



(12) 发明专利

(10) 授权公告号 CN 101292044 B

(45) 授权公告日 2012. 11. 07

(21) 申请号 200680036976. 8

(22) 申请日 2006. 08. 08

(30) 优先权数据

11/202, 023 2005. 08. 12 US

(85) PCT申请进入国家阶段日

2008. 04. 03

(86) PCT申请的申请数据

PCT/SG2006/000224 2006. 08. 08

(87) PCT申请的公布数据

W02007/021250 EN 2007. 02. 22

(73) 专利权人 新加坡科技研究局

地址 新加坡新加坡市

(72) 发明人 C·W·C·黄 宋永健 李华星

L·D·米勒

(74) 专利代理机构 北京润平知识产权代理有限公司

公司 11283

代理人 周建秋 王凤桐

(51) Int. Cl.

C12Q 1/68 (2006. 01)

C12M 1/34 (2006. 01)

G06F 19/20 (2011. 01)

(56) 对比文件

US 20040259124 A1, 2004. 12. 23, 全文.

US 20030236633 A1, 2003. 12. 25, 全文.

SUNG, W-K. et al.. Fast and accurate probe selection algorithm for large genomes. 《Proceedings IEEE Computer Society Bioinformatics Conference》. 2003, 第 2 卷 65-74.

审查员 李娟娟

权利要求书 6 页 说明书 55 页

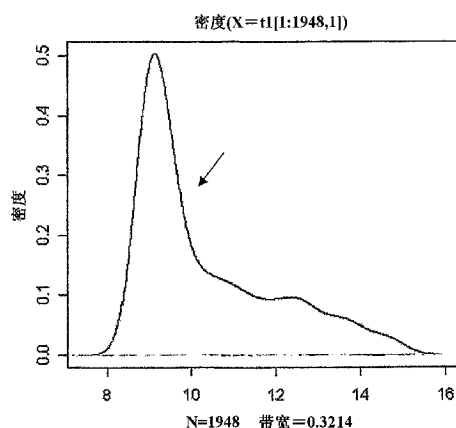
序列表 5 页 附图 19 页

(54) 发明名称

寡核苷酸设计和 / 或核酸检测的方法和 / 或装置

(57) 摘要

本发明提供了设计至少一种用于核酸检测的寡核苷酸的方法, 该方法包括任意顺序的以下步骤: (I) 确定和 / 或选择至少一种靶核酸的待扩增区域, 该区域的扩增效率 (AE) 高于平均 AE; 和 (II) 设计至少一种能够与所选区域进行杂交的寡核苷酸。本发明还提供了检测至少一种靶核酸的方法, 该方法包括以下步骤: (i) 提供至少一种生物样品; (ii) 对该生物样品中所包含的核酸进行扩增; (iii) 提供能够与至少一种假定存在于该生物样品中的靶核酸进行杂交的至少一种寡核苷酸; 和 (iv) 使所述寡核苷酸与扩增的核酸进行接触并检测与所述靶核酸杂交的寡核苷酸。尤其是, 本方法用于检测在至少一种人生物样品中存在的至少一种病原体, 例如病毒。探针可以被放置的支持物上, 例如微阵列上。



1. 设计至少一种用于核酸检测的寡核苷酸的方法,该方法包括以下步骤:

(I) 对靶核酸 v_a 的每个位置 i 计算扩增效率得分:

$$AES_i = \sum_{j=i-Z}^i \left\{ P^f(j) \times \sum_{k=\max(i+1, j+500)}^{j+Z} P^r(k) \right\}$$

$$\text{其中 } \sum_{k=\max(i+1, j+500)}^{j+Z} P^r(k) = P^r(i+1) + P^r(i+2) + \dots + P^r(j+Z);$$

$P^f(i)$ 和 $P^r(i)$ 为随机引物 r_i 作为正向引物和反向引物分别与 v_a 的位置 i 结合的概率,并且 $Z \leq 10000\text{bp}$ 为 v_a 的想要扩增的区域;

(II) 确定和 / 或选择至少一种靶核酸的至少一个待扩增区域,该区域的扩增效率高于平均扩增效率;和

(III) 设计至少一种能够与所确定和 / 或选择的区域进行杂交的寡核苷酸。

2. 根据权利要求 1 所述的方法,其中,至少一种能够与所述选择的区域进行杂交的寡核苷酸是根据以下标准中的至少一种进行选择和设计的:

(a) 选择 CG 含量为 40% 至 60% 的寡核苷酸;

(b) 选择具有根据最邻近模型计算的最高自由能的寡核苷酸;

(c) 假定寡核苷酸 s_a 和寡核苷酸 s_b 分别为靶核酸 v_a 和 v_b 的子串,则根据 s_a 与来自靶核酸 v_b 的长度为 m 的任意子串 s_b 之间的汉明距离和 / 或根据 s_a 与寡核苷酸 s_b 的最长公共子串来选择 s_a ;

(d) 就靶核酸 v_a 特异性的长度为 m 的任何寡核苷酸 s_a 而言,若寡核苷酸 s_a 与异于靶核酸的核酸的任何区域无任何标的,则选择该寡核苷酸 s_a ;若长度为 m 的寡核苷酸 s_a 与异于靶核酸的核酸有标的,则选择长度为 m 的具有最小的最大比对长度和 / 或具有最低标的数的寡核苷酸 s_a ;和

(e) 如果预测 p_i 杂交到所扩增的靶核酸的位置 i 上,则选择在靶核酸的位置 i 上的寡核苷酸 p_i 。

3. 根据权利要求 2 所述的方法,其中,所述寡核苷酸是根据标准 (e) 选择的。

4. 根据权利要求 2 所述的方法,其中,根据标准 (e),若 $P(p_i | v_a) > \lambda$, 其中 λ 为 0.5 并且 $P(p_i | v_a)$ 为 p_i 杂交到靶核酸 v_a 的位置 i 上的概率,则选择在靶核酸 v_a 的位置 i 上的寡核苷酸 p_i , 其中, $P(p_i | v_a) \approx P(X \leq x_i) = \frac{c_i}{k}$, 其中 X 为代表 v_a 的全部寡核苷酸的扩增效率得分值的随机变量, k 为 v_a 中的寡核苷酸的数目, 并且 c_i 为扩增效率得分值 $\leq x_i$ 的寡核苷酸的数目。

5. 根据权利要求 2 所述的方法,其中,根据标准 (e),若 $P(p_i | v_a) > \lambda$, 其中 λ 为 0.8 并且 $P(p_i | v_a)$ 为 p_i 杂交到靶核酸 v_a 的位置 i 上的概率,则选择在靶核酸 v_a 的位置 i 上的寡核苷酸 p_i , 其中, $P(p_i | v_a) \approx P(X \leq x_i) = \frac{c_i}{k}$, 其中 X 为代表 v_a 的全部寡核苷酸的扩增效率得分值的随机变量, k 为 v_a 中的寡核苷酸的数目, 并且 c_i 为扩增效率得分值 $\leq x_i$ 的寡核苷酸的数目。

6. 根据权利要求 1 所述的方法,其中,该方法还包括制备所选择和 / 或设计的寡核苷酸

的步骤。

7. 根据权利要求 1 所述的方法,其中,所述至少一种寡核苷酸为至少一种寡核苷酸探针或引物。

8. 根据权利要求 1 所述的方法,其中,所述扩增使用至少一种正向随机引物和 / 或至少一种反向随机引物。

9. 根据权利要求 1 所述的方法,其中,所述扩增为逆转录 - 聚合酶链式反应。

10. 根据权利要求 2 所述的方法,其中,所述寡核苷酸是根据标准 (a) 到 (d) 中的至少一个和标准 (e) 选择和 / 或设计的。

11. 根据权利要求 8 所述的方法,其中,该正向随机引物和反向随机引物中的每种引物在 5' -3' 方向上包含固定的引物头部和可变的引物尾部,并且其中至少该可变的尾部与靶核酸 v_a 的一部分进行杂交。

12. 检测至少一种靶核酸的方法,该方法不以疾病诊断和 / 或治疗为目的,且该方法包括以下步骤:

(i) 提供至少一种生物样品;

(ii) 对该生物样品中包含的核酸进行扩增;

(iii) 提供能够与至少一种假定存在于该生物样品中的靶核酸进行杂交的至少一种寡核苷酸,其中,该寡核苷酸是根据权利要求 1 至 11 中的任意一项所述的方法进行设计和 / 或制备的;和

(iv) 使该寡核苷酸与扩增的核酸进行接触和 / 或检测与靶核酸杂交的寡核苷酸。

13. 根据权利要求 12 所述的方法,其中,相对于生物样品的核酸,待检测的靶核酸为外源的核酸。

14. 根据前述权利要求 12 所述的方法,其中,所述待检测的靶核酸至少为病原体基因组或基因组片段。

15. 根据权利要求 14 所述的方法,其中,病原体核酸至少为来自病毒、寄生虫或细菌的核酸或核酸片段。

16. 根据权利要求 12 所述的方法,其中,所述生物样品从人类得到,并且假定存在于所述生物样品中的靶核酸为非人源的。

17. 根据权利要求 12 所述的方法,其中,所述探针被放置在不溶性支持物上。

18. 根据权利要求 17 所述的方法,其中,所述不溶性支持物是微阵列。

19. 根据权利要求 12 所述的方法,其中,检测步骤 (iv) 包括通过计算加权相对熵得分的分布而对靶核酸 v_a 评估每个标签探针组中的探针信号强度:

$$WKL(P_a | \bar{P}_a) = \sum_{j=0}^{k-1} \frac{Q_a(j) \log\left(\frac{Q_a(j)}{Q_a(j)}\right)}{\sqrt{Q_a(j)[1-Q_a(j)]}}$$

其中, $Q_a(j)$ 是在组 b_j 中所找到的 P_a 内的探针的信号强度的累积分布函数; $Q_a(j)$ 是在组 b_j 中所找到的 $\bar{P}_a$ 内的探针的信号强度的累积分布函数, P_a 为病毒 v_a 的探针组,

$$\bar{P}_a = P - P_a$$

20. 根据权利要求 19 所述的方法, 其中, 在检测步骤 (iv) 中, 加权相对熵值 ≥ 1.0 时, 表明所述生物样品内存在至少一种靶核酸。

21. 根据权利要求 19 所述的方法, 其中, 在检测步骤 (iv) 中, 加权相对熵值 ≥ 5.0 时, 表明所述生物样品内存在至少一种靶核酸。

22. 根据权利要求 19 所述的方法, 其中, 每个标签探针组具有正态分布的信号强度和 / 或小于 5 的加权相对熵得分, 则表明不存在靶核酸 v_a ; 并且,

其中, 每个标签探针组具有正偏的信号强度分布和 / 或大于 5 的加权相对熵得分, 则表明存在至少一种靶核酸 v_a 。

23. 根据权利要求 19 所述的方法, 该方法还包括对加权相对熵得分的分布进行安德森 - 塔林检验, 其中, $P > 0.05$ 的结果表示不存在靶核酸 v_a ; 并且,

其中, $P < 0.05$ 的结果表示存在靶核酸 v_a 。

24. 根据权利要求 23 所述的方法, 其中, $P < 0.05$ 表明加权相对熵得分的分布为非正态且 $P > 0.05$ 表明加权相对熵得分的分布为正态。

25. 根据权利要求 24 所述的方法, 其中, 如果加权相对熵得分的分布为非正态, 确定具有最高的加权相对熵得分的标签探针组为所述生物样品中出现的靶。

26. 根据权利要求 25 所述的方法, 该方法进一步包括: 从所述加权相对熵得分中去除所述最高的加权相对熵得分, 并且对剩余的加权相对熵得分的分布重复进行安德森 - 塔林检验, 以确定剩余的加权相对熵得分的分布是否为正态。

27. 根据权利要求 26 所述的方法, 其中, 如果所述剩余的加权相对熵得分的分布为非正态, 确定具有下一个最高的所述加权相对熵得分的标签探针组为所述生物样品中出现的下一个靶。

28. 根据权利要求 27 所述的方法, 其中, 具有下一个最高的所述加权相对熵得分的靶核酸分子为共感染病原体的指示。

29. 根据权利要求 27 所述的方法, 该方法包括: 从所述加权相对熵得分中重复去除所述下一个最高的加权相对熵得分, 并且重复进行安德森 - 塔林检验直至加权相对熵得分的分布变为正态的步骤, 从而检测任何其它靶核酸分子和 / 或共感染病原体的存在。

30. 用于设计用于核酸检测的至少一种寡核苷酸的装置, 该装置包括至少一个计算处理单元, 该单元被配置为:

(I) 对靶核酸 v_a 的每个位置 i 计算扩增效率得分:

$$AES_i = \sum_{j=i-Z}^i \left\{ P^f(j) \times \sum_{k=\max(i+1, j+500)}^{j+Z} P^r(k) \right\}$$

$$\text{其中 } \sum_{k=\max(i+1, j+500)}^{j+Z} P^r(k) = P^r(i+1) + P^r(i+2) + \dots + P^r(j+Z);$$

$P^f(i)$ 和 $P^r(i)$ 为随机引物 r_i 作为正向引物和反向引物分别与 v_a 的位置 i 结合的概率, 并且 $Z \leq 10000\text{bp}$ 为 v_a 的想要扩增的区域;

(II) 确定和 / 或选择至少一种靶核酸的至少一个待扩增区域, 该区域的扩增效率高于平均扩增效率; 和

(III) 设计能够与所确定和 / 或选择的区域进行杂交的至少一种寡核苷酸。

31. 根据权利要求 30 所述的装置,其中,所述能够与所选择的区域进行杂交的寡核苷酸是根据以下标准中的至少一种进行选择 and / 或设计的:

(a) 选择 CG 含量为 40% 至 60% 的寡核苷酸;

(b) 选择具有根据最邻近模型计算的最高自由能的寡核苷酸;

(c) 假定寡核苷酸 s_a 和寡核苷酸 s_b 分别为靶核酸 v_a 和 v_b 的子串,则根据 s_a 与来自靶核酸 v_b 的长度为 m 的任意子串 s_b 之间的汉明距离和 / 或根据 s_a 与寡核苷酸 s_b 的最长公共子串选择 s_a ;

(d) 就靶核酸 v_a 特异性的长度为 m 的任何寡核苷酸 s_a 而言,若寡核苷酸 s_a 与异于靶核酸的核酸的任何区域无任何标的,则选择寡核苷酸 s_a ;若长度为 m 的寡核苷酸 s_a 与异于靶核酸的核酸有标的,则选择长度为 m 的具有最小的最大比对长度和 / 或具有最低标的数的寡核苷酸 s_a ;和

(e) 若预测 p_i 杂交到所扩增靶核酸的位置 i 上,则选择在靶核酸的位置 i 上的至少一种寡核苷酸 p_i 。

32. 根据权利要求 31 所述的装置,其中,所述寡核苷酸是根据标准 (e) 选择的。

33. 根据权利要求 31 所述的装置,其中,根据标准 (e),若 $P(p_i | v_a) > \lambda$, 其中 λ 为 0.5 并且 $P(p_i | v_a)$ 为 p_i 杂交到靶核酸 v_a 的位置 i 上的概率,则选择在靶核酸 v_a 的位置 i 上的寡核苷酸 p_i ,其中, $P(p_i | v_a) \approx P(X \leq x_i) = \frac{c_i}{k}$, 其中 X 为代表 v_a 的全部寡核苷酸的扩增效率得分值的随机变量, k 为 v_a 中的寡核苷酸的数目,并且 c_i 为扩增效率得分值 $\leq x_i$ 的寡核苷酸的数目。

34. 根据权利要求 31 所述的装置,其中,根据标准 (e),若 $P(p_i | v_a) > \lambda$, 其中, λ 为 0.8 并且 $P(p_i | v_a)$ 为 p_i 杂交到靶核酸 v_a 的位置 i 上的概率,则选择在靶核酸 v_a 的位置 i 上的寡核苷酸 p_i ,其中, $P(p_i | v_a) \approx P(X \leq x_i) = \frac{c_i}{k}$, 其中 X 为代表 v_a 的全部寡核苷酸的扩增效率得分值的随机变量, k 为 v_a 中的寡核苷酸的数目,并且 c_i 为扩增效率得分值 $\leq x_i$ 的寡核苷酸的数目。

35. 根据权利要求 30 所述的装置,其中,所述扩增使用至少一种正向随机引物和 / 或至少一种反向随机引物。

36. 根据权利要求 30 所述的装置,其中,所述扩增为逆转录 - 聚合酶链式反应。

37. 根据权利要求 35 所述的装置,其中,所述寡核苷酸是根据标准 (a) 到 (d) 中的至少一个和标准 (e) 选择和 / 或设计的。

38. 根据权利要求 35 所述的装置,其中,该正向随机引物和反向随机引物中的每种引物在 5' -3' 方向上包含固定的引物头部和可变的引物尾部,其中至少该可变的尾部与靶核酸 v_a 的一部分杂交。

39. 一种装置,该装置配置为进行检测至少一种靶核酸的方法,该装置包括不溶支持物、设置于所述不溶支持物上的至少一种寡核苷酸和检测器;所述寡核苷酸能够与生物样品中的至少一个靶核酸杂交;其中,所述寡核苷酸是根据权利要求 1 至 11 中的任意一项进行设计和 / 或制备的;该方法包括以下步骤:

(i) 提供至少一种生物样品;

(ii) 对该生物样品中包含的核酸进行扩增；

(iii) 使在所述不溶支持物上的所述寡核苷酸探针与扩增的核酸进行接触；和

(iv) 用所述检测器检测与所述生物样品中靶核酸杂交的寡核苷酸探针。

40. 根据权利要求 39 所述的装置，其中，相对于所述生物样品的核酸，待检测的靶核酸为至少一种外源的核酸。

41. 根据权利要求 39 所述的装置，其中，待检测的靶核酸为至少一种病原体基因组或基因组片段。

42. 根据权利要求 39 所述的装置，其中，病原体核酸为至少一种来自病毒、寄生虫或细菌的核酸或核酸片段。

43. 根据权利要求 39 所述的装置，其中，所述生物样品从人类得到，并且假定存在于所述生物样品中的靶核酸为非人源的。

44. 根据权利要求 39 所述的装置，其中，所述不溶性支持物为微阵列。

45. 根据权利要求 39 所述的装置，其中，所述检测器检测与所述生物样品中靶核酸杂交的寡核苷酸的信号强度。

46. 根据权利要求 45 所述的装置，其中，所述检测器进一步包括计算处理单元，该计算处理单元配置为通过计算加权相对熵得分的分布而对靶核酸评估每个标签探针组内的探针信号强度：

$$WKL(P_a | \bar{P}_a) = \sum_{j=0}^{k-1} \frac{Q_a(j) \log\left(\frac{Q_a(j)}{\bar{Q}_a(j)}\right)}{\sqrt{Q_a(j)[1-Q_a(j)]}}$$

其中， $Q_a(j)$ 为在组 b_j 中所找到的 P_a 内的探针的信号强度的累积分布函数； $\bar{Q}_a(j)$ 为在组 b_j 中所找到的 $\bar{P}_a$ 内的探针的信号强度的累积分布函数， P_a 为病毒 v_a 的探针组， $\bar{P}_a = P - P_a$ 。

47. 根据权利要求 46 所述的装置，其中，在加权相对熵值 ≥ 1.0 时，表明所述生物样品内存在靶核酸。

48. 根据权利要求 46 所述的装置，其中，在加权相对熵值 ≥ 5.0 时，表明所述生物样品内存在靶核酸。

49. 根据权利要求 46 所述的装置，其中，代表不存在靶核酸 v_a 的每个标签探针组具有正态分布的信号强度和 / 或小于 5 的加权相对熵得分，并且，

其中，代表存在至少一种靶核酸的每个标签探针组具有正偏的信号强度分布和 / 或大于 5 的加权相对熵得分。

50. 根据权利要求 46 所述的装置，其中，所述计算处理单元进一步配置为对加权相对熵得分的分布进行安德森-塔林检验，其中 $P > 0.05$ 的结果表明加权相对熵得分的分布为正态和 / 或不存在靶核酸，并且，

其中 $P < 0.05$ 的结果表明存在靶核酸 v_a 。

51. 根据权利要求 50 所述的装置，其中，如果加权相对熵得分的分布为非正态，所述计算处理单元配置为确定具有最高的加权相对熵得分的标签探针组为所述生物样品中出现

的靶。

52. 根据权利要求 51 所述的装置,其中,所述计算处理单元配置为从所述加权相对熵得分中去除所述最高的加权相对熵得分,并且对剩余的加权相对熵得分的分布重复进行安德森-塔林检验,以确定剩余的加权相对熵得分的分布是否为正态。

53. 根据权利要求 52 所述的装置,其中,如果所述剩余的加权相对熵得分的分布为非正态,所述计算处理单元配置为确定具有下一个最高的所述加权相对熵得分的标签探针组为所述生物样品中出现的下一个靶。

54. 根据权利要求 52 所述的装置,其中,所述计算处理单元配置为从所述加权相对熵得分中重复去除所述下一个最高的加权相对熵得分,并且重复进行安德森-塔林检验直至加权相对熵得分的分布变为正态的步骤,从而检测任何其它靶核酸分子和 / 或共感染病原体的存在。

寡核苷酸设计和 / 或核酸检测的方法和 / 或装置

技术领域

[0001] 本发明涉及寡核苷酸设计和 / 或核酸检测的领域。本发明的方法、装置和 / 或产物可以用于检测病原体,例如用于检测病毒。

背景技术

[0002] 精确而迅速地检测人类患者和人群中的病毒病原体和细菌病原体具有极其重要的医学及流行病学意义。以往,诊断技术依赖于细胞培养传代和多种免疫学测定法或染色方法。目前精确并灵敏地检测传染因子仍然是困难的,尽管该领域内的进展相当大。培养和基于抗体检测的常规方法仍然在微生物学实验室中发挥核心作用,尽管存在疾病表现与诊断间的滞后问题,以及由这些方法可以检测的生物种类数目有限的问题。更迅速地诊断感染将降低发病率和死亡率,例如,通过更早的实施适宜的抗微生物治疗。在过去数十年间,已经提出多种方法以实现此目的;基于核酸检测的那些方法,包括基于聚合酶链式反应(PCR)及微阵列的技术,似乎最有前景。尤其是,基于PCR的测定法已经得到实施,这使得能以更高的检测灵敏度、更迅速地诊断可疑病原体。然而在临床实践中,病因学因子常常无法确定,以复杂方式规避检测。例如,某些病毒不可培养。而有时候,患者样品的质量可能很差或对于通过常规技术检测病原体而言,其滴度不足。此外,基于PCR的方法和基于抗体的方法均仅因天然遗传多样性导致了PCR引物结合位点的改变及抗原漂移可能不能识别可疑病原体。

[0003] 具有同时检测多种病原体能力的DNA微阵列和寡核苷酸微阵列已经得到描述(Wang等,2002;Urisman等,2005)。然而,未解决的技术问题妨碍它们在临床情况下的例行使用。例如,如何根据扩增和交叉杂交的人为产物来选择提供包含最多病原体“标签”的信息的探针?何种强度的荧光信号及标签探针才构成检测到的病原体?何谓优化检测算法的精确性和灵敏性?(Striebel等,2003;Bodrossy和Sessitsch,2004;Vora等,2004)。

[0004] 因此,本技术领域内需要检测核酸的替代方法及改良方法。尤其是,需要用于检测病原体的替代诊断方法和 / 或改良诊断方法。

发明内容

[0005] 本发明的目的是为了解决上述问题,并且具体提供设计寡核苷酸的方法、装置和 / 或产物。尤其是,本发明提供设计寡核苷酸探针和 / 或引物的方法、装置和 / 或产物。还提供核酸检测的方法、装置和 / 或产物。

[0006] 根据第一方面,本发明提供了设计至少一种用于核酸检测的寡核苷酸的方法,该方法包括任意顺序的以下步骤:

[0007] (I) 确定和 / 或选择至少一种靶核酸的至少一个待扩增区域,该区域的扩增效率(AE)高于平均AE;和

[0008] (II) 设计至少一种能够与所选择的区域进行杂交的寡核苷酸。

[0009] 所述至少一种寡核苷酸可以是至少一种探针和 / 或引物。

[0010] 尤其是,在步骤(I)中,对全长靶核酸或其至少一个区域内的每个位置 i 确定 AE 得分,并且随后得到平均 AE 得分。可以选择 AE 得分高于平均 AE 得分的那些区域作为靶核酸的待扩增区域。更尤其是,所选区域的 AE 可以被计算为扩增效率得分 (AES),AES 是正向引物 r_i 可能与靶核酸的位置 i 结合并且反向引物 r_j 可能在靶核酸的位置 j 处结合的概率, $|i-j|$ 是靶核酸的想要扩增的区域。区域 $|i-j|$ 可以优选为 $\leq 10000\text{bp}$,更优选为 $\leq 5000\text{bp}$,或 $\leq 1000\text{bp}$,例如 $\leq 500\text{bp}$ 。尤其是,正向引物和反向引物可以是随机引物。

[0011] 根据另一方面,步骤(I)包括确定几何级扩增偏倚对靶核酸内每个位置的影响,和选择至少一个待扩增区域作为扩增效率高于平均扩增效率的区域。例如,几何级扩增偏倚是 PCR 偏倚。

[0012] 可以根据本领域已知的任何寡核苷酸设计技术实现设计至少一种能够与在步骤(I)中所选择的区域杂交的寡核苷酸的步骤(II)。尤其是,可以根据以下标准中的至少一种选择和设计能够与所选择的区域杂交的寡核苷酸:

[0013] (a) 选择 CG 含量为 40% 至 60% 的寡核苷酸;

[0014] (b) 选择具有根据最邻近模型 (Nearest-Neighbor model) 计算的最高自由能的寡核苷酸;

[0015] (c) 假定寡核苷酸 s_a 和寡核苷酸 s_b 分别是靶核酸 v_a 和 v_b 的子串,则根据 s_a 与长度为 m 的任意子串 s_b 之间的汉明距离和 / 或 s_a 与寡核苷酸 s_b 的最长公共子串来选择 s_a 。

[0016] (d) 就靶核酸 v_a 特异性的长度为 m 的任何寡核苷酸 s_a 而言,若寡核苷酸 s_a 与异于靶核酸的核酸的任何区域无任何标的 (hit),则选择寡核苷酸 s_a ;如果长度为 m 的寡核苷酸 s_a 与异于靶核酸的核酸有标的,则选择长度为 m 的具有最小的最大比对长度和 / 或具有最小标的数的寡核苷酸 s_a ;和

[0017] (e) 如果预测 p_i 能杂交到所述扩增的靶核酸的位置 i 上,则选择在靶核酸的位置 i 上的寡核苷酸 p_i 。

[0018] 尤其是,所述寡核苷酸可以是探针和 / 或引物。

[0019] 因此,可以采用两项或多项上述标准设计所述寡核苷酸。例如,可以通过应用 (a) 至 (e) 中所有的标准设计所述寡核苷酸。也可以使用在本文中没有明确地提及但为本领域技术人员所知的其它标准。

[0020] 尤其是,根据标准 (e),如果 $P(p_i | v_a) > \lambda$, 其中 λ 为 0.5 并且 $P(p_i | v_a)$ 为 p_i 杂交到靶核酸 v_a 的位置 i 上的概率,则选择在靶核酸 v_a 的位置 i 上的寡核苷酸 p_i 。 λ 更优选为 0.8。

[0021] 尤其是, $P(p_i | v_a) \approx P(X \leq x_i) = \frac{c_i}{k}$, 其中 X 为代表 v_a 的全部寡核苷酸的扩增效率得分 (AES) 的随机变量, k 为 v_a 中寡核苷酸的数目,并且 c_i 为 AES 值 $\leq x_i$ 的寡核苷酸的数目。

[0022] 根据本发明的另一方面,设计如上所述寡核苷酸的方法还包括制备选择及设计的寡核苷酸的步骤。可以根据本领域已知的任何标准方法,例如通过化学合成或光蚀刻技术,可以制备可以是至少一种探针和 / 或引物的寡核苷酸。

[0023] 根据另一方面,本发明提供检测至少一种靶核酸的方法,该方法包括以下步骤:

[0024] (i) 提供至少一种生物样品;

[0025] (ii) 对该生物样品中包含的核酸进行扩增；

[0026] (iii) 提供能够与至少一种假定存在于该生物样品中的靶核酸进行杂交的至少一种寡核苷酸,其中,使用根据本文中所述的本发明的任何方面的方法设计和 / 或制备该寡核苷酸 ;和

[0027] (iv) 使该寡核苷酸接触扩增的核酸和 / 或检测与靶核酸杂交的寡核苷酸。

[0028] 尤其是,所述寡核苷酸是探针。

[0029] 扩增步骤 (ii) 可以在随机引物的存在下实施。例如,扩增步骤 (ii) 是在存在至少一种正向随机引物和 / 或至少一种反向随机引物的条件下进行的。可以使用本领域已知的任何扩增方法。例如,扩增方法是逆转录 - 聚合酶链式反应 (RT-PCR)。

[0030] 尤其是,与靶核酸 v_a 的位置 i 结合的正向随机引物和与靶核酸 v_a 的位置 j 结合的反向随机引物选自对靶核酸 v_a 的每个位置 i 具有如下扩增效率得分 (AES_i) 的引物 :

$$[0031] \quad AES_i = \sum_{j=i-Z}^i \left\{ P^f(j) \times \sum_{k=\max(i+1, j+500)}^{j+Z} P^r(k) \right\},$$

$$[0032] \quad \text{其中,} \quad \sum_{k=\max(i+1, j+500)}^{j+Z} P^r(k) = P^r(i+1) + P^r(i+2) + \dots + P^r(j+Z),$$

