(19)中华人民共和国国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 103492588 A
(43)申请公布日 2014.01.01

(21)申请号 201280010224.X

(22)申请日 2012.02.24

(30)优先权数据
61/446,890 2011.02.25 US
61/509,960 2011.07.20 US

(85)PCT国际申请进入国家阶段日
2013.08.23

(86)PCT国际申请的申请数据
PCT/US2012/026623 2012.02.24

(87)PCT国际申请的公布数据
WO2012/116331 EN 2012.08.30

(71)申请人 伊路敏纳公司
地址 美国加利福尼亚州

(72)发明人 J-B. 范 J.S. 费希尔 F. 凯珀

(74)专利代理机构 中国专利代理(香港)有限公司 72001
代理人 徐晶 万雪松

(51)Int.Cl.
C12Q 1/68(2006.01)
C12N 15/11(2006.01)

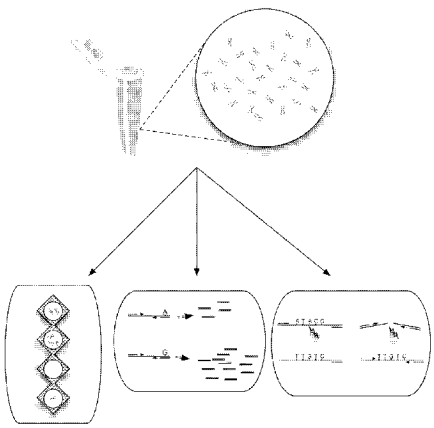
权利要求书3页 说明书20页 附图15页

(54)发明名称

用于单体型测定的方法和系统

(57)摘要

本发明的实施方案提供用于测定生物样品的单体型的方法和系统。特定的实施方案提供用于基因组的远程单倍体分型的方法。



CN 103492588 A

1. 一种用于测定核酸样品的单体型的方法，所述方法包括提供核酸样品的一个或多个部分，其中母系和父系染色体的贡献不相等，检测核酸样品的一个或多个部分中关注的两个或更多个序列之间的不平衡，并基于所述可检测的不平衡测定所述核酸样品的单体型。

2. 权利要求 1 的方法，其中所述核酸样品来自基因组或其片段。

3. 权利要求 2 的方法，其中所述基因组来自一个或多个细胞。

4. 权利要求 3 的方法，其中所述一个或多个细胞为约 10-100 个细胞。

5. 权利要求 1 的方法，其中所述核酸样品来自哺乳动物。

6. 权利要求 5 的方法，其中所述哺乳动物为人。

7. 权利要求 1 的方法，其中所述母系和父系染色体包括选自单核苷酸多态性、拷贝数目变异体、基因组插入和基因组缺失的一种或多种变异序列。

8. 权利要求 1 的方法，其中母系和父系染色体的所述不相等贡献包括除了 1:1 比例的染色体比例。

9. 权利要求 1 的方法，其中所述单体型通过荧光进行测定。

10. 权利要求 1 的方法，其中所述单体型通过核酸测序技术进行测定。

11. 权利要求 1 的方法，其中所述单体型通过在微阵列上实施的基因分型技术进行测定。

12. 权利要求 1 的方法，其中所述单体型通过定量聚合酶链反应进行测定。

13. 一种制备用于单体型测定的部分的方法，所述方法包括 ：

a) 提供包含对样品为天然的一定比例的母系和父系染色体组分的核酸样品，和

b) 产生多个部分，其中一个或多个部分包含偏倚比例的母系和父系染色体组分，其中所述偏倚比例基本上不同于对所述个体为天然的比例，从而制备用于单体型测定的部分。

14. 权利要求 13 的方法，其中所述产生包括向多个部分中的一个或多个部分不对称地分布母系和父系染色体组分。

15. 权利要求 13 的方法，其中所述产生包括在所述多个部分的一个或多个部分中差异性地降解母系或父系染色体组分中的一种或多种。

16. 权利要求 13 的方法，其中所述产生包括在所述多个部分的一个或多个部分中差异性地扩增母系或父系染色体组分中的一种。

17. 权利要求 13 的方法，其中所述核酸样品来自哺乳动物。

18. 权利要求 17 的方法，其中哺乳动物为人。

19. 权利要求 13 的方法，其中所述核酸样品来自多个细胞。

20. 权利要求 19 的方法，其中所述多个细胞为中期同步的。

21. 权利要求 19 的方法，其中所述多个细胞为约 5- 约 300 个细胞。

22. 权利要求 19 的方法，其中所述多个细胞为约 10- 约 100 个细胞。

23. 一种用于对样品中关注的多个序列测定单体型的方法，所述方法包括 ：

a) 提供来自权利要求 13 的一个或多个部分，

b) 自所述一个或多个部分创建一个库，

c) 对所述多个关注的序列检测可检测的信号，

d) 基于可检测的信号的所述差异测定关注的多个序列的单体型。

24. 权利要求 23 的方法，其中所述关注的两个或更多个序列在同一染色体上。

25. 权利要求 23 的方法，其中所述关注的两个或更多个序列位于同一染色体的两个或更多个不同位点上。

26. 权利要求 24 的方法，其中同一染色体的两个或更多个不同位点由至少 10000 个核苷酸分开。

27. 权利要求 24 的方法，其中所述两个或更多个不同位点位于同一染色体上，并由至少 100000 个核苷酸分开。

28. 权利要求 24 的方法，其中所述两个或更多个不同位点位于同一染色体上，并由至少 100000000 个核苷酸分开。

29. 权利要求 24 的方法，其中所述两个或更多个不同位点位于同一染色体上，并由至少 200000000 个核苷酸分开。

30. 权利要求 23 的方法，其中所述一个或多个部分来自个体生物。

31. 权利要求 23 的方法，其中所述一个或多个部分来自哺乳动物。

32. 权利要求 23 的方法，其中所述一个或多个部分来自人。

33. 权利要求 23 的方法，所述方法进一步包括在步骤 b) 之前测定母系和父系染色体的比例。

34. 权利要求 23 的方法，其中所述测定单体型包括部分的定量聚合酶链反应分析。

35. 权利要求 23 的方法，其中所述测定单体型包括部分的微阵列分析。

36. 权利要求 23 的方法，其中所述测定单体型包括对多个关注序列中的每一个检测序列读取数目的差异，匹配具有相似序列读数的关注的序列，并基于所匹配的关注的序列测定单体型。

37. 权利要求 23 的方法，其中所述可检测的信号为荧光。

38. 权利要求 36 的方法，其中所述可检测的信号为荧光。

39. 权利要求 23 的方法，其中所述两个或更多个关注的序列选自等位基因、单核苷酸多态性、拷贝数目变异体、基因组插入和基因组缺失。

40. 权利要求 23 的方法，其中所述检测包括核酸测序技术。

41. 权利要求 23 的方法，其中所述检测包括在微阵列上实施的基因分型技术。

42. 权利要求 23 的方法，其中所述检测包括定量聚合酶链反应基因分型技术。

43. 权利要求 40 的方法，其中所述测序技术检测自多个关注序列的读取总数扣除多个关注序列的读取数目的差值。

44. 权利要求 43 的方法，其中检测读取数目包括检测多个关注的序列产生的荧光信号的数目。

45. 一种测定多个位点的等位基因的取相的方法，所述方法包括 ：

a) 提供核酸分子的不对称分布，其中不对称分布包含多个部分，其中各个部分包含等位基因的多份拷贝，和其中各个部分包含不同数量的等位基因 ；

b) 区分存在于一个或多个各个部分中的核酸分子拷贝中的等位基因 ；

c) 评价存在于一个或多个各单独的部分中等位基因的不同数量 ；和

d) 自等位基因的区分和自不同数量的等位基因的评价确定多个位点的等位基因的取相。

46. 权利要求 45 的方法，其中所述评价包括检测自读取总数扣除多个位点的等位基因

的荧光测序读取数目的差值。

47. 权利要求 45 的方法，其中所述核酸分子来自个体生物。

48. 权利要求 45 的方法，其中所述不同数量的评价包括测定多个位点的等位基因的比例。

49. 权利要求 45 的方法，其中所述等位基因的区分包括测定存在于多个位点的一个或多个核苷酸的同一性。

50. 权利要求 45 的方法，其中所述等位基因的区分包括核酸测序技术。

51. 权利要求 45 的方法，其中所述等位基因的区分包括在微阵列上实施的基因分型技术。

52. 权利要求 45 的方法，其中所述多个位点位于同一染色体上，并由至少 10000 个核苷酸分开。

53. 权利要求 45 的方法，其中所述多个位点位于同一染色体上，并由至少 100000 个核苷酸分开。

54. 权利要求 45 的方法，其中所述多个位点位于同一染色体上，并由至少 100000000 个核苷酸分开。

55. 权利要求 45 的方法，其中所述多个位点位于同一染色体上，并由至少 200000000 个核苷酸分开。

56. 一种用于测定单体型的核酸部分，其中所述核酸部分包含不对称地分布的母系和父系染色体组分，其中所述不对称分布的染色体组分为偏倚比例的母系与父系染色体组分，这种偏倚比例不同于对个体为天然的比例。

用于单体型测定的方法和系统

[0001] 本申请要求2011年2月25日递交的美国临时专利申请系列号61/446890和2011年6月20日递交的美国临时专利申请系列号61/509960的优先权，其两者通过引用以其全部结合到本文中。

[0002] 背景

人类基因组计划的努力开辟了更广阔的人类遗传密码的窗口。例如使用高通量测序技术进一步解开人类基因组的工作正在不断的进行中。HapMap（单体型图）计划(HapMap (Haplotype Map) Project)为通过比较没有特定疾病的人群与具有所述疾病的人群的基因组信息，针对发现导致疾病的基因变异的全球性的科学努力。等位基因，对于特定基因的DNA序列的一种或多种形式，可含有一个或多个不同的基因变异和识别的单体型，或者特定染色体上的不同位置或位点的等位基因的组合为HapMap计划(HapMap Project)的主要焦点。所确认的其中两组不同的单体型可能与引起疾病的基因异常的位置相关。这样，HapMap结果将有助于描述在人类基因变异的常见模式以及这些变异是否潜在地与疾病相关。

[0003] 从这些努力获得的信息，即使序列是不完整的，并且存在差距和有时是错误的，在帮助破译疾病和障碍背后的遗传学方面提供有价值的工具。不幸地，进行这样大规模测序的成本仍然非常高，并且提供更深入的信息的技术比如单染色体单倍体分型、等位基因或引导序列的取相(phasing)为虚幻的。所需要的是从人类基因组解开更多信息的另外的工具和技术。

[0004] 概述

目前的基因分型技术可给研究者提供受试者的基因组成。然而，关于提供方便和可扩展的手段的技术有限，这种手段用来测定一个染色体上的什么序列相对于另一个染色体上的相邻或邻近的那些序列彼此相邻或邻近。图2举例说明一种困境，其中受试者的基因型可被测定，然而为测定关注的序列（例如等位基因、单核苷酸多态性(SNP)、拷贝数目变异体(CNV)、基因插入或缺失（插入/缺失(indel)等）是否位于与另一个关注的序列相同的染色体上所获得的信息不足。例如，对于采自受试者的样本中的染色体的混合群体（图2A)，可以能自数据测定示例性的基因型（图2B)。然而，对于测定杂合性等位基因如何在染色体上组合在一起（单倍体分型）提供的信息不足。例如，不知是否母体A (P_a) 提供等位基因α和γ，母体B (P_b) 提供等位基因α'和γ'（图2C)，或者是否它们为混合的（图2D)。当那些序列在染色体上彼此相隔很远或位于远端或远程时，甚至更加难以测定哪些关注的序列存在于相同的染色体上，从而测定染色体的长单倍型或等位基因取相。

[0005] 本公开的实施方案提供用于测定取相的(phased)等位基因而不管其彼此在染色体上的位置（例如近端或远端）的新的解决方案。在针对解决当前的单倍体分型挑战的实验期间，发现提供遗传物质的不平衡或不对称分布，对于受试者的准确单倍体分型问题提供一种新的解决方案。在不平衡分布后的引导序列的任选扩增是特别有用的。本发明不限于特定的机制。的确，理解机制对于实践本发明是没有必要的。但是，考虑部分基于不平衡物质的差分扩增(differential amplification)，扩增信号强度确定染色体的单体型。例如，不同等位基因信号的比例确定哪一个存在于单染色体上，从而确定样本的取相的单体

型。图 3 举例说明这样的实施方案。原始样本分布的不平衡（如在 3B 和 3D 所见）被利用，并且差分扩增证实，α 等位基因被取相或与 P_a 的 γ’等位基因组合在一起，和 α’在 P_b 上取相，与 γ 组合在一起 (3E)。进一步地，实施方案不限于单倍体样本，而是当采用二倍体样本（例如配对的染色体、DNA 插入、YACs、BACs、粘粒、F 粘粒 (fosmids) 等）或单倍体样本（例如来自精子、卵子、完整的水泡样胎块 (hydatiform mole) 等的遗传互补）时有效。

[0006] 发现从通过实践本文描述的方法提供的基因组中的等位基因取相获得的信息，在一般性研究和发现努力以及例如疾病检测、治疗和用于降低移植排斥反应的 HLA 相容性的更高信心方面具有用途。例如，已知的单体型可能与药物代谢、药物发现、疾病状态、癌症、障碍、移植排斥反应的风险和指定极少数的个性化的卫生保健计划相关。的确，关于个性化的卫生保健，一旦受试者的个人单体型为已知，那么受试者的特定疾病相关性和治疗选择可专门地进行设计，以满足所述受试者的需要。

[0007] 本公开的一个实施方案包括用于通过提供部分样品（在核酸样品中包含关注的两个或更多个序列之间可检测的不平衡），并基于所述可检测的不平衡测定核酸样品的单体型，测定核酸样品的单体型的方法。在一些实施方案中，核酸样品来自基因组或其片段，其中所述基因组源于一个或多个细胞，例如约 1-100 个细胞。在一些实施方案中，核酸样品来自哺乳动物，优选地来自人。在其它的实施方案中，核酸样品来自非人哺乳动物、植物或病毒。在一些实施方案中，核酸样品包含关注的序列的野生型序列，而在其它的实施方案中，核酸样品包含关注的序列的变异序列。在一些实施方案中，关注的序列包含关注的一个序列的野生型序列和关注的另一个序列的变异序列或其组合。在一些实施方案中，变异序列选自单核苷酸多态性、拷贝数目变异体、基因组插入和基因组缺失。在一些实施方案中，样品中关注的两个或更多个序列之间可检测的不平衡通过荧光进行测定。在一些实施方案中，样品中关注的两个或更多个序列之间可检测的不平衡通过核酸测序技术、通过例如在微阵列实施的基因分型技术或通过定量聚合酶链反应进行测定。

[0008] 本公开的一个实施方案包括制备用于单体型测定的部分的方法，所述方法包括提供包含染色体组分的核酸样品，并把染色体组分不对称地分布成多个部分，从而制备用于单体型测定的部分。在一些实施方案中，染色体组分的不对称分布包括把不等量的染色体组分递送至多个部分中的不同部分中。在一些实施方案中，不对称地分布的染色体组分的比例与初始细胞群体中的染色体组分的比例不同。在一些实施方案中，染色体组分的不对称分布包括在多个部分中的不同部分中差异性地降解染色体组分。在一些实施方案中，染色体组分的不对称分布包括在多个部分中的不同部分中差异性地扩增染色体组分。在一些实施方案中，核酸样品来自哺乳动物，优选地来自人。在其它的实施方案中，核酸样品来自非人哺乳动物、植物或病毒。在一些实施方案中，核酸样品来自多个细胞，例如约 5-300 个细胞或约 10-100 个细胞。在一些实施方案中，多个细胞为中期同步的，而在其它的实施方案中，多个细胞不为中期同步的。在一些实施方案中，染色体组分包含在不同位点的两个或更多个等位基因，其中这些等位基因进一步包含关注的一个或多个序列。

[0009] 本公开的一个实施方案包括用于测定关注的两个或更多个序列的取相 (phasing) 的方法，所述方法包括提供其中在所述部分中的染色体组分不对称地分布的部分，从所述部分创建一个库，对库中关注的两个或更多个序列检测可检测的信号，并基于可检测的信号中的所述差异测定关注的两个或更多个序列的取相。在一些实施方案中，可检测的信号

为荧光信号。在一些实施方案中，关注的两个或更多个序列为在同一染色体上，并且进一步地位于同一染色体的两个或更多个不同位点上。在一些实施方案中，位于同一染色体的两个或更多个不同位点由至少 10000、至少 100000、至少 100000000 或至少 200000000 个核苷酸隔开。在一些实施方案中，所述部分来自个体生物。在一些实施方案中，所述部分来自哺乳动物，例如来自人。在其它的实施方案中，所述部分来自非人哺乳动物、植物或病毒。在一些实施方案中，在提供所述部分用于相测定之前，测定所述部分中关注的两个或更多个序列之间的不对称度。在一些实施方案中，测定不对称度包括所述部分的定量聚合酶链反应分析。在一些实施方案中，测定不对称度包括所述部分的微阵列分析。在一些实施方案中，测定不对称度包括测定所述部分中关注的两个或更多个序列之间的信噪比。在一些实施方案中，所述部分中关注的两个或更多个序列之间的信噪比大于其它部分中的信噪比。在一些实施方案中，信噪比通过荧光检测来测定。

