(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 108064273 A
(43)申请公布日 2018.05.22

(21)申请号 201580074801.5
(22)申请日 2015.01.30
(85)PCT国际申请进入国家阶段日
2017.07.21
(86)PCT国际申请的申请数据
PCT/CN2015/071896 2015.01.30
(87)PCT国际申请的公布数据
WO2016/119191 EN 2016.08.04
(71)申请人 深圳华大基因研究院
地址 518083 广东省深圳市盐田区北山工业区综合楼
(72)发明人 冯强 梁穗莎 贾慧珏 王俊
(74)专利代理机构 中国国际贸易促进委员会专利商标事务所 11038
代理人 林远成
(51)Int.Cl.
C12N 1/20(2006.01)
C12Q 1/04(2006.01)
C12Q 1/68(2006.01)

权利要求书6页 说明书48页
序列表（电子公布） 附图7页

(54)发明名称

结直肠癌相关疾病的生物标志物

(57)摘要

本发明提供了用于预测与微生物相关的疾病(特别是结直肠癌和结直肠中的进展性腺癌)的风险的生物标志物和方法。

CN 108064273 A

1.一种用于在受试者中预测或诊断与微生物群相关的疾病或确定受试者是否具有形成所述疾病的风险的生物标志物组,其包含以下生物标志物:

(1)MLG 5045或其一个或多个特异性片段,所述MLG 5045由SEQ ID NO:3732-3918组成;和

(2)MLG 121或其一个或多个特异性片段,所述MLG 121由SEQ ID NO:3919-6548组成;

任选地,所述生物标志物组还包含以下生物标志物中的一种或多种:

(3)MLG 75或其一个或多个特异性片段,所述MLG 75由SEQ ID NO:1350-1527组成;

(4)MLG 109或其一个或多个特异性片段,所述MLG 109由SEQ ID NO:6549-7235组成;

(5)MLG 317或其一个或多个特异性片段,所述MLG 317由SEQ ID NO:7581-7700组成;

(6)MLG 135或其一个或多个特异性片段,所述MLG 135由SEQ ID NO:2230-3731组成;

(7)MLG 223或其一个或多个特异性片段,所述MLG 223由SEQ ID NO:9892-11298组成;

(8)MLG 100或其一个或多个特异性片段,所述MLG 100由SEQ ID NO:9596-9891组成;

(9)MLG 219或其一个或多个特异性片段,所述MLG 219由SEQ ID NO:7701-8028组成;

(10)MLG 114或其一个或多个特异性片段,所述MLG 114由SEQ ID NO:1528-2089组成;

(11)MLG 84或其一个或多个特异性片段,所述MLG 84由SEQ ID NO:1-165组成;

(12)MLG 166或其一个或多个特异性片段,所述MLG 166由SEQ ID NO:7236-7580组成;

(13)MLG 2985或其一个或多个特异性片段,所述MLG 2985由SEQ ID NO:8029-9595组成;

(14)MLG 131或其一个或多个特异性片段,所述MLG 131由SEQ ID NO:166-1349组成;和

(15)MLG 1564或其一个或多个特异性片段,所述MLG 1564由SEQ ID NO:2090-2229组成。

优选地,所述生物标志物组包含如(1)-(13)中定义的生物标志物。

2.权利要求1的生物标志物组,其中所述特异性片段的长度为至少30bp,或至少40bp,或至少50bp,或至少60bp,或至少70bp,或至少80bp,或至少90bp,或至少100bp,或至少150bp,或至少200bp,或至少250bp,或至少300bp,或至少350bp,或至少400bp,或至少450bp,或至少500bp,或至少600bp,或至少700bp,或至少800bp,或至少900bp,或至少1000bp,或至少1500bp,或至少2000bp。

3.权利要求1或2的生物标志物组,其特征还在于以下项中的任一项或多项:

(1)所述MLG 84的一个或多个特异性片段选自SEQ ID NO:1-165或其任意组合;

(2)所述MLG 131的一个或多个特异性片段选自SEQ ID NO:166-1349或其任意组合;

(3)所述MLG 75的一个或多个特异性片段选自SEQ ID NO:1350-1527或其任意组合;

(4)所述MLG 114的一个或多个特异性片段选自SEQ ID NO:1528-2089或其任意组合;

(5)所述MLG 1564的一个或多个特异性片段选自SEQ ID NO:2090-2229或其任意组合;

(6)所述MLG 135的一个或多个特异性片段选自SEQ ID NO:2230-3731或其任意组合;

(7)所述MLG 5045的一个或多个特异性片段选自SEQ ID NO:3732-3918或其任意组合;

(8)所述MLG 121的一个或多个特异性片段选自SEQ ID NO:3919-6548或其任意组合;

(9)所述MLG 109的一个或多个特异性片段选自SEQ ID NO:6549-7235或其任意组合;

(10)所述MLG 166的一个或多个特异性片段选自SEQ ID NO:7236-7580或其任意组合;

(11)所述MLG 317的一个或多个特异性片段选自SEQ ID NO:7581-7700或其任意组合;

(12)所述MLG 219的一个或多个特异性片段选自SEQ ID NO:7701-8028或其任意组合;

(13)所述MLG 2985的一个或多个特异性片段选自SEQ ID NO:8029-9595或其任意组合;

(14)所述MLG 100的一个或多个特异性片段选自SEQ ID NO:9596-9891或其任意组合;和

(15)所述MLG 223的一个或多个特异性片段选自SEQ ID NO:9892-11298或其任意组合。

4.权利要求1-3中任一项的生物标志物组,其中所述疾病是结直肠癌。

5.权利要求1-4中任一项的生物标志物组,其中所述受试者是哺乳动物,诸如灵长类动物,优选为人。

6.权利要求1-5中任一项的生物标志物组,其中所述生物标志物组用于区分患有结直肠癌的患者与健康受试者和患有进展性腺瘤的患者。

7.一种用于在受试者中预测或诊断与微生物群相关的疾病,或确定受试者是否具有形成所述疾病的风险的试剂盒,其包含用于测定权利要求1-6中任一项的生物标志物组的每种生物标志物在样品中的水平或其量的试剂。

8.权利要求7的试剂盒,其中所述试剂选自:

(a)引物组,其包含:

(a1)能够特异性扩增MLG 5045或其一个或多个特异性片段的一种或多种引物,所述MLG 5045由SEQ ID NO:3732-3918组成;和

(a2)能够特异性扩增MLG 121或其一个或多个特异性片段的一种或多种引物,所述MLG 121由SEQ ID NO:3919-6548组成;

任选地,所述引物组还包含以下引物的一种或多种:

(a3)能够特异性扩增MLG 75或其一个或多个特异性片段的一种或多种引物,所述MLG 75由SEQ ID NO:1350-1527组成;

(a4)能够特异性扩增MLG 109或其一个或多个特异性片段的一种或多种引物,所述MLG 109由SEQ ID NO:6549-7235组成;

(a5)能够特异性扩增MLG 317或其一个或多个特异性片段的一种或多种引物,所述MLG 317由SEQ ID NO:7581-7700组成;

(a6)能够特异性扩增MLG 135或其一个或多个特异性片段的一种或多种引物,所述MLG 135由SEQ ID NO:2230-3731组成;

(a7)能够特异性扩增MLG 223或其一个或多个特异性片段的一种或多种引物,所述MLG 223由SEQ ID NO:9892-11298组成;

(a8)能够特异性扩增MLG 100或其一个或多个特异性片段的一种或多种引物,所述MLG 100由SEQ ID NO:9596-9891组成;

(a9)能够特异性扩增MLG 219或其一个或多个特异性片段的一种或多种引物,所述MLG 219由SEQ ID NO:7701-8028组成;

(a10)能够特异性扩增MLG 114或其一个或多个特异性片段的一种或多种引物,所述MLG 114由SEQ ID NO:1528-2089组成;

(a11)能够特异性扩增MLG 84或其一个或多个特异性片段的一种或多种引物,所述MLG 84由SEQ ID NO:1-165组成;

(a12)能够特异性扩增MLG 166或其一个或多个特异性片段的一种或多种引物,所述MLG 166由SEQ ID NO:7236-7580组成;

(a13)能够特异性扩增MLG 2985或其一个或多个特异性片段的一种或多种引物,所述MLG 2985由SEQ ID NO:8029-9595组成;

(a14)能够特异性扩增MLG 131或其一个或多个特异性片段的一种或多种引物,所述MLG 131由SEQ ID NO:166-1349组成;和

(a15)能够特异性扩增MLG 1564或其一个或多个特异性片段的一种或多种引物,所述MLG 1564由SEQ ID NO:2090-2229组成;

优选地,所述引物组包含如(a1)-(a13)中定义的引物;

(b)探针组,其包含:

(b1)能够与MLG 5045或其一个或多个特异性片段特异性杂交的一种或多种探针,所述MLG 5045由SEQ ID NO:3732-3918组成;和

(b2)能够与MLG 121或其一个或多个特异性片段特异性杂交的一种或多种探针,所述MLG 121由SEQ ID NO:3919-6548组成;

任选地,所述探针组还包含以下探针的一种或多种:

(b3)能够与MLG 75或其一个或多个特异性片段特异性杂交的一种或多种探针,所述MLG 75由SEQ ID NO:1350-1527组成;

(b4)能够与MLG 109或其一个或多个特异性片段特异性杂交的一种或多种探针,所述MLG 109由SEQ ID NO:6549-7235组成;

(b5)能够与MLG 317或其一个或多个特异性片段特异性杂交的一种或多种探针,所述MLG 317由SEQ ID NO:7581-7700组成;

(b6)能够与MLG 135或其一个或多个特异性片段特异性杂交的一种或多种探针,所述MLG 135由SEQ ID NO:2230-3731组成;

(b7)能够与MLG 223或其一个或多个特异性片段特异性杂交的一种或多种探针,所述MLG 223由SEQ ID NO:9892-11298组成;

(b8)能够与MLG 100或其一个或多个特异性片段特异性杂交的一种或多种探针,所述MLG 100由SEQ ID NO:9596-9891组成;

(b9)能够与MLG 219或其一个或多个特异性片段特异性杂交的一种或多种探针,所述MLG 219由SEQ ID NO:7701-8028组成;

(b10)能够与MLG 114或其一个或多个特异性片段特异性杂交的一种或多种探针,所述MLG 114由SEQ ID NO:1528-2089组成;

(b11)能够与MLG 84或其一个或多个特异性片段特异性杂交的一种或多种探针,所述MLG 84由SEQ ID NO:1-165组成;

(b12)能够与MLG 166或其一个或多个特异性片段特异性杂交的一种或多种探针,所述MLG 166由SEQ ID NO:7236-7580组成;

(b13)能够与MLG 2985或其一个或多个特异性片段特异性杂交的一种或多种探针,所述MLG 2985由SEQ ID NO:8029-9595组成;

(b14)能够与MLG 131或其一个或多个特异性片段特异性杂交的一种或多种探针，所述MLG 131由SEQ ID NO:166-1349组成；和

(b15)能够与MLG 1564或其一个或多个特异性片段特异性杂交的一种或多种探针，所述MLG 1564由SEQ ID NO:2090-2229组成；

优选地，所述探针组包含如(b1)-(b13)中定义的探针；

(c)包含(a)的引物组和/或(b)的探针组的微阵列；

(d)进行第二代测序方法或第三代测序方法的试剂；和

(e)(a)-(d)的任意组合。

9.权利要求7或8的试剂盒，其中所述试剂盒通过包括以下步骤的方法，在受试者中预测或诊断与微生物群相关的疾病，或确定受试者是否具有形成所述疾病的风险：

(1)使用所述试剂盒来测定来自所述受试者的样品中的根据权利要求1至6中任一项的生物标志物组的每种生物标志物的水平或其量；和

(2a)通过使用多元统计模型(诸如随机森林模型)将所述样品中的每种生物标志物的水平或其量与训练数据集进行比较来计算所述疾病的概率；

优选地，所述训练数据集包含关于多个患有所述疾病的受试者和多个健康受试者的每种生物标志物的水平或其量的数据；

其中当所述疾病的概率大于临界值时，表明所述受试者患有所述疾病或具有形成所述疾病的风险；或

(2b)将所述样品中的每个生物标志物的水平或其量与所述对照中的相应生物标志物的水平或其量进行比较；优选地，所述对照是多个健康受试者；

其中，与所述对照相比，当所述样品中所述生物标志物的水平或其量升高时，表明所述受试者患有所述疾病或具有形成所述疾病的风险。

10.权利要求9的试剂盒，其中所述训练数据集包含表4-1和4-2中的数据，并且当所述疾病的概率大于临界值0.5时，表明所述受试者患有所述疾病或具有形成所述疾病的风险。

11.权利要求7-10中任一项的试剂盒，其中所述受试者是哺乳动物，例如灵长类动物，优选为人。

12.权利要求7-11中任一项的试剂盒，其中所述样品是粪便样品。

13.权利要求7-12中任一项的试剂盒，其中所述每种生物标志物的水平或其量是所述样品中每种生物标志物的相对丰度。

14.权利要求7-13中任一项的试剂盒，其中所述疾病是结直肠癌。

15.权利要求7-14中任一项的试剂盒，其中所述试剂盒还包含另外的试剂，诸如用于处理所述样品的试剂(例如无菌水)，用于进行PCR扩增的试剂(例如聚合酶、dNTP和扩增缓冲液)，以及用于进行杂交的试剂(诸如标记缓冲液、杂交缓冲液和洗涤缓冲液)。

16.用于测定权利要求1至6中任一项的生物标志物组的每种生物标志物的水平或其量的试剂在制备试剂盒中的用途，所述试剂盒用于在受试者中预测或诊断与微生物群相关的疾病或确定受试者是否具有形成所述疾病的风险。

17.权利要求16的用途，其中所述试剂是如权利要求8中所定义的。

18.权利要求16或17的用途，其中所述试剂盒通过包括以下步骤的方法，在受试者中预测或诊断与微生物群相关的疾病，或确定受试者是否具有形成所述疾病的风险：

(1)使用所述试剂盒测定来自所述受试者的样品中的根据权利要求1-6中任一项的生物标志物组的每种生物标志物的水平或其量;和

(2a)通过使用多元统计模型(诸如随机森林模型)将所述样品中的每种生物标志物的水平或其量与训练数据集进行比较来计算所述疾病的概率;

优选地,所述训练数据集包含关于多个患有所述疾病的受试者和多个健康受试者的每种生物标志物的水平或其量的数据;

其中当所述疾病的概率大于临界值时,表明所述受试者患有所述疾病或具有形成所述疾病的风险;或

(2b)将所述样品中的每个生物标志物的水平或其量与所述对照中的相应生物标志物的水平或其量进行比较;优选地,所述对照是多个健康受试者;

其中,与所述对照相比,当所述样品中所述生物标志物的水平或其量升高时,表明所述受试者患有所述疾病或具有形成所述疾病的风险。

19.权利要求18的用途,其中所述训练数据集包含表4-1和4-2中的数据,并且当所述疾病的概率大于临界值0.5时,表明所述受试者患有所述疾病或具有形成所述疾病的风险。

20.权利要求16-19中任一项的用途,其中所述试剂者是哺乳动物,例如灵长类动物,优选为人。

21.权利要求16-20中任一项的用途,其中所述样品是粪便样品。

22.权利要求16-21中任一项的用途,其中所述每种生物标志物的水平或其量是所述样品中每种生物标志物的相对丰度。

23.权利要求16-22中任一项的用途,其中所述疾病是结直肠癌。

24.权利要求16-23中任一项的用途,其中所述试剂盒还包含另外的试剂,诸如用于处理所述样品的试剂(例如无菌水),用于进行PCR扩增的试剂(例如聚合酶、dNTP和扩增缓冲液),以及用于进行杂交的试剂(诸如标记缓冲液、杂交缓冲液和洗涤缓冲液)。

25.一种用于在受试者中预测或诊断与微生物群相关的疾病或确定受试者是否具有形成所述疾病的风险的方法,其包括以下步骤:

(1)测定来自所述受试者的样品中的根据权利要求1至6中任一项所述的生物标志物组的每种生物标志物的水平或其量;和

(2a)通过使用多元统计模型(诸如随机森林模型)将所述样品中的每个生物标志物的水平或其量与训练数据集进行比较来计算所述疾病的概率;

优选地,所述训练数据集包含关于多个患有所述疾病的受试者以及多个健康受试者的每种生物标志物的水平或其量的数据;

其中当所述疾病的概率大于临界值时,表明所述受试者患有所述疾病或具有形成所述疾病的风险;或

(2b)将所述样品中的每个生物标志物的水平或其量与所述对照中的相应生物标志物的水平或其量进行比较;优选地,所述对照是多个健康受试者;

其中,与所述对照相比,当所述样品中所述生物标志物的水平或其量升高时,表明所述受试者患有所述疾病或具有形成所述疾病的风险。

26.权利要求25的方法,其中在步骤(1)中,使用权利要求7-14任一项所述的试剂盒或如权利要求8中定义的试剂。

27.权利要求25或26的方法，其中所述训练数据集包含表4-1和4-2中的数据，并且当所述疾病的概率大于临界值0.5时，表明所述受试者患有所述疾病或具有形成所述疾病的风险。

28.权利要求25-27的方法，其中所述受试者是哺乳动物，例如灵长类动物，优选为人。

29.权利要求25-28的方法，其中所述样品是粪便样品。

30.权利要求25-29的方法，其中所述每种生物标志物的水平或其量是所述样品中每种生物标志物的相对丰度。

31.权利要求25-30的方法，其中所述疾病是结直肠癌。

32.权利要求25-31的方法，其中在体外进行所述方法。

结直肠癌相关疾病的生物标志物

[0001] 相关申请的交叉引用

[0002] 无

[0003] 领域

[0004] 本发明涉及用于预测与微生物相关的疾病，特别是结直肠癌和结直肠中的进展性腺瘤的风险的生物标志物和方法。

[0005] 背景

[0006] 结直肠癌(CRC)是全球前三位最常诊断的癌症之一，是癌症死亡的主要原因。其在较发达国家的发病率较高，但在诸如东亚、西班牙和东欧等历史低风险地区，由于所谓的西方生活方式，发病率正在迅速上升。在结直肠癌的发展中，遗传学改变累积了多年，通常涉及肿瘤抑制基因腺瘤性结肠息肉病基因(APC)的丧失，以及随后分别发生在KRAS、PIK3CA和TP53中的激活或失活突变(Brenner,H.,Kloor,M.&Pox,C.P.Colorectal cancer.Lancet383,1490-502(2014)，通过引用并入本文)。虽然大多数CRC病例是散发性的，但在其出现之前通常发生异常腺瘤，该异常腺瘤可进展为恶性形式，这称为腺瘤-癌顺序。结直肠腺瘤和结直肠癌的早期诊断不仅有助于防止死亡，而且也有助于降低手术干预的费用。

[0007] CRC是研究最多的与肠道微生物群相关的疾病之一。然而，该疾病的因果关系通常通过施用抗生素混合剂疗法来研究，所述抗生素混合剂疗法清除肠道微生物群而无法获知起作用的确切微生物菌株和基因。相比于正常结肠组织，在结直肠癌中检测到了梭杆菌属(Fusobacterium)，并且发现其富集在腺瘤中。具核梭杆菌(Fusobacterium nucleatum)(一种牙周病病原体)，被发现能够促进$Apc^{Min/+}$小鼠中肠道肿瘤的骨髓浸润，并与小鼠和人中的促炎基因诸如Ptgs2(COX-2)、Scyb1(IL8)、Il6、Tnf(TNFα)和Mmp3的表达上调相关(Kostic,A.D.等,Fusobacterium nucleatum potentiates intestinal tumorigenesis and modulates the tumor-immune microenvironment.Cell Host Microbe14,207-215(2013)，通过引用并入本文)。然而，目前尚不清楚，是否有更多的细菌或古细菌可作为结直肠癌病因的标志物或促成病因。

