(19)中华人民共和国国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 102348979 A
(43)申请公布日 2012.02.08

(21)申请号 201080011326.4
(22)申请日 2010.02.19
(30)优先权数据
61/158,683 2009.03.09 US
61/241,347 2009.09.10 US
(85)PCT申请进入国家阶段日
2011.09.09
(86)PCT申请的申请数据
PCT/US2010/024830 2010.02.19
(87)PCT申请的公布数据
WO2010/104662 EN 2010.09.16
(71)申请人 乔治亚大学研究基金公司
地址 美国乔治亚州
申请人 吉林大学
(72)发明人 崔娟 李凡 大卫·普特 C·洪 徐鹰
(74)专利代理机构 北京三友知识产权代理有限公司 11127
代理人 丁香兰 庞东成
(51)Int.Cl.
G01N 33/48(2006.01)
G01N 33/53(2006.01)

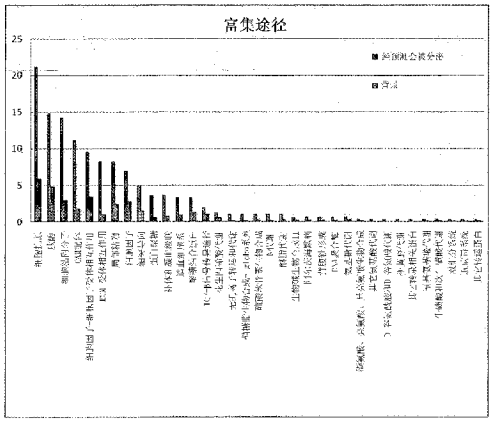
权利要求书 3 页 说明书 56 页 附图 26 页

(54)发明名称

胃癌诊断用蛋白标记的鉴定

(57)摘要

本发明提供了通过检测分泌到生物液中的蛋白而检测癌的方法以及诊断癌的方法。首先将本发明应用于检测分泌到血清和尿中的蛋白。但是，应该理解的是，所述方法具有更广泛的应用，以开发用于检测分泌到其它生物液中的蛋白的工具和系统，所述其它生物液例如，但不限于唾液、脊髓液、精液、阴道液和眼内液。通过该方法的实施方式提供的对分泌到生物液中的蛋白进行的可信检测能够更及时准确地检测和诊断癌。



CN 102348979 A

1. 确定用于检测癌的血清蛋白标记的方法，所述方法包括：

(a) 获得癌样品和参照样品；

(b) 确定在所述癌样品和所述参照样品之间差异性表达的一个或多个基因；

(c) 鉴定作为所述一个或多个基因的产物的一个或多个蛋白；

(d) 预测所述一个或多个蛋白被分泌到生物液中的可能性；和

(e) 检测据预测会分泌到所述生物液中的所述一个或多个蛋白在所述生物液中的存在，

其中所述生物液中的所述一个或多个蛋白的检测构成癌的检测。

2. 如权利要求 1 所述的方法，其中所述癌样品或所述参照样品包含组织样品。

3. 如权利要求 1 所述的方法，其中在所述癌样品和所述参照样品之间所述一个或多个基因的表达存在至少 1.5 倍的变化。

4. 如权利要求 1 所述的方法，其中在所述癌样品和所述参照样品之间所述一个或多个基因的表达存在至少 2 倍的变化。

5. 如权利要求 1 所述的方法，其中，与参照样品相比，所述一个或多个基因的表达增加。

6. 如权利要求 1 所述的方法，其中，与参照样品相比，所述一个或多个基因的表达减少。

7. 如权利要求 1 所述的方法，其中所述确定在所述癌样品和所述参照样品之间差异性表达的一个或多个基因的步骤包括从所述癌样品和所述参照样品分离总 RNA。

8. 如权利要求 7 所述的方法，其中所述确定在所述癌样品和所述参照样品之间差异性表达的一个或多个基因的步骤进一步包括对从所述癌样品和所述参照样品分离的 RNA 进行微阵列分析。

9. 如权利要求 1 所述的方法，所述方法还包括鉴定在所述癌样品和所述参照样品之间差异性产生的一个或多个蛋白的特征。

10. 如权利要求 9 所述的方法，其中鉴定在所述癌样品和所述参照样品之间差异性产生的一个或多个蛋白的特征的步骤包括鉴定在所述癌样品中相对于所述参照样品差异性表达的基因。

11. 如权利要求 9 所述的方法，其中鉴定在所述癌样品和所述参照样品之间差异性产生的一个或多个蛋白的特征的步骤包括鉴定在癌样品中相对于参照样品差异性表达的基因剪接变体。

12. 如权利要求 9 所述的方法，其中鉴定在所述癌样品和所述参照样品之间差异性产生的一个或多个蛋白的特征的步骤包括鉴定能够区分所述癌样品和所述参照样品的标记基因。

13. 如权利要求 9 所述的方法，其中所述预测包括使用所鉴定的在所述癌样品和所述参照样品之间差异性产生的一个或多个蛋白的特征，并且其中所述特征对应于在已知被分泌到所述生物液中的蛋白的集合中呈现的性质。

14. 如权利要求 13 所述的方法，其中在已知被分泌到所述生物液中的蛋白的集合中存在的性质包括：一般序列特征、物化性质、结构性质、以及结构域和基序。

15. 如权利要求 14 所述的方法，其中所述一般序列特征包括：氨基酸组成、序列长度、

二肽组成、序列顺序、标准化 Moreau-Broto 自相关指数和 Geary 自相关指数。

16. 如权利要求 14 所述的方法，其中所述物化性质包括：疏水性、标准化范德华体积、极性、极化率、电荷、二级结构、溶剂可进入性、溶解度、不可折叠性、非稳定区、全局电荷和亲水性。

17. 如权利要求 14 所述的方法，其中所述结构性质包括：二级结构含量和形状。

18. 如权利要求 14 所述的方法，其中所述结构域和基序包括：信号肽、跨膜域、糖基化和双 - 精氨酸信号肽基序 (TAT)。

19. 如权利要求 1 所述的方法，其中所述检测包括对所述生物液进行质谱分析。

20. 如权利要求 1 所述的方法，其中所述检测包括对所述生物液进行蛋白质印迹分析。

21. 如权利要求 1 所述的方法，其中所述检测包括对所述生物液进行 MS/MS 分析。

22. 如权利要求 1 所述的方法，所述方法还包括在所述检测之前除去在所述生物液中存在的最丰富的蛋白。

23. 如权利要求 22 所述的方法，所述方法包括使用抗体柱除去在所述生物液中存在的最丰富的蛋白。

24. 如权利要求 23 所述的方法，所述方法还包括在除去所述生物液中存在的最丰富的蛋白之后从所述抗体柱洗脱非特异性结合的蛋白。

25. 如权利要求 23 所述的方法，所述方法还包括从所述抗体柱洗脱特异性结合的蛋白以用于进一步分析。

26. 如权利要求 22 所述的方法，其中所述生物液中存在的最丰富的蛋白包括白蛋白、IgG、α 1- 酸糖蛋白、α 2- 巨球蛋白、HDL(载脂蛋白 A-1 和 A-II) 和纤维蛋白原。

27. 如权利要求 1 所述的方法，其中所述生物液是血清、唾液、血液、尿、脊髓液、精液、阴道液、羊膜液、龈沟液或眼内液中的一种或多种。

28. 如权利要求 1 所述的方法，其中所述癌包括胃癌、胰腺癌、肺癌、卵巢癌、肝癌、结肠癌、结直肠癌、乳癌、鼻咽癌、肾癌、子宫颈癌、脑癌、膀胱癌、肾癌和前列腺癌、黑素瘤以及鳞状细胞癌。

29. 如权利要求 1 所述的方法，其中所述蛋白为人类蛋白。

30. 诊断罹患癌的患者的方法，所述方法包括：

(a) 从所述患者获得生物液；和

(b) 检测所述生物液中一个或多个标记蛋白的存在，

其中所述一个或多个标记蛋白是在癌样品和参照样品之间差异性表达的一个或多个基因的产物，其中所述一个或多个标记蛋白据预测且经实验证实会分泌到所述生物液中，并且其中所述生物液中的所述一个或多个标记蛋白的检测构成癌的检测。

31. 诊断罹患癌的受试对象的方法，所述方法包括：

(a) 从所述受试对象获得生物液；和

(b) 测定所述生物液中一个或多个标记蛋白的水平，

其中所述一个或多个标记蛋白是在癌样品和参照样品之间差异性表达的一个或多个基因的产物，其中所述一个或多个标记蛋白据预测且经实验证实会分泌到所述生物液中，并且其中所述生物液中的所述一个或多个标记蛋白相对于标准水平的差异性表达指示癌。

32. 如权利要求 31 所述的方法，其中所述差异性表达包括所述生物液中的所述一个或

多个蛋白的水平相对于所述标准水平增加。

33. 如权利要求 31 所述的方法，其中所述差异性表达包括所述生物液中的所述一个或多个蛋白的水平相对于所述标准水平减少。

34. 如权利要求 31 所述的方法，其中一个或多个标记蛋白选自由 MUC13、GKN2、COL10A、AZTP1、CTSB、LIPF、GIF、EL 和 TOP2A 组成的组。

35. 用于癌鉴定的标记，所述标记包含选自由 MUC13、GKN2、COL10A、AZTP1、CTSB、LIPF、GIF、EL 和 TOP2A 组成的组的一个或多个蛋白，其中获自受试对象的生物液中的所述一个或多个蛋白相对于标准水平的差异性表达指示所述受试对象中癌的出现。

36. 如权利要求 32 所述的标记，其中所述差异性表达包括所述生物液中的所述一个或多个蛋白的水平相对于所述标准水平增加。

37. 如权利要求 32 所述的标记，其中所述差异性表达包括所述生物液中的所述一个或多个蛋白的水平相对于所述标准水平减少。

38. 一种用于检测受试对象中的癌的试剂盒，所述试剂盒包含 ：

(a) 与生物液中的蛋白特异性结合的一种或多种一抗，其中所述蛋白选自由 MUC13、GKN2、COL10A、AZTP1、CTSB、LIPF、GIF、EL 和 TOP2A 组成的组 ；

(b) 与所述一种或多种一抗特异性结合的二抗 ；以及可选的是，

(c) 参照样品。

胃癌诊断用蛋白标记的鉴定

[0001] 发明背景

技术领域

[0002] 本发明主要涉及用于检测和 / 或诊断癌的检测患者的生物液中的蛋白标记的方法。

背景技术

[0003] 癌领域中的主要挑战之一是检测处于早期的癌的能力。早期癌检测方面的挑战主要由于大多数癌在其早期不具有明显的可以暗示癌的身体症状所致。经证明如乳房造影法或结肠镜检查等身体检查是有效的，但是仅限于特定类型的癌，例如乳癌或结直肠癌。此外，当通过所述身体检查进行检测时，即使定期进行所述身体检查，癌可能已经超过了早期。非常常见的是当癌已经处于晚期时才被诊断，显然，需要用于早期癌检测的更有效的技术。

[0004] 基因和蛋白表达的变化提供了关于组织或器官的生理状态的重要线索。恶性转化期间，肿瘤细胞中的基因变化可以干扰自分泌信号传导网络和旁分泌信号传导网络，引起例如生长因子、细胞因子或可以被分泌到癌细胞外部的激素等某类蛋白的过表达 (Hanahan 和 Weinberg，2000 ；Sporn 和 Roberts，1985)。这些分泌蛋白以及其它分泌蛋白可以通过复杂的分泌途径进入血清、唾液、血液、尿、脑脊液 (脊髓液)、精液、阴道液、眼内液、或其它生物液。

[0005] 虽然如果检测出癌，组织标记基因可用于对癌进行分级，但是它们不可直接用于癌诊断，除非疑似为特定的癌并且对相关组织进行探测。来自生物液的蛋白标记确实是用于标记鉴定的最终目标，因为它们允许通过简单的分析测试来进行癌检测。

[0006] 但是，生物液 (例如，血清) 中癌标记 (蛋白、肽或其它分子) 的鉴定与癌组织的基因表达研究相比，由于分子组成的复杂性更高和人类血清中分子丰度的动态范围较宽 (可能高达 6 个数量级，差异范围从 mg/ml 至 ng/ml)，因此代表了更有挑战性的问题。例如，人类血清蛋白组是高丰度的天然血清蛋白的非常复杂的混合物，所述天然血清蛋白例如白蛋白和免疫球蛋白、以及由不同病变组织或正常组织分泌的或者从遍及人体的细胞渗漏的蛋白和肽。诸如疾病、饮食、甚至精神状态等许多因素都能相当迅速地改变血清中的分子组成及其丰度。将这些组织综合，大多数循环性天然血液蛋白的丰度比大多数经分泌的蛋白的丰度高出几个数量级。这些组织使得极其难以对来自患者群体和参照群体的生物液的蛋白组进行直接比较分析以用于生物标记鉴定。

[0007] 基因组技术和蛋白组技术的最近进展使得对于鉴定用于癌早期检测的有效标记产生了极大热情和新的希望。通过使用诸如微阵列芯片等技术对癌组织与参照组织中的基因表达模式进行比较分析，即使对于非常早期的癌，也可以检测某些基因在癌组织中相对于正常组织的表达模式的持续变化。这是可行的，因为随着癌经过关键的发育阶段的发展，会获得许多新能力，例如 (a) 生长信号的自足性，(b) 对于抗生长信号的不敏感性，(c) 躲避

凋亡，(d) 无限复制潜能，(e) 持续的血管生成和 (f) 组织入侵和转移，每一种都会改变某些基因的“正常”表达模式，例如，增加其表达水平以产生所获能力所需的相关蛋白；并且这些蛋白中的一些能够分泌到血液循环中，提供用于通过血液测试进行癌检测的可能痕迹。

[0008] 使用组学 (omics) 技术，已经提出了同时位于癌组织和血清中的许多标记。质谱法一直是用于对诸如血清等生物液中的蛋白进行蛋白组学研究的主要技术，特别是用于对诸如血清等生物液中的蛋白的鉴定和定量 (Tolson 等，2004)。

[0009] 表达蛋白的全局模式可用于某些病例，但是由于表达蛋白的全局模式的高度复杂性，显然它们不是良好的标记。

[0010] 本领域的普遍共识是现有标记未有效地起作用，并且需要根本性的新观点以鉴定更有效的癌检测用标记，特别是对于早期癌检测。

[0011] 本领域存在的另一问题是为了诊断癌和其它疾病，必须对以下情况做出准确的预测，即何种来自病变组织中（例如癌）中异常表达基因的蛋白可以被分泌到生物液中。与解决该问题相关的困难在于，目前对蛋白被分泌到细胞外部后的下游定位的理解非常有限，现有知识不足以提供关于蛋白到生物液的分泌方面的有用提示。因此，所需要的是用于预测何种蛋白可能被分泌到生物液中的数据分类方法。

[0012] 本发明人认为将可源自癌组织的微阵列数据的信息与使用计算方法对生物液进行的蛋白组学研究结合，呈现出一种以更系统的方式发现新颖且更为有效的标记的新颖且更为有效的方法。

发明内容

[0013] 本发明公开了用于检测癌的方法以及通过检测分泌到生物液中的蛋白来诊断癌的方法。通过本发明的实施方式提供的对分泌到生物液中的蛋白进行的可信检测会允许更及时准确地检测和诊断癌。

[0014] 在一个实施方式中，本发明公开了确定用于癌检测的蛋白标记的方法，所述方法包括：a) 获得癌样品和参照样品；b) 确定在所述癌样品和所述参照样品之间差异性表达的一个或多个基因；c) 鉴定作为所述一个或多个基因的产物的一个或多个蛋白；d) 预测所述一个或多个蛋白被分泌到生物液中的可能性；和 e) 在所述生物液中检测经预测会分泌到所述生物液中的所述一个或多个蛋白的存在，其中所述生物液中的所述一个或多个蛋白的检测构成癌的检测。

[0015] 在另一实施方式中，本发明公开了诊断罹患癌的患者的方法，所述方法包括：a) 从所述患者获得生物液；和 b) 检测所述生物液中一个或多个标记蛋白的存在，其中所述一个或多个标记蛋白是在癌样品和参照样品之间差异性表达的一个或多个基因的产物，其中所述一个或多个标记蛋白据预测且经实验验证会分泌到所述生物液中，并且其中所述生物液中的所述一个或多个标记蛋白的检测构成癌的检测。

[0016] 在第三实施方式中，本发明公开了诊断罹患癌的受试对象的方法，所述方法包括：a) 从所述受试对象获得生物液；和 b) 测定所述生物液中一个或多个标记蛋白的水平，其中所述一个或多个标记蛋白是在癌样品和参照样品之间差异性表达的一个或多个基因的产物，其中所述一个或多个标记蛋白据预测且经实验证实会分泌到所述生物液中，并且其中所述生物液中的所述一个或多个标记蛋白相对于标准水平的差异性表达指示癌。

[0017] 在又一实施方式中，本发明公开了用于癌鉴定的标记，所述标记包括选自由以下蛋白组成的组中的一个或多个蛋白：MUC13、GKN2、COL10A、AZTP1、CTSB、LIPF、GIF、EL 和 TOP2A，其中获自受试对象的生物液中的所述一个或多个蛋白相对于标准水平的差异性表达指示所述受试对象中癌的出现。

[0018] 在另一实施方式中，本发明公开了用于检测受试对象中的癌的试剂盒，所述试剂盒包含：(a) 与生物液中的蛋白特异性结合的一种或多种一抗，其中所述蛋白选自由 MUC13、GKN2、COL10A、AZTP1、CTSB、LIPF、GIF、EL 和 TOP2A 组成的组；(b) 与所述一种或多种一抗特异性结合的二抗；以及可选的是，(c) 参照样品。

[0019] 为了说明本发明，首先将本发明应用于检测分泌到血清和尿中的蛋白。但是，应该理解，可将本发明更广泛地应用到开发用于检测分泌到其它生物液中的蛋白的工具和系统，所述其它生物液例如，但不限于，唾液、脊髓液、精液、阴道液和眼内液。

附图说明

[0020] 图 1 显示 (a) 在转录物的全长上选择探针选择区 (PSR) 的示意图。PSR 下面的短划线表示用于各 PSR 的各个探针（来源：Affymetrix：人、小鼠及大鼠用GeneChip® Exon 阵列系统）。浅色区表示外显子，深色区表示在剪接期间被除去的内含子。(b) 三个所预测剪接同种型的 PCR 数据。x 轴是组织样品轴 (12 个组织样品)，其中 NC 是阴性对照。Y 轴是质量轴。(i) 略过外显子 2 的一个同种型；和 (ii) 分别是略过替代性外显子 2（下方）和略过外显子 1（上方）的两个同种型。(c) 外显子同种型和探针的示意图。长的水平线表示人类基因组的部分，最窄的矩形表示外显子，三个较宽的矩形表示三个外显子同种型，位于底部的较短的黑线表示探针。

[0021] 图 2 描述了 (a) 在癌组织中相对于参照组织差异性表达的总共 2,540 个基因和在早期癌中差异性表达的 1,276 基因的维恩图 (Venn diagram)。(b) 在癌组织和参照组织之间所述 2,540 个基因的表达差异性的分布。

[0022] 图 3 描述了 (a) 所述 2,540 个差异性表达的基因、911 个癌相关基因和 1,276 个在早期癌中差异性表达的基因的功能家族分布。(b) 以上三组基因的亚细胞位置分布 (*Cyt.：细胞质；Nuc.：细胞核；E.R.：内质网；Pla.：质膜；Ext.：细胞外间隙)。

[0023] 图 4 描述了（上部）癌组织中 MUC1 的表达水平作为年龄的函数而改变，其与性别无关；（下部）THY1 的表达与年龄和性别都无关。

[0024] 图 5 描述了在基因的子集的 80 个样品上鉴定的双基因簇 (bi-cluster)，其中各行表示基因，各列表示一对癌组织 / 参照组织，(a)C1（上部）具有 244 个在癌组织中相对于参照组织一致性上调的基因；C2（中部）具有 95 个基因，其大多数下调；C3（下部）具有 53 个显示复合模式的基因。要注意的是用于不同双基因簇的组织样品的顺序不必相同，因为所述算法会将组织样品的顺序重排。(b) 可能具有亚型特异性的双基因簇，由 42 个基因组成。已知以竖线标记的 6 个基因与胃癌的亚型相关。

[0025] 图 6 描述了一个盒式图，显示了在出现所预测的外显子 - 略过事件时的紧邻上游内含子区 (-150nt，+30nt) 中的匹配基序的分布。

[0026] 图 7(a) 以竖线标记的曲线表示 k 基因标记 (k = 1，..，100) 的总精度，其是 500 个随机选择的子集的最佳精度的平均值；以十字交叉标记的曲线表示通过穷举搜索鉴定出的

k 基因标记 (k = 1,..,8) 的 5 倍交叉验证 (5-cross validation) 精度。(b) 最佳 28 个基因标记的热图，其包括 13 个上调基因和 15 个下调基因。其中，NKAP、TMEM185B、C14orf104 和 Clorf96 上调，而 KLF15、PI16 和 GADD45B 在 > 89％的早期患者中下调。

[0027] 图 8 描述了从对照组和癌组收集的血清样品的 MS 总离子色谱图。(a) 对照组的基峰位于左侧，癌组的基峰位于右侧；(b) 不同的分子量范围。

[0028] 图 9 描述了以下 8 个蛋白的蛋白质印迹 (SDS-PAGE 之后转移至硝酸纤维素以随后用抗体进行印迹）：MUC13、GKN2、COL10A1、AZTP1、CTSB、LIPF、GIF 和 TOP2A，显示了对照组和胃癌组之间丰度的差异。1)MUC13(1μg，稀释度：一抗 1 ： 200；抗兔二抗，1 ： 10,000)；2)GKN2(150μg，稀释度：一抗 1 ： 1,000；抗兔二抗，1 ： 30,000)；3)COL10A1(1μg，稀释度：一抗 1 ： 500；抗兔二抗，1 ： 10,000)；4)AZTP1(120μg，稀释度：一抗 1 ： 500；抗鼠二抗，1 ： 3,000)；5)CTSB(5μg，稀释度：一抗 1 ： 1,500；抗兔二抗，1 ： 20,000)；6)LIPF(120μg，稀释度：一抗 1 ： 500；抗羊二抗，1 ： 10,000)；7)GIF(120μg，稀释度：一抗 1 ： 5,00；抗鼠二抗，1 ： 3,000)；和 8)TOP2A(60μg，稀释度：一抗 1 ： 350；抗羊二抗，1 ： 10,000)。

[0029] 图 10 描述了 d 值和 p 值之间的统计关系 = P(TP)，d 表示离位于阳性训练数据和阴性训练数据之间的分离超平面的距离。

[0030] 图 11 描述了由注释、可视化及综合发现用数据库 (Database for Annotation, Visualization and Integrated Discovery(DAVID)) 富集的功能组。DAVID 提供了一套全面的功能注释工具以理解大的基因列表所隐藏的生物学意义。x 轴表示功能组，y 轴表示富集度。

[0031] 图 12 使用 KEGG 直系同源类注释系统 (Orthology-based Annotation System(KOBAS)) 网络服务器描述了 480 个所预测尿蛋白的富集途径。KOBAS 鉴定了与背景分布相比所查询序列中经常出现（或显著富集）的途径。各组中较短的条形表示所述 480 个蛋白的百分比，各组中较长的条形表示所有人类蛋白；x 轴表示途径名称；以及 y 轴表示百分比。

[0032] 图 13 描述了 480 个蛋白的代表性不足 (underrepresented) 的途径。各组中较短的条形表示所述 480 个蛋白的百分比，各组中较长的条形表示所有人类蛋白；x 轴表示途径名称；以及 y 轴表示百分比。

[0033] 图 14 描述了 3 个正常样品 (N1、N2、N3) 和 3 个胃癌样品 (SC1、SC5、SC11) 的 274 个细胞因子的抗体阵列。人类 G6 阵列显示 Fit3- 配体（白色矩形）；人类 G7 阵列显示 EGF-R（深灰色矩形）、SGP-130（白色矩形）；人类 G8 阵列显示 PDGF-AA（白色矩形）；人类 G9 阵列显示 Trappin-2（浅灰色矩形）、黄体化激素（白色矩形）、TIM-1（深灰色矩形）；人类 G10 阵列显示 CEACAM1（浅灰色矩形）、FSH（白色矩形）、CEA（深灰色矩形）。

[0034] 图 15 描述了三个癌样品 (GC) 和三个对照样品 (CTRL) 的粘蛋白 13(Mucin13) 的蛋白质印迹。各泳道含有 1μg 的尿蛋白。Santa Cruz Mucin 13(M-250) 兔多克隆抗体以 1 ： 200 稀释使用；抗兔二抗以 1 ： 10,000 稀释使用。

[0035] 图 16 描述了三个对照样品 (CTRL) 和三个癌样品 (GC) 的 COL10A1 的蛋白质印迹。各泳道含有 1μg 的尿蛋白。Calbiochem 的抗胶原 X 型 Rabbit pAb 以 1 ： 200 稀释使用；抗兔二抗以 1 ： 10,000 稀释使用。

[0036] 图 17(上部)三个对照样品 (CTRL) 和三个胃癌样品 (GC) 的内皮脂肪酶 (EL) 的蛋白质印迹。各泳道含有 1μg 的尿蛋白。用于 EL 的抗体是 Santa Cruz EL(C-19) 亲和纯化羊多克隆抗体 (1 ： 200 稀释);抗羊二抗以 1 ： 15,000 稀释使用。(下部)前 7 条泳道对应于正常样品;后 7 条泳道是癌样品。

[0037] 图 18 描述了对前列腺癌和对照数据通过最佳 1- 基因标记和 2- 基因标记得到的分类表现。y 轴是分类精度,x 轴是通过其分类精度分选的前 100 个最佳标记的列表。

[0038] 图 19 显示使用基于生物素标志的抗体阵列进行的蛋白阵列实验的结果。图 19 描述了癌血清和参照血清之间在 103 个蛋白中的蛋白丰度差异性的分布,x 轴表示以其丰度差异性的 log 值的升序分选的 103 个蛋白的列表,y 轴是丰度差异性的 log 值。

[0039] 现在参照附图描述本发明。应该理解的是本申请的附图不必按比例绘出,并且这些图和图解仅是说明性的,并不限制本发明。

具体实施方式

[0040] 本发明涉及检测癌的方法,所述方法通过以下步骤进行 :预测蛋白是否被分泌到生物液中,以及通过在蛋白组学研究中确定所述生物液中所述蛋白的存在来验证所述预测,所述生物液例如但不限于血清、唾液、血液、尿、脊髓液、精液、阴道液和眼内液,其中所述生物液中所述蛋白的检测构成了癌的检测。本发明包括诊断罹患癌的患者的方法的实施方式,所述实施方式通过以下步骤进行 :检测所述患者的生物液中由癌组织中的异常表达基因表达的一个或多个标记蛋白的存在,其中所述标记蛋白据预测并经实验验证会分泌到所述生物液中,并且其中所述生物液中的所述标记蛋白的检测构成癌的检测。

[0041] 各种生物液中的任一种都适于使用本发明的装置和方法进行分析。所述生物液包括脑脊液、滑液、血液、血清、血浆、唾液、肠液、精液、眼泪、鼻分泌物等。应该意识到根据本发明可同样地使用任何流体生物样品 (例如,组织提取物或活组织检查提取物、粪便提取物、痰等)。

[0042] 在以下出于说明目的的描述中,所陈述的具体数值、参数和试剂是为了对本发明提供全面的理解。但是,应该理解的是,本发明无需这些具体细节即可实施。在某些情况下,为了不使本发明模糊,可以省略或简述公知特征。

[0043] 说明书中所述的实施方式和参考文献提到“一种实施方式”、“本发明的实施方式”、“实施方式”、“示例性实施方式”等,表示所述的实施方式可以包括特定的特征、结构或特性,但是每一个实施方式可以不必包括该特定的特征、结构或特性。此外,以上术语不必指同一实施方式。另外,当将特定的特征、结构或特性结合实施方式进行描述时,应该理解,无论是否明确指出,在本领域中已知都可以结合其它实施方式实现所述特征、结构或特性。

[0044] 本文的描述“a”或“an”物品可以指单数物品或复数物品。例如,某特征、蛋白、生物液或分类器可以是单个的特征、蛋白、生物液或分类器。作为另一种选择,某特征、蛋白、生物液或分类器可以是多个的特征、蛋白、生物液或分类器。因此,如本文所用,“a”或“an”可以是单数或复数的。类似地,对于复数项目的提及或描述可以指代单个项目。

[0045] 应该理解的是,在本文无论何处以语言“包含”来描述实施方式,也就另外提供了以术语“由 ..... 组成”和 / 或“基本上由 ..... 组成”描述的类似实施方式。

[0046] 说明书描述了通过检测生物液中标记蛋白的存在来检测和诊断癌的通常方法。本

文提供了用于检测血清中的标记蛋白的具体示例性实施方式。本说明书公开了一个或多个并入本发明的特征的实施方式。所公开的实施方式仅仅是对本发明的举例说明。本发明的范围不限于所公开的实施方式。本发明由所附的权利要求限定。

[0047] 虽然说明书中所要求保护的方法及其对应的描述通常要求保护的特征是对癌检测用蛋白标记的检测，应该理解的是，针对所述蛋白标记的存在对样品进行分析、发现没有所述标记蛋白并由此未诊断出癌仍然是对所述蛋白标记的存在性的检测。

[0048] 定义

[0049] 术语“多肽”、“肽”、“蛋白”和“蛋白片段”在本文中可相互替换地使用以指代氨基酸残基的聚合物。这些术语适用于其中一个或多个氨基酸残基是相应天然存在的氨基酸的人工化学模拟物的氨基酸聚合物，以及天然存在的氨基酸聚合物和非天然存在的氨基酸聚合物。如本文所用，“蛋白”或“肽”通常是指大于约 200 个氨基酸至最大为从基因翻译的全长序列的蛋白；多肽为约 100 个氨基酸～ 200 个氨基酸；和 / 或“肽”为约 3 个氨基酸～约 100 个氨基酸，但并不限于以上定义。如本文所用，“氨基酸”是指任何天然存在的氨基酸、本领域已知的任何氨基酸衍生物或任何氨基酸模拟物。在某些实施方式中，蛋白或肽的残基是连续的，没有任何非氨基酸打断氨基酸残基的序列。在其它实施方式中，所述序列可以包含一个或多个非氨基酸部分。在特定实施方式中，蛋白或肽的残基的序列可以被一个或多个非氨基酸部分打断。

[0050] 术语“氨基酸”是指天然存在的氨基酸和合成的氨基酸，以及与天然存在的氨基酸功能类似的氨基酸类似物和氨基酸模拟物。天然存在的氨基酸是由遗传密码编码的那些氨基酸，以及被稍后修饰的那些氨基酸，例如羟基脯氨酸、γ- 羧基谷氨酸和 O- 磷酸丝氨酸。氨基酸类似物是指与天然存在的氨基酸具有相同的基本化学结构（例如与氢结合的 α 碳、羧基、氨基和 R 基）的化合物，例如高丝氨酸、正亮氨酸、蚕氨酸亚砜、蚕氨酸甲基锍。所述类似物可以具有经修饰的 R 基（例如正亮氨酸）或经修饰的肽主链，但是保留与天然存在的氨基酸相同的基本化学结构。氨基酸模拟物是指具有与氨基酸的一般化学结构不同的结构但是其功能与天然存在的氨基酸类似的化合物。

[0051] 如本文所用，受试对象或患者中的“癌”是指拥有致癌细胞的典型特性的细胞的存在，所述典型特性例如不受控的增殖、永生化、转移潜能、快速生长和增殖速率和某些特征性形态学特征。通常，癌细胞是肿瘤的形式，但是此类细胞可以在受试对象内单独存在，或可以是非致瘤性癌细胞，例如白血病细胞。在某些情况下，癌细胞是肿瘤的形式，此类细胞可以在动物内局部存在，或在血流中作为独立细胞循环，例如白血病细胞。癌的实例包括但不限于乳癌、黑色素瘤、肾上腺癌、胆管癌、膀胱癌、脑癌或中枢神经系统癌、支气管癌、母细胞瘤、癌 (carcinoma)、软骨肉瘤、口腔癌或咽癌、子宫颈癌、结肠癌、结直肠癌、食道癌、胃肠癌、成胶质细胞瘤、肝癌、肝细胞瘤、肾癌、白血病、肝癌、肺癌、淋巴瘤、非小细胞肺癌、骨肉瘤、卵巢癌、胰腺癌、外周神经系统癌、前列腺癌、肉瘤、唾液腺癌、小肠癌或阑尾癌、小细胞肺癌、鳞状细胞癌、胃癌、睾丸癌、甲状腺癌、膀胱癌、子宫癌或子宫内膜癌和外阴癌。

[0052] 如本文所用，“样品”是指从患者、优选从人类患者获得的生物材料的样品，包括组织、组织样品、细胞样品，例如活组织检查（例如抽吸活组织检查、刷拭活组织检查、表面活组织检查、针吸活组织检查、钻取活组织检查、切除活组织检查、切开活组织检查、切取活组织检查或内窥镜活组织检查），肿瘤样品或从所述组织样品提取的 RNA。样品还可以是生物

液样品，包括但不限于尿、血液、血清、血小板、唾液、脑脊液、乳头抽吸液和细胞裂解物（例如全细胞裂解物的上清液、微粒体级分、膜级分或细胞质级分）。可以使用任何本领域已知的方法获得所述样品。

[0053] “生物样品”是指从个体获得的任何生物样品，包括但不限于，粪便（大便）样品、生物液（例如血液）、细胞、组织样品、RNA 样品或组织培养物。从哺乳动物获得大便样品、组织活组织检查或其它生物样品的方法是本领域公知的。

[0054] 如本文所用，“组织样品”是指从受试对象的完整组织获得或移取的组织的部分、碎片、局部、片段或级分。

[0055] 术语“基因”是指包含产生多肽、前体或 RNA（例如 rRNA、tRNA）所需的编码序列的核酸（例如，DNA）序列。术语“基因”包括基因的 cDNA 和基因组形式。

[0056] 基因的基因组形式或克隆物含有被命名为“内含子”或“插入区”或“插入序列”的非编码序列打断的编码区或“外显子”。内含子从核转录物或初级转录物中被除去或“剪除”；因此在信使 RNA(mRNA) 转录物中不存在内含子。除了含有内含子之外，基因的基因组形式还包括位于存在于 RNA 转录物上的序列的 5′ 和 3′ 末端的序列。将这些序列称为“侧接”序列或“侧接”区（这些侧接序列处在相对存在于 mRNA 转录物上的非翻译序列的 5′ 或 3′ 处）。

[0057] 应该理解的是，对于特定的 mRNA 剪接变体而言“内含子”和“外显子”是相对的，一种剪接变体的外显子可以是另一种剪接变体的内含子，反之亦然。但是，在一个剪接变体内，“内含子”不能是“外显子”，反之亦然。这些术语“内含子”和“外显子”在本文是为方便和清楚起见而使用的，并非意在限制。

[0058] 如本文所用，术语“基因表达”是指通过内源基因、其 ORF 或部分、或植物中的转基因的“转录”（例如，经由 RNA 聚合酶的酶促作用），将在内源基因、其 ORF 或部分、或植物中的转基因中编码的遗传信息转换为 RNA（例如 mRNA、rRNA、tRNA 或 snRNA）的过程，并且对于蛋白编码基因而言，通过 mRNA 的“翻译”转换为蛋白的过程。另外，表达是指正义 (mRNA) 或功能性 RNA 的转录和稳定累积。在该过程中的许多阶段可以调节基因表达。“上调”或“激活”是指增加基因表达产物（例如，RNA 或蛋白）的产生的调节，而“下调”或“阻遏”是指减少产生的调节。涉及上调或下调的分子（例如转录因子）经常分别称为“激活子”或“阻遏子”。

[0059] 术语“差异性表达的基因”、“差异性基因表达”及其同义词可相互替换地使用，是指相对于所述基因在正常受试对象或对照受试对象中的表达，其在罹患疾病、特别是癌（例如胃癌）的受试对象中的表达被激活至更高水平或更低水平的基因。这些术语还包括其表达在相同疾病的不同阶段被激活至更高水平或更低水平的基因。还应该理解的是，差异性表达的基因可以在核酸水平或蛋白水平被激活或抑制，或可以经受替代性剪接以产生不同的多肽产物。所述差异可以由例如 mRNA 水平、多肽的表面表达、分泌或其它配分的改变而证明。差异性基因表达可以包括两个或多个基因或其基因产物之间的表达的比较，或两个或多个基因或其基因产物之间的表达比例的比较，或甚至是相同基因的两种不同加工产物的比较，所述两种不同加工产物在正常受试对象和罹患疾病（特别是癌）的受试对象之间不同、或在相同疾病的不同阶段之间不同。差异性表达包括定量以及定性差异，例如正常细胞和病变细胞之间、或经历不同疾病事件或疾病阶段的细胞之间的时间上或基因或其

