(19)国家知识产权局

(12)发明专利

(10)授权公告号 CN 115558716 B
(45)授权公告日 2023.11.03

(21)申请号 202211203394.1
(22)申请日 2022.09.29
(65)同一申请的已公布的文献号
申请公布号 CN 115558716 A
(43)申请公布日 2023.01.03
(73)专利权人 南京医科大学
地址 210000 江苏省南京市汉中路140号
(72)发明人 汪强虎 吴玲祥 吴维 张若寒
(74)专利代理机构 杭州信与义专利代理有限公司 33450
专利代理师 万景旺
(51)Int.Cl.
C12Q 1/6886(2018.01)
C12Q 1/6869(2018.01)
G16H 50/30(2018.01)
G16H 50/70(2018.01)

(56)对比文件
CN 112410422 A,2021.02.26
CN 112805563 A,2021.05.14
CN 113195741 A,2021.07.30
CN 113421608 A,2021.09.21
CN 113728116 A,2021.11.30
CN 114974430 A,2022.08.30
US 2021043275 A1,2021.02.11
US 2021172019 A1,2021.06.10
WO 2022040163 A1,2022.02.24

审查员 宋顺意

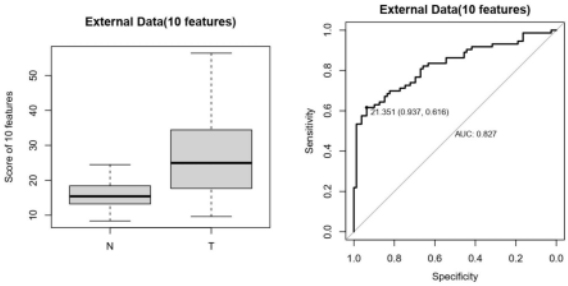
权利要求书1页 说明书14页 附图7页

(54)发明名称

一种用于预测癌症的cfDNA片段特征组合、系统及应用

(57)摘要

本发明公开了一种用于预测癌症的cfDNA片段特征组合、系统及应用，属于癌症基因组学技术领域。所述cfDNA片段特征组合包括第一cfDNA片段特征子组合和/或第二cfDNA片段特征子组合，所述第一cfDNA片段特征子组合包括落在60bp～200bp之间并且片段数量比例在群体癌症样本中增加的cfDNA片段特征，所述第二cfDNA片段特征子组合包括落在300～400bp之间并且片段数量比例在群体癌症样本中减少的cfDNA片段特征。利用本发明的cfDNA片段特征组合和系统进行癌症预测，既降低了基于cfDNA片段分析预测癌症的方法对于上游实验端的要求和依赖，又显著拓宽了其他组学测序数据的可解读性和利用率，因此，极大的降低了基于cfDNA诊断肿瘤的实验成本，同时提高了基于cfDNA预测癌症的准确性。



CN 115558716 B

1.一种cfDNA片段特征组合的检测试剂在制备用于预测受试者是否患有癌症的试剂盒中的应用,其特征在于,所述cfDNA片段特征组合包括第一cfDNA片段特征子组合和第二cfDNA片段特征子组合,所述第一cfDNA片段特征子组合由163-164bp、163-165bp、161-165bp、161-164bp和165-166bp组成,所述第二cfDNA片段特征子组合由339-340bp、341-342bp、343-344bp、337-339bp和340-342bp组成,所述癌症为结直肠癌、肝癌、胃癌、胰腺癌、食管癌或胶质母细胞瘤。

2.一种预测受试者是否患有癌症的系统,其特征在于,包括以下模块:

数据输入模块,用于输入受试者cfDNA片段长度和数量数据;

分布谱分析模块,与所述数据输入模块连接,用于获得权利要求1所述cfDNA片段特征组合中各cfDNA片段特征的片段数量比例;

癌症预测模块,与所述分布谱分析模块连接,用于根据所述cfDNA片段特征的片段数量比例判断受试者是否患有癌症,其中,所述癌症为结直肠癌、肝癌、胃癌、胰腺癌、食管癌或胶质母细胞瘤,