[0008] 目前CRC的检测，诸如可屈性乙状结肠镜检查和结肠镜检查，都是侵入性的，并且患者可能会在该检测过程和肠道准备过程中感到不舒服或不愉快。肠道微生物群与免疫系统之间的相互作用在肠内和肠外的许多疾病中具有重要作用(Cho,I.&Blaser,M.J.The human microbiome:at the interface of health and disease.Nature Rev.Genet.13,260-270(2012)，通过引用并入本文)。粪便DNA的肠道微生物群分析有潜力被用作无创性检测，以发现可用作CRC患者早期诊断的筛选工具的特异性生物标志物，从而获得更长的生存时间和更好的生活质量。

[0009] 概要

[0010] 本公开内容的实施方案旨在至少一定程度地解决现有技术中存在的至少一个问题。

[0011] 本发明基于发明人的以下发现：

[0012] 肠道微生物群的评估和表征已成为人类疾病(包括结直肠癌)的主要研究领域。本发明人首次针对来自健康对照、结直肠癌和腺瘤患者的粪便样品进行宏基因组全基因组鸟枪法测序。为了对结直肠癌和腺瘤患者中的肠道微生物含量进行分析,本发明人进行了宏基因组关联分析(Metagenome-Wide Association Study)(MGWAS)方案(Qin,J.等A metagenome-wide association study of gut microbiota in type 2diabetes.Nature 490,55-60(2012),通过引用并入本文)。为了比较健康对照组、进展性腺瘤组和癌症患者组的粪便微生物群系,鉴定了相对丰度在任意两组之间展现出显著差异的基因($p<0.05$,Kruskal-Wallis检验)。随后根据其在所有样品中的丰度变化,这些标记基因被聚类形成MLG(宏基因组连锁群)(Qin等,2012,同上),并且本发明人鉴定了这些肿瘤的MLG特征。然后本发明人鉴定并验证了15个用于结直肠癌的早期和无创性诊断的MLG,以及10个用于结直肠腺瘤的早期和无创性诊断的MLG。为了利用这些基于肠道微生物群的CRC分类的潜力,本发明人通过分别基于15个MLG和10个MLG的随机森林模型计算了疾病的概率。本发明人的数据为表征与CRC风险相关的肠道宏基因组提供了具有洞察力的见解,也为以后研究肠道宏基因组在其它相关病症的病理生理学中的作用提供了一个范例,同时还揭示了基于肠道-微生物群的方法用于评估处于此类病症风险中的个体的潜在用途。

[0013] 据信,上述15个MLG和10个MLG对于改善CRC的早期检测具有重要价值,原因如下。第一,与常规标志物相比,本发明的标志物更特异和灵敏。第二,粪便分析具备准确性、安全性、经济可承受性和患者依从性。粪便样品是可运输的。与需要肠道制备的结肠镜检查相比,本发明涉及舒适且无创的体外方法,因此人们更容易参与给定的筛查程序。第三,本发明的标志物也可用作CRC患者的治疗监测工具,以检测对治疗的反应。

[0014] 因此,在第一方面,本发明提供了用于在受试者中预测或诊断与微生物群相关的疾病或确定受试者是否具有形成所述疾病的风险的生物标志物组。

[0015] 在第二方面,本发明提供了用于在受试者中预测或诊断与微生物群相关的疾病,或确定受试者是否具有形成所述疾病的风险的试剂盒,其包含用于测定样品中的本发明的生物标志物组的每种生物标志物的水平或其量的试剂。

[0016] 在第三方面,本发明提供了用于测定本发明的生物标志物组的每种生物标志物的水平或其量的试剂在制备试剂盒中的用途,所述试剂盒用于在受试者中预测或诊断与微生物群相关的疾病或用于确定受试者是否具有形成所述疾病的风险。

[0017] 在第四方面,本发明提供了用于在受试者中预测或诊断与微生物群相关的疾病或确定受试者是否具有形成所述疾病的风险的方法。

附图说明

[0018] 根据以下描述,并结合附图,本公开内容的各个方面及优点将变得明显并且易于理解,其中:

[0019] 图1示出了肠道MLG能够将结直肠癌样品与健康对照样品进行分类。(a)随着MLG数量的增加,癌的随机森林分类中5次10折交叉验证的误差的分布情况。使用MLG(>100个基因)在对照和癌症样品($n=55$和41)中的相对丰度训练该模型。黑色曲线表示5次验证的平均值(灰线)。黑色直线标示最优集中的MLG数目(15个MLG)(表2-1,表2-2,表3)。即使将年龄和BMI因素与MLG一起考虑,仍然筛选得到相同的MLG。(b)根据(a)中的模型的交叉验证训练

集中的癌的概率的盒须图(box-and-whisker plot)。(c)训练集的接受者工作曲线(ROC)。在临界值(cut-off)为0.5时,AUC为98.34%,95%的置信区间(CI)为96.29-100%。(d)由8个对照样品(黑色方块)、47个进展性腺瘤样品(空心圆)和5个癌症样品(实心黑圈)组成的测试集的分类结果。(e)测试集的ROC。在临界值为0.5时,AUC为96%,95%的CI为87.88-100%。如果癌的概率≥0.5,则该受试者处于患结直肠癌的风险中。图1的结果表明,上述15个MLG可用作诊断结直肠癌和/或确定患结直肠癌的风险的生物标志物,且具备高灵敏度和高特异性。

[0020] 图2示出了肠道MLG能够将进展性腺瘤样品与健康对照样品进行分类。(a)随着MLG数量增加,进展性腺瘤的随机森林分类中5次10折交叉验证的误差的分布情况。使用MLG(>100个基因)在对照组和进展性腺瘤样品(n=55和42)中的相对丰度训练该模型。黑色曲线表示5次验证的平均值(灰线)。黑色直线标示最优集中的MLG数目(10个MLG)(表7-1,表7-2)。即使将年龄和BMI因素与MLG一起考虑,仍然筛选得到相同的MLG。(b)根据(a)中的模型的交叉验证训练集中的进展性腺瘤的概率的盒须图(box-and-whisker plot)。(c)训练集的接受者工作曲线(ROC)。在临界值为0.5时,AUC为87.38%,95%的置信区间(CI)为80.21-94.55%。(d)由15个对照样品(空心圆)和15个进展性腺瘤样品(实心黑圈)组成的测试集的分类结果。(e)测试集的ROC。在最佳临界值为为0.4572时,AUC为90.67%,真阳性率(TPR)为1,假阳性率(FPR)为0.2667。如果结直肠腺瘤的概率≥0.4572(最佳临界值),则该受试者处于患结直肠腺瘤的风险中。图2的结果表明,上述10个MLG可用作诊断进展性腺瘤和/或确定患进展性腺瘤的风险的生物标志物,且具备高灵敏度和高特异性。

[0021] 详述

[0022] 本文使用的术语具有本发明相关领域的普通技术人员通常理解的含义。然而,为了更好地理解本发明,相关术语的定义和解释如下。

[0023] 术语诸如“一个/一种(a)”、“一个/一种(an)”和“该(the)”并不旨在仅指单个实体,而且还包括可以使用具体示例来说明的一个种类。

[0024] 根据本发明,术语“生物标志物”(也称为“生物学标志物”),是指受试者的生物学状态或状况的可测量指标。此类生物标志物可以是受试者中的任何物质,例如核酸标志物(例如DNA)、蛋白质标志物、细胞因子标志物、趋化因子标志物、糖类标志物、抗原标志物、抗体标志物、物种标记(种/属标志物)和功能标志物(KO/OG标志物)等,只要它们与受试者的特定生物学状态或状况(例如疾病)相关。通常通过测量和评估生物标志物以检测正常生物过程、病理过程或对治疗干预的药理学应答,并且生物标志物在许多科学领域中都是有用的。

[0025] 根据本发明,术语“生物标志物组”是指一组生物标志物(即,两种或更多种生物标志物的组合)。

[0026] 根据本发明,术语“与微生物群相关的疾病”是指与肠道中的微生物群的失衡相关的疾病。例如,所述疾病可由肠道中的微生物群的失衡引起、诱发或加剧。这种疾病可以是结直肠癌的进展性腺瘤或结直肠恶性肿瘤/癌。

[0027] 根据本发明,术语“受试者”是指动物,特别是哺乳动物,诸如灵长类动物,优选人。

[0028] 根据本发明,表述“结直肠癌(colorectal cancer)”具有与“结直肠癌(colorectal carcinoma)”相同的含义。

[0029] 根据本发明，表述“进展性腺瘤”和“结直肠进展性腺瘤”具有与“结直肠癌中的进展性腺瘤”相同的含义。

[0030] 根据本发明，表述“临界值(cutoff)”和“临界值(cut-off)”具有相同的含义，是指预测的临界值。可以通过常规实验(例如通过平行检测来自已知生理状态的受试者的样品中的生物标志物的相对丰度)获得该预测的临界值。

[0031] 根据本发明，术语“MLG”被定义为宏基因组中的一组遗传物质，其在物理上可能连接形成一个单元而不是独立分布(参见，Qin,J.等A metagenome-wide association study of gut microbiota in type 2diabetes.Nature 490,55-60(2012)，其全部内容通过引用并入本文)。MLG使得不再需要完全确定存在于宏基因组中的特定微生物种类，这一点是非常重要的，因为目前还存在大量未知生物并且细菌之间存在频繁的侧向基因转移(LGT)。在本发明中，MLG是指具有一致丰度水平和分类学分配的一组基因。

[0032] 根据本发明，术语MLG的“特异性片段”是MLG的一个片段，其对于该MLG是独特的。可使用常规方法来确定片段对于其所源自的MLG是否是独特的。例如，可将该片段的序列输入公共数据库(诸如GenBank)并执行BLAST程序。如果该片段仅存在于数据库中的一个物种中(在这种情况下，这个MLG将代表或对应该物种)，或者如果数据库中不存在与所述片段具有至少90％同一性(诸如95％同一性)的同源物(在这种情况下，该MLG将指未知物种)，则该片段可被认为是独特的。如上所论述的，一个MLG通常是指一个特定的微生物物种(已知或未知的)，因此MLG的“特异性片段”也可被认为是特定微生物物种的一个独特的基因组片段(即，该片段仅存在于特定微生物物种中)。

[0033] 根据本发明，术语“同一性”是指两个多肽之间或两个核酸之间的匹配度。当用于比较的两个序列在某一位点具有相同的碱基或氨基酸单体亚单元时(例如，两个DNA分子的每一个中在某个位点都为腺嘌呤，或者两个多肽的每一个中的某一位点都为赖氨酸)，所述两个分子在该位点是同一的。两个序列之间的百分比同一性是由两个序列共有的匹配位点的数目除以用于比较的位点总数×100的函数。例如，如果两个序列的10个位点中有6个匹配，则这两个序列具有60％的同一性。例如，DNA序列:CTGACT和CAGGTT共有50％的同一性(6个位点中有3个匹配)。通常，以产生最大同一性的方式进行两个序列的比较。这种比对可通过使用基于Needleman等人(J.Mol.Biol.48:443-453,1970)的方法的计算机程序(诸如Align程序(DNAstar,Inc.))来进行。

[0034] 根据本发明，表述“用于测定生物标志物的水平或其量的试剂”是指可用于定量或测量样品中的生物标志物的水平或其量的试剂。基于本发明所提供的生物标志物的序列，这样的试剂可通过本领域公知的常规方法容易地设计或获得。例如，这样的试剂包括但不限于，可用于通过例如实时PCR来定量或测量生物标志物的水平或其量的PCR引物；可用于通过例如定量Southern印迹来定量或测量生物标志物的水平或其量的探针；可用于定量或测量生物标志物的水平或其量的微阵列(例如，基因芯片)等。另外，如本领域已知的，第二代测序方法或第三代测序方法也可用于定量或测量生物标志物的水平或其量。因此，这样的试剂也可以是可商购的用于进行第二代测序方法或第三代测序方法的试剂。

[0035] 根据本发明，表述“能够特异性扩增”特定核酸或特定序列的引物是指当用于扩增(例如PCR扩增)时，所述引物与所述特定核酸或序列特异性退火，以及产生独特的扩增产物(即，不与其它核酸或序列退火或产生其他副产物)。

[0036] 根据本发明，表述“能够与特定核酸或特定序列特异性杂交的探针”是指当在严格条件下用于杂交或检测时，所述探针与所述特定核酸杂交酸或序列特异性退火并与其杂交，但不与其它核酸或序列退火或与其杂交。

[0037] 基于特定序列(诸如特定MLG或其特异性片段)设计所述引物或探针，是本领域技术人员的公知常识。例如，此类公知常识可见于各种教科书(参见例如，J.Sambrook等，Molecular Cloning:Laboratory Manual，第二版，Cold Spring Harbor Laboratory Press，1989；F.M.Ausubel等，Short Protocols in Molecular Biology，第三版，John Wiley&Sons，Inc.；以及许多论文，如Buck等(1999)，Lowe等(1990)，等等。

[0038] 根据本发明，术语“第二代测序方法”是指近些年开发的新一代DNA测序方法，包括例如Illumina GA，Roche 454，ABI Solid；并且与传统的测序方法(诸如，Sanger测序方法)不同。第二代测序方法与传统测序方法(诸如，Sanger测序方法)的区别在于第二代测序方法通过边合成边测序的方式来分析DNA序列。第二代测序方法具有以下有利方面：1)成本低，为传统测序方法成本的1％；2)高通量，能够同时对多个样品进行测序，并且一次Solexa测序即可产生约500亿(50G)碱基的数据；3)高精度(大于98.4％)，有效解决了多重重复序列读出的问题。另一方面，当要测序的序列的数量已被预先确定时，高测序通量又提高了序列的测序深度(例如，每个序列可被测序多次)，从而确保测序结果的可信性。

[0039] 根据本发明，术语“第三代测序方法”是指最近开发的新一代单分子测序技术。第三代测序技术提供优于当前测序技术的有利方面，包括(i)更高的通量；(ii)更短的周转时间(例如在数分钟内以高倍覆盖度测序后生动物基因组)；(iii)更长的测序长度以增强从头组装(de novo assembly)，并使得能够直接检测单体型(haplotypes)和甚至全染色体定相(whole chromosome phasing)；(iv)更高的一致准确度，以使得能够进行稀有变异检测；(v)少量起始材料(理论上只需要单个分子即可进行测序)；和(vi)低成本，其中以低于100美元的价格实现对人类基因组的高倍覆盖度测序已成为社会的合理目标。关于第三代测序方法的更多细节，参见例如，Eric E.Schadt等，A window into third-generation sequencing，Human Molecular Genetics，2010，第19卷，Review Issue 2，R227-R240，通过引用并入本文。

[0040] 根据本发明，术语“相对丰度”具有本领域已知的常规含义，并且可通过本领域已知的方法计算。例如，可通过Qin，J.等A metagenome-wide association study of gut microbiota in type 2diabetes.Nature 490，55-60(2012)(通过引用并入本文)所公开的方法来测定或计算基因(即生物标志物)或MLG的相对丰度。

[0041] 本领域技术人员将理解，提供上述术语定义以更好地理解本发明，但上述术语定义无意限定本发明，除了如权利要求中所概述的外。

[0042] 在第一方面，本发明提供了用于在受试者中预测或诊断与微生物群相关的疾病或确定受试者是否具有形成所述疾病的风险的生物标志物组，其包含以下生物标志物(所有生物标志物列于表3中)：

[0043] (1)MLG 5045或其一个或多个特异性片段，所述MLG 5045由SEQ ID NO:3732-3918组成；和

[0044] (2)MLG 121或其一个或多个特异性片段，所述MLG 121由SEQ ID NO:3919-6548组成；

[0045] 任选地，所述生物标志物组还包含以下生物标志物中的一种或多种：

[0046] (3)MLG 75或其一个或多个特异性片段，所述MLG 75由SEQ ID NO:1350-1527组成；

[0047] (4)MLG 109或其一个或多个特异性片段，所述MLG 109由SEQ ID NO:6549-7235组成；

[0048] (5)MLG 317或其一个或多个特异性片段，所述MLG 317由SEQ ID NO:7581-7700组成；

[0049] (6)MLG 135或其一个或多个特异性片段，所述MLG 135由SEQ ID NO:2230-3731组成；

[0050] (7)MLG 223或其一个或多个特异性片段，所述MLG 223由SEQ ID NO:9892-11298组成；

[0051] (8)MLG 100或其一个或多个特异性片段，所述MLG 100由SEQ ID NO:9596-9891组成；

[0052] (9)MLG 219或其一个或多个特异性片段，所述MLG 219由SEQ ID NO:7701-8028组成；

[0053] (10)MLG 114或其一个或多个特异性片段，所述MLG 114由SEQ ID NO:1528-2089组成；

[0054] (11)MLG 84或其一个或多个特异性片段，所述MLG 84由SEQ ID NO:1-165组成；

[0055] (12)MLG 166或其一个或多个特异性片段，所述MLG 166由SEQ ID NO:7236-7580组成；

[0056] (13)MLG 2985或其一个或多个特异性片段，所述MLG 2985由SEQ ID NO:8029-9595组成；

[0057] (14)MLG 131或其一个或多个特异性片段，所述MLG 131由SEQ ID NO:166-1349组成；和

[0058] (15)MLG 1564或其一个或多个特异性片段，所述MLG 1564由SEQ ID NO:2090-2229组成。

[0059] 在优选实施方案中，本发明的生物标志物组包含如(1)-(13)中定义的生物标志物。

[0060] 如本领域技术人员已知的，特异性片段可具有任何长度，只要这样的片段对于其所源自的MLG或由该MLG表示的物种是独特的(即，片段不存在于其它MLG或其它物种中)。然而，为方便起见，所述特异性片段的长度可以是至少30bp，或至少40bp，或至少50bp，或至少60bp，或至少70bp，或至少80bp，或至少90bp，或至少100bp，或至少150bp，或至少200bp，或至少250bp，或至少300bp，或至少350bp，或至少400bp，或至少450bp，或至少500bp，或至少600bp，或至少700bp，或至少800bp，或至少900bp，或至少1000bp，或至少1500bp，或至少2000bp。

[0061] 例如，在优选实施方案中，本发明的生物标志物组还可由以下项中的任一项或多项表征：

[0062] (1)所述MLG 84的一个或多个特异性片段选自SEQ ID NO:1-165或其任意组合；

[0063] (2)所述MLG 131的一个或多个特异性片段选自SEQ ID NO:166-1349或其任意组

合；

[0064] (3)所述MLG 75的一个或多个特异性片段选自SEQ ID NO:1350-1527或其任意组合；

[0065] (4)所述MLG 114的一个或多个特异性片段选自SEQ ID NO:1528-2089或其任意组合；

[0066] (5)所述MLG 1564的一个或多个特异性片段选自SEQ ID NO:2090-2229或其任意组合；

[0067] (6)所述MLG 135的一个或多个特异性片段选自SEQ ID NO:2230-3731或其任意组合；

[0068] (7)所述MLG 5045的一个或多个特异性片段选自SEQ ID NO:3732-3918或其任意组合；

[0069] (8)所述MLG 121的一个或多个特异性片段选自SEQ ID NO:3919-6548或其任意组合；

[0070] (9)所述MLG 109的一个或多个特异性片段选自SEQ ID NO:6549-7235或其任意组合；

[0071] (10)所述MLG 166的一个或多个特异性片段选自SEQ ID NO:7236-7580或其任意组合；

[0072] (11)所述MLG 317的一个或多个特异性片段选自SEQ ID NO:7581-7700或其任意组合；

[0073] (12)所述MLG 219的一个或多个特异性片段选自SEQ ID NO:7701-8028或其任意组合；

[0074] (13)所述MLG 2985的一个或多个特异性片段选自SEQ ID NO:8029-9595或其任意组合；

[0075] (14)所述MLG 100的一个或多个特异性片段选自SEQ ID NO:9596-9891或其任意组合；和

[0076] (15)所述MLG 223的一个或多个特异性片段选自SEQ ID NO:9892-11298或其任意组合。

[0077] 在优选实施方案中，所述疾病是结直肠中的结直肠癌。

[0078] 在优选实施方案中，所述受试者是哺乳动物，诸如灵长类动物，优选为人。

[0079] 在优选实施方案中，本发明的生物标志物组用于区分患有结直肠癌的患者与健康受试者和患有进展性腺瘤的患者。

[0080] 在第二方面，本发明提供了试剂盒，所述试剂盒用于在受试者中预测或诊断与微生物群相关的疾病，或确定受试者是否具有处于形成所述疾病的风险，其包含用于测定根据本发明的生物标志物组的每种生物标志物在样品中的水平或其量的试剂。