表达产物中的细胞表达模式上的定量及定性差异。出于本发明的目的，当在正常受试对象和病变受试对象中或在病变受试对象的疾病发展的不同阶段中给定基因的表达之间的差异至少为约 1.5 倍、2 倍，优选至少约 4 倍、更优选至少约 6 倍、最优选至少约 10 倍时，认为存在“差异性基因表达”。

[0060] 如本文所用，术语“受试对象”或“患者”是指疑似患有癌或待要经受特定诊断的任何动物（例如，哺乳动物），包括但不限于人类、非人类灵长类和啮齿动物等。通常，提及人类受试对象时，在本文术语“受试对象”或“患者”可相互替换地使用。

[0061] 如本文所用，“正常受试对象”或“对照受试对象”是指未罹患疾病的受试对象。

[0062] 诸如“治疗中”、或“治疗”或“待治疗”、或“缓解”或“待缓解”等术语是指 1) 治愈、减慢、减轻所诊断病理性病况或病症的症状和 / 或暂停发展的治疗性措施，以及 2) 预防和 / 或减慢所针对的病理性病况或病症的发展的预防性或防止性措施。因此需要治疗的那些包括已经罹患所述病症的那些对象、倾向于罹患所述病症的那些对象和其中待预防所述病症的那些对象。如果患者显示出以下情况中的一种或多种，则已根据本发明的方法成功地“治疗”了受试对象 ：癌细胞的数量减少或完全不存在 ；肿瘤尺寸的减小 ；浸润到周围器官的癌细胞（包括例如癌至软组织和骨的扩散）的抑制或不存在 ；肿瘤转移的抑制或不存在 ；肿瘤生长的抑制或不存在 ；与特定癌相关的一种或多种症状的缓解 ；发病率和致死率减少 ；生活品质提高 ；或某些效果组合。

[0063] 如本文所用，术语“分类器”是指用于执行数据分类的方法、算法、计算机程序或系统。

[0064] 如本文所用，术语“分类”是学习将数据点分成不同类别的过程，其通过发现在已知类别内所收集的数据点之间的共同特征而进行。可以使用神经网络、回归分析或其它技术完成分类。

[0065] 如本文所用，术语“数据分类方法”表示一种一般性计算方法的类别，其试图基于所提供的各数据要素的特征值，确定给定数据集合中的各数据要素属于哪种预定义类别。

[0066] 术语“基于抗体的结合部分”或“抗体”包括免疫球蛋白分子和免疫球蛋白分子的免疫活性决定簇，例如含有特异性结合蛋白（与蛋白发生免疫反应）的抗原结合位点的分子。术语“基于抗体的结合部分”试图包括完整抗体，例如任何同型（IgG、IgA、IgM、IgE 等）的完整抗体，并且包括其也与抑制蛋白或其片段特异性反应的其片段。可以使用常规技术将抗体片段化。因此，该术语包括抗体分子的蛋白水解 - 切割的部分或重组制备的部分的区段 (segment)，其能够与特定蛋白选择性地反应。所述蛋白水解片段和 / 或重组片段的非限制性实例包括 Fab、F(ab′)2、Fab′ 、Fv、dAbs 和含有通过肽连接子连接的 VL 域和 VH 域的单链抗体 (scFv)。scFv 可以共价连接或非共价连接以形成具有两个或多个结合位点的抗体。因此，“基于抗体的结合部分”包括多克隆抗体、单克隆抗体或抗体和重组抗体的其它纯化制品。术语“基于抗体的结合部分”还试图包括人源化抗体、双特异性抗体和具有至少一个源自抗体分子的抗原结合决定簇的嵌合抗体。在优选实施方式中，对基于抗体的结合部分进行可检测标记。

[0067] 如本文所用，“经标记抗体”包括通过可检测手段标记的抗体，并且包括但不限于被酶促、放射性、荧光和化学发光标记的抗体。还可以用诸如 c-Myc、HA、VSV-G、HSV、FLAG、V5 或 HIS 等可检测标记将抗体标记。

[0068] 本发明的一个方面中，提供了确定癌检测用血清蛋白标记的方法，所述方法包括：a) 获得癌样品和参照样品；b) 确定在所述癌样品和所述参照样品之间差异性表达的一个或多个基因；c) 鉴定作为所述一个或多个基因的产物的一个或多个蛋白；d) 预测所述一个或多个蛋白被分泌到生物液中的可能性；和 e) 在所述生物液中检测据预测会分泌到所述生物液中的所述一个或多个蛋白的存在，其中所述生物液中的所述一个或多个蛋白的检测构成癌的检测。

[0069] 癌样品和参照样品可以从相同受试对象或从不同受试对象获得。“参照样品”是指含有基线量的一个或多个基因的表达的样品，该基线量在一个或多个不患有癌的受试对象中确定。基线可以从至少一个受试对象获得，并且优选从平均量的受试对象（例如，n = 2 ～ 100 或更多）获得，其中所述受试对象之前没有癌病史。基线还可以从来自疑似罹患癌的受试对象的一个或多个正常样品获得。例如，基线可以从至少一个正常样品获得，并且优选从平均量的正常样品（例如，n = 2 ～ 100 或更多）获得，其中所述受试对象疑似罹患癌。在一个方面，与参照样品相比，一个或多个基因的表达在癌样品中可以增加。在另一方面，与参照样品相比，一个或多个基因的表达在癌样品中可以减少。

[0070] 基因表达的分析

[0071] 对在癌样品和参照样品之间差异性表达的一个或多个基因的确定包括从癌样品和参照样品分离核酸。核酸样品可以是总 RNA、cDNA 样品、聚 (A)RNA、不含一种或多种 RNA 的 RNA 样品，例如不含 rRNA 的 RNA 样品或 RNA 的扩增产物。在一个方面，所述样品来自哺乳动物，例如人类、大鼠或小鼠。所述样品还可以分离自组织，包括例如血液、肺、心脏、肾、胰腺、前列腺、睾丸、子宫、大脑或皮肤。

[0072] 在癌样品和参照样品之间差异性表达的基因可以通过本领域已知的任何手段检验，包括但不限于微阵列图谱、聚合酶链式反应 (PCR)、基于多核苷酸的杂交分析的方法、基于多核苷酸的测序的方法、基于选择性基因剪接的分析的方法和基于蛋白组学的方法。

[0073] 用于通过将生物液中的 RNA 定量而研究基因表达的本领域已知的广泛应用的方法包括微阵列分析、RNA 印迹分析 (Harada，1990) 和原位杂交 (Parker&Barnes，1999)；核糖核酸酶保护检验 (Hod，1992)；S1 核酸酶作图 (Fujita 等，1987) 和基于 PCR 的方法，例如逆转录聚合酶链式反应 (RT-PCR)(Weis 等，1992)、定量 RT-PCR 和连接酶链式反应 (LCR)(Barany，1991)，这些都是本领域的常规方法。作为另一选择，可以使用能够识别具有序列特异性的双链体（包括 DNA 双链体、RNA 双链体和 DNA-RNA 杂交双链体或 DNA- 蛋白双链体）的抗体。基于测序的基因表达分析的代表性方法包括基因表达系列分析 (SAGE) 和通过大规模平行特征序列 (parallel signature) 测序 (MPSS) 进行的基因表达分析。

[0074] 在一个实施方式中，确定在癌样品和参照样品之间差异性表达的一个或多个基因包括从癌样品和参照样品分离总 RNA。用于总 RNA 提取的通常方法是本领域已知的，并且记载于分子生物学的标准课本中，包括 Ausubel 等，Current Protocols ofMolecularBiology，John Wiley 和 Sons(1997)。

[0075] 在优选实施方式中，对分离自癌样品和参照样品的总 RNA 使用微阵列分析来研究在癌样品中相对于参照样品差异性表达的基因。

[0076] 在另一实施方式中，使用 RNA 印迹分析研究在癌样品中相对于参照样品差异性表达的基因。

[0077] 在又一实施方式中，使用 RNA 酶保护检验研究在癌样品中相对于参照样品差异性表达的基因。

[0078] 在另一实施方式中，通过使分离的细胞 RNA 与经放射性标记的合成 DNA 序列杂交来评估 RNA 的表达，以便确定在癌样品中相对于参照样品差异性表达的基因，所述经放射性标记的合成 DNA 序列与所关注 RNA 的 5′ 末端具有同源性。

[0079] 在另一实施方式中，使用聚合酶链式反应 (PCR) 研究在癌样品中相对于参照样品差异性表达的基因。

[0080] 在另一实施方式中，使用 RT-PCR 研究在癌样品中相对于参照样品差异性表达的基因。

[0081] RT-PCR 技术的最近变化形式是实时定量 PCR，其通过经双标记的荧光发生探针（即 TaqMan™ 探针）测定 PCR 产物的累积。实时 PCR 与以下 PCR 均相容：其中将各靶序列的内部竞争物用于标准化的定量竞争性 PCR，以及与使用包含在样品内的标准化基因或 RT-PCR 用管家基因的定量比较 PCR。详细资料参见例如 Held 等，1996。

[0082] 可以使用代替 PCR 的替代性方法，例如“连接酶链式反应”(″ LCR″) 来研究基因表达 (Barany，1991)。

[0083] 另外的基于 PCR 的技术例如包括：差异性展示 (Liang 和 Pardee，1992)；扩增片段长度多态性 (iAFLP)(Kawamoto 等，1999)；BeadArray™ 技术 (Illumina，San Diego，Calif.；Oliphant 等，Discovery of Markers for Disease(Supplement to Biotechniques)，2002 年 6 月；Ferguson 等，2000)；在基因表达用快速检验中使用商购 Luminex100LabMAP 系统和多色编码的微球 (Luminex Corp.，Austin，Tex.) 的用于检测基因表达的珠阵列 (BADGE)(Yang 等，2001)；和高覆盖表达图谱 (HiCEP) 分析 (Fukumura 等，2003)。

[0084] 在本发明的另一实施方式中，通过基因表达系列分析 (SAGE) 研究在癌样品中相对于参照样品差异性表达的基因。

[0085] 在本发明的另一实施方式中，通过大规模平行特征序列测序 (MPSS) 研究在癌样品中相对于参照样品差异性表达的基因。关于该方法的描述，参见 Brenner 等，(2000)。

[0086] 迄今，此前关于癌标记的研究一直不能检查全人类转录物组，由于缺乏有效研究手段而未能检查大多数人类转录物组、由基因的选择性剪接生成的剪接变体。因此，在本发明的另一实施方式中，通过鉴定在癌样品中相对于参照样品差异性表达的剪接变体来研究在癌样品中相对于参照样品差异性表达的基因。

[0087] 选择性剪接是这样的真核细胞过程，通过其经由包含外显子的不同部分和 / 或经由保留内含子而可以从同一前 mRNA 产生多种成熟的 mRNA 转录物。据估计至少 40％～ 75％的人类基因在不同条件下经受选择性剪接 (Modrek 和 Lee，2002)。选择性剪接是造成人类转录物组和蛋白组的复杂性的主要原因。此前的估计表明，人类蛋白组具有由约 20,000 个基因编码的至少约 100,000 个、可能至多约 150,000 个不同的蛋白，表明每个人类基因平均编码 5 ～ 7 个蛋白。因此，人类细胞中大多数功能蛋白是剪接同种型，强调了当研究基因表达和蛋白（在本案中为生物液中的标记蛋白）时研究剪接变体的需要。

[0088] 已知选择性剪接涉及人类的许多生物过程 (Nakao 等，2005)，在正常和异常的功能过程中都涉及。异常剪接可对细胞的正常功能具有严重的影响。最近的调查回顾了在 12 种癌类型中出现在 p53 剪接位点处的 29 个突变 (Holmila 等，2003)。另一最近研究发现约

200 个基因的 464 个剪接变体在人类前列腺癌中差异性表达 (Li 等,2006)。

[0089] 在一个实施方式中,由 Affymetrix 进行的新兴外显子阵列技术为研究选择性剪接提供了有力工具。

[0090] 外显子阵列数据的分析代表了一个具有挑战性的问题,因为所述阵列的基本单元是外显子而不是基因。使用诸如鲁棒多芯片平均法 (Robust Multichip Average, RMA) (Irizary 等,2003) 和探针对数强度误差 (Probe Logarithmic Intensity Error, PLIER) 估计法 (Affymetrix,2005) 等方法,可以从外显子阵列数据评估个体外显子的表达水平,而从所述表达水平并基于外显子的表达水平的相似性,可以推断出主要的剪接同种型。挑战在于在给定组织中,对于各个基因,可以存在具有不同表达水平的超过一种的表达剪接同种型,因此各外显子的所观察到的表达水平是含有该外显子的所有表达剪接同种型的总的表达水平。计算问题在于算出哪些剪接同种型被表达和以何种水平被表达,并且预测结果应该与外显子表达数据一致,但外显子表达数据通常具有噪音。虽然存在诸如 ANOVA(Affymetrix,2005) 等设计用于解读外显子阵列数据的计算机程序,因为外显子阵列从 2006 年才开始广泛应用,该问题提出了新的难题。关于外显子阵列数据的解读仍然存在许多挑战和未解决的问题。其中的关键问题是可信地预测主要的剪接同种型及其表达水平。

[0091] 能够被从组织分泌到血液循环中的蛋白的预测

[0092] 使用基因表达数据分析技术,已经鉴定或提出与诸如肝癌 (Smith 等,2003)、肾癌 (Young 等,2003)、乳癌 (van der Vijver 等、2002)、结直肠癌 (Resnick,2004) 和其它主要的癌 (Sallimen 等,2000 ;Hendrix 等,2001) 等特定的癌相关的许多基因。另外,已经提供用于评估癌阶段的几个标记。但是,通过将基于差异性基因表达数据得出的组织中的标记基因和通过蛋白组学分析发现的血清中的标记蛋白进行比较,观察到它们的关联相当弱,表明分别对癌组织和血清使用基因组学和蛋白组学技术得到的信息之间的无关联。

[0093] 因此,虽然如果检测出癌,组织标记基因可用于对癌进行分级,但是它们不直接用于癌诊断,除非疑似为具体的癌并且对相关组织进行探测。获自生物液的标记确实是用于标记鉴定的最终目标,因为它们允许通过简单的分析测试来进行癌检测。将此成功完成的关键在于发现有效的途径来最大限度地利用源自在癌组织上进行的基因表达研究的信息,从而指导生物液中的癌标记鉴定。

[0094] 具有预测病变组织中的哪些蛋白能够被分泌到生物液中的能力会在将可源自微阵列表达数据的信息与生物液中标记蛋白的鉴定连接起来方面提供关键的联系。

[0095] 基于如信号肽、特定长度的跨膜域、氨基酸组成和蛋白功能等蛋白序列信息 (Mott 等,2002 ;Guda 等,2006),已经进行了许多研究来预测蛋白的亚细胞定位,所述蛋白包括能够被运输到细胞表面或被分泌到胞外环境中的蛋白 (Menne 等,2000 ;Nair 和 Rost,2005 ; Guda 等,2006 ;Horton 等,2007)。虽然这些程序能够预测蛋白是否能够由细胞分泌,但是它们不涉及所述蛋白在离开细胞后最终在何处。

[0096] 本发明中,该问题已经使用数据挖掘方法得以解决,所述数据挖掘方法通过以下过程进行 :首先收集已知由于各种病理性病况被分泌到生物液中的人类蛋白,所述生物液例如但不限于血清、尿、唾液、脊髓液、精液、阴道液、羊膜液、龈沟液和眼内液,所述蛋白可通过蛋白组学研究进行检查,然后就可用于预测这些蛋白的其物化性质以及其序列和结构

特征方面，鉴定在这些蛋白中存在的共同特征。使用该策略，已经开发并据报道用于预测能够从组织分泌到生物液中的蛋白的计算机程序。参见 PCT 申请第 PCT/US2009/053309 号，本文并入其全部内容作为参考。

[0097] 该算法的基本思路如下。通过广泛的文献检索产生大人类蛋白集合，如通过之前的蛋白组学研究所检测已知会由于各种病理性病况而分泌到血流中的人类蛋白。绘出这些分泌蛋白共有的特征的列表，所述特征包括其物化性质、氨基酸序列和基序，以及结构特征（表 1）。使用这些特征，对分类器进行训练来将能够被分泌到生物液中的蛋白与不能被分泌到生物液中的蛋白区分开。然后使用该算法来预测所述组织基因标记中的哪些可以被分泌到生物液中。

[0098] 在一个实施方式中，所述算法包括以下步骤：选择蛋白的阳性分泌类别；选择阴性集的代表性蛋白；映射 (mapping) 蛋白特征以构建特征集；对分类器进行训练以识别蛋白的类别的特性；确定所映射特征的精度和相关性；除去最不重要的特征以产生经再训练的分类器；接收蛋白序列；载体生成和扩增；预测所接收蛋白序列的类别；和返回所接收蛋白序列的预测结果。该算法的详细描述在共同未决的申请 PCT/US2009/053309 中提供。

[0099] 表 1：预测血液分泌蛋白的初始特征的列表

[0100]

性质的类型	特征	来源
一般性序列特征	氨基酸组成、序列长度、二肽组成	本地计算
	标准化 Moreau-Broto 自相关指数、Moran 自相关指数、Geary 自相关指数、序列顺序、假氨基酸组成	使用由新加坡国立大学、理学院、计算科学系内的生物信息和药物设计组(BIDD)开发的 Protein Feature Server (PROFEAT)来计算。
物化性质	疏水性、标准化范德瓦华体积、极性、极化率、电荷、二级结构和溶剂可及性	使用三种描述符进行本地计算：组成(C)、转移(T)和分布(D)。
	溶解度、不可折叠性、非稳定区(disorder region)、全局电荷	使用来自 Stockholm Bioinformatics Centre 的基于序列的 PROtein SOlubility 求值程序(PROSO)(Smialowski 等，2007)和结合跨膜拓扑学以及信号肽预测器(Phobius)来确定
结构性质	二级结构含量、形状(回转半径)	使用来自 European Molecular Biology Laboratory 的 Secondary Structural Content Prediction (SSCP) 工具和来自 Supercomputing Facility for Bioinformatics & Computational Biology, Indian Institute of Technology (IIT), Delhi 的用于球状蛋白的回转半径过滤器进行确定
结构域和基序	信号肽、跨膜域(α 螺旋和 β 桶)、糖基化(N-连接和 O-连接)、双精氨酸信号肽基序(TAT)	使用来自位于丹麦技术大学的生物序列分析中心的 SignalP 工具和基于氨基酸组成的 TransMembrane Barrel-Hunt(TMB-Hunt) 工具(Garrow 等，2005)来确定。 使用来自位于丹麦技术大学的生物序列分析中心的 NetOglyc，NetNgly and Twin-arginine signal peptide (TatP)服务器来计算

[0101] 应该理解，对于不同的生物液而言蛋白特征可以不同。因此对于不同的生物液而言表 1 中所列出的特征可以不同。表 1 中所列出的蛋白特征可以粗分为四类：(i) 一般性序列特征，例如氨基酸组成、序列长度和二肽组成 (Bhasin 和 Raghava，2004；Reczko and Bohr，1994)；(ii) 物化性质，例如溶解度，非稳定区、疏水性、标准化范德华体积、极性、极化率和电荷；(iii) 结构特征，例如二级结构含量、溶剂可及性和回转半径，和 (iv) 结构域 / 基序，例如信号肽，跨膜域和双精氨酸信号肽基序 (TAT)。

[0102] 在一个实施方式中，选择注释为分泌蛋白并且从已知的蛋白数据库 (例如 Swiss-Prot and Secreted Protein Database(SPD) 数据库) 收集的人类蛋白，以及通过之前研究已经在血液中经过实验检测的蛋白。Chen 等 (2005) 描述了基于网络的 SPD。

[0103] 根据本发明的实施方式，以 FASTA 格式接收与从生物液收集的蛋白相符的蛋白序列。

[0104] 在本发明的其它实施方式中，以其它已知的格式接收与从生物液收集的蛋白相符

的蛋白序列，所述其它已知的格式包括但不限于仅包含字母字符的‘raw’文本格式。根据本发明的实施方式，在 raw 文本格式中所接收的蛋白序列中的任何空格符，例如空格、回车或 TAB 字符都被忽略。

[0105] 对于数据分离和回归模型可以广泛地执行各种受监督的学习方法，例如支持向量机 (SVM)、人工神经网络 (ANN)、决策树、回归模型和其它算法。基于已知数据（形式为训练数据集的知识），这些受监督的学习方法能够使计算机自动学习识别复杂的模式和开发分类器，其接下来可用于作出明智的决定和预测未知数据的类别（独立集）。

[0106] 在本发明的一个实施方式中，分类器是支持向量机 (SVM)。常规的 SVM 是基于定义判定边界的判定超平面的概念。判定超平面是将具有不同类别成员资格的目标的集合分开的超平面。例如，所收集的目标可以属于第一类或第二类，并且诸如 SVM 等分类器可用于确定（即预测）待分类的任何新目标的类别（例如，第一类或第二类）。常规的 SVM 是初级的分类器方法，其通过在分开不同类别标记的案例的多维空间中构建超平面来执行分类任务。SVM 可以支持回归任务和分类任务，并且可以处理多个连续的分类变量。在本发明的实施方式中，训练基于 SVM 的分类器来预测蛋白序列的类别是被分泌到生物液中还是不被分泌到生物液中。

[0107] 在本发明的另一实施方式中，分类器是专门化的、经改良的基于 SVM 的分类器。使用经改良的基于 SVM 的分类器来有效地计算蛋白被分泌到生物液中的可能性。高斯径向基函数核提供比用于 SVM 中的其它更常规的核（诸如线性核和多项式核）更优的性能。因此，在实施方式中，将高斯核 SVM 用于训练所述分类器。

[0108] 在本发明的另一实施方式中，对基于 SVM 的分类器进行进一步训练来预测通过微阵列基因表达实验检测到的异常高表达的基因是否将其蛋白分泌到血流中。研究已经鉴定了在诸如癌等各种病理状况的患者中显示异常高表达水平的许多此类基因。配备有该知识后，基于 SVM 的分类器可用于基于计算某些蛋白被排泄到患者血流中的可能性来诊断各种癌。

[0109] 在一个实施方式中，基于初始训练的各分类器的性能，使用命名为递归特征排除法 (RFE)(Tang 等，2007) 的特征选择方法来除去与分类目的无关或可以忽略的特征。

[0110] 根据一个实施方式，基于以上提出的多个数据集的结合，通过基于 SVM 的分类器产生的预测的总体预测精度为 79.5％～ 98.1％，对于独立评价测试和额外的血液蛋白测试，至少 80％已知的血液 - 分泌蛋白预测正确。从独立的负面评价测试可知，假阳性率经计算为约 10％（合理的经误分类为非血液 - 分泌蛋白的百分比），这有助于减轻与低精度相关的疑虑。

[0111] 分泌蛋白标记的验证

[0112] 一旦使用以上算法预测被分泌到生物液中的蛋白，则通过使用蛋白组学方法评估癌患者的生物液中这些蛋白标记的存在来验证这些蛋白标记。

[0113] 可以通过本领域已知的任何手段测定生物液中所述蛋白标记的存在，包括但不限于竞争结合检验、质谱、蛋白印迹、荧光激活细胞分选 (FACS)、酶联免疫吸附检验 (ELISA)、抗体阵列、高压液相色谱、光生物传感器和表面等离子共振。

[0114] 在一个实施方式中，对生物液样品进行处理以防止蛋白降解。抑制或预防蛋白降解的方法包括但不限于用蛋白酶处理生物液样品、将生物液样品冷冻、或将生物液样品置

于冰上。优选的是，在分析之前，将生物液样品持续地保持在防止蛋白降解的条件下。

[0115] 在一个实施方式中，生物液是血清，并且通过测定血清中的蛋白水平来确定蛋白水平。

[0116] 在一个实施方式中，生物液是血液，并且通过测定血液样品的血小板中的蛋白水平来确定蛋白水平。

[0117] 在一个实施方式中，生物液是尿，并且通过测定尿中的蛋白水平来确定蛋白水平。

[0118] 在一个实施方式中，在测定生物液中的蛋白水平之前除去生物液中存在的最丰富的蛋白。在一个方面，生物液中存在的最丰富的蛋白包括白蛋白、IgG、α1-酸糖蛋白、α2-巨球蛋白、HDL(载脂蛋白A-1和A-II)和纤维蛋白原。

[0119] 在一个实施方式中，使用抗体柱除去生物液中存在的最丰富的蛋白。

[0120] 在一个实施方式中，在除去生物液中存在的最丰富的蛋白之后将非特异性结合的蛋白从抗体柱洗脱。

[0121] 在一个实施方式中，将特异性结合的蛋白从抗体柱洗脱以用于进一步分析。

[0122] 在一个实施方式中，本发明的方法可以与检测其它分析物的方法一起进行，所述检测其它分析物例如检测mRNA或与癌有关的其它蛋白标记(例如，P-糖蛋白、β-微管蛋白、β-微管蛋白基因的突变或β-微管蛋白同型的过表达)。

[0123] 在一个实施方式中，通过使生物液与基于抗体的结合部分接触来检测蛋白，所述基于抗体的结合部分与该蛋白或和该蛋白的片段特异性结合。然后检测抗体-蛋白复合物的形成并对其进行测定以指示蛋白水平。抗-蛋白抗体可商购获得(例如来自明尼阿波利斯的R&D Systems，Inc.的人类蛋白亲和纯化的多克隆抗体和单克隆抗体，MN55413；AVIVA Systems Biology，圣地亚哥，CA 92121；还参见美国专利第5,463,026号)。作为另一选择，可以建立针对全长蛋白或蛋白的一部分的抗体。还可以使用生产抗体的标准方法生产用于本发明的抗体，例如通过单克隆抗体产生。

[0124] 在使用基于抗体的结合部分以检测分泌蛋白的本发明方法中，存在于生物液中的所关注蛋白的水平与从经可检测标记的抗体发出的信号强度相关。

[0125] 在一个优选实施方式中，通过将抗体与酶连接来对基于抗体的结合部分进行可检测标记。化学发光是可用于检测基于抗体的结合部分的另一方法。还可以使用各种免疫检验中的任一种来实现检测。例如，通过对抗体进行放射性标记，可以通过使用放射免疫检验来检测抗体。还可以使用荧光化合物来标记抗体。最常使用的荧光标记化合物是CYE染料、异硫氰酸荧光素、罗丹明、藻红蛋白、藻蓝蛋白、别藻蓝蛋白、邻苯二甲醛和荧光胺。还可以使用诸如^{52}Eu或镧系元素等荧光发射金属对抗体进行可检测标记。

[0126] 在一个实施方式中，可以通过免疫检验测定生物液中的蛋白水平，所述免疫检验例如酶联免疫吸附(ELISA)、放射免疫检验(RIA)、免疫放射检验(IRMA)、蛋白质印迹或免疫组织化学。还可以使用抗体阵列或蛋白芯片，参见例如美国专利申请：20030013208A1；20020155493A1；20030017515和美国专利：6,329,209；6,365,418，本文并入其全部内容作为参考。

[0127] 广泛使用的酶免疫检验是“酶联免疫吸附检验(ELISA)”。存在不同形式的ELISA，例如本领域公知的“夹心ELISA”和“竞争性ELISA”。本领域已知的ELISA标准技术记载于″Methods in Immunodiagnosis″，第二版，Rose和Bigazzi编著，John Wiley&Sons，

1980 ;Campbell 等， ″ Methods and Immunology ″ ，W.A.Benjamin，Inc.，1964 ； 和 Oellerich，1984。

[0128] 作为另一选择，可以通过将针对蛋白的经标记抗体导入受试对象中而在受试对象中体内检测细胞和 / 或肿瘤内的蛋白水平。例如，可以对抗体用放射性标记进行标记，所述放射性标记在受试对象中的存在和位置可以通过标准成像技术来检测。

[0129] 在一个实施方式中，使用免疫组织化学（″ IHC″ ）和免疫细胞化学（″ ICC″ ）技术。

[0130] 对于直接标记技术，使用经标记抗体。对于间接标记技术，使样品进一步与经标记物质反应。

[0131] 基于现有的公开内容，根据从业者的偏好可以使用其它技术来检测蛋白水平。一种此类技术是蛋白质印迹 (Towbin 等，1979)，其中经适当处理的生物液在 SDS-PAGE 凝胶上运行，然后被转移至诸如硝酸纤维素滤纸等固相载体上。在一个实施方式中，使用蛋白质印迹来检测血清或尿中的蛋白水平。在一个实施方式中，使用蛋白质印迹来检测血清或尿中的蛋白水平。然后使用经可检测标记的抗体来检测和 / 或评估蛋白水平，其中来自可检测标记的信号强度对应于蛋白的量。该水平可以例如通过光密度法定量。

[0132] 另外，可以使用质谱法检测蛋白水平，所述质谱法例如 MALDI/TOF（飞行时间）、SELDI/TOF、液相色谱 - 质谱 (LC-MS)、气相色谱 - 质谱 (GC-MS)、高效液相色谱 - 质谱 (HPLC-MS)、毛细管电泳 - 质谱、核磁共振光谱法或串联质谱（例如 MS/MS、MS/MS/MS、ESI-MS/MS 等）。参见例如，美国专利申请 :20030199001、20030134304、20030077616，本文并入它们作为参考。

[0133] 质谱法是本领域公知的，并且一直用于定量和 / 或鉴定诸如蛋白等生物分子（参见例如 Li 等 2000 ;Rowley 等，2000 ;以及 Kuster 和 Mann，1998)。此外，一直在开发允许对分离蛋白进行至少部分地从头测序的质谱技术（参见例如 Chait 等 1993 ;Keough 等，1999 ; Bergman 的综述，2000)。

[0134] 在某些实施方式中，使用气相离子分光光度法。在其它实施方式中，使用激光解吸 / 离子化质谱来分析生物液。现代的激光解吸 / 离子化质谱（″ LDI-MS″ ）可以以两种主要变化形式来运行 :基质辅助激光解吸 / 离子化（″ MALDI″ ）质谱和表面增强激光解吸 / 离子化（″ SELDI″ ）。

[0135] 关于额外的与质谱法有关的信息，参见例如 Principles of Instrumental Analysis， 第 3 版，Skoog，Saunders College Publishing，Philadelphia，1985 ； 和 Kirk-OthmerEncyclopedia of Chemical Technology，第 4 版第 15 卷 (John Wiley&Sons，New York1995)，第 1071-1094 页。

[0136] 检测蛋白标记的存在通常会包括检测信号强度。这反过来能够反映与底物结合的多肽的量和特性。例如，在某些实施方式中，可以比较来自第一样品和第二样品的光谱的峰值信号强度（例如，目视、通过计算机分析等），以确定具体生物分子的相对量。可以使用诸如 Biomarker Wizard 程序 (Ciphergen Biosystems，Inc，Fremont，Calif.) 等软件程序来辅助分析质谱。质谱及其技术是本领域技术人员公知的。

[0137] 应该理解的是，诸如解吸源、质量分析器、检测器等质谱仪的任何组件，以及各种样品制剂可以与本文所述或本领域已知的其它合适的组件或制剂组合。例如，在一些实施

方式中，对照样品可以含有重原子，例如 ^{13}C，由此允许在同一次质谱分析中将测试样品与已知的对照样品混合。

[0138] 在一个优选实施方式中，使用激光解吸飞行时间 (TOF) 质谱法。

[0139] 在一些实施方式中，部分地通过利用可编程数字计算机执行算法，来确定存在于生物液的第一样品或第二样品中的一个或多个蛋白的相对量。该算法鉴定第一质谱和第二质谱中的至少一个峰值。然后该算法将质谱中第一质谱的峰值强度与第二质谱的峰值强度进行比较。相对信号强度是存在于第一样品和第二样品中的蛋白的量的指示。可以对含有已知量的蛋白的标准物作为第二样品进行分析，以更好地对存在于第一样品中的蛋白的量进行定量。在某些实施方式中，还可以确定第一样品和第二样品中蛋白的身份。

[0140] 在本发明的一个实施方式中，通过 MALDI-TOF 质谱检测生物液中的蛋白水平。

[0141] 检测生物液中的蛋白的方法还包括使用表面等离子共振 (SPR)。

[0142] SPR 生物传感技术也已经与 MALDI-TOF 质谱结合以用于生物分子的解吸和鉴定。

[0143] 在一个实施方式中，使用抗体阵列检测生物液中的蛋白。在优选实施方式中，使用能够基于生物素标记的抗体阵列来检测蛋白。

[0144] 在一个实施方式中，本发明公开了诊断受试对象中的癌的方法，所述方法包括检测获自所述受试对象的生物液中的一个或多个标记蛋白。

[0145] 在另一实施方式中，本发明公开了诊断受试对象中的癌的方法，所述方法包括检测一个或多个标记蛋白在获自所述受试对象的生物液中相对于标准水平的差异性表达。在一个方面，所述一个或多个标记蛋白的差异性表达包括生物液中的所述一个或多个标记蛋白的水平相对于标准水平增加。在另一方面，所述一个或多个标记蛋白的差异性表达包括生物液中的所述一个或多个标记蛋白的水平相对于标准水平减少。

[0146] 在一个实施方式中，本发明公开了用于癌鉴定的标记，所述标记包括选自由以下蛋白组成的组中的一个或多个蛋白 :MUC13、GKN2、COL10A、AZTP1、CTSB、LIPF、GIF、EL 和 TOP2A，其中获自受试对象的生物液中的所述一个或多个蛋白相对于标准水平的差异性表达指示所述受试对象中癌的出现。

[0147] 在一个实施方式中，使用单基因标记来检测早期癌。

[0148] 在另一实施方式中，使用 2 基因标记来检测早期癌。

[0149] 在另一实施方式中，使用 k 基因标记 (k = 1...8) 来检测早期癌。

[0150] 在另一实施方式中，本发明公开了用于检测受试对象中的癌的试剂盒，所述试剂盒包含 :(a) 包含获自正常受试对象的生物液的参照样品 ;(b) 包含一种或多种与生物液中的蛋白特异性结合的一抗的溶液，其中所述蛋白选自由 MUC13、GKN2、COL10A、AZTP1、CTSB、LIPF、GIF、EL 和 TOP2A 组成的组 ;和 (c) 包含与所述一种或多种一抗特异性结合的二抗的溶液。

[0151] 根据以下对某些优选实施方式进行的更详细描述和权利要求，本发明的具体优选实施方式会变得明显。

[0152] 实施例

[0153] 以下实施例说明了本发明的具体实施方式及其各种应用。它们的描述仅仅是出于说明目的，而不应理解为对本发明的限制。

[0154] 实施例 1

[0155] 样品收集

[0156] 从相同的 80 名患者（肿瘤局限在粘膜或粘膜下层）收集总共 80 个胃癌组织（4 个 I 期、7 个 II 期、54 个 III 期以及 15 个 IV 期，来自 27 名女性和 53 名男性患者）和相同数量的相邻胃部但非癌性的组织。为了确保阵列实验中使用的 mRNA 的完整性，将所有组织在切除后 20 分钟内急速冷冻并贮存在液氮中。另外，还在外科手术前从每名癌患者收集血液样品。所有样品在中国长春的吉林大学医学院的 3 所附属医院和吉林省癌症医院收集。根据 WHO 标准和国际抗癌联盟的 TNM 分类系统由有经验的病理学家确定各个组织的组织分类和病理分期。根据肿瘤深度将癌分成早期（I 期和 II 期）和晚期胃癌（III 期和 IV 期）。诸如年龄、性别、组织分化、病理阶段以及饮酒 / 吸烟史等详细患者信息列于表 2。

[0157] 表 2 :(a) 患者统计信息，(b) 所收集样品的详细信息

[0158] (a)

[0159]

特性		患者	
		病例数量	百分比(%)
性别(n=80)	女性	27	33.8
	男性	53	66.2
阶段(n=80)	I	4	5.0
	II	7	8.8
	III	54	67.5
	IV	15	18.8
年龄(n=77)	>=55	53	68.8
	<55	24	31.2
抽烟(n=64)	是	18	28.1
	否	46	71.9
喝酒(n=64)	是	11	17.2
	否	53	82.8

[0160] (b)

[0161]

患者编号	年龄	性别	阶段	抽烟	喝酒	体重 (kg)
1	41	F	IV	0	0	43
2	62	F	III	0	0	70
3	54	F	III	0	0	70
4	62	F	IIIA	0	0	60
5	63	M	IIIB	1	1	-
6	56	M	IIIB	1	1	-
7	71	M	IIIB	1	0	-
8	55	F	IIIB	0	0	63
9	53	M	IIIB	0	0	60
10	-	M	IV	-	-	-
11	55	M	IIIB	0	0	60
12	51	M	IIIB	1	0	-
13	64	M	IIIB	0	0	55
14	53	F	IIIB	0	0	77
15	56	M	IIIB	1	0	55
16	54	M	III	0	0	70
17	53	M	III	0	0	62
18	71	M	III	0	0	60
19	57	M	IIIA	-	-	65
20	58	M	III	0	0	50
21	42	M	IB	0	0	52
22	73	M	IB	0	0	63
23	69	F	III	0	0	50
24	65	F	IIIA	0	0	-
25	50	M	III	1	0	47
26	47	M	IB	1	1	65
27	59	M	III	0	0	57
28	75	M	III	0	0	65
29	40	M	III	0	1	80
30	69	M	III	0	0	55
31	41	M	II	-	-	--
32	76	F	II	0	0	-
33	51	F	III	1	0	52
34	36	M	IIIA	1	0	60
35	67	F	IV	0	0	48
36	42	M	III	0	0	60