[0033] $P^f(i)$ 和 $P^r(i)$ 为随机引物 r_i 作为正向引物和反向引物分别与 v_a 的位置 i 结合的概率,并且 $Z \leq 10000\text{bp}$ 是想要扩增的 v_a 的区域。更尤其是, Z 可以是 $\leq 5000\text{bp}$ 、 $\leq 1000\text{bp}$ 或 $\leq 500\text{bp}$ 。

[0034] 扩增步骤可以包括正向引物和反向引物,并且正向引物和反向引物中的每种引物在 $5' - 3'$ 方向上包含固定的引物头部和可变的引物尾部,并且其中至少可变的尾部与靶核酸 v_a 的一部分杂交。尤其是,扩增步骤可以包括具有 SEQ ID NO :1 所示的核苷酸序列或该核苷酸序列的变体或衍生物的正向随机引物和 / 或反向随机引物。

[0035] 生物样品可以是采自哺乳动物的任何样品,例如来自人类。生物样品可以是组织、血清、鼻咽冲洗液、唾液、任何其它体液、血液、尿、粪便等。生物样品可以是在实施扩增步骤前被处理以释放生物样品中包含的核酸。靶核酸可以是想要检测的任何核酸。待检测的靶核酸可以是至少对生物样品的核酸为外源的核酸。因此,若生物样品来自人,则待检测的外源靶核酸 (若其存在于生物样品中) 是非人源的核酸。根据本发明的方面,待检测的靶核酸至少是病原体的基因组或基因组片段。病原体核酸可以是来自病毒、寄生虫或细菌的至少一种核酸或核酸片段。

[0036] 因此,本发明提供检测假定存在于生物样品中的至少一种靶核酸的方法。该方法可以是用于检测生物样品中病原体存在的诊断方法。例如,若生物样品从人类得到,则假定存在于生物样品中的靶核酸为非人源的。

[0037] 根据本发明的任何方法所设计和 / 或制备的寡核苷酸可以在溶液中被使用或可以被放置不溶性支持物上。例如,寡核苷酸探针可以根据本领域已知的任何技术而涂到、滴到或刷到 (print) 不溶性支持物上。支持物可以是微阵列、生物芯片、膜 / 合成性表面、固体支持物或凝胶。

[0038] 随后将探针与生物样品的核酸接触,若靶核酸存在,则它与探针杂交,从而检测出

存在的靶核酸。尤其是,在检测步骤 (iv) 中,与 v_a 杂交的探针的信号强度的均值在统计学上高于探针 $\notin v_a$ 的信号强度的均值,则表明生物样品中存在 v_a 。

[0039] 更尤其是,在检测步骤 (iv) 中,与 v_a 杂交的探针的信号强度的均值在统计学上高于探针 $\notin v_a$ 的信号强度的均值,并且该方法还包括如下步骤:计算具有高信号强度的探针 $\notin v_a$ 的比例与在检测方法内所用的具有高信号强度的探针的比例的相对差异,探针 v_a 的信号强度的密度分布比探针 $\notin v_a$ 的信号强度的密度分布更为正偏,则表明生物样品中存在 v_a 。

[0040] 例如,在检测步骤 (iv) 中,生物样品中的至少一种靶核酸得到检测,若靶核酸的探针信号强度的密度分布不是正态性的,即更为正偏,这是由安德森-塔林 (Anderson-Darling) 检验值 ≤ 0.05 和 / 或 t- 检验值 ≤ 0.1 和 / 或加权相对熵 (Weighted Kullback-Leibler, WKL) ≥ 1.0 、优选 ≥ 5.0 表示。最尤其是, t- 检验值 ≤ 0.05 。

[0041] 尤其是,检测步骤 (iv) 的方法还包括通过计算加权相对熵得分的分布,评估在每个病原体特异性标签探针组 (SPS) 中的探针对靶核酸 v_a 的探测信号强度:

[0042]

$$WKL(P_a | \bar{P}_a) = \sum_{j=0}^{k-1} \frac{Q_a(j) \log\left(\frac{Q_a(j)}{Q_a(j)}\right)}{\sqrt{Q_a(j)[1-Q_a(j)]}}$$

[0043] 其中 $Q_a(j)$ 是在 P_a 中的探针的信号强度的累积分布函数,且 P_a 是在组 b_j 中找到的; $Q_a(j)$ 是在 $\bar{P}_a$ 中的探针的信号强度的累积分布函数,且 $\bar{P}_a$ 是组 b_j 中所找到。 P_a 是病毒 v_a 的探针组并且 $\bar{P}_a = P - P_a$ 。

[0044] 每个标签探针组 (SPS) 具有正态性分布的信号强度 (由安德森-塔林检验值 ≤ 0.05 评定) 和 / 或得分小于 5 的加权相对熵 ($WKL < 5$), 则表明不存在靶核酸 v_a 。每个标签探针组 (SPS) 具有正偏的信号强度分布和 / 或大于 5 的加权相对熵得分 ($WKL > 5$), 则表明存在至少一种靶核酸 v_a 。

[0045] 所述方法还可以包括对 WKL 得分的分布进行安德森-塔林检验,其中, $P > 0.05$ 的结果表示不存在靶核酸 v_a , 或 $P < 0.05$ 的结果表示存在靶核酸 v_a 。此外,还可以进行安德森-塔林检验,表明存在其它共感染性靶核酸。根据另一方面,本发明提供确定靶核酸 v_a 存在的方法,该方法包括检测至少一种寡核苷酸探针 (该探针是根据本领域内的任何已知方法且不必要限于本发明方法选择和设计的) 与至少一种靶核酸 v_a 的杂交,并且,其中与 v_a 杂交的探针的信号强度的均值在统计学上高于探针 $\notin v_a$ 的信号强度的均值,则表明存在 v_a 。尤其是,与 v_a 杂交的探针的信号强度的均值在统计学上高于探针 $\notin v_a$ 的信号强度的均值,并且该方法还包括以下步骤:计算具有高信号强度的探针 $\notin v_a$ 的比例与在检测方法中所用的具有高信号强度的探针的比例的相对差异,探针 v_a 的信号强度的密度分布比探针 $\notin v_a$ 的信号强度的密度分布更为正偏,则表明生物样品中存在 v_a 。更尤其是, t- 检验值 ≤ 0.1 和 / 或安德森-塔林检验值 ≤ 0.05 和 / 或加权相对熵 ≥ 1.0 、优选 ≥ 5.0 时表示生物样品中存在靶核酸。例如, t- 检验值可以是 ≤ 0.05 。

[0046] 根据另一方面,本发明提供检测至少一种靶核酸的方法,该方法包括以下步骤:

[0047] (i) 提供至少一种生物样品;

[0048] (ii) 对该生物样品中包含的至少一种核酸进行扩增;

[0049] (iii) 提供能够与至少一种假定存在于该生物样品中的靶核酸杂交的至少一种寡核苷酸 ; 和

[0050] (iv) 使所述寡核苷酸与扩增的核酸进行接触并且检测与靶核酸杂交的寡核苷酸, 其中, 与 v_a 杂交的寡核苷酸的信号强度的均值在统计学上高于寡核苷酸 $\neq v_a$ 的均值, 则表明在生物样品中存在 v_a 。

[0051] 尤其是, 所述寡核苷酸是寡核苷酸探针。

[0052] 在步骤 (iv) 内, 与 v_a 杂交的探针的信号强度的均值在统计学上高于探针 $\neq v_a$ 的信号强度的均值, 并且该方法还包括以下步骤: 计算具有高信号强度的探针 $\neq v_a$ 的比例与在检测方法内所用的具有高信号强度的探针的比例的相对差异, 探针 v_a 的信号强度的密度分布比探针 $\neq v_a$ 的信号强度的密度分布更为正偏, 则表明生物样品中存在 v_a 。尤其是, 在步骤 (iv) 中, t - 检验值 ≤ 0.1 和 / 或安德森 - 塔林检验值 ≤ 0.05 和 / 或加权相对熵值 ≥ 1.0 、优选 ≥ 5.0 时表示生物样品中存在至少一种靶核酸。 t - 检验值可以是 ≤ 0.05 。相对于生物样品的核酸, 待检测的核酸为外源的核酸。待检测的靶核酸可以是至少一种病原体基因组或基因组片段。病原体核酸可以是来自病毒、寄生虫或细菌的至少一种核酸或核酸片段。尤其是, 当生物样品从人类得到时, 且如果在该生物样品中存在靶核酸, 则该靶核酸为非人源的基因组。探针可以被置于不溶性支持物上。支持物可以是微阵列、生物芯片或 / 合成的表面。

[0053] 本发明提供一种装置, 该装置包括用于实施本发明的方法的装置。尤其是, 所述装置可以用于设计用来检测和 / 或扩增核酸的寡核苷酸, 其中, 所述装置被配置成确定和 / 或选择至少一种靶核酸的至少一个待扩增区域, 其中, 所述区域的扩增效率 (AE) 高于平均的 AE ; 并且设计能够与所确定和 / 或选择的区域进行杂交的至少一种寡核苷酸。更尤其是, 所述装置可以被配置成检测至少一种靶核酸, 包括以下步骤中的任一步骤: 提供至少一种生物样品 ; 扩增该生物样品中包含的核酸 ; 提供至少一种能够与至少一种假定存在生物样品中的靶核酸杂交的寡核苷酸, 其中寡核苷酸根据本发明所配置的装置进行设计和 / 或制备 ; 并使寡核苷酸与扩增的核酸进行接触和 / 或检测与靶核酸杂交的寡核苷酸。

[0054] 本发明还提供被配置成用于实施本发明的方法的至少一种计算机程序产品。本发明还提供至少一种存贮本发明装置的配置的电子存储介质。根据一个方面, 本发明提供被配置成包括实施本发明方法的软件的可移动的电子存储介质。尤其是, 所述可移动的电子存储介质可以包含软件, 其中该软件被配置成确定 WKL 得分和 / 或安德森 - 塔林检验以便设计至少一种寡核苷酸探针和 / 或引物, 和 / 或检测至少一种靶核酸。更尤其是, 包含软件编译程序的可移动的电子存储介质可以包含如本发明定义的 WKL、安德森 - 塔林检验、探针的设计和 / 或靶核酸的检测。因此, 本发明还提供被配置成如上所述的设计的软件。

附图说明

[0055] 图 1 表示一对随机引物对病毒序列 (SEQ ID NOS :1 至 9) 的 RT-PCR 结合过程。图 1 的标记如下所示 :

[0056] A : 逆转录 (RT)。引物与模板结合。

[0057] B : 生成标记的 RT 产物 (具体是用假定的病毒序列模板和假定的特定的随机引物)。

[0058] C:第二链合成完成,掺入标签。

[0059] D:使用 PCR 引物 GTTCCAGTCACGATA (SEQ ID NO:8) 扩增标记的 RT 产物;

[0060] 图 2 表示用于 RSV (呼吸道合胞体病毒) B 基因组的扩增效率得分 (AES) 图;

[0061] 图 3 表示实施例 1 的寡核苷酸探针对 RSV B 的信号强度;

[0062] 图 4(A,B)。图 4A 表示被测样品的病毒的信号强度的密度分布。箭头指出该分布的正偏性。虽然存在噪音,但是也显著存在真实信号。图 4B 表示样品中不存在病毒的信号强度的密度分布。该分布是噪音占优势的;

[0063] 图 5 表示病原体检测芯片数据的分析流程图;

[0064] 图 6 是寡核苷酸探针设计示意图。该示意图说明由 NC_001781 人呼吸道合胞体病毒 (RSV) 的基因组所产生的嵌搭探针 (tiling probe)。数字代表每个探针的起始位置和终止位置。合成了 1948 个探针以覆盖完整的 15225bpRSV 基因组。该过程重复用于其余 34 种病毒的基因组;

[0065] 图 7 是 (A、B、C) 微阵列短线标记的检索表:

[0066]

病毒科	病毒属 / 种
正粘病毒科	重症急性呼吸综合症 (Sars) 新加坡变种 2500 OC41 型病毒 229E 型病毒
冠状病毒科	流感病毒 (Flu)A 流感病毒 B
微 RNA 病毒科	肠道病毒 (Enterovirus) D 肠道病毒 C 致肠细胞病变人孤儿病毒 (ECHO) 1 肠道病毒 B 肠道病毒 A 鼻病毒 (Rhinovirus) 89 鼻病毒 B 甲型肝炎病毒 (Hepatitis A) 口蹄疫病毒 C
布尼亚病毒科	汉坦病毒 (Hantavirus) 辛诺柏病毒 (Sin Nombre)
黄病毒科	西尼罗病毒

[0067]

	日本脑炎病毒 登革病毒 3 登革病毒 1 登革病毒 2 登革病毒 4 黄热病毒
副粘病毒科	副流感病毒 1 副流感病毒 3 Nipah(聂帕病毒) 副流感病毒 2 新城疫病毒 呼吸道合胞病毒 (RSV) (B1) 偏肺炎病毒
其它病毒科	人乳头瘤病毒 (HPV) 型 10 人免疫缺陷症病毒 (HIV) 1 乙型肝炎病毒 (Hep B) 风疹病毒 淋巴细胞脉络丛脑膜炎病毒 (LCMV)-S LCMV-L 辣椒温和斑点病毒 (PMMV) 人对照

[0068] 从 SARS Sin850- 感染的细胞系 (A) 或登革病毒 I- 感染的细胞系 (B) 中分离的 RNA 杂交到病原体微阵列上, 随后分别进行 SARS- 特异性 RT-PCR 或登革病毒 I- 特异性 RT-PCR。SARS 与其它冠状病毒科基因组交叉杂交 (以黑色表示), 特别与基因组中高度保守的中央部分交叉杂交 (Ruan 等, 2003)。登革病毒 I 与探针交叉杂交, 其中所述的探针是根据登革病毒 I 与黄病毒科和其它基因组序列相似性而从黄病毒科和其它基因组获得的。通过检查汉明距离 (HD) 和最大连续匹配 (MCM) 得分, 我们建立了预测交叉杂交是否会发生的阈值并利用该信息生成计算机模拟 (in silico) 杂交标签。(C) 使用随机 RT-PCR 进行扩增从诊断患有 RSV 的临床患者中分离的 RNA 并杂交到病原体微阵列上;

[0069] 图 8: 探针汉明距离 (HD)、探针最大连续匹配 (MCM) 与探针信号强度间的关系。平均探针信号强度随 HD 增加及 MCM 降低而减少。这与可检测探针的百分数降低相关 (信号强度 > 均值 + 2SD)。在最佳交叉杂交阈值 $HD \leq 4$ 或 $MCM \geq 18$ (阴影部分) 上, 可以检测到大于 98% 的探针。在 $HD = 5$ 或 $MCM = 17$ 上, 检测率下降到 85%;

[0070] 图 9 (A, B): 从 RSV 感染的患者中分离的 RNA 杂交到病原体检测阵列上。(A) 全部 53, 555 个探针的探测信号强度的分布显示出正态分布 (灰色实线)。当在基因组特异性水平上检查时, 非 RSV 探针例如副流感病毒 -1 (灰色点线) 也显示出正态分布。RSV- 特异性探针的信号强度具有正偏性, 在分布的尾部具有更高信号强度 (黑色实线)。(B) 对 35 个 SPS 的 WKL 得分的分布频率, 大部分分布在 -5 与 3 之间。然而, 对 RSV 基因组的 WKL 得分

是 17, 因此该分布不是正态性的 (安德森 - 塔林检验 $P < 0.05$)。排除异常 (outlier) 基因组产生正态分布。从该计算中, 得出的结论是 RSV 存在于杂交的样品中;

[0071] 图 10 : AES 指示探针的扩增效率。可在高于超过 5 个实验的信号强度阈值上检测到更高比例的具有高 AES 的探针;

[0072] 图 11 表示使用微阵列检测病原体所需的过程的示意图;

[0073] 图 12 : 杂交信号强度与扩增效率得分 (AES) 的相关, $P = 2.2 \times 10^{-16}$ 。将 RSV 患者的样品杂交至微阵列上, 并且将每个探针的信号强度与计算的 AES 作图。用于在一般阵列上高置信度检测的信号阈值由绿线表示;

[0074] 图 13 : 使用 AES- 优化的引物标签进行随机 RT-PCR, AES 增加 10-30 倍。预测优化的引物对表现在微阵列上的所有 35 个基因组都具有相同性能。大部分患者的样品使用 AES- 优化的引物 A2 进行扩增;

[0075]

SEQ ID NO :	引物	核苷酸序列
10	A1	GTTTCCCAGTCACGATA
11	A2	GATGAGGGAAGATGGGG
12	A3	CTCATGCACGACCCAAA
13	A4	AGATCCATTCCACCCCA

[0076] 图 14 (A, B) 是微阵列短线标记的检索表:

[0077]

病毒科	病毒属 / 种
正粘病毒科	重症急性呼吸综合症新加坡变种 2500 OC41 229E
冠状病毒科	流感病毒 A 流感病毒 B

[0078]

微 RNA 病毒科	肠道病毒 D 肠道病毒 C 致肠细胞病变人孤儿病毒 1 肠道病毒 B 肠道病毒 A 鼻病毒 89 鼻病毒 B Hep A 口蹄疫病毒 C
布尼安病毒科	汉坦病毒 辛诺柏病毒
黄病毒科	西尼罗病毒 日本脑炎病毒 登革病毒 3 登革病毒 1 登革病毒 2 登革病毒 4 黄热病毒
副粘病毒科	副流感病毒 1 副流感病毒 3 聂帕病毒 副流感病毒 2 新城疫病毒 RSV(B1) 偏肺炎病毒
其它病毒科	HPV 型 10 HIV 1 Hep B 风疹病毒 LCMV-S LCMV-L PMMV 人对照

[0079] 在随机 RT-PCR 中引物标签的选择对 PCR 效率有显著影响。热图显示探针与临床的 hMPV 样品杂交,随后使用 Bohlander 等 1992 所述的原始引物 (A) 或使用根据 PCR 建模设计的旨在确保高效扩增表现在微阵列上的所有基因组 (高 AES) 的引物 (B) 进行 RT-PCR;

[0080] 图 15 :对 RSV 患者 #412 的诊断性 PCR 结果证实该患者没有被冠状病毒感染。(A)

使用泛冠状病毒引物的 PCR。泳道 1:OC43 冠状病毒阳性对照;泳道 2:229E 冠状病毒阳性对照;泳道 3:RSV 患者 #412,泳道 4:PCR 引物和仅作为阴性对照的试剂。1kb 序列梯。(B) 使用 OC43 特异性引物的 PCR。泳道 1:OC43 冠状病毒阳性对照;泳道 2:RSV 患者 #412;泳道 3:来自 ATCC 的纯化 RSV;泳道 4:PCR 阴性对照。50bp 序列梯。(C) 使用 229E 特异性引物的 PCR。泳道 1:229E 冠状病毒阳性对照;泳道 2:RSV 患者 #412,泳道 3:PCR 阴性对照。1kb 序列梯。

具体实施方式

[0081] 为方便起见,本说明书中提及的参考文献以参考文献列表形式列出并附加在实施例的结尾。这些参考文献的完整内容引入本文作为参考。

[0082] 本发明解决了现有技术的问题,特别是提供设计寡核苷酸的至少一种方法、装置和/或产物。具体而言,本发明提供设计探针和/或引物的方法、装置和/或产物。本发明还提供核酸检测的方法、装置和/或产物。

[0083] 尽管已经提出了使用寡核苷酸杂交微阵列作为确定病原体存在的工具的概念,然而仍存在明显的障碍,从而妨碍这些微阵列的常规使用 (Striebel, H. M., 2003)。这些障碍包括探针设计和数据分析 (Striebel, H. M., 2003; Bodrossy, L. & Sessitsch, A., 2004; Vora, G. J., 等, 2004)。本发明的发明人在一种试验性微阵列中观察到尽管仔细选择探针,然而计算机模拟设计的最佳探针不一定与患者样品良好地杂交。本发明人认识到要产生与患者材料始终良好地杂交的探针,必须开发新的和/或探针设计改良的方法以便确定最佳设计预测值。尤其是,如实施例部分中所述,本发明人创造了包含重叠性 40 聚体 (mer) 探针的微阵列,其中所述探针在 35 种病毒基因组范围内嵌搭 (tiled)。不过,本发明不限于这种具体应用、探针长度和靶核酸类型。

[0084] 根据本发明的具体方面,本发明的发明人描述如何优化支持物,尤其是微阵列平台,以使该支持物在靶核酸检测,尤其是病原体检测中成为可行的工具。本发明人还确定了探针设计预测值,包括解链温度、探针的 GC 含量、二级结构、汉明距离、与人基因组的相似性、PCR 引物标签在随机 PCR 扩增效率方面的影响和/或序列多态性的影响。将这些结果视作和/或纳入探针和/或引物设计方法和标准的开发中。根据更具体的方面,本发明人开发了可以精确预测靶核酸存在的数据分析算法,其中靶核酸可能是病原体或可能不是病原体。例如病原体可以是,但不限于病毒,细菌和/或寄生虫。即使探针不是理想地设计的,也可以使用该算法。这种结合探针设计方法学检测算法显著改善预测的置信度水平 (见表 6 和 7)。

[0085] 根据具体方面,本发明的方法可以不需要预测可能的病原体,但是或许能够以非偏倚方式检测到大部分已知的人病毒、细菌和/或寄生虫,以及一些新物种。将基因组或基因组片段定义为生物染色体中的全部遗传材料。来源于特定生物染色体的遗传材料中的 DNA 是基因组 DNA。基因组文库是克隆的集合,其中克隆由一套随机生成的代表生物完整基因组的重叠性 DNA 片段产生的。在本发明的这种检测平台背后的逻辑基础是病毒、细菌和/或寄生虫的每个物种在它们基因组的原始序列中均含有独特的分子标签。鉴定这些识别性区域使设计的合理的寡核苷酸探针能够用于特异性地鉴定单个物种以及 (在某些情况下) 单个株。同时设计和/或制备代表科成员及属成员间最高度保守性区域的寡核苷酸 (寡)

探针将能够检测并部分表征某些新的病原体。此外,在单个支持物上包含此类全部探针可以使检测同时共感染临床样品的多种病毒、细菌和 / 或寄生虫。支持物可以是不溶性支持物,尤其是固体支持物,例如微阵列或生物分析芯片。

[0086] 根据具体方面,本发明可以用作诊断工具,这取决于设计寡核苷酸探针的方式,和 / 或如何解读并分析由微阵列生成的数据。

[0087] 扩增效率的确定

[0088] 根据第一方面,本发明提供设计用于核酸检测的寡核苷酸探针的方法,该方法包括任意顺序的以下步骤:

[0089] (i) 确定和 / 或选择至少一种靶核酸的至少一个待扩增区域,该区域的扩增效率(AE) 高于平均的 AE ;和

[0090] (ii) 设计至少一种能够与所确定和 / 或选择的区域进行杂交的寡核苷酸探针。

[0091] 尤其是,在步骤 (I) 中,对全长靶核酸上或其区域内的每个位置 i 确定 AE 得分并且获得平均 AE。选择 AE 高于平均值的那些区域作为靶核酸的待扩增区域。尤其是,所选区域的 AE 可以被计算为扩增效率得分 (AES),其中 AES 是正向引物 r_i 可能与靶核酸的位置 i 结合并且反向引物 r_j 可能在靶核酸的位置 j 处结合的概率, $|i-j|$ 是靶核酸的想要的扩增区域。区域 $|i-j|$ 可以优选为 $\leq 10000\text{bp}$,更优选为 $\leq 5000\text{bp}$ 或 $\leq 1000\text{bp}$,例如 $\leq 500\text{bp}$ 。尤其是,正向引物和 / 或反向引物可以是随机引物。根据另一方面,确定和 / 或选择靶核酸的待扩增区域的步骤 (i) 包含确定几何级扩增偏倚对靶核酸中每个位置的影响,并且选择待扩增区域作为扩增效率高于平均扩增效率的区域。几何级扩增偏倚可以被定义为核酸内某些区域的扩增的能力比其它区域的扩增的能力更高。例如,几何级扩增偏倚是 PCR 偏倚。

[0092] 扩增效率的建模

[0093] 由于不知道何种靶核酸(例如病原体)存在于患者样品内,因此可以在扩增步骤和 / 或逆转录(RT) 过程中使用随机引物,以确保使存在的全部 RNA 无偏倚地逆转录成 DNA。本领域已知的任何随机扩增方法可以用于本发明的目。在本说明书中,随机扩增方法可以是 RT-PCR。然而,技术人员将明白本发明的方法不限于 RT-PCR。RT-PCR 方法可能易受 RT-PCR 方法中引物二聚体结合及较差扩增效率所致的信号不精确性的影响(Bustin, S. A. 等,2004)。为克服此问题,本发明人已经通过使用随机引物对 RT-PCR 方法建模。

[0094] 根据本发明的具体方面,扩增步骤包括正向引物和反向引物,并且正向引物和反向引物中的每种引物在 5' -3' 方向上包含固定的引物头部和可变的引物尾部,并且其中至少可变的尾部与靶核酸 v_a 的一部分杂交。固定的引物头部和可变的引物尾部的尺寸可以是以 mer 为单位的适合本发明方法的目的任何尺寸。固定的头部可以是 10-30mer,优选为 5-25mer,例如 17mer。可变尾部可以是 1-20mer,优选为 5-15mer,例如 9mer。这些正向引物和反向引物的实例如图 1 所示。更尤其是,扩增步骤可以包括具有核苷酸序列 5'-GTTTCCAGTCACGATANNNNNNNN-3' (SEQ ID NO:1) 的正向随机引物和 / 或反向随机引物,其中,N 是 A、T、C 和 G 及它们的衍生物中的任意一种。

[0095] 根据具体实施方案,仍如图 1 所示,本发明的发明人按照下列方法建立了随机 RT-PCR 方法模型。设 v_a 是样品中实际病毒。在 RT-PCR 方法中所用的随机引物优选为具有固定的 17mer 头部和可变的 9mer 尾部的 26mer 引物,即 (5'-GTTTCCAGTCACGATANNNNNNNN-3') (SEQ ID NO:1,尤其为 SEQ ID NOS:2-7)。然而,对于技术人员来说,本发明的引物显

然不限于 SEQID NOS :1-7 和图 1 的序列。实际上,引物的核苷酸数,尤其是头部及可变尾部的核苷酸数,可以变化并且在以上所讨论的范围内加以选择。为在 v_a 的位置 i 与 j 间的区域内得到 RT-PCR 产物,本发明人需要 (1) 结合至位置 i 的正向引物, (2) $|i-j| \leq 10000$ 以及 (3) 结合至位置 j 的反向引物。作为靶核酸内想要扩增区域的 $|i-j|$ 优选为 $\leq 5000\text{bp}$, 更优选为 ≤ 1000 , 例如 $\leq 500\text{bp}$ 。RT-PCR 产物的质量取决于正向引物和 / 或反向引物与 v_a 结合得的情况。一些随机引物可以比其它随机引物更好地与 v_a 结合。确定此类引物并确定它们在哪里与 v_a 结合提示了 v_a 的特定区域有多大可能被扩增。使用这种方法,提供对 v_a 的每个位置计算扩增效率得分 (AES) 的扩增效率模型。

[0096] 对于靶核酸 v_a 的特定位置 i , $P^f(i)$ 和 $P^r(i)$ 是随机引物 r_i 作为正向引物和反向引物分别可以与 v_a 的位置 i 结合的概率。为简便起见,假设随机引物仅当该随机引物的最后 9 个核苷酸是 v_a 的反向互补物 (正向引物) 的子串或是 v_a (反向引物) 的子串时才可以与 v_a 结合。这如图 1 所示。基于非常确定的引物设计标准 (Wu, D. Y., 等, 1991), 若 r_i 形成明显的引物二聚体或具有极端的解链温度, 则 $P^f(i)$ 被评定为低。另一方面, 若 r_i 未形成任何明显的引物二聚体并具有最佳的解链温度, 则 $P^f(i)$ 将被评定为高。需要注意的是, 若随机引物的头部与 v_a 相似, 这也可以辅助结合并因此产生更高的 $P^f(i)$ 。类似地, 计算了 $P_r(i)$ 。

[0097] 随机引物 r_i 作为正向引物在 v_a 的位置 i 的结合影响位置 i 上游至少 10000 个核苷酸的 RT-PCR 产物的质量。随机引物 r_i 作为反向引物在 v_a 的位置 i 的结合影响位置 i 下游至少 10000 个核苷酸的 RT-PCR 产物的质量。因此, 对 v_a 的每个位置 i 的扩增效率得分 AES_i 可以通过考虑用于扩增位置 i 的所有正向引物和反向引物对的总效果计算:

$$[0098] \quad AES_i = \sum_{j=i-Z}^i \left\{ P^f(j) \times \sum_{k=\max(i+1, j+500)}^{j+Z} P^r(k) \right\}$$

$$[0099] \quad \text{其中} \quad \sum_{k=\max(i+1, j+500)}^{j+Z} P^r(k) = P^r(i+1) + P^r(i+2) + \dots + P^r(j+Z)$$

[0100] $P^f(i)$ 和 $P^r(i)$ 是随机引物 r_i 作为正向引物和反向引物分别与 v_a 的位置 i 结合的概率, 并且 $Z \leq 10000\text{bp}$ 是 v_a 的欲扩增的区域。

[0101] 因此, Z 可以是 $\leq 10000\text{bp}$ 、 $\leq 5000\text{bp}$ 、 $\leq 1000\text{bp}$ 或 $\leq 500\text{bp}$ 。

[0102] 为验证由病毒的不同区域所表现的信号强度上的变异是否与所述区域的相应扩增效率得分直接相关, 进行了几个对常见的感染人的病原体即人呼吸道合胞体病毒 B (RSV B) 的微阵列实验 (在这种具体情况下, 总计 5 个微阵列实验)。