[0081] 在优选实施方案中，用于测定所述生物标志物组的每种生物标志物的水平或其量的试剂选自：

[0082] (a)引物组，其包含：

[0083] (a1)能够特异性扩增MLG 5045或其一个或多个特异性片段的一种或多种引物，所述MLG 5045由SEQ ID NO:3732-3918组成；和

[0084] (a2)能够特异性扩增MLG 121或其一个或多个特异性片段的一种或多种引物，所述MLG 121由SEQ ID NO:3919-6548组成；

[0085] 任选地，所述引物组还包含以下引物的一种或多种：

[0086] (a3)能够特异性扩增MLG 75或其一个或多个特异性片段的一种或多种引物，所述MLG 75由SEQ ID NO:1350-1527组成；

[0087] (a4)能够特异性扩增MLG 109或其一个或多个特异性片段的一种或多种引物，所述MLG 109由SEQ ID NO:6549-7235组成；

[0088] (a5)能够特异性扩增MLG 317或其一个或多个特异性片段的一种或多种引物，所述MLG 317由SEQ ID NO:7581-7700组成；

[0089] (a6)能够特异性扩增MLG 135或其一个或多个特异性片段的一种或多种引物，所述MLG 135由SEQ ID NO:2230-3731组成；

[0090] (a7)能够特异性扩增MLG 223或其一个或多个特异性片段的一种或多种引物，所述MLG 223由SEQ ID NO:9892-11298组成；

[0091] (a8)能够特异性扩增MLG 100或其一个或多个特异性片段的一种或多种引物，所述MLG 100由SEQ ID NO:9596-9891组成；

[0092] (a9)能够特异性扩增MLG 219或其一个或多个特异性片段的一种或多种引物，所述MLG 219由SEQ ID NO:7701-8028组成；

[0093] (a10)能够特异性扩增MLG 114或其一个或多个特异性片段的一种或多种引物，所述MLG 114由SEQ ID NO:1528-2089组成；

[0094] (a11)能够特异性扩增MLG 84或其一个或多个特异性片段的一种或多种引物，所述MLG 84由SEQ ID NO:1-165组成；

[0095] (a12)能够特异性扩增MLG 166或其一个或多个特异性片段的一种或多种引物，所述MLG 166由SEQ ID NO:7236-7580组成；

[0096] (a13)能够特异性扩增MLG 2985或其一个或多个特异性片段的一种或多种引物，所述MLG 2985由SEQ ID NO:8029-9595组成；

[0097] (a14)能够特异性扩增MLG 131或其一个或多个特异性片段的一种或多种引物，所述MLG 131由SEQ ID NO:166-1349组成；和

[0098] (a15)能够特异性扩增MLG 1564或其一个或多个特异性片段的一种或多种引物，所述MLG 1564由SEQ ID NO:2090-2229组成；

[0099] (b)探针组，其包含：

[0100] (b1)能够与MLG 5045或其一个或多个特异性片段特异性杂交的一种或多种探针，所述MLG 5045由SEQ ID NO:3732-3918组成；和

[0101] (b2)能够与MLG 121或其一个或多个特异性片段特异性杂交的一种或多种探针，所述MLG 121由SEQ ID NO:3919-6548组成；

[0102] 任选地，所述探针组还包含以下探针的一种或多种：

[0103] (b3)能够与MLG 75或其一个或多个特异性片段特异性杂交的一种或多种探针，所述MLG 75由SEQ ID NO:1350-1527组成；

[0104] (b4)能够与MLG 109或其一个或多个特异性片段特异性杂交的一种或多种探针，所述MLG 109由SEQ ID NO:6549-7235组成；

[0105] (b5)能够与MLG 317或其一个或多个特异性片段特异性杂交的一种或多种探针，所述MLG 317由SEQ ID NO:7581-7700组成；

[0106] (b6)能够与MLG 135或其一个或多个特异性片段特异性杂交的一种或多种探针，所述MLG 135由SEQ ID NO:2230-3731组成；

[0107] (b7)能够与MLG 223或其一个或多个特异性片段特异性杂交的一种或多种探针，所述MLG 223由SEQ ID NO:9892-11298组成；

[0108] (b8)能够与MLG 100或其一个或多个特异性片段特异性杂交的一种或多种探针，所述MLG 100由SEQ ID NO:9596-9891组成；

[0109] (b9)能够与MLG 219或其一个或多个特异性片段特异性杂交的一种或多种探针，所述MLG 219由SEQ ID NO:7701-8028组成；

[0110] (b10)能够与MLG 114或其一个或多个特异性片段特异性杂交的一种或多种探针，所述MLG 114由SEQ ID NO:1528-2089组成；

[0111] (b11)能够与MLG 84或其一个或多个特异性片段特异性杂交的一种或多种探针，所述MLG 84由SEQ ID NO:1-165组成；

[0112] (b12)能够与MLG 166或其一个或多个特异性片段特异性杂交的一种或多种探针，所述MLG 166由SEQ ID NO:7236-7580组成；

[0113] (b13)能够与MLG 2985或其一个或多个特异性片段特异性杂交的一种或多种探针，所述MLG 2985由SEQ ID NO:8029-9595组成；

[0114] (b14)能够与MLG 131或其一个或多个特异性片段特异性杂交的一种或多种探针，所述MLG 131由SEQ ID NO:166-1349组成；和

[0115] (b15)能够与MLG 1564或其一个或多个特异性片段特异性杂交的一种或多种探针，所述MLG 1564由SEQ ID NO:2090-2229组成；

[0116] (c)包含(a)的引物组和/或(b)的探针组的微阵列；

[0117] (d)进行第二代测序方法或第三代测序方法的试剂；和

[0118] (e)(a)-(d)的任意组合。

[0119] 在优选实施方案中，所述引物组包含如(a1)-(a13)中定义的引物。

[0120] 在优选实施方案中，所述探针组包含如(b1)-(b13)中定义的探针。

[0121] 在优选实施方案中，该试剂盒通过包括以下步骤的方法，在受试者中预测或诊断与微生物群相关的疾病，或确定受试者是否具有形成所述疾病的风险：

[0122] (1)使用所述试剂盒来测定来自所述受试者的样品中的根据本发明的生物标志物组的每种生物标志物的水平或其量；和

[0123] (2a)通过使用多元统计模型(诸如随机森林模型)将所述样品中的每种生物标志物的水平或其量与训练数据集进行比较来计算所述疾病的概率；

[0124] 其中当所述疾病的概率大于临界值时，表明所述受试者患有所述疾病或具有形成所述疾病的风险；或

[0125] (2b)将所述样品中的每个生物标志物的水平或其量与所述对照中的相应生物标志物的水平或其量进行比较；优选地，所述对照是多个健康受试者；

[0126] 其中，与所述对照相比，当所述样品中所述生物标志物的水平或其量升高时，表明所述受试者患有所述疾病或具有形成所述疾病的风险。

[0127] 在优选实施方案中，所述训练数据集包含关于多个患有所述疾病的受试者和多个健康受试者的每种生物标志物的水平或其量的数据。

[0128] 在优选实施方案中，所述训练数据集包含表4-1和4-2中的数据，并且当概率大于临界值0.5时，表明所述受试者患有所述疾病或具有形成所述疾病的风险。

[0129] 在优选实施方案中，所述受试者是哺乳动物，例如灵长类动物，优选为人。

[0130] 在优选实施方案中，所述样品是粪便样品。

[0131] 在优选实施方案中，所述每种生物标志物的水平或其量是所述样品中每种生物标志物的相对丰度。

[0132] 在优选实施方案中，所述疾病是结直肠癌。

[0133] 在优选实施方案中，所述试剂盒还包含另外的试剂，诸如用于处理所述样品的试剂(例如无菌水)，用于进行PCR扩增的试剂(例如聚合酶、dNTP和扩增缓冲液)，以及用于进行杂交的试剂(诸如标记缓冲液、杂交缓冲液和洗涤缓冲液)。

[0134] 在第三方面，本发明提供了用于测定根据本发明的生物标志物组的每种生物标志物的水平或其量的试剂在制备试剂盒中的用途，所述试剂盒用于在受试者中预测或诊断与微生物群相关的疾病或确定受试者是否具有形成所述疾病的风险。

[0135] 在优选实施方案中，所述用于测定所述生物标志物组的每种生物标志物的水平或其量的试剂是如上所定义的。

[0136] 在优选实施方案中，所述试剂盒通过包括以下步骤的方法，在受试者中预测或诊断与微生物群相关的疾病，或确定受试者是否具有形成所述疾病的风险：

[0137] (1)使用所述试剂盒来测定样品中根据本发明的生物标志物组的每种生物标志物的水平或其量；和

[0138] (2a)通过使用多元统计模型(诸如随机森林模型)将所述样品中的每种生物标志物的水平或其量与训练数据集进行比较来计算所述疾病的概率；

[0139] 其中当所述疾病的概率大于临界值时，表明所述受试者患有所述疾病或具有形成所述疾病的风险；或

[0140] (2b)将所述样品中的每个生物标志物的水平或其量与所述对照中的相应生物标志物的水平或其量进行比较；优选地，所述对照是多个健康受试者；

[0141] 其中，与所述对照相比，当所述样品中所述生物标志物的水平或其量升高时，表明所述受试者患有所述疾病或具有形成所述疾病的风险。

[0142] 在优选实施方案中，所述训练数据集包含关于多个患有所述疾病的受试者和多个健康受试者的每种生物标志物的水平或其量的数据。

[0143] 在优选实施方案中，所述训练数据集包含表4-1和4-2中的数据，并且当所述疾病的概率大于临界值0.5时，表明所述受试者患有所述疾病或具有形成所述疾病的风险。

[0144] 在优选实施方案中，所述受试者是哺乳动物，例如灵长类动物，优选为人。

[0145] 在优选实施方案中，所述样品是粪便样品。

[0146] 在优选实施方案中，所述每种生物标志物的水平或其量是所述样品中每种生物标志物的相对丰度。

[0147] 在优选实施方案中，所述疾病是结直肠癌。

[0148] 在优选实施方案中，所述试剂盒还包含另外的试剂，诸如用于处理样品的试剂(例

如无菌水)，用于进行PCR扩增的试剂(例如聚合酶、dNTP和扩增缓冲液)，以及用于进行杂交的试剂(诸如标记缓冲液、杂交缓冲液和洗涤缓冲液)。

[0149] 在第四方面，本发明提供了用于在受试者中预测或诊断与微生物群相关的疾病或确定受试者是否具有形成所述疾病的风险的方法，其包括以下步骤：

[0150] (1)测定来自所述受试者的样品中的根据本发明的生物标志物组的每种生物标志物的水平或其量；和

[0151] (2a)通过使用多元统计模型(如随机森林模型)将所述样品中的每个生物标志物的水平或其量与训练数据集进行比较来计算所述疾病的概率；

[0152] 其中当所述疾病的概率大于临界值时，表明所述受试者患有所述疾病或具有形成所述疾病的风险；或

[0153] (2b)将所述样品中的每个生物标志物的水平或其量与所述对照中的相应生物标志物的水平或其量进行比较；优选地，所述对照是多个健康受试者；

[0154] 其中，与所述对照相比，当所述样品中所述生物标志物的水平或其量升高时，表明所述受试者患有所述疾病或具有形成所述疾病的风险。

[0155] 在优选实施方案中，所述训练数据集包含关于多个患有所述疾病的受试者以及多个健康受试者的每种生物标志物的水平或其量的数据。

[0156] 在优选实施方案中，所述训练数据集包含表4-1和4-2中的数据，并且当所述疾病的概率大于临界值0.5时，表明所述受试者患有所述疾病或具有形成所述疾病的风险。

[0157] 在优选实施方案中，在步骤(1)中使用如上定义的试剂盒或如上定义的试剂。

[0158] 在优选实施方案中，所述受试者是哺乳动物，例如灵长类动物，优选为人。

[0159] 在优选实施方案中，所述样品是粪便样品。

[0160] 在优选实施方案中，所述每种生物标志物的水平或其量是所述样品中每种生物标志物的相对丰度。

[0161] 在优选实施方案中，所述疾病是结直肠癌。

[0162] 在优选实施方案中，在体外进行所述方法。

[0163] 在以下非限制性实施例中进一步举例说明本发明。除非另有说明，否则部分和百分比以重量计，度为摄氏度。所用的试剂皆是商购可得的。对于本领域普通技术人员而言显而易见的是，这些实施例虽然表示本发明的优选实施方案，但仅以说明的方式给出。

实施例

[0164] 实施例1.鉴定和验证用于评估CRC相关疾病风险的生物标志物

[0165] 1.样品收集和测序

[0166] 1.1受试者和患者

[0167] 在依照CRC国家筛选建议(Stadlmayr,A.et al.Nonalcoholic fatty liver disease:an independent risk factor for colorectal neoplasia.J Intern Med270,41-49(2011)，通过引用并入本文)进行的一个健康筛查程序中的那些参与者以及2010年至2012年期间在Oberndorf医院内科(奥地利萨尔斯堡Paracelsus医科大学的教学医院)进行过结肠镜检查(作为临床检查的部分)的那些疑似患有CRC的患者中进行研究。本研究获得当地伦理委员会(Ethikkommission des Landes Salzburg，批准号415-E/1262/2-2010)的

批准，并获得所有参与者的知情同意书。

[0168] 将泻药**Klean-Prep®**(含有聚乙二醇59.0g、硫酸钠5.68g、碳酸氢钠1.68g、NaCl 1.46g和氯化钾0.74g;Norgine,Marburg,德国)用于肠道准备，然后进行结肠镜检查。基于肉眼检查和组织学检测结果的组合分析，结肠镜检查结果被分为管状腺瘤、进展性腺瘤(即绒毛状或管状绒毛状特征，大小≥1cm或高度发育异常)或癌(Bond,J.H.Polyp guideline: diagnosis,treatment,and surveillance for patients with colorectal polypsACG Colorectal Polyp Guideline.Am.J.Gastroenterol.95,3053-3063(2000),Winawer SJ& AG.,Z.The advanced adenoma as the primary target of screening.Gastrointest Endosc Clin N Am12,1-9(2002)，通过引用并入本文)。根据位置(即右结肠(包括盲肠、升结肠和横结肠)，左结肠(从脾曲到乙状结肠)，以及单独的直肠)对病灶进行分类。

[0169] 初始分析囊括了来自147名年龄在45至86岁之间的白种人的数据，其中包括57名健康对照(24名男性，33名女性)，44例进展性腺瘤患者(女性22例，男性22例)和46例癌患者(18例男性，28例女性)(表1-1)。另外9个样品(6个健康对照，3个进展性腺瘤样品，表1-1)也被用于基于MLG的癌分类器的测试集(图1d)。到目前为止，还没有研究以可比较的方式探究过上述给定的主题;因此，无法进行用于样品量计算的正式效能分析(formal power analysis)。但是，根据以前的16S和宏基因组鸟枪法测序对病人的粪便微生物的研究来判断，这是合理的样品量。将受试者在性别、年龄和体重指数(BMI)方面进行分层，以使得三组(对照组、进展性腺瘤组、癌组)在这些变量上可比较。在进展性腺瘤组中，14例的病灶位于右结肠(包括盲肠、升结肠和横结肠)，15例的病灶位于左结肠(从脾曲至乙状结肠)，15例的病灶位于直肠。在癌组中，8例的病灶位于右结肠，11例的病灶位于左结肠，27例的病灶位于直肠。结直肠癌由美国癌症联合委员会(AJCC)TNM分期系统(Greene,F.L.Current TNM staging of colorectal cancer.Lancet.Oncol.8,572-3(2007)，通过引用并入本文)进行分类。

[0170] 表1-1:所有156个样品的临床资料

[0171]

样品 ID	临床数据					
	状态	年龄(岁)	BMI	癌的 TNM-分类	组织学	结肠中的定位(右/左/直肠)
147 个样品						
31766	对照	73	32	n.a.	n.a.	n.a.
31537	对照	70	31	n.a.	n.a.	n.a.
31600	对照	70	30	n.a.	n.a.	n.a.
31428	对照	69	30	n.a.	n.a.	n.a.
530167	对照	72	30	n.a.	n.a.	n.a.
530315	对照	75	31	n.a.	n.a.	n.a.
530050	对照	71	29	n.a.	n.a.	n.a.
31160	对照	67	30	n.a.	n.a.	n.a.
530177	对照	71	29	n.a.	n.a.	n.a.
31700	对照	72	28.7	n.a.	n.a.	n.a.
31714	对照	69	28.3	n.a.	n.a.	n.a.
31219	对照	74	28.3	n.a.	n.a.	n.a.
31557	对照	74	28.5	n.a.	n.a.	n.a.
530154	对照	74	28.5	n.a.	n.a.	n.a.
530444	对照	62	22.1	n.a.	n.a.	n.a.
31021	对照	60	22.1	n.a.	n.a.	n.a.

