

[19] 中华人民共和国国家知识产权局

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



[12] 发明专利申请公布说明书

[21] 申请号 200880008027.8

[43] 公开日 2010 年 3 月 31 日

[11] 公开号 CN 101688237A

[22] 申请日 2008.1.10

[21] 申请号 200880008027.8

[30] 优先权

[32] 2007.1.11 [33] US [31] 60/999,750

[32] 2007.10.5 [33] US [31] 60/977,900

[86] 国际申请 PCT/IB2008/000625 2008.1.10

[87] 国际公布 WO2008/084405 英 2008.7.17

[85] 进入国家阶段日期 2009.9.11

[71] 申请人 伊拉兹马斯大学医疗中心

地址 荷兰鹿特丹

[72] 发明人 沃特·德拉特

弗兰克·格罗斯维尔德

[74] 专利代理机构 北京市柳沈律师事务所

代理人 封新琴

权利要求书 7 页 说明书 97 页 附图 41 页

[54] 发明名称

环状染色体构象捕获 (4C)

[57] 摘要

本发明的一个方面涉及分析靶核苷酸序列与一个或多个目的核苷酸序列(如一个或多个基因组基因座)相互作用的频率的方法,其包括如下步骤:

(a)提供交联的 DNA 样品;(b)用第一限制酶消化交联的 DNA;(c)连接交联的核苷酸序列;(d)解除交联;(e)任选用第二限制酶消化核苷酸序列;(f)任选将已知核苷酸组成的一个或多个 DNA 序列与在一个或多个目的核苷酸序列侧翼的可用的第二限制酶消化位点连接;(g)利用至少两个寡核苷酸引物扩增一个或多个目的核苷酸序列,其中每个引物与在目的核苷酸序列侧翼的 DNA 序列杂交;(h)将扩增的序列与阵列杂交;和(i)确定 DNA 序列间相互作用的频率。

1. 分析靶核苷酸序列与一个或多个目的核苷酸序列(如一个或多个基因组基因座)相互作用的频率的方法, 其包括以下步骤:

(a)提供交联 DNA 的样品;

(b)用第一限制酶消化交联的 DNA;

(c)连接交联的核苷酸序列;

(d)解除交联;

(e)任选用第二限制酶消化核苷酸序列;

(f)任选将核苷酸组成已知的一个或多个 DNA 序列与在一个或多个目的核苷酸序列侧翼的可用的第二限制酶消化位点连接;

(g)利用至少两个寡核苷酸引物扩增一个或多个目的核苷酸序列, 其中每个引物与目的核苷酸序列侧翼的 DNA 序列杂交;

(h)将扩增的序列与阵列杂交; 并

(i)确定 DNA 序列间相互作用的频率。

2. 根据权利要求 1 所述的方法, 其中步骤(c)或(f)中的连接反应导致 DNA 环的形成。

3. 根据权利要求 1 或权利要求 2 所述的方法, 其中步骤(h)包括用测序方法(如高通量测序)分析靶序列和目的交联序列之间的连接产物。

4. 根据前述权利要求之任一项分析两个或更多个靶核苷酸序列与一个或多个目的核苷酸序列相互作用的频率的方法, 其包括在步骤(g)中使用多重 PCR。

5. 根据前述权利要求之任一项分析两个或更多个靶核苷酸序列与一个或多个目的核苷酸序列相互作用的频率的方法, 其包括在步骤(g)中为每个靶序列汇集部分或全部获得的 PCR 产物并随后同时分析它们的 DNA 相互作用。

6. 根据权利要求 5 所述的方法, 其中在汇集和通过与阵列杂交进行分析之前将两个或更多个扩增的序列不同地标记。

7. 根据权利要求 5 或权利要求 6 所述的方法, 其中当两个或更多个扩增的序列处在不同染色体上时将所述序列相同地标记并通过与阵列杂交进行分析。

8. 根据权利要求5所述的方法,其中当两个或更多个扩增的序列处在同一染色体上足够远的距离从而使DNA-DNA相互作用信号之间的重叠最小化时将所述序列相同地标记。

9. 根据前述权利要求之任一项所述的方法,其中将高通量测序用于分析靶序列与捕获的目的序列之间形成的连接接头。

10. 根据权利要求9所述的方法,其中通过向扩增序列的末端添加测序需要的衔接序列而将测序定向于靶序列与捕获的目的序列之间形成的连接接头。

11. 根据权利要求10所述的方法,其中通过向用于扩增一个或多个目的核苷酸序列的寡核苷酸引物添加测序需要的完整或部分衔接序列作为5'突出端而将测序定向于靶序列与捕获的目的序列之间形成的连接接头。

12. 根据权利要求10所述的方法,其中通过将生物素物质或其它模块与用以扩增一个或多个目的核苷酸序列的寡核苷酸引物缀合,随后是链霉抗生物素蛋白或其它介导的对PCR扩增物质的纯化而将测序定向于靶序列与捕获的目的序列之间形成的连接接头。

13. 根据权利要求9到12之任一项所述的方法,其中通过设计用于扩增在距所分析的第一和/或第二限制酶识别位点400、300、200、150、100、90、80、70、60、50、40、30、20、10、9、8、7、6、5、4、3、2、1核苷酸内的一个或多个目的核苷酸序列的寡核苷酸引物而将测序定向于靶序列与捕获的目的序列之间的连接接头。

14. 根据权利要求9到12之任一项所述的方法,其中通过设计用于扩增一个或多个目的核苷酸序列的寡核苷酸引物从而使它们部分地或完全地与所分析的第一和/或第二限制酶识别位点重叠而将测序定向于靶序列与捕获的目的序列之间的连接接头。

15. 根据权利要求9到14之任一项所述的方法,其中序列读过连接接头从而当分析多重或汇集的PCR样品时,在连接接头的每一侧获得充分的序列信息(例如12个核苷酸或更多)以明确地鉴定每个靶序列和每个捕获的目的序列。

16. 根据前述权利要求之任一项所述的方法,其中靶核苷酸序列选自由基因组重排、启动子、增强子、沉默基因、隔离子、基质附着区、基因座控制区、转录单位、复制起点、重组热点、易位断点、着丝粒、端粒、基因密

集区、基因稀少区、重复元件和(病毒)整合位点组成的组。

17. 根据前述权利要求之任一项所述的方法, 其中靶核苷酸序列是与疾病相关的或造成疾病的核苷酸序列, 或在线性 DNA 模板上位于距离与疾病相关的或造成疾病的基因座多至 15Mb 或大于 15Mb 处。

18. 根据前述权利要求之任一项所述的方法, 其中靶核苷酸序列选自下组: *AML1*、*MLL*、*MYC*、*BCL*、*BCR*、*ABL1*、*IGH*、*LYL1*、*TAL1*、*TAL2*、*LMO2*、*TCR α/δ* 、*TCR β* 和 *HOX* 或其它与疾病相关的基因座, 所述其它与疾病相关的基因座描述于“Catalogue of Unbalanced Chromosome Aberrations in Man”第 2 版. Albert Schinzel. 柏林: Walter de Gruyter, 2001. ISBN 3-11-011607-3 中。

19. 根据前述权利要求之任一项的方法, 其中靶序列沿着线性基因组模板分布从而使相互作用的序列覆盖整个染色体或基因组。

20. 根据前述权利要求之任一项所述的方法, 其中第一限制酶是识别 6-8 bp 识别位点的限制酶。

21. 根据权利要求 20 所述的方法, 其中第一限制酶选自由 *Bgl*III、*Hind*III、*Eco*RI、*Bam*HI、*Spe*I、*Pst*I 和 *Nde*I 组成的组。

22. 根据权利要求 20 或权利要求 21 所述的方法, 其中第一限制酶是基于其在重复序列中不存在或不具有代表性而选择的。

23. 根据前述权利要求之任一项所述的方法, 其中第二限制酶是识别 4 或 5 bp 核苷酸序列识别位点的限制酶。

24. 根据前述权利要求之任一项所述的方法, 其中第二限制酶识别位点在靶核苷酸序列中位于距离第一限制酶位点大于约 350bp 处。

25. 分析靶核苷酸序列与一个或多个核苷酸序列(如一个或多个基因组基因座)相互作用的频率的方法, 其包括以下步骤:

- (a)提供交联 DNA 的样品;
- (b)用第一限制酶消化交联的 DNA;
- (c)连接交联的核苷酸序列;
- (d)解除交联;
- (e)任选地用第二限制酶消化核苷酸序列;
- (f)环化核苷酸序列;
- (g)扩增与靶核苷酸序列连接的一个或多个核苷酸序列;

(h)任选地将扩增的序列与阵列杂交或通过测序(如高通量测序)分析扩增的序列; 并

(i)确定 DNA 序列间相互作用的频率。

26. 鉴定一种或多种对特定疾病状态起指示作用的 DNA-DNA 相互作用的方法, 其包括执行权利要求 1-24 的步骤(a)–(i)的步骤, 其中在步骤(a)中由患病和未患病的细胞提供交联 DNA 的样品, 而且其中来自患病和未患病的细胞的 DNA 序列间相互作用频率之间的差异表明染色体模板线性组构的差异(如基因组重排), 其指示特定性状或疾病状态。

27. 诊断或预测由 DNA-DNA 相互作用的变化造成的或与 DNA-DNA 相互作用的变化相关的疾病或综合征的方法, 其包括执行权利要求 1-24 之任一项的步骤(a)–(i)的步骤, 其中步骤(a)包括提供来自受试者的交联的 DNA 样品; 而且其中步骤(i)包括将 DNA 序列间相互作用的频率与未受影响的对照进行比较; 其中得自对照的值和得自受试者的值之间的差异指示受试者正罹患所述疾病或综合征或指示受试者将患上所述疾病或综合征。

28. 根据权利要求 27 所述的方法, 其中相互作用频率从低向高转变指示平衡的和/或不平衡的遗传重排的位置。

29. 根据权利要求 27 所述的方法, 其中受试者样品的 DNA-DNA 相互作用频率与对照相比呈倒转模式指示平衡的和/或不平衡的倒位。

30. 根据权利要求 27 所述的方法, 其中受试者样品的 DNA-DNA 相互作用频率与对照相比的降低与更远区域的 DNA-DNA 相互作用频率的增加的组合指示平衡的和/或不平衡的缺失。

31. 根据权利要求 27 所述的方法, 其中受试者样品的 DNA-DNA 相互作用频率与对照相比的增加或减少指示平衡的和/或不平衡的重复或插入。

32. 根据权利要求 27-31 之任一项所述的方法, 其中在进行所述方法前使用光谱核型分析和/或 FISH。

33. 根据权利要求 27-32 之任一项所述的方法, 其中所述疾病是遗传疾病。

34. 根据权利要求 27-33 之任一项所述的方法, 其中所述疾病是癌症。

35. 诊断或预测由 DNA-DNA 相互作用的变化造成的或与 DNA-DNA 相互作用的变化相关的疾病或综合征的方法, 其包括以下步骤: 进行权利要求 1-24 之任一项的步骤(a)–(i), 其中步骤(a)包括提供来自受试者的交联 DNA

的样品；而且其中所述方法包括额外步骤：(j)鉴定一个或多个经历了与疾病相关的基因组重排的基因座。

36. 鉴定一种或多种调节 DNA-DNA 相互作用的试剂的测定方法，其包括以下步骤：

(a)将样品与一种或多种试剂接触；和

(b)进行权利要求 1-24 之任一项的步骤(a)至(i)，其中步骤(a)包括提供来自样品的交联的 DNA；

其中(i)在存在所述试剂的情况下的 DNA 序列间相互作用的频率和(ii)在所述试剂不存在的情况下的 DNA 序列间相互作用的频率之间的差异指示试剂调节 DNA-DNA 相互作用。

37. 检测平衡的和/或不平衡的重排(如易位)的位置的方法，其包括以下步骤：

(a)进行权利要求 1-24 之任一项的步骤(a)至(i)；和

(b)将 DNA 序列间相互作用的频率与对照的频率进行比较；

其中样品的 DNA-DNA 相互作用频率与对照相比从低至高的转变指示断点的位置。

38. 检测平衡的和/或不平衡的倒位的位置的方法，其包括以下步骤：

(a)进行权利要求 1-24 之任一项的步骤(a)至(i)；和

(b)将 DNA 序列间相互作用的频率与对照的频率进行比较；

其中样品的 DNA-DNA 相互作用频率与对照相比呈倒转模式指示倒位。

39. 检测缺失的位置的方法，其包括以下步骤：

(a)进行权利要求 1-24 之任一项的步骤(a)至(i)；和

(b)将 DNA 序列间相互作用的频率与对照的频率进行比较；

其中样品的 DNA-DNA 相互作用频率与对照相比的降低指示缺失。

40. 检测重复的位置的方法，其包括以下步骤：

(a)进行权利要求 1-24 之任一项的步骤(a)至(i)；和

(b)将 DNA 序列间相互作用的频率与对照的频率进行比较；

其中受试者样品的 DNA-DNA 相互作用频率与对照相比的增加或减少指示重复或插入。

41. 根据前述权利要求之任一项的方法，其中扩增与两个或更多个靶序列相互作用的核苷酸序列。

42. 根据权利要求 41 所述的方法, 其中靶序列位于或靠近已知与疾病状态相关的基因组基因座。

43. 根据权利要求 42 所述的方法, 其中在对重排位置无现有知识的情况下选择靶序列并且使靶序列间隔开从而使相互作用的序列覆盖整个染色体或基因组, 并且其中鉴定的相互作用序列允许重建线性染色体图和发生在染色体与染色体间的基因组重排。

44. 根据权利要求 41 到 43 之任一项所述的方法, 其中扩增的序列是标记的。

45. 根据权利要求 44 所述的方法, 其中扩增的序列根据它们在基因组中的位置而不同地标记。

46. 根据权利要求 41 到 45 之任一项所述的方法, 用于检测平衡的和/或不平衡的重排、易位、倒位、缺失、重复或插入。

47. 通过根据权利要求 35 的测定方法得到或可得到的试剂。

48. 分析一个或多个靶核苷酸序列与一个或多个目的核苷酸序列(如一个或多个基因组基因座)相互作用的频率的方法, 其包含以下步骤:

- (a)提供交联 DNA 的样品;
- (b)用第一限制酶消化交联 DNA;
- (c)连接交联的核苷酸序列;
- (d)解除交联; 和
- (e)对连接的核苷酸序列测序。

49. 确定样品中存在基因组重排的方法, 其包括以下步骤:

(a)提供核酸样品(如基因组 DNA), 其中所述核酸在邻近疑似基因组重排的位置处包含已知序列的核苷酸序列;

- (b)用第一限制酶消化 DNA 以形成多个限制性片段;
- (c)任选地, 纯化限制性片段;
- (d)连接限制性片段以形成环化的 DNA;
- (e)任选地, 纯化环化的 DNA;
- (f)用第二限制酶消化环化的 DNA 以形成多个限制性片段;
- (g)连接限制性片段以形成环化的 DNA;

(h)用一个或多个与已知序列的核苷酸序列杂交的引物扩增疑似的基因组重排; 和

(i)对疑似的基因组重排进行测序。

50. 根据权利要求 1 到 46 之任一项所述的方法,其中阵列杂交步骤被测序步骤代替。

51. 根据权利要求 48 或 49 所述的方法,其中靶核苷酸序列和目的核苷酸序列都通过测序鉴定。

52. 根据权利要求48到51之任一项所述的方法,其中将衔接序列与PCR产物连接。

53. 约 6-50 个碱基对的核酸序列的数据库,所述核酸序列直接位于每个靶序列的第一限制酶识别位点或第二限制酶识别位点的侧翼,和任选包括每个靶序列的第一限制酶识别位点或第二限制酶识别位点。

54. 约 12-50 个碱基对的核酸序列的数据库,所述核酸序列直接位于基因组中所有相关的第一和第二限制酶识别位点的侧翼。

55. 权利要求 53 或权利要求 54 中的核酸序列数据库用于确定所鉴定的每个捕获序列的基因组位置的用途。

56. 基本上如本文描述的和参考任意实施例或附图的方法或试剂或数据库或用途。

环状染色体构象捕获(4C)

发明领域

本发明涉及分析两个或更多个核苷酸序列在核空间中相互作用的频率。利用相互作用的改变作为检测基因组重排的手段用于诊断和预测。

发明背景

对哺乳动物核结构的研究目的在于理解2米长的DNA如何折叠成10 μ m直径的核，同时允许精确表达指定细胞类型的基因，以及在每个细胞周期中如何如实地增殖。在该领域的进展大多来自于显微镜研究，其揭示了基因组非随机地排列在核空间中。例如，紧密包装的异染色质与更加展开的常染色质分离，染色体在核空间中占据不同的区域。核定位和转录活性间存在复杂的关系。尽管转录的发生遍及核内部，但在染色体上簇集的活性基因优先位于它们的染色体区域的边界或外侧。如针对相对较大的核界标(如染色体区域、着丝粒或核周缘)所测得的，独立的基因在它们的转录状态改变后可能迁移。而且如最近对 β -球蛋白基因座和一些选择的其它基因进行的荧光原位杂交(fluorescence in situ hybridization, FISH)所证实的，在染色体上相距上千万个碱基的活性转录基因可在核中集合到一起。除了转录，基因组组构还与协调复制、重组和基因座转座的可能性(其能导致恶性肿瘤)和外遗传程序(epigenetic program)的设置和重置有关。根据这些观察结果，认为细胞核中DNA结构性组构对基因组功能是关键的因素。

开发出了不同的测定法可以在体内洞察基因组基因座的空间组构。已经开发出了一种称为RNA-TRAP的测定法(Carter等.(2002)Nat. Genet. 32, 623)，其涉及将辣根过氧化物酶(HRP)靶向于新生的RNA转录物，然后定量HRP催化的生物素在附近染色质上的沉积。

另一种已开发出的测定法被称为染色体构象捕获(3C)技术，其提供了一种研究基因组区域的结构性组构的工具。3C技术涉及用定量PCR分析两个给定DNA限制性片段间的交联频率，其能测量它们在核空间中的接近度(参见图1)。该技术原来被开发用来分析酵母中染色体的构象(Dekker等., 2002)，

已经使该技术适用于研究复杂哺乳动物基因簇中基因表达和染色质折叠间的关系(参见, 诸如 Tolhuis 等., 2002; Palstra 等., 2003; 和 Drissen 等., 2004)。简而言之, 3C 技术涉及在体内用甲醛交联细胞并用限制酶对染色质进行核消化, 然后连接交联成一个复合物的 DNA 片段。然后用 PCR 定量连接产物。PCR 扩增步骤需要知道每个要扩增的 DNA 片段的序列信息。因此, 3C 技术提供了对所选 DNA 片段间相互作用频率的测量。

已经开发了 3C 技术来鉴定基因组的选定部分之间相互作用的元件, 两种技术都要求为所分析的限制性片段设计引物。最近, 已经开发了新的策略, 其允许在整个基因组中以无偏的方式筛选与所选 DNA 片段物理地相互作用的 DNA 区段。它们基于 3C 技术并被合称为“4C 技术”。4C 技术允许在整个基因组中以不偏好的方式筛选与所选 DNA 片段物理地相互作用的 DNA 片段。4C 技术基于交联的 DNA 片段与所选限制性片段(“诱饵”)的选择性连接。在 4C 技术中, 细胞群中所有被诱饵捕获的 DNA 片段是经由反向 PCR 同时扩增的, 该 PCR 使用两个从环化的连接产物扩增的诱饵特异性引物。

基本上有两种策略可用以获得这些 DNA 环。一种策略依赖标准 3C 连接步骤中环的形成, 即在 DNA 仍交联时(Zhao 等, (2006) *Nat Genet* **38**, 1341-7)。这里, 环的形成要求诱饵片段的两端连接到捕获的限制性片段的两端。如果多个限制性片段交联在一起, 环仍可形成, 但它们可以包含多于一个捕获的片段, 因此将更大。去交联之后, 捕获的 DNA 片段用朝外的诱饵特异性的引物通过反向 PCR 直接扩增。在这一设计中可以使用识别四个或六个碱基对的限制酶。而在此方法中 4-切酶(four-cutters)是优选的, 因为它们产生更小的限制性片段(平均大小 256bp, 相对于 6-切酶的~4kb), 而对捕获的 DNA 片段的线性 PCR 扩增要求平均产物大小是小的。因此本方法实质上包括以下步骤: (a)提供交联 DNA 的样品; (b)用第一限制酶-例如 4bp 切酶或 5bp 切酶消化交联的 DNA; (c)连接交联的核苷酸序列; (d)解除交联(reversing the cross linking)和(e)用至少两个寡核苷酸引物扩增一个或多个目的核苷酸序列, 其中每个引物与目的核苷酸序列两侧的 DNA 序列杂交。扩增的序列可以与阵列杂交以辅助确定 DNA 序列之间相互作用的频率。

第二种策略依赖如本文中和我们共同待审的申请 WO2007/004057 中所述的在染色质去交联之后的 DNA 环的形成。如其中所述, 4C 技术允许对与所选基因座相互作用的 DNA 片段进行无偏的基因组范围的搜索。简而言之,

3C 分析如常进行,但是省略了 PCR 步骤。3C 模板包括连接到许多不同的目的核苷酸序列(代表此基因的基因组环境)的靶序列或“诱饵”(如包含选定基因的所选限制性片段)。此模板由另一个第二限制酶切割然后再连接以形成小 DNA 环。有益的是,将连接到靶核苷酸序列的一个或多个目的核苷酸序列用至少两个寡核苷酸引物扩增,其中至少一个引物与靶序列杂交。优选地,第二个引物也与靶序列杂交,因而两个引物位于目的核苷酸的侧翼。或者,第二个引物与连接到第二限制性位点的衔接序列(adapter sequence)杂交,因而两个引物位于目的核苷酸的侧翼。这通常产生在独立扩增反应之间高度可重复且对特定组构特异性的 PCR 片段模式。*HindIII* 和 *DpnII* 可用作第一和第二限制酶。接下来,可将扩增的片段标记并任选地与阵列杂交,通常是与包含用相同的限制酶组合消化的基因组 DNA 的对照样品杂交。3C 技术因此得到改进从而使与靶核苷酸序列相互作用的所有目的核苷酸序列得到扩增。在实践中,这意味着不使用针对人们想要分析的片段有特异性的引物进行扩增反应,而是使用与目的核苷酸序列侧翼的 DNA 序列杂交的寡核苷酸引物进行扩增。有益的是,4C 对包含在 PCR 扩增步骤中的 PCR 引物的设计没有偏倚,因而能够用以在完整的基因组中搜索相互作用的 DNA 元件。

对于这样的高通量技术有重大需求,所述技术能以一种对于核空间中彼此接触的 DNA 基因座无偏的方式系统地筛选完整基因组。

而且,存在改善此技术的需求,其允许同时分析基因组中多个序列发生的多个相互作用,并且允许分析基因组中发生在未知位点的和可能与疾病相关的插入、缺失、异位、倒位和重排。

本发明寻求提供对 3C 和 4C 技术及其相关技术的改善。

发明简述

当前所用的 3C 技术仅能够分析有限数量的所选的 DNA-DNA 相互作用,这归咎于 PCR 扩增步骤的局限,其需要知道每个要分析的片段的特定序列信息。而且,选择限制性片段作为长程 DNA 相互作用的候选物需要对目的基因座大量的现有知识(如超敏感位点的位置),这通常是不可用的。考虑到迄今描述过的许多长距离 DNA-DNA 相互作用的功能相关性,对 DNA 元件与目的序列(如基因启动子、增强子、隔离子(insulator)、沉默基因、复制起点或 MAR/SAR)或目的基因组区域(如基因密集或基因稀少区或重复元

件)成环的随机筛选能力能极大方便对调节网络中所涉及的序列进行作图。

本发明涉及 4C 技术(即 capture and characterise colocalised chromatin(捕获和表征共定位的染色质))及其改进,其提供高通量分析核空间中两个或更多个核苷酸序列相互作用的频率的方法。4C 技术既可用于鉴定长程 DNA-DNA 相互作用(如研究染色体折叠)也可鉴定平衡的和不平衡的基因组重排-例如异位、倒位、缺失、扩增等-可能引起人受试者的性状或疾病。

通常 4C 技术涉及用微阵列分析由单一所选靶序列(“诱饵”)捕获的 DNA 片段(Simonis 等, *Nature Genetics* 2006)。微阵列具有缺点,即由于阵列上存在的探针可能饱和,所以它们提供的动态范围有限,这使信号强度的量化分析更加困难。高通量测序由于其提供无限的动态范围而回避了这一问题。它也是可量化的因为它提供序列的绝对数量。

另外,优选同时用多个靶序列分析 DNA 相互作用。对所有基于 4C 的应用,并且尤其是对于基于 4C 的基因组重排分析这都属实。4C 技术可用作诊断工具以允许以无偏的方式在整个基因组扫描基因组重排的存在。可以使用在每条染色体上共同捕获所有序列(即限制性片段)的一系列靶序列。随后,为了鉴定基因组重排,对哪个“诱饵”捕获的序列进行鉴定。

捕获的片段可以分别针对每个靶序列(“诱饵”)测序。但是优选地,同时分析所有靶序列形成的所有连接产物。为此,每次读取都需要定位于连接接头并提供足够的序列信息以明确地鉴定靶序列和捕获的序列。

本发明的简要方面

本发明的各方面在所附的权利要求书中提出。

在第一个方面,提供了分析靶核苷酸序列与一个或多个目的核苷酸序列(如一个或多个基因组基因座)相互作用的频率的方法,其包括以下步骤:(a)提供交联 DNA 的样品;(b)用第一限制酶消化交联的 DNA;(c)连接交联的核苷酸序列;(d)解除交联;(e)任选用第二限制酶消化核苷酸序列;(f)任选将已知核苷酸组成的一个或多个 DNA 序列与在一个或多个目的核苷酸序列侧翼的可用的一个或多个第二限制酶消化位点连接;(g)利用至少两个寡核苷酸引物扩增一个或多个目的核苷酸序列,其中每个引物与在目的核苷酸序列侧翼的 DNA 序列杂交;(h)将扩增的一个或多个序列与阵列杂交;和(i)确定 DNA 序列间相互作用的频率。

在第二个方面, 提供了分析靶核苷酸序列与一个或多个核苷酸序列(如一个或多个基因组基因座)相互作用的频率的方法, 其包括以下步骤: (a)提供交联 DNA 的样品; (b)用第一限制酶消化交联的 DNA; (c)连接交联的核苷酸序列; (d)解除交联; (e)任选用第二限制酶消化核苷酸序列; (f)环化核苷酸序列; (g)扩增与靶核苷酸序列连接的一个或多个核苷酸序列; (h)任选将扩增的序列与阵列杂交或用高通量测序来分析扩增的序列; 和(i)确定 DNA 序列间相互作用的频率。

在第三个方面, 提供了鉴定一种或多种指示特定疾病状态的 DNA-DNA 相互作用的方法, 其包括执行根据第一方面的方法的步骤(a)–(i)的步骤, 其中步骤(a)中交联 DNA 的样品由患病和未患病的细胞提供, 而且其中来自患病和未患病的细胞的 DNA 序列间相互作用的频率之间的差异表明染色体模板线性组构的差异(例如基因组重排), 其指示特定性状或疾病状态。

在第四个方面, 提供了诊断或预测由 DNA-DNA 相互作用中的变化造成的或与 DNA-DNA 相互作用中的变化相关的疾病或综合征的方法, 其包括执行根据第一方面的方法的步骤(a)–(i)的步骤, 其中步骤(a)包括提供来自受试者的交联 DNA 的样品; 而且其中步骤(i)包括将 DNA 序列间相互作用的频率与未受影响的对照的频率进行比较; 其中得自对照的值和得自受试者的值之间的差异指示受试者正罹患该疾病或综合征或指示受试者将患上该疾病或综合征。

在第五个方面, 提供了诊断或预测由 DNA-DNA 相互作用中的变化造成的或与 DNA-DNA 相互作用中的变化相关的疾病或综合征的方法, 其包括以下步骤: 进行依据第一个方面的方法的步骤(a)–(i), 其中步骤(a)包括由受试者提供交联 DNA 的样品; 而且其中所述方法包括额外步骤: (j)鉴定一个或多个经历了与疾病相关的基因组重排的基因座。

在第六个方面, 提供了鉴定一种或多种调节 DNA-DNA 相互作用的试剂的测定方法, 其包括以下步骤: (a)将样品与一种或多种试剂接触; 和(b)进行依据第一个方面的方法的步骤(a)至(i), 其中步骤(a)包括由样品提供交联的 DNA; 其中在(i)存在试剂的情况下 DNA 序列间相互作用的频率和(ii)无试剂的情况下 DNA 序列间相互作用的频率之间的差异指示该试剂调节 DNA-DNA 相互作用。

在第七个方面, 提供了检测平衡的和/或不平衡的重排(如易位)的位置的

方法,其包括以下步骤:(a)进行依据第一个方面的方法的步骤(a)至(i);和(b)将 DNA 序列间相互作用的频率与对照的频率进行比较;其中样品中 DNA-DNA 相互作用频率与对照相比从低至高的转变指示断点的位置。

在第八个方面,提供了检测平衡的和/或不平衡的倒位的位置的方法,其包括以下步骤:(a)进行依据第一个方面的方法的步骤(a)至(i);和(b)将 DNA 序列间相互作用的频率与对照的频率比较;其中样品的 DNA-DNA 相互作用频率与对照相比呈倒转模式指示倒位。

在第九个方面,提供了检测缺失位置的方法,其包括以下步骤:(a)进行根据第一个方面的方法的步骤(a)至(i);和(b)将 DNA 序列间相互作用的频率与对照的比较;其中样品的 DNA-DNA 相互作用频率与对照比较的降低指示缺失。

在第 10 个方面,提供了检测重复(duplication)的位置的方法,其包括以下步骤:(a)进行依据第一个方面的方法的步骤(a)至(i);和(b)将 DNA 序列间相互作用的频率与对照的频率比较;其中受试者样品的 DNA-DNA 相互作用频率与对照相比的增加或减少表明重复或插入。

在第 11 个方面,提供了由或可由本文所述的测定方法得到的试剂。

在第 12 个方面,提供了分析一个或多个靶核苷酸序列与一个或多个目的核苷酸序列(如一个或多个基因组基因座)的相互作用的频率的方法,其包括如下步骤:(a)提供交联 DNA 样品;(b)用第一限制酶消化所述交联的 DNA;(c)连接交联的核苷酸序列;(d)解除交联;和(e)对连接的核苷酸序列测序。

在第13个方面,提供确定样品中基因组重排存在的方法,其包括如下步骤:(a)提供核酸样品(如基因组DNA),其中所述核酸包含与疑似有基因组重排的位置相邻的已知序列的核苷酸序列;(b)用第一限制酶消化所述DNA,形成多个限制性片段;(c)任选地纯化限制性片段;(d)连接限制性片段以形成环化DNA;(e)任选地纯化环化DNA;(f)用第二限制酶消化环化DNA以形成多个限制性片段;(g)连接限制性片段以形成环化DNA;(h)用一个或多个与已知序列的核苷酸序列杂交的引物扩增疑似的基因组重排;和(i)对疑似的基因组重排测序。

在第14个方面,提供直接位于每个靶序列第一限制酶识别位点或第二限制酶识别位点侧翼或任选包括第一限制酶识别位点或第二限制酶识别位点的约6-50个碱基对的核酸序列的数据库。

在第 15 个方面, 提供了直接位于基因组中所有相关的第一和第二限制酶识别位点侧翼的约 12-50 碱基对的核酸序列的数据库。

在第 16 个方面, 提供了用于确定每个鉴定过的捕获序列的基因组位置的核酸序列的数据库的用途。

在第 17 个方面, 提供了基本上如本文所述的和参考任意实施例或附图的方法或试剂或数据库或用途。

发明的实施方案

适宜的是, 步骤(c)或(f)中的连接反应导致 DNA 环的形成。

适宜的是, 步骤(h)包括用测序的方法(例如高通量测序)分析靶序列和目的交联序列间的连接产物。

适宜的是, 所述方法用于用一个或多个目的核苷酸序列分析两个或更多个靶核苷酸序列相互作用的频率, 包括步骤(g)中多重 PCR 的使用。

适宜的是, 所述方法用于用一个或多个目的核苷酸序列分析两个或更多个靶核苷酸序列相互作用的频率, 包括汇集步骤(g)中为每个靶序列得到的 PCR 产物的部分或全部和随后对它们的 DNA 相互作用的同时分析。

适宜的是, 将两个或更多个扩增的序列在通过与阵列杂交进行汇集和分析之前进行不同标记。

适宜的是, 当两个或更多个扩增的序列位于不同染色体上时, 将所述序列进行相同的标记并通过与阵列杂交进行分析。

适宜的是, 当两个或更多个扩增的序列位于相同染色体上、其距离足够远以使得 DNA-DNA 相互作用信号间的重叠最小时, 将这些扩增的序列进行相同的标记。

适宜的是, 将高通量测序用于分析靶序列和捕获的目的序列间形成的连接接头。

适宜的是, 通过加入测序至扩增序列的末端所需的衔接序列将测序定向于靶序列和捕获的目的序列的连接接头。

适宜的是, 通过加入测序需要的完整的或部分的衔接序列作为用于扩增一个或多个目的核苷酸序列的寡核苷酸引物的 5'突出端, 将测序定向于靶序列与捕获的目的序列间形成的连接接头。

适宜的是, 通过将生物素物质或其它部分连接到用于扩增一个或多个目

的核苷酸序列的寡核苷酸引物,随后通过链霉抗生物素蛋白或者其它介导的对 PCR 扩增产物的纯化将测序定向于靶序列与捕获的目的序列间形成的连接接头。

适宜的是,通过设计用于扩增在距所分析的第一和/或第二限制酶识别位点 400、300、200、150、100、90、80 70、60、50、40、30、20、10、9、8、7、6、5、4、3、2、1 核苷酸范围内的一个或多个目的核苷酸序列的寡核苷酸引物将测序定向于靶序列与捕获的目的序列之间的连接接头。

适宜的是,通过设计用于扩增一个或多个目的核苷酸序列的寡核苷酸引物因而使它们部分地或完全地与所分析的第一和/或第二限制酶识别位点重叠,将测序定向于靶序列与捕获的目的序列间的连接接头。

适宜的是,序列读过连接接头因而当分析多重或汇集的 PCR 样品时,在连接接头的每一侧获得足够的序列信息(如 12 个核苷酸或更多)以明确地确定每个靶序列和每个捕获的目的序列。

适宜的是,靶核苷酸序列选自由基因组重排、启动子、增强子、沉默基因、隔离子、基质附着区、基因座控制区、转录单位、复制起点、重组热点、易位断点(translocation breakpoint)、着丝粒、端粒、基因密集区(gene-dense region)、基因稀少区(gene-poor region)、重复元件和(病毒)整合位点组成的组。

适宜的是,靶核苷酸序列是与疾病相关的或造成疾病的核苷酸序列,或在线性 DNA 模板上位于与疾病相关的或造成疾病的基因座相距多至或大于 15Mb 处。

适宜的是,靶核苷酸序列选自由 *AML1*, *MLL*, *MYC*, *BCL*, *BCR*, *ABL1*, *IGH*, *LYL1*, *TAL1*, *TAL2*, *LMO2*, *TCR α / δ* , *TCR β* 和 *HOX* 或其它与疾病相关的基因座组成的组,所述其它与疾病相关的基因座描述于“Catalogue of Unbalanced Chromosome Aberrations in Man”第 2 版. Albert Schinzel.柏林: Walter de Gruyter, 2001. ISBN 3-11-011607-3 中。

适宜的是,靶序列沿着线性基因组模板分布因而使相互作用的序列覆盖整个染色体或基因组。

适宜的是,第一限制酶是识别 6-8 bp 识别位点的限制酶。

适宜的是,第一限制酶选自由 *Bgl*II、*Hind*III、*Eco*RI、*Bam*HI、*Spe*I、*Pst*I 和 *Nde*I 组成的组。

适宜的是,第一限制酶基于其在重复序列中不存在或不具有代表性

(under representation)而被选择。

适宜的是，第二限制酶是识别 4 或 5 bp 核苷酸序列识别位点的限制酶。

适宜的是，第二限制酶识别位点位于与靶核苷酸序列中第一限制酶位点相距大于约 350bp 处。

适宜的是，相互作用频率从低至高的转变指示平衡的和/或不平衡的基因重排的位置。

适宜的是，受试者样品的 DNA-DNA 相互作用频率与对照的相互作用频率相比呈倒转模式指示平衡的和/或不平衡的倒位。

适宜的是，受试者样品的 DNA-DNA 相互作用频率与对照的相互作用频率相比的降低，与更远区域的 DNA-DNA 相互作用频率的增加的组合，指示平衡的和/或不平衡的缺失。

适宜的是，受试者样品的 DNA-DNA 相互作用频率与对照的相互作用频率相比的增加或减少指示平衡的和/或不平衡的重复或插入。

适宜的是，在进行所述方法前使用光谱核型分析和/或 FISH。

适宜的是，疾病是遗传疾病。

适宜的是，疾病是癌症。

适宜的是，对与两个或更多个靶序列相互作用的核苷酸序列进行扩增。

适宜的是，靶序列位于已知与疾病状态相关的基因组基因座或在其附近。

适宜的是，靶序列在对重排的位置缺乏现有知识的情况下选择并且隔开从而使相互作用的序列覆盖整个染色体或基因组，并且其中鉴定的相互作用序列允许重建线性染色体图和发生在染色体内部和染色体间的基因组重排。

适宜的是，扩增的序列是标记的。

适宜的是，扩增的序列根据其在基因组中的位置而不同地标记。

适宜的是，本方法是为了检测平衡的和/或不平衡的重排、易位、倒位、缺失、重复或插入。

适宜的是，阵列杂交步骤由测序步骤代替。

适宜的是，靶核苷酸序列和目的核苷酸序列都由测序来鉴定。

适宜的是，衔接序列连接到 PCR 产物。

适宜的是，与两个或更多个靶序列相互作用的序列每个都在分别的 PCR 反应中扩增。

适宜的是，与两个或更多个靶序列相互作用的序列每个都在分别的 PCR 反应中扩增并随后汇集进行同步分析。

适宜的是，与两个或更多个靶序列相互作用的序列由多重 PCR 扩增。

优点

本发明有许多优点。这些优点将在以下描述中明确。

举例来说，4C 技术可以是多重化的(multiplexed)，从而使与两个或更多个靶序列的相互作用可以在单个实验中分析，例如在单个阵列上。

进一步举例来说，多重化的 4C 技术可用以筛选整个基因组中未知位置上基因组 DNA 的重排。

进一步举例来说，高通量测序可以用于代替微阵列来分析捕获的 DNA 片段。多重化和测序的改善甚至可以组合。

进一步举例来说，代替多重化，由不同靶序列捕获的目的序列可以针对每个靶序列单独地扩增，然后汇集以在微阵列上进行同时分析。

进一步举例来说，代替多重化，由不同靶序列捕获的目的序列可以为每个靶序列单独地扩增，然后汇集以通过高通量测序进行同时分析。

举例来说，本发明是有益的，这是因为它提供了可商用的核苷酸序列、方法、探针和阵列等。

进一步举例来说，本发明是有益的，这是因为它提供了以高通量分析核空间中两种或更多种核苷酸序列的相互作用的频率的方法。

进一步举例来说，本发明有益的，这是因为利用常规的 3C 技术，每种单一的 DNA-DNA 相互作用必须通过包括独特引物对的独特 PCR 反应来分析。因此，只有将 PCR 自动化，高通量分析才有可能，但是如此多的引物的成本会很高。因此，用常规 3C 技术进行高通量(基因组广度的)DNA-DNA 相互作用分析是不可行的。相反，本发明现在能同时筛选上千个 DNA-DNA 相互作用。本发明所述的 DNA-DNA 相互作用的高通量分析将极大增加分析规模和解析度。

进一步举例来说，本发明是有益的，这是因为利用常规 3C 技术，筛选将偏向于那些其寡核苷酸引物经设计、排列和包括在分析中的 DNA 序列。选择这些寡核苷酸引物通常是基于对据信会与正被研究的核苷酸序列交联的诸如(远处的)增强子和/或其它调节元件/超敏感位点的位置的认识。因此，

常规 3C 偏向于包括在 PCR 扩增步骤中的 PCR 引物的设计, 而 4C 是无偏的而且能用于针对完整的基因组搜索相互作用的 DNA 元件。这是因为在 4C 中扩增交联的序列不是基于对与正被研究的核苷酸序列交联的序列的预先认识。更确切地, 在一个 4C 的实施方案中, 利用与该核苷酸序列杂交的 PCR 引物可扩增与第一(靶)核苷酸序列交联的序列。因此, 本发明能无偏地在基因组广度筛选 DNA-DNA 相互作用。

进一步举例来说, 本发明是有益的, 这是因为利用常规 3C 技术仅能选择性扩增单一 DNA-DNA 相互作用。这在与阵列杂交时是无法提供出信息的。该技术已被改善, 从而使所有与第一(靶)核苷酸序列相互作用的片段在现在被扩增出, 如选择性地扩增出。

进一步举例来说, 本发明是有益的, 这是因为 4C 技术能用于检测核酸(例如, 染色体)中平衡的或不平衡的遗传异常——如所有类型的易位、缺失、倒位、重复和其它基因组重排。4C 技术(其测量 DNA 片段的接近度)甚至能确定受试者获得某些易位、缺失、倒位、重复和其它基因组重排(如平衡的或不平衡的易位、缺失、倒位、重复和其它基因组重排)的倾向性。较当前策略的优势是它无需知道改变的确切位置, 因为 4C 技术的解析度使它即使在‘4C-诱饵’(如由被分析的第一和第二限制酶识别位点所限定的)远离变化(如多至一百万个碱基或甚至更多)时也可用于检测重排。另一个相对于目前策略的优势是其允许同步的、无偏的对平衡的和不平衡的基因组重排进行基因组广度的搜索。另一个优势是 4C 技术能对变化进行准确作图, 这是因为它能用于限定改变发生处之间的两个(第一)限制酶位点。另一个优势是细胞不需要在固定前进行培养。因此, 也可分析诸如实体瘤的基因组重排。

进一步举例来说, 本发明是有益的, 这是因为 4C 技术也能在恶变前状态中(即在所有细胞包含这些变化前)检测改变(如重排)。因此, 该技术不仅可用于诊断疾病, 而且用于疾病的预测。

进一步举例来说, 本发明所述的阵列设计比现有的基因组嵌合阵列(genomic tiling arrays)(如 Nimblegen 基因组嵌合阵列)尤其有优势, 这是因为该设计能在每个单一阵列中代表大得多的基因组部分。举例来说, 对于识别六个核苷酸的序列的限制酶, 例如, 3 个阵列(每个带有约 385,000 个探针)将足以覆盖完整的人或小鼠基因组。对于识别超过 6bp 的限制酶, 例如, 约 385,000 个探针的单一阵列可用于覆盖例如完整的人或小鼠基因组。阵列设

计的优势有：(1)每个探针能提供很多信息，这是因为每个探针分析独立的连接事件，极大方便了结果的解释；和(2)基因组的大代表物可在单一阵列上被发现，其具有成本效率。

4C技术可有益地用于对最初由细胞遗传学方法(光学显微术、FISH、SKY等)检测到而没有很好表征的重排进行细致作图。

4C技术可有益地用于在单一阵列上同时筛选发生在多个基因座附近的重排的组合。

附图简述

图 1

3C 技术的原理

图 2

(a)4C 技术的一个实施方案的原理。如常用诸如 *HindIII*(H)作为限制酶，进行 3C 分析。解除交联后，DNA 混合物将包含第一(靶)核苷酸序列，其连于许多不同的片段。用扩增方法(如反向 PCR)利用第一(靶)核苷酸序列特异性引物在诸如 *DpnII* 环上扩增并标记这些片段。标记的扩增产物可与本文所述的阵列杂交。以 *HindIII* 和 *DpnII* 做例子，但也可使用其它限制酶组合(如 6 或 8-和 4 或 5-切割酶)。(b)来自两个独立的胎儿肝(L1, L2)和脑(B1, B2)样品的通过凝胶电泳分离的 PCR 结果。(c)示意性表示微阵列探针位置。探针被设计在 *HindIII* 位点的 100 bp 内。因此，每个探针分析一个可能的连接配体(partner)。

