



(12)发明专利申请

(10)申请公布号 CN 110382752 A

(43)申请公布日 2019.10.25

(21)申请号 201780084031.1

陈思远

(22)申请日 2017.11.17

(74)专利代理机构 北京安信方达知识产权代理有限公司 11262

(30)优先权数据

62/424,302 2016.11.18 US

62/548,307 2017.08.21 US

62/558,666 2017.09.14 US

代理人 刘晓杰

(85)PCT国际申请进入国家阶段日

2019.07.18

(51)Int.Cl.

C40B 50/14(2006.01)

C12Q 1/6874(2018.01)

C12N 15/10(2006.01)

(86)PCT国际申请的申请数据

PCT/US2017/062391 2017.11.17

(87)PCT国际申请的公布数据

W02018/094263 EN 2018.05.24

(71)申请人 特韦斯特生物科学公司

地址 美国加利福尼亚州

权利要求书5页 说明书57页

(72)发明人 拉姆齐·易卜拉欣·泽图恩

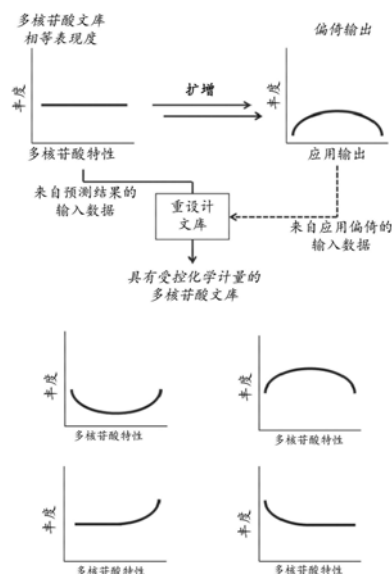
序列表4页 附图46页

(54)发明名称

具有受控化学计量的多核苷酸文库及其合成

(57)摘要

本文提供了与多核苷酸文库相关的组合物、方法和系统,所述多核苷酸文库对于多核苷酸的种类具有预选的化学计量,使得所述文库允许预定的应用结果,例如扩增后的受控表现度(representation)和与靶序列结合后的一致性富集。本文还提供了多核苷酸探针及其用于一致性和准确的下一代测序的应用。



1. 一种多核苷酸文库, 所述多核苷酸文库包含至少5000种多核苷酸, 其中所述至少5000种多核苷酸中的每一种多核苷酸以一定的量存在, 使得在与基因组片段杂交并对杂交的基因组片段测序后, 所述多核苷酸文库在基因组片段碱基的高达55倍理论读段深度的条件下提供至少90%的基因组片段碱基的至少30倍读段深度。

2. 权利要求1所述的多核苷酸文库, 其中所述多核苷酸文库在基因组片段碱基的高达55倍理论读段深度的条件下提供至少95%的基因组片段碱基的至少30倍读段深度。

3. 权利要求1所述的多核苷酸文库, 其中所述多核苷酸文库在基因组片段碱基的高达55倍理论读段深度的条件下提供至少98%的基因组片段碱基的至少30倍读段深度。

4. 权利要求1所述的多核苷酸文库, 其中所述多核苷酸文库提供基因组片段碱基的至少90%的独特读段。

5. 权利要求1所述的多核苷酸文库, 其中所述多核苷酸文库提供基因组片段碱基的至少95%的独特读段。

6. 权利要求1所述的多核苷酸文库, 其中所述多核苷酸文库提供至少90%的具有在平均读段深度的约1.5倍内的读段深度的基因组片段碱基。

7. 权利要求1所述的多核苷酸文库, 其中所述多核苷酸文库提供至少95%的具有在平均读段深度的约1.5倍内的读段深度的基因组片段碱基, 所述基因组片段碱基。

8. 权利要求1所述的多核苷酸文库, 其中所述多核苷酸文库提供至少90%的如下基因组片段: 所述基因组片段具有GC百分比为10%至30%或者70%至90%, 具有在平均读段深度的约1.5x内的读段深度。

9. 权利要求1所述的多核苷酸文库, 其中所述多核苷酸文库提供至少约80%的如下基因组片段: 所述基因组片段具有10%至30%或70%至90%的重复序列或二级结构序列百分比, 具有在平均读段深度的约1.5x内的读段深度。

10. 权利要求1所述的多核苷酸文库, 其中每个基因组片段长度为约100个碱基至约500个碱基。

11. 权利要求1所述的多核苷酸文库, 其中所述至少5000种多核苷酸中至少约80%以所述多核苷酸文库平均表现度的至少约1.5倍内的量表现。

12. 权利要求1所述的多核苷酸文库, 其中所述至少5000种多核苷酸中的至少30%包含具有GC百分比为10%至30%或70%至90%的多核苷酸。

13. 权利要求1所述的多核苷酸文库, 其中所述至少5000种多核苷酸中的至少约15%包含具有重复序列或二级结构序列百分比为10%至30%或70%至90%的多核苷酸。

14. 权利要求1所述的多核苷酸文库, 其中所述至少5000种多核苷酸编码至少1000种基因。

15. 权利要求1所述的多核苷酸文库, 其中所述多核苷酸文库包含至少100,000种多核苷酸。

16. 权利要求1所述的多核苷酸文库, 其中所述多核苷酸文库包含至少700,000种多核苷酸。

17. 权利要求1所述的多核苷酸文库, 其中所述至少5000种多核苷酸包含至少一个外显子序列。

18. 权利要求16所述的多核苷酸文库, 其中所述至少700,000种多核苷酸包含至少一组

多核苷酸,所述至少一组多核苷酸共同包含单个外显子序列。

19. 权利要求18所述的多核苷酸文库,其中所述至少700,000种多核苷酸包含至少150,000组。

20. 一种多核苷酸文库,所述多核苷酸文库包含至少5000种多核苷酸,其中每种多核苷酸长度约为20至200个碱基,其中多于一个多核苷酸编码来自至少1000种预选基因的各外显子的序列,其中每个多核苷酸包含分子标签,其中所述至少5000种多核苷酸中的每种多核苷酸以一定的量存在,使得在与基因组片段杂交并对杂交的基因组片段测序之后,所述多核苷酸文库在基因组片段碱基的高达55倍理论读段深度的条件下提供至少90%基因组片段碱基的至少30倍读段深度。

21. 权利要求20所述的多核苷酸文库,其中所述多核苷酸文库在基因组片段碱基的高达55倍理论读段深度的条件下,提供至少95%的基因组片段碱基的至少30倍读段深度。

22. 权利要求20所述的多核苷酸文库,其中所述多核苷酸文库在基因组片段碱基的高达55倍理论读段深度的条件下,提供至少98%的基因组片段碱基的至少30倍读段深度。

23. 权利要求20所述的多核苷酸文库,其中所述多核苷酸文库提供基因组片段碱基的至少90%的独特读段。

24. 权利要求20所述的多核苷酸文库,其中所述多核苷酸文库提供基因组片段碱基的至少95%的独特读段。

25. 权利要求20所述的多核苷酸文库,其中所述多核苷酸文库提供至少90%的具有在平均读段深度的约1.5倍内的读段深度的基因组片段碱基。

26. 权利要求20所述的多核苷酸文库,其中所述多核苷酸文库提供至少95%的具有在平均读段深度的约1.5倍内的读段深度的基因组片段碱基。

27. 权利要求20所述的多核苷酸文库,其中所述多核苷酸文库提供大于90%的如下基因组片段:所述基因组片段具有GC百分比为10%至30%或者70%至90%,具有在平均读段深度的约1.5倍内的读段深度。

28. 权利要求20所述的多核苷酸文库,其中所述多核苷酸文库提供大于约80%的如下基因组片段:所述基因组片段具有10%至30%或70%至90%的重复序列或二级结构序列百分比,具有在平均读段深度的约1.5倍内的读段深度。

29. 权利要求20所述的多核苷酸文库,其中每个基因组片段长度为约100个碱基至约500个碱基。

30. 权利要求20所述的多核苷酸文库,其中所述至少5000种多核苷酸中的大于约80%以所述多核苷酸文库平均表现度的至少约1.5倍内的量表现。

31. 权利要求20所述的多核苷酸文库,其中所述至少5000种多核苷酸中的大于约30%包含具有GC百分比为10%至30%或70%至90%的多核苷酸。

32. 权利要求20所述的多核苷酸文库,其中所述至少5000种多核苷酸中的大于约15%包含具有重复序列或二级结构序列百分比为10%至30%或70%至90%的多核苷酸。

33. 权利要求20所述的多核苷酸文库,其中所述多核苷酸文库包含至少100,000种多核苷酸。

34. 权利要求20所述的多核苷酸文库,其中所述多核苷酸文库包含至少700,000种多核苷酸。

35. 权利要求34所述的多核苷酸文库,其中所述至少700,000种多核苷酸包含至少一组多核苷酸,所述至少一组多核苷酸共同包含单个外显子序列。

36. 权利要求35所述的多核苷酸文库,其中所述至少700,000种多核苷酸包含至少150,000组。

37. 一种用于产生多核苷酸文库的方法,所述方法包括:

- a. 提供编码至少5000种多核苷酸的预定序列;
- b. 合成所述至少5000种多核苷酸;和
- c. 用聚合酶扩增所述至少5000种多核苷酸以形成多核苷酸文库,其中所述至少5000种多核苷酸中的大于约80%以所述多核苷酸文库平均表现度的至少约2倍内的量表现。

38. 权利要求37所述的方法,其中所述至少5000种多核苷酸中的大于约80%以所述多核苷酸文库平均表现度的至少约1.5倍内的量表现。

39. 权利要求37所述的方法,其中所述至少5000种多核苷酸中的大于30%包含具有GC百分比为10%至30%或70%至90%的多核苷酸。

40. 权利要求37所述的方法,其中所述至少5000种多核苷酸中的大于约15%包含具有重复序列或二级结构序列百分比为10%至30%或70%至90%的多核苷酸。

41. 权利要求37所述的方法,其中与没有校正错误的预定序列相比,所述多核苷酸文库具有在800个碱基中少于1个的总错误率。

42. 权利要求37所述的方法,其中所述预定序列编码至少700,000种多核苷酸。

43. 权利要求37所述的方法,其中所述至少5000种多核苷酸的合成发生在具有表面的结构上,其中所述表面包括多于一个簇,其中每个簇包括多于一个位点;并且其中所述至少5000种多核苷酸中的每一个都从所述多于一个位点的不同位点延伸。

44. 权利要求43所述的方法,其中所述多于一个位点包括每个簇高达1000个位点。

45. 权利要求43所述的方法,其中所述多于一个位点包括每个簇高达200个位点。

46. 一种用于多核苷酸文库扩增的方法,所述方法包括:

- a. 获得至少5000种多核苷酸的扩增分布;
- b. 基于至少一个序列特征将所述扩增分布的至少5000种多核苷酸聚类到两个或更多个箱中,其中所述序列特征是GC含量百分比、重复序列含量百分比或二级结构含量百分比;
- c. 调整在至少一个箱中多核苷酸的相对频率,以产生具有预选表现度的多核苷酸文库;
- d. 合成具有所述预选表现度的多核苷酸文库;和
- e. 扩增具有所述预选表现度的多核苷酸文库。

47. 权利要求46所述的方法,其中所述至少一个序列特征是GC含量百分比。

48. 权利要求46所述的方法,其中所述至少一个序列特征是二级结构含量百分比。

49. 权利要求46所述的方法,其中所述至少一个序列特征是重复序列含量百分比。

50. 权利要求49所述的方法,其中所述重复序列含量包括具有3个或更多腺嘌呤的序列。

51. 权利要求49所述的方法,其中所述重复序列含量包括在所述多核苷酸的至少一个末端上的重复序列。

52. 权利要求46所述的方法,其中所述多核苷酸基于一个或更多个多核苷酸序列结合

靶序列的亲合力而被聚类到箱中。

53. 权利要求46所述的方法,其中当与调整前的箱中的较低的30%的序列数量相比时,调整后的箱中的较低的30%的序列数量在下游应用中具有至少多50%的表现度。

54. 权利要求46所述的方法,其中,当与调整前的箱中的较高的30%的序列数量相比时,调整后的箱中的较高的30%的序列数量在下游应用中具有至少多50%的表现度。

55. 一种用于对基因组DNA测序的方法,所述方法包括:

(a) 将权利要求1-36中任一项所述的文库与多于一个基因组片段接触;

(b) 富集至少一个与所述文库结合的基因组片段,以产生至少一种富集的靶多核苷酸;

和

(c) 对所述至少一种富集的靶多核苷酸测序。

56. 权利要求55所述的方法,其中,所述多于一种富集的靶多核苷酸包括cDNA文库。

57. 权利要求55所述的方法,其中所述至少5000种多核苷酸的长度为约80至约200个碱基。

58. 权利要求55所述的方法,其中每个基因组片段长度为约100个碱基至约500个碱基。

59. 权利要求55所述的方法,其中接触发生在溶液中。

60. 权利要求55所述的方法,其中所述至少5000种多核苷酸至少部分与所述基因组片段互补。

61. 权利要求55所述的方法,其中分离包括(i)在固体支持物上捕获多核苷酸/基因组片段杂交对;和(ii)释放所述多于一个基因组片段以产生富集的靶多核苷酸。

62. 权利要求55所述的方法,其中测序导致在基因组片段碱基的高达55倍理论读段深度的条件下至少95%的基因组片段碱基的至少30倍读段深度。

63. 权利要求55所述的方法,其中测序导致在基因组片段碱基的高达55倍理论读段深度的条件下至少98%的基因组片段碱基的至少30倍读段深度。

64. 权利要求55所述的方法,其中测序导致基因组片段碱基的至少90%的独特读段。

65. 权利要求55所述的方法,其中测序导致基因组片段碱基的至少95%的独特读段。

66. 权利要求55所述的方法,其中测序导致至少90%的基因组片段碱基具有在平均读段深度的约1.5x内的读段深度。

67. 权利要求55所述的方法,其中测序导致至少95%的基因组片段碱基具有在平均读段深度的约1.5x内的读段深度。

68. 权利要求55所述的方法,其中测序导致至少90%的基因组片段具有10%至30%或者70%至90%的GC百分比、具有在平均读段深度的约1.5x内的读段深度。

69. 权利要求55所述的方法,其中测序导致至少约80%的基因组片段具有10%至30%或70%至90%的重复序列或二级结构序列百分比、具有在平均读段深度的约1.5x内的读段深度。

70. 权利要求55所述的方法,其中所述至少5000种多核苷酸中的至少约80%以所述多核苷酸文库平均表现度的至少约1.5倍的量表现。

71. 权利要求55所述的方法,其中所述至少5000种多核苷酸中的至少30%包含具有GC百分比为从10%至30%或70%至90%的多核苷酸。

72. 权利要求55所述的方法,其中所述至少5000种多核苷酸中的至少15%包含具有重

复序列或二级结构序列百分比为10%至30%或70%至90%的多核苷酸。

73. 权利要求55所述的方法,其中所述至少5000种多核苷酸编码至少1000种基因。

74. 权利要求55所述的方法,其中所述多核苷酸文库包含至少100,000种多核苷酸。

75. 权利要求55所述的方法,其中所述多核苷酸文库包含至少700,000种多核苷酸。

76. 权利要求55所述的方法,其中所述至少5000种多核苷酸包含至少一个外显子序列。

77. 权利要求75所述的方法,其中所述至少700,000种多核苷酸包含至少一组多核苷酸,所述至少一组多核苷酸共同包含单个外显子序列。

78. 权利要求77所述的方法,其中所述至少700,000种多核苷酸包含至少150,000组。

具有受控化学计量的多核苷酸文库及其合成

[0001] 交叉引用

[0002] 本申请要求2016年11月18日提交的第62/424,302号美国临时专利申请、2017年8月21日提交的第62/548,307号美国临时专利申请和2017年9月14日提交的第62/558,666号美国临时专利申请的权益,这些申请的每一个都通过引用以其整体并入本文。

[0003] 序列表

[0004] 本申请包含以ASCII格式电子提交的序列表,并特此通过引用将其全部并入本文。所述ASCII副本创建于2017年11月13日,命名为44854-730_601_SL.txt,且大小为5,304字节。

[0005] 背景

[0006] 具有高保真性和低成本的高效化学基因合成在生物技术和医学以及基础生物医学研究中具有核心作用。从头基因合成是用于基础生物学研究和生物技术应用的有效工具。虽然已知多种用于以小规模合成相对较短的片段的方法,但是这些技术通常受限于可扩展性、自动化、速度、准确性和成本。

[0007] 通过引用并入

[0008] 本说明书中提到的所有出版物、专利和专利申请特此通过引用并入,如同指示了每个单独的出版物、专利或专利申请被具体地和单独地通过引用并入一样。

[0009] 简要概述

[0010] 本文提供了多核苷酸文库,所述多核苷酸文库包含至少5000种多核苷酸,其中所述至少5000种多核苷酸中的每一种多核苷酸以一定的量存在,使得在与基因组片段杂交并对杂交的基因组片段测序后,所述多核苷酸文库在基因组片段碱基的高达55倍理论读段深度(read depth)的条件下提供至少90%的基因组片段碱基的至少30倍读段深度。本文还提供了多核苷酸文库,其中多核苷酸文库在基因组片段碱基理论读段深度高达55倍的条件下,提供至少95%的基因组片段碱基的至少30倍读段深度。本文还提供了多核苷酸文库,其中多核苷酸文库在基因组片段碱基理论读段深度高达55倍的条件下,提供至少98%的基因组片段碱基的至少30倍读段深度。本文还提供了多核苷酸文库,其中多核苷酸文库提供基因组片段碱基的至少90%的独特读段。本文还提供了多核苷酸文库,其中多核苷酸文库提供基因组片段碱基的至少95%的独特读段。本文还提供了多核苷酸文库,其中多核苷酸文库提供了至少90%的基因组片段碱基,所述基因组片段碱基具有在平均读段深度的约1.5倍内的读段深度。本文还提供了多核苷酸文库,其中多核苷酸文库提供了至少95%的基因组片段碱基,所述基因组片段碱基具有在平均读段深度的约1.5倍内的读段深度。本文还提供了多核苷酸文库,其中多核苷酸文库提供了至少90%的基因组片段,所述基因组片段具有10%至30%或70%至90%的GC百分比、具有在平均读段深度的约1.5x内的读段深度。本文还提供了多核苷酸文库,其中多核苷酸文库提供了至少约80%的如下基因组片段:所述基因组片段具有从10%至30%或70%至90%的重复序列或二级结构序列百分比、具有在平均读段深度的约1.5x内的读段深度。本文还提供了多核苷酸文库,其中每个基因组片段长度为约100个碱基至约500个碱基。本文还提供了多核苷酸文库,其中所述至少5000种多核

核苷酸中的至少约80%以多核苷酸文库平均表现度 (representation) 的至少约1.5倍的量表现 (represent)。本文还提供了多核苷酸文库,其中所述至少5000种多核苷酸中的至少30%包含具有10%至30%或70%至90%的GC百分比的多核苷酸。本文还提供了多核苷酸文库,其中所述至少5000种多核苷酸中的至少约15%包含具有10%至30%或70%至90%的重复序列或二级结构序列百分比的多核苷酸。本文还提供了多核苷酸文库,其中所述至少5000种多核苷酸编码至少1000种基因。本文还提供了多核苷酸文库,其中多核苷酸文库包含至少100,000种多核苷酸。本文还提供了多核苷酸文库,其中多核苷酸文库包含至少700,000种多核苷酸。本文还提供了多核苷酸文库,其中所述至少5000种多核苷酸包含至少一个外显子序列。本文还提供了多核苷酸文库,其中所述至少700,000种多核苷酸包含至少一组多核苷酸,所述至少一组多核苷酸共同包含单个外显子序列。本文还提供了多核苷酸文库,其中所述至少700,000种多核苷酸包含至少150,000组。

[0011] 本文提供了多核苷酸文库,所述多核苷酸文库包含至少5000种多核苷酸,其中每种多核苷酸长度为约20至200个碱基,其中多于一种 (the plurality of) 多核苷酸编码来自至少1000种预选基因的各外显子的序列,其中每个多核苷酸包含分子标签,其中所述至少5000种多核苷酸中的每种多核苷酸以一定的量存在,使得在与基因组片段杂交并对杂交的基因组片段测序后,多核苷酸文库在基因组片段碱基的高达55倍理论读段深度条件下提供至少90%基因组片段碱基的至少30倍读段深度。本文还提供了多核苷酸文库,其中多核苷酸文库在基因组片段碱基的高达55倍理论读段深度的条件下,提供至少95%的基因组片段碱基的至少30倍读段深度。本文还提供了多核苷酸文库,其中多核苷酸文库在基因组片段碱基的高达55倍理论读段深度的条件下,提供至少98%基因组片段碱基的至少30倍读段深度。本文还提供了多核苷酸文库,其中多核苷酸文库提供基因组片段碱基的至少90%的独特读段。本文还提供了多核苷酸文库,其中多核苷酸文库提供基因组片段碱基的至少95%的独特读段。本文还提供了多核苷酸文库,其中多核苷酸文库提供了至少90%的基因组片段碱基,所述基因组片段碱基具有在平均读段深度的约1.5倍内的读段深度。本文还提供了多核苷酸文库,其中多核苷酸文库提供了至少95%的基因组片段碱基,所述基因组片段碱基具有在平均读段深度的约1.5倍内的读段深度。本文还提供了多核苷酸文库,其中多核苷酸文库提供了大于90%的如下基因组片段:所述基因组片段具有10%至30%或70%至90%的GC百分比、具有在平均读段深度的约1.5倍内的读段深度。本文还提供了多核苷酸文库,其中多核苷酸文库提供了大于约80%的基因组片段,所述基因组片段具有10%至30%或70%至90%的重复序列或二级结构序列百分比、具有在平均读段深度的约1.5倍内的读段深度。本文还提供了多核苷酸文库,其中每个基因组片段长度为约100个碱基至约500个碱基。本文还提供了多核苷酸文库,其中所述至少5000种多核苷酸中的大于约80%以多核苷酸文库平均表现度的至少约1.5倍内的量表现。本文还提供了多核苷酸文库,其中所述至少5000种多核苷酸中的大于30%包含具有10%至30%或70%至90%的GC百分比的多核苷酸。本文还提供了多核苷酸文库,其中所述至少5000种多核苷酸中的大于约15%包含具有10%至30%或70%至90%的重复序列或二级结构序列百分比的多核苷酸。本文还提供了多核苷酸文库,其中多核苷酸文库包含至少100,000种多核苷酸。本文还提供了多核苷酸文库,其中多核苷酸文库包含至少700,000种多核苷酸。本文还提供了多核苷酸文库,其中所述至少种700,000多核苷酸包含至少一组多核苷酸,所述至少一组多核苷酸共同包含单个

外显子序列。本文还提供了多核苷酸文库,其中所述至少700,000种多核苷酸包含至少150,000组。

[0012] 本文提供了用于产生多核苷酸文库的方法,该方法包括:提供编码至少5000种多核苷酸的预定序列;合成所述至少5000种多核苷酸;以及用聚合酶扩增所述至少5000种多核苷酸以形成多核苷酸文库,其中所述至少5000种多核苷酸中大于约80%的多核苷酸以多核苷酸文库平均表现度的至少约2倍内的量表现。本文还提供了方法,其中所述至少5000种多核苷酸中的大于约80%以多核苷酸文库平均表现度的至少约1.5倍内的量表现。本文还提供了方法,其中所述至少5000种多核苷酸中的大于30%包含具有10%至30%或70%至90%的GC百分比的多核苷酸。本文还提供了方法,其中所述至少5000种多核苷酸中的大于约15%包含具有10%至30%或70%至90%的重复序列或二级结构序列百分比的多核苷酸。本文还提供了方法,其中与没有校正错误的预定序列相比,多核苷酸文库具有在800个碱基中小于1个的总错误率。本文还提供了方法,其中预定序列编码至少700,000种多核苷酸。本文还提供了方法,其中所述至少5000种多核苷酸的合成发生在具有表面的结构上,其中所述表面包括多于一个簇,其中每个簇包括多于一个位点;并且其中所述至少5000种多核苷酸中的每一种都从所述多于一个位点的不同位点延伸。本文还提供了方法,其中所述多于一个位点在每个簇中包含高达1000个位点。本文还提供了方法,其中所述多个位点包括每个簇高达200个位点。

[0013] 本文提供了用于多核苷酸文库扩增的方法,该方法包括:获得至少5000种多核苷酸的扩增分布;基于至少一个序列特征将该扩增分布的至少5000种多核苷酸聚类到两个或更多个箱中,其中所述序列特征是GC含量百分比、重复序列含量百分比或二级结构含量百分比;调整在至少一个箱中多核苷酸的相对频率,以产生具有预选表现度的多核苷酸文库;合成具有预选表现度的多核苷酸文库;以及扩增具有预选表现度的多核苷酸文库。本文还提供了方法,其中所述至少一个序列特征是GC含量百分比。本文还提供了方法,其中至少一个序列特征是二级结构含量百分比。本文还提供了方法,其中所述至少一个序列特征是重复序列含量百分比。本文还提供了方法,其中重复序列含量包括具有3个或更多腺嘌呤的序列。本文还提供了方法,其中重复序列含量包括多核苷酸的至少一个末端上的重复序列。本文还提供了方法,其中基于一个或更多个多核苷酸序列结合靶序列的亲合力,将所述多核苷酸聚类至箱中。本文还提供了方法,其中当与调整前的箱中的较低的30%的序列数量相比时,调整后的箱中的较低的30%的序列数量在下游应用中具有至少多50%的表现度。本文还提供了方法,其中当与调整前的箱中的较高的30%的序列数量相比时,调整后的箱中的较高的30%的序列数量在下游应用中具有至少多50%的表达。

[0014] 本文提供了用于对基因组DNA测序的方法,所述方法包括:将本文所述的任何多核苷酸文库与多于一种基因组片段接触;富集至少一种与文库结合的基因组片段,以产生至少一种富集的靶多核苷酸;和对所述至少一种富集的靶多核苷酸测序。本文还提供了方法,其中所述多于一种富集的靶多核苷酸包括cDNA文库。本文还提供了方法,其中所述至少5000种多核苷酸的长度为约80至约200个碱基。本文还提供了方法,其中每种基因组片段长度为约100个碱基至约500个碱基。本文还提供了方法,其中接触发生在溶液中。本文还提供了方法,其中所述至少5000种多核苷酸至少部分与基因组片段互补。本文还提供了分离方法,其中分离包括(i)在固体支持物上捕获多核苷酸/基因组片段杂交对;和(ii)释放所述

多于一种基因组片段以产生富集的靶多核苷酸。本文还提供了方法,其中测序导致在基因组片段碱基的高达55倍理论读段深度的条件下至少95%的基因组片段碱基的至少30倍读段深度。本文还提供了方法,其中测序导致在基因组片段碱基的高达55倍理论读段深度的条件下至少98%的基因组片段碱基的至少30倍读段深度。本文还提供了方法,其中测序导致基因组片段碱基的至少90%的独特读段。本文还提供了方法,其中测序导致基因组片段碱基的至少95%的独特读段。本文还提供了方法,其中测序导致至少90%的基因组片段碱基具有在平均读段深度的约1.5x内的读段深度。本文还提供了方法,其中测序导致至少95%的基因组片段碱基具有在平均读段深度的约1.5x内的读段深度。本文还提供了方法,其中测序导致至少90%的基因组片段具有10%至30%或70%至90%的GC百分比、具有在平均读段深度的约1.5x内的读段深度。本文还提供了方法,其中测序导致至少约80%的基因组片段具有10%至30%或70%至90%的重复序列或二级结构序列百分比、具有在平均读段深度的约1.5x内的读段深度。本文还提供了方法,其中所述至少5000种多核苷酸中的至少约80%以多核苷酸文库平均表现度的至少约1.5倍的量表达。本文还提供了方法,其中所述至少5000种多核苷酸中的至少30%包含具有10%至30%或70%至90%的GC百分比的多核苷酸。本文还提供了方法,其中所述至少5000种多核苷酸中的至少15%包含具有10%至30%或70%至90%的重复序列或二级结构序列百分比的多核苷酸。本文还提供了方法,其中至少5000种多核苷酸编码至少1000种基因。本文还提供了方法,其中多核苷酸文库包含至少100,000种多核苷酸。本文还提供了方法,其中多核苷酸文库包含至少700,000种多核苷酸。本文还提供了方法,其中所述至少5000种多核苷酸包含至少一个外显子序列。本文还提供了方法,其中所述至少700,000种多核苷酸包含至少一组多核苷酸,所述至少一组多核苷酸共同包含单个外显子序列。本文还提供了方法,其中所述至少700,000种多核苷酸包含至少150,000组。

附图说明

[0015] 图1A描绘了示意性的工作流程,包括应用第一多核苷酸文库,测量来自应用输出的偏倚,设计和合成第二受控化学计量的多核苷酸文库,以及应用第二多核苷酸文库以产生期望的表现度输出。

[0016] 图1B描绘了用靶结合多核苷酸文库富集靶多核苷酸的示意图。

[0017] 图1C描绘了用于核酸样品的富集和测序的示例性工作流程。

[0018] 图2描绘了从簇扩增产生多核苷酸文库的示意图。

[0019] 图3A描绘了用于靶向和富集的一对多核苷酸。所述多核苷酸包含互补性靶结合(插入)序列,以及引物结合位点。

[0020] 图3B描绘了用于靶向和富集的一对多核苷酸。所述多核苷酸包含互补性靶序列结合(插入)序列、引物结合位点和非靶序列。

[0021] 图4A描绘了多核苷酸与较大的多核苷酸的靶序列的结合构造。靶序列比多核苷酸结合区短,并且多核苷酸结合区(或插入序列)相对于靶序列偏移(offset),并且还与相邻序列的一部分结合。

[0022] 图4B描绘了多核苷酸与较大的多核苷酸的靶序列的结合构造。靶序列长度小于或等于多核苷酸结合区,并且多核苷酸结合区在靶序列的中心,并且还与相邻序列的一部分

结合。

[0023] 图4C描绘了多核苷酸与较大的多核苷酸的靶序列的结合构造。靶序列比多核苷酸结合区稍长,并且多核苷酸结合区在靶序列的中心,且每侧有缓冲区。

[0024] 图4D描绘了多核苷酸与较大的多核苷酸的靶序列的结合构造。靶序列比多核苷酸结合区长,并且两个多核苷酸的结合区重叠以跨越靶序列。

[0025] 图4E描绘了多核苷酸与较大的多核苷酸的靶序列的结合构造。靶序列比多核苷酸结合区长,并且两个多核苷酸的结合区重叠以跨越靶序列。

[0026] 图4F描绘了多核苷酸与较大的多核苷酸的靶序列的结合构造。靶序列比多核苷酸结合区长,并且两个多核苷酸的结合区不重叠以跨越靶序列,留下空位405。

[0027] 图4G描绘了多核苷酸与较大的多核苷酸的靶序列的结合构造。靶序列比多核苷酸结合区长,并且三个多核苷酸的结合区重叠以跨越靶序列。

[0028] 图5呈现了展示如本文公开的基因合成的示例性处理工作流程的步骤图。

[0029] 图6示出了计算机系统。

[0030] 图7是示出计算机系统体系结构的框图。

[0031] 图8是展示被配置为结合了多于一个计算机系统、多于一个蜂窝电话和个人数据助理以及网络附加存储(NAS)的网络的图。

[0032] 图9是使用共享虚拟地址存储空间的多处理器计算机系统的框图。

[0033] 图10是具有256个簇的板的图像,每个簇具有121个位点且多核苷酸从其延伸。

[0034] 图11A是来自240个簇(每个簇具有121种多核苷酸)的29,040种独特多核苷酸的合成的遍及板的多核苷酸表现度的图(多核苷酸频率对丰度,按测量的吸光度)。

[0035] 图11B是遍及每个单独的簇的多核苷酸频率的测量值对丰度吸光度(按测量的吸光度)的图,且对照簇以方框标识。

[0036] 图12是遍及四个单独的簇的多核苷酸频率对丰度(按测量的吸光度)的测量值的图。

[0037] 图13A是来自240个簇(每个簇具有121种多核苷酸)的29,040种独特多核苷酸的合成的遍及板的中靶(on)频率对错误率的图。

[0038] 图13B是遍及每个单独簇的多核苷酸错误率对频率的测量值的图,且对照簇以方框标识。

[0039] 图14是遍及四个簇的多核苷酸频率对错误率的测量值的图。

[0040] 图15是作为多核苷酸数量的量度的GC含量对每个多核苷酸百分比的图。

[0041] 图16提供了来自用两种不同的聚合酶进行PCR的结果的图。每个图描绘了多核苷酸的数量(0至2,000)对观察到的频率(每100,000个中测量的计数“0至35”)。

[0042] 图17提供了记录的扩增后多核苷酸群体一致性的定量图。

[0043] 图18描绘了展示过度扩增对序列失落(dropout)的影响的图。

[0044] 图19描绘了汇集的未扩增的和扩增的多核苷酸群体中每种多核苷酸频率(每100,000个读段)的GC含量百分比的图。

[0045] 图20是扩增簇后两次独立运行的每种多核苷酸频率(每100,000个读段)的GC含量百分比的图。

[0046] 图21A是GC平衡的多核苷酸文库的每种多核苷酸频率的GC含量百分比的图。

- [0047] 图21B是重度高和低GC偏倚的多核苷酸文库的每种多核苷酸频率的GC含量百分比的图。
- [0048] 图21C是轻度高和低GC偏倚的多核苷酸文库的每种多核苷酸频率的GC含量百分比的图。
- [0049] 图21D是低GC偏倚的多核苷酸文库的每种多核苷酸频率的GC含量百分比的图。
- [0050] 图21E是高GC偏倚的多核苷酸文库的每种多核苷酸频率的GC含量百分比的图。
- [0051] 图22是理论上13,000丛(plex)多核苷酸文库(序列包含15%至85%GC含量)的每种多核苷酸频率的GC含量百分比的图。
- [0052] 图23是GC平衡的多核苷酸文库的多核苷酸数量对多核苷酸频率(每100,000个读段)的图。
- [0053] 图24A是显示与单分散的文库的理论最大值相比,GC平衡的多核苷酸文库获得80%测序覆盖所需的取样量的图。
- [0054] 图24B是显示与单分散的文库的理论最大值相比,GC平衡的多核苷酸文库获得90%测序覆盖所需的取样量的图。
- [0055] 图25是包含长度为80个核苷酸的多核苷酸的文库的多核苷酸数量对多核苷酸频率(每1,000,000个读段中的计数)的图。
- [0056] 图26是包含长度为120个核苷酸的多核苷酸的文库的多核苷酸数量对多核苷酸频率(每1,000,000个读段中的计数)的图。
- [0057] 图27描绘了显示长度为80个和120个核苷酸的GC平衡的多核苷酸文库的多核苷酸的平均频率(每1,000,000个读段)的图。
- [0058] 图28描绘了显示PCR扩增循环数、GC含量和DNA聚合酶的选择对多核苷酸序列表现度的影响的图。
- [0059] 图29是序列失落作为两种不同的高保真性聚合酶的扩增循环的函数的图。
- [0060] 图30描绘了示出不同的DNA聚合酶对序列表现度的影响。用DNA聚合酶1或DNA聚合酶2将同一多核苷酸文库扩增15个循环。
- [0061] 图31A描绘了使用没有受控化学计量的外显子组探针文库来实现靶序列的特定读段深度所需的过度测序的量。
- [0062] 图31B描绘了当与没有受控化学计量的外显子组探针文库相比时,使用具有受控化学计量的外显子组探针文库来实现靶序列的特定读段深度所需的过度测序的减少。
- [0063] 图32A是比较物外显子组探针试剂盒A和受控化学计量的探针文库1的具有1x、20x或30x测序读段深度(X覆盖)的碱基百分比的图。
- [0064] 图32B是比较物外显子组探针试剂盒和受控化学计量的探针文库1的组的具有1x或10x测序读段深度(X覆盖)(以4.5Gb的测序归一化)的碱基百分比的图。
- [0065] 图33描绘了不同规模的多核苷酸探针文库的合成随文库中多核苷酸数量的变化。
- [0066] 图34描绘了覆盖(碱基数量)随着基于比较物阵列的探针文库的读段深度的变化与覆盖(碱基数量)随着受控化学计量的探针文库2的读段深度的变化之间的比较。
- [0067] 图35A描述了对于GC含量在10-30%之间和30-50%之间的靶,覆盖随着基于比较物阵列的探针文库的读段深度的变化与覆盖随着受控化学计量的探针文库2的读段深度的变化之间的比较。

[0068] 图35B描绘了对于GC含量在大于50-70%之间和大于70-90%之间的靶,覆盖随着基于比较物阵列的探针文库的读段深度的变化与覆盖随着受控化学计量的探针文库2的读段深度的变化之间的比较。

[0069] 图36A描绘了基于比较物阵列的探针文库的中靶率百分比(%)与0.1x、1x和3x浓度的受控化学计量的探针文库3的中靶率百分比(%)之间的比较。

[0070] 图36B描绘了基于比较物阵列的探针文库的读段深度与0.1x、1x和3x浓度的受控化学计量的探针文库3的读段深度之间的比较。

[0071] 图37描绘了比较物外显子组相比于受控化学计量的探针文库4之间的独特读段和靶碱基在1X、20X和30X的读段深度的百分比的比较。

[0072] 图38描绘了比较物外显子组相比于受控化学计量的探针文库4之间的覆盖的碱基和靶碱基在1X、20X和30X的读段深度的百分比的比较。

[0073] 详细描述

[0074] 除非另有指示,本公开内容采用本领域技术范围内的常规分子生物学技术。除非另有定义,本文使用的所有技术和科学术语具有与本领域普通技术人员通常理解的相同的含义。

[0075] 本文提供了用于设计、合成和控制大型多核苷酸文库的化学计量的方法。当对第一群体的多核苷酸进行初步应用步骤,例如用于扩增,作为用于富集和基因合成的捕获探针时,多核苷酸群体的后续扩增反应可以由于多核苷酸序列的差异而导致偏倚的代表性输出,从而导致某些多核苷酸的表现比其他多核苷酸更丰富。图1A。测量了从该初步应用输出观察到的结果偏倚,并将其用于控制具有预选化学计量的多核苷酸的第一群体,该预选化学计量例如,群体中多核苷酸的相对频率,考虑任何数量的序列特征,例如GC含量、重复序列、尾随腺嘌呤、二级结构、对靶序列结合的亲和力或修饰的核苷酸。在修改多核苷酸的化学计量后,用预选的化学计量设计和合成第二多核苷酸群体,以校正与应用步骤相关的不期望的偏倚效应。在一些情况下,对第二受控化学计量的多核苷酸群体进行应用步骤,例如PCR扩增,导致平衡的输出,例如具有高度一致性表现度或者具有预选的表现度偏移的非一致性表现度的扩增的多核苷酸群体。参见图1A下图。在一些情况下,本文描述的方法包括控制多核苷酸探针的序列表现度,使得多核苷酸群体提供高度一致的靶序列捕获频率(图1B)。例如,多核苷酸100的样品包含靶多核苷酸101。样品100与靶结合多核苷酸103在适当条件102下的接触导致杂交对104的形成,将杂交对104与样品100中的非靶多核苷酸分离。对104的变性和分离释放富集的靶多核苷酸107用于下游应用,例如测序。本文还提供了用于与基因组DNA杂交的从头合成的多核苷酸,例如在测序过程的情况下。在示例性测序工作流程的第一步骤中(图1C),通过机械或酶的剪切将包含靶多核苷酸的核酸样品108片段化,以形成片段文库109。连接任选地包含引物序列和/或条形码的衔接子115以形成衔接子标记的文库110。然后任选地扩增该文库,并与杂交到靶多核苷酸的靶结合多核苷酸117以及阻止靶结合多核苷酸117和衔接子115之间杂交的封闭多核苷酸116杂交。捕获靶多核苷酸-靶结合多核苷酸杂交对112并去除靶结合多核苷酸117允许分离/富集靶多核苷酸113,然后任选地将靶多核苷酸113扩增和测序114。

[0076] 定义

[0077] 贯穿本公开内容,数值特征以范围格式呈现。应该理解,范围格式的描述仅仅是为

了方便和简洁,不应该被解释为对任何实施方案的范围的僵化限制。因此,范围的描述应该被认为已经具体公开了所有可能的子范围以及在该范围内到下限单位的十分之一的单个数值,除非上下文另有明确指示。例如,范围如1至6的描述应被认为已经具体公开了子范围如从1至3、从1至4、从1至5、从2至4、从2至6、从3至6等,以及该范围内的单个值,例如1.1、2、2.3、5和5.9。无论范围的宽度如何这都适用。这些介于其间的范围的上限和下限可以被独立地包括在较小的范围内,并且也包括在本发明内,受所述范围内任何特定排除的限制。在所述范围包括一个或两个限制的情况下,排除这些包括的限制中的任一个或两个之外的范围也包括在本发明中,除非上下文另有明确规定。

[0078] 本文使用的术语仅用于描述特定实施方案的目的,并不旨在限制任何实施方案。如本文所使用的,单数形式“一个(a)”、“一个(an)”和“该(the)”旨在也包括复数形式,除非上下文另有明确指示。还应当理解,当在本说明书中使用时,术语“包括/包含(comprise)”和/或“括/包含(compring)”指定所述特征、整数、步骤、操作、要素和/或组成部分的存在,但是不排除一个或更多个其他特征、整数、步骤、操作、要素、组成部分和/或其组的存在或添加。如本文所使用的,术语“和/或”包括相关列出的项目的一项或更多项的任何和所有组合。

[0079] 除非特别说明或从上下文中显而易见,如本文所使用的,关于数量或数量的范围的术语“约”应理解为意指所述数量及其 $\pm 10\%$ 的数量,或者对于一个范围所列的值,低于所列的下限 10% ,高于所列的上限 10% 。

[0080] 如本文所使用的,术语“预选序列”、“预定义序列”或“预定序列”可互换使用。这些术语意指在合成或组装聚合物之前,聚合物的序列是已知的和选择的。具体而言,本文描述的主要关于核酸分子制备的本发明的各个方面,寡核苷酸或多核苷酸的序列在核酸分子的合成或组装之前是已知的和选择的。

[0081] 术语核酸包括双链或三链核酸以及单链分子。在双链或三链核酸中,核酸链不必为共延伸的(coextensive)(即,双链核酸不需要沿着两条链的整个长度为双链的)。除非另有说明,核酸序列,当提供时,以5'至3'方向列出。本文描述的方法提供了分离的核酸的产生。本文描述的方法还提供了分离和纯化的核酸的产生。多核苷酸的长度,当提供时,描述为碱基的数量并缩写,例如nt(核苷酸)、bp(碱基)、kb(千碱基)或Gb(千兆碱基)。

[0082] 本文提供了用于产生合成(即从头合成或化学合成)多核苷酸的方法和组合物。术语寡核苷酸(oligonucleic acid)、寡核苷酸(oligonucleotide)、寡核苷酸(oligo)和多核苷酸在全文中被定义为同义的。本文所述的合成多核苷酸文库可以包含共同编码一种或更多种基因或基因片段的多于一种多核苷酸。在一些情况下,多核苷酸文库包含编码或非编码序列。在一些情况下,多核苷酸文库编码多于一种cDNA序列。作为cDNA序列基础的参考基因序列可能包含内含子,而cDNA序列不包含内含子。本文描述的多核苷酸可以编码来自生物体的基因或基因片段。示例性生物体包括但不限于原核生物(例如细菌)和真核生物(例如小鼠、兔、人类和非人类灵长目动物)。在一些情况下,多核苷酸文库包含一种或更多种多核苷酸,所述一种或更多种多核苷酸中的每一种编码多于一种外显子的序列。本文所述文库中的每种多核苷酸可以编码不同的序列,即非相同的序列。在一些情况下,本文所述文库中的每种多核苷酸包含至少一个与文库中另一多核苷酸序列互补的部分。除非另有说明,本文描述的多核苷酸序列可以包含DNA或RNA。本文描述的多核苷酸文库可以包含至少

10、20、50、100、200、500、1,000、2,000、5,000、10,000、20,000、30,000、50,000、100,000、200,000、500,000、1,000,000或多于1,000,000种多核苷酸。本文描述的多核苷酸文库可以具有不多于10、20、50、100、200、500、1,000、2,000、5,000、10,000、20,000、30,000、50,000、100,000、200,000、500,000或不多于1,000,000种多核苷酸。本文描述的多核苷酸文库可以包含10至500、20至1000、50至2000、100至5000、500至10,000、1,000至5,000、10,000至50,000、100,000至500,000或50,000至1,000,000种多核苷酸。本文描述的多核苷酸文库可包含约370,000;400,000;500,000个或更多种不同的多核苷酸。

[0083] 本文提供了用于产生合成(即从头合成)基因的方法和组合物。包含合成基因的文库可以通过本文其他地方更详细描述的各种方法构建,例如PCA、非PCA基因组装方法或分级基因组装,组合(“拼接(stitching)”)两个或更多个双链多核苷酸以产生更大的DNA单元(即底架(chassis))。大型构建体的文库可以包含至少1、1.5、2、3、4、5、6、7、8、9、10、15、20、30、40、50、60、70、80、90、100、125、150、175、200、250、300、400、500kb长度或更长的多核苷酸。大的构建体可以由独立选择的大约5000、10000、20000或50000碱基对的上限来限定。任何数量的编码多肽区段的核苷酸序列的合成,包括编码非核糖体肽(NRP)的序列,编码非核糖体肽合成酶(NRPS)模块和合成变体的序列,其它模块蛋白的多肽区段,例如抗体,来自其它蛋白家族的多肽区段的序列;包括非编码DNA或RNA,例如调控序列,例如启动子、转录因子、增强子、siRNA、shRNA、RNAi、miRNA、微RNA衍生的核仁小RNA,或任何感兴趣的功能性或结构性DNA或RNA单元。以下是多核苷酸的非限制性实例:基因或基因片段的编码或非编码区、基因间DNA、由连锁分析定义的基因座(loci)(基因座(locus))、外显子、内含子、信使RNA(mRNA)、转移RNA、核糖体RNA、短干扰RNA(siRNA)、短发夹RNA(shRNA)、微RNA(miRNA)、核仁小RNA、核酶、互补DNA(cDNA)(互补DNA是mRNA的DNA表示,通常通过信使RNA(mRNA)的逆转录或通过扩增获得)、合成或通过扩增产生的DNA分子、基因组DNA、重组多核苷酸、支链多核苷酸、质粒、载体、任何序列的分离的DNA、任何序列的分离的RNA、核酸探针和引物。编码本文提及的基因或基因片段的cDNA可以包含至少一个编码在对应的基因组序列中不存在介于其间的内含子序列的外显子序列的区域。可选择地,cDNA的对应基因组序列可能首先缺少内含子序列。

[0084] 用于扩增反应的小多核苷酸群体的从头合成

[0085] 本文描述了从表面例如板合成多核苷酸的方法。在一些情况下,多核苷酸在用于多核苷酸延伸的位点的簇上合成,被释放、并且然后进行扩增反应,例如PCR。图2中描绘了从簇合成多核苷酸的示例性工作流程。硅板201包括多于一个簇203。在每个簇中有多于一个位点221。多核苷酸在板201上从簇203从头合成207。多核苷酸被切割211并从板上取下213,以形成释放的多核苷酸群体215。然后,释放的多核苷酸群体215被扩增217,以形成扩增的多核苷酸文库219。

[0086] 本文提供了方法,其中在簇上合成的多核苷酸的扩增与没有这种簇布置的结构的整个表面上的多核苷酸的扩增相比,提供了对多核苷酸表现度的增强控制。在一些情况下,从具有用于多核苷酸延伸的位点的簇布置的表面合成的多核苷酸的扩增提供了克服由于大多核苷酸群体的重复合成而对表现度的负面影响。由于大多核苷酸群体的重复合成而对表现度的示例性负面影响包括但不限于由高/低GC含量、重复序列、尾随腺嘌呤、二级结构、对靶序列结合的亲和力或多核苷酸序列中的修饰核苷酸导致的扩增偏倚。

[0087] 与在没有簇布置的遍及板上扩增多核苷酸对照,簇扩增可以导致平均值周围更紧密的分布。例如,如果随机取样100,000个读段,每个序列平均8个读段将产生平均值的约1.5X的分布的文库。在一些情况下,单个簇扩增导致平均值的最多约1.5X、1.6X、1.7X、1.8X、1.9X或2.0X。在一些情况下,单个簇扩增导致平均值的至少约1.0X、1.2X、1.3X、1.5X、1.6X、1.7X、1.8X、1.9X或2.0X。

[0088] 当与遍及板的扩增相比时,本文所述的簇扩增方法可以导致需要较少的测序来获得等同的序列表现度的核苷酸文库。在一些情况下,需要少至少10%、至少20%、至少30%、至少40%、至少50%、至少60%、至少70%、至少80%、至少90%或至少95%的测序。在一些情况下,需要少高达10%、高达20%、高达30%、高达40%、高达50%、高达60%、高达70%、高达80%、高达90%或高达95%的测序。有时,与遍及板的扩增相比,簇扩增后需要少30%的测序。在一些情况下,多核苷酸的测序通过高通量测序如通过下一代测序来验证。测序文库的测序可以用任何合适的测序技术进行,包括但不限于单分子实时 (SMRT) 测序、Polony 测序、连接测序、可逆终止子测序、质子检测测序、离子半导体测序、纳米孔测序、电子测序、焦磷酸测序、Maxam-Gilbert测序、链终止 (例如Sanger) 测序、+S测序或合成测序。单个核苷酸或多核苷酸被识别或“读段”的次数被定义为测序深度或读段深度。在某些情况下,读段深度被称为倍覆盖 (bold coverage), 例如55倍 (或55X) 覆盖,任选择地描述碱基的百分比。

[0089] 在一些情况下,与遍及板的扩增相比,来自簇布置的扩增导致较少的失落或扩增产物测序后未检测到的序列。失落可以是AT和/或GC。在一些情况下,失落数最多约为多核苷酸群体的1%、2%、3%、4%或5%。在一些情况下,失落数为零。

