



[12] 发明专利申请公开说明书

[21] 申请号 200510009558.7

[43] 公开日 2005 年 10 月 12 日

[11] 公开号 CN 1680594A

[22] 申请日 2005.2.16
[21] 申请号 200510009558.7
[30] 优先权
[32] 2004.2.13 [33] US [31] 60/544,844
[32] 2004.12.3 [33] US [31] 60/633,062
[71] 申请人 阿菲梅特利克斯公司
地址 美国加利福尼亚州
[72] 发明人 罗伯特·J·利普舒茨
史蒂芬·P.A.·福多尔
希瓦尼·诺蒂亚尔

[74] 专利代理机构 中原信达知识产权代理有限责任
公司
代理人 杨 青 樊卫民

权利要求书 10 页 说明书 43 页

[54] 发明名称 使用核酸阵列分析甲基化状态
[57] 摘要

本方面公开了扩增核酸样品同时保留胞嘧啶的甲基化状态的方法。在某些方面，扩增的甲基化样品被甲基化敏感的修饰方法所修饰，并且通过与阵列杂交来分析，从而鉴定在起始材料中被甲基化的胞嘧啶和在起始材料中没有被甲基化的胞嘧啶。本发明还公开了检测甲基化状态的方法。在一个实施方案中，在扩增反应中引入 DNA 甲基转移酶活性，这种活性使用甲基化的基因组模板链作为指导，对新合成的 DNA 进行甲基化。

I S S N 1 0 0 8 - 4 2 7 4

1. 一种扩增基因组 DNA 同时维持甲基化状态的方法，包括：

(a) 获得含有甲基化基因组 DNA 的样品；

5 (b) 将一条或多条引物与基因组 DNA 杂交；

(c) 用 DNA 聚合酶延伸所述一条或多条引物，产生含有新合成的未甲基化的 cDNA 链和甲基化的模板链的半甲基化的杂合体；

(d) 用 DNA 甲基转移酶活性在存在甲基供体的情况下处理杂合体，其中的 DNA 甲基转移酶活性使双链 DNA 中的半甲基化位点甲基化，产生含有新合成的甲基化的 cDNA 链和甲基化的模板链的甲基化的杂合体；

(e) 将甲基化的杂合体变性；以及

(f) 将一条或多条引物与步骤 (e) 的产物杂交，并且重复步骤 (c) 和 (d) 至少 1 次以产生甲基化的扩增产物。

15

2. 权利要求 1 的方法，其中的 DNA 聚合酶是热稳定的聚合酶。

3. 权利要求 1 的方法，其中的一条或多条引物包含随机序列引物的集合。

20

4. 权利要求 1 的方法，其中的一条或多条引物包含多个位点特异性引物，它们与 CpG 岛的 1,000 个碱基中至少 15 个碱基的区域完全互补。

25

5. 权利要求 1 的方法，其中的 DNA 聚合酶是链置换酶。

6. 权利要求 5 的方法，其中的链置换酶是 phi29。

30

7. 权利要求 1 的方法，其中的 DNA 甲基转移酶活性包括 Dnmt 1 酶。

8. 权利要求 7 的方法，其中的 Dnmt 1 选自小鼠 Dnmt 1 和人类 Dnmt 1。

5 9. 权利要求 7 的方法，其中的 Dnmt 1 是重组的。

10. 权利要求 1 的方法，其中的甲基供体是 S-腺苷甲硫氨酸。

11. 权利要求 1 的方法，其中的 DNA 甲基转移酶活性是人类或
10 小鼠 Dnmt 1 的变体形式，相对于天然酶它增加了对半甲基化 DNA 的特异性。

12. 一种在基因组 DNA 样品中分析至少一个基因组区域的甲基化状态以鉴定被甲基化的区域的方法，包括：

15 (a) 按照权利要求 1 的方法扩增基因组 DNA 样品，产生含有扩增的甲基化产物的扩增样品；

 (b) 将从 (a) 得到的扩增样品用第一个限制性酶进行片段化，并且将接头与片段连接产生连接了接头的片段；

 (c) 用第二个限制性酶处理从 (b) 获得的样品的第一个等份
20 试样，并用第三个限制性酶处理从 (b) 获得的样品的第二个等份试样，其中第二和第三个限制性酶是同裂酶，其中第二个限制性酶是甲基化敏感的，而第三个限制性酶是甲基化不敏感的；

 (d) 通过 PCR 使用与接头互补的引物扩增从 (c) 获得的第一
25 和第二个等份试样，以便从第一个等份试样产生第一个扩增样品，而从第二个等份试样产生第二个扩增样品；

 (e) 将 (d) 的产物进行标记，并将标记的产物平行地与探针
阵列杂交，以便从第一个扩增样品产生第一个杂交模式，而从第二个
扩增样品产生第二个杂交模式，其中的阵列含有多个探针，每个探针
至少 15 个碱基长，并且都与选定的生物体基因组的 200 到 2,000 个碱
30 基对的限制性片段完全互补，这些限制性片段是选定的生物体的基因

组用所述第一个限制性酶消化产生的；以及

(f) 将第一个和第二个杂交模式进行比较，以鉴定在第一个模式中存在而在第二个模式中不存在的片段，其中在第一个模式中存在的片段被鉴定为甲基化的区域。

5

13. 权利要求 1 的方法，进一步包括在扩增样品中分析多个基因组区域的甲基化状态以鉴定多个没有被甲基化的区域，使用的方法包括：

将扩增的样品片段化产生片段；

10

使接头连接到片段上产生含有连接了接头的片段的样品；

把含有连接了接头的片段的样品分成至少第一个和第二个等份试样；

用甲基化不敏感的限制性酶处理第一个等份试样；

用甲基化依赖的限制性酶处理第二个等份试样；

15

使用与接头互补的引物扩增处理的第一个和第二个等份试样；

标记扩增产物，并将扩增产物同含有与已知的基因组区域互补的探针的探针阵列进行杂交，以便对于所述第一个等份试样产生第一个杂交模式，而对于所述第二个等份试样产生第二个杂交模式；

20

将第一个和第二个杂交模式进行比较，并且通过鉴定在第二个等份试样中存在而在第一个等份试样中不存在的基因组片段来鉴定未甲基化的基因组区域。

14. 权利要求 12 的方法，其中的第一个酶的识别位点中不含 CpG。

25

15. 权利要求 12 的方法，其中的第一个酶选自 XbaI、HindIII 和 BglII。

30

16. 权利要求 12 的方法，其中的第二个限制性酶是 HpaII，而第三个限制性酶是 MspI。

17. 权利要求 13 的方法，其中的甲基化依赖性酶是 McrBC。

5 18. 权利要求 12 的方法，其中的探针阵列含有至少 100,000 个不同的连接到固相支持物上的探针，其中探针的位置是确定的或可确定的，以及其中阵列的探针由计算机系统选择，所述计算机系统为阵列选择探针的方法包括：

 模拟扩增样品的片段化，产生来自片段化的第一个片段列表，其中第一个列表包括预计长度的片段；

10 通过鉴定长度在 200 到 2,000 个碱基对之间的片段从第一个列表产生第二个片段列表；

 通过鉴定含有第二和第三个限制性酶的识别位点的片段从第二个列表产生第三个片段列表；以及

 选择与第三个列表中的多个片段互补的探针。

15

 19. 权利要求 13 的方法，其中的探针阵列含有至少 100,000 个不同的连接到固相支持物上的探针，其中探针的位置是确定的或可确定的，以及其中阵列的探针由计算机系统选择，所述计算机系统为阵列选择探针的方法包括：

20 模拟扩增样品的片段化，产生来自片段化的第一个片段列表，其中第一个列表包括预计长度的片段；

 通过鉴定长度在 200 到 2,000 个碱基对之间的片段从第一个列表产生第二个片段列表；

25 通过鉴定含有甲基化依赖性酶的识别位点的片段从所述第二个列表产生第三个片段列表；以及

 选择与第三个列表中的多个片段互补的探针。

20. 权利要求 18 的方法，其中的固相支持物是多个珠子。

30 21. 权利要求 19 的方法，其中的固相支持物是多个珠子。

22. 权利要求 1 的方法，进一步包括对扩增样品进行分析以鉴定多个在基因组样品中被甲基化了的胞嘧啶和多个在基因组样品中没有被甲基化的胞嘧啶，使用的方法包括通过不将甲基化的胞嘧啶转化为尿嘧啶的方法将扩增样品中未甲基化的胞嘧啶转化为尿嘧啶，以及确定存在于多个胞嘧啶位置上的序列。

23. 权利要求 22 的方法，其中所述转化步骤选自亚硫酸氢盐处理或用胞苷脱氨酶处理。

24. 权利要求 22 的方法，其中所述转化步骤包含了亚硫酸氢盐处理和用胞苷脱氨酶处理。

25. 权利要求 23 或 24 的方法，其中的胞苷脱氨酶是活化诱导的胞苷脱氨酶。

26. 权利要求 22、23、24 或 25 的方法，其中所述确定步骤包括将扩增样品标记，并且使标记样品与探针阵列杂交，其中探针阵列含有质询多个胞嘧啶序列的探针，以确定基因组样品中正在被质询的胞嘧啶是否被甲基化了。

27. 权利要求 26 的方法，其中被质询的胞嘧啶是 CpG 二核苷酸的一部分。

28. 权利要求 27 的方法，其中的阵列质询至少 10,000 个胞嘧啶的甲基化。

29. 一种探针阵列，含有至少 100,000 个不同的连接到固相支持物上的探针，其中探针的位置是确定的或可确定的，以及其中阵列的探针由计算机系统选择，所述计算机系统为阵列选择探针的方法包

括：

模拟第一个核酸样品被第一个限制性酶的片段化，产生来自片段化的第一个片段列表，其中第一个列表包括预计长度的片段；

5 通过鉴定在选定的大小范围内的片段从第一个列表产生第二个片段列表；

通过鉴定第二个列表中含有第二个限制性酶的识别位点的片段从所述第二个列表产生第三个片段列表，其中所述第二个限制性酶是甲基化依赖的限制性酶或是甲基化敏感的限制性酶；以及

10 为阵列选择至少 100,000 个不同的探针，其中每个探针至少 15 个碱基，并且与第三个列表中的片段完全互补。

30. 权利要求 29 的阵列，其中第一个限制性酶是两种或多种限制性酶的组合。

15 31. 权利要求 29 的阵列，其中第二个限制性酶是 McrBC。

32. 权利要求 29 的阵列，其中第二个限制性酶是 Hpa II。

20 33. 在第一个含有基因组 DNA 的核酸样品中分析多个不同 CpG 位点的甲基化的方法，包括：

(a) 按照权利要求 1 的方法扩增基因组 DNA 样品，产生含有扩增的甲基化产物的第一个扩增样品；

(b) 将第一个扩增样品用第一个限制性酶片段化，并且使接头与片段连接产生含有连接了接头的片段的第二个样品；

25 (c) 用亚硫酸氢钠处理第二个样品产生第三个样品；

(d) 通过 PCR 使用与接头互补的引物扩增第三个样品中的至少一些连接了接头的片段；

(e) 将步骤 (d) 的产物片段化，并把片段进行末端标记；

30 (f) 将标记的产物与探针阵列杂交产生杂交模式，其中的阵列含有多个探针对，其中每个探针对中的第一个探针在 C 是甲基化的

情况下与亚硫酸氢钠处理后的第一个 CpG 位点互补，而第二个探针在 C 是未甲基化的情况下与同样的区域互补；以及

(g) 对杂交模式进行分析，以确定对于多个 CpG 位点中的每一个来说，在第一个核酸样品中所述位点是否是未甲基化的。

5

34. 一种减少第一个含有甲基化基因组 DNA 的核酸样品的复杂性以产生复杂性降低的样品的方法，包括：

将第一个核酸样品用第一个限制性核酸内切酶片段化，产生含有限制性片段的第二个核酸样品；

10

将接头连接到第二个核酸样品的限制性片段上，产生含有连接了接头的片段的第三个核酸样品；

将第三个核酸样品用甲基化依赖性核酸内切酶片段化，产生第四个核酸样品；以及

15

用与接头互补的引物通过 PCR 对第四个核酸样品进行扩增，从而产生复杂性降低的样品。

35. 权利要求 34 的方法，其中的甲基化依赖性酶是 McrBC。

20

36. 权利要求 34 的方法，其中第一个核酸样品选自血液样品、组织样品和肿瘤样品。

37. 一种获得样品的杂交模式特征的方法，包括：

从所述样品获得基因组 DNA 样品；

按照权利要求 34 的方法减少核酸样品的复杂性；

25

将复杂性减少的样品片段化，并用可检测的标记物标记片段；以

及

将标记的片段与核酸探针阵列杂交，从而获得杂交模式。

30

38. 一种将未知核酸样品与已知核酸样品进行比较的方法，包括：

按照权利要求 30 的方法为所述未知样品产生第一个杂交模式；
为所述已知样品获得第二个杂交模式，其中的第二个杂交模式是
按照权利要求 37 的方法产生的；以及
将第一个杂交模式与第二个杂交模式进行比较。

5

39. 一种将肿瘤归类为已知肿瘤类型的方法，包括：

按照权利要求 37 的方法为所述未知样品产生第一个杂交模式；

从多个已知类型的肿瘤获得多个第二个杂交模式，其中第二个杂交模式是分别按照权利要求 37 的方法从来自已知类型肿瘤样品产生的；

10

将第一个杂交模式与各种第二个杂交模式进行比较，以鉴定与第一个杂交模式最匹配的第二个杂交模式；以及

将肿瘤归类于杂交模式最匹配的已知肿瘤的类型。

15

40. 一种检测基因组 DNA 样品中甲基化的基因组区域的方法，包括下列步骤：

a、用亚硫酸氢盐处理基因组 DNA 样品；

b、将基因组 DNA 样品片段化；

c、将接头连接到片段上；

20

d、对连接了接头的片段进行扩增；

e、用可检测的标记物标记扩增的片段，并将标记的片段与阵列杂交，以产生杂交模式；以及

f、将上述杂交模式与对照进行比较以鉴定甲基化的基因组区域。

25

41. 权利要求 40 的方法，其中步骤 (a) 在步骤 (b) 之前进行，以及其中步骤 (c) 包括使用 RNA 连接酶将接头连接到片段的 5' 末端，除去 3' 磷酸，并且将接头连接到片段的 3' 末端。