图 3

4C 技术检测 Rad23A 的基因组环境(染色体 8)。显示了未处理过的比率(针对 Rad23A 的 4C 信号除以对照样品所得的信号)，其针对位于小鼠染色体 10、11、12、14、15、7 和 8(从上至下；显示的区域处于距离每个相应的着丝粒的相同距离处)上~15 Mb 或更远的基因组区的探针。注意到染色体 8(第 7 行)上(Rad23A)诱饵周围有大簇强信号，其证实了 4C 技术检测到线性染色体模板附近的基因组片段(符合以下事实，即相互作用频率与基因组位点分隔成反比)。注意到显示高信号强度的诱饵周围顺式相连的区域是大的(>5Mb)，这暗示诸如易位甚至能用距离断点超过 1MB 的诱饵来检测。

图 4

转录组(胎儿肝)和非转录组(胎儿脑)中染色体 7(~135Mb)上 β -球蛋白的 4C 相互作用(由连续平均(running mean)法分析而得)。注意到与 β -球蛋白的长程相互作用在组间有差异(可能依赖于基因转录状态)。强 4C 信号在诱饵周围划分了大区域(>5 Mb), 而与组织无关。

图 5

在胎儿肝细胞中, *Uros* 和 *Eraf* 与 β -球蛋白相互作用。4C 法揭示两种基因(*Eraf* 和 *Uros*), 与位于~30 Mb 以外的 β -球蛋白基因座相互作用超过>30 Mb。这两种相互作用以前被不同的技术(荧光原位杂交)所发现, 其描述于 Osborne 等., *Nature Genetics* 36, 1065(2004)中。该例子显示 4C 技术检测到的长程相互作用可通过 FISH 验证并真实地反映了核接近情况。

图 6

4C 技术准确地鉴定出顺式相连的不相关基因组区域间的转换。对于这些实验, 使用转基因小鼠, 其包含人 β -球蛋白基因座控制区(LCR)表达盒(~20 kb), 该表达盒(通过同源重组)插入到小鼠染色体 8 上的 *Rad23A* 基因座中。4C 技术在转基因小鼠 E14.5 胚胎肝上进行, 该转基因小鼠中就这种插入物而言是纯合的。整合表达盒(HS2)内的 *HindIII* 片段被用作‘4C 诱饵’。数据显示, 4C 技术准确地确定了转基因表达盒的两个末端(底下一行: 仅人 LCR(~20kb)中的探针得到 4C 信号, 而余下~380 kb 人 β -球蛋白序列中的探针不行), 清楚地揭示了小鼠染色体 8 上的整合位置(上排: 将染色体 8 上的信号(对于整合位置, 参见箭头)与 6 个其它小鼠染色体上的信号作比较)(描述了完整染色体)。该例子显示, 4C 技术可用于检测异位整合的 DNA 片段(病毒、转基因等)的基因组位置。它显示, 顺式相连的不相关基因组区域间的转换能被准确鉴定出, 其可用于鉴定基因组断点和易位配偶体。

图 7

4C 技术产生了可重复的数据, 这是因为针对 HS2 和 β -球蛋白的图谱(profile)非常相似。4 个生物学独立的 4C 实验在 E14.5 胎儿肝上进行, 其使用了 β -球蛋白基因 β -major(上 2 行)或 β -球蛋白 HS2(底下 2 行)作为诱饵。这些诱饵在线性染色体模板上相距~40 kb, 但以前显示在核空间中接近(Tolhuis 等, *Molecular Cell* 10, 1453(2002))。所描述的是小鼠染色体 7 上的~5 Mb 区域, 其与 β -球蛋白基因座相距 20-20Mb。数据在独立实验间显示出高度可重复性, 证实了在核空间中接近的 2 个片段共有位于基因组其它地方的相互

作用配体。

图 8

将 4C 技术用于测量在来自健康人(顶部)和带有易位的患者(A; B)(底部)的细胞中(染色体 A 上的)序列 X 的 DNA-DNA 相互作用频率。代表 DNA-DNA 相互作用频率的信号强度(Y 轴)相对于在线性染色体模板上排列的探针(X 轴)进行作图。在正常细胞中,在序列 X 周围的染色体 A 上检测到了频繁的 DNA-DNA 相互作用。在患者细胞中,对于位于断点(BP)另一边的染色体 A 上的探针,观察到相互作用频率降低 50%(将灰色曲线(患者)与黑线(健康人)作比较)。而且,易位使部分染色体 B 在物理上接近于序列 X,而且对于染色体 B 上的该区域,现在观察到了频繁的 DNA-DNA 相互作用。该染色体上的相互作用频率从低至高的突然转变标志着其断点的位置。

图 9

4C 技术可检测一个或多个(平衡的)倒位。相较于非患病的(点曲线)受试者,在患病的(实曲线)受试者中观察到了 DNA-DNA 相互作用频率的倒转模式(由 4C 技术以杂交信号强度测量),这揭示了倒位的存在和大小。

图 10

由 4C 技术进行的杂合缺失检测。相比于非患病的(黑曲线)受试者,在患病的(灰曲线)中 DNA-DNA 相互作用频率降低(由 4C 技术以杂交信号强度测量)的探针揭示了缺失区域的位置和大小。患病的受试者的缺失区域中的剩余杂交信号来自完整的等位基因(杂合缺失)。缺失通常伴随着直接位于缺失区域外的探针的信号强度的增加(注意到灰曲线在缺失的右边位于黑曲线之上),这是因为这些区域在物理上更接近于 4C 序列(诱饵)。

图 11

由 4C 技术检测到的重复。与正常(黑曲线)受试者相比,在患者(灰曲线)中杂交信号增加的探针指明了重复的位置和大小。由 4C 技术检测到的重复通常伴随着与非患病的受试者相比在患病的受试者中在重复区域之外的探针的杂交信号的降低(重复增加了它们与 4C 序列的基因组位点分离)。

图 12

4C 技术揭示的与 β -球蛋白的长程相互作用。**a**, 4C 相对于对照杂交信号的未处理的比率揭示了染色体 7 内的 β -球蛋白 HS2 与 2 种无关的染色体(8 和 14)的相互作用。**b-c**, 针对 2 种独立的胎儿肝(顶部, 红色)和胎儿脑样品(底

部,蓝色)的未处理的数据沿着染色体 7 上 2 个不同的 1-2 Mb 区域进行绘图。在 2 个胎儿肝样品(b)或 2 个脑样品(c)中观察到了高度可重复的相互作用簇。**d-e**, 针对相同区域的连续平均数据。错误发现率被设置在 5%(点线)。**f**, 与染色体 7 上活性的(胎儿肝, 顶部)和失活的(胎儿脑, 底部) β -球蛋白相互作用的区域的示意图。

图 13

活性的和失活的 β -球蛋白分别与活性的和失活的染色体区域相互作用。**a**, 对胎儿肝中 β -球蛋白长程相互作用之间作比较(4C 连续平均, 顶部), 在胎儿肝中进行微阵列表达分析(对数坐标, 中间)和沿着含基因 Uros(距离 β -球蛋白~30 Mb)的 4 Mb 区域作图的基因位置(底部), 表明活性 β -球蛋白优先与其它活性转录基因相互作用。**b**, 同样比较胎儿脑中距离球蛋白~38 Mb 处的 OR 基因簇的周围, 显示失活的 β -球蛋白优先与失活的区域相互作用。**c**, 根据基因成分和活性来表征与胎儿肝(左)和脑(右)中 β -球蛋白相互作用的区域。

图 14

普遍表达的 Rad23A 与胎儿肝和脑中非常相似的活性区域相互作用。**a**, 与胎儿肝(顶部, 红色)和脑(底部, 蓝色)中活性 Rad23A 相互作用的染色体 8 上区域的示意图。**b**, 比较 Rad23A 长程相互作用(4C 连续平均)和胎儿肝中微阵列表达分析(对数坐标)作比较(顶部两排), Rad23A 长程相互作用(4C 连续平均)和胎儿脑中微阵列表达分析(对数坐标)(第 3 和 4 排), 和沿着染色体 8 的 3 Mb 区域作图的基因位置(底部一排)。**c**, 根据基因成分和活性来表征与胎儿肝(左)和脑(右)中 Rad23A 相互作用的区域。

图 15

冷冻 FISH (cryo-FISH)确证了 4C 技术真正能鉴定出相互作用的区域。**a**, 部分的冷冻切片(200 nm)的例子显示出超过 10 个核, 其中一些含 β -球蛋白基因座(绿色)和/或 Uros(红色)。由于切片的缘故, 许多核不含针对这两个基因座的信号。**b-d**, 完全(b)和部分(c)重叠信号和接触信号(d)的例子, 这些都被评为相互作用阳性。**e-g**, 含非接触性等位基因的核(e-f)和仅含 β -球蛋白的核(g)的例子, 将其全部评分为相互作用阴性。**h-i**, 冷冻 FISH 结果的示意图。与 β -球蛋白(h)和 Rad23A(i)相互作用的百分比在染色体上方标明, 用于指示通过 4C 技术鉴定为阳性(红色箭头)和鉴定为阴性(蓝色箭头)的区域。同一

BAC 用于两种组构。通过冷冻 FISH 测量的胎儿肝和脑中两个远离的 OR 基因簇之间的相互作用频率写在染色体下方。

图 16

4C 分析 HS2 和 β -major, 得出高度相似的结果。(a)4 个独立 E14.5 肝样品的未处理的 4C 数据显示出与 HS2 的相互作用(顶部)和与 β -major 的相互作用(底部)之间非常相似的模式。(b)大量重叠存在于在 HS-2 实验中被评为相互作用阳性的探针和 β -major 实验中被评为相互作用阳性的探针之间。

图 17

与 β -球蛋白相互作用的区域也频繁互相接触。包含活跃转录的基因、并由 4C 技术鉴定为与胎儿肝中 β -球蛋白相互作用的 2 个区域(几乎相距 60 Mb), 通过冷冻 FISH 显示共定位频率为 5.5%, 其显著大于背景共定位频率。

图 18

通过使用单个染料标记与多个靶序列相互作用的 DNA 片段而进行的多重 4C 来展示杂合缺失的实例。观察到的患者(样品)对健康人(对照)的相互作用频率的比率在右侧描述。

图 19

用第一个断点上游(‘左侧’)2Mb(A)或 1.3Mb(B)处的靶核苷酸序列, 通过 4C 揭示出现在白血病患者中的缺失的存在。注意, 缺失导致缺失区 DNA 相互作用信号的减弱, 但是对最后断点直接下游(‘右侧’)的序列也导致 DNA:DNA 相互作用频率的增加。当仔细研究与靶核苷酸序列 B 的相互作用时(参见底部两张图), 这一现象特别明显。在缺失区的每一侧根据 4C 数据设计引物, 并通过测序确定断点: 常规字体是缺失上游的序列, 粗体表示插入的核苷酸, 下划线是缺失下游的序列。

图 20

通过使用单个染料以标记与多个靶序列相互作用的 DNA 片段的多重 4C 展示的杂合倒位。观察到的患者(样品)对健康人(对照)的相互作用频率的比率在右侧描述。注意, 当相对于靶核苷酸序列的断点的位置不同时断点附近的比率可能不同。

图 21

颜色在相邻靶核苷酸序列之间交替, 其允许检测靶核苷酸序列(蓝色)附近的因杂交信号的饱和而不能检测的缺失(红色)。在阵列上的探针量没有饱

和的情况下,蓝色信号也会在缺失处减弱。

图 22

颜色在相邻靶核苷酸序列之间交替,其允许倒位的检测(红色)。观察到的患者(样品)对健康人(对照)的相互作用频率的比率在右侧描述。注意,与单个染料实验相比(参见图 2)交替染料的使用有助于重排的检测,例如倒位。另外注意,当相对于靶核苷酸序列的断点的位置不同时断点附近的比率可能不同。

图 23

通过使用不同染料以标记与不同靶序列相互作用的 DNA 片段来进行的多重 4C 所揭示的杂合倒位的实例。倒位的断点由患者的红色和绿色信号(对照样品中没有)的位置显示。注意,多个颜色的引入有助于重排的检测(比较例如图 4 和 5)。

图 24

平衡易位的检测。每个染色体用两个独特的染料标记从而使线性染色体模板上相邻的靶序列之间染料不同。如果易位是平衡的,两种染色体特异的染料每个都应该在不相关染色体的线性模板上直接彼此相邻的互斥的一组探针上给出强烈的杂交信号。这种不相关的染色体上的断点位于显示正杂交信号的两组探针之间。在断点上和越过断点时亲本染色体的信号将是对照的一半强度(图中未显示)。

图 25

检测平衡易位的原理论证。(R. Burnett 等, Blood, Vol 84, No 4(August 15), 1994:第 1232-1236 页)中描述的 t(1;7)易位的检测。染色体 7 上的 TCR β 基因座侧翼的靶核苷酸序列,其中红色信号代表与位于 TCR β 基因座上游靶序列的 DNA:DNA 相互作用,而蓝色信号代表与位于 TCR β 基因座下游靶序列的 DNA:DNA 相互作用。描述的是在染色体 1 上找到的相互作用的 DNA 信号。顶部一排显示理论信号分布。中下排显示实际信号分布。底部的一排显示染色体模板上并列的单个探针解析度的信号。注意,在平衡易位的情况下,断点侧翼的靶核苷酸序列将显示互斥的一组染色体间 DNA 相互作用信号,其在易位的配体染色体上的线性染色体模板上直接彼此相邻。测序的断点的位置(在 Burnett 等, 1994 中描述)在底部的一排用箭头表示。

图 26

检测不平衡易位的理论性实例。每个染色体用两个独特的染料标记从而使线性染色体模板上相邻的靶序列之间染料不同。如果易位发生时在断点丢失 DNA (即不平衡的易位), 两种染色体特异的染料每个都将在不相关的染色体上相互排斥的一组探针上给出强烈的杂交信号, 其在不相关的染色体的线性模板上彼此不直接相邻。指示了缺失区。

图 27

不平衡易位的检测。如(RJ Galjaard 等, Am J Med Genet A. 2003 Aug 30;121(2):168-73)中的描述的 t(4;7)易位的检测。靶核苷酸序列位于染色体 7; 描述的相互作用的 DNA 信号位于染色体 4。使用了位于染色体 7 上断点上游(5')和下游(3')的两个靶序列。指示了位于染色体 4 上的相互作用的 DNA 信号(对两个靶序列均为蓝色)。在这个患者中染色体 4 上相互作用的 DNA 片段簇之间的区域已经缺失。顶部: 完整的染色体 4 的信号。底部行 4C 数据: 染色体 4 上断点周围 11.5MB 区域处的信号。根据这些 4C 数据, 包括易位断点在内的染色体 4 上的 HindIII 限制性片段被鉴定并通过测序用其对断点作图。在图的底部提供了序列, 其中下划线序列来自染色体 4, 粗体的在 7 和 4 上都存在, 而常规字体的序列来自染色体 7。

图 28

染色体特异的 DNA 相互作用的标记。染色体 3 上出现的蓝色信号和染色体 1 出现的桔色信号揭示易位配体染色体和断点的大体位置。

图 29

对连接到靶核苷酸序列(红色: 家庭地址(home address))的 PCR 扩增的目的核苷酸序列(蓝色: 旅行地址(travel address))的测序。扩增用引物(红色)进行, 其中至少一个与靶核苷酸序列互补。任选地, 可用不同方法将衔接子(绿色)加至 PCR 产物的末端, 如图所示。

图 30

4C精确地检测平衡的易位和倒位(A-B)。4C技术检测平衡的易位t(1;7)。(A)在健康的对照样品中, 位于染色体7 *TCRB*基因座相对两侧的靶片段a (红色)和b (蓝色)不捕获染色体1上的区域。(B)在包含平衡易位t(1;7)(p35;q35)的HSB-2细胞系中, 每个*TCRB*靶片段都捕获染色体1的区域。所述捕获的区域为几个兆碱基大小(缩放1), 彼此直接相邻(缩放2)并处在以前克隆的断点侧翼(箭头)。其它染色体的结果参见图S1。(C-D)4C检测平衡倒位。(C)在健康

的对照样品中,位于染色体7 *TCRB*基因座相对两侧的靶片段a(红色)和b(蓝色)不捕获染色体7上它处的大区域。(D)在T-ALL患者样品中,每个*TCRB*靶片段捕获染色体7另一端的额外的区域。最5'端的靶片段(a; 红色)捕获*HOXA*基因簇5'方向的区域,最3'端的靶片段(b; 蓝色)捕获*HOXA*基因簇3'方向的区域,证明倒位。捕获的区域是几个兆碱基大小(缩放1)和彼此直接相邻(缩放2),显示倒位是平衡的。靶片段都鉴定*HOXA9*附近6kb区域内的断点(箭头)。用~60kb的窗口大小将连续平均数据作图。放大(zoom)显示未处理的信号强度。

图31

4C准确地检测未平衡的重排。(A)从具有先天畸形的未出生儿童的细胞系中4C精确地检测易位t(4;7)(p15.2;q35)与相伴的小缺失(即未平衡的重排)。位于染色体7断点相对两侧的靶片段a(红色)和b(蓝色)均捕获染色体4上延续(spanning)几个兆碱基的片段(其它染色体上的信号,参见图35)。具有高信号的两个捕获的区域不直接彼此相邻,显示易位伴有染色体4上的缺失。(B)一个断点的序列(箭头),染色体7和4的序列分别是小写和大写字母。(C)4C精确地鉴定T-ALL患者样品中的纯合缺失。与对照样品(顶部)相比,位于染色体9上19.3Mb处的靶片段鉴定患者样品(底部)中缺少高信号的区域(箭头之间),显示9p21上的~2Mb的缺失。缺失的3'方向上的信号在患者中比对照高,因为由于缺失这一区域更加接近靶片段。(D)由(C)中的箭头指示的跨越断点的序列,确证了缺失。绘在此图中的是未处理的信号强度。缺失区中罕见的高信号指示这些探针显示非特异性杂交。

图32

4C筛选确定*LMO3*是*TCRB*的新的易位配体。用染色体7上*TCRB*基因座附近的靶片段通过4C筛选了五个未表征的T-ALL患者样品。(A)在一个患者样品中,高信号特异性地出现在染色体12上,揭示了易位t(7;12)(q35;p12.3)。关于所有其它染色体上的信号,参见图S4。缺失出现在染色体12上的易位位点(箭头)的几个兆碱基处(缩放1)。易位位点出现在*LMO3*基因附近的6kb区域(缩放2)。(B)断点t(7;12)(q35;p12.3)的序列;大写的核苷酸来自12;小写的来自7而斜体的来自未知区域。(C)易位位点t(7;12)(q35;p12.3)的图示。*TCRB*的增强子位于*LMO3*基因的下游70kb。用~60kb的窗口大小将连续平均数据作图。放大显示未处理的信号强度。

图33

健康对照和携带t(1;7)(p35;q35)的样品中跨越所有染色体的4C信号。黑色箭头显示靶序列的位置。红色箭头显示易位位点的位置。用~60kb的窗口大小将连续平均数据作图。Y轴的刻度(绝对单位)对所有染色体都相同。

图34

限制性-片段-配对-末端-测序。(A)限制性-片段-配对-末端-测序的图示。(B)染色体7上TCRB和HOXA之间倒位的断点序列(参见图30)。黑色的序列是用4C技术找到的,其位于HOXA中标记从捕获的向非捕获的片段转化的探针之间的位置。红色的序列是用限制性-片段-配对-末端-测序从黑色的序列中找到的,其位于TCRB-基因座中。

图35

携带t(4;7)(p15.2;q35)的样品中用两个不同的染色体7靶序列获得的跨越所有染色体的4C信号。黑色的箭头指示靶序列的位置。红色的箭头指示易位位点的位置。用~60kb的窗口大小将连续平均数据作图。Y轴的刻度(绝对单位)对所有染色体都相同。

图36

两个T-ALL患者样品中染色体7上TCRB基因座附近的靶序列获得的跨越所有染色体的4C信号,其中一个携带t(7;12)易位。黑色箭头指示靶序列的位置。红色箭头指示易位位点的位置。用~60kb的窗口大小将连续平均数据作图。Y轴的刻度(绝对单位)对所有染色体都相同。

图37

T-ALL患者样品中LMO3的表达。基因表达用affymetrix基因表达阵列测定。LMO3在携带t(7;12)(q35;p12.3)的患者中表达,而不在其它患者中表达。

图38

要用Solexa测序分析的4C PCR引物和产物。Solexa产生的序列(箭头)首先阅读“DpnII引物”(18个核苷酸,包括GATC(即DpnII识别位点)),随后是捕获的序列。

图39

使用具有Solexa接头序列的5'突出端的引物的PCR结果。作为对比,也显示了用标准引物(道1、5、9从左到右;无突出端)获得的结果。

图40

4C-测序的结果。

图41

越过HSB-2 T-ALL细胞系中的断点(在染色体7的~142Mb处, 即距诱饵3Mb)引物组3(139Mb)捕获染色体1序列。作为对比, 显示了微阵列结果。注意, 用于微阵列实验的靶序列(诱饵)距断点更近(<1Mb), 解释了为何它在染色体1上更好地在图上定位了(map)断点。

发明详述

3C 技术

3C 方法已经详细描述于 Dekker 等.(2002), Tolhuis 等.(2002), Palstra 等.(2003), Splinter 等.(2004)和 Drissen 等.(2004)中。简而言之, 3C 的进行是通过: 用第一限制酶消化交联的 DNA, 然后以非常低的 DNA 浓度进行连接。在这些条件下, 交联片段的分子内连接大大优于随机片段的分子间连接。然后, 解除交联并通过聚合酶链式反应(PCR)利用基因座特异性引物检测和定量各个连接产物。两个特异性基因座的交联频率(X)由定量 PCR 反应利用对照和交联模板来确定, 而且 X 被表示为用交联模板和用对照模板得到的产物量的比率。

根据本发明, 利用 Splinter 等, (2004)*Methods Enzymol.* 375, 493-507 所述的方法制备 3C 模板。(即甲醛固定、(第一)限制酶消化、重连接交联的 DNA 片段并纯化 DNA)。简而言之, 样品(如细胞、组构或核)利用交联剂(如甲醛)固定。然后, 进行第一限制酶消化, 从而在交联的核的范围内消化 DNA。然后, 以低 DNA 浓度(例如, 约 3.7ng/ μ l)进行分子内连接, 其偏好交联的 DNA 片段间的连接(即分子内连接)胜于非交联的 DNA 片段间的连接(即分子间或随机连接)。接下来, 解除交联并纯化 DNA。产生的 3C 模板包含被连接的限制性片段, 这是因为它们原来在核空间中是接近的。

由于在分子内连接步骤前将第一限制酶用于消化 DNA, 第一限制酶的酶识别位点将分隔第一(靶)核苷酸序列和已经连接的核苷酸序列。因此, 第一识别位点位于第一(靶)核苷酸序列和连接的核苷酸序列(即连接的第二序列)之间。

核苷酸序列

本发明涉及核苷酸序列(如 3C 模板、4C 模板、DNA 模板、扩增模板、

DNA 片段和基因组 DNA)的用途, 其可从数据库中获得。

核苷酸序列可以是基因组的、合成或重组来源的 DNA 或 RNA, 如 cDNA。例如, 重组核苷酸序列可用 PCR 克隆技术来制备。这将包括制备在需要克隆的序列区域侧翼的引物对, 将引物与得自诸如哺乳动物(如动物或人细胞)或非哺乳动物细胞的 mRNA 或 cDNA 接触, 在能扩增所需区域的条件下进行聚合酶链式反应(PCR), 分离扩增的片段(如通过在琼脂糖凝胶上纯化反应混合物)并回收扩增的 DNA。可将引物设计成包含合适的限制酶识别位点, 从而使扩增的 DNA 能被克隆进合适的克隆载体中。

核苷酸序列可以是双链的或单链的, 无论其代表有义或反义链或其组合。

对于一些方面, 优选核苷酸序列是单链 DNA——如单链引物和探针。

对于一些方面, 优选核苷酸序列是双链 DNA——如双链 3C 和 4C 模板。

对于一些方面, 优选核苷酸序列是基因组 DNA ——如一个或多个基因组基因座。

对于一些方面, 优选核苷酸序列是染色体 DNA。

核苷酸序列可包含第一(靶)核苷酸序列和/或第二核苷酸序列。

第一和第二限制酶识别位点将会互不相同, 并通常将在核苷酸序列中只出现一次。

在一个方面, 提供了环化的核苷酸序列, 其包含第一核苷酸序列和(如连接于)第二核苷酸序列, 所述第一和第二核苷酸序列由第一和第二限制酶识别位点分隔(如分开或隔开), 其中所述第一核苷酸序列是靶核苷酸序列而且所述第二核苷酸序列可由交联基因组 DNA(如在体内或体外)而得到。第一和第二限制酶识别位点将会互不相同并通常将在核苷酸序列中仅出现一次。

在另一方面, 提供了环化的核苷酸序列, 其包含第一核苷酸序列和(如连接于)第二核苷酸序列, 所述第一和第二核苷酸序列由第一和第二限制酶识别位点分隔(如分开或隔开), 其中所述第一核苷酸序列是靶核苷酸序列, 而且其中所述第一和第二核苷酸序列可由以下过程而得, 其包括以下步骤: (a)交联基因组 DNA(如在体内或体外); (b)用第一限制酶消化交联的 DNA; (c)连接交联的核苷酸序列; (d)解除交联; 和(e)用第二限制酶消化核苷酸序列来环化核苷酸序列。

优选, 第二核苷酸序列切分(如对切)第一(靶)核苷酸序列。因此, 核苷

酸序列包含第二核苷酸序列，其将第一(靶)核苷酸序列分隔成两个部分或片段——如2个大小大致相等的部分或片段。通常该部分或片段的长度将至少为约16个核苷酸。

在另一方面，提供6-50个碱基对的序列的数据库，所述序列直接从侧翼包围并任选包括含有的每个靶序列的第一限制酶识别位点并且能够用于本文描述的方法以鉴定每个靶序列。

在另一方面，提供12-50个碱基对的序列的数据库，所述序列直接从侧翼包围基因组中所有第一限制酶识别位点并且可用于本文描述的方法以确定鉴定的每个捕获序列的基因组位置。

在另一方面，提供6-50个碱基对的序列的数据库，所述序列直接从侧翼包围并任选包括含有的每个靶序列的相关第二限制酶识别位点并且能够用于本文描述的方法以鉴定每个靶序列的位点。

在另一方面，提供12-50个碱基对的序列的数据库，所述序列直接从侧翼包围基因组中所有相关第二限制酶识别位点并且可用于本文描述的方法以确定鉴定的每个捕获序列的基因组位置。

第一核苷酸序列

第一核苷酸序列是靶核苷酸序列。

本文所用的术语“靶核苷酸序列”指用作诱饵序列的序列，以此鉴定与它交联的一个或多个序列(如一个或多个目的核苷酸序列或核苷酸序列组成未知的一个或多个序列)。

靶核苷酸序列的序列已知。

交联表明，靶核苷酸序列和与其交联的序列最初在核空间中接近。通过确定序列相互接近的频率，例如，有可能去理解核空间关系(如在体内或体外)中染色体和染色体区域的构象。而且，例如当增强子或其它转录调节元件与顺式或甚至反式定位的远距离启动子互相联络时，有可能理解基因组中复杂的结构组构。此外，甚至有可能去理解给定的基因组区相对于出现在相同染色体上(顺式)的核苷酸序列以及在其它染色体上(反式)的核苷酸序列的定位。因此，有可能对频繁共有核空间中位点的不同染色体上的核苷酸序列作图。此外，甚至有可能检测平衡的和/或不平衡的遗传异常——如平衡的和/或不平衡的易位、缺失、倒位、重复和其它基因组重排(如一个或多个染色

体中的缺失或易位)。在这方面,遗传异常会导致 DNA-DNA 相互作用在变化发生的位置上发生改变,这是可以被检测的。

根据本发明所述的第一(靶)核苷酸序列可以是任何希望确定其与一个或多个其它序列在核空间中相互作用的频率的序列。

在一个实施方案中,第一(靶)核苷酸序列的长度将大于约 350 bp,这是因为所选择的第二限制酶在距离第一限制性位点约 350 bp 或更远处切割第一(靶)核苷酸序列。这可使因拓扑学约束而对环形成带来的偏倚最小化(Rippe 等.(2001)*Trends in Biochem. Sciences* 26, 733-40)。

适宜的是,扩增后的第一(靶)核苷酸序列包含至少约 32 bp,这是因为用于扩增第二核苷酸序列的至少 2 个扩增引物的最小长度分别约为 16 个碱基。

在优选的实施方案中,第一(靶)核苷酸序列可包含完整或部分(如其片段)的如下序列或与其接近(如临近): 启动子、增强子、沉默基因、隔离子、基质附着区、基因座控制区、转录单位、复制起点、重组热点、易位断点、着丝粒、端粒、基因密集区、基因稀少区、重复元件、(病毒)整合位点、其缺失和/或突变与某种效应(如疾病、生理学上的、功能性的或结构性的效应——如 SNP(单核苷酸多态性))相关的核苷酸序列、或含这样的缺失和/或突变的核苷酸序列、或任何其中需要确定核空间中与其它序列相互作用的频率的序列。

如上所述,第一(靶)核苷酸序列可包含完整或部分(如片段)的其中遗传异常(如缺失和/或突变)与某种效应(如疾病)相关的核苷酸序列,或与这样的核苷酸序列接近(如临近)。因此根据本发明的这个实施方案,第一(靶核苷酸序列)可以是这样的核苷酸序列(如基因或基因座),该序列与其中的变化与疾病(如遗传或先天疾病)相关或有关的基因组区(在实体的 DNA 模板上)邻近或处在这样的基因组区中。换句话说,第一(靶)核苷酸序列可以是临床表型或者以其与临床表型的关联性为基础来选择。在优选的实施方案中,变化是在一个或多个染色体中的变化,而且疾病可以是诸如其中一个或多个缺失、一个或多个易位、一个或多个重复、和/或一个或多个倒位等的结果。这些基因/基因座的非限制性例子是 *AML1*、*MLL*、*MYC*、*BCL*、*BCR*、*ABL1*、免疫球蛋白基因座、*LYL1*、*TAL1*、*TAL2*、*LMO2*、*TCR α/δ* 、*TCR β* 、*HOX* 和其它在各种成淋巴细胞白血病中的基因座。

其它例子在电子数据库中有描述,如:

<http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=cancerchromosomes>

<http://cgap.nci.nih.gov/Chromosomes/Mitelman>

<http://www.progenetix.net/progenetix/P14603437/ideogram.html>

<http://www.changbioscience.com/cytogenetics/cyto1.pl?query=47,xy>

<http://www.possum.net.au/>

<http://www.lmdatabases.com/>

<http://www.wiley.com/legacy/products/subject/life/borgaonkar/index.html>

<http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=OMIM>

<http://www.sanger.ac.uk/PostGenomics/decipher/>

<http://agserver01.azn.nl:8080/ecaruca/ecaruca.jsp>

其它例子描述于“Catalogue of Unbalanced Chromosome Aberrations in Man”第2版. Albert Schinzel. 柏林: Walter de Gruyter, 2001. ISBN 3-11-011607-3 中。

在一个实施方案中, 术语“邻近”指“直接相邻”, 从而使在2个相邻序列间没有插入的核苷酸。

在另一个实施方案中, 在核酸序列和第一限制酶识别位点的上下文中的术语“邻近”指“直接相邻”, 从而使在核酸序列和第一限制酶识别位点间没有插入的核苷酸。

第二核苷酸序列

第二核苷酸序列是通过交联基因组DNA(如体内或体外)而可以得到的、得到的、鉴定到的、可以鉴定到的。

第二核苷酸序列(如目的核苷酸序列)在用交联剂处理样品并消化/连接交联的DNA片段后变成与第一(靶)核苷酸序列连接。该序列与第一(靶)核苷酸序列交联, 这是因为它们最初在核空间中接近并连接于第一(靶)核苷酸序列上, 因为连接条件相对于随机连接事件而言更倾向于交联DNA片段间的(分子内)连接。

基于例如易位、缺失、倒位、重复和其它基因组重排的改变的疾病一般由异常的DNA-DNA相互作用造成。4C技术测量出DNA-DNA相互作用的频率, 其主要是基因组位点分离的函数, 即DNA-DNA相互作用频率与出现在相同实体DNA模板上的2个DNA基因座间的线性距离(以千碱基计)成反

比(Dekker 等., 2002)。因此,产生新的和/或物理上不同的 DNA 模板的改变伴随着 DNA-DNA 相互作用的改变,而且这可通过 4C 技术来测量。

适宜的是,第二核苷酸序列至少有 40 个碱基对。

交联剂(如甲醛)可用于将蛋白质与其它邻近蛋白质和核酸交联。因此,两种或更多种核苷酸序列可以仅通过结合于这些核苷酸序列(之一)的蛋白质来交联。除了甲醛之外的交联剂也可根据本发明所述而使用,包括那些直接交联核苷酸序列的交联剂。交联 DNA 的试剂的例子包括但不限于紫外光、丝裂霉素 C、氮芥、苯丙氨酸氮芥、1, 3-丁二烯二环氧化物、顺式二氨基二氯铂(*cis* diaminedichloroplatinum) (II)和环磷酰胺。

适宜的是,交联剂将形成桥接相对短的距离(如约 2 Å)的交联,由此选出可逆转的密切相互作用。

交联的进行可以是诸如,于室温在 2%甲醛中孵育细胞——如通过在 10 ml 补充了 2%甲醛的 DMEM-10% FCS 中于室温孵育 1×10^7 个细胞 10 分钟。

第一限制酶

本文所用的术语“第一限制酶”指用于消化交联的 DNA 的第一限制酶。

第一限制酶的选择取决于要分析的靶序列(如基因座)的类型。希望进行预实验以优化消化条件。

第一限制酶可选自识别至少 8 bp 序列或更多碱基的 DNA 的限制酶。第一限制酶可选自识别至少 7 bp 序列或更多碱基的 DNA 的限制酶。第一限制酶可选自识别至少 6 bp 序列或更多碱基的 DNA 的限制酶。对于一些实施方案,第一限制酶可选自识别 4 bp 和/或 5 bp 序列的 DNA 的限制酶。

使用不常见的切割酶将增加每个诱饵捕获的(覆盖的)基因组距离。

识别 6 个 bp 序列 DNA 的限制酶包括但不限于 AclI、HindIII、SspI、BspLU11I、AgeI、MluI、SpeI、BglII、Eco47III、StuI、ScaI、ClaI、AvaIII、VspI、MfeI、PmaCI、PvuII、NdeI、NcoI、SmaI、SacII、AvrII、PvuI、XmaIII、SphI、XhoI、PstI、AflII、EcoRI、AatII、SacI、EcoRV、SphI、NaeI、BsePI、NheI、BamHI、NarI、ApaI、KpnI、SnaI、SalI、ApaLI、HpaI、SnaBI、BspHI、BspMII、NruI、XbaI、BclI、MstI、Bali、Bsp1407I、PsiI、AsuII 和 AhaIII。

识别超过 6 bp 的序列的 DNA 的限制酶包括但不限于 BbvC I、AscI、

AsiS I、Fse I、Not I、Pac I、Pme I、Sbf I、SgrA I、Swa I、Sap I、Cci NI、FspA I、Mss I、Sgf I、Smi I、Srf I 和 Sse8387 I。

对于本发明的一些方面，对于识别 6 bp 序列的限制酶来说，优选 *Bgl*III、*Hind*III 或 *Eco*RI。

识别 4 或 5 个 bp 序列 DNA 的限制酶包括但不限于 *Tsp*EI、*Mae*II、*Alu*I、*Nla*III、*Hpa*II、*Fnu*DII、*Mae*I、*Dpn*I、*Mbo*I、*Hha*I、*Hae*III、*Rsa*I、*Taq*I、*Cvi*RI、*Mse*I、*Sth*132I、*Acc*I、*Dpn*II、*Sau*3AI 和 *Mnl*II。在一个实施方案中，第二限制酶是 *Nla*III 和/或 *Dpn*II。

术语“第一限制酶识别位点”指被第一限制酶识别和切割的核苷酸序列中的位点。

对于一些实施方案，限制酶不消化重复的 DNA 或重复的 DNA 中相对不具代表性的 DNA。这可以增加可解读的读取的数量。

第二限制酶

本文所用的术语“第二限制酶”指第一限制酶消化、连接交联的 DNA、去交联并(任选)纯化 DNA 之后任选使用的第二限制酶。在一个实施方案中，第二限制酶被用于为目的核苷酸序列提供确定的 DNA 末端，从而能将已知核苷酸组成的序列与目的核苷酸序列侧翼的第二限制酶识别位点连接。

在一个实施方案中，将已知核苷酸组成的序列与目的核苷酸序列侧翼(如在每一侧或每一端)的第二限制酶识别位点连接，涉及在稀释的条件下连接以促进在靶核苷酸序列侧翼的第二限制酶识别位点和连接的目的核苷酸序列之间的分子内连接。这有效地导致了DNA环的形成，其中已知靶核苷酸序列位于目的未知序列的侧翼。

在另一个实施方案中，将已知核苷酸组成的序列连接于在目的核苷酸序列侧翼(如在每一侧或每一端)的第二限制酶识别位点上涉及添加核苷酸组成已知的独特DNA序列，然后在促进在目的核苷酸序列侧翼的第二限制酶识别位点和导入的已知核苷酸组成的独特DNA序列之间的分子间连接的条件下进行连接。

在一个实施方案中，选择第二限制酶从而不使第二限制酶位点在距离第一限制性位点约 350bp(如 350-400bp)以内处。

在另一个实施方案中，选择第二限制酶从而使同一第二限制酶位点很可

能位于连接的核苷酸序列(即连接的交联序列)中。由于第一(靶)核苷酸序列和连接的核苷酸序列的末端可以是匹配的粘(或平)末端,因此所述序列甚至可以连接从而使 DNA 环化。因此,消化步骤后,在促进分子内相互作用的稀释条件下连接,并任选通过匹配的末端使 DNA 环化。

优选地,第二限制酶识别位点是 4 或 5 bp 核苷酸序列识别位点。识别 4 或 5 bp 序列的 DNA 的酶包括但不限于 TspEI、MaeII、AluI、NlaIII、HpaII、FnuDII、MaeI、DpnI、MboI、HhaI、HaeIII、RsaI、TaqI、CviRI、MseI、Sth132I、AciI、DpnII、Sau3AI 和 MnlI。

在优选的实施方案中,第二限制酶是 NlaIII 和/或 DpnII。

术语“第二限制酶识别位点”指核苷酸序列中被第二限制酶识别和切割的位点。

用第二限制酶消化后,进行进一步的连接反应。在一个实施方案中,该连接反应将已知核苷酸序列组成的 DNA 序列与一个或多个与靶核苷酸序列相连的序列上的第二限制酶消化位点连接。

对一些实施方案,本方法不包括用第二限制酶消化核苷酸序列的步骤。

对一些实施方案,本方法不包括将一个或多个已知核苷酸组成的 DNA 序列连接到一个或多个目的靶序列侧翼的可用第二限制酶位点。

第三限制酶

本文所用的术语“第三限制酶”指第二限制酶步骤后为了在扩增前线性化环化的 DNA 而可任选使用的第三限制酶。

第三限制酶优选是识别 6bp 或更多的核苷酸识别位点的酶。

第三限制酶优选消化第一和第二限制酶识别位点之间的第一(靶)核苷酸序列。如普通技术人员所能理解的,希望第三限制酶在消化第一(靶)核苷酸序列时不太靠近第一和第二限制酶识别位点,以防使扩增的引物不再杂交。因此,优选第三限制酶识别位点至少与第一和第二限制酶识别位点相距与所用的引物长度相同的距离,从而使一个或多个扩增引物仍能杂交。

在优选的实施方案中,第三限制酶是识别 6-bp DNA 序列的限制酶。

术语“第三限制酶识别位点”指核苷酸序列中被第三限制酶识别并切割的位点。

识别位点

限制酶是裂解 DNA 糖-磷酸骨架的酶。在最实用的配置中，给定的限制酶切割一段仅几个碱基长的双链 DNA 的两条链。限制酶的底物是被称为识别位点/序列的双链 DNA 序列。

限制性识别位点的长度可变化，这取决于所用的限制酶。识别序列的长度控制酶将如何频繁地在 DNA 序列中进行切割。

举例来说，许多限制酶识别 4 bp 的 DNA 序列。序列和识别 4 bp 的 DNA 序列的酶包括但不限于 AATT(TspEI)、ACGT(MaeII)、AGCT(AluI)、CATG(NlaIII)、CCGG(HpaII)、CGCG(FnuDII)、CTAG(MaeI)、GATC(DpnI、DpnII、Sau3AI & MboI)、GCGC(HhaI)、GGCC(HaeIII)、GTAC(RsaI)、TCGA(TaqI)、TGCA(CviRI)、TTAA(MseI)、CCCG(Sth132I)、CCGC(AciI) 和 CCTC(MnlI)。

进一步举例来说，许多限制酶识别 6 bp 的 DNA 序列。序列和识别 6 个碱基对的 DNA 序列的酶包括但不限于 AACGTT(AclI)、AAGCTT(HindIII)、AATATT(SspI)、ACATGT(BspLU11I)、ACCGGT(AgeI)、ACGCGT(MluI)、ACTAGT(SpeI)、AGATCT(BglII)、AGCGCT(Eco47III)、AGGCCT(StuI)、AGTACT(ScaI)、ATCGAT(ClaI)、ATGCAT(AvaIII)、ATTAAT(VspI)、CAATTG(MfeI)、CACGTG(PmaCI)、CAGCTG(PvuII)、CATATG(NdeI)、CCATGG(NcoI)、CCCGGG(SmaI)、CCGCGG(SacII)、CCTAGG(AvrII)、CGATCG(PvuI)、CGGCCG(XmaIII)、CGTACG(SplI)、CTCGAG(XhoI)、CTGCAG(PstI)、CTTAAG(AflII)、GAATTC(EcoRI)、GACGTC(AatII)、GAGCTC(SacI)、GATATC(EcoRV)、GCATGC(SphI)、GCCGGC(NaeI)、GCGCGC(BsePI)、GCTAGC(NheI)、GGATCC(BamHI)、GGCGCC(NarI)、GGGCCC(ApaI)、GGTACC(KpnI)、GTATAC(SnaI)、GTCGAC(SalI)、GTGCAC(ApaLI)、GTTAAC(HpaI)、TACGTA(SnaBI)、TCATGA(BspHI)、TCCGGA(BspMII)、TCGCGA(NruI)、TCTAGA(XbaI)、TGATCA(BclI)、TGCGCA(MstI)、TGGCCA(BalI)、TGTACA(Bsp1407I)、TTATAA(PsiI)、TTCGAA(AsuII)和 TTTAAA(AhaIII)。

进一步举例来说，许多限制酶识别 7 bp 的 DNA 序列。序列和识别 7 bp 的 DNA 序列的酶包括但不限于 CCTNAGG(SauI)、GCTNAGC(EspI)、GGTNACC BstEII和 TCCNGGA PfoI。

进一步举例来说,许多限制酶识别 8 bp 的 DNA 序列。序列和识别 8 bp 的 DNA 序列的酶包括但不限于 ATTTAAAT(SwaI)、CCTGCAGG(Sse8387I)、CGCCGGCG(Sse232I)、CGTCGACG(SgrDI)、GCCCGGGC(SrfI)、GCGATCGC(SgfI)、GCGGCCGC(NotI)、GGCCGGCC(FseI)、GGCGCGCC(AscI)、GTTTAAAC(PmeI)和 TTAATTAA(PacI)。

许多这些序列包含序列 CG, 序列 CG 可在体内甲基化。许多限制酶对这种甲基化敏感, 而且将不会切割甲基化的序列, 如 HpaII 不会切割序列 CC^mGG, 而其同裂酶 MspI 对该修饰不敏感, 能切割甲基化的序列。因此, 在一些情况下, 不使用真核甲基化敏感的酶。

在一个实施方案中, 识别位点是消化位点。

在一个实施方案中, 限制酶识别位点是限制酶消化位点。

环化

根据本发明的一个实施方案, 用于 4C 的材料通过产生 DNA 环来制备, 其中用第二限制酶消化 3C 模板, 然后连接。

优选地, 选择在距离第一限制性位点大于约 350bp(如 350-400bp)处切割第一(靶)核苷酸序列的第二限制酶。有益的是, 这使因拓扑学约束而存在于环形成中的偏倚最小化(Rippe 等.(2001)*Trends in Biochem. Sciences* 26, 733-40)。

第二限制酶优选是识别 4 或 5 bp 限制酶识别位点的频繁切割分子。因此可能获得对扩增期间所有连接的片段有相等扩增效率的最小限制性片段。

在第二限制酶消化和连接之前, DNA 模板将包含第一(靶)核苷酸序列中的一个第二酶识别位点, 其与第一限制酶位点相距大于约 350-400bp, 和另一个第二酶识别位点, 其位于已经连接的核苷酸序列中(即在第二核苷酸序列中)。

第二限制酶消化步骤优选进行 1 小时以上至过夜, 然后使酶热失活。

该反应混合物中的 DNA 优选使用本领域已知的常规方法/试剂盒来纯化。

在第二限制酶消化步骤之后, 第二限制酶位点将与第一(靶)核苷酸序列中的第一限制酶位点相距大于 350-400bp, 而且另一个第二限制酶位点将位于被连接的核苷酸序列(即第二核苷酸序列)中。由于第一(靶)核苷酸序列和被连接的核苷酸序列的末端具有匹配的末端, 因此所述序列可连接以使 DNA

环化。

消化步骤后,在有利于分子内相互作用的稀释的条件下连接,并通过匹配的末端使 DNA 环化。

连接反应优选以约 1-5 ng/ μ l 的 DNA 浓度来进行。

连接反应优选在约 16-25 °C 进行超过 1 小时(如 2、3、4 或更多小时)。

因此,连接反应后,可制备环化的 DNA。环化的 DNA 将包含至少第二限制酶或第一和第二限制酶的识别位点。在包含第一(靶)核苷酸序列的环化的 DNA 中,第一限制酶识别位点和第二限制酶识别位点将限定第一(靶)核苷酸序列和被连接的核苷酸序列(即第二核苷酸序列)的末端。因此,第一限制酶识别位点和第二限制酶识别位点将使第一(靶)核苷酸序列和被连接的核苷酸序列隔开(如分开)。

扩增

可进行一个或多个扩增反应以扩增 4C DNA 模板。

可利用许多本领域已知的不同方法来进行 DNA 扩增。例如,可利用聚合酶链式反应(Saiki 等., 1988); 连接介导的 PCR、Qb 复制酶扩增(Cahill, Foster 和 Mahan, 1991; Chetverin 和 Spirin, 1995; Katanaev, Kurnasov 和 Spirin, 1995); 连接酶链式反应(LCR)(Landegren 等., 1988; Barany, 1991); 自动维持序列复制系统(Fahy, Kwoh 和 Gingeras, 1991)和链置换扩增(Walker 等., 1992)来扩增 DNA。

适宜用 PCR 来扩增 DNA。“PCR”指 K. B. Mullis 的美国专利 4,683,195、4,683,202 和 4,965,188 的方法,其描述了一种不进行克隆或纯化而增加基因组 DNA 混合物中核苷酸序列片段浓度的方法。

在一个实施方案中,使用反向 PCR。(Ochman 等(1988)*Genetics* 120(3), 621-3 所述的)反向 PCR(IPCR)是一种用以快速体外扩增已知序列区域侧翼的 DNA 序列的方法。该方法使用聚合酶链式反应(PCR),但是它具有与常规方向相反方向的引物。反向引物的模板是自身连接成环的限制性片段。反向 PCR 在分子遗传学中有许多应用,例如,扩增和鉴定在转座元件侧翼的序列。为了增加扩增的有效性和可重复性,优选在扩增之前将 DNA 环用第三限制酶线性化。优选使用第三限制酶,其是 6 bp 或更多 bp 的切割分子。第三限制酶优选切割第一和第二限制酶位点之间的第一(靶)核苷酸序列。

用第二限制酶消化 3C 模板,任选进行环化,连接(如在稀释条件下连接)并任选将含第一(靶)核苷酸序列的环线性化,产生用于扩增的 DNA 模板(“4C DNA 模板”).

