



(10) 授权公告号 CN 113207299 B

(45) 授权公告日 2024. 10. 29

(21) 申请号 201980080514.3

(22) 申请日 2019.10.04

(65) 同一申请的已公布的文献号
申请公布号 CN 113207299 A

(43) 申请公布日 2021.08.03

(30) 优先权数据
62/741,466 2018.10.04 US
62/832,560 2019.04.11 US

(85) PCT国际申请进入国家阶段日
2021.06.04

(86) PCT国际申请的申请数据
PCT/US2019/054843 2019.10.04

(87) PCT国际申请的公布数据
W02020/072990 EN 2020.04.09

(73) 专利权人 阿克生物公司
地址 美国马萨诸塞州

(72) 发明人 E·哈尼斯 V·纳格什
L·G·本利 M·L·卡彭特

(74) 专利代理机构 北京戈程知识产权代理有限公司 11314

专利代理师 程伟

(51) Int.Cl.
C12Q 1/6806 (2006.01)

(56) 对比文件
WO 2015073080 A1,2015.05.21
WO 2016001736 A1,2016.01.07
WO 2017165864 A1,2017.09.28
Mauro D. Locati等.“Improving small RNA-seq by using a synthetic spike-in set for size-range quality control together with a set for data normalization”.《Nucleic Acids Research》.2015,第43卷(第14期),第1-10页.

Michael A Quail等.“SASI-Seq_ sample assurance Spike-Ins, and highly differentiating 384 barcoding for Illumina sequencing”.《BMC Genomics》.2014,第15卷第1-12页.

审查员 宋如生

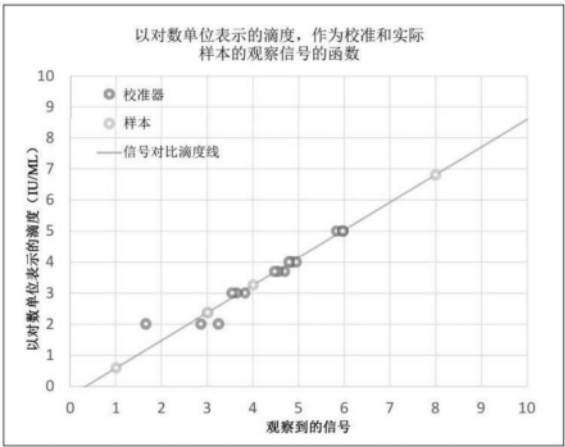
权利要求书5页 说明书114页
序列表6页 附图43页

(54) 发明名称

用于管理下一代测序中的低样本输入的归一化对照

(57) 摘要

本公开提供了用于使用下一代测序对样本中的一个或多个靶核酸序列进行定量的组合物和方法。本公开的方法可用于确定样本中的一种或多种靶生物体的滴度。



1. 一种对样本中的靶生物体的滴度进行定量的方法,所述方法包括:
 - a. 提供包含所述靶生物体的样本,其中所述靶生物体包含至少一个靶序列;
 - b. 提供包含至少三种物种的生物体的已知滴度的多分析物对照,其中所述生物体已被灭活;
 - c. 将已知量的归一化对照物与所述样本和所述多分析物对照混合;所述归一化对照包括至少三组多核苷酸,其中每个组内的多核苷酸具有相同的长度;
 - d. 从所述样本和所述多分析物对照制备高通量测序文库;
 - e. 对所述文库测序以产生样本读段集合和多分析物对照读段集合;
 - f. 使用所述归一化对照将所述样本读段集合和来自(e)的所述多分析物对照读段集合归一化;
 - g. 确定所述多分析物对照中的所述至少三种物种的生物体的归一化读段与已知滴度之间的关系;以及
 - h. 使用(g)中确定的关系计算所述样本中的所述靶生物体的滴度。
2. 根据权利要求1所述的方法,其中所述归一化对照中的所有多核苷酸的长度是相同的。
3. 根据权利要求1所述的方法,其中每个组内的多核苷酸当与任何其他组内的多核苷酸相比时具有不同的长度。
4. 根据权利要求3所述的方法,其中所述多核苷酸组的所述长度以线性序列或几何序列分布,其中所述线性序列是数字序列,其中所述序列中的下一个数字每次相对于前一个数字增加或减少相同的量;所述几何序列是数字序列,其中第一项之后的每个项都可以通过将前一项乘以固定的非零数来找到。
5. 根据权利要求1-4中任一项所述的方法,其中所述至少三个组中的所述多核苷酸的所述长度介于15bp与50,000bp之间。
6. 根据权利要求1-4中任一项所述的方法,其中所述至少三个组中的所述多核苷酸的所述长度介于100bp与1500bp之间。
7. 根据权利要求1-4中任一项所述的方法,其中每个组内的多核苷酸包含相同的序列。
8. 根据权利要求1-4中任一项所述的方法,其中每个组内的多核苷酸不包含相同的序列。
9. 根据权利要求8所述的方法,其中每组多核苷酸包含至少三个多核苷酸亚组,其中每个亚组内的多核苷酸包含相同的序列,并且其中每个亚组内的多核苷酸不包含与任何其他亚组相同的序列。
10. 根据权利要求7所述的方法,其中至少一组多核苷酸中的至少一个序列包含组分序列。
11. 根据权利要求7所述的方法,其中每组多核苷酸中的每个序列均包含组分序列。
12. 根据权利要求10所述的方法,其中每组多核苷酸的组分序列与任何其他组的组分序列不同。
13. 根据权利要求9所述的方法,其中至少一组多核苷酸的至少一个亚组的序列包含组分序列。
14. 根据权利要求9所述的方法,其中至少一个组的所述至少三个亚组中的每一个亚组

的序列包含组分序列。

15. 根据权利要求9所述的方法, 其中每个组的每个亚组的序列包含组分序列。

16. 根据权利要求14或15所述的方法, 其中每个亚组的组分序列与任何其他亚组的组分序列不同。

17. 根据权利要求10所述的方法, 其中所述组分序列包含介于6bp与3000bp之间的序列。

18. 根据权利要求10所述的方法, 其中所述组分序列包含随机序列。

19. 根据权利要求7所述的方法, 其中至少一组多核苷酸的序列包含分离的序列。

20. 根据权利要求7所述的方法, 其中每组多核苷酸的序列包含分离的序列。

21. 根据权利要求9所述的方法, 其中至少一组多核苷酸的至少一个亚组的序列包含分离的序列。

22. 根据权利要求9所述的方法, 其中至少一个组的每个亚组的序列包含分离的序列。

23. 根据权利要求9所述的方法, 其中每个组的每个亚组的序列包含分离的序列。

24. 根据权利要求22所述的方法, 其中每个亚组的分离的序列与任何其他亚组的分离的序列不同。

25. 根据权利要求19所述的方法, 其中所述分离的序列包含介于6bp与200,000bp之间。

26. 根据权利要求19所述的方法, 其中所述分离的序列分离自或来源于病毒、细菌、真菌或真核寄生虫。

27. 根据权利要求19所述的方法, 其中所述分离的序列与测序样本中的至少一个靶序列不同。

28. 根据权利要求19所述的方法, 其中所述归一化对照中的多核苷酸组的至少一个亚组的至少一个序列与所述测序样本中的所述至少一个靶序列共享至少一种序列特性。

29. 根据权利要求28所述的方法, 其中所述至少一个序列特性包括GC含量百分比、熵、复杂性、长度、电子-离子相互作用势(EIIP), 转座元件序列、病毒序列、细菌序列、真菌序列、真核寄生虫序列、一个或多个人类基因序列的序列特性, 或它们的组合。

30. 根据权利要求28所述的方法, 其中所述测序样本包含宿主和非宿主核酸的混合物。

31. 根据权利要求30所述的方法, 其中所述宿主是真核生物。

32. 根据权利要求30所述的方法, 其中所述宿主是昆虫、植物或动物。

33. 根据权利要求32所述的方法, 其中所述动物是人。

34. 根据权利要求30所述的方法, 其中所述非宿主包括共生生物、寄生虫或病原体, 其中所述方法不以疾病诊断为目的。

35. 根据权利要求34所述的方法, 其中所述非宿主包括多个物种。

36. 根据权利要求30所述的方法, 其中所述非宿主包括病毒物种、细菌物种、真菌物种或真核寄生虫物种, 或它们的组合, 其中所述方法不以疾病诊断为目的。

37. 根据权利要求27所述的方法, 其中所述至少一个靶序列是所述测序样本中的非宿主序列。

38. 根据权利要求27所述的方法, 其中所述至少一个靶序列包含多个参考序列。

39. 根据权利要求38所述的方法, 其中所述多个参考序列包括至少2个参考序列。

40. 根据权利要求39所述的方法, 其中所述参考序列包含非宿主序列。

41. 根据权利要求39或40所述的方法, 其中所述参考序列包含病毒序列、细菌序列、真菌序列, 或它们的组合, 其中所述方法不以疾病诊断为目的。

42. 根据权利要求38所述的方法, 其中所述多个参考序列包括来自多种生物的参考序列。

43. 根据权利要求42所述的方法, 其中所述参考序列包含基因组DNA序列或cDNA序列, 或它们的组合。

44. 根据权利要求38所述的方法, 其中所述分离的序列包含至少一个参考序列的至少一个片段, 所述至少一个片段已被改组。

45. 根据权利要求44所述的方法, 其中所述分离的序列包含至少第一参考序列和第二参考序列的至少两个片段, 所述至少两个片段已被改组和连接。

46. 根据权利要求45所述的方法, 其中所述第一参考序列和第二参考序列在生物体的基因组中不相邻。

47. 根据权利要求45所述的方法, 其中所述第一参考序列和第二参考序列来自不同的生物体。

48. 根据权利要求44所述的方法, 其中所述参考序列的所述至少一个片段包含所述参考序列的10-100个连续核苷酸。

49. 根据权利要求39所述的方法, 其中:

通过参考序列片段的分布表示所述多个参考序列;

将所述分布划分为至少5个箱; 并且

从所述至少5个箱中的至少3个箱选择多个参考序列片段, 进行改组, 并任选地连接, 从而生成分离的序列。

50. 根据权利要求49所述的方法, 其中所述参考序列片段是关于参考序列长度和以下中的至少一者分布的: (1) GC含量百分比、(2) 熵、(3) 复杂性、(4) EIIP, 或它们的组合。

51. 根据权利要求49所述的方法, 其中每个箱代表所述分布的1%、2%、5%、10%、15%、或20%。

52. 根据权利要求49所述的方法, 其中每个箱代表GC含量的分布的5%、10%、15%、或20%。

53. 根据权利要求50所述的方法, 其中所述归一化对照的每个组内的多核苷酸是通过改组和任选地连接选自所述至少3个箱中的一个箱的参考序列片段而生成的。

54. 根据权利要求51所述的方法, 其中每个组内的多核苷酸具有以下项中的至少三者: 与来自所述参考序列分布中的对应箱的参考序列片段 (1) 相似的GC含量百分比、(2) 相似的熵、(3) 相似的EIIP、(4) 相似的长度、或 (5) 相似的复杂性。

55. 根据权利要求50所述的方法, 其中参考序列片段选自来自所述GC含量百分比的分布的第10、第30、第50、第70、第80和第90百分位箱。

56. 根据权利要求38所述的方法, 其中所述分离的序列包含连接的至少一个参考序列的至少两个片段, 并且其中所述至少两个片段在所述至少一个参考序列中不相邻。

57. 根据权利要求38所述的方法, 其中所述分离的序列包含至少第一参考序列的至少第一片段和至少第二参考序列的至少第二片段, 其中所述至少第一和第二片段是连接的, 并且其中所述第一和第二参考序列不是相同的参考序列。

58. 根据权利要求57所述的方法, 其中所述至少第一和第二片段是所述对应参考序列的15-60个连续核苷酸。

59. 根据权利要求57所述的方法, 其中所述至少第一和第二片段是所述对应参考序列的29个、30个、31个、32个、33个或34个连续核苷酸。

60. 根据权利要求57所述的方法, 其中所述至少第一和第二片段小于测序读段的平均长度的一半。

61. 根据权利要求56所述的方法, 其中:

通过使用滑动窗口生成的重叠参考序列片段的分布表示所述至少一个参考序列;

将所述分布划分为至少5个箱; 并且

从所述至少5个箱中的至少3个箱中选择多个参考序列片段并进行连接, 从而生成分离的序列。

62. 根据权利要求61所述的方法, 其中所述参考序列片段是关于参考序列长度和以下中的至少一者分布的: (1) GC含量百分比、(2) 熵、(3) 复杂性、(4) EIIP, 或它们的组合。

63. 根据权利要求61所述的方法, 其中每个箱代表所述分布的1%、2%、5%、10%、15%、或20%。

64. 根据权利要求61所述的方法, 其中每个箱代表GC含量的分布的5%、10%、15%、或20%。

65. 根据权利要求61所述的方法, 其中每个组内的多核苷酸是通过连接选自所述至少3个盒中的一个盒的参考序列片段而生成的。

66. 根据权利要求61所述的方法, 其中每个组内的多核苷酸具有以下项中的至少三者: 与来自所述参考序列分布中的对应箱的参考序列片段 (1) 相似的GC含量百分比、(2) 相似的熵、(3) 相似的EIIP、(4) 相似的长度、或 (5) 相似的复杂性。

67. 根据权利要求61所述的方法, 其中所述滑动窗口包括1bp、2bp、3bp、4bp或5bp的滑动窗口。

68. 根据权利要求62所述的方法, 其中参考序列片段选自来自所述GC含量百分比的分布的第10、第30、第50、第70、第80和第90百分位箱。

69. 根据权利要求27所述的方法, 其中所述测序样本包括来自被诊断为患有癌症的受试者的样本。

70. 根据权利要求27所述的方法, 其中所述测序样本包含宏基因组样本。

71. 根据权利要求70所述的方法, 其中所述至少一个靶序列包含所述宏基因组样本内的物种的序列。

72. 根据权利要求1-4中任一项所述的方法, 其中每个组的多核苷酸具有相同的浓度。

73. 根据权利要求1-4中任一项所述的方法, 其中每个组的多核苷酸不具有相同的浓度。

74. 根据权利要求1-4中任一项所述的方法, 其中一个组内的多核苷酸具有相同的GC含量百分比。

75. 根据权利要求74所述的方法, 其中一个组内的多核苷酸不与任何其他组内的多核苷酸具有相同的GC含量百分比。

76. 根据权利要求1-4中任一项所述的方法, 其中所述多核苷酸包含编码唯一分子标识

符的序列。

77. 根据权利要求76所述的方法, 其中编码所述唯一分子标识符的所述序列包括第一索引序列或第二索引序列。

78. 根据权利要求1-4中任一项所述的方法, 其中所述多核苷酸包含脱氧核糖核酸(DNA)分子、核糖核酸(RNA)分子、或DNA-RNA杂交分子。

79. 根据权利要求78所述的方法, 其中在所述归一化对照中, 所述多核苷酸的浓度介于0.005皮克(pg)/微升(μL)与500pg/ μL 之间。

80. 根据权利要求1-4中任一项所述的方法, 所述归一化对照还包含至少一种蛋白质。

81. 根据权利要求80所述的方法, 其中所述至少一种蛋白质分离自或来源于细菌、病毒或真核寄生虫。

用于管理下一代测序中的低样本输入的归一化对照

[0001] 相关申请

[0002] 本申请要求于2018年10月4日提交的美国临时申请号62/741,466和2019年4月11日提交的美国临时申请号62/832,560的优先权和权益,所述美国临时申请中的每一者的内容以全文引用方式并入本文。

技术领域

[0003] 本发明涉及使用高通量测序对样本中的核酸进行定量的方法和组合物。

[0004] 通过引用并入序列表

[0005] 本申请与电子格式的序列表一起提交。序列表作为名为ARCB-006-02W0_SeqList.txt的文件提供,创建于2019年10月3日,并且大小为5943字节。序列表的电子格式信息以引用方式整体并入。

背景技术

[0006] 下一代测序(next generation sequencing,NGS),也称为高通量测序,是用于研究和诊断的强大工具。NGS技术允许大规模的核酸高通量测序。使用NGS技术,可以廉价高效地对来自单个样本的数百万个DNA或RNA分子(多核苷酸)进行测序。在NGS测序期间取样多核苷酸的程度与样本中多核苷酸的初始浓度成正比。因此,NGS可用于确定测序样本中单个多核苷酸或多核苷酸群体的数量。然而,由NGS生成的读段数量和样本中多核苷酸的起始浓度不一定很好地相关。当从低样本输入开始时,使NGS读段数和起始浓度相关就特别成问题,因为样本制备和测序期间的随机过程在起始样本输入低时会对NGS过程中测序的核酸分子的数量具有不成比例的影响。在各种临床应用中,例如,当尝试检测样本中以低滴度存在的病原体时,起始样本输入量可能很低。

[0007] 在本领域中需要在使用NGS时准确定量样本中多核苷酸的丰度的组合物和方法。本文提供了可用于使用NGS对样本中的多核苷酸进行这种定量的方法和组合物。

发明内容

[0008] 本文提供了用于下一代测序的归一化对照、所述归一化对照的制备方法,以及使用所述归一化对照的方法。

[0009] 在一个方面中,本文提供了归一化对照,所述归一化对照包括至少三组多核苷酸,其中每个组内的多核苷酸具有相同的长度。在一些实施方式中,所述归一化对照中的所有多核苷酸的长度是相同的。在一些实施方式中,每个组内的多核苷酸当与任何其他组内的多核苷酸相比时具有不同的长度。

[0010] 在本公开的归一化对照的一些实施方式中,每个组内的多核苷酸包含相同的序列。在一些实施方式中,每个组内的多核苷酸不包含相同的序列。在一些实施方式中,每组多核苷酸包含多核苷酸的至少三个亚组,其中每个亚组内的多核苷酸包含相同的序列,并且其中每个亚组内的多核苷酸不包含与任何其他亚组相同的序列。

[0011] 在本公开的归一化对照的一些实施方式中,至少一组多核苷酸的序列包含组分序列。在一些实施方式中,每组多核苷酸的序列包含组分序列。在一些实施方式中,组分序列包含随机序列或由随机序列组成。

[0012] 在本公开的归一化对照的一些实施方式中,至少一组多核苷酸的序列包含分离的序列。在一些实施方式中,每组多核苷酸的序列包含分离的序列。在一些实施方式中,分离的序列是分离的或来源于病毒、细菌、真菌或真核寄生虫。在一些实施方式中,分离的序列与测序样本中的靶序列不同。

[0013] 在本公开的归一化对照的一些实施方式中,归一化对照中的至少一个亚组或至少一组多核苷酸的序列与测序样本中的靶序列共享至少一种序列特性。所述至少一个序列特性是例如转座元件序列、病毒序列、细菌序列、真菌序列、真核寄生虫序列、或一个或多个个人类基因序列的序列特性。

[0014] 在本公开的归一化对照的一些实施方式中,所述测序样本包含宿主核酸和非宿主核酸的混合物。在一些实施方式中,宿主是真核生物,例如昆虫、植物或动物。在一些实施方式中,宿主是人。在一些实施方式中,非宿主包含共生体、共生生物、寄生虫或病原体。非宿主可以包含多个物种。在一些实施方式中,靶序列是测序样本中的非宿主序列。在示例性实施方式中,非宿主是病毒、细菌或真菌,并且靶序列包含非宿主的毒力因子的序列。

[0015] 在另一方面中,本文提供了制备归一化对照的方法,所述方法包括:(a)合成至少三组DNA分子,其中一组内的DNA分子具有相同的序列,并且其中所述DNA分子中的每个DNA分子的序列从5'至3'包含第一组分序列、第一限制性酶切位点、第二组分序列、第二限制性位点和第三组分序列;(b)对溶液中各组DNA分子中的每个组的量进行定量;(c)用切割第一和第二限制位点的限制性酶消化所述多组DNA分子中的每个组,以产生至少三组经消化的DNA分子;以及(d)将预定量的所述多组经消化的DNA分子中的各个组混合,从而产生归一化对照。

[0016] 在另一方面中,本文提供了制备包含RNA分子的归一化对照的方法,所述方法包括:(a)合成至少三组DNA分子,其中所述至少三组DNA分子中的每一组从5'至3'包含第一组分序列、启动子序列、第一限制性酶切位点、第二组分序列、第二启动子序列、第二限制性位点、第三组分序列和第三启动子序列;(b)用切割第一和第二限制性酶切位点的限制性酶消化每组DNA分子,以产生至少三组消化的DNA分子;(c)在体外转录所述至少三组经消化的DNA分子中的每个组,以产生至少三个RNA分子群体;(d)对所述RNA分子群体中的每个RNA分子群体中产生的RNA的量进行定量;以及(e)混合预定量的所述RNA分子群体中的每个RNA分子群体以产生包含RNA分子的归一化对照。

[0017] 在另一方面中,本文提供了制备包含至少三组多核苷酸的归一化对照的方法,其中每个组内的多核苷酸具有相同的长度,所述方法包括:(a)从样本中提取DNA;(b)用限制性酶消化DNA以产生DNA片段集合;(c)分离所述DNA片段集合;(d)纯化DNA片段以产生至少三组多核苷酸,其中每个组内的多核苷酸具有相同的长度;以及(e)混合预定量的所述各组多核苷酸中的每个组以产生归一化对照。

[0018] 在另一方面中,本文提供了对样本中至少一个靶核酸分子的表达水平进行定量的方法,所述方法包括:(a)将已知量的本公开的归一化对照与样本混合,(b)制备高通量测序文库,(c)对所述文库测序以产生读段集合,(d)将每个读段映射至样本或归一化对照,(e)

确定由归一化对照中的所述组或亚组多核苷酸中的每个组或亚组产生的读段的数量, (f) 计算 (a) 中与样本混合的归一化对照中的多种核酸分子中的每种核酸分子的起始浓度与 (c) 中产生的读段的数量之间的关系, (g) 确定读段与样本中浓度之间的关系, 以及 (h) 使用 (g) 中的模型根据至少一个靶核酸分子产生的读段的数量计算样本中所述靶核酸分子的初始浓度。

[0019] 本文还提供了用于下一代测序的多分析物对照、所述多分析物对照的制备方法, 以及使用所述多分析物对照的方法。在一些实施方式中, 本文所述的归一化对照用于归一化来自多分析物对照和样本的读段, 并且归一化读段用于确定样本中的一种或多种靶生物体的滴度。

[0020] 在一个方面中, 本文提供了包含至少三种不同物种的生物体的混合物的多分析物对照, 其中所述生物体已被灭活。在一些实施方式中, 多分析物对照包含至少十种不同物种的生物。

[0021] 在另一方面中, 本文提供了对样本中至少一个靶序列进行定量的方法, 其中所述样本包含多分析物对照, 所述对照包含至少三种不同物种的生物体的混合物, 其中所述生物体已被灭活。

[0022] 在另一方面中, 本文提供了对样本中的至少一个靶序列进行定量的方法, 所述方法包括将已知量的多分析物对照与样本混合, 其中所述多分析物对照包括至少三种不同物种的生物体的混合物, 并且其中所述生物体已被灭活。

[0023] 在另一方面, 本文提供了对样本中的靶生物体体的滴度进行定量的方法, 所述方法包括: (a) 提供包含所述靶生物体体的样本, 其中所述靶生物体体包含至少一个靶序列; 以及 (b) 提供包含已知滴度的所述至少三种物种的生物体的多分析物对照, 其中所述生物体已被灭活; (c) 将已知量的本公开的归一化对照物与样本和多分析物对照混合; (d) 从样本和多分析物对照制备高通量测序文库; (e) 对所述文库测序以产生样本读段集合和多分析物对照读段集合; (f) 使用归一化对照对来自 (e) 的样本读段集合和多分析物对照读段集合进行归一化; (g) 确定归一化读段与来自多分析物对照中的至少三种物种的生物体的生物体的已知滴度之间的关系; 以及 (h) 使用 (g) 中确定的关系计算样本中的靶生物体体的滴度。

[0024] 在一些实施方式中, 步骤 (f) 处的归一化包括: (i) 将来自样本读段集合的读段映射至样本或归一化对照; (ii) 将来自多分析物对照读段集合的读段映射至多分析物对照或归一化对照; (iii) 针对样本读段集合和多分析物对照读段集合确定归一化对照中的所述组或亚组多核苷酸中的每个组或亚组产生的读段的数量; (iv) 计算归一化对照中的所述组或亚组多核苷酸中的每个组或亚组的起始浓度与在步骤 (e) 中产生的样本和多分析物对照两者的读段的数量之间的关系; 以及 (v) 确定样本和多分析物对照中的读段与浓度之间的关系。

[0025] 本公开提供了制备多个归一化对照寡核苷酸的方法, 所述方法包括: (a) 从至少一个参考序列产生多个参考序列片段; (b) 根据参考序列片段的数量产生至少一个参数的分布; (c) 将所述分布划分成至少5个箱; (d) 从至少5个箱中的至少3个箱中选择至少一个参考序列片段; (e) 改组至少3个参考序列以产生经改组的序列; 以及 (f) 合成包含经改组的序列的寡核苷酸; 从而产生多个归一化对照寡核苷酸。

[0026] 本公开提供了制备多个归一化对照寡核苷酸的方法,所述方法包括:(a)使用滑动窗从至少一个参考序列产生多个参考序列片段;(b)根据参考序列片段的数量产生至少一个参数的分布;(c)将所述分布划分为至少5个箱;(d)从至少5个箱中的至少3个箱中选择至少两个参考序列片段,其中所述至少两个参考序列片段在参考序列中是不连续的,或者是从不同的参考序列中选择的;(e)将来自至少3个箱中的每一个箱的至少两个参考序列片段连接起来;以及(f)合成包含连接的参考序列片段的寡核苷酸;从而产生多个归一化对照寡核苷酸。在一些实施方式中,根据权利要求203-206中任一项所述的方法,其中所述滑动窗口包括1bp、2bp、3bp、4bp或5bp的滑动窗口。在一些实施方式中,所述参考序列片段是对应参考序列的约15-60个、约20-40个、约20-30个、约15-32个、20-32个或约25-35个连续核苷酸。在一些实施方式中,参考序列片段是对应参考序列的29个、30个、31个、32个、33个或34个连续核苷酸。

[0027] 在本公开的方法的一些实施方式中,所述参数包括以下各项中的至少一项:(1)GC含量百分比、(2)熵、(3)复杂性、(4)EIIP,或它们的组合。

[0028] 在本公开的方法的一些实施方式中,所述至少一个参考序列包含至少2个、至少10个、至少20个、至少50个、至少100个、至少200个、至少300个、至少400个、至少500个、至少600个、至少700个、至少800个、至少900个、至少1,000个、至少1,200个、至少1,300个、至少1,400个、至少1,500个、至少1,600个、至少1,700个、至少1,800个、至少1,900个、至少2,000个、至少2,200个、至少2,400个、至少2,600个、至少2,800个、至少3,000个、至少4,000个、至少5,000个、至少6,000个、至少7,000个、至少8,000个、至少9,000或至少10,000个参考序列。在一些实施方式中,至少一个参考序列包含约2,000个参考序列。

[0029] 在本公开内容的方法的一些实施方式中,每个归一化对照寡核苷酸具有以下中的至少三者:与来自参考序列分布中的对应箱的参考序列片段(1)类似的GC含量百分比、(2)类似的熵、(3)类似的EIIP、(4)类似的长度、或(5)类似的复杂性。

[0030] 本公开提供了富集样本中感兴趣的靶核酸以供与本文提供的归一化对照和多分析物对照组合使用的方法。在一些实施方式中,所述方法包括基于核苷酸修饰的富集方法。在一些实施方式中,所述方法包括基于核酸引导核酸酶的富集方法。在一些实施方式中,所述方法包括基于核苷酸修饰的富集方法和基于核酸引导核酸酶的富集方法两者。

[0031] 在本公开的方法的一些实施方式中,富集样本中感兴趣的核酸的方法包括:(a)提供包含感兴趣的核酸和靶向耗尽的核酸的样本,其中所述感兴趣的核酸的至少一个子集或所述靶向耗尽的核酸的子集包含用于第一修饰敏感的限制性酶的多个第一识别位点;(b)使样本中的多种核酸末端去磷酸化;(c)在允许切割样本中的核酸中的第一修饰敏感限制性位点中的至少一些第一修饰敏感限制性位点的条件下,使来自(b)的样本与第一修饰敏感限制性酶接触;以及(d)在允许衔接子与多个感兴趣的核酸的5'末端和3'末端连接的条件下,使来自(c)的样本与衔接子接触;从而产生富集感兴趣的核酸的样本,所述感兴趣的核酸在其5'末端和3'末端与衔接子连接。

[0032] 在本公开的方法的一些实施方式中,所述方法使样本与多种核酸引导的核酸酶引导核酸(nuclease-guide nucleic acid,gNA)复合物接触,其中所述gNA与靶向耗尽的核酸中的靶位点互补,从而产生对在一个末端处与衔接子连接的靶向耗尽的核酸和在5'末端和3'末端处均与衔接子连接的感兴趣的靶核酸的切割。在一些实施方式中,该方法包括使样

本与至少102个独特的核酸引导的核酸酶-gNA复合物、至少103个独特的核酸引导的核酸酶-gNA复合物、104个独特的核酸引导的核酸酶-gNA复合物、或105个独特的核酸引导的核酸酶-gNA复合物接触。在一些实施方式中,核酸引导的核酸酶是Cas9、Cpf1或它们的组合。

[0033] 在其他方面中,本文提供了包含本公开的多分析物对照的试剂盒、包含本公开的归一化对照的试剂盒,以及包含本公开的归一化对照和多分析物对照的试剂盒。在一些实施方式中,试剂盒还包含用于感兴趣的靶核酸的基于核苷酸修饰的富集的试剂、用于感兴趣的靶核酸的基于核酸引导核酸酶的富集的试剂,或两者。

[0034] 本公开提供了包含本公开的归一化对照的试剂盒。

[0035] 本公开提供了包含本公开的多分析物对照的试剂盒。

[0036] 本公开提供了试剂盒,所述试剂盒包含归一化对照、多分析物对照、试剂和使用说明书。在一些实施方式中,试剂盒还包含用于耗尽靶向耗尽的序列的试剂和使用说明书。

[0037] 本公开提供了用于设计多个归一化对照多核苷酸序列的系统。在一些实施方式中,该系统包括计算机可读存储介质,该计算机可读存储介质存储计算机可执行指令,该计算机可执行指令包括:(i)用于导入至少一个参考序列的指令;(ii)从至少一个参考序列产生多个参考序列片段的指令;(iii)根据参考序列片段数目产生至少一个参数的分布的指令;(iv)将分布划分成箱的指令;(v)从箱的至少一个子集中选择多个参考序列片段的指令;和(vi)改组多个参考序列片段以产生改组序列的指令;从而产生多个归一化多核苷酸序列。在一些实施方式中,该系统包括计算机可读存储介质,该计算机可读存储介质存储计算机可执行指令,该计算机可执行指令包括:(i)用于导入至少一个参考序列的指令;(ii)从至少一个参考序列产生多个参考序列片段的指令;(iii)根据参考序列片段数目产生至少一个参数的分布的指令;(iv)将分布划分成箱的指令;(v)从箱的至少一个子集中的每一个中选择至少两个参考序列片段的指令,其中所述至少两个参考序列片段在参考序列中是不连续的,或来自不同的参考序列;(vi)连接来自每个箱的至少两个参考序列片段的指令;从而产生多个归一化多核苷酸序列。在一些实施方式中,该系统还包括处理器,该处理器被配置成执行包括以下步骤的步骤:(a)接收包括至少一个参考序列的一组输入文件;和(b)执行存储在计算机可读存储介质中的计算机可执行指令。在一些实施方式中,该参数包括以下各项中的至少一项:(1)GC含量百分比、(2)熵、(3)复杂性、(4)EIIP、(5)长度,或它们的组合。

附图说明

[0038] 图1是示出归一化对照(normalization control,NC)读段的质量控制分析的表。

[0039] 图2是绘制巨细胞病毒(Cytomegalovirus,CMV)丰度(x轴)对比滴度(y轴)的图。如果不进行归一化,则R2值为0.89。

[0040] 图3是绘制归一化对照(NC)归一化的CMV丰度(x轴)对比滴度(y轴)的图。通过归一化,R2值为0.98。

[0041] 图4是条形图,示出了用三个假想的下一代测序文库分析的样本中病毒的丰度百分比(y轴,任意单位),以证明使用不同转化计算得出的病毒的丰度百分比。对于每个文库,条形从左到右分别代表已知的病毒载量、使用下一代测序后的原始读段迹象计算的病毒载量以及使用f(r)读段证据和f(r)+ Δr 读段迹象计算的病毒载量。

[0042] 图5是示出当使用 $f(r)$ 计算病毒载量对比使用原始读段计算病毒载量时预期的可变速性降低的图。

[0043] 图6是用于定量样本中靶序列水平的方案的概述,其中在RNA提取过程中添加归一化对照。

[0044] 图7是示出精确诊断阴性血浆样本文库中归一化对照读段百分比的图,其中在样本提取期间以低、中和高量向样本中添加NC。

[0045] 图8是用于将双索引引物与归一化对照退火的方案的表。

[0046] 图9是示出在添加归一化对照的情况下文库中的归一化对照读段百分比的图,其中DNA NC以高(左两栏)、中(右两栏)和低(中两栏)的量添加。

[0047] 图10是示出示例性归一化对照的GC含量、总计数和长度的表。

[0048] 图11是示出本公开的示例性归一化对照的浓度和比率的表。从左至右,各列示出归一化对照组合物中的各组多核苷酸(在图11中称为“寡核苷酸”,并具有对应的寡核苷酸编号)、每组多核苷酸的最终浓度($\text{ng}/\mu\text{L}$),以及对于组合物中归一化对照多核苷酸的群体,归一化对照组合物中特定组的多核苷酸的DNA的量与归一化对照组合物中所有多核苷酸的所有DNA总量之比。

[0049] 图12是示出本公开的示例性RNA归一化对照的大小、输入分子数量和浓度的表。从左至右,各列示出了归一化对照中的单独RNA片段、归一化对照中的各组RNA、RNA分子的长度(以kb和bp为单位)、输入分子的数量、每个RNA分子在其原始gBlock中的阿摩尔(attomole)数,以及在gBlock片段化后每个RNA分子的所得浓度(以阿摩尔/微升计)。

[0050] 图13是示出以log为单位的滴度作为用于校准和实验样本的观察信号的函数的图。深灰色圆圈表示用于校准的多分析物对照样本中的物质,而浅灰色圆圈表示实验样本。

[0051] 图14A是示出用于设计本公开的归一化对照的示例性过程的流程图。

[0052] 图14B是示出本公开的归一化对照的一种设计的信号(y轴)和对数浓度(x轴)的图。

[0053] 图14C是示出本公开的一组归一化对照的浓度的表。每个寡核苷酸是不同的序列。

[0054] 图14D是示出本公开的归一化对照的一种设计的信号(y轴)和对数浓度(x轴)的图。浓度的每个数量级由6个浓度表示,每个对数浓度具有至少一个不同的寡核苷酸,并且覆盖了7-8个数量级。

[0055] 图15是示出关于大约1200个参考基因组的熵、电子-离子相互作用电势(electron-ion interaction potential, EIIP)和GC含量之间的相互作用的三元图。

[0056] 图16是一系列三个曲线图,示出了来自使用改组方法产生的参考序列的Kmer的分布。从左至右:电子EIIP、熵和GC含量。EIIP、熵和GC含量的值分别在x轴上从左至右示出。在y轴上示出了Kmer密度。该线表示分布,并且直方图条表示给定箱中Kmer数量的实际计数。

[0057] 图17A是一系列三个曲线图,示出了使用改组方法产生的1000万个Kmer的分布。从左至右:电子EIIP、熵和GC含量。EIIP、熵和GC含量的值分别在x轴上从左至右示出。在y轴上示出了Kmer密度。

[0058] 图17B是一系列三个曲线图,示出了使用改组方法产生的130万个Kmer的分布。从左至右:电子EIIP、熵和GC含量。EIIP、熵和GC含量的值分别在x轴上从左至右示出。在y轴上示出了Kmer密度。

[0059] 图18是示出在改组之前将Kmer连接以产生归一化对照序列的过程的图。

[0060] 图19A是一系列三个图,示出了使用Cantor改组方法生成的归一化对照与BK多瘤病毒分离株CH-1之间的EIIP(左上)、GC含量(右上)和复杂性(底部行)对应关系。未改组的Kmer绘制在x轴上,机构Cantor改组的Kmer绘制在y轴上,线表示1:1对应。

[0061] 图19B是一系列三个图,示出了使用Cantor改组方法生成的归一化对照与JC多瘤病毒株NIID 12-31之间的EIIP(左上)、GC含量(右上)和复杂性(底部行)对应关系。未改组的Kmer绘制在x轴上,机构Cantor改组的Kmer绘制在y轴上,线表示1:1对应。

[0062] 图19C是一系列三个图,示出了使用Cantor改组方法产生的归一化对照与人多瘤病毒1菌株BK 2之间的EIIP(左上)、GC含量(右上)和复杂性(底部行)对应关系。未改组的Kmer绘制在x轴上,机构Cantor改组的Kmer绘制在y轴上,线表示1:1对应。

[0063] 图19D是一系列三个图,示出了使用Cantor改组方法生成的归一化对照与的人腺病毒C之间的EIIP(左上)、GC含量(右上)和复杂性(底部行)对应关系。未改组的Kmer绘制在x轴上,机构Cantor改组的Kmer绘制在y轴上,线表示1:1对应。

[0064] 图20A是基本局部比对搜索工具(Basic Local Alignment Search Tool, BLAST)比对,示出了使用NCBI BLAST核苷酸序列时,经Cantor改组的归一化对照序列不与BK多瘤病毒分离株CH-1进行比对。

[0065] 图20B是blast比对,示出了使用NCBI BLAST核苷酸序列时,经Cantor改组的归一化对照序列不与JC多瘤病毒株NIID 12-31比对。

[0066] 图20C是blast比对,示出了使用NCBI BLAST核苷酸序列时,经Cantor改组的归一化对照序列不与人多瘤病毒1菌株BK 2比对。

[0067] 图20D是BLAST比对,示出了使用NCBI BLAST核苷酸序列时,经Cantor改组的归一化对照序列不与人腺病毒C比对。

[0068] 图21A是一系列三个曲线图,示出了具有选定的GC含量百分比的Kmer的EIIP、熵和GC含量。

[0069] 图21B是一系列曲线图,示出了具有选定的GC含量百分比的Kmer的EIIP、熵和GC含量。

[0070] 图21C是一系列曲线图,示出了具有选定的GC含量百分比的Kmer的EIIP、熵和GC含量。

[0071] 图22是一对曲线图,示出了EIIP、GC含量和熵的第一和第二主分量(分别为x轴和y轴)。计算了PC1和PC2的参考基因组序列(左)和使用滑动窗口/穷举Kmer方法生成的归一化对照序列(右)。

[0072] 图23是一系列三个Q-Q曲线图,比较了参考基因组序列(x轴)和使用滑动窗/穷举Kmer方法生成的归一化对照寡核苷酸(y轴)的GC含量(顶部)、EIIP(左下)和熵(右下)。

[0073] 图24是Kolmogorov-Smirnov(KS)检验曲线图,比较了参考基因组序列和使用滑动窗/穷举Kmer方法生成的归一化对照序列的随机子取样群体中的熵。ECDF=经验累积分布函数。箭头指示归一化对照的熵。点表示K-S统计。

[0074] 图25A是BLAST比对,示出了使用滑动窗/穷举Kmer方法生成的785,000个归一化对照序列的文件头部的序列在约750个碱基对(bp)的位置与假单胞菌属(Pseudomonas)的约50-80bp比对。

[0075] 图25B是BLAST比对,示出了使用BLAST时,使用滑动窗/穷举Kmer方法生成的785,000个归一化对照序列的文件尾部的序列不与任何序列进行比对。

[0076] 图26是比较使用改组和穷举Kmer/滑动窗方法设计的归一化对照的表。

[0077] 图27是示出本发明的示例性方法的图。样本中的核酸被去磷酸化,然后用限制性酶消化,该限制性酶通过限制性酶识别位点处的修饰的存在而被封端。然后将来自所得消化的暴露的磷酸酯用于将衔接子连接至感兴趣的核酸。

[0078] 图28是示出本发明的示例性方法的图。将样本中的核酸去磷酸化,然后用识别限制性酶切位点的限制性酶消化,该限制性酶切位点包含一个或多个经修饰的核苷酸。然后,将切出的核酸用核酸外切酶消化,所述核酸外切酶使用暴露的末端磷酸酯,并将衔接子与其余的感兴趣的核酸连接。

[0079] 图29是示出本发明的示例性方法的图。将样本中的核酸与衔接子连接,然后用限制性酶消化,所述限制性酶识别包含一个或多个经修饰的核苷酸的限制性酶切位点,从而产生在两个末端处均与衔接子连接的感兴趣的核酸。

[0080] 图30是示出了本公开的示例性方法的图。将样本中的核酸与衔接子连接,然后用核酸引导的核酸酶进行切割,该核酸引导的核酸酶切割靶向耗尽的核酸,从而产生在两个末端上均与衔接子连接的感兴趣的核酸。该方法可以与本公开的基于核苷酸修饰的方法结合使用。

具体实施方式

[0081] 本文提供了用作下一代测序(next generation sequencing,NGS)中样本的定量的归一化对照的组合物。在一些实施方式中,当使用低样本输入工作以进行NGS时使用归一化对照。在一些实施方式中,归一化对照用于计算样本中生物体的滴度,例如使用如本文所述的多分析物对照。在一些实施方式中,如本文所述,将归一化对照与富集样本中的靶序列的方法组合。

[0082] 本文提供了制备归一化对照的方法,以及在NGS期间使用归一化对照的方法。

[0083] 除非本文另外定义,否则本文使用的所有技术术语和科学术语的含义与如本公开所属领域的普通技术人员通常所理解的含义相同。尽管在本公开的实践或测试中可以使用类似于或等同于本文所述的方法和材料的任何方法和材料,但是描述了示例性方法和材料。

[0084] 术语“下一代测序”(NGS)是指所谓的并行合成测序或连接测序平台,例如,目前由Illumina、Life Technologies和Roche采用的那些平台。下一代测序方法还可以包括纳米孔测序方法或基于电子检测的方法,例如由Life Technologies商售的Ion Torrent技术。NGS在本文中也称为“高通量测序”。

[0085] “测序样本”是包含使用NGS测序的核酸的样本。核酸可以是脱氧核糖核酸(deoxyribonucleic acid,DNA)、核糖核酸(ribonucleic acid,RNA)或两者。

[0086] “靶序列”是指测序样本中的单独序列或序列组,所述样本中的丰度是通过使用本公开的NC在NGS反应中测定的。靶序列的丰度可以为非零(即,靶序列以一定的丰度存在于样本中)或零(即,不存在靶序列或低于检测水平)。

[0087] 如本文所用,“参考序列”是指与任何已知靶序列相同或相似的一个或多个序列。

靶序列可以包含一个参考序列或多个参考序列或由其组成。

[0088] “靶生物体”是指样本中包含靶序列的生物体。

[0089] 如本文所用,术语“组分序列”是指在归一化对照中发现的多核苷酸的一部分或全部。归一化对照多核苷酸包含组分序列或由其组成。可以从天然存在的序列、经工程化的质粒或合成物中分离组分序列。术语“分离的序列”是指已经分离或来源于生物体或其他先前存在的序列的类型的组分序列。分离的序列可以类似于它们所分离自或来源于的生物体的序列,或者也可以经受一种或多种转化,例如改组或连接。

[0090] 术语“低样本”或“低输入样本”是指包含比常规文库制备方案中使用的核酸量低的核酸量的样本。低样本输入量可包含样本中总核酸中的至多100ng的核酸,以及少至1皮克核酸。在某些实施方式中,术语“低样本”或“低样本输入”是指编码样本中的一个或多个靶序列的核酸的量。低样本输入可包含至多100ng包含靶序列的核酸,或少于1飞克的包含靶序列的核酸。对于特别罕见的靶序列,飞摩尔和阿摩尔(attomolar)范围内的靶序列的浓度被认为在本公开的范围內。

[0091] 如在整个本公开中所使用的,“序列同一性”或“序列相似性”可以通过使用默认参数,使用用于对两个序列进行blast的独立可执行BLAST引擎程序(bl2seq)来确定,所述程序可以从美国国家生物技术信息中心(National Center for Biotechnology Information,NCBI)ftp站点取得(Tatusova和Madden,FEMS Microbiol Lett.,1999,174,247-250;该文献以引用方式整体并入本文)。当在两个或更多个核酸或多肽序列的情境中使用,术语“同一”或“同一性”是指在每个序列的指定区域上相同的残基的指定百分比。可以通过以下方法计算出百分比:对两个序列进行最佳比对;在指定区域内比较两个序列;确定两个序列中出现相同残基的位置数,以得到匹配的位置数;将匹配的位置数除以指定区域中的位置总数;以及将结果乘以100以得出序列同一性百分比。在两个序列具有不同的长度。或者比对产生一个或多个交错的末端并且指定的比较区域仅包含单个序列的情况下,单个序列的残基包含在计算的分母,而不是分子中。比较DNA和RNA时,胸腺嘧啶(T)和尿嘧啶(U)可以认为是等效的。同一性可以手动执行,或通过使用计算机序列算法(例如BLAST或BLAST 2.0)执行。

[0092] 如果通过比对序列是至少50%、至少60%、至少70%、至少75%、至少80%、至少85%、至少90%、至少91%、至少92%、至少93%、至少94%、至少95%、至少96%、至少97%、至少98%、至少99%、至少99.5%或至少99.9%同一的,则所述序列是“类似的”。

[0093] 本公开提供了要添加到用于NGS的样本中的归一化对照。任选地,该样本是低输入样本。

[0094] 本公开提供了制备归一化对照的方法以及使用归一化对照的方法。

[0095] 在一些实施方式中,将归一化对照以相同的量添加至每个样本中。

[0096] 在一些实施方式中,不将归一化对照以相同的量添加至每个样本中。

[0097] 在一些实施方式中,在核酸提取之前将归一化对照添加至样本中。在其他实施方式中,在核酸提取之后和文库制备期间添加归一化对照。在其他替代方案中,在文库制备之后和测序之前将归一化对照添加到样本中。

[0098] 本公开提供了包含至少三组多核苷酸的归一化对照,其中所述至少三组中的每个组内的多核苷酸具有相同的长度。在一些实施方式中,归一化对照多核苷酸包含组分序列,

在一些实施方式中,所述组分序列被设计为类似于或概括待测序的靶样本的一些特征或特性。在一些实施方式中,组分序列是随机序列。在一些实施方式中,归一化对照多核苷酸包含分离的序列。在一些实施方式中,归一化对照多核苷酸包含分离的序列和组分序列两者。在一些实施方式中,归一化对照多核苷酸的序列与测序样本中的靶序列共享至少一种序列特性。

[0099] 本公开提供了制备归一化对照的方法,所述方法包括:(a)合成至少三个DNA分子群体,其中一个群体内的每个DNA分子具有相同的序列,并且其中所述DNA分子中的每个DNA分子的序列从5'至3'包含第一组分序列、第一限制性酶切位点、第二组分序列、第二限制性位点和第三组分序列;(b)对溶液中DNA分子群体中的每个群体的量进行定量;(c)用切割第一和第二限制位点的限制性酶消化所述DNA分子群体中的每个群体,以产生至少三个经消化的DNA分子群体;以及(d)将预定量的所述经消化的DNA分子群体中的各个群体在单个归一化对照组合物中混合。

[0100] 在一些实施方式,例如其中样本是RNA样本并且归一化对照是RNA归一化对照的那些实施方式中,本公开提供了制备归一化对照的方法,所述方法包括:(a)合成具有至少三个DNA分子的群体;其中所述至少三个合成DNA分子中的每一个DNA分子从5'至3'包含第一组分序列、与启动子序列互补的序列、第一限制性酶切位点、第二组分序列、与启动子序列互补的第二序列、第二限制性位点、第三组分序列和与启动子序列互补的第三序列;(b)用切割第一和第二限制性酶切位点的限制性酶消化每个DNA分子群体,以产生至少三个经消化的DNA分子群体;(c)在体外转录所述至少三个经消化的DNA分子群体中的每一个群体,以产生至少三个RNA分子群体;(d)对每个RNA分子群体中产生的RNA数量进行定量;以及(e)混合预定量的各个RNA分子群体以产生单一的归一化对照组合物。

[0101] 本公开提供了制备归一化对照的方法,所述方法包括:(a)从样本中提取DNA;(b)用限制性酶消化DNA以产生DNA片段集合;以及(c)分离所述DNA片段集合;(d)纯化至少三个DNA片段以产生至少三组多核苷酸,其中所述至少三组多核苷酸中的每个组均包含多个多核苷酸;以及(e)将预定量的所述至少三组多核苷酸中的各个组混合以产生归一化对照。

[0102] 本公开还提供了制备归一化对照组合物的方法,其中所述归一化对照包含使用本公开的方法合成的多核苷酸和包含使用本公开的方法分离和/或得到的分离的序列的多核苷酸两者。

[0103] 本公开提供了使用归一化对照对样本中的至少一种靶核酸分子的表达水平进行定量的方法,所述方法包括:(a)将已知量的本公开的归一化对照与样本混合,(b)制备高通量测序文库,(c)对所述文库进行测序以产生读段集合,(d)将每个读段映射至样本或归一化对照,(e)确定由归一化对照中的多个核酸分子中的每个核酸分子产生的读段的数量,(f)计算(a)中与样本混合的归一化对照中的多个核酸分子中的每个核酸分子的起始浓度与(c)中产生的读段的数量之间的关系,(g)对读段与样本中的浓度之间的关系进行建模,以及(h)使用(g)中的模型,根据靶核酸分子产生的读段的数量计算样本中的至少一种靶核酸分子的初始浓度。

[0104] 归一化对照组合物

[0105] 本公开提供了归一化对照组合物(在本文中可互换地称为归一化对照,或(NC),其可用于对NGS应用中的样本输入进行定量。如本文所提供的,本公开的NC可包含至少两组多

核苷酸,其中在所述至少两组中的每个组内的多核苷酸具有相同的长度。在一些实施方式中,将NC从头合成为多核苷酸。在一些实施方式中,NC是分离自或来源于样本,例如来自生物体的DNA或RNA样本,或载体。在一些实施方式中,NC包括从头合成的多核苷酸和分离自或来源于样本的多核苷酸两者。在一些实施方式中,NC包含至少三组多核苷酸,其中所述至少三组中的每个组内的多核苷酸具有相同的长度。

[0106] 在一些实施方式中,NC包含至少3组不同长度的多核苷酸序列,并且每组多核苷酸以不同的浓度存在于归一化对照中。

[0107] 在一些实施方式中,NC包含至少15组不同长度的多核苷酸序列,并且每组多核苷酸以不同的浓度存在于归一化对照中。

[0108] 在一些实施方式中,NC包含15组不同长度的多核苷酸序列,并且每组多核苷酸以不同的浓度存在于归一化对照中。

[0109] NC中的多核苷酸长度

[0110] 可以在整个浓度范围内改变归一化对照内的多核苷酸的长度,以允许片段化率或聚类的偏倚(例如有时在基于Illumina的NGS方法的情况下看到的)。例如,归一化对照包括多组多核苷酸,并且每组多核苷酸具有不同的长度并且在归一化对照内以不同的浓度存在。

[0111] 设想具有不同长度和长度的组合的各组多核苷酸在本公开的归一化对照的范围内。给定组内的所有单独多核苷酸均具有相同的长度。然而,在包含至少三个组的归一化对照组合物中,一个组内的多核苷酸的长度可以与归一化对照组合物中的任何其他组内的多核苷酸的长度相同或不同。

[0112] 因此,在一些实施方式中,所述归一化对照中的所有多核苷酸的长度是相同的。在一些实施方式中,至少三个组中的每一个组内的多核苷酸具有与任何其他组中的多核苷酸不同的长度。

[0113] 在一些实施方式中,至少三组多核苷酸的长度以线性序列或几何序列分布。

[0114] 线性序列是数字序列,其中所述序列中的下一个数字每次相对于前一个数字增加或减少相同的量。示例性线性序列包括序列“1、2、3、4、5、6”。通常,线性序列由下式表示:

[0115] $un = dxn + c$

[0116] 其中d是序列中连续项之间的第一个差异,n是序列中的项,并且c是常数。

[0117] 几何序列是数字序列,其中第一项之后的每个项都可以通过将前一项乘以固定的非零数(称为“公共比率”)来找到。示例性的几何序列包括“2、4、8、16”(每个数字是前一个的2倍)。通常,几何序列由下式表示:

[0118] $un = u1r^{n-1}$

[0119] 其中r是公共比率。

[0120] 在一些实施方式中,例如在其中多核苷酸包含组分序列的那些实施方式中(如本文进一步描述的),所述组中的每个组中的多核苷酸的长度介于约15bp与约3000bp之间、介于约50bp与约3000bp之间、介于约100bp与约3000bp之间、介于约1000bp与约3000bp之间、介于约1200与约3000bp之间、介于约1500bp与约3000bp之间、介于约15bp与约50bp之间、介于约15bp与约100bp之间、介于约15bp与约150bp之间、介于约15bp与约200bp之间、介于约15bp与约300bp之间、介于约15bp与约400bp之间、介于约15与约500bp之间、介于约50bp与

约1200bp之间、介于约100bp与约1200bp之间、介于约150bp与约1200bp或介于约150与约1100bp之间。在示例性实施方式中,各组多核苷酸中的多核苷酸的长度介于约15bp与约3000bp之间。在示例性实施方式中,各组多核苷酸中的多核苷酸的长度介于约500bp与约1500bp之间。在示例性实施方式中,各组多核苷酸中的多核苷酸的长度介于约100bp与约1200bp之间。在示例性实施方式中,各组多核苷酸中的多核苷酸的长度介于约150bp与约600bp之间。

[0121] 在一些实施方式中,至少三组多核苷酸包含具有选自以下组成的长度的多核苷酸:175bp、250bp和450bp;192bp、250bp和450bp;200bp、300bp和500bp;217bp、300bp和517bp;436bp、552bp和974bp;450bp、612bp和1034bp;510bp、626bp和1048bp;以及450bp、612bp和1034bp。

[0122] 如本文所提供的,各组多核苷酸可以在体外合成,或者可以分离自或来源于已合成的核酸,或者可以分离自或来源于天然存在的核酸。预期在体外合成的多核苷酸可具有约3kb的最大长度,但是本领域技术人员将理解该长度可取决于当前的寡核苷酸合成技术。预期与使用合成方法制备的归一化对照相比,分离自或来源于已经合成的或天然存在的核酸的归一化对照多核苷酸可以包含更宽的长度范围,但是在此,所述长度范围也是取决于当前寡核苷酸合成技术。

[0123] 在一些实施方式中,例如在归一化对照多核苷酸包含分离的序列(如本文进一步描述)的情况下,至少三个不同组的多核苷酸的长度介于约15bp与约500kb之间。在一些实施方式中,至少三个不同组的多核苷酸的长度介于约15bp与约50kb之间。在一些实施方式中,至少三个不同组的多核苷酸的长度介于约50bp与约50kb之间。在一些实施方式中,至少三个不同组的多核苷酸的长度介于约500bp与约50kb之间。在一些实施方式中,至少三个不同组的多核苷酸的长度介于约1000bp与约50kb之间。在一些实施方式中,至少三个不同组的多核苷酸中的每一组中的最长的是至少15kb的长度、至少20kb的长度、至少25kb的长度、至少30kb的长度、至少35kb的长度、或至少40k的长度,包括端点。

[0124] NC中的多核苷酸序列

[0125] 设想本公开的各组多核苷酸包含具有相同序列的多核苷酸,或包含其中该组中的所有多核苷酸的序列不完全相同的多核苷酸。

[0126] 在一些实施方式中,每个组内的多核苷酸由相同序列组成。

[0127] 在一些实施方式中,每个组内的多核苷酸包含相同的序列。对于每个组中的所有多核苷酸可以相同的序列的非限制性示例包括衔接子序列、引物序列、索引序列、限制性位点、启动子序列、组分序列、分离的序列,或它们的组合。

[0128] 在一些实施方式中,每个组内的多核苷酸包含在所有多核苷酸中相同的序列和在所有多核苷酸中不同的序列两者。例如,每个组内的所有多核苷酸可包含相同的衔接子和/或索引序列,但是它们的组分序列不同。在其他实施方式中,每个组内的所有多核苷酸可包含不同的衔接子和/或索引序列,但是共享相同的组分序列或分离的序列。

[0129] 在一些实施方式中,每个组内的多核苷酸的序列是至少5%同一、至少10%同一、至少20%同一、至少30%同一、至少40%同一、至少50%同一的,是至少60%同一、至少70%同一、至少80%同一的,是至少90%同一、至少95%同一、至少97%同一、至少98%同一或至少99%同一的。

[0130] 在一些实施方式中,每个组内的多核苷酸不包含相同的序列。

[0131] 在一些实施方式中,一个或多个多核苷酸或每组多核苷酸包含多核苷酸的至少三个亚组,其中每个亚组内的多核苷酸包含相同的序列,并且每个亚组内的多核苷酸不包含与任何其他亚组相同的序列。

[0132] 在本公开的归一化对照的一些实施方式中,至少三个组中的每个组内的每个多核苷酸包含相同的序列。

[0133] 在一些实施方式中,至少三个组中的每个组内的多个多核苷酸的序列不包含与任何其他组中的多个多核苷酸的序列相同的序列。

[0134] 在一些实施方式中,至少三个组中的每个组内的多个多核苷酸中的每一个多核苷酸包含不具有相同序列的多核苷酸。

[0135] 在一些实施方式中,至少三组多核苷酸中的每一个组包含多核苷酸的至少三个亚组,并且多核苷酸的每个亚组包含具有相同序列的多个多核苷酸,并且每个亚组中的多个多核苷酸的序列不与该组中的其他亚组相同。

[0136] 组分序列

[0137] 在本公开的归一化对照的一些实施方式中,至少一组多核苷酸的序列包含被设计、选择或已知为具有某些特性(例如长度、序列或GC含量)的组分序列。这可以导致那组多核苷酸具有所表示的那些特性的经设计、选择或已知的范围。组分序列的性质可经设计或选择以匹配测序样本中的靶序列的特性。

[0138] 如本文所用,术语“组分序列”是指在归一化对照中发现的多核苷酸的一部分或全部。归一化对照多核苷酸包含组分序列或由组分序列组成。在一些实施方式中,组分序列是计算机模拟设计的,并且不是分离自或来源于任何生物体的。

[0139] 在一些实施方式中,组分序列包括随机序列。在一些实施方式中,将组分序列设计为模拟或共享测序样本中的靶序列的一种或多种序列特性。

[0140] 在一些实施方式中,每组多核苷酸的序列包含组分序列。在一些实施方式中,每组多核苷酸的组分序列与任何其他组的组分序列不同。

[0141] 在一些实施方式中,至少一组多核苷酸的至少一个亚组的序列包含组分序列。在一些实施方式中,至少一个组的至少三个亚组中的每一个亚组的序列包含组分序列。在一些实施方式中,每个组的至少三个亚组中的每一亚组的序列包含组分序列。在一些实施方式中,每个亚组的组分序列与任何其他亚组的组分序列不同。

[0142] 例如,在具有三个组的归一化对照组合中,每个组具有三个亚组,每个亚组具有不同的组分序列,所述组分序列没有一个是相同的。在该示例中,归一化对照组合中的多核苷酸群体包含具有九种不同组分序列的多核苷酸。

[0143] 在一些实施方式中,组分序列包含介于约15bp与约3000bp之间、介于约50bp与约3000bp之间、介于约100bp与约3000bp之间、介于约1000bp与约3000bp之间、介于约1200与约3000bp之间、介于约1500bp与约3000bp之间、介于约15bp与约50bp之间、介于约15bp与约100bp之间、介于约15bp与约150bp之间、介于约15bp与约200bp之间、介于约15bp与约300bp之间、介于约15bp与约400bp之间、介于约15与约500bp之间、介于约50bp与约1200bp之间、介于约100bp与约1200bp之间、介于约150bp与约1200bp或介于约150与约1100bp之间的序列。在一些实施方式中,组分序列包含介于约6bp与约3000bp之间的序列。在一些实施方式

中,组分序列包含介于约150bp与约500bp之间的序列。

[0144] 在本公开的方法的一些实施方式中,组分序列与测序样本中的靶序列不同。在一些实施方式中,组分序列具有与测序样本中的靶序列的小于或等于1%、2%、3%、4%、5%、6%、7%、8%、9%、10%、11%、12%、13%、14%、15%、16%、17%、18%、19%、20%、30%、40%、50%、60%、70%、80%、85%、90%、95%、96%、97%、98%、99%或99.5%、99.8%或99.9%的同一性。在一些实施方式中,组分序列可与靶序列差异每20bp约1bp、每50bp约1bp、每150bp约1bp、每250bp约1bp、或每300bp约1bp。

[0145] 在一些实施方式中,组分序列包含随机序列或由随机序列组成。在一些实施方式中,使用随机序列生成器来生成随机序列。在一些实施方式中,随机序列不映射到参考序列。在一些实施方式中,随机序列不映射到任何感兴趣的生物体(例如NCBI核苷酸数据库中的任何生物体,或将使用归一化对照来检测、定量或以其他方式分析的任何生物体)的基因组。在一些实施方式中,随机序列不由在被设计用于切割人DNA的90k gRNA阵列中的任何与CRISPR相关联的(CRISPR-associated,Cas)指导RNA(guide RNA,gRNA)切割。在一些实施方式中,可以将随机序列设计为排除特定的序列特征。例如,随机序列可以排除诸如CRISPR/Cas gRNA识别位点、限制性位点,转录因子结合位点或重复序列的特征。在一些实施方式中,可以将随机序列设计为模拟靶序列的特性。例如,随机序列可以被设计为具有与靶序列相同的GC含量或相同的特定序列基序(例如三核苷酸重复)频率。

[0146] 分离的序列

[0147] 在本公开的归一化对照的一些实施方式中,归一化对照多核苷酸包含分离的序列。如本文所用,“分离的序列”是一种类型的组分序列。术语“分离的序列”是指已经分离自或来源于生物体或其他先前存在的序列的序列。分离的序列可包含基因组、线粒体或叶绿体DNA序列。分离的序列可以包含RNA序列,例如信使RNA(messenger RNA,mRNA)、转移RNA(transfer RNA,tRNA)、非编码RNA(non-coding RNA,ncRNA)或微型RNA。分离的序列可以分离自生物体或预先存在的序列,并且包含预先存在的序列的一个或多个片段。替代地或另外,可以对分离的序列进行一种或多种转化,以从预先存在的序列中得到它们,所述一种或多种转化例如序列改组、序列连接、反转核苷酸的顺序、或嵌套在其他序列中。

[0148] 在一些实施方式中,至少一组多核苷酸的序列包含分离的序列。在一些实施方式中,每组多核苷酸的序列包含分离的序列。

[0149] 在一些实施方式中,至少一组多核苷酸的至少一个亚组的序列包含分离的序列。在一些实施方式中,至少一个组的每个亚组的序列包含分离的序列。在一些实施方式中,每个组的每个亚组的序列包含分离的序列。在一些实施方式中,每个亚组的分离序列与任何其他亚组的分离序列不同。

[0150] 在一些实施方式中,分离的序列的长度介于约15bp与约500kb之间。在一些实施方式中,分离的序列的长度介于约15bp与约100kb之间。在一些实施方式中,分离的序列的长度介于约15bp与约50kb之间。在一些实施方式中,分离的序列的长度介于约50bp与约50kb之间。在一些实施方式中,分离的序列的长度介于约500bp与约50kb之间。在一些实施方式中,分离的序列的长度介于约1000bp与约50kb之间。

[0151] 在一些实施方式中,分离的序列介于约6bp与约200,000bp之间、介于约15bp与约50,000bp之间、介于约500bp与约1500bp之间、介于约100bp与约1200bp之间,或介于约

150bp与约600bp之间。

[0152] 在一些实施方式中,分离的序列为至少约50bp、至少约100bp、至少约150bp、至少约200bp、至少约250bp、至少约300bp、至少约350bp、至少约400bp、至少约450bp、至少约500bp、至少约550bp、至少约600bp、至少约650bp、至少约700bp、至少约750bp、至少约800bp、至少约850bp、至少约900bp、至少约950bp、至少约1000bp、至少约1250bp、至少约1500bp、至少约2000bp、至少约2500bp或至少约3000bp。

[0153] 在一些实施方式中,分离的序列为约100bp、约200bp、约300bp、约400bp、约500bp、约600bp、约700bp、约800bp、约900bp、约1000bp、约1100bp、约1200bp、约1300bp、约1400bp或约1500bp。

[0154] 在一些实施方式中,分离的序列是分离的或来源于病毒、细菌、真菌或真核寄生虫。在一些实施方式中,病毒是T4噬菌体(T4)或巨细胞病毒(CMV)。在一些实施方式中,分离的序列分离自克隆载体(例如质粒)。

[0155] 在一些实施方式中,从其纯化归一化对照的分离的序列的核酸样本是克隆载体。在一些实施方式中,克隆载体是细菌组分染色体(bacterial component chromosome, BAC)、酵母人工染色体(yeast artificial chromosome, YAC)、粘粒、福斯质粒或质粒。在一些实施方式中,从其纯化分离的归一化对照的核酸样本是质粒。在一些实施方式中,将归一化对照序列从感兴趣的物种(例如病毒、细菌或真核寄生虫)克隆到克隆载体中,并使用本公开的方法从克隆载体中纯化归一化对照。

[0156] 在一些实施方式中,分离的序列与测序样本中的靶序列不同。在一些实施方式中,归一化对照中的任何多核苷酸的序列具有与测序样本中的靶序列小于或等于1%、2%、3%、4%、5%、6%、7%、8%、9%、10%、11%、12%、13%、14%、15%、16%、17%、18%、19%、20%、30%、40%、50%、60%、70%、80%、85%、90%、95%、96%、97%、98%或99%的同一性。

[0157] 在一些实施方式中,分离的序列分离自或来源于多个参考序列。“参考序列”是指与本领域中已知的靶序列相同或相似的一个或多个序列。参考序列可以指单个靶序列,例如靶生物体的基因或基因组。

[0158] 作为另一个示例,“参考序列”可以指来自多种不同生物体的多种序列。所述生物体可以是病毒、细菌、真菌或单细胞真核生物,例如寄生或致病性真核生物,或它们的组合。参考序列可以是基因组DNA序列、cDNA序列或它们的组合。

[0159] 在一些实施方式中,归一化对照包含分离自或来源于单个参考序列的多个分离的序列。

[0160] 在一些实施方式中,归一化对照包括多个分离的序列,所述多个分离的序列分离自或来源于多个参考序列,例如来自多种不同的生物体。在一些实施方式中,多个参考序列包括来自通常被认为是非宿主生物体的生物体的多个序列,所述生物体例如病毒、细菌或真菌(例如,当宿主是哺乳动物、植物或多细胞真核生物时)。多个参考序列可用于对本文所述的归一化对照中的靶生物体组的群体水平特征进行建模。在测序样本中可以存在任何一组靶生物体,并且它们的存在和滴度可以使用本文所述的方法进行测量。在其中测序样本中的靶序列或靶生物体在测序之前未知的那些实施方式中,使用对多种生物体进行建模的归一化对照可以提高鉴定和滴度测量的准确性。

