[19]中华人民共和国国家知识产权局

[51]Int. Cl[7]
C12Q 1/68
C12Q 1/70 C07H 21/04
C12P 19/34

[12]发明专利申请公开说明书

[21]申请号 00811616.4

[43]公开日 2002年9月18日

[11]公开号 CN 1370242A

[22]申请日 2000.6.15 [21]申请号 00811616.4
[30]优先权
[32]1999.6.15 [33]US [31]09/333,110
[86]国际申请 PCT/US00/16465 2000.6.15
[87]国际公布 WO00/77260 英 2000.12.21
[85]进入国家阶段日期 2002.2.9
[71]申请人 基因描绘系统有限公司
地址 美国麻萨诸塞州
[72]发明人 D·斯特劳斯
[74]专利代理机构 中国专利代理(香港)有限公司
代理人 姜建成

权利要求书7页 说明书85页 附图页数11页

[54]发明名称 基因组分布分析:一种检测复杂生物样品中多种类型生物的存在的快速方法

[57]摘要

本发明提供了称为基因组分布分析的一种方法,该方法同时扫描 复杂的生物样品中多种不同类型生物特征性的核酸序列(包括基因组 差异序列、类群特异性序列和DNA多态性)的存在。本发明还包括用 于本发明的方法中的探针、检测集合和相关分子。

ISSN1008-4274

知识产权出版社出版

权 利 要 求 书

1. 一种从可能含有靶核酸分子的生物样品获取遗传信息的方法，所述方法包括下列步骤:

a) 提供下列核酸分子: (i) 在所述样品中的靶核酸分子，或(ii) 与所述样品中的靶核酸分子杂交的探针，或(iii) (i)或(ii)的扩增产物，或(iv) (i)的基因组代表；然后

b) 通过将(a)的核酸分子与最小基因组起源大于 5 的检测集合相接触或比较，检测靶核酸分子，其中所述检测集合包括能够检测靶核酸分子的检测序列。

2. 权利要求 1 的方法，所述方法还包括步骤(c)鉴定在步骤(b)检测到的核酸分子。

3. 权利要求 1 的方法，其中所述检测集合的最小基因组起源大于 11。

4. 权利要求 1 的方法，其中步骤(a)的核酸分子并不作为根据大小分级分离的片段固定化到基质或固相支持体上。

5. 权利要求 1 的方法，所述方法还包括下列步骤：假如在所述样品中存在靶核酸分子，就使用少于四对的扩增序列来产生扩增产物。

6. 权利要求 5 的方法，其中使用一对扩增序列进行扩增。

7. 权利要求 1 的方法，其中使用所述方法通过原位杂交来定量所述生物样品中的靶生物。

8. 权利要求 1 的方法，其中在步骤(a)之前，使所述样品的核酸分子与一个 ID 探针集合同时杂交以产生步骤(a)(ii)的探针。

9. 权利要求 1 的方法，其中步骤(a)(ii)的探针包括(i)能够与靶核酸分子杂交的第一个区，和(ii)扩增序列。

10. 权利要求 1 的方法，其中将所述样品的所述核酸分子固定到固相支持体上。

11. 权利要求 1 的方法，其中步骤(a)的所述核酸分子处于液相中。

12. 权利要求 1 的方法，其中步骤(a)的至少一些核酸分子包含一种或多种寡核苷酸标记。

13. 权利要求 1 的方法，其中步骤(a)(ii)的至少一些探针包含：(i)当与靶核酸分子杂交时可以相互连接的两种或更多种寡核苷酸，和(ii)扩增序列。

14. 权利要求 1 的方法，其中所述检测集合的所述检测序列在固相支持体上作为以两维的点排列或作为平行带排列。

15. 权利要求 8 的方法，其中所述 ID 探针集合包括探针，所述探针与来自至少十种不同病毒的每一种的至少两种不同核酸分子杂交，其中所述病毒的每一种都属于不同的属。

16. 权利要求 1 的方法，其中所述生物样品是胃肠道样品，并且所述遗传信息是对所述样品中来自 6 种或更多种以下生物的核酸分子的鉴定：大肠埃希氏菌(*Escherichia coli*)、沙门氏菌属(Salmonella)、志贺氏菌属(Shigella)、小肠结肠炎耶尔森氏菌(*Yersinia enterocolitica*)、霍乱弧菌(*Vibrio cholera*)、粪弯曲杆菌(*Campylobacter fecalis*)、艰难梭菌(*Clostridium difficile*)、轮状病毒属(Rotavirus)、诺沃克病毒(Norwalk virus)、星状病毒属(Astrovirus)、腺病毒属(Adenovirus)、冠状病毒属(Coronavirus)、兰氏贾第鞭毛虫(*Giardia lamblia*)、溶组织内阿米巴(*Entamoeba histolytica*)、人酵母菌(*Blastocystis hominis*)、隐孢子虫属(Cryptosporidium)、Microsporidium、美洲板口线虫(*Necator americanus*)、人蛔虫(*Ascaris lumbricoides*)、毛首鞭虫(*Trichuris trichiura*)、蛲虫(*Enterobius vermicularis*)、粪类圆线虫(*Strongyloides stercoralis*)、麝后睾吸虫(*Opsthorchis viverrini*)、华支睾吸虫(*Clonorchis sinensis*)和短膜壳绦虫(*Hymenoplepis nana*)。

17. 权利要求 1 的方法，其中所述生物样品是呼吸道样品，并且所述遗传信息是对所述样品中来自 6 种或更多种以下生物的核酸

分子的鉴定：白喉棒杆菌(*Cornybacterium diphtheriae*)、结核分枝杆菌(*Mycobacterium tuberculosis*)、肺炎支原体(*Mycoplasma pneumoniae*)、沙眼衣原体(*Chlamydia trachomatis*)、肺炎衣原体(*Chlamydia pneumoniae*)、百日咳博德特氏菌(*Bordetella pertussis*)、军团菌(Legionella spp.)、诺卡氏菌(Nocardia spp.)、肺炎链球菌(*Streptococcus pneumoniae*)、流感嗜血菌(*Haemophilus influenzae*)、鹦鹉热衣原体(*Chlamydia psittaci*)、铜绿假单胞菌(*Pseudomonas aeruginosa*)、金黄色葡萄球菌(*Staphylococcus aureus*)、荚膜组织胞浆菌(*Histoplasma capsulatum*)、*Coccidoides immitis*、新型隐球酵母(*Cryptococcus neoformans*)、皮炎芽生菌(*Blastomyces dermatitidis*)、卡氏肺囊虫(*Pneumocystis carinii*)、呼吸道合胞病毒、腺病毒属、单纯疱疹病毒、流感病毒、副流感病毒和鼻病毒属(Rhinovirus)。

18. 权利要求 1 的方法，其中所述生物样品是血液样品，并且所述遗传信息是对所述样品中来自 6 种或更多种以下生物的核酸分子的鉴定：凝固酶阴性葡萄球菌、金黄色葡萄球菌、*Viridans streptococci*、肠球菌(Enterococcus spp.)、β溶血性链球菌、肺炎链球菌、埃希氏菌(Escherichia spp.)、克雷伯氏菌(Klebsiella spp.)、假单胞菌(Pseudomonas spp.)、肠杆菌(Enterbater spp.)、变形虫(Proteus spp.)、拟杆菌(Bacteroides spp.)、梭菌(Clostridium spp.)、铜绿假单胞菌、棒杆菌(Cornybacterium spp.)、疟原虫(Plasmodium spp.)、杜氏利什曼原虫(*Leishmania donovani*)、弓形虫(Toxoplasma spp.)、微丝蚴(Microfilariae)、真菌、荚膜组织胞浆菌、*Coccidoides immitis*、新型隐球酵母、假丝酵母(Candida spp.)、HIV、单纯疱疹病毒、丙型肝炎病毒、乙型肝炎病毒、巨细胞病毒属(Cytomegalovirus)和 EB 病毒。

19. 权利要求 1 的方法，其中所述遗传信息是对所述样品中来自 6 种或更多种以下生物的核酸分子的鉴定：柯萨奇病毒 A、单纯疱疹病毒、圣•路易脑炎病毒、EB 病毒、粘液病毒、JC 病毒、柯萨奇病毒 B、披膜病毒、麻疹病毒、肝炎病毒、副粘病毒、艾可病毒、

布尼亚病毒、巨细胞病毒、水痘-带状疱疹病毒、HIV、腮腺炎病毒、马脑炎病毒、淋巴细胞性脉络丛脑膜炎病毒、狂犬病病毒和 BK 病毒。

20. 权利要求 8 的方法，其中至少 50%的组成所述核酸探针集合的探针能够与可能存在于所述样品中或存在于所述样品的基因组代表中的预先确定的基因组差异序列杂交。

21. 一种用于从生物样品获取遗传信息的试剂盒，所述试剂盒包含:

a) 多种 ID 探针和/或 SNP 探针；和

b) 一个包含与(a)的探针相应的检测序列的检测集合，其中所述检测集合的最小基因组起源大于五。

22. 权利要求 21 的试剂盒，其中(a)包含多于十种的不同的可扩增探针。

23. 权利要求 22 的试剂盒，其中(a)包含多于五十种的不同的可扩增探针。

24. 权利要求 23 的试剂盒，其中(a)包含多于二百五十种的不同的可扩增探针。

25. 权利要求 21 的试剂盒，其中所述检测集合的最小基因组起源大于 11。

26. 权利要求 21 的试剂盒，其中(a)包含多于五个的可扩增探针家族。

27. 权利要求 21 的试剂盒，其中(a)的探针对于至少两个不同的分类单位具有特异性。

28. 权利要求 27 的试剂盒，其中(a)的探针对于至少两个不同的物种具有特异性。

29. 权利要求 27 的试剂盒，其中(a)的探针对于至少两个不同的属具有特异性。

30. 权利要求 27 的试剂盒，其中(a)的探针对于至少两个不同的界具有特异性。

31. 权利要求 21 的试剂盒，其中(a)的探针包括包含以下的探针：(i)当与靶核酸分子的 ID 序列杂交时可以相互连接的两种或更多种寡核苷酸，和(ii)扩增序列。

32. 权利要求 21 的试剂盒，其中(a)的探针和/或(b)的检测序列物理结合于固相支持体上的不同位置。

33. 权利要求 21 的试剂盒，其中至少 50%的(a)的探针包含来自至少三个不同物种的基因组差异序列。

34. 权利要求 32 的试剂盒，其中检测(i)一个分类群的成员和(ii)密切相关的分类群的所述检测集合所包含的检测序列在所述支持物上相互邻近定位。

35. 一个 ID 探针集合，所述 ID 探针集合可以使用少于四对的扩增序列扩增，并且包含多于三个的 ID 探针家族和多于十种的不同的 ID 探针。

36. 权利要求 35 的集合，所述集合包含多于五十种的不同的可扩增 ID 探针。

37. 权利要求 36 的集合，所述集合包含多于两百五十种的不同的可扩增 ID 探针。

38. 权利要求 35 的集合，所述集合包含多于十个的可扩增 ID 探针家族。

39. 权利要求 35 的集合，所述集合包含多于二十五个的可扩增 ID 探针家族。

40. 权利要求 35 的集合，其中所述可扩增探针家族中多于两个的家族对于非重叠的分类单位具有特异性。

41. 权利要求 35 的集合，其中所述可扩增探针家族中多于两个的家族对于不同物种具有特异性。

42. 权利要求 35 的集合，其中所述可扩增探针家族中多于两个的家族对于不同属具有特异性。

43. 权利要求 35 的集合，其中所述可扩增探针家族中多于两个

的家族对于不同界具有特异性。

44. 权利要求 35 的集合，其中(a)的探针包括包含以下的探针：(i)当与靶核酸分子的 ID 序列杂交时可以相互连接的两种或更多种寡核苷酸，和(ii)扩增序列。

45. 权利要求 35 的集合，其中至少 50%的所述探针包含来自三个不同物种的基因组差异序列。

46. 权利要求 35 的试剂盒，其中检测(i)一个分类群的成员和(ii)密切相关的分类群的所述检测集合所包含的检测序列在支持物上相互邻近定位。

47. 一种从可能含有靶核酸分子的生物样品获得遗传信息的方法，所述方法包括下列步骤：

a) 提供最小基因组起源大于五的核酸探针集合；

b) 使所述探针集合同时与所述样品的核酸分子接触；

c) 检测在所述探针和所述样品的任何靶核酸分子之间的杂交；以及

d) 鉴定在步骤(c)中检测到的核酸分子。

48. 权利要求 13 的方法，其中所述可以连接的寡核苷酸是 SNP 探针。

49. 权利要求 48 的方法，其中至少某些所述 SNP 探针包含可以与检测集合中标记序列杂交的标记序列，其中所述检测集合包含与所述 SNP 探针相应的标记序列集合。

50. 权利要求 48 的方法，其中所述检测集合的最小基因组起源大于 20。

51. 权利要求 50 的方法，其中所述检测集合的最小基因组变异大于 50。

52. 权利要求 1 的方法，其中通过使用不多于四对的扩增序列来扩增步骤(a)(i)的靶核酸分子，产生步骤(a)(iv)的扩增产物。

53. 权利要求 52 的方法，其中所述扩增序列指导使用 Alu 特异

性引物的对位于 Alu 重复序列之间的序列的扩增。

54. 权利要求 52 的方法，其中(b)的检测集合包含与可能在步骤(a)(iv)中扩增的 ID 探针相应的 ID 位点。

55. 一种用于从生物样品中获取遗传信息的试剂盒，所述试剂盒包括:

a) 多种核酸引物，所述核酸引物能够在生物样品中引发靶基因组 DNA 中与重复序列邻接的 DNA 序列的扩增，产生 ID 探针；以及

b) 一个包含检测序列的检测集合，所述检测与使用(a)的引物可能扩增出的 ID 探针相应，其中所述检测集合的最小基因组起源大于五。

56. 权利要求 55 的试剂盒，其中所述检测集合的最小基因组起源大于 20。

57. 权利要求 55 的试剂盒，其中所述重复序列是人 Alu 重复序列，并且所述引物是 Alu 特异性引物。

说　明　书

基因组分布分析：一种检测复杂生物样品中多种类型生物的存在的快速方法

发明背景

本发明涉及从复杂生物样品如机体样品(如血液、尿、痰和粪便)中获取遗传信息。在医学上，鉴定所述样品中的传染性生物对于感染的最佳治疗和保持公共卫生是重要的。确定患者是否患有遗传性疾病和法医鉴定也极大依赖于对机体样品的遗传信息的分析。

虽然目前用于诊断传染因子的程序包括整套复杂的几百种测试，但很大一部分传染性生物常常没有被检测出来。例如，在鉴定肺炎患者体内的传染因子的尝试中，成功率仅约一半，而肺炎是美国由传染病引起的死亡的最常见的死因。

许多疾病如肺炎、脑膜炎和急性胃肠疾病的特征在于可以由多种传染因子引起的一组症状(“表现(presentation)”)。还不存在扫描所有通常引起这样的疾病的病原体的单一测试。(我称这样的测试为“表现特异性测试”。)目前的程序常常仅测试一种类型致病生物的存在。这中间存在问题，因为常常必须对一个样品进行许多不同的测试，这增加了费用、鉴定所需时间以及错误的可能性。

另外，许多程序对于日常应用来说太昂贵。例如，可能需要几百美元来对一种特定病毒进行测试。卫生保健提供者必须权衡这种费用，尤其是考虑到鉴定传染因子可能需要多项测试。

大多数目前的诊断测试需要培养传染因子以获得大量的生物。不幸的是，许多类型的生物无法在医院实验室内进行常规培养。大多数病毒和寄生虫以及许多细菌属于这种类型。对于可以培养的生物，培养可能需要的几天或甚至几周，这就浪费了宝贵的时间。因此，患有例如细菌性脑膜炎的患者的生命非常依赖于立即治疗，但最佳的治疗

可能需要由于培养而引起的耗时和威胁生命的延迟。其它传染因子，如导致肺结核的细菌，一般需要几周以在培养物中生长。在鉴定(以及最佳治疗)上的延迟可能导致患有肺结核的患者将这种高度传染性的疾病传染给许多其他人。

目前在医院中实行的诊断测试仅产生在样品中存在的生物种类的粗略鉴定。在许多情况下，很难将一种致病生物与一种密切相关的非病原体区别开来。

此外，为鉴定一种病原体，一个样品可能需要在几个不同的实验室由几组人员进行许多测试，而每一组人员都接受不同类型的专业训练。配备必要专业人员所需的费用是诊断学实验室的预算的一项主要支出。同时，在不同实验室之间分配样品引入了另一个误差源，另外，如果测试需要病原体存活，那么运输可能成为问题。

因此，需要一种新类型的测试，所述测试是表现特异性的(即全面的)、有效率地检测来自各种不同类群的大量生物的存在、能够在相当短的时间内进行(例如几个小时)、使用单一测试的形式、并导致高分辨率的病原体鉴定。

从生物样品中获取精确的遗传信息可以提供关于在所述样品中存在的生物的身份和医学上的相关属性的信息。这是因为由于进化趋异，每一种类型的生物都具有独特的基因组 DNA 序列。DNA 序列随时间的流逝发生变化的原因包括宇宙射线的冲击、化学诱变剂的修饰、正常 DNA 复制中的错误、遗传重组引起的重排、以及病毒、质粒和转座遗传因子的入侵。结果，单个碱基的改变积累，序列区段缺失，序列区段插入，并且染色体重排。因此，基因组是保守序列(即对于不同分类单位是共同的序列)以及作为上文枚举的改变类型的结果的趋异序列的嵌合体。因此，测试独特基因组笔迹(genomic signature)或基因组指纹的方法对于鉴定生物是有用的。

已经发展了多种方法以获得传染性生物的 DNA 指纹。这些方法包括限制性片段长度多态性(RFLP)分析、扩增片段长度多态性(AFLP)

分析、脉冲电场凝胶电泳、任意引物聚合酶链式反应(AR-PCR)、基于重复序列的 PCR、ribotyping 和比较性核酸测序。这些方法一般太慢、太昂贵、无可重复性、并且对技术过分要求，以致于不能在大多数诊断环境中使用。所有上面提到的方法一般要求使用麻烦的凝胶电泳步骤，需要在培养物中培养病原体，需要纯化病原体的基因组 DNA，以及要求所述样品不包含多于一种类型的生物(这排除了直接测试复杂医学样品的可能性)。最近发展的依赖样品与高密度微阵列(microarray)进行杂交的高分辨率株鉴定法(Salazar 等, Nucleic Acids Res. 24: 5056-5057, 1996; Troesch 等, J. Clin. Microbiol. 37: 49-55, 1999; Lashkari 等, Proc. Natl. Acad. Sci. U.S.A. 94: 13057-13062, 1997)也具有相同的限制(除了对凝胶电泳的需要外)。此外，这些新的杂交方法可能对技术过分要求，因为它们一般要求将与小寡核苷酸的杂交和不同程度的错配区分开来。基于更大 DNA 序列的存在与否的方法将提供更健全(robust)、并因此在临床上更有用的诊断测定。采用 DNA 指纹形式的精确的基于遗传学的鉴定对于追踪和控制在地区和在医院中的传染病爆发是至关重要的。在治疗上，指纹分析，尤其是当能够以快速、不依赖于培养的测试提供指纹分析时，就能够通过比目前的实践更快地确定给予何种抗生素而挽救生命。

也已经发展了一次测试样品中几种不同类型生物的存在的方法。注意到：目前这样的方法一般尚不适用于指纹分析，也就是说，不适用于在一个物种内的密切相关的生物之间进行区分。不需要培养而一次测试几种生物的存在的方法是多重 PCR。多重 PCR 和其它多重扩增方法的一个主要问题在于很难同时扩增许多序列(当包括更多引物序列时，扩增假象(amplification artifact)开始积累)。由于可以使用多重 PCR 测试的序列数目的限制，很难建立健全的多重测试，检测在多种不同类型生物中出现的多种不同序列。因此，应用多重 PCR 同时测试在系统发生上全异的生物的最好的例子之一仅检查九个序列，这远不足以提供表现特异性的测试(Grondahl 等, J. Clin. Microbiol. 37: 1-7,

1999)。此外，由于可以使用的诊断探针的数量的限制(仅测试每种类型生物的一种序列)，该测试缺乏冗余性(这对于可重复性是重要的)，仅提供传染因子的粗略鉴定。多重 PCR 还对于在大多数医学样品中存在的抑制剂敏感，需要对技术过分要求的样品处理以获得健全的结果。

在遗传上鉴定生物的一种方法涉及测试对于特定类型生物独特的序列(或序列组)的存在。这样的序列称为标识(ID)序列。例如，为检测人类免疫缺陷病毒的存在，人们测试在该病毒类群的成员中独特存在的 DNA 序列的存在。在另一个例子中，一个大肠埃希氏菌(*Escherichia coli*)菌株当存在于人类胃肠道中时可能是无害的，而另一个大肠埃希氏菌菌株的存在可能是威胁生命的。虽然这样的菌株非常密切相关，但可以通过检测在它们的 DNA 序列中的变异而将它们区分开来。

为将一种生物与其密切相关的亲缘生物区分开来，测试在来自每个类群的每个株中以独特组合出现的一组 DNA 序列的成员的存在是有用的。这样的序列称为基因组差异序列，已经在文献中描述，如在 Straus ("基因组扣除", 在 PCR Strategies, Innes 等编辑, 第 220-236 页 (Academic Press Inc., San Diego, 1995))，该文献特此通过引用结合到本文中。基因组差异序列是与一种生物的基因组杂交，但不与另一种不同但密切相关的生物的基因组杂交的 DNA 序列。如在 Straus (1995, 见上文)中所述，例如，可以通过用两种不同生物的基因组进行扣除杂交，制备基因组差异序列。得到的基因组差异序列组成一组核酸序列，该组序列在一种基因组扣除样品中存在，但在另一组基因组扣除样品中不存在。例如，在一个大肠埃希氏菌病原株和一个大肠埃希氏菌非病原株的基因组之间进行扣除，分离出一组基因组差异序列，该组差异序列中的每一种序列都与所述致病株的核酸杂交，但不与所述非致病株的核酸杂交。

已经将多种不同的基因组扣除方法应用于成对的相关株，以分离

病原体特异性基因组差异序列(例如，Mahairas 等，Journal of Bacteriology 178: 1274-1282, 1996; Tinsley 等, Proc. Natl. Acad. Sci. U.S.A. 93: 11109-11114, 1996)。已经应用这样的序列作为诊断标记以鉴定其它密切相关的株和对所述株进行指纹分析(见，例如，Darrasse 等, Applied and Environmental Microbiology 60: 298-306, 1994)。简要地说，应用基因组扣除于两个相关株的基因组 DNA，并分离基因组差异序列。将一组基因组差异序列与来自同一类群的其它株的基因组杂交(每种序列的杂交都在单独的杂交反应中进行)。与所述基因组杂交的基因组差异序列亚组在株与株之间各不相同，并因此构成鉴定指纹。虽然已经显示这种方法是鉴定一个生物类群内密切相关的成员的有力方法，但该方法对技术过分要求、耗时、麻烦，而无法在临床设置中执行。此外，在这些实验中的基因组差异序列通常来自单一的病原株，因此仅适用于对单一类群内的非常密切相关的株进行分型。因此，现有技术不能利用基因组差异序列在一个表现特异性测试中同时测试来自不同生物的多种序列。

这对于将一种生物鉴定为更大的生物类群的成员也是有用的。例如，可能重要的是确定下呼吸道感染是否是由于物种百日咳博德特氏菌(*Bordetella pertussis*)的任一成员引起的。在这种情况下，人们可以通过核酸杂交，测试在该物种的所有菌株中出现、但不在任何其它物种出现的序列的存在。这样的将一个类群的成员与其它类群的成员区分开来的 ID 序列称为类群特异性序列。

许多最具有医学意义和在诊断上最有用的遗传变异是单核苷酸多态性(SNP)。例如，在珠蛋白基因上的单碱基对改变是镰状细胞贫血的原因。结核分枝杆菌(*Mycobacterium tuberculosis*)中 RNA 聚合酶基因的单个碱基对改变是利福平抗性的原因，其中利福平是用于治疗肺结核的最重要的抗生素。已经发展出一次检测许多 SNP 的基于杂交的方法，但这些方法一般由于难以区分完全匹配和包含单个核苷酸错配的匹配而缺乏健全性(Gingeras 等, Genome Res. 8: 435-438, 1998; Wan

等, Science 280: 1077-1082, 1998)。一些用于辨别 SNP 的基因型的方法仅测试在单个基因上的突变(Gingeras 等, 1998, 见上文)。其它方法依赖于，不具有可重现性的多重 PCR 法。因此，需要利用健全的杂交和扩增方法学一次辨别许多 SNP 的基因型的方法。

因此，为鉴定生物，测试 ID 序列的存在是有用的，所述 ID 序列可以包括基因组差异序列和/或类群特异性序列。不用培养医学样品而测试 ID 序列需要检测少量基因组(如 100-1000 个基因组)的方法。已经发展出依赖于核酸扩增的灵敏方法，但一般地说，如同上文关于多重 PCR 所描述的，这些方法仅能可靠地一次应用于非常少量的序列。因此，已经批准用于临床使用的基于扩增的灵敏方法一次仅测试一种或二种病原体。这些测试比在临床实验室中进行的标准微生物测试昂贵得多(通常是约 100 倍)。因此，基于扩增的测定的商业化发展一直局限于一个小亚组的导致常见和严重的感染、并且不能在培养物中容易地生长的生物(如 HIV、结核分枝杆菌、和沙眼衣原体(*Chlamydia trachomatis*))。需要扩展这种技术对于日常诊断的能力和灵敏度。

最后，定量生物样品中的病原体数量常常是重要的。例如，用于诊断下呼吸道感染(如肺炎)的样品常常受到来自上呼吸道的正常共生菌群的污染。许多在上呼吸道无害的物种在破坏呼吸系统的正常防御后可能成为下呼吸道感染的原因，这进一步增加了诊断复杂性。在这种情况下，关于在下呼吸道样品中的生物数量的知识对于区别上呼吸道污染和下呼吸道感染是重要的。

假如能够培养所述生物，那么定量分析临床样品中的病原体是相对简单的。然而，许多在医学上重要的生物难以或不可能培养(如大多数病毒、寄生虫、衣原体和厌氧性细菌)。此外，定量培养通常需要几天，在某些情况下需要一个月以上，如培养引起肺结核的结核分枝杆菌。在有限的情况下，通过不需要培养的方法可以获得定量数据，例如直接免疫荧光测定。用于定量分析病原体的新的分子生物学方法，如定量聚合酶链式反应(PCR)已经对于监测 AIDS 患者体内的病毒水平

非常重要。然而，定量扩增方法极其难于正确设计，可能是没有重复性的，目前一次仅能应用于单个物种。

因此，需要测定生物样品或临床样品中病原体数量的方法。这样的方法最好是快速而普遍适用的，即该方法不需要培养并且可以定量在样品中可能存在的多种类型生物。

总的来说，需要健全而灵敏的鉴定方法，快速而准确地测试未经培养的样品中大量病原体特异性序列(基因组差异序列和类群特异性序列和单核苷酸多态性)，所述病原体特异性序列是可以引起特定表现(如肺炎)的一组不同传染因子的鉴别。也需要这样一种测试以提供关于该样品所来自的个体的医学和法医信息。

发明概述

在一方面，本发明提供了称为基因组分布分析(Genomic profiling)的方法，同时测试未知生物样品中多种(如多于5种)不同类型生物的诊断性核酸序列(包括基因组差异序列、类群特异性序列和DNA多态性)的存在。基因组分布分析代表了对现有方法的显著改良，因为该方法(1) 同时扫描样品中广谱的生物(如病毒、细菌、真菌、寄生虫和人类细胞)的存在，(2) 提供高分辨率遗传鉴定信息，(3) 测试特定突变(如那些隐藏的遗传疾病或抗生素抗性)，(4) 提供速度和简单性，(5) 不需要限制性的并耗时的培养步骤，(6) 使得有可能灵敏地测试复杂“原始”样品中比以前可能测试的数量更大数量的鉴别序列，(7) 通过引入高度冗余性和内部对照获得健全性，以及(8) 提供定量样品中靶生物的数量的方法。这种属性的组合使得能够对传染病进行新型的全面、表现特异性诊断测试。例如，基因组分布分析使得有可能为患有呼吸系统症状的个体提供单一测试，所述测试同时并且快速地扫描所有常见呼吸系统病原体的存在，所述呼吸系统病原体包括不同的病原体如细菌、病毒和真菌。

因此，本发明的一个方面是从可能包含靶核酸分子的生物样品中

获取遗传信息的方法，该方法包括: (a) 提供这样的核酸分子，即(i) 样品中的靶核酸分子，或(ii) 与样品中靶核酸分子杂交的探针，或(iii) (i)或(ii)的扩增产物，或(iv) (i)的基因组代表(genomic representation); (b)通过将(a)的核酸分子与最小基因组起源(genomic derivation)大于 5 (如大于 11)并且包括能够检测靶核酸分子的检测序列的一个检测集合(ensemble)相接触或比较，检测靶核酸分子。该方法还可以包括步骤(c): 鉴定在步骤(b)中检测到的核酸分子。

