[19] 中华人民共和国国家知识产权局

[51] Int. Cl.
C12Q 1/68 (2006.01)
C12N 15/00 (2006.01)
C12P 19/34 (2006.01)

[12] 发明专利申请公布说明书

[21] 申请号 200780007408.X

[43] 公开日 2009 年 3 月 25 日

[11] 公开号 CN 101395281A

[22] 申请日 2007.1.4
[21] 申请号 200780007408.X
[30] 优先权
[32] 2006. 1. 4 [33] US [31] 60/756,417
[32] 2006. 4.17 [33] US [31] 60/792,926
[32] 2006. 6.15 [33] US [31] 60/814,378
[86] 国际申请 PCT/CN2007/000001 2007.1.4
[87] 国际公布 WO2007/076726 英 2007.7.12
[85] 进入国家阶段日期 2008.9.1
[71] 申请人 骆树恩
地址 香港薄扶林沙湾道 25 号
[72] 发明人 骆树恩
[74] 专利代理机构 中国专利代理(香港)有限公司
代理人 梁 谋 黄可峻

权利要求书 4 页 说明书 39 页

[54] 发明名称

用于核酸作图和鉴定核酸的精细结构变化的方法以及用途

[57] 摘要

一般地讲，本发明涉及用于高通量分析核酸精细结构变化的方法。 具体地说，本发明涉及生产连接核酸的标签对的新策略、载体和载体组分，其中连接核酸的标签－对的组成成员处于用户限定的间隔距离，和/或为沿着靶核酸分子的长度分界一个或多个不同限制性内切核酸酶的邻近切割位点的核酸位置标记。

1. 一种并列序列标签(GVT)的方法，其中标签对(GVT-对)的两个组成成员是靶核酸分子中限定间隔距离的独特位置标记，所述方法包括:

将具有一个或多个限制性内切核酸酶识别位点的 DNA 连接物连接至片段化的靶 DNA 插入片段的两个末端;

使用限制性内切核酸酶在识别位点消化所述连接物，以在距靶 DNA 插入片段的每个末端的限定距离切割靶 DNA 插入片段，产生两个序列标签(GVT)，这两个序列标签含有靶 DNA 插入片段的末端序列，所述末端序列与质粒载体连接；和

再环化连接 GVT 的质粒载体，以获得含有具有两个并列 GVT 的 GVT 对的环化质粒。

2. 一种并列序列标签(GVT)的方法，其中标签对的两个组成成员为沿着靶核酸分子群长度侧接一个或多个给定限制性内切核酸酶的两个邻近并可切割的限制性内切核酸酶位点的独特位置标签，所述方法包括:

将消化的靶 DNA 插入片段连接入载体中侧接针对 IIS 型、IIG 型或 III 型限制性内切核酸酶的位点对的位置;

在距靶 DNA 插入片段的每个末端的限定距离切割插入 DNA，由此产生两个序列标签(GVT)，这两个序列标签含有靶 DNA 插入片段的末端序列，所述末端序列与载体骨架连接；和

再环化连接 GVT 的载体骨架，以形成环形质粒，每个环形质粒均携带含有两个并列 GVT 的 GVT-对。

3. 一种通过受控且有序的短 DNA 单体连接产生 DNA 寡聚物的方法，所述短 DNA 单体具有旋转等同的回文粘性末端，以产生以启动连接物的两个末端为边界的寡聚产物，所述方法包括以下步骤:

形成由启动连接物启动的 DNA 单体的寡聚物，其中一个连接物

末端具有不能自连接但可以粘附载体的非回文粘性末端，而另一个连接物末端具有这样的粘性末端：其未被磷酸化，从而防止形成连接物二聚体，并与 DNA 单体的粘性末端互补，用于连接单体，以启动寡聚物形成；和

在游离启动连接物与通过加入 DNA 单体形成的寡聚物连接时或通过与由启动连接物启动的另一个寡聚物连接终止寡聚物生长；

其中如此形成的寡聚物具有的平均长度受寡聚物开始形成时检测到的 DNA 单体与启动连接物的摩尔比率调节。

4. 一种用于制备环形组件载体的方法，所述环形组件载体能够生产连接的序列标签、增殖一个或多个独立的 DNA 插入片段和启动至少 4 种测序反应，所述方法包括：

提供两个组件载体区段或组件，第一个组件包含药物选择标记，第二个组件包含用于质粒复制的复制子，每个组件的末端部分都具有 IIS 型内切核酸酶切割位点，该末端部分产生独特的非回文粘性末端，用于载体组件的切离和靶向置换，以产生新的载体功能性；

将第一个和第二个组件的一个末端连接在含有识别位点的 DNA 表达盒中，所述识别位点在用内切核酸酶切割时在用于连接 DNA 插入片段的载体上产生非回文粘性末端对；和

连接第一个和第二个组件的另一个末端，以在第二个 DNA 表达盒中产生环形分子，所述第二个 DNA 表达盒含有第二个克隆位点，该克隆位点含有另一对限制性内切核酸酶识别位点，该识别位点在用内切核酸酶消化时在与第一个克隆位点不同的载体上产生非回文粘性末端对，用于连接第二个且不同的受体插入片段，该克隆位点两侧侧接不同 DNA 测序引物结合位点，以启动受体 DNA 插入片段中的桑格双脱氧测序反应；其中所述载体没有 *Mme* I、*Cst*M I、*Nme*A III、*Eco*P15 I、*Pst* II、*Bam*H I、*Pst* I、*Bsp*T I 或 *Kas* I 的识别位点，所述载体插入片段克隆位点包含 *Eco*31 I 和 *Esp*3 I 识别位点。

5. 权利要求 **1** 的方法，其中所述标签对的两个组成成员在靶核酸

分子中位于一个或多个限制性内切核酸酶的两个邻接并可切割的限制性内切核酸酶位点的侧翼。

6. 权利要求 1 的方法，其中所述靶 DNA 插入片段选自：基因组 DNA、cDNA、病毒 DNA、微生物 DNA、质体 DNA、化学合成的 DNA、核酸扩增的 DNA 产物和由 RNA 转录的 DNA。

7. 权利要求 1 的方法，其中所述靶 DNA 通过施加机械力或用一种或多种酶部分消化而被随机片段化。

8. 权利要求 1 和 2 的方法，其中所述靶 DNA 通过使用一种或多种单独的或组合的限制性内切核酸酶完全消化而被片段化。

9. 权利要求 1 和 2 的方法，其中所述片段化的靶 DNA 被大小分级分离。

10. 权利要求 1 和 2 的方法，其中所述片段化的靶 DNA 没有被大小分级分离。

11. 权利要求 1 和 2 的方法，其中产生 GVT 的限制性内切核酸酶为选自以下的 IIS 型或 IIG 型限制性内切核酸酶：*Mme* I、*Nme*A III、*Cst*M I、*Bce*A I、*Bpm* I、*Bpu*E I、*Bsg* I、*Bsm*F I、*Bst*V1 I、*Eco*57 I、*Eco*57M I 和 *Gsu* I。

12. 权利要求 1 和 2 的方法，其中所述 IIS 型或 IIG 型限制性内切核酸酶为 *Mme* I。

13. 权利要求 1 和 2 的方法，其中所述 IIS 型或 IIG 型限制性内切核酸酶为 *Cst*M I。

14. 权利要求 1 和 2 的方法，其中所述 IIS 型或 IIG 型限制性内切核酸酶为 *Nme*A III。

15. 权利要求 1 和 2 的方法，其中所述产生 GVT 的限制性内切核酸酶为选自以下的 III 型限制性内切核酸酶：*Eco*P15 I、*Eco*P1 I、*Pst* II、*Hind* fIII、*Sty*LT I、*Lla*F I、*Bce*S I、*Hine* I、*Pha*B I、*Hpy*790545P、*Hpy*790639 I 和 *Hpy*AXIP。

16. 权利要求 1 和 2 的方法，其中所述 III 型限制性内切核酸酶为

*Eco*P15 I。

17. 权利要求 **1** 和 **2** 的方法，其中所述 III 型限制性内切核酸酶为 *Pst* II。

18. 权利要求 **1** 和 **2** 的方法，其中所述产生 GVT 的 IIS 型或 IIG 型限制性内切核酸酶识别 6 个以上碱基对的不间断识别序列。

19. 权利要求 **1** 和 **2** 的方法，其中所述产生 GVT 的 III 型限制性内切核酸酶识别 6 个以上碱基对的不间断识别序列。

20. 权利要求 **2** 的方法，其中靶 DNA 插入片段选自：基因组 DNA、cDNA、病毒 DNA、微生物 DNA、质体 DNA、化学合成的 DNA、核酸扩增的 DNA 产物和由 RNA 转录的 DNA。

21. 权利要求 **4** 的组合物，其中所述选择标记没有 *Mme* I、*Cst*M I、*Nme*A III、*Eco*P15 I、*Pst* II、*Bam*H I、*Pst* I、*Bsp*T I 或 *Kas* I 限制性内切核酸酶位点。

22. 权利要求 **4** 的组合物，其中所述选择标记为 *Kan* 基因。

23. 权利要求 **4** 的组合物，其中所述选择标记为 *Amp* 基因。

24. 权利要求 **4** 的组合物，其中所述质粒复制子没有 *Mme* I、*Cst*M I、*Nme*A III、*Eco*P15 I、*Pst* II、*Bam*H I、*Pst* I、*Bsp*T I 或 *Kas* I 限制性内切核酸酶位点。

25. 权利要求 **4** 的组合物，其中所述质粒复制子为 *P15A*。

26. 权利要求 **4** 的组合物，其中所述质粒复制子为 *ColE1*。

27. 权利要求 **4** 的组合物，其中所述质粒复制子为 *pUC* 的 *ColE1* 衍生物。

28. 权利要求 **4** 的组合物，其中所述质粒掺入片段克隆位点通过用识别 6 个以上碱基对的不间断序列的 II 型、IIS 型或 IIG 型限制性内切核酸酶消化产生。

用于核酸作图和鉴定核酸的精细结构变化的方法以及用途

相关申请的交叉参考

[0001] 本申请要求保护基于2006年1月4日申请的美国临时专利申请美国序号60/756,417、2006年4月17日申请的美国临时专利申请美国序号60/792,926和2006年6月15日申请的美国临时专利申请美国序号60/814,378的优先权。前述临时申请的完整内容通过引用结合到本文中。

发明领域

[0002] 一般地讲，本发明涉及高通量分析核酸的精细结构变化的方法。具体地说，本发明涉及产生连接核酸的标签对的新策略、载体和载体组分，其中连接的核酸标签对的组成成员处于用户限定的分隔距离，和/或为沿着靶核酸分子长度分界一个或多个不同限制性内切核酸酶的邻近切割位点的核酸位置的标记。

发明背景

[0003]尽管最丰富且研究最深入的人类基因组变体类型是单核苷酸多态性(SNP)，但日益清楚的是，含有拷贝数(插入、缺失和复制)改变、倒位、易位和其它序列重排的所谓“精细结构变化”为人类基因组和其它基因组的整体特征。这些类型的变化似乎以比最初设想高得多的频率存在于一般人群中。建立的证据表明，结构变化可在每个基因组中包含上百万的核苷酸异质性。理解精细结构变化在基因组进化、与环境的相互作用、表型多样性和疾病或疾病易感性中的作用是当前基因组研究中最活跃的研究领域之一。关于综述，参见Bailey等(*Science 297*:1001 (2002)), Check (*Nature 437*:1094 (2005)), Cheng等(*Nature 437*:88 (2005))和Feuk等(*Nat Reviews 7*:85 (2006), Redon等