对于扩增步骤,使用至少 2 种寡核苷酸引物,其中每一种引物与在目的核苷酸序列的侧翼的 DNA 序列杂交。在优选的实施方案中,使用至少 2 种寡核苷酸引物,其中每一种引物与在目的核苷酸序列的侧翼的靶序列杂交。

在一个实施方案中,在引物杂交的上下文中的术语“在……侧翼/侧翼包围”指至少一种引物与和目的核苷酸序列的一个末端(如 5'末端)相邻的 DNA 序列杂交,而且至少一种引物与在目的核苷酸序列的另一个末端(如 3'末端)上的 DNA 序列杂交。优选至少一种正向引物与和目的核苷酸序列的一个末端(如 5'末端)相邻的 DNA 序列杂交,而且至少一种反向引物与在目的核苷酸序列的另一个末端(如 3'末端)上的 DNA 序列杂交。

在优选的实施方案中,在引物杂交的上下文中的术语“在……侧翼/侧翼包围”指至少一种引物与和目的核苷酸序列的一个末端(如 5'末端)相邻的靶序列杂交,而且至少一种引物与在目的核苷酸序列的另一个末端(如 3'末端)上的靶序列杂交。优选至少一种正向引物与和目的核苷酸序列的一个末端(如 5'末端)相邻的靶序列杂交,而且至少一种反向引物与在目的核苷酸序列的另一个末端(如 3'末端)上的靶序列杂交。

本文所用的术语“引物”指寡核苷酸,无论是以纯化的限制性消化物的形式天然存在的还是合成生产的,其都能在置于诱导合成与核酸链互补的引物延伸产物的条件下(即,在核苷酸和诱导剂(如 DNA 聚合酶)存在时,并且以合适的温度和 pH)时用作合成起始点。引物优选是单链的,从而有最大扩增效率,但可以是双链的。如果是双链的,则首先处理引物以分开其链,然后用于制备延伸产物。引物优选是寡脱氧核糖核苷酸。引物必须足够长到能在诱导剂存在的条件下引发延伸产物的合成。引物的确切长度将取决于许多因素,包括温度、引物来源和所用的方法。

适宜的是,引物长度将是至少 15、优选至少 16、17、18、19 或 20、例如至少 25、30 或 40 个核苷酸。优选扩增引物长度为 16 至 30 个核苷酸。

优选将引物设计成尽可能靠近分隔第一(靶)核苷酸序列和第二(捕获的)核苷酸序列的第一和第二限制酶识别位点。可设计引物从而使它们相距第一和第二限制酶识别位点在约 100 个核苷酸之内——如约 90、80、70、60、50、

40、30、20、10、9、8、7、6、5、4、3、2或1个核苷酸。

适宜的是，可设计引物从而使它们部分或完全与第一和第二限制酶识别位点重叠。

适宜的是，设计扩增引物以使它们的3'末端向外朝着第二核苷酸序列。

在一个实施方案中，扩增可与向扩增产物的末端附加额外的序列相结合。优选地，这些额外的序列是高通量测序所需的衔接序列。引物适宜包含突出端(如5'突出端)。突出端适宜地附加高通量测序所需的部分或完整的衔接序列。引物适宜地包含附加部分或完整序列的突出端(如5'突出端)，该序列在高通量测序中用于引导(priming)测序反应。因此，在一个实施方案中，引物部分或完全地与第一和/或第二限制酶识别位点重叠，衔接子和测序引导序列作为5'突出端附加在引物上。另外，或作为代替，引物包含允许PCR产物随后从基因组4C模板上分离的缀合模块(conjugated moiety)(如生物素)。

如果所用的扩增方法是反向PCR，则优选对约100-400 ng 4C模板DNA(在每个约50 μ l PCR反应混合物中)或对重复PCR反应能得到可重复结果(参见图1)并在每个PCR反应中包括最大数量的连接事件的其它量的DNA进行扩增反应。

优选根据制造商说明利用缓冲液1，用Expand Long Template PCR系统(扩展长模板PCR系统)(Roche)，进行反向PCR扩增反应。

样品

本文所用的术语“样品”具有其正常的含义。样品可以是任何物理学实体，其包含交联或能交联的DNA。样品可以是或可以源自生物材料。

样品可以是或可以源自一个或多个实体——如一个或多个细胞、一个或多个核、或一个或多个组织样品。实体可以是或可以源自其中存在DNA，如染色质的任何实体。样品可以是或可以源自一个或多个分离的细胞或一个或多个分离的组织样品、或一个或多个分离的核。

样品可以是或可以源自活细胞和/或死细胞和/或核溶解产物和/或分离的染色质。

样品可以是或可以源自患病和/或没有患病的受试者。

样品可以是或可以源自疑似罹患疾病的受试者。

样品可以是或可以源自要测试他们将来将患上疾病的可能性的受试者。

样品可以是或可以源自活的或非活体的患者材料。

Splinter 等., (2004) *Methods Enzymol.* 375, 493-507 中详细描述了固定细胞和组织用于制备 3C 模板。

标记

优选, 核苷酸序列(如扩增的 4C DNA 模板、引物或探针等)是标记的, 从而辅助它们的下游应用——如阵列杂交。举例来说, 4C DNA 模板可利用随机引发或切口平移来标记。

可用许多种标记物(如报告分子)来标记本文所述的核苷酸序列, 尤其在扩增步骤中。合适的标记包括放射性核素、酶、荧光、化学发光、或显色剂以及底物、辅助因子、抑制剂、磁性颗粒等。教导应用这些标记的专利包括 US-A-3817837 ; US-A-3850752 ; US-A-3939350 ; US-A-3996345 ; US-A-4277437; US-A-4275149 和 US-A-4366241。

其它标记包括而不限于 β -半乳糖苷酶、转化酶、绿色荧光蛋白、萤光素酶、氯霉素、乙酰转移酶、 β -葡糖醛酸糖苷酶、外切葡聚糖酶和葡糖淀粉酶。也可用荧光标记、以及带有特定化学性质的特异合成的荧光试剂。可用许多种测量荧光的方式。例如, 一些荧光标记展示出激发或吸收光谱的变化, 一些当其中一种荧光报告分子放出荧光而第二种吸收荧光时展示出共振能量转移, 一些展示出荧光丧失(淬灭)或显示荧光, 而一些报告旋转运动。

为了获得足以进行标记的材料, 可汇集多次扩增, 而不用增加每个反应中的扩增循环数量。另外, 标记的核苷酸可被纳入最后几个扩增反应循环中(如 30 个循环的 PCR(无标记)+ 10 个循环的 PCR(有标记))。

阵列

在特别有益的实施方案中, 可根据本文所述方法制备的 4C DNA 模板与阵列杂交。因此, 可用阵列(如微阵列)技术来鉴定频繁与第一(靶)核苷酸序列共有核位点的核苷酸序列(如基因组片段)。

根据本发明, 可用现有的阵列, 如表达和基因组阵列。可是, 本发明还寻求提供如本文所述的新阵列(如 DNA 阵列)。

“阵列”是有意产生的核酸集合, 其可通过合成方法或生物合成方法制备, 并可以各种不同形式(如, 可溶性分子文库; 和连在(tether to)树脂珠、

硅片、或其它固相支持物上的寡聚物文库)来筛选生物活性。另外,术语“阵列”包括那些通过将基本上为任意长度(如,从1至约1000个核苷酸单体长度)的核酸点样在基质上而制备的核酸的文库。

阵列技术和与其相关的各种技术和应用一般在大量教科书和文献中有描述。这些包括 Lemieux 等., 1998, *Molecular Breeding* 4, 277-289, Schena 和 Davis. *Parallel Analysis with Biological Chips*. 摘自 *PCR Methods Manual*(M. Innis, D. Gelfand, J. Sninsky 编), Schena 和 Davis, 1999, *Genes, Genomes and Chips*. 摘自 *DNA Microarrays: A Practical Approach*(M. Schena 编), Oxford University Press, Oxford, 英国, 1999), *The Chipping Forecast*(*Nature Genetics* 特刊; 1999 年 1 月增刊), Mark Schena(编), *Microarray Biochip Technology*, (Eaton Publishing Company), Cortes, 2000, *The Scientist* 14[17]: 25, Gwynne 和 Page, *Microarray analysis: the next revolution in molecular biology*, *Science*, 1999 年 8 月 6 日; 和 Eakins 和 Chu, 1999, *Trends in Biotechnology*, 17, 217-218。

阵列技术克服了传统分子生物学方法的缺点,传统分子生物学方法一般以“一个实验中一个基因”为基础进行作业,是低通量的并且不能形成基因功能的“全景图”。当前,阵列技术的主要应用包括鉴定序列(基因/基因突变)和确定基因表达水平(丰度)。基因表达谱可运用阵列技术,任选还合并运用蛋白质组技术(Celis 等, 2000, *FEBS Lett*, 480(1): 2-16; Lockhart 和 Winzeler, 2000, *Nature* 405(6788): 827-836; Khan 等., 1999, 20(2): 223-9)。阵列技术的其它应用也是本领域已知的;例如,基因发现、癌症研究(Marx, 2000, *Science* 289: 1670-1672; Scherf, 等, 2000, *Nat Genet*; 24(3): 236-44; Ross 等, 2000, *Nat Genet*. 2000 Mar; 24(3): 227-35)、SNP 分析(Wang 等, 1998, *Science*, 280(5366): 1077-82)、药物发现、药物基因组学、疾病诊断(例如,利用微流体学设备: *Chemical & Engineering News*, 1999 年 2 月 22 日, 77(8): 27-36)、毒理学(Rockett 和 Dix(2000), *Xenobiotica*, 30(2): 155-77; Afshari 等., 1999, *Cancer Res* 1; 59(19): 4759-60)和毒理基因组学(综合了功能基因组学和分子毒理学的学科)。

一般而言,任何文库都可通过在空间上分隔文库的成员来以有序方式排列成阵列。合适的阵列文库的例子包括核酸库(包括 DNA、cDNA、寡核苷酸等文库)、肽、多肽和蛋白质文库、以及包含任何分子的文库,如配体库,

以及其它。

样品(如,文库的成员)一般被固定或固定化在固相上,优选在固体基质上,由此限制样品的扩散和混合。在优选的实施方案中,可制备结合配体的DNA文库。尤其可将文库固定在基本上平的固相上,包括膜和无孔基质(如塑料和玻璃)上。另外,样品优选以方便标引(即,提供对特定样品的参考或提取)的方式排列。通常样品被用作网格形式中的点。为此可修改普通测试系统。例如,阵列可固定在微量平板表面上,一个孔中有多个样品,或每个孔中有单个样品。此外,固体基质可以是膜,如硝化纤维素或尼龙膜(例如,用于印迹实验中的膜)。其它可选的基质包括基于玻璃或硅的基质。因此,样品可用任何合适的本领域已知的方法来固定,例如,通过电荷相互作用,或通过化学偶联于孔壁上或孔底部、或膜表面上。可用其它排列和固定方法,例如,用吸管点、滴落接触(drop-touch)、压电法、喷墨和喷泡沫技术、静电应用等。对于基于硅的芯片来说,可用光版印刷来将样品排列并固定于芯片上。

可以通过“点”在固体基质上来排列样品;这可以手工进行或通过使用机器人技术来点样样品。一般而言,阵列可被描述成大阵列或微阵列,其区别是样品点的大小。大阵列通常含约300微米或更大的大小的样品点并可方便地通过现有的凝胶和印迹扫描仪来成像。微阵列中的样品点直径大小通常小于200微米,而且这些阵列通常含上千个点。因此,微阵列需要专门的机器人技术和成像设备,其可能需要定制。所用的仪器在Cortese, 2000, *The Scientist* 14[11]: 26中有一般性的综述。

产生固定的DNA分子文库的技术在现有技术中有过描述。一般而言,大多数现有技术方法描述了如何合成单链核酸分子文库,其利用诸如隐蔽技术在固体基质上的各种离散位置处产生各种序列改变。美国专利5,837,832描述了一种基于非常大规模的整合技术的改进方法来产生固定于硅基质上的DNA阵列。具体而言,美国专利5,837,832描述了一种被称为“铺瓦”的策略,在基质上空间确定的位置上合成特定探针组,其可用于产生本发明固定的DNA文库。美国专利5,837,832也提供了对也可使用的较早技术的参考。

也可用光沉积化学(photo deposition chemistry)来制造阵列。

肽(或肽模拟物)阵列也可以在表面上合成,其方式是将每个独特的文库成员(如,独特的肽序列)置于离散的、预先确定的阵列位置上。每个文库成

员的身份是由其在阵列中的空间位置确定的。确定阵列中预先确定的分子(如, 靶或探针)和反应性文库成员之间发生的结合性相互作用的位置, 由此基于空间位置来鉴定反应性文库成员的序列。这些方法描述于美国专利 5.143.854; WO90/15070 和 WO92/10092; Fodor 等.(1991)*Science*, 251: 767; Dower 和 Fodor(1991)*Ann. Rep. Med. Chem.*, 26: 271 中。

为了帮助检测, 通常使用(如上所述的)标记, 如任何方便检测的报道分子, 例如, 荧光、生物发光、磷光、放射性物质等报道分子。这些报道分子、其检测、与靶/探针的偶联等在本文的其它地方有讨论。探针和靶的标记还公开于 Shalon 等., 1996, *Genome Res* 6(7): 639-45 中。

DNA 阵列的特定例子如下:

I 型: 利用机器人点样将探针 cDNA(500~5,000 个碱基长)固定于固体表面(如玻璃), 并将其暴露于一组分离或混合的靶。该方法被广泛认为是由斯坦福大学开发的(Ekins 和 Chu, 1999, *Trends in BioTechnology*, 1999, 17, 217-218)。

II 型: 将寡核苷酸(20-25 聚寡聚物, 优选是 40-60 聚寡聚物)或肽核酸(PNA)探针阵列原位(芯片上)合成或常规合成, 之后固定在芯片上。将阵列暴露于标记的样品 DNA, 杂交, 并确定互补序列的身份/丰度。该 DNA 芯片由 Affymetrix, Inc.以 GeneChip®商标销售。Agilent 和 Nimblegen 也提供了合适的阵列(如基因组嵌合阵列(genomic tiling array))。

一些可商购的微阵列型号的例子在下表 1 中列出(也可参见 Marshall 和 Hodgson, 1998, *Nature Biotechnology*, 16(1), 27-31)。

公司	产品名	排列方法	杂交步骤	读数方式
Affymetrix, Inc., Santa Clara, California	GeneChip®	原位(芯片上)光版 印刷合成~20-25 聚寡聚物于硅片 上, 将其切割成 1.25 cm ² 或 5.25 cm ² 见方的芯片	10,000-260,000 个寡 聚物的特征用样品 cDNA 或反义 RNA 的标记的 30-40 个核 苷酸片段检测	荧光
Brax, Cambridge, UK		短的合成的寡聚 物, 脱离芯片合成	“通用芯片”上的 1000 个寡聚物用标 记的核酸检测	质谱
Gene Logic,	READS™			

<u>Inc.,</u> Columbia, Maryland				
<u>Genometrix</u> <u>Inc.,</u> The Woodlands, Texas	Universal Arrays™			
<u>GENSET,</u> Paris, 法国				
<u>Hyseq Inc.,</u> Sunnyvale, California	HyChip™	500-2000 nt DNA 样品印在 0.6 cm ² (HyGnostics) 或~18 cm ² (Gene Discovery)膜上 制成的 5 聚寡聚物 以 1,15 cm ² 阵列印 在玻璃(HyChip)上	64 个样品 cDNA 点用 8,000 个 7 聚寡聚物 检测(HyGnostics)或 <=55,000 个样品 cDNA 点用 300 个 7 聚寡聚物检测(Gene Discovery) 通用的 1024 个寡聚 物点可检测 10 kb 样 品 cDNA, 标记的 5 聚寡聚物、和连接酶	放射性同 位素 荧光
<u>Incyte</u> <u>Pharmaceuticals</u> <u>, Inc.,</u> Palo Alto, California	GEM	压电印制点样 PCR 片段并在芯 片上合成寡聚物	<=1000(最多为 10,000)个寡聚物 /PCR 片段点用标记 的 RNA 检测	荧光和放 射性同位 素
<u>Molecular</u> <u>Dynamics, Inc.,</u> Sunnyvale, California	Storm® FluorImager®	500-5000 nt cDNA 用笔印在~10 cm ² 的玻璃片上	~10,000 个 cDNA 点 用 200-400 nt 标记的 样品 cDNA 检测	荧光
<u>Nanogen, San</u> <u>Diego,</u> California	半导体 微芯片	预制的~20 聚寡聚 物, 其被捕获于硅 片上的电话活性点 上, 其被切成<=1 cm ² 见方的芯片	25、64、400(并最多 为 10,000)个寡聚物 点被极化以增强与 200-400 nt 标记的样 品 cDNA 的杂交	荧光
<u>Protogene</u> <u>Laboratories,</u> Palo Alto, California		通过印在表面张 力阵列上来将 40-50 聚寡聚物用 芯片合成在 9 cm ² 玻璃芯片上	<=8,000 个寡聚物点 用 200-400 nt 标记的 样品核酸检测	荧光
<u>Sequenom,</u>	大阵列	胶印阵列; 约	通过激光解吸附和质	质谱

Hamburg, Germany, 和 San Diego, California	光谱芯片 (SpectroChip)	20-25 聚寡聚物	谱检测每个 SpectroChip 的 250 个 位置	
Synteni, Inc., Fremont, California	UniGEM™	500-5,000 nt cDNA 用尖状物印 在~4 cm ² 玻璃芯 片上	<=10,000 个 cDNA 点 可用 200-400 nt 标记 的样品 cDNA 检测	荧光
Nimblegen Systems Inc., Madison	智人全基因 组 60 聚微阵 列	38,000 个转录物, 每基因 5 个探针 17.4mm x 13mm		5-微米扫 描平台
The German Cancer Institute, Heidelberg, 德 国		原型 PNA 大芯 片, 其利用 f-moc 或 t-moc 化学而在 芯片上合成的探 针	8 x 12 cm 芯片上有 约 1,000 个点	荧光/质谱

表 1: 当前可用的杂交微阵列型号的例子

为了从基于阵列的测定中产生数据, 检测表明探针和核苷酸序列间存在或缺少杂交的信号。本发明还关注直接和间接标记技术。例如, 直接标记将荧光染料直接纳入与同阵列相连的探针杂交的核苷酸序列上(如, 在存在标记的核苷酸或 PCR 引物的情况下通过酶催化合成将染料纳入核苷酸序列)。直接标记方案能产生强杂交信号, 其通常利用化学结构和特征相似的荧光染料家族, 并易于实现。在优选的包括直接标记核酸的实施方案中, 将菁蓝或 alexa 类似物用于多重荧光比较阵列分析中。在其它实施方案中, 间接标记方案可用于在与微阵列探针杂交之前或之后将表位纳入核酸上。一种或多种染色过程和试剂可用于标记杂交的复合物(如, 结合表位的荧光分子, 由此通过与杂交物种的表位缀合的染料分子提供荧光信号)。

数据分析也是涉及阵列的实验中的重要部分。阵列实验中得到的未加工的数据通常是图像, 其需要转成矩阵——表, 其中行代表诸如基因, 列代表诸如各种样品(如组构)或实验条件, 而且每格中的数字表征诸如特定序列(优选是与第一(靶)核苷酸序列连接的第二序列)在特定样品中的表达。如果要提取出任何有关的潜在生物学过程的知识, 这些矩阵必须进一步被分析。数据分析方法(包括有指导的和无指导的数据分析以及生物信息学方法)被公开于

Brazma 和 Vilo J(2000)FEBS Lett 480(1): 17-24 中。

如本文所述, 标记的并之后与阵列杂交的一个或多个核苷酸序列(如 DNA 模板)包含富含带有独特特征的小段序列的核苷酸序列, 即横跨在 3C 过程期间与第一(靶)核苷酸序列相连的第一限制酶识别位点和它们各自相邻的第二限制酶识别位点之间的核苷酸序列。

单个阵列可包含多种(如两种或更多种)诱饵序列。

染色体图

本文还描述了涉及由于它们的空间或物理接近度而交联的DNA片段的表征方法, 该方法以(重新)建立染色体图(如线性染色体图)和诊断性鉴定这些染色体图中的变化为目的。

有益的是, 此类技术可依诊断目的而应用-例如重建物理染色体图并鉴定由于基因组重排而在这些图中产生的变化。

此方法也可用于鉴定不一定与疾病相关但可使受试者易出现某些性状(如精神或行为性状)的(天然的)基因组变异。

在一个方面, 提供了建立个体的至少一个线性染色体图的方法, 其包含以下步骤: (a)提供来自所述个体的核酸样品; (b)交联所述样品内的核酸; (c)用第一限制酶消化核酸; (d)连接交联的核苷酸序列; (e)解除交联; (f)分析连接产物; (g)建立至少一个线性染色体图; 和(h)鉴定所述个体线性染色体图上的一个或多个基因组重排。

在另一方面, 提供了诊断疾病或鉴定由染色体上的一个或多个基因组重排而引起的性状的方法, 其包括以下步骤: (a)用第一限制酶消化交联DNA的样品; (b)连接交联的核苷酸序列; (c)解除交联; (d)分析连接产物; (e)建立至少一个线性染色体图; (f)鉴定线性染色体图上的一个或多个基因组重排; 和(g)将一个或多个染色体重排与疾病或性状相关联。

探针

本文所用的术语“探针”指分子(如, 寡核苷酸, 无论其是以纯化的限制性消化物形式天然产生的还是以合成方法、重组方法或通过 PCR 扩增而产生的), 其能够与另一种目的分子(如, 另一种寡核苷酸)杂交。当探针是寡核苷酸时, 它们可以是单链或双链的。探针可用于检测、鉴定和分离特定的靶(如,

基因序列)。如本文所述,关注的是,本发明所用的探针可以是用标记物标记的,从而可在任何检测系统中检测,其包括但不限于酶(如,ELISA、以及基于酶的组织化学测定法)、荧光、放射性和发光系统。

至于阵列和微阵列,术语“探针”用于指任何为了检测已经与所述探针杂交的核苷酸序列而固定于阵列上的可杂交材料。这些探针优选是 25-60 聚或更长的。

探针设计策略被描述于 WO95/11995、EP 717,113 和 WO97/29212 中。

由于 4C 能对相互作用进行无偏的基因组广度的搜索,因此其能有益地制备阵列,其中带有的探针能探测基因组中每个可能的(如独特/非重复的)第一限制酶识别位点。因此,阵列设计仅依赖于第一限制酶的选择,而不依赖于实际的第一或第二核苷酸序列。

尽管可根据本发明所述使用现有的阵列,优选使用其它构型。

在一种构型中,设计阵列上的一个或多个探针,从而使它们能杂交于由第一限制酶消化的位点附近。一个或多个探针更优选是在第一限制酶识别位点约 20 bp 内。一个或多个探针更优选是在第一限制酶识别位点约 50 bp 内。

适宜的是,一个或多个探针是在第一限制酶识别位点约 100 bp (如约 0-100 bp、约 20-100 bp)内。

在一种优选的构型中,单个、独特的探针被设计在由第一限制酶消化的位点之每一侧的 100 bp 内。

在另一种优选的构型中,相对于第一限制酶消化的位点位置的第二限制酶消化的位点位置也纳入考量。在这种构型中,单个、独特的探针仅被设计在第一限制酶消化的位点之每一侧上,其与最接近的第二限制酶识别位点距离大到足够在第一和第二限制酶识别位点间设计给定长度的探针。例如,在这种构型中,没有探针被设计成位于特定第一限制酶识别位点的一侧而该侧的第二限制酶识别位点在 10 bp 之内。

在另一种构型中,设计阵列上的探针,从而使它们能与第一限制酶消化的位点之任一侧杂交。适宜的是,可使用第一限制酶识别位点每一侧上的单个探针。

在又一种构型中,两种或更多种(如 3、4、5、6、7 或 8 或更多)探针可被设计在第一限制酶识别位点之每一侧上,然后其可用于研究同一连接事件。对于探针相对于每一个第一限制酶识别位点的数量和位置,可考量其相

邻的第二限制酶识别位点的确切基因组位置。

在又一种构造中,两种或更多种(如3、4、5、6、7或8或更多)探针可被设计成位于每个第一限制酶识别位点附近,而不考虑最接近的第二限制酶识别位点。在这种构型中,所有探针应仍旧靠近第一限制酶识别位点(优选在该限制性位点的300 bp内)。

有益的是,后者的设计,还有每个第一限制酶识别位点(的一边)使用1个探针的设计,都能应用不同的第二限制酶与给定的第一限制酶组合。

有益的是,每个第一限制酶识别位点应用多个(如2、3、4、5、6、7或8或更多)探针可使由于单个探针的不良表现而得到假阴性结果的问题最小化。而且,它也可增加单个芯片实验中得到的数据的可靠性,并减低得到统计学上可信结论所需的阵列数量。

用于阵列中的探针可大于40个核苷酸长而且可以是等温的。

优选排除含重复DNA序列的探针。

用于侦测在第一核苷酸序列紧邻的侧翼或附近的限制酶位点的探针预计能带来非常强的杂交信号,并也可排除在探针设计之外。

阵列可涵盖任何基因组,包括哺乳动物(如人、小鼠(如染色体7))、脊椎动物(如斑马鱼))、或非脊椎动物(如细菌、酵母、真菌或昆虫(如果蝇))基因组。

在进一步优选的实施方案中,阵列含在每个独特的第一限制酶位点周围、而且尽可能地接近限制酶消化位点的2-6个探针。

距限制酶消化位点的最大距离优选是约300 bp。

在本发明进一步优选的实施方案中,提供了针对限制酶(如 *HindIII*、*EcoRI*、*BglII* 和 *NotI*)的阵列,其涵盖了哺乳动物或非哺乳动物基因组。有益的是,本文所述的阵列设计克服了对每个靶序列重新设计阵列的要求,只要在相同物种中进行分析。

探针组

本文所用的术语“探针组”指探针组合或集合,所述探针与基因组中第一限制酶的每一个第一限制酶识别位点杂交。

因此,在另一方面,提供了探针组,所述探针在序列上与和基因组DNA中第一限制酶的每一个第一限制酶识别位点相邻的核酸序列互补。

适宜的是, 探针组在序列上与和基因组 DNA 中每一个第一限制酶识别位点相邻的最初 25-60(如 35-60、45-60、或 50-60)或更多个核苷酸互补。探针组在序列上可与第一限制酶识别位点之一侧(如任一侧)或两侧互补。因此, 探针在序列上可与基因组 DNA 中每一个第一限制酶识别位点之每一侧相邻的核酸序列互补。

也有可能确定在其中设计该组的一个或多个探针的窗口(如距离第一限制酶识别位点 300bp 或更少, 如 250bp、200bp、150bp 或 100bp)。对于确定在其中设计探针的窗口而言重要的因素为: 诸如 GC-含量、不存在能形成发夹结构的回文序列、单一类型核苷酸区段的最大尺寸。因此, 探针组在序列上可与距离基因组 DNA 中每一个第一限制酶识别位点小于 300bp 的核酸序列互补。

也有可能确定距离第一限制酶识别位点约 100 bp 的窗口以鉴定出在每个限制性位点附近的最优探针。

在本发明的其它实施方案中, 探针组与距离基因组 DNA 中每一个第一限制酶识别位点小于 300bp 的序列互补, 与距离基因组 DNA 中每一个第一限制酶识别位点 200-300bp 的序列互补和/或与距离基因组 DNA 中每一个第一限制酶识别位点 100-200bp 的序列互补。

在本发明的其它实施方案中, 探针组与距离基因组 DNA 中每一个第一限制酶识别位点 0-300bp 的序列互补, 与距离基因组 DNA 中每一个第一限制酶识别位点 0-200bp 的序列互补和/或与距离基因组 DNA 中每一个第一限制酶识别位点 0-100bp 的序列互补(如距离基因组 DNA 中每一个第一限制酶识别位点约 10、20、30、40、50、60、70、80 或 90 bp)。

甚至可设计两种或更多种能够与和基因组 DNA 中每个第一限制酶识别位点相邻的序列杂交的探针。

探针可重叠或部分重叠。如果探针重叠, 则优选重叠小于 10 个核苷酸。

也可使用代表在每个第一限制酶识别位点侧翼的最初 1-300 个核苷酸(如 1-20、1-40、1-60、1-80、1-100、1-120、1-140、1-160、1-180、1-200、1-220、1-240、1-260 或 1-280 个核苷酸)的 PCR 片段。

PCR 片段也可用作探针, 其严格对应于其两侧分别为第一限制酶识别位点和第一个相邻的第二限制酶识别位点的每个基因组位点。因此, 探针序列可对应于每一个第一限制酶识别位点和每一个第一个相邻的第二限制酶识

别位点之间的全部或部分序列。

通常, 探针、探针阵列或探针组将固定于支持物上。支持物(如固相支持物)可由各种材料制成, 如玻璃、硅石(silica)、塑料、尼龙或硝化纤维素。支持物优选是刚性的并具有平坦的表面。支持物通常有约 1-10,000,000 个离散的可空间编址的区域、或单元。具有约 10-1,000,000 或约 100-100,000 或约 1000-100,000 个单元的支持物是常见的。单元密度通常至少为每平方厘米内约 1000、10,000、100,000 或 1,000,000 个单元。在一些支持物中, 所有单元都被汇集的探针混合物或探针组占据。在另一些支持物中, 一些单元被汇集的探针混合物或探针组占据, 而其它单元被至少用合成方法可得的纯度程度的单一类型的寡核苷酸占据。

本文所述的阵列优选每个第一限制酶识别位点包含超过一种探针, 例如在切割 6 bp 的限制酶的情况下, 其在每个人或小鼠基因组中出现约 750,000 次。

例如对于识别>6 bp 的识别序列的限制酶, 可用约有 $2 \times 750,000$ 个探针的单个阵列来涵盖完整的人或小鼠基因组, 在每个限制性位点的每一侧上有 1 个探针。

在优选的阵列设计中, 存在于阵列上的给定核苷酸序列的探针分子总数大大超过 4C 样品中存在的要与该阵列杂交的同源片段。考虑到 4C 技术的性质, 代表与线性染色质模板上要分析的核苷酸序列接近的基因组区域的片段将在 4C 杂交样品中大大过量(如图 2 所述)。为了得到有关该丰富片段杂交效率的定量信息, 可能必须减少要杂交的样品量和/或增加阵列上给定的寡核苷酸序列探针的分子数。

因此, 为了检测频繁接触诸如基因启动子元件的 DNA 调节元件, 可能必须使用其中探针仅代表所选的基因组区域(如约 0.5-10 Mb)的阵列, 但其中每个独特探针出现在阵列上多个(如约 100、200、1000 个)位置处。该设计也可优选用于诊断目的以检测位点(如目的基因)周围局部(如在约 10 Mb 内)的基因组重排(如缺失、倒位、重复等)。

阵列可包含约 $3 \times 750,000$ 个探针、 $4 \times 750,000$ 个探针、 $5 \times 750,000$ 个探针、或优选 $6 \times 750,000$ 个探针。阵列更优选包含 $6 \times 750,000$ 个探针, 其中每个限制性位点的每一侧上有 2、3、4、5、6、7 或 8 或更多个探针。阵列最优选包含 $6 \times 750,000$ 个探针, 其中每个限制性位点之每一侧有 3 个探针。

探针阵列或探针组可在支持物上以按步就班的方式合成,或可以预先合成的形式来附着。一种合成方法是 VLSIPS.TM.(如 US 5,143,854 和 EP 476,014 所述),其必须用光指导寡核苷酸探针在高密度、小型化的阵列中合成。用来设计减少合成循环数的掩蔽物(mask)的算法在 US 5,571,639 和 US. 5,593,839 中描述。如 EP 624,059 所述,阵列也可以组合方式通过机械限制的轨道将单体递送到支持物单元来合成。阵列也可通过用喷墨打印机将试剂点样到支持物上来合成(例如参见,EP 728,520)。

在本发明的内容中,术语“实质上的探针组”、“实质上的探针阵列”指探针组或阵列包括至少约 50、60、70、80、90、95、96、97、98 或 99% 的全部或完整的探针组或阵列。探针组或阵列优选是全部或完整的探针组(即 100%)。

在优选的实施方案中,阵列在每个出现于给定基因组的第一限制酶识别位点之每一侧包括单个独特的探针。如果这一探针数超过单个阵列所能包含的探针数,则阵列可优选仍包含给定物种完整基因组的代表,但解析度较低,例如存在于阵列上按序排列在线性染色体模板上的每 2、3、4、5、6、7、8、9、10、 10^2 、 10^3 、 10^4 个等探针中有一个。例如在要找易位配体的情况下,这类以次优化解析度涵盖完整的人或其它基因组的阵列可能比涵盖相同基因组的一部分的高解析度阵列更优选。

较低解析度的给定物种完整基因组的代表优选通过阵列上的探针来获得,所述探针每个代表用第一限制酶消化后得到的单个限制性片段。这优选通过每隔二个、三个、四个、五个、六个、七个、八个、九个、十个、二十个、三十个、四十个、五十个、六十个、七十个、八十个、九十个、或一百个如二个到一百个)等等与相同限制性片段杂交的探针忽略一个而得到。

给定物种完整基因组较低解析度的代表优选包括沿线性染色体模板平均分布的探针。这优选通过忽略那些显示最高探针密度的基因组区域中一个或多个探针来获得。

杂交

本文所用的术语“杂交”应当包括“核酸链通过碱基配对与互补链结合的过程”以及在例如聚合酶链式反应(PCR)技术中进行的扩增过程。

能选择性杂交的核苷酸序列一般将与相应的互补核苷酸序列在至少 20、

优选至少 25 或 30、例如至少 40、60 或 100 或更多个连续核苷酸的区域上是至少 75%、优选至少 85 或 90%和更优选至少 95%或 98%同源的。

“特异杂交”指在严紧条件(如 65°C 和 0.1xSSC {1xSSC = 0.15 M NaCl, 0.015 M 柠檬酸钠 pH 7.0})下使分子只与特定核苷酸序列结合、二聚(duplex)、或杂交。严紧条件是探针将与其靶序列杂交但不与其它序列杂交的条件。严紧条件是序列依赖性的,而且在不同情况下是不同的。较长的序列在较高的温度时特异杂交。一般而言,在确定的离子强度和 pH 下,所选的严紧条件低于特定序列的热熔点(T_m)约 5 °C。 T_m 是(在确定的离子强度、pH、和核酸浓度的条件下)其中与靶序列互补的探针中 50%与靶序列平衡杂交的温度(因为靶序列一般过量存在,在 T_m 时,50%的探针被平衡占据)。通常,严紧条件包括盐浓度在 pH 7.0-8.3 时至少为约 0.01-1.0 M Na(或其它盐)离子浓度,而且对于短探针,温度至少为约 30 °C。严紧条件也可添加去稳定剂(如甲酰胺或四烷基季铵盐)来获得。

如所属领域技术人员将会理解的,可用最大严紧性的杂交来鉴定或检测相同的核苷酸序列,而可用中等(或低)严紧性的杂交来鉴定或检测相似或相关的多核苷酸序列。

将探针阵列与标记的或未标记的核苷酸序列杂交的方法也有描述。可控制特定杂交反应条件来改变杂交(如,增加或减少探针/靶结合严紧性)。例如,反应温度、阴离子和阳离子浓度、去污剂的加入等都可改变阵列探针和靶分子的杂交特征。

相互作用频率

对限制性片段连接频率定量可测量它们的交联频率。适宜的是,这可利用如 Splinter 等.(2004)(见上)所述的常规 3C 技术中使用的 PCR 来获得。简而言之,通过在溴化乙锭染色的琼脂糖凝胶上分离、然后用 Typhoon 9200 成像仪(Molecular Dynamics, Sunnyvale, CA)扫描信号强度来测量 PCR 产物的形成。适宜的是,也如 Splinter 等.(2004)(见上)所述,用几个对照来正确解释数据。

由于本文所述的 4C 技术提供了高通量分析核空间中两个或更多个核苷酸序列相互作用的频率的方法,因此优选利用本文所述的阵列来定量限制性片段的连接频率。

为了定量, 可以将 4C 样品中所得的信号相对于对照样品所得的信号来标准化。将 4C 样品和一个或多个对照样品用不同的且可分辨的标记物(如染料)标记, 并将同时与阵列杂交。一个或多个对照样品通常将包含等摩尔量的所有 DNA 片段(即与第一(靶)核苷酸序列连接了的所有潜在的第二核苷酸序列), 并且为了排除杂交效率的偏倚, 它们应与一个或多个第二核苷酸序列大小相似。因此, 对照模板通常将包含(与用于获得 4C 模板的基因组 DNA 具有相同遗传背景的)基因组 DNA, 其用第一和第二限制酶消化, 并用与 4C 模板相同的方法(如随机引发)来标记。这样的对照模板有可能校正杂交效率中探针-与-探针间的差异。将 4C 阵列信号相对于对照阵列信号标准化使得以富集方式而不是随机事件方式表述结果成为可能。

标记的 4C 模板甚至可以与带有或不带有不同标记的对照样品和带有或不带有一个或多个不同标记的其它 4C 模板的阵列杂交。其它 4C 模板可以与该 4C 模板无关, 例如它可以得自不同的组织和/或用不同组反向 PCR 引物来获得。例如, 第一 4C 模板可以是患者的材料, 而第二 4C 模板可来自健康的受试者或对照样品。

考虑到遗传重排所预计到的惊人的杂交模式, 不总是必须将患病的受试者与健康的受试者作比较。因此, 多个(如两个或更多个)4C 模板(其每一个都可研究来自相同患者或受试者的不同基因座)可与一个(如一个或多个)阵列杂交。

4C 模板可以是不同标记的(如用两种或多种颜色杂交), 和/或在该基因座正常情况下位于不同染色体上或在相同染色体上的距离足够远使得 DNA-DNA 相互作用信号间的重叠最小的情况下, 可以是相同标记的。例如, 可以处理来自患有 T 细胞白血病的受试者的材料, 以获得针对 *TCR α / δ* (其用一种颜色标记, 从而能检测易位)、和 *MLL*、*TAL1*、*HOX11* 和 *LMO2* (每个用相同的第二种颜色标记从而能检测其它遗传重排)的 4C 模板。这 5 种 4C 模板可与一种阵列杂交, 从而能在多个基因座处同时分析与疾病相关的基因组重排。

为了定量相互作用的频率, 也可考量相对于对照样品的绝对信号强度或比率。另外, 线性染色体模板上相邻的探针信号可用于鉴定相互作用的染色体区域。该位置信息优选如下分析, 通过在线性染色体模板上按序排列探针并通过滑动窗口的方法, 使用例如连续平均或连续中位数法, 来分析绝对信

号强度或相对于对照模板信号的比率。

依据本文描述的方法获得的一个或多个靶核苷酸序列与一个或多个目的核苷酸序列(如一个或多个基因组基因座)的相互作用频率可用以重建部分或整个线性染色体图并鉴定染色体内和染色体间发生的平衡的和不平衡的基因组重排,而这种重排指示性状或疾病。

测定方法

在本发明的另一方面,提供了鉴定一种或多种调节 DNA-DNA 相互作用的试剂的测定方法。

本文所用的术语“调节”指防止、减小、抑制、恢复、抬升、增加或以其它方式影响 DNA-DNA 相互作用。

在一些情况下,希望评估两种或多种试剂在一起用于调节 DNA-DNA 相互作用。在这些情况下,测定可通过在第一试剂同时或之后加入这样的一种或多种额外的试剂来方便地修改。