[0010] 本公开的一个实施方案包括用于测定两个或更多个不同位点的等位基因的相的方法，所述方法包括提供在两个或更多个不同位点包含等位基因的核酸分子的不对称分布，其中不对称分布包含多个部分，其中各独立的部分包含等位基因的多份拷贝，和其中各独立的部分包含不同数量的等位基因，区分存在于一个或多个独立的部分中的核酸分子拷贝中的等位基因，评价存在于一个或多个独立的部分中的不同数量的等位基因，并且对于两个或更多个不同位点的等位基因自等位基因的区分和自不同数量的等位基因的评价测定取相。在一些实施方案中，评价包括检测两个或更多个不同位点的等位基因的读取总数减去两个或更多个不同位点的等位基因的荧光测序读取数目的差值。在一些实施方案中，不对称分布的核酸分子来自个体生物。在一些实施方案中，评价等位基因的不同数量包括测定两个或更多个不同位点的等位基因的比例。在一些实施方案中，评价不同数量包括计数两个或更多个不同位点的等位基因。在一些实施方案中，区分等位基因包括核酸测序技术，而在其它的实施方案中，区分等位基因包括在微阵列是实施的基因分型技术。在特殊情况下，可使用核酸测序技术和基于阵列的基因分型技术。在一些实施方案中，两个或更多个不同位点在同一染色体上并由至少 10000 个核苷酸分开。在一些实施方案中，位于同一染色体的两个或更多个不同位点由至少 100000、至少 100000000 或至少 200000000 个核苷酸分开。

[0011] 定义

本文使用的术语“单体型”指的是单倍体基因型、在染色体的不同位置或位点发现的等位基因或 DNA 序列的组合或组，其通常作为一个单位遗传而得和例如在易位事件期间被连接。单体型可提供个体的独特遗传模式。单体型可依在给定组的位点之间发生的重组事件的数目而定对于一个位点、几个位点或整个染色体进行测定。等位基因或 DNA 序列不限于任何特定的类型，并且包括例如正常的基因序列（即非变异的）或变异的基因序列。例如单核苷酸多态性 (SNPs)、短串联重复序列 (STRs) 等可被考虑为变异的基因序列。术语“取相的等位基因”指的是在单染色体上的特定等位基因的分布。因此，两个等位基因的“取相”可指表征或测定等位基因是位于单染色体上，还是位于两个独立的染色体（例如母系或父系遗传的染色体）上。除非另作说明，“单体型”和“取相的等位基因”被认为是同义词。

[0012] 本文使用的术语“分离的”、“纯化的”或“纯化”指的是自样品去除组分（例如污染物）的产品或行为。例如，核酸通过去除污染宿主细胞或其它蛋白质、用于自其存在的环

境分离核酸的盐、酶、缓冲剂等，被分离或分离远离细胞碎屑或分离试剂。

[0013] 本文使用的术语“样品”与其在生物学和化学领域的含义一致进行使用。在某种意义上，其意指包括来自从任何来源比如生物和环境样品得到的样本或培养物的核酸。生物样品可得自动物，所述动物包括但不限于人、非人灵长类动物和非人动物，所述非人动物包括但不限于脊椎动物比如啮齿动物、绵羊、牛科动物、反刍动物、兔类动物、猪、山羊、马、犬科动物、猫科动物、鸟类等。生物样品包括但不限于流体比如血液制品、组织、细胞等。生物样品可进一步属于植物来源，单子叶植物的或双子叶植物的、落叶性或常绿的、草本或木本的，包括但不限于农业植物、景观植物、苗圃植物等。环境样品可为细菌、病毒、真菌等起源的。优选的样品为真核生物起源的。基本上，研究者在测定取相的等位基因中关注的任何生物核酸样品来源适用于本发明。样品也可包括合成的核酸。核酸的衍生物或产品比如扩增的拷贝或化学改性的种类也包括在内。

[0014] 本文使用的术语“核酸”例如可为核苷酸的聚合物或多核苷酸。该术语可用于指定单分子或分子的集合。核酸可为单链或双链，并可包括编码区和各种控制元件的区域、非编码区、整个染色体、部分染色体、其片段和变体。

[0015] 本文使用的术语“不对称的”、“不平衡的”、“不等的”或“有偏倚的”，当用于指类似项目的分布时，被认为是同义词，除非另外说明。所述术语指的是类似项目例如染色体或染色体组分的集合，其跨多个部分、等分试样、亚组等分布，使得在两个或更多个独立的部分存在不同数量的类似项目。多个部分中的两个或更多个独立的部分可具有类似项目。然而，不是多个部分中的所有部分需要具有项目，相反一个或多个部分、等分试样、亚组等可能没有项目。独立的部分关于存在的项目可为均匀的，或者作为选择可在独立的部分存在项目的不均匀集合，使得与一种或多种不同项目一起存在多个类似项目。类似项目可为基本上类似或相同的。例如，类似项目可为具有共有序列的染色体、具有共有序列的染色体的片段、具有共有序列的染色体的至少一部分的拷贝或具有共有序列的其它核酸分子。类似项目的不对称或不平衡样品可通过把样品离散成其组分的比例与初始群体中的比例不相同的部分、等分试样、亚组等进行制备。类似项目的不对称分布为例如两个亲代染色体贡献的分布（例如一个母源染色体和一个父源染色体），这种分布导致部分中两个亲代染色体贡献的不相等分布例如 0.5:1、1:1.5、1:2、1:3、2:3 等比例。部分、等分试样、亚组等可为例如管、孔（例如在微量滴定板中）、微阵列的特征、表面或基底的斑点、珠或颗粒等。

[0016] 应该理解，样品的不对称、不平衡或偏倚可为相对特征，或者可以相对的方式测定。例如，样品可具有染色体或染色体组分的不对称、不平衡或偏倚，其特征为染色体或染色体组分的量不同于存在于所述样品源自的个体、组织或细胞的染色体或染色体组分的量。这样，应该理解，样品源自的个体、组织或细胞可具有至少一种染色体或染色体组分数量的天然存在的不对称、不平衡或偏倚，而样品可偏倚以具有至少一种染色体或染色体组分数量的非天然存在的不对称、不平衡或偏倚。

[0017] 图示

图 1 显示用于产生包含不平衡分布的父系和母系染色体组分的遗传物质池的实施方案。

[0018] 图 2 显示来自父母两者的染色体混合群体的实例和测定混合群体的单体型的挑战。

[0019] 图 3 显示示例性的染色体群体及其在测定单体型方面的用途。

[0020] 图 4 证实对于实践本文描述的方法可得到的包括遗传物质的不平衡分布的示例性基因分型信息。

[0021] 图 5 证实与自给定的试验产生有用信息的可能性（即可测量差异的概率）相比较的示例性加载百分数（预期加载的目标分子数目 / 试验孔或位置 x 100)。

[0022] 图 6 证实用于产生具有两个代表性等位基因即等位基因 A 和等位基因 B 的遗传物质的不平衡分布的偏倚扩增方法的实施方案。

[0023] 图 7 证实用于产生遗传物质的不平衡分布的模板偏倚降解的方法的实例。

[0024] 图 8 证实用于产生具有两个代表性的等位基因即等位基因 A 和等位基因 B 的遗传物质的不平衡分布的偏倚降解的方法的实施方案。

[0025] 图 9 显示正常二倍体个体的荧光原始强度 (raw intensities) 的示例性散点图和本文描述的方法把杂合 SNPs 拆分为其单倍体组分的能力。

[0026] 图 10 显示从源自图 9 的二倍体样品的 6 个 12 倍稀释的样品任意指定 A （在 Y 轴）和 B （在 X 轴）的两个位点的荧光原始强度的一系列示例性散点图。

[0027] 图 11 显示来自从顶面板 (top panel) 的细胞 HG01377 （顶部）和 NA18507 （底部）和底面板 (bottom panel) 的融合的单体型模块 (blocks) （分别为 HG01377 和 NA28507）衍生的不平衡遗传物质池的比对区段 (aligned segments)。

[0028] 图 12 显示来自从细胞 NA18506 （顶面板）和底面板的合并单体型域衍生的正常个体的整个人基因组的不平衡遗传物质池的匹配段。

[0029] 实施方案的详述

本公开的实施方案提供用于测定生物样品的单体型的方法和系统。特定的实施方案提供用于基因组的远程单倍体分型的方法。单倍体分型基因组的重要性例如在有助于和驱动个性化的卫生保健系统以及有助于成功的器官和组织移植方面具有深远的意义。

[0030] 常规的基因分型方法（例如微阵列、测序、PCR 等）在测定单染色体的单体型中，特别是当关注的序列位于染色体上距离很远处时面临困难。例如，微阵列和 PCR 分析如目前实践的那样一般不提供单倍体分型信息，只是序列的存在或不存在。第一代测序技术如目前实践的那样，比如基于毛细管的序列分析方法，可以能够依系统而定检测近端例如 1000bp 或者更少范围内的关注的序列。下一代测序如目前实践的那样，落在关于测定远程单体型的下一代测序 (NGS) 方法的可量测性之间的某处，已经受到相对短的测序读取（例如依系统而定几百个碱基对）的限制。本文描述的实施方案通过在基因组中提供相邻或近端和远端或远程等位基因的取相，填补由这些以上提及的技术留下的缺口。的确，本文描述的实施方案特别适合于鉴定远程单体型。这些方法特别是很好地适合于鉴定具有长于以所使用的特定技术检测的核酸片段长度的范围的单体型。例如，本文阐述的方法的基于 NGS 的实施方案可用于鉴定具有长于所采用的 NGS 技术的读取长度的范围的单体型。发现从通过实践本文描述的方法提供的取相等位基因获得的信息，在例如疾病检测和个性化的卫生保健 (PHC) 方面具有用途。例如个体的单体型可能与药物代谢、药物发现、疾病状态、癌症、障碍、移植排斥反应的风险等相关。的确，关于个性化的卫生保健，一旦受试者的取相的单体型为已知，那么受试者的特定疾病相关性和治疗选择可进行专门设计，以满足所述受试者的需要。

[0031] 本文描述的实施方案与其它用于单倍体分型的方法相比较提供更好的选择。本公开提供例如易于使用、适合于高通量应用和具有取相远程等位基因的能力的方法，而不管样品为单倍体还是二倍体，和不管样品对于关注的等位基因是纯合的还是杂合的。

[0032] 在图 1 中举例说明产生用于单体型测定的遗传物质池的实施方案。产生用于很大一部分基因组或染色体，具有不平衡分布的母系和父系染色体组分的遗传物质池的方法的一个实施方案包括利用泊松随机性 (Poisson randomness)，以产生遗传物质的不相等分布（左箭头）。例如，正常的 DNA 样品具有 1:1 比例的母系：父系染色体。该样品可通过实践本文公开的方法分开，以产生除了 1:1 比例之外例如至少 1:0.5、至少 1:2、至少 1:3、至少 1:4、至少 2:1、至少 2:3 等的母系：父系染色体（或反之亦然），因此为不平衡分布的染色体。

[0033] 在图 2 和 3 中举例说明包括利用泊松随机性，以产生不相等分布的遗传物质的本公开的实施方案。基因型分型样本可由来自双亲两者的染色体的混合群体组成（图 2A)。尽管可能对患者测定基因型（图 2B)，这种类型的分析将不显示杂合性等位基因如何在染色体上组在一起。在该实施例中，不知亲代 A (Parent A) 是否在基因 α 和 γ 提供示例性的 (-) 等位基因两者，和亲代 B (Parent B) 提供示例性的 (+) 等位基因（图 2C)，或者是否它们为混合的（图 2D)。测定单体型的一种方法包括把每一个染色体分离到其自己的隔室（图 3D)，并将其作为单独的样品处理。这样，每一个样品在所有的等位基因为纯合的，因为仅在隔室中存在每个基因的一个拷贝。然而，该方法的不利条件是将存在许多空的试验孔（图 3C)（然而，空孔对于用作阴性试验对照可为有利的)，并且来自具有单染色体的孔的信号可能很低。本文阐述的方法以较高浓度和不对称分布在那些部分提供以部分比如试验孔或隔室存在的染色体样品。只要例如与显示相等数目的亲代染色体的图 3A 形成对比，存在来自每个双亲的不等数目的染色体（或具有源于染色体的序列的核酸分子）（图 3B)，来自具有更大数目的染色体的等位基因可呈现更高的检测信号（例如荧光、发光等)，并且从而相互关联，允许测定不同染色体的单体型（图 3E)。

[0034] 可以预见的是，实践本分开的具体方法所估计的改进可导致与现有技术相比较，加载密度增加至 2-3x 和来自给定试验的总可用数据增加至 5-6x（图 4 和 5)。例如，图 4A 证实可自标准稀释法测定得到的基因型分型信息的程度，其中染色体在测定中被稀释至单分子水平。仅有其中存在一个染色体的那些试验孔将提供有用的数据，例如 P_a=1，P_b=0 或反之亦然。相反，有用信息量的大幅度增加起因于实践本文描述的方法的实施方案，因为，例如，可使用任何数目的染色体 / 每体积，只要两个不同等位基因之间的检测差异大于测量阈值 θ (theta)（图 4B)。

[0035] 因为实践所公开的方法的实施方案可导致对于给定数目的部分加载密度较大和每体积或部分产生数据的概率较高，单倍体（即单倍体基因组）的覆盖范围与实践其它方法比如 0-1 稀释法相比较应更高（图 5)。例如，可在 24% 加载下发现对于 0 或 1 稀释情况下（例如如在图 5A 中举例说明的那样）的最大值，仅有 36% 的试验孔产生有用的数据。或者，图 5B 证实，如本文公开的不对称加载方法可提供最多 100% 加载，76% 的试验孔产生有用的数据。考虑到检测系统的分辨率或灵敏度影响需要提供有用数据的试验部分的数目。目标分子（即染色体组分）包括整个染色体、染色体的片段、克隆的染色体插入物比如于 BACs、YACs、MACs、F 粘粒、粘粒等中发现的那些。进一步地，所公开的方法与 0-1 稀释法相

比较，可有效地提供给较少的部分同等覆盖范围的单倍体。

[0036] 在一个实施方案中，偏倚或不平衡的扩增方法包括用于扩增等位基因，具有不同效率的引物和 / 或扩增条件，使得一组取相的等位基因在扩增的群体中是可区别的，考虑用于产生遗传物质的不平衡分布（图 1， 中间箭头）。偏倚或不平衡的扩增比如偏倚的或不平衡的聚合酶链反应 (PCR)，可通过例如阻断（部分地）其中一个等位基因的扩增，用于产生两个等位基因的不平衡分布。例如一个实施方案包括使用阻断探针，比如在 Rex et al. (2009, J. Virol. Meth. 158:24-29) 和 Senescau et al. (2005, J. Clin. Micro. 43:3304-3308) 中描述的探针（其两者通过引用以其全部结合到本文中）。例如阻断探针可为其中一个等位基因的补体（图 6A， 顶端反应 ; 阻断探针显示跨越 A 核苷酸），具有与 PCR 的延伸温度 (extension temperature) 适配的 Tm，和具有防止其通过 DNA 聚合酶延长的 3’阻断基团。一旦 DNA 聚合酶（例如非链置换 (non-strand displacing)）遇到探针，链延长 (strand elongation) 停止，导致最终 PCR 产物混合物中的一个等位基因表现度减少。相反，其它等位基因的链延长将不会由于存在阻断探针受到阻碍，从而导致最终 PCR 产物混合物中的所述等位基因表现度正常，从而造成 PCR 产物混合物中的一个等位基因的表现度偏倚（图 6A， 等位基因 B 比等位基因 A 更多）。

[0037] 在另一个实施方案中，偏倚或不平衡的扩增方法包括热稳定的 MutS 蛋白和等位基因 - 特异性探针，例如等位基因特异性阻断探针，这种探针在扩增反应中产生不平衡的遗传物质池（图 6B)。MutS 为 DNA 错配结合蛋白，其在 Mg^{2+} 存在下强烈结合于异源双链 DNA (Lishanski et al., 1994, Proc. Natl. Acad. Sci. 91:2674-2678; Stanislawska-Sachadyn and Sachadyn, 2005, Acta Biochim. Pol. 52:575-583; 其两者通过引用以其全部结合到本文中）。例如，为一个等位基因的补体的等位基因特异性阻断探针可退火以与模板 DNA 分子结合，与两个等位基因模板形成同源双链 DNA 和异源双链 DNA 两者。MutS 可优先结合于已与非补体等位基因配对的阻断探针（图 6B 顶端反应 ;异源双链形成显示在 B 等位基因中和 MutS 结合作为圆形显示在底端反应中）。通过使用链置换 DNA 聚合酶（例如 phi29 DNA 聚合酶、BST DNA 聚合酶大片段、Vent® （外 -) DNA 聚合酶、Deep Vent® （外 -) DNA 聚合酶、9°N_m DNA 聚合酶等），可去除未通过 MutS 结合的探针（例如通过使用抗 -MutS 的阴性抗体选择），以允许完美匹配的模板分子的链延长，而 MutS- 复合的探针依然存在于适当的位置，从而停止错配模板分子的链延长，从而在最终产物混合物中产生等位基因的不平衡表现度（图 6B， 等位基因 A 比等位基因 B 更多）。

[0038] 在另一个实施方案中，偏倚或不平衡的扩增方法通过图 6C 举例说明。在图 6C （顶部组的等位基因）中，短探针可杂交至位点的任何一侧。对于那些匹配特定等位基因的探针，可发生探针的延伸和连接。然而，当探针和等位基因为非同源时，没有或存在探针的最小延伸和连接（来自顶部第二组等位基因）。在延伸和连接后，可升高温度，使得已经延伸和连接的那些探针将保持杂交至模板，而没有延伸的短探针将自模板释放（第三组等位基因）。杂交和延伸的探针可交联至模板，从而阻断 PCR 扩增，导致一个等位基因比另一个更多（在这种情况下，等位基因 B 比等位基因 A 更多）。