[0172]

530074	对照	68	22.1	n.a.	n.a.	n.a.
530227	对照	68	22.4	n.a.	n.a.	n.a.
530394	对照	67	22.58	n.a.	n.a.	n.a.
530295	对照	70	22.67	n.a.	n.a.	n.a.
530368	对照	70	22.8	n.a.	n.a.	n.a.
31300	对照	68	21.7	n.a.	n.a.	n.a.
31452	对照	66	21.7	n.a.	n.a.	n.a.
31723	对照	69	22.4	n.a.	n.a.	n.a.
531424	对照	46	28.74	n.a.	n.a.	n.a.
31009	对照	68	32	n.a.	n.a.	n.a.
530251	对照	72	31.8	n.a.	n.a.	n.a.
31416	对照	72	31.8	n.a.	n.a.	n.a.
530119	对照	70	31.6	n.a.	n.a.	n.a.
31749	对照	70	31.25	n.a.	n.a.	n.a.
530364	对照	63	31.17	n.a.	n.a.	n.a.
31328	对照	63	30.9	n.a.	n.a.	n.a.
530163	对照	63	30.4	n.a.	n.a.	n.a.
31379	对照	63	30.4	n.a.	n.a.	n.a.
31112	对照	66	30.3	n.a.	n.a.	n.a.
31750	对照	66	34.04	n.a.	n.a.	n.a.
530451	对照	68	29.8	n.a.	n.a.	n.a.
31129	对照	73	29.7	n.a.	n.a.	n.a.
530052	对照	64	29.38	n.a.	n.a.	n.a.
31512	对照	64	29.3	n.a.	n.a.	n.a.
31236	对照	65	29.06	n.a.	n.a.	n.a.
31360	对照	71	29.04	n.a.	n.a.	n.a.
31343	对照	72	29.03	n.a.	n.a.	n.a.
530215	对照	65	31.22	n.a.	n.a.	n.a.
31519	对照	65	28.90	n.a.	n.a.	n.a.
31450	对照	67	23.14	n.a.	n.a.	n.a.
531414	对照	46	23.15	n.a.	n.a.	n.a.
31071	对照	68	23.45	n.a.	n.a.	n.a.
530331	对照	70	23.67	n.a.	n.a.	n.a.
530258	对照	65	22.03	n.a.	n.a.	n.a.
31333	对照	68	22.2	n.a.	n.a.	n.a.
31232	对照	70	22.64	n.a.	n.a.	n.a.
530055	对照	64	22.78	n.a.	n.a.	n.a.
31267	对照	67	22.79	n.a.	n.a.	n.a.
31711	对照	66	30.02	n.a.	n.a.	n.a.
530075	对照	63	34.14	n.a.	n.a.	n.a.
31637	对照	66	29.98	n.a.	n.a.	n.a.
31398	进展性腺癌	68	34.01	n.a.	n.a.	右
531403	进展性腺癌	59	29.2	n.a.	n.a.	左
530168	进展性腺癌	62	32.8	n.a.	n.a.	右

[0173]

530185	进展性腺癌	52	32.8	n.a.	n.a.	右
530600	进展性腺癌	70	25.2	n.a.	n.a.	右
31477	进展性腺癌	78	31.24	n.a.	n.a.	直肠
530403	进展性腺癌	75	30.12	n.a.	n.a.	右
530002	进展性腺癌	60	30	n.a.	n.a.	右
530697	进展性腺癌	63	30.47	n.a.	n.a.	左
31455	进展性腺癌	72	27.43	n.a.	n.a.	直肠
530756	进展性腺癌	70	29.38	n.a.	n.a.	左
31424	进展性腺癌	68	24.84	n.a.	n.a.	右
31501	进展性腺癌	69	24.4	n.a.	n.a.	右
31256	进展性腺癌	69	24.13	n.a.	n.a.	左
530297	进展性腺癌	78	24	n.a.	n.a.	直肠
530142	进展性腺癌	61	22.8	n.a.	n.a.	左
530026	进展性腺癌	83	21.94	n.a.	n.a.	左
530558	进展性腺癌	56	22.8	n.a.	n.a.	直肠
530054	进展性腺癌	64	20.52	n.a.	n.a.	左
530743	进展性腺癌	66	22.65	n.a.	n.a.	直肠
530867	进展性腺癌	58	27.02	n.a.	n.a.	右
530705	进展性腺癌	77	37.56	n.a.	n.a.	直肠
31337	进展性腺癌	62	35.43	n.a.	n.a.	左
31030	进展性腺癌	70	34.11	n.a.	n.a.	直肠
31282	进展性腺癌	63	34.08	n.a.	n.a.	直肠
530028	进展性腺癌	67	33.23	n.a.	n.a.	右
31449	进展性腺癌	63	30.77	n.a.	n.a.	左
31275	进展性腺癌	62	30.1	n.a.	n.a.	右
530018	进展性腺癌	56	29.41	n.a.	n.a.	右
530262	进展性腺癌	64	29.3	n.a.	n.a.	左
530039	进展性腺癌	67	28.34	n.a.	n.a.	左
530172	进展性腺癌	63	27.4	n.a.	n.a.	直肠
31137	进展性腺癌	67	27.14	n.a.	n.a.	右
31431	进展性腺癌	71	27.04	n.a.	n.a.	左
31233	进展性腺癌	62	26.88	n.a.	n.a.	左
530398	进展性腺癌	80	26.47	n.a.	n.a.	左
31582	进展性腺癌	77	25.7	n.a.	n.a.	直肠
530450	进展性腺癌	48	25.61	n.a.	n.a.	直肠
530623	进展性腺癌	62	24.3	n.a.	n.a.	直肠
530323	进展性腺癌	71	24.8	n.a.	n.a.	直肠
530348	进展性腺癌	58	21.7	n.a.	n.a.	直肠
530041	进展性腺癌	69	21.46	n.a.	n.a.	左
31705	进展性腺癌	84	24.7	n.a.	n.a.	直肠
530840	进展性腺癌	74	24.13	n.a.	n.a.	右
31446	癌	84	31.22	pTis	原位癌	右
31881	癌	60	30.08	T4N1M0	腺癌	直肠
31866	癌	73	29.75	T3N0M0	腺癌	左

[0174]

31874	癌	74	29.17	T4N1M1	腺癌	左
31549	癌	84	27.77	T1N0M0	腺癌	左
31878	癌	52	27.14	T3N1M0	腺癌	直肠
531281	癌	72	27.12	pTis	原位癌	左
31877	癌	72	26.67	T3N1M0	腺癌	直肠
530373	癌	47	25.78	T2N0M0	腺癌	直肠
531155	癌	72	24.22	T1N0M0	腺癌	右
31868	癌	64	24.09	T2N0M0	腺癌	直肠
531361	癌	72	23.83	T4N1M0	腺癌	直肠
531775	癌	77	23.50	T3N0M0	腺癌	右
31865	癌	55	23.34	T2N0M0	腺癌	直肠
31223	癌	65	23.31	T4N0M0	腺癌	直肠
531469	癌	45	22.15	T2N0M0	腺癌	直肠
31876	癌	56	20.07	T3N2M0	腺癌	直肠
531416	癌	63	33.56	T2N0M0	腺癌	直肠
31872	癌	46	30.86	pTis	原位癌	左
31276	癌	82	30.47	pTis	原位癌	右
531333	癌	69	29.94	T3N1M0	腺癌	左
531382	癌	65	29.74	T3N0M0	腺癌	直肠
530890	癌	77	29.41	pTis	原位癌	直肠
31237	癌	67	29.40	T3N0M0	腺癌	右
31004	癌	64	29.35	T1N0M0	腺癌	直肠
531277	癌	64	29.07	T1N0M0	腺癌	直肠
31871	癌	72	28.41	T2N0M0	腺癌	左
31188	癌	65	28.03	n.a.	腺癌	直肠
31875	癌	58	27.76	T2N0M0	腺癌	直肠
31884	癌	54	27.34	pTis	原位癌	直肠
31285	癌	84	27.28	T3N0M0	腺癌	直肠
31489	癌	60	26.64	T2N0M0	腺癌	直肠
31685	癌	74	26.37	T1N0M0	腺癌	左
531248	癌	79	26.30	T3N1M0	腺癌	左
31870	癌	63	25.88	T1N0M0	腺癌	左
531128	癌	52	25.73	T3N1M0	腺癌	直肠
31873	癌	73	24.82	T3N0M0	腺癌	直肠
31880	癌	74	24.34	T4N2M0	腺癌	右
531766	癌	72	24.34	T2N0M0	腺癌	右
31883	癌	43	22.68	T3N0M0	腺癌	直肠
531352	癌	66	22.60	T1N0M0	原位癌	左
31367	癌	86	20.50	T2N0M0	腺癌	右
31493	癌	68	32.25	pTis	原位癌	直肠
31879	癌	71	29.74	T4N2M0	腺癌	直肠
531274	癌	77	19.16	T3N2M0	腺癌	直肠
31159	癌	73	17.99	T3N0M0	腺癌	直肠
另外 9 个样品						

[0175]

532749	对照	55	27.76	n.a.	n.a.	n.a.
532779	对照	70	28.72	n.a.	n.a.	n.a.
532796	对照	73	26.56	n.a.	n.a.	n.a.
532802	对照	68	23.53	n.a.	n.a.	n.a.
532826	对照	78	31.22	n.a.	n.a.	n.a.
532832	进展性腺癌	68	27.55	n.a.	n.a.	n.a.
532305	进展性腺癌	56	43.58	n.a.	n.a.	n.a.
532915	对照	43	22.65	n.a.	n.a.	n.a.
531663	进展性腺癌	63	25.65	n.a.	n.a.	n.a.

[0176] 1.2粪便样品

[0177] 从所有患者和受试者收集新鲜粪便样品。样品用无菌刮刀机械匀化，然后使用Sarstedt粪便取样系统(Sarstedt,Nümbrecht,德国)取4份等分试样。每个等分试样含有1g粪便并放置在无菌12ml冻存管中。然后将粪便等分试样储存在-20℃家用冰箱中，并在收集后的48小时内将其放置在冷藏包中运送至实验室，然后立即将其储存在-80℃。所有患者和受试者在过去3个月内没有接受过益生菌或抗生素。

[0178] 1.3DNA的提取

[0179] 将粪便样品在冰上解冻，并按照制造商的说明书使用Qiagen QIAamp DNA Stool Mini试剂盒(Qiagen)进行DNA提取。提取物用不含DNA酶的RNA酶处理以消除RNA污染。使用NanoDrop分光光度计，Qubit荧光计(使用Quant-iTTMdsDNA BR Assay试剂盒)和凝胶电泳测定DNA量。

[0180] 1.4宏基因组测序和基因目录的构建

[0181] 在Illumina平台(插入片段大小为350bp，读段(read)长度为100bp)上进行双末端宏基因组测序(paired-end metagenomic sequencing)，并且如前所述(Qin等，2012，同上)使用SOAPdenovo v2.04(除对于-K 51-M3-F-u外，使用缺省参数)(Luo,R.等SOAPdenovo2:an empirically improved memory-efficient short-read de novo assembler.Gigascience1,18(2012)，通过引用并入本文)对测序读段(read)进行质量控制并从头组装成重叠群(contig)。从头组装高质量的测序读段(平均每个样品含有5GB数据量)，将鉴定的基因编入3.5M非冗余基因集，这使得平均每个样品中有76.3%的读段可以匹配上。

[0182] 使用GeneMark v2.7d对组装的重叠群进行基因预测。使用BLAT(Kent,W.J.BLAT--the BLAST-like alignment tool.Genome Res.12,656-64(2002)，通过引用并入本文)，除去冗余基因，其中以90%重叠度和95%同一性(不允许有缺口)作为临界值。通过使用与Qin等2012(同上)中相同的程序，将高质量的测序读段与基因目录进行比对来测定基因的相对丰度。

[0183] 根据IMG数据库(v400)，并使用先前详述(Qin等2012，同上)的内部流程，利用80%的重叠和65%的同一性，前10%的评分(BLASTN v2.2.24,-e 0.01-b 100-K 1-F T-m 8)对预测基因进行分类学分配。分配至门时，临界值为65%的同一性，分配至属时，临界值为85%同一性，分配至种时，临界值为95%的同一性；如果存在多次命中(multiple hits)，则对于存在该疑问的分类群，其临界值为≥50%的一致性。

[0184] 2.宏基因组关联分析(MGWAS)

[0185] 为了比较健康对照、进展性腺瘤和癌症患者的粪便微生物群系，鉴定了相对丰度在上述任意两组之间展现出显著差异的基因(Benjamin-Hochberg q值<0.1，Kruskal-Wallis检验)。然后将这些标志物基因根据其在所有三组样品中的丰度变化聚类成MLG，这使得能够鉴定每组的微生物物种特征(Qin等，2012，同上)。147个样品中有9个含有超过20％的埃希氏杆菌属(Escherichia)(2个对照、2个腺瘤、5个癌症样品)，随后该样品仅用在用于基于MLG的癌分类器的测试集中(图1d)。另外9个样品(6个健康对照、3个进展性腺瘤样品，表1-1)也用在用于上述分类器的测试集中。

[0186] 如前所述，根据分类学及其组成基因的相对丰度进行MLG的分类学分配和丰度特征谱表征(Qin等2012，同上)。简而言之，将MLG分配至种需要MLG中超过90％的基因能够以超过95％的同一性，以及超过70％的查询重叠度比对到该种的基因组。将MLG分配至属，需要该MLG中超过80％的基因能够在DNA和蛋白质序列上均以至少85％同一性比对到该属的基因组上。

[0187] 为了探索健康或肿瘤样品中的肠道微生物群系的特征，本发明人鉴定了在三组的任意两组中均显示出显著的丰度差异的130,715个基因(Kruskal-Wallis检验，Benjamin-Hochberg q值<0.1)。除了血清铁蛋白和对红肉的摄取状况外，除肿瘤状态以外没有一种表型在对照、腺瘤和癌症患者中均显示出显著的差异($p<0.05$，Kruskal-Wallis检验，表1-2)。与健康和进展性腺瘤样品相比，58.9％的基因标志物在癌症样品中显著升高，表明它们对结直肠癌是特异的；另外24.3％的基因在癌症样品中的丰度显著高于对照样品，但在进展性腺瘤样品中具有中等水平。在具有下降趋势的基因中，与健康和进展性腺瘤样品相比，5388个基因(占总数的4.1％)在癌症样品中显著降低；2601个基因(占总数的2.0％)的丰度在癌症样品中显著低于对照样品，在进展性腺瘤样品中具有中等水平。这些在对照样品中富集的基因，而非那些在腺瘤或癌样品中富集的基因，被更多地匹配至京都基因与基因组百科全书(KEGG)通路。递增和递减基因数目的差异表明，在癌发展过程中病生菌(pathobionts)的增加比有益细菌的减少更为明显。根据各个基因在所有样品中丰度的共变化，将显著不同的基因聚类成126个MLG，这使得能够鉴定每组的微生物物种特征(Qin等，2012，同上)。

[0188] 3.结直肠癌或腺瘤的基于MLG的分类

[0189] 为了评价结直肠癌的粪便微生物群系的诊断价值，本发明人构建了可检测癌症样品的随机森林分类器。使用训练队列(集)的MLG丰度特征谱来训练随机森林模型(R.2.14，randomForest4.6-7软件包)(Liaw，Andy&Wiener，Matthew.Classification and Regression by randomForest，R News(2002)，第2/3卷，第18页，通过引用并入本文)，从而选择MLG标志物的最优集。在测试集上测试该模型，并测定了预测误差。关于该随机森林模型，通过使用R vision 2.14中的“randomForest4.6-7包”，输入训练数据集(即训练样品中所选MLG的相对丰度特征谱)、样品疾病状态(训练样品的样品疾病状态为矢量，1表示病例，0表示对照)和测试数据集(即在测试集中所选MLG的相对丰度特征谱)。然后本发明人使用R软件中的randomForest包的随机Forest函数来构建分类，并且使用预测函数来预测测试集。输出的是预测结果(患病的概率；临界值是指最佳临界值，如果疾病的概率≥最佳临界值，则受试者处于疾病的风险中)。

[0190] 使用对照、进展性腺瘤或癌症样品的MLG丰度特征谱，对随机森林模型(R 3.0.2，randomForest4.6-7包)进行10折交叉验证。对获得自5次10折交叉验证的交叉验证误差曲线(每条曲线为10个测试集的平均值)进行平均，并将该平均曲线中的最小误差加上该点处的标准偏差得到的数值用作临界值。列出误差小于该临界值的MLG标志物的所有集合(≤50)，并选择具有最少数目的MLG的集合作为最优集。使用该MLG集计算腺瘤或癌的概率，并绘制ROC(R 3.0.2，pROC3包)。在测试集上进一步测试该模型，并测定了预测误差。

[0191] 通过对由55个对照和41个癌症样品(表1-1)组成的训练集的5次重复的10折交叉验证(即50次测试)，从而获得15个MLG标志物的最优选择(表2-1，表2-2，表3)。简而言之，在由55个对照和41个癌样品组成的训练集中进行5次重复的10折交叉验证(即50次测试)。每次测试，随机森林测试均对每个MLG的重要性进行排序。本发明人挑选了前15个重要的MLG，并按照出现次数对mlg进行了排序。上述前15个MLG用于构建分类器。表6列出了MLG的重要性的排序。

[0192] 我们的研究结果表明，上述15个MLG在训练集上表现良好，接受者工作曲线的曲线下面积(AUC)为98.34％(临界值＝0.5，图1a、1b、1c，表4-1，表4-2，表5和表6)。测试集(8个对照样品，47个进展性腺瘤样品和5个癌样品)的分类误差较低，接受者操作曲线的曲线下面积(AUC)为96％(进展性腺瘤被认为是非癌，临界值＝0.5，图1d、1e，表5)，这与他们(进展性腺瘤)大多为良性这一性质一致。上述MLG标志物当中包括可能为口腔厌氧菌的mlg-75和mlg-84，前者显示出对腺瘤的高优势比(odds ratio)(表2-2)，表明其在发病机制中的早期作用。其它MLG标志物包括马赛拟杆菌(Bacteroides massiliensis)，mlg-2985、mlg-121和10种另外的分类学未定义的MLG(表2-2)。因此，由癌分类器选择出的MLG显示在腺瘤和癌中导致疾病恶化的肠道微生物群系的重要特征，并对这些肿瘤的早期和无创性诊断具有很大的潜力。

[0193] 另外，表6中的结果显示，对于前2个重要的MLG(MLG 5045和MLG 121)的组合，AUC为0.91751663；对于前3个重要的MLG(MLG5045、MLG 121和MLG 75)的组合，AUC为0.970731707；对于前4个MLG的组合，AUC为0.959645233；对于前5个MLG的组合，AUC为0.975609756；对于前6个MLG的组合，AUC为0.978713969；对于前7个MLG的组合，AUC为0.980044346；对于前8个MLG的组合，AUC为0.985365854；对于前9个MLG，AUC为0.984035477；对于前10个MLG，AUC为0.981818182；对于前11个MLG，AUC为0.980931264；对于前12个MLG，AUC为0.979157428；对于前13个MLG，AUC为0.987583149；对于前14个MLG，AUC为0.986696231；以及对于前15个MLG，AUC为0.983370288。这些结果表明，MLG 5045和MLG 121是15个MLG中最重要的生物标志物，MLG 5045和MLG121的组合足以诊断结直肠癌和/或确定患结直肠癌的风险，且具有高灵敏度和高特异性(AUC＝0.91751663)；该15个MLG中的其它MLG生物标志物的并入可以在一定程度上提高诊断或预测的灵敏度和特异度。特别地，这些结果还表明，前13个MLG的组合可以被认为是最优生物标志物组，其对结直肠癌的诊断或预测具有最优灵敏度和特异性(AUC＝0.987583149)。

[0194] 这些结果完全支持本文所鉴定的MLG(特别是MLG 5045和MLG121，任选地与这15个MLG中另外的MLG中的一个或多个组合)可用作诊断结直肠癌和/或确定患结直肠癌的风险的生物标志物，且具有高灵敏度和高特异性。

[0195] 本发明人进一步直接研究了肠道MLG用于鉴定腺瘤的效用，其比结直肠癌更难筛

选，但对于早期干预是重要的。

[0196] 类似地，在由55个对照和42个腺瘤样品组成的训练集中进行5次重复的10折交叉验证(即50次测试)。每次测试，随机森林测试均对每个MLG的重要性进行了排序。本发明人挑选了前10个重要的MLG，并按照出现次数对MLG进行了排序。这前10个MLG用于构建分类器。表11中列出了MLG的重要性的排序。

[0197] 在5次重复的10折交叉验证之后，随机森林模型选择了允许对训练集进行最优分类(55个对照和42个进展性腺瘤，表9和表10，图2b)的10个MLG(表7-1，表7-2，图2a)，AUC为0.8738(临界值＝0.5，图2c)。在测试集(由15个对照和15个进展性腺瘤组成的、且未使用的新样品)中，所有进展性腺瘤样品均被正确分类(临界值＝0.4572，图2d，2e，表10)。因此，粪便MLG为结直肠进展性腺瘤的无创性检测提供了新的机会。

[0198] 另外，表11中的结果显示，对于前2个重要MLG(MLG 317和MLG3770)的组合，AUC为0.782251082；对于前3个重要的MLG(MLG 317、MLG 3770和MLG 3840)的组合，AUC为0.805194805；对于前4个MLG的组合，AUC为0.773160173；对于前5个MLG的组合，AUC为0.795238095；对于前6个MLG的组合，AUC为0.780952381；对于前7个MLG的组合，AUC为0.895670996；对于前8个MLG的组合，AUC为0.896536797；对于前9个MLG的组合，AUC为0.884848485；对于前10个MLG的组合，AUC为0.873809524。这些结果表明，MLG 317、MLG 3770和MLG 3840是这10个MLG中最重要的生物标志物，并且MLG 317、MLG 3770和MLG 3840的组合足以诊断结直肠进展性腺瘤和/或确定患结直肠进展性腺瘤的风险，且具有高灵敏度和高特异性(AUC＝0.805194805)；以及这10个MLG中的其它MLG生物标志物的并入可在一定程度上提高诊断或预测的敏感度和特异性。特别地，这些结果还表明，前8个MLG的组合可被认为是最优生物标志物组，其对结直肠进展性腺瘤的诊断或预测具有最优灵敏度和特异性(AUC＝0.896536797)。

[0199] 这些结果完全支持本文鉴定的MLG(特别是MLG 317、MLG 3770和MLG 3840，任选地与上述10个MLG中另外的MLG中的一个或多个组合)可用作诊断结直肠进展性腺瘤和/或确定患结直肠进展性腺瘤的风险的生物标志物，且具有高灵敏度和高特异性。

[0200]