所述癌症预测模块基于以下方法预测受试者是否患有癌症:

利用下面公式获得判断值:

$$Score = \frac{\sum_{i=1}^{m} T_i}{\sum_{j=1}^{n} N_j}$$

其中,

Score为判断值,

m为所述第一cfDNA片段特征子组合中cfDNA片段特征的数量,n为所述第二cfDNA片段特征子组合中cfDNA片段特征的数量;

Ti为第一cfDNA片段特征子组合中第i个cfDNA片段特征的片段数量比例;

Nj为第一cfDNA片段特征子组合中第j个cfDNA片段特征的片段数量比例,

若Score大于预设阈值,则判断所述受试者患有癌症。

一种用于预测癌症的cfDNA片段特征组合、系统及应用

技术领域

[0001] 本发明属于癌症基因组学技术领域，具体地，涉及一种用于预测癌症的cfDNA片段特征组合、系统及应用。

背景技术

[0002] 血液中的游离DNA(cfDNA,Circulating free DNA or Cell free DNA)能够随着组织损伤、癌症和炎症反应等发生浓度变化，在疾病的早期诊断、预后、监测等方面具有重要潜在价值。近年来，cfDNA已被广泛用于癌症早筛等研究领域。研究表明，可以利用特定的cfDNA片段特征对肿瘤组织来源进行分类，cfDNA片段的长度也可以揭示组织起源或肿瘤来源。

[0003] 然而，目前大多数液体活检方法都专注于检测血液中的基因突变或染色体异常，且已有的片段组学方法多依赖于全基因组测序(WGS)的方法，无法充分开发利用其他组学测序数据信息。

发明内容

[0004] 为解决上述技术问题中的至少一个，本发明开发了一种可基于多种组学数据分析片段组学的系统，以识别cfDNA片段分布肿瘤标志物，进而鉴别样本是否为肿瘤样本。具体地，本发明采用的技术方案如下：

[0005] 本发明第一方面提供一种cfDNA片段特征组合，包括第一cfDNA片段特征子组合和/或第二cfDNA片段特征子组合，所述第一cfDNA片段特征子组合包括落在60bp～200bp之间并且片段数量比例在群体癌症样本中增加的cfDNA片段特征，所述第二cfDNA片段特征子组合包括落在300～400bp之间并且片段数量比例在群体癌症样本中减少的cfDNA片段特征，所述增加或减少是指相对于群体正常样本相应片段特征的片段数量比例的代表值而言。

[0006] 在本发明中，相关术语的定义如下：

[0007] 片段特征:是指将cfDNA片段按不同长度划分为不同的片段区间，每个片段区间内的所有cfDNA片段即为一个片段特征。例如片段特征为:61-65bp，包括片段长度为61bp、62bp、63bp、64bp和65bp的cfDNA片段。例如片段特征为:74-75bp，包括片段长度为74bp和75bp的cfDNA片段。

[0008] 片段数量比例:是指一个片段特征中的cfDNA片段数占总片段数的比例。

[0009] 在本发明中，所述cfDNA片段长度和数量数据是指利用测序方法得到的数据，所述测序选自包括WGS测序、WES测序、MeDIP和MBD-Seq的组中的任意一种。事实上，本领域技术人员可能使用任意测序的或非测序的方法，只要能够获得cfDNA片段的长度及数量即可。

[0010] 在本发明中，所述第一cfDNA片段特征子组合中的每个片段特征所包含的cfDNA片段相对都比较短，本发明的发明人意外地发现，落在60bp～200bp之间的片段特征可以用来识别癌症，并且具有较高的精准度。更加令人惊喜地，发明人发现落在130bp～175bp之间的