30

42. 权利要求 40 的方法，进一步包括通过将亚硫酸氢盐处理的

样品与针对 5-甲基胞嘧啶的抗体或与结合 5-甲基胞嘧啶的蛋白和针对所述蛋白的抗体一起保温，并且分离抗体复合物，从而使获得的样品富集含有 5-甲基胞嘧啶的片段，其中所述分离步骤在步骤（c）之前进行。

5

43. 权利要求 40 的方法，其中步骤（b）在步骤（a）之前进行。

44. 权利要求 40 的方法，其中的亚硫酸氢盐处理包括与 8-10M 的亚硫酸氢盐保温 5 分钟到 1 小时。

10

45. 一种分析核酸样品中一个或多个胞嘧啶的甲基化状态的方法，所述方法包括：扩增核酸样品中的至少一部分序列，其中起始核酸样品中至少一部分序列的甲基化模式在扩增步骤中被拷贝了，从而产生甲基化的扩增样品；将甲基化的扩增样品进行区别修饰甲基化的胞嘧啶和未甲基化的胞嘧啶的处理；以及通过与核酸探针阵列的杂交检测扩增样品中至少一个胞嘧啶的甲基化状态。

15

46. 权利要求 45 的方法，其中区别修饰甲基化的胞嘧啶和未甲基化的胞嘧啶的处理是亚硫酸氢盐处理。

20

47. 权利要求 45 的方法，其中区别修饰甲基化的胞嘧啶和未甲基化的胞嘧啶的处理是使用活化诱导的胞苷脱氨酶的处理。

48. 权利要求 45 的方法，其中区别修饰甲基化的胞嘧啶和未甲基化的胞嘧啶的处理包括使用活化诱导的胞苷脱氨酶的处理和亚硫酸氢盐处理。

25

49. 权利要求 46、47 或 48 的方法，其中所述探针阵列包含与用亚硫酸氢盐处理或用活化诱导的胞苷脱氨酶处理后产生的多个不同序列完全互补的探针。

30

50. 权利要求 49 的方法，其中的探针阵列包含与用亚硫酸氢盐处理或用活化诱导的胞苷脱氨酶处理多个选定的基因组区域后产生的所有可能的序列组合完全互补的探针。

5

使用核酸阵列分析甲基化状态

5 相关申请

本申请要求 2004 年 2 月 13 日提交的美国临时申请 No. 60/544,844 和 2004 年 12 月 3 日提交的美国临时申请 No. 60/633,062 的优先权。上述申请的全部公开内容在此以其全文引为参考。

10 发明领域

本发明涉及扩增样品以保存后生信息的方法以及使用核酸阵列检测甲基化的方法。

发明背景

15 高等真核生物的基因组含有修饰的核苷 5-甲基胞嘧啶 (5-meC)。该修饰通常作为二核苷酸 CpG 的一部分。将胞嘧啶转化为 5-甲基胞嘧啶的反应包括将完整双螺旋中的靶胞嘧啶翻转出来，然后通过甲基转移酶从 S-腺苷甲硫氨酸中转移甲基基团 (Klimasauskas 等, Cell 76:357-369, 1994)。这个酶促转化反应是已知的脊椎动物中存在的唯一后生的 DNA 修饰，对于正常的胚胎发育是必要的 (Bird, Cell 70:5-8, 1992; Laird 和 Jaenisch, Human Mol. Genet. 3:1487-1495, 1994; 以及 Li 等, Cell 69:915-926, 1992)。

25 CpG 二核苷酸在人类基因组中的出现频率仅仅是统计预期频率的大约 20%，这可能是因为 5-甲基胞嘧啶到胸腺嘧啶的自发脱氨基作用 (Schorer 等, Proc. Natl. Acad. Sci. USA 89:957-961, 1992)。CpG 以大约是预期的频率水平出现的区域被称为 “CpG 岛” (Bird, A.P. (1986) Nature 321, 209-213)。这些区域占脊椎动物基因组的大约 1%，占总的 CpG 二核苷酸数量的大约 15%。CpG 岛的长度一般在 0.2 到 1kb 之间，并位于许多看家基因和组织特异性基因的上游。CpG 岛通常位于被转

30

录区域的上游，但是也可能延伸到被转录区域内。大约 2-4%的位于鸟嘌呤 5'端的胞嘧啶是甲基化的。

5 DNA 甲基化是基因表达的后生决定因素。CpG 甲基化的模式是可遗传和组织特异性的，并与基因表达相关。DNA 甲基化也与其它
细胞过程相关，这些过程包括染色质结构、基因组印记、雌性中体细胞 X 染色体的失活和 DNA 复制的定时。当基因高度甲基化时，它被
表达的可能性降低，这可能是因为 CpG 甲基化阻止了转录因子识别它们的同源结合位点。与甲基化 DNA 结合的蛋白也可以召集组蛋白脱
10 乙酰酶以缩合临近的染色质。一般来说，无转录活性的基因含有 5-甲基胞嘧啶，而有转录活性的基因不含 5-甲基胞嘧啶。因此，鉴别基因组中含有 5-甲基胞嘧啶的位点，对于理解在正常发育和疾病例如癌症
的过程中，基因表达的细胞类型特异性程序以及基因表达图如何改变都是非常重要的。CpG 岛中 DNA 甲基化模式的精确作图，对于理解
15 多种多样的生物过程，例如印记基因的调控，X 染色体的失活，以及在人类癌症中由甲基化增加而引起的肿瘤抑制基因的沉默，已经变得
非常重要了。

胞嘧啶的甲基化可以降低基因的表达，这可以通过例如破坏局部的
20 染色质结构，抑制转录因子与 DNA 的结合，或者通过召集能够与
甲基化序列发生特异性相互作用并阻止转录因子结合的蛋白来进行。
甲基化模式的改变已显示与癌症相关。肿瘤抑制基因启动子中的 CpG
寡聚核苷酸的甲基化可以导致它们的失活。正常甲基化过程的改变也
已显示与基因组不稳定性有关 (Lengauer 等 Proc. Natl. Acad. Sci. USA
25 94:2545-2550, 1997)。这种异常的后生性改变可以在许多类型的癌症
中发现，并能用作致癌转化的潜在标记。

在本公开中引用的所有文件，即出版物和专利申请，包括上述引
用文件，在此为所有的目的以其全文引为参考，其程度如同每个单独
30 的文件被特定和分别地指明在此以其全文引为参考。

发明概述

本发明的一个方面公开了一种扩增基因组 DNA 的方法，它拷贝了胞嘧啶的甲基化。甲基化的基因组 DNA 样品被获得后，使用酶法延伸与 DNA 杂交的引物来扩增，引物可以是随机序列，也可以是位点特异性的。引物与甲基化的 DNA 杂交，并使用 DNA 聚合酶进行延伸，从而产生半甲基化的杂交体，它包括新合成的未甲基化的 cDNA 链和甲基化的模板链。然后可以用 DNA 甲基转移酶在存在甲基供体的情况下对半甲基化的杂交体进行处理。DNA 甲基转移酶可以对双链 DNA 中的半甲基化位点进行甲基化，产生甲基化的杂交体，它包括新的甲基化的 cDNA 和甲基化的模板链。甲基化的杂交体可以被变性，然后如上述再次扩增。获得的产物是扩增的甲基化的样品。扩增的甲基化的样品可以被直接分析以检测甲基化的和未甲基化的胞嘧啶。

DNA 聚合酶可以是热稳定的聚合酶或链置换聚合酶，例如 phi29。在有些方面，DNA 甲基转移酶活性包括 Dnmt 1 酶，其可以是例如人类或小鼠的 Dnmt 1，并且可以从生物来源纯化，也可以是重组蛋白或融合蛋白。Dnmt 1 活性可以是人类或小鼠的 Dnmt 1 的变体形式，与天然酶相比增加了对半甲基化 DNA 的特异性。在优选情况下，甲基供体是 S-腺苷甲硫氨酸。

在一个方面，按照上述方法扩增的 DNA 可以用甲基化特异的限制性酶处理以检测选定的胞嘧啶的甲基化情况。扩增的甲基化的 DNA 可以用第一种限制性酶片段化，然后可以给片段连接上接头。与接头连接的片段的等分试样可以平行地采用对甲基化具有不同敏感性的同裂酶进行片段化，例如一种酶对甲基化敏感而另一种酶对甲基化不敏感，或者一种酶是甲基化依赖性的而另一种酶是甲基化不敏感的或甲基化敏感的。在与接头连接的片段被片段化后，片段可以用接头序列的引物进行扩增。在接头之间被切开的片段将不能被扩增。扩增的片段可以被标记，并与检测是否存在不同片段的阵列杂交。可以对杂交

模式进行分析以确定哪个片段被扩增了以及哪个片段被切开了从而没有被有效地扩增。扩增可以使用任何引物介导的延伸方法，例如 PCR 或 MDA。

5 在一个方面，按照上述方法扩增 DNA 样品，然后使用一种将扩增样品中未甲基化的胞嘧啶转化为尿嘧啶但是对甲基化的胞嘧啶不进行修饰的方法来分析胞嘧啶的甲基化。接着通过测定产物在目的位置处的序列来确定特定胞嘧啶的甲基化。如果它仍然是胞嘧啶，它就被甲基化了，而如果它现在是尿嘧啶，它就没有被甲基化。转化的方法
10 包括化学和酶学的方法，或二者的组合。检测可以通过与探针阵列的杂交。探针可以被设计以查询选定的胞嘧啶。

 在有些方面，使用了针对 5-甲基胞嘧啶的抗体或针对与 5-甲基胞嘧啶选择性结合的蛋白的抗体。样品中的甲基化 DNA 可以通过免疫学的方法富集，例如用针对 5-甲基胞嘧啶的抗体或与 5-甲基胞嘧啶结合
15 的蛋白和针对该蛋白的抗体进行免疫沉淀。

 在一个方面，探针阵列包括针对预计的片段的探针。生物体的基因组可以通过计算机模拟（in silico）消化进行分析以预测限制性片段的大小和序列，并鉴定具有 CpG 的片段。在阵列上可以包括针对目的
20 片段的探针。在设计阵列时也可以把扩增方法考虑在内。例如，接头介导的 PCR 扩增一般来说对大约 200 到 2000 碱基对的片段扩增最为有效，所以在这个大小范围内的片段可以作为探针的靶。依赖于被比较的酶的组合可以确定片段中的甲基化。例如，如果在已经用甲基化
25 依赖性的酶消化的样品中存在含有相应限制性位点的片段，该片段极可能没有被甲基化。HpaII 和 MspI 是一对对甲基化具有不同敏感性的同裂酶的例子，而 McrBC 是一种甲基化依赖性的酶。

 在一个方面，公开了按照本发明的方法设计的阵列。阵列可以在
30 固相支持物上已知的或可确定的位置处存在超过 1,000,000、2,000,000

或 5,000,000 个不同的探针。阵列中的探针可以被设计用来检测在按照本发明的方法扩增和处理后，在样品中预计会存在的片段。

按照本发明的方法扩增的基因组样品可以被分析，以确定样品中一个或多个 CpG 位点的甲基化状态。可以使用的方法包括亚硫酸氢钠处理，然后检测没有被处理所修饰仍然保留为 C/G 碱基对的甲基化的位点，以及未被甲基化但被处理所修饰从而导致引入了 A/T 碱基对的位点。C/G 或 A/T 碱基对的存在可以通过与探针阵列的杂交来检测，该阵列的探针与带有或不带有由亚硫酸氢钠处理引入的变化的含 CpG 的区域完全互补。A/T 或 C/G 的检测以类似于 SNP 的等位基因特异性杂交检测的方式进行。

在另一方面，本方面公开了一种通过用甲基化依赖性的核酸内切酶将样品片段化从而减少复杂性的方法。样品用其切割不依赖于甲基化的限制性酶进行片段化，然后给片段连接上接头。连接了接头的片段再用甲基化依赖性的酶进行消化，对甲基化依赖性的核酸内切酶的消化具有抗性的那部分片段被扩增。扩增的片段可以与阵列杂交以产生表明了该样品甲基化状态的特征的杂交图。来自血液、组织和肿瘤

的样品可以被分析以产生杂交图。这样产生的杂交图可以与从已知样品同样产生的杂交图进行比较。通过对杂交图的比较可以对未知样品进行分类。

发明详述

a) 总则

本发明有许多优选实施方案，在细节方面正如本领域的专业人员所知依赖于许多专利、申请或其它参考文献。因此，当在下文中引用或重复专利、申请或其它参考文献时，应该理解它为所有的目的以及被重新引用的命题而以其全部引为参考。

除非上下文中有清楚的说明，用于本申请中的单数形式“一个”，

“一种”和“该”包括了复数的指称。例如术语“一种试剂”包括多种试剂，还包括其混合物。

5 个体不限于人，也可以是其它的生物体，包括但不限于哺乳动物、植物、细菌或来自上述任何生物体的细胞。

10 贯穿本申请，本发明的各个方面以范围的形式加以陈述。应该理解以范围的形式进行的描述仅仅是为了方便和简便，不应该被解释为对发明范围的硬性限制。因此，范围的描述应该被视为特异性地公开了所有可能的子范围，以及范围内所有单个数值。例如，对范围的描述，从1到6应该被视为特异性地公开了子范围，例如从1到3、从1到4、从1到5、从2到4、从2到6、从3到6等，以及该范围内单个数值，例如1、2、3、4、5和6。不论范围有多宽都是这样。

15 除非特别地说明，本发明的实施都可以使用本领域技术范围内的有机化学、聚合物技术、分子生物学（包括重组技术）、细胞生物学、生物化学和免疫学常规技术和描述。这样的常规技术包括聚合物阵列合成、杂交、连接和使用标记物检测杂交。适当技术的特别的说明可以参考下面的事实例。但是，其它等同的常规步骤当然也可以使用。