表 1-2：表 1-1 中 3 个组之间的表型一致性

表型	对照组的平均值	进展性腺瘤组的平均值	癌组的平均值	p-值 (Kruskal-Wallis)	p 值校正(BH)	使用的样品数量(最大为 138)
年龄(岁数)	67.44	66.19	66.10	0.3936	0.6332	138
BMI	27.50	27.82	26.82	0.4735	0.6332	138
腰围(cm)	100.78	99.53	99.56	NA	NA	108
臀围(cm)	105.35	105.97	102.29	NA	NA	100
WHR (腰-臀比)	0.94	0.93	0.97	0.0968	0.4768	132
GGT (U/L)	45.07	49.98	38.22	0.8623	0.8623	138
GOT (AST) (U/L)	23.45	23.69	24.24	0.6200	0.7188	138
GPT (ALT) (U/L)	26.13	27.86	23.59	0.2877	0.6332	138
空腹胰岛素(μU/mL)	10.44	10.48	9.53	NA	NA	68
空腹血糖(mg/L)	109.73	107.45	103.48	0.3640	0.6332	137
HOMA 指数	2.97	2.83	2.55	NA	NA	68
Hba1C (%)	5.83	5.96	5.93	NA	NA	95
CRP (mg/L)	0.66	0.61	1.02	0.3177	0.6332	138
血清铁蛋白(ng/mL)	234.41	238.48	126.94	0.0004	0.0080	136
Hb (g/L)	14.67	16.88	14.07	0.2585	0.6332	138
TG (mg/L)	123.49	133.52	128.63	0.6829	0.7188	136
HDL (mg/L)	59.64	65.24	57.88	0.1148	0.4768	136
LDL (mg/L)	146.79	143.19	138.46	0.6825	0.7188	136
红肉(g/wk)	110.85	150.64	141.00	0.0483	0.4768	132
白肉(g/wk)	77.26	81.79	64.92	0.4619	0.6332	132
总的肉(g/wk)	188.87	234.67	207.53	0.1192	0.4768	132
鱼(g/wk)	151.85	171.43	166.67	0.6123	0.7188	132
蔬菜(g/wk)	1740.74	1690.48	1461.11	0.4749	0.6332	132
水果(g/wk)	1798.15	1604.76	1397.22	0.3733	0.6332	132
纤维摄入量(g/wk)	50.17	64.08	66.69	0.2600	0.6332	132

[0201]

表 2-1: 与结直肠癌相关的 15 个最具判别性的 MLG(物种标志物)(富集:分别在对照(CTRL)和腺瘤(AA)、在对照(CTRL)和癌症(CRC)、以及在腺瘤(AA)和癌症(CRC)的对比中的 MLG 富集方向，其中+表示在后组中的富集，-表示前组中的富集，0 表示两组之间无差异(p≥0.05，Wilcoxon 秩和检验，用于多重检验的针对对照的 Bonferroni 校正)

MLG ID	p-值(Kruskal)	p 值校正(BH)	p-值(Wilcoxon)			富集	分类器		平均秩			出现率		
			CTRL 对比 AA	CTRL 对比 CRC	AA 对比 CRC		CRC	AA	CTRL	AA	CRC	CTRL	AA	CRC
317	3.15E-07	9.64E-06	1.65E-03	9.00E-07	4.12E-02	+,+,+	是	是	49	74	93	0.51	0.88	1.00
2985	1.16E-05	9.17E-05	4.35E-01	2.44E-06	2.77E-02	0,+,+	是	非	54	67	92	0.47	0.57	0.93
75	3.81E-09	6.05E-07	1.13E-01	1.65E-08	1.83E-04	0,+,+	是	非	54	64	95	0.11	0.26	0.63
84	3.19E-08	2.55E-06	3.63E-01	1.38E-07	8.98E-05	0,+,+	是	非	54	63	97	0.24	0.38	0.73
121	1.90E-09	3.68E-07	7.38E-01	2.88E-09	5.85E-05	0,+,+	是	非	53	62	100	0.27	0.36	0.90
131	1.84E-08	1.74E-06	3.60E-01	4.15E-08	4.81E-04	0,+,+	是	非	55	64	95	0.13	0.24	0.66
100	7.75E-06	7.27E-05	2.54E-01	6.80E-06	1.29E-02	0,+,+	是	非	57	67	90	0.15	0.29	0.59
109	1.60E-10	6.66E-08	7.99E-01	7.93E-10	1.50E-05	0,+,+	是	非	53	61	100	0.18	0.26	0.80
219	1.90E-09	3.68E-07	7.36E-01	5.61E-09	7.33E-05	0,+,+	是	非	54	62	98	0.18	0.26	0.76
223	5.22E-11	3.18E-08	8.37E-02	7.66E-10	4.97E-05	0,+,+	是	非	54	63	96	0.04	0.17	0.63
1564	6.39E-09	8.59E-07	7.16E-02	4.05E-08	2.95E-04	0,+,+	是	非	55	65	94	0.05	0.21	0.59
5045	1.31E-10	5.75E-08	2.89E-02	3.47E-10	2.24E-04	+,+,+	是	非	52	65	97	0.05	0.24	0.73
114	9.32E-09	1.06E-06	1.76E-01	4.83E-09	1.08E-03	0,+,+	是	非	52	65	97	0.20	0.36	0.83
166	3.29E-12	5.21E-09	8.63E-01	6.92E-10	4.11E-07	0,+,+	是	非	55	59	100	0.09	0.17	0.73
135	8.34E-13	1.65E-09	1.00E+00	3.95E-09	2.36E-07	0,+,+	是	非	57	58	98	0.04	0.05	0.63

表 2-2: 与结直肠癌相关的 15 个最具判别性的 MLG (物种标志物)

MLG	丰度平均值	优势比(95% CI)	基因数目	MLG 注释	最佳匹配的菌株(核苷	以大于 65%同一	匹配基因的平均同

[0202]

ID	CTRL	AA	CRC	CTRL 与 AA 比较	CTRL 与 CRC 比较		马赛拟杆菌 (Bacteroides massiliensis)	酸)	性匹配的基因所占的分数	一性(%)
317	4.13 E-07	3.77 E-07	8.77 E-07	1.02 (0.68-1.53)	1.28 (0.8-2.05)	120		马赛拟杆菌	0.9583	99.54
2985	1.44 E-06	1.20 E-06	3.03 E-06	1.07 (0.71-1.59)	1.5 (0.93-2.43)	1567	mlg-2985	Dialister invisus	0.8826	98.49
75	2.80 E-09	1.00 E-10	1.55 E-07	111.87 (0.13-99028.45)	4.789676036621 82e+243 (5.043587818 46618e+37-Inf)	178	mlg-75	梭杆菌属某种 (Fusobacterium sp.)口腔分类群 370	0.7753	99.36
84	2.30 E-09	1.10 E-09	8.44 E-08	1.37 (0.84-2.25)	2541299739464 95 (36942.45-1.7481797484976e+24)	165	mlg-84	麻疹孪生球菌 (Gemella morbillorum)	0.5212	97.88
121	5.08 E-07	4.81 E-08	6.93 E-06	2.54 (0.71-9.08)	4011933781285.55 (18.76-8.580342394829 51e+23)	2630	mlg-121	普氏菌 (Prevotella copri)	0.7293	98.07
131	3.59 E-08	1.10 E-09	1.66 E-06	6.76 (0.2-225.78)	2.615725180853 32e+253 (9.330172269978 7e+44-Inf)	1184	mlg-131	未分类的		
100	4.19 E-08	1.10 E-09	3.00 E-06	5.41 (0.13-232.75)	Inf (1.181974014542 76e+50-Inf)	296	mlg-100	未分类的		
109	6.14 E-08	3.10 E-09	2.87 E-06	5.12 (0.43-61.37)	9.228438669811 52e+91 (189.69-4.489686777590	687	mlg-109	未分类的		

[0203]

					76e+181)					
219	4.04 E-08	1.80 E-09	2.53 E-07	4.31 (0.25-75.58)	6464869070139582 (0.7-5.9327349907021e+31)	328	mlg-219	未分类的		
223	2.90 E-09	2.00 E-10	5.31 E-07	3.23 (0.33-31.58)	4.18468389007774e+201 (868257309.25-Inf)	1407	mlg-223	未分类的		
1564	1.52 E-08	5.00 E-10	8.54 E-07	1.77 (0.11-29.19)	2.2802941587178e+217 (82198762116180.2-Inf)	140	mlg-1564	未分类的		
5045	1.30 E-07	2.08 E-08	2.37 E-07	1.7 (0.86-3.39)	3.13 (0.96-10.21)	187	mlg-5045	未分类的		
114	2.17 E-07	5.14 E-08	9.26 E-07	1.39 (0.73-2.64)	4.77 (0.62-36.99)	562	mlg-114	未分类的		
166	3.00 E-10	7.00 E-10	3.10 E-07	0.88 (0.53-1.45)	4.50269368364809e+58 (189.1-1.0721382304530 8e+115)	345	mlg-166	未分类的		
135	4.00 E-09	4.82 E-08	1.02 E-06	0.59 (0.08-4.51)	3.31 (0.71-15.34)	1502	mlg-135	未分类的		

表 3: 15 个 MLG 种类的序列

MLG ID	SEQ ID NO:	基因数目
mlg_id:84	1-165	165
mlg_id:131	166-1349	1184
mlg_id:75	1350-1527	178
mlg_id:114	1528-2089	562
mlg_id:1564	2090-2229	140
mlg_id:135	2230-3731	1502
mlg_id:5045	3732-3918	187
mlg_id:121	3919-6548	2630
mlg_id:109	6549-7235	687
mlg_id:166	7236-7580	345
mlg_id:317	7581-7700	120

[0204]

mlg_id:219	7701-8028	328
mlg_id:2985	8029-9595	1567
mlg_id:100	9596-9891	296
mlg_id:223	9892-11298	1407

表 4-1: 96 个样品(训练集)中的 15 个 MLG 的相对丰度特征谱

编号 (NO. 1-41: CRC; NO. 42-96: CTRL)	样品 ID	MLG						
		121	2985	135	223	131	109	114
1	31446	3.64E-08	3.71E-08	1.70E-08	0	6.57E-09	1.64E-09	9.39E-10
2	31881	4.53E-07	2.82E-09	4.47E-06	2.92E-11	8.90E-08	2.70E-08	6.33E-09
3	31866	9.87E-07	2.32E-06	1.79E-07	4.84E-08	4.27E-07	2.66E-07	1.14E-07
4	31874	5.28E-09	2.72E-09	8.47E-11	4.33E-06	1.79E-09	7.03E-11	3.88E-10
5	31549	1.56E-08	1.41E-06	1.37E-08	8.54E-10	3.31E-09	4.59E-09	3.27E-09
6	31878	4.35E-07	3.17E-06	0	3.40E-06	1.14E-08	9.08E-08	3.67E-08
7	531281	8.32E-09	6.04E-09	1.18E-09	3.73E-10	3.71E-09	3.97E-11	2.86E-10
8	31877	9.23E-09	8.07E-09	5.15E-08	1.38E-08	3.67E-09	0	2.72E-10
9	530373	4.42E-09	1.67E-06	0	8.96E-11	1.11E-09	0	0
10	531155	1.19E-07	4.10E-06	1.91E-08	1.02E-09	5.84E-08	2.61E-08	1.51E-08
11	31868	4.12E-05	2.20E-06	7.21E-06	4.96E-08	1.82E-05	1.23E-05	4.39E-06
12	531361	0.00011555	2.34E-08	1.70E-05	1.69E-09	4.75E-05	3.28E-05	1.21E-05
13	531775	4.10E-08	8.73E-06	6.04E-09	0	1.81E-08	7.30E-09	4.68E-09
14	31865	2.08E-08	1.25E-08	2.17E-09	3.67E-09	9.18E-09	1.41E-09	1.83E-09
15	31223	4.96E-07	6.64E-10	2.63E-07	0	7.32E-08	9.57E-08	2.75E-08
16	531469	4.07E-08	3.22E-08	4.27E-09	5.17E-07	1.71E-08	6.47E-09	4.59E-09
17	31876	3.29E-08	1.22E-08	3.55E-09	3.59E-07	1.53E-08	7.31E-09	4.81E-09
18	531416	1.77E-08	1.16E-05	3.74E-09	1.94E-07	6.71E-09	1.25E-09	1.40E-09
19	31872	8.12E-08	1.66E-05	1.82E-08	0	3.81E-08	1.93E-08	1.18E-08
20	31276	1.03E-06	3.80E-08	4.49E-06	0	2.72E-07	3.28E-07	5.37E-07

[0205]

21	531333	2.06E-08	2.07E-08	4.14E-09	0	1.05E-08	2.75E-09	1.86E-09
22	531382	0	7.51E-09	0	0	0	0	0
23	530890	1.28E-09	0	0	1.46E-09	0	0	0
24	31237	1.64E-06	0	5.71E-08	0	5.17E-07	2.04E-07	2.66E-07
25	31004	0	0	0	0	0	0	0
26	531277	1.75E-08	1.03E-08	0	1.12E-08	0	1.40E-09	1.07E-09
27	31871	5.00E-09	1.46E-06	0	0	0	2.38E-09	2.14E-10
28	31188	1.37E-06	6.23E-09	1.36E-07	0	9.90E-09	3.48E-07	1.19E-07
29	31875	5.42E-10	3.45E-06	0	1.20E-08	0	0	0
30	31884	1.48E-08	2.63E-06	0	0	0	8.15E-09	4.46E-09
31	31285	3.27E-08	8.48E-09	4.04E-09	5.89E-09	0	1.35E-08	9.36E-09
32	31489	1.53E-06	2.71E-06	1.81E-11	5.33E-09	1.54E-09	1.16E-06	2.83E-07
33	31685	2.45E-08	1.73E-09	0	1.16E-09	0	1.83E-08	6.27E-09
34	531248	8.26E-05	1.06E-05	5.69E-10	0	1.18E-07	6.05E-05	1.47E-05
35	31870	9.25E-08	2.10E-08	3.45E-10	0	0	7.15E-08	2.13E-08
36	531128	2.88E-06	2.20E-08	7.93E-06	2.61E-08	8.09E-07	4.41E-07	2.64E-06
37	31873	0	1.45E-06	0	4.47E-06	0	0	0
38	31880	0	2.62E-05	0	0	0	0	0
39	531766	6.15E-08	2.31E-07	0	3.24E-08	0	1.46E-08	4.21E-09
40	31883	3.34E-05	1.14E-05	0	8.27E-06	2.44E-08	8.85E-06	2.65E-06
41	531352	1.25E-08	1.19E-05	0	2.17E-09	0	8.91E-10	9.75E-10
42	31766	0	6.09E-09	0	0	0	0	0
43	31537	3.10E-08	4.51E-06	0	0	0	1.60E-09	9.02E-10
44	31600	0	0	0	0	0	0	0
45	31428	3.09E-08	8.25E-10	0	0	6.76E-12	3.73E-09	9.41E-10
46	530167	2.88E-08	1.20E-09	0	2.61E-10	0	1.55E-09	8.91E-10
47	530315	3.71E-07	0	0	0	4.20E-09	8.85E-09	1.63E-09
48	530050	0	5.32E-06	0	9.12E-09	0	0	0
49	31160	1.42E-09	2.19E-09	0	0	0	0	1.78E-09
50	530177	0	2.02E-09	0	0	0	0	0
51	31700	0	4.27E-09	0	0	0	0	0
52	31714	6.40E-08	1.41E-08	0	0	2.07E-09	1.40E-09	9.32E-07

[0206]

53	31219	0	2.64E-05	0	0	0	0	0
54	31557	0	3.95E-09	0	0	0	0	0
55	530154	3.16E-09	4.13E-08	0	0	0	0	0
56	530444	1.57E-06	0	0	0	1.50E-08	1.02E-07	7.99E-09
57	31021	0	0	0	0	0	0	0
58	530074	0	0	0	0	0	0	0
59	530227	0	0	0	0	0	0	0
60	530394	0	6.97E-07	0	0	0	0	0
61	530295	2.05E-08	0	0	0	2.88E-10	5.88E-10	1.46E-08
62	530368	0	1.27E-06	0	0	0	0	0
63	31300	0	0	0	0	0	0	0
64	31452	0	0	0	0	0	0	0
65	31723	0	0	0	0	0	0	0
66	531424	0	0	0	0	0	0	0
67	31009	0	0	0	0	0	0	0
68	530251	0	0	0	0	0	0	0
69	31416	0	1.10E-07	0	0	0	0	0
70	530119	0	0	0	0	0	0	0
71	31749	0	0	2.31E-06	0	0	0	0
72	530364	0	4.45E-06	0	0	0	0	0
73	31328	0	2.72E-07	0	0	0	0	0
74	530163	0	0	0	0	0	0	0
75	31379	1.55E-08	0	0	0	0	2.24E-10	2.15E-09
76	31112	1.93E-08	0	0	0	0	2.14E-10	2.61E-09
77	31711	0	2.96E-08	0	0	0	0	0
78	31750	0	0	0	0	0	0	0
79	530451	4.29E-07	1.26E-10	0	0	3.98E-08	4.88E-08	1.86E-06
80	31129	5.68E-08	2.63E-09	0	0	1.05E-09	0	0
81	530052	0	2.67E-09	0	0	0	0	0
82	31512	0	9.36E-06	0	0	0	0	0
83	31236	0	1.73E-09	0	0	0	0	0
84	31360	0	1.35E-05	0	0	0	0	0

[0207]

85	31343	0	1.66E-09	0	0	0	0	0
86	530215	0	7.59E-10	0	0	0	0	0
87	31519	0	0	3.41E-07	0	0	0	0
88	31450	0	0	0	0	0	0	0
89	531414	0	0	0	0	0	0	0
90	31071	0	0	0	0	0	0	0
91	530331	4.72E-10	0	0	0	0	0	0
92	530258	0	0	0	0	0	0	0
93	31333	0	0	0	0	0	0	0
94	31232	0	0	0	0	0	0	0
95	530055	4.15E-10	0	0	0	0	0	0
96	31267	0	0	0	0	0	0	0

表4-2：96个样品(训练集)中的15个MLG的相对丰度特征谱

编号 (NO. 1-41: CRC; NO. 42-96: CTRL)	样品 ID	MLG							
		166	219	100	5045	75	84	1564	317
1	31446	8.76E-10	1.46E-09	1.83E-10	0	3.78E-09	3.79E-10	5.59E-09	2.30E-08
2	31881	6.59E-10	3.26E-08	2.20E-09	0	8.19E-08	3.65E-08	1.05E-09	4.40E-07
3	31866	6.64E-09	1.39E-08	3.84E-10	0	9.07E-08	3.16E-08	1.68E-06	1.06E-08
4	31874	2.40E-09	4.88E-07	0	5.24E-09	1.13E-08	4.12E-08	2.45E-09	3.18E-08
5	31549	0	0	1.21E-08	4.68E-08	2.55E-10	1.92E-09	3.57E-09	1.16E-07
6	31878	4.95E-07	6.70E-08	1.29E-07	0	7.98E-09	6.24E-08	0	1.27E-08
7	531281	4.38E-09	0	0	6.03E-09	6.30E-09	8.64E-09	6.32E-09	3.00E-08
8	31877	4.31E-10	1.98E-08	0	1.27E-06	2.26E-09	0	5.71E-09	4.16E-06
9	530373	0	1.95E-10	0	1.03E-09	5.36E-06	1.73E-06	6.19E-09	6.78E-09
10	531155	1.04E-10	0	0	2.06E-10	6.59E-09	5.60E-09	2.90E-07	6.04E-08
11	31868	1.01E-06	4.41E-09	2.31E-09	8.17E-09	3.01E-08	8.16E-09	2.62E-05	4.45E-07

[0208]

