部序列 3' 端之侧的 3' 侧接序列,各侧接序列包含引物对的引物识别位点和限制性酶识别位点。在一些实施方式中,各侧接区可包含共同的引物识别位点。在一些实施方式中,可扩增多个构建寡核苷酸。在一些实施方式中,寡核苷酸可包含序列错误或错配。在一些实施方式中,可对多个经扩增的寡核苷酸进行错误去除。例如,可使多个经扩增的寡核苷酸接触错配结合剂,所述错配结合剂选择性结合并剪切包含错配的双链寡核苷酸。

[0021] 在一些实施方式中,可在适于促进消化和连接的条件将限制性酶和连接酶加入单个经扩增的寡核苷酸的池中,从而生成包含组装的目标核酸序列和侧接区的混合物。在一些实施方式中,所述限制性酶可以是 IIS 型限制性酶且使用 IIS 型限制性酶进行的消化可生成多个粘性末端双链寡核苷酸,其中所述多个粘性末端双链寡核苷酸以特殊的线性排列进行连接。

[0022] 在一些实施方式中,该方法还包括使用能够识别位于第一寡核苷酸 5' 端和第二寡核苷酸 3' 端的引物识别位点的引物对扩增目标核酸。

[0023] 在一些实施方式中,该方法还包括对目标核酸进行测序以确认其序列准确性,例如通过高通量测序。

[0024] 在一些实施方式中,该方法还包括从核酸序列池中分离至少一种具有预定序列的目标核酸。

[0025] 在一些实施方式中,该方法还包括加工目标核酸。

[0026] 在一些实施方式中,该方法还包括提供第二混合物,其包含 (i) 限制性酶和 (ii) 第二寡核苷酸池,所述第二寡核苷酸池包含:包含与目标核酸 5' 端相同序列的第一寡核苷酸、包含与目标核酸 3' 端相同的序列的第二寡核苷酸和包含与目标核酸序列另外部分相同序列的多寡核苷酸组,各寡核苷酸具有对应于下一寡核苷酸中序列区域的重叠序列区域,第二池中的寡核苷酸合起来包含第二目标核酸。在一些实施方式中,使第二混合物接触连接酶,从而生成第二目标核酸。在一些实施方式中,第一池中的第二寡核苷酸包含限制性内切酶的限制性内切酶识别位点且第二池中的第一寡核苷酸包含限制性内切酶的限制性内切酶识别位点。

[0027] 在一些实施方式中,该方法还包括组装至少两个目标核酸。在一些实施方式中,组装的步骤是通过分级组装。在一些实施方式中,对所述至少两个目标核酸进行限制性内切

酶消化和连接,从而形成长的目标核酸构建体。在一些实施方式中,所述长的目标核酸构建体的长度为至少约 10 千碱基或至少约 100 千碱基。

[0028] 在一些方面,本发明涉及用于组装具有预定序列的目标核酸的组合物,所述组合物包含多个寡核苷酸,其包含:包含与目标核酸 5' 端相同序列的第一寡核苷酸、包含与目标核酸 3' 端相同的序列的第二寡核苷酸和包含与目标核酸序列另外部分相同序列的一个或多个寡核苷酸,各寡核苷酸具有对应于下一寡核苷酸中序列区域的重叠序列区域,多个寡核苷酸合起来包含目标核酸;多个共有序列包含引物对的引物识别位点和限制性内切酶识别位点。在一些实施方式中,所述组合物还包含限制性内切酶和 / 或连接酶。所述限制性内切酶可以是 IIS 型限制性内切酶。

[0029] 在一些实施方式中,可对多个寡核苷酸进行扩增和 / 或错误校正。在一些实施方式中,所述组合物还可包含线性化的载体,所述载体包含与第一寡核苷酸相容的 5' 端和与第二寡核苷酸相容的 3' 端。

[0030] 附图简要说明

[0031] 图 1 显示了根据本发明的一个实施方式的高保真核酸组装的示例性过程。

[0032] 图 2 显示了具有预定序列的多核苷酸的组装方法的非限制性示例。

[0033] 图 3 显示了将具有预定序列的多核苷酸组装至载体中的组装方法的非限制性示例。

[0034] 图 4 显示了具有预定序列的多核苷酸的分级组装方法的非限制性示例。

[0035] 图 5 显示了具有限制性内切酶识别位点(下划线)的质粒 pG9-1 的核苷酸序列。

[0036] 图 6 显示了测序验证的非限制性示例性方法。

[0037] 发明详述

[0038] 本发明的各方面对于优化核酸组装反应和减少不准确组装的核酸数目是有用的。本发明的方法和组合物可促进获得具有预定序列的目标序列的方法。因此,本发明的方法和组合物可增加获得正确组装的核酸的可能性,从而减少生产具有预定序列的核酸相关的成本和时间。

[0039] 本发明的各方面可用于提高一种或多种初始或中间组装反应的产量。在一些实施方式中,本发明的方法和组合物可通过避免需要分离多个组装步骤(例如酶消化、纯化和连接步骤)来提高总组装过程的效率。因此,本发明的一些方面允许可预测和 / 或可靠的组装策略并可显著减少基因合成所需的时间和步骤并提高中间产物或最终核酸产物的产量和 / 或准确性。

[0040] 在本发明的一些方面中,组装过程包括设计和实施核酸组装策略,该核酸组装策略能够调和已知或预测会干扰一个或多个组装步骤的序列特征。例如,可分析待合成的核酸序列中会干扰一个或多个组装步骤的序列特征,如重复的序列,具有显著高或低 GC 含量的序列和 / 或与二级结构相关的其他序列。本领域技术人员应理解,某些序列特征会干扰多重组装反应(例如基于聚合酶的延伸反应)和 / 或促进不需要的组装产物的形成,从而减少或阻止正确的核酸产物的组装。在一些实施方式中,如果在目标核酸序列中鉴定到多个干扰序列特征,有用的策略可涉及在组装期间分离干扰序列特征。例如,可在涉及多个中间片段或结构模块(其被设计以仅包含少量的干扰序列(例如 0、1、2 或 3 个))的过程中组装目标核酸。在一些实施方式中,各中间片段或结构模块可含有最多一种干扰序列特征。

因此,可有效地组装各中间片段。在一些实施方式中,核酸片段或结构模块的设计可排除来自其 5' 和 / 或 3' 端的干扰序列特征。因此,可从设计用于组装反应的相邻起始核酸之间的互补重叠区域中排除干扰序列特征。这会防止或减少对序列特异性杂交反应的干扰,该序列特异性杂交反应对核酸的正确组装是重要的。在一些实施方式中,从中间体结构模块的 3' 和 / 或 5' 端排除干扰序列特征是足够的。例如,干扰序列特征可位于来自结构模块 3' 端和 / 或 5' 端的至少一个核苷酸处,优选位于来自结构模块 3' 端和 / 或 5' 端的 2、3、4、5 或更多个核苷酸(例如 5-10、10-15、15-20 或更多个核苷酸)处。

[0041] 本发明的各方面可与体外和 / 或体内核酸组装步骤联用。

[0042] 本文提供的方法和组合物的各方面对增加核酸合成和组装反应的精确性、产率、通量、和 / 或成本效益是有用的。本文使用的术语“核酸”、“多核苷酸”、“寡核苷酸”可互换使用,并且指核苷酸的天然产生或合成的聚合物形式。本发明所述寡核苷酸和核酸分子可以从天然产生的核苷酸形成,例如形成脱氧核糖核酸(DNA)或核糖核酸(RNA)分子。或者,该天然产生的寡核苷酸可以包含改变其性质的结构修饰,例如肽核酸(PNA)或锁核酸(LNA)。有天然产生碱基或人工碱基的寡核苷酸和核酸分子的固相合成为本领域熟知。应理解这些术语包含从核苷酸类似物中生成的 RNA 或 DNA 的等同物、类似物和应用用于要描述的实施方式时的单链或双链多核苷酸。本发明中可用的核苷酸包含例如天然产生的核苷酸(例如核糖核苷酸或脱氧核糖核苷酸),或者核苷酸的天然或合成修饰、或者人工碱基。本文使用的术语单体指小分子组成员,其是并且能结合在一起以形成低聚物、聚合物或由两个或更多成员构成的化合物。聚合物中所述单体的特定顺序在本文中称为聚合物的“序列”。该单体组包含但不限于例如常见 L-氨基酸组、D-氨基酸组、合成和 / 或天然氨基酸组、核苷酸组及戊糖和己糖组。本文所述的本发明的方面主要关于制备寡核苷酸,但是易于应用于制备其他聚合物例如肽或多肽、多糖、磷脂、杂聚物、聚酯、聚碳酸酯、聚脲、聚酰胺、聚乙烯亚胺、聚亚芳基硫醚、聚硅氧烷、聚酰亚胺、聚乙酸酯或任何其他聚合物。

[0043] 目标核酸

[0044] 本文使用的术语“预定序列”指在所述聚合物合成或组装前已知且已选择的聚合物的序列。具体地,本文所述本发明的各方面主要涉及核酸分子的制备,在核酸分子合成或组装前已知且已选择寡核苷酸或多核苷酸的序列。在本文所提供技术的一些实施方式中,固定的寡核苷酸或多核苷酸被用作材料来源。在多个实施方式中,本文所述方法使用多个寡核苷酸,各序列基于待合成的最终多核苷酸构建体的序列确定。在一个实施方式中,寡核苷酸是短核酸分子。例如,寡核苷酸的长度可以是 10 至约 300 个核苷酸、20 至约 400 个核苷酸、30 至约 500 个核苷酸、40 至约 600 个核苷酸,或超过约 600 个核苷酸。然而,可以使用更短或更长的寡核苷酸。寡核苷酸可设计成具有不同长度。在一些实施方式中,多核苷酸构建体序列可以分成更短序列组,该序列能使用本文所述方法平行合成并组装成单个或多个所需多核苷酸构建体。

[0045] 在一些实施方式中,目标核酸可具有天然产生的基因和 / 或其他天然产生的核酸(例如天然产生的编码序列、调节序列、非编码序列、染色体结构序列(如端粒或中心粒序列)等,其任意片段或其两种或多种的任意组合)的序列或者非天然产生的序列。在一些实施方式中,可将目标核酸设计为具有在一个或多个位置上不同于天然序列的序列。在其他实施方式中,可将目标核酸设计为具有全新的序列。然而,应理解目标核酸可包括一种或

多种天然产生的序列、非天然产生的序列或其组合。

[0046] 在一些实施方式中,本文提供了组装包含具有预定序列差异的核酸的文库的方法。本文提供的组装方案能用于生成代表很多不同感兴趣核酸序列的非常大的文库。例如,本文提供的方法可用于组装具有超过 10 种不同序列变体的文库。在一些实施方式中,核酸文库是序列变体文库。序列变体可以是单个天然产生蛋白编码序列的变体。然而,在一些实施方式中,序列变体可以是多个不同蛋白编码序列的变体。因此,本发明的一个方面涉及用于制备精确的高密度核酸文库的组装策略的设计。本文提供的技术的另一个方面涉及组装精确的高密度核酸文库。本文提供的技术的各方面还涉及精确的高密度核酸文库。高密度核酸文库可以包含超过 100 种不同序列变体(如约 10^2 至 10^3 、约 10^3 至 10^4 、约 10^4 至 10^5 、约 10^5 至 10^6 、约 10^6 至 10^7 、约 10^7 至 10^8 、约 10^8 至 10^9 、约 10^9 至 10^{10} 、约 10^{10} 至 10^{11} 、约 10^{11} 至 10^{12} 、约 10^{12} 至 10^{13} 、约 10^{13} 至 10^{14} 、约 10^{14} 至 10^{15} 或更多不同序列),其中与随机序列相反,高百分比的不同序列是特异性序列(如,大于约 50%、大于约 60%、大于约 70%、大于约 75%、大于约 80%、大于约 85%、大于约 90%、约 91%、约 92%、约 93%、约 94%、约 95%、约 96%、约 97%、约 98%、约 99% 或更多的序列是预定的感兴趣的序列)。

[0047] 在某些实施方式中,目标核酸可包括功能序列(例如蛋白结合序列、调节序列、编码功能蛋白的序列等,或其组合)。然而,在一些实施方式中,目标核酸可缺少特定功能序列(例如目标核酸可仅包括蛋白结合序列、调节序列或蛋白编码序列的非功能片段或变体,或任何其他非功能的天然产生的或合成的序列,或其任意非功能的组合)。某些目标核酸可包括功能和非功能序列。本文中更详细地描述了目标核酸及其用途的这些和其他方面。