[0039] 在另一个实施方案中，偏倚或不平衡的扩增方法通过图 6D 举例说明。图 6D 显示等位基因特异性的 PCR 的使用，其中引物之一在靠近多形态位点（即 SNP 或其它多态性的位置）于其 3’末端退火。错配的引物将不引发复制，而匹配的引物可以复制，这样导致一

个等位基因比另一个更多（图 6D，等位基因 A 比等位基因 B 更多）(Newton, 1989, Nucl. Acid. Res. 17:2503-2516; 通过引用以其全部结合到本文中）。

[0040] 在一个实施方案中，产生遗传物质的不平衡分布包括等位基因的偏倚降解（图 1，右箭头）。例如，模板可在引物之间的两个位点（例如示例性位点包括 ATACC 和 TTGTC）上于等位基因 - 特异性位置消化，使得仅有一个等位基因（例如未消化的等位基因）扩增，并且扩增链上的所有等位基因因此共享相同的相（图 7）。可把样品分成几个独立的部分 (A、B 和 C)。一些位点在等位基因靶标 (A 和 G) 为杂合的 (7A)，其中在降解之后生成的群体将超过代表的单一单倍体组分（在该实例中为位点 TTGTC 和等位基因 G)，从而允许区域中的所有等位基因在例如把单独的反应索引和排序后取相。一些位点在等位基因靶标（等位基因 T) 为纯合的（例如 7B 和 C)，或者在两个单倍体染色体贡献之间产生同等扩增的群体 (7B) 或者很少甚或没有扩增 (7C, 等位基因 C)。

[0041] 图 8 证实用于偏倚降解方法的几个示例性实施方案。作为图 6B 的示例性修饰，图 8A 证实，完全匹配的双链分子可用例如双链特异性核酸酶 DSN 选择性地破坏，而 MutS- 结合的错配双链被保护免于裂解。图 8A 证实热稳定的 MutS 蛋白（圆形）、等位基因特异性探针和双链特异性核酸酶（剪刀）的使用，其中双链特异性核酸酶可对等位基因 B 超过对等位基因 A 的偏倚扩增裂解同源双链 DNA。

[0042] 在另一个实施方案中，偏倚降解方法包括对于单核苷酸错配具有强的靶标位点倾向的噬菌体 Mu 转座子 (Yanagihara and Mizuuchi, 2002, Proc. Natl. Acad. Sci. 99:11317-11321; 通过引用以其全部结合到本文中）和等位基因特异性探针。Mu 本身可伴随错配优先插入异源双链 DNA 中，使得其在例如库制备方案方面的用途（图 8B, 作为圆形显示的 Mu 转座子）可用于使错配等位基因的模板分子破裂，而完美匹配的等位基因的模板分子保持完整并用作 PCR 扩增的模板，从而产生偏倚或不平衡的基因池用于单体型测定（图 8B, 等位基因 A 比等位基因 B 更多）。

[0043] 在另一个实施方案中，偏倚或不平衡的扩增方法通过图 8C 举例说明，其为图 8B 的修饰。在图 8C 中，生物素化的等位基因特异性探针（对于 B) 被显示杂交于模板 DNA。链霉抗生物素转座子融合蛋白（例如如在来自 Epicentre Biotechnologies 的 NextEra DNA 样品制备试剂盒中举例说明的用圆形指定的 Mu 转座子）可通过链霉抗生物素 - 生物素相互作用募集到双链杂交位点，从而导致完美匹配的等位基因破裂和一个等位基因比另一个更多（图 8C, 等位基因 B 比等位基因 A 更多）。

[0044] 在另一个实施方案中，偏倚降解方法可包括限制性内切核酸酶，如在图 8D 中证实的那样。例如，可选择一种或多种限制性内切核酸酶，使得存在约一个限制位点 / 每个扩增子对（例如通过靶向已知的杂合位点或通过基于扩增子长度的统计学）。包含靶向位点的扩增子可被降解（即通过在圆形指定的限制性内切核酸酶受到限制)，使得扩增为不可能的。未消化的等位基因可优先扩增，产生表现度不等的等位基因用于单体型测定（图 8D，等位基因 A 比等位基因 B 更多）。

[0045] 本公开提供用于测定基因组的单体型的方法。在一个实施方案中，本公开的方法自受试者的二倍体或单倍体基因组样品产生遗传物质（即染色体组分）的不平衡分布。用标准方法（例如微阵列、测序、PCR、基于凝胶等）对不平衡的遗传物质进行基因型分型，使得能够对于远程单倍体分型在大的基因组区域测定单体型。例如，当对于遗传物质的不

对称或不平衡分布采用本文描述的方法用于单倍体分型时，如果特定基因组区域中关注的一组引导序列比另一组等位基因扩增信号强度更高 (3x) （例如通过微阵列）或读取更多 (3x) （测序），那么推断两个相应的组对应于两个不同的单体型。不平衡的遗传物质池中关注的每一个引导序列的相对量一旦测定，与自正常二倍体基因组或汇集的正常基因组测定的量进行比较，从而测定受试样品中的异常现象。

[0046] 本公开提供包括样品的不相等、不平衡、偏倚或不对称分布，用于单体型测定的方法。不相等分布可为例如稀释、不对称 PCR、靶标降解等的结果。特别是，本文描述的实施方案在各部分比如基底上的测试位置（例如板上的孔、玻片上的区域、多个毛细管、柔性带中 / 上的孔等）之间提供分布不均的来自受试者的遗传物质。在某些实施方案中，样品的遗传物质的不均匀分布代表位于基底上一个或多个测试位置的染色体的分布不等。考虑一些测试位置不含遗传物质，并且发现这些位置在如在图 3C 中举例说明的试验中作为阴性对照品具有用途。基底包括但不限于微阵列基底比如二氧化硅或高密度塑料玻片、芯片等、板比如 96、384、1536 孔测定板、毛细管例如如用于流过 PCR 的毛细管、柔性的高通量测试条（例如 Douglas Scientific 的 Array Tape™）、珠粒、纳米颗粒等。本文描述的方法不受在其上或其中实施测试的基底的限制。

[0047] 本文描述的方法的特定实施方案可用于例如测定染色体上彼此近端和远端两者的关注序列的单体型。考虑关注的序列不被任何特定的距离分开，例如关注的序列可在染色体上为彼此相邻或者近端的。相反，考虑关注的序列在染色体上为彼此远端分离的或远程的。的确，实践本文描述的实施方案在测定远程单体型时可为特别有益的。关注的序列之间的距离不打算限制所述方法，例如关注的序列可由至少 100、200、300、400、500、750 或至少 1000 个碱基对分开。然而，实施方案发现，当关注的序列在染色体上间隔离得很远，并且由例如至少 10000、至少 100000、至少 1000000、至少 10000000、至少 100000000、至少 150000000、至少 200000000、至少 247000000 或者更多个碱基对分开时，对测定其单体型特别有用。这样，本文描述的实施方案可提供特别适合于个体基因组的远程单倍体分型的方法，而不管被提供的用于测定的样品是单倍体还是二倍体。

[0048] 在本公开的实施方案中，提供用于测定单体型，特别是位于染色体上远侧的关注序列的方法。在一些实施方案中，关注的序列为单核苷酸多态性，或 SNPs。在一些实施方案中，SNPs 为彼此相邻的或接近的，而在其它的实施方案中，SNPs 为彼此离得很远或远程的。在一些实施方案中，关注的序列为基因组中序列的插入或缺失，或者插入 / 缺失 (indels)。在一些实施方案中，关注的序列为基因组拷贝数目变异，或者 CNVs。在其它的实施方案中，关注的序列为等位基因，或者位于染色体上特定位置的基因或序列的替代形式。在一些实施方案中，等位基因为野生型或正常的识别序列，而在其它的实施方案中，等位基因与野生型相比较可隐匿一个或多个突变，比如 SNPs、CNVs、插入 / 缺失等。

[0049] 这样的突变可被确定为与疾病状态比如癌症、遗传疾病等直接相关。突变的等位基因对于研究者具有特别意义，并且实践本公开的实施方案可在使得研究者能够研究等位基因突变及其单体型方面提供有价值的工具。单体型在定义个体的二倍体基因组的基因组成方面是有价值的。单倍体分型信息可导致更多的理解，并且在许多科学研究领域发现具有更广泛的用途，这些领域包括但不限于药物代谢、药物发现、个性化的卫生保健计划、移植成功群体遗传学的 HLA 分型、复杂疾病连锁、遗传人类学、疾病和癌症的医学遗传学、癌

症和其它疾病的结构变化、等位基因的特异性表达和修饰比如等位基因特异性甲基化模式以及更始基因组(de novo genome)组装。当用于单倍体分型的关注的等位基因来自小的基因组区域时,包括偏倚扩增和偏倚降解的实施方案是特别有利的。这样,临床应用比如其中需要超过几千个碱基或者一个或多个基因组区域的单体型测定的HLA基因型分型(例如HLA-A、HLA-B、HLA-C、HLA-DRB1、HLA-DQB1、HLA-DQA1等),将极大地得益于实践本文公开的方法。

[0050] 把等位基因分配到染色体(即单倍体分型)的能力强大,因为其可例如通过提供关于基因组中重组事件的信息来提供临床相关性的信息。这种信息对于确定引起疾病的突变的位置可为重要的,并可有助于确定连锁不平衡,或者基因组中两个多态性的存在之间的统计关联性,此为疾病基因组广泛疾病关联性研究的一种关键特性。例如,如果两种多态性之间的关联性(即连锁不平衡)高,已知一种多态性(即SNP)的基因型可有助于预测另一种多态性(即SNP)的基因型。通过测定其单体型更完全匹配人白细胞抗原(HLA)的能力将极大地改善例如移植接受者的临床结果(Crawford and Nickerson, 2004, Ann. Rev. Med. 56:303-320, 通过引用以其全部结合到本文中)。例如,通过实践本文公开的方法,移植接受者和潜在供者可沿着主要组织相容性复合体对多个标记进行基因型分型,并可自产生的数据测定单体型。这样的匹配的实例可见于本文公开的实施例中。这样的匹配可提供移植接受者与供者之间高度准确的HLA匹配,导致比不是如此匹配的患者与供者更好的移植结果。

[0051] 另外,存在一些疾病,其中单体型而不是在特定位点的基因型可预测疾病的严重性,这样准确的单体型将不仅对于确定具体患者的疾病严重性具有广泛用途,而且也提供给临床医生基于诊断和/或预后确定有效的治疗选择方面的信息,因为不同的治疗选择可能与不同的疾病状态和/或严重性水平相关。例如,特定的镰状细胞性贫血β-球蛋白位点单体型与不太严重的镰状细胞性贫血有关,并且IL10启动子区域的单体型与移植物抗宿主病和接受细胞移植的患者死亡的发生率较低有关。这样,提供基因组样品的单倍体分型的方法可对例如疾病相关性的研究、疾病诊断和预后实践以及治疗方案的应用具有很大影响。然而,单倍体分型也在农业和其它园艺领域具有重要意义,特别是在其中疾病或有利的性质可能与动物或植物中的特定单体型有关的牲畜饲养和农作物方面。

[0052] 本文提供的实施方案描述用于测定样品中取相的等位基因的方法。通常,样品包括核酸样品。在一些实施方案中,核酸样品源于体液,例如来自受试者的血液、痰液、尿液、脊髓液等。在其它的实施方案中,生物样品源于固体,例如来自受试者的组织、活组织切片检查、细胞刮取、细胞学或细胞样品等。在一个实施方案中,生物样品为纯化的单染色体或其片段,或者例如在粘粒、F粘粒、质粒、酵母人工染色体(YAC)、细菌人工染色体(BAC)、哺乳动物人工染色体(MAC)、植物克隆系统(例如农杆菌(*Agrobacterium tumefacians*)T-DNA克隆系统、双元载体克隆系统等)或其片段中的DNA插入等。在优选的实施方案中,生物样品为如在一种或多种细胞中发现的二倍体DNA样品。然而,本文描述的方法的实施方案不限于二倍体样品,因为单倍体样品(例如源于卵子、精子、水泡样胎块(hydatiform mole)的核酸,和机械分开和/或分离的染色体、其片段、克隆的DNA片段等)同样适用于实践本文描述的方法。

[0053] 在一个实施方案中,样品为细胞样品或组织样品。细胞或组织样品可来自任何来

源，例如来自解离组织的细胞、来自血液或其它体液的细胞、来自细胞学样本的细胞、来自非人动物的细胞、来自植物的细胞等。在优选的实施方案中，细胞为哺乳动物起源的，优选地为人起源的。然而，本文描述的方法不限于细胞样品的来源。在一些实施方案中，用于实践本文描述的方法的基因组材料源于多个细胞。在一些实施方案中，多个细胞为至少2-1000 个细胞之间、至少 5-500 个细胞之间、至少 10-300 个细胞之间、至少 10-100 个细胞之间。除非特别相反地指出，实践本文阐述的方法可采用本领域技术范围内的病毒学、免疫学、微生物学、分子生物学和 DNA 重组技术的常规方法。这种技术在以下文献中得到充分说明：参见例如 1995, Ausubel et al., 精编分子生物学实验指南 (Short Protocols in Molecular Biology), (第3版), Wiley & Sons; 2001, Sambrook and Russell, 分子克隆：实验室手册 (Molecular Cloning: A Laboratory Manual) (第3版); 1982, Maniatus et al., 分子克隆：实验室手册 (Molecular Cloning: A Laboratory Manual); DNA 克隆：一种实用方法 (DNA Cloning: A Practical Approach), 第 I 和 II 卷 (D. Glover 编辑); 1984, 寡核苷酸合成 (Oligonucleotide Synthesis) (N. Gait 编辑); 1985, 核酸杂交 (Nucleic Acid Hybridization) (B. Hames 和 S. Higgins 编辑); 1986, 动物细胞培养 (Animal Cell Culture) (R. Freshney 编辑); 1984, Perbal, 分子克隆的实用指南 (A Practical Guide to Molecular Cloning)。基因组材料可通过本领域已知的方法收获，并且本文描述的方法不一定限于用于分离基因组材料的任何具体方法。技术人员应理解，对于这种分离存在大量的市售和自产的 (homebrew) 替代品。

[0054] 在一个实施方案中，由受试者提供用于单倍体分型的样品。受试者可为对希望测定来自所述实体的单体型的研究者关注的任何生物实体。这样，用于测试的样品不一定限于特定受试者，并且受试者可为例如动物或植物起源的。例如，提供样品的受试者可为动物（人或非人）或植物，例如相关的经济作物等。在优选的实施方案中，受试者为人。在其它优选的实施方案中，受试者为经济相关的动物或其衍生物。在其它的实施方案中，受试者为经济相关的植物或其衍生物。

[0055] 通过实践本公开的方法提供的不对称分布的样品易于应用于下游应用。在一些实施方案中，考虑在测序或其它仪器相关的单体型测定之前对样品实施下游过程。在一些实施方案中，不对称分布的样品的等分试样或部分用于制备群集 (clustering) 准备下一代测序的 DNA 库。例如通过在 Nextera™ DNA 样品制备试剂盒 (Nextera™ DNA Sample Prep Kit) (Epicentre® Biotechnologies, Madison WI)、GL FLX 钛库制备试剂盒 (GL FLX Titanium Library Preparation Kit) (454 Life Sciences, Branford CT)、SOLiD™ 库制备试剂盒 (SOLiD™ Library Preparation Kits) (Applied Biosystems™ Life Technologies, Carlsbad CA) 等实施所描述的方法产生这种库。本文描述的样品一般通过例如多重链置换扩增 (MDA) 技术进行进一步扩增用于测序或微阵列分析。对于 MDA 后的测序，例如通过以如在配对库制备试剂盒 (Mate Pair Library Prep kit)、基因组 DNA 样品制备试剂盒 (Genomic DNA Sample Prep kits) 或 TruSeq™样品制备或外显子组富集试剂盒 (TruSeq™ Sample Preparation or Exome Enrichment kits) (Illumina®, Inc., San Diego CA) 产生所描述的 DNA 库，制备扩增的样品库。有用的群集扩增 (cluster amplification) 方法描述在例如美国专利第 5641658 号、美国专利公布号 2002/0055100、美国专利第 7115400 号、美国专利公布号 2004/0096853、美国专利公布号 2004/0002090、

美国专利公布号 2007/0128624 和美国专利公布号 2008/0009420 中，其每一个通过引用以其全部结合到本文中。另一种用于在表面扩增核酸的有用方法为例如如在 Lizardi et al., Nat. Genet. 19:225-232 (1998) 和 US 2007/0099208 中描述的滚环扩增 (RCA)，其每一个通过引用以其全部结合到本文中。乳液 PCR 方法也是有用的，示例性方法被描述于 Dressman et al., Proc. Natl. Acad. Sci. USA 100:8817-8822 (2003)、WO 05/010145 或美国专利公布号 2005/0130173 或 2005/0064460 中，其每一个通过引用以其全部结合到本文中。本公开的方法不一定受到任何具体的库制备或扩增方法的限制，因为考虑本文描述的样品的不对称分布适用于本领域已知和 / 或对此目的市售可得到的各种方法中的任何一种。