在优选的实施方案中，步骤(a)的核酸分子在步骤(a)之前并不作为以大小分级的片段固定化在基质或固相支持体上; 所述扩增步骤使用少于四对(如一对)扩增序列进行，如果靶核酸分子在所述样品中存在，则将产生扩增产物; 以及通过原位杂交用所述方法定量在生物样品中的靶生物。

在下面实施例 2 中作为例子展示的该方法的优选形式涉及在步骤(a)之前，使所述样品的核酸分子同时与用于产生上面步骤(a)(ii)的探针的 ID 探针集合杂交的步骤。

步骤(a)(ii)的探针最好包括(i) 能够与靶核酸分子杂交的第一个区，和(ii)扩增序列。可以进行杂交，以便使步骤(a)中的所有核酸分子都处于液相中，或者使得步骤(a)中的至少一部分核酸分子固定到固相支持体上。此外，至少步骤(a)的一些核酸分子可以包括一个或多个寡核苷酸标记。

至少步骤(a)(ii)的一些探针可以包括(i) 在与靶核酸分子杂交时可以相互连接的两种或更多种寡核苷酸，和(ii) 扩增序列。

在另一实施方案中，所述核酸探针集合的至少 50%的探针能够与在所述样品或所述样品的基因组代表中可能存在的预定的基因组差异序列杂交。

在一个优选的实施方案中，如上面所述可以与另一寡核苷酸连接的寡核苷酸是 SNP 探针。至少部分所述 SNP 探针可以包括标记序列，所述标记序列能够与包含标记序列集合的检测集合中的一种标记序列

杂交。在这些实施方案中所述检测集合的最小基因组起源可以是，例如，大于二十(如大于五十)。

在一些优选实施方案中，所述检测集合的检测序列作为两维的点或作为平行带(strip)在固相支持体上排列。

在另一实施方案中，通过使用不多于四对的扩增序列扩增步骤(a)(i)的靶核酸分子，产生步骤(a)(iv)的扩增产物，所述扩增序列如指导使用 Alu 特异性引物扩增处于 Alu 重复序列间的序列的扩增序列。在这些实施方案中，(b)的检测集合可以包括与在步骤(a)(iv)中可能扩增的 ID 探针相应的 ID 位点。

本发明可以用于检测和定量任何类型的生物。例如，在一个优选实施方案中，ID 探针集合包括与来自分别属于不同属的至少十种不同的病毒、每种病毒至少两种不同核酸分子杂交的探针。

本发明可以连同许多类型的生物样品使用，所述生物样品包括临床样品。在一个实施例中，所述生物样品是来自人类胃肠道的样品，并且使用本发明的方法所获得的遗传信息可以鉴定所述样品中来自六种或更多种以下生物的核酸分子：大肠埃希氏菌、沙门氏菌属(Salmonella)、志贺氏菌属(Shigella)、小肠结肠炎耶尔森氏菌(*Yersinia enterocolitica*)、霍乱弧菌(*Vibrio cholera*)、粪弯曲杆菌(*Campylobacter fecalis*)、艰难梭菌(*Clostridium difficile*)、轮状病毒属(Rotavirus)、诺沃克病毒(Norwalk virus)、星状病毒属(Astrovirus)、腺病毒属(Adenovirus)、冠状病毒属(Coronavirus)、兰氏贾第鞭毛虫(*Giardia lamblia*)、溶组织内阿米巴(*Entamoeba histolytica*)、人酵母菌(*Blastocystis hominis*)、隐孢子虫属(Cryptosporidium)、Microsporidium、美洲板口线虫(*Necator americanus*)、人蛔虫(*Ascaris lumbricoides*)、毛首鞭虫(*Trichuris trichiura*)、蛲虫(*Enterobius vermicularis*)、粪类圆线虫(*Strongyloides stercoralis*)、麝后睾吸虫(*Opsthorchis viverrini*)、华支睾吸虫(*Clonorchis sinensis*)和短膜壳绦虫(*Hymenoplepis nana*)。

在另一实施方案中，所述生物样品是呼吸道样品，并且所述遗传

信息可以鉴定来自以下六种或更多种生物的核酸分子：白喉棒杆菌(*Cornybacterium diphtheriae*)、结核分枝杆菌(*Mycobacterium tuberculosis*)、肺炎支原体(*Mycoplasma pneumoniae*)、沙眼衣原体(*Chlamydia trachomatis*)、肺炎衣原体(*Chlamydia pneumoniae*)、百日咳博德特氏菌(*Bordetella pertussis*)、军团菌(Legionella spp.)、诺卡氏菌(Nocardia spp.)、肺炎链球菌(*Streptococcus pneumoniae*)、流感嗜血菌(*Haemophilus influenzae*)、鹦鹉热衣原体(*Chlamydia psittaci*)、铜绿假单胞菌(*Pseudomonas aeruginosa*)、金黄色葡萄球菌(*Staphylococcus aureus*)、荚膜组织胞浆菌(*Histoplasma capsulatum*)、*Coccidoides immitis*、新型隐球酵母(*Cryptococcus neoformans*)、皮炎芽生菌(*Blastomyces dermatitidis*)、卡氏肺囊虫(*Pneumocystis carinii*)、呼吸道合胞病毒、腺病毒属(Adenovirus)、单纯疱疹病毒、流感病毒、副流感病毒和鼻病毒属(Rhinovirus)。

另一种可以根据本发明测试的生物样品是血液样品，其中鉴定来自至少六种以下生物的核酸分子：凝固酶阴性葡萄球菌、金黄色葡萄球菌、*Viridans streptococci*、肠球菌属(Enterococcus spp.)、β溶血性链球菌、肺炎链球菌、埃希氏菌(Escherichia spp.)、克雷伯氏菌(Klebsiella spp.)、假单胞菌(Pseudomonas spp.)、肠杆菌(Enterbater spp.)、变形虫(Proteus spp.)、拟杆菌(Bacteroides spp.)、梭菌(Clostridium spp.)、铜绿假单胞菌、棒杆菌(Cornybacterium spp.)、疟原虫(Plasmodium spp.)、杜氏利什曼原虫(*Leishmania donovani*)、弓形虫(Toxoplasma spp.)、微丝蚴(Microfilariae)、真菌、荚膜组织胞浆菌、*Coccidoides immitis*、新型隐球酵母、假丝酵母(Candida spp.)、HIV、单纯疱疹病毒、丙型肝炎病毒、乙型肝炎病毒、巨细胞病毒属(Cytomegalovirus)和EB病毒。

本发明还可用于鉴定在任何类型生物样品中的核酸分子，其中所鉴定的核酸分子是以下生物中的六种或更多种的核酸分子：柯萨奇病毒A、单纯疱疹病毒、圣•路易脑炎病毒、EB病毒、粘液病毒、JC病毒、柯萨奇病毒B、披膜病毒、麻疹病毒、肝炎病毒、副粘病毒、艾

可病毒、布尼亚病毒、巨细胞病毒、水痘-带状疱疹病毒、HIV、腮腺炎病毒、马脑炎病毒、淋巴细胞性脉络丛脑膜炎病毒、狂犬病病毒和BK病毒。

本发明还包括用于从可能包含靶核酸分子的生物样品获取遗传信息的方法，所述方法包括(a) 提供最小基因组起源大于五的核酸探针集合；(b) 使所述探针集合同时与所述样品的核酸分子接触；(c) 检测在所述探针和所述样品中任何靶核酸分子间的杂交；和(d) 鉴定在步骤(c)中检测到的核酸分子。

本发明还包括用于从生物样品中获取遗传信息的试剂盒，所述试剂盒包括：(a) 多种ID探针和/或SNP探针；和(b) 包括与(a)的探针相应的检测序列并且最小基因组起源大于五(如大于十一)的检测集合。

在优选实施方案中，(a)的探针包括多于十种(如多于五十种或多于两百五十种)不同的可扩增探针；(a)的至少50%的探针包括来自至少三种不同物种的基因组差异序列；(a)的探针包括多于五个家族的可扩增探针；并且(a)的探针对于至少两个不同分类单位、两个不同物种、两个不同属或两个不同界是特异性的。

在其它优选实施方案中，(a)的探针包括包含以下的探针：(i) 在与靶核酸分子的ID序列杂交时可以相互连接的两种或更多种寡核苷酸，和(ii)扩增序列。

在其它实施方案中，(a)的探针和/或(b)的检测序列物理性附着到固相支持体的不同位点。在这些实施方案中，检测集合的检测序列可以在所述支持物上彼此相邻定位，其中所述检测序列检测(i) 分类群的成员(ii) 密切相关的分类群。

本发明还包括用于从生物样品中获取遗传信息的试剂盒，所述试剂盒包括：(a)能够引发生物样品中的靶基因组DNA中由重复序列(如人类Alu重复序列)邻接的DNA序列的扩增以产生四探针的多种核酸引物(如Alu特异性引物)；和(b) 检测集合，所述检测集合包括与使用(a)的引物可能扩增的ID探针相应的检测序列，所述检测集合的最小基

因组起源大于 5 (如大于二十)。

本发明还包括 ID 探针集合，所述 ID 探针集合可以使用少于四对扩增序列扩增，包括多于三个(如多于十个或多于二十五个) ID 探针家族以及多于十种(如多于五十种或多于两百五十种)不同的 ID 探针。

在优选实施方案中，多于两个可扩增探针家族对于不重叠的分类单位、不同物种、不同属或不同界具有特异性。至少 50%的所述探针可以包括来自至少三个不同物种的基因组差异序列。

在其它优选实施方案中，(a)的探针包括包含以下的探针：(i) 在与靶核酸分子的 ID 序列杂交时可以相互连接的两种或更多种寡核苷酸，和(ii) 扩增序列。

在其它优选实施方案中，检测集合中包括的检测序列在支持物上相互邻接定位，其中所述检测序列检测(i) 一个分类群内的成员和(ii) 密切相关的分类群。

在本发明中使用的程序和试剂是通用的，即一组试剂可以用于鉴定许多不同类型的生物。所述测试是快速的，并且可以简单地加入阳性内部对照和阴性内部对照。本发明的方法可以产生高分辨率遗传指纹，鉴定用常规方法无法分辨的株。所述方法适合于自动化形式，并且不需大量人员培训就可进行。

本发明具有广泛的应用性，包括对微生物(如细菌、真菌和原生动物)分型；鉴定高等生物(包括人类)的基因型；以及在流行病学中，监测医院和地理遥远地区的传染病爆发(infection outbreak)。本发明的方法还可用于环境测试、农业(以进行家畜育种和分析)以及如在种子产业中进行植物分型。人类法医学代表着本发明的又一个应用。

本发明的一个关键特征在于其能够在一次测定中，测试可用于鉴定复杂生物样品中的生物的 ID 序列集合。该组 ID 序列包含多种区分一个分类群内的成员(如不同的大肠埃希氏菌株)的基因组差异序列，以及在不同分类群(如不同物种或属)之间进行区分的多种类群特异性序列。这样，每个集合可以包括非常大的一系列不同 ID 序列，所有这

些 ID 序列都可以在一个快速、不基于凝胶的测定中同时使用。不需要培养样品的事实增强了所述测试的快速性。

根据下面的详细描述、附图和权利要求书，本发明的其它方面和好处将变得显而易见。

定义

“基因组”是指在一种生物中作为该生物可遗传遗传信息的最终来源的核酸分子。对于大多数生物，基因组主要由染色体 DNA 组成，但基因组也可以包括质粒、线粒体 DNA 等等。对于一些生物如 RNA 病毒，基因组由 RNA 组成。

“核酸”是指 DNA、RNA 或其它可以包括相似部分的取代的相关物质组合物。例如，核酸可以包括不在 DNA 或 RNA 中发现的碱基，所述碱基包括但不限于 DNA 中的黄嘌呤、肌苷、尿嘧啶，RNA 中的胸腺嘧啶，次黄嘌呤等等。核酸还可以包括磷酸或糖部分的化学修饰，可以引入所述化学修饰以改善稳定性、对酶降解的抗性、或一些其它有用的特性。

“寡核苷酸”或“寡核苷酸序列”是指长度从 6 个碱基到 150 个碱基的核酸。寡核苷酸一般但不一定在体外合成。6 个碱基到 150 个碱基长、并且是更大序列的亚序列的核酸区段也可称为寡核苷酸序列。

“靶序列”或“靶核酸序列”是所指设计的探针所要检测的核酸序列。对于 ID 探针，靶序列可以是 ID 序列中的 ID 位点。对于 SNP 探针，靶序列可以是单核苷酸多态性。

“靶生物”或“靶类群”是指诊断测试所设计要检测的一类生物或生物类群(分类单位)。

“杂交”是指由碱基对的氢键介导的核酸分子非共价结合。

“有意义的杂交”是指一种探针分子或多种探针分子与所述探针所设计检测的核酸序列的杂交，其中所述杂交导致检测出信号。

“比较杂交条件”是指如国际系统细菌学委员会(International Committee on Systematic Bacteriology)所推荐的，用于将物种相互区分开的条件(Wayne 等, Internat. J. System. Bacteriol. 37: 463-464, 1987)。比较杂交条件在本文中是指由 Hartford 等(Int. J. Syst. Bacteriol. 43: 26-31, 1993)使用的条件。

“扣除杂交条件”是指在严格性上等同于如下反应的严格性的条件：所述反应在 65℃下，在由 10 mM EPPS, pH 8.0 和 1 M NaCl 组成的缓冲液中进行。

“发现于”、“存在于”、“出现于”、“对应于”、“杂交于”或“处于”另一核酸序列、核酸分子、寡核苷酸、探针或基因组的核酸序列、核酸分子、寡核苷酸或探针，是指可以与另一序列、寡核苷酸、探针或基因组形成杂交体的序列、寡核苷酸或探针，并且与由进行比较的两种核酸分子中较短的一种核酸分子与其完全互补物在由 10 mM EPPS, pH 8.0 和 1 M NaCl 构成的缓冲液中组成的双链 DNA 片段相比，所述杂交体的解链温度(T_m)比所述双链 DNA 片段的 T_m 低 20℃(对于大于 30 bp 的序列)、12℃(对于 15 bp 到 20 bp 的序列)或 8℃(对于 8 bp 到 14 bp 的序列)。“不存在于”另一核酸序列、核酸分子、寡核苷酸、探针或基因组的核酸序列、核酸分子、寡核苷酸或探针，是指没有在另一核酸序列、核酸分子、寡核苷酸、探针或基因组中发现的核酸序列、核酸分子、寡核苷酸或探针。

“ID 序列”或“鉴定序列”是指这样一种核酸序列：当在基因组或富集的基因组(见下文)中，通过杂交使用如上文定义所述的长度特异性解链温度标准测定所述核酸序列的存在时，所述核酸序列是特定生物或生物类群的诊断性序列。ID 序列对应于基因组或富集的基因组中长度大于等于 30 bp、可用于将一种类型生物与另一类型生物区分开来的序列。例如，当重要的是将密切相关的类群的成员相互区分开来时，基因组差异序列可以用作 ID 序列。“类群特异性序列”是可用于将一个类群的所有成员与其它类群区分开来的一种类型的 ID 序列。

“基因组差异序列”是指在一种生物的基因组(或富集的基因组)中发现、而未在密切相关的生物的基因组(或富集的基因组)中发现的核酸序列或核酸序列集合体。通过杂交/扣除技术、通过使用计算机比较基因组序列、或通过多种其它技术中的任何一种，可以发现基因组差异序列。比较基因组(或富集基因组)的生物必须是密切相关的。如果一对生物是同一属的成员，或者如果它们的基因组满足下面特定的杂交标准(请注意国际系统细菌学委员会推荐使用比较杂交建立相关性(Wayne 等, 1987, 见上文))，就认为它们是“密切相关的”。假如使用 Hartford 等(1993, 见上文)描述的方法，在比较杂交条件下，一对生物的多于 70%的基因组 DNA 片段(在具有 RNA 基因组的病毒的情况下，是基因组 cDNA 片段)可以相互杂交，那么就认为它们是“密切相关的”。基因组差异序列的长度大于等于 30 bp。基因组差异序列的一个例子是出现在大肠埃希氏菌 O157:H7 的一个致病株、但不出现在大肠埃希氏菌 O157:H7 的另相应病株中的 DNA 片段。

“类群特异性序列”是指这样的核酸序列或核酸序列集合体：当在比较杂交条件下进行杂交时，所述核酸序列或核酸序列集合体是一个系统发生类群中生物的基因组的特征，而不是另一个分类单位或系统发生类群的基因组的特征。类群特异性序列的长度大于等于 30 bp。例如，在大肠埃希氏菌 O157:H7 类群的 99%以上的分离物中出现、但不在 99%以上的沙门氏菌属分离物中出现的片段是类群特异性序列。相似地，在 99%以上的轮状病毒分离物中出现(如在比较杂交条件下所鉴定)、但不在于 99%以上的人免疫缺陷病毒分离物中出现的片段是类群特异性序列。类群特异性序列可以用于鉴定更低水平的分类群，如亚种或通过世代相关联的杂种繁殖群体(如人类)的成员。注意：为了诊断目的，类群特异性序列在出现于一个分类群内，而不出现于相似分类学水平的姐妹类群(sister group)内时最有用。

类群特异性序列的一个例子是在肠沙门氏菌鼠伤寒血清型(*Salmonella enterica* serotype Typhimurium)的基本所有分离物中发现、

但基本在肠沙门氏菌乙型副伤寒血清型(*Salmonella enterica* serotype Paratyphi B)的分离物中未发现的序列(见图 6)。请注意，类群特异性序列也可以是基因组差异序列(也就是说，该组类群特异性序列与该组基因组差异序列重叠)。例如，在所有大肠埃希氏菌 O157:H7 菌株中出现、但在大肠埃希氏菌的非 O157:H7 菌株中未发现的序列既是基因组差异序列，也是类群特异性序列。

“保守序列”是指这样的核酸序列或核酸序列集合体：按照杂交标准，所述核酸序列或核酸序列集合体是跨越同一分类学水平上多个独立分类群的生物的基因组的特征。保守序列的长度大于等于 30 bp。因此，编码人类 RNA 聚合酶的基因的许多片段的序列是保守序列，因为它们可以在比较杂交条件下与黑猩猩基因组杂交。保守序列不可用于区分带有所述保守序列的类群的成员。

“ID 探针”是指用于与生物样品中的 ID 序列杂交的寡核苷酸或一对寡核苷酸或一组寡核苷酸。为进行杂交，所述探针寡核苷酸的一部分必须能够与对应的 ID 序列进行碱基配对。所述探针的该部分通常长度在 8 个碱基到 120 个碱基之间。ID 探针也可以具有其它部分，所述部分包括扩增位点(例如，对应于用于 PCR 扩增的引物结合位点的序列)和作为检测时的标记的序列(见下文)。

“基因组差异探针”是指与基因组差异序列对应、即与其杂交的 ID 探针。

“类群特异性探针”是指与基因组差异序列对应、即与其杂交的 ID 探针。

“ID 探针位点”或“探针位点”是指 ID 序列中在序列上对应于 ID 探针的部分。

“ID 序列家族”是指可以与一种(非重组)生物的基因组杂交(在比较杂交条件下)的包含 2 个或更多成员的一组 ID 序列。在所述家族的 ID 序列中，至少 2 种 ID 序列在它们天然和通常出现的基因组中在图谱中距离大于 3,000 碱基。一个 ID 序列家族可以包括类群特异性序列

和基因组差异序列的组合，可以仅包括类群特异性序列，或可以仅包括基因组差异序列。

例如，考虑可用于追踪传染性大肠埃希氏菌 O157:H7 的爆发的 ID 序列家族。该 ID 序列家族可以包括所有下面类型的有诊断用途的 ID 序列：物种大肠埃希氏菌的所有成员所共有并且限于该物种所有成员的多种类群特异性序列；仅包含大肠埃希氏菌 O157:H7 菌株的系统发生类群的所有成员所共有并且限于所述系统发生类群所有成员的多种类群特异性序列；仅包含大肠埃希氏菌 O157:H7 的系统发生类群的所有成员所共有并且限于所述系统发生类群所有成员的多种类群特异性序列，其中所述大肠埃希氏菌 O157:H7 经多酶电泳分析发现具有电泳型 3 (DEC3 类群; Whittam 等, Infect. Immun. 61: 1619-1629, 1993)；以及在大肠埃希氏菌 O157:H7 参考菌株 DEC3B 中存在，但在大肠埃希氏菌 O157:H7 参考菌株 DEC4C 中不存在的多种基因组差异序列。

请注意，在上面的例子中，所述 ID 序列家族都可以在比较杂交条件下与一种生物即大肠埃希氏菌 O157:H7 参考菌株 DEC3B 的基因组杂交。这是表达方式“ID 序列家族”的定义方面。

“寡核苷酸家族”或“探针家族”是指对应于 ID 序列家族的寡核苷酸或探针的集合体。在寡核苷酸或探针家族中的所有寡核苷酸或探针序列对应于特定 ID 序列家族中所有或部分成员的序列。

“多态性探针”或“单核苷酸多态性探针”或“SNP 探针”是指这样一组寡核苷酸：当该组寡核苷酸与基因组杂交时，邻接一个多态性位点，并且该组寡核苷酸具有在该位点与一段特定的在该位点出现的基因组序列发生精确碱基配对的序列。当一组这样的寡核苷酸邻近地与基因组杂交时，只有在靶位点的等位基因或基因型符合所述多态性探针的寡核苷酸的邻接序列时，这些寡核苷酸才可以相互连接。SNP 探针的结构和应用显示于图 10。一般地说，合成一组多态性探针以使其对应于特定位点的每个等位基因。多态性探针可以包含 ID 探针所包含的同样部分(如扩增位点和标记)。具有标记序列的多态性探针的集

合可用于产生包含差异的富集的基因组样本，其中所述差异可以通过与包含标记集合的检测集合杂交而检测出。

多态性探针或“单核苷酸多态性探针”或“SNP 探针”“家族”的定义与 ID 序列家族和 ID 探针家族的定义类似，只是在这种情况下，探针和基因组 DNA 之间的对应性在于成对半边探针(probe-half)与多态性基因组位点(如单碱基对多态性)杂交并且与所述位点精确邻接的能力，而不是基于针对 ID 序列使用的杂交标准(见图 10)。为了定义 SNP 探针家族，仅考虑用每种 SNP 探针测试的一个等位基因。仅考虑用具有最小等位基因频率的特定 SNP 探针测试的 SNP 等位基因。该等位基因定义为“最罕见的 SNP 等位基因靶”。“等位基因频率”是在一个物种的群体中，针对基因组中在特定基因座的特定等位基因定义。等位基因频率是在群体中，在该基因座的所有等位基因中特定等位基因所占的分数(King, 等人, *A dictionary of genetics* (Oxford University Press, New York, 1990)。用于确定等位基因频率的群体样本必须包括至少 100 个(不是纯系相关的(non-clonally related))个体。SNP 探针家族是一组 SNP 探针，该组 SNP 探针中最罕见的 SNP 等位基因靶都出现在一个个体的基因组中。

“标记”或“标记序列”是指可以掺入更大寡核苷酸或探针中的非生物的寡核苷酸序列。标记序列可以用作检测序列。例如，在检测阵列中的标记序列可以用于通过杂交而检测在所扩增的探针中的(互补)标记序列。当不同的诊断序列不能用其它方法通过杂交进行区分时(如 SNP 探针；见下文)，可以使用标记序列通过杂交将探针相互区分开来。

同样，“标记序列家族”或“标记家族”是指对应于一个探针家族的一组标记序列。例如，在下面的实施例 5 中，将多态性探针或 SNP 探针的集合与人类基因组 DNA 样品杂交。可以被连接和扩增的 SNP 探针集合的亚组是一个 SNP 探针家族。由于一个 SNP 探针家族对应于一个人类个体的基因型，因此该家族的定义与 ID 探针家族相似。所述

SNP 探针家族包含一个标记序列家族(一般构建 SNP 探针时加入识别标记序列)。因此，该 SNP 探针家族与所述标记探针家族相应，并且可以通过与在检测集合中的相应标记序列家族杂交而鉴定。

相应的序列组是指在各组的元件之间存在一一对应。例如，考虑与一个 ID 序列集合相应的 ID 探针集合。每种 ID 探针包含位于一种 ID 序列中的一个 ID 位点，而每种 ID 序列对应于一种 ID 探针。或者，考虑由与一个多态性探针集合相应的标记集合组成的检测集合。在该检测集合中的每种标记对应于在所述多态性探针集合中一种多态性探针中的一种标记。相似的，一个标记序列家族可以与一个多态性探针家族相应。

“最小基因组起源”是指一组序列、探针、寡核苷酸或标记可以杂交的不同基因组的最小数目(或不同基因组代表的最小数目)。例如，一组 ID 序列的最小基因组起源等同于由一组 ID 序列可以构建的家族的最小数目。因此，例如，一组 ID 序列，该组中每种序列对应于一种不同人类基因的一个蛋白编码区段，该组 ID 序列的最小基因组起源是一，因为整组序列可以与一个人的基因组杂交。作为另一个例子，考虑由一对类群特异性腺病毒序列和一对类群特异性呼吸道合胞病毒序列组成的一组序列。这样一组序列的最小基因组起源是 2，因为 2 个基因组的序列，即腺病毒和呼吸道合胞病毒的序列是在比较杂交条件下足以与所有 4 种序列杂交的最小基因组数目。该组 4 种 ID 序列组成 2 个 ID 序列家族，只要每对病毒 ID 序列在来源的基因组中被分开大于等于 3000 bp(见上面“家族”的定义)。

考虑在表 1 中举例说明的一个更复杂的例子也是有帮助的，在该例子中，一组 ID 序列可以用于测试患有急性胃肠疾病的患者中某些病原体的存在。注意：在表 1 每个格子中的序列组可以与单个个体的基因组 DNA 杂交。(在表 1 中有 9 个这样的格子。)同时，注意不可能使表 1 所述 9 个格子中包含的所有序列与少于 9 个个体的基因组 DNA 杂交。因此，表 1 中 ID 序列组的最小基因组起源是 9。

表 1. 一个最小基因组起源为 9 的 ID 序列集合。下表中的每个格子包括一个 ID 序列“家族”(即可以与一个基因组杂交的一组序列)。

大肠埃希氏菌 O157:H7 基因组差异序列 2 (存在于大肠埃希氏菌 O157:H7 X 菌株中，但不存在于大肠埃希氏菌 O157:H7 Y 菌株中) 大肠埃希氏菌 O157:H7 类群特异性序列 A 大肠埃希氏菌 O157:H7 类群特异性序列 B 大肠埃希氏菌类群特异性序列 A 大肠埃希氏菌类群特异性序列 B
大肠埃希氏菌 O157:H7 基因组差异序列 3 (存在于大肠埃希氏菌 O157:H7 Y 菌株中，但不存在于大肠埃希氏菌 O157:H7 X 菌株中) 大肠埃希氏菌 O157:H7 基因组差异序列 4 (存在于大肠埃希氏菌 O157:H7 Y 菌株中，但不存在于大肠埃希氏菌 O157:H7 X 菌株中) 大肠埃希氏菌 O157:H7 类群特异性序列 A 大肠埃希氏菌 O157:H7 类群特异性序列 B 大肠埃希氏菌类群特异性序列 A 大肠埃希氏菌类群特异性序列 B
大肠埃希氏菌 O55:H6 基因组差异序列(存在于一个大肠埃希氏菌 O55:H6 菌株中，但不存在于另一个大肠埃希氏菌 O55:H6 菌株中) 大肠埃希氏菌类群特异性序列 A
肠沙门氏菌鼠伤寒血清型基因组差异序列 1 (存在于一个肠沙门氏菌鼠伤寒血清型菌株中，但不存在于另一个肠沙门氏菌鼠伤寒血清型菌株中) 肠沙门氏菌鼠伤寒血清型基因组差异序列 2 (存在于一个肠沙门氏菌鼠伤寒血清型菌株中，但不存在于一个肠沙门氏菌乙型副伤寒血清型菌株中) 肠沙门氏菌类群特异性序列 肠沙门氏菌鼠伤寒血清型类群特异性序列

肠沙门氏菌乙型副伤寒血清型基因组差异序列 1 (存在于一个肠沙门氏菌鼠伤寒血清型菌株中，但不存在于另一个肠沙门氏菌乙型副伤寒血清型菌株中) 肠沙门氏菌乙型副伤寒血清型基因组差异序列 2 (存在于一个肠沙门氏菌鼠伤寒血清型菌株中，但不存在于另一个肠沙门氏菌鼠伤寒血清型菌株中) 肠沙门氏菌类群特异性序列 肠沙门氏菌乙型副伤寒血清型类群特异性序列
粪弯曲杆菌基因组差异序列 1 (存在于粪弯曲杆菌 X 菌株中，但不存在于粪弯曲杆菌 Y 菌株中) 粪弯曲杆菌基因组差异序列 2 (存在于粪弯曲杆菌 X 菌株中，但不存在于粪弯曲杆菌 Z 菌株中)
轮状病毒类群特异性序列 1 轮状病毒类群特异性序列 2 轮状病毒类群特异性序列 3
诺沃克病毒类群特异性序列 1 诺沃克病毒类群特异性序列 2 诺沃克病毒类群特异性序列 3
兰氏贾第鞭毛虫基因组差异序列 1 兰氏贾第鞭毛虫基因组差异序列 2