[0162]

患者编号	年龄	性别	阶段	抽烟	喝酒	体重 (kg)
37	68	M	III	0	0	50
38	65	M	III	0	1	50
39	59	M	III	1	1	51
40	68	M	IV	0	0	48
41	74	M	IB	0	0	62
42	65	F	IIIA	0	0	53
43	50	M	III	0	0	62
44	49	M	III	1	1	60
45	58	M	IV	0	0	66
46	-	F	IV	-	-	-
47	53	F	IIIA	1	0	60
48	84	M	IV	1	1	70
49	60	F	IIIB	0	0	60
50	55	M	III	0	0	50
51	70	M	II	1	0	59
52	56	F	III	0	0	45
53	43	F	III	0	0	55
54	71	F	III	0	0	42
55	56	F	IV	-	-	-
56	81	M	III	1	0	56
57	65	M	III	0	0	70
58	55	M	III	0	0	69
59	56	F	II	0	0	74
60	76	M	II	0	0	70
61	78	F	III	0	0	39
62	55	M	III	0	0	74
63	65	M	III	0	1	70
64	68	M	III	1	1	69
65	63	M	IV	0	0	-
66	-	M	IV	-	-	-
67	57	F	III	0	0	61
68	68	F	III	-	-	-
69	54	M	III	1	1	49
70	51	M	II	-	-	70
71	34	M	III	0	0	90
72	75	F	IV	-	-	40
73	61	M	III	1	0	70
74	54	M	IV	-	-	-
75	55	M	III	-	-	-
76	67	F	II	-	-	-
77	62	F	IV	-	-	-
78	50	F	III	-	-	-
79	71	M	IV	-	-	-

[0163]

患者编号	年龄	性别	阶段	抽烟	喝酒	体重 (kg)
80	58	M	IV	-	-	-

[0164] 实施例 2

[0165] RNA 制备和微阵列实验

[0166] 使用 Trizol 试剂 (Invitrogen) 从癌组织和参照组织提取总 RNA，然后使用 RNeasyMini 试剂盒 (QIAGEN) 根据制造商的建议进行纯化。使用 $A_{260}/A_{280} > 1.9$ 的比例和 28S/18S rRNA 等于 2，确保 RNA 样品是高度纯化的且未经降解。按照用于阵列实验的基因芯片表达分析技术手册 (Genechip Expression Analysis Technical Manual)(P/N900223) 中详述的策略，使用基因芯片人外显子 1.0ST(Affymetrix) 对 RNA 样品进行分析。简言之，在 rRNA 减少和 RNA 浓缩后使用 1μg 总 RNA 作为模板以合成 cDNA。通过体外逆转录，获得 cRNA 并将其用作第二轮循环中 cDNA 合成用模板。接着利用 RNA 酶 H 将 cRNA 水解，通过两种核酸内切酶将正义链 DNA 消化。使用 DNA 标记试剂将片段化的样品标记。使经标记样品与杂交混合物 (hybridization cocktail) 混合，在 45℃以 60rpm 杂交至微阵列，并温育 17 小时。杂交后，在将阵列插入到 Affymetrix 自动进样器圆盘传送带中并使用 GeneChip® Scanner 3000 利用GeneChip®操作软件 (GCOS) 进行扫描之前，使用合适的射流轨迹 (fluidics script)，将阵列洗涤并在GeneChip® Fluidics Station 450 上进行染色。

[0167] 除了 RNA 品质控制评估之外，定期对基因芯片 QC 和数据 QC 报告进行分析。根据 Affymetrix 基因芯片品质控制文档的要求和建议，对各个杂交阵列的品质量度，即平均背景、噪音 (Raw Q)、换算因子、呼叫进行 (present call) 的百分比和内部对照基因 (杂交和聚 A 对照) 进行评估以确保各个阵列生成高品质的基因表达数据。使用 Expression Console™ 软件来计算品质评估量度。利用主成份分析 (PCA) 来评估数据品质。生成两份报告来分别总结基因芯片品质控制和数据品质控制的评估结果。在基因芯片品质控制和数据品质控制分析中都未检测到离群芯片。

[0168] 阵列设计。基因芯片人外显子 1.0ST 阵列设计为在外显子水平尽可能包含较大范围，源自范围为从根据经验确定的、经高度恢复 (curated) 的 mRNA 序列到从头算的预测结果的注释。该阵列含有约 540 万个 5-μm 探针，所述探针分组为 140 万个探针集，其询问超过 100 万个外显子基因簇。对于每个外显子，使用一个或数个探针选择区 (PSR)，每个探针选择区都是外显子的连续且不重叠的区段，并且具有不同的长度 (图 1)。PSR 表示被预测为完整连贯的转录行为单元的基因组区域 (组件 HG18、构建块 38)。在许多情况下，每个 PSR 都是外显子；在其它情况下，由于可能存在的重叠性外显子结构，数个 PSR 可以形成真生物外显子的连续而不重叠的子集。选择各个外显子内的 PSR 的位置的关键考量在于它们能够潜在地揭示在所表达剪接变体中使用的选择性剪接位点。为此，在基因的内含子内也使用一些 PSR 以捕获内含子保留。对于各 PSR，通常使用 4 个探针，每个探针的长度为 25 碱基对，其通常是唯一的 (图 1)。约 90％的 PSR 由 4 个探针表示 (“探针集”)。所述冗余允许将鲁棒统计算法用于评估信号的存在、选择性剪接的相关表达和存在。Affymetrix 外显子阵列包括一组 1195 个阳性对照探针集以及 2904 个阴性对照探针集，所述阳性对照探针集代表 100 个通常在大部分组织中高度表达的管家基因的外显子。

[0169] 在各探针和提取自癌组织和参照组织的表达 mRNA 之间进行杂交，各探针附有荧光分子。将各 PSR 的表达水平估计作为置于该区域中的 4 个探针的平均强度。在本研究中，使用由 Affymetrix 推荐的算法 PLIER(Affymetrix，2005) 来进行估计。

[0170] 实施例 3

[0171] 差异性表达的基因的鉴定

[0172] 使用四分位数标准化方法对各外显子的原始探针强度进行标准化，并利用 PLIER 程序 (Affymetrix,2005) 程序来将探针信号总结成外显子水平表达和基因水平表达。除去在癌样品和参照样品中表达非常低的基因，具体而言，如果一个基因的表达水平低于 10（标准化信号强度）则将其除去。为了检测在癌组织中相对于参照组织具有一致性差异性表达模式的基因，如下对表达数据应用简单的统计检验：对于各个基因，对癌组织 / 参照组织对的数目 K_{exp} 进行确定，所述癌组织 / 参照组织对的表达倍数变化大于 k(k 取决于具体问题而设定为 1.25 ～ 4)；如果所观察的 K_{exp} 的 p 值小于 0.05，则认为该基因在大多数癌和参照组织对之间具有差异性表达。同样，使用另外的统计分析，即 ANOVA 检验和 Wilcoxon 符号秩检验，以确保所选择基因在整个癌组织和参照组织对中一致性地具有差异性表达模式。

[0173] 实施例 4

[0174] 基于外显子阵列数据的剪接变体的预测

[0175] 开发了基于所评估的外显子表达水平来预测剪接变体的新算法。该算法依赖于 ECgene 数据库 (Lee 等，2007)，该数据库是最全面的人类转录物的数据库，其含有 181,848 个高可信度的剪接变体和 129,209 中等可信度的变体，所有都源自人类 EST 数据。假定各基因的所有转录物都在 ECgene 中，因此该算法需要确定对于给定阵列数据哪些转录物是最可能的。首先使用 ANOVA 来鉴定在癌组织和参照组织之间所有差异性表达的探针选择区 (PSR) 模式。然后该算法解决了以下优化问题。

[0176] 对于具有 n 个外显子和 m 个已知剪接变体（所有都在 ECgene 中）的给定基因，需要计算 m 个剪接变体的子集和其表达水平，从而使得其总外显子表达水平与所观察到的外显子表达数据尽可能接近。设 I 为 m×n 的二元矩阵，各行表示剪接变体，各列表示外显子，当且仅当变体 i 不含有外显子 j 时 $I_{ij} = 0$。设 $(e_1, e_2, \ldots, e_n)$ 为 n 个外显子的所观察到的表达值。需要计算使以下（二次方程）函数最小的 $\{x_i,\}$ 和 $\{y_i,\}$。

[0177] $$\min \sum_{j=1}^{n} (e_j - \sum_{i=1}^{m} I_{ij} x_i y_i)$$

[0178] 条件为：

$$\begin{cases} \sum_{i=1}^{m} I_{ij} x_i y_i \le e_j, \quad j = 1,\ldots,n \\ x_i = 0,1, \quad i = 1,\ldots,m; \\ y_j > 0, \quad j = 1,\ldots,n. \end{cases}$$

（方程式 1）

[0179] 其中 x_i 是二进制变量，y_i 是实变量。使用以下启发式策略解决该问题。首先假设所有已知剪接变体正用于当前基因，即将所有 $\{x_i\}$ 设定为 1。现在该问题缩为（方程式 1 中 $\{y_i\}$ 变量的）线性规划 (LP) 程序，其可以使用任何现有的用于最佳 $\{y_i\}$ 值的 LP 解算器来解决，所述最佳 $\{y_i\}$ 值是相应转录物的预测表达水平。为了评价该假设的可行性，针对基于所有可能的 2^n-1 剪接变体区间获得的 100,000 个方案测试所观察的 LP 方案。如果统计显著性高 (p 值小于 0.05)，则可认为其是可信的预测方案。否则，这表明 Ecgene 所含的转录物不足以代表某些基因结构，在该情况下对于选择剪接变体需要一套特定标准。该信息可能是外显子 / 内含子长度、外显子存在频率或诸如基序、二级结构等其它类型的特性，

其可以与选择性剪接机制相关并且需要更多的探索。

[0180] 该算法已经作为计算机程序执行，所述计算机程序中使用 Matlib (Dantzig 等，1999) 中提供的 LP 解算器解决各个 LP 问题。该程序使用根据经验确定的截留值来确定一组选定的剪接同种型是否给出了对于所观察到的外显子表达数据而言足够接近的方案。已经在利用根据经验验证的剪接同种型获得的一组外显子阵列数据上对该程序进行了检验 (Xi 等，2008)，其中使用 qRT-PCR 确认了 11 个基因的 17 个剪接同种型。对于这 11 个基因，该方案覆盖了 81.8％的根据经验证实的剪接同种型，表明该程序是高度可信的。

[0181] 使用该计算方法，已经鉴定了总共 2,540 个在所收集的 80 个癌组织和 80 个参照组织之间差异性表达的剪接同种型（包括全长基因）。使用 PCR 和同种型特异性引物（图 1）对数个所预测的剪接同种型进行简单的验证实验。例如，针对 THY1 基因的 3 个所预测的剪接同种型制备同种型 - 特异性引物，以检查所述 3 个所预测的同种型中的任一种是否可以通过相关引物进行检测。如图 1(c) 所示，从 THY1 的表达的剪接同种型的库中鉴定出与所述三种所预测同种型质量相同的剪接变体。

[0182] 在替代性的方法中，对外显子阵列数据应用 MIDAS(Affymetrix，2005) 以检测某基因是否具有选择性剪接变体。基本思路是在对某基因没有选择性剪接的零假设的条件下，该基因中的所有外显子应该具有统计学一致的表达水平。接下来，对于所有样品使用单向 ANOVA 法，以通过检验恒定效果模型 $\log(p_{i,j,k}) = 0$ 来检验所述零假设 ($0 \leqslant P_{i,j,k} \leqslant 1$ 是第 k 基因的第 j 样品的第 i 外显子的成比例的表达）。

[0183] 对以上确定的具有剪接变体的各基因，应用该新算法以及各剪接变体的预测表达水平以预测剪接变体的最可能的集合，所述预测表达水平与从阵列数据观察到的外显子表达水平的一致性最高。具体而言，首先该算法使用 ECgene 数据库 (Lee 等，2007) 中的基因的已知剪接变体以及各变体的最可能的表达水平的估计，来检查所述基因的所观察的外显子表达数据是否能够良好地近似。如果答案为是，然后该算法基于 ECgene 数据库对剪接变体的可能集合作出预测。否则，该算法试图鉴定新剪接变体的最小集合，并结合 ECgene 中的某些已知转录物，给出最简约意义上的对所观察到的外显子表达数据的良好近似。该剪接变体预测问题用公式表示为线性规划 (LP) 问题，并且使用公共 LP 解算器解决 (Dantzig 等，1999)。

[0184] 对于剪接变体的各预测集，使用以下方法来评估其统计学显著性。在不丧失一般性的情况下，假设所有的剪接变体来自 ECgene 数据库。对于由 n 个外显子组成的基因，设 S 是剪接变体的预测集，v 是来自微阵列数据的各外显子的所观察到的表达值和所有预测的剪接变体的累积表达值以及在所有 n 个外显子上的它们的预测表达水平之间的总差异。如下对该预测的剪接变体基以及表达水平的 p 值进行评估。从 ECgene 数据库中的相应基因入口随机选择 |S| 剪接变体，并且对于各剪接变体指定基因表达值，从而其使用与以上相同的步骤从整体上给出所观察到的外显子表达值的最佳拟合。将以上最佳拟合的差记为 v′ 。执行该过程 10,000 次。如果 v 小于 v 值的 95％，则承认预测的 S 是可信的，否则，拒绝该预测。对认为具有剪接变体的各基因使用该方法进行剪接变体预测。然后在所有 80 对组织上对各预测变体的频率计数。如果至少 30％的组织具有该预测变体，则认为该剪接变体是可信的。

[0185] 实施例 5

[0186] 在胃癌组织中相对于参照组织差异性表达的基因

[0187] 收集总共80个胃癌组织和相同数量的邻近胃部但非癌性的组织(参见表2)。使用覆盖17,800个人类基因的Affymetrix基因芯片人外显子1.0ST Array平台对这些组织进行外显子阵列实验。使用一套以上所讨论的标准,发现总共2,540个基因在癌组织和参照组织之间显示差异性表达模式,其中715个显示至少2倍的表达变化,如图(a)所示。基因是指所有其外显子的集合,应该注意的是各个外显子的表达水平不必相同。在癌组织相对于参照组织差异性表达的基因是指癌组织相对于参照组织中的综合基因表达不同的基因。在癌中2,540个基因中的大多数上调,五分之一下调。另外,1,276个基因在早期癌(I期和II期)中差异性表达,其中935个上调,341个下调。1,276个基因中,208个在所有早期胃癌样品中差异性表达,其中186个上调,22个下调,其中48个为胃肠疾病相关的(图2)。

[0188] 1,276个基因中,469个仅在早期癌组织中差异性表达,即在晚期癌组织中不具有实质性差异。此前所提出的标记基因中的大多数在癌中都上调(Takeno等,2008)。与集中在被上调的基因的此前研究相反,在本研究中发现了大量下调基因对胃癌具有高度特异性。这些包括GIF、GNK1、GNK2、TFF1、GHL1、LIPF和ATP4A,提供了癌中丰度减少的不同类型的标记。

[0189] 对通过精细途径分析(Ingenuity Pathways Analysis(IPA))注释定义的2,540个基因的功能家族进行分析。其中,911个基因是癌相关的,219个与抗原呈递或免疫响应相关,414个是胃肠疾病相关的。13个主要的IPA功能家族中,当与全人类基因组相比时,分别发现第9和10家族在(2,540个的)2,094个IPA-注释的基因中显著富集,911个是癌相关的。从图3(a)中可见,诸如蛋白激酶、肽酶、细胞因子、生长因子、跨膜受体和转录调节子等蛋白家族在癌相关基因中是高度富集的,其中酶和转运蛋白在差异性表达的基因中更丰富。从图3(b)中可见,2,540个基因的蛋白产物通常位于细胞质、质膜、细胞外间隙、或细胞核中。类似地在468个仅在早期癌组织中差异性表达的基因中,129个基因是癌相关的,37个与与抗原呈递或免疫响应相关,54个是胃肠疾病相关的。发现3个功能家族在这些基因中显著富集,即酶、转录调节子和转运蛋白。

[0190] 已经将在本研究中发现的差异性表达的基因与之前报道的胃癌相关基因进行比较。通过广泛的文献检索,发现77个基因是胃癌相关的,并且在癌发生和肿瘤进展期间具有显著差异性表达(参见表3)。对于77个基因中的64个(83.1%),在本研究中提出的表达数据与之前的发现一致,包括例如以下基因:TOP2A、CDK4和CKS2(El-Rifai等,2001)、E-钙粘蛋白(Becker等,1994)、GKN1、GKN2和TFF1(Hippo等2002;Moss等,2008)。对于其它13个基因,本研究中提出的数据是新的。例如,发现与染色体扩增、转录调节和信号转导相关的基因(如cyclinE1、POP4、RMP、UQCRFS和DKFZP762D096)在本研究中在80个癌组织中的55个(约68.7%)中具有差异性表达,而在之前研究中126个癌组织中仅约10%具有差异性表达(Chen等,2003)。另一实例是发现在不超过半数的本研究所分析的患者中发现致癌基因JUN(Dar等,2009)的上调和肿瘤抑制基因,TP53的下调(Kim等,2007;Katayama等2004)。这些差异的一个可能原因可能是本研究所用样品相对于之前研究中的患者群体的癌阶段、亚型、年龄和性别的不同分布。

[0191] 表3:通过在胃癌上的转录组学研究和蛋白组学研究获得的生物标记的最新关键发现

[0192]

参考文献	基因(发现)	技术	样品详细信息	类别
Chen 等，2008	TSPAN1、Ki67、CD34	免疫组织化学	86 个癌组织	癌相关基因
Long 等，2008	核因子 κ	免疫组织化学	60 个癌组织	IV 期的基因标记
Yamada 等，2008	PDCD6	微阵列分析	40 个组织+19 个独立	预后基因生物标记
Silva 等，2008	E-钙粘蛋白、β-连环蛋白和粘蛋白 (MUC1、MUC2、MUC5AC 和 MUC6)	微阵列+免疫组织化学	62 个年轻患者+453 个年老患者	基因标记
Xu 等，2009	MUC1 和 MUC5AC	定量夹心酶免疫检验	104 个癌患者和 120 个健康患者	基因标记
Takeno 等，2008	NEK6 和 INHBA	微阵列	222 个癌组织	基因/蛋白水平
Kon 等，2008	胃蛋白酶原 C、胃蛋白酶 A	蛋白组学	来自 24 个癌患者和 29 个良性胃炎患者的胃液	蛋白组学模式
Bernal 等，2008	reprimo	甲基化特异性 PCR	75 个癌组织、43 个癌血浆和 31 个对照	DNA 甲基化模式
Taddei 等，2008	NF2	RT-PCR	5 个胃肠基质瘤	基因标记
Ebert 等，2005	组织蛋白酶 B	蛋白组学	上皮细胞和血清	肿瘤细胞/血清标记
Stefatic 等，2008	CEA、CA19-9、CA15-3、CA125、ecPKA、NNMT	--	--	血清标记综述
Jin 等，2009	MG7-Ag	ELISA	来自 257 个癌患者+50 个正常患者的血清	有用的诊断标记
Ren 等，2006	HSPB1、葡萄糖调节蛋白、PHB、PDIA	SELDI-TOF-MS	来自 46 个癌患者+40 个正常患者的血清	蛋白模式标记

[0193] 还使用 1-、2-、3-、4- 和 5 个基因的组合鉴定了一组“标记”基因，其表达模式在癌组织和参照组织之间能最好地区分。为此，本发明人已经在本团队具有完全权限的计算机集群上使用 R 中的线性辨别分析（并且使用基于线性 SVM 的分类进行验证），通过所述 2,540 基因中的所有 k- 基因组合检索癌组织和参照组织之间的最佳标记。通过使用总体分类精度 P = (TP+TN)/(TP+TN+FP+FN) 对表现进行评价。表 4 给出了针对每个 k 的前几个

k- 基因标记。

[0194] 表 4. 使用 1-、2-、3-、4- 和 5- 基因标记的在癌样品和参照样品之间的分类精度，其中精度定义为“真阳性”和“真阴性”预测与组织总数的比

[0195]

基因标记		精度 (%)
1	TTYH3	80.1
	LIPG	78.7
	MMP1	72.0
2	LIPG-WNT2	83.9
	LIPF- CD276	82.2
	COL10A1- LIPG	80.8
3	AGTRL1-DPT-MMP1	89.7
	TIMP2-DPT-COL10A1	89.1
	DPT-THY1- LIPF	88.4
4	SLC5A5- ANGPTL3-MMP1- DPT	93.1
	COL10A1-LIPG-DTP-HOXB13	92.0
	CLDN1- MMP1- SULT2A1-TRIM	90.6
5	COL10A1-LIPG-DTP-HOXB13-VIL1	95.7
	CLDN1-MMP1- SULT2A1-TRIM29-CDH17	93.7
	CLDN2-DPT- COL10A1-LIPG-DTP-HOXB13	92.7

[0196] 实施例 6

[0197] 年龄和性别对基因表达数据的影响

[0198] 已经通过使用 ANOVA 的多变量分析 (Affymetrix,2005) 和 Cox 比例风险回归模型 (Proportional Hazard Regress Model)(Peduzzi 等,1995) 评估了年龄和性别对 2,540 个差异性表达的基因的影响。关键发现总结如下 (详细内容参见表 5)。据发现年龄显著地影响 2,540 个基因中的 143 个的表达水平,其中大多数 (143 中的 113 个) 进一步增加了在癌组织和参照组织之间其表达水平的差异,这是一个对生物标记选择可能具有重要影响的观察。例如,发现平均 MUC1 表达水平在 55 岁以上的胃癌患者中相对于低于 55 岁的患者显著更高。对于例如 Mucin 家族的其它成员 UBFD1 和 MDK 等数个其它基因类似地观察也成立,而与之相反一些其它潜在标记 (例如 THY1) 不具有年龄依赖性 (图 4)。

[0199] 表 5. 对多种因素因子以及通过 ANOVA 和 Cox 比例风险回归分析 (p 值 < 0.05) 鉴定的其高度相关的基因的统计

[0200]

参数	高度相关的基因	
	基因数量	实例
年龄	143	OLFM4,ABP1, DUOX2, TRIM31, GABRA3, PRSS3, KRT17,GCNT3, LOXL2, TACSTD2
性别	59	SCNN1G, FGA, IL1A, CYP2B6, FAM19A4, WNT2, ARSE, KCNN2, PCSK5, TTLL6,HIST1H2BJ
阶段	27	MT1A, LIF, B3GNT6, HIST1H3J, MT1M
抽烟	113	TRIM29, PI3, FLJ42875, CKS2, DNER, DUOX2, ANGPTL3, HRASLS2, PKM2, DUOXA2, DSG3, APOBEC2
喝酒	63	KIAA1199, DSC3, COL11A1, C1orf125, COL12A1, SULT1C2, LRRC15, SLCO1B3, RPESP, GJB2, ADHFE1, RNF186, ANGPTL3, ADRB2, APOBEC2, MT1L, PTK7, CKMT2
年龄+性别	118	SDS, C1orf125, EGFL6, COL1A1, THY1, REG4, ADH1A, CPS1, SORBS2, GPR68, TIMP1, ADH1C
年龄+阶段	379	ALDH3A1, GSTM5, SORBS2, ADH1A, CDH13, RASL12,GPM6B, PCOLCE2, CAB39L, CASQ2, ACADL, MAMDC2, ZBTB16, C8orf42, MT1A, ADAMTSL3, CNTN1, GPX3

[0201] 还对所提出的表达数据中可能的性别特异性偏向进行了检查，已知胃癌发生的男女比例为约 2 ： 1(Chandanos 和 Lagergen，2008)。据发现诸如 WNT2、ARSE 和 KCNN2 等 59 个基因的表达水平是性别依赖性的（对于全部列表参见表 5)。一个令人感兴趣的观察是年龄和性别的组合对包括 COL1A1、THY1、REG4、ADH1A 和 CPS1 在内的 118 个基因的基因表达水平具有更显著的影响。对于如 TIMP1 和 ADH1A 等基因，老年女性患者比年轻女性患者具有更高的表达水平。还发现，在早期癌所特有的差异性表达的基因中，28 个基因和 9 个基因分别是年龄依赖性和性别依赖性的，其中如 P2RY6 和 NSUN5 等基因同时属于两个组。

[0202] 实施例 7

[0203] 癌组织中的共表达基因和富集途径

[0204] 出于发现具有特定亚型的基因与胃癌的发展阶段的新关联的目的，使用双基因簇分析对基因表达数据进行分析。对于该研究使用双基因簇程序 QUBIC(Li 等，2009)。该算法的基本思路是发现癌组织的某些（待鉴定）子集中具有相似（或相关）表达模式的基因的所有亚群。QUBIC 程序的独特之处在于其检测复杂关系的能力（不仅是仅享有相似的表达模式)，以及对即使含有数以万计基因和数以千计组织样品的数据集也能以非常有效的方式进行检测的能力。该算法在 Li 等，2009 中详细提出。

[0205] 利用双基因簇程序 QUBIC，已经鉴定和分析了 14 个具有统计学显著性的双基因簇，其具有癌特异性、阶段特异性、亚型特异性或性别特异性。首先强调 3 个所鉴定的双基因簇，C1、C2 和 C3。图 5(c) 在所有 80 个癌组织 - 参照组织对中的大多数、特别是在所有的早期癌中的组织对上总结了 C1 和 C2 中的基因及其相关的表达模式。

[0206] 对这两个双基因簇 (C1 和 C2) 进行的详细分析揭示，(a) 诸如转录调节子、生长因子以及参与细胞周期 (STMN 和 CDCA8)、转录调节 (TCF 19 和 BRIP1)、血管发生 (IL8)、染色

体整合 (TOP2A) 和胞外基质重塑 (MMP) 的酶等基因在胃癌的非常早期就被激活 (C1 中),而参与代谢的基因失活 (C2 中);和 (b)C1 和 C2 中的大多数基因甚至在 I 期就显示区分癌组织和参照组织的能力。实例包括在所有早期癌和约 80％的所有癌组织中上调的 HOXB 13、TOP2A、CDC6 和 CLDN7,以及在所有早期癌和 79.1％的所有癌组织中下调的 CHIA。C3 基因中的一些显示出特定癌阶段所特有的不同表达模式。例如,SPP1、SPRP4、COLBA1、INHBA、CTHRC1、COL1A1、THBS2、SULF1 和 COL12A1 在大多数 III 期和 IV 期癌组织中过表达,而在 I 期和 II 期癌组织中未观察到一致的模式 (图 5)。这组基因可以提供潜在的用于为测定胃癌的标记。

[0207] 如图 5(b) 所示,另一经鉴定的双基因簇提供了关于亚型方面的有用信息,图 5(b) 中将 80 名患者分成两个不同组 (左边的绿色部分和右边的红色部分),其与阶段无关。该双基因簇由 42 个基因和 80 名患者组成。42 个基因中的 6 个,即 CNN1、MYH11、LMOD1、MAOB、HSPB8 和 FHL1,之前已经报道在胃癌的肠亚型和扩散亚型之间差异性表达 (Kim 等,2007)。这似乎表明这 42 个基因可以区分胃癌的两种可能亚型。

[0208] 实施例 8

[0209] 途径富集分析

[0210] 还已经检查了差异性表达的基因富集的途径。使用两个程序 DAVID (Dennis 等,2003) 和 KOBAS(Wu 等,2006) 完成给定基因集的途径富集分析。DAVID 基于 GOBiological Processes 和 BIOCARTA 途径计算 EASE 评分 (改良的费歇尔精确 P 值) 以评价相关基因的富集比,而 KOBAS 使用所有 KEGG 途径和 KEGG 直系同源性 (KO) 计算 4 个统计学评分以评估富集途径。除了这些来源之外,将来自 UCSC 癌症途径数据库 (Zhu 等,2009) 的信息整合,所述数据库包括由 NCI-Nature 维护的的人类途径相互作用数据库 (human Pathway Interaction Database)。然后在针对人类基因组中的所有基因的受询基因上基于费歇尔精确检验对各富集途径计算改良 p 值。表 6 列出了 13 条此类途径。

[0211] 表 6 :差异性表达基因利用的 13 条富集途径,↑表示上调,↓表示下调。对于所有阶段中富集的途径计算 P 值,例外的是用 * 标记的 P 值仅用于早期

[0212]

途径	基因数量		P 值
	I-II 期(特异性)	所有阶段	
细胞周期	22↑(9↑)	49↑	1.59E-21
p53 信号传导途径	10↑(3↑)	27↑	2.66E-12
ECM-受体相互作用	4↑(-)	31↑	8.18E-13
细胞通讯	6↑(-)	34↑	4.70E-04
细胞粘附分子(CAM)	4↑(2↑)	31↑	5.13E-04
BRCA1、BRCA2 和 ATR 在癌易感性中的作用	4↑(-)	10↑	2.90E-03
E2F1 破坏途径	4↑(-)	6↑	8.00E-03
Wnt 信号传导途径	4↑(-)	17↑	2.22E-02
局部粘附	4↑(3↑) 3↓(3↓)	41↑ 4↓	1.32E-09 9.81E-02*
通过细胞色素 P450 对外源物的代谢	4↓(-)	16↓	7.21E-04*
精氨酸和脯氨酸代谢	3↓(-)	3↓	1.16E-03*
脂肪酸代谢	3↓(-)	7↓	2.56E-03*
胰岛素信号传导途径	5↓(-)	7↓	9.37E-04*

[0213] 从表 6 可看出，参与细胞增殖、细胞周期和 DNA 复制的基因在大多数癌样品中一致性上调，而参与脂肪酸代谢、消化和离子转运的基因一致性下调。这些途径中的大多数在早期癌中上调 / 下调，并且在晚期癌中高度富集。除了诸如细胞周期和调节、DNA 损坏和修复、细胞生长、死亡和调节以及雌激素受体调节途径等一般性癌相关途径之外，还揭示了一些胃癌特异性过程。例如，新的甲状腺激素介导的胃癌发生信号传导途径在癌组织中与上调基因 (TTHY、PKM2、GRP78、FUMH、ALDOA 和 LDHA) 一起富集 (Liu 等，2009)，所述上调基因中的大多数在晚期。另一令人感兴趣的观察是某些途径仅存在于男性或女性的组织样品中并且在其中更为富集。例如，Ran 在有丝分裂纺锤体调节中的作用、Wnt 信号传导途径和双酚 A 降解在男性但不在女性中富集，而胃促生长素 (Ghrelin)、3- 氯丙烯酸降解、补体旁路途径和组氨酸 / 酪氨酸 / 氮 / 半胱氨酸代谢在女性中更富集。这些发现可以为研究胃癌形成和进展提供新角度。

[0214] 实施例 9

[0215] 在癌组织中相对于参照组织中基因的选择性剪接变体

[0216] 使用特征选择方法来鉴定可以基于随机取样和基因排序一致性的多步评价来区分癌组织和参照组织的多基因标记 (Bell 等，1991)。基本思路如下：使用基于 SVM 的递归特征消除 (RFE) 法来发现基因 (特征) 的最小子集，所述最小子集在随机选择样品的 500 个大小相等的子集上获得 500 个经训练 SVM 的最佳分类表现。如果基因满足以下两条标准则将其消除：(1) 对于本发明的分类，500 个分类器中超过 80％一致性地将其排序为 10％最不重要的基因；和 (2) 它们从未在 (1) 中排序至最重要的 50％之内。继续该基因选择过程直至在不低于分类精度的预定义截止值的同时基因的剩下集合不能进一步缩减。

[0217] 2,540 个差异性表达的基因中，通过如以上实施例 4 中所讨论的新算法将 1,875 个

鉴定为具有选择性剪接变体。基于该预测，在参照组织和癌组织中 1,875 个基因中的分别 69.2％和 72.8％具有实质上的剪接结构改变。1,875 个基因中，预测了总共 11,757 个不同的剪接变体，其中 6,532 个和 6,827 个分别存在于超过 30％的癌组织和参照组织中，将这认为是可信预测。虽然低于该截止值的剪接变体也可能是真的，所述数据可信度较低，更加难以解读。因此，在本研究中不考虑低于该截止值的剪接变体。所述剪接变体中的 6,114 个似乎同时在癌组织和参照组织中出现，其中 3,933 个在胃癌组织中相对于参照组织差异性表达，94 个仅在早期胃癌中差异性表达。已经对在这些预测的剪接变体中所预测的外显子 - 略过事件进行了检查，并且据发现在所预测的选择性剪接变体中略过频率更高的外显子倾向于与具有更多用于剪接调节的顺式调节性基序的内含子区有关，这与如图 6 所示的之前观察 (Wang 等，2008) 一致，为所预测的剪接变体提供了一个支持证据，但需要实质实验来验证所有的剪接变体。

[0218] 对剪接变体进行的所述分析揭示 :(a) 通过将其与 Ensemble 数据库中的已知转录物 (Eyras 等，2004) 进行比较，预测了总共 4,733 个新剪接变体，所述 Ensemble 数据库是最全面的人类剪接变体数据库 ;(b) 具有表达差异性最大的剪接变体的基因是癌相关的，包括 COL11A1、CTSC、CDH11 和 WNT5A ;(c) 不同剪接变体的数量随着癌从 I 期至 IV 期进展而增加 ;和 (d) 发现了分别为女性和男性所特有的 1,690 和 1,377 个剪接变体，其中 364 个和 126 个分别在癌组织中相对于参照组织差异性表达。

[0219] 早期癌特异性剪接变体中，其亲本基因中的 84 个涉及诸如紧密连接、钙信号传导、嘧啶代谢、Wnt 信号传导和上皮细胞信号传导等已知与幽门螺旋杆菌感染相关的途径 (Kanehisa 和 Kegg，2000)。另外，在所有差异性表达的剪接变体中，其亲本基因包括以下途径的成员 :Wnt 途径 (CTNNB 1、WNT2、SFRP4、WISP1、WNT5A)、整联蛋白信号传导 (ITGAX)、p53 信号传导 (E2F1、CDK2、PCNA、TP53、BAX、CDK4) 和胞外基质蛋白 (FN1、COL6A3) 以及诸如 VEGFC、FGFR4、CEACAM6、CDH3、NCAM1、MSH2、VCL 和 ANLN 等其它基因。还注意到 10 个转录因子已经具有表达的剪接变体 (但不是在早期)，即 TFAP2A、NOC2L、MYBL2、MSC、HOXA13、H2AFY、ETV4、E2F4、CCNA1 和 BRD8，其可以充当细胞生长和存活、增殖、分化或凋亡的重要指示物。

[0220] 实施例 10

[0221] 胃癌和阶段的特征基因

[0222] 如以上实施例 9 所讨论，已经鉴定了其表达模式通过使用有效 RFE-SVM 法可以良好地区分癌组织和参照组织的许多基因。图 7(a) 总结了对于所选择最佳 k- 基因标记 (k 为 1 ～ 100) 标记的分类精度。从该图可以看出，28- 基因标记组在所有 k 中是最佳的，分别与癌组织和参照组织具有 95.9％和 97.9％的一致性 (关于其基因名称参见表 7)。

[0223] 基于 RFE-SVM 的方法的设计考虑分类精度、稳定性和可再现性，因此结果具有高度的通用性。对于所有的 k ＜＝ 8，还已经使用线性 SVM 方法 (Vapnik，1995)，通过检查所有的 k- 基因组合对最佳 k- 基因标记组进行了穷举检索，这保证以损失 RPE-SVM 法的计算效率的代价发现全局最优标记。使用留一验证法和 5 倍交叉验证法评价了所鉴定 k- 基因标记的表现。如图 7(a) 所示，如此鉴定的 k- 基因标记 (k ＝ 1，…，8) 的最佳精度始终比通过 RFE-SVM 法得到的最佳精度更好。该分析表明，这些最佳标记基因与以下已知途径相关 :细胞周期、ECM- 受体相互作用、DNA 复制的 CDK 调节以及 TNFR1 信号传导途径 (详细资

料参见表 7)。

[0224] 令人感兴趣的观察是一些标记对于某些患者组表现非常良好,但对诸如不同性别和年龄的其它患者组表现并不好。这与以上实施例 6 中存在的观察一致,即年龄和性别对基因表达水平具有显著的影响。为了解决该问题,已经对不同性别单独进行了标记检索。两个性别组的标记的详细列表在表 7 中给出,表 7 列出了性别特异性最高的标记,包括对于女性的 LIPG、INHBA、MFAP2 和 TTYH3 和对于男性的 WNT2、CD276 和 MFAP2。

