(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 109504778 A
(43)申请公布日 2019.03.22

(21)申请号 201910026672.2

(22)申请日 2019.01.11

(71)申请人 复旦大学附属中山医院
地址 200032 上海市徐汇区枫林路180号

(72)发明人 许剑民 常文举 刘天宇 韦烨 任黎 何国栋 吉美玲 朱德祥 陈竟文 冯青阳

(74)专利代理机构 上海卓阳知识产权代理事务所(普通合伙) 31262
代理人 周春洪

(51)Int.Cl.
C12Q 1/6886(2018.01)
G16H 50/20(2018.01)

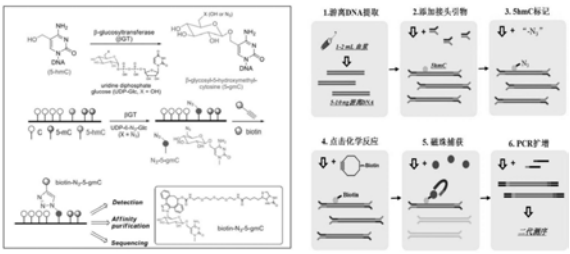
权利要求书2页 说明书9页 附图3页

(54)发明名称

一种基于表观修饰的5hmC多分子标志物及结直肠癌早期诊断模型

(57)摘要

本发明涉及临床分子诊断技术领域，具体公开一种基于表观修饰的5hmC多分子标志物及结直肠癌早期诊断模型。本发明通过5-hmC高通量测序技术，对结直肠癌患者与健康人外周血浆DNA全基因组5-hmC表达量的检查，对比两组样本中全基因组基因位点5-hmC表达差异，筛选表达差异最显著的部分基因位点作为该诊断模型的基因标志物，进而构建结直肠癌的早期诊断模型。经验证，本发明诊断的灵敏度、特异度均较高，与目前市场上的粪便隐血检测试剂盒、spetin9试剂盒相比，灵敏度和特异度分别提高约20%、8%。本发明的诊断方法具有检测无创、便捷等优点，可应用于临床一线，用于结直肠癌的早期筛查。



CN 109504778 A

1.一种基于表观修饰的5hmC多分子标志物结直肠癌早期诊断模型，其特征在于，包括检测受试者外周血浆中基因标志物5-hmC表达量的试剂，所述基因标志物包括：

GBX2 FAM84A FAM25B LCE1F FBXL7 DBX1 KRTAP27-1 AL353791.1 CEBPD LTB4R2 RP4-583P15.14 OR5B2 RPRM RNASE4 INSL5 AURKC IL36A AC017081.1 SPA17 NBPF12 FABP1 CST8。

2.根据权利要求1所述的诊断模型，其特征在于，还包括以下试剂：正常人外周血样本。

3.根据权利要求1所述的诊断模型，其特征在于，还包括记载有如下模型和评判方法的载体：

$$P=\frac{e^{(b_0+b_1x_1+b_2x_2+\cdots+b_kx_k)}}{1+e^{(b_0+b_1x_1+b_2x_2+\cdots+b_kx_k)}}$$

通过评分值判定样本来源是否为结直肠患者，当评分值大于0.5，则该受试者患有结直肠癌，当评分值小于或等于0.5，则该受试者为正常；

其中x_1,x_2,x_3...x_k分别为以上有意义基因标志物5-hmC表达量标准化结果；b_0为建模过程输出的常数项；b_1,b_2,b_3...b_k为对应基因标志物5-hmC表达量的系数，目前k＝22。

4.权利要求1-3任一所述诊断模型在制备结直肠癌早期诊断试剂盒中的应用。

5.一种基于表观修饰的5hmC多分子标志物结直肠癌早期诊断试剂盒，其特征在于，包括检测受试者外周血浆中基因标志物5-hmC表达量的试剂，所述基因标志物包括：

GBX2 FAM84A FAM25B LCE1F FBXL7 DBX1 KRTAP27-1 AL353791.1 CEBPD LTB4R2 RP4-583P15.14 OR5B2 RPRM RNASE4 INSL5 AURKC IL36A AC017081.1 SPA17 NBPF12 FABP1 CST8。

6.根据权利要求5所述的试剂盒，其特征在于，所述的试剂盒还包括记载有如下模型和评判方法的载体：

$$P=\frac{e^{(b_0+b_1x_1+b_2x_2+\cdots+b_kx_k)}}{1+e^{(b_0+b_1x_1+b_2x_2+\cdots+b_kx_k)}}$$

通过评分值判定样本来源是否为结直肠患者，当评分值大于0.5，则该受试者患有结直肠癌，当评分值小于或等于0.5，则该受试者为正常；

其中x_1,x_2,x_3...x_k分别为以上基因标志物5-hmC表达量标准化结果。

7.根据权利要求3所述的诊断模型、权利要求6所述的试剂盒，其特征在于，其中b_0,b_1,b_2,b_3...b_k的取值如下：

标志物	B系数
GBX2	1.31894902
FAM84A	1.21974841
FAM25B	1.15502531
LCE1F	1.10125199
FBXL7	1.08682839
DBX1	1.01982759

KRTAP27-1	0.52487632
AL353791.1	0.40010969
CEBPD	0.38147623
LTB4R2	0.34240677
RP4-583P15.14	0.24499059
OR5B2	0.08550743
RPRM	0.02997599
RNASE4	-0.0157149
INSL5	-0.0274637
AURKC	-0.1550417
IL36A	-0.1906636
AC017081.1	-0.4367674
SPA17	-0.5486784
NBPF12	-0.9815598
FABP1	-1.0037715
CST8	-1.035181
常数项	-23.890494