[0161] 在一些实施方式中,多个分离的序列是分离自或来源于至少2个、至少3个、至少4个、至少5个、至少6个、至少7个、至少8个、至少9个、至少10个、至少20个、至少50个、至少100个、至少200个、至少300个、至少400个、至少500个、至少600个、至少700个、至少800个、至少900个、至少1,000个、至少1,200个、至少1,300个、至少1,400个、至少1,500个、至少1,600个、至少1,700个、至少1,800个、至少1,900个、至少2,000个、至少2,200个、至少2,400个、至少2,600个、至少2,800个、至少3,000个、至少4,000个、至少5,000个、至少6,000个、至少7,000个、至少8,000个、至少9,000个、至少10,000个参考序列、至少11,000个参考序列、或至少12,000个参考序列。在一些实施方式中,多个分离的序列是分离自或来源于至少10个参考序列。在一些实施方式中,多个分离的序列是分离自或来源于至少100个参考序列。在一些实施方式中,多个分离的序列是分离自或来源于至少500个参考序列。来源于在一些实施方式中,多个分离的序列是分离自或来源于至少1000个参考序列。来源于在一些实施方式中,多个分离的序列是分离自或来源于至少1,200个参考序列。来源于在一些实施方式中,多个分离的序列是分离自或来源于至少1,500个参考序列。来源于在一些实施方式中,多个分离的序列是分离自或来源于至少2,000个参考序列。在一些实施方式中,每个参考序列来自不同生物体或同一生物体的不同品系或亚种。在一些实施方式中,多个参考序列是来自不同生物体或同一生物体的不同品系或亚种的基因组DNA序列。

[0162] 在一些实施方式中,多个参考序列包括来自多种生物体的参考序列。在一些实施方式中,多个参考序列包含来自至少2种、至少10种、至少20种、至少50种、至少100种、至少200种、至少300种、至少400种、至少500种、至少600种、至少700种、至少800种、至少900种、至少1,000种、至少1,200种、至少1,300种、至少1,400种、至少1,500种、至少1,600种、至少1,700种、至少1,800种、至少1,900种、至少2,000种、至少2,200种、至少2,400种、至少2,600种、至少2,800种、至少3,000种、至少4,000种、至少5,000种、至少6,000种、至少7,000种、至少8,000种、至少9,000种、或至少10,000种生物体的序列。在一些实施方式中,多个参考序列包含来自约1,000种生物体、来自约1,200种生物体、来自约1,400种生物体、来自约1,500种生物体、来自约1,600种生物体、来自约1,700种生物体、来自约1,800种生物体、来自约2,000种生物体、来自约2,200种生物体、来自约2,400种生物体、来自约2,500种生物体、来自约3,000种生物体、来自约5,000种生物体、来自约7,000种生物体、或来自约10,000种生物体的序列。

[0163] 在一些实施方式中,多个分离的序列包含分离自或来源于参考序列的片段的序列。在一些实施方式中,参考序列片段是参考序列的介于约10-100个之间、介于约10-80个之间、介于约10-70个之间、介于约10-60个之间、介于约10-50个之间、介于约10-40个之间、介于约10-30个之间、介于约10-20个之间、介于约20-100个之间、介于约20-80个之间、介于约20-60个之间、介于约20-50个之间、介于约20-40个之间、介于约20-30个之间、介于约10-35个之间、介于约20-35个之间、或介于约20-25个之间的连续核苷酸。在一些实施方式中,所述参考序列片段是参考序列的介于约15-60个之间、介于约20-40个之间、介于约20-30个之间、介于约15-32个之间、介于20-32个之间或介于约25-35个之间的连续核苷酸。在一些实施方式中,参考序列片段包含参考序列的20个、21个、22个、23个、24个、25个、26个、27个、28个、29个、30个、31个、32个、33个、34个、35个、36个、37个、38个、39个、40个、41个、42个、43个、44个、45个、46个、47个、48个、49个、50个、51个、52个、53个、54个、55个、56个、57

个、58个、59个、60个、61个、62个、63个、64个或65个连续碱基对。在一些实施方式中,参考序列片段包含参考序列的25个、26个、27个、28个、29个、30个、31个、32个、33个、34个或35个连续bp。在一些实施方式中,参考序列片段包含参考序列的31个或32个连续bp。

[0164] 在一些实施方式中,参考序列片段小于测序读段的平均长度的一半。例如,如果NGS测序方法产生的平均读段均为120bp,则参考序列片段小于60bp。不同的NGS平台可产生不同长度的测序读段。例如,HiSeq和HiSeq X的平均读段长度为约300bp,而MiSeq平台的平均读段长度介于约300bp与600bp之间。在一些实施方式中,平均NGS读段长度为约100bp、约150bp、约200bp、约250bp、约300bp、约350bp、约400bp、约450bp、约500bp、约550bp、约600bp、约650bp、约700bp、约750bp、约800bp、约850bp、约900bp、约950bp、约1000bp、约1100bp、约1200bp、约1300、约1400bp、约1500bp、约1600bp、约1700bp、约1800bp、约1900bp或约2000bp。

[0165] 在一些实施方式中,多个分离的序列包含已经改组的参考序列的片段。改组序列的方法在本领域中是已知的,并且包括随机改组和Cantor改组。各种程序可用于改组核苷酸序列,并且将是本领域技术人员已知的(参见例如,www.bioinformatics.org/sms2/shuffle_dna.html)。如本文中所预期的,经改组的参考序列片段保留了参考序列的一种或多种特征,而与此同时与样本中的靶序列不同,并且因此期望通过例如blast的方法与靶序列相区分。

[0166] 在一些实施方式中,多个分离的序列包含已连接的参考序列片段。在一些实施方式中,将参考序列片段改组并连接,例如以产生具有期望的一个或多个长度的分离的序列。在一些实施方式中,分离的序列包含至少第一参考序列和第二参考序列的至少两个片段,所述至少两个片段已被改组和连接。在一些实施方式中,第一参考序列和第二参考序列在生物体的基因组中不相邻。在一些实施方式中,第一参考序列和第二参考序列来自不同的生物体。分离的序列可以包含至少1、2、3、4、5、6、7、8、9、10、11、12、13、14、15、16、17、18、19或20个参考序列片段,所述参考序列片段已经被改组并连接,其中所述参考序列片段中的任何一个参考序列片段都可能是来自相同参考序列或参考生物体,或来自不同参考序列或参考生物体的不相邻片段。

[0167] 在一些实施方式中,分离的序列是通过以下过程而来源于参考序列的:(a)通过参考序列片段的分布来表示至少一个参考序列;(b)划分成箱;(c)从所述箱的至少一个子集中选择多个参考序列片段;以及(d)对参考序列片段进行改组并任选地连接以生成分离的序列。

[0168] 在一些实施方式中,通过以下方式生成归一化对照的分离的序列:(a)生成在至少一个参考序列上使用滑动窗生成的重叠参考序列片段的分布;(b)将所述分布划分成箱;(c)从所述箱的至少一个子集中选择多个参考序列片段;以及(d)连接所述参考片段序列,从而生成分离的序列。

[0169] 在一些实施方式中,滑动窗包含1个碱基对(bp)、2bp、3bp、4bp、5bp、6bp、7bp、8bp、9bp、10bp、11bp、12bp、13bp、14bp、15bp、16bp、17bp、18bp、19bp或20bp的滑动窗。滑动窗是长度为N的窗,其可沿参考序列滑动以生成参考序列片段,所述参考序列片段会偏移所述窗的长度达例如1bp、2bp、3bp、4bp、5bp或更多。

[0170] 该分布可以是以下各项中任一项的分布:(1)GC含量百分比、(2)熵、(3)复杂性、

(4) 电子-离子相互作用电势 (EIIP)、(5) 长度, 或它们的组合。为了生成分布, 将参考序列分解为具有一个或多个指定大小的片段 (有时在本文中称为Kmer), 并将参考序列片段的数量针对给定的参数 (例如GC含量百分比、EIIP、长度或熵) 进行分箱。每个箱代表参数分布的百分比, 例如1%、2%、3%、4%、5%、6%、7%、8%、9%、10%、15%、20%、或25%。

[0171] 分布可以划分成任意数量的箱。例如, 可以将分布划分成2、3、4、5、6、7、8、9、10、20、50、100或1000个箱。在一些实施方式中, 每个箱代表参数 (例如GC含量百分比、熵、复杂性、长度或EIIP) 的分布的0.1%、0.5%、1%、2%、5%、10%、15%或20%。

[0172] 在一些实施方式中, 使用选自划分成5个箱的分布的至少3个箱中的一个箱的参考序列片段, 来生成归一化对照的每个组内的多核苷酸。在一些实施方式中, 通过使用选自分布中的至少1个、2个、3个、4个、5个、6个、7个、8个、9个、10个、11个、12个、13个、14个、15个、16个、17个、18个、19个或20个箱的参考序列片段来生成归一化对照中的每个组内的多核苷酸。在一些实施方式中, 将分布划分成10个箱, 并且参考序列片段选自来自GC含量百分比分布的第10、第30、第50、第70、第80、和第90百分位箱。在一些实施方式中, 通过使用选自分布中每个箱的参考序列片段来生成归一化对照的每个组内的多核苷酸。

[0173] 在一些实施方式中, 一个或多个参考序列片段选自分布 (例如GC含量百分比分布) 的一个箱。在一些实施方式中, 从分布中的每个箱中选择一个或多个参考序列片段。在一些实施方式中, 从分布的箱的子集中选择一个或多个参考序列片段。在一些实施方式中, 将从相同箱中选择的参考序列片段连接起来以形成具有期望长度的分离的序列。这允许归一化对照中的分离的序列通过对选定的箱进行子取样来对整个分布范围内的各种参数 (例如GC含量、EIIP、长度和熵) 的分布进行建模, 而无需以参考序列与归一化对照之间的1:1对应关系, 使用高得离谱的数量的归一化对照寡核苷酸来对整个分布参考序列进行建模。

[0174] 在一些实施方式中, 每个组内的多核苷酸具有以下项中的至少三者: 与来自所述参考序列分布中的对应箱的参考序列片段 (1) 相似的GC含量百分比、(2) 相似的熵、(3) 相似的EIIP、(4) 相似的长度、或 (5) 相似的复杂性。

[0175] 相似的GC含量百分比、EIIP、长度、复杂性或熵可以指参考序列和归一化对照的平均GC含量、EIIP、长度、复杂性或熵的比较。例如, 如果参考序列的平均GC含量、EIIP、长度、复杂性或熵在归一化对照的平均GC含量、EIIP、长度、复杂性或熵的0.1、0.5、1、2、3、4、5、6、7、8、9、10、11、12、13、14或15个百分点内, 则归一化对照和参考序列可具有类似的GC含量百分比、EIIP、长度、复杂性或熵。替代地或另外, 类似的GC含量、EIIP、长度、复杂性或熵可以指参考序列和归一化对照序列中的一些或全部序列上的GC含量、EIIP、长度、复杂性或熵的分布。确定相似程度的方法将是本领域中是已知的, 包括但不限于, Kolmogorov-Smirnov检验、Z检验、Q-Q取向和方差分析 (Analysis of Variance, ANOVA)。普通技术人员将能够基于所讨论的分布的特征 (例如, 正态、泊松、高斯或偏斜) 选择适当的检验。

[0176] 在一些实施方式中, 多个分离的序列包含已连接的参考序列片段。在一些实施方式中, 参考序列片段被连接而不被改组。如果经连接以形成分离的序列的参考序列片段在参考序列中不相邻, 或者来自不同参考序列, 则在不相邻序列之间或来自不同来源的序列之间存在“接头”应允许在NGS后通过blast比对鉴定分离的序列。接头的存在允许归一化对照中的分离的序列与样本中的靶序列区分开。不希望受理论的束缚, 预期对于这种方法, 介于约15bp与约50bp之间, 介于约20bp与约35bp之间的序列片段可为优选的, 因为太小的序

列,例如低于比对器的种子长度的序列,架构可能不足以独特地产生稳健的比对,而太大的序列将不会具有通过单独NGS读段捕获的接头。

[0177] 在一些实施方式中,分离的序列包含连接的至少一个参考序列的至少两个片段,并且所述至少两个片段在至少一个参考序列中不相邻。即,至少两个片段分开至少1bp、至少2bp、至少3bp、至少4bp、至少5bp、或任何数量的bp。

[0178] 在一些实施方式中,分离的序列包含至少第一参考序列的至少第一片段和至少第二参考序列的至少第二片段,其中所述至少第一和第二片段是连接的,并且其中所述第一和第二参考序列不是相同的参考序列。例如,第一片段来自人腺病毒,并且第二片段来自白色念珠菌(*Candida albicans*)。作为进一步的示例,分离的序列可包含嵌套的参考序列片段:例如,来自第二参考生物体的参考序列片段内的来自第一参考生物体的参考序列片段,任选地与来自另一种不同生物体的参考序列片段连接。分离的序列可包含至少1、2、3、4、5、6、7、8、9、10、11、12、13、14、15、16、17、18、19或20个参考序列片段,所述参考序列片段已经被连接,其中所述参考序列片段中的任何一个参考序列片段都可能是来自相同参考序列或参考生物体,或来自不同参考序列或参考生物体的不相邻片段。

[0179] 在一些实施方式中,分离的序列(例如,改组或连接的分离的序列)不被设计用于切割人DNA的90k gRNA阵列中的任何CRISPR相关(Cas)指导RNA(gRNA)切割。在一些实施方式中,分离的序列可以被设计为排除特定的序列特征。例如,可以选择不包含诸如CRISPR/Cas gRNA识别位点、限制性位点、转录因子结合位点或重复序列的特征的改组或连接的分离的序列。

[0180] 序列特性

[0181] 在一些实施方式中,至少三组多核苷酸中的每个组的序列具有与测序样本中的靶序列共有的至少一种特性。在一些实施方式中,将组分序列设计为具有与测序样本中的靶序列共有的至少一种特性。在一些实施方式中,分离的序列是从已经选择的生物体或克隆载体中的序列中分离的,因为已知该分离的序列具有与测序样本中的靶序列共有的至少一种特性。

[0182] 在一些实施方式中,至少一种特性包括转座元件序列、病毒序列、细菌序列、真菌序列、真核寄生虫序列、或一个或多个人类基因序列的序列特性。

[0183] 在一些实施方式中,转座元件的特性包括转座元件序列。转座元件是可以改变基因组中的位置的DNA序列。转座元件(转座子)通常由一对反向末端重复序列(inverted terminal repeat, ITR)组成,该ITR位于编码转座酶的开放阅读框的侧翼。因此,转座元件的示例性特性包括ITR序列、转座酶序列或转座子的任何序列。转座子的任何序列或转座子的序列特性可被并入本公开的组分序列中。本公开的分离的序列可包含转座子的任何序列或转座子的序列特性。

[0184] 在一些实施方式中,病毒序列的特性包括GC含量百分比、重复序列元件、反向末端重复(ITR)、内部核糖体进入位点(IRES)、蛋白质编码序列、转录后调控元件、转录调控元件、启动子序列、顺式作用RNA元件、RNA结构元件、基因组包装信号、5'非翻译区(5'UTR)序列、3'非翻译区(3'UTR)序列,或它们的组合。可以将病毒的任何序列或病毒的序列特性并入本公开的组分序列中。

[0185] 在一些实施方式中,细菌序列的特性包括GC含量百分比、重复序列、微型RNA结合

位点、内部核糖体进入位点(IRES)、蛋白质编码序列、转录调控元件、启动子序列、5'UTR序列、3'UTR序列,或它们的组合。在一些实施方式中,重复序列元件包含polyA基序、polyT基序、polyG基序、polyC基序、二核苷酸基序、三核苷酸基序、四核苷酸基序、五核苷酸基序、六核苷酸基序、七核苷酸基序、八核苷酸基序、非核苷酸基序、散布的重复序列元件、核糖体RNA序列、转移RNA(tRNA)序列、末端反向重复序列(TIR)、非自主微型反向重复转座元件(non-autonomous miniature inverted repeat,MITE)、成簇规律间隔短回文重复序列(Clustered Regularly Interspaced Short Palindromic Repeat,CRISPR),或它们的组合。polyA基序、polyT基序、polyG基序和polyC基序是一系列A、T、G和C(例如,4-50个A、T、G或C)。细菌的任何序列或细菌的序列特性都可以并入本公开的组分序列中。本公开的分离的序列可以包含细菌的序列或细菌的序列特性。

[0186] 在一些实施方式中,真菌序列的特性包括GC含量百分比、重复序列元件、微型RNA结合位点、内部核糖体进入位点(IRES)、蛋白质编码序列、转录调控元件、启动子序列、5'UTR序列、3'UTR序列、着丝粒序列、端粒序列、子端粒序列、线粒体序列,或它们的组合。可以将真菌的任何序列或真菌的序列特性并入本公开的组分序列中。本公开的分离的序列可以包含真菌的序列或真菌的序列特性。

[0187] 在一些实施方式中,一个或多个人类基因的特性包括GC含量百分比、重复序列、蛋白质编码序列、内含子序列、5'UTR序列、3'UTR序列、转录调控元件序列、启动子序列、微型RNA结合序列,或它们的组合。在示例性实施方式中,一个或多个人类基因的特性是编码由一个家族中的多个蛋白质共享的保守蛋白质结构域的序列。蛋白质结构域的非限制性示例是DNA结合结构域,例如锌指DNA结合结构域或同源结构域。在示例性实施方式中,一个或多个人类基因的特性是由同源微型RNA调控的所有基因共享的微型RNA结合位点。可以将一个或多个人类基因的任何序列或一个或多个人类基因的序列特性并入本公开的组分序列中。本公开的分离的序列可以包含一个或多个人类基因的序列或一个或多个人类基因的序列特性。

[0188] 在一些实施方式中,真核寄生虫序列的特性包括GC含量百分比、重复序列、微型RNA结合位点、蛋白质编码序列、转录调控元件、启动子序列、5'UTR序列、3'UTR序列、着丝粒序列、端粒序列、子端粒序列、线粒体序列,或它们的组合。真核寄生虫的任何序列或序列特性都可以并入本公开的组分序列中。本公开的分离的序列可以包含真核寄生虫的序列或真核寄生虫的序列特性。

[0189] 在一些实施方式中,归一化对照序列和靶序列共享的至少一种特性包括熵。熵是核苷酸序列的信息含量和复杂性的量度。计算熵的方法对于本领域普通技术人员而言将是已知的(参见例如,Bioinformatics,第27卷,2011,第1061-1067页)。

[0190] 在一些实施方式中,归一化对照序列和靶序列共享的至少一种特性包括复杂性。复杂性(用N表示)是在给定区段的多核苷酸或基因组的组件中的独特或非重复核苷酸的碱基对的数量。如果多核苷酸中的一些多核苷酸被重复,则这与序列的长度(L)不同。

[0191] 在一些实施方式中,归一化对照序列和靶序列共享的至少一种特性包括GC含量百分比。

[0192] 在本公开的归一化对照的一些实施方式中,多核苷酸的一个或多个组或亚组的GC含量与测序样本中的靶序列相同。在一些实施方式中,该组或亚组中的整个多核苷酸的GC

含量与靶序列相同。在一些实施方式中,组分序列和/或分离的序列的GC含量与靶序列相同。已知GC含量会影响NGS期间的读段计数(GC偏差)。具有低或高GC含量的序列趋向于在NGS期间产生的读段数量不足。因此,在NC组合物中模拟靶序列的GC含量增加了NC对NGS期间的靶序列的行为进行建模的能力。

[0193] 在一些实施方式中,一个组内的多核苷酸具有相同的GC含量百分比。在一些实施方式中,一个组内的多核苷酸具有与每个其他组内的多核苷酸相同的GC含量百分比。在一些实施方式中,一个组内的多核苷酸不具有与任何其他组内的多核苷酸相同的GC含量百分比。

[0194] 在一些实施方式中,每个组内的多核苷酸的GC含量百分比介于40%与60%之间,包括端点。在一些实施方式中,每个组内的多核苷酸的GC含量百分比介于43%与56%之间,包括端点。

[0195] 在一些实施方式中,归一化对照序列和靶序列共享的至少一种特性包括电子-离子干扰电势(EIIP)。EIIP可以定义为核苷酸序列中的核苷酸的离域电子的平均能量。通过将每个核苷酸替换为对应的EIIP值,可以将核苷酸序列转换为数字EIIP序列。例如, $A=0.1260$, $C=0.1340$, $T=0.1335$, $G=0.0806$ 。在一些实施方式中,EIIP值可以在整个或部分序列上求和,例如在分离的序列或参考序列片段上求和。

[0196] 在一些实施方式中,归一化对照序列和靶序列共享的至少一种特性包括GC含量、EIIP和熵。GC含量、EIIP和熵均取决于序列,因此共同变化。在一些实施方式中,归一化对照与靶序列共享这些参数中的所有三个参数。

[0197] 可以通过本领域中已知的多种统计方法来确定归一化对照序列与靶序列或参考序列之间的参数(例如GC含量百分比、EIIP、熵或复杂性)的相似程度。这些统计方法包括但不限于方差分析(ANOVA)、t检验(例如学生t检验)、Kolmogorov Smirnov(KS)检验和Q-Q曲线。在一些实施方式中,归一化对照包含经修饰的核苷酸。设想在本公开的范围内的任何类型的核苷酸修饰。下文描述本公开的核苷酸修饰的示例性但非限制性示例。

[0198] 由本公开的方法使用的核苷酸修饰可以在任何核苷酸(例如腺嘌呤、胞嘧啶、鸟嘌呤、胸腺嘧啶或尿嘧啶)上发生。这些核苷酸修饰可发生在脱氧核糖核酸(DNA)或核糖核酸(RNA)上。这些核苷酸修饰可以发生在双链或单链DNA分子上,或发生在双链或单链RNA分子上。

[0199] 在一些实施方式中,核苷酸修饰包括腺嘌呤修饰或胞嘧啶修饰。

[0200] 在一些实施方式中,腺嘌呤修饰包括腺嘌呤甲基化。在一些实施方式中,腺嘌呤甲基化包括N6-甲基腺嘌呤(6mA)。N6-甲基腺嘌呤(6mA)存在于原核基因组和真核基因组两者中。

[0201] 在一些实施方式中,腺嘌呤甲基化包括EcoKI甲基化。EcoKI甲基化是由EcoKI甲基化酶进行的一种类型的DNA核苷酸修饰。EcoKI甲基化酶修饰序列AAC(N6)GTGC和GCAC(N6)GTT中的腺嘌呤残基。在原核生物中发现了EcoKI甲基化酶和EcoKI甲基化。

[0202] 在一些实施方式中,腺嘌呤修饰包括在N6处通过甘氨酸修饰的腺嘌呤(momylation)。甲基化改变N6-(1-乙酰氨基)-腺嘌呤的腺嘌呤。甲基化作用发生在病毒例如噬菌体中。

[0203] 在一些实施方式中,修饰包括胞嘧啶修饰。在一些实施方式中,基因组中胞嘧啶修

饰的丰度和类型基于物种而变化。在一些实施方式中,基因组中胞嘧啶修饰的位置(在例如特定的限制性内切酶识别位点内)基于物种而变化。

[0204] 在一些实施方式中,胞嘧啶修饰包括5-甲基胞嘧啶(5mC)、5-羟基甲基胞嘧啶(5hmC)、5-甲酰基胞嘧啶(5fC)、5-羧基胞嘧啶(5caC)、5-葡萄糖基羟甲基胞嘧啶(5ghmC)或3-甲基胞嘧啶(3mC)。

[0205] 在一些实施方式中,胞嘧啶修饰包括胞嘧啶甲酰化。在一些实施方式中,胞嘧啶甲基化包括5-甲基胞嘧啶(5mC)或N4-甲基胞嘧啶(4mC)。在一些实施方式中,在细菌中发现了4mC胞嘧啶甲基化。在一些实施方式中,胞嘧啶甲基化包括Dcm甲基化。在一些实施方式中,胞嘧啶甲基化包括DNMT1甲基化,DNMT3A甲基化或DNMT3B甲基化。

[0206] 在一些实施方式中,胞嘧啶甲基化包括CpG甲基化、CpA甲基化、CpT甲基化、CpC甲基化,或它们的组合。在一些实施方式中,胞嘧啶甲基化包括CpG甲基化。

[0207] 在一些实施方式中,胞嘧啶修饰包括5-羟甲基胞嘧啶(5hmC)、5-甲酰基胞嘧啶(5fC)、5-羧基胞嘧啶(5caC)、5-葡萄糖基羟甲基胞嘧啶、或3-甲基胞嘧啶。

[0208] 靶生物体

[0209] 本公开提供了用于确定样本中的一种或多种靶生物体的滴度的组合物和方法。靶生物体是包含一种或多种本文所述靶序列的生物体。

[0210] 在一些实施方式中,样本包含宿主生物体和一种或多种非宿主生物体的混合物,并且靶生物体是一种或多种非宿主生物体。在一些实施方式中,宿主生物体是靶生物体。

[0211] 在一些实施方式中,非宿主靶生物体是病原体。在一些实施方式中,靶生物体是病毒、细菌、真菌或真核病原体。例如,在一些实施方式中,靶生物体是人类临床样本中的病毒病原体,并且靶序列是该病毒病原体的序列。

[0212] 在一些实施方式中,靶生物体是共生体或共生生物体。

[0213] 在一些实施方式中,靶生物体是宏基因组样本中的物种混合物中的物种。

[0214] 靶序列

[0215] 在一些实施方式中,将NC设计为考虑影响测序样本中靶序列向NGS后的测序读段的转化的可能变量。例如,NC序列可经设计以匹配靶序列的核酸的GC含量,所述靶序列的在起始样本中的丰度是使用本公开的归一化对照和方法确定的。

[0216] 在本公开的归一化对照的一些实施方式中,组分序列和/或分离的序列具有与测序样本中的靶序列共有的至少一种特性。在靶序列与组分序列之间共享的共同特性使得NC的组分序列能够在样本和/或文库制备和测序期间模拟靶序列。这允许NC对从样本中的靶序列的初始浓度至NGS之后靶序列的多个读段的转化进行建模。

[0217] 在一些实施方式中,归一化对照中的多核苷酸的至少一个亚组的序列与测序样本中的靶序列共享至少一种序列特性。在一些实施方式中,至少一种序列特性包括转座元件序列、病毒序列、细菌序列、真菌序列、真核寄生虫序列,或一种或多种人类基因序列的序列特性。

[0218] 预期靶序列的所有序列特性都在本公开的归一化对照组合物的范围内。在一些实施方式中,至少一种特性包括转座元件序列、病毒序列、细菌序列、真菌序列、真核寄生虫序列、或一个或多个人类基因序列的序列特性。在一些实施方式中,至少一种特性包括GC含量、熵、复杂性、长度、EIIP,或它们的组合。

[0219] 在一些实施方式中,测序样本包含宿主核酸和非宿主核酸的混合物。在一些实施方式中,非宿主包含一个或多个靶序列。

[0220] 在一些实施方式中,宿主是真核生物。在一些实施方式中,宿主是昆虫、动物或植物。在一些实施方式中,动物是人。

[0221] 在一些实施方式中,宿主是任何哺乳动物生物体。在特定的实施方式中,哺乳动物是人。在其他实施方式中,哺乳动物是牲畜动物,例如马、绵羊、牛、猪或驴。在其他实施方式中,哺乳动物生物体是家养宠物,例如猫、狗、沙鼠、小鼠或大鼠。在其他实施方式中,哺乳动物是一种类型的猴子。

[0222] 在一些实施方式中,宿主是任何鸟类或禽类生物体。禽类生物体包括但不限于鸡、火鸡、鸭和鹅。

[0223] 在一些实施方式中,宿主是昆虫。昆虫包括但不限于蜜蜂、独居蜂、蚂蚁、苍蝇、黄蜂和蚊子。

[0224] 在一些实施方式中,宿主是植物。在一些特定的实施方式中,植物是稻、玉米、小麦、玫瑰、葡萄、咖啡、水果、番茄、马铃薯或棉花。

[0225] 在一些实施方式中,非宿主包含共生体、共生生物体、寄生虫或病原体。在一些实施方式中,非宿主包含一种或多种物种的共生体、共生生物体、寄生虫或病原体。

[0226] 在一些实施方式中,非宿主是病原体。在一些实施方式中,非宿主是病毒物种、细菌物种、真菌物种、或真核寄生虫物种。在一些实施方式中,非宿主是藻类物种。

[0227] 在一些实施方式中,真核寄生虫是哺乳动物寄生虫。在一些实施方式中,该寄生虫是蠕虫。在其他实施方式中,该寄生虫是引起疟疾的寄生虫。在其他实施方式中,该寄生虫是引起利什曼病的寄生虫。在其他实施方式中,该寄生虫是变形虫。

[0228] 在一些实施方式中,非宿主是细菌物种。在特定的实施方式中,细菌是引起结核的细菌。

[0229] 在一些实施方式中,靶序列是测序样本中的非宿主序列。

[0230] 在一些实施方式中,靶序列包含参考序列。在一些实施方式中,靶序列包含多个参考序列。在一些实施方式中,靶序列包括至少2个、至少10个、至少20个、至少50, least 100个、至少200个、至少300个、至少400个、至少500个、至少600个、至少700个、至少800个、至少900个、至少1,000个、至少1,200个、至少1,300个、至少1,400个、至少1,500个、至少1,600个、至少1,700个、至少1,800个、至少1,900个、至少2,000个、至少2,200个、至少2,400个、至少2,600个、至少2,800个、至少3,000个、至少4,000个、至少5,000个、至少6,000个、至少7,000个、至少8,000个、至少9,000个或至少10,000个参考序列。在一些实施方式中,每个参考序列来自不同的生物体。在一些实施方式中,多个参考序列是来自不同生物体的基因组DNA或cDNA序列。在一些实施方式中,靶序列包含来自至少2种、至少10种、至少20种、至少50, least 100种、至少200种、至少300种、至少400种、至少500种、至少600种、至少700种、至少800种、至少900种、至少1,000种、至少1,200种、至少1,300种、至少1,400种、至少1,500种、至少1,600种、至少1,700种、至少1,800种、至少1,900种、至少2,000种、至少2,200种、至少2,400种、至少2,600种、至少2,800种、至少3,000种、至少4,000种、至少5,000种、至少6,000种、至少7,000种、至少8,000种、至少9,000种或至少10,000种生物体的多个参考序列。

[0231] 在一些实施方式中,靶序列包含非宿主序列,并且非宿主是病毒物种、细菌物种、

真菌物种或真核寄生物种。

[0232] 在一些实施方式中,非宿主包含共生体。共生体,有时称为互惠或共生生物体,是与另一种生物体有关系的生物体,所述关系可为双方带来互利。

[0233] 在一些实施方式中,非宿主包含共生生物体。在共生生物体相互作用中,一种生物体(例如,非宿主)受益,而另一种(例如,宿主)则不受影响。共生生物体或互惠生物体可能取决于其宿主的食物栖息地(food shelter)、支撑、运输或它们的组合。宿主可以从相互作用中获得多种益处,包括但不限于防止感染、改善消化和免疫调节。互惠和/或共生关系的非限制性示例是人类宿主与肠道内居住的微生物群体(肠道微生物群、或微生物群系)的互惠和/或共生关系。肠道菌群帮助维持肠道粘膜屏障,提供营养物质(例如维生素),预防病原体,并有助于维持适当的免疫功能。

[0234] 在一些实施方式中,非宿主包括多个物种,或来自相同物种的不同生物体。在一些实施方式中,非宿主包含至少5种、至少10种、至少50种、至少100种、至少200种、至少500种、至少1000种、至少1,500种、或至少2,000种物种,或来自相同物种的不同生物体,或它们的组合。在一些实施方式中,非宿主包含至少2,000种物种,例如至少2,000种肠道微生物群系物种或微生物群系生物体。

[0235] 在一些实施方式中,靶序列是与致病性相关的非宿主序列。与致病性相关的示例性非宿主序列包括但不限于与真核寄生虫相关的抗生素抗性基因,毒力因子或标志物致病性。

[0236] 毒力因子是在感染宿主时有助于微生物如细菌,病毒或真菌致病性的基因或一组基因。毒力是指微生物对宿主造成的损害程度,更大的毒力导致更大的损害。

[0237] 示例性但非限制性的毒力因子包括有助于宿主中利基定殖的基因(例如,非宿主与宿主细胞的附着),免疫逃避(逃避宿主免疫应答),免疫抑制(抑制肝细胞毒性)。宿主的非宿主免疫应答),(细胞内非宿主的)细胞进入和退出,以及从宿主获得营养。用于辅助和促进宿主定殖的示例性毒力因子包括但不限于粘附素,侵袭素和抗吞噬因子。

[0238] 示例性毒力因子包括增加对宿主损害的因子,包括但不限于调节宿主GTP酶的毒素,破坏性酶和蛋白质。示例性毒素包括内毒素和外毒素。内毒素(脂多糖或LPS)是一种细菌细胞壁成分,能够触发宿主的炎症。外毒素是由某些细菌主动分泌的毒素,对宿主具有广泛的作用,包括但不限于抑制特定的生化途径。示例性外毒素包括但不限于破伤风梭菌分泌的破伤风毒素,肉毒梭菌分泌的肉毒杆菌毒素,以及大肠杆菌,产气荚膜梭菌,炭疽芽孢杆菌和艰难梭菌分泌的毒素。在一些实施方式中,真菌分泌内毒素(称为霉菌毒素)。示例性的霉菌毒素包括曲霉产生的黄曲霉毒素。

[0239] 在一些实施方式中,毒力因子是破坏性酶。破坏性酶是通过破坏宿主细胞和组织而引起毒力的酶。示例性破坏性酶包括但不限于蛋白酶,脂肪酶和DNA酶。例如,由细菌(如大肠杆菌和金黄色葡萄球菌)产生的溶血素分解宿主细胞,包括红细胞。

[0240] 在一些实施方式中,毒力因子是调节宿主GTP酶活性的蛋白质。毒力因子可以充当鸟嘌呤核苷酸交换因子(GEF)或GTP酶激活蛋白(GAP)来修饰宿主GTP酶的活性,或者共价修饰宿主GTP酶本身。

[0241] 示例性病毒毒力基因包括控制病毒潜伏期的病毒基因。另外的示例性病毒毒力基因包括有助于逃避宿主免疫应答的病毒基因以及病毒生命周期的各个方面,例如结合和进

入宿主细胞或病毒体产生和释放。

[0242] 在一些实施方式中,毒力因子是染色体的和固有的。例如,细菌毒力因子例如脂多糖(内毒素)是染色体的。

[0243] 在一些实施方式中,毒力因子由通过水平基因转移传播的流动遗传元件编码。在一些实施方式中,移动遗传元件是质粒,转座元件,噬菌体元件或核酶。在一些实施方式中,在获取毒力因子之前,毒力因子的水平基因转移将作为共生或共生(共生)微生物的微生物转化为病原微生物。

[0244] 在一些实施方式中,非宿主是病毒,细菌或真菌,并且靶序列包含毒力因子序列。

[0245] 在一些实施方式中,非宿主是细菌,并且靶序列包含抗生素抗性基因的序列。在一些实施方式中,抗生素抗性基因由通过水平基因转移传播的流动遗传元件编码。在一些实施方式中,移动遗传元件是质粒,转座元件,噬菌体元件或核酶。

[0246] 在一些实施方式中,抗生素抗性基因传达对喹诺酮类, β -内酰胺类,大环内酯类,四环素类,磺酰胺类,氨基糖苷类或万古霉素类的抗性。

[0247] 在一些实施方式中,测序样本包括宏基因组学样本。如本文所用,术语“元基因组学”表示例如使用本公开的组合物和方法对混合生物样本进行直接分析。

[0248] 在一些实施方式中,宏基因组学样本是环境样本。示例性环境样本包括但不限于土壤,水和空气样本。在一些实施方式中,环境样本包括从人造表面,例如从医院的表面获取的样本。

[0249] 在一些实施方式中,样本包括环境样本,并且靶序列包括毒力因子序列,抗生素抗性基因或与致病性相关的序列。例如,通过侧向基因转移产生抗生素抗性是主要的公共卫生问题。因此,使用本公开的组合物和方法,对从使用抗生素的环境例如医院或农场采集的宏基因组样本中的抗生素抗性基因的频率进行量化是有益的。

[0250] 在一些实施方式中,本公开的归一化对照被设计成在样本提取和/或文库制备期间模拟测序样本中靶序列即非宿主序列的特性。在一些实施方式中,归一化对照的多核苷酸群是从核酸样本中纯化的,并且该核酸样本是从病毒,细菌,真菌或真核寄生虫中分离或衍生的。在一些实施方式中,从其纯化NC的病毒,细菌,真菌或真核生物与非宿主的病毒,细菌,真菌或真核寄生虫相似但不相同。

[0251] 在一些实施方式中,非宿主是病毒,并且归一化对照的多核苷酸是从相关病毒或具有与非宿主相似性质的病毒中纯化的。例如,如果非宿主是慢病毒,则从相关物种慢病毒中纯化归一化对照的多核苷酸。在一些实施方式中,病毒是T4噬菌体或巨细胞病毒(CMV)。或者,如果非宿主是细菌,则从相关细菌种类或具有与非宿主性质相似的细菌种类中纯化归一化对照多核苷酸。

[0252] NC可以为一种或多种生物。例如,NC可以包含从一种或多种生物或生物物种(例如病毒,细菌,真菌或真核生物)提取的多核苷酸,其模拟混合的宿主/非宿主样本中非宿主序列的一种或多种特性。

[0253] 在某些实施方式中,NC用于确定样本中不同核酸的相对量。例如,归一化对照可用于确定来自被感染宿主(例如感染病毒或细菌的人类宿主)样本中病原体(非宿主)的丰度值。可以通过将映射到病原体的NGS读段除以映射到NC的NGS读段的数量来补偿病原体的丰度值,以补偿背景宿主的不同含量。然后将此值映射到标准曲线,该曲线以相同的方式

生成和标准化,并且可以从标准曲线推断出病原体滴度。使用这样的方法,则不一定需要确定绝对病原体负荷。

[0254] 在某些实施方式中,NC用于确定样本中不同核酸的绝对量。例如,NC可以用于确定来自被感染宿主的样本(例如,来自被病毒或细菌感染的人类宿主的样本)中病原体(非宿主)的绝对量。可以通过将映射到病原体的NGS读段的数目除以映射到NC来补偿不同背景宿主含量的NGS读段的数目来确定病原体的丰度值。基于该比率和NC的输入量(例如,噬菌体或多核苷酸的拷贝),可以确定相对病原体基因组拷贝。然后可以针对病原体基因组大小校正该值,以在起始材料中得出“绝对”病原体丰度。

[0255] 在一些实施方式中,从其纯化归一化对照多核苷酸的感兴趣的物种与非宿主物种相似但不相同。

[0256] 在一些实施方式中,测序样本包括来自已被诊断患有某种癌症的受试者的癌症相关样本。癌症的进展是一个多步骤的过程,需要随着时间的推移积累遗传变化。在一些实施方式中,导致癌症的遗传改变包括激活原癌基因并且使肿瘤抑制基因和DNA修复基因失调的遗传改变。导致癌症的遗传变化使以前正常的细胞逃避了细胞分裂的调控,并不受限制地分裂。因此,癌症包括具有不同突变的异质细胞群体。因此,在一些实施方式中,癌症相关样本包括细胞群体,该细胞群体包含相对于受试者的非癌细胞或细胞中的核酸具有相对于核酸至少一种遗传变化的核酸(例如基因,DNA或RNA)。来自没有癌症的受试者,或来自患有不同类型癌症的受试者。

[0257] 在一些实施方式中,仅在癌症样本中的异质细胞群体中很少发生感兴趣的突变。但是,癌症样本中的特定突变可预测对治疗的临床反应。因此,使用本公开的组合物和方法鉴定和量化异质癌症样本中突变的发生是有益的。当细胞类型和/或突变很少见时,本公开的组合物和方法特别有益并且有用。

[0258] 在一些实施方式中,癌细胞中的至少一种遗传变化包括单核苷酸多态性(SNP),插入,缺失,倒位或染色体重排。

[0259] 在一些实施方式中,至少一种遗传机会与原癌基因的活化或肿瘤抑制基因的失活相关。示例性原癌基因包括但不限于人表皮生长因子受体2(HER2),RAS,MYC原癌基因(MYC),SRC原癌基因(SRC),端粒酶,BCL-2,细胞凋亡调节剂(BCL-2)和表皮生长因子受体(EGFR)。示例性的肿瘤抑制基因包括但不限于肿瘤蛋白p53(p53),成视网膜细胞瘤(Rb),腺瘤性息肉病大肠杆菌(APC),乳腺癌敏感性基因1(BRCA1)和乳腺癌敏感性基因2(BRCA2)。

[0260] 在一些实施方式中,靶序列包括在癌症中具有至少一种遗传改变的序列。

[0261] 在一些实施方式中,从中纯化归一化对照多核苷酸的核酸样本中的核酸序列与测序样本中的靶序列不同。在一些实施方式中,核酸样本中核酸的序列具有小于或等于1%,2%,3%,4%,5%,6%,7%,8%,9%,10%,11%,12%,13%,14%,15%,16%,17%,18%,19%,20%,30%,40%,50%,60%,70%,80%,与测序样本中的靶序列具有85%,90%,95%,96%,97%,98%或99%的同一性。

[0262] 浓度

[0263] 在一些实施方式中,NC被设计为在DNA提取期间添加至宿主(例如患者)样本中,以作为非宿主(例如病原体)鉴定管线的一部分对序列读段进行归一化。归一化对照片段共同设计为使得它们代表从样本获得的总序列读段的一小部分。归一化对照片段总体上应该代

表从样本获得的序列读段的一小部分(<5%)。在一些实施方式中,归一化对照片段的频率与样本中靶序列的频率匹配。

[0264] 设想在多核苷酸组之间的浓度的每种可能的排列都在本公开的范围内。

[0265] 在一些实施方式中,每组的多核苷酸处于相同浓度。例如,在一些实施方式中,在归一化对照组合物中总浓度为3pg/ μ L,三组中的每组的相同浓度为每微升(μ L)1皮克(pg)。

[0266] 在一些实施方式中,每组的多核苷酸的浓度不同。在一些实施方式中,多核苷酸的每组可以不同的浓度存在。例如,在某些实施方式中,如果三个组以1:2:4的浓度比存在,并且第一组的浓度为1pg/ μ L,则第二组的浓度为2pg/ μ L,第三组的浓度为4pg/ μ L,在归一化对照组合物中的总浓度为7pg/ μ L。

[0267] 设想在组内的亚组之间以及在不同组的亚组之间的浓度的每种可能的排列都在本公开的范围内。

[0268] 在一些实施方式中,至少三个组包括至少一组,其中该组中的至少三个亚组中的多核苷酸的浓度是相等的,以及至少一组,其中该组中的亚组的浓度是不相等的。

[0269] 在一些实施方式中,至少三个组中的每个中的至少三个亚组中的每个中的多核苷酸的浓度不等于该组中其他亚组的浓度。在一些实施方式中,组中至少三个亚组中每一个的浓度以线性序列或几何序列分布。线性序列是数字序列,其中所述序列中的下一个数字每次相对于前一个数字增加或减少相同的量。通常,线性序列由以下公式表示:

[0270] $u_n = dx_n + c$

[0271] 其中d是序列中连续项之间的第一个差异,n是序列中的项,并且c是常数。在其中浓度为线性序列的组中存在三个亚组的示例性非限制性实施方式中,这三个亚组的浓度为例如1pg/ μ L,2pg/ μ L和3pg/ μ L。

[0272] 几何序列是数字序列,其中第一项之后的每个项都可以通过将前一项乘以固定的非零数(称为“公共比率”)来找到。通常,几何序列由下式表示:

[0273] $u_n = u_1 r^{n-1}$

[0274] 其中r是公共比率。在一组示例性的非限制性实施方式中,其中组中的三个亚组的浓度是几何序列,则这三个亚组的浓度为例如2pg/ μ L,4pg/ μ L和8pg/ μ L。每组多核苷酸内亚组的相对浓度是本公开归一化对照的有利特征。在NGS之后,映射回多核苷酸的每个亚组和/或组的测序读段的相对丰度将反映归一化对照组合物以及因此样本中NGS之前的亚组和/或组的相对浓度。因此,计算组或亚组内读段的相对丰度可提供一种内部质量控制措施,并可以为测序样本中靶序列的初始浓度提供更准确的模型。

[0275] 在一些实施方式中,组中的一个亚组具有基线浓度,并且该组中非基线亚组的浓度是基线的整数倍。在一些实施方式中,组中的一个亚组具有基线浓度,并且该组中每个亚组的浓度是该组中不包括基线的另一亚组的浓度的两倍。在一些实施方式中,该组中的至少三个亚组以1:2:4的浓度比存在。

[0276] 在一些实施方式中,以允许单个NC序列的浓度的宽动态范围的方式混合多种多核苷酸。浓度范围可以从10¹到10⁷,某些片段的浓度处于中间水平。使用包括QPCR在内的多种方法对归一化对照进行验证。

[0277] 在一些实施方式中,在所述归一化对照组合物中,多核苷酸群体的浓度介于约0.005皮克(pg)/微升(μ L)与约5000pg/ μ L之间、介于约0.005pg/ μ L与约1000pg/ μ L之间、介

于约0.005pg/ μ L与约700pg/ μ L之间、介于约0.005pg/ μ L与约500pg/ μ L之间、介于约0.005pg/ μ L与约100pg/ μ L之间、介于约0.005pg/ μ L与约50pg/ μ L之间、介于约0.05pg/ μ L与约5000pg/ μ L之间、介于约0.05pg/ μ L与约1000pg/ μ L之间、介于约0.05pg/ μ L与约700pg/ μ L之间、介于约0.05pg/ μ L与约500pg/ μ L之间、介于约0.05pg/ μ L与约100pg/ μ L之间、介于约0.05pg/ μ L与约50pg/ μ L之间、介于约0.5pg/ μ L与约1000pg/ μ L之间、介于约0.5pg/ μ L与约700pg/ μ L之间、介于约0.5pg/ μ L与约500pg/ μ L之间、介于约0.5pg/ μ L与约100pg/ μ L之间、介于约0.05pg/ μ L与约50pg/ μ L之间、介于约5pg/ μ L与约1000pg/ μ L之间、介于约5pg/ μ L与约700pg/ μ L之间、介于约5pg/ μ L与约500pg/ μ L之间、介于约5pg/ μ L与约100pg/ μ L之间或介于约5pg/ μ L与约50pg/ μ L之间。在一些实施方式中,在归一化对照组合物中,多核苷酸群体的浓度介于约0.005皮克 (pg) /微升 (μ L) 与500pg/ μ L之间。在一些实施方式中,在归一化对照组合物中多核苷酸群体的浓度介于约0.5pg/ μ L与约50pg/ μ L之间。在一些实施方式中,在归一化对照组合物中多核苷酸群体的浓度介于约10pg/ μ L与约20pg/ μ L之间。在一些实施方式中,在归一化对照组合物中,多核苷酸群体的浓度为介于约0.1pg/ μ L与约0.2pg/ μ L之间。在一些实施方式中,在归一化对照组合物中,多核苷酸群体的浓度为1.4pg/ μ L,包括端点在内。

[0278] 索引编制

[0279] 在一些实施方式中,归一化对照多核苷酸包含编码唯一分子标识符的序列。在一些实施方式中,编码唯一分子标识符的序列包括第一索引序列或第二索引序列。

[0280] 本公开的示例性非限制性索引策略使用两个索引,第一索引和第二索引。在一些实施方式中,编码独特分子标识符的序列包括共同的引物序列,索引序列和衔接子序列。在一些实施方式中,共同引物序列包含NGS的流动池附着位点。

[0281] 在一些实施方式中,索引序列是独特的4-20bp序列。在一些实施方式中,索引序列是独特的6bp、8bp或10bp序列。在一些实施方式中,索引序列是独特的6bp序列。索引的长度将取决于文库的复杂性。合适的索引序列长度对于本领域技术人员将是显而易见的。在一些实施方式中,包含第一引物序列、第一索引序列和第一衔接子序列的序列附接至多核苷酸的5'端,并且包含第二引物序列、第二索引序列和第二衔接子的序列附接到同一多核苷酸的3'端。所述多核苷酸是例如本公开的归一化对照组合物的多核苷酸。在一些实施方式中,每组多核苷酸在该组中的多个多核苷酸的5'和3'末端上包含唯一的双索引。在该实施方式中,双索引对于该组内的所有多个多核苷酸是相同的,并且对于组之间的多核苷酸是不同的。在一些实施方式中,多核苷酸的每个亚组在该亚组中的多个多核苷酸的5'和3'末端上包含唯一的双索引。在该实施方式中,双索引对于亚组内的所有多个多核苷酸是相同的,但是对于不同亚组中的多核苷酸是不同的。在一些实施方式中,每个组内和每个亚组内的多个多核苷酸具有对该组和/或亚组特异的独特的双重索引。

[0282] 本公开的非限制性示例性双索引系统是Illumina i5和i7双索引系统。

[0283] 在一些实施方式中,归一化对照被设计为测量DNA或RNA的转化。将测序样本中的靶序列靶向可测序的文库分子。在一些实施方式中,在文库制备期间起始材料到终止材料的转化率很差,并且难以在所有样本中可靠地测量。通过双索引NC多核苷酸的末端,可以跟踪在文库制备过程中的任何时间转换的分子以及每个样本的转换方式。这允许计算样本类

型不可知库转换因子。这也允许将质量控制内置到样本制备中。使用此技术,可以比较试剂批次,酶促反应的效率以及影响NGS的许多其他指标。这些比较可以按照测序反应的步骤进行,通常需要等到文库制备过程结束才能确定是否出现问题。

[0284] 蛋白质

[0285] 归一化对照多核苷酸包含脱氧核糖核酸(DNA)分子,核糖核酸(RNA)分子或DNA-RNA杂合分子。

[0286] 与蛋白质的结合是从NGS读段计数计算靶序列浓度时变异和错误的重要来源。例如,在一些实施方式中,靶序列是病毒序列,并且包含病毒序列的多核苷酸与衣壳的缔合防止了在样本和文库制备期间靶序列的纯化和/或片段化。因此,将蛋白质添加到本公开的归一化对照组合物中以模拟测序样本中靶序列的核酸/蛋白质相互作用,从而增加了NC在NGS期间模拟靶序列行为的能力。

[0287] 在本公开的归一化对照的一些实施方式中,归一化对照还包含至少一种蛋白质。在一些实施方式中,至少一种蛋白质是分离或衍生自细菌,病毒或真核寄生虫。在一些实施方式中,至少一种蛋白是病毒衣壳蛋白或细菌细胞壁蛋白。示例性但非限制性的病毒衣壳蛋白包括分离或衍生自腺病毒,腺伴随病毒(AAV),慢病毒或逆转录病毒的病毒衣壳蛋白。

[0288] 在一些实施方式中,多核苷酸群和蛋白质可操作地连接。在示例性的实施方式中,蛋白质是病毒衣壳蛋白,并且多核苷酸被包封在病毒衣壳蛋白内。在其他实施方式中,蛋白质和多核苷酸被结合。

[0289] 人工序列

[0290] 在一些实施方式中,归一化对照多核苷酸包含启动子序列。在一些实施方式中,启动子序列包含选自T7启动子序列,SP6启动子序列或T3启动子序列的启动子序列。在一些实施方式中,特别是其中启动子是T7启动子的那些实施方式,T7启动子包含序列5'-TAATACGACTCACTATAGG-3'(SEQ ID NO:14)。在一些实施方式中,T7启动子包含序列5'-TAATACGACTCACTATAGGG-3'(SEQ ID NO:15)。在一些实施方式中,T7启动子包含序列5'-GCCTCGAGCTAATACGACTCACTATAGAG-3'(SEQ ID NO:16)。在一些实施方式中,SP6启动子包含序列5'-ATTTAGGTGACACTATAG-3'((SEQ ID NO:17)。在一些实施方式中,SP6启动子包含序列5'-CATACGATTTAGGTGACACTATAG-3'(SEQ ID NO:18)。在一些实施方式中,T3启动子包含序列5'AATTAACCCTCACTAAAG 3'的(SEQ ID NO:19)。

[0291] 在一些实施方式中,归一化对照多核苷酸包含另外的引物序列。引物序列可以位于组分和/或分离的序列的5'和3'。引物序列可以作为组分序列的计算机设计的一部分被添加,或者其次被添加到组分和/或分离的序列中。引物序列可以通过例如衔接子连接或组分和/或分离的序列的随机引物和延伸而添加到组分和/或分离的序列。

[0292] 混合组合物

[0293] 在一些实施方式中,本公开内容的归一化对照包括合成多核苷酸和已分离或衍生自一种或多种生物的多核苷酸的混合物。例如,样本中的一组靶序列可以覆盖一定范围的长度,该范围超出了合成多核苷酸的当前制造方法的长度(例如,10kb的长度)。在该示例中,为了用归一化对照完全模拟靶序列的长度范围,归一化对照包括组分和分离的多核苷酸的混合物。

[0294] 在一些实施方式中,使用本公开的方法从生物中纯化归一化对照中的至少一组多

核苷酸。在一些实施方式中,生物体是病毒、细菌、真菌或真核寄生虫。在一些实施方式中,生物体与混合宿主/非宿主测序样本中的非宿主共享一个或多个特征。在一些实施方式中,病毒是噬菌体T4。

[0295] 在一些实施方式中,还以已知浓度将噬菌体T4添加或添加到基于合成的多核苷酸的NC(基于寡核苷酸的NC)中。

[0296] 将噬菌体T4添加到归一化对照中可以使归一化对照来控制提取变异。噬菌体T4的额外好处是,与基于寡核苷酸的NC相比,它更便宜且更易于制造,尤其是对于长序列而言。

[0297] 制备归一化对照的方法

[0298] 从头合成

[0299] 本公开内容提供了进行归一化对照的方法,所述归一化对照包括至少三组多核苷酸,其中每组内的多核苷酸具有相同的长度。在一些实施方式中,至少三组多核苷酸中的每一个进一步包含多核苷酸的至少三个亚组。

[0300] 在一些实施方式中,至少三个组包含相同序列的多核苷酸。在一些实施方式中,至少三个组包含序列不同的多核苷酸

[0301] 在本公开的方法的一些实施方式中,所述方法包括合成每个独特组和/或亚组的多核苷酸作为寡核苷酸,对每个独特组和/或亚组的浓度进行定量,以及混合预定量的每个独特组和/或亚组以产生NC组合物。

[0302] 在一些实施方式中,多核苷酸是DNA多核苷酸。在一些实施方式中,多核苷酸在约10bp至约3000bp之间。在一些实施方式中,多核苷酸在约10bp至约1200bp之间。在一些实施方式中,多核苷酸在10bp与250bp之间。在一些实施方式中,多核苷酸在约150bp和约1500bp之间。在一些实施方式中,多核苷酸在约400bp和约1100bp之间。在一些实施方式中,多核苷酸在约500bp至约1500bp之间。

[0303] 在一些实施方式中,多核苷酸具有与测序样本中的靶序列相同的至少一种性质。预期靶序列的所有序列性质都在本公开的归一化对照的范围内。在一些实施方式中,靶序列包含转座元件序列,病毒序列,细菌序列,真菌序列,真核寄生虫序列或一种或多种人类基因的序列。在一些实施方式中,至少一种特性包括转座元件序列、病毒序列、细菌序列、真菌序列、真核寄生虫序列、或一个或多个人类基因序列的序列特性。

[0304] 在一些实施方式中,归一化对照的多核苷酸包含RNA。在一些实施方式中,所述DNA多核苷酸进一步包含编码启动子的序列,并且所述方法进一步包括体外转录每个DNA多核苷酸以产生多个RNA,定量所述多个RNA中的每个的浓度,以及在所述RNA中混合多个RNA。预定量以产生归一化对照。示例性的启动子包括但不限于T3,SP6和T7。

[0305] 设想在多核苷酸的组和亚组之间的浓度的每种可能的排列都在本公开的范围。

[0306] 设想在本组的范围内,在一个组内的各亚组之间以及在不同组的各亚组之间的浓度的每种可能的布置。

[0307] 浓度可以表示为每体积的重量(例如,象形图(每微升pg(pg/微升)或每μL纳克(ng)))或每体积的多核苷酸分子数(摩尔浓度)。对于本领域的普通技术人员而言,如何转换浓度的度量将是显而易见的。

[0308] 在一些实施方式中,将每个多核苷酸合成为DNA寡核苷酸,并且任选地,在体外转录以产生多个RNA多核苷酸。定量每个组或亚组的多个多核苷酸中的每一个多核苷酸,并以

本公开的浓度比混合多个多核苷酸。然后将最终的归一化对照混合物稀释至介于约0.001皮克 (pg) /微升 (μL) 与约5000pg/ μL 之间、介于约0.005pg/ μL 与约1000pg/ μL 之间、介于约0.005pg/ μL 与约700pg/ μL 之间、介于约0.005pg/ μL 与约500pg/ μL 之间、介于约0.005pg/ μL 与约100pg/ μL 之间、介于约0.005pg/ μL 与约50pg/ μL 之间、介于约0.05pg/ μL 与5000pg/ μL 之间、介于约0.05pg/ μL 与约1000pg/ μL 之间、介于约0.05pg/ μL 与约700pg/ μL 之间、介于约0.05pg/ μL 与约500pg/ μL 之间、介于约0.05pg/ μL 与约100pg/ μL 之间、介于约0.5pg/ μL 与约1000pg/ μL 之间、介于约0.5pg/ μL 与约700pg/ μL 之间、介于约0.5pg/ μL 与约500pg/ μL 之间、介于约0.5pg/ μL 与约100pg/ μL 之间、介于约0.05pg/ μL 与约50pg/ μL 之间、介于约5pg/ μL 与约1000pg/ μL 之间、介于约5pg/ μL 与约700pg/ μL 之间、介于约5pg/ μL 与约500pg/ μL 之间、介于约5pg/ μL 与约100pg/ μL 之间,或介于约5pg/ μL 与约50pg/ μL 之间的浓度。在一些实施方式中,归一化对照的浓度介于约0.005pg/ μL 与约500pg/ μL 之间。在一些实施方式中,归一化对照的浓度介于约0.5pg/ μL 与约50pg/ μL 之间。在一些实施方式中,归一化对照的浓度介于约10pg/ μL 与约20pg/ μL 之间。在一些实施方式中,归一化对照的浓度介于约1pg/ μL 与约2pg/ μL 之间。在一些实施方式中,归一化对照的浓度介于约0.1pg/ μL 与约0.2pg/ μL 之间。在一些实施方式中,归一化对照的浓度为1.4pg/ μL 。

[0309] 本公开提供了制备多个归一化对照寡核苷酸的方法,所述方法包括:(a)从至少一个参考序列产生多个参考序列片段;(b)根据参考序列片段的数量产生至少一个参数的分布;(c)将所述分布划分成箱;(d)从所述箱中的至少子集中选择至少一个参考序列片段;(e)改组参考序列以产生经改组的序列;以及(f)合成包含经改组的序列的寡核苷酸;从而产生多个归一化对照寡核苷酸。

[0310] 本公开提供了制备多个归一化对照寡核苷酸的方法,所述方法包括:(a)使用滑动窗从至少一个参考序列产生多个参考序列片段;(b)根据参考序列片段的数量产生至少一个参数的分布;(c)将所述分布划分成箱;(d)从所述箱中的至少子集中选择至少两个参考序列片段,其中所述至少两个参考序列片段在参考序列中是不连续的,或者是从不同的参考序列中选择的;(e)将来自至少3个箱中的每一个箱的至少两个参考序列片段连接起来;以及(f)合成包含连接的参考序列片段的寡核苷酸;从而产生多个归一化对照寡核苷酸。在一些实施方式中,滑动窗包括1个碱基对 (bp)、2bp、3bp、4bp或5bp的滑动窗。在一些实施方式中,所述参考序列片段是对应参考序列的约15-60个、约20-40个、约20-30个、约15-32个、20-32个或约25-35个连续核苷酸。在一些实施方式中,参考序列片段是对应参考序列的29、30、31、32、33或34个碱基对。在一些实施方式中,参考序列片段小于测序读段的平均长度的一半。

[0311] 在一些实施方式中,该参数包括以下各项中的至少一项:(1)GC含量百分比、(2)熵、(3)复杂性、(4)EIIP,或它们的组合。在一些实施方式中,该参数包括GC含量百分比。

[0312] 根据归一化对照的所需分辨率,可以将分布(例如GC含量的分布)划分为任意数量的合适的箱。例如,每个箱可以代表所述分布的1%、2%、5%、10%、15%、或20%。参考序列片段可以从每个箱或从所述箱的子集中选择。参考序列片段选自来自所述GC含量百分比分布的第10、第30、第50、第70、第80和第90百分位箱。

[0313] 在一些实施方式中,步骤(d)包括从每个箱中选择至少2个参考序列片段。例如,从每个箱中选择至少2、3、4、5、6、7、8、9或10个参考序列片段。在一些实施方式中,步骤(e)还

包括连接来自每个箱的经改组的参考序列。在一些实施方式中,仅连接来自给定箱的序列,以模拟对应箱中的参考序列片段的特性,并因此对沿分布的多个点建模。

[0314] 在一些实施方式中,所述至少一个参考序列包含至少2个、至少10个、至少20个、至少50个、至少100个、至少200个、至少300个、至少400个、至少500个、至少600个、至少700个、至少800个、至少900个、至少1,000个、至少1,200个、至少1,300个、至少1,400个、至少1,500个、至少1,600个、至少1,700个、至少1,800个、至少1,900个、至少2,000个、至少2,200个、至少2,400个、至少2,600个、至少2,800个、至少3,000个、至少4,000个、至少5,000个、至少6,000个、至少7,000个、至少8,000个、至少9,000或至少10,000个参考序列。在一些实施方式中,至少一个参考序列包含约2,000个参考序列。在一些实施方式中,至少一个参考序列包含基因组序列或cDNA序列或它们的组合。例如,对应于第一生物的参考序列可以是基因组序列,并且对应于第二生物的参考序列可以是cDNA序列。或者,多种生物(所有生物)的参考序列可以是基因组序列或cDNA序列。

[0315] 在一些实施方式中,参考序列片段是参考序列的介于约10-100个之间、介于约10-80个之间、介于约10-70个之间、介于约10-60个之间、介于约10-50个之间、介于约10-40个之间、介于约10-30个之间、介于约10-20个之间、介于约20-100个之间、介于约20-80个之间、介于约20-60个之间、介于约20-50个之间、介于约20-40个之间、介于约20-30个之间、介于约10-35个之间、介于约20-35个之间、或介于约20-25个之间的连续核苷酸。在一些实施方式中,所述参考序列片段是参考序列的介于约15-60个之间、介于约20-40个之间、介于约20-30个之间、介于约15-32个之间、介于20-32个之间或介于约25-35个之间的连续核苷酸。在一些实施方式中,参考序列片段包含参考序列的20个、21个、22个、23个、24个、25个、26个、27个、28个、29个、30个、31个、32个、33个、34个、35个、36个、37个、38个、39个、40个、41个、42个、43个、44个、45个、46个、47个、48个、49个、50个、51个、52个、53个、54个、55个、56个、57个、58个、59个、60个、61个、62个、63个、64个或65个连续碱基对。

[0316] 在一些实施方式中,每个归一化对照寡核苷酸具有以下中的至少三者:与来自参考序列分布中的对应箱的参考序列片段(1)类似的GC含量百分比、(2)类似的熵、(3)类似的EIIP、(4)类似的复杂性、或(5)类似的长度。

[0317] 多重合成

[0318] 在一些实施方式,归一化对照包括合成DNA多核苷酸的混合物。

[0319] 本公开提供了制备合成多核苷酸群体的方法,所述方法包括:(a)合成至少三个DNA分子群体,其中所述至少三个群体中的每个群体内的每个DNA分子具有相同的序列,并且其中所述DNA分子中的每个DNA分子的序列从5'至3'包含第一组分序列、第一限制性酶切位点、第二组分序列、第二限制性位点和第三组分序列;(b)对溶液中DNA分子群体中的每个群体的量进行定量;(c)用切割第一和第二限制性位点的限制性酶消化所述至少三个DNA分子群体中的每个群体,以产生至少三个经消化的DNA分子群体;以及(d)将预定量的所述至少三个经消化的DNA分子群体中的各个群体在单个组合物中混合。在一些实施方式中,所述至少三个合成DNA分子中的每一个合成DNA分子包含与所述至少三个合成DNA分子中的任何其他合成DNA分子的第一、第二和第三组分序列具有相同长度的第一、第二和第三组分序列。

[0320] 在本公开的方法的一些实施方式中,第一、第二和第三组分序列均介于约15bp与

约3000bp之间。在一些实施方式中,第一、第二和第三组分序列均介于约15bp与约1200bp之间。在一些实施方式中,第一、第二和第三组分序列均介于约100bp与约600bp之间。在一些实施方式中,第一、第二和第三组分序列均介于约150bp与约500bp之间。在一些实施方式中,第一、第二和第三组分序列分别是175bp、250bp和450bp。

[0321] 在一些实施方式中,至少三个DNA分子中的任何一个DNA分子的第一、第二和第三组分序列都不是相同的序列。在一些实施方式中,第一、第二和第三组分序列是随机序列。在一些实施方式中,使用随机序列生成器来生成随机序列。可以在www.bioinformatics.org/sms2/random_dna.html找到示例性但非限制性的随机序列生成器,并且通过引用将其并入本文。在一些实施方式中,随机序列不映射到NCBI核苷酸数据库中的任何生物体的基因组。在一些实施方式中,可以设计包括序列的随机序列。重要的序列属性(例如GC含量),并排除其他(例如,常见的限制性酶切位点)。

[0322] 在一些实施方式中,至少三个DNA分子中的每一个的第一、第二和第三组分序列具有与测序样本中的靶序列相同的至少一个特性。

[0323] 在一些实施方式中,第一和第二限制性位点的序列是相同的。在一些实施方式中,限制性位点是被II型限制性核酸内切酶识别和切割的限制性位点。合适的限制性位点的选择将取决于第一、第二和第三组分序列,并且对于本领域普通技术人员而言将是显而易见的。在一些示例性实施方式中,选择在第一、第二和第三组分序列中的任何一者中均不出现的限制性位点。

[0324] 在一些实施方式中,所述第一和第二限制位点的序列选自由以下项组成的组:EcoRI位点、BamHI位点、HindIII位点、NotI位点、PvuII位点、SmaI位点、EcoRV位点、KpnI位点、PstI位点、SalI位点、ScaI位点、SpeI位点、SphI位点、StuI位点、和XbaI位点的序列。在一些实施方式中,用于消化至少三个合成DNA分子的至少三个群体中的每一个群体的限制性酶包括EcoRI、BamHI、HindIII、NotI、PvuII、SmaI、EcoRV、KpnI、PstI、SalI、ScaI、SpeI、SphI、StuI或XbaI。在一些实施方式中,第一和第二限制性位点的序列包含XhoI位点,而限制性酶包含XhoI。