[0048] 在一些实施方式中,可在单个多重组装反应(例如单核苷酸组装反应)中组装目标核酸。然而,也可由多个核酸片段组装目标核酸,其中可在独立的多重寡核苷酸组装反应中生成各核酸片段。应理解在一些实施方式中,可将通过多重寡核苷酸组装生成的一个或多个核酸片段与获自另一来源的一个或多个核酸分子(例如限制性片段、核酸扩增产物等)混合以形成目标核酸。在一些实施方式中,在第一反应中组装的目标核酸可以用作后续组装反应的输入核酸片段以生成较大的目标核酸。本文所用术语“多重组装”和“多重寡核苷酸组装反应”通常指涉及多个起始核酸(例如多个至少部分重叠的核酸)的组装反应,该起始核酸经组装以生成较大的目标核酸。

[0049] 组装方法

[0050] 图 1 显示了根据本发明的一个实施方式用于组装核酸的方法。首先,获得序列信息。序列信息可以是待组装的预定目标核酸的序列。在一些实施方式中,可以来自客户的指令的形式接受序列。在一些实施方式中,可以核酸序列(例如, DNA 或 RNA)的形式接受序列。在一些实施方式中,可以蛋白序列的形式接受序列。该序列可转化为 DNA 序列。例如,如果所得的序列是 RNA 序列,可以用 T 来代替 U 以得到相应的 DNA 序列。如果所得的序列是蛋白序列,可以使用合适的氨基酸密码子将该蛋白序列转化为 DNA 序列。

[0051] 在一些实施方式中,可对序列信息进行分析以确定组装策略(如待组装片段(在本文中也称作结构模块、寡核苷酸或中间片段)的数目和序列)以生成目标核酸的预定序列。在一些实施方式中,序列分析可包括扫描一种或多种干扰序列特征是否存在,该干扰序列特征已知或预计干扰寡核苷酸合成、扩增或组装。例如,干扰序列结构可以是在至少 10 个碱基(例如,10-20、20-50、50-100 或超过 100 个碱基)的长度上具有低 GC 含量(例如,

低于 30% GC、低于 20% GC、低于 10% GC 等) 的序列, 或者可以形成二级结构或茎环结构的序列。一旦通过该过滤, 可将核酸序列分为较小碎片, 如寡核苷酸结构模块。

[0052] 在一些实施方式中, 在构建体定量和解析步骤之后, 可设计用于组装的合成的寡核苷酸 (例如序列、大小和数量)。可使用标准 DNA 合成化学 (例如, 亚磷酰胺方法) 来生产合成的寡核苷酸。可使用本领域已知的任何合适技术或如本文详述在固体支持物 (例如微阵列) 上合成合成的寡核苷酸。可在扩增之前将寡核苷酸从微阵列上洗脱下来或在微阵列上扩增寡核苷酸。应理解可将不同的寡核苷酸设计为具有不同的长度。

[0053] 在一些实施方式中, 可扩增各目标序列的结构模块寡核苷酸。例如, 可对寡核苷酸进行设计以在其 3' 端和 5' 端上具有引物结合序列并可使用合适的引物对通过聚合酶链式反应 (PCR) 对寡核苷酸进行扩增。

[0054] 应理解合成的寡核苷酸可具有序列错误。因此, 可如本文中更详细描述的那样选择或筛选寡核苷酸制备物以除去含有错误的分子。含有错误的寡核苷酸可以是在两条链上都具有错误的同源双链体 (即在两条链上都有不正确的互补核苷酸、缺失或添加)。在一些实施方式中, 可使用涉及使双链核酸变性并再退火的技术除去序列错误。在一些实施方式中, 如果含有各单个错误的核酸以比在相同位置具有正确序列的核酸低的频率存在于核酸制备物中, 则含有互补错误的单链核酸可能无法再退火并相互结合。相反, 含有错误的单链可能与不含错误或含有一种或多种不同错误的互补链进行再退火。结果是, 最终含有错误的链在再退火的反应产物中以异源双链分子的形式存在。无错误的核酸链可与含有错误的链或者其他无错误的链发生再退火。再退火的无错误的链在再退火样品中形成同源双链体。因此, 通过从再退火的寡核苷酸制备物中去除异源双链分子, 可减少含有错误的核酸的量或频率。可使用本领域已知的除去异源双链分子的任何适当方法, 包括色谱、电泳、异源双链分子的选择性结合等。在一些实施方式中, 可使用选择性 (例如特异性) 结合异源双链核酸分子的错配结合蛋白。在一些实施方式中, 该错配结合蛋白可用于溶液中或者固定在支持物上的双链寡核苷酸或多核苷酸上。

[0055] 在一些实施方式中, 使用 MutS 过滤方法 (例如使用 MutS、MutS 同源物或其组合) 除去含有错误的寡核苷酸。在大肠杆菌中, 似乎以二聚体形式起作用的 MutS 蛋白作为错配识别因子发挥作用。在真核细胞中, 已鉴定到至少三种 MutS 同源 (MSH) 蛋白, 即 MSH2、MSH3 和 MSH6, 且其形成异源二聚体。例如在酿酒酵母中, MSH2-MSH6 复合物 (也称作 MutS α) 识别碱基错配和单核苷酸插入 / 缺失环, 而 MSH2-MSH3 复合物 (也称作 MutS β) 识别最多 12-16 个核苷酸的插入 / 缺失, 但其起基本冗余的作用。错配结合蛋白可获自重组或天然来源。错配结合蛋白可以是热稳定的。在一些实施方式中, 可以使用来自嗜热微生物的热稳定错配结合蛋白。热稳定 DNA 错配结合蛋白的示例包括但不限于: Tth MutS (来自嗜热栖热菌 (*Thermus thermophilus*))、Taq MutS (来自水生栖热菌 (*Thermus aquaticus*))、Apy MutS (来自嗜火液菌 (*Aquifex pyrophilus*))、Tma MutS (来自海栖热袍菌 (*Thermotoga maritima*))、其同源物、任何其他合适的 MutS 或其中两种或多种的任意组合。

[0056] 已证明获自不同物种的 MutS 可对特定的错配或对不同的错配具有不同的亲和性。在一些实施方式中, 可使用对不同错配具有不同亲和性的不同 MutS 的组合。

[0057] 在一些实施方式中, 可使用使用一种或多种修复蛋白的酶复合物。修复蛋白的示例包括但不限于: 用于错配识别的 MutS、用于在目标链上导入缺口的 MutH 和用于介导 MutH

和 MutS 之间相互作用的 MutL、其同源物或其任意组合。在一些实施方式中,错配结合蛋白复合物是 MutHLS 酶复合物。

[0058] 在一些实施方式中,滑动夹 (sliding clamp) 技术可用于富集无错误的双链寡核苷酸。在一些实施方式中,MutS 或其同源物可与 DNA 夹蛋白 (DNA clamp protein) 相互作用。DNA 夹蛋白的示例包括但不限于由 dnaN 基因编码的细菌滑动夹蛋白 DnaN,其以同源二聚体的形式发挥作用。在一些实施方式中,MutS 蛋白 (或其同源物) 与夹蛋白之间的相互作用可提高 MutS 在结合错配中的有效性。

[0059] 在一些实施方式中,可使用来自 S1 蛋白家族的酶 (例如 CELI, CELII 或其同源物,如 RESI,或其组合) 去除含有错误的寡核苷酸。来自 S1 蛋白家族的酶可识别碱基错配、插入和缺失环。在一些实施方式中,通过仅有的一条或两条 DNA 链,这类酶可优先结合霍利迪连结体,之后剪切识别位点。在一些实施方式中,可以使用 S1 蛋白的热稳定等价物。

[0060] 在一些实施方式中,可使用结合错配碱基位点的小分子、化学或无机材料去除含有错误的寡核苷酸。在错配位点处,核苷酸碱基是超螺旋的且易于进行化学修饰反应。可在化学剪切方法中使用诸如高锰酸、羟胺、赖氨酸和 / 或五胺钉的材料来分别修饰错配的胸腺嘧啶和胞嘧啶。随后使用哌啶处理所得修饰的 DNA 以在脱碱基位点处引起剪切。在一些实施方式中,可使用二价盐监测剪切的特异性。

[0061] 在一些实施方式中,在下一步骤中,通过依次除去共有序列并随后连接成更长的多寡核苷酸构建体的方式混合经错误校正的寡核苷酸。

[0062] 在本发明的一些方面中,酶促消化共有序列去除步骤与连接步骤联用。本领域技术人员应理解,本发明的方法允许同时去除共有序列并连接成目标核酸构建体并无需酶促去除、基于珠的捕获和连接连续步骤。此外,本领域技术人员应理解,相对于标准基因组装方法,本发明的方法具有许多优势,例如:

[0063] 增加生产效率:使用标准的独立酶促去除共有序列,反应在设定的时间点后停止,其中未反应的底物或未消化的寡核苷酸仍然存在,作为进一步去除的对象。本领域技术人员应理解,由于连接反应产生了不是酶促去除底物的所需产物,所以去除和连接步骤的组合能够不可逆地驱动反应产生所需产物。

[0064] 成本节约:根据本发明的一些方面的方法是成本节约的,因为在去除共有序列和连接之间不再需要纯化步骤。由于删去了纯化步骤,本发明的各方面也不需要生物素标记的引物。相关的节约还有:收到非生物素化引物的订货时间比含生物素的引物短。

[0065] 时间效率:通过删去酶促共有序列去除和连接之间的纯化步骤,减少了基因合成所需的时间和步骤数目。

[0066] 容易地添加其他序列的机会,这与序列大小无关。由于用于除去不需要的序列的部分纯化步骤是基于大小,删去纯化步骤可除去对任何其他用于基因合成的待添加序列大小的限制。这可包括一步连接至载体,或添加共有的侧接序列。

[0067] 该方法允许在基因中使用限制性位点,其用于基因合成过程本身。在先前的方法中,不能使用这些限制性位点,因为切割位点会导致形成会在纯化步骤中被除去的小 DNA 碎片。使用这些限制性位点可允许进行递推 (分级) 基因合成以构建更长的核酸。

[0068] 本领域技术人员应理解,寡核苷酸组装后,组装产物 (例如最终目标核酸和中间体核酸片段) 可含有不需要的序列。错误可能由在寡核苷酸合成期间,或在寡核苷酸组装

成较长的核酸期间引入的序列错误所致。在一些实施方式中,可将具有正确预定序列的核酸与其他核酸序列分离(在本文中也可称作制备型体外克隆)。在一些实施方式中,可通过将正确序列与其他不正确序列选择性地分离来分离正确序列。例如,可将具有正确序列的核酸选择性地移动或转移至支持物的不同特征或另一块板。或者,可将具有不正确序列的核酸选择性地从包含感兴趣的核酸的特征中移除(参见例如 PCT/US2007/011886,其通过引用全文纳入本文)。

[0069] 在一些实施方式中,在寡核苷酸加工和连接后,可通过克隆分离来分离组装构建体或组装的构建体的拷贝。可使用例如高通量测序来对组装构建体进行序列验证。在一些实施方式中,可使用对个体分子进行测序(如单一分子测序)或对目标核酸序列的扩增群的测序(如聚合酶克隆测序)来实施目标核酸序列的序列测定。可使用任何合适的测序方法,如通过杂交测序、通过连接测序或通过合成测序。

[0070] 本方面的一些方面涉及使用本文所述方法的基因合成平台。在一些实施方式中,基因合成平台可与下一代测序平台(例如通过杂交测序、通过合成测序或通过连接测序或任何其他合适的测序方法)联用。

[0071] 在一些实施方式中,组装方法可以包括数种平行和/或顺序反应步骤,其中多种不同核酸或寡核苷酸组被合成或固定、扩增并且合并以组装(如通过本文所述延伸或连接)生成更长核酸产物,从而用于进一步组装、克隆或其他应用(参见 PCT 申请 PCT/US09/55267,其通过引用全文纳入本文)。

[0072] 寡核苷酸合成

[0073] 在一些实施方式中,本文提供的方法和设备使用固定在表面或基底上的寡核苷酸(如结合于支持物的寡核苷酸)。本文所用术语“支持物”和“基底”可互换使用,并且指聚合物(如核酸)在其上合成或固定的多孔或非多孔溶剂不溶性材料。本文使用的“多孔”指材料包含有基本一致直径(例如 nm 范围)的孔。多孔材料包括纸、合成过滤器等。在这种多孔材料中,反应可以在孔中进行。所述支持物能具有很多形状中的任何一种,例如销型、条、板、平盘、杆状、弯曲、圆柱形结构、颗粒(包含珠、纳米颗粒)等。所述支持物可具有不同宽度。所述支持物可以是亲水性的,或可以制成亲水性的,并且包含无机粉末(如二氧化硅、硫酸镁和氧化铝)、天然聚合材料(特别是纤维素材料和纤维素衍生材料,例如包含纤维的纸(如滤纸、色谱纸等))、合成或改性天然产生的聚合物(如硝酸纤维素、乙酸钠纤维素、聚(氯乙烯)、聚丙烯酰胺、交联的葡聚糖、琼脂糖、聚丙烯酸酯、聚乙烯、聚丙烯、聚(4-甲基丁烯)、聚苯乙烯、聚甲基丙烯酸酯、聚(对苯二甲酸乙二酯)、尼龙、聚(丁酸乙烯酯)、聚偏二氟乙烯(PVDF)膜、玻璃、可控孔度玻璃、磁性可控孔度玻璃、陶瓷、金属等),或者单独使用或与其他材料联用。在一些实施方式中,在阵列形式上合成寡核苷酸。例如,在常见支持物上原位合成单链寡核苷酸,其中各寡核苷酸在底物的单独或不连续特征(或点)上合成。在优选实施方式中,单链寡核苷酸结合在支持物或特征的表面。本文所用术语“阵列”是指用于存储、路径选择、扩增和释放寡核苷酸或互补寡核苷酸用于进一步反应的离散特征的排列。在优选实施方式中,所述支持物或阵列是可寻址的:所述支持物包含在支持物上特定预定位置(即“地址”)上的两个或更多个离散的可寻址特征。因此,阵列上的各寡核苷酸分子位于支持物上已知并确定的位置。各寡核苷酸序列能从支持物上其位点处测定。