[0056] 例如，包含不平衡分布的遗传物质的 DNA 库可被固定在基底比如流动池上，并在对例如通过合成方法学得到的序列进行测序之前对固定化的多核苷酸实施桥式扩增 (bridge amplification)。在桥式扩增中，固定化的多核苷酸 (例如来自 DNA 库) 被杂交至固定化的寡核苷酸引物。固定化的多核苷酸分子的 3’末端提供给模板自固定化的寡核苷酸引物延伸的，聚合酶催化的，模板定向的伸长反应 (例如引物延伸)。生成的双链产物“桥接”两个引物，并且两个链共价连接于载体 (support)。在下一个周期中，在产生固定于固体载体的一对单链 (固定化的模板和延伸的引物产物) 的变性之后，两个固定化的链可用作用于新的引物延伸的模板。因此，第一和第二部分可被扩增，以产生多个群集。术语“群集”和“集落”可互换使用，并且指的是核酸序列和 / 或其附着于表面的补体的多个拷贝。通常地，群集包含核酸序列和 / 或其通过其 5’末端附着于表面的补体的多个拷贝。示例性桥式扩增和群集方法学被描述在例如国际专利公布号 WO00/18957 和 WO98/44151、美国专利第 5641658 号、美国专利公布号 2002/0055100、美国专利第 7115400 号、美国专利公布号 2004/0096853、美国专利公布号 2005/0100900、美国专利公布号 2004/0002090、美国专利公布号 2007/0128624 和美国专利公布号 2008/0009420 中，其每一个通过引用以其全部结合到本文中。本文描述的组合物和方法在采用包含群集的流动池通过合成方法学得到的序列中为特别有用的。

[0057] 用于在测序之前扩增核酸的乳液 PCR 方法也可与本文描述的方法和系统组合使用。乳液 PCR 包括衔接子侧面鸟枪 DNA 库在油包水乳液中的 PCR 扩增。PCR 为多模板 PCR，仅使用单引物对。PCR 引物中的一个系于微尺度珠 (microscale beads) 的表面 (5’附着)。低的模板浓度导致存在不多于一个模板分子，含有大多数珠粒的乳液微泡。在生产乳液微泡 (其中存在珠粒和模板分子两者的乳液微泡) 中，PCR 扩增子可被捕获于珠粒的表面。在破乳后，可选择性地富集带有扩增产物的珠粒。每一个克隆扩增的珠粒将在其表面带有对应于来自模板库的单分子扩增的 PCR 产物。乳液 PCR 方法的各种实施方案被阐述于例如 Dressman et al., Proc. Natl. Acad. Sci. USA 100:8817-8822 (2003)、国际专利公布号 WO 05/010145、美国专利公布号 2005/0130173、2005/0064460 和 US2005/0042648 中，其每一个通过引用以其全部结合到本文中。

[0058] DNA 纳米球也可与本文描述的方法和系统组合使用。产生和采用用于基因组测序的 DNA 纳米球的方法可见于例如美国专利和出版物 7910354、2009/0264299、2009/0011943、2009/0005252、2009/0155781、2009/0118488，以及如在例如 Drmanac et al., 2010, Science 327(5961): 78-81 中描述的那样，其全部通过引用以其全部结合到

本文中。简言之，在衔接子连接的基因组DNA片段连续来回之后，扩增和消化导致被环化为单链DNA（例如通过用圆形连接酶(circle ligase)连接）和滚环扩增（例如如在Lizardi et al., Nat. Genet. 19:225-232 (1998)和US 2007/0099208 A1中描述的那样，其每一个通过引用以其全部结合到本文中）的圆形基因组DNA模板/衔接子序列的多个拷贝的首尾相接的串联体。所述串联体的衔接子结构促进单链DNA的盘绕，从而产生紧密的DNA纳米球。DNA纳米球可被捕获于基底上，优选地产生有序或图形排列，使得保持每一个纳米球之间的距离，从而使得能够将单独的DNA纳米球测序。

[0059] 在一些实施方案中，一旦不对称分布的样品得到进一步处理，将其应用于测序、微阵列分析、基因型分型或其它下游应用。例如，测序可按照制造商的方案，在系统比如由Illumina, Inc. (HiSeq 1000, HiSeq 2000, 基因组分析仪(Genome Analyzers), MiSeq, HiScan, systems（系统））、454 Life Sciences (FLX基因组测序仪(FLX Genome Sequencer), GS Junior)、Applied Biosystems™ Life Technologies (ABI PRISM® 序列检测系统(Sequence detection systems), SOLiD™ System)、Ion Torrent® Life Technologies（个人基因组机械测序仪(Personal Genome Machine sequencer)）提供的那些系统、进一步如在例如美国专利和专利申请5888737、6175002、5695934、6140489、5863722、2007/007991、2009/0247414、2010/0111768和PCT申请号WO2007/123744中描述的那些系统上进行，其每一个通过引用以其全部结合到本文中。

[0060] 在一些实施方案中，发现本文描述的用于测定单体型的方法在用于测序，例如合成测序(SBS)技术时具有特别的用途。合成测序通常包括使用聚合酶依序增添一个或多个标记的核苷酸，以使多核苷酸链在5’至3’方向生长。延伸的多核苷酸链与可附着于基底（例如流动池、芯片、玻片等）上，并含有引导序列的核酸模板互补。用于SBS的标记的核苷酸可包括各种荧光团、质量标记、可电子检测的标记或其它类型标记中的任何一种。用于SBS的标记的核苷酸也可包括可逆性的终止基团，使得每个SBS循环仅增添一个核苷酸。在所结合的核苷酸被检测之后可加入解封剂，以提供增添的适合于在随后的循环中延伸的核苷酸。SBS方法对于核酸样品的不同序列片段的平行分析特别有用。例如数百、数千、数百万或者更多的不同序列片段可使用已知的SBS技术在单一基底上同时进行测序。示例性的测序方法被描述于例如Bentley et al., Nature 456:53-59 (2008)、WO 04/018497、US 7057026、WO 91/06678、WO 07/123744、US 7329492、US 7211414、US 7315019、US 7405281和US 2008/0108082中，其每一个通过引用以其全部结合到本文中。

[0061] 也发现所公开的用于测定单体型的方法在用于连接法测序、杂交测序及其它测序技术时具有用途。示例性的连接法测序方法学为应用生物系统公司的(Applied Biosystems’) SOLiD™测序系统采用的二元化编码（例如色彩空间测序）(Voelkerding et al., 2009, Clin Chem 55:641-658; 通过引用以其全部结合到本文中）。

[0062] 用于本文公开的单倍体分型的方法可通过杂交技术用于测序。杂交测序包括使用向其增添分裂成碎片的标记的目标DNA的一些列短序列的核苷酸探针（例如，如在Drmanac et al., 2002, Adv Biochem Eng Biotechnol 77:75-101; Lizardi et al., 2008, Nat Biotech 26:649-650, 美国专利7071324中描述的；通过引用以其全部结合到本文中）。对杂交测序的进一步改进可见于例如美国专利申请出版物2007/0178516、2010/0063264和2006/0287833中（通过引用以其全部结合到本文中）。结合杂交与连接生物化学的测

序方法已得到开发和商业化，比如由完整的染色体组，高原病展望 (Complete Genomics, Mountain View), CA) 实践的基因组测序技术。例如，组合的探针 - 锚定序列连接方法或 cPAL™ (Drmanac et al., 2010, Science 327(5961): 78-81) 采用连接生物化学，同时利用杂交测序的优势。单分子测序技术，例如如在 Pushkarev et al. (2009, Nat. Biotechnol. 27:847-52; 通过引用以其全部结合到本文中）描述的和如由 HeliScope™单分子测序器 (Helicos, Cambridge, MA) 实践的单分子测序技术，也可利用所公开方法的优势用于测定单体型。

[0063] 本文描述的方法不受到任何特定测序样品制备方法的限制，并且备选方法对技术人员是显而易见的，并考虑在本公开的范围内。然而，发现在本文的方法应用于以下测序装置时具有特殊的用途：比如流动池或阵列，其用于实践合成测序方法学或其它相关的测序技术，比如聚合酶测序技术 (polony sequencing technology) (Dover Systems)、通过杂交荧光平台测序 (Complete Genomics)、sTOP 技术 (Industrial Technology Research Institute) 和合成测序 (Illumina, Life Technologies) 中的一种或多种实践的那些测序技术。

[0064] 在一些实施方案中，本文描述的不对称分布的样品经 MDA 处理，并进行进一步处理用于微阵列和 / 或其它基因型分析试验。例如，在一些实施方案中，样品经定量 PCR (qPCR) 处理，以信噪比表征各部分或等分试样（例如通过采用 Eco PCR 系统 (Illumina®, Inc.))。这种表征在定义自下游测序或微阵列分析潜在提供最高概率的可判断数据的部分或等分试样方面是有用的。在一些实施方案中，进行进一步处理用于微阵列分析之前的制备。例如，不对称分布的样品在经 MDA 扩增和 / 或经 qPCR 表征之后进行制备，用于经各种方法进行微阵列分析，所述方法包括但不限于以上对库样品制备先前描述的那些。

[0065] 有用的示例性微阵列包括但不限于可得自 Illumina®, Inc. (San Diego, CA) 的 Sentrix® Array 或 Sentrix® BeadChip Array，或者其它孔中包含珠粒的微阵列，比如在例如美国专利第 6266459、6355431、6770441 和 6859570 号和 PCT 公布号 WO 00/63437（其每一个通过引用以其全部结合到本文中）中描述的那些微阵列。

[0066] 其它表面上具有颗粒的阵列包括在 US 2005/0227252、US 2006/0023310、US 2006/006327、US 2006/0071075、US 2006/0119913、US 6489606、US 7106513、US 7126755、US 7164533、WO 05/033681 和 WO 04/024328（其每一个通过引用以其全部结合到本文中）中阐述的那些微阵列。用于测试如通过实践本公开的方法提供的不对称分布的样品的一系列珠粒也可呈流动格式 (fluid format)，比如流式细胞分析仪或类似装置的液流。用于区分珠粒的市售可得到的流动格式包括例如用于来自 Luminex 的 XMAP™技术或来自 Lynx Therapeutics 的 MPSS™ 方法的那些流动格式。

[0067] 可与通过实践本公开的方法提供的样品一起使用的，市售可得到的微阵列的其它实例包括例如 Affymetrix® GeneChip® 微阵列，或按照如例如在以下文献描述的有时称为 VLSIPS™（极大尺度的固定化聚合物合成 (Very Large Scale Immobilized Polymer Synthesis)）技术的技术合成的其它微阵列：美国专利第 5324633、5744305、5451683、5482867、5491074、5624711、5795716、5831070、5856101、5858659、5874219、5968740、5974164、5981185、5981956、6025601、6033860、6090555、6136269、6022963、6083697、6291183、6309831、6416949、6428752 和 6482591（其每一个通过引用以其全部结合到本文

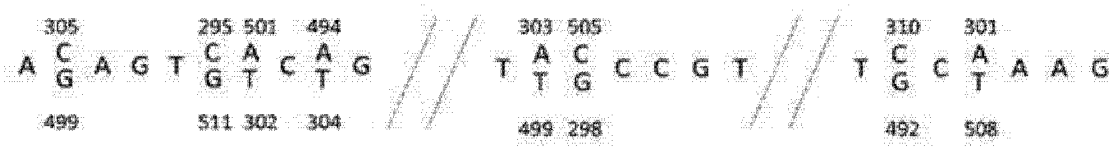
中）。

[0068] 点样微阵列也可与通过实践本公开的方法提供的样品一起使用。示例性的点样微阵列为可得自安玛西亚公司（Amersham Biosciences）的CodeLink™ Array （阵列）。有用的另一种微阵列为使用喷墨印刷法比如可得自安捷伦科技（Agilent Technologies）的SurePrint™ Technology制作的微阵列。可使用的其它微阵列包括但不限于在Butte, 2002, Nature Reviews Drug Discov. 1:951-60或美国专利第5429807、5436327、5561071、5583211、5658734、5837858、5919523、6287768、6287776、6288220、6297006、6291193和6514751号及WO 93/17126 和WO 95/35505 （其每一个通过引用以其全部结合到本文中）中描述的那些微阵列。

[0069] 来自测序、微阵列或其它基因分型方法学或仪器的输出可具有任何方式。例如，一些技术采用生成可读输出的光，比如荧光或发光，而其它技术测量电子或离子的释放。然而，本发明不限于可读输出的类型，只要可对关注的特定序列测定输出信号的差异。可用于表征源于实践本文描述的方法的输出的分析软件的实例包括但不限于Pipeline, CASAVA, 基因组Studio数据分析（Genome Studio Data Analysis), BeadStudio Genotyping and KaryoStudio数据分析软件（Illumina®, Inc.）、SignalMap and NimbleScan数据分析软件（Roche NimbleGen）、GS Analyzer分析软件（454 Life Sciences）、SOLiD™, DNASTAR® SeqMan® NGen® and Partek® Genomics Suite™数据分析软件（Life Technologies）、特征提取和Agilent染色体组工作台（Feature Extraction and Agilent Genomics Workbench）数据分析软件（Agilent Technologies）、Genotyping Console™, 染色体分析研究和基因芯片序列分析（Chromosome Analysis Suite and GeneChip® Sequence Analysis）数据分析软件（Affymetrix®）。技术人员应了解用于微阵列、测序和PCR产生的输出的数据分析的另外众多的商业和学术上可用的备选软件。本文描述的实施方案不限于任何数据分析方法。

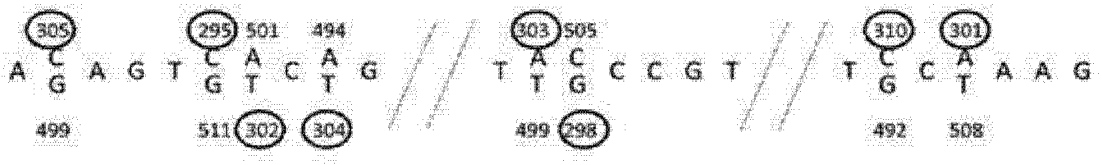
[0070] 本公开的示例性方法不一定受到任何特定的测序、微阵列或基因分型系统的限制，因为考虑对于特定仪器要求的特定样品制备适合用于本文描述的不对称分布的样品。然而，考虑任何给定检测系统的分辨率或灵敏度可影响可被测试以产生可判断的结果的部分的数目。在图3B （κ）和图4B （θ）中举例说明分辨率差异。

[0071] 以下实施例描述用于通过采用不对称产生的样品进行测序测定SNP单体型的方法。在该具体实施例中，采用低输入DNA水平（例如10-100pg）的制备方法比如Nextera™ DNA样品制备试剂盒是特别有用的，因为用这种试剂盒处理的样品适合准备测序，并且不需进一步处理，比如多链置换扩增。另外，可需要另外的扩增步骤，比如MDA。所制备的样品可例如在Illumina, Inc. 基因组分析仪（Genome Analyzer), HiSeq, MiSeq, TruSeq或其中产生对应于每个荧光标记的核苷酸的荧光读数用于分析的其它测序平台上进行测序。对于该实施例的目的，自不对称分布的样品制备得到以下测序结果：



在该实施例中，用双散列线（double hash lines）自不连续和可能远离地位于染色体

区域分离单个位点的核酸。对一个位置列出的两个核苷酸代表杂合序列变异或关注的序列中的单核苷酸多态性 (SNPs)。核苷酸上面和下面的数目代表出自读取总数，例如在这种情况下读取约 800 的特定核苷酸位置的读取数目。远程 SNP 取相通过匹配具有如下相似读数的 SNP 位置进行测定：



在该实施例中，圆圈中的数字代表不同 SNP 位置的相似读数，并因此确定哪一个 SNPs 位于同一染色体或染色体片段或分段上并因此为同相，从而测定样品的单体型。这样，对多个 SNPs 计数读取的数目并匹配那些读取计数，可用于测定样品单体型。对两个亲代染色体贡献（例如顶部序列为母系贡献和底部序列为父系贡献）的单体型被测定为：

为了解释的目的，可以类似的方式举例说明通过微阵列进行的单倍体分型，除了代替核酸序列输出，可提供对应于每一个 SNP 和内含子的杂交强度计算的模拟值的数字衍生的颜色读取。

[0072] 在一些实施方案中，不对称分布的样品可在任何另外的处理比如库制备或扩增之前，或在如先前举例说明的那样采用 Nextera™试剂盒进行库制备之前进行表征。例如，样品如在实施例 1 中见到的那样分成多个部分或等分，并且每个部分在测序或微阵列分析中进行单独处理。如在实施例 1 中描述的那样，如果样品被离散成 10 个部分，那么对一份原始样品可进行 10 个下游处理。把样品分成多个部分或等分试样提供许多优势，包括但不限于对一份样品的多次分析以及较低的成本和精力（例如试剂和其它耗材、研究者时间等）。为了进一步降低成本和精力，可在对具有最高不对称性、最高信噪比和所需目标的最高覆盖范围的那些样品进行分析之前对多个部分进行表征和 / 或量化。例如，对多个序列进行部分的基于 qPCR 的基因型分型应足以测定部分的不对称性和信噪比。进一步地，微阵列分析或低深度测序方法也可用于测定部分的不对称性和信噪比。考虑仅有例如具有最高信噪比的那些部分的单体型分析，以提供产生可判断结果最高的概率，从而节省时间、精力和金钱。

[0073] 技术人员应意识到，用于单倍体分型（例如测序、微阵列分析、qPCR、PCR 等）的不同检测系统的分辨率不同。因此，可以预见的是，当决定不对称程度、信噪比等将提供由任何给定系统产生的最有用数据时考虑给定系统的分辨极限。

[0074] 提供以下实施例以证实和进一步说明本发明的某些实施方案和方面，并且不视为对其范围的限制。

实施例

[0075] 实施例 1 :样品的不对称分布

在评价用于测定基因组的远程单倍体分型的方法中，确定样品以来自染色体的根本贡献（例如源于母系的染色体贡献和源于父系的染色体贡献）的信号相互区别的这样一种方式分布，在基因组的成功远程单倍体分型方面提供优越的结果。

[0076] 以下方法为如何产生可用于测定远程单体型的样品的不对称分布的示例。染色体 6 (Chr 6) 为示例性染色体，然而应该理解，可使用来自任何组织、细胞类型、细胞系（永生或原代的）等的任何染色体来源的遗传物质。