[0325] 在一些实施方式中,至少三个DNA分子中的每一个DNA分子的序列还包含第一引物序列和与第二引物序列互补的序列,所述第一引物序列位于第一随机序列的5',并且与所述第二引物序列互补的序列位于第三随机序列的3'。在一些实施方式中,第一引物序列包含序列5'-TGAAGAACTGCGGCAGG-3'(SEQ ID NO:12)(Mito阵列引物)。在一些实施方式中,第二引物序列包含序列5'-AGCTGGAAGTGCAGACC-3'(Ribo阵列引物序列),与第二引物序列互补的序列包含序列5'-GGTCTGCACTTCCAGCT-3'的(SEQ ID NO:13)。

[0326] 在一些实施方式中,本公开的方法包括对溶液中至少三个DNA分子中的每一个DNA分子的量进行定量。在一些实施方式中,本公开的方法进一步对溶液中至少三个经消化的DNA分子群体中的每一个群体的量进行定量。对溶液中每个DNA分子群体的量进行定量和/或对溶液中至少三个经消化的DNA分子群体中的每一个群体的量进行定量的方法对于本领域普通技术人员来说将是显而易见的。DNA定量方法的非限制性示例包括实时PCR、液滴数字PCR(ddPCR)和分光光度法。

[0327] 在一些实施方式中,本公开的方法包括将预定量的至少三个经消化的DNA分子群体中的每一个群体在单个组合物中混合。

[0328] 在一些实施方式中,至少三个经消化的DNA分子群体的预定量是相等的。在一些实施方式中,其中至少三个经消化的DNA分子群体的预定量不相等。在一些实施方式中,至少三个经消化的DNA分子群体的预定量是线性序列或几何序列。

[0329] 在一些实施方式中,一个数量的消化的合成DNA分子的预定量是基线,并且至少两个另外的数量的消化的DNA分子的预定量是基线的整数倍。在一些实施方式中,一个消化的DNA分子群体的预定量是基线,并且其中每个另外的消化的DNA分子群体的预定量是另一个消化的DNA分子群体的预定量,不包括基线。

[0330] 在一些实施方式中,至少三个消化的合成DNA分子群体的预定量为1:2:4的比例。

[0331] 在示例性实施方式中,将三个群体的消化的DNA分子以A:B:C的比例混合,其中“A”,“B”和“C”代表任何预定的浓度比例(例如1:2:4、1:3:5或任何线性或几何顺序)。在该实施方式中,三个DNA分子中的每个的第一组分序列是相同的长度,三个DNA分子中的每个的第二组分序列是相同的长度,并且三个合成DNA分子中的每个的第三组分序列是相同的长度。相同的长度。所得的NC组成由三组组成,每组由三个亚组组成。第一组的每个亚组由包含每个DNA分子的第一组分序列的多核苷酸组成,第二组的每个亚组由三个DNA分子中的每个包含第二组分序列的多核苷酸组成,第三组的每个亚组包括包含第三组分的多核苷酸序列每个DNA分子。当三个消化的DNA分子群体以A:B:C的比例混合时,第一组三个亚组的三种多核苷酸以A:B:C的比例存在。由于每个消化的DNA分子群体对每个组的贡献均等,因此每个组以相同的浓度存在于组合物中。

[0332] 体外转录

[0333] 在一些实施方式中,归一化对照包括RNA多核苷酸。本公开提供了进行这种RNA归一化对照的方法。

[0334] 本公开内容提供了体外转录方法,以转录为聚合酶的启动子序列的3'的序列(例如,T3,SP6或T7聚合酶)。设想具有5'启动子序列的任何序列是用于本公开内容的体外转录方法的合适模板。

[0335] 本公开内容提供了进行归一化对照的方法,包括合成至少三个DNA多核苷酸,每个多核苷酸包含启动子序列和组分序列,在体外转录每个多核苷酸以产生多个RNA,定量多个中的每个的浓度混合RNA,并以预定量混合多个RNA以产生NC组合物。

[0336] 本公开内容提供了制备合成核酸分子群体的方法,其包括:(a)合成至少三组DNA分子;其中每个DNA分子包含5'至3'的第一组分序列,启动子序列,第一限制性酶切位点,第二组分序列,第二启动子序列,第二限制性位点,第三组分序列第三启动子序列;(b)用限制性酶消化每一组DNA分子,所述限制性酶切割第一和第二限制性酶切位点,以产生至少三个被消化的合成DNA分子群体;(c)在体外转录每组DNA分子以产生至少三个RNA分子种群;(d)定量每个RNA分子群体中产生的RNA数量;(e)混合预定量的每个RNA分子群体以产生单一的归一化组合物。

[0337] 在一些实施方式中,每个DNA分子包含与任何其他DNA分子的第一,第二和第三组分序列相同长度的第一,第二和第三组分序列。

[0338] 在一些实施方式中,第一,第二和第三组分序列各自在约15bp和约1200bp之间。在一些实施方式中,第一、第二和第三组分序列均介于约150bp与约500bp之间。在一些实施方式中,第一、第二和第三组分序列均介于约100bp与约600bp之间。

[0339] 在一些实施方式中,任何DNA分子的第一,第二和第三组分序列都不是相同序列。在一些实施方式中,每个DNA分子的第一,第二和第三组分序列是随机序列。在一些实施方式中,每个DNA分子的第一,第二和第三组分序列具有与样本中的靶序列相同的至少一个特性。

[0340] 本公开提供了用限制性酶消化至少三个合成DNA分子中的每一个的方法,所述限制性酶例如II型限制核酸内切酶,其切割第一和第二限制性酶切位点以产生至少三个被消化的DNA分子群体。

[0341] 在一些实施方式中,第一和第二限制位点的序列是相同的。在一些实施方式中,限制性位点是被II型限制性核酸内切酶识别和切割的限制性位点。合适的限制性位点的选择将取决于第一、第二和第三组分序列,并且对于本领域普通技术人员而言将是显而易见的。在一些示例性实施方式中,选择在第一、第二和第三组分序列中的任何一者中均不出现的限制性位点。

[0342] 在一些实施方式中,第一和第二限制性位点的序列选自EcoRI位点,BamHI位点,HindIII位点,NotI位点,PvuII位点,SmaI位点,EcoRV位点的序列。KpnI站点,PstI站点,SalI站点,ScaI站点,SpeI站点,SphI站点,StuI站点和XbaI站点。在一些实施方式中,用于消化至少三个合成DNA分子的至少三个群体中的每一个群体的限制性酶包括EcoRI、BamHI、HindIII、NotI、PvuII、SmaI、EcoRV、KpnI、PstI、SalI、ScaI、SpeI、SphI、StuI或XbaI。在一些实施方式中,第一和第二限制性位点的序列包含XhoI位点,而限制性酶包含XhoI。

[0343] 在一些实施方式中,所得的消化的DNA分子群体包含具有用于体外转录的5'启动子和组分序列的DNA分子。例如,至少三个合成DNA分子的每个群体在消化时产生三个DNA片段,每个片段包含可操作地连接至第一,第二或第三组分序列的5'启动子。这些DNA片段然后可以用作体外转录的模板以产生本公开的RNA多核苷酸。在一些实施方式中,在体外转录之前分别纯化和分离片段。在一些实施方式中,片段库用作体外转录的模板。

[0344] 本公开提供了体外转录每个DNA分子群体以产生至少三个RNA分子群体的方法。在一些实施方式中,第一启动子序列,第二启动子序列和第三启动子序列包含相同的启动子序列。在一些实施方式中,启动子序列包含选自T7启动子序列,SP6启动子序列或T3启动子序列的启动子序列。在一些实施方式中,特别是在其中启动子是T7启动子的那些实施方式中,T7启动子包含序列5'-TAATACGACTCACTATAGG-3'(SEQ ID NO:14)。在一些实施方式中,T7启动子包含序列5'-TAATACGACTCACTATAGGG-3'(SEQ ID NO:15)。在一些实施方式中,T7启动子包含序列5'-GCCTCGAGCTAATACGACTCACTATAGAG-3'(SEQ ID NO:16)。在一些实施方式中,SP6启动子包含序列5'-ATTTAGGTGACACTATAG-3'(SEQ ID NO:17)。在一些实施方式中,SP6启动子包含序列5'-CATACGATTTAGGTGACACTATAG-3'(SEQ ID NO:18)。在一些实施方式中,T3启动子包含序列5'-AATTAACCCTCACTAAAG 3'(SEQ ID NO:19)。

[0345] 本公开的聚合酶可以是RNA聚合酶II或RNA聚合酶III聚合酶。在一些实施方式中,聚合酶是T7聚合酶、SP6聚合酶或T3聚合酶。本公开的RNA聚合酶可以是野生型聚合酶、组分聚合酶,或已被优化或工程化(例如,用于体外转录)的聚合酶。对于给定的启动子序列(例如,用于T7启动子的T7聚合酶、用于SP6启动子的SP6聚合酶或用于T3启动子的T3聚合酶),本公开的聚合酶的活性可以是高度特异性的。

[0346] T7启动子被T7噬菌体RNA聚合酶识别并支持其转录。本公开的T7聚合酶可以是已

被优化或工程改造(例如,用于体外转录)的野生型T7聚合酶、组分T7聚合酶或T7聚合酶。T7聚合酶是DNA依赖性RNA聚合酶,其可催化DNA模板在5'至3'方向上形成RNA。DNA模板可以是双链或单链的。T7聚合酶对T7启动子表现出高特异性,可以在体外产生强大的转录,并且能够将修饰的核苷酸(例如,标记的核苷酸)掺入新生的RNA转录物中。T7聚合酶的这些特征使其成为用于合成本公开内容的RNA的优异的聚合酶。

[0347] 本公开提供了定量在至少三个合成RNA分子群体的每一个中产生的RNA的量的方法。定量RNA的方法对于本领域普通技术人员将是显而易见的。定量RNA的非限制性示例性方法包括分光光度法或用染料例如溴化乙锭进行荧光标记。

[0348] 本公开提供了混合预定量的每个RNA分子群体以产生单一组合物的方法。

[0349] 在一些实施方式中,每个RNA分子群体的预定量是相等的。在一些实施方式中,每个RNA分子群体的预定量不相等。在一些实施方式中,预定量的RNA分子群体为线性序列或几何序列。

[0350] 在一些实施方式中,一个RNA分子群体的预定量是基线,并且其中另外的RNA分子群体的预定量是基线的整数倍。在一些实施方式中,一个RNA分子群体的预定量是基线,并且其中每个另外的RNA分子群体的预定量是不包括基线的另一RNA分子群体的浓度的两倍。在一些实施方式中,至少三个RNA分子群体的预定量为1:2:4的比例。

[0351] 在一些实施方式中,该方法进一步包括将归一化对照稀释至介于约0.001pg/ μ L与约500pg/ μ L之间的浓度。在一些实施方式中,该方法进一步包括将归一化对照稀释至介于约0.5pg/ μ L与约500pg/ μ L之间的浓度。在一些实施方式中,该方法进一步包括将归一化对照稀释至介于约0.5pg/ μ L与约50pg/ μ L之间的浓度。在一些实施方式中,所述方法进一步包括将归一化对照稀释至介于约10pg/ μ L与约20pg/ μ L之间的浓度。在一些实施方式中,该方法进一步包括将归一化对照稀释至介于约1.0pg/ μ L与约2.0pg/ μ L之间的浓度。在一些实施方式中,该方法进一步包括将归一化对照稀释至介于约0.1pg/ μ L与约0.2pg/ μ L之间的浓度。在一些实施方式中,归一化对照的浓度为1.4pg/ μ L。

[0352] 分离的序列

[0353] 本公开提供了制备归一化对照组合物的方法,其中所述归一化对照包含分离的序列。在一些实施方式中,分离的序列是从天然存在的序列中分离的,而没有另外的步骤,例如序列改组或与另外的分离的或组分序列的连接。

[0354] 因此,本公开提供了进行归一化对照的方法,所述方法包括:(a)从样本中提取DNA;(b)用限制性酶消化DNA以产生DNA片段集合;(c)分离DNA片段集合;(d)从DNA片段集合中纯化至少三个DNA片段,以产生至少三组多核苷酸,其中所述至少三组多核苷酸中的每个组均包含多个多核苷酸;以及(e)将预定量的至少三组多核苷酸中的每个组混合以产生组合物。

[0355] 本公开提供了从样本提取DNA的方法。在一些实施方式中,样本分离自或衍生自病毒,细菌,真菌或真核寄生虫。在一些实施方式中,样本包含克隆载体。在一些实施方式中,克隆载体是细菌人工染色体(BAC),酵母人工染色体(YAC),粘粒,粘粒或质粒。在一些实施方式中,样本是质粒。在一些实施方式中,将归一化对照序列从感兴趣的物种例如病毒,细菌或真核寄生虫克隆到克隆载体中,并从产生克隆载体的细胞培养物中提取DNA。在一些实施方式中,细胞是细菌(例如大肠杆菌(*Escherichia coli*))。DNA提取的示例性方法包括试

剂盒(例如Qiagen MiniPrep)或基于酚的提取方法。合适的DNA提取方法对于本领域普通技术人员将是显而易见的。

[0356] 本公开提供了用限制性酶消化DNA以产生DNA片段集合的方法。在一些实施方式中,限制性酶选自以下项组成的组:EcoRI、BamHI、HindIII、NotI、PvuII、SmaI、EcoRV、KpnI、PstI、SalI、ScaI、SpeI、SphI、StuI和XbaI。在一些实施方式中,限制性酶选自以下项组成的组:EcoRI、BamHI、HindIII、PvuII、SmaI、EcoRV、KpnI、PstI、SalI、ScaI、SpeI、SphI、StuI、XbaI、NotI、AscI、FseI、PacI、PmeI、BglII、BstBI、HincII和SgfI。

[0357] 在一些实施方式中,通过用限制性酶消化DNA而产生的片段介于约15bp与约100kb之间。在一些实施方式中,片段介于约15bp与约50kb之间。在一些实施方式中,片段介于约50bp与约50kb之间。在一些实施方式中,片段介于约500bp与约50kb之间。在一些实施方式中,片段介于约1000bp与约50kb之间。在一些实施方式中,至少三个不同组的多核苷酸中的每一个组的最长是至少15kb的长度、至少20kb的长度、至少25kb的长度、30kb的长度、至少35kb的长度、或至少40kb的长度。

[0358] 本公开提供了从DNA片段集合中纯化至少三个DNA片段以产生至少三组多核苷酸的方法,其中至少三组多核苷酸中的每一个都包含多个多核苷酸。在一些实施方式中,分离DNA片段集合包括基于片段大小,片段电荷或其组合的分离。在一些实施方式中,分离DNA片段集合包括基于片段大小的分离。在一些实施方式中,基于大小分离DNA片段包括凝胶电泳,色谱法或切向流过滤(TFF)。

[0359] 在一些实施方式中,基于大小分离DNA片段包括凝胶电泳。凝胶电泳根据其大小分离多核苷酸。将多核苷酸样本装入凝胶一端的孔中,并施加电流以使其通过凝胶。多核苷酸带负电,因此向正极移动。多核苷酸的大小决定了其通过凝胶孔迁移的速率。可以通过改变凝胶的琼脂糖浓度(例如,介于在0.8%与3%之间的琼脂糖)来分离大小不同的多核苷酸。或者,可以使用丙烯酰胺凝胶分离较小的多核苷酸。在一些实施方式中,特别是其中分离了大的多核苷酸的那些实施方式,改变电磁场的角度,强度和持续时间可以帮助分离。在一些实施方式中,凝胶电泳是脉冲场凝胶电泳。

[0360] 在一些实施方式中,基于大小分离DNA片段包括色谱法。一种示例性的色谱方法包括阴离子交换色谱。阴离子交换色谱法使用含有带正电基团的离子交换树脂根据多核苷酸的电荷分离多核苷酸,该基团与带负电的分子结合。多核苷酸与柱或基质结合的紧密度基于多核苷酸的负电荷的整体强度,因此取决于大小。

[0361] 在一些实施方式中,基于大小分离DNA片段包括TFF。TFF是基于膜的过滤过程。基于膜的孔隙率,可以将其分类为微滤或超滤过程。超滤膜的孔径通常在0.001至0.1 μm 之间。切向流过滤(TFF),也称为错流过滤,使进料流平行于膜表面。一部分物流通过膜(渗透物),而其余部分(截留物)再循环回到进料贮存器。多核苷酸通过穿过膜的孔而分离。

[0362] 在一些实施方式中,所述方法进一步包括将衔接子与至少三个不同组的多核苷酸的每个多核苷酸的至少一个末端连接。

[0363] 在一些实施方式中,所述方法进一步包括将衔接子连接至所述至少三个不同组的多核苷酸的每个末端。

[0364] 在一些实施方式中,归一化对照中的至少三个不同组的多核苷酸的序列进一步包含合成的多核苷酸序列。示例性但非限制性的合成序列包括衔接子序列,用于PCR扩增的引

物,用于体外转录的启动子序列以及在NGS过程中用于流通池附着的结合位点。

[0365] 在一些实施方式中,衔接子序列包含编码独特分子标识符的序列。在一些实施方式中,编码唯一分子标识符的序列包括第一索引序列或第二索引序列。本公开的示例性非限制性索引策略使用两个索引,第一索引和第二索引。在一些实施方式中,编码独特分子标识符的序列包括共同的引物序列,索引序列和衔接子序列。在一些实施方式中,共同引物序列包含NGS的流通池附着位点。

[0366] 在一些实施方式中,索引序列是独特的4-20bp序列。在一些实施方式中,索引序列是独特的6bp、8bp或10bp序列。在一些实施方式中,索引序列是独特的6bp序列。索引的长度将取决于文库的复杂性。合适的索引序列长度对于本领域技术人员将是显而易见的。在一些实施方式中,包含第一引物序列、第一索引序列和第一衔接子序列的序列附接至多核苷酸的5'端,并且包含第二引物序列、第二索引序列和第二衔接子的序列附接到同一多核苷酸的3'端。所述多核苷酸是例如本公开的归一化对照组合物的多核苷酸。在一些实施方式中,每组多核苷酸在该组中的多个多核苷酸的5'和3'末端上包含唯一的双索引。在该实施方式中,双索引对于该组内的所有多个多核苷酸是相同的,并且对于组之间的多核苷酸是不同的。

[0367] 在一些实施方式中,合成序列包含启动子序列。在一些实施方式中,启动子序列包含选自T7启动子序列,SP6启动子序列或T3启动子序列的启动子序列。

[0368] 分离与从头合成相结合

[0369] 本公开提供了使用本公开的方法制备归一化对照组合物的方法,其中归一化对照既包含包含合成的序列的多核苷酸,也包含包含从样本中分离的序列的多核苷酸。

[0370] 本公开内容提供了进行归一化对照的方法,包括:(a)合成至少三组DNA分子,其中一组内的DNA分子具有相同的序列,并且其中每个DNA分子的序列包括5个以上的'至3',第一成分序列,第一限制性酶切位点,第二成分序列,第二限制位点和第三成分序列;(b)定量溶液中每组DNA分子的数量;(c)用限制性酶消化每一组DNA分子,所述限制性酶切割第一和第二限制位点以产生至少三组消化的DNA分子;(d)将预定量的每组经消化的DNA分子混合,从而产生单一的归一化对照组合物。

[0371] 在一些实施方式中,(d)的混合步骤还包括将预定量的至少另一组DNA分子与来自步骤(a)-(c)的至少三组消化的DNA分子混合以产生至少单个归一化对照中包含四组DNA分子,其中至少另一组DNA分子是从样本中分离或衍生的。在某些实施方式中,使用本公开的方法从样本中分离或衍生出至少另一组DNA分子。

[0372] 在一些实施方式中,至少四组DNA分子的预定量是相等的。

[0373] 在一些实施方式中,至少四组DNA分子的预定量不相等。在某些实施方式中,至少四组DNA分子的预定量为线性序列或几何序列。

[0374] 本公开内容提供了进行归一化对照的方法,包括:(a)合成至少一组DNA分子,其中该组中的DNA分子具有相同的序列,并且其中每个DNA分子的序列包括5个以上的'至3',第一成分序列,第一限制性酶切位点,第二成分序列,第二限制位点和第三成分序列;(b)定量溶液中DNA分子的数量;(c)用限制性酶消化至少一组DNA分子,该酶切割第一和第二限制位点以产生至少一组消化的DNA分子;(d)将预定量的消化的DNA分亚组与预定量的至少两个另外的DNA分亚组混合,以在单个归一化对照组合物中产生至少三个DNA分亚组,其中从样

本中分离或衍生出至少另外两组DNA分子。

[0375] 在一些实施方式中,至少三组DNA分子的预定量是相等的。

[0376] 在一些实施方式中,至少三组DNA分子的预定量不相等。在一些实施方式中,至少三组DNA分子的预定量为线性序列或几何序列。

[0377] 在一些实施方式中,一组DNA分子的预定量是基线,并且预定量的DNA分子的其他组是基线的整数倍。在一些实施方式中,一组消化的DNA分子的预定量是基线,并且其中每另外一组消化的DNA分子的预定量是另一组消化的DNA分子的预定量,不包括基线。在一些实施方式中,至少四组消化的DNA分子的预定量的比例为1:2:4,任选地1:2:4:8或1:2:4:8:16等。

[0378] 在一些实施方式中,样本包含质粒DNA,线粒体DNA,叶绿体DNA或基因组DNA。在某些实施方式中,样本包含从克隆载体分离的DNA。在一些实施方式中,克隆载体是细菌组分染色体(bacterial component chromosome,BAC)、酵母人工染色体(yeast artificial chromosome,YAC)、粘粒、福斯质粒或质粒。

[0379] 样本制备

[0380] 本公开提供了定量样本中至少一种靶核酸分子的表达水平的方法,其包括:(a)将已知量的本公开的归一化对照组合物与样本混合,(b)制备高通量测序文库,(c)对所述文库测序以产生读段集合,(d)将每个读段映射至样本或组合物,(e)确定由组合物中的多个核酸分子中的每个核酸分子产生的读段数目,(f)计算(a)中与样本混合的组合物中多个核酸分子各自的起始浓度与(c)中产生的读段数之间的关系,(g)建模(确定)该关系在读段和样本浓度之间,以及h.使用(g)中的模型,由靶核酸分子产生的读段的数目计算样本中至少一种靶核酸分子的初始浓度。

[0381] 在一些实施方式中,样本包含核酸。核酸可以是DNA,RNA或DNA和RNA的混合物。核酸可以是单链的(例如,单链RNA,单链的DNA),双链的(例如,双链的DNA,DNA:RNA杂交体)或其混合物。

[0382] 在一些实施方式中,样本包含全血,血浆,血清,眼泪,唾液,粘液,脑脊液,牙齿,骨骼,指甲,粪便,尿液,组织或取自哺乳动物的活检物。在一些实施方式中,样本包括全血,血浆,血清,眼泪,唾液,粘液,脑脊液,牙齿,骨骼,指甲,粪便,尿液,组织或取自人的活检物。在一些实施方式中,样本包含来自植物的叶,花或其部分,茎,根,坚果或种子。在一些实施方式中,样本包含来自昆虫的身体部位,组织,粪便或血淋巴。

[0383] 在一些实施方式中,该方法进一步包括从样本中提取核酸。在一些实施方式中,核酸包含DNA,RNA或其混合物。从样本中提取核酸的方法是本领域普通技术人员众所周知的。示例性的提取方法包括但不限于基于苯酚/氯仿的提取方法,以及诸如Qiagen Miniprep试剂盒(Cat.No.27104)和Qiagen RNeasy mini kit(Cat.74104)。

[0384] 在一些实施方式中,样本是哺乳动物样本。在一些实施方式中,靶核酸分子包含病毒,细菌,真菌或真核寄生虫的序列(例如,在哺乳动物样本中发现的病毒,细菌,真菌或真核寄生虫)。在一些实施方式中,靶核酸分子包含一种或多种哺乳动物基因。

[0385] 在一些实施方式中,样本是人样本。在一些实施方式中,靶核酸分子包含病毒,细菌,真菌或真核寄生虫的序列(例如在人样本中发现的病毒,细菌,真菌或真核寄生虫)。在一些实施方式中,靶核酸分子包含一个或多个人类基因。

[0386] 本公开内容提供了将已知量的本公开内容的归一化对照与样本混合的方法。在一些实施方式中,将归一化对照组合物与样本混合,使得归一化对照占从样本获得的总序列读段的一小部分。在一些实施方式中,归一化对照总体上应该代表使用NGS从样本获得的序列读段的一小部分(<5%)。

[0387] 在一些实施方式中,相对于归一化对照中的多核苷酸的总量,归一化对照组合物中的多核苷酸的每个组或亚组的比率介于约0.001至约0.500:1之间。在一些实施方式中,相对于归一化对照中的多核苷酸的总量,归一化对照组合物中的多核苷酸的每个组或亚组的比率为约0.001至约0.420:1。在一些实施方式中,归一化对照组合物中的多核苷酸的每个组或亚组的比率介于约0.001至约0.450:1之间、介于约0.001至约0.400:1之间、介于约0.001至约0.400:1之间、介于约0.001至约0.350:1之间、介于约0.001至约0.300:1之间、介于约0.001至约0.250:1之间、介于约0.001至约0.200:1之间、介于约0.001至约0.150:1之间、介于约0.001至约0.100:1之间、介于约0.001至约0.05:1之间、介于约0.005至约0.500:1之间、介于约0.010至约0.500:1之间、介于约0.050至约0.500:1之间、介于约0.100至约0.500:1之间、介于约0.010至约0.400:1之间、介于约0.010至约0.300:1之间、介于约0.010至约0.200:1之间、介于约0.010至约0.100:1之间、介于约0.050至约0.400:1之间、介于约0.050至约0.300:1之间、介于约0.050至约0.200:1之间、或介于约0.050至约0.100:1之间。

[0388] 在一些实施方式中,以介于约0.0001pg/ μ L与约45pg/ μ L之间、介于约0.001pg/ μ L与约45pg/ μ L之间、介于约0.001pg/ μ L与约10pg/ μ L之间、介于约0.005pg/ μ L与约5pg/ μ L或介于约0.009pg/ μ L与约5pg/ μ L之间的终浓度将组合物中的多核苷酸的每个组或亚组加入样本中。在一些实施方式中,以介于约0.009pg/ μ L与约5pg/ μ L之间的终浓度将组合物中的多核苷酸的每个组或亚组加入样本中。

[0389] 在一些实施方式中,以介于约0.001阿摩尔/ μ L与约50阿摩尔/ μ L之间的终浓度将组合物中的多核苷酸的每个组或亚组加入样本中。在一些实施方式中,以介于约0.01阿摩尔/ μ L与50阿摩尔/ μ L之间的终浓度将组合物中的多核苷酸的每个组或亚组加入样本中。在一些实施方式中,以介于约0.05阿摩尔/ μ L与25阿摩尔/ μ L之间的终浓度将组合物中的多核苷酸的每个组或亚组加入样本中。

[0390] 在一些实施方式中,该方法还包括从样本中提取核酸,并且在从样本中提取核酸之前进行归一化对照与样本的混合。在一些实施方式中,在核酸提取之前将归一化对照与样本混合允许归一化对照在样本制备期间测量核酸提取中的偏差。例如,与蛋白质复合的靶核酸可能对核酸提取有抵抗力,因此在NGS之后的阅读丰度中代表性不足。该过程可以通过添加适当的归一化对照来模仿,例如,被病毒衣壳蛋白封装的归一化对照。

[0391] 在一些实施方式中,将归一化对照与样本混合在制备高通量测序文库之前进行。在一些实施方式中,特别是其中混合步骤在文库制备之前进行的那些实施方式中,NC测量测序样本中DNA或RNA靶序列向可测序文库分子的转化。与样本多核苷酸同时对NC多核苷酸的末端进行双索引可在文库制备过程中的任何时间跟踪每个样本转换的分子以及如何转换分子。这允许计算样本类型不可知库转换因子。这也允许将质量控制内置到样本制备中。使用此技术,可以比较试剂批次,酶促反应的效率以及许多其他有关测序的指标。这些比较可以按照测序反应的步骤进行,通常需要等到文库制备过程结束才能确定是否出现问题。

[0392] 在一些实施方式中,将归一化对照与样本混合在与文库制备步骤同时发生。在库

准备过程的任何时候都可以添加带有适当衔接子的归一化对照。

[0393] 在一些实施方式中,在文库制备步骤之后发生归一化对照与样本的混合。具有适当的索引和/或测序衔接子的归一化对照可以在文库制备后但在NGS之前添加到文库中。在一些实施方式中,在该过程控制中在这一点上添加的归一化对照用于序列反应本身中的偏差。例如,具有高或低GC含量的归一化对照可以控制焦磷酸测序方法中已知的GC偏差。

[0394] 文库

[0395] 本公开的测序方法涉及测序文库的制备。测序文库的制备又涉及到一系列经衔接子修饰的DNA片段集合的生产,所述集合已经准备好进行测序。多核苷酸的测序文库可以由DNA或RNA制备,包括DNA或cDNA的等同物,类似物,其是互补的或复制RNA模板产生的DNA,例如通过逆转录酶的作用。所述多核苷酸可以以双链DNA(dsDNA)形式(例如基因组DNA片段、PCR和扩增产物)起源,或者多核苷酸可以以单链形式起源,如DNA或RNA,并已转换为dsDNA形式。举例来说,可以将mRNA分子复制到适合用于制备测序文库的双链cDNA中。

[0396] 用于某些NGS测序平台的测序文库的制备要求多核苷酸具有特定范围的片段大小,例如0-1200bp。因此,可能需要将大的多核苷酸片段化。无论多核苷酸是被强制片段化还是以片段形式自然存在,它们都将被转化为具有5'-磷酸和3'-羟基的平末端DNA。文库制备过程中的片段化过程是使用NGS定量靶序列时潜在的错误来源,可以通过在文库制备中的片段化步骤之前加入归一化对照来进行建模。

[0397] 通常,使用本领域已知的方法或试剂盒对片段末端进行末端修复,例如平末端修饰。可通过酶处理,例如使用多核苷酸激酶,将平末端片段磷酸化。在一些实施方式中,例如通过某些类型的DNA聚合酶(例如Taq聚合酶或Klenow exo⁻聚合酶)的活性将单个脱氧核苷酸(例如脱氧腺苷(A))添加至多核苷酸的3'末端。具有dA尾的产物与在随后的步骤中连接至其的衔接子的每个双链体区域的3'末端上存在的'T'突出端兼容。dA尾部阻止了两个平末端多核苷酸的自连接,使得存在朝向衔接子连接的序列的形成的偏倚。将具有dA尾的多核苷酸连接至双链衔接子多核苷酸序列。相同的衔接子可以用于多核苷酸的两个末端,或者可以使用两组衔接子。连接方法是本领域中已知的,并利用连接酶(例如DNA连接酶)将衔接子共价连接至具有dA尾的多核苷酸。衔接子可包含5'-磷酸部分以促进与靶3'-OH的连接。具有dA尾的多核苷酸包含5'-磷酸部分,该部分是剪切过程中残留的,或者是使用酶处理步骤添加的,并已被末端修复,并任选地通过一个或多个突出的碱基进行延伸,以得到适用于连接的3'-OH。纯化连接反应的产物以去除未连接的衔接子、可能已经相互连接的衔接子,并选择用于簇产生的模板的大小范围,所述选择可以在扩增(例如PCR扩增)之前进行。连接产物的纯化可以通过包括凝胶电泳和固相可逆固定化(SPRI)的方法获得。

[0398] 标准协议,例如使用例如Illumina平台进行测序的协议,可以指示用户在dA尾部之前纯化最终修复的产品,并在文库的衔接子连接步骤之前纯化dA尾部产品。准备。末端修复产物和具有dA尾的产物的纯化去除了酶、缓冲液、盐等,以为随后的酶促步骤提供有利的反应条件。在一些实施方式中,末端修复、dA加尾和衔接子连接的步骤不包括纯化步骤。因此,在一些实施方式中,本公开的方法包括制备测序文库,所述制备测序文库包括末端修复、dA加尾和衔接子连接的连续步骤。在制备不需要dA加尾步骤的测序文库的实施方式(例如,使用Roche 454和SOLIDTM 3平台进行测序的方案)中,末端修复和衔接子连接的步骤不包括在衔接子连接之间进行末端修复产物的纯化步骤。

[0399] 在一些实施方式中,制备扩增反应。扩增步骤将与流动池杂交所需的多核苷酸序列引入至衔接子连接的模板分子。扩增反应的内容物是本领域技术人员已知的,并且包括扩增反应所需的合适的底物(例如dNTP),酶(例如DNA聚合酶)和缓冲液组分。任选地,可以省略衔接子连接的多核苷酸的扩增。通常,扩增反应需要至少两个扩增引物,即引物寡核苷酸,它们可以是相同的,并且包括能够与待扩增的多核苷酸分子中的引物结合序列退火的“衔接子特异性部分”(如果互补,则可以互补)模板在退火步骤中被视为单链)。一旦形成,根据上述方法制备的模板文库可用于固相核酸扩增。如本文所用,术语“固相扩增”是指在固相支持物上或与固相支持物结合进行的任何核酸扩增反应,使得全部或部分扩增产物在形成时被固定在固相支持物上。特别地,该术语涵盖固相聚合酶链反应(固相PCR)和固相等温扩增,其类似于标准溶液相扩增的反应,除了将正向和反向扩增引物之一或两者固定在其上。坚实的支持。固相PCR涵盖了诸如乳液之类的系统,其中一种引物固定在珠粒上,另一种引物固定在游离溶液中,固相凝胶基质中的菌落形成,其中一种引物固定在表面,另一种引物固定在游离溶液中。扩增后,可通过微流体毛细管电泳分析测序文库,以确保文库中不含衔接子二聚体或单链DNA。模板多核苷酸分子的文库特别适用于固相测序方法。除了提供用于固相测序和固相PCR的模板外,文库模板还提供了用于全基因组扩增的模板。

[0400] 在一些实施方式中,使与衔接子连接的多核苷酸文库进行大规模平行测序,其包括用于对数百万个核酸片段进行测序的技术,例如,使用随机片段化的基因组DNA附着于平面,光学透明的表面并进行固相扩增的技术。创建具有数百万个簇的高密度测序流动池。Solexa/Illumina方法依赖于随机片段化的基因组DNA与平面光学透明表面的连接。延伸连接的DNA片段并进行桥扩增,以创建具有数百万个簇的超高密度测序流动池,每个簇包含同一模板的数千个拷贝。使用强大的四色DNA合成测序技术对簇模板进行测序,该技术采用可逆的终止子和可移动的荧光染料。或者,可以在珠粒上扩增文库,其中每个珠粒包含正向和反向扩增引物。

[0401] 扩增文库的测序可以使用本文所述的任何合适的测序技术进行。

[0402] 测序

[0403] 本公开提供了使用下一代测序对样本进行测序的方法,该样本包含归一化对照。

[0404] 在一些实施方式中,本文描述的方法采用下一代测序技术(NGS),其中克隆扩增的DNA模板或单个DNA分子在流动池中以大规模平行的方式进行测序(例如,如Volkerding等人,Clin Chem 55:641-658[2009];Metzker M Nature Rev 11:31-46[2010]中所述)。除了高通量序列信息外,NGS还提供数字定量信息,因为每个读段的序列都是可计数的“序列标签”,代表单个克隆DNA模板或单个DNA分子。NGS的测序技术包括焦磷酸测序,使用可逆染料终止子的合成测序,通过寡核苷酸探针连接进行测序和实时测序。

[0405] 测序技术中的一些测序技术是可商购的,例如来自Affymetrix Inc. (Sunnyvale, Calif.)的的杂交测序平台,以及来自454 Life Sciences (Bradford, Conn.)、Illumina/Solexa (Hayward, Calif.)和Helicos Biosciences (Cambridge, Mass.)的合成测序平台,以及来自Applied Biosystems (Foster City, Calif.)的连接测序平台。除了使用Helicos Biosciences的合成测序进行的单分子测序外,本公开内容的方法还包括其他单分子测序技术,包括Pacific Biosciences的SMRT™技术,Ion Torrent™技术和纳米孔测序例如,由牛津纳米孔技术公司开发。然而,本公开被设想为与任何未来开发的高通量测序技术兼容,

该技术固有地具有低输入样本的定量问题。

[0406] 虽然自动的Sanger方法被认为是“第一代”技术,但是包括本公开的自动Sanger测序在内的Sanger测序也可以采用。

[0407] 在本公开的方法的一些实施方式中,使用的DNA测序技术是454测序(Roche)(例如,如Margulies, M等人, Nature 437:376-380[2005]中所述)。在454测序中,约300-800个碱基对的DNA片段是平末端的,并将衔接子连接至所述片段的末端。衔接子用作片段的扩增和测序的引物。可以使用例如含有5'-生物素标签的衔接子B将片段附接到DNA捕获珠,例如链霉亲和素包被的珠。在油水乳状液的液滴内PCR扩增与珠粒相连的片段。结果是有每个珠粒上克隆扩增的DNA片段的多个拷贝。在第二步骤中,将珠粒捕获在孔(皮升大小)中。对每个DNA片段并行进行焦磷酸测序。一个或多个核苷酸的添加产生光信号,该光信号由测序仪器中的CCD相机记录。信号强度与掺入的核苷酸数成比例。焦磷酸测序利用焦磷酸(PPi),所述焦磷酸在添加核苷酸后被释放。在腺苷5'磷酸硫酸盐存在下,PPi通过ATP硫化酶转化为ATP。荧光素酶使用ATP将荧光素转化为氧化荧光素,该反应产生的光可以被识别和分析。454测序可产生长度介于100-1000bp之间的读段。

[0408] 在本公开的方法的一些实施方式中,使用的DNA测序技术是SOLiD™技术(Applied Biosystems)。在SOLiD™连接测序中,将5'和3'附接到DNA片段以生成片段文库。接下来,在含有微珠,引物,模板和PCR成分的微反应器中制备克隆微珠种群。PCR之后,使模板变性,并富集珠粒以用延伸的模板分离珠粒。对所选珠粒上的模板进行3'修饰,使其可以结合到载玻片上。可以通过将部分随机的寡核苷酸与由特定荧光团鉴定的中心确定的碱基(或碱基对)进行顺序杂交和连接来确定序列。记录颜色后,将连接的寡核苷酸切割并去除,然后重复该过程。SOLiD技术可以产生约50bp的读段。

[0409] 在本公开的方法的一些实施方式中,使用的DNA测序技术是Pacific Biosciences的单分子实时(SMRT™)测序技术。在SMRT测序中,在DNA合成过程中对染料标记核苷酸的连续掺入进行了成像。单个DNA聚合酶分子连接到各个零模波长标识符(ZMW标识符)的底表面,该标识符可在磷酸化核苷酸被掺入正在生长的引物链中时获得序列信息。ZMW是一种限制结构,其使得能够在荧光核苷酸的背景下观察到DNA核苷酸通过DNA聚合酶掺入单个核苷酸,所述荧光核苷酸迅速扩散出ZMW(以微秒为单位)。将核苷酸整合到一条正在生长的链中需要花费几毫秒的时间。在这段时间内,荧光标记被激发并产生荧光信号,并且荧光标记被切割掉。染料的对应该荧光的鉴定表明掺入了哪个碱基。重复该过程。RS仪器的平均读段长度约为1100bp,有时读段的最大长度为2500或2900bp。RS II可以产生8,500-60,000bp的读段长度。

[0410] 在本公开的方法的一些实施方式中,所使用的DNA测序技术是纳米孔测序(例如,如在Soni GV和Meller A. Clin Chem 53:1996-2001[2007]中所描述的)。纳米孔测序DNA分析技术正在由包括牛津纳米孔技术公司(英国牛津)在内的许多公司进行工业开发。纳米孔测序是一种单分子测序技术,通过该技术,单分子DNA在通过纳米孔时就可以直接测序。纳米孔是直径约为1纳米的小孔,纳米孔浸入导电流体中,由于离子通过纳米孔的传导,在纳米孔上施加电势(电压)会产生少量电流。流动的电流对纳米孔的大小和形状敏感。当DNA分子穿过纳米孔时,DNA分子上的每个核苷酸都会以不同程度阻塞纳米孔,从而以不同程度改变通过纳米孔的电流大小。因此,当DNA分子通过纳米孔时电流的这种变化代表了DNA序

列的读段。纳米孔MinION可以产生超过10,000bp的读段长度。

[0411] 在一些实施方式中,DNA测序技术是离子激流单分子测序,其将半导体技术与简单的测序化学方法相结合,以将化学编码的信息(A、C、G、T)直接翻译成数字信息(0、1)。半导体芯片。实际上,当核苷酸通过聚合酶掺入DNA链时,氢离子作为副产物释放出来。离子激流使用高密度的微型加工孔阵列以大规模并行方式执行此生化过程。每个孔中都保持不同的DNA分子。孔下面是离子敏感层,在其下方是离子传感器。当将核苷酸(例如C)添加到DNA模板中,然后整合到DNA链中时,氢离子将被释放。来自该离子的电荷将改变溶液的pH值,可以通过Ion Torrent的离子传感器进行识别。测序仪实际上是世界上最小的固态pH计—调用碱基,直接从化学信息转变为数字信息。然后,离子个人基因组机(PGM™)测序仪将一个核苷酸一个接一个地顺序注入芯片。如果下一个充满芯片的核苷酸不匹配。则将不记录任何电压变化,也不会调用任何碱基。如果DNA链上有两个相同的碱基,则电压将加倍,并且芯片将记录两个相同的碱基被调用。直接鉴定允许在几秒钟内记录核苷酸掺入。离子激流可产生约400bp的读段长度。

[0412] 在一些实施方式中,该方法采用Illumina的基于合成的测序和基于可逆终止子的测序化学对数百万个DNA片段进行大规模平行测序(例如,如Bentley等,Nature 6:53-59 [2009]中所述)。Illumina的测序技术依赖于将片段化的基因组DNA连接到一个平面的,光学透明的表面,该表面上结合了寡核苷酸锚。模板DNA进行末端修复,以生成5'-磷酸化的平末端,而Klenow片段的聚合酶活性用于将单个A碱基添加到平末端磷酸化DNA片段的3'末端。该添加制备了用于与寡核苷酸衔接子连接的DNA片段,该寡核苷酸衔接子在其3'端具有单个T碱基的突出端以增加连接效率。衔接子寡核苷酸与流动池锚互补。在有限稀释条件下,将衔接子修饰的单链模板DNA添加到流通池中,并通过与锚点杂交的方法固定。延伸附着的DNA片段并进行桥扩增,以创建具有数亿个簇的超高密度测序流动池,每个簇包含约1,000个相同模板的拷贝。使用强大的四色DNA合成测序技术对模板进行测序,该技术采用可逆的终止子和可移动的荧光染料。使用激光激发和全内反射光学器件可以实现高灵敏度的荧光识别。将短序列读段与重复屏蔽的参考基因组进行比对,并使用专门开发的数据分析管道软件调用遗传差异。第一读段完成后,可以在原位重新生成模板,以从片段的相对端进行第二读段。因此,根据该方法使用DNA片段的单端或双端测序。Illumina的读段长度取决于特定的平台,但通常介于50-400bp之间。

[0413] 定量

[0414] “式”、“算法”或“模型”是采用一个或多个连续或分类输入并计算输出值(有时称为“索引”或“索引值”)的任何数学方程式、算法、分析或编程过程或统计技术。“式”的非限制性示例包括总和、比率和回归算子(例如系数或指数)、生物标志物值的转换和归一化、统计分类模型,以及在历史群体上训练的神经网络。在确定样本中初始浓度与NGS之后的输出读段之间的关系时,线性和非线性方程式以及统计分类分析特别用于确定样本中初始浓度与NGS之后的序列读段之间的关系。

[0415] 特别令人感兴趣的是利用模式识别功能的结构和句法统计分类算法,包括已建立的技术,例如互相关,主成分分析(PCA),因子旋转,对数回归(LogReg),线性判别分析(LDA),支持向量机器(SVM),随机森林(RF),递归分区树(RPART)以及其他相关的决策树分类技术,收缩质心(SC),StepAIC,Kth最近邻算法,Boosting,决策树,神经网络,贝叶斯网络

和隐马尔可夫模型等。这些可以与诸如Akaike的信息标准 (AIC) 或贝叶斯信息标准 (BIC) 的信息标准结合,以便量化附加序列数据与模型改进之间的折衷,并有助于最小化过度拟合。生成的预测模型可以使用Bootstrap, Leave-One-Out (LOO) 和10折交叉验证 (10折CV) 等技术在其他研究中验证,或在最初接受过培训的研究中交叉验证。在各个步骤,可以根据本领域中已知的技术通过值排列来评估错误发现率。

[0416] 线性相关

[0417] 在本公开的方法的一些实施方式中,归一化对照的组和/或亚组的浓度包含线性序列。在一些实施方式中,归一化对照的设计的线性允许库中读书之间的相关性。NC分子的浓度当加到样本中或在制备文库时完全相关并遵循线性模型。对于被这些这种NC多核苷酸共享的输出读段,这种关系仍然成立。当NC的初始浓度相对于对应归一化的读段计数绘制时,确定回归的R2值,R2大于0.8表示在没有任何错误的情况下遵循提取和下游文库制备过程。

[0418] 将NC以一定的浓度和不同的大小加到库中。不同多核苷酸的大小范围意在表示要在文库中识别的靶DNA或RNA的插入物大小。由于预先确定在制备期间加到样本或文库中的NC的浓度,可以通过将靶的读段计数与通过NC确定的线性模型拟合来确定样本中存在的靶序列的浓度。在此模型中,在文库制备过程期间靶序列、样本中的背景核酸或NC的偏差最小。每个NC线性具有其自己的线方程, $y = mx + c$ 。Y是给出的靶序列/NC的对数(每百万的读段,RPM);M是线的斜率并且C是截距。对于靶序列,X不确定。从理论上讲,NC遵循线性,可以计算为对数(浓度),以阿摩尔s/ μL 为单位。该浓度转换为每 μL 在文库中存在多少个靶序列片段。

[0419] 这种模式的成功取决于R2大于0.8的NC的线性开始靶滴度评估。R2低于0.8表示在样本制备和/或高通量测序期间NC的质量控制失败。在该模型中,NC的线性基于绝对浓度建立。它进一步假设插入大小仅为一种,即存在100%基因组覆盖率和最小的样本变化。可以通过计算映射到背景样本序列(而非“靶”的样本序列)的读段百分比,映射到NC的读段百分比以及理论上的消耗百分比来确定样本变化。尽管需要靶序列的单个插入物大小,但这很少可以实现。理想地,在直方图中,选择靶插入物大小的3-5个最具代表性的箱,然后计算每个箱中的读段百分比。这样做本质上可以解决基因组覆盖问题。然后,插入分子的数量取决于这些箱中每个箱中的数量。

[0420] 可以使用线性关系 $Y = mx + C$ 来计算靶插入分子的数量或浓度。靶分子数量(是以下的函数)滴度;和滴度f(插入大小箱,基因组覆盖率,%人映射,%NC映射)。

[0421] 例如,为了进行质量控制,可以以彼此的比率添加归一化对照。这允许计算观察到的比率,并确认在整个过程/测定中核酸按预期进行转化。如果比率如预期的那样出来(例如,将NC1、NC2和NC3以1:2:4的浓度比率加到样本中,则读段计数/证据将恢复该比率-100个读段:200个读段:400个读段)。图1示出示例性归一化对照质量控制检查。例如,为了包含基因组DNA而不是合成的DNA多核苷酸例如噬菌体T4归一化NC的归一化对照的质量控制,寻找从该基因组(例如噬菌体T4基因组)中回收的最小读段数。

[0422] 在一些情况下,使用NC对标准曲线进行归一化。在简单的情况下,要确定其丰度的观察读段(在本文中称为“靶”读段)除以观察到的NC读段(组合或单独读段,具体取决于哪种方式可以生成更好的标准曲线)。在这种情况下,归一化靶读段丰度如下计算:

[0423] 归一化读段的丰度 = [(观察到的靶) / (观察到的NC)] * C

[0424] (如果需要, 其中C=常数)

[0425] 图2显示在下一代DNA文库测序中巨细胞病毒CMV (靶) 病毒滴度与读段丰度的图, 其中从向其添加CMV DNA (靶) 的血浆中提取DNA。虚线表示CMV滴度对CMV读段丰度的线性回归。如果不进行归一化, 则R2值为0.89。相反, 在归一化的情况下, R2值为0.98 (图3)。因此, 归一化对照改善初始样本中的靶 (此处为CMV) 的丰度与文库制备和下一代测序后映射到靶的读段的百分比之间的相关性。

[0426] 在文库制备之前将归一化对照加到样本中可以解释从影响样本中的核酸的不同实验条件和系统/随机误差出现的错误。当设计为尽可能接近实际生物学时, 可以解释和归一化制备、转化、测序等方面的许多差异。

[0427] 然而, 在一些情况下, 一些生物学因素无法模拟, 并且因此无法被归一化对照解释。在这种情况下, 原始读段到靶浓度的转换可以通过以下公式描述:

$$[0428] \quad \tilde{r} = f(r) + \Delta r$$

[0429] 其中r是原始靶读段证据 (或计数), f (r) 是转换函数, 该函数允许人们对所有所占变量的读段进行归一化, 而 Δ 函数是由于无法模拟或解释的过程而导致的残留误差。理想地, Δ (r) 变为零, 并且该转换将原始靶读段证据转换为初始初始浓度 (例如每毫升病毒拷贝)。简单函数f (r) 的示例是简单的乘法因子。在这种情况下, f (r) 由归一化对照的实际输入百分比/丰度和观察到的归一化对照的丰度 (经由读段计数/证据) 计算:

$$[0430] \quad f(r) = f \cdot r$$

[0431] 其中f可以计算为

$$[0432] \quad f = \frac{\text{实际丰度}}{\text{观察到的丰度}}$$

[0433] 其他形式的f (r) 可以解释并模拟靶性质, 例如靶GC含量, 靶大小, 片段化条件, 靶是否与蛋白质 (例如衣壳中的病毒基因组) 相关或为游离DNA分子, 等等。如果归一化对照旨在解释所有这些变量, 则 Δ 会越来越小。在图4的一个假设的示例中概述它的视觉表示。在图4中, 使用下一代测序一式三份地分析从感染病毒的宿主分离的核酸样本。在图4中, 每个下一代测序文库的最左边的条是起始样本中已知的非宿主浓度 (病毒载量) (即样本中“靶”的起始丰度)。在文库制备后, 下一代测序和读段比对后, 对读段计数进行转化, 以使它们恢复病毒载量。每个文库左侧第二个条是仅使用获得的病毒证据来计算病毒的丰度百分比的示例。图4证明, 仅从获得的病毒证据计算起始样本中病毒的丰度百分比 (病毒载量) 时, 每个文库之间存在巨大差异。该计算也不能反映原始样本中的实际病毒载量。使用f (r) 函数, 可以将原始读段证据转换为在文库之间更具可比性的量 (图4, 每个文库左起第三条)。使用f (r) 函数, 可以减少文库中计算出的病毒丰度百分比的可变性, 并且所得的值更接近于实际的病毒载量。唯一直接影响病毒载量的剩余转化是 Δ (r)。准确计算 Δ (r) 甚至可以进一步减少文库之间的可变性, 并导致病毒载量 (误差范围内)。

[0434] 仔细构造归一化对照可以为每个文库和每个靶估算f (r)。可以使用随机序列, 但也可以添加更多的生物学信息。这种方法的一个特点是, 如果可以计算f (r), 则可以避免运行多个样本, 以将原始读段直接转换为初始样本中的靶丰度。f (r) 缩小读段计数和靶丰度之间的差距, 并降低可变性。可以通过设计实验来评估 Δ (r), 该实验考虑到归一化对照中

未包含的所有变量。例如, $\Delta(r)$ 可以归因于细胞壁, 例如细菌细胞壁, 或病毒的衣壳环境, 其作用可以通过实验确定。通过用 $f(r)$ 缩小初始样本中靶丰度与输出读段计数之间的差距, 可以将 Δ 最小化。可以从多个样本中计算可能是乘数因子 ($y=ax+b$) 的 Δ 。在一些实施方式中, 用于 Δ 的高阶模型也可能是合适的。图5描绘使用 $f(r)$ 而不是原始读段来计算病毒载量时预期的可变性降低。在运行许多样本以计算 Δr 以获得图5中所示的线的方程式之前, 将可变性通过 $f(r)$ 尽可能减小。

[0435] $f(r)$ 和 Δr 的计算后可变性的一个潜在来源是测定的执行位点。在一些情况下, 可能是在次要位点执行测定时 (例如, 另一家医院计算患者样本中的病毒滴度), 测定的执行递送并不完全落入针对该组靶和归一化对照最初生成的线的数据。在这种情况下, 可以在次要位点验证该测定。例如, 次要位点可以运行足够的样本来计算病毒载量转换, 并将其方程式输入到Web应用程序中, 该Web应用程序在根据实际样本计算病毒载量时使用。

[0436] 影响样本到测序读段转化的一些可能变量的表在下文表示为表1。表1概述NC控制设计中要考虑的可能变量。表1概述当靶是病毒时在设计NC时考虑的可能变量。然而, 其中一些可能的变量可能适用于其他靶。

[0437] 表1.NC设计中的变量

变量	类别/噪音	NC 设计策略	有助于	测试说明
GC 含量	生物学	计算所有 TPx 病毒中的 GC 含量, 并使 NC 模拟该范围	$f(r)$	用这些 NC 运行库, 计算读段证据, 查看是否可以将 NC 计数的平均值用于 $f(r)$, 或具有与病毒接近的 GC 含量的 NC 组分产生更好的 $f(r)$ 乘法因子 (或观察 $f(r)$ 是否更复杂, 但将原始读段的丰度以较小的可变性移近病毒负载。
位置偏差	随机	无	$\Delta(r)$	在各个位置发送足够的样本以评估位置可变性的影响。
[0438] 病毒片段化	生物学	具有大到足以片段化的 NC	$f(r)$	用这些 NC 运行库, 计算读段证据, 查看是否可以将 NC 计数的平均值用于 $f(r)$, 或具有与病毒接近的片段化模式的 NC 组分产生更好的 $f(r)$ 乘法因子 (或观察 $f(r)$ 是否更复杂, 但将原始读段的丰度以较小的可变性移近病毒负载。
从衣壳环境中提取 (提取偏差)	生物学	无/将 NC 加到相当于病毒的衣壳中	$\Delta(r)$	运行足够的样本以确定是否观察到线性范围 (R^2 高), 并获得线的方程式。
NC 与病毒的丰度	生物学	在对病毒的临床相关水平下 NC 的掺入	$f(r)$	用这些 NC 运行库, 计算读段证据, 查看是否可以将 NC 计数的平均值用于 $f(r)$, 或具有与病毒接近的丰度的 NC 组分产生更好的 $f(r)$ 乘法因子 (或观察 $f(r)$ 是否更复杂, 但将原始读段的丰度以较小的可变性移近病毒负载。

[0439] 因此, 本公开提供以下方法: 计算与样本混合的归一化对照中多个核酸分子中的每个核酸分子的初始浓度与通过 NGS 产生的读段数目之间的关系, 对读段与样本中浓度之间的关系进行建模, 和使用该模型从通过靶核酸分子产生的读段的数目计算样本中的至少一个靶核酸分子的初始浓度。

[0440] 在一些实施方式中, 模型是线性模型。在一些实施方式中, 样本中至少一种靶核酸分子的初始浓度与通过靶核酸分子产生的读段的数目的线性回归具有在归一化后大于

0.95,大于0.96,大于0.97,大于0.98或大于0.99的R²值。在一些实施方式中,线性回归的R²值提高至少0.01、0.03、0.05、0.07、0.09、0.1、0.13、0.15、0.17、0.19、0.2、0.23、0.25、0.27、0.29、0.3、0.33、0.35、0.37、0.39、0.4、0.43、0.45、0.47、0.49、0.5、0.53、0.55、0.57、0.59、0.6、0.63、0.65、0.67、0.69、0.7、0.8、0.9、1.0或以下归一化之间的任何值。

[0441] 使用机器学习确定滴度

[0442] 支持向量机(SVM)是一种监督学习方法的已知机制,以预测药物毒性和抗药性。它们是用于分类、回归和离群值检测的一组方法。为了使SVM能够对数据进行预测,它必须与数据匹配。给定一组训练数据,标记为属于两个类别之一的每个数据点,SVM训练算法建立一个将新数据分配到其中一个类别的模型。SVM是使用监督的机器学习方法训练分类器的非概率二进制线性分类器。

[0443] 在一些实施方式中,归一化对照数据用于拟合SVM模型。在一些实施方式中,80%的数据用于训练SVM,并且其余20%的数据用于测试SVM。对模型的拟合充当反馈过程,其中仅考虑了线性模型。具有最接近进入NGS文库的归一化对照的实际浓度的预测的模型确定了用于预测样本中的靶序列滴度的模型。

[0444] 因此,本公开提供以下方法:计算与样本混合的归一化对照中多个核酸分子中的每个核酸分子的初始浓度与通过NGS产生的读段数目之间的关系,和对读段和样本中的浓度之间的关系进行建模,和使用模型从靶核酸分子产生的读段数目计算样本中至少一个靶核酸分子的初始浓度。

[0445] 在一些实施方式中,使用机器学习分类器创建模型。在一些实施方式中,监督机器学习分类器。在一些实施方式中,机器学习分类器是向量支持机器。

[0446] 除非另有说明,否则所有百分比和比率均以分子的重量或数量(例如摩尔比)计算。适当的测量单位对于本领域的普通技术人员将是显而易见的。

[0447] 除非另有说明,否则所有百分比和比率均基于总归一化对照组合物计算。

[0448] 多分析物对照

[0449] 本文提供多分析物对照、包括多分析物对照的组合物、制备多分析物对照的方法和使用多分析物对照的方法。如本文所述,多分析物对照包括已灭活的生物物种的混合物。在不限制可包含在多分析物对照中的生物物种的情况下,病毒、细菌、真菌和真核寄生虫(单细胞和多细胞二者)均被设想为可包含在多分析物对照中的混合物或物种中的生物物种。在一些实施方式中,多分析物对照还包含合适的载体、稀释剂或赋形剂。多分析物对照以及包含多分析物对照的组合物在本文中可互换地指代。

[0450] 不希望受限于任何特定的应用,本公开的多分析物对照在高通量测序和样本分析中具有许多应用。在一些实施方式中,自主使用多分析物对照,即不使用本文所述的归一化对照。例如,在样本处理步骤(例如文库制备和测序)期间,可以将多分析物对照用作阳性对照。在一些实施方式中,多分析物对照与本文所述的归一化对照一起使用。例如,归一化对照可用于归一化来自多分析物对照中的一个或多个物种的读段。因此,在一些实施方式中,将归一化对照加到多分析物对照,其如本文所述进行处理和测序。由于预先确定多分析物对照中物质的浓度(滴度),多分析物对照可用作本文所述归一化对照的阳性对照。替代地或另外地,可以使用归一化对照来归一化多分析物对照中的一种或多种物种的读段,并且归一化的读段计数和滴度之间的关系用于生成校准曲线。该校准曲线可用于从读段计数中

确定样本中靶生物的滴度,该读段计数也已使用本文所述的归一化对照进行归一化。在一些实施方式中,将归一化对照加到多分析物对照,并且对多分析物对照进行处理,测序并且与样本平行地归一化读段以产生校准曲线。在一些实施方式中,多分析物对照和归一化对照用于产生参考校准曲线。在一些实施方式中,将归一化对照和多分析物对照都加到样本中(即掺入)。在一些实施方式中,校准曲线用于计算样本中靶生物的滴度。

[0451] 在一些实施方式中,本公开的归一化对照用于归一化来自多分析物对照的读段。在一些实施方式中,来自多分析物对照的归一化读段用于确定样本中的一种或多种靶生物的滴度。

[0452] 本公开提供包含至少三种不同物种的生物的混合物的多分析物对照。在一些实施方式中,多分析物对照包括至少十种生物物种。在一些实施方式中,多分析物对照包含至少4种生物物种,至少5种生物物种,至少6种生物物种,至少7种生物物种,至少8种生物物种,至少9种生物物种,至少10种生物物种,至少11种生物物种,至少12种生物物种,至少13种生物物种,至少14种生物物种,至少15种生物物种,至少16种生物物种,至少18种生物物种,至少19种生物物种,至少20种生物物种,至少25种生物物种,至少30种生物物种,至少35种生物物种,至少40种生物物种,至少45种生物物种,至少50种生物物种,至少75种生物物种,至少100种生物物种,至少150种生物物种或至少200种生物物种。在一些实施方式中,多分析物对照基本上由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或40种生物物种组成。在一些实施方式中,多分析物对照基本上由10种生物物种组成。在一些实施方式中,多分析物对照基本上由11种生物物种组成。在一些实施方式中,多分析物对照中的生物已被灭活。在一些实施方式中,多分析物对照还包含可接受的载体、稀释剂或赋形剂。

[0453] 在一些实施方式中,多分析物对照包含至少4种病毒,至少5种病毒,至少6种病毒,至少7种病毒,至少8种病毒,至少9种病毒,至少10种病毒,至少11种病毒,至少12种病毒,至少13种病毒,至少14种病毒,至少15种病毒,至少16种病毒,至少18种病毒,至少19种病毒,至少20种病毒,至少25种病毒,至少30种病毒,至少35种病毒,至少40种病毒,至少45种病毒,至少50种病毒,至少75种病毒,至少100种病毒,至少150种病毒或至少200种病毒。在一些实施方式中,多分析物对照基本上由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或40种病毒组成。在一些实施方式中,多分析物对照基本上由10种病毒组成。在一些实施方式中,多分析物对照基本上由11种病毒组成。在一些实施方式中,多分析物对照中的生物已被灭活。在一些实施方式中,多分析物对照还包含可接受的载体、稀释剂或赋形剂。

[0454] 在本公开的多分析物对照的一些实施方式中,多分析物对照中的物种混合物中的每种物种以相同的滴度存在。

[0455] 在一些实施方式中,每种物种以不同的滴度存在于多分析物对照中。在一些实施方式中,多分析物对照中的生物物种的滴度以预定的期望比率存在。如果预定的期望比率为1:1,则每种物种以相同的滴度存在。在替代示例中,在包含三种物种的多分析物对照中的每种生物可以以1:5:35或1:20:90的比率存在。设想在本公开的范围内的任何预定的期望滴度比。在一些实施方式中,在多分析物对照中的生物物种的滴度为线性序列、几何序列或对数序列。例如,在包含三种物种的多分析物对照中的物种可以以1:2:3或1:10:100的比

率存在。本领域普通技术人员将能够选择适合于本文公开的多分析物对照的特定用途的预定期望比率。

[0456] 在一些实施方式中,至少两种物种以相同的滴度存在于多分析物对照中。在一些实施方式中,例如其中多分析物对照包含至少三种物种的那些实施方式,至少两种物种以相同的滴度存在于多分析物对照中,而至少两种物种以不同的滴度存在。

[0457] 在包含病原体物种的本公开的多分析物对照的一些实施方式中,混合物中的病原体的每种不同物种以每mL (IU/mL) 约5至1000万个感染单位的浓度存在。在一些实施方式中,混合物中每种不同的物种以约50至500万IU/mL,约500至500万IU/mL,约500至100万IU/mL,约500至100,000IU/mL,约500至10,000IU/mL,或约1,000至10,000IU/mL的浓度存在。在一些实施方式中,每种物种是以0IU/mL,100IU/mL,1,000IU/mL,5,000IU/mL,10,000IU/mL,100,000IU/mL或1,000,000IU/mL存在的多分析物对照。

[0458] 在一些实施方式中,以100IU/mL至1,000,000IU/mL,100IU/mL至100,000IU/mL,1,000IU/mL至1,000,000IU/mL,1,000IU/mL至100,000IU/mL,100IU/mL至10,000IU/mL或1,000IU/mL至10,000IU/mL的终浓度将多分析物对照加到样本中。在一些实施方式中,以100IU/mL、150IU/mL、1,000IU/mL、5,000IU/mL、10,000IU/mL或100,000IU/mL的终浓度将多分析物对照加到样本。在一些实施方式中,可以将多分析物对照以一定浓度,例如0IU/mL,100IU/mL,1,000IU/mL,5,000IU/mL,10,000IU/mL和100,000IU/mL范围加到多个样本。作为另一个示例,可以将多分析物对照以1IU/mL,100IU/mL,150IU/mL,1,000IU/mL,5,000IU/mL,10,000IU/mL和100,000IU/mL的浓度加到样本中。在一些实施方式中,该浓度范围用于计算用于确定样本中靶生物的滴度的校准曲线。

[0459] 如本文所用,“滴度”或“物理滴度”是指溶液中生物的浓度。对于感染性生物,“滴度”或“感染滴度”是指每体积溶液的浓度感染单位(例如,感染单位或IU,每mL或 μ L)。测量滴度的方法对于本领域普通技术人员将是显而易见的。例如,可以通过测量溶液中生物的浓度来测量物理滴度。这可以使用直接方法(例如计数)或间接方法(例如通过测量特定于所讨论生物的蛋白质水平)来完成。物理滴度可以表示为每体积溶液中细胞,生物,颗粒或单位的数量(例如单位/mL)。测量感染滴度的方法是本领域普通技术人员已知的。例如,可以使用本领域常规的测定例如噬斑形成测定来确定病毒的感染滴度。在噬斑形成测定中,制备病毒原液的稀释液并铺在易感细胞单层上,并计数形成病毒噬菌斑的被感染细胞的数量。细菌滴度可以通过铺板并测量形成的菌落的数量,通过计数方法,分光光度法或通过本领域已知的任何其他方法来确定。

[0460] 在本公开的多分析物对照的一些实施方式中,多分析物对照中的生物物种包括病毒,细菌,真菌,真核寄生虫或其组合的物种。

[0461] 在一些实施方式中,在多分析物对照中的生物物种基本上由病毒物种组成。在一些实施方式中,多分析物对照中的生物物种基本上由至少3种,至少4种,至少5种,至少6种,至少7种,至少8种,至少9种,至少10种,至少11种,至少12种,至少13种,至少14种,至少15种,至少20种,至少20种,至少25种,至少30种,至少35种,至少40种,至少45种,至少50种,至少75种,至少100种,至少150或至少200种病毒物种组成。在一些实施方式中,多分析物对照中的生物物种基本上由3、4、5、6、7、8、9、10、11、12、13、14、15、16、17、18、19或20种病毒物种组成。在一些实施方式中,在多分析物对照中的生物物种基本上由11种病毒物种组成。在一