[0074] 在一些实施方式中,寡核苷酸在表面或阵列的离散特征上连接、点样、固定、表面

结合、支持或合成。寡核苷酸可以共价连接到表面或在表面上沉积。阵列可以构建、定制购买或购自商业销售商（如安捷伦公司 (Agilent)、昂飞公司 (Affymetrix)、尼姆布拉齐 (Nimblegen) 公司）。各种构建方法为本领域熟知，如无掩膜阵列合成器、利用掩膜的光定向方法、流动通道方法、点样法等。在一些实施方式中，可以使用无掩膜阵列合成器 (MAS) 在固体支持物上合成构建和 / 或选择寡核苷酸。无掩膜阵列合成器描述于例如 PCT 申请号 WO 99/42813 和相应的美国专利号 6,375,903 中。其他示例为已知的无掩膜设备，其能制造定制 DNA 微阵列，其中所述阵列中的各特征中有所需序列的单链 DNA 分子。其他合成构建和 / 或选择寡核苷酸的方法包含例如利用掩膜的光导向的方法、流动通道方法、点样法、销型方法和利用多个支持物的方法。用于合成寡核苷酸的利用掩膜的光导向的方法（例如，VLSIPS™方法）描述于美国专利号 5,143,854、5,510,270 和 5,527,681 中。这些方法涉及活化固体支持物上的预定区域，并随后使支持物接触预选择的单体溶液。选定区域能用通过掩膜的光源辐射来活化，方法大多按照集成电路制造中使用的光刻法技术。支持物的其他区域保持失活，因为照射被掩膜封闭并且其仍保持化学保护。因此，光模式定义支持物上的哪个区域与给定的单体反应。通过重复活化不同预定区域组并使不同单体溶液接触支持物，在支持物上产生聚合物的不同阵列。能任选使用其他步骤，例如从支持物上洗涤未反应的单体溶液。其他可应用方法包括机械技术如描述于美国专利号 5,384,261 中的那些。可应用于在单个支持物上合成构建和 / 或选择寡核苷酸的其他方法描述于例如美国专利号 5,384,261 中。例如，可通过 (1) 在预定区域上确定的通道中流动，或者 (2) 在预定区域上“点样”的方式将试剂递送到支持物上。也可以使用其他方法以及点样和流动的组合。在各情况下，在将单体溶液递送到多个反应位点时，将支持物上的某些活化区域与其他区域机械分开。流动通道方法包括，例如微流体系统以控制在固体支持物上寡核苷酸的合成。例如，可以通过在固体支持物的表面上形成流动通道的方式在固体支持物的选定区域上合成不同的聚合物序列，其中合适的试剂流过该流动通道或者位于该流动通道中。用于在固体支持物上制备寡核苷酸的点样法涉及通过直接置于选定区域来以相对少量递送反应物。在一些步骤中，能用溶液喷洒或覆盖整个支持物表面，前提是这样做更有效。可以通过从一个区域到另一个区域移动的分配器逐滴加入精确测量的单体溶液等分样品。用于在固体支持物上合成寡核苷酸的销型方法描述于例如美国专利号 5,288,514 中。销型方法利用有多个销型或其他延伸的支持物。该销型各自同时插入到浅盘中的单个试剂容器中。96 销型的阵列通常利用 96 容器的浅盘，例如 96 孔微量滴定皿。各个浅盘填充有在单个销型上特定化学反应中偶联的特定试剂。因此，所述浅盘会经常包含不同试剂。由于已对化学反应进行优化从而使各反应能在较相似的反应条件组下进行，因此可同时进行多个化学偶联的步骤。

[0075] 在另一个实施方式中，可在多个支持物上合成定多个寡核苷酸。一个示例是描述于例如美国专利号 5,770,358、5,639,603 和 5,541,061 的基于珠的合成方法。为了在珠上合成分子（如寡核苷酸），将大量的珠悬浮于容器中的合适运载体（如水）中。提供了带有可选间隔分子的珠，所述分子有其复合的活性位点，可选为保护基团。在合成的各步骤中，分开所述珠以使其偶联至多种容器中。新生寡核苷酸链脱保护后，不同单体溶液加入到各容器中，从而在给定容器内的所有珠上发生相同的核苷酸加成反应。随后用过量试剂洗涤珠，收集入单个容器中，混合并重新分布到另外的多个容器以准备下一轮合成。应注意到由

于在开始利用的大量珠,在容器内同样会随机分布大量珠,在多轮随机加入碱基后,各自在其表面具有合成的独特寡核苷酸序列。可以用序列来对单个珠加标签,该序列对其上的双链寡核苷酸是唯一的,从而允许在使用过程中进行鉴定。

[0076] 预合成的寡核苷酸和 / 或多核苷酸序列可以连接到支持物上或使用下列方法原位合成:光导向方法、流动通道和点样法、喷墨法、销型方法和珠基方法,示于下列参考文献中:McGall 等,(1996)Proc. Natl. Acad. Sci. U. S. A. 93:13555 ;Synthetic DNA Arrays In Genetic Engineering(《在基因工程操作中合成 DNA 阵列》),卷 20:111,普莱南出版社(Plenum Press)(1998);Duggan 等,(1999)Nat. Genet. S21:10 ;Microarrays:Making Them and Using Them In Microarray Bioinformatics(《微阵列:在微阵列生物信息学中制作和使用》),剑桥大学出版社(Cambridge University Press),2003 ;美国专利申请公开号 2003/0068633 和 2002/0081582 ;美国专利号 6,833,450、6,830,890、6,824,866、6,800,439、6,375,903 和 5,700,637 ;和 PCT 公开号 WO 04/031399、WO 04/031351、WO 04/029586、WO 03/100012、WO 03/066212、WO 03/065038、WO 03/064699、WO 03/064027、WO 03/064026、WO 03/046223、WO 03/040410 和 WO 02/24597 ;上述公开内容通过引用全文纳入本文以用于所有目的。在一些实施方式中,预合成的寡核苷酸连接到支持物上或使用点样方法合成,其中通过从一个区域到另一个区域移动的分配器(如喷墨)逐滴加入单体溶液。在一些实施方式中,使用例如机械波驱动的分配器在支持物上点样寡核苷酸。

[0077] 扩增

[0078] 在一些实施方式中,可以使用合适的引物对来扩增寡核苷酸,引物对中的一条引物对应于寡核苷酸的每个末端(例如,一条与寡核苷酸的 3' 末端互补而一条与寡核苷酸的 5' 末端相同)。在一些实施方式中,寡核苷酸可以被设计成含有由 5' 扩增序列(例如,5' 通用序列或 5' 共有扩增序列)和 3' 扩增序列(例如,3' 通用序列或 5' 共有扩增序列)侧接的中心或内部组装序列(对应于目标序列,设计为纳入终产物)。

[0079] 在一些实施方式中,合成寡核苷酸可以包含由 5' 和 3' 扩增序列侧接的中心组装序列。中心组装序列被设计纳入到组装的核酸中。侧接序列被设计用于扩增并且不旨在被纳入到组装的核酸中。可以使用侧接扩增序列作为引物序列来扩增多个有相同扩增序列但是不同中心组装序列的不同组装寡核苷酸。在一些实施方式中,在扩增后去除侧接序列来得到仅含有组装序列的寡核苷酸。

[0080] 可以使用对应于侧接扩增序列的扩增引物(例如,10-50 个核苷酸长,15-45 个核苷酸长,约 25 个核苷酸长等)来扩增寡核苷酸(例如,一条引物可与 3' 扩增序列互补而一条引物具有与 5' 扩增序列相同的序列)。在一些实施方式中,多个具有不同中心组装序列的不同的寡核苷酸(例如,约 5、10、50、100 或更多个)可以具有相同的 5' 扩增序列和 / 或相同的 3' 扩增序列。这些寡核苷酸都可以在相同的反应中使用相同的扩增引物来扩增。然后可以使用任意合适的技术从经扩增的寡核苷酸中去除扩增序列来产生仅含有组装序列的寡核苷酸。在一些实施方式中,通过本文中较详细描述的限制性酶去除扩增序列。

[0081] 在一些实施方式中,可以扩增寡核苷酸,同时其仍然与支持物连接。在一些实施方式中,可以在扩增之前从支持物上移去或切下寡核苷酸。

[0082] 在一些实施方式中,该方法包含在链延伸反应中使用第一组多个单链寡核苷酸作为模板合成多个寡核苷酸或多核苷酸。如前所述,可以首先在表面的离散特征组上合成寡

核苷酸,或可将寡核苷酸置于支持物的多个特征上。在一些实施方式中,寡核苷酸与支持物共价连接。在一些实施方式中,第一组多个寡核苷酸固定在固体表面上。在一些实施方式中,固体表面的各特征包含具有不同预定序列的高密度寡核苷酸(如每个特征约 10^6 - 10^8 个分子)。支持物可以包含至少100、至少1,000、至少 10^4 、至少 10^5 、至少 10^6 、至少 10^7 、至少 10^8 个特征。在一些实施方式中,在扩增后,可在溶液中洗脱双链寡核苷酸和/或对其进行错误减少和/或对其进行组装以形成较长的核酸构建体。

[0083] 错误减少

[0084] 在一些实施方式中,在进行加工以生成粘性末端前,对各片段进行组装和保真度优化以去除含有错误的核酸(例如使用本文所述的一种或多种组装后保真度优化技术)。序列错误可能包含一个或多个核苷酸缺失、插入、取代(例如,颠换或转换)、倒位、重复或其两种或更多种的任意组合。在寡核苷酸合成期间可能产生寡核苷酸错误。不同的合成技术可能易于出现不同的错误分布和频率。在一些实施方式中,根据使用的合成方案,错误率可能在每个碱基1/10-1/200个错误之间变化。然而,在一些实施方式中,可以实现较低的错误率。另外,错误的类型可能取决于使用的合成技术。例如,与基于柱的合成技术相比,基于微阵列的寡核苷酸合成可能产生相对更多的缺失。

[0085] 本发明的一些方面涉及多核苷酸组装方法,其中对合成的寡核苷酸进行设计并用于将多核苷酸组装成为较长的多核苷酸构建体。在酶促扩增或链延伸反应期间,序列中的错误如实地得到复制。结果是,通过该方法合成的多核苷酸群体含有无错误序列和易错序列。在一些实施方式中,由于寡核苷酸合成期间导入的错误,合成的寡核苷酸可含有不正确序列,所以去除在组装或延伸期间整合了一种或多种含有错误的寡核苷酸的多核苷酸是有用的。在一些实施方式中,可对一种或多种组装的多核苷酸进行测序以测定其是否包含预测定的序列。这个方法能鉴定有正确序列的片段。在其他实施方式中,可以使用其他技术以除去包含错误的核酸片段。这类核酸片段可以是初始合成的寡核苷酸或者组装的核酸聚合物。应理解,含有错误的核酸可以是在两条链上都含有错误的同源双链体(即两条链上不正确的互补核苷酸、缺失或添加),因为组装方法可涉及一轮或多轮聚合酶延伸(例如在组装期间或组装后以扩增组装的产物)。聚合酶延伸期间,含有错误的输入核酸可用作模板从而生成包含互补错误的互补链。在某些实施方式中,双链核酸片段或双链体的制备物可能含有核酸混合物,该核酸混合物包含具有正确预定序列的核酸以及含有一种或多种组装期间整合的序列错误的核酸。术语“双链体”指至少部分双链的核酸分子。“稳定的双链体”指在给定的杂交条件组下与互补序列维持杂交状态的可能性相对较大的双链体。在一个示例性实施方式中,稳定的双链体指不含有碱基对错配、插入或缺失的双链体。“不稳定的双链体”指在给定的杂交条件组(如严谨熔融)下与互补序列维持杂交状态的可能性较小的双链体。在一个示例性实施方式中,不稳定的双链体指含有至少一个碱基对错配、插入或缺失的双链体。本文所用术语“严谨性”指温度、离子强度和是否存在其他化合物(如有机溶剂)等的条件,核酸杂交在该条件下进行。在杂交过程期间,杂交严谨性随着温度和/或溶液化学特性(如杂交溶液中盐和/或甲酰胺的量)而增加。在“高严谨性”条件下,核酸碱基配对只发生在具有高频率互补碱基序列的核酸片段之间。可对严谨性条件进行选择使其比限定的离子强度和pH下给定的多核苷酸双链体的热解链点(T_m)低约 5°C 。互补多核苷酸链的长度和GC含量决定了双链体的 T_m ,并因此决定了获得所需杂交特异性所必需