[0077] 样品含有分别称为 M6 和 P6 的 Chr 6 的母系和父系贡献的混合物。对于该实施例的目的，样品源于同步到中期的细胞。中期同步不需要实施所描述的方法，然而因为该实施例证实对两个亲代贡献测定取相的等位基因为要完成的一种方式，这开始于同步到中期的细胞。

[0078] 样品细胞数目通过荧光激活的细胞分选或 FACS、细胞计数测定或其它已知方法进行测定。一旦测定了样品细胞数目，把样品以约 10 μl 的最终总细胞体积稀释至约 100 细胞 /μl。因为平均起来细胞每一个 M6 和 P6 含有一个拷贝，两个贡献的比例为 1:1。在稀释之后，把细胞经技术人员已知的建立的技术进行细胞溶解并收获 DNA。对于该实施例的目的，把样品分成 10 个部分，把 DNA 样品离散成 15 nl 的 10 个部分，得到最佳概率的含有 1.5 个染色体的受试样品（图 5），提供总数约 670 的潜在受试部分。最终部分的体积将例如依 DNA 收获之前的细胞浓度差异、目标染色体组分（例如在这种情况下为 M6 和 P6）的差异以及用于下游分析（例如测序、微阵列分析、PCR 等）的测量技术的灵敏度而变化。用于把样品等分或分成几个部分的方法可以不同，例如对于该实施例的目的，微流体装置用于把样品分流到所需数目的测试室用于下游应用，然而，如果要求的体积在所述方法的范围内，任何手动或自动的分流样品的方法均为合适的。

[0079] 如果不需要来自每个染色体的单体型数据，例如如果需要来自几个染色体的单倍体分型数据，图 5 可用于测定下游处理要求的部分数目，以得到概率最高的可判断数据。在本实施例中，对于含有 1.5 个来自 Chr 6 的染色体的样品部分，可以预见 72% 的部分将产生具有足够不对称性的数据，以允许区别 M6 等位基因与 P6 等位基因检测信号（例如用于取相等位基因），或反之亦然。因此，少至 10 个 15 nl 的部分在这些部分中的至少一个具有 99.99% 的概率提供可用数据（例如基本组分的不对称性大于下游过程比如测序、微阵列分析、PCR 等的分辨率）。然而，考虑对于测试将产生可判断的数据的图表字符或部分的数目依用于数据采集的方法的分辨率而变化（例如测序方法与微阵列方法与 PCR 方法等相比较）。

[0080] 一旦分流，样品按照研究者的需要（例如测序、微阵列分析、外遗传分析、PCR 等）进行处理。如先前描述的那样，不对称分布方法的结果提供许多样品等分试样或部分，其全部可直接由研究者进行分析，或者适当存储供以后分析。研究者可能仅希望分析含有最高程度不对称性的样品等分试样，从而提供最高的信噪比，此时研究者可对所述特性表征这些部分，并仅使用在这一点上满足研究者的需要的那些部分。可用于分析不对称分布的样品部分的示例性下游应用包括但不限于如先前讨论的 DNA 库制备、扩增（例如 PCR、qPCR、MDA 等）、微阵列分析、测序、基因型分型和单倍体分型。

[0081] 实施例 2 ：使用不对称的样品分布方法的单体型的测定

把来自正常个体的人类基因组 DNA 稀释至 0.5 个单倍体拷贝每 3 μl 水 (5.00E-07 μg/

μl)。把所稀释的基因组 DNA (3 μl) 等分到多个管中,导致每管平均含有 0.5 个人类基因组的单倍体拷贝。向每个管中加入 3 μl 的缓冲液 D2 (含有 0.25 μl 1M DTT 的 2.75 μl DLB 缓冲液) (Qiagen REPLI-g® UltraFast Mini Handbook, Catalog # 150035),随后在 BioRad DNA Engine 热循环仪 (BioRad Part # PTC-0200G) 中于 4℃下温育 10 分钟。加入 3 μl 的 REPLI-g UltraFast 停止液,随后加入 33 μl 的 Mastermix。Mastermix 含有 30 μl REPLI-g UltraFast 反应缓冲液、2 μl REPLI-g UltraFast DNA 聚合酶和 1 μl 含有 6000 寡核苷酸 (最终浓度为 0.03 μM 每寡核苷酸) 的 7.56 mM 人 9-mer 池。把反应物在 BioRad Tetrad2 热循环仪 (BioRad Part # PTC-0240G) 中于 30℃下温育 90 分钟,随后通过在 65℃下把样品加热 3 分钟,使 REPLI-g UltraFast DNA 聚合酶热灭活。

[0082] 多重置换扩增 (MDA) 产物使用 DNA Clean & Concentrator™-5 离心柱 (Zymo Research Catalog # D4003),按照制造商的方案进行纯化。使用 2:1 的 DNA 结合缓冲液与 MDA 产物体积比例。所纯化的 MDA 产物以 12 μl 水洗脱。

[0083] 使用 4 μl 每一个纯化的产物进行 Infinium® 基因型分型测试,这种基因型分型测试使用 Illumina® 300K HumanCytoSNP-12 BeadChips,按照 Illumina 方案 11230143 Rev A 中的指南实施。在于 iScan 上用 iCS 3.3.28 (Infinium II Assay Lab Setup and Procedures Guide, Illumina part # 11207963) 扫描 BeadChips 后,把数据输入到使用 GenomeStudio™ 基因型分型模块 (Genotyping Module) v1.0 (Illumina Part # 11318815) 的 GenomeStudio™ 2008.1 Framework (架构) 中。

[0084] 原始 X 和 Y 强度每样品的散点图用于表明起始物料中存在任意称为 A 和 B 的两个位点的半合子,例如 (X,0) 和 (0,Y),而 A/A 或 A/0 基因型将导致沿着 X 轴的数据点,B/B 或 B/0 基因型将导致沿着 Y 轴的数据点和 A/B 基因型将导致沿着 X 和 Y 轴之间的斜线的数据点。合并的全基因组 (X,Y) 数据点表明,如果在起始物料中存在大于一个人类基因组的单倍体拷贝,则将存在杂合性等位基因。

[0085] 图 9-10 证实把杂合性 SNPs 分解成其单倍体组分的方法的能力。图 9 散点图代表正常二倍体个体的示例性原始强度,证实 B/B 基因型位点强度沿着 Y 轴 (0,Y) 集中,A/A 基因型位点强度沿着 X 轴 (X,0) 集中,和 A/B 基因型位点强度在 X 和 Y 轴 (X/Y) 之间的中途集中。图 10 代表 12 个稀释样品中的 6 个对图 9 的二倍体样品中的 A & B 位点的示例性原始强度。A/A 位点强度数据沿着 X 轴集中,而 B/B 位点强度数据沿着 Y 轴集中。

[0086] 实施例 3:两个雄性基因组 DNAs 的混合物中的 X- 染色体杜兴肌营养不良症基因 (DMD) 单体型使用不对称样品分布方法和下一代测序法的测定

把含有测序的基因组 NA18507 (Bentley et al., 2008, Nature 456: 53-59) 和 HG01377 (Durbin et al., 2010, Nature 467: 1061-1073) (Coriell Cell Repositories, Camden, New Jersey) 的两个正常雄性基因组 DNA 样品以等比例合并,得到含有已知单体型的二倍体 X- 染色体的人工样品。把样品稀释并以 0.2 个单倍体拷贝每等分试样 (6.00E-07 μg/μl) 分配成 96 个等分试样。每一个稀释的模板 DNA 的等分试样如在实施例 2 中描述的那样用 MDA 单独扩增。为了评价等分试样的半合子状态和基因组覆盖率,通过在 Illumina® 300K HumanCytoSNP-12 BeadChip 上的 Infinium® 基因型分型测试 4 μl 所扩增的材料。把 100 毫微克纯化的 MDA 产物或 50 ng 未稀释的基因组 DNA 使用 Nextera™技术,按照制造商的方案转化为测序库 (Illumina, Inc., San Diego, CA)。每

个样品在有限的周期 PCR 期间被应用条码技术 (barcoded)。

[0087] 把最多 12 个测序库在对总计 8 个池测序之前进行合并。测序库按照制造商的指南，以 0.6 比例用 AMPure XP 珠粒进行纯化 (Beckman Coulter Genomics, Danvers, MA)。对 DMD 基因的 1Mb 连续区域的定向下拉设计探针池。生物素化的探针为 80nt 长，并被设计以 190-370bp 间隔杂交于 DMD 基因的 5'区域。合并后，测序库按照 TruSeq™定制富集试剂盒 (TruSeq™ Custom Enrichment Kit) (Illumina, San Diego, CA) 的方案富集 1Mb DMD 基因区域。所富集的索引化的库使用双端测序对 75 + 35 或 75 + 75 读取长度，在基因组分析仪 (Genome Analyzer) IIx (Illumina, San Diego, CA) 上进行测序。每个道含有 12 个样品中的一个池。每个雄性单独富集和测序，以证实混合 DNA 中的真正单体型结构，并且混合的样品被独立地富集和测序，以确定区域中的所有杂合 SNPs 和评价 DMD 寡核苷酸富集池的性能。

[0088] 把序列读数使用 Illumina CASAVA v1.8.1 软件包进行多路解编并与人类基因组比对，对每一个索引化的稀释样品创建比对的欺骗文件 (bam files)。用 SAMtools (Li et al., 2009, Bioinformatics 25: 2078-2079) 和目标削减 (target cut) (Kitzman et al., 2011, Nat. Biotechnol 29: 59-63) 自欺骗文件提取连续区域。在每一个连续的片段中，在已知 SNPs 的位置自“二倍体”测序数据进行碱基判定 (base calls)。片段被分解成连续的纯合子区段，即去除重叠的 DNA 片段，并运行 ReFHap (Duitama et al., 2011, Nucleic Acids Res. doi:10.1093/nar/gkr1042)，以把单体型化的片段融合成单体型模块。生成的单体型与两个单一雄性 gDNAs 中的已知单体型进行比较。

[0089] 图 11 显示作为加载到加州大学圣克鲁兹分校基因组浏览器的定制序列段 (custom tracks)，自顶面板 (top panel) 的 HG01377 (顶部) 和 NA18507 (底部) 和底面板 (bottom pane) 的合并单体型模块衍生的个体连续的纯合子比对区段。两个合并的单体型模块之间的间隙是由于人类基因组中的非可比对区 (unalignable region)。总的单体型化区域为 989 kb，并且平均单体型模块大小为 494kb。

[0090] 实施例 4：整个人类基因组使用不对称样品分布方法和下一代测序的单倍体分型

将来自正常个体的人类基因组 DNA，NA18506 (Coriell Cell Repositories, Camden, New Jersey)，稀释至 0.5 或 1.0 个单倍体拷贝每 1 μl 水 (1.50E-06 μg/μl 或 3.00E-06 μg/μl)。把每份稀释物所稀释的基因组 DNA (1 μl) 等分至 24 个管中，平均导致每管中含有 0.5 或 1.0 个人类基因组的单倍体拷贝。向每管中加入 1 μl 缓冲液 D1 (含有 0.875 μl 水的 0.125 μl DLB 缓冲液) (Qiagen REPLI-g® UltraFast Mini Handbook, Catalog # 150035)，随后在室温下温育 3 分钟。加入 1 微升缓冲液 N1 (含有 1.8 μl 水的 0.2 μl REPLI-g UltraFast 停止液)，随后加入 17μl 的 Mastermix。Mastermix 含有 15 μl REPLI-g UltraFast 反应缓冲液、1 μl REPLI-g UltraFast DNA 聚合酶和 1 μl 水。把反应物在 BioRad Tetrad2 热循环仪 (BioRad Part # PTC-0240G) 中于 30℃下温育 90 分钟，随后通过在 65℃下把样品加热 3 分钟使 REPLI-g UltraFast DNA 聚合酶热灭活。

[0091] MDA 产物使用 DNA Clean & Concentrator™-5 离心柱 (Zymo Research Catalog # D4003)，按照制造商的方案进行纯化。使用 2:1 的 DNA 结合缓冲液与 MDA 产物的体积比例。所纯化的 MDA 产物以 17 μl 水洗脱。把纯化的 MDA 产物 (15 μl) 使用 Nextera™技术，按照制造商的方案转化为测序库 (Illumina, Inc., San Diego, CA)，除了 Nextera 酶被

稀释 100 倍，以补偿向标记化 (tagmentation) 反应中的低 DNA 模板输入量和增加 dsDNA/Nextera 酶的比例。这防止产生太小的库插入大小。每个样品在有限的周期 PCR 期间被应用条码技术。把最多 12 个测序库在对总计 4 个池测序之前进行合并。测序库按照制造商的指南，以 0.6 比例用 AMPure XP 珠粒进行纯化 (Beckman Coulter Genomics, Danvers, MA)。这些库使用双端测序对 100 + 100 读取长度，在 HiSeq 2000 测序系统 (Sequencing System) (Illumina, San Diego, CA) 上测序。12 个样品中的每一个池以 2 个道进行测序。如在实施例 4 中描述的那样进行序列读取的分析。通过比较经亲本基因型的统计计算得到的分解的单体型得到证实。

[0092] 图 12 显示作为加载到加州大学圣克鲁兹分校基因组浏览器的定制序列段的顶面板的个体连续的纯合子区段和底面板的合并单体型模块的实例。该实例证实可对整个基因组二倍体样品实施单倍体分型。所得到的最大准确单体型模块为 303.5kb，并且总计 1.27Gb 为单体型 - 分离的 (haplotype-resolved)。

[0093] 在本申请中提及的所有出版物和专利通过引用结合到本文中。本发明所描述的方法和组合物的各种修饰和变化对本领域技术人员而言是显而易见的，而不背离本发明的范围和精神。尽管本发明已经结合具体优选的实施方案进行描述，应该理解，所要求保护的本发明不应过度地限于这样的具体实施方案。的确，所描述的用于实施对相关领域的技术人员显而易见的本发明的模式的各种修饰打算处于以下权利要求书的范围内。