些实施方式中,多分析物对照中的生物物种包括至少3种、至少4种、至少5种、至少6种、至少7种、至少8种、至少9种、至少10种、至少11种、至少12种、至少13种、至少14种、至少15种、至少20种、至少20种、至少25种、至少30种、至少35种、至少40种、至少45种、至少50种、至少75种、至少100种、至少150种或至少200种病毒物种。在一些实施方式中,多分析物对照中的生物物种包括3、4、5、6、7、8、9、10、11、12、13、14、15、16、17、18、19或20种病毒物种。在一些实施方式中,多分析物对照中的生物物种包括11种病毒物种。

[0462] 在一些实施方式中,多分析物对照中的生物物种基本上由细菌物种组成。在一些实施方式中,多分析物对照中的生物物种基本上由至少3种,至少4种,至少5种,至少6种,至少7种,至少8种,至少9种,至少10种,至少11种,至少12种,至少13种,至少14种,至少15种,至少20种,至少20种,至少25种,至少30种,至少35种,至少40种,至少45种,至少50种,至少75种,至少100种,至少150种或至少200种细菌物种组成。在一些实施方式中,多分析物对照中的生物物种基本上由3、4、5、6、7、8、9、10、11、12、13、14、15、16、17、18、19或20种细菌物种组成。在一些实施方式中,多分析物对照中的生物物种包括至少3种,至少4种,至少5种,至少6种,至少7种,至少8种,至少9种,至少10种,至少11种,至少12种,至少13种,至少14种,至少15种,至少20种,至少20种,至少25种,至少30种,至少35种,至少40种,至少45种,至少50种,至少75种,至少100种,至少150或至少200种细菌物种。在一些实施方式中,多分析物对照中的生物物种包括3、4、5、6、7、8、9、10、11、12、13、14、15、16、17、18、19或20种细菌物种。

[0463] 在一些实施方式中,在多分析物对照中的生物物种基本上由真菌物种组成。在一些实施方式中,多分析物对照中的生物物种基本上由至少3种,至少4种,至少5种,至少6种,至少7种,至少8种,至少9种,至少10种,至少11种,至少12种,至少13种,至少14种,至少15种,至少20种,至少20种,至少25种,至少30种,至少35种,至少40种,至少45种,至少50种,至少75种,至少100种,至少150种或至少200种真菌物种组成。在一些实施方式中,多分析物对照中的生物物种基本上由3、4、5、6、7、8、9、10、11、12、13、14、15、16、17、18、19种或20种真菌物种组成。在一些实施方式中,多分析物对照中的生物物种包括至少3种,至少4种,至少5种,至少6种,至少7种,至少8种,至少9种,至少10种,至少11种,至少12种,至少13种,至少14种,至少15种,至少20种,至少20种,至少25种,至少30种,至少35种,至少40种,至少45种,至少50种,至少75种,至少100种,至少150种或至少200种真菌物种。在一些实施方式中,多分析物对照中的生物物种包括3、4、5、6、7、8、9、10、11、12、13、14、15、16、17、18、19或20种真菌物种。

[0464] 在一些实施方式中,多分析物对照中的生物物种基本上由病毒和细菌物种组成。在一些实施方式中,多分析物对照中的生物物种基本上由至少3种,至少4种,至少5种,至少6种,至少7种,至少8种,至少9种,至少10种,至少11种,至少12种,至少13种,至少14种,至少15种,至少20种,至少20种,至少25种,至少30种,至少35种,至少40种,至少45种,至少50种,至少75种,至少100种,至少150种或至少200种病毒和细菌物种组成。在一些实施方式中,多分析物对照中的生物物种基本上由3、4、5、6、7、8、9、10、11、12、13、14、15、16、17、18、19或20种病毒和细菌物种组成。在一些实施方式中,多分析物对照中的生物物种包括至少3种,至少4种,至少5种,至少6种,至少7种,至少8种,至少9种,至少10种,至少11种,至少12种,至少13种,至少14种,至少15种,至少20种,至少20种,至少25种,至少30种,至少35种,至少40种,至少45种,至少50种,至少75种,至少100种,至少150种或至少200种病毒和细菌物种。在一

些实施方式中,多分析物对照中的生物物种包括3、4、5、6、7、8、9、10、11、12、13、14、15、16、17、18、19或20种病毒和细菌物种。

[0465] 在一些实施方式中,在多分析物对照中的生物物种基本上由病毒和真菌物种组成。在一些实施方式中,多分析物对照中的生物物种基本上由至少3种,至少4种,至少5种,至少6种,至少7种,至少8种,至少9种,至少10种,至少11种,至少12种,至少13种,至少14种,至少15种,至少20种,至少20种,至少25种,至少30种,至少35种,至少40种,至少45种,至少50种,至少75种,至少100种,至少150种或至少200种病毒和真菌物种组成。在一些实施方式中,多分析物对照中的生物物种基本上由3、4、5、6、7、8、9、10、11、12、13、14、15、16、17、18、19或20种病毒和真菌物种组成。在一些实施方式中,多分析物对照中的生物物种包括至少3种,至少4种,至少5种,至少6种,至少7种,至少8种,至少9种,至少10种,至少11种,至少12种,至少13种,至少14种,至少15种,至少20种,至少20种,至少25种,至少30种,至少35种,至少40种,至少45种,至少50种,至少75种,至少100种,至少150种或至少200种病毒和真菌物种。在一些实施方式中,多分析物对照中的生物物种包括3、4、5、6、7、8、9、10、11、12、13、14、15、16、17、18、19或20种病毒和真菌物种。

[0466] 在一些实施方式中,在多分析物对照中的生物物种基本上由细菌和真菌组成。在一些实施方式中,多分析物对照中的生物物种基本上由至少3种,至少4种,至少5种,至少6种,至少7种,至少8种,至少9种,至少10种,至少11种,至少12种,至少13种,至少14种,至少15种,至少20种,至少20种,至少25种,至少30种,至少35种,至少40种,至少45种,至少50种,至少75种,至少100种,至少150种或至少200种细菌和真菌物种组成。在一些实施方式中,多分析物对照中的生物物种基本上由3、4、5、6、7、8、9、10、11、12、13、14、15、16、17、18、19或20种细菌和真菌物种组成。在一些实施方式中,多分析物对照中的生物物种包括至少3种,至少4种,至少5种,至少6种,至少7种,至少8种,至少9种,至少10种,至少11种,至少12种,至少13种,至少14种,至少15种,至少20种,至少20种,至少25种,至少30种,至少35种,至少40种,至少45种,至少50种,至少75种,至少100种,至少150种或至少200种细菌和真菌物种。在一些实施方式中,多分析物对照中的生物物种包括3、4、5、6、7、8、9、10、11、12、13、14、15、16、17、18、19或20种细菌和真菌物种。

[0467] 在一些实施方式中,在多分析物对照中的生物物种基本上由病毒、细菌和真菌物种组成。在一些实施方式中,多分析物对照中的生物物种基本上由至少3种,至少4种,至少5种,至少6种,至少7种,至少8种,至少9种,至少10种,至少11种,至少12种,至少13种,至少14种,至少15种,至少20种,至少20种,至少25种,至少30种,至少35种,至少40种,至少45种,至少50种,至少75种,至少100种,至少150种或至少200种病毒、细菌和真菌物种组成。在一些实施方式中,多分析物对照中的生物物种基本上由3、4、5、6、7、8、9、10、11、12、13、14、15、16、17、18、19或20种病毒、细菌和真菌物种组成。在一些实施方式中,多分析物对照中的生物物种包括至少3种,至少4种,至少5种,至少6种,至少7种,至少8种,至少9种,至少10种,至少11种,至少12种,至少13种,至少14种,至少15种,至少20种,至少20种,至少25种,至少30种,至少35种,至少40种,至少45种,至少50种,至少75种,至少100种,至少150种或至少200种病毒、细菌和真菌物种。在一些实施方式中,多分析物对照中的生物物种包括3、4、5、6、7、8、9、10、11、12、13、14、15、16、17、18、19或20种病毒、细菌和真菌物种。

[0468] 在一些实施方式中,多分析物对照中的生物物种包括一种或多种病毒物种。设想

所有病毒物种都有资格包括在本公开的多分析物对照中。病毒可以是DNA病毒、RNA病毒或其组合。在一些实施方式中,该病毒是DNA病毒或RNA病毒。在一些实施方式中,DNA病毒包含双链DNA病毒基因组。在一些实施方式中,DNA病毒包含单链DNA病毒基因组。在一些实施方式中,RNA病毒包含双链RNA病毒基因组。在一些实施方式中,RNA病毒包含单链RNA病毒基因组。在一些实施方式中,单链RNA病毒基因组包含有义链。在一些实施方式中,单链RNA病毒基因组包含反义链。在一些实施方式中,病毒是通过Baltimore分类系统确定的I型,II型,III型,IV型,V型,VI型或VII型病毒。

[0469] 在一些实施方式中,多分析物对照包含作为人类病原体的病毒物种。在下表2中显示作为人类病原体的示例性但非限制性病毒:

[0470] 表2. 示例性人类病毒病原体

[0471]

病毒	属和科	疾病示例
腺相关病毒	依赖病毒属, 细小病毒科	无
爱知病毒	嗜病毒属, 微小核糖核酸病毒科	肠胃炎
阿留申病病毒	细小病毒	浆细胞增多
澳大利亚蝙蝠狂犬病病毒	狂犬病病毒属, 弹状病毒科	致命性脑炎
BK 多瘤病毒	多瘤病毒, 多瘤病毒科	无
班纳病毒	Seadornavirus, 呼肠孤病毒科	脑炎
巴马森林病毒	甲病毒属, 披膜病毒科	发烧, 关节痛
布尼姆韦拉病毒	正本雅病毒属, 布尼亚病毒科	脑炎
Bunyavirus La Crosse	正本雅病毒属, 布尼亚病毒科	脑炎
Bunyavirus snowshoe hare	正本雅病毒属, 布尼亚病毒科	脑炎
弥猴疱疹病毒	淋巴病毒, 疱疹病毒科	脑炎
金迪普拉病毒	水疱病毒属, 弹状病毒科	脑炎
基孔肯雅病毒	甲病毒, 披膜病毒科	发烧, 关节痛
柯萨病毒 A	柯萨病毒, 微小核糖核酸病毒科	-
牛痘病毒	正痘病毒, 痘病毒科	无
柯萨奇病毒	肠病毒, 微小核糖核酸病毒科	脑膜炎, 心肌炎, 瘫痪
克里米亚-刚果出血热病毒	内罗病毒属, 布尼亚病毒科	出血热

[0472]

病毒	属和科	疾病示例
登革热病毒	黄病毒，黄病毒科	出血热
多里病毒	托高土病毒属，正粘病毒科	发烧，脑炎
杜比病毒	内罗病毒属，布尼亚病毒科	血小板减少症
杜温黑致病毒（Duvénhage virus）	狂犬病病毒属，弹状病毒科	致命性脑炎
东方马脑炎病毒	甲病毒，披膜病毒科	脑炎
伊波拉病毒	伊波拉病毒，丝状病毒科	出血热
艾柯病毒	肠病毒，微小核糖核酸病毒科	普通感冒
脑心肌炎病毒	心脏病毒，微小核糖核酸病毒科	脑炎
巴尔病毒	淋巴病毒，疱疹病毒科	单核细胞增多症
欧洲蝙蝠狂犬病病毒	狂犬病病毒，弹状病毒	致命性脑炎
GB 病毒 C/G 肝炎病毒	佩吉病毒，黄病毒科	无
汉坦病毒	汉坦病毒，布尼亚病毒科	肾或呼吸综合症
亨德拉病毒	肝炎病毒，副粘病毒科	脑炎
甲型肝炎病毒	肝病毒，微小核糖核酸病毒科	肝炎
乙型肝炎病毒	Orthohepadnavirus，嗜肝病毒科	肝炎
丙型肝炎病毒	肝炎病毒，黄病毒科	肝炎
戊型肝炎病毒	肝炎病毒，未分配	肝炎
丁型肝炎病毒	丁型病毒，未分配	肝炎
单纯疱疹病毒 1	疱疹病毒科	疱疹
单纯疱疹病毒 2	疱疹病毒科	疱疹
马痘病毒	正痘病毒，痘病毒科	无
人腺病毒	腺病毒，腺病毒科	呼吸道，胃肠道
人类星状病毒	乳腺病毒，星状病毒科	肠胃炎
人冠状病毒	冠状病毒，冠状病毒科	呼吸道
人巨细胞病毒	巨细胞病毒，疱疹病毒科	单核细胞增多症，肺炎
人肠病毒 68、70	肠病毒，微小核糖核酸病毒科	腹泻，神经系统疾病
人疱疹病毒 1	单纯疱疹病毒，疱疹病毒科	皮肤病变

[0473]

病毒	属和科	疾病示例
人疱疹病毒 2	单纯疱疹病毒, 疱疹病毒科	皮肤病变
人疱疹病毒 6 (A 和 B)	玫瑰疹病毒属, 疱疹病毒科	皮肤病变
人疱疹病毒 7	玫瑰疹病毒属, 疱疹病毒科	皮肤病变
人疱疹病毒 8	猴病毒属, 疱疹病毒科	皮肤淋巴瘤
人类免疫缺陷病毒	慢病毒, 逆转录病毒科	艾滋病
人乳头瘤病毒 1	丑型乳头瘤病毒属, 乳头瘤病毒科	皮肤疣
人乳头瘤病毒 2	甲型乳头瘤病毒属, 乳头瘤病毒科	皮肤疣
人乳头瘤病毒 16,18	甲型乳头瘤病毒属, 乳头瘤病毒科	尖锐湿疣
人副流感	呼吸道病毒, 副粘病毒科	呼吸道
人细小病毒 B19	红病毒属, 细小病毒科	皮肤病变
人呼吸道合胞病毒	肺病毒, 副粘病毒科	呼吸道
人鼻病毒	肠病毒, 微小核糖核酸病毒科	呼吸道
人 SARS 冠状病毒	乙型冠状病毒, 冠状病毒科	呼吸道
人泡沫反转录病毒	泡沫病毒属, 逆转录病毒科	无
人 T 淋巴病毒	逆转录病毒属, 逆转录病毒科	白血病
人朊病毒属	朊病毒属, 冠状病毒科	肠胃炎
甲型流感病毒	甲型流感病毒, 正粘病毒科	流感
乙型流感病毒	乙型流感病毒, 正粘病毒科	流感
丙型流感病毒	甲型流感病毒, 正粘病毒科	流感
伊斯法罕病毒	水疱性病毒, 弹状病毒科	脑炎
JC 多瘤病毒 (JC 病毒, 人多瘤病毒 2)	多瘤病毒, 多瘤病毒科	脑炎
日本脑炎病毒	黄病毒, 黄病毒科	脑炎
胡宁沙粒病毒	沙粒病毒, 沙粒病毒科	出血热
KI 多瘤病毒	多瘤病毒, 多瘤病毒科	脑炎

[0474]

病毒	属和科	疾病示例
昆金病毒	黄病毒，黄病毒科	脑炎
拉各斯蝙蝠病毒	狂犬病病毒	致命性脑炎
维多利亚湖马尔堡病毒	马尔堡病毒，丝状病毒科	出血热
兰加特病毒	黄病毒，黄病毒科	脑炎
拉萨病毒	沙粒病毒，沙粒病毒科	出血热
洛兹代尔病毒	诺如病毒，杯状病毒科	肠胃炎
流行病病毒	黄病毒，黄病毒科	脑炎
淋巴细胞性脉络膜脑膜炎病毒	沙粒病毒，沙粒病毒科	脑炎
马丘波病毒	沙粒病毒，沙粒病毒科	脑炎
马亚罗病毒	甲病毒，披膜病毒科	发烧，关节痛
MERS 冠状病毒	乙型冠状病毒属，冠状病毒科	呼吸道
麻疹病毒	胆汁病毒，副粘病毒科	发烧，皮疹
芒果脑心肌炎病毒	心脏病毒，微小核糖核酸病毒科	脑炎
默克尔细胞多瘤病毒	多瘤病毒，多瘤病毒科	默克尔细胞癌
莫科拉病毒	狂犬病病毒属，弹状病毒科	脑炎
传染性软疣病毒	软疣痘病毒属，痘病毒科	皮肤病变
猴痘病毒	正痘病毒，痘病毒科	皮肤病变
腮腺炎病毒	风疹病毒，副粘病毒科	腮腺炎
墨累山谷脑炎蚊传病毒	黄病毒，黄病毒科	脑炎
纽约病毒	汉坦病毒，布尼亚病毒	出血热
尼帕病毒	肝炎病毒，副粘病毒科	脑炎
诺沃克病毒	诺如病毒，杯状病毒科	肠胃炎
阿尼昂尼昂病毒	甲病毒，披膜病毒科	发烧，关节痛
口疮病毒	副痘病毒，痘病毒科	皮肤病变
奥罗波奇病毒	正本雅病毒属，布尼亚病毒科	发烧，关节痛
皮钦德病毒	沙粒病毒，沙粒病毒科	出血热
泡沫病毒属	肠病毒，微小核糖核酸病毒科	脊髓灰质炎

[0475]

病毒	属和科	疾病示例
蓬塔托静脉病毒	白蛉病毒属，布尼亚病毒科	出血热
普马拉病毒	汉坦病毒，布尼亚病毒	出血热
狂犬病病毒	狂犬病病毒属，弹状病毒科	致命性脑炎
裂谷热病毒	白蛉病毒属，布尼亚病毒科	出血热
Rosavirus A	Rosavirus，微小核糖核酸病毒科	
罗斯河病毒	甲病毒，披膜病毒科	发烧，关节痛
隆病毒属 A	隆病毒属，呼肠孤病毒科	肠胃炎
隆病毒属 B	隆病毒属，呼肠孤病毒科	肠胃炎
隆病毒属 C	隆病毒属，呼肠孤病毒科	肠胃炎
风疹病毒	风疹病毒，多病毒科	风疹
山病毒	甲病毒，披膜病毒科	发烧，关节痛
唾液病毒 A	唾液病毒，微小核糖核酸病毒科	肠胃炎
白蛉热西西里病毒	白蛉病毒属，布尼亚病毒科	出血热
札幌病毒	杯状病毒，杯状病毒科	肠胃炎
塞姆利基森林病毒	甲病毒，披膜病毒科	发烧，关节痛
汉城病毒	汉坦病毒，布尼亚病毒	出血热
猿猴泡沫病毒	泡沫病毒属，逆转录病毒科	无
猿猴病毒 5	风疹病毒，副粘病毒科	没有记载的
信德比斯病毒	甲病毒，披膜病毒科	发烧，关节痛
南安普敦病毒	诺如病毒，杯状病毒科	肠胃炎
圣路易斯脑炎病毒	黄病毒，黄病毒科	脑炎
婢传播的波瓦桑病毒	黄病毒，黄病毒科	脑炎
输血型传播病毒	细环病毒属，指环病毒科	无
托斯卡纳病毒	白蛉病毒属，布尼亚病毒科	出血热
Uukuniemi 病毒	白蛉病毒属，布尼亚病毒科	出血热
牛痘病毒	正痘病毒，痘病毒科	无
水痘带状疱疹病毒	水痘病毒，疱疹病毒科	水痘
天花病毒	正痘病毒，痘病毒科	天花

[0476]

病毒	属和科	疾病示例
委内瑞拉马脑炎病毒	甲病毒，披膜病毒科	发烧，关节痛
水疱性口炎病毒	水疱病毒属，弹状病毒科	脑炎
西部马脑炎病毒	甲病毒，披膜病毒科	发烧，关节痛
WU 多瘤病毒	多瘤病毒，多瘤病毒科	无
西尼罗河病毒	黄病毒，黄病毒科	出血热
雅巴猴肿瘤病毒	正痘病毒，痘病毒科	无
雅巴样疾病病毒	正痘病毒，痘病毒科	无
黄热病病毒	黄病毒，黄病毒科	出血热
寨卡病毒	黄病毒，黄病毒科	发烧，关节痛，皮疹

[0477] 在一些实施方式中，多分析物对照中的生物物种包括一种或多种细菌物种。设想所有细菌物种都适合包含在本公开的多分析物对照中。在一些实施方式中，细菌是革兰氏阳性细菌或革兰氏阴性细菌。

[0478] 在一些实施方式中，多分析物对照包括作为人类病原体的细菌物种。在下表3中显示作为人类病原体的示例性但非限制性细菌：

[0479] 表3. 示例性人类细菌病原体

[0480]

疾病	细菌物种
痤疮	痤疮丙酸杆菌
炭疽病	炭疽芽孢杆菌
细菌性血管瘤病/肝硬化	亨氏巴尔通体，昆塔纳巴尔通体
细菌性败血症	无乳链球菌（b 组链球菌），大肠杆菌，肺炎链球菌
细菌性阴道病	阴道加德纳菌，阴道阿托波菌属，人型支原体，各种厌氧菌，包括动弯杆菌属，普雷沃菌属
小袋虫病	结肠小袋虫
生疖	葡萄球菌属
肉毒杆菌	肉毒梭菌

[0481]

疾病	细菌物种
支气管炎	支原体肺炎, 肺炎衣原体, 化脓性链球菌
布鲁氏菌病	布鲁氏菌属, 鼠疫耶尔森氏菌
黑死病	鼠疫耶尔森菌
布鲁氏溃疡	溃疡分枝杆菌
猫抓病	亨氏巴尔通体
蜂窝组织炎	链球菌和葡萄球菌属
软性下疳	杜克伊嗜血杆菌
衣原体	沙眼衣原体
霍乱	霍乱弧菌
索氏梭菌感染	索氏梭菌
腹泻病	空肠弯曲杆菌
埃希氏菌病	查菲埃立克体, 埃文氏埃立克体, <i>E. muris caudalensis</i>
地方性复发热-	疏螺旋体 (15 种不同物种; 包括赫姆斯疏螺旋体, 帕克疏螺旋体)
黄曲霉病毒感染	米黄杆菌
毛囊炎	金黄色葡萄球菌; 葡萄球菌属
食物中毒	大肠杆菌, 金黄色葡萄球菌, 蜡状芽孢杆菌, 产气荚膜梭菌, 肉毒杆菌
鼻疽病	鼻疽伯克霍尔德氏菌
淋病	淋球菌
腹股沟肉芽肿	肉芽孢杆菌
B 组链球菌病	B 组链球菌 (GBS) 或无乳链球菌
嗜血杆菌流感	嗜血杆菌流感
汉森氏病	麻风分枝杆菌
卫生保健相关感染	鲍氏不动杆菌, 脆弱拟杆菌, 洋葱伯克霍尔德菌, 艰难梭菌, 索氏梭菌, 耐碳青霉烯肠杆菌, 粪肠球菌, 克雷伯氏肺炎杆菌, 耐甲氧西林金黄色葡萄球菌

[0482]

疾病	细菌物种
	萄球菌，摩氏摩根菌，脓肿分枝杆菌铜绿假单胞菌，金黄色葡萄球菌，嗜麦芽寡养单胞菌，结核分支杆菌，耐万古霉素金黄色葡萄球菌，耐万古霉素肠球菌，肺炎链球菌
人粒细胞埃希氏菌病	嗜吞噬细胞
人单核细胞埃希氏菌病	查菲埃立克体
脓疱疮	葡萄球菌属
肠道感染	空肠弯曲杆菌，大肠杆菌（EIEC），大肠杆菌（EHEC），鼠伤寒沙门氏菌，鼠伤寒沙门氏菌，痢疾志贺氏菌 1 型，索氏志贺氏菌/弗氏杆菌小肠结肠炎耶尔森氏菌，艰难梭菌，霍乱弧菌
克雷伯菌感染	克雷伯菌属
军团病	军团菌，嗜肺军团菌
勒米尔综合症	坏死镰刀菌
李斯特菌感染	李斯特菌
莱姆病	伯氏疏螺旋体
性腺淋巴肉芽肿	沙眼衣原体
类鼻疽	假伯克霍尔德氏菌
脑膜炎	结核分枝杆菌，无乳链球菌（b 组链球菌），大肠杆菌，单核细胞增生李斯特菌，无乳链球菌，大肠杆菌，肺炎链球菌，流感嗜血杆菌
菌丝瘤	放线菌属，白乐杰放线菌（Actinomadura pelletier），马杜拉放线菌
分枝杆菌感染	脓肿分枝杆菌，鸟分枝杆菌，胞内分枝杆菌，龟分枝杆菌，偶发分枝杆菌，戈登分枝杆菌，堪萨斯分枝杆菌，莫尔门分枝杆菌，海鱼分枝杆菌，结核分枝杆菌，异种分支杆菌

[0483]

疾病	细菌物种
坏死性筋膜炎	A 组链球菌和其他菌属；金黄色葡萄球菌，化脓性链球菌，梭状芽孢杆菌
非结核分枝杆菌性肺疾病	分枝杆菌属
骨髓炎	金黄色葡萄球菌，链球菌，肠杆菌科
多杀性巴氏杆菌感染	多杀巴斯德氏菌
心包炎	金黄色葡萄球菌，肺炎链球菌等链球菌，结核分枝杆菌
品他病	斑点密螺旋体
脓性肌炎	金黄色葡萄球菌
Q 热病	伯尼氏杆菌
风湿热	A 组链球菌
落基山斑疹热	立克次体
沙门氏菌	鼠伤寒沙门氏菌
森里特苏热	<i>Noerickettsia sennetsu</i>
志贺氏菌病	志贺氏菌属
葡萄球菌食物中毒	葡萄球菌属
嗜麦芽窄食单胞菌感染	嗜麦芽窄食单胞菌
链球菌 A 组侵袭性疾病	链球菌 A 组
链球菌 B 组浸润性疾病	链球菌 B 组
梅毒	梅毒螺旋体
破伤风	破伤风梭菌属，破伤风梭菌
结核	结核分枝杆菌
兔热病	图拉弗朗西斯菌
斑疹伤寒	普氏立克次体，恙虫病东方体，伤寒立克次体
溃疡	幽门螺杆菌
尿道炎，宫颈炎	沙眼衣原体，淋病奈瑟菌
弧菌感染	弧菌属，副溶血性弧菌，创伤性弧菌，溶藻弧菌

[0484]

疾病	细菌物种
沃特豪斯·弗里德里希森综合症	脑膜炎双球菌
鞭毛病	惠氏螺旋体
百日咳（Whooping cough）/百日咳（Pertussis）	百日咳博代氏杆菌
雅司病	雅司螺旋体

[0485] 在下表4中显示可包含在多分析物对照中的其他细菌物种,包括代表性的共生、共栖、致病性和抗药性物种,以及宏基因组学样本中发现的物种。

[0486] 表4. 示例性其他细菌物种

[0487]

细菌属和种	细菌属和种
鲍氏不动杆菌	加氏乳杆菌
约翰逊不动杆菌	詹森乳杆菌
不动杆菌属	植物乳杆菌
溶齿放线菌	单核细胞增多性李斯特氏菌
嗜盐别样芽胞杆菌	黄褐微球菌
蜡状芽孢杆菌	人型支原体
枯草芽孢杆菌	脑膜炎奈瑟菌
脆弱拟杆菌	牙龈卟啉单胞菌
普通拟杆菌	二路普氏杆菌
青春双歧杆菌	黑色素丙酸杆菌
紫色杆菌	痤疮丙酸杆菌
艰难梭菌	奇异变形杆菌
拜氏梭菌	雷氏普罗威登斯菌
产气荚膜梭菌	嗜盐假单胞菌
凝固酶阴性葡萄球菌	铜绿假单胞菌
纹状棒状杆菌	荧光假单胞菌
痤疮杆菌	假单胞菌属

[0488]	放射球菌	球形红细菌
	粪肠球菌	肠沙门氏菌
	阴沟肠杆菌	液体粘质沙雷氏菌
	肠杆菌属	粘质沙雷氏菌
	大肠杆菌	沙雷氏菌属
	核梭状芽胞杆菌	金黄色葡萄球菌
	阴道加德纳菌	金黄色葡萄球菌 (MRSA, MSSA)
	副流感嗜血杆菌	表皮葡萄球菌
	嗜盐杆菌	无乳链球菌
	火山嗜盐杆菌	链球菌炎
	幽门螺杆菌	变形链球菌
	<i>Imtechella halotolerans</i>	肺炎链球菌
	肺炎克雷伯菌	化脓性链球菌
	克雷伯菌属	链球菌属
	发酵乳杆菌	小韦荣球菌
		小肠结肠炎耶尔森菌

[0489] 在一些实施方式中,多分析物对照包含一种或多种真核寄生虫。在一些实施方式中,该寄生虫是人寄生虫。在一些实施方式中,真核寄生虫是多细胞寄生虫(例如,圆虫物种)。在一些实施方式中,真核寄生虫是单细胞生物。在一些实施方式中,真核寄生虫是酵母或真菌。在一些实施方式中,真核寄生虫是未被分类为酵母或真菌(例如变形虫)的物种。

[0490] 在一些实施方式中,多分析物对照包括一种或多种作为人类病原体的真菌。下表5显示作为人类病原体的示例性但非限制性真菌:

[0491] 表5. 示例性人类真菌病原体

真菌物种	疾病
曲霉属，烟曲霉，黄曲霉，棒曲霉等	曲霉病
毛癣菌，表皮癣菌，小孢子菌，其他菌属。	脚气
芽孢杆菌属	芽孢杆菌病
加蒂隐球菌	加蒂隐球菌感染
新型隐球菌	新型隐球菌感染
新型隐球菌	新型隐球菌感染
白色念珠菌，耳道念珠菌和其他念珠菌属	念珠菌病
念珠菌和曲霉属属	内膜炎
念珠菌属，其他菌属	真菌眼部感染
各种菌属	指甲真菌感染
组织胞浆菌属，荚膜组织胞浆菌	组织胞浆菌病
新型隐球菌，罗伦隐球菌，浅白隐球菌	脑膜炎
毛霉菌	毛霉菌病
笔直顶孢霉，构巢曲霉，罗萨氏裸藻， <i>Phaeoacremonium krajdenii</i> ，波氏假阿利什霉，土曲霉，新月弯孢，斑替支孢瓶霉，甄氏外瓶霉，塞内加尔小球腔菌，汤普金小球腔菌，灰马杜拉分枝菌，足马杜拉分枝菌， <i>Pyrenochaeta romeroi</i> ，其他菌属	足分支菌病
耶氏肺孢子虫	肺孢子虫肺炎
耶氏肺孢子虫	肺炎
葡萄穗霉	呼吸系统的问题
红毛毛癣菌，扁桃体毛癣菌，叉指毛癣菌和/或薄荷毛癣菌，犬小孢子菌，絮状表皮藻	癣
孢粉	孢子菌病
马尔尼菲蓝状菌	塔拉菌病
球虫科	谷热

[0493] 可以包括在本公开的多分析物对照中的其他真菌物种包括但不限于新隐球菌和酿酒酵母。

[0494] 在一些实施方式中，多分析物对照包括真核寄生虫物种。示例性但非限制性的真

核寄生虫示于下表6中:

[0495] 表6. 示例性人类真核寄生虫

[0496]	真核寄生虫	疾病
	曼氏血吸虫 (<i>Chistosoma mansonii</i>), 埃及血吸虫 (<i>S. haematobium</i>), 日本血吸虫 (<i>S. japonicum</i>), 湄公血吸虫 (<i>S. mekongi</i>), 间插血吸虫 (<i>S. intercalatum</i>)	血吸虫病
	丝虫目菌属, 班氏丝虫, 马来丝虫, 帝纹丝虫	丝虫病
	布氏锥虫	锥虫病
	利什曼原虫	利什曼病
	溶组织性变形虫	变形虫病
	牛巴贝虫, 分歧巴贝虫	巴贝西虫病
	克氏锥虫	恰加斯病
	兰伯氏贾第虫 (<i>Giradia lamblia</i>)	贾第虫病
	卡伯特森氏棘阿米巴	肉芽肿性阿米巴性脑炎
	美洲钩虫	钩虫感染
	马来丝虫	淋巴丝虫病
	镰状疟原虫, 间日疟原虫	疟疾
	刚地弓形虫	弓形虫病
	阴道毛滴虫	毛滴虫病

[0497] 在示例性实施方式中, 表7 (改编自CDC的“应具报疾病的概述-美国, 2019”, 于2016年2月18日在 www.cdc.gov/nndss/conditions/notifiable/2019/) 访问) 提供进一步的示例性致病病原体, 其可以使用本文提供的方法和组合物进行量化。

[0498] 表7. 示例性致病性病原体

[0499]

炭疽病原体	李斯特菌病
虫媒病毒病原体，神经侵袭性和非神经侵袭性	卢霍病毒
Babiosis	莱姆病
肉毒杆菌病病原体	疟疾
布鲁氏菌病病原体	马尔堡病毒
加州血清群病毒	麻疹
加州血清群病毒病	脑膜炎球菌病
弯曲杆菌病	腮腺炎
假丝酵母	新世界沙粒病毒
软性下疳	新型甲型流感病毒感染
基孔肯雅病毒病	百日咳
沙眼衣原体感染	瘟疫
霍乱	脊髓灰质炎，麻痹
霍乱	脊髓灰质炎病毒感染，非麻痹性
球孢子菌病	玻瓦桑病毒
克里米亚-刚果出血热病毒	牛皮癣
隐孢子虫病	Q 热病
环孢子虫病	狂犬病
登革热	风疹
登革出血热	风疹，先天性综合症
登革热休克综合症	沙门氏菌病
登革热病毒感染	严重急性呼吸系统综合症相关冠状病毒（SARS-CoV）疾病
白喉	产生志贺毒素的大肠杆菌（STEC）
东方马脑炎病毒	志贺氏菌病
伊波拉病毒	天花
埃希氏菌病/无形体病	点状发热立克次体病
食源性病原体	圣路易斯脑炎病毒

[0500]	贾第虫病	链球菌中毒性休克综合症
	淋病	肺炎链球菌，侵袭性疾病
	流感嗜血杆菌，侵袭性疾病	梅毒
	汉森病（麻风病）	先天性梅毒
	汉坦病毒性肺综合症	破伤风
	溶血性尿毒症综合症，腹泻后	毒性休克综合症（链球菌除外）
	急性甲型肝炎	旋毛虫病
	乙肝病毒，围产期感染	结核
	急性乙型肝炎	兔热病
	慢性乙型肝炎	伤寒
	急性丙型肝炎	万古霉素中度（VISA）感染
	慢性丙型肝炎	耐万古霉素（VRSA）感染
	病毒性肝炎	水痘
	人类免疫缺陷病毒（HIV）	弧菌病
	流感相关的小儿死亡率	病毒性出血热
	拉萨病毒	西尼罗河病毒
	军团菌病	西部马脑炎病毒
	钩端螺旋体病	黄热病
		寨卡病毒

[0501] 在本公开的多分析物对照的一些实施方式中，多分析物对照中的生物的不同物种是人病原体。人类病原体的示例性物种包括表2-3和5-7中列出的任何物种。在多分析物对照中使用人类已知病原体允许多分析物对照对要使用本公开的方法确定其滴度的样本中的靶物种的行为进行建模。核酸提取的有效性，病原体基因组大小，文库制备效率，序列读段数和样本组成均会影响高通量测序文库的读段计数反映样本中靶生物的初始滴度的程度。包含与样本中的靶物种类似（例如，相同属）或相同物种的多分析物对照允许多分析物对照有效控制影响归一化读段反映样本中滴度的准确性的参数。在一些实施方式中，多分析物对照包括样本中的一种或多种靶生物。可以知道靶生物存在于样本中，或被怀疑存在于样本中（例如，多分析物对照包括一组常见的人类病原体）。

[0502] 在一些实施方式中，人类病原体包括通常在血液或组织移植中使用的组织中发现的人类病原体。移植获得性感染是移植接受者的重要原因或发病率和死亡率。因此，本公开的多分析物对照可用作筛选一种或多种人类病原体的移植组织和血液中的阳性对照，从而降低被移植组织或血液感染的风险。通常在人类血液或组织移植中发现的示例性病原体包括但不限于巨细胞病毒（CMV），爱泼斯坦-巴尔病毒（EBV），腺病毒（ADV），BK病毒（BKV），JC多瘤病毒（JCV），人疱疹病毒6A（HHV6A），人疱疹病毒6B（HHV6B），单纯疱疹病毒1型（HSV1），单纯疱疹病毒2型（HSV2），水痘带状疱疹病毒（VZV）和人类细小病毒B19（B19），人类免疫缺陷病毒（HIV），乙型肝炎，丙型肝炎，I型和II型人类T淋巴细胞性病毒（HTLV），西尼罗河病

毒,寨卡病毒,链球菌种,金黄色葡萄球菌,粪肠球菌,蜡状芽孢杆菌,痤疮丙酸杆菌,粘质沙雷氏菌,沙雷氏菌粘菌,小肠结肠炎耶尔森菌,肠杆菌属,不动杆菌属,假单胞菌属,大肠杆菌,肺炎克雷伯菌,奇异变形杆菌,鲍氏不动杆菌,蜡状芽孢杆菌,凝固阴性葡萄球菌,链球菌,克雷伯菌属,沙雷氏菌属雷氏普罗威登斯菌,苍白密螺旋体,克鲁斯锥虫和小巴贝虫中的一种或多种。

[0503] 在一些实施方式中,多分析物对照包含或基本上由11种病毒的混合物组成。示例性多分析物对照包括或基本上由以下组成:巨细胞病毒(CMV),爱泼斯坦-巴尔病毒(EBV),腺病毒(ADV),BK病毒(BKV),JC多瘤病毒(JCV),人疱疹病毒6A(HHV6A),人疱疹病毒6B(HHV6B),单纯疱疹病毒1型(HSV1),单纯疱疹病毒2型(HSV2),水痘带状疱疹病毒(VZV)和人细小病毒B19(B19)。在一些实施方式中,多分析物对照包括或基本上由10种病毒的混合物组成。示例性多分析物对照包括或基本上由以下组成:CMV(例如,菌株AD169),EBV,ADV(例如,1型),BKV(例如,亚型1b-2),JCV(例如,1a型),HHV6A(例如,菌株GS),HHV6B(例如Z-29菌株),HSV1(例如95菌株),HSV2(例如09菌株)和VZV(例如9/84菌株)。在一些实施方式中,多分析物对照包括或基本上由以下组成:CMV,EBV,ADV,BKV,JCV,HHV6A,HHV6B,HSV1,HSV2和VZV。另一个示例性多分析物对照包括或基本上由以下组成:麻疹病毒,西尼罗河病毒,寨卡病毒,黄热病病毒,辛德比斯病毒,天花病毒,诺沃克病毒,狂犬病病毒和人鼻病毒。另一个示例性多分析物对照包括或基本上由以下组成:爱泼斯坦-巴尔病毒,人类巨细胞病毒,人类呼吸道合胞病毒,天花病毒,流感病毒,风疹病毒,腮腺炎病毒和人类SARS冠状病毒。

[0504] 可以构成本公开的多分析物对照的示例性生物混合物如下表8所示:

[0505] 表8. 示例性多分析物对照

[0506]

示例性多分析物对照	物种
病毒面板	巨细胞病毒 (CMV), 爱泼斯坦-巴尔病毒 (EBV), 腺病毒 (ADV), BK 病毒 (BKV), JC 多瘤病毒 (JCV), 人疱疹病毒 6A (HHV6A), 人疱疹病毒 6B (HHV6B), 单纯疱疹病毒 1 型 (HSV1), 单纯疱疹病毒 2 型 (HSV2), 水痘带状疱疹病毒 (VZV), 人细小病毒 B19 (B19)
病毒面板	CMV, 乙型肝炎, 丙型肝炎, 牛痘病毒, 风疹病毒, 隆病毒属 A, 隆病毒属 B, 腮腺炎病毒, 麻疹病毒, 甲型流感, 乙型流感, 丙型流感, 人疱疹病毒 6
病毒面板	埃博拉病毒, 天花病毒, 黄热病病毒, 狂犬病病毒, 马丘波病毒, 拉萨病毒, HIV, 登革热病毒, 奇琴古尼亚病毒
病毒面板	CMV, EBV, ADV, BKV, JCV, HHV6A, HHV6B, HSV1, HSV2 和 VZV
病毒+细菌面板	CMV, EBV, ADV, BKV, JCV, HHV6A, HHV6B, HSV1, HSV2, VZV, B19, 金黄色葡萄球菌, 粪肠球菌, 蜡状芽孢杆菌, 痤疮丙酸杆菌, 粘质沙雷氏菌, 沙雷氏菌粘菌, 小肠结肠炎耶尔森氏菌
病毒、细菌和真菌小组	CMV, EBV, ADV, BKV, JCV, HHV6A, HHV6B, HSV1, HSV2, VZV, B19, 金黄色葡萄球菌, 粪肠球菌, 蜡状芽孢杆菌, 痤疮丙酸杆菌, 粘质沙雷氏菌, 沙雷氏菌粘菌, 小肠结肠炎耶尔森氏菌, 白色念珠菌, 曲霉菌, 新生隐球菌
多生物小组	CMV, EBV, ADV, BKV, JCV, HHV6A, HHV6B, HSV1, HSV2, VZV, B19, 金黄色葡萄球菌, 粪肠球菌, 蜡状芽孢杆菌, 痤疮丙酸杆菌, 粘质沙雷氏菌, 沙雷氏菌粘菌, 小肠结肠炎耶尔森氏菌, 白色念珠菌, 曲霉菌, 新生隐球菌, 布氏锥虫, 利什曼原虫, 蓝氏贾第鞭毛虫

[0507] 在本公开的多分析物对照的一些实施方式中,多分析物对照中的生物的至少三个不同物种不是人类病原体。在一些实施方式中,多分析物对照中的至少三种不同物种的生物不是人类病原体,而是与作为人类病原体的物种处于相同属。相同属中的物种通常具有类似的物理特性,例如,相同属中的细菌可能具有类似的细胞壁,而病毒可能具有类似的病毒蛋白。当经历本文所述的样本处理方法时,可以期望相同属的物种表现类似。通过将多分析物对照中的物种与相同属的人类病原体(例如常见于移植组织或血液中的人类病原体)相匹配,多分析物对照可以模拟样本处理期间样本中的那些人病原性物种的行为。这可以产生本公开的方法的优异能力来计算样本中的一种或多种靶生物的滴度,例如使用本公开

的多分析物对照和归一化对照的人临床样本的病原体。

[0508] 在本公开的多分析物对照的一些实施方式中,多分析物对照中的物种包括人类病原体的物种和不是人类病原体的物种的混合物。

[0509] 在一些实施方式中,多分析物对照还包括血浆。血浆可以来自任何脊椎动物。在一些实施方式中,血浆来自人,非人灵长类,小鼠,大鼠,兔,狗,猫,沙鼠,绵羊,猪,马,牛或驴。在一些实施方式中,血浆是人血浆。在一些实施方式中,血浆是合成产生的。

[0510] 本文提供从本公开的任何一种多分析物对照中提取的核酸的集合。可以从本文所述的多分析物对照中提取核酸的方法对于本领域普通技术人员而言将是显而易见的。示例性方法包括但不限于苯酚/氯仿萃取,然后用乙醇或类似的合适溶剂沉淀,以及可商购的试剂盒,例如Qiagen和Zymo mini-prep试剂盒。从多分析物对照中提取的核酸可用于本文所述的任何实施方式中。例如,从多分析物对照中提取的核酸可以在文库制备步骤中加到样本中,或与样本平行处理和测序,并用作阳性对照或生成校准曲线。

[0511] 制备多分析物对照的方法

[0512] 本公开提供制备本文所述的多分析物对照的方法。在一些实施方式中,该方法包括:(a)选择一组生物以包括在多分析物对照中;(b)在适当的培养条件下培养每种生物;(c)使每种生物灭活,和(d)将预定量的每种灭活生物与合适的载体、稀释剂或赋形剂混合以产生多分析物对照。

[0513] 在多分析物对照中使生物失活的适当方法是本领域普通技术人员已知的。例如,许多病毒可以通过干热、蒸汽或低pH(例如小于4.0的pH)灭活。包膜病毒可以通过加入溶剂或去污剂使其失活,从而破坏病毒的包膜。也可以通过暴露于化学交联剂(例如甲醛)使病毒灭活。使细菌失活的方法包括紫外线,辐射,加热,低PH和用化学物质(例如环氧乙烷,甲醛和硫氰酸胍)处理。单细胞真核生物(例如变形虫或酵母)可以通过加热或化学交联剂(例如甲醛)进行灭活。

[0514] 在一些实施方式中,生物已通过化学处理、热处理、pH处理或紫外线照射灭活。在一些实施方式中,化学处理包括甲醛。

[0515] 在一些实施方式中,将多分析物对照中的已知量的灭活生物加到血浆中。血浆可以例如通过离心和/或过滤去除血细胞、血小板和其他细胞组分而与血液分离。血浆可以来自任何脊椎动物。在一些实施方式中,血浆来自人,非人灵长类,小鼠,大鼠,兔子,沙鼠,狗,猫,牛,绵羊,猪,马或驴。在一些实施方式中,血浆来自人。在样本处理期间,将灭活的生物加到血浆中可使多分析物对照紧密模拟实验样本的行为,例如包含人血浆的人临床样本。

[0516] 使用多分析物对照的方法

[0517] 本公开提供使用本文所述的多分析物对照的方法。在一些实施方式中,本公开的多分析物对照可以与样本,例如临床或实验样本并行处理。与样本并行处理的多分析物对照可用于产生校准曲线,本领域普通技术人员可使用该校准曲线确定包含生物混合物的样本中一种或多种生物的滴度。例如,样本可以是临床样本,其包含来自人宿主的细胞和来自非宿主例如病毒、细菌或真菌病原体的细胞或颗粒。在替代实施方式中,将多分析物对照直接加到样本,例如临床或实验样本,并用作内标以计算包含生物混合物的样本中一种或多种生物体的滴度。在一些实施方式中,多分析物对照用作本公开的方法的阳性对照。

[0518] 在一些实施方式中,与样本平行但与样本分开处理多分析物对照,以产生高通量

测序文库和来自其的读段的集合。在一些实施方式中,加到多分析物对照和样本的归一化对照用于归一化来自多分析物对照和样本的读段。在其中多分析物对照和样本被并行处理的那些实施方式中(即,没有将多分析物对照“掺入”样本中),多分析物对照包括作为人类病原体的物种、不是人类病原体的物种或其组合。在一些实施方式中,来自多分析物对照的归一化读段用于产生校准曲线,并且该校准曲线用于计算样本中靶生物的滴度。

[0519] 在本公开的方法的一些实施方式中,所述方法用于量化多分析物对照中的每种生物物种的至少一个靶序列的表达水平。在一些实施方式中,所述方法包括(a)将已知量的归一化对照与多分析物对照混合,(b)制备高通量测序文库,(c)对所述文库测序以产生读段集合,(d)将来自读段集合的读段映射到多分析物对照或归一化对照,(e)确定被归一化对照中的多核苷酸的每个组或亚组产生的读段数量,(f)计算与(a)中的多分析物对照混合的归一化对照中的多核苷酸的每个组或亚组的初始浓度和(c)中产生的读段数之间的关系,(g)建模读段与多分析物对照中的每个物种的至少一个靶序列的浓度之间的关系,和(h)使用(g)中的模型通过靶序列产生的读段数,计算多分析物对照中的每个物种的至少一个靶序列的初始浓度。在一些实施方式中,该方法进一步包括从多分析物对照中提取核酸。在一些实施方式中,(a)的混合步骤在(b)的文库制备步骤之前,与(b)的文库制备步骤同时或在(b)的文库制备步骤之后进行。在一些实施方式中,(a)的混合步骤在从样本提取核酸之前进行。

[0520] 在一些实施方式中,多分析物对照中的生物体的每个不同物种包括至少一个靶序列。在一些实施方式中,至少一种靶序列是多分析物对照中的每种生物物种的不同靶序列。在一些实施方式中,多分析物对照中的每个物种包括一个或多个不同的靶序列。在一些实施方式中,在多分析物对照中来自每个生物物种的至少一个靶序列与来自多分析物对照中任何一种或多种其他生物的靶序列具有小于99%同一性,小于95%同一性,小于90%同一性,小于85%同一性,小于80%同一性,小于75%同一性,小于70%同一性,小于60%同一性,小于50%同一性,小于40%同一性,小于30%同一性或小于20%同一性。

[0521] 在一些实施方式中,在多分析物对照中来自每个生物物种的至少一个靶序列与来自样本的靶序列相同,具有小于99%同一性,小于95%同一性,小于90%同一性,小于85%同一性,小于80%同一性,小于75%同一性,小于70%同一性,小于60%同一性,小于50%同一性,小于40%同一性,小于30%同一性,小于20%同一性,小于10%同一性或小于5%同一性。在一些实施方式中,在多分析物对照中来自生物物种的至少一个靶序列与来自样本的靶序列相同。在一些实施方式中,多分析物对照包括包含与来自样本的靶序列相同的靶序列的至少一种物种,和包含与来自样本的靶序列不相同的靶序列的至少一种物种。

[0522] 在一些实施方式中,多分析物对照中的每种生物物种的至少一个靶序列包含基因组序列。在一些实施方式中,在多分析物对照中来自每个生物物种的至少一个靶序列包含该物种的参考基因组或由该物种的参考基因组组成。在一些实施方式中,在多分析物对照中来自每个生物物种的至少一个靶序列包含该物种的转录组或由该物种的转录组组成。在一些实施方式中,来自多分析物对照中的每个生物物种的至少一个靶序列包含该物种的参考基因组的一部分。所述至少一个靶序列可以是蛋白质编码序列或非编码序列。

[0523] 在一些实施方式中,在步骤(d)将来自读段集合的读段映射到多分析物对照或归一化对照还包括将读段映射到多分析物对照(d)中的每个生物物种。来自高通量测序读段

的集合的读段的映射方法对于本领域普通技术人员将是显而易见的。例如,在多分析物对照中的每个生物具有参考基因组的情况下,来自通过高通量序列生成的读段的集合的读段可以经由BWA,Novoalign,Bowtie,SOAP2,BFAST,SSAHA2,MPscan,GASSST或PerM算法映射到参考基因组。在一些实施方式中,将读段映射到多分析物对照中的每个生物参考基因组包括将读段映射到多分析物对照中的每个生物物种的至少一个靶序列。

[0524] 在一些实施方式中,在(h)计算至少一个靶序列的初始浓度包括在多分析物对照中为每种物种计算至少一个靶序列的初始浓度。多分析物对照中的每种物种的至少一种靶序列的初始浓度可以从读段计数与归一化对照中的每个组或亚组的初始浓度之间的关系计算,该关系使用归一化对照确定本公开的组合物和方法。

[0525] 本公开提供量化样本中靶生物的滴度的方法。在一些实施方式中,所述方法包括(a)提供包含靶生物样本,其中所述靶生物包含至少一个靶序列;(b)提供包含至少三种生物物种的已知滴度的多分析物对照,其中所述生物已被失活;(c)将已知量的本公开的归一化对照与样本和多分析物对照混合;(d)从样本和至少一个多分析物对照制备高通量测序文库;(e)对所述文库测序以产生样本读段集合和多分析物对照读段集合;(f)使用归一化对照使来自(e)的样本读段集合和多分析物对照读段集合归一化;(g)确定多分析物对照中归一化读段与已知生物物种滴度之间的关系;和(h)使用(g)中确定的关系计算样本中靶生物的滴度。

[0526] 在一些实施方式中,生物物种各自包含至少一种物种特异性靶序列。在一些实施方式中,至少一种物种特异性靶序列在多分析物对照中的每个物种中是不同的。

[0527] 在一些实施方式中,归一化来自多分析物对照的读段包括:(i)将来自样本读段集合的读段映射到样本或归一化对照;(ii)将来自多分析物对照读段集合的读段映射到多分析物对照或归一化对照;(iii)确定样本读段集合和多分析物对照读段集合的归一化对照中的多核苷酸的每一组或亚组产生的读段的数目;(iv)针对样本和多分析物对照,计算在归一化对照中的核苷酸的每个组或亚组的初始浓度和步骤(e)中产生的读段的数目之间的关系;以及(v)确定样本和多分析物对照中的读段与浓度之间的关系。

[0528] 在一些实施方式中,确定多分析物对照和样本中的读段和滴度之间的关系包括:(i)从多分析物对照中的每个物种计算至少一种物种特异性靶序列的初始浓度;(ii)计算样本中的至少一个靶序列的初始浓度。在一些实施方式中,步骤(i)的映射进一步包括将样本读段集合映射到参考基因组。在一些实施方式中,参考基因组不包含靶序列。在一些实施方式中,参考基因组是人类基因组。例如,在包含人类宿主细胞和病原体,即包含靶序列的病原体的人类临床样本中,由样本的高通量测序文库产生的读段被映射到人类基因组作为参考基因组。映射到人类基因组并与人类基因组对齐的读段来自从人类细胞分离的核酸,并且不包含靶序列。未映射到人参考基因组的读段来自靶序列,并使用本公开的方法进行归一化。这种方法的优点是不一定需要提前知道产生靶序列的靶生物的身份。本公开的方法可以用于有效地确定样本中的未知靶生物的存在和滴度,例如人类临床样本中未知病原体。

[0529] 在一些实施方式中,在步骤的映射进一步包括将样本读段集合映射到多于一个参考基因组。在一些实施方式中,第一参考基因组不包含靶序列,而第二参考基因组包含靶序列。例如,在包含人宿主细胞和病原体即包含靶序列的病原体的人临床样本中,将从样本的

高通量测序文库产生的读段映射到不包含靶序列的人基因组,和对应于包含靶序列的病原体的第二参考基因组。在一些实施方式中,映射还包括针对多分析物对照中的每个生物物种,映射到物种特异性参考基因组。

[0530] 在一些实施方式中,在多分析物对照中归一化读段与生物的已知滴度之间的关系是线性关系。在一些实施方式中,该关系是对数的。在多分析物对照中确定归一化读段与生物的已知滴度物种之间的关系的方法对于本领域普通技术人员而言将是显而易见的。

[0531] 在一些实施方式中,多分析物对照用于产生校准曲线,该校准曲线用于从归一化读段计数确定样本中生物的滴度。示例性校准曲线在图13中显示并显示为相对于用于生成校准曲线的多分析物对照中生物的对数滴度的信号(归一化读段)。在一些实施方式中,校准曲线与一个或多个样本(例如临床或实验样本)平行产生。在一些实施方式中,校准曲线由在多种不同浓度下分析的多分析物对照产生。例如,可以从以1:10、1:100、1:1000等稀释的多分析物对照生成示例性校准曲线,然后进行测序。在一些实施方式中,校准曲线由以2个浓度,3个浓度,4个浓度,5个浓度,6个浓度,7个浓度,8个浓度,9个浓度或10个浓度提供的多分析物对照生成。在一些实施方式中,校准曲线包括参考校准曲线。例如,可以将样本中的靶序列的归一化读段与参考校准曲线进行比较,以确定样本中与靶序列相对应的生物的滴度。在一些实施方式,在例如采用参考校准曲线的那些实施方式中,与样本平行地处理一个或多个多分析物对照以验证校准曲线和/或用作阳性对照。例如,将高浓度和低浓度的两种多分析物对照与样本并行处理。

[0532] 在本公开的方法的一些实施方式中,所述方法进一步包括将已知量的多分析物对照与样本混合(掺入方法)。多种分析物的对照用作内标以计算包含生物混合物的样本中的一种或多种生物的滴度。例如,样本可以是包含宿主细胞(人细胞)和一种或多种病毒颗粒或来自细菌或真菌病原体(非宿主)的细胞的人类临床样本。多分析物对照包括不是人类病原体的生物,因此可以使用序列差异将其与样本中的人类病原体区分开。

[0533] 在一些实施方式中,多分析物对照包括与人类病原体相同属的生物。与人类病原体有关的生物在处理期间的行为将与样本中的人类病原体类似,因此为计算样本中的人类病原体的滴度提供优越的指标。在一些实施方式中,多分析物对照中的生物包含具有与人类病原体序列的小于99%同一性,小于95%同一性,小于90%同一性,小于85%同一性,小于80%同一性,小于75%同一性,小于70%同一性,小于60%同一性,小于50%同一性,小于40%同一性,小于30%同一性或小于20%同一性的靶序列。

[0534] 制剂

[0535] 本公开的归一化对照和多分析物对照组合物可以配制在本领域已知的任何可接受的载体、稀释剂或赋形剂中。

[0536] 可以在溶液中以任何可接受的浓度、沉淀或冻干的组合物形式提供归一化对照。核酸例如归一化对照可以悬浮在溶液中,包括但不限于具有至多0.5M NaCl的水,Tris缓冲液,Tris-EDTA缓冲液和Tris-EDTA。

[0537] 在本公开的制剂中使用的示例性缓冲剂包括三-(羟甲基)氨基甲烷(TRIS)的乙酸盐、硫酸盐、盐酸盐、磷酸盐或游离酸形式,尽管可以使用与TRIS相同的近似离子强度和pKa的替代缓冲剂,其结果相同。除了缓冲盐之外,在组合物中还包括辅因子盐,例如钾(例如氯化钾或乙酸钾)和镁(例如氯化镁或乙酸镁)的那些。

[0538] 可以将多分析物对照中的灭活生物体以溶液形式提供,冷冻干燥或冻干。如果冷冻干燥或冻干,可以将多分析物对照重新悬浮在适当体积的任何可接受的载体中。可接受的载体包括但不限于水,磷酸盐缓冲盐水(PBS)或林格氏溶液。

[0539] 消耗或富集的方法

[0540] 归一化对照、多分析物对照和使用它们的方法可以与富集感兴趣的靶序列的样本或消耗靶向耗尽的序列的样本的方法组合。在一些实施方式中,富集或消耗的方法包括基于核酸引导的核酸酶的方法,基于核苷酸修饰的方法或其组合。

[0541] 因此,本公开提供量化或识别样本中的靶序列的方法,该方法包括消耗靶向消耗的序列的样本,加入本文所述的归一化对照和使用NGS对样本进行测序。可以在富集感兴趣的靶序列或靶向耗尽的核酸消耗之前、之中或之后加入归一化对照。

[0542] 在一些实施方式中,在富集感兴趣的靶序列或靶向耗尽的序列消耗之前,将归一化对照加到测序样本中。在一些实施方式中,可以将归一化对照设计为模拟感兴趣的靶序列的那些特性,所述特性用于将那些序列与靶向耗尽的序列区分开。例如,如果使用核酸引导的核酸酶和靶向靶向耗尽的序列的多个引导核酸(gNA)富集感兴趣的靶序列,则归一化对照不包含被多个gNA靶向的序列。作为进一步的示例,如果使用基于核苷酸修饰的方法消耗靶向耗尽的序列,则归一化对照可以包括与富集的感兴趣的靶向序列的修饰类似的核苷酸修饰。

[0543] 基于核苷酸修饰的方法

[0544] 本文提供相对于靶向耗尽的核酸富集感兴趣的靶核酸的样本的方法,包括使用感兴趣的靶核酸与靶向耗尽的核酸之间核苷酸修饰的差异。在一些实施方式中,靶核酸包含测序样本中的靶序列,其存在和/或滴度将使用本公开的归一化对照和任意的多分析物对照来确定。可以在基于修饰的消耗方法之前将包含与感兴趣的靶序列相同的核苷酸修饰的归一化对照加到样本中。可替代地,可以在基于修饰的消耗后将归一化对照加到样本中。

[0545] 设想在本公开的范围内的任何类型的核苷酸修饰。下文描述本公开的核苷酸修饰的示例性但非限制性示例。

[0546] 由本公开的方法使用的核苷酸修饰可以在任何核苷酸(例如腺嘌呤、胞嘧啶、鸟嘌呤、胸腺嘧啶或尿嘧啶)上发生。这些核苷酸修饰可发生在脱氧核糖核酸(DNA)或核糖核酸(RNA)上。这些核苷酸修饰可以发生在双链或单链DNA分子上,或发生在双链或单链RNA分子上。

[0547] 在一些实施方式中,核苷酸修饰包括腺嘌呤修饰或胞嘧啶修饰。

[0548] 在一些实施方式中,腺嘌呤修饰包括腺嘌呤甲基化。在一些实施方式中,腺嘌呤甲基化包括N⁶-甲基腺嘌呤(6mA)。在一些实施方式中,腺嘌呤甲基化包括由脱氧腺苷甲基化酶进行的Dam甲基化。在一些实施方式中,腺嘌呤甲基化包括EcoKI甲基化。在一些实施方式中,腺嘌呤修饰包括在N6处通过甘氨酸修饰的腺嘌呤(momylation)。

[0549] 在一些实施方式中,修饰包括胞嘧啶修饰。在一些实施方式中,胞嘧啶修饰包括5-甲基胞嘧啶(5mC),5-羟基甲基胞嘧啶(5hmC),5-甲酰基胞嘧啶(5fC),5-羧基胞嘧啶(5caC),5-葡萄糖基羟甲基胞嘧啶(5ghmC)或3-甲基胞嘧啶(3mC)。在一些实施方式中,胞嘧啶甲基化包括5-甲基胞嘧啶(5mC)或N⁴-甲基胞嘧啶(4mC)。在一些实施方式中,胞嘧啶甲基化包括Dcm甲基化,DNMT1甲基化,DNMT3A甲基化或DNMT3B甲基化。在一些实施方式中,胞嘧

啶甲基化包括CpG甲基化、CpA甲基化、CpT甲基化、CpC甲基化,或它们的组合。在一些实施方式中,胞嘧啶甲基化包括CpG甲基化。例如,使用本公开的方法,CpG甲基化可以用于选择性地靶向哺乳动物基因组中的活性区域以进行消耗。

[0550] 在一些实施方式中,胞嘧啶修饰包括5-羟甲基胞嘧啶(5hmC),5-甲酰基胞嘧啶(5fC)。5-甲酰基胞嘧啶是5mC,5-羧基胞嘧啶(5caC),5-葡萄糖基羟甲基胞嘧啶或3-甲基胞嘧啶的氧化衍生物。

[0551] 在本公开的方法的一些实施方式中,所述方法至少使用第一修饰敏感的限制性酶和第二修饰敏感的限制性酶。在一些实施方式中,第一和第二修饰敏感的限制性酶是相同的。在一些实施方式中,第一和第二修饰敏感的限制性酶是不同的。在一些实施方式中,第一或第二修饰敏感的限制性酶是单一物种的限制性酶(例如,AluI或McrBC,但不是两者)。在一些实施方式中,第一或第二修饰敏感的限制性酶是2种或更多种修饰敏感的限制性酶的混合物(例如,FspEI和AbaSI的混合物)。在本公开的方法的一些实施方式中,组合两种以上的不同方法,每种方法使用不同的修饰敏感的限制性酶或修饰敏感的限制性酶的混合物。

[0552] 如本文所用,术语“修饰敏感的限制性酶”是指对限制性酶的识别位点内或附近的修饰核苷酸的存在敏感的限制性酶。可替代地或另外,修饰敏感的限制性酶可以对识别位点本身内的修饰核苷酸敏感。修饰敏感的限制性酶可以对与识别位点相邻的修饰核苷酸敏感,例如在识别位点的1-50个核苷酸,5'或3'内。本公开的核苷酸修饰可以在识别位点本身内,或包含与识别位点相邻的核苷酸(例如,在识别位点的1-50个核苷酸,5'或3',或两者内)。

[0553] 能够阻断或降低修饰敏感的限制性酶的活性的示例性修饰包括但不限于N6-甲基腺嘌呤,5-甲基胞嘧啶(5mC),5-羟基甲基胞嘧啶(5hmC),5-甲酰基胞嘧啶(5fC),5-羧基胞嘧啶(5caC),5-葡萄糖基羟甲基胞嘧啶,3-甲基胞嘧啶(3mC),N4-甲基胞嘧啶(4mC)或其组合。能够阻断修饰敏感的限制性酶的示例性修饰包括由Dam,Dcm,EcoKI,DNMT1,DNMT3A,DNMT3B和TET酶介导的修饰。

[0554] 在一些实施方式中,修饰包括Dam甲基化。被Dam甲基化阻断的限制性酶包括但不限于AlwI,BcgI,BclI,BsaBI,BspDI,BspEI,BspHI,ClaI,DpnII,HphI,Hpy188I,Hpy188III,MboI,MboII,NruI,Nt.AlwI,Taq α I和XbaI。

[0555] 在一些实施方式中,修饰包括Dcm甲基化。被Dcm甲基化阻断的限制性酶包括但不限于Acc65I,AlwNI,ApaI,AvaI,AvaII,BanI,BsaI,BsaHI,BslI,BsmFI,BssKI,BstXI,EaeI,Esp3I,Eco0109I,MscI,NlaIV,PflMI,PspGI,PspOMI,Sau96I,ScrFI,SexAI,SfiI,SfoI和StuI。被CpG甲基化阻断的限制性酶包括但不限于AatII,AccII,AciI,AcII,AfeI,AgeI,Aor13HI,Aor51HI,AscI,AsiSI,AluI,AvaI,BceAI,BmgBI,BsaI,BsaHI,BsiEI,BsiWI,BsmBI,BspDI,BspT104I,BsrFalphaI,BssHII,BstBI

[0556] BstUI,Cfr10I,ClaI,CpoI,EagI,Esp3I,Eco52I,FauI,FseI,FspI,HaeII,HgaI,HhaI,HpaII,HpyCH4IV,Hpy99I,KasI,MluI,NaeI,NgoMIV,NotI,NruI,NtB.CviPII,NsbI,PmaCI,Psp1406I,PluTI,PmlI,PvuI,RsrII,SacII,SalI,SmaI,SnaBI,SfoI,SgrAI,SmaI,SrfI,Sau3AI,TspMI和ZraI。

[0557] 在一些实施方式中,修饰敏感的限制性内切酶在包含至少一个修饰核苷酸的识别

位点处是活性的,而在不包含至少一个修饰核苷酸的识别位点处不是活性的。在包含一个或多个修饰核苷酸的识别位点处裂解的修饰敏感的限制性酶识别的示例性修饰包括但不限于N6-甲基腺嘌呤,5-甲基胞嘧啶(5mC),5-羟基甲基胞嘧啶(5hmC),5-甲酰基胞嘧啶(5fC),5-羧基胞嘧啶(5caC),5-葡萄糖基羟甲基胞嘧啶,3-甲基胞嘧啶(3mC),N4-甲基胞嘧啶(4mC)或其组合。特异性裂解包含一个或多个修饰核苷酸的识别位点的示例性修饰识别的修饰敏感的限制性酶包括由Dam,Dcm,EcoKI,DNMT1,DNMT3A,DNMT3B和TET酶介导的修饰。

[0558] 在包含在识别位点内或附近的一个或多个修饰核苷酸的识别位点处裂解的示例性但非限制性的修饰敏感的限制性酶包括但不限于AbaSI,DpnI,FspEI,LpnPI,MspJI和McrBC。

[0559] 在一些实施方式中,修饰包括5-葡萄糖基羟甲基胞嘧啶,并且修饰敏感的限制性酶包含AbaSI。AbaSI切割包含葡萄糖基羟甲基胞嘧啶的AbaSI识别位点,并且不切割不包含葡萄糖基羟甲基胞嘧啶的AbaSI识别位点。