应用于 SNP 探针集合和标记序列集合的最小基因组起源的定义如下文所定义。一个 SNP 探针的集合包括多个 SNP 探针家族，并且每个 SNP 探针家族对应于一个个体的基因型。然而，与 ID 序列集合不同，一个 *SNP 探针集合的最小基因组起源一般是一*。这是因为 SNP 探针一般可以与任何靶物种的基因组以不多于一个碱基对错配进行杂交。

现在考虑一个人类 SNP 探针集合，所述 SNP 探针集合的每种探

针都包括一种独特的标记序列部分。同时，考虑包含与所述 SNP 探针集合中的标记序列相应的一个标记集合的检测阵列。所述 SNP 探针集合的最小基因组起源一般是一，因为所有成员都可以与任何特定人类基因组杂交。然而注意：与此不同，对应的标记集合可能具有大的最小基因组起源。为理解这一明显自相矛盾的说法，认识到以下事实是有帮助的：所述 SNP 探针集合由多个 SNP 探针家族组成，其中每一个 SNP 探针家族对应于一个个体的基因型。在 SNP 探针家族中的标记序列组是标记序列的对应家族。在所述检测阵列中的对应标记序列家族可以与这样一个 SNP 探针家族杂交。然而，在所述标记集合中的其它标记序列不能与该 SNP 探针家族杂交。因此，*与一个 SNP 探针集合相应的一个标记序列集合的最小基因组起源等于所述 SNP 探针组合中的家族数目，即使所述 SNP 探针组合本身的最小基因组起源是 1*。

最小基因组起源的定义在应用于标记集合时依赖于下面的定义。回忆针对特定 SNP 探针的“最罕见 SNP 等位基因靶”的定义(见上面“SNP 探针家族”的定义)。我以相似方式定义“最常见 SNP 等位基因靶”。因此，对于用特定 SNP 探针测试的等位基因靶，一个等位基因被确认在一个物种内是最罕见的，而一个等位基因被确认是最普遍的。一种 SNP 探针的“平均等位基因频率”定义为最常见的等位基因靶和最罕见的等位基因靶的等位基因频率的平均值。例如，假如用一种 SNP 探针可以检测到的等位基因以 0.85、.06 和 0.002 的频率出现，那么平均等位基因频率就是 0.426 (即，(0.85 + 0.002) ÷ 2))。“平均等位基因频率的乘积”(*P*)定义为在所述 SNP 集合中所有 SNP 的等位基因频率的乘积。因此，例如，考虑一个假设的测试，其中用 SNP 探针测试 36 个人类疾病突变，每个人类疾病突变都以 0.001 的等位基因频率出现，并且所述每个突变都与一个以 0.999 的等位基因频率出现的正常等位基因相关。对于所述 36 种 SNP 中的每一种来说，平均等位基因频率是 0.5 (即，(0.001 + 0.999) ÷ 2))。因此，平均等位基因频率的乘积(P)是 $0.5^{36} = 1.46 \text{ X } 10^{-11}$。(注意：对于实际的 SNP 探针集合

来说，等位基因频率和平均等位基因频率的值将随着不同探针而各不相同。此外，注意一种 SNP 探针的等位基因频率不一定要加到 1.0，因为并不是所有出现的等位基因都要用 SNP 探针进行测定)。

由于在实践中可能难以确定对于一个特定物种的包含一组 SNP 探针的最小家族数，我以下面方式定义与一个 SNP 探针集合相应的标记集合的最小基因组起源。一个标记集合的最小基因组起源定义为 $(10^{-10})(P)^{-1}$，其中 P 是平均等位基因频率的乘积。因此，在前面的例子中，对应于人类疾病突变 SNP 探针集合的标记集合的最小基因组起源是 $(10^{-10})(1.46 \text{ X } 10^{-11})^{-1} = 6.9$。与此不同，如上文所解释的，对应 SNP 探针集合的最小基因组起源是一。

我提供下面的例子，帮助理解与一组 SNP 探针相应的一组标记的最小基因组起源的定义的生物学解释。考虑一个 33 种标记的组，该组标记与一组非连接的人类 SNP 探针相应，其中每种 SNP 探针检测两个等位基因，这两个等位基因的等位基因频率都是 0.5。该组标记的最小基因组起源是 $(10^{-10})(P)^{-1} = (10^{-10})(0.5^{33})^{-1} = 0.85$，接近于一。注意：最有可能发现的基因型是在这 33 个 SNP 基因座中的每一个都是杂合的个体(在这样一个基因座上杂合的概率是 0.5)。发现具有最有可能的基因型的个体的概率是 $0.5^{33} = 1.2 \text{ X } 10^{-10}$。预期这样一个个体出现的概率稍稍小于在 2000 年总人口中出现一个(约 $6 \text{ X } 10^{9}$)。

检测集合可以包含与包括 ID 探针和 SNP 探针的探针集合相应的检测序列(即所述检测集合具有 ID 位点序列和标记序列)。这样一个集合的最小基因组起源是所述 ID 位点的最小基因组起源加上所述标记序列的最小基因组起源的总和。假如所述标记集合覆盖多于一种的物种，那么所述集合的最小基因组起源是对应于每个物种的最小基因组起源的总和。

“ID 序列集合”是指对应于多个 ID 序列家族的一组 ID 序列。也就是说，一个 ID 序列集合的最小基因组起源大于 1。此外，由于每个家族最少包含 2 种(完全分离的) ID 序列，故一个 ID 序列集合最少具

有 4 个 ID 序列成员。一个 ID 序列集合的特征是：一种生物的基因组不足以给出与所有个别 ID 序列的阳性杂交信号。ID 序列集合不一定与样品在物理上分开。而且可以仅仅将这样一个集合概念化，以方便设计 ID 探针用于构建探针集合(见下文)。图 1 图示了在表 1 中描述的最小基因组起源为 9 的 ID 序列集合。

“ID 寡核苷酸集合”或“ID 探针集合”是指寡核苷酸或探针的集合体，其中每种寡核苷酸或探针对应于一个特定 ID 序列集合中一种 ID 序列的全部或部分的核苷酸序列。这样的集合设计用于通过杂交，检测在样品中存在的对应于两种或更多种不同基因组的核酸序列(见下文)。最好在探针集合中，探针的序列和/或探针在水溶液中的浓度是已知的。

“SNP 探针集合”或“单核苷酸多态性探针集合”或“多态性探针集合”是指包含多于一个 SNP 探针家族的一组 SNP 探针。

“标记序列集合”或“标记集合”是指与一个探针集合相应的一组标记序列。也就是说，在一个标记序列集合中每种标记序列与一个探针集合的一种标记序列(或与一种标记序列的反向互补物)互补。标记序列集合可用于在基因组分布分析中将单核苷酸多态性基因型(难以通过杂交检测)转变为健全的杂交基因型(见下面的实施例 5)。

某些物理特性或化学特性的“集合”是指与核酸序列集合相应、涉及所述物理特性或化学特性的一组值。例如，存在与一个 ID 探针集合的分子量一一对应的一个分子量集合。这样一个分子量集合可以用作检测集合或检测数列，以确定一个 ID 探针集合中样品选择的亚组的元件的身份。可以通过质谱，分析所述探针亚组，并将观察到的分子量与所述分子量集合(即原始 ID 探针集合的分子量)比较。

“检测集合”或“检测序列的集合”是指称为“检测序列”的序列集合体，其中所述序列集合体中的所有序列都对应于一个序列、探针、寡核苷酸或标记的集合(如一个 ID 探针集合或 SNP 探针集合)的所有或部分成员。也就是说，检测集合与序列集合、探针集合、寡核苷

酸集合或标记集合相应。这样的集合设计用于检测(通常是通过杂交，但不一定通过杂交)下列集合中在诊断上能提供信息的亚组：ID 探针集合、ID 序列集合、多态性探针集合或其它包含在诊断上有用序列的基因组代表的集合。如下文所提到的，检测集合的成分(即检测序列)可以排列成为二维阵列，以利于诊断探针(如，已经与样品的核酸分子内的 ID 序列杂交的 ID 探针)的鉴定。或者，所述检测集合的元件可以与诊断探针在液体中接触。如下文所提到的，可以在接触检测集合之前，扩增已经与样品的核酸分子内的 ID 序列杂交的 ID 探针。

检测集合也可以是与序列集合、探针集合、寡核苷酸集合或标记集合一一对应(即与之相应)的一组物理或化学特性的值。例如，ID 探针集合的成员的分子量表或分子量数列是一种类型的检测集合。这样一个检测集合可以用于质谱分析鉴定 ID 探针集合的特定亚组。可以使用质谱确定临床样品所选择的 ID 探针家族的分子量。然后将该 ID 探针家族的分子量与分子量检测集合(即原始未经选择的 ID 探针集合的分子量)相比较。用这种方法，鉴定选定的 ID 探针，这进而导致鉴定所述临床样品中的基因组。或者，如下面的实施例 3 所述，可以通过与一个寡核苷酸检测集合的杂交检测探针家族。然后可以通过确定所述寡核苷酸的分子量，并将所述分子量与另一个检测集合相比较，鉴定所述探针选定的检测寡核苷酸亚组，所述检测集合是所述寡核苷酸检测集合的元件的分子量数列。

“二维检测阵列”是指 ID 序列、ID 寡核苷酸、ID 探针或检测序列的集合，所述 ID 序列、ID 寡核苷酸、ID 探针或检测序列已经通过非电泳方法排列到基本上两维的(即平面的)固相支持体上，例如尼龙滤膜或聚赖氨酸包被的玻璃载玻片上。

“基因组分布分析测定”是指本发明的某些方法。

“基因组分布分析指纹”或“指纹”是指根据通过基因组分布分析扩增和检测的诊断探针，推测在生物样品中存在的诊断序列(如 ID 探针或 SNP 探针)亚组。

“分类单位”或“系统发生类群”是指单系群的集体成员，所述单系群是从一种共同祖先生物类型(或者是已知的，或者是假设的)遗传下来并且包括所述共同祖先生物类型的生物类型类群。注意：为本发明的目的，分类单位以并不暗示任何分类学水平的一般意义使用。因此，例如，分类单位在亚种等级上定义，也在属、纲、门等的等级上定义。

“独立分类群”或“独立分类单位”是指没有重叠成员的分类单位。因此，细菌肠杆菌属和沙门氏菌属是独立分类单位。然而，肠杆菌属和由大肠埃希氏菌 O157:H7 病原体组成的分类群不是独立的分类单位，因为该致病菌株的所有成员也都是该属的成员。

“分类学等级”是指一个分类单位在系统发生等级体系中的位置。术语分离物、生态型、亚种、物种、属、科、纲、目、门、界和超界是分类学等级的例子。

生物的“界“是指下面列举的其中一种：病毒、细菌、古细菌、真菌、原生动物、植物和动物。

“独特基因组”是指具有与所有其它基因组的核酸序列(除了遗传上相同的生物的基因组的核酸序列)不同的特定核酸序列的基因组。具有独特基因组的不同生物可以是不相关或密切相关的。认为纯系亲缘体(clonal relatives)具有相同的独特基因组，所述纯系亲属如在一个细菌菌落内在遗传上同源的生物，

“样品”是指由其制备核酸并测试特定核酸序列的存在的材料集合体。例如，样品可以是粪便样品、尿样品、血液样品或痰样品，或者可以是其它这样的在医院内常规收集的样品。或者，样品可以是在培养皿中培养的微生物单个菌落。样品也可以是人类法医学样品、食品样品、环境样品或纯核酸。

“扩增方法学”或“扩增方法”是指用于线性或指数增加核酸分子拷贝数的技术。扩增方法的例子包括连接酶链式反应、PCR、依赖于连接的 PCR、转录介导的扩增、链置换扩增、自身支持性序列扩增、

Qβ-复制酶介导的扩增、滚环扩增等等。

“扩增产物”是指应用扩增方法得到的核酸分子。

“扩增位点”或“扩增序列”是指在一种扩增方法中，介导复制或复制需要的核酸分子区。扩增位点对的例子是在 PCR 反应的特异性引发过程中寡核苷酸引物所结合的 DNA 片段或染色体上的位点对。在某些扩增方法中使用的针对 RNA 聚合酶如 Qβ-复制酶或噬菌体 T7 聚合酶的启动子序列，构成另一种类型的扩增位点。

“基因组扣除”是指导致分离基因组差异序列的方法。例如这样的杂交方法：其中“+” DNA 基因组差异样品(见下文)与“-” DNA 基因组差异样品退火，随后分离出剩余的非退火“+”序列。另外一个例子是使用计算机比较两个序列组，找到在第一个序列组存在但在第二个序列组不存在的序列。假如所述“+”样品中的一段序列(30 个碱基长)不能在扣除杂交条件下与所述“-”样品杂交，那么就认为该段序列在所述“-”样品中不存在。也就是说，在扣除杂交条件下，该序列不能与所述“-”样品中的序列形成解链温度(Tm)比所述扣除杂交条件的温度减 5℃要高的的杂交体。可以根据试验确定杂交，或者可以根据已知序列预测杂交。

“基因组差异样品对”是指用于发现基因组差异序列、对应于基因组 DNA 或 RNA 的两组核酸序列。例如，在基因组扣除实验中，“+” DNA 样品和“-” DNA 样品是基因组差异样品。当通过计算机分析比较两个基因组时，每个基因组就是一个基因组差异样品。基因组差异样品可以来自于一种生物或来自于一个生物类群；基因组差异样品可以包含已扩增或未扩增的核酸，例如聚合酶链式反应(PCR)扩增的 DNA；基因组差异样品可以由已经分级分离的核酸，例如大小级分或扩增级分组成；基因组差异样品可以是推导的核酸序列，如来自完全测序或几乎完全测序的基因组的序列的计算机代表；而且基因组差异样品可以由 RNA、DNA 或任何其它密切相关的核酸分子组成。只有在所述“+”样品中的许多但不是所有序列也在所述“-”样品中存在

时，基因组差异样品才有意义。

“富集的基因组”、“富集的基因组级分”、“富集的基因组差异样品”或“基因组代表”是指经过一个富集程序的基因组、基因组级分或基因组差异样品，所述富集程序产生原始基因组或基因组差异样品的选定部分。为基因组分布分析的目的，富集的基因组具有两个重要特性：(1) 它们提供健全的基于杂交的诊断学(与通过杂交检测SNP 的方法相比)，以及(2) 通过扩增产生的富集的基因组级分是从小样品(例如法医样品)产生材料的有效途径。例如，可以通过基因组分布分析，测试通过 Alu-PCR 产生的在富集的基因组中位于 Alu 重复序列之间的大量多态性序列(见实施例 4)，从而鉴定法医头发样品的来源。所述基因组富集可以基于大小分级分离、差异扩增(如 Alu-PCR 或 SNP 探针的差异扩增)、或任何其它分级分离方法。

表 2. 基因组代表的例子和它们用于检测序列的用途

基因组代表	代表类别	检测序列的类型的例子
限制性消化基因组 DNA 的经扩增的大小级分	限制性片段的物理特性(大小)	限制性片段长度多态性(RFLP)，即在一个菌株的一个大小级分中存在、但在另一个菌株同一的大小级分中不存在的序列
在重复序列之间的序列的扩增	依赖于重复序列排列的扩增的差异扩增	alu-形态(alu-morphs)(由于多态性，可以从一条染色体上扩增，但不能从一条同源染色体上扩增的处于 alu 重复序列间的序列
用 SNP 探针集合扩增	SNP 的扩增家族(即代表一个个体的基因型的 SNP)	在扩增的 SNP 上的标记
扩增与样品杂交的 ID 探针	扩增的 ID 探针家族	ID 序列集合

附图简述

图 1 是最小基因组起源为 9 的 ID 序列集合的示意性说明。

图 2A 是一个系统树的示意性说明，展示了一个假设的、但典型的菌株类群的祖先关系，其中所述菌株类群包括致病菌株(如菌株 1)和非致病菌株(如菌株 8)。

图 2B 是本发明一种方法的示意性说明，其中使用一个相关菌株类群内两种生物(如菌株 1 和菌株 8)的基因组扣除，产生可以用于该类群内任何菌株(如菌株 2-7)的指纹分析的基因组差异序列。

图 2C 是本发明的一种方法的示意性说明，其中通过从几种生物汇集基因组核酸分子而产生基因组差异序列。例如，汇集几种病原体

的基因组核酸分子可以产生“+”样品，汇集几种非病原体的基因组核酸分子可以产生“-”样品。通过该扣除实验获得的基因组差异序列包含至少在一种致病(“+”)菌株中出现但不在任何一种非致病(“-”)菌株中出现的序列。

图 3 是一种能够用于本发明方法的二元 ID 探针的示意性说明。在与染色体 ID 序列杂交后，将左半边 ID 探针和右半边 ID 探针相互连接。然后使用对应于引物位点 L 和引物位点 R 的引物扩增所述连接产物。随后通过与包含所述 ID 探针或标记序列的检测阵列的杂交，鉴定所扩增的 ID 探针产物。

图 4 是不同类型检测阵列的例子的示意性说明。

图 5 是本发明一种方法的示意性说明，在所述方法中，使用样品对 ID 探针的选择，通过基因组分布分析扫描临床样品中的多种病原体。在该方法中，将来自样品的 DNA 淀积在固相支持体如尼龙滤膜上。随后使多对半边探针与结合的样品 DNA 杂交，然后连接正确杂交的探针，将探针从所述滤膜上洗脱下来，扩增以在检测阵列中进行检测。

图 6 是用于从肠沙门氏菌获得基因组差异序列的基因组扣除策略的示意性说明。在该策略中，将肠沙门氏菌的亚种分为两个亚群，即 X 群和 Y 群。进行交互扣除，获得每一群的基因组差异序列。

图 7A 是大肠埃希氏菌类群的部分系统树的示意性说明。病原体标为黑色，非病原体标为白色。

图 7B 是用于获得大肠埃希氏菌 O157:H7 的基因组差异序列的策略的示意性说明，其中在大肠埃希氏菌 O157:H7 (“+”基因组差异样品)和非致病菌株(“-”基因组差异样品)之间进行基因组扣除。

图 7C 是用于获得弗氏志贺氏菌(*Shigella flexneri*)的基因组差异序列的策略的示意性说明，其中在弗氏志贺氏菌(“+”基因组差异样品)和非致病菌株(“-”基因组差异样品)之间进行基因组扣除。

图 8A 是用于滚环扩增的一种 ID 探针(包含一种带缺口的环状探

针和一种缺口探针)的示意性说明。

图 8B 是在对连接的滚环模板进行滚环扩增时使用的成对引物(一种生物素化滚环引物和一种生物素化分支引物)的示意性说明。

图 8C 是使用图 8B 举例说明的引物和连接的滚环模板进行高分支滚环扩增(hyperbranched rolling circle amplification)的示意性说明。

图 9A 是一对生物素化 DNA 捕获探针、一对扩增探针以及一种缺口探针的示意性说明，如所指出的，所述每种探针都与一种 ID 序列杂交。

图 9B 是使用一对生物素化引物扩增三联连接的探针的示意性说明。

图 9C 是在一种缺口探针序列和一种用于质谱检测的寡核苷酸之间杂交的示意性说明。

图 10 是 SNP 探针杂交选择的示意性说明，其中连接和扩增依赖于在 SNP 位点的匹配。

图 11 是本发明三类一般性基因组分布分析方法的共有特征的示意性说明。

发明详述

基因组分布分析是用于鉴定生物和对生物分型的方法，与现有技术相比，该方法提供几个显著的好处。在医学诊断学中，该方法可以在临床诊断设置中实施，提供了治疗的好处和流行病学的好处。可以同时、快速并且灵敏地扫描复杂生物样品中大量病原体特异性序列的存在。基因组分布分析产生高分辨率的遗传指纹，使得能够用该方法区分非常相似的菌株。这对于在病原体和密切相关的非病原体之间进行区分、在涉及疾病分别爆发(separate outbreak)的相似病原体之间进行区分、在相同病原体的抗生素敏感菌株和抗生素抗性菌株之间进行区分是重要的。对于扫描患者体内多种遗传标记的应用和在遗传鉴定中的应用而言，本发明扫描许多诊断序列的能力是重要的。

基因组分布分析使得能够进行一种新型的表现特异性测试，检测患者样品中全面的致病病原体组。例如，基因组分布分析使得可以为患有呼吸系统症状(respiratory symptom)的个体提供快速扫描所有常见呼吸系统病原体存在的单一测试，所述常见呼吸系统病原体包括不同病原体如细菌、病毒和真菌。

目前用于对生物分型的方法常常涉及培养所述生物，这需要使所述生物生长的时间，需要不同的培养条件，并且在医院设置中对于许多生物(包括一些细菌、大多数病毒和真核寄生虫)可能是不可行的。由于所述新方法不需要培养，该方法使得能够在几小时内获得结果(而不是目前方法所需的几天和有时几周)。

基因组分布分析的其它好处有：该方法需要最少的临床样品处理、产生以前未鉴定生物的指纹、简单地实现阳性内部对照和阴性内部对照、不需要凝胶电泳以及该方法适用于自动化的形式。

基因组分布分析将高度平行、基于杂交的筛选与灵敏的核酸扩增方法结合起来，使得能够在一次测定中鉴定广泛范围的生物类型。一次测试可以扫描生物样品中一类有用的 DNA 序列多态性，即 ID 序列的存在。ID 序列是特定类群内生物基因组所特有的核酸序列。一次测试也可以同时扫描多种单核苷酸多态性(SNP)，即另外一种类型的基因组变异。此外，基因组分布分析可以在一次测试中检测 ID 序列和 SNP 的混合物。

两类 ID 序列可以用于鉴定生物：类群特异性序列和基因组差异序列。在相关生物类群的所有成员中存在的 ID 序列称为类群特异性序列。类群特异性序列可用于确定某个类群的成员是否存在于生物样品中。例如，HIV 类群特异性序列的存在指出 HIV 类群中一种病毒的存在。可以通过基因组数据库的计算机比较，或通过用于分离保守序列的分子方法如符合克隆(coincidence cloning)，可以分离类群特异性序列。

仅在一个相关生物类群的某些成员中存在的 ID 序列称为基因组

差异序列。基因组差异序列组对于获得生物的高分辨率指纹尤其有用。因此，这种类型的 ID 序列有利于将一个类群中的一个成员与该类群内的另一个成员区分开来。对生物进行指纹分析对于流行病学、法医学、以及快速确定细菌是否可能对某种抗生素有抗性是重要的。基因组差异序列可以如下制备：例如，用两种不同生物的基因组进行扣除杂交程序，或对两组不同生物的汇集的基因组进行扣除杂交(见下文)。

基因组分布分析扫描复杂生物样品中的 ID 序列，ID 序列是 DNA 片段，其存在是特定类型生物的指示。两种类型 ID 序列可以用于确定一种生物的存在。类群特异性序列是特定分类群中(即在成员通过谱系密切相关的生物类群中)基本所有生物都共有的。与此不同，基因组差异序列将特定分类群内的生物区分开来。一个基因组差异序列家族有用的诊断属性在于：在一个类群中的密切相关菌株的基因组中存在该家族成员的独特亚组。

基因组分布分析的诊断能力部分是由于它能够测试 ID 序列的复杂混合物，所述 ID 序列是庞大并且不同组生物类型所特征性拥有的。因此，扩展这样的诊断 ID 序列组的早前提出的定义是有用的。

ID 序列“家族”是可用于鉴定特定生物类群的成员的一组类群特异性序列和/或基因组差异序列。在一个家族内 ID 序列组的定义特性在于所有的成员都能够与一个“独特基因组”杂交(见表 1 和上文的定义)。例如，一个 ID 序列家族可以由 100 个 ID 序列组成，其中包括 80 个鉴别大肠埃希氏菌 O157:H7 病原体类群的菌株(来源于菌株 DEC3B 的菌株除外)的基因组差异序列、18 个存在于所有大肠埃希氏菌 O157:H7 菌株中的类群特异性序列、以及 2 个存在于大肠埃希氏菌所有菌株中的类群特异性序列。注意：虽然这些序列可用于专门鉴定大肠埃希氏菌 O157:H7 类群中的病原体，但所有这些序列都可以与一个独特基因组，即大肠埃希氏菌 O157:H7 DEC3B 菌株的基因组杂交。

基因组分布分析的独特特征是：该方法可以用于一次扫描样品中

许多不同家族的存在。由多于一个家族组成的一组ID序列称为一个ID序列“集合”。一个集合中家族的数量反映出该集合可以测试的不同生物类群的数量。一个集合内的家族数量又可以用称为集合“最小基因组起源”的数量来准确定义。“最小基因组起源”是组成该集合的所有序列可以杂交的“独特基因组”的最小数量。例如，基因组分布分析可以用最小基因组起源为5的一个集合同时测试痰样品中结核分枝杆菌、军团菌、*Coccidoides immitus*、流感病毒和呼吸道合胞病毒的存在。因此，基因组分布分析在一次测试中鉴定广泛范围生物的能力是该方法扫描样品中具有大“最小基因组起源”集合的ID序列存在的能力的结果。

相似的，在非传染病的应用例如人类遗传筛选和法医学中，可以使用基因组分布分析扫描样品中单核苷酸多态性的集合。与ID序列集合的定义相似，SNP集合定义为一组多家族SNP。一个SNP家族就如一个ID序列家族一样，反映一个个体的基因型。注意：ID序列家族根据成员ID序列与单个个体的基因组杂交的能力来定义，而SNP家族则是根据与单个生物的基因型的对应来定义。

应用基因组分布分析进行基因型分析(genotyping)的一个好处是可以使用健全的杂交测定来检测SNP。在一些大规模SNP基因型分析应用中，检测区分形成完全配对双链体(perfect duplex)的寡核苷酸杂交体和形成带有单碱基对错配的双链体的寡核苷酸杂交体的SNP基因型。与此不同，基因组分布分析可以测试寡核苷酸标记序列的存在或不存在，这是一个更容易的工作。为完成这种更健全的杂交测试，可以将独特的非生物学标记序列掺入每种SNP探针。因此，这样的SNP探针集合与标记序列集合相应，并且每个SNP家族与一个标记序列家族相应。在基因组分布分析测定的检测步骤中，可以使用由一个标记序列集合构成的一个检测集合，检测一个对应于从单个个体分离的基因组DNA样品的基因型的扩增的SNP探针家族(包括对应的标记序列家族)(见图3)。

优选的基因组分布分析方法通用设置包括以下步骤:

步骤 1: 指定一个包括基因组差异序列和类群特异性序列的 ID 序列集合，其中将在给定测试中探测所述集合。该步骤涉及选择需要检测的生物和选择诊断 ID 序列的家族。

步骤 2: 设计和制备一个对应于要在生物样品中检测的 ID 序列集合的探针集合。同时设计和制备对照探针。

步骤 3: 设计和制备一个对应于所述 ID 探针集合的检测集合。同时设计和制备对应于对照探针的对照序列。在一个优选的实施方案中，制备两维的检测阵列。

步骤 4: 制备生物样品。该步骤涉及裂解样品中的生物，以便所述生物的核酸分子能够进行杂交。例如，处理样品如大便样品或呼吸系统样品，以便来自所述样品中的生物的核酸分子结合到固相支持体上。

步骤 5: 从所述 ID 探针组合中选出与所制备样品中的基因组序列杂交(结合)的 ID 探针。然后通过洗涤除去未杂交、未结合的探针。

步骤 6: 扩增与所述样品中的基因组序列结合的 ID 探针。

步骤 7: 通过所扩增的探针序列与检测集合的杂交，鉴定样品选定的 ID 探针。

步骤 8: 通过所述样品选定的 ID 探针与所述生物样品的原位杂交，定量所述生物样品中的靶生物。

(注意: 为了简单化，优选通用设置的步骤根据使用 ID 序列的基因组分布分析描述。对于用于使用 SNP 的基因组分布分析的该方法的修改，见实施例 5)

这些步骤的每一个步骤如下更详细描述。

步骤 1: 指定一个包括基因组差异序列和类群特异性序列的 ID 序列集合，其中将在给定测试中探测所述集合。该步骤涉及选择需要检测的生物和选择诊断 ID 序列的家族。

基因组分布分析的第一个步骤涉及选择需要检测的生物类型。例如，对于医学应用，可以选择人类病原体；为检测食物腐败，可以选择导致食物毒性的细菌；为法医学目的，可以选择多个人类个体等等。为特定测试选择的生物可以在它们的遗传组成上相差极大，例如不同界的成员(即病毒、细菌、古细菌、真菌、原生动物、植物和动物)；或者，所选择的生物可以是一个更小的类群如一个种的成员。基因组分布分析的一个重要的应用是检测人类体液样品中或大便中的病原体，所述人类体液样品如血液、尿、脑脊液或痰。(本方法对于应用于多种其它组织样品也是重要的。) 根据组织样品的来源以及患者的症状，决定需要鉴定的重要生物类型。例如，可以选择检测通常是肺炎的病因的病毒、细菌和真核寄生虫。