[0225] 还对早期癌样品 (I 期和 II 期) 进行了类似分析,并鉴定了早期胃癌所特有的许多有前景的标记。例如,诸如 HOXB9、HIST1H3F、MEM25 和 CLDN3 等基因一致地在所有早期癌组织中显示出差异性表达,但是在晚期癌中未观察到类似的差异性表达。表 7 给出了用于早期癌的最佳 k- 基因标记组以及其分类精度。总之,据发现最佳单基因标记可以获得至多 94.4% 的分类一致性,对于癌组织和参照组织分别为 100% 和 88.9%。当使用最佳 2 基因标记时,该数值提高至 97.3%。

[0226] 为了检查所预测基因标记的通用性,在之前由其它团队公开的胃癌用大型微阵列数据集上对其分类精度进行检查。在 Xin 等,2003 的 GSE2701 数据集上,当 k 为 1 ～ 7 时本研究的 k- 基因标记的成功率为 81.7% ～ 100%。当对来自 Kim 数据集 (Kim 等,2007) 的早期样品进行评价时,诸如 TFF3、CLDN4、MDK 和 MUC13 等本研究的单基因标记在其早期样品的 80% (15 个中的 12 个) 上显示出一致性的差异性表达。总体上这些结果表明所鉴定的组织标记是通用的。

[0227] 已经对所预测基因标记的剪接变体进行了检查,并且已经基于所鉴定的基因标记及其预测的剪接变体 (在癌组织中相对于参照组织过表达或表达不足),预测了作为可能标记的许多剪接变体。虽然详细结果在表 7 中给出,此处列出了数个剪接变体标记 :过表达的剪接变体 LMNB2 :000111111111、WNT2 :11111、WNT :00111、LIPG :1111111110 和 LIPG :1111110000,以及表达不足的剪接变体 AQP4 :111110、GRIA4 :0001111110000000 和 ESRRG :0111110110000000,其中位于第 i- 位的“1”表示剪接变体中基因的第 i 个外显子的存在,“0”表示其不存在。

[0228] 表 7 :为不同类别预测的前 5 个 1-、2-、3- 和 4- 基因标记的最佳检测精度,包括通用标记、早期特异性标记和性别特异性标记。将精度 (Acc.) 测定为 100 次 5 倍交叉验证 (CV) 检测精度的平均值

[0229]

所预测标记的检测精度(5-CV)								
	通用标记	精度	仅早期I-II期	精度	仅男性	精度	仅女性	精度
1	CD276	80.1	HIST1H3F	94.4	WNT2	79.8	LIPG	91.3
	TTYH3	80.1	CCL20	94.4	CD276	78.7	INHBA	86.9
	LIPG	78.7	HIST1H3F	94.4	MFAP2	77.7	MFAP2	86.9
	LMNB2	78.7	C2orf40*	94.4	TTYH3	77.7	TTYH3	86.9

[0230]

	所预测标记的检测精度(5-CV)							
	通用标记	精度	仅早期I-II期	精度	仅男性	精度	仅女性	精度
	WNT2	78.1	HOXB13	88.9	PON2	76.6	RUNX1	86.9
	COL1A1	77.4	CLDN3	88.9	HOXB9	75.5	GPER*	86.9
	PON2	77.4	HOXB9	88.9	CDH3	75.5	GKN1*	86.9
2	CST1- ITGB8	81.5	SCN7A-IKIP	94.4	MYOC-BHLHB2	90.4	INTU-LIPG	97.8
	CST1- AGT	81.5	HIST1H4I-TFCP2L1	94.4	DPT-VASH1	88.3	C16orf53-LIPG	97.8
	MMP1-INHBA	80.8	FAM129A-TREM1	94.4	MAMDC2-MMP2	87.2	Gcom1-GPRIN3	97.8
	MMP1-COL1A1	80.1	MYO1B-MYH11	94.4	CFD-THY1	86.2	CST7-LIPG	95.6
	LIPG-WNT2	83.9	WNT3-NUDCD1	94.4	DGKB-WNT2	86.2	CRABP2-UCKL1	95.6
	LIPF-CD276	82.2	TMEM25-HOXB5	94.4	C2orf40-PLXDC1	85.1	HOXB9-LIPG	95.6
	COL10A1-LIPG	80.8	MMP1-MFAP2	88.9	DPT-COL1A1	85.1	CLDN1-LIPG	95.6
3	AGTRL1-DPT-MMP1	89.7	SCN7A-IKIP-HIST1H3F	94.4	CD44-DPT-AGTRL1	93.6	GIF*-PID1-LRRIQ1	**100**
	TIMP2-DPT-COL10A1	89.1	SCN7A-IKIP-C2orf40	94.4	GGTLA1-DPT-NID1	92.5	FCGR3A-C16orf53- LIPG	100
	DPT-THY1-LIPF	88.4	HIST1H4I-TFCP2L1	94.4	LOC202051-CGNL1-THY1	92.5	SLC15A3-PAICS-FAM123A	100
	THBS2-DPT-C19orf40	88.4	SCN7A-IKIP-RYR2	88.9	FRMD1-MAMDC2-RASAL2	92.5	SLC15A3-LIPG-TPD52	97.8
	TIMP2- DPT-CLIC1	88.4	SCN7A-IKIP-C2orf40	88.9	HOXB9-RYR2-CD109	91.5	SLC15A3-LIPG-SPON2	95.7

[0231]

	所预测标记的检测精度(5-CV)							
	通用标记	精度	仅早期I-II期	精度	仅男性	精度	仅女性	精度
	MYOC-CD44-HIST2H2AB	88.4	SCN7A-IKIP-CCL20	88.9	PDZRN4-INHBA-AGTRL1	91.5	SLC15A3-MYOC-CD3EAP	95.7
4	CXorf36-DPT-CD44-BST2	94.5	GAL3ST4-PPA1-HOXA13-HIST1H3F	**94.4**	RYR2-HMCN1-HOXB9-MT1M	95.7	EPDR1-GIF*-TEAD4-OR1L1	100
	PDGFRB-MYOC-HFM1-PGRMC2	93.8	-	-	TGM2-PARK2-RASGRF2-PI16	95.7	KIAA1199-DUSP10-LYCAT-ADHFE1	100
	SLC5A5-ANGPTL3-MMP1-DPT	93.1	-	-	MEX3D-DPT-C10orf72-C10orf129	95.7	FCGR3A-PGRMC2-GLIS3-TMEM40	100
	COL10A1-LIPG-DTP-HOXB13	92.0	-	-	NR0B2-BTG2-CTSA-DBT	95.7	CKMT2-CCL18-MICALL1-LRRIQ1	100
	CLDN1-MMP1-SULT2A1-TRIM	90.6	-	-	IRX3-ADCYAP1R1-FADS2-RUNX1	95.7	PTGIR-GAL3ST4-PTPRS-XAF1	100

[0232] （用＊标记的基因是在癌中相对于参照下调的基因；“-”：如果具有较小 k 值的组合标记已经对本发明的样品具有 100％或不发生变化的最佳检测精度，则此处省略 k- 基因标记）

[0233] 实施例 11

[0234] 用于预测血液分泌蛋白的计算方法的开发

[0235] 已经为了预测能够被分泌到循环中的人类蛋白开发了计算技术 (Cui 等，2008)。该方法的基本思路是收集已知血液分泌蛋白的集合和与已经在人类血清中检测到的任何蛋白不具有同源性的蛋白的集合。然后训练分类器以区别这两个集合。已经对从蛋白序列可计算的大量特征进行检查，并且已经鉴定了能够在所述两个集合之间提供最高辨别力的特征。

[0236] 用于收集训练数据的起点是含有约 16,000 个由血浆蛋白组项目 (PPP) (Omenn 等，2005) 汇集的在人类血清中已检测出的蛋白。还从 Swissprot 和 SPD 数据库 (Chen 等，2005) 收集了 1,620 个人类分泌蛋白。通过将该列表与 PPP 比较，发现了属于两个集合的 305 个蛋白不在天然血液蛋白之内。因此，认为这 305 个蛋白被分泌到血液中，并且用作阳性集。然后从不与 PPP 重叠的 Pfam 各家族 (Bateman 等，2002) 中选择代表，并且收集了 26,962 个蛋白作为阴性集。然后将阳性集和阴性集分成训练集和测试集。

[0237] 为了发现可以区分所述两个集合的特征，对 50 个特征进行检查，这 50 个特征大致落入 4 个类别 :(i) 诸如氨基酸组成和二肽组成等一般性序列特征 (Reczko 等，1994 ; Bhasin 等，2004) ;(ii) 诸如溶解度、非稳定区和电荷等物化特征 ;(iii) 诸如二级结构含量和溶剂可及性等结构特征 ;和 (iv) 诸如信号肽、跨膜区和双精氨酸信号肽基序 (TAT) 等特异性结构域 / 基序。

[0238] 使用这些特征，对基于支持向量机 (SVM) 的分类器进行训练以使用高斯核区分从阴性训练数据区分阳性训练数据 (Platt 等，1999 ;Keerthi 等，2001)。基于起始 SVM 的性能，使用被称为递归特征消除 (RFE) 的特征选择方法来除去与分类目标无关或可忽略的特征。基于一致性评分方案和基因排序一致性评价 (Tang 等，2007)，该特征选择方法反复地除去无关特征。具体而言，在各次重复中，从特征列表消除由 RFE 给出的具有最低评分 (排序最低) 的特征。继续该方法直到在维持分类表现的水平的同时获得特征的最小集。整个训练中，一直使用随机取样 (Bell 等，1991) 来生成训练集和测试集，并且基于给定的训练集和测试集对分类器进行训练。该方法执行 500 次，并挑选出最具代表性的集合 (Cui 等，2008) 作为选定集合。经过该过程，发现对于分类而言最重要的特征包括跨膜区、电荷、TatP 基序、溶解度、信号肽和 O- 连接的糖基化基序。

[0239] 基于所选择的特征，保留了基于 SVM 的分类器并对其进行交叉验证，在独立评价集上测试了其表现，其可以正确地分类 90％的血液分泌蛋白和 98％的非血液分泌蛋白。使用 7 个额外数据集来进一步评估该分类器的表现，每个数据集含有最新鉴定的血液分泌蛋白和文献中报道的蛋白。测试结果给出了与对所述评价集进行的相当的表现统计。例如，通过广泛地文献检索将通过质谱获得的人类血清中检测的 122 个蛋白的列表汇集。这些蛋白在 14 种人类癌中的至少一种中过表达，并且它们都不包括在本发明的训练集中。使用上述方法正确地预测了 122 个蛋白中的 97 个 (79.5％)。

[0240] 实施例 12

[0241] 血液分泌蛋白的预测

[0242] 在所有差异性表达的基因中，集中于能够被分泌到血流中作为可能的血清标记的那些基因。已经为所述分泌蛋白的预测开发了计算方法 (Cui 等，2008)。该实施例描述了用于预测蛋白向血清的分泌的方法。但是，基于本文存在的教导和指导，应该理解，本领域已知可以容易地采取本文所述方法来预测蛋白向其它生物液的分泌，所述其它生物液例如但不限于唾液、脊髓液、精液、阴道液、羊膜液、龈沟液和眼内液。

[0243] 已经基于所鉴定的其在癌组织中的差异性表达和血液分泌预测而预测了胃癌的许多血清蛋白标记 (Cui 等，2008)。将这些预测的血清标记分成 3 类 :(a) 胃癌的通用标记，(b) 对早期癌具有特异性的标记，和 (c) 性别特异性标记。表 8 显示了被认为单独或组合成组时最有前景的蛋白。表 9 中给出了关于这些和其它有前景的标记蛋白的详细信息。

[0244] 这些预测的血清标记中，MMP1、MUC13 和 CTSB 是有效的区分癌组织和参照组织的基因区分物，但是由于它们在诸如乳癌、卵巢癌、肺癌和结肠癌等其它癌中的过表达 (Poola 等，2008)，它们对胃癌不具有特异性。然而，LIPF、GAST、GIF、GHRL 和 GKN2 具有胃组织特异性，因此使得它们成为有前景的用于胃癌的血清标记，特别是当与其它标记结合使用时。

[0245] 表 8 :用于胃癌的最有前景的预测标记的实例

[0246]

血清标记		阶段功效		性别特异性	
		通用	早期	女性	男性
MMP1	基质金属蛋白酶 1 前蛋白	√			
MUC13	粘蛋白-13	√			
CTSB	组织蛋白酶 B	√		√	
GKN2	胃动蛋白-2		√	√	
GHRL	食欲调节激素(胃促生长素)		√		
LIPF	胃三酰甘油脂肪酶（胃脂肪酶）		√	√	
LIPG	内皮脂肪酶	√		√	
LIMK1	LIM 结构域激酶 1		√	†	†
GAST	胃泌激素		√		
GIF	胃内因子	√			
AZGP1	锌-α-2-糖蛋白	√			

[0247] （†表示基因具有良好的分类精度但非性别依赖性）

[0248] 表 9 :18 个预测标记以及其功能注释、在癌中的表达特异性和相关疾病的详细信息

[0249]

基因符号	蛋白[AC]	质量(kDa)	FC	亚细胞定位&血液中的存在(注释*/本发明预测)	AS	有报道的癌中表达(相对于正常)	相关疾病
MMP1	基质金属蛋白酶 1 前原蛋白 [Q53G97]	44.8	7	胞外间隙 &(1/1)	√	乳癌；结肠癌；舌癌；头颈癌中中度过表达；肺癌；膀胱癌	癌、心血管疾病、肝系统疾病、炎性疾病、神经疾病
COL10A1	胶原蛋白 α-1(X)链 [Q03692]	6.2	3	分泌的；胞外基质 &(1/1)		结肠癌；乳癌	结缔组织障碍、皮肤性疾病、炎性疾病、骨骼和肌肉障碍
CLDN1	封闭蛋白-1	22.7	4	质膜 &(0/1)	√	精原细胞瘤和卵巢癌中中度过表达	癌、皮肤疾病和病况、胃肠疾病
TOP2A	DNA 拓扑异构酶 2-α EC=5.99.1.3 [P11388]	174.4	3	细胞质；细胞核 &(1/0)	√	膀胱癌；脑癌；肝癌	抗原呈递、癌、皮肤疾病和病况、胃肠疾病
CST1	胱抑素-SN 前体 [P01037]	16.4	12	分泌的 &(0/1)		在膀胱癌中中度过表达；头颈癌；精原细胞瘤	癌、神经疾病
COL1A1	胶原蛋白 α-1(I)链 [P02452]	138.9	3	胞外间隙 &(1/1)	√	精原细胞瘤；在脑癌中中度过表达；头颈癌；胃癌	抗原呈递、耳科疾病、癌、心血管疾病、结缔组织障碍、肝系统疾病、炎性响应
MUC13	粘蛋白-13 [Q9H3R2]	54.6	2	分泌的 &(1/1)		在上皮癌组织，特别是胃肠道和呼吸道的上皮癌组织中高度表达	癌、胃肠疾病
CTSB	组织蛋白酶 B [P07858]	37.8	1.8	溶酶体 &(1/1)	√	在子宫颈癌、子宫内膜癌、肝黑素瘤和胰腺癌中高度表达	癌、心血管疾病、结缔组织障碍、皮肤疾病、内分泌系统障碍、胃肠疾病、血液疾病、肝系统疾病、传染病、炎性响应、神经疾病、肾和泌尿系疾病、呼吸道疾病、骨骼和肌肉障碍
GKN2	胃动蛋白-1 [Q86XP6]	22.0	3	分泌的 &(0/1)	√	在脑癌中略微上调，在肺癌中略微下调	胃癌、克罗恩氏病
GHRL	食欲调节激素(胃促生长素) [Q9UBU3]	12.9	9	分泌的 &(0/1)	√	在结直肠癌、肝癌和胰腺癌中中度过表达	抗原呈递、癌、心血管疾病、肝系统疾病、炎性疾病、炎性响应、神经疾病、营养性疾病、机体损伤和异常、心理性障碍、生殖系统疾病、骨骼和肌肉障碍

[0250]

LIPF	胃三酰甘油酯脂肪酶(胃脂肪酶) [P07098]	45.2	5	分泌的&(0/1)	√	在卵巢癌中略微上调，在乳癌中下调	心血管疾病、内分泌系统障碍、代谢性疾病、营养性疾病、呼吸性疾病
LIPG	内皮脂肪酶 [Q9Y5X9]	56.8	3	分泌的&(1/1)	√	在脑癌、卵巢癌和头颈中略上调，在白血病中下调	抗原呈递、心血管疾病、炎性响应
LIMK1	LIM 域激酶 1 [P53667]	72.6	1.8	细胞质&(0/1)	√	在淋巴瘤癌和黑素瘤中中度上调	癌、心血管疾病、皮肤性疾病、发育障碍、内分泌系统障碍、遗传性疾病、血液疾病、神经疾病、生殖系统疾病
GAST	胃泌激素 [P01350]	11.4	1.1	分泌的&(0/1)		在胃癌中表达	癌、克罗恩氏病、卓-艾综合征
TIP47 (M6PRBP1)	甘露糖-6-磷酸受体结合蛋白 1 [O60664]	47.0	1.3	细胞质、内体膜&(1/1)		乳癌、子宫颈癌、结直肠癌、胰腺恶性癌、肾癌、睾丸癌、胃癌和和恶性神经胶质瘤	子宫颈发育异常、癌
PDGFRB	β-型血小板衍生生长因子受体 [P09619]	124.0	2	膜&(1/1)	√	恶性神经胶质瘤、在卵巢癌中中等	癌、心血管疾病、皮肤性疾病、内分泌系统障碍、胃肠疾病、血液疾病、肝系统疾病、免疫性疾病、炎性疾病、神经疾病、眼科疾病、肾和泌尿疾病、生殖系统疾病、呼吸性疾病、骨骼和肌肉障碍
GIF	胃内因子 [P27352]	45.4	12	分泌的&(0/1)	√	在大多数癌组织中上调，但在平滑肌肉瘤中中度下调	遗传障碍、血液疾病、代谢性疾病
AZGP1	锌-α-2 糖蛋白 [P25311]	33.9	3	分泌的&(1/1)	√	在前列腺癌和乳癌中高度表达	炎性疾病、呼吸性疾病

[0251] (FC :倍数变化 ;注释 * 是基于 IPA 注释 ;AS :检测到选择性剪接变体。癌表达信息获自 Oncomine 网站和 Proteinatlas 网站检索)

[0252] 实施例 13

[0253] 所预测血清标记的实验验证

[0254] 使用质谱和蛋白质印迹分析的组合方法来验证所预测的血清蛋白标记。使用抗体柱 (来自 Beckman Coulter 的 ProteomeLab™ IgY-12 高容量蛋白组配分试剂盒) 对血清样品进行加工以除去 12 种最丰富的蛋白 (白蛋白、IgG、α 1- 抗胰蛋白酶、IgA、IgM、转铁蛋白、结合珠蛋白、α 1- 酸糖蛋白、α 2- 巨球蛋白、HDL(载脂蛋白 A-1&A-II) 和纤维蛋白原)。这 12 种高丰度的蛋白的特异性去除从人类血清或血浆中除去了 96％的总蛋白质量。所预测的生物标记存在于剩下的 4％总蛋白质量中，因此易于作为分离步骤的结果而鉴定。

[0255] 免疫捕获12种最丰富的血清蛋白后，从所述柱洗脱和收集非特异性结合蛋白。还从所述柱洗脱特异性结合蛋白以用于进一步分析，以检查它们是否充当潜在的生物标记的载体。

[0256] 对于蛋白（印迹）分析，在100℃温育蛋白样品5分钟，通过4％～20％的梯度聚丙烯酰胺凝胶(Bio-Rad)利用SDS-PAGE将其分离，然后转移到PVDF膜上。用3％在TBST中的脱脂奶粉(10mM Tris HCl、pH 7.5、150mM NaCl、0.05％聚氧乙烯山梨糖醇单月桂酸酯(Tween-20)[重量/体积])于室温封闭非特异性结合位点后，使膜与一抗一起在4℃于1.5％的TBST中的脱脂奶粉中温育过夜。用TBST洗涤3次后，在室温在含有二抗的1.5％的TBST中的脱脂奶粉中使所述膜温育2小时。然后使用增强型蛋白印迹放电化学发光试剂(Perkin Elmer，USA)使膜进行增强化学发光反应。使用MagicMark蛋白印迹蛋白标准物(Invitrogen，Karlsruhe，德国）来鉴定分子量。使用ImageJ 1.34软件（可从NIH网址上获得）的凝胶分析(Gel Analysis)功能就蛋白浓度的定量评价ECL膜图像。所述抗体来自Abnova，Inc.（台北，台湾），Santa CruzBiotechnology，Inc.(Santa Cruz，CA)和Abeam，Inc.(Cambridge，MA)。在抗体选择中使用所预测的剪接变体。如果最丰富的剪接同种型过短而不能覆盖任何抗原性区（表位），通过特别设计用于全长蛋白的抗体可能不会检测到标记。因此，基于所预测的剪接变体的分析，选择其表位区被大多数转录物覆盖的那些抗体。

[0257] 对通过两种不同方法从所述凝胶提取的蛋白进行MS实验。用测序级改良胰蛋白酶消化之后，使用Agilent 1100系列HPLC对蛋白样品进行在线HPLC分析，所述Agilent 1100系列HPLC具有直接偶联到配备有Apollo II纳米电喷雾源的9.4T BrukerApex IV QeFTMS(Billerica，MA)上的75μm C-18反相柱。碰撞激活解离(CAD)用于离子解离，并且使用氩作为碰撞气体完成蛋白片段化，然后将其注射到ICR分析仪小室。对于蛋白鉴定使用在Protein Prospector网站上的Bruker数据分析软件和MS-Tag程序实现数据分析。同时，用蛋白组学级胰蛋白酶(Promega)将同一样品消化，并在与LTQ线性离子阱质谱仪(Thermo Electron，San Jose，CA)直接连接的Agilent1100毛细管LC(Pal Alto，CA)上进行分析。通过充满5-μm直径的C18珠的50-μm柱(New Objective，Woburn，MA)对PicoFrit 8-cm施加N2正压来将肽样品上样。以200nL/分钟的流速在55分钟的线性梯度期间将肽从所述柱洗脱到质谱仪中，所述线性梯度为从5％至60％的由流动相B组成的总溶液。将仪器设定为在9个来自各MS的最丰富的前体离子上采集MS/MS谱，重复数为3，重复持续时间15秒。使动态排除进行20秒，并通过Mascot（参见matrixscience网站）进行数据分析（图8）。

[0258] 验证集由来自9名胃癌患者(4名早期癌，5名晚期癌）和5名年龄和性别匹配的对照成。该验证集包括除汇集用于质谱分析的样品之外的若干额外样品，其作为独立的评价集。基于本发明的计算预测选择了20个最有前景的候选物以用于蛋白质印迹分析，其中4个通过上述MS分析检测。在血清样品中发现这些蛋白中的15种，包括通过基于MS分析检测的2种(TOP2A和AZGP1)。其中，如图9所示，7种(GKN2、MUC13、LIPF、GIF、AZGP1、CTSB和COL10A1)在癌患者的血清和对照样品之间显示出某种程度的差异性丰度。

[0259] 从图9中可以看出，存在两种潜在的标记：(1)在晚期癌中丰度增加/减少的蛋白。例如，在晚期癌血清中显示丰度增加的粘蛋白-13，其是覆盖气管和胃肠道的顶部表

面的糖蛋白，在数种影响癌发生、运动性和细胞形态的信号传导途径中起作用。其可用作通常的癌标记，但是对于早期癌检测可能不太有效。胃脂肪酶 (LIPF) 和 DNA 拓扑异构酶 2-α (TOP2A) 在晚期癌血清中也差异性表达，其表达分别减少和增加。(2) 在早期癌中具有差异性表达的蛋白，即 GKN2、COL10A1 和 AZTP1。在癌血清中表达减少的 GKN2 对于检测早期癌是有效的，因为在本发明测试中一半早期样品的丰度改变，包括一个 I 期癌。

[0260] 在这些有前景的标记中，已经提出 CTSB 作为潜在的胃癌标记 (Ebert 等，2005；Poon 等，2006)，其显示出差异性丰度，但在本发明的样品上不一致；之前已经提出 MMP1 和 TOP2A 通常是癌相关的 (Poola，2005)；这得到本文提出的数据支持。GKN2 和 LIPF 是胃组织特异性的；COL10A1 和 GAST 通常可与其它疾病或免疫响应相关。

[0261] 这些个体蛋白的组合也被认为是潜在的组合标记。虽然由于缺乏这些蛋白的精确量测定而使组合标记的详细定量评估较为困难，但已基于来自蛋白质印迹数据的所评估蛋白丰度对分类精度进行了粗略评价。如表 4 所示，列出了 k- 蛋白标记的集合，其比个体血清标记给出了明显提高的分类精度。表 10 给出了 k- 蛋白血清标记的详细列表。

[0262] 表 10：经验证的 k- 蛋白标记的血清精度，基于 5 倍交叉验证精度在基因水平和蛋白水平上对所述经验证的 k- 蛋白标记进行了验证。

[0263]

k	标记	检测精度	
		蛋白-水平	基因-水平
1	GIF	0.867	0.726
	GKN2	0.80	0.705
	MUC13	0.667	0.613
2	GIF+LIPF	0.933	0.746
	GIF+COL10A1	0.867	0.732
	GIF+TOP2A	0.80	0.732
3	GIF+LIPF+MUC13	0.933	0.733
	LIPF+GIF+AZGP1	0.867	0.719
	COL10A1+GKN2+GIF	0.80	0.753
4	LIPF+GIF+MUC13+AZGP1	0.933	0.767
	LIPF+GIF+MUC13+COL10A1	0.933	0.788
	LIPF+GIF+MUC13+GKN2	0.80	0.740

[0264] 应该注意的是某些因素可能影响蛋白质印迹结果。例如，一个此类因素是不同的剪接同种型可以不必具有针对每种相关蛋白的全长常见形式设计的抗体类似的结合亲和力。基于所提出的预测，诸如 MMP1、LIPG、LIPF 和 CTSB 等标记都具有剪接变体。因此，基于所选择的剪接变体选择合适的抗体。

[0265] 实施例 14

[0266] 尿中癌标记的鉴定

[0267] 训练数据和测试数据的收集。将由主要的尿蛋白组学研究 (Adachi 等，2006) 鉴定的 1500 个蛋白的集合用作阳性训练数据。在利用 SwissProt 登录 ID 的该蛋白组学研究中鉴定了总共 1,313 个人类蛋白，并包括在该训练集中。对于独立的测试集，使用来自三个其它主要尿蛋白组学研究 (Pieper 等，2004；Castagna 等，2005；Wang 等，2006) 的数据，包括不与训练集重叠的总共 460 个人类蛋白。

[0268] 对于阴性训练集和测试数据集，在进行 Cui 等，2008 中所述的选择步骤后，从不与阳性数据重叠的 Pfam 家族中选择蛋白，以确保所选择的蛋白遵循相同的家族 - 大小分布

(Finn 等,2008)。结果,对于训练集和测试集分别选择了 2,627 和 2,148 个蛋白,所述训练集和测试集之间无任何重叠。

[0269] 特征计算和选择。对于从 SwissProt 数据库检索的各蛋白序列,对 18 个特征进行计算。这些特征中的一些需要多个特征值来表示它们,例如,需要 20 个特征值来表示蛋白序列中的氨基酸组成;因此使用 243 个特征值表示 18 个特征。表 11 列出了该 18 个特征以及用于表示它们中每一个的特征值的数值。使用内部程序或如果可在互联网上获得则使用预测服务器对 18 个特征进行计算。

[0270] 基于可获得的关于尿分泌的信息进行选择,该特征列表可潜在地用于区分尿分泌的蛋白和非尿分泌的蛋白。为了检查它们中哪些是确实有用的,使用支持向量机用文库(LIBSVM)中提供的特征选择工具来选择 243 个特征值中有用的特征。LIBSVM 是用于支持向量分类(C-SVC,nu-SVC)、回归(ε-SVR,nu-SVR)和分布估算(一类 SVM)的积分软件。该特征选择工具计算 F 评分(Chang&Lin 2001)来测定本发明的分类问题的各特征值的相关性的排序。除去所有 F 评分低于预选阈值的特征,认为剩下的特征对于分类问题有用。

[0271] 表 11:用于起始分类模型的总结

[0272]

特征种类	特征名称和特征值	用于计算特征的程序
序列特征	序列长度(1)、AA 组成(20)	Fldbin (Prilusky 等, 2005)、Profeat (Li 等, 2006)
物化性质	疏水性(21)、标准化范德华体积(21)、极性(21)、极化率(21)、电荷(21)、二级结构(21)、溶剂可及性(21)、伪氨基酸描述符(50)	本地计算、Profeat (Li 等, 2006):使用三个描述符:组成、转移和分布
	不可折叠性(1)、电荷(1)、疏水性(1)、非稳定区的数目、最长非稳定区(1)、非稳定残基的数目(1)、PI(1)、MW(1)、电荷(2)、非稳定区的百分比(1)	Fldbin (Prilusky 等, 2005)、Swiss (Gasteiger 等, 2003),本地计算
基序	跨膜域(1)、双精氨酸信号肽(1)、跨膜域(α 螺旋或 β 桶)(2)、糖基化数量&存在(N&O 连接的)(4)	TMB-Hunt (Bendtsen 等, 2005; Garrow 等, 2005)、TatP (Bendtsen 等, 2005)、phobius (Kall 等, 2007)、NetOgly (Julenius 等, 2005)、NetNGly (Gupta 等, 2004)
结构选项 2.243	二级结构含量(4)、回转半径(1)、半径(1)、	SSCP(Eisenhaber 等, 1995)、回转半径,本地计算

[0273] 使用 DAVID 生物信息学资源网络服务器来完成对所有所预测的尿分泌蛋白进行的功能性富集分析。使用人类蛋白作为背景进行功能注释基因簇分析。对于各个基因簇通过 EASE 评分确定总富集评分(Dennis 等,2003;Huang 等,2009)。

[0274] 使用 KOBAS 网络服务器(Mao 等,2005;Wu 等,2006)来计算所预测尿分泌蛋白中的统计学上富集的和代表性不足(underrepresented)的途径。KOBAS 读取序列集合并且基于 BLAST 序列相似性对 KEGG 直系同源术语(orthology term)进行注释。然后针对所有人类

蛋白比较经注释的 KO 术语。如果在百分比组成方面存在至少 2 倍的变化则认为途径是富集的或代表性不足的。

[0275] 在中国长春吉林大学医学院收集来自 10 名处于转移期的胃癌患者 (7 名男性,3 名女性) 如 10 名性别匹配的健康人的尿样品。立即将这些样品冻干并在准备使用前贮存。使这些样品复原并在 4℃于 3,000 相对离心力下旋转 25 分钟,以除去细胞成分。收集上清液并将其冷冻在 -80℃直到进一步使用。然后使用 Slide-A-Lyzer 透析盒 (Thermo Fisher Scientific,Rockford,IL) 针对 Millipore 超纯水 (更换三次缓冲液,然后进行过夜透析) 在 4℃对所述样品进行透析。使用 Bio-Rad 蛋白测定 (Bio-Rad,Hercules,CA) 利用牛血清白蛋白作为标准品测定蛋白浓度。

[0276] 信号肽和二级结构是尿分泌蛋白的关键特征。使用基于 F 评分的特征选择,当特征值数值为 74 时观察到最高精度。使用这 74 个特征值,对基于 SVM 的分类器进行再训练。所选择的特征中,对于分泌蛋白的最有辨别力的特征是信号肽的存在。已知通过 ER 分泌的蛋白具有信号肽,并且根据特定的信号肽被运送到其目的地 ;因此大多数分泌蛋白具有该特征。另一突出的特征是二级结构的类型 ;数个与二级结构有关的特征值包括在前 74 个最佳特征中,并且 α 螺旋的百分比排在 74 个中的第 2 位。

[0277] 对于分泌蛋白,蛋白的电荷在排在前几名的特征中。这与电荷实际上是确定哪些蛋白过滤透过肾中的肾小球膜的因素的通常理解一致。但是,发现排在第 232 位的蛋白的分子大小对于所述分类问题是无关的。

[0278] 如表 12 所示,对两个分类器进行训练。模型 1 的特异性较高但敏感性较低,而模型 2 展示出更平衡的表现。由于阳性训练数据和阴性训练数据的不平衡数量,精度可能不是确定模型的性能的最佳度量。因此,使用马修相关系数作为分类品质的度量。

[0279] 表 12 :训练时所训练模型的表现

[0280]

集合	模型	TP	TN	FP	FN	SEN	SP	ACC	MCC
训练	1	792	2493	134	341	0.7403	0.9490	0.8794	0.5228
训练	2	1164	2230	297	149	0.8865	0.8869	0.8868	0.5697
独立	1	360	1983	165	100	0.7826	0.9232	0.8984	0.4500
独立	2	404	1838	310	56	0.87820	0.85567	0.85966	0.39358

[0281] 在预测置信度和蛋白距分离超平面的距离之间存在直接相关性,所述分离超平面存在于由基于 SVM 训练导出的阳性训练数据和阴性训练数据之间。具体而言,离分离超平面的距离越远,正确预测的可能性越高 (图 10)。使用置信区间作为指导,可以选择少量蛋白用于实验验证。

[0282] 将经训练分类模型应用至胃癌数据。在致力于鉴定尿中的用于胃癌的潜在生物标记时,在 Affymetrix 人外显子测定 1.0(Cui 等,2009) 上将本文开发的经训练模型应用于 2048 个差异性表达基因的集合,所述差异性表达基因基于来自相同的 80 名患者的 80 个

胃癌组织及 80 个匹配的非癌性胃组织上的 160 个外显子阵列而鉴定出。在所述 2,048 个蛋白中，预测 480 个通过模型 1 被分泌到尿中，这 480 个蛋白中，11 个蛋白的置信水平高于 98％，表明它们非常有可能被分泌到尿中。480 个蛋白中的总共 203 个蛋白具有至少 92％的置信水平，这也被认为是高度可信的预测。

[0283] 对所有 480 个蛋白进行功能和途径富集分析以帮助确定哪些类型的蛋白可以在尿中发现。具体而言，如果分析表明某具体的功能组或途径被富集，则在该组中发现生物标记的机会增加。分别使用 DAVID(Dennis 等，2003) 和 KOBAS(Wu 等，2006) 网络服务器，利用完整的人类蛋白作为背景对功能和途径富集分析进行分析。

[0284] 通过 DAVID 进行的功能富集分析揭示，480 个蛋白中的大多数富集的功能组涉及胞外基质 (ECM)。ECM 在癌进展中通过影响细胞增殖和移动性起重要作用。细胞表面受体与 ECM 中的配体之间的相互作用不仅影响细胞脱附和移动，而且 ECM 还充当细胞可以在其上粘附和生长的模板 (Ashkenas 等，1996 ；McKinnell 等，2006)。ECM 分子的组成、细胞类型和细胞表面受体组成可以通过经由整联素发送信号而促进或抑制细胞增殖 (Stein&Pardee 2004)。因此，涉及 ECM 的蛋白不仅对于胃癌，而且对于所有其它类型的癌也是重要的尿生物标记。总之，480 个蛋白中的 164 个在该组中。

[0285] 下一最重要的富集组是涉及细胞粘附的蛋白。众所周知，细胞粘附是有助于癌生长的因素。例如，细胞彼此之间粘附或粘附到 ECM 上，但是当肿瘤形成时，细胞必须从原发瘤脱离，并且入侵淋巴系统以进行转移。因此，癌细胞不表达诸如 E- 钙粘蛋白等细胞粘附分子，并且失去其特征性形态以及变得具有入侵性 (Frixen 等，1991)。所鉴定的 480 个蛋白中，93 个位于该组，因此为发现尿中的细胞粘附生物标记提供了谨慎的优化。其它富集功能组包括涉及发育、细胞移动、防御性 / 炎症性响应和血管发育 / 血管发生的蛋白。图 11 显示了功能富集分析的综合结果。

[0286] 对 480 个蛋白进行的途径富集分析揭示，某些途径与背景 (全人类集合) 相比是统计学上富集的 (图 12) 或代表性不足的 (图 13)。480 个蛋白中，超过 20％涉及细胞抗原途径，其可以通过免疫系统响应于癌形成和发育而触发。免疫系统在癌发育中的作用尚不明确，很大程度上地因为其对癌发育和进展具有自相矛盾的作用。例如，抗肿瘤适应性免疫响应的激活可以抑制肿瘤生长和发育，而浸润的淋巴细胞的丰度与更有利的预后有关，浸润的先天免疫细胞的丰度增加与血管发生和不良的预后有关 (de Visser 等，2006)。

[0287] 由于蛋白容易进入血流，蛋白在抗原途径中的富集并不令人惊讶。而在血液循环中，所述蛋白与胞内蛋白不同，它们可以容易地过滤通过肾小球。这表明存在留待发现的更多的抗原癌标记。根据肽酶、细胞粘附分子和 CAM 配体在癌进展中的作用来预期，肽酶、细胞粘附分子和 CAM 配体在该途径分析中被过度代表 (overrepresented)。