[0560] 在一些实施方式中,核苷酸修饰包含5-羟甲基胞嘧啶,且修饰敏感的限制性酶包含AbaSI和T4噬菌体 β -葡萄糖基转移酶。T4噬菌体 β -葡萄糖基转移酶可将尿苷二磷酸葡萄糖(UDP-Glc)的葡萄糖部分特异性转移到双链DNA中的5-羟甲基胞嘧啶(5-hmC)残基,例如在AbaSI识别位点内,从而制备糖基羟甲基胞嘧啶修饰的AbaSI识别位点。AbaSI切割包含葡萄糖基羟甲基胞嘧啶的AbaSI识别位点,并且不切割不包含葡萄糖基羟甲基胞嘧啶的AbaSI识别位点。

[0561] 在一些实施方式中,核苷酸修饰包含甲基胞嘧啶,而修饰敏感的限制性酶包含McrBC。McrBC切割包含甲基胞嘧啶的McrBC位点,而不切割不包含甲基胞嘧啶的McrBC位点。可以用一条或两条DNA链上的甲基胞嘧啶修饰McrBC位点。在一些实施方式中,McrBC还切割一条或两条DNA链上的包含羟甲基胞嘧啶的McrBC位点。在一些实施方式中,McrBC半位点被至多3,000个核苷酸隔开。在一些实施方式中,McrBC半位点被55-103个核苷酸隔开。

[0562] 在一些实施方式中,修饰包括腺嘌呤甲基化,并且所述方法包括用DpnI消化。当GATC识别的两条链上的腺嘌呤被甲基化时,DpnI会切割GATC识别位点。在一些实施方式中,同时包含腺嘌呤甲基化和胞嘧啶修饰的DpnI GATC识别位点出现在细菌DNA中,而不存在于哺乳动物DNA中。包含甲基化腺嘌呤和修饰的胞嘧啶的这些识别位点可以通过样本中的DpnI选择性切割(例如,混合的细菌和哺乳动物DNA),然后用T4聚合酶处理以用未修饰的腺嘌呤和胞嘧啶替换切割末端的甲基化腺嘌呤和修饰的胞嘧啶。在存在模板、引物和核苷酸的情况下,T4聚合酶催化5'至3'方向的DNA合成。T4聚合酶会将未修饰核苷酸整合到新合成的DNA中。这产生现在感兴趣的核酸中包含未修饰的胞嘧啶并且在靶向耗尽的核酸中包含修饰的胞嘧啶的样本。使用本公开的方法,修饰的胞嘧啶中的这些差异可用于富集感兴趣的核酸。

[0563] 在本公开的方法的一些实施方式中,样本中的核酸被末端去磷酸化,从而使样本中的核酸与修饰敏感的限制性酶接触产生具有暴露的末端磷酸酯的感兴趣的核酸或靶向耗尽的核酸,这些暴露的末端磷酸酯可以在本公开的方法中使用以富集感兴趣的核酸的样本。例如,这些暴露的末端磷酸酯可用于靶向核酸被核酸外切酶降解而消耗(图28)或感兴趣的核酸用于衔接子连接(图27)。

[0564] 如本文所用,术语“末端去磷酸化”是指具有从核酸分子的5'和3'末端去除的末端

磷酸基团的核酸。在一些实施方式中,使用磷酸酶例如碱性磷酸酶将样本中的核酸末端去磷酸化。本公开的示例性磷酸酶包括但不限于虾碱性磷酸酶(SAP),重组虾碱性磷酸酶(rSAP),小牛肠碱性磷酸酶(CIP)和南极磷酸酶。

[0565] 如本文所用,术语“核酸外切酶”是指从核酸分子的3'或5'末端连续去除核苷酸的一类酶。核酸分子可以是DNA或RNA。DNA或RNA可以是单链或双链的。示例性核酸外切酶包括但不限于λ核酸酶,核酸外切酶I,核酸外切酶III和BAL-31。核酸外切酶可以用于使用本公开的方法选择性地降解靶向消耗的核酸(例如,图28)。

[0566] 本公开提供与样本或感兴趣的核酸以及可选的归一化对照中的核酸的5'和3'末端连接的衔接子。在一些实施方式中,衔接子与归一化对照序列连接。在其他实施方式中,从头合成包含衔接子序列的归一化对照。

[0567] 在本公开的方法的一些实施方式中,将衔接子连接到样本中的所有核酸,然后使用核苷酸修饰的差异来选择性地切割靶向耗尽的核酸,从而产生在两端被衔接子连接的感兴趣的核酸和在一端被衔接子连接的靶向耗尽的核酸(图29,图30)。在一些实施方式中,核苷酸修饰的差异用于选择性地消耗靶向耗尽的核酸,然后将衔接子连接到感兴趣的靶核酸(图28)。在一些实施方式中,核苷酸修饰的差异用于产生具有暴露的末端磷酸酯的感兴趣的核酸,所述末端磷酸酯用于将衔接子连接到感兴趣的靶核酸(图27)。

[0568] 在本公开的方法的一些实施方式中,将衔接子连接到样本中的核酸的5'和3'末端。在一些实施方式中,衔接子还包括在5'末端和/或3'末端之间的插入序列。

[0569] 在一些实施方式中,衔接子是可与双链DNA分子的两条链连接的核酸。

[0570] 在一些实施方式中,在消耗/富集之前连接衔接子。在其他实施方式中,衔接子在随后的步骤中被连接。

[0571] 衔接子的非限制性示例包括线性,线性Y形或发夹衔接子。在一些实施方式中,衔接子包含polyG序列。

[0572] 在各种实施方式中,衔接子可以是发夹衔接子,即,一个与其自身碱基配对以形成具有双链茎和环的结构分子,其中该分子的3'和5'末端分别连接到片段的双链DNA分子的5'和3'末端。

[0573] 可替代地,衔接子可以是连接到片段的一端或两端的Y-衔接子,也称为通用衔接子。可替代地,衔接子本身可以由彼此碱基配对的两个不同的寡核苷酸分子组成。另外,衔接子的可连接末端可以设计成与通过限制性酶切割产生的突出端相容,或它可以具有平末端或5'T突出端。在一些实施方式中,限制性酶是修饰敏感的限制性酶。

[0574] 衔接子可以包括双链以及单链分子。因此,衔接子可以是DNA或RNA,或两者的混合物。含有RNA的衔接子可以通过RNase处理或碱性水解来裂解。

[0575] 衔接子的长度可以为10至100bp,尽管在不偏离本公开的情况下可以使用在该范围之外的衔接子。

[0576] 衔接子可以配置用于下一代测序平台,例如,在Illumina测序平台(例如HiSeq或MiSeq)上使用,或在IonTorrent平台上使用,或与Nanopore技术一起使用。在一些实施方式中,衔接子包括测序衔接子(例如,Illumina测序衔接子)。在一些实施方式中,衔接子包含独特的分子标识符(UMI)序列,有时称为条形码。在一些实施方式中,UMI序列包含对于每个原始核酸分子而言唯一的序列(例如,随机序列)。在一些实施方式中,衔接子包含多个不同

的序列,例如每个核酸分子特有的UMI,来自特定来源的核酸分子之间共享的条形码和测序衔接子。

[0577] 靶向消耗的核酸可通过差异衔接子连接而消耗。在一些实施方式中,将衔接子连接到样本的核酸,并且随后基于其修饰状态从靶向消耗的核酸中去除一个或多个衔接子。例如,可以通过修饰敏感的限制性酶切割两端连接有衔接子的靶向消耗的核酸,从而产生在仅一端连接有衔接子的靶向消耗的核酸。后续步骤(例如,扩增)可用于仅靶向两端连接有衔接子的核酸,从而消耗靶向消耗的核酸。在另一个实施例中,样本的核酸被处理(例如,通过去磷酸化),使得仅被切割的核酸能够具有连接的衔接子;随后,感兴趣的核酸可以被修饰敏感的限制性酶(例如,从而暴露磷酸酯基团)切割,并且可以连接衔接子。后续步骤(例如,扩增)可用于仅靶向连接有衔接子的核酸,从而消耗靶向消耗的核酸。

[0578] 靶向消耗的核酸可以通过消化(例如用核酸外切酶消化)来消耗。

[0579] 靶向消耗的核酸可以通过大小选择来消耗。例如,修饰敏感的限制性内切酶可用于切割感兴趣的核酸或靶向耗尽的核酸,并且随后可基于由于切割导致的大小差异将感兴趣的核酸与靶向耗尽的核酸分离。

[0580] 在一些情况下,不使用大小选择就消耗靶向消耗的核酸。

[0581] 方案1:在图27中描述本文描述的本申请的示例性方法。对包含感兴趣的靶核酸(2701)和靶向耗尽的核酸(2702)以及可选的归一化对照的核酸样本进行末端去磷酸化(2705),以产生未磷酸化的感兴趣的核酸(2706)和靶向耗尽的核酸(2707)。在一些实施方式中,核酸在去磷酸化之前被片段化。在一些实施方式中,样本中的核酸被磷酸酶例如重组虾碱性磷酸酶(rSAP)末端去磷酸化。在一些实施方式中,感兴趣的核酸和靶向耗尽的核酸都包括一个或多个用于修饰敏感的限制性酶(分别为2703,2704)的识别位点。在感兴趣的核酸中,修饰敏感的限制性内切酶的识别位点不包含修饰核苷酸(2703),或可替代地,比靶向耗尽的核酸的对应识别位点较不频繁地包含修饰核苷酸。在靶向耗尽的核酸中,修饰敏感的限制性酶的识别位点包含在限制位点内或附近的修饰核苷酸(2704),或可替代地,比感兴趣的核酸的对应识别位点更频繁地包含修饰核苷酸。修饰敏感的限制性酶(2709)的活性被其同源识别位点(2708)内或附近的修饰核苷酸的存在阻断,从而将修饰敏感的限制性酶的活性靶向到感兴趣的核酸(比较2710与2711)。在一些实施方式中,修饰敏感的限制性酶(2709)包括AatII,AccII,Aor13HI,Aor51HI,BspT104I,BssHII,Cfr10I,ClaI,CpoI,Eco52I,HaeII,HapII,HhaI,MluI,NaeI,NotI,NruI,NsbI,PmaCI,Psp1406I,PvuI,SacII,SalI,SmaI,SnaBI,AluI或Sau3AI。在一些实施方式中,修饰敏感的限制性酶(2709)包含AluI或Sau3AI。用修饰敏感的限制性内切酶消化样本(2713)产生在末端磷酸酯(2714)的5'和3'端具有末端磷酸酯的感兴趣的核酸。这些末端磷酸酯用于将衔接子(2715,连接步骤;116,衔接子)连接到感兴趣的核酸的末端,产生两端均衔接子连接的感兴趣的核酸(2717)。相反,靶向消耗的核酸没有衔接子连接(2711)。这些衔接子可用于下游应用,例如衔接子介导的PCR扩增,测序(例如高通量测序)和样本中靶核酸的量化和/或克隆。通过将衔接子与感兴趣的核酸连选择性地接,从而消耗靶向耗尽的核酸。无需使用大小选择即可完成该消耗。可替代地,将感兴趣的衔接子连接的核酸进行本文所述的一种或多种其他富集方法。例如,使衔接子连接的核酸经受本公开的另外的依赖修饰的富集方法(例如,图29中描绘的方法)。替代地或另外地,使衔接子连接的核酸经受本公开的基于核酸引导的核酸酶的富集方

法(例如,图30中描绘的方法)。

[0582] 方案2:在图28中描绘在此描述的本申请的示例性方法。对包含感兴趣的靶核酸(2801)和靶向耗尽的核酸(2802)以及任选的归一化对照的核酸的样本进行末端去磷酸化(2805)以产生未磷酸化的感兴趣的核酸(2806)和靶向耗尽的核酸(2807)。在一些实施方式中,核酸在去磷酸化之前被片段化。在一些实施方式中,样本中的核酸被磷酸酶例如重组虾碱性磷酸酶(rSAP)末端去磷酸化。在一些实施方式中,感兴趣的核酸和靶向耗尽的核酸都包含一个或多个修饰敏感的限制性内切酶的识别位点(分别为2803和2804)。在感兴趣的核酸中,修饰敏感的限制性内切酶的识别位点不包含修饰核苷酸(2803),或可替代地,比靶向耗尽的核酸的对应识别位点较不频繁地包含修饰核苷酸。在靶向耗尽的核酸中,修饰敏感的限制性内切酶的识别位点包含在限制位点内或附近的修饰核苷酸(2804),或可替代地,比感兴趣的核酸的对应识别位点更频繁地包含修饰核苷酸。当在识别位点(2808)内或附近存在一个或多个修饰核苷酸时,修饰敏感的限制性内切酶(2809)切割其同源识别位点,而当识别位点不包含一个或多个修饰核苷酸(2808)时,则不切割其同源识别位点,从而将修饰敏感的限制性酶的活性靶向到靶向消耗的核酸(比较2810与2811)。在一些实施方式中,修饰敏感的限制性酶包括AbaSI, FspEI, LpnPI, MspJI或McrBC。在一些实施方式中,修饰敏感的限制性内切酶是FspEI。在一些实施方式中,修饰敏感的限制性内切酶是MspJI。用修饰敏感的限制性内切酶(2812)消化样本产生在核酸的一端(2813)或5'和3'末端(2814)具有末端磷酸酯的靶向消耗的核酸。相反,未被修饰敏感的限制性酶切割的感兴趣的核酸在核酸的5'和/或3'末端没有暴露的末端磷酸酯(比较2810与2813-2814)。然后用核酸外切酶消化样本(2815, 消化步骤; 2816核酸外切酶),该酶使用靶向消耗的核酸中的末端磷酸酯以从核酸分子末端去除连续的核苷酸,从而从样本消耗靶向消耗的核酸。无需使用大小选择即可完成该消耗。在核酸外切酶消化后,将衔接子连接到感兴趣的核酸(2817),该核酸缺少末端磷酸酯,尚未被核酸外切酶消化。这产生在两端均衔接子连接的感兴趣的核酸(2818)。这些衔接子可用于下游应用,例如衔接子介导的PCR扩增,测序(例如高通量测序)和样本中靶核酸的量化和/或克隆。可替代地,将感兴趣的衔接子连接的核酸进行本文所述的一种或多种其他富集方法。例如,使衔接子连接的核酸经受本公开的另外的依赖修饰的富集方法(例如,图29中描绘的方法)。替代地或另外地,使衔接子连接的核酸经受本公开的基于核酸引导的核酸酶的富集方法(例如,图30中描绘的方法)。

[0583] 方案3:在图29中描绘在此描述的本申请的示例性方法。将包含感兴趣的核酸(2901)和靶向耗尽的核酸(2902)以及任选地归一化对照的核酸样本进行衔接子连接(2905),或使其经受本公开的富集方法(2906)(例如,图27或图28中描绘的方法,其产生感兴趣的衔接子连接的核酸(2907)和靶向耗尽的衔接子连接的核酸(2908)。在一些实施方式中,感兴趣的核酸和靶向耗尽的核酸都包含一个或多个修饰敏感的限制性酶的识别位点(分别为2903和2904)。在感兴趣的核酸中,修饰敏感的限制性内切酶的识别位点不包含修饰核苷酸(2903),或可替代地,比靶向耗尽的核酸的对应识别位点较不频繁地包含修饰核苷酸。在靶向耗尽的核酸中,修饰敏感的限制性内切酶的识别位点在限制位点内或附近包含修饰核苷酸(2904),或可替代地,比感兴趣的核酸的对应识别位点更频繁地包含修饰核苷酸。当在识别位点(2908)内或附近存在一个或多个修饰核苷酸时,修饰敏感的限制性内切酶(2909)切割其同源识别位点,而当识别位点不包含一个或多个修饰核苷酸(2908)时,

则不切割其同源识别位点,从而将修饰敏感的限制性酶的活性靶向到靶向消耗的核酸(比较2910与2911)。在一些实施方式中,修饰敏感的限制性酶包括AbaSI, FspEI, LpnPI, MspJI或McrBC。在一些实施方式中,修饰敏感的限制性内切酶是FspEI。在一些实施方式中,修饰敏感的限制性内切酶是MspJI。样本用修饰敏感的限制性内切酶(2911)消化,产生未进行衔接子连接的靶向消耗的核酸(2912)或仅在一端进行衔接子连接(2913)。通过从靶向消耗的核酸中选择性地去除衔接子,从而消耗靶向消耗的核酸。无需使用大小选择即可完成该消耗。相反,未被修饰敏感的限制性酶切割的感兴趣的核酸在两端均被衔接子连接(对比2910与2912-2913)。这些衔接子可用于下游应用,例如衔接子介导的PCR扩增,测序(例如高通量测序)和样本中感兴趣的核酸的量化和/或克隆。

[0584] 方案4:在图30中描述本文所述的示例性方法。多个gNA(3001)用于将核酸引导的核酸酶(3002)靶向到衔接子连接的核酸样本中的靶向消耗的核酸(3003)。通过本文所述的任何富集方法均可产生衔接子连接的核酸,该方法在初始衔接子连接之前或之后,使用修饰敏感的限制性内切酶从样本中消耗靶向消耗的核酸。在这种方法中,gNAs专门靶向到靶向耗尽的核酸(3003),而不是感兴趣的核酸(3004),因此不会被核酸引导的核酸酶(3002)切割。被核酸引导的核酸酶的切割导致靶向耗尽的核酸在一端被衔接子连接(3005),并且感兴趣的核酸在两端被衔接子连接(3003)。这些衔接子可用于下游应用,例如衔接子介导的PCR扩增,测序(例如高通量测序),样本中感兴趣的核酸的量化和克隆。

[0585] 本文所述的任何方法均可用作独立方法以从样本中消耗靶向耗尽的核酸,从而富集靶核酸。可替代地,与单独的任何单个方法相比,本文所述的方法可以组合以获得更大程度的富集。

[0586] 尽管本文描述方法的特定组合以及方法的组合的顺序,这些绝不旨在限制可以组合本公开的方法的方式。富集产生作为该方法的产物的感兴趣的衔接子连接的核酸的本公开的感兴趣的核酸的样本的任何方法可以与使用衔接子连接的核酸作为其起始底物的本公开的任何其他方法结合。

[0587] 基于核酸引导的核酸酶的富集方法

[0588] 本公开提供可与归一化对照和多分析物对照结合的用于富集靶序列的基于核酸引导的核酸酶的方法,以及任选地本文所述的其他富集或消耗方法。基于核酸引导的核酸酶的富集方法是采用核酸引导的核酸酶以富集感兴趣的序列的样本的方法。在W0/2016/100955、W0/2017/031360、W0/2017/100343、W0/2017/147345和W0/2018/227025中描述基于核酸引导的核酸酶的富集方法,其内容各自通过引用整体并入本文。

[0589] 可以在对样本进行基于核酸引导的核酸酶的方法之前或之后,将归一化和任选的多分析物对照加到样本中。在一些实施方式中,归一化对照包括与本文所述的gNA的靶向序列的序列不同的序列。

[0590] 术语“核酸引导的核酸酶-gNA复合物”是指包含核酸引导的核酸酶蛋白和引导核酸(gNA,例如gRNA或gDNA)的复合物。例如,“Cas9-gRNA复合物”是指包含Cas9蛋白和引导RNA(gRNA)的复合物。核酸引导的核酸酶可以是任何类型的核酸引导的核酸酶,包括但不限于野生型核酸引导的核酸酶,催化死亡的核酸引导的核酸酶或核酸引导的核酸酶-切口酶。当核酸引导的核酸酶是CRISPR/Cas核酸引导的核酸酶时,该复合物可以称为“CRISPR/Cas系统蛋白-gNA复合物”。

[0591] 本公开的方法可以利用核酸引导的核酸酶。如本文所用,“核酸引导的核酸酶”是切割DNA、RNA或DNA/RNA杂交体并且使用一种或多种引导核酸(gNA)赋予特异性的任何核酸酶。核酸引导的核酸酶包括CRISPR/Cas系统蛋白以及非CRISPR/Cas系统蛋白。本文提供的核酸引导的核酸酶可以是DNA引导的DNA核酸酶;例如,DNA引导的RNA核酸酶;RNA引导的DNA核酸酶;或RNA引导的RNA核酸酶。核酸酶可以是核酸内切酶。核酸酶可以是核酸外切酶。在一个实施方式中,核酸引导的核酸酶是核酸引导的DNA核酸内切酶。在一个实施方式中,核酸引导的核酸酶是核酸引导的RNA核酸内切酶。

[0592] 在一些实施方式中,本公开的基于修饰的富集方法和基于核酸引导的核酸酶的富集方法消耗样本中的不同核酸,从而与单独的任一方法相比,实现所感兴趣的核酸更大程度的富集。

[0593] 本文提供多个引导核酸(gNA)(可互换地称为文库或集合)。在一些实施方式中,归一化对照不包含与多个gNA中的任何gNA的序列相同或高度类似的序列。

[0594] 术语“引导核酸”是指能够与核酸引导的核酸酶和任选的其他核酸形成复合物的引导核酸(gNA)。gNA可以作为分离的核酸或作为核酸引导的核酸酶-gNA复合物,例如Cas9-gRNA复合物的一部分存在。

[0595] 如本文所用,多个gNA表示包含至少102个独特gNA的gNA的混合物。在一些实施方式中,多个gNA包含至少102个独特gNA,至少103个独特gNA,至少104个独特gNA,至少105个独特gNA,至少106个独特gNA,至少107个独特gNA,至少108个独特gNA,至少109个独特gNA或至少1010个独特gNA。在一些实施方式中,gNA的集合包含总共至少10²个独特gNA,至少10³个独特gNA,至少10⁴个独特gNA或至少10⁵个独特gNA。

[0596] 在一些实施方式中,gNA的集合包含含有靶向序列的第一NA片段;和包含核酸引导的核酸酶系统(例如,CRISPR/Cas系统)蛋白结合序列的第二NA片段。在一些实施方式中,第一和第二片段为5'至3'顺序。在一些实施方式中,第一和第二片段为3'至5'顺序。

[0597] 在一些实施方式中,第一段的大小在多个gNA中是12-100bp,12-75bp,12-50bp,12-30bp,12-25bp,12-22bp,12-20bp,12-18bp,12-16bp,14-250bp,14-100bp,14-75bp,14-50bp,14-30bp,14-25bp,14-22bp,14-20bp,14-18bp,14-17bp,14-16bp,15-250bp,15-100bp,15-75bp,15-50bp,15-30bp,15-25bp,15-22bp,15-20bp,15-18bp,15-17bp,15-16bp,16-250bp,16-100bp,16-75bp,16-50bp,16-30bp,16-25bp,16-22bp,16-20bp,16-18bp,16-17bp,17-250bp,17-100bp,17-75bp,或17-50bp,17-30bp,17-25bp,17-22bp,17-20bp,17-18bp,18-250bp,18-100bp,18-75bp,18-50bp,18-30bp,18-25bp,18-22bp,18-20bp,19-250bp,19-100bp,19-75bp,或19-50bp,19-30bp,19-25bp,或19-22bp。在一些特定的实施方式中,第一片段的大小是15bp,16bp,17bp,18bp,19bp或20bp。

[0598] 在一些实施方式中,多个中的至少10%,或至少15%,或至少20%,或至少25%,或至少30%,或至少35%,或至少40%,或至少45%,或至少50%,或至少55%,或至少60%,或至少65%,或至少70%,或至少75%,或至少80%,或至少85%,或至少90%,或至少95%或100%的第一片段是15-50bp。

[0599] 在一些实施方式中,多个gNA包含靶向序列,所述靶向序列可与靶向消耗的核酸中的靶序列碱基配对,其中靶向消耗的核酸中的靶序列在样本中的靶向消耗的基因组或转录组中至少每1bp,至少每2bp,至少3bp,至少每4bp,至少每5bp,至少每6bp,至少每7bp,至少

每8bp,至少每9bp,至少每10bp,至少每11bp,至少每12bp,至少每13bp,至少每14bp,至少每15bp,至少每16bp,至少每17bp,至少每18bp,至少每19bp,20bp,至少每25bp,至少每30bp,至少每40bp,至少每50bp,至少每100bp,至少每200bp,至少每300bp,至少每400bp,至少每500bp,至少每600bp,至少每700bp,至少每800bp,至少每900bp,至少每1000bp,至少每2500bp,至少每5000bp,至少每10,000bp,至少每15,000bp,至少每20,000bp,至少每25,000bp,至少每50,000bp,至少每100,000bp,至少每250,000bp,至少每500,000bp,至少每750,000bp,或甚至至少每1,000,000bp被间隔开。

[0600] 在一些实施方式中,多个gNAs包含含有靶向序列的第一NA片段;和包含核酸引导的核酸酶系统(例如,CRISPR/Cas系统)蛋白结合序列的第二NA片段,其中多个中的gNA可以具有多种第二NA片段,这些片段对核酸引导的核酸酶系统(例如,CRISPR/Cas系统)的蛋白质成员具有不同的特异性。例如,本文所提供gNAs的集合可以包括其第二片段包含对第一核酸引导的核酸酶系统(例如,CRISPR/Cas系统)蛋白特异的核酸引导的核酸酶系统(例如,CRISPR/CAS系统)蛋白结合序列的成员;并且还包含其第二片段包含对第二核酸引导的核酸酶系统(例如,CRISPR/CAS系统)蛋白特异的核酸引导的核酸酶系统(例如,CRISPR/CAS系统)蛋白结合序列的成员,其中所述第一第二核酸引导的核酸酶系统蛋白(例如CRISPR/Cas系统)并不相同。在一些实施方式中,本文提供的多个gNA包含对Cas9蛋白和选自以下的另一种蛋白表现出特异性的成员:Cpf1,Cas3,Cas8a-c,Cas10,CasX,CasY,Cas13,Cas14,Cse1,Csy1,Csn2,Cas4,Csm2和Cm5。包含靶向序列的第一NA片段和包含核酸引导的核酸酶系统蛋白结合序列的第二NA片段将取决于核酸引导的核酸酶系统蛋白。第一和第二NA片段的合适的5'至3'排列以及核酸引导的核酸酶系统蛋白的选择对于本领域普通技术人员而言将是显而易见的。

[0601] 在一些实施方式中,gNA包含DNA和RNA。在一些实施方式中,gNA由DNA(gDNA)组成。在一些实施方式中,gNA由RNA(gRNA)组成。

[0602] 在一些实施方式中,gNA包含gRNA,并且gRNA包含两个编码crRNA和tracrRNA的子段。在一些实施方式中,crRNA不包含靶向序列加上可以与tracrRNA杂交的额外序列。在一些实施方式中,crRNA包含可以与tracrRNA杂交的额外序列。在一些实施方式中,两个子段被独立地转录。在一些实施方式中,两个子段被转录为单个单元。在一些实施方式中,编码crRNA的DNA包含序列GTTTTAGAGCTATGCTGTTTTG(SEQ ID NO:1)的靶向序列5'。在一些实施方式中,编码tracrRNA的DNA包含序列GGAACCATTCAAAACAGCATAGCAAGTTAAAATAAGGCTAGTCCGTTATCAACTTGAAAAAGTGGCACCGAGTCGGTCTTTTTTTTTT(SEQ ID NO:2)。

[0603] 如本文所用,靶向序列是将gNA引导至在样本中靶向消耗的核酸中的靶序列的靶向序列。例如,靶向序列靶向本文所述的任何非宿主序列。

[0604] 本文提供包含片段的gNA和多个gNA,所述片段包含靶向序列。

[0605] 在一些实施方式中,靶向序列包含DNA或由DNA组成。在一些实施方式中,靶向序列包含RNA或由RNA组成。

[0606] 在一些实施方式中,靶向序列包含RNA,并且与感兴趣的序列上的PAM序列的序列5'共有至少70%序列同一性,至少75%序列同一性,至少80%序列同一性,至少85%序列同一性,至少90%序列同一性,至少95%序列同一性,或具有100%序列同一性,除了RNA包含尿嘧啶而不是胸腺嘧啶外。在一些实施方式中,靶向序列包含RNA,并且与感兴趣的序列上

的PAM序列的序列3'共有至少70%序列同一性,至少75%序列同一性,至少80%序列同一性,至少85%序列同一性,至少90%序列同一性,至少95%序列同一性,或共有100%序列同一性,除了RNA包含尿嘧啶而不是胸腺嘧啶外。在一些实施方式中,PAM序列是AGG、CGG、TGG、GGG或NAG。在一些实施方式中,PAM序列是TTN、TCN或TGN。

[0607] 在一些实施方式中,靶向序列包含RNA并且与PAM序列的核苷酸5'的核苷酸序列相反的链互补。在一些实施方式中,靶向序列与PAM序列5'核苷酸序列相反的链至少70%互补,至少75%互补,至少80%互补,至少85%互补,至少90%互补,至少95%互补或100%互补。在一些实施方式中,靶向序列包含RNA并且与PAM序列的核苷酸3'的序列相反的链互补。在一些实施方式中,靶向序列与PAM序列的核苷酸3'序列相反的链至少70%互补、至少75%互补、至少80%互补、至少85%互补、至少90%互补、至少95%互补或100%互补。在一些实施方式中,PAM序列是AGG、CGG、TGG、GGG或NAG。在一些实施方式中,PAM序列是TTN、TCN或TGN。

[0608] 不同的CRISPR/Cas系统蛋白识别不同的PAM序列。PAM序列可以位于靶向序列的5'或3'。例如,Cas9可以识别位于靶向序列的紧邻3'末端的NGG PAM。Cpf1可以识别位于靶向序列5'末端的TTN PAM。被所有CRISPR/Cas系统蛋白识别的所有PAM序列被设想为在本公开的范围内。对于本领域普通技术人员而言,显而易见的是,PAM序列与特定的CRISPR/Cas系统蛋白相容。

[0609] 本文提供gNA和多个gNA,其包含含有核酸引导的核酸酶蛋白质结合序列的片段。核酸引导的核酸酶可以是核酸引导的核酸酶系统蛋白(例如,CRISPR/Cas系统)。核酸引导的核酸酶系统可以是RNA引导的核酸酶系统。核酸引导的核酸酶系统可以是DNA引导的核酸酶系统。

[0610] 核酸引导的核酸酶蛋白质结合序列是与核酸引导的核酸酶系统的任何蛋白质成员结合的核酸序列。例如,CRISPR/Cas蛋白质结合序列是结合CRISPR/Cas系统的任何蛋白质成员的核酸序列。

[0611] 在一些实施方式中,CRISPR/Cas系统蛋白可以来自任何细菌或古细菌物种。在一些实施方式中,CRISPR/Cas系统蛋白是分离的,重组产生的或合成的。在一些实施方式中,CRISPR/Cas系统蛋白的示例可以是天然存在的或工程改造的版本。

[0612] 在一些实施方式中,核酸引导的核酸酶系统蛋白(例如,CRISPR/Cas系统蛋白)可以来自任何细菌或古细菌物种。在一些实施方式中,天然存在的CRISPR/Cas系统蛋白可以属于CAS I类I,III或IV型,或CAS II类II或V型,并且可以包括Cas9,Cas3,Cas8a-c,Cas10,CasX,CasY,Cas13,Cas14,Cse1,Csy1,Csn2,Cas4,Csm2,Cmr5,Csf1,C2c2,和Cpf1。在一个示例性实施方式中,CRISPR/Cas系统蛋白包含Cas9。在一个示例性实施方式中,CRISPR/Cas系统蛋白包含Cpf1。

[0613] 在一些实施方式中,核酸引导的核酸酶系统蛋白(例如,CRISPR/Cas系统蛋白)来自或衍生自化脓性链球菌,金黄色葡萄球菌,脑膜炎奈瑟氏球菌,嗜热链球菌,树突密螺旋体,土拉弗朗西斯菌,多杀性巴斯德氏菌,空肠弯曲杆菌,海鸥弯曲杆菌,鸡败血支原体,Nitratifractor salsuginis,Parvibaculum lavamentivorans,肠道罗斯拜瑞氏菌,灰质奈瑟氏球菌,重氮葡萄糖醋杆菌,固氮螺菌,Sphaerochaeta globus,柱状黄杆菌,流动型塔夫杆菌,嗜粪拟杆菌,运动支原体,法氏乳杆菌,巴氏链球菌,约氏乳杆菌,伪中间型葡萄球菌,

Filifactor alocis,嗜肺军团菌,Suterella wadsworthensis和白喉杆菌,氨基酸球菌,毛螺科菌或普雷沃氏菌的核酸引导的核酸酶系统蛋白(例如,CRISPR/Cas系统蛋白)。

[0614] 在一些实施方式中,核酸引导的核酸酶系统蛋白结合序列包含gNA(例如,gRNA)茎环序列。不同的CRISPR/Cas系统蛋白与不同的核酸引导的核酸酶系统蛋白结合序列兼容。对于本领域普通技术人员而言,显而易见的是,CRISPR/Cas系统蛋白与哪些核酸引导的核酸酶系统蛋白结合序列相容。

[0615] 在一些实施方式中,编码gNA(例如,gRNA)茎环序列的双链DNA序列在一条链上包含以下DNA序列(5'>3',GTTTATAGAGCTAGAAATAGCAAGTTAAAATAAGGCTAGTCCGTTATCAACTTGAAA AAGTGGCACCAGAGTCGGTGCTTTTTTTTTT)(SEQ ID NO:3),及其在另一链上的反向互补DNA(5'>3',AAAAAAGCACCAGACTCGGTGCCACTTTTTCAAGTTGATAACGGACTAGCCTTATTTTAACTTGCTATTTCTA GCTCTAAAAC)。

[0616] 在一些实施方式中,编码gNA(例如,gRNA)茎环序列的单链DNA序列包含以下DNA序列:(5'>3',AAAAAAGCACCAGACTCGGTGCCACTTTTTCAAGTTGATAACGGACTAGCCTTATTTTAACTTGCT ATTTCTAGCTCTAAAAC)(SEQ ID NO:4),其中单链DNA充当转录模板。

[0617] 在一些实施方式中,gNA(例如,gRNA)茎环序列包含以下RNA序列:(5'>3',GUUUU AGAGCUAGAAAUAGCGCGUUAUUAAAGGCUAGUCCGUUAUCAACUUGAAAAAGUGGCACCAGUCGGUGCUCUUU UUUUU)(SEQ ID NO:5)。

[0618] 在一些实施方式中,编码gNA(例如,gRNA)茎环序列的双链DNA序列在一条链上包含以下DNA序列(5'>3',GTTTATAGAGCTATGCTGGAAACAGCATAGCAAGTTAAAATAAGGCTAGTCCGTTAT CAACTTGAAAAAGTGGCACCAGAGTCGGTGCTTTTTTTC)(SEQ ID NO:6),及其在另一链上的反向互补 DNA(5'>3',GAAAAAAGCACCAGACTCGGTGCCACTTTTTCAAGTTGATAACGGACTAGCCTTATTTTAACTTGC TATGCTGTTTCCAGCATAGCTCTAAAAC)。

[0619] 在一些实施方式中,编码gNA(例如,gRNA)茎环序列的单链DNA序列包含以下DNA序列:(5'>3',GAAAAAAGCACCAGACTCGGTGCCACTTTTTCAAGTTGATAACGGACTAGCCTTATTTTAACTTGC TATGCTGTTTCCAGCATAGCTCTAAAAC)(SEQ ID NO:7),其中单链DNA充当转录模板。

[0620] 在一些实施方式中,gNA(例如,gRNA)茎环序列包含以下RNA序列:(5'>3',GUUUUA GAGCUAUGCUGGAAACAGCAUAGCAAGUUAUUAAAGGCUAGUCCGUUAUCAACUUGAAAAAGUGGCGCCGAGUCG GUGUGCUUUUUUC)(SEQ ID NO:8)。

[0621] 在一些实施方式中,CRISPR/Cas系统蛋白是Cpf1蛋白。在一些实施方式中,所述 Cpf1蛋白分离或衍生自Franciscella物种或Acidaminococcus物种。在一些实施方式中, gNA(例如,gRNA)CRISPR/Cas系统蛋白结合序列包含以下RNA序列:(5'>3', AAUUUCUACUGUUGUAGAU)(SEQ ID NO:9)。

[0622] 在一些实施方式中,CRISPR/Cas系统蛋白是Cpf1蛋白。在一些实施方式中,所述 Cpf1蛋白分离或衍生自Franciscella物种或Acidaminococcus物种。在一些实施方式中,编 码gNA(例如,gRNA)CRISPR/Cas系统蛋白结合序列的DNA序列包含以下DNA序列:(5'>3', AATTTCTACTGTTGTAGAT)(SEQ ID NO:10)。在一些实施方式中,DNA是单链的。在一些实施方 式中,DNA是双链的。

[0623] 在一些实施方式中,本文提供包含第一NA片段和第二NA片段的gNA(例如,gNA),所 述第一NA片段包含靶向序列,所述第二NA片段包含核酸引导的核酸酶(例如CRISPR/Cas)系

统蛋白结合序列。在一些实施方式中,第一片段的大小是15bp,16bp,17bp,18bp,19bp或20bp。在一些实施方式中,第二片段包含单个片段,该单个片段包含gRNA茎环序列。在一些实施方式中,gRNA茎环序列包含以下RNA序列:(5'>3',GUUUUAGAGCUAGAAAUAGCAAGUUA AAUAAGGCUAGUCCGUUAUCAACUUGAAAAAGUGGCACCGAGUCGGUGCUUUUUUUU) (SEQ ID NO:5)。在一些实施方式中,gRNA茎环序列包含以下RNA序列:(5'>3',GUUUUAGAGCUAUGCUGGAAACAGCAUA GCAAGUUAUUAAAGGCUAGUCCGUUAUCAACUUGAAAAAGUGGCACCGAGUCGGUGCUCUUUUUUUUC) (SEQ ID NO:8)。在一些实施方式中,第二片段包含两个子段:与第二RNA子段(tracrRNA)形成杂合体的第一RNA子段(crRNA),其共同作用以指导核酸引导的核酸酶(例如,CRISPR/Cas)系统蛋白结合。在一些实施方式中,第二子段的序列包含GUUUUAGAGCUAUGCUGUUUUG。在一些实施方式中,第一RNA片段和第二RNA片段一起形成crRNA序列。在一些实施方式中,将与第二RNA片段形成杂交体的其他RNA是tracrRNA。在一些实施方式中,tracrRNA包含5'>3'的序列,GGAACCAUUCAAAACAGCAUAGCAAGUUAUUAAAGAGUAUAGCUAGUCCGUUAUCAACUUGAAAAAGUGGCACCGAGUCGGUGCUUUUUUUU (SEQ ID NO:11)。

[0624] 在一些实施方式中,本文提供包含第一NA片段和第二NA片段的gNA(例如,gNA),所述第一NA片段包含靶向序列,所述第二NA片段包含核酸引导的核酸酶(例如CRISPR/Cas)系统蛋白结合序列。在一些实施方式中,例如其中CRISPR/Cas系统蛋白是Cpf1系统蛋白的那些实施方式,第二片段是第一片段的5'。在一些实施方式中,第一片段的大小是20bp。在一些实施方式中,第一片段的大小大于20bp。在一些实施方式中,第一片段的大小大于30bp。在一些实施方式中,第二片段包含单个片段,该单个片段包含gRNA茎环序列。在一些实施方式中,gRNA茎环序列包含以下RNA序列:(5'>3',AAUUUCUACUGUUGUAGAU) (SEQ ID NO:9)。

[0625] CRISPR/Cas系统蛋白可以与野生型CRISPR/Cas系统蛋白至少60%相同(例如,至少70%,至少80%或90%相同,至少95%相同或至少98%相同或至少99%相同)。CRISPR/Cas系统蛋白可能具有野生型CRISPR/Cas系统蛋白的所有功能,或只有一种或一些功能,包括结合活性和核酸酶活性。

[0626] 术语“CRISPR/Cas系统蛋白相关的引导NA”是指引导NA(gNA)。CRISPR/Cas系统蛋白相关的引导NA可以作为分离的NA存在,或作为CRISPR/Cas系统蛋白-gNA复合体的一部分存在。

[0627] 在一些实施方式,CRISPR/Cas系统蛋白是RNA引导的DNA核酸酶。在一些实施方式中,被CRISPR/Cas系统蛋白切割的DNA是双链的。切割双链DNA的示例性RNA引导的DNA核酸酶包括但不限于Cas9,Cpf1,CasX和CasY。其他示例性RNA引导的DNA核酸酶包括Cas10,Csm2,Csm3,Csm4和Csm5。在一些实施方式中,Cas10,Csm2,Csm3,Csm4和Csm5与gRNA形成核糖核蛋白复合物。

[0628] 在一些实施方式中,CRISPR/Cas系统蛋白核酸引导的核酸酶是或包含Cas9。本公开的Cas9可以是分离的,重组产生的或合成的。在一些实施方式中,Cas9蛋白是热稳定的。可在本文的实施方式中使用的Cas9蛋白的一些示例可以在FA Ran,L.Cong,WX Yan,DA Scott,JS Gootenberg,AJ Kriz,B.Zetsche,O.Shalem,X.Wu,KS Makarova,EV Koonin,PA Sharp和F.Zhang;“使用金黄色葡萄球菌Cas9进行体内基因组编辑”,《自然》520,186-191(2015年4月9日)doi:10.1038/nature14299中找到,在此通过引用并入本文。在一些实施方式中,Cas9是衍生自化脓性链球菌,金黄色葡萄球菌,脑膜炎奈瑟氏球菌,嗜热链球菌,树突

密螺旋体,土拉弗朗西斯菌,多杀性巴斯德氏菌,空肠弯曲杆菌,海鸥弯曲杆菌,唾液分枝杆菌,唾液支原体,沙门氏菌,食管支原体,鸡败血支原体,Nitratifractor salsuginis, Parvibaculum lavamentivorans,肠道罗斯拜瑞氏菌,灰质奈瑟氏球菌,重氮葡萄糖醋杆菌,固氮螺菌,Sphaerochaeta globus,柱状黄杆菌,流动型塔夫杆菌,嗜粪拟杆菌,运动支原体,法氏乳杆菌,巴氏链球菌,约氏乳杆菌,伪中间型葡萄球菌,Filifactor alocis,嗜肺军团菌,Suterella wadsworthensis和白喉杆菌的II型CRISPR系统。

[0629] 在一些实施方式中,Cas9是衍生自化脓性链球菌的II型CRISPR系统,并且PAM序列是NGG,其位于靶特异性引导序列的紧邻3'末端。来自示例性细菌物种的II型CRISPR系统的PAM序列还可以包括:化脓性链球菌(NGG),金黄色葡萄球菌(NNGRRT),脑膜炎奈瑟氏球菌(NNNGATT),嗜热链球菌(NNAGAA)和树突密螺旋体(NAAAAC),这些在不脱离本公开的情况下都是可用的。

[0630] 在一个示例性实施方式中,Cas9序列可以例如从pX330质粒(可得自Addgene)获得,通过PCR再扩增,然后克隆至pET30(得自EMD biosciences)以在细菌中表达并纯化重组的6His标记的蛋白。

[0631] “Cas9-gNA复合物”是指包含Cas9蛋白和引导NA的复合物。Cas9蛋白可以与野生型Cas9蛋白例如化脓链球菌Cas9蛋白至少60%相同(例如,至少70%,至少80%或90%相同,至少95%相同或至少98%相同或至少99%相同)。Cas9蛋白可能具有野生型Cas9蛋白的所有功能,或只有一种或一些功能,包括结合活性、核酸酶活性和核酸酶活性。

[0632] 术语“Cas9相关的引导的NA”是指如上所述的引导的NA。与Cas9相关的引导的NA可以单独存在,或作为Cas9-gNA复合体的一部分存在。

[0633] 在一些实施方式中,CRISPR/Cas系统蛋白核酸引导的核酸酶是或包含Cpf1系统蛋白。本公开的Cpf1系统蛋白可以是分离的、重组产生的或合成的。在一些实施方式中,Cpf1蛋白是热稳定的。

[0634] Cpf1系统蛋白是II类、V型CRISPR系统蛋白。在一些实施方式中,Cpf1蛋白分离或衍生自土拉弗朗西斯菌。在一些实施方式中,Cpf1蛋白分离或衍生自酸氨基球菌、乳酸杆菌科细菌或普雷沃氏菌。

[0635] Cpf1系统蛋白质结合到包含核酸引导的核酸酶系统蛋白-结合序列(例如,茎环)和靶向序列。Cpf1靶向序列包含紧邻定位在靶核酸中的Cpf1 PAM序列的3'的序列。与Cas9不同,Cpf1核酸引导的核酸酶系统蛋白-结合序列位于Cpf1 gRNA中靶向序列的5'。Cpf1还可以在靶核酸中产生交错而不是平末端的切割。将Cpf1蛋白-gRNA蛋白复合物靶向靶核酸后,弗朗西斯菌衍生的Cpf1,例如以交错形式切割靶核酸离子,在靶向序列的3'端距PAM 18-23个碱基形成约5个核苷酸的5'突出端。相比之下,被野生型Cas9切割在Cas9 PAM的上游3个核苷酸处产生平末端。

[0636] 示例性Cpf1 gRNA茎环序列包含以下RNA序列:(5' > 3', AAUUUCUACUGUUGUAGAU)(SEQ ID NO:9)。

[0637] “Cpf1蛋白-gNA复合物”是指包含Cpf1蛋白和引导NA(例如,gRNA)的复合物。当gNA是gRNA时,gRNA可以由单个分子,即与靶杂交并提供序列特异性的一个RNA(“crRNA”)组成。

[0638] Cpf1蛋白与野生型Cpf1蛋白可以具有至少60%同一性(例如,至少70%,至少80%或90%同一性,至少95%同一性或至少98%同一性或至少99%同一性)。Cpf1蛋白可能具有

野生型Cpf1蛋白的所有功能,或只有一种或一些功能,包括结合活性和核酸酶活性。

[0639] Cpf1系统蛋白可识别多种PAM序列。被Cpf1系统蛋白识别的示例性PAM序列包括但不限于TTN,TCN和TGN。另外的Cpf1 PAM序列包括但不限于TTTN。Cpf1 PAM序列的一个特征是,它们比被Cas9蛋白使用的NGG或NAG PAM序列具有更高的A/T含量。

[0640] 计算机系统和软件

[0641] 本文描述的方法可以在计算机系统的上下文中使用,或用作存储在计算机可读存储介质中的软件或计算机可执行指令的一部分。

[0642] 在一些实施方式中,系统(例如,计算机系统)可以用于实现本发明的一些实施方式的某些特征。

[0643] 在一些实施方式中,该系统可以包括一个或多个存储器和/或存储设备。存储器和存储设备可以是一种或多种计算机可读存储介质,所述计算机可读存储介质可以存储实现本发明的各种实施方式的至少一部分的计算机可执行指令。

[0644] 本文提供一种用于设计多个归一化对照多核苷酸序列的系统。在一些实施方式中,该系统包括计算机可读存储介质,该计算机可读存储介质存储计算机可执行指令,该计算机可执行指令包括:(i)用于导入至少一个参考序列的指令;(ii)从至少一个参考序列产生多个参考序列片段的指令;(iii)根据参考序列片段数目产生至少一个参数的分布的指令;(iv)将分布划分成箱的指令;(v)从箱的至少一个子集中选择多个参考序列片段的指令;和(vi)改组多个参考序列片段以产生改组序列的指令;从而产生多个归一化多核苷酸序列。在一些实施方式中,该系统包括计算机可读存储介质,该计算机可读存储介质存储计算机可执行指令,该计算机可执行指令包括:(i)用于导入至少一个参考序列的指令;(ii)从至少一个参考序列产生多个参考序列片段的指令;(iii)根据参考序列片段数目产生至少一个参数的分布的指令;(iv)将分布划分成箱的指令;(v)从箱的至少一个子集中的每一个中选择至少两个参考序列片段的指令,其中所述至少两个参考序列片段在参考序列中是不连续的,或来自不同的参考序列;(vi)连接来自每个箱的至少两个参考序列片段的指令;从而产生多个归一化多核苷酸序列。在一些实施方式中,该系统还包括处理器,该处理器被配置成执行包括以下步骤的步骤:(a)接收包括至少一个参考序列的一组输入文件;和(b)执行存储在计算机可读存储介质中的计算机可执行指令。在一些实施方式中,该参数包括以下各项中的至少一项:(1)GC含量百分比、(2)熵、(3)复杂性、(4)EIIP、(5)长度,或它们的组合。

[0645] 该计算系统可以包括一个或多个连接到互连的中央处理单元(“处理器”),存储器,输入/输出设备(例如键盘和指点设备),触摸设备,显示设备,存储设备(例如磁盘驱动器)和网络衔接子(例如网络接口)。互连是表示通过适当的桥、衔接子或控制器连接的一个或多个单独的物理总线、点对点连接或两者的抽象概念。因此,互连可以包括例如系统总线,外围组件互连(PCI)总线或PCI-Express总线,HyperTransport或行业标准结构(ISA)总线,小型计算机系统接口(SCSI)总线,通用串行总线(USB),IIC(12C)总线或电气与电子工程师协会(IEEE)标准1394总线,也称为Firewire。

[0646] 另外,数据结构和消息结构可以经由数据传输介质(例如,通信链路上的信号)来存储或传输。可以使用各种通信链路,例如Internet、局域网、广域网或点对点拨号连接。因此,计算机可读介质可以包括计算机可读存储介质,例如非暂时性介质,以及计算机可读传

输介质。

[0647] 可以将存储在存储器中的指令实现为软件和/或固件以对一个或多个处理器进行编程以执行上述动作。在本发明的一些实施方式中,可以通过例如经由网络衔接子通过计算机系统从远程系统下载这些软件或固件来将其最初提供给处理系统。

[0648] 本文介绍的本发明的各种实施方式可以通过例如可编程电路(例如一个或多个微处理器)来实现,该可编程电路完全在专用的硬连线即非可编程电路中,或在这种形式的组合中用软件和/或固件进行编程。专用硬连线电路可以采用例如一个或多个ASIC、PLD、FPGA等形式。

[0649] 可以根据算法来呈现详细描述的一些部分,该算法可以是对计算机存储器内的数据位的操作的符号表示。这些算法描述和表示是数据处理领域的技术人员用来最有效地将其工作的实质传达给本领域其他技术人员的那些方法。这里,算法通常被认为是导致所需结果的自洽操作序列。这些操作是需要对物理量进行物理操纵的操作。通常,尽管不是必须的,这些量采取能够被存储、传输、组合、比较和以其他方式操纵的电或磁信号的形式。主要出于通用的原因,有时已经证明将这些信号称为位,值,元件,符号,字符,项,数字等是方便的。

[0650] 本文提出的算法和显示与任何特定的计算机或其他设备不是固有地相关。各种通用系统可以与根据本文的教导的程序一起使用,或可以证明构造更专用的装置以执行一些实施方式的方法是方便的。

[0651] 此外,尽管已经在功能全面的计算机和计算机系统的上下文中描述实施方式,本领域技术人员将理解,各种实施方式能够以各种形式作为程序产品来分发,并且本公开同样适用,这与实际影响分发的机器或计算机可读介质的特定类型无关。

[0652] 机器可读存储介质、机器可读介质或计算机可读(存储)介质的其他示例包括但不限于可记录类型的介质,例如易失性和非易失性存储设备,软盘和其他可移动磁盘,硬盘驱动器,光盘(例如光盘只读存储器(CD ROMS),数字通用磁盘,(DVD)等)以及传输类型的介质(例如,数字和模拟通信链路)。

[0653] 试剂盒和制品

[0654] 本文提供的试剂盒和制品包括本文所述的任何一种或多种归一化对照组合物和/或用于制备或使用本文所述的任何归一化对照组合物的试剂。在一些实施方式中,示例性试剂盒包含本文所述的DNA归一化对照组合物。在一些实施方式中,示例性试剂盒包含本文所述的RNA归一化对照组合物。在一些实施方式中,示例性试剂盒包含本文所述的DNA和RNA归一化对照组合物的组合。在一些实施方式中,试剂盒进一步包括使用归一化对照组合物的指令,例如稀释、浓度以及将归一化对照加到样本的指令。

[0655] 在本文描述的试剂盒的一些实施方式中,试剂盒包含本公开的多分析物对照中的任何一种或多种,以及用于其使用的试剂。在一些实施方式中,该试剂盒还包括用于使用多分析物对照的指令,例如用于稀释、浓缩以及将多分析物对照加到样本,或与样本平行地处理多分析物对照的指令。本公开的试剂盒提供的示例性多分析物对照包括包含至少三种已经灭活的生物的多分析物对照。在一些实施方式中,试剂盒还包含用于稀释多分析物对照的试剂。在一些实施方式中,试剂盒进一步包括使用多分析物对照的指令。

[0656] 在一些实施方式中,试剂盒包含归一化对照,多分析物对照以及试剂和使用它们

的指令。在一些实施方式中,试剂盒进一步包含阳性对照。示例性阳性对照包括可以用于阳性对照或校准本文所述的归一化对照和多分析物对照的已知浓度的灭活生物(例如1IU/mL,10IU/mL,100IU/mL,1,000IU/mL,10,000IU/mL,100,000IU/mL或1,000,000IU/mL)。在一些实施方式中,阳性对照包括病毒,细菌,真菌,单细胞真核生物或其任何组合。

[0657] 如本文所述,本文提供的试剂盒和制品包含任何一种或多种用于消耗靶向耗尽的序列的样本的试剂或用于在NGS测序之前富集感兴趣的序列的样本的试剂。在一些实施方式中,试剂盒包含任何一种或多种本文所述的组合物,不限于衔接子,gNAs(例如,gRNAs或gDNAs),gNA集合(例如,gRNA或gDNA的多个),改性敏感的限制性酶,对照和类似物。

[0658] 在示例性实施方式中,试剂盒包含gRNA,其中gRNA靶向本文所述的任何宿主基因组或cDNA序列。在进一步的示例性实施方式中,试剂盒包含gRNA,其中gRNA靶向人基因组或DNA序列的其他来源。

[0659] 如本文所述,本公开还提供用于进行使用核苷酸修饰的差异来富集感兴趣的核酸的样本的方法的所有必需试剂和指令。

[0660] 本公开还提供试剂盒,所述试剂盒包含归一化对照,多分析物对照,用于消耗靶向耗尽的序列的样本的试剂或用于在NGS测序之前富集感兴趣的序列的样本的试剂,以及使用它们的试剂和指令。

[0661] 本文还提供在使用本文提供的方法富集样本之前和之后监测信息的计算机软件。在一个示例性实施方式中,该软件可以在应用本文描述的方法之前和之后计算并报告样本中靶向耗尽的核酸序列的丰度,以评估脱靶消耗水平,并且其中该软件可以通过在使用本文提供的富集方法处理样本之前和之后比较感兴趣的序列的丰度,检查靶向耗尽/富集/捕获/分区/标记/调节/编辑的功效。

[0662] 在本公开的试剂盒的一些实施方式中,试剂盒包含归一化对照,多分析物对照,用于样本富集和/或消耗的试剂,使用它们的指令,以及使用它们例如可接受的稀释剂的试剂。

[0663] 在本公开的试剂盒的一些实施方式中,多分析物对照包含11种病毒物种的混合物或由11种病毒物种的混合物组成。在一些实施方式中,病毒由巨细胞病毒(CMV),爱泼斯坦-巴尔病毒(EBV),腺病毒(ADV),BK病毒(BKV),JC多瘤病毒(JCV),人疱疹病毒6A(HHV6A),人疱疹病毒6B(HHV6B)组成,1型单纯疱疹病毒(HSV1),2型单纯疱疹病毒(HSV2),水痘带状疱疹病毒(VZV)和人细小病毒B19(B19)组成。

[0664] 在本公开的试剂盒的一些实施方式中,多分析物对照包含10种病毒物种的混合物或由10种病毒物种的混合物组成。在一些实施方式中,病毒由CMV,EBV,ADV,BKV,JCV,HHV6A,HHV6B,HSV1,HSV2和VZV组成。

[0665] 贯穿本公开给出的每个最大值限制包括每个较低的值限制,就像在本文中明确写出这些较低的值限制一样。贯穿本公开给出的每个最小值限制将包括每个更高的值限制,就像在本文中明确写出这些更高的值限制一样。贯穿本公开给出的每个数值范围将包括落入该较宽的数值范围内的每个较窄的数值范围,就好像这些较窄的数值范围均在本文中明确写出一样。

[0666] 本文公开的值不应将其理解为严格限于所列举的确切数值。相反,除非另有说明,否则每个这样的值旨在表示所列举的值和围绕该值的功能上等效的范围。例如,公开为“20

μm”的值旨在表示“约20μm”。

[0667] 除非明确排除或以其他方式限制,否则本文引用的每个文件,包括任何交叉引用的或相关的专利或申请,均通过引用整体并入本文。引用任何文件并不意味着承认本文是本文公开或要求保护的发明的现有技术,或单独或与任何其他参考文献组合,教导,建议或公开任何这样的发明。此外,在本文档中术语的任何含义或定义与通过引用并入的文档中相同术语的任何含义或定义冲突的程度上,以在本文档中分配给该术语的含义或定义为准。

[0668] 虽然已经图示和描述本公开的特定实施方式,在不脱离本公开的精神和范围的情况下,可以做出各种其他改变和修改。所附权利要求的范围包括在本公开的范围内所有的这样的改变和修改。

[0669] 列举的实施方式

[0670] 可以通过参考以下列举的说明性实施方式来定义本发明:

[0671] 1.一种归一化对照,所述归一化对照包括至少三组多核苷酸,其中每个组内的多核苷酸具有相同的长度。

[0672] 2.实施方式1的归一化对照,其中所述归一化对照中的所有多核苷酸的长度是相同的。

[0673] 3.实施方式1的归一化对照,其中当与任何其他组内的多核苷酸相比时,每组中的多核苷酸具有不同的长度。

[0674] 4.实施方式3的归一化对照,其中所述多核苷酸组的长度以线性序列或几何序列分布。

[0675] 5.实施方式1-4中任一项的归一化对照,其中所述至少三个组中的多核苷酸的长度在约15bp至约50,000bp之间。

[0676] 6.实施方式1-4中任一项的归一化对照,其中所述至少三个组中的多核苷酸的长度在约500bp至约1500bp之间。

[0677] 7.实施方式1-4中任一项的归一化对照,其中所述至少三个组中的多核苷酸的长度在约100bp至约1200bp之间。

[0678] 8.实施方式1-4中任一项的归一化对照,其中所述至少三个组中的多核苷酸的长度在约150bp至约600bp之间。

[0679] 9.实施方式1-4中任一项的归一化对照,其中所述至少三个组包含具有选自以下的长度的多核苷酸:175bp,250bp和450bp;192bp,250bp和450bp;200bp,300bp和500bp;217bp,300bp和517bp;436bp,552bp和974bp;450bp,612bp和1034bp;510bp,626bp和1048bp;450bp,612bp和1034bp。

[0680] 10.实施方式1-9中任一项的归一化对照,其中每个组内的多核苷酸包含相同的序列。

[0681] 11.实施方式1-9中任一项的归一化对照,其中每个组内的多核苷酸不包含相同序列。

[0682] 12.实施方式11的归一化对照,其中每组多核苷酸包含至少三个多核苷酸亚组,

[0683] 其中每个亚组内的多核苷酸包含相同的序列,并且其中每个亚组内的多核苷酸不包含与任何其他亚组相同的序列。

- [0684] 13.实施方式10-12中任一项的归一化对照,其中所述至少一组多核苷酸中的至少一个序列包含组分序列。
- [0685] 14.实施方式10-12中任一项的归一化对照,其中每组多核苷酸中的每个序列均包含组分序列。
- [0686] 15.实施方式13或14的归一化对照,其中每组多核苷酸的组分序列与任何其他组的组分序列不同。
- [0687] 16.实施方式12的归一化对照,其中所述至少一组多核苷酸的至少一个亚组的序列包含组分序列。
- [0688] 17.实施方式12的归一化对照,其中所述至少一组的所述至少三个亚组中的每一个的序列包含组分序列。
- [0689] 18.实施方式12的归一化对照,其中每个组的每个亚组的序列包含组分序列。
- [0690] 19.实施方式17或18的归一化对照,其中每个亚组的组分序列与任何其他亚组的组分序列不同。
- [0691] 20.实施方式13-19中任一项的归一化对照,其中所述组分序列包含约6bp至约3000bp的序列。
- [0692] 21.实施方式13-20的归一化对照,其中所述组分序列包含随机序列。
- [0693] 22.实施方式10-13的归一化对照,其中所述至少一组多核苷酸的序列包含分离的序列。
- [0694] 23.实施方式10-12中任一项的归一化对照,其中每组多核苷酸的序列包含分离的序列。
- [0695] 24.实施方式12、13或15-17中任一项的归一化对照,其中所述至少一组多核苷酸的至少一个亚组的序列包含分离的序列。
- [0696] 25.实施方式12、13或15-17中任一项的归一化对照,其中所述至少一组的每个亚组的序列包含分离的序列。
- [0697] 26.实施方式12的归一化对照,其中每个组的每个亚组的序列包含分离的序列。
- [0698] 27.实施方式25或26中任一项的归一化对照,其中每个亚组的分离的序列与任何其他亚组的分离的序列不同。
- [0699] 28.实施方式22-26中任一项的归一化对照,其中所述分离的序列包含约6bp至约200,000bp,约15bp至约50,000bp,约500bp至约1500bp,约100bp至约1200bp,或约150bp至约600bp的序列。
- [0700] 29.根据实施方式22-28中任一项所述的归一化对照,其中所述分离的序列是分离的或来源于病毒、细菌、真菌或真核寄生虫。
- [0701] 30.实施方式29的归一化对照,其中所述病毒是T4噬菌体(T4)或巨细胞病毒(CMV)。
- [0702] 31.实施方式22-30中任一项的归一化对照,其中所述分离的序列是从质粒分离的。
- [0703] 32.实施方式22-31中任一项的归一化对照,其中所述分离的序列与测序样本中的至少一个靶序列不同。
- [0704] 33.实施方式1-32中任一项的归一化对照,其中归一化对照中的任何多核苷酸的

序列具有小于或等于与测序样本中的至少一种靶序列的1%,2%,3%,4%,5%,6%,7%,8%,9%,10%,11%,12%,13%,14%,15%,16%,17%,18%,19%,20%,30%,40%,50%,60%,70%,80%,85%,90%,95%,96%,97%,98%或99%同一性。

[0705] 34.实施方式12-33中任一项的归一化对照,其中所述归一化对照中的多核苷酸的至少一个亚组的至少一个序列与所述测序样本中的所述至少一个靶序列具有至少一种序列特性。

[0706] 35.实施方式34的归一化对照,其中所述至少一个序列特性包括GC含量百分比,熵,复杂性,长度,电子-离子相互作用电势(EIIP),转座元件序列的序列特性,病毒序列的序列特性,细菌序列的序列特性,真菌序列的序列特性,真核寄生虫序列,一个或多个人类基因序列的序列特性或其任何组合。

[0707] 36.实施方式35的归一化对照,其中病毒序列的特性包括GC含量百分比,重复序列,反向末端重复(ITR)序列,内部核糖体进入位点(IRES),蛋白质编码序列,转录后调控元件序列,转录调控元件,启动子序列,顺式作用RNA元件,RNA结构元件的序列,基因组包装信号,5'非翻译区(5'UTR)序列,3'非翻译区域(3'UTR)序列或其组合。

[0708] 37.实施方式35的归一化对照,其中细菌序列的特性包括GC含量百分比,重复序列,microRNA结合位点,内部核糖体进入位点(IRES),蛋白质编码序列,转录调控元件序列,启动子序列,5'UTR序列,3'UTR序列或其组合。

[0709] 38.实施方式37的归一化对照,其中所述重复序列元件包含polyA基序,polyT基序,polyG基序,polyC基序,二核苷酸基序,三核苷酸基序,四核苷酸基序,五核苷酸基序,六核苷酸基序,七核苷酸基序,八核苷酸基序,九核苷酸基序,散布的重复序列元件,核糖体RNA序列,tRNA序列或其组合。

[0710] 39.实施方式35的归一化对照,其中所述真菌序列的特性包括GC含量百分比,重复序列,microRNA结合位点,内部核糖体进入位点(IRES)序列,蛋白质编码序列,转录调节元件,启动子序列,5'UTR序列,3'UTR序列,着丝粒序列,端粒序列,亚端粒序列,线粒体序列或其组合。

[0711] 40.实施方式35的归一化对照,其中一个或多个人类基因的特性包括GC含量百分比,重复序列,蛋白质编码序列,内含子序列,5'UTR序列,3'UTR序列,转录调控元件,启动子序列,microRNA结合位点或其组合。

[0712] 41.实施方式34-40中任一项的归一化对照,其中测序样本包含宿主和非宿主核酸的混合物。

[0713] 42.实施方式41的归一化对照,其中所述宿主是真核生物。

[0714] 43.实施方式41的归一化对照,其中所述宿主是昆虫,植物或动物。

[0715] 44.实施方式43的归一化对照,其中所述动物是人。

[0716] 45.实施方式41-44中任一项的归一化对照,其中所述非宿主包括共生体,共生生物,寄生虫或病原体。

[0717] 46.实施方式45的归一化对照,其中所述非宿主包括多种物种。

[0718] 47.实施方式41-44中任一项的归一化对照,其中所述非宿主包括病毒物种,细菌物种,真菌物种,真核寄生虫物种或其组合。

[0719] 48.实施方式32-47中任一项的归一化对照,其中所述至少一个靶序列是测序样本

中的非宿主序列。

[0720] 49.实施方式48的归一化对照,其中所述非宿主是病毒,细菌或真菌,并且所述至少一个靶序列包含毒力因子序列。

[0721] 50.实施方式49的归一化对照,其中所述毒力因子包含一种或多种基因编码内毒素,外毒素,溶血素,蛋白酶,脂肪酶,DNA酶,粘附素,侵袭素,抗吞噬剂,破坏性酶或GTP酶活性的调节剂。

[0722] 51.实施方式48的归一化对照,其中所述非宿主是细菌,并且所述至少一个靶序列包含抗生素抗性基因的序列。

[0723] 52.实施方式32-51中任一项的归一化对照,其中所述至少一个靶序列包含多个参考序列。

[0724] 53.实施方式52的归一化对照,其中多个参考序列包含至少2个、至少10个、至少50个、至少100个、至少200个、至少300个、至少400个、至少500个、至少600个、至少700个、至少800个、至少900个、至少1,000个、至少1,200个、至少1,300个、至少1,400个、至少1,500个、至少1,600个、至少1,700个、至少1,800个、至少1,900个、至少2,000个、至少2,200个、至少2,400个、至少2,600个、至少2,800个、至少3,000个、至少4,000个、至少5,000个、至少6,000个、至少7,000个、至少8,000个、至少9,000个或至少10,000个参考序列。