一旦决定了需要通过基因组分布分析测定鉴定的生物类型，就为该测定选择一个 ID 序列集合。由多个 ID 序列家族组装所述集合，其中每个 ID 序列家族都是在所述测定中需要检测的一种生物类型的诊断性序列。所述 ID 序列集合不一定在物理上是分离的。当然，可以仅仅将这样一个集合概念化，以利于设计用于构建探针集合的ID探针(见下文)。

如上文所述，所述 ID 序列集合包括两种有用的序列类型：基因组差异序列和类群特异性序列。对于任何特定靶生物类型来说，是否包括类群特异性序列、基因组差异序列或二者都包括的选择取决于与所述特定生物类型相关的诊断用组织。

当重要的是需要知道一个生物类群的任一成员是否存在于样品中时，类群特异性序列在诊断上是最有用的。例如，如果重要的是需要知道肠沙门氏菌类群的任一成员是否存在于胃肠样品中，则类群特异性样品是有帮助的。当测试病毒如丙型肝炎病毒时，也可能选择类群特异性样品。

与类群特异性样品不同，当需要在一个类群内区分密切相关的菌株时，基因组差异序列尤其有用。例如，当重要的病原体(如大肠埃希

氏菌 O157:H7)与出现在同一组织中的菌株(如共生的大肠埃希氏菌)密切相关时，就是这种情况。当需要传染因子的指纹时，基因组差异序列也是有价值的。指纹分析或高分辨率菌株鉴定是追踪和遏制传染病爆发(包括基于医院的感染)的有力流行病学工具。在治疗上，指纹分析，尤其是在快速、不依赖于培养的测试中的指纹分析，提供了比目前实践快得多地确定需要给予何种抗生素的可能救命的机会。

对于需要在基因组分布分析测定中检测的每一种生物类型，使用标准方法选择包含类群特异性序列和/或基因组差异序列的 ID 序列家族，所述标准方法如在下文和在实施例中描述的那些方法。假如新分离出的 ID 序列的序列还是未知的，就通过标准方法测定该序列。然后将对应于不同并且可能不相关的生物类型的各种 ID 序列家族组织成为一个集合。

然后使用商业化可得的寡核苷酸合成方法或服务，通过从质粒合成重组 DNA，或通过用于产生足量纯 DNA 分子的任何其它方法，设计并合成对应于所选择的 ID 序列的探针集合。给定 ID 序列的探针可以包括一个、两个或几个寡核苷酸以及用于检测的附加部分。至少所述探针的一部分即 ID 位点设计用于与来自测试生物的 ID 序列核酸分子杂交。

使用基因组扣除分离基因组差异序列。基因组差异序列用于将一个菌株与一个密切相关的菌株区分开来。基因组差异序列家族具有这样的特性：该家族中不同序列亚组存在于不同菌株中。基因组分布分析可以确定在临床样品中出现的基因组差异序列家族的亚组。这样就准确鉴定了在样品中存在的一个菌株。基因组分布分析优于现有测定的一个好处是：可以同时调查*许多不同*家族，其中每个家族都能够对一个特定生物类群进行指纹分析。

可以通过对致病菌株和相关的非致病菌株进行基因组扣除而分离可用于临床诊断的基因组差异序列。一些基因组差异序列具有重大的临床重要性。例如，近年来逐渐了解致病细菌常常带有“致病性岛

(pathogenicity island)”，即包含致病性所需的多个毒性基因的连续 DNA 序列段。密切相关的非致病菌株一般缺乏致病性岛。因此，致病性岛是有用的基因组差异序列。其它(可能大多数)基因组差异序列没有临床重要性，但对于菌株鉴定仍然是非常有价值的。值得注意的是：在类群特异性序列和基因组差异序列间的区别有时是不清楚的。例如，可以将大肠埃希氏菌 O157:H7 致病性岛看作基因组差异序列，因为它出现在大肠埃希氏菌的一些菌株中，但不出现在其它菌株中。或者，同一序列可以看作是类群特异性序列，因为它出现在由大肠埃希氏菌 O157:H7 菌株组成的分类单位的所有成员中。不考虑有时出现的不明确性，这些序列是有用的诊断 ID 序列。

可以通过使用几种基因组扣除方法中的一种，分离基因组差异序列家族(如 Straus, 1995, 见上文; Diatchenko 等, Proc. Natl. Acad. Sci. U.S.A. 93: 6025-6030, 1996; Tinsley 等, Proc. Natl. Acad. Sci. U.S.A. 93: 11109-11114, 1996)。基因组扣除分离在一个菌株(“+” 菌株)的基因组中出现，但不在相关菌株(“-” 菌株)的基因组中出现的 DNA 序列。基因组扣除的产物是基因组差异序列家族：整个组与所述 “+” 菌株杂交，没有一个序列与所述 “-” 菌株杂交，并且独特的亚组与密切相关的菌株杂交。基因组差异序列家族的一个普遍特性是：在与用于制造所述基因组差异样品的菌株(即用于基因组扣除的菌株)密切相关的菌株的基因组中，所述成员以不同组合出现。该基因组差异序列家族的存在于个别菌株内的独特亚组构成了高分辨率指纹。但是，注意：来自基因组扣除的整个基因组差异序列家族可以与一个菌株杂交，即用于制造 “+” 基因组扣除样品的菌株。(在使用多于一个菌株制造所述 “+” 基因组差异样品的情况下，扣除的产物可以构成多于一个家族。)

基因组扣除一般使用扣除杂交和亲和层析，从 “+” 和 “-” 基因组差异样品中纯化基因组差异序列(Straus, 1995, 见上文)。首先制备来自两个相关菌株(“+” 菌株和 “-” 菌株)的基因组 DNA。用限制酶切割来自所述 “+” 菌株的 DNA，随机剪切来自所述 “-” 菌株的 DNA

并用生物素修饰，生物素是亲和性标记，允许通过与其配体抗生物素蛋白的结合而随后除去所述“-”菌株 DNA。通过使来自所述“+”菌株和所述“-”菌株的变性 DNA 片段复性，完成对基因组差异样品的富集。复性后，通过与抗生物蛋白包被的珠粒的结合，取出生物素化序列以及所有已经与所述生物素化序列杂交的序列。然后重复该扣除过程几次。在每一个循环中，来自前一轮扣除的来自所述“+”菌株的未结合 DNA 与新鲜的来自所述“-”菌株的生物素化 DNA 杂交。将来自最后一个循环的来自所述“+”菌株的未结合 DNA 连接到连接物上，并在聚合酶链式反应中通过使用所述连接物的一条链作为引物进行扩增。然后可以克隆所扩增的序列。注意：进行交互扣除(即转换“+”菌株和“-”菌株)产生一组不同的基因组差异序列。这样的可以用于产生基因组差异序列的扣除方法是重组 DNA 技术领域内一般技术人员已知的，并且这样的方法已经广泛发表。在下面的实施例中提供其它细节。

基因组扣除的全面评述在图 2 中图解说明。图 2A 显示了具有共同祖先的一个生物类群(“分类单位”)的假设的系统树。其中一些生物是病原体，而另一些是非病原体。图 2B 图解说明了用于分离基因组差异序列的一种策略。可以选择一个相关菌株类群中的两种生物(如菌株 1 和菌株 8)制备基因组差异序列。病原体菌株 1 用于制备“+”基因组差异样品，而非病原体菌株 8 用于制造“-”基因组差异样品。所述扣除(图 2B)的产物是出现在菌株 1 中、但不出现于菌株 8 中的基因组差异序列。这些基因组差异序列可以用于对该类群内的任何菌株(即包括菌株 2-7)进行指纹分析。使用菌株 1 和菌株 8 的基因组扣除(图 2A)可以从菌株 1 产生数百种不出现于菌株 8 中的序列。菌株 2 具有这些基因组差异序列中的一些，但缺乏其它的基因组差异序列。菌株 5 携带有所述基因组差异序列的一个独特亚组，菌株 7 也一样，依此类推。重要并且普遍性的发现是：当应用基因组扣除于一个类群内的两个菌株(图 2 中的菌株 1 和菌株 8 以及本文描述的实施例)时，相关菌株(如

菌株 2 和菌株 5)携带有所得基因组扣除产物的不同亚组。

如图 2C 所举例说明的，也可以通过从几种生物汇集基因组核酸而产生基因组差异序列。例如，可以通过汇集几种病原体而产生“+”样品，可以通过汇集几种非病原体而产生“-”样品(图 2C)。在这种情况下，通过基因组扣除分离的基因组差异序列是在所述“+”基因组差异样品的至少一种病原体基因组中出现、但不在任何一种所述“-”基因组差异样品的非病原体基因组中出现的序列。

不用扣除杂交，而可以使用计算机和序列比较软件比较两种生物或两组生物的基因组，并因此产生基因组差异序列。例如，当靶生物基因组的序列完成或基本完成时，该方法是实用的。例如，已经报道了其序列最近已经完成的幽门螺杆菌(*Helicobacter pylori*)的相关菌株的基于计算机的比较(Alm 等, Nature 397: 176-180, 1999)。已经公开的分析和公众可获得的数据提供了对于一种或另一种菌株独特的多种基因组差异序列。则这种分析构成了一种类型的“虚拟(virtual)”基因组扣除分析，由所述分析确定了基因组差异序列。

分离类群特异性序列。当重要的是仅仅确定某个类群的任一成员是否存在于生物样品中时(与确定来自某个类群的哪种个别菌株不同)，在通过基因组分布分析测定评估的 ID 序列集合中包括类群特异性序列。可以用多种方法分离类群特异性序列，包括通过基因组扣除和通过分析公共数据库。例如，基因组扣除使用来自作为“+”基因组差异样品的致病性结核分枝杆菌菌株的 DNA，以及来自作为“-”菌株的非致病性分枝杆菌菌株的 DNA，所述基因组扣除产生类群特异性序列，其中包括在所有致病性肺炎分枝杆菌菌株中共有的毒性基因。这些类群特异性序列对于测试引起肺结核的菌株的存在的是价值的 ID 序列。作为另一个例子，可以通过在公共数据库如 GenBank 中扫描病毒基因组 DNA 序列，筛选在所有单纯疱疹病毒的已知分离物中出现、但不在该数据库其它类型病毒中出现的序列，从而分离针对单纯疱疹病毒的类群特异性序列。

步骤 2: 设计和制备对应于要在生物样品中检测的 ID 序列集合的 ID 探针集合。同时设计和制备对照探针。

在基因组分布分析的第二个步骤中，设计 ID 探针集合，以便该集合中的 ID 探针可以与步骤 1 中选定用于基因组分布分析的 ID 序列集合的成员杂交。一个 ID 探针可以包括单个寡核苷酸，或者在优选的实施方案中，ID 探针可以包括两个或更多个寡核苷酸。ID 探针和任何其组成寡核苷酸可以包含一个或多个功能部分。

一种 ID 探针的一个部分即 ID 位点对应于一种 ID 序列。在本方法优选的实施方案中，ID 探针集合包含多功能 ID 探针，其中探针序列的第一个部分对应于在步骤 1 中组装的 ID 序列集合中的一个序列。因此，如下文所述，一个这样的 ID 探针包括对应于一个 ID 序列的一部分的一个序列或一组序列，并且所述 ID 探针可以与包括所述 ID 序列在内的核酸分子杂交。该部分称为 ID 位点。例如，这样一种 ID 探针可以包含对应于一种基因组差异序列或一种类群特异性序列的 ID 位点。

对应于扩增序列的 ID 探针的部分。基因组分布分析的一个重大好处是它能够一次完成许多序列的健全的无假象扩增的能力。通过使用非常少量的扩增序列指导大量独特 ID 探针的扩增，基因组分布分析测定避免了在多重扩增中通常出现的扩增假象。为此，所述 ID 探针的第二个部分(除所述第一个部分外，对应于一种 ID 序列)可以包括一个或多个扩增序列。例如，该第二部分可以对应于一个或多个引物结合位点，或对应于核酸聚合酶如 Qβ复制酶的结合位点。所述扩增部分是该集合内(包括对照序列)要扩增的大多数或所有探针所共有的。因此，可以在同一反应中有效扩增的包括 ID 探针和对照序列的集合(见下文)的探针组。所述探针可选的第三个部分可以包括用于检测所扩增探针的标记序列。标记的使用在下面的步骤 3 中讨论。

对照序列。在 ID 探针集合中可以包括阳性对照和阴性对照。在

所述集合中可以包括并不对应于实际基因组中的序列、而对应于在样品制备过程中加入所述样品中的对照核酸分子的阳性对照序列。在基因组分布分析测定中，检测到阳性对照序列指示整个测定工作正确。(当在样品中没有检测到 ID 序列时，重要的是知道所述样品中是否确实不存在 ID 序列，或者是否测试由于某种原因失败。)

在所述 ID 序列探针集合中也可以包括阴性对照序列。这些阴性对照序列并不对应于天然出现的序列，并且与阳性对照序列不同，这些阴性对照序列并不加入所述生物样品中。通过基因组分布分析测定检测到的阴性对照序列的水平指示出在所述测定中，由于不依赖于 ID 序列的选择和 ID 探针的扩增而产生的背景水平。

二元探针(半边探针)。在一个实施方案中，一个 ID 探针由一对寡核苷酸组成，即左半边 ID 探针和右半边 ID 探针(图 3)。每个左半边探针和右半边探针的内部部分包括对应于一种 ID 序列的邻近部分的序列，所述 ID 序列如基因组差异序列或类群特异性序列。当所述半边探针与变性 ID 序列杂交时，可以通过核酸连接酶连接各探针部分。如下文所描述的，半边探针的依赖于样品的连接导致形成可以扩增和检测的更大分子。

在本实施方案中，每个半边探针的外部部分包括一个扩增序列，所述扩增序列例如对应于用于聚合酶链式反应的引物结合位点的位点。在这样的 ID 探针集合中，每个探针具有一个独特 ID 序列和标记序列，但具有一对共有的引物结合位点。假如存在标记序列，则该标记序列位于其中一个半边探针的内部部分和外部部分之间。

图 3 图解说明了一个实施方案，该实施方案使用了半边探针、依赖于 ID 序列的连接、标记、以及 PCR 扩增与样品杂交的半边探针。在这个实施例中，PCR 的左引物与引物位点-L 序列相同，而右引物是引物位点-R 序列的反向互补物。在该检测阵列中可以包括四种不同的标记序列(tag-R、tag-R'、tag-L 和 tag-L')(见下文)。所述四种标记序列与两种互补序列杂交，所述互补序列每一个都包含在所扩增的 ID 探针

中的两种标记序列。

ID 探针的合成和浓缩。通过标准核酸合成技术制备 ID 探针。确定所述 ID 探针的序列和所述 ID 探针在水溶液中的浓度。根据需要，所述 ID 探针在水溶液中的浓度可以不同。例如，在一个 ID 探针集合中，每种寡核苷酸可以以等摩尔量存在。在一个可替代的实施方案中，ID 探针存在的量与包含所述对应生物的典型生物样品中其对应 ID 序列的预期丰度负相关。例如，假如一个人同时受到轮状病毒和寄生性线虫的胃肠感染，则在大便样品中的轮状病毒基因组拷贝数可能比所述大便样品中的线虫基因组拷贝数更多。因此，使针对轮状病毒序列的探针以有限量存在是有用的。

步骤 3: 设计和制备对应于所述 ID 探针集合的检测集合。同时设计和制备对应于对照探针的对照序列。在一个优选的实施方案中，制备二维检测阵列。

检测集合的作用是检测和鉴定通过与生物样品中的 ID 序列杂交而选定的 ID 探针集合亚组。所述检测集合包含对应于在步骤 2 中组装的 ID 探针集合的序列(以及对应于对于该测试中不同类型生物的存在是诊断性的 ID 序列的序列)。换句话说，所述检测集合与所述 ID 探针集合相应。所述检测集合中也包括对应于所述对照探针的对照序列。

所述检测集合由可以用于检测探针-样品杂交事件的核酸分子组成。所述检测集合可以包括对应于 ID 序列或所述探针内序列标记的序列。在基因组分布分析方法的一个实施方案中，使所述检测集合的 DNA 序列变性并固定到固相支持体上，以便所述检测集合的 DNA 序列可以与所加入的 ID 探针杂交。当在平面固相支持体上构建所述检测集合时，该检测集合称为二维检测阵列。将所述检测序列 DNA 置于所述支持物上的不同位置。将 DNA 分子以这种方式固定到固相支持体上的方法是基因组学领域内的技术人员已知的。例如，在实施例中提到的方法可以用于该目的。或者，可以在液相中进行所述样品选定的 ID

探针与所述检测阵列的杂交，如在下面的实施例 3 所述。

在阵列设计的一个优选实施方案中，对应于一个类群或相关类群的检测序列在所述阵列上相互相临排列。这样，检测序列家族，即那些对给定类型生物(例如，在大肠埃希氏菌 O157:H7 类群中的病原体)特异性的检测序列家族就作为一组相邻点放置在一起。此外，将对应于密切相关家族(例如大肠埃希氏菌 O157:H7 和志贺氏菌属)的检测序列家族放置在所述阵列的同一区。这种组织方便了杂交结果的读取。

所述 ID 探针集合所包括的阳性对照序列和阴性对照序列(见上文)也可以掺入所述检测集合中。如上文所讨论的，也将所述阳性对照序列与所述生物样品混合，并用于指示所述测定的正确运行。所述阳性对照探针序列与所述生物样品中的靶对照序列杂交，扩增所述阳性对照探针序列，然后使所述阳性对照探针序列与所述检测阵列中的对应对照序列杂交。

阴性对照序列是所述测定中不依赖于病原体的背景信号的有用量度(即，尽管在所述生物样品中不存在对应病原体，但仍被扩增的 ID 探针的量的量度)。与阳性对照序列不同，阴性对照序列并不与所述生物样品混合。这样，阴性对照序列在所述生物样品中没有要杂交的靶序列。所述阴性对照序列与所述生物样品或样品基质的非特异性结合，使得这些序列随后被扩增并与所述检测阵列中的对应序列杂交。

构建包含一个检测序列集合的阵列。可以使用各种类型的检测阵列来检测诊断序列。图 4 图解说明了用于下文描述的实施例的检测阵列的一些设计。

已经描述了多种用于构建核酸分子阵列的方法。用于本发明的一种优选方法是这样一种方法：其中核酸分子以高密度放置在聚赖氨酸处理过的玻璃载玻片上(见，如，Schena 等，Science 270: 467-470, 1995)。对应于 ID 序列的检测序列可以作为克隆 DNA (如作为质粒载体中的插入片段)、作为扩增的 DNA (如由克隆序列的扩增得到的 PCR 产物)或作为合成寡核苷酸放置在所述阵列中。

或者，所述检测集合可以包括一组可寻址的合成寡核苷酸标记，而不是 ID 序列。在这种情况下，所述标记对应于所述 ID 探针(如下文所述)或 SNP 探针(如在实施例 5 中所述)中的标记元件。所述阵列中的每种可寻址标记对应于在接受杂交选择的探针集合中与特定探针序列结合的标记(见下文)。在阵列元件和探针集合之间的一一对应关系使得有可能通过观察哪些寡核苷酸标记阵列元件与混合物中的分子杂交，鉴定所述混合物中的所述 ID 序列。该方法的好处是可以使用预制的阵列，因为包含同一组可寻址标记的阵列可以用于不同组探针。例如，用于检测呼吸系统病原体的一组探针和用于检测胃肠病原体的一组探针可以使用同一组标记。这样，可以使用一种阵列检测呼吸系统样品或胃肠道样品中的病原体。

或者，所述检测阵列可以是在液体中与所述样品或探针杂交的检测序列组。检测阵列也可以是诊断产物所比较的一组物理特性，如分子量。

步骤 4: 制备生物样品。该步骤涉及裂解样品中的生物，以便所述生物的核酸分子可以用于杂交。例如，处理样品如大便样品或呼吸系统样品，使得来自所述样品中生物的核酸分子结合到固相支持体上。

通过下面的样品制备策略达到的目标是:

(a) 将来自广泛来源(如培养物、菌落、痰、血液、尿和粪便)的样品转化成为与所述测定的随后步骤相匹配的共有形式。裂解生物，并使它们的基因组核酸分子可以用于杂交。

(b) 浓缩所述样品，因此增加所述测定在测试稀形式的生物(如在尿样品或血液样品的情况下)时的灵敏度。

(c) 通过去除或固定化抑制性物质，除去或减弱所述样品中酶抑制剂的效应。

可以使用几种样品制备方法中的任何一种制备用于本方法中的

样品。样品制备的一般概念是使核酸分子释放和变性，以及除去可能干扰随后步骤的污染蛋白质和其它物质。可以可选地用样品制备方法选择性保留 DNA、RNA 或同时保留二者。

在制备前，可以通过标准过滤装置过滤，浓缩稀的样品类型如尿样品。假如样品来源包含大于目标生物的颗粒性物质，那么在执行样品浓缩步骤前，通过使所述样品过滤通过孔径大于目标生物的滤膜，从所述样品中除去所述颗粒。当测试微生物时，例如，通过用平均孔径为 20 到 30 微米的膜预过滤，将微生物与大颗粒分离开来。

或者，可以使用离心步骤将微生物与具有不同大小或密度的材料分离开来。例如，可以通过离心步骤，以导致大颗粒而不是微生物沉积在沉淀中的速度，将大颗粒物质与微生物分离开来。如在培养的微生物样品的情况下，可选地通过离心由液相分离微生物。使用过滤和离心的组合来浓缩和富集怀疑的测试生物。然后进一步制备从通过离心处理的样品回收的沉淀。过滤和离心都有潜在的缺点：病毒可能从样品中丢失。该步骤也可以包括其它富集方法，如亲和层析、细胞分选和基于抗原的富集。

在一个优选的实施方案中，将实验样品(通过过滤或离心获得的，以及有高含量微生物的粗制样品如粪便样品)淀积并固定到固相支持体上，所述固相支持体如尼龙滤膜、颗粒性基质或珠粒(图 5)。使用固相支持体提供了优于其它方法的几种好处。将样品 DNA 固定到固相支持体上并使其变性，准备与单链核酸分子探针杂交。通过固定和洗涤粗制 DNA 样品，酶促步骤(如连接和扩增)的抑制剂或者被固定到基质上，或者从包含结合 DNA 的滤膜上洗涤下来。这是一个重要的好处，因为对临床样品的 PCR 测试有时由于样品成分的抑制而缺乏灵敏度。最后，包括内部对照以检测假阴性结果是简单的。

优选的支持物是尼龙滤膜，尼龙滤膜耐用但柔韧，广泛用于固定包含核酸分子的样品以进行杂交测定(Church 等, Proc. Natl. Acad. Sci. USA 81: 1991-1995, 1984)。将粗制样品如痰样品或粪便样品涂抹到固

相支持体上，如同目前当使用“抗酸涂片”测定(Koneman 等, *Color Atlas and Textbook of Diagnostic Microbiology* (Lippincott-Raven, Philadelphia, 1997))来测试痰样品中的结核分枝杆菌时的实践一样。相似的，可以将在培养皿的半固体培养基上生长的细菌或真菌菌落“转移”到尼龙滤膜上，或从培养皿涂抹到滤膜上涂抹到固相支持体上。

在一个优选的实施方案中，随后使用破开样品中的细胞并变性任何双链 DNA 的程序，将样品固定到固相支持体上。已经发展了用于破开细胞的多种方法。这些方法包括机械破碎和用碱、离液剂、热以及有机溶剂处理。本发明的该步骤可以加入一个或多个这样的方法以破碎细胞。一种涉及碱处理以及随后的中和及洗涤的简单方法是将样品中的变性 DNA 固定到固相支持体上的优选方法(Hanahan 等, Methods Enzymol. 100: 333-42, 1983; Grunstein 等, Proc. Natl. Acad. Sci. USA 72: 3961-3965, 1975; Ausubel, 1987, 见上文)。

假如测定产生了阴性结果，重要的是知道所述样品是否确实不含有来自测试微生物的基因组 DNA，或者是否所述测试本身失败，即该结果是否是假阴性。由于实验样品中阻断所述测定中一个酶促步骤的抑制剂的存在，可能出现假阴性。

为鉴定假阴性结果，可以在所述实验样品中加入一个或多个阳性对照 DNA 样品。所述阳性对照 DNA 样品包含不在所测试的生物范围内出现的 DNA 序列。在所述探针集合中包括对应于所述阳性对照 DNA 样品的探针。这些探针将在所有的测定中被扩增和检测到，除非一个或多个测定步骤是不成功的。不能检测到来自阳性对照的信号将由此可以指示假阴性结果。

图 5 图解说明了样品制备、杂交-选择、扩增以及检测所选定的探针。在该实施方案中，通过将样品裂解到尼龙滤膜上而制备样品，以便使所述样品的核酸分子变性并结合到所述滤膜上。阳性对照 DNA 样品也结合到所述滤膜上。然后使可连接的半边探针与结合的核酸分子杂交。假如一种探针的两半都结合到一种 ID 序列上，则它们被连接

起来以产生全长的探针，因为在所述全长探针的每个末端存在引物结合位点，所以所述全长探针可以用 PCR 扩增。不正确结合的半边探针不能通过 PCR 扩增。

步骤 5: 从与所制备样品中的基因组序列杂交(结合)的 ID 探针集合中选择 ID 探针。通过洗涤除去未杂交、未结合的探针。

使所述探针集合与已固定的样品杂交的目的是：选择对应于所述已固定样品中的基因组 DNA、并因此能够用于鉴定所述基因组 DNA 的探针，以及将这些杂交探针与非杂交探针分离开来。各种靶生物的基因组 DNA 与所述 ID 探针的独特亚组杂交。因此，选定的 ID 探针特定亚组构成特定生物的基因组的指纹。所述 ID 探针杂交步骤设计是快速、特异性、并用来测试广泛范围的生物。包括阳性对照和阴性对照便利了确定所述杂交是否如所需要地起作用。

在该步骤中，使 ID 探针集合与变性的核酸样品杂交。如上所述，杂交可以在水溶液中完成，或者可以用固定化在固相支持体上的核酸分子完成。通过将探针集合与所制备的生物样品混合，并最好温育直到至少度过一个 $C_0t_{1/2}$ 时间，进行杂交。随后洗涤、稀释或用其它方法处理所述探针/样品混合物，以便从已杂交的探针和所述样品中分离出未杂交或非特异性杂交的探针分子。可以对已杂交的探针进行酶处理，如连接或核酸聚合。最后，如下一个步骤所述，从样品核酸分子中分离已杂交的探针并进行扩增。

在一个优选的实施方案中，将样品(包括阳性对照核酸分子)固定到固相支持体上(图 5)。使该样品与探针集合杂交，所述探针集合包括 ID 探针和阳性及阴性对照。所述探针由与 ID 序列的相邻部分杂交的成对寡核苷酸组成。洗涤已杂交的样品以除去未结合的探针，然后用核酸分子连接酶处理已杂交的样品，连接左半探针和右半探针。最后，从所述样品中取出连接的左半探针和右半探针，并进行扩增。下面是该优选实施方案的特定版本的描述。

i. 将所述 ID 探针杂交混合物置于所述实验样品上，所述实验样品固定在固相支持体如玻璃载玻片或尼龙滤膜上。所述优选的杂交混合物包括:

a) 一个 ID 探针集合，其中包括基因组差异序列和/或类群特异性序列探针。在这种情况下，所述 ID 探针是由两个可连接半边探针组成的成对寡核苷酸。在优选的体积 10-100 μl 中，每个半探针的优选浓度是 1-10 nM。在优选的复性条件下，该探针浓度导致几分钟内与所固定的样品的可接受水平的杂交(Britten 等，Meth. Enzym. XXIX: 363-418,1972)。

b) 一对或更多对阳性对照半边探针，其浓度与所述 ID 序列的浓度相当。这些探针的序列对应于固定到固相支持体上的阳性对照 DNA (固相支持体上面还结合了所述生物样品)。

c) 一对或更多对阴性对照半边探针，其浓度与所述 ID 序列的浓度相当。这些探针序列在已固定的 DNA 样品中没有对应物。

d) 1 M NaCl/10 mM EPPS/1 mM EDTA, pH8.0。用标准杂交溶液取代也是可接受的(Ausubel, 1987, 见上文; Church, 1984, 见上文)。

ii. 用玻璃盖玻片覆盖所述杂交混合物，最好用垫片(如 Cenegator™, 目录号 #009917, BioWorld Fine Research Chemicals)将所述玻璃盖玻片与所述样品分离开来。

iii. 在约 65℃温育 5-30 分钟。

iv. 洗掉未结合的探针。这可以通过除去盖玻片并在严格条件下洗涤所述固定的样品而完成，使得仅有无错配或仅少数错配而复性的 ID 探针保持与固定化的互补基因组 DNA 结合。所选择的条件依赖于几个因素，包括所述探针中 ID 序列的长度

以及错配可以接受的程度。

v. 连接退火的成对半边探针。使用 T4 DNA 连接酶(如来自 New England Biolabs)连接已经退火到所固定的实验样品中互补基因组 DNA 上的相邻半边探针。按照厂家的指示进行连接。

vi. 从所述实验样品取出已连接的半边探针。通过在变性条件下的短暂温育，从所述样品中洗脱已经退火到所固定的实验样品中互补基因组序列的探针。释放已结合的探针的优选方法是应用 10 mM EPPS/1 mM EDTA，盖上盖玻片并短暂加热到 100℃。