的杂交条件。 T_m 是（在确定离子强度和 pH 的条件下）50% 的多核苷酸序列与完全匹配的互补链杂交时的温度。在某些情况下，可能需要将杂交条件的严谨性提高至约等于特定双链体的 T_m 。合适的严谨性条件是本领域技术人员已知的或者可以由本领域技术人员通过实验来确定。参见例如 *Current Protocols in Molecular Biology*（《新编分子生物学实验指南》），纽约州威立公司（John Wiley&Sons）（1989），6. 3. 1-12. 3. 6；Sambrook 等，1989，*Molecular Cloning, A Laboratory Manual*（《分子克隆：实验室手册》），纽约州冷泉港出版社（Cold Spring Harbor Press）；Agrawal 编，*Methods in Molecular Biology*（《分子生物学方法》），第 20 卷；Tijssen（1993）*Laboratory Techniques in biochemistry and molecular biology-hybridization with nucleic acid probes*（《生物化学和分子生物学实验技术——使用核酸探针进行杂交》），例如第 1 部分第 2 章“Overview of principles of hybridization and the strategy of nucleic acid probe assays（杂交原理概述和核酸探针试验策略）”，纽约爱斯韦尔公司（Elsevier）。

[0086] 在一些实施方式中，可使用涉及使双链核酸变性并再退火的技术除去序列错误。在一些实施方式中，如果含有各单个错误的核酸以比在相同位置具有正确序列的核酸低的频率存在于核酸制备物中，则含有互补错误的单链核酸可能无法再退火并相互结合。相反，含有错误的单链可与无错误的互补链或者含有一个或多个不同错误或在不同位置含有错误的互补链发生再退火。结果是，最终含有错误的链在再退火的反应产物中以异源双链分子的形式存在。无错误的核酸链可与含有错误的链或者其他无错误的链发生再退火。再退火的无错误的链在再退火样品中形成同源双链体。因此，通过从再退火的核酸片段制备物中去除异源双链体分子，可减少含有错误的核酸的量或频率。

[0087] 因此通过被理解为改组的过程形成异源双链体，其中来自不同群体的核酸链可彼此杂交从而形成完全匹配和含有错配的双链体。用于去除异源双链体分子的合适的方法包括色谱、电泳、优先结合在两条链之间具有序列错配的双链 DNA 的异源双链体分子的选择性结合。术语“错配”或“碱基对错配”表示根据沃森和克里克碱基配对规则通常不会形成核酸的碱基对组合。例如，对于 DNA 中常见的碱基（即腺嘌呤、鸟嘌呤、胞嘧啶和胸腺嘧啶），碱基对错配指 DNA 中常见的 A-T 和 G-C 配对以外的那些碱基组合。如本文中所述，错配可以表示为：例如 C/C 表示胞嘧啶残基与另一个胞嘧啶残基相对，这不同于其正确配对的配体鸟嘌呤。

[0088] 在一些实施方式中，可如本文中更详细描述的那样选择或筛选寡核苷酸制备物以除去含有错误的分子。在一些实施方式中，可使用本文所述的错配结合剂对寡核苷酸进行错误校正。

[0089] 在一个方面，本发明涉及在固体支持物上生产高保真多核苷酸的方法。合成的多核苷酸长度为至少约 1、2、3、4、5、8、10、15、20、25、30、40、50、75、或 100 千碱基 (kb)，或 1 兆碱基 (mb)，或更长。在示例性实施方式中，合成的多核苷酸的组合物包含至少约 1%、2%、3%、4%、5%、6%、7%、8%、9%、10%、15%、20%、25%、50%、60%、70%、80%、90%、95% 或更多的无错误拷贝（如具有非衍生自预定序列的序列）。无错误拷贝的百分比是基于与旨在具有正确（如预定或预测的）序列的组合物中的多核苷酸拷贝总数相比，组合物中无错误拷贝的数目。

[0090] 本发明的一些方面涉及用于高保真多核苷酸组装的寡核苷酸设计。本发明的各方

面可用于提高核酸组装过程的生产速率和 / 或减少用于生成正确组装的核酸的步骤数目或试剂量。在某些实施方式中,本发明的各方面可用于自动化核酸组装以减少各正确核酸组装所需的时间、步骤数目、试剂量和其他因素。因此,本发明的这些和其他方面可用于降低一个或多个核酸组装过程的成本和时间。

[0091] 单链突出端

[0092] 在本发明的一些方面中,组装的核酸片段被设计为具有重叠的互补序列。在一些实施方式中,核酸片段是具有 3' 和 / 或 5' 单链突出端的双链 DNA 片段。这些突出端可以是能够与不同核酸片段上的互补粘性末端发生退火的粘性末端。按照本发明的各方面,两个核酸片段上存在的互补序列(和特定互补粘性末端)促进了它们的共价组装。在一些实施方式中,能够组装具有不同重叠互补单链粘性末端的多个核酸片段并且通过在各片段上的粘性末端的相同性确定它们在组装的核酸产物中的顺序。例如,可对核酸片段进行设计使得第一核酸具有与载体的第一粘性末端互补的第一粘性末端以及与第二核酸的第一粘性末端互补的第二粘性末端。第二核酸的第二粘性末端可以与第三核酸的第一粘性末端互补。第三核酸的第二粘性末端可以与第四核酸的第一粘性末端互补。依此类推直至与倒数第二个核酸上的第二粘性末端互补的第一粘性末端的最后一个核酸。

[0093] 在某些实施方式中,将相邻核酸片段之间的重叠互补区域设计(或选择)为具有足够的差异来促进(例如,热力学上有利于)唯一排比的核酸片段(例如,选择或设计的片段排比)的组装。应理解可以使用不同长度的重叠区域。在一些实施方式中,当组装较大数量的核酸片段时,可以使用较长的粘性末端。较长的粘性末端可以为设计或选择充分区别的序列提供更多的灵活性,从而区分正确的粘性末端退火(例如,涉及经设计退火互相结合的粘性末端)与不正确的粘性末端退火(例如,在非互补粘性末端之间)。

[0094] 在一些实施方式中,可以设计或选择两对或更多对不同核酸片段之间的互补粘性末端使其具有相同或相似的序列来促进含有相对随机排列(和 / 或数量)的具有相似或相同粘性末端的片段的产物的组装。这可以用于产生具有某些内部序列区域的不同序列排列和 / 或不同拷贝数量的核酸产物的库。

[0095] 在一些实施方式中,最后的核酸的第二粘性末端与载体的第二粘性末端互补。按照本发明的各方面,可以使用该方法来产生含有以预定线性顺序(例如,第一、第二、第三、第四……最后)组装的核酸片段的载体。在一些实施方式中,可对两个末端核酸片段(例如组装的产物各端处的末端片段)中的每个片段进行设计以使其具有与载体上(例如线性化载体上)的粘性末端互补的粘性末端。这些粘性末端可与能够与线性化载体上相同互补末端序列发生退火的粘性末端相同。然而,在一些实施方式中,末端片段上的粘性末端是不同的且载体含有两个不同的粘性末端,其各自位于线性化载体的一端,各自与末端片段粘性末端之一互补。因此,该载体可以是具有两个粘性末端的线性化质粒,各粘性末端与组装的核酸片段的一端互补。

[0096] 本发明的一些方面包含具有单链突出端的双链核酸。可以使用任意合适的技术来产生突出端。在一些实施方式中,可以用合适的限制性酶来消化双链核酸片段(例如,在多重组装中组装的片段)来产生末端的单链突出端。在一些实施方式中,可以用相同的酶消化被设计为在组装的产物中互相邻接的片段来暴露互补突出端。在一些实施方式中,可以使用 IIS 型限制性酶产生突出端。IIS 型限制性酶是在称为识别位点的位点上与双链核酸

结合的酶,并在识别位点外侧进行单个双链切割。称为剪切位点的双链切割处通常位于距离识别位点 0-20 个碱基处。识别位点长度通常是约 4-7bp。所有的 IIS 型限制性酶都表现出至少部分的不对称识别。不对称识别指核酸的每条链的 5' 3' 识别序列是不同的。酶活性也显示极性,表明剪切位点仅位于识别位点一侧。因此,通常每个识别位点仅对应于一个双链切割。剪切通常在 5' 或 3' 末端产生 1-5 个核苷酸的单链突出端,但是一些酶产生钝末端。两种切割都用于本发明的内容中,虽然在一些情况下产生单链突出端。迄今为止,已经鉴定了大约 80 种 IIS 型酶。示例包括但不限于 BstF5I、BtsC I、BsrD I、Bts I、Alw I、Bcc I、BsmA I、Ear I、Mly I(钝末端)、Ple I、Bmr I、Bsa I、BsmB I、Fau I、Mnl I、Sap I、Bbs I、BciV I、Hph I、Mbo II、BfuA I、BspCN I、BspM I、SfaN I、Hga I、BseR I、Bbv I、Eci I、Fok I、BceA I、BsmF I、BtgZ I、BpuE I、Bsg I、Mme I、BseG I、Bse3D I、BseM I、AcIW I、Alw26I、Bst6I、BstMA I、Eam1104 I、Ksp632I、Pps I、Sch I(钝末端)、Bfi I、Bso31I、BspTN I、Eco31I、Esp3I、Smu I、Bfu I、Bpi I、BpuA I、BstV2I、AsuHP I、Acc36I、Lwe I、Aar I、BseM II、TspDT I、TspGW I、BseX I、BstV1I、Eco57I、Eco57M I、Gsu I 和 Bcg I。这类酶和关于它们的识别位点和剪切位点的信息可从供应商如新英格兰生物实验室公司(New England Biolabs, Inc., 美国马萨诸塞州伊普斯威奇)处得到。

[0097] 在一些实施方式中,可以使用商品化或工程改造的限制性酶。在一些实施方式中,可对 IIS 型限制性酶进行设计和工程改造以产生更长的突出端。对限制性酶进行设计和工程改造以产生更长的单链突出端能够使得更多的寡核苷酸结合在一起以形成更长的核酸构建体。例如,可对产生 4 个核苷酸单链突出端的 BsaI 进行工程改造以产生 5 个或 6 个或更长的单链突出端。通过使用该工程改造的 BsaI 增加单链突出端长度,能够增加可结合的 17 个核酸或寡核苷酸的理论限值。

[0098] 在一些实施方式中,设计用于核酸组装的多个核酸片段中的每个都可以在各末端具有 IIS 型限制性位点。可以导向 IIS 型限制性位点使得剪切位点与识别序列内部相关。结果,酶消化暴露了内部序列(例如,内部序列中的突出端)并且从末端去除识别序列。因此,相同的 IIS 型位点可用于所有经制备用于组装的核酸片段的两个末端和/或可用于线性化合适的载体。然而,也可以使用不同的 IIS 型位点。经设计在组装的产物中邻接的两个片段各自可以包含相同的重叠末端序列和适当位置的侧接 IIS 型位点以在限制性酶消化后暴露重叠序列内的互补突出端。因此,可以用不同的互补突出端来产生多个核酸片段。可以对核酸片段的每个末端上的限制性位点进行定位,使得用合适的 IIS 型酶进行的消化去除了限制性位点并且暴露了与核酸片段上的单链区域互补的单链区域,该核酸片段被设计与组装的核酸产物相邻。在一些实施方式中,可对两个末端核酸片段中每一个的末端进行设计以使其具有与线性化载体核酸的单链突出端互补的单链突出端(例如在使用合适的限制性酶进行消化后)。因此,可将所得的核酸片段和载体直接转化至宿主细胞内。或者,可将核酸片段和载体进行孵育以在转化进宿主细胞前促进互补序列的杂交和退火。应理解,可使用本文所述任意技术或者产生与末端核酸片段之一的一端互补的单链突出端的任意其他适当技术来制备载体。

[0099] 使用 IIS 型或位点特异性限制性酶的 DNA 酶消化通常产生 4 至 6 个核苷酸的突出端。这些短粘性末端足以连接两个含有互补末端的核酸片段。然而,当多个核酸片段连接在一起时,优选较长的互补粘性末端以促进组装并确保特异性。例如,粘性末端可以足够长以