20 这些常规的技术和描述可以在标准的实验室手册中发现，例如 *Genome Analysis: A Laboratory Manual Series (Vols. I-IV)*（基因组分析：实验室手册丛书，第1-4卷），*Using Antibodies: A Laboratory Manual*（使用抗体：实验室手册），*Cells: A Laboratory Manual*（细胞：实验室手册），*PCR Primer: A Laboratory Manual*（PCR引物：实验室手册）以及

25 *Molecular Cloning: A Laboratory Manual*（分子克隆实验指南）（都来自冷泉港实验室出版社），Stryer, L. (1995)主编 *Biochemistry* (4th Ed.)（生物化学，第4版）Freeman, New York, Gait, "Oligonucleotide Synthesis: A Practical Approach"（寡聚核苷酸合成：实用方法）1984, IRL Press, London, Nelson and Cox (2000), Lehninger, *Principles of*

30 *Biochemistry* 3rd Ed.（生物化学原理，第3版），W.H. Freeman Pub., New

York, NY 以及 Berg 等(2002) Biochemistry, 5th Ed. (生物化学, 第 5 版), W.H. Freeman Pub., New York, NY, 所有这些在此为所有目的以其全文引为参考。

5 本发明可以使用固相基底, 包括在一些优选实施方案中的阵列。
可用于聚合物(包括蛋白)阵列合成的方法和技术已经在下列专利中
描述了, 包括美国系列号 09/536,841、WO 00/58516、美国专利 Nos.
5,143,854、5,242,974、5,252,743、5,324,633、5,384,261、5,405,783、
5,424,186、5,451,683、5,482,867、5,491,074、5,527,681、5,550,215、
10 5,571,639、5,578,832、5,593,839、5,599,695、5,624,711、5,631,734、
5,795,716、5,831,070、5,837,832、5,856,101、5,858,659、5,936,324、
5,968,740、5,974,164、5,981,185、5,981,956、6,025,601、6,033,860、
6,040,193、6,090,555、6,136,269、6,269,846 和 6,428,752, 以及 PCT
申请号 PCT/US99/00730(国际公布号 WO 99/36760)和 PCT/US01/04285
15 (国际公布号 WO 01/58593), 在此为所有目的以其全文引为参考。

在特定的实施方案中描述合成技术的专利包括美国专利 Nos.
5,412,087、6,147,205、6,262,216、6,310,189、5,889,165 和 5,959,098。
在许多上述专利中也描述了核酸阵列, 但是同样的技术也可用于多肽
20 阵列。

用于本发明的核酸阵列包括那些可以从 Affymetrix 公司(Santa
Clara, CA)购买到的商标名为 GeneChip®的产品。示例的阵列显示在
affymetrix.com 的网点上。

25 本发明也考虑了结合到固相基底上的聚合物的许多应用。这些应
用包括基因表达监控、分布、文库筛选、基因分型和诊断。基因表达
监控和分布的方法示于美国专利 Nos. 5,800,992、6,013,449、6,020,135、
6,033,860、6,040,138、6,177,248 和 6,309,822 中。基因分型及其应用
30 示于美国系列号 10/442,021、10/013,598 (美国专利申请公布号

20030036069), 以及美国专利 Nos. 5,856,092、6,300,063、5,858,659、6,284,460、6,361,947、6,368,799 和 6,333,179 中。其它的应用体现在美国专利 Nos. 5,871,928、5,902,723、6,045,996、5,541,061 和 6,197,506 中。

5

本发明也考虑了某些优选实施方案中的样品制备方法。在基因分型之前或同时, 基因组样品可以通过多种机制进行扩增, 其中有些使用了 PCR。参见例如 PCR Technology: Principles and Applications for DNA Amplification (PCR 技术: DNA 扩增的原理和应用) (H.A. Erlich 主编, Freeman Press, NY, NY, 1992)、PCR Protocols: A Guide to Methods and Applications (PCR 方案: 方法和应用指南) (Innis 等主编, Academic Press, San Diego, CA, 1990)、Mattila 等, Nucleic Acids Res. 19, 4967 (1991)、Eckert 等, PCR Methods and Applications 1, 17 (1991)、PCR (McPherson 等主编, IRL Press, Oxford) 和美国专利 Nos. 4,683,202、4,683,195、4,800,159、4,965,188 和 5,333,675, 在此每一个为所有的目的以其全文引为参考。样品可以在阵列上扩增。参见例如美国专利 No. 6,300,070 和美国系列号 No. 09/513,300, 在此引为参考。

10
15

其它适当的扩增方法包括连接酶链反应 (LCR) (例如 Wu 和 Wallace, Genomics 4, 560 (1989), Landegren 等, Science 241, 1077 (1988) 和 Barringer 等 Gene 89:117 (1990))、转录扩增 (Kwoh 等, Proc. Natl. Acad. Sci. USA 86, 1173 (1989) 和 WO88/10315)、自持序列复制 (Guatelli 等, Proc. Nat. Acad. Sci. USA, 87, 1874 (1990) 和 WO90/06995)、靶多核苷酸序列的选择性扩增 (美国专利 No. 6,410,276)、共有序列引发的聚合酶链反应 (CP-PCR) (美国专利 No. 4,437,975)、随意引发的聚合酶链反应 (AP-PCR) (美国专利 Nos. 5, 413,909、5,861,245) 以及基于核酸的序列扩增 (NABSA) (参见美国专利 Nos. 5,409,818、5,554,517 和 6,063,603, 每一个在此引为参考)。其它可以使用的扩增方法在美国专利 Nos. 5,242,794、5,494,810、4,988,617 和美国系列号 No. 09/854,317 中有描述, 每一个在此引为参考。

20
25
30

其它的样品制备方法和减少核酸样品复杂性的技术描述在 Dong 等, Genome Research 11, 1418 (2001)、美国专利 No. 6,361,947、6,391,592 和 6,107,023、以及美国系列号 09/916,135、09/920,491 (美国专利申请公布 20030096235)、09/910,292 (美国专利申请公布 20030082543) 和 10/013,598 中。

进行多核苷酸杂交分析的方法在本领域已经很成熟。杂交分析的步骤和条件依赖于具体的应用而变化, 并按照公知的通用结合方法来选择, 包括在下面文献中提到的方法: Maniatis 等 Molecular Cloning: A Laboratory Manual (分子克隆实验指南, 第 2 版) (2nd Ed. Cold Spring Harbor, N.Y, 1989); Berger 和 Kimmel Methods in Enzymology, Vol. 152, Guide to Molecular Cloning Techniques (分子克隆技术指南) (Academic Press, Inc., San Diego, CA, 1987); Young 和 Davism, P.N.A.S, 80: 1194 (1983)。实施重复和可控的杂交反应的方法和装置已经在美国专利 Nos. 5,871,928、5,874,219、6,045,996 和 6,386,749、6,391,623 中有描述, 在此将每一个引为参考。

本发明也考虑到在某些优选实施方案中配体间的杂交信号检测。参见美国专利 Nos. 5,143,854、5,578,832、5,631,734、5,834,758、5,936,324、5,981,956、6,025,601、6,141,096、6,185,030、6,201,639、6,218,803 和 6,225,625、美国系列号 10/389,194 和 PCT 申请 PCT/US99/06097 (公布号 WO99/47964), 在此为所有的目的以其全文引为参考。

信号检测和强度数据处理的方法和装置公开在例如美国专利 Nos. 5,143,854、5,547,839、5,578,832、5,631,734、5,800,992、5,834,758、5,856,092、5,902,723、5,936,324、5,981,956、6,025,601、6,090,555、6,141,096、6,185,030、6,201,639、6,218,803 和 6,225,625、美国系列号 10/389,194、60/493,495 和 PCT 申请 PCT/US99/06097 (公布号 WO

99/47964) 中, 在此为所有的目的以其全文引为参考。仪器和软件也可以从多种来源购买到, 包括 Affymetrix。

5 本发明的实施也可以使用常规的生物学方法、软件和系统。本发明的计算机软件产品一般包括计算机可读介质, 其中带有用于执行本发明方法的逻辑步骤的计算机可执行指令。适当的计算机可读介质包括软盘、CD-ROM/DVD/DVD-ROM、硬盘驱动器、闪存、ROM/RAM、磁带等。计算机可执行指令可以用适当的计算机语言或几种语言的组合来撰写。基本的计算生物学方法描述在例如 Setubal 和 Meidanis 等, 10 Introduction to Computational Biology Methods (计算生物学方法导论) (PWS Publishing Company, Boston, 1997); Salzberg、Searles、Kasif 主编, Computational Methods in Molecular Biology (分子生物学中的计算方法) (Elsevier, Amsterdam, 1998); Rashidi 和 Buehler, Bioinformatics Basics: Application in Biological Science and Medicine (生物信息学基础: 在生物科学和医学中的应用) (CRC Press, London, 2000) 以及 15 Ouelette 和 Bzevanis Bioinformatics: A Practical Guide for Analysis of Gene and Proteins (生物信息学: 基因和蛋白分析实用指南) (Wiley & Sons, Inc., 第 2 版, 2001) 中。参见美国专利 No. 6,420,108。

20 本发明也可以将各种计算机程序产品和软件应用于各种目的, 例如探针设计、数据管理、分析和仪器操作。参见美国专利 Nos. 5,593,839、5,795,716、5,733,729、5,974,164、6,066,454、6,090,555、6,185,561、6,188,783、6,223,127、6,229,911 和 6,308,170。

25 此外, 本发明在一些优选实施方案中包括了通过网络例如英特网提供遗传信息的方法, 如同在美国系列号 10/197,621、10/063,559 (美国公布号 20020183936)、10/065,856、10/065,868、10/328,818、10/328,872、10/423,403 和 60/482,389 中显示的那样。

30 b) 定义

“接头序列”或“接头”一般是长度为至少 5、10 或 15 个碱基的寡核苷酸，在优选情况下长度不超过 50 或 60 个碱基，但是它们可以更长，长达 100 或 200 个碱基。接头序列可以使用任何本领域专业人员所熟知的方法进行合成。出于本发明的目的，作为选择，它们可以含有引物结合位点、核酸内切酶识别位点、常见序列和启动子。接头可以是完全或基本上双链的或完全单链的。双链接头可以包含两个至少部分互补的寡核苷酸。接头在一条或两条链上可以是磷酸化或未磷酸化的。

接头如果含有基本上是双链的区域以及与限制性酶消化产生的单链区互补的短的单链区域，则与片段的连接将更有效。例如，当 DNA 用限制性酶 EcoR I 消化时，产生的双链片段在每端都带有突出的 5'-AATT-3'单链，带有突出的 5'-AATT-3'单链的接头将与片段通过突出区域之间的互补性而杂交。接头与片段的这种“粘性末端”杂交可以利于接头与片段的连接，但是平末端的连接也是可能的。使用 Klenow 片段的核酸外切酶活性可以将平末端转化为粘性末端。例如，当 DNA 用 PvuII 消化时，通过片段与 Klenow 在存在 dTTP 和 dCTP 的情况下保温，平末端可以被转化为带有两个碱基对的突出。通过在突出上进行填充或去除突出，突出也可以被转化为平末端。

20

连接的方法为本领域的专业人员所熟知，在例如 Sambrook 等(2001)的文献和 New England BioLabs 的目录中有描述，它们在此为所有的目的引为参考。方法包括使用 T4 DNA 连接酶，它催化在具有平末端或粘性末端的双链 DNA 或 RNA 中毗连的 5'磷酸和 3'羟基末端之间形成磷酸二酯键；Taq DNA 连接酶，它催化两个相邻的与互补的靶 DNA 杂交的寡核苷酸中毗连的 5'磷酸和 3'羟基末端之间形成磷酸二酯键；大肠杆菌 (E. Coli) DNA 连接酶，它催化在含有粘着末端的双链 DNA 中毗连的 5'磷酸和 3'羟基末端之间形成磷酸二酯键；以及 T4 RNA 连接酶，它通过形成 3' 5'磷酸二酯键催化 5'-磷酸基末端的核酸供体与 3'-羟基末端的核酸受体的连接，底物包括单链 RNA 和 DNA 以及二核

30

苷焦磷酸；或者本领域描述的其它任何方法。片段化的 DNA 在将接头连接到一个或两个末端之前，可以用一种或多种酶，例如核酸内切酶，进行处理，以产生与连接相容的末端而方便连接。

5 接头也可以掺有能够修饰接头序列性质的修饰的核苷酸。例如硫代磷酸酯基团可以被引入接头的一条链中。硫代磷酸酯基团是一种修饰的磷酸基团，其中有一个氧原子被硫原子所取代。在硫代磷酸酯化的寡聚物（通常称为“S-Oligo”）中，一些或所有的核苷酸间的磷酸基团被硫代磷酸酯基团取代。S-Oligo 的修饰的骨架对大多数核酸外切酶和核酸内切酶的作用具有抗性。硫代磷酸酯可以被引入接头链的所有残基之间，也可以引入到序列内的特定位置处。有用的选择是仅仅硫化寡聚物每端的最后几个残基。这样产生的寡聚物对核酸外切酶具有抗性，但是具有天然的 DNA 中心。

15 此处所用的术语“阵列”是指有意产生的分子的集合体，这些分子是通过合成或生物合成制备的。阵列中的分子彼此之间可以是相同的，也可以是不同的。阵列可以采取各种格式，例如可溶性分子的文库、束缚在树脂珠、硅芯片或其它固相支持物上的化合物的文库。

20 此处所用的术语“阵列板”是指具有多个阵列并形成区域或空间的物体，在所述阵列中每个微阵列被物理障碍分隔开来，以阻挡液体的流动，而所述区域或空间称为孔，能够包含与探针阵列接触的液体。

25 此处所用的术语“生物单体”是指生物聚合物的单个单位，它可以与相同的或其它的生物单体连接形成生物聚合物（例如，带有两个连接基团的单个氨基酸或核苷酸，连接基团中的一个或两个可以具有可去除的保护基团），或是指不是生物聚合物的一部分的单个单位。因此，例如在寡核苷酸生物聚合物中核苷酸是生物单体，而在蛋白或肽生物聚合物中氨基酸是生物单体；例如亲和素、生物素、抗体、抗体片段等也是生物单体。