步骤 6：扩增结合到样品中基因组序列的 ID 探针。

该扩增步骤是基因组分布分析测定的高灵敏度的基础。(然而，并不是在所有的应用中都需要扩增。) 在取出(通过热变性或化学变性)任何已经与所述生物样品杂交的 ID 探针后，使用核酸聚合酶以及核酸分子前体扩增所述 ID 探针。可以使用在所述探针中存在的引物结合位点，用引物驱动扩增。或者，扩增可以是由特异性核酸聚合酶(如 Qβ 复制酶或 T7 RNA 聚合酶)结合到掺入所述探针的特异性结合位点而驱动的。可以使用几种扩增方法中的任何一种，包括连接酶链式反应、PCR、依赖于连接的 PCR、转录介导的扩增、链置换扩增、自身支持性序列复制、滚环扩增等。

可以在扩增期间标记所述扩增产物。例如，可以或者通过使用经合成带有化学标记(如生物素或碱性磷酸酶)或荧光标记的引物，或者通过使用标记的 dNTP 前体，标记所述扩增产物。一种特别有用的方法是使用合成的带有生物素末端标记的引物。

在包括连接的本方法的一个优选实施方案中(图 3 和图 5)，存在左引物和右引物，这两种引物对应着所述探针寡核苷酸的外部部分。所述左引物与所述左半边探针的外部部分相同，而所述右引物是所述右半边探针的外部部分的反向互补物。在反应混合物中未连接的半边探

针并不扩增到显著程度。(所述探针对的未连接的左半部分没有互补的引物，不被扩增；所述探针对的未连接的右半部分被线性扩增。)

步骤 7: 鉴定所述样品选择的 ID 探针: 使所扩增的探针序列与检测集合杂交。

为产生在所述实验样品中存在的基因组的代表性指纹，必须鉴定所述样品选定的扩增的 ID 探针。通过与由对应于(与之相应)原始未经选择的探针混合物中 ID 探针的 ID 序列或 ID 寡核苷酸或标记的集合的杂交，推导所选定 ID 探针的身份。所述集合中的序列可以对应于 ID 序列的部分或对应于掺入探针内部部分和外部部分之间的标记序列。上文的步骤 3 描述了检测集合的设计和构建。

可以使用多种方法中的任何一种进行所扩增的 ID 探针的鉴定。在一个实施方案中，使用扩增的 ID 探针，通过在液体介质中杂交而选择检测集合的成员。随后通过使用质谱确定分子量，鉴定所选定的检测集合成员。然后通过与完整检测序列集合的分子量表相比较，鉴定所选定的序列。在一个优选的实施方案中，通过与二维检测阵列的杂交，鉴定标记的扩增 ID 探针(见上文的步骤 3)。使用标准程序杂交和检测核酸分子(Ausubel 等, 1987, 见上文)。用于鉴定所扩增的 ID 探针的方法在下面的实施例中进一步描述。

步骤 8: 通过样品选定的 ID 探针与所述生物样品的原位杂交，定量所述生物样品中的靶生物。

定量生物样品中的靶生物常常是重要的。在医药领域中，例如，关于人类免疫缺陷病毒在血液中浓度的知识(也称为病毒载量，或滴度)对于估计疾病阶段以及对治疗的反应是重要的。当在样品的意外污染和真实感染之间进行区分时，对样品中靶生物的数量的了解也是重要的。

在步骤 7 中使用的标记 ID 探针可以用于通过使用原位杂交方

法，定量所述生物样品中的靶生物数量。使一部分经标记、扩增的、样品所选定的 ID 探针混合物变性，并用于与已固定的(可选已染色的)生物样品杂交。或者，可以使用前面步骤检测到的对于要检测的生物类型具有特异性的任何类群特异性序列作为探针。对于原位杂交，优选使用灵敏并且易于实施的方法(如 Huang 等, Modern Pathology 11: 971-977, 1998)，所述灵敏的方法如使用催化的报道分子沉积的方法，该方法足以使用单一拷贝序列检测单一细胞/病毒。所述固定的样品可以是在步骤 4 中使用的同样的样品，或者可以通过熟悉本领域的人员已知的其它标准方法制备(如 Nuovo 等，见上文)。

这些方法在下面的实施例中描述:

实施例 1　　测试胃肠样品中病原体的存在

肠胃炎。肠胃疾病是主要的国际健康问题。每年在儿童中出现约十亿病例，导致约五百万人死亡。该疾病的某些类型在症状出现的几小时内可能是致命的。很多种病原体引起胃肠疾病，其中包括细菌、病毒和原生动物。快速而准确地鉴定引起胃肠疾病的病原体对于选择合适的抗微生物疗法、鉴定医院获得性感染以及追踪食物传染的病原体的爆发是重要的，其中所述食物传染的病原体如新出现的病原体大肠埃希氏菌 O157:H7。

目前诊断胃肠疾病的方法还远远不够理想。由于可能的病原体(如病毒病原体、细菌病原体和寄生病原体)的数量和范围，确定感染因子的身份常常是困难、耗时(通常需要至少几天，有时甚至是几周)并且昂贵的。在正常消化道中不同微生物的存在加剧了鉴定肠胃炎的病因的难度。测试原生动物感染、病毒感染和细菌感染，以及检查样品中特征性人类细胞的存在，需要不同的专业化实验室设备。此外，进行这些测试必须雇佣高度训练有素的人员。

目标和好处。在本实施例中，我使用单一的基因组分布分析测定来测试来自患有胃肠疾病的患者的样品中，广泛范围胃肠病原体的存在。通过同时并且快速地(如几小时)测试常见的细菌病原体、病毒病

原体和原生动物病原体，以及测试特征性人类细胞的存在，本方法提供了优于目前实践的显著改进。本测试帮助确定合适并且及时的治疗。此外，由于基因组分布分析能够产生高分辨率指纹，因此该方法是用于流行病学分析的强有力的工具。

注意：在本实施例中描述的用于测试临床样品中胃肠病原体的基因组分布分析，对于食品检验工业也是有价值的工具。检验食物中的胃肠病原体对于预防胃肠疾病是重要的。

本实施例的总结。发展了一种基因组分布分析测定，所述测定在一次测试中，扫描胃肠样品中一组广泛的胃肠病原体的存在。我从各种胃肠病原体中分离了一个 ID 序列集合。对于细菌病原体和寄生虫，使用基因组扣除分离基因组差异序列和类群特异性序列。使用计算机分析来分离用于鉴定胃肠病毒的类群特异性序列。在给定病原体的 DNA 中存在的所述 ID 序列集合的亚组，构成了所述病原体的基因组分布分析指纹。通过确定在来自每一胃肠病原体类群的代表性菌株中存在的基因组差异序列亚组，构建指纹数据库。通过将临床样品中的基因组分布分析指纹与所述指纹数据库相比较，确定所述临床样品中病原体的身份。

在本实施例中使用的方法的总结。我使用了 Straus 等(Proc. Natl. Acad. Sci. USA 87: 1889-1893, 1990)的基因组扣除方法的改变形式，鉴定引起胃肠疾病的细菌和寄生虫的病原体特异性 ID 序列。可以使用其它可选的方法分离基因组差异序列，因此这些方法可以替代下面概述的扣除技术。对于引起胃肠疾病的病毒，我使用对序列数据库的计算机搜索，鉴定了类群特异性序列。通过使 ID 探针集合与已固定的特定样品的基因组 DNA 杂交，鉴定所述样品中的 ID 序列。一个 ID 探针亚组将与所述已固定的基因组 DNA 杂交，并因此被所述固定的基因组 DNA 保留下来。使用依赖于连接的 PCR 策略扩增已杂交的 ID 探针。通过使扩增的 ID 探针与检测集合杂交，鉴定它们的身份，在这种情况下，所述检测集合是完整的、未经选择的 ID 序列组的有序两维阵列。

在所述阵列上可见的杂交信号模式构成了基因组分布分析指纹。

从引起胃肠疾病的细菌中分离基因组差异序列

用于从细菌中分离 ID 序列的策略。为诊断胃肠疾病，最有用的诊断 ID 序列是那些在消化道病原体中存在、但在几百种居住在健康肠中的物种中不存在的 ID 序列。对于许多细菌性胃肠病原体来说，可以使用基因组扣除有效地分离这样的 ID 序列。如上文所讨论的(在详细描述部分的步骤 2)，所使用的基因组扣除策略取决于特定的病原体。本部分举例说明用于分离代表性胃肠病原体肠沙门氏菌和大肠埃希氏菌的基因组差异序列的两种不同策略。

从肠沙门氏菌分离基因组差异序列的策略。99%以上的沙门氏菌属临床分离物是肠沙门氏菌亚种的成员。肠沙门氏菌的所有菌株都被认为是人类病原体。因此，该类群是那些分类单位(生物学上相关的类群)的代表：对于那些分类单位来说，诊断目标是鉴定该类群的任一成员并将该类群的任一成员与该类群的任何其它成员区分开来。有许多使用现有的菌株分离用于高分辨率鉴定的标记的方法；本实施例使用在图 6 中图解说明的策略。

对于这种方法，将肠沙门氏菌的亚种分为两个亚群，即 X 群和 Y 群。汇集来自每个亚群的代表性成员的 DNA，构建 X 群的基因组差异序列和 Y 群的基因组差异序列。从 SARB 参考物保藏中心(SARB reference collection)获得来自每个分支的菌株(Boyd 等，J. Gen. Microbiol. 139: 1125-1132, 1993)。进行使用所述基因组差异样品的交互扣除。在一次扣除中，使用所述 X 基因组差异样品作为“+”样品，所述 Y 基因组差异样品作为“-”样品。该扣除的产物是在 X 群的至少一个成员中发现、但未在 Y 群的任何成员中发现的序列。在交互扣除实验中，使用所述 Y 基因组差异样品作为“+”样品，所述 X 基因组差异样品作为“-”样品。该扣除的产物是在 Y 群的至少一个成员中发现、但未在 X 群的任何成员中发现的序列。

通过该基因组扣除策略分离的基因组差异序列构成一个或多个家族。一般地说，该策略产生多于一个的家族，即一般不是所有的ID序列扣除产物都能与任何单个基因组杂交。因此，对汇集的生物的基因组扣除是从一个相关生物类群产生多个ID序列家族的有效方法。

从大肠埃希氏菌分离基因组差异序列的策略。图7A显示了大肠埃希氏菌类群的部分系统树。注意：该类群中的病原体(黑色)(大肠埃希氏菌 O157:H7 和弗氏志贺氏菌)具有非常密切相关的非致病性胞亲分类单位(sibling taxa)(白色)。对于未在该图中显示的大肠埃希氏菌系统树部分来说，这也是普遍的情况。在健康个体的消化道中多种非致病性或共生性大肠埃希氏菌的存在可能混淆对大肠埃希氏菌致病菌株的诊断。大肠埃希氏菌代表了在人体内发现的包含病原体和非病原体的生物类群。

为分离对这样的类群进行指纹分析的基因组差异序列，应用在图7B 和图 7C 中描述的策略。汇集来自非致病分类单位(分支)的代表性菌株，用它们的 DNA 制备“-”基因组差异样品。汇集来自致病分类单位(分支)的代表性菌株，用它们的 DNA 制备“+”基因组差异样品。

基因组扣除的产物是至少在病原体类群的至少一个成员(或者大肠埃希氏菌，或者弗氏志贺氏菌)中发现，但未在所述扣除的任何非致病菌株中发现的序列。注意：该基因组扣除将分离基因组差异序列，其中一些基因组差异序列也是类群特异性序列，因为它们出现在一个类群(如大肠埃希氏菌 O157:H7)的所有成员中，但不出现在相关类群的成员中。出现在致病性大肠埃希氏菌中(但不出现在非致病性大肠埃希氏菌中)的毒性基因(即涉及感染过程的那些基因)属于这一类产物。

用于本实验的菌株来自 Thomas Whittman 博士(Penn. State University)提供的 ECOR (非致病性)和 DEC (致病性)菌株保藏物。

表 3. 引起急性胃肠疾病的病原体。

细菌	寄生虫
大肠埃希氏菌	兰氏贾第鞭毛虫
沙门氏菌属	溶组织内阿米巴
志贺氏菌属	人酵母菌
小肠结肠炎耶尔森氏菌	隐孢子虫属
霍乱弧菌	*Microsporidium*
粪弯曲杆菌	美洲板口线虫
艰难梭菌	人蛔虫
病毒	毛首鞭虫
轮状病毒属	蛲虫
诺沃克病毒	粪类圆线虫
星状病毒属	麝猫后睾吸虫
腺病毒	华支睾吸虫
冠状病毒属	短膜壳绦虫

引起胃肠疾病的细菌病原体。表 3 列出了引起胃肠疾病的常见细菌类群。由某些这些病原体(包括霍乱弧菌和肠出血性大肠埃希氏菌(如大肠埃希氏菌 O157:H7))引起的感染甚至在健康个体中都可能是致命的。快速诊断是实现合适治疗和抑制爆发的关键。为从表 3 列出的细菌类群中分离 ID 序列家族，我使用上文所述应用于大肠埃希氏菌和沙门氏菌属的策略。

制备基因组 DNA 用于扣除。为制备 DNA 以制造基因组扣除样品，将表 3 所列出的菌株在液体培养基(500 ml)中培养直至饱和，并制备基因组 DNA (Ausubel 等, 1987, 见上文)。通过上文关于大肠埃希氏菌和沙门氏菌属的同样考虑来选择“+”菌株和“-”菌株。混合来自每个“+”菌株的 DNA (50 μg)(此后称为“+” DNA)。相似地混合来自所述“-”基因组差异样品菌株的 DNA (50 μg) (此后称为“-” DNA)。

制备基因组差异样品。为制备“-”基因组扣除样品，如以前所述(Straus, 1995, 见上文)，剪切“-” DNA，使其与乙酸光生物素反应，然后以 2.5 mg/ml 重悬浮。如下制备“+”基因组扣除样品：用限制酶 Sau3A 切割“+” DNA (2 μg)，产生具有粘性末端的片段。用乙醇沉淀后，将所述 DNA 片段以 0.1 μg/μl 重悬浮于 10 mM EPPS/1 mM EDTA, pH 8.0 (EE) (Straus, 1995, 见上文)。

基因组扣除。如以前所述(Straus, 1995, 见上文)进行基因组扣除。为分离病原体特异性 DNA 片段，使用来自致病菌株的“+”基因组扣除样品和来自非致病菌株的生物素化“-”基因组扣除样品进行基因组扣除实验。三个扣除杂交循环纯化了病原体特异性基因组差异序列。

克隆所述基因组差异序列。在将连接物连接到所述基因组差异序列后，使用 PCR 对它们进行扩增(Straus, 1995, 见上文; Straus 等, 1990, 见上文)。然后通过用 Sau3A 切割，从所扩增的基因组差异序列中除去所述连接物。将所述样品溶于 0.3 M 醋酸钠(NaOAc)，用苯酚/氯仿(1: 1)提取，然后用乙醇沉淀。将部分样品(20 ng)连接到用 BamH I 消化、去磷酸化的载体 pBluescript II KS+ (100 ng; Stratagene)，并将连接后的产物转化进大肠埃希氏菌中(Ausubel 等, 1987, 见上文)。

对所述基因组差异产物进行测序。使用 ABI DNA 合成仪，按照生产厂家的建议(Perkin-Elmer)，通过循环测序法，对单个克隆的插入片段进行测序。

从引起胃肠疾病的细菌分离基因组差异序列集合。通过对由表 3 列出的细菌类群中的生物制备的基因组差异样品进行如上文所概述的基因组扣除，从通常引起胃肠疾病的不同病原体类群分离基因组差异序列。每次扣除产生一个菌株类群内的病原体所特有的大量基因组差异序列。例如，在致病性大肠埃希氏菌菌株和非致病性大肠埃希氏菌菌株之间的一次扣除产生了几百种基因组差异序列(Juang, “取样调查大肠埃希氏菌 K1 分离物和 K2 分离物之间的基因组差异(Sampling

Genomic Differences Between *Escherichia coli* K1 and K2 isolates)," Harvard University, 1990)。

使用 DNA 序列数据库的基因组扣除。基因组扣除一般意义指扫描整个基因组寻找基因组差异序列，但也可以通过将已经完全测序(或近乎完全测序)的基因组的DNA序列与另一基因组(或另外多个基因组)的全部或部分进行比较而完成基因组扣除(见，例如，Alm 等, 1999, 见上文)。

制备对应于所述基因组差异序列的探针集合和检测集合

用如上文所述通过基因组扣除鉴定的病原体特异性 ID 序列集合，确定用于基因组分布分析的 ID 探针的结构。合成两个 ID 寡核苷酸集合。一个组成所述 ID 探针(或半边 ID 探针)的集合与生物样品杂交。连接与实验样品中的病原体基因组退火的半边 ID 探针，将其扩增并进行标记。另一个 ID 寡核苷酸集合构成一个检测集合。所述检测集合中的 ID 寡核苷酸对应于所述 ID 探针集合中的序列。也就是说，所述检测集合与所述 ID 探针集合相应。将所述检测集合寡核苷酸淀积到固相支持体上，构成一个可寻址阵列。通过与所述可寻址寡核苷酸阵列的杂交，鉴定与所述临床样品中的病原体基因组杂交的已标记、扩增的探针。

合成对应于所述 ID 序列的 ID 探针。从计划包括在基因组分布分析测定中的每个 ID 序列、人 mRNA (见下文)和对照序列中选出约 30 个碱基长的序列，该序列称为 ID 探针位点。合成对应于每种 30 个碱基 ID 探针位点的两个半边 ID 探针(图 3)。所述左半边 ID 探针包含所述 ID 探针位点的左边 15 个碱基和一个引物位点，即引物位点-L (“左”引物位点)。所述右半边 ID 探针包含所述 ID 探针位点的右边 15 个碱基和一个引物位点，即引物位点-R (“右”引物位点)。所述引物位点是对应于为 PCR 扩增所需要使用的引物类型的扩增位点。

所述引物位点-L (“左”引物位点)具有序列：5’-GACACTCTC-

GAGACATCACCGTCC-3'。所述引物位点-R ("右" 引物位点)具有序列: 5'-GTTGGTTTAAGGCGCAAGAATT-3'。因此，对于在上面部分鉴定的每种 30 个碱基序列，合成两个半边 ID 探针: 一个半边探针具有序列 5'-GACACTCTCGAGACATCACCGTCC-<ID 探针位点 $_{1\text{-}15}$>-3'，一个半边探针具有序列 5'-<ID 探针位点 $_{16\text{-}30}$>-GTTGGTTTAAGGCGCAAGAATT-3'。设计所述半边 ID 探针，使得当它们退火到包含所述 30 bp ID 探针位点的模板时相互邻接。当以这种方式退火时，可以连接所述半边探针，并因此转化为可以用引物 L (5'-GACACTCTCGAGACATCACCGTCC-3' 和引物 R (5'-AATTCTTGCGCCTTAAACCAAC-3')扩增的形式，其中所述引物 L 和引物 R 分别对应于所述左引物位点和所述右引物位点。

构建用于基因组分布分析的检测阵列。为检测哪些半边探针与临床样品杂交，可以通过杂交查询一个可寻址的 ID 序列检测集合。该集合的元件是对应于所述 ID 探针集合中的 ID 探针位点的合成 ID 序列寡核苷酸。也就是说，每种检测寡核苷酸约 30 个碱基长，并且与通过连接和扩增一对半边 ID 探针得到的 ID 探针位点序列的一条链互补。

在本实施例中，我按照 DiRisi 等(Science 278: 680-686, 1997)的程序，使用一台带有打印头的阵列形成机器(arraying machine)将每个寡核苷酸点样(Shalon 等, Genome Res. 6: 639-645, 1996)，构建了一个二维检测阵列。将每种约 30 个碱基长的寡核苷酸约 2.5 ng 点样到已经用聚 L-丝氨酸包被的 40 片载玻片的每一片上，其中在相邻寡核苷酸点之间的距离是 500 μm (Schena 等, 1995, 见上文)。

构建指纹的基因组分布分析数据库

基因组分布分析通过将患者样品的基因组分布分析指纹与包含已知生物的指纹的数据库相比较，鉴定所述样品中的病原体。(一种指纹对应于与特定类型生物杂交的 ID 探针集合的亚组)。构建指纹数据库需要从每个靶类群的一组参考菌株获取基因组分布分析指纹。

最好根据靶类群所属的两个诊断类别考虑构建所述数据库。大多数鉴定计划分为两类(根据靶类群)：简单测试在一个类群中的成员资格的鉴定计划，以及测试在一个类群中的成员资格并且将一个类群中的成员相互区分的鉴定计划。

将主要由类群特异性序列组成的指纹输入所述指纹数据库。当在一个类群中的成员资格是主要的考虑时，我在选定用于鉴定靶生物的 ID 序列家族中主要包括类群特异性序列。当一个类群的一个成员的存在几乎总是与疾病相关，并且当流行病学信息不具有很大价值时，测试作为该类群的成员的病原体的存在(不用在该类群的成员之间进行区分)常常是最佳的诊断策略。例如，为鉴定危险并且致病力强的胃肠病原体霍乱弧菌，该病原体引起危及生命的疾病霍乱，可以在所述集合中包括大部分由类群特异性序列组成的一个 ID 序列家族。注意：可以通过基因组扣除分离类群特异性序列，在所述基因组扣除中“+”菌株是病原体，“-”菌株是非病原体。这样的 ID 序列既是基因组差异序列，也是类群特异性序列。测试可能的类群特异性序列的特异性：使每一序列与来自该类群内代表性成员的基因组 DNA 杂交，并使每一序列与广谱的其它类群的成员杂交(见，例如，美国专利第 5,714,321 号)。这样，假如实验样品产生由对应于类群特异性 ID 序列的阳性信号组成的基因组分布分析指纹，则指示出在所述样品中存在靶类群的成员。将这样的指纹包括在指纹数据库中。

将主要由基因组差异序列组成的指纹输入所述指纹数据库。对于某些类型生物而言，诊断目标可能是鉴定作为一个类群的成员的一个菌株，同时将该菌株与该类群中的其它菌株区分开来。例如，在追踪医院获得性感染的爆发和食物传染的病原体的爆发时，这样的亚菌株鉴定是重要的。这种类型的高分辨率鉴定需要比仅仅鉴定作为靶类群成员的病原体(如在前面的段落描述)更为详细的指纹。通过基因组扣除分离的基因组差异序列是用于获得高分辨率指纹最有用的 ID 序列。

为从靶类群构建指纹数据库，我从该类群代表性的一组参考菌株获得指纹。为产生指纹，对包含单个参考菌株的基因组的样品(常常是单个细菌菌落)应用基因组分布分析测定。扫描所述基因组中该靶类群特征性的一个或多个 ID 序列家族的成员(通常是对应于基因组扣除产物的基因组差异序列)的存在。将所获得的指纹储存在所述数据库中。根据所述指纹，使用标准分析建立所述参考菌株的系统发生关系(Hillis 等, *Molecular Systematics* (Sinauer Associates, Sunderland, 1996))。

构建用于对食物传染的病原体(如大肠埃希氏菌 O157:H7)进行高分辨率指纹分析的数据库是用于追踪爆发的重要工具。例如，我通过获取大肠埃希氏菌和志贺氏菌属菌菌株的参考保藏物的基因组分布分析指纹，建立了代表大肠埃希氏菌/志贺氏菌属类群中生物范围的指纹数据库。从疾病控制中心和美国典型培养物保藏中心可以获得大量这样的菌株。使用所述指纹作为特征组，构建该类群的系统发生(相关性的进化树)。该方法的一个强有力特征是：当使用在临床样品中发现的相关病原体的新指纹更新该类群的指纹数据库时，该数据库逐渐变得更加完全。

制备用于使用基因组分布分析测定进行指纹分析的细菌菌菌株。为获得指纹，我首先将细菌菌落固定在尼龙滤膜上，并使用简单和标准的方法(Grunstein 等, 1975, 见上文)，使所述菌落的基因组 DNA 能够用于杂交。将所述菌落涂布在尼龙滤膜(1 cm^2)上，使其干燥，然后用 0.5 M NaOH, 1 M Tris, pH 8/3 M NaCl, 1 M Tris, pH8 顺序处理(每种处理 5 分钟)。将固定在所述尼龙滤膜上的样品在 1 M NaCl 中于 65 ℃振荡下洗涤 3 次，每次 5 分钟，以除去未固定的化学制剂和颗粒性物质。在碱处理前，用特定的酶或化学制剂预处理所述涂布的生物，可以增强某些细菌(和其它生物)的有效裂解。例如，通过用包含磷脂酶和溶菌酶的溶液处理滤膜，帮助裂解革兰氏阳性细菌(Graves, L.等(1993), "通用细菌 DNA 分离程序," 载于 *Diagnostic Molecular Microbiology, Principles and Applications*, D. Persing 等编辑(Washington,

D. C. ASM Press), 第 617-621 页)。

选择与一种细菌菌株的 DNA 杂交的基因组差异序列亚组。基因组分布分析测定选择与结合于尼龙滤膜的基因组 DNA 杂交的病原体特异性 ID 探针亚组。相比之下，可以容易地从滤膜上除去在已固定的细菌 DNA 中没有对应物的基因组差异探针。在随后的连接步骤中，任何通过与滤膜或样品的非特异性相互作用而保持附着于所述滤膜的残余半边 ID 探针将是不可扩增的。

在 36℃下(或在比在 1 M NaCl 中所有半边探针的最低 T_m 低 5℃的温度下)，在 0.5 ml 杂交缓冲液(1 M NaCl/50 mM EPPS/2 mM EDTA, pH 8)中，使对应于来自特定细菌类群的病原体特异性基因组差异序列的一组半边探针(每种半边探针 1 nM)与所述滤膜杂交。将所述杂交反应物温育 30 分钟，然后通过在 2 ml 洗涤缓冲液(1 M NaCl/50 mM EPPS/2 mM EDTA, pH 8)中于 36℃(或在比在 1 M NaCl 中所有半边探针的最低 T_m 低 5℃的温度下)伴随振荡的五个洗涤步骤，每个洗涤步骤 30 秒钟，除去未结合的半边探针。随后用 1 ml 连接缓冲液(10 mM $MgCl_2$/50 mM Tris-HCl/10 mM 二硫苏糖醇/1 mM ATP/25 μg/μl 牛血清白蛋白)，在 30℃下连续洗涤所述滤膜 3 次。在连接步骤前，除去所述滤膜上的多余液体。在各步骤之间不能使所述滤膜干燥。

连接与所述细菌样品杂交的成对半边探针。消除由于非特异性结合的探针分子引起的背景对于基因组分布分析是至关重要的，尤其是当应用于临床样品时更是如此，因为如在下面的部分所述，在这样的样品中检测未经培养的病原体需要高度的灵敏度。回想起要求连接邻近结合的半边探针是有效的方法，保证仅有的可以被扩增的探针是那些已经与所述样品中的病原体基因组杂交的探针。

通过加入含 1,600 粘性末端单位(等于 25 Weiss 单位)的 T4 DNA 连接酶(New England Biolabs)的 200μl 连接酶缓冲液(10 mM $MgCl_2$/50 mM Tris-HCl/10 mM 二硫苏糖醇/1 mM ATP/25 μg/μl 牛血清白蛋白)，连接与所述固定样品杂交的半边探针。使所述连接反应在 30℃进行 1

小时。

扩增与所述细菌样品杂交的基因组差异序列。通过加热，从所述滤膜释放与所述细菌样品中的基因组杂交的成对已连接的半边探针。然后使用聚合酶链式反应和对应于在所述已连接探针分子末端的引物结合位点的引物，扩增所述已连接的半边探针。

在连接所述半边探针后，用 2 ml 10 mM EPPS/1 mM EDTA, pH 8.0 洗涤滤膜，从所述滤膜上除去液体，然后在所述滤膜加上 500 µl 10 mM EPPS/1 mM EDTA, pH 8.0，随后在 100℃温育 5 分钟。在将溶液与滤膜分离开后，加入 50 µl 3 M 醋酸钠和 20 µg 酵母 tRNA。通过乙醇沉淀纯化核酸：将 1 ml 乙醇与所述样品混合，然后将所述样品在 12,000 g 离心 5 分钟。用 100%乙醇洗涤所述核酸沉淀，干燥，并重悬浮于 10 µl 10 mM EPPS/1 mM EDTA, pH 8.0 中。

使用 10X PCR 缓冲液(Boehringer Mannheim)、200 µM 每种 dNTP (dATP、TTP、dCTP 和 dGTP)、1 µM 生物素化寡核苷酸引物 L (5'-(生物素-dX)GACACTCTCGAGACATCACCGTCC-3') (Midland Certified Reagent)、1 µM 生物素化寡核苷酸引物 R (5'-(生物素-dX)AATTCTTGCGCCTTAAACCAAC-3')和 0.1 单位/µl Taq 聚合酶 (Promega)，将一半(5µl)包含所洗脱探针的样品溶于总反应体积为 50µl 的 1X PCR 缓冲液中。使用如下 PCR 模式扩增所述所洗脱的探针：30 个循环(94℃30 秒钟，55℃30 秒钟，72℃1 分钟)，然后是 72℃10 分钟。