[0103] 对扩增效率建立 RT-PCR 模型

[0104] 本发明的方法是采用引物用于改良的逆转录的方法 (Sung 等, 2003, CSB), 该方法包括固定的寡核苷酸标签 (头部) 和随机寡核苷酸尾部。理论上, 随机寡核苷酸尾部应当不加区分地与患者样品内的全部核酸结合, 启动第一链合成。在第二链合成后, 全部逆转录的序列将在两端具有固定的寡核苷酸标签 (头部)。这些序列使用固定的寡核苷酸标签 (头部) 作为引物, 通过 PCR 扩增以生成长度至少为 10000bp 的 PCR 产物。尤其是, 所扩增的 PCR 产物的长度大部分在 $500-1000\text{bp}$ 之间。根据所述特定实施方案, 用于逆转录的 (RT) 26mer 引物包含固定的 17mer 标签和 9mer 的随机尾部: $5'-\text{GTTTCCCAGTCACGATANNNNNNNNN}-3'$ (SEQ ID NO:1)。

[0105] 在我们的模型中, v_a 代表临床样品中的病原体。为了在基因组内的任何区域中生成至少一种由 v_a 的位置 i 和 j 所定义的 PCR 产物, 例如 500-1000bp, 需要结合至位置 i 的正向引物和以反义方向结合至位置 j 的反向引物, 且使 $500 \leq |i-j| \leq 10000$, 尤其是 $500 \leq |i-j| \leq 1000$ 。引物的结合亲和力由至少两种因素决定: (1) 引物二聚体形成, 和 (2) 引物对病毒 v_a 的杂交亲和力。通过对如图 1 所示的 v_a 的每个位置计算扩增效率得分 (AES) 可以预测在 10000 个核苷酸范围内, 尤其在 1000 个或 500 个核苷酸范围内因具有理想引物结合位置而被成功扩增的基因组区域。

[0106] 扩增效率得分 (AES)

[0107] 对 v_a 的每个位置 i , 设 $P^f(i)$ 和 $P^r(i)$ 是随机引物 r_i 作为正向引物和反向引物分别与 v_a 的位置 i 可能结合的概率。为简便起见, 我们假设随机引物仅当该引物的随即尾部 (例如, 如图 1 内所示随机引物的最后 9 个核苷酸) 是 v_a 的反向互补物 (正向引物) 的子串或是 v_a (反向引物; 图 1) 的子串时才可以与 v_a 结合。基于非常确定的引物设计标准 (Wu, and Ugozzoli, 1991), 若 r_i 形成明显的引物二聚体或具有极端的解链温度, 则我们评定 $P^f(i)$ 为低。另一方面, 若 r_i 未形成任何明显的引物二聚体并具有最佳的解链温度, 则 $P^f(i)$ 将被评定为高。若随机引物的固定的寡核苷酸标签 (头部) (例如, 图 1 所示固定的 17mer 标签) 与 v_a 相似, 这也可以辅助结合并且因此产生更高的 $P^f(i)$ 。类似地, 我们计算了 $P^r(i)$ 。

[0108] 随机引物 r_i 作为正向引物在 v_a 的位置 i 的结合影响位置 i 上游的核苷酸 (例如位置 i 上游 500 至 1000 个核苷酸) 的 RT-PCR 产物的质量。类似地, 随机引物 r_i 作为反向引物在 v_a 的位置 i 的结合影响位置 i 下游的核苷酸 (例如位置 i 下游 500 至 1000 个核苷酸) 的 RT-PCR 产物的质量和覆盖度。假定 v_a 有位置 x 。位于位置 i 和 j 内的全部有效引物对分别影响在 x 处的 RT-PCR 产物的质量。需要注意的是 $i \geq x \geq j$ 并且 $i-j \leq 10000$ 。例如, $500 \leq i-j \leq 1000$, 因为我们的 RT-PCR 产物的长度是 500 至 1000 碱基对。因此, 对 v_a 的每个位置 x 的扩增效率得分 AES_x 可以通过考虑用于扩增位置 x 的全部引物对的总效果计算:

$$[0109] \quad AES_x = \sum_{j=x-1000}^x \left\{ P^f(j) \times \sum_{k=\max(x+1, j+500)}^{j+1000} P^r(k) \right\}$$

[0110] 预测成功的 RT-PCR 的 AES 阈值

[0111] 用于病毒 v_a 的探针选择的扩增效率得分的阈值由 v_a 的 AES 值的累积分布函数测定。设 X 是代表 v_a 的所有探针的 AES 值的随机变量。设 k 是 v_a 探针的数目。随后, 我们指定 AES 值小于或等于 x 的概率是 $P(X \leq x) = c/k$, 其中 c 是具有小于或等于 x 的 AES 值的探针的数目。对于在 v_a 的位置 i 的探针 p_i , 设 x_i 是该探针的相应 AES 值。由于探针的信号强度与它的 AES 值高度相关, 我们认为 $P(p_i | v_a)$, 即 p_i 在 v_a 存在下具有高信号强度的概率, 是 $P(X \leq x_i)$ 。因此,

$$[0112] \quad P(p_i | v_a) \approx P(X \leq x_i) = \frac{c_i}{k}$$

[0113] 其中 c_i 是 AES 值小于或等于 x_i 的探针的数目。

[0114] 对于探针的选择而言, 若 $P(p_i | v_a) > \lambda$, 则选择探针 p_i 。在我们的实验中, 设定 $\lambda = 0.8$ 。在该阈值 (AES 的最高 20%), 观察到多于 50% 的预期探针可重复地与不同临床样品杂交。当使用具有更高 AES (例如, AES 的最高 10%) 的探针将改善可重复性, 与此同时

这将在物种水平上降低对某些基因组仍为独特的探针的数目至 < 10 , 因此有损此阵列特异性鉴定病原体的能力。因此使用 AES 的最高 20%。

[0115] 在病原体检测微阵列上实验性地测定交叉杂交阈值：

[0116] 探针的设计

[0117] 设计能够与所选区域杂交的寡核苷酸探针的步骤 (ii) 可以选择为任何一种本领域内已知的探针设计技术。以下描述涉及探针设计, 然而, 对于技术人员来说, 显然还可采用相同的原理用于设计引物, 尤其用于 RT-PCR 设计引物。

[0118] 例如, 给定一套靶核酸 (例如, 病毒基因组) $V = \{v_1, v_2, \dots, v_n\}$, 对于每个 $v_i \in V$, 可以考虑设计一套长度为 m 的满足以下条件, 例如至少一种以下条件, 的探针 (其是 v_i 的子串)：

[0119] (a) 已建立的探针设计标准, 即同质性、灵敏性和特异性 (Sung, W. K. 等, 2003, CSB)；

[0120] (b) 与人基因组无显著的序列相似性；和

[0121] (c) 使用 AE 得分高效扩增, 例如通过如本文中所述的 RT-PCR。

[0122] 由交叉杂交人为产物所致的噪音性信号对于解读微阵列数据、特别对鉴定存在于复杂核酸混合物内的稀有病原体序列设置了巨大障碍。例如, 在临床标本中, 那些来自宿主组织中的核酸序列等污染的核酸序列, 将在序列互补性的某阈值以上与病原体特异性微阵列探针交叉杂交。这可能产生导致错误结论的假阳性信号。类似地, 病原体序列除了结合它的特异性探针外, 还可以与其它非靶探针 (即设计旨在检测其它病原体的探针) 交叉杂交。尽管后一现象似乎带来问题, 却可能为病原体鉴定提供有用信息以可以精确预测交叉杂交。由于微阵列探针具有评定复性潜能和序列特异性的诸多优点, 因此通常将它设计为确保 (对已知靶的) 最大的特异性杂交, 同时具有 (对非特异性序列的) 最小的交叉杂交。然而实践中, 我们发现尽管使用最佳计算机模拟参数来设计, 因未知原因, 众多探针未如预期那样表现性能。

[0123] 为了系统研究基于阵列的病原体检测的动力学, 我们使用宁布勒根 (Nimblegen) 阵列合成技术 (Nuwaysir 等, 2002) 创造寡核苷酸阵列。使用 40mer 探针而设计的所述阵列检测多达 35 种 RNA 病毒, 其中, 所述的探针以平均 8 个碱基的解析度在每种基因组全部长度范围内嵌搭 (tilled) (53, 555 个探针; 图 6, 表 1)。

[0124] 表 1 在病原体检测微阵列上代表的基因组列表。

[0125] (第 1 列) 对微阵列上合成的每种基因组的探针的数目。(第 2 列) 应用探针设计筛选程序 (filter) 后剩下的对每种基因组的探针的数目。(第 3 列) 对每种基因组的探针的数目, 其中所述探针对所述基因组是独特的并且不与人的基因组交叉杂交。

[0126]

基 因 组	探 针 的 原 始 数 目 (1)	筛 选 后 探 针 的 数 目 (2)	独 特 探 针 (3)	NCBI GI 编号	参 考 类 型	编 号	描 述
1	1948	537	271	9629198	RefSeq	NC_001781.1	人呼吸道合胞体病毒， 全基因组
2	1995	550	295	19718363	RefSeq	NC_003461.1	人副流感病毒 1 株 华盛顿 / 1964, 全基因组
3	2002	762	474	19525721	RefSeq	NC_003443.1	人副流感病毒 2, 全基因组
4	1979	701	345	10937870	RefSeq	NC_001796.2	人副流感病毒 3, 全基因组
5	3805	588	444	30468042	Genbank	AY283794.1	SARS 冠状病毒 Sin2500, 全基因组

6	3937	604	356	38018022	RefSeq	NC_005147.1	人冠状病毒 OC43, 全基因组
7	3495	182	112	12175745	RefSeq	NC_002645.1	人冠状病毒 229E, 全基因组
8	1705	292	177	46852132	RefSeq	NC_004148.2	人偏肺炎病毒, 全基因组
9	296	118	101	8486138	RefSeq	NC_002023.1	甲型流感病毒 RNA 片段 1, 全序列
10	282	69	42	8486136	RefSeq	NC_002022.1	甲型流感病毒 RNA 片段 3, 全序列
10	296	81	54	8486134	RefSeq	NC_002021.1	甲型流感病毒 RNA 片段 2, 全序列
10	110	69	57	8486131	RefSeq	NC_002020.1	甲型流感病毒 RNA 片段 8, 全序列

[0127]

10	196	71	62	8486129	RefSeq	NC_002019.1	甲型流感病毒 RNA 片段 5, 全序列
----	-----	----	----	---------	--------	-------------	-------------------------

10	177	75	59	8486127	RefSeq	NC_002018.1	甲型流感病毒 RNA 片段 6, 全序列
10	225	70	51	8486125	RefSeq	NC_002017.1	甲型流感病毒 RNA 片段 4, 全序列
10	300	105	48	8486164	RefSeq	NC_002204.1	乙型流感病毒 RNA-1, 全序列
10	293	113	74	8486148	RefSeq	NC_002205.1	乙型流感病毒 RNA-2, 全序列
10	279	94	59	8486150	RefSeq	NC_002206.1	乙型流感病毒 RNA-3, 全序列
10	237	70	53	8486152	RefSeq	NC_002207.1	乙型流感病毒 RNA-4, 全序列
10	232	90	82	8486154	RefSeq	NC_002208.1	乙型流感病毒 RNA-5, 全序列

10	195	64	32	8486156	RefSeq	NC_002209.1	乙型流感病毒 RNA-6, 全序列
10	150	47	37	8486159	RefSeq	NC_002210.1	乙型流感病毒 RNA-7, 全序列
10	136	59	50	8486161	RefSeq	NC_002211.1	乙型流感病毒 RNA-8, 全序列
11	1401	85	54	11528013	RefSeq	NC_001563.2	西尼罗病毒,全基因组
12	1389	145	123	9627244	RefSeq	NC_002031.1	黄热病毒,全基因组
13	2335	235	171	13559808	RefSeq	NC_002728.1	聂帕病毒 (Nipah), 全基因组
14	1943	244	211	11545722	RefSeq	NC_002617.1	新城疫病毒,全基因组
15	1174	208	128	9629357	RefSeq	NC_001802.1	人免疫缺陷症病毒 1, 全基因组
16	409	134	106	21326584	RefSeq	NC_003977.1	乙型肝炎病毒,全基因组

[0128]

17	1011	169	135	9627257	RefSeq	NC_001576.1	人乳头瘤病毒型 10, 全基因组
18	1036	325	299	10445391	RefSeq	NC_002554.1	口蹄疫病毒 C, 全基因组
19	1246	211	209	9790308	RefSeq	NC_001545.1	风疹病毒, 全基因组
20	955	309	172	9626732	RefSeq	NC_001489.1	甲型肝炎病毒, 全基因组

21	834	103	29	38371716	RefSeq	NC_005222.1	汉坦病毒,全基因组
22	837	188	98	38371727	RefSeq	NC_005217.1	辛诺柏病毒,全基因组
23	430	100	86	23334588	RefSeq	NC_004294.1	淋巴细胞脉络丛脑膜炎 病毒片段 S,全序列
23	853	455	286	23334585	RefSeq	NC_004291.1	淋巴细胞脉络丛脑膜炎 病毒片段 L,全序列
24	1404	204	122	9626460	RefSeq	NC_001437.1	日本脑炎病毒,基因组
25	1370	284	91	51850386	日本 DNA 数 据库	ABI89128.1	登革病毒病毒 3 型基因组 RNA, 全基因组, 株 : 98902890 DF DV-3
26	1361	130	57	12659201	基因银行 (Genbank)	AF326573.1	登革病毒病毒 4 型, 株 814669, 全基因组
27	1370	142	21	19744844	Genbank	AF489932.1	登革病毒病毒 2 型, 株 BR64022, 全基因组

28	1370	152	52	323660	Genbank	M87512.1	DENTISEQ 登革病毒病 毒 1 型, 全基因组
29	944	175	87	9626436	RefSeq	NC_001430.1	人肠道病毒 D, 全基因组
30	945	183	122	9626433	RefSeq	NC_001428.1	人肠道病毒 C, 全基因组
31	946	196	148	9627719	RefSeq	NC_001612.1	人肠道病毒 A, 全基因组
32	945	364	154	21363125	RefSeq	NC_003986.1	致肠细胞病变人孤儿病 毒 1, 全基因组
33	944	94	12	9626677	RefSeq	NC_001472.1	人肠道病毒 B, 全基因组
34	913	283	190	9627730	RefSeq	NC_001617.1	人鼻病毒 89, 全基因组
35	920	426	291	9626735	RefSeq	NC_001490.1	人鼻病毒 B, 全基因组

[0129] 包括对每种病毒探针的 7 个重复探针, 和用于阵列合成及杂交 (如下所述) 的对照序列, 该阵列总共含有 390, 482 个探针。

[0130] 同质性、灵敏性和特异性

[0131] 同质性需要选择具有相似解链温度的探针。据发现低 CG 含量的探针不会产生可

靠的杂交信号强度,而高 CG 含量的探针通过非特异性结合易于产生高信号强度。因此,可能建立的是所选探针的 CG 含量应当是 40%至 60%。

[0132] 因此,本发明提供设计用于核酸检测的寡核苷酸探针的方法,该方法包括选择 CG 含量为 40%至 60%的探针。

[0133] 术语“杂交”是指其中寡探针非共价地与靶核酸或其部分结合以形成稳定双链的过程。三链杂交在理论上也是可能的。杂交探针是能够以碱基特异性方式与靶核酸的互补链结合的寡核苷酸。特异性杂交是指,当序列存在于 DNA 或 RNA 的复杂混合物中(例如总细胞的)时,在严格条件下分子基本上与或仅与特定的核苷酸序列或序列结合、形成双链体或杂交。杂交,例如等位基因特异性探针杂交,通常在严格条件下进行。例如,其中盐浓度不高于约 1 摩尔浓度(M)并且温度是至少 25℃的条件是严格条件,所述盐浓度为,例如 750mM NaCl,50mM 磷酸钠,5mM EDTA,pH 7.4(5 倍浓度 SSPE);所述温度为约 25℃至约 30℃。杂交通常在严格条件下进行,例如在盐浓度不高于 1M 并且温度至少 25℃下。对于严格条件,还可参考,例如,Sambrook 和 Russel,Molecular Cloning :A Laboratory Manual,Cold Springs Harbor Laboratory, New York(2001),为以上的所有目的将该文献全部内容引入本文作为参考。

[0134] 灵敏性需要选择不能形成大量的二级结构的探针,以便检测低丰度的 mRNA。因此,根据最邻近模型计算的最高自由能选择探针(SantaLucia, J., Jr. 等,1996)。

[0135] 因此本发明提供设计至少一种用于核酸检测的寡核苷酸探针的方法,其中,根据最邻近模型计算的最高自由能选择探针。

[0136] 特异性需要选择对病毒基因组最独特的探针。这将使探针与其它非靶核酸(例如,病毒基因组)的交叉杂交最小化。假定探针 s_a 和探针 s_b 分别是靶核酸 v_a 和 v_b 的子串,则根据 s_a 与来自靶核酸 v_b 的长度为 m 的任意子串 s_b 之间的汉明距离和 / 或 s_a 与探针 s_b 的最长公共子串来选择 s_a 。尤其,设 s_a 和 s_b 分别是来自病毒基因组 v_a 和 v_b 的长度为 m 的子串,其中 $v_a \neq v_b$ 。

[0137] 待设计的探针的长度可以是对本发明的目的有用的任何长度。探针可以小于 100mer,例如 20 至 80mer;25 至 60mer,例如 40mer。汉明距离和 / 或最长公共子串也可以变化。

[0138] 根据凯恩氏(Kane's)标准(Kane, M. D., 等,2000), s_a 对 v_a 是特异性的,若:

[0139] (a) s_a 与来自病毒基因组 v_b 的长度为 m 的任意子串 s_b 间的汉明距

[0140] 离大于 $0.25m$;

[0141] (b) s_a 和 s_b 的最长公共子串小于 15。

[0142] 用于汉明距离的界限值可以根据所需的严格性进行选择。对于任何技术人员来说,如何根据所需的特定严格性选择汉明距离界限是显而易见的。根据本文中所述的探针设计的具体实例,本发明人对特异性探针使用相对其它靶核酸的 > 10 的汉明距离界限值,并且对保守性探针使用 < 10 、优选 < 5 的汉明距离界限值。对于特异性探针而言,这表示仅与特异性靶核酸杂交的探针,而对于保守性探针而言,这表示可以与靶核酸家族的任何成员杂交的探针。

[0143] 因此,本发明还提供设计用于核酸检测的寡核苷酸探针的方法,其中,假定探针 s_a 和探针 s_b 分别是生物样品中包含的靶核酸 v_a 和 v_b 的子串,若 s_a 与来自靶核酸 v_b 的长度为

m 的任意子串 s_b 间的汉明距离大于 $0.25m$, 并且 s_a 与探针 s_b 的最长公共子串小于 15, 则选择 s_a 。

[0144] 为了在没有来自人 RNA 的交叉杂交复杂情况下研究阵列杂交动力学, 将 SARS 冠状病毒及登革病毒血清型 1 的病毒 RNA 从感染细胞系的培养基中纯化、逆转录并使用病毒特异性引物进行 PCR 扩增 (Wong, 等, 2004)。每种基因组 cDNA 被完整地扩增 (通过测序证实), 用 Cy3 标记并在微阵列上分别杂交。SARS 样品很好地与 SARS 嵌搭探针杂交, 全部 3,805 个 SARS 特异性探针显示远高于检测阈值 (检测阈值由探针信号强度在阵列信号强度均值之上 > 2 标准差而确定; 图 7A) 的荧光 (Cy3) 信号。与其它病原体探针组的交叉杂交极少, 仅对冠状病毒科其它成员和微 RNA 病毒科和副粘病毒科的几个物种观察到交叉杂交, 这同 SARS 与其它已知病毒很少有序列同源性的观察相一致 (Ksiazek 等, 2003)。另一方面, 登革病毒 1 的杂交模式更为复杂 (图 7B)。首先, 由于序列多态性, 我们观察到与登革病毒 1 探针组的杂交是局部不完整的 (即区域无信号)。在阵列上杂交的登革病毒 1 样品是从 1944 年的夏威夷分离株 (ATCC 登记号 #VR-1254) 中培养出来, 而阵列探针组以 1990 年在新加坡分离的 S275/90 株 (Fu 等, 1992) 的序列为基础。未与 cDNA 靶杂交的登革病毒 1 探针分别含有对靶序列的至少 3 个错配 (在 15 碱基的片段内)。其次, 我们观察到在某种程度上与阵列上存在的几乎全部病毒探针组, 尤其与其它黄病毒科成员的探针发生交叉杂交, 这与 4 种登革病毒血清型共享 60-70% 同源性的事实相一致。为理解杂交信号输出与复性特异性间的关系, 我们使用 2 种相似性度量: 即探针汉明距离 (HD) 与最大连续匹配 (MCM), 首先将全部探针序列与每种病毒基因组比较。HD 测量两种序列的整体相似性距离, 对相似序列产生低得分 (Hamming, 1950)。MCM 测量完全匹配的连续碱基数目, 对相似序列产生高分 (Kane 等, 2000)。

[0145] 我们计算每个探针相对于夏威夷登革病毒 1 分离株的 HD 和 MCM 得分并且观察到这些得分分别反向地并且直接地与探测信号强度相关。阵列上与夏威夷登革病毒 I 基因组具有高相似性即 $HD = 2 (n = 942)$ 或 $MCM = 27 (n = 627)$ 的所有探针以高于背景 3 个对数值的中等信号强度发生杂交。虽然 98% 的探针在 0-4 的低 HD 范围内或 18-40 的高 MCM 范围内是可检测到的, 然而中等探测信号强度随序列距离的每个增量而下降。中等信号强度在 $HD = 7$ 和 $MCM = 15$ 急剧下降至背景水平, 分别具有 43% 和 46% 的可检测探针。绝大多数探针 ($> 96\%$, $n > 51,000$) 具有 8-21 的 HD 得分和 / 或 0-15 的 MCM 得分, 其中可分别检测到 1.23% 和 1.57% 的这些探针。

[0146] 理想的交叉杂交相似性阈值为这样的值, 即其中鉴定特定病原体的所有探针总是具有高于背景噪音的可检测信号强度, 即使在病原体序列内存在多态性时也是如此。在最佳的相似性阈值 $HD = 4$ 和 $MCM = 18$, $> 98\%$ 的探针可以以高于背景 2 个对数值的中等信号强度被检测到, 而调整阈值下降 1 级至 $HD = 5$ 和 $MCM = 17$ 将仅产生 $\sim 85\%$ 探针检测和高于背景 ~ 1.2 个对数值的中等信号强度 (图 8)。

[0147] 使用这些最佳 HD 和 MCM 阈值来预测交叉杂交, 我们将全部探针分成最有可能检测给定病原体的组。我们将这些组称作特异性标签探针组 (SPSs), 并且我们为阵列上所代表的 35 种病原体基因组中的每种基因组定义了 SPSs (表 2)。

[0148] 表 2 每种病原体标签探针组 (SPS) 包含具有最高的 20% 的 AES 的探针 [第 (1) 列]。剔除了 GC 含量不为 40-60% 的探针 [第 (2) 列] 或与人基因组具有高度相似性的

探针 [第 (3) 列]。将来自于其它病原体的、根据 HD 和 MCM 将与该病原体交叉杂交的探针 [第 (4) 列] 添加至 SPS [第 (5) 列]。

[0149]

	病原体	科	总的 嵌搭 探针	AES (1)	GC 含量 筛选 程序 (2)	人 基 因 组 筛 选 程 序 (3)	筛选 留下 的探 针的 数目	预测的交叉 杂交性探针 的数目 ($HD \leq 4$ 和 $MCM \geq 18$) (4)	SPS 内的 探针 的数 目 (5)
1	LCMV	砂粒病毒科	1283	574	1	18	555	0	555
2	汉坦病毒	布尼安病毒科	834	131	6	22	103	2	105
3	辛诺柏病毒	布尼安病毒科	837	225	8	29	188	3	191
4	229E	冠状病毒科	3495	196	2	12	182	2	184
5	OC43	冠状病毒科	3937	663	16	43	604	3	607
6	SARS	冠状病毒科	3805	672	6	78	588	3	591
7	登革病毒血清型 1	黄病毒科	1370	201	2	47	152	50	202
8	登革病毒血清型 2	黄病毒科	1370	178	0	36	142	71	213
9	登革病毒血清型 3	黄病毒科	1370	336	1	51	284	69	353
10	登革病毒血清型 4	黄病毒科	1361	172	1	41	130	44	174
11	日本肝炎病毒	黄病毒科	1404	274	6	64	204	40	244
12	西尼罗病毒	黄病毒科	1401	111	4	22	85	22	107
13	黄热病毒	黄病毒科	1389	151	0	6	145	10	155
14	乙型肝炎病毒	嗜肝 DNA 病毒科	409	146	2	10	134	0	134
15	甲型流感病毒	正粘病毒科	1582	601	2	46	553	0	553
16	乙型流感病毒	正粘病毒科	1822	718	7	69	642	2	644
17	人乳头瘤病毒型 10	乳多空病毒科	1011	177	1	7	169	0	169
18	hMPV	副粘病毒科	1705	375	23	60	292	8	300
19	新城疫病毒	副粘病毒科	1943	252	0	8	244	0	244
20	Nipah	副粘病毒科	2335	274	22	17	235	0	235
21	副流感病毒 1	副粘病毒科	1995	625	13	62	550	3	553
22	副流感病毒 2	副粘病毒科	2002	838	31	45	762	0	762
23	副流感病毒 3	副粘病毒科	1979	834	29	104	701	9	710
24	RSV B	副粘病毒科	1948	655	52	66	537	4	541
25	致肠道细胞病变人孤儿病毒 1	微 RNA 病毒科	945	439	3	72	364	59	423
26	肠道病毒 A	微 RNA 病毒科	946	205	0	9	196	21	217
27	肠道病毒 B	微 RNA 病毒科	944	109	0	15	94	47	141
28	肠道病毒 C	微 RNA 病毒科	945	202	0	19	183	31	214

29	肠道病毒 D	微 RNA 病毒科	944	191	0	16	175	15	190
30	口蹄疫病毒	微 RNA 病毒科	1036	356	26	5	325	0	325
31	甲型肝炎病毒	微 RNA 病毒科	955	355	9	37	309	0	309
32	鼻病毒 A(型 89)	微 RNA 病毒科	913	333	2	48	283	13	296
33	鼻病毒 B	微 RNA 病毒科	920	464	3	35	426	11	437
34	HIV 1	逆转录病毒科	1174	229	4	17	208	0	208
35	风疹病毒	披盖病毒科	1246	748	534	3	211	0	211
		总计	53555				10955		11497

[0150] 每种病原体的 SPS 包含从其基因组序列中衍生的嵌搭探针 (HD = 0, MCM = 40) 以及从其它病原体中衍生的交叉杂交性探针 (HD = 4, MCM = 18)。

[0151] 随后,我们考虑了可能影响我们的 SPS 探针性能的其它非特异性杂交现象。例如,我们观察到探针信号与 % GC 含量间的一般关系。与先前观察一致,我们发现 GC 含量 < 40% 的探针产生减弱的信号强度,而 GC 含量 > 60% 的探针显示更高的信号强度 (Wong 等, 2004; Maskos 和 Southern, 1993)。因此,我们利用 % GC 含量作为额外的选择筛选程序,而从我们的 SPS 中排除 GC 含量 < 40% 及 GC 含量 > 60% 的探针,尽管这些探针具有最佳的 HD 和 MCM 值。

[0152] 与人基因组的序列相似性

[0153] 在待检测的靶核酸从人提取的情况下 (例如,含有病毒基因组的人样品),还应当避免使用与人基因组具有高度同源性的探针。因此,就对靶核酸 v_a 为特异的长度为 m 的任何探针 s_a 而言,若探针 s_a 与异于靶核酸的核酸的任何区域无任何标的,则选择探针 s_a ,若长度为 m 的探针 s_a 与异于靶核酸的核酸有标的,则选择长度为 m 的具有最小的最大比对长度和 / 或具有最低标的数的探针 s_a 。尤其是,对于任何长度为 m 的探针 s_a ,用 BLAST 算法找到 s_a 对人基因组的标的 (Altschul, S. F. 等, 1997)。使用 BLAST 字段大小 ($W = 15$) 和期望值 100 来找到全部标的。若 s_a 与人基因组无任何标的,则选择 s_a ,也就是说, s_a 对 v_a 是特异性的。然而,若 v_a 的长度为 m 的全部子串与人基因组有标的,则选择具有最小的最大比对长度和具有最低标的数的那些子串。

[0154] 此外由于与人序列的交叉杂交也可能使结果混乱,我们通过 BLAST 使用字段大小为 15 (Altschul 等, 1997) 将全部探针与人基因组装配物 (build17) (International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. Nature 409 (6822), 860-921 (2001).) 进行比较。从 SPS 中进一步筛选具有期望值为 100 的探针 (见上表 2)。

[0155] 因此,本发明提供设计用于核酸检测的寡核苷酸探针的方法,其中,就对靶核酸 v_a 特异的长度为 m 的任何探针 s_a 而言,若探针 s_a 与异于靶核酸的核酸的任何区域无任何标的,则选择探针 s_a ,并且若长度为 m 的探针 s_a 与异于靶核酸的核酸有标的,则选择长度为 m 的具有最小的最大比对长度和 / 或具有最低标的数的探针 s_a 。

[0156] 此外,寡核苷酸探针设计还可以通过本发明的 AES 实施。尤其是,本发明提供选择和 / 或设计探针的方法,其中,若预测 p_i 与所扩增靶核酸的位置 i 杂交,则选择在靶核酸的位置 i 的探针 p_i 。

[0157] 尤其是,能够与所选区域杂交的寡核苷酸探针可以根据至少一种以下标准进行选

择和 / 或设计：

[0158] (a) 选择 CG 含量为 40% 至 60% 的探针；

[0159] (b) 选择具有根据最邻近模型计算的最高自由能的探针；

[0160] (c) 假定探针 s_a 和探针 s_b 分别是靶核酸 v_a 和 v_b 的子串，则根据 s_a 与来自靶核酸 v_b 的长度为 m 的任意子串 s_b 的汉明距离和 / 或 s_a 与探针 s_b 的最长公共子串来选择 s_a ；