12	531361	5.88E-09	4.01E-10	5.35E-09	0	0	0	5.94E-06	7.43E-09
13	531775	3.08E-08	3.43E-07	0	2.41E-09	0	3.02E-08	4.37E-08	3.53E-08
14	31865	6.51E-09	0	0	2.93E-07	1.57E-08	1.18E-08	1.69E-08	2.72E-08
15	31223	5.75E-08	9.01E-09	1.43E-07	4.10E-10	2.09E-09	2.09E-07	2.07E-08	7.67E-09
16	531469	1.62E-08	2.03E-09	0	5.96E-09	7.31E-08	1.92E-08	3.45E-08	3.08E-08
17	31876	1.03E-05	1.07E-08	0	0	7.33E-09	0	8.15E-08	2.26E-06
18	531416	6.36E-07	8.23E-10	0	1.80E-09	5.87E-08	6.25E-08	7.67E-08	1.10E-06
19	31872	7.96E-09	1.09E-09	0	1.53E-09	0	2.60E-08	2.21E-07	5.11E-09
20	31276	2.48E-08	0	5.47E-07	9.37E-11	0	0	2.29E-07	1.38E-08
21	531333	2.10E-08	0	0	1.17E-06	0	0	1.25E-07	1.43E-08
22	531382	3.82E-10	1.61E-08	0	9.65E-10	0	3.49E-10	0	3.75E-08
23	530890	5.16E-10	5.35E-09	5.43E-10	1.92E-10	0	0	6.17E-11	6.91E-07
24	31237	0	0	2.36E-08	0	0	1.50E-08	0	8.43E-09
25	31004	0	5.58E-09	0	1.78E-10	4.00E-09	2.04E-09	0	3.28E-08
26	531277	1.82E-10	6.93E-07	4.05E-09	8.23E-09	0	2.80E-09	0	4.19E-06
27	31871	0	0	1.05E-08	0	0	2.01E-09	0	3.95E-09
28	31188	0	1.49E-08	5.21E-07	8.76E-10	0	4.50E-09	0	1.99E-08
29	31875	1.15E-09	3.84E-09	0	1.02E-06	9.90E-08	1.13E-07	0	1.82E-07
30	31884	0	1.38E-09	1.81E-08	5.92E-09	0	0	8.93E-10	6.67E-09
31	31285	0	1.54E-09	3.09E-08	1.20E-06	0	0	0	1.18E-09
32	31489	3.22E-10	0	2.00E-06	4.04E-06	1.02E-08	4.46E-07	0	2.35E-08
33	31685	0	7.77E-10	3.58E-08	5.37E-09	0	0	0	1.07E-09
34	531248	7.31E-10	2.04E-10	0.000109	3.96E-09	5.12E-08	1.50E-07	0	6.12E-07
35	31870	0	0	1.34E-07	2.19E-08	1.26E-08	1.12E-08	4.60E-10	9.16E-11
36	531128	4.90E-10	1.78E-06	2.66E-07	8.40E-09	1.40E-09	3.65E-09	0	6.15E-09
37	31873	3.04E-09	3.69E-08	0	0	1.76E-07	1.77E-07	0	8.26E-07
38	31880	0	6.88E-08	0	0	5.68E-08	4.19E-08	0	4.58E-06
39	531766	6.19E-10	6.55E-06	2.00E-08	5.76E-07	3.27E-10	0	0	1.38E-05
40	31883	4.45E-08	1.94E-07	1.05E-05	2.46E-08	1.95E-07	2.09E-07	0	2.01E-06
41	531352	3.49E-11	2.64E-09	2.84E-09	0	0	0	5.66E-09	2.27E-08
42	31766	0	0	0	0	0	0	0	0
43	31537	0	0	7.30E-09	0	0	9.21E-09	0	0

[0209]

44	31600	2.87E-08	0	0	0	0	0	0	0	0
45	31428	0	0	9.30E-09	6.89E-07	1.09E-10	0	0	0	0
46	530167	4.00E-09	1.31E-09	5.04E-09	0	0	0	1.37E-08	9.13E-09	
47	530315	0	8.68E-10	0	0	0	0	0	6.18E-09	
48	530050	2.96E-10	0	0	0	0	0	1.06E-09	2.83E-09	
49	31160	0	0	0	0	0	0	0	0	
50	530177	0	0	0	0	0	0	0	0	
51	31700	1.98E-10	0	0	0	2.29E-10	1.97E-09	0	5.58E-08	
52	31714	0	0	0	0	0	5.69E-09	0	0	
53	31219	0	0	0	2.74E-08	0	0	0	2.05E-10	
54	31557	0	0	0	0	0	0	0	0	
55	530154	0	0	0	0	0	1.18E-08	0	5.25E-10	
56	530444	0	0	2.68E-09	0	0	1.75E-09	0	1.28E-09	
57	31021	0	0	0	0	0	0	0	0	
58	530074	0	0	0	0	0	0	0	0	
59	530227	0	0	0	0	0	0	0	0	
60	530394	0	7.62E-10	0	4.28E-07	0	0	0	0	
61	530295	0	0	6.27E-10	0	0	0	0	2.03E-07	
62	530368	0	0	0	0	0	0	0	0	
63	31300	0	1.81E-11	0	0	0	0	0	2.77E-09	
64	31452	0	0	0	0	0	0	0	0	
65	31723	0	1.86E-08	0	0	0	1.35E-09	0	0	
66	531424	0	0	0	0	2.73E-09	0	0	1.22E-07	
67	31009	0	0	0	0	0	0	0	0	
68	530251	0	0	0	0	0	0	0	0	
69	31416	0	0	0	0	0	0	0	0	
70	530119	0	7.31E-08	0	0	0	0	0	9.56E-10	
71	31749	0	0	0	0	0	0	0	2.00E-09	
72	530364	0	0	0	0	0	0	0	9.83E-10	
73	31328	0	0	0	0	8.27E-11	0	0	9.12E-07	
74	530163	0	0	0	0	0	0	0	0	
75	31379	0	0	2.84E-09	0	3.54E-10	5.94E-09	0	1.75E-06	

[0210]

76	31112	0	0	4.52E-09	0	0	1.20E-08	0	5.45E-08
77	31711	0	0	0	0	0	0	0	0
78	31750	0	0	0	0	0	7.99E-10	1.12E-08	4.99E-07
79	530451	0	0	2.99E-08	0	4.00E-10	8.52E-10	0	0
80	31129	0	4.58E-11	0	0	0	3.01E-09	0	0
81	530052	0	0	0	0	0	0	0	0
82	31512	0	0	0	0	0	0	0	0
83	31236	0	2.56E-09	0	0	0	0	0	1.48E-06
84	31360	3.08E-09	1.08E-10	0	0	0	0	0	8.08E-10
85	31343	0	0	0	0	0	0	0	5.12E-10
86	530215	0	3.67E-10	0	0	0	0	0	0
87	31519	0	0	0	0	0	6.02E-10	0	7.03E-08
88	31450	0	0	0	0	0	0	0	1.47E-05
89	531414	0	0	0	0	0	0	0	5.08E-07
90	31071	0	0	0	0	0	0	0	2.25E-10
91	530331	0	0	0	0	0	0	0	7.25E-09
92	530258	0	0	0	0	0	3.09E-09	0	2.48E-07
93	31333	0	0	0	0	0	0	0	0
94	31232	0	0	0	0	0	0	0	0
95	530055	0	0	0	0	0	0	0	8.87E-10
96	31267	0	0	0	0	0	0	0	5.32E-08

表 5：156 个样品中的 15 个 MLG 的预测结果

样品	样品 ID	状态	癌的概率	集
训练集 (96 个样品)	31766	对照	0.0054	训练
	31537	对照	0.3462	训练
	31600	对照	0.1117	训练
	31428	对照	0.4737	训练
	530167	对照	0.6935	训练

[0211]

530315	对照	0.3333	训练
530050	对照	0.4316	训练
31160	对照	0.0183	训练
530177	对照	0.0000	训练
31700	对照	0.2552	训练
31714	对照	0.2000	训练
31219	对照	0.2599	训练
31557	对照	0.0000	训练
530154	对照	0.1211	训练
530444	对照	0.3535	训练
31021	对照	0.0000	训练
530074	对照	0.0000	训练
530227	对照	0.0000	训练
530394	对照	0.2500	训练
530295	对照	0.1307	训练
530368	对照	0.0123	训练
31300	对照	0.0000	训练
31452	对照	0.0000	训练
31723	对照	0.2102	训练
531424	对照	0.2034	训练
31009	对照	0.0000	训练
530251	对照	0.0000	训练
31416	对照	0.0000	训练
530119	对照	0.2022	训练
31749	对照	0.1514	训练
530364	对照	0.0378	训练
31328	对照	0.0578	训练
530163	对照	0.0000	训练
31379	对照	0.0778	训练

[0212]

31112	对照	0.0824	训练
31750	对照	0.1568	训练
530451	对照	0.3918	训练
31129	对照	0.0511	训练
530052	对照	0.0000	训练
31512	对照	0.0272	训练
31236	对照	0.1353	训练
31360	对照	0.1530	训练
31343	对照	0.0000	训练
530215	对照	0.0211	训练
31519	对照	0.1852	训练
31450	对照	0.0734	训练
531414	对照	0.0000	训练
31071	对照	0.0000	训练
530331	对照	0.0244	训练
530258	对照	0.0594	训练
31333	对照	0.0000	训练
31232	对照	0.0000	训练
530055	对照	0.0052	训练
31267	对照	0.0000	训练
31711	对照	0.0000	训练
31446	癌	0.6963	训练
31881	癌	0.8757	训练
31866	癌	0.9789	训练
31874	癌	0.9244	训练
31549	癌	0.8722	训练
31878	癌	0.9162	训练
531281	癌	0.8226	训练
31877	癌	0.8514	训练

[0213]

530373	癌	0.6813	训练
531155	癌	0.9144	训练
31868	癌	0.9946	训练
531361	癌	0.7990	训练
531775	癌	0.9492	训练
31865	癌	0.9694	训练
31223	癌	0.8619	训练
531469	癌	0.9944	训练
31876	癌	0.9211	训练
531416	癌	0.9585	训练
31872	癌	0.9063	训练
31276	癌	0.8421	训练
531333	癌	0.7713	训练
531382	癌	0.4703	训练
530890	癌	0.5179	训练
31237	癌	0.4641	训练
31004	癌	0.2961	训练
531277	癌	0.8225	训练
31871	癌	0.1320	训练
31188	癌	0.8525	训练
31875	癌	0.8128	训练
31884	癌	0.6685	训练
31285	癌	0.7967	训练
31489	癌	0.9053	训练
31685	癌	0.5916	训练
531248	癌	0.9179	训练
31870	癌	0.6898	训练
531128	癌	0.9459	训练
31873	癌	0.7644	训练

[0214]

31880	癌	0.5561	训练	
531766	癌	0.9372	训练	
31883	癌	0.9412	训练	
531352	癌	0.4737	训练	
530075	对照	0.0540	测试	
31637	对照	0.0320	测试	
31398	进展性腺癌	0.7140	测试	
531403	进展性腺癌	0.7520	测试	
530168	进展性腺癌	0.2980	测试	
530185	进展性腺癌	0.2220	测试	
530600	进展性腺癌	0.7080	测试	
31477	进展性腺癌	0.5560	测试	
530403	进展性腺癌	0.0600	测试	
530002	进展性腺癌	0.0860	测试	
530697	进展性腺癌	0.2620	测试	
31455	进展性腺癌	0.5800	测试	
530756	进展性腺癌	0.0920	测试	
31424	进展性腺癌	0.1300	测试	
31501	进展性腺癌	0.3240	测试	
31256	进展性腺癌	0.1420	测试	
530297	进展性腺癌	0.5340	测试	
530142	进展性腺癌	0.5360	测试	
530026	进展性腺癌	0.4320	测试	
530558	进展性腺癌	0.2140	测试	
530054	进展性腺癌	0.3220	测试	
530743	进展性腺癌	0.7300	测试	
530867	进展性腺癌	0.2860	测试	
530705	进展性腺癌	0.2820	测试	
31337	进展性腺癌	0.0140	测试	
测试集 (60 个样品)				

[0215]

31030	进展性腺瘤	0.4040	测试
31282	进展性腺瘤	0.0480	测试
530028	进展性腺瘤	0.5340	测试
31449	进展性腺瘤	0.2800	测试
31275	进展性腺瘤	0.4340	测试
530018	进展性腺瘤	0.0200	测试
530262	进展性腺瘤	0.0540	测试
530039	进展性腺瘤	0.4000	测试
530172	进展性腺瘤	0.0180	测试
31137	进展性腺瘤	0.0000	测试
31431	进展性腺瘤	0.0440	测试
31233	进展性腺瘤	0.0000	测试
530398	进展性腺瘤	0.1640	测试
31582	进展性腺瘤	0.1500	测试
530450	进展性腺瘤	0.0820	测试
530623	进展性腺瘤	0.1700	测试
530323	进展性腺瘤	0.1180	测试
530348	进展性腺瘤	0.0060	测试
530041	进展性腺瘤	0.0000	测试
31705	进展性腺瘤	0.2420	测试
530840	进展性腺瘤	0.0080	测试
31367	癌	0.9220	测试
31493	癌	0.9840	测试
31879	癌	0.4680	测试
531274	癌	0.9480	测试
31159	癌	0.8940	测试
532749	对照	0.0820	测试
532779	对照	0.5240	测试
532796	对照	0.2520	测试

	532802	对照	0.0860	测试
	532826	对照	0.3700	测试
	532832	进展性腺瘤	0.2500	测试
	532305	进展性腺瘤	0.4640	测试
	532915	对照	0.4940	测试
	531663	进展性腺瘤	0.3600	测试

表 6: 15 个 MLG 重要性的排序

重要性的排序	mlg ID	mlg 的数目	AUC
1	5045		
2	121	2	0.91751663
3	75	3	0.970731707
4	109	4	0.959645233
5	317	5	0.975609756
6	135	6	0.978713969
7	223	7	0.980044346
8	100	8	0.985365854
9	219	9	0.984035477
10	114	10	0.981818182
11	84	11	0.980931264
12	166	12	0.979157428
13	2985	13	0.987583149
14	131	14	0.986696231
15	1564	15	0.983370288

表 7-1: 与进展性腺瘤相关的 10 个最具判别性的 MLG(物种标志物)(富集: 分别在对照(CTRL)和腺瘤(AA)、对照(CTRL)和癌症(CRC)、腺癌(AA)和癌症(CRC)的对比中的 MLG 富集方向，其中+表示后组的富集，-表示前组的富集，0 表示两组之间无差异(p≥0.05，Wilcoxon 等级和检验，用于多重检验的针对对照的 Bonferroni 校正)

[0217]

MLG ID	p-值 (Kruskal)	p 值校正 (BH)	p-值 (Wilcoxon)			富集	分类器		平均秩			出现率		
			CTRL 对比 AA	CTRL 对对 CRC	AA 对比 CRC		CRC	AA	CTRL	AA	CRC	CTRL	AA	CRC
317	0.0000	0.0000	0.0016	0.0000	0.0412	+,+,+	是	是	49	74	93	0.51	0.88	1.00
1340	0.0000	0.0000	0.2241	0.0001	0.0000	0,+,+	非	是	64	52	95	0.76	0.62	0.98
3840	0.0000	0.0000	0.0076	0.0000	0.0910	-,-,0	非	是	89	64	48	0.80	0.40	0.20
665	0.0011	0.0017	0.0013	0.0111	1.0000	+,+,0	非	是	56	81	76	0.20	0.52	0.46
3770	0.0001	0.0004	0.0002	0.0143	0.8448	-,-,0	非	是	86	54	63	0.71	0.33	0.46
721	0.0000	0.0000	0.0003	0.0000	0.0929	+,+,0	非	是	51	74	90	0.13	0.50	0.63
711	0.0000	0.0000	0.0022	0.0000	0.2128	+,+,0	非	是	51	74	89	0.20	0.52	0.71
1738	0.0004	0.0008	0.0002	0.0905	0.2183	+,0,0	非	是	56	85	71	0.22	0.62	0.44
4668	0.0000	0.0002	0.0091	0.0000	0.8270	+,+,0	非	是	51	76	87	0.51	0.76	0.95
5954	0.0000	0.0001	0.0132	0.0000	0.2240	+,+,0	非	是	52	74	89	0.44	0.74	0.88

表 7-2：与进展性腺瘤相关的 10 个最具判别性的 MLG (物种标志物)

MLG ID	丰度平均值			优势比(95% CI)		基因数目	MLG 注释	最佳匹配的菌株 (核苷酸)	以高于65%同一性匹配的基因所占的分数	匹配基因的平均同一性(%)
	CTRL	AA	CRC	CTRL 对比 AA	CTRL 对比 CRC					
317	4.13 E-07	3.77 E-07	8.77 E-07	1.02 (0.68-1.53)	1.28 (0.8-2.05)	120	马赛拟杆菌	马赛拟杆菌	0.9583	99.54
1340	1.07 E-07	7.03 E-08	3.00 E-07	1.14 (0.76-1.72)	2.56 (1.16-5.63)	175	mlg-1340	解木聚糖拟杆菌 (Bacteroides xylanisolvens)	0.5600	99.33
3840	2.59 E-07	4.54 E-07	4.42 E-08	0.84 (0.53-1.32)	0.21 (0.02-2)	402	动物双歧杆菌 (Bifidobacteri	动物双歧杆菌	1.0000	99.41

[0218]

							um animalis)			
665	1.01 E-07	5.80 E-09	1.68 E-07	69.15 (1.94-2459.39)	81.33 (0.38-17357.41)	293	Paraprevotella clara	Paraprevotella clara	0.9078	98.55
3770	8.50 E-09	1.89 E-07	2.24 E-08	0 (0-0.99)	0.07 (0-1.63)	229	变异链球菌 (Streptococcus mutans)	变异链球菌	0.9825	99.20
721	7.07 E-07	2.60 E-09	9.04 E-07	6.2655720525973e+22 (0-1.852698400015441e+58)	4.61983727398773e+68 (390818561448.17-5.46107543076802e+125)	118	mlg-721	未分类的		
711	7.60 E-07	1.40 E-08	8.72 E-07	4397459.31 (0-16170363812236702)	4047307274.13 (8.33-1966686969666667776)	189	mlg-711	未分类的		
1738	7.77 E-07	1.07 E-07	3.01 E-07	3.46 (1-11.92)	1.36 (0.83-2.24)	966	mlg-1738	未分类的		
4668	1.25 E-06	3.24 E-07	2.19 E-06	2.18 (1.07-4.45)	3.45 (0.72-16.5)	456	mlg-4668	未分类的		
5954	1.63 E-05	4.84 E-06	5.07 E-06	1.24 (0.75-2.03)	1.01 (0.67-1.51)	119	mlg-5954	未分类的		

表 9：97 个样品(训练集)中的 10 个 MLG 的相对丰度特征谱

编号 NO.1-55: CTRL; NO. 56-97: AA	样品 ID	MLG									
		1738	4668	3840	665	3770	711	1340	317	5954	721
1	31766	0	0	1.23E-07	6.70E-09	2.44E-08	0	1.10E-08	0	6.84E-07	0
2	31537	1.12E-08	2.65E-06	5.07E-06	0	1.70E-07	0	3.53E-08	0	0	0
3	31600	0	5.44E-09	1.44E-06	0	5.45E-06	0	3.97E-09	0	2.13E-10	0
4	31428	0	0	9.22E-10	0	0	0	5.64E-10	0	0	0

[0219]