[0090] 如本文所述的簇包括用于多核苷酸合成的离散的、非重叠的位点的集合。簇可以包括大约50-1000、75-900、100-800、125-700、150-600、200-500或300-400个位点。在一些情况下,每个簇包括121个位点。在一些情况下,每个簇包括大约50-500、50-200、100-150个位点。在一些情况下,每个簇包括至少大约50、100、150、200、500、1000或更多个位点。在一些情况下,单个板包括100、500、10000、20000、30000、50000、100000、500000、700000、100000或更多个位点。位点可以是点、孔、微孔、通道或柱 (post)。在一些情况下,每个簇具有至少1X、2X、3X、4X、5X、6X、7X、8X、9X、10X或更多冗余的支持具有相同序列的多核苷酸延伸的独立特征。

[0091] 具有受控化学计量的多核苷酸文库的设计

[0092] 本文提供了设计和合成多核苷酸文库的方法,其中将每种多核苷酸种类 (即,与文库中的另一种多核苷酸具有不同的序列) 的量 (或化学计量) 调节到预定量,以便在下游应用中控制期望的结果。因此,本文提供了受控和预定的修改多核苷酸种类化学计量的方法。例如,扩增反应后的多核苷酸种类分布可以通过使用本文所述的方法来控制。预先选择多核苷酸种类分布,以便提供靶序列的高度一致性捕获,例如,使用一组多核苷酸进行基于杂交的分析,例如测序分析。此外,本文描述的方法提供了设计具有一个或多个序列特征的序列的多核苷酸文库,该序列特征通常由于某些“有问题的”多核苷酸序列的结构特征,会导致非一致性的扩增产物或捕获产物,其中“有问题的”多核苷酸序列包括与在多核苷酸文库的应用中产生偏倚相关的一个或多个性质。使用本文所述方法控制化学计量的示例性“有问题的”多核苷酸序列性质包括但不限于高或低GC或AT含量、重复序列、尾随腺嘌呤、二级结构、对靶序列结合 (用于扩增、富集或检测) 的亲合力、稳定性、解链温度、生物活性、组

装成更大片段的能力、包含修饰核苷酸或核苷酸类似物的序列、或产生基于预测或经验数据的序列的第二多核苷酸文库的多核苷酸的任何其他性质。在一些情况下,获得控制化学计量的序列文库,并基于一个或更多个序列特征将序列文库组织或聚类(分箱)成两个或更多个预定义组(箱)。在一些情况下,两个或更多个箱表示单独的独特序列。在一些情况下,箱表示基于所定义的一个或更多个序列特征的值的范围,每个序列特征包含总序列的至少10%、20%、30%、40%、50%、60%、70%、80%、90%、95%或至少99%。在一些情况下,箱代表每个包含最多10%、20%、30%、40%、50%、60%、70%、80%、90%、95%或少于100%的总序列的值的范围。在一个实例中,箱可以按%GC含量定义,多于一个箱以5%的增量表示25-75%的范围(例如,25-29%、30-34%、35-39%等),一个箱表示小于25%,而一个箱表示大于75%的GC含量。指定表示每个箱中所有序列的分子化学计量的每个箱的丰度值。在一些情况下,丰度值最初被设置为100,导致每个箱中序列的相等表达。在一些情况下,化学计量的控制是通过使用获得的应用偏倚数据来增加、减少或保持每个箱的丰度值来实现的。还采用了符合本说明书的其他调整序列丰度的方法。在一些情况下,先前获得的分布被用于确定初始丰度值。

[0093] 在一些情况下,应用偏倚数据是通过预测算法获得的。应用偏倚数据可以凭经验获得,或者从未受控制的或先前受控制的化学计量文库中获得。例如,应用偏倚数据是从多核苷酸文库的扩增中获得的;将扩增后每个箱的多核苷酸频率相对于%GC箱作图,以建立作为%GC含量的函数的扩增偏倚。在另一个实例中,在用多核苷酸探针文库富集靶序列后,从下一代测序(NGS)数据中获得应用偏倚数据;每个靶基因的读段被用于将探针序列分类到箱中;将每个靶基因的读段针对NGS读段箱的数量作图,以建立作为多核苷酸探针序列的函数的NGS测序偏倚。在另一个实例中,在用含有多核苷酸文库的载体处理细胞后,可以从细胞测定输出如荧光中获得应用偏倚数据;在荧光细胞中识别的每个序列的读段被用于将探针序列分类到箱中;将每个序列的读段相对于读段箱的数量作图,以建立作为多核苷酸探针序列的函数的偏倚。

[0094] 在控制化学计量后,合成修改的序列文库以产生受控化学计量的多核苷酸文库。在一些情况下,受控化学计量的文库被用于下游应用。在一些情况下,来自具有受控化学计量的多核苷酸文库的下游应用的数据被用于对文库进行额外轮次的化学计量修改。

[0095] 具有受控化学计量的GC含量的多核苷酸文库的产生

[0096] 本文提供了合成多核苷酸文库的方法,所述多核苷酸文库具有定义的性质,例如GC含量、重复序列、尾随腺嘌呤、二级结构、对靶序列结合的亲和力或修饰的核苷酸,以产生基于预测或经验数据的多核苷酸的第二群体。例如,在选择多核苷酸文库进行合成以导致扩增后定义的GC含量的情况下,在合成阶段根据GC含量调整文库中多核苷酸的种类表现度导致扩增后改善的多核苷酸表达。多核苷酸文库中的GC含量可以是至少10%、20%、30%、40%、50%、60%、70%、80%、90%、95%或多于95%。在一些情况下,多核苷酸文库中的GC含量最多为10%、20%、30%、40%、50%、60%、70%、80%、90%、95%或小于100%。在一些情况下,GC含量在约5-95%、10-90%、30-80%、40-75%或50-70%的范围内。

[0097] 本文描述的多核苷酸文库可以针对它们的GC含量进行调整。在一些情况下,多核苷酸文库偏向高GC含量。例如,设计了一种文库,其中增加的多核苷酸频率具有在约40%至约90%的范围内的GC含量。在一些情况下,多核苷酸文库含有较低的GC含量。例如,设计了

一种文库,其中增加的多核苷酸频率具有在约10%至约60%的范围内的GC含量。文库可以被设计成偏向高和低GC含量。例如,文库可以被设计成其中增加的多核苷酸频率具有主要在约10%和约30%的范围内,以及在约70%至约90%的范围内的GC含量。在一些情况下,文库偏向于一致性的GC含量。例如,多核苷酸频率是一致性的,且GC含量在约10%至约90%的范围内。在一些情况下,文库包含具有约10%至约95%GC百分比的多核苷酸。在一些情况下,本文描述的文库包含具有大于30%的具有10%至30%或70%至90%的GC百分比的不同多核苷酸的多核苷酸。在一些情况下,本文所述的文库包含具有少于约15%的具有10%至30%或60%至90%的GC百分比的多核苷酸的多核苷酸。

[0098] 在一些情况下,通过组合至少两个具有不同GC含量的多核苷酸文库进行具有指定GC含量的多核苷酸文库的产生。在一些情况下,至少2、3、4、5、6、7、10或多于10个多核苷酸文库被组合以产生具有指定GC含量的多核苷酸群体。在一些情况下,不多于2、3、4、5、6、7或10个多核苷酸文库被组合以产生具有指定GC含量的非相同的多核苷酸的群体。

[0099] 在一些情况下,GC含量是通过合成每簇更少或更多的多核苷酸来调整的。例如,在单个簇上合成至少25、50、100、200、300、400、500、600、700、800、900、1000或多于1000个非相同的多核苷酸。在一些情况下,在单个簇上合成不多于约50、100、200、300、400、500、600、700、800、900、1000个非相同的多核苷酸。在一些情况下,在单个簇上合成50至500个非相同的多核苷酸。在一些情况下,在单个簇上合成100至200个非相同的多核苷酸。在一些情况下,在单个簇上合成约100、约120、约125、约130、约150、约175或约200个非相同的多核苷酸。

[0100] 在一些情况下,通过合成不同长度的非相同的多核苷酸来调整GC含量。例如,每个合成的非相同的多核苷酸的长度可以是至少或约至少10、15、20、25、30、35、40、45、50、100、150、200、300、400、500、2000个核苷酸或更多。合成的非相同的多核苷酸的长度可以是最多或大约最多2000、500、400、300、200、150、100、50、45、35、30、25、20、19、18、17、16、15、14、13、12、11、10个核苷酸或更少。合成的每个非相同的多核苷酸的长度可以落入从10-2000、10-500、9-400、11-300、12-200、13-150、14-100、15-50、16-45、17-40、18-35和19-25。

[0101] 具有受控化学计量的重复序列含量的多核苷酸文库的产生

[0102] 本文描述的多核苷酸文库可以以指定的重复序列分布合成。在一些情况下,调整多核苷酸文库的重复序列含量导致改善的多核苷酸表现度。

[0103] 重复序列可以是单个核苷酸的重复或两个或更多个核苷酸的区段重复。在一些情况下,重复序列为至少2、3、4、5、6、7、8、9或至少10个核苷酸。在一些情况下,重复序列为最多2、3、4、5、6、7、8、9或最多10个核苷酸。在一些情况下,核苷酸区段包含至少2、3、4、5、10、15、25、50、100、200、500或至少1000个核苷酸。在一些情况下,核苷酸区段包含最多2、3、4、5、10、15、25、50、100、200、500或最多1000个核苷酸。重复序列可以位于较大的合成多核苷酸的内部或末端位置。末端位置可以在多核苷酸的5'末端、3'末端、或5'末端和3'末端两者附近。在一些情况下,重复序列位于末端的至少1、2、3、4、5、6、7、8、9或至少10个核苷酸内。在一些情况下,重复序列位于末端的最多1、2、3、4、5、6、7、8、9或最多10个核苷酸内。在一些情况下,重复核苷酸是腺嘌呤。在一些情况下,重复序列位于多核苷酸末端,例如聚腺苷尾部。

[0104] 多核苷酸文库中重复序列的含量可以为至少10%、20%、30%、40%、50%、60%、

70%、80%、90%、95%或多于95%。在一些情况下,多核苷酸文库中重复序列的含量为最多10%、20%、30%、40%、50%、60%、70%、80%、90%、95%或不多于100%。在一些情况下,重复序列含量在约5-95%、10-90%、30-80%、40-75%或50-70%的范围内。

[0105] 多核苷酸文库可以被调整其重复序列含量。在一些情况下,多核苷酸文库偏向于高重复序列含量。例如,设计了一种文库,其中增加的多核苷酸频率具有在约40%至约90%范围内的重复序列含量。在一些情况下,多核苷酸文库包含低重复序列含量。例如,设计了一种文库,其中增加的多核苷酸频率具有在约10%至约60%范围内的重复序列含量。文库可以被设计成偏向高和低重复序列含量。例如,文库可以被设计成其中增加的多核苷酸频率具有主要在约10%和约30%的范围内,以及在约70%至约90%的范围内的重复序列含量。在一些情况下,文库偏向于一致性的重复序列含量。例如,多核苷酸频率是一致性的,且重复序列含量在约10%至约90%的范围内。在一些情况下,文库包含具有重复序列百分比为约10%至约95%的多核苷酸。在一些情况下,本文描述的文库包含具有大于30%的具有10%至30%或70%至90%的重复序列百分比的不同多核苷酸的多核苷酸。在一些情况下,本文所述的文库包含具有少于约15%的有10%至30%或60%至90%的重复序列百分比的多核苷酸的多核苷酸。

[0106] 在一些情况下,产生具有指定重复序列含量的多核苷酸文库通过组合至少两个具有不同重复序列含量的多核苷酸文库来发生。在一些情况下,至少2、3、4、5、6、7、10或多于10个多核苷酸文库被组合以产生具有指定重复序列含量的多核苷酸群体。在一些情况下,不多于2、3、4、5、6、7或10个多核苷酸文库被组合以产生具有指定重复序列含量的非相同的多核苷酸的群体。

[0107] 在一些情况下,通过合成每个簇更少或更多的多核苷酸来调整重复序列含量。例如,在单个簇上合成至少25、50、100、200、300、400、500、600、700、800、900、1000或多于1000种非相同的多核苷酸。在一些情况下,在单个簇上合成不多于约50、100、200、300、400、500、600、700、800、900、1000种非相同的多核苷酸。在一些情况下,在单个簇上合成50至500种非相同的多核苷酸。在一些情况下,在单个簇上合成100至200种非相同的多核苷酸。在一些情况下,在单个簇上合成约100、约120、约125、约130、约150、约175或约200种非相同的多核苷酸。

[0108] 在某些情况下,通过合成不同长度的非相同的多核苷酸来调整重复序列含量。例如,每个合成的非相同的多核苷酸的长度可以为至少或约至少10、15、20、25、30、35、40、45、50、100、150、200、300、400、500、2000个核苷酸或更多。合成的非相同的多核苷酸的长度可以为最多或约最多2000、500、400、300、200、150、100、50、45、35、30、25、20、19、18、17、16、15、14、13、12、11、10个核苷酸或更少。合成的每个非相同的多核苷酸的长度可以落入从10-2000、10-500、9-400、11-300、12-200、13-150、14-100、15-50、16-45、17-40、18-35和19-25。

[0109] 具有受控化学计量的二级结构含量的多核苷酸文库的产生

[0110] 本文描述的多核苷酸文库可以以指定的二级结构含量合成。在一些情况下,调整多核苷酸文库的二级结构含量导致改善的多核苷酸的表现度。

[0111] 二级结构可以在一条或更多条多核苷酸链中包含形成结构诸如螺旋(例如 α 螺旋)、 β 折叠、茎环、假结、同二聚体或异二聚体的三个或更多个核苷酸。茎环可以是发夹环、内环、凸起或多环。二级结构类型及其形成潜力可以从序列数据中预测。线性序列折叠或杂

交成二级结构可以发生在多核苷酸附接到固体支持物时,或裂解到溶液中后。

[0112] 多核苷酸文库中的二级结构含量可以为至少10%、20%、30%、40%、50%、60%、70%、80%、90%、95%或多于95%。在一些情况下,多核苷酸文库中的二级结构含量为最多10%、20%、30%、40%、50%、60%、70%、80%、90%、95%,或不多于100%。在一些情况下,二级结构含量在约5-95%、10-90%、30-80%、40-75%或50-70%的范围内。

[0113] 多核苷酸文库可以被调整其二级结构含量。在一些情况下,多核苷酸文库偏向高二级结构含量。例如,设计了一种文库,其中增加的多核苷酸频率具有在约40%至约90%范围内的二级结构含量。在一些情况下,多核苷酸文库包含低二级结构含量。例如,设计了一种文库,其中增加的多核苷酸频率具有在约10%至约60%范围内的二级结构含量。文库可以被设计成偏向高和低二级结构含量。例如,文库可以被设计成其中增加的多核苷酸频率具有主要在约10%和约30%范围内和在约70%至约90%范围内的二级结构含量。在一些情况下,文库偏向一致性的二级结构含量。例如,多核苷酸频率是一致性的,且二级结构含量在约10%至约90%的范围内。在一些情况下,文库包含具有二级结构百分比为约10%至约95%的多核苷酸。在一些情况下,本文描述的文库包含具有大于30%的具有10%至30%或70%至90%的二级结构百分比的不同多核苷酸的多核苷酸。在一些情况下,本文描述的文库包含具有少于约15%的具有10%至30%或60%至90%的二级结构百分比的多核苷酸的多核苷酸。

[0114] 在一些情况下,具有指定二级结构含量的多核苷酸文库的产生通过组合至少两个具有不同二级结构含量的多核苷酸文库发生。在一些情况下,至少2、3、4、5、6、7、10或多于10个多核苷酸文库被组合以产生具有指定二级结构含量的多核苷酸群体。在一些情况下,不多于2、3、4、5、6、7或10个多核苷酸文库被组合以产生具有指定二级结构含量的非相同的多核苷酸的群体。

[0115] 在一些情况下,二级结构含量通过合成每个簇更少或更多的多核苷酸来调整。例如,在单个簇上合成至少25、50、100、200、300、400、500、600、700、800、900、1000或多于1000种非相同的多核苷酸。在一些情况下,在单个簇上合成不多于约50、100、200、300、400、500、600、700、800、900、1000种非相同的多核苷酸。在一些情况下,在单个簇上合成50至500种非相同的多核苷酸。在一些情况下,在单个簇上合成100至200种非相同的多核苷酸。在一些情况下,在单个簇上合成约100、约120、约125、约130、约150、约175或约200种非相同的多核苷酸。

[0116] 在一些情况下,通过合成不同长度的非相同的多核苷酸来调整二级结构含量。例如,每个合成的非相同的多核苷酸的长度可以为至少或大约至少10、15、20、25、30、35、40、45、50、100、150、200、300、400、500、2000个核苷酸或更多。合成的非相同的多核苷酸的长度可以为最多或大约最多2000、500、400、300、200、150、100、50、45、35、30、25、20、19、18、17、16、15、14、13、12、11、10个核苷酸或更少。合成的每种非相同的多核苷酸的长度可以落入从10-2000、10-500、9-400、11-300、12-200、13-150、14-100、15-50、16-45、17-40、18-35和19-25。

[0117] 具有受控化学计量的序列含量的多核苷酸文库的产生

[0118] 在一些情况下,多核苷酸文库以指定分布的期望的多核苷酸序列合成。在一些情况下,调整多核苷酸文库以富集特定的期望的序列导致改善的下游应用结果。

[0119] 一种或更多种特定序列可以基于它们在下游应用中的评估来选择。在一些情况下,评估是与用于扩增、富集或检测的靶序列的结合亲和力、稳定性、解链温度、生物活性、组装成更大片段的能力或多核苷酸的其他性质。在一些情况下,评估是经验性的或从先前的实验和/或计算机算法预测的。

[0120] 多核苷酸文库中选择的序列可以是序列的至少10%、20%、30%、40%、50%、60%、70%、80%、90%、95%或多于95%。在一些情况下,多核苷酸文库中选择的序列是序列的最多10%、20%、30%、40%、50%、60%、70%、80%、90%、95%或最多100%。在一些情况下,选择的序列在序列的约5-95%、10-90%、30-80%、40-75%或50-70%的范围内。

[0121] 多核苷酸文库的每个选择的序列的频率可以被调整。在一些情况下,多核苷酸文库偏向较高数数量的选择的序列。例如,文库被设计成其中选择的序列的增加了的多核苷酸频率在约40%至约90%的范围内。在一些情况下,多核苷酸文库包含较低数数量的选择的序列。例如,文库被设计成其中选择的序列的增加了的多核苷酸频率在约10%至约60%的范围内。文库可以被设计成偏向较高和较低频率的选择的序列。在一些情况下,文库偏向一致性的序列表现度。例如,多核苷酸频率相对于选择的序列频率是一致性的,在约10%至约90%的范围内。在一些情况下,文库包含具有序列的约10%至约95%的选择的序列频率的多核苷酸。

[0122] 在一些情况下,具有特定的选择的序列频率的多核苷酸文库的产生通过组合至少两个具有不同的选择的序列频率含量的多核苷酸文库发生。在一些情况下,至少2、3、4、5、6、7、10或多于10个多核苷酸文库被组合以产生具有特定的选择的序列频率的多核苷酸群体。在一些情况下,不超过2、3、4、5、6、7或10个多核苷酸文库被组合以产生具有特定的选择的序列频率的非相同的多核苷酸群体。

[0123] 在一些情况下,通过合成每个簇更少或更多的多核苷酸来调整选择的的序列频率。例如,在单个簇上合成至少25、50、100、200、300、400、500、600、700、800、900、1000或多于1000种非相同的多核苷酸。在一些情况下,在单个簇上合成不多于约50、100、200、300、400、500、600、700、800、900、1000种非相同的多核苷酸。在一些情况下,在单个簇上合成50至500种非相同的多核苷酸。在一些情况下,在单个簇上合成100至200种非相同的多核苷酸。在一些情况下,在单个簇上合成约100、约120、约125、约130、约150、约175或约200种非相同的多核苷酸。

[0124] 在一些情况下,通过合成不同长度的非相同的多核苷酸来调整选择的的序列频率。例如,每个合成的非相同的多核苷酸的长度可以是至少或大约至少10、15、20、25、30、35、40、45、50、100、150、200、300、400、500、2000个核苷酸或更多。合成的非相同的多核苷酸的长度可以是最多或大约最多2000、500、400、300、200、150、100、50、45、35、30、25、20、19、18、17、16、15、14、13、12、11、10个核苷酸或更少。合成的每个非相同的多核苷酸的长度可以落入从10-2000、10-500、9-400、11-300、12-200、13-150、14-100、15-50、16-45、17-40、18-35和19-25。

[0125] 多核苷酸探针结构

[0126] 多核苷酸探针文库可以用于富集较大样品多核苷酸群体中的特定靶序列。在一些情况下,多核苷酸探针各自包含与一个或更多个靶序列互补的靶结合序列、一个或更多个非靶结合序列以及一个或更多个引物结合位点,例如通用引物结合位点。在一些情况下互

补或至少部分互补的靶结合序列与靶序列结合(杂交)。引物结合位点,如通用引物结合位点,有助于探针文库的所有成员或成员子群体的同时扩增。在一些情况下,探针还包括条形码或索引序列。条形码是核酸序列,允许识别与条形码缔合的多核苷酸的一些特征。测序后,条形码区域提供用于识别与编码区域或样品源相关的特征的指示物。条形码可以被设计成合适的长度,以允许足够程度的识别,例如,长度为至少约3、4、5、6、7、8、9、10、11、12、13、14、15、16、17、18、19、20、21、22、23、24、25、26、27、28、29、30、31、32、33、34、35、36、37、38、39、44、45、46、47、48、49、50、51、52、53、54、55或更多个碱基。可以在同一分子上使用多于一个条形码,例如大约2、3、4、5、6、7、8、9、10或更多个条形码,它们任选地由非条形码序列分开。在一些实施方案中,多于一个条形码中的每个条形码与该多于一个条形码中的每个其他条形码在至少三个碱基位置诸如至少约3、4、5、6、7、8、9、10或更多个位置中不同。在一些情况下,多核苷酸被连接到一个或更多个分子(或亲和)标签,例如小分子、肽、抗原、金属或蛋白,以形成用于随后捕获感兴趣的靶序列的探针。在一些情况下,两个具有能够杂交的互补靶结合序列的探针形成双链探针。

[0127] 这里描述的探针可以与为基因组中的序列的靶序列互补。这里描述的探针可以与为基因组中的外显子组序列的靶序列互补。这里描述的探针可以与为基因组中的内含子序列的靶序列互补。在一些情况下,探针包括与靶序列互补的靶结合序列,以及至少一个与靶不互补的非靶结合序列。在一些情况下,探针的靶结合序列长度为约120个核苷酸,或至少10、15、20、25、50、75、100、110、120、125、140、150、160、175、200、300、400、500或多于500个核苷酸。在一些情况下,靶结合序列的长度为不多于10、15、20、25、50、75、100、125、150、175、200或不多于500个核苷酸。探针的靶结合序列在一些情况下长度为约120个核苷酸,或者长度为约10、15、20、25、40、50、60、70、80、85、87、90、95、97、100、105、110、115、117、118、119、120、121、122、123、124、125、126、127、128、129、130、135、140、145、150、155、157、158、159、160、161、162、163、164、165、166、167、168、169、170、175、180、190、200、210、220、230、240、250、300、400或约500个核苷酸。在一些情况下,靶结合序列的长度为约20至约400个核苷酸、或约30至约175个核苷酸、约40至约160个核苷酸、约50至约150个核苷酸、约75至约130个核苷酸、约90至约120个核苷酸、或约100至约140个核苷酸。探针的非靶结合序列在一些情况下长度为至少约20个核苷酸,或者长度为至少约1、5、10、15、17、20、23、25、50、75、100、110、120、125、140、150、160、175个核苷酸,或者多于约175个核苷酸。非靶结合序列的长度通常不多于约5、10、15、20、25、50、75、100、125、150、175个核苷酸,或者不多于约200个核苷酸。探针的非靶结合序列长度通常为约20个核苷酸,或者长度为约1、2、3、4、5、6、7、8、9、10、11、12、13、14、15、16、17、18、19、20、21、22、23、25、40、50、60、70、80、90、100、110、120、130、140、150或约200个核苷酸。或在一些情况下,非靶结合序列的长度为约1至约250个核苷酸,或者长度为约20至约200个核苷酸、约10至约100个核苷酸、约10至约50个核苷酸、约30至约100个核苷酸、约5至约40个核苷酸、或约15至约35个核苷酸。非靶结合序列通常包含与靶序列不互补的序列,和/或包含不用于结合引物的序列。在一些情况下,非靶结合序列包含单个核苷酸的重复,例如聚腺苷或聚胸腺嘧啶。探针通常不包含非靶结合序列或至少包含一个非靶结合序列。在一些情况下,探针包含一个或两个非靶结合序列。非靶结合序列可以与探针中的一个或更多个靶结合序列相邻。例如,非靶结合序列位于探针的5'或3'末端上。在一些情况下,非靶结合序列附接到分子标签或间隔区。

[0128] 在一些情况下,非靶结合序列可以是引物结合位点。引物结合位点的长度通常为每个至少约20个核苷酸,或至少约10、12、14、16、18、20、22、24、26、28、30、32、34、36、38个核苷酸,或至少约40个核苷酸。在一些情况下,每个引物结合位点的长度不多于约10、12、14、16、18、20、22、24、26、28、30、32、34、36、38或不多于约40个核苷酸。在一些情况下,每个引物结合位点的长度为约10至约50个核苷酸,或约15至约40个核苷酸、约20至约30个核苷酸、约10至约40个核苷酸、约10至约30个核苷酸、约30至约50个核苷酸、或约20至约60个核苷酸。在一些情况下,多核苷酸探针包含至少两个引物结合位点。在一些情况下,引物结合位点可以是通用引物结合位点,其中所有探针在这些位点包含相同的引物结合序列。在一些情况下,靶向特定序列及其反向互补序列(例如基因组DNA的区域)的一对多核苷酸探针在图3A中以300表示,包括第一靶结合序列301、第二靶结合序列302、第一非靶结合序列303和第二非靶结合序列304。例如,与特定序列(例如基因组DNA的区域)互补的一对多核苷酸探针。

[0129] 在一些情况下,第一靶结合序列301是第二靶结合序列302的反向互补序列。在一些情况下,两个靶结合序列在扩增前都是化学合成的。在可选择的布置中,靶向特定序列及其反向互补序列(例如基因组DNA的区域)的一对多核苷酸探针在图3B中以305表示,包括第一靶结合序列301、第二靶结合序列302、第一非靶结合序列303、第二非靶结合序列304、第三非靶结合序列306和第四非靶结合序列307。在一些情况下,第一靶结合序列301是第二靶结合序列302的反向互补序列。在一些情况下,一个或更多个非靶结合序列包含聚腺苷或聚胸腺嘧啶。

[0130] 在一些情况下,对中的两个探针都被用至少一个分子标签标记。在一些情况下,在扩增期间,PCR被用于将分子标签(通过包含分子标签的引物)引入到探针上。在一些情况下,分子标签包括一种或更多种生物素、叶酸、多组氨酸、FLAG标签、谷胱甘肽或符合本说明书的其他分子标签。在一些情况下,探针在5'末端被标记。在一些情况下,探针在3'末端被标记。在一些情况下,5'和3'末端两者都用分子标签标记。在一些情况下,一对中第一探针的5'末端用至少一个分子标签标记,而该对中第二探针的3'末端用至少一个分子标签标记。在一些情况下,在一个或更多个分子标签和探针的核酸之间存在间隔区。在一些情况下,间隔区可以包括烷基、多元醇或多氨基链、肽或多核苷酸。在一些情况下,用于捕获探针-靶核酸复合物的固体支持物是珠或表面。在一些情况下,固体支持物包括玻璃、塑料或能够包含将结合分子标签的捕获部分的其他材料。在一些情况下,珠是磁珠。例如,用生物素标记的探针被包含链霉亲和素的磁珠捕获。使探针与核酸文库接触,以允许探针与靶序列结合。在一些情况下,添加封闭多核苷酸以防止探针结合到一个或更多个附接于靶核酸的衔接子序列。在一些情况下,封闭多核苷酸包含一种或更多种核酸类似物。在一些情况下,封闭多核苷酸具有在一个或更多个位置取代胸腺嘧啶的尿嘧啶。

[0131] 本文描述的探针可以包含互补性靶结合序列,其结合到一个或更多个靶核酸序列。在一些情况下,靶序列是任何DNA或RNA核酸序列。在一些情况下,靶序列可以比探针插入片段长。在一些情况下,靶序列可以比探针插入片段短。在一些情况下,靶序列可以与探针插入片段长度相同。例如,靶序列的长度可以是至少或大约至少2、10、15、20、25、30、35、40、45、50、100、150、200、300、400、500、1000、2000、5,000、12,000、20,000个核苷酸或更多。靶序列的长度可以是最多或大约最多20,000、12,000、5,000、2,000、1,000、500、400、300、200、150、100、50、45、35、30、25、20、19、18、17、16、15、14、13、12、11、10、2个核苷酸或更少。

靶序列的长度可以落入从2-20,000、3-12,000、5-5,500、10-2,000、10-1,000、10-500、9-400、11-300、12-200、13-150、14-100、15-50、16-45、17-40、18-35和19-25。探针序列可以靶向与特定基因、疾病、调控途径或符合本说明书的其他生物功能相关的序列。

[0132] 在一些情况下,单个探针插入片段403与较大的多核酸400中的一个或多个靶序列互补402(图4A-4G)。示例性的靶序列是外显子。在一些情况下,一个或多个探针靶向单个靶序列(图4A-4G)。在一些情况下,单个探针可以靶向多于一个靶序列。在一些情况下,探针的靶结合序列靶向靶序列402和相邻序列401两者(图4A和4B)。在一些情况下,第一探针靶向靶序列的第一区域和第二区域,并且第二探针靶向靶序列的第二区域和第三区域(图4D和图4E)。在一些情况下,多于一个探针靶向单个靶序列,其中所述多于一个探针的靶结合序列包含在靶序列区域的互补性方面重叠的一个或多个序列(图4G)。在一些情况下,探针插入片段在与靶序列区域的互补性方面不重叠。在一些情况下,至少2、10、15、20、25、30、35、40、45、50、100、150、200、300、400、500、1000、2000、5,000、12,000、20,000或多于20,000个探针靶向单个靶序列。在一些情况下,不多于4个定向于单个靶序列的探针重叠,或者不多于3、2、1个,或者没有靶向单个靶序列的探针重叠。在一些情况下,一个或多个探针不靶向靶序列中的所有碱基,留下一个或多个空位(图4C和图4F)。在一些情况下,空位靠近靶序列405的中间(图4F)。在一些情况下,空位404位于靶序列的5'或3'末端(图4C)。在一些情况下,空位长度为6个核苷酸。在一些情况下,空位长度不多于1、2、3、4、5、6、7、8、9、10、20、30、40或不多于50个核苷酸。在一些情况下,空位长度为至少1、2、3、4、5、6、7、8、9、10、20、30、40或至少50个核苷酸。在一些情况下,空位长度落在1-50、1-40、1-30、1-20、1-10、2-30、2-20、2-10、3-50、3-25、3-10或3-8个核苷酸的长度范围内。在一些情况下,当与互补序列杂交时,靶向序列的一组探针在该组中的探针之间不包括重叠区域。在一些情况下,当与互补序列杂交时,靶向序列的一组探针在该组中的探针之间没有任何空位。探针可以被设计成使与靶序列的一致性结合最大化。在一些情况下,探针被设计成使可能干扰探针结合到靶的靶结合序列的高或低GC含量、二级结构、重复/回文序列或其他序列特征最小化。在一些情况下,单个探针可以靶向多于一个靶序列。

[0133] 本文描述的探针文库可以包括至少10、20、50、100、200、500、1,000、2,000、5,000、10,000、20,000、50,000、100,000、200,000、500,000、1,000,000或多于1,000,000种探针。探针文库可以具有不多于10、20、50、100、200、500、1,000、2,000、5,000、10,000、20,000、50,000、100,000、200,000、500,000或不多于1,000,000种探针。探针文库可以包括10至500、20至1000、50至2000、100至5000、500至10,000、1,000至5,000、10,000至50,000、100,000至500,000或50,000至1,000,000种探针。探针文库可以包括约370,000、400,000、500,000种或更多不同的探针。

[0134] 下一代测序应用

[0135] 多核苷酸文库的下游应用可以包括下一代测序。例如,用受控化学计量的多核苷酸探针文库富集靶序列导致更有效的测序。多核苷酸文库捕获靶或与靶杂交的性能可以由描述效率、准确性和精确度的许多不同度量来定义。例如,Picard度量包括变量,例如HS文库大小(文库中对应于靶区域的独特分子的数量,从读段对计算)、平均靶覆盖(达到特定覆盖水平的碱基百分比)、覆盖深度(包括特定核苷酸的读段数)、富集倍数(唯一地映射(mapping)到靶的序列读段/映射到总样品的读段,乘以总样品长度/靶长度),脱饵碱基百

分比(与探针/饵的碱基不对应的碱基百分比)、中靶的可用碱基、AT或GC失落率、80倍的碱基罚分(将80%的非零靶提高到平均覆盖水平所需的过度覆盖倍数)、零靶覆盖百分比、PF读段(通过质量过滤器的读段数)、选择的碱基的百分比(上饵碱基和近饵碱基的总和除以总比对碱基)、重复百分比或符合本说明书的其他变量。

[0136] 读段深度(测序深度或取样)代表对于一个序列获得的测序的核酸片段(“读段”)的总次数。理论读段深度定义为假设读段在整个理想化的基因组中完美分布,相同核苷酸被读取的预期次数。读段深度表示为%覆盖(或覆盖宽度)的函数。例如,完美分布的1百万个碱基的基因组的1000万次读取,理论上导致100%的序列的10X读段深度。实验上,可能需要更大数量的读取(更高的理论读段深度或过度取样)来获得一定百分比的靶序列的期望的读段深度。用受控化学计量的探针文库富集靶序列增加了下游测序的效率,因为获得具有超过期望的%靶序列的可接受读段数的实验结果需要较少的总读段。例如,在一些情况下,靶序列的55x理论读段深度导致至少90%的序列的至少30x覆盖。在一些情况下,靶序列的不多于55x理论读段深度导致至少80%的序列的至少30x读段深度。在一些情况下,靶序列的不多于55x理论读段深度导致至少95%的序列的至少30x读段深度。在一些情况下,靶序列的不多于55x理论读段深度导致至少98%的序列的至少10x读段深度。在一些情况下,靶序列的55x理论读段深度导致至少98%的序列的至少20x读段深度。在一些情况下,靶序列的不多于55x理论读段深度导致至少98%的序列的至少5x读段深度。在与靶杂交期间增加探针浓度可以导致读段深度的增加。在一些情况下,探针的浓度增加至少1.5x、2.0x、2.5x、3x、3.5x、4x、5x或多于5x。在一些情况下,增加探针浓度导致读段深度增加至少1000%,或增加20%、30%、40%、50%、60%、70%、80%、90%、100%、200%、300%、500%、750%、1000%或多于1000%。在一些情况下,将探针浓度提高3x导致读段深度增加1000%。

[0137] 中靶率代表与期望的靶序列相对应的测序读段的百分比。在一些情况下,受控化学计量的多核苷酸探针文库导致至少30%或至少35%、40%、45%、50%、55%、60%、65%、70%、75%、80%、85%或至少90%的中靶率。在与靶核酸接触期间增加多核苷酸探针的浓度导致中靶率的增加。在一些情况下,探针的浓度增加至少1.5x、2.0x、2.5x、3x、3.5x、4x、5x或多于5x。在一些情况下,探针浓度的增加导致中靶结合增加至少20%,或增加10%、20%、30%、40%、50%、60%、70%、80%、90%、100%、200%、300%或至少500%。在一些情况下,将探针浓度增加3x导致中靶率增加20%。

[0138] 在一些情况下,覆盖一致性被计算为作为靶序列同一性的函数的读段深度。较高的覆盖一致性导致获得期望的读段深度所需的较低的测序读段数。例如,靶序列的性质可能影响读段深度,例如高或低GC或AT含量、重复序列、尾随腺嘌呤、二级结构、对靶序列结合(用于扩增、富集或检测)的亲和力、稳定性、解链温度、生物活性、组装成更大片段的能力、包含修饰核苷酸或核苷酸类似物的序列、或多核苷酸的任何其他性质。用受控化学计量的多核苷酸探针文库富集靶序列导致测序后更高的覆盖一致性。在一些情况下,95%的序列具有在平均文库读段深度的1x内,或在平均文库读段深度的约0.05、0.1、0.2、0.5、0.7、1、1.2、1.5、1.7或约2x内的读段深度。在一些情况下,80%、85%、90%、95%、97%或99%的序列具有在平均值的1x内的读段深度。

[0139] 用多核苷酸探针文库富集靶核酸

[0140] 本文描述的探针文库可以用于富集样品多核苷酸群体中存在的靶多核苷酸,用于

多种下游应用。在一些情况下,从一个或多个来源获得样品,并且使用本领域已知的常规技术分离样品多核苷酸群体。样品是从生物来源如唾液、血液、组织、皮肤或完全合成的来源获得的(通过非限制性实例的方式)。从样品中获得的多于一种多核苷酸被片段化、末端修复和腺苷化以形成双链样品核酸片段。在一些情况下,末端修复是通过在适当的缓冲液中用一种或更多种酶如T4 DNA聚合酶、克列诺酶(klenow enzyme)和T4多核苷酸激酶进行处理来完成的。在一些情况下,用不具有3'至5'外切酶活性的克列诺片段和dATP添加了有助于连接到衔接子的核苷酸突出端。

[0141] 可以将衔接子用连接酶如T4连接酶连接到样品多核苷酸片段的两端,以产生衔接子标记的多核苷酸链文库,并将衔接子标记的多核苷酸文库用引物如通用引物扩增。在一些情况下,衔接子是Y形衔接子,包括一个或多个引物结合位点、一个或多个接枝区和一个或多个索引区(index region)。在一些情况下,一个或多个索引区存在于衔接子的每一链上。在一些情况下,接枝区与流通池表面互补,并有助于样品文库的下一代测序。在一些情况下,Y形衔接子部分地包含互补序列。在一些情况下,Y形衔接子包含单个胸苷突出端,该突出端与双链的衔接子标记的多核苷酸链的突出的腺嘌呤杂交。Y形衔接子可以包含抗裂解的修饰核酸。例如,硫代磷酸酯骨架被用于将突出的胸苷附接到衔接子的3'末端。双链样品核酸片段的文库随后在衔接子封闭剂的存在下变性。衔接子封闭剂使探针与衔接子标记的多核苷酸链上存在的衔接子序列(而不是靶序列)的脱靶杂交最小化。在一些情况下,变性在96℃或在约85℃、87℃、90℃、92℃、95℃、97℃、98℃或约99℃进行。在一些情况下,多核苷酸靶向文库(探针文库)在杂交溶液中,在96℃、在约85℃、87℃、90℃、92℃、95℃、97℃、98℃或℃变性。将变性的衔接子标记的多核苷酸文库和杂交溶液在合适的温度孵育合适的量的时间,以允许探针与其互补的靶序列杂交。在一些情况下,合适的杂交温度为约45℃至80℃,或至少45℃、50℃、55℃、60℃、65℃、70℃、75℃、80℃、85℃或90℃。在一些情况下,杂交温度是70℃。在一些情况下,合适的杂交时间是16小时,或至少4、6、8、10、12、14、16、18、20、22小时,或多于22小时,或约12至20小时。然后将结合缓冲液添加到杂交的衔接子标记的多核苷酸探针中,并且使用包含捕获部分的固体支持物来选择性地结合杂交的衔接子标记的多核苷酸探针。在添加洗脱缓冲液以从固体支持物释放富集的标记多核苷酸片段之前,用缓冲液洗涤固体支持物以去除未结合的多核苷酸。在一些情况下,固体支持物被洗涤2次,或1、2、3、4、5或6次。扩增富集的衔接子标记的多核苷酸片段文库,并对富集的文库测序。

[0142] 可以从样品中获得多于一种核酸(即基因组序列),并进行片段化、任选的末端修复和腺苷酸化。将衔接子连接到多核苷酸片段的两端以产生衔接子标记的多核苷酸链文库,并将衔接子标记的多核苷酸文库扩增。然后使衔接子标记的多核苷酸文库在衔接子封闭剂的存在下在高温优选地96℃变性。使多核苷酸靶向文库(探针文库)在杂交溶液中在高温优选地约90℃至99℃变性,并在杂交溶液中在约45℃至80℃与变性的标记的多核苷酸文库结合约10至24小时。然后将结合缓冲液添加到杂交的标记的多核苷酸探针中,并且使用包含捕获部分的固体支持物来选择性地结合杂交的衔接子标记的多核苷酸探针。在添加洗脱缓冲液以从固体支持物释放富集的衔接子标记的多核苷酸片段之前,用缓冲液洗涤固体支持物一次或更多次,优选地约2次和5次,以去除未结合的多核苷酸。扩增富集的衔接子标记的多核苷酸片段文库,并且然后对文库测序。该方法中还采用了可选择的实验变量,诸如

孵育时间、温度、反应体积/浓度、洗涤次数或符合本说明书的其他变量。

[0143] 多核苷酸群体可以在衔接子连接之前被富集。在一个实例中,从样品中获得多于一种多核苷酸,片段化,任选地末端修复,并在高温,优选90-99℃变性。使多核苷酸靶向文库(探针文库)在杂交溶液在高温优选约90℃至99℃变性,并在杂交溶液中在约45℃至80℃与变性的标记的多核苷酸文库结合约10至24小时。然后将结合缓冲液添加到杂交的标记的多核苷酸探针中,并且使用包含捕获部分的固体支持物来选择性地结合杂交的衔接子标记的多核苷酸探针。在添加洗脱缓冲液以从固体支持物释放富集的衔接子标记的多核苷酸片段之前,用缓冲液洗涤固体支持物一次或更多次,优选地约2次和5次,以去除未结合的多核苷酸。然后将富集的多核苷酸片段聚腺苷酸化,将衔接子连接到多核苷酸片段的两端以产生衔接子标记的多核苷酸链文库,并且将衔接子标记的多核苷酸文库扩增。然后对衔接子标记的多核苷酸文库测序。

[0144] 多核苷酸靶向文库也可以用于通过与不期望的片段杂交,从多于一种多核苷酸中过滤不期望的序列。例如,从样品中获得多于一种多核苷酸,并片段化、任选地末端修复和腺苷酸化。将衔接子连接到多核苷酸片段的两端以产生衔接子标记的多核苷酸链文库,并将衔接子标记的多核苷酸文库扩增。可选择地,腺苷酸化和衔接子连接步骤改为在样品多核苷酸的富集后进行。然后使衔接子标记的多核苷酸文库在衔接子封闭剂的存在下在高温优选地90-99℃变性。使设计用于去除不期望的非靶序列的多核苷酸过滤文库(探针文库)在杂交溶液在高温优选地约90℃至99℃变性,并在杂交溶液中在约45℃至80℃与变性的标记的多核苷酸文库结合约10至24小时。然后将结合缓冲液添加到杂交的标记的多核苷酸探针中,并且使用包含捕获部分的固体支持物来选择性地结合杂交的衔接子标记的多核苷酸探针。用缓冲液洗涤固体支持物一次或更多次,优选地约1次和5次,以洗脱未结合的衔接子标记的多核苷酸片段。将富集的未结合的衔接子标记的多核苷酸片段文库扩增,并且然后对扩增的文库测序。

[0145] 高度并行的从头核酸合成

[0146] 本文描述了一种平台方法,该方法利用小型化、并行化和垂直整合的从多核苷酸合成到硅上纳米孔内基因组装的端到端过程来创建革命性的合成平台。本文描述的装置提供了具有与96孔板相同的占位面积(footpring)的硅合成平台,与传统合成方法相比,硅合成平台能够将产量提高100至1,000倍,且在单次高度并行化的运行中产生高达约1,000,000个多核苷酸。在一些情况下,本文描述的单个硅板提供了约6,100个非相同的多核苷酸的合成。在一些情况下,每种非相同的多核苷酸位于一个簇内。一个簇可以包含50至500个非相同的多核苷酸。

[0147] 本文描述的方法提供了各自编码至少一个预定参考核酸序列的预定变体的多核苷酸的文库的合成。在一些情况下,预定参考序列是编码蛋白的核酸序列,并且变体文库包括编码至少单个密码子的变体的序列,使得由合成的核酸编码的随后的蛋白中的单个残基的多于一个不同变体通过标准翻译过程产生。核酸序列中合成的特定改变可以通过将核苷酸变化掺入到重叠或钝端多核苷酸引物中来引入。可选择地,多核苷酸群体可以共同编码长核酸(例如,基因)及其变体。在这种布置中,多核苷酸群体可以杂交并经受标准分子生物学技术以形成长核酸(例如基因)及其变体。当长核酸(例如基因)及其变体在细胞中表达时,产生变体蛋白文库。相似地,这里提供了用于合成编码RNA序列(例如,miRNA、shRNA和

mRNA)或DNA序列(例如,增强子、启动子、UTR和终止子区域)的变体文库的方法。这里还提供了从使用这里描述的方法合成的文库中选择出的变体的下游应用。下游应用包括鉴定具有增强的生物学相关功能,例如生物化学亲和力、酶活性、细胞活性变化,以及用于疾病状态的治疗或预防的变体核酸或蛋白序列。

[0148] 基板

[0149] 本文提供了包含多于一个簇的基板,其中每个簇包含多于一个支持多核苷酸附接和合成的位点。如本文使用的术语“位点”指结构上的离散区域,其为编码从表面延伸的单个预定序列的多核苷酸提供支持。在一些情况下,位点在二维表面,例如基本平坦的表面。在一些情况下,位点指表面上的离散升高或降低的位点,例如孔、微孔、通道或柱。在一些情况下,位点的表面包含活性官能化的材料,以附接到至少一个用于多核苷酸合成的核苷酸,或者优选地附接到用于多核苷酸群体合成的相同核苷酸的群体。在一些情况下,多核苷酸指编码相同核酸序列的多核苷酸群体。在一些情况下,装置的表面包括基板的一个或多于一个表面。

[0150] 本文提供了可以包括表面的结构,所述表面支持在共同支持物上的可寻址位置合成具有不同预定序列的多于一种多核苷酸。在一些情况下,一个装置为多于2000、5,000、10,000、20,000、30,000、50,000、75,000、100,000、200,000、300,000、400,000、500,000、600,000、700,000、800,000、900,000、1,000,000、1,200,000、1,400,000、1,600,000、1,800,000、2,000,000、2,500,000、3,000,000、3,500,000、4,000,000、4,500,000、5,000,000、10,000,000或更多非相同的多核苷酸的合成提供支持。在一些情况下,装置为多于2000、5,000、10,000、20,000、30,000、50,000、75,000、100,000、200,000、300,000、400,000、500,000、600,000、700,000、800,000、900,000、1,000,000、1,200,000、1,400,000、1,600,000、1,800,000、2,000,000、2,500,000、3,000,000、3,500,000、4,000,000、4,500,000、5,000,000、10,000,000或更多编码不同序列的多核苷酸的合成提供支持。在一些情况下,至少一部分多核苷酸具有相同的序列,或者被配置为用相同的序列合成。

[0151] 本文提供了用于制备和生长长度为约5、10、20、30、40、50、60、70、80、90、100、125、150、175、200、225、250、275、300、325、350、375、400、425、450、475、500、600、700、800、900、1000、1100、1200、1300、1400、1500、1600、1700、1800、1900或2000个碱基的多核苷酸的方法。在一些情况下,形成的多核苷酸的长度为约5、10、20、30、40、50、60、70、80、90、100、125、150、175、200或225个碱基。多核苷酸的长度可以为至少5、10、20、30、40、50、60、70、80、90或100个碱基。多核苷酸可以是长度为10至225个碱基,长度为12至100个碱基,长度为20至150个碱基,长度为20至130个碱基,或长度为30至100个碱基。

[0152] 在一些情况下,多核苷酸在基板的位点上合成,其中每个位点支持多核苷酸群体的合成。在一些情况下,每个位点支持合成具有与在另一位点上生长的多核苷酸群体不同的序列的多核苷酸群体。在一些情况下,装置的位点位于多于一个簇内。在一些情况下,装置包括至少10,500,1000,2000,3000,4000,5000,6000,7000,8000,9000,10000,11000,12000,13000,14000,15000,20000,30000,40000,5000或更多簇。在一些情况下,装置包含多于2000、5,000、10,000、100,000、200,000、300,000、400,000、500,000、600,000、700,000、800,000、900,000、1,000,000、1,100,000、1,200,000、1,300,000、1,400,000、1,500,000、1,600,000、1,700,000、1,800,000、1,900,000、2,000,000、300,000、400,000、

500,000、600,000、700,000、800,000、900,000、1,000,000、1,200,000、1,400,000、1,600,000、1,800,000、2,000,000、2,500,000、3,000,000、3,500,000、4,000,000、4,500,000、5,000,000或者10,000,000个或更多不同的位点。在一些情况下,装置包括大约10,000个不同的位点。在不同的情况下,单个簇中的位点的量是不同的。在一些情况下,每个簇包括1、2、3、4、5、6、7、8、9、10、20、30、40、50、60、70、80、90、100、120、130、150、200、300、400、500、1000或更多个位点。在一些情况下,每个簇包括大约50-500个位点。在一些情况下,每个簇包括约100-200个位点。在一些情况下,每个簇包括大约100-150个位点。在一些情况下,每个簇包括大约109、121、130或137个位点。在一些情况下,每个簇包括大约19个、20个、61个、64个或更多个位点。