[0161] (d) 对靶核酸 v_a 特异性长度为 m 的任何探针 s_a 而言，若探针 s_a 与异于靶核酸的核酸的任何区域无任何标的，则选择探针 s_a ，并且若长度为 m 的探针 s_a 与异于靶核酸的核酸有标的，则选择长度为 m 的具有最小的最大比对长度和 / 或具有最低命标的数的探针 s_a ；和 / 或

[0162] (e) 若预测 p_i 与所扩增靶核酸的位置 i 杂交，则选择在靶核酸的位置 i 的探针 p_i 。

[0163] 根据本发明的具体方面，可以使用以上所述标准中的两项或多项设计寡核苷酸探针。例如，探针可以通过应用全部标准 (a) 至 (e) 进行设计。也可以使用在本文中沒有明确提及但本领域技术人员知道的其它标准。

[0164] 尤其是，根据标准 (e)，选择在靶核酸 v_a 的位置 i 的探针 p_i ，若 $P(p_i | v_a) > \lambda$ ，其中 λ 是 0.5 并且 $P(p_i | v_a)$ 是 p_i 应当与靶核酸 v_a 的位置 i 杂交的概率。更尤其是， λ 是 0.8。

[0165] 根据另一方面，本发明提供如上所述的方法，其中， $P(p_i | v_a) \approx P(X \leq x_i)$ 其中， X 是代表 v_a 的所有探针的扩增效率得分 (AES) 值的随机变量， k 是 v_a 中探针的数目，并且 c_i 是其 AES 值 $\leq x_i$ 的探针的数目。

[0166] 根据另一方面，AES 还可以用来设计随机引物标签以辅助通过随机 PCR 对样品进行随机扩增（这种应用如检测病原体、检测基因表达、构建克隆性 DNA 文库，和技术人员会采用随机 PCR 的其它应用）。

[0167] 在支持物上合成寡核苷酸探针

[0168] 根据本发明的另一方面，选择和 / 或设计如上所述的至少一种寡核苷酸探针的方法还包括制备选择的和 / 或设计的探针的步骤。设计探针包含通过任何合适的手段，例如通过使用软件，了解该探针的序列和 / 或设计该探针。制备探针的步骤包含实际地制备探针。可以根据本领域已知的任何标准方法制备探针。例如，探针可以是化学合成的或通过克隆法制备的。例如，如 Sambrook 和 Russel, 2001 所述。

[0169] 本发明还提供根据本发明的任何实施方案所制备的支持物，例如微阵列或生物芯片。

[0170] 根据本发明的任何方法所设计和制备的探针可以在溶液中使用或可以被置于不溶性支持物上。例如，可以根据本领域已知的任何技术将探针施加到、点滴到或刷在不溶性支持物上。支持物可以是固体支持物或凝胶。被施加探针的支持物可以是微阵列或生物芯片。

[0171] 更具体地，本发明提供用于从 PCR 扩增的 cDNA 中快速检测并鉴定病原体（例如病毒和 / 或细菌病原体）的基于寡微阵列杂交的方法，其中所述的 cDNA 是从原始组织样品中制备的。尤其是，该方法应用于从随机 PCR 扩增的 cDNA 中快速检测并鉴定病原体。

[0172] 在以下描述中，探针的制备具体参考微阵列而进行。然而，支持物以及探针可以根据本申请的全部内容中的任何描述制备。尤其是，“阵列”是有目的创造的可以通过合成或生物合成制备的分子集合。阵列中的分子可以是彼此相同或不同的。阵列可以采用多种形

式,例如,可溶性分子的文库;与树脂珠连接的化合物的文库、二氧化硅芯片或其它固体支持物。阵列平台或平台是具有多个阵列的物体,在该物体上每个阵列由抗液体通过的物理屏障与其它阵列分隔并形成区域及空间,称为“孔”。

[0173] 样品制备和杂交到微阵列上

[0174] 生物样品可以是采自哺乳动物的任何样品,例如来自人类。生物样品可以是血液、体液、唾液、尿、粪便等。生物样品可以是在实施扩增步骤前被处理以释放生物样品中包含的核酸。靶核酸可以是想要检测的任何核酸。待检测的靶核酸可以至少是对生物样品的核酸为外源的核酸。因此,若生物样品来自人,待检测的外源靶核酸(若其存在于生物样品内)是非人源的核酸。根据本发明的方面,待检测的靶核酸至少是病原体的基因组或基因组片段。病原体核酸可以至少是来自病毒、寄生虫或细菌的核酸或核酸片段。

[0175] 根据本发明的一个方面,本发明提供靶核酸检测分析的方法。该方法可以是用于检测生物样品内病原体存在的诊断方法。来自生物样品的想要检测的靶核酸可以是任何靶核酸、RNA 和 / 或 DNA。例如,mRNA 和 / 或 cDNA。更具体地,待检测的靶核酸可以是病原体或非病原体。例如,它可以是至少一种病毒、至少一种细菌和 / 或至少一种寄生虫的基因组或基因组片段。对于本领域技术人员来说,可以根据已知的任何标准技术将选择和 / 或制备的探针放置、施加和 / 或固定在支持物上。支持物可以是不溶性支持物,例如固体支持物,尤其是微阵列和 / 或生物芯片。

[0176] 根据具体实施例,使用已建立的方法和商业试剂盒从患者样品,例如组织、血清、鼻咽冲洗液、粪便,中提取 RNA 和 DNA。例如,可以使用用于提取核酸的 Qiagen Kit。或者,酚 / 氯仿法也可以用于提取 DNA 和 / 或 RNA。可以使用本领域已知的任何技术,例如,如在 Sambrook 和 Russel,2001 中所述的技术。根据 Bohlander 等,1992 和 Wang 等,2003 描述的方法,使用标记的随机引物,将 RNA 逆转录成 cDNA。cDNA 随后通过随机 PCR 进行扩增。按照 Wong 等 2004 所述进行样品的片段化、标记以及将样品杂交至微阵列上。

[0177] 微阵列合成

[0178] 根据实施例部分所述的具体实验,本发明的发明人选择了代表在新加坡的病毒疾病中最常见病因的数种病毒基因组。使用从 Genbank 下载的全基因组序列,合成在全基因组范围内嵌搭并以 5 碱基解析度重叠的 40mer 探针。使用宁布勒根技术(Nuwaysir,E.F.,等,2002)在微阵列上直接合成每种病毒探针的 7 个复制品。探针在微阵列上随机分布,旨在使杂交人为产物的影响最小化。为控制样品对探针的非特异性杂交,设计并在微阵列上合成 10,000 个寡核苷酸探针。这 10,000 个寡核苷酸与人基因组或与病原体基因组没有任何序列相似性。它们是具有 40-60% CG 含量的随机探针。测量这些探针背景信号强度。作为阳性对照,在阵列上合成针对人基因的 400 个寡核苷酸探针,其中所述的人基因在免疫应答中具有已知或推测的功能。包含植物病毒 PMMV 作为对总计大约 380,000 个探针的阴性对照。在以下描述中,本发明将参考病原体检测芯片分析(又称作 PDC)进行更具体地说明。然而,分析(方法)不限于这种具体实施方案,但包括在本申请的全部内容中所述的本发明数个方面。

[0179] 检测靶核酸的方法

[0180] 根据另一方面,本发明提供检测至少一种靶核酸的方法,该方法包括以下步骤:

[0181] (i) 提供生物样品;

[0182] (ii) 对该生物样品中所包含的核酸进行扩增；

[0183] (iii) 提供能够与至少一种假定存在于生物样品中的靶核酸杂交的至少一种寡核苷酸, 其中该探针是通过使用本文中所述的本发明任何方面的方法制备的；

[0184] (iv) 使该探针与扩增的核酸接触和 / 或检测与至少一种靶核酸杂交的探针。

[0185] 扩增步骤 (ii) 可以在随机引物、部分随机引物 (即包含固定的部分和随机的部分) 或特异性引物存在下进行。尤其是, 扩增步骤 (ii) 可以在至少一种随机引物存在下进行。更尤其是, 在至少一种随机正向引物和 / 或至少一种随机反向引物存在下进行。例如, 扩增步骤 (ii) 可以在多于两种随机引物的存在下进行。可以使用本领域已知的任何扩增方法。例如, 扩增方法是 RT-PCR。

[0186] 尤其是, 本发明的发明人基于扩增效率得分 (AES) 开发了检测与靶核酸杂交的探针的方法。在本文中, 该方法也可以称作本发明的算法。尤其是, 与靶核酸 v_a 的位置 i 结合的正向随机引物和与靶核酸 v_a 的位置 j 结合的反向随机引物选自对靶核酸 v_a 的每个位置 i 具有以下扩增效率得分 (AESI) 的引物：

$$[0187] \quad AES_i = \sum_{j=i-Z}^i \left\{ P^f(j) \times \sum_{k=\max(i+1, j+500)}^{j+Z} P^r(k) \right\},$$

$$[0188] \quad \text{其中} \quad \sum_{k=\max(i+1, j+500)}^{j+Z} P^r(k) = P^r(i+1) + P^r(i+2) + \dots + P^r(j+Z)$$

[0189] $P^f(i)$ 和 $P^r(i)$ 是随机引物 r_i 作为正向引物和反向引物分别与 v_a 的位置 i 结合的概率, 并且 $Z \leq 10000\text{bp}$ 是 v_a 内欲扩增的区域。更尤其是, Z 可以是 $\leq 5000\text{bp}$ 、 $\leq 1000\text{bp}$ 或 $\leq 500\text{bp}$ 。

[0190] 扩增步骤可以包含正向引物和反向引物, 并且正向引物和反向引物中的每种引物可以在 5'-3' 方向上包含固定的引物头部和可变的引物尾部, 并且, 其中至少可变的尾部与靶核酸 v_a 的一部分杂交。尤其是, 扩增步骤可以包含具有 SEQ ID NO: 1-7 中任意的核苷酸序列或该核苷酸序列的变体或衍生物的正向随机引物和 / 或反向随机引物。

[0191] 生物样品可以采自哺乳动物的任何样品, 例如来自人类。生物样品可以是组织、血清、鼻咽冲洗液、唾液、任何其它体液、血液、尿、粪便等。生物样品可以在进行扩增步骤前被处理以释放生物样品中包含的核酸。靶核酸可以是想要检测的任何核酸。待检测的靶核酸可以至少是对生物样品的核酸为外源的核酸。因此, 若生物样品来自人, 则待检测的外源靶核酸 (若其存在于生物样品中) 是非人源的核酸。根据本发明的方面, 待检测的靶核酸至少是病原体的基因组或基因组片段。病原体核酸可以至少是来自病毒、寄生虫或细菌的核酸或核酸片段。

[0192] 因此, 本发明提供检测生物样品中的至少一种靶核酸 (若存在) 的方法。该方法可以是用于检测生物样品中存在病原体的诊断方法。例如, 若生物样品从人类得到, 且如果该生物样品存在靶核酸, 则该靶核酸为非人源的。

[0193] 根据本发明的任何方法所设计和 / 或制备的探针可以在溶液中使用或可以被放置在不溶性支持物上。例如, 可以根据本领域已知的任何技术将探针施加到、点滴到或刷在不溶性支持物上。所述支持物可以是固体支持物或凝胶。尤其是, 被施加探针的支持物可

以是微阵列或生物芯片。

[0194] 随后将探针与生物样品的核酸接触,并且靶核酸(若存在)与探针杂交,并且检测靶核酸的存在。尤其是,在检测步骤(iv)中,与 v_a 杂交的探针的信号强度的均值在统计学上高于探针 $\notin V_a$ 的信号强度的均值,则表明生物样品中存在 v_a 。

[0195] 更尤其是,在检测步骤(iv)中,与 v_a 杂交的探针的信号强度的均值在统计学上高于探针 $\notin V_a$ 的信号强度的均值,并且该方法还包括以下步骤:计算具有高信号强度的探针 $\notin V_a$ 的比例与检测方法中所用的具有高信号强度的探针的比例的相对差异,探针 v_a 的信号强度的密度分布比探针 $\notin V_a$ 的信号强度的密度分布更为正偏,则表明生物样品中存在 v_a 。

[0196] 例如,在检测步骤(iv)中, t -检验值 ≤ 0.1 和/或安德森-塔林检验值 ≤ 0.05 和/或加权相对熵 ≥ 1.0 、优选 ≥ 5.0 时表示生物样品中存在靶核酸。尤其是, t -检验值 ≤ 0.05 。

[0197] 根据另一方面,本发明提供确定靶核酸 v_a 存在的方法,该方法包括检测探针与靶核酸 v_a 的杂交,并且其中与 v_a 杂交的探针的信号强度的均值在统计学上高于探针 $\notin V_a$ 的信号强度的均值,则表明存在 v_a 。尤其是,与 v_a 杂交的探针的信号强度的均值在统计学上高于探针 $\notin V_a$ 的信号强度的均值,并且该方法还包括以下步骤:计算具有高信号强度的探针 $\notin V_a$ 的比例与检测方法中所用的具有高信号强度的探针的比例的相对差异,探针 v_a 的信号强度的密度分布比探针 $\notin V_a$ 的信号强度的密度分布更为正偏,则表明生物样品中存在 v_a 。更尤其是, t -检验值 ≤ 0.1 和/或安德森-塔林检验值 ≤ 0.05 和/或加权相对熵 ≥ 1.0 、优选 ≥ 5.0 时表示生物样品中存在靶核酸。例如, t -检验值可以是 ≤ 0.05 。

[0198] 根据另一方面,本发明提供检测至少一种靶核酸方法,该方法包括以下步骤:

[0199] (i) 提供至少一种生物样品;

[0200] (ii) 对该生物样品中包含的核酸进行扩增;

[0201] (iii) 提供能够与至少一种假定存在于该生物样品中的靶核酸杂交的至少一种寡核苷酸;

[0202] (iv) 使该探针与扩增的核酸接触并且检测与靶核酸杂交的探针,其中与 v_a 杂交的探针的信号强度的均值在统计学上高于探针 $\notin V_a$ 的信号强度的均值,则表明生物样品中存在 v_a 。

[0203] 在步骤(iv)中,与 v_a 杂交的探针的信号强度的均值在统计学上高于探针 $\notin V_a$ 的信号强度的均值,并且该方法还包括以下步骤:计算具有高信号强度的探针 $\notin V_a$ 的比例与检测方法内所用的具有高信号强度的探针的比例的相对差异,探针 v_a 的信号强度的密度分布比探针 $\notin V_a$ 的信号强度的密度分布更为正偏,则表明生物样品中存在 v_a 。尤其是,在检测步骤(iv)中, t -检验值 ≤ 0.1 和/或安德森-塔林检验值 ≤ 0.05 和/或加权相对熵 ≥ 1.0 、优选 ≥ 5.0 时表示生物样品中存在靶核酸。 t -检验值可以是 ≤ 0.05 。待检测的靶核酸可以至少是对生物样品的核酸为外源的核酸。待检测的靶核酸可以至少是病原体的基因组或基因组片段。病原体核酸可以至少是来自病毒、寄生虫或细菌的核酸或核酸片段。尤其是,当样品从人类得到时,且该生物样品中存在靶核酸,则该靶核酸为非人源的基因组。探针可以被放置在不溶性支持物上。支持物可以是微阵列或生物芯片。

[0204] 使用 RSV B 模板序列的测试

[0205] 为验证在病毒的不同区域所表现的信号强度上的变异是否与它们相应的扩增效率得分直接相关,对感染人的常见病原体,人呼吸道合胞体病毒 B(RSV B),上进行总计 5 个

微阵列实验。

[0206] 其次,将上述的探针设计标准应用到从NCBI得到的RSV B的模板序列(NC_001781)。这产生了在每个阵列上点滴的1948个探针。在实际实验前还对RSV B的扩增效率图进行计算并表示在图2中。该图表示出具有高于平均AES的AES并显示具有更高扩增概率的RSV B的区域的峰。

[0207] 使用5种含有人呼吸道合胞体病毒B(RSV B)的样品,进行独立的微阵列实验。将每个这种实验得到的信号强度表示在图3中。

[0208] 对于每一个实验,将1948个探针的信号强度以递减顺序排序并与探针的对应AES值建立联系。发现p-值平均 $< 2.2 \times 10^{-16}$ 。这表明在RSV B的位置i的探针的信号强度与AES_i间的相关性并不是随机性的。进一步研究揭示在全部5个实验中始终产生高信号强度的约300个探针具有第90百分位水平上的扩增效率得分。

[0209] 在已经说明所述描述的扩增效率模型对RSV B基因组工作良好后,还需要说明的是本发明的模型可以延伸用于其它病毒基因组。对人偏肺炎病毒(HMPV)进行另一个微阵列实验。此次在微阵列上存在1705个探针。再次计算对HMPV的扩增效率图。在本次实验中,信号强度与扩增效率得分的相关性检验产生 1.335×10^{-9} 的p-值。

[0210] 因此,本发明的扩增效率模型能够预测在所述的实验方案中由病毒基因组的不同区域所产生的信号的相对强度。来自扩增效率得分低的区域中的探针极易于不产生信号强度或产生低信号强度。这将在微阵列上导致假阴性。此类探针将使微阵列数据的分析变得复杂,并且因为低信号强度的探针可能由于其靶基因组不存在或仅由于该探针未被扩增将使微阵列数据的分析变得更复杂。因此,应当选择区域内具有合理高的扩增效率得分的探针,以便使由于使用随机引物的RT-PCR方法所致的不精确性最小化。

[0211] 用于病毒 v_a 的探针选择的扩增效率得分阈值由 v_a 的AES值的累积分布函数确定。设X是代表 v_a 的全部探针的AES值的随机变量。设k是 v_a 中探针的数目。随后,我们指定AES值小于或等于x的概率是 $P(X \leq x) = c/k$,其中,c是具有小于或等于x的AES值的探针的数目。对于在 v_a 的位置i的探针 p_i ,设 x_i 是该探针的相应AES值。由于探针的信号强度与它的AES值高度相关,我们估计 $P(p_i | v_a)$,即 p_i 在 v_a 存在下具有高信号强度的概率,是 $P(X \leq x_i)$ 。因此,

$$[0212] \quad P(p_i | v_a) \approx P(X \leq x_i) = \frac{c_i}{k}$$

[0213] 其中 c_i 是其AES值小于或等于 x_i 的探针的数目。

[0214] 对于探针选择而言,若 $P(p_i | v_a) > \lambda$,则选择探针 p_i 。在本实验中,将 λ 设定为 $\lambda = 0.8$ 。

[0215] 因此,本发明还提供探针设计的方法和/或靶核酸检测的方法,其中,若 $P(p_i | v_a) > \lambda$,其中 λ 是0.75并且 $P(p_i | v_a)$ 是 p_i 在 v_a 存在下具有高信号强度的概率,则选择在靶核酸 v_a 的位置i上的探针 p_i 。更尤其是, $P(p_i | v_a) \approx P(X \leq x_i) = \frac{c_i}{k}$,其中X是代表 v_a 的全部探针的扩增效率得分(AES)值的随机变量,k是 v_a 的探针的数目并且 c_i 是其AES值小于或等于 x_i 的探针的数目。

[0216] 靶核酸检测分析

[0217] 在以下描述中,将参考病原体检测芯片分析(又称作PDC)更具体地说明本发明。

然而,分析(方法)不局限于这种具体实施方案,但包括在本申请全部内容中所述的本发明的数个方面。因此,尤其是,假定使用一套长度为 m 的探针 $P = \{p_1, p_2, \dots, p_1\}$ 的 PDC,其中所述探针设计用于一套病毒基因组 $V = \{v_1, v_2, \dots, v_n\}$,则病原体检测芯片分析问题将是根据芯片数据检测样品中存在的病毒。这里的芯片数据是指由 PDC 上的探针信号提供的综合信息。因此,芯片数据 $D = \{d_1, d_2, \dots, d_x\}$ 是 PDC 上的探针组 P 的对应信号组。

[0218] 给定一种样品,不知道何种病原体存在该样品中,存在多少种不同病原体(如果确实存在)。然而如果病毒 v_a 确实存在于样品内,则 v_a 的探针的信号强度应当与来自其它病毒的探针的信号强度明显不同。具体而言,与其它病毒相比,更高比例的 v_a 的探针应当具有高信号强度。因此,可以预测 v_a 的探针的信号强度的均值在统计学上应当高于探针 $\notin V_a$ 的信号强度的均值。

[0219] 因此,本发明提供与 v_a 杂交的探针的信号强度的均值在统计学上高于探针 $\notin V_a$ 的信号强度的均值从而可能表示生物样品中存在 v_a 的方法。

[0220] 然而,具有统计学上更高的均值仍可能不足以得出 v_a 存在于样品中的结论。优选地,可能需要额外的步骤。我们需要计算具有高信号强度的探针 $\in V_a$ 的比例与检测方法内所用的具有高信号强度的探针的比例的相对差异。这是根据观察到探针 $\in v_a$ 的信号强度的密度分布比探针 $\notin V_a$ 的信号强度的密度分布更为正偏的现象(见图 4A 中的箭头。为了比较,见图 4B)。

[0221] 基于以上观察,对病毒存在的芯片数据 D 按照以下分析。对于每种病毒 $v_a \in V$,我们使用单尾 T 检验 (Goulden, C. H., 1956) 来确定探针 $\in v_a$ 的信号强度的均值是否在统计学上高于探针 $\notin V_a$ 的信号强度的均值。因此,计算 t -统计量:

$$[0222] \quad t_i = \frac{\mu_a - \mu_{a'}}{\sqrt{\frac{\sigma_a^2}{n_a} + \frac{\sigma_{a'}^2}{n_{a'}}}}$$

[0223] 其中 μ_a 、 σ_a^2 和 n_a 分别是探针 $\in v_a$ 的信号强度的均值、方差和大小并且 $\mu_{a'}$ 、 $\sigma_{a'}^2$ 和 $n_{a'}$ 分别是探针 $\notin V_a$ 的信号强度的均值、方差和大小。

[0224] 为检验差异的显著性,将显著性水平设定为 0.05。这意指当 t_a 的 p -值 < 0.05 ,才可接受探针 $\in v_a$ 的信号强度的均值高于探针 $\notin V_a$ 的信号强度的均值的假定。在此情况下, v_a 有可能存在于样品内。

[0225] 单凭使本发明人知道病毒的信号强度分布是否不同于其它病毒的信号强度分布的 t -检验可能不足以确定特定的病毒是否存在于样品内。还必须知道两种分布相似或不同到何种程度。可以用来测量真实分布与模型分布之间相似性的度量是相对熵 (Kullback-Leiber, KL) (Kullback 和 Leiber, 1951)。在本申请中, v_a 中的探针的信号强度的概率分布是真实分布,而 P 中全部探针的信号强度的概率分布是模型分布。设 P_a 是 v_a 中的探针组。 P_a 和 P 的信号强度的概率分布的相对熵是:

$$[0226] \quad KL(P_a \parallel P) = \sum_{\mu \leq x \leq \max(D)} f_a(x) \log\left(\frac{f_a(x)}{f(x)}\right)$$

[0227] 其中 μ 是 P 内的探针的信号强度的均值; $f_a(x)$ 是 P_a 内的具有信号强度 x 的探针的分数;并且 $f(x)$ 是 P 内的具有信号强度 x 的探针的分数。接着,若 $KL(P_a \parallel P) = 0$,则 P_a 的概率分布与 P 的概率分布完全相同。否则,它们不相同。

[0228] 由于样品内存在的病毒具有高于群体信号强度的信号强度,这表明若 $KL(P_a \parallel P)$

> 0 , 则 v_a 有可能存在于样品内。因此, KL 值 ($P_a \parallel P$) 越大, 两种概率分布差异越大并且 v_a 确实存在于样品内的可能性越高。

[0229] 值得注意的是, 相对熵是在两种概率分布的全部 x 范围内的整体差异 (collective difference)。因此, 尽管相对熵擅长找到概率分布内的漂移, 但它并非总是善于找到对概率分布的尾部影响更大的发散。如图 4(A, B) 所述, 概率分布的尾部提供病毒是否存在于样品中的最多信息。因此, 相对熵统计量必须进行改良以更精确地反映此类观察。

[0230] 为提高相对熵在尾部上的灵敏性, 我们向相对熵中引入稳定化或加权的统计量即安德森-塔林统计量 (Stephens, M. A. (1974). EDF Statistics for Goodness of Fit and Some Comparisons, Journal of the American Statistical Association, 第 69 卷, 第 730-737 页)。因此加权相对熵是:

[0231]

$$WKL(P_a \parallel P) = \sum_{\mu \leq x \leq \max(D)} \frac{f_a(x) \log \frac{f_a(x)}{f(x)}}{\sqrt{Q(x)[1-Q(x)]}}$$

[0232] 其中 $Q(x)$ 是 P 内探针的信号强度的累积分布函数。

[0233] 实验检测表明在不存在病毒的样品中, 通过显著水平 0.05 的 t -检验的病毒具有 $WKL < 5.0$ 。在确实存在病毒的样品中, 实际病毒不仅通过显著水平 0.05 的 t -检验, 而且正是具有 $WKL \geq 5.0$ 的病毒。因此, 我们对存在于样品内的病毒设定加权相对熵阈值为 5.0。该分析流程图如图 5 所示。

[0234] 实施本发明方法的装置和 / 或产品

[0235] 对本领域技术人员来说, 如何配置可以实施本发明所提供的算法和 / 或方法的软件是熟知的。因此, 本发明还提供被配置成实施根据本发明任何实施方案的算法和 / 或方法的软件和 / 或计算机程序产品。还提供至少一种电子存储介质。电子存储介质可以是计算机硬盘驱动器、光盘驱动器 (CD-ROM)、闪存装置 (例如, 通用串行总线 (USB) 拇指驱动器 (thumbdrive))、软盘或本领域内任何其它电子存储介质。软件可以在个人计算机、电脑主机和任何计算处理单元上运行, 并且具体的配置是本领域技术人员公知的。

[0236] 应当理解的是仅通过举例的方式描述了本发明, 在设计中可以进行各种修改并不违背本发明的宗旨和范围。

[0237] 在对本发明进行概括地描述后, 将通过参考以下以说明方式提供的实施例使本发明更容易理解, 但这并不是用来限制本发明的。

[0238] 实施例

[0239] 本领域内已知并且未进行具体描述的标准分子生物学技术通常遵循如 Sambrook 和 Russel, Molecular Cloning: A Laboratory Manual, Cold Springs Harbor Laboratory, New York (2001) 中所述的内容。

[0240] 微阵列合成

[0241] 我们选择代表在新加坡病毒疾病的最常见病因的 35 种病毒基因组 (见上表 1)。

[0242] 全基因组序列从 NCBI 分类学数据库 (NCBI Taxonomy Database) (<http://www.ncbi.nlm.nih.gov/Taxonomy/taxonomyhome.html/>) 下载以产生在全基因组范围内嵌搭的并以平均 8 碱基解析度重叠的 40mer 探针序列。使用 Nimblegen 专利技术 (Nuwaysir, E. F.,

等,2002) 在微阵列上直接合成每种病毒探针的 7 个复制品。探针在微阵列上随机分布,旨在使杂交人为产物的影响最小化。为控制样品对探针的非特异性杂交并测量背景信号,设计并在微阵列上合成 10,000 个寡核苷酸探针。它们是具有 40-60% CG 含量的与人基因组或与病原体基因组无序列相似性的随机探针。作为阳性对照,在阵列上合成针对人基因的 400 个寡核苷酸探针,其中所述的人基因在免疫应答中具有已知或推测的功能。包含植物病毒 PMMV 作为对总计 390,482 个探针的阴性对照。

[0243] 样品制备、微阵列杂交和染色

[0244] 登革病毒细胞系(ATCC#VR-1254)根据 ATCC 推荐进行培养,而 Sin850SARS 细胞系如 Vega 等(Vega 等 2004)所述进行培养。临床标本(鼻咽冲洗液)从印度尼西亚小儿群体中得到并储存在 -80°C 的 RNAzol 中(Leedo Medical Laboratories, Inc., Friendswood, TX)。全部疑似肺炎患者的年龄在 7 至 38 月龄间,表现呼吸道疾病的特异性临床征状。RNA 用 RNAzol 根据制造商说明书进行提取(Smalling 等 2002; Tang 等 1999)。提取的 RNA 重悬于 RNA 贮藏液(Ambion, USA)中并在 -80°C 储存直至需要。根据 Bohlander 等和 Wang 等(Wang 等 2002; Bohlander 等 1992)描述的方法,使用标记的随机引物将 RNA 逆转录成 cDNA。随后, cDNA 如前所述(Wong 等 2002)通过随机 PCR 进行扩增、片段化、用生物素标记末端、杂交至微阵列上以及染色。在初始实验中,我们发现探针的 GC 含量可能在信号强度检测中产生人为现象,即信号的增加直接与探针的 GC 含量成正比。添加 0.82M 氯化四甲基铵(TMAC)至 Nimblegen 的专利 TMAC 杂交缓冲液内则消除这种人为现象。