cfDNA片段特征具有更高的癌症识别精准度。

[0011] 进一步地，如果在落在130bp～175bp之间的cfDNA片段特征中，选择163-164bp、163-165bp、161-165bp、161-164bp、165-166bp、159-165bp、157-164bp、155-156bp、163-168bp、160-168bp、157-158bp、154-156bp、161-170bp、156-160bp、161-162bp、157-159bp、165-168bp、157-162bp、157-160bp、151-160bp、152-158bp、160-162bp、153-156bp、151-159bp、166-168bp、148-150bp、149-150bp、149-156bp、159-160bp、151-156bp、167-168bp、147-148bp、146-150bp、165-172bp、166-170bp、151-155bp、153-154bp、149-152bp、145-150bp、145-151bp、166-172bp、145-148bp、151-153bp、151-152bp、169-170bp、145-147bp、169-171bp、142-150bp、169-172bp和141-150bp中的至少一个，能够精准地预测受试者是否患有癌症或者是否具有患癌症的风险。

[0012] 发明人进一步发现，上述片段特征的选择也不是越多越好，当选择163-164bp、163-165bp、161-165bp、161-164bp和165-166bp作为标志物时，具有非常好的癌症识别效果。

[0013] 在本发明中，所述第二cfDNA片段特征子组合中的每个片段特征所包含的cfDNA片段相对都比较长，本发明的发明人意外地发现，落在300bp～400bp之间的片段特征可以用来识别肿瘤，并且具有较高的精准度。更加令人惊喜地，发明人发现落在330bp～360bp之间的cfDNA片段特征具有更高的肿瘤识别精准度。

[0014] 进一步地，如果在落在330bp～360bp之间的cfDNA片段特征中，选择339-340bp、341-342bp、343-344bp、337-339bp、340-342bp、337-340bp、341-344bp、336-340bp、341-345bp、337-342bp、337-338bp、343-345bp、341-347bp、340-348bp、334-340bp、341-348bp、345-346bp、343-348bp、341-350bp、345-348bp、346-348bp、333-340bp、347-348bp、346-350bp、335-336bp、331-340bp、334-336bp、349-350bp、349-351bp、331-339bp、349-352bp、333-336bp、348-354bp、351-352bp、349-354bp、349-356bp、352-354bp、351-355bp、333-334bp、331-336bp、349-357bp、353-354bp、331-335bp、353-356bp、351-360bp、355-356bp、355-357bp、331-333bp、355-360bp和357-358bp中的至少一个，能够精准地预测受试者是否患有癌症或者是否具有患癌症的风险。

[0015] 同样地，发明人进一步发现，上述片段特征的选择同样不是越多越好，当选择339-340bp、341-342bp、343-344bp、337-339bp和340-342bp作为标志物时，具有非常好的癌症识别效果。

[0016] 本发明的第二方面提供一种预测受试者是否患有癌症或者是否具有患癌症的风险的系统，包括以下模块：

[0017] 数据输入模块，用于输入受试者cfDNA片段长度和数量数据；

[0018] 分布谱分析模块，与所述数据输入模块连接，用于获得所述cfDNA片段特征组合中各cfDNA片段特征的片段数量比例；

[0019] 癌症预测模块，与所述分布谱分析模块连接，用于根据所述cfDNA片段特征的片段数量比例判断受试者是否患有癌症或者是否具有患癌症的风险。

[0020] 在本发明的一些实施方案中，所述第一cfDNA片段特征子组合中至少一个cfDNA片段特征的片段数量比例增加和/或所述第二cfDNA片段特征子组合中至少一个cfDNA片段特征的片段数量比例减少，则判断所述受试者患有癌症或者具有患癌症的风险。

[0021] 在本发明的另一些实施方案中，所述cfDNA片段特征组合包括第一cfDNA片段特征子组合和第二cfDNA片段特征子组合，所述癌症预测模块利用下面公式获得判断值：