本发明的方法也可以是筛选方法,由此测试许多试剂能否调节 DNA-DNA 相互作用的活性。

预计本发明的测定方法适于小和大规模筛选试剂以及进行定量测定。

这些治疗剂的医学应用包括在本发明的范围内,药物开发程序本身和包含这些试剂的药物组合物也包括在其中。例如,药物开发程序可包括取得由或可由本文所述的方法鉴定的试剂,任选修饰它(如修饰它的结构和/或提供包含所述模块的新的组合物)并进行进一步研究(如毒性研究和/或对于活性、结构或功能的研究)。可在非人动物上进行试验并最终在人上进行。这些试验一般会包括确定不同剂量水平的一种或多种效果。药物开发程序可利用计算机来分析由筛选方法鉴定的模块(如预测结构和/或功能,鉴定可能的激动剂或拮抗剂,搜索可能具有相似结构或功能的其它模块等)。

诊断测试

当前,各种基因组重排仍难以通过可用的分子-细胞遗传技术来检测。尽管阵列比较基因组杂交技术(阵列-CGH)是新近开发的技术用以以 35-300 Kb 的解析度检测染色体扩增和/或缺失,但是该技术不适于检测平衡的易位和染色体倒位。在另一方面,光谱核型分析(Spectral karyotyping, SKY)或常

规核型分析经常在患者材料上进行来检测染色体易位以及数量改变,但是确定易位断点的解析度低,通常分别为 10-50 Mb 和 5-10 Mb。因此,这两种方法(尤其是 SKY)得到的结果将要导致进行费时、费力的验证实验,像荧光原位杂交(FISH)和分子断点克隆策略。

4C 技术包括基于物理连接的 DNA 序列间相互作用频率的改变可检测任何染色体重排的过程。所以,4C 技术可用于为大多数恶性肿瘤/多种先天畸形或智力迟钝鉴定(复发的)染色体重排。4C 技术的重要优势是它能非常准确地将断点作图在仅几千个碱基对的区域上。4C 技术的另一个优势是不需要事先知道确切的断点位置,这是因为即使 4C-诱饵序列距离断点 1-5 Mb 时也能检测到断点。这也具有相同的诱饵序列可用于检测覆盖大断点区域的特定染色体重排的优势。通过 4C 技术将基因组重排准确作图将大大方便鉴定涉及疾病或遗传病的一个或多个异常表达的基因,这对于更好地理解基因型-表型的相互关系将发挥重要贡献,帮助做出治疗决定,并增加重要的预测信息。

在本发明的一个实施方案中,为了提供诊断或预测疾病的基础,要建立受试者的正常或标准值。这通过测试取自正常受试者(如动物或人)的样品来获得。DNA-DNA 相互作用的频率可通过将它与一系列稀释度的阳性对照比较来定量。然后,得自正常样品的标准值可与得自受到或潜在受到疾病或病症影响的受试者样品的值来比较。标准和受试者值之间的偏差确立了疾病状态的存在。

可修改这些诊断测试来评估特定治疗方案的功效并可用于动物研究、临床试验、或用于监测单个患者的治疗。为了提供诊断疾病的基础,可确立 DNA-DNA 相互作用的正常或标准图谱。得自正常样品的标准值可与得自潜在受到疾病或病症影响的受试者样品的值来比较。标准和受试者值之间的偏差确立了疾病状态的存在。如果确定了疾病,则可给药现有的治疗剂,并可产生治疗图谱或值。最终,有规律地重复该方法来评估该值是否朝正常或标准模式进行或回归正常或标准模式。连续治疗图谱可用于显示几天或几个月的时间内治疗的功效。

4C 技术准确地检测了与要分析的核苷酸序列顺式连接的至少 5Mb 的基因组 DNA(参见图 2-3 和 5)。有益的是,4C 技术可用于检测任何伴随着重排序列和所选的 4C 序列(诱饵)之间基因组位点分离的改变的基因组异常。例

如,该改变可以是基因组位点分离的增加或减小,或可以是与 4C 序列(诱饵)邻近(如距离多至或大于 15 Mb)的序列的不充分代表(如在缺失中)或过度代表(如在复制中)。通常,该基因组异常或重排是疾病(如癌症(如白血病)和如本文所述的其它遗传或先天疾病)的成因或与之相关。

基因异常(如基因组或染色体异常——如平衡的和/或不平衡的基因组或染色体异常)包括但不限于核酸(如染色体)的重排、易位、倒位、插入、缺失和其它突变以及丢失或获得部分或完整的染色体。它们是遗传病症或疾病(包括先天疾病和获得性病症,如恶性肿瘤)的主要成因。在许多重排中,涉及 2 条不同的染色体。用这种方式,基因(或基因片段)从特定染色体的正常生理内容物中被去除,而且定位于受体染色体上,邻近不相关的基因或基因片段(通常是致癌基因或原癌基因)。

恶性肿瘤可包括急性白血病、恶性淋巴瘤和实体瘤。改变的非限制性实例有: $t(14;18)$, 其通常发生在 NHL 中; $t(12;21)$, 其通常在儿童期的前体-B-ALL 中找到; 和急性白血病中出现的 $11q23$ (MLL(髓细胞样-淋巴样白血病或混合谱系的白血病)基因)异常。

染色体区域 $11q23$ 中的 MLL 基因涉及 ALL 和急性髓细胞样白血病(AML)中的几种易位。至今,至少鉴定出了 10 种配体基因。这些易位中的一些(如 $t(4;11)(q21;q23)$ 、 $t(11;19)(q23;p13)$ 和 $t(1;11)(p32;q23)$)主要发生在 ALL 中;而其它的,像 $t(1;11)(q21;q23)$ 、 $t(2;11)(p21;q23)$ 、 $t(6;11)(q27;q23)$ 和 $t(9;11)(p22;q23)$, 则更常在 AML 中被观察到。涉及 $11q23$ 区域的重排非常频繁地发生在婴儿急性白血病中(约 60-70%), 并发生在儿童和成人白血病中的程度低得多(分别约为 5%)。

淋巴细胞恶性肿瘤中的重排通常涉及 Ig 或 TCR 基因。实例包括三类在 Burkitt 氏淋巴瘤中发现的易位($t(8;14)$, $t(2;8)$, 和 $t(8;22)$), 其中 MYC 基因分别与 Ig 重链(IGH)、Ig κ (IGK)、或 Ig λ (IGL)基因片段偶联。此类中的另一种常见类型的易位是 $t(14;18)(q32;q21)$, 其在约 90%的滤泡状淋巴瘤(其为主要 NHL 类型之一)中被观察到。在该易位中,将 BCL2 基因重排至 JH 基因片段内或与之相邻的 IGH 基因座内的区域。该染色体异常的结果是过量表达 BCL2 蛋白质,其在生长控制中通过抑制程序性细胞死亡来起存活因子的作用。

BCL2 基因由三个外显子组成,但这些分散在大区域中。其中最后一个

外显子编码大的 3'非翻译区域(3' UTR)。该 3' UTR 是 2 个其中有许多 t(14;18) 断点成群聚集的区域之一, 并被称为“主要断点区”; 另一个涉及 t(14; 18) 易位的断点区域位于 BCL2 基因座下游 20-30 kb 处, 并被称为“次要成群区”。第三个 BCL2 断点区——VCR(变体成群区)位于 BCL2 基因座的 5' 侧, 并处于其它涉及变体易位的区域中, 这些其它涉及变体易位的区域即 t(2;18)和 t(18;22), IGK 和 IGL 基因片段是其配偶体基因。

因此, 举例来说, 4C 技术可用于筛选针对在基因座中或附近的遗传异常的患者材料, 这些基因座基于它们与给定的临床表型的频繁相关性而选择。这些基因座的其它非限制性的例子有 *AML1*、*MLL*、*MYC*、*BCL*、*BCR*、*ABL1*、免疫球蛋白基因座、*LYL1*、*TAL1*、*TAL2*、*LMO2*、*TCR α/δ* 、*TCR β* 、*HOX* 和各种淋巴母细胞白血病中的其它基因座。

有益的是, 如果疑似有遗传异常, 则 4C 技术可用作最初和仅有的筛选方法来验证本文所解释的异常的存在并对其作图。

检测基因组重排

在本发明特别优选的实施方案中, 本文所述的方法可用于检测基因组重排。

当前, 基因组重排(如易位断点)非常难以检测。例如, 比较基因组杂交 (CGH)微阵列能检测几类重排, 但不能检测易位, 如果怀疑患者中有易位但不知道染色体配体, 则可进行光谱核型分析(SKY)来找出易位配体并粗略估计断点位置。可是, 解析度非常低(通常不超过~50 Mb), 并且通常需要额外的精细作图(其是费时和昂贵的)。这通常利用荧光原位杂交(FISH)来进行, 其提供的解析度也是有限的。利用 FISH, 断点可以最大的解析度定位于+/-50 kb 的区域中。

DNA-DNA 相互作用的频率主要是基因组位点分隔距离(separation)的函数, 即 DNA-DNA 相互作用的频率与出现在相同物理 DNA 模板上的 2 个 DNA 基因座之间的线性距离(以千碱基对计)成反比(Dekker 等., 2002)。因此, 产生一种或多种新物理 DNA 模板的易位伴随着断点附近的 DNA-DNA 相互作用的变化, 而且这可通过 4C 技术来测量。基于易位的疾病通常由异常的 DNA-DNA 相互作用造成, 这是因为易位是断的染色体(DNA)臂物理连接(相互作用)的结果。

因此, 为了检测易位。4C 技术可用于鉴定那些在患病和未患病受试者间不同的 DNA-DNA 相互作用。

举例来说, 4C 技术可用于针对基因座附近的易位筛选患者材料, 所述基因座基于它们与如本文所述的给定的临床表型的频繁相关性来选择。

如果怀疑患者中有易位但不知道染色体配体, 则可利用当前可用的方法, 像光谱核型分析(SKY), 来进行初始作图。这可鉴定易位配体并非常粗略地评估断点位置(通常不优于 ~50 Mb 的解析度)。然后可用 4C 技术, 利用该区域中位于例如每 2 Mb、5Mb、10Mb、20Mb(或如本文所述的其它间隔)处的‘诱饵’-序列来对断点精细作图, 并鉴定诸如由于易位而错误表达的一个或多个基因。

通常, 易位鉴定的方法是通过在含 4C-诱饵序列的染色体之外的染色体上或在同一染色体上其它地方的相互作用频率从低到高的突然变化来进行。

在优选的实施方案中, 受试者的样品处于恶变前的状态。

在优选的实施方案中, 受试者的样品由产前诊断所用的羊膜穿刺术获得的培养的或未培养的羊水细胞组成。

在优选的阵列设计中, 单个阵列上出现的探针以最大解析度代表给定物种的完整基因组。因此, 通过 4C 技术检测易位等的阵列包含如本文所述的与给定物种(如人)的基因组中每个第一限制酶识别位点之每一侧互补的探针。

在另一个优选的设计中, 单个阵列上出现的探针代表给定物种的完整基因组, 但解析度不是最大的。因此, 通过 4C 技术检测易位等的阵列包含如本文所述的仅与给定物种(如人)的基因组中每个第一限制酶识别位点的一侧互补的探针。

在另一个优选的设计中, 单个阵列上出现的探针代表给定物种的完整基因组, 但解析度不是最大的。因此, 通过 4C 技术检测易位、缺失、倒位、重复和其它基因组重排的阵列包含如本文所述的探针, 其与沿给定物种(如人)基因组线性模板排列的每隔一个第一限制酶识别位点的一边互补。

因此, 通过 4C 技术检测易位、缺失、倒位、重复和其它基因组重排的阵列包含如本文所述的探针, 其每一个代表用第一限制酶消化后得到的单一限制性片段。这优选通过在每隔二个、三个、四个、五个、六个、七个、八个、九个、十个、二十个、三十个、四十个、五十个、六十个、七十个、八

十个、九十个或一百个等与相同限制性片段杂交的探针中忽略一个来实现。通过 4C 技术检测易位、缺失、倒位、重复和其它基因组重排的阵列可包含如本文所述的沿着线性染色体模板平均分布的探针。这优选通过忽略那些显示出最高探针密度的基因组区中的一个或多个探针来获得。

在另一个优选的设计中，单个阵列上出现的探针代表给定物种的完整基因组，但不是以最大解析度。因此，通过 4C 技术检测易位、缺失、倒位、重复和其它基因组重排的阵列包含如本文所述的探针，其与沿着给定物种(如人)基因组线性模板按序排列的每隔三个、四个、五个、六个、七个、八个、九个、十个、二十个、三十个、四十个、五十个、六十个、七十个、八十个、九十个或一百个等第一限制酶识别位点中的一个的一边互补。通过 4C 技术检测易位、缺失、倒位、重复和其它基因组重排的阵列可包含如本文所述的代表完整的基因组的探针，每十万个碱基一个探针。通过 4C 技术检测易位、缺失、倒位、重复和其它基因组重排的阵列可包含如本文所述的探针，其代表基因组中可由独特的探针序列代表的每个单个第一限制酶识别位点。

在另一种优选的阵列设计中，如本文所述的在单一阵列上的探针代表已知涉及易位、缺失、倒位、重复和其它基因组重排的所有基因座周围的给定大小——如约 50 kb、100 kb、200 kb、300 kb、400 kb、500 kb、1 Mb、2 Mb、3Mb、4Mb、5Mb、6Mb、7Mb、8Mb、9Mb 或 10Mb(如约 50kb-10Mb)——的基因组区域。

在另一种优选的阵列设计中，如本文所述的在单一阵列上的探针代表已知涉及易位、缺失、倒位、重复和其它基因组重排的所选基因座周围的给定大小——如约 50 kb、100 kb、200 kb、300 kb、400 kb、500 kb、1 Mb、2 Mb、3Mb、4Mb、5Mb、6Mb、7Mb、8Mb、9Mb 或 10Mb(如约 50kb-10Mb)——的基因组区域。选择可以按教导的标准来进行，例如它们可仅代表给定的疾病类型中包含的基因座。

在另一种优选的阵列设计中，如本文所述的在单一阵列上的探针代表(部分)染色体或多个染色体的例如 100 kb、200 kb、300 kb、400 kb、500 kb、600 kb、700 kb、800 kb、900 kb、1 Mb、2 Mb、3 Mb、4 Mb、5 Mb、6 Mb、7 Mb、8 Mb、9 Mb、10 Mb、20 Mb、30 Mb、40 Mb、50 Mb、60 Mb、70 Mb、80 Mb、90 Mb，或 100 Mb(如 100kb-10Mb)感兴趣的基因组区域，其中每个探针被代表多次(如 10、100、1000 次)从而可以定量测量每个探针序列处的

杂交信号强度。

在优选的实验性设计中, 4C 序列(诱饵)处于距离实际重排序列(即易位情况下的断点)约 0kb、10kb、20kb、30kb、40kb、50kb、100kb、200kb、300kb、400 kb、500 kb、1 Mb、2 Mb、3Mb、4Mb、5Mb、6Mb、7Mb、8Mb、9Mb、10Mb、11Mb、12Mb、13Mb、14Mb 或 15Mb(如约 0-15Mb)或更远范围之内。

在优选的杂交中, 用来自患病和非患病的受试者的 1 种序列(4C 诱饵)得到的 2 种区别标记的 4C 模板与相同的阵列同时杂交。DNA-DNA 相互作用中的差异能检测顺式(与 4C-诱饵在相同染色体上)和反式(在易位配体上)的断点。

在优选的杂交中, 用来自患病和非患病的受试者的 1 种序列(4C 诱饵)得到的多种区别标记的 4C 模板与相同的阵列同时杂交。DNA-DNA 相互作用中的差异能检测顺式(与 4C-诱饵在相同染色体上)和反式(在易位配体上)的断点。

有益的是, 可用微阵列上的多颜色分析代替双颜色分析来使超过 2 个样品同时与单个阵列杂交。因此, 可在 4C 技术中使用多颜色杂交。

在优选的杂交中, 用来自患病的受试者的 1 种序列(4C 诱饵)得到的多种区别标记的 4C 模板和来自非患病的受试者的 1 种区别标记的 4C 模板与相同的阵列同时杂交。DNA-DNA 相互作用中的差异能检测顺式(与 4C-诱饵在相同染色体上)和反式(在易位配体上)的断点。

在另一种优选的杂交中, 用分别代表另一种可能的易位配体的 2 种不同序列(4C-诱饵)获得的、来自同一非患病的受试者的 2 种区别标记的 4C 模板与相同的阵列同时杂交。在与带有目的序列(4C-诱饵)的染色体无关的染色体线性模板上观察到的强杂交信号簇将鉴定出易位配体染色体和在易位配体上的断点。

在另一种优选的杂交中, 用分别代表另一种可能的易位配体的多种不同序列(4C-诱饵)获得的、来自相同非患病受试者的多种区别标记的 4C 模板与相同的阵列同时杂交。在与带有目的序列(4C-诱饵)的染色体无关的染色体线性模板上观察到的强杂交信号簇将鉴定出易位配体染色体和其针对目的序列的断点。

用于通过 4C 技术检测易位、缺失、倒位、重复和其它基因组重排的材料可通过交联(并如所述的, 进一步处理)来自患病和/或非患病的受试者的活

细胞和/或死细胞和/或核溶解产物和/或分离(如本文所述)的染色质等来获得。

检测倒位

倒位(如平衡的倒位)不能通过诸如比较基因组杂交技术的方法来检测,但能通过 4C 技术来检测,尤其在(平衡的)倒位接近 4C 序列(诱饵)(如多至约 1-15 Mb 或更多)的时候能检测。

(平衡的)倒位的检测基于鉴定那些在患病和非患病的受试者之间有差异的 DNA-DNA 相互作用。倒位将改变重排区域所有(但除外位于最中心的)序列在物理 DNA 模板上相对于当作 4C 序列(诱饵)的在相同染色体上的附近序列的相对位置(以千碱基计)。由于 DNA-DNA 相互作用频率与基因组位点分隔距离成反比,因此患病的受试者与非患病的受试者相比将给出对于所有位于重排基因组区域中的探针呈倒转模式的杂交强度。因此,4C 技术能鉴定(平衡的)倒位的位置和大小。

根据本发明的这一方面,优选的专门阵列设计包括在单一阵列上的探针,其代表怀疑有倒位或其它重排的基因座周围给定大小——如约 50 kb、100 kb、200 kb、300 kb、400 kb、500 kb、1 Mb、2 Mb、3Mb、4Mb、5Mb、6Mb、6Mb、7Mb、8Mb、9Mb 或 10Mb)(如 50kb-10Mb)——的基因组区域。

在另一种优选的专门阵列设计中,在单一阵列上的探针代表怀疑有倒位或其它重排的基因座周围的给定大小(50 kb、100 kb、200 kb、300 kb、400 kb、500 kb、1 Mb、2 Mb 等)的基因组区域。为了可靠地定量分析信号强度,存在于阵列上的探针量通常大大过量于与阵列杂交的关联片段(cognate fragment)的量。所以,可能必须使每种探针在阵列上出现多次(如 10、20、50、100、1000 次等)。另外,可能必须滴定测量与阵列杂交的模板量。

检测缺失

缺失的检测基于鉴定那些在患病和非患病的受试者之间有差异的 DNA-DNA 相互作用。缺失将造成与位于缺失区域附近(如约 1、2、3、4、5、6、7、8、9、10、11、12、13、14 或 15 Mb 或更多)的 4C 序列(诱饵)的 DNA 相互作用的缺乏。如果缺失存在于两个等位基因上(纯合),则可造成完全缺乏针对位于重排区域中所有探针的杂交信号,或者如果缺失存在于仅一个等

位基因上(杂合),可造成与非患病的受试者相比患病的受试者信号强度减少。缺失使物理 DNA 模板上更远处的序列与要分析的 4C 序列(诱饵)更接近,其将导致针对直接位于缺失区旁边的探针更强的杂交信号。

检测重复

重复的检测通常基于鉴定那些在患病和非患病的受试者之间有差异的 DNA-DNA 相互作用。与来自对照的非患病的受试者的信号相比,在重复区域中的探针将显示出与位于重排区域附近(如约 1、2、3、4、5、6、7、8、9、10、11、12、13、14 或 15 Mb 或更多)的 4C 序列(诱饵)的杂交信号增强。重复区旁边的探针更远离于 4C 序列,因此将比来自对照的非患病的受试者的信号显示出降低的杂交信号。

受试者样品与对照相比 DNA-DNA 相互作用频率的增加或降低优选指示重复或插入。

受试者样品与对照相比 DNA-DNA 相互作用频率的增加和/或针对更远距离区域的 DNA-DNA 相互作用频率的降低优选指示重复或插入。

产前诊断

有益的是,4C 技术也可用于产前诊断。

利用各种本领域已知的方法可从胎儿获得核酸。举例来说,可用羊膜穿刺来获得羊水,由其提取胎儿细胞悬浮液并培养几天(Mercier & Bresson(1995)*Ann. Gnt.*, 38, 151-157)。然后从细胞中提取核酸。收集绒毛膜绒毛有可能省却培养步骤并避免收集羊水。这些技术可更早应用(对于收集绒毛膜绒毛,在妊娠的多至 7 周时间;而对于羊膜穿刺,在 13-14 周),但微微增加流产的风险。

直接在脐带水平收集胎儿血也可用于获得核酸,但这通常需要在该技术上专业化的临床医生团队(Donner 等.(1996)*Fetal Diagn. Ther.*, 10, 192-199)。

有益的是,遗传异常(如基因组或染色体异常)——如染色体和核酸中的重排、易位、倒位、插入、缺失和其它突变——可在该阶段被检测出。

优选可检测遗传异常(如基因组或染色体异常)——如染色体 21、18、13、X 或 Y 中的重排、易位、倒位、插入、缺失和其它突变以及丧失或获得部分或完整染色体 21、18、13、X 或 Y,这是因为这些染色体是大多数胎儿异

常发生在其中的染色体。

确定基因组整合位点

当多拷贝被插入在基因组不同位置时, 4C 技术也能确定病毒和转基因等的基因组整合位点(如图 3 所述)。

确定获得某种易位的倾向

有益的是, 4C 技术也可用于非患病的受试者来测量频繁涉及遗传异常的基因座的基因组环境。以这种方式, 有可能确定受试者获得某种遗传异常的倾向。

因此, 除了本文所述的医学应用, 本发明可用于诊断。

多重 4C

本发明允许同时分析多个靶核苷酸序列与一个或多个目的核苷酸序列的相互作用频率。扩增可用多重 PCR 获得。此方法允许对平衡的和不平衡的基因组重排(即基因组任意位置已发生的易位、倒位、缺失和重复)的无偏筛选。本方法可用于以非常高解析度鉴定重排的断点, 特别是在 2 万碱基(平均 5kb)之内。本方法可用于上述的诊断应用, 例如产前诊断、产后诊断和分析肿瘤与其它患病样品以鉴定涉及疾病或预测受试者疾病的基因组重排。扩增的目的核苷酸序列能够在修改的 4C 微阵列(如上所述)或基因组嵌合微阵列上, 或通过本文进一步描述的测序进行分析。

已知与疾病相关的多个基因座处重排的同时筛选

对于许多疾病、症状或表型, 多个可能成为原因的 DNA 改变是已知的, 需要筛选受试者以鉴定涉及疾病/症状/表型的确切重排。例如, 在涉及 AML1、MLL、MYC、BCL、BCR、ABL、免疫球蛋白基因座、LYL1、TAL1、TAL2、TCR α/δ 、TCR β 、HOX 和其它可能的经常涉及该疾病的基因座的成淋巴细胞白血病重排的情况下, 4C 技术可用以鉴定哪个基因座和重排与患者给定的疾病相关。

在这个实施方案中, 4C 技术定向于每个目的基因座。每个基因座能够单独分析, 但是多个基因座也能够在单个阵列上同时分析。因此, 4C 筛选

可以包括(反向)PCR扩增与一个或多个靶核苷酸序列相互作用的DNA元件,所述序列靠近需要分析重排的每个基因座。处在或靠近这些基因座的靶序列基于本文提到的准则选择。

目的核苷酸序列的扩增能够针对每个靶核苷酸序列单独地进行,或者能够通过多重(反向)PCR在一个反应体积内同时进行。当与不同靶核苷酸序列相互作用的目的核苷酸序列能够相同地标记而不影响分析时,后一种方法在这种情况下优选。这可以是以下这样的情况,例如当基因座处在不同染色体上或当基因座处在同一染色体上足够远的距离从而使DNA-DNA相互作用信号的重叠最小时,或当从不同靶核苷酸序列扩增的DNA-DNA相互作用的重叠不干扰基因组重排的检测时。

当每个DNA-DNA相互作用组需要不同标记或当反向PCR引物组干扰彼此的扩增效率时,与不同靶核苷酸序列相互作用的目的核苷酸序列的扩增优选单独进行。

如上所述,将和不同靶核苷酸序列相互作用的不同或相同标记的目的序列与包含代表完整基因组(如高密度阵列或嵌合阵列)或选择的部分基因组的探针的单个或多个微阵列杂交。杂交信号将与那些得自对照样品的信号进行比较,其中测得的受试对对照样品的DNA-DNA相互作用频率的增加或降低指示受试样品中DNA的重排。

对于在基因组未知位置的重排的无偏基因组广度筛选

在第二个实施方案中,4C技术用于无偏基因组广度的筛选以鉴定来自患病的(或非患病的)受试者样品的重排,其中所述重排是以前未知的和/或发生在未知的位置。在这个实施方案中,不能选择重排附近的靶序列。因此靶序列未知或未被怀疑与疾病相关。相反,选择它们广泛分布在整个所选的基因组或基因组部分因而提供对要研究的染色质的充分覆盖。优选整个基因组被覆盖。

例如,选择靶序列从而使与它们相互作用的目的序列(大部分是直接围绕靶序列的序列,即在距靶序列5、10、15、20、25、30、35、40或45到50Mb之内)覆盖完整的或实质性部分的基因组或染色体或部分染色体。这将允许重建存在于任何受试者或细胞类型的物理染色体图。

为了使线性染色体模板上相邻的两个靶序列具有重叠的相互作用的

DNA 序列(即重叠的基因组环境), 它们需要彼此相距例如 5、10、15、20、25、30、35、40、45 或 50 到 100Mb。因此, 为了以分隔~10Mb 的靶序列覆盖完整的基因组(2-3 x 10⁹bp), 需要 200-300 个靶序列。为了以分隔~50 Mb 的靶序列覆盖完整的基因组, 需要 40-60 个靶序列。在最终的实施方案中, 每个染色体仅需要一个靶序列来覆盖完整的基因组。

在 4C 中与靶序列相互作用的所有目的核苷酸序列都被扩增, 或者通过多重(反向)PCR 在单个反应混合物(或有限数量的反应混合物)中共同进行, 或者在可以随后汇集的单个(反向)PCR 反应中进行。当与不同靶核苷酸序列相互作用的目的核苷酸序列能够在不影响分析的情况下相同地标记时, 优选多重 PCR。当属于给定靶序列的每组 DNA-DNA 相互作用需要不同标记和/或当反向 PCR 引物组干扰彼此的扩增效率时, 每个靶核苷酸序列的扩增将单独进行。

染色体内重排如缺失、插入和重复和(平衡的和不平衡的)倒位的鉴定

在一个优选的实施方案中, 将与来自受试样品(如患者样品)的靶核苷酸组相互作用的所有扩增的 DNA 序列相同地标记, 且将基因组广度的相互作用频率与对照样品(如来自健康的受试者)的这些相互作用频率比较。对照和受试样品可用不同的颜色与同一阵列杂交, 或者它们能够与不同阵列杂交和比较。受试样品与对照样品相比 DNA-DNA 相互作用频率的增加或降低指示受试样品中的重复/插入或缺失。它也可以指示倒位。

图 18 和 19 显示了以这种方式通过 4C 鉴定的缺失。

在另一个优选的实施方案中, 将与来自受试样品(如患者样品)的靶序列组相互作用的扩增的 DNA 序列以两种颜色标记, 在线性染色体模板上相邻的靶序列的颜色相互交替。相邻的靶序列在线性染色体模板上足够靠近从而使它们的相互作用序列重叠。因此相邻的靶序列可以是彼此相距, 例如 5、10、15、20、25、30、35、40、45 到 50Mb。这样的实验设计防止由于饱和的杂交信号靠近给定的靶序列而错过重排。将基因组广度的相互作用频率与对照样品(如来自健康受试者)的这些相互作用频率比较。对照和受试样品可以不同地标记并与同一阵列杂交, 或者它们能够与不同阵列杂交并比较。受试样品与对照样品相比 DNA-DNA 相互作用频率的增加或降低指示受试样品中的重复/插入或缺失。它也可指示倒位。

随后可进行针对位于基因组重排部分侧翼或内部的靶序列的 4C 分析以鉴定倒位。与对照样品相比杂交信号的倒转模式鉴定受试样品中的倒位。这在图 20 中描述。

在另一个优选的实施方案中,与染色体模板上并列的不同靶序列相互作用的目的核苷酸序列用不同的染料标记。通过患者样品与对照样品相比相互作用的 DNA 信号的出现或消失来检测重排。这在图 21 到 23 中描述。

平衡的和不平衡的易位的鉴定

在另一个优选的实施方案中,多种染料可用(如 48 种染料)而每个染色体用两种独特的染料标记从而使线性染色体模板上的相邻靶序列的颜色交替。所有的 DNA 片段能够一起与包含代表完整基因组的探针的阵列杂交。发生在受试样品而非对照样品的染色体间 DNA-DNA 相互作用的鉴定指示易位并鉴定两个重排的染色体。不相关的染色体上从低到高信号的转变鉴定 DNA 断点附近的第一限制酶识别位点。

如果易位是平衡的,两个染色体特异的染料每个都应当在不相关染色体的线性模板上彼此直接相邻的互斥探针组上给出强烈的杂交信号。这一不相关染色体上的断点位于显示正杂交信号的两组探针之间。参见图 24。

当测试包含染色体 1: 染色体 7 易位的患者样品时,得到图 25 所示的结果。在这个特定的实例中阵列上的靶序列代表对位于 *HindIII* 位点旁序列的选择(参见上面,靶序列的衍生)。

如果易位的发生伴随着断点上 DNA 的丢失(即不平衡的易位),两种染色体特异的染料中的每一个将在不相关染色体的线性模板上彼此不直接相邻的互斥探针组上给出强烈的杂交信号。位于不相关的染色体上,显示染色体间 DNA-DNA 相互作用的两组探针之间的探针代表丢失的基因组区域。这一区域最外侧的探针标示不相关染色体上断点附近的限制性位点(参见图 26)。

当在断点处用未知量的缺失 DNA 测试包含染色体 4:染色体 7 易位的患者样品时,得到图 27 所示的结果。在这个特定的样品中,阵列上的靶序列同样代表位于 *HindIII* 位点旁的序列的选择。结果显示易位和缺失(约 2Mb)都检测到了。

在另一个优选的实施方案中,多种染料(如 24 种染料)可用并用以不同地

标记每条染色体(图 28)。因此,与同一染色体上出现的靶序列相互作用的所有 DNA 序列都相同地标记并与那些位于其它染色体的不同。发生在受试样品而不是对照样品中的染色体间 DNA-DNA 相互作用的鉴定指示易位并鉴定两个重组的染色体。

随后特异性针对涉及易位的染色体的 4C 分析允许鉴定断点。这里,一个或每个染色体用两种独特的染料标记,从而使线性染色体模板上相邻的与靶序列相互作用的 DNA 片段之间的染料相互交替。

如果易位是平衡的,两个染色体特异的染料每个都应当在不相关染色体的线性模板上彼此直接相邻的互斥探针组上给出强烈的杂交信号。这一不相关染色体上的断点位于显示正杂交信号的两组探针之间。

如果易位的发生伴随着断点上 DNA 的丢失(即不平衡的易位),两种染色体特异的染料中的每一个将在不相关染色体上的线性模板上彼此不直接相邻的不相关染色体上的互斥探针组上给出强烈的杂交信号。位于不相关的染色体上,显示染色体间 DNA-DNA 相互作用的两组探针之间的探针代表丢失的基因组区域。这一区域最外侧的探针标示不相关的染色体上断点附近的限制性位点。

在另一个优选的实施方案中,少于 24 种染料可用且每种染料独特地用于标记与同一染色体上存在的靶序列相互作用的所有 DNA 片段。因此,用 1、2、3、4、6、8、10、12 等种染料可以独特地标记 1、2、3、4、6、8、10、12 等个染色体。属于这些染色体中每一个的标记的 DNA 片段能够共同与包含代表完整基因组的探针的阵列杂交。不相关的染色体上代表 DNA-DNA 相互作用的杂交信号的出现鉴定此染色体为易位配体。随后特异性针对涉及易位的染色体的 4C 分析允许鉴定断点(如上所述)。

在另一个优选的实施方案中,少于 48 种染料可用且每个染色体用两种独特的染料标记,从而使与线性染色体模板上相邻的靶序列相互作用的 DNA 片段之间的染料相交替。因此,用 2、4、6、8、10、12、14、16、18、20、22、24 等种染料可以独特地标记 1、2、3、4、5、6、7、8、9、10、11、12 等个染色体。所有的 DNA 片段能够共同与包含代表完整基因组的探针的阵列杂交。不相关染色体上的杂交信号代表染色体间的 DNA-DNA 相互作用并鉴定此染色体为易位配体。不相关的染色体上从低到高信号的转变鉴定 DNA 断点附近的第一限制酶识别位点。

在另一个优选的实施方案中,少于48种染料可用且每个染色体用多于两种独特的染料标记,从而使与线性染色体模板上相邻的靶序列相互作用的DNA片段之间的染料相交替。所有的DNA片段能够共同与包含代表完整基因组的探针的阵列杂交。不相关染色体上的杂交信号代表染色体间的DNA-DNA相互作用并鉴定此染色体为易位配体。不相关的染色体上从低到高信号的转变鉴定DNA断点附近的第一限制酶识别位点。

如何鉴定其它易位断点

依据上面提到的任一策略在一个染色体上鉴定的断点可以随后用于鉴定易位配体染色体上的断点。这可以例如通过如下方法实现:使用对一个染色体特异的、读入其它染色体序列的引物对经由如下PCR得到的产物进行测序,跨越DNA接头的长程PCR或在经限制酶消化和(未交联的)基因组DNA的再连接而形成的DNA环上进行反向PCR或连接介导的(LM-)PCR。

在一个优选的实施方案中,每个前述的对基因组重排的筛选都随后进行专用的针对候选断点附近靶序列的4C实验以明确地鉴定它们。

上述检测不同种类基因组重排的方法的方面可以与同时在基因组筛选它们的发生组合。

如果使用基因组嵌合阵列而不是4C阵列,在患者样品中观察到,但没有在对照样品中观察到的线性染色体模板上并列的探针信号强度的转变指示与基因组重排有关的断点的位置(而不是DNA断点附近的第一限制酶的识别位点)。

测序

高通量DNA测序一定会成为可负担的和更加量化的微阵列的替代物以分析大的DNA序列集合。高通量测序方法的实例在*E.Y. Chan, Mutation Research 573(2005)13-40*中进行了列举,其包括但不限于短期(near-term)测序方法例如环延伸方法、聚合酶阅读方法和外切核酶测序、革命性的测序方法例如DNA扫描和毫微孔测序(nanopore sequencing)和直接线性分析。现有高通量测序方法的实例是454(热)测序(pyrosequencing)、Solexa基因组分析系统、Agencourt SOLiD测序方法(Applied Biosystems)、MS-PET测序(Ng等, 2006, <http://nar.oxfordjournals.org/cgi/content/full/34/12/e84>)。

测序在对 4C 和基于检测基因组相互作用的其它方法的结果进行高通量分析方面能够代替阵列杂交。序列发生的频率指示基因组中结合的频率，并能够如前述分析杂交结果一样进行分析。

如权利要求 1 所述、步骤 a-g 提供的那样，对模板进行测序。或者，测序可以对由 Lomvardas 等, Cell 126, 403-413, July 28, 2006 或由 Ling 等, Science 312, 14 April 2006, 269-272 所描述的方法获得的 PCR 产物进行。

测序起始于 PCR 产物的一个或两个末端。PCR 产物的两端由已知核苷酸组成的核苷酸序列组成，其至少一端是靶核苷酸序列，并位于与靶核苷酸序列相互作用并连接到靶核苷酸序列的目的核苷酸序列的侧翼。根据所用的测序方法，衔接子可能需要附加到 PCR 产物的一端或两端。衔接子可以是目的测序方法所需的寡核苷酸序列，其可能或可能不包含例如允许它们被捕获的模块。衔接子可直接地或在平端化 PCR 产物的末端之后连接到 PCR 产物。或者，如步骤 g(权利要求 1)所使用的那样，PCR 引物组可以包含代表衔接序列的突出端或可以含有引入限制酶切割位点的突出端，所述切割位点可用于随后连接对 PCR 产物的每个末端特异性或非特异性的衔接子。

在旨在鉴定与单个核苷酸靶序列相互作用的 DNA 序列的 4C 实验中，测序需要读过第一连接事件(步骤 c)和/或第二连接事件(步骤 f)从而获得足够的序列信息以鉴定目的核苷酸序列。通常情况下，这需要测序并鉴定连接接头旁边、进入目的核苷酸序列之内的最少 8-30 个核苷酸的序列段(stretch)(参见图 28)。

在多重 4C 中，测序需要读过第一连接事件(步骤 c)和/或第二连接事件(步骤 f)从而获得足够的序列信息以鉴定共同形成连接产物的靶核苷酸序列和目的核苷酸序列。通常情况下，这需要在连接接头每侧测序和鉴定 8-30 个核苷酸的序列段。靶核苷酸序列的鉴定将为每个连接产物提供“家庭地址”。在步骤 g 使用的与靶核苷酸序列杂交的引物需要位于与第一和第二连接接头足够远从而明确地鉴定此“家庭地址”的距离。根据测序方法，此距离最小可以距离连接接头 0、10、20、30、40、50、60、70、80、90 或 100 个核苷酸。

在单个和多重 4C 中，当 PCR 产物的组成为靶核苷酸序列每侧侧翼的目的核苷酸序列时，从 PCR 产物的一端读过第一连接接头，和从 PCR 产物的另一端读过第二连接接头提供同样的信息。因此，从 PCR 产物的一侧获得的序列足以分析 DNA-DNA 相互作用。从 PCR 产物的另一侧获得的序列可

用于补充或确认数据。参见图 29。

实际测序反应的起始可以用相应的高通量筛选平台常规使用的标准引物完成。它也可以由与用以扩增被给定靶序列捕获的序列的反向 PCR 引物部分或完全重叠的定制测序引物完成。这将防止重复阅读整个反向 PCR 引物序列并允许阅读捕获片段的更多核苷酸，其因此可以更易被鉴定和绘制至基因组。其后这些定制测序引物也可以部分退火到通常用作测序引物互补 DNA 链的衔接序列。

在多重设置中，可以同时使用与衔接序列和反向PCR扩增的DNA产物的外端重叠的多个测序引物，从而使分析中包含的每个靶片段都具有它自己的独特测序引物。与反向PCR扩增的DNA产物(即反向PCR引物序列)的外端重叠应当使靶序列仍可为测序提供足够的核苷酸并使每个靶序列(家庭地址)都能明确地鉴定。根据同时分析的靶序列的性质和数量，这意味着在一个实施方案中，测序引物可以与形成靶序列与捕获序列之间接头的限制性位点之外的1到20个核苷酸杂交。也可以用双标记测序来完成每个靶序列和捕获序列的组合的鉴定，其提供每个所分析DNA片段两端的序列信息。

在另一个实施方案中，为每个靶序列设计测序引物从而使它们靠近所分析的第一和/或第二限制酶识别位点，并可以部分或完全地与所分析的第一和/或第二限制酶识别位点重叠。

在一个实施方案中，使用的高通量测序技术是Solexa (Illumina)测序。

在一个实施方案中，测序可定向于第一限制酶识别位点的侧面。这可以防止读取在第二连接步骤过程中发生的随机连接事件。

在一个实施方案中，测序可定向于第二限制酶识别位点的侧面。

数据分析

由4C技术分析的连接产物由“家庭地址”(靶核苷酸序列)和相互作用的目的核苷酸序列(“旅行地址”)组成。在单个4C情况下此“家庭地址”已知(为所分析的一个靶核苷酸序列)。

在多重4C的情况下，“家庭地址”是通过用电子方法将获得的靶核苷酸序列与含有所有该分析包含的靶核苷酸序列的文库/数据库进行比较而鉴定的。获得的“旅行地址”(即目的核苷酸序列)是通过用电子方法将其序列与含有位于所选第一与第二限制酶识别位点之间的所有基因组DNA片段的文库/数据

库进行比较而鉴定的。

为了使数据可视化并有助于分析,在优选的实施方案中沿着染色体的线性模板将每个测序的连接产物绘制在“旅行地址”(目的核苷酸序列)的基因组位置上。独特的颜色代码揭示每个连接产物的“家庭地址”,并且每个连接产物的检测频率也在图上显示出来。基因组重排可以通过用一个样品(如得自患者的)中测得的每个相互作用频率与其它样品(如来自健康受试者)中测得的那些相互作用频率相除进行检测。这些数值也可以沿着线性的染色体模板绘制。

受试样品中DNA-DNA相互作用频率的降低指示缺失。通常这与从靶序列测得的、在最远断点之外的序列的DNA-DNA相互作用频率的增加同时发生。受试样品中DNA-DNA相互作用频率的增加指示重复。通常这与从靶序列测得的、在最远断点之外的序列的DNA-DNA相互作用频率的降低同时发生。受试样品中DNA-DNA相互作用频率的倒转指示基因组倒位。跨越染色体的DNA-DNA相互作用的检测指示易位。如微阵列分析中所述那样检测断点。

生物标记

鉴定与疾病相关的重排-例如易位、倒位和缺失允许能够用于诊断疾病的生物标记的鉴定。例如,可以设计检测给定的重排并用于诊断患者疾病的杂交探针或PCR引物。可以根据本领域公知的技术设计PCR探针,从而使怀疑在疾病状态下重排的区域用引物得到扩增;扩增产物的性质将指示疾病的有无。另外,可以设计在有无重排时将排他地杂交的杂交探针或引物。源自重排的融合蛋白可通过如下技术检测,例如使用根据本领域公知的技术设计的抗体的抗体检测或质谱。

受试者

术语“受试者”包括哺乳动物——如动物和人。

试剂

试剂可以是有机化合物或其它化学物质。试剂可以是化合物,其可由任何合适的、无论是天然的还是人工的来源获得或产生。试剂可以是氨基酸分

子、多肽、或其化学衍生物、或其组合物。试剂甚至可以是多核苷酸分子，其可以是有义或反义分子、或抗体，例如，多克隆抗体、单克隆抗体或单克隆人源化抗体。

已经开发出了各种策略来产生带有人特征的单克隆抗体，其不需要产生抗体的人细胞系。例如，有用的小鼠单克隆抗体通过连接啮齿动物可变区和人恒定区来进行“人源化”(Winter, G.和 Milstein, C.(1991)Nature 349, 293-299)。这降低了抗体的人抗小鼠免疫原性，但残余的免疫原性仍由外来V区框架而保留。而且，抗原结合特异性主要是鼠供体的。CDR移植和框架操作(EP 0239400)将抗体操作改进并优化到可能产生可在人中进行治疗应用的人源化鼠抗体的程度。人源化抗体可利用本领域公知的方法获得(例如US-A-239400中所述的)。

试剂可以通过接头与实体(如有机分子)相连，所述接头可以是可水解类型的双功能接头。

可以设计或由化合物文库中获取实体，所述化合物文库包括肽、以及其他化合物，如小的有机分子。

举例来说，实体可以是天然物质、生物大分子、或由诸如细菌、真菌、或动物(尤其是哺乳动物)细胞或组织中制备的提取物、有机或无机分子、合成剂、半合成剂、结构性或功能性模拟物、肽、肽模拟物、从完整蛋白质中切割的肽、或用合成方法(如，举例来说，利用肽合成仪或通过重组技术或其组合)合成的肽、重组剂、抗体、天然或非天然剂、融合蛋白质或其等价物和突变体、衍生物或其组合。