[0245] 用于 RSV 和 hMPV 的实时诊断性 RT-PCR

[0246] $20\mu\text{l}$ 反应混合物含有 $2\mu\text{l}$ 纯化的患者 RNA, 5U MuLV 逆转录酶, 8U 重组 RNA 酶抑制剂, $10\mu\text{l}$ 无 UNG 的 2X 通用 PCR 主混合物(均来自 Applied Biosystems), $0.9\mu\text{M}$ 引物和 $0.2\mu\text{M}$ 探针。实时 RT-PCR 在 ABI Prism 7900HT Sequence Detection System (Applied Biosystems) 中进行。RT 在 48°C 进行 30 分钟,随后为激活 DNA 聚合酶,在 95°C 维持 10 分钟。RT 产物的扩增通过 95°C 保持 15 秒并在 60°C 保持 1 分钟的 40 个循环实现。每个 PCR 测定法内包括阴性对照和质粒克隆的系列稀释物(阳性对照)。扩增期间,在每个热循环上监测荧光发射。阈值(CT)代表首次检测到明显荧光的循环。使用浓度已知的对照质粒,将 CT 值转换成拷贝数。对于 RSV, 2.61×10^9 个拷贝具有 CT 值为 11.897,而对于 hMPV, 7.51×10^9 个拷贝具有 CT 值为 10.51。

[0247] 用于冠状病毒和鼻病毒的 1-步诊断性 RT-PCR

[0248] 人冠状病毒 OC43、229E 和鼻病毒 16 的冷冻活培养物从 ATCC(货号 VR-1558、VR-740、VR-283)购买,用作阳性对照。使用 RNAMini Kit(Qiagen, Germany)根据制造商说明书从这些培养物中提取 RNA。使用以下诊断性引物对:泛冠状病毒(Cor-FW, Cor-RV), OC43(OC43-FW, OC43-RV), 229E(229E-FW, 229E-RV), 鼻病毒(扩增引物 1(Amplimer 1), 扩增引物 2(Amplimer 2))(Moës 等 2005; Deffernez 等 2004),如前所述扩增样品。

[0249] 病原体微阵列数据的分析

[0250] 我们的病原体微阵列含有一套 40mer 探针 $P = \{p_1, p_2, \dots, p_s\}$, 分组为针对 35 种病毒基因组 $V = \{v_1, v_2, \dots, v_{35}\}$ 的明显不同的探针杂交标签。在杂交病原体核酸时,产生与探针组 P 对应的一组探测信号强度数据 $D = \{d_1, d_2, \dots, d_s\}$ 。

[0251] 单尾 T 检验

[0252] 若病毒 v_a 存在, 则包含病毒的杂交标签的探针 (探针 $\in v_a$) 应当在统计学上具有比探针 $\notin v_a$ 更高的由以下 t - 统计量 (单尾 T 检验) 所确定的信号强度:

$$[0253] \quad t_i = \frac{\mu_a - \mu_{a'}}{\sqrt{\frac{\sigma_a^2}{n_a} + \frac{\sigma_{a'}^2}{n_{a'}}}}$$

[0254] 其中 μ_a 、 σ_a^2 和 n_a 分别是探针 $\in v_a$ 的信号强度的均值、方差和大小并且 $\mu_{a'}$ 、 $\sigma_{a'}^2$ 和 $n_{a'}$ 分别是探针 $\notin v_a$ 的信号强度的均值、方差和大小。

[0255] 将显著性水平设定成 0.05。这意味当 t_a 的 p - 值 < 0.05 , 我们才会接受探针 $\in v_a$ 的信号强度的均值高于探针 $\notin v_a$ 的信号强度的均值的假定。在此情况下, v_a 有可能存在于样品内。然而, 检测的 T - 检验法产生了许多假阳性信号 (calls)。

[0256] PDA v. 1

[0257] PDA v. 1 包含一系列统计检验, 始于加权相对熵检验和 Z - 得分转换 (WKL 得分), 随后是用于正态性的安德森 - 塔林检验。

[0258] 假定病毒为 v_a 。设 P_a 是病毒 v_a 的探针组并且 $\overline{P}_a = P - P_a$ 。设 $[r_{\text{低}}, r_{\text{高}}]$ 是信号强度范围。我们将信号强度对 $j = 0, 1, \dots, c-1$ 分成 c 个组 $[r_{\text{低}} + j(\frac{r_{\text{高}} - r_{\text{低}}}{c}), r_{\text{低}} + (j+1)(\frac{r_{\text{高}} - r_{\text{低}}}{c})]$ 。未改良的加权相对熵可以由下式计算

$$[0259] \quad KL(P_a | \overline{P}_a) = \sum_{j=0}^{c-1} f_a(j) \log\left(\frac{f_a(j)}{f_{\overline{P}_a}(j)}\right)$$

[0260] 其中 n_a^j 和 $n_{\overline{P}_a}^j$ 分别是包含于组 b_j 内的在 P_a 内探针的数目和在 $\overline{P}_a$ 内探针的数目。

$f_a(j) = \frac{n_a^j}{\sum_{h=0}^{c-1} n_a^h}$ 是在组 b_j 中所找到 P_a 内探针的分数; 和 $f_{\overline{P}_a}(j) = \frac{n_{\overline{P}_a}^j}{\sum_{h=0}^{c-1} n_{\overline{P}_a}^h}$ 是在组 b_j 中所找到 $\overline{P}_a$ 内探针的分数。

[0261] 为比较概率分布的尾部的信号差异, 我们设定 $r_{\text{低}} = \overline{\mu}_a$, 即 $\overline{P}_a$ 内探针的平均信号强度, 并且 $r_{\text{高}} =$ 最大信号强度。我们设定组的默认数目 $c = 20$ 。

[0262] 为进一步稳定和 / 或增加相对熵在概率分布的尾部上的灵敏性, 做了两处改良。首先, 我们将安德森 - 塔林型加权函数导入相对熵。这赋予分布的尾部比分布的中间部分更多的权重。其次, 我们对这两种相应的累积分布函数使用统计量代替它们的概率密度函数。我们将改良的相对熵称作加权相对熵 (WKL 得分):

[0263]

$$WKL(P_a | \overline{P}_a) = \sum_{j=0}^{c-1} \frac{Q_a(j) \log\left(\frac{Q_a(j)}{Q_{\overline{P}_a}(j)}\right)}{\sqrt{Q_a(j)[1 - Q_a(j)]}}$$

[0264] 其中 $Q_a(j)$ 是在组 b_j 中找到 P_a 内的探针的信号强度的累积分布函数; $Q_{\overline{P}_a}(j)$ 是在

组 b_j 中所找到 $\overline{P}_a$ 内的探针的信号强度的累积分布函数。

[0265] 因此,对于每种杂交的样品,我们计算每种病毒 $v_a \in V$ 的 WKL 得分。其次,我们断言若样品内不存在病毒,则全部病毒 $v_a \in V$ 的 WKL 得分的分布大致是正态性的。我们通过引导方法 (bootstrapping process) 实验性验证了我们的断言是否正确:设 n 是 V 内病毒的数目。对于每种病毒 $v_k \in V$, 其中 $k = 1, \dots, n$, 我们从真实数据组 D 中选择 $|v_k|$ 探针信号强度 $|v_k|$ 以随机替换方式形成 v_k 的“扰动性”信号强度分布。这种分布可以模拟病毒 v_k 不存在于样品 D 中的情况。此后,产生对 n 个病毒组的 n 个 WKL 得分。其次,我们通过对正态性在 95% 置信度区间的安德森-塔林检验核对 n 个 WKL 得分是否服从正态分布。重复引导方法 100,000 次。在大于 99% 的次数中发现分布是正态性的 (注意:由于在我们的微阵列上代表 35 种病毒基因组,故 $n = 35$)。

[0266] 根据以上讨论,我们通过建立以下无假设和择一假设可以测试样品是否含有病毒:

[0267] H_0 :WKL 得分的分布是正态性的,即病毒不存在于样品内。

[0268] H_1 :WKL 得分的分布不是正态性的,即至少 1 个病毒存在于样品内。

[0269] 定义 将安德森-塔林检验定义为:

[0270] H_0 :数据服从特定的分布。

[0271] H_a :数据不服从特定的分布。

[0272] 检验统计量:将安德森-塔林检验统计量定义为

[0273] $A^2 = -N-S$

[0274] 其中

[0275]
$$S = \sum_{i=1}^N \frac{(2i-1)}{N} [\ln F(Y_i) + \ln (1 - F(Y_{N+1-i}))]$$

[0276] F 是特定分布的累积分布函数。注意 Y_i 是有序数据。

[0277] 显著水平: α

[0278] 临界区:安德森-塔林检验的临界值取决于被检验的具体分布。用表格表示的值和公式已经发表 (Stephens, 1974, 1976, 1977, 1979) 用于几种特定分布 (正态分布、对数正态分布、指数分布、Weibull、对数分布、I 型极值分布)。检验是单侧检验并且若检验统计量 A 大于临界值,则分布具有特定形式的假设被放弃。

[0279] 我们开始应用安德森-塔林检验验证 WKL 得分的分布的正态性以放弃具有 95% 置信度区间的 H_0 。若 WKL 得分的分布不是正态性的,则我们排除具有异常 (outlying) WKL 得分的病毒,并再次应用安德森-塔林检验。重复该过程 (以鉴定共感染性病原体的存在) 直到接受 H_0 为止。

[0280] 我们指定接受 H_0 时的 WKL 得分的分布是背景 WKL 分布。因此,被排除的病毒很有可能存在于样品中,因为它们 WKL 得分不服从背景 WKL 分布。

[0281] 在我们的实验中,观察到含有病毒的样品内的 P , 即非正态分布因随机偶然性对给定 WKL 得分发生的概率,是极低的,即 $P < 1.0 \times 10^{-6}$ (通过 WKL 得分的 Z -得分转换而得到)。表 1 显示用于我们的病毒检测算法的伪码。

[0282] 表 1:病毒检测算法

[0283]

给予病毒组 V 和探针组 P 病原体微阵列数据 D

设 $V_{存在} = F$

设 D_{WKL} 是对全部 $v \in V$ 的 $WKL(P_v \parallel P_v)$ 组;

1. 以安德森-塔林检验来验证正态性以确定 D_{WKL} 的正态性。若 D_{WKL} 具有显著水平 0.05 的正态分布, 返回 $V_{存在}$ 。否则, 进入步骤 2。
2. 从 D_{WKL} 中找到具有最高 $WKL(P_a \parallel P_a)$ 的病毒 v_a 。
 设 $V_{存在} = V_{存在} \cup \{v_a\}$; $D_{WKL} = D_{WKL} - \{WKL(P_a \parallel P_a)\}$; 进入步骤 1。
3. 剔除检测到的 SPS 并验证 WKL 分布是正态性的。
4. 若该分布不是正态性的, 返回步骤 2 以找到共感染性病原体。

[0284] 预测全基因组的扩增偏倚

[0285] 为鉴定临床标本内的未知病原体, 优选采用随机引物扩增, 而不是引物特异性扩增。然而, 在使用随机引发扩增来鉴定已知病原体的初始实验中, 我们经常观察到通过序列多态性无法解释的遍及基因组区域的不完全杂交 (图 7C)。基因组二级结构、探针二级结构和探针 GC 含量也无法解释这些低信号强度的探针。因此, 我们假设不完全杂交可能归因于 PCR 偏倚, 其中所述的 PCR 偏倚来源于随机引物在逆转录 (RT) 步骤中与病毒基因组结合的差异性能力。我们实验中所用的随机引物是 26mer, 其包含以固定的 17mer 序列 (5' -GTTTCCAGTCACGATA) (SEQ ID NO:1) 为标签的随机九聚物 (3') (仍见图 1), 其中固定的 5' 标签的目的是为了促进 RT 产物的 PCR, 产生小于 10000bp 的 PCR 片段, 尤其是 500-1000bp PCR 片段 (Pang 等 2005; Wang 等 2002; Wang 等 2003)。为研究此现象, 我们设计一种算法 (AES) 使用实验数据以模拟 RT-PCR 方法。成功的 RT-PCR 取决于引物与模板的结合。已知引物内二级结构的形成, 如两个模板间形成的二聚体和发夹, 引物内二级结构的形成, 如标签与九聚体间形成的二聚体和发夹, 及探针的解链温度影响结合效率 (Nguyen 和 Southern, 2000; Ratushna 等 2005)。

[0286] 假设随机引物混合物内的九聚体与病毒基因组序列完美互补, 则该算法确定可以从基因组内每个可能的起始位置中产生 500-1000bp 产物的概率。因此, 对于 1000 碱基的滑动窗口内的每个核苷酸而言, 该核苷酸成功地被扩增的概率在它的扩增效率得分 (AES; 见上文“扩增效率得分”) 中被反映出来。为证实算法的有效性, 我们将针对 RSV 基因组的全部 1,948 个 SPS 探针的杂交信号强度分级, 并将 SPS 探针的杂交信号强度与 SPS 探针的 AES 值比较。在整个 RSV 基因组范围内, 我们观察到 AES 与杂交信号强度充分相关 (Fisher's 精确概率法检验 $P = 2.2 \times 10^{-16}$), 这表明 AES 与探针检测间的强相关性 (图 12)。另一个使用针对偏肺炎病毒的 1,705 个 SPS 探针的比较显示了类似结果, $P = 1.3 \times 10^{-9}$ 。图 10 表示了 AES 在预测临床样品内的 SPS 探针检测中的重要性。值得注意的是, 我们观察到更高的 AES 值与更高比例的可检测探针相关, 尤其是 AES 值的最高的 20%。因此, 尽管 HD、MCM、% GC 及序列独特性是探针性能重要的参数, 然而它们没有考虑 PCR 偏倚, 因此在缺少 AES 下

考虑时,它们不能充分预测探针性能。在选择病原体 SPS 中使用最高的第 20 百分位的 AES 作为第一筛选程序明显地改进了病原体的预测,这由更高 WKL 得分和消除假阳性信号证明(表 3)。

[0287] 表 3:仅使用平均探针信号强度检测病原体(T-检验)产生大量的假阳性信号。优化的杂交标签和与人基因组交叉杂交的探针的剔除(筛选)减少了假阳性信号,但是这对检测精确性仍是不充分的。使用完整的未筛选的探针组,PDA v.1 能够作出准确诊断。若 WKL 得分 > 5,则“检测”到病毒。使用优化的杂交标签(筛选)增加了 WKL 得分,这与诊断置信度提高相对应。病毒 CT 值:检测到病毒时的实时 PCR 循环(见上文)。

[0288]

芯片 #	病原体	使用 PDA v.1 的检测			病毒 CT 值	病毒拷贝数
		Max WKL 得分(无筛选程序)	Max WKL 得分(经过筛选的)	检测到的病毒数目		
32272	纯 SARS	5.007	5.803	1	--	--
34959	纯登革病毒	14.351	20.373	1	--	--
35259	RSV 患者 324	18.288	20.611	1	21.4366	9.8×10^7
35179	hMPV 患者 122	1.747	8.439	1	25.5388	50384
35253	RSV 患者 841	12.056	12.069	1	20.8619	14×10^7
36042	RSV 患者 412	16.466	17.531	1	23.5804	2.5×10^7
36053	RSV 患者 483	12.089	12.168	1	24.8340	1.2×10^7
35915	非肺炎患者(阴性对照)	3.916	4.284	0	0	0

[0289] 将阵列上所杂交的全部患者样品的数据表示在下文表 4 中。

[0290] 表 4. 杂交到病原体微阵列上的临床患者的完整列表

[0291]

阵列	患者 ID	WKL	P-值	PDA v.1 诊断	临床诊断 *	初始 PCR 诊 断	PCR CT 值	病毒拷 贝数	RT-PCR 引物
35179	122	8.439216	1.34×10^{-71}	hMPV	LRTI	hMPV	24.8	5.0×10^4	A1
35887	122	18.312077	2.98×10^{-22}	hMPV	LRTI	hMPV	24.8	5.0×10^4	A2
71180	133	17.359597	2.42×10^{-37}	hMPV	LRTI	hMPV	25.1159	4.0×10^4	A2
66691	165	8.56786	1.84×10^{-4}	hMPV	肺炎	hMPV	27.9	3.9×10^3	A2
70935	254	21.348515	8.70×10^{-30}	hMPV	LRTI	hMPV	21.9518	5.4×10^5	A2
63781	283	16.680752	3.97×10^{-12}	hMPV	肺炎	未知			A2
73067	769	24.006323	1.34×10^{-51}	hMPV	LRTI	hMPV	25.6715	2.5×10^4	A2

[0292]

66690	853			未检测到	肺炎	hMPV	36	0.5	A2
68359	892	12.534284	5.66×10^{-5}	鼻病毒属	肺炎	hMPV	33.8	27	A2
35915	111			未检测到	阴性对照 1	None			A1
70927	818			未检测到	阴性对照 1	None			A2
66701	312			未检测到	肺炎	RSV A	33.7	44	A2
71006	321			未检测到	肺炎	RSV A	31.1	340	A2
66702	368			未检测到	肺炎	未知			A2
71025	414	25.406289	3.80×10^{-24}	RSV B	肺炎	RSV A	22.3	3.9×10^5	A2
71027	478			未检测到	肺炎	RSV A	34.8	18	A2
73068	832	59.275233	1.91×10^{-102}	RSV 属	LRTI	RSV A	23.7681	1.2×10^5	A2
71028	913	25.897084	3.23×10^{-30}	RSV B	肺炎	RSV A	19.1	4.7×10^6	A2
66703	924	12.673149	9.71×10^{-6}	RSV 属	肺炎	RSV A	31.5	250	A2
35259	324	20.61147	3.55×10^{-94}	RSV B	LRTI	RSV B	21.4366	3.0×10^6	A1
35662	355	17.999418	2.97×10^{-40}	RSV B	LRTI	RSV B	20.2642	6.7×10^6	A1
66695	374			未检测到	肺炎	RSV B	34.1	500	A2
70933	378	13.81578	7.77×10^{-17}	RSV B	LRTI	RSV B	23.9204	5.4×10^5	A2
36042	412	17.531234	4.58×10^{-55}	RSV B	LRTI	RSV B	23.5804	6.9×10^5	A1
35890	412	17.214556	1.05×10^{-43}	RSV B	LRTI	RSV B	23.5804	6.9×10^5	A2+A3
36053	483	12.168025	1.47×10^{-12}	RSV B	LRTI	RSV B	24.834	2.9×10^5	A1

[0293]

70997	554	76.547183 54.013223	1.83×10^{-119} 2.45×10^{-61}	鼻病毒 属; 肠道 病毒科	肺炎	RSV B	35.1	240	A2
35253	841	12.069138	4.86×10^{-26}	RSV B	肺炎	RSV B	20.8619	4.4×10^6	A1
73070	841	22.10857 5.708560	6.80×10^{-50} 5.66×10^{-6}	RSV B, hMPV 共同感 染	肺炎	RSV B/ hMPV	20.8619 35.4	4.4×10^6 8	A2
68360	841	21.369516 9.647188	2.09×10^{-25} 1.23×10^{-8}	RSV B, hMPV 共同感 染	肺炎	RSV B/ hMPV	20.8619 35.4	4.4×10^6 8	A2
66696	185			未检测 到	肺炎	未知			A2
66697	261			未检测 到	肺炎	未知			A2
66698	331			未检测 到	肺炎	未知			A2
71189	393			未检测 到	肺炎	未知			A2
66699	461			未检测 到	肺炎	未知			A2
66700	573	41.397051 27.444893	3.97×10^{-23} 1.34×10^{-11}	鼻病毒 属; 肠 道病毒 科	肺炎	未知			A2
71182	639			未检测 到	肺炎	未知			A2
71007	699			未检测 到	肺炎	未知			A2
71188	859			未检测 到	肺炎	未知			A2

[0294] *LRTI : 下呼吸道感染

[0295] AES 的重要性表明扩增效率和随后的探针检测可以通过使用优化的 RT-PCR 引物标签被改善。因此, 我们使用随机生成的 17mer 标签序列计算 AES 得分, 并选择前三位差异最大的导致最全面地增加 AES 得分的引物 (图 13)。使用 AES 优化的引物, 我们从临床样品中扩增了偏肺炎病毒和 RSV, 具有改良的 PCR 效率和检测灵敏性 (图 14, 表 5)。

[0296] 表 5 :用患者样品 #412 和 #122 比较 E-Predict 与 PDA v.1 算法。阵列 35179 使用在结果中所述的原始 PCR 引物进行扩增。阵列 36731 和 35887 使用引物 A2 进行扩增,阵列 35890 使用两种引物 A2 和 A3 进行扩增。PDAv.1 在所有情况下都只反馈了正确的病原体。E-Predict 的作者使用 $P < 0.01$ 作为其平台的显著性界限值 (Urisman 等 2005)。若使用该算法来分析我们的阵列数据,则似乎需要更低界限值。使用两种算法 (阵列 35179 与 35887) 中的任何一种算法,由 PCR 建模所设计的新引物均产生更好的预测得分。在 PCR 过程期间使用了第二种引物在 WKL 得分和 P- 值方面表现出增加的改良 (阵列 36731 与 35890)。

[0297]

阵列	患者	PCR 扩增引物	E-预测 (E-predict) 算法			GIS 病原体算法	
			基因组	相似性得分	P-值	基因组	WKL
36042	412(RSV)	原始引物 A1	RSV	0.35128	0	RSV	21.526316
			OC43 冠状病毒	0.350264	6.84E-20		
			229E 冠状病毒	0.323503	1.77E-10		
			乙型肝炎病毒	0.134825	3.03E-04		
			SARS 冠状病毒	0.338911	0.00299		
			甲型肝炎病毒	0.229589	0.00847		
36731	412(RSV)	A2	RSV	0.335389	0	RSV	21.836754
			OC43 冠	0.348043	2.29E-13		

[0298]

			状病毒				
			229E 冠状病毒	0.322055	2.00E-09		
			乙型肝炎病毒	0.135222	1.02E-06		
			风疹病毒	0.164332	0.00919		

35890	412 (RSV)	A2+A3	RSV	0. 334602	0	RSV	22. 093258
			OC43 冠 状病毒	0. 348969	3. 63E-23		
			229E 冠 状病毒	0. 322805	3. 20E-14		
			乙型肝 炎病毒	0. 13436	6. 74E-04		
			SARS 冠 状病毒	0. 338609	0. 03060		
35179	122 (hMPV)	原始引 物 A1	hMPV	0. 26011069 5	5. 01E-28	hMP V	9. 763149
			风疹病 毒	0. 16478498 1	1. 20E-17		
			口蹄疫 病毒 C	0. 20674781 6	4. 66E-11		
			日本脑 炎病毒	0. 20134722 2	1. 65E-04		
			乙型肝 炎病毒	0. 13340762 2	1. 98E-04		
			黄热病 毒	0. 20050056 4	0. 00567		
			致肠道 细胞病 变人孤 儿病毒 1	0. 22200202 5	0. 01740		
			新城疫 病毒	0. 23448168 6	0. 01820		
35887	122 (hMPV)	A2	hMPV	0. 299655	0	hMP V	39. 677149

			风疹病毒	0.169626	3.40E-19		
			乙型肝炎病毒	0.137703	5.84E-12		
			OC43 冠状病毒	0.347685	5.06E-10		
			229E 冠状病毒	0.321702	1.72E-06		
			SARS 冠状病毒	0.340504	1.76E-06		
			口蹄疫病毒 C	0.2075	1.31E-04		
			新城疫病毒	0.23453	0.04310		

[0299] PDA v. 1- 用于检测病原体的算法

[0300] 对基因组扩增而言,临床标本往往是次优的:它们可能具有低病毒滴度,具有与阵列上参考株不同的序列多态性,或具有共感染性病原体。微阵列还具有来自非特异性杂交和其它人为产物的固有噪音。因此,解读微阵列数据不仅仅是使探测信号强度曲线与 SPS 匹配,或使用简单的统计方法(例如,T- 检验,ANOVA 等)。为解决此问题,我们建立了加强的统计软件 PDA v. 1,相对于计算机模拟预测的 SPS,该软件分析探针信号强度的分布来鉴定在杂交的样品内存在的病原体(见上文)。

[0301] 基于我们的观察,即尽管对阵列上全部探针的信号强度处于正态分布,然而包含存在于样品内的病原体 SPS 的大部分探针具有非常强烈的产生向右偏斜的分布信号强度;我们推断通过分析探针信号强度的分布可能检测到病原体的存在(图 9A)。检验每个 SPS 的信号强度分布的尾部还能够使我们鉴定样品内共感染性病原体的存在。

[0302] 因此,PDA v. 1 包含 2 个部分:(1) 用于评估每种病原体 SPS 内探针的探针信号强度的增强的加权相对熵(WKL;我们的增强相对熵检验),和(2) 用于确定每种 SPS 的 WKL 得分的分布是否是正态的安德森-塔林检验。

[0303] 原来的相对熵不能可靠地确定概率分布尾部内的差异,并且高度依赖探针/基因组的数目及每个信号强度组的大小(Kullback 和 Leibler,1951)。我们通过以下方式克服这些缺陷,即通过引入安德森-塔林统计量以给予每个分布的尾部更大权重,和通过使用累积分布函数替代原有的概率分布(Anderson 和 Darling,1952)。我们称增强的 KL 发散度为加权相对熵(WKL):

[0304]

$$WKL(P_a | \bar{P}_a) = \sum_{j=0}^{k-1} \frac{Q_a(j) \log(\frac{Q_a(j)}{\bar{Q}_a(j)})}{\sqrt{Q_a(j)[1-Q_a(j)]}}$$

[0305] 其中 $Q_a(j)$ 是在组 b_j 中找到 P_a 内的探针的信号强度的累积分布函数； $\bar{Q}_a(j)$ 是在组 b_j 中所找到 $\bar{P}_a$ 内的探针的信号强度的累积分布函数。代表无病原体的 SPS 应当具有正态的信号强度分布并且因此具有相对低的 WKL 得分，而代表有病原体的那些 SPS 应当具有高的统计学显著的异常 WKL 得分（图 9B）。在 PDA v. 1 的第二部分中，使 WKL 得分的分布接受安德森-塔林检验处理以评价正态性。若 $P < 0.05$ ，认为 WKL 分布不是正态的，表明具有异常 WKL 得分的病原体存在。鉴定病原体时，在缺少病原体的 WKL 得分下进行独立的安德森-塔林检验以检测共感染性病原体的存在。以这种方式，反复进行该过程直至只剩下正态分布（即 $P > 0.05$ ；见上表 3 和表 4）。PDA v. 1 是极其迅速的，能够在约 10 秒内从杂交的微阵列中作出诊断。

[0306] 对 33 份临床患者样品的病原体诊断

[0307] 根据图 11 中所示的工作流程，我们通过将 33 份临床标本杂交至病原体微阵列平台上评估我们的平台。这些临床标本中的 27 份标本先前已经被诊断为感染 RSVA、RSV B 或偏肺炎病毒。我们的平台从 21/27 样品中精确地检测到病原体。没有检测到病毒（假阴性）的 6 份样品处在通过实时 PCR（ < 10 个病毒拷贝 / 反应）所确定的检测界限值上，并且如此低的病毒负荷不可能成为导致患者严重疾病的病因学病原。这 6 份样品中的 2 份由微阵列正确地诊断为感染鼻病毒。在另一个因未知病原体所致的严重呼吸道疾病的 6 位患者的筛查中，微阵列鉴定出在一份样品中存在病因学病原（鼻病毒）（上表 4）。这些结果通过实时 PCR 进行验证。如所预期，当我们杂交从无病毒病因学的肺炎患者中提取的样品时，未检测到任何病原体。

[0308] 数据分析

[0309] 使用 Axon 4000b 扫描仪和 Genepix 4 软件 (Axon Instruments) 以 $5 \mu m$ 解析度扫描微阵列。使用 NimbleScan 2.1 软件 (NimbleGen Systems) 提取信号强度，使用自动化脚本，我们从每个探针的 7 个复制品中计算出中等信号强度（以消除杂交人为产物）和标准差。探针信号强度根据基因组进行分组并按序列顺序排列，随后再格式化成 CDT 格式，以便在 Java Treeview (<http://jtreeview.sourceforge.net>) 中以图形方式观察使信号强度。平行地，使用 PDA v. 1 分析探针的中等信号强度来确定哪种病原体存在和相关的预测置信度水平。本发明的发明人通过实验来证实探针设计对实验结果的影响并随后显示本发明分析算法鲁棒性。

[0310] 探针设计对实验结果的影响

[0311] 将含有来自 35 种病毒的 53555 个 40mer 探针的 PDC 用于 4 个独立的微阵列实验。这 53555 个探针是根据每种病毒的 5-bp 嵌搭 (tiling) 进行选择的并且不经过我们的任何探针设计标准处理。因此，我们预期因 CG 含量、交叉杂交和低效率扩增所产生的错误显著多于设计良好的探针的 PDC 的错误。我们将 4 个实验处于这种不利的环境下测试我们的分析算法。

[0312] 在该实施例中，具有未知病原体的人样品使用随机探针通过 RT-PCR 方法进行扩增并随后杂交至 PDC 上。我们使 PDC 上 35 种病毒中每种病毒的探针接受显著水平 0.05

的单尾 T- 检验, 并计算探针的信号强度相对于芯片上全部探针的信号强度的加权相对熵, 以对每一个实验确定哪种病毒存在于样品内。证实我们程序的分析精确性是通过增压舱 (wet-lab) PCR 以鉴定样品内的实际病毒实现。在表 6 中我们提供了对表 6 内 4 个实验的分析结果以及它们相应的 PCR 验证结果。

[0313] 表 6 : 对未应用探针设计标准的 PDC 进行分析产生的结果。由我们的分析算法确定是样品中实际病毒的病毒用浅灰颜色标出, 其中所述的样品对每个实验进行测试的。

[0314]