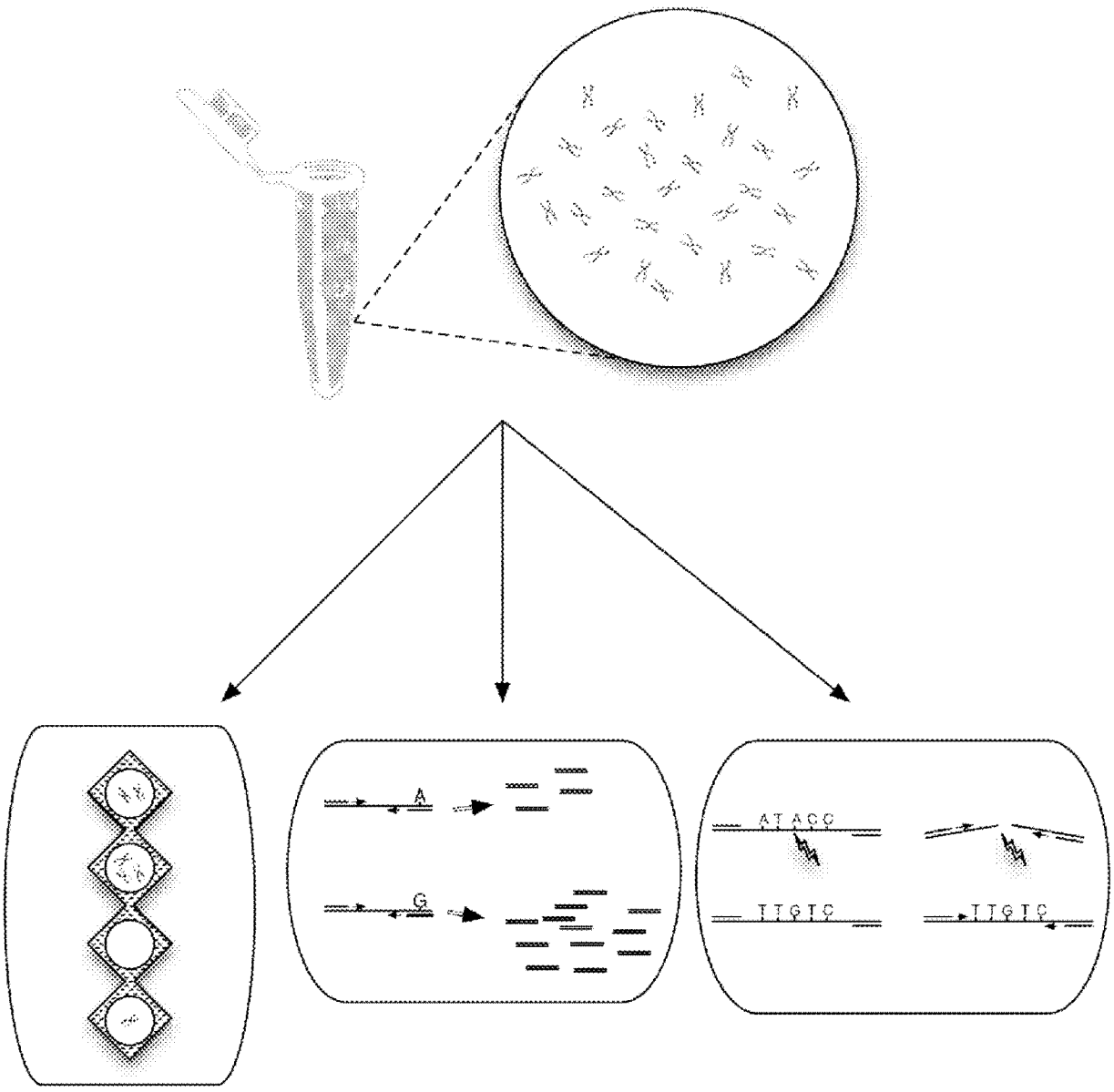


图 1

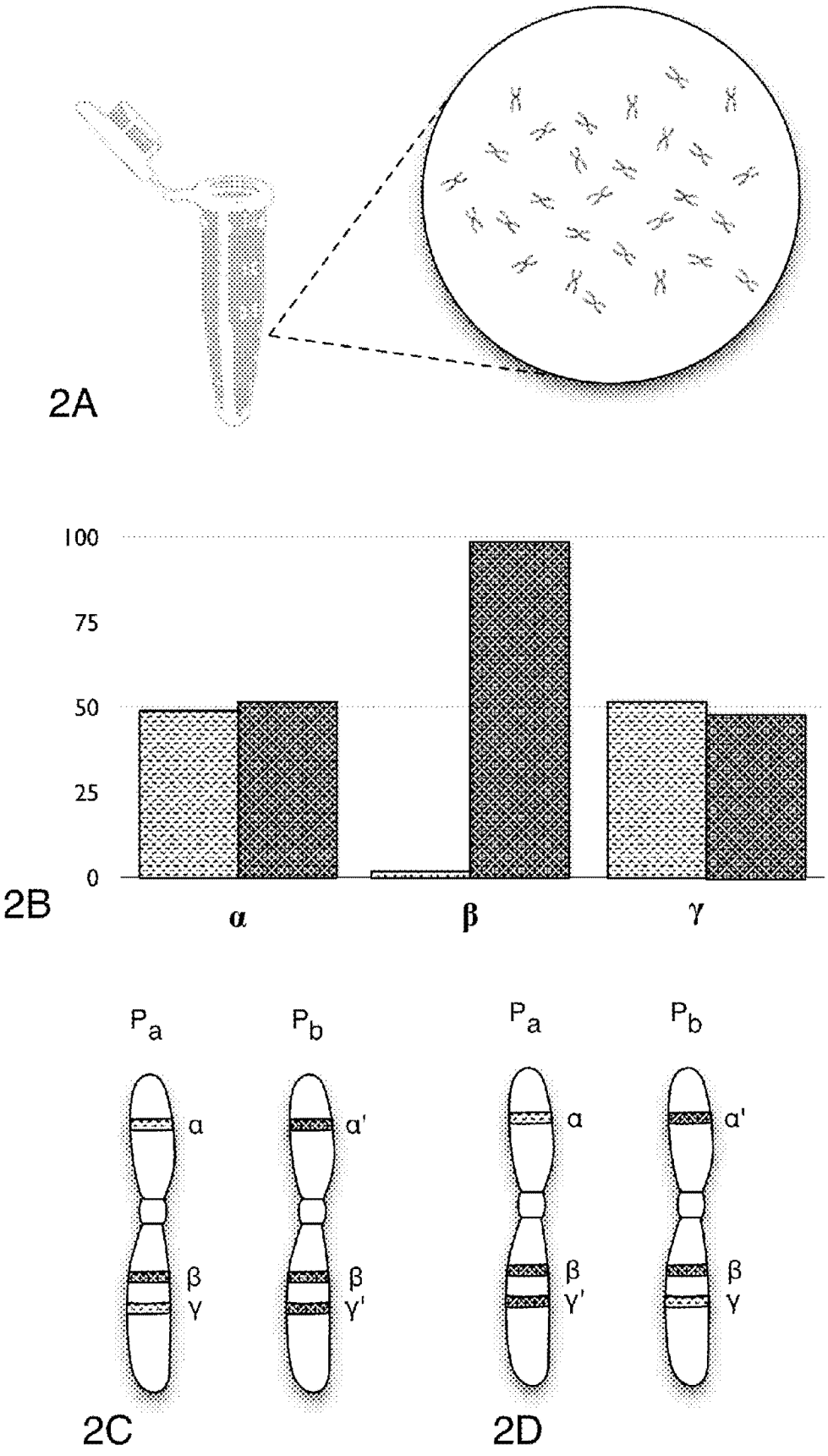
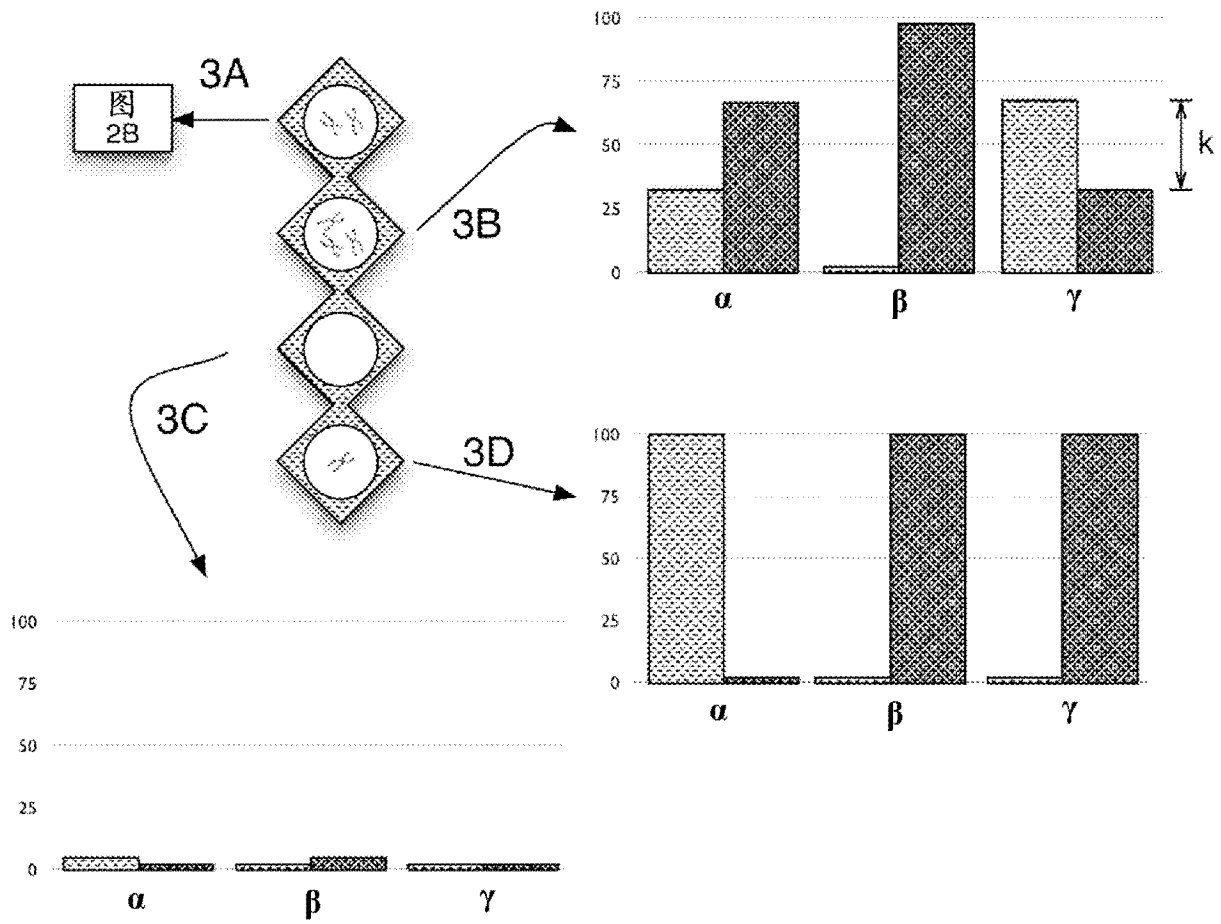


图 2



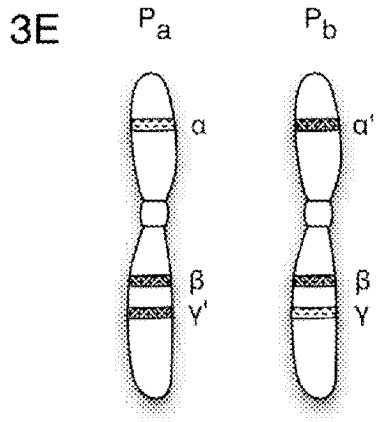
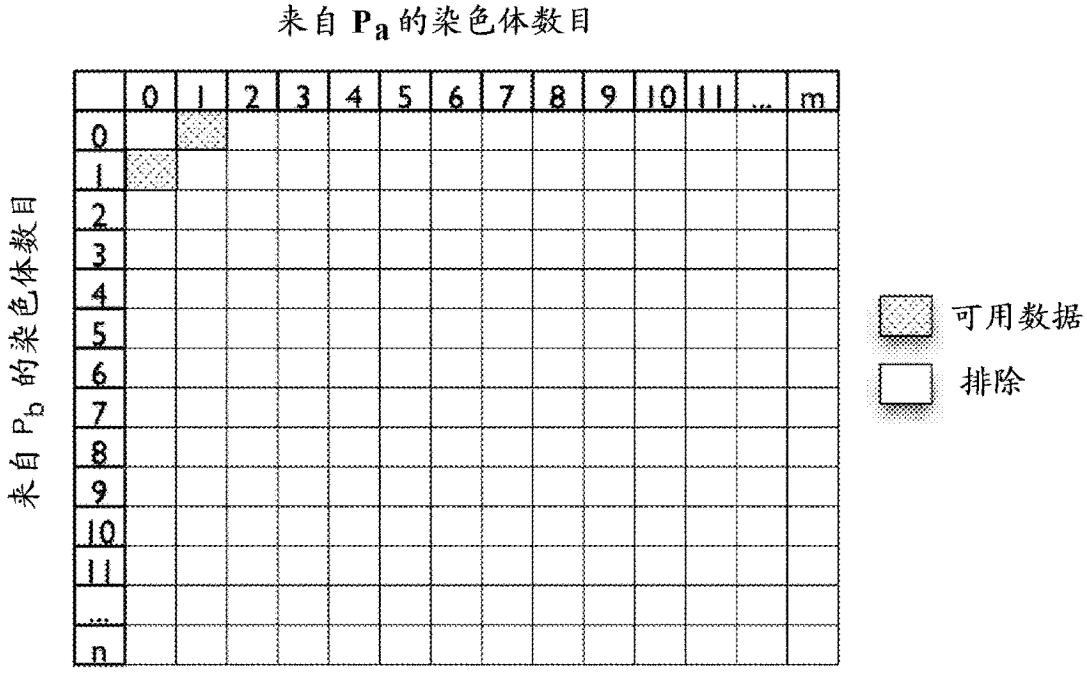


图 3

4A


来自 P_a 的染色体数目
来自 P_b 的染色体数目
0 1 2 3 4 5 6 7 8 9 10 11 ... m
0 1 2 3 4 5 6 7 8 9 10 11 ... n
可用数据
排除

4B

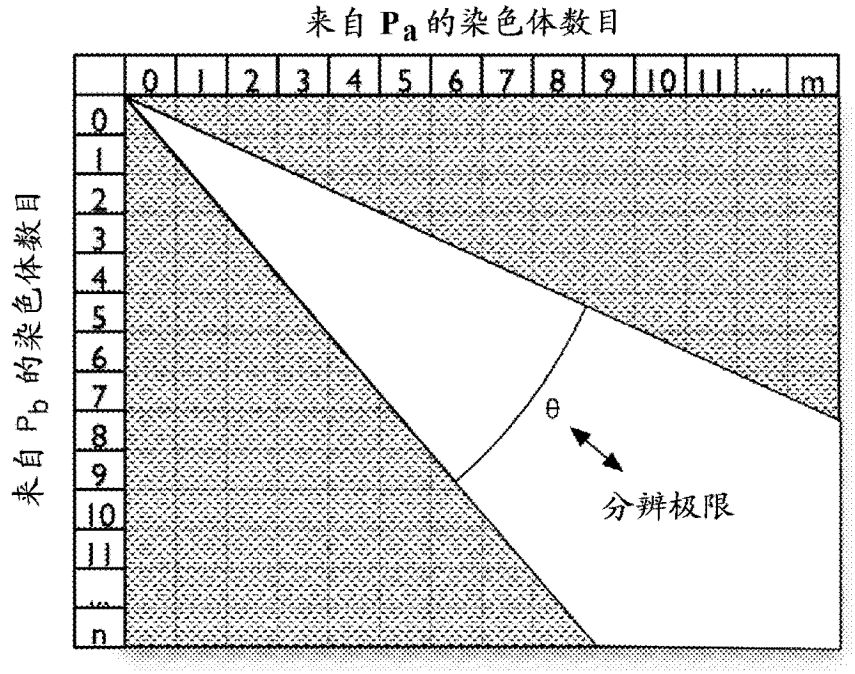

来自 P_a 的染色体数目
来自 P_b 的染色体数目
0 1 2 3 4 5 6 7 8 9 10 11 ... m
0 1 2 3 4 5 6 7 8 9 10 11 ... n
θ
分辨极限

图 4

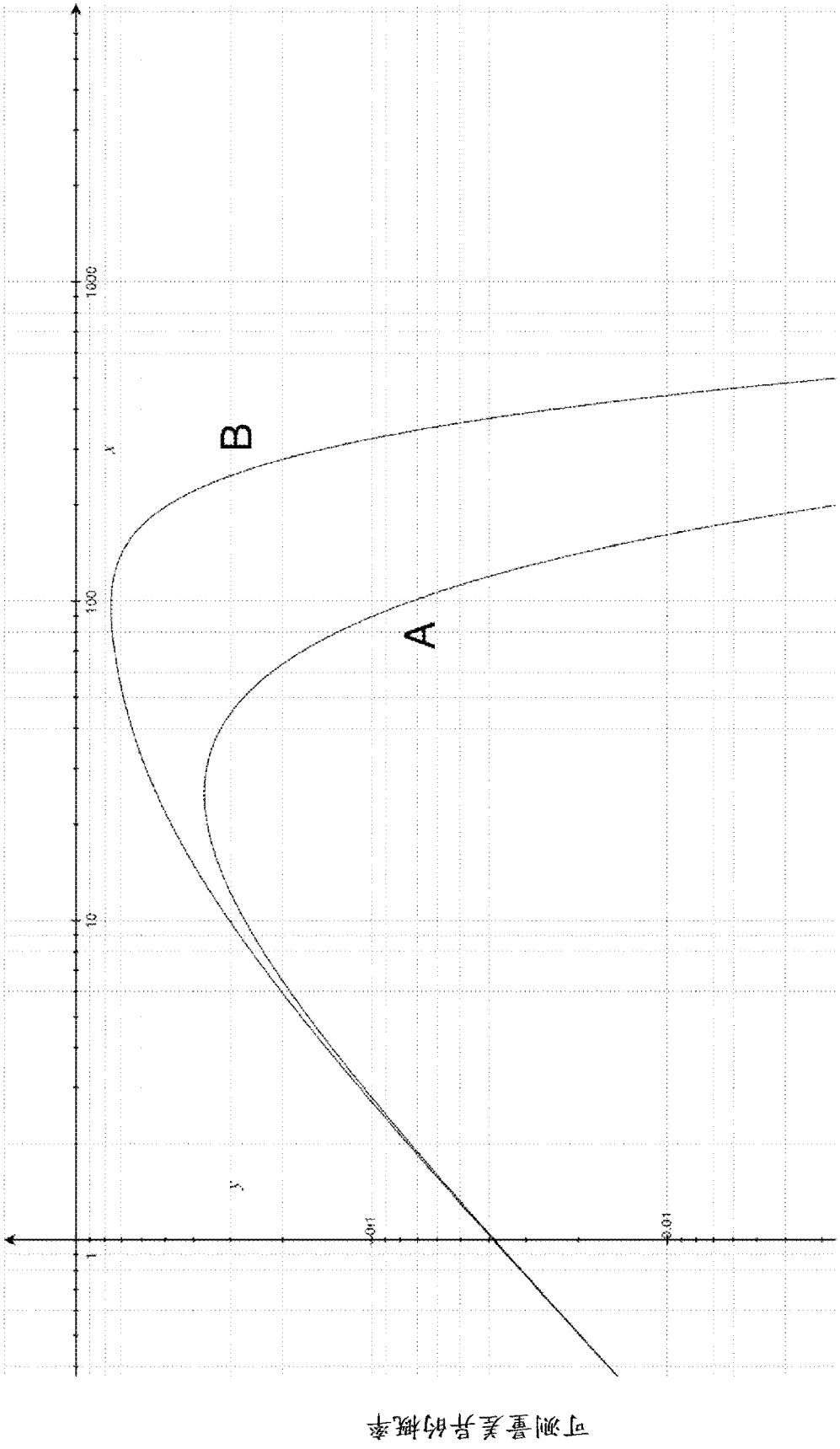
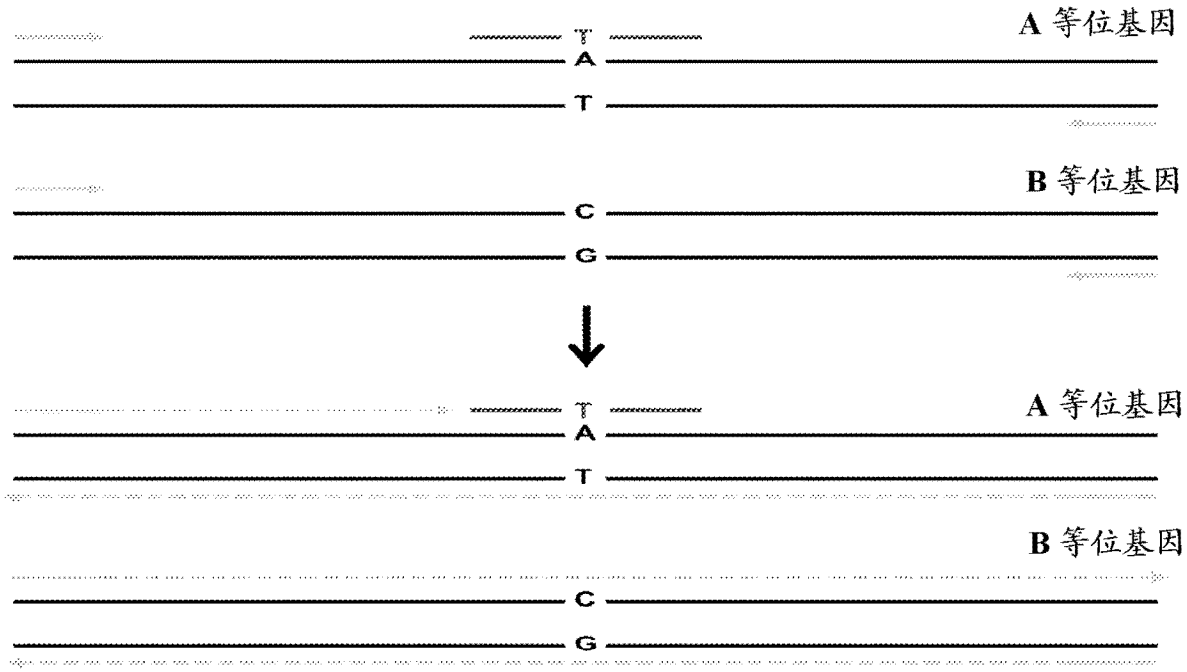