具有足够不同的序列以防止或减少相似粘性末端之间的错配。然而,优选其长度不会长到稳定相似粘性序列之间的错配。在一些实施方式中,可以使用约 9 至约 15 个碱基的长度。然而,对于用于产生粘性突出端的区域,可以选择任意合适的长度。特异性的的重要性可以取决于同时组装的不同片段的数量。另外,为避免使错配的区域稳定所需要的合适的长度可取决于用于退火不同粘性末端的条件。

[0100] 基于连接酶的组装

[0101] 基于连接酶的组装技术可以涉及一种或多种合适的连接酶,所述连接酶能催化邻近 3' 和 5' 核酸末端的共价连接(如核酸的 5' 磷酸和 3' 羟基在互补模板核酸上退火从而使 3' 末端紧邻 5' 末端)。因此,连接酶可以催化第一核酸的 5' 磷酸和第二核酸的 3' 羟基之间的连接反应,前提是第一和第二核酸在模板核酸上彼此相邻发生退火,连接酶可以获自重组或天然来源。连接酶可以是热稳定的连接酶。在一些实施方式中,可以使用来自嗜热微生物的热稳定连接酶。热稳定 DNA 连接酶的示例包括但不限于:Tth DNA 连接酶(来自嗜热栖热菌 (*Thermus thermophilus*)),可购自例如友基公司 (Eurogentec) 和基因工艺公司 (GeneCraft)、Pfu DNA 连接酶(来自激烈火球菌 (*Pyrococcus furiosus*))的超嗜热连接酶)、Taq 连接酶(来自水生栖热菌 (*Thermus aquaticus*)),任何其他合适的热稳定连接酶,或其任意组合。在一些实施方式中,可以使用一种或多种较低温度的连接酶(如 T4DNA 连接酶)。较低温度的连接酶可以用于在较高温度下可能不稳定的更短的突出端(如约 3、约 4、约 5 或约 6 个碱基的突出端)。

[0102] 在一些实施方式中,可对连接酶进行设计和工程改造以使其具有更大程度的特异性以最小化形成的不需要的连接产物。在一些实施方式中,连接酶可与蛋白联用或者可与能够促进连接酶与核酸分子之间相互作用和/或增加连接特异性的蛋白融合。

[0103] 非酶技术能用于连接核酸。例如,一个或多个核酸的 5' 末端(如 5' 磷酸基团)和 3' 末端(如 3' 羟基)可以不用酶(如不使用连接酶)而共价连接在一起。在一些实施方式中,非酶技术可以提供超过酶连接的某些优势。例如,非酶技术可以对核酸底物中非天然核苷酸类似物具有高耐受性,可以用于连接短的核酸底物,可以用于连接 RNA 底物,和/或对某些自动(如高通量)应用而言更便宜和/或更合适。

[0104] 因此,化学连接可以用于形成第一核酸末端的 5' 末端和第二核酸末端的 3' 末端的共价连接,其中所述第一和第二核酸末端可以是单个核酸末端或单独的核酸末端。一方面,化学连接可以涉及具有修饰的末端(例如修饰的 5' 和/或 3' 末端)的至少一个核酸底物,该修饰的末端包括帮助或促进连接形成的一种或多种化学反应部分。在一些实施方式中,当使得一个或多个核酸末端紧密相邻时(例如由于互补核酸序列之间的退火使得末端结合在一起时),发生化学连接。因此,互补 3' 或 5' 突出端(如由双链核酸的限制性酶剪切生成的突出端)之间或者引起 3' 末端与 5' 末端紧密相邻的任意互补核酸组合之间(如当核酸与互补模板核酸发生退火时,3' 和 5' 末端彼此相邻)的退火可以促进模板定向的化学连接。化学反应的示例可以包含但不限于缩合、还原、和/或光-化学连接反应。应理解在一些实施方式中,化学连接可用于生成天然产生的磷酸二酯核苷酸间连接、非天然产生的磷酸胺焦磷酸核苷酸间连接和/或其他非天然产生的核苷酸间连接。

[0105] 同时酶促去除共有寡核苷酸序列并将经加工的寡核苷酸连接成更长的构建体

[0106] 图 2 显示了根据本发明的一个实施方式用于组装核酸的方法。在一些实施方式

中,该方法包括同时酶促去除共有寡核苷酸序列并将经加工的寡核苷酸序列连接成更长的构建体。在一些实施方式中,如本文所述通过 PCR 扩增寡核苷酸并对其进行错误校正。通过合适的限制性内切酶 (40) 对由共有启动 (扩增) 序列 (20) 和构建体特异性载荷 (payload) 或内部序列区域 (30) 组成的经扩增的寡核苷酸 (10) 进行加工。在一些实施方式中,第一和最后的寡核苷酸含有用于扩增目标构建体的独特的启动序列 (25)。限制性内切酶催化所有寡核苷酸 (50) 共有的末端共有区域 (本文中称作扩增区域或引物识别区域) 的剪切,留下具有末端单链 DNA 序列 (60) 的内部区域 (本文中称作游离载荷 (free payload))。在一些实施方式中,该限制性内切酶是 IIS 型限制性内切酶。对这些单链序列进行设计以指导一个寡核苷酸与另一个的特异性相互作用,使得多个寡核苷酸线性排列成预定序列 (70)。因此,末端单链 DNA 序列可将寡核苷酸的适当相互作用指导为正确的顺序,从而连接酶 (80) 催化单个寡核苷酸的连接,生成最后的目标核酸构建体 (90) 或中间核酸构建体。

[0107] 本领域技术人员应理解,如果初始的共有序列向后连接在一起 (例如使用与 (60) 互补的末端序列的 (50)), 则限制性内切酶的存在可确保能够再次对其进行切割以生成游离末端 (60)。然而,由于限制性内切酶的选择,适当连接的连接部 (例如 1' 和 2' 之间) 不会被识别为限制性位点且不会是解开的。天然状态下反应应朝所需产物 (90) 的方向进行。

[0108] 在一些实施方式中,对于一种识别用于去除共有序列的限制性位点的方法的变体现在也可以是待合成基因的一部分。该约束性去除允许基因合成方法的递推 (分级) 应用以构建较长的核酸序列 (如图 4 所示)。先前的方法中,在以独立的步骤进行去除和连接时,由于在去除和连接步骤之间必须有部分基于大小选择的纯化步骤,该设计是不被允许的。在这类方法中,纯化过程中可能丢失所需目标序列的切割碎片,导致无法构建所需目标序列。在一些实施方式中,通过同时使用本发明的去除和连接步骤,可以一致地切割那些切割序列并进行重新连接,使得一些感兴趣的目标序列得以存在。在一些实施方式中,所需序列的量取决于对限制性酶和连接酶的相对活性的调节。

[0109] 如图 4 所示,可由寡核苷酸组 (310) 和 (311) 组装基因合成碎片 (390) 和 (391)。可对寡核苷酸组进行设计使其具有匹配的限制性内切酶位点 (340), 从而能够使用相同的消化和连接步骤连接基因合成碎片 (390) 和 (391) (随后进行扩增)。在一些实施方式中,可对第二轮进行设计使其具有使用第二限制性酶的限制性内切酶位点 (340)。然而,由于在该方法中使用多种酶的复杂性,可能无法进行上述操作。此外,在不同步进行消化和连接的情况下,使用两种限制性酶会导致不允许目标序列具有两个限制性酶位点,这将进一步约束能够被合成的基因。

[0110] 仍然根据图 4,可使用引物 (325) 扩增核酸片段 (390), 并可使用引物 (326) 扩增核酸片段 (391)。随后将核酸片段混合在一起并以类似于先前合成步骤的方式进行加工以生成合并的核酸片段 (392), 其中限制性位点 (340) 以与先前轮次中的位点 (350) 类似的方式发挥作用。可使用来自 (325) 的 5' 引物和来自 (326) 的 3' 引物扩增合并的目标序列 (392)。

[0111] 在一些实施方式中,可根据本文所公开的方法使用分级组装策略。本领域技术人员应理解,本方法可规模化至多种核酸片段,使得随后轮次中的核酸片段数目能够与第一轮中的核酸片段数目类似。分级组装方法可以是几何的,从而在较少轮次中构建非常大的目标序列。例如,可由池 (310) 或 (311) 之一构建 1000 碱基 (1kbp) 的目标序列。第二轮

的类似于 (390) 或 (391) 的 10 个核酸片段会形成 10kbp 碱基的目标核酸序列。使用 10kbp 核酸序列的第三轮会形成 100kbp 的目标核酸序列,其来源于初始的 100 源池。

[0112] 在一些实施方式中,可在单独的池中进行多个组装反应。随后可混合来自组装反应的组装构建体以形成更长的核酸序列。在一些实施方式中,可使用限制性内切酶进行分级组装以形成粘性末端,该粘性末端可以所需顺序结合在一起。可设计并合成构建寡核苷酸以使其含有识别和剪切位点,用于使一种或多种限制性内切酶在位点处促进特定顺序的连接。在一些实施方式中,可将一种或多种 IIS 型内切酶识别位点整合至构建寡核苷酸的末端以允许 IIS 型限制性内切酶进行剪切。可通过互补粘性末端的杂交确定连接的顺序。

[0113] 在一些实施方式中,第一寡核苷酸池包含经设计在其 3' 端具有额外的限制性酶识别位点的 3' 端寡核苷酸且第二寡核苷酸池包含经设计在其 5' 端具有额外的限制性酶识别位点的 5' 端寡核苷酸。在一些实施方式中,限制性酶是相同的。在各池中进行寡核苷酸组装后,可根据本文所述方法使用限制性内切酶和连接酶处理两种亚组装构建体。

[0114] 本领域技术人员应理解,可通过本发明的各方面显著(几何地)提高合成的可用组装空间。先前,为生成两倍序列数目(2n)的构建体,需要使寡核苷酸数目翻倍。例如,为生成构建体(390),需要使构建体(310)的数目翻倍,因此需要使相容的单链末端(360)的数目翻倍。使用如图4所示的方法,用于(310)和(311)的连接部仅需与连接部(340)相容,从而确保大小翻倍的核酸的组装过程中仅使用一个额外的连接部。因此,如果寡核苷酸(310)和(311)具有干扰或不相容的末端,仍可通过本文所述方法(消化(340)和连接)将其连接在一起以制备目标核酸(392),而仅将寡核苷酸池(310)和(311)混合在一起是无法实现连接的。

[0115] 图3显示了一种变体形式,其中同时进行寡核苷酸加工和组装成目标构建体并同时进入质粒。质粒 pG9-1 (SEQ ID NO. 1) 的详细信息如图5所示。该质粒含有允许限制性内切酶(在该示例中是 BsaI) 在两个位点处切割质粒的限制性内切酶识别位点(带下划线的文本,图5),得到限定的单链序列(图5-反向文本)。根据图3,质粒(100)(例如 pG9-1) 被导入包含寡核苷酸(110)混合物的池中,该寡核苷酸(110)已如本文所述进行扩增和错误校正。在一些实施方式中,这些寡核苷酸序列(110)可具有被特定的限制性内切酶(140)识别的共有序列(120)。在一些实施方式中,质粒(130)可具有被相同限制性内切酶(140)识别的序列。限制性内切酶(140)对这些序列的作用导致从寡核苷酸((310), (311))和质粒(150)中去除共有序列,暴露单链 DNA 序列(160)。在一些实施方式中,所述限制性酶可以是 IIS 型限制性酶。在一些实施方式中,对所述单链序列进行设计以指导一个寡核苷酸与另一个的特定相互作用,使得多个寡核苷酸排列成限定的序列并使该排列的寡核苷酸序列(170)进入质粒(100)。在一些实施方式中,连接酶(180)催化单个寡核苷酸的共价连接。终产物是含有来源于连接寡核苷酸(190)的特定构建体的质粒(例如 pG9-1)。随后可将该质粒(190)转化进细菌并进行测序验证。

[0116] 本发明的各方面涉及根据本发明的方法对组装的构建体进行测序验证。构建体的测序验证如图6所示。在该方法中,可如图3所示产生多个构建体(200, C1至C4)并将其转化至细菌中。可使用适当的用于选择的抗生素抗性在固体生长平板上选择含有质粒 DNA 的细菌转化子。生长后,可挑取单克隆并置于池中,各构建体平板(220)取一个,生成构建体池,各池含有各构建体的一个拷贝。在一些实施方式中,池的数量可取决于待测序以鉴定

具有完美序列的构建体的单个构建体的数量。如图 6 所示,生成了 4 个构建体的 4 个池,允许分析各构建体的 4 个成员。随后可从收集的材料 (230) 中制备质粒 DNA。随后可对各质粒 DNA 分子池进行准备以进行测序。该准备可使用多种方法之一,该方法导致 DNA 断裂成小片段并使用测序 (例如下一代高通量测序) 所需的共有序列进行连接。这些共有序列中含有短的 DNA 碎片,所述 DNA 碎片对生成的 4 个池中的每个都是独特的。这些独特的 DNA 碎片可用于鉴定各测序的构建体是来源于哪个池。随后可通过回到初始的细菌生长平板并再次生长相应的含有具有所需构建体的质粒的克隆来恢复具有正确序列的构建体。