	实验 1		实验 2		实验 3		实验 4	
样品名称	35259_324		35179_122		35253_841		35915_111	
ID	53555		53555		53555		53555	
病毒(编号)	t-检验 p-值	WKL	t-检验 p-值	WKL	t-检验 p-值	WKL	t-检验 p-值	WKL
NC_001781.1	0	16.391	1	未分析 (NA)	0	10.8563 5	1	NA
NC_003461.1	1	NA	1	NA	1	NA	1	NA
NC_003443.1	0.999324	NA	0.873017	NA	0.99802	NA	0.99996 1	NA
NC_001796.2	1	NA	1	NA	1	NA	1	NA
AY283794.1	0	0.5435	0.108141	NA	0	0.77595 9	0	0.435427
NC_005147.1	0	1.2896	1	NA	0	1.39959 1	0	1.762912
NC_002645.1	0	1.2943	0.999847	NA	0	1.65588 8	0	2.079334
NC_004148.2	1	NA	0.002733	5.76290 7	1	NA	1	NA
NC_002023.1	1	NA	0.579561	NA	1	NA	1	NA
NC_002022.1								
NC_002021.1								
NC_002020.1								
NC_002019.1								
NC_002018.1								
NC_002017.1								
NC_002204.1	1	NA	1	NA	1	NA	1	NA
NC_002205.1								

[0315]

NC_002206.1								
NC_002207.1								
NC_002208.1								
NC_002209.1								
NC_002210.1								
NC_002211.1								
NC_001563.2	1	NA	0.000001	0.53782 6	1	NA	0.99501 3	NA
NC_002031.1	1	NA	0.000005	0.75875 8	0.99887 3	NA	0.36394 7	NA
NC_002728.1	1	NA	0.999062	NA	1	NA	1	NA
NC_002617.1	0.999994	NA	0	0.57184 4	1	NA	0.76909 8	NA
NC_001802.1	1	NA	0.999966	NA	1	NA	1	NA
NC_003977.1	0	2.7424	0	2.18982 7	0	3.97874 7	0	1.490665
NC_001576.1	0.371224	NA	0.004643	0.94841	0.00959 9	1.25704 1	0	3.961532
NC_002554.1	0.000062	0.7146	0	1.52729 2	0.29933 4	NA	0.00000 2	0.166239
NC_001545.1	0	1.4545	0	2.43855 8	0	0.86978 2	0	0.989592
NC_001489.1	0	1.7088	0.319125	NA	0	2.59306 5	0	1.510399
NC_005222.1	0.999757	NA	0.646314	NA	0.77391 2	NA	0.80787 5	NA
NC_005217.1	0.60477	NA	0.999903	NA	0.35435 8	NA	0.00087 1	0.626818
NC_001294.1	0	1.8411	0.000523	0.90239 9	0	2.43215	0.00000 7	0.537531
NC_004291.1	0.662386	NA	0.954137	NA	0.25542 2	NA	0.09914 8	NA
NC_001437.1	1	NA	0	0.59309 3	1	NA	1	NA
AB189128.1	1	NA	0.906213	NA	1	NA	1	NA
AF326573.1	1	NA	0.038503	0.53978 3	1	NA	1	NA
AF489932.1	1	NA	0.899797	NA	1	NA	1	NA
M87512.1	1	NA	0.759668	NA	1	NA	1	NA
NC_001430.1	1	NA	0.912496	NA	1	NA	0.99991 2	NA
NC_001428.1	0.999988	NA	0.284792	NA	0.99934 6	NA	0.95716 4	NA
NC_001612.1	0.970379	NA	0.000001	0.55786 5	0.99887 8	NA	0.06122 6	NA
NC_003986.1	1	NA	0.000012	0.60447 4	1	NA	0.99794 5	NA
NC_001472.1	0.999999	NA	0.0046	0.45519 4	0.99957 9	NA	0.14340 4	NA
NC_001617.1	0.721465	NA	0.98373	NA	0.17873 3	NA	0.41420 9	NA
NC_001490.1	0.999808	NA	0.995029	NA	0.99736 9	NA	0.85902 5	NA
推断病毒	NC_001781.1(RSV)		NC_004148.2(HM PV)		NC_001781.1(RS V)		无	

[0316]

证实病毒 (PCR)	NC_001781.1(RSV)	NC_004148.2(HM PV)	NC_001781.1(RS V)	无
---------------	----------------------	-----------------------	----------------------	---

[0317] 这些结果表明该分析算法精确推断出前 3 个实验中的所测试样品内的实际病毒（结果示表示在上表 6 内）。此外，我们能够推断在最后一个实验中的样品无病毒。值得注意的是若我们仅使用显著性水平 0.05 的 t- 检验，则在下表 7 中表示出对每份样品中检测存在的病毒的数目。

[0318] 表 7：仅使用 t- 检验推断出的病毒的假阳性检测

[0319]

样品名称	35259_324	35179_122	35253_841	35915_111
使用 T- 检验所检 测的病毒	9	14	9	10
假阳性	8	13	8	10
最大 KL 发散度 (> 5.0)	16.391	5.76	10.85	—
使用 T- 检验随后 用 KL 发散度所 检测的病毒	1	1	1	0

[0320] 对通过 t- 检验的病毒使用加权相对熵，我们能够剔除全部假阳性病毒并鉴定实际病毒。因此，我们的分析算法可以在高水平噪音下稳定地确定病毒。

[0321] 然后，我们研究使用具有探针设计标准的 PDC 对我们的分析结果带来的影响。首先，对 35 种病毒的每种病毒计算扩增效率图。随后，使原始 PDC 上的实际 53555 个探针经受探针设计标准处理。从芯片中剔除具有极端水平 CG 含量、与人或非靶病毒相似性高和扩增效率得分低的探针。为第二组实验留下总计 10955 个探针。使用在第一组实验内所用的样品，我们用新芯片重复下表 8 内的 4 个实验。实验结果列在表 8 中。

[0322] 表 8：对应用探针设计标准的 PDC 分析产生的结果。由我们的分析算法确定是样品内实际病毒的病毒用浅灰颜色标出，其中所述的样品对每个实验进行测试。

[0323]

	实验 1		实验 2		实验 3		实验 4	
样品名称	35259_324		35179_122		35253_841		35915_111	
D1	10955		10955		10955		10955	
病毒(编号)	t-检验 p-值	WKL	t-检验 p-值	WKL	t-检验 p-值	WKL	t-检验 p-值	WKL
NC_001781.1	0	18.54859	1	NA	0	11.17914	1	NA
NC_003461.1	1	NA	1	NA	1	NA	1	NA
NC_003443.1	0.548718	NA	0.53727	NA	0.002783	0.837121	0.020436	0.603552
NC_001796.2	1	NA	0.999907	NA	1	NA	1	NA
AY283794.1	0	1.347801	0.024116	0.858364	0	1.523272	0	1.128637
NC_005147.1	0	1.604381	0.999697	NA	0	2.150019	0	2.893555
NC_002645.1	0	2.802742	0.999895	NA	0	4.612482	0	3.635771
NC_004148.2	1	NA	0.000003	9.324785	1	NA	1	NA
NC_002023.1	1	NA	0.124517	NA	1	NA	0.999163	NA
NC_002022.1								
NC_002021.1								
NC_002020.1								
NC_002019.1								
NC_002018.1	1	NA	0.998724	NA	1	NA	1	NA
NC_002017.1								
NC_002204.1								
NC_002205.1								
NC_002206.1								
NC_002207.1	1	NA	0.998724	NA	1	NA	1	NA
NC_002208.1								
NC_002209.1								
NC_002210.1								
NC_002211.1								
NC_001563.2	0.986443	NA	0.428418	NA	0.76002	NA	0.112011	NA
NC_002031.1	0.998103	NA	0.003435	2.52162	0.278672	NA	0.409527	NA
NC_002728.1	0.999375	NA	0.30951	NA	0.969492	NA	0.297244	NA
NC_002617.1	0.63418	NA	0.003578	0.965856	0.247148	NA	0.025188	0.861163
NC_001802.1	1	NA	0.998118	NA	1	NA	1	NA
NC_003977.1	0	3.062956	0.000028	3.027442	0	4.574591	0	3.277708
NC_001576.1	0.579342	NA	0.101093	NA	0.155219	NA	0.026417	3.280335
NC_002554.1	0.6722	NA	0	2.289379	0.80654	NA	0.106683	NA
NC_001545.1	0	2.225817	0	2.794877	0.000019	1.674329	0	1.97064
NC_001489.1	0.099427	NA	0.999985	NA	0.000366	1.829543	0.000006	3.023235
NC_005222.1	0.999735	NA	0.294141	NA	0.974031	NA	0.356952	NA
NC_005217.1	0.916186	NA	0.994358	NA	0.600759	NA	0.032616	2.105628

[0324]

NC_004294.1	0.867625	NA	0.235197	NA	0.100961	NA	0.05275 9	NA
NC_004291.1	0.992032	NA	0.964128	NA	0.714211	NA	0.20642 2	NA
NC_001437.1	1	NA	0.001058	1.563913	1	NA	0.85722 8	NA
AB189128.1	1	NA	0.732737	NA	0.999997	NA	0.98859	NA
AF326573.1	1	NA	0.435629	NA	0.999986	NA	0.90539 3	NA
AF489932.1	1	NA	0.322655	NA	0.999996	NA	0.99683 7	NA
M87512.1	0.999617	NA	0.057346	NA	0.999758	NA	0.93793 7	NA
NC_001430.1	1	NA	0.865038	NA	1	NA	0.88233 9	NA
NC_001428.1	1	NA	0.522986	NA	0.999351	NA	0.74941 2	NA
NC_001612.1	0.991708	NA	0.751091	NA	0.990929	NA	0.25763 5	NA
NC_003986.1	0.999997	NA	0.02014	0.93616	0.937996	NA	0.70898 5	NA
NC_001472.1	0.99959	NA	0.977242	NA	0.957869	NA	0.69293 6	NA
NC_001617.1	0.435562	NA	0.474076	NA	0.028549	1.699567	0.07967 6	NA
NC_001490.1	1	NA	0.90881	NA	0.996231	NA	0.51866 2	NA
推断病毒	NC_001781.1(RSV)	NC_004148.2(HMPV)	NC_001781.1(RSV)	无				
证实病毒(PCR)	NC_001781.1(RSV)	NC_004148.2(HMPV)	NC_001781.1(RSV)	无				

[0325] 在以下的实验组中,分析算法正确地检测到 3 份样品内的实际病毒并且还检测出阴性样品。在为我们的芯片设计优异的探针后,实验 1、2 和 3 内实际病毒的加权相对熵大于无探针设计的相应实验的实际病毒的加权相对熵。这意味来自实际病毒的信号强度比 PDC 内的背景噪音相对更高。这证实我们的探针设计标准已经从 PDC 中剔除一些不良探针,这产生更精确的分析。

[0326] 若我们仅使用显著性水平 0.05 的 T- 检验,我们再一次将 4 个实验的结果列在表 9 中。此时,对每份样品中检测存在的病毒的数目列在表 9 中:

[0327] 表 9:在具有探针设计的 PDC 中仅使用 t- 检验推断的病毒的假阳性检测。

[0328]

样品名称	35259_324	35179_122	35253_841	35915_111
使用 T- 检验所检测的病毒	6	9	9	10
假阳性	5	8	8	10
最大 KL 发散度 (> 5.0)	18.54859	9.324785	11.17914	-

使用 T- 检验随后用 KL 发散度所检测的 病毒	1	1	1	0
---------------------------------	---	---	---	---

[0329] 从表 9 中,可以看到探针设计已经减少由 t- 检验对样品 35259_324 和 35179_122 所检测到的假阳性病毒的数目。更重要的观察是实际病毒的加权相对熵对全部 4 个样品均增加。这意味当探针设计标准应用于 PDC 时,实际病毒的信号比背景信号具有更大差别。

[0330] 总之,我们证实使用显著水平 0.05 的单尾 t- 检验,随后对每种病毒的信号强度计算加权相对熵,能够精确分析 PDC 上的数据并高概率地确定样品内的实际病原体。尽管该分析算法甚至在高水平噪音下运行良好,然而我们证实该通过使用上述探针设计标准为 PDC 选择优异探针组可以改善分析的精确性。

[0331] 用于探针设计和病原体检测的备选方法

[0332] 非常少的算法可用于预测微阵列上的交叉杂交,仅 1 种算法,即 E- 预测法 (E-predict) 已经被报道并被验证在微阵列上检测病原体的有效性 (Urisman 等 2005 ;Li 等 2005)。E- 预测使杂交标签与预测的标签匹配,其中预测的标签从对每种微阵列探针的杂交的理论自由能中产生。然而,使用 E- 预测法分析我们的微阵列产生许多假阳性信号 (见上表 5)。例如, E- 预测法在 RSV 患者 412 内检测到冠状病毒 (图 15)。使用泛冠状病毒引物以及 OC43 和 229E 冠状病毒特异性诊断引物的诊断性 PCR 证实患者 412 中无冠状病毒 (见上表 4)。我们假设使用 E- 预测法导致的假阳性信号由与人基因组或 RSV 基因组交叉杂交的冠状病毒探针中产生。实际上,预测 85% 的具有最高信号强度的 50 个冠状病毒探针与人基因组交叉杂交和 65% 的所述探针具有相对于 RSV 的小于 17 的 HD,其中所述 HD 仅高于我们用于家族的交叉杂交的 HD 阈值 12。此外,优化 E- 预测法以在含有病毒基因组区域间高度保守的探针的微阵列上工作,而不是在其中与人基因组交叉杂交将是重要考虑因素的嵌搭 (tiling) 阵列上工作。因此,有可能这两种因素——不同的微阵列设计策略和与人基因组交叉杂交——造成 E- 预测法在我们的平台上表现不佳。从我们使用 E- 预测法的经验来看,将 PDA v1 与其它算法比较对我们而言是不合理的,因为后者针对不同的探针长度进行设计并针对其它的应用和平台进行优化的。

[0333] 结论

[0334] 通过实验地确定交叉杂交阈值,我们产生了计算机模拟的病原体标签探针组,该探针组仅包含与临床样品内存在的特定病毒良好杂交的探针。AES 算法使我们可以设计高效扩增完整病毒基因组的通用引物标签。连同 PDAv. 1 检测算法,我们可以确定无疑地鉴定来自于临床样品中的由微阵列上所代表的任何病原体。这种方法不需要实验性验证每种病原体杂交标签的有效性并使含有针对 > 10000 种病原体的探针的其它微阵列成为病原体鉴定的有力诊断性平台。

[0335] 我们已经优化了对病原体检测微阵列的设计和分析,便于它们在医院环境中使用。我们发现随机 PCR 内例行所用的引物标签是偏倚性的,这导致病原体基因组的非均一性扩增。这种偏倚可以通过使用我们的 AES 算法设计引物被避免。我们的计算机模拟的标签探针组使我们精确预测哪些探针将与阵列上所代表的任何病原体杂交。连同 PDA v. 1 检测算法,这种方法不需要实验性验证每种病原体杂交标签的有效性并使含有针对 > 10000

种病原体的探针的其它微阵列成为病原体鉴定的有力诊断性平台。

[0336] 本文中,我们报道了使用定制设计的微阵列平台以系统地研究病毒扩增效率、杂交信号输出、靶-探针复性特异性与病原体检测的可重复性间复杂关系的结果。我们的发现形成了用于计算机模拟预测最佳病原体标签探针组 (SPS) 的新方法学的基础,(在微阵列杂交前)揭示决定病毒扩增效率的因素并且证实病毒扩增效率得分 (AES) 与最佳探针选择间的重要联系。最后,我们描述了新的基于统计的病原体检测算法 (PDA),它可以快速并可重复地鉴定临床标本中一系列病毒滴度范围内的病原体。

[0337] 我们已经证实当存在至少 4000 个病毒拷贝时,使用从公众可用的数据库中得到的病毒基因组序列以高度的确定性检测临床样品内病毒的可行性(见上表 3)。该算法的灵敏性接近抗原检测法的灵敏性,从而使本算法成为临床上有关的检测工具 (Liu 等 2005; Marra 等 2003)。通过计算机模拟精确预测病原体杂交标签的能力的确具有胜过现有微阵列方法的明显优势,其中现有微阵列方法需要通过首先将阵列与纯的病原体样品杂交而实验性验证有效性。除了特异性鉴定阵列上所代表的病原体以外, PDA v. 1 还使对阵列上没有得到具体代表的那些基因组鉴定病原体的纲、科或属(通过放宽 HD 和 MCM 的阈值)。这种信息对于临床中的治疗决策往往是足够的。使用 AES 优化的标签,我们能够鉴定在使用非 AES 优化的标签进行扩增时从临床样品中先前不能检测到的病毒。因此,选择 AES 优化的标签提高 PCR 效率和灵敏性。本发明的算法可以适用于其它基于标签的 PCR 应用,如生成 DNA 文库和富集用于再测序的 RNA。

[0338] 参考文献

[0339] Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, et al. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* 25:3389-3402.

[0340] Anderson TW, Darling DA (1952) Asymptotic theory of certain goodness of fit criteria based on stochastic processes. *Annals of Mathematical Statistics* 23: 192-212.

[0341] Bodrossy L, Sessitsch A (2004) Oligonucleotide microarrays in microbial diagnostics. *Curr Opin Microbiol* 7:245-254.

[0342] Bohlander SK, Espinosa I, Rafael, Le Beau MM, Rowley JD, Diaz MO (1992) A method for the rapid sequence-independent amplification of microdissected chromosomal material. *Genomics* 13:1322-1324.

[0343] Bustin, S. A. & Nolan, T. (2004) Pitfalls of quantitative real-time reverse-transcription polymerase chain reaction. *J Biomol Tech* 15, 155-166.

[0344] Deffernez C, Wunderli W, Thomas Y, Yerly S, Perrin L, et al. (2004) Amplicon Sequencing and Improved Detection of Human Rhinovirus in Respiratory Samples 10.1128/JCM.42.7.3212-3218. 2004. *J Clin Microbiol* 42:3212-3218.

[0345] Fu J, Tan BH, Yap EH, Chan YC, Tan YH (1992) Full-length cDNA sequence of dengue type 1 virus (Singapore strain S275/90). *Virology* 188:953-958.

[0346] Goulden, C. H. *Methods of Statistical Analysis*, End. 2nd. (John Wiley & Sons,

Inc., New York ;1956).

[0347] Hamming RW(1950)Error Detecting and Error Correcting Codes.BellSystem Technical Journal 29 :147-160.

[0348] International Human Genome Sequencing Consortium.Initial sequencingand analysis of the human genome.Nature 409(6822),860-921(2001).

[0349] Kane MD, Jatkoe TA, Stumpf CR, Lu J, Thomas JD, et al. (2000)Assessment of the sensitivity and specificity of oligonucleotide(50mer)microarrays.Nucleic Acids Res 28 :4552-4557.

[0350] Kane, M.D. et al.Assessment of the sensitivity and specificity ofoligonucleotide(50mer)microarrays.Nucleic Acids Res 28,4552-4557(2000).

[0351] Ksiazek TG, Erdman D, Goldsmith CS, Zaki SR, Peret T, et al. (2003)A novel coronavirus associated with severe acute respiratory syndrome.N Engl JMed 348 : 1953-1966.

[0352] Kullback S, Leibler RA(1951)On information and sufficiency.Annals ofMathematical Statistic 22 :79-86.

[0353] Li X, He Z, Zhou J(2005)Selection of optimal oligonucleotide probes formicroarrays using multiple criteria, global alignment and parameter estimation.Nucl Acids Res 33 :6114-6123.

[0354] Liu J, Lim SL, Ruan Y, Ling AE, Ng LF, et al. (2005)SARS transmission pattern in Singapore reassessed by viral sequence variation analysis.PLoS Med2(2), 162-168.

[0355] Marra MA, Jones SJ, Astell CR, Holt RA, Brooks-Wilson A, et al. (2003)The Genome sequence of the SARS-associated coronavirus.Science 300 :1399-1404.

[0356] Maskos U, Southern EM(1993)A study of oligonucleotide reassociation using large arrays of oligonucleotides synthesised on a glass support.NucleicAcids Res 21 :4663-4669.

[0357] **Moës** E, Vijgen L, Keyaerts E, Zlateva K, Li S, et al. (2005)A novelpancoronavirus RT-PCR assay:frequent detection of human coronavirus NL63in children hospitalized with respiratory tract infections in Belgium. BMCInfect Dis 5 :6.

[0358] Nguyen HK, Southern EM(2000)Minimising the secondary structure of DNA targets by incorporation of a modified deoxynucleoside :implications for nucleic acid analysis by hybridisation.Nucleic Acids Res 28 :3904-3909.

[0359] Nuwaysir EF, Huang W, Albert TJ, Singh J, Nuwaysir K, et al. (2002) Gene expression analysis using oligonucleotide arrays produced by masklessphotolithography.Genome Res 12 :1749-1755.

[0360] Pang XL, Preiksaitis JK, Lee B(2005)Multiplex real time RT-PCR for thedetection and quantitation of norovirus genogroups I and II in patients withacute gastroenteritis.J Clin Virol 33 :168-171.

- [0361] Ratushna VG, Weller JW, Gibas CJ(2005)Secondary structure in the target as a confounding factor in synthetic oligomer microarray design. *BMC Genomics* 6 : 31.
- [0362] Ruan YJ, Wei CL, Ee AL, Vega VB, Thoreau H, et al. (2003) Comparative full-length genome sequence analysis of 14 SARS coronavirus isolates and common mutations associated with putative origins of infection. *Lancet* 361 : 1779-1785.
- [0363] Sambrook and Russel, (2001) *Molecular Cloning : A Laboratory Manual*, Cold Springs Harbor Laboratory, New York.
- [0364] SantaLucia, J. , Jr. , Allawi, H. T. & Seneviratne, P. A. (1996) Improved nearest-neighbor parameters for predicting DNA duplex stability. *Biochemistry* 35, 3555-3562.
- [0365] Smalling TW, Sefers SE, Li H, Tang YW (2002) Molecular approaches to detecting herpes simplex virus and enteroviruses in the central nervous system. *J Clin Microbiol* 40 : 2317-2322.
- [0366] Stephens, M. A. (1974). EDF Statistics for Goodness of Fit and Some Comparisons, *Journal of the American Statistical Association*, Vol. 69, pp. 730-737.
- [0367] Striebel HM, Birch-Hirschfeld E, Egerer R, Foldes-Papp Z (2003) Virus diagnostics on microarrays. *Curr Pharm Biotechnol* 4 : 401-415.
- [0368] Sung, W. K. & Lee, W. H. Fast and Accurate Probe Selection Algorithm for Large Genomes. *CSB* (2003).
- [0369] Sung, W. K. & Lee, W. H. (2003) in *IEEE Computational Systems Bioinformatics Conference* Stanford University, Stanford, CA.
- [0370] Urisman A, Fischer KF, Chiu CY, Kistler AL, Beck S, et al. (2005) E-Predict : a computational strategy for species identification based on observed DNA microarray hybridization patterns. *Genome Biol* 6 : R78.
- [0371] Vega VB, Ruan Y, Liu J, Lee WH, Wei CL, et al. (2004) Mutational dynamics of the SARS coronavirus in cell culture and human populations isolated in 2003. *BMC Infect Dis* 4 : 32.
- [0372] Vora GJ, Meador CE, Stenger DA, Andreadis JD (2004) Nucleic acid amplification strategies for DNA microarray-based pathogen detection. *Appl Environ Microbiol* 70 : 3047-3054.
- [0373] Wang D, Coscoy L, Zylberberg M, Avila PC, Boushey HA, et al. (2002) Microarray-based detection and genotyping of viral pathogens. *Proc Natl Acad Sci U S A* 99 : 15687-15692.
- [0374] Wang D, Urisman A, Liu YT, Springer M, Ksiazek TG, et al. (2003) Viral discovery and sequence recovery using DNA microarrays. *PLoS Biol* 1 : E2.
- [0375] Wong CW, Albert TJ, Vega VB, Norton JE, Cutler DJ, et al. (2004) Tracking

the Evolution of the SARS Coronavirus Using High-Throughput, High-Density Resequencing Arrays. *Genome Res* 14:398-405.

[0376] Wu, D. Y., Ugozzoli, L., Pal, B. K., Qian, J. & Wallace, R. B. (1991) The effect of temperature and oligonucleotide primer length on the specificity and efficiency of amplification by the polymerase chain reaction. *DNA Cell Biol* 10, 233-238.

SEQUENCE LISTING

〈110〉 新加坡科技研究局

〈120〉寡核苷酸设计和/或核酸检测的方法和/或装置

<130>FP3143

 $\langle 140 \rangle_{11/202, 023}$

<141>2005-08-12

<160>13

<170>PatentIn version 3.3

 $\langle 210 \rangle 1$

<211>26

$\langle 212 \rangle$ DNA

〈213〉 人工序列

 $\langle 220 \rangle$

〈223〉 正向和 / 或反向随机引物 (图 1A)

 $\langle 220 \rangle$

<221>misc_feature

 $\langle 222 \rangle (18) \dots (26)$

$\langle 223 \rangle_n$ 为 a, c, g, 或 t

 $\langle 400 \rangle 1$

gtttcccagt cacgatannn nnnnnn

26

 $\langle 210 \rangle_2$

<211>26

<212>DNA

〈213〉 人工序列

 $\langle 220 \rangle$

〈223〉 随机正向引物 (图 1B)

 $\langle 400 \rangle_2$

gtttcccagt cacgatagca tgaggg

26

<210>3

<211>26

<212>DNA

<213> 人工序列

<220>

<223> 随机反向引物（图 1B）

<400>3

gtttcccagt cacgatacga atagct

26

<210>4

<211>22

<212>DNA

<213> 人工序列

<220>

<223> 病毒序列的片段（图 1B 中的病毒序列的上面最左边的链）

<400>4

acgatatccg cgaatagcta ga

22

<210>5

<211>23

<212>DNA

<213> 人工序列

<220>

<223> 病毒序列的片段（图 1B 中的病毒序列的上面最右边的链）

<400>5

catccctcat gcatggggca att

23

<210>6

<211>22

<212>DNA

<213> 人工序列

<220>

<223> 病毒序列的片段（图 1B 中的病毒序列的下面最左边的链）

<400>6

tgctataggc gcttatcgat ct

22

<210>7

<211>23

<212>DNA

<213> 人工序列

<220>

<223> 病毒序列的片段（图 1B 中的病毒序列的下面最右边的链）

<400>7

gtagggagta cgtaccccg taa

23

<210>8

<211>17

<212>DNA

<213> 人工序列

<220>

<223> 随机引物标签（图 1C 和 1D 的上面的链）

<400>8

gtttcccagt cacgata

17

<210>9

<211>17

<212>DNA

<213> 人工序列

<220>

<223> 随机引物标签（图 1C 和 1D 的下面的链）

<400>9

caaagggtca gtgctat

17

<210>10

<211>17

<212>DNA

<213> 人工序列

<220>

<223> 引物 A1(图 13)

<400>10

gtttcccagt cacgata

17

<210>11

<211>17

<212>DNA

<213> 人工序列

<220>

<223> 引物 A2(图 13)

<400>11

gatgaggga gatgggg

17

<210>12

<211>17

<212>DNA

<213> 人工序列

<220>

<223> 引物 A3(图 13)

<400>12

ctcatgcacg acccaaa

17

<210>13

<211>17

<212>DNA

<213> 人工序列

<220>

<223> 引物 A4(图 13)

<400>13

agatccattc cacccca

17

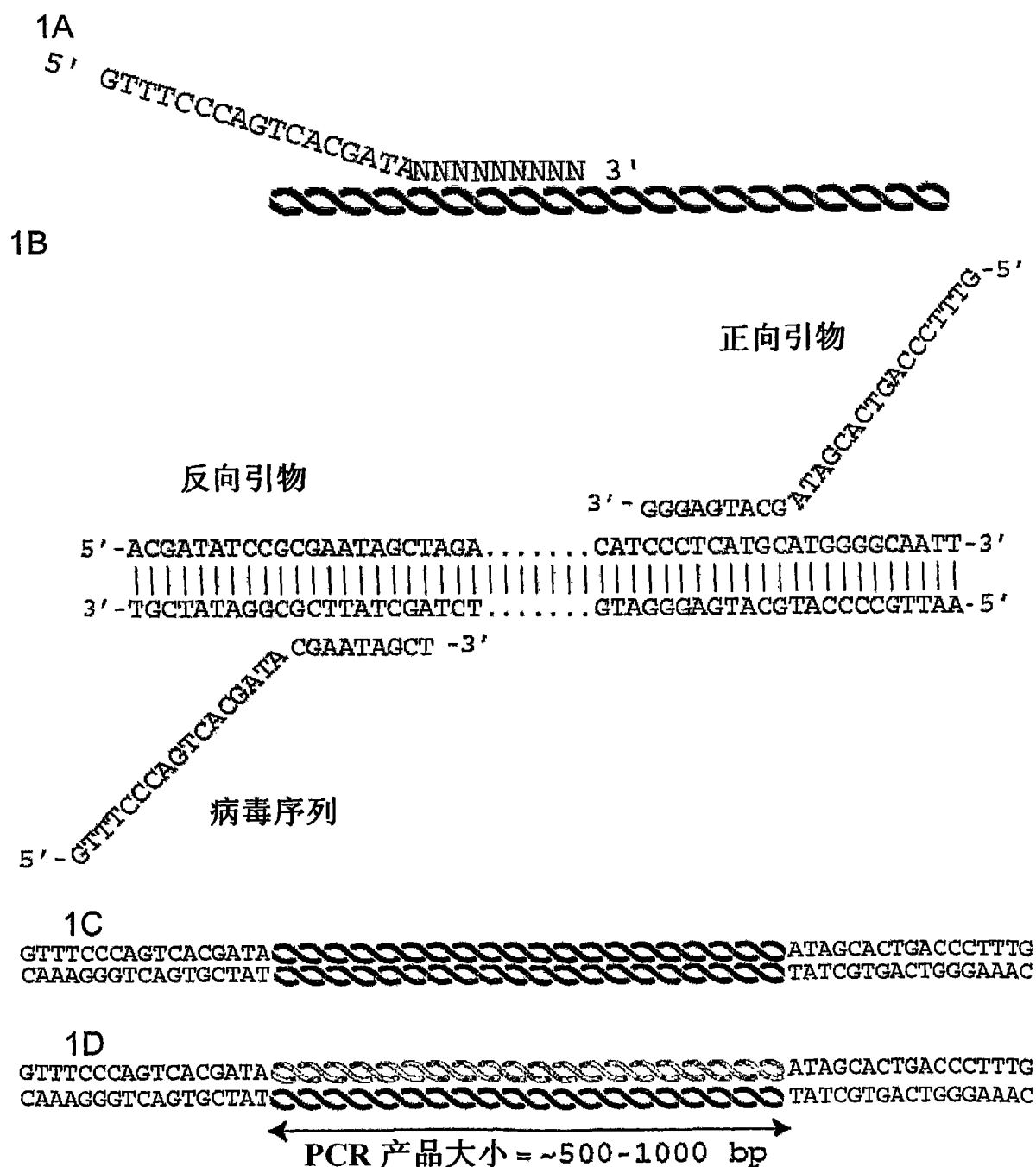


图 1

扩增效率图 (RSV)