一个菌株的基因组分布分析指纹：通过与一个阵列的杂交，鉴定扩增的细菌 DNA 所选定的探针分子。鉴定通过与所述菌株的固定化 DNA 杂交而选定的 ID 探针，建立菌株的指纹。在本实施例中，我通过使扩增的选定 ID 探针与检测阵列杂交，鉴定了由细菌基因组 DNA 选定的 ID 探针。该阵列是一个二维的可寻址序列阵列，与用于与所述生物样品杂交的 ID 探针集合相应。这样，该集合中的每种 ID 探针都可以与在该检测阵列确定位点的 DNA 序列杂交。通过与所述阵列的杂交，鉴定通过与所述细菌样品的结合而选定的探针。只有选定探针通

过与所述阵列上的对应点结合，产生信号(图 5)。

通过在 100℃加热 1 分钟，我使得代表与所述细菌样品杂交的序列的扩增探针变性。将所述已变性的探针加入 25 ml 2X 杂交缓冲液(2 M NaCl/100 mM EPPS, pH 8/10 mM EDTA/0.2%十二烷基硫酸钠)中。将所述探针/杂交混合物置于所述阵列上，用玻璃盖玻片覆盖，并在 50℃温育 20 分钟(如 Schena 等, 1995(见上文)所述)。通过在 2 ml 洗涤缓冲液(0.4 M NaCl/50 mM EPPS/2 mM EDTA, pH 8)中于 50℃伴随振荡的五个各 30 秒钟的洗涤步骤，除去未结合的探针。

并如已公开的报道所述(DiRisi 等, 1997, 见上文; Schena 等, 1995, 见上文)，用激光荧光扫描仪扫描微阵列，并处理和记录信号。将每个菌株的指纹记录为 1 和 0 的二进制字符串，每个数字代表在微阵列上的一种基因组差异序列。假如在该微阵列的一个位点获得信号，一个“1”就出现在代表该基因组分布分析指纹的字符串中的对应数字。

使用基因组分布分析指纹和系统发生分析对类群中的菌株分型。可以使用针对类群中代表性菌株的指纹数据库鉴定未知菌株。如上文所述编制指纹数据库，并如 Hillis 等(见上文)所述，使用标准方法进行所述指纹的系统发生分析。通过将未知指纹与以系统发生排序的指纹数据库相比较(使用 Hillis 等(见上文)所述的方法)，确定未知病原体如在患者样品中未知病原体的身份。

从引起胃肠疾病的寄生虫分离 ID 序列

引起胃肠疾病的寄生虫。根据地理位置、气候、社会经济因素和免疫活性，在患者体内发现的肠寄生虫范围有所不同。表 3 列出了北美洲通常在患有胃肠疾病的患者体内发现的原生动物和蠕虫类群。目前准确诊断肠寄生虫的方法按最好来说也是困难的。基因组分布分析极大改善了胃肠寄生虫的检测。

从引起胃肠疾病的寄生虫分离 ID 序列。为分离表 3 中每一种寄生虫所独有的 ID 序列组，我使用了在上文概述的针对细菌病原体的相同策略和方法，只有下面一些小的改动。因为寄生虫一般与在消化道

中通常发现的生物不相关，所以从来自目的分类单位内隔开最远的两个菌株的基因组DNA构建基因组差异样品常常就足够了。进行交互杂交，即每一个菌株在一个扣除中作为“+”菌株，但在另一个扣除中作为“-”菌株。与细菌扣除的温育时间相比，增加扣除杂交反应的温育时间对于补偿真核生物基因组复杂性的增加是必要的。我使用的复性时间是一半单拷贝序列重退火所需时间的四十到五十倍(Straus, 1995, 见上文)。

构建寄生虫指纹的数据库。如上文关于细菌病原体的指纹分析所述，使用寄生虫ID序列构建用于鉴定表3所列出的生物的ID探针家族。也如针对细菌病原体所述，进行对参考菌株的指纹分析和构建指纹数据库。

鉴定引起胃肠疾病的病毒的类群特异性序列

引起胃肠疾病的病毒。据认为病毒性肠胃炎是美国第二最常见的疾病病因。儿童和免疫妥协患者尤其易感。诊断病毒引起的胃肠疾病是有问题的，因为大多数常见因子不能培养并且很少特征鉴定。已经发展出的测试一般非常昂贵。由于可用测试的费用、严重并发症的不常见性、普通支持性治疗、以及缺乏抗病毒治疗，一般不进行诊断测试。然而，对病毒全面并且不昂贵的测试对流行病学、对排除其它病因、对排除抗生素的使用以及对指示恰当地给予新型抗病毒治疗是有用的。表3列出了通常引起胃肠疾病的病毒病原体。

鉴定来自引起胃肠疾病的病毒的类群特异性序列。对于引起胃肠疾病的病毒，从已公开的DNA序列数据推导出类群特异性序列。在一些情况下，病毒类群特异性序列已经在文献中描述。在其它情况下，在将公共数据库中的病毒基因组序列与所述数据库中其它病毒的序列比较后，从所述病毒基因组序列选出序列。使用标准方法进行序列比较(Ausubel 等, 1987, 见上文)。选择至少30 bp长的病毒类群特异性序列作为测试探针的靶。

构建病毒指纹的数据库。如上文针对细菌病原体的指纹分析所

述，使用寄生虫 ID 序列构建用于鉴定表 3 中的病毒的 ID 探针家族。除样品制备外，对参考病毒株的指纹分析和构建病毒指纹数据库也如上文针对细菌病原体所述进行。对于包含 RNA 基因组的病毒，样品制备必须保证 RNA 的完整性。我通过高压灭菌处理滤膜(Allday 等，Nucleic Acids Res. 15: 10592, 1987)或将滤膜放在微波炉中烘烤(Buluwela 等, Nucleic Acids Res. 17: 452, 1989)，使核酸变性，将其固定到滤膜上，并使其可以接触探针。

用于诊断胃肠疾病的人类序列

基因组分布分析测定的一个好处是：可以在筛选病原体的同一测试中测定在诊断上有用的人类细胞类型。例如，在胃肠疾病中，重要的是知道白细胞和红细胞是否在临床样品中过高。为测试特定细胞类型，获得细胞类型特异性 mRNA 的序列(通常得自已公开的报告或遗传数据库)。表 4 指出了已知的在某些细胞类型中表达并且在诊断胃肠疾病中重要的序列的细胞类型特异性 mRNA。

合成与 ID 探针类似的探针(即作为带有扩增位点的二元半边探针)，并将所述探针包括在用于接触所制备的生物样品的杂交混合物中。对应的检测序列包括在检测阵列中。

表 4. 用于对诊断胃肠疾病重要的人类细胞的探针

转录物	转录物的特征
乳铁蛋白	白细胞的产物－指示侵袭性感染
LCA、CD45	白细胞特异性的
珠蛋白	红细胞的产物－指示出血
肌动蛋白	对于所有人类细胞是共有的(用作人类特异性探针)

可用于评估基因组分布分析测定的内部对照序列

内部对照。在基因组分布分析测定中包括内部对照改善了测试结果的置信度并允许进行有效的故障检查。对照探针、寡核苷酸和检测

序列包含非生物学序列。

假如技术起作用的话，阳性对照序列在每个实验中都给出阳性信号。假如，例如，其中一种试剂不正常作用，将缺乏来自阳性对照的预期信号。缺乏来自所述阳性对照的信号保证避免了由于技术失败引起的假阴性。

包括阴性对照，以监测所述探针中不在所述临床样品中的序列是否在所述诊断检测测定中导致信号。设计所述基因组分布分析测定，以便只有当所述ID探针集合中的ID探针对应于所述临床样品中的ID序列时，才能在所述检测阵列上获得信号。阴性对照的使用与阳性对照相似，只是没有将对应序列与所述临床样品一起点样(即它与所述ID探针集合一起包括在所述杂交混合物中，并且是所述检测阵列的元件)。这样，阴性对照序列应该不能被所述固定的样品选择，并且不能被连接和扩增。来自检测阵列中阴性对照序列的阳性信号，指示选择ID探针与靶序列的杂交的步骤并未适当地运作。

我在所述测定中包括了另一种对照探针，所述探针允许监测连接酶反应。该探针不作为半边探针合成，而是作为以左连接物和右连接物标记的连续序列合成。另外，将该序列与阳性对照探针一样使用(即将其与所述临床样品平行点样，其包括于所述探针中，并且是所述检测阵列的元件)。假如所述检测阵列的阳性对照元件是阴性的，但该检测阵列的连接酶对照元件是阳性的，那么所述测定中的连接酶步骤是值得怀疑的。

表 5. 用于基因组分布分析测定的内部对照。

对照类型	对照功能	在带有样品的滤膜上存在的对照序列	在探针中存在的对照序列
阴性对照	指示由与样品中的 DNA 不匹配的探针获得的信号的背景水平	不存在	存在
连接对照	假如在测定中的所有非连接步骤工作，将给出阳性信号	存在	存在
阳性对照	假如在测定中的所有步骤工作，将给出阳性信号	存在	存在

鉴定在临床样品中存在的病原体

制备临床样品。为使基因组分布分析在临床设置中最有效，优选一种用于制备临床样品以与半边探针杂交的简单方法。为了实验室工作人员的安全，患者样品的制备最好也应当快速中和所述样品中存在的病原体，并且样品的制备应当有效除去随后酶促反应如探针扩增的抑制剂。

我使用一种普遍用于制备用于杂交的在生化上复杂的生物样品的简单、通用但有有效的方法(Grunstein 等, 1975, 见上文)，固定所述临床样品、将核酸分子变性并中和任何病原体。在尼龙滤膜(1 cm^2)上涂抹胃肠样品(0.5 ml 液体粪便样品、成形的大便样品、或直肠药签样品)，使其干燥，并如上文关于制备病毒样品所述进行处理。将所述固定在尼龙滤膜上的样品在 65℃下振荡洗涤几次，以除去未固定的化学制剂和颗粒性物质。

通过杂交扫描临床样品中基因组差异序列的存在。通过使所述ID探针集合、人类诊断序列和对照序列与胃肠样品杂交，我扫描了所述样品中广泛的相关病原体组。该方法与用于对参考菌株进行指纹分析以建立细菌指纹数据库的方法基本相同(见上文)，不同之处在于所述ID探针集合的广泛组成以及使用临床样品(如前面的段落所述制备)作为生物样品。

获得临床样品的基因组分布分析指纹。根据如上文针对细菌所详细描述的相同方法(见“构建指纹的基因组分布分析数据库”)，进行连接、扩增和指纹显示(阵列检测)，与该方法的不同之处在于所述阵列包含代表表3指出的所有病原体的检测集合。所述检测阵列中的检测序列对应于与所述临床样品杂交的所述ID探针集合、人类诊断序列和对照序列。

定量分析：所述临床样品中的病原体滴度是多少？基因组分布分析测定的一个强有力特征是能够定量生物样品中的病原体。一旦已经通过指纹鉴定了靶生物，就可以通过与根据标准方法(如 Huang 等, Modern Pathology 11: 971-977, 1998)制备的一部分原始生物样品进行原位杂交，定量它们的存在。我使用一种足以检测单个生物中的核酸序列的单个分子的灵敏但简单的方法(Huang 等，见上文，1998)。该方法与用于与所述检测阵列杂交的标记探针一起使用。或者，可以使用任何所述生物特征性的、可以通过与所述阵列的杂交而检测的类群特异性探针进行原位杂交。

实施例2. 检测呼吸系统样品中病原体的存在

肺炎。肺炎是美国由于传染病死亡的最常见的死因。该疾病的病因学依赖于年龄和免疫状况。病毒引起大部分的儿童肺炎，而细菌病原体是引起成人肺炎的最常见病原体。在免疫妥协寄主中引起肺炎的病原体谱变化很大，并且对于癌症影响免疫系统或保护性表面(粘膜表面或皮肤)的患者、移植物受体和HIV感染患者有所不同。

为成功治疗肺炎，最基本的是快速鉴定病原体。但是，所有确定肺炎病因的诊断努力几乎有一半不能鉴定病因因子。(这还不包括没有尝试鉴定病原体的大部分病例。)引起下呼吸道感染的许多细菌病原体和所有病毒病原体不能通过常规微生物培养方法鉴定。例如，鉴定引起肺结核、百日咳、军团病和支原体引起的肺炎的病原体需要特殊方法。患有下呼吸道感染的患者使用了美国处方开出的75%的抗生素。由于目前的诊断不能鉴定大多数下呼吸道感染中的病原体，故每年约10亿美元浪费在无用的抗生素上。因此，对于测试广泛的下呼吸道病原体组的单次诊断测定有极大的需求。

目的和好处。在本实施例中，我使用一次基因组分布分析测定，检测来自表现出下呼吸道疾病症状的患者的样品中呼吸系统病原体的存在。通过同时并快速地(如在几小时内)测试常见的细菌病原体、病毒病原体和原生动物病原体，本方法提供了比目前实践显著的改进。所述测试帮助确定合适和及时的治疗。此外，由于基因组分布分析测定能产生高分辨率指纹，因此它是用于流行病学分析的有力工具。

本实施例概述。在细菌病原体和寄生虫的情况下，我使用基因组扣除从各种下呼吸道病原体分离ID序列，或者在病毒的情况下，我使用计算机分析分离ID序列。在给定菌株的DNA中存在的基因组差异序列亚组构成了其基因组分布分析指纹。通过确定在每个呼吸系统病原体类群的代表株中存在的ID序列亚组，构建指纹数据库。通过将临床样品的基因组分布分析指纹与指纹数据库相比较，确定所述临床样品中的病原体身份。

在本实施例中使用的方法概述。在本实施例中，我使用抑制扣除杂交来分离病原体特异性基因组差异序列，而不使用在实施例1中使用的基因组扣除。如前面实施例所述，通过使用特定样品的基因组DNA通过杂交来选择一组ID探针，鉴定所述样品中ID序列的身份。随后使用高分支滚环扩增方法(hRCA)(Lizardi等, Nat. Genet. 19: 225-232, 1998)，扩增所选定的ID探针。通过使用与实施例1中所述不同

的检测阵列技术，我确定了由所述样品选定的 ID 探针的身份。

从引起下呼吸系统疾病的病原体分离 ID 序列。表 6 列出了一些引起下呼吸道感染的常见病原体。使用得自 Clontech 的抑制扣除杂交试剂盒(Diatchenko 等，Proc. Natl. Acad. Sci. USA 93: 6025-6030, 1996)，根据厂家推荐的方法，从非病毒(即细菌和真菌)病原体分离 ID 序列。如实施例 1 中一样，选择用于从不同类群分离 ID 序列的扣除计划(如，选择使用汇集的基因组差异样品或者单菌株基因组差异样品)。如实施例 1 中所述，表 6 中列出的特定类群的“+”基因组差异样品由来自该类群的一种或多种代表性病原体的 DNA 组成，而“-”菌株由来自一种或多种密切相关的非致病性生物的 DNA 组成。(对于所有已知代表都是病原体的类群，所述“+”和“-”样品包括来自致病菌株亚类群的汇集的 DNA。)对通过基因组扣除分离的基因组差异序列进行测序，以准备用于合成滚环扩增探针和引物(见下文)。

对于引起下呼吸道疾病的病毒，从已公开的 DNA 序列数据推导出类群特异性序列。合成对应于在一个病毒类群内保守、但未在其它病毒类群中发现的序列的 ID 探针。我通过将可能的类群特异性序列与病毒序列数据库(如 Genbank)相比较，选出符合所述比较标准的序列。

表 6. 引起下呼吸道疾病的病原体。

细菌	真菌
白喉棒杆菌	荚膜组织胞浆菌
结核分枝杆菌	*Coccidoides immitis*
肺炎支原体	新型隐球酵母
沙眼衣原体	皮炎芽生菌
肺炎衣原体	卡氏肺囊虫
百日咳博德特氏菌	**病毒**
军团菌属	呼吸道合胞病毒
诺卡氏菌属	腺病毒
肺炎链球菌	单纯疱疹病毒
流感嗜血菌	流感病毒
鹦鹉热衣原体	副流感病毒
铜绿假单胞菌	鼻病毒
金黄色葡萄球菌	

可用于判断呼吸系统样品质量的组织特异性序列。众所周知呼吸系统样品的质量不均一。方便并且非侵袭性收集的痰样品常常由于受到上呼吸道生物的污染而被丢弃。已经根据显微镜观察到的扁平上皮细胞与多形核白细胞的比例，发展出判断样本质量的系统。在我的呼吸系统测定中，我包括了一种基于内部杂交的测试，以根据这两种细胞类型的相对丰度判断下呼吸道样品的质量。通过测试来自多形核白细胞的细胞类型特异性转录物(编码蛋白 LCA 和 CD45)和来自扁平上皮细胞的细胞类型特异性转录物(编码蛋白 spr 1)的转录物相对水平，完成这一工作。

使用用于构建对应于 ID 序列的 ID 探针的同样方法，合成具有对应于所述组织特异性序列的探针位点的组织特异性序列探针，不同之处在于从 GenBank 数据库获得所述序列。这些探针与所述 ID 探针集

合一起包括在所述杂交混合物中，并且这些探针包括在所述检测阵列上。

还包括用于定量所述组织特异性 mRNA 的代表的对照序列。所述对照序列是以不同量加入所述生物样品中的一系列独特非生物的 RNA 序列。在所述杂交混合物和检测阵列中包括对应的探针和检测序列。通过在具有已知数量的扁平上皮细胞和多形核白细胞的样品上进行所述测定，完成这些定量对照的校准。

用于滚环扩增的 ID 探针和引物。对于所述呼吸系统基因组分布分析测定中的每一 ID 序列，合成一对 ID 探针(图 8A)和一对引物(图 8B)。ID 探针和引物基于 Lizardi 等(1998, 见上文)的缺口寡核苷酸方法的那些探针和引物。然而，所述缺口 ID 探针(约 15 个碱基)和所述带有缺口的环状 ID 探针(约 15 个碱基)对应于一个 ID 序列。同时，在本实施例中，我使用 5'生物素化引物以用于滚环扩增(图 8C)。相似地，合成对应于实施例 1 中所述的实验对照序列的 ID 探针和对应于组织特异性 RNA 的 ID 探针。

构建用于基因组分布分析测定的两维检测阵列。为确定哪些 ID 探针与样品杂交，我使扩增的选定 ID 探针与一个检测阵列(包含一个检测序列的集合的可寻址阵列)杂交。该阵列的元件包括对应于滚环扩增探针对的缺口探针部分的寡核苷酸和对应于实验对照序列的寡核苷酸。在本实施例中，我使用光刻法，如以前所述构建微阵列(Chee 等, Science 274: 610-614, 1996; Lockhart 等, Nat. Biotech. 14: 1675-1680, 1996)。

对呼吸系统病原体进行指纹分析

为鉴定引起下呼吸道感染的病原体，我将临床样品的基因组分布分析指纹与来自以前特征鉴定的生物的指纹数据库相比较。如在涉及胃肠基因组分布分析测定的实施例 1 中一样，我首先从来自每个病原体类群的参考菌株的基因组分布分析指纹组装了指纹数据库。然后将临床样品的指纹与该数据库相比较，确定所述样品中病原体的身份。

获得参考菌株的指纹并组装数据库。样品制备、与所述 ID 序列集合的杂交以及洗涤步骤与实施例 1 中描述的那些步骤相同，不同之处在于所述 ID 探针集合的组成和结构。当与已固定的样品中的 DNA 退火的成对带缺口环状 ID 探针和缺口 ID 探针相互连接时，就产生了用于高分支滚环扩增(HRCA)的模板。如图 8 图解并如以前所述(Lizardi 等, 1998, 见上文)，进行连接和 HRCA。如以前所述(Lockhart 等, 1996, 见上文)完成与微阵列的杂交、用链霉抗生物素-藻红蛋白染色、以及扫描。从所述微阵列数据获得指纹，并使用实施例 1 中所述的方法，组装和分析由每一呼吸系统病原体类群获得的指纹数据库。

鉴定在临床样品中存在的病原体。使用实施例 1 中所述的方法，将各种类型和质量的呼吸系统样品(如痰样品、支气管肺泡灌洗样品和支气管刷缘样品)加样并固定到尼龙滤膜上。如实施例 1 一样，对临床样品以及参考菌株进行指纹分析，不同之处在于所述杂交反应中包括来自表 6 中所有呼吸系统病原体类群的 ID 探针。通过将获得的指纹与参考菌株的指纹数据库中的指纹相比较，鉴定在临床样品中存在的病原体。

实施例 3 – 测试血液样品中的病原体

血流感染。心血管系统的致病性侵袭是最严重的传染病之一。在美国每年发生的约 200,000 例血流感染中，20%到 50%是致命的。尤其危险的是免疫妥协患者、太幼小的儿童和太老的老人、患有皮肤或软组织感染和带有伤口的患者、以及侵入性医疗程序的接受者。所有主要病原体类型都可以感染血流，其中包括细菌、病毒、真菌和寄生虫。快速鉴定血流感染中的病原体对于制定合适的(可能是救命的)治疗是至关重要的。

目前的方法一般是病原体特异性的。因此，确定感染来源可能需要许多测试和大量费用。存在对于快速确定广泛范围常见血流病原体的身份的单次测试的需求。

目标和好处。在本实施例中，我使用单次基因组分布分析测定来测试在临床样品中广泛范围的血流病原体的存在。通过同时并且快速地(如在几小时内)测试常见细菌病原体、病毒病原体和原生动物病原体，本方法提供了比目前实践显著的改进。该测试的快速性使得其对于快速诊断血流病原体以及制定合适和及时治疗的关键任务特别有用。此外，由于基因组分布分析测定能够产生高分辨率指纹，因此它是进行流行病学分析的强有力工具。

本实施例概述。我使用基因组扣除(细菌病原体和寄生虫)或计算机分析(病毒)从各种血流病原体分离 ID 序列。在给定株的 DNA 中存在的 ID 序列亚组构成该株的基因组分布分析指纹。通过确定在每个血流病原体类群的代表株中存在的 ID 序列亚组，构建指纹数据库。通过将临床血流样品的基因组分布分析指纹与指纹数据库相比较，确定该临床样品中病原体的身份。

在本实施例中使用的方法的概述。在本实施例中，我使用 Tinsley 等(Proc. Natl. Acad. Sci. USA 93: 11109-11114, 1996)的改良的表现度差异分析(representational difference analysis)基因组扣除方法，分离病原体特异性 ID 序列，而不是在以前实施例中使用的方法。如前面实施例所述，通过使用特定样品中的基因组 DNA 通过杂交选择一组 ID 探针，确定在所述样品中的 ID 序列的身份。然而，在本实施例中，通过液相杂交-捕获方法，分离所选定的探针。同时，在本实施例中，我使用质谱法鉴定所选定的扩增 ID 探针，而不是使用前面实施例中所述的微阵列方法。

从引起血流感染的病原体分离 ID 序列。表 7 列出了一些引起血流感染的常见病原体。使用 Tinsley 等(1996, 见上文)所改良的表现度差异分离方法，从所述非病毒(即细菌、真菌和寄生虫)病原体分离 ID 序列。如在实施例 1 中所述，针对表 7 中列出的特定类群的“+”基因组差异样品由来自该类群的代表性病原体的 DNA 组成，而“-”基因组差异样品由来自密切相关的非致病性生物的 DNA 组成。(对于其中

所有已知代表都是病原体的类群，所述“+”和“-”样品包括由致病菌株亚类群汇集的 DNA。)对于引起血流感染的病毒，如在前面的实施例所述，从已公开的 DNA 序列数据推导出 ID 序列。

表 7. 引起血流感染的病原体。

细菌	真菌
凝固酶阴性葡萄球菌	疟原虫(Plasmodium spp.)
金黄色葡萄球菌	杜氏利什曼原虫
Viridans streptococci	弓形虫(Toxoplasma spp.)
肠球菌(Enterococcus spp.)	微丝蚴
β溶血性链球菌	真菌
肺炎链球菌	荚膜组织胞浆菌
埃希氏菌(Escherichia spp.)	*Coccidoides immitis*
克雷伯氏菌(Klebsiella spp.)	新型隐球酵母
假单胞菌(Pseudomonas spp.)	假丝酵母(Candida spp.)
肠杆菌(Enterbater spp.)	**病毒**
变形菌(Proteus spp.)	HIV
拟杆菌(Bacteroides spp.)	单纯疱疹病毒
梭菌(Clostridium spp.)	丙型肝炎病毒
铜绿假单胞菌	乙型肝炎病毒
棒杆菌(*Cornybacterium spp.*)	巨细胞病毒
	EB 病毒

用于捕获 ID 序列的 ID 探针、扩增和质谱检测。对于所述血流基因组分布分析测定中的每个 ID 序列，合成一对 DNA 捕获 ID 探针、两种扩增 ID 探针、一种缺口 ID 探针和一种质谱检测寡核苷酸(图 9A-9C)。每种捕获 ID 探针具有两个部分：一个生物素化臂(约 10 个碱基长)和对应于一种 ID 序列的一部分的一个臂(约 15 碱基长)。所述左

扩增探针和右扩增探针也具有两个部分：一个部分包含对应于扩增探针的序列(约 20 碱基长)，一个部分与一种 ID 序列互补(约 15 个碱基长)。合成在 5'末端生物素化的引物，以便可以扩增已连接的三联探针(图 9B)并进行亲和纯化。所述缺口 ID 探针(约 20 个碱基长)与一种 ID 序列互补，并且当所述缺口 ID 探针退火至对应的 ID 探针时，它与所述左扩增 ID 探针和所述右扩增 ID 探针相邻。与实施例 1 中描述的那些方法相似地合成阳性对照探针和阴性对照探针并使用它们，不同之处在于本实施例中的样品溶液包括实施例 1 中与所述滤膜结合的阳性对照探针。

为确定哪些 ID 探针与样品杂交，我使扩增的选定 ID 探针与对应于需要测定的 ID 探针集合的质谱检测寡核苷酸杂交。每种质谱检测寡核苷酸约 8-15 个核苷酸长(质谱获得小寡核苷酸的非常高分辨率的区别)，并且每种质谱检测寡核苷酸与一种探针的缺口探针部分互补(图 9C)。在该集合中的各种质谱检测寡核苷酸应当都具有独特的分子量，以便可以通过质谱鉴定它们的身份。为增强具有相似分子量的寡核苷酸间的分子量区别，在某些情况下，包括化学修饰的寡核苷酸是有用的。具有各种各样的化学修饰以及具有最少改变的复性特征的寡核苷酸是商业化可得的。

对血流病原体进行指纹分析

如前面实施例所述，为鉴定引起血流感染的病原体，我将临床样品的基因组分布分析指纹与来自以前特征鉴定的生物的指纹数据库相比较。如以前所述，我首先从来自表 7 列出的每个血流病原体类群的参考菌株的基因组分布分析指纹组装指纹数据库。然后将临床血液样品的指纹与该数据库相比较，确定在所述样品中任何病原体的身份。

捕获和扩增与参考菌株的 DNA 杂交的 ID 探针。在本实施例中，我使用液相杂交-捕获方法(Hsuih 等, J. Clin. Microbiol. 34: 501-507, 1996)来亲和纯化在一个参考菌株的核酸分子中存在的病原体特异性 ID 序列。通过在 5 M 硫氰酸胍中温育(在 90℃5 分钟，然后在 65℃10

分钟)并短时间涡旋混合，裂解生物并使该生物的核酸分子可以用于杂交。根据要检测的生物，可以如下修改所述程序，例如，包括在更高温度的热处理、酶处理(如用溶菌酶、几丁质酶或磷脂酶)、用去垢剂(如CTAB 或 SDS)处理或有机提取(如用苯酚或氯仿)。然后我根据 Hsuih 等(1996, 见上文)的方法，进行用探针(捕获探针、扩增探针和缺口探针)的杂交、亲和纯化、连接和扩增所述三联连接的扩增/缺口探针(图9B)(Hsuih 等, 1996, 见上文)。

纯化对应于扩增的 ID 探针的质谱检测寡核苷酸。扩增的探针对应于所述参考菌株中病原体特异性的 ID 序列。对于这些序列的基于质谱的鉴定，我使用生物素化的扩增产物来亲和纯化对应的质谱检测寡核苷酸(图 9C)。使扩增反应物(50 μl)溶于 10 mM EDTA，并与在 10 mM EPPS, pH 8.0/1 mM EDTA 中包含 10 ng 每种质谱检测寡核苷酸的 10 μl 溶液混合，然后在 100℃变性 2 分钟。在加入 15μl 5 M NaCl 并在 30℃温育 15 分钟后，加入 30μl 链霉抗生物素包被的顺磁珠(Promega)，并如以前所述进行亲和层析(Hsuih 等, 1996, 见上文)。所述珠用 500μl 10 mM EPPS, pH 8.0/1 mM EDTA 洗 3 次。通过在 100μl 10 mM EPPS, pH 8.0/1 mM EDTA 中加热所述溶液到 50℃(或比在 1 M NaCl 中所述检测寡核苷酸的最高 T_m 高 10℃)，回收亲和纯化的质谱检测寡核苷酸。从所述磁珠取出包含所述质谱检测寡核苷酸的上清液，而用磁铁将所述磁珠保留在管中。