(*Nature 444: 444* (2006))。

[0004]与 SNP 分析相比，用于分析精细结构变化的有效高通量方法还没有被充分开发。重要的第一步是阵列比较基因组杂交(阵列 CGH)技术(Pinkel 等, *Nat Genet 20*:207 (1998); Pinkel 等, 美国专利第 *5,830,645* 号和第 *6,159,685* 号)，该技术能够定量靶 DNA 和参比 DNA 之间的相对拷贝数。阵列 CGH 允许以单个排列的细菌人工染色体(BAC)克隆水平的分辨率可靠地检测 DNA 或基因组样品之间的脱氧核糖核酸(DNA)拷贝数差异(Pinkel 等, *Nat Genet 20*: 207 (1998); Albertson 等, *Nat Genet 25*:144 (2000); Snijders 等, *Nat Genet 29*:263 (2001))。针对 cDNA (Heiskanen 等, *Cancer Res 60*:799 (2000); Pollack 等, *Nat Genet 23*:41 (1999))和高密度寡核苷酸阵列平台(Brennan 等, *Cancer Res 64*:4744 (2004); Lucito 等, *Genome Res 13*:2291 (2003); Bignell 等, *Genome Res 14*:287 (2004); Hung 等, *Hum Genomics 1*:287 (2004))修改阵列 CGH 进一步扩展了该方法的分辨率和应用性。通过其应用，阵列 CGH 已能够鉴定与肿瘤(Inazawa 等, *Cancer Sci 7*:559 (2004); Pinkel 和 Albertson, *Nat Genet 37* 增刊:S11 (2005); Pollack 等, *Proc Natl Acad Sci USA 99*:12963 (2002); Albertson 和 Pinkel, *Hum Mol Genet 12 Spec No 2*: R145 (2003))和疾病发展(Gonzalez 等, *Science 307*:5714 (2005))相关的基因拷贝数变化。

[0005] 尽管对拷贝数测定有用，但阵列 CGH 不适合于解决其它类型的基因组结构变化，最显著地，不适于倒位、易位和其它类型的核酸重排。Tuzun 等(*Nat Genet 37*:727 (2005))尝试用称为“fosmid 配对末端作图”的方法解决这些限制。该方法依靠 fosmid 包装的头部满装(head-full)机制，产生具有相当均一的约 40 kb 大小的测试者基因组插入片段的基因组 DNA 文库。随机选择的约 40 kb 文库插入片段的末端终止测序产生成对的短序列标签，其中每个标签-对标记两个基因组位置，这两个基因组位置沿着靶 DNA 长度间隔约 40 kb。然后用计算机比对标签-对和参比基因组组装，在它们的预期方向或它们的约 40

kb 间隔距离方面的任何不一致性都应表明在靶和参比核酸之间跨越该区域存在至少一个结构差异。作图位置间隔 40 kb 以上的的标签-对表示相比于参比在靶 DNA 上存在缺失; 间隔低于 40 kb 的作图位置表示在靶中有 DNA 插入片段。已作图的标签对方向的不一致性表示潜在的 DNA 倒位或其它复杂的染色体重排。标签-对分配给参比序列上的两个不同染色体表示染色体易位。Tuzun 等(*Nat Genet 37*:727 (2005)) 分析超过 1.1×10^6 个 fosmid 克隆插入片段，能够在测试者和参比基因组组装之间鉴定出接近 300 个结构变化位置。

[0006] 尽管 fosmid 配对末端作图是鉴定人类基因组中的精细结构变化的有用开始，但对于每个测试者，都需要巨大的成本和后勤工作来纯化和测序百万以上的 fosmid 插入片段末端，这阻碍了其在广泛人群和队列调查中鉴定基因组变化的应用，所述基因组变化可能与复杂疾病有关或响应于环境因素等。此外，fosmid 载体及其变体一般以非常低的拷贝数在宿主细胞中增殖，使得难以保持可靠的自动化 DNA 生产和测序。因此，需要用于基因组和相关研究的有效的、高通量的且低成本的鉴定精细结构变化的方法，从而将这些遗传元件与疾病、疾病发展和疾病易感性联系起来。本发明提供这些和其它的基本利益。

发明概述

[0007] 本发明提供筛选和鉴定核酸群的精细结构变化的改进的高通量方法、载体和载体组分。本发明创造了称为基因组变化标签(GVT)的短并列序列标签对，其中 GVT-对的组成成员处于用户限定的间隔距离，和/或为沿着所研究的核酸分子长度分界一个或多个不同限制性内切核酸酶的邻近位点的位置的标记。

[0008] 当用计算机比对 GVT-对的单个 GVT 和参比序列时，在它们的预期同一性、间隔距离和/或方向方面与参比序列的任何不一致性都表示靶和参比核酸之间在 GVT-对跨越的区域中存在一个或多个

精细结构差异。以此方式，GVT-对综合文库提供了可用于产生高分辨率结构作图的基因组分析，以鉴定核酸群之间的精细结构变化。本发明的另一方面能使用户确定和改变以 GVT-对做标签的核酸群上的间隔距离，使得可以产生 GVT-对文库，这些文库适合于以不同的空间分辨率和覆盖率检测精细结构变化。本发明的另一方面产生为 GVT-对的位置标记，所述位置紧邻沿着所研究核酸长度的一个或多个不同限制性内切核酸酶的邻近识别位点对。本发明的另一方面产生为位置标记的 GVT-对，所述位置紧邻沿着核酸长度的一个或多个不同限制性内切核酸酶的邻近识别位点对，所述标记沿着所研究核酸的长度间隔用户限定的距离。本发明的又一方面提供有效寡聚化产生的 GVT-对并在优化的载体和宿主系统中稳定增殖所获寡聚物的方法，以利于GVT-对的有效的高通量序列测定。

[0009] 按照本发明，待分析的目标群的 DNA 被随机地或在限定位点被片段化。在某些实施方案中，纯化片段化的 DNA 样品至预定大小，该大小限定了设置用于分析的分辨率水平的空间窗。片段化 DNA 的末端连接短的合成 DNA 连接物，该连接物含有合适的粘性突出端，有利于将连接物所连接的样品 DNA 克隆入适宜的载体中。连接物以某一方向掺入合适的 IIS 型、IIG 型或 III 型限制性内切核酸酶(例如：Mme I、NmeA III、CstM I、EcoP15 I、Pst II、Hpy790545P 或它们的优选功能等效物)的识别位点，使得用前述限制性内切核酸酶消化带有插入片段的质粒的文库以距离每个插入末端有用和限定的距离切割 DNA 插入片段，引起间插序列释放，产生与载体连接的基因组变异标签(GVT)对。通过将 GVT 连接在一起，产生代表原始靶 DNA 插入片段的两个末端区的 GVT-对，再环化新的线性化载体-GVT 复合物。将环化重组质粒转染入宿主细胞中，产生含有各自携带 GVT-对的单个质粒克隆的初始 GVT-对文库。扩增初始文库，用第二个限制性内切核酸酶消化纯化的质粒，该第二个内切核酸酶在 GVT 对侧翼的位点切割，以将 GVT 对由质粒载体中释放出来。纯化释放的 GVT-

对，寡聚化至适宜的大小，并亚克隆入适宜的载体中，用于寡聚 GVT-对的有效的高通量 DNA 序列测定。当用计算机比对 GVT 对的单个 GVT 序列和参比序列时，在它们的预期同一性、间隔距离或方向方面与和它们进行比对的参比的任何不一致性都标志着靶和参比核酸之间在 GVT-对跨越的区域中存在一个或多个精细结构差异。因此，相对于参比序列，多种 GVT 对的列表序列构成了目标核酸群的详细基因组分析。本发明的这些和其它方面在参考以下的详述时将变得显而易见。另外，以下标示了多个参考文献(包括专利、专利申请和期刊文章)，这些参考文献通过引用结合到本文中。

[0010] 本发明提供的有用用途包括但不限于快速建立高分辨率基因图，该图可用于：(1)鉴定基因组的精细尺度变化，该精细尺度变化促成人类多样性，可引起疾病、疾病发展或疾病易感性以及所观察到的用作诊断剂或治疗干预靶的其它性状；(2)能设计和建立寡核苷酸微阵列或其它测定方法，用于快速和大量地平行探询 DNA 样品的精细结构变化，该变化用于医学诊断、基因分型和其它这样的有用用途；(3)有利于由完整基因组或鸟枪 DNA 测序法精确并快速地进行 DNA 组装；(4)鉴定由差异 RNA 加工产生的 RNA 转录物的精细结构变化，以帮助基因组注释、功能基因组研究和潜在疾病诊断；(5)建立基因组分析，以利于比较基因组和系统发生研究，帮助差异鉴定密切相关的生物；和(6)建立相关品系、种族、生物型、变体、品种或物种的基因组分析，以鉴定可能引起任何可观察到的理论、医学或商业目标表型的基因组元件。

优选实施方案的详述

[0011] 以下方法提供了实施本发明的背景，并扩展和组合了先有技术的若干方面，以产生所述的并用于所示用途的改进新方法。

1. FOSMID 配对-末端作图

[0012] Tuzun 等(*Nat Genet 37*:727 (2005))描述了 fosmid 配对末端作图法，其中短序列标签对间隔约 40 kb，通过对来源于人类 fosmid 基因组文库的约 40 kb 随机基因组插入片段进行末端终止测序产生。在比对标签-对和参比基因组组装后，以预期的标记间隔距离和/或方向与和它们比对的参比序列的不一致性鉴定标签-对跨越的靶 DNA 中的结构变异。Tuzan 等概述的方法依靠 fosmid 包装，产生在基因组 DNA 上间隔距离约 40 kb 的标签对(根据试验，实际上片段在 32-48 kb 的范围内，＜平均值的 3 个标准偏差，39.9 ± 2.76 kb)。作者没有讲述或公开建立标签-对、建立不同间距以改变分析的空间分辨率的标签-对、改善插入片段长度在它们的文库中的均一性的其它方法，他们也没有讲述或公开生产其它类型的序列标签-对的方法，所述其它类型的序列标签-对例如为本发明的那些可基于邻近内切核酸酶切割位点对之间的位置和/或间隔距离而分界基因组位置的标签-对。

[0013] 许多类型的精细结构变化不能由以 fosmid 配对末端作图法固定的约 40 kb 分辨率窗来分辨。Fosmid 配对末端作图具有其它限制。Fosmid 载体在宿主细胞中以非常低的拷贝数增殖，该特性用于使某些基因组序列在细菌宿主中增殖的过程中遇到的潜在重组、重排和其它人为构造最少。尽管目前应用可扩增形式的 fosmid 载体(Szybalski，美国专利第 *5,874,259* 号)，但是末端测序 fosmid 克隆来产生标签的经济性仍非常差，原因在于与常规质粒相比 DNA 产量低，使得难以保持高通量的自动化模板生产和测序。此外，由单个 fosmid DNA 模板产生标签-对序列需要两个单独的测序反应，由此进一步降低了经济性。本发明通过以下几项克服了这些限制：(1)生产 GVT-对的能力，由此可将靶 DNA 上的标签-对成员的间隔由 50 bp 以下工程至几百千碱基对以上，以使检测分辨率适于分析不同类型的核酸和适于任何给定的实验设计；(2)标签-对成员之间明显更精确和均一的间隔，用于更高的分析精度；(3)基于除了间隔距离之外的其它标准生产基因组标签-对的能力，例如建立基于标签-对的邻近内切核酸酶位点

的位置和/或相对间隔距离，用于改善靶核酸样品的探询；和(4)寡聚GVT-对，并将GVT-对寡聚物亚克隆入载体中，载体针对高通量DNA测序进行了优化，以降低操作成本，由此使本发明可用于广泛的群体和队列研究。

2. 用于产生基因组标签的方法

[0014] 本领域已描述了多种基于DNA的、表征和对比基因组的指纹图谱法(Schloter等, *Microbiol Rev 24*:647, (2000)；Kozdroj和van Elsas, *J Microbiol Meth 43*:197, (2001)；Bouillard等, *Genome Res 11*:1453, (2001)；Wimmer等, *Genes Chromosomes Cancer 33*:285, (2002))。所有这些方法都使用靶DNA的限制性消化、PCR扩增或凝胶电泳分离的某些组合。通常，需要由用于DNA测序的凝胶提取候选DNA片段严重阻碍了这些方法。Dunn等的近期工作取得进步，其中，他们描述了一种使用IIS型/IIG型限制性内切核酸酶Mme I的方法，以产生用于分析基因组DNA的“基因组信号标签”(GST)(Dunn等, *Genome Research, 12*:1756 (2002))。通过将具有Mme I识别位点的连接物连接至基因组DNA片段产生GST，所述基因组DNA片段最初如下产生：通过用II型限制性酶初始消化靶DNA，接着用屡次切割的标签酶进行第二次消化。用Mme I消化连接物连接的DNA，产生21 bp的标签(GST)，该标签在DNA中的位置相对于初始限制性酶消化识别的位点固定。在通过PCR扩增后，寡聚纯化的GST，用于克隆和测序。所述标签及其相对丰度的鉴定用于建立基因组DNA的高分辨率“GST序列分析”，其可用于鉴定和定量给定的复杂DNA分离物中的初始基因组。使用鼠疫耶尔森氏菌(*Yersinia pestis*)作为模型系统，Dunn等能够确定相对简单的基因组中可经受添加或缺失限制性位点的改变的区域。然而，Dunn等的方法在复杂的基因组如人基因组中的用途有限，在复杂的基因组中，许多结构变化不能通过简单的获得或失去所研究的少量限制性内切核酸酶位点来揭示。此外，