[0153] 在装置上合成的不同多核苷酸的数量可以取决于基板中可用的不同位点的数量。在一些情况下,装置的簇内的位点密度为至少或约每 mm^2 1个位点、每 mm^2 10个位点、每 mm^2 25个位点、每 mm^2 50个位点、每 mm^2 65个位点、每 mm^2 75个位点、每 mm^2 100个位点、每 mm^2 130个位点、每 mm^2 150个位点、每 mm^2 175个位点、每 mm^2 200个位点、每 mm^2 300个位点、每 mm^2 400个位点、每 mm^2 500个位点、每 mm^2 1,000个位点或更多。在一些情况下,装置包括约10个位点每 mm^2 至约500个每 mm^2 ,约25个位点每 mm^2 至约400个每 mm^2 ,约50个位点每 mm^2 至约500个每 mm^2 ,约100个位点每 mm^2 至约500个每 mm^2 ,约150个位点每 mm^2 至约500个每 mm^2 ,约10个位点每 mm^2 至约250个每 mm^2 ,约50个位点每 mm^2 至约250个每 mm^2 ,约10个位点每 mm^2 至约200每 mm^2 ,或约50个位点每 mm^2 至约200个每 mm^2 。在一些情况下,簇内两个相邻位点的中心之间的距离为约 $10\mu\text{m}$ 至约 $500\mu\text{m}$ 、约 $10\mu\text{m}$ 至约 $200\mu\text{m}$ 、或约 $10\mu\text{m}$ 至约 $100\mu\text{m}$ 。在一些情况下,两个相邻位点的中心之间的距离大于约 $10\mu\text{m}$ 、 $20\mu\text{m}$ 、 $30\mu\text{m}$ 、 $40\mu\text{m}$ 、 $50\mu\text{m}$ 、 $60\mu\text{m}$ 、 $70\mu\text{m}$ 、 $80\mu\text{m}$ 、 $90\mu\text{m}$ 或 $100\mu\text{m}$ 。在一些情况下,两个相邻位点的中心之间的距离小于约 $200\mu\text{m}$ 、 $150\mu\text{m}$ 、 $100\mu\text{m}$ 、 $80\mu\text{m}$ 、 $70\mu\text{m}$ 、 $60\mu\text{m}$ 、 $50\mu\text{m}$ 、 $40\mu\text{m}$ 、 $30\mu\text{m}$ 、 $20\mu\text{m}$ 或 $10\mu\text{m}$ 。在一些情况下,每个位点具有约 $0.5\mu\text{m}$ 、 $1\mu\text{m}$ 、 $2\mu\text{m}$ 、 $3\mu\text{m}$ 、 $4\mu\text{m}$ 、 $5\mu\text{m}$ 、 $6\mu\text{m}$ 、 $7\mu\text{m}$ 、 $8\mu\text{m}$ 、 $9\mu\text{m}$ 、 $10\mu\text{m}$ 、 $20\mu\text{m}$ 、 $30\mu\text{m}$ 、 $40\mu\text{m}$ 、 $50\mu\text{m}$ 、 $60\mu\text{m}$ 、 $70\mu\text{m}$ 、 $80\mu\text{m}$ 、 $90\mu\text{m}$ 或 $100\mu\text{m}$ 的宽度。在一些情况下,每个位点具有约 $0.5\mu\text{m}$ 至 $100\mu\text{m}$ 、约 $0.5\mu\text{m}$ 至 $50\mu\text{m}$ 、约 $10\mu\text{m}$ 至 $75\mu\text{m}$ 或约 $0.5\mu\text{m}$ 至 $50\mu\text{m}$ 的宽度。

[0154] 在一些情况下,装置内的簇密度为至少或约每 100mm^2 1簇、每 10mm^2 1簇、每 5mm^2 1簇、每 4mm^2 1簇、每 3mm^2 1簇、每 2mm^2 1簇、每 1mm^2 1簇、每 1mm^2 2簇、每 1mm^2 3簇、每 1mm^2 4簇、每 1mm^2 5簇、每 1mm^2 10簇、每 1mm^2 50簇或更多。在一些情况下,装置包括从约每 10mm^2 1簇至约每 1mm^2 10簇。在一些情况下,两个相邻簇的中心之间的距离小于约 $50\mu\text{m}$ 、 $100\mu\text{m}$ 、 $200\mu\text{m}$ 、 $500\mu\text{m}$ 、 $1000\mu\text{m}$ 或 $2000\mu\text{m}$ 或 $5000\mu\text{m}$ 。在一些情况下,两个相邻簇的中心之间的距离为从约 $50\mu\text{m}$ 和约 $100\mu\text{m}$ 、从约 $50\mu\text{m}$ 和约 $200\mu\text{m}$ 、从约 $50\mu\text{m}$ 和约 $300\mu\text{m}$ 、从约 $50\mu\text{m}$ 和约 $500\mu\text{m}$ 以及约 $100\mu\text{m}$ 至约 $2000\mu\text{m}$ 。在一些情况下,两个相邻簇的中心之间的距离是从约 0.05mm 到约 50mm 、从约 0.05mm 到约 10mm 、从约 0.05mm 和约 5mm 、从约 0.05mm 和约 4mm 、从约 0.05mm 和约 3mm 、从约 0.05mm 和约 2mm 、从约 0.1mm 和约 10mm 、从约 0.2mm 和约 10mm 、从约 0.3mm 和约 10mm 、从约 0.4mm 和约 10mm 、从约 0.5mm 和约 10mm 、从约 0.5mm 和约 5mm 、或从约 0.5mm 和约 2mm 。在一些情况下,每个簇具有约 0.5mm 至 2mm 、约 0.5mm 至 1mm 或约 1mm 至 2mm 的沿一个维度的直径或宽度。在一些情况下,每个簇具有约 0.5 、 0.6 、 0.7 、 0.8 、 0.9 、 1 、 1.1 、 1.2 、 1.3 、 1.4 、 1.5 、 1.6 、 1.7 、 1.8 、 1.9 或 2mm 的沿一个维度的直径或宽度。在一些情况下,每个簇具有约 0.5 、 0.6 、 0.7 、 0.8 、 0.9 、 1 、 1.1 、 1.15 、 1.2 、 1.3 、 1.4 、 1.5 、 1.6 、 1.7 、 1.8 、 1.9 或 2 的沿一个维度的直径或宽度。

[0155] 装置可以是约标准96孔板的大小,例如从约100mm和200mm乘以从约50mm和150mm。在一些情况下,装置具有小于或等于约1000mm、500mm、450mm、400mm、300mm、250mm、200mm、150mm、100mm或50mm的直径。在一些情况下,装置的直径为从约25mm和1000mm,从约25mm和约800mm,从约25mm和约600mm,从约25mm和约500mm,从约25mm和约400mm,从约25mm和约300mm,或从约25mm和约200mm。装置大小的非限制性实例包括约300mm、200mm、150mm、130mm、100mm、76mm、51mm和25mm。在一些情况下,装置具有至少约100mm²、200mm²、500mm²、1,000mm²、2,000mm²、5,000mm²、10,000mm²、12,000mm²、15,000mm²、20,000mm²、30,000mm²、40,000mm²、50,000mm²或更大的平坦表面面积。在一些情况下,装置的厚度为从约50mm和约2000mm、从约50mm和约1000mm、从约100mm和约1000mm、从约200mm和约1000mm、或从约250mm和约1000mm。装置厚度的非限制性实例包括275mm、375mm、525mm、625mm、675mm、725mm、775mm和925mm。在一些情况下,装置的厚度随直径变化并取决于基板的构成而变化。例如,包括除硅以外的材料的装置与相同直径的硅装置具有不同的厚度。装置厚度可以由所用材料的机械强度确定,并且装置必须足够厚以支撑其自身重量,且在操作期间不会破裂。在一些情况下,结构包括本文描述的多于一个装置。

[0156] 表面材料

[0157] 本文提供了一种包含表面的装置,其中所述表面被修饰以支持在预定位置的多核苷酸合成,并具有由此产生的低错误率、低失落率、高产率和高寡核苷酸表现度。在一些实施方案中,本文提供的用于多核苷酸合成的装置的表面由能够修饰以支持从头多核苷酸合成反应的多种材料制成。在一些情况下,装置足够导电,例如能够在装置的全部或一部分上形成一致性的电场。本文描述的装置可以包括柔性材料。示例性柔性材料包括但不限于改性尼龙、未改性尼龙、硝酸纤维素和聚丙烯。本文描述的装置可以包括刚性材料。示例性刚性材料包括但不限于玻璃、熔融二氧化硅、硅、二氧化硅、氮化硅、塑料(例如聚四氟乙烯、聚丙烯、聚苯乙烯、聚碳酸酯及其混合物)和金属(例如金、铂)。本文公开的装置可以由包括硅、聚苯乙烯、琼脂糖、葡聚糖、纤维素聚合物、聚丙烯酰胺、聚二甲基硅氧烷(PDMS)、玻璃或其任何组合的材料制成。在一些情况下,本文公开的装置用本文列出的材料或本领域已知的任何其他合适的材料的组合制成。

[0158] 本文描述的示例性材料的拉伸强度列表如下提供:尼龙(70MPa)、硝酸纤维素(1.5MPa)、聚丙烯(40MPa)、硅(268MPa)、聚苯乙烯(40MPa)、琼脂糖(1-10MPa)、聚丙烯酰胺(1-10MPa)、聚二甲基硅氧烷(PDMS)(3.9-10.8MPa)。本文描述的固体支持物可以具有1至300、1至40、1至10、1至5或3至11MPa的拉伸强度。本文描述的固体支持物可具有约1、1.5、2、3、4、5、6、7、8、9、10、11、20、25、40、50、60、70、80、90、100、150、200、250、270或更多MPa的拉伸强度。在一些情况下,本文描述的装置包括用于多核苷酸合成的固体支持物,所述固体支持物呈能够以连续环或卷轴储存的柔性材料,诸如胶带或柔性片的形式。

[0159] 杨氏模量测量材料在负载下对弹性(可恢复)变形的抵抗力。本文描述的示例性材料的硬度的杨氏模量列表如下提供:尼龙(3GPa)、硝酸纤维素(1.5GPa)、聚丙烯(2GPa)、硅(150GPa)、聚苯乙烯(3GPa)、琼脂糖(1-10GPa)、聚丙烯酰胺(1-10GPa)、聚二甲基硅氧烷(PDMS)(1-10GPa)。本文描述的固体支持物可以具有1至500、1至40、1至10、1至5或3至11GPa的杨氏模量。本文描述的固体支持物可以具有约1、1.5、2、3、4、5、6、7、8、9、10、11、20、25、40、50、60、70、80、90、100、150、200、250、400、500GPa或更多的杨氏模量。由于柔性和硬度之

间的关系是彼此相反的,所以柔性材料具有低杨氏模量,并且在负载下显著改变其形状。

[0160] 在一些情况下,本文公开的装置包括二氧化硅基底和氧化硅表面层。可选择地,装置可以具有氧化硅基底。这里提供的装置的表面可以有纹理,导致用于多核苷酸合成的总表面积增加。本文公开的装置可以包括至少5%、10%、25%、50%、80%、90%、95%或99%的硅。本文公开的装置可以由绝缘体上硅(SOI)晶片制成。

[0161] 表面结构(Architecture)

[0162] 本文提供了包括升高和/或降低特征的装置。具有这些特征的一个益处是增加了支持多核苷酸合成的表面积。在一些情况下,具有升高和/或降低特征的装置被称为三维基板。在一些情况下,三维装置包括一个或更多个通道。在一些情况下,一个或更多个位点包括通道。在一些情况下,通道通过沉积装置如多核苷酸合成仪进行试剂沉积。在一些情况下,试剂和/或流体收集在流体连通一个或更多个通道的较大的孔中。例如,装置包括对应于簇内多于一个位点的多于一个通道,并且多于一个通道与簇的一个孔流体连通。在一些方法中,多核苷酸文库在簇的多于一个位点中合成。

[0163] 在一些情况下,结构被配置为允许表面上用于多核苷酸合成的受控流动和传质途径。在一些情况下,装置的配置允许在多核苷酸合成期间传质途径、化学暴露时间和/或洗涤效率的受控且均等的分布。在一些情况下,装置的配置允许增加的波及系数,例如通过为生长的多核苷酸提供足够的体积,使得生长的多核苷酸的排除体积不占用可用或适用于生长多核苷酸的初始可用体积的50%、45%、40%、35%、30%、25%、20%、15%、14%、13%、12%、11%、10%、9%、8%、7%、6%、5%、4%、3%、2%、1%或更少。在一些情况下,三维结构允许流体的受控流动,以允许化学暴露的快速交换。

[0164] 本文提供了合成1fM、5fM、10fM、25fM、50fM、75fM、100fM、200fM、300fM、400fM、500fM、600fM、700fM、800fM、900fM、1pM、5pM、10pM、25pM、50pM、75pM、100pM、200pM、300pM、400pM、500pM、600pM、700pM、800pM、900pM或更多的量的DNA的方法。在一些情况下,多核苷酸文库可以跨越基因的大约1%、2%、3%、4%、5%、10%、15%、20%、30%、40%、50%、60%、70%、80%、90%、95%或100%的长度。基因可以变异高达约1%、2%、3%、4%、5%、10%、15%、20%、30%、40%、50%、60%、70%、80%、85%、90%、95%或100%。

[0165] 非相同的多核苷酸可以共同编码基因的至少1%、2%、3%、4%、5%、10%、15%、20%、30%、40%、50%、60%、70%、80%、85%、90%、95%或100%的序列。在一些情况下,多核苷酸可以编码基因的50%、60%、70%、80%、85%、90%、95%或更多的序列。在一些情况下,多核苷酸可以编码基因的80%、85%、90%、95%或更多的序列。

[0166] 在一些情况下,隔离是通过物理结构实现的。在一些情况下,隔离是通过产生用于多核苷酸合成的活性(active)和惰性(passive)区域的表面的差异官能化来实现的。差异官能化也可以通过改变遍及装置表面的疏水性来实现,从而产生引起沉积的试剂成珠或润湿的水接触角效应。采用较大的结构可以减少不同多核苷酸合成位置与相邻位点的试剂的飞溅和交叉污染。在一些情况下,诸如多核苷酸合成仪的装置被用于将试剂沉积到不同的多核苷酸合成位置。具有三维特征的基板以允许合成具有低错误率(例如,小于约1:500、1:1000、1:1500、1:2,000;1:3,000;1:5,000;或者1:10,000)的大数量的多核苷酸(例如,多于约10,000)的方式来配置。在一些情况下,装置包括具有约或多于约1、5、10、20、30、40、50、60、70、80、100、110、120、130、140、150、160、170、180、190、200、300、400或500特征每 mm^2 的

密度的特征。

[0167] 装置的孔可以具有与基板的另一个孔相同或不同的宽度、高度和/或体积。装置的通道可以具有与基板的另一个通道相同或不同的宽度、高度和/或体积。在一些情况下,簇的宽度为约0.05mm至约50mm、约0.05mm至约10mm、从约0.05mm和约5mm、从约0.05mm和约4mm、从约0.05mm和约3mm、从约0.05mm和约2mm、从约0.05mm和约1mm、从约0.05mm和约0.5mm、从约0.05mm和约0.1mm、从约0.1mm和10mm、从约0.2mm和10mm、从约0.3mm和约10mm、从约0.4mm和约10mm、从约0.5mm和10mm、从约0.5mm和约5mm、或从约0.5mm和约2mm。在一些情况下,包括簇的孔的宽度为从约0.05mm至约50mm、从约0.05mm至约10mm、从约0.05mm和约5mm、从约0.05mm和约4mm、从约0.05mm和约3mm、从约0.05mm和约2mm、从约0.05mm和约1mm、从约0.05mm和约0.5mm、从约0.05mm和约0.1mm、从约0.1mm和10mm、从约0.2mm和10mm、从约0.3mm和约10mm、从约0.4mm和约10mm、从约0.5mm和约10mm、从约0.5mm和约5mm、或从约0.5mm和约2mm。在一些情况下,簇的宽度为小于或约5mm、4mm、3mm、2mm、1mm、0.5mm、0.1mm、0.09mm、0.08mm、0.07mm、0.06mm或0.05mm。在一些情况下,簇的宽度为从约1.0mm和1.3mm。在一些情况下,簇的宽度为约1.150mm。在一些情况下,孔的宽度为小于或约5mm、4mm、3mm、2mm、1mm、0.5mm、0.1mm、0.09mm、0.08mm、0.07mm、0.06mm或0.05mm。在一些情况下,孔的宽度为从约1.0mm和1.3mm。在一些情况下,孔的宽度为约1.150mm。在一些情况下,簇的宽度为约0.08mm。在一些情况下,孔的宽度为约0.08mm。簇的宽度可以指二维或三维基板内的簇。

[0168] 在一些情况下,孔的高度为约20 μ m至约1000 μ m、约50 μ m至约1000 μ m、约100 μ m至约1000 μ m、约200 μ m至约1000 μ m、约300 μ m至约1000 μ m、约400 μ m至约1000 μ m或约500 μ m至约1000 μ m。在一些情况下,孔的高度为小于约1000 μ m、小于约900 μ m、小于约800 μ m、小于约700 μ m或小于约600 μ m。

[0169] 在一些情况下,装置包括对应于簇内多于一个位点的多于一个通道,其中通道的高度或深度为约5 μ m至约500 μ m、约5 μ m至约400 μ m、约5 μ m至约300 μ m、约5 μ m至约200 μ m、约5 μ m至约100 μ m、约5 μ m至约50 μ m或约10 μ m至约50 μ m。在一些情况下,通道的高度为小于100 μ m、小于80 μ m、小于60 μ m、小于40 μ m或小于20 μ m。

[0170] 在一些情况下,通道、位点(例如,在基本平坦的基板中)或通道和位点两者(例如,在其中位点对应于通道的三维装置中)的直径为从约1 μ m至约1000 μ m、约1 μ m至约500 μ m、约1 μ m至约200 μ m、约1 μ m至约100 μ m、约5 μ m至约100 μ m,或者从约10 μ m至约100 μ m,例如,约90 μ m、80 μ m、70 μ m、60 μ m、50 μ m、40 μ m、30 μ m、20 μ m或10 μ m。在一些情况下,通道、位点或通道和位点两者的直径为小于约100 μ m、90 μ m、80 μ m、70 μ m、60 μ m、50 μ m、40 μ m、30 μ m、20 μ m或10 μ m。在一些情况下,两个相邻通道、位点或通道和位点的中心之间的距离为约1 μ m至约500 μ m、约1 μ m至约200 μ m、约1 μ m至约100 μ m、约5 μ m至约200 μ m、约5 μ m至约100 μ m、约5 μ m至约50 μ m、或约5 μ m至约30 μ m,例如约20 μ m。

[0171] 表面修饰

[0172] 在多种情况下,表面修饰被用于表面的化学和/或物理改变,通过添加或减过程来改变装置表面或装置表面的选择的位置或区域的一个或更多个化学和/或物理性质。例如,表面修饰包括但不限于:(1)改变表面的润湿性质,(2)使表面官能化,即提供、修饰或取代表面官能团,(3)使表面去官能化,即去除表面官能团,(4)以其他方式改变表面的化学组成,例如通过蚀刻,(5)增加或减少表面粗糙度,(6)在表面上提供涂层,例如表现出与表面

润湿性质不同的润湿性质的涂层,和/或(7)在表面上沉积颗粒。

[0173] 在一些情况下,在表面顶上添加化学层(称为粘附促进剂)有助于基板表面上位点的结构图案化。用于应用粘附促进的示例性表面包括但不限于玻璃、硅、二氧化硅和氮化硅。在一些情况下,粘附促进剂是具有高表面能的化学物质。在一些情况下,在基板的表面上沉积第二化学层。在一些情况下,第二化学层具有低表面能。在一些情况下,涂覆在表面上的化学层的表面能支持液滴在表面上的定位。取决于所选择的图案化布置,位点的接近度和/或位点处的流体接触面积是可变的。

[0174] 在一些情况下,核酸或其它部分沉积在其上(例如用于多核苷酸合成)的装置表面或解析位点(resolved loci)是光滑的或基本平坦的(例如二维的),或者具有不规则性,诸如升高或降低的特征(例如三维特征)。在一些情况下,装置表面被用一层或更多层不同的化合物层修饰。这种感兴趣的修饰层包括但不限于无机和有机层,例如金属、金属氧化物、聚合物、小有机分子等。非限制性聚合物层包括肽、蛋白、核酸或其模拟物(例如肽核酸等)、多糖、磷脂、聚氨酯、聚酯、聚碳酸酯、聚脲、聚酰胺、聚乙烯胺、聚亚芳基硫醚、聚硅氧烷、聚酰亚胺、聚乙酸酯以及本文描述的或另外本领域已知的任何其他合适的化合物。在一些情况下,聚合物是杂聚物。在一些情况下,聚合物是均聚物。在一些情况下,聚合物包含官能部分或者是缀合的。

[0175] 在一些情况下,装置的解析位点被用一个或更多个增加和/或减少表面能的部分官能化。在一些情况下,部分是化学上惰性的。在一些情况下,部分被配置为支持期望的化学反应,例如多核苷酸合成反应中的一个或更多个过程。表面的表面能或疏水性是决定核苷酸附接于表面上的亲和力的一个因素。在一些情况下,用于使装置官能化的方法可以包括:(a)提供具有包含二氧化硅的表面的装置;和(b)使用本文描述的或本领域已知的合适的硅烷化剂,例如有机官能烷氧基硅烷分子,使表面硅烷化。

[0176] 在一些情况下,有机官能烷氧基硅烷分子包括二甲基氯-十八烷基-硅烷、甲基二氯-十八烷基-硅烷、三氯-十八烷基-硅烷、三甲基-十八烷基-硅烷、三乙基-十八烷基-硅烷或其任何组合。在一些情况下,装置表面包括官能化的聚乙烯/聚丙烯(通过 γ 辐射或铬酸氧化官能化,并还原成羟烷基表面)、高度交联的聚苯乙烯-二乙烯基苯(通过氯甲基化衍生,并胺化成苄胺官能表面)、尼龙(末端氨基己基基团是直接反应性的),或用还原的聚四氟乙烯蚀刻。美国专利第5474796号中描述了其他方法和官能化剂,该专利通过引用以其整体并入本文。

[0177] 在一些情况下,在将硅烷偶联到装置表面的有效反应条件下,典型地通过存在于装置表面上的反应性亲水部分使装置表面通过与包含硅烷混合物的衍生组合物接触而官能化。硅烷化通常通过与有机官能烷氧基硅烷分子的自组装覆盖表面。

[0178] 多种硅氧烷官能化试剂可以进一步如本领域当前已知的被使用,例如用于降低或增加表面能。有机官能烷氧基硅烷可以根据它们的有机功能进行分类。

[0179] 本文提供了可以包含能够偶联到核苷的剂的图案化的装置。在一些情况下,装置可以用活性剂涂覆。在一些情况下,装置可以用惰性剂涂覆。本文描述的涂覆材料中包括的示例性活性剂包括但不限于N-(3-三乙氧基甲硅烷基丙基)-4-羟基丁酰胺(HAPS)、11-乙酰氧基十一烷基三乙氧基硅烷、N-癸基三乙氧基硅烷、(3-氨基丙基)三甲氧基硅烷、(3-氨基丙基)三乙氧基硅烷、3-缩水甘油氧基丙基三甲氧基硅烷(GOPS)、3-碘-丙基三甲氧基硅烷、

丁基-醛-三甲氧基硅烷、二聚仲氨基烷基硅氧烷、(3-氨基丙基)-二乙氧基-甲基硅烷、(3-氨基丙基)-二甲基-乙氧基硅烷、和(3-氨基丙基)-三甲氧基硅烷、(3-缩水甘油氧基丙基)-二甲基-乙氧基硅烷、缩水甘油氧基-三甲氧基硅烷、(3-巯基丙基)-三甲氧基硅烷、3-4环氧基环己基-乙基三甲氧基硅烷、和(3-巯基丙基)-甲基-二甲氧基硅烷、烯丙基三氯氯硅烷、7-辛-1-烯基三氯氯硅烷、或双(3-三甲氧基甲硅烷基丙基)胺。本文描述的涂覆材料中包括的示例性惰性剂包括但不限于全氟辛基三氯硅烷；十三氟-(1,1,2,2-四氢辛基)三氯硅烷；1H,1H,2H,2H-氟辛基三乙氧基硅烷(FOS)；三氯(1H,1H,2H,2H-全氟辛基)硅烷；叔丁基-[5-氟-4-(4,4,5,5-四甲基-1,3,2-二氧杂戊硼烷-2-基)吡啶-1-基]-二甲基硅烷；CYTOP™；Fluorinert™；全氟辛基三氯硅烷(PFOTCS)；全氟辛基二甲基氯硅烷(PFODCS)；全氟癸基三乙氧基硅烷(PFDTES)；五氟苯基-二甲基丙基氯-硅烷(PFP TES)；全氟辛基三乙氧基硅烷；全氟辛基三甲氧基硅烷；辛基氯硅烷；二甲基氯-十八烷基硅烷；甲基二氯十八烷基硅烷；三氯十八烷基硅烷；三甲基十八烷基硅烷；三乙基十八烷基硅烷；或者十八烷基三氯硅烷。

[0180] 在一些情况下，官能化剂包括烃硅烷，例如十八烷基三氯硅烷。在一些情况下，官能化剂包括11-乙酰氧基十一烷基三乙氧基硅烷、N-癸基三乙氧基硅烷、(3-氨基丙基)三甲氧基硅烷、(3-氨基丙基)三乙氧基硅烷、缩水甘油氧基丙基/三甲氧基硅烷和N-(3-三乙氧基硅烷基丙基)-4-羟基丁酰胺。

[0181] 多核苷酸合成

[0182] 当前公开内容的用于多核苷酸合成的方法可以包括涉及亚磷酰胺化学的过程。在一些情况下，多核苷酸合成包括将碱基与亚磷酰胺偶联。多核苷酸合成可以包括在偶联条件下通过亚磷酰胺的沉积来偶联碱基，其中相同的碱基被任选地与亚磷酰胺沉积多于一次，即双偶联。多核苷酸合成可以包括未反应位点的加帽。在一些情况下，加帽是任选的。多核苷酸合成还可以包括氧化或一个氧化步骤或多于一个氧化步骤。多核苷酸合成可以包括脱封闭、脱三苯甲基化和硫化。在一些情况下，多核苷酸合成包括氧化或硫化。在一些情况下，在多核苷酸合成反应期间的一个或每个步骤之间，使用例如四唑或乙腈洗涤该装置。亚磷酰胺合成方法中任何一步的时间范围可以小于约2min、1min、50sec、40sec、30sec、20sec和10sec。

[0183] 使用亚磷酰胺方法的多核苷酸合成可以包括随后向生长的多核苷酸链中添加亚磷酰胺构件单元(building block)(例如核苷亚磷酰胺)，以形成亚磷酸三酯键。亚磷酰胺多核苷酸合成以3'至5'方向进行。亚磷酰胺多核苷酸合成允许在每个合成循环向生长的核酸链中受控添加一个核苷酸。在一些情况下，每个合成循环包括偶联步骤。亚磷酰胺偶联包括在活化的核苷亚磷酰胺和通过例如接头结合到基板的核苷之间形成亚磷酸三酯键。在一些情况下，核苷亚磷酰胺被提供至活化的装置。在一些情况下，核苷亚磷酰胺与活化剂一起被提供至装置。在一些情况下，核苷亚磷酰胺被以多于基板结合的核苷1.5、2、3、4、5、6、7、8、9、10、11、12、13、14、15、16、17、18、19、20、25、30、35、40、50、60、70、80、90、100倍过量或更多提供给装置。在一些情况下，核苷亚磷酰胺的添加在无水环境中例如在无水乙腈中进行。加入核苷亚磷酰胺后，任选地洗涤装置。在一些情况下，偶联步骤另外重复一次或更多次，且任选地在核苷亚磷酰胺添加到基板之间进行洗涤步骤。在一些情况下，本文使用的多核苷酸合成方法包括1、2、3或更多个顺序偶联步骤。在偶联之前，在许多情况下，结合到装置上的核苷通过去除保护基团而被脱保护，其中保护基团起到防止聚合的作用。常见的保护

基团是4,4'-二甲氧基三苯甲基(DMT)。

[0184] 偶联后,亚磷酰胺多核苷酸合成方法任选地包括加帽步骤。在加帽步骤中,用加帽剂处理生长的多核苷酸。加帽步骤可用于在偶联后阻止未反应的基板结合的5'-羟基基团进一步链延伸,防止形成具有内部碱基缺失的多核苷酸。此外,用1H-四唑活化的亚磷酰胺可以在很小程度上与鸟苷的O6位置反应。不受理论的约束,在用I₂/水氧化时,这种副产物(可能通过O6-N7迁移)可能经历脱嘌呤作用。在多核苷酸的最终脱保护过程中,无嘌呤(apurinic)位点可能最终被裂解,从而减少全长产物的产量。O6修饰可以通过在用I₂/水氧化之前用加帽剂处理来去除。在一些情况下,在多核苷酸合成期间包括加帽步骤与没有加帽的合成相比降低了错误率。例如,加帽步骤包括用乙酸酐和1-甲基咪唑的混合物处理与基板结合的多核苷酸。在加帽步骤之后,任选地洗涤装置。

[0185] 在一些情况下,在添加核苷亚磷酰胺之后,并且任选地在加帽和一个或多个洗涤步骤之后,装置结合的生长的核酸被氧化。氧化步骤包括将亚磷酸三酯氧化成四配位磷酸三酯,这是天然存在的磷酸二酯核苷间键的受保护前体。在一些情况下,生长的多核苷酸的氧化是通过任选地在弱碱(例如吡啶、二甲基吡啶(lutidine)、可力丁(collidine))存在下用碘和水处理实现的。氧化可以在无水条件下,使用例如叔丁基过氧化氢或(1S)-(+)-(10-樟脑磺酰基)-氧杂吡丙啉(CSO)进行。在一些方法中,在氧化之后进行加帽步骤。第二个加帽步骤允许装置干燥,因为可能持续存在的从氧化残余的水可以抑制随后的偶联。氧化后,任选地洗涤装置和生长的多核苷酸。在一些情况下,氧化步骤被硫化步骤替代,以获得多核苷酸硫代磷酸酯,其中任何加帽步骤可以在硫化之后进行。许多试剂能够有效地转移硫,包括但不限于3-(二甲基氨基亚甲基)氨基)-3H-1,2,4-二噻唑-3-硫酮、DDTT、3H-1,2-苯并二硫-3-酮1,1-二氧化物、也称为Beaucage试剂、和二硫化N,N,N'-四乙基秋兰姆(TETD)。

[0186] 为了核苷掺入的随后循环通过偶联发生,装置结合的生长的多核苷酸的受保护的5'末端被去除,使得伯羟基基团是与下一个核苷亚磷酰胺反应性的。在一些情况下,保护基团是DMT,并且在二氯甲烷中用三氯乙酸进行脱封闭。持续延长的时间或用比推荐更强的酸溶液进行脱三苯甲基化可能导致固体支持物结合的多核苷酸的脱嘌呤作用增加,并且从而降低期望的全长产物的产量。本文描述的公开内容的方法和组合物提供了限制不期望的脱嘌呤反应的受控的脱封闭条件。在一些情况下,脱封闭后洗涤装置结合的多核苷酸。在一些情况下,脱封闭后的有效洗涤有助于合成具有低错误率的多核苷酸。

[0187] 用于合成多核苷酸的方法通常包括以下步骤的重复序列:将受保护单体施加到活性官能化表面(例如位点),以与活化表面、接头或与先前脱保护的单体连接;将施加的单体脱保护,使得其与随后施加的受保护单体是反应性的;以及施加用于连接的另一种受保护单体。一个或多个中间步骤包括氧化或硫化。在一些情况下,一个或多个洗涤步骤在一个或所有步骤之前或之后。

[0188] 用于基于亚磷酰胺的多核苷酸合成的方法包括一系列化学步骤。在一些情况下,合成方法的一个或多个步骤涉及试剂循环,其中该方法的一个或多个步骤包括将可用于步骤的试剂应用于装置。例如,试剂通过一系列液体沉积和真空干燥步骤循环。对于包括诸如孔、微孔、通道等三维特征的基板,试剂任选地经由孔和/或通道通过装置的一个或多个区域。

[0189] 本文描述的方法和系统涉及用于合成多核苷酸的多核苷酸合成装置。合成可以是并行的。例如至少或大约至少2、3、4、5、6、7、8、9、10、11、12、13、14、15、16、17、18、19、20、21、22、23、24、25、30、35、40、45、50、100、150、200、250、300、350、400、450、500、550、600、650、700、750、800、850、900、1000、10000、50000、75000、100000种或更多多核苷酸可以并行合成。可以并行合成的多核苷酸的总数可以是2-100000、3-50000、4-10000、5-1000、6-900、7-850、8-800、9-750、10-700、11-650、12-600、13-550、14-500、15-450、16-400、17-350、18-300、19-250、20-200、21-150、22-100、23-50、24-45、25-40、30-35。本领域技术人员理解，并行合成的多核苷酸的总数可以落入由这些值中的任何一个限定的任何范围内，例如25-100。并行合成的多核苷酸的总数可以落入由用作范围端点的任何值定义的任何范围内。装置内合成的多核苷酸的总摩尔质量或每种多核苷酸的摩尔质量可以是至少或至少约10、20、30、40、50、100、250、500、750、1000、2000、3000、4000、5000、6000、7000、8000、9000、10000、25000、50000、75000、100000皮摩尔或更多。装置内每个多核苷酸的长度或多核苷酸的平均长度可以是至少或大约至少10、15、20、25、30、35、40、45、50、100、150、200、300、400、500个核苷酸或更多。装置内每个多核苷酸的长度或多核苷酸的平均长度可以是最多或大约最多500、400、300、200、150、100、50、45、35、30、25、20、19、18、17、16、15、14、13、12、11、10个核苷酸或更少。装置内每个多核苷酸的长度或多核苷酸的平均长度可以落入从10-500、9-400、11-300、12-200、13-150、14-100、15-50、16-45、17-40、18-35、19-25。本领域技术人员理解，装置内每个多核苷酸的长度或多核苷酸的平均长度可以落入由这些值中的任何一个限定的任何范围内，例如100-300。装置内每个多核苷酸的长度或多核苷酸的平均长度可以落入由用作范围端点的任何值定义的任何范围内。

[0190] 本文提供的用于在表面上合成多核苷酸的方法允许快速合成。作为实例，每小时合成至少3、4、5、6、7、8、9、10、11、12、13、14、15、16、17、18、19、20、21、22、23、24、25、26、27、28、29、30、35、40、45、50、55、60、70、80、90、100、125、150、175、200个核苷酸或更多。核苷酸包括腺嘌呤、鸟嘌呤、胸腺嘧啶、胞嘧啶、尿苷构件单元或其类似物/修饰形式。在一些情况下，多核苷酸文库在基板上并行合成。例如，包括大约或至少大约100个、1,000、10,000、30,000、75,000、100,000、1,000,000、2,000,000、3,000,000、4,000,000、或者5,000,000个解析位点的装置能够支持至少相同数量的不同多核苷酸的合成，其中编码不同序列的多核苷酸在解析位点上合成。在一些情况下，在少于约三个月、两个月、一个月、三周、15、14、13、12、11、10、9、8、7、6、5、4、3、2天、24小时或更少的时间内，在本文描述的具有低错误率的装置上合成多核苷酸文库。在一些情况下，从使用本文所述的基板和以低错误率合成的多核苷酸文库组装的较大核酸在少于约三个月、两个月、一个月、三周、15、14、13、12、11、10、9、8、7、6、5、4、3、2天、24小时或更少的时间内制备。

[0191] 在一些情况下，本文描述的方法提供了多核苷酸文库的产生，该文库包含在多于一个密码子位点不同的变体多核苷酸。在一些情况下，多核苷酸可以具有1个位点、2个位点、3个位点、4个位点、5个位点、6个位点、7个位点、8个位点、9个位点、10个位点、11个位点、12个位点、13个位点、14个位点、15个位点、16个位点、17个位点、18个位点、19个位点、20个位点、30个位点、40个位点、50个位点或更多变体密码子位点。

[0192] 在一些情况下，变体密码子位点的一个或更多个位点可以相邻。在一些情况下，变体密码子位点的一个或更多个位点可以不相邻，并且被1、2、3、4、5、6、7、8、9、10或更多个密

码子分开。

[0193] 在一些情况下,多核苷酸可以包含变体密码子位点的多于一个位点,其中所有变体密码子位点彼此相邻,形成变体密码子位点的链段。在一些情况下,多核苷酸可以包含变体密码子位点的多于一个位点,其中所有变体密码子位点彼此不相邻。在一些情况下,多核苷酸可以包括变体密码子位点的多于一个位点,其中一些变体密码子位点彼此相邻,形成变体密码子位点的链段,并且一些变体密码子位点彼此不相邻。

[0194] 参考附图,图5示出了从较短的多核苷酸合成核酸(例如基因)的示例性过程工作流程。工作流程通常分为几个阶段:(1)从头合成单链多核苷酸文库,(2)连接多核苷酸以形成更大的片段,(3)错误校正,(4)质量控制,和(5)运送。在从头合成之前,预先选择意图的核酸序列或核酸序列的组。例如,一组基因被预先选择用于产生。

[0195] 在选择了用于产生的大的多核苷酸后,设计了用于从头合成的预定的多核苷酸文库。已知用于产生高密度多核苷酸阵列的多种合适的方法。在工作流程实例中,提供了装置表面层501。在该实例中,为了改善多核苷酸合成过程,改变了表面的化学性质。产生低表面能的区域来排斥液体,同时产生高表面能的区域来吸引液体。表面本身可以呈平坦表面的形式,或者包含形状的变化,例如增加表面积突起或微孔。在工作流程实例中,选择的高表面能分子用于支持DNA化学的双功能,如国际专利申请公布WO/2015/021080中公开的,其通过引用以其整体并入本文。

[0196] 多核苷酸阵列的原位制备在固体支持物上产生,并利用单个核苷酸延伸过程并行延伸多于一个寡聚体。材料沉积装置,例如多核苷酸合成仪,被设计成以分步方式释放试剂,使得多于一个多核苷酸一次并行延伸一个残基,以产生具有预定核酸序列的寡聚体502。在一些情况下,多核苷酸在这一阶段被从表面裂解。裂解包括气体裂解,例如用氨或甲胺裂解。

[0197] 产生的多核苷酸文库被放置在反应室中。在该示例性工作流程中,反应室(也称为“纳米反应器”)是硅涂覆的孔,包含PCR试剂,并降低到多核苷酸文库503上。在多核苷酸密封504之前或之后,添加试剂以从基板释放多核苷酸。在示例性工作流程中,多核苷酸在纳米反应器505密封后释放。释放后,单链多核苷酸片段杂交,以跨越全长范围的DNA序列。部分杂交505是可能的,因为每个合成的多核苷酸被设计成具有与群体中至少一个其他多核苷酸重叠的小部分。

[0198] 杂交后,开始PCR反应。在聚合酶循环期间,多核苷酸与互补片段退火,并通过聚合酶填充空位。每个循环随机增加各种片段的长度,这取决于哪种多核苷酸找到彼此。片段之间的互补性允许形成完整的大跨度双链DNA 506。

[0199] PCR完成后,将纳米反应器与装置507分开,并放置用于与具有PCR的引物的装置508相互作用。密封后,纳米反应器经受PCR 309,且较大的核酸被扩增。PCR 510之后,打开纳米室511,添加错误校正试剂512,将室密封513,并且发生错误校正反应以从双链PCR扩增产物514中去除错配的碱基对和/或具有弱互补性的链。将纳米反应器打开和分开515。然后错误校正的产物经受额外的处理步骤,诸如PCR和分子条形码,并且然后包装522用于运送523。

[0200] 在一些情况下,采取质量控制措施。错误校正后,质量控制步骤包括例如与具有用于扩增错误校正产物的测序引物的薄片516相互作用,将薄片密封到包含错误校正的扩增

产物的室517中,并进行额外一轮扩增518。将纳米反应器打开519,并将产物汇集520并测序521。在做出可接受的质量控制确定之后,批准包装的产物522用于运送523。

[0201] 在一些情况下,通过诸如图5中的工作流程产生的核酸经受使用本文公开的重叠引物的诱变。在一些情况下,引物文库通过在固体支持物上原位制备而产生,并利用单个核苷酸延伸过程来并行延伸多于一个寡聚体。沉积装置,诸如多核苷酸合成仪,被设计成以分步式的方式释放试剂,使得多于一个多核苷酸一次并行延伸一个残基,以产生具有预定核酸序列的寡聚体502。

[0202] 具有低错误率的大型多核苷酸文库

[0203] 使用所提供的系统和方法在文库中合成的多核苷酸的平均错误率可以小于1000分之1、小于1250分之1、小于1500分之1、小于2000分之1、小于3000分之1或更少比率。在一些情况下,使用所提供的系统和方法在文库中合成的多核苷酸的平均错误率小于1/500、1/600、1/700、1/800、1/900、1/1000、1/1100、1/1200、1/1250、1/1300、1/1400、1/1500、1/1600、1/1700、1/1800、1/1900、1/2000、1/3000或更少。在一些情况下,使用所提供的系统和方法在文库中合成的多核苷酸的平均错误率小于1/1000。

[0204] 在一些情况下,与预定序列相比,使用所提供的系统和方法在文库中合成的多核苷酸的总错误率小于1/500、1/600、1/700、1/800、1/900、1/1000、1/1100、1/1200、1/1250、1/1300、1/1400、1/1500、1/1600、1/1700、1/1800、1/1900、1/2000、1/3000或更少。在一些情况下,使用所提供的系统和方法在文库中合成的多核苷酸的总错误率小于1/500、1/600、1/700、1/800、1/900或1/1000。在一些情况下,使用所提供的系统和方法在文库中合成的多核苷酸的总错误率小于1/1000。

[0205] 在一些情况下,错误校正酶可以被用于使用所提供的系统和方法在文库中合成的多核苷酸。在一些情况下,与预定序列相比,具有错误校正的多核苷酸的总错误率可以小于1/500、1/600、1/700、1/800、1/900、1/1000、1/1100、1/1200、1/1300、1/1400、1/1500、1/1600、1/1700、1/1800、1/1900、1/2000、1/3000或更少。在一些情况下,使用所提供的系统和方法在文库中合成的多核苷酸的错误校正的总错误率可以小于1/500、1/600、1/700、1/800、1/900或1/1000。在一些情况下,使用所提供的系统和方法在文库中合成的多核苷酸的错误校正的总错误率可以小于1/1000。

[0206] 错误率可能限制基因合成对于产生基因变体文库的价值。对于1/300的错误率,1500个碱基对的基因中约0.7%的克隆将是正确的。由于多核苷酸合成中的大多数错误都会导致移码突变,因此这种文库中多于99%的克隆将不产生全长蛋白。使错误率降低75%,将使正确克隆的分数增加40倍。本公开内容的方法和组合物允许以低于通常观察到的基因合成方法的错误率快速从头合成大型多核苷酸和基因文库,这是由于合成质量的改善和以大规模并行且时间高效方式实现的错误校正方法的适用性。因此,可以合成具有以下特征的文库:遍及文库或遍及文库的多于80%、85%、90%、93%、95%、96%、97%、98%、99%、99.5%、99.8%、99.9%、99.95%、99.98%、99.99%或更多,碱基插入、缺失、取代或总错误率低于1/300、1/400、1/500、1/600、1/700、1/800、1/900、1/1000、1/1250、1/1500、1/2000、1/2500、1/3000、1/4000、1/5000、1/6000、1/7000、1/8000、1/9000、1/10000、1/12000、1/15000、1/20000、1/25000、1/30000、1/40000、1/50000、1/60000、1/70000、1/80000、1/90000、1/100000、1/125000、1/150000、1/200000、1/300000、1/400000、1/500000、1/

600000、1/700000、1/800000、1/900000、1/1000000或更少。本公开内容的方法和组合物还涉及与预定/预选序列相比具有低错误率的大型合成多核苷酸和基因文库,所述低错误率与涉及无错序列的文库的至少一个子集中的至少30%、40%、50%、60%、70%、75%、80%、85%、90%、93%、95%、96%、97%、98%、99%、99.5%、99.8%、99.9%、99.95%、99.98%、99.99%或更多的多核苷酸或基因相关。在一些情况下,文库内分离体积中的至少30%、40%、50%、60%、70%、75%、80%、85%、90%、93%、95%、96%、97%、98%、99%、99.5%、99.8%、99.9%、99.95%、99.98%、99.99%或更多的多核苷酸或基因具有相同的序列。在一些情况下,涉及多于95%、96%、97%、98%、99%、99.5%、99.6%、99.7%、99.7%或更多相似性或同一性的任何多核苷酸或基因的至少30%、40%、50%、60%、70%、75%、80%、85%、90%、93%、95%、96%、97%、98%、99%、99.5%、99.8%、99.9%、99.95%、99.98%、99.99%或更多具有相同的序列。在一些情况下,涉及多核苷酸或基因上的指定位点的错误率被优化。因此,作为大文库的一部分的一个或更多个多核苷酸或基因的特定位置或多于一个选择的位点可以各自具有小于1/300、1/400、1/500、1/600、1/700、1/800、1/900、1/1000、1/1250、1/1500、1/2000、1/2500、1/3000、1/4000、1/5000、1/6000、1/7000、1/8000、1/9000、1/10000、1/12000、1/15000、1/20000、1/25000、1/30000、1/40000、1/50000、1/60000、1/70000、1/80000、1/90000、1/100000、1/125000、1/150000、1/200000、1/300000、1/400000、1/500000、1/600000、1/700000、1/800000、1/900000、1/1000000或更小的错误率。在多种情况下,这种错误优化的位点可以包括至少1、2、3、4、5、6、7、8、9、10、11、12、13、14、15、16、17、18、19、20、25、30、35、40、45、50、60、70、80、90、100、200、300、400、500、600、700、800、900、1000、1500、2000、2500、3000、4000、5000、6000、7000、8000、9000、10000、30000、50000、75000、100000、500000、1000000、2000000、3000000个或更多位点。错误优化的位点可以分布到至少1、2、3、4、5、6、7、8、9、10、11、12、13、14、15、16、17、18、19、20、25、30、35、40、45、50、60、70、80、90、100、200、300、400、500、600、700、800、900、1000、1500、2000、2500、3000、4000、5000、6000、7000、8000、9000、10000、30000、75000、100000、500000、1000000、2000000、3000000个或更多多核苷酸或基因。

[0207] 该错误率可以在有或没有错误校正的情况下实现。该错误率可以在遍及文库或遍及文库的多于80%、85%、90%、93%、95%、96%、97%、98%、99%、99.5%、99.8%、99.9%、99.95%、99.98%、99.99%或更多中实现。

[0208] 计算机系统

[0209] 本文描述的任何系统都可以可操作地连接到计算机,并且可以通过本地或远程的计算机来自动化。在多种情况下,本公开内容的方法和系统还可以包括计算机系统上的软件程序及其使用。因此,用于分配/真空/再填充功能的同步化,诸如编排和同步材料沉积装置的运行、分配动作和真空致动的计算机化控制,都在本公开内容的范围内。计算机系统可以被编程以在用户指定的碱基序列和材料沉积装置的位置之间进行联系,以将正确的试剂递送到基板的指定区域。

[0210] 图6中所示的计算机系统600可以被理解为可以从介质611和/或网络端口605读取指令的逻辑装置,网络端口605可以任选地连接到具有固定介质612的服务器609。诸如图6中所示的系统可以包括CPU601、磁盘驱动器603、任选的输入装置,诸如键盘615和/或鼠标616以及任选的显示器607。数据通信可以通过指示的通信媒介到本地或远程位置的服务器

来实现。通信媒介可以包括发送和/或接收数据的任何手段。例如,通信媒介可以是网络连接、无线连接或互联网连接。这种连接可以提供通过万维网的通信。设想的是,如图6中所示,与本公开内容相关的数据可以通过这样的网络或连接传输,以供一方622接收和/或查看。

[0211] 图7是示出可以结合本公开内容的示例性情况使用的计算机系统700的第一示例结构的框图。如图7中描绘的,示例计算机系统可以包括用于处理指令的处理器702。处理器的非限制性实例包括:英特尔至强处理器、AMD皓龙处理器、三星32位RISC ARM 1176JZ (F)-S v1.0TM处理器、ARM Cortex-A8三星S5PC100TM处理器、ARM Cortex-A8苹果A4TM处理器、Marvell PXA 930TM处理器或功能等同的处理器。多线程的执行可以被用于并行处理。在一些情况下,也可以使用多于一个处理器或具有多于一个核的处理器,无论是在单个计算机系统中、在集群中,还是通过包括多于一个计算机、蜂窝电话和/或个人数据助理装置的网络遍及系统分布。

[0212] 如图7中所示,高速缓存704可以连接到处理器702,或者并入到处理器702中,以便为处理器702最近使用或者经常使用的指令或数据提供高速存储器。处理器702通过处理器总线708连接到北桥706。北桥706通过存储器总线712连接到随机存取存储器 (RAM) 710,并管理处理器702对RAM710的访问。北桥706还通过芯片组总线716连接到南桥714。南桥714转而连接到外围总线718。外围总线可以是例如PCI、PCI-X、PCI Express或其他外围总线。北桥和南桥通常被称为处理器芯片组,并管理处理器、RAM和外围总线718上的外围组件之间的数据传输。在一些可选择的结构中,北桥的功能可以并入到处理器中,而不是使用单独的北桥芯片。在一些情况下,系统700可以包括附接到外围总线718的加速器卡722。加速器可以包括现场可编程门阵列 (FPGA) 或用于加速某些处理的其他硬件。例如,加速器可以被用于自适应数据重组或评估扩展集处理中使用的代数表达式。

[0213] 软件和数据存储在外部存储器724中,并且可以被加载到RAM710和/或高速缓存704中供处理器使用。系统700包括用于管理系统资源的操作系统;操作系统的非限制性实例包括:Linux、WindowsTM、MACOSTM、BlackBerry OSTM、iOSTM和其他功能等同的操作系统,以及在操作系统之上运行的用于根据本公开内容的示例性实例管理数据存储和优化的应用软件。在这个例子中,系统700还包括连接到外围总线的网络接口卡 (NIC) 720和721,用于提供网络接口到外部存储器诸如网络附接存储器 (NAS) 和可以用于分布式并行处理的其他计算机系统。