[0022] $$Score=\frac{\sum_{i=1}^{m}T_i}{\sum_{j=1}^{n}N_j}$$

[0023] 其中，

[0024] Score为判断值，

[0025] m为所述第一cfDNA片段特征子组合中cfDNA片段特征的数量，n为所述第二cfDNA片段特征子组合中cfDNA片段特征的数量；

[0026] Ti为第一cfDNA片段特征子组合中第i个cfDNA片段特征的片段数量比例；

[0027] N_j为第一cfDNA片段特征子组合中第j个cfDNA片段特征的片段数量比例，

[0028] 若Score大于预设阈值，则判断所述受试者患有癌症或者具有患癌症的风险。

[0029] 在本发明的一些实施方案中，所述预测阈值是根据群体癌症样本Score值和/或群体正常样本Score值进行确定的。

[0030] 任选地，所述预测阈值是根据群体癌症样本Score值的代表值确定的。

[0031] 任选地，所述预测阈值是根据群体正常样本Score值的代表值确定的。

[0032] 任选地，所述预测阈值是根据群体癌症样本Score值相对于群体正常样本Score值的增加值的代表值确定的。这里的癌症样本和正常样本为配对样本，以使得增加值具有临床意义。

[0033] 在本发明的一些具体实施方案中，所述群体癌症样本是指10个以上癌症样本，例如10个、20个、50个、100个、200个、500个或更多。

[0034] 在本发明的一些具体实施方案中，所述代表值是指平均数、众数、中位数、1/4分位数和3/4分位数中的一种。

[0035] 在本发明中，所述癌症包括但不限于实体瘤和血癌，如纤维肉瘤、肌肉瘤、脂肪肉瘤、软骨肉瘤、成骨肉瘤、脊索瘤、血管肉瘤、内皮肉瘤、淋巴管肉瘤、淋巴管内皮肉瘤、滑膜瘤、间皮瘤、尤因瘤、平滑肌肉瘤、横纹肌肉瘤、结肠癌、胰腺癌、前列腺癌、鳞状细胞癌、基底细胞癌、腺癌、汗腺癌、皮脂腺癌、乳头状癌、乳头腺癌、囊腺癌、髓样癌、支气管癌、肝细胞瘤、胆管癌、绒毛膜癌、精原细胞瘤、胚胎癌、肾母细胞瘤、宫颈癌、睾丸瘤、肺癌、小细胞肺癌、上皮癌、胶质瘤、星形细胞瘤、髓母细胞瘤、颅咽管瘤、室管膜瘤、松果体瘤、成血管细胞瘤、听神经瘤、少突神经胶质瘤、脑膜瘤、黑素瘤、神经母细胞瘤、胶质母细胞瘤、视网膜母细胞瘤；白血病，如急性淋巴细胞性白血病和急性成髓细胞性白血病(成髓细胞、前髓细胞、髓单核细胞、单核细胞和红细胞白血病)；慢性白血病(慢性髓细胞(粒细胞)白血病和慢性淋巴细胞性白血病)；和真性红细胞增多、淋巴瘤(霍奇金病和非霍奇金病)、多发性骨髓瘤、瓦尔登斯特伦巨球蛋白血症和重链病。

[0036] 本发明第三方面提供本发明第一方面所述的cfDNA片段特征组合的检测试剂在制备用于预测受试者是否患有癌症或者是否具有患癌症的风险的试剂盒中的应用。

[0037] 在本发明的一些实施方案中，所述检测试剂包括捕获试剂和/或测序试剂。

[0038] 在本发明的一些实施方案中，所述试剂盒还包括cfDNA提取试剂。

[0039] 本发明的有益效果

[0040] 相对于现有技术,本发明具有以下有效效果:

[0041] 利用本发明的cfDNA片段特征组合和系统进行癌症预测,不仅可以利用选自包括WGS测序、WES测序、MeDIP和MBD-Seq的组中的任意一种测序方法的数据,也可以使用任意测序的或非测序的方法得到的数据,只要能够获得cfDNA片段的长度及数量即可。

[0042] 利用本发明的cfDNA片段特征组合和系统进行癌症预测,能够利用cfDNA片段综合特征分析,对于癌症的预测性能更优。

[0043] 利用本发明的cfDNA片段特征组合和系统进行癌症预测,既降低了基于cfDNA片段分析预测癌症的方法对于上游实验端的要求和依赖,又显著拓宽了其他组学测序数据的可解读性和利用率,因此,极大的降低了基于cfDNA诊断肿瘤的实验成本,同时提高了基于cfDNA预测癌症的准