[0288] 大多数代表性不足的蛋白是胞内蛋白 (图 3)。例如，在 480 个蛋白中蛋白激酶途径明显代表性不足。蛋白激酶涉及诸如离子转运、细胞增殖、激素响应、细胞凋亡、代谢、转录和细胞骨架重组以及细胞移动等胞内过程 (Malumbres&Barbacid，2007)。激酶活性的失调经常导致肿瘤生长。例如，有证据表明许多激酶突变是促进癌发育的“驱动”突变 (Greenman 等，2009) ；此外，突变蛋白激酶的抑制在癌治疗中已经显示出功效 (Sawyers，2004)。虽然其在癌进展中具有关键作用，蛋白激酶途径的代表性不足是由于这些蛋白是胞内蛋白，因此不可能被分泌到尿中。

[0289] 抗体阵列筛选。2,048 个在胃癌组织和正常组织之间差异性表达的基因中,26 个蛋白包括 274 个抗体的阵列中(图 14)。这 26 个蛋白中,通过我们的模型预测 7 个(FGF7、CD14、MMP9、MMP2、MMP10、TREM1、CEACAM1)会被分泌。所述抗体阵列数据确认,在至少一个或多个样品中经预测被分泌的 7 个蛋白中的 6 个存在于尿中。但是,在 6 个样品中的任一个中都未检测到 MMP10,表明其是假阳性。尽管如此,该模型在预测分泌尿蛋白方面是精确的。

[0290] 从抗体阵列中,发现 10 个蛋白(Fit3- 配体、EGF-R、sgpBO、PDGF AA、黄体化激素、Tim-3、Trappin-2、CEA、CEACAM1、FSH)在所有癌样品中与正常样品相比基本上下调(图 14),表明这些可以作为可能的新的生物标记,但是在胃癌中的浓度减少。这 10 个蛋白中,CEACAM1 是唯一包括在 2048 个在胃癌样品和参照样品之间差异性表达的基因的数据集中的蛋白(Cui 等,2009)。据预测该蛋白被该模型分泌,这表明了该模型在鉴定尿中潜在的生物标记方面的成功。

[0291] 对数个所预测的尿分泌蛋白进行蛋白质印迹分析。基于尿分泌预测评级和蛋白功能选择了 3 个蛋白 MUC13、COL10A1 和 EL。跨膜粘蛋白 MUC13 已经在胃癌组织中显示出上调,并且已经被建议作为潜在的诊断和治疗靶标(Shimamura 等,2005)。其具有 3 个可能涉及细胞粘附、调节、细胞信号传导、趋化性、伤口愈合和粘蛋白 / 生长因子相互作用的 EGF 样结构域(Williams 等,2001 ;N′ Dow 等,2004)。

[0292] 据预测 MUC13(58kD)被分泌到尿中,并且蛋白质印迹确认了该预测。如图 15 所示,MUC13 同时存在于胃癌患者和对照的尿样品中。使用 ImageJ 软件确定条带的相对定量,其中对各泳道进行分析,并且确定和比较峰下的面积。虽然微阵列数据揭示 MUC13 显示了 mRNA 水平上的差异,蛋白质印迹条带的定量未显示在 58kD 的条带的癌样品和对照样品之间显示显著性差异。由于该条带位于 55K ～ 75K 之间,这些结果表明该蛋白以完整形式或接近完整的形式被分泌到尿中。

[0293] COL10A1 是同源三聚型胶原,具有较大的 C 端和 N 端结构域(Gelse 等,2003)。据认为其参与较低的肥大区中的钙化过程,并且发现其位于透明软骨的推定矿化区(Schmid&Linsenmayer,1987 ;Kwan 等,1989 ;Kirsch&Mark,99 ;Alini 等,1994)。已经发现其在乳癌和卵巢癌中过表达(Ferguson 等,2005)。本发明的微阵列数据还显示 COL10A1 在胃癌组织中过表达。

[0294] 对 COL10A(66kD)进行的蛋白质印迹显示了一条 37kD ～ 50kD 之间的较清楚的条带,表明该蛋白可能由于一次或多次切割而以不完整形式主要出现在尿中(图 16)。当比对照样品相比时胃癌样品的平均强度高出约 50%。

[0295] 内皮脂肪酶(EL)(55kD)由内皮细胞产生,并且在通常的脂质代谢中在合成位点处发挥作用(Choi 等,2002 ;shida 等,2003)。数个研究已经表明,该蛋白是控制 HDL 水平的决定因素,并且在 EL 和 HDL 的表达之间存在反相关(Ishida 等,2003 ;Jin 等,2003 ;Ma 等,2003)。EL 还与人类动脉粥样硬化损伤中的巨噬细胞有关,EL 的抑制减少了人类巨噬细胞中促炎症细胞因子的表达,并且减少了胞内脂质浓度(Oiu 等,2007)。

[0296] 该蛋白尚未与任何癌相联系,但是基于本发明的微阵列数据分析发现该蛋白在胃癌组织中上调(Cui 等,2009)。令人感兴趣的是,用于 EL 的蛋白质印迹显示了在胃癌患者的尿样品中相对于对照样品其丰度明显减少(图 17)。具体而言,对于所有 3 个对照样品都

检测到EL，而胃癌样品显示几乎没有或没有EL。令人吃惊的是，检测到100kD以上的条带，表明EL以活性形式（头尾衔接构象的同源；聚体）(Griffon等，2009)被分泌到尿中；对于任何样品没有观察到其它条带。

[0297] 实施例15

[0298] 用于标记鉴定的抗体阵列实验

[0299] 还使用基于生物素标记的抗体阵列对来自3个胃癌个体和3个对照的血清样品进行了蛋白阵列实验。对于基于生物素标记的阵列实验，对各血清样品进行透析，然后根据制造商说明(Pierce，Rockford，IL，USA)进行生物素标记步骤，其中将蛋白的伯胺生物素化。然后将经生物素标记的蛋白(50μl血清样品)与(抗体芯片RayBio®基于生物素标记的抗体阵列，RayBiotech，Inc.U.S.A)在室温一起温育2小时。与HRP-链霉亲和素或荧光染料-链霉亲和素一起温育后，通过化学发光或荧光使信号可视化，然后通过扫描阵列激光共聚焦幻灯片扫描器(PerkinElmer Life Science)成像。所有阵列实验重复3次。

[0300] 测定507个已知人类蛋白的丰度，包括(抗)炎症性细胞因子、趋化因子、脂肪细胞激素、基质金属蛋白酶、血管发生因子、生长和分化因子、细胞粘附分子和可溶性受体。所述分析鉴定了103个在胃癌样品和对照样品之间具有非常显著的表达差异性的蛋白，其中28个蛋白在癌样品中丰度更高，而其它的蛋白在癌样品中相对于对照样品显示较低的丰度。丰度差异性的分布示于图19中，并且这些蛋白名称的列表在表13中给出。

[0301] 这103个蛋白中只有一个蛋白(CCL28)通过本发明的质谱分析检测到，这可能归因于样品中的信号传导蛋白的丰度相对较低。基于本研究，可以总结出虽然抗体阵列可潜在地检测蛋白标记，其特异性可能成为问题。

[0302] 表13：通过基于生物素标记的抗体阵列鉴定的在癌血清中相对于对照血清具有丰度差异性的103个蛋白

[0303]

蛋白标识符	平均对照	平均癌	倍数变化
胰岛素酶 / IDE	96.7	747.3	7.7
IL-20 R α	199.0	1314.0	6.6
IL-31 RA	41.3	263.0	6.4
IL-16	244.3	1404.3	5.7
SDF-1 / CXCL12	1584.3	7729.3	4.9
SCF	585.3	2782.7	4.8
IL-17RC	29.0	120.0	4.1
TECK / CCL25	49.0	195.0	4.0
RELT / TNFRSF19L	73.7	262.0	3.6
IL-18 BPa	1622.3	5707.0	3.5
TGF- α	54.7	185.3	3.4
FGF-12	101.7	344.3	3.4
IL-17RD	1039.0	3473.0	3.3
GRO	1057.7	3534.0	3.3
DR3 / TNFRSF25	43.3	142.3	3.3
EGF R / ErbB1	145.7	406.3	2.8
IL-12 R β 1	177.7	473.0	2.7
IL-1 α	1360.0	3331.0	2.4
IL-17R	832.0	1945.3	2.3
IL-4 R	8509.3	19494.3	2.3
IL-8	1766.7	3823.3	2.2
MCP-1	725.0	1548.3	2.1
RANTES	158.0	290.0	1.8
粒酶 A	1019.0	1717.0	1.7
IL-5	1205.3	1996.3	1.7
Kremen-2	391.0	622.0	1.6
骨保护素 / TNFRSF11B	4484.7	7127.3	1.6
Siglec-9	43881.7	64277.7	1.5
MIP-1b	233.3	151.3	-1.5
抑制素 A	210.0	134.0	-1.6
MCP-2	551.7	338.0	-1.6
TGF- β 2	941.3	546.3	-1.7
TRAIL R1 / DR4 / TNFRSF10A	862.7	495.3	-1.7
NGF R	217.3	123.3	-1.8
BMP-15	562.0	314.7	-1.8
BAFF R / TNFRSF13C	413.7	228.7	-1.8

[0304]

蛋白标识符	平均对照	平均癌	倍数变化
痕迹蛋白	270.3	147.7	-1.8
B7-1 /CD80	961.3	508.7	-1.9
神经纤毛蛋白-2	565.0	294.7	-1.9
NT-4	415.0	209.0	-2.0
FGF 碱性	896.7	450.7	-2.0
MCP-3	587.7	291.7	-2.0
CTLA-4 /CD152	557.3	271.3	-2.1
BD-1	250.0	117.3	-2.1
EGF	1850.7	867.7	-2.1
IFN- α / β R1	352.7	163.3	-2.2
VE- 钙粘蛋白	412.0	187.7	-2.2
IL-2 R α	1129.3	508.3	-2.2
内皮蛋白. / CD105	1140.3	510.0	-2.2
PARC / CCL18	488.7	217.7	-2.2
CCR1	556.3	243.7	-2.3
淋巴细胞趋化因子/ XCL1	301.0	130.3	-2.3
TLR3	1029.3	445.3	-2.3
淋巴毒素 β R / TNFRSF3	271.0	116.3	-2.3
TIMP-4	477.7	201.0	-2.4
脂联素 / Acrp30	4485.0	1860.3	-2.4
CCR2	510.3	209.3	-2.4
FADD	282.0	115.7	-2.4
Vasorin	372.0	152.0	-2.4
TRAIL / TNFSF10	513.7	208.7	-2.5
CXCR5 /BLR-1	600.7	239.3	-2.5
IL-1 R4 /ST2	1342.0	532.3	-2.5
LIF	267.7	103.3	-2.6
VEGF-C	430.7	165.0	-2.6
CCR4	639.0	244.7	-2.6
IL-2 R γ	396.3	151.3	-2.6
MMP-3	207.3	78.7	-2.6
神经秩蛋白(neurturin)	1021.7	381.3	-2.7
BMP-3	1039.0	387.3	-2.7
ICAM-1	100.7	36.3	-2.8
HVEM / TNFRSF14	123.3	43.7	-2.8
IL-22 R	243.0	84.7	-2.9
WIF-1	882.7	301.3	-2.9
PDGF-BB	203.7	67.7	-3.0
IFN- α / β R2	509.3	164.7	-3.1
E- 选择蛋白	341.7	109.0	-3.1
Tie-1	231.7	73.3	-3.2
IGF-I SR	932.0	287.3	-3.2
IL-1 R6 / IL-1 Rrp2	501.3	154.0	-3.3
IL-3 R α	610.7	174.7	-3.5

[0305]

蛋白标识符	平均对照	平均癌	倍数变化
CCL28 / VIC	682.0	193.7	-3.5
IL-15 R α	282.0	80.0	-3.5
NT-3	648.7	178.3	-3.6
Tie-2	5343.7	1468.0	-3.6
血管生成素-1	814.7	219.7	-3.7
MIP-3 α	766.3	202.7	-3.8
GFR α-3	307.3	75.3	-4.1
Glut1	165.0	40.3	-4.1
PDGF-AB	526.0	124.7	-4.2
CXCR3	1713.3	384.3	-4.5
DANCE	395.7	86.7	-4.6
MFRP	736.3	146.7	-5.0
CCR3	1279.0	240.0	-5.3
VEGF-B	996.0	166.0	-6.0
CXCR4 (融合素)	1138.3	183.3	-6.2
PLUNC	137.0	20.3	-6.7
BLC / BCA-1 / CXCL13	5564.3	422.7	-13.2
sFRP-4	173.3	12.7	-13.7
EMAP-II	6165.7	383.0	-16.1
RANK / TNFRSF11A	381.7	20.3	-18.8
CXCR2 / IL-8 RB	27292.0	1048.3	-26.0
IL-22 BP	37.7	1.3	-28.3
VEGF-D	13874.7	320.0	-43.4

[0306] 实施例 16

[0307] 用于其它癌的标记鉴定

[0308] 除了胃癌之外，已经使用可公开获得的癌微阵列数据将上文概述的计算技术和额外的工具应用至其它癌。对于本研究，从互联网上的数据库已经收集了用于 8 种癌的微阵列基因表达数据：肝癌 (Chen 等，2002)、前列腺癌 (Lapointe 等，2004)、肺癌 (Garber 等，2001)、肾癌 (Sarwal 等，2001)、结直肠癌 (Giacomini 等，2005)、乳癌 (Dairkee 等，2004)、卵巢癌 (Schaner 等，2003) 和胰腺癌 (lacobuzio-Donahue 等，2003)，其中每一个都具有相对较大的样本尺寸。

[0309] 对于各数据集，使用 1-、2-、3-、4- 和 5- 基因作为标记，使用同上文概述的步骤，预测能够区分癌组织和参照组织的前 100 个标记。图 18 分别显示了通过最佳的 1- 基因和 2- 基因标记在区分 83 个前列腺癌组织和 50 个参照前列腺组织时的分类精度 (2/3 的数据用于训练，剩下 1/3 的数据用于测试，使用 5 倍交叉验证)。对于前列腺癌，最佳的 3 个 1- 基因标记是 AMACR、ITPR1 和 ACPP，分类精度分别是 88.0％、86.1％和 85.7％，最佳的 3 个 2- 基因标记是 ITGA9-SPG3A、CREB3L4-ITGA9 和 BLNK-ITGA9，分类精度都是 98.0％。令人感兴趣地观察到，在本发明的 1- 基因标记列表中在其区分癌组织和参照组织的辨别力方面广泛使用的 PSA 排在第 167 位。这与公认的 PSA 在区分前列腺癌和良性前列腺肥大上所具有的限制相一致。最近数个团队已经将 AMACR 从最佳的标记候选物中鉴定为用于前列腺癌的潜在血清标记 (Bradford 等，2006)。在以上列表中还对 7 个其它癌类型完成了类似

的分析。

[0310] 实施例 17

[0311] 通过针对公共微阵列数据的检索来对所预测的基因标记的特异性分析

[0312] 为了检查所预测的基因标记对于胃癌是否具有特异性，开发了生物标记评价系统，针对用于人类疾病的 GEO(Barrett 等，2005)、Oncomine(Rhodes 等，2004) 和 SMD(Sherlock 等，2001) 中的公共微阵列数据集检索各个预测标记。对于各预测标记、个体基因或基因的组以及其表达倍数变化信息，进行了以下检索。如果基因标记在多种疾病上给出大致阳性的预测（目前设定为 30％），则认为该标记对于胃癌不具有特异性，并因此从候选物列表中将其除去。

[0313] 实施例 18

[0314] 用于检测差异性表达的基因 / 转录物的算法

[0315] 本研究的目标在于测试假设 (H_0)，该假设为在大多数患者中，某特定基因在表达水平上不显示出 k 倍以上变化 (p 值＜ 0.05)。对假设 H_0（即特定基因在癌中不显示特定的表达水平变化）的检查以及对该假设的否定将意味着对癌的选择性支持。设 N[i] 和 C[i] (i ＝ 1,...,m) 是第 i 个患者的参照组织和癌组织中的基因表达，m 是所有患者的数量。如果假设 H_0 为真，假定基因表达是连续随机变量，则概率 P(N[i] ＞ C[i]) ＝ P(N[i] ＜ C[i]) ＝ 0.5。设 K 为具有 N[i]/C[i] ＞ 0.5 的患者的数量，则基于中心极限定理，随机变量 K/m 大致为正态的，平均值＝ 0.5，并且标准偏差=0 5/$\sqrt{m}$ ，或x=2k/$\sqrt{m}$ 具有标准正态分布 N(0，1)。因此可将 p 值估计为$P(X > 2K_{exp}/\sqrt{m})$，其中是 K_{exp} 是具有 P(N[i] ＜ C[i]) 的患者的实验观察数。

[0316] 实施例 19

[0317] 胃癌的公共微阵列数据

[0318] 为了避免由样品分布的偏差引起的矛盾，从 GEO 数据库下载了用于胃癌的两个公共微阵列数据集用于进行比较性研究：一个 (Kim 数据集)(Kim 等，2007) 测定了韩国的不同阶段、癌类型和癌分化程度的 50 名癌患者的基因表达谱。对于各肿瘤相对于正常样品的平均值通过计算 log2 倍数变化值给出原始数据；另一个 (Xin 数据集，GSE2701)(Chen 等，2003) 使用针对常见对照 (CRG) 的 44K 人类阵列进行评估，测定了从香港收集的总共 126 名胃癌患者肿瘤的基因表达。第一集合已经进行标准化和对数转化，并且我们通过按照 (Sharma 等，2008) 中所述的相同步骤对 Xin 数据集进行了预处理。

[0319] 将具有韩国 50 名胃癌患者的基因表达数据的 Kim 数据集，用于评价早期标记，将具有 100 个胃癌组织和 24 个参照组织的基因表达数据的 Xin 数据集，用于评估本发明所提出的基因标记的通用性。

[0320] 实施例 20

[0321] 将已知的剪接用顺式调控基序映射到紧邻于被略过的外显子之前的内含子

[0322] 已经收集了据认为参与剪接调节的 362 个内含子顺式调控基序 (Wang 等，2008)。Wang 等，2008 中的研究表明，外显子的紧邻上游内含子区（相对于 5′ 剪接位点的 -150nt ～ -30nt) 富集有所述顺式调控基序通常表明该外显子可以被选择性剪接。进一步的分析表明，所述顺式调控基序的更高的出现次数与更高的所述外显子的外显子略过事件的发生次数相关。因此，对于各外显子，对这些调控基序 (100％序列匹配）在如上限定的

内含子区中的出现进行计数。

[0323] 本文通过援引将以上说明书中提到的所有出版物和专利并入。考虑到本文公开的本发明的说明书和实践，对本领域技术人员而言本发明的其它实施方式会变得显而易见。说明书和实例旨在仅被视作示例性的，而本发明的真实范围和主旨由后附权利要求所指定。

[0324] 参考文献

[0325] Adkins JN, Varnum SM, Auberry KJ, Moore RJ, Angell NH, Smith RD 等 . Toward a human blood serum proteome :analysis by multidimensional separation coupled with mass spectrometry. Mol Cell Proteomics. 2002 ;1(12) :947-55.

[0326] Schrader M, Schulz-Knappe P. Peptidomics technologies for human body fluids. Trends Biotechnol. 2001 ;19(10Suppl) :S55-60.

[0327] Tolson J, Bogumil R, Brunst E, Beck H, Eisner R, Humeny A 等 . Serum protein profiling by SELDI mass spectrometry :detection of multiple variants of serum amyloid alpha in renal cancer p atients. Lab Invest. 2004 ;84(7) :845-56.

[0328] Holmila R, Fouquet C, Cadranel J, Zalcman G, Soussi T. Splice mutations in the p53gene :case report and review ofthe literature. Hum Mutat. 2003 ;21(1) : 101-2.

[0329] Li HR, Wang-Rodriguez J, Nair TM, Yeakley JM, Kwon YS, Bibikova M 等 . Two-dimensional transcriptome profiling :identification of messenger RNA isoform signatures in prostate cancer from archived paraffin-embedded cancer specimens. Cancer Res. 2006 ;66(8) :4079-88.

[0330] Smith MW, Yue ZN, Geiss GK, Sadovnikova NY, Carter VS, Boix L 等 . Identification of novel tumor markers in hepatitis C virus-associated hepatocellular carcinoma. Cancer Res. 2003 ;63(4) :859-64.

[0331] Young AN, de Oliveira Salles PG, Lim SD, Cohen C, Petros JA, Marshall FF 等 . Betadefensin-1, parvalbumin, and vimentin :a panel of diagnostic immunohistochemical markers for renal tumors derived from gene expression profiling studies using cDNAmicroarrays. Am J Surg Pathol. 2003 ;27(2) :199-205.

[0332] van de Vijver MJ, He YD, van ′ t Veer LJ, Dai H, Hart AA, Voskuil DW 等 . Agene-expression signature as a predictor of survival in breast cancer. N Engl J Med. 2002 ;347(25) :1999-2009.

[0333] Resnick MB, Routhier J, Konkin T, Sabo E, Pricolo VE. Epidermal growth factor receptor, c-MET, beta-catenin, and p53expression as prognostic indicators in stage IIcolon cancer :a tissue microarray study. Clin Cancer Res. 2004 ;10(9) : 3069-75.

[0334] Sallinen SL, Sallinen PK, Haapasalo HK, Helin HJ, Helen PT, Schraml P 等 . Identification of differentially expressed genes in human gliomas by DNA microarray and tissue chip techniques. Cancer Res. 2000 ;60(23) :6617-22.

[0335] Hendrix MJ, Senor EA, Meltzer PS, Gardner LM, Hess AR, Kirschmann DA

等.Expression and functional significance of VE-cadherin in aggressive human melanoma cells:role in vasculogenic mimicry.Proc Natl Acad Sci U S A.2001;98(14):8018-23.PMCID:35460.

[0336] Menne KM,Hermj akob H,Apweiler R.A comparison of signal sequence prediction methods using a test set of signal peptides.Bioinformatics.2000;16(8):741-2.

[0337] Nair R,Rost B.Mimicking cellular sorting improves prediction of subcellular localization.J MoI Biol.2005;348(1):85-100.

[0338] Horton P,Park KJ,Obayashi T,Fujita N,Harada H,Adams-Collier CJ等.WoLFPSORT:protein localization predictor. Nucleic Acids Res. 2007;35(Web Server issue):W585-7.

[0339] Guda C.pTARGET:a web server for predicting protein subcellular localization.Nucleic Acids Res.2006;34(Web Server issue):W210-3.

[0340] Mott R,Schultz J,Bork P,Ponting CP.Predicting protein cellular localization using a domain projection method.Genome Res.2002;12(8):1168-74.

[0341] Smialowski P,Martin-Galiano AJ,Mikol ajka A,Girschick T,Holak TA,F rishman D.Protein solubility:sequence based prediction and experimental verification.Bioinformatics,2007;23(19):2536-42.

[0342] Chen Y,Zhang Y,Yin Y,Gao G,Li S,Jiang Y等.SPD--a web-based secreted protein database.Nucleic Acids Res.2005;33(Database issue):D 169-73.

[0343] Tang ZQ,Han LY,Lin HH,Cui J,Jia J,Low BC等.Derivation of stable microarray cancer-differentiating signatures using consensus scoring of multiple random sampling and gene-ranking consistency evaluation.Cancer Res.2007;67(20):9996-10003.

[0344] Lee Y,Kim B,Shin Y,Nam S,Kim P,Kim N等.ECgene:an alternative splicing database update.Nucleic Acids Res.2007;35(Database issue):D99-103.PMCID:1716719.

[0345] Dantzig GB,A.Orden,and P.Wolfe.Generalized Simplex Method for Minimizing a Linear from Under Linear Inequality Constraints.Pacific Journal Math.1999;Vol.5:183-95.

[0346] Takeno,A. 等.Integrative approach for differentially overexpressed genes in gastric cancer by combining large-scale gene expression profiling and network analysis.Br J Cancer99,1307-1315(2008).

[0347] El-Rifai,W.,Frierson,H.F.,Jr.,Harper,J.C,Powell,S.M.&Knuutila,S.Expression profiling of gastric adenocarcinoma using cDNA array.Int J Cancer92,832-838(2001).

[0348] Becker,K.F.等.E-cadherin gene mutations provide clues to diffuse type gastriccarcinomas.Cancer Res 54,3845-3852(1994).

[0349] Hippo,Y. 等.Global gene expression analysis of gastric cancer by

oligonucleotide microarrays.Cancer Res 62,233-240(2002).

[0350] Moss,S.F.等.Decreased expression of gastrokine 1and the trefoil factor interacting protein TFIZ 1/GKN2in gastric cancer:influence of tumor histology and relationship to prognosis.Clin Cancer Res14,4161-4167(2008).

[0351] Chen,X.等.Variation in gene expression patterns in human gastric cancers.Mol Biol Cell14,3208-3215(2003).

[0352] Dar,A.A.,Belkhiri,A.&El-Rifai,W.The aurora kinase A regulates GSK-3beta in gastric cancer cells.Oncogene 28,866-875(2009).

[0353] Kim,K.R.等.[Gene expression profiling using oligonucleotide microarray in atrophic gastritis and intestinal metaplasia].Korean J Gastroenterol49,209-224(2007).

[0354] Katayama,H.等.Phosphorylation by aurora kinase A induces Mdm2-mediated destabilization and inhibition of p53.Nat Genet 36,55-62(2004).

[0355] Chen,L.等.,Clinicopathological significance of overexpression of TSPAN1,Ki67and CD34in gastric carcinoma.Tumori,2008.94(4):p.531-8.

[0356] Long,Y.M.等.,Nuclear factor kappa B:a marker of chemotherapy for human stage IV gastric carcinoma.World J Gastroenterol,2008.14(30):p.4739-44.

[0357] Yamada,Y.等.,Identification of prognostic biomarkers in gastric cancer using endoscopic biopsy samples.Cancer Sci,2008.99(11):p.2193-9.

[0358] Silva,E.M.等.,Cadherin-catenin adhesion system and mucin expression:a comparison between young and older patients with gastric carcinoma.Gastric Cancer,2008.11(3):p.149-59.

[0359] Xu,Y.,L.Zhang,and G.Hu,Potential application of alternatively glycosylated serum MUC1 and MUC5AC in gastric cancer diagnosis.Biologicals,2009.37(1):p.18-25.

[0360] Takeno,A.等.,Integrative approach for differentially overexpressed genes in gastric cancer by combining large-scale gene expression profiling and network analysis.Br J Cancer,2008.99(8):p.1307-15.

[0361] Kon,O.L.等.,The distinctive gastric fluid proteome in gastric cancer reveals a multi-biomarker diagnostic profile.BMC Med Genomics,2008.1:p.54.

[0362] Bernal,C等.,Reprimo as a potential biomarker for early detection in gastric cancer.Clin Cancer Res,2008.14(19):p.6264-9.

[0363] Taddei,A.等.,NF2expression levels of gastrointestinal stromal tumors:a quantitative real-time PCR study.Tumori,2008.94(4):p.551-5.

[0364] Ebert,M.P.等.,Overexpression of cathepsin B in gastric cancer identified by proteome analysis.Proteomics,2005.5(6):p.1693-704.

[0365] Stefatic,D.等.,Optimization of diagnostic ELISA-based tests for the detection of autoantibodies against tumor antigens in human serum.Bosn J Basic Med Sci,2008.8(3):p.245-50.

[0366] Jin, B. 等., Detection of serum gastric cancer-associated MG7-Ag from gastric cancer patients using a sensitive and convenient ELISA method. Cancer Invest, 2009. 27(2) :p. 227-33.

[0367] Ren, H. 等., Analysis of variabilities of serum proteomic spectra in patients with gastric cancer before and after operation. World J Gastroenterol, 2006. 12(17) :p. 2789-92.

[0368] Peduzzi P, C. J., Feinstein AR, Holford TR Importance of events per independent variable in proportional hazards regression analysis. II. Accuracy and precision of regression estimates. Journal of ClinicalEpidemiology 48, 1503-1510(1995).

[0369] Chandanos, E. &Lagergren, J. Oestrogen and the enigmatic male predominance of gastric cancer. Eur J Cancer 44, 2397-2403(2008).

[0370] Guojun Li, Q. M., Haibao Tang, Ying Xu. QUBIC :A Qualitative Biclustering Algorithm for Analyses of Gene Expression Data. (2009).

[0371] Dennis, G., Jr. 等 .DAVID :Database for Annotation, Visualization, and Integrated Discovery. Genome Biol4, P3(2003).

[0372] Wu, J., Mao, X., Cai, T., Luo, J. &Wei, L KOBAS server :a web-based platform forautomated annotation and pathway identification. Nucleic Acids Res 34, W720-724(2006).

[0373] Zhu, J. 等. The UCSC Cancer Genomics Browser. NatMethods 6, 239-240(2009).

[0374] Schaefer, C. F. 等 .PID :the Pathway Interaction Database. Nucleic Acids Res 37, D674-679(2009).

[0375] Liu, R. 等 .Mechanism of cancer cell adaptation to metabolic stress : proteomics identification of a novel thyroid hormone-mediated gastric carcinogenic signaling pathway. MolCell Proteomics 8, 70-85(2009).

[0376] Bell, G. I. 等 .Facilitative glucose transport proteins :structure and regulation of expression in adipose tissue. Int J Obes 15Suppl 2, 127-132(1991).

[0377] Wang, ET. 等 .Alternative isoform regulation in human tissue transcriptomes. Nature 456, 470-476(2008).

[0378] Eyras, E., Caccamo, M., Curwen, V. &Clamp, M. ESTGenes :alternative splicing from ESTs in Ensembl. Genome Res 14, 976-987(2004).

[0379] Kanehisa, M. a. G., S. KEGG :Kyoto Encyclopedia of Genes and Genomes. Nucleic AcidsRes. 28, 27-30(2000).

[0380] Cui, J., Liu, Q., Puett, D. &Xu, Y. Computational Prediction of Human Proteins That Can Be Secreted into the Bloodstream. Bioinformatics(2008).

[0381] Omenn GS, States DJ, Adamski M, Blackwell TW, Menon R, Hermj akob H 等 .Overview of the HUPO Plasma Proteome Project :results from the pilot phase with 35collaborating laboratories and multiple analytical groups, generating a core dataset of 3020proteins and a publicly-available database.

Proteomics. 2005 ;5(13) :3226-45.

[0382] Chen Y, Zhang Y, Yin Y, Gao G, Li S, Jiang Y 等 . SPD-a web-based secreted protein database. Nucleic Acids Res. 2005 ;33(Database issue) :D169-73.

[0383] Bateman A, Birney E, Cerruti L, Durbin R, Etwiller L, Eddy S 等 . The Pfam protein families database. Nucleic acids research. 2002 ;30(1) :276-80.

[0384] Reczko M, Bohr H. The DEF data base of sequence based protein fold class predictions. Nucleic Acids Res. 1994 ;22(17) :3616-9.

[0385] Bhasin M, Raghava GP. Classification of nuclear receptors based on amino acid composition and dipeptide composition. J Biol Chem. 2004 ;279(22) :23262-6.

[0386] Platt JC. Fast Training of Support Vector Machines using Sequential Minimal Optimization. Advances in kernel methods :support vector learning. Cambridge, MA, USA :MIT Press 1999. p. 185-208.

[0387] S. S. Keerthi SKS, C. Bhattacharyya, K. R. K. Murthy. Improvements to Platt' s SMOAlgorithm for SVM Classifier Design Neural Computation. 2001 ;13 :637-49.

[0388] Poola, L 等 . Identification of MMP-I as a putative breast cancer predictive marker by global gene expression analysis. Nat Med 11, 481-483(2005).

[0389] Ebert, M. P. 等 . Overexpression of cathepsin B in gastric cancer identified by proteome analysis. Proteomics 5, 1693-1704(2005).

[0390] Poon, T. C. 等 . Diagnosis of gastric cancer by serum proteomic fingerprinting. Gastroenterology 130, 1858-1864(2006).

[0391] Pieper R, Gatlin C, McGrath A, Makusky A, Mondal M, Seonarain M, Field E, Schatz C, Estock M, Ahmed N, al e(2004). Characterization of the human urinary proteome :a method for high-resolution display of urinary proteins on two-dimensional electrophoresis gels with a yield of nearly 1400nearly protein spots. Proteomics, 1159-1174.

[0392] Castagna A, Cecconi D, Sennels L, Rappsilber J, Guerrier L, Fortis F, Boschetti E, Lomas L, Righetti P(2005). Exploring the hidden human urinary proteome via ligand library beads. JProteome Res, 1917-1930.

[0393] Wang L, Li F, Sun W, Wu S, Wang X, Zhang L, Zheng D, Wnag J, Gao Y(2006). Concanavalin A captured glycoproteins in healthy human urine. Mol Cell Proteomics, 560-562.

[0394] Chang C-C, Lin C-J(2001). LIB SVM :a library for support vector machines.

[0395] Li ZR, Lin HH, Han LY, Jiang L, Chen X, Chen YZ(2006). PROFEAT :a web server for computing structural and physicochemical features of proteins and peptides from amino acid sequence. Nucleic AcidsRes. 34, W32-37.

[0396] Prilusky J, Felder CE, Zeev-Ben-Mordehai T, Rydberg EH, Man O, Beckmann JS, Silman I, Sussman JL(2005). FoldIndex :a simple tool to predict whether a given protein sequence is intrinsically unfolded. Bioinformatics. 21, 3435-3438.

[0397] Gasteiger E, Gattiker A, Hoogland C, Ivanyi I, Appel RD, Bairoch A(2003).

ExPASy :The proteomics server for in-depth protein knowledge and analysis. Nucleic Acids Res. 31,3784-3788.

[0398] Bendtsen JD,Nielsen H,Widdick D,Palmer T,Brunak S(2005). Prediction of twin-arginine signal peptides. BMC Bioinformatics. 6,167.

[0399] Kail L, Krogh A, Sonnhammer EL(2007). Advantages of combined transmembrane topology and signal peptide prediction-the Phobius web server. Nucleic Acids Res. 35, W429-432.

[0400] Julenius K, Molgaard A, Gupta R, Brunak S(2005). Prediction, conservation analysis, and structural characterization of mammalian mucin-type O-glycosylation sites. Glycobiology. 15,153-164.

[0401] Gupta R, Jung E, Brunak S(2004). Prediction of N-glycosylation sites in human proteins eds).

[0402] Eisenhaber F, Imperiale F, Argos P, Froemmel C(1995). Prediction of Secondary Structural Content of Proteins from Their Amino Acid Comosition Alone Utilizing Analytic Vector Decompositioned eds).

[0403] Mao X, Cai T, Olyarchuk JG, Wei L(2005). Automated Genome Annotation and Pathway Identification Using the KEGG Orthology(KO)As a Controlled Vocabulary. Bioinformatics,3787-3793.

[0404] Ashkenas J, Muschler J, Bissell M(1996). The extracellular matrix in epithelial biology :Shared molecules and common themes in distant phyla. Dev Biol. 180,433-444.

[0405] McKinnell RG, Parchment RE, Perantoni A, Damj anov I, Pierce GB(2006). TheBiological Basis of Cancer. 2.

[0406] Stein GS,Pardee AB (2004). Cell cycle and Growth Control :Biomolecular Regulation and Cancer. 2.

[0407] Frixen U, Behrens J, Sachs M, Elberle G, Voss B, Warda A, Lochner D, Birchmeier W (1991). E-Cadherin-mediated cell-cell adhesion prevents invasiveness of human carcinoma cells. J Cell Biology. 113,173-185.

[0408] de Visser KE, Eichten A, Coussens LM(2006). Paradoxical roles of the immune system during cancer development. Nat Rev Cancer. 6,24-37.

[0409] Malumbres M, Barbacid M(2007). Cell cycle kinases in cancer. Curr Opin Genet Dev. 17,60-65.

[0410] Greenman C, Stephens P, Smith R(2009). Patterns of Somatic Mutation in Human Cancer Genomes. Nature. 446,153-158.

[0411] Sawyers C(2004). Targeted cancer therapy. Nature. 432,294-297.

[0412] Cui J, Chen Y, Chou J, Sun L(2009). Biomarker Identification for Gastric Cancered eds) :The University of Georgia.

[0413] Shimamura T, Ito H, Shibahara J, Watanabe A, Hippo Y, Taniguchi H, Chen Y, Kashima T, Ohtomo T, Tanioka F, Iwanari H, Kodama T, Kazui T, Sugimura H, Fukayama

M,Aburatani H(2005).Overexpression of MUC 13is associated with intestinal-type gastric cancer.Cancer Sci.96,265-273.

[0414] Williams SJ,Wreschner DH,Tran M,Eyre HJ,Sutherland GR,McGuckin MA(2001).Mucl3,a novel human cell surface mucin expressed by epithelial and hemopoietic cells.J Biol Chem.276,18327-18336.

[0415] N′ Dow J,Pearson J,Neal D(2004).Mucus production after transposition of intestinal segments into the urinary tract.World J Urol.22,178-185.

[0416] Gelse K,Poschl E,Aigner T(2003).Collagens-structure,function,and biosynthesis.Adv Drug DelivRev.55,1531-1546.