。

8.根据权利要求3所述的诊断模型、权利要求6所述的试剂盒，其特征在于，所述标准化是指将获得的测序结果进行初步质控评估，清除低质量测序位点后，将达到测序质量标准的读段利用Bowtie2工具与人类标准基因组参考序列进行比较。然后利用featureCounts和HtSeq-Count工具来统计读段数量以确定各基因标志物的5-hmC含量。

9.根据权利要求5-6任一所述试剂盒，其特征在于，还包括以下试剂：正常人外周血样本。

一种基于表观修饰的5hmC多分子标志物及结直肠癌早期诊断模型

技术领域

[0001] 本发明涉及临床生物医学技术领域，具体地说，是一种基于表观修饰的5hmC多分子标志物及结直肠癌早期诊断模型。

背景技术

[0002] 结直肠癌(colorectal cancer,CRC)是消化道常见的恶性肿瘤之一，全世界结直肠癌的发病率和死亡率处于恶性肿瘤的第三位。我国情况亦类似，结直肠癌是死亡率分别为男性第五位和女性第四位的恶性肿瘤。并且由于结直肠静脉回流的解剖学特点等原因，约有25％的患者在就诊同时即合并肝转移，使结直肠癌患者5年生存率明显下降。在结直肠癌早期阶段即确诊并接受治疗的患者5年生存率明显高于IV期患者。因此，在结直肠癌发展的早期阶段确诊在提高患者的长期生存中尤为关键。

[0003] 结直肠癌的筛查工作在早期诊断、提高患者长期生存中尤为重要。以往结直肠癌筛查方法包括结肠镜检查、软式乙状结肠镜检查(SIG)、计算机断层结肠成像(CTC)、基于愈创树脂的粪便隐血检测(gFOBT)、粪便免疫化学检测(FIT)、多靶点粪便DNA检测(sDNA) 和血浆DNA中sept9甲基化检测(mSEPT9)。临床研究证实接受SIG筛查与无筛查的人群相比可明显降低结直肠癌发生率和死亡率，但获益更多局限于远端的结直肠癌，受试者的依从性为58％-83.5％并且有穿孔、出血等风险。CTC对直径大于等于10mm的新生物有相对较好的检出灵敏度和特意度，几项临床研究证实在肠道准备后的CTC检测中灵敏度为 67％-100％，特异度为96％-98％；而在没有肠道准备的患者中灵敏度和特异度分别为 67％-90％和85％-97％，但接受CTC检测的患者要接受辐射。在包括FIT和sDNA的粪便筛查试验表明，由于研究方式、人群的不同，检出的灵敏度、特异度波动范围较大，检出结果稳定性相对较差。美国的一项临床研究表明mSEPT9对结直肠癌患者的检出率灵敏度为 48.2％和91.5％。虽然肠镜检查是结直肠癌诊断的金标准，但由于本检测需要清理肠道、预约时间长、存在出血穿孔等风险，患者对于结肠镜的依从性偏低。而其他目前常用的筛查方法虽然经济、便捷，但存在灵敏度、特异度偏低、检测结果不稳定等诸多缺陷。使结直肠癌早期诊断困难重重，影响患者的早期检出率，致使结直肠癌患者生存率降低。结直肠癌的早期筛查工作亟待解决，迫切需要一种微创、灵敏、高效的早期筛查手段。

[0004] 现有的结直肠癌的筛查手段均具有其局限性，高效的诊断方法是研究的重点。随着肿瘤研究的进展，DNA表观修饰在肿瘤发生发展中的作用逐渐受到关注。5-methylcytosine (5-mC)是DNA胞嘧啶环中第五碳通过脱氧核苷酸甲基转移酶甲基化所形成，在哺乳动物胚胎发育以及疾病发生发展中起着重要的作用。5-mC经过TET酶的氧化作用形成5-羟甲基胞嘧啶(5-hydroxymethylcytosine,5-hmC)。研究显示羟甲基化在正常基因组中表达量相对稳定，表明羟甲基化有独特的表观修饰作用，不仅仅是5-mC代谢的中间产物。5-hmC与包括癌症在内的多种疾病相关，在多种肿瘤组织中低水平的5-hmC得到证实。5-hmC水平的高低与肿瘤的类型及临床分期也有着密切的关系，如肺癌中，随着病理分期发

展，5-hmC的水平进行性下降。然而，5-hmC在肠癌中的表达模式如何改变尚缺乏研究。肿瘤的液态活检因其无创、便捷等特点迅速发展，通过外周血的检测实现对肿瘤的诊断及预测成为研究热点。既往对cfDNA的研究中存在cfDNA含量低、检测困难等问题。上海易毕恩公司低输入量cfDNA的5hmC测序技术(Nano-hmC-Seal)，克服了以往DNA表观修饰检出率低的缺陷。然而微量5-hmC全基因组测序技术是否能实现结直肠癌患者早期诊断尚缺乏研究。因此本发明，针对结直肠癌患者早期诊断设计一种基于液态活检5-hmC修饰多标志物模型用于结直肠癌的早期诊断模型。

[0005] 中国专利文献201710662851.6公开了一种用于检测胰腺癌的基因标志物、试剂盒及胰腺癌检测方法，基因标志物包括一个或两个以上以下基因：麦芽糖酶、C型外源凝集素家族4成员C、分选蛋白7、间质同源框2、FAT非典型钙粘蛋白1、黄素包含单氧酶3、囊性纤维化跨膜传导调节蛋白、磷脂磷酸酶相关蛋白3、α白蛋白和胶原蛋白V型α2链，通过高通量测序检测胰腺癌基因标志物中5-hmC的含量，从而判定胰腺癌是否存在。中国专利文献201710662852.0公开了用于检测肝肿瘤良恶性的基因标志物、试剂盒及检测方法，基因标志物包括一个或两个以上以下基因：FAT非典型钙粘蛋白1、雌激素相关受体γ、性别决定基因y染色体区域盒家族成员9、纤毛状络合物子单元1、1号染色体开放阅读框125抗体、络丝蛋白、转铁蛋白、TNF受体超家族成员、胰腺和十二指肠同圆框1、α-酸性糖蛋白 2，通过高通量测序检测肝肿瘤基因标志物中5-hmC的含量，从而判定肝肿瘤良恶性。现有技术中，关于本发明基于表观修饰的5hmC多分子标志物及结直肠癌早期诊断模型，目前还未见报道。