[0214] 图8是示出具有多于一个计算机系统802a和802b、多于一个蜂窝电话和个人数据助理802c以及网络附接存储器 (NAS) 804a和804b的网络800的图。在示例性实例中,系统802a、802b和802c可以管理数据存储并优化存储在网络附接存储器 (NAS) 804a和804b中的数据的数据访问。数学模型可以用于数据,并且可以使用跨计算机系统802a和802b以及蜂窝电话和个人数据助理系统802c的分布式并行处理来评估。计算机系统802a和802b以及蜂窝电话和个人数据助理系统802c也可以为存储在网络附接存储器 (NAS) 804a和804b中的数据的自适应数据重组提供并行处理。图8仅示出了实例,并且各种其他计算机结构和系统可以与本公开内容的各种实例一起使用。例如,刀片服务器可以用于提供并行处理。处理器刀片可以通过背板连接,以提供并行处理。存储器也可以连接到背板或通过单独的网络接口作为网络附接存储器 (NAS)。在一些示例性实例中,处理器可以维持单独的存储空间,并通

过网络接口、背板或其他连接器传输数据,以供其他处理器并行处理。在其他情况下,一些或所有处理器可以使用共享虚拟地址存储空间。

[0215] 图9是根据示例性实例使用共享虚拟地址存储空间的多处理器计算机系统900的框图。该系统包括可以访问共享存储器子系统904的多于一个处理器902a-f。该系统在存储器子系统904中并入了多于一个可编程硬件存储器算法处理器 (MAP) 906a-f。每个MAP 906a-f可以包括存储器908a-f和一个或更多个现场可编程门阵列 (FPGA) 910a-f。MAP提供可配置的功能单元,并且特定的算法或算法的一部分可以被提供给FPGA 910a-f,用于与各自的处理器紧密协作进行处理。例如,MAP可以用于评估关于数据模型的代数表达式,并在示例性实例中执行自适应数据重组。在这个实例中,出于这些目的,所有处理器都可以全局访问每个MAP。在一种配置中,每个MAP可以使用直接存储器访问 (DMA) 来访问相关的存储器908a-f,这允许其独立于各自的微处理器902a-f并与其异步地执行任务。在这种配置中,MAP可以将结果直接馈送到另一个MAP,用于流水线操作和算法的并行执行。

[0216] 以上计算机体系结构和系统仅是实例,并且多种其他计算机、蜂窝电话和个人数据助理结构和系统可以结合示例性实例使用,包括使用通用处理器、协处理器、FPGA和其他可编程逻辑装置的任何组合的系统、芯片上系统 (SOC)、专用集成电路 (ASIC) 以及其他处理和逻辑元件。在一些情况下,计算机系统的全部或部分可以用软件或硬件来实现。任何种类的数据存储介质都可以与示例性实例结合使用,包括随机存取存储器、硬盘驱动器、闪存、磁带驱动器、磁盘阵列、网络附接存储器 (NAS) 以及其他本地或分布式数据存储装置和系统。

[0217] 在示例性实例中,计算机系统可以使用在任何以上或其他体系结构和系统上执行的软件模块来实现。在其他情况下,系统的功能可以部分或完全在固件、可编程逻辑装置诸如图9中提及的现场可编程门阵列 (FPGA)、芯片上系统 (SOC)、专用集成电路 (ASIC) 或其他处理和逻辑元件中实现。例如,集处理器和优化器可以通过使用硬件加速器卡,诸如图7中所示的加速器卡722,以硬件加速来实现。

[0218] 另外的方法和组合物

[0219] 本文提供了用于产生多核苷酸文库的方法,所述方法包括:

[0220] 提供编码至少约5000种非相同的多核苷酸的预定序列;

[0221] 提供具有表面的结构,其中该表面包括多于一个簇;合成至少约5000种非相同的多核苷酸,其中至少5000种非相同的多核苷酸中的每一种延伸不同的位点;以及扩增至少5000种非相同的多核苷酸以形成多核苷酸文库,其中至少5000种非相同的多核苷酸中的大于约80%以多核苷酸文库平均表现度的至少约2倍内的量表现。本文还提供了方法,其中至少5000种非相同的多核苷酸中的大于约80%以多核苷酸文库平均表现度的至少约1.5倍内的量表现。本文还提供了方法,其中至少5000种非相同的多核苷酸中大于约90%以多核苷酸文库平均表现度的至少约2倍内的量表现。本文还提供了方法,其中至少5000种非相同的多核苷酸中大于约90%以多核苷酸文库平均表现度的至少约1.5倍内的量表现。本文还提供了方法,其中与来自使用具有非聚簇位点的表面的结构的方法的扩增产物相比,多核苷酸文库包含较少的失落。本文还提供了方法,其中至少约5000种非相同的多核苷酸包含具有至少约10%的GC百分比的多核苷酸。本文还提供了方法,其中至少约5000种非相同的多核苷酸包含具有最多95%的GC百分比的多核苷酸。本文还提供了方法,其中至少约5000种

非相同的多核苷酸包含具有约10%至约95%的GC百分比的多核苷酸。本文还提供了方法,其中大于30%的至少约5000种非相同的多核苷酸包含具有10%至30%或70%至90%的GC百分比的多核苷酸。本文还提供了方法,其中少于约15%的至少约5000种非相同的多核苷酸包含具有10%至30%或60%至90%的GC百分比的多核苷酸。本文还提供了方法,其中至少约5000种非相同的多核苷酸包含具有至少约10%的重复序列百分比的多核苷酸。本文还提供了方法,其中至少约5000种非相同的多核苷酸包含具有最多95%的重复序列百分比的多核苷酸。本文还提供了方法,其中至少约5000种非相同的多核苷酸包含具有约10%至约95%的重复序列百分比的多核苷酸。本文还提供了方法,其中大于30%的至少约5000种非相同的多核苷酸包含具有10%至30%或70%至90%的重复序列百分比的多核苷酸。本文还提供了方法,其中少于约15%的至少约5000种非相同的多核苷酸包含具有10%至30%或60%至90%的重复序列百分比的多核苷酸。本文还提供了方法,其中至少约5000种非相同的多核苷酸包含具有至少约10%的二级结构百分比的多核苷酸。本文还提供了方法,其中至少约5000种非相同的多核苷酸包含具有最多95%的二级结构百分比的多核苷酸。本文还提供了方法,其中至少约5000种不相同的多核苷酸包含具有约10%至约95%的二级结构百分比的多核苷酸。本文还提供了方法,其中大于30%的至少约5000种非相同的多核苷酸包含具有10%至30%或70%至90%的二级结构百分比的多核苷酸。本文还提供了方法,其中少于约15%的至少约5000种非相同的多核苷酸包含具有10%至30%或60%至90%的二级结构百分比的多核苷酸。本文还提供了方法,其中多核苷酸文库编码变体文库。本文还提供了方法,其中至少5000种非相同的多核苷酸文库编码至少一种基因。本文还提供了方法,其中至少5000种非相同的多核苷酸文库编码至少50种基因。本文还提供了方法,其中多核苷酸文库编码至少一种基因。本文还提供了方法,其中多核苷酸文库编码抗体、酶或肽的至少一部分。本文还提供了方法,其中与预定序列相比,没有校正错误的多核苷酸文库具有500个碱基中少于1个碱基的总错误率。本文还提供了方法,其中与预定序列相比,没有校正错误的多核苷酸文库具有1000个碱基中少于1个碱基的总错误率。本文还提供了方法,其中预定序列编码至少700,000种非相同的多核苷酸。本文还提供了方法,其中每个簇包含50至约500个用于多核苷酸合成的位点。本文还提供了方法,其中每个簇包含高达约500个用于多核苷酸合成的位点。

[0222] 本文提供了用于产生多核苷酸文库的方法,所述方法包括:提供编码至少约5000种非相同的多核苷酸的预定序列;提供具有表面的结构,其中该表面包括多于一个簇;合成至少约5000种非相同的多核苷酸,其中至少5000种非相同的多核苷酸中的每一种延伸不同的位点;以及扩增至少5000种非相同的多核苷酸以形成多核苷酸文库,其中扩增反应后多核苷酸文库具有大于75%的正确序列率。本文还提供了方法,其中多核苷酸文库在扩增反应后具有大于85%的正确序列率。本文还提供了方法,其中至少约5000种非相同的多核苷酸包含具有至少约10%的GC百分比的多核苷酸。本文还提供了方法,其中至少约5000种非相同的多核苷酸包含具有最多95%的GC百分比的多核苷酸。本文还提供了方法,其中至少约5000种非相同的多核苷酸包含具有约10%至约95%的GC百分比的多核苷酸。本文还提供了方法,其中大于30%的至少约5000种非相同的多核苷酸包含具有10%至30%或70%至90%的GC百分比的多核苷酸。本文还提供了方法,其中少于约15%的至少约5000种非相同的多核苷酸包含具有10%至30%或60%至90%的GC百分比的多核苷酸。本文还提供了方

法,其中与使用具有非集簇位点的的表面的结构的方法的扩增产物相比,多核苷酸文库包含较少的失落。本文还提供了方法,其中多核苷酸文库编码变体文库。本文还提供了方法,其中多核苷酸文库编码至少一种基因。本文还提供了方法,其中多核苷酸文库编码抗体、酶或肽的至少一部分。本文还提供了方法,其中与预定序列相比,没有校正错误的多核苷酸文库具有500个碱基中少于1个碱基的总错误率。本文还提供了方法,其中与预定序列相比,没有校正错误的多核苷酸文库具有1000个碱基中少于1个碱基的总错误率。本文还提供了方法,其中预定序列编码至少700,000种非相同的多核苷酸。本文还提供了方法,其中至少5000种非相同的多核苷酸编码至少一种基因。本文还提供了方法,其中至少5000种非相同的多核苷酸编码至少50种基因。本文还提供了方法,其中每个簇包含50至约500个用于多核苷酸合成的位点。本文还提供了方法,其中每个簇包含高达约500个用于多核苷酸合成的位点。

[0223] 本文提供了包含至少5000种非相同多核苷酸的核酸文库,其中至少5000种非相同的多核苷酸是合成的多核苷酸的扩增产物,并且其中至少5000种非相同的多核苷酸中的大于约80%以核酸文库平均表现度的至少约2倍内的量表现。本文还提供了核酸文库,其中至少5000种非相同的多核苷酸中的大于约80%以核酸文库平均表现度的至少约1.5倍内的量表现。本文还提供了核酸文库,其中至少5000种非相同的多核苷酸中的大于约90%以核酸文库平均表现度的至少约2倍内的量表现。本文还提供了核酸文库,其中至少5000种非相同多核苷酸中的大于约90%以核酸文库平均表现度的至少约1.5倍内的量表现。本文还提供了核酸文库,其中与来自使用具有非集簇位点的的表面的结构的方法的扩增产物相比,核酸文库包含较少的失落。本文还提供了核酸文库,其中至少约5000种非相同的多核苷酸包含具有至少约10%的GC百分比的多核苷酸。本文还提供了核酸文库,其中至少约5000种非相同的多核苷酸包含具有最多95%的GC百分比的多核苷酸。本文还提供了核酸文库,其中至少约5000种不相同的多核苷酸包含具有约10%至约95%的GC百分比的多核苷酸。本文还提供了核酸文库,其中大于30%的至少约5000种非相同的多核苷酸包含具有10%至30%或70%至90%的GC百分比的多核苷酸。本文还提供了核酸文库,其中少于约15%的至少约5000种非相同的多核苷酸包含具有10%至30%或60%至90%的GC百分比的多核苷酸。本文还提供了核酸文库,其中至少约5000种非相同的多核苷酸包含具有至少约10%的重复序列百分比的多核苷酸。本文还提供了核酸文库,其中至少约5000种不相同的多核苷酸包含具有最多95%的重复序列百分比的多核苷酸。本文还提供了核酸文库,其中至少约5000种非相同的多核苷酸包含具有约10%至约95%的重复序列百分比的多核苷酸。本文还提供了核酸文库,其中大于30%的至少约5000种非相同的多核苷酸包含具有10%至30%或70%至90%的重复序列百分比的多核苷酸。本文还提供了核酸文库,其中少于约15%的至少约5000种非相同的多核苷酸包含具有10%至30%或60%至90%的重复序列百分比的多核苷酸。本文还提供了核酸文库,其中至少约5000种非相同的多核苷酸包含具有至少约10%的二级结构百分比的多核苷酸。本文还提供了核酸文库,其中至少约5000种非相同的多核苷酸包含具有最多95%的二级结构百分比的多核苷酸。本文还提供了核酸文库,其中至少约5000种非相同的多核苷酸包含具有约10%至约95%的二级结构百分比的多核苷酸。本文还提供了核酸文库,其中大于30%的至少约5000种非相同的多核苷酸包含具有10%至30%或70%至90%的二级结构百分比的多核苷酸。本文还提供了核酸文库,其中少于约15%的至少约5000种非相同的多核苷酸中包含具有10%至30%或60%至90%的二级结构百分比的

多核苷酸。本文还提供了核酸文库,其中多核苷酸文库编码变体文库。本文还提供了核酸文库,其中至少5000种非相同的多核苷酸编码至少一种基因。本文还提供了核酸文库,其中至少5000种非相同的多核苷酸编码至少50种基因。本文还提供了核酸文库,其中多核苷酸文库编码抗体、酶或肽的至少一部分。本文还提供了核酸文库,其中预定序列编码至少700,000种非相同的多核苷酸。

[0224] 本文提供了包含至少5000种非相同的多核苷酸的核酸文库,其中GC含量是受控的,并且其中文库在扩增反应后提供大于75%的正确序列率。本文还提供了具有在扩增反应后大于85%的正确序列率的核酸文库。本文还提供了核酸文库,其中至少约5000种非相同的多核苷酸包含具有至少约10%的GC百分比的多核苷酸。本文还提供了核酸文库,其中至少约5000种非相同的多核苷酸包含具有最多95%的GC百分比的多核苷酸。本文还提供了核酸文库,其中至少约5000种非相同的多核苷酸包含具有约10%至约95%的GC百分比的多核苷酸。本文还提供了核酸文库,其中大于30%的至少约5000种非相同的多核苷酸包含具有10%至30%或70%至90%的GC百分比的多核苷酸。本文还提供了核酸文库,其中少于约15%的至少约5000种非相同的多核苷酸包含具有10%至30%或60%至90%的GC百分比的多核苷酸。

[0225] 本文提供了包含至少5000种非相同的多核苷酸的核酸文库,其中重复序列含量是受控的,并且其中文库在扩增反应后提供大于75%的正确序列率。本文还提供了核酸文库,其中扩增反应后多核苷酸文库具有大于85%的正确序列率。本文还提供了核酸文库,其中至少约5000种非相同的多核苷酸包含具有至少约10%的重复序列百分比的多核苷酸。本文还提供了核酸文库,其中至少约5000种非相同的多核苷酸包含具有最多95%的重复序列百分比的多核苷酸。本文还提供了核酸文库,其中至少约5000种非相同的多核苷酸包含具有约10%至约95%的重复序列百分比的多核苷酸。本文还提供了核酸文库,其中大于30%的至少约5000种非相同的多核苷酸包含具有10%至30%或70%至90%的重复序列百分比的多核苷酸。本文还提供了核酸文库,其中少于约15%的至少约5000种非相同的多核苷酸包含具有10%至30%或60%至90%的重复序列百分比的多核苷酸。

[0226] 本文提供了包含至少5000种非相同的多核苷酸的核酸文库,其中至少5000种非相同的多核苷酸编码的二级结构含量是预选的,并且其中文库在扩增反应后提供了大于75%的正确序列率。本文还提供了核酸文库,其中核酸文库在扩增反应后具有大于85%的正确序列率。本文还提供了核酸文库,其中至少约5000种非相同的多核苷酸包含具有至少约10%的二级结构百分比的多核苷酸。本文还提供了核酸文库,其中至少约5000种非相同的多核苷酸包含具有最多95%的二级结构百分比的多核苷酸。本文还提供了核酸文库,其中至少约5000种非相同的多核苷酸包含具有约10%至约95%的二级结构百分比的多核苷酸。本文还提供了核酸文库,其中大于30%的至少约5000种非相同的多核苷酸包含具有10%至30%或70%至90%的二级结构百分比的多核苷酸。本文还提供了核酸文库,其中少于约15%的至少约5000种非相同的多核苷酸包含具有10%至30%或60%至90%的二级结构百分比的多核苷酸。本文还提供了编码变体文库的核酸文库。本文还提供了编码至少一种基因的核酸文库。本文还提供了编码抗体、酶或肽的至少一部分的核酸文库。本文还提供了在没有校正错误的情况下与预定序列相比,具有500个碱基中少于1个碱基的总错误率的核酸文库。本文还提供了在没有校正错误的情况下与预定序列相比,具有1000个碱基中少于1个

碱基的总错误率的核酸文库。本文还提供了核酸文库,其中预定序列编码至少700,000种非相同的多核苷酸。本文还提供了核酸文库,其中至少5000种非相同的多核苷酸编码至少一种基因。本文还提供了核酸文库,其中至少5000种非相同的多核苷酸编码至少50种基因。

[0227] 本文提供了用于多核苷酸文库扩增的方法,所述方法包括:获得至少5000种非相同的多核苷酸的扩增分布;基于至少一个序列特征将扩增分布的至少5000种非相同的多核苷酸聚类到两个或更多个箱中;并且基于两个或更多个箱中的每个箱中的至少5000种非相同的多核苷酸的数量的频率来调整每种非相同的多核苷酸的合成的表现度,以产生具有预选表现度的多核苷酸文库;合成具有预选表现度的多核苷酸文库;以及扩增具有预选表现度的多核苷酸文库。本文还提供了方法,其中至少一个序列特征是GC含量百分比。本文还提供了方法,其中至少一个序列特征是重复序列含量百分比。本文还提供了方法,其中至少一个序列特征是二级结构含量百分比。本文还提供了方法,其中重复序列包含3个或更多腺嘌呤。本文还提供了方法,其中重复序列在多核苷酸的一个或两个末端上。本文还提供了方法,其中基于一个或更多个多核苷酸序列结合靶序列的亲合力,将所述多核苷酸聚类到箱中。本文还提供了方法,其中当与调整之前的箱中的较低的30%的序列的数量相比时,调整之后的箱中的较低的30%的序列的数量在下游应用中具有至少多50%的表现度。本文还提供了方法,其中当与调整之前的箱中的较高的30%的序列的数量相比时,调整之后的箱中的较高的30%的序列的数量在下游应用中具有至少多50%的表现度。本文还提供了方法,其中所述扩增分布是凭经验获得的。本文还提供了方法,其中通过预测算法获得所述扩增分布。在一些情况下,调整包括控制文库中多核苷酸的化学计量。

[0228] 本文提供了包含至少100,000种非相同的多核苷酸的核酸文库,其中每种非相同的多核苷酸编码至少一种不同的外显子序列,并且其中至少100,000种非相同的多核苷酸中的至少约80%各自以文库中每种非相同的多核苷酸的平均频率的2x内的量存在于多核苷酸文库中。本文还提供了核酸文库,其中所述核酸文库是扩增子文库,并且其中多于一种非相同的多核苷酸中的至少约80%各自以文库中每种非相同的多核苷酸的平均频率的2x内的量存在于扩增子文库中。本文还提供了核酸文库,其中以高达55倍的理论读段深度对文库测序导致至少90%的碱基具有至少30倍的读段深度。本文还提供了核酸文库,其中以高达55倍的理论读段深度对文库测序导致至少98%的碱基具有至少10倍的读段深度。

[0229] 本文提供了用于合成多核苷酸文库的方法,所述方法包括:(a)提供至少100,000种非相同的多核苷酸的预定序列,其中每种非相同的多核苷酸编码基因组DNA的一个或更多个部分;(b)合成至少100,000种非相同的多核苷酸;和(c)扩增至少100,000种非相同的多核苷酸以产生多核苷酸文库,其中与至少100,000种非相同的多核苷酸的预定序列相比,文库中至少约75%的多核苷酸是无错误的。本文还提供了方法,其中多核苷酸文库是扩增子文库,并且其中多于一种非相同的多核苷酸中的至少约80%各自以文库中每种非相同的多核苷酸的平均频率的2x内的量存在于扩增子文库中。本文还提供了方法,其中每种非相同的多核苷酸编码一种或更多种外显子。本文还提供了方法,其中扩增的非相同的多核苷酸各自包含至少一个分子标签。

[0230] 本文提供了合成多核苷酸文库的方法,所述方法包括:(a)扩增至少2000种非相同的多核苷酸的第一文库;(b)根据一个或更多个序列特征识别第一文库中序列的分布;和(c)基于序列的分布改变第一文库中序列的相对比例,以产生第二文库,使得对第二文库的

不多于2.5x的取样导致至少80%的测序覆盖。本文还提供了方法,其中一个或更多个序列特征包括GC含量百分比。本文还提供了方法,其中一个或更多个序列特征包括重复序列含量百分比。本文还提供了方法,其中一个或更多个序列特征包括二级结构含量百分比。本文还提供了方法,其中一个或更多个序列特征包括测序覆盖。本文还提供了方法,其中不多于1.7x的取样导致至少80%的测序覆盖。本文还提供了方法,其中不多于2.5x的取样导致至少90%的测序覆盖。本文还提供了方法,其中该方法还包括合成第二文库。本文还提供了方法,其中该方法还包括扩增第二文库。本文还提供了方法,其中文库包含至少5,000种多核苷酸。本文还提供了方法,其中文库包含至少10,000种多核苷酸。本文还提供了方法,其中文库包含至少30,000种多核苷酸。

[0231] 本文提供了用于靶富集的方法,所述方法包括:使至少2000种非相同的双链多核苷酸的文库与包含靶核酸的样品多核苷酸群体接触,其中至少2000种非相同的多核苷酸中的每一个包括:(从5'至3'):第一非靶序列和第二非靶序列;和与一个或多个靶核酸序列互补的插入序列;捕获在固体支持物上与至少2000种非相同的多核苷酸中的一种或更多种杂交的靶核酸序列;以及释放捕获的靶核酸以产生富集的靶多核苷酸文库。本文还提供了方法,其中每个多核苷酸还包含至少一个分子标签。本文还提供了方法,其中每个非靶序列还包含引物结合位点。本文还提供了方法,其中第一非靶序列位于多核苷酸的5'末端,且第二非靶序列位于多核苷酸的3'末端。本文还提供了方法,其中将一个或更多个分子标签附接到多核苷酸的5'末端。本文还提供了方法,其中将一个或更多个分子标签附接到多核苷酸的3'末端。本文还提供了方法,其中一个或更多个分子标签和多核苷酸通过间隔区连接。本文还提供了方法,其中插入序列与至少一个外显子互补。本文还提供了方法,其中一个或更多个分子标签是生物素、叶酸、多组氨酸、FLAG标签或谷胱甘肽。本文还提供了方法,其中一个或更多个分子标签是两个生物素分子。本文还提供了方法,其中固体支持物是磁珠。本文还提供了方法,其中第一非靶序列和第二非靶序列的长度在20至40个碱基之间。本文还提供了方法,其中插入序列的长度在90至200个碱基之间。本文还提供了方法,其中文库包含至少5,000种多核苷酸。本文还提供了方法,其中文库包含至少10,000种多核苷酸。本文还提供了方法,其中文库包含至少30,000种多核苷酸。

[0232] 本文提供了包含多于一种部分互补性的双链多核苷酸的探针文库,每个探针文库包含:第一多核苷酸,其包含:第一非靶序列和第二非靶序列;和与一个或多个靶核酸序列互补的第一插入序列;第二多核苷酸,其包含:第一非靶序列和第二非靶序列;和与第一插入序列互补的第二插入序列;其中第一多核苷酸和第二多核苷酸部分杂交。本文还提供了文库,其中双链多核苷酸的每条链还包含至少两个分子标签。本文还提供了文库,其中第一非靶序列和第二非靶序列不互补。本文还提供了文库,其中第一非靶序列位于多核苷酸的5'末端,第二非靶序列位于多核苷酸的3'末端。本文还提供了文库,其中一个或更多个分子标签附接到多核苷酸的5'末端。本文还提供了文库,其中一个或更多个分子标签附接到多核苷酸的3'末端。本文还提供了文库,其中一个或更多个分子标签和多核苷酸通过间隔区连接。本文还提供了文库,其中插入序列与至少一个外显子互补。本文还提供了文库,其中一个或更多个分子标签是生物素、叶酸、多组氨酸、FLAG标签或谷胱甘肽。本文还提供了文库,其中一个或更多个分子标签是两个生物素分子。本文还提供了文库,其中固体支持物是磁珠。本文还提供了文库,其中第一非靶序列和第二非靶序列的长度在20至40个碱基之

间。本文还提供了文库,其中插入序列的长度在90至200个碱基之间。

[0233] 本文提供了用于设计探针文库的方法,所述方法包括:获得靶序列文库;以及设计与靶序列互补的插入序列文库,其中设计包括:如果靶序列的长度比插入序列短,则产生与靶序列互补的插入序列;如果靶序列的长度比插入序列+X短,则产生与靶序列至少部分互补的插入序列;或者如果靶序列比插入序列+X长,则产生与共同靶序列至少部分互补的一组插入序列,其中X是不被插入序列靶向的连续碱基的数量;对文库中的每个靶序列重复步骤(b),以产生插入序列文库。本文还提供了方法,其中X少于30个核苷酸。本文还提供了方法,其中X少于10个核苷酸。本文还提供了方法,其中X为约6个核苷酸。

[0234] 本文提供了用于下一代测序的方法,所述方法包括使本文描述的文库与包含多于一种靶多核苷酸的样品接触;富集至少一种与文库结合的靶多核苷酸;和对至少一种富集的靶多核苷酸测序。

[0235] 本文提供了用于下一代测序的方法,所述方法包括:使本文描述的文库与包含多于一个多核苷酸的样品接触;将样品中的至少一种与文库结合的多核苷酸与至少一种不与文库结合的多核苷酸分离;以及对至少一种不与文库结合的多核苷酸测序。

实施例

[0236] 给出以下实施例是为了说明本发明的各种实施方案的目的,并不意味着以任何方式限制本发明。本发明的实施例以及本文描述的方法目前是优选实施方案的代表,是示例性的,并不意图作为对本发明的范围的限制。本领域技术人员将会想到如本权利要求的范围所限定的本发明的精神内包含的其中的变化和其他用途。

[0237] 实施例1:基板表面的官能化

[0238] 使基板官能化以支持多核苷酸文库的附接和合成。首先使用包含90% H_2SO_4 和10% H_2O_2 的食人鱼溶液(piranha solution)湿法清洁基板表面20分钟。将基板在数个烧杯中用DI水冲洗,在DI水鹅颈水龙头下保持5分钟,并用 N_2 干燥。随后将基板浸泡在 NH_4OH (1:100; 3mL:300mL毫升)中5分钟,使用手枪用DI水冲洗,在三个连续烧杯中用DI水浸泡各1分钟,并且然后使用手枪用DI水再次冲洗。然后通过将基板表面暴露于 O_2 ,等离子体清洁基板。在下游模式下,使用SAMCO PC-300仪器以250瓦等离子体蚀刻 O_2 1分钟。

[0239] 使用具有以下参数:0.5至1托、60min、70°C、135°C蒸发器的YES-1224P气相沉积炉系统,用包含N-(3-三乙氧基甲硅烷基丙基)-4-羟基丁酰胺的溶液对清洁的基板表面进行活性官能化。使用Brewer Science 200X旋转涂布机对基板表面进行抗蚀剂涂布。将SPR™ 3612光致抗蚀剂以2500rpm持续40sec旋转涂覆在基板上。将基板在Brewer热板上于90°C预烘烤30min。使用Karl Suss MA6掩模对准器仪器使基板经受光刻。将基板曝光2.2sec,并在MSF 26A中显影1min。用手枪冲洗剩余的显影剂,并将基板浸泡在水中5min。将基板在烘箱中于100°C烘烤30min,随后使用Nikon L200目视检查光刻缺陷。使用SAMCO PC-300仪器以250瓦的 O_2 等离子体蚀刻1min,使用除渣处理去除残留的抗蚀剂。

[0240] 将基板表面用与10 μ L轻质矿物油混合的100 μ L全氟辛基三氯硅烷溶液惰性官能化。将基板放置在室中,泵送10min,并且然后关闭泵的阀门,并静置10min。使室通向空气。通过在70°C在500mLNMP中伴随最大功率超声波处理(在Crest系统上的9)进行两次5min的浸泡对基板抗蚀剂剥离。然后在室温将基板在500mL异丙醇中伴随最大功率的超声波处理

浸泡5min。将基板在300mL的无水乙醇(200-proof ethanol)中浸泡,并用N₂吹干。使官能化表面活化以用作多核苷酸合成的支持物。

[0241] 实施例2:在多核苷酸合成装置上合成50-mer序列

[0242] 将二维多核苷酸合成装置组装到流动池中,该流动池与流动池相连(Applied Biosystems (ABI394 DNA合成仪))。使用本文描述的多核苷酸合成方法,采用用N-(3-三乙氧基甲硅烷基丙基)-4-羟基丁酰胺(Gelest)均匀官能化的多核苷酸合成装置来合成50bp的示例性多核苷酸(“50-mer多核苷酸”)。

[0243] 50-mer的序列如SEQ ID NO.:1.

[0244] 5'AGACAATCAACCATTTGGGGTGGACAGCCTTGACCTCTAGACTTCGGCAT##TTTTTTTTTT3' (SEQ ID NO.:1)中所述,其中#表示胸苷-琥珀酰己酰胺CED亚磷酰胺(来自ChemGenes的CLP-2244),它是一种可裂解的接头,能够在脱保护期间从表面释放多核苷酸。

[0245] 根据表1中的方案,使用标准的DNA合成化学(偶联、加帽、氧化和脱封闭)和ABI合成仪完成合成。

[0246] 表1

[0247]

表 1		
通用 DNA 合成 过程名称	过程步骤	时间(sec)
洗涤(乙腈洗涤流)	乙腈系统冲洗	4
	乙腈至流动池	23
	N ₂ 系统冲洗	4
	乙腈系统冲洗	4
DNA 碱基添加 (亚磷酰胺+活化剂流)	活化剂歧管冲洗	2
	活化剂至流动池	6
	活化剂+亚磷酰胺至流动池	6
	活化剂至流动池	0.5
	活化剂+亚磷酰胺至流动池	5
	活化剂至流动池	0.5
	活化剂+亚磷酰胺至流动池	5
	活化剂至流动池	0.5
	活化剂+亚磷酰胺至流动池	5
	孵育 25 sec	25
洗涤(乙腈洗涤流)	乙腈系统冲洗	4
	乙腈至流动池	15

通用 DNA 合成 过程名称	表 1	
	过程步骤	时间(sec)
DNA 碱基添加 (亚磷酸胺+活化剂流)	N2 系统冲洗	4
	乙腈系统冲洗	4
	活化剂歧管冲洗	2
	活化剂至流动池	5
	活化剂+亚磷酸胺至流动池	18
洗涤(乙腈洗涤流)	孵育 25 sec	25
	乙腈系统冲洗	4
	乙腈至流动池	15
	N2 系统冲洗	4
加帽(CapA+B, 1:1, 流)	乙腈系统冲洗	4
	CapA+B 至流动池	15
	乙腈系统冲洗	4
	乙腈至流动池	15
氧化(氧化剂流)	乙腈系统冲洗	4
	氧化剂至流动池	18
	乙腈系统冲洗	4
	乙腈至流动池	15
[0248] 洗涤(乙腈洗涤流)	乙腈系统冲洗	4
	N2 系统冲洗	4
	乙腈系统冲洗	4
	乙腈至流动池	15
	乙腈系统冲洗	4
	乙腈至流动池	15
	N2 系统冲洗	4
	乙腈系统冲洗	4
	乙腈至流动池	23
	N2 系统冲洗	4
	乙腈系统冲洗	4
	乙腈至流动池	15
脱封闭(脱封闭流)	脱封闭至流动池	36
	乙腈系统冲洗	4
洗涤(乙腈洗涤流)	N2 系统冲洗	4
	乙腈系统冲洗	4
	乙腈至流动池	18
	N2 系统冲洗	4.13
	乙腈系统冲洗	4.13
	乙腈至流动池	15
	乙腈系统冲洗	4

[0249] 亚磷酸胺/活化剂组合递送与通过流动池递送主要试剂相似。没有执行干燥步骤，因为环境始终保持用试剂“湿润”。

[0250] 将限流器从ABI 394合成仪中移除，以实现更快的流动。没有限流器，亚磷酸胺(amidites) (0.1M于ACN中)、活化剂(ACN中的0.25M苯甲酰硫四唑(“BTT”; 30-3070-xx来自GlenResearch)) 和Ox (20%吡啶、10%水以及70%THF中的0.02M I₂) 的流速大约为~100μL/sec;对于乙腈(“ACN”) 和加帽试剂(CapA和CapB的1:1混合物,其中CapA是THF/吡啶中的乙酸酐,CapB是THF中的16%1-甲基咪唑),约为~200μL/sec;以及对于脱封闭剂(甲苯中的3%二氯乙酸),约为~300μL/sec (与具有限流器时所有试剂的~50μL/sec相比)。遵守完全推出氧化剂的时间,相应地调整化学物质流动时间的定时,并在不同的化学物质之间引入

额外的ACN洗涤。多核苷酸合成后,使芯片在75psi的气态氨中脱保护过夜。向表面施加五滴水以回收多核苷酸。然后在BioAnalyzer小RNA芯片上分析回收的多核苷酸(数据未显示)。

[0251] 实施例3:在多核苷酸合成装置上合成100-mer序列

[0252] 与实施例2中描述的合成50-mer序列相同过程被用于在两个不同的硅芯片上合成100-mer多核苷酸(“100-mer”;5’CGGGATCCTTATCGTCATCGTCGTACAGATCCCCGACCCATTTGCTGTCACCCAGTCATGCTAGCCATACCATGATGATGATGATGATGAGAACCCCGCAT##TTTTTTTTTT3’,其中#表示胸苷-琥珀酰己酰胺CED亚磷酰胺(来自ChemGenes的CLP-2244);SEQ ID NO.:2),第一个硅芯片用N-(3-三乙氧基甲硅烷基丙基)-4-羟基丁酰胺均匀官能化,并且第二个硅芯片用11-乙酰氧基十一烷基三乙氧基硅烷和正癸基三乙氧基硅烷的5/95混合物官能化,以及从表面提取的多核苷酸在BioAnalyzer仪器上分析(数据未显示)。

[0253] 来自两个芯片的所有10个样品都使用正向引物(5’ATGCGGGTTCTCATCATC3’;SEQ ID NO.:3)和反向引物(5’CGGGATCCTTATCGTCATCG3’,SEQ ID NO.4)在50μLPCR混合物(25μL NEB Q5 Master Mix、2.5μL 10μM正向引物、2.5μL 10μM反向引物、从表面提取的1μL多核苷酸和水至高达50μL)中进一步PCR扩增,使用以下热循环程序:

[0254] 98°C,30sec

[0255] 98°C,10sec;63°C,10sec;72°C,10sec;重复12个循环

[0256] 72°C,2min

[0257] PCR产物也在BioAnalyzer上运行(数据未示出),证明在100-mer位置有尖锐的峰。然后,克隆PCR扩增的样品,并Sanger测序。表2概括了从芯片1的点(spot)1-5采集的样品和从芯片2的点6-10采集的样品的Sanger测序结果。

[0258] 表2

[0259]

点	错误率	循环效率
1	1/763bp	99.87%
2	1/824bp	99.88%
3	1/780bp	99.87%
4	1/429bp	99.77%
5	1/1525bp	99.93%
6	1/1615bp	99.94%
7	1/531bp	99.81%
8	1/1769bp	99.94%
9	1/854bp	99.88%
10	1/1451bp	99.93%

[0260] 因此,合成的多核苷酸的高质量 and 一致性在具有不同表面化学性质的两个芯片上重复。总体上,89%,对应于测序的100-mer中262个中的233个是没有错误的完美序列。

[0261] 最后,表3总结了从来自点1-10的多核苷酸样品获得的序列的错误特征。

[0262] 表3

[0263]

样品 ID/点 no.	OSA_00 46/1	OSA_00 47/2	OSA_00 48/3	OSA_00 49/4	OSA_00 50/5	OSA_00 51/6	OSA_00 52/7	OSA_00 53/8	OSA_00 54/9	OSA_005 5/10
总序列	32	32	32	32	32	32	32	32	32	32
测序质量	28 个中 25 个	27 个中 27 个	30 个中 26 个	23 个中 21 个	26 个中 25 个	30 个中 29 个	31 个中 27 个	31 个中 29 个	29 个中 28 个	28 个中 25 个
寡核苷酸质量	25 个中 23 个	27 个中 25 个	26 个中 22 个	21 个中 18 个	25 个中 24 个	29 个中 25 个	27 个中 22 个	29 个中 28 个	28 个中 26 个	25 个中 20 个
ROI 匹配计数	2500	2698	2561	2122	2499	2666	2625	2899	2798	2348
ROI 突变	2	2	1	3	1	0	2	1	2	1
ROI 多碱基缺失	0	0	0	0	0	0	0	0	0	0
ROI 小插入片段	1	0	0	0	0	0	0	0	0	0
ROI 单个碱基缺失	0	0	0	0	0	0	0	0	0	0
大量缺失计数	0	0	1	0	0	1	1	0	0	0
突变: G>A	2	2	1	2	1	0	2	1	2	1
突变: T>C	0	0	0	1	0	0	0	0	0	0
ROI 错误计数	3	2	2	3	1	1	3	1	2	1
ROI 错误率	错误: ~ 834 个中	错误: ~ 1350 个	错误: ~ 1282 个	错误: ~ 708 个中	错误: ~ 2500 个	错误: ~ 2667 个	错误: ~ 876 个中	错误: ~ 2900 个	错误: ~ 1400 个	错误: ~ 2349 个

[0264]

	1 个	中 1 个	中 1 个	1 个	中 1 个	中 1 个	1 个	中 1 个	中 1 个	中 1 个
ROI 负引物错误率	MP 错误: ~ 763 个中 1 个	MP 错误: ~ 824 个中 1 个	MP 错误: ~ 780 个中 1 个	MP 错误: ~ 429 个中 1 个	MP 错误: ~ 1525 个 中 1 个	MP 错误: ~ 1615 个 中 1 个	MP 错误: ~ 531 个中 1 个	MP 错误: ~ 1769 个 中 1 个	MP 错误: ~ ~854 个 中 1 个	MP 错误: ~ 1451 个 中 1 个

[0265] 实施例4:29,040个独特多核苷酸的并行组装

[0266] 如图10中所示,制备了在平坦硅板1001上包括256个簇1005的结构,每个簇1005包括121个位点。在1010中示出了具有121个位点的簇的扩展视图。来自256个簇中的240个簇的位点为具有不同序列的多核苷酸的合成提供了附接和支持。使用实施例3的一般方法,通过亚磷酰胺化学进行多核苷酸合成。256个簇中的16个簇的位点是对照簇。合成的29,040个独特多核苷酸(240×121)的总体分布在图11A中示出。多核苷酸文库以高一致性合成。90%的序列以平均值的4倍内的信号存在,允许100%的表现度。如图11B中所示,测量每个簇的分布。在4个代表性簇中合成的独特多核苷酸的分布在图12中示出。在总体水平上,运行中的所有多核苷酸都存在,并且99%的多核苷酸具有在指示合成一致性的平均值的2x内的丰度。相同的观察结果在每个簇水平上是一致的。

[0267] 使用Illumina MiSeq基因测序仪确定每种多核苷酸的错误率。29,040种独特多核苷酸的错误率分布在图13A中示出,并且平均约500个碱基中1个碱基,且一些错误率低至800个碱基中1个碱基。如图13B中所示,测量了每个簇的分布。四个代表性簇中独特多核苷酸的错误率分布在图14中示出。在少于20小时内合成了29,040种独特多核苷酸的文库。

[0268] 对遍及所有29,040种独特多核苷酸的GC百分比对多核苷酸表现度的分析表明,不管GC含量如何,合成是一致的,图15。

[0269] 实施例5:合成的多核苷酸文库的PCR扩增

[0270] 实施例3中描述了在具有相似布置的结构上设计和合成的9,996种多核苷酸,每100个碱基的长度的随机序列具有不同的GC含量,为20-80%GC。为了确定PCR扩增对GC表现度的影响,用高保真性的DNA聚合酶(DNA聚合酶1)扩增多核苷酸群体6或20个循环。或者,使用两种其他的高保真PCR酶扩增多核苷酸群体6、8、10或15个循环,以确定聚合酶选择是否对扩增后的总体序列表现度具有影响。在PCR扩增后,样品被准备用于下一代测序,并在Illumina MiSeq平台上测序。产生了150bp的SE读段,读段深度约为100x。分析了原始FASTQ文件。用任一聚合酶进行6、10或15个循环的多核苷酸表现度在图16中描绘。评估了不同条件下多核苷酸表现度的一致性,并在表4中总结。

[0271] 表4

[0272]

	循环数	在 1.5x 内的%	在 2x 内的%
聚合酶 1	6	72.1%	92.6%
	8	76.1%	90.3%
	10	70.9%	86.6%
	15	64.1%	82.7%
聚合酶 2	6	91.9%	98.9%
	8	89.9%	98.1%
	10	90.1%	98.4%
	15	89.2%	97.9%

[0273] 每个扩增的多核苷酸群体的失落的数量被量化,如图15中所示,扩增循环对低于平均阈值的10%的群体的分数。聚合酶1失落增长迅速,而聚合酶2失落保持相对恒定。

[0274] 评估了过度扩增对GC分布的影响通常,图18。通常,GC含量30%至70%的多核苷酸遵循趋势线, $Y=X$,并且频率随着更多循环而增加。GC含量大于70%的多核苷酸通常在20个循环后稍微更频繁,而GC含量低于30%的多核苷酸通常在6个循环后稍微更频繁。

[0275] 实施例6.全平板扩增与并行多核苷酸簇扩增的多核苷酸表现度的比较

[0276] 多核苷酸在如图10中所示制造的平坦硅板上的包括256个簇的结构上合成,每个簇包括121个位点。使用实施例3的一般方法,通过亚磷酰胺化学进行多核苷酸合成。该结构上的多核苷酸被裂解并合并。

[0277] 将整个板的多核苷酸合并并扩增。扩增后,如在图19中所见的,与线中所见的平均值相比,存在显著的GC偏倚和方差。因此,需要更多的测序,并且存在更多的失落。

[0278] 来自簇扩增的多核苷酸的分布见图20中。在运行1和运行2中,平均值(线)的频率分布约为8,并且平均值的方差约为 $1.7x$ 。GC百分比在17%和94%的范围内。图23和图20说明了存在可再现性,并且多核苷酸群体显示出GC偏倚的显著降低(图20)。此外,失落为零,并需要少30%的测序。

[0279] 实施例7.具有不同GC含量的合成的多核苷酸文库

[0280] 预选包含约15%至约85%的GC含量的13,000种多核苷酸序列的文库用于合成(图22)。在结构上合成第一多核苷酸文库,并且使用实施例3的一般方法通过亚磷酰胺化学进行合成。裂解并合并结构上的多核苷酸,随后扩增以产生PCR偏倚的多核苷酸文库。根据GC含量对文库中的多核苷酸序列进行分箱,并调整每个箱的化学计量,以考虑由PCR扩增产生的观察到的GC偏倚。例如,含有较高或较低GC含量的多核苷酸具有较高的初始浓度,这导致扩增后一致性的化学计量表现度。这有效地减少或消除了来自扩增步骤的PCR GC偏倚。在结构上合成第二多核苷酸文库,并且使用实施例3的一般方法通过亚磷酰胺化学进行合成。裂解并合并结构上的多核苷酸,随后扩增,以产生扩增后具有一致性的GC表现度(图21A)的高度一致性的多核苷酸文库(图23)。GC平衡的文库的一个优点是,对于期望的取样覆盖,它需要较少的取样。例如,对于单分散文库,80%和90%文库覆盖的取样率接近理论最小值(图24A和图24B)。还合成了偏向高和低GC含量两者的不同程度的多核苷酸文库(分别为图21B和图21C)。还合成了偏向低GC含量(图21D)或高GC含量(图21E)的多核苷酸文库。

[0281] 实施例8.合成的具有80-mer和120-mer多核苷酸长度的GC平衡的多核苷酸文库

[0282] 使用实施例7的一般方法设计并GC平衡包含约20,000种每种长度为80个核酸的独特多核苷酸的文库,并在结构上合成;使用实施例3的一般方法通过亚磷酰胺化学进行合成。还合成了含有每种长度为120个核酸的多核苷酸的相似文库(图26)。两个文库都显示出高度一致性的分布,且识别出>99%的独特序列。这些文库在GC含量的方差之间也表现出一致性,在重复之间具有高度一致性(图27),在经受噪声的尾部具有低数量的多核苷酸。

[0283] 实施例9.重复扩增多核苷酸文库后的GC含量评估

[0284] 使用实施例3的一般方法,通过亚磷酰胺化学在结构上合成由9,996种独特多核苷酸组成的多核苷酸文库,所述独特多核苷酸含有20-80%的GC含量,每种长度为100个碱基。用两种不同的高保真性的DNA聚合酶将文库扩增8或15个PCR循环,并比较这两种条件之间的群体中的多核苷酸的频率(图28)。一致性线(黑色虚线)表示在8或15个PCR循环后群体中具有相同频率的多核苷酸。相对于8个循环的扩增,在15个循环后,高于一致性线的序列是群体中表现度过多的,而在15个循环后,低于一致性线的序列是群体中表现度不足的。在这种情况下,聚合酶1随着PCR循环的增加表现出GC偏倚。与8个循环相比,在15个循环的扩增后,观察到高GC序列(GC多于70%,中等灰色)表现度过多,而低GC序列(GC低于30%,以最深灰色显示)表现度不足。此外,在相似的GC百分比内的富集的大幅度变化表明,除了GC含量之外的因素,诸如发夹形成或均聚物链段,可以影响扩增偏倚。聚合酶2未表现出相同的序列表现度偏倚。

[0285] 实施例10.重复扩增多核苷酸文库后的失落和表现度评估

[0286] 研究了通过用不同的DNA聚合酶扩增引入的偏倚。将实施例9的多核苷酸文库用DNA聚合酶1扩增(图29,深色线)或用DNA聚合酶2扩增(图29,浅色线)6、8、10或15个循环。增加的PCR循环与增加的多核苷酸序列失落频率相关,其中失落频率被定义为具有小于平均值的10%的丰度的序列。这种影响的程度取决于用于扩增的DNA聚合酶。与用DNA聚合酶2扩增相比,用DNA聚合酶1扩增后更大比例的序列失落(在15个循环的情况下失落多约20倍)。取决于GC含量、长度和序列复杂性,对于扩增不同文库序列,不同的聚合酶可能是最佳的。

[0287] 更详细地研究了用每种DNA聚合酶扩增15个PCR循环的文库,以评估多核苷酸序列

的表现度(图30)。用DNA聚合酶1扩增多核苷酸文库,导致在15个PCR循环后多20倍的序列失落(图29)。用DNA聚合酶1扩增的多核苷酸的分布大于用DNA聚合酶2扩增的多核苷酸的分布。用DNA聚合酶1扩增的文库的多核苷酸分布具有64%的序列存在于平均值的1.5倍内。当用DNA聚合酶2扩增同一个文库时,>89%的序列存在于平均值的1.5倍内,这表明DNA聚合酶2扩增文库的偏倚比DNA聚合酶1低得多。在用DNA聚合酶1扩增的文库中引入的偏倚增加了覆盖多核苷酸文库所需的筛选努力。

[0288] 实施例11.受控化学计量的多核苷酸文库用于用下一代测序(NGS)靶向外显子组的用途

[0289] 使用实施例3的一般方法,通过亚磷酰胺化学在结构上设计和合成第一多核苷酸cDNA靶向文库(探针文库),其包含高达370,000种或更多种与一个或多个基因外显子重叠的非相同的多核苷酸。使用PCR(或直接在固相合成期间)将多核苷酸连接到分子标签诸如生物素上,以形成用于随后捕获感兴趣的靶外显子的探针。将探针与基因组核酸文库中的序列杂交,并与非结合序列分离。未结合的探针被洗掉,留下富集cDNA序列的靶文库。然后使用NGS对富集的文库测序,并根据用于靶向基因的cDNA探针测量每种预期基因的读段。

[0290] 在一些情况下,靶序列的读段频率受到靶序列丰度、探针结合、二级结构或在靶序列测序后尽管富集还是降低表现度的其他因素的影响表现度。通过修改第一多核苷酸cDNA靶向文库的化学计量来进行多核苷酸文库化学计量控制,以获得第二多核苷酸cDNA靶向文库,且导致较少的读段的多核苷酸探针序列的化学计量增加。使用实施例3的一般方法,通过亚磷酰胺化学在结构上设计和合成第二cDNA靶向文库,并如前所述用于富集靶基因组DNA文库的序列外显子。

[0291] 实施例12.对外显子组探针文库进行化学计量控制的多次迭代

[0292] 使用实施例11的一般方法合成并测试外显子组探针文库。进行化学计量修改的多次迭代,产生受控的化学计量探针文库,文库1。与数个比较物外显子组富集试剂盒相比,这导致获得期望的靶覆盖测序读段显著更少。为了准确测序,至少90%的靶外显子组碱基的30x读段深度是理想的,并且通常需要过度测序(理论读段深度,多于30x读段深度)来补偿一致性问题。受控化学计量的外显子组探针文库能够实现55x理论读段深度情况下90%的靶碱基的30x读段深度(图32A),这比另一个比较物外显子组富集试剂盒所需的测序覆盖显著更少,并且测序通量(每次运行的样品,表5)更快。当针对4.5Gb的测序归一化时,受控化学计量的探针文库提供了>95%的所有靶碱基的10x读段深度,并且这显著高于比较的所有其他比较物外显子组探针试剂盒(图32B)。

[0293] 表5

[0294]

需要的平均测序覆盖	每次运行的样品
比较物外显子组富集试剂盒A	4
受控化学计量的外显子组探针(文库1)	17

[0295] 实施例13:杂交组(panel)的生产

[0296] 使用实施例11的一般方法制备多核苷酸靶向文库,其靶向特定基因、疾病、组的组合或定制外显子组。反应大小的量级跨越 10^3 ,并且探针组大小的范围为从约80至约900,000个探针(图33)。