30

此处所用的术语“生物聚合物”是指生物或化学部分的重复单位。代表性的生物聚合物包括但不限于核酸、寡核苷酸、氨基酸、蛋白、肽、激素、寡糖、脂类、糖脂、脂多糖、磷脂、上述生物聚合物的合成类似物，包括但不限于反向核苷酸、肽核酸、Meta-DNA，以及上述物质的组合。

此处所用的术语“组合合成策略”指一种组合性合成策略，是通过顺序加入试剂平行地合成多种聚合物序列的有序策略，它可以用反应物矩阵和转换矩阵来代表，其积是产物矩阵。反应物矩阵是待加入积木的 1 列 m 行的矩阵。转换矩阵是成列排列的 1 和 m 之间的二元数字的全部或其子集，优选情况下是有序的。“二元策略”是指这样的策略，其中至少两个相连的步骤照亮了基底上目的区域的一部分，通常是一半。在二元合成策略中，从有序的反应物组能够形成的所有可能的化合物都形成了。在最优选实施方案中，二元合成是指也把以前的加入步骤作为因素考虑在内的合成策略。例如，在一个策略中，掩蔽策略的转换矩阵将以前照亮的区域分为两半，照亮大约一半以前照亮的区域，保护剩下的一半（同时也保护大约一半以前保护的区域并照亮大约一半以前保护的区域）。应该认识到二元循环中可以散布有非二元循环，只有一部分基底可以经受二元方案。组合的“掩蔽”策略是合成策略，它使用光或其它空间选择性去保护或激活试剂来除去物质上的保护基团，以便添加其它的物质例如氨基酸。

此处所用的术语“互补性”是指核苷酸或核酸之间的杂交或碱基配对，例如在双链 DNA 分子的两条链之间或在寡核苷酸引物和被测序或扩增的单链核酸上的引物结合位点之间。互补的核苷酸一般是 A 和 T（或 A 和 U）或 C 和 G。两个单链 RNA 或 DNA 分子，当其中一条链的核苷酸通过最适化比对和比较，并进行适当的核苷酸插入或缺失，与另一条链的核苷酸有至少 80% 配对，通常至少约 90% 到 95%，更优选情况下从大约 98 到 100% 配对时，它们被说成是互补的。此外，

当 RNA 或 DNA 链在选择性杂交条件下与其互补链杂交时，互补性也存在。一般来说，当在一段至少 14-25 个核苷酸的跨度上存在至少大约 65%的互补性，优选至少大约 75%，更优选至少大约 90%互补性时，选择性杂交可以发生。参见 M. Kanehisa, *Nucleic Acids Res.* 12:203 (1984)，在此引为参考。

此处所用的术语“后生的”是指除了基因组的一级序列之外影响生物体的发育或功能的因素，它们可以不改变生物体的基因型而影响其表型。后生的因素包括对由 DNA 甲基化和染色质结构中可遗传但潜在可逆的变化所控制的基因表达进行修饰。已知甲基化模式与基因表达相关，并且一般而言，高度甲基化的序列表达很差。

此处所用的术语“基因组”是生物体的染色体中所有的遗传物质。来自特定生物体染色体中的遗传物质的 DNA 是基因组 DNA。基因组文库是从代表生物体完整基因组的一套随机产生的重叠 DNA 片段中制成的克隆的集合。

此处所用的术语“杂交”是指两个单链多核苷酸非共价地结合形成稳定的双链多核苷酸的过程；三链的杂交在理论上也是可能的。产生的（通常）双链多核苷酸是“杂交体”。杂交通常在严谨条件下进行，例如在盐浓度不超过大约 1M 和温度至少 25 °C 的情况下。例如，对于等位基因特异的探针杂交来说，5X SSPE (750 mM NaCl, 50 mM 磷酸钠, 5 mM EDTA, pH 7.4)和温度 25-30 °C 的条件是适当的，或者是 100 mM MES, 1 M [Na⁺], 20 mM EDTA, 0.01% Tween-20 和温度 30-50°C，优选为大约 45-50°C 的条件。杂交可以在存在试剂例如大约 0.1 mg/ml 鲑鱼精子 DNA、大约 0.5 mg/ml 乙酰化 BSA 的情况下进行。因为其它的因素包括互补链的碱基组成和长度、有机溶剂的存在和碱基误配的程度可以影响杂交的严谨性，参数的组合比单独任何一个参数的绝对值更为重要。适合于微阵列的杂交条件描述在 *Gene Expression Technical Manual*(基因表达技术手册)2004 和 *GeneChip Mapping Assay*

Manual（基因芯片作图分析手册）2004 中，可以从 Affymetrix.com 获得。

5 此处所用的术语“杂交探针”是指能够以碱基特异性方式与互补的核酸链结合的寡核苷酸。这样的探针包括肽核酸，如在 Nielsen 等, Science 254, 1497-1500 (1991)中所描述, LNAs, 如在 Koshkin 等, Tetrahedron 54:3607-3630, 1998 和美国专利 No. 6,268,490 中所描述, 以及其它的核酸类似物和核酸模拟物。

10 此处所用的术语“分离的核酸”是指目的核酸在存在的核酸中居主导地位（即在摩尔基础上它比组合物中任何其它单独核酸种类更丰富）。在优选情况下，分离的核酸占存在的所有大分子物质的至少大约 50、80 或 90%（在摩尔基础上）。在更优选情况下，目的核酸种类被纯化为基本上均一（通过常规的检测方法不能在组合物中检测到污
15 染物质）。

此处所用的术语“标记物”是指荧光标记物、光散射标记物或放射性标记物。其中，荧光标记物包括可以购买到的亚磷酰胺荧光素，例如 Fluoreprime (Pharmacia)、Fluoredite (Millipore)和 FAM (ABI)。参
20 见美国专利 6,287,778。

此处所用的术语“配体”是指被特定的受体识别的分子。与受体结合或反应的试剂被称为“配体”，该术语只有根据其对应的受体定义才有意义。除了表明该物质能够与受体结合或相互作用之外，术语
25 “配体”不暗含任何特定的分子大小或其它的结构或组成特点。此外，配体既可以用作与受体结合的天然配体，也可以作为功能类似物用作激动剂或拮抗剂。在本发明中研究的配体的例子包括但不限于细胞膜受体的激动剂或拮抗剂、毒素和毒液、病毒表位、激素（例如阿片剂、类固醇等）、激素受体、肽、酶、酶的底物、底物类似物、过渡态类似物、辅助因子、药物、蛋白和抗体。
30

此处所用的术语“混合种群”或有时称为“复杂种群”是指任何含有所需和无需核酸的样品。作为非限制的例子，核酸的复杂种群可以是总的基因组 DNA、总的基因组 RNA 或其组合。此外，核酸的复杂种群可以富集了给定的种群但是仍包括其它不需要的种群。例如，核酸的复杂种群可以是富集了所需的信使 RNA (mRNA) 序列但是仍包括一些不需要的核糖体 RNA 序列 (rRNA) 的样品。

此处所用的术语“mRNA”或有时称为“mRNA 转录本”包括但不限于 mRNA 转录本前体、转录本加工中间体、备译的成熟 mRNA 以及来自 mRNA 转录本的基因或核酸的转录本。转录本的加工包括剪接、编辑和降解。如本文所用，来自 mRNA 转录本的核酸是指为其合成，mRNA 转录本或其子序列最终用作模板的核酸。因此，从 mRNA 反转录得到的 cDNA、从 cDNA 转录得到的 RNA、从 cDNA 扩增得到的 DNA、从扩增的 DNA 转录得到的 RNA 等都是来自于 mRNA 转录本，对这些衍生产物的检测可以指示样品中原始转录本的存在和/或丰度。因此，mRNA 衍生样品包括但不限于基因的 mRNA 转录本、从 mRNA 反转录得到的 cDNA、从 cDNA 转录得到的 cRNA、从基因扩增得到的 DNA、从扩增的 DNA 转录得到的 RNA 等。

此处所用的术语“核酸文库”是指有意产生的核酸的集合，可以通过合成或生物合成制备，以各种不同的形式用于生物活性的筛选（例如可溶性分子的文库、束缚在珠、芯片或其它固相支持物上的寡聚物的文库）。此外，术语“阵列”的意思包括了那些可以通过将几乎任何长度的核酸（例如长度从 1 到大约 1000 个核苷酸单体）点样在基底上所制备的核酸文库。此处所用的术语“核酸”是指任何长度的核苷酸的聚合形式，核苷酸可以是核糖核苷酸、脱氧核糖核苷酸或肽核苷酸 (PNAs)，它们含有嘌呤和嘧啶碱基，或其它天然的、化学或生物化学修饰的、非天然的或衍生的核苷酸碱基。多核苷酸的骨架可以含有在 RNA 或 DNA 中通常所能发现的那些糖和磷酸基团，或者修饰的

或取代的糖或磷酸基团。多核苷酸可以含有修饰的核苷酸，例如甲基化的核苷酸和核苷酸类似物。核苷酸的序列可以被非核苷酸的成分所打断。因此术语核苷、核苷酸、脱氧核苷、脱氧核苷酸一般都包括例如此处所描述的那些类似物。这些类似物是哪些与天然存在的核苷或核苷酸具有一些相同结构特点分子，因此当它们掺入到核酸或寡核苷序列中时，可以在溶液中与天然存在的核酸序列杂交。一般来说，这些类似物是通过对碱基、核糖或磷酸二酯部分进行取代和/或修饰而从天然存在的核苷和核苷酸衍生而来的。这些改变可以被定制以便稳定或去稳定杂交体的形成，或增强与所需的互补核酸序列杂交的特异性。

此处所用的术语“核酸”可以包括任何嘧啶和嘌呤碱基的多聚物或寡聚物，碱基分别优选为胞嘧啶、胸腺嘧啶、尿嘧啶以及腺嘌呤和鸟嘌呤。参见 Albert L. Lehninger, PRINCIPLES OF BIOCHEMISTRY (生物化学原理)，第 793-800 页 (Worth Pub. 1982)。事实上，本发明考虑到了任何脱氧核糖核苷酸、核糖核苷酸或肽核酸成分，以及它们的任何化学变体，例如这些碱基的甲基化、羟甲基化或糖基化形式等。组合物中的多聚体或寡聚体可以是非均质的，也可以是均质的，可以从天然存在的来源分离，也可以人工或合成生产。此外，核酸可以是 DNA，也可以是 RNA 或它们的混合物，并且可以永久或暂时地以单链或双链的形式存在，包括同源双链体、异源双链体和杂交体状态。

此处所用的术语“寡核苷酸”或有时称为“多核苷酸”是指长度范围至少为 2、优选至少为 8、和更优选至少为 20 个核苷酸的核酸，或能够与多核苷酸特异性杂交的化合物。本发明的多核苷酸包括可以从天然来源分离、重组生产或人工合成和模拟的脱氧核糖核酸 (DNA) 或核糖核酸 (RNA) 的序列。本发明的多核苷酸的另一个例子可以是肽核酸 (PNA)。本发明还包括存在非常规的碱基配对的情况，例如 Hoogsteen 碱基配对，它已经在某些 tRNA 分子中被识别并推断存在于三螺旋中。在本申请中，“多核苷酸”和“寡核苷酸”可以互换使用。

此处所用的术语“引物”是指单链寡核苷酸，能够在适当的条件例如缓冲液和温度、在存在四种不同的核苷三磷酸以及聚合试剂例如DNA或RNA聚合酶或反转录酶的情况下，充当模板指导的DNA合成的起始点。在任何给定的情况下，引物的长度依赖于例如引物的使用目的，一般来说，范围从15到30个核苷酸。短的引物分子通常需要较低的温度以与模板形成足够稳定的杂交体复合物。引物不必须反映模板的精确序列，但是必须具有足够的互补性以与该模板杂交。引物位点是模板上引物杂交的区域。引物对是一套引物，包括与被扩增序列的5'端杂交的5'上游引物和与被扩增序列的3'端的互补序列杂交的3'端下游引物。

此处所用的术语“探针”是指被固定在表面上的分子，它能够被特定的靶识别。对于由具有10、12和更多个碱基的探针的所有可能的组合所构成的阵列的例子，参见美国专利 No. 6,582,908。本发明所研究的探针的例子包括但不限于细胞膜受体的激动剂或拮抗剂、毒素和毒液、病毒表位、激素（例如阿片样肽、类固醇等）、激素受体、肽、酶、酶的底物、辅助因子、药物、外源凝集素、糖、寡核苷酸、核酸、寡糖、蛋白和单克隆抗体。

此处所用的术语“受体”是指与给定的配体具有亲和性的分子。受体可以是天然存在的或人造的分子。同样，它们可以以其未改变的状态使用，也可以与其它物质结合使用。受体可以直接地或通过特定的结合物以共价或非共价形式连接到结合部分上。本发明中使用的受体的例子包括但不限于抗体、细胞膜受体、与特定的抗原决定簇（例如位于病毒、细胞或其它物质上）具有反应性的单克隆抗体和抗血清、药物、多核苷酸、核酸、肽、辅助因子、外源凝集素、糖、多糖、细胞、细胞膜和细胞器。受体在本技术领域中有时称为抗配体。在此所用的术语受体并不打算有不同意义。当两个大分子通过分子识别结合形成复合物时就形成了“配体受体对”。本发明中可以使用的其它受

体的例子包括但不限于那些在美国专利 No. 5,143,854 中所显示的分子，该专利在此以其全文引为参考。

5 此处所用的术语“固相支持物”、“支持物”和“基底”可以互换使用，是指一种或一组具有刚性或半刚性表面的材料。在许多实施方案中，固相支持物的至少一个表面应该是基本上平的，尽管在某些实施方案中可能希望用例如孔、升高的区域、针、刻蚀的沟渠等将合成区域进行物理上的分隔用于不同的化合物。按照其它的实施方案，固相支持物将采取珠子、树脂、凝胶、微球的形式或其它的几何构型。
10 示例性的基底参见美国专利 No. 5,744,305。