发明内容

[0006] 本发明基于5-hmC修饰特征性变化检测的结直肠癌早期诊断模型是通过对外周血中游离DNA的5-hmC表观修饰进行检测，通过对比结直肠癌患者与健康人5-hmC修饰特征性变化差异性，建立早期诊断模型，可广泛用于结直肠癌患者的早期筛查、诊断等临床工作中。与现有传统结直肠癌早期筛查手段相比，本发明的检测方法仅需采集患者外周静脉血，采集过程与常规血液检查无明显差别，检测便捷、安全，无需患者肠道准备过程，可明显提高患者的依从性；此外，本方法尚有稳定性强、灵敏度和特异度高等优点，是一种较为理想的结直肠癌筛查手段。

[0007] 本发明的第一个目的是，针对现有技术中的不足，提供一种基于表观修饰的5hmC多分子标志物结直肠癌早期诊断模型。

[0008] 本发明的第二个目的是，提供如上所述诊断模型在制备结直肠癌早期诊断试剂盒中的应用。

[0009] 本发明的第三个目的是，提供一种基于表观修饰的5hmC多分子标志物结直肠癌早期诊断试剂盒。

[0010] 为实现上述第一个目的，本发明采取的技术方案是：

[0011] 一种基于表观修饰的5hmC多分子标志物结直肠癌早期诊断模型，包括检测受试者外周血浆中基因标志物5-hmC表达量的试剂，所述基因标志物包括：GBX2FAM84AFAM25B LCE1F FBXL7 DBX1 KRTAP27-1 AL353791.1 CEBPD LTB4R2 RP4-583P15.14 OR5B2 RPRM RNASE4 INSL5 AURKC IL36AAC017081.1 SPA17 NBPF12 FABP1 CST8。

[0012] 在上述基于表观修饰的5hmC多分子标志物结直肠癌早期诊断模型中，作为一个优

选技术方案，所述模型中还包括正常人外周血样本。

[0013] 在上述基于表观修饰的5hmC多分子标志物结直肠癌早期诊断模型中，作为一个优选技术方案，所述模型中还包括记载有如下模型和评判方法的载体：

[0014] Logistic回归模型

[0015] $$P=\frac{e^{(b_0+b_1x_1+b_2x_2+\cdots+b_kx_k)}}{1+e^{(b_0+b_1x_1+b_2x_2+\cdots+b_kx_k)}}$$

[0016] 通过评分值判定样本来源是否为结直肠患者，当评分值大于0.5，则该受试者患有结直肠癌，当评分值小于或等于0.5，则该受试者为正常；

[0017] 其中$x_1,x_2,x_3...x_k$分别为以上基因标志物5-hmC表达量标准化结果；b0为建模过程输出的常数项；$b_1,b_2,b_3...b_k$为对应基因标志物5-hmC表达量的系数，目前k＝22。

[0018] 为实现上述第二个目的，本发明采取的技术方案是：

[0019] 如上任一所述诊断模型在制备结直肠癌早期诊断试剂盒中的应用。

[0020] 为实现上述第三个目的，本发明采取的技术方案是：

[0021] 一种基于表观修饰的5hmC多分子标志物结直肠癌早期诊断试剂盒，所述试剂盒包括检测受试者外周血浆中基因标志物5-hmC表达量的试剂，所述基因标志物包括：GBX2 FAM84A FAM25B LCE1F FBXL7 DBX1 KRTAP27-1 AL353791.1 CEBPD LTB4R2 RP4-583P15.14 OR5B2 RPRM RNASE4 INSL5 AURKC IL36A AC017081.1 SPA17 NBPF12 FABP1 CST8。

[0022] 在上述基于表观修饰的5hmC多分子标志物结直肠癌早期诊断试剂盒中，作为一个优选技术方案，所述的试剂盒还包括记载有如下模型和评判方法的载体：

[0023] Logistic回归模型

[0024] $$P=\frac{e^{(b_0+b_1x_1+b_2x_2+\cdots+b_kx_k)}}{1+e^{(b_0+b_1x_1+b_2x_2+\cdots+b_kx_k)}}$$

[0025] 通过评分值判定样本来源是否为结直肠患者，当评分值大于0.5，则该受试者患有结直肠癌，当评分值小于或等于0.5，则该受试者为正常；

[0026] 其中$x_1,x_2,x_3...x_k$分别为以上基因标志物5-hmC表达量标准化结果；b_0为建模过程输出的常数项；$b_1,b_2,b_3...b_k$为对应基因标志物5-hmC表达量的系数，目前k＝22。

[0027] 在上述基于表观修饰的5hmC多分子标志物结直肠癌早期诊断试剂盒中，作为一个优选技术方案，其中$b_0,b_1,b_2,b_3...b_{22}$的取值如下：

[0028]

标志物	B 系数
GBX2	1.31894902
FAM84A	1.21974841
FAM25B	1.15502531
LCE1F	1.10125199

[0029]

FBXL7	1.08682839
DBX1	1.01982759
KRTAP27-1	0.52487632
AL353791.1	0.40010969
CEBPD	0.38147623
LTB4R2	0.34240677
RP4-583P15.14	0.24499059
OR5B2	0.08550743
RPRM	0.02997599
RNASE4	-0.0157149
INSL5	-0.0274637
AURKC	-0.1550417
IL36A	-0.1906636
AC017081.1	-0.4367674
SPA17	-0.5486784
NBPF12	-0.9815598
FABP1	-1.0037715
CST8	-1.035181
常数项	-23.890494