[0117] 载体和宿主细胞

[0118] 可使用任何合适的载体,本发明不限于此。例如,载体可以是质粒、细菌载体、病毒载体、噬菌体载体、昆虫载体、酵母载体、哺乳动物载体、BAC、YAC 或任何其他合适的载体。在一些实施方式中,载体可以是仅在一类生物体 (例如细菌、酵母、昆虫、哺乳动物等) 内或仅在一种生物体内复制的载体。一些载体可具有广泛的宿主范围。一些载体可具有在不同生物体中发挥功能的不同的功能序列 (例如起点或复制、可选择标记等)。这些可用于在两种不同类型的生物体之间 (例如细菌和哺乳动物之间、酵母和哺乳动物之间等) 运输载体 (以及任何克隆至载体中的核酸片段)。在一些实施方式中,可通过所选宿主细胞类型确定所用载体类型。

[0119] 应理解,载体可编码可检测标记 (如可选择标记,例如抗生素抗性等),使得转化的细胞能够选择性地生长并可分离载体并可表征任何插入以确定其是否含有所需的组装的核酸。可使用任何合适的技术 (例如大小分析、限制性片段分析、测序等) 表征该插入。在一些实施方式中,可通过确定宿主细胞中是否表达预测由正确组装的核酸编码的功能来测定载体中是否存在正确组装的核酸。

[0120] 在一些实施方式中,可通过使用一种或多种其他可检测或可选择标记选择或富集包含含有核酸插入的载体的宿主细胞,所述标记仅当正确的 (例如经设计的) 末端核酸片段被克隆至载体中时发挥功能。

[0121] 因此,宿主细胞应具有适当的表型以选择一种或多种载体上编码的抗药标记 (或者检测一种或多种载体上编码的可检测标记)。然而,任何合适的宿主细胞类型都可以使用 (例如原核、真核、细菌、酵母、昆虫、哺乳动物等)。例如,宿主细胞可以是细菌细胞 (例如大肠杆菌 (*Escherichia coli*)、枯草芽孢杆菌 (*Bacillus subtilis*)、结核分枝杆菌属 (*Mycobacterium* spp.)、结核分枝杆菌 (*M. tuberculosis*) 或其他合适的细菌细胞)、酵母细胞 (例如酵母菌属 (*Saccharomyces* spp.)、毕赤酵母菌属 (*Picchia* spp.)、假丝酵母菌属 (*Candida* spp.) 或其他合适的酵母物种,如酿酒酵母 (*S. cerevisiae*)、白色念珠菌 (*C. albicans*)、粟酒裂殖酵母 (*S. pombe*) 等)、爪蟾细胞、小鼠细胞、猴细胞、人细胞、昆虫细胞 (例如 SF9 细胞和果蝇细胞)、蠕虫细胞 (例如隐杆线虫属 (*Caenorhabditis* spp.))、植物细胞或其他合适的细胞 (包括例如转基因或其他重组细胞系)。此外,可使用多种异源细胞系,如中国仓鼠卵巢细胞 (CHO)。

[0122] 应用

[0123] 本发明的各方面可用于多种应用,涉及合成的核酸的生成和 / 或使用。如本文所述,本发明提供了以提高的效率组装合成的核酸的方法。可对所得组装的核酸进行体外扩增 (如使用 PCR、LCR 或任何合适的扩增技术)、体内扩增 (如通过克隆到合适的载体中)、

分离和 / 或纯化。可将 (单独或克隆到载体中的) 组装的核酸转化入宿主细胞 (如原核、真核、昆虫、哺乳动物或其他宿主细胞)。在一些实施方式中, 该宿主细胞可以用于增殖该核酸。在某些实施方式中, 该核酸可以整合到该宿主细胞的基因组中。在一些实施方式中, 该核酸可以取代细胞基因组上的对应核酸区域 (如通过同源重组)。因此, 核酸可以用于生成重组生物体。在一些实施方式中, 目标核酸可以是整个基因组或用于取代全部或部分宿主生物体基因组的基因组大片段。重组生物体也可用于多种研究、工业、农业和 / 或医学应用。

[0124] 本文所述的许多技术可以一起使用, 在一个或多个点上应用一种或多种基于延伸和 / 或基于连接的组装技术的组合以生成核酸分子。例如, 可以使用协同组装来组装寡核苷酸双链体和低于 100 至超过 10000 个碱基对长度的核酸片段 (例如, 100 聚体至 500 聚体、500 聚体至 1000 聚体、1000 聚体至 5000 聚体、5000 聚体至 10000 聚体、25000 聚体、50000 聚体、75000 聚体、100000 聚体等)。在一个示例性实施方式中, 本文所述方法可以在组装生物体 (如病毒、细菌、酵母或其他原核或真核生物) 的整个基因组 (或其大片段, 如约 10%、20%、30%、40%、50%、60%、70%、80%、90% 或更多) 的过程中使用, 可选在一个或多个所需位置处将特异性修饰整合到序列中。

[0125] 可将使用本发明的方法生成的核酸分子整合到载体中。该载体可以是克隆载体或表达载体。载体可包含一个复制起始和一个或多个可选择标记 (例如抗生素抗性标记、营养缺陷标记等)。在一些实施方式中, 该载体可以是病毒载体。病毒载体可包含能够感染目标细胞的核酸序列。类似地, 在一些实施方式中, 可操作地与合适的启动子系统连接的原核表达载体可用于转化目标细胞。在其他实施方式中, 可操作地与合适的启动子系统连接的真核载体可以用于转染目标细胞或组织。

[0126] 本文所述的构建体的转录和 / 或翻译可以在体外 (即使用不含细胞的系统) 或体内 (即在细胞中表达) 进行。在一些实施方式中, 可以制备细胞裂解物。在某些实施方式中, 可以分离或纯化表达的 RNA 或多肽。

[0127] 本文提供的方法和设备的各方面可以包括自动进行本文所述的一个或多个操作。在一些实施方式中, 扩增和 / 或组装反应中的一个或多个步骤可以使用一个或多个自动化样品处理装置 (如一个或多个自动化液体或流体处理装置) 来自动操作。自动化装置和方法可用于递送反应试剂, 所述反应试剂包括下列一种或多种: 起始核酸、缓冲液、酶 (如一种或多种连接酶和 / 或聚合酶)、核苷酸、盐和任何其他合适试剂 (如稳定剂)。自动化装置和方法也可用于控制反应条件。例如, 自动化热循环仪可用于控制反应温度和可用的任何温度循环。在一些实施方式中, 可对扫描激光器进行自动化以提供适于孵育多核苷酸的一个或多个反应温度或温度循环。类似地, 组装的多核苷酸产物的后续分析可以自动进行。例如, 可以使用测序装置和自动化测序方案自动进行测序。也可以使用一种或多种合适装置和相关方案自动进行其他步骤 (如扩增、克隆等)。应理解, 本文所述的一种或多种装置或装置组件可以在某一系统 (如机器人系统) 或微环境 (如微流体反应室) 中组合。使用自动化装置和方法 (如样品和 / 或样品容器的机械化操作和 / 或转移, 包含自动化移液装置、微系统等) 可将组装反应混合物 (如液体反应样品) 从系统的一个组件转移至另一个组件。该系统及其任何组件可以通过控制系统来控制。

[0128] 因此, 本文所提供装置的方法步骤和 / 或方面可以使用例如计算机系统 (如计算

机控制系统)自动进行。能实施本文所提供技术方面的计算机系统可以包括用于任何处理类型的计算机(如本文所述序列分析和/或自动化装置控制)。然而,应理解某些处理步骤可以通过作为组装系统的一部分的一种或多种自动化装置来提供。在一些实施方式中,计算机系统可包括两台或更多台计算机。例如,一台计算机可以通过网络与第二台计算机偶联。一台计算机可进行序列分析。第二台计算机可以控制系统中的一个或多个自动化合成和组装装置。在其他方面中,其他计算机可以包含在网络中以控制一个或多个分析或处理操作。各计算机可以包含内存和处理器。该计算机可采用任何形式,因为本文提供的技术方面并不限于在任何特定计算机平台上实施。类似地,该网络可采用任何形式,包含专用网络或公共网络(如互联网)。显示设备能与一个或多个装置和计算机连接。或者,或另外,显示设备可以位于远端位点,并且根据本文提供的技术连接以显示分析输出。该系统的不同组件之间的连接可以通过有线、光纤、无线传送,卫星传送,任何其他合适的传送,或者上述两种或多种的任意组合。

[0129] 本文所提供技术的各个不同方面、实施方式、或操作能以任意的多种方式独立自动进行和实施。例如,各个方面、实施方式或操作能使用硬件、软件或其组合独立实施。当以软件实施时,能在任何合适的处理器或处理器集合上执行软件代码,该处理器集合在单独计算机中提供或分布在多个计算机上。应理解,通常可将完成上述功能的任何组件或组件集合视作控制上文所讨论功能的一个或多个控制器。所述一个或多个控制器能以多种方式实施,例如使用微码或软件程序控制的专用硬件或通用目的硬件(如一个或多个处理器)以完成上述功能。

[0130] 在这方面,应理解,本文所提供技术的实施方式的一种实现中包含了用计算机程序(如多种指令)编码的至少一种计算机可读介质(如计算机内存、软盘、光盘、磁带等),其能够在处理器上运行时完成本文所提供技术的一种或多种上述功能。该计算机可读介质可运输,从而其上存储的程序能加载到任何计算机系统来源以运行本文所提供技术的一种或多种功能。此外,应理解,执行时完成上文所讨论功能的计算机程序不限于在主机上运行的应用程序。相反,术语计算机程序在本文以一般意义使用,指任何类型的计算机代码(如软件或微码),该代码能用于编程处理器以进行上文讨论的本文所提供技术方面。

[0131] 应理解,与处理器存储于计算机可读介质上的数个本文所提供技术的实施方式一致,计算机实施的处理在其执行过程中可以接受手工输入(如来自用户)。

[0132] 因此,本文所述组装装置或组件的整体系统水平控制可以通过系统控制器进行,该系统控制器可以向下列装置提供控制信号:相关的核酸合成器、液体处理装置、热循环仪、测序装置、相关的机械化组件,以及用于运行所需输入/输出或其他控制功能的其他合适系统。因此,该系统控制器与任何装置控制器一起形成控制核酸组装系统运行的控制器。该控制器可以包含通用目的数据处理系统和其他相关装置,该通用目的数据处理系统能是通用目的计算机或通用目的计算机的网络,所述其他相关设备包含通信装置、调制解调器、和/或其他回路或组件,以进行所需输入/输出或其他功能。该控制器也能至少部分作为单个特定目的集成电路(如ASIC)或ASIC阵列实施,各自具有用于整体、系统水平控制的主要或中央处理器部分以及专用的分离部分以在中央处理器部分控制下进行多种不同特定计算、功能和其他处理。该控制器也能使用多种分离的专用程序集成或其他电子回路或装置实施,例如硬连线电子装置或逻辑回路如分开的元件电路或程序逻辑装置。该控制器也

能包含任何其他组件或装置,如用户输入/输出装置(监控器、显示器、打印机、键盘、用户点击装置、触摸屏、或其他用户界面等)、数据存储装置、驱动马达、连接、阀控制器、机械化装置、真空和其他泵、压力传感器、检测器、电源供应器、脉冲源、通信装置或其他电子电路或组件等。该控制器也可以控制系统其他部分的运行,如自动化客户订单处理、质量控制、包装、运输、开票等,以进行本领域已知而本文没有详述的其他合适功能。

[0133] 本发明的多个方面可以单独使用、联用或以前述实施方式未特定讨论的各种排列来使用,并且因此并不限制其应用于前面描述或附图说明所示组件的细节和排列。例如,一个实施方式中描述的各方面可与其他实施方式中描述的各方面以任何方式组合。

[0134] 权利要求中修饰权利要求元件使用的顺序术语“第一”、“第二”、“第三”等本身并不暗指任何优先、级别高低或一个权利要求项高于另一个的顺序或实行方法作用的暂时顺序,而是仅仅用作标记把有某一名称的一个权利要求元件与有相同名称的另一元件(但是就顺序术语使用而言)区分开以区别权利要求元件。

[0135] 而且,本文所用的词语和术语是为了描述目的,而不是限制性的。本文使用“包含”、“包括”、或“具有”、“含有”、“涉及”及其变体表示涵盖其后列出的项目及其等价物,以及其他项目。

[0136] 等同形式

[0137] 本发明提供了用于高保真基因组装的新方法和装置等。尽管讨论了本发明的具体实施方式,但以上说明书仅为说明性而非限制性的。本领域的技术人员在阅读本说明书后将清楚了解本发明的许多变化。本发明的全部范围应该通过参考所附权利要求书连同其等同物的全部范围,以及说明书连同此类变化来确定。