[0297] 实施例14:产生70,000种探针的组

[0298] 使用实施例3的一般方法通过亚磷酰胺化学在结构上设计和合成包含70,000种非相同的多核苷酸的多核苷酸靶向文库(探针文库),并使用实施例11的一般方法进行GC控制以产生文库2。测序后的读段分布在图34中示出,并且GC分箱的靶覆盖在图35A和图35B中示出。

[0299] 实施例15:产生2,544种探针的组

[0300] 使用实施例3的一般方法,通过亚磷酰胺化学在结构上设计和合成包含2,544种非相同的多核苷酸的多核苷酸靶向文库(探针文库),并且使用实施例11的一般方法控制化学计量以产生文库3。中靶率在图36A中示出,并且覆盖率在图36B中示出。利用文库3进行靶富集导致了中靶率和覆盖率均高于基于比较物阵列的试剂盒#2。

[0301] 实施例16:样品制备和用多核苷酸靶向文库的富集

[0302] 基因组DNA(gDNA)是从样品中获得的,并在片段化缓冲液中酶促片段化、末端修复及3'腺苷酸化。将衔接子连接到基因组DNA片段的两端,以产生衔接子标记的gDNA链文库,并且将衔接子标记的DNA文库用高保真性聚合酶扩增。然后,在衔接子封闭剂的存在下,使gDNA文库在96°C变性成单链。使多核苷酸靶向文库(探针文库)在96°C的杂交溶液中变性,并在70°C的杂交溶液中与变性的标记的gDNA文库结合16小时。然后将结合缓冲液加入杂交的标记的gDNA-探针中,并使用包含链霉亲和素的磁珠捕获生物素化的探针。使用磁体将珠从溶液中分离,并在加入洗脱缓冲液以从珠释放富集的、标记的gDNA片段之前,将珠用缓冲液洗涤三次以去除未结合的衔接子、gDNA和衔接子封闭剂。用高保真性聚合酶扩增富集的标记的gDNA片段文库,以获得足够用于簇产生的产量,并且然后使用NGS仪器对文库测序。

[0303] 实施例17:一般样品制备和用多核苷酸靶向文库的富集

[0304] 从样品获得多于一种多核苷酸,并将其片段化、任选地末端修复和腺苷酸化。将衔接子连接到多核苷酸片段的两端以产生衔接子标记的多核苷酸链文库,并且将衔接子标记的多核苷酸文库扩增。然后使衔接子标记的多核苷酸文库在衔接子封闭剂的存在下在高温变性,优选96°C。使多核苷酸靶向文库(探针文库)在杂交溶液中进行高温变性,优选约90°C至99°C,并在杂交溶液中进行约45°C至80°C与变性的标记的多核苷酸文库结合约10至24小时。然后将结合缓冲液添加到杂交的标记的多核苷酸探针中,并且使用包含捕获部分的固体支持物来选择性地结合杂交的衔接子标记的多核苷酸-探针。在添加洗脱缓冲液以从固体支持物释放富集的衔接子标记的多核苷酸片段之前,用缓冲液洗涤固体支持物一次或更多次,优选地约2次和5次,以去除未结合的多核苷酸。扩增富集的衔接子标记的多核苷酸片段文库,然后对文库测序。

[0305] 实施例18:用多核苷酸靶向文库标记前的一般富集

[0306] 从样品获得多于一种多核苷酸,片段化并任选地末端修复。然后使片段化的多核苷酸样品在高温变性,优选96°C。使多核苷酸靶向文库(探针文库)在杂交溶液中进行高温变性,优选地约90°C至99°C,并在杂交溶液中进行约45°C至80°C与变性的多核苷酸文库结合约10至24小时。然后将结合缓冲液添加到杂交的多核苷酸探针中,并且使用包含捕获部分的固体支持物来选择性地结合杂交的片段化的多核苷酸探针。在加入洗脱缓冲液以从固体支持物释放富集的多核苷酸片段之前,用缓冲液洗涤固体支持物一次或更多次,优选地约2至5次,以去除未结合的多核苷酸。将富集的多核苷酸腺苷酸化,将衔接子连接到多核苷酸的

两端以产生富集的衔接子标记的多核苷酸链文库,并且将衔接子标记的多核苷酸文库扩增。然后对富集的衔接子标记的多核苷酸片段文库测序。

[0307] 实施例19:一般样品制备和用多核苷酸靶向文库的过滤

[0308] 从样品获得多于一种多核苷酸,并将其片段化、任选地末端修复和腺苷酸化。将衔接子连接到多核苷酸片段的两端以产生衔接子标记的多核苷酸链文库,并且将衔接子标记的多核苷酸文库扩增。然后使衔接子标记的多核苷酸文库在衔接子封闭剂的存在下在高温变性,优选96℃。使设计成去除不期望的非靶序列的多核苷酸过滤文库(探针文库)在杂交溶液中在高温变性,优选地约90℃至99℃,并在杂交溶液中在约45℃至80℃与变性的标记的多核苷酸文库结合约10至24小时。然后将结合缓冲液添加到杂交的标记的多核苷酸探针中,并且使用包含捕获部分的固体支持物来选择性地结合杂交的衔接子标记的多核苷酸探针。用缓冲液洗涤固体支持物一次或更多次,优选地约1至5次,以洗脱靶衔接子标记的多核苷酸片段。扩增富集的靶衔接子标记的多核苷酸片段文库,并且然后对文库测序。

[0309] 实施例20:160-mer探针文库的制备

[0310] 在结构上合成包含至少1,000种探针的文库,并且使用实施例3的一般方法通过亚磷酰胺化学进行合成。每个探针是双链的,并且探针的每条链包含与靶互补的120个核苷酸的靶结合序列。每个探针还包含20个核苷酸的正向引发位点和20个核苷酸的反向引发位点。将探针的每条链在5'位置用两个生物素分子标记。

[0311] 实施例21:包含非靶结合序列的210-mer探针文库的制备

[0312] 在结构上合成包含至少1,000种探针的文库,并且使用实施例3的一般方法通过亚磷酰胺化学进行合成。每个探针是双链的,并且探针的每条链包含与靶互补的120个核苷酸的靶结合序列。每个探针还包含20个核苷酸的正向引物结合位点、20个核苷酸的反向引物结合位点、包含聚腺苷的25个核苷酸的5'非靶结合序列和包含聚腺苷的25个核苷酸的3'非靶结合序列。将探针的每条链在5'位置用两个生物素分子标记。

[0313] 实施例22:靶向人类HLA的外显子1的210-mer探针

[0314] 在结构上合成包含至少1,000种探针的文库,并且使用实施例3的一般方法通过亚磷酰胺化学进行合成。每个探针是双链的,并且探针的每条链包含与人类HLA的外显子1的区域互补的120个核苷酸的靶结合序列。每个探针还包含20个核苷酸的正向引物结合位点、20个核苷酸的反向引物结合位点、包含聚腺苷的25个核苷酸的5'非靶结合序列和包含聚腺苷的25个核苷酸的3'非靶结合序列。将探针的每条链在5'位置用两个生物素分子标记。

[0315] 实施例22:非重叠探针文库的设计方法

[0316] 提供了至少100种靶序列,并基于与互补性探针靶结合序列的期望长度相比的长度,将其分类为离散的类别。例如,类别包括但不限于(a)比插入片段长度短的靶,(b)比插入片段长度+X短或等于插入片段长度+X的靶,以及(c)比插入片段长度+X长的靶,其中X是探针不靶向的期望的空位长度。类别(a)中的靶序列被用靶结合序列靶向,所述靶结合序列在靶序列的中心或在靶序列的左侧或右侧对齐,这取决于插入片段也将与其互补的非靶区域的复杂性(重复、高/低GC、回文序列等)。以与类别(a)相同的方式靶向类别(b)中的靶序列,其中X是靶结合序列不靶向的期望的空位长度。对于类别(c)中的靶,将靶的总长度除以靶结合序列的长度,并向上取整到最接近的整数值,该整数值表示完全靶向所有靶序列以产生靶序列的插入片段集所需的靶结合序列的数量。任选地,可以减少靶结合序列的数量,

其中在减少后,靶结合序列之间的空位小于期望的空位长度Y。然后对每个靶序列重复这整个过程,形成插入片段文库。然后通过添加包含一个或更多个引发序列的一个或更多个非靶序列来修改文库中的靶结合序列,在结构上合成文库,使用实施例3的一般方法通过亚磷酰胺化学进行合成,并用分子标签标记探针。

[0317] 实施例23:重叠探针文库的设计方法

[0318] 提供了至少100种靶序列,并基于与互补性探针靶结合序列的期望的长度相比的长度,将其分类为离散的类别。例如,类别包括但不限于(a)比插入片段长度短的靶,(b)比插入片段长度+X短或等于插入片段长度+X的靶,以及(c)比插入片段长度+X长的靶,其中X是探针不靶向的期望的空位长度。类别(a)中的靶序列被用靶结合序列靶向,所述靶结合序列在靶序列的中心或在靶序列的左侧或右侧对齐,这取决于插入片段也将与其互补的非靶区域的复杂性(重复、高/低GC、回文序列等)。以与类别(a)相同的方式靶向类别(b)中的靶序列,其中X是靶结合序列不靶向的期望的空位长度。对于类别(c)中的靶,将靶的总长度除以靶结合序列的长度,并向上取整到最接近的整数值,该整数值表示完全靶向所有靶序列所需的靶结合序列的数量。然后将互补性靶结合序列在靶序列上间隔开(任选地均匀间隔开),允许重叠以产生靶序列的插入片段集。然后对每个靶序列重复这整个过程,形成插入片段文库。然后通过添加包含一个或更多个引发序列的一个或更多个非靶序列来修改文库中的靶结合序列,在结构上合成文库,使用实施例3的一般方法通过亚磷酰胺化学进行合成,并用分子标签标记探针。

[0319] 实施例24:混合探针文库的设计方法

[0320] 探针文库按照具有修改的实施例22和实施例23的一般方法合成。为每个靶序列产生包括非重叠插入片段、重叠插入片段或混合(重叠和非重叠)插入片段的集。

[0321] 实施例25:用于外显子靶向的多核苷酸探针

[0322] 多核苷酸探针可以靶向基因组中的外显子,且基因可以包含多于一个外显子。例如,人类白细胞抗原(HLA)基因包含七个外显子,其中三个在表6中列出。

[0323] 表6

[0324]

Seq ID	描述	外显子长度 (bp)	包含外显子(下划线)的基因组序列
5	人类 HLA 基因 外显子 1	63	CCCACCGGGACTCAGATTCTCCCCAGACGCCGAGGAT GGTGCTC <u>ATGGCGCCCCGAACCCTCCTCCTGCTGCTC</u> <u>TCAGGGGGCCCTGGCCCTGACCCAGACCTGGGCGCGT</u> GAGTGCAGGGTCTGCAGGGAAATGGGCGCGTGAGTG CAGGGTCTGC

[0325]

6	人类 HLA 基因 外显子 2	270	<u>GCTCCAGGTTCCCACTCCATGAGGTATTTCTACACC</u> <u>ACCATGTCCCGGCCCGGCGCCGGGAGCCCCGCTTCA</u> <u>TCTCCGTCGGCTACGTGGACGATACGCAGTTCGTGCG</u> <u>GTTCGACAGCGACGACGCGAGTCCGAGAGAGGAGCC</u> <u>GCGGGCGCCGTGGATGGAGCGGGAGGGGCCAAAGTA</u> <u>TTGGGACCGGAACACACAGATCTGCAAGGCCCAGGC</u> <u>ACAGACTGAACGAGAGAACCTGCGGATCGCGCTCCG</u> <u>CTACTACAACCAGAGCGAGGGCGGTGAGTTGACCCC</u> <u>GG</u>
7	人类 HLA 基因 外显子 5	117	<u>TAGCAGGGTCAGGGTTCCCTCACCTTCCCCCTTTTCC</u> <u>CAGCCATCTTCCCAGCCCACCGTCCCCATCGTGGGCA</u> <u>TCGTTGCTGGCTTGGTTCTACTTGTAGCTGTGGTCACT</u> <u>GGAGCTGTGGTTCGCTGCTGTAATGTGGAGGAAGAAG</u> <u>AGCTCAGGTAAGGAAGGGGT</u>

[0326] 对于特定外显子,靶结合序列和非靶结合序列的多种组合可以用于设计多种构型和长度的探针。靶向HLA外显子1的非限制性探针设计如表7中的实例所示,并且仅示出了探针的一条链的序列。表8中列出了非靶结合序列、靶结合序列和总探针长度的大小。

[0327] 表7

[0328]

Seq ID	靶	5'非靶 结合序列	靶结合序列	3'非靶 结合序列
8	人类 HLA 外显子1	<u>GTTACCCA</u> <u>AGAACGCA</u> GCTGATTC TCCCCAGA CGCCGAGG ATGGTGCT C	ATGGCGCCCCGAACCTCCTCCTGC TGCTCTCAGGGGCCCTGGCCCTGAC CCAGACCTGGGCG	CGTGAGTGCAG GGTCTGCAGGG AAATGGTAGTG <u>TCGGAGGTCGT</u> <u>TCCT</u>
9	人类 HLA	<u>GTTACCCA</u> <u>AGAACGCA</u>	GGTTCCCACTCCATGAGGTATTTCT ACACCACCATGTCCCGGCCCGGCG	<u>TAGTGTCCGAG</u> <u>GTCGTTCCCT</u>

[0329]

	外显子2	<u>GCTG</u>	CCGGGGAGCCCCGCTTCATCTCCGT CGGCTACGTGGACGATACGCAGTT CGTGCGGTTTCGACAGCGACGAC	
10	人类 HLA 外显子2	<u>GTTACCCA</u> <u>AGAACGCA</u> <u>GCTG</u>	TGGATGGAGCGGGAGGGGCCAAAG TATTGGGACCGGAACACACAGATC TGCAAGGCCCAGGCACAGACTGAA CGAGAGAACCTGCGGATCGCGCTC CGCTACTACAACCAGAGCGAGGGC	<u>TAGTGTCGGAG</u> <u>GTCGTTTCCT</u>
11	人类 HLA 外显子2	<u>GTTACCCA</u> <u>AGAACGCA</u> <u>GCTG</u>	GGCTACGTGGACGATACGCAGTTC GTGCGGTTTCGACAGCGACGACGCG AGTCCGAGAGAGGAGCCGCGGGCG CCGTGGATGGAGCGGGAGGGGCCA AAGTATTGGGACCGGAACACACAG	<u>TAGTGTCGGAG</u> <u>GTCGTTTCCT</u>
12	人类 HLA 外显子1	<u>GTTACCCA</u> <u>AGAACGCA</u> <u>GCTG</u>	ATGGCGCCCCGAACCTCCTCCTGC TGCTCTCAGGGGGCCCTGGCCCTGAC CCAGACCTGGGCG	<u>TAGTGTCGGAG</u> <u>GTCGTTTCCT</u>
13	人类 HLA 外显子2	AAAAAAAA AAAAAAAA AAAAAAAA <u>AGTTACCC</u> <u>AAGAACGC</u> <u>AGCTG</u>	GGTTCCTCCATGAGGTATTTCT ACACCACCATGTCCCGGCCCGGCG CCGGGGAGCCCCGCTTCATCTCCGT CGGCTACGTGGACGATACGCAGTT CGTGCGGTTTCGACAGCGACGAC	<u>TAGTGTCGGAG</u> <u>GTCGTTTCCTAAA</u> AAAAAAAAAAAA AAAAAAAAAAAA
14	人类 HLA 外显子2	<u>GTTACCCA</u> <u>AGAACGCA</u> <u>GCTGAAAA</u> AAAAAAAA AAAAAAAA AAAAA	GGTTCCTCCATGAGGTATTTCT ACACCACCATGTCCCGGCCCGGCG CCGGGGAGCCCCGCTTCATCTCCGT CGGCTACGTGGACGATACGCAGTT CGTGCGGTTTCGACAGCGACGAC	AAAAAAAAAAAA AAAAAAAAAAAA AAAT <u>TAGTGTCG</u> <u>GAGGTCGTTTCCT</u>
15	人类 HLA 外显子5	<u>GTTACCCA</u> <u>AGAACGCA</u> <u>GCTG</u>	AGCCATCTTCCCAGCCCACCGTCCC CATCGTGGGCATCGTTGCTGGCTTG GTTCTACTTGTA	<u>TAGTGTCGGAG</u> <u>GTCGTTTCCT</u>
16	人类 HLA 外显子5	<u>GTTACCCA</u> <u>AGAACGCA</u> <u>GCTG</u>	GCTGTGGTCACTGGAGCTGTGGTCG CTGCTGTAATGTGGAGGAAGAAGA GCTCAG	<u>TAGTGTCGGAG</u> <u>GTCGTTTCCT</u>

[0330] 非靶结合序列的引发序列加下划线。

[0331] 表8

[0332]

Seq ID	靶	5'非靶结合序列长度(bp)	靶结合序列长度(bp)	3'非靶结合序列长度(bp)	总探针长度长度(bp)
8	人类 HLA 外显子 1	49	63	48	160
9	人类 HLA 外显子 2	20	120	20	160
10	人类 HLA 外显子 2	20	120	20	160
11	人类 HLA 外显子 2	20	120	20	160
12	人类 HLA 外显子 1	20	63	20	103
13	人类 HLA	45	120	45	210

[0333]

	外显子 2				
14	人类 HLA 外显子 2	45	120	45	210
15	人类 HLA 外显子 5	20	62	20	102
16	人类 HLA 外显子 5	20	55	20	95

[0334] 多于一种探针的多种布置可以用于覆盖特定外显子,例如人类HLA外显子2。包含SEQ ID:9或10的探针包含集并靶向人类HLA外显子2,但一起在靶外显子中留下一个空位,并且不包含重叠序列(参见图4F)。包含SEQ ID:11的探针靶向HLA外显子2,并包含与SEQ ID:9和10重叠的靶结合序列(参见图4G)。各自包含SEQ ID:15或16的探针包含集,靶向人类HLA外显子5,并且不靶向靶或非靶区域的任何重叠区域。对应于SEQ ID 8-16的探针在结构上合成,并且使用实施例3的一般方法通过亚磷酰胺化学进行合成,并且任选地用至少一个分子标签诸如生物素来标记。

[0335] 实施例26:用多核苷酸探针文库捕获基因组DNA

[0336] 使用实施例3的一般方法,通过亚磷酰胺化学在结构上设计和合成包含至少500,000种靶向人类外显子的非相同的多核苷酸的多核苷酸靶向文库,并且使用实施例11的一般方法控制化学计量以产生文库4。然后用生物素标记多核苷酸,并且然后溶解以形成外显子组探针文库溶液。使用实施例16的一般方法,从基因组DNA(gDNA)样品获得干燥的索引文库汇集。

[0337] 通过脉冲涡旋2sec混合外显子组探针文库溶液、杂交溶液、封闭剂混合物A和封闭剂混合物B。将杂交溶液在65℃加热10分钟,或者直到所有沉淀物溶解,并且然后在工作台上持续额外5分钟恢复至室温。将20μL杂交溶液和4μL外显子组探针文库溶液加入到薄壁PCR 0.2mL条管中,并通过温和移液混合。在具有105℃盖子的热循环仪中将混合的杂交溶液/外显子组探针溶液加热至95℃持续2分钟,并立即在冰上冷却至少10分钟。然后允许溶液在工作台上持续5分钟冷却至室温。当杂交溶液/外显子组探针文库溶液冷却时,将水加入到9μL的每个基因组DNA样品中,并将5μL封闭剂混合物A和2μL封闭剂混合物B加入到薄壁PCR 0.2mL条管中的干燥索引文库汇集物中。然后通过温和移液混合溶液。在具有105℃盖

子的热循环仪中,将汇集的文库/封闭剂管在95℃加热5分钟,然后在进行下一步之前,在工作台上持续不多于5分钟恢复至室温。通过移液混合杂交混合物/探针溶液,并将其添加到整个24μL的汇集文库/封闭剂管中。通过温和移液将整个捕获反应孔混合,以避免产生气泡。将样品管脉冲旋转,确保管密封紧密。使捕获/杂交反应在具有85℃的盖子温的PCR热循环器中于70℃加热16小时。

[0338] 将结合缓冲液、洗涤缓冲液1和洗涤缓冲液2在48℃加热,直至所有沉淀物溶解到溶液中。每个捕获等分700μL洗涤缓冲液2,并预热至48℃。将链霉亲和素结合珠和DNA纯化珠在室温平衡至少30分钟。将聚合酶,诸如KAPA HiFi HotStart ReadyMix和扩增引物在冰上融解。试剂融解后,将其通过脉冲涡旋2sec混合。每个捕获反应制备500μL的80%乙醇。将链霉亲和素结合珠在室温预平衡,并涡旋直至均质化。每个捕获反应,将100μL链霉亲和素结合珠加入清洁的1.5mL微量离心管中。向每个管中加入200μL的结合缓冲液,并通过移液混合每个管,直至均质化。将管放置在磁性支架上。链霉亲和素结合珠粒在1分钟内沉淀。取出管,弃去澄清的上清液,确保不扰动珠沉淀。将管从磁性支架取出,并且另外重复洗涤两次。第三次洗涤后,取出管,并弃去澄清上清液。加入最后200μL的结合缓冲液,并通过涡旋将珠重悬浮直至均匀。

[0339] 杂交反应完成后,打开热循环仪盖子,并将全部体积的捕获反应液快速转移(36-40μL)到洗涤过的链霉亲和素结合珠。将混合物在室温在振荡器、摇床或旋转器上以足以保持捕获反应/链霉亲和素结合珠溶液均质化的速度混合30分钟。从混合器中取出捕获反应/链霉亲和素结合珠溶液并脉冲旋转,以确保所有溶液都在管的底部。将样品放置在磁性支架上,并使链霉亲和素结合珠沉淀,在1分钟内留下澄清的上清液。取出澄清上清液并弃去。将管从磁性支架中取出,并在室温加入200μL的洗涤缓冲液,然后通过移液混合直至均质化。对管进行脉冲旋转,以确保所有溶液都在管的底部。热循环仪按以下条件编程(表9)。

[0340] 将加热的盖子的温度设定为105℃

[0341] 表9

[0342]

步骤	温度	时间	循环数
1	98℃	45 秒	1
2	98℃	15 秒	9
	60℃	30 秒	
	72℃	30 秒	
3	72℃	1 分钟	1
4	4℃	保持	

[0343]

[0344] 将扩增引物(2.5μL)和聚合酶(如KAPA HiFi HotStart ReadyMix(25μL))加入到含有水/链霉亲和素结合珠浆的管中,并通过移液混合管。然后将管分成两个反应。将管脉冲旋转并转移到热循环仪,并开始表9中的循环程序。当热循环仪程序完成时,从模块中取出样品并立即进行纯化。将在室温预平衡的DNA纯化珠涡旋直至均质化。将90μL(1.8x)均质化的DNA纯化珠加入到管中,并通过涡旋充分混合。将管在室温孵育5分钟,并放置在磁性支架上。DNA纯化珠沉淀,在1分钟内留下澄清的上清液。弃去澄清的上清液,并将管留在磁性支架上。用200μL新制备的80%乙醇洗涤DNA纯化珠沉淀,孵育1分钟,然后将乙醇取出并弃去。重复清洗一次,总共清洗两次,同时将试管保持在磁性支架上。用10微升移液管取出并

弃去所有剩余的乙醇,确保不扰动DNA纯化珠沉淀。将该DNA纯化珠沉淀在磁性支架上空气干燥5-10分钟,或者直到沉淀干燥。将管从磁性支架中取出,添加32μL水,通过移液混合直至均质化,并在室温下孵育2分钟。将管放在磁性支架上3分钟,或者直到珠完全沉淀。回收30μL澄清上清液并转移到清洁的薄壁PCR 0.2mL条管中,确保不扰动DNA纯化珠沉淀。在分析仪器上使用150bp至1000bp的范围设置,平均片段长度在约375bp至约425bp之间。理想情况下,最终浓度值为至少约15ng/μL。使用下一代测序(NGS)对每个捕获进行量化和验证。

[0345] 表10、表11和图37中示出了与比较物外显子组捕获试剂盒(比较物试剂盒D)相比的NGS度量的概括。文库4具有探针(饵),其对应于比比较物试剂盒D更高百分比的外显子靶。这导致使用文库4获得相当的质量和覆盖的靶序列的测序更少。

[0346] 表10

[0347]

NGS 度量	比较物试剂盒 D	文库 4
靶区域	38.8 Mb	33.2 Mb
饵区域	50.8 Mb	36.7 Mb
饵设计效率	76.5%	90.3%

[0348]

捕获丛	8 丛	8 丛
PF 读段	57.7M	49.3M
归一化覆盖	150X	150X
HS 文库大小	30.3M	404.0M
重复百分比	32.5%	2.5%
富集倍数	43.2	48.6
80 倍碱基罚分	1.84	1.40

[0349] 表11

[0350]

NGS 度量	比较物试剂盒 D	文库 4
通过过滤的独特读段百分比 (PCT PF UQ READ)	67.6%	97.5%
1X 的靶碱基百分比	99.8%	99.8%
20X 的靶碱基百分比	90.3%	99.3%
30X 的靶碱基百分比	72.4%	96.2%

[0351] 表12和图37中示出了试剂盒D和文库4的重叠靶区域的比较(总读段针对96X覆盖归一化)。文库4按每个杂交8个样品来处理,试剂盒D以每个杂交2个样品来处理。此外,对于这两个文库,将来自重叠区域的单核苷酸多态性和框内缺失检出与从“Genome in a Bottle”NA12878参考数据中鉴定的高置信度区域进行了比较(表13)。文库4在鉴定SNP和插入缺失方面表现与试剂盒D相似或更佳(更高的插入缺失精确度)。

[0352] 表12

[0353]

NGS 度量	比较物试剂盒 D	文库 4
通过过滤的读段百分比 (PCT PF UQ READ)	94.60%	97.7%
选择的碱基百分比	79%	80%
1X 的靶碱基百分比	100%	100%
20X 的靶碱基百分比	90%	96%
30X 的靶碱基百分比	71%	77%
富集倍数	44.9	49.9
80 倍碱基罚分	1.76	1.4
HS 文库大小	122M	267M

[0354] 表13

[0355]

变体	比较物试剂盒 D		文库 4	
	精确度	灵敏度	精确度	灵敏度
单核苷酸多态性(SNP)	98.59%	99.23%	99.05%	99.27%
框内缺失(Indel)	76.42%	94.12%	87.76%	94.85%
总计	98.14%	99.15%	98.85%	99.20%

[0356] 精确度表示真阳性检出与总阳性检出(真和假)的比率。敏感性表示真阳性检出与总真值(真阳性和假阴性)的比率。

[0357] 虽然本文已经示出并描述了本发明的优选实施方案,但是对于本领域技术人员将明显的是,这些实施方案仅以实例的方式提供。在不脱离本发明的情况下,本领域技术人员现在将会想到许多变化、改变和替换。应当理解,在实践本发明时,可以采用本文描述的本发明实施方案的各种替代方案。所附权利要求意图限定本发明的范围,并且这些权利要求及其等同物范围内的方法和结构由此被覆盖。

序列表

<110> 特韦斯特生物科学公司

<120> 具有受控化学计量的多核苷酸文库及其合成

<130> 44854-730.601

<140>

<141>

<150> 62/558,666

<151> 2017-09-14

<150> 62/548,307

<151> 2017-08-21

<150> 62/424,302

<151> 2016-11-18

<160> 16

<170> PatentIn 3.5版

<210> 1

<211> 62

<212> DNA

<213> 人工序列(Artificial Sequence)

<220>

<223> 人工序列(Artificial Sequence)的描述:合成寡核苷酸

<220>

<221> 修饰的碱基

<222> (51) .. (52)

<223> 胸苷-琥珀酰六酰胺CED亚磷酰胺

<400> 1

agacaatcaa ccatttgggg tggacagcct tgacctctag acttcggcat tttttttttt 60
tt 62

<210> 2

<211> 112

<212> DNA

<213> 人工序列(Artificial Sequence)

<220>

<223> 人工序列(Artificial Sequence)的描述:合成多核苷酸

<220>

<221> 修饰的碱基

<222> (101) .. (102)

<223> 胸苷-琥珀酰六酰胺CED亚磷酰胺

<400> 2

cgggatacctt atcgtcatcg tcgtacagat cccgacccat ttgctgtcca ccagtcacgc 60
tagccataacc atgatgatga tgatgatgag aacccccgcat tttttttttt tt 112

<210> 3

<211> 19

<212> DNA

<213> 人工序列 (Artificial Sequence)

<220>

<223> 人工序列 (Artificial Sequence) 的描述:合成引物

<400> 3

atgcgggggtt ctcatcatc 19

<210> 4

<211> 20

<212> DNA

<213> 人工序列 (Artificial Sequence)

<220>

<223> 人工序列 (Artificial Sequence) 的描述:合成引物

<400> 4

cgggatacctt atcgtcatcg 20

<210> 5

<211> 156

<212> DNA

<213> 智人 (Homo sapiens)

<400> 5

cccaccggga ctcagattct ccccagacgc cgaggatggt gtcacatggcg ccccgaaacc 60
tcctcctgct gctctcaggg gccctggccc tgaccagac ctgggcgcgt gactgcaggg 120
tctgcaggga aatgggcgcg tgactgcagg gtctgc 156

<210> 6

<211> 293

<212> DNA

<213> 智人 (Homo sapiens)

<400> 6

gctcccaggt tcccaactcca tgaggtatct ctacaccacc atgtcccggc ccggcgccgg 60
ggagccccgc ttcatctccg tcggtacgt ggacgatacg cagttcgtgc ggttcgacag 120
cgacgacgcg agtccgagag aggagccgcg ggcgccgtgg atggagcggg aggggccaaa 180
gtattgggac cggaacacac agatctgcaa ggcccaggca cagactgaac gagagaacct 240
gcggatcgcg ctccgctact acaaccagag cgagggcggt gatttgacct cgg 293

<210> 7

<211> 169

<212> DNA

<213> 智人 (Homo sapiens)

<400> 7

tagcagggtc agggttccct caccttcccc ctttttccca gccatcttcc cagcccaccg 60
tccccatcgt gggcatcggt gctggcttgg ttctacttgt agctgtggtc actggagctg 120
tggtcgctgc tgtaatgtgg aggaagaaga gctcaggtaa ggaaggggt 169

<210> 8

<211> 160

<212> DNA

<213> 智人 (Homo sapiens)

<400> 8

gttaccceaag aacgcagctg atttctccca gacgccgagg atggtgctca tggcgccccg 60
aaccctcctc ctgctgctct caggggccct ggccctgacc cagacctggg cgcgtgagtg 120
cagggctctgc agggaaatgg tagtgtcgga ggtcgttcct 160

<210> 9

<211> 160

<212> DNA

<213> 智人 (Homo sapiens)

<400> 9

gttaccceaag aacgcagctg ggttcccact ccatgaggtg tttctacacc accatgtccc 60
ggcccggcgc cggggagccc cgcttcatct ccgtcggcta cgtggacgat acgcagttcg 120
tgcggttcga cagcgacgac tagtgtcgga ggtcgttcct 160

<210> 10

<211> 160

<212> DNA

<213> 智人 (Homo sapiens)

<400> 10

gttaccceaag aacgcagctg tggatggagc gggaggggcc aaagtattgg gaccggaaca 60
cacagatctg caaggcccag gcacagactg aacgagagaa cctgcggatc gcgctccgct 120
actacaacca gagcgagggc tagtgtcgga ggtcgttcct 160

<210> 11

<211> 160

<212> DNA

<213> 智人 (Homo sapiens)

<400> 11

gttaccceaag aacgcagctg ggctacgtgg acgatacgca gttcgtgcgg ttcgacagcg 60
acgacgcgag tccgagagag gagccgcggg cgccgtggat ggagcgggag gggccaaagt 120
attgggaccg gaacacacag tagtgtcgga ggtcgttcct 160

<210> 12

<211> 103

<212> DNA

<213> 智人 (Homo sapiens)

<400> 12

gttacccaag aacgcagctg atggcgcccc gaaccctcct cctgctgctc tcagggggccc 60
tggccctgac ccagacctgg gcgtagtgct ggaggtcgctt cct 103

<210> 13

<211> 210

<212> DNA

<213> 智人 (Homo sapiens)

<400> 13

aaaaaaaaaa aaaaaaaaaa aaaaagttac ccaagaacgc agctgggttc ccactccatg 60
aggtattttct acaccacccat gtcccgcccc ggcgccgggg agccccgctt catctccgctc 120
ggctacgtgg acgatacgca gttcgtgcgg ttcgacagcg acgactagtg tcggaggtcg 180
ttcctaataaa aaaaaaaaaa aaaaaaaaaa 210

<210> 14

<211> 210

<212> DNA

<213> 智人 (Homo sapiens)

<400> 14

gttacccaag aacgcagctg aaaaaaaaaa aaaaaaaaaa aaaaaggttc ccactccatg 60
aggtattttct acaccacccat gtcccgcccc ggcgccgggg agccccgctt catctccgctc 120
ggctacgtgg acgatacgca gttcgtgcgg ttcgacagcg acgacaaaaa aaaaaaaaaa 180
aaaaaaaaaaa tagtgctcga ggtcgttcct 210

<210> 15

<211> 102

<212> DNA

<213> 智人 (Homo sapiens)

<400> 15

gttacccaag aacgcagctg agccatcttc ccagcccacc gtccccatcg tgggcatcgt 60
tgctggcttg gttctacttg tatagtgtcg gaggtcgctt ct 102

<210> 16

<211> 95

<212> DNA

<213> 智人 (Homo sapiens)

<400> 16

gttacccaag aacgcagctg gctgtggtca ctggagctgt ggtcgtgct gtaatgtgga 60
ggaagaagag ctgagtagtg tcggaggtcg ttcct 95

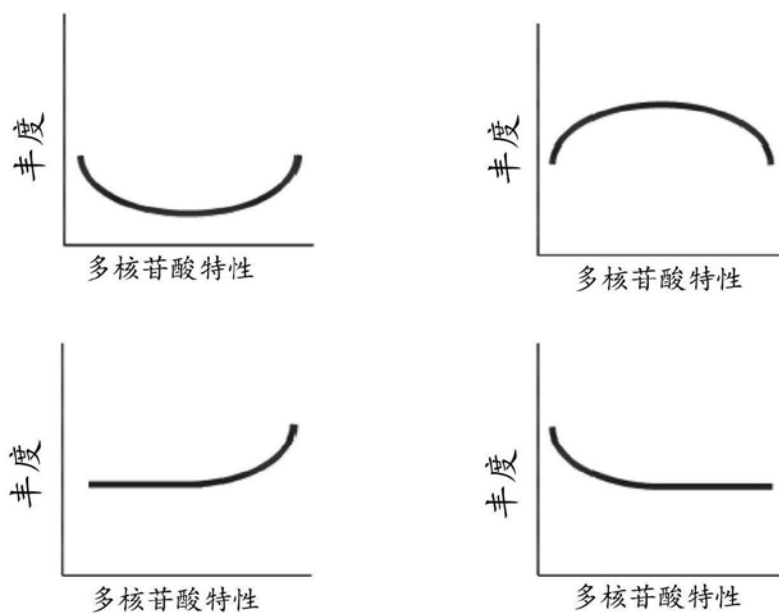
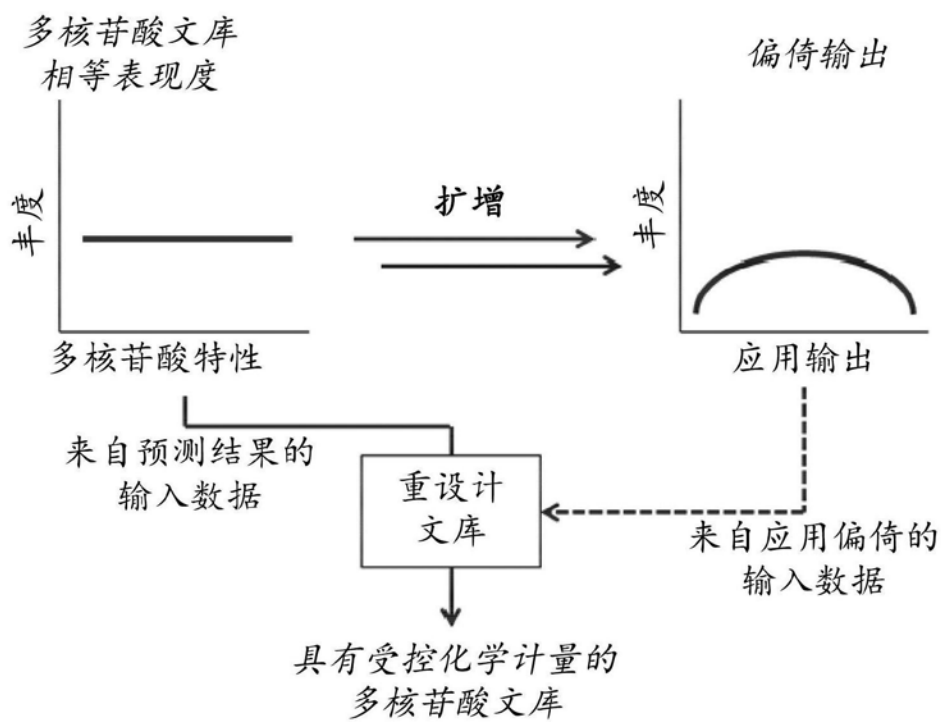