实体通常会是有有机化合物。对于一些情况来说，有机化合物将包含两种或多种烃基。在本文中，术语“烃基”指基团，其至少包含C和H并任选可包含一种或多种其他合适的取代基。这些取代基的例子可包括卤素、烷氧基、硝基、烷基、环基团等。除了取代基可能是环基团之外，取代基的组合可形成环基团。如果烃基包含超过一个C，则那些碳不必互相连接。例如，至少可通过合适的元素或基团连接2个碳。因此，烃基可包含杂原子。合适的杂原子对于所属领域技术人员来说是显而易见的，例如包括硫、氮和氧。对于一些应用来说，实体优选包含至少一个环基团。环基团可以是多环基团，如非稠合的多环基团。对于一些应用来说，实体至少包含与另一个烃基相连的所述环基团中的一个。

实体可包含卤素基团——如氟、氯、溴或碘基。

实体可包含一种或多种烷基、烷氧基、烯基、亚烷基和亚链烯基(alkenylene)——其可以是直链或支链的。

疾病

本发明的方面可用于治疗和/或预防和/或诊断和/或预测疾病——如列于 WO-A-98/09985 中的那些。

为了便于参考，现在提供该列表的一部分：巨噬细胞抑制和/或 T 细胞抑制活性以及由此产生的抗炎症活性；抗免疫活性，即抗细胞和/或体液免疫应答的抑制效应，包括不与炎症相关的应答；与病毒和/或其它细胞内病原体相关的疾病；抑制巨噬细胞和 T 细胞与细胞外基质成份和纤连蛋白粘附的能力，以及上调 T 细胞中 fas 受体表达；抑制不需要的免疫反应和炎症，包括关节炎，包括类风湿性关节炎，与超敏性相关的炎症，过敏反应，哮喘，全身性红斑狼疮，胶原病和其它自身免疫病，与动脉粥样硬化相关的炎症，动脉硬化，动脉粥样硬化心脏病，再灌注损伤，心搏停止，心肌梗塞，血管炎症疾病，呼吸窘迫综合征或其它心肺疾病，与消化性溃疡相关的炎症，溃疡性结肠炎和其它胃肠道疾病，肝纤维化，肝硬化或其它肝病，甲状腺炎或其它腺体疾病，肾小球性肾炎或其它肾和泌尿疾病，耳炎或其它耳鼻喉病，皮炎或其它皮肤病，牙周病或其它牙齿疾病，睾丸炎或附睾睾丸炎(epididimo-orchitis)，不孕症，睾丸损伤或其它免疫相关的睾丸疾病，胎盘机能障碍，胎盘机能不全，习惯性流产，子痫，子痫前期和其它免疫和/或炎症相关的妇科病，后眼色素层炎，中间眼色素层炎，前眼色素层炎，结膜炎，脉络膜视网膜炎，眼色素层视网膜炎，视神经炎，眼内炎症，如视网膜炎或囊状黄斑水肿，交感性眼炎，巩膜炎，色素性视网膜炎，退化性眼底疾病(degenerative fondus disease)的免疫和炎症成份，眼损伤的炎症成份，感染造成的眼炎，增殖性玻璃体视网膜炎，急性缺血性视神经病，过度瘢痕，诸如在青光眼滤过手术后的过度瘢痕，抗眼移植物的免疫和/或炎症反应和其它免疫和炎症相关的眼病，与自身免疫病或病况或病症相关的炎症(其中无论在中枢神经系统(CNS)还是在任何其它器官中抑制免疫和/或炎症会是有益的)，帕金森氏症，来自帕金森氏症治疗的并发症和/或副作用，AIDS 相关性痴呆复合 HIV 相关性脑病，Devic 氏症，Sydenham 舞蹈病，阿尔茨海默病和其

它 CNS 退行性疾病、病况或病症，中风的炎症成份，脊髓灰质炎后综合征，精神性疾病的免疫和炎症成份，脊髓炎，脑炎，亚急性硬化性全脑炎，脑脊髓炎，急性神经病，亚急性神经病，慢性神经病，Guillain-Barre 综合征，Sydenham 舞蹈病，重症肌无力，脑假瘤，唐氏综合征，亨廷顿氏症，肌萎缩性侧索硬化，CNS 压迫或 CNS 外伤或 CNS 感染的炎症成份，肌萎缩和肌营养不良的炎症成份，和中枢和外周神经系统的免疫和炎症相关疾病，病况或病症，外伤后的炎症，败血病性休克，传染病，外科手术的炎症并发症或副作用，骨髓移植或其它移植的并发症和/或副作用，例如由于病毒载体感染而造成的基因治疗的炎症和/或免疫并发症和副作用，或与 AIDS 相关的炎症，从而阻抑或抑制体液和/或细胞免疫应答，通过降低单核细胞或淋巴细胞的数量来治疗或改善单核细胞或白细胞增殖性疾病(如白血病)，用于在移植天然或人造细胞、组织和器官(如角膜、骨髓、器官、晶状体、起搏器、天然或人工的皮肤组织)的情况下预防和/或治疗移植排斥。特定癌相关性病症包括但不限于：实体瘤；血液产生的肿瘤，如白血病；肿瘤转移；良性肿瘤，例如血管瘤、听神经瘤、神经纤维瘤、沙眼(trachomas)、和生脓性肉芽肿；类风湿性关节炎；牛皮癣；眼血管生成病，例如糖尿病性视网膜病、早产儿视网膜病变、黄斑变性、角膜移植排斥、新生血管性青光眼、晶状体后纤维组织生成、潮红；Osler-Webber 综合征；心肌血管生成；斑块新生血管化(plaque neovascularization)；毛细血管扩张；血友病关节(hemophiliac joints)；血管纤维瘤；伤口肉芽形成；冠状侧支；脑侧支；动静脉畸形；缺血性肢体血管形成；新生血管性青光眼；晶状体后纤维形成；糖尿病性新生血管化；幽门螺旋杆菌相关的疾病、骨折、血管发生、造血作用、排卵、月经和胎盘形成。

疾病优选是癌，如急性淋巴细胞白血病(ALL)、急性髓细胞样白血病(AML)、肾上腺皮质癌、肛门癌、膀胱癌、血癌、骨癌、脑肿瘤、乳腺癌、女性生殖系统癌、男性生殖系统癌、中枢神经系统淋巴瘤、子宫颈癌、儿童横纹肌肉瘤、儿童肉瘤、慢性淋巴细胞白血病(CLL)、慢性髓细胞样白血病(CML)、结肠和直肠癌、结肠癌、子宫内膜癌、子宫内膜肉瘤、食管癌、眼癌、胆囊癌、胃癌、胃肠道癌、毛细胞白血病、头和颈癌、肝细胞癌、霍奇金氏病、咽下部癌、Kaposi 氏肉瘤、肾癌、喉癌、白血病、肝癌、肺癌、恶性纤维性组织细胞瘤、恶性胸腺瘤、黑素瘤、间皮瘤、多发性骨髓瘤、骨髓瘤、鼻腔和鼻旁窦癌、鼻咽癌、神经系统癌、成神经细胞瘤、非霍奇金氏淋

巴瘤、口腔癌、口咽癌、骨肉瘤、卵巢癌、胰腺癌、甲状旁腺癌、阴茎癌、咽癌、垂体肿瘤、浆细胞瘤、原发性 CNS 淋巴瘤、前列腺癌、直肠癌、呼吸系统、成视网膜细胞瘤、唾液腺癌、皮肤癌、小肠癌、软组织肉瘤、胃癌、胃癌、睾丸癌、甲状腺癌、泌尿系统癌、子宫肉瘤、阴道癌、血管系统、瓦尔登斯特伦巨球蛋白血症和 Wilms 氏肿瘤。

试剂盒

用于本发明方法中的材料是理想地适用于制备试剂盒的。

这样的试剂盒可包括多个容器，各自带有用于本文所述的方法中的各种试剂(通常以浓缩的形式)中的一种或多种，例如包括，第一限制酶、第二限制酶、交联剂、用于连接的酶(如连接酶)和解交联的试剂(如蛋白酶 K)。

寡核苷酸也可在容器中被提供，其可以是任意形式的，如冻干或溶液(如，蒸馏水或缓冲溶液)等。

在本发明优选的方面中，提供了试剂盒，其包括如本文所述的探针组、阵列并任选有一种或多种标记。

通常还包括一套说明书。

应用

有益的是，为了获得关于核苷酸序列(如基因组基因座)空间组构信息而在体外或体内使用本发明。

举例来说，4C 技术可用于研究一个或多个基因座的三维组构。该技术尤其可用于研究一种或多种转录因子在一个或多个基因座三维组构中的作用。

进一步举例来说，4C 技术可用于研究反式作用因子和顺式调节 DNA 元件的作用。

进一步举例来说，4C 技术可用于在体外或体内研究长程基因调节。

进一步举例来说，4C 技术可用于研究染色体内邻近区域和相互作用。

进一步举例来说，4C 技术可用于研究染色体间邻近区域和相互作用。

进一步举例来说，4C 技术可用于鉴定与启动子、增强子、沉默基因、隔离子、基因座控制区、复制起点、MAR、SAR、着丝粒、端粒或任何其它在调节网络中的目的序列一起作用的核苷酸序列。

进一步举例来说, 4C 技术可用于鉴定在下述情况下造成表型(疾病)的基因, 其中突变和/或缺失碰巧影响远距离调节元件并因此对它们作图不能提供这样的信息的情况下。

进一步举例来说, 4C 技术可用于最终重建基因座、大基因组区域或甚至完整染色体的空间构象。

进一步举例来说, 4C 技术可用于确定在核空间中将某些染色体保持在一起的潜在锚定序列。

进一步举例来说, 4C 技术可用于最终以高解析度重建染色体相互间的位置。

进一步举例来说, 4C 技术可用于诊断(如产前诊断)以检测或鉴定基因组重排和/或异常, 如易位、缺失、倒位、重复。

一般性的重组 DNA 方法技术

除非另外指定, 本发明使用常规化学、分子生物学、微生物学、重组 DNA 和免疫学技术, 其在所属领域普通技术人员的能力范围内。这些技术在文献中有解释。例如参见, J. Sambrook, E. F. Fritsch, 和 T. Maniatis, 1989, *Molecular Cloning: A Laboratory Manual*, 第二版, Books 1-3, Cold Spring Harbor Laboratory Press; Ausubel, F. M. 等.(1995 和定期补录; *Current Protocols in Molecular Biology*, 第 9、13、和 16 章, John Wiley & Sons, New York, N.Y.); B. Roe, J. Crabtree, 和 A. Kahn, 1996, *DNA Isolation and Sequencing: Essential Techniques*, John Wiley & Sons; M. J. Gait(编者), 1984, *Oligonucleotide Synthesis: A Practical Approach*, Irl Press; 和, D. M. J. Lilley and J. E. Dahlberg, 1992, *Methods of Enzymology: DNA Structure Part A: Synthesis and Physical Analysis of DNA Methods in Enzymology*, Academic Press。这些普通课本中的每一本都纳入本文参考。

本发明现在将进一步举例来描述, 其目的是用于帮助所属领域普通技术人员实施发明, 而并不想以任何方式限制本发明的范围。

实施例 1

与图 2、13、14、15、16、17、19 相配的材料与方法部分

4C 技术

3C 技术过程的初始步骤如前述(Splinter 等(2004). *Methods Enzymol* **375**, 493-507(2004))进行, 在 *Hind*III 片段间产生连接产物。该 *Hind*III 连接的 3C 模板(~50 µg)以 100ng/µl 用 50U 第二、频繁切割的限制酶消化过夜, 所述酶是 DpnII(HS2, Rad23A)或 *Nla*III(β-major)。为了避免 DNA 环形成受限制(Rippe 等.(1995)*Trends Biochem Sci* **20**, 500-6), 注意选择第二限制酶, 其不在距离划分目的限制性片段(即‘诱饵’)的 *Hind*III 限制酶位点约 350-400 bp 内切割。第二限制酶消化后, 用苯酚抽提 DNA, 用乙醇沉淀, 然后以低浓度连接(用 200 U 连接酶(Roche)于 16⁰C 连接溶于 14 ml 中的 50 µg 样品 4 小时), 从而促进 DpnII-或 DpnII-环形成。用苯酚抽提连接产物并用乙醇沉淀, 将糖原(Roche)用作载体(20µg/ml)。用 50U 在第一和第二限制酶识别位点间切割诱饵的第三限制酶消化过夜, 使目的环线性化; 第三限制酶为: *Spe*I(HS2)、*Pst*I(Rad23A)和 *Pfl*mI(β-major)。进行该线性化步骤以帮助接下来在第一轮 PCR 扩增期间进行引物杂交。用 QIAquick 核苷酸分离(250)柱(Qiagen)纯化消化的产物。

用 Expand Long Template PCR 系统(Roche)进行 PCR 反应, 使用经仔细优化的条件来保证最长 1.2 kb 片段的线性扩增(80%的 4C-PCR 片段小于 600 bp)。PCR 条件如下: 94°C 进行 2 分钟, 94°C 15 秒、55°C 1 分钟和 68°C 3 分钟进行 30 个循环, 然后最后步骤为 68°C 进行 7 分钟。确定仍旧显示线性扩增范围的最大模板量。为此, 向 PCR 反应中加入连续稀释的模板, 扩增的 DNA 材料在琼脂糖凝胶上分离并用 ImageQuant 软件定量 PCR 产物。通常, 每 50 µl PCR 反应中用 100-200 ng 模板能在线性扩增范围内产生出产物。合并 16 至 32 个 PCR 反应, 并用 QIAquick 核苷酸分离(250)系统(Qiagen)纯化该 4C 模板。将纯化的 4C 模板标记并根据标准 ChIP-芯片规程(Nimblegen Systems of Iceland, LLC)与阵列杂交。区别标记的基因组 DNA(其用 4C 过程中所用的第一和第二酶消化)用作对照模板来校正杂交效率的差异。对于每个实验, 用交替的染料定位(orientation)标记 2 个独立处理的样品。

所用的 4C-引物序列:

HS2: 5'-ACTTCCTACACATTAACGAGCC-3',
5'-GCTGTTATCCCTTTCTCTTCTAC-3'

Rad23A: 5'- TCACACGCGAAGTAGGCC-3',
5'- CCTTCCTCCACCATGATGA-3'

β -major: 5'-AACGCATTTGCTCAATCAACTACTG-3',
5'-GTTGCTCCTCACATTTGCTTCTGAC-3'

4C 阵列

阵列和分析基于 NCBI 建立的 m34。探针(60-聚体)选自 *Hind*III 位点上和下游 100 bp 处的序列。将 CG 含量朝 50%优化,用于使杂交信号均一化。为了避免交叉杂交,从探针组中去除与高丰度重复序列(RepBase 10.09)³具有任何相似性的探针。另外,在基因组中有超过两个 BLAST 命中情况的探针也去除出探针组。用 MegaBLAST(Zhang 等.(2000)*J Comput Biol* 7, 203-14)利用标准设置进行序列比对。命中被定义为有 30 nt 或更长的比对排列(alignment)。

4C 数据分析

为每个探针计算 4C-样品/基因组 DNA 的信号比率,并用 Nimblegen 系统提供的 SignalMap 软件使数据可视化。用 R 软件包(<http://www.r-project.org>)、Spotfire 和 Excel 分析数据。未处理的杂交比率显示沿着染色体模板有由 20-50 个阳性 4C-信号组成的簇。为了确定这些簇,应用了连续平均法。使用各种窗口大小,范围从 9-39 个探针,其都能鉴定同一簇。显示的结果基于 29 个探针的窗口大小(平均 60 kb),并与在随机数据上进行的连续平均进行比较。每个阵列分别如此进行。因而,所有测量值都相对于特定阵列的振幅(amplitude)和噪声加以评价。假发现率(False Discovery Rate, FDR)被定义为(假阳性数)/(假阳性数 + 真阳性数),其以如下公式确定:(随机化组中的阳性数)/(数据中的阳性数)。利用由上到下的方法(top down approach)确立阈值水平,以建立 FDR<0.05 的最小值。

接着对生物学重复实验进行了比较。在两个重复实验中都达到阈值的窗口被认为是阳性的。当比较随机化的数据时,在两个重复实验中都没有高于阈值的窗口。将染色体模板上直接邻接的阳性窗口连接(不允许有缺口),从而产生阳性区域。

表达分析

对于每个组织, 根据 Affymetrix 规程(小鼠 430_2 阵列)进行了 3 个独立微阵列操作。用 RMA *ca*-工具(www.bioconductor.org)使数据标准化, 并对于每个探针组将 3 个微阵列的测量值加以平均。另外, 当多个探针组代表相同的基因时, 也将它们平均。用 Mas5calls(Affy 库: www.bioconductor.org)确定“存在”、“不存在”和“边缘”访问(call)。在所有 3 个阵列中都表示为“存在”访问且表达值大于 50 的基因被称为表达的基因。将“胎儿肝特异性基因”归类为达到我们对在胎儿肝中表达的标准、并且表达值超过胎儿脑 5 倍的基因。为了测量每个基因周围的整体转录活性, 运用连续总值。为此, 我们使用了转换成对数的表达值。对于每个基因, 我们计算了在基因起点上游 100 kb 处和末端下游 100 kb 处的窗口中发现的所有基因(包括基因本身)的表达总值。将在阳性 4C 区域中找到的活性基因的结果值(对于肝中的 HS2、脑中的 Rad23A 和肝中的 Rad23A, 分别为 $n = 124$ 、 123 和 208)与阳性 4C 区域外活性基因获得的值(分别 $n = 153$ 、 301 和 186 , 其中 $n=153$ 对应于染色体 7 中最着丝粒相互作用区(the most centromeric interacting region)和端粒之间存在的有活性、无相互作用的基因数)作比较; 利用单尾 Wilcoxon 秩和检验来比较这两组。

FISH 探针

使用以下 BAC 克隆(BACPAC Resources Centre); 针对 Hbb-1 的 RP23-370E12, 针对染色体 7 的 80.1Mb 处(OR 基因簇)的 RP23-317H16, 针对 Uros 的 RP23-334E9, 针对染色体 7 的 118.3 Mb 处的 RP23-32C19, 针对染色体 7 的 130.1Mb 处的 RP23-143F10, 针对染色体 7 的 73.1Mb 处的 RP23-470N5, 针对染色体 7 的 135.0Mb 处(OR 基因簇)的 RP23-247L11, 针对 Rad23A 的 RP23-136A15, 针对染色体 8 的 21.8 Mb 处的 RP23-307P24, 和针对染色体 8 的 122.4 Mb 处的 RP23-460F21。对于染色体 7 的着丝粒特异性探针, 我们使用 P1 克隆 5279(Genome Systems Inc.), 其与 DNA 区段 D7Mit21 退火。用 BioPrime Array CGH Genomic Labeling System(Invitrogen)制备随机引物标记的探针。标记前, 用 DpnII 消化 DNA 并用 DNA 清洁和浓缩剂-5 试剂盒(Zymo research)纯化。消化的 DNA(300 ng)用 SpectrumGreen

dUTP(Vysis)或 Alexa fluor 594 dUTP(Molecular probes)标记,并通过 GFX PCR DNA 和凝胶条带纯化试剂盒(Amersham Biosciences)纯化,来去除未掺入的核苷酸。在由鼠胚胎干细胞制备的中期涂片上测试标记的探针的特异性。

冷冻 FISH

如前所述进行冷冻 FISH⁵。简而言之, E14.5 肝和脑固定于 4% 多聚甲醛/250 mM HEPES(pH 7.5)中 20 分钟并切成小组构块,然后在 8% 多聚甲醛中于 4°C 再固定 2 小时。固定的组构块于室温浸泡在 2.3 M 蔗糖中 20 分钟,装在样品座上并在液氮中急速冻结。组构块储存于液氮中,直至进行切片。用带有冷冻附件的 Reichert 超薄切片器 E(Leica)切成约 200 nm 厚的超薄冷冻切片。用充满蔗糖的环,将切片转移至盖玻片上并储存于 -20°C。为了进行杂交,用 PBS 洗切片去除蔗糖,于 37°C 用溶于 2xSSC 中的 250 ng/ml RNA 酶处理 1 小时,在 0.1 M HCL 中孵育 10 分钟,在连续稀释的乙醇中脱水并在 70%甲酰胺/2xSSC(pH 7.5)中于 80°C 变性 8 分钟。临进行探针杂交前,再次对切片脱水。500 ng 标记的探针与 5 μ g 小鼠 Cot1 DNA(Invitrogen)共沉淀并溶解于杂交混合液(50% 甲酰胺, 10% 硫酸葡聚糖, 2xSSC, 50 mM 磷酸盐缓冲液, pH 7.5)中。探针于 95°C 变性 5 分钟,于 37°C 重新退火 30 分钟并于 37°C 杂交至少 40 小时。杂交后洗涤之后,用溶于 PBS/0.05%吐温-20 的 20 ng/ml DAPI(Sigma)复染核,并在 Prolong Gold 抗退色试剂(Molecular Probes)中固定。

用装有 CCD 照相机和 Isis FISH 成像系统软件(Metasystems)的 Zeiss Axio Imager Z1 落射荧光显微镜(x100 倍平面复消色差透镜, 1.4 倍油物镜)收集图像。最少分析 250 个 β -球蛋白或 Rad23A 等位基因,并由不知道用于切片的探针组合的人将其评为与位于基因组其它地方的 BAC 重叠或不重叠。进行重复的吻合度测试(G-统计)⁶来评估 4C-阳性相对于 4C 阴性区域的测量值之间差异的显著性。表 2 中提供了结果的总述。

尽管我们在背景(0.4-3.9%)和真实的(5-20.4%)相互作用频率之间发现了有统计学显著性差异,但是可以清楚的是由冷冻 FISH 测得的频率比其它人用不同的 FISH 规程测得的那些频率更低。切片可能分开了一些相互作用的基因座,因此冷冻 FISH 测量将轻微低估真实的相互作用频率。在另一方面,由于在 z-方向上有限的解析度,因此当前的 2D-和 3D FISH 规程将高估这些

百分比。将来,改进的显微镜技术与更特异的 FISH 探针的组合将更好地揭示真实的相互作用频率。

实施例 2

基本如所述的(Splinter 等., (2004)*Methods Enzymol.* 375: 493-507)进行 3C 过程(即用甲醛固定,用(第一)限制酶消化,重新连接交联的 DNA 片段并进行 DNA 纯化),产生含限制性片段的 DNA 混合物('3C 模板'),这些限制性片段由于它们原来在核空间中接近而被连接。

进行反向 PCR 来扩增所有与给定的限制性片段("诱饵";因为它含启动子、增强子、隔离子、基质附着区、复制起点或任何其它第一(靶)核苷酸序列而被挑选)连接的片段。

为此,通过用第二限制酶(优选识别四个或五个核苷酸的序列的频繁切割的酶)消化 3C 模板,然后在有利于分子内相互作用的稀释的条件下连接,由此产生 DNA 环。为了使由于拓扑学限制而造成的在环形成中的偏倚(Rippe 等, (2001)*Trends in Biochem. Sciences* 26, 733-40)最小化,应该挑选优选在距离第一限制性位点>350-400bp 处切割诱饵的第二限制酶。为了增加反向 PCR 扩增的效率和可重复性,最好用在诊断性第一和第二限制性位点之间切割诱饵的限制酶(如识别 6 或更多 bp 的限制性酶)使环线性化,然后进行 PCR 扩增。

用第二限制酶消化 3C 模板,通过在稀释条件下连接来环化,并使含诱饵的环线性化,这些步骤在这些 DNA 操作的标准条件下进行,以产生用于进行反向 PCR 扩增的 DNA 模板('4C 模板')。

因此,在 100 μ l 中用 20U 第二限制酶消化 10 μ g 3C 模板(过夜),然后使酶热失活并纯化 DNA。在 10 ml 中(1 ng/ μ l DNA)用 50U T4 连接酶进行连接(于 16 $^{\circ}$ C 4 小时,于 RT 30 分钟),然后进行 DNA 纯化。最后,在 100 μ l 中用 20U 限制酶使目的环线性化(过夜),然后再次进行 DNA 纯化。

对于反向 PCR,设计两个诱饵特异性引物,每个分别尽可能接近第一限制酶识别位点并与第二限制酶识别位点直接相邻,每种的 3'末端朝外,从而使延伸立即穿过限制酶位点进入与诱饵相连的片段。优选(每 50 μ l PCR 反应混合物中)使用 100-400 ng DNA 的 4C 模板实施用这些引物进行的反向 PCR,从而使每个 PCR 反应包括了最大数量的连接事件。我们根据制造商的方法

利用缓冲液 1 应用 Expand Long Template PCR 系统(Roche)进行了反向 PCR。

进行了以下 PCR 循环:

1. 94⁰C 2 分钟
2. 94⁰C 15 秒
3. 55⁰C 1 分钟
4. 68⁰C 3 分钟
5. 重复步骤 2-4 共 29 次(或 25-40 次间的任意次数)
6. 68⁰C 7 分钟
7. 结束

进行凝胶电泳来分析各 PCR 反应间的可重复性。通常应当获得一致的产物模式。

为了通过随机引发和阵列杂交获得足以进行标记的材料,合并多个 PCR 反应(每个都是 30 个循环的 PCR 后所获得的), (而不是在每个反应中增加 PCR 循环数)。作为随机引发标记的可选方法,可将标记的核苷酸加入 PCR 的最后几个循环中(如 30 个循环(无标记)+ 10 个循环(有标记))。

实施例 3

利用 4C 技术检测易位

用 4C 技术测量针对来自健康受试者的细胞中和来自患者的细胞中给定的染色体 A 上出现的给定的序列 X 的相互作用频率,所述患者在染色体 A 和 B 间带有单一的交互易位,其中断点接近于序列 X(如图 8 所示)。

在正常细胞中,该分析揭示出针对(几乎)每个位于染色体 A 上序列 X 的 0.2-10Mb 内的探针的杂交信号得到了提高(即与 X 的频繁相互作用)(显示出强交联信号的染色体区域的实际大小主要依赖于与阵列杂交的样品的复杂性)。在相同染色体 A 上的其它地方以及在其它染色体上,没有观察到这样带有提高的杂交信号的(在线性 DNA 模板上)的大的探针区域。

可是在患者细胞中,用位于断点另一边的所有染色体 A 探针得到的杂交信号减少了~50%(染色体 A 的一个拷贝仍旧是完整的,并将产生正常的信号),而对于染色体 B 上断点边缘的探针,则观察到了提高了的杂交信号的独特集中(即不存在于正常细胞中)。事实上,染色体 B 上显示从无杂交信号探针到强杂交信号探针的突然转变揭示了染色体 B 上断点的位置。

实施例 4

分析 4C 技术结果

用 4C 技术表征小鼠 β -球蛋白基因座控制区(locus control region, LCR)的基因组环境,关注含其超敏位点 2(hypersensitive site2, HS2)的限制性片段。LCR 是强红血球特异性转录调节元件,对于高 β -球蛋白基因表达水平是必需的。 β -球蛋白基因座出现在染色体 7 上 97 Mb 位置处,其中它位于只在嗅觉神经元中转录的嗅觉受体基因的大的 2.9 Mb 的簇中。分析 2 种组织中的相互作用: E14.5 胎儿肝(其中 LCR 是有活性的而且 β -球蛋白基因高度转录),和 E14.5 胎儿脑(其中 LCR 是无活性的而且球蛋白基因是沉默的)。在两种组织中,绝大多数相互作用在染色体 7 上的序列中被发现,而 6 个无关的染色体(8、10、11、12、13、14)只检测到很少的 LCR 相互作用(图 12a)。染色体 7 上的最强的信号在以 β -球蛋白染色体位置为中心的周围 5-10 Mb 区域内被发现,符合相互作用频率与物理相连的 DNA 序列间的距离(以碱基对计)成反比的观点。不可能定量说明该区域中的相互作用。我们的理由是这些邻近的序列太频繁地与 β -球蛋白在一起,使得它们在我们的杂交样品中大大地过量表达,使相应探针都饱和了。我们用 1: 10 和 1: 100 稀释的样品进行杂交,发现了信号强度对于在外边和边缘的探针都降低了,但在该区域内的探针则没有降低(数据未显示),这确证了这个理由。

4C 过程生成了高度可重复的数据。图 2b-c 显示了针对染色体 7 上两个 1.5 Mb 区域(大约距离 β -球蛋白基因 25 Mb 和 80 Mb)的 4C-信号相对于对照杂交信号的未处理的比率。在该解析度水平时,独立处理的样品的结果几乎是相同的。在胎儿肝和脑中,都有阳性信号簇在染色体 7 上被鉴定出,其通常位于距离 β -球蛋白上千万碱基的染色体位置上。这些簇通常由最少 20-50 个并列排在染色体模板上的探针组成,所述探针的信号比率增加(图 12b-c)。每个在阵列上的探针分析独立连接事件。而且,每个细胞仅有 2 个拷贝的 HS2 限制性片段,其每一个仅能连接于一个其它限制性片段。因此,检测到与 20 或更多个邻近限制性片段的独立连接事件强有力地显示了,在多种细胞中,相应基因座与 β -球蛋白 LCR 接触。

为了确定这些簇的统计学显著性,各实验的数据按序排列在染色体图谱上并用窗口大小约为 60 kb 的连续平均算法分析。用随机改组的数据的连续

平均分布来设定阈值,允许有 5%的假发现率。该分析鉴定了胎儿肝中的 66 个簇和脑中的 45 个簇,它们在重复的实验中被重复发现(图 12d-f)。确实,高解析度的 FISH 确认了这些簇真实地代表了频繁相互作用的基因座(见下)。

因此,4C 技术通过检测在染色体位置簇集的多个限制性片段的独立连接事件来鉴定长程相互作用基因座。

用不同的反向 PCR 引物组进行一系列完全独立的 4C 实验,研究 β -major 基因的基因组环境, β -major 基因位于 HS2 下游~50 kb 处。在胎儿肝中, β -major 基因是高度转录的并与 LCR 频繁接触。在胎儿肝和脑中,发现与 β -major 长程相互作用的簇和与 HS2 的几乎完全相同,进一步证实了这些基因座频繁与 β -球蛋白基因座接触(图 16)。

实施例 5

活性和失活的 β -球蛋白基因座占据不同的基因组环境。

2 种组织间的比较揭示,胎儿肝中活跃转录的 β -球蛋白基因座和其脑中转录沉默的对应物相比与完全不同组的基因座相互作用($\tau=-0.03$; Spearman 氏秩相关)(图 12f)。这排除了结果受探针序列组成的影响。在胎儿肝中,相互作用的 DNA 片段位于以 β -球蛋白基因座为中心的周围 70 Mb 区域内,其大多数(40/66)朝向染色体 7 的端粒分布。在胎儿脑中,与胎儿肝相比,相互作用的基因座在距离 β -球蛋白相似或甚至更远处被发现,而且最主要的相互作用(43/45)朝向染色体 7 的端粒分布。这些数据证实了有活性的和无活性的 β -球蛋白基因座接触染色体 7 的不同部分。

微阵列上有另外 6 个染色体(8、10、11、12、13 和 14)的代表物。这些染色体上的强杂交信号很少,通常显示为在线性 DNA 模板上呈分离状态,并经常在重复实验中不存在。另外,沿着这些染色体的连续平均水平从未可重复地接近于针对染色体 7 计分的水平(图 17)。因此,我们的数据显示, β -球蛋白基因座大多与同一染色体上其它地方的基因座接触,符合该基因座在其自身染色体区域内部的优选位置。我们注意到, α -球蛋白基因座也出现在阵列上(染色体 11),并没有呈现出与 β -球蛋白的阳性相互作用,符合近来由 FISH 的证明结果,即小鼠 α -和 β -球蛋白在核空间中不频繁相遇(Brown, J. M. 等.(2006)*J Cell Biol* 172, 177-87)。

为了更好地理解染色体 7 上观察到的长程相互作用的相关性,我们比较了相互作用的基因座与基因的染色体位置。另外,进行 Affymetrix 表达阵列分析以确定在两种组构中在这些位置上的转录活性。尽管胎儿肝和脑中相互作用区域的平均大小是类似的(分别为 183 kb 和 159 kb),但是在它们的基因成分和活性上观察到了巨大的差异。在胎儿肝中,80% 的 β -球蛋白相互作用基因座含一个或多个活跃转录的基因,而在胎儿脑中,大多数(87%)显示没有可检测的基因活性(图 14)。因此在两种组构中, β -球蛋白基因座包含在非常不同的基因组环境中。在脑中(其中该基因座是无活性的),它主要接触朝向染色体 7 着丝粒定位的转录沉默基因座。在胎儿肝中(其中该基因座是有高活性的),它优先与更显著地朝着染色体 7 端粒侧定位的活跃转录的区域相互作用。重要的是,4C 技术能将 Uros 和 Eraf (距离 β -球蛋白~30Mb)鉴定为在胎儿肝中与有活性的 β -球蛋白基因座相互作用的基因,符合以前由 FISH 得到的观察结果(Osborne, C. S. 等.(2004)*Nat Genet* **36**, 1065-71(2004))。有趣的是,观察到在脑中另外两个嗅觉受体基因簇有接触,其存在于染色体 7 上,各自位于 β -球蛋白的两边,并分别距离 β -球蛋白 17 和 37 Mb。

在胎儿肝中,并非染色体 7 上所有的转录区域都与有活性的 β -球蛋白基因座相互作用。所以,我们寻找专门由相互作用的基因座共享、而不由胎儿肝中其它活性区域共享的共同特征(denominator)。 β -球蛋白基因——Uros 和 Eraf 都是可由相同的转录因子组调节的红细胞特异性基因,而且吸引人的观点是,这些因子协调了它们的靶基因在核空间中的表达。我们比较了来自 E14.5 胎儿肝的与胎儿脑的 Affymetrix 表达阵列数据,以鉴定胎儿肝中优先表达(>5 倍以上)的基因。由此,染色体 7 上 28% 的有活性的基因被归类为“胎儿肝特异性的”,其中 25% 在共定位区域中被发现。因此,我们发现在共定位区域中“胎儿肝特异性”基因并不丰富。更重要的是,66 个相互作用的区域中的 49 个(74%)不含有“胎儿肝特异性”基因,因此断定我们的数据没有显示核空间中组构特异性基因协调表达的证据。 β -球蛋白基因以异常高的速率被转录,而且接下来询问基因座是否优先与其它高转录活性区域相互作用,不管这些区域是高表达的基因抑或带有高密度活性基因的区域。利用 Affymetrix 计数来测定基因活性,我们进行了连续总和算法(running sum algorithm)来测量活跃转录的基因周围 200 kb 区域内的总体转录活性。该分析揭示在相互作用的基因周围的转录活性不高于在染色体 7 上无相互作用活

性的基因周围的转录活性($p = 0.9867$; Wilcoxon 秩和)。

实施例 6

管家基因的基因组环境在组织之间很大程度上是保守的

接着研究在两种组织中相似表达的基因是否也转换其基因组环境。Rad23A 是普遍表达的基因，其位于染色体 8 上主要由管家基因组成的基因密集簇中。在 E14.5 胎儿肝和脑中，该基因和其直接相邻的许多基因都是有活性的。进行 4C 分析并鉴定了与 Rad23A 距离长达 70 Mb 处的基因座的许多长程相互作用。重要的是，与 Rad23A 的相互作用在胎儿肝和脑中高度相关($\tau=0.73$; Spearman 氏秩相关)(图 14a)。另外，这些基因座共有的特点是它们含有活性转录的基因。因此，在这两种组织中都约有 70% 含至少一种有活性的基因(图 14b-c)。如连续总和算法所确定的(对两种组织, $p < 0.001$)，与染色体其它地方的活性基因相比，相互作用的基因周围的区域显示出统计学上显著更高的基因活性水平。因此，与 β -球蛋白基因座不同，位于基因丰富区域中的 Rad23A 基因优先与其它增加转录活性的染色体区域跨距离相互作用。通过 FISH 观察到，含 Rad23A 的染色体区域大多位于其染色体区域的边缘(90%)或外边(10%)(未发表, D. Noordermeer, M. Branco, A. Pombo 和 W. de Laat)。可是，4C 分析仅揭示了染色体内的相互作用，染色体 7、10、11、12、13 或 14 上则没有区域可重复地达到我们严格的相互作用标准。因此，Rad23A 主要参与在两种非常不同的组织中相似的染色体内相互作用。如果 Rad23A 在这些不相关的染色体上有优选的邻近基因座，则它们的相互作用的频繁程度还不足以在本文 4C 技术所用的条件下被检测。

实施例 7

通过高解析度显微镜验证 4C 技术

为了验证 4C 技术得到的结果，进行了冷冻 FISH 实验。冷冻 FISH 是近来开发的显微镜技术，其相对于现有 3D-FISH 规程的优点在于，它更好地保留核超结构，并通过制备超薄冷冻切片来改善 z 轴解析度(Branco, M. R. & Pombo, A(2006). *PLoS Biol* 4, e138)。4C 数据的验证是通过在由 E14.5 肝和脑制备的 200 nm 超薄切片中测量 β -球蛋白或 Rad23A 等位基因(通常 $n>250$)如何频繁地与 15 个以上的选择的染色体区域共定位而进行的。重要

的是,所有通过冷冻 FISH 测量的相互作用频率完美地符合 4C 结果(图 16)。例如,通过 4C 技术鉴定为与 β -球蛋白有相互作用的远距离区域比未由 4C 检测到的介于其间的区域更频繁地共定位(分别为 7.4%和 9.7%,相对于 3.6%和 3.5%)。另外,由 4C 技术鉴定出与胎儿脑而不是肝中 β -球蛋白相互作用的两个远距离嗅觉受体基因簇在脑中的共定位频率分别被评为 12.9%和 7%,而肝切片中为 3.6%和 1.9%。总之,4C 技术鉴定为阳性的基因座所测得的共定位频率都显著高于背景基因座所测得的频率($p < 0.05$; G-检验)。我们断定 4C 技术如实地鉴定出相互作用的 DNA 基因座。最后,我们使用冷冻 FISH 证明被鉴定为与 β -球蛋白相互作用的基因座也频繁地相互接触。对于胎儿肝中跨越大染色体距离的 2 个活性区域来说(图 17)以及对于脑中在染色体上相距较远的两个无活性的 OR 基因簇来说(图 16),都是如此。有趣的是,这两个远距离 OR 基因簇间频繁的接触也在胎儿肝中被发现,在这里它们不与含有活跃转录的 β -球蛋白基因座的 OR 基因簇相互作用。这些数据显示,相距较远的 OR 基因簇间的核相互作用不是所分析的胎儿脑组构所独有的。这似乎在推测,这种空间接触对许多 OR 基因间的联系提供了帮助,这种联系对于确保每个嗅觉神经元中仅一个等位基因被转录是必需的(Shykind, B.(2005)*Hum Mol Genet* 14 Spec No 1, R33-9。

实施例 8

有活性和无活性的染色质结构域的核组织

本文所述的观察结果证实了不仅有活性的、而且无活性的基因组区域也在核空间中形成涉及许多长程接触的独特区域,这有力地提示了每种 DNA 片段具有其自身优选的相互作用组。我们的数据提示,当 β -球蛋白基因座被开启时,它脱离转录沉默基因组环境,并进入有益于与活性结构域相互作用的核区域。预计这种在转录活化后戏剧性的再定位很可能仅为达到了某种表达水平、并且更重要的是在线性染色体模板上与其它活性基因分隔开(如对于 β -球蛋白的情况)的组织特异性基因的标志。这提示,在无活性的基因组基因座之间和有活性的基因组基因座之间都鉴定到的长程相互作用的广泛网络,反映出细胞与细胞在染色体构象上的差异,而不是间期的动态移动的结果(Chakalova 等.(2005)*Nat Rev Genet* 6, 669-77(2005)。推测起来,细胞分裂后不同程度的去凝聚作用驱使有活性的基因组区域远离无活性的染色质

(Gilbert, N. 等.(2004)*Cell* 118, 555-66(2004)), 并通过染色质结合蛋白质间的亲和性使具有相似染色质组成的远距离基因座间的接触稳定。远距离基因座间的空间相邻可以是功能性的, 但也可简单地是染色体展开模式的结果。尽管单个基因座能在有限的核体积中移动, 染色体的一般构象将在整个细胞周期内被极大地保持并需要细胞分裂才能被重置。该观点符合活细胞成像研究的结果(其显示带标签的 DNA 基因座在核内受限地运动(Chubb 等.(2002)*Curr Biol* 12, 439-45(2002))), 而且很好地符合以下研究结果, 即显示了核染色质位置信息在细胞分裂过程中被频繁传递而在细胞群体中并不被保存(Essers, J. 等. *Mol Biol Cell* 16, 769-75(2005); Gerlich, D. 等. *Cell* 112, 751-64(2003))。

实施例 9

原理论证: 4C 技术精确检测患者样品中的缺失(图 19)

通过距离第一断点上游(左侧)2Mb(A)或 1.3Mb(B)的靶核苷酸序列使用 4C 揭示的白血病患者中出现的缺失的存在。注意, 缺失引起缺失区 DNA 相互作用信号的降低, 但是也引起最后断点直接下游(右侧)序列的 DNA:DNA 相互作用频率的增加。当仔细研究与靶核苷酸序列 B 的相互作用时(参见底部二图)这尤其明显。基于 4C 数据在缺失区的每侧都设计引物并通过测序鉴定断点: 一般字体是缺失上游的序列, 粗体指示插入核苷酸, 下划线的是缺失下游的序列。

实施例 10

原理论证: 4C 技术精确地检测患者样品中的平衡易位(图 25)

对平衡易位的检测的原理论证。(R. Burnett 等, *Blood*, Vol 84, No 4(August 15), 1994: 1232-1236 页)中描述的 t(1;7)易位的检测。染色体 7 上 TCRb 基因座侧翼的靶核苷酸序列, 红色信号代表与位于 TCRb 基因座上游的靶序列的 DNA:DNA 相互作用, 而蓝色信号代表与位于 TCRb 基因座下游的靶序列的 DNA:DNA 相互作用。描述的是染色体 1 上找到的相互作用的 DNA 信号。顶部一排显示理论信号分布。中下一排显示实际信号分布。底部一排显示并列在染色体模板上的单个探针解析度的信号。注意, 在平衡易位的情况下, 断点侧翼的靶核苷酸序列将显示互斥的染色体间 DNA 相互作用

用信号组,其在易位配体染色体的线性染色体模板上彼此直接相邻。测序的断点位置(在 Burnett 等, 1994 中描述)在底部一排用箭头指示。

实施例 11

原理论证: 4C 技术精确地检测患者样品中不平衡的易位(图 27)

不平衡易位的检测。(RJ Galjaard 等, Am J Med Genet A. 2003 Aug 30;121(2):168-73)中描述的 t(4;7)易位的检测。定位于染色体 7 的靶核苷酸序列;描述的相互作用的 DNA 信号定位于染色体 4。使用了位于染色体 7 上断点上游(5')和下游(3')的两个靶序列。指示了位于染色体 4 上的相互作用的 DNA 信号(将蓝色用于两个靶序列)。在这个患者中检测到了染色体 4 上相互作用的 DNA 片段簇之间的区域。顶部:完整的染色体 4 的信号。底部一排 4C 数据:染色体 4 上断点周围 11.5MB 区域的信号。基于这些 4C 数据,鉴定了染色体 4 上包含易位断点的 *HindIII* 限制性片段并将其用于通过测序来将该断点作图。在图的底部提供了序列,其中下划线的序列来自染色体 4,粗体能在 7 和 4 上找到而一般字体的序列来自染色体 7。

实施例 12

通过 4C 技术快速高解析度鉴定平衡的基因组重排

简述

研究遗传变异的现有技术不足以(fail)或不能精确地鉴定频繁发生在人群中并能导致疾病的平衡的染色体重排(倒位、易位)。本文中我们证明 4C 技术在允许直接对断点进行测序的解析度(~7千碱基)检测平衡的倒位和易位,以及不平衡的易位和缺失。4C 技术用于表征涉及先天畸形和白血病的重排。在 T 细胞急性成淋巴细胞白血病(T-ALL)中 *LMO3* 基因被鉴定为 T 细胞受体 β 基因(*TCRB*)的新的易位配体。这些结果确立 4C 技术为基因组重排的精确分析的强有力的新临床研究工具,对诊断疾病、预测和最终最优的患者护理很重要。