此处所用的术语“靶”是指对给定的探针具有亲和性的分子。靶可以是天然存在的或人工制造的分子。同样，它们可以以其未改变的状态使用，也可以与其它物质结合使用。靶可以直接地或通过特定的结合物以共价或非共价方式连接到结合部分上。本发明中使用的靶的例子包括但不限于抗体、细胞膜受体、与特定的抗原决定簇（例如位于病毒、细胞或其它物质上）具有反应性的单克隆抗体和抗血清、药物、寡核苷酸、核酸、肽、辅助因子、外源凝集素、糖、多糖、细胞、细胞膜和细胞器。靶在本技术领域中有时称为抗探针。在此所用的术语靶并不打算有不同意义。当两个大分子通过分子识别结合形成复合物时就形成了一个“探针靶对”。
15
20

此处所用的术语“晶片”是指具有可结合多个阵列的表面的基底。在优选实施方案中，阵列在底物的表面上合成，产生了多个在物理上分离的阵列。在晶片的一个优选实施方案中，阵列物理分离的距离为至少大约 0.1、0.25、0.5、1 或 1.5 毫米。位于晶片上的阵列可以是相同的，每一个可以是不同的，或可以是它们的组合。特别优选的晶片是大约 8” x 8”，使用照相平板印刷工艺制造。
25

30 个体不仅限于人类，它也可以包括其它的生物体，包括但不限于

哺乳动物、植物、真菌、细菌或来源于上述生物体的细胞。

甲基化分析

5 DNA 中胞嘧啶残基的甲基化在基因调控中发挥重要作用。DNA 甲基化为正常的胚胎发育所必需，甲基化的变化常常与疾病有关。基因组印记、X 染色体失活、染色质修饰和内源的反转录病毒的沉默都依赖于适当的甲基化模式的建立和维持。基因的表达可以被甲基化的模式所调控。甲基化异常是癌细胞的标志，肿瘤抑制基因的沉默被认为是许多人类癌症的因果基础。使用基于微阵列的方法进行甲基化作图可以被用来例如描述癌细胞的大体情况，发现 DNA 甲基化的模式可以用来例如诊断恶性肿瘤、预测治疗结果或监测疾病的进展。真核生物中的甲基化也可以用来抑制病毒和转座子的活性，参见 Jones 等，EMBO J. 17:6385-6393 (1998)。

15 在优选方面，本发明公开了利用与核酸探针阵列杂交分析基因组样品中胞嘧啶的甲基化状态的方法。在许多方面，方法包括了辨别甲基化和未甲基化胞嘧啶的步骤。这个步骤可以是例如化学或酶法的。例如，在某些方面，含有甲基化胞嘧啶的样品用亚硫酸氢盐处理，它可以选择性地修饰未甲基化的胞嘧啶。在其它方面，使用能够辨别甲基化和未甲基化 DNA 的酶，例如，DNA 可以平行地用同裂酶消化，其中一种酶是甲基化敏感的，而另一种是甲基化不敏感的。

25 在某些方面，本方面公开了扩增基因组 DNA 样品同时保留 DNA 中胞嘧啶甲基化状态的信息的方法。然后，可以分析被扩增样品中一个或多个胞嘧啶的甲基化状态。当与例如亚硫酸氢盐修饰这样的检测方法结合使用时，这种保留甲基化状态的扩增方法可能特别有用，因为亚硫酸氢盐修饰和其它相似的处理可能损坏 DNA，结果使得在修饰后扩增 DNA 的方法可能是无效的。在修饰或处理前扩增样品可以允许直接检测被修饰的样品，而不需要在修饰或处理后扩增样品。此外本方面还公开了检测胞嘧啶甲基化状态的方法。在优选实施方案中，

该方法能够同时分析大量的胞嘧啶，例如超过 1000、5000、10,000 或 100,000 个胞嘧啶。

5 许多分析基因组信息的方法在分析前使用扩增步骤，一般来说，这些方法会导致与起始材料的甲基化状态有关的信息的丧失，这是因为起始物质的序列被再生了，但是后生的信息例如甲基化的存在丢失了。在本发明的一个方面，公开了扩增核酸同时保留至少某些后生信息的方法。

10 扩增可以通过多种方式进行，例如聚合酶链反应（PCR）、滚环扩增（RCA）、多重置换扩增（MDA）、或其它全基因组扩增方法、代表性的基因组扩增或位点特异性扩增。方法可以包括例如用一个或多个限制性核酸内切酶或通过随机的片段化方法将核酸样品片段化，然后连接通用引发位点，这可以通过例如连接接头、或延伸具有通用 5' 15 末端和 3'端是随机或简并序列的引物来实现。

对许多扩增方法来说，引物与基因组模板杂交，然后被延伸产生 cDNA 链，它是基因组 DNA 的拷贝。第一条链 cDNA 的合成导致在体外合成的 cDNA 和体内合成的基因组 DNA 链之间形成杂交体，但 20 是一般来说，新合成的 cDNA 是没有甲基化的，因为在体内负责维持甲基化状态的系统在体外反应中不存在。在本发明的一个实施方案中，基因组 DNA 的扩增是在存在能够通过识别亲本模板链上的甲基化并且在新合成的子代链的相应位点处进行甲基化从而保留甲基化的活性的情况下来实现的。

25 在优选实施方案中，第一链 cDNA 在与作为模板产生 cDNA 的基因组 DNA 链形成复合物时被甲基化。然后，这个现在已经甲基化的第一链 cDNA（子代）和模板基因组 DNA 链（亲本）被分开，并且可以被用作模板合成其它的链，甲基化的子代作为模板和亲本用于随后新合成的子代链。各个随后产生的链都使用其模板（亲本）链作为 30

甲基化的指导来进行甲基化。

哺乳动物的甲基化模式是复杂的，在发育过程中会变化，参见 van Steensel 和 Henikoff *BioTechniques* 35: 346-357 (2003)。甲基化一般发生在胞嘧啶的 5' 位置，产生 5-甲基胞嘧啶 (5-meC)。大多数甲基胞嘧啶发现在 CpG 二核苷酸中。启动子区中的甲基化一般伴随着基因沉默，甲基化的丧失或与甲基化的 CpG 结合的蛋白的丧失可以导致人类疾病，例如免疫缺陷性颅面综合症和 Rett 综合症，Bestor (2000) *Hum. Mol. Genet.* 9:2395-2402。

已经进化出在多轮的细胞分裂甚至种系分裂过程中维持甲基化的方法，参见 Raykan 和 Whitelaw (2003) *Curr. Biol.* 13:R6。在动物中，甲基 CpG 由 DNA 甲基转移酶 1 (Dnmt 1) 维持。Dnmt 1 跟随在复制叉后面，并且使每个 CpG 中未甲基化的胞嘧啶残基甲基化，所述 CpG 是与甲基 CpG 配对的碱基，参见 Leonhardt 等 (1992) *Cell* 71:865-873。Dnmt 1 对半甲基化的位点具有特异性，因此通过连续的细胞分裂循环能够保存甲基化位点。DNA 甲基化是基因特异性的，在整个基因组都可以发生。

有两种不同的过程调节甲基化，在 DNA 复制后维持甲基化和在以前未甲基化的区域重新加上甲基基团。在哺乳动物中，一类甲基转移酶负责对未修饰的 DNA 进行甲基化，被命名为重新合成的酶。另一类酶在 DNA 复制过程中维持子代链的甲基化状态，被称为维持性 DNA 甲基转移酶。两类酶都催化甲基基团从 S-腺苷-L-甲硫氨酸 (AdoMet) 转移到 DNA 中的胞嘧啶碱基上。维持性甲基化负责为复制和细胞分裂后新合成的 DNA 中的胞嘧啶加上甲基基团。亲本 DNA 中的甲基化位点作为模板用来校正在复制后马上识别半甲基化的子代链的维持性甲基转移酶活性所产生的甲基化，参见 Riggs, A.D. (1989) *Cell Biophys.* 15:1-13。相反，重新合成的甲基化活性产生新的甲基化模式。DNA 甲基转移酶 (DNMTs) 在许多生物体包括哺乳动物、植

物和细菌中发现 (Bestor 等 *Curr. Opin. Cell Biol.*, 6:380-389 (1994))。在哺乳动物中, 已经鉴定了三种活性 DNA 甲基转移酶, Dnmt 1、Dnmt 3a 和 Dnmt 3b。Dnmt 2 最近也已经被鉴定。Dnmt 3a 和 Dnmt 3b 的功能主要作为重新合成的甲基转移酶, 而 Dnmt 1 负责维持性甲基化。已经显示出 Dnmt 1 在体外对半甲基化的 CpG 位点的甲基化具有高度偏好性, 参见 Pradhan 等 (1997) *Nucl. Acids Res.* 25:4666-4673 和 (1999) *J. Biol. Chem.* 274:33002-33010。

Dnmt 1、Dnmt 3a 和 Dnmt 3b 在存在底物 DNA 和 AdoMet 辅助因子的情况下体外具有活性。Dnmt 1 是哺乳动物细胞中最丰富的 DNA 甲基转移酶, 已经从天然来源纯化。来自几种生物体包括小鼠和人的重组 Dnmt 1 也已经获得。该酶在体内对半甲基化的底物的活性比对未甲基化的底物的活性高 7 到 20 倍。已经显示出 Dnmt 1 具有几种天然存在的同工型, 包括拼接变体形式 Dnmt 1b 和卵母细胞特异形式 Dnmt 1o。对 Dnmt 1 蛋白的突变分析已经鉴定出羧基末端的催化结构域和含有核定位信号、复制叉靶向肽和锌结合区域的调控结构域。在该蛋白中也已经鉴定出了许多特异蛋白相互作用区域。综述可参见 Pradhan 和 Esteve, *Clinical Immunology* 109:6-16 (2003)。

人类 DNA (胞嘧啶-5) 甲基转移酶 (Dnmt 1) 可以从 New England Biolabs 购买到。该蛋白使用杆状病毒表达系统从人类的 cDNA 表达。随酶同时提供 10x 反应缓冲液、BSA 和 S-腺苷甲硫氨酸 (SAM)。推荐的反应条件是 1x Dnmt1 反应缓冲液 (1X 是 59 mM Tris-HCl, 1 mM DTT, 1 mM EDTA, 5% 甘油, pH7.8, 25°C), 100 µg/ml BSA 和 160 µM SAM, 在 37°C 进行反应。1 单位的酶定义为在 37°C 下, 30 分钟内, 在 25 µl 总反应体积中将 1 pmol 甲基基团催化转移到 poly dI.dC 底物上所需的酶量。针对该酶的抗体也是可用的。

对 Dnmt3a 的最新研究显示该酶的催化活性在体外受 DMSO 的刺激, 参见 Yokochi 和 Robertson, *Bioorganic Chem.* 32:234-243 (2004)。

该研究表明 DMSO 的刺激效应依赖于 DMSO 与酶反应底物 DNA 和 AdoMet 之间的相互作用，而不是酶本身。在某些方面，DMSO 可以包括在 Dnmt1 反应中。

5 在本发明的一个方面，获得了含有甲基化胞嘧啶的基因组 DNA 样品，并且将一个或多个引物与 DNA 杂交。引物可以是例如一个或多个位点特异引物或一个集合体或随机或部分简并的引物。引物用 DNA 聚合酶延伸形成双链 DNA 杂交体。杂交体含有一条模板 DNA 链和一条新合成的 cDNA 链。基因组模板链可以含有甲基化位点，而
10 新合成的链可以是未甲基化的。用 DNA 甲基转移酶活性处理杂交体，该活性识别模板链上的甲基化位点，并在模板链中的甲基化位点处对新合成的链进行甲基化。甲基转移酶和 DNA 聚合酶活性可以同时或不同时间作用。在新合成的链的甲基化完成后，这个新合成的甲基化的链可以用作第二轮扩增和甲基化中的模板链。这个步骤可以重复多
15 次，以产生起始许多拷贝的基因组模板，其中包括甲基化位点。在优选实施方案中，模板甲基化的拷贝在下一轮扩增前完成。在某些实施方案中，在下一轮扩增之前，反应可以用切割半甲基化位点的酶进行处理。

20 在一个实施方案中，PCR 用于扩增，DNA 甲基转移酶活性是热稳定性的，并且在连续使双链 DNA 加热变性的循环过程中将保持活性。在另一个实施方案中，扩增使用链置换酶例如 phi29 进行（参见例如美国专利 Nos. 6,617,137 和 6,323,009），DNA 甲基转移酶可以是也可以不是热稳定的。当使用链置换聚合酶时，在优选情况下甲基转
25 移酶活性与聚合酶结合起作用，以便在子代链被随后的子代链取代之前甲基化位点被拷贝到子代链中。在一个方面，扩增是等温的。

 扩增后，甲基化状态保留下来的扩增材料可以被分析，以确定一个或多个胞嘧啶的甲基化状态。任何可用来确定甲基化状态的方法都
30 可以使用。对于检测甲基化状态的方法，参见例如美国专利申请 No.