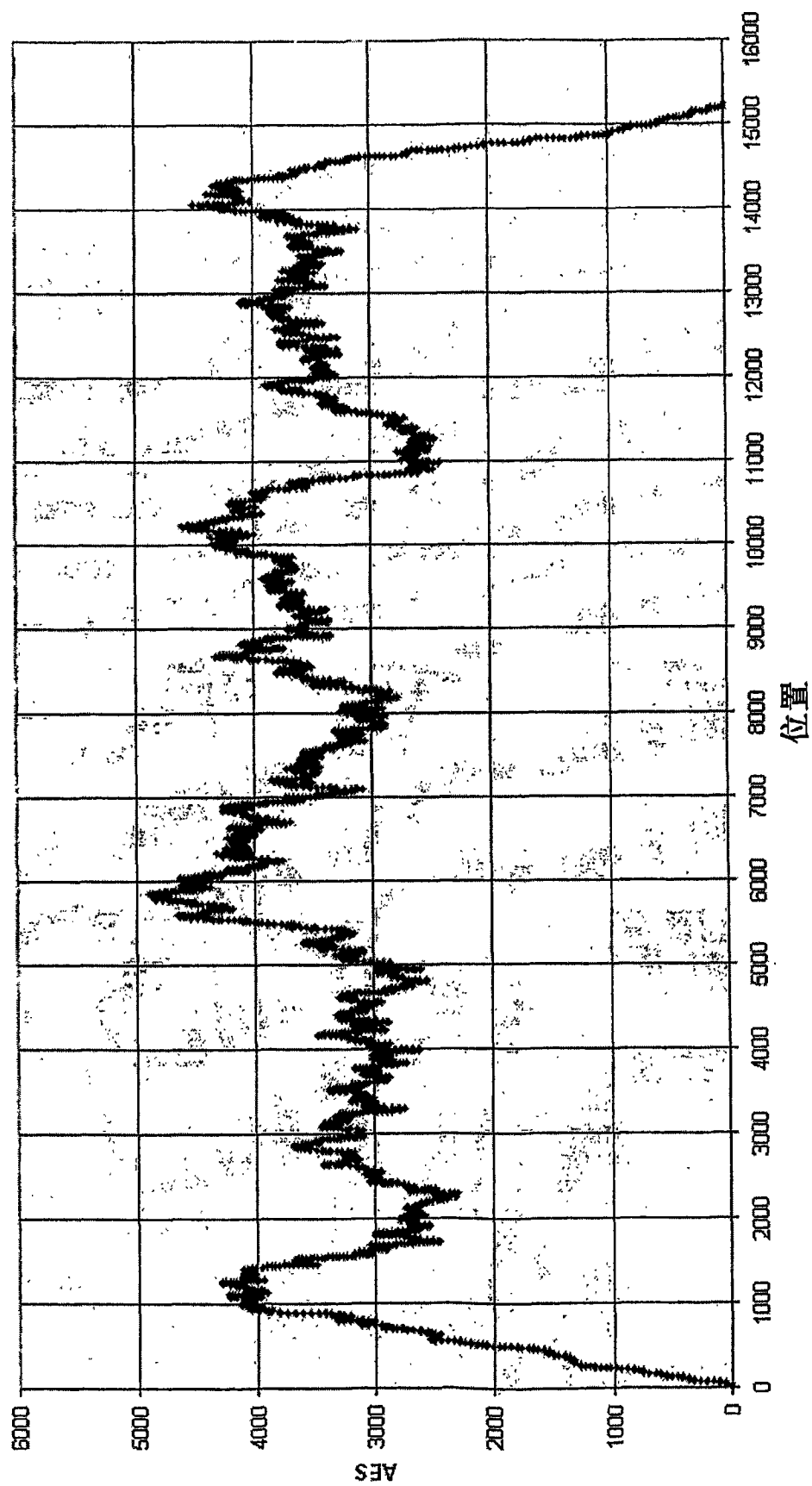


图 2

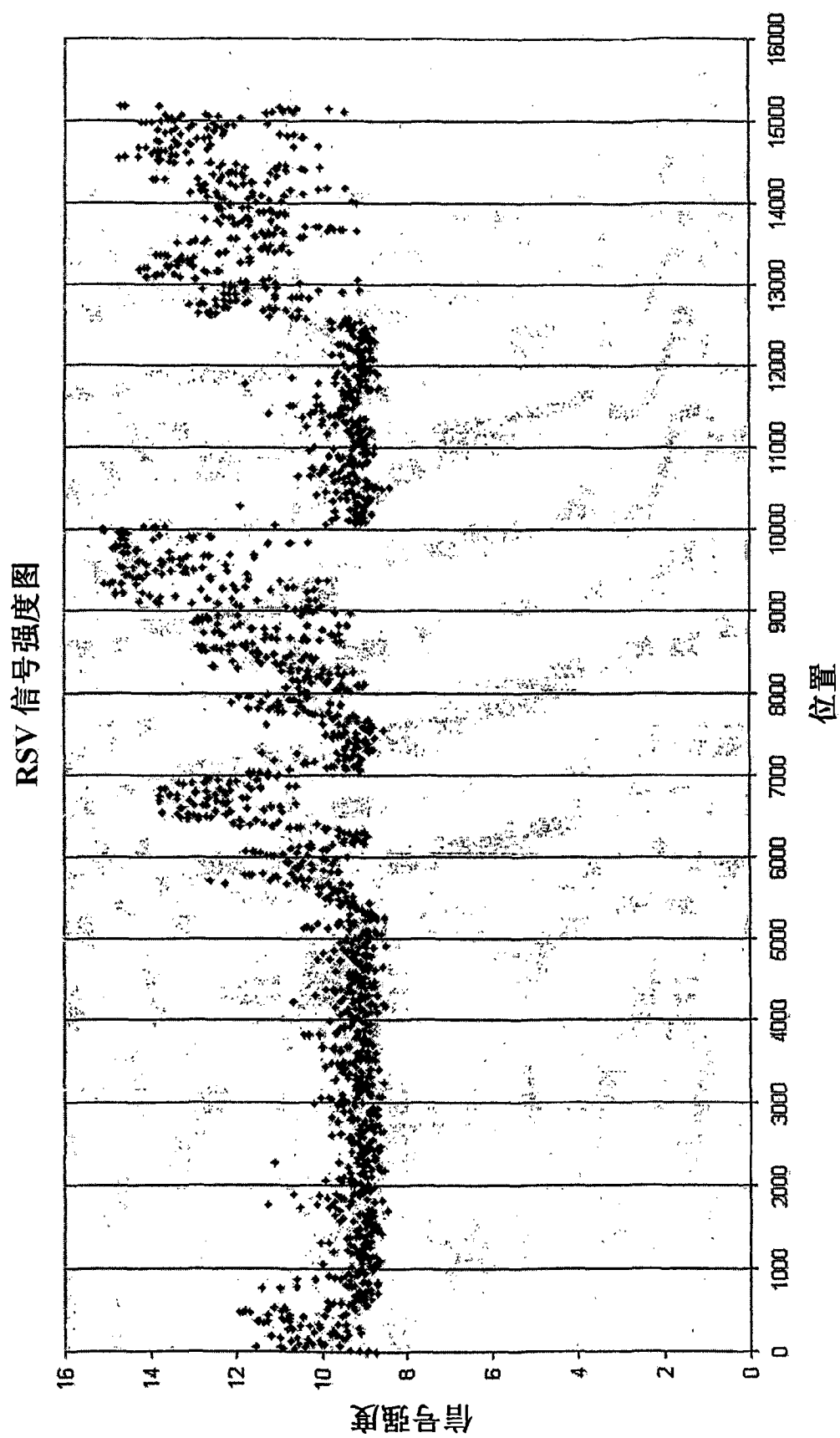


图 3

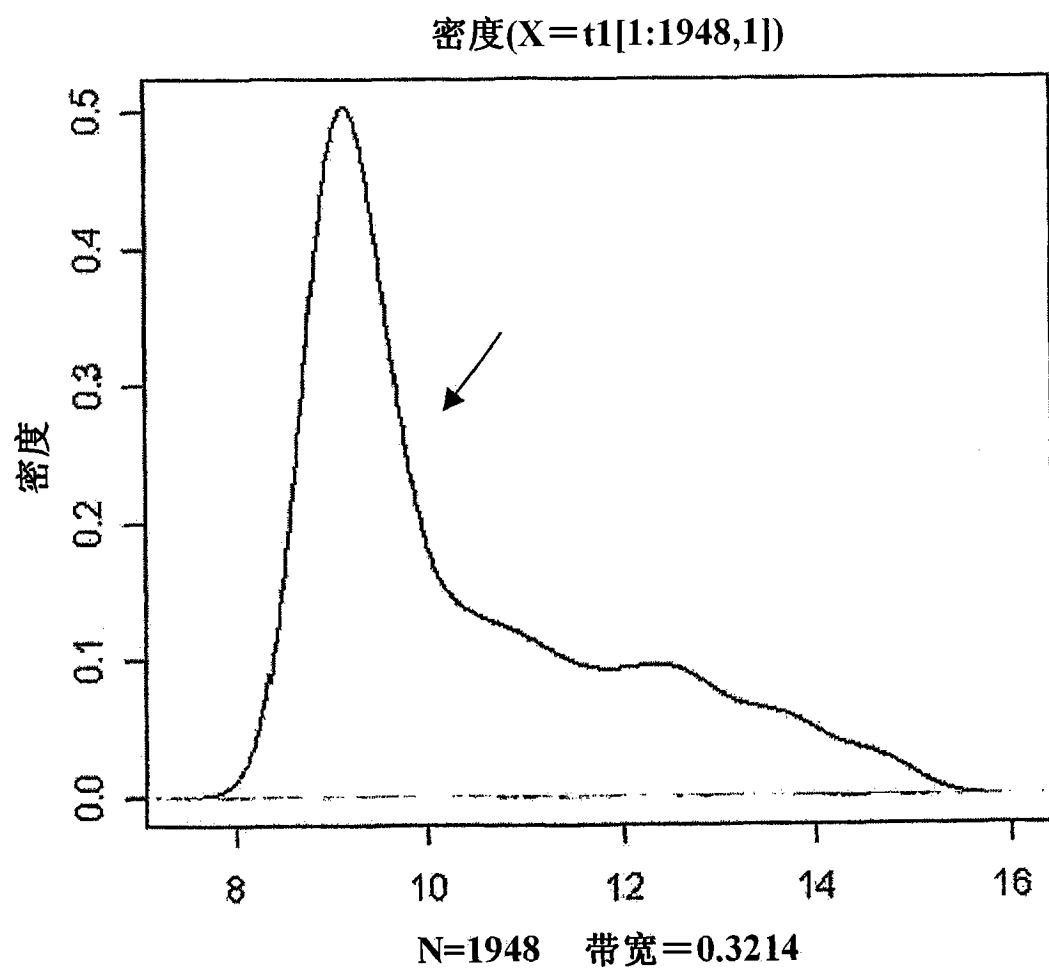


图 4A

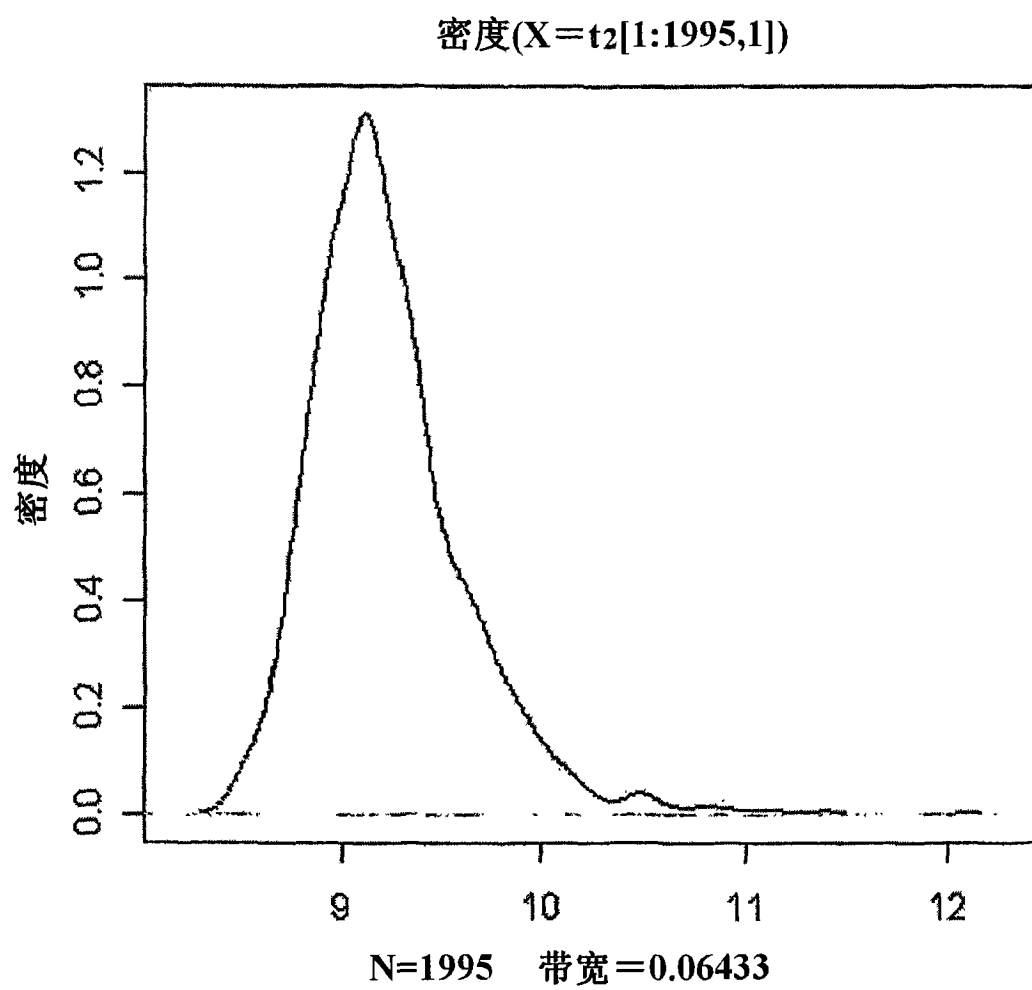


图 4B

假定 PDC 数据 D 具有病毒组 V 和探针组 P

设 $V' = \emptyset$

对于每个 $v_a \in V$,

- 计算 v_a 内探针的显著水平为 0.05 的单尾 t-检验
若(t_a 的 p -值 < 0.05)
- 若 $KL(P_a \parallel P) \geq 0.1$, $V' = V' \cup \{v_a\}$, 接受
- 否则, 放弃;
- 否则, 放弃;

返回 V'

图 5

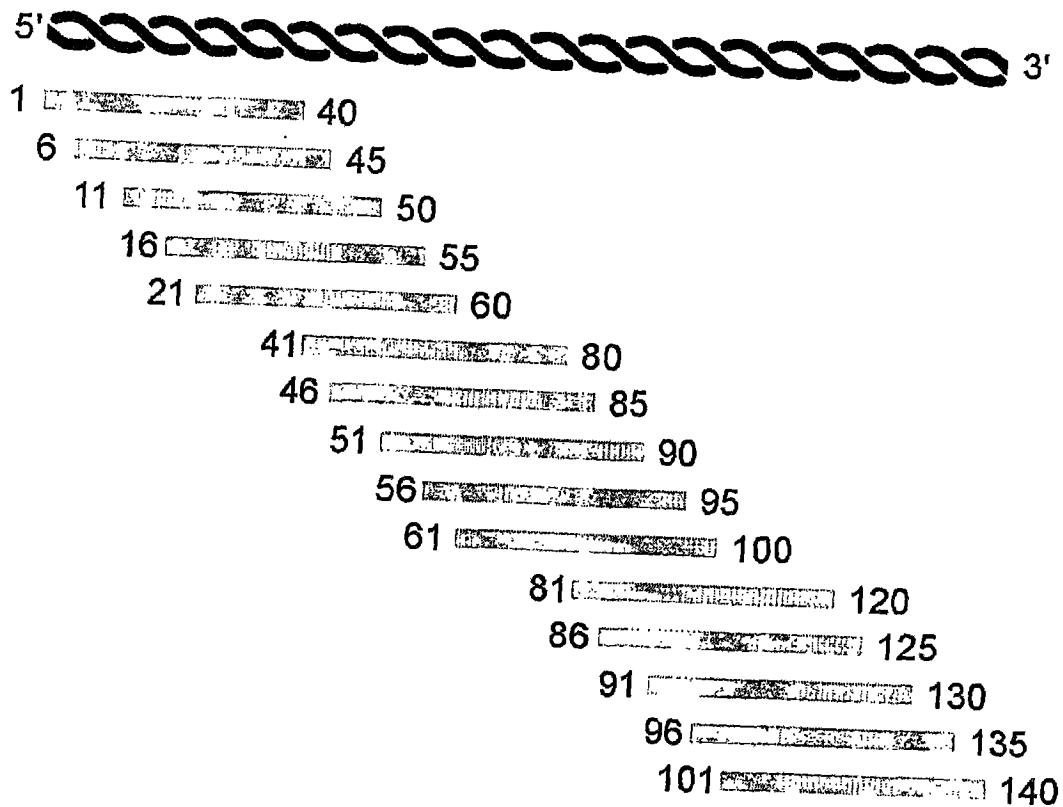
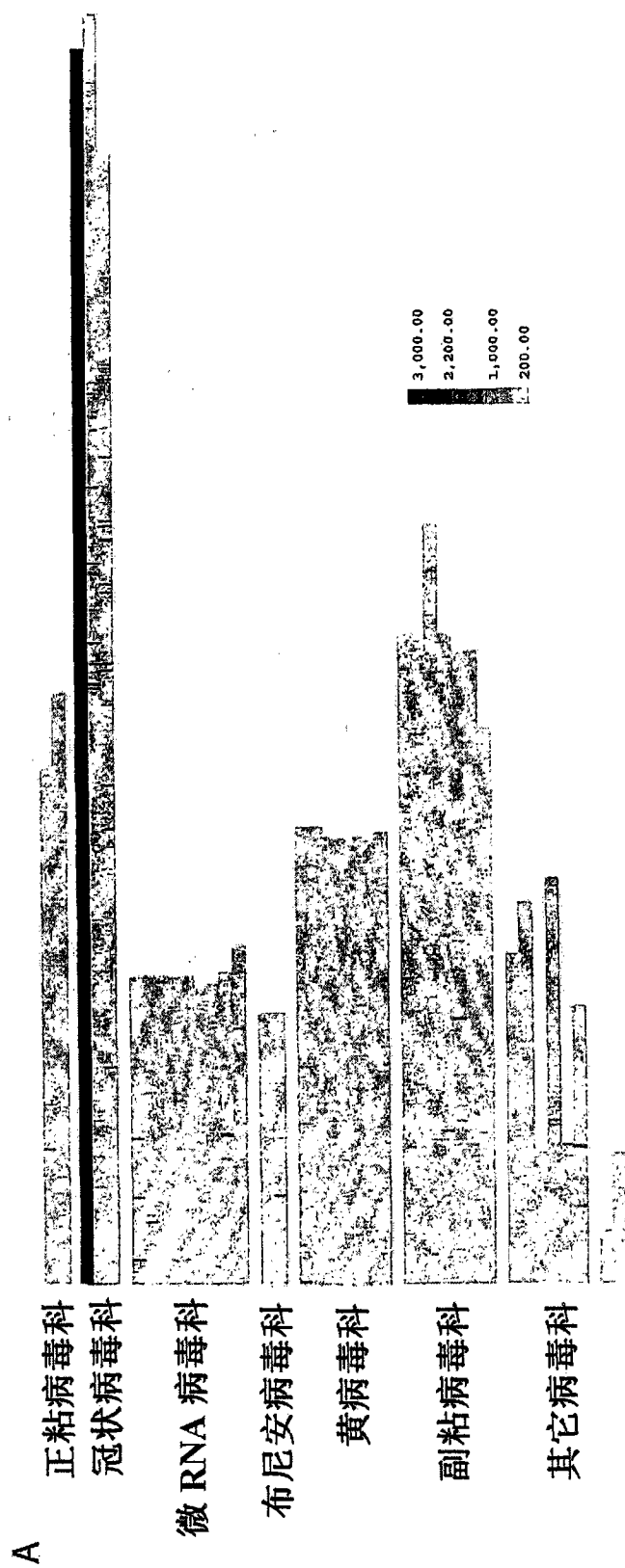