[0138] 通过引用纳入

[0139] 参考2013年4月24日提交的美国申请13/986,368、2012年6月15日提交的美国申请13/524,164和PCT公开PCT/US2009/055267。本文提到的所有发表物、专利、专利申请和序列数据库条目在此通过引用全文纳入,就好像各个单独发表物或专利特定和单独地表明通过引用纳入。

[0001]

序列表

- <110> GEN9股份有限公司 (Gen9, Inc.)
M·E·赫德森 (Hudson, Michael E.)
L·Y·A·昆 (Kung, Li-yun A.)
D·辛德勒 (Schindler, Daniel)
S·阿彻 (Archer, Stephen)
I·萨奥尔默 (Saaem, Ishtiaq)

<120> 用于核酸组装和高通量测序的方法

<130> 127662-013402

<140> 未编号

<141> 2013-06-24

<150> 61/664, 118

<151> 2012-06-25

<150> 61/731, 627

<151> 2012-11-30

<160> 1

<170> PatentIn version 3.5

<210> 1

<211> 2608

<212> DNA

<213> 人工序列

<220>

<223> 合成的构建体

<400> 1

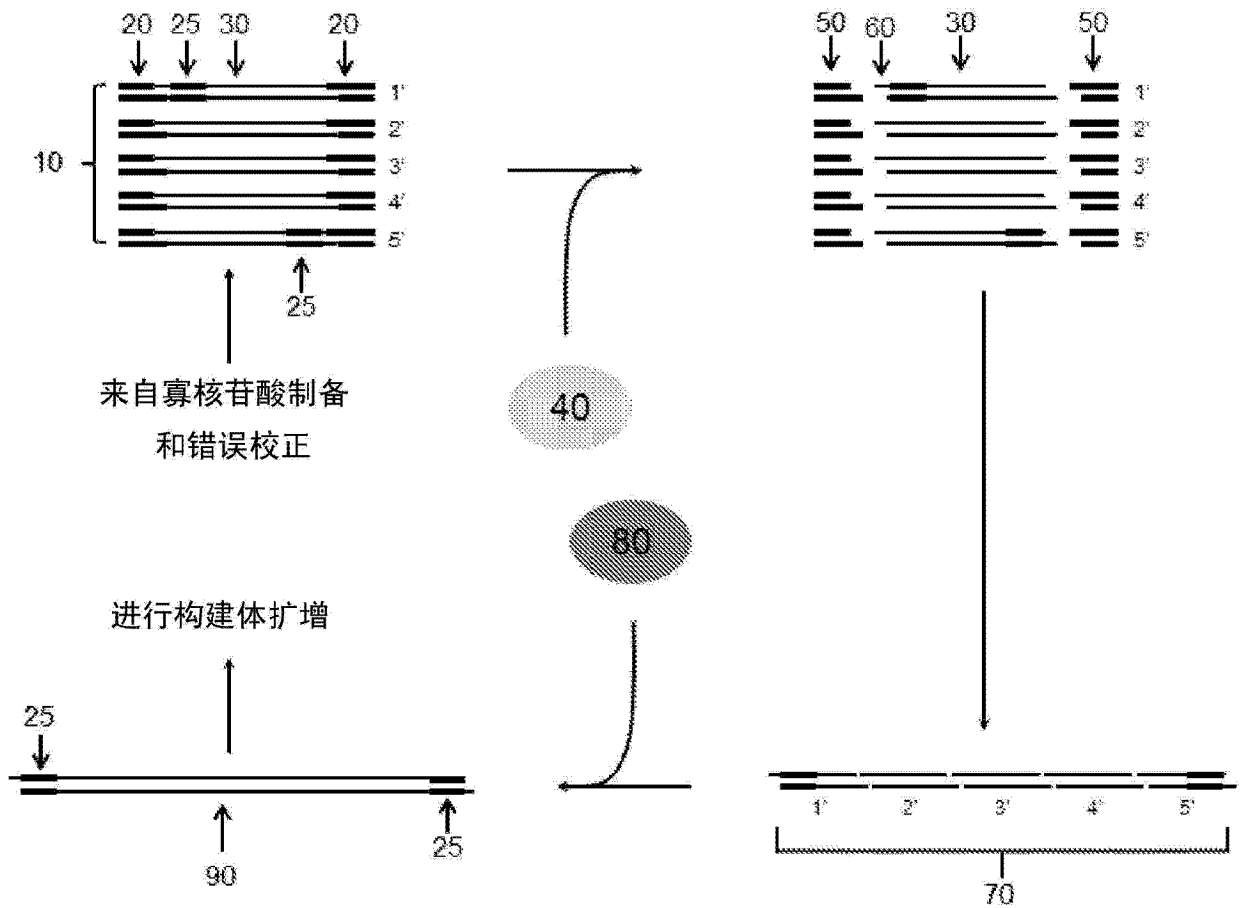
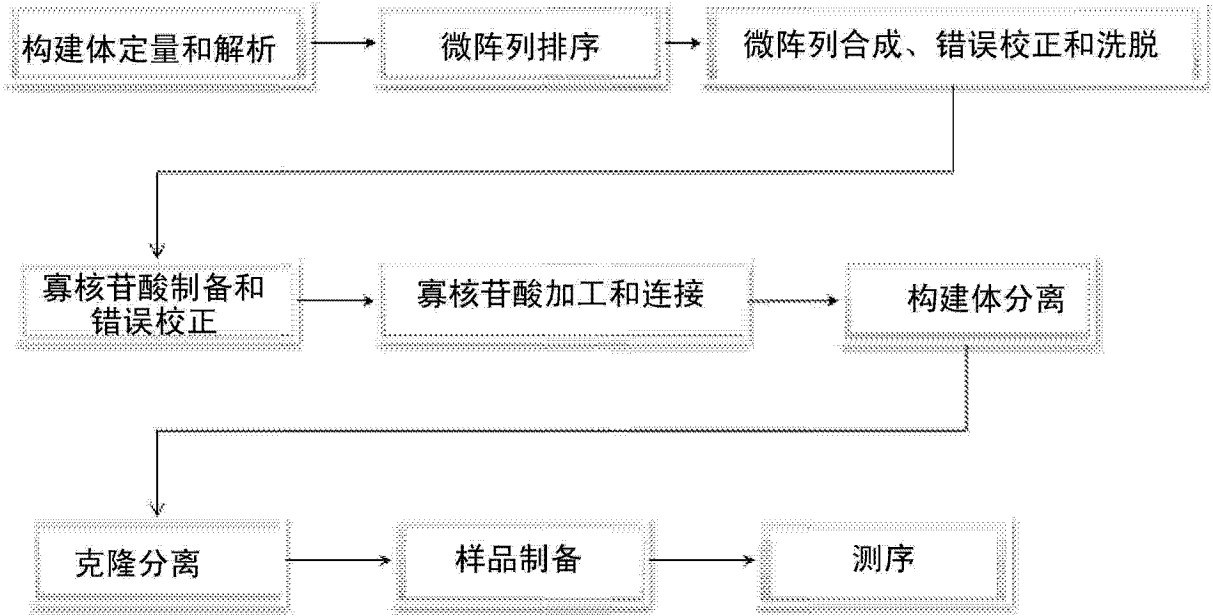
```

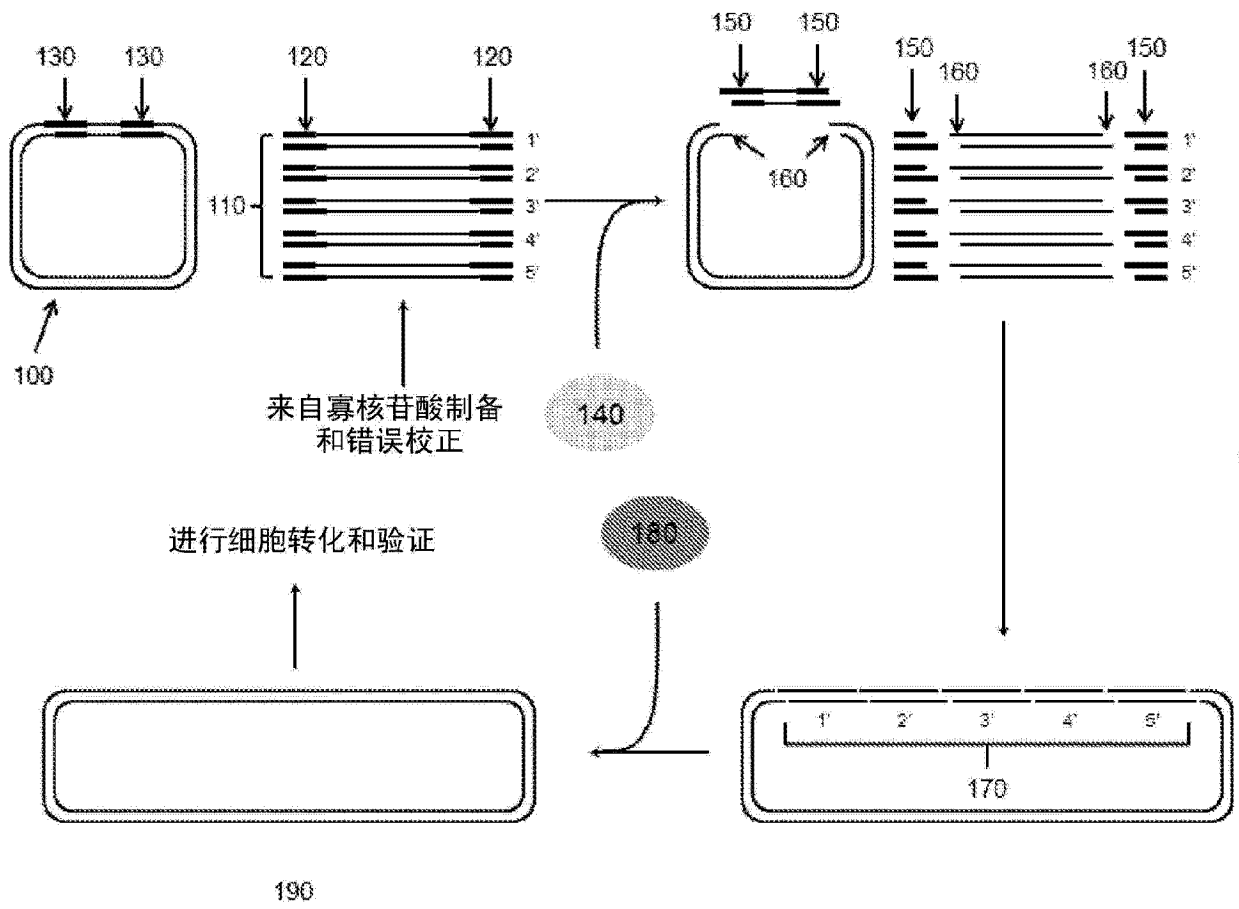
tcgcgcgttt cggtgatgac ggtgaaaacc tctgacacat gcagctccc gagacggcca      60
cagcttgtct gtaagcggat gccgggagca gacaagccc tcagggcgcg tcagcgggtg      120
ttggcgggtg tcgggctgg cttaactatg cggcatcaga gcagattgta ctgagagtgc      180
accatattgc gtgtgaaata ccgcacagat gcgtaaggag aaaataccgc atcagggccc      240
attgccatt caggetgcgc aactgltggg aaggcgatc ggtgcgggce tcttcgctat      300
tacgccagct gccgaaagg gtagtgctg caaggcgatt aagttgggta acgccagggt      360
tttcccagtc acgacgttgt aaaacgacgg ccagtgtaatt agtgttgaga ccattcagct      420
ccggtctcga cactgagctt ggcgtaatca tggatcagc tgtttcctgt gtgaaattgt      480
tatccgctca caattccaca caacatacga gccggaagca taaagtgtaa agcctggggt      540
gcctaattgag tgagctaaact cacattaatt gcgttgccct cactgcccgc tttccagtcg      600
ggaaacctgt cgtgccagct gcattaatga atcgccaac gcccggggag aggcggtttg      660
cgtattgggc getcttcgc ttcctcctc actgaactgc tgcctcggg cgttcggctg      720
cggcgagcgg tatcagctca ctcaaaggcg gtaatacggc tatccacaga atcaggggat      780
ancgcaggaa agaacatgtg agcaaaaggc cagcaaaagg ccaggaaacc taaaaaggcc      840
gcgttgctgg cgtttttcca taggctccgc ccccctgacg agcatcaca aaatcgacgc      900
tcaagtcaga ggtggcgaaa ccgacagga ctataaagat accaggcgtt tccccctgga      960
agctcccteg tgcgctctcc tgttcgacc ctgcccctta ccggatacct gtcccctttt      1020
ctcccttggg gaagcgtggc gctttctcat agctcagctt gtaggtatct cagttcgggt      1080
taggtcgttc getccaagct gggctgtgtg caagaacccc ccgttcagcc cgaccgctgc      1140
gccttatccg gtaactatcg tcttagctcc aaccggtaa gacacactt atcgccaactg      1200
gcagcagcca ctggtaacag gattagcaga gcgaggtatg taggcgggtc tacagagttc      1260

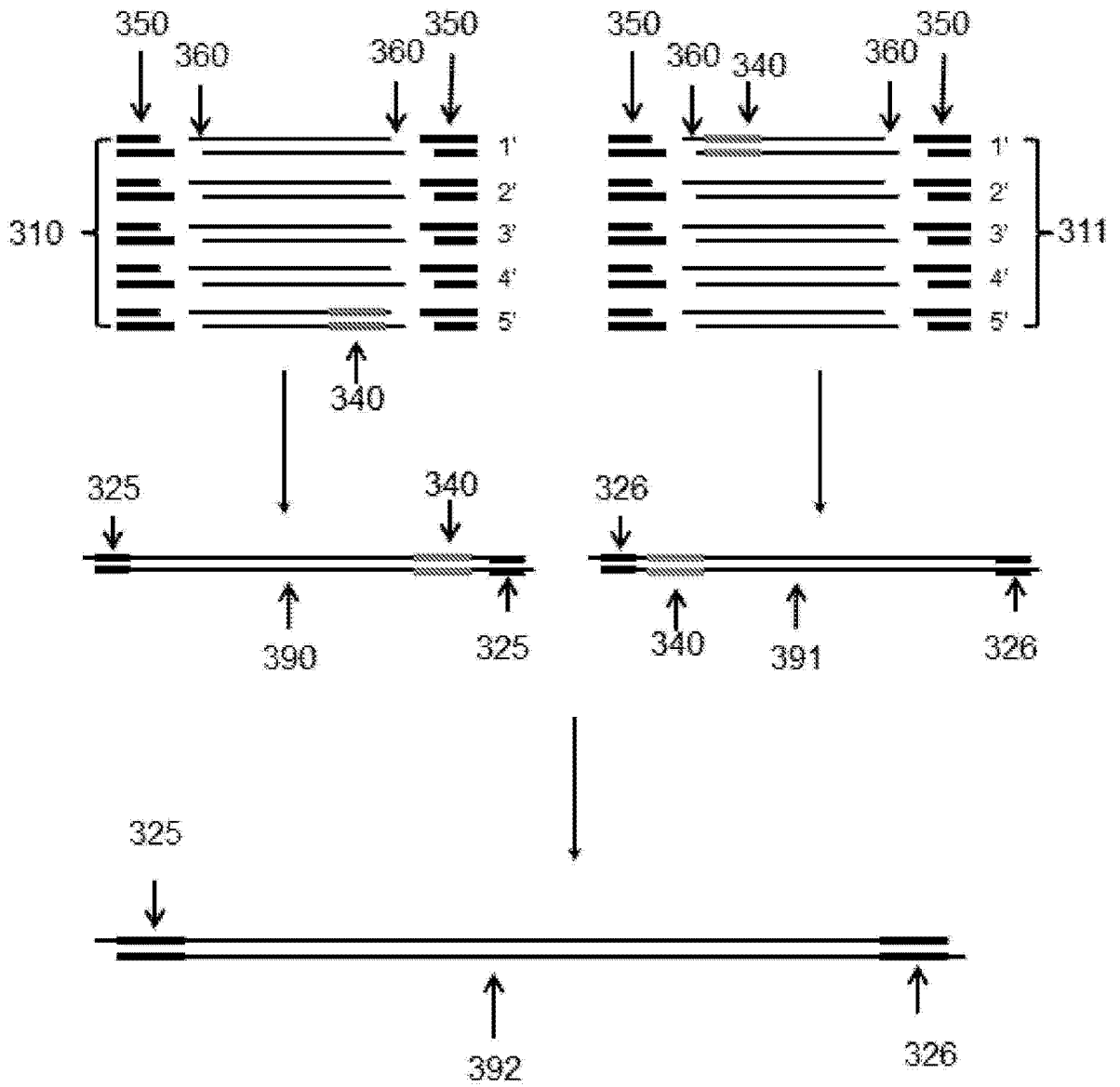
```