10/841,027、美国专利 Nos. 6,214,556、5,786,146、6,017,704、6,265,171、6,200,756、6,251,594、5,912,147、6,331,393、6,605,432 和 6,300,071，以及美国专利申请公布 20030148327、20030148326、20030143606 和 20030082609。对于一些甲基化检测方法的综述，参见 Oakeley, E.J.,
5 Pharmacology & Therapeutics 84:389-400 (1999)。可用的方法包括：反相 HPLC、薄层层析、掺入标记的甲基基团的 SssI 甲基转移酶、氯乙醛反应、不同敏感性的限制性酶、胍或高锰酸盐处理（5-甲基胞嘧啶被高锰酸盐处理切割但不被胍处理切割）、亚硫酸氢钠、组合的亚硫酸氢盐-限制性分析以及甲基化敏感的单核苷酸引物延伸。这些方法分别描述在 Oakeley (1999)中。对于这些方法中的许多来说，扩增过程中甲基化状态的保留利于下游的分析方法，这些分析方法依赖于对甲基化敏感的处理。许多检测甲基化的方法在甲基化敏感性修饰步骤之后使用扩增步骤，在某些方面，通过在处理前扩增的同时保留甲基化状态，对修饰后样品进行扩增的需要就被消除了。这可以提高扩增和
10 检测的效率。由于在修饰步骤中对 DNA 的损坏，修饰后的扩增可能是无效的。

在本发明的一个方面，胞嘧啶的甲基化状态通过限制性消化加以分析，使用两种识别同样识别位点但对甲基化具有不同敏感性的限制性酶来进行。这样的酶对的例子是 HpaII 和 MspI。HpaII 和 MspI 是同裂酶，在识别位点 CCGG 处切割（参见 New England Biolab 目录，在此以其全文引为参考）。HpaII 的切割被中央 C 的甲基化所阻断，而 MspI 的切割不依赖于中央 C 的甲基化。在本发明的一个方面，具有保留的甲基化信息的被扩增样品用 HpaII 和 MspI 进行消化，然后可以分析
20 切割产物以确定甲基化状态。如果目的位点是甲基化的，则它将不被 HpaII 切割但是可以被 MspI 切割。

在示例性实施方案中，样品首先用限制性酶例如 XbaI、HindIII 或 BglII 片段化，然后将接头连接到片段上。在优选的实施方案中，
30 第一个限制性酶的识别位点不含有 CpG。有多种酶可以用于第一次片

段化。在连接了接头后，样品被分成组分或等份，然后进行不同的处理，可以通过例如平行地与探针阵列杂交来进行分析。在一个方面，一个组分用 Hpa II 片段化，而第二个组分用 MspI 片段化。也可以包括第三个组分，它既不用 Hpa II 也不用 MspI 片段化。Hpa II 消化的组分可以与 MspI 消化的组分或未消化的组分或二者同时进行比较。此处所用的 Hpa II 和 MspI 是作为示例性的酶，但是任何一对能够识别同样的限制性位点并对甲基化具有不同敏感性的酶都可以使用。

具有 CCGG 识别位点的片段，如果 CpG 是未甲基化的，它在 Hpa II 和 MspI 组分中都将切割，但是如果 CpG 是甲基化的，它将在 MspI 组分但不在 Hpa II 组分中被切割。切割后样品用一个或多个与接头互补的引物来扩增。在这个扩增步骤中，并不必需维持关于甲基化状态的信息。用 MspI 或 HpaII 消化过的连接了接头的片段在 PCR 反应过程中将不会扩增，这是因为位点没有甲基化。如果片段已经被 MspI 或 HpaII 切割了，该片段将不能在 PCR 反应中被扩增，因为产生的片段只在一端有接头序列和引发位点。仅仅具有甲基化的 CCGG 位点的片段在 HpaII 反应中将被切割，它们在 PCR 过程中被扩增。在 HpaII 反应中出现但是不在 MspI 反应中出现的片段，可以通过将每一个 PCR 扩增反应的产物与探针阵列杂交来进行鉴定。探针只在 HpaII 反应中检测到高于背景的杂交，是甲基化片段的指示。

在另一方面，可以使用一种只切割甲基化的 DNA 但不切割非甲基化的 DNA 的酶。如上述产生的连接了接头的片段，用只切割甲基化的 DNA 的甲基化依赖性酶进行消化，未消化的连接了接头的片段用接头序列的引物进行扩增。被甲基化依赖性酶消化的连接了接头的片段不能被扩增，并在随后的检测步骤中不能被检测到。扩增的产物与阵列杂交，获得杂交模式，并与从平行处理的、但没有用甲基依赖性酶消化的样品中获得的杂交模式进行比较。两种模式之间的区别指示样品中的片段被甲基化。甲基依赖性酶包括例如 McrBC。

扩增产物可以用任何本领域所熟知的方法进行检测，例如通过与探针阵列杂交。阵列可以含有针对选定区域的探针，也可以含有平铺的代表完整的染色体、完整的基因组、基因组的一个或多个大区域的探针。阵列也可以被设计有针对含有预计或已知的甲基化位点的区域的探针。示例性阵列包括在美国专利申请 Nos. 10/264,945、09/916,135 和 60/417,190 中所公开的阵列，分别在此引为参考。

在一个方面，计算机被用来模拟第一次限制性酶消化的产物，以预测片段的大小和序列。然后计算机可用来鉴定那些还含有一个或多个甲基化敏感的限制性酶的识别位点的片段。计算机还可以用来鉴定在 PCR 条件下易于扩增的片段。在许多实施方案中，PCR 条件优选扩增有限大小范围的片段，例如 200 到 2,000 个碱基对或 200 到 1,000 个碱基对。在预期的大小范围内同时也含有甲基化敏感性酶位点的片段被鉴定出来，并且可以对阵列进行设计，使之含有与多个被鉴定的片段互补的探针。

在示例性方面，第一个限制性酶是 XbaI，计算机被用来模拟人类基因组被 XbaI 的消化，以鉴定 200 到 1,000 个碱基对之间的 XbaI 片段。计算机被用来分析被鉴定的片段的序列，从而鉴定出在片段内至少具有一个 CCGG 位点的片段的子集。阵列中的探针可以被设计用来质询那些满足这两个标准的片段。探针可以针对片段的任何区域，在优选情况下，每个片段用多个不同序列探针进行质询，这些探针与片段的不同、但任选重叠的区域完全互补。在一个方面，来自 HpaII 消化的片段的杂交模式与 MspI 消化反应相比。含有 CCGG 位点但是在 HpaII 消化后可以被扩增的片段指示该片段中 CCGG 位点的甲基化。该片段在 MspI 消化反应中不应该被检测到。在另一方面，HpaII 杂交模式被用来与没有进行第二次消化处理的样品的杂交模式进行比较。没有第二次消化时，甲基化和未甲基化的片段都能被扩增，并通过与阵列杂交而被检测，这可以作为扩增的阳性对照。只带有未甲基化的 CCGG 序列的片段将在 HpaII 反应中被消化，所以在随后的扩增

步骤中将不能被扩增，针对那些片段的探针应该不产生高于背景的信号。没有用第二个消化步骤处理的样品也可以被用来估计甲基化的水平。如果片段仅仅是部分被甲基化，并且存在甲基化和未甲基化片段的混合物，信号的强度可以与未处理样品的信号强度进行比较，以估计甲基化的片段量。

在一个方面，基于第一次片段化反应的预计产物和显示杂交的探针的序列，计算机系统被用来对基因组中的甲基化片段进行定位和作图。此外，计算机还可以被用来鉴定被鉴定片段中的 CCGG 位点。

在本发明的一个方面，探针阵列包括与含有 CpG 岛的基因组区域互补的探针。探针可以被设计成与某区域互补，该区域与 CpG 岛位于同样的限制性片段中，但是可以与不含 CpG 二核苷酸的区域互补。

在一个实施方案中，扩增的甲基化产物中的甲基化状态可以用亚硫酸氢钠处理方法进行分析。在亚硫酸氢钠修饰过程中未甲基化的胞嘧啶通过三步过程转化为尿嘧啶。这些步骤是将胞嘧啶转化为胞嘧啶磺酸盐的磺化反应，将胞嘧啶磺酸盐转化为尿嘧啶磺酸盐的脱氨基作用，以及将尿嘧啶磺酸盐转化为尿嘧啶的碱性脱磺酸作用。转化不发生在甲基化的胞嘧啶上。参见 Clark 等 *Nucleic Acids Res.*, 22(15):2990-7 (1994)。如果胞嘧啶被甲基化了，它将保留为胞嘧啶。如果胞嘧啶未甲基化，它将被转化为尿嘧啶。当修饰的链通过例如位点特异的引物、随机或简并引物或针对接头的引物的延伸进行拷贝时，如果 C 被甲基化，则在质询位置将掺入 G（与质询的 C 相对），而如果 C 是未甲基化的，则在质询位点将会掺入 A。当双链延伸产物被扩增时，那些被转化为尿嘧啶并导致在延伸的引物中掺入腺嘌呤的胞嘧啶，将在扩增过程中被胸腺嘧啶所代替。那些没有被修饰并导致鸟嘌呤掺入的胞嘧啶，仍会保留成胞嘧啶。

用于 DNA 亚硫酸氢盐修饰的试剂盒可以商购自，例如 Human

Genetic Signatures 的 Methyleasy 和 Chemicon 的 CpGenome 修饰试剂盒。也可参见 WO04096825A1, 它描述了亚硫酸氢盐修饰方法, 以及 Olek 等 Nuc. Acids Res. 24:5064-6 (1994), 它公开了在包被于琼脂糖珠中的物质上进行亚硫酸氢盐处理和随后扩增的方法。在一个方面, 诸如二亚乙基三胺的催化剂可以与亚硫酸氢盐处理结合使用, 参见 Komiyama 和 Oshima, Tetrahedron Letters 35:8185-8188 (1994)。二亚乙基三胺已经被显示在 pH 5 时能有效催化亚硫酸氢盐离子诱导的从 2'-脱氧胞苷到 2'-脱氧尿苷的脱氨基作用。其它的催化剂包括氨、亚乙基二胺、3',3'-二氨基二丙胺和精胺。在某些方面, 脱氨基作用的进行采用 3-5M 亚硫酸氢钠溶液, 在大约 50°C 保温 12-16 小时。更快的步骤也已经报道, 使用 pH 5.4 的 9-10M 亚硫酸氢盐, 在 90°C 保温大约 10 分钟, 参见 Hayatsu 等, Proc. Jpn. Acad. Ser. B 80:189-194 (2004)。

亚硫酸氢盐处理允许用多种方法检测胞嘧啶的甲基化状态。例如, 任何可以被用来检测 SNP 的方法都可以使用, 例如参见 Syvanen, Nature Rev. Gen. 2:930-942 (2001)。可以使用例如单碱基延伸 (SBE) 的方法, 或与等位基因特异性杂交方法相似的序列特异探针的杂交。

在一个方面, DNA 样品用一种或多种限制性酶片段化, 在用亚硫酸氢盐处理前与一个或多个接头序列连接。然后, 亚硫酸氢盐处理的样品使用与接头互补的引物, 可以通过 PCR 进行扩增。扩增的条件可以被选择以优选扩增选定大小的片段, 例如 200 到 2000bp 的片段, 从而减少样品的复杂性。

在一个方面, 接头在亚硫酸氢盐处理后与 DNA 连接。在优选的方面, T4 RNA 连接酶被用于接头的连接。对于 PCR 引物的添加, 引物可以连接到 5'末端 (优选情况下, 引物是末端保护的), 然后将片段的 3'端脱磷酸化, 并且将 5'磷酸化的引物连接到 3'末端。亚硫酸氢盐处理可以降解 DNA, 所以在亚硫酸氢盐处理前连接的接头可能通过处理被破坏或切割掉, 使得片段不能被扩增。亚硫酸氢盐处理也可以

使 DNA 成单链，因此为了连接接头序列可能必需使用 T4 RNA 连接酶。亚硫酸氢盐处理也会在 DNA 中引入误配。将未甲基化的胞嘧啶转化为尿嘧啶，产生了 G:U 碱基对代替 G:C 碱基对，从而导致双链 DNA 的变性。

5

除了未甲基化的胞嘧啶的脱氨基作用外，亚硫酸氢盐处理可导致 DNA 的损坏，使得 DNA 片段化。在某些方面，亚硫酸氢盐处理在 pH 大约 5 时需要较长的保温（大约 4-16 小时）。在这个步骤中，胞嘧啶被磺化，然后发生脱氨。这个步骤也可能有不需要的副反应，导致 DNA 的部分脱嘌呤。在脱氨后，硫酸基团通过碱处理除去。碱处理可能导致链在发生脱嘌呤的位点断裂。产生的片段连接到接头上，但是可能需要对片段进行化学或酶法的处理以产生适合连接的末端。在某些方面，脱嘌呤位点的碱水解可以产生适合于连接接头的 5'磷酸化末端和因为缺少 3'羟基而不适合于连接的 3'末端。3'末端可以被处理以除去阻碍连接的修饰。在一个方面，片段在连接接头前用 AP 核酸内切酶处理。在另一方面，接头可以通过两个反应连接到片段上，在第一个反应中将接头连接到可用于连接的末端，然后用例如激酶处理反应液以从 3'端除去磷酸，再用于第二个连接反应。在脱嘌呤和链断裂后形成的末端可以根据切割的机制而变化。在某些方面，产生 3'磷酸化的核糖，但是在某些方面，产生具有不同末端的混合物，包括具有末端核糖的片段。在优选的方面，对 3'末端进行化学或酶法的加工以产生适合于接头连接的末端。

10

15

20

25

30

在另一方面，亚硫酸氢盐处理的 DNA 的扩增可以用随机的引物引发，例如随机六聚体。其它的扩增方法也可以使用，例如恒温链置换扩增、滚环扩增（Lizardi 等, Nat. Genet. 19:225-232, 1998）、多重置换扩增（Dean 等, Proc. Natl. Acad. Sci. 99:5261-5266, 2002）以及那些在美国专利公布 Nos. 20040209298 和 20040209299 中所描述的方法。亚硫酸氢盐处理会破坏 DNA，损坏的 DNA 很难被扩增。能使被降解的样品，例如那些从福尔马林固定的石蜡包埋的（FFPE）样品中获得

5 的样品扩增的方法，可以被用来扩增亚硫酸氢盐处理的 DNA。对于被降解的样品来说，优选的扩增方法包括那些在美国专利公布 Nos. 20040209298 和 20040209299、Wang 等, Gen. Res. 14:2357-2366, 2004 和 Wang 等, Nuc. Acids Res. 32:e76, 2004 中所公开的方法。在优选的方面，用于扩增的引物对亚硫酸氢盐转化的 DNA 有偏好性，因为它们具有减少数量的 G/C 碱基对。在扩增的第一轮中，未甲基化的胞嘧啶一般已经被转化为尿嘧啶了，所以引物可以是有偏好的，含有较少的或没有鸟嘌呤。

10 在一个方面，亚硫酸氢盐处理的 DNA 与针对 5-meC 的抗体或与 5-meC 结合蛋白和针对该蛋白的抗体一起保温，然后分离结合了抗体的复合物。从分离的复合物得到的 DNA 如上所述通过连接接头和 PCR 扩增可加以扩增。