介绍

染色体重排(缺失、扩增、倒位、易位)可能是疾病的起因,特别是当它们由于下列原因影响基因表达时:获得或丢失基因、形成融合基因或转录调

控DNA元件的重置。种系中出现的重排会增加先天缺陷，体细胞组织中的那些会导致肿瘤形成。人类基因组测序计划完成之后，表征人类基因组中的结构变化已经成为了主要的任务，因为已经越来越清楚基因组多样性在人群中自然发生并与对疾病的易感性相关(1-6)。

基于微阵列的比较基因组杂交(阵列-CGH)是广泛使用的能够在几千碱基或甚至更小解析度下检测染色体扩增或缺失的高通量基因组方法。CGH依赖于测量DNA拷贝数的变化，例如在缺失或扩增中可见的，因而不足以鉴定没有DNA内容缺失或获得时发生的易位和倒位。此类平衡时间多频繁地发生是未知的但它们预计构成多至20%的所有结构变化(7)。目前它们的检测主要依靠例如染色体核型分析的细胞遗传学方法，其具有缺点：它们遗漏事件(约20%)且提供的解析度有限(最大5-10兆碱基)。这使得鉴定涉及疾病的实际基因异常的其它劳动密集型的分析成为必需。这里，我们证明芯片上的染色质构象捕获(4C)技术(8)在允许直接克隆并测序染色体断点的解析度(~7kb，见下)鉴定平衡的和不平衡的基因组重排。重要的是，此高解析度策略要求使用单个微阵列以筛选整个基因组因而是成本高效的。

我们最初的4C结果显示，不论染色质的折叠和与给定基因座的长距离相互作用，线性染色体模板上靠近的DNA片段总是最高效地被捕获，导致靶序列周围至少5百万-1千万碱基范围内探针上强烈而且经常甚至饱和的杂交信号。这与灵活的染色质纤维上位置靠近的DNA片段将更频繁地相互作用的想法相吻合(9)。本地片段也比频繁地向靶序列卷曲的百万碱基之外的片段更频繁地被捕获(8)。限制性片段的随机捕获是少见的，如位于不相关的染色体上探针的非常低频的信号所证明的那样。因此，4C技术允许重建靶序列周围染色体模板的物理图，并且因此它也应当能够鉴定作为基因组重排结果的这些图中的变化。

材料与amp;方法

样品制备

如前所述(Vlierberghe等, *Leukemia* **20**, 1245(Jul, 2006); Simonis等, *Nat Genet* **38**, 1348(Nov, 2006))处理T-ALL患者样品与健康的对照T细胞样品。对源自PAP患者的EBV-转化细胞系进行培养并如前所述进行处理(Simonis等, *Nat Genet* **38**, 1348(Nov, 2006); Galjaard等, *Am J Med Genet A* **121**, 168(Aug 30,

2003)。

4C阵列设计

用前述的标准例如仅选择独特的DNA序列(Simonis等, *Nat Genet* **38**, 1348(Nov, 2006)), 在距离*HindIII*位点100bp之内设计了60bp的探针。为了能够用装载在Nimblegen微阵列的400,000个探针覆盖整个基因组而进行了选择。首先通过为每个*HindIII*片段仅保留一个而不是每侧一个探针减少了探针的数量。第二, 对探针进行了选择以使探针的间隔在整个基因组中尽可能相等。

4C分析

4C分析如前述(2)进行, 使用下列引物

序列:

*TCRB*的5'端

CATGAAGAAACGAGCACCC CCTTGATGTTTCTCCCTTTACC

*TCRB*的3'端

TGTCAGGCTCTTCTCCTACAC GTCGTCCAGAACACTCACC

着丝粒t(4;7)AATCCAGGGCTACTTCCAG CCGTGATGCTATCTGCCA

端粒t(4;7)TGTTGGAAGACCAGGTGAAG TGTCGTGGAAAGCGAGTG

缺失9

CAATCCCAGATACATTCCTCATACAAATACTTTCCAAGACTGGAC

*TCRA*的3' GAATATGTTATGCTTGATCC TTCCATGAGAGAAGTCTAG

用SignalMap软件将4C数据可视化。为了生成4C数据的整个染色体视图, 用R软件包(<http://www.r-project.org>)计算29个探针窗口大小的连续平均。

限制性-片段-配对-末端-测序

先将10 µg基因组DNA在500µl中用10U的识别6碱基的酶(*HindIII*、*BglII*或*EcoRI*)消化(37 °C, 2小时)。样品通过酚-氯仿提取和乙醇沉淀进行纯化。然后, 样品在2ml中用40U的连接酶(Roche)在16 °C连接4小时, 然后是20 °C 30分钟。

连接样品通过酚-氯仿提取和乙醇沉淀进行纯化。

在与为识别6碱基的酶描述的条件相同条件下，用识别4碱基的限制酶(如NlaIII或DpnII)进行第二次消化。随后的连接也如上所述。样品通过酚-氯仿提取和乙醇沉淀进行纯化。从50-100ng的DNA中用如下条件进行PCR扩增所选的片段：94 °C 3分钟，随后是94 °C 15秒、55 °C 1分钟和72 °C 2分钟的30个循环和72 °C 7分钟的一个最终步骤。

结果

鉴定易位断点

为了测试此项，4C技术首先应用到HSB-2 T-ALL细胞系，包括在7q35的T细胞受体 β 基因座(*TCRB*)与1p35的*LCK*基因座之间的相互易位t(1;7)(p35;q35)(10)。进行了两个独立的4C实验，每个实验分析与染色体7上位于*TCRB*基因座中断点任一侧并分别相距462kb和239kb的不同限制性片段的相互作用。在两种情况下，在健康的对照样品和HSB-2样品中观察到了染色体7上*TCRB*基因座周围的强烈杂交信号。对照样品在所有其它染色体上显示没有信号或显示背景信号(图30A；图33)。相反，HSB-2样品显示额外的非常强烈的信号，特别是在1p35上的百万碱基区域内(图30B)。

这些信号代表被HSB-2中染色体7上的两个片段捕获的染色体1上的限制性片段，显示这些染色体的部分在物理上接近。染色体7上最端粒一侧的*TCRB*靶序列捕获了染色体1上*LCK*基因朝着着丝粒一侧的限制性片段。相反地，最着丝粒一侧的*TCRB*靶序列捕获了*LCK*朝向端粒一侧的片段。这与t(1;7)易位的方向相吻合。而且，两个实验中染色体1捕获的第一限制性片段直接位于前面鉴定的染色体断点侧翼。因此，4C将易位断点定位于代表从非捕获到捕获的限制性片段的探针对之间的位置。在这种情况下，分析与染色体7靶序列相互作用的两个4C实验每个都将涉及平衡的t(1;7)易位的染色体1上的断点鉴定于27kb之内。

鉴定倒位

接下来，我们测试了4C是否也能够鉴定倒位。我们对小儿(pediatric) T-ALL患者样品应用了4C技术，基于FISH和微阵列表达研究疑似该患者在染色体7上携带倒位inv(7)(p15q35)。这一异常导致*TCRB*基因座重排入*HOXA*基因簇，如前面为其它患者所描述的那样(11, 12)。也用上述鉴定t(1;7)易位的

相同 $TCRB$ 靶片段组进行两个实验。两个靶序列高效地捕获仅4个患者样品中染色体7另一侧 $HOXA$ 基因座周围覆盖百万碱基的许多片段(图30 C-D)。而且,每个靶序列捕获了 $HOXA$ 基因簇周围不同的染色体区域,指示两个片段与染色体7的不同片段连接。最3'的 $TCRB$ 靶片段捕获最3'的 $HOXA$ 片段,而最5'的 $TCRB$ 靶序列捕获最5'的 $HOXA$ 片段,因而揭示基因座之间的倒位。 $HOXA$ 周围的两个捕获的区域直接位于彼此侧翼,显示倒位是平衡的且并不伴有(大范围的) $HOXA$ 序列的丢失。标示非捕获的和捕获的片段之间转换的两个探针揭示断点的位置,其位于 $HOXA$ 簇的 $HOXA9$ 基因附近的6kb区域(图30D)。对这一区域确实携带断点的确证是用限制性-片段-配对-末端测序方法经过测序得到的(图34)。因此,我们得出结论,4C技术是第一个能够检测平衡易位和倒位的高通量基因组方法。4C技术提供的解析度允许立刻克隆断点。4C技术因此是第一个能够以如此高解析度检测平衡的遗传事件的技术。

鉴定非平衡的易位

4C技术的潜力通过将其应用于源自罹患轴后性多指症(Postaxial Polydactyly(PAP))的患者的EBV转化细胞系得到进一步的探索。PAP是通过腓骨指上的多余尺骨表征的常染色体显性遗传疾病。以前通过核型分析和FISH表征患者细胞包含带有小缺失的染色体4和7之间的非平衡易位 $t(4;7)(p15.2;q35)$ 。但是,FISH的有限解析度阻碍了缺失范围和断点确切位置的确定(13)。进行两个4C实验,每个都分析与位于染色体7重排部分另一侧的靶片段的DNA相互作用,其中一个距最近的断点4百万碱基(见下)。在两个实验中,基因组片段不仅在染色体7上也在染色体4的4p15.2上捕获(图31A;图35)。与平衡易位找到的不同,两个靶序列捕获的染色体4片段不直接彼此相邻。一个断点位于17.28 Mb位置(NCBI 36)而另一个在染色体4的20.08Mb位置找到,显示 $t(4;7)$ 易位伴有染色体4上的2.8Mb缺失。为了确认从捕获的到非捕获的限制性片段的转换确实标记了断点的位置,将4C数据中最不明显的位于20.08Mb的断点克隆并测序(图31B)。这确证了断点的位置在染色体4上20.08Mb处 $SLIT2$ 基因内部并揭示它与染色体7上、与用于鉴定断点的靶序列相距4Mb的基因间序列重组。这说明即使当断点相距几百万碱基时,4C靶序列也能够捕获DNA片段并鉴定重排。当4C分析定位于基因组断点的两侧时,它能够立刻鉴定易位或倒位是否伴有额外的例如缺失(即不平衡的)

的重组。

鉴定缺失

接下来我们研究了4C技术是否可鉴定不与易位关联的缺失。为此，我们分析了另一个小儿T-ALL患者的样品，基于阵列-CGH数据其包含5染色体9p21上*p15/p16*基因座的同源缺失。缺失的确切大小和实际断点未知。我们确定了位于距离估计的断点之一~2 Mb远的靶片段。如希望的那样，观察到了缺少探针信号的区域，界定了缺失区域(图31C)。与健康的对照样品相比，在患者样品中观察到了在紧邻缺失的下游区域杂交信号的增加。这是因为由于缺失，该区域在线性模板上与靶片段更加接近。基于4C数据，设计了~2Mb缺失区域侧翼的PCR引物，其允许扩增越过断点；PCR产物的测序确认了位于缺失侧翼的两个断点的位置(图31D)。我们得出结论4C技术能够鉴定纯合缺失。缺失揭示其自身为包含降低的杂交信号的区域连同显示增加的杂交信号的更多下游序列。

未表征的患者样品

将4C技术用于筛选未表征的T-ALL患者样品中与*TCRB*基因座或T细胞受体 α (*TCRA*)基因座相关的遗传重组。在T-ALL中，染色体易位在*TCR*基因座尝试VDJ重组时频繁发生。将以前显示没有携带任何与T-ALL相关的反复出现的遗传异常(数据未显示)的来自五个T-ALL患者的样品用靠近*TCRB*的靶序列和靠近*TCRA*的靶序列通过4C技术进行了分析。没有一个样品显示了与*TCRA*基因座的重排(数据未显示)，五个患者样品中的四个也显示没有与*TCRB*重排(图36)，其随后被FISH确认。但是一个患者样品显示了*TCRB*和染色体12的p臂间的易位(图32A；图36)。除此之外，也发现该患者携带染色体12上的大的缺失，其被寡阵列CGH实验所确认(数据未显示)。这一缺失位于距离易位断点~3 Mb处，也显示4C靶序列能够在大距离上鉴定重排。易位t(7;12)(q35;p12.3)以前未在T-ALL中描述过。标示捕获的和未捕获的限制性片段之间转换的染色体12上的两个探针相距6kb并正好位于Lim-domainonly基因*LMO3*的下游。限制性-片段-配对-末端-测序被用来确证这些探针界定了包含断点的区域。出现在两个衍生染色体上的断点的测序证明染色体12没有丢失一个核苷酸而进行了重排，而染色体7中的断裂伴随有近400 kb的*TCRB*

序列的缺失(图32B)。这一缺失可能源于与*TCRB*基因座的尝试性VDJ重组相关的缺失事件。两个断点也包含未知来源的干扰碱基对(相应地为4和18 bp),其可能代表通常也在VDJ重组过程中插入的随机核苷酸。有趣的是,易位将*TCRB*的增强子置于*LMO-3*基因的70 kb下游(图32C),这与其相对于*TCRB*的正常位置相似。微阵列表达数据显示,尽管*LMO3*在T-ALL患者样品中通常关闭,而在这个T-ALL样品中高度表达(图37)。以前已经发现,蛋白质家族成员*LMO-1*和*LMO-2*,而非*LMO3*是T-ALL中*TCR*基因座的致癌基因易位配体。有趣的是,最近发现*LMO3*在成神经细胞瘤中发挥致癌基因的作用(14)。因此,4C技术应用于筛选未表征患者样品中的遗传异常导致发现以前未检测到的易位并将*LMO-3*确立为推定的新的T细胞致癌基因。

结论

这些数据将4C技术确立为第一个能够鉴定平衡的遗传异常例如相互易位和倒位的基因组方法。此外,很清楚,4C技术能够鉴定纯合缺失和与易位相关的缺失。4C技术也可以基于捕获DNA拷贝数的变化鉴定杂合的不平衡的事件。4C技术相对于配对-末端测序方法(15)或甚至是全基因组(高通量)测序的主要优点是对平衡重排的鉴定不依赖于捕获携带断点的单个序列片段;取而代之,4C技术基于捕获覆盖经过断点的几百万碱基的许多片段来鉴定平衡的重排。例如,相互易位是基于通过位于另一染色体上的靶序列捕获许多位于一个染色体部分的许多片段来鉴定的。这使该技术比其它方法更加强大。4C技术的解析度高并允许直接克隆和测序遗传断点,如对四个不同重排所证明的那样。甚至能够轻松地鉴定距离几百万碱基远的断点。解析度基本上与限制酶形成片段的平均大小相等。形成更小片段的酶的使用甚至将进一步提高此技术的解析度。只要能够分离(或培养)含有完整DNA的足够的细胞,4C技术可用于所有种类的细胞类型,例如血癌细胞、实体瘤、为产前诊断收集的羊膜细胞等。我们现在从约1千万个细胞开始,但是如果我们杂交源自~50万基因组当量的PCR扩增的连接接头,这一数量能被降低。现在4C技术要求选择靶序列因而特别适合筛选靠近经常涉及疾病(例如T-ALL中的T细胞受体基因或人淋巴瘤中B细胞受体(*BCR*)重和轻链基因座)的基因座的重排。本技术对于未充分表征的重排的精细作图很有用,例如基于染色体核型分析找到的易位或倒位。

实施例 13

多重 4C-测序

实验设计

本实施例基于几种高通量测序技术中的一种(Solexa),但也可修改以适应其它平台。为了对 4C 结果的分析使用 Solexa(Illumina)测序,测序针对连接接头。因此,对每个靶序列设计反向 PCR 引物以使它们靠近所分析的第一和第二限制酶识别位点。这里我们设计了短反向 PCR 引物(18 聚体),每个都与所分析的第一(*Hind*III)和第二(*Dpn*II)限制酶识别位点部分或完全地重叠。将 Solexa 衔接子和测序引导序列作为 5'突出端附加到反向 PCR 引物上(图 38)。

分别在人染色体 7 上的 85、105 和 139 Mb 处选择三个靶序列(诱饵)。用于扩增这些诱饵中的每个捕获的片段的引物组如下:

85 Mb:

*Dpn*II 引物: atgtgactcctctagatc

*Dpn*II 引物与衔接子:

aatgatacggcgaccaccgaacactctttccctacacgacgctcttccgatct-atgtgactcctctagatc

*Hind*III 引物: ccctgaacctcttgaagct

*Hind*III 引物与衔接子: caagcagaagacggcatcacga-ccctgaacctcttgaagct

105 Mb:

*Dpn*II 引物: cggcctccaattgtgatc

*Dpn*II 引物与衔接子:

aatgatacggcgaccaccgaacactctttccctacacgacgctcttccgatct-cggcctccaattgtgatc

*Hind*III 引物: gaattgcttttggtgaagctt

*Hind*III 引物与衔接子: caagcagaagacggcatcacga-gaattgcttttggtgaagctt

139 Mb:

*Dpn*II 引物: ttttagccctgacagatc

*Dpn*II 引物与衔接子:

aatgatacggcgaccaccgaacactctttccctacacgacgctcttccgatct-ttttagccctgacagatc

*Hind*III 引物: agtcaaacataagcctaagc

*Hind*III 引物与衔接子: caagcagaagacggcatacga-agtcaaacataagcctaagc

每个引物组(与接头)用于单独的 PCR 反应; 在标准条件下(在 Simonis 等, *Nature Methods* 2007, vol.4, 895-901 中描述)进行了 3 个 PCR 反应(每个反应使用 200 ng 模板)。4C 模板从 HSB-2 T-ALL 细胞系中制备, 包括 7q35 上 T 细胞受体- β (TCRB)基因座与 1p35 上 LCK 基因座之间的相互易位 t(1;7)(p35;q35)(Burnett, R. C., 等, *Blood* 84, 1232-6(1994))。图 39 显示了 PCR 产物。

随后, 将每个引物组的 PCR 反应合并并在 Amersham 清洗柱上纯化。测定 DNA 浓度并将相等量的每个引物组的扩增物质混合。用 Solexa 测序分析该混合物。

Solexa 测序结果

读数的总数(一列): 4.9×10^6 个序列。

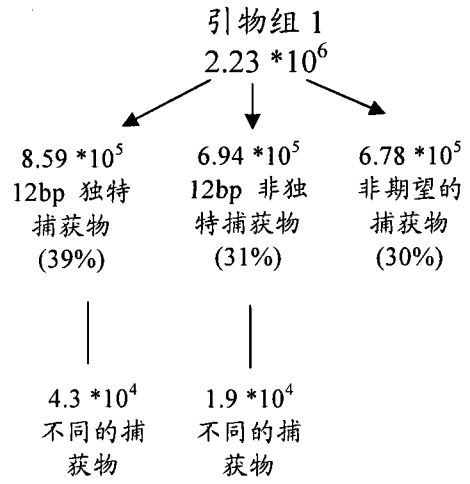
93%的序列起始于希望的引物序列之一。

当分析经过 DpnII 连接接头的前 12 个碱基对(bp)并将这 12 个碱基对与包含基因组中相关的 DpnII 位点(即那些与 *Hind*III 位点直接相邻的)侧翼的所有 12 bp 基因组片段的本地数据库比较时我们发现: 37%包含独特的 12bp 捕获物(catch), 35%包含非独特的 12bp 捕获物, 而 29%不包含期望的 12bp 捕获物。

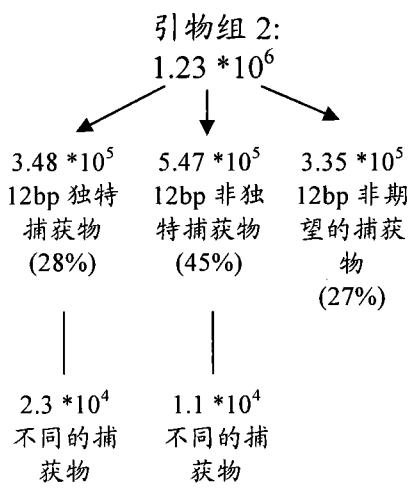
在所有 4.9×10^6 个序列中: 35%包含引物+独特捕获物, 32%包含引物+非独特捕获物。因此总共 66%包含期望的前 30bp 序列。

每引物组的序列的总数和这些序列的性质如下:

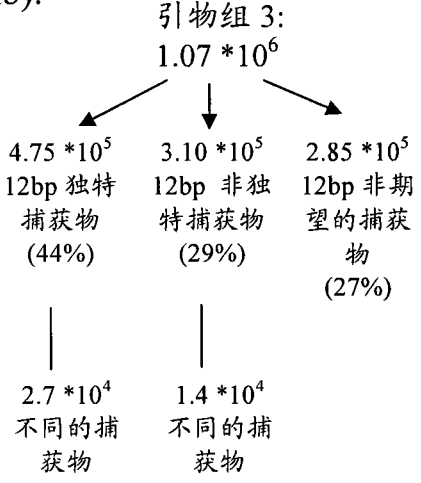
引物组 1(85 Mb):



引物组 2(105 Mb):



引物组 3(139 Mb):



当捕获的序列绘制在它们的染色体位置上时,数据显示在图 40 和 41 中。

简而言之,数据对多重 4C 测序提供了原理验证。对于每个序列,都可以鉴定诱饵和捕获的序列。本技术可进一步改进:

1. 测序可以更好地定位于 *HindIII* 侧(而不是 *DpnII* 侧);这防止读取在第二连接步骤过程中发生的随机连接事件。
2. 使用不频繁的切割酶(7-或 8-切割酶而不是 6-切割酶)将增加由每个诱饵捕获(覆盖)的基因组距离。
3. 使用不消化重复 DNA(或在重复 DNA 中相对而言不具有代表性)的限制酶将增加可分析的读取的数量。

其它方面 1

1. 分析靶核苷酸序列与一个或多个核苷酸序列(如一个或多个基因组基因座)相互作用的频率的方法,其包括使用如本文所述的核苷酸序列或探针阵列或探针组或阵列。
2. 鉴定一种或多种指示特定疾病状态或携带者状态的 DNA-DNA 相互作用的方法,其包括使用如本文所述的核苷酸序列或探针阵列或探针组或阵列。
3. 诊断或预测由 DNA-DNA 改变造成的或与 DNA-DNA 改变相关的疾病或综合征或携带者状态的方法,其包括使用如本文所述的核苷酸序列或探针阵列或探针组或阵列。
4. 鉴定一种或多种调节 DNA-DNA 相互作用的试剂的测定方法,其包括使用如本文所述的核苷酸序列或探针阵列或探针组或阵列。
5. 检测断点(如易位)位置的方法,其包括使用如本文所述的核苷酸序列或探针阵列或探针组或阵列。
6. 检测倒位的位置的方法,其包括使用如本文所述的核苷酸序列或探针阵列或探针组或阵列。
7. 检测缺失的位置的方法,其包括使用如本文所述的核苷酸序列或探针阵列或探针组或阵列。
8. 检测重复的位置的方法,其包括使用如本文所述的核苷酸序列或探针阵列或探针组或阵列。

9.分析靶核苷酸序列与一个或多个目的核苷酸序列(如一个或多个基因组基因座)相互作用的频率的方法,其包括以下步骤:

- (a)提供交联DNA的样品;
- (b)用第一限制酶消化交联的DNA;
- (c)连接交联的核苷酸序列;
- (d)解除交联;
- (e)任选用第二限制酶消化核苷酸序列;
- (f)任选连接核苷酸序列;

(g)利用至少两个寡核苷酸引物扩增一个或多个与靶核苷酸序列相连的目的核苷酸序列,其中每个引物与在目的核苷酸序列的侧翼的已知的DNA序列杂交;

(h)将被扩增的一个或多个序列与阵列杂交或对扩增的序列测序;和

(i)确定DNA序列间相互作用的频率。

其它方面 2

本发明又进一步的方面在以下编号的段落中提出。

1. 环化的核苷酸序列,其包含由第一限制酶识别位点分隔的第一和第二核苷酸序列,其中所述第一核苷酸序列是靶核苷酸序列而且所述第二核苷酸序列可通过交联基因组DNA而获得。

2. 段落1所述的环化的核苷酸序列,其中靶核苷酸序列选自由启动子、增强子、沉默基因、隔离子、基质附着区、基因座控制区、转录单位、复制起点、重组热点、易位断点、着丝粒、端粒、基因密集区、基因稀少区、重复元件和(病毒)整合位点所组成的组。

3. 段落1所述的环化的核苷酸序列,其中靶核苷酸序列是与疾病相关的或造成疾病的核苷酸序列,或在线性DNA模板上位于距离与疾病相关的或造成疾病的基因座少于15Mb处。

4. 段落1-3之任一段所述的环化的核苷酸序列,其中靶核苷酸序列选自以下组成的组: *AML1*, *MLL*, *MYC*, *BCL*, *BCR*, *ABL1*, *IGH*, *LYL1*, *TAL1*, *TAL2*, *LMO2*, *TCR α / δ* , *TCR β* 和 *HOX* 或其它与疾病相关的基因座,所述其它与疾病相关的基因座如“Catalogue of Unbalanced Chromosome Aberrations in Man”第2版. Albert Schinzel. 柏林: Walter de Gruyter, 2001. ISBN

3-11-011607-3 中所述。

5. 段落 1-4 之任一段所述的环化的核苷酸序列, 其中第一限制酶识别位点是 6-8 bp 识别位点, 优选选自 *Bgl*II、*Hind*III、*Eco*RI、*Bam*HI、*Spe*I、*Pst*II 和 *Nde*I 所组成的组。

6. 前述段落之任一段所述的环化的核苷酸序列, 其中第二限制酶识别位点是 4 或 5 bp 核苷酸序列识别位点。

7. 前述段落之任一段所述的环化的核苷酸序列, 其中第二限制酶识别位点位于距离第一限制性位点大于约 350bp 处。

8. 前述段落之任一段所述的环化的核苷酸序列, 其中核苷酸序列是标记的。

9. 制备环化的核苷酸序列的方法, 其包括以下步骤:

- (a) 提供交联 DNA 的样品;
- (b) 用第一限制酶消化交联的 DNA;
- (c) 为环化连接交联的核苷酸序列;

10. 段落 9 所述的方法, 其中用 PCR 扩增交联的核苷酸序列。

11. 段落 10 所述的方法, 其中用反向 PCR 扩增交联的核苷酸序列。

12. 段落 10 或段落 11 所述的方法, 其中使用 Expand Long Template PCR System (扩展长模板 PCR 系统) (Roche)。

表 2

相互作用	在 4C 中	N	% 重叠	在冷冻 FISH 中	P 值
B-球蛋白 - 染色体 7 73.1 Mb	+	258	7.4	+	P < 0.001
B-球蛋白 - 染色体 7 80.1 Mb(OR)	-	254	3.6	-	
B-球蛋白 - 染色体 7 118.3 Mb	-	255	3.5	-	
B-球蛋白 - 染色体 7 127.9 Mb(Uros)	+	259	6.6	+	P < 0.001
B-球蛋白 - 染色体 7 130.1 Mb	+	413	9.7	+	P < 0.001
B-球蛋白 - 染色体 7 135.0 Mb(OR)	-	261	1.9	-	
B-球蛋白 - D7Mit21	x	258	0.4	-	
染色体 7 80.1 Mb - 染色体 7 135.0 Mb	x	253	5.9	+	P < 0.05
染色体 7 73.1 Mb - 染色体 7 130.1 Mb	x	254	5.5	+	P < 0.05
Rad23A - 染色体 8 21.8 Mb	+	255	5.9	+	P < 0.05
Rad23A - 染色体 8 122.4 Mb	+	261	8	+	P < 0.001

相互作用	在 4C 中	N	% 重叠	在冷冻 FISH 中	P 值
B-球蛋白 - 染色体 7 73.1 Mb	-	256	3.9	-	
B-球蛋白 - 染色体 7 80.1 Mb(OR)	+	256	12.9	+	P < 0.001
B-球蛋白 - 染色体 7 118.3 Mb	-	242	4.1	-	
B-球蛋白 - 染色体 7 130.1 Mb	-	263	3	-	

B-球蛋白 - 染色体 7 135.0 Mb(OR)	+	256	7	+	P < 0.05
B-球蛋白 - D7Mit21		258	6.2	+	P < 0.05
染色体 7 80.1 Mb - 染色体 7 135 Mb		261	5	+	P < 0.1
Rad23A - 染色体 8 21.8 Mb	-	260	3.8	-	
Rad23A - 染色体 8 122.3 Mb	+	258	8.1	+	P < 0.001

参考文献

Blanton J, Gaszner M, Schedl P. 2003. Protein:protein interactions and the pairing of boundary elements in vivo. *Genes Dev* 17:664-75.

Dekker, J., Rippe, K., Dekker, M., 和 Kleckner, N. 2002. Capturing chromosome conformation. *Science* 295: 1306-11.

Drissen R, Palstra RJ, Gillemans N, Splinter E, Grosveld F, Philipsen S, de Laat W. 2004. The active spatial organization of the beta-globin locus requires the transcription factor EKLF. *Genes Dev* 18:2485-90.

Horike S, Cai S, Miyano M, Cheng JF, Kohwi-Shigematsu T. 2005. Loss of silent-chromatin looping and impaired imprinting of DLX5 in Rett syndrome. *Nat Genet* 37:31-40.

Murrell A, Heeson S, Reik W. 2004. Interaction between differentially methylated regions partitions the imprinted genes Igf2 and H19 into parent-specific chromatin loops. *Nat Genet* 36:889-93.

Palstra, R.J., Tolhuis, B., Splinter, E., Nijmeijer, R., Grosveld, F., 和 de Laat, W. 2003. The beta-globin nuclear compartment in development and erythroid differentiation. *Nat Genet* 35: 190-4.

Patrinos, G.P., de Krom, M., de Boer, E., Langeveld, A., Imam, A.M.A, Strouboulis, J., de Laat, W., and Grosveld, F.G.(2004). Multiple interactions between regulatory regions are required to stabilize an active chromatin hub. *Genes & Dev.* 18: 1495-1509.

Spilianakis CG, Flavell RA. 2004. Long-range intrachromosomal interactions in the T helper type 2 cytokine locus. *Nat Immunol* 5:1017-27.

Tolhuis, B., Palstra, R.J., Splinter, E., Grosveld, F.,和 de Laat, W. 2002. Looping and interaction between hypersensitive sites in the active beta-globin locus. *Molecular Cell* 10: 1453-65.

Vakoc CR, Letting DL, Gheldof N, Sawado T, Bender MA, Groudine M,

Weiss MJ, Dekker J, Blobel GA. 2005. Proximity among distant regulatory elements at the beta-globin locus requires GATA-1 and FOG-1. *Mol Cell*. 17:453-62

实施例 12 的参考文献

1. *Nature* **447**, 661(Jun 7, 2007).
2. D. F. Easton等, *Nature* **447**, 1087(Jun 28, 2007).
3. L. Feuk, A. R. Carson, S. W. Scherer, *Nat Rev Genet* **7**, 85(Feb, 2006).
4. A. J. Sharp, Z. Cheng, E. E. Eichler, *Annu Rev Genomics Hum Genet* **7**, 407(2006).
5. A. J. Iafrate 等, *Nat Genet* **36**, 949(Sep, 2004).
6. M. R. Mehan, N. B. Freimer, R. A. Ophoff, *Hum Genomics* **1**, 335(Aug, 2004).
7. E. E. Eichler等, *Nature* **447**, 161(May 10, 2007).
8. M. Simonis等, *Nat Genet* **38**, 1348(Nov, 2006).
9. K. Rippe, *Trends Biochem Sci* **26**, 733(Dec, 2001).
10. R. C. Burnett, M. J. Thirman, J. D. Rowley, M. O. Diaz, *Blood* **84**, 1232(Aug 15, 1994).
11. F. Speleman等, *Leukemia* **19**, 358(Mar, 2005).
12. J. Soulier等, *Blood* **106**, 274(Jul 1, 2005).
13. R. J. Galjaard等, *Am J Med Genet A* **121**, 168(Aug 30, 2003).
14. M. Aoyama等, *Cancer Res* **65**, 4587(Jun 1, 2005).
15. E. Tuzun等, *Nat Genet* **37**, 727(Jul, 2005).

以上说明书所述的所有公开文献都纳入本文参考。在不偏离本发明范围和精神的情况下,所述的本发明方法和系统的各种修改和变型对于所述领域技术人员来说是显而易见的。尽管本发明联系特别优选的实施方案来描述,但是应当理解,本发明所要求保护的内容不应被不正当地局限于这些特定实施方案中。事实上,所述对分子生物学或相关领域技术人员显而易见的执行本发明的模式的各种修改也要在以下权利要求的范围内。