[0002]

ttgaagtgg	ggcctaacta	cgctacaact	agaagaacag	tatttggat	ctgcgctctg	1320
ctgaagccag	ttaccttcgg	aaaaagagtt	gtagctctt	gatccgcaa	acaaaccacc	1380
gctggtagcg	gtggtttttt	fgtttgcaag	cagcagatta	cgcgagaaa	aaaaggatct	1440
caagaagate	ctttgatctt	ttctacgggg	ctgacgcte	agtggaacga	aaactcacgt	1500
taagggat	ttgcatgag	attatcaaaa	aggatcttca	cctagatcct	tttaaatata	1560
aaatgaagtt	ttaatcaat	ctaaagtata	tatgagtaaa	cttggcttga	cagtcagaag	1620
aaactgtaaa	gaaggcgata	gaaggcgatg	cgctgcgaat	egggagcggc	gataccgtaa	1680
agcagcagga	agcggtcagc	ccattcggcc	ccaagctctt	cagcaatata	acgggtagcc	1740
aacgctatgt	ccgatagcgc	gtccgcaca	eccagcggc	cacagtcgat	gaatecagaa	1800
aagcggccat	ttccaccat	gatattcgcc	aagcagcct	cgccatgggt	cagcagcaga	1860
tcctcgccgt	cgggcatgct	cgccctgagc	ctggcgaaca	gttcggctgg	cgcgagcccc	1920
tgatgctctt	cgccagatc	atcctgatcg	acaagaccgg	cttccatccg	agtacgtgct	1980
cgctcgatgc	gatgtttcgc	ttggtggctg	aatgggcagg	tagccggatc	aagcgtatgc	2040
agccgcccga	ttgcatcagc	catgatggat	actttctcgg	caggagcaag	gtgagatgac	2100
aggagatcct	gccccgac	ttcgcccaat	agcagccagt	cccttcccgc	ttcagtgaca	2160
acgtcgagca	cagctgcgca	aggaacgccc	gtcgtggcca	gccacgatag	ccgcgctgcc	2220
tcgtcttgca	gttcattcag	ggcaccggac	aggtcggctt	tgacaaaaag	aaccgggcgc	2280
ccctgcgctg	acagccggaa	cacggcggca	tcagagcagc	cgattgtctg	ttgtgccag	2340
tcatagccga	atagcctctc	caccaagcgc	gccggagaac	ctgcgtgcaa	tccatcttgt	2400
tcaateatac	tcttctttt	tcaatattat	tgaagcattt	atcagggtta	ttgtctcatg	2460
agcggataca	tatttgaatg	tatttagaaa	aataaataaa	taggggttcc	gcgcacattt	2520
ccccgaaaag	tgccacctga	cgtctaagaa	accattatta	tcatgacatt	aacctataaa	2580
aataggcgta	tcacgaggcc	ctttcgtc				2608







TCGCGCGTTTTGGGTGATGACGGTGAAAACCTCTGACACATGCAGCTCCCGGAGACGGTCACAGCTTGTCTGTAAGC
GGATGCCGGGAGCAGACAAAGCCCGTCAGG GCGCGTCAGCGGGTGTGGCGGGTGTCGGGGCTGGCTTAACTATG
CGGCATCAGAGCAGATTG TACTGAGAGTGCACCATATGCGGTGTGAAATACCGCACAGATGGGTAAAGGAGAAAATA
CCGCATCAGGCGCCATTCCGCCATTCAGGCTCC GCAACTGTTGGGAAGGGCGATCGGTGCGGSCCTCTTCGCTATTA
CGCCAGCTGGCGAAAGGGGGATGTGCTGCAAGGC GATTAAAGTTGGGTAACGCCAGGTTTTCCAGTCACGAGCTT
GTA AACGACGGCCAGTGAATTA**GTGTTGAGACC**ATTAGCTCCGGTCT**CGACACT**AGCTTGGCGTAAATCATGGTC
ATAGCTGTTTCCTGTGTGAAATTTTATCCGCTCACAAATCCACACAACATACGAGCCGGAAGCATAAAGTGTAAAGC
CTGGGGTGCCTAATGAGTGAGCTAACTCACATTAATTGC GTTGCCTCACTGCCCGCTTTCCAGTCGGGAAACCTGT
CGTGCCAGCTGCATTAATGAATCGGCCAACGCGGGGGAGAGCGCGTTT GCGTATTGGGGCTCTTCGGCTTCCTC
GCTCACTGACTCGCTGCGCTCGGTCTGGGTCTGGGTGCGGCGAGCGGTATCAGCTCACTCAAAGGCGGTAAATACGGTTA
TCCACAGAATCAGGGGATAACGCAGGAAAGAACATGTGAGCAAAAGGCCAGCAAAAAGGCCAGGAACCGTAAAAAGG
CCGGGTTGCTGGCGTTTTTCCATAGGCTCCGCCCCCTGACGAGCATCACAAAAATCGACGCTCAAGTCAGAGGTG
GCGAAACCCGACAGGACTATAAAGATACCAGGCGTTTCCCCCTGGAAAGCTCCCTCGTGCCTCTCCTGTTCCGACC
CTGCCGCTTACCGGATACCTGTCCGCTTTCTCCCTTCG GGAAGCGTG GCGCTTCTCATAGCTCACGCTGTAGGTA
TCTCAGTTGGGTGTAGGTCGTTCCCTCAAAGCTGGGCTGTGTGCACGAACCCCGCTTCAGCCCGACCGCTGGCC
CTTATCCGGTAACTATCGTCTT GAGTCCAAACCCGGTAAAGACAGGACTTATCGCCACTGGCAGCAGCCACTG GTAAACA
GGATTAGCAGAGCGAGGTATGTAGGCGGTGCTACAGAGTTCTTGAAGTGTG GCTAACTACGGCTACACTAGAAAG
AACAGTATTTGGTATCTGCGCTCTGCTGAAGCCAGTTACCTTCGGAAAAAGAGTTGGTAGCTCTTGATCCGGCAAAC
AAACCACCGCTGGTAGCGGTG GTTTTTTGTGGTTCAGCAGCAGATTACGCGCAGAAAAAAGGATCTCAAGAAGAT
CCTTTGATCTTTCTACGGGGTCTGACGCTCAGTGGAACGAAAACCTACGTTAAGGGATTTTGGTCATGAGATTATCA
AAAAGGATCTTCACTAGATCCTTTTTAAATTA AAAATGAAGTTTTAAATCAATCTAAAGTATATATGAGTAAACTTGGTC
TGACAGTCAGAAAGAACTCGTCAAGAAGGGGATAGAAGGGCGATGCGCTGGGAATCGGGAGCGGCGATACCGTAAAG
CAGGAGGAAGCGGTCAGCCCATTCGCGCCAAAGCTCTTCAGCAATATCAGCGGTAGCCAAAGCTATGTCTGTATAG
CGGTCCGCCACCCAGCCGGCCACAGTCCGATGAATCCAGAAAAGCGGCCATT TTCCACCATGATATTCCGGCAAGC
AGGCATCGCCATGGGTCACGACGAGATCCTCGCGTCCGGGCATGCTCGCCTTGAGCCTGGCGAACAGTTCCGCTG
GCGCGAGCCCTGATGCTCTTCGTCCAGATCATCCTGATCGACAAGACCGGC TTCCATCCGAGTACGTGCTCGCTC
GATGCGATGTTCCGTTG GTGGTCCGAATGGGCAGGTAGCCGGATCAAGCGTATGCAGCCGCCCGCATTGCATCAGCC
ATGATGGATACTTCTCGGCAGGAGCAAGGTGAGATGACAGGAGATCCTGCCCGGCCACTTCGCCCAATAGCAGCC
AGTCCCTTCCCGCTTCAGTGACAACGTCGAGCACAGCTGCGCAAGGAACGCCCGTGTGGCCAGCCACGATAGCC
GGCTGGCTCGTCTTTCAGTTTCAATCAGGCGACCGGACAGGTCGGTCTTGACAAAAAGAACCGGGCGCCCTGCG
CTGACAGCCCGAACCGCGCGGCATCAGAGCAGCCGATTGTCTGTTGTGCCAGTCATAGCCGAATAGCCTCTCCAC
CCAAGCGGCCGGAGAACCTGCGTGCAATCCATCTTGTTCAATCATACTCTTCTTTTTCAATATTATTGAAGCATTTAT
CAGGGTATTGTCTCATGAGCGGATACATATTTGAATGTATTTAGAAAAATAAACAAATAGGGGTTCGCGGCACATTT
CCCCGAAAAGTGCCACCTGACGCTCAAGAAACCATTATATCATGACATTAACCTATAAAAAATAGGCGTATCACGAGG
CCCTTTCGTC