[0417] Schmid TM,Linsenmayer TF(1987).Type X collagen.Orlando :Academic Press.

[0418] Ferguson DA,Muenster MR,Zang Q,Spencer JA,Schageman JJ,Lian Y,Garner HR,Gaynor RB,Huff JW,Pertsemlidis A,Ashfaq R,Schorge J,Becerra C,Williams NS,Graff JM(2005).Selective identification of secreted and transmembrane breast cancer markers using Escherichia coli ampicillin secretion trap.CancerRes.65,8209-8217.

[0419] Choi SY,Hirata K,Ishida T,Quertermous T,Cooper AD(2002).Endothelial lipase :a new lipase on the block.J Lipid Res.43,1763-1769.

[0420] Ishida T,Choi S,Kundu RK,Hirata K,Rubin EM,Cooper AD,Quertermous T(2003).Endothelial lipase is a major determinant of HDL level.J Clin Invest.111,347-355.

[0421] Jin W,Millar JS,Broedl U,Glick JM,Rader DJ(2003).Inhibition of endothelial lipase causes increased HDL cholesterol levels in vivo.J ClinInvest.111,357-362.

[0422] Ma K,Cilingiroglu M,Otvos JD,Ballantyne CM,Marian AJ,Chan L(2003).Endothelial lipase is a major genetic determinant for high-density lipoprotein concentration,structure,and metabolism.Proc Natl Acad Sci USA.100,2748-2753.

[0423] Qiu G,Ho AC,Yu W,Hill JS(2007).Suppression of endothelial or lipoprotein lipase in THP-I macrophages attenuates proinflammatory cytokine secretion.J LipidRes.48,385-394.

[0424] Griffon N,Jin W,Petty TJ,Millar J,Badellino KO,Saven JG,Marchadier DH,Kempner ES,Billheimer J,Glick JM,Rader DJ(2009).Identification of the Active Form of Endothelial Lipase,a Homodimer in a Head-to-Tail Conformation.J Biol Chem.284,23322-23330.

[0425] Chen X,Cheung ST,So S,Fan ST,Barry C,Higgins J等.Gene expression patterns in human liver cancers.MoI Biol Cell.2002 ;13(6) :1929-39.PMCID :117615.

[0426] Lapointe J,Li C,Higgins JP,van de Rij n M,Bair E,Montgomery K等 .Geneexpression profiling identifies clinically relevant subtypes of prostate cancer.Proc Natl Acad Sci U S A.2004 ;101(3) :811-6.PMCID :321763.

[0427] Garber ME, Troyanskaya OG, Schluens K, Petersen S, Thaesler Z, Pacyna-Gengelbach M 等. Diversity of gene expression in adenocarcinoma of the lung. Proc Natl Acad Sci U S A. 2001 ;98(24) :13784-9. PMCID :61119.

[0428] Sarwal M, Chang S, Barry C, Chen X, Alizadeh A, Salvatierra O 等 .Genomicanalysis of renal allograft dysfunction using cDNA microarrays. Transplant Proc. 2001 ;33(1-2) :297-8.

[0429] Giacomini CP,Leung SY,Chen X,Yuen ST,Kim YH,Bair E 等. A gene expression signature of genetic instability in colon cancer. Cancer Res. 2005 ;65(20) : 9200-5.

[0430] Dairkee SH, Ji Y, Ben Y, Moore DH, Meng Z, Jeffrey S S. A molecular′ signature′ of primary breast cancer cultures ;patterns resembling tumor tissue. BMC Genomics. 2004 ;5(1) :47. PMCID :509241.

[0431] Schaner ME, Ross DT, Ciaravino G, Sorlie T, Troyanskaya O, Diehn M 等 .Geneexpression patterns in ovarian carcinomas. MoI Biol Cell. 2003 ;14(1 1) : 4376-86. PMCID :266758.

[0432] Iacobuzio-Donahue CA, Maitra A, Olsen M, Lowe AW, van Heek NT, Rosty C 等. Exploration of global gene expression patterns in pancreatic adenocarcinoma using cDNAmicroarrays. Am J Pathol. 2003 ;162(4) :1151-62. PMCID :1851213.

[0433] Bradford TJ, Tomlins SA, Wang X, Chinnaiyan AM. Molecular markers of prostate cancer. Urol Oncol. 2006 ;24(6) :538-51.

[0434] Barrett T, Suzek TO, Troup DB, Wilhite SE, Ngau WC, Ledoux P 等 . NCBI GEO : mining millions of expression profiles-database and tools. Nucleic Acids Res. 2005 ;33(Database issue) :D562-6. PMCID :539976.

[0435] Rhodes DR, Yu J, Shanker K, Deshpande N, Varambally R, Ghosh D 等. ONCOMINE :a cancer microarray database and integrated data-mining platform. Neoplasia. 2004 ;6(1) :1-6. PMCID :1635162.

[0436] Sherlock, G. 等 . The Stanford Microarray Database. Nucleic Acids Res 29, 152-155(2001).

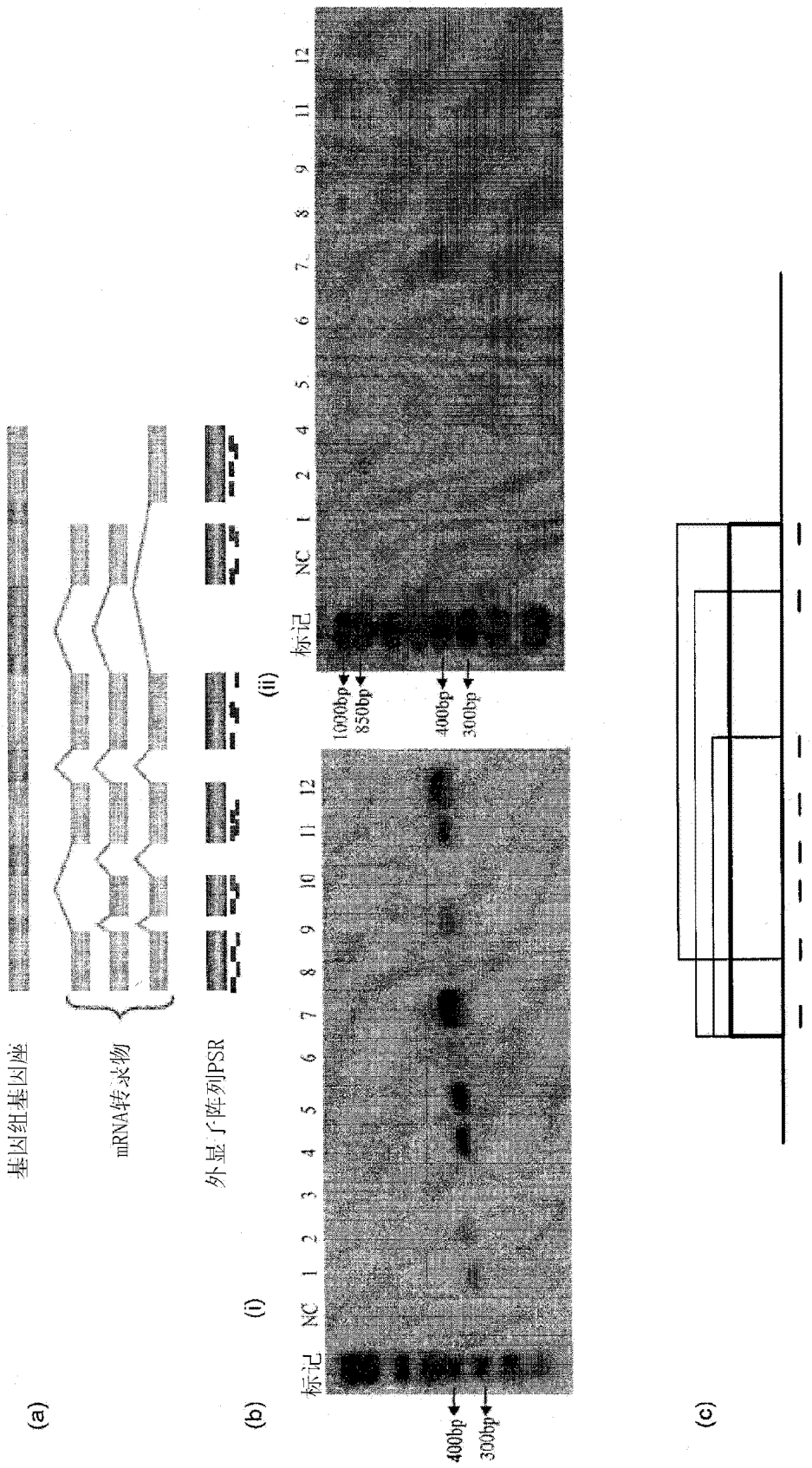
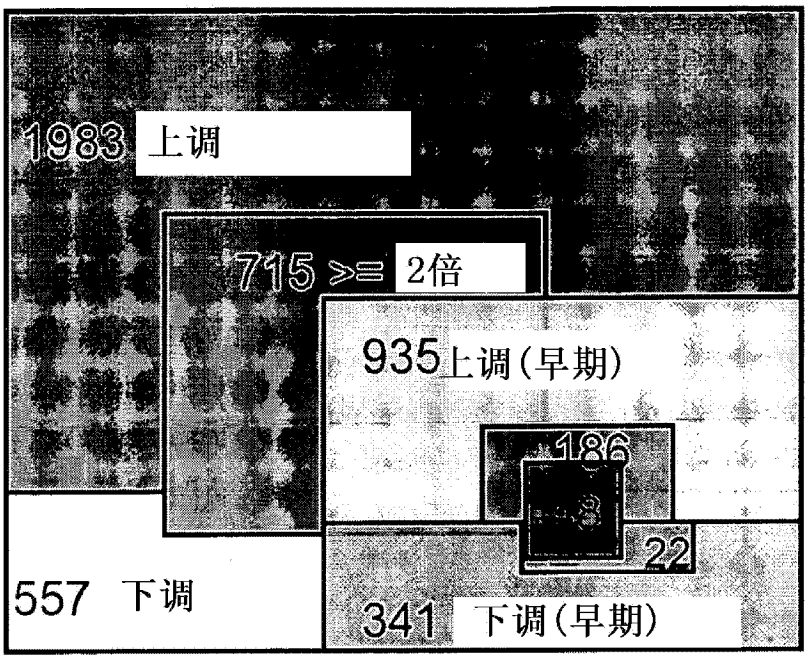

(a)
基因组基因座
mRNA转录物
外显子阵列PSR
(b)
(i)
标记
NC
1
2
3
4
5
6
7
8
9
10
11
12
400bp
300bp
(ii)
标记
NC
1
2
4
5
6
7
8
9
11
12
1000bp
850bp
400bp
300bp
(c)

图 1

(a)


1983 上调
715 >= 2倍
935 上调(早期)
186
22
557 下调
341 下调(早期)

(b)

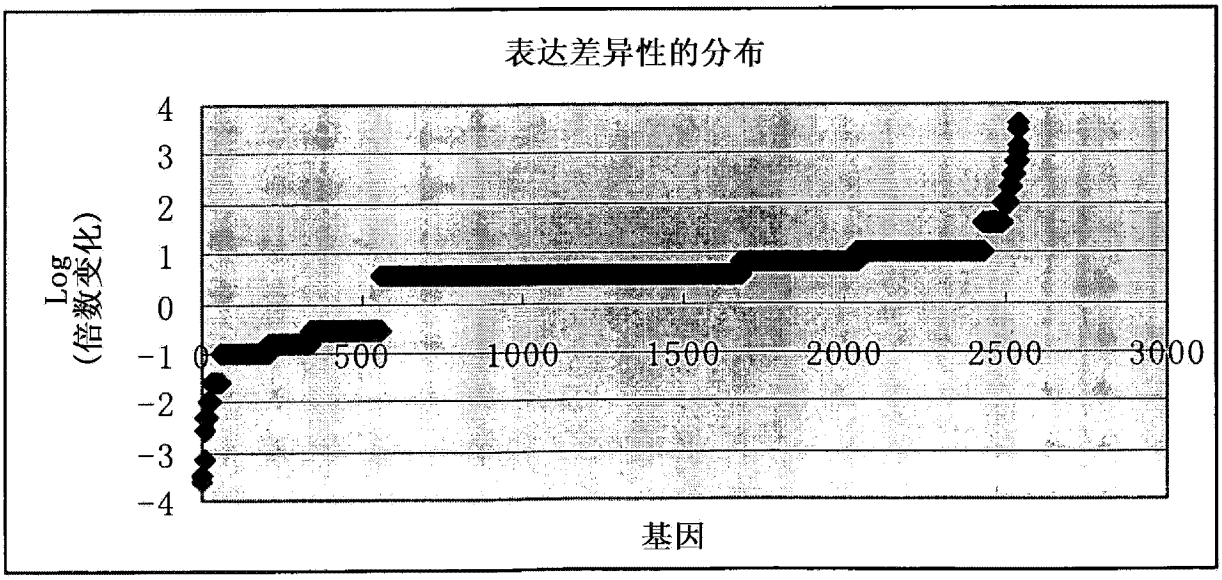
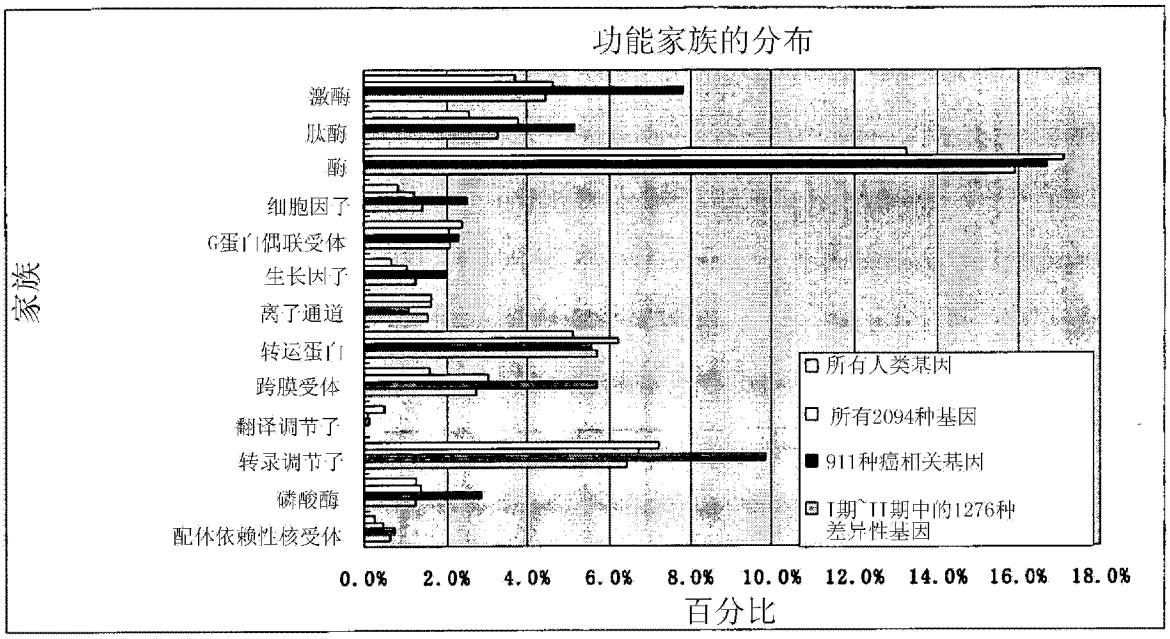

表达差异性的分布
Log (倍数变化)
4
3
2
1
0
-1
-2
-3
-4
0
500
1000
1500
2000
2500
3000
基因

图 2

(a)


功能家族的分布
家族
激酶
肽酶
酶
细胞因子
G蛋白偶联受体
生长因子
离子通道
转运蛋白
跨膜受体
翻译调节子
转录调节子
磷酸酶
配体依赖性核受体
所有人类基因
所有2094种基因
911种癌相关基因
I期~II期中的1276种差异性基因
0.0% 2.0% 4.0% 6.0% 8.0% 10.0% 12.0% 14.0% 16.0% 18.0%
百分比

(b)

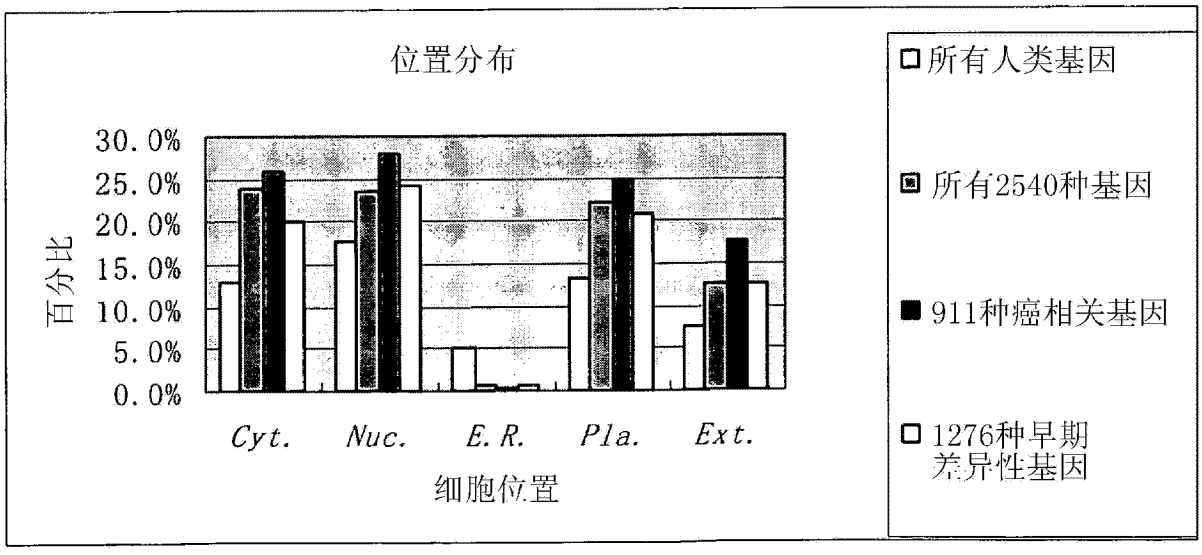
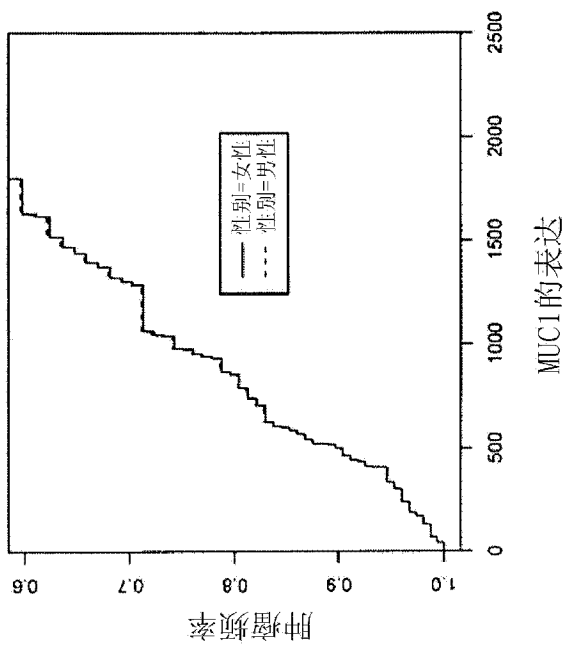
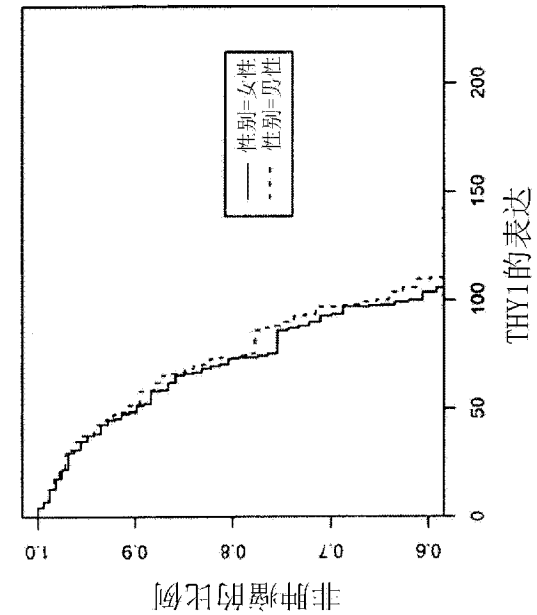
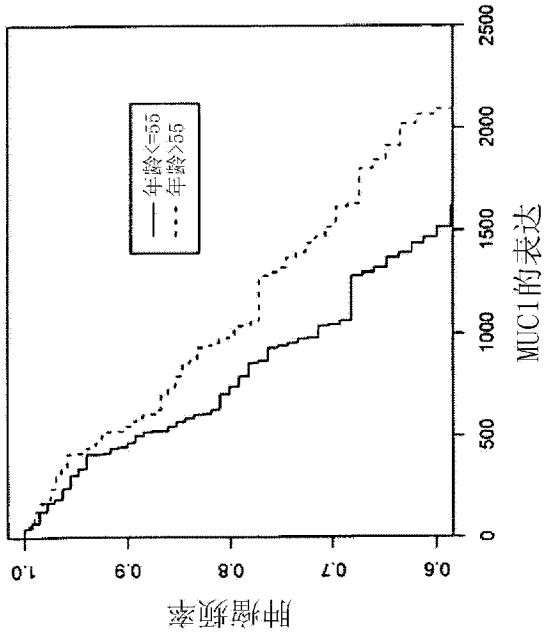

位置分布
百分比
30.0%
25.0%
20.0%
15.0%
10.0%
5.0%
0.0%
Cyt. Nuc. E.R. Pla. Ext.
细胞位置
所有人类基因
所有2540种基因
911种癌相关基因
1276种早期差异性基因

图 3


肿瘤频率
MUC1的表达
性别=女性
性别=男性


非肿瘤的比例
THY1的表达
性别=女性
性别=男性


肿瘤频率
MUC1的表达
年龄<=55
年龄>55

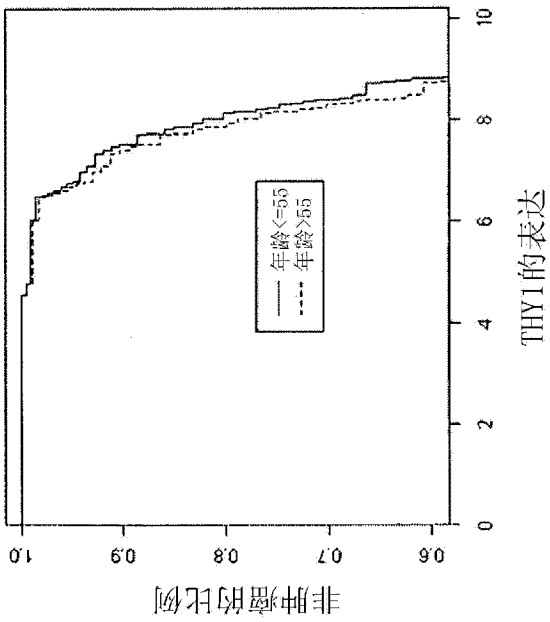

非肿瘤的比例
THY1的表达
年龄<=55
年龄>55

图 4

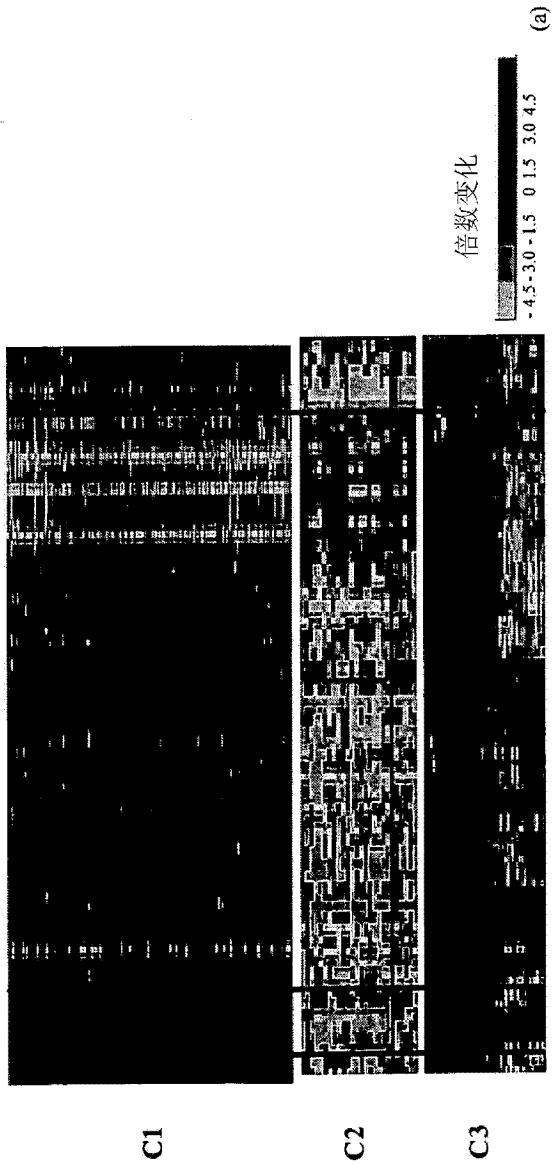


图 5

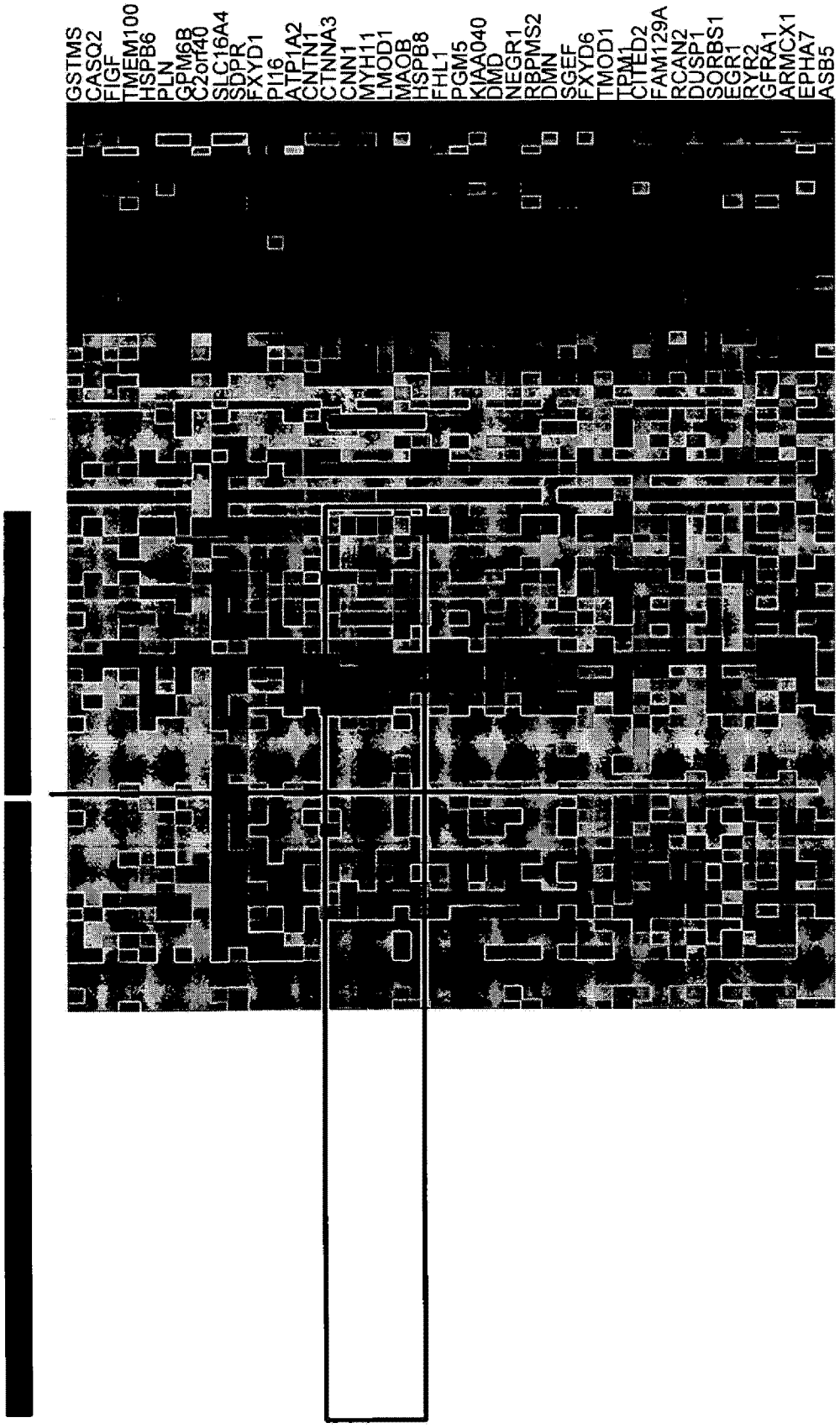

GSTM5
CASQ2
FIGF
TMEM100
HSPB6
PLN
GPM6B
C2orf40
SLC16A4
SDPR
FXYD1
PI16
ATP1A2
CNTN1
CTNNA3
CNN1
MYH11
LMOD1
MAOB
HSPB8
FHL1
PGM5
KIAA0040
DMD
NEGR1
RBPMS2
DMN
SGEF
FXYD6
TMOD1
TPM1
CITED2
FAM129A
RCAN2
DUSP1
SORBS1
EGR1
RYR2
GFRA1
ARMCX1
EPHA7
ASB5

图 5(续)

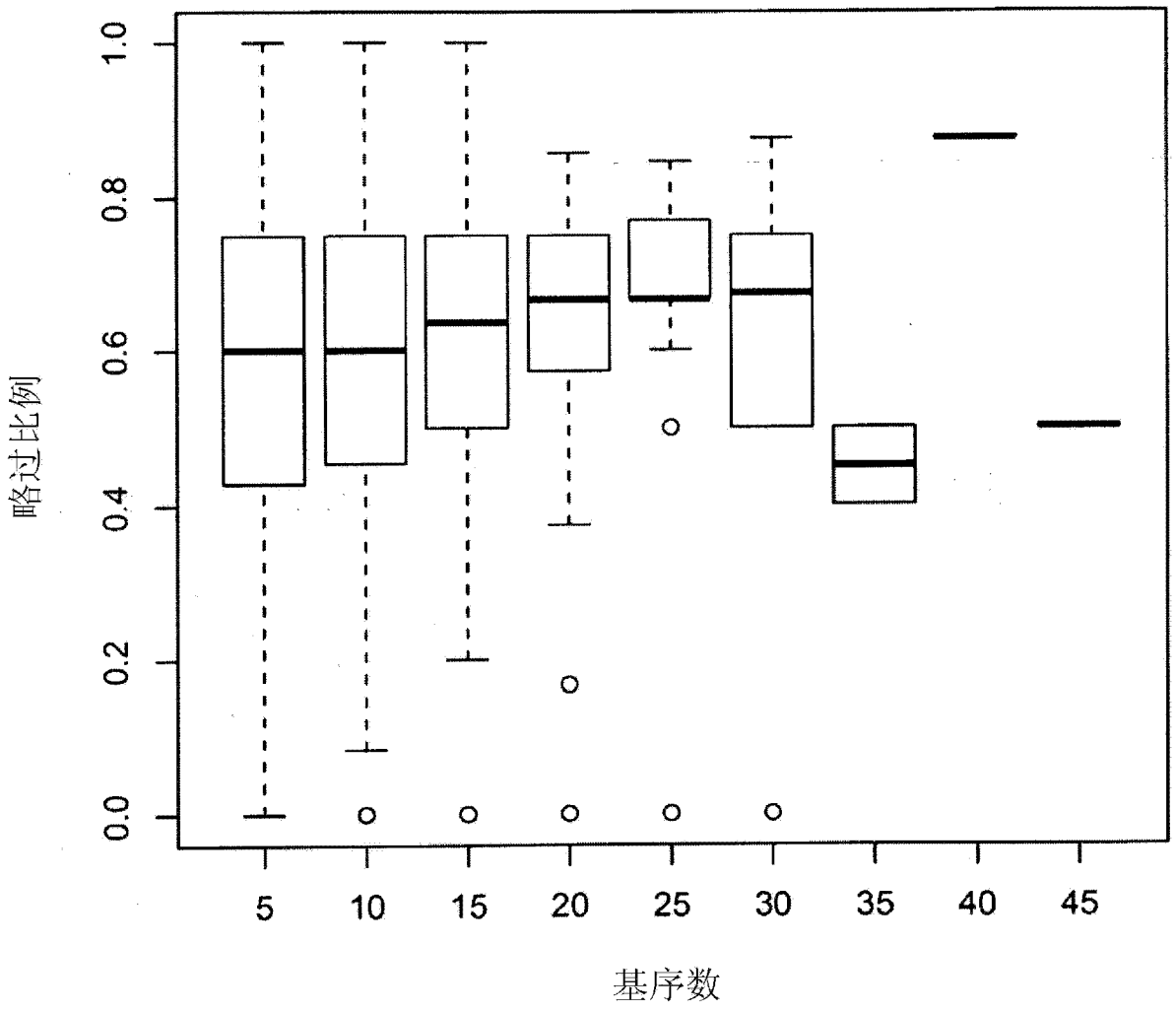


图 6

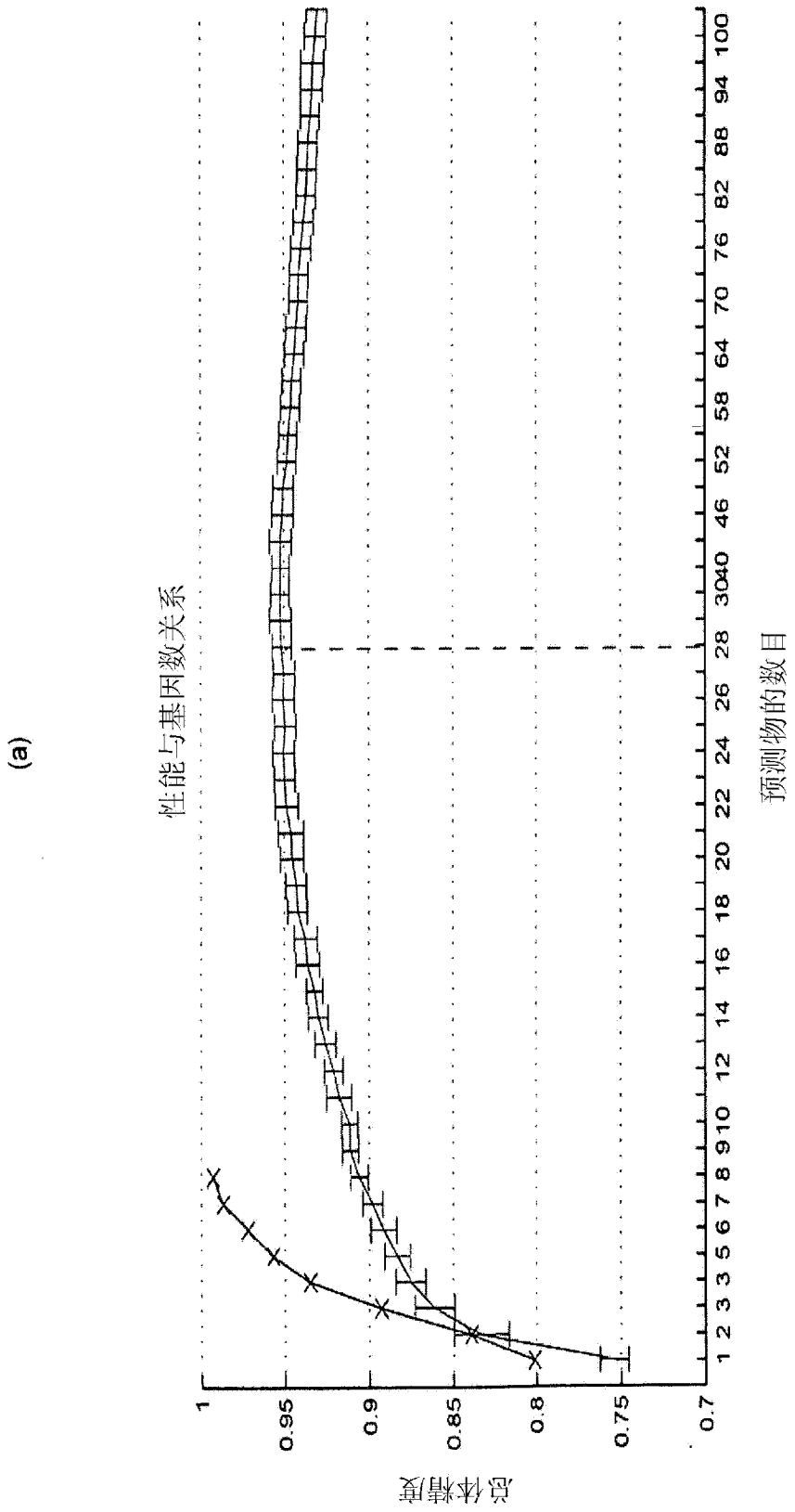

(a)
性能与基因数关系
总体精度
预测物的数目
1
0.95
0.9
0.85
0.8
0.75
0.7
1 2 3 5 6 7 8 9 10 12 14 16 18 20 22 24 26 28 3040 46 52 58 64 70 76 82 88 94 100