。

[0030] 在上述基于表观修饰的5hmC多分子标志物结直肠癌早期诊断试剂盒中，作为一个优选技术方案，所述标准化是指用受试者样品除以正常人样品中同一基因标志物的5-hmC表达量，得到的比值即受试者样品该基因标志物5-hmC表达量标准化结果。

[0031] 在上述基于表观修饰的5hmC多分子标志物结直肠癌早期诊断试剂盒中，作为一个优选技术方案，所述试剂盒还包括以下试剂：正常人外周血样本。

[0032] 本发明还提供如上所述基因标志物在结直肠早期癌诊断中的应用，所述诊断中通过高通测序检测受试者外周血浆中上述基因标志物的5-hmC的表达量，从而判定结直肠癌是否存在。

[0033] 本发明还提供一种结直肠癌早期诊断方法，包括以下步骤：

[0034] 1)测定健康人样品和受试者样品中如上所述基因标志物的5-hmC的表达量；

[0035] 2)用健康人样品中所述基因标志物的5-hmC表达量作为参照，将该受试者样品中对应的基因标志物的5-hmC表达量标准化；

[0036] 3)对步骤2)中经标准化的基因标志物的5-hmC表达量进行数学关联，并获得评分P；

[0037] 4)根据评分P的数值大小得到受试者是否患有结直肠癌的诊断结果。

[0038] 优选地，所述样品是来自于健康人或受试者外周血浆中的游离DNA。

[0039] 优选地，步骤2)中所述标准化是指将获得的测序结果进行初步质控评估，清除低质量测序位点后，将达到测序质量标准的读段利用Bowtie2工具与人类标准基因组参考序列进行比较。然后利用featureCounts和HtSeq-Count工具来统计读段数量以确定各基因标志物的5-hmC含量。

[0040] 根据本发明，在测定各基因标志物上5-hmC含量之后，用正常样品中所述基因标志

物的5-hmC含量作为参照，将受试者样品中对应的基因标志物的5-hmC含量标准化。

[0041] 在本发明中，基因标志物的5-hmC含量数据标准化后，对各基因标志物的标准化5-hmC 含量进行数学关联以获得评分，从而根据所述评分获得诊断结果。本发明中，“数学关联”是指将来自生物样品的基因标志物的5-hmC含量与结直肠癌诊断结果相关联的任何计算方法或机器学习方法。优选的，所述“数学关联”是指逻辑回归数学关联。

[0042] 在本发明中，作为一个优选技术方案，对各基因标志物的标准化5-hmC含量进行数学关联并获得评分的具体步骤如下：将各基因标志物的标准化5-hmC表达量x乘以加权系数b，获得该基因标志物的预测因子，将各基因标志物的预测因子相加，获得总预测因子，将总预测值经过Logistic转换获得评分P；若P>0.5，则该受试者样品患有结直肠癌，若P≤0.5，则该受试者样品健康。

[0043] 本发明所述的基因标志物的5-hmC表达量可通过本领域技术人员已知的任何方法进行测定，例如包括但不限于：葡糖基化法、限制性内切酶法、化学标记法、高通量测序方法、单分子实时测序法(SMRT)、氧化重亚硫酸盐测序法(OxBS-Seq)等。在本发明的一个技术方案中，优选采用化学标记法测定基因标志物的5-hmC表达量。

[0044] 其中，葡糖基化法的原理是采用T4噬菌体β-葡萄糖转移酶(β-GT)，在葡萄糖供体底物尿核苷二磷酸葡萄糖存在下，将葡萄糖转移至羟基位置，从而生成β-葡萄糖基-5-羟甲基胞嘧啶。同时可采用同位素标记底物进行定量。化学标记法的原理是：将酶反应底物上的葡萄糖进行化学修饰转变成UDP-6-N3-glucose，将6-N3-glucos转移到羟甲基位置，生成N3-5ghmC。随后，通过点击化学反应在每个5-hmC上添加一分子生物素，结合下一代高通量DNA测序技术或单分子测序技术，可分析5-hmC在基因组DNA中的分布情况。

[0045] 本发明的一个技术方案中，利用化学标记法结合高通量测序来测定本发明的基因标志物的5-hmC含量。在该具体的实施方案中，测定本发明的基因标志物的5-hmC含量的方法包括以下步骤：将来自患者和正常人的样品的DNA片段化；将所述片段化的DNA末端修复并末端补齐；将末端补齐的DNA与测序接头连接，获得连接产物；通过标记反应对连接产物中的5-羟甲基胞嘧啶进行标记；富集含有5-羟甲基胞嘧啶标记的DNA片段，获得富集产物；对富集产物进行PCR扩增，获得测序文库；对测序文库进行高通量测序，获得测序结果；根据测序结果确定5-羟甲基胞嘧啶在基因上的含量。其中，标记反应包括：i)利用糖基转移酶将带有修饰基团的糖共价连接到5-羟甲基胞嘧啶的羟甲基上，和ii)将直接或间接连有生物素的点击化学底物与带有修饰基团的5-羟甲基胞嘧啶反应。在该方案中，所述糖基转移酶包括但不限于：T4噬菌体β-葡糖基转移酶(β-GT)、T4噬菌体α-葡糖基转移酶及其具有相同或相似活性的衍生物、类似物、或重组酶；所述带有修饰基团的糖包括但不限于：带有叠氮修饰的糖类(例如6-N3-葡萄糖)或带有其他化学修饰(例如羰基、巯基、羟基、羧基、碳-碳双键、碳-碳三键、二硫键、胺基、酰胺基、双烯等)的糖类，其中优选带有叠氮修饰的糖类；所述用于间接连接生物素和点击化学底物的化学基团包括但不限于：羰基、巯基、羟基、羧基、碳-碳双键、碳-碳三键、二硫键、胺基、酰胺基、双烯。在该实施方案中，优选通过固相材料来富集含有5-hmC标记的DNA片段。具体地，可以通过固相亲和反应或其他特异性结合反应将含有5-羟甲基胞嘧啶标记的DNA片段结合在固相材料上，然后通过多次洗涤去除未结合的DNA片段。固相材料包括但不限于带有表面修饰的硅片或其他芯片，例如人工高分子小球(优选直径为1nm-100um)、磁性小球(优选直径为 1nm-100um)、琼脂糖小球等(优选直径为