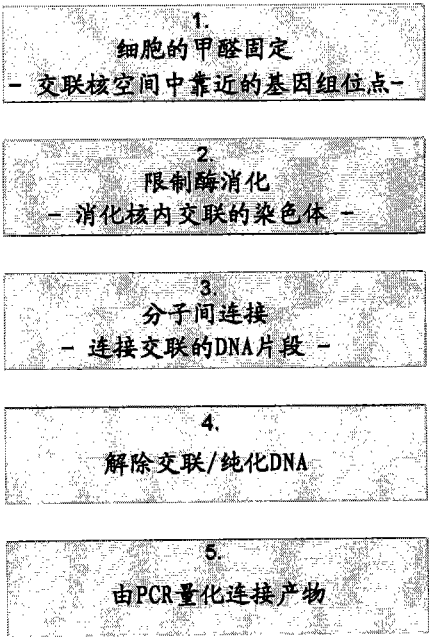


图 1

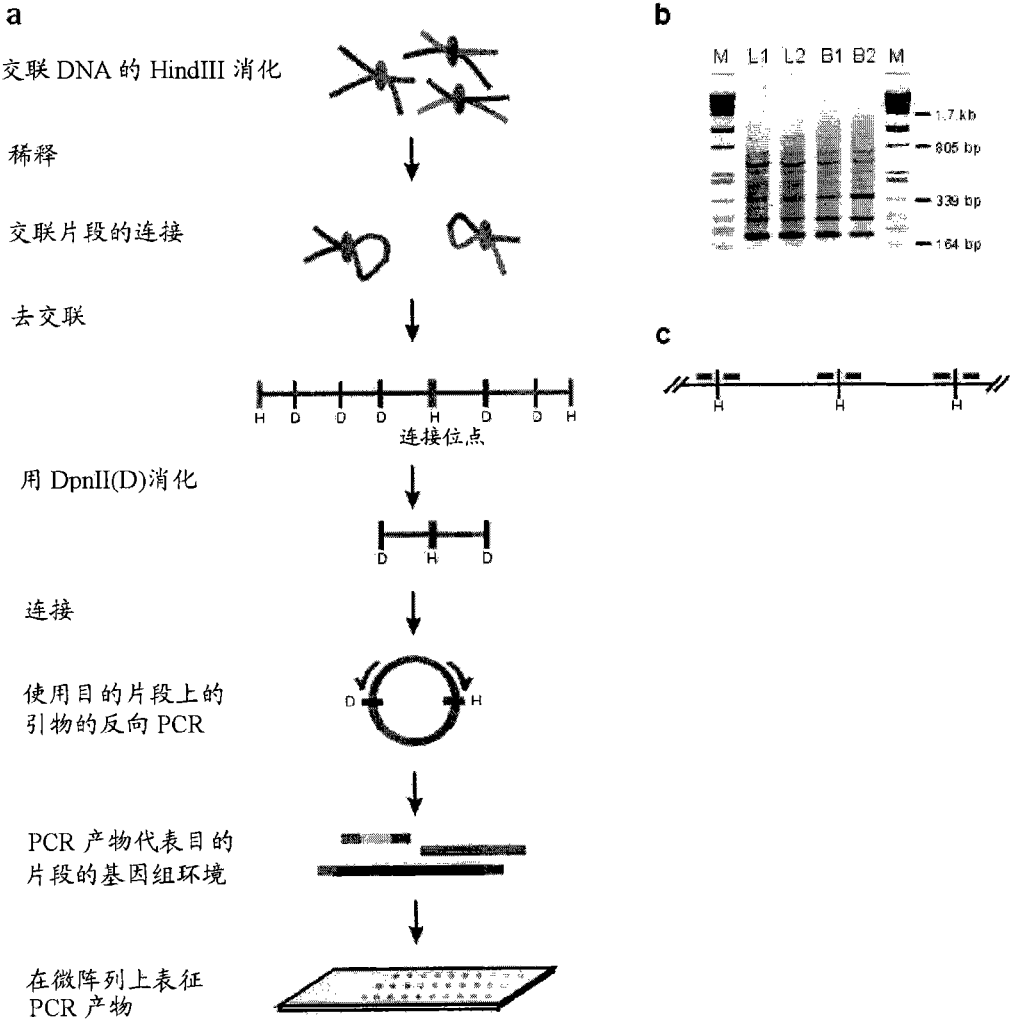


图 2

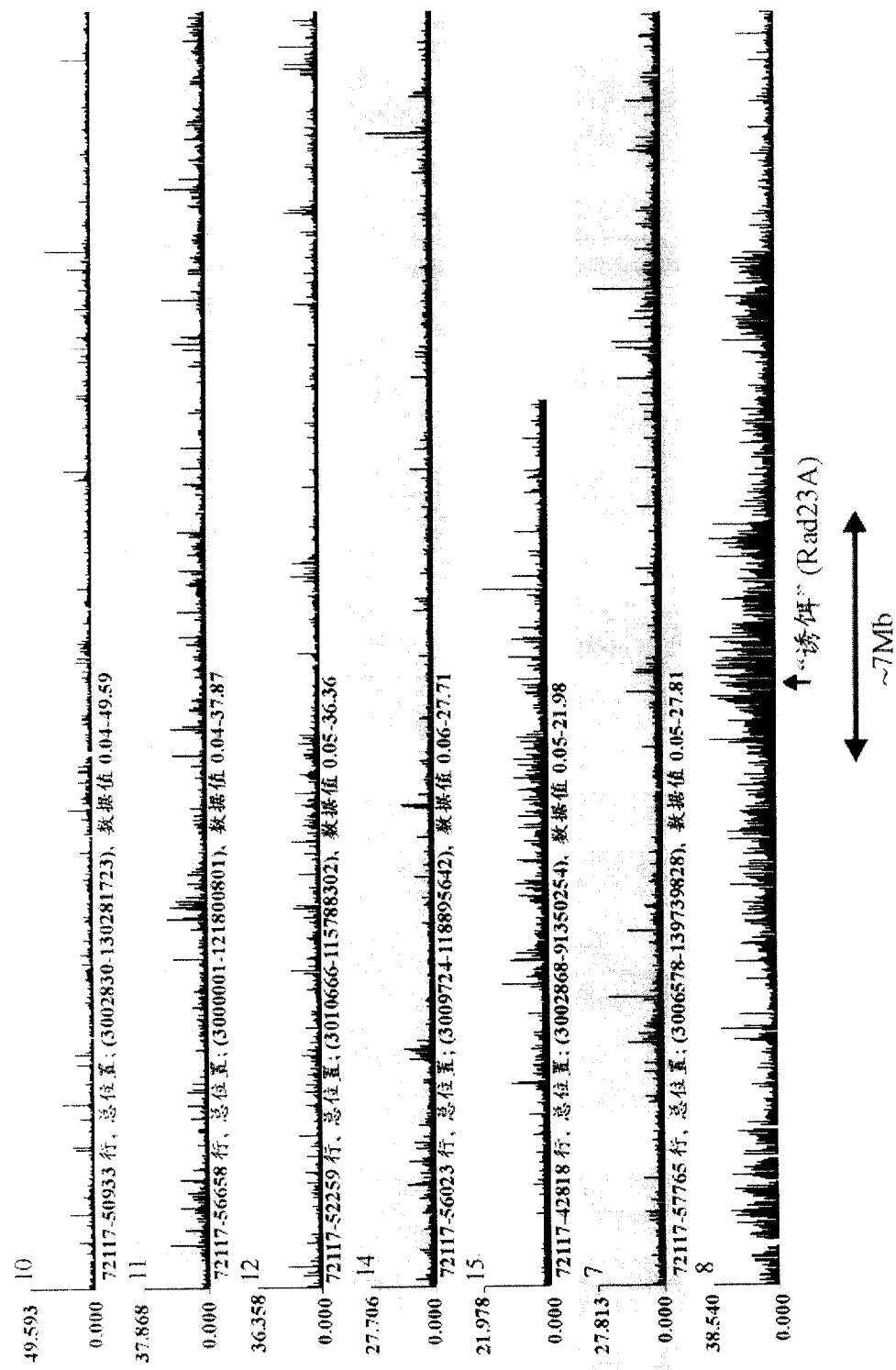


图 3

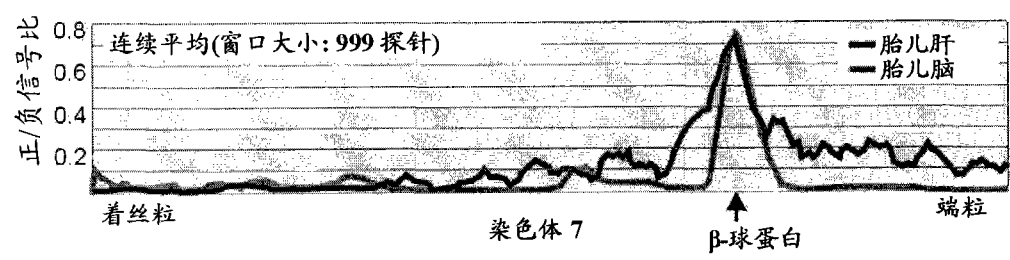


图 4

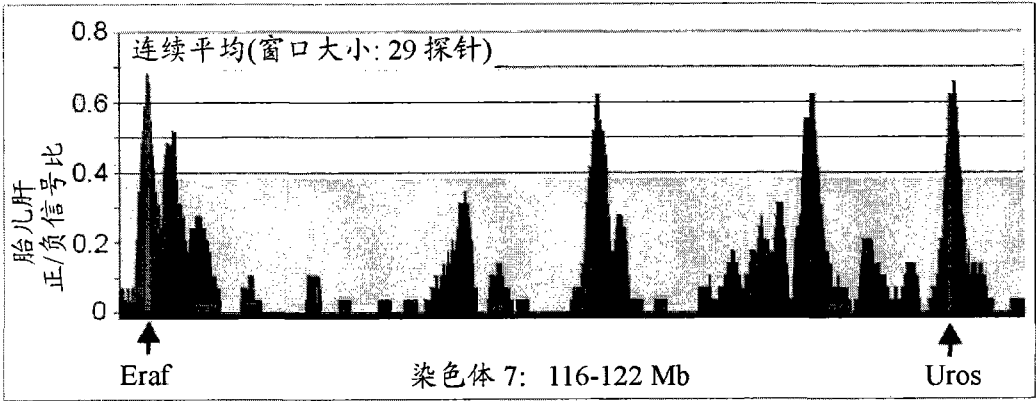


图 5

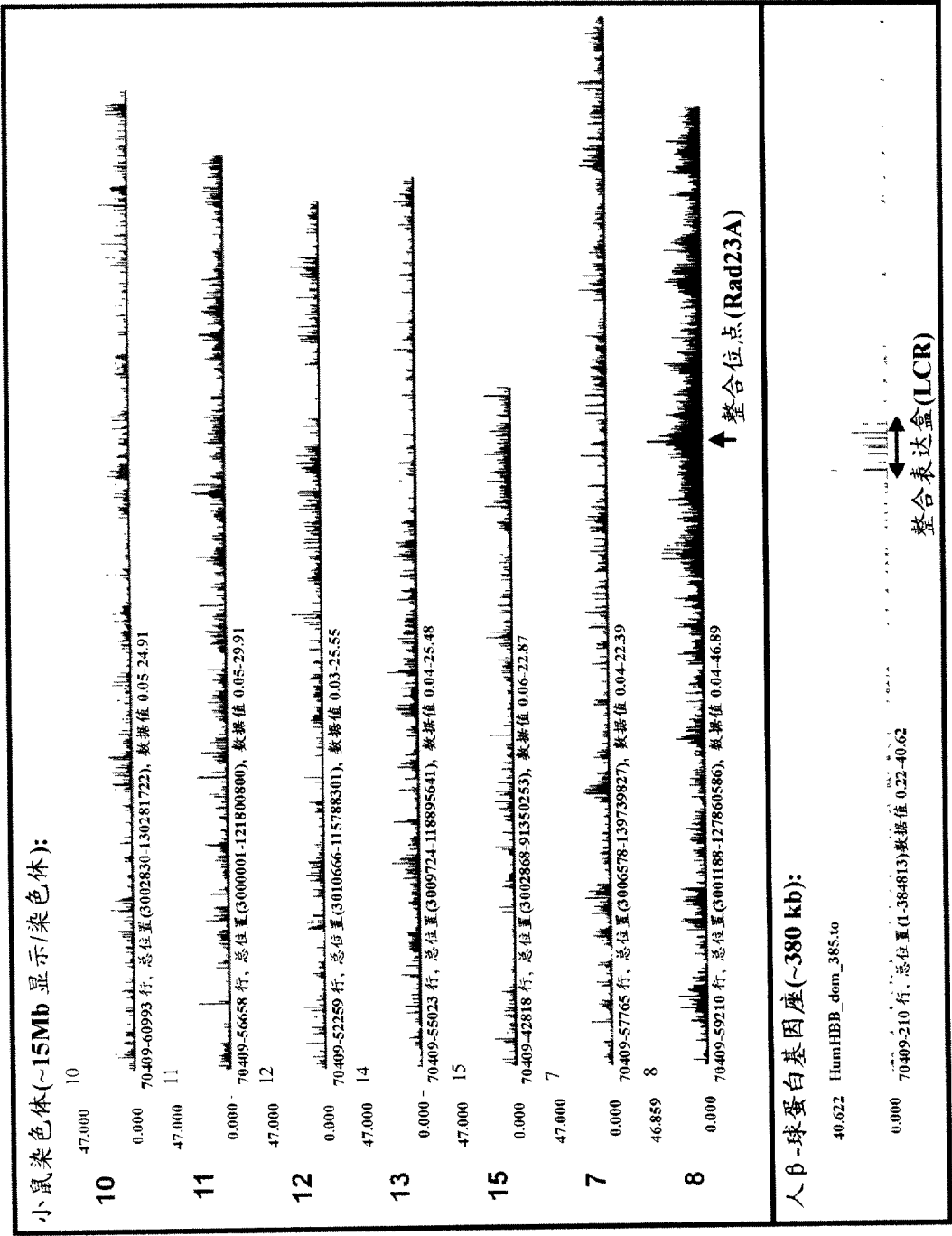


图 6



图 7

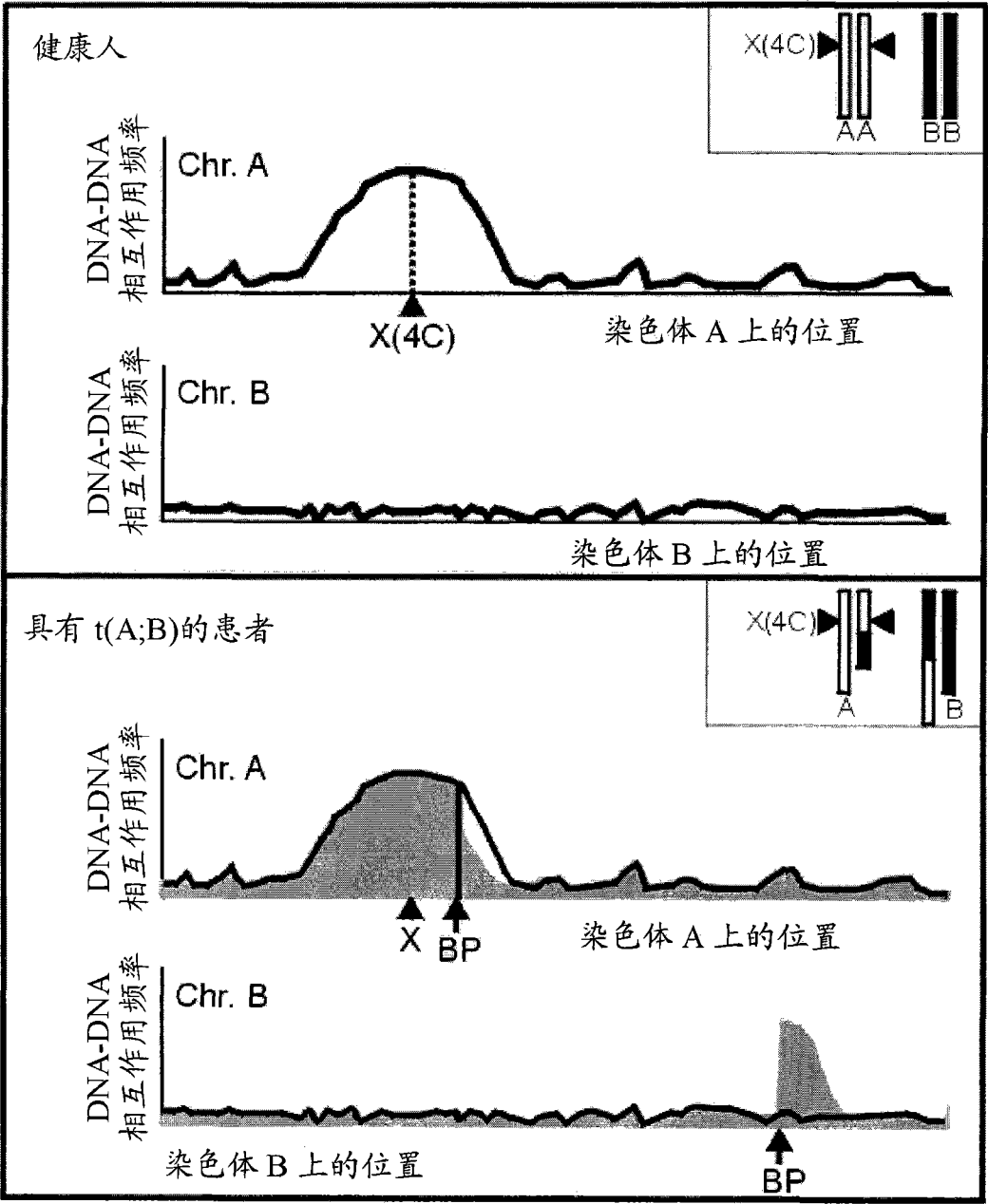


图 8

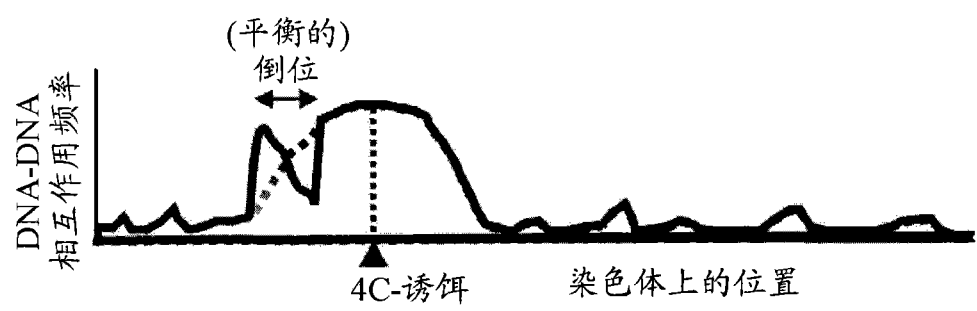


图 9

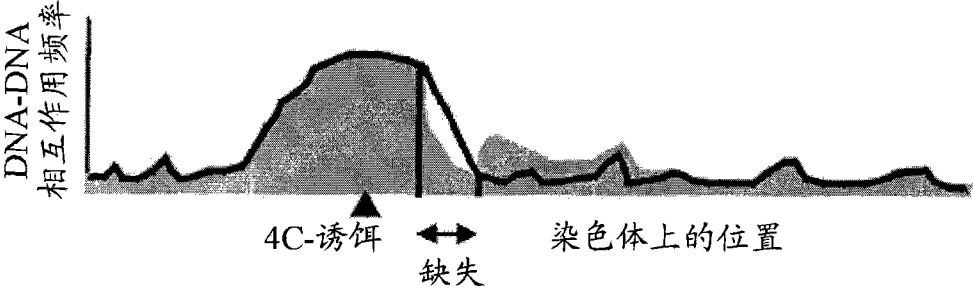


图 10

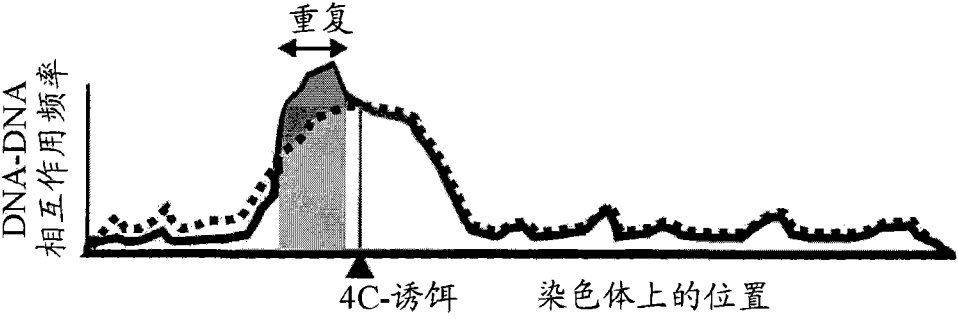


图 11

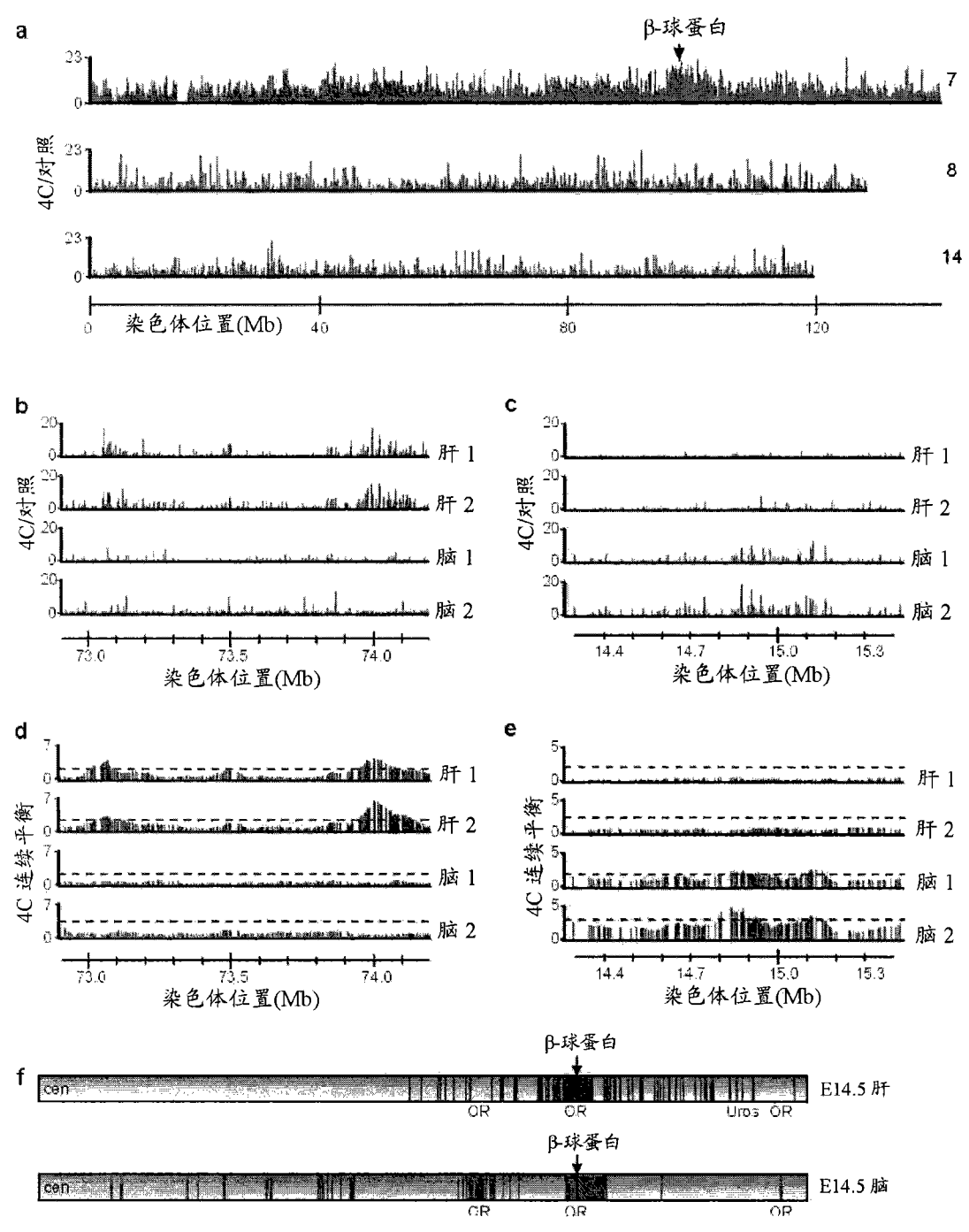


图 12

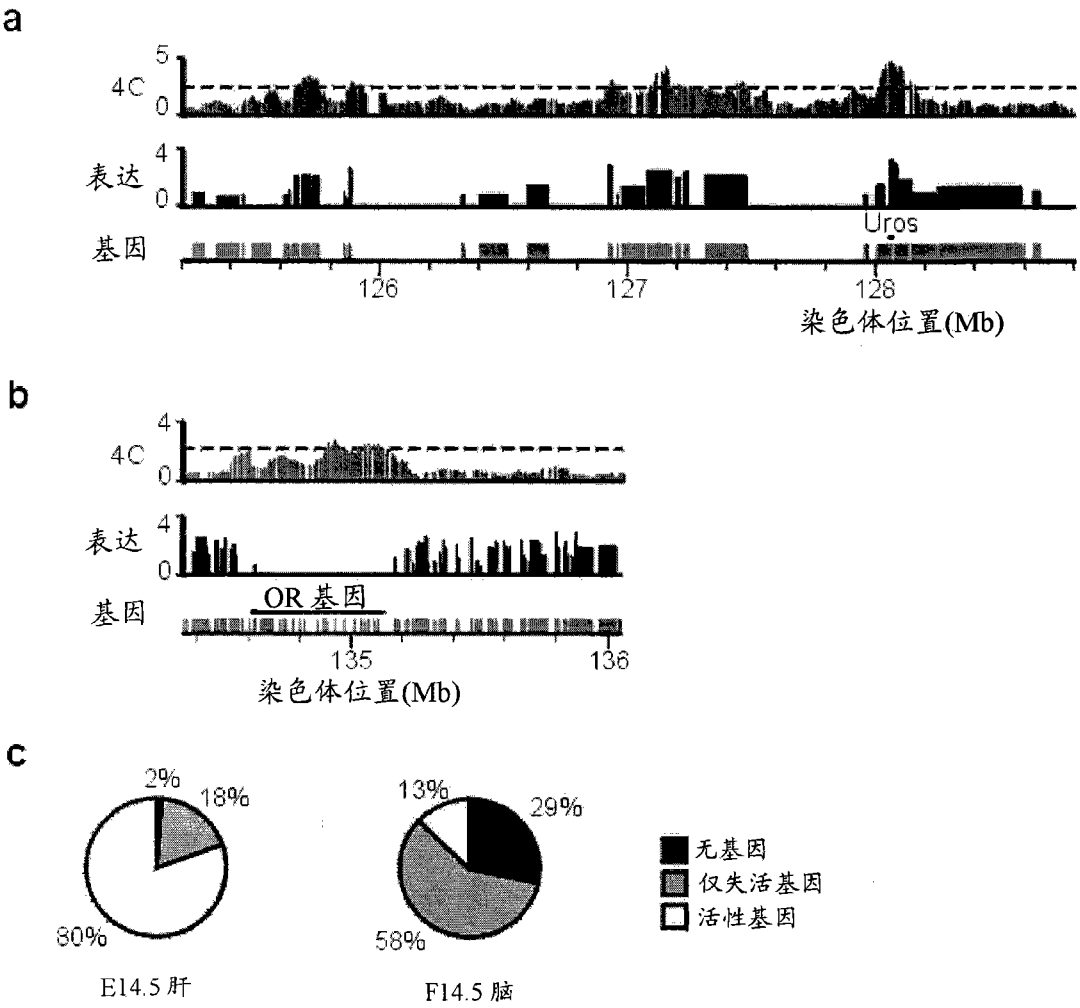


图 13

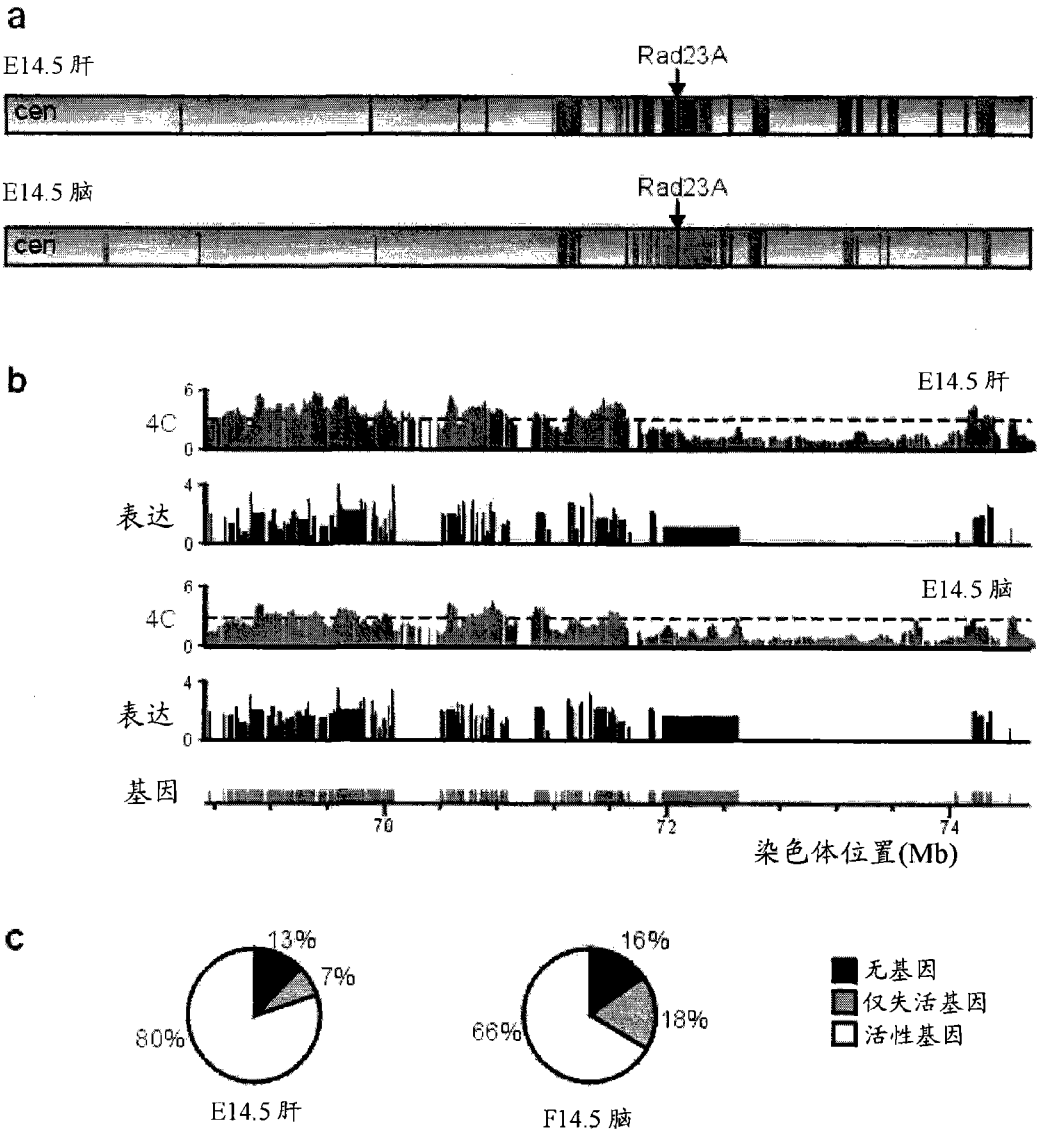


图 14

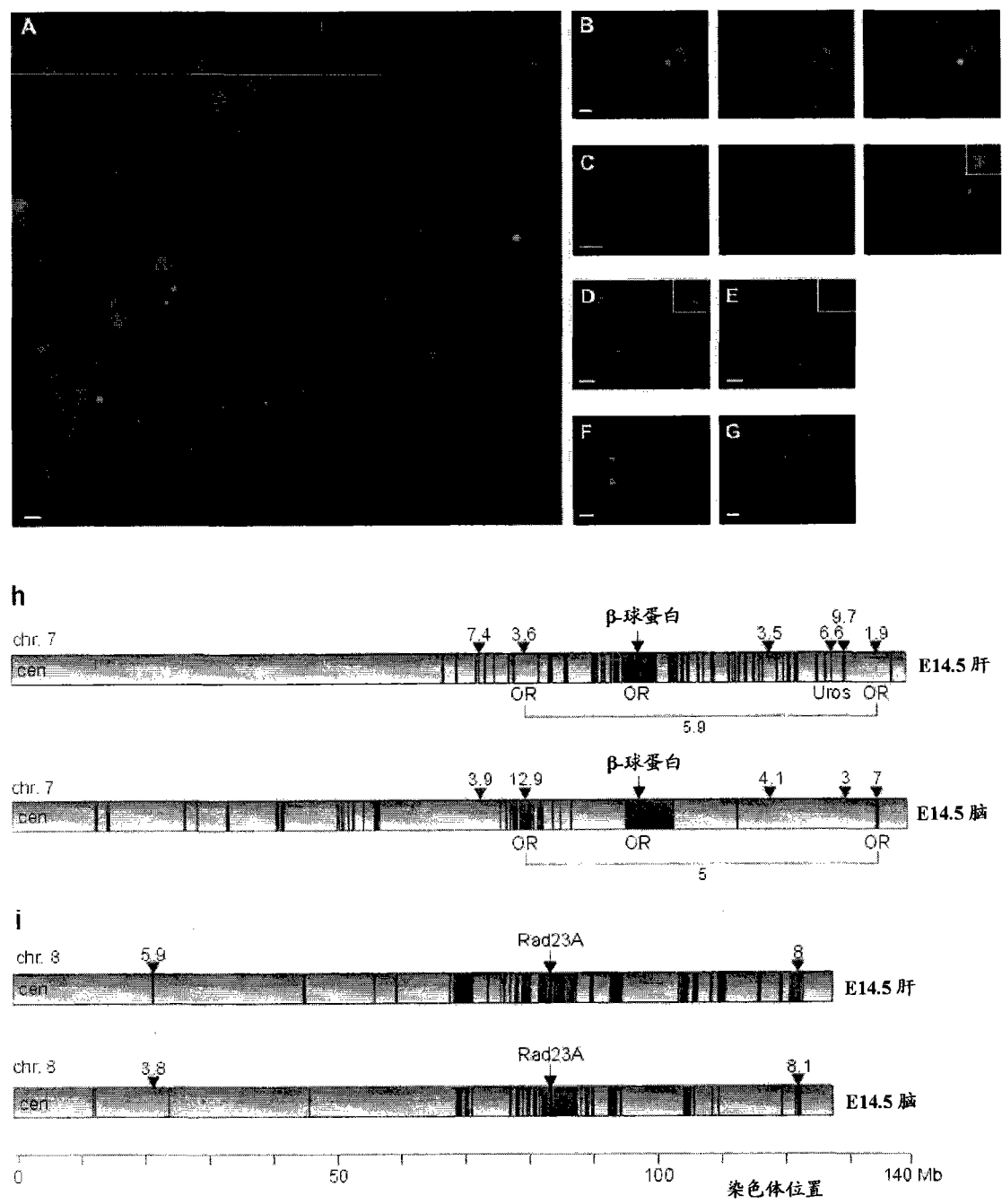


图 15

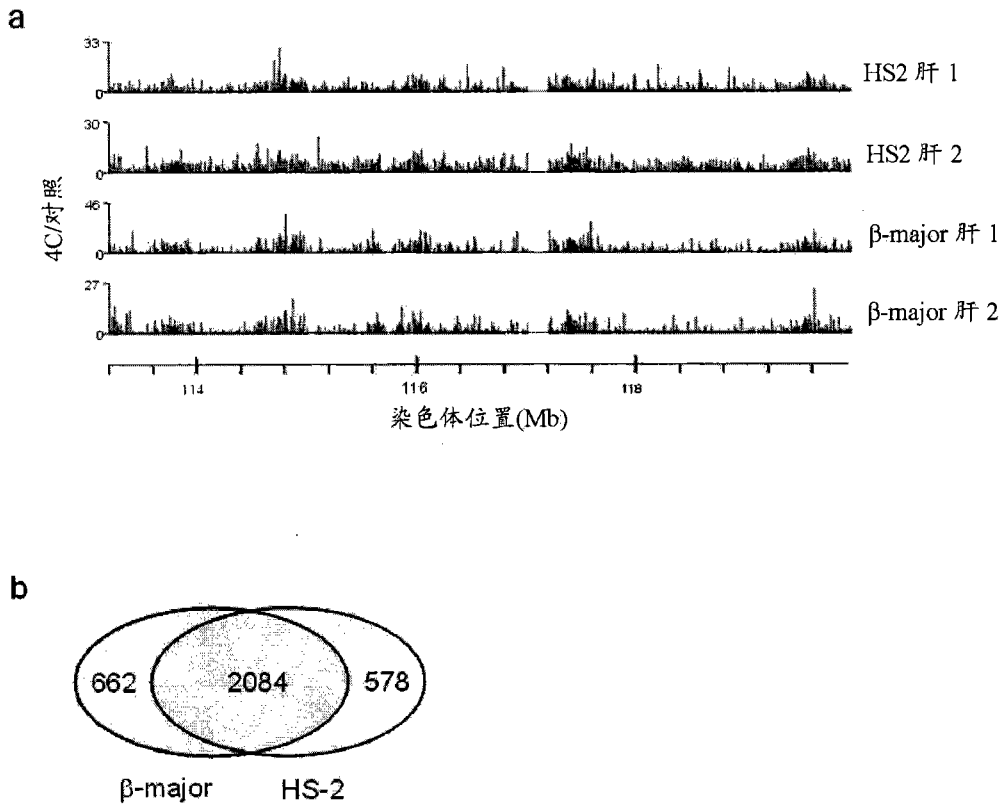


图 16

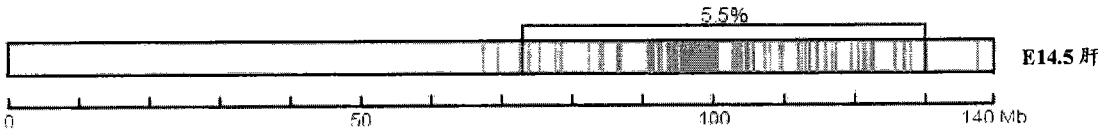


图 17

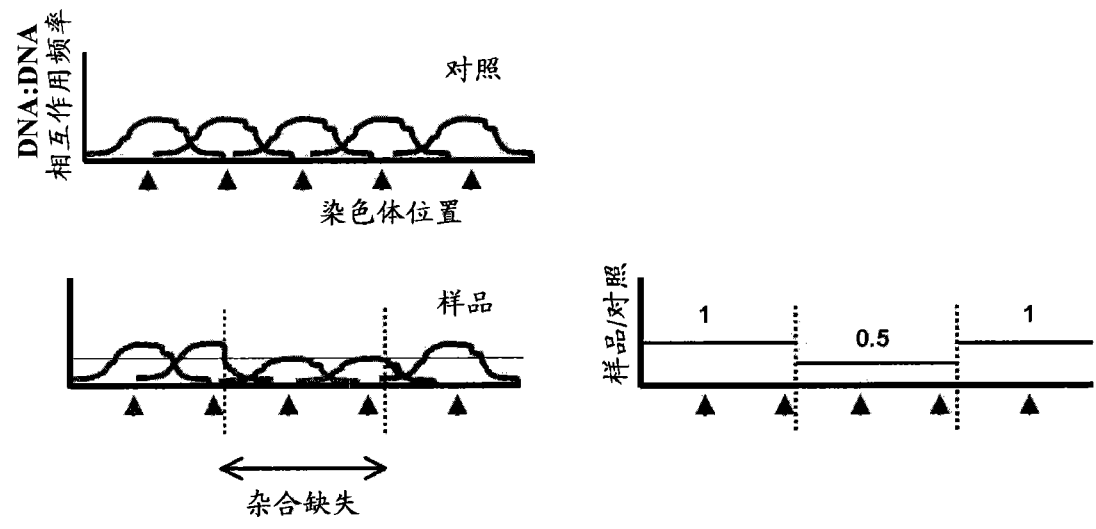
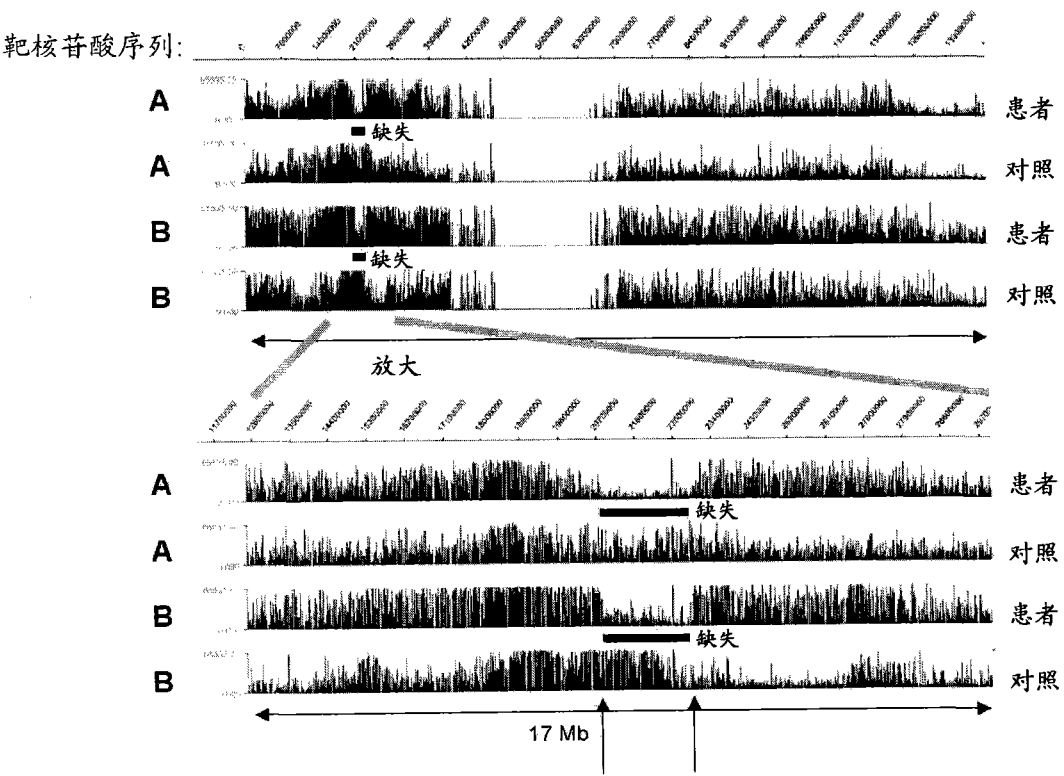


图 18



经过断点的序列:

AAGGAAACCCCATGCCCATAGACGTCACCTAATTCTGAACTCTTGTTTTTTTTTTTTTTTTTTT
CAAGTAGTTCTCATCTAAGTAGTTGTTTTTTGTCATGAGAAAATCAGATATGTTGCTAAAAATTCAAA
CTATTGCAAGAAAAAATAAAAGAC

图 19

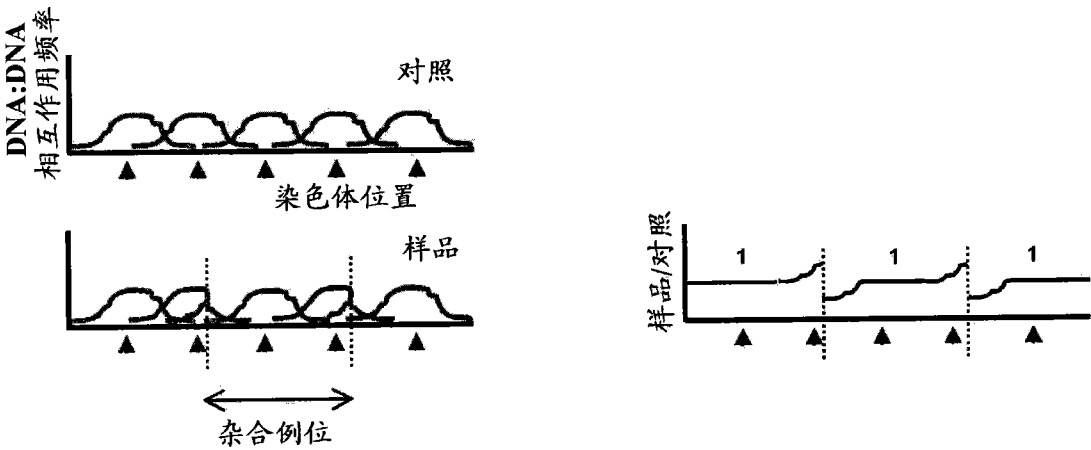


图 20

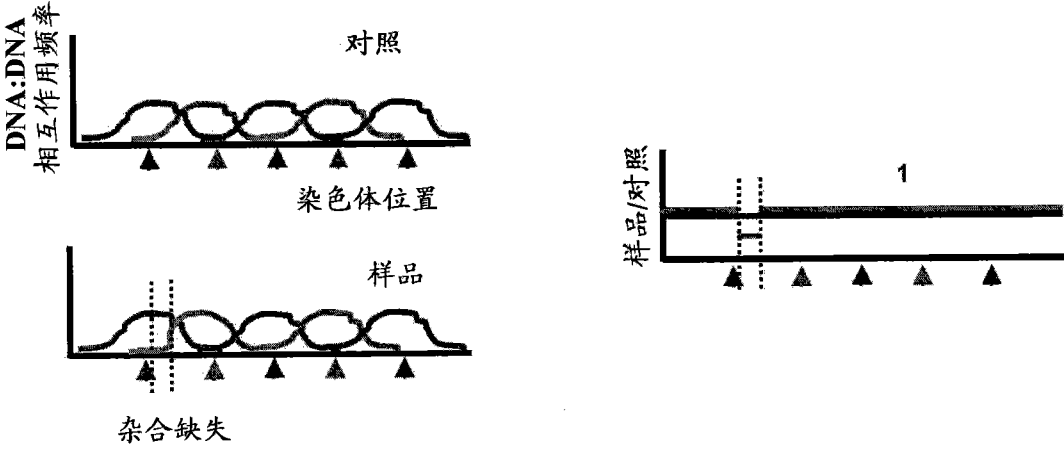


图 21

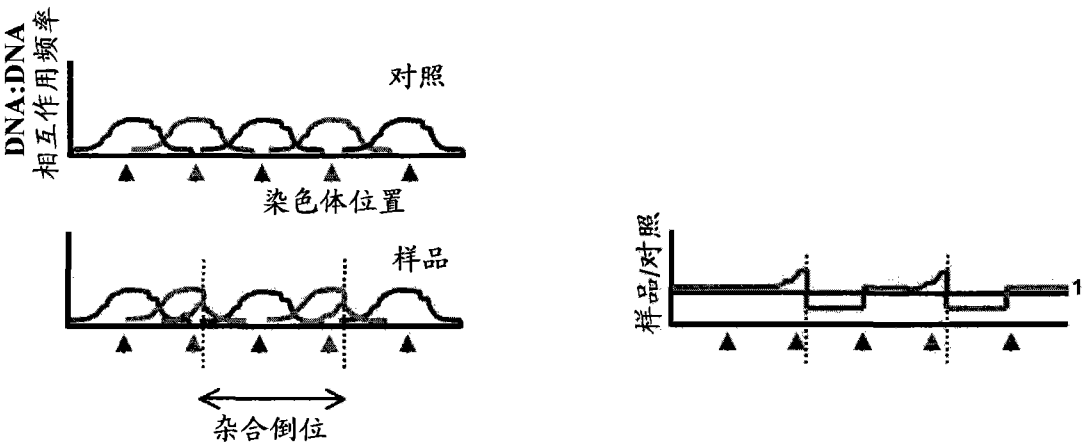


图 22

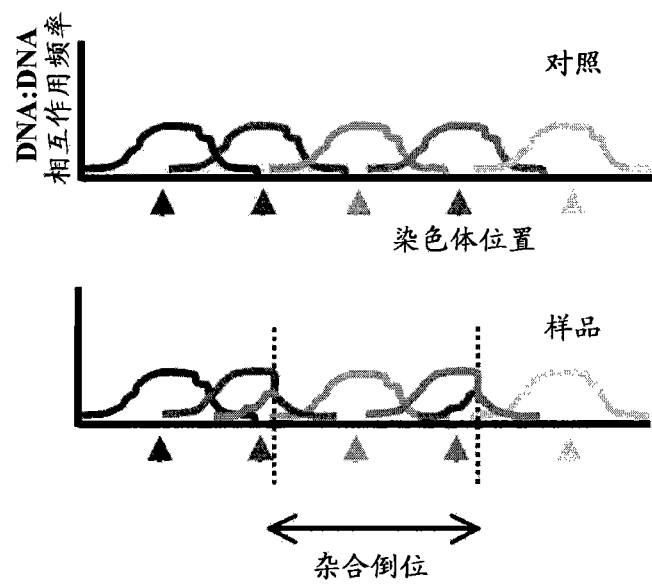


图 23

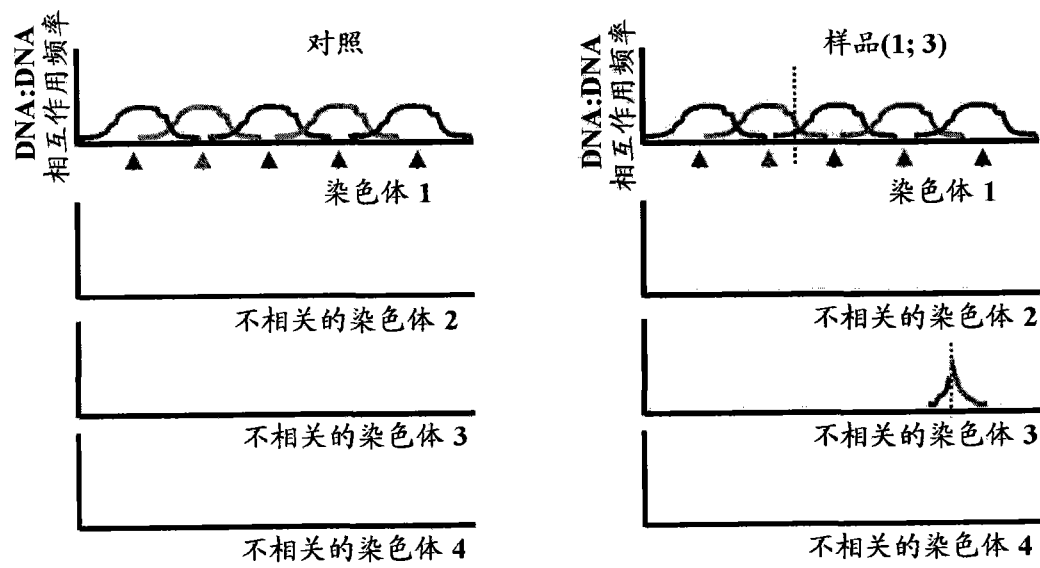


图 24

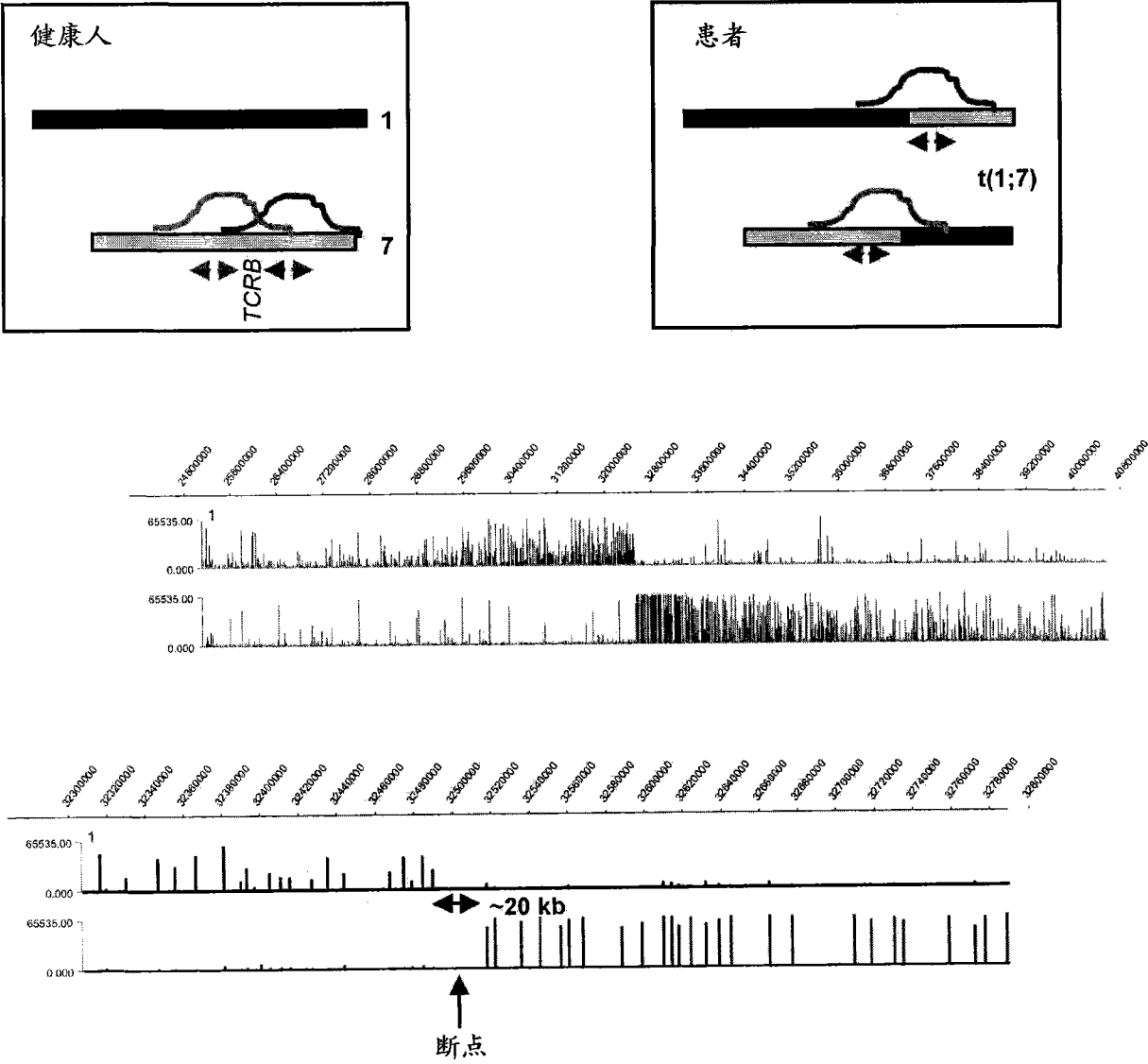


图 25

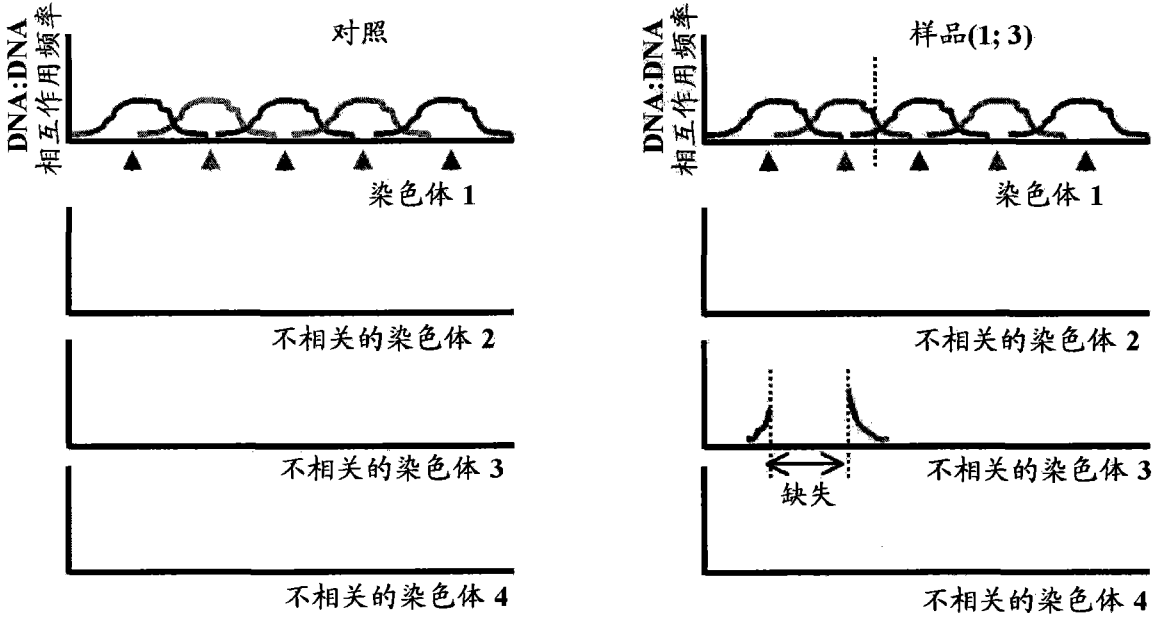
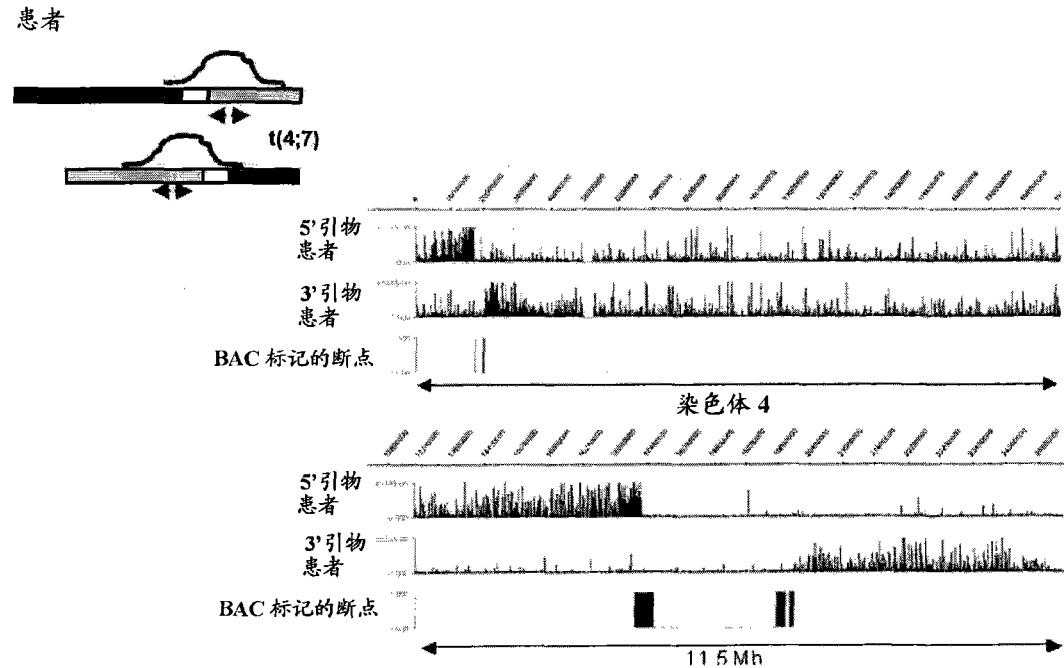