对于即便 1 个限制性位点，跨越大基因组或分析多个样品所需要的 GST 的数量也是非常高的。相比于 Dunn 等的方法，本发明的 GVT-对具备经济性，提供了分析复杂基因组或扩展分析多个 DNA 样品的分析能力。

[0015] 一种称为基因表达的连续分析(SAGE)的方法的多种形式首先由 Velculesu 等(*Science 270*:484 (1995)和 Kinzler 等(美国专利第 *5,695,937* 号)描述，也使用 IIS 型或 IIG 型限制性内切核酸酶来产生 DNA 标签(Saha 等, *Nat Biotechnol 20*:508 (2002); Ng 等, *Nat Methods 2*:105 (2005); Wei 等, *Proc Natl Acad Sci USA 101*:11701 (2004))。所谓的“SAGE 标签”由 cDNA 模板产生，以提供对生物样品中的 cDNA 种类的复杂性和相对丰度的评价。最新形式的 SAGE 方法称为“LongSAGE”，其利用 Mme I 消化，产生长 21 bp 的标签，以标记 mRNA 转录物(Saha 等, *Nat Biotechnol 20*:508, (2002))。最新的精修称为“SuperSAGE”，其利用 III 型限制性内切核酸酶 EcoP15 I，产生 26 bp 的较长标签，用于改善 mRNA 对基因组的分配(Matsumura 等, *Proc Acad Sci USA 100*:15718-15723, (2003))。尽管本发明也利用 IIS 型、IIG 型或 III 型限制性内切核酸酶产生序列标签，但就生产方法和改善的信息内容而言，产生的本发明的 GVT-对与前述 SAGE 和 GST 标签根本不同。就产生尤其可用于表征新基因组或注释基因组和 DNA 样品的精细结构变化的高分辨率物理图谱而言，相对于使用未连接标签，使用连接的标签对提供了显著的效力和分析能力的改善。

[0016] Ng 等(*Nat Methods* 2:105 (2005))的近期工作描述了 SAGE 法的进一步发展。研究者利用 Collins 和 Weissman (*Proc Natl Acad Sci USA 81*:6812 (1984))倡导的方法，在该方法中，使用 DNA 片段环化(也称为分子内 DNA 连接)将远端 DNA 区段一起连接入载体中，产生所谓的“基因组跳跃文库”(Collins 等, *Science 235*:2046 (1987))。Ng 等环化单个 cDNA，将其 5’和 3’来源的 SAGE 标签连接在一起，产生“配对末端双标签”(PET)，然后将 PET 寡聚化，以利于有效测序。通过

鉴定转录单元的转录起始位点和聚腺苷酸化位点，以确定基因边界和帮助鉴定它们的侧翼调节序列，PET 可用于基因组注释。尽管生产本发明的 GVT-对和通过 Ng 等的方法生产 PET 均依靠分子内连接来实现 DNA 标记连接，但只有本发明的 GVT-对整合了 DNA 标记之间的精确物理距离和其它有用信息，由此使 GVT-对可用于详细的基因组结构分析。Ng 等没有讲述产生限定的空间间隔或其它标准的标签-对的方法，他们也没有描述如何使用他们的方法获得例如由 mRNA 加工或基因组的精细结构变化产生的结构变化。

3. 多重测序载体

[0017] 本文使用的术语多重测序载体是指为用于高通量桑格双脱氧测序而进行了优化的质粒载体，其具有携带两个或更多个独立插入片段的能力，导致由单个模板产生多个测序读数，由此通过经济性使用材料而节约成本。

[0018] 一般实施的技术是一个质粒载体增殖一个 DNA 插入片段。此构型的代表性质粒模板可由 DNA 插入片段侧翼的两个载体引物结合位点的每一个产生两个测序读数。Mead 和 Godiska (美国专利第 *6,709,861* 号)描述了“多重克隆载体”，借此将 DNA 插入片段克隆入克隆载体的分散位点中，从而允许随后在单个 DNA 测序反应中同时测序插入序列，或者在平行反应中使用同一模板制备物测序插入序列。

[0019] Mead 和 Godiska 描述的多重克隆载体可以 pLEXX-AK (Lucigen Corporation, Middleton, WI)商购，其为 CLONEPLEX™文库构建系统的主要组分。质粒载体 pLEXX-AK 作为两个去磷酸化的平端载体 DNA 区段由销售商提供。每个载体区段都具有单独的药物选择标记和用于 DNA 测序的测序引物结合位点对。提升载体系统，以降低用于高通量测序应用的材料成本。在实际实施时，用于 DNA 测序的主要高通量应用是鸟枪法基因组测序，pLEXX-AK 载体系统对其不

是特别适合。原则上，将磷酸化的平端 DNA 插入片段加入含两个去磷酸化 pLEXX-AK 载体区段的连接反应应产生这样的构型：其中 DNA 插入片段连接在两个载体区段的每一个之间，以产生功能性环形分子。在实践中，实际上产生了复杂背景的连接产物，其中只有少量产物含有期望的环形分子，由此单个 DNA 插入片段连接在两个不同载体区段之间。尽管两个载体区段的每一个上的药物抗性标记都允许由背景中选择生产性物质，但系统先天低效，原因是组成载体和插入片段的无方向的随机平端连接。大量的输出 DNA 插入片段在非生产性连接事件中扩增，需要相对大量的起始 DNA 来弥补损失。最关键的是，绝对要求将磷酸化平端 DNA 插入片段克隆到 pLEXX-AK 的两个位点中为应用设置了严重限制，其中原始 DNA 插入片段的序列连续性例如对构建用于鸟枪法测序的基因组 DNA 文库是关键。对于该应用，在文库构建过程中连接至其它基因组插入片段的任何基因组插入片段(所谓的嵌合插入片段)会严重破坏随后由序列数据建立的基因组组装。此外，尽管研究者要求保护的是他们的方法可被扩展至在载体上的 3 个以上分散位点具有独立插入片段的载体构建，以进一步增加效力，但对平端连接的依赖以及为保留每个载体区段而需要多个选择标记使该权利要求在实际执行时不切实际。

[0020] 本发明克服了 Mead 和 Godiska (美国专利第 *6,709,861* 号)所述用于构建多重测序载体的方法的前述局限，并提供用于直接组装更复杂的 DNA 分子、载体和载体组分的改进材料、方法和策略，以促进有效的多重 DNA 测序和其它应用。具体地说，本发明描述了组件载体系统，由此单个载体组分位于独特的 IIS 型限制酶位点侧翼，产生不对称粘性末端，以引导有序的载体组件组装，并以高效间插 DNA 元件至任何需要的构型，获得新功能性。由本发明获得的质粒 pSLGVT-3 是高拷贝数的质粒，为进行高通量 DNA 测序进行了优化，并可以携带至少两个独立插入片段，以能够由单个模板获得 4 个独立的测序读数。第二个质粒 pSLGVT-2 是 pSLGVT-3 的低拷贝数质粒变

体，其为增殖长 DNA 区段或在不重排或重组的情况下在微生物宿主中可能难以增殖的那些插入片段进行了优化。pSLGVT-2 和 pSLGVT-3 上的两个独立的克隆位点利用独特的非不对称互补粘性末端组，用于在两个克隆位点有序和特异性连接独立的插入片段，由此解除对平端克隆的需要和对磷酸化 DNA 插入片段的需求，磷酸化 DNA 插入片段是在文库构建过程中产生插入片段嵌合体的主因。来自 Mead 和 Godiska (美国专利第 *6,709,861* 号)的 pLEXX-AK 的 pSLGVT 系列-质粒的另一个分辨性特征是质粒复制子作为正确质粒组装的生物选择的应用，由此降低载体的材料大小，以增加携带插入片段大小的能力。如有需要，pSLGVT 载体的组件构建和载体组件之间的不对称粘性末端的应用允许快速重构载体系统，以携带 3 个或更多个独立的 DNA 插入片段。

1. 用于生产 GVT-对的核酸的制备和片段化

[0021] 如本文所述，本发明提供产生高分辨率基因组图谱的方法，该图谱可用于表征未知基因组或鉴定靶核酸群和参比序列之间的精细结构变化。适于分析的靶核酸包括但不限于：真核生物和原核生物的基因组 DNA、微生物 DNA、质体 DNA、质粒和噬菌粒 DNA、病毒 DNA 和 RNA、来源于核糖核酸(RNA)的互补 DNA (cDNA)，以及通过体外扩增如尤其通过 PCR 产生的 DNA。用于由前述来源分离 DNA、由 RNA 合成 cDNA 和用于扩增核酸的方法是本领域技术人员已知的。

[0022] 对于本发明的某些实施方案，GVT-对跨越的基因组距离决定了分析的分辨率水平。GVT 之间的间隔越小，所获得的用于作图和用于检测靶核酸群的精细结构变化的空间分辨率就越高。大 GVT 间隔需要较少的 GVT-对，以涵盖给定复杂性的 DNA 样品，但空间分辨率伴随下降。对于 mRNA 加工变体的鉴定，50 或 100 bp 的 GVT 间隔提供了足以检测 cDNA 群中的大部分可变剪切产物的分辨率水

平。对于人类全基因组勘测，10、25、50 或 100 kb 的 GVT 间隔在分辨率和经济性之间提供了生产力妥协。GVT 间隔、检测不同类型的 DNA 结构变化所需要的分辨率水平和涵盖给定序列复杂性至需要的深度需要的 GVT-对数量之间的功能性折衷可用计算机建模，以得到对给定应用最佳的实验设计。

[0023] 如上所述，用于构建 GVT-对的靶 DNA 插入片段的材料长度控制 GVT-对的残余 GVT 之间的间隔距离，由此设定用于分析的分辨率水平。产生和纯化接近均一大小的片段化核酸群的方法在本领域已有描述。片段化靶 DNA 至需要的长度可在用多种限制性内切核酸酶部分或完全消化的条件下酶促完成。使用具有 6 个以上碱基对的识别位点的限制性内切核酸酶对生产更长的 DNA 片段有用。屡次切割的 II 型内切核酸酶如 Mbo I、Hae III 等平均每 256 bp 切割 DNA 一次，这些酶在本领域已知通过部分消化生产可变大小的 DNA 片段。在放宽的条件下使用限制性内切核酸酶 CviJ I 于 GC 二核苷酸位置切割 DNA (Fitzgerald 等, *Nucleic acid Res 20*:3753 (1992))，这对在部分消化条件下生产 DNA 片段大小的有用连续体特别有用。在某些实施方案中，随机产生的 DNA 片段有用。用于随机生产 DNA 片段的方法包括：(1)用牛胰腺脱氧核糖核酸核酸酶 I (DNA 酶 I)消化，该酶在镁离子存在下在 DNA 中进行随机双链切割(Melgar 和 Goldwait *J Biol Chem 243*:4409 (1968)；Heffron 等, *Proc Natl Acad Sci USA 75*:6012 (1978))；(2)物理剪切(Shriefer 等, *Nucleic acid Res 18*:7455 (1990))；和(3)超声(Deininger *Anal Biochem 129*:216 (1983))。期望长度的随机片段化 DNA 片段还可以通过在 cDNA 合成过程中使用随机引物或者通过使用单独的或与描述的其它片段化方法组合的 PCR 产生。

[0024] 用于部分酶促消化的条件凭经验确定，改变反应体积、酶浓度以及酶对底物的比率、温育时间或温度的一个或多个参数。对于需要约 5 kb 以下的 GVT 间隔的高分辨率分析，优选非序列依赖性的片段化方法。牛胰腺 DNA 酶 I 在镁离子存在下在 DNA 中进行随机双