构建一个病原体类群的指纹数据库：使用质谱来鉴定所选定的质谱检测寡核苷酸。制各样品并且使用仪器(PerSeptive Biosystems)和以前描述的方法(Roskey 等, Proc. Natl. Acad. Sci. USA 93: 4724-4729, 1996)，通过基质辅助激光解吸电离飞行时间质谱(延迟提取)(MALDI-TOF(DE))制备和分析样品。将亲和纯化的寡核苷酸的质量与以前确定的整个质谱检测寡核苷酸集合的元件的质量相比较。这样，鉴定了选定的质谱检测寡核苷酸，该选定的质谱检测寡核苷酸进而又指出在受测试的参考菌株中 ID 序列的身份。

在所述参考菌株中存在的 ID 序列亚组构成其基因组分布分析指纹。收集在表 7 列出的每个类群中参考菌株的指纹数据库。

鉴定在血液样品中存在的病原体。如上文针对参考菌株所述，裂解血液样品并进行指纹分析，不同之处在于所述杂交反应中包括来自表 7 所有血流病原体类群的 ID 探针。通过将所获得的指纹与那些参考菌株的指纹数据库中的指纹相比较，鉴定在血液样品中存在的病原体。

实施例 4. 使用基因组分布分析测定的法医学鉴定

法医学鉴定的概述。鉴定细胞样品的来源是现代法医分析的一个重要方面。法医学样品的遗传鉴定需要扩增常常仅以微观量可得的细胞材料中的 DNA，并将所述 DNA 与其他个体的 DNA 相比较。目前遗传鉴定的方法一般要求分析型凝胶电泳，该步骤极其消耗时间，并且在技术上对于许多法医学实验室是不合适的。本实施例提供了使用基因组分布分析进行法医学鉴定的快速、简单并且健全的方法。

本实施例概述。我使用富集的基因组差异样品，分离了可用于鉴定人类法医学样品的来源的 ID 序列集合。在本实施例中，所述富集的基因组样品是经扩增的人类基因组亚组，根据扩增过程的本质，所述人类基因组亚组包含一些可重现地从某些个体的基因组中扩增、但不从其他个体的基因组扩增的序列。这些差异扩增的序列构成基因组差异序列：它们在一个富集的基因组差异样品中存在，但不在另一个富集的基因组差异样品中存在。在来自某一个体的 DNA 中存在的这样的序列集合的亚组构成一个基因组分布分析指纹。通过将所述样品指纹与其他个体的样品指纹相比较，获得所述样品来源的身份。

在本实施例中使用的方法的概述。本实施例与前面的实施例在几个方面有所不同。通过选择性扩增人类基因组 DNA，构建用于获得人类 ID 序列集合的富集的基因组差异样品。本实施例使用 Alu-PCR 选择性扩增人类 DNA，但也可以使用其它方法进行选择性扩增，如用

于扩增根据大小分级分离的 DNA 的 AFLP 方法(Lisitsyn 等, Mol. Gen. Microbiol. Virus. 3: 26-29, 1993; Rosenberg 等, Proc. Natl. Acad. Sci. USA 91: 6113-6117, 1994)，或在实施例 5 中描述的方法。进行多次基因组扣除以产生多个人类 ID 序列家族。用对应于基因组扣除产物的检测序列构建一个检测阵列。为鉴定人类法医学样品，使用选择性扩增(在这种情况下是 Alu-PCR)来扩增样品 DNA。得到的人类基因组 DNA 在所述样品中的“代表”由标记的扩增产物组成。通过与所述检测阵列杂交，测试所述产物中特征性 ID 序列的存在。不同人类个体的基因组将产生不同的基因组分布分析指纹。

使用 Alu-PCR 选择性扩增人类 DNA。 Alu-PCR 方法扩增在 Alu 重复序列之间的 DNA，所述 Alu 重复序列频繁出现于人类基因组中(平均每几千个碱基一个)。由于 Alu 重复序列具有多态性，一些扩增的片段存在于一个人体内，而不存在于另一个人体内(Stoneking 等, Genome Res. 7: 1061-1071, 1997; Zietkiewicz 等, Proc. Natl. Acad. Sci. USA 89: 8448-8451, 1992)。

通过标准方法(Ausubel 等, 1987, 见上文)纯化用于制备基因组扣除样品的人类基因组 DNA。如以前所详述(Lincoln 等,“法医 DNA 分布分析方法,” 载于 *Methods in Molecular Biology* (Humana Press, Totowa, New Jersey) 1998)，通过应用适于该样品类型的方法，制备法医学样品以进行扩增。使用 Zietkiewicz 等(1992, 见上文)的方法，进行 Alu-PCR 反应，改动之处在于 PCR 扩增用作“+”基因组差异样品的 DNA，并且使用 5'-末端生物素化的寡核苷酸引物，对法医学样品进行 PCR 扩增。

分离 ID 序列并构建检测集合阵列。 通过基因组扣除(Straus 等, 1990, 见上文)，分离上文所述的通过富集的基因组差异序列定义的一个人类 ID 序列家族。如上所述，使用来自个体的样品或通过汇集来自几个个体的 Alu-PCR 产物，制备富集的基因组差异样品(所述样品可以根据遗传和/或地区标准分组)。对所述基因组差异序列进行克隆、测

序、并如以前所述扩增(Rosenberg 等, 1994, 见上文; Straus 等, 1990, 见上文)。为构建所述检测集合阵列，使用 Maier 等(J. Biotechnol. 35: 191-203, 1994)的基于机器人的方法，将扩增的扣除产品，即基因组差异序列在尼龙膜上排成阵列。

对法医学样品进行指纹分析。通过以前描述的方法(Lincoln, 1998, 见上文)，制备法医学样品以进行指纹分析。通过使法医学样品的生物素化 Alu-PCR 扩增产物与所述检测集合阵列杂交，获得所述法医学样品中人类 DNA 的指纹。在通常少于 1 ml 的体积中，在 65℃进行所述杂交反应(1 M NaCl/50 mM EPPS/2 mM EDTA, pH 8) 30 分钟。通过在 65℃的 2 ml 洗涤缓冲液(50 mM NaCl/50 mM EPPS/2 mM EDTA, pH 8)中五个 30 秒钟洗涤步骤(伴随振荡)，除去未结合的扩增产物。使用 Phototope-Star 检测系统(New England Biolabs)，根据厂家的建议，使所述指纹(杂交模式)显现。

实施例 5. 扫描样品中的多种人类遗传标记

现代医学遗传学和药物基因组学(pharmacogenomics)的一个重要目标是快速获得患者的基因组分布。遗传标记可以作为疾病(如乳腺癌和亨廷顿舞蹈病)的早期警报，或可以指示患者可能对哪种药疗法有利地反应。本实施例展示了在一个快速的基于杂交的测试中，使用基因组分布分析测定来调查大量人类遗传标记的基因型。

本实施例概述。在本实施例中，同时调查一个人类基因组在多个多态位点的基因型。如在前三个实施例中所述，使一个探针(在这种情况下是 SNP 探针)集合与基因组 DNA 杂交。如以前所述，所述探针集合的选择性扩增产生了该集合的一个诊断信息亚组。然后通过与检测阵列的杂交，鉴定所扩增亚组的成员。在本实施例中，与以前的实施例不同，根据在样品基因组中存在的特定 SNP 等位基因，选择性连接半边 SNP 探针，从而完成选择性扩增。使用 SNP 探针进行基因组分型的方法图解于图 10。

合成多态性探针集合和检测集合。在本实施例中，使用已知的人类 DNA 多态性设计多态性探针。当所述多态性探针退火于包含等位基因的一个版本的基因组 DNA 时，可以连接所述多态性探针，但当基因组包含所述基因的不同版本时，就不能连接所述多态性探针。等位基因特异性 SNP 探针连接的使用图解于图 10。所靶向的 DNA 多态性可以是对应于用于对人类基因组进行作图的标记的单核苷酸多态性(SNP)(如, Landegren 等, Genome Res. 8: 769-776, 1998)或对应于具有医学重要性的突变的单核苷酸多态性(如引起遗传病镰状细胞贫血的单碱基对突变)。在所述测定中也可以包括任何其它类型的核酸序列多态性(包括插入、缺失和重排)。

一旦选择了所述 DNA 多态性，则可以基本如实施例 1 中制造 ID 探针一样合成多态性探针。SNP 探针的优选设计利用 T4 DNA 连接酶的能力以鉴别在要连接的 3'末端的单碱基对错配。然而，在本实施例中，设计所述半边多态性探针，以便成对的探针在所述 DNA 多态性位点邻接。一般合成对应于每个靶 DNA 多态性的两种多态性探针：一种探针检测在所述多态性位点的一种基因型，而另一种探针检测另一种可能的基因型。对于出现几种基因型的基因座，合成另外的多态性探针。

因此，对于每个要进行基因组分型的 SNP，SNP 探针包含几种半边探针。一种半边探针(图 10 中的右半边探针)是不变的。在所述测定中也包括了左半边 SNP 探针的几种版本。每个版本在所述基因组 SNP 位点具有对应于所述等位基因的不同 3'末端核苷酸。只有在所述 3'位点与所述基因组等位基因匹配的左半边探针才被连接并随后扩增。如前面的实施例，可以通过在扩增反应中使用生物素化引物而标记扩增产物。

因为每种独特的左半边探针都具有一种独特的标记(见图 10)，所以有可能通过使所述标记的扩增 SNP 探针与包含标记集合的检测阵列杂交，检测哪些等位基因已经被连接并成功地扩增，其中所述标记集

合对应于 SNP 探针原始集合。也就是说，所述阵列中的每一种标记对应于所述原始 SNP 探针集合中其中一种左半边 SNP 探针中的标记(或其互补物)。

如实施例 1 构建所述检测阵列，不同之处在于，在这种情况下，所述阵列的元件是对应于所述多态性探针集合的标记序列。

选择性扩增人类 DNA 多态性并进行指纹分析。如实施例 4 制备包含人类 DNA 的样品。假如使用纯化的 DNA，就简单地将其点样在 0.5 M NaOH 中的尼龙滤膜上，使其风干，并用紫外光使其交联到滤膜上(使用来自 Stratagene 的 Stratalinker 仪器，按照厂家的说明书)。注意：对于法医学样品，预扩增 DNA 样品，即制备基因组代表可能是有用的。例如，可以使用实施例 4 中描述的 Alu-PCR 方法，从单个人类毛囊扩增 DNA。当使用代表作为样品测试 SNP 多态性时，设计所述 SNP 探针，使其对应于从所有样品扩增的区段中的多态性。(注意：这与前面的实施例不同，在前面的实施例中在诊断上有用的序列是差异扩增的序列，即 ID 探针)。

如实施例 1 所述(关于该实施例的 ID 探针)，使所述多态性探针集合与所述样品杂交、洗涤、连接、扩增、标记、与检测阵列杂交、并使指纹显现。与所述检测阵列的杂交模式指出了通过所述多态性探针集合的调查，在所述样品的基因组 DNA 中在每个多态性位点呈现的等位基因。

实施例 6. 扫描脑脊液样品中的大量病毒

本实施例概述。中枢神经系统(CNS)的感染被认为是医疗急症。快速诊断传染因子对于最佳的治疗效果是至关重要的。诊断病毒感染尤其存在问题，并且常常是昂贵的。本实施例描述的方法可以用于同时测试脑脊液(CSF)样品中各种类型病毒的存在。通过用 ID 探针集合进行液相杂交捕获，然后扩增样品选定的 ID 探针，选定在 CSF 样品中的病毒特异性 ID 序列。使用所扩增的 ID 探针探测检测集合阵列，

以确定存在哪些病毒(假如存在的话)。本实施例描述了针对 CSF 中的病毒的测试，但采用合适的样品制备，可以对其它类型样品进行类似测试，所述样品包括血液样品和固体组织样品。

组装病毒特异性 ID 序列、探针和引物。选择对于表 8 列出的病毒表中每个病毒类群特异性的类群特异性序列。在某些情况下，文献中已经描述了病毒特异性 ID 序列。在其它情况下，在将公共数据库中的病毒基因组序列与该数据库中的其他病毒相比较，选定序列。使用标准方法进行序列比较(Ausubel 等, 1987, 见上文)。选择至少 30 个碱基长的病毒特异性序列，并如实施例 3 (血流病原体测定)所述，如图 9A-9C 所示，合成对应的 ID 探针集合和引物集合。然而，我合成了与所述缺口探针互补的更长的(约 20 个碱基)检测集合寡核苷酸，而不是图 9C 所示的小质谱检测寡核苷酸。如实施例 2 所述，通过光刻法构建检测集合阵列。如实施例 3 所述合成和使用阳性对照探针和阴性对照探针。

表 8. 引起 CNS 感染的病毒

柯萨奇病毒 A	柯萨奇病毒 B
单纯疱疹病毒	披膜病毒
圣•路易脑炎病毒	麻疹病毒
EB 病毒	肝炎
粘液病毒	副粘病毒
JC 病毒	腮腺炎病毒
艾可病毒	马脑炎病毒
布尼亚病毒	淋巴细胞性脉络丛脑膜炎病毒
巨细胞病毒	狂犬病病毒
水痘-带状疱疹病毒	BK 病毒
HIV	

扫描样品寻找所述病毒组的成员。如实施例 3 所述制备 CSF 样品、与探针集合杂交、通过磁力分离纯化靶序列、连接所选定的探针、以及扩增。然后如实施例 4 所述，使所述生物素化的扩增产物与所述病毒检测集合阵列杂交并使其显现。

其它实施方案包括在下面的权利要求书中。

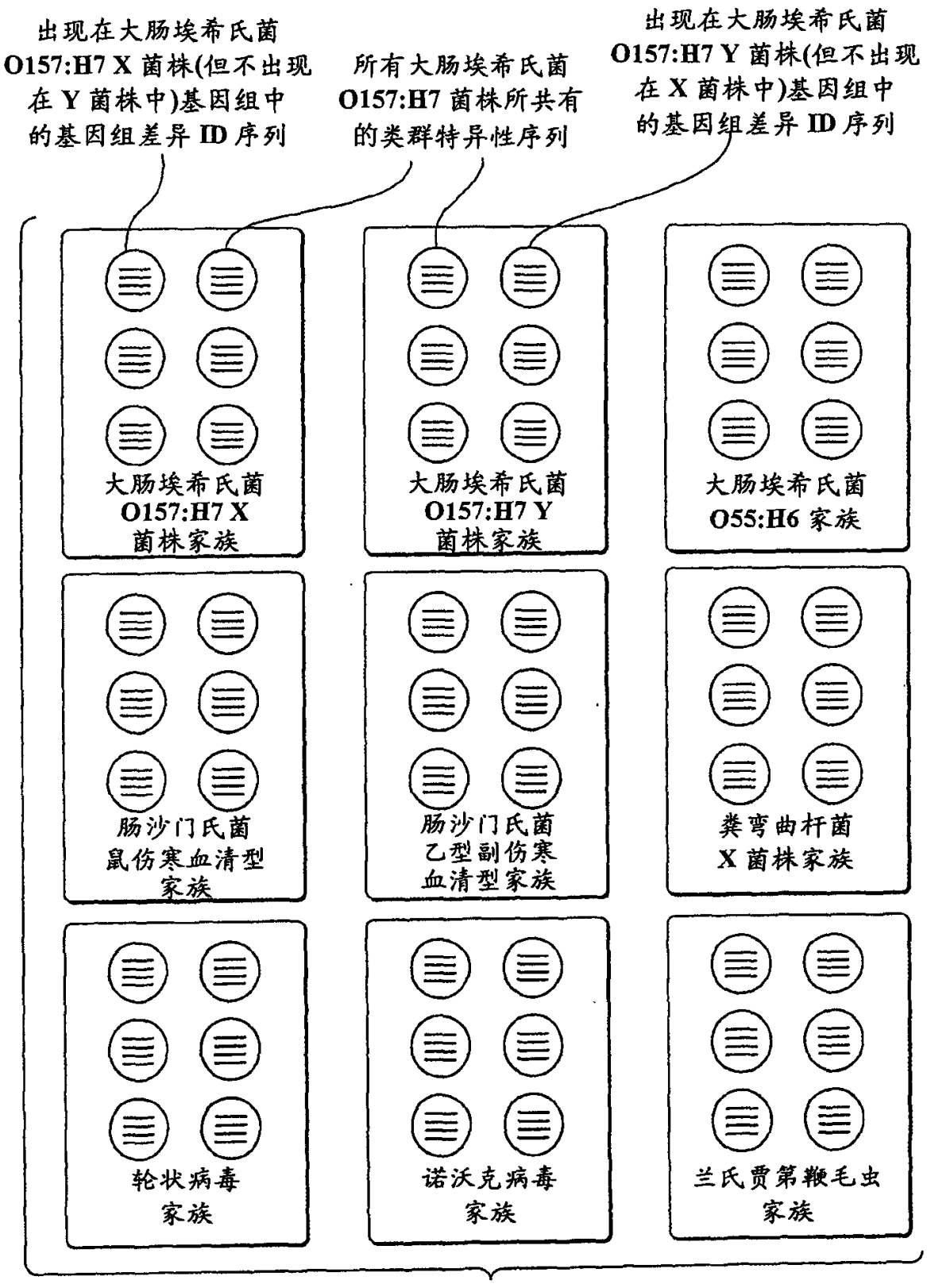
最小基因组起源为 9 的 ID 序列集合
出现在大肠埃希氏菌 O157:H7 X 菌株(但不出现在 Y 菌株中)基因组中的基因组差异 ID 序列
所有大肠埃希氏菌 O157:H7 菌株所共有的类群特异性序列
出现在大肠埃希氏菌 O157:H7 Y 菌株(但不出现在 X 菌株中)基因组中的基因组差异 ID 序列
大肠埃希氏菌 O157:H7 X 菌株家族
大肠埃希氏菌 O157:H7 Y 菌株家族
大肠埃希氏菌 O55:H6 家族
肠沙门氏菌鼠伤寒血清型家族
肠沙门氏菌乙型副伤寒血清型家族
粪弯曲杆菌 X 菌株家族
轮状病毒家族
诺沃克病毒家族
兰氏贾第鞭毛虫家族

图 1

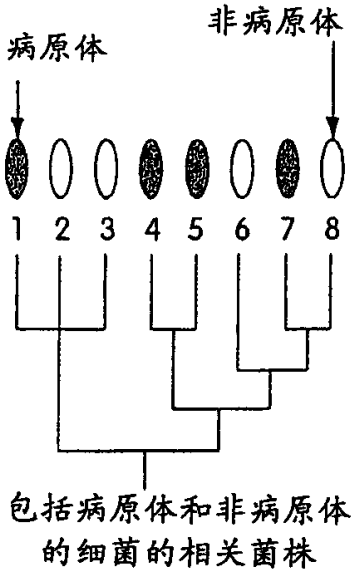
病原体
非病原体
1 2 3 4 5 6 7 8
包括病原体和非病原体
的细菌的相关菌株

图 2A

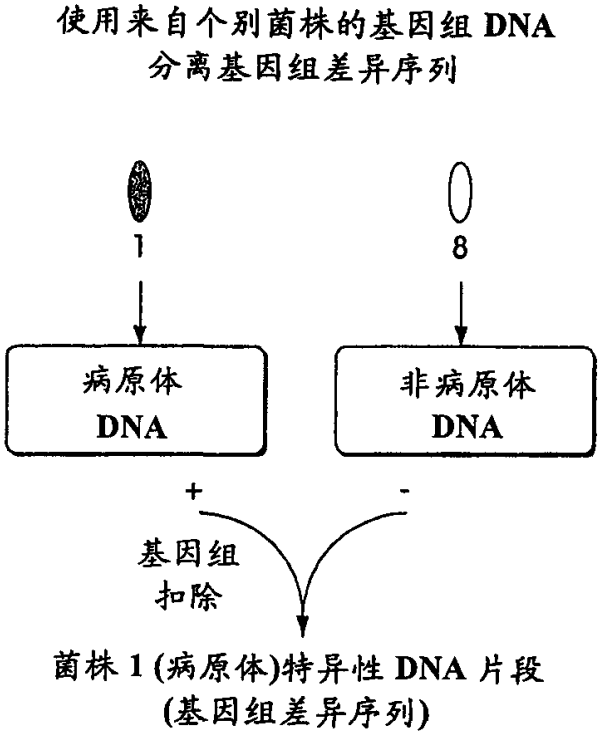
使用来自个别菌株的基因组 DNA
分离基因组差异序列
1
8
病原体
DNA
非病原体
DNA
+
-
基因组
扣除
菌株 1 (病原体)特异性 DNA 片段
(基因组差异序列)

图 2B

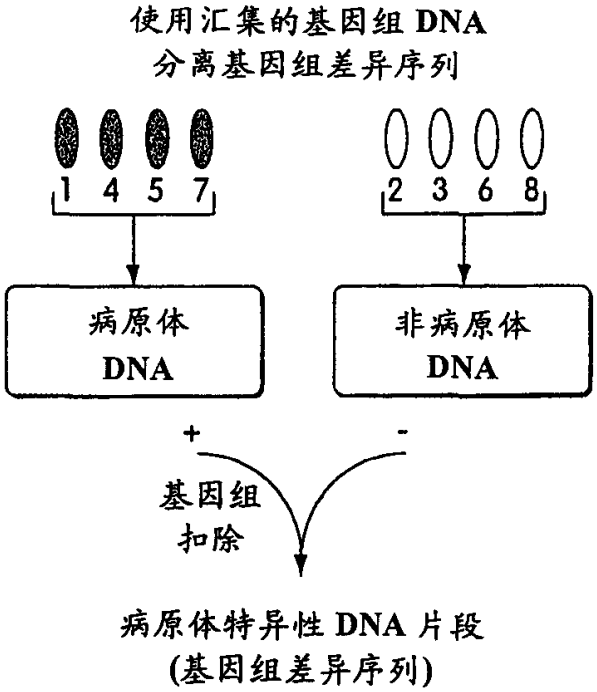
使用汇集的基因组 DNA
分离基因组差异序列
1 4 5 7
2 3 6 8
病原体
DNA
非病原体
DNA
+
-
基因组
扣除
病原体特异性 DNA 片段
(基因组差异序列)

图 2C

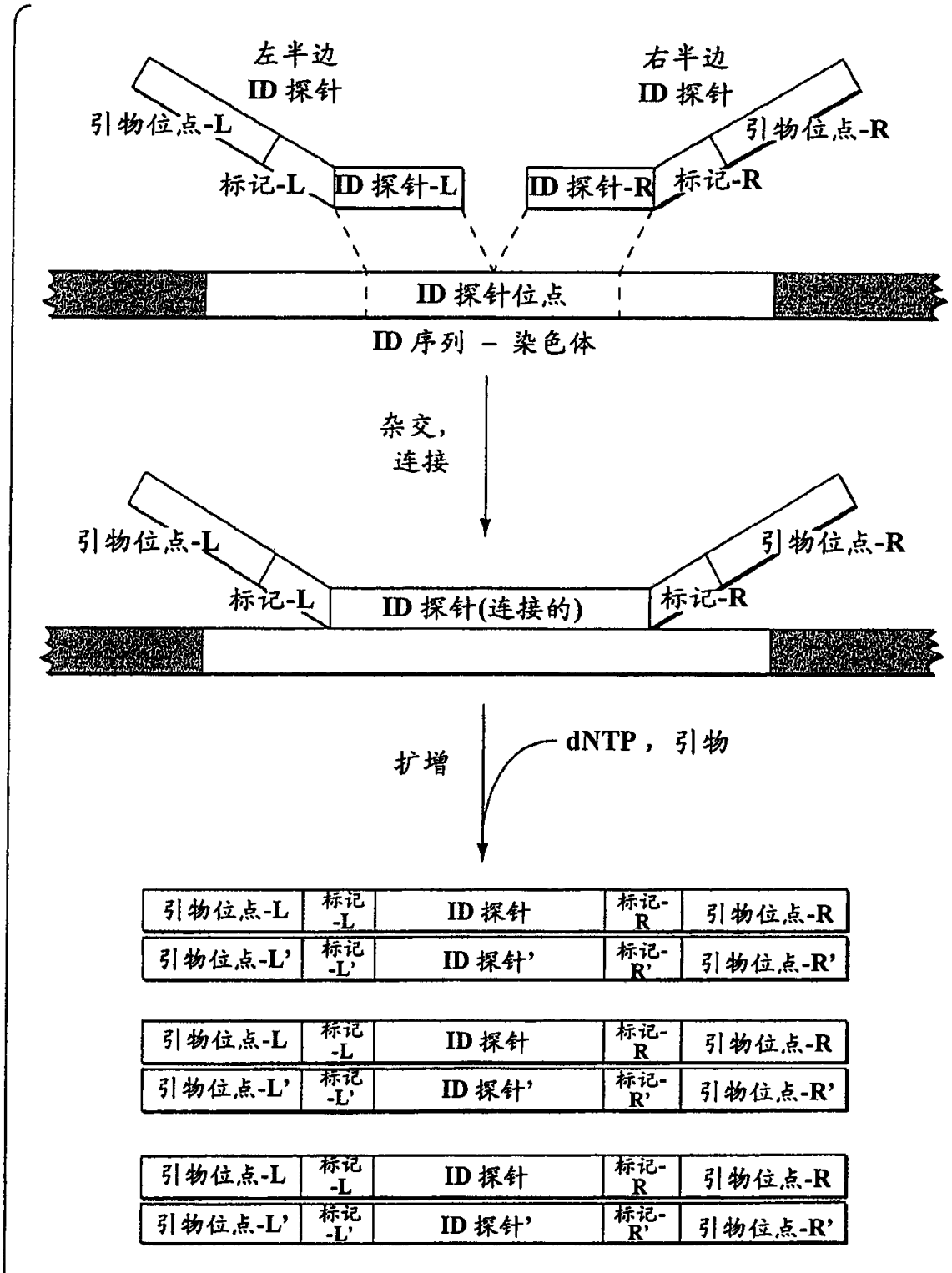
左半边
ID 探针
右半边
ID 探针
引物位点-L
引物位点-R
标记-L
ID 探针-L
ID 探针-R
标记-R
ID 探针位点
ID 序列 – 染色体
杂交，
连接
引物位点-L
引物位点-R
标记-L
ID 探针(连接的)
标记-R
扩增
dNTP，引物
引物位点-L
标记-L
ID 探针
标记-R
引物位点-R
引物位点-L'
标记-L'
ID 探针'
标记-R'
引物位点-R'