图 5

A


A 等位基因
T
A
T
B 等位基因
C
G
A 等位基因
T
A
T
B 等位基因
C
G

B

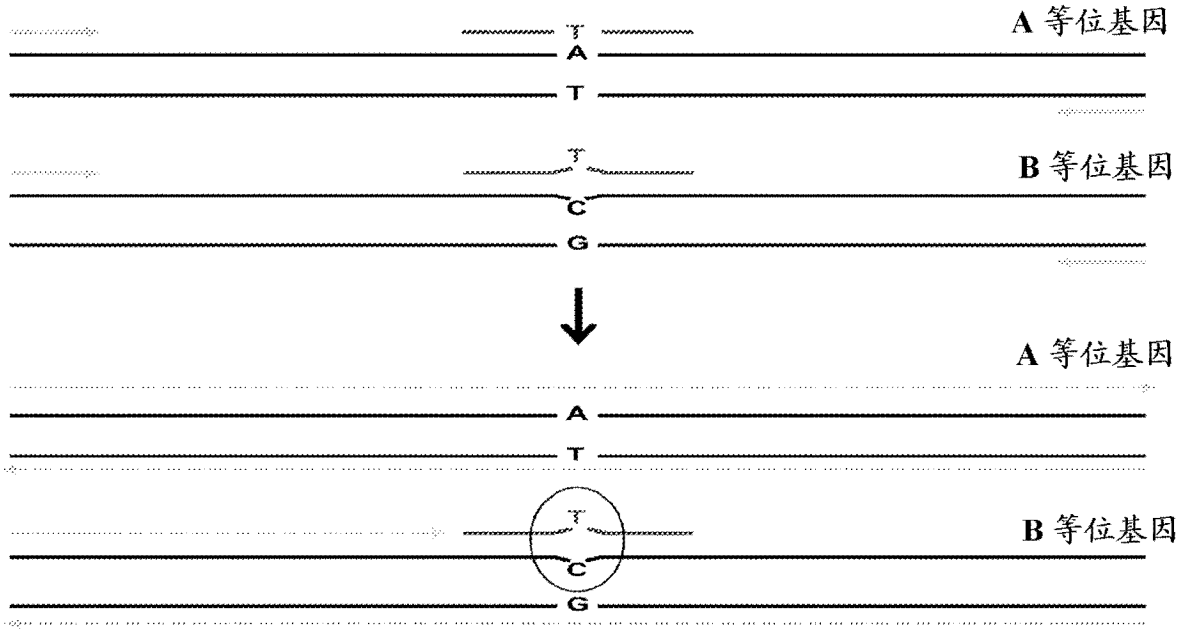

A 等位基因
T
A
T
B 等位基因
T
C
G
A 等位基因
T
A
T
B 等位基因
T
C
G

图 6

C

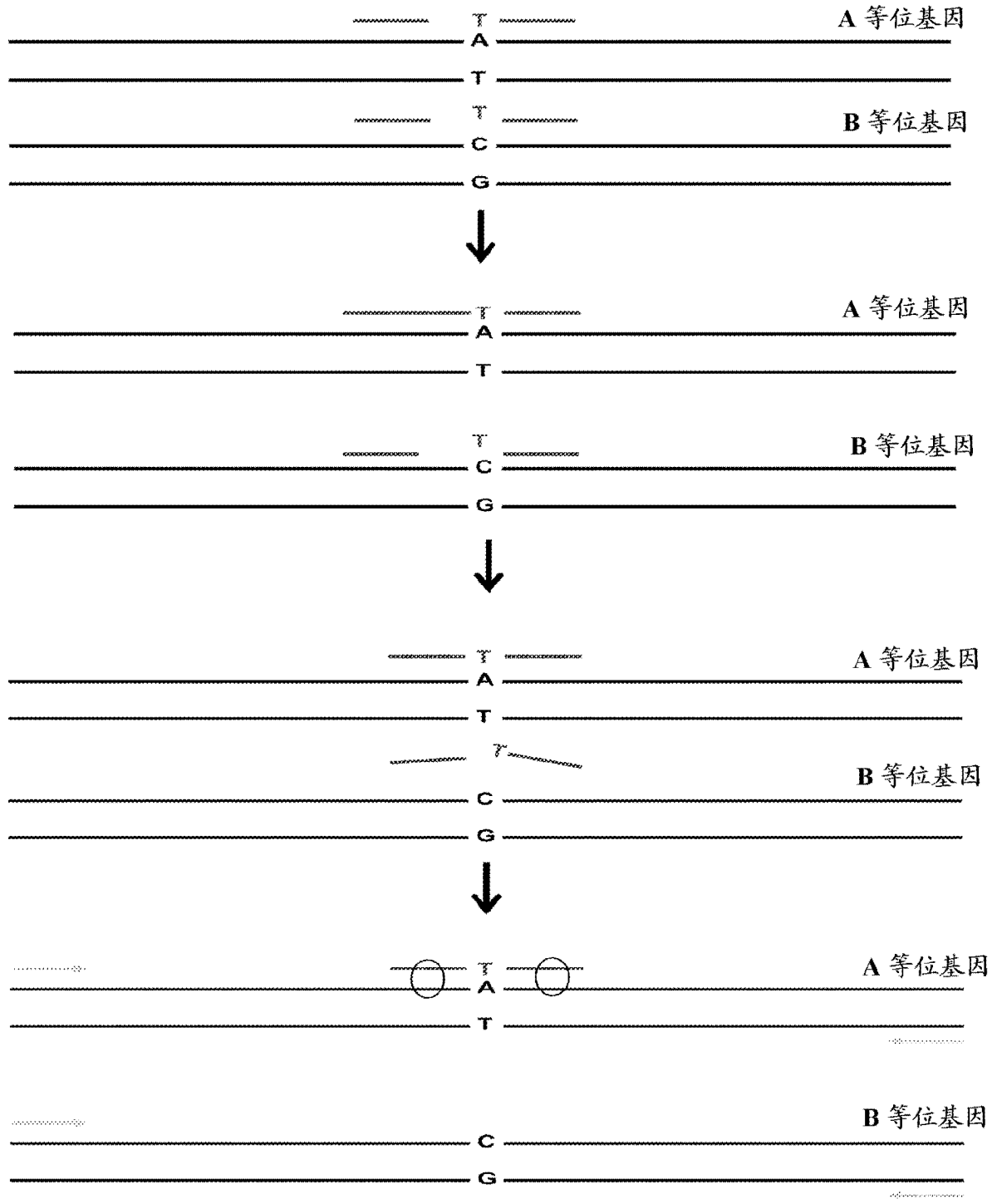

A 等位基因
T
A
T
B 等位基因
T
C
G
A 等位基因
T
A
T
B 等位基因
T
C
G
A 等位基因
T
A
T
B 等位基因
C
G
A 等位基因
T
A
T
B 等位基因
C
G

图 6(续)

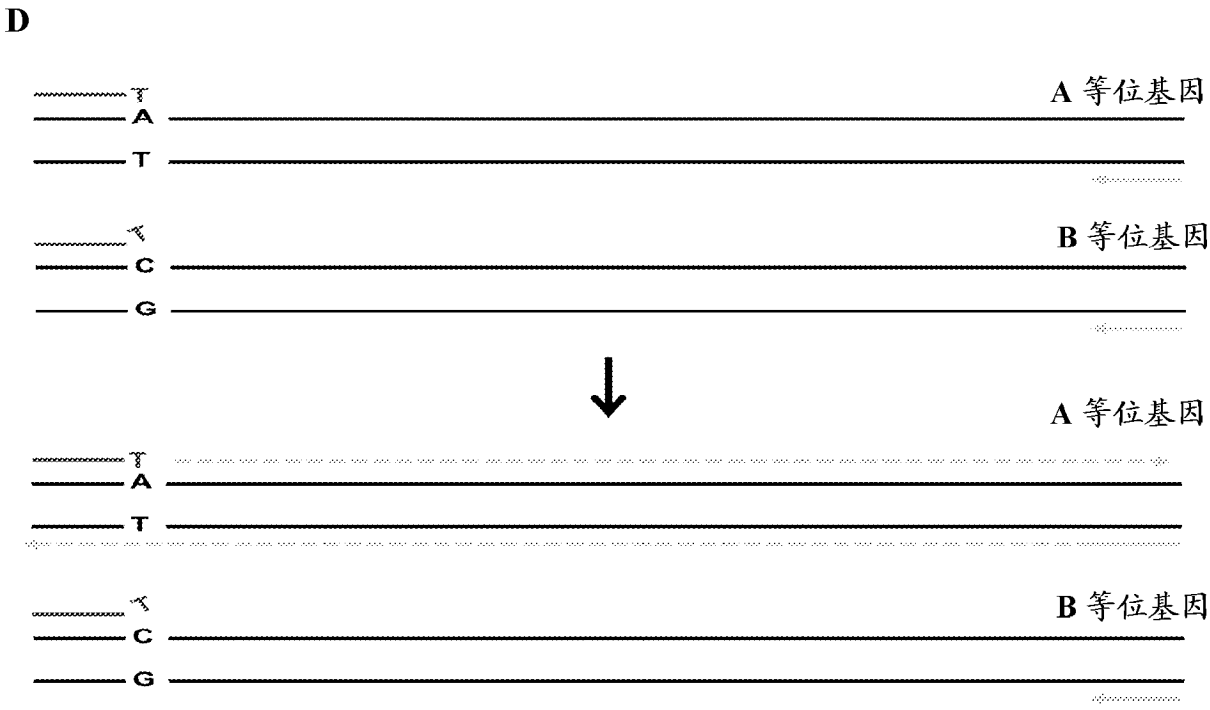


图 6(续)

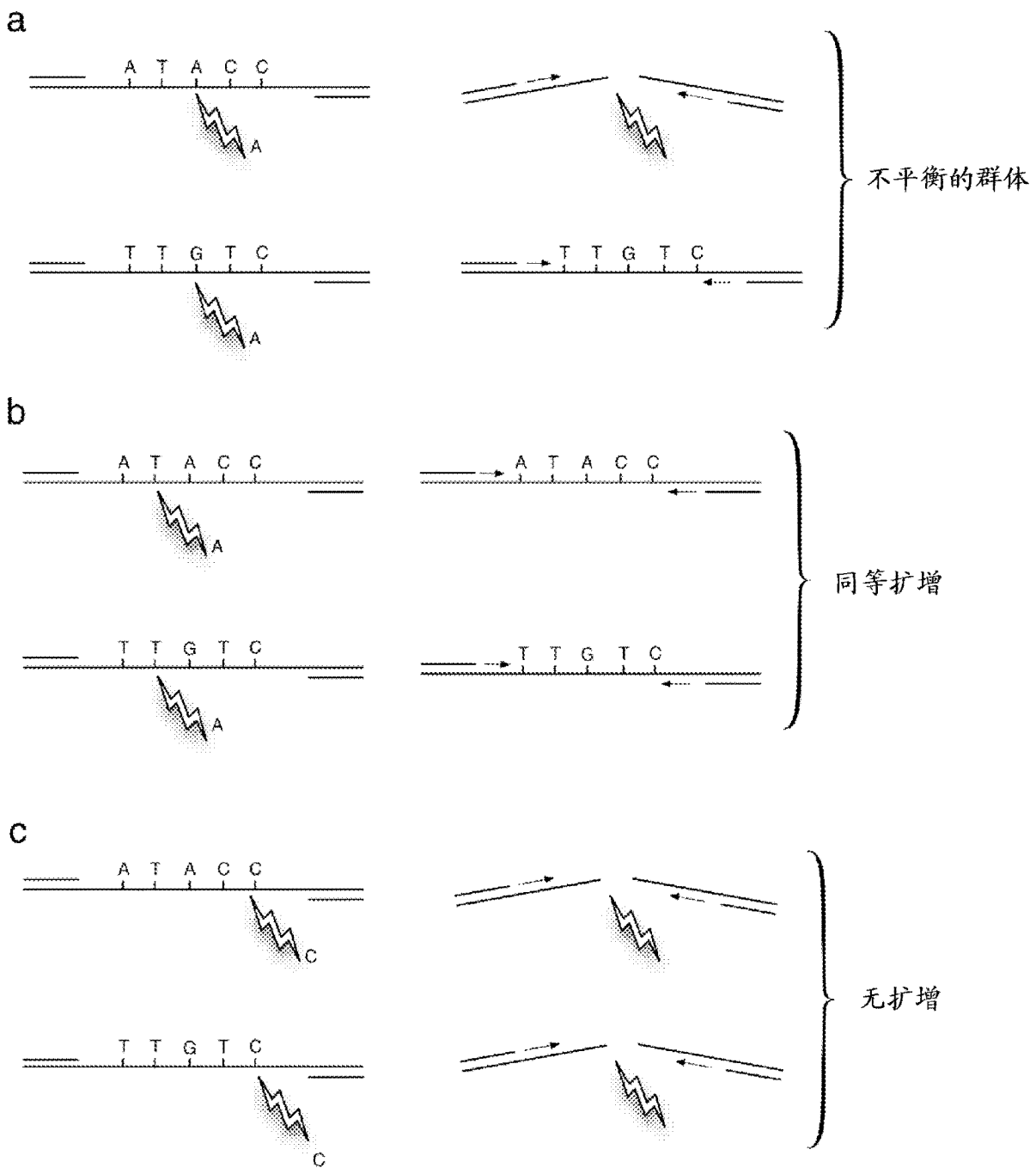

a
A T A C C
A
T T G T C
A
T T G T C
不平衡的群体
b
A T A C C
A
A T A C C
T T G T C
A
T T G T C
同等扩增
c
A T A C C
C
T T G T C
C
无扩增

图 7

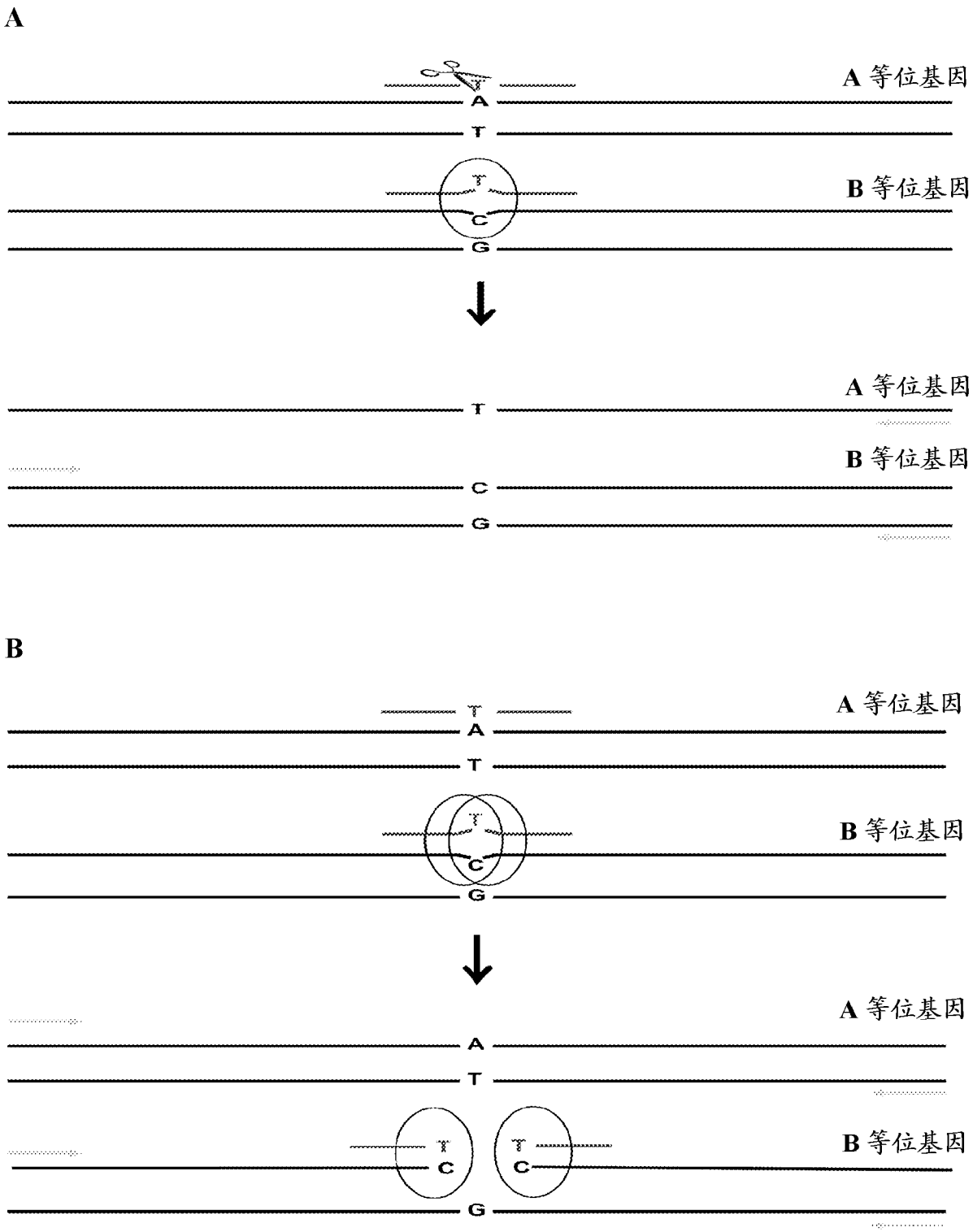

A
A 等位基因
A
T
B 等位基因
T
C
G
A 等位基因
T
B 等位基因
C
G
B
A 等位基因
A
T
B 等位基因
C
G
A 等位基因
A
T
B 等位基因
C
C
G