图 6



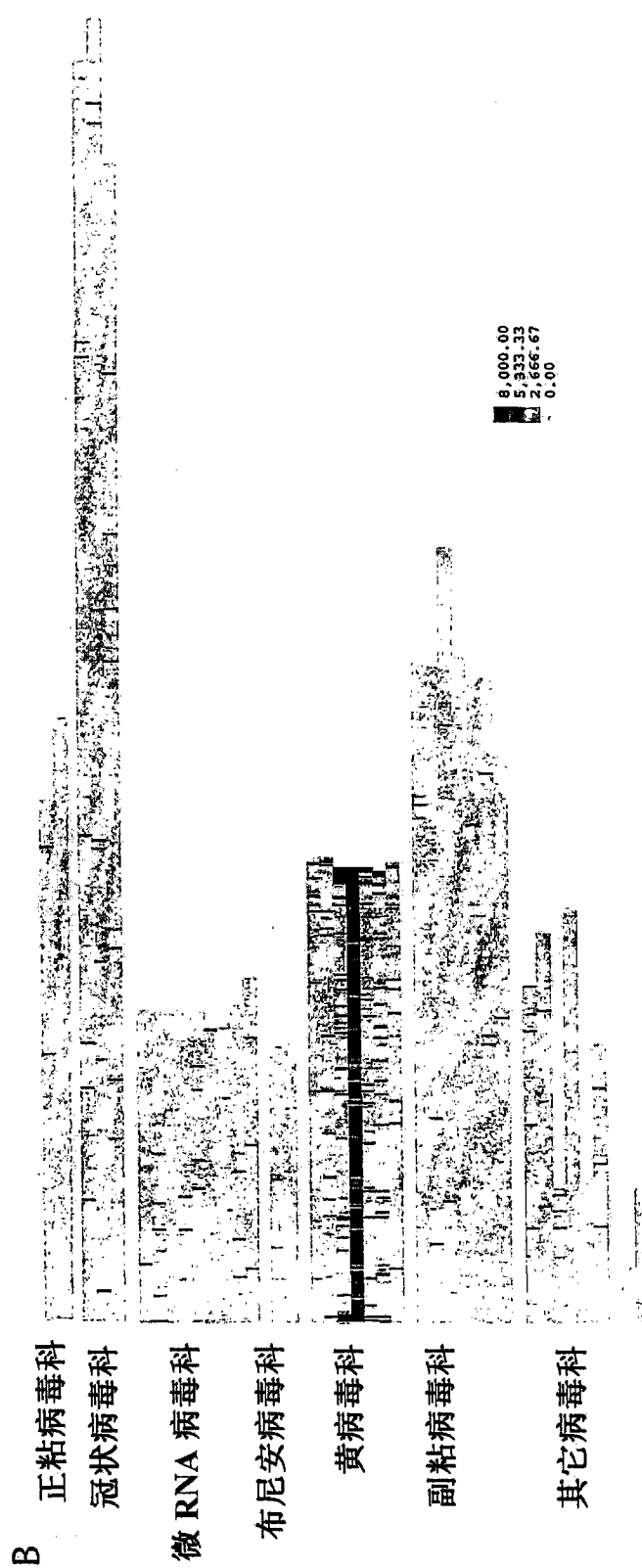


图 7B



图 7C

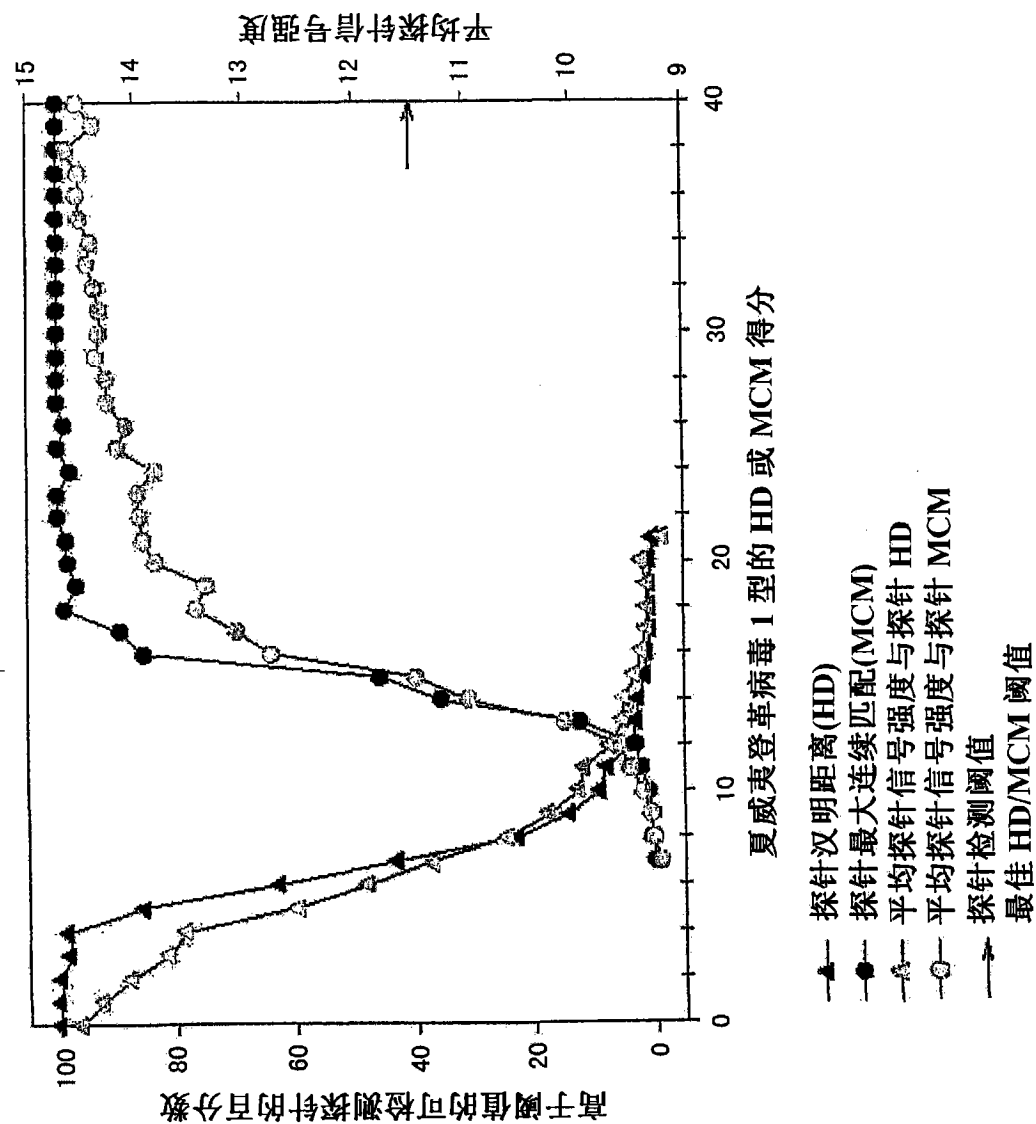


图 8

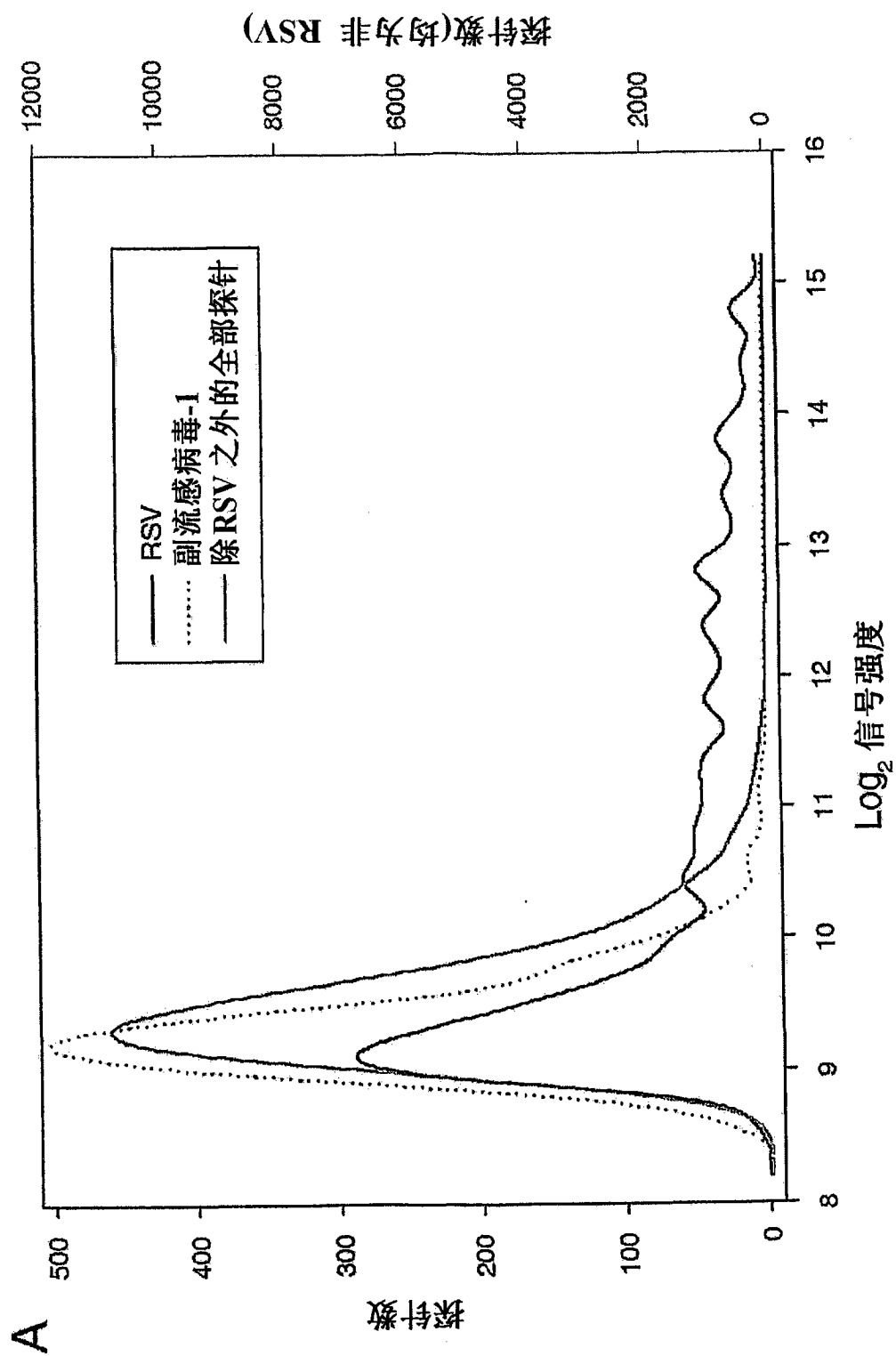


图 9A

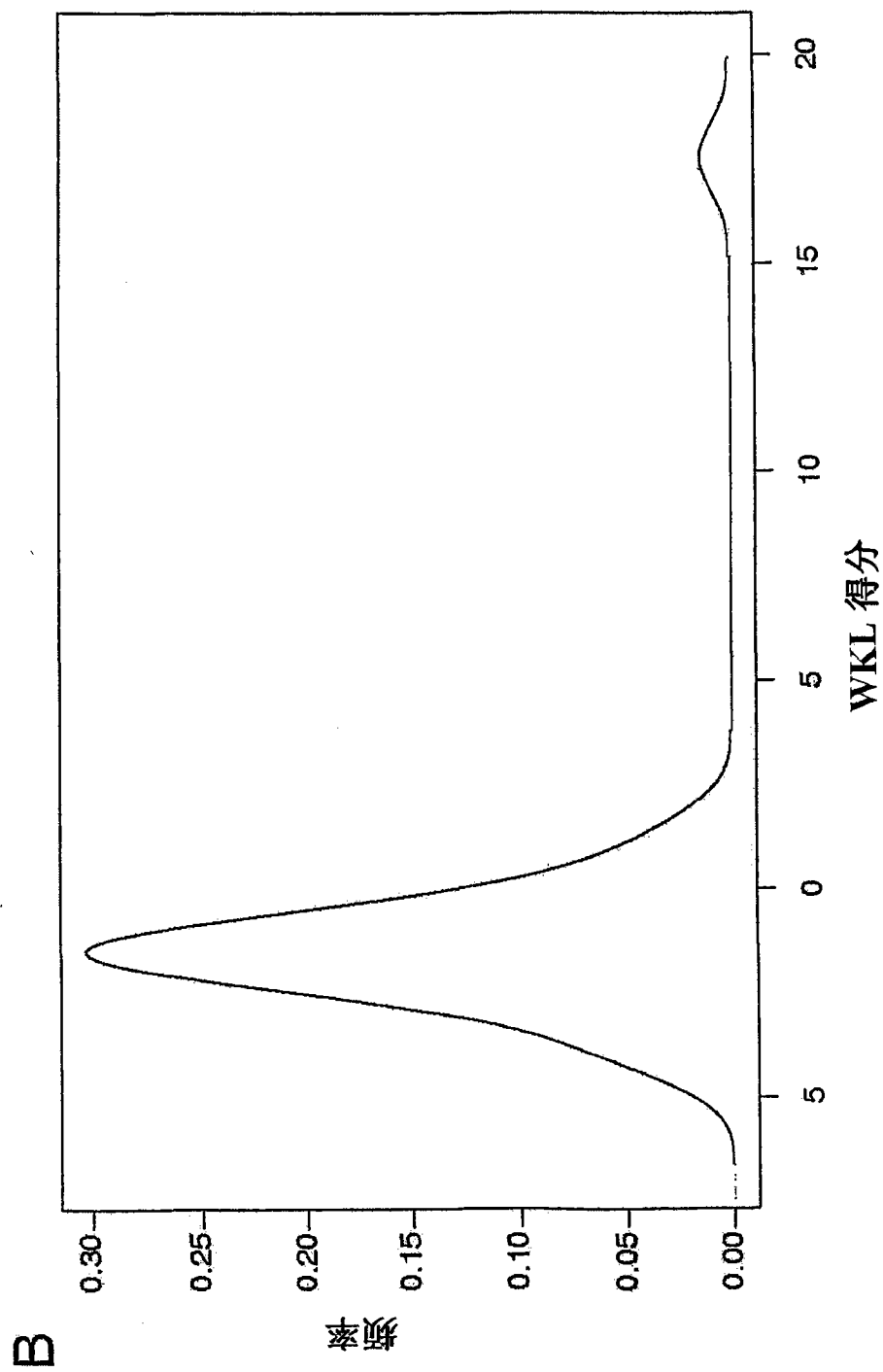


图 9B

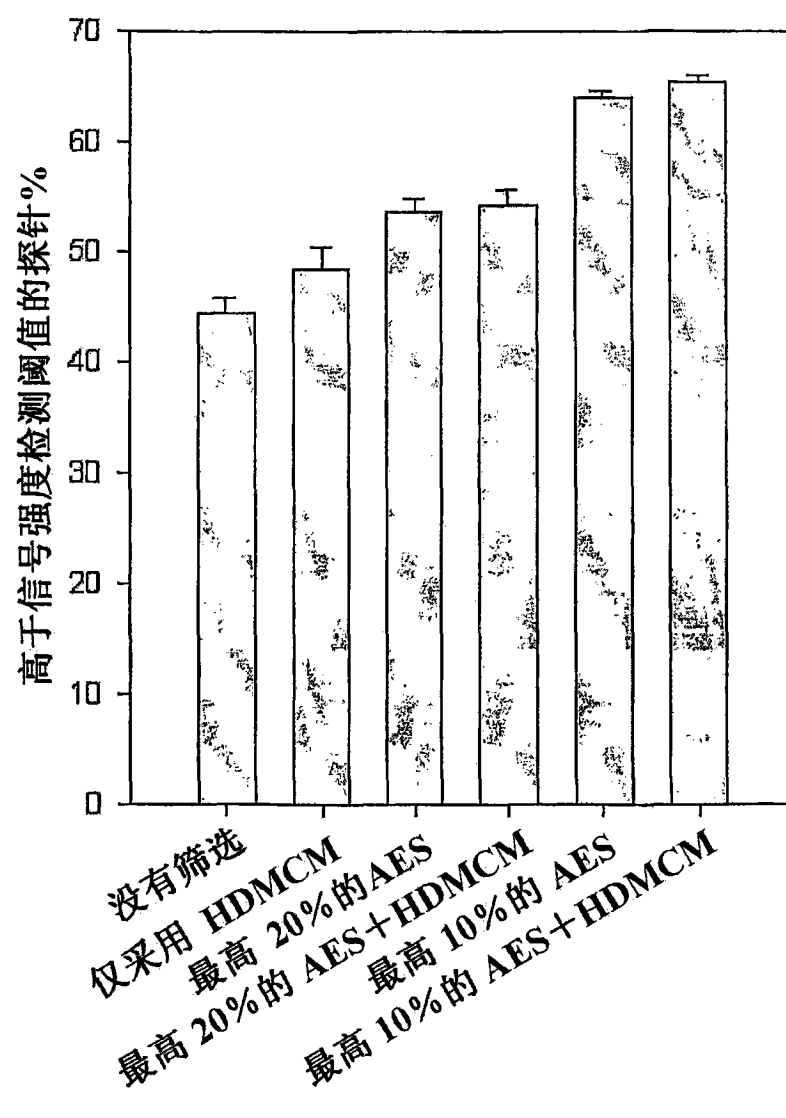


图 10

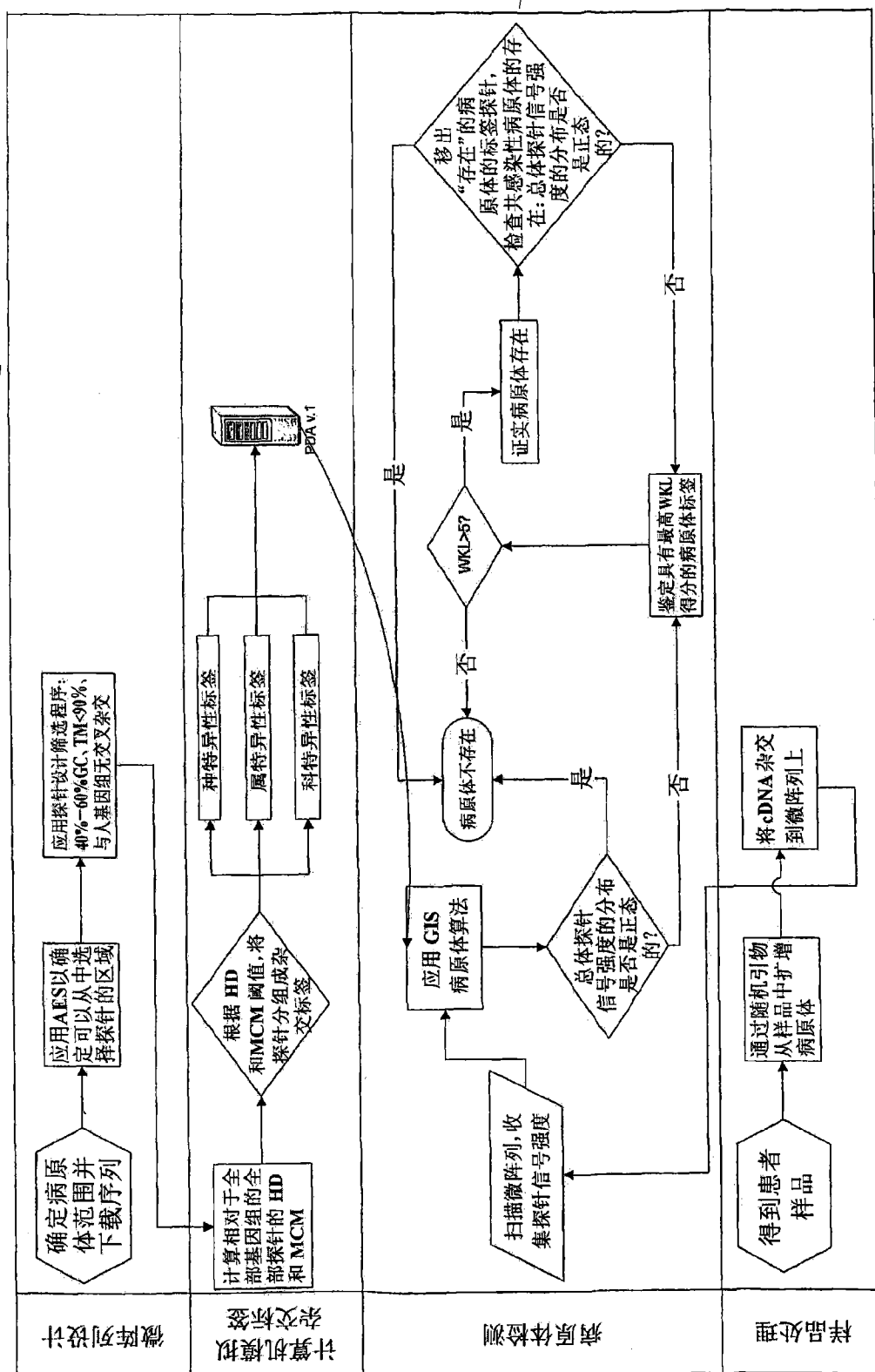


图 11

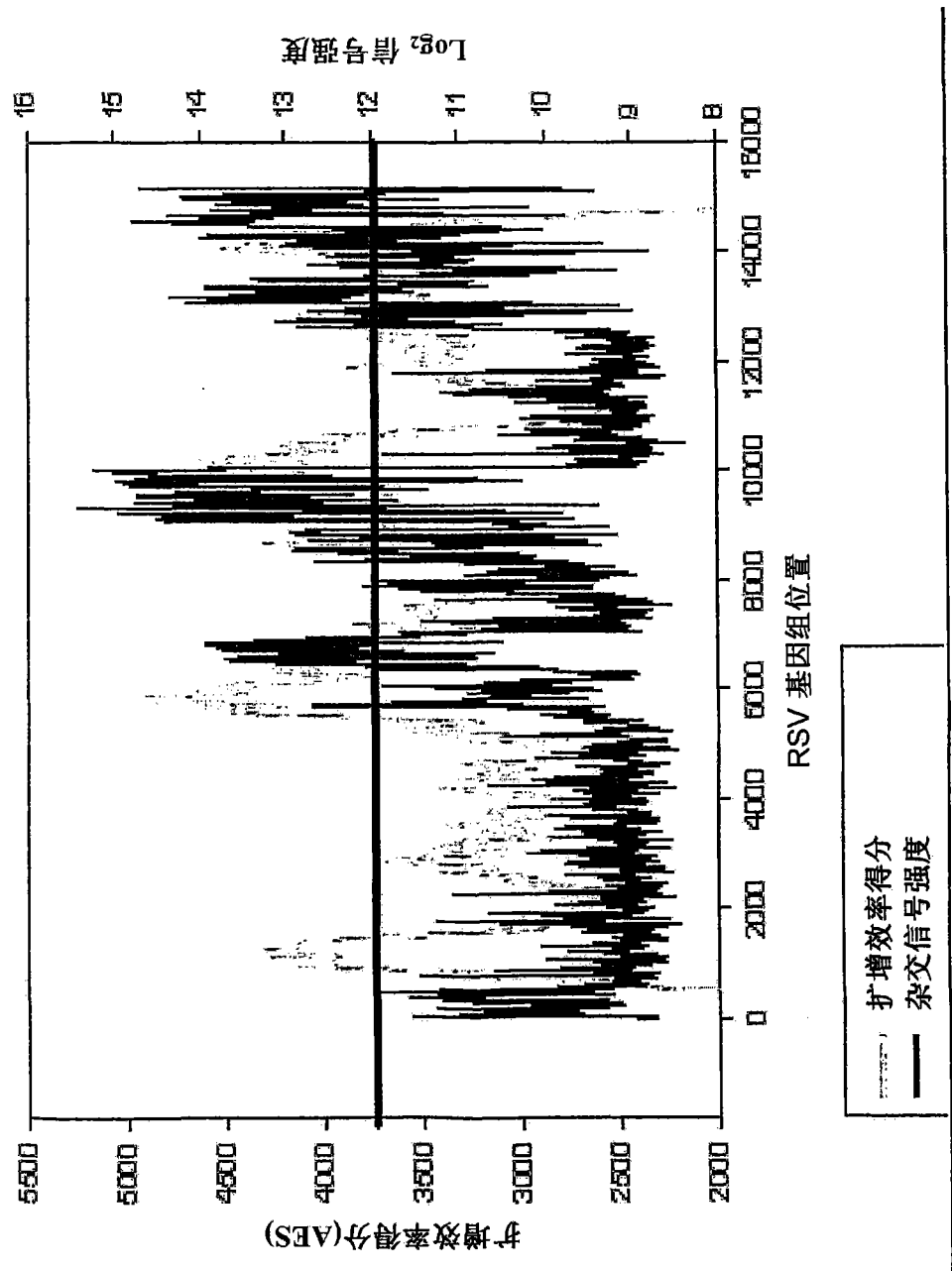


图 12

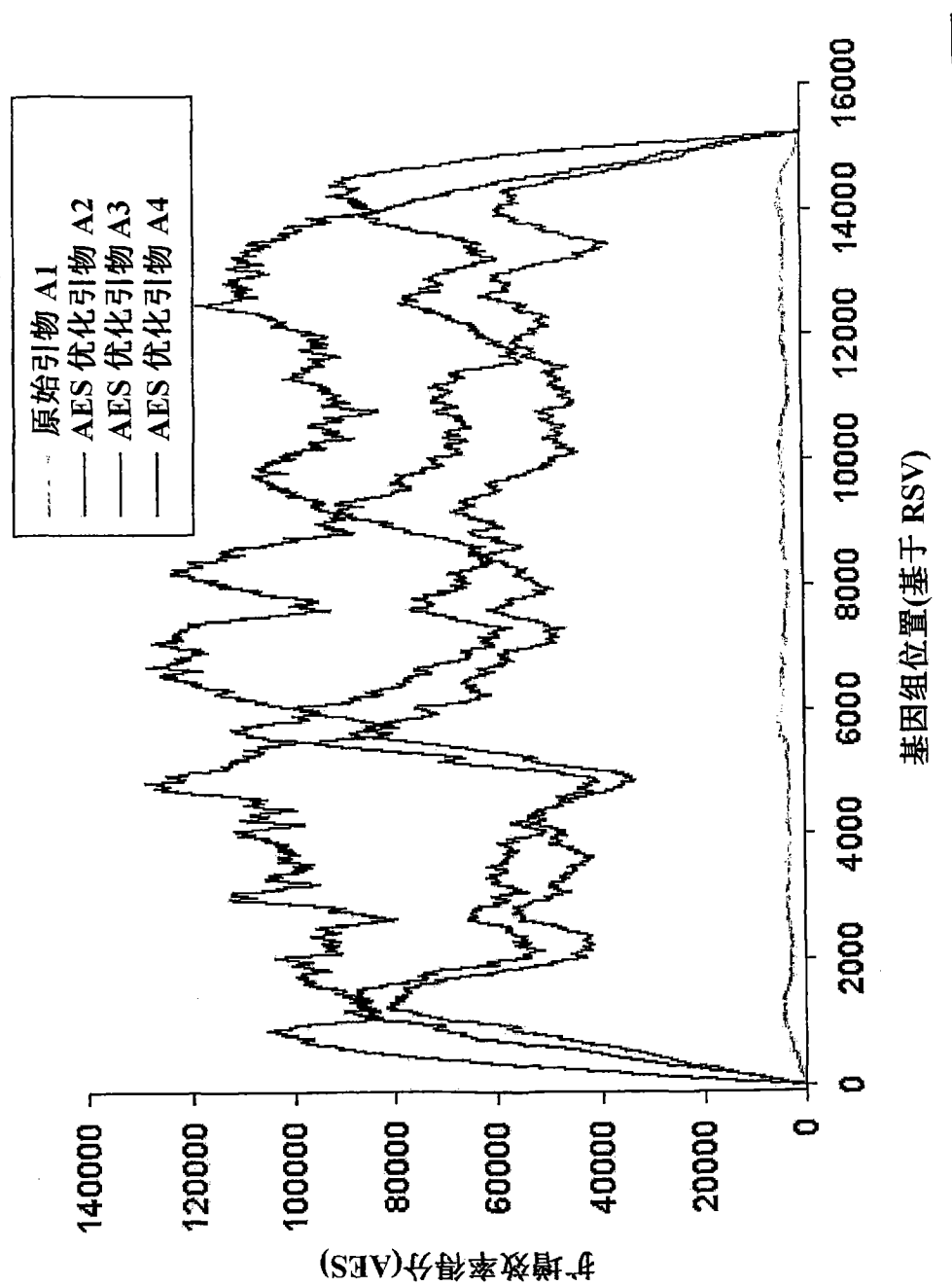


图 13

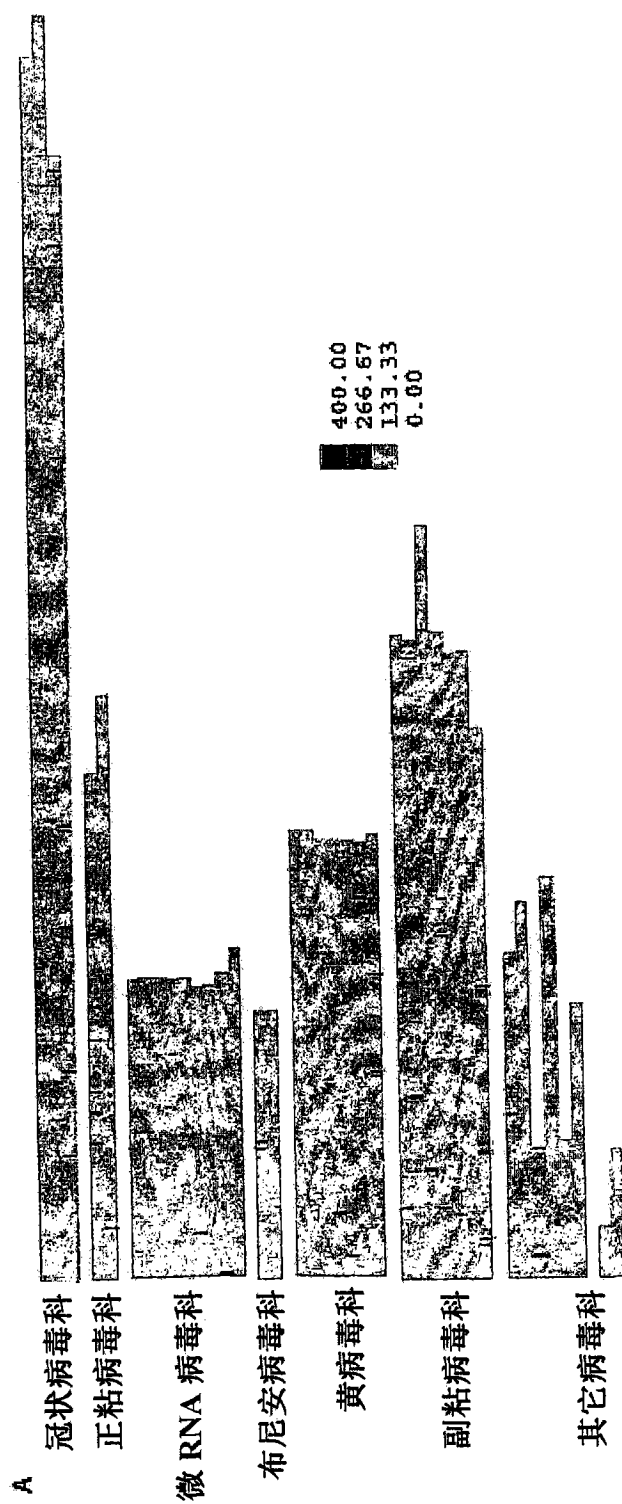


图 14A

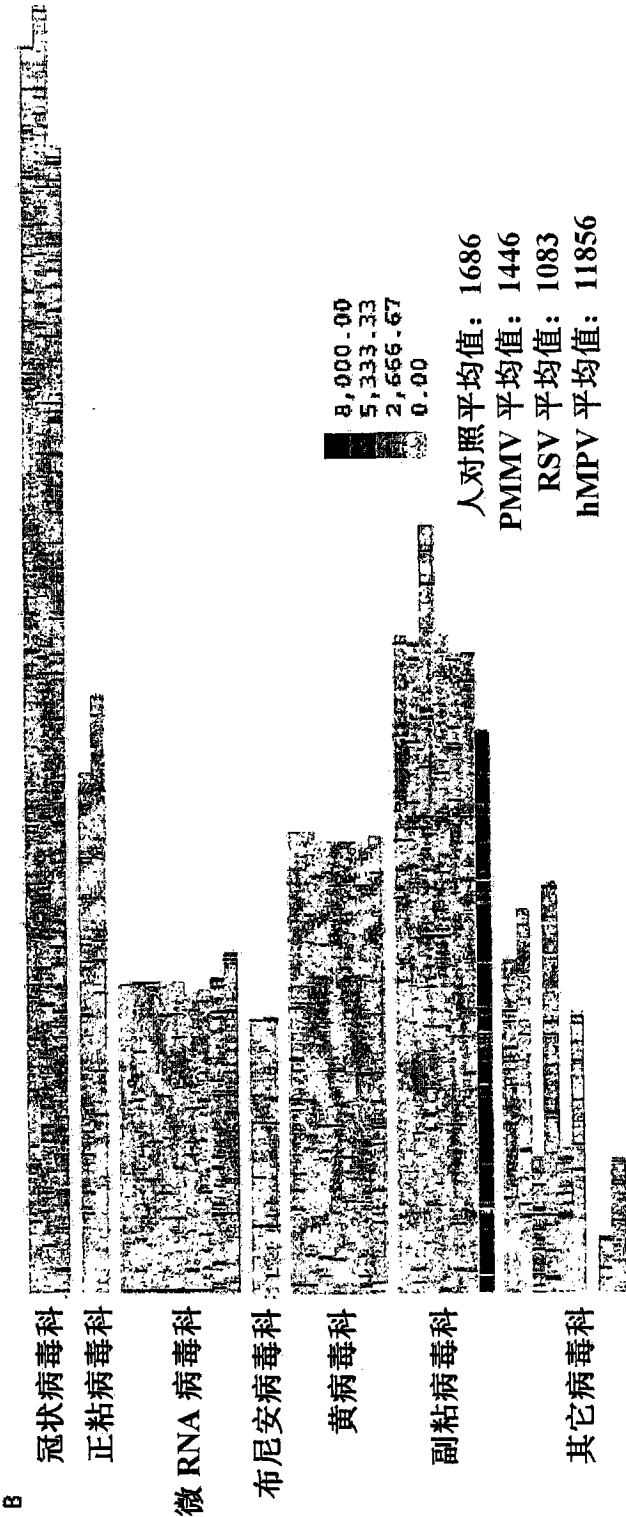


图 14B

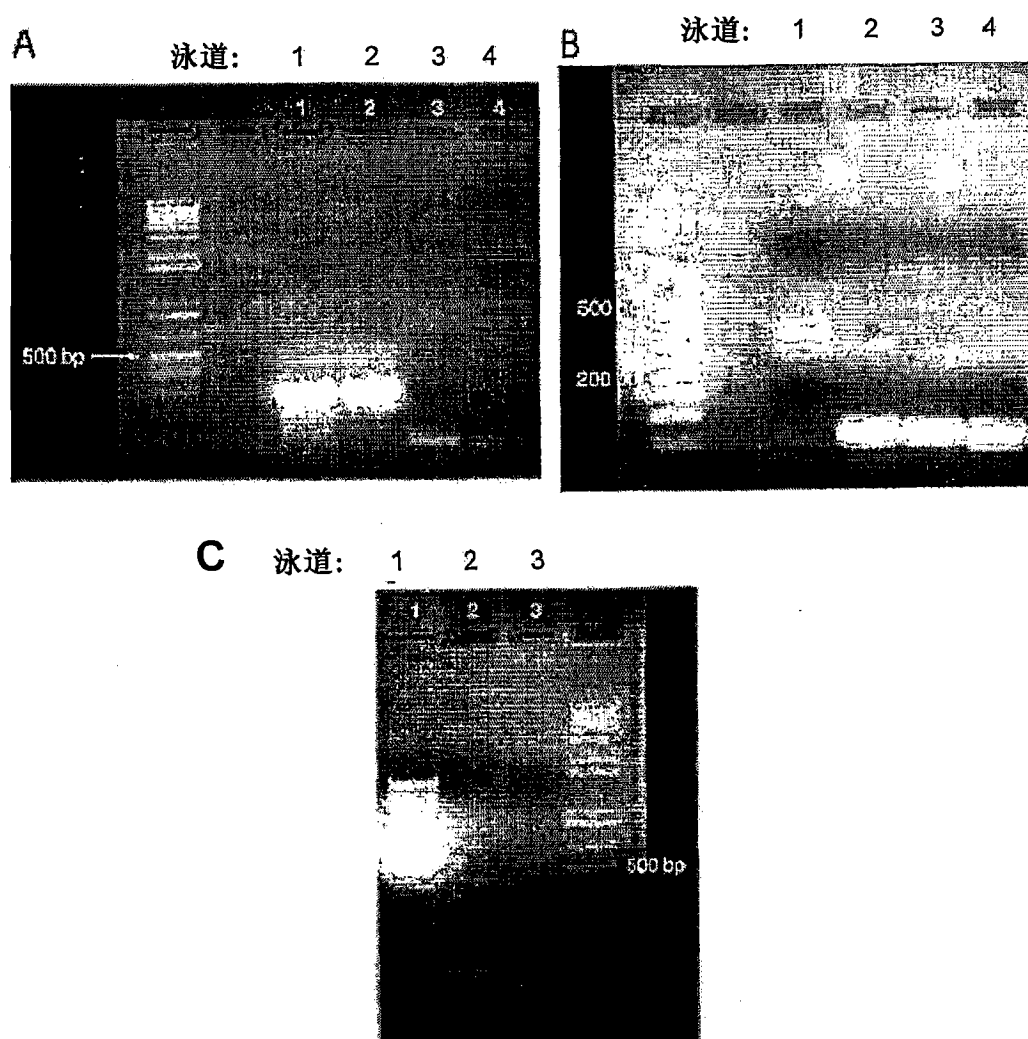


图 15 (A、B、C)