5	530167	0	2.66E-09	1.41E-09	0	8.32E-08	5.62E-09	1.94E-07	9.13E-09	0	2.39E-09
6	530315	0	0	1.73E-11	9.90E-09	3.54E-08	0	0	6.18E-09	1.27E-09	0
7	530050	0	0	0	0	0	0	2.62E-08	2.83E-09	0	0
8	31160	0	1.10E-08	0	1.99E-08	8.34E-10	5.39E-09	1.94E-08	0	0	0
9	530177	0	0	0	0	0	0	3.63E-09	0	0	0
10	31700	1.65E-07	3.43E-06	2.50E-08	0	0	0	6.77E-08	5.58E-08	0	0
11	31714	0	0	2.76E-09	0	1.18E-07	0	0	0	2.75E-07	0
12	31219	0	1.32E-06	3.99E-08	0	0	0	0	2.05E-10	0	0
13	31557	0	0	4.50E-07	0	0	0	2.44E-08	0	0	0
14	530154	0	0	1.02E-08	0	1.68E-08	0	0	5.25E-10	0	0
15	530444	0	0	0	0	4.28E-07	2.07E-08	6.32E-10	1.28E-09	8.93E-10	1.50E-09
16	31021	0	0	1.35E-09	0	1.04E-08	0	1.16E-09	0	0	0
17	530074	0	0	6.84E-10	3.78E-10	1.25E-06	0	0	0	0	0
18	530227	5.38E-07	0	0	0	9.70E-08	0	0	0	0	0
19	530394	1.02E-06	1.92E-07	0	0	0	5.88E-08	1.93E-06	0	0	1.54E-09
20	530295	0	1.25E-07	9.32E-07	0	0	7.94E-11	1.15E-08	2.03E-07	0	0
21	530368	0	0	3.89E-07	0	3.61E-07	0	0	0	0	0
22	31300	0	2.60E-06	8.30E-08	0	5.92E-09	0	2.09E-07	2.77E-09	0	0
23	31452	2.11E-07	1.19E-08	3.05E-08	0	2.44E-09	0	3.20E-09	0	0	0
24	31723	0	3.26E-08	3.58E-07	1.02E-08	2.82E-07	0	1.78E-08	0	7.15E-07	0
25	531424	0	2.49E-09	2.90E-11	0	4.13E-09	0	1.44E-07	1.22E-07	6.81E-11	0
26	31009	0	5.17E-08	7.20E-09	0	6.08E-11	2.74E-08	9.31E-09	0	0	1.98E-08
27	530251	0	0	2.81E-07	0	4.70E-10	0	0	0	0	0
28	31416	0	0	6.63E-10	0	0	0	1.47E-09	0	0	0
29	530119	3.98E-08	1.26E-10	1.01E-10	0	6.33E-09	0	5.49E-08	9.56E-10	0	0
30	31749	0	0	2.07E-09	0	2.33E-08	0	7.32E-10	2.00E-09	0	0
31	530364	0	5.23E-11	1.16E-09	0	0	0	2.31E-09	9.83E-10	1.03E-08	0
32	31328	1.92E-06	8.48E-10	1.89E-07	0	1.10E-08	0	5.55E-08	9.12E-07	2.38E-08	0
33	530163	0	1.96E-09	9.16E-10	0	8.78E-07	0	0	0	1.50E-08	0
34	31379	0	0	6.52E-07	8.22E-08	2.58E-07	0	2.34E-08	1.75E-06	0	0
35	31112	0	0	3.30E-10	4.88E-08	2.40E-08	0	1.19E-07	5.45E-08	0	0
36	31711	0	0	6.29E-06	0	2.06E-07	0	5.64E-09	0	7.28E-06	0

[0220]

37	31750	0	3.11E-06	2.07E-08	0	1.81E-08	0	6.31E-08	4.99E-07	3.32E-08	0
38	530451	0	2.24E-06	4.42E-09	9.76E-09	8.80E-08	5.44E-07	0	0	1.06E-07	1.69E-08
39	31129	0	3.35E-10	1.39E-06	0	2.11E-08	3.15E-09	0	0	0.000252	0
40	530052	0	0	5.93E-08	0	0	0	2.52E-10	0	5.44E-08	0
41	31512	2.21E-08	0	9.43E-10	0	1.64E-08	1.71E-09	6.74E-11	0	5.07E-08	0
42	31236	0	0	0	7.57E-08	6.07E-09	0	1.73E-07	1.48E-06	2.84E-08	0
43	31360	0	0	4.66E-11	0	7.28E-10	0	1.07E-07	8.08E-10	1.27E-06	0
44	31343	0	0	2.48E-09	0	4.38E-07	0	0	5.12E-10	3.49E-08	0
45	530215	0	1.95E-08	0	0	9.43E-09	0	8.44E-09	0	3.74E-08	0
46	31519	1.86E-07	4.90E-07	5.28E-10	0	3.23E-08	9.03E-08	1.14E-07	7.03E-08	0	9.42E-08
47	31450	0	3.62E-09	0	0	0	0	1.03E-08	1.47E-05	0	0
48	531414	0	0	8.88E-10	0	0	0	1.95E-08	5.08E-07	0	0
49	31071	4.25E-09	1.21E-08	1.79E-09	9.59E-09	1.79E-08	0	0	2.25E-10	0	0
50	530331	1.78E-06	5.17E-08	0	0	3.77E-09	0	3.10E-08	7.25E-09	3.18E-09	0
51	530258	0	1.42E-06	1.09E-07	0	7.71E-09	0	8.25E-08	2.48E-07	4.13E-06	0
52	31333	0	3.68E-09	6.95E-06	0	1.72E-09	0	1.28E-08	0	5.85E-11	0
53	31232	0	0	9.64E-10	0	0	0	5.29E-08	0	0	0
54	530055	2.94E-09	0	0	0	0	1.33E-08	6.80E-08	8.87E-10	9.44E-10	7.18E-09
55	31267	0	1.99E-08	6.98E-08	4.45E-08	0	0	1.54E-07	5.32E-08	0	0
56	31398	2.48E-07	3.46E-08	0	4.71E-09	0	0	2.31E-09	5.78E-08	7.14E-07	0
57	531403	0	0	0	2.70E-09	0	0	0	1.35E-08	2.36E-08	0
58	530168	9.72E-07	3.83E-07	2.29E-07	0	1.81E-09	9.48E-08	0	7.12E-09	0.000544	9.45E-08
59	530185	2.61E-07	4.76E-06	3.83E-09	4.22E-07	0	3.55E-07	1.25E-06	1.37E-06	1.33E-07	1.89E-08
60	530600	8.63E-06	6.37E-07	3.38E-08	9.69E-08	0	1.14E-08	1.03E-07	4.93E-09	6.24E-10	1.48E-09
61	31477	1.47E-09	6.39E-07	3.74E-09	9.08E-09	0	0	3.05E-10	1.71E-09	1.55E-09	0
62	530403	1.63E-09	0	0	0	2.87E-10	0	9.81E-10	7.04E-10	1.39E-09	0
63	530002	0	0	0	0	3.94E-09	0	0	5.69E-09	1.63E-07	0
64	530697	6.19E-07	1.22E-09	0	0	3.83E-08	8.23E-08	2.29E-08	8.19E-09	1.34E-07	8.76E-09
65	31455	0	3.49E-07	0	0	0	1.23E-07	1.10E-06	1.25E-06	1.50E-06	1.27E-08
66	530756	0	0	8.98E-09	0	4.47E-09	0	0	0	0	0
67	31424	4.60E-09	1.03E-08	0	6.31E-09	0	0	0	0	0	0
68	31501	0	0	0	3.19E-08	7.73E-09	0	1.62E-08	6.41E-10	0	0

[0221]

69	31256	0	4.76E-09	0	1.64E-08	0	0	0	6.17E-08	0	0
70	530297	0	2.24E-06	5.49E-07	4.05E-09	0	1.23E-07	4.28E-09	5.81E-09	1.25E-08	7.78E-09
71	530142	0	1.74E-09	0	4.31E-08	0	0	7.92E-09	1.33E-07	1.92E-08	0
72	530026	2.22E-08	2.95E-08	7.29E-07	6.84E-08	0	2.19E-08	1.92E-09	1.27E-09	0	1.37E-10
73	530558	4.67E-06	2.37E-06	0	0	0	0	2.53E-09	7.67E-10	0	0
74	530054	2.05E-07	4.02E-06	0	3.62E-09	0	1.38E-07	4.00E-09	5.49E-09	0	1.37E-07
75	530743	1.37E-06	2.21E-07	0	1.21E-06	0	0	2.54E-07	9.28E-09	0	0
76	530867	6.15E-07	4.25E-06	0	1.46E-06	0	1.50E-07	2.21E-07	7.23E-06	0	1.10E-08
77	530705	7.05E-06	4.75E-09	0	0	5.90E-08	1.31E-09	0	1.78E-08	1.49E-05	8.96E-10
78	31337	0	5.33E-09	2.31E-07	0	0	0	0	2.98E-09	0	0
79	31030	1.19E-06	3.45E-06	0	1.87E-07	9.99E-08	1.35E-09	1.78E-08	4.46E-09	1.65E-08	2.47E-09
80	31282	2.15E-07	1.28E-08	0	1.86E-07	0	8.23E-10	1.88E-09	9.17E-07	1.90E-08	2.01E-09
81	530028	7.69E-10	6.34E-06	4.31E-08	0	0	1.72E-09	1.13E-06	2.58E-06	2.17E-08	2.02E-09
82	31449	4.31E-08	4.85E-09	0	5.09E-08	0	3.01E-09	1.38E-09	1.01E-07	2.63E-08	2.29E-09
83	31275	0	1.52E-05	0	2.33E-08	0	3.04E-05	0	5.10E-08	1.30E-06	2.93E-05
84	530018	0	2.24E-10	0	0	1.35E-08	1.97E-09	1.53E-10	4.02E-10	9.62E-05	1.26E-09
85	530262	0	0	8.81E-10	1.23E-07	3.83E-09	0	1.60E-09	7.68E-07	3.09E-08	0
86	530039	0	0	0	0	0	0	0	2.23E-09	0	0
87	530172	0	0	2.20E-06	0	3.89E-09	0	0	0	4.74E-09	0
88	31137	0	0	8.47E-10	0	0	0	0	0	1.21E-09	0
89	31431	8.66E-07	2.36E-06	1.18E-08	0	1.97E-09	1.66E-07	3.05E-07	3.20E-07	3.43E-06	1.39E-08
90	31233	0	6.05E-11	5.75E-06	0	0	0	7.06E-10	0	1.06E-05	0
91	530398	3.92E-10	1.09E-06	0	6.78E-08	0	1.63E-09	3.35E-09	3.19E-08	1.71E-06	3.06E-09
92	31582	5.03E-07	3.47E-09	3.65E-10	0	1.14E-07	4.48E-08	0	1.68E-08	3.70E-09	3.49E-09
93	530450	3.03E-06	4.06E-07	1.09E-06	0	0	2.66E-10	0	8.70E-09	5.67E-10	0
94	530623	1.57E-06	2.92E-07	0	0	0	1.60E-07	1.89E-08	2.15E-06	9.73E-10	1.51E-08
95	530323	1.80E-09	0	0	3.08E-08	0	0	0	2.17E-09	6.75E-10	0
96	530348	2.35E-09	1.73E-08	1.94E-10	0	3.60E-09	0	0	7.32E-10	1.14E-05	0
97	530041	5.14E-07	3.45E-06	0	2.08E-07	0	5.03E-08	1.52E-08	1.78E-07	2.41E-08	4.41E-09

表 10: 127 个样品中的 10 个 MLG 的预测结果

[0222]

样品	样品 ID	状态	进展性腺瘤的概率	集
训练集 (97 个样品)	31766	对照	0.0688	训练
	31537	对照	0.0699	训练
	31600	对照	0.0856	训练
	31428	对照	0.2889	训练
	530167	对照	0.2959	训练
	530315	对照	0.5099	训练
	530050	对照	0.3613	训练
	31160	对照	0.4444	训练
	530177	对照	0.2912	训练
	31700	对照	0.4061	训练
	31714	对照	0.0914	训练
	31219	对照	0.5274	训练
	31557	对照	0.0760	训练
	530154	对照	0.1684	训练
	530444	对照	0.5155	训练
	31021	对照	0.1229	训练
	530074	对照	0.1190	训练
	530227	对照	0.4082	训练
	530394	对照	0.4686	训练
	530295	对照	0.3122	训练
	530368	对照	0.0611	训练
	31300	对照	0.1630	训练
	31452	对照	0.2462	训练
	31723	对照	0.0387	训练
	531424	对照	0.0884	训练
	31009	对照	0.2821	训练

[0223]

530251	对照	0.4615	训练
31416	对照	0.2607	训练
530119	对照	0.1828	训练
31749	对照	0.1818	训练
530364	对照	0.5621	训练
31328	对照	0.3667	训练
530163	对照	0.1838	训练
31379	对照	0.1946	训练
31112	对照	0.1508	训练
31750	对照	0.0941	训练
530451	对照	0.4346	训练
31129	对照	0.4010	训练
530052	对照	0.5401	训练
31512	对照	0.2486	训练
31236	对照	0.5000	训练
31360	对照	0.3492	训练
31343	对照	0.1288	训练
530215	对照	0.3660	训练
31519	对照	0.6313	训练
31450	对照	0.4874	训练
531414	对照	0.1297	训练
31071	对照	0.3240	训练
530331	对照	0.6867	训练
530258	对照	0.2193	训练
31333	对照	0.2036	训练
31232	对照	0.0209	训练
530055	对照	0.5901	训练
31267	对照	0.3929	训练
31711	对照	0.2209	训练

[0224]

31398	进展性腺瘤	0.8505	训练
531403	进展性腺瘤	0.8989	训练
530168	进展性腺瘤	0.8743	训练
530185	进展性腺瘤	0.8811	训练
530600	进展性腺瘤	0.6477	训练
31477	进展性腺瘤	0.6719	训练
530403	进展性腺瘤	0.5775	训练
530002	进展性腺瘤	0.6600	训练
530697	进展性腺瘤	0.5615	训练
31455	进展性腺瘤	0.6541	训练
530756	进展性腺瘤	0.1170	训练
31424	进展性腺瘤	0.4033	训练
31501	进展性腺瘤	0.1404	训练
31256	进展性腺瘤	0.7368	训练
530297	进展性腺瘤	0.6480	训练
530142	进展性腺瘤	0.6126	训练
530026	进展性腺瘤	0.4890	训练
530558	进展性腺瘤	0.4153	训练
530054	进展性腺瘤	0.7677	训练
530743	进展性腺瘤	0.5167	训练
530867	进展性腺瘤	0.7956	训练
530705	进展性腺瘤	0.6721	训练
31337	进展性腺瘤	0.3369	训练
31030	进展性腺瘤	0.7073	训练
31282	进展性腺瘤	0.9665	训练
530028	进展性腺瘤	0.5202	训练
31449	进展性腺瘤	0.9461	训练
31275	进展性腺瘤	0.8391	训练
530018	进展性腺瘤	0.2874	训练

[0225]

530262	进展性腺瘤	0.4877	训练	
530039	进展性腺瘤	0.4916	训练	
530172	进展性腺瘤	0.4525	训练	
31137	进展性腺瘤	0.3476	训练	
31431	进展性腺瘤	0.7656	训练	
31233	进展性腺瘤	0.2500	训练	
530398	进展性腺瘤	0.9209	训练	
31582	进展性腺瘤	0.5062	训练	
530450	进展性腺瘤	0.6867	训练	
530623	进展性腺瘤	0.7914	训练	
530323	进展性腺瘤	0.8802	训练	
530348	进展性腺瘤	0.6344	训练	
530041	进展性腺瘤	0.9672	训练	
V1	对照	0.0187	测试	测试集 (30 个样品)
V2	对照	0.0180	测试	
V3	对照	0.0218	测试	
V4	对照	0.0521	测试	
V5	对照	0.1012	测试	
V6	对照	0.1130	测试	
V7	对照	0.1425	测试	
V8	对照	0.1965	测试	
V9	对照	0.1997	测试	
V10	对照	0.2947	测试	
V11	对照	0.3476	测试	
V12	对照	0.5040	测试	
V13	进展性腺瘤	0.4572	测试	
V14	进展性腺瘤	0.4587	测试	
V15	进展性腺瘤	0.4743	测试	
V16	进展性腺瘤	0.4744	测试	

[0226]

V17	进展性腺瘤	0.4876	测试
V18	对照	0.5011	测试
V19	对照	0.5080	测试
V20	进展性腺瘤	0.5370	测试
V21	对照	0.5474	测试
V22	进展性腺瘤	0.6626	测试
V23	进展性腺瘤	0.6816	测试
V24	进展性腺瘤	0.6923	测试
V25	进展性腺瘤	0.6934	测试
V26	进展性腺瘤	0.8108	测试
V27	进展性腺瘤	0.9218	测试
V28	进展性腺瘤	0.9245	测试
V29	进展性腺瘤	0.9499	测试
V30	进展性腺瘤	0.9662	测试

表 11：10 个 MLG 重要性的排序

重要性的排序	mlg ID	mlg 的数目	AUC
1	317		
2	3770	2	0.782251082
3	3840	3	0.805194805
4	665	4	0.773160173
5	721	5	0.795238095
6	1738	6	0.780952381
7	1340	7	0.895670996
8	5954	8	0.896536797
9	711	9	0.884848485
10	4668	10	0.873809524

[0227] 因此，本发明人通过基于相关基因标志物的随机森林模型，鉴定并验证了用于结直肠癌的早期和无创性诊断的15个MLG和用于结直肠腺瘤的早期和无创性诊断的10个MLG。并且本发明人构建了基于这些相关肠道微生物群的评估结直肠癌和腺瘤的风险的方法。

[0228] 虽然已详细地显示和描述了说明性实施方案，但本领域技术人员将理解，上述实施方案是说明性的，并且不旨在以任何方式限制本公开，并且可在不脱离本公开的精神、原理和范围的情况下对实施方案进行改变、替代和修改。

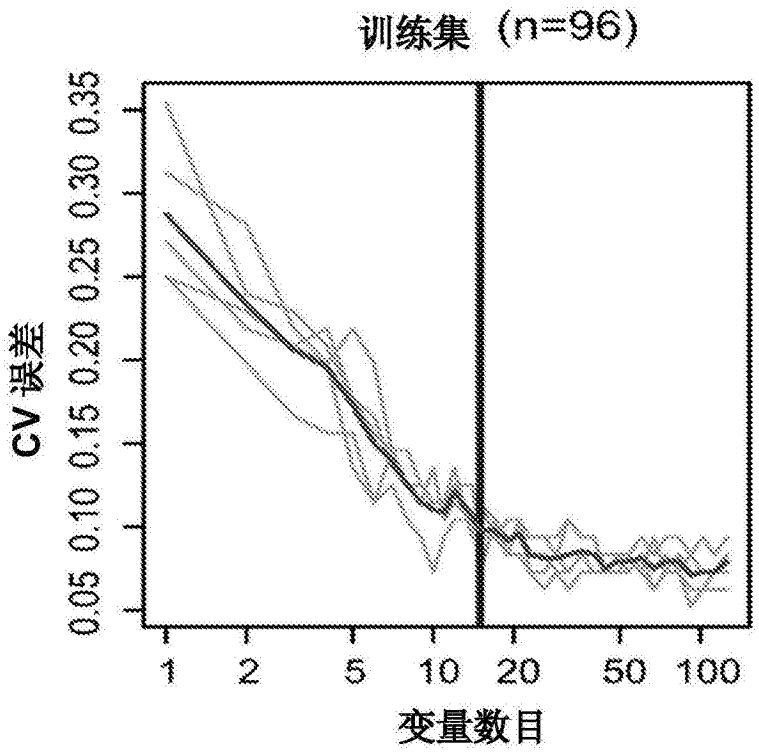
训练集 (n=96)
CV 误差
0.05 0.10 0.15 0.20 0.25 0.30 0.35
1 2 5 10 20 50 100
变量数目