图1A

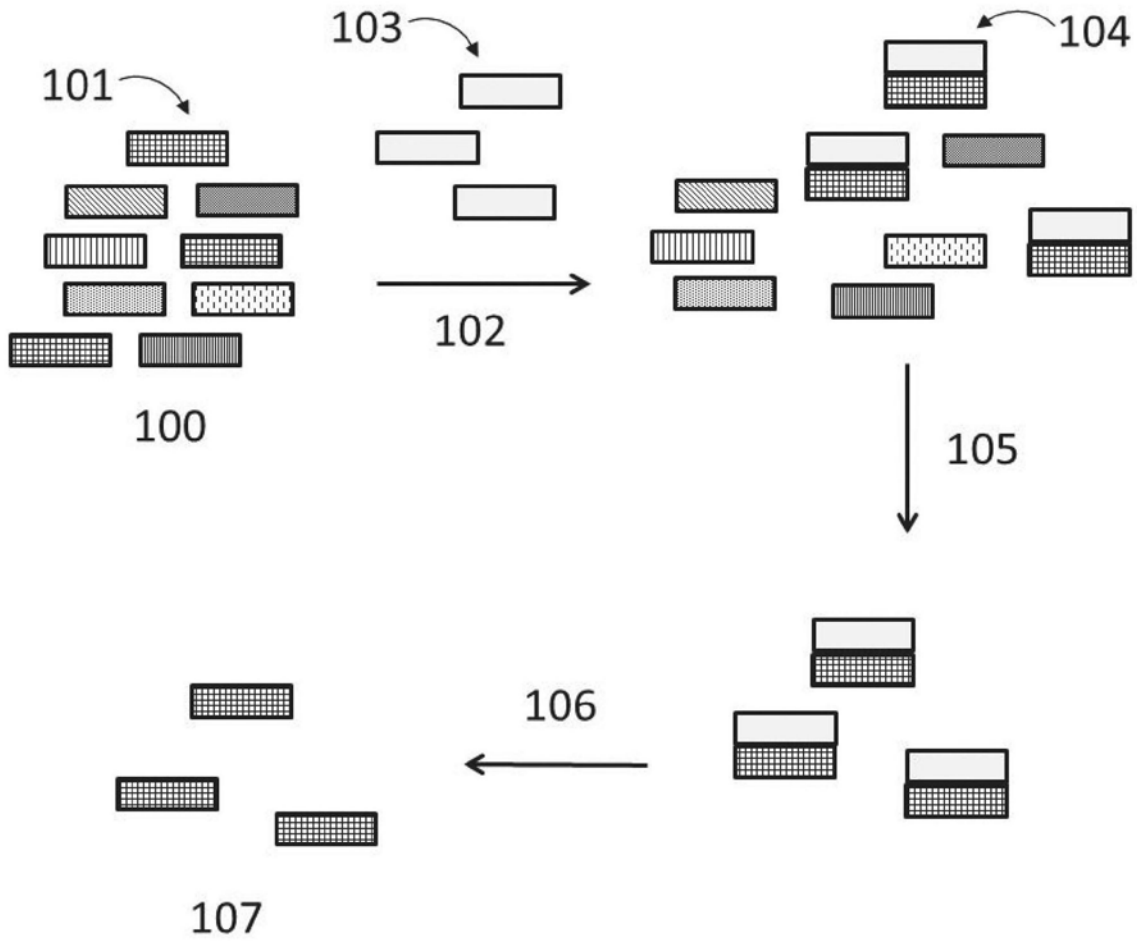


图1B

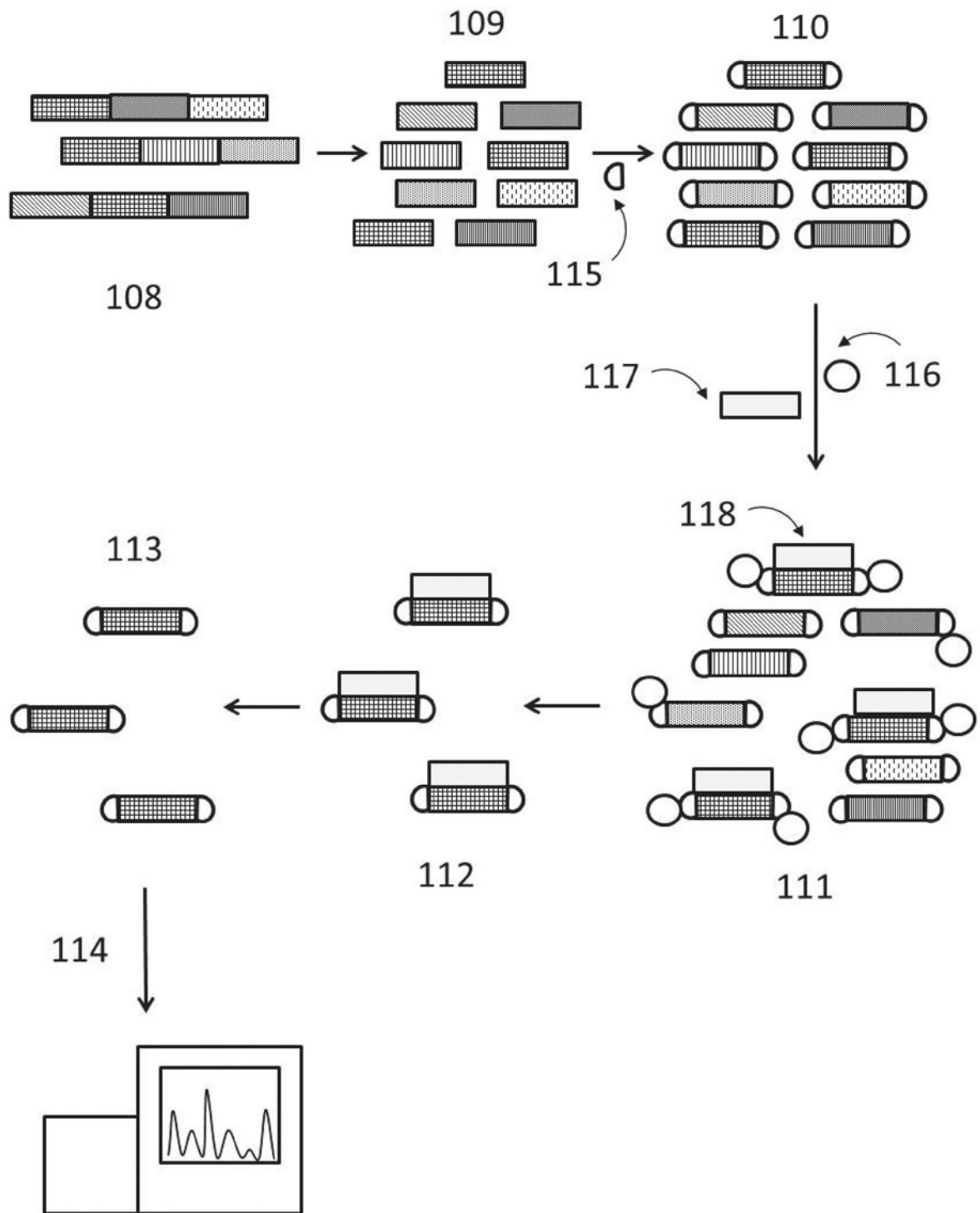


图1C

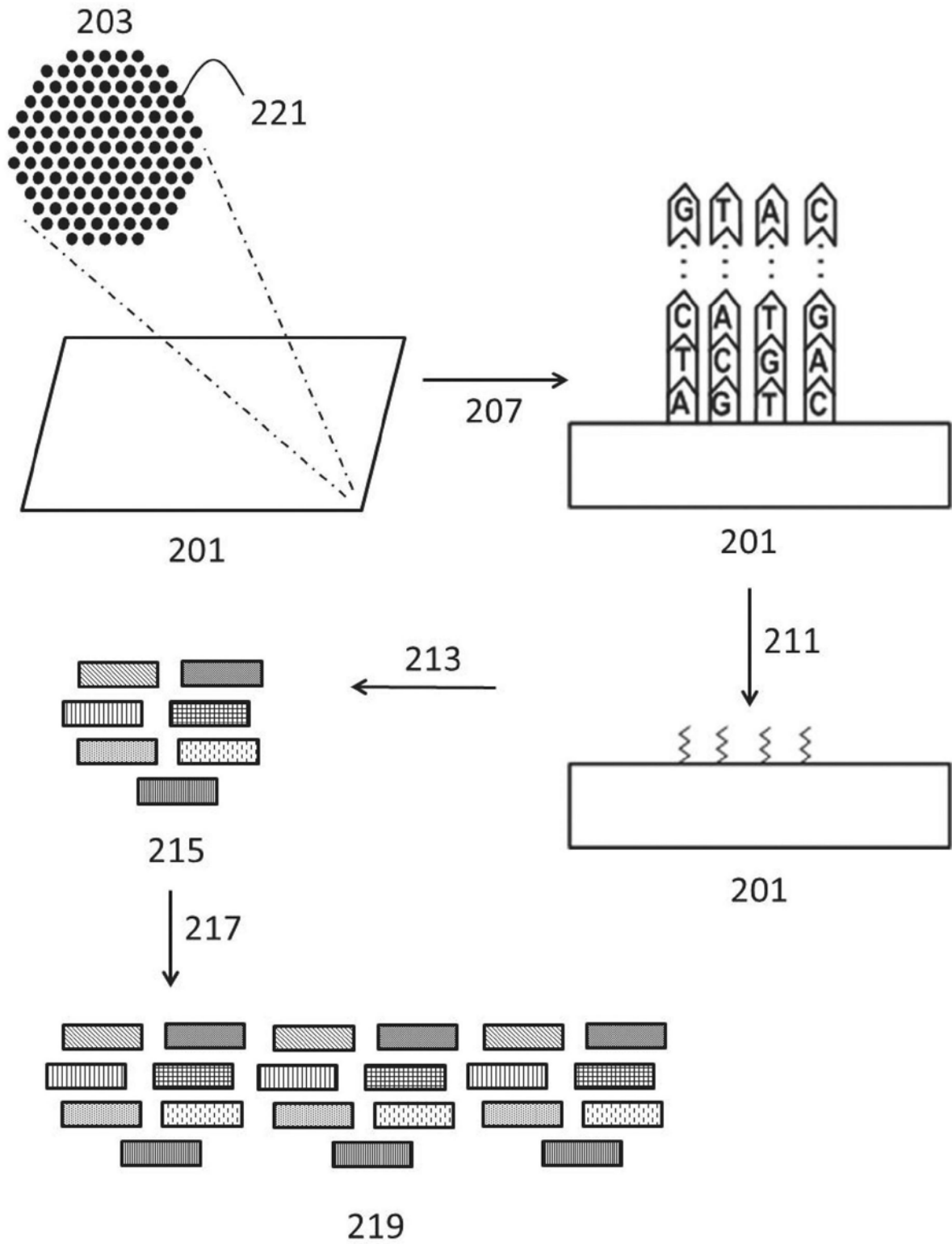


图2

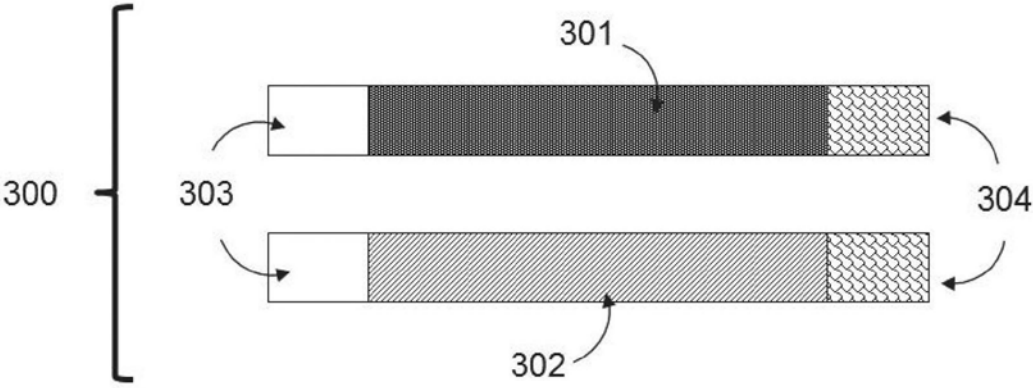


图3A

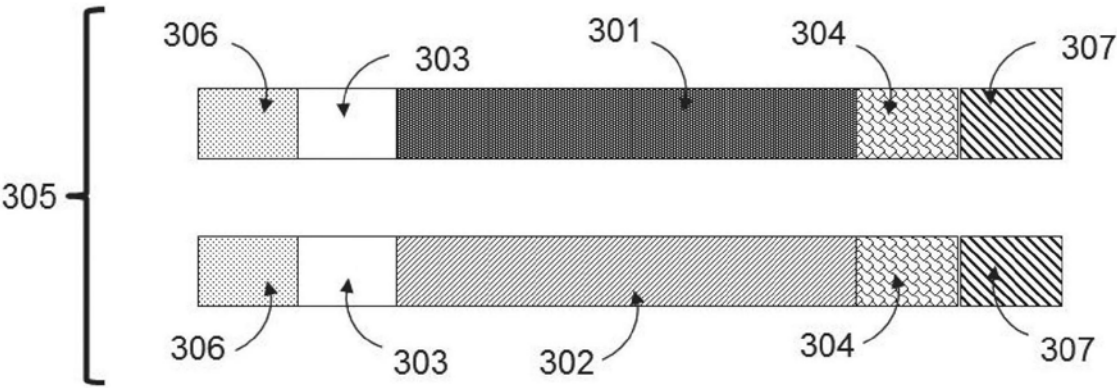


图3B

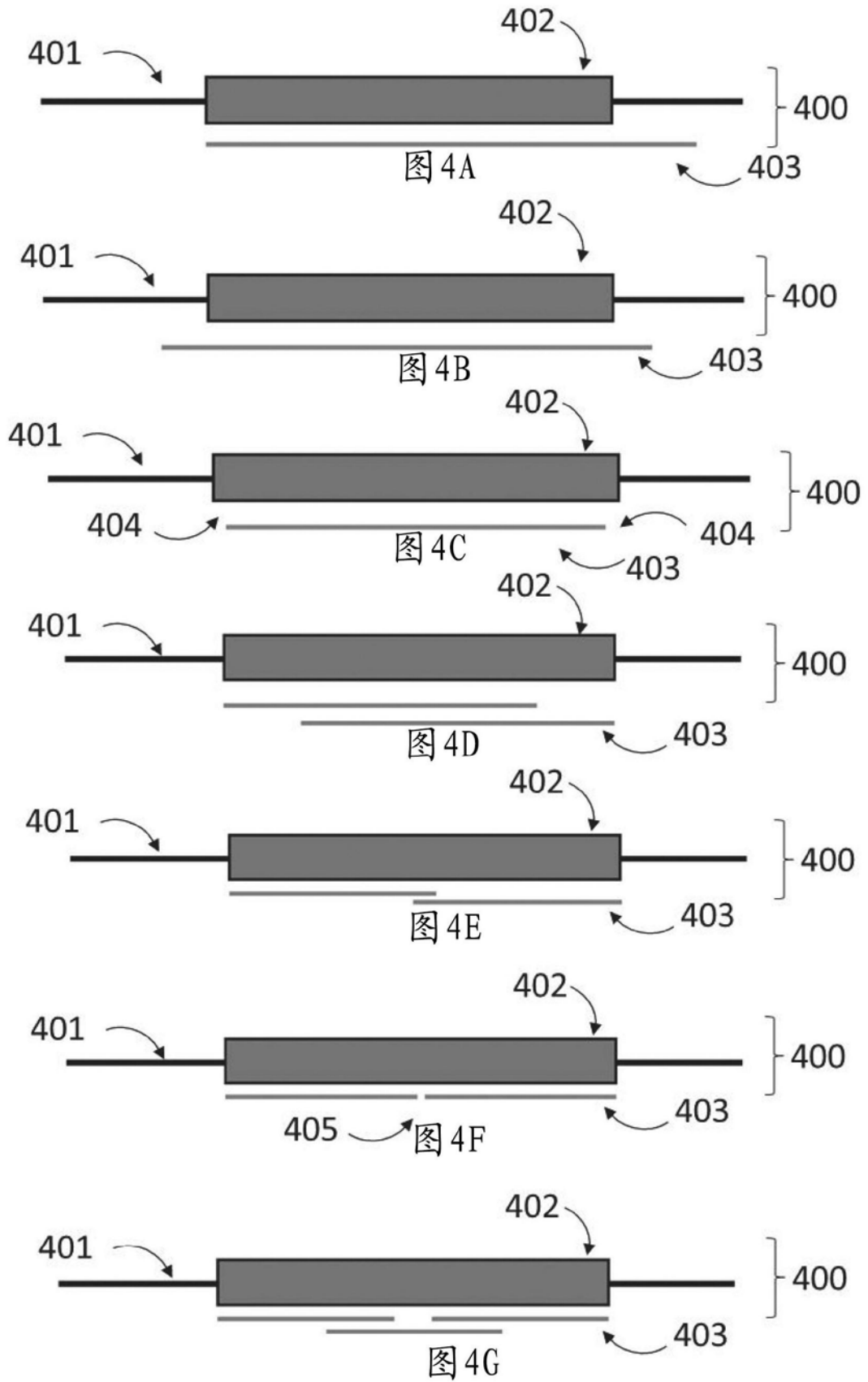


图4A-4G

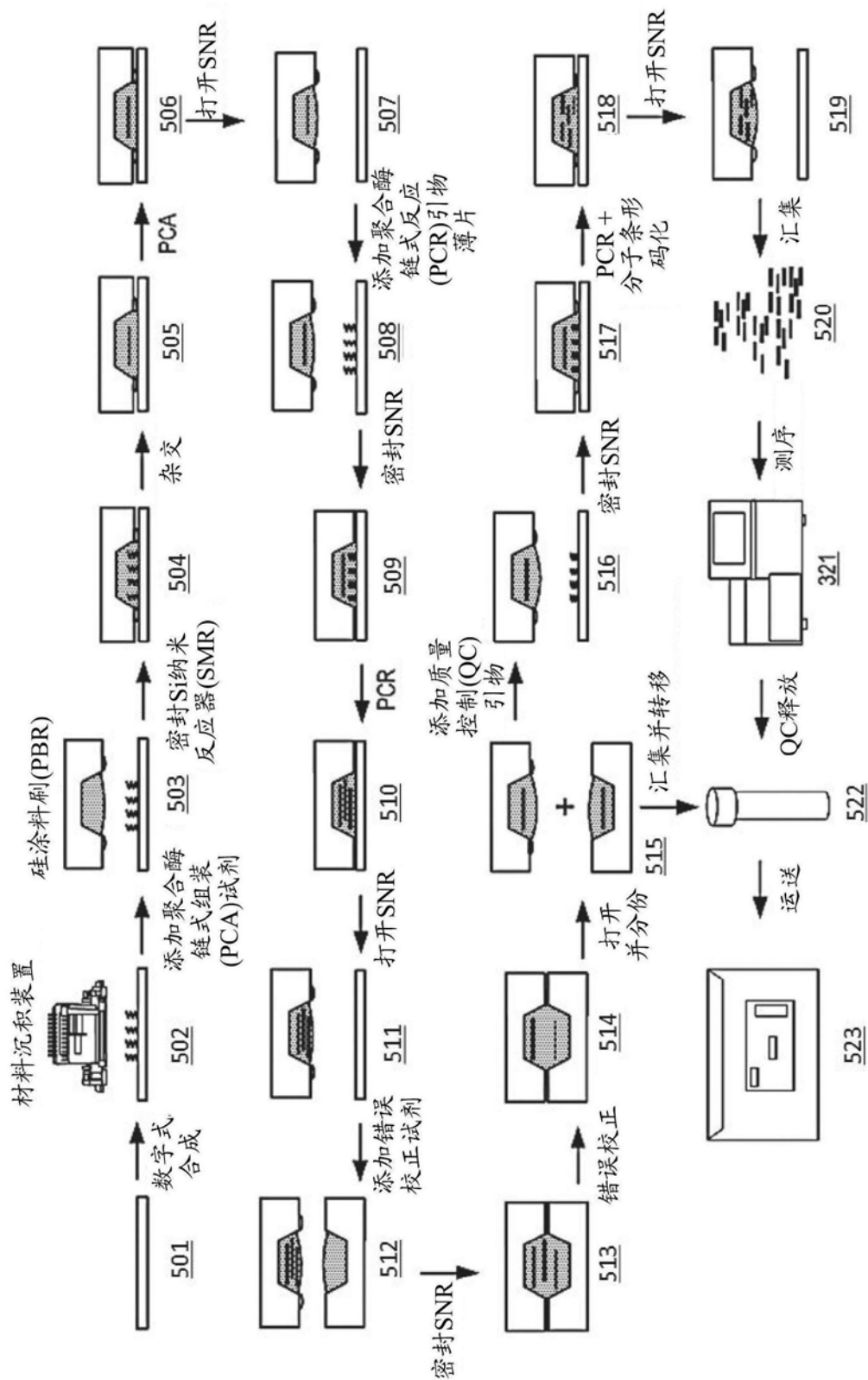


图5

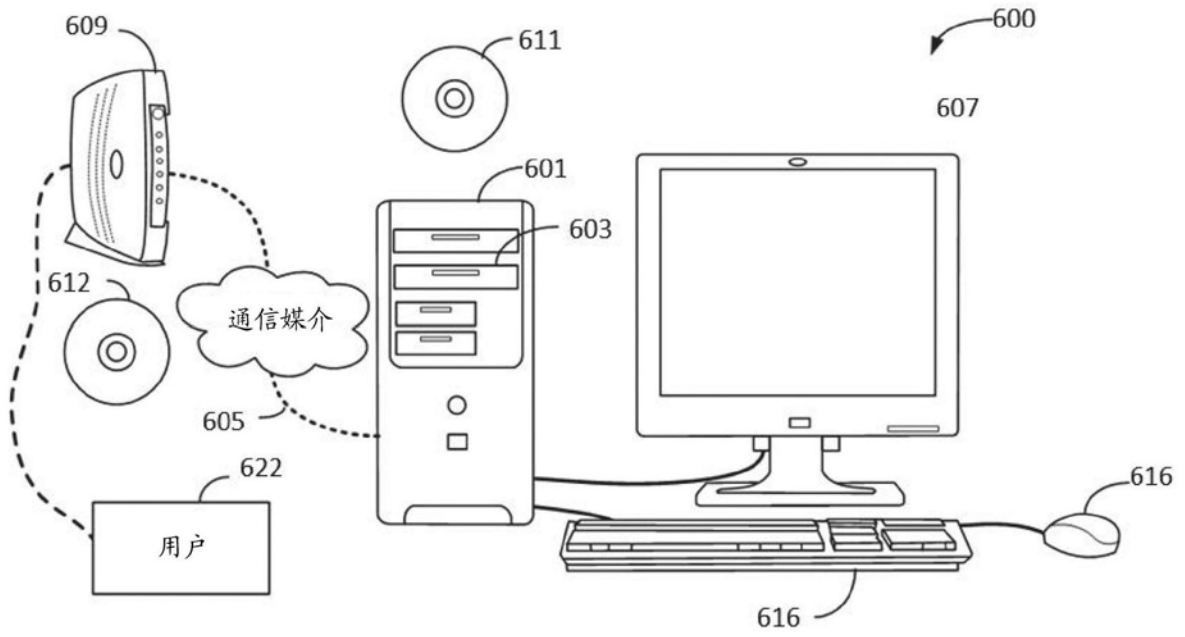


图6

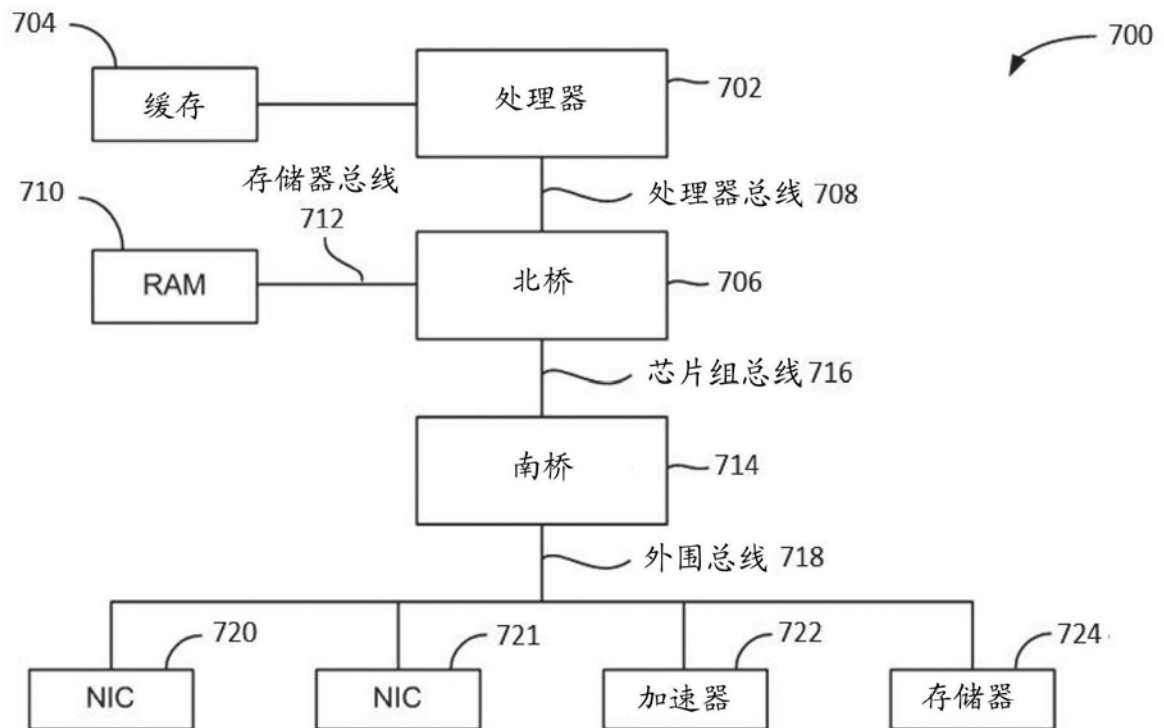


图7

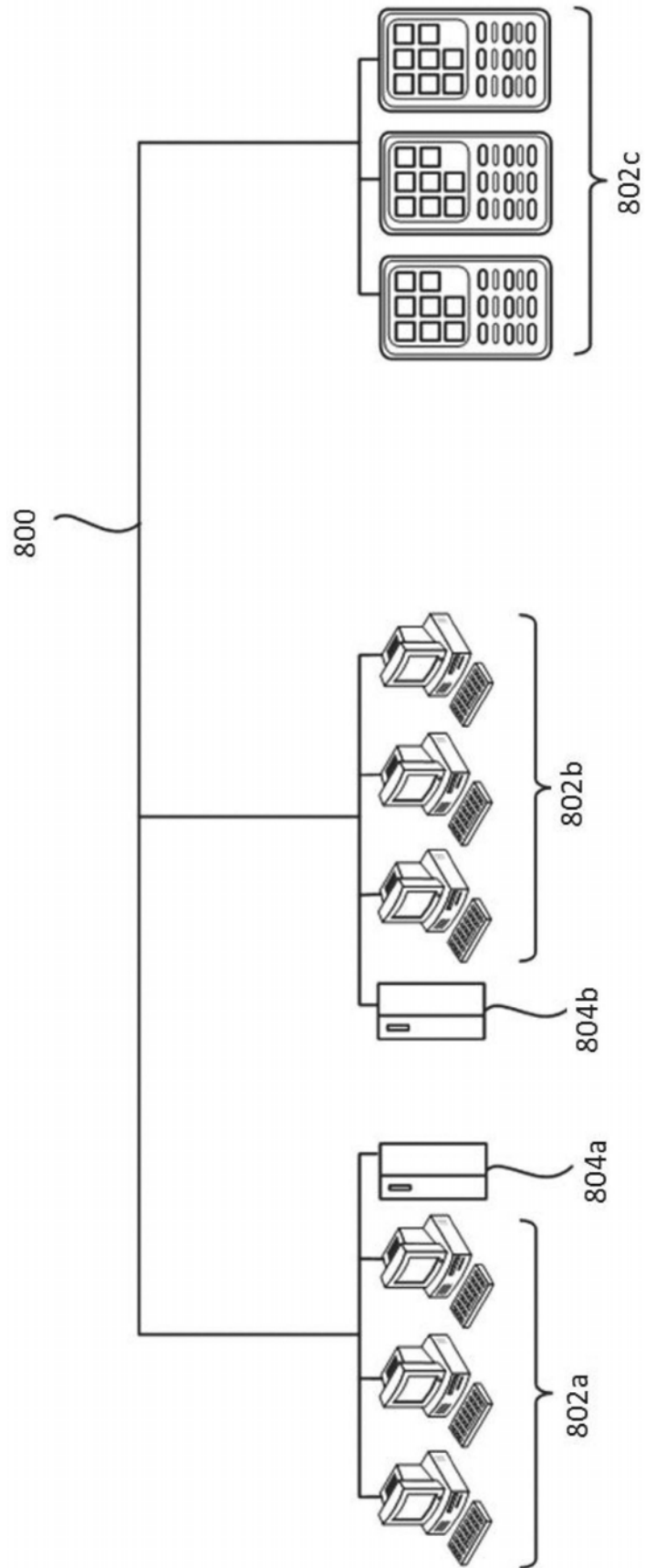


图8

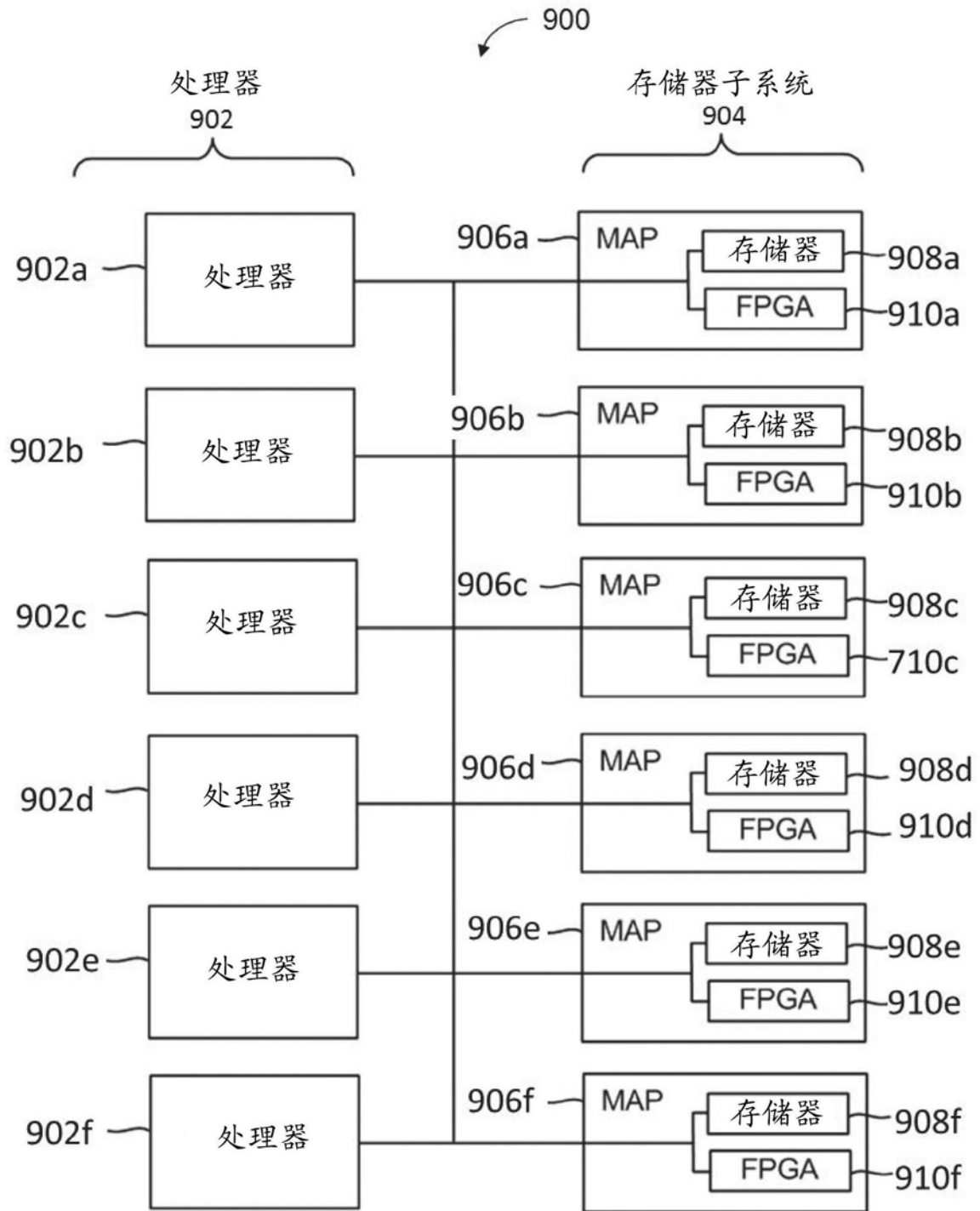


图9

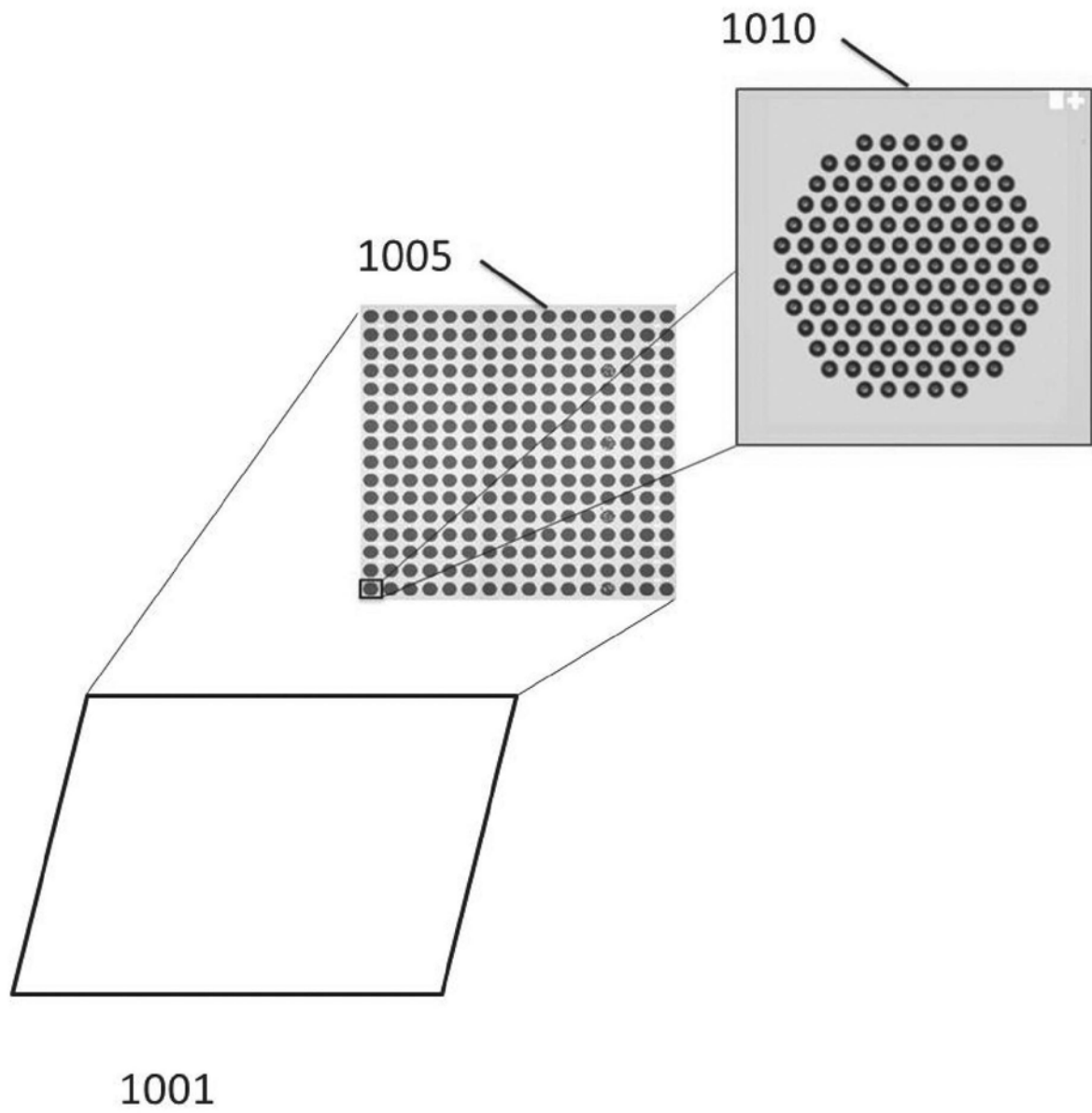


图10

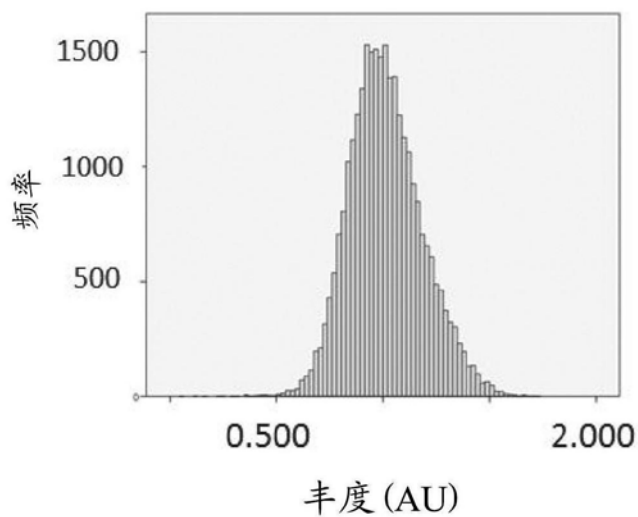


图11A

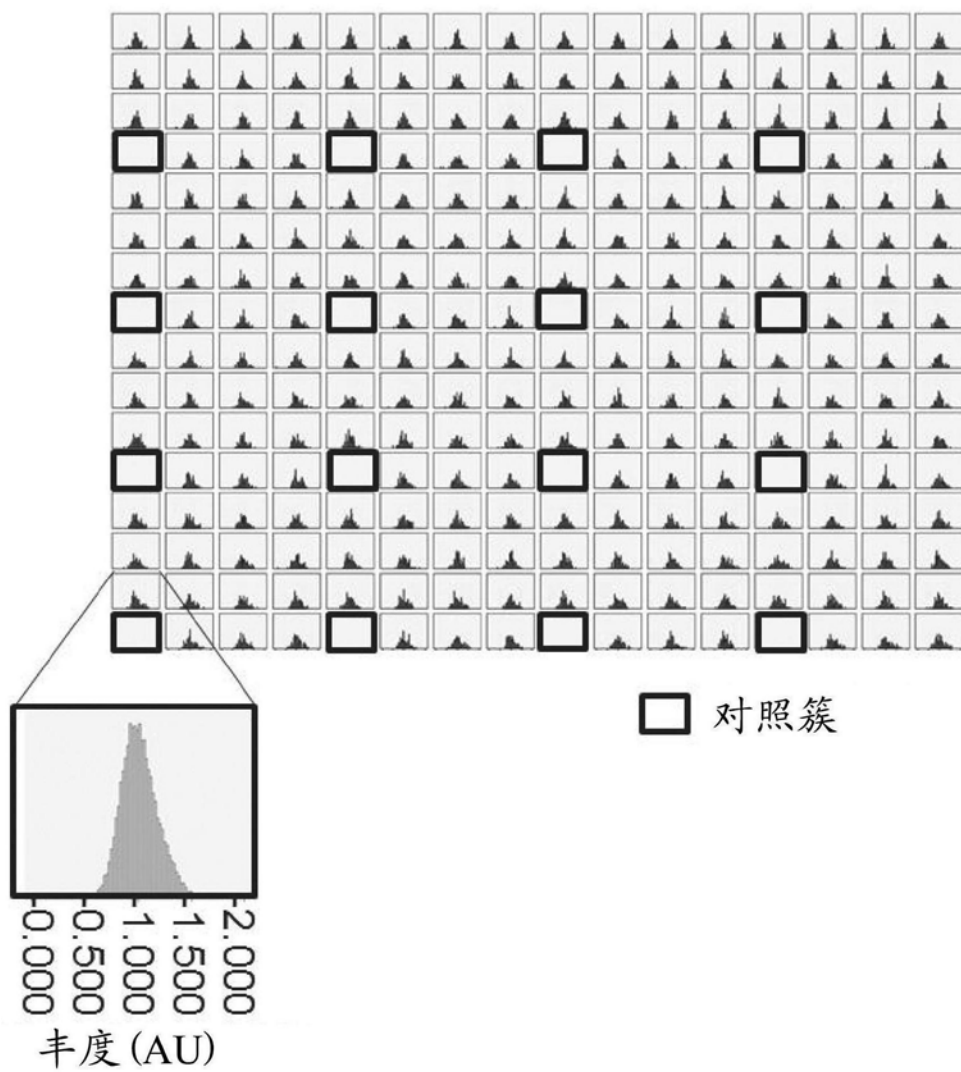


图11B

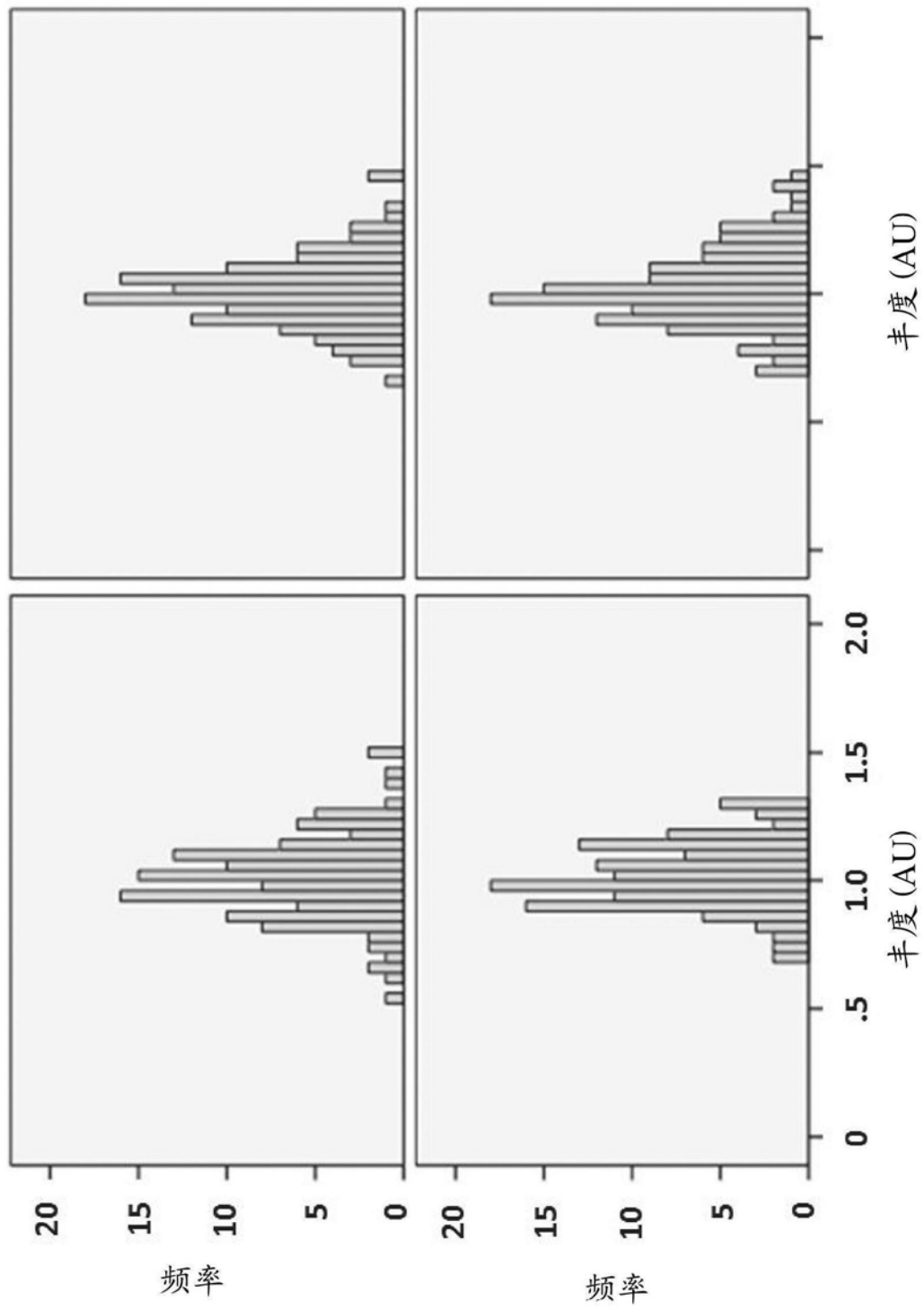


图12

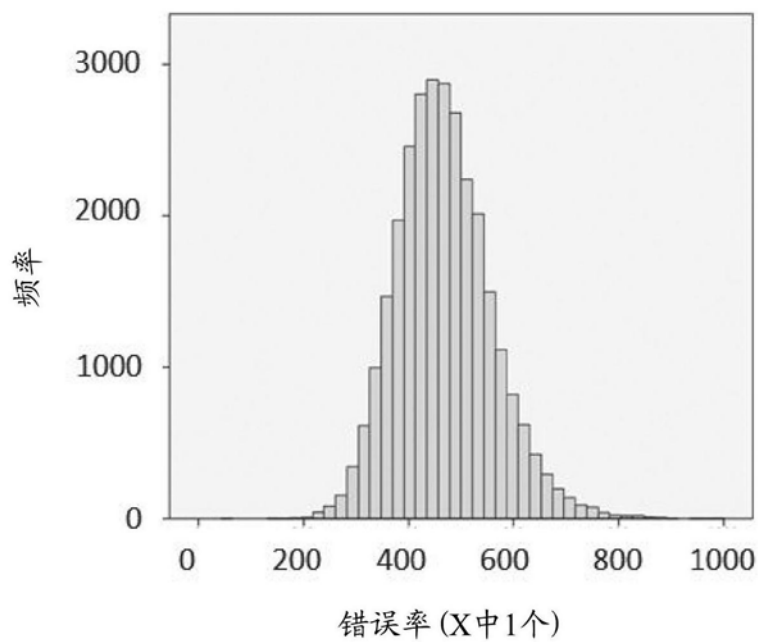
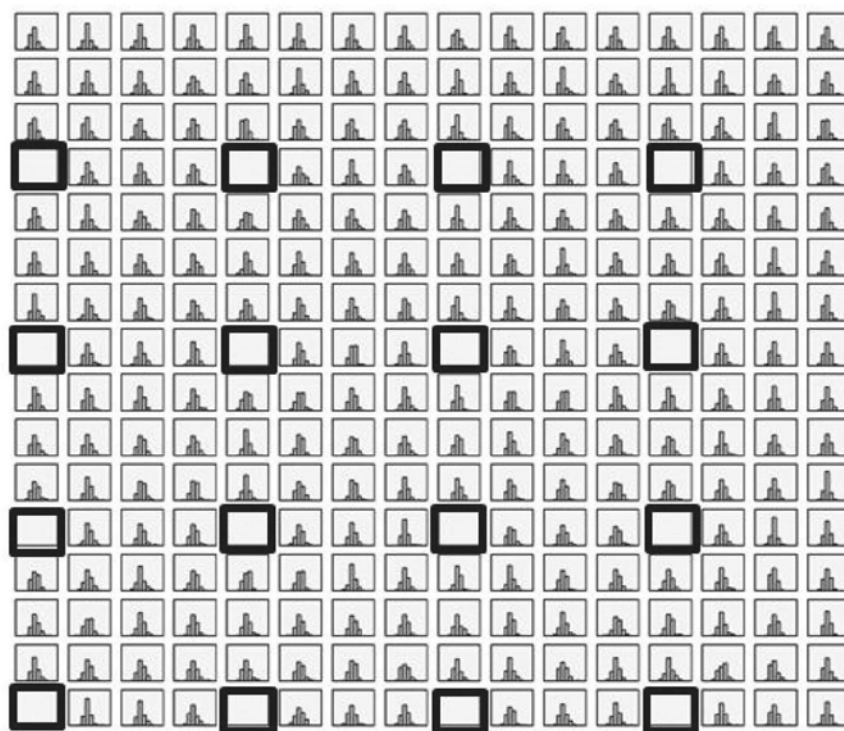