1nm-100um)。固相富集中所用的洗涤液是本领域技术人员熟知的缓冲液，包括但不限于：含有Tris-HCl、MOPS、HEPES(pH＝6.0-10.0，浓度在1mM到1M之间)、NaCl(0-2M)或表面活性剂如Tween20(0.01％-5％)的缓冲液。在该实施方案中，优选直接在固相上进行PCR扩增从而制备测序文库。本领域已知的各种二代测序平台及其相关的试剂可用于本发明。

[0046] 在本发明中，测定基因标志物的5-hmC含量是指测定该基因标志物全长上的5-hmC含量或测定该基因标志物上某一片段的5-hmC含量或其组合。

[0047] 在本发明中，用于测定基因标志物的5-hmC含量的试剂是本领域技术人员已知的，例如T4噬菌体β-葡萄糖转移酶和同位素标记(对于葡糖基化法)、限制性内切酶(对于限制性内切酶法)、糖基转移酶和生物素(对于化学标记法)、PCR和测序所用试剂等。

[0048] 本发明通过5-hmC高通量测序技术，对结直肠癌患者与健康人外周血浆DNA全基因组5-hmC表达量的检查，对比两组样本中全基因组基因位点5-hmC表达差异，应用Deseq 方法寻找表达差异最显著的部分基因位点作为该诊断模型的构建标志物。进而构建结直肠癌的早期诊断模型。

[0049] 本发明优点在于：

[0050] 1、本发明基于表观修饰的5hmC多分子标志物结直肠癌早期诊断模型经验证，检测的灵敏度、特异度均较高，内部验证的AUC曲线下面积为0.9488(灵敏度和特异度取值分别89.4％和93.4％)；外部验证的AUC曲线下面积为0.9501(灵敏度和特异度取值分别85.1％和90.4％)。与目前市场上的粪隐血检测试剂盒、spetin9试剂盒相比，灵敏度和特异度分别提高约20％、8％。

[0051] 2、本发明的诊断方法具有检测无创、便捷等特点，仅一次采血10ml，即可完成全基因组的5hmc检测，并通过早诊模型实施一站式实现结直肠癌的诊断和报告。

[0052] 3、本发明采用的5-hmC检测方法为成熟方法，检测结果可重复性高，检测成本低，仅为septin9试剂盒的66％。

附图说明

[0053] 附图1为低输入量DNA的5-hmC测序技术检测流程示意图。

[0054] 附图2为实施例的检测结果。左图为结直肠癌患者与健康人全基因组位点5-hmC表达差异性PCI降维区分，右图为选取前80个表达差异最显著的基因(即建模过程所涉及的基因)做两组的PCI降维区分，图中红色点代表一例健康人；蓝色点代表一例结直肠癌患者。

[0055] 附图3为表达差异最显著的前80个基因位点5-hmC表达情况火山图。图中右侧77个点代表表达量上调；图中左侧3个点代表表达下调。

[0056] 附图4为诊断模型内部验证结果。左图中每点代表一例样本，纵坐标为评分值，横坐标中0-100代表结直肠癌患者，101-200代表健康人，100例结直肠癌患者中92例诊断为结直肠癌，8例判断错误，100例健康人中90例判断为非结直肠癌，10例判断错误。右图为内部验证的ROC曲线，AUC取值0.9488，灵敏度和特异度分别为89.4％和93.4％。

[0057] 附图5为诊断模型外部验证结果。左图中每点代表一例样本，纵坐标为评分值，横坐标中0-47代表结直肠癌患者，48-99代表健康人，47例结直肠癌患者中42例诊断为结直肠癌，5例判断错误，52例健康人中44例判断为非结直肠癌，8例判断错误。右图为外部验证的ROC曲线，AUC取值0.9501，灵敏度和特异度分别为85.1％和90.4％。

具体实施方式

[0058] 下面结合具体实施方式，进一步阐述本发明。应理解，这些实施例仅用于说明本发明而不用于限制本发明的范围；此外应理解，在阅读了本发明记载的内容之后，本领域技术人员可以对本发明作各种改动或修改，这些等价形式同样落于本申请所附权利要求书所限定的范围。

[0059] 实施例1结直肠癌早期诊断模型的构建

[0060] 一、基因标志物的筛选

[0061] 本发明入组100例I-III期结直肠癌患者与100例肠镜检查正常的健康人。在征得知情同意的情况下采集外周静脉血10ml用于检测。检测过程与上海易毕恩基因科技有限公司合作，采用低输入量DNA的5-hmC测序(Nano-hmC-Seal)技术，检测过程如图1所示。检测得到全基因组20000余基因位点的5-hmC表达量。如图2(左)所示，两组全基因组5-hmC 表达量PCI降维区分无明显差别，随后，我们采用Deseq方法对比两组各基因位点5-hmC 表达差异，选取差异最大的前80个基因位点($p<4.0\times10^{-8}$)用于模型建立，采用所选取的 80个基因位点再次进行两组的PCI降维区分，可见两组有明显差别，如图2(右)所示。所选取的80个基因位点中77个是5-hmC高表达，3个5-hmC低表达，如图3所示。随后，采用该80个基因位点，利用逻辑回归和弹性网络正则化的方法构建诊断模型，获得各基因标志物的加权系数，建模过程中实际输出1个常数项和22个有意义的基因项(各位点权重见下文)。本诊断模型可针对每个样本的5-hmC表达变化进行计算评分，通过评分值判断其是否为结直肠癌患者。