链切割(Melgar 和 Goldwait *J Biol Chem 243*:4409 (1968); Heffron 等, *Proc Natl Acad Sci USA 75*:6012)), 可用于该用途。同样, 还可以使用通过机械手段如超声进行的 DNA 片段化或剪切力的选择性用途。HydroShear 设备(Genomic Solutions Inc, Ann Arbor, MI)尤其可用于产生限定大小范围的随机 DNA 片段。还可以通过在 cDNA 合成过程中使用随机引物或通过使用单独的或与描述的其它片段化方法组合的 PCR 产生随机 DNA 片段。通过凝胶电泳最容易监测产生期望长度的产物的片段化的发展。在产生适宜的 DNA 大小分布后, 使用 T_4 DNA 聚合酶修复或制造 DNA 平端, 以准备平端连接 GVT-连接物, 用于生产本发明的 GVT-对。对于用一种或多种内切核酸酶部分或完全消化片段化 DNA 而留下粘性末端的情况, 修复不是必需的, 但需要设计 GVT-连接物来适应片段化酶产生的粘性末端。因为插入片段与其它插入片段的连接破坏了靶 DNA 的共线性, 并破坏了基因组图谱的建立, 所以通过磷酸酶去除插入 DNA 的 5'磷酸基团, 以防止插入 DNA 在与 GVT-连接物连接的过程中与其它插入 DNA 连接。

2. 选定大小的 DNA 的大小分级分离和纯化

[0025] 对于某些实施方案, 通过凝胶电泳分级分离去磷酸化的 DNA 插入片段, 并纯化, 以产生目标大小的 DNA 插入片段。丙烯酰胺凝胶最好用于分级分离 50 bp 至 1 kb 的 DNA。对于约 250 bp 至 20 kb 的片段大小, 0.4%至 3%琼脂糖凝胶是适宜的。脉冲场凝胶电泳适于分级分离约 10 kb 至几百 kb 大小的 DNA。这些方法描述于其中的参考文献(Rickwood 和 Hames (编辑), *Gel electrophoresis of nucleid acid: A practical approach* (Oxford University Press, New York, 1990); Hamelin 和 Yelle *Appl Theor Electrophor 1*:225 (1990); Birren 和 Lai, *Pulse field electrophoresis: A practical guide* (Academic Press, San Diego, 1993))。DNA 使用与样品平行电泳的适宜尺寸标志物确定大小, 并通过染色显现。用手术刀切下含有期望大小的 DNA 的凝胶切片, 其后

通过电洗脱或通过酶促或化学降低凝胶基质由凝胶基质回收 DNA。用于分析的回收 DNA 片段应接近均一大小。用于最大化分离分辨率的凝胶系统和电泳条件是本领域已知的。使用两轮以上的凝胶电泳获得更高的样品大小均一性。大小与平均长度偏差 2.5%以上的样品可能导致对本发明使用不可接受的噪音。

3. GVT-连接物的设计和与靶 DNA 的连接

[0026] 本领域技术人员会认识到，存在多种适用于本发明的 GVT-连接物设计。总之，适宜的 GVT-连接物包含以下材料特性：(1) 不等长度的 5'磷酸化寡核苷酸的短上链(top strand)和短下链(bottom strand)，其能够稳定互补碱基配对，产生双链结构；(2) GVT-连接物的一条链具有短非回文单链突出，其可与具有互补序列的载体连接；(3)另一连接物末端具有平端结构或其它适宜的末端结构，使得能够与去磷酸化的靶 DNA 片段有效连接；(4)位于靶 DNA 侧翼的连接物末端带有适宜的 IIS 型、IIG 型或 III 型限制性内切核酸酶识别位点，其方向使得该位点引导在靶 DNA 上以固定和有用的距离切割，以产生 GVT；和(5)邻近的或重叠的 IIS 型、IIG 型或 III 型酶识别位点是第二个限制性内切核酸酶位点，用于由载体切下产生的 GVT-对。适宜的 GVT 连接物的说明性实例如下所示(实施例 1-4)。

实施例 1：用于平端连接去磷酸化靶 DNA 的 GVT (Mme I)-连接物。

5'-pGACACAGAGGA TCCAAC (Seq ID No: 1)

GTCTCCT AGGTTGp-5' (Seq ID No: 2)

Mme I

[0027] 说明性实施例 1 (Seq ID No 1)的序列 5'pGACA-3'为粘性末端，用于将连接物连接的 DNA 插入片段亚克隆入具有一对突出的 5'-TGTC-3'序列的载体中。粘性末端是非回文的，以防止形成连接物二聚体和带有连接的连接物的 DNA 多聚体，并防止产生没有插入片段的载体。Seq ID No 1 的 5'-CAGAGGA-3'序列及其在 Seq ID No 2

的反向互补物 5′-TCCTCTG-3′描述了能够稳定互补碱基配对以帮助形成功能性双链连接物的短序列。Seq ID No 1 的 5'-TCCAAC-3′序列及其 Seq ID No 2 的反向互补物 5′-GTTGGA-3′为 IIS 型内切核酸酶 Mme I 的识别位点(Boyd 等, *Nucleid acid Res 14*:5255 (1986))。Mme I 切割其 5'-TCCAAC-3'识别位点下游(即为 5'至 3'方向) 20 bp 的 DNA 和在相对链上其反向互补物上游(即为 3'至 5'方向) 18 bp 的 DNA，以产生具有 2 bp 的突出 3′突出端的 20 bp GVT。与 Mme I 识别位点重叠的是 BamH I 的识别位点 5'-GGATTC-3'。BamH I 切割用于由载体释放产生的 GVT-对。BamH I 位点重叠 Mme I 位点，以便最小化外来的连接物序列，使寡聚化 GVT-对序列测定过程中的经济性更强。为在其它连接物设计中获得相同末端，重叠的 BspT I 位点可用于切除通过 CstM I 消化产生的 GVT-对。同样，Kas I 可用于切除通过用 NmeA III 消化产生的 GVT-对。

实施例 2: GVT (Mme I)-连接物连接 Xba I 消化的去磷酸化靶 DNA。

5'-pGACACAGAGGA TCCAAC (Seq ID No: 1)
GTCTCCT GGTTGGATCp-5' (Seq ID No: 3)
Mme I

说明性实施例 2 的 GVT (Mme I)-连接物的显著特征与说明性实施例 1 的连接物特征相同，额外掺入 5'-pCTAG-3'突出端(Seq ID No 3)，以引导连接物与 Xba I 消化的去磷酸化靶 DNA 片段连接。本领域技术人员会认识到，实施例 2 的连接物只是一个变体。存在通过掺入适宜的突出端产生的其它功能性连接物变体，这些变体与用其它限制性内切核酸酶消化的靶 DNA 连接，以适于不同的实验设计。

实施例 3: 用于平端连接去磷酸化靶 DNA 的 GVT (EcoP15 I)-连接物。

5'-pGACACAGACTG CAGCAG (Seq ID No: 4)
GTCTGAC GTCGTCp-5' (Seq ID No: 5)
EcoP15 I

实施例 4: 用于粘性末端连接用 Xba I 消化的去磷酸化靶 DNA 的 GVT(EcoP15 I)-连接物。

5'-pGACACAGACTG CAGCAG (Seq ID No: 4)
GTCTGAC GTCGTCGATCp-5' (Seq ID No: 6)
EcoP15 I

说明性实施例 3 和 4 描述了使用 III 型限制性内切核酸酶 EcoP15 I 产生 27 bp GVT 的连接物设计。用于切除 GVT-对的 Pst I 的限制性内切核酸酶位点(5'-CTGCAG-3')重叠 EcoP15 I 位点(5'-CAGCAG-3')。Pst I 位点与 EcoP15 I 位点重叠使 GVT-对中的外来连接物序列最小，使测序过程中的经济性更强。说明性实施例 4 的连接物掺入 Xba I 粘性末端，以引导连接物与 Xba I 消化的去磷酸化靶 DNA 片段连接。本领域技术人员会认识到，实施例 4 的连接物仅是一个变体。存在通过掺入适宜的突出端产生的其它功能性连接物变体，这些变体与用其它限制性内切核酸酶消化的靶 DNA 连接，以适于不同的实验设计。

[0028] 说明性实施例 1 和 2 的说明性 GVT-连接物通过用 Mme I 消化可以产生 18 bp 或 20 bp 的长 GVT。用 T_4 DNA 聚合酶去除由 Mme I 切割产生的 3'-突出端，之后平端连接所连接的 GVT，产生 36 bp 的 GVT-对，此时产生 18 bp 的 GVT。使用具有 16 倍变性 5'-突出端(与由 Mme I 消化产生的所有可能的 2 碱基 3'突出端相适)的连接物将 GVT 连接在一起，产生 GVT-对，此时产生 20 bp 的 GVT。与 Mme I 相比，EcoP15 I 切割产生 2 bp 的 3'凹缺末端，该末端由 DNA 聚合酶延伸，以产生 27 bp 的平端 GVT，通过平端连接由该平端 GVT 产生 54 bp 的 GVT 对。

[0029] 识别不间断的核苷酸序列并切割距其识别位点至少 10 个碱基距离的任何 IIS 型或 IIG 型限制性内切核酸酶均适用于产生 GVT。这些酶包括: BceA I、Bpm I、BpuE I、Bsg I、BsmF I、BstV1 I、Eco57 I、Eco57M I、Gsu I、CstM I、NmeA III 和 Mme I。其中，本发明优选使用 Mme I、NmeA III 或 CstM I，因为它们的切割位点在迄今为止

描述的 IIS 型内切核酸酶中距离其 DNA 识别位点最远，由此产生最长长度的 GVT。预期在以后将发现距离其识别位点具有更长的限定切割距离的其它 IIS 型或 IIG 型内切核酸酶，本发明可使用这些酶。关于 IIS 型和 IIG 型限制性内切核酸酶的综述，参见 Sistla 和 Rao (*Critical Rev Biochem Biol 39*:1, (2004))和 Bujnick (*Acta Biochimica Polonica 48*:935, (2001))。

[0030] 最初将 III 型限制性内切核酸酶描述为需要两个反向的不对称识别位点，体内切割发生在随机选定的两对反向识别位点中的仅一对的远端。关于综述，参见 Sistla 和 Rao, *Critical Rev Biochem Biol 39*:1, (2004))和 Bujnick (*Acta Biochimica Polonica 48*:935, (2001))。这些特性对本发明没用。然而，表征原型 III 型酶 EcoP15 I 表明，重组或纯化的天然酶在钾离子存在下以 2 倍至 3 倍高的浓度使用时能够在体外于单个位点混杂切割(Mucke 等, *J Mol Biol 312*:287, (2001); Peakman 等, *J Mol Biol 333*:321, (2003); Raghavendra 和 Rao, *Nucleid acid Res 32*:5703, (2004); Sistla 和 Rao, *Critical Rev Biochem Biol 39*:1, (2004))。开发 EcoP15 I 的该新描述的特性，以由 cDNA 生产 SAGE 标签(Matsumura 等, *Proc Acad Natl Sci USA 100*:15718, (2003))。EcoP15 I 酶可商购(New England Biolabs, Ipswich, MA)，本发明使用其生产 27 bp 的 GVT 和随后的 54 bp 的 GVT-对。本发明可使用在距其识别位点有用的距离切割 DNA 的其它 III 型内切核酸酶。

[0031] 本领域技术人员已知用于连接连接物与 DNA 插入片段和用于核酸分子的通用连接的方法。参见例如 Ausubel 等, (编辑), *Short Protocols in Molecular Biology*, 第 3 版, (John Wiley & Sons 1995)。用于将连接物与 DNA 插入片段平端连接的典型连接条件需要对靶 DNA 约 50 至 500 倍摩尔过量的连接物、高 T_4 连接酶浓度或包含诸如聚乙二醇的体积排阻剂(Pheiffer 和 Zimmerman, *Nucleid acid Res 11*:7853 (1983); Zimmerman 和 Pheiffer, *Proc Natl Acad Sci USA 80*:5852 (1983); Harrison 和 Zimmerman, *Nucleid acid Res 12*:8235 (1984); Hayahi 等,

Nucleid acid Res 14:7617 (1986))。连接物与粘性末端靶 DNA 的连接需要约 5 倍摩尔过量。使连接 GVT-连接物的 DNA 插入片段通过 ChromoSpin 柱(Clontech, Mountain View, CA)，以去除过量的连接物，然后通过凝胶电泳纯化和选择大小。为通过分子内连接产生 GVT-对，将纯化的产物连接入下述几种质粒载体中的一种。

4. 用于 GVT-对产生的载体构建体

[0032] 本发明的一个方面提供生产克隆载体的通用方法，所述方法通过用 IIS 型、IIG 型或 III 型内切核酸酶消化插入片段，接着进行分子内连接，能够产生 GVT-对。使用含有下述材料特征的 GVT-DNA 克隆表达盒，通过将表达盒连接入适宜的位点，修饰既有克隆载体。