图13A



 对照簇

图13B

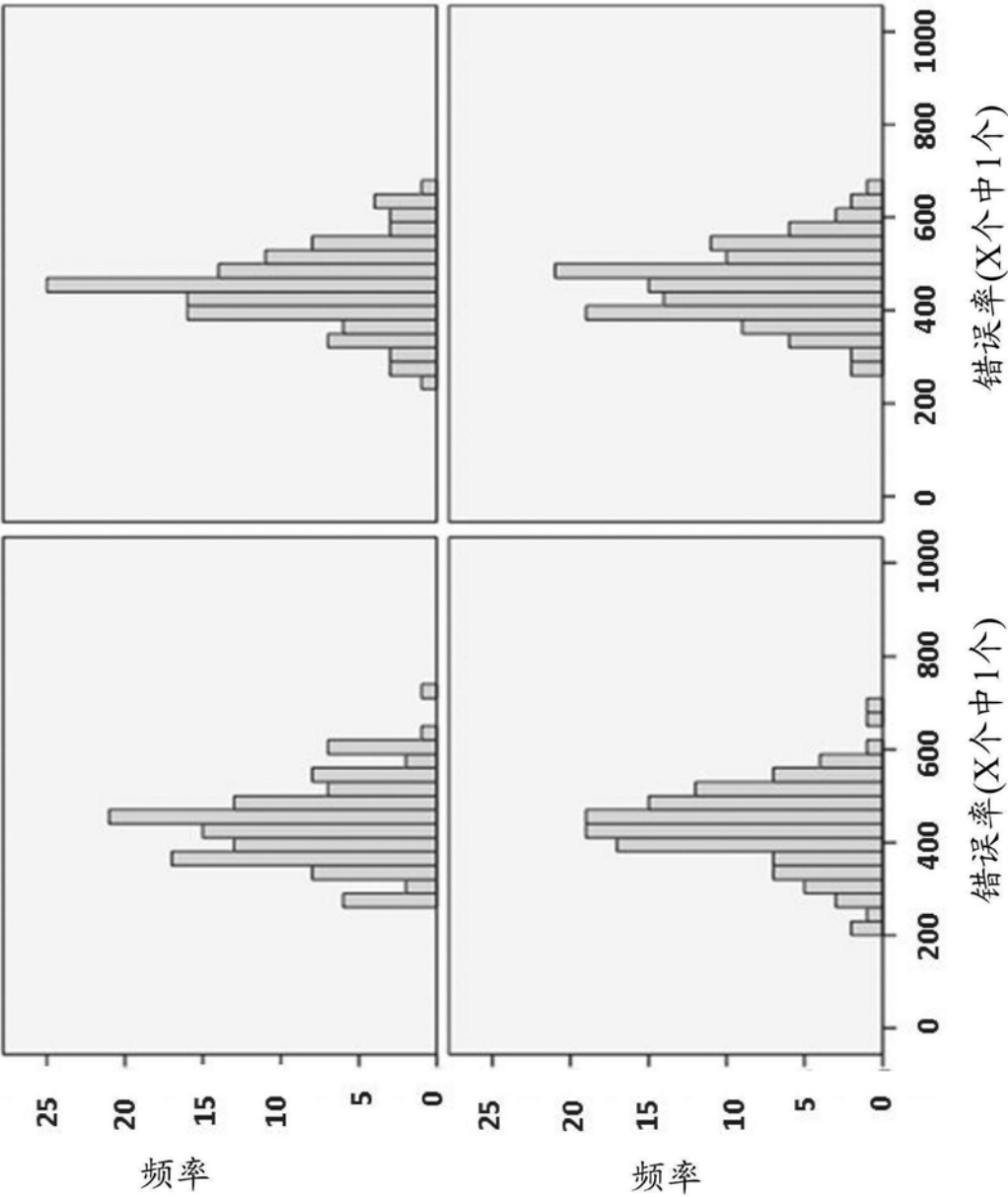


图14

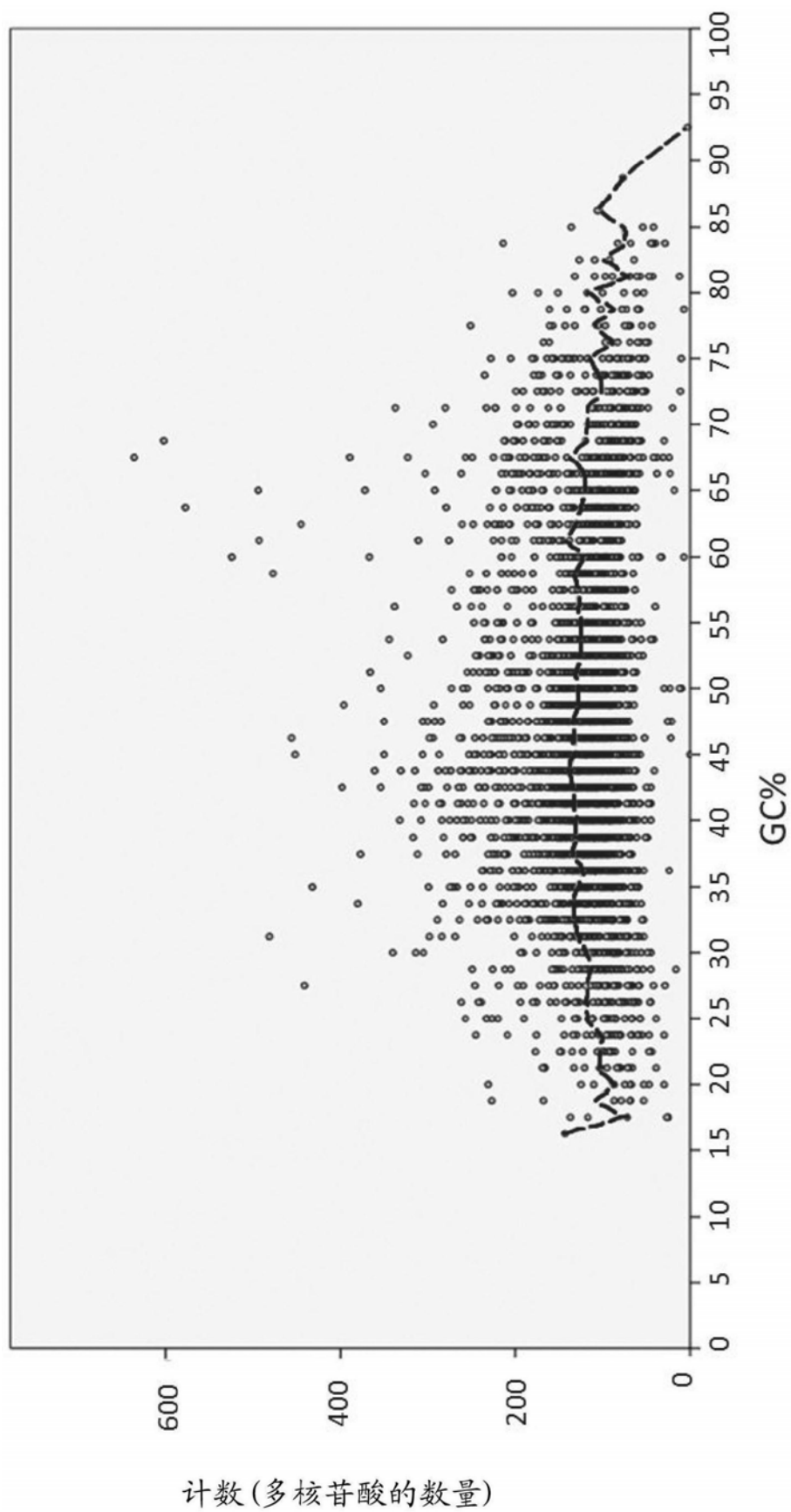


图15

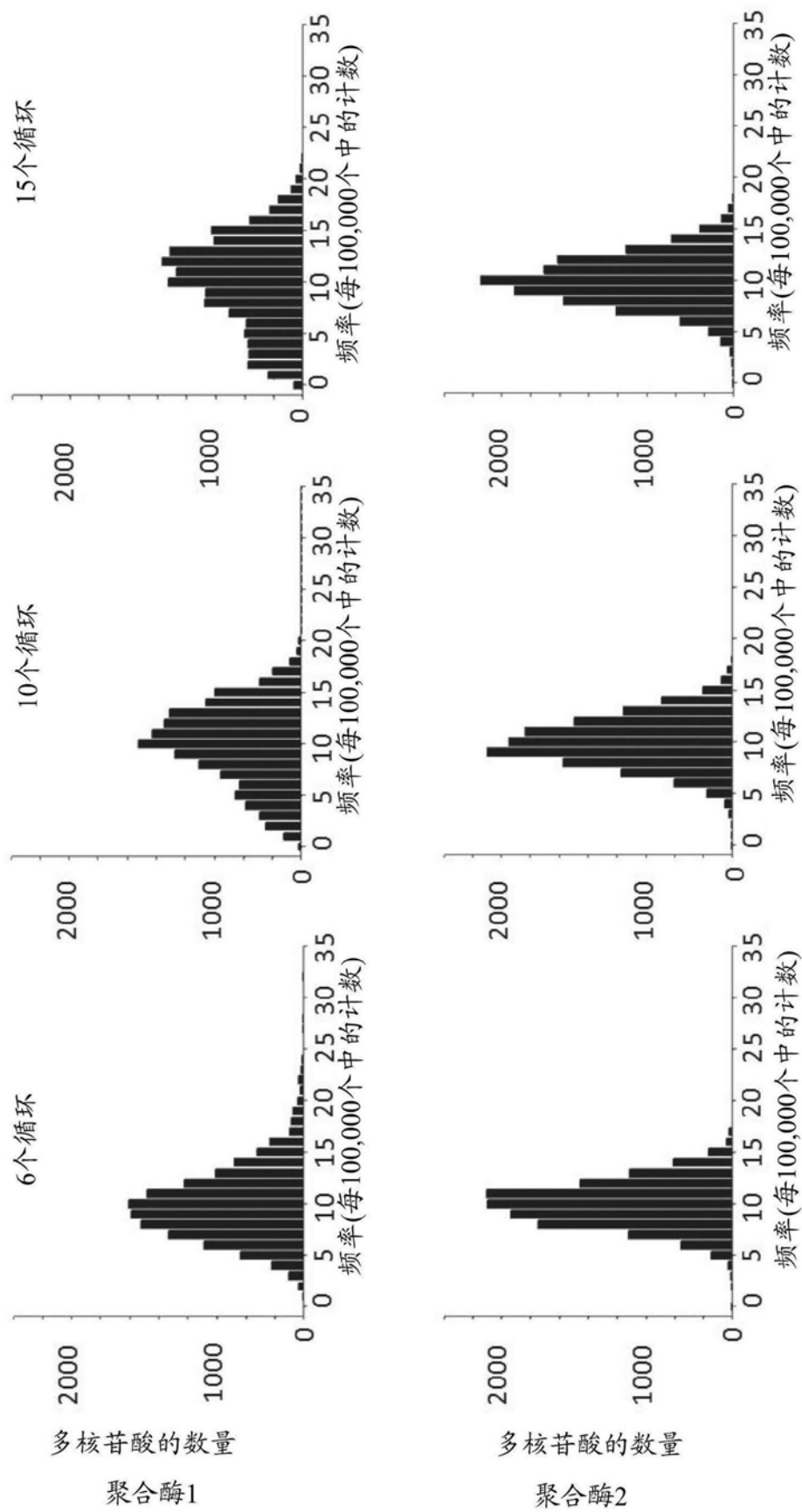


图16

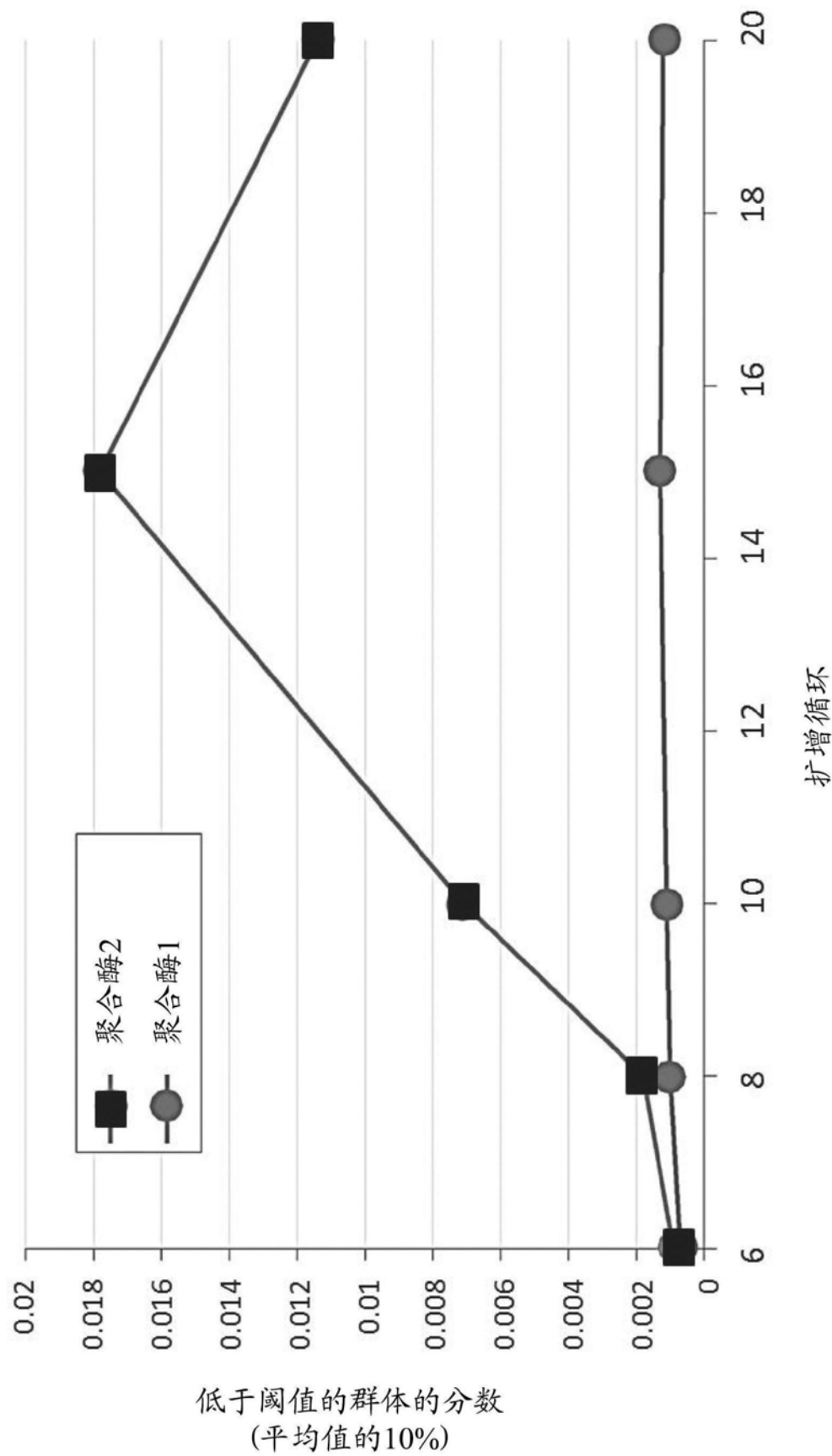


图17

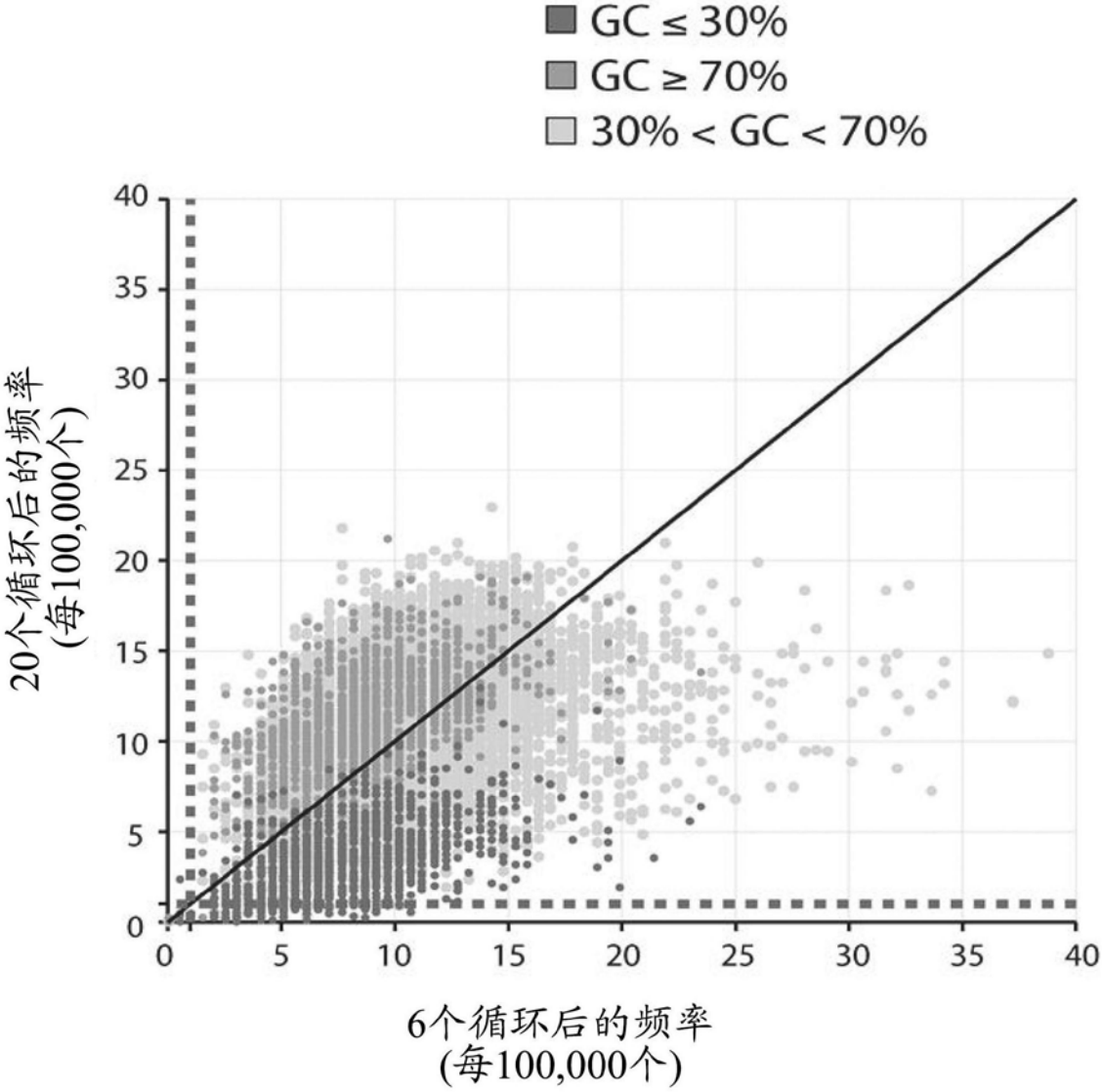


图18

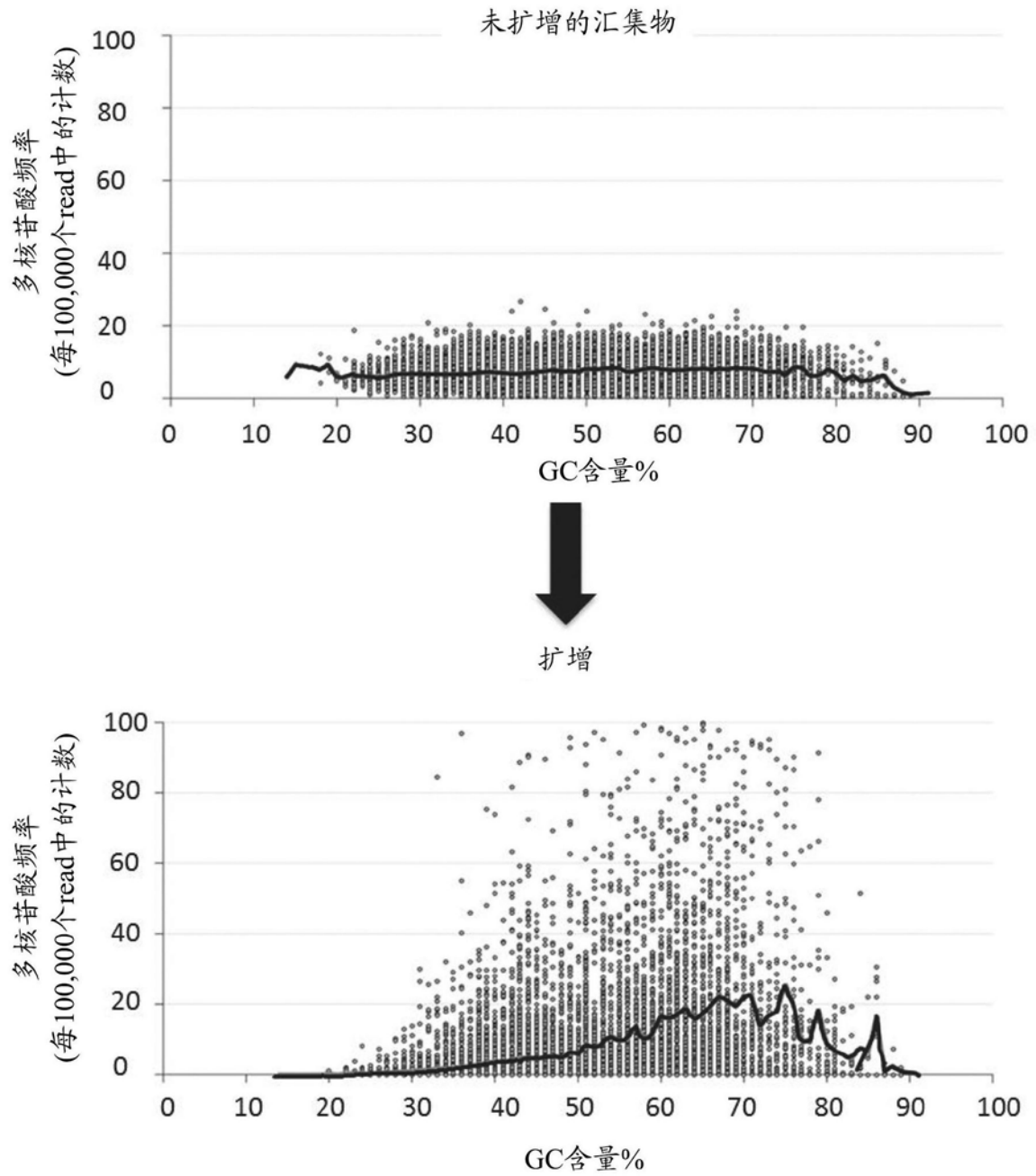


图19

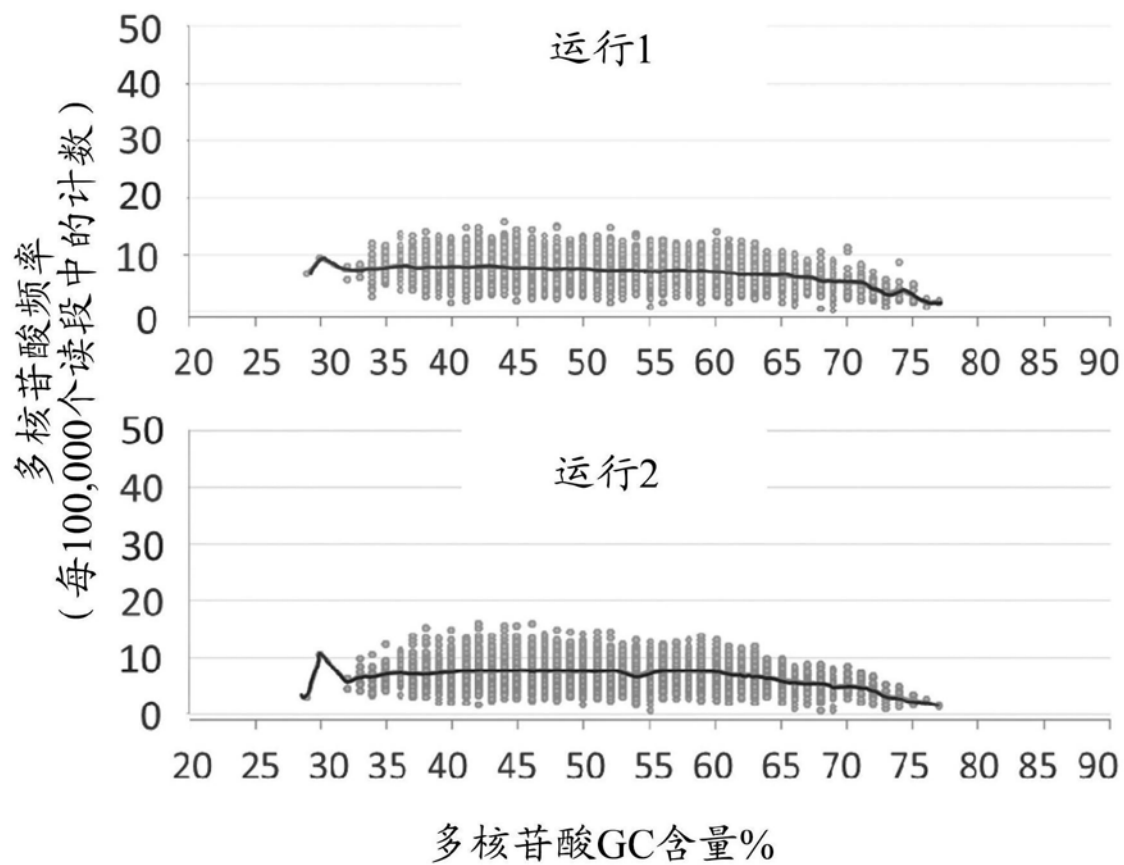


图20

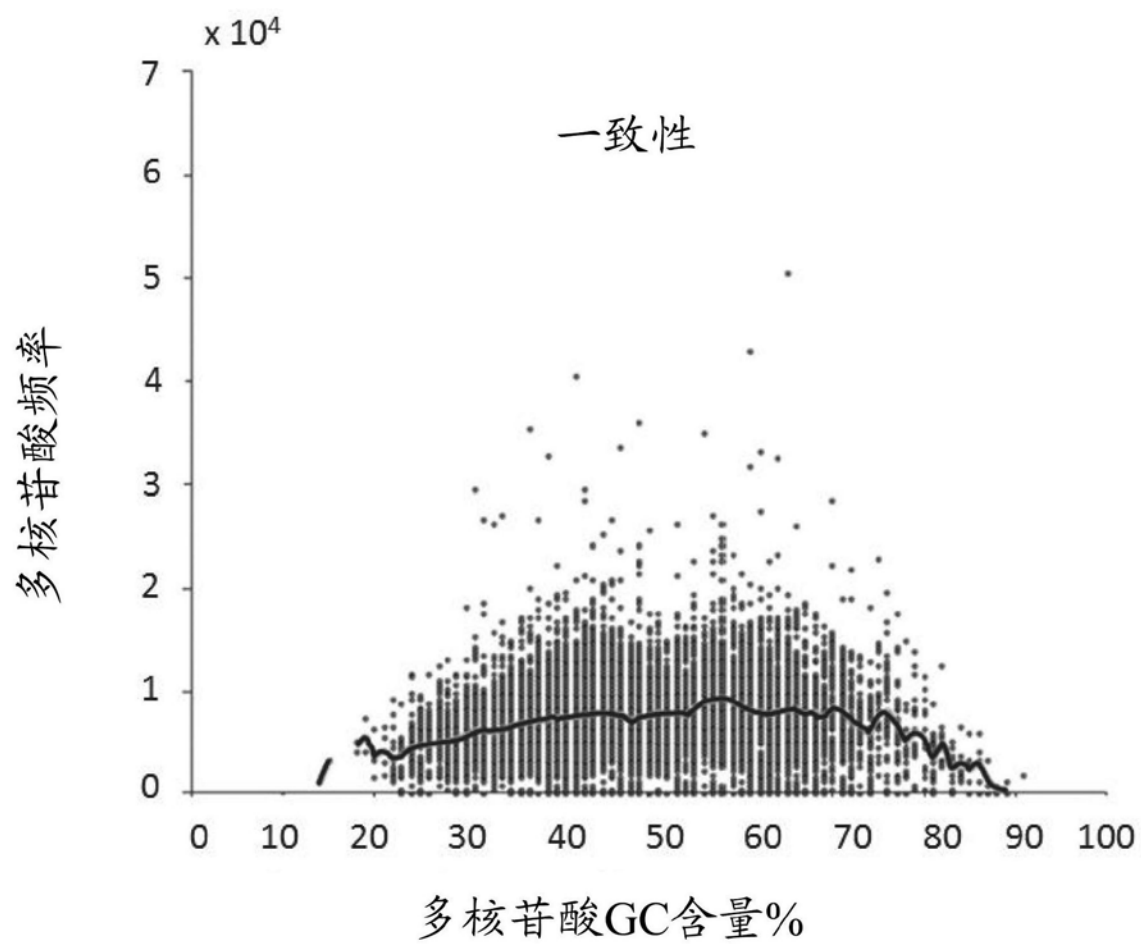


图21A

重度偏向高/低GC

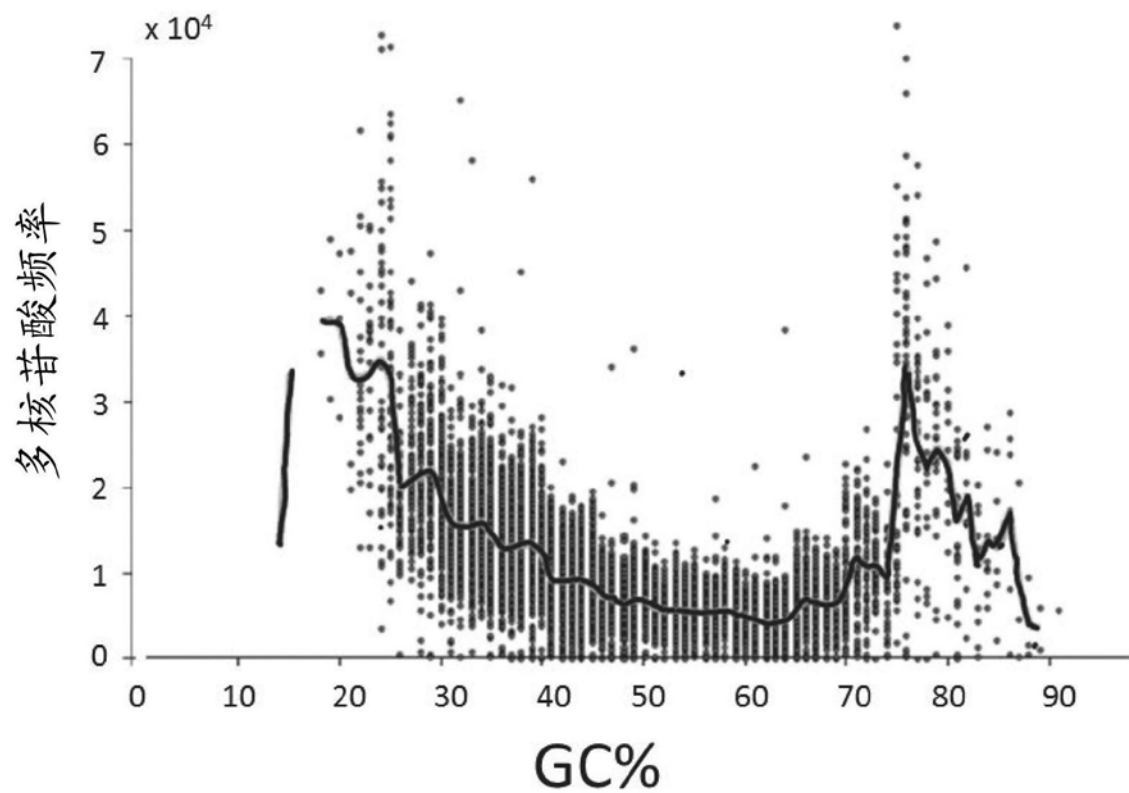


图21B

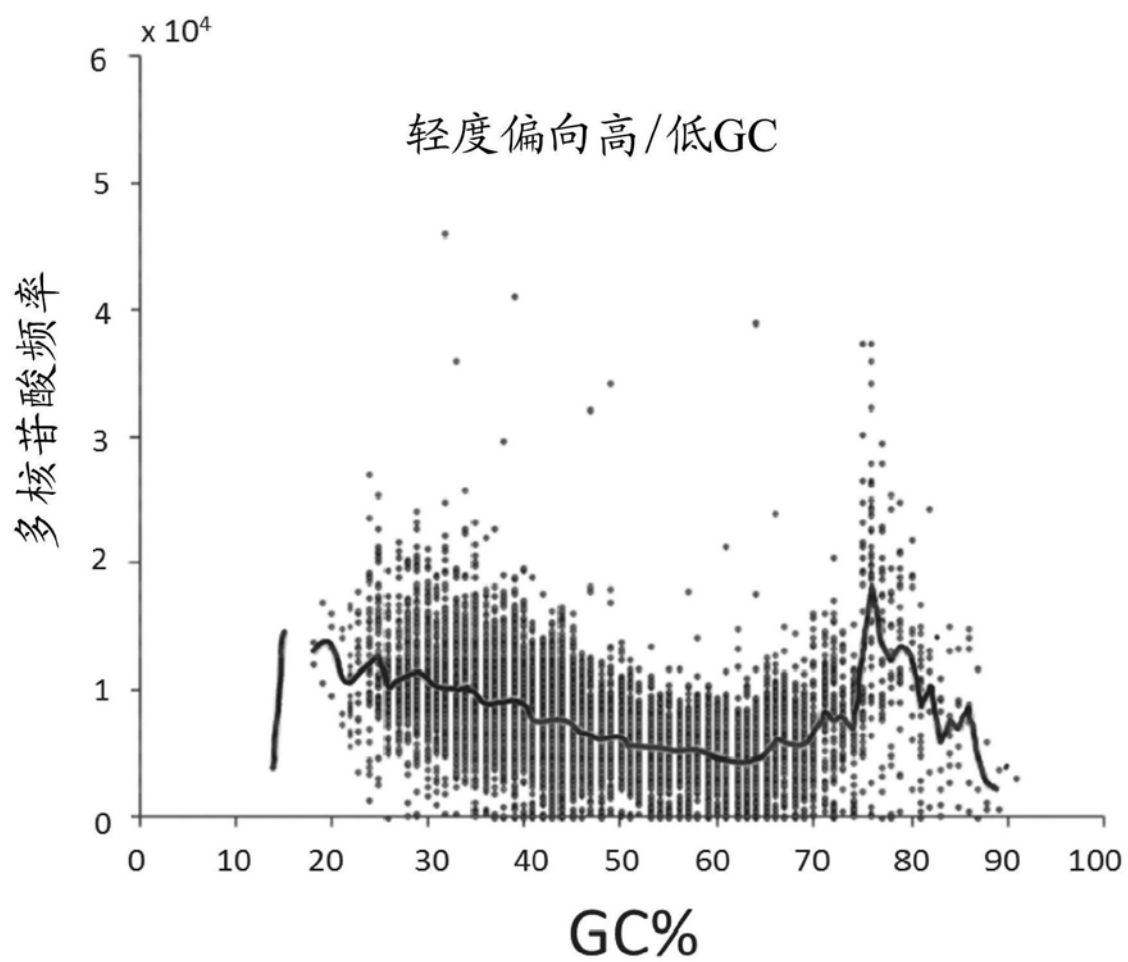


图21C

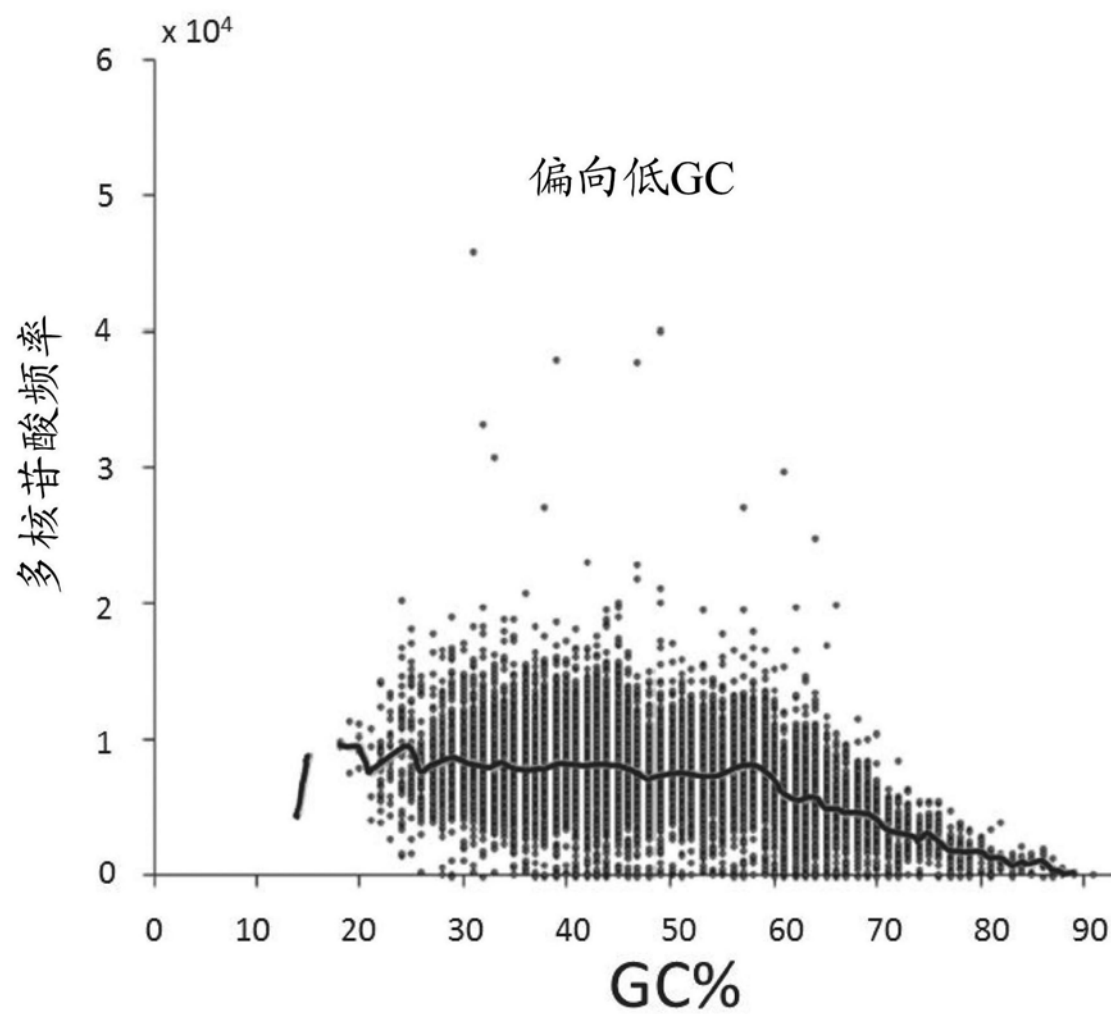


图21D

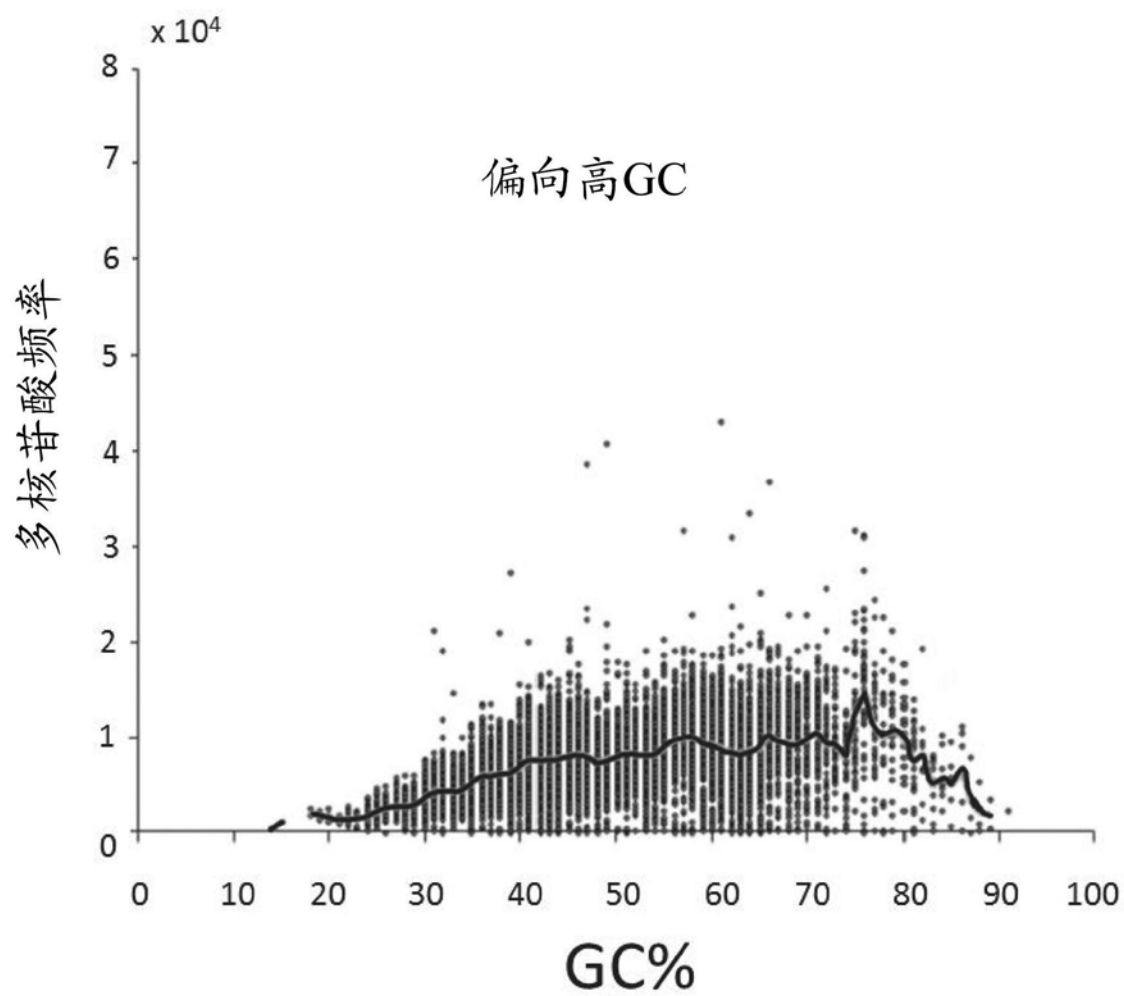


图21E

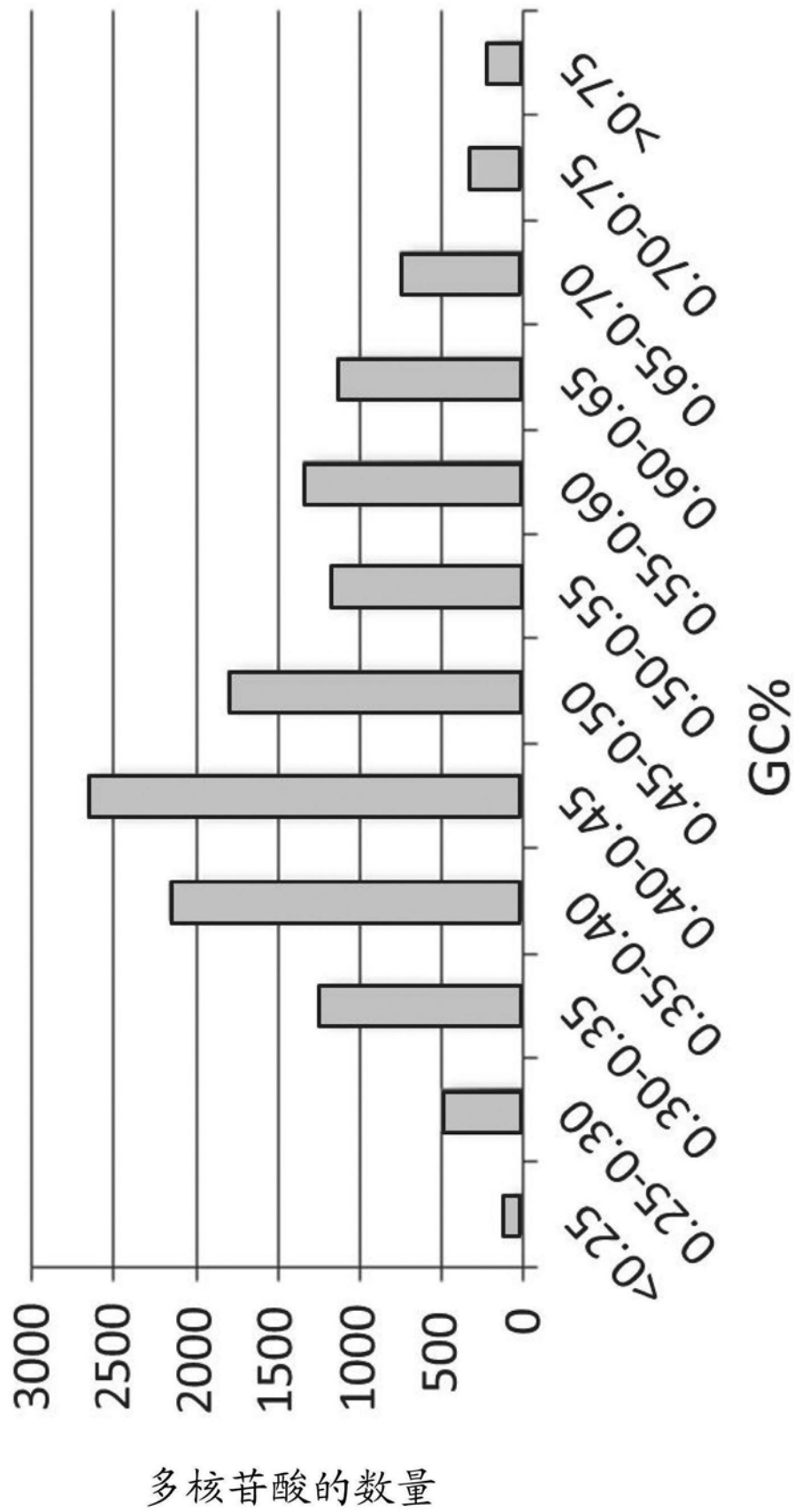


图22

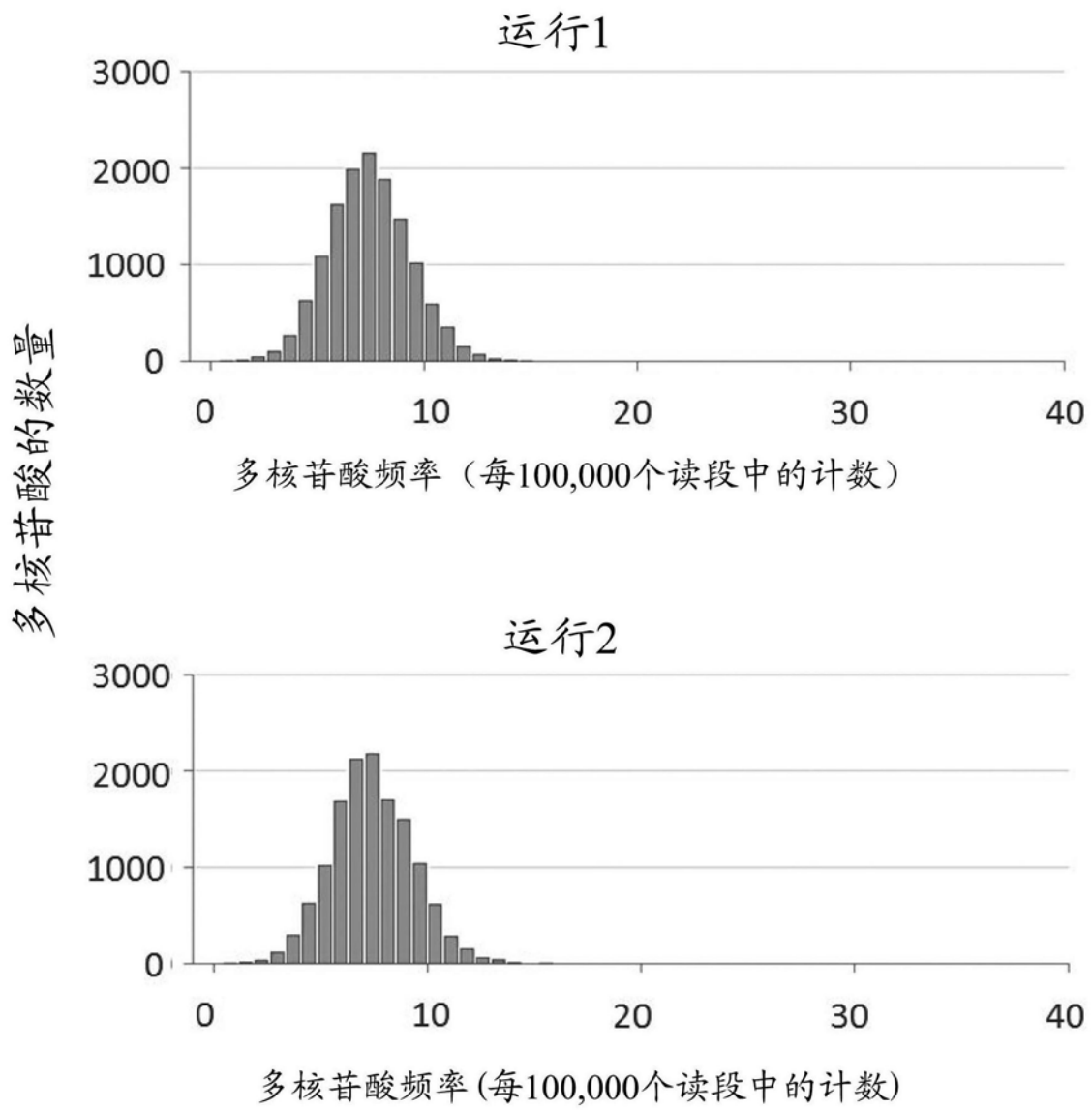


图23

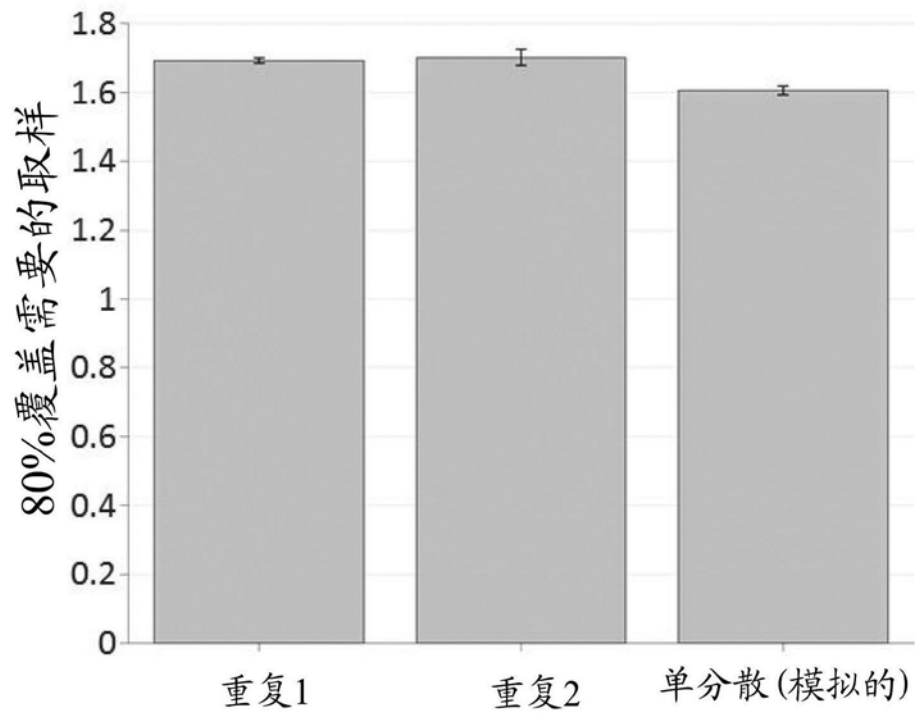


图24A

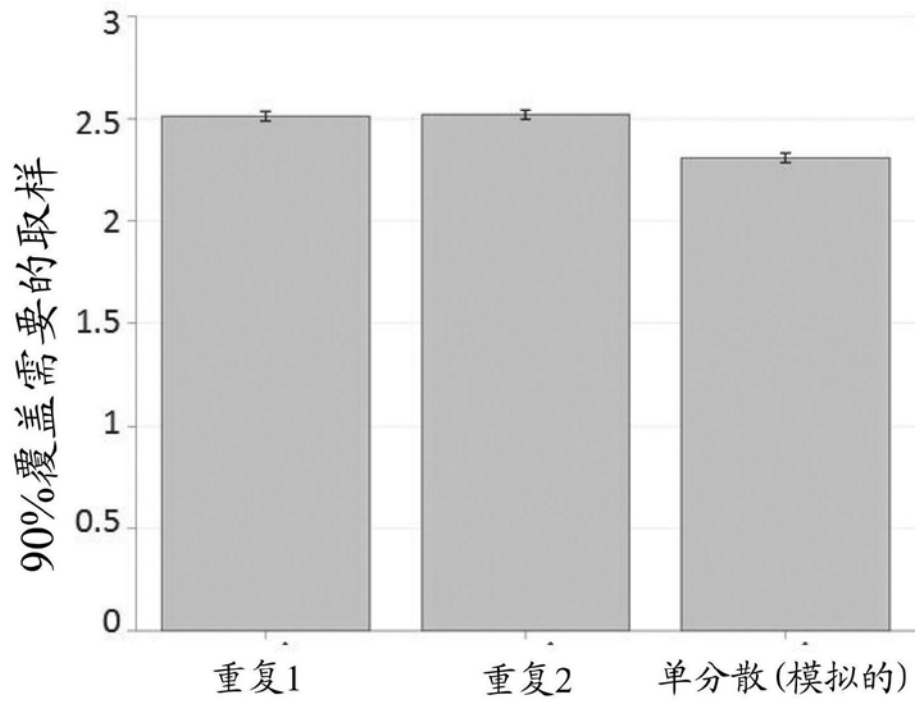


图24B

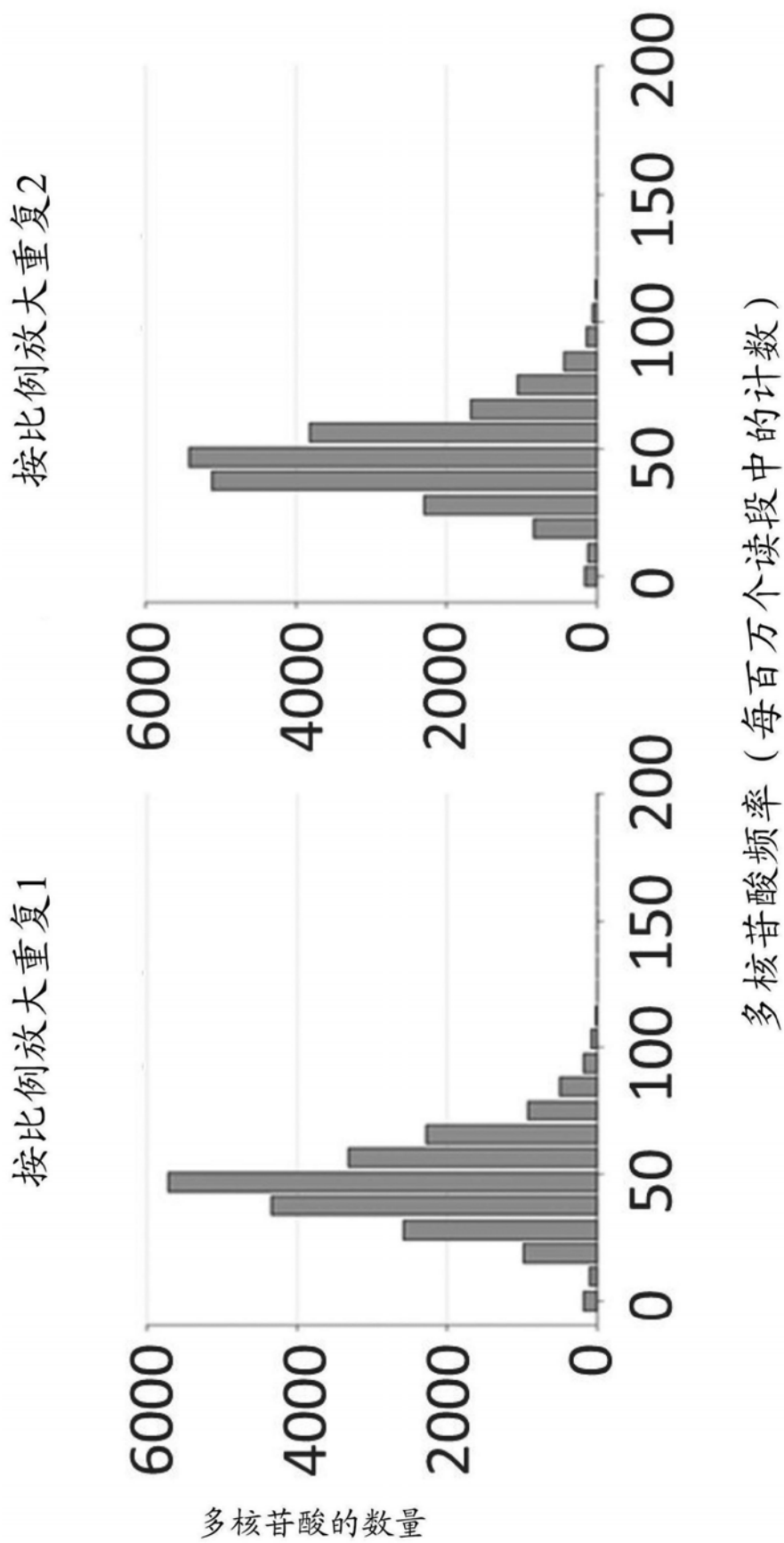