15 在另一方面，活化诱导的胞苷脱氨酶（AID）被用来代替亚硫酸氢盐处理。AID 可以将未甲基化的胞嘧啶脱氨，而甲基化的 CpG 基元免受 AID 介导的脱氨作用，参见 Larijani 等, Mol Immunol. 42(5):599-604 (2005)。AID 处理的 DNA 可以采用与亚硫酸氢盐处理的 DNA 的同样方法来加以分析。AID 分析的优点在于，它可以在较短的时间内进行，大约 30 分钟，而与此相比，典型的亚硫酸氢盐处理需要 12 个小时以上，此外，它比复杂的亚硫酸氢盐处理所用的步骤少，使用的有毒化学物质也较少。在某些方面，DNA 可以用 AID 处理和亚硫酸氢盐处理相结合的方法来处理。这两种方法的组合可以被用来改进 AID 处理的效率，缩短亚硫酸氢盐处理的时间，并且减少 DNA 的降解。

25 在一个方面，特定的胞嘧啶的甲基化水平可以被定量。杂交模式可以被分析以测量甲基化的水平，杂交的强度与甲基化程度相关。例如，如果特定的胞嘧啶在样品 DNA 中有 80%是甲基化的，则在亚硫酸氢盐处理后 C “等位基因”的标准化强度应该是 T “等位基因”的标准化强度的约 4 倍。使用亚硫酸氢盐处理对特定胞嘧啶甲基化水平

30

进行定量的方法已经公开了,例如在 Thomassin 等, Nuc. Acids Res. 32:e 168 (2004)中。

在优选的方面,产物通过与阵列的杂交进行分析。在一个示例性
5 实施方案中,阵列被设计来检测亚硫酸氢盐处理的产物,其原理与可
商购的 Affymetrix 10K 作图阵列相同。10K 阵列对超过 11,000 个不同
的人类 SNPs 中的每一个具有探针组。每个探针组的第一多个探针与
SNP 的第一种等位基因完全互补,而第二多个探针与 SNP 的第二种等
10 位基因完全互补。如果存在第一种等位基因,信号就被第一多个探针
检测,而如果存在第二种等位基因,信号就被第二多个探针检测。杂
合子则将被两者都检测到。探针组可以包括对照探针,例如错配探针,
质询位置相对于探针的中央位置有偏移的探针也可以包括,例如 SNP
位置可以位于中央位置,也可以偏离探针的中央向 5'或 3'移动 1 个或
15 多个位置。对于有疑问的甲基化位点也可以设计类似的探针组,处理
该位置,就好象它是带有 C/G 或 T/A 等位基因的 SNP。两条链都可以
被分析。示例的探针和阵列在美国专利申请 No. 10/681,773 和美国专
利 Nos. 5,733,729、6,300,063、6,586,186 和 6,361,947 中有描述。亚硫
酸氢盐处理可以修饰片段中任何未甲基化的 C,包括引物结合位点中
20 的 C 和质询位置周围区域中的 C。在优选实施方案中,接头的设计应
将此考虑在内,例如接头可以设计成在引物结合位点不含有 C,引物
也可以用对亚硫酸氢盐修饰有抗性的修饰碱基进行合成以致引物结合
位点的序列不会被处理所改变,例如可以将 C 甲基化,或者引物可以
被设计成假定接头中的 C 会变成 U。

25 可以从序列中检测新的 SNPs 的重新测序阵列也可以用来检测亚
硫酸氢盐处理的产物。重新测序阵列和重新测序方法在例如 Cutler 等
Genome Res. 2001 Nov;11(11):1913-25 和美国专利申请 No.
20030124539 中有描述,在此以其全文引为参考。一般来说,重新测
序阵列检测参比序列中所有可能的单核苷酸变异。与参比序列完全互
30 补的探针被包括在内,并且逐个地质询序列中的多个位点以寻找参比

序列中的变异。对于每个可能的变异来说，与变体序列完全互补的探针都被包括在内。阵列可以被平铺开以检测一个或多个参比序列中所有可能的单核苷酸变异。被阵列质询的参比序列可以是例如一个或多个完整的染色体、一个或多个完整的基因组、一个或多个线粒体基因组、或者是从一个或多个基因组内选择的目的区域。在一个实施方案中，重新测序阵列被平铺有已知或怀疑被甲基化的区域。在某些实施方案中，CpG 位点可能靠在一起，因此阵列的探针可以与重叠的 CpG 位点互补。例如，如果探针是 25 个聚体，第 13 位的质询位置与第一个胞嘧啶位置互补，在第一个胞嘧啶上游 12 个碱基对或下游 12 个碱基对的范围内可能存在第二个 CpG。第二个胞嘧啶可以是甲基化的也可以是未甲基化的。探针可以被设计来检测两种可能性，即都是甲基化的（都是 C）、都是未甲基化的（都是 T）、一个是甲基化的（C）另一个是未甲基化的（T）。可以设计与每一种可能的结果完全互补的探针。

在本发明的另一方面，扩增的甲基化的靶相对于未甲基化的靶得到了富集。在一个示例的实施方案中，使用限制性酶将怀疑含有 m5C 的核酸样品片段化，并将接头与片段连接。针对 m5C 的抗体被用来分离含有 m5C 的连接了接头的片段。此外，核酸也可以与能够特异性结合 m5C 的蛋白一起保温，然后针对那些蛋白的抗体可以被用来分离甲基化的片段。针对 5-meC 的抗体是可得到的，例如来自 Abcam (Cambridge, UK) 的 ab 1884。使用与接头互补的引物通过 PCR 将分离的片段进行扩增，扩增的片段可以与探针阵列杂交。在一种优选方面，阵列的探针与基因组的一个或多个区域互补。阵列中示出高于背景的杂交的区域指明了基因组中被甲基化的区域。在优选实施方案中，阵列含有针对基因组的富含 CpG 的区域、基因内区域或已知或推测是调控区的区域的探针。在另一个实施方案中，被免疫沉淀的片段用亚硫酸氢盐处理，以便可以鉴定甲基化胞嘧啶的精确位置。样品可以通过与上述的序列特异性探针的阵列杂交来进行分析。

在本发明的一个方面，甲基结合蛋白例如 MeCP2 和 SAP18/30（结合 Sin3 的多肽 18/30）与基因组 DNA 样品混合，用于富集甲基化的序列。针对甲基 CpG 结合结构域蛋白（MBDs）例如 MBD2 和 MBD3 的抗体可以被用来分离含有甲基化的 DNA。抗 5-meC 结合蛋白的抗体是可得到的，例如针对 MeCP2 (IMG-297) 的抗体来自 Imgenex Corp. (San Diego, CA)。在另一方面，识别 5-meC 的抗体可以被用来富集甲基化的序列。在优选情况下在抗体结合前将 DNA 变性。

在本发明的另一方面，甲基化被用作以相对可再现的方式使基因组分成亚类的手段，以便在进一步分析前减少样品的复杂性。基因组的某些区域稳定地被甲基化，同时其他区域稳定地被未甲基化。区分甲基化和未甲基化 DNA 的机制可以被用来获得富集了甲基化 DNA 或未甲基化 DNA 的样品的组分。通过这种方法样品的复杂性得以减少。如果扩增采用的是保留甲基化信息的方法，则分离可以在扩增之前进行。通常都希望在杂交之前减少含有复杂的核酸混合物的样品的复杂度，以便改善检测的灵敏度，并使背景最小化。

在本发明的一个方面，公开了在分离甲基化和未甲基化组分之后分析核酸样品的方法。被分析的组分可以是甲基化的组分也可以是未甲基化的组分，或者可以进行甲基化的和未甲基化的组分的比较。在许多实施方案中，未甲基化的组分通过例如从未甲基化的 DNA 中分离出甲基化的 DNA，或者通过优先扩增未甲基化的 DNA 而富集。对富集有亚类起始核酸的组分的分离可以用作减少样品复杂性的方法，或者用作测量甲基化组分和未甲基化组分之间区别的方法。在一个实施方案中，该方法尤其用来分析样品，以鉴定出在未甲基化组分中存在的基因组区域和甲基化组分中存在的基因组区域。在许多实施方案中，分离甲基化和未甲基化核酸的方法与用探针阵列分析核酸的方法相结合。

在一个方面，通过用酶例如 Mse I 消化 DNA 样品，接着大小选

择来富集 CpG 岛。Mse I 将基因组 DNA 切成小片段，但是在 CpG 岛中切割很罕见。较大的富含 CpG 岛的片段可以通过任何现有的大小分离方法从较小的片段中分离出来。

5 复杂性减少了的样品代表了更复杂的样品，例如基因组，它们可以用于多种分析方法，包括那些含有杂交步骤的方法。复杂性减少了的样品可以被用于例如测序、基因分型、拷贝数的定量评估、LOH 分析和 CGH 分析。在许多实施方案中，分析是通过将复杂性减少了的样品与探针阵列杂交来进行的。用于表达分析、重新测序和基因分型的阵列可以得自，例如 Affymetrix, Inc., Santa Clara, CA。

从未甲基化的核酸分离甲基化核酸的方法已经有描述，参见例如美国专利公布 Nos. 20010046669、20030157546 和 20030180775，在此分别以其全文引为参考。

15

在植物和哺乳动物基因组中重复序列通常以高的拷贝数存在，具有高水平的胞嘧啶和低的转录活性(参见，例如 Martienssen, R. A. (1998) Trends Genet. 14:263; Kass, S. U.等 (1997) Trends Genet. 13:335; SanMiguel, P.等 (1996) Science 274:765; Timmermans, M. C.,等 (1996) Genetics 143:1771; Martienssen, R. A.和 E. J. Richards, (1995) Curr. Opin. Genet. Dev. 5:234-242; Bennetzen, J. L.,等 (1994) Genome 37:565; White, L. F.,等 (1994) Proc. Natl. Acad. Sci. U.S.A. 91:11792; Moore, G.等 Genomics 15:472)。高拷贝的 DNA 序列通常是甲基化的，并且一般不出现在表达的基因的区域。从文库或从核酸样品中消除或减少这样的高拷贝甲基化 DNA 的方法可以被用来富集目的靶序列，产生复杂性被降低的样品，利于进一步分析。通常未甲基化的区域是含有基因的区域，是最希望被分析的区域。

25

通过将核酸文库，例如可能是部分文库的基因组文库在甲基化限制性宿主，例如大肠杆菌菌株 JM101、JM107 和 JM109 中繁殖，可以

30

从核酸样品中富集出未甲基化的序列。这种方法、甲基化过滤最近才被用来对玉米的基因组进行测序，参见 Palmer 等 Science 302:2115-2117 (2003)。该方法阻止了带有甲基化插入物的克隆的繁殖，导致与对照文库相比基因富集了 5 到 7 倍。

5

在另一个实施方案中，核酸样品用甲基化敏感的酶，例如只切割未甲基化的 DNA 或只切割甲基化的 DNA 的酶，或切割甲基化或未甲基化 DNA 的甲基化不敏感的酶进行消化。不同消化的样品被扩增，扩增的片段可以被标记，然后使用微阵列进行检测。样品可以平行地
10 用甲基化敏感和甲基化不敏感的酶消化，并在每种处理后分析以确定存在哪些序列。在第一种样品中出现但是在第二种样品中不出现的序列表明该序列是甲基化的。

在一个示例的实施方案中，核酸样品从某来源，例如从个体获得，
15 核酸可以通过例如用一个或多个限制性酶消化而片段化，然后接头序列连接到片段上产生连接了接头的片段。连接了接头的片段可以用切割甲基化 DNA 但是不切割未甲基化 DNA 的酶，例如 McrBC 来进行消化。接着，样品可以用与接头序列杂交的引物进行扩增。那些已经被甲基特异性酶切开的甲基化片段不会被扩增，因为它们只在一端有
20 接头，从而使未甲基化的 DNA 得到选择性的扩增。

扩增的产物可以通过例如与微阵列杂交来检测。McrBC 消化的样品与没有用 McrBC 消化的平行样品进行比较，以鉴定被甲基化的区域。如果产物平行地与探针阵列杂交，针对样品中被甲基化的区域的
25 探针应在没有用 McrBC 消化的样品中显示出杂交，但是在用 McrBC 消化的样品中不显示出杂交。因为片段中甲基化的存在是通过检测是否存在限制性片段来检测的，在设计合适的探针时有相当的灵活性，这将是合适的。例如，被检测的片段一般在 200 和 1,000 个碱基对之间，探针可以被靶向到片段的任何区域。探针无需与甲基化的位点互补，但是可以与上游或下游位点互补。探针可以被靶向到片段的一个
30

区域或片段中的多个区域，它们可以被靶向到片段上的特定特征，例如片段中的 SNP 或片段中的一个或多个 CpG。在一个实施方案中，可以使用含有均匀分隔在整个基因组的探针的探针阵列。

5 在一个示例的实施方案中，扩增的产物被标记，并且与基因分型阵列，例如作图的 10K 或 100K 阵列(Affymetrix, Santa Clara)杂交。GeneChip 作图阵列(WGSA)可以被用来减少样品的复杂性。基本分析步骤如下：总基因组 DNA(250ng)用限制性酶（例如 Xba I）消化，与能够识别 4 个粘着碱基对突出的接头连接。所有的限制性酶消化片段，
10 不论大小，都是接头连接的底物。识别接头序列的通用引物被用来扩增连接了接头的 DNA 片段。扩增使用了最适于优先扩增选定的大小范围（例如 250 到 2000bp）的片段的 PCR 条件。对于不同的大小范围，例如 200 到 1,000 碱基对，可以选择最优化条件。然后，扩增的 DNA 被片段化、标记并与作图 10K 阵列杂交。与由计算机模拟消化
15 所预测的存在于选定大小范围内的片段（例如当基因组用 Xba I 消化时长度为 250 到 1000bp）中的基因组区域互补的探针被选择出来包含在阵列中。通过这种方法，扩增富集了亚类片段，同亚类的片段可以被重复地富集，阵列被设计用来质询这些片段中的至少一部分。作图 10K 和 100K 阵列质询存在于预计的片段上的已知 SNPs 的基因型，但是
20 是在其他实施方案中，阵列可以被设计用来质询片段是否存在。对于关于作图 10K 阵列和分析的其它信息参见 GeneChip 人类作图 10K 阵列和分析试剂盒数据单，零件号码 701366 Rev. 4, Affymetrix, Inc., 以及作图 10K 手册。