[0062] 建模过程中所输入的80个基因位点，即上文所述基因位点如下所示：OR5J2 TGIF2 SATB1 HIST1H2BO OR5D14 OR5AK2 H1FX C21orf59 SPN LCE1F OR5P2 OR8J1 OR5B2 LTB4R2 FAM84A CEBPD HIST1H1E RPL5 FERD3L CEBPB OR14A2 KRT18 AURKC RP11-1212A22.4 OR3A3 OR4F6 CCT8L2 GOLGA6L22 REG1B DHRS7 REG1A NPIPB9 TAS2R39 OR14A16 OR7C2 SPA17 HNRNPCL2 OR52E2 NBPF11 AC017081.1 RUNX1T1 FBXL7 TLL1 SULF1 PTPN14 PTPRB RHOJ NOVA2 HIC1 RBMS3 LDB2 ADGRL4 PDE10A LRRC3B LIFR PLCXD3 NRN1 CDH11 ADAMTSL1 SNTB2 AM155A TMC7 GBX2 EPB41L4A PCDH17 CCDC177 CCDC85A TNFRSF11B GULP1 SNTB1 HIF3A TUSC3 RNF43 SNTG2 NPNT GJA1 DBX1 TRIML1 CNTNAP3 GNA14。

[0063] 二、诊断模型的建立

[0064] 采用逻辑回归模型的方法构建函数模型，对每例样本的检测结果计算模型得分(得分P 取值范围为[0,1])，通过得分判定样本来源为结直肠患者(得分≥0.5)或健康人(得分<0.5)。

[0065] ①检测结果标准化处理与参数选定

[0066] 将每例样本中各基因位点5-hmC表达量检测结果进行标准化处理，以便后续计算。具体如下：将获得的测序结果进行初步质控评估，清除低质量测序位点后，将达到测序质量标准的读段利用Bowtie2工具与人类标准基因组参考序列进行比较。然后利用featureCounts和HtSeq-Count工具来统计读段数量以确定各基因标志物的5-hmC含量。对比训练集中两组样本检测结果，选定以上80位差异最显著的基因位点用于诊断模型构建。

[0067] ②模型构建与计算机学习

[0068] 利用高通量测序结果(以上80个基因位点)，将可能影响5-hmC含量的因素作为共

变量,通过逻辑回归和弹性网络正则化获得各基因标志物的加权系数。模型如下：

[0069] Logistic回归模型

[0070] $$P=\frac{e^{(b_0+b_1x_1+b_2x_2+\cdots+b_kx_k)}}{1+e^{(b_0+b_1x_1+b_2x_2+\cdots+b_kx_k)}}$$

[0071] 将训练集中100例结直肠癌患者与100例健康人检测结果输入。建模过程中实际输出1 个常数项和22个有意义的基因项。

[0072] 有意义的基因位点如下：

[0073] GBX2 FAM84A FAM25B LCE1F FBXL7 DBX1 KRTAP27-1 AL353791.1 CEBPD LTB4R2 RP4-583P15.14 OR5B2 RPRM RNASE4 INSL5 AURKC IL36A AC017081.1 SPA17 NBPF12 FABP1 CST8

[0074] 式中b_0,b_1,b_2...b_k(k＝22)值如下表所示。

[0075]

标志物	B 系数
GBX2	1.31894902
FAM84A	1.21974841
FAM25B	1.15502531
LCE1F	1.10125199
FBXL7	1.08682839
DBX1	1.01982759
KRTAP27-1	0.52487632

[0076]

AL353791.1	0.40010969
CEBPD	0.38147623
LTB4R2	0.34240677
RP4-583P15.14	0.24499059
OR5B2	0.08550743
RPRM	0.02997599
RNASE4	-0.0157149
INSL5	-0.0274637
AURKC	-0.1550417
IL36A	-0.1906636
AC017081.1	-0.4367674
SPA17	-0.5486784
NBPF12	-0.9815598
FABP1	-1.0037715
CST8	-1.035181
常数项	-23.890494

[0077] 实施例2结直肠癌早期诊断模型的验证

[0078] 对构建的诊断模型进行内部验证。利用实施例1构建的模型对200例样本进行评分， cutoff值取0.5,结果大于等于0.5诊断为结直肠癌,反之认为是健康人,判定结果如图

4(左) 所示。100例结直肠癌患者中92例诊断为结直肠癌,8例判断错误;100例健康人中90例判断为非结直肠癌,10例判断错误,内部验证的灵敏度和特异度分别为92%和90%。内部验证的AUC曲线下面积为0.9488,灵敏度和特异度取值分别89.4%和93.4%,见图4(右)。