图25

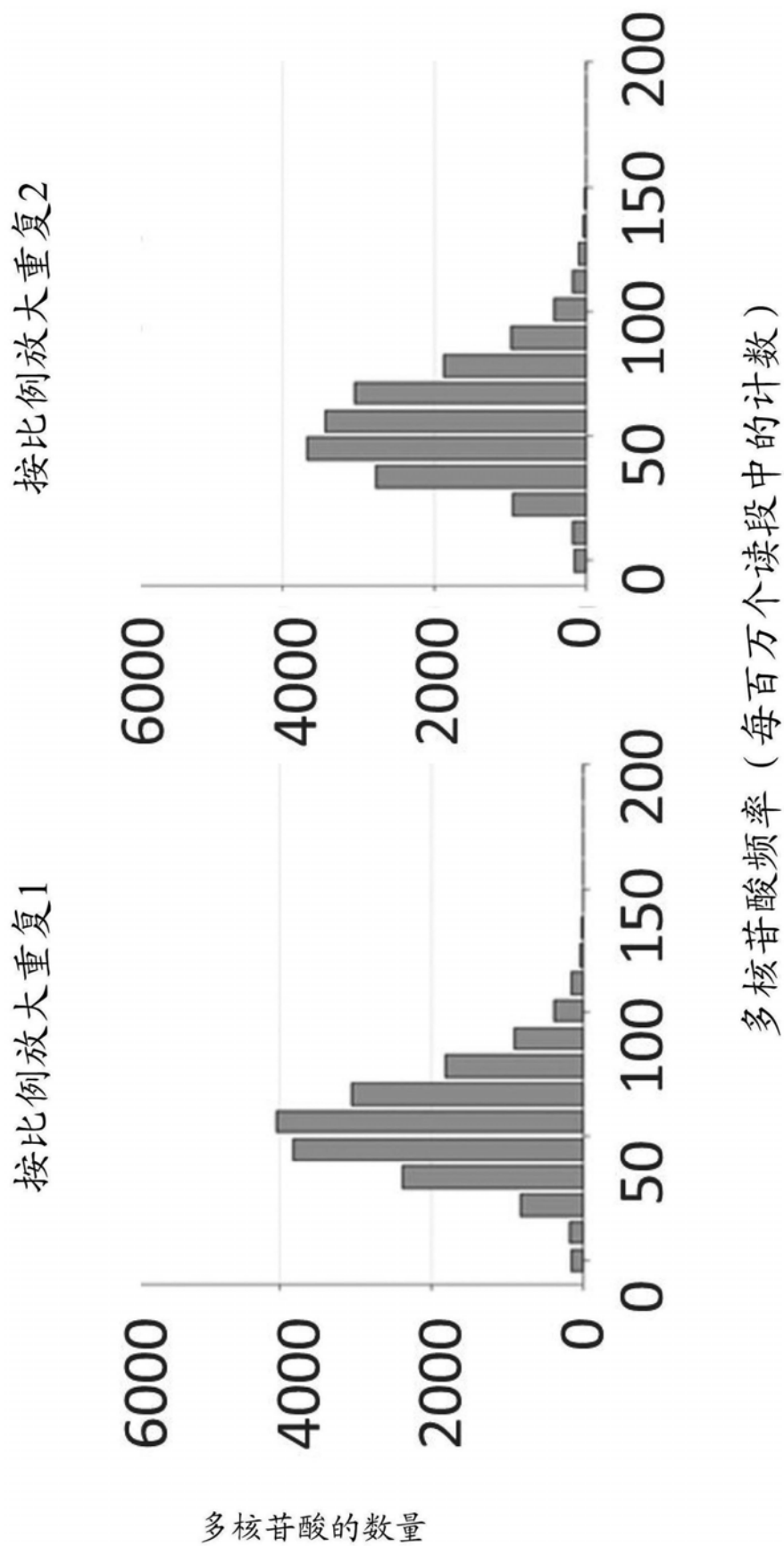


图26

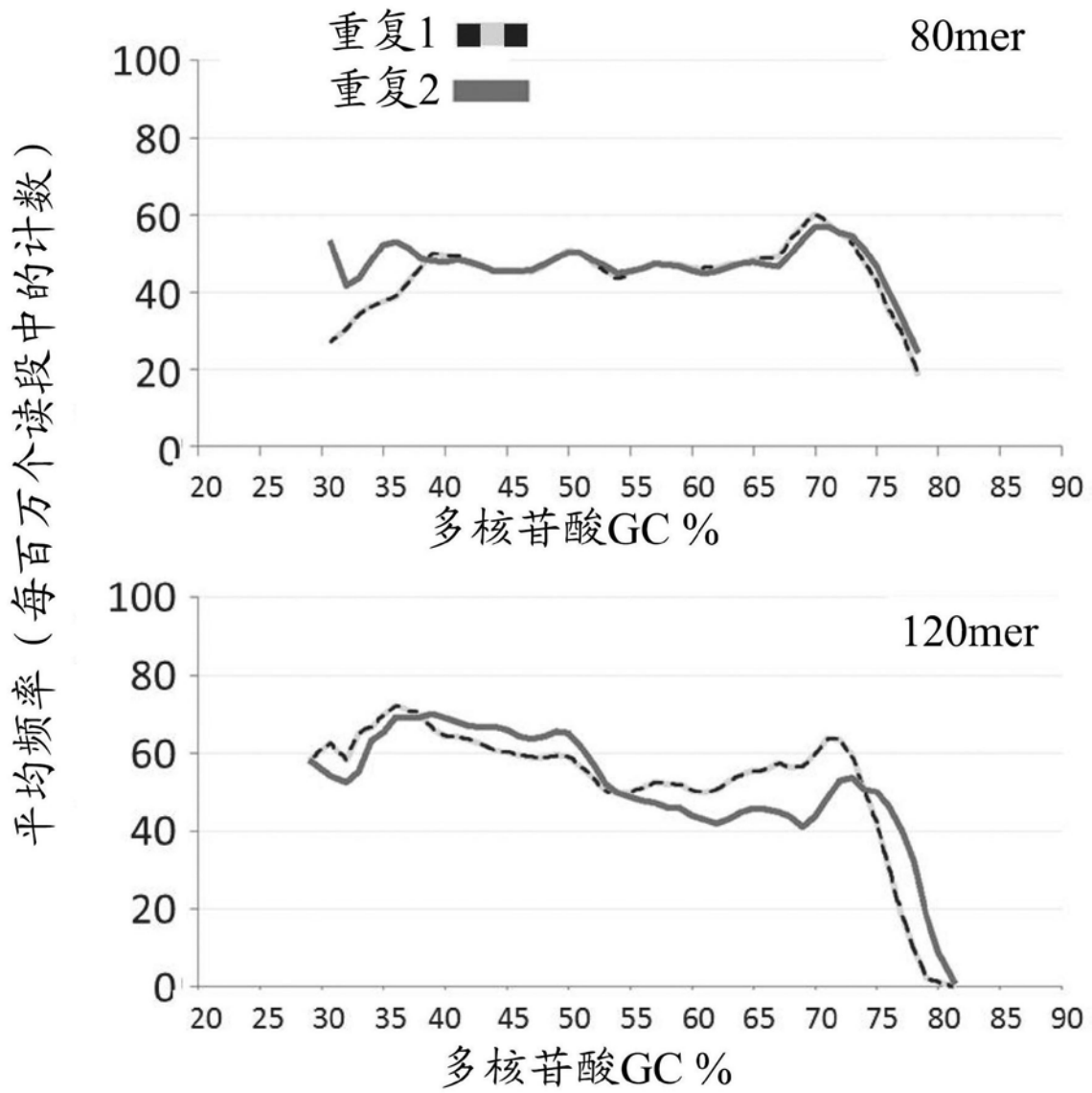


图27

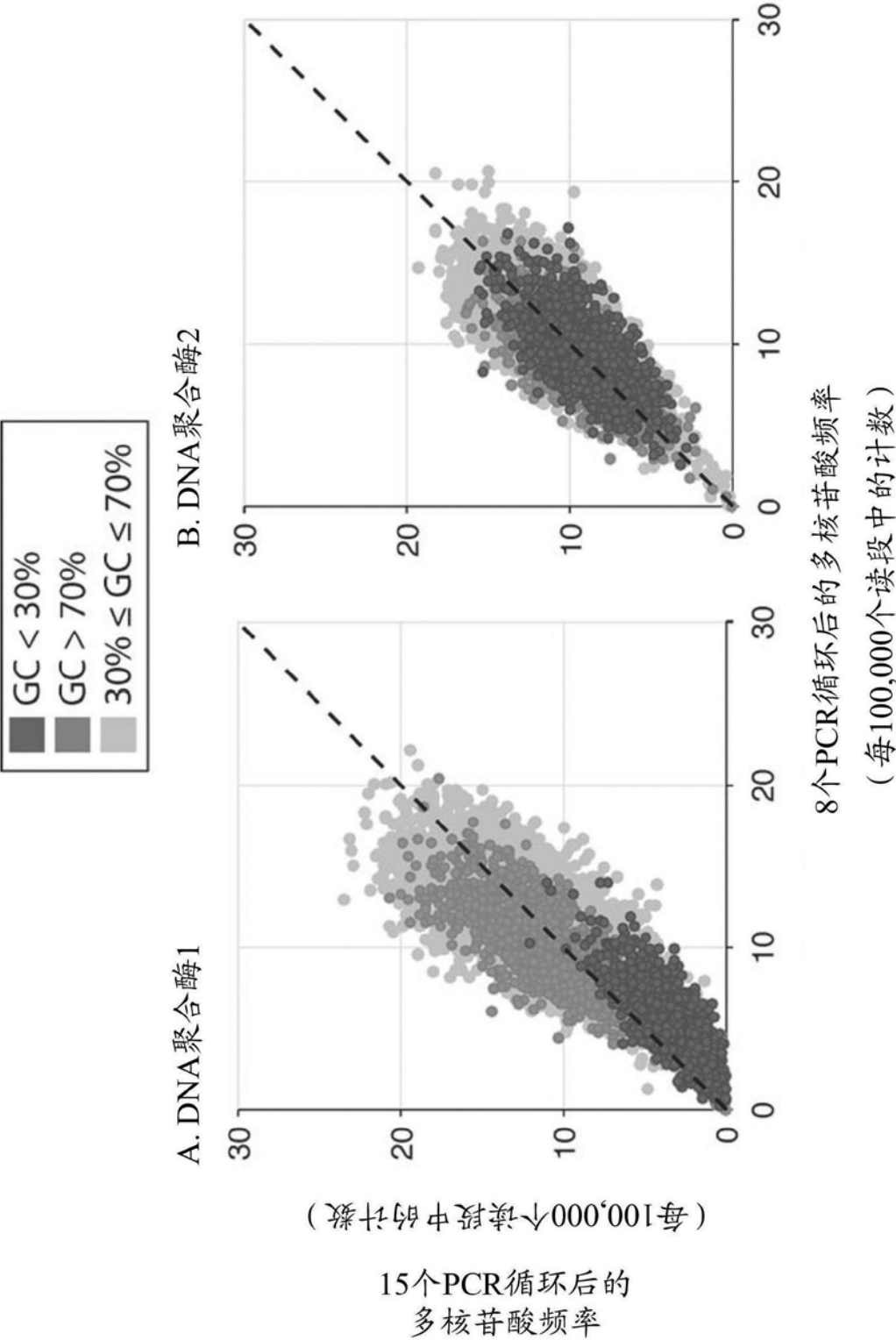


图28

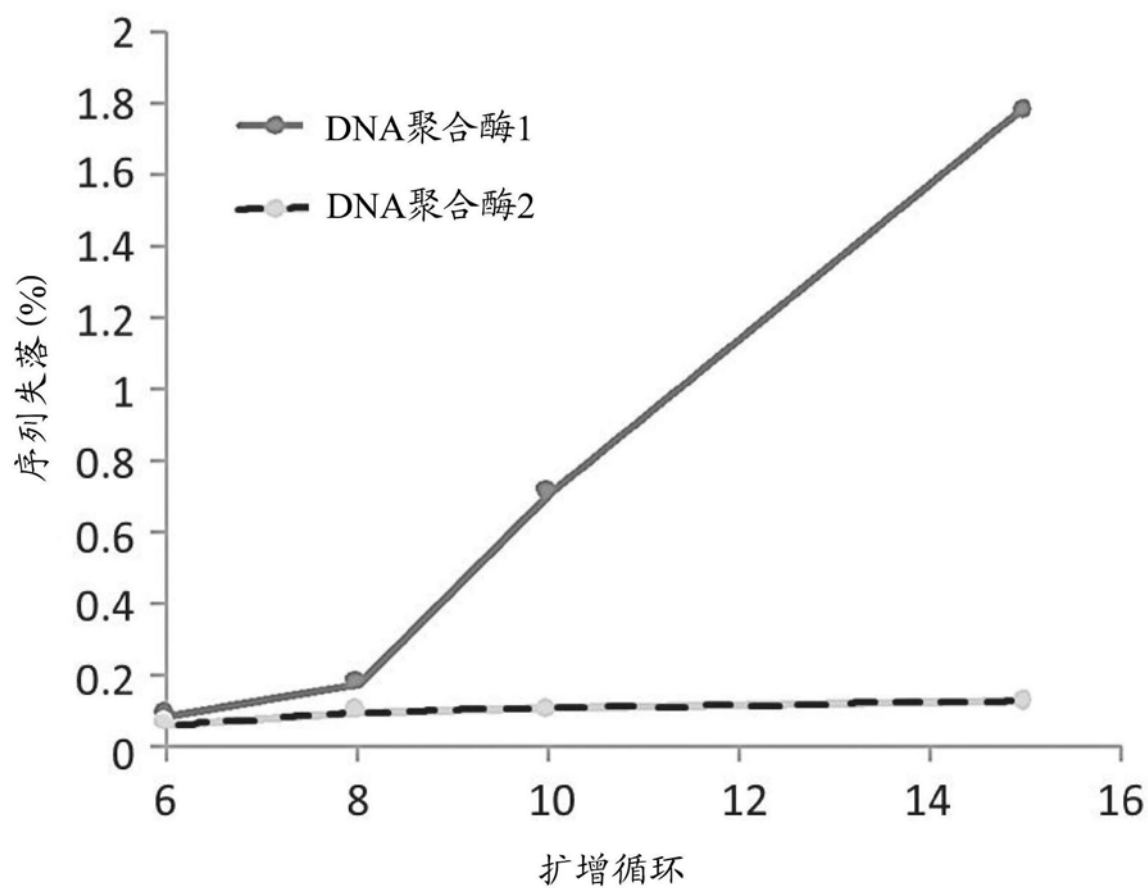


图29

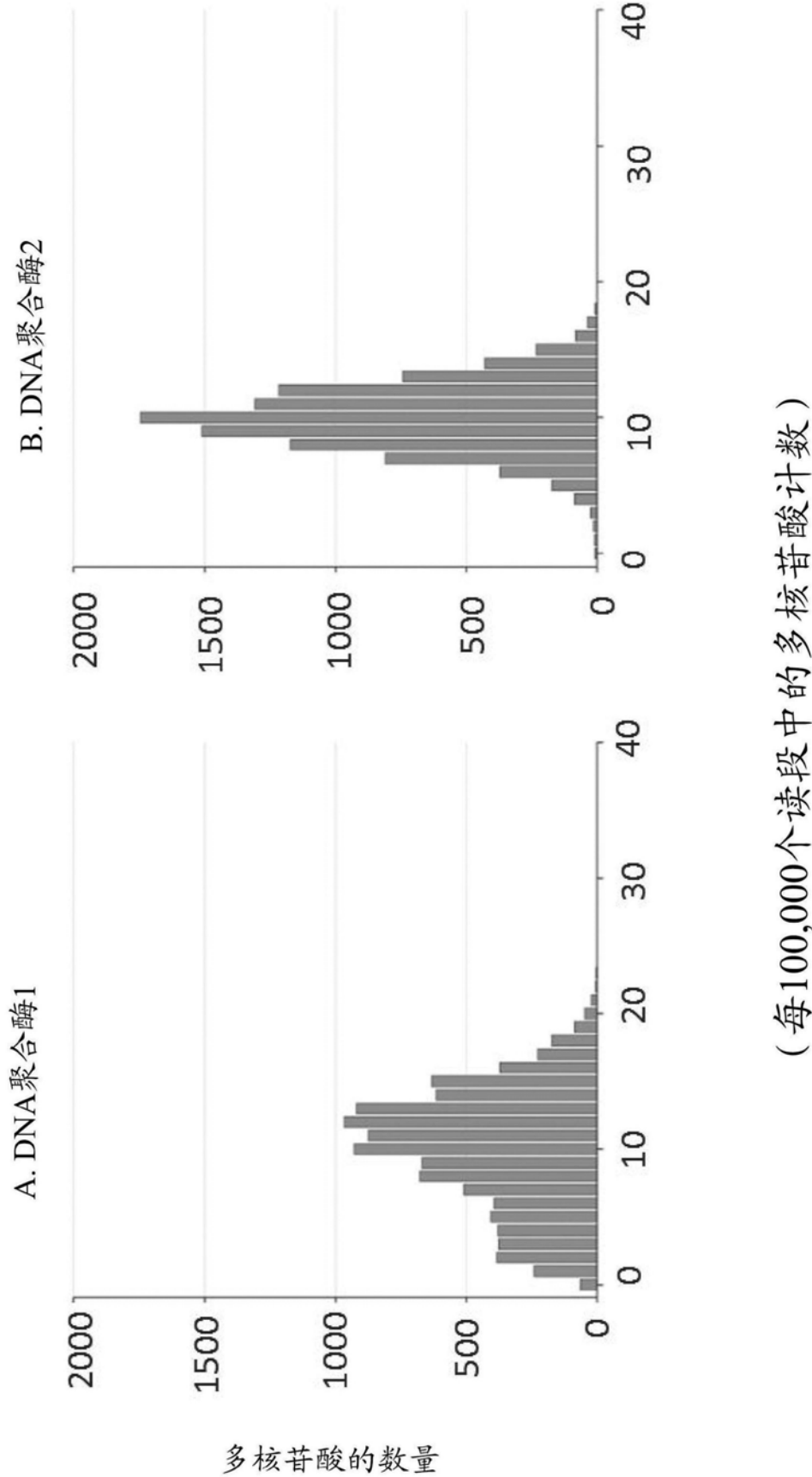


图30

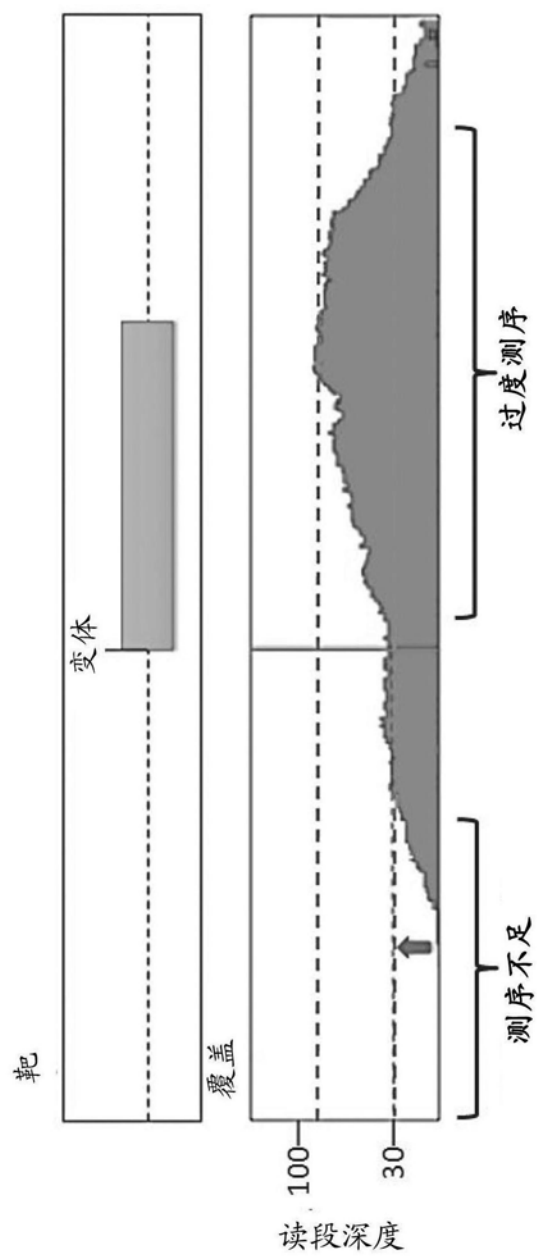


图31A

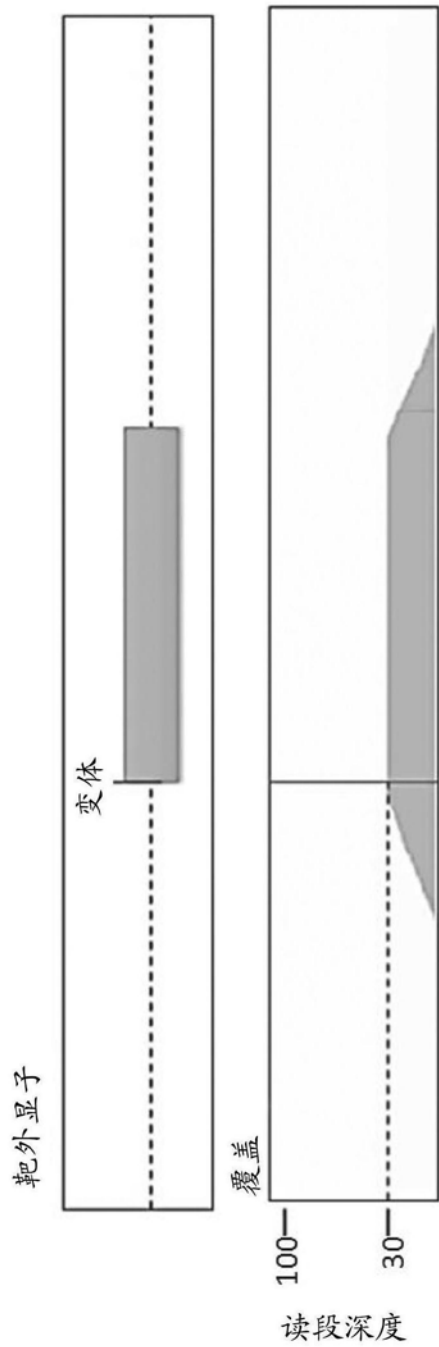


图31B

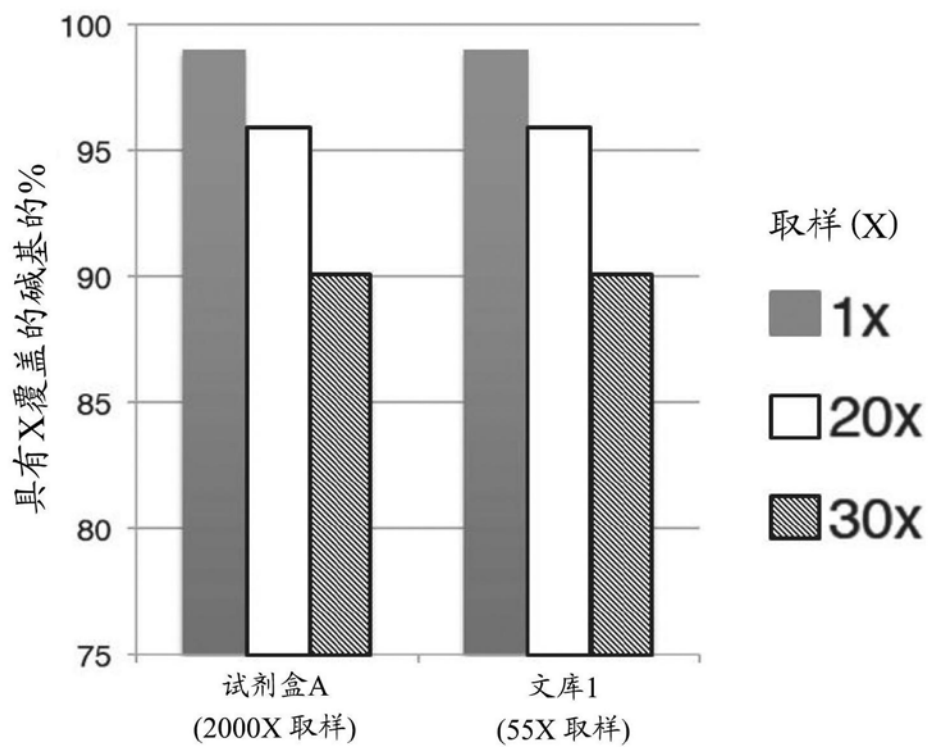


图32A

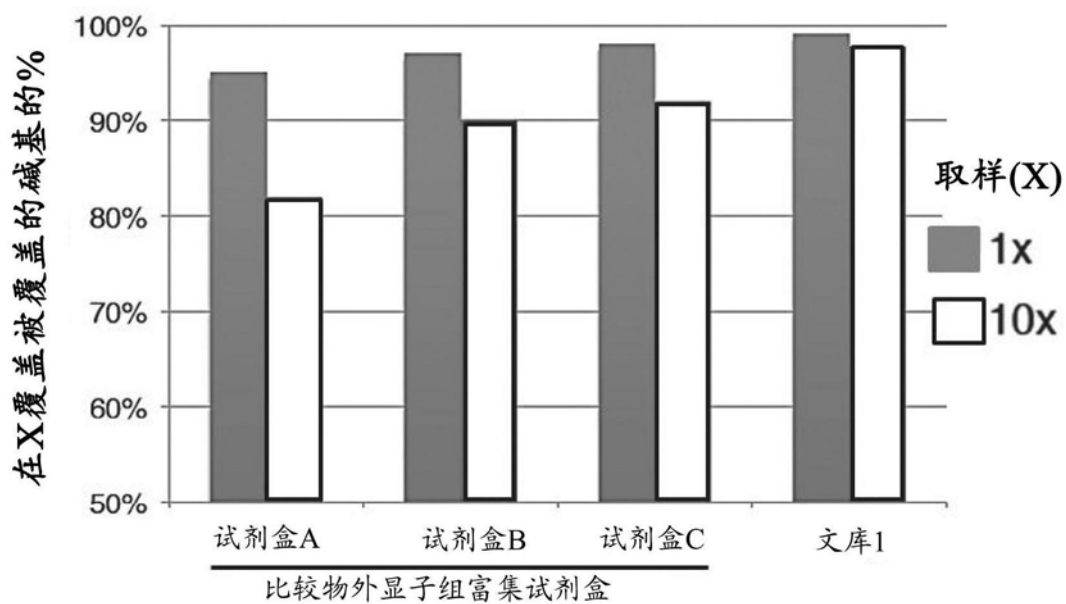


图32B

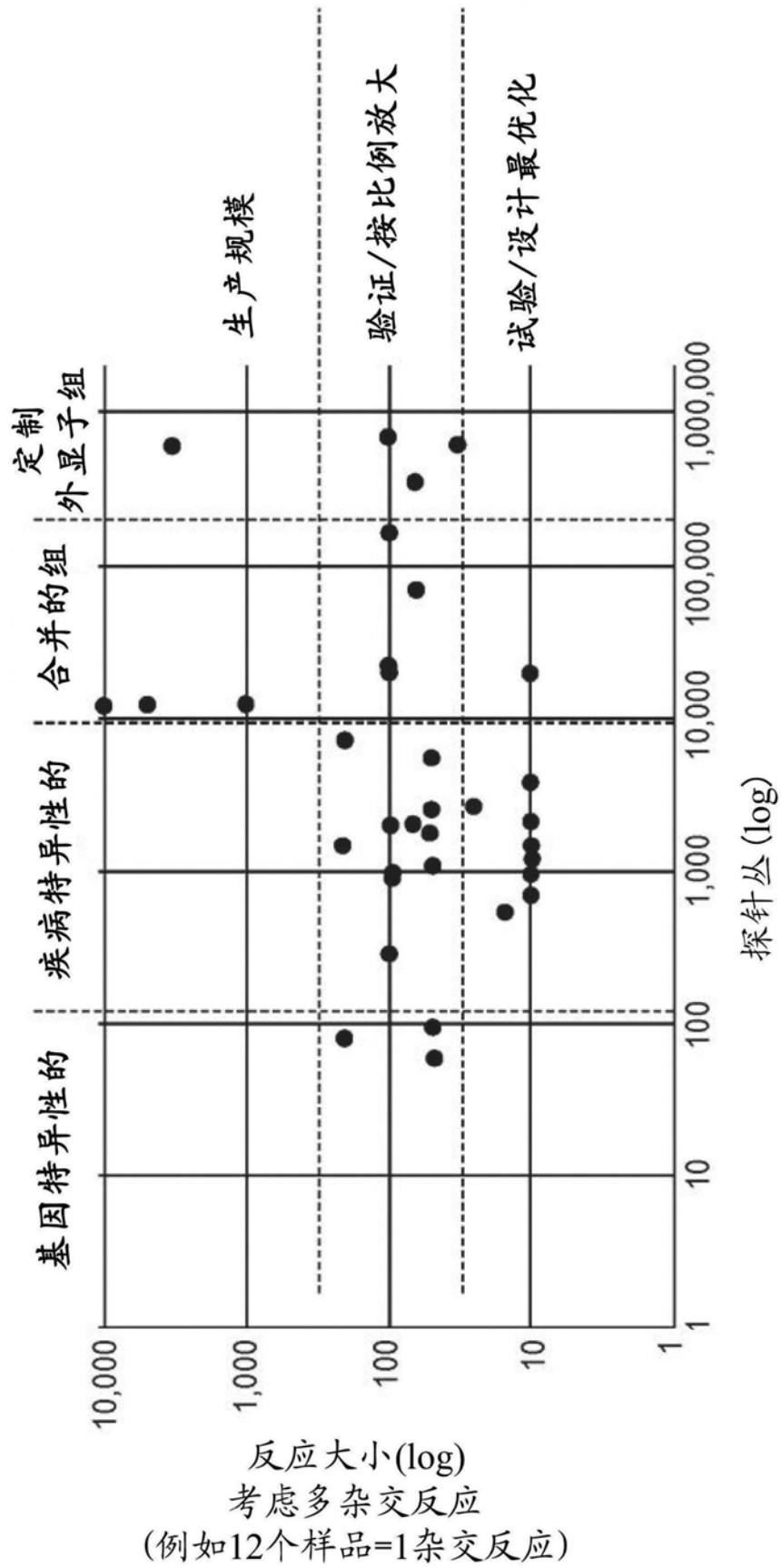


图33

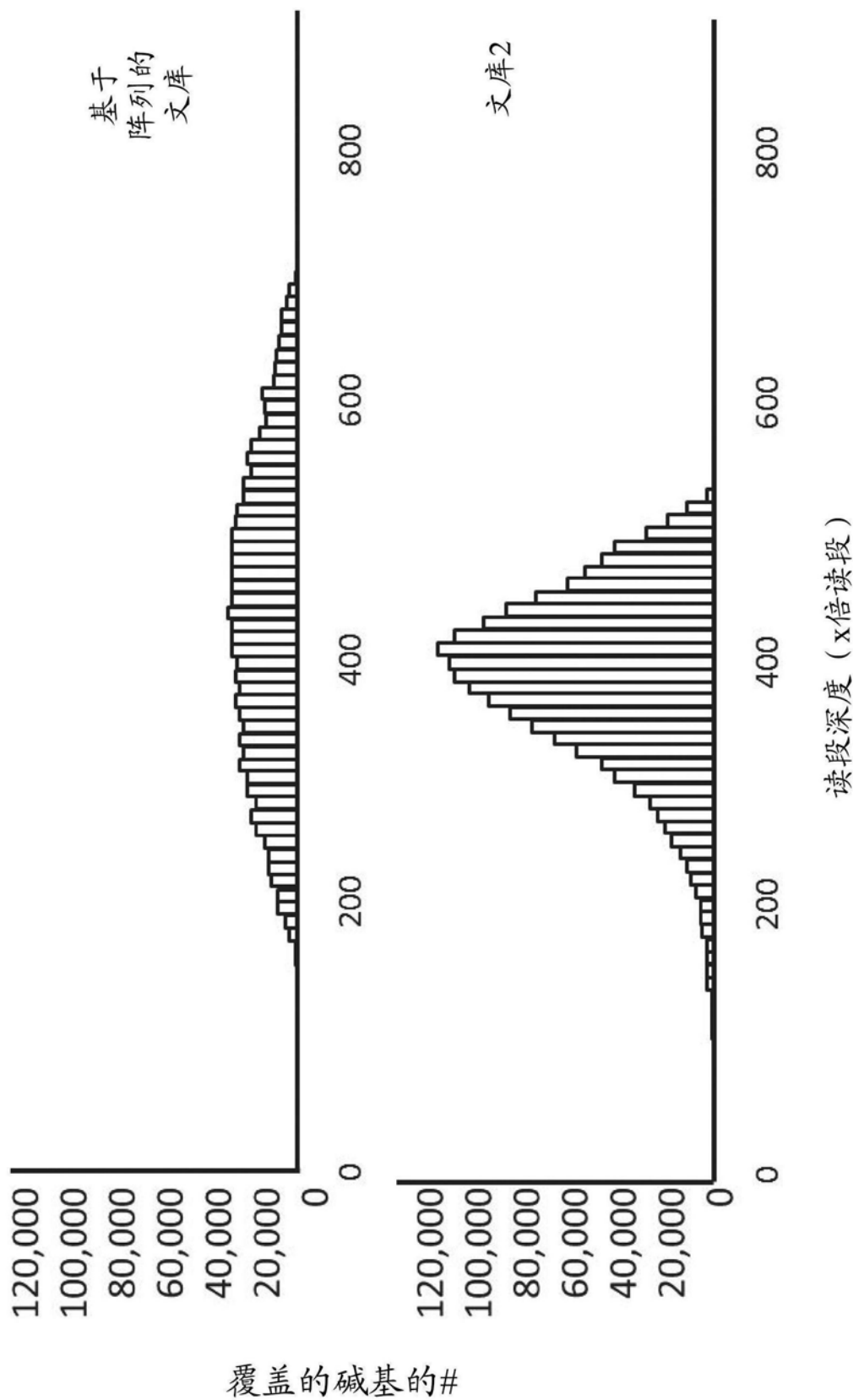


图34

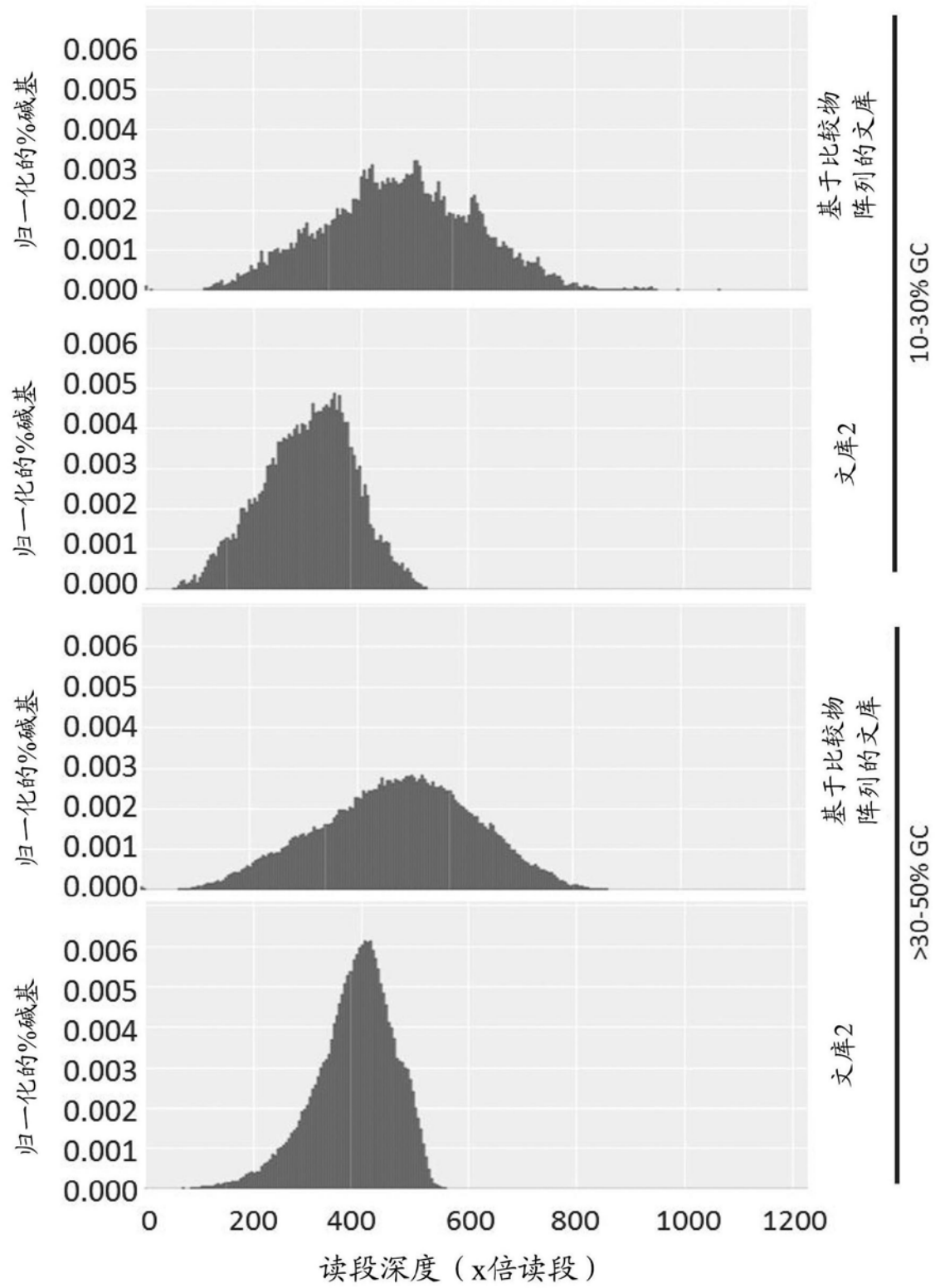


图35A

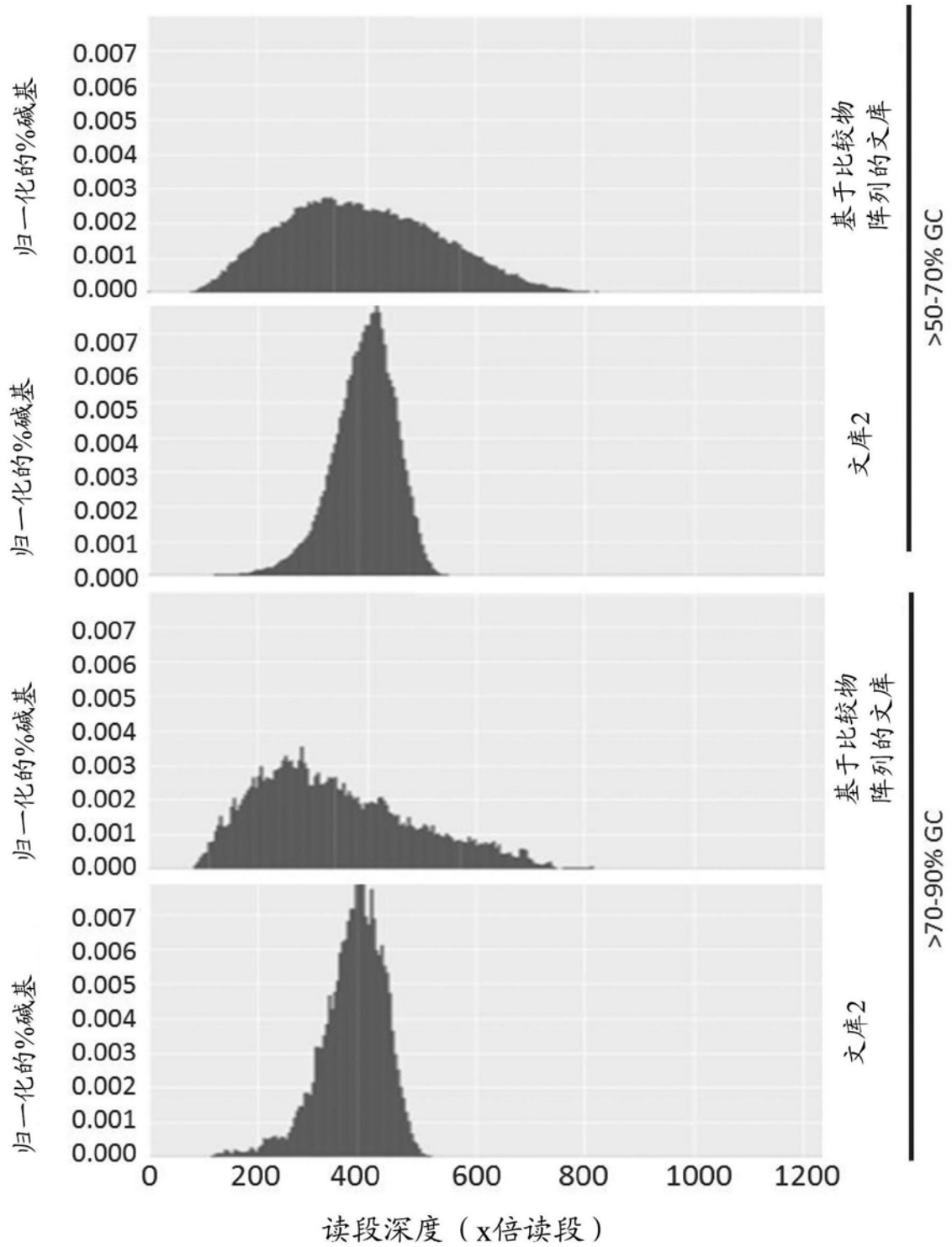


图35B

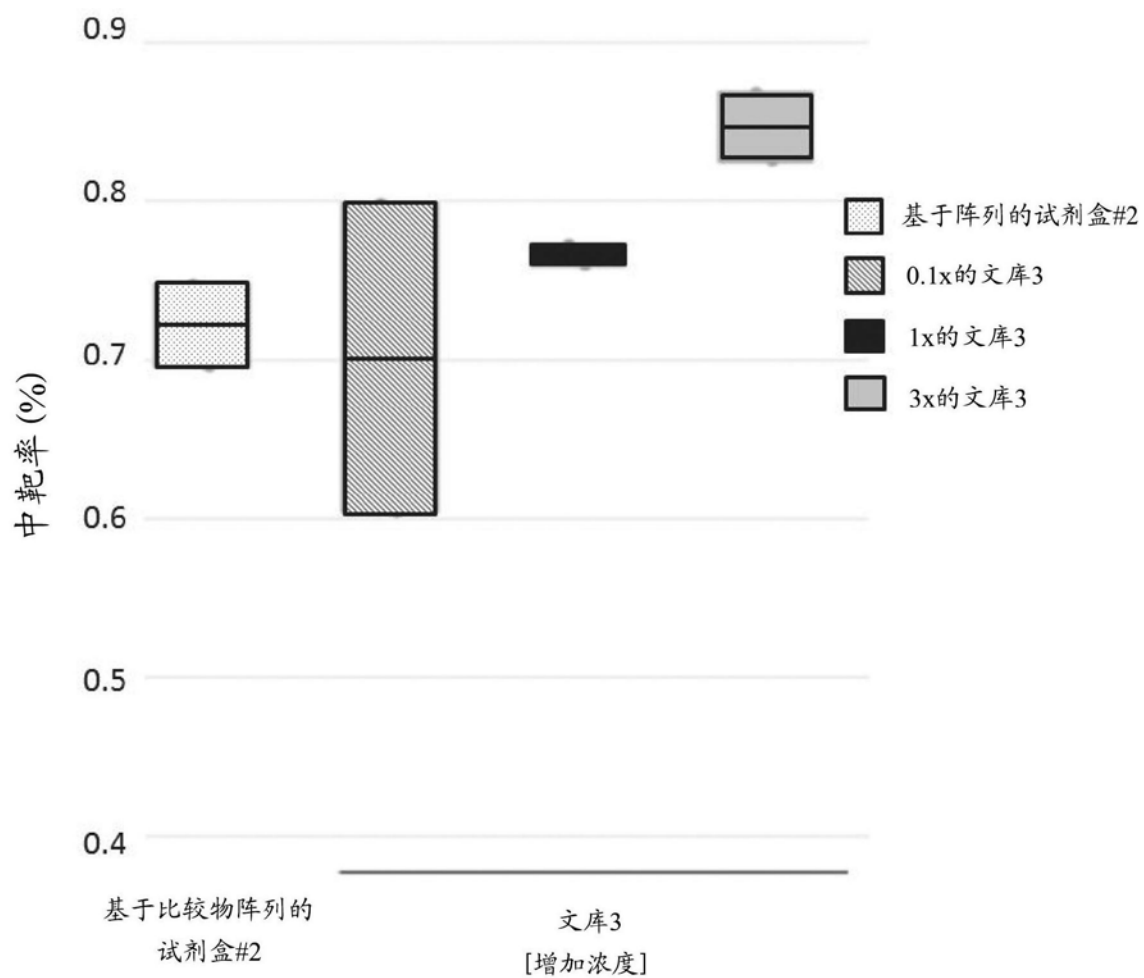


图36A

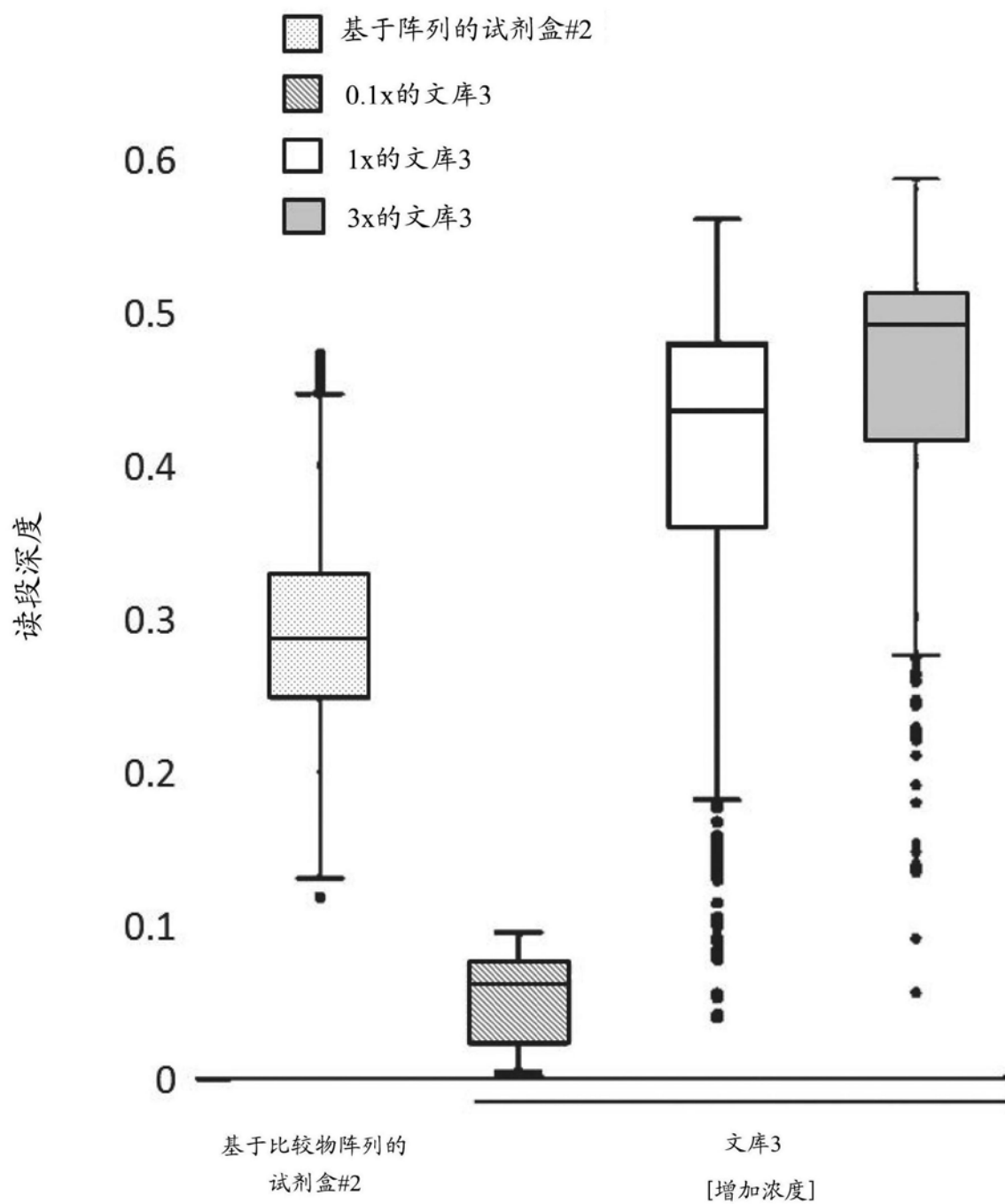


图36B

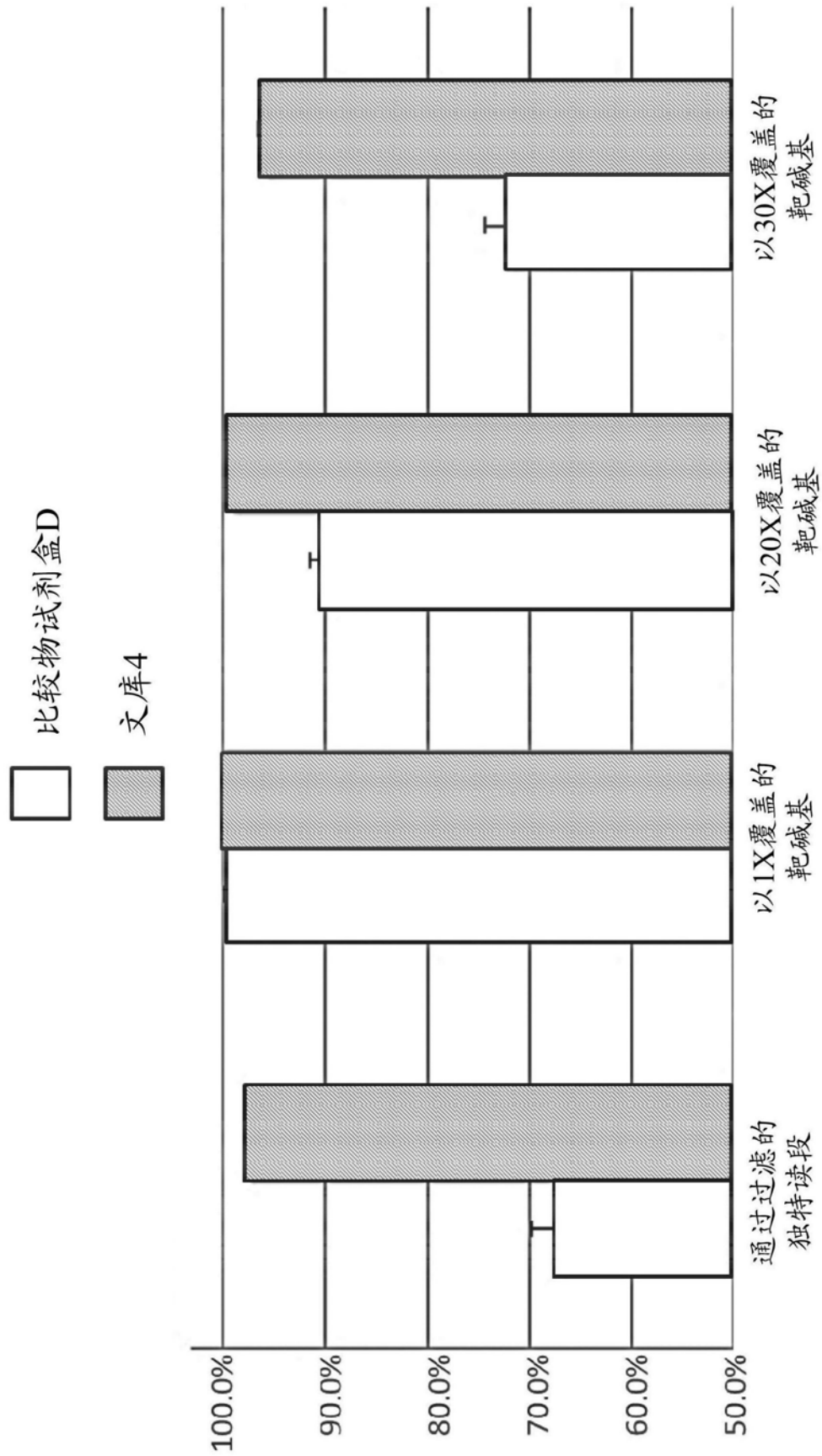


图37

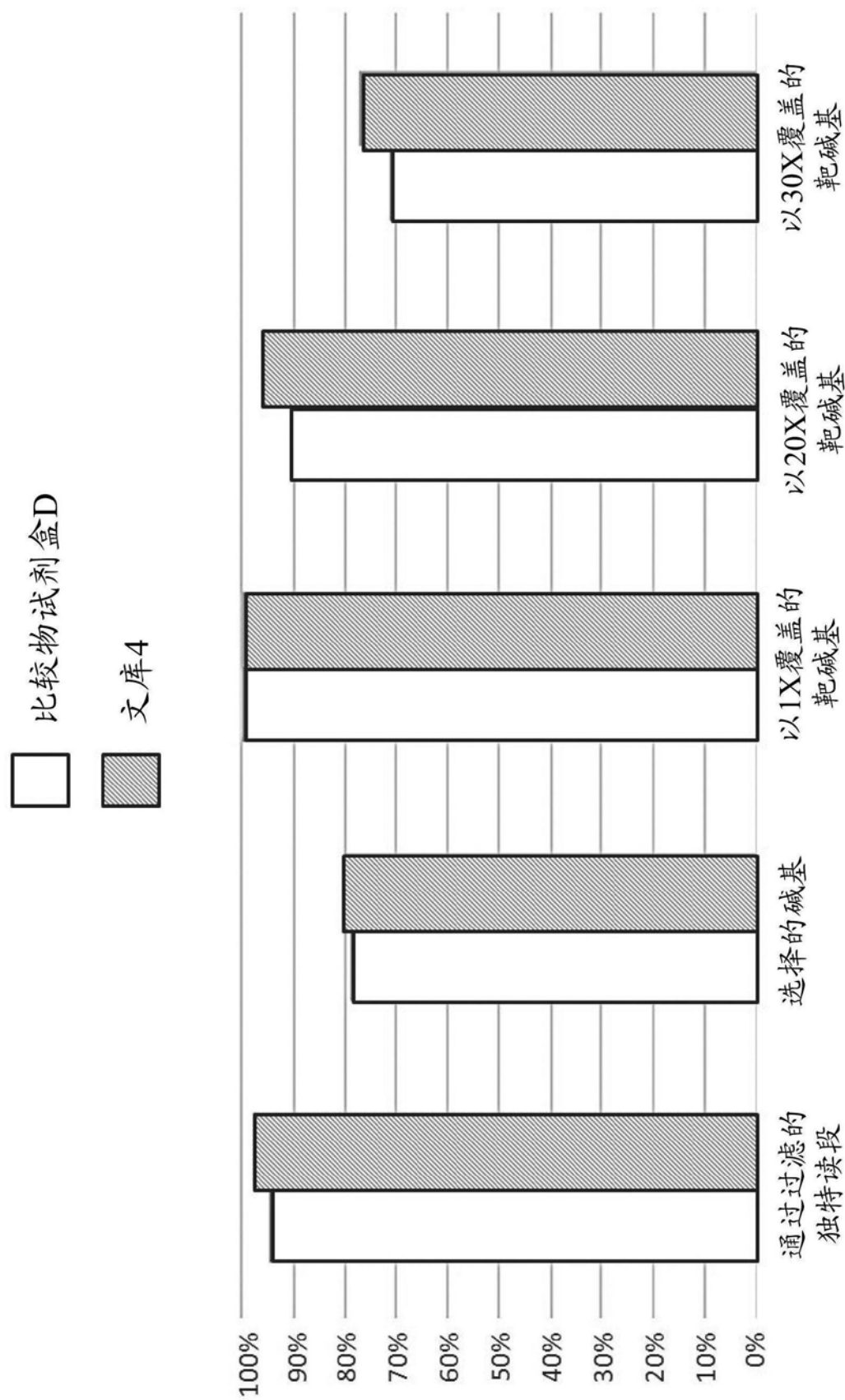


图38