图 7

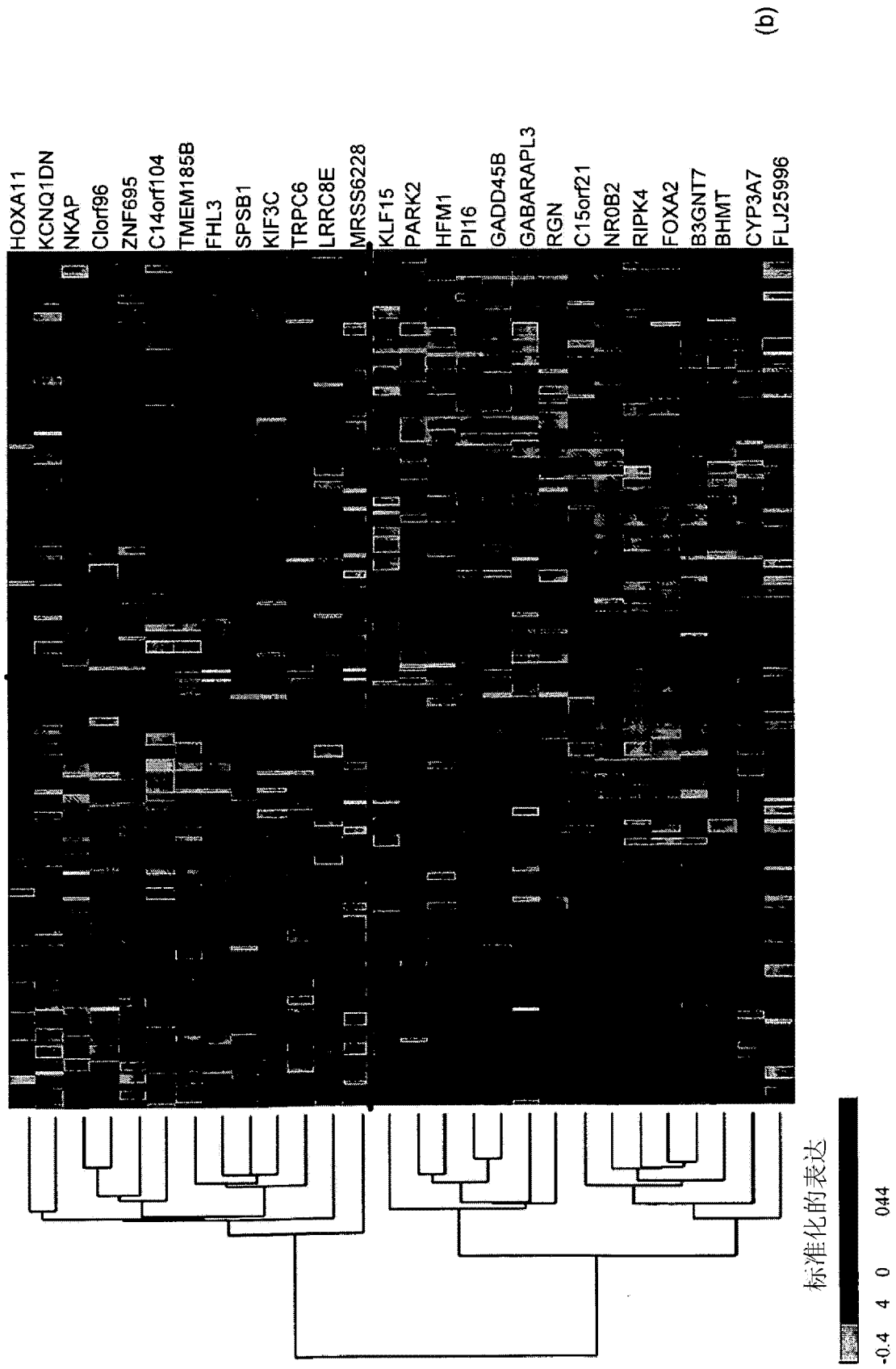

(b)
HOXA11
KCNQ1DN
NKAP
Clorf96
ZNF695
C14orf104
TMEM185B
FHL3
SPSB1
KIF3C
TRPC6
LRRC8E
MRSS6228
KLF15
PARK2
HFM1
PI16
GADD45B
GABARAPL3
RGN
C15orf21
NR0B2
RIPK4
FOXA2
B3GNT7
BHMT
CYP3A7
FLJ25996
标准化的表达
-0.4 4 0 044

图 7（续）

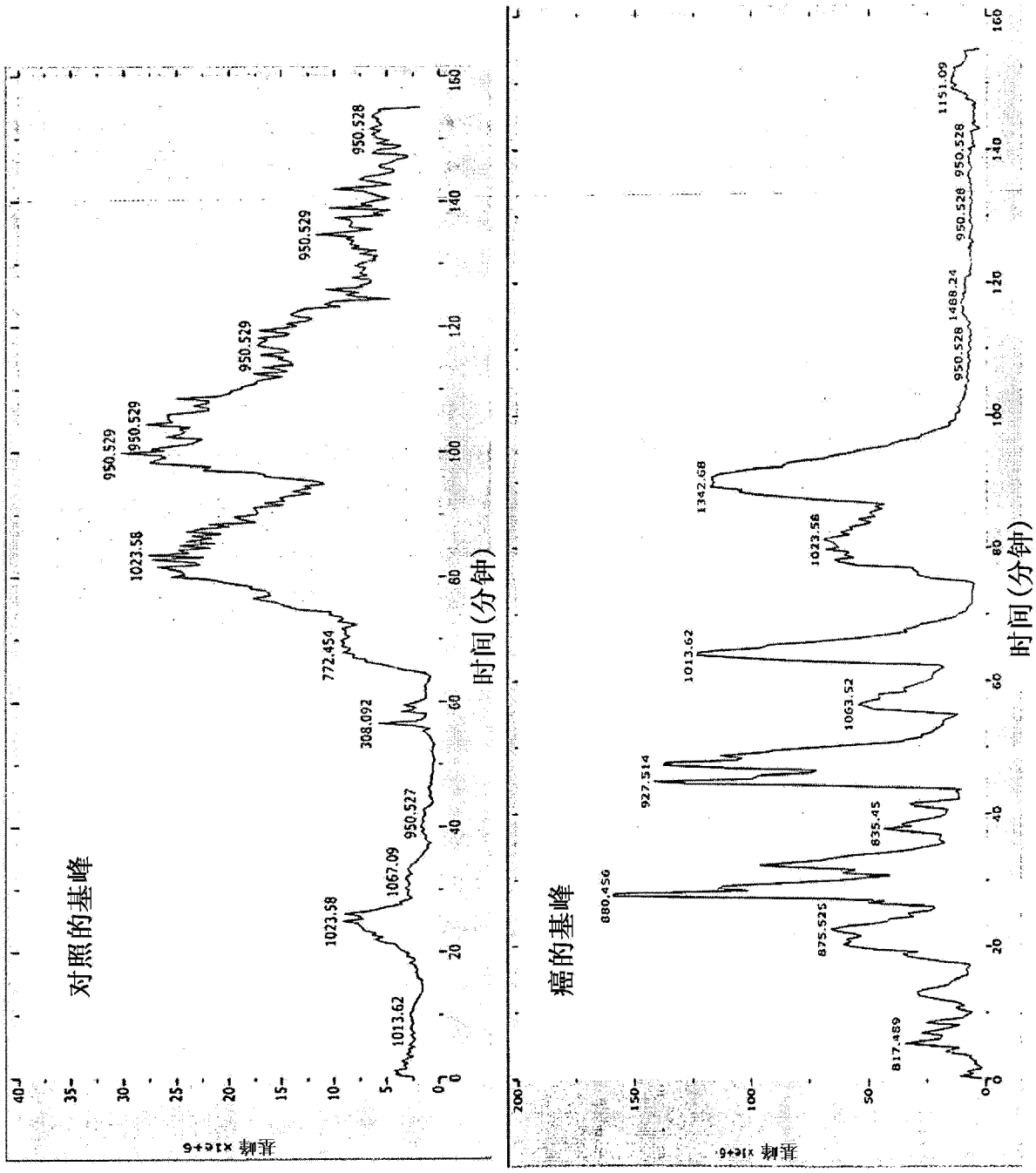

a)
对照的基峰
基峰 x1e+6
时间(分钟)
1013.62
1023.58
1067.09
950.527
308.092
772.454
950.529
950.529
1023.58
950.529
950.529
950.528
癌的基峰
基峰 x1e+6
时间(分钟)
817.489
875.525
880.456
927.514
835.45
1063.52
1013.62
1023.58
1342.68
950.528
1488.24
950.528
950.528
1151.09

图 8

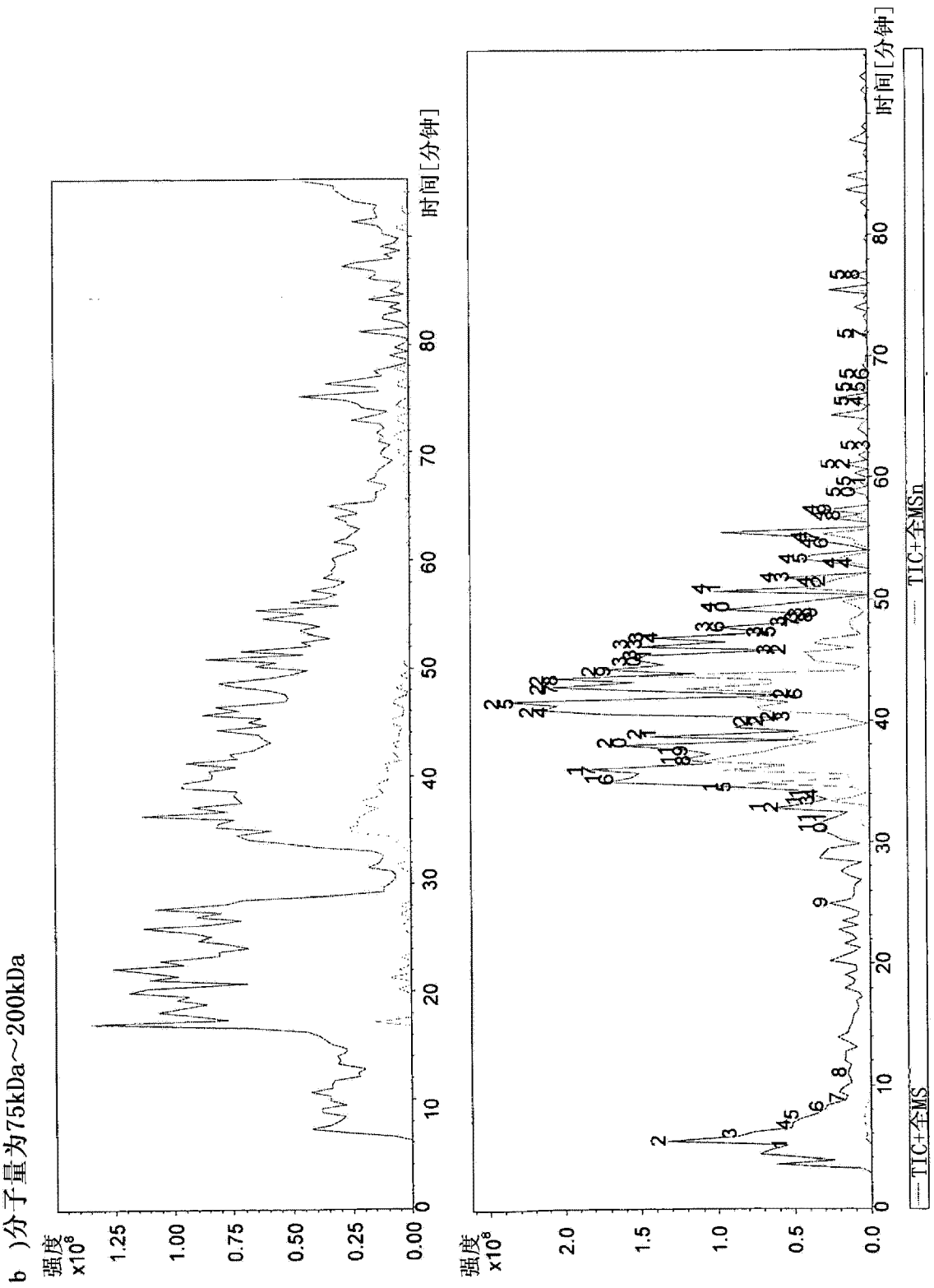


图 8 续

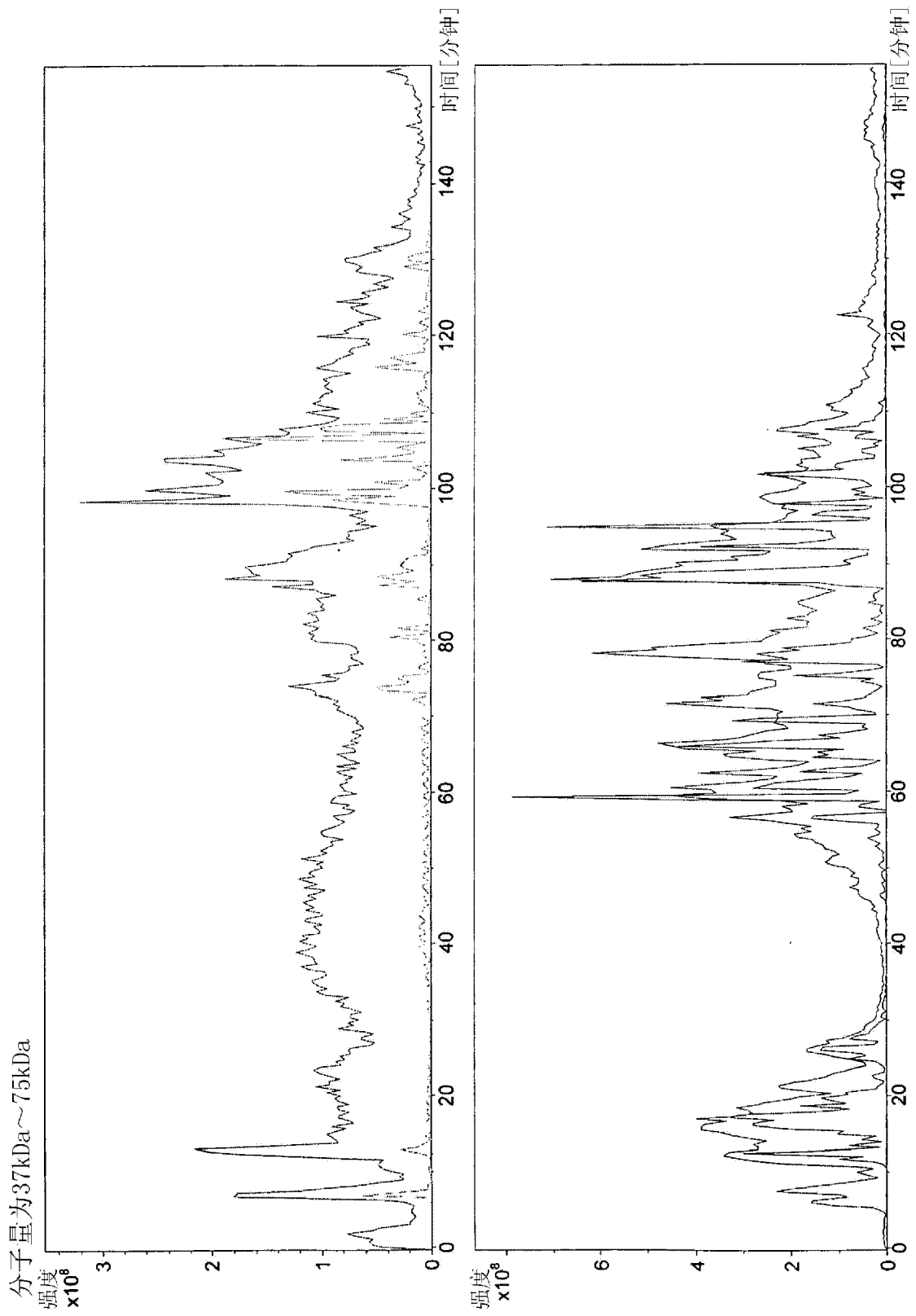

分子量为37kDa～75kDa
强度
x10^8
时间[分钟]
0
1
2
3
0
20
40
60
80
100
120
140
强度
x10^8
时间[分钟]
0
2
4
6
0
20
40
60
80
100
120
140

图 8 续

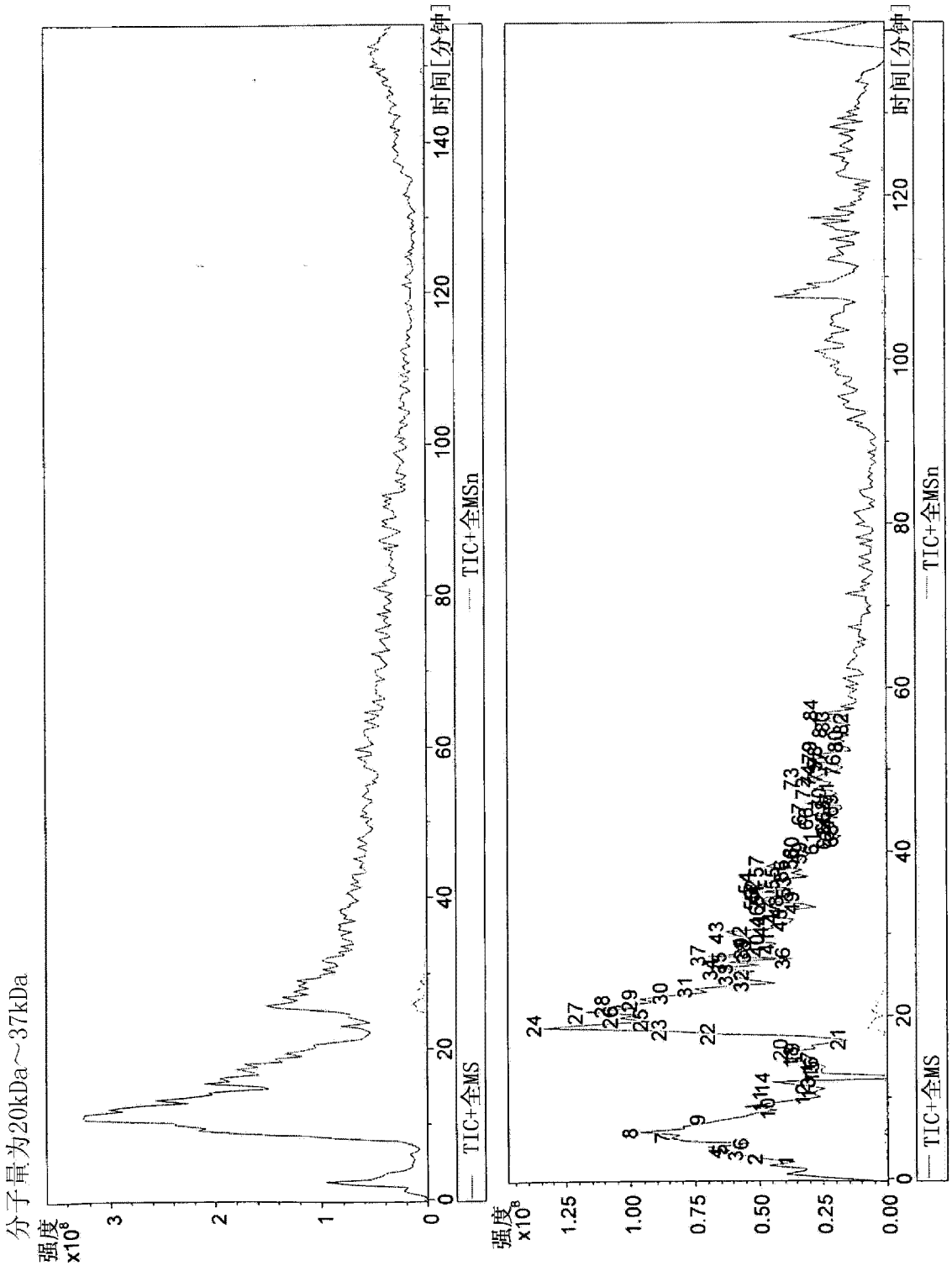


图 8 续

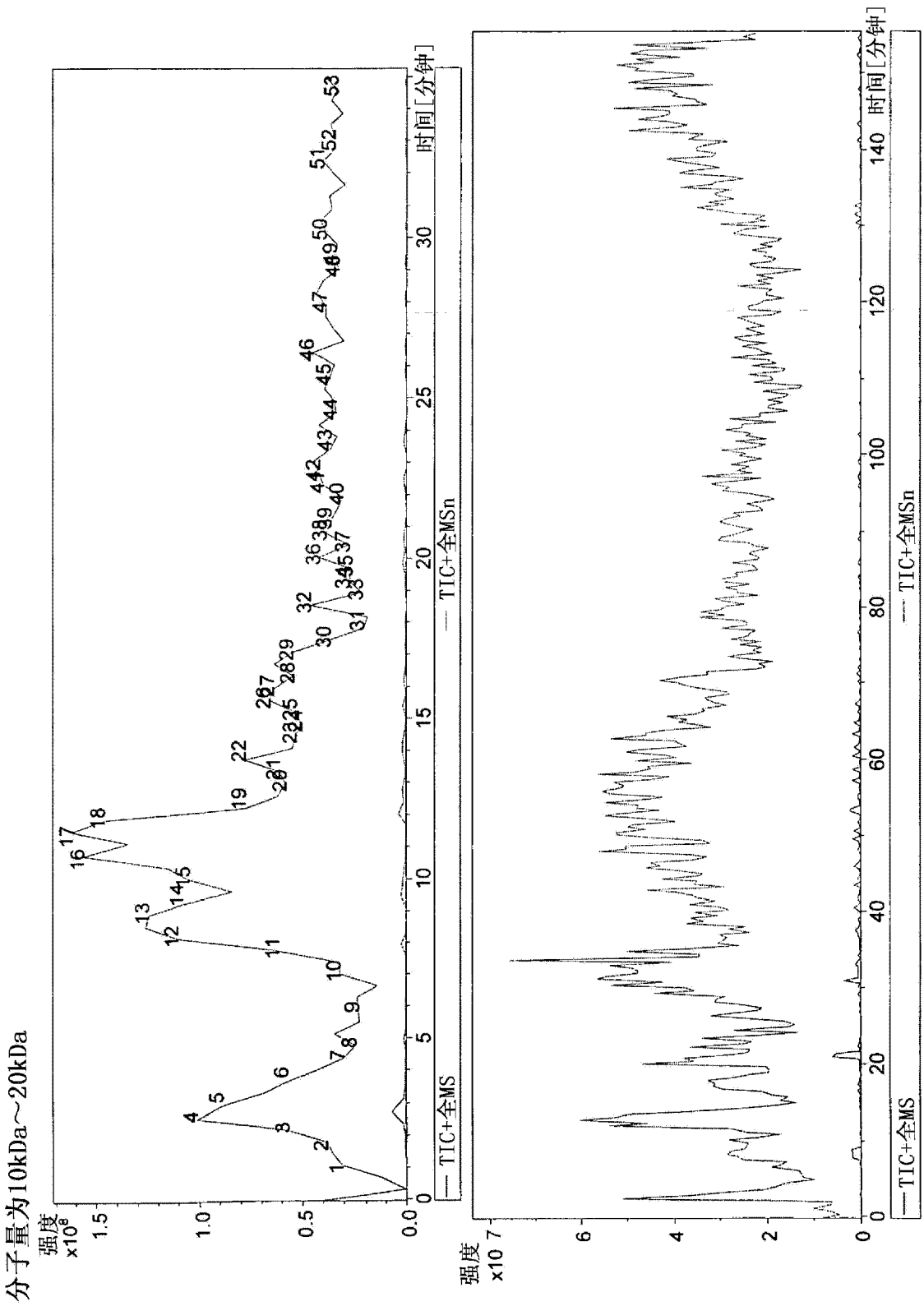

分子量为10kDa～20kDa
强度
x10⁸
1.5
1.0
0.5
0.0
0
5
10
15
20
25
30
时间[分钟]
TIC+全MS
TIC+全MSn
强度
x10 7
6
4
2
0
0
20
40
60
80
100
120
140
时间[分钟]
TIC+全MS
TIC+全MSn

图 8 续

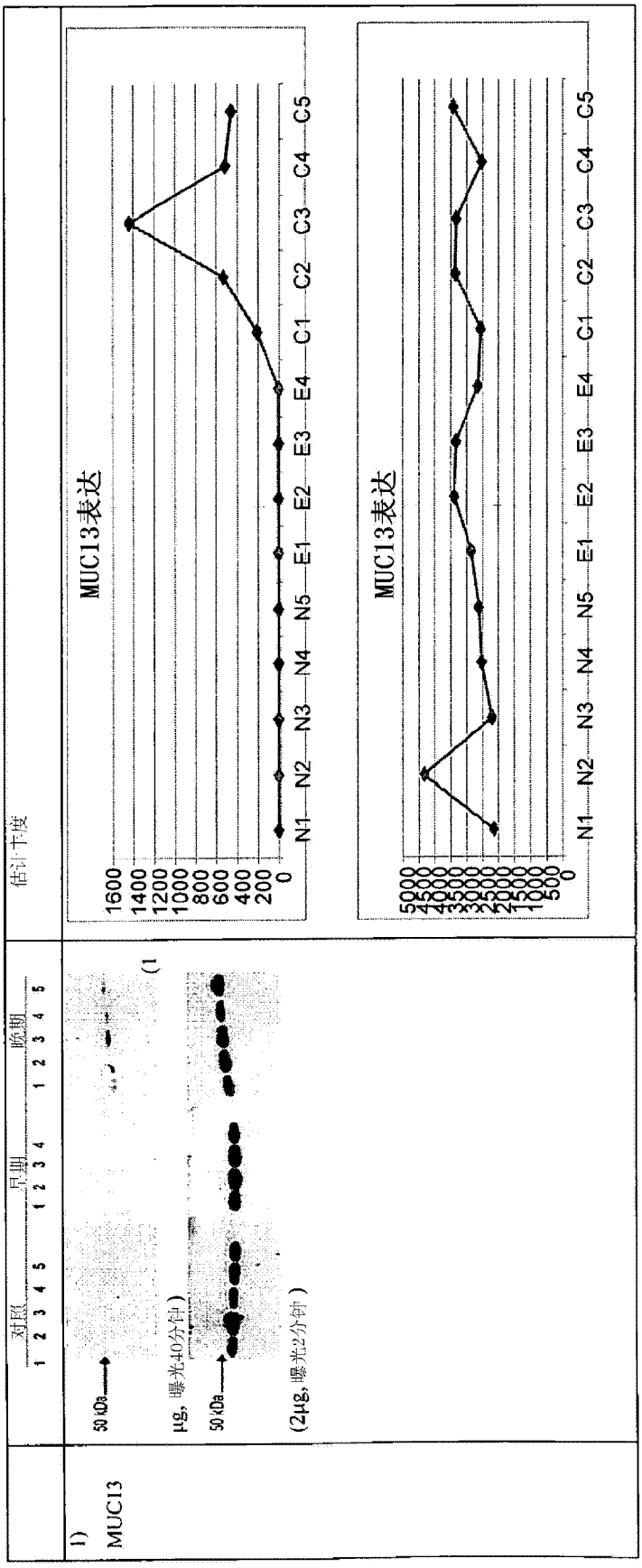

1)
MUC13
对照
早期
晚期
1 2 3 4 5 1 2 3 4 5
50 kDa
(1
μg，曝光40分钟)
50 kDa
(2μg，曝光2分钟)
估计干度
MUC13表达
1600
1400
1200
1000
800
600
400
200
0
N1 N2 N3 N4 N5 E1 E2 E3 E4 C1 C2 C3 C4 C5
MUC13表达
5000
4500
4000
3500
3000
2500
2000
1500
1000
500
0
N1 N2 N3 N4 N5 E1 E2 E3 E4 C1 C2 C3 C4 C5

图 9

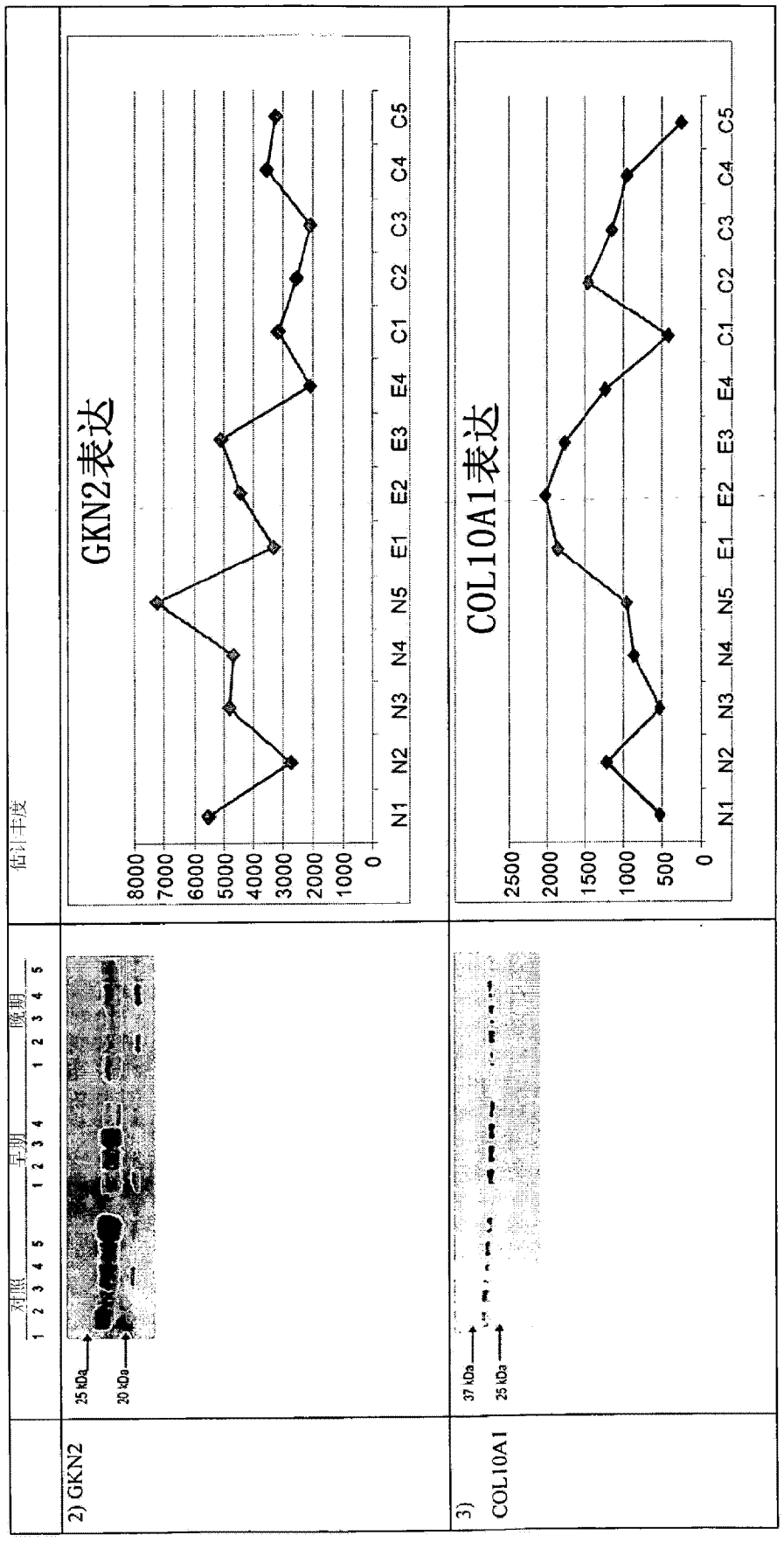

2) GKN2
3) COL10A1
对照
早期
晚期
1 2 3 4 5 1 2 3 4 5 1 2 3 4 5
25 kDa
20 kDa
37 kDa
25 kDa
估计丰度
GKN2表达
COL10A1表达
N1 N2 N3 N4 N5 E1 E2 E3 E4 C1 C2 C3 C4 C5
8000
7000
6000
5000
4000
3000
2000
1000
0
2500
2000
1500
1000
500
0

图 9 续

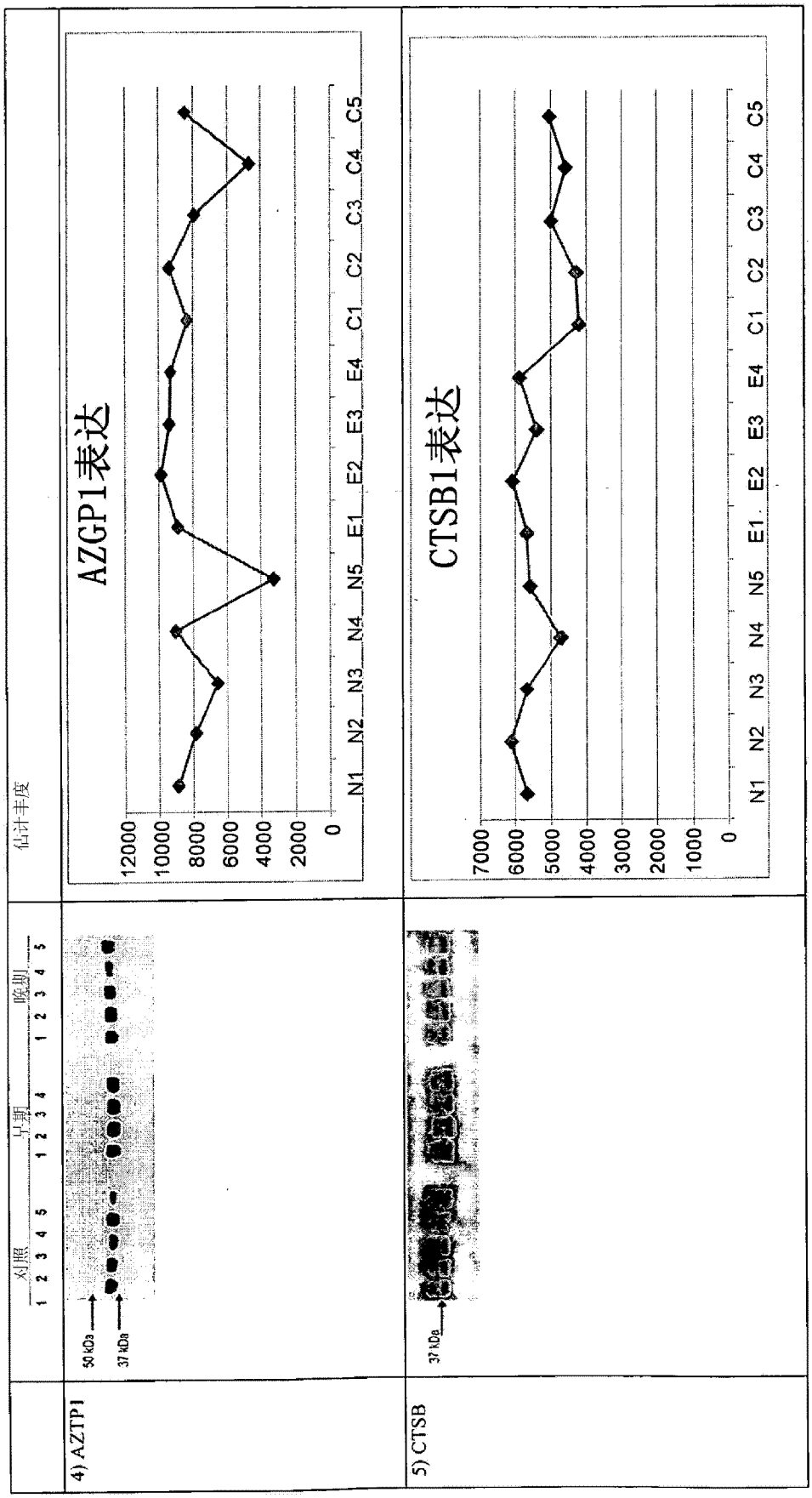

4) AZTP1
5) CTSB
对照
早期
晚期
1 2 3 4 5 1 2 3 4 1 2 3 4 5
50 kDa
37 kDa
37 kDa
仙计丰度
AZGP1表达
12000
10000
8000
6000
4000
2000
0
N1 N2 N3 N4 N5 E1 E2 E3 E4 C1 C2 C3 C4 C5
CTSB1表达
7000
6000
5000
4000
3000
2000
1000
0
N1 N2 N3 N4 N5 E1 E2 E3 E4 C1 C2 C3 C4 C5