实施例 5: GVT-DNA 克隆表达盒的说明性实例

5'AATTGGACAA (GAGACG)GAATAT(TCTAGA)ACGATA(CGTCTC)CTGTCC Seq ID:7
CCTGTT (CTCTGC) CGTATA (AGATCT)TGCTAT (GCAGAG)GACAGGTTAA5' Seq ID:8
EcoR I Esp3 I Xba I Esp3 I EcoR I

[0033] GVT-DNA 表达盒由两个互补寡核苷酸(Seq ID No 7 和 Seq ID No 8)合成生产，这两个互补寡核苷酸退火形成双链结构，该双链结构具有适于连接入既有载体中的末端粘性末端。作为说明性实例，以上的 DNA 表达盒显示出具有 EcoR I 的突出粘性末端，用于置换 pSMART VC 或 pSMART-cDNA 的多克隆位点区(Lucigen, Middleton, WI)，以使这些质粒载体能够产生 GVT-对连同先前描述的连接 GVT-连接物的靶 DNA。DNA 克隆表达盒的材料特征为反向的 IIS 型或 IIG 型内切核酸酶位点对；在说明性实施例中的 Esp3 I。用 Esp3 I 消化修饰载体在载体上产生一对非回文突出端(5′-TGTC-3′)，该突出端特异性连接互补突出端(5′-GACA-3′)，该互补突出端由 GVT-连接物连接的靶 DNA 插入片段伸出。载体和插入片段突出端的非旋转对称性基本上消除了没有插入片段的质粒和带有多个拷贝的连接物连接的靶 DNA 的质粒的产生，这些质粒会损害靶 DNA 插入片段的共线性和随后的基因组分析的建立。DNA 表达盒还掺入限制性内切核酸酶位点

Xba I，其位于反向的 Esp3 I 位点对之间。在 Xba I 位点克隆的适宜大小的“填充 DNA”片段能够监测载体制备过程中的 Esp3 I 消化。选择填充 DNA 片段的长度，使得可容易地通过凝胶电泳分离 Esp3 I 单一消化的、双重消化的和未消化的载体物质，仅纯化双重消化的片段待用。

[0034] 本领域技术人员会认识到，如同先前描述的适宜的 GVT-连接物的实施例一样，上述 DNA 克隆表达盒仅是多种功能等效设计中的一个。例如，DNA 表达盒中的 Esp3 I 位点可被其它 IIS 型或 IIG 型内切核酸酶的 Esp3 I 位点取代，其中 DNA 切割远离连续的识别位点。适宜的 IIS 型或 IIG 型酶包括：Alw I、Alw26 I、AsuHP I、Bbv I、Bcc I、BseG I、BseMi I、BsmA I、BsmF I、BsoMA I、BspCN I、BspM I、BspP I、BspTN I、BstF5 I、BstV1 I、Fau I、Fok I、Hga I、Hph I、Lwe I、Ple I、Pps I、Sfa I、Smu I、TspDT I、TspGW I、Bbs I、BciV I、Bfi I、Bfu, I、Bmr I、Bpi I、Bpm I、BpuA I、BpuE I、Bsa I、Bse3D I、BseM I、BseR I、BseX I、Bsg I、BsmF I、Bso31 I、BsrD I、Eco31 I、Esp3 I、BstV2 I、Bve I、Eam1104 I、Eci I、Eco57 I、Eco57M I、Faq I、Gsu I、Ksp632 I、CstM I、Mme I、NmeA III、Taq II、Sap I、它们的同切点酶和 Szybalski 等(*Gene 100*:13 (1991))描述的其它实例。优选的酶具有 6 个碱基对或更长的识别位点，(例如：BspM I、Eco31 I、Esp3 I、Sap I 及其同切点酶)，因为这些酶的位点不大可能存在于载体骨架中，降低了在载体构建过程中进行定点诱变以消除这些位点的需要。对本领域技术人员还显而易见的是，可改变通过前述酶产生的粘性末端的精确序列，只要它们可与它们预期的连接配偶体形成功能性和特异性的碱基对。DNA 表达盒上的末端结构可被修饰，以使表达盒适于连接入先存载体上的期望位点或连接至分离的载体组分，产生本发明可以使用的新载体。

[0035] 在宿主细胞中稳定增殖 DNA 区段的能力对基因组分析是非常重要的。含有富 AT 或 GC 区、重复序列、发夹、强启动子、毒

性基因和其它问题序列的 DNA 区段在宿主细胞中增殖时的重排或丢失是精细基因组变化研究的重要考虑因素。DNA 重排和其它克隆人为构造可被错认为是靶核酸的结构变化。而且，克隆偏好可限制插入片段的大小，并可能未足够地反映所研究基因组的重要区域。最近通过用条件扩增系统开发 fosmid 和 BAC 载体解决了该问题(Szybalski，美国专利第 *5,874,259* 号)，其中 DNA 的增殖保持在每个宿主细胞 1-2 个拷贝，直至为进行分析而被诱导至较高水平。报告了 15 kb 至 100 kb 以上的基因组插入片段的改善的稳定性，条件化扩增载体现在常规用于基因组研究。条件化扩增 fosmid/BAC 载体，如 pCC1FOS (Epicentre, Madison, WI)和 pSMART VC (Lucigen, Middleton, WI)以及它们的变体，适用于 10 kb 至 200 kb GVT-间隔的 GST-对生产。然而，常规低拷贝质粒载体的使用似乎足以稳定维持大 DNA 片段，而不需要 BAC、PAC 或 fosmid 型载体(Feng 等, *BioTechniques 32*:992, (2002); Tao 和 Zhang, *Nucleid acid Res 26*:4901, (1998))。pSMART 系列载体提供低拷贝数增殖，并具有在载体上具有转录终止子的额外特征，以降低转录干扰的潜在作用，这可能进一步改善DNA稳定性(Mead和Godiska, 美国专利第 6,709,861 号)。对于 50 bp 至 10 kb 以上 GVT-间隔的 GVT-对生产，多种已建立并广泛使用的低拷贝质粒型载体适于进行修饰，以生产 GVT-对，这些载体包括: pBR322 (Bolivar 等, *Gene 2*:95, (1977)) 和 pACYC177 (Chang 和 Cohen, *J Bacteriol 134*:1141, (1978))。

[0036] 通过将 GVT-DNA 表达盒于合适的克隆位点插入合适的载体骨架中生产用于 GVT-对生产的载体。用于连接核酸分子的通用方法是本领域技术人员已知的。参见例如 Ausubel 等(编辑), *Short Protocols in Molecular Biology*, 第 3 版, (John Wiley & Sons, New York, 1995)。为了使用，必须使载体骨架没有以下几类酶的识别位点: (1) II 型、IIS 型或 IIG 型限制性内切核酸酶，这些酶用于产生 DNA 克隆表达盒上的粘性末端，这些末端用于定向克隆靶 DNA 或连接物连接的靶 DNA; (2) IIS 型、IIG 型或 III 型内切核酸酶，这些酶用于由克隆的

靶 DNA 插入片段产生 GVT；和(3)用于切除质粒中新产生的 GVT-对的酶。对于 GVT-DNA 表达盒和 GVT-连接物的说明性实例，载体骨架需要没有 Esp3 I、Eco31 I、CstM I、Mme I、NmeA III、Pst II、EcoP15 I、BamH I、Pst I、BspT I 或 Kas I 位点的特定组合，实际需要取决于所用 GVT-DNA 表达盒和连接物的精确构型。如有需要，可通过使用标准方法的定点诱变使载体骨架没有前述那些位点。参见例如：McPherson (编辑), *Directed mutagenesis: A practical approach* (Oxford University Press, New York, 1991)和 Lok (美国专利第 6,730,500 号)。通常，可通过单碱基对变化改变大部分载体 DNA，以消除不需要的限制性内切核酸酶识别位点，而对载体功能性没有不适当影响。在蛋白编码序列中，将单核苷酸变化靶向密码子摇摆位置，以保持天然蛋白编码。在载体骨架上的它处实施的改变应需要在使用前进行功能验证。

5. GVT-对生产载体 pSLGVT-1 和 pSLGVT-2

[0037] 本发明的质粒 pSLGVT-1 和 pSLGVT-2 分别是专门设计用于使用 Mme I 或 EcoP15 I 生产 GVT 和 GVT-对的优化通用载体。pSLGVT-1 和 pSLGVT-2 也没有 CstM I 和 NmeA III 位点，可用于按照本发明的方法使用这两种酶生产 GVT 和 GVT-对。基础载体含有两种化学合成的 DNA 组件，以分别提供药物选择和质粒复制的基本维持功能。连接两个 DNA 组件产生的环形分子为 DNA 表达盒，其为基础质粒骨架提供特定实用功能。载体组件带有末端独特的 IIS 型限制性内切核酸酶位点，其产生独特的不对称粘性末端，以允许在以后快速重构载体组分，从而加入或取代针对新功能的组件或 DNA 表达盒。

[0038] 第一个载体组件含有修饰的 P15A 复制起点。带有 P15A 复制子的质粒以每个宿主细胞约 15 个拷贝的低数目增殖(Sambrook 等, *Molecular Cloning: A Laboratory Manual*, 第 2 版, CSH Laboratory Press, Cold Spring Harbor, NY, (1989))，由此优化克隆的基因组插入片段的稳定性。相比之下，高拷贝数质粒，例如 pUC 或 pBluescript，可

达到每个细胞几千个拷贝。P15A 复制子中的两个 Mme I 位点各自通过单核苷酸改变而被消除，产生用于构建质粒 pSLGVT-1 的“P15A-m 复制子组件”。预期这两个位点的突变不改变二级结构或调节质粒复制所需的 RNA II 或 RNA I 的转录。以相同方式消除在 P15A 复制子中的单 EcoP15 I 位点，以产生用于构建质粒 pSLGVT-2 的“P15A-e 组件”。两种形式的 p15A 组件在所述组件的 RNA II 启动子末端侧接独特的 Bpi I 位点，产生 5’ GTGA-突出端，以利于 DNA 表达盒的连接。出于相同目的，复制组件的复制叉末端侧接 Faq I 位点，产生 5’ TCTC-突出端。

[0039] 第二个载体组件包含来自转座子 Tn903 的修饰形式的 Kan 基因，该基因赋予针对抗生素卡那霉素的抗性(Grindley 等, *Proc Natl Acad Sci USA 77*:7176, (1980))。利用摇摆位置并无论何时都尽可能与大肠杆菌中的优化密码子选择一致，去除 Kan 基因编码区中的 4 个 Mme I 位点连同 2 个 Nci I 和 Nsi I 位点以及针对 Esp3 I、Pst II 和 Hind III 的单个位点，以产生“Kan 组件”。Kan 组件在组件的 Kan 启动子末端侧接独特的 Sap I 位点，产生 5’ TTG-突出端，用于 DNA 表达盒连接。在 Kan 组件的另一端的独特 BspM I 产生 5’ ACTG-突出端，用于相同目的。一般公认，卡那霉素药物选择为维持带有特别长的和/或难的插入片段的质粒提供最佳稳定性，在许多情况下，其应用还会允许在液体培养物中有限但便利的扩增质粒文库，而没有可使质粒文库的组成失真的不适当克隆选择。

[0040] pSLGVT 系列质粒的核心组分是两个 DNA 克隆表达盒，其提供特定插入片段克隆功能性，用于将 Kan 组件和复制子组件连接在一起，产生环形质粒。质粒 pSLGVT-1、-2 和-3 具有通用结构，该结构在顺时针方向的环形图上包含以下材料特征：(1)复制子组件；(2) DNA 克隆表达盒 1；(3) Kan 组件；和(4) DNA 克隆表达盒 2。Kan 基因的质粒复制和转录以顺时针方向进行。以下显示了 DNA 克隆表达盒 1 和 2 的结构：

实施例 6: DNA 克隆表达盒 1 和 2

DNA 克隆表达盒 1

5'GAGA(T7>) GACAA(GAGACG)GCATCTCAGTAG(TCTAGA)AGTGCACGATAG(CGTCTC)CTGTC(T3)
(T7) CTGTT (CTCTGC) CGTAGAGTCATC (AGATCT)TCACGTGCTATC(GCAGAG)GACAG(<T3)CAA
Esp3 I Xba I Esp3 I