[0725] 54.实施方式53的归一化对照,其中所述参考序列包含非宿主序列。

[0726] 55.实施方式53或54的归一化对照,其中所述参考序列包含病毒序列,细菌序列,真菌序列或其组合。

[0727] 56.实施方式52-55中任一项的方法,其中所述多个参考序列包含来自多种生物参考序列。

[0728] 57.实施方式56的归一化对照,其中所述参考序列包含基因组DNA序列或cDNA序列或其组合。

[0729] 58.实施方式52-57中任一项的归一化对照,其中所述分离的序列包含至少一个已经改组的参考序列的至少一个片段。

[0730] 59.实施方式58的归一化对照,其中所述分离的序列包含至少第一参考序列和第二参考序列的至少两个片段,所述至少两个片段已经被改组和连接。

[0731] 60.实施方式58的归一化对照,其中所述第一参考序列和第二参考序列在生物的基因组中不相邻。

[0732] 61.实施方式58的归一化对照,其中第一和第二参考序列来自不同的生物。

[0733] 62.实施方式58-61中任一项的归一化对照,其中所述参考序列的至少一个片段包含参考序列的约10-100,约10-80,约20-60,约20-40,约20-30或约20-25个连续核苷酸。

[0734] 63.实施方式52-62中任一项的归一化对照,其中:

[0735] a.所述至少一个参考序列由参考序列片段的分布表示;

[0736] b.分布划分为至少5个箱;和

[0737] c.多个参考序列片段选自至少5个箱中的至少3个,将其改组并任选地连接,从而产生分离的序列。

[0738] 64.实施方式63的归一化对照,其中所述参考序列片段相对于参考序列长度和(1) GC含量百分比、(2) 熵、(3) 复杂性、(4) EIIP、(5) 长度或其组合中的至少一个而分布。

[0739] 65.实施方式63或64的归一化对照,其中每个箱代表所述分布的1%,2%,5%,10%,15%或20%。

[0740] 66.实施方式63或64的归一化对照,其中每个箱代表GC含量分布的5%,10%,15%或20%。

[0741] 67.实施方式63-66中任一项的归一化对照,其中归一化对照的每组内的多核苷酸通过改组和任选地连接选自至少3个箱之一的参考序列片段而产生。

[0742] 68.实施方式65-67中任一项的归一化对照,其中每个组中的多核苷酸具有以下中的至少三者:(1)与参考序列分布中对应箱的参考序列片段类似的GC含量百分比,(2)类似的熵,(3)类似的EIIP,(4)类似的长度,(5)类似的复杂性。

[0743] 69.实施方式63-68中的任一项的归一化对照,其中参考序列片段选自GC含量分布百分比的第10,第30,第50,第70,第80和第90个百分位箱。

[0744] 70.实施方式52-57中任一项的归一化对照,其中所述分离的序列包含连接的至少一个参考序列的至少两个片段,并且其中所述至少两个片段在所述至少一个参考序列中不相邻。

[0745] 71.实施方式52-57中任一项的归一化对照,其中所述分离的序列包含至少第一参考序列的至少第一片段和至少第二参考序列的至少第二片段,其中所述至少第一和第二片段是连接的,并且其中第一和第二参考序列不是相同的参考序列。

[0746] 72.实施方式70或71的归一化对照,其中所述至少第一和第二片段是对应的参考序列的约15-60,约20-40,约20-30,约15-32、20-32或约25-35个连续核苷酸。

[0747] 73.实施方式70或71的归一化对照,其中所述至少第一和第二片段是对应参考序列的29、30、31、32、33或34个连续核苷酸。

[0748] 74.实施方式70-73中任一项的归一化对照,其中所述至少第一和第二片段小于测序读段的平均长度的一半。

[0749] 75.实施方式70-74中任一项的归一化对照,其中:

[0750] a.至少一个参考序列由使用滑动窗口产生的重叠参考序列片段的分布表示;

[0751] b.分布划分为至少5个箱;和

[0752] c.从至少5个箱中的至少3个中选择多个参考序列片段并连接,从而产生分离的序列。

[0753] 76.实施方式75的归一化对照,其中所述参考序列片段相对于参考序列长度和(1)GC含量百分比、(2)熵、(3)复杂性、(4)EIIP或其组合中的至少一个而分布。

[0754] 77.实施方式75或76的归一化对照,其中每个箱代表分布的1%,2%,5%,10%,15%或20%。

[0755] 78.实施方式75或76的归一化对照,其中每个箱代表GC含量分布的5%,10%,15%或20%。

[0756] 79.实施方式75-78中任一项的归一化对照,其中每个组内的多核苷酸通过连接选自至少3个箱之一的参考序列片段而产生。

[0757] 80.实施方式75-79中任一项的归一化对照,其中每个组中的多核苷酸具有以下至少三个:(1)与参考序列分布中的对应箱的参考序列片段类似的GC含量百分比,(2)类似的熵,(3)类似的EIIP,(4)类似的长度,(5)类似的复杂性。

- [0758] 81.实施方式75-80中任一项的归一化对照,其中所述滑动窗包括1bp,2bp,3bp,4bp或5bp的滑动窗。
- [0759] 82.实施方式75-81中的任一项的归一化对照,其中参考序列片段选自GC含量分布百分比的第10,第30,第50,第70,第80和第90个百分位箱。
- [0760] 83.实施方式32-80中任一项的归一化对照,其中所述测序样本包括来自被诊断患有癌症的受试者的样本。
- [0761] 84.实施方式83的归一化对照,其中所述癌症包括细胞群体,所述细胞群体包含相对于所述受试者的非癌性细胞或来自健康受试者的非癌性细胞具有至少一个遗传改变的序列。
- [0762] 85.实施方式84的归一化对照,其中所述至少一个遗传改变包括单核苷酸多态性(SNP),插入,缺失,倒位或染色体重排。
- [0763] 86.实施方式83或84的归一化,其中所述靶序列包含至少一种遗传改变。
- [0764] 87.实施方式32-82中任一项的归一化对照,其中所述测序样本包括宏基因组样本。
- [0765] 88.实施方式87的归一化对照,其中所述宏基因组样本是环境样本。
- [0766] 89.实施方式88的归一化对照,其中所述至少一个靶序列包含宏基因组学样本内的物种的序列。
- [0767] 90.实施方式88的归一化对照,其中所述至少一个靶序列包含毒力因子,抗生素抗性基因的序列或与致病性相关的序列。
- [0768] 91.实施方式1-90中任一项的归一化对照,其中每个组的多核苷酸处于相同浓度。
- [0769] 92.实施方式1-90中任一项的归一化对照,其中每组的多核苷酸不具有相同的浓度。
- [0770] 93.实施方式92的归一化对照,其中所有组的多核苷酸处于线性序列或几何序列的浓度。
- [0771] 94.实施方式93的归一化对照,其中一组多核苷酸具有基线浓度,并且其中剩余的非基线组的浓度是基线浓度的整数倍。
- [0772] 95.实施方式93的归一化对照,其中一组多核苷酸具有基线浓度,并且其中每组多核苷酸的浓度是归一化对照中的另一组的浓度的两倍,不包括基线浓度。
- [0773] 96.实施方式93的归一化对照,其中所述至少三组多核苷酸以1:2:4的浓度比率存在。
- [0774] 97.实施方式1-96中任一项的归一化对照,其中组内的多核苷酸具有相同的GC含量百分比。
- [0775] 98.实施方式97的归一化对照,其中组内的多核苷酸与每个其他组内的多核苷酸具有相同的GC含量百分比。
- [0776] 99.实施方式97的归一化对照,其中组内的多核苷酸与任何其他组内的多核苷酸具有不同的GC含量百分比。
- [0777] 100.实施方式97-99中任一项的归一化对照,其中每个组内的多核苷酸的GC含量百分比在40%至60%之间,包括端点。
- [0778] 101.实施方式97-99中任一项的归一化对照,其中每个组内的多核苷酸的GC含量

百分比在43%至56%之间,包括端点。

[0779] 102.实施方式1-101中任一项的归一化对照,其中所述多核苷酸包含编码唯一分子标识符的序列。

[0780] 103.实施方式102的归一化对照,其中编码唯一分子标识符的序列包含第一索引序列或第二索引序列。

[0781] 104.实施方式1-103中任一项的归一化对照,其中所述多核苷酸包含脱氧核糖核酸(DNA)分子,核糖核酸(RNA)分子或DNA-RNA杂合分子。

[0782] 105.实施方式104的归一化对照,其中多核苷酸在归一化对照中的浓度在每微升(μL)约0.005至约500皮克(pg)之间。

[0783] 106.实施方式105的归一化对照,其中所述多核苷酸在归一化对照中的浓度为约0.5至约50 $\text{pg}/\mu\text{L}$ 。

[0784] 107.实施方式105的归一化对照,其中所述多核苷酸在归一化对照中的浓度为1.4 $\text{pg}/\mu\text{L}$ 。

[0785] 108.实施方式1-107中任一项的归一化对照,所述归一化对照还包含至少一种蛋白质。

[0786] 109.实施方式108的归一化对照,其中所述至少一种蛋白质分离或衍生自细菌,病毒或真核寄生虫。

[0787] 110.实施方式108的归一化对照,其中所述至少一种蛋白质是病毒衣壳蛋白或细菌细胞壁蛋白。

[0788] 111.实施方式108-110中任一项的归一化对照,其中所述多核苷酸群和所述蛋白质可操作地连接。

[0789] 112.一种制备归一化对照的方法,所述方法包括:

[0790] a.合成至少三组DNA分子,其中一组内的DNA分子具有相同的序列,并且其中每个DNA分子的序列在5'至3'之间包含第一组分序列,第一限制性酶切位点,第二组分序列,第二限制性位点和第三组分序列;

[0791] b.对溶液中所述多组DNA分子中的每个组的量进行定量;

[0792] c.用切割第一和第二限制位点的限制性酶消化所述多组DNA分子中的每个组,以产生至少三组经消化的DNA分子;以及

[0793] d.将预定量的所述多组经消化的DNA分子中的各个组混合,从而产生归一化对照。

[0794] 113.实施方式112的方法,其中每个DNA分子包含与所有其他DNA分子的第一、第二和第三组分序列相同长度的第一、第二和第三组分序列。

[0795] 114.实施方式112的方法,其中所述第一、第二和第三组分序列各自在约15bp和约1200bp之间。

[0796] 115.实施方式112的方法,其中所述第一、第二和第三组分序列各自在约100bp至约600bp之间。

[0797] 116.实施方式112-115中任一项的方法,其中所述至少三组DNA分子中的任一个的第一、第二和第三组分序列都不是相同的序列。

[0798] 117.实施方式112的方法,其中所述第一和第二限制位点的序列选自EcoRI位点、BamHI位点、HindIII位点、NotI位点、PvuII位点、SmaI位点、EcoRV位点、KpnI位点、PstI位

点、SalI位点、ScaI位点、SpeI位点、SphI位点、StuI位点和XbaI位点的序列。

[0799] 118.实施方式112的方法,其中所述第一和第二限制位点的序列包含XhoI位点,所述限制性酶包含XhoI。

[0800] 119.实施方式112-118中任一项的方法,其中所述至少三个DNA分子中的每一个的序列进一步包含第一引物序列和与第二引物序列互补的序列,其中所述第一引物序列位于所述第一组分序列的5',其中与第二引物序列互补的序列位于第三组分序列的3'。

[0801] 120.实施方式119的方法,其中所述第一引物序列包含序列5'-TGAAGAACTGCGGCAGG-3'(SEQ ID NO:12)。

[0802] 121.实施方式119或120的方法,其中与第二引物序列互补的序列包含序列5'-GGTCTGCACTTCCAGCT-3'(SEQ ID NO:13)。

[0803] 122.实施方式112-121中任一项的方法,其中所述至少三组消化的DNA分子的预定量是相等的。

[0804] 123.实施方式112-122中任一项的方法,其中所述至少三组消化的DNA分子的预定量不相等。

[0805] 124.实施方式123的方法,其中所述至少三组消化的DNA分子的预定量为线性序列或几何序列。

[0806] 125.实施方式123的方法,其中一组消化的DNA分子的预定量是基线,并且所述至少两组另外的消化的DNA分子的预定量是基线的整数倍。

[0807] 126.实施方式123的方法,其中一组DNA分子的预定量是基线,并且其中每个其他组的消化的DNA分子的预定量是另一组消化的DNA分子的预定量的两倍,不包括基线。

[0808] 127.实施方式123的方法,其中所述至少三组消化的DNA分子的预定量的比率为1:2:4。

[0809] 128.实施方式112-127中任一项的方法,其中(d)的混合步骤还包括将预定量的至少另一组DNA分子与来自步骤(a)-(c)的至少三组消化的DNA分子混合以在单个归一化对照中产生至少四组DNA分子,并且其中至少另一组DNA分子从样本中分离或衍生。

[0810] 129.实施方式128的方法,其中所述样本包含质粒DNA,线粒体DNA,叶绿体DNA或基因组DNA。

[0811] 130.实施方式128或129的方法,其中所述至少四组DNA分子的预定量是相等的。

[0812] 131.实施方式128或129的方法,其中所述至少四组DNA分子的预定量不相等。

[0813] 132.实施方式131的方法,其中所述至少四组DNA分子的预定量为线性序列或几何序列。

[0814] 133.实施方式131的方法,其中一组DNA分子的预定量是基线,并且所述其他组的DNA分子的预定量是基线的整数倍。

[0815] 134.实施方式131的方法,其中一组消化的DNA分子的预定量是基线,并且其中每个其他组的消化的DNA分子的预定量是另一组消化的DNA分子的预定量的两倍,不包括基线。

[0816] 135.实施方式131的方法,其中所述至少四组消化的DNA分子的预定量的比率为1:2:4:8。

[0817] 136.一种制备包含RNA分子的归一化对照的方法,所述方法包括:

[0818] a.合成至少三组DNA分子,其中所述至少三个DNA分子中的每一个DNA分子从5'至3'包含第一组分序列、与启动子序列互补的序列、第一限制性酶切位点、第二组分序列、与启动子序列互补的第二序列、第二限制性位点、第三组分序列和与启动子序列互补的第三序列;

[0819] b.用切割第一和第二限制位点的限制性酶消化所述多组DNA分子中的每个组,以产生至少三组经消化的DNA分子;

[0820] c.体外转录至少三组消化的DNA分子中的每一组,以产生至少三组RNA分子;

[0821] d.对RNA分子群体中的每个群体中产生的RNA的量进行定量;以及

[0822] e.混合预定量的所述RNA分子群体中的各群体,从而产生包含RNA分子的归一化对照。

[0823] 137.实施方式136的方法,其中所述至少三个DNA分子中的每一个包含与所述至少三个DNA分子中的任何其他分子的第一、第二和第三组分序列具有相同长度的第一、第二和第三组分序列。

[0824] 138.实施方式136的方法,其中第一启动子序列,第二启动子序列和第三启动子序列包含相同的启动子序列。

[0825] 139.实施方式138的方法,其中所述启动子序列包含选自T7启动子序列,SP6启动子序列或T3启动子序列的启动子序列。

[0826] 140.实施方式139的方法,其中所述T7启动子序列包含序列5'-TAATACGACTCACTATAGGG-3'(SEQ ID NO:15)。

[0827] 141.实施方式136-140中任一项的方法,其中所述第一和第二限制位点的序列相同。

[0828] 142.实施方式136-141中任一项的方法,其中第一、第二和第三组分序列均介于约15bp与约1200bp之间。

[0829] 143.实施方式136-141中任一项的方法,其中第一、第二和第三组分序列均介于约100bp与约600bp之间。

[0830] 144.实施方式136-141中任一项的方法,其中所述至少三组DNA分子中的任何一个的第一、第二和第三组分序列都不是相同的序列。

[0831] 145.实施方式136-144中任一项的方法,其中所述第一和第二限制位点的序列选自EcoRI位点,BamHI位点,HindIII位点,NotI位点,PvuII位点,SmaI位点,EcoRV位点,KpnI位点,PstI位点,SalI位点,ScaI位点,SpeI位点,SphI位点,StuI位点和XbaI位点的序列。

[0832] 146.实施方式145的方法,其中所述第一和第二限制位点的序列包含XhoI位点,并且所述限制性酶包含XhoI。

[0833] 147.实施方式136-146中任一项的方法,其中所述至少三个RNA分子群体中的每一个的预定量是相等的。

[0834] 148.实施方式136-146中任一项的方法,其中所述至少三个RNA分子群体中的每一个的预定量不相等。

[0835] 149.实施方式148的方法,其中所述至少三个RNA分子群体的预定量为线性序列或几何序列。

[0836] 150.实施方式148的方法,其中一个合成RNA分子群体的预定量是基线,并且其中

至少两个另外的RNA分子群体的预定量是基线的整数倍。

[0837] 151.实施方式148的方法,其中一个RNA分子群体的预定量是基线,并且其中每个其他RNA分子群体的预定量是除基线之外的另一个RNA分子群体的浓度的两倍。

[0838] 152.实施方式148的方法,其中所述至少三个RNA分子群体的预定量为1:2:4的比率。

[0839] 153.实施方式136-152中任一项的方法,所述方法进一步包括将归一化对照稀释至约0.005至约500pg/ μ L之间的浓度。

[0840] 154.实施方式136-152中任一项的方法,所述方法进一步包括将归一化对照稀释至约0.5至约50pg/ μ L之间的浓度。

[0841] 155.实施方式136-152的方法,其中所述归一化对照浓度为1.4pg/ μ L。

[0842] 156.一种制备归一化对照的方法,所述归一化对照包括至少三组多核苷酸,其中每组中的多核苷酸具有相同的长度,该方法包括:

[0843] i.从样本中提取DNA;

[0844] ii.用限制性酶消化DNA以产生DNA片段集合;

[0845] iii.分离DNA片段集合;

[0846] iv.纯化DNA片段以产生至少三组多核苷酸,其中每组内的多核苷酸具有相同的长度;和

[0847] v.将预定量的所述多组经消化的DNA分子中的各个组混合,从而产生归一化对照。

[0848] 157.实施方式156的方法,其中所述限制性酶选自EcoRI, BamHI, HindIII, PvuII, SmaI, EcoRV, KpnI, PstI, SalI, ScaI, SpeI, SphI, StuI, XbaI, NotI, AscI, FseI, PacI, PmeI, BglII, BstBI, HincII和SgfI。

[0849] 158.实施方式156或157的方法,其中分离DNA片段集合包括基于片段大小的分离。

[0850] 159.实施方式158的方法,其中基于大小分离DNA片段包括凝胶电泳,色谱法或切向流过滤(TFF)。

[0851] 160.实施方式156-159中任一项的方法,所述方法还包括将衔接子连接到所述至少三个不同组的多核苷酸中的每一个。

[0852] 161.通过实施方式112-160中任一项的方法生成的归一化对照。

[0853] 162.试剂盒,其包含实施方式1-111中任一项的归一化对照。

[0854] 163.一种多分析物对照,其包含至少三种不同物种的生物的混合物,其中所述生物已被灭活。

[0855] 164.实施方式163的多分析物对照,其中每种物种以相同的滴度存在于所述多分析物对照中。

[0856] 165.实施方式163的多分析物对照,其中每种物种以不同的滴度存在于多分析物对照中。

[0857] 166.实施方式163的多分析物对照,其中所述滴度为线性序列或几何序列。

[0858] 167.实施方式163的多分析物对照,其中至少两种物种以相同的滴度存在于所述多分析物对照中,并且至少两种物种以不同的滴度存在于所述混合物中。

[0859] 168.实施方式163-167中任一项的多分析物对照,其中所述多分析物对照中的每个不同物种以约5至约1000万单位/mL(U/mL)的滴度存在。

- [0860] 169.实施方式168的多分析物对照,其中生物的物种包括病毒,细菌,真菌或真核寄生虫的物种。
- [0861] 170.实施方式169的多分析物对照,其中所述病毒是DNA病毒。
- [0862] 171.实施方式169的多分析物对照,其中所述病毒是RNA病毒。
- [0863] 172.实施方式169的多分析物对照,其中所述细菌是革兰氏阳性细菌或革兰氏阴性细菌。
- [0864] 173.实施方式163-172中任一项的多分析物对照,其中生物的物种是人类病原体。
- [0865] 174.实施方式173的多分析物对照,其中所述多分析物对照中的每个不同物种的滴度为每mL约5至约1000万感染单位(IU/mL)。
- [0866] 175.实施方式163的多分析物对照,其中所述多分析物对照包含10种或11种病毒或基本上由其组成。
- [0867] 176.实施方式175的多分析物对照,其中所述多分析物对照中的物种选自表8中所述的物种的集合。
- [0868] 177.实施方式173或175的多分析物对照,其中所述人类病原体包括在血液或组织移植物中发现的人类病原体。
- [0869] 178.实施方式177的多分析物对照,其中在血液或组织移植物中发现的人类病原体包括巨细胞病毒(CMV),爱泼斯坦-巴尔病毒(EBV),腺病毒(ADV),BK病毒(BKV),JC多瘤病毒(JCV),人疱疹病毒6A(HHV6A),人疱疹病毒6B(HHV6B),单纯疱疹病毒1型(HSV1),单纯疱疹病毒2型(HSV2),水痘带状疱疹病毒(VZV)和人类细小病毒B19(B19),人类免疫缺陷病毒(HIV),乙型肝炎,丙型肝炎,I型和II型人类T淋巴细胞性病毒(HTLV),西尼罗河病毒,寨卡病毒,链球菌种,金黄色葡萄球菌,粪肠球菌,蜡状芽孢杆菌,痤疮丙酸杆菌,粘质沙雷氏菌,沙雷氏菌粘菌,小肠结肠炎耶尔森菌,肠杆菌属,不动杆菌属,假单胞菌属,大肠杆菌,肺炎克雷伯菌,奇异变形杆菌,鲍氏不动杆菌,蜡状芽孢杆菌,凝固阴性葡萄球菌,链球菌,克雷伯菌属,沙雷氏菌属雷氏普罗威登斯菌,苍白密螺旋体,克鲁斯锥虫和小巴贝虫中的一种或多种。
- [0870] 179.实施方式173或174的多分析物对照,其中所述多分析物对照中的人类病原体选自表2-3或5-7。
- [0871] 180.实施方式163-172中任一项的多分析物对照,其中所述物种不是人类病原体。
- [0872] 181.实施方式180的多分析物对照,其中所述物种与作为人类病原体的物种在相同属中。
- [0873] 182.实施方式163-172中任一项的多分析物对照,其中所述物种包括作为人类病原体的物种和不是人类病原体的物种的混合物。
- [0874] 183.实施方式163-182中任一项的多分析物对照,其中所述生物已通过化学处理,热处理,pH处理或紫外线照射而灭活。
- [0875] 184.实施方式183的多分析物对照,其中所述化学处理包括甲醛。
- [0876] 185.实施方式163-184中任一项的多分析物对照,其还包含血浆。
- [0877] 186.实施方式183的多分析物对照,其中所述血浆是人血浆。
- [0878] 187.包含从实施方式163-186中任一项的多分析物对照中提取的核酸的集合。
- [0879] 188.实施方式163-186中任一项的多分析物对照或实施方式185的集合,还包括实

施方式1-111中任一项的归一化对照。

[0880] 189.一种试剂盒,所述试剂盒包含实施方式163-186中任一项的多分析物对照或实施方式187的集合。

[0881] 190.实施方式189的试剂盒,所述试剂盒进一步包含实施方式1-111中任一项的归一化对照。

[0882] 191.一种量化样本中的至少一个靶序列表达水平的方法,所述方法包括:

[0883] a.将已知量的实施方式1-111中任一项的归一化对照与样本混合,

[0884] b.制备高通量测序文库,

[0885] c.对所述文库进行测序以产生读段的集合,

[0886] d.将来自读段集合的读段映射到样本或归一化对照,

[0887] e.确定通过归一化对照中的多核苷酸的每个组或亚组产生的读段的数目,

[0888] f.计算与(a)中的样本混合的归一化对照中的多核苷酸的每个组或亚组的初始浓度与(c)中产生的读段的数目之间的关系,

[0889] g.模拟样本中的至少一个靶序列的读段与浓度之间的关系,和

[0890] h.使用(g)中的模型,由通过靶序列产生的读段的数目计算样本中的至少一种靶多核苷酸的初始浓度。

[0891] 192.实施方式191的方法,所述方法进一步包括从样本中提取核酸。

[0892] 193.实施方式191的方法,其中所述核酸包含DNA,RNA或其混合物。

[0893] 194.实施方式191-193中任一项的方法,其中(a)的混合步骤发生在(b)的文库制备步骤之前,与(b)的文库制备步骤同时或在(b)的文库制备步骤之后。

[0894] 195.实施方式191-194中任一项的方法,所述方法进一步包括从样本中提取核酸,并且其中(a)的混合步骤在从样本中提取核酸之前进行。

[0895] 196.实施方式191-195中任一项的方法,其中将归一化对照中的至少三组多核苷酸中的每组以约0.001:0.420:1的比率加到样本中。

[0896] 197.实施方式191-196中任一项的方法,其中将归一化对照中的至少三组多核苷酸中的每组以约0.001pg/ μ L至约5pg/ μ L之间的最终浓度加到样本。

[0897] 198.实施方式191-197中任一项的方法,其中所述样本是人样本。

[0898] 199.实施方式191-198中任一项的方法,其中所述靶序列包含病毒,细菌,真菌,真核寄生虫或一个或多个人类基因的序列。

[0899] 200.实施方式191-198中任一项的方法,其中所述靶序列包含毒力因子序列,抗生素抗性基因或与致病性相关的序列。

[0900] 201.实施方式191-198中任一项的方法,其中所述样本包括来自诊断为患有癌症的受试者的样本。

[0901] 202.实施方式201的方法,其中所述癌症包括细胞群体,所述细胞群体包含相对于所述受试者的非癌细胞具有至少一种遗传变化的序列。

[0902] 203.实施方式202的方法,其中所述至少一个遗传改变包括单核苷酸多态性(SNP),插入,缺失,倒位或重排。

[0903] 204.实施方式201-203中任一项的方法,其中所述靶序列包含具有至少一个遗传改变的序列。

- [0904] 205.实施方式191-197中任一项的方法,其中所述样本包括宏基因组样本。
- [0905] 206.实施方式205的方法,其中所述宏基因组样本是环境样本。
- [0906] 207.实施方式205或206的方法,其中所述靶多核苷酸包含在宏基因组学样本内的物种的序列。
- [0907] 208.实施方式205-207中任一项的方法,其中所述靶多核苷酸包含毒力因子序列,抗生素抗性基因或与致病性相关的序列。
- [0908] 209.实施方式191-193中任一项的方法,其中所述样本包括多分析物对照,所述多分析物对照包括至少三种不同物种的生物的混合物,其中所述生物已被灭活。
- [0909] 210.实施方式209的方法,其中所述至少三种不同物种的生物中的每一种均包含至少一种靶序列。
- [0910] 211.实施方式210的方法,其中所述至少一种靶序列是每种生物物种中的不同靶序列。
- [0911] 212.实施方式209的方法,其中
- [0912] (i) 在步骤(d)中将来自读段集合的读段映射到样本还包括将读段映射到多分析物对照(d)中的至少三种生物的每一种;和
- [0913] (ii) 在(h)计算至少一个靶序列的初始浓度包括为每种物种计算至少一个靶序列的初始浓度。
- [0914] 213.实施方式209-212中任一项的方法,其中每种物种以相同的滴度存在于多分析物对照中。
- [0915] 214.实施方式209-212中任一项的方法,其中每种物种以不同的滴度存在于多分析物对照中。
- [0916] 215.实施方式214的方法,其中所述滴度为线性序列或几何序列。
- [0917] 216.实施方式209-212中任一项的方法,其中至少两种物种以相同的滴度存在于多分析物对照中,并且至少两种物种以不同的滴度存在。
- [0918] 217.实施方式209-216中任一项的方法,其中所述多分析物对照中的每种物种的滴度为5至1000万单位/mL(U/mL)。
- [0919] 218.实施方式209-217中任一项的方法,其中所述生物是病毒,细菌,真菌,真核寄生虫或其组合。
- [0920] 219.实施方式218的方法,其中所述病毒是DNA病毒。
- [0921] 220.实施方式218的方法,其中所述病毒是RNA病毒。
- [0922] 221.实施方式218的方法,其中所述细菌是革兰氏阳性细菌或革兰氏阴性细菌。
- [0923] 222.实施方式209-221中任一项的方法,其中所述生物的物种是人类病原体。
- [0924] 223.实施方式222的方法,其中所述多分析物对照中的每个不同物种以约5至约1000万感染单位/mL(IU/mL)的滴度存在。
- [0925] 224.实施方式222或223的方法,其中所述多分析物对照包括10种或11种病毒物种或基本上由其组成。
- [0926] 225.实施方式224的方法,其中所述多分析物对照中的物种选自表8中列出的物种的集合。
- [0927] 226.实施方式222或223的方法,其中所述人类病原体包括在血液或组织移植植物中

发现的人类病原体。

[0928] 227.实施方式226的方法,其中在血液或组织移植物中发现的人类病原体包括巨细胞病毒(CMV),爱泼斯坦-巴尔病毒(EBV),腺病毒(ADV),BK病毒(BKV),JC多瘤病毒(JCV),人疱疹病毒6A(HHV6A),人疱疹病毒6B(HHV6B),单纯疱疹病毒1型(HSV1),单纯疱疹病毒2型(HSV2),水痘带状疱疹病毒(VZV)和人类细小病毒B19(B19),人类免疫缺陷病毒(HIV),乙型肝炎,丙型肝炎,I型和II型人类T淋巴细胞性病毒(HTLV),西尼罗河病毒,寨卡病毒,链球菌种,金黄色葡萄球菌,粪肠球菌,蜡状芽孢杆菌,痤疮丙酸杆菌,粘质沙雷氏菌,沙雷氏菌粘菌,小肠结肠炎耶尔森菌,肠杆菌属,不动杆菌属,假单胞菌属,大肠杆菌,肺炎克雷伯菌,奇异变形杆菌,鲍氏不动杆菌,蜡状芽孢杆菌,凝固阴性葡萄球菌,链球菌,克雷伯菌属,沙雷氏菌属雷氏普罗威登斯菌,苍白密螺旋体,克鲁斯锥虫和小巴贝虫中的一种或多种。

[0929] 228.实施方式222或223的方法,其中多分析物对照中的人类病原体选自表2-3或5-7。

[0930] 229.实施方式209-224中任一项的方法,其中所述生物已通过化学处理,热处理,pH处理或紫外线照射而灭活。

[0931] 230.实施方式229的方法,其中化学处理包括甲醛。

[0932] 231.实施方式209-230中任一项的方法,其中所述多分析物对照还包括血浆。

[0933] 232.实施方式231的方法,其中所述血浆是人血浆。

[0934] 233.实施方式191-232中任一项的方法,其中所述方法包括将已知量的多分析物对照与样本混合,其中所述多分析物对照包括至少三种不同物种的生物的混合物,并且其中生物已被灭活。

[0935] 234.实施方式233的方法,其中所述生物是病毒,细菌,真菌,真核寄生虫或其组合。

[0936] 235.实施方式234的方法,其中所述病毒是DNA病毒。

[0937] 236.实施方式234的方法,其中所述病毒是RNA病毒。

[0938] 237.实施方式234的方法,其中所述细菌是革兰氏阳性细菌或革兰氏阴性细菌。

[0939] 238.实施方式234-237中任一项的方法,其中所述生物不是人类病原体。

[0940] 239.实施方式238的方法,其中所述生物是与人类病原体的物种在相同属中的物种。

[0941] 240.实施方式239的方法,其中所述人类病原体包括在血液或组织移植物中发现的人类病原体。

[0942] 241.实施方式240的方法,其中在血液或组织移植物中发现的人类病原体包括巨细胞病毒(CMV),爱泼斯坦-巴尔病毒(EBV),腺病毒(ADV),BK病毒(BKV),JC多瘤病毒(JCV),人疱疹病毒6A(HHV6A),人疱疹病毒6B(HHV6B),单纯疱疹病毒1型(HSV1),单纯疱疹病毒2型(HSV2),水痘带状疱疹病毒(VZV)和人类细小病毒B19(B19),人类免疫缺陷病毒(HIV),乙型肝炎,丙型肝炎,I型和II型人类T淋巴细胞性病毒(HTLV),西尼罗河病毒,寨卡病毒,链球菌种,金黄色葡萄球菌,粪肠球菌,蜡状芽孢杆菌,痤疮丙酸杆菌,粘质沙雷氏菌,沙雷氏菌粘菌,小肠结肠炎耶尔森菌,肠杆菌属,不动杆菌属,假单胞菌属,大肠杆菌,肺炎克雷伯菌,奇异变形杆菌,鲍氏不动杆菌,蜡状芽孢杆菌,凝固阴性葡萄球菌,链球菌,克雷伯菌属,沙雷氏菌属雷氏普罗威登斯菌,苍白密螺旋体,克鲁斯锥虫和小巴贝虫中的一种或多种。

- [0943] 242.实施方式234-240中任一项的方法,其中所述生物已通过化学处理,热处理,pH处理或紫外线照射灭活。
- [0944] 243.实施方式234-242中任一项的方法,其中将已知量的多分析物对照与样本混合在步骤(b)中的文库制备之前进行。
- [0945] 244.实施方式234-243中任一项的方法,其中将所述多分析物对照以所述多分析物对照中的每种物种的约5至1000万U/mL之间的浓度加到所述样本。
- [0946] 245.实施方式191-244中任一项的方法,其中所述模型是线性模型。
- [0947] 246.实施方式191-244中的任一项的方法,其中所述模型用机器学习分类器创建。
- [0948] 247.实施方式246的方法,其中所述机器学习分类器是受监督的。
- [0949] 248.实施方式247的方法,其中所述机器学习分类器是矢量支持机器。
- [0950] 249.实施方式246的方法,其中样本中的至少一种靶核酸分子的初始浓度与通过靶核酸分子产生的读段的数目的线性回归具有在归一化后大于0.95,大于0.96,大于0.97,大于0.98或大于0.99的R²值。
- [0951] 250.实施方式249的方法,其中线性回归的R²值在归一化后提高至少0.01、0.03、0.05、0.07、0.09、0.1、0.13、0.15、0.17、0.19、0.2、0.23、0.25、0.27、0.29、0.3、0.33、0.35、0.37、0.39、0.4、0.43、0.45、0.47、0.49、0.5、0.53、0.55、0.57、0.59、0.6、0.63、0.65、0.67、0.69、0.7或其之间的任何值。
- [0952] 251.一种量化样本中的靶生物滴度的方法,所述方法包括:
- [0953] a.提供包含靶生物的样本,其中所述靶生物包含至少一个靶序列;
- [0954] b.提供包含至少三种生物物种的已知滴度的多分析物对照,其中所述生物已被灭活;
- [0955] c.将已知量的实施方式1-111中任一项的归一化对照与样本和多分析物对照混合;
- [0956] d.从样本和多分析物对照中制备高通量测序文库;
- [0957] e.对所述文库测序以产生样本读段集合和多分析物对照读段集合;
- [0958] f.使用归一化对照将样本读段集合和来自(e)的多分析物对照读段集合归一化;
- [0959] g.确定多分析物对照中的至少三种生物物种的归一化读段与已知滴度之间的关系;和
- [0960] h.使用(g)中确定的关系计算样本中靶生物的 α 。
- [0961] 252.实施方式251的方法,其中所述至少三种生物各自包含至少一种种特异性靶序列。
- [0962] 253.实施方式252的方法,其中至少一种物种特异性靶序列在多分析物对照中的每个物种中是不同的。
- [0963] 254.实施方式251的方法,其中在步骤(f)进行归一化包括:
- [0964] i.将样本读段集合中的读段映射到样本或归一化对照;
- [0965] ii.将来自多分析物对照读段集合的读段映射到多分析物对照或归一化对照;
- [0966] iii.确定通过归一化对照中的多核苷酸的每个组或亚组产生的读段的数目,以收集样本读段和收集多分析物对照读段;
- [0967] iv.计算归一化对照中的多核苷酸的每个组或亚组的初始浓度与样本和多分析物

对照在步骤(e)产生的读段的数目之间的关系;和

[0968] v.确定样本和多分析物对照中的读段与浓度之间的关系。

[0969] 255.实施方式251-254中任一项的方法,其中确定(g)的读段与滴度之间的关系包括:

[0970] vi.从多分析物对照中的至少三种物种的每一种计算至少一种物种特异性靶序列的初始浓度;和

[0971] vii.计算样本中的至少一个靶序列的初始浓度。

[0972] 256.实施方式251或255的方法,其中步骤(i)的映射包括将样本读段集合映射到参考基因组。

[0973] 257.实施方式256的方法,其中参考基因组不包含靶序列。

[0974] 258.实施方式256或257的方法,其中参考基因组是人基因组。

[0975] 259.实施方式254-258中任一项的方法,其中步骤(ii)的映射进一步包括针对多分析物对照中的至少三种生物物种中的每一种,对物种特异性参考基因组进行映射。

[0976] 260.实施方式251-259中任一项的方法,所述方法进一步包括在步骤(b)中以2个浓度,3个浓度,4个浓度,5个浓度,6个浓度,7个浓度,8个浓度,9个浓度或10个浓度提供多分析物对照,然后将步骤(c)-(g)应用于步骤(b)中提供的多分析物对照的每种浓度。

[0977] 261.实施方式251-260中任一项的方法,其中在步骤(g)的多分析物对照中,归一化读段与至少三种生物物种的已知滴度之间的关系是线性关系。

[0978] 262.实施方式251-261中任一项的方法,其中每种物种以相同的滴度存在于多分析物对照中。

[0979] 263.实施方式251-261中任一项的方法,其中每种物种以不同的滴度存在于多分析物对照中。

[0980] 264.实施方式263的方法,其中所述滴度为线性序列或几何序列。

[0981] 265.实施方式251-261中任一项的方法,其中至少两种物种以相同的滴度存在于多分析物对照中,并且至少两种物种以不同的滴度存在。

[0982] 266.实施方式251-265中任一项的方法,其中每种物种以5至1000万单位/mL(U/mL)的浓度存在于所述多分析物对照中。

[0983] 267.实施方式251-266中任一项的方法,其中所述生物是病毒,细菌,真菌或其组合。

[0984] 268.1]实施方式267的方法,其中所述病毒是DNA病毒。

[0985] 269.实施方式267的方法,其中所述病毒是RNA病毒。

[0986] 270.实施方式267的方法,其中所述细菌是革兰氏阳性细菌或革兰氏阴性细菌。

[0987] 271.实施方式251-270中任一项的方法,其中所述生物是人类病原体。

[0988] 272.实施方式271的方法,其中所述多分析物对照中的每个不同物种以约5至约1000万感染单位/mL(IU/mL)的滴度存在。

[0989] 273.实施方式271或272的方法,其中所述多分析物对照包括10种病毒或11种病毒或基本上由其组成。

[0990] 274.实施方式273的方法,其中所述多分析物对照中的物种选自表8中列出的物种的集合。

[0991] 275.实施方式271或272的方法,其中所述人类病原体包括在血液或组织移植植物中发现的人类病原体。

[0992] 276.实施方式275的方法,其中在血液或组织移植植物中发现的人类病原体包括巨细胞病毒(CMV),爱泼斯坦-巴尔病毒(EBV),腺病毒(ADV),BK病毒(BKV),JC多瘤病毒(JCV),人疱疹病毒6A(HHV6A),人疱疹病毒6B(HHV6B),单纯疱疹病毒1型(HSV1),单纯疱疹病毒2型(HSV2),水痘带状疱疹病毒(VZV)和人类细小病毒B19(B19),人类免疫缺陷病毒(HIV),乙型肝炎,丙型肝炎,I型和II型人类T淋巴细胞性病毒(HTLV),西尼罗河病毒,寨卡病毒,链球菌种,金黄色葡萄球菌,粪肠球菌,蜡状芽孢杆菌,痤疮丙酸杆菌,粘质沙雷氏菌,沙雷氏菌粘菌,小肠结肠炎耶尔森菌,肠杆菌属,不动杆菌属,假单胞菌属,大肠杆菌,肺炎克雷伯菌,奇异变形杆菌,鲍氏不动杆菌,蜡状芽孢杆菌,凝固阴性葡萄球菌,链球菌,克雷伯菌属,沙雷氏菌属雷氏普罗威登斯菌,苍白密螺旋体,克鲁斯锥虫和小巴贝虫中的一种或多种。

[0993] 277.实施方式271或272的方法,其中所述多分析物对照中的人类病原体选自表2-3或5-7。

[0994] 278.实施方式251-270中任一项的方法,其中所述生物包括人类病原体的物种和不是人类病原体的物种的混合物。

[0995] 279.实施方式251-278中任一项的方法,其中所述生物已通过化学处理,热处理,pH处理或紫外线照射灭活。

[0996] 280.实施方式279的方法,其中化学处理包括甲醛。

[0997] 281.实施方式251-280中任一项的方法,其中所述多分析物对照还包括血浆。

[0998] 282.实施方式281的方法,其中血浆是人血浆。

[0999] 283.一种制备多个归一化对照寡核苷酸的方法,所述方法包括:

[1000] a.从至少一个参考序列产生多个参考序列片段;

[1001] b.根据参考序列片段的数目产生至少一个参数的分布;

[1002] c.将分布划分为至少5个箱;

[1003] d.从至少5个箱中的至少3个箱中选择至少一个参考序列片段;

[1004] e.改组至少3个参考序列以产生改组序列;和

[1005] f.合成包含改组序列的寡核苷酸;

[1006] 从而产生多个归一化对照寡核苷酸。

[1007] 284.实施方式283的方法,其中所述参数包括以下各项中的至少一项:(1)GC百分比,(2)熵,(3)复杂性,(4)EIIP或其组合。

[1008] 285.实施方式283的方法,其中所述参数包括GC含量百分比。

[1009] 286.实施方式283-285中任一项的方法,其中每个箱代表分布的1%,2%,5%,10%,15%或20%。

[1010] 287.实施方式283-285中任一项的方法,其中所述参考序列片段选自GC含量分布百分比的第10,第30,第50,第70,第80和90百分位箱。

[1011] 288.实施方式283-287中任一项的方法,其中步骤(d)包括从至少3个箱中的每一个中选择至少2个参考,并且其中步骤(e)还包括将来自每个箱的改组的参考序列连接。

[1012] 289.实施方式283-288中任一项的方法,其中所述至少一个参考序列包含至少2,至少10,至少20,至少50,至少100,至少200,至少300,至少400,至少500,至少600,至少700,

至少800,至少900,至少1,000,至少1,200,至少1,300,至少1,400,至少1,500,至少1,600,至少1,700,至少1,800,至少1,900,至少2,000,至少2,200,至少2,400,至少2,600,至少2,800,至少3,000,至少4,000,至少5,000,至少6,000,至少7,000,至少8,000,至少9,000或至少10,000个参考序列。

[1013] 290.实施方式283-288中任一项的方法,其中所述至少一个参考序列包含约2,000个参考序列。

[1014] 291.实施方式283-290中任一项的方法,其中所述至少一个参考序列包含基因组序列或cDNA序列或其组合。

[1015] 292.实施方式283-291中任一项的方法,其中所述多个参考序列片段包含片段参考序列的约10-100,约10-80,约20-60,约20-40,约20-30或约20-25个连续核苷酸。

[1016] 293.实施方式283-292中任一项的方法,其中每个归一化对照寡聚物具有至少以下三个:(1)与来自参考序列分布中对应箱的参考序列片段类似的GC含量百分比,(2)类似的熵,(3)类似的EIIP,(4)类似的长度或(5)类似的复杂性。

[1017] 294.一种用于制备多个归一化对照寡核苷酸的方法,所述方法包括:

[1018] a.使用滑动窗口从至少一个参考序列产生多个参考序列片段;

[1019] b.根据参考序列片段的数目产生至少一个参数的分布;

[1020] c.将分布划分为至少5个箱;

[1021] d.从至少5个箱中的至少3个中选择至少两个参考序列片段,其中所述至少两个参考序列片段在参考序列中是不连续的,或来自不同的参考序列;

[1022] e.从至少3个箱中的每个箱中连接至少两个参考序列片段;和

[1023] f.合成包含连接的参考序列片段的寡核苷酸;

[1024] 从而产生多个归一化对照寡核苷酸。

[1025] 295.实施方式294的方法,其中所述参数包括以下各项中的至少一项:(1)GC含量百分比,(2)熵,(3)复杂性,(4)EIIP或其组合。

[1026] 296.实施方式294的方法,其中所述参数包括GC含量百分比。

[1027] 297.实施方式294-296中任一项的方法,其中每个箱代表分布的1%,2%,5%,10%,15%或20%。

[1028] 298.实施方式294-296中任一项的方法,其中所述参考序列片段选自GC含量百分比分布的第10,第30,第50,第70,第80和第90个百分位箱。

[1029] 299.实施方式294-298中任一项的方法,其中所述滑动窗包括1bp,2bp,3bp,4bp或5bp的滑动窗。

[1030] 300.实施方式294-299中任一项的方法,其中所述参考序列片段是对应的参考序列的约15-60,约20-40,约20-30,约15-32,20-32或约25-35个连续核苷酸。

[1031] 301.实施方式294-299中任一项的方法,其中所述参考序列片段是对应参考序列的29、30、31、32、33或34个连续核苷酸。

[1032] 302.实施方式294-301中任一项的方法,其中所述参考序列片段小于测序读段的平均长度的一半。

[1033] 303.实施方式294-302中任一项的方法,其中所述至少一个参考序列包含至少2,至少10,至少20,至少50,至少100,至少200,至少300,至少400,至少500,至少600,至少700,

至少800,至少900,至少1,000,至少1,200,至少1,300,至少1,400,至少1,500,至少1,600,至少1,700,至少1,800,至少1,900,至少2,000,至少2,200,至少2,400,至少2,600,至少2,800,至少3,000,至少4,000,至少5,000,至少6,000,至少7,000,至少8,000,至少9,000或至少10,000个参考序列。

[1034] 304.实施方式294-302中任一项的方法,其中所述至少一个参考序列包含约2,000个参考序列。

[1035] 305.实施方式294-304中任一项的方法,其中参考序列包含基因组序列或cDNA序列。

[1036] 306.实施方式294-305中任一项的方法,其中每个归一化对照寡核苷酸具有以下至少三个:(1)与来自参考序列分布中的对应箱的参考序列片段类似的GC含量百分比,(2)类似的熵,(3)类似的EIIP,(4)类似的长度或(5)类似的复杂性。

[1037] 307.通过实施方式283-306中任一项的方法产生的归一化对照。

[1038] 308.一种试剂盒,所述试剂盒包含实施方式307的归一化对照。

[1039] 309.一种用于设计多个归一化对照多核苷酸序列的系统,所述系统包括:

[1040] a.存储计算机可执行指令的计算机可读存储介质,包括:

[1041] i.导入至少一个参考序列的指令;

[1042] ii.从至少一个参考序列产生多个参考序列片段的指令;

[1043] iii.根据参考序列片段的数目产生至少一个参数的分布的指令;

[1044] iv.将分布划分为箱的指令;

[1045] v.从箱的至少一个子集中选择多个参考序列片段的指令;

[1046] vi.改组多个参考序列片段以产生改组序列的指令;和

[1047] 从而产生多个归一化多核苷酸序列;和

[1048] b.处理器,被配置成执行包括以下步骤的步骤:

[1049] i.接收包括至少一个参考序列的一组输入文件;和

[1050] ii.执行存储在计算机可读存储介质中的计算机可执行指令。

[1051] 310.一种用于设计多个归一化对照多核苷酸序列的系统,所述系统包括:

[1052] 存储计算机可执行指令的计算机可读存储介质,包括:

[1053] i.导入至少一个参考序列的指令;

[1054] ii.从至少一个参考序列产生多个参考序列片段的指令;

[1055] iii.根据参考序列片段的数目产生至少一个参数的分布的指令;

[1056] iv.将分布划分为箱的指令;

[1057] v.从箱的至少一个子集中的每一个中选择至少两个参考序列片段的指令,其中所述至少两个参考序列片段在参考序列中是不连续的,或是从不同的参考序列中选择的;和

[1058] vi.连接来自每个箱的至少两个参考序列片段的指令;

[1059] 从而产生多个归一化多核苷酸序列;和

[1060] b.处理器,被配置成执行包括以下步骤的步骤:

[1061] i.接收包括至少一个参考序列的一组输入文件;和

[1062] ii.执行存储在计算机可读存储介质中的计算机可执行指令。

[1063] 311.实施方式309或310的系统,其中所述参数包括以下各项中的至少一项:(1)GC

含量百分比, (2) 熵, (3) 复杂性, (4) EIIP或其组合。

[1064] 实施例

[1065] 为了可以更有效地理解在此公开的发明, 下面提供实施例。应当理解, 这些实施例仅用于说明性的目的, 并且不应解释为以任何方式限制本发明。

[1066] 实施例1: RNA归一化对照

[1067] RNA归一化对照可以通过转录随机生成DNA序列进行。随机生成序列被合成为gBlock, 具有用于转录的末端T7启动子序列和XhoI限制性位点以线性化随机DNA模板。创建以下所示设计的三个版本, 分别称为RNA_NC_1, RNA_NC_2和RNA_NC_3。每个版本都是按照以下设计创建的, 但是具有不同的随机生成序列:

[1068] 5'-TGAAGAACTGCGGCAGG (SEQ ID NO:12) (Mito阵列引物)+175bp随机序列 (使用Random DNA Generator设计)+T7 RNA聚合酶启动子+XhoI位点+250bp随机序列 (使用随机DNA发生器设计)+T7 RNA聚合酶启动子+XhoI位点+450bp随机序列 (使用随机DNA发生器设计)+T7 RNA聚合酶启动子+(Ribo阵列引物)GGTCTGCACTTCCAGCT-3' (SEQ ID NO:13)。

[1069] Ribo阵列引物的序列包含序列5'-AGCTGGAAGTGCAGACC-3' (SEQ ID NO:20)。

[1070] T7启动子的序列包含5'-TAATACGACTCACTATAGGG-3'的序列 (SEQ ID NO:15)。

[1071] 设计的总长度为981bp。具有不同随机序列的三个设计版本按gBlocks (基因块) 测序, 分别称为RNA_NC_1, RNA_NC_2和RNA_NC_3。用XhoI消化RNA_NC_1, RNA_NC_2, RNA_NC_3gBlock, 以产生三个线性DNA片段, 并用T7 RNA聚合酶转录得到的三个片段的混合物, 以产生三个RNA NC。gBlock RNA_NC_1产生3个RNA片段: RNA_NC_1.1, RNA_NC_1.2, RNA_NC_1.3。gBlock RNA_NC_2产生3个RNA片段: RNA_NC_2.1, RNA_NC_2.2, RNA_NC_2.3。gBlock RNA_NC_3产生3个RNA片段: RNA_NC_3.1, RNA_NC_3.2, RNA_NC_3.3。对应于RNA_NC_1.1, RNA_NC_2.1和RNA_NC_3.1 NC RNA的DNA片段的T7转录产生的NC RNA包含与Mito阵列引物互补的序列, 与175bp随机DNA序列互补的序列和在转录期间由T7聚合酶引入的末端GGG序列。对应于NC RNA RNA_NC_1.2, RNA_NC_2.2和RNA_NC_3.2的DNA片段的T7转录产生的NC RNA包含与250bp随机序列互补的序列和在转录期间引入的末端GGG序列。对应于NC RNA RNA_NC_1.3, RNA_NC_2.3和RNA_NC_3.3的DNA片段的T7转录产生的NC RNA包含与450bp随机序列互补的序列和在转录期间引入的末端GGG序列。

[1072] 用于在样本提取期间添加NC对照的方案在图6中示出。稀释三个NC转录反应 (RNA归一化对照集1、2和3, 也称为NC 1, NC 2和NC 3或RNA_NC_1, RNA_NC_2和RNA_NC_3, 每个都包含上述三个转录本) 的输出20,000X至0.1ng/ μ L。然后通过混合25 μ L RNA_NC_1、50 μ L RNA_NC_2和50 μ L RNA_NC_3以1:2:4的比率混合三组NC转录本, 最终体积为175 μ L。该NC混合物总共包含9个不同的转录本, 3个包含3个不同的175bp随机序列, 3个包含3个不同的250bp随机序列, 而3个包含3个不同的450bp随机序列。然后, 通过将30 μ L混合物加到270 μ L AVE缓冲液中, 将NC混合物再稀释10倍, 以产生10pg/ μ L NC混合物。

[1073] 使用EZ1病毒迷你试剂盒 (Qiagen) 从阴性血浆样本 (精确诊断) 中提取核酸。将8 μ L 10pg/ μ L NC混合物和52 μ L AVE缓冲液加到EZ1试剂盒第3行1.5mL试管中。从第3行试管中取出50 μ L等分试样, 并将其加到400 μ L样本中, 从而在提取期间将70pg NC混合物加到样本中。

[1074] 分析了在提取期间向阴性血浆样本中添加三种不同量的NC混合物的结果。对于“高”量的NC, 每个样本添加38.10pg NC混合物 (NC 1:5.4pg, NC2:10.8pg, NC 3:21.8pg)。对

于“中等”量的NC,每个样本添加19.04pg NC混合物(NC 1:2.7pg,NC2:5.4pg,NC 3:10.8pg)。对于“低”量的NC,9.每个样本添加52pg NC混合物(NC 1:1.4pg,NC2:2.7pg,NC 3:5.4pg)。“低”数量导致NextSeq 500下一代测序后的读段为4%。基于重复数据删除的读段,转换为文库的“低”NC 1的质量为1.076pg,占输入的76%。

[1075] 图7总结了从精确诊断阴性血浆制备的RNA-Seq文库中RNA归一化对照的表示,其中在核酸提取期间添加NC。每个文库的平均读段输出为23,000,000次读段,人类平均代表率为80%。RNA归一化对照(38.10pg)的“高”水平添加使读段平均代表这些文库中的总读段的16%。RNA归一化对照(19.04pg)的“中等”水平添加使读段平均代表这些文库中的总读段的6%。RNA归一化对照的“低”水平添加(9.52pg)使读段平均代表这些文库中的总读段的4%。在“低”输入下,RNA_NC 1片段的转化率为76%(1.067pg)。

[1076] 实施方式2:DNA归一化对照

[1077] 可以使用随机生成序列制备DNA NC。下面是DNA NC设计的示例。使用不同的随机生成序列制作该设计的三个版本:

[1078] 5'-TGAAGAACTGCGGCAGG (SEQ ID NO:12) (Mito阵列引物)+使用随机DNA产生器的随机序列(200bp)+XhoI位点+使用随机DNA产生器的随机序列(300bp)+XhoI位点+使用随机DNA产生器的随机序列(500bp)+T7 RNA聚合酶启动子+(Ribo阵列引物)GGTCTGCACTTCCAGCT-3' (SEQ ID NO:13)。

[1079] Ribo阵列引物的从5'至3'的序列是5'-AGCTGGAAGTGCAGACC-3' (SEQ ID NO:20)。设计的总长度为1048bp。设计的三个版本按gBlocks(基因块)的顺序订购,分别称为gBlock DNA_NC_1,gBlock DNA_NC_2和gBlock DNA_NC_3。gBlock DNA_NC_1的XhoI消化产生3个片段:DNA_NC_1.1,DNA_NC_1.2和DNA_NC_1.3。gBlock DNA_NC_2的消化产生3个片段:DNA_NC_2.1,DNA_NC_2.2和DNA_NC_2.3。gBlock DNA_NC_3的消化产生3个片段:DNA_NC_3.1,DNA_NC_3.2和DNA_NC_3.3。片段DNA_NC_1.1、DNA_NC_2.1和DNA_NC_3.1分别包含Mito阵列引物的DNA序列,200bp的随机序列(各自不同)和XhoI半位点。片段DNA_NC_1.2、DNA_NC_2.2和DNA_NC_3.2包含300bp的随机序列,侧翼有两个XhoI半位点。片段DNA_NC_1.3、DNA_NC_2.3和DNA_NC_3.3包含XhoI半位点,500bp随机序列,T7启动子序列和Ribo阵列引物序列。

[1080] 在提取基因组DNA之前,将归一化对照加到样本中。然后将归一化对照与测序文库的其余部分并行索引。

[1081] 具有归一化对照的DNA文库使用i7和i5衔接子的双重索引策略制作。如下所述的具有i7和i5序列的引物以200μM的浓度重新悬浮。5μL的i5和i7引物在NEB缓冲器2中以20μL的总体积混合(图8)。然后将i5和i7引物混合物放入热循环仪中,升温至95℃持续1分钟,然后缓慢冷却至20℃。然后将i5/i7混合物在1200μL的总体积中稀释至15μM的浓度(用于1个反应)。将i7/i5衔接子进一步按1:20稀释,以达到750nM的总浓度。

[1082] 双索引引物

[1083] i701-UM16引物:5' phos/GAT CGG AAG AGC ACA CGT CTG AAC TCC AGT CAC TAC AGG TCN NNN NNA TCT CGT (SEQ ID NO:21)

[1084] i702-UM16引物:5' phos/GAT CGG AAG AGC ACA CGT CTG AAC TCC AGT CAC AGT TAC ATN NNN NNA TCT CGT (SEQ ID NO:22)

[1085] i501-UM16引物:AAT GAT ACG GCG ACC GAG ATC TAC ACN NNN NNC CAG TAC AAC

ACT CTT TCC CTA (SEQ ID NO:23)

[1086] i502-UM16引物:5' AAT GAT ACG GCG ACC ACC GAG ATG TAC CAN NNN NNG TGC TTA TAC ACT CTT TGG CTA N (SEQ ID NO:24)

[1087] 图9显示在示例性DNA文库中用包括的NC测序的NC读段百分比。将NC对照以图12所示的浓度加到样本。

[1088] 实施例3:量化样本中的非宿主的水平

[1089] 在该实施例中,确定非宿主与NC的比率用于减去混合的宿主和非宿主核酸样本中非宿主量化的不同宿主背景量的作用。

[1090] NC由一组核酸组成,分别是DNA或RNA或DNA和RNA二者的混合物。混合各种大小的NC以提供不同长度和基础成分的线性分布,以模拟NGS文库内的不同可变性。NC中包含的这些核酸可以合成制造,或由天然核酸的培养混合物制造。片段可以代表非常特定的大小,核酸组成,或涵盖各种浓度的非常大范围的片段大小。可以包含序列基序以了解文库处理是否因序列重复或重复序列而有偏差

[1091] NC是一种或多种生物。例如,NC包含从一种或多种生物或生物物种,例如病毒,细菌,真菌或真核生物中提取的多核苷酸,其模拟混合的宿主/非宿主样本中非宿主序列的一种或多种特性。

[1092] NC用于确定样本中的不同核酸的相对量。例如,归一化对照用于确定来自被感染宿主(例如,被病毒或细菌感染的人类宿主)的样本中病原体(非宿主)的丰度值。通过将映射到病原体的NGS读段除以映射到NC的NGS读段的数目来确定病原体的丰度值,以补偿背景宿主的不同含量。该值被映射到标准曲线,该曲线以相同的方式生成和归一化,并从标准曲线推断出病原体滴度。使用这样的方法,不一定需要确定绝对病原体负荷。

[1093] NC用于确定样本中的不同核酸的绝对量。例如,NC用于确定来自被感染宿主的样本(例如,来自被病毒或细菌感染的人类宿主的样本)中病原体(非宿主)的绝对量。通过将映射到病原体的NGS读段的数目除以映射到NC的NGS读段的数目来补偿不同的背景宿主含量,可以确定病原体的丰度值。基于该比率和NC的输入量(例如,噬菌体或多核苷酸的拷贝),确定相对病原体基因组拷贝。然后针对病原体基因组大小校正该值以在起始材料中得出“绝对”病原体丰度。

[1094] 在一些情况下,NC包含T4噬菌体核酸序列,而病原体包含巨细胞病毒。由于T4噬菌体和CMV基因组都在200kb左右,两者的大小大致相同,并且假定噬菌斑形成单位/mL (pfu/mL, T4) 和拷贝/mL (cp/mL, CMV) 为相等的。如果在下一代测序后,每1份T4噬菌体NC读段存在2份CMV读段,并且以100pfu/mL的浓度加入NC,则CMV的浓度必须为200cp/mL。如果CMV是T4基因组大小的1/2,则CMV和T4在初始样本中的浓度大致相同(100pfu/mL或cp/mL)。

[1095] 实施例4:使用多分析物对照确定病毒滴度

[1096] 本公开的归一化对照可以用于确定多分析物对照中的一组病毒滴度,用于计算实验样本中的病毒滴度。

[1097] 在该实施例中,多分析物对照是人类移植样本中常见的11种病毒的混合物:巨细胞病毒(CMV),爱泼斯坦-巴尔病毒(EBV),腺病毒(ADV),BK病毒(BKV),JC多瘤病毒(JCV),人疱疹病毒6A(HHV6A),人疱疹病毒6B(HHV6B),单纯疱疹病毒1型(HSV1),单纯疱疹病毒2型(HSV2),水痘带状疱疹病毒(VZV)和人细小病毒B19(B19)。使用11种病毒的混合物来构建不

同滴度(例如0、100、1,000和10,000感染单位(IU)/mL)的线性面板,并掺入血浆中。可以将多分析物对照掺入不同水平的血浆中,以生成多分析物对照的校准面板。将归一化对照加到校准面板中的样本和多分析物对照,提取核酸,然后从实验样本和多分析物对照中生成一个或多个高通量测序文库。在多分析物对照面板或归一化对照中,将读段映射到不同的病毒,并且归一化对照用于对来自多分析物对照中的每种病毒的读段计数进行归一化。然后针对多分析物对照将归一化读段计数相对于滴度绘图以产生标准曲线(图13),并计算归一化读段与 $\log(\text{IU}/\text{mL})$ 之间的关系方程。在这个实施例中(图13),这种关系是线性的。将来自实验样本的病毒信号类似地用归一化对照进行归一化,并且从多分析物对照计算的关系用于计算实验样本中的病毒滴度。

[1098] 该多分析物对照也用作阳性外部对照,它与实验样本分开进行整个样本制备过程。高和低浓度的多分析物对照也可以与实验样本一起包含以确保先前的校准面板(例如参考面板)仍适用于当前条件。

[1099] 实施例5:使用序列改组生成归一化对照序列

[1100] 设计归一化对照的一种方法是通过基础改组方法(Cantor改组)。该设计过程的图在图14A中显示。在改组方法中,靶序列的所有相同碱基均以相同比率保留,并且靶序列的参数被保留,但归一化对照的序列被改组且未映射到靶序列。

[1101] 代表潜在病原体的约1200-2000个参考基因组用于生成GC含量、熵、复杂性和电子-离子相互作用潜能(EIIP)相对于基因组长度的分布。在图15中显示三元图,其显示参考基因组的GC含量、熵和EIIP之间的相互作用。

[1102] 序列从长度箱获得,长度箱相对于不同分布(GC含量,熵,复杂性,EIIP)在x轴(长度)的第10,30,50,70,80和90百分位。每个序列在其内部通过20-25个mer窗口改组,并放置在序列库中以形成潜在阵列。每个序列还通过100个mer窗口在其内部改组,并放置在序列库中,以用作“映射对照”。

[1103] 将潜在阵列和“映射对照”与参考基因组进行分区,并针对NCBI非冗余序列数据库(NR)和引导序列进行blast比对。消除潜在阵列中映射的序列。归一化对照设计为55个交错浓度。

[1104] 归一化对照通过信息学方法掺入Probit(1000cp/ml),临床样本,识别和量化寡核苷酸,识别和量化病原体。将“映射对照”序列与参考进行映射以证明改组有效。

[1105] 在图14B-14C中显示以每对数浓度1-3个寡核苷酸设计的归一化对照的示例。在图14D中显示示例性归一化对照,其设计为浓度的每个数量级由6个浓度表示,每个对数浓度具有至少一个不同的寡核苷酸,并且覆盖7-8个数量级。

[1106] 图16-17显示使用参考基因组面板生成的Kmer中EIIP、熵和GC含量的分布。生成32亿个Kmer,并使用20亿个Kmer(图16)、1000万个Kmer(图17A)和130万个Kmer(图17B)构建初始配置文件。将Kmer改组以模拟参考基因组中的EIIP分布、熵和GC含量。

[1107] python中的多重处理可用于对Kmer数据集(例如Process和Pool)执行处理。Pandas和Process用于对大文件执行类似的操作。python中的Datamash用于12亿行Kmer(分为7个文件)以获取功能的百分位值。AWK用于执行选择并将选择的Kmer写入文件,该文件用于构建寡核苷酸。

[1108] 在图18中显示用于处理Kmer并产生输出寡核苷酸的示例性管线。对Kmer进行索引

以显示它们已在归一化对照寡核苷酸构建中使用了多少次并启动全局变量DIST_RANGE={1KB...10KB}。检查(归一化对照寡核苷酸的)输出数据帧,以查看长度和位置是否在DIST_RANGE内,以及生成的寡核苷酸的数量。如果为“否”,则选择未覆盖的长度箱,然后从步骤1中创建的索引中选择使用最少的Kmer。将这些连接,直到达到设计的归一化对照长度,然后重复该过程,直到生成所需数量的寡核苷酸。然后将序列传递给Shuffle函数。

[1109] 图19A-19D显示由Cantor改组产生的归一化对照模拟BK多瘤病毒分离株CH-1,JC多瘤病毒株NIID 12-31,人多瘤病毒1株BK 2和人腺病毒C的EIIP、复杂性和GC含量。图20A-20D显示使用生物技术信息中心(NCBI)BLAST核苷酸序列服务器(blast.ncbi.nlm.nih.gov/Blast.cgi?PAGE_TYPE=BlastSearch),归一化对照不会对BK多瘤病毒分离株CH-1,JC多瘤病毒株NIID 12-31,人多瘤病毒1株BK 2和人腺病毒C产生Blast。

[1110] 实施例6:使用滑动窗口生成归一化对照序列(穷举Kmer方法)

[1111] 实施例5中描述的Cantor改组的替代方法是使用来自靶序列的自然序列。在这种方法中,在自然序列上采用滑动窗口以生成交错的、重叠的Kmer。然后将不相邻的Kmer连接以产生归一化对照序列。如果使用约31或32bp的Kmer大小,则约150bp的测序读段将包含多个Kmer,并且可以通过在靶序列中不相邻的Kmer的并列识别归一化对照序列。

[1112] 在这种方法中,可以通过选择分布在整个基因组中的多个序列来表示单个基因组,该基因组代表该基因组如何片段化。此外,各个归一化对照可以包含来自多种生物的Kmer,并匹配多种生物的参数。例如,归一化对照寡核苷酸可以包含长Kmer,并且在长Kmer内可以是包含来自其他生物的序列的较短Kmer。

[1113] 约1200-2000个参考基因组用于生成交错的Kmer,并且如上文实施例5所述计算Kmer分布。

[1114] 首先基于GC含量百分比选择Kmer:即,选择一定数量的GC含量百分比以将其包括在归一化对照中。接下来,将处于选择GC含量百分比的Kmer连接以在GC箱中生成1千碱基(KB)的寡核苷酸。检查连接序列的GC含量,以及熵。然后将序列通过均聚物过滤器,并进行blast比对(Blastn和guide blast)以产生785,000个寡核苷酸,每个寡核苷酸的长度为1Kb。

[1115] 图21A-21C显示在选择的GC含量百分比下产生的Kmer的EIIP、熵和GC含量的分布。当对参考基因组序列和使用这些方法生成的归一化对照序列进行主成分分析(PCA)时,两者在熵、EIIP和GC含量之间都显示出类似的相关性(图22)。此外,使用该方法产生的归一化对照的熵、EIIP和GC含量与参考基因组序列的熵、EIIP和GC含量相关(图23)。另外,进行Kolmogorov Smirnov (KS) 测试,比较参考基因组序列与归一化对照的熵的概率分布(图24)。