25 在一个实施方案中，含有与生物体中的基因互补的探针的阵列可以被用来分析甲基化的或未甲基化的组分。例如，可得自 Affymetrix 的表达阵列，例如人类基因组 U133 Plus 2.0 阵列，可以被使用。有多种生物体，包括小鼠和大鼠的表达阵列可以获得，并且对于选择的生物体，其表达阵列也可以被定制。含有针对预测或已知的外显子或剪
30 接连接（内含子-外显子或外显子-外显子剪接点）的探针的阵列也可

以使用。

5 在一个实施方案中，平铺了完整的基因组，一个或多个完整的染色体或完整的基因组的代表或一个或多个完整的染色体的高密度阵列可以被用来分析通过分离甲基化和未甲基化 DNA 而制备的样品。例如，可以使用这样的阵列，它含有的探针沿着一个或多个染色体或完整基因组分隔在平均每 35 个碱基对上。参见，例如 Kapranov 等 Science 296:916-919 (2002)。也可参见美国专利申请 Nos. 10/741,193、10/736,054、10/714,253 和 10/712,322。在一个实施方案中，已经被富集了未甲基化的序列的样品可以通过转录因子结合亲和性进行分析。与转录因子结合的序列可以通过与转录因子的亲和而纯化，然后通过阵列分析得以鉴定。通过富集甲基化的序列，用仅切割未甲基化 DNA 的酶消化，同样可以降低复杂性。

15 众多甲基依赖的限制性酶对于本领域的专业技术人员来说是众所周知的，可以从例如 New England Biolabs 商购到。甲基依赖的限制性酶的例子包括 McrBC、McrA、MrrA 和 DpnI。McrBC 是核酸内切酶，可以切割在一条或两条链上含有甲基胞嘧啶（例如，5-甲基胞嘧啶或 5-羟甲基胞嘧啶或 N4-甲基胞嘧啶，在 Raleigh, E. A. (1992) Mol. Microbiol. 6, 1079-1086 中有综述）的 DNA。McrBC 不作用于未甲基化的 DNA (Sutherland, E. 等 (1992) J. Mol. Biol. 225, 327-334)。McrBC 的识别位点是 5' ... Pu^mC (N₄₀₋₃₀₀₀) Pu^mC ... 3'。被 McrBC 识别的 DNA 上的位点由两个 (G/A)^mC 形式的半位点组成。这些半位点可以彼此分开多达 3kb，但是最适的分离距离为 55-103 个碱基对 (Stewart, F. J. 和 Raleigh E. A. (1998) Biol. Chem. 379, 611-616 以及 Panne, D. 等 (1999) J. Mol. Biol. 290, 49-60.)。McrBC 的切割需要 GTP，但是在存在非可水解的 GTP 类似物的情况下，酶将特异性地与甲基化的 DNA 结合，但不切割 (Stewart, F. J. 等 (2000) J. Mol. Biol. 298, 611-622)。重组 McrBC 可以得自例如 New England Biolabs。McrBC 可以被用来确定 CpG 二核苷酸的甲基化状态。McrBC 作用于一对 Pu^mCG 序列元件，

20

25

30

但是不识别内部的胞嘧啶被甲基化了的 Hpa II/Msp I 位点 (CCGG)。这个非常短的半位点共有序列(Pu^mC)使得存在的大部分甲基胞嘧啶能够被检测到。

5 在一个实施方案中, 用 McrBC 消化的条件是 50 mM NaCl, 10 mM Tris-HCl, 10 mM MgCl₂, 1 mM 二硫苏糖醇 (25°C, pH 7.9), 以及 100 µg/ml BSA 和 1mM GTP。37°C 保温。条件可以改变。NEB 定义一个单位为 37°C 下在 50 µl 总反应体积中, 1 小时内切割 1 µg 含有单个 McrBC 位点的质粒所需的酶量。切割基因组 DNA 时可以使用 5-10 倍过量的酶。酶可以通过加热至 65°C 20 分钟来失活。McrBC 在每一对半位点之间造成一次切割, 切割靠近一个半位点或另一个半位点, 但是切割位置分布在距离甲基化的碱基大约 30 个碱基对的几个碱基对中。也参见, Bird, A.P. (1986) Nature 321, 209-213 和 Gowher, H.等 (2000) EMBO J. 19, 6918-6923。

15 对 McrBC 的研究或利用在文献中已有报道, 例如 Gast 等 Biol Chem. 378(9):975-82, (1997), Pieper 等, Rabinowicz, Methods Mol Biol. 236:21-36 (2003), Badal 等 J Virol. 77(11):6227-34 (2003)以及 Chotai 和 Payne, J Med Genet. 35(6):472-5 (1998)。也可参见 Lyko, F. 等 Nat. Genet., 23, 363-366 (2000), 其中使用了 McrBC 作为在果蝇中富集未甲基化的 DNA 的工具。

25 在一个方面, 基因组 DNA 通过本领域所熟知的任何方法被分成甲基化的组分和未甲基化的组分。每个组分可以单独与阵列杂交, 或者每个组分可以用不同的可检测的标记物, 例如不同颜色的荧光染料进行标记 (例如, 未甲基化的 DNA 可以用绿色标记, 而甲基化的 DNA 可以用红色标记), 然后, 将上述两者与同样的探针阵列杂交。参见, 例如美国专利 No. 6,576,424, 在此引为参考。如果基因组的区域没有甲基化, 那么对应于该基因组区域的特征将以绿色被检测。如果该区域是甲基化的, 那么这些特征将以红色被检测。如果红色和绿色都被

30

检测到了，样品中该区域可能已被部分甲基化，红色与绿色的比率可以被用来确定甲基化的程度。

5 在一个方面，公开的方法被用来获得肿瘤或组织的甲基化特征或分布图。在癌症的诊断、治疗和结果预测方面，甲基化受到特别的关注，参见 Jones 和 Baylin, *Nat. Rev. Genet.* 3:415-428 (2002)以及 Bird, *Genes Dev.* 16:6-21 (2002)。甲基化的模式可能与特定的肿瘤相关。来自特定类型肿瘤的样品可以被分离，并且使用公开的方法进行分析，获得肿瘤类型或肿瘤阶段的甲基化模式特征。在一个实施方案中，来自个体或肿瘤的样品可以与已知类型或阶段的肿瘤的甲基化模式进行比较，以确定未知的样品在甲基化模式上是否与一个或多个已知的肿瘤类型相似。按照这种方法获得的模式可以用来诊断疾病、疾病阶段、
10 监控治疗、预测治疗结果以及监控疾病进展。在许多实施方案中，分析通过直接比较杂交模式进行，而模式与任何特定序列的存在或不存在无关。从被分析的未知样品获得的模式和从已知样品获得的模式之间的不同或相同点可以被用来确定未知样品是否在甲基化模式上可能与已知样品相匹配。

20 在一个实施方案中，血液样品被分析来检测脱落到血流中的肿瘤细胞的甲基化模式的变化。异常的甲基化或脱甲基化模式是肿瘤类型的特征，可通过血样的分析进行鉴定。在一个示例的实施方案中，样品用第一个限制性酶片段化，并且片段被连接到接头上。然后，连接了接头的片段用甲基化依赖或甲基化敏感的酶消化。不被消化的连接了接头的片段通过 PCR 使用针对接头的引物进行扩增。PCR 扩增的
25 产物与探针的阵列杂交以产生杂交模式。杂交模式可以与来自另一个被同样处理的样品的杂交模式进行比较。杂交模式之间的区别指示两个样品之间甲基化模式的区别。疾病状态、正常状态或组织类型的特征的杂交模式可以产生数据库，并用来比较未知样品的杂交模式，以鉴定出相似的模式。参见，例如美国专利 No. 6,228,575，它公开了基于比较杂交模式对样品表征的方法。有多种阵列可用于这个目的，阵
30

列没必要特异地设计来检测来自被分析生物体的特定的基因组序列。

5 在一个实施方案中，未甲基化的 DNA 的富集与比较基因组杂交（CGH）相结合来分析肿瘤细胞，以鉴定肿瘤 DNA 和正常 DNA 之间的区别。参见，例如 Kallioniemi 等 *Methods* 9(1):113-121 (1996)。等量的不同标记的肿瘤 DNA 和正常的参比 DNA（例如一个可以用生物素标记，而另一个用地高辛标记）可以与探针阵列杂交，信号强度被定量，并且肿瘤相对于正常细胞中的过量或不足信号可以被定量。在一个实施方案中，分析甲基化状态的方法可以与基因组中一个或多个区域的拷贝数的估计方法相结合。许多癌症与基因组中一个或多个区域的拷贝数的增加有关。拷贝数的增加可以通过与阵列的杂交而检测。拷贝数的增加被检测为杂交强度的增加。使用寡聚核苷酸阵列分析拷贝数的方法已公开在，例如美国专利公布 No. 20040157243 中，其中公开了特定的计算机方法，使用例如 Affymetrix 10K 作图阵列和分析进行拷贝数分析。

20 在另一方面，在存在或不存在甲基化的基础上，利用组分的分离来减少复杂性的方法被用来富集样品中的目的序列，该样品是宿主和病原基因组 DNA 的混合物。一些生物体在它们的基因组中缺少 5-meC 修饰或水平降低的 5-meC。例如，病原体如支原体缺少 5-meC 或水平很低。其它的例子参见，例如 Razin 和 Razin, *NAR* 8:1383-1390 (1980)。未甲基化的病原体 DNA 可以通过用甲基依赖性酶如 McrBC 消化样品来富集。未甲基化的病原体 DNA 也可以通过使用针对 5-meC 的抗体或 5-meC 结合蛋白与针对结合蛋白的抗体相结合来消耗甲基化的 DNA 而得以富集。在一个方面，样品首先用其识别位点中没有 CpG 的限制性酶进行片段化，再将接头连接到片段上。连接有接头的片段用甲基化依赖性酶消化，以便被甲基化并且含有酶识别位点的片段被片段化。没有被甲基化依赖性酶片段化的连接了接头的片段通过 PCR 使用针对接头的引物进行扩增。这样产生了相对甲基化的 DNA 来说未甲基化的 DNA 被富集了的扩增产物。

在一个实施方案中，利用那些优先扩增未甲基化的核酸的方法来减少基因组样品的复杂性的方法，可以被用来在复杂的混合物中富集病原体 DNA。例如，如果核酸样品是从被认为感染有病原体的病人中分离出来的，该核酸样品可能含有病人的 DNA 和病原体的 DNA 的混合物。许多原核生物的病原体与它们感染的生物体相比具有较低的甲基化水平，因此在扩增前用优先降解甲基化 DNA 的酶处理混合样品可以被用来相对于宿主 DNA 来说富集病原体 DNA。然后可以通过例如与核酸探针阵列的杂交来分析扩增的样品以检测病原体 DNA。由于宿主 DNA 的存在而引起的潜在的干扰效应被减小了，从而对病原体 DNA 的检测得到改善。

在一个实施方案中，被怀疑含有病原体 DNA 的核酸样品被片段化以产生片段，以及接头被连接到片段的末端。然后，接头修饰的片段用切割甲基化 DNA 但是不切割未甲基化 DNA 的酶，例如 McrBC 进行处理。含有 McrBC 识别位点的片段将被切割为较小的片段，它们只在一个末端有接头序列。接着，样品使用与接头中的序列互补的引物通过 PCR 进行扩增。被 McrBC 切割的片段将不被扩增，因为它们只有一个接头，因而只在一个末端有引发位点。由于病原体的序列没有被甲基化，它不被 McrBC 切割，并会被扩增。

可以被用来检测甲基化的阵列包括例如平铺阵列，该阵列中的探针针对目的区域而言，在亚硫酸氢盐处理后，与多个可能的 CpG 和 5-meCpG 的组合完全互补。两条链或一条链上的甲基化都可以被分析。如果探针针对一条链设计，它们也可以被设计用来质询任意一条链。在某些方面，选择有待质询的链是含有胞嘧啶的链，但是在其它方面，那条链已经在修饰后被扩增了，以至于得到的扩增的双链产物含有 A:T 碱基对代替了 C:G 碱基对，任何一条链都可以被质询。所有未甲基化的胞嘧啶可以被转化为尿嘧啶，探针和引物在设计时可以将这考虑在内，例如，在与胞嘧啶位置互补的位置上，探针与基因组中胞嘧啶的

位置互补的位置应有腺嘌呤。

据估计在人类基因组中有大约 2800 万到 2900 万个 CpG，其密度在低密度的 CpG 区域预计是大约每 50-100 个碱基对 1 个 CpG，而在高密度区域大约是每 20 个碱基对 1 个 CpG。在一个方面，阵列被设计成质询超过 50,000、超过 100,000、超过 500,000、超过 1,000,000、超过 2,500,000 或超过 5,000,000 个这些 CpG 的甲基化状态。在某些实施方案中，阵列还含有质询也能够在胞嘧啶上被甲基化的 CNG 位置的探针。质询可以例如类似于检测胞嘧啶位点的多态性，反映出由于化学，例如亚硫酸氢盐或酶学，例如 AID 的机制产生的从胞嘧啶到尿嘧啶的变化。基于相邻 CpG 二核苷酸的定位可以选择出特定的 CpG 来质询。当在探针互补的区域内，例如探针的 25 个碱基中，有多于 1 个 CpG 时，为质询中央 CpG 所需的探针的完全互补性可能受到探针区域内第二、第三或第四个 CpG 的甲基化状态的影响。在某些方面，用于质询第一个 CpG（质询的 CpG）的探针组可以被设计成包含了由于次级（非质询的）CpG 的甲基化状态的改变而引起的所有可能的序列变化的组合。对于每一个可能的序列变化来说，这需要附加的探针。在另一方面，在上游或下游 12、15、20 或 30 个碱基内没有另外的 CpG 被选择来质询。

结论

本发明公开了保留后生信息的基因组 DNA 扩增方法。在一个方面，DNA 甲基转移酶活性被包括在基因组 DNA 的扩增过程中。该甲基转移酶活性识别半甲基化的位点，并在扩增过程中使用甲基化的模板 DNA 为指导对新合成的 DNA 进行甲基化。

上面的描述是说明性的而非限制性的。对于本领域的专业技术人员来说，在浏览了本公开后，显然可以对本发明做出许多修改。因此，本发明的范围将不是参照上面的描述而确定，而是参照随附的权利要求及其所有等同物的范围而确定。