[0079] 对构建的诊断模型进行外部验证。采用同样的检测方法对另外47例结直肠癌患者及52 例肠镜检查无异常的健康人外周静脉血(10ml)进行检测。将检测结果标准化处理,取上述22位基因位点检测结果,利用实施例1构建的结直肠癌诊断模型每例样本的结果进行评分。47例结直肠癌患者样本评分如下:0.76、0.76、0.68、0.94、0.76、0.90、0.51、0.67、0.49、0.91、0.70、0.66、0.90、0.83、0.90、0.70、0.71、0.76、0.68、0.59、0.39、0.87、0.68、0.60、0.80、0.34、0.53、0.66、0.64、0.66、0.89、0.59、0.61、0.90、0.77、0.47、0.59、0.28、0.72、0.79、0.73、0.85、0.77、0.71、0.90、0.72、0.65。52例健康人样本评分结果如下:0.12、0.21、0.09、0.30、0.26、0.14、0.33、0.16、0.10、0.71、0.60、0.19、0.18、0.31、0.46、0.23、0.19、0.19、0.07、0.15、0.22、0.26、0.17、0.44、0.25、0.14、0.37、0.14、0.51、0.25、0.71、0.63、0.33、0.34、0.21、0.19、0.30、0.29、0.35、0.08、0.11、0.65、0.54、0.22、0.24、0.18、0.17、0.36、0.12、0.56、0.15、0.35。如图5(左)所示。47例结直肠癌患者中42 例评分≥0.50,诊断为结直肠癌,5例判断错误;52例健康人中44例评分<0.50,判断为非结直肠癌,8例判断错误,检测的灵敏度和特异度分别为89.3%和84.6%。外部验证的AUC 曲线下面积为0.9501,灵敏度和特异度取值分别85.1%和90.4%。

[0080] 本发明以成功构建了结直肠癌诊断模型,将本发明应用于后续样本的检测过程如下:

[0081] ①采集外周静脉血10ml;

[0082] ②采用5-hmC高通量测序技术(Nano-hmC-Seal)进行检测;

[0083] ③将标准化处理后的检测结果(以上22位基因位点)输入本模型,得出评分;

[0084] ④基于本模型的评分结果对每例样本进行诊断判定。

[0085] 以上所述仅是本发明的优选实施方式,应当指出,对于本技术领域的普通技术人员,在不脱离本发明方法的前提下,还可以做出若干改进和补充,这些改进和补充也应视为本发明的保护范围。

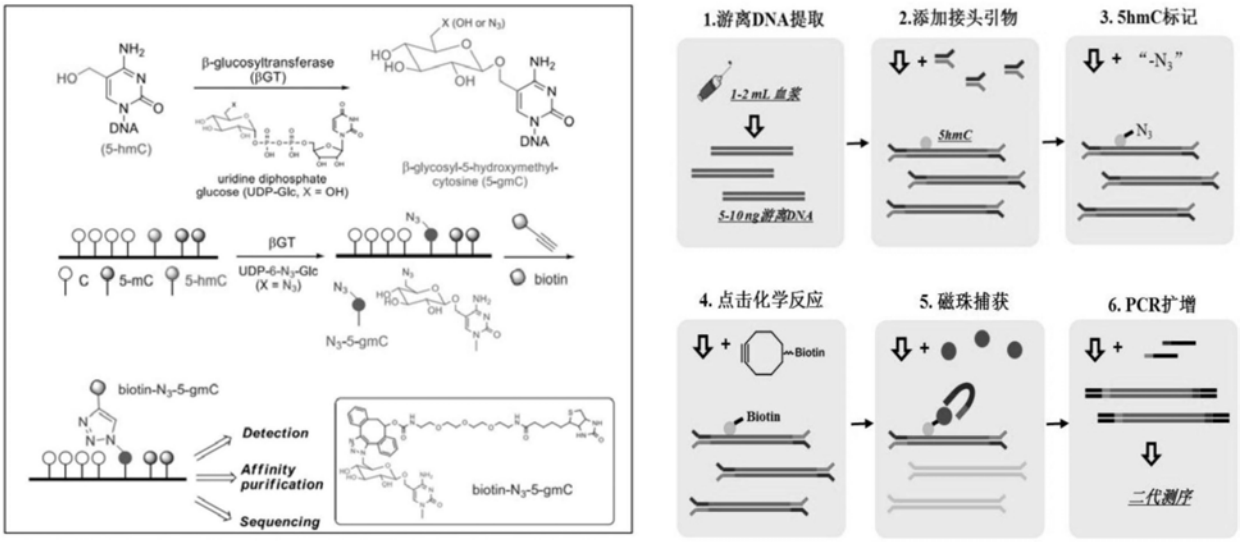

β-glucosyltransferase
(βGT)
uridine diphosphate
glucose (UDP-Glc, X = OH)
β-glycosyl-5-hydroxymethyl-
cytosine (5-gmC)
(5-hmC)
βGT
UDP-6-N3-Glc
(X = N3)
C
5-mC
5-hmC
biotin
N3-5-gmC
biotin-N3-5-gmC
Detection
Affinity
purification
Sequencing
1.游离DNA提取
2.添加接头引物
3. 5hmC标记
1-2 mL 血浆
5-10 ng游离DNA
5hmC
4. 点击化学反应
5. 磁珠捕获
6. PCR扩增
Biotin
二代测序

图1

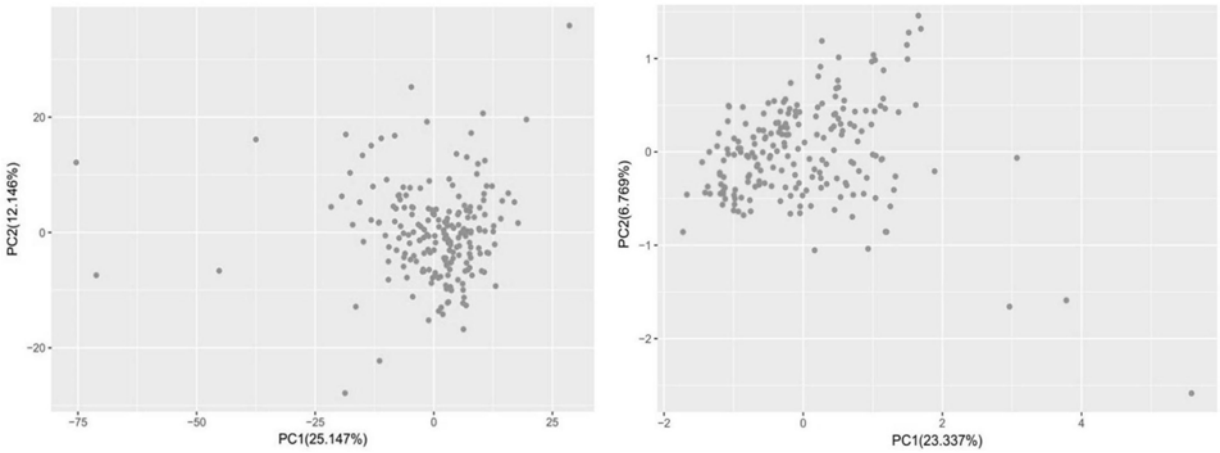

PC2(12.146%)
PC1(25.147%)
PC2(6.769%)
PC1(23.337%)