图 3

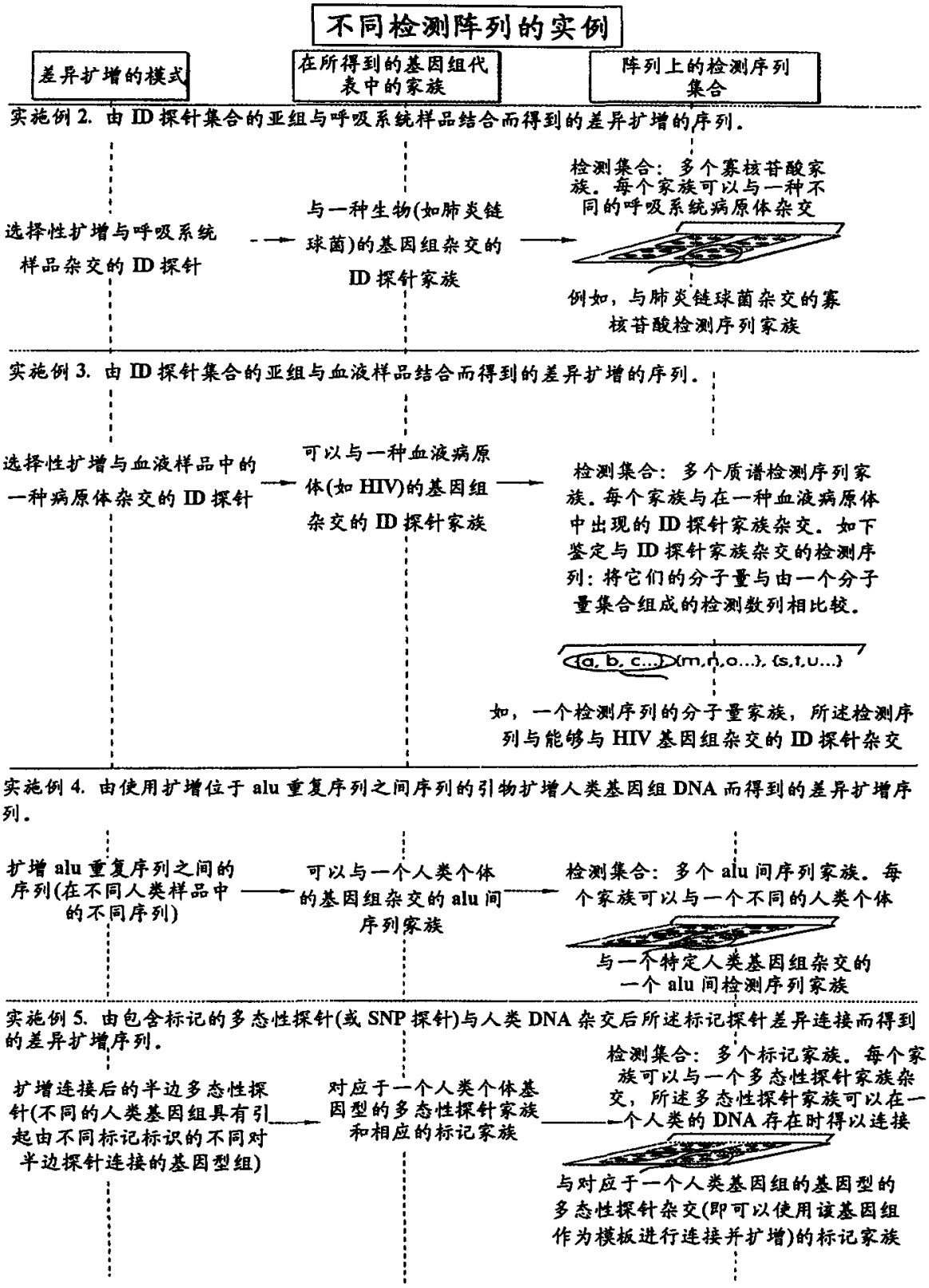
不同检测阵列的实例
差异扩增的模式
在所得到的基因组代表中的家族
阵列上的检测序列集合
实施例 2. 由 ID 探针集合的亚组与呼吸系统样品结合而得到的差异扩增的序列。
选择性扩增与呼吸系统样品杂交的 ID 探针
与一种生物(如肺炎链球菌)的基因组杂交的 ID 探针家族
检测集合：多个寡核苷酸家族。每个家族可以与一种不同的呼吸系统病原体杂交
例如，与肺炎链球菌杂交的寡核苷酸检测序列家族
实施例 3. 由 ID 探针集合的亚组与血液样品结合而得到的差异扩增的序列。
选择性扩增与血液样品中的一种病原体杂交的 ID 探针
可以与一种血液病原体(如 HIV)的基因组杂交的 ID 探针家族
检测集合：多个质谱检测序列家族。每个家族与在一种血液病原体中出现的 ID 探针家族杂交。如下鉴定与 ID 探针家族杂交的检测序列：将它们的分子量与由一个分子量集合组成的检测数列相比较。
{a, b, c...} {m,n,o...}, {s,t,u...}
如，一个检测序列的分子量家族，所述检测序列与能够与 HIV 基因组杂交的 ID 探针杂交
实施例 4. 由使用扩增位于 alu 重复序列之间序列的引物扩增人类基因组 DNA 而得到的差异扩增序列。
扩增 alu 重复序列之间的序列(在不同人类样品中的不同序列)
可以与一个人类个体的基因组杂交的 alu 间序列家族
检测集合：多个 alu 间序列家族。每个家族可以与一个不同的人类个体
与一个特定人类基因组杂交的一个 alu 间检测序列家族
实施例 5. 由包含标记的多态性探针(或 SNP 探针)与人类 DNA 杂交后所述标记探针差异连接而得到的差异扩增序列。
扩增连接后的半边多态性探针(不同的人类基因组具有引起由不同标记标识的不同对半边探针连接的基因型组)
对应于一个人类个体基因型的多态性探针家族和相应的标记家族
检测集合：多个标记家族。每个家族可以与一个多态性探针家族杂交，所述多态性探针家族可以在一个人类的 DNA 存在时得以连接
与对应于一个人类基因组的基因型的多态性探针杂交(即可以使用该基因组作为模板进行连接并扩增)的标记家族

图 4

扫描临床样品中的多种病原体
使用 ID 探针的样品选择的基因组分布分析

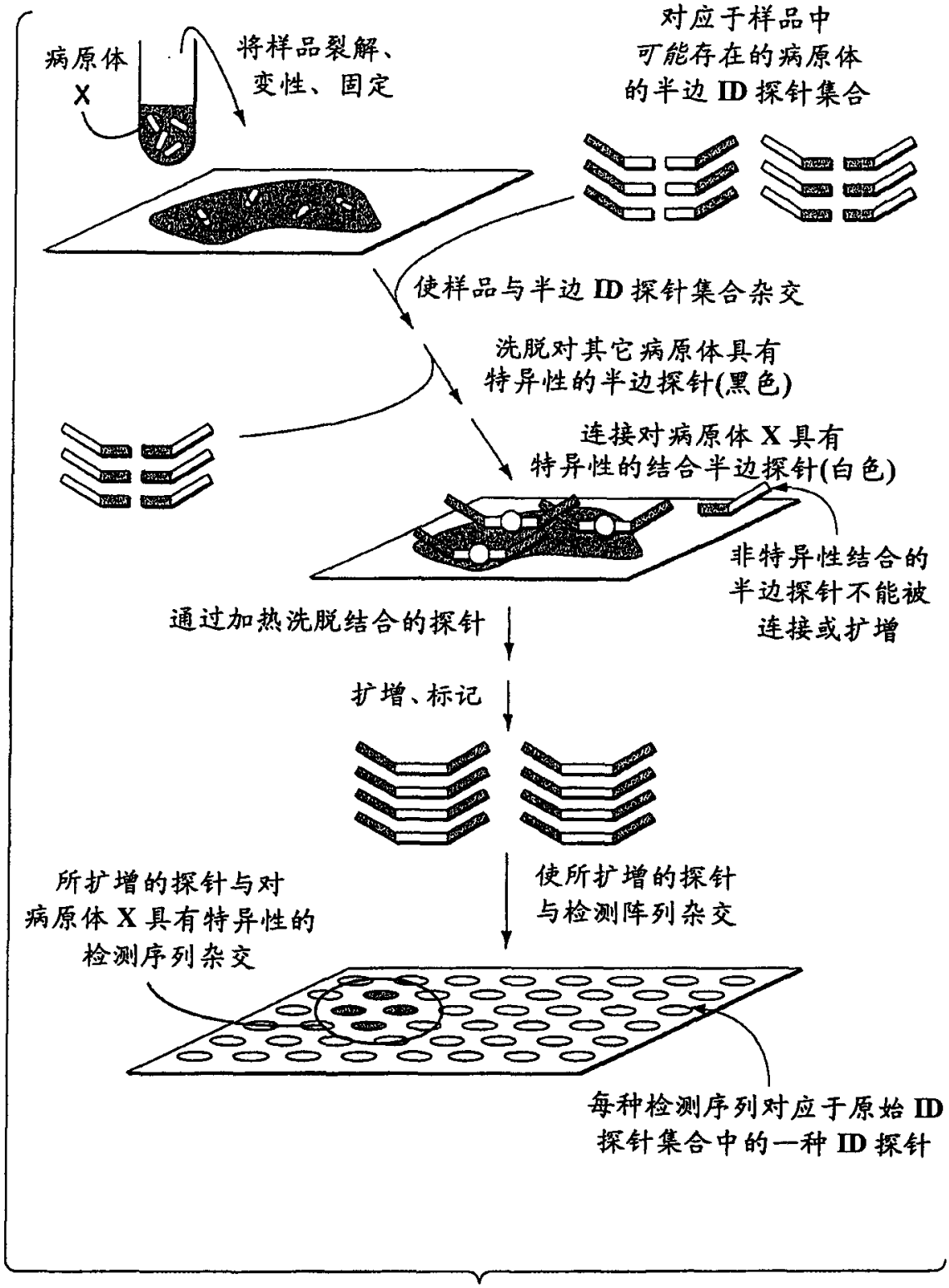

图 5

肠沙门氏菌亚种 I

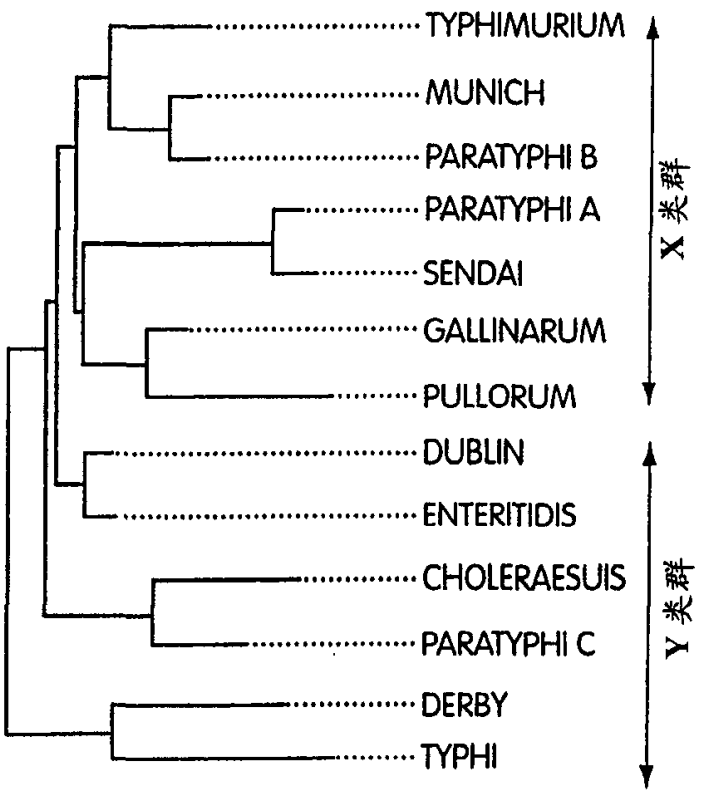
TYPHIMURIUM
MUNICH
PARATYPHI B
PARATYPHI A
SENDAI
GALLINARUM
PULLORUM
DUBLIN
ENTERITIDIS
CHOLERAESUIS
PARATYPHI C
DERBY
TYPHI
X 类群
Y 类群

图 6

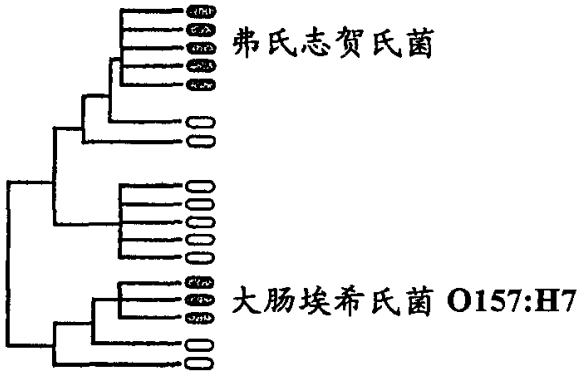
弗氏志贺氏菌
大肠埃希氏菌 O157:H7

图 7A

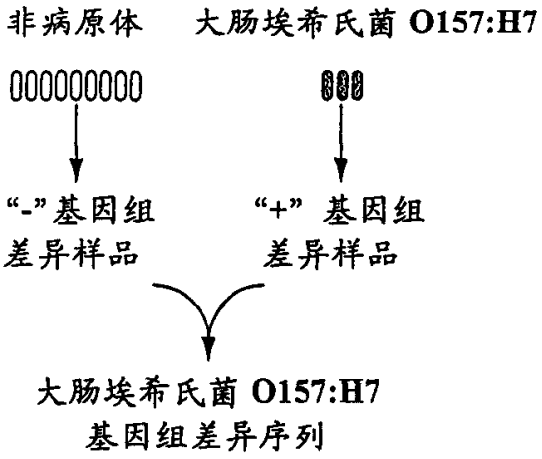
非病原体
大肠埃希氏菌 O157:H7
"-"基因组
差异样品
"+" 基因组
差异样品
大肠埃希氏菌 O157:H7
基因组差异序列

图 7B

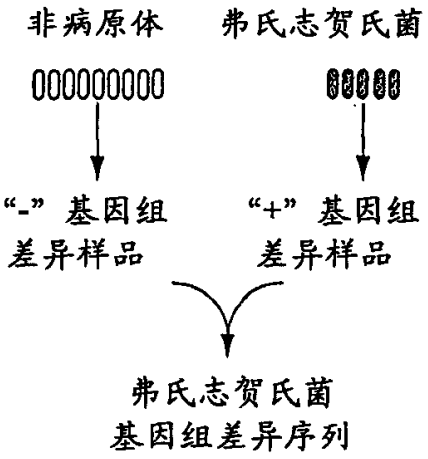
非病原体
弗氏志贺氏菌
"-" 基因组
差异样品
"+" 基因组
差异样品
弗氏志贺氏菌
基因组差异序列

图 7C

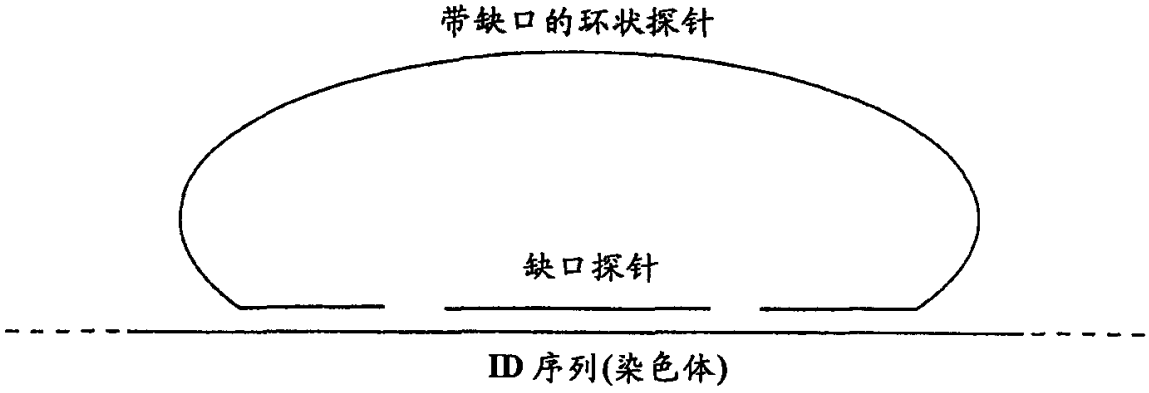
带缺口的环状探针
缺口探针
ID 序列(染色体)

图 8A

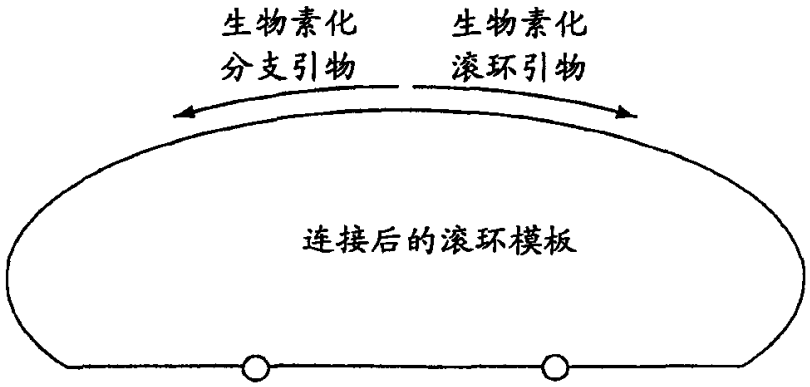
生物素化
分支引物
生物素化
滚环引物
连接后的滚环模板

图 8B

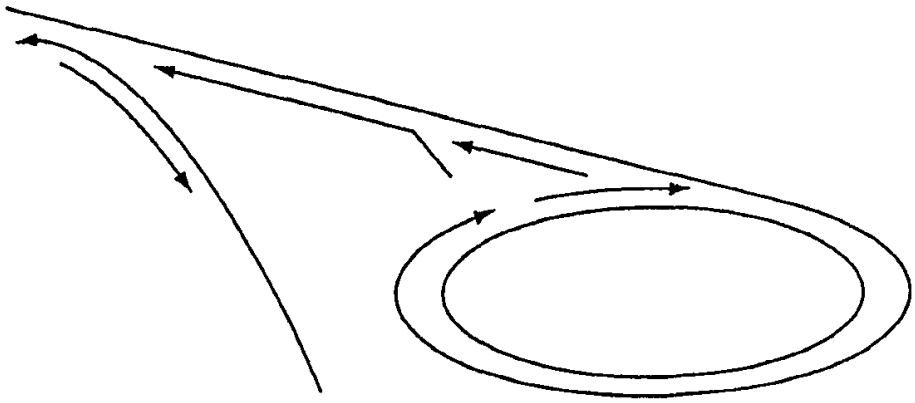
高分支滚环扩增

图 8C

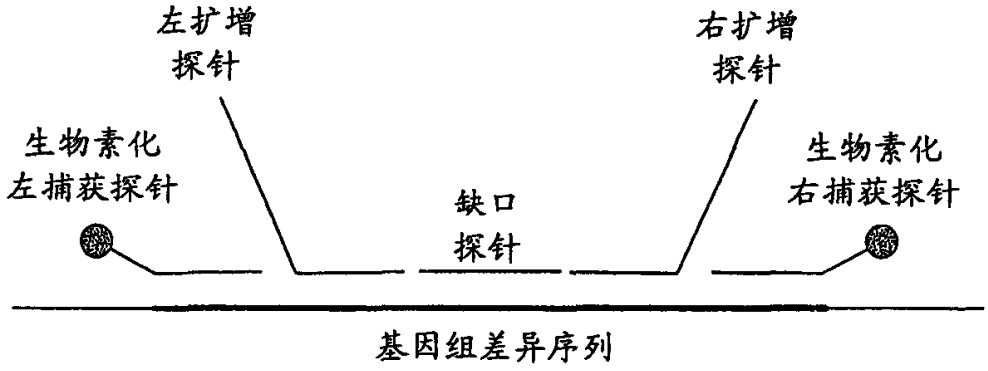
左扩增
探针
右扩增
探针
生物素化
左捕获探针
缺口
探针
生物素化
右捕获探针
基因组差异序列

图 9A

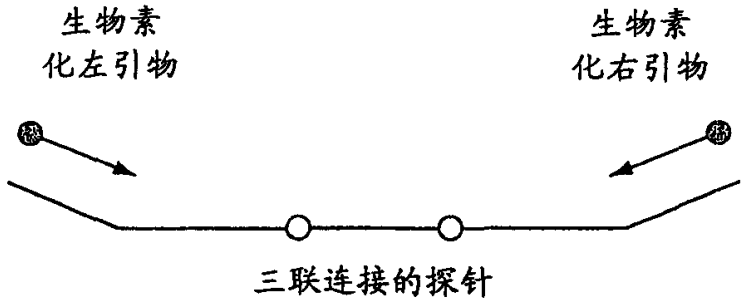
生物素
化左引物
生物素
化右引物
三联连接的探针

图 9B

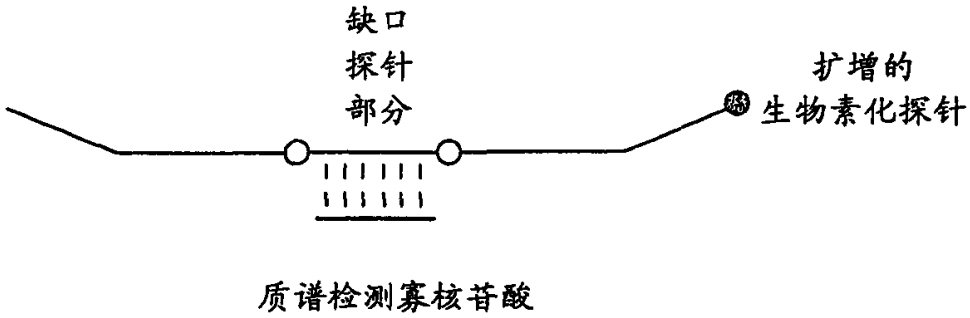
缺口
探针
部分
扩增的
生物素化探针
质谱检测寡核苷酸

图 9C

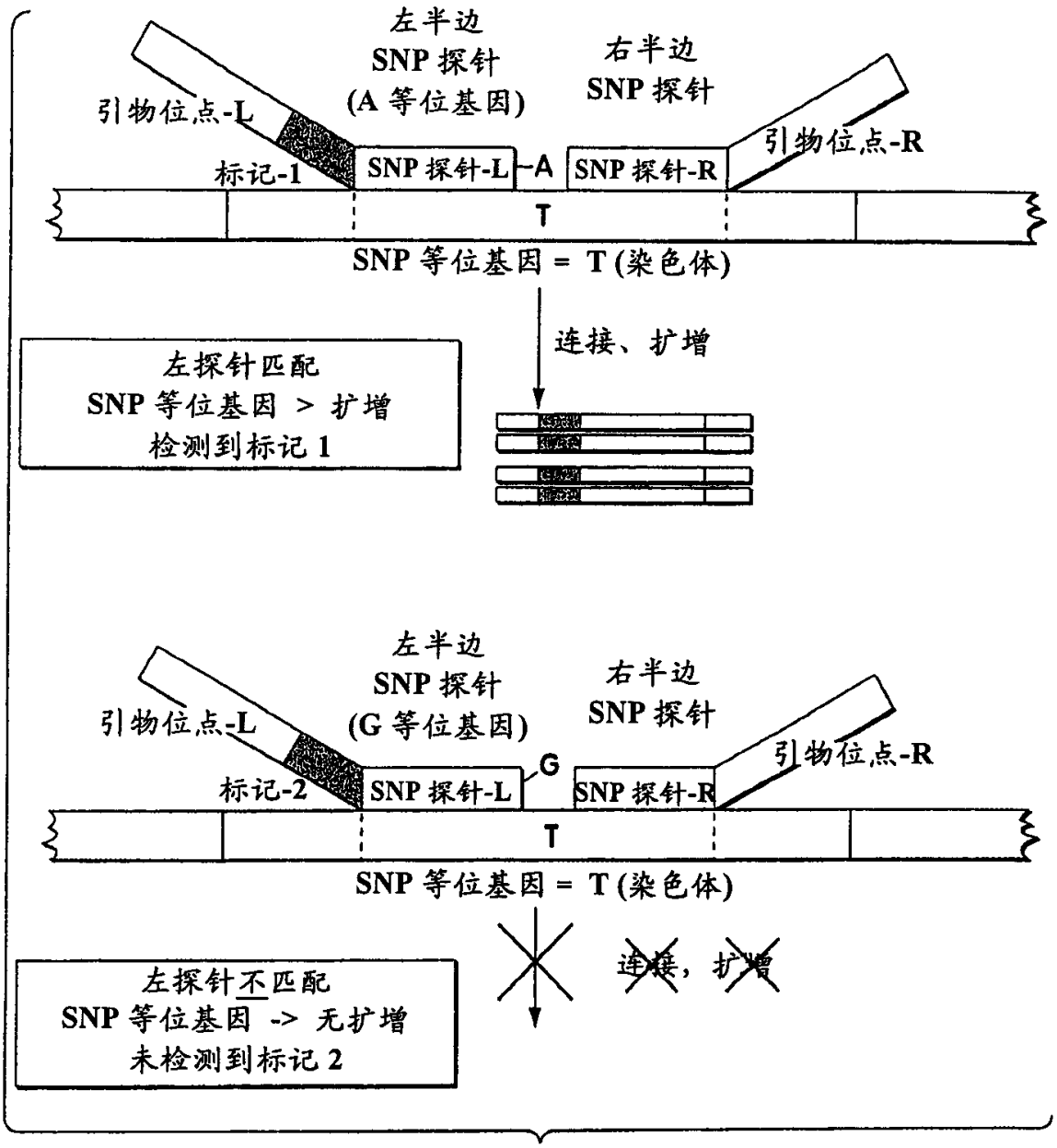

图 10

SNP 探针杂交-选择；连接和扩增依赖于在 SNP 位点匹配

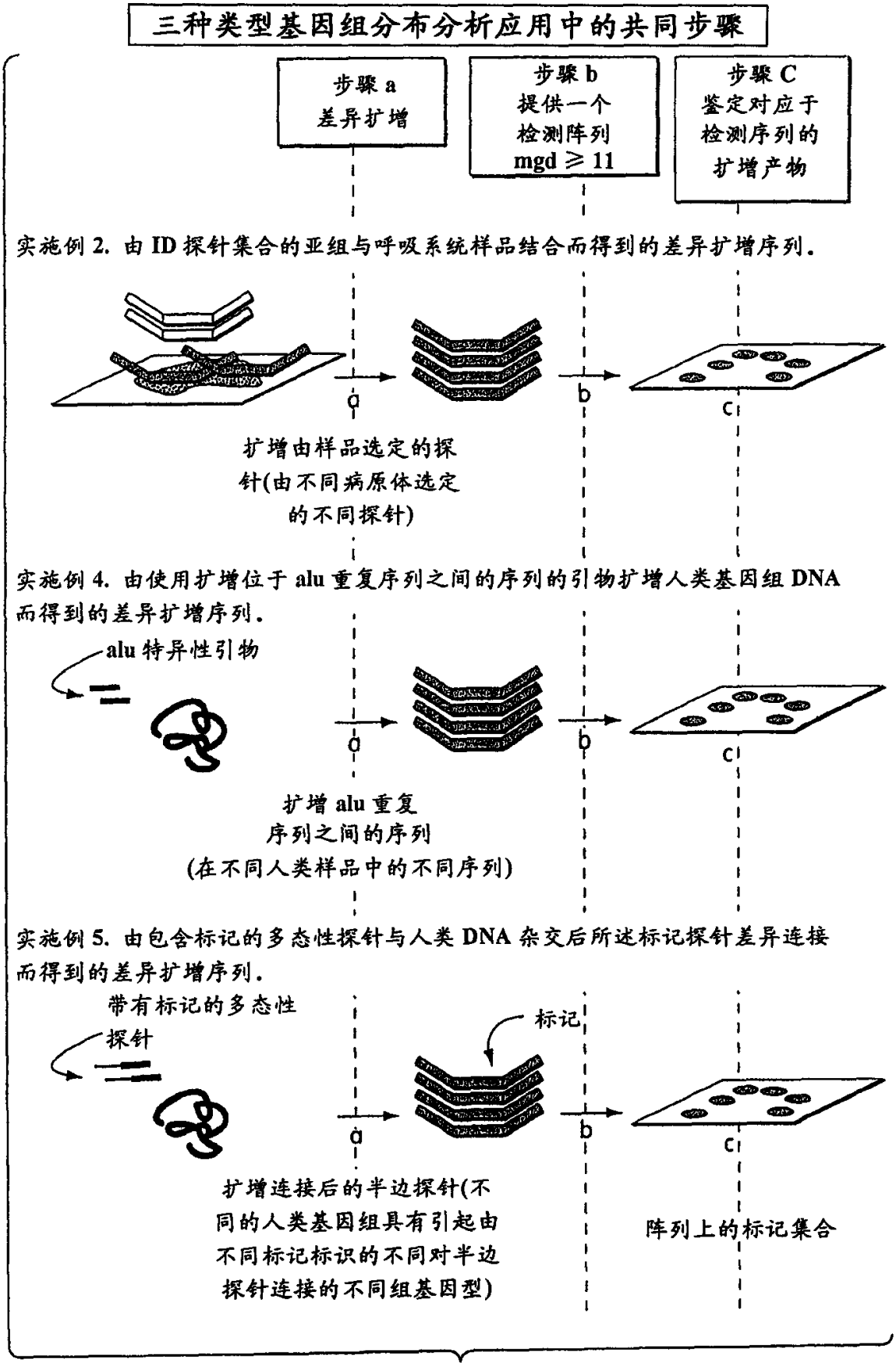
三种类型基因组分布分析应用中的共同步骤
步骤 a
差异扩增
步骤 b
提供一个
检测阵列
mgd ≥ 11
步骤 C
鉴定对应于
检测序列的
扩增产物
实施例 2. 由 ID 探针集合的亚组与呼吸系统样品结合而得到的差异扩增序列.
a
b
c
扩增由样品选定的探
针(由不同病原体选定
的不同探针)
实施例 4. 由使用扩增位于 alu 重复序列之间的序列的引物扩增人类基因组 DNA
而得到的差异扩增序列.
alu 特异性引物
a
b
c
扩增 alu 重复
序列之间的序列
(在不同人类样品中的不同序列)
实施例 5. 由包含标记的多态性探针与人类 DNA 杂交后所述标记探针差异连接
而得到的差异扩增序列.
带有标记的多态性
探针
标记
a
b
c
扩增连接后的半边探针(不
同的人类基因组具有引起由
不同标记标识的不同对半边
探针连接的不同组基因型)
阵列上的标记集合

图 11

专利名称(译)	基因组分布分析:一种检测复杂生物样品中多种类型生物的存在的快速方法		
公开(公告)号	CN1370242A	公开(公告)日	2002-09-18
申请号	CN00811616.4	申请日	2000-06-15
[标]申请(专利权)人(译)	基因描绘系统有限公司		
申请(专利权)人(译)	基因描绘系统有限公司		
当前申请(专利权)人(译)	基因描绘系统有限公司		
[标]发明人	D斯特劳斯		
发明人	D·斯特劳斯		
IPC分类号	G01N33/53 C12N15/09 C12Q1/68 C12R1/01 G01N33/566 C12Q1/70 C07H21/04 C12P19/34		
CPC分类号	C12Q1/6827 Y02A50/52		
代理人(译)	姜建成		
优先权	09/333110 1999-06-15 US		
外部链接	Espacenet SIPO		

摘要(译)

本发明提供了称为基因组分布分析的一种方法,该方法同时扫描复杂的生物样品中多种不同类型生物特征性的核酸序列(包括基因组差异序列、类群特异性序列和DNA多态性)的存在。本发明还包括用于本发明的方法中的探针、检测集合和相关分子。