图 8

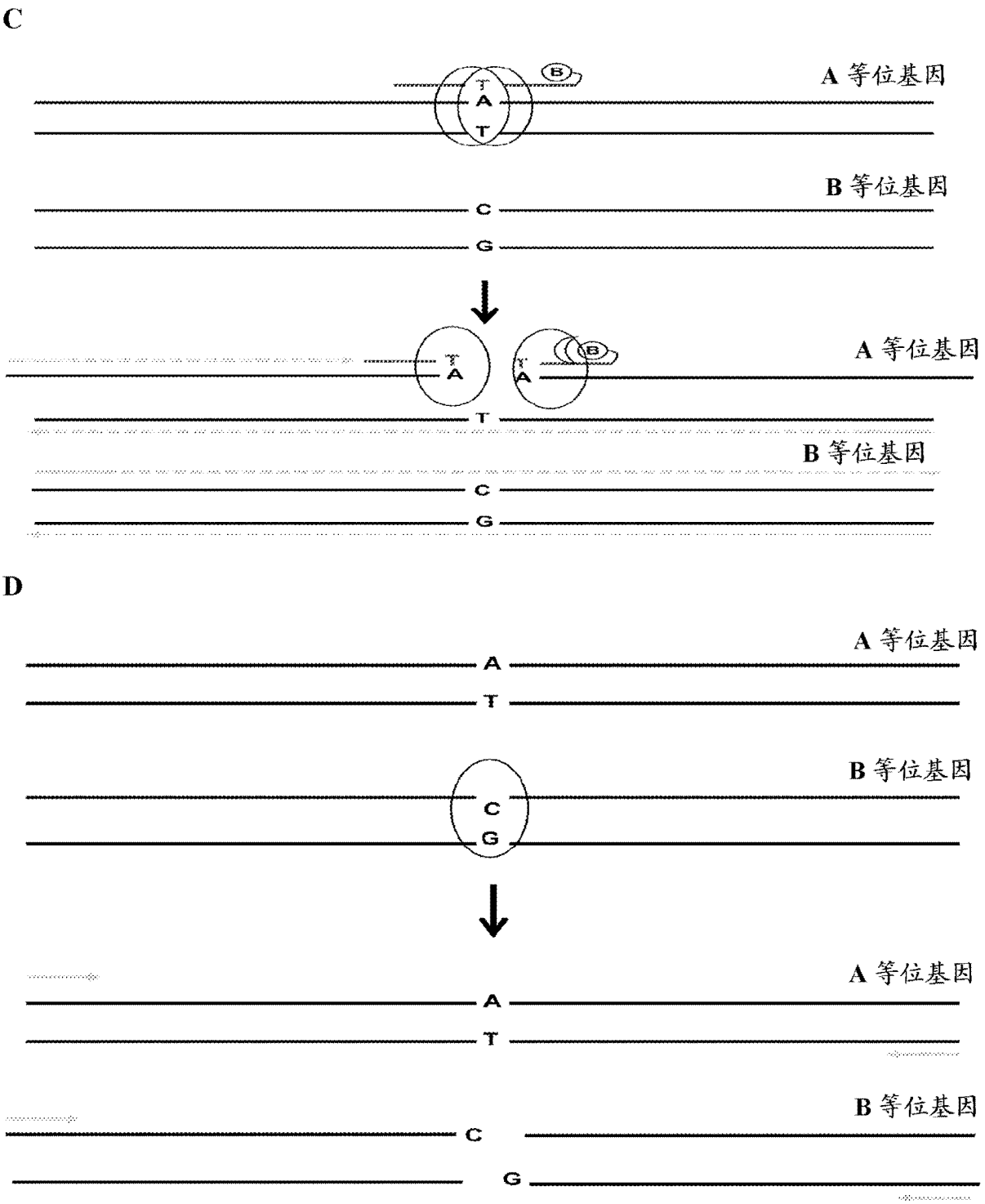

C
A 等位基因
T
A
T
B
B 等位基因
C
G
A 等位基因
T
A
B
T
A
T
B 等位基因
C
G
D
A 等位基因
A
T
B 等位基因
C
G
A 等位基因
A
T
B 等位基因
C
G

图 8(续)

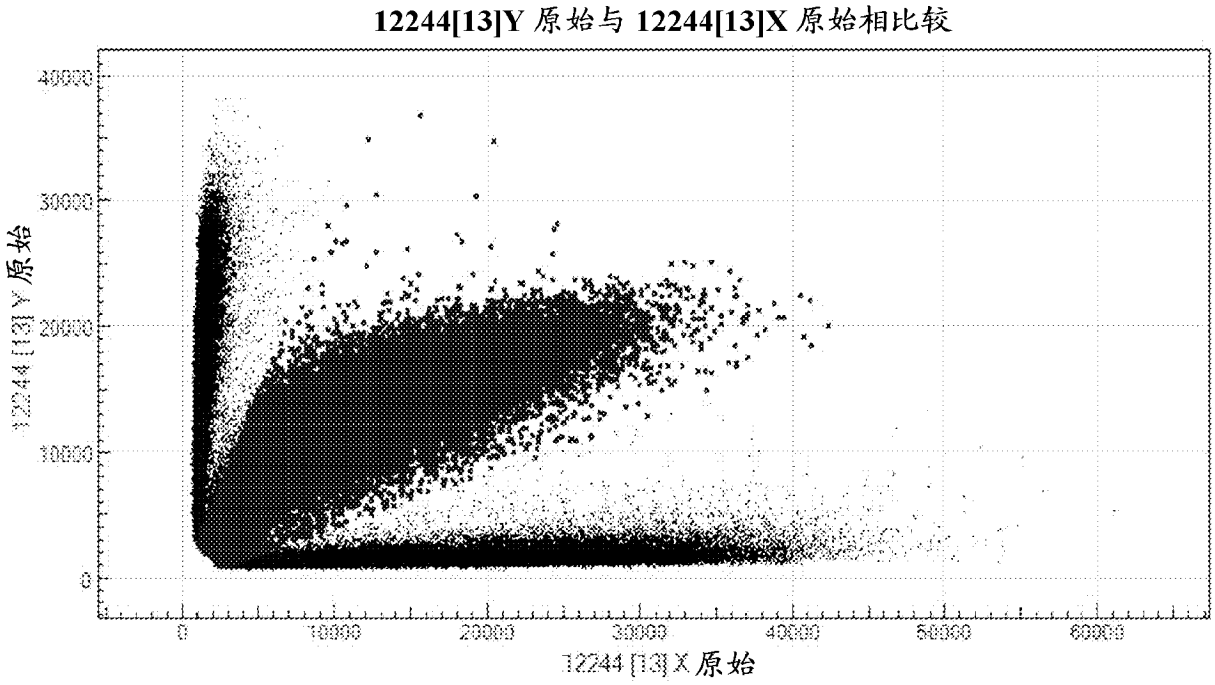

12244[13]Y 原始与 12244[13]X 原始相比较
40000
30000
20000
10000
0
12244 [13] Y 原始
0
10000
20000
30000
40000
50000
60000
12244 [13] X 原始

图 9

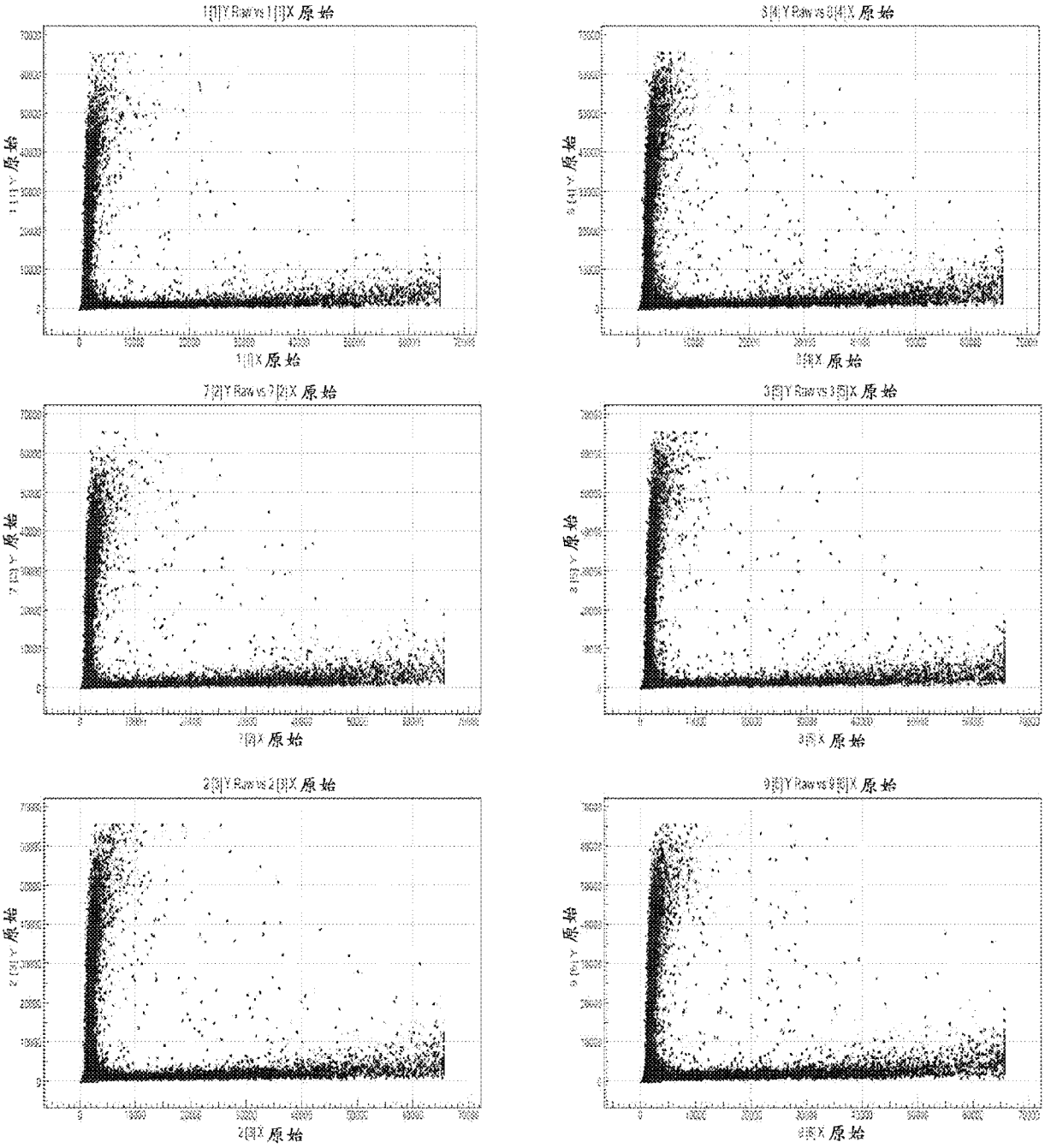


图 10

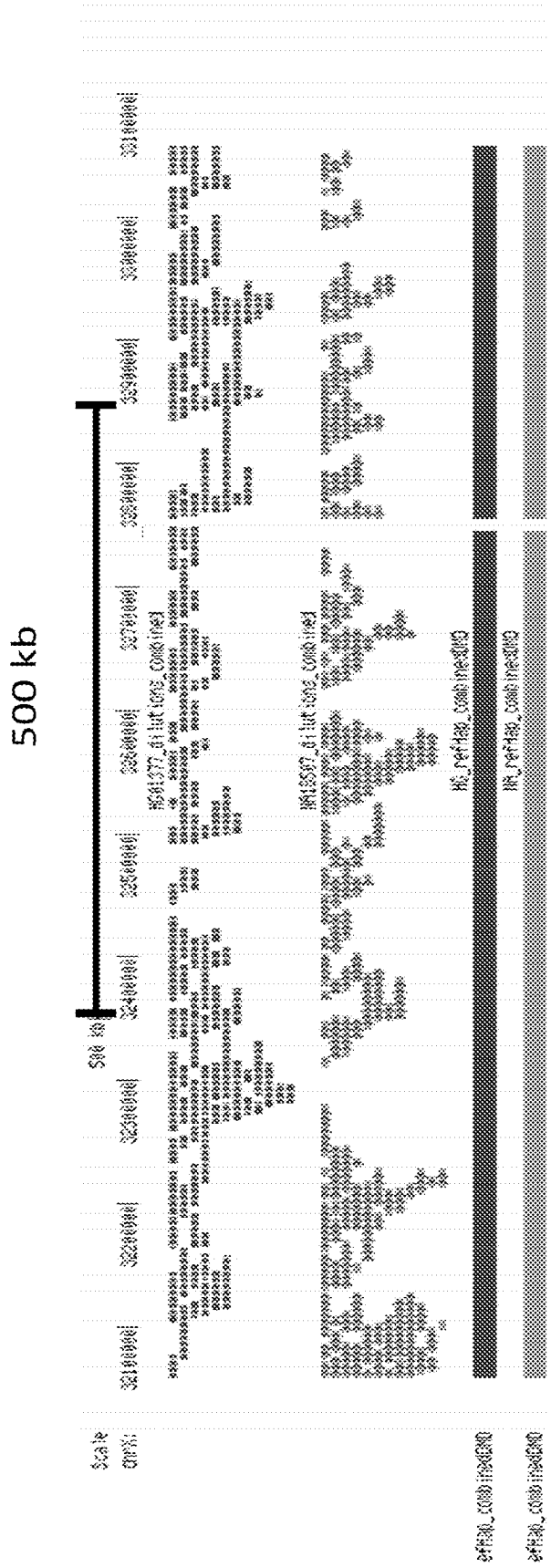


图 11

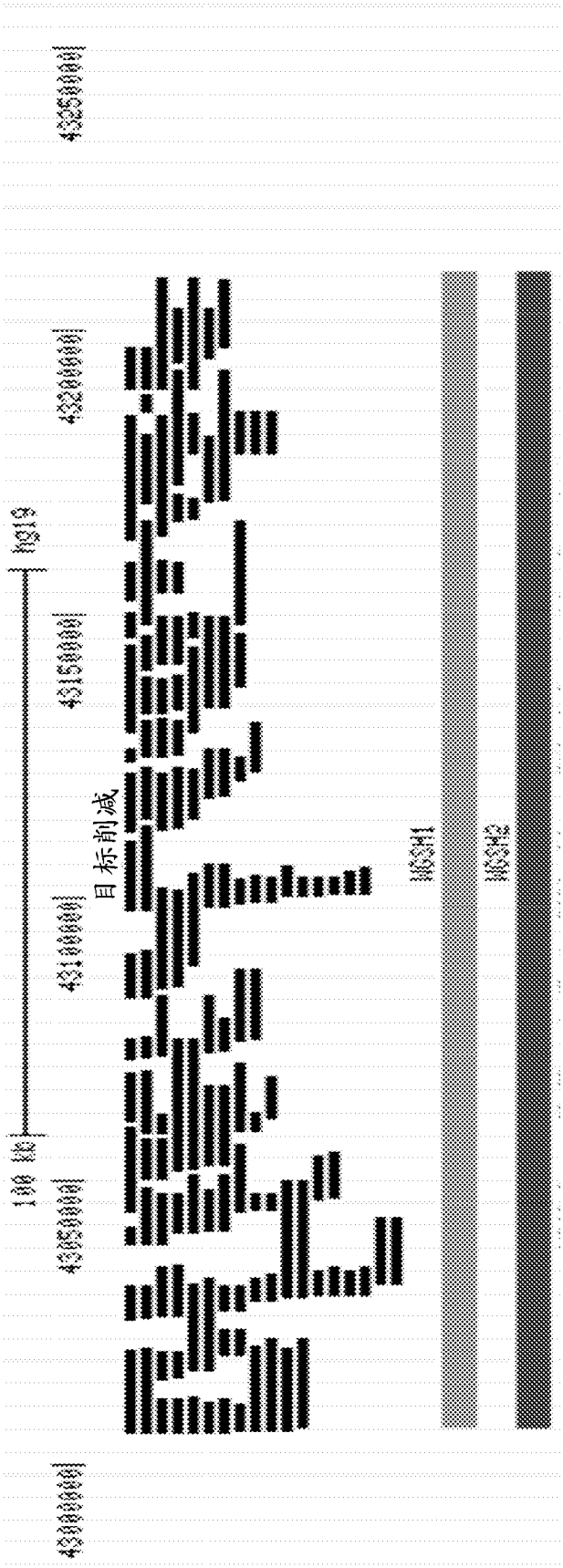

43250000
43200000
43150000
43100000
43050000
43000000
hg19
100 kb
目标削减
BS#1
BS#2

图 12