图 9 续

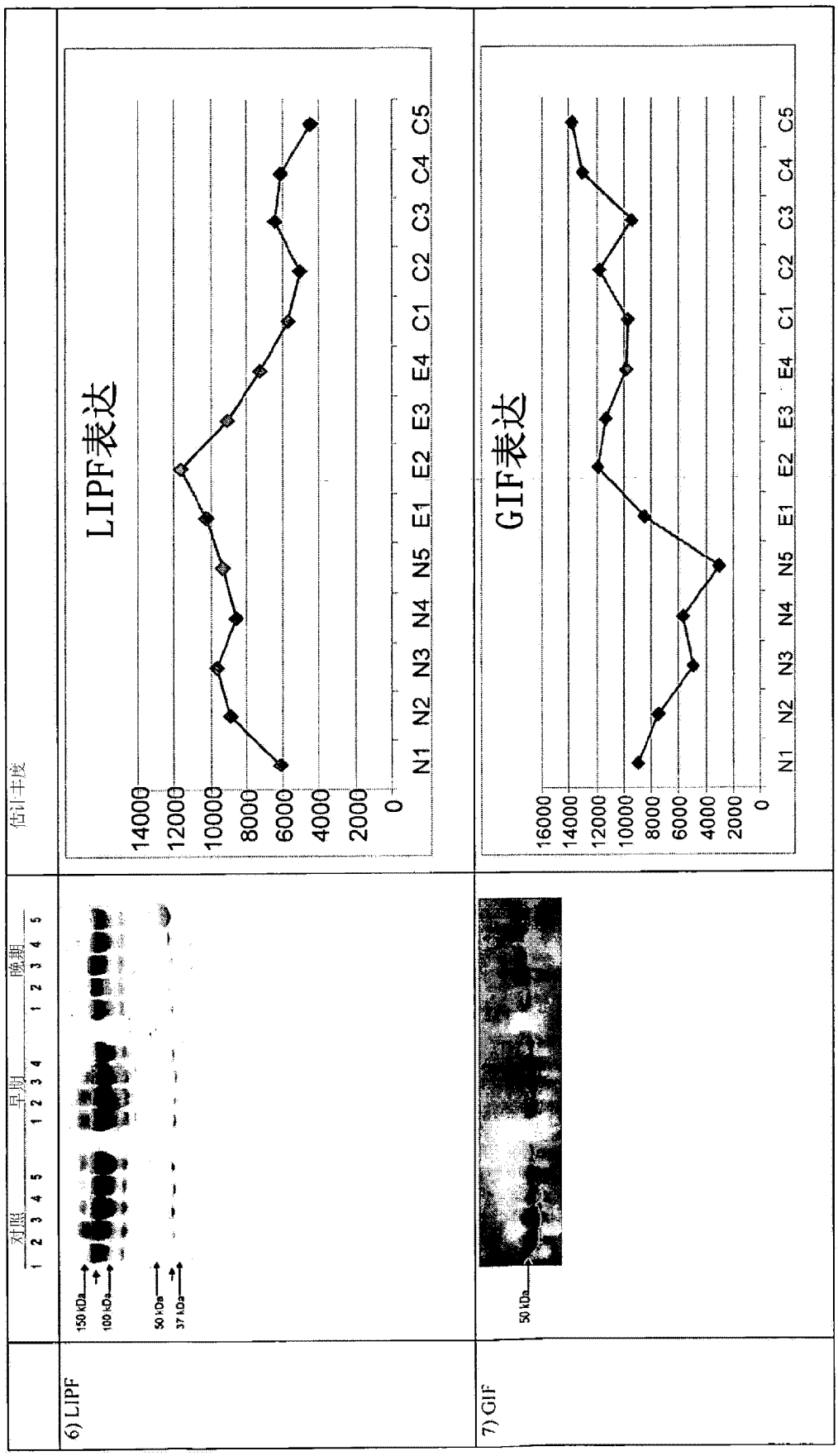

估计丰度
6) LIPF
对照
早期
晚期
1 2 3 4 5 1 2 3 4 1 2 3 4 5
150 kDa
100 kDa
50 kDa
37 kDa
LIPF表达
14000
12000
10000
8000
6000
4000
2000
0
N1 N2 N3 N4 N5 E1 E2 E3 E4 C1 C2 C3 C4 C5
7) GIF
50 kDa
GIF表达
16000
14000
12000
10000
8000
6000
4000
2000
0
N1 N2 N3 N4 N5 E1 E2 E3 E4 C1 C2 C3 C4 C5

图 9 续

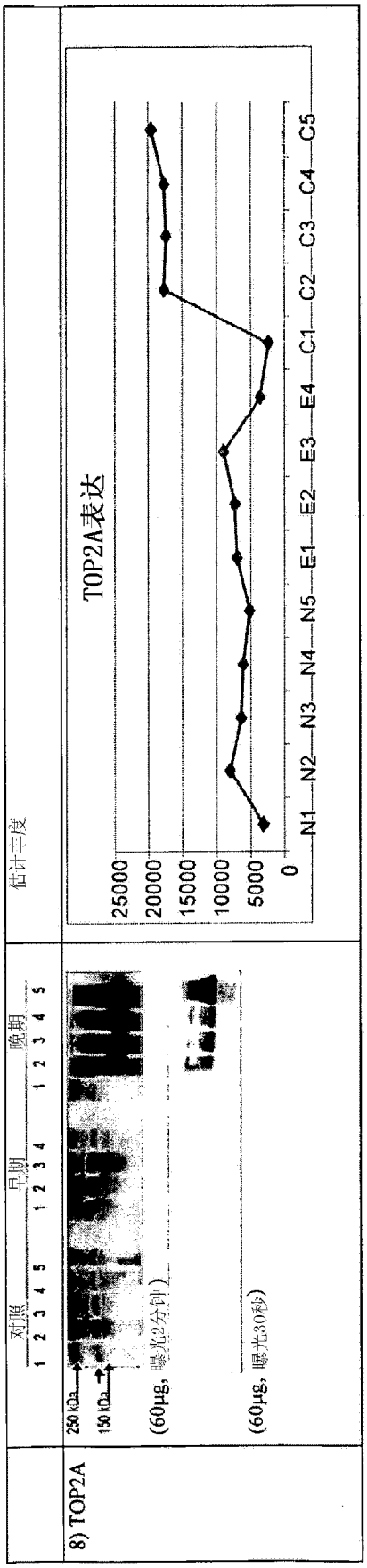

8) TOP2A
对照
早期
晚期
1 2 3 4 5 1 2 3 4 1 2 3 4 5
250 kDa
130 kDa
(60μg，曝光2分钟)
(60μg，曝光30秒)
估计丰度
TOP2A表达
25000
20000
15000
10000
5000
0
N1 N2 N3 N4 N5 E1 E2 E3 E4 C1 C2 C3 C4 C5

图 9 续

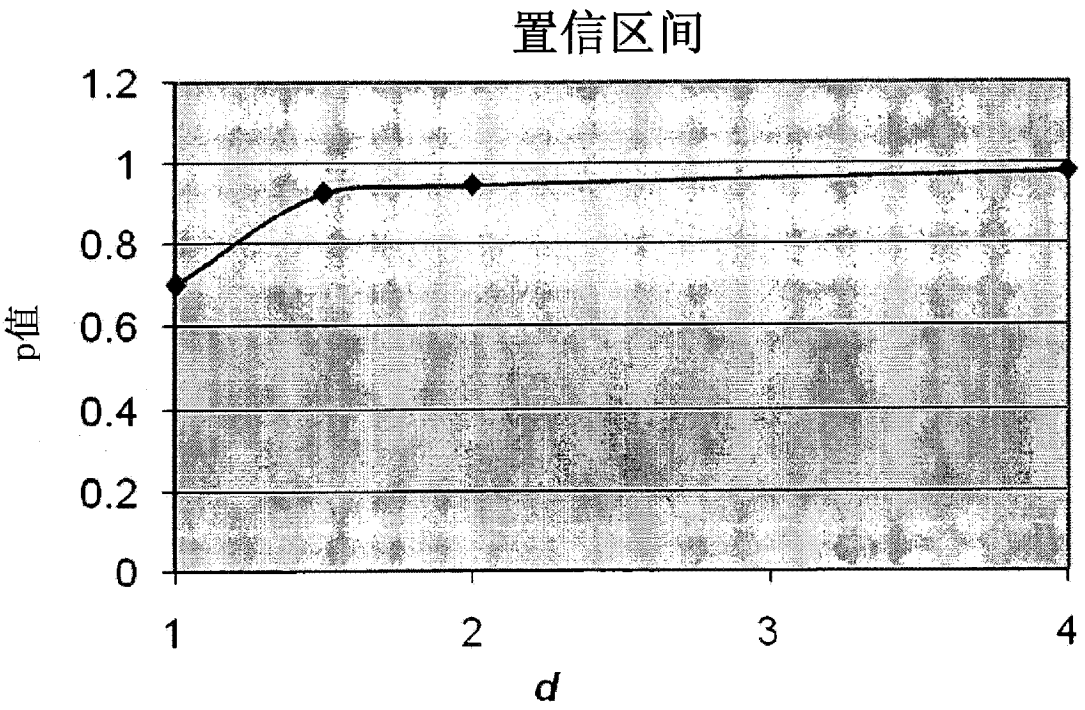

置信区间
1.2
1
0.8
0.6
0.4
0.2
0
p值
1
2
3
4
d

图 10

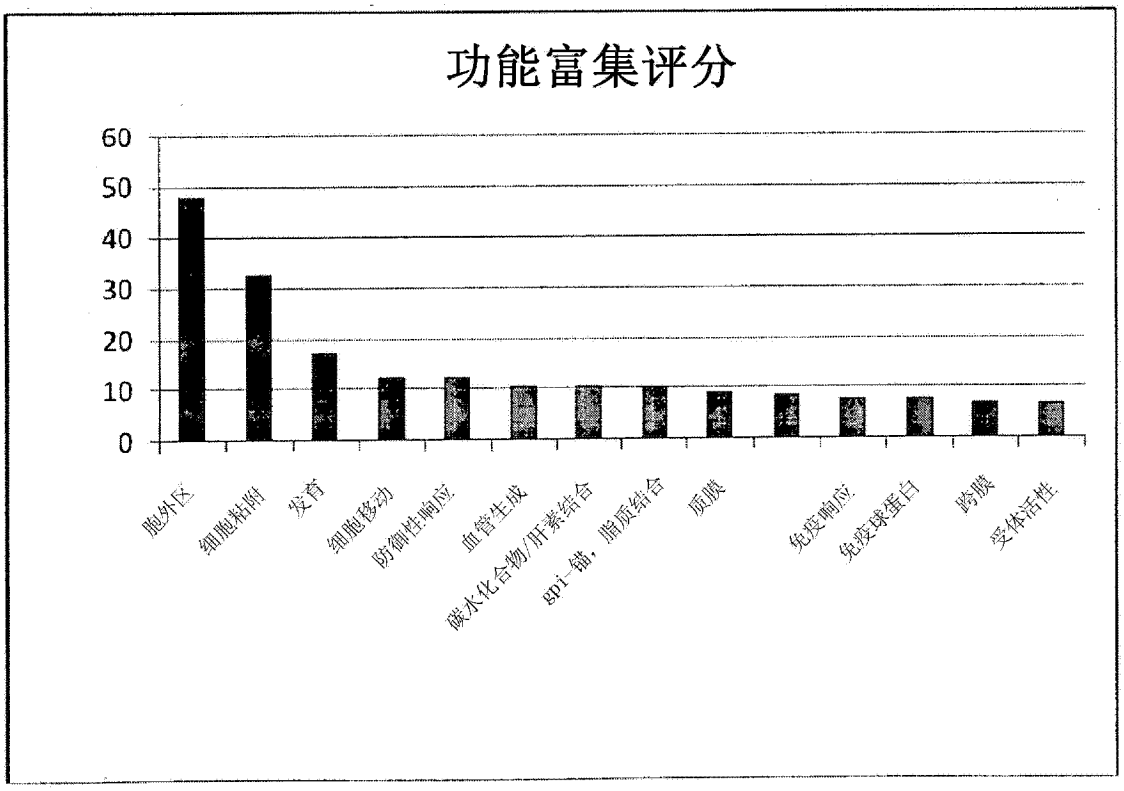

功能富集评分
60
50
40
30
20
10
0
胞外区
细胞粘附
发育
细胞移动
防御性响应
血管生成
碳水化合物/肝素结合
gpi-锚，脂质结合
质膜
免疫响应
免疫球蛋白
跨膜
受体活性

图 11

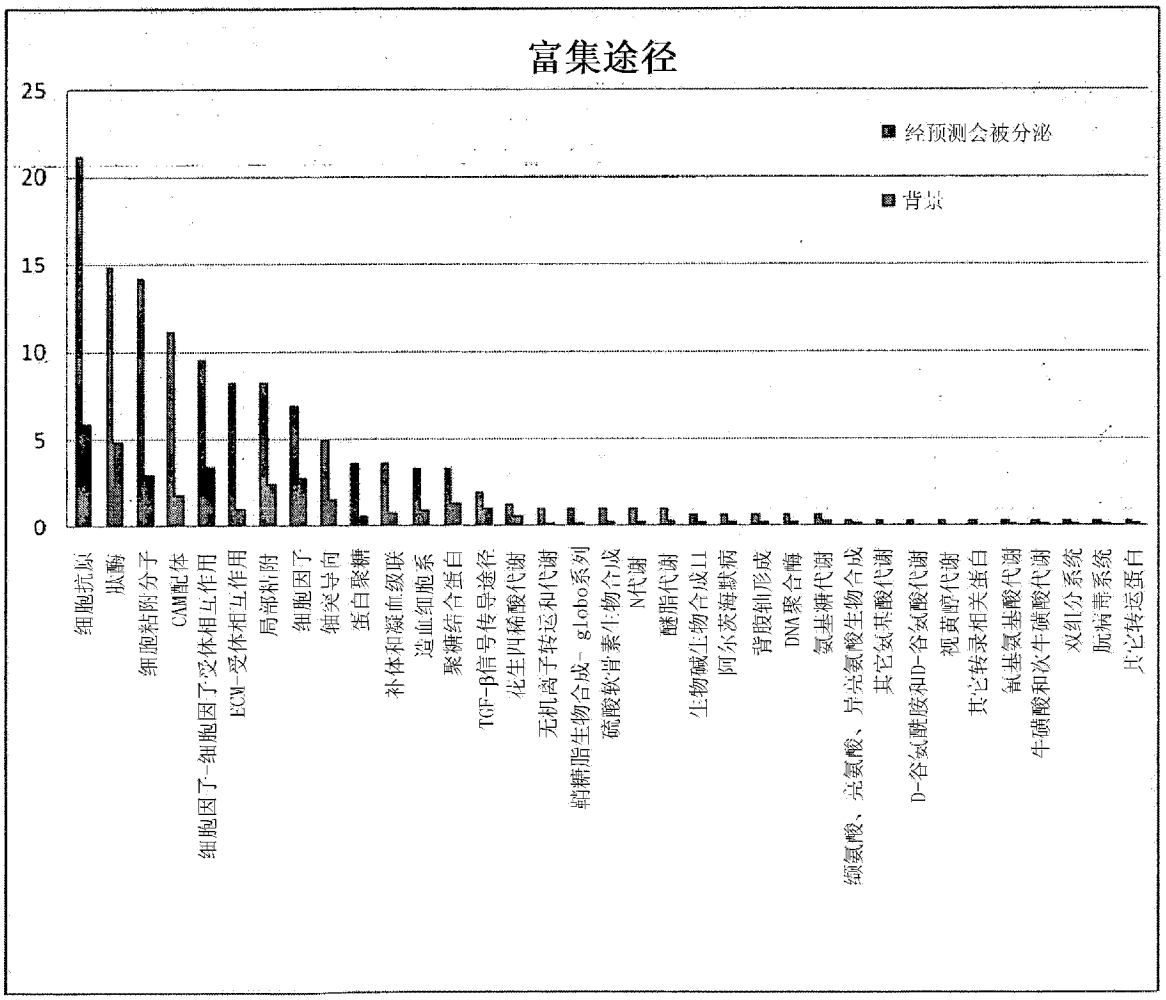

富集途径
25
20
15
10
5
0
经预测会被分泌
背景
细胞抗原
肽酶
细胞粘附分子
CAM配体
细胞因子-细胞因子受体相互作用
ECM-受体相互作用
局部粘附
细胞因子
轴突导向
蛋白聚糖
补体和凝血级联
造血细胞系
聚糖结合蛋白
TGF-β信号传导途径
花生四稀酸代谢
无机离子转运和代谢
鞘糖脂生物合成 - gLobo系列
硫酸软骨素生物合成
N代谢
醚脂代谢
生物碱生物合成11
阿尔茨海默病
背腹轴形成
DNA聚合酶
氨基糖代谢
缬氨酸、亮氨酸、异亮氨酸生物合成
其它氨基酸代谢
D-谷氨酰胺和D-谷氨酸代谢
视黄醇代谢
其它转录相关蛋白
氨基氨基酸代谢
牛磺酸和次牛磺酸代谢
双组分系统
脏病毒系统
其它转运蛋白

图 12

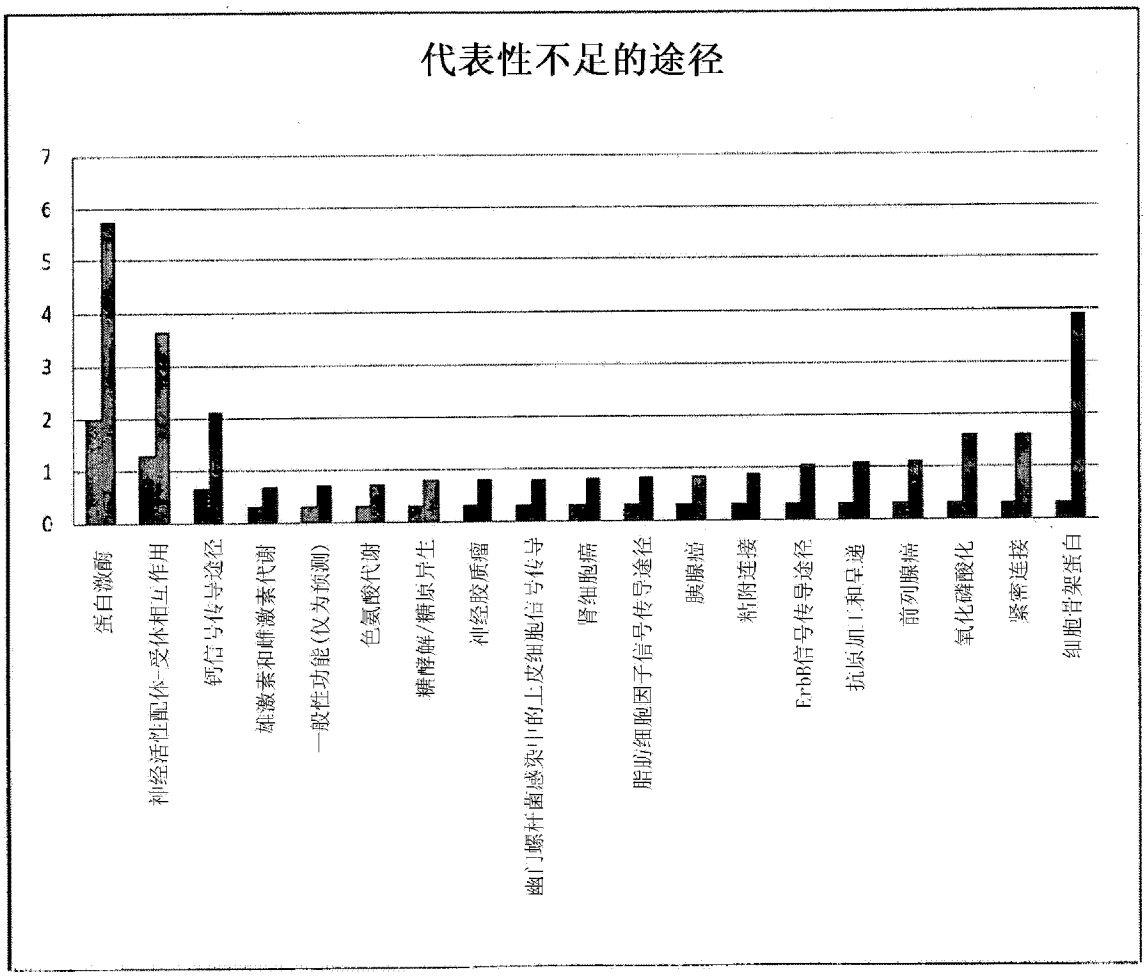

代表性不足的途径
7
6
5
4
3
2
1
0
蛋白激酶
神经活性配体-受体相互作用
钙信号传导途径
雄激素和雌激素代谢
一般性功能(仅为预测)
色氨酸代谢
糖酵解/糖原异生
神经胶质瘤
幽门螺杆菌感染中的上皮细胞信号传导
肾细胞癌
脂肪细胞因子信号传导途径
胰腺癌
粘附连接
ErbB信号传导途径
抗原加工和呈递
前列腺癌
氧化磷酸化
紧密连接
细胞骨架蛋白

图 13

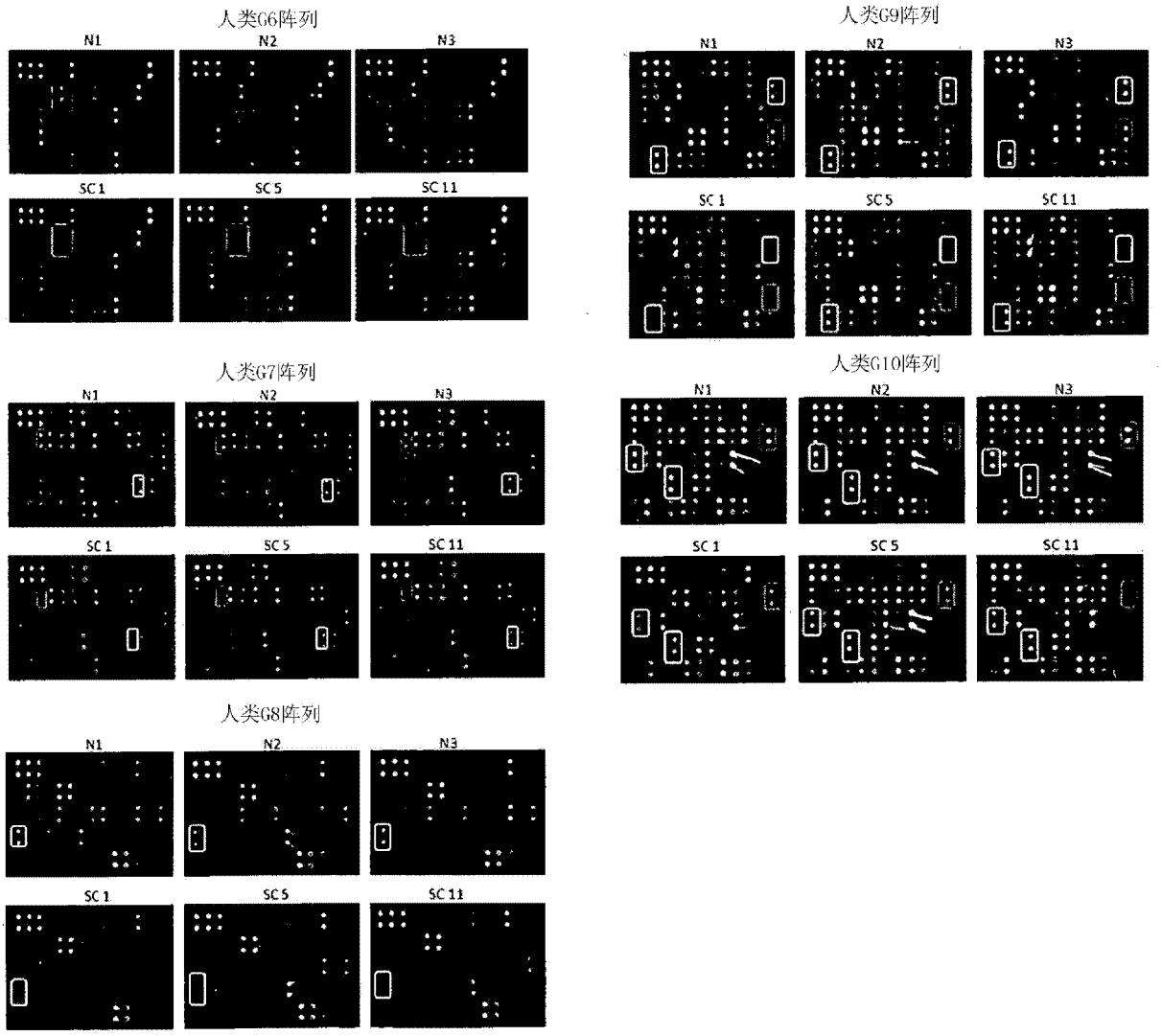

人类G6阵列
N1
N2
N3
SC1
SC5
SC11
人类G9阵列
N1
N2
N3
SC1
SC5
SC11
人类G7阵列
N1
N2
N3
SC1
SC5
SC11
人类G10阵列
N1
N2
N3
SC1
SC5
SC11
人类G8阵列
N1
N2
N3
SC1
SC5
SC11

图 14

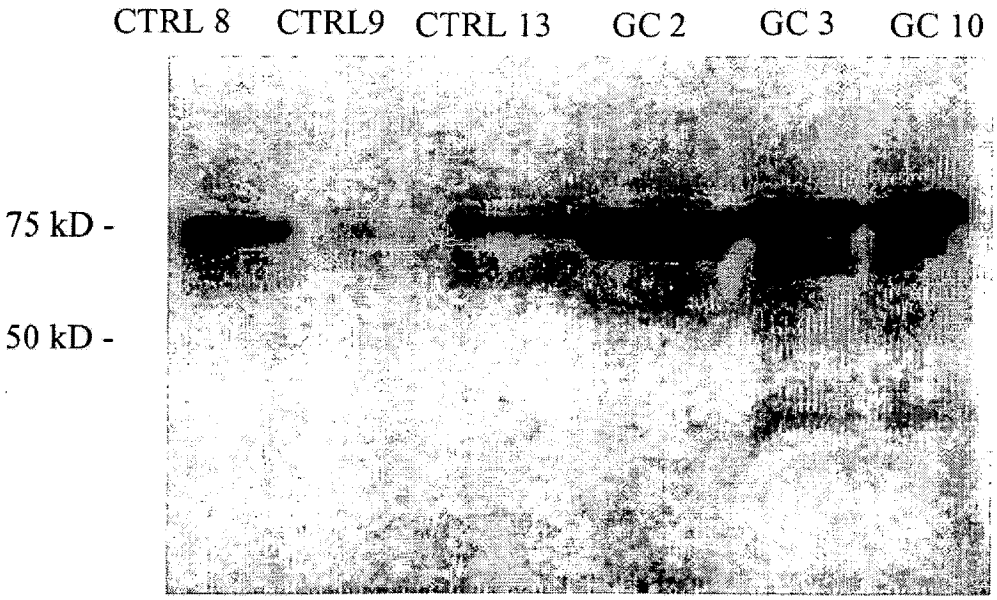

CTRL 8
CTRL9
CTRL 13
GC 2
GC 3
GC 10
75 kD -
50 kD -

图 15

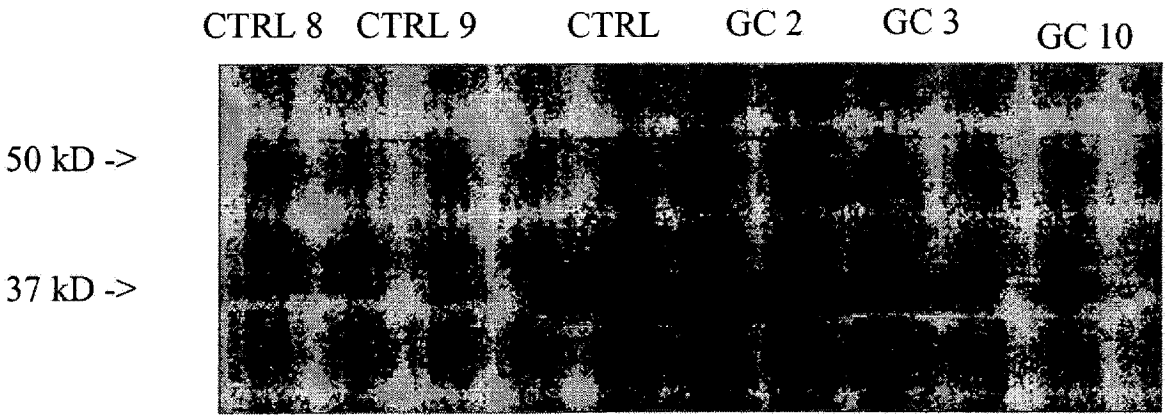

CTRL 8
CTRL 9
CTRL
GC 2
GC 3
GC 10
50 kD ->
37 kD ->

图 16

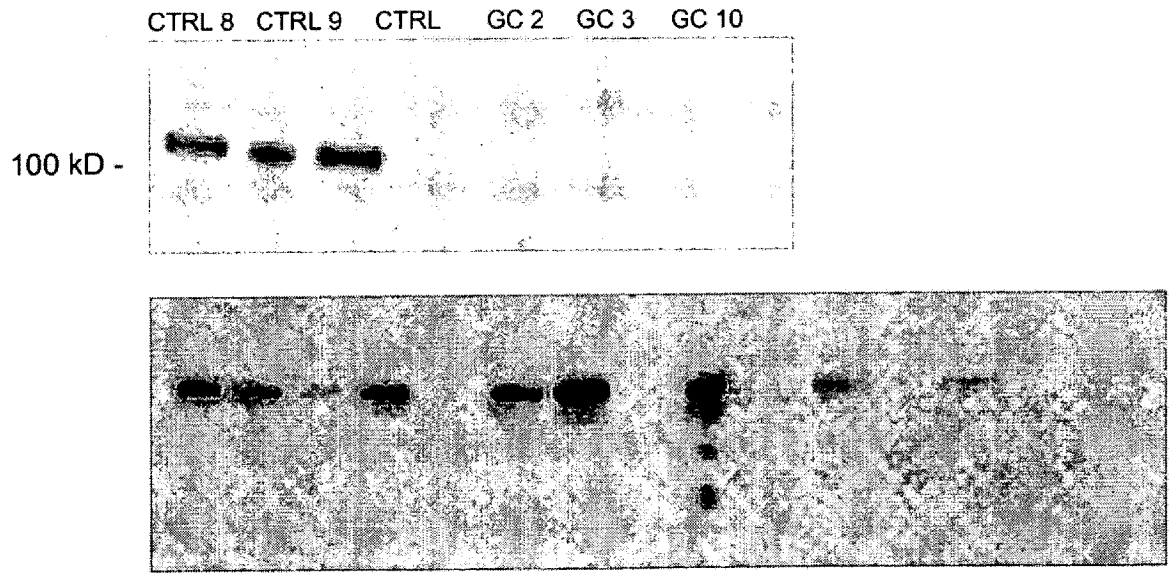

CTRL 8
CTRL 9
CTRL
GC 2
GC 3
GC 10
100 kD -

图 17

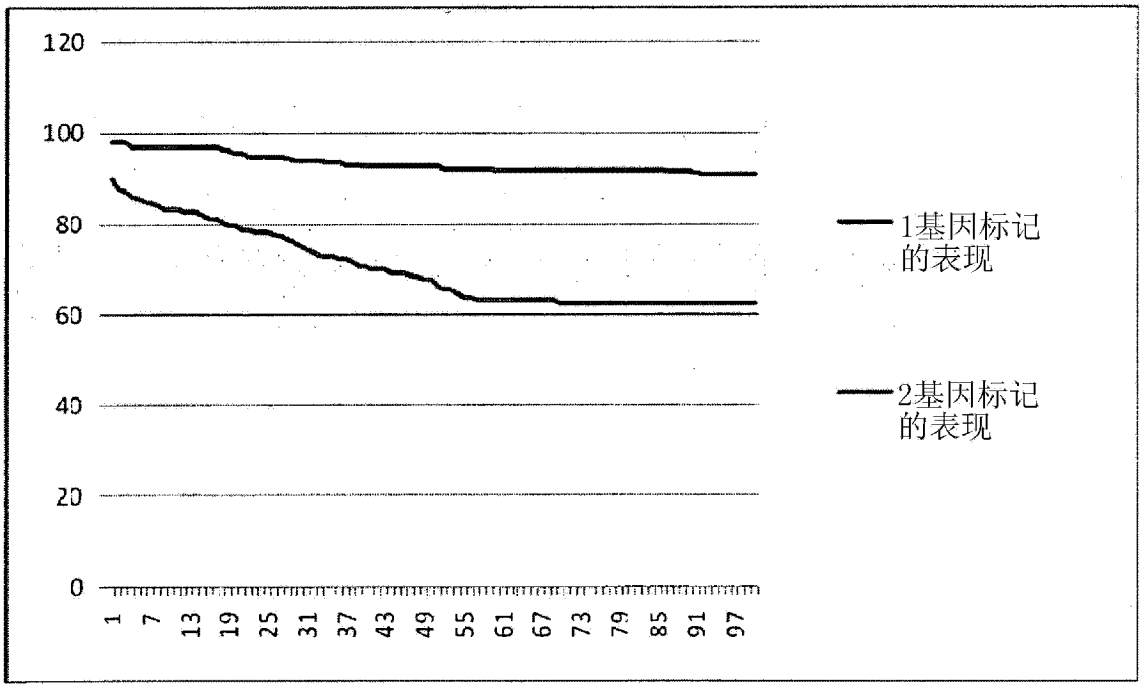

120
100
80
60
40
20
0
1
7
13
19
25
31
37
43
49
55
61
67
73
79
85
91
97
1基因标记的表现
2基因标记的表现

图 18

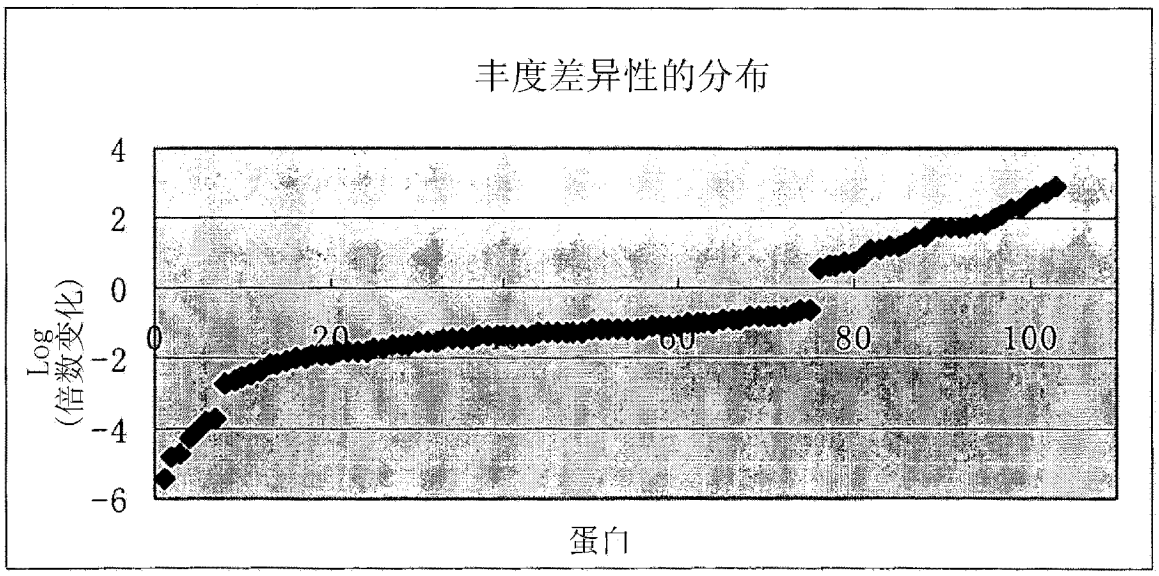


图 19

专利名称(译)	胃癌诊断用蛋白标记的鉴定		
公开(公告)号	CN102348979A	公开(公告)日	2012-02-08
申请号	CN201080011326.4	申请日	2010-02-19
[标]申请(专利权)人(译)	吉林大学		
申请(专利权)人(译)	乔治亚大学研究基金公司 吉林大学		
当前申请(专利权)人(译)	UNIV GEORGIA		
[标]发明人	崔娟 李凡 大卫普特 C洪 徐鹰		
发明人	崔娟 李凡 大卫·普特 C·洪 徐鹰		
IPC分类号	G01N33/48 G01N33/53		
CPC分类号	G01N33/57484		
优先权	61/241347 2009-09-10 US 61/158683 2009-03-09 US		
外部链接	Espacenet SIPO		

摘要(译)

本发明提供了通过检测分泌到生物液中的蛋白而检测癌的方法以及诊断癌的方法。首先将本发明应用于检测分泌到血清和尿中的蛋白。但是，应该理解的是，所述方法具有更广泛的应用，以开发用于检测分泌到其它生物液中的蛋白的工具和系统，所述其它生物液例如，但不限于唾液、脊髓液、精液、阴道液和眼内液。通过该方法的实施方式提供的对分泌到生物液中的蛋白进行的可信检测能够更及时准确地检测和诊断癌。