DNA 克隆表达盒 2

5'GAGT(M13F>) CTGAT(GAGACC)CTAGCCTCTTGA(GTCGAC)CACTATACATCA(GGTCTC)CTCAG(M13R)
(M13F)GACTA(CTCTGG)GATCGGAGAACT(CAGCTG)GTGATATGTAGT(CCAGAG)GAGTC(<M13R)CACT
Eco31 I Sal I Eco31 I

T7 测序引物: 5'-TAA TAC GAC TCA CTA TAG GG-3'

T3 测序引物: 5'-ATTAACCCTCACTAA AGG GA-3'

M13 F 测序引物: 5'-CAC GAC GTT GTA AAA CGA C-3'

M13 R 测序引物: 5'-GGA TAA CAA TTT CAC ACA GG-3'

[0041] DNA 克隆表达盒 1 由两个化学合成的互补寡核苷酸产生，这两个寡核苷酸退火形成双链结构，具有两个末端不对称的 5'突出粘性末端 5'-GAGA-3'和 5'-AAC-3'，用于将表达盒分别定向连接至复制子组件(P15A-m 或 P15-e)的 5'-TCTC-3'突出端和 Kan 组件的 5'-GTT-3'突出端。显示了 DNA 克隆表达盒 1 和 2 上针对 T7、T3、M13 正向和 M13 反向测序引物的结合位点。本领域技术人员会知晓，其它测序引物结合位点也适用于本发明。DNA 克隆表达盒 1 上的反向 Esp3 I 位点对在载体上产生 5'-TGTC-3'突出端对，以接受连接 GVT-连接物的靶 DNA，以便生产 GVT-对产物。Xba I 位点位于 Esp3 I 位点组之间，用于克隆填充 DNA 片段，以在制备载体时帮助监测 Esp3 I 消化进展，以接受连接 GVT-连接物的靶 DNA。Esp3 I 位点侧翼是 T7 和 T3 测序引物的引物结合位点。这些引物位点用于测序部分靶 DNA 插入片段，以便对文库构建进行质量控制。如本文公开内容的以下和以后章节所述，pSLGVT-质粒系列的变体 pSLGVT-3 利用这些引物位点对寡聚化的 GVT-对进行高通量的多重 DNA 测序。

[0042] DNA 克隆表达盒 2 由两个化学合成的互补寡核苷酸产生，这两个寡核苷酸退火形成双链结构，具有两个末端不对称 5'突出粘性

末端 5'-GAGT-3'和 5'-TCAC-3'，用于将表达盒分别定向连接至 Kan 组件的 5'-ACTC-3'突出端和复制子组件(P15A-m 或 P15-e)的 5'-GTGA-3'突出端。DNA 克隆表达盒 2 上的反向 Eco31 I 位点对在载体上产生 5'-TCAG-3'突出端对，并提供可替代位点，以接受连接 GVT-连接物的靶 DNA，以便生产 GVT-对。Sal I 位点位于 Eco31 I 位点组之间，用于克隆填充 DNA 片段，以在制备载体时帮助监测 Eco31 I 消化进展，以接受靶 DNA。Eco31 I 位点侧翼是 M13 正向和 M13 反向测序引物的引物结合位点。这些引物位点用于测序部分靶 DNA 插入片段，以便对文库构建进行质量控制。如以下和本文公开内容的以后章节所述，pSLGVT-质粒系列的变体 pSLGVT-3 利用这些引物位点对寡聚化的 GVT-对进行高通量的多重 DNA 测序。

[0043] 质粒 pSLGVT-1 通过两步连接策略构建。P15A-m 复制子组件与 DNA 克隆表达盒 1 温育。在单独的连接反应中，Kan 组件与 DNA 克隆表达盒 2 温育。在 1 小时温育后，合并两个连接反应，以组装需要的环形产物。质粒 pSLGVT-2 通过类似方式生产，但在初始连接反应中用 P15A-e 复制子组件取代 P15A-m 复制子组件。

[0044] 构建 pSLGVT 系列质粒的替代途径是通过化学合成，借此由一系列化学合成的寡核苷酸组装质粒。

[0045] 本发明的质粒 pSLGVT-3 代表用于有效构建多重 DNA 测序载体家族的新方法，所述载体用于测序寡聚的 GVT-对和其它 DNA 区段。通过用含有来源于 pUC 质粒的复制子的那些位点终止的片段替代 pSLGVT-2 的 Bpi I-Fag I 片段上的 P15A 复制子组件，构建质粒 pSLGVT-3。pUC 复制子来源于低拷贝数的 ColE1 复制子，其中在与缺失 rop 调节物组合的 *Ori* 中的单碱基突变导致质粒拷贝数由每个细胞约 20 个拷贝增加至超过 1000 个拷贝(Vieira 和 Messing, *Gene 19*:259, (1982))。高拷贝数的 pSLGVT-3 应有利于寡聚化 GVT-对的高通量 DNA 测序的模板制备。显著特征性的 pSLGVT-3 为位于 DNA 表达盒 1 和 2 中的前述反向 IIS 型限制性酶位点对。用 Esp3 I 和 Eco31 I 消化

pSLGVT-3 产生两个 DNA 载体区段，其具有不对称粘性末端，用于 2 个独立的寡聚 GVT-对区段组的靶向和定向的连接，这允许由存在于 DNA 表达盒 1 和 2 中的 4 个引物结合位点的每一个获得 4 个独立的测序读数。常规测序载体通常携带 1 个插入片段，并可以支持仅 2 个测序读数。

6. GVT-对生产

[0046] 本文使用的 fosmid、BAC 和其它游离型元件被统称为质粒，以下描述的用于产生 GVT-对的方法基于先前描述的 GVT-DNA 表达盒和 GVT-连接物的说明性实施例。在某些实施方案中，通过机械或酶促方法随机片段化用于生产 GVT-对的靶 DNA，产生需要大小的片段，用于 GVT-对生产。在其它实施方案中，以单独的反应或与在特定位点切割靶 DNA 组合，用一种或多种限制性内切核酸酶完全消化靶 DNA，产生 DNA 片段群，用于生产如本文公开内容所述的 GVT-对。对于用产生粘性末端的酶消化的靶 DNA，可直接将去磷酸化的插入片段 DNA 克隆入适当修饰的载体的 IIS 型或 IIG 型位点对之间的位点，无需连接物。在又一个实施方案中，用一种或多种限制性内切核酸酶完全消化靶 DNA，并分级分离至需要的大小，用于生产 GVT-对。

[0047] 使用 T_4 DNA 聚合酶修复用于生产 GVT 的、具有“不齐”末端的靶 DNA，并去磷酸化，以防止在插入片段与 GVT-连接物连接的过程中出现插入片段的自连接。同样，带有粘性末端的靶 DNA 被去磷酸化，之后与带有互补末端的适宜 GVT-连接物连接。使连接 GVT-连接物的 DNA 通过适宜的 Chroma Spin 柱(Clontech, Mountain View, CA)，以去除未连接的连接物，之后将连接连接物的靶 DNA 连接至 GVT 生产载体。在某些实施方案中，通过凝胶电泳或其它方法选择为期望长度的靶DNA大小，之后将插入片段与GVT-连接物连接，随后连接入 GVT-生产载体，例如在本发明中描述的 pSLGVT-1 和

pSLGVT-2。

[0048] 针对在一定片段长度范围内的DNA区段，已描述了用于优化载体与插入片段的分子间连接继之以分子内连接以产生环形分子的连接条件(Wang和Davidson, *J Mol Biol 19*:469 (1966); Dugaiczyk等, *J Mol Biol 96*:171 (1975); Collins和Weissman, *Proc Natl Acad Sci USA 81*:6812 (1984))。用于连接核酸分子、转染入宿主细胞中和构建基于质粒的文库的通用方法是本领域技术人员已知的。参见例如Sambrook等, *Molecular Cloning: A laboratory manual* 第2版, (CSH press, New York, 1989); Ausubel等(编辑), *Short Protocols in Molecular Biology*, 第3版, (John Wiley & Sons, New York 1995); Birren等, *Bacterial artificial chromosomes in genome analysis: A laboratory manual* (CSH Press, New York, 1999)。通过电穿孔或转染将连接的DNA导入宿主细胞中。甲基化的靶DNA的增殖需要具有失活的*mcr*和*mrr*等位基因的宿主细胞菌株，所述甲基化的靶DNA例如为基因组DNA或cDNA，通过某些利用甲基化核苷酸类似物的方法合成。适宜的宿主菌株包括: *10G* (Lucigen, Middleton, WI); *XL1-Blue MR*和*XL2Blue MRF'* (Stratagene, La Jolla, CA)。将电穿孔或转染的细胞以约20,000个菌落/板的密度铺板在处于适宜药物选择下的10 cm直径琼脂板上，以产生初始文库。替代方法是在液体培养基中培养转染细胞，同时小心使细胞不过度生长，从而促进克隆选择。处于培养中的克隆总数应反映出研究设计所需要的GVT-对数。收获细胞，并分离质粒，用于下述的后续步骤。

[0049] 作为通用步骤，用Mme I、CstM I、NmeA III或EcoP15 I (New England Biolabs, Ipswich, MA)消化带有靶DNA插入片段的纯化质粒，以产生符合实验设计的GVT。新产生的GVT的末端用T_4 DNA聚合酶修复，以使消化的末端平端。通过凝胶电泳将连接新产生的GVT的线性化质粒与切离的间插插入片段残余部分纯化开来，纯化的产物通过平端连接环化，产生初始GVT-对文库。用于再环化质粒的

替代方法避免了对 DNA 末端修复的需要，利用携带所有 16 倍双碱基对变性的 3'-突出端或 5'-突出端的连接物，所述突出端分别通过 Mme I、CstM I、NmeA III 或 EcoP15 I 消化产生。所述方法应将通过 Mme I 消化产生的 GVT 长度由 18 bp 增加至 20 bp，但不应增加 EcoP15 I 产生的 GVT 长度，因为 EcoP15 I 消化产生 2 bp 的 3′-凹缺末端，该凹缺末端在修复过程中被 T_4 DNA 聚合酶补平，之后质粒再环化，产生 GVT-对。使用连接物再环化质粒会增加所获的具有外来序列的 GVT-对的总体单位长度，对寡聚 GVT-对的测序经济性产生负面影响。

[0050] 将环化质粒导入到宿主细胞中，并以约 20,000 个菌落/10 cm 板的密度铺板，或在液体培养基中在选择下培养，以产生初始 GVT-对文库。用切割 GVT-对两侧的酶消化初始 GVT-对文库的纯化质粒，以将 GVT-对切离质粒。在用于文库构建的 GVT-连接物的说明性实施例中，分别使用 BamH I 或 Pst I 由 Mme I 或 EcoP15 I 产生的 GVT-对文库中切下 GVT-对。使用类似的连接物设计，酶 BspT I 或 Kas I 分别可用于由 CstM I 或 NmeA III 产生的 GVT 对文库中切下 GVT-对。以下显示了通过 Mme I 或 EcoP15 I 消化后平端末端连接产生的切离 GVT-对的一般性结构:

实施例 7: 通过 Mme I 消化、分子内连接和经 BamH I 消化切除产生的 GVT-对单体的结构:

5'pGATCCAAC-18N-18N-GTTG

GTTG-18N-18N-CAACCTAGp

Mme I Mme I

"18N-18N" 代表 GVT-对的 2 个并列的 18 bp GVT，由用 Mme I 消化的靶 DNA 产生。单体上的 Mme I 识别位点对标以下划线。余下的 52 bp 单体部分，包括标有下划线的 Mme I 位点，包含通用"构架"。52 bp 的 GVT-对单体在 5%聚丙烯酰胺凝胶上通过电泳分离，并纯化和寡聚化，用于测序。

实施例 8: EcoP15 I 消化、分子内连接和经 Pst I 消化切除产生的 GVT-对单体的结构:

5' pGCAG-27N-27N-CTGCTGCA
ACGTCGTC-27N-27N-GACGp 5'
EcoP15 I EcoP15 I

"27N-27N"代表 GVT-对的 2 个并列的 27 bp GVT, 由用 EcoP15 I 消化的靶 DNA 产生。单体上的 EcoP15 I 识别位点对标以下划线。余下的 70 bp 单体部分, 包括标有下划线的 EcoP15 I 位点, 包含通用"构架"。70 bp GVT-对单体在 5%聚丙烯酰胺凝胶上通过电泳分离, 并纯化和寡聚化, 用于测序。