图2

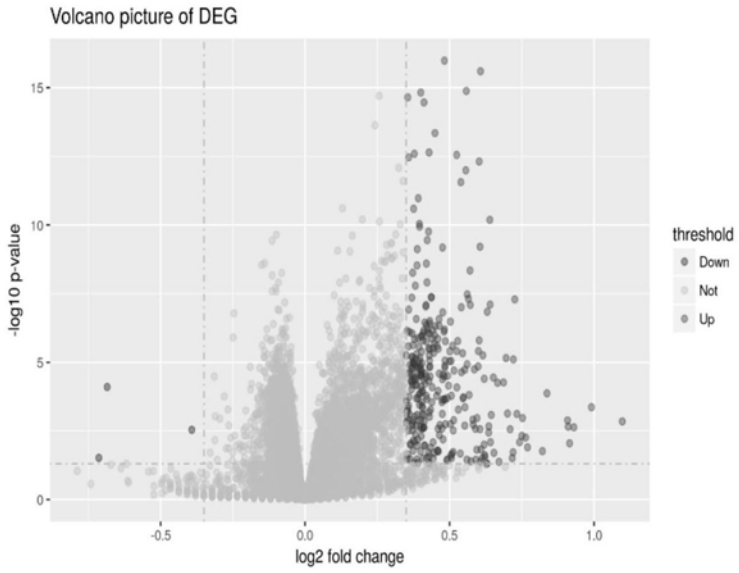

Volcano picture of DEG
-log10 p-value
log2 fold change
threshold
Down
Not
Up

图3

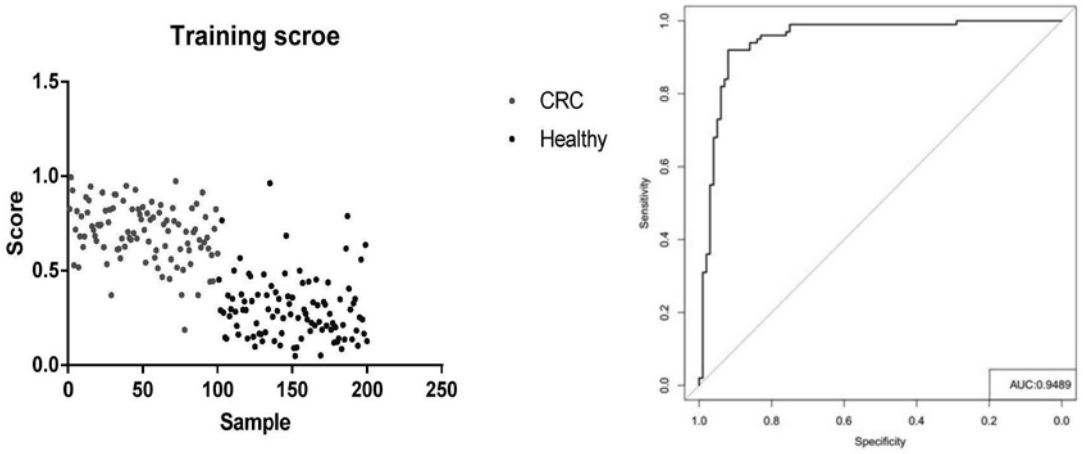

Training scroe
Score
Sample
CRC
Healthy
Sensitivity
Specificity
AUC:0.9489

图4

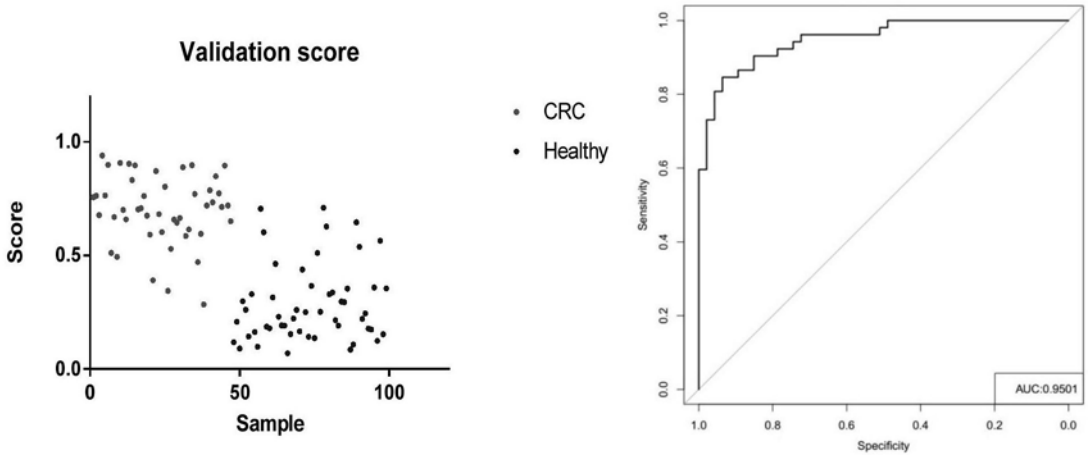

Validation score
Score
1.0
0.5
0.0
0
50
100
Sample
CRC
Healthy
Sensitivity
Specificity
AUC:0.9501

图5