[1116] 使用NCBI BLAST核苷酸序列服务器通过BLAST验证该785,000个归一化对照寡核苷酸序列与参考基因组序列不匹配(图25A-25B)。对于50-80个碱基,归一化对照寡核苷酸的头部总是以约750bp对Pseudomonas进行BLAST处理(图25A)。文件的尾部没有BLAST命中(图25)。可以对所有785,000个归一化对照寡核苷酸序列进行大规模BLAST处理,并且只有那些没有BLAST命中的序列才能被选择用作归一化对照。

- [0001] 序列表
- [0002] <110> 阿克生物公司(Arc Bio, LLC)
- [0003] <120> 用于管理下一代测序中的低样本输入的归一化对照
- [0004] <130> ARCB-006/02W0
- [0005] <150> 62/741,466
- [0006] <151> 2018-10-04
- [0007] <150> 62/832,560
- [0008] <151> 2019-04-11
- [0009] <160> 24
- [0010] <170> PatentIn第3.5版
- [0011] <210> 1
- [0012] <211> 22
- [0013] <212> DNA
- [0014] <213> 人工序列(Artificial Sequence)
- [0015] <220>
- [0016] <223> crRNA编码序列
- [0017] <400> 1
- [0018] gtttttagagc tatgctgttt tg 22
- [0019] <210> 2
- [0020] <211> 86
- [0021] <212> DNA
- [0022] <213> 人工序列(Artificial Sequence)
- [0023] <220>
- [0024] <223> tracrRNA编码序列
- [0025] <400> 2
- [0026] ggaaccattc aaaacagcat agcaagttaa aataaggcta gtccgttatc aacttgaaaa 60
- [0027] agtggcaccg agtcggtgct tttttt 86
- [0028] <210> 3
- [0029] <211> 83
- [0030] <212> DNA
- [0031] <213> 人工序列(Artificial Sequence)
- [0032] <220>
- [0033] <223> gNA茎环
- [0034] <400> 3
- [0035] gtttttagagc tagaaatagc aagttaaaat aaggctagtc cgttatcaac ttgaaaaagt 60
- [0036] ggcaccgagt cgggtgctttt ttt 83
- [0037] <210> 4
- [0038] <211> 83

- [0039] <212> DNA
- [0040] <213> 人工序列(Artificial Sequence)
- [0041] <220>
- [0042] <223> gNA茎环反向互补序列
- [0043] <400> 4
- [0044] aaaaaaagca ccgactcggg gccacttttt caagttgata acggactagc cttattttta 60
- [0045] cttgctatatt ctagctctaa aac 83
- [0046] <210> 5
- [0047] <211> 83
- [0048] <212> RNA
- [0049] <213> 人工序列(Artificial Sequence)
- [0050] <220>
- [0051] <223> gNA茎环
- [0052] <400> 5
- [0053] guuuuagagc uagaaauagc aaguuaaaau aaggcuaguc cguuaucaac uugaaaaagu 60
- [0054] ggcaccgagu cggugcuuuu uuu 83
- [0055] <210> 6
- [0056] <211> 94
- [0057] <212> DNA
- [0058] <213> 人工序列(Artificial Sequence)
- [0059] <220>
- [0060] <223> gNA茎环
- [0061] <400> 6
- [0062] gtttttagagc tatgctggaa acagcatagc aagttaaaat aaggctagtc cgttatcaac 60
- [0063] ttgaaaaagt ggcaccgagt cgggtgctttt tttc 94
- [0064] <210> 7
- [0065] <211> 94
- [0066] <212> DNA
- [0067] <213> 人工序列(Artificial Sequence)
- [0068] <220>
- [0069] <223> gNA茎环反向互补序列
- [0070] <400> 7
- [0071] gaaaaaaagc accgactcgg tgccactttt tcaagttgat aacggactag ccttattttta 60
- [0072] acttgctatg ctgtttccag catagctcta aaac 94
- [0073] <210> 8
- [0074] <211> 94
- [0075] <212> RNA
- [0076] <213> 人工序列(Artificial Sequence)
- [0077] <220>

- [0078] <223> gNA茎环
- [0079] <400> 8
- [0080] guuuuagagc uaugcuggaa acagcauagc aaguuaaaau aaggcuaguc cguuaucaac 60
- [0081] uugaaaaagu ggcaccgagu cggugcuuuu uuuc 94
- [0082] <210> 9
- [0083] <211> 19
- [0084] <212> RNA
- [0085] <213> 人工序列(Artificial Sequence)
- [0086] <220>
- [0087] <223> Cpf1蛋白结合序列
- [0088] <400> 9
- [0089] aauuucuacu guuguagau 19
- [0090] <210> 10
- [0091] <211> 19
- [0092] <212> DNA
- [0093] <213> 人工序列(Artificial Sequence)
- [0094] <220>
- [0095] <223> 编码Cpf1蛋白结合序列
- [0096] <400> 10
- [0097] aatttctact gttgtagat 19
- [0098] <210> 11
- [0099] <211> 86
- [0100] <212> RNA
- [0101] <213> 人工序列(Artificial Sequence)
- [0102] <220>
- [0103] <223> tracrRNA序列
- [0104] <400> 11
- [0105] ggaaccuuc aaaacagcau agcaaguuaa aaauaggcua guccguuauac aacuugaaaa 60
- [0106] aguggcaccg agucggugcu uuuuuu 86
- [0107] <210> 12
- [0108] <211> 17
- [0109] <212> DNA
- [0110] <213> 人工序列(Artificial Sequence)
- [0111] <220>
- [0112] <223> 引物序列
- [0113] <400> 12
- [0114] tgaagaactg cggcagg 17
- [0115] <210> 13
- [0116] <211> 17

[0117]	<212> DNA
[0118]	<213> 人工序列(Artificial Sequence)
[0119]	<220>
[0120]	<223> 引物序列
[0121]	<400> 13
[0122]	ggtctgcact tccagct 17
[0123]	<210> 14
[0124]	<211> 19
[0125]	<212> DNA
[0126]	<213> 人工序列(Artificial Sequence)
[0127]	<220>
[0128]	<223> T7启动子
[0129]	<400> 14
[0130]	taatacgact cactatagg 19
[0131]	<210> 15
[0132]	<211> 20
[0133]	<212> DNA
[0134]	<213> 人工序列(Artificial Sequence)
[0135]	<220>
[0136]	<223> T7启动子
[0137]	<400> 15
[0138]	taatacgact cactataggg 20
[0139]	<210> 16
[0140]	<211> 29
[0141]	<212> DNA
[0142]	<213> 人工序列(Artificial Sequence)
[0143]	<220>
[0144]	<223> T7启动子
[0145]	<400> 16
[0146]	gcctcgagct aatacgactc actatagag 29
[0147]	<210> 17
[0148]	<211> 18
[0149]	<212> DNA
[0150]	<213> 人工序列(Artificial Sequence)
[0151]	<220>
[0152]	<223> SP6启动子
[0153]	<400> 17
[0154]	atttaggtga cactatag 18
[0155]	<210> 18

- [0156] <211> 24
[0157] <212> DNA
[0158] <213> 人工序列(Artificial Sequence)
[0159] <220>
[0160] <223> SP6启动子
[0161] <400> 18
[0162] catacgattt aggtgacact atag 24
[0163] <210> 19
[0164] <211> 18
[0165] <212> DNA
[0166] <213> 人工序列(Artificial Sequence)
[0167] <220>
[0168] <223> T3启动子
[0169] <400> 19
[0170] aattaaccct cactaaag 18
[0171] <210> 20
[0172] <211> 17
[0173] <212> DNA
[0174] <213> 人工序列(Artificial Sequence)
[0175] <220>
[0176] <223> 引物序列
[0177] <400> 20
[0178] agctggaagt gcagacc 17
[0179] <210> 21
[0180] <211> 54
[0181] <212> DNA
[0182] <213> 人工序列(Artificial Sequence)
[0183] <220>
[0184] <223> i701-UM16引物
[0185] <220>
[0186] <221> misc_特征
[0187] <222> (42) .. (47)
[0188] <223> n是a、c、g、或t
[0189] <400> 21
[0190] gatcggaaga gcacacgtct gaactccagt cactacaggt cnnnnnnatc tcgt 54
[0191] <210> 22
[0192] <211> 54
[0193] <212> DNA
[0194] <213> 人工序列(Artificial Sequence)

[0195] <220>
[0196] <223> i702-UM16引物
[0197] <220>
[0198] <221> misc_特征
[0199] <222> (42) .. (47)
[0200] <223> n是a、c、g、或t
[0201] <400> 22
[0202] gatcggaaga gcacacgtct gaactccagt cacagttaca tnnnnnnatc tcgt 54
[0203] <210> 23
[0204] <211> 54
[0205] <212> DNA
[0206] <213> 人工序列(Artificial Sequence)
[0207] <220>
[0208] <223> i501-UM16引物
[0209] <220>
[0210] <221> misc_特征
[0211] <222> (27) .. (32)
[0212] <223> n是a、c、g、或t
[0213] <400> 23
[0214] aatgatacgg cgaccgagat ctacacnnnn nccagtaca acactctttc ccta 54
[0215] <210> 24
[0216] <211> 58
[0217] <212> DNA
[0218] <213> 人工序列(Artificial Sequence)
[0219] <220>
[0220] <223> i502-UM6引物
[0221] <220>
[0222] <221> misc_特征
[0223] <222> (30) .. (35)
[0224] <223> n是a、c、g、或t
[0225] <220>
[0226] <221> misc_特征
[0227] <222> (58) .. (58)
[0228] <223> n是a、c、g、或t
[0229] <400> 24
[0230] aatgatacgg cgaccaccga gatgtaccan nnnnngtget tatacactct ttggetan 58

归一化对照

对照	读段	预期比率	实际	QC合格/不合格
GalileoNormC.100	124	-	-	-
GalileoNormC.1000	1,108	10x	8.9x	合格
GalileoNormC.10000	9,978	10x	9.0x	合格
GalileoNormC.100000	101,390	10x	10.2x	合格

图1

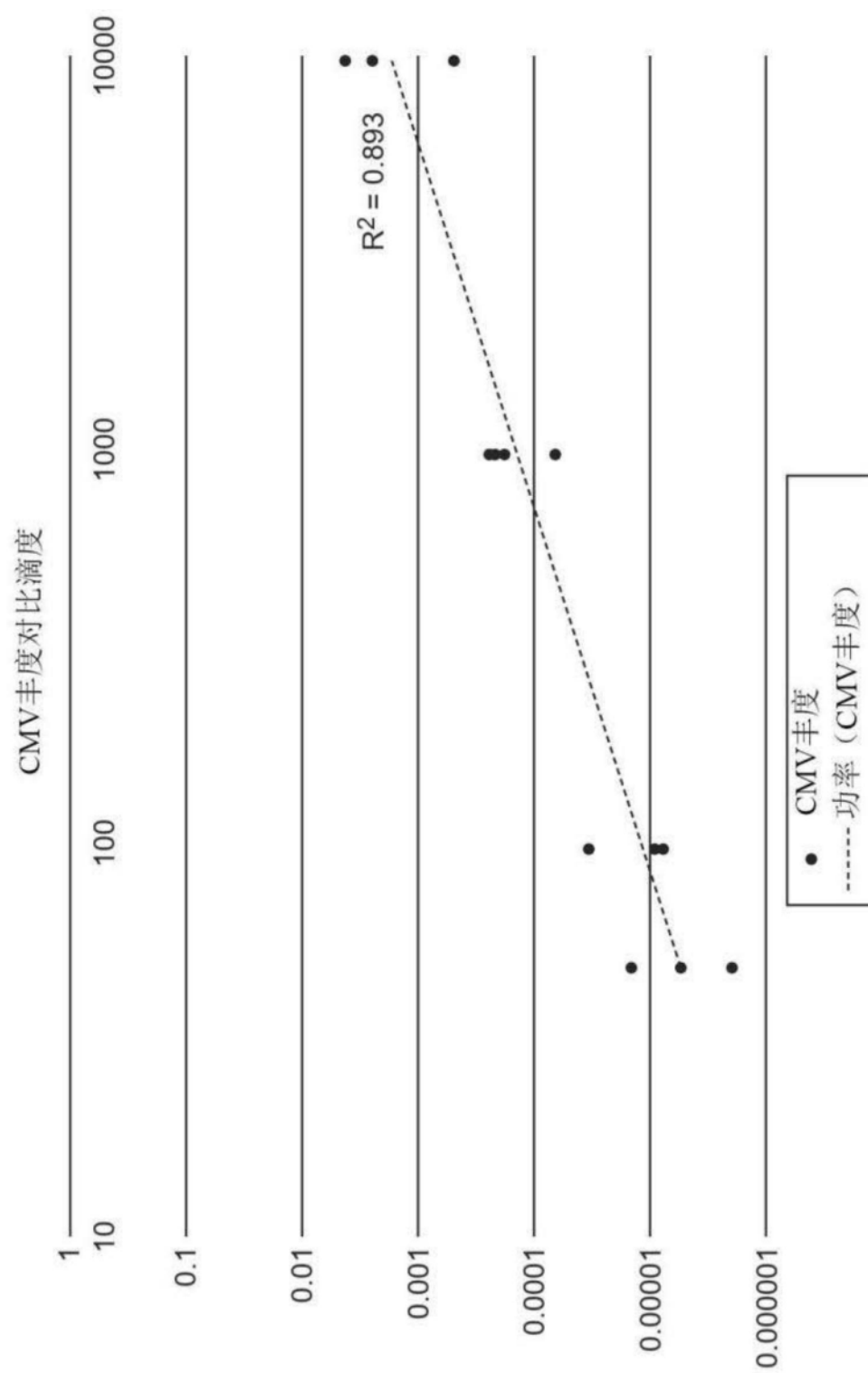


图2

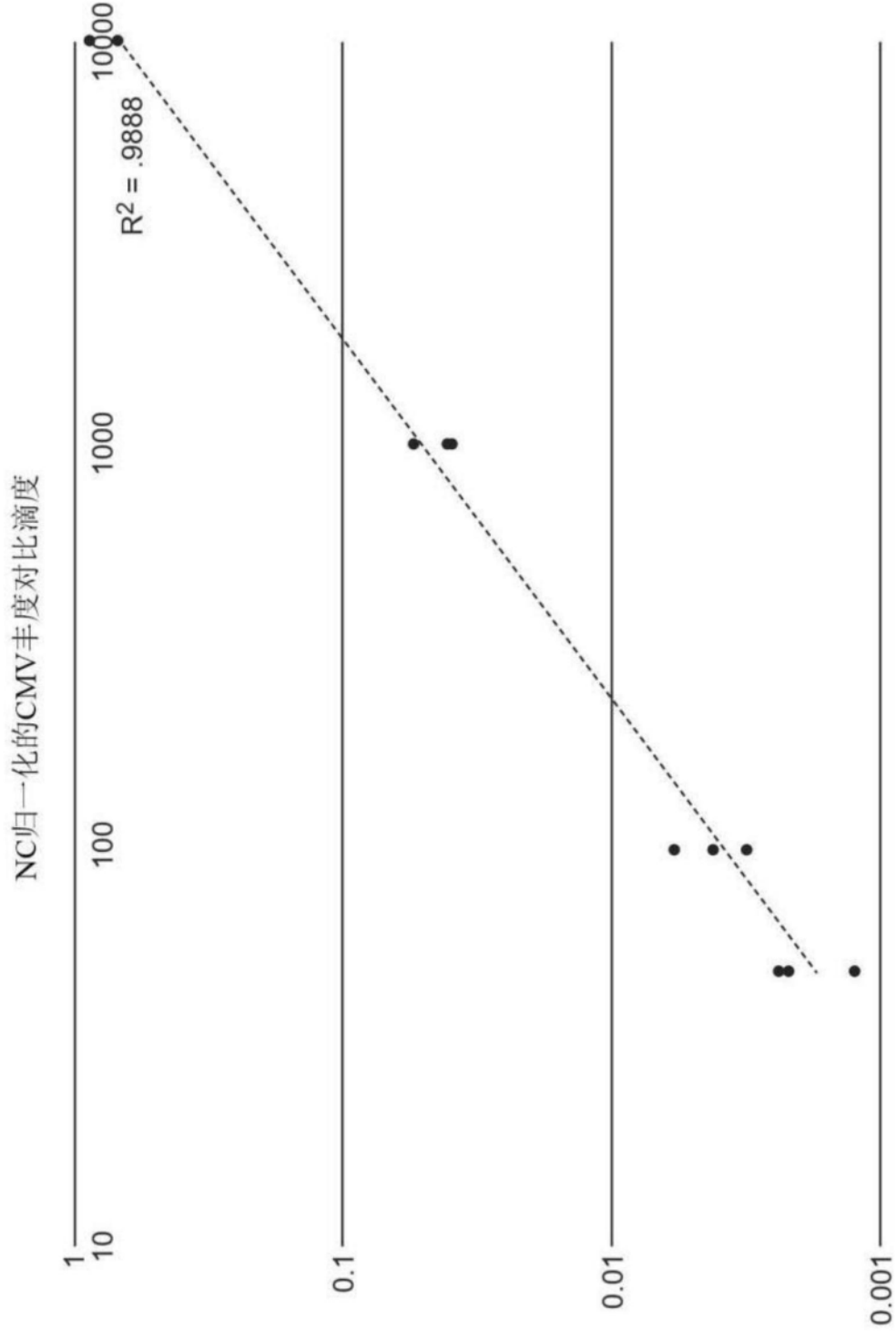


图3

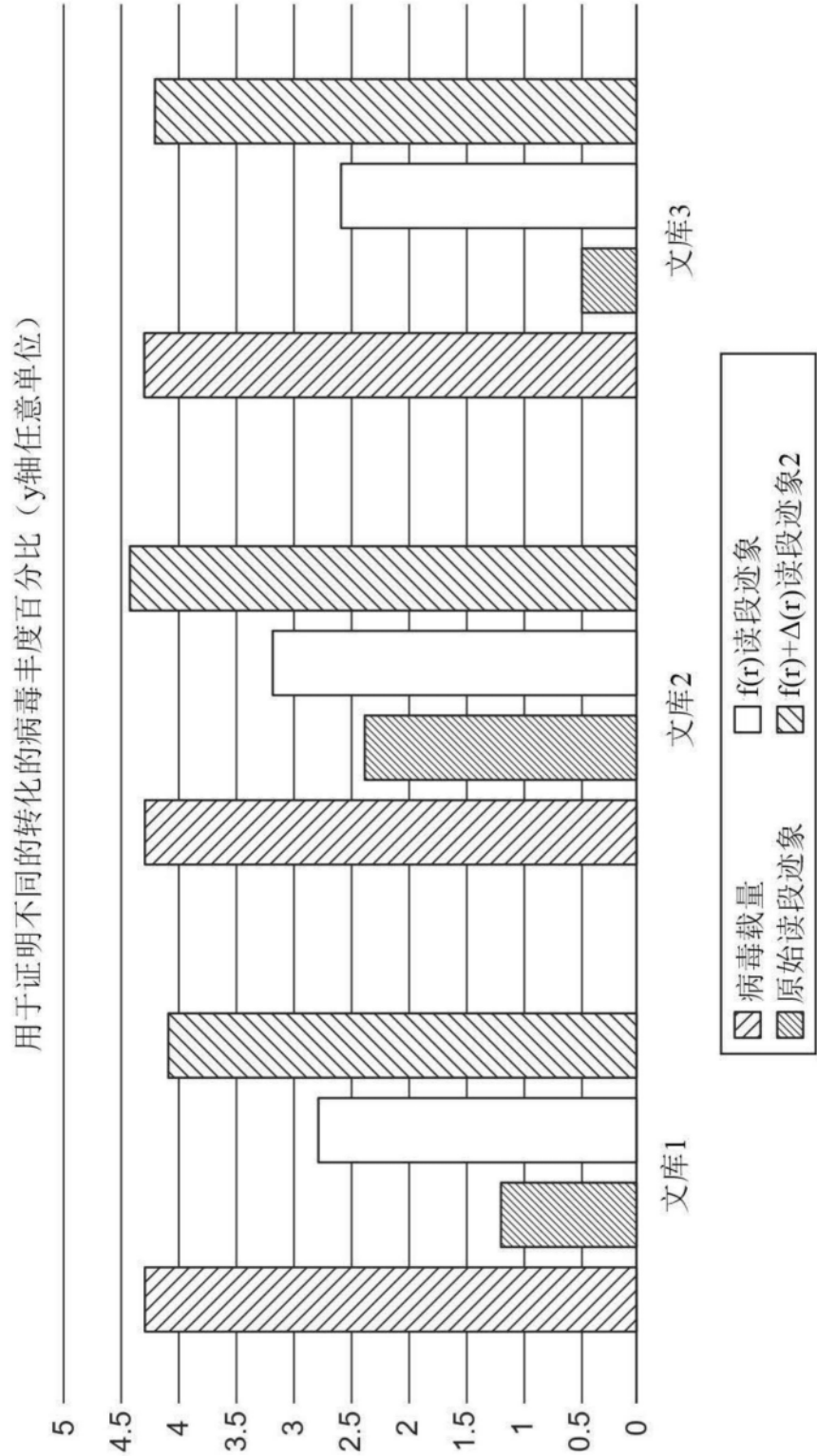


图4

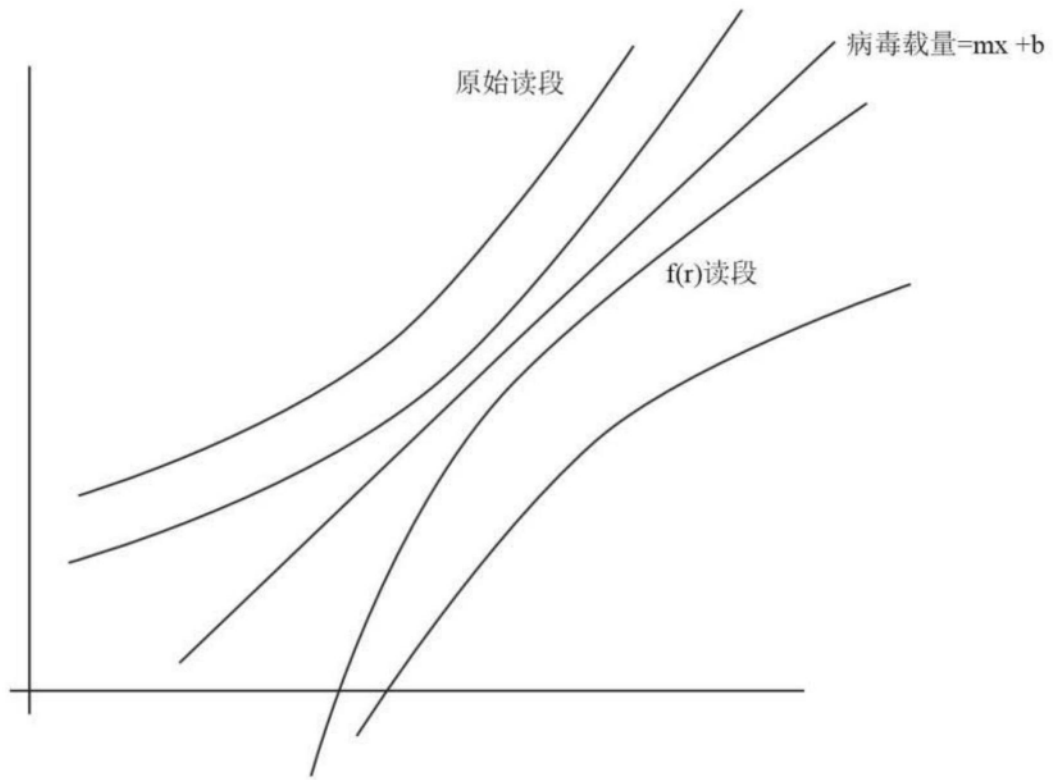


图5

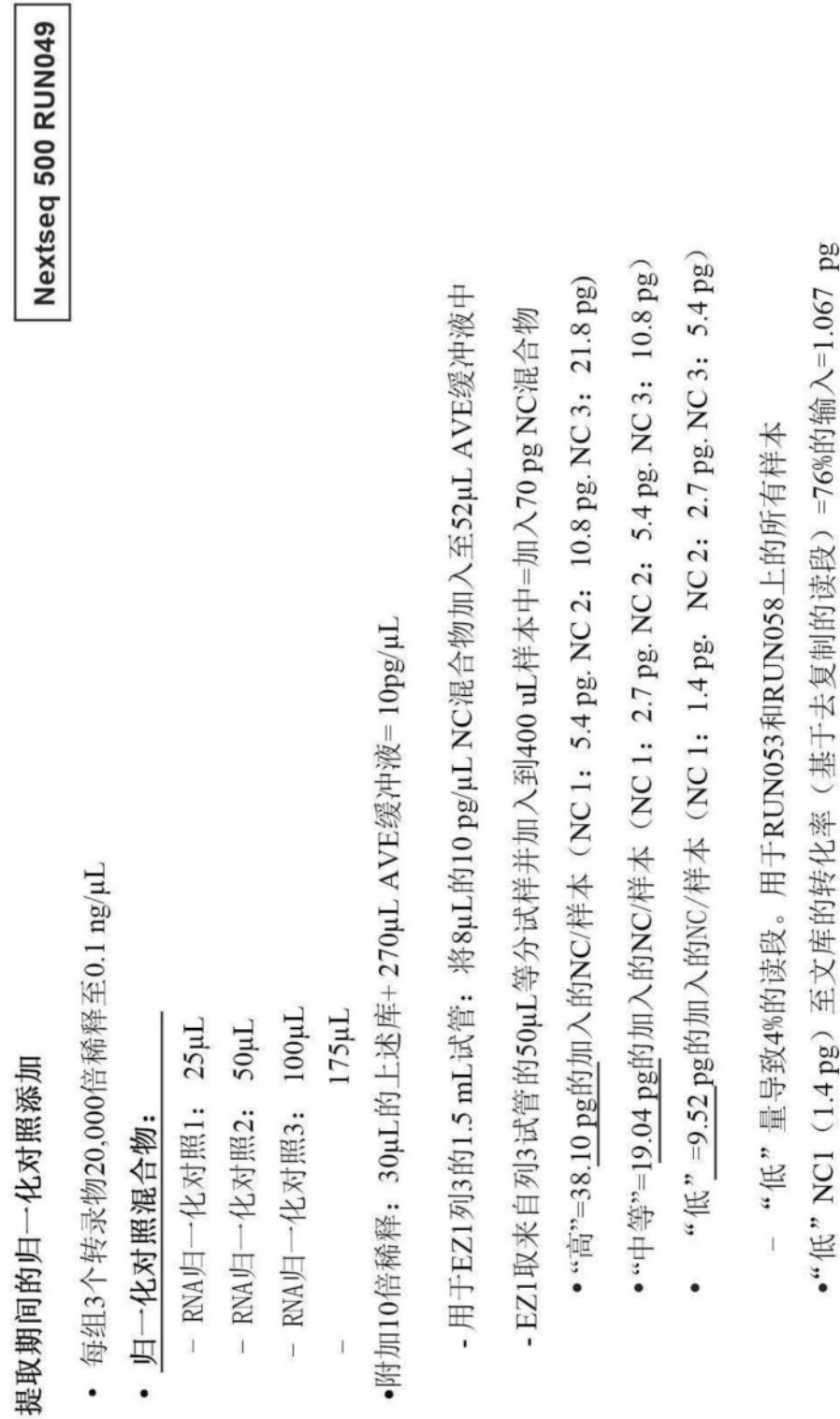


图6

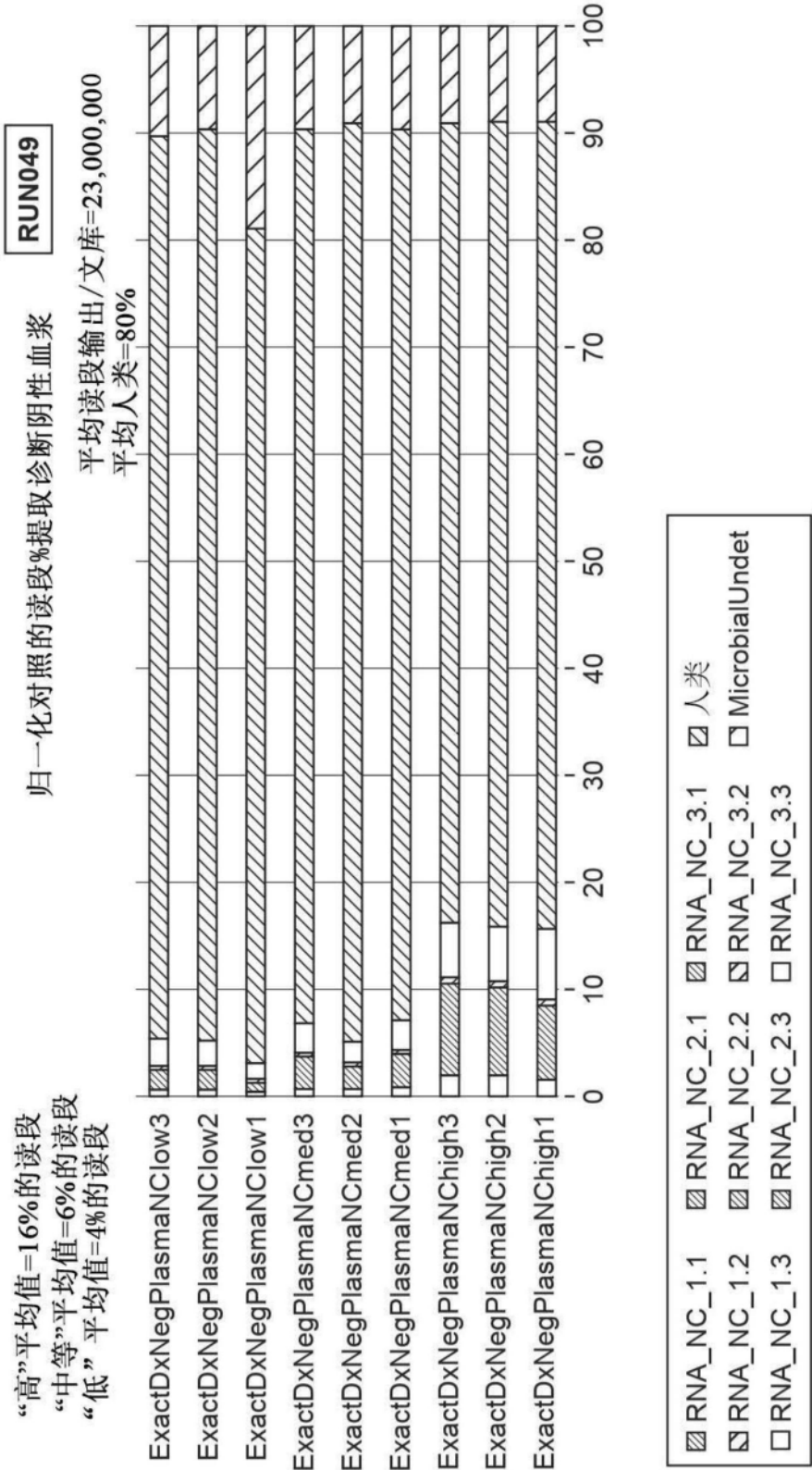


图7

衔接子混合										反应	2			
储备试剂	供应商	P/N编号	MW	[初始液体] 或克固体	UOM			RXN 1x 体积	UOM		生产规模 RXN	UOM	最终浓度 1X RXN	UOM
Mol Bio Wate	Sigma	793566-25KG		1	x			8	uL		16	uL	0.4	x
i5衔接子				200	uM			5	uL		10	uL	50	uM
i7衔接子				200	uM			5	uL		10	uL	50	uM
NEB 2				10	x			2	uL		4	uL	1	x
						总计		20	uL		40	uL		
从95℃缓慢上升1分钟至20℃														
稀释至15 uM														
归一化对照										反应	2			
储备试剂	供应商	P/N编号	MW	[初始液体] 或克固体	UOM			RXN 1x 体积	UOM		生产规模 RXN	UOM	最终浓度 1X RXN	UOM
Mol Bio Wate	Sigma	793566-25KG		1	x			42	uL		84	uL	0.7	x
i5 i7衔接子				50	uM			18	uL		36	uL	15	uM
						总计		60	uL		120	uL		
稀释至750 nM														
归一化对照										反应	2			
储备试剂	供应商	P/N编号	MW	[初始液体] 或克固体	UOM			RXN 1x 体积	UOM		生产规模 RXN	UOM	最终浓度 1X RXN	UOM
Mol Bio Wate	Sigma	793566-25KG		1	x			1140	uL		2280	uL	19	x
i5 i7衔接子				15	uM			60	uL		120	uL	0.75	uM
						总计		1200	uL		2400	uL		

图8

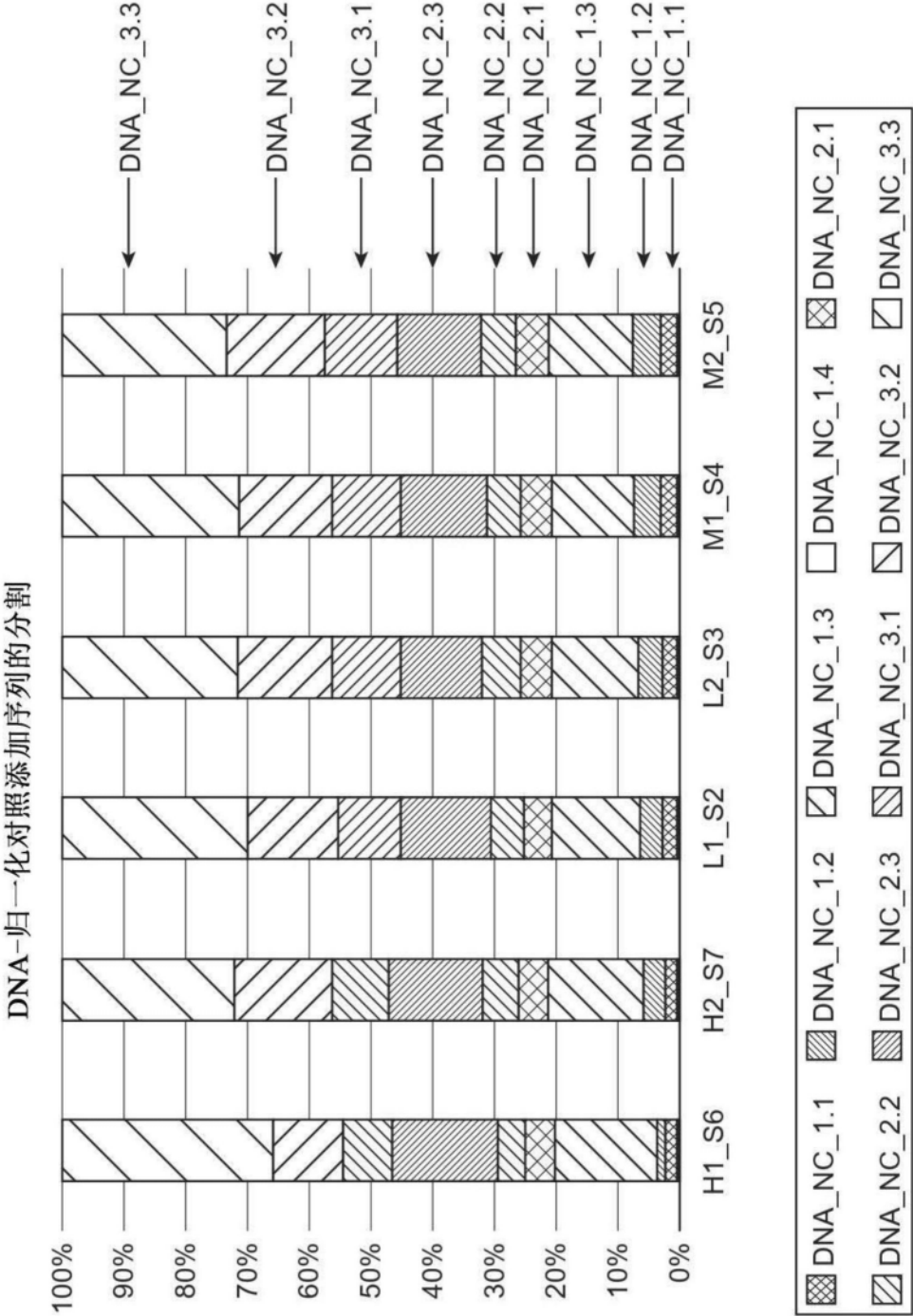


图9

Fasta表头	GC含量	总计数	长度
>Oligo177_RNA_Normalization_Controls_1_with_T7_RNA_polymerase_promoter_for_IVT	45.41284404	218	436
>Oligo178_RNA_Normalization_Controls_2_with_T7_RNA_polymerase_promoter_for_IVT	46.78899083	218	436
>Oligo179_RNA_Normalization_Controls_3_with_T7_RNA_polymerase_promoter_for_IVT	43.57798165	218	436
>Oligo180_RNA_Normalization_Controls_4_with_T7_RNA_polymerase_promoter_for_IVT	46.73913043	276	552
>Oligo181_RNA_Normalization_Controls_5_with_T7_RNA_polymerase_promoter_for_IVT	43.84057971	276	552
>Oligo182_RNA_Normalization_Control_6_with_T7_RNA_polymerase_promoter_for_IVT	40.57971014	276	552
>Oligo183_RNA_Normalization_Control_7_with_T7_RNA_polymerase_promoter_for_IVT	44.35318275	487	974
>Oligo184_RNA_Normalization_Control_8_with_T7_RNA_polymerase_promoter_for_IVT	43.32648871	487	974
>Oligo185_RNA_Normalization_Control_9_with_T7_RNA_polymerase_promoter_for_IVT	43.53182752	487	974
>Oligo186_RNA_Normalization_Control_1	49.77777778	225	450
>Oligo187_RNA_Normalization_Control_2	52	225	450
>Oligo188_RNA_Normalization_Control_3	49.77777778	225	450
>Oligo189_RNA_Normalization_Control_4	42.15686275	306	612
>Oligo190_RNA_Normalization_Control_5	43.46405229	306	612
>Oligo191_RNA_Normalization_Control_6	41.50326797	306	612
>Oligo192_RNA_Normalization_Control_7	45.64796905	517	1034
>Oligo193_RNA_Normalization_Control_8	46.61508704	517	1034
>Oligo194_RNA_Normalization_Control_9	44.87427466	517	1034
>Oligo241_rnanorm1seg1	47.05882353	255	510
>Oligo242_rnanorm2seg1	48.62745098	255	510
>Oligo243_rnanorm3seg1	53.7254902	255	510
>Oligo244_rnanorm1seg2	51.7571885	313	626
>Oligo245_rnanorm2seg2	51.7571885	313	626
>Oligo246_rnanorm3seg2	44.7284345	313	626
>Oligo247_rnanorm2seg3	50.57251908	524	1048
>Oligo248_rnanorm3seg3	49.42748092	524	1048
>Oligo249_rnanorm4seg3	48.85496183	524	1048
>Oligo250_dnanorm1seg1	46.66666667	225	450
>Oligo251_dnanorm2seg1	55.55555556	225	450
>Oligo252_dnanorm3seg1	54.22222222	225	450
>Oligo253_dnanorm1seg2	49.34640523	306	612
>Oligo254_dnanorm2seg2	50.65359477	306	612
>Oligo255_dnanorm3seg2	49.01960784	306	612
>Oligo256_dnanorm1seg3	49.9032882	517	1034
>Oligo257_dnanorm2seg3	48.54932302	517	1034
>Oligo258_dnanorm3seg3	48.54932302	517	1034

图10

寡核苷酸编号	最终浓度ng/ul	与DNA总量的比率
寡核苷酸186	4.20E-03	0.420
寡核苷酸187	4.21E-04	0.042
寡核苷酸188	2.11E-04	0.021
寡核苷酸189	2.86E-03	0.286
寡核苷酸190	4.29E-04	0.043
寡核苷酸191	1.43E-04	0.014
寡核苷酸192	9.68E-04	0.097
寡核苷酸193	4.84E-04	0.048
寡核苷酸194	9.68E-05	0.010
寡核苷酸250	4.22E-05	0.004
寡核苷酸251	2.11E-05	0.002
寡核苷酸253	4.31E-05	0.004
寡核苷酸254	1.43E-05	0.001
寡核苷酸256	4.84E-05	0.005
寡核苷酸257	9.71E-06	0.001
文库大小		

图11

组别	长度 (kb)	长度 (bp)	分子数	总attomole/片段	浓度 (attomole/uL)
RNA_NC_1.1 A	0.175	175	263158	0.44	0.07
RNA_NC_1.2 B	0.25	250	1144737	1.90	0.29
RNA_NC_1.3 C	0.45	450	5131579	8.52	1.28
RNA_NC_2.1 A	0.175	175	2789474	4.63	0.69
RNA_NC_2.2 B	0.25	250	4513157	7.49	1.12
RNA_NC_2.3 C	0.45	450	44473684	73.84	11.08
RNA_NC_3.1 A	0.175	175	4605263	7.65	1.15
RNA_NC_3.2 B	0.25	250	17500000	29.06	4.36
RNA_NC_3.3 C	0.45	450	77894737	129.33	19.40

图12

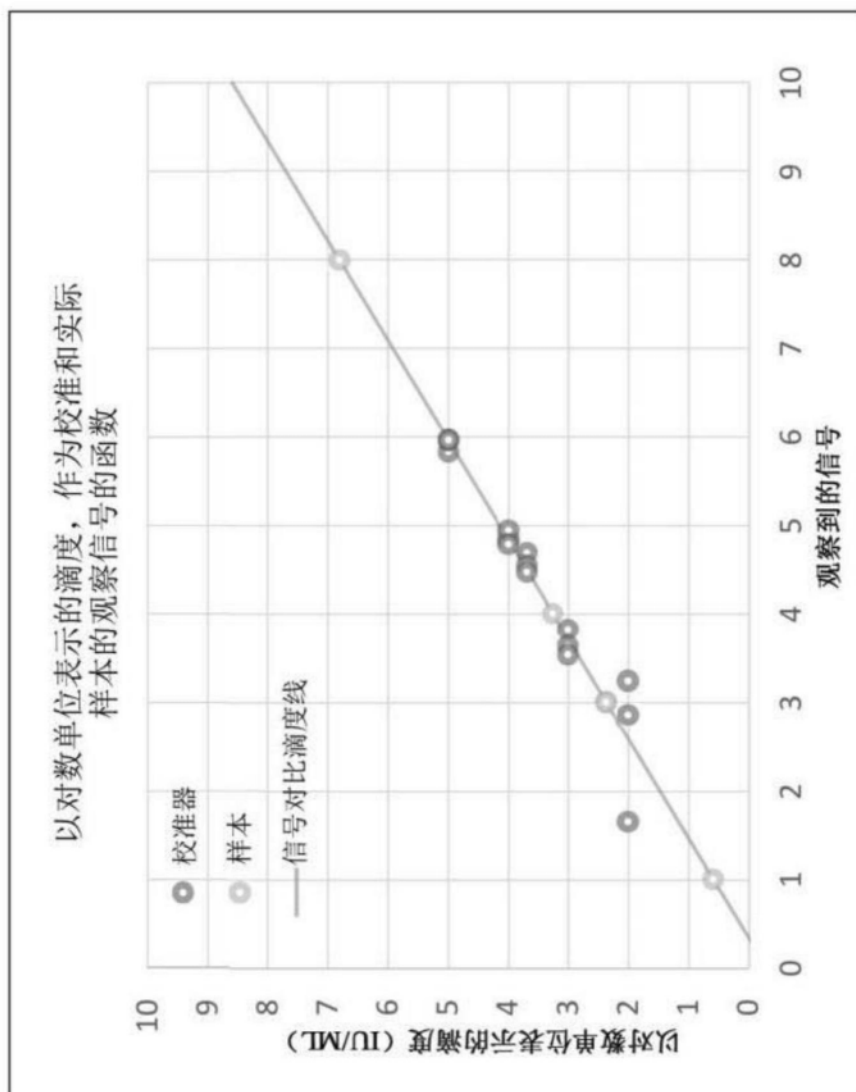


图13

示例性归一化对照设计管线

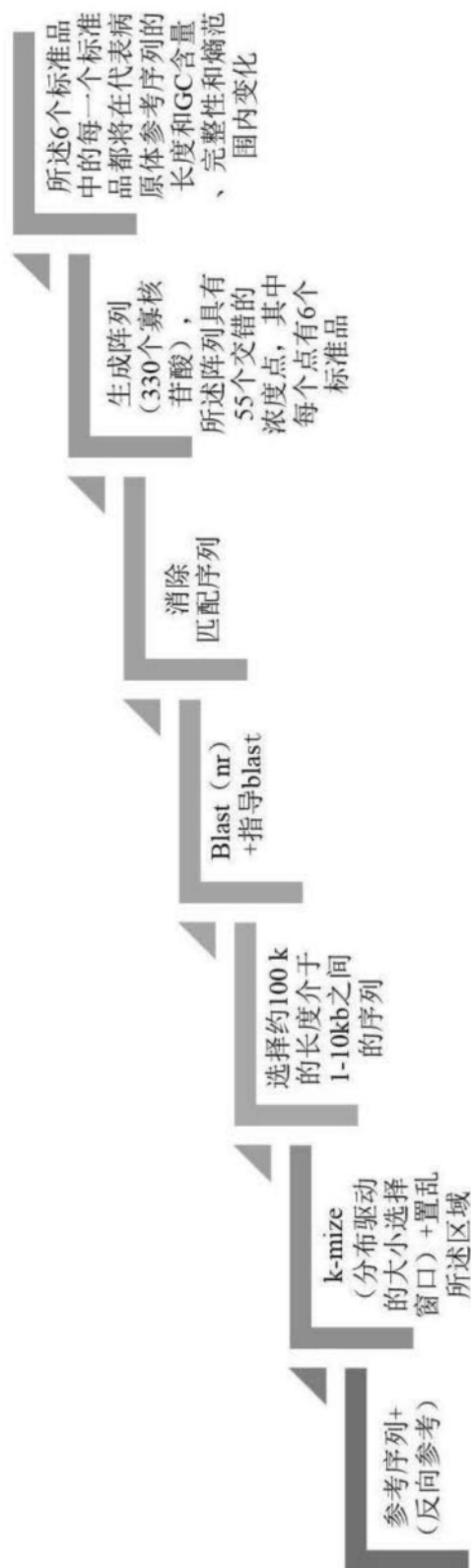


图14A

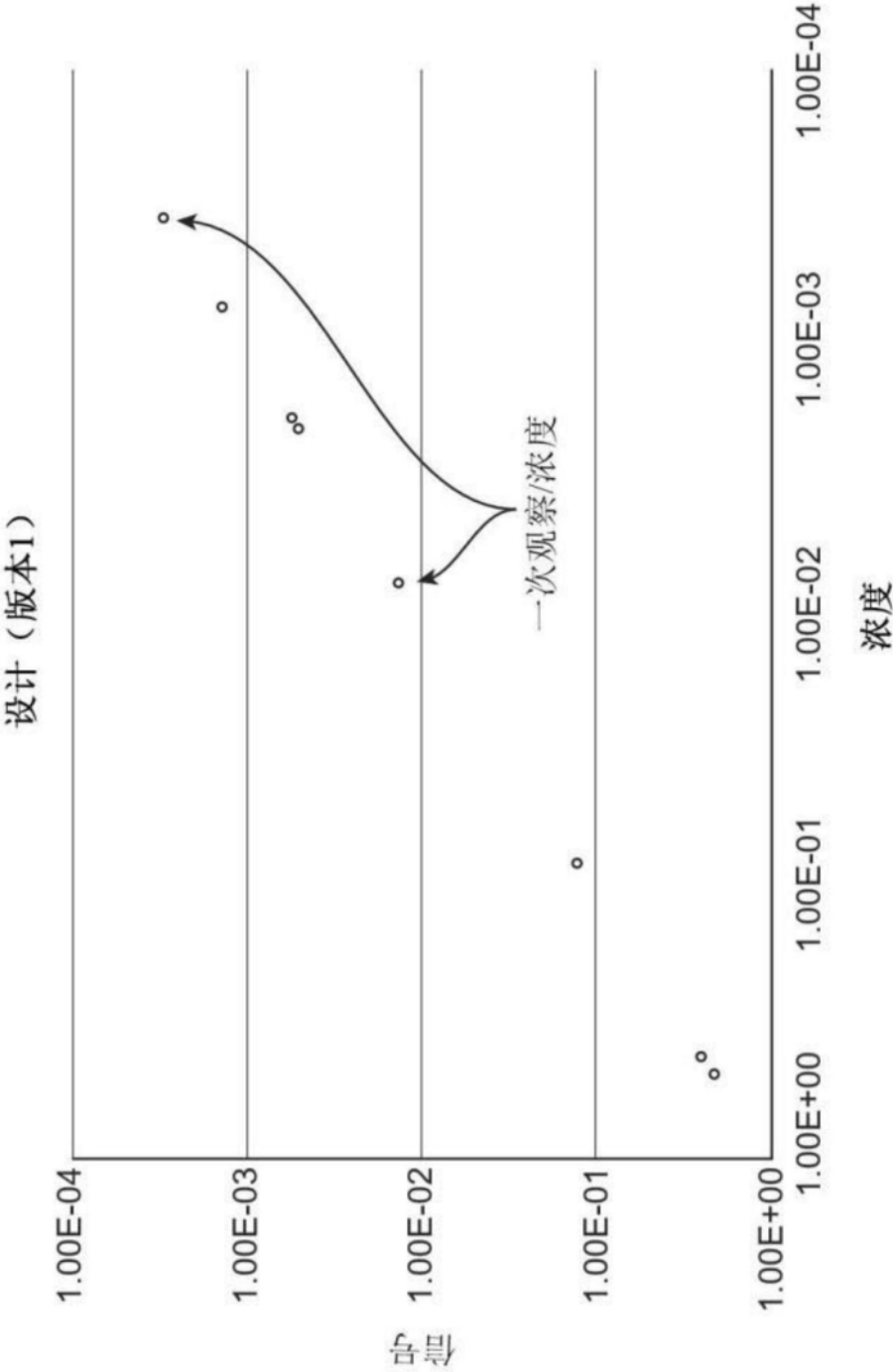


图14B

名称	文库	浓度输入	底部输出	浓度输出
Oligo298.dna	298	4.88E-01	无	4.88E-01
Oligo299.dna	299	4.19E-01	无	4.19E-01
Oligo301.dna	301	8.13E-02	无	8.13E-02
Oligo302.dna	302	2.04E-03	无	2.04E-03
Oligo303.dna	303	7.63E-03	无	7.63E-03
Oligo304.dna	304	1.87E-03	无	1.87E-03
Oligo305.dna	305	7.39E-04	底部输出	7.39E-04
Oligo306.dna	306	3.45E-04	底部输出	3.45E-04