图 26



测序的断点 #2:
CTCCAATGTAACGTGGATTACACCTAAAAGAGCCAGAAAACACAGACTC
TCTGT**GGA**ACCATGACACAACAGTGCTTGGTATTATTTTTTCCTAGTTAG

图 27

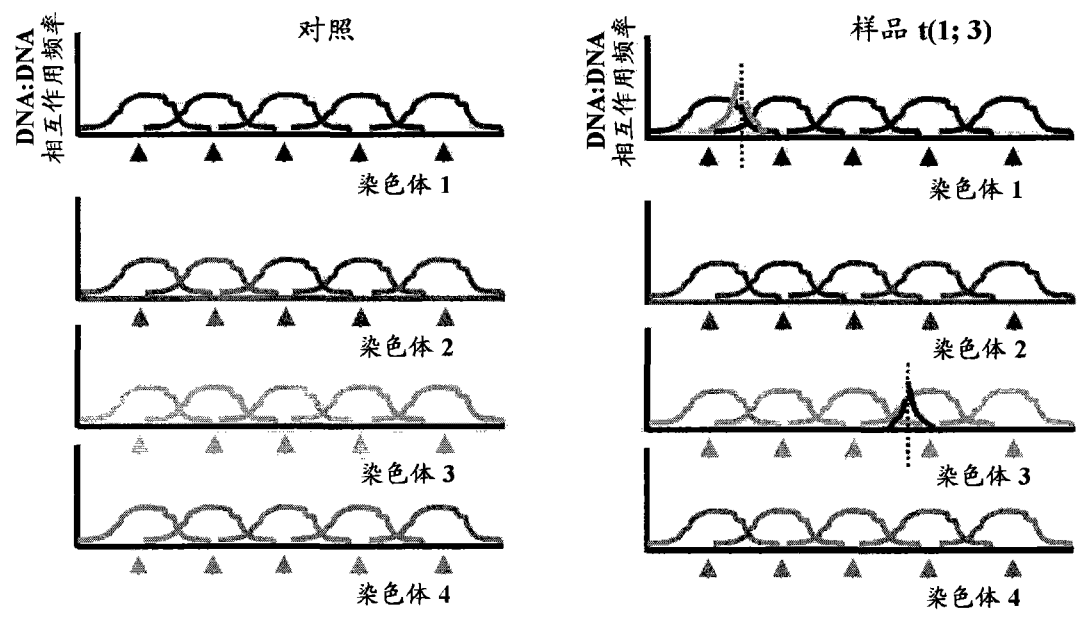


图 28

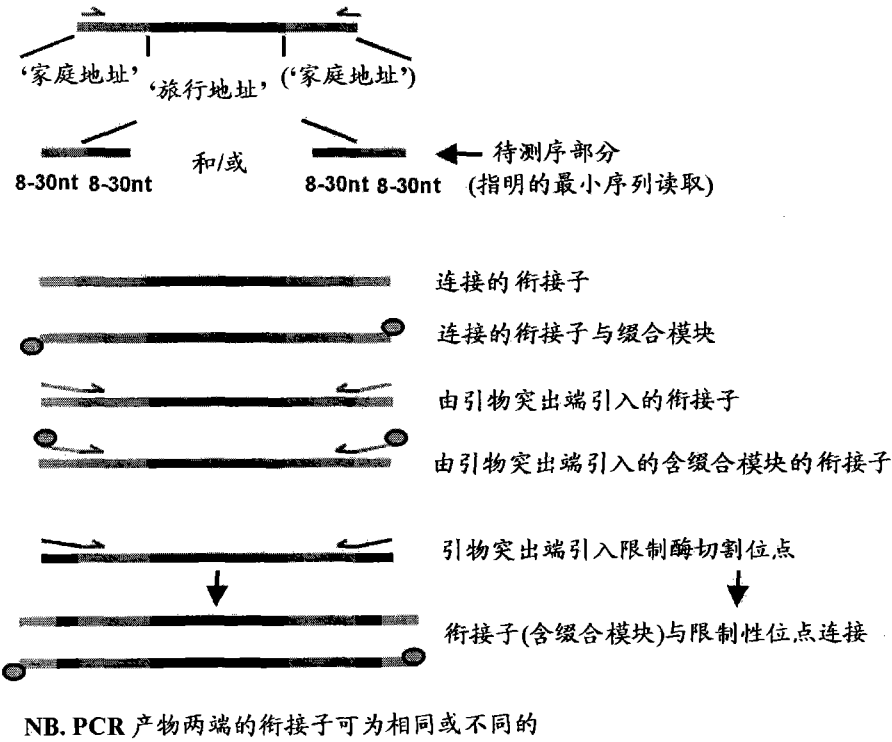


图 29

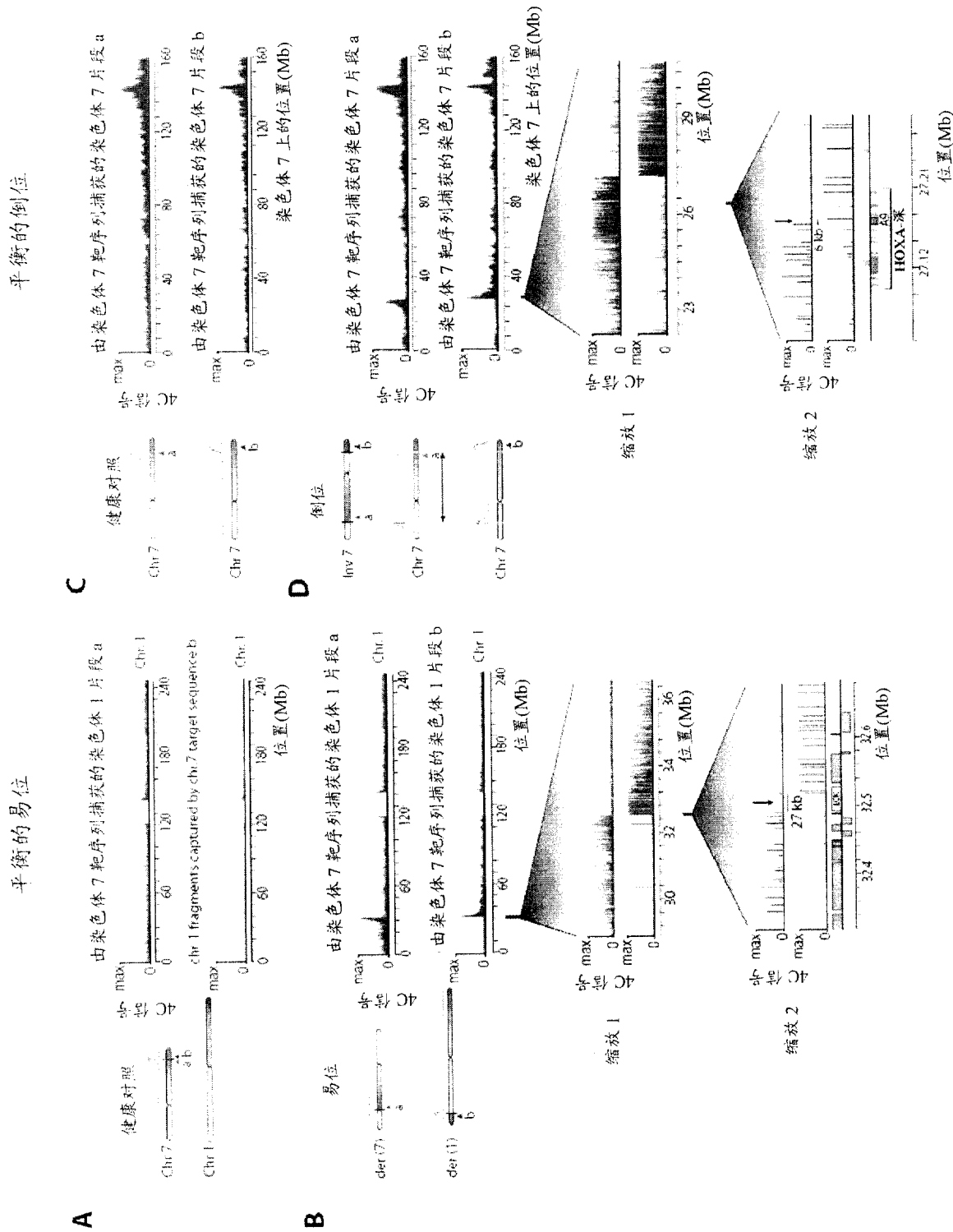


图 30

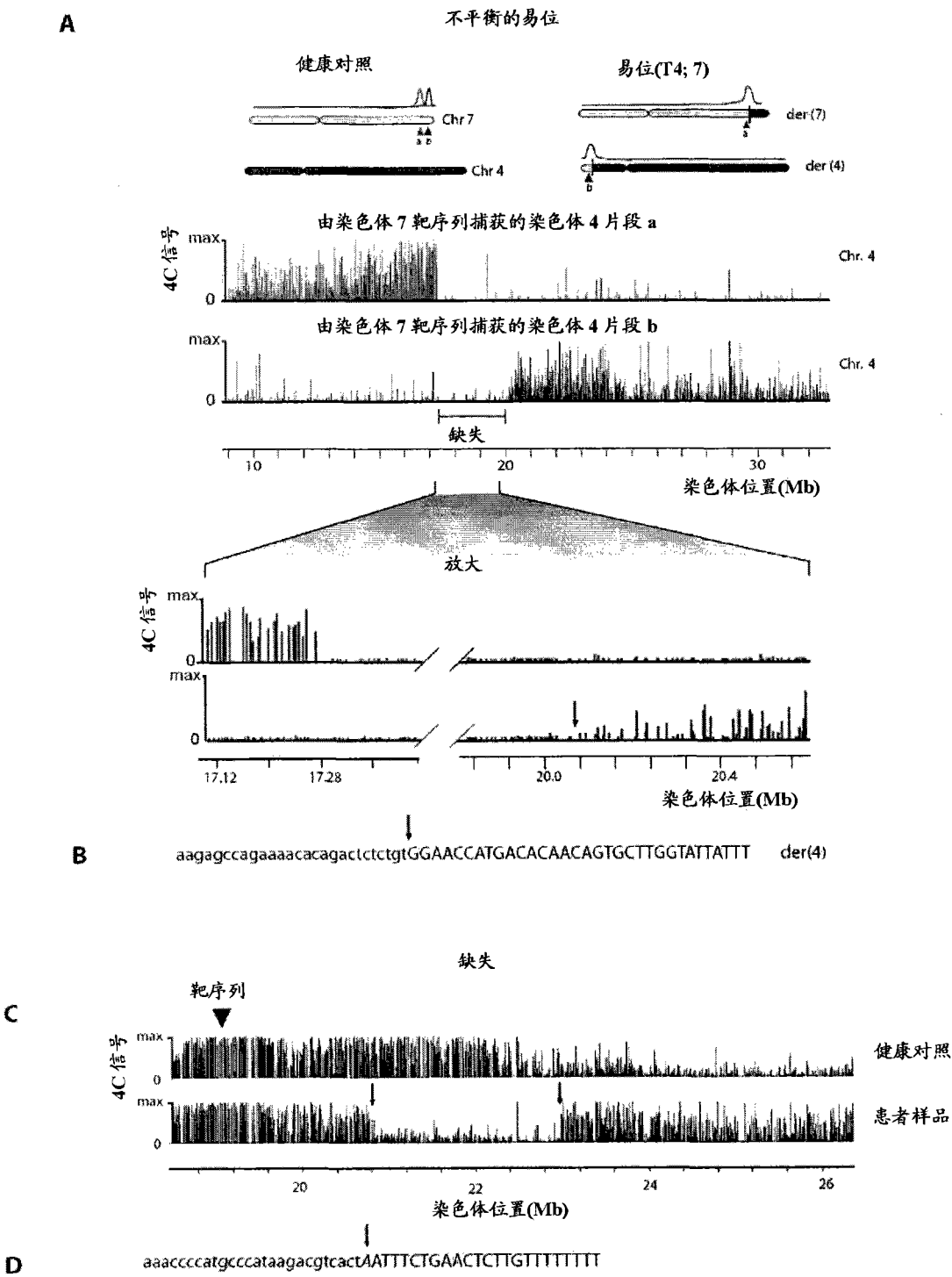


图 31

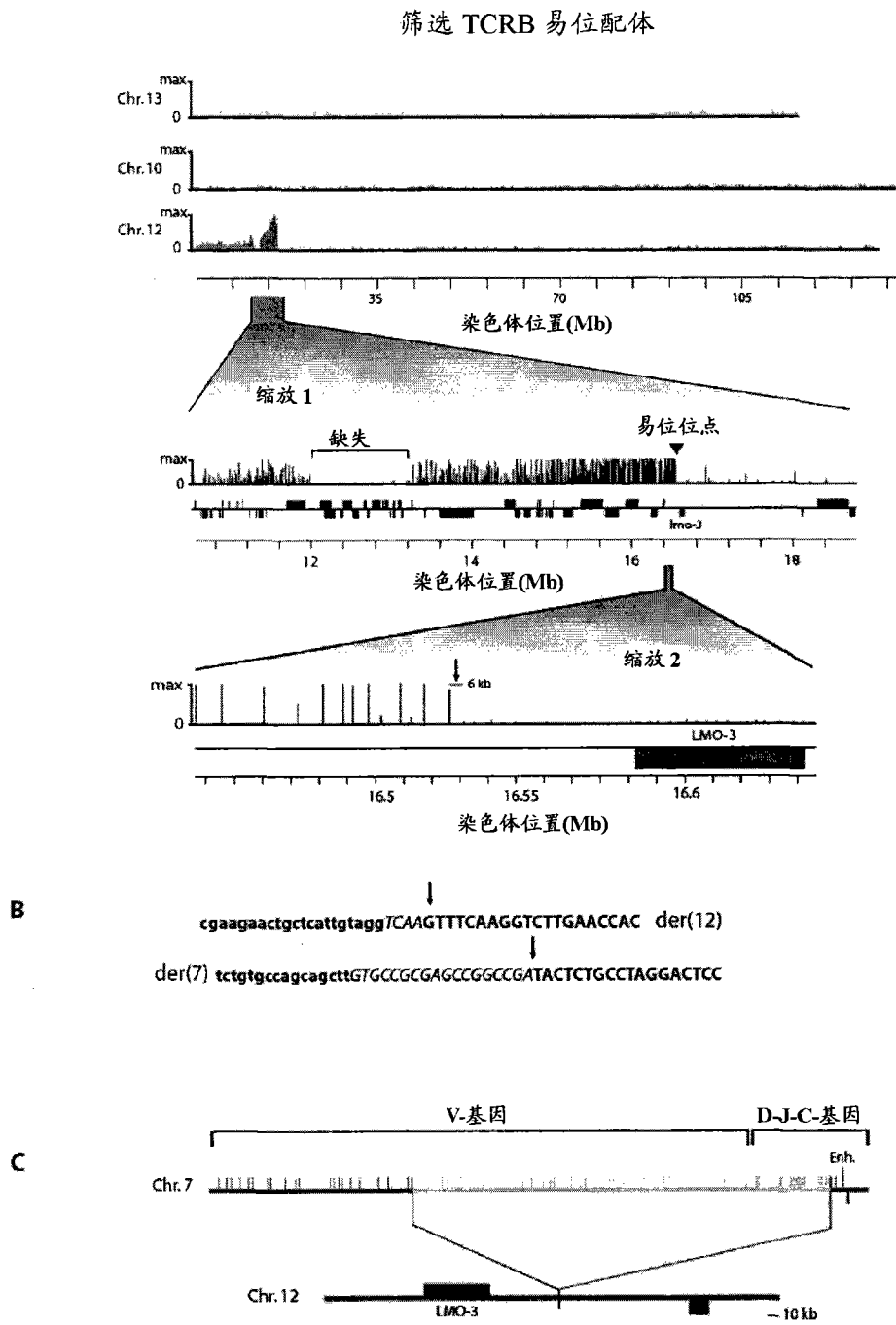


图 32

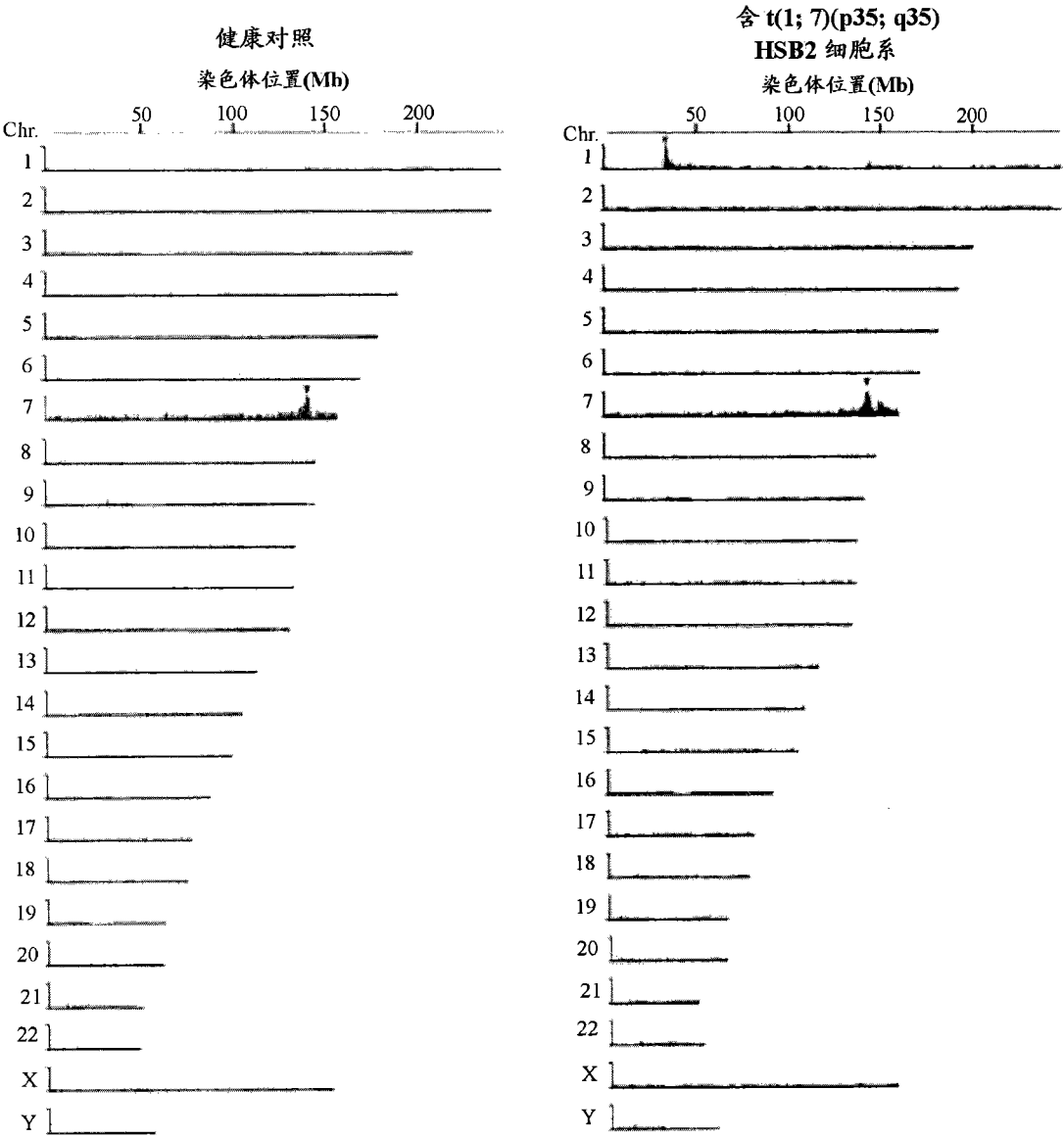


图 33

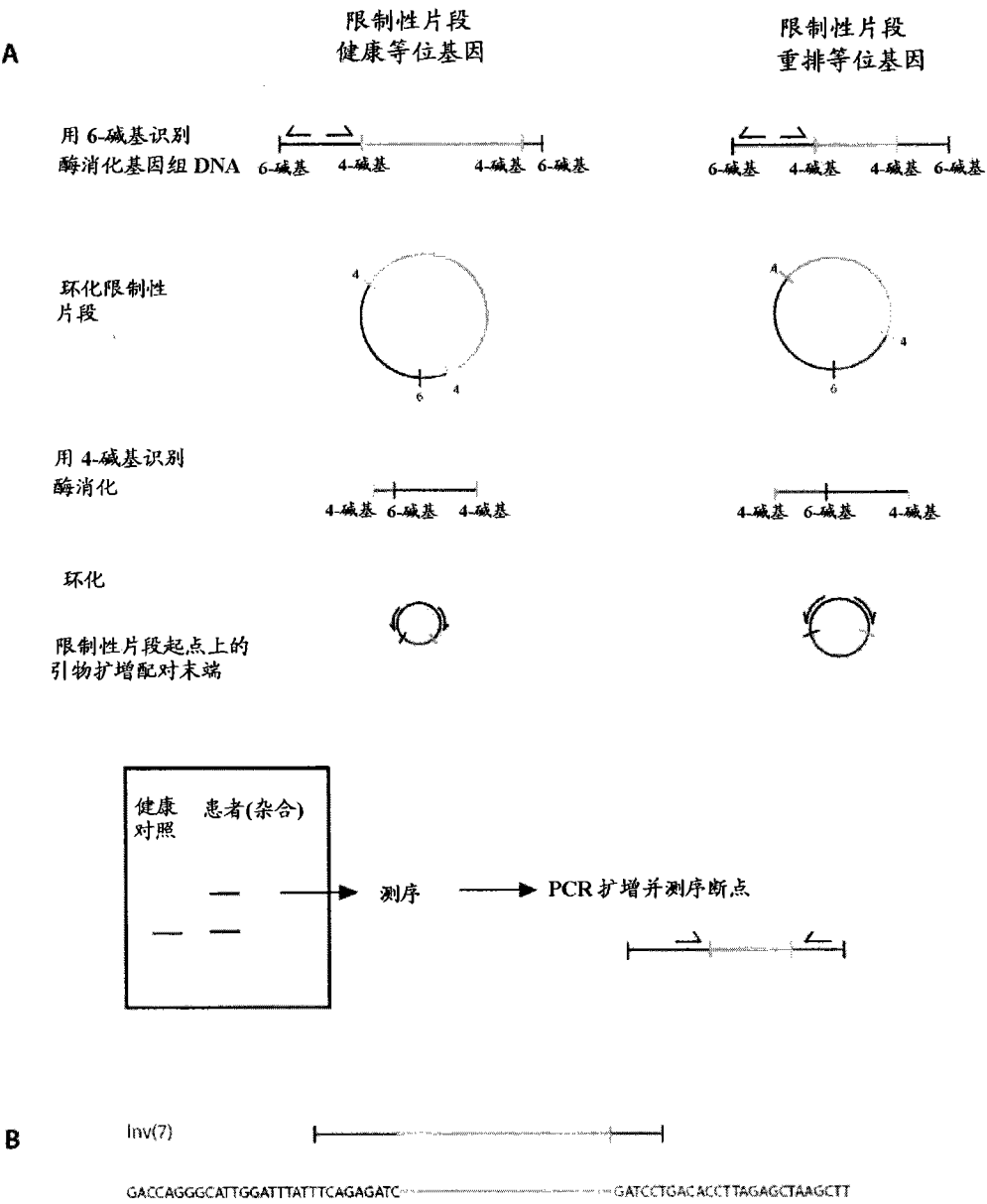


图 34

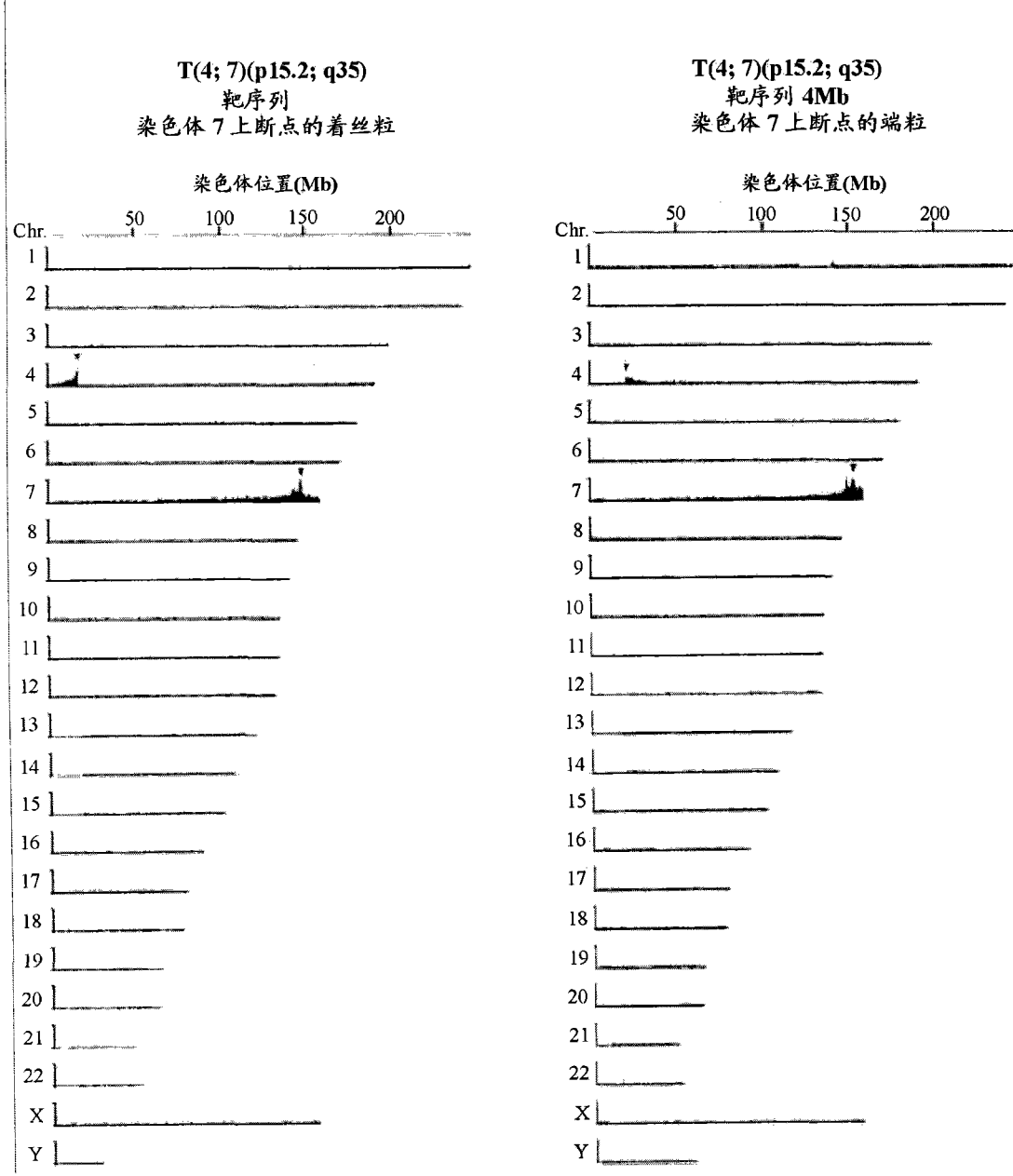


图 35

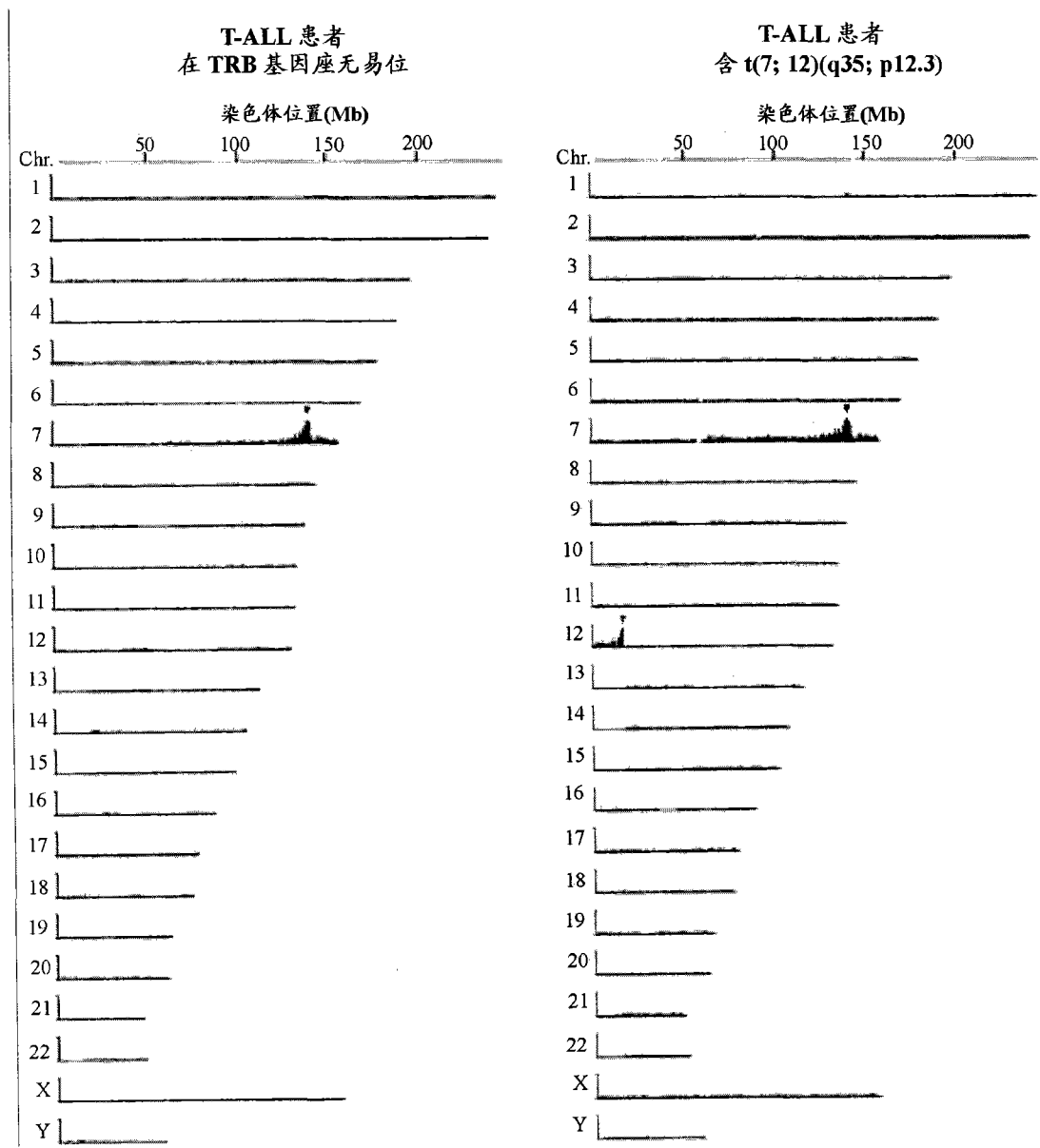


图 36

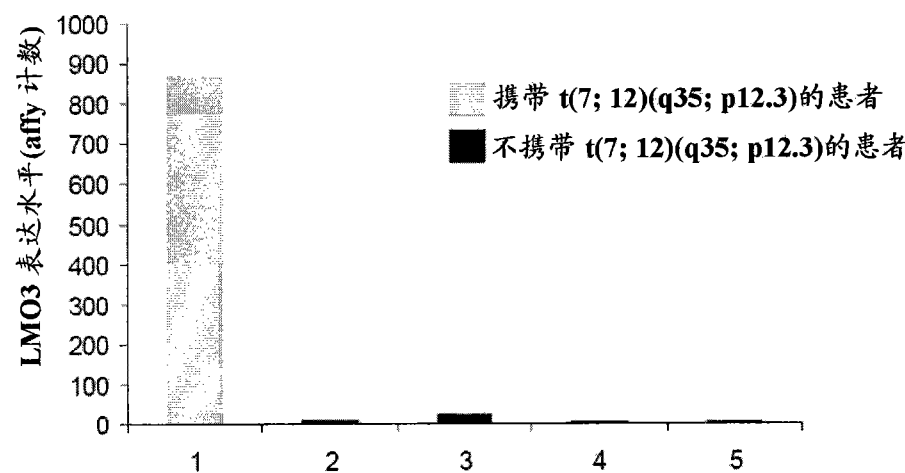


图 37

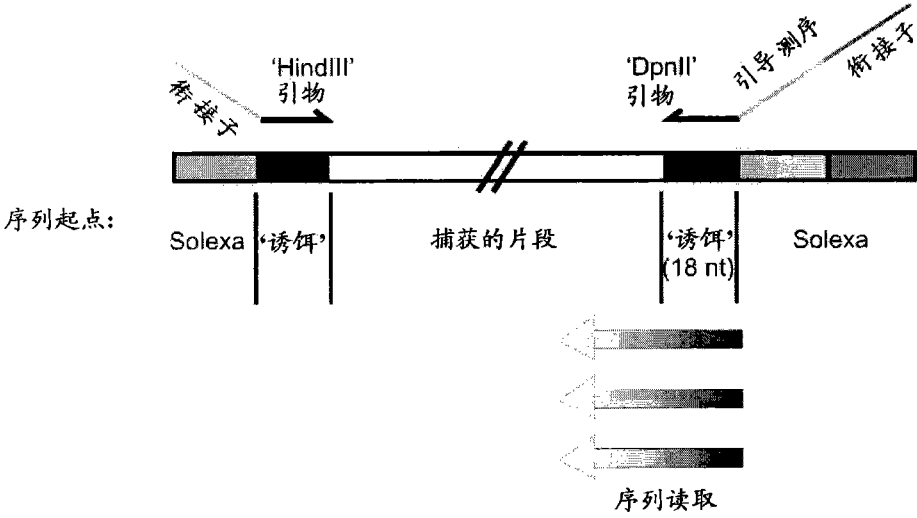


图 38

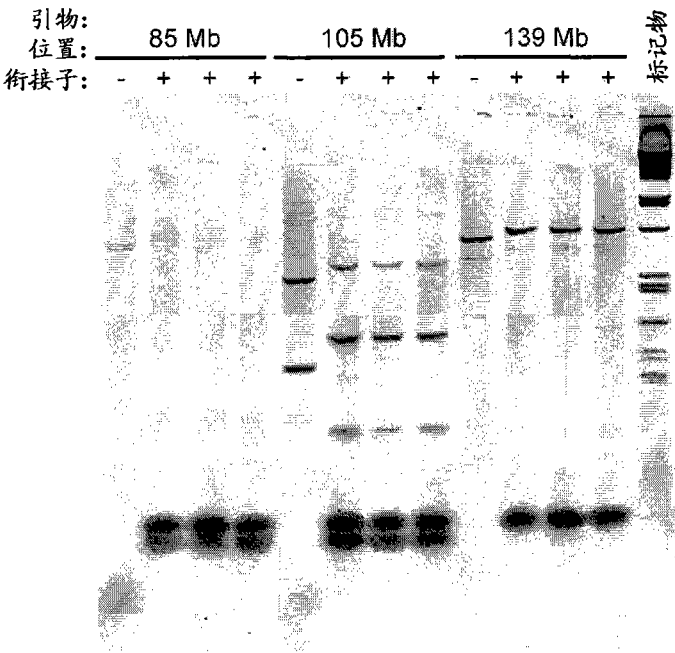


图 39

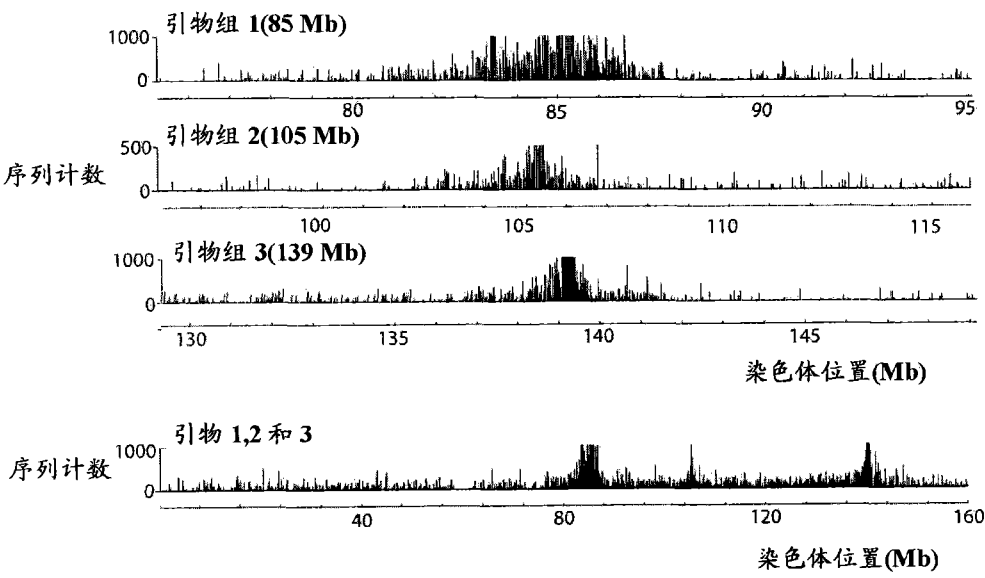


图 40

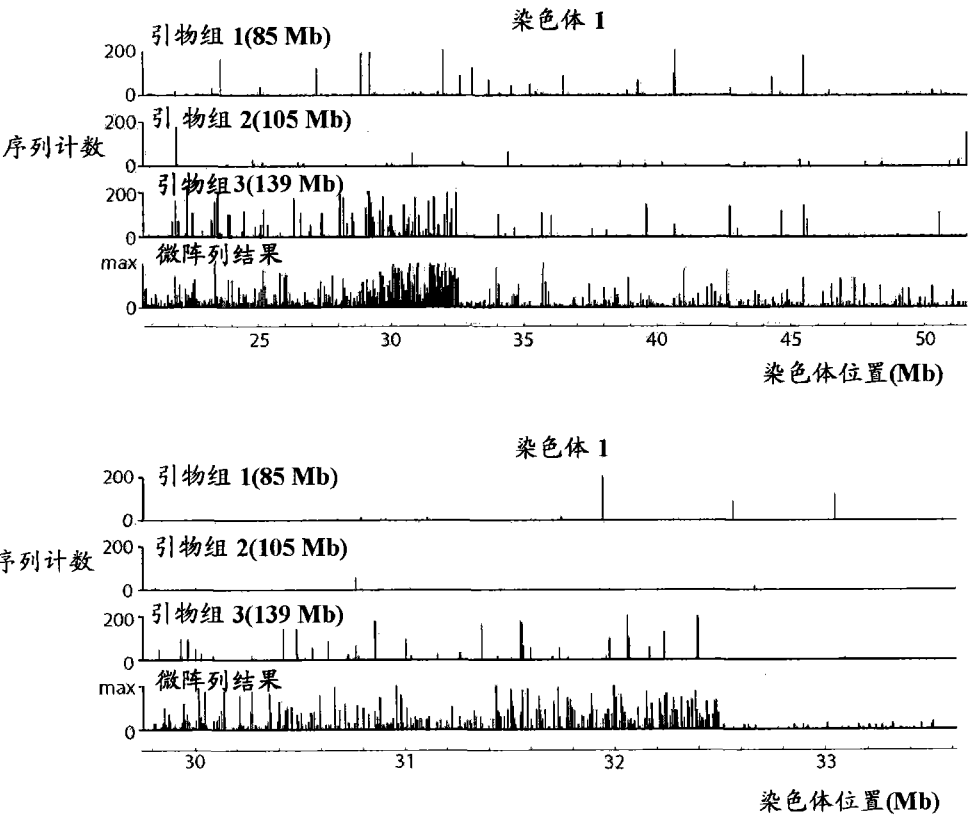


图 41