图1a

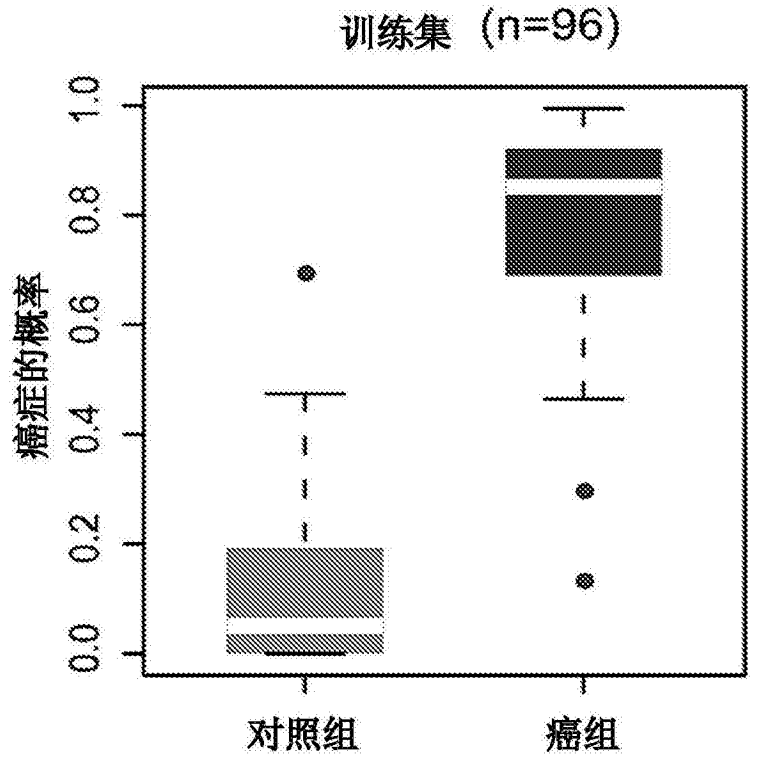
训练集 (n=96)
癌症的概率
0.0 0.2 0.4 0.6 0.8 1.0
对照组
癌组

图1b

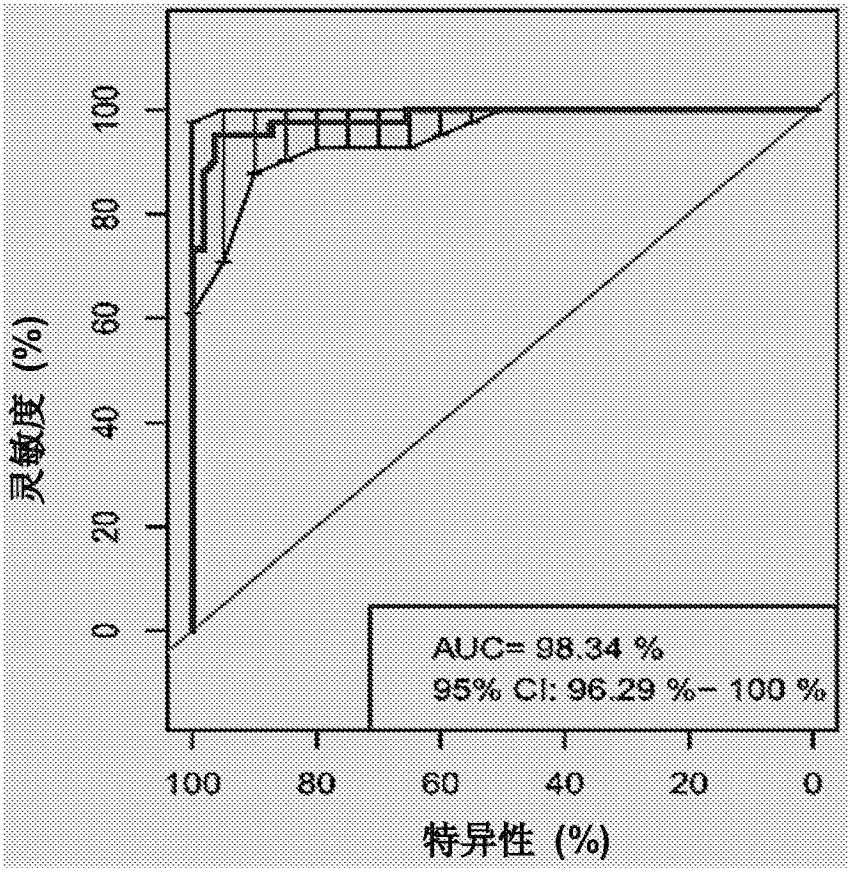

图1c

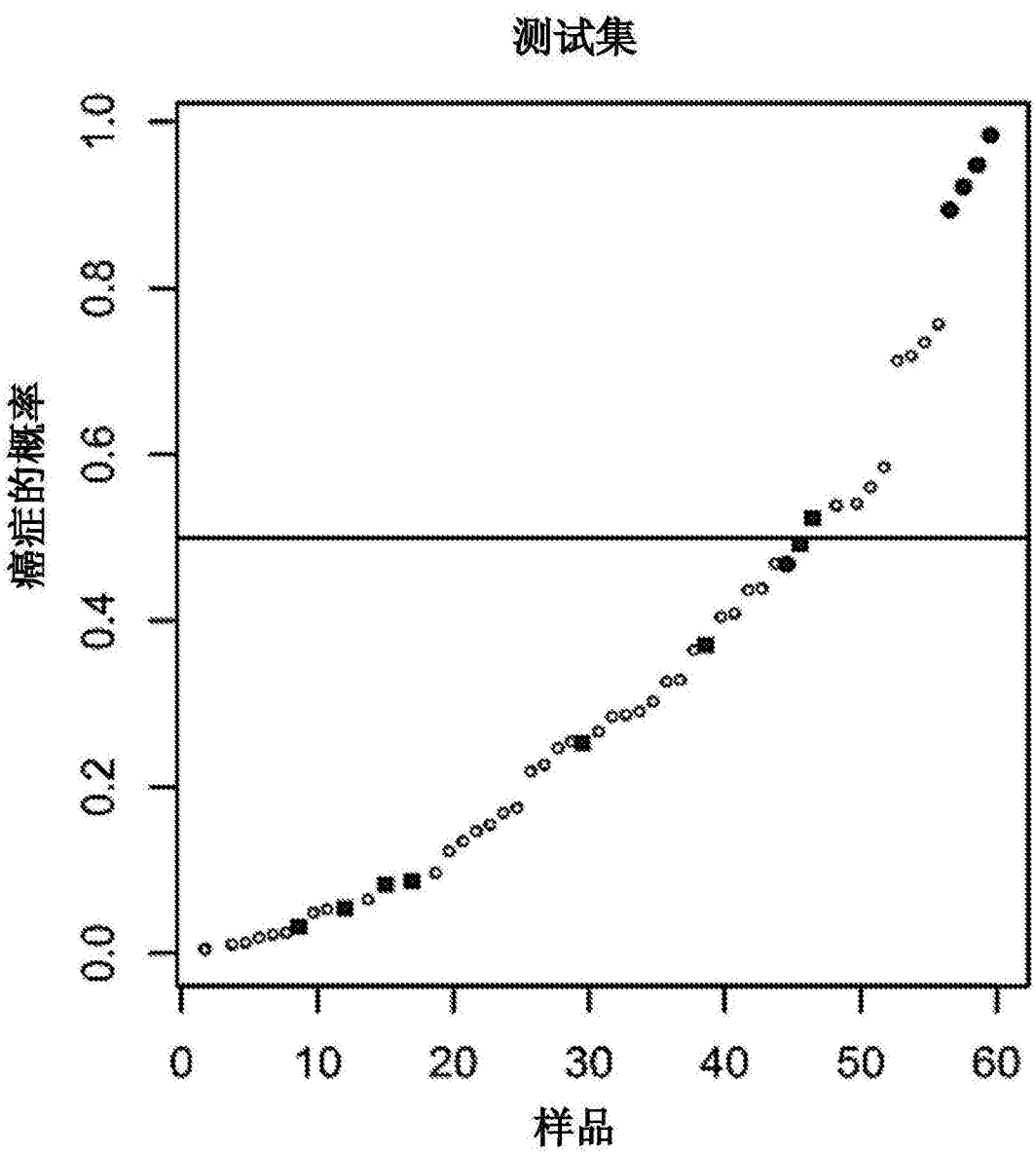

图1d

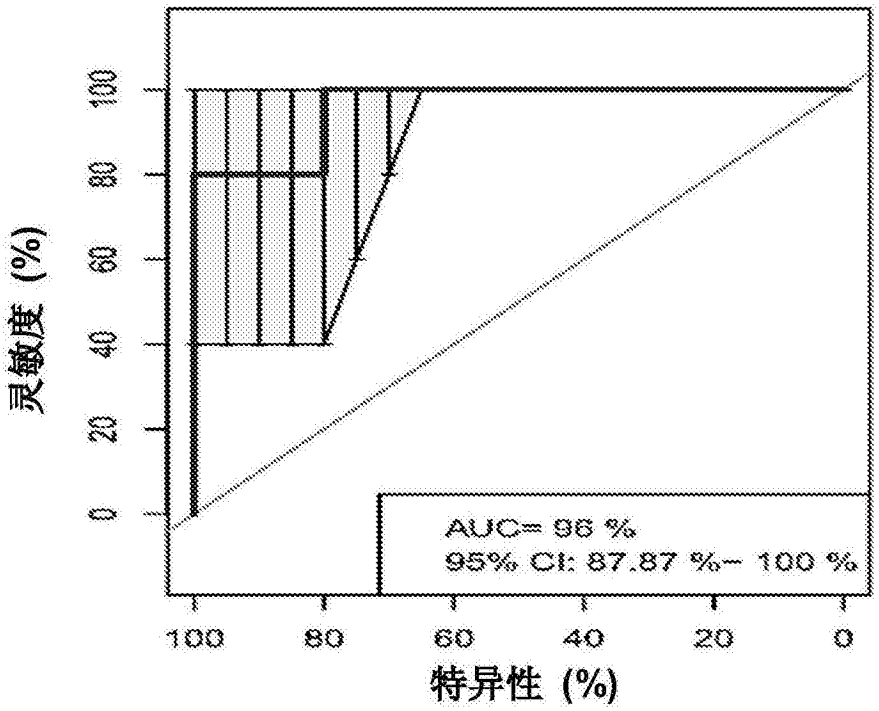
灵敏度 (%)
0
20
40
60
80
100
AUC= 96 %
95% CI: 87.87 %– 100 %
100
80
60
40
20
0
特异性 (%)

图1e

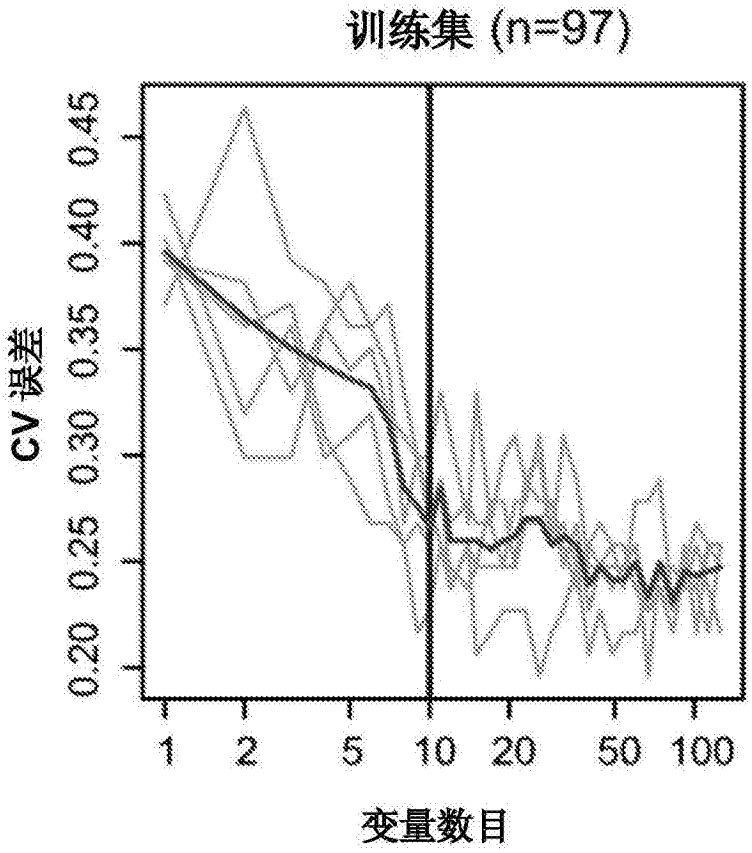
训练集 (n=97)
CV 误差
0.20
0.25
0.30
0.35
0.40
0.45
1
2
5
10
20
50
100
变量数目

图2a

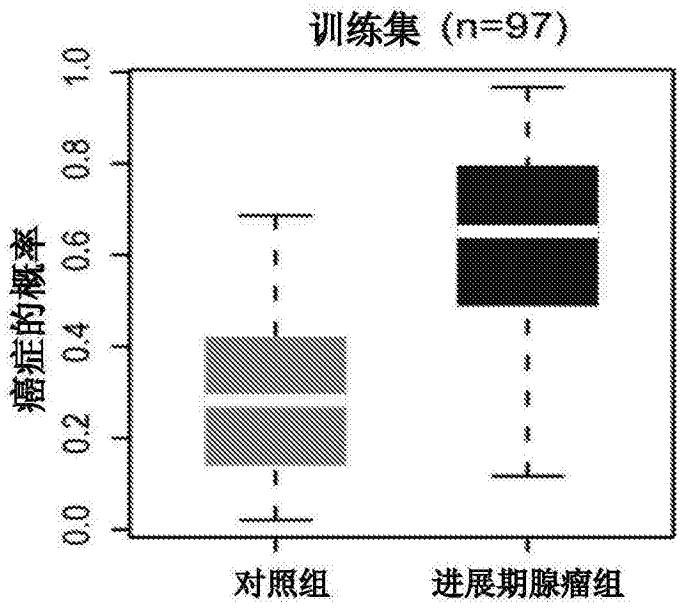
训练集 (n=97)
癌症的概率
0.0
0.2
0.4
0.6
0.8
1.0
对照组
进展期腺瘤组

图2b

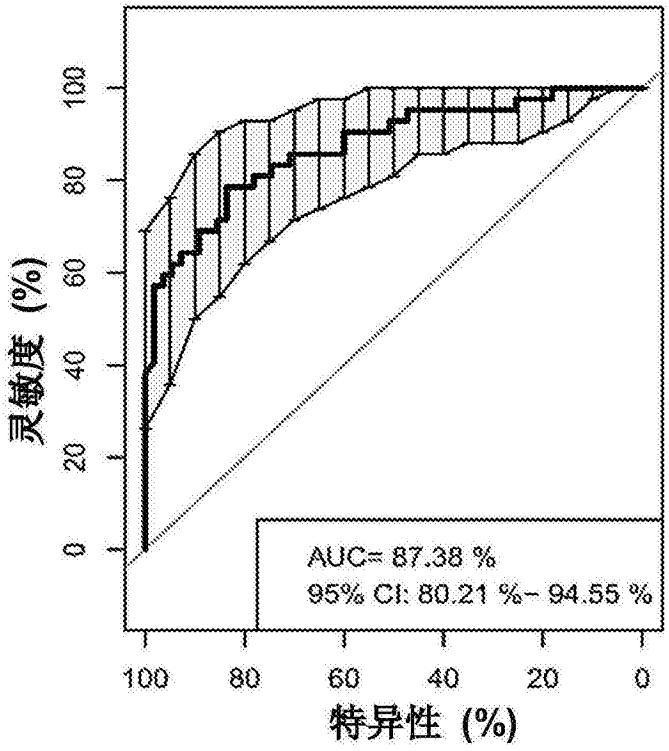
灵敏度 (%)
100
80
60
40
20
0
AUC= 87.38 %
95% CI: 80.21 %- 94.55 %
特异性 (%)

图2c

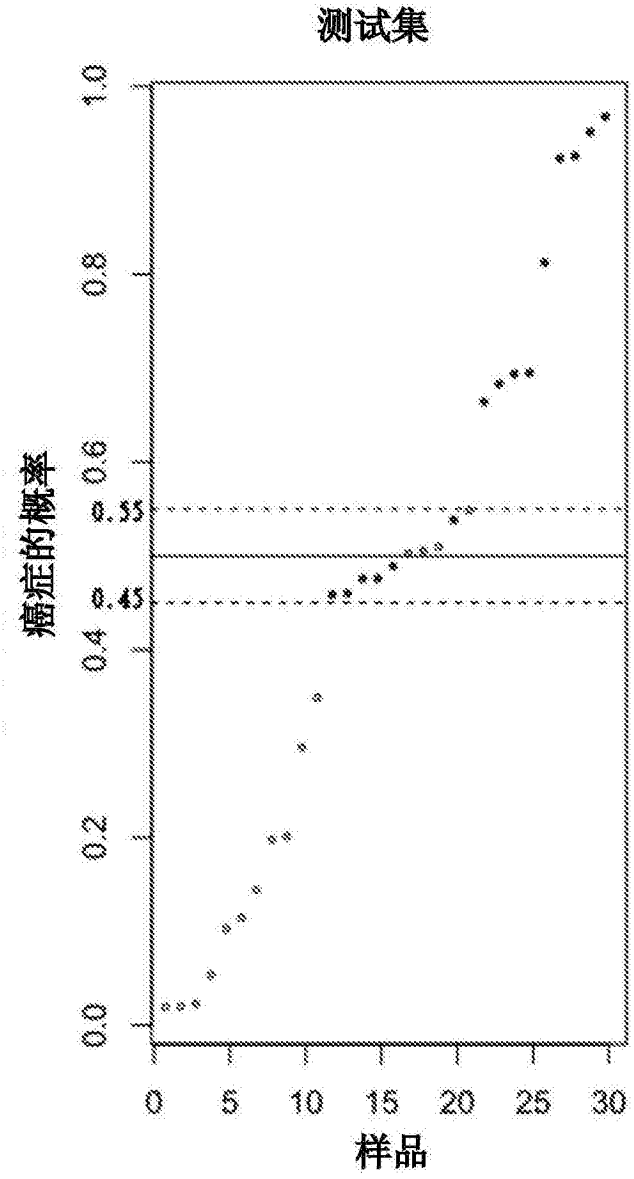

图2d

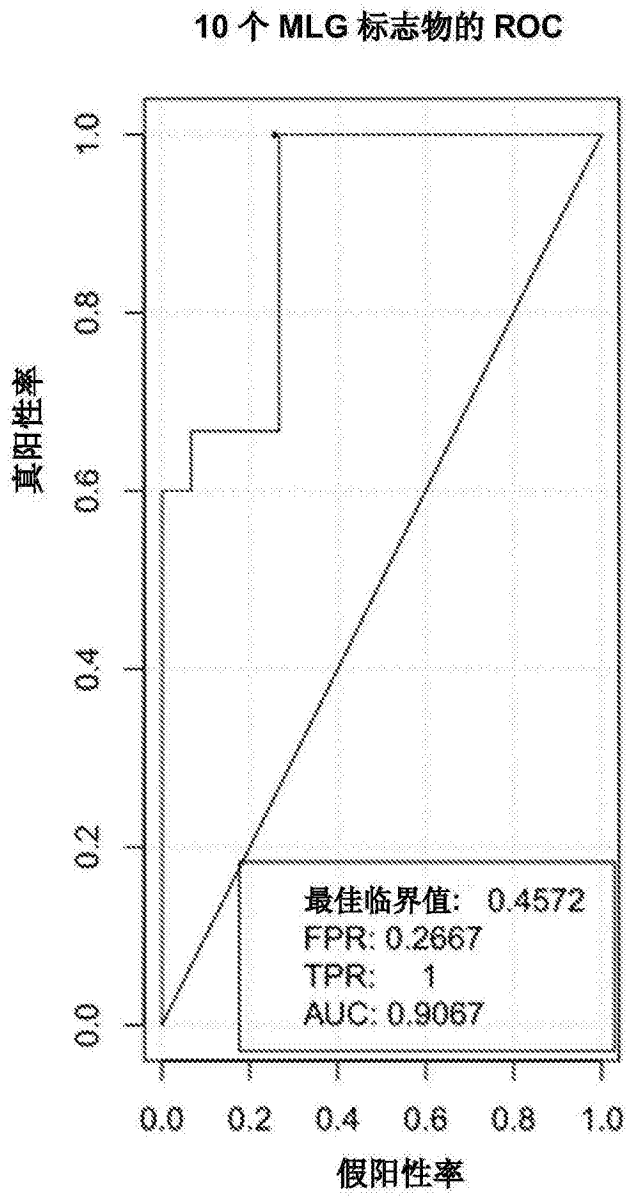
10 个 MLG 标志物的 ROC
真阳性率
假阳性率
0.0
0.2
0.4
0.6
0.8
1.0
最佳临界值: 0.4572
FPR: 0.2667
TPR: 1
AUC: 0.9067

图2e