7. 用于有效 DNA 测序的寡聚化 GVT-对单体的产生

[0051] 为经济性使用 DNA 测序资源, DNA 序列标签通常被寡聚化, 并作为延长的寡聚物克隆入序列载体中。本发明提供有效的方法来产生 DNA 标签的寡聚物, 并将寡聚化的 DNA 区段组装成改进的测序载体。通常, 构建具有末端携带对称粘性末端(例如在所示实施例中的 BamH I 或 Pst I)的 DNA 序列标签单体。然而, 通常用于生产和克隆寡聚化序列标签单体的方法先天低效, 原因是在寡聚反应当中和插入片段连接入载体的过程中产生非生产性的环形产物。如本文所述, 以下概述了生产和克隆寡聚序列标签的新的和优选的方法。改进的方法利用"启动连接物", 其可启动单体的寡聚化, 并允许将寡聚产物克隆入载体中, 但同时防止寡聚 DNA 环化。以下显示了适宜的启动连接物的 4 个说明性实施例:

实施例 9: 用于 BamH I 寡聚物的启动连接物 GACA-Bam

5' pGACACACGTGCTAGTCCG
GTGCACGATCAGGCCTAG-5'

实施例 10: 用于 Pst I 寡聚物的启动连接物 GACA-Pst

5’ pGACACACGTGCTAGTCCCTGCA
GTGCACGATCAGGG-5’

实例 11: 用于 BamH I 寡聚物的启动连接物 CTGA-Bam

5’ pCTGACACGTGCTAGTCCG
GTGCACGATCAGGCCTAG-5’

实施例 12: 用于 Pst I 寡聚物的启动连接物 CTGA-Pst:

5’ pCTGACACGTGCTAGTCCCTGCA
GTGCACGATCAGGG-5’

[0052] 启动连接物由 2 个化学合成的互补寡核苷酸产生，这 2 个寡核苷酸退火形成说明性的双链连接物。在一个末端，连接物具有回文粘性互补末端，用于连接 BamH I 或 Pst I 产生的序列标签单体，并启动寡聚物形成。非不对称粘性末端(5’-GACA-3’或 5’-CTGA-3’)存在于另一个连接物末端，用于特异性连接入多重测序载体 pSLGVT-3 上的一个或另一个克隆位点中。pSLGVT-3 和 pSLGVT 系列的其它质粒的独特设计具有携带两个独立 DNA 插入片段的能力。

[0053] 在启动连接物的仅 1 个末端处针对单体的互补粘性末端将单体的连接和寡聚物的增长限制在一个方向，由此使形成的非生产性环形分子最少。启动连接物的下链未被磷酸化，以防止形成连接物二聚体。在连接反应中，在对启动连接物过量的 GVT-对单体存在下进行寡聚物形成，这允许反应进行到完成。产生的主产物为在两个末端被启动连接物“加帽”的寡聚单体的集合。DNA 单体与启动连接物的比率表明最终寡聚化产物的总体大小范围。使用一份启动连接物对 N 份单体作为起点通过滴定获得生产性比率；其中 N 等于(在终产物中需要的单体平均数+2)/2。如有需要，可合并使用一系列启动连接物对单体比率的若干连接反应，通过凝胶电泳纯化期望长度的产物。对条件进行选择，以由 GAGC 启动连接物和 GTGA 启动连接物产生寡聚物质，该寡聚物质含有约 25-30 个拷贝(约长 1.6 至 2 kb)，在 1.5%

琼脂糖凝胶上纯化，并克隆入测序载体 pSLGVT-3 的两个位点中。

8. 将寡聚化 GVT-对单体克隆入多重测序载体 pSLGVT-3 中

[0054] 本文使用的术语多重测序载体指为进行高通量桑格双脱氧测序而进行了优化的质粒载体，具有在两个 DNA 克隆表达盒的每一个中携带独立插入片段的能力，由 4 个引物结合位点的每一个都获得 4 个测序读数。

[0055] 用 Eco31 I 和 Esp3 I 消化 pSLGVT-3 (或其低拷贝数变体 pSLGVT-2)，以产生两个载体区段，所述区段通过凝胶电泳纯化待用。载体区段 1 含有质粒复制子组件，并具有 5'TCAG-3'和 5'-TGTC-3' 粘性末端。载体区段2含有Kan组件，并具有5'-TGTC-3'和5'-TCAG-3' 突出端。载体区段 1 与通过启动连接物 GACA-产生的等摩尔当量的寡聚 GVT-对连接。在独立的反应中，载体区段 2 与通过启动连接物 CTGA-产生的等摩尔当量的寡聚 GVT-对连接。在 1 小时温育后，合并两个连接反应，并再温育，以组装需要的环形产物，该产物含有两个独立获得的寡聚 GVT-对的插入片段，连接在两个载体区段之间。

[0056] 600-800 bp 的典型序列读取长度足以确定至少 10 个 GVT-对的序列。基于对每个测序读数 10 个 GVT-对和单个模板的 4 个测序读数的测定结果，本发明的单个质粒模板应产生 40 个以上 GVT-对的序列。采用 40 kb 的末端配对间隔的 Fosmid 配对末端作图需要假定末端至末端间隔 75,000 个 fosmid 末端配对，以 75,000 个 fosmid 模板制备物和 150,000 个测序读数的成本支付人类基因组的费用。相比之下，本发明使用的 GVT 之间以类似的 40 kb 间隔 1 倍覆盖人类基因应需要 75,000 个 GVT-对，其以仅 7,500 个测序读数和 1,875 个质粒模板制备物的成本产生。对于类似的基因组覆盖和分辨率水平，与 Tuzun 等(*Nat Genet 37*:727 (2005))的 fosmid 配对末端方法相比，本发明的方法使用降低 20 倍以上的测序读数和降低 40 倍以上的模板制备物。

本发明的优选实施方案

[0057] 证据表明，遗传结构变化在人中含有成百万的碱基配对异质性，是我们的遗传多样性的主要组分，其中一些几乎肯定牵涉我们与环境的相互作用，并在疾病、疾病易感性或发展中起作用。本发明涉及产生连锁基因组序列标签对的系统、方法、组合物、载体、载体组分和试剂盒，所述标签对用于快速产生高分辨率遗传图谱，以鉴定这些基因组变化。

[0058] 在一个优选实施方案中，本发明通过产生多种 GVT-对鉴定靶基因组中的精细结构变化，所述 GVT-对为限定的空间距离和方向的独特基因组位置鉴定物。GVT-对共同地代表受试者的基因组分析，然后将该基因组分析与参比序列对比，或与类似产生的其它靶基因组的基因组分析对比，表明在核酸群之间存在精细结构差异。本发明可检测的基因组精细结构变化包括：缺失和插入、复制、翻转、易位和其它染色体重排。本发明提供以用户限定的、取决于实验设计的分辨率水平鉴定这些基因组特征的方法。

[0059] 假定 4 个碱基均一分布，本发明的 18 bp 或 27 bp 的 GVT 应碰巧分别平均每 4^{18} 和 4^{27} 个碱基出现 1 次，并应代表在人和其它复杂基因组中的独特序列标识符(unique sequence identifiers)。在考虑 GVT 之间的间隔距离时，GVT 对基因组的明确分配变好。例如，由大小分级分离的靶 DNA 群产生的、含有两个空间连接的 18 bp GVT 的 GVT-对是有效的 36 bp 序列标签。同样，连接的 27 bp GVT 对的功能是 54 bp 序列标签。不管标签长度，应当不可能将非常小的 GVT 或 GVT-对组分配至独特的基因组位置，例如完全处于重复元件中的那些位置。预期本发明可撤销分析的基因组区域很小，可通过本领域已知的计算机方法建模。

[0060] 在每个 GVT-对单体上存在的通用构架序列允许由高通量测序数据明确的提取 GVT-对序列。使用 MEGABLAST (Zhang 等, *J Comput Biol* 7:203 (2000))或类似的计算机程序通过比对揭示 GVT-对

与一个或多个参比序列之间的不一致性。在阈值水平内 GVT-对间隔距离或方向与参比的不一致性预示在靶和参比 DNA 之间存在结构差异。阈值水平由实验设计确定，相对于平均 GVT 间隔距离的两个标准偏差为合理的默认值。在与参比序列相比时，靶 DNA 中的缺失可由 2 个或更多个 GVT-对限定，所述 GVT-对跨越平均间隔距离的 2 个标准偏差以上。因此，靶 DNA 中的插入片段可被限定为这样的位点：其中在与参比序列相比时，两个或更多个 GVT-对跨越平均间隔的两个标准偏差以下。在靶 DNA 中的翻转被定义为这样的位点：其中二个或更多个 GVT-对的 GVT 方向不一致。人工维护(curate)和评价不一致的 GVT-对，之后继续通过 PCR、DNA 印迹杂交分析或通过插入片段分离和测序来验证。

[0061] 本发明的靶基因组核酸可来源于任何来源，包括：真核生物、原核生物、微生物、质体和病毒的基因组 DNA。本发明的靶基因组核酸还可以来源于生物的 RNA 基因组，例如通过逆转录过程将 RNA 转变为 DNA 的 RNA 病毒。用于研究的靶核酸的选择可受到在科学文献中描述的特定染色体或染色体区域与某些病症相关的先有知识影响。本发明可利用来自分离的染色体或染色体区域的靶 DNA。本发明可以适于研究设计的分辨率范围用于广泛的全基因组范围的患者队列扫描。纯化染色体、染色体区段、基因组 DNA 和 RNA 的方法是本领域已知的。本领域还已知通过 PCR 或通过其它方法扩增核酸的方法，以产生经由本发明进行分析的靶 DNA。

[0062] 在本文公开内容的较早部分描述了切割靶 DNA 和分级分离靶 DNA 至需要大小的方法，用于确定 GVT-对的 GVT 之间的空间距离。用屡次切割的酶动力学剪切或部分酶促消化 DNA 可用于产生具有高度重叠片段的 DNA 片段群，用于最大化覆盖靶 DNA 的每个区。或者，可用几种限制性内切核酸酶在单独的切割反应中完全消化靶 DNA，然后大小分级分离至用于 GVT-对生产所需要的大小类别。由用单一限制性内切核酸酶完全消化制备的、选择过大小的靶 DNA 产

生的 GVT-对是不重叠的，仅覆盖了一部分靶 DNA 复杂性。用其它限制性内切核酸酶完全酶促消化获得的、选择过大小的 DNA 片段可用于覆盖空位。随机地或与完全酶促消化组合切割靶 DNA，以覆盖给定复杂性的基因组，此切割可由本领域技术人员通过计算机方法建模，以取得使资源得到最佳利用的研究设计。诸如 BamH I、Hind III、Pst I、Spe I 和 Xba I 的酶对 CpG 甲基化不敏感，并应在每个位点切割哺乳动物基因组 DNA，以产生精确地代表那些酶的邻近识别位点对的 GVT-对。对 CpG 甲基化、重叠 CpG 甲基化或可影响本发明的核酸分析的其它种类的 DNA 修饰的作用不敏感的其它适宜的酶已由文献 (May 等, *J Bacteriol 123*:768, (1975); Hattman 等, *J Mol Biol 126*:367, (1978); Buryanov 等, *FEBS Letters 88*:251, (1978); Geier 等, *J Biol Chem 254*:1408, (1979); Kan 等, *J Mol Biol 130*:191, (1979); McClelland 等, *Nucleid acid Res 22*:3640, (1994))和主要的限制性内切核酸酶供应商 (Fermentas, Hanover, MD; New England Biolabs, Ispwich, MA)描述。在某些实施方案中，其靶 DNA 的切割对 DNA 修饰敏感的酶可用于分界靶 DNA 中的修饰位点。例如，本发明可鉴定已知调节基因表达的 DNA 甲基化位点。对于该应用，用甲基化敏感的限制性酶完全消化靶 DNA，并由消化的 DNA 产生 GVT-对。通过所获 GVT-对在与参比序列上的邻近限制性位点相比时的不一致性鉴定甲基化位点。