图14C

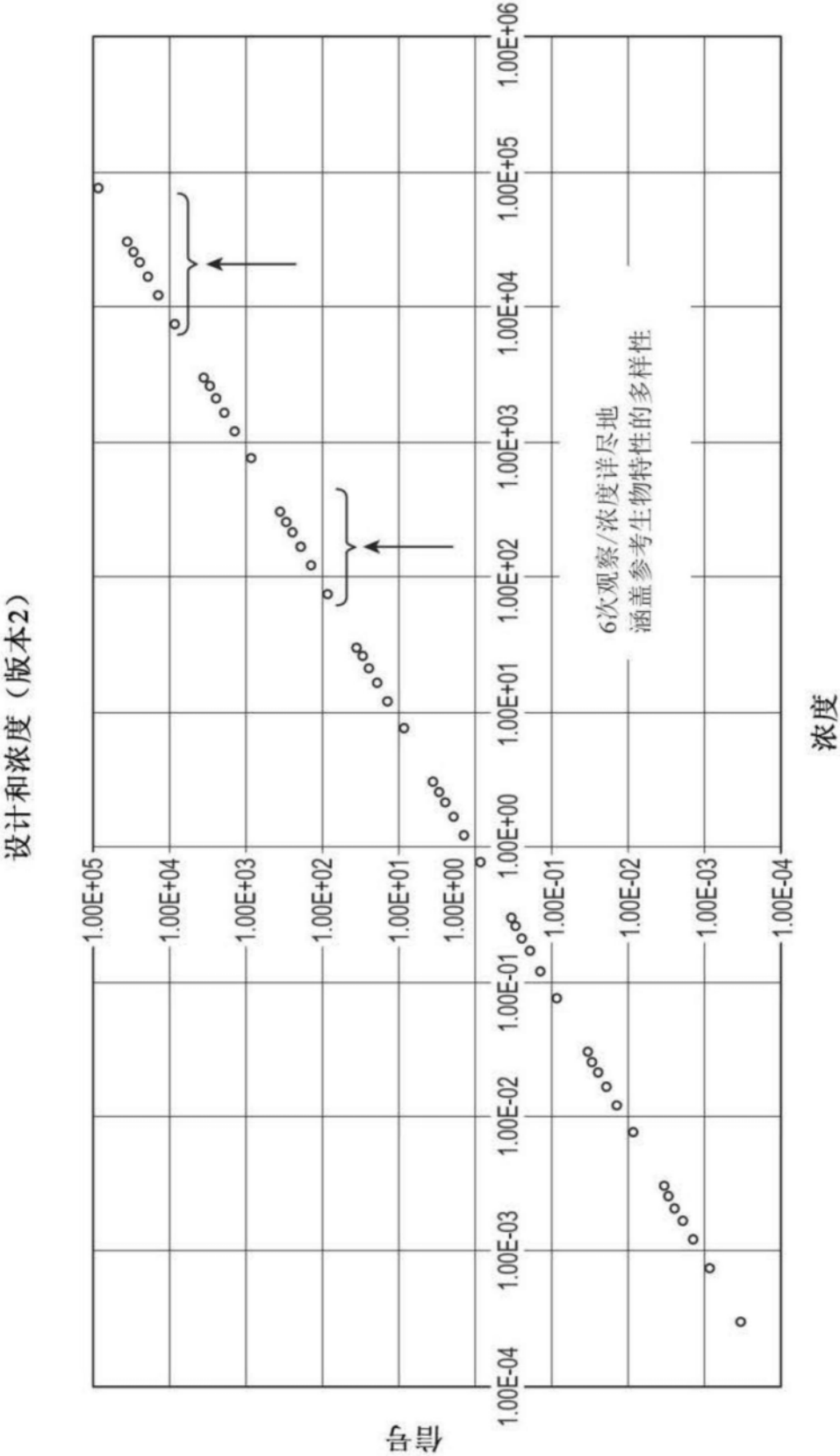


图14D

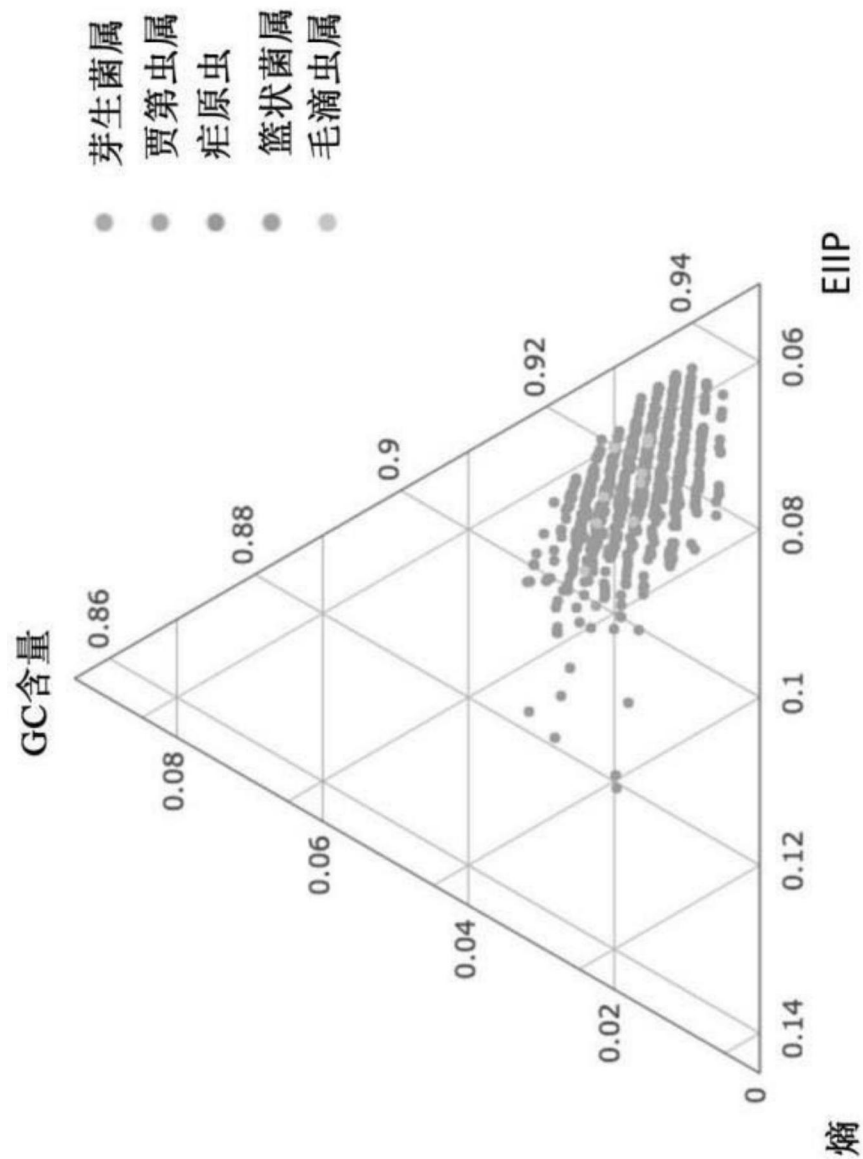


图15

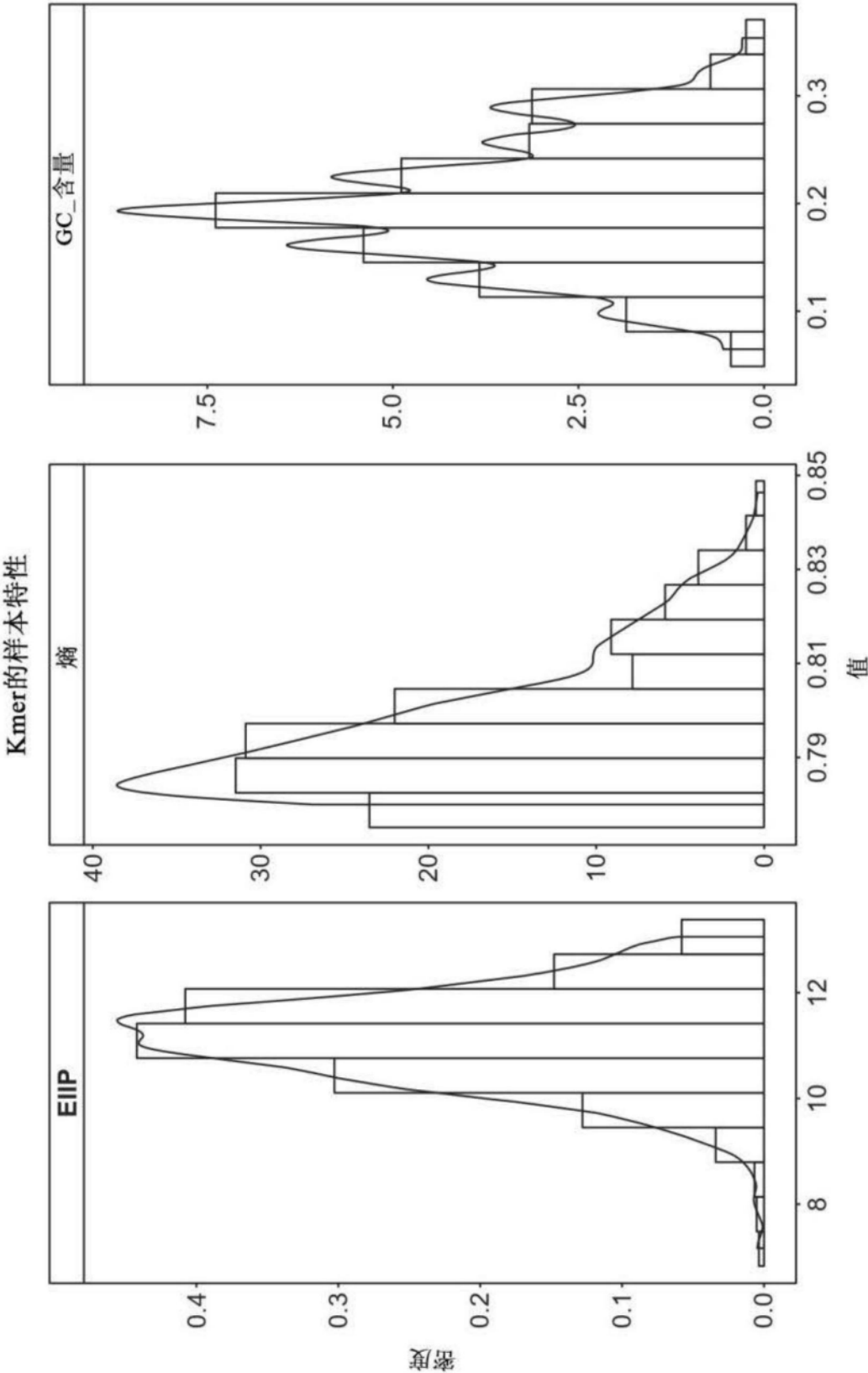


图16

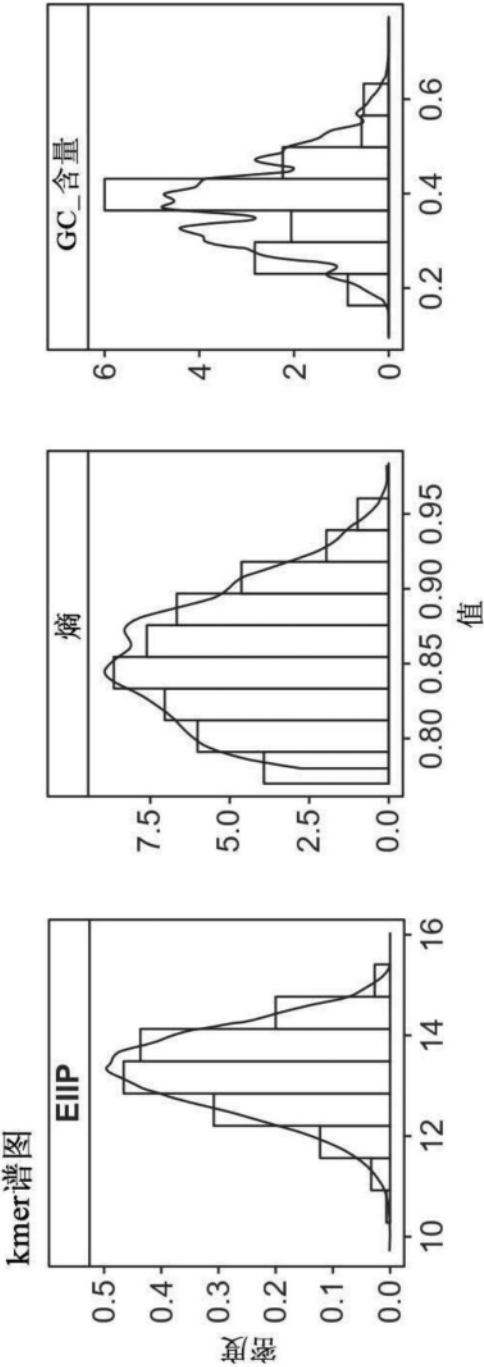


图17A

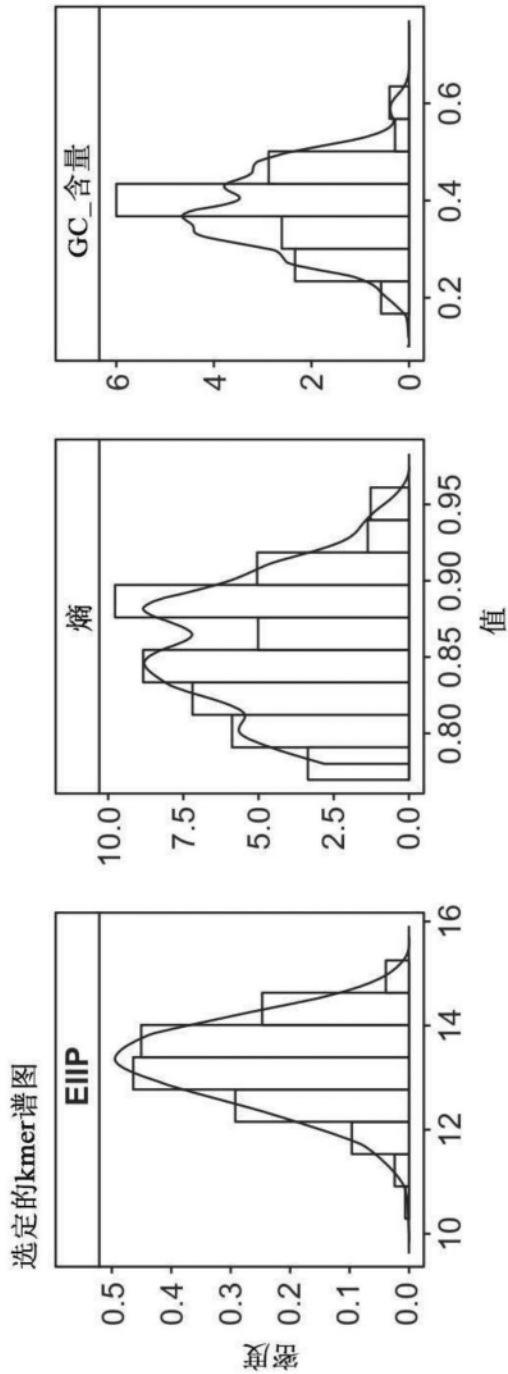


图17B

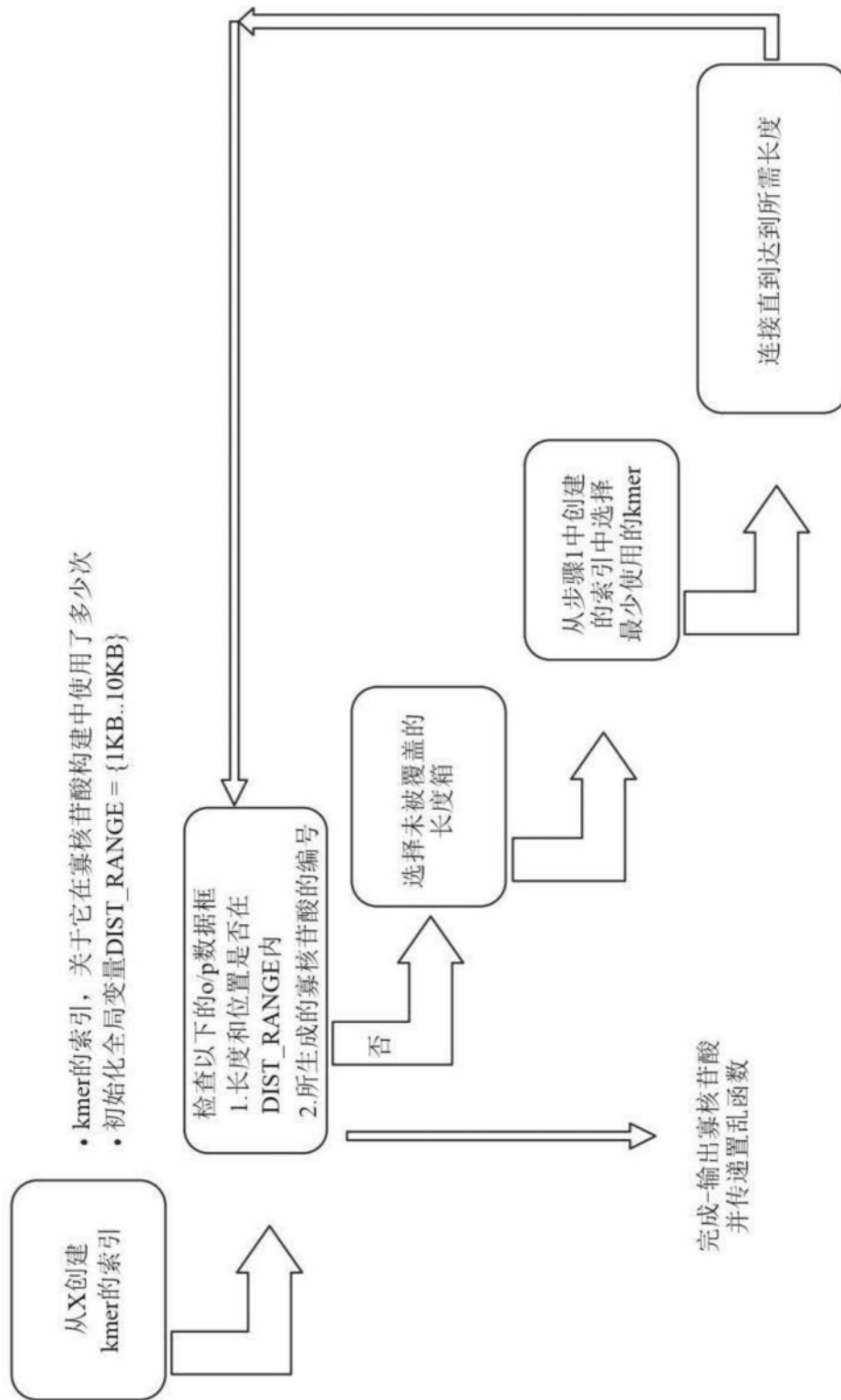


图18

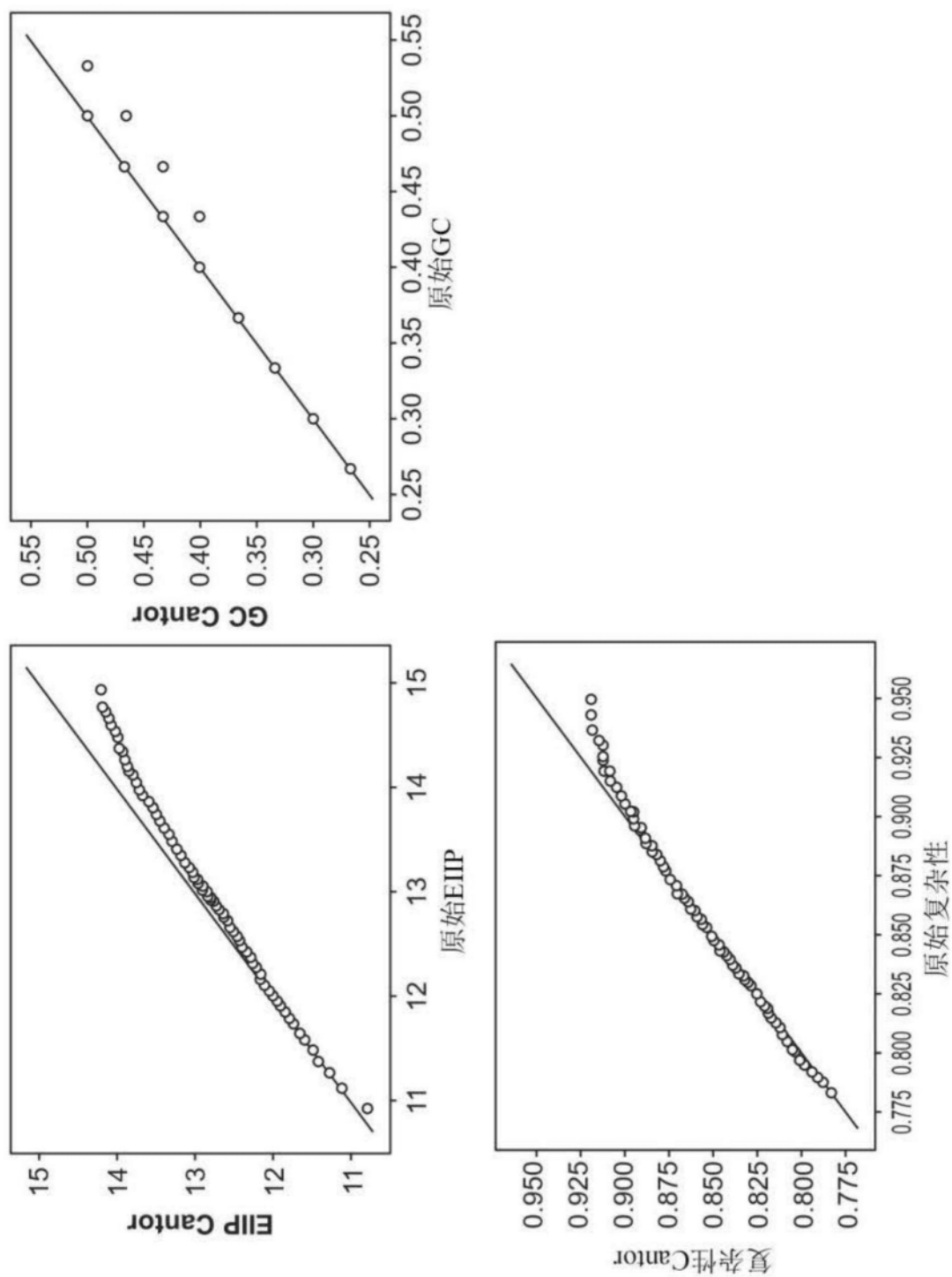


图19A

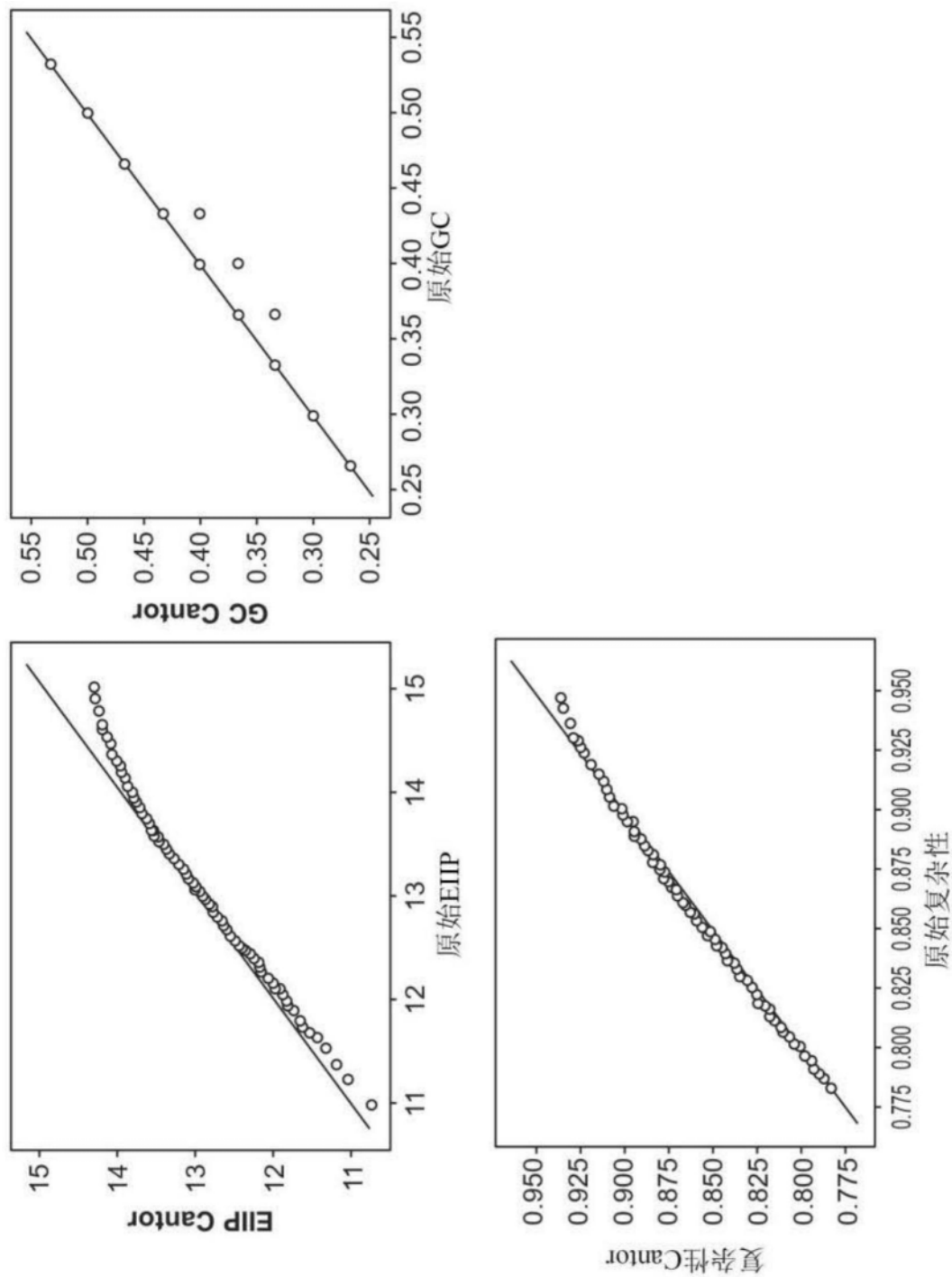


图19B

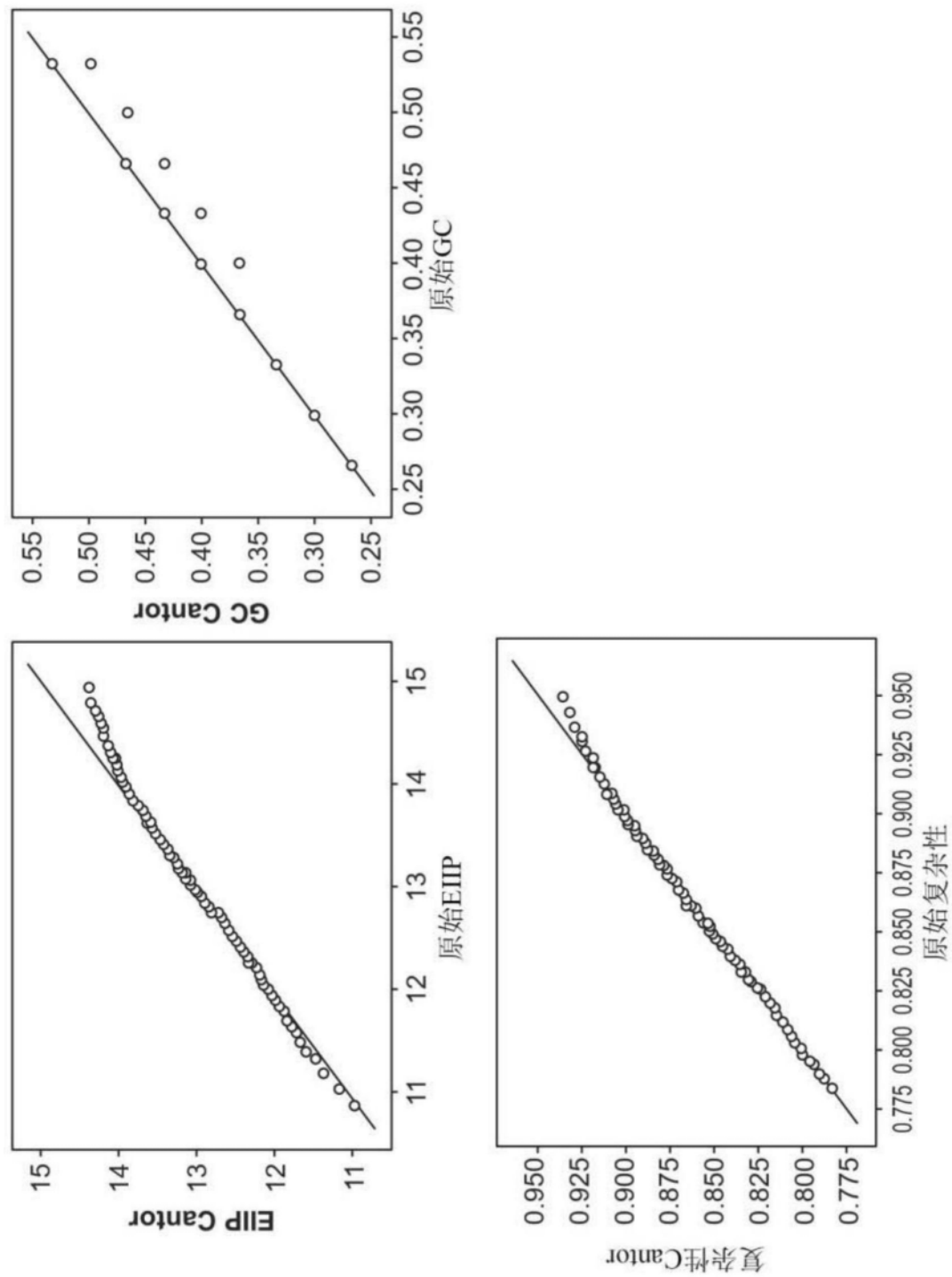


图19C

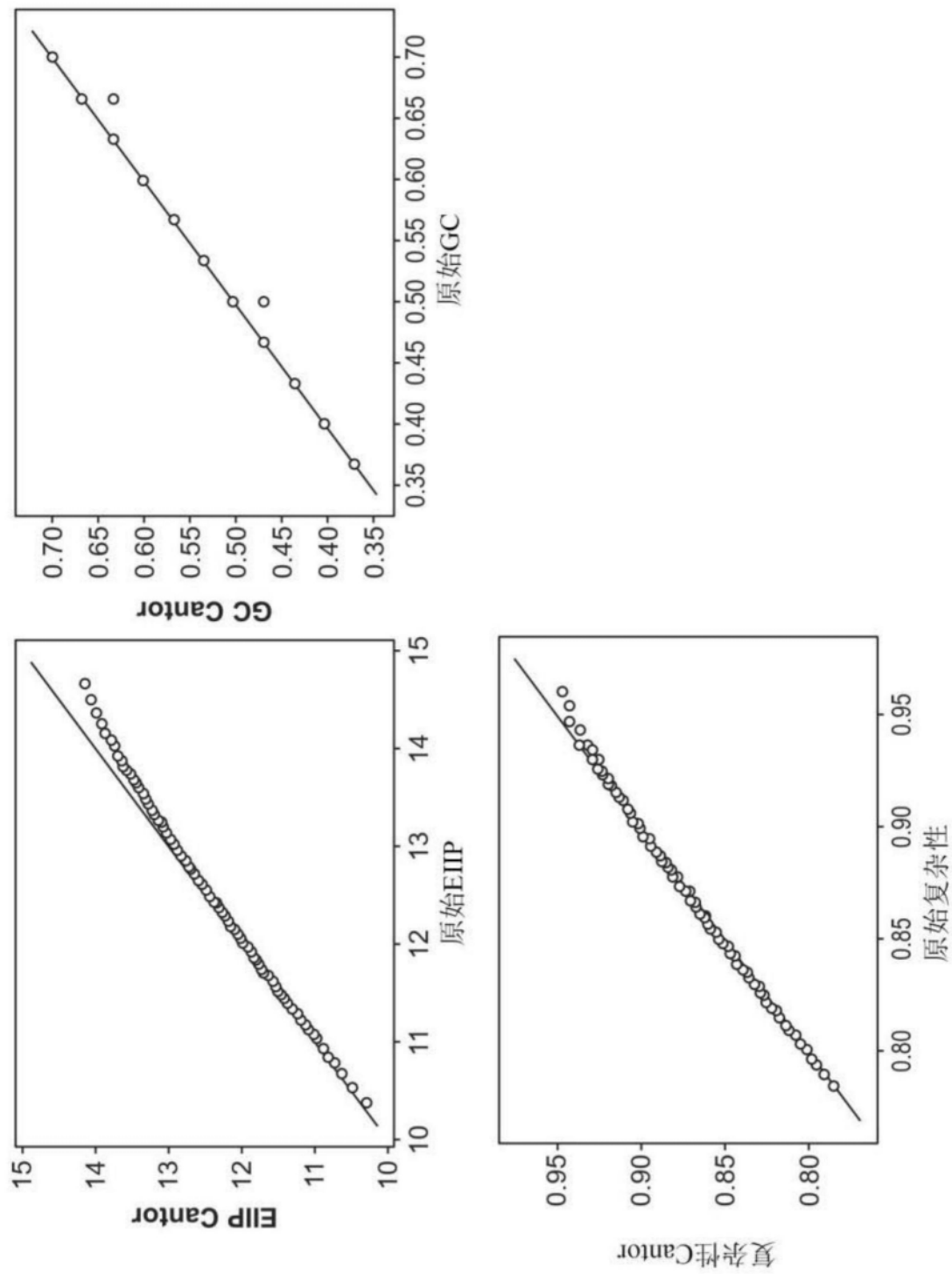


图19D

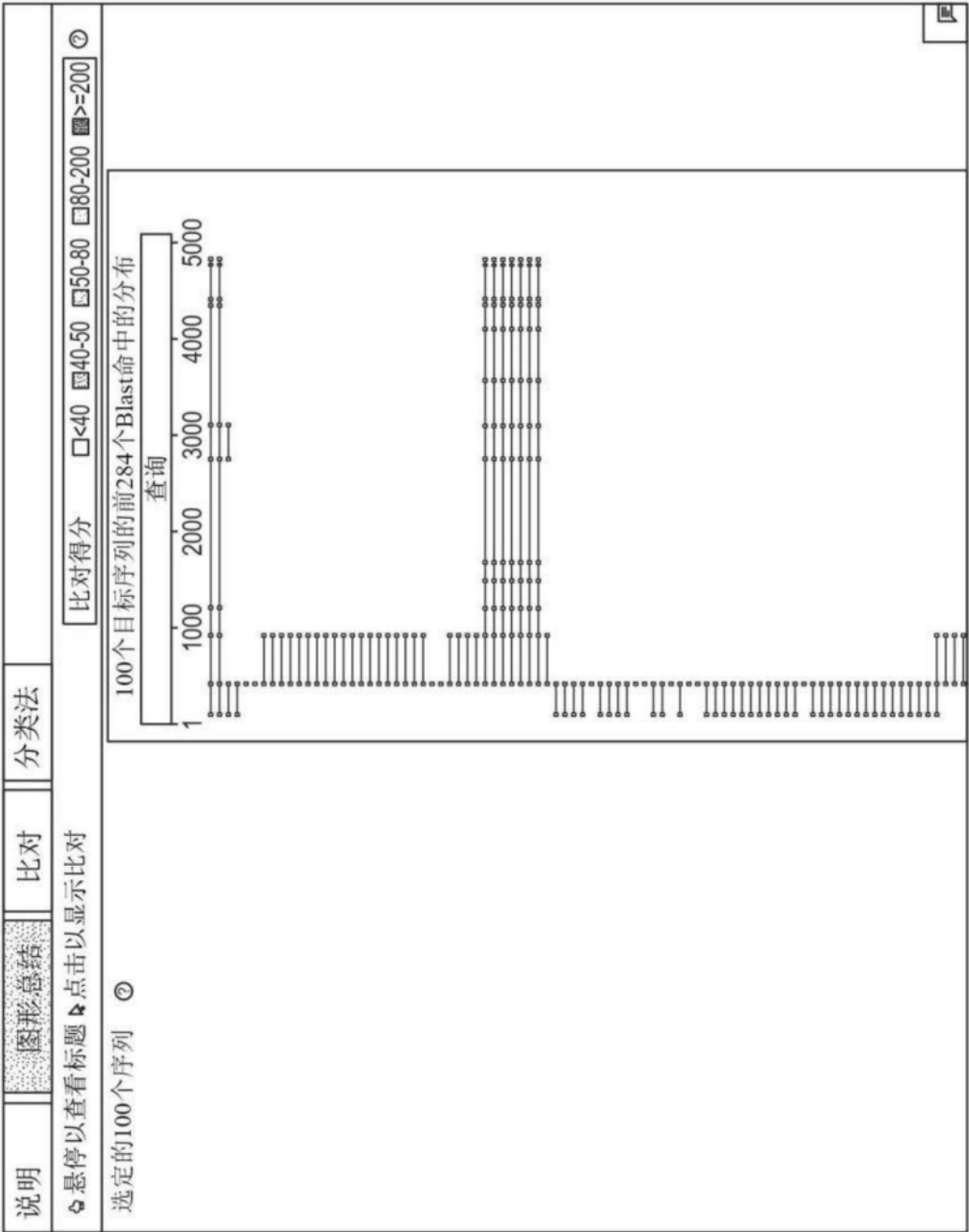


图20A

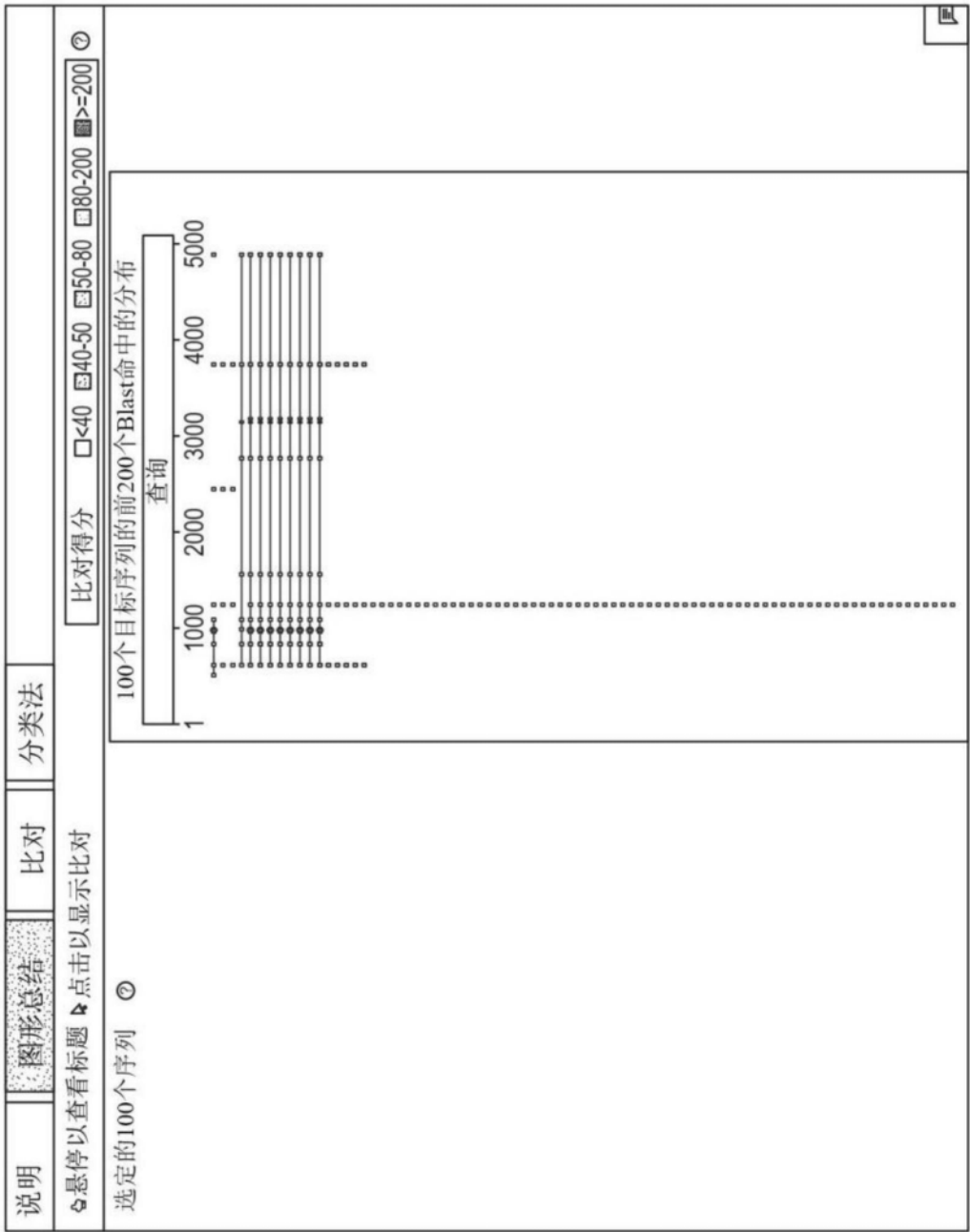


图20B

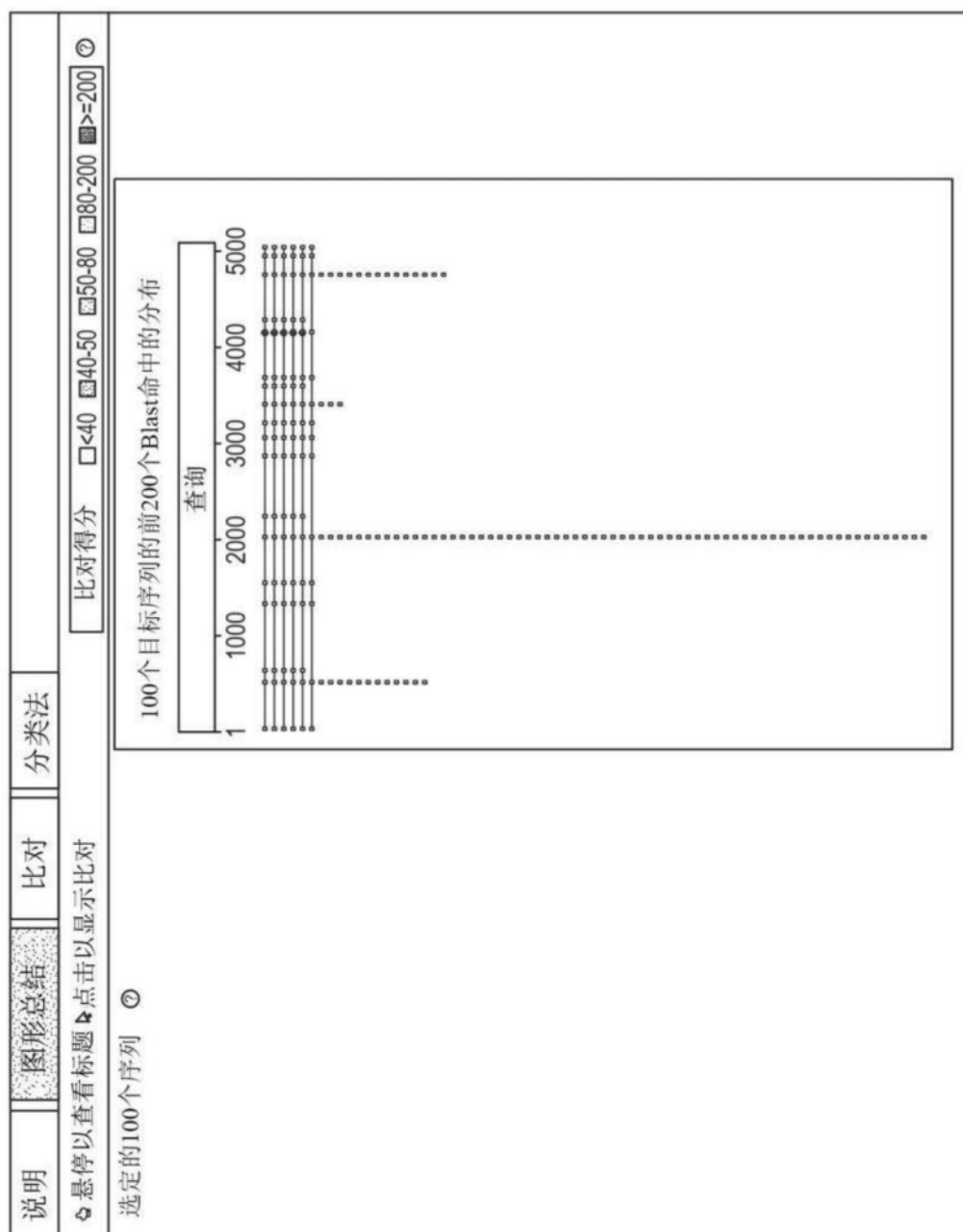


图20C

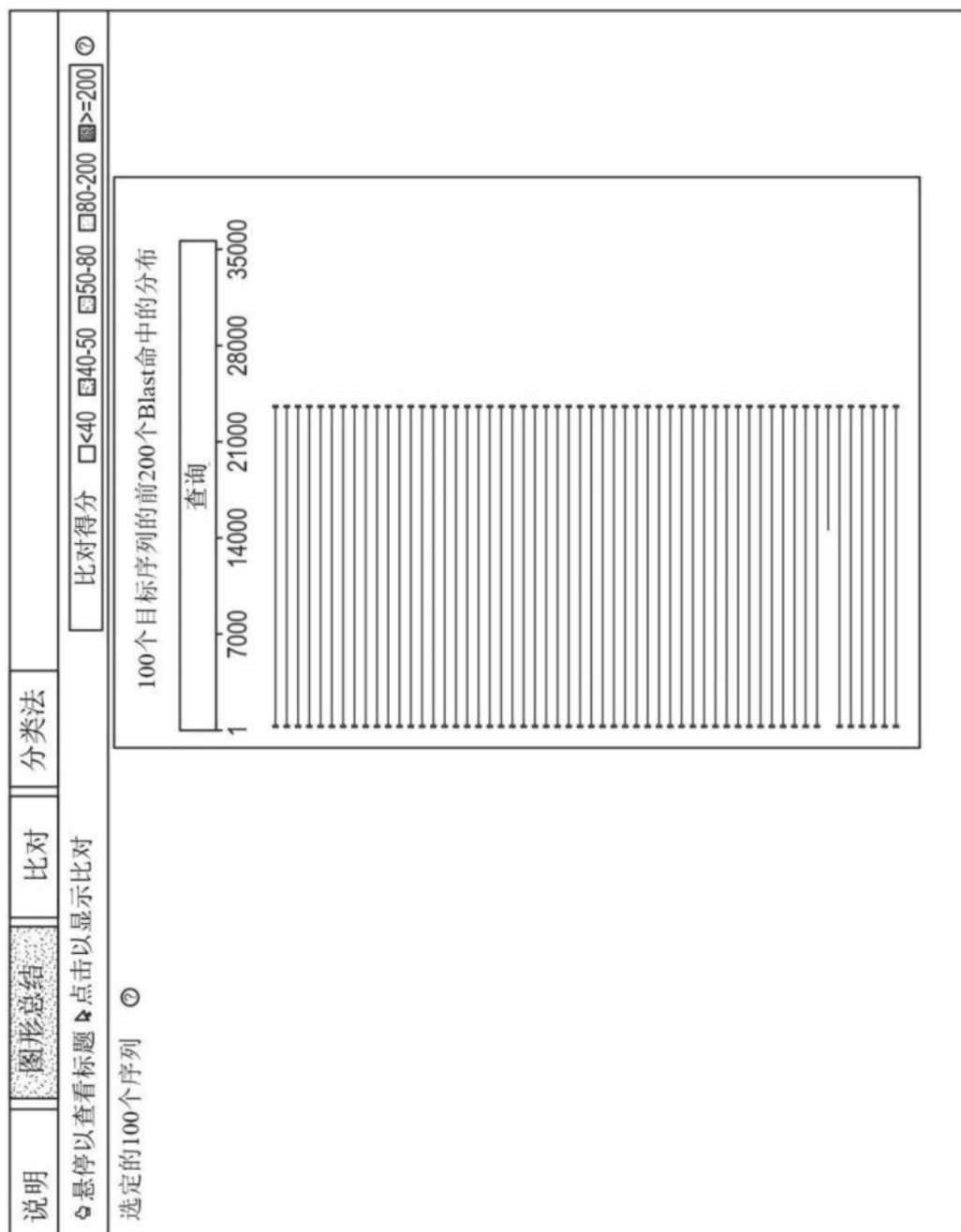


图20D

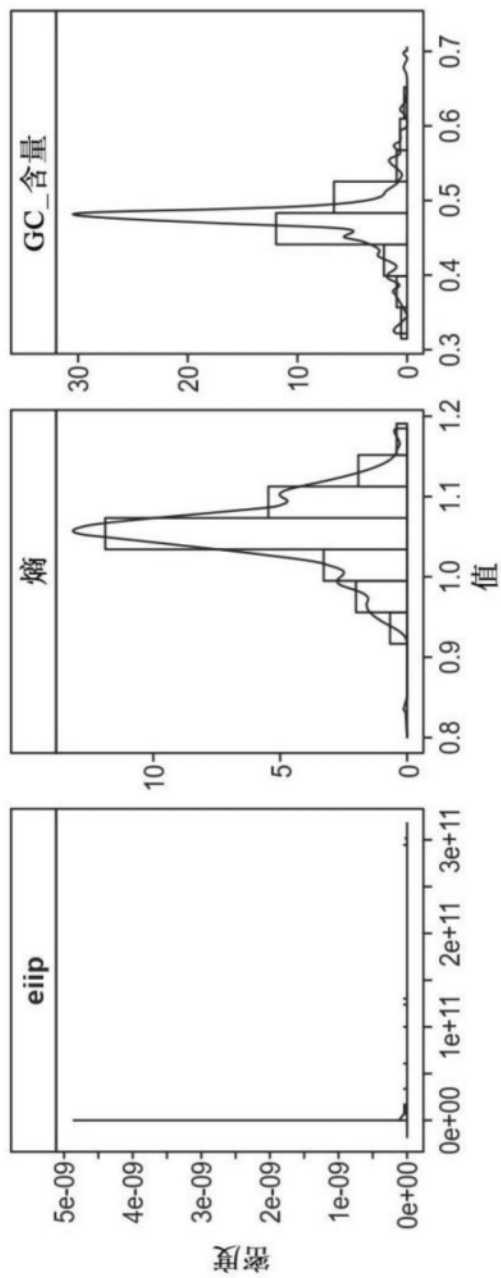


图21A

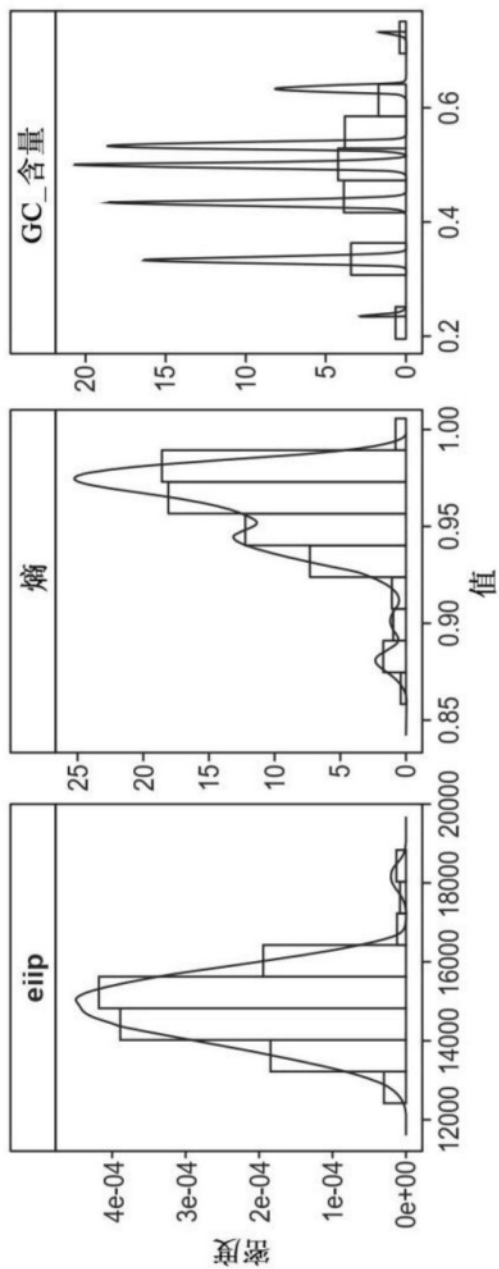


图21B

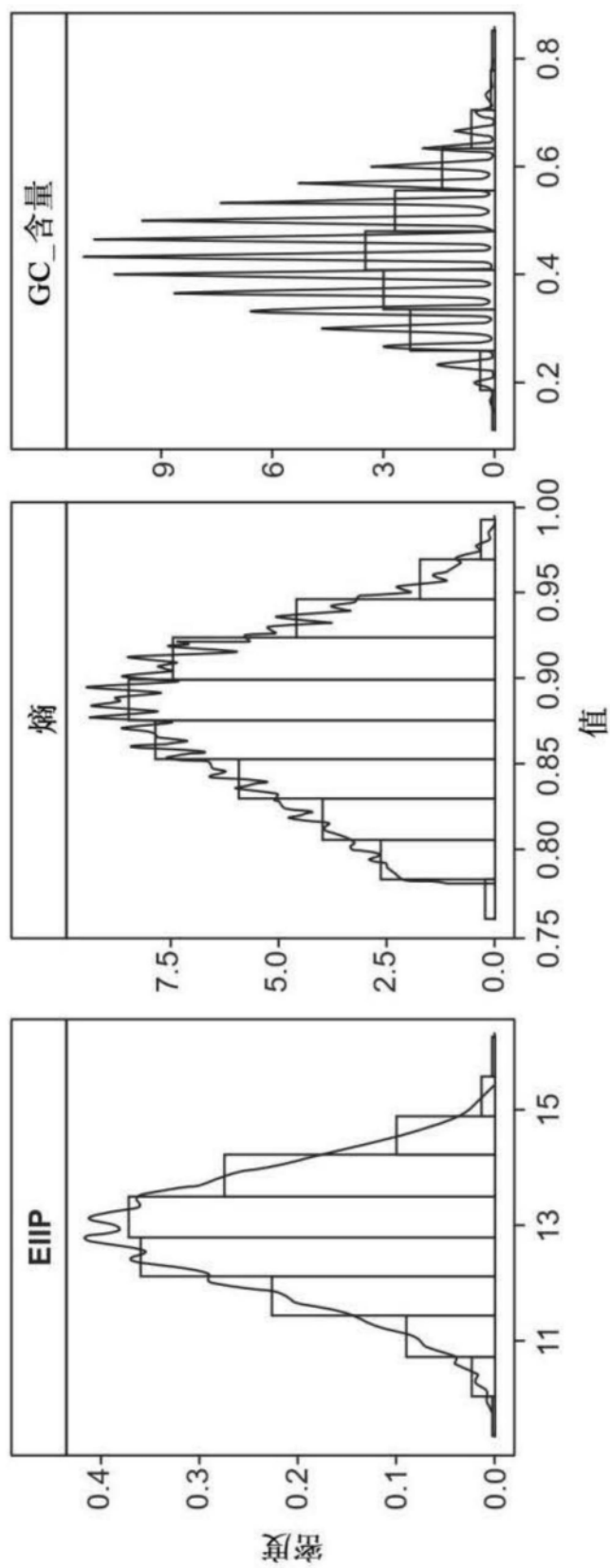


图21C

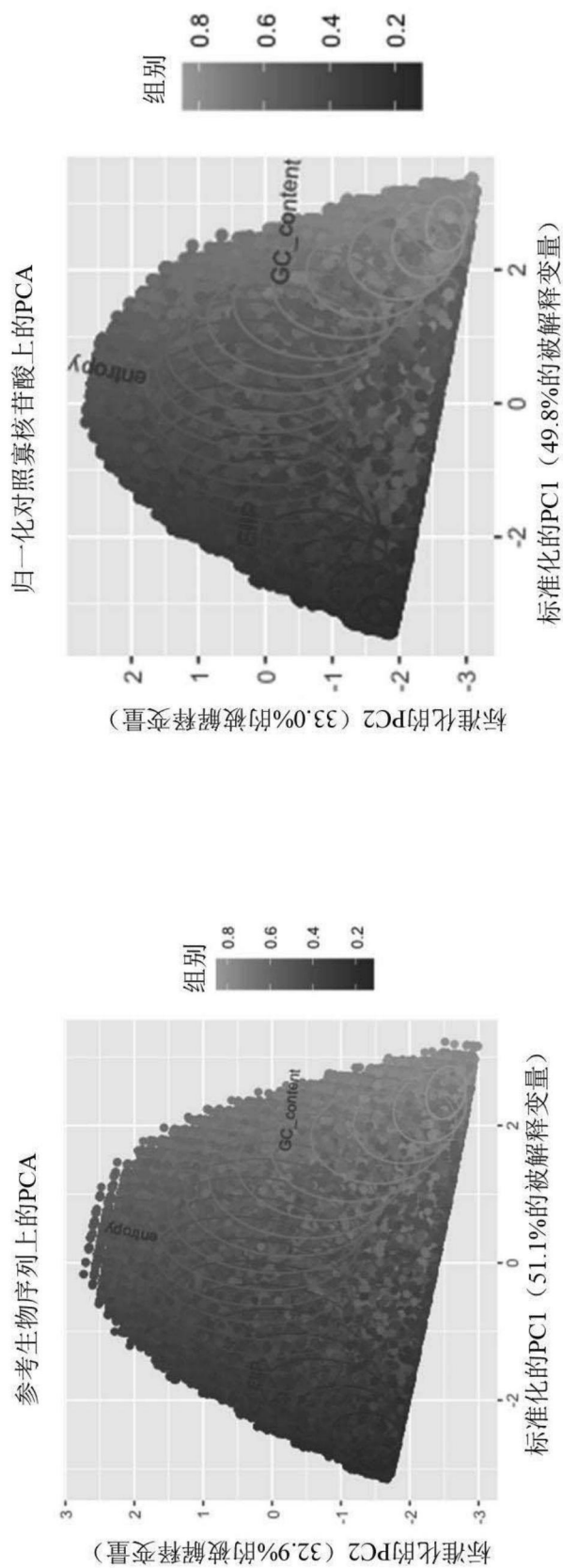


图22

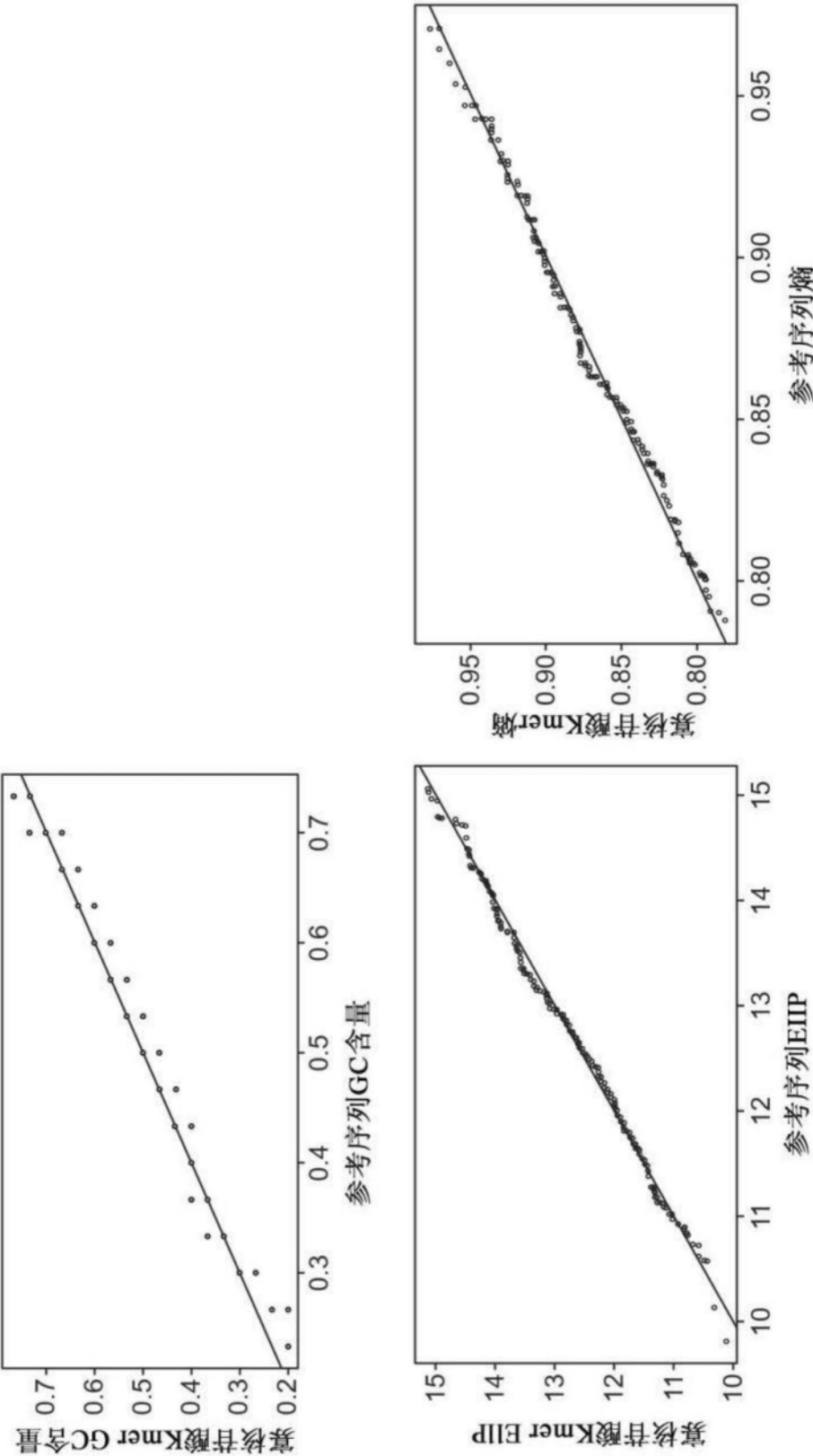


图23

K-S测试: 样本1/样本2

— fasra_熵 — 寡核苷酸_熵

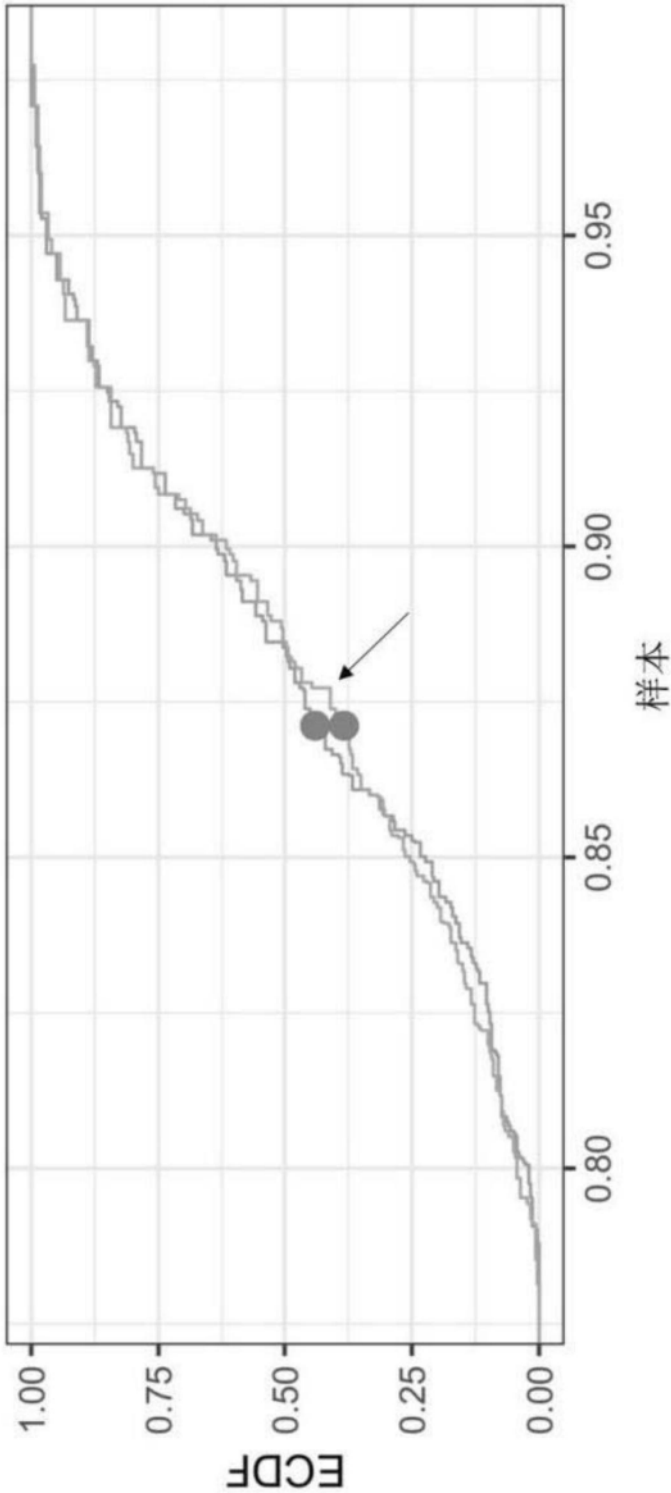


图24

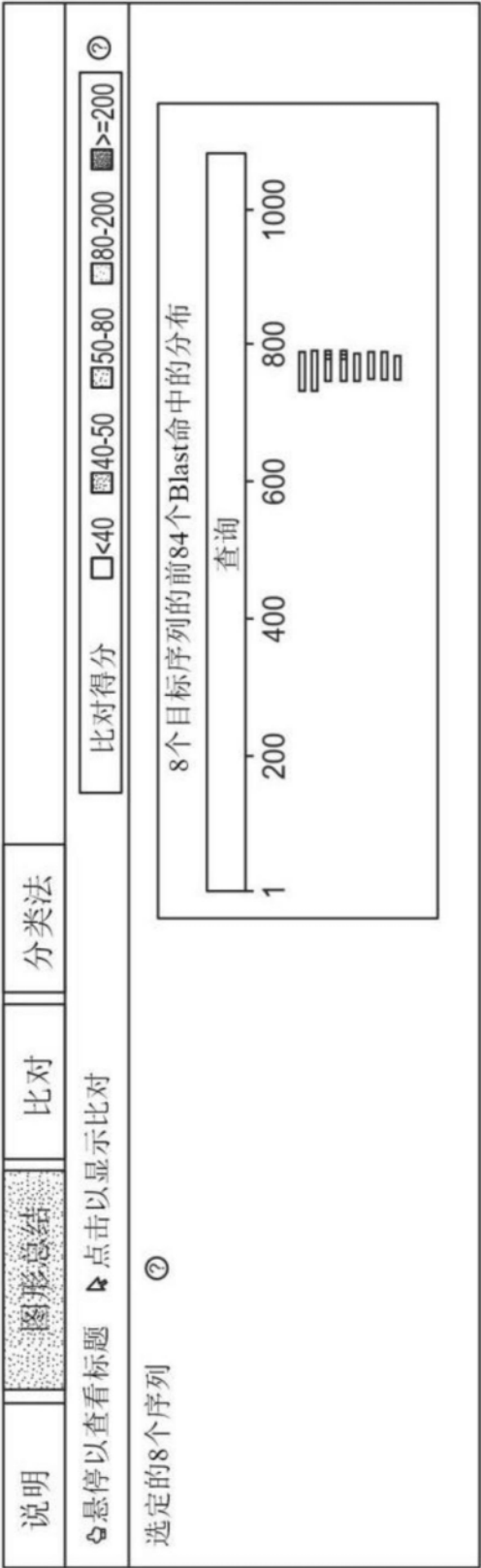


图25A

工作标题

核苷酸序列

RID

MHK0H18B014

搜索于08-07 08:38上午到期

下载所有

程序

② 引用

数据库

nr 查看细节

查询ID

lcl|查询_226969

说明

无

分子类型

dna

查询长度

1080

其他报告

②

过滤器结果

生物体将仅显示前20名

键入俗名、双名法名称、taxid名称或群名称

+添加生物体

同一性百分比

至

至

E值

至

过滤器

重置

排除

没有找到发现明显相似之处。要想知道原因是什么，请点击这里

图25B

Cantor置乱法与穷举Kmer法的比较

参数	Cantor置乱	穷举搜索
p值和D统计 (ks检验-熵)	1.0, 0.06 - 0.08	0.7212,0.0566
p值和D统计 (ks检验-eiip)	1.0,0.01 - 0.044	0.58, 0.0633
p值 (曼-惠特尼检验-GC含量)	0.0039 - 0.49	0.2672
对所有参考基因组都完成了?	否	是
建造所耗费的时间		3-4天
需要任何进一步置乱?	是-至遮罩区域	否
对所有参考基因组计算的特征?	否	是

图26

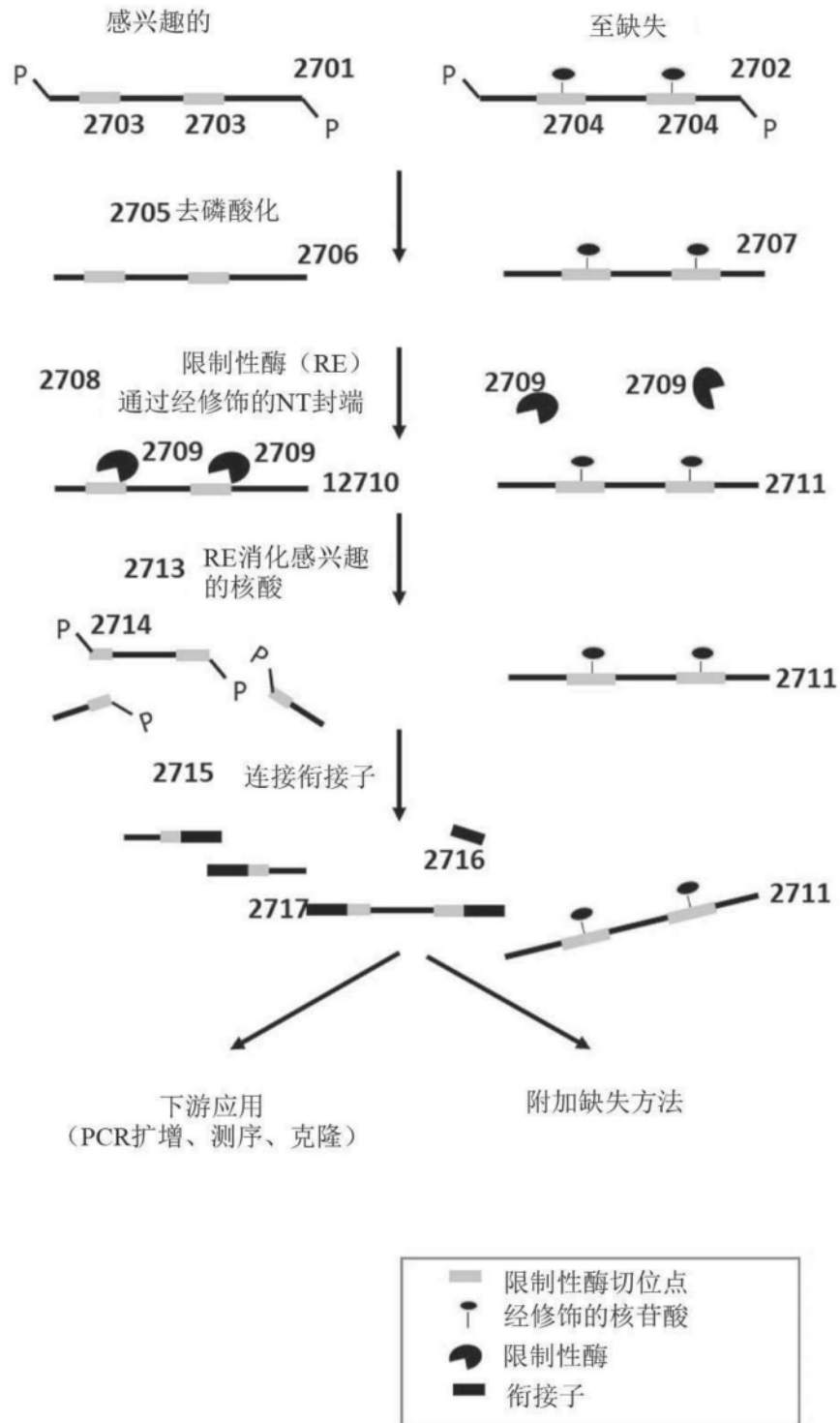


图27

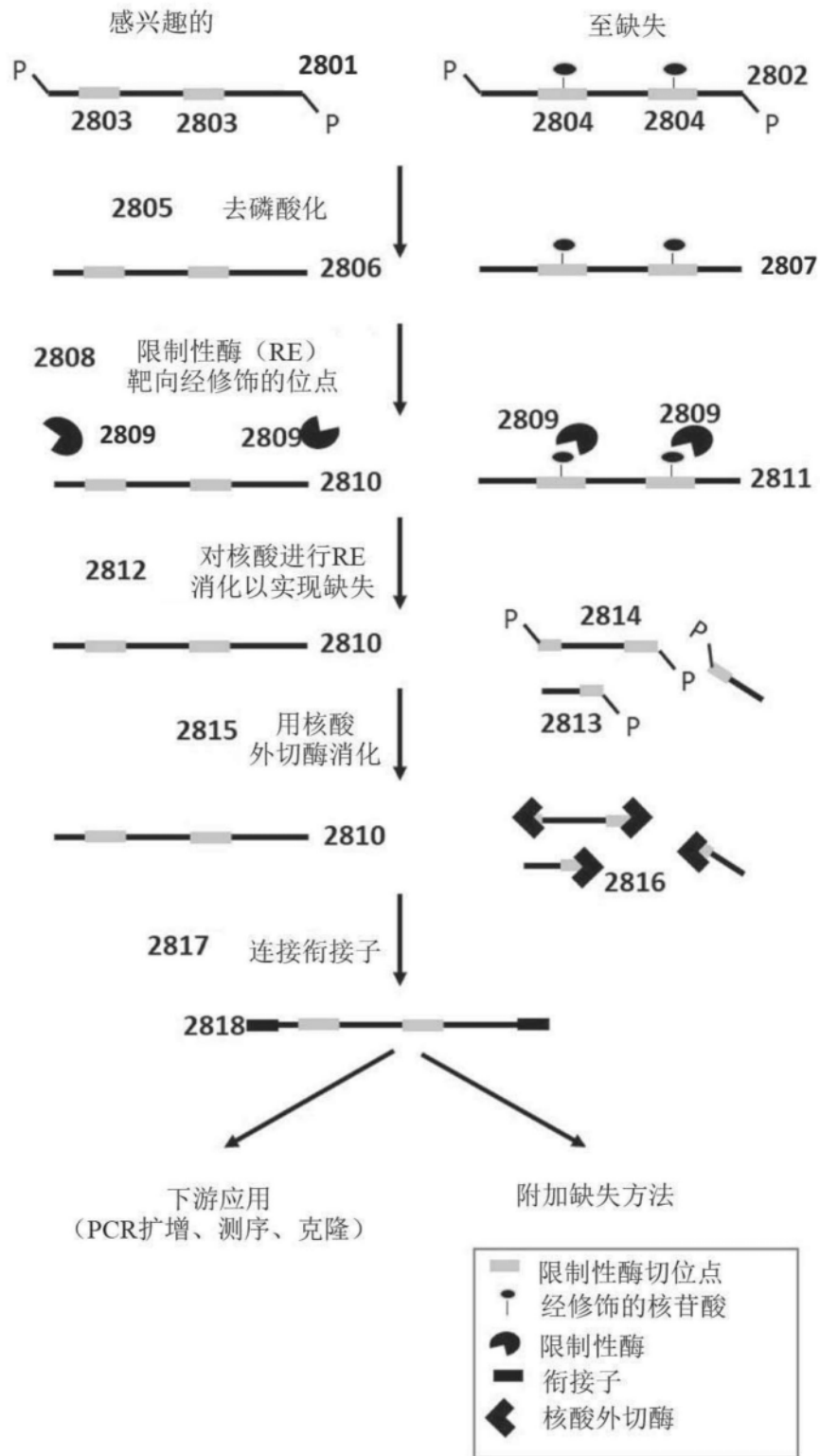


图28

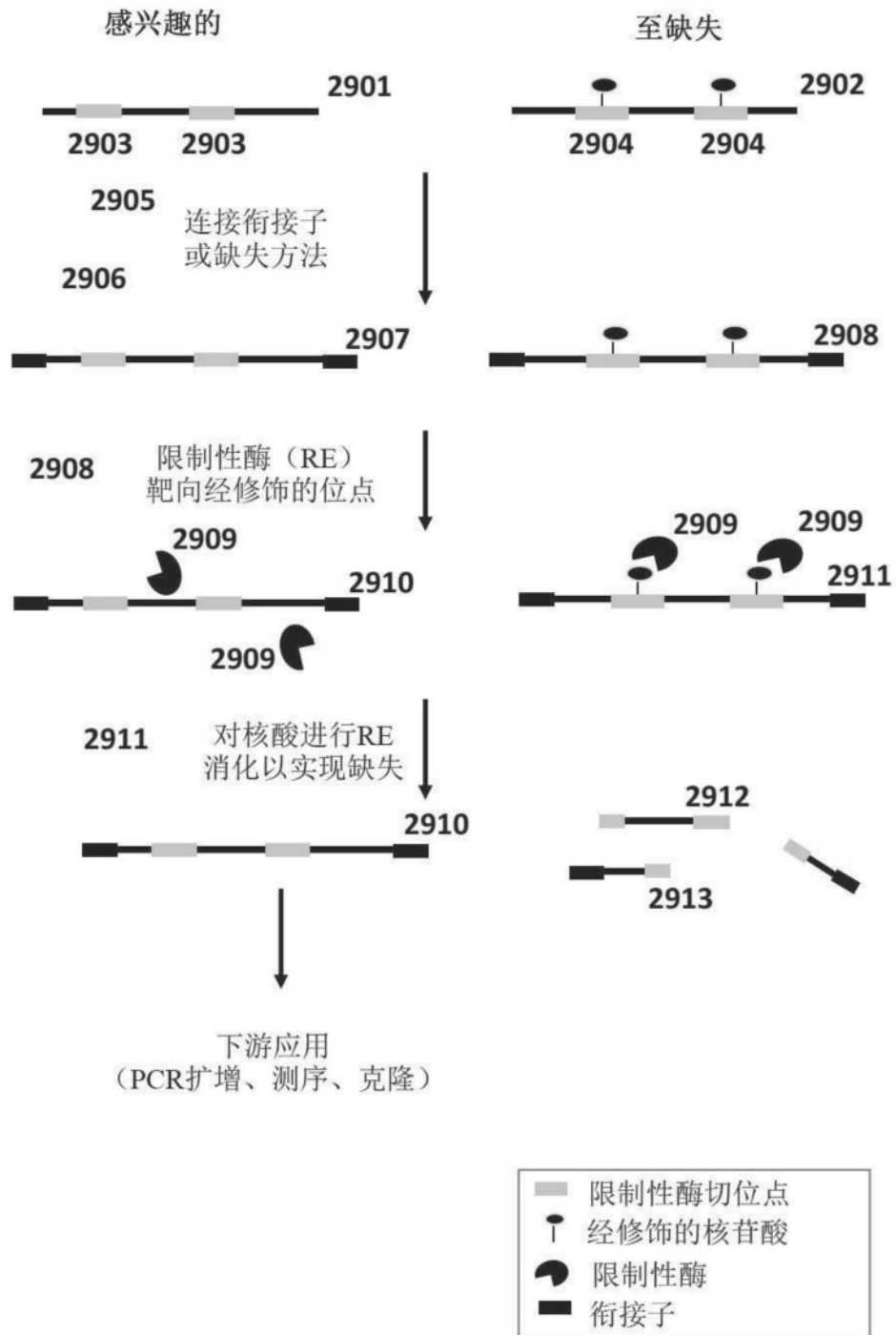


图29

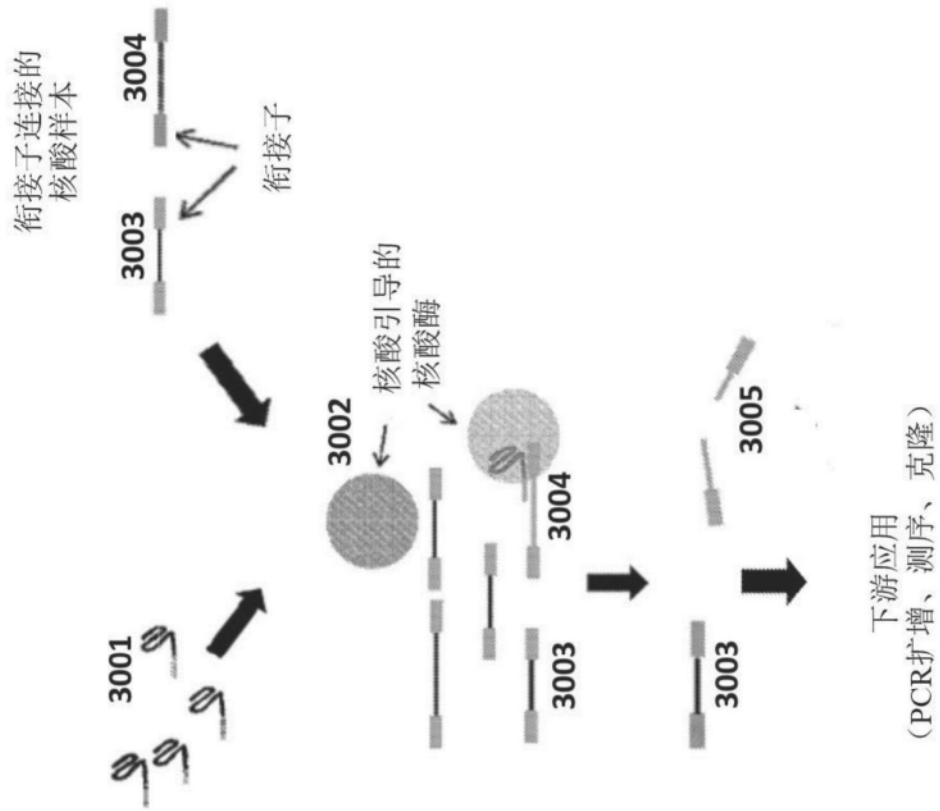


图30