[0063] 首先人工维护不一致的 GVT-对，之后进行一系列的分级过滤，以便检验。在其中不一致的 GVT-对由来源于完全限制性内切核酸酶消化的、选择过大小的 DNA 产生的情况下，采用相同限制性内切核酸酶消化的靶 DNA 和参比 DNA 的 DNA 印迹分析可用于验证靶 DNA 和参比 DNA 之间的标记距离的差异。GVT 的长度足以用作 PCR 引物，以便分离间插基因组序列进行鸟枪法测序，以确定结构变化的精确性质。

[0064] 一般认为，结构变化的研究将进一步阐明复杂疾病，例如肥胖和糖尿病，这些疾病的发展由基因、基因元件和环境的相互作用

触发。本发明分析的核酸的选择可受到在科学文献中描述的特定染色体或染色体区域与某些病症相关的先有知识的影响。本发明可以高分辨率针对来自分离的染色体或染色体区域或组织样品的 DNA。或者，本发明可以适于研究设计的分辨率范围用于广泛的全基因组范围的患者队列扫描。现行的 fosmid 配对末端测序技术需要成百万的序列读数来以中等的分辨率和覆盖率水平分析每个个体，由此限制了其作为平台扫描大群体的应用，所述大群体用于关联研究，以发现对疾病结果为诊断性或预后性的生物标记以及为用于药物干预的潜在药物靶的生物标记。本发明提供了这些限制的解决方法，因此，本发明具有产生新的药物诊断方法和帮助药物发现的潜力。

[0065] 在另一个优选实施方案中，本发明鉴定的精细结构变化用于设计寡核苷酸阵列测定、微阵列测定、基于 PCR 的测定和本领域中的其它诊断测定，以检测核酸群之间的差异。本发明的微阵列和寡核苷酸阵列是用于检测核酸拷贝数改变以及单个或少数核苷酸多态性的有效平台，但不适于检测可能导致或引起疾病的其它基因组改变。本发明的鉴定产物能够设计寡核苷酸和微阵列测定和本领域的其它诊断测定，以筛选分界本发明鉴定的精细结构变化的易位、插入、缺失和翻转接合处。这些测定然后可用于筛选一般群体和大的患者队列，以确定精细结构变化在复杂疾病中的作用，所述疾病例如为肥胖、糖尿病和许多癌症，这些疾病的发展由多种遗传和环境因素的相互作用触发。这些测定的其它应用包括但不限于诊断或区分在医学诊断和工业微生物领域中使用的微生物的密切相关的物种、品系、种族或生物型。

[0066] 在另一个优选实施方案中，本发明用于产生高分辨率基因组图谱，以帮助由鸟枪法 DNA 测序进行基因组组装。限定间隔距离或邻近限制性内切核酸酶位点的广泛的独特遗传标记组通过提供用于基因组组装的骨架应极大促进全基因组测序工作。预期本发明产生的与人类基因组组装的当前版本(35 版, 2004 年 5 月)不一致的大量

GVT-对实际上可能不代表靶DNA的精细结构变化，而是反映了当前人类基因组组装中的错误或空位。使问题更复杂的是现行的基因组组装来源于合并的多个供体的DNA。需要来源于单个个体的、代表人类多样性范围的参比序列，以推动基因组领域前进。本发明提供的用途提供了实施此工作的方法。

[0067] 在另一个优选的实施方案中，本发明用于产生高分辨率的基因组图谱，以利于系统发生研究和测定密切相关的生物之间的遗传和功能关联。本发明的一个方面尤其适于该用途，这方面利用由靶DNA产生的GVT-对，所述靶DNA用单独的或者在用于GVT-对生产的有用组合中的一种或多种限制性内切核酸酶完全消化，没有 DNA 大小分级步骤。基本上，如此产生的 GVT-对构成了含有位置标记对的基因组分析，所述位置标记沿着靶 DNA 长度分界邻近的限制性内切核酸酶位点。GVT-对的鉴定及其相对丰度可用于产生高分辨率基因组分析，该基因组分析可用于鉴定、区分和定量复杂医学或环境DNA分离物中的原始基因组。如此产生的 GVT-对还可用于工业微生物领域，用于鉴定引起期望性状的基因组差异，例如在密切相关的品系、生物型或种族或遗传修饰的生物中有利的生长速率和生产有用的次级代谢物和重组蛋白。因此，本发明可用作工具，以在微生物来源产物的工业化生产中帮助改良菌株。本发明产生的高分辨率基因组图谱还提供了低成本和有效的方法来研究密切相关的病原体核酸，以鉴定变化区域，以这些区域为目标进行详细的序列分析，以鉴定可用于诊断和作为医学干预的药物靶的病原决定因素。

[0068] 在另一个优选实施方案中，本发明可用于遗传解剖家畜和农业作物的表型多样性，以利于标记辅助性育种。家畜特别令人有兴趣进行复杂遗传元件的鉴定，所述遗传元件有助于控制生长、能量代谢、发育、机体组成、生育和行为以及通过经典育种研究的其它性状。关于综述参见 Andersson (*Nat Rev Genet 2*:130 (2001))。大部分目标农业性状是多因素的，经常受未知数量的数量性状基因座(QTL)控制。

基因组扫描的微卫星图谱已被开发用于大部分家畜。使用这些标记的相关研究和候选基因方法是用于鉴定 QTL 的两种主要策略。QTL 的克隆具有挑战性，因为基因型和表型之间的关联被认为比单基因性状更复杂。然而，有可能通过后代测验间接确定 QTL，其中 QTL 的分离使用由子代之间的遗传标记和表型变化获得的数据来推断。目前，大部分 QTL 的分子基础仍是未知的。果蝇中的 QTL 作图提示，QTL 经常与非编码区中的序列变化相关(MacKay *Nat Rev Genet* 2:11 (2001))。如在人中一样，预期家畜和作物基因组中的精细结构变化在表型表达以及基因组与环境的相互作用方面可能起重要作用。本发明提供以低成本将家畜和作物中的广泛范围的基因组结构多样性制表的方法。然后，制表的信息应能够产生寡核苷酸微阵列和其它诊断平台，用于关联和连锁研究，以鉴定和表征导致标记辅助育种的实际 QTL。

[0069] 作为主要的传粉者，蜜蜂在农业当中和世界上的许多地区起关键作用。养蜂是由本发明获益的另一个领域。蜜蜂是一种在经济上重要的物种，适于在育种发育中使用遗传技术。蜜蜂传代时间短，产生大量子代。家系还容易通过人工授精增殖。蜜蜂品系在生育、抗病性和行为性状方面表现出广泛的表型变化，其中许多处于复杂的遗传控制之下。处于遗传控制之下的重要行为性状包括：以许多非洲品系为代表的攻击、觅食习性、产蜜量和所谓的“卫生”行为。“卫生”性状由至少 7 个至今还没确定的基因座调节，这些基因座合在一起导致蜂房成员去除死亡或患病群体的清洁行为，作为抵御 funal 和小虫侵袭的主要防御，funal 和小虫是两种主要的经济性蜜蜂病原体。主要目标是开发可信赖的诊断分子标记，这些标记可用于标记辅助育种，以快速有效地鉴定需要的子代品系，而不需要复杂且耗时的育种试验和大田试验。本发明可使用意大利蜂(*Apis mellifera*)品系 DH4 的 200 兆碱基大小基因组的遗传图谱和参比序列(The Honeybee Genome Sequencing Consortium *Nature 443*:931, (2006))来提供有效且低成本的

方法，以高分辨率研究多个蜜蜂品系基因组的精细结构变化，从而关联期望的表型和基因型。成本有效地研究多个品系的能力是本发明提供的关键优势。例如，以 10 kb 分辨率窗 5 倍覆盖 200 兆碱基的蜜蜂基因组应仅需要 10,000 轮测序和 2,500 个测序模板制备物。成本估计基于每轮测序 10 个寡聚化的 GVT-对的序列测定结果以及每个载体模板支持 4 个独立的测序反应。

[0070] 在另一个优选实施方案中，本发明可用于鉴定神经疾病和性状的基础性遗传病因。一般认为，许多神经障碍(如孤独症、双相型障碍和精神分裂症)的至少一种组分具有复杂的非孟德尔遗传组分(Holzman 和 Matthysse, *Psycholology Sci 1*:270 (1990); Owen 和 Craddock, *Mol Psychiatry 1*: 21 (1996); Craddock 和 Jones, *Br J Psychiatry 178*:s128 (2001))。互补连锁和相关性研究目前用于鉴定基因组组分，本发明提供了评价基因组精细结构变化在神经疾病中的促进性作用的方法，并可以产生用于诊断、预后和患者管理的新方法。

[0071] 在另一个优选实施方案中，本发明可用于鉴定癌症的基础性遗传病因，由此产生用于诊断、预后和治疗干预的方法。实际上，所有的癌症都是缘于 DNA 序列的异常性，这些异常性或者是固有的，或者是通过生命当中的体细胞突变获得的。肿瘤生成的主要原则在于，累积的 DNA 突变与环境因素一起改变了基因表达，或者基因功能越过了允许克隆扩增、细胞侵入周围组织和启动转移的关键功能阈。在西方国家有 1/3 的人将出现癌症，1/5 将死亡，这使癌症称为最常见的遗传疾病。在历史上，该领域以鉴定有效的癌症或肿瘤抑制基因开始，其中由于基因座的少量核苷酸改变而简单失去或获得功能是癌症的主要促成因素。该领域后来扩展到基因剂量，其中导致基因拷贝数改变的 DNA 区段的复制或缺失是癌症发生的推测病因。应用阵列 CGH 对检测 DNA 拷贝数的改变以及癌细胞系和原发性肿瘤的杂合性的丧失特别有用。癌症中的拷贝数分析的全面综述和癌症中的体细胞突变目录以及其中的参考文献可见于桑格研究所的“癌症基因组计

划”(http://www.sanger.ac.uk/genetics/CGP/)。

[0072] 最近，知晓了基因组精细结构变化在癌症发生中的重要作用。在癌症发生过程中，肿瘤基因组累积了大量重排，包括扩增、缺失、易位、翻转等，其中许多直接促成肿瘤发展(Gray 和 Collins, *Carcinogenesis 21*: 443 (2000))。Volik 等(*Genome Research 16*: 394 (2006))利用 fosmid 配对末端作图的功能变化，以检测发展中的肿瘤的基因组结构的所有改变，尤其是不能通过阵列 CGH 检测的易位和翻转事件。他们的解析乳癌基因组的方法是最多信息的，但被研究者公认受限于获得每个样品的大量 BAC 克隆的末端终止序列所需要的费用和资源。本发明提供低成本的、高分辨率的方法来克服这些缺陷，鉴定不适于通过阵列 CGH 检测的基因组精细结构变化。本发明具有足够低的成本，能够用于广泛的癌症患者队列研究，能够用于跟踪个体患者的肿瘤发展中的基因组变化累积。跟踪肿瘤发生过程中的基因组变化的能力在临床结果上应具有意义深远的预测价值，提供了患者管理的显著改善。

[0073] 在又一个优选实施方案中，本文所述方法可用于鉴定 mRNA 加工变体。一个基因编码一个蛋白的概念被一个基因编码多个蛋白取代，其中一些蛋白具有在医学上相关的不同功能。该过程似乎是高度可调的，部分通过 mRNA 的可变加工以及启动子、转录终止子和翻译后加工的不同用途来介导。其中两个不同 mRNA 转录物重组的反式剪接的过程又增加了转录组复杂性。所用靶 mRNA 的选择可受到其中某些 mRNA 变体可能很重要的某些疾病情形、细胞类型、器官或发育阶段的先有知识的影响。

[0074] 本领域技术人员熟知用于 mRNA 分离和将 mRNA 转变为 cDNA 的方法。在本发明的一方面中，通过逆转录或逆转录与 PCR 偶联将分离的 RNA 转变为 cDNA，所述 PCR 利用的方法包括使用随机引物，所随机述引物含有限制性内切核酸酶，例如 Mme I、CstM I、NmeA III 或 EcoP15 I。限制性位点位于引物上，使得用所述内切核酸

酶消化所获的双链cDNA去除了cDNA中的引物序列。调节引物浓度，以产生300-500 bp平均大小的产物，或符合实验设计大小的产物。在使用T_4 DNA聚合酶修复cDNA末端后，cDNA被去磷酸化，连接至合适的 GVT-连接物，并在 5%丙烯酰胺凝胶上选择大小，用于生产GVT-对。鉴定mRNA加工变体的GVT-对与NCBI参比序列(RefSeq)或其它数据库的不一致性。加工变体通过使用来源于不一致的 GVT-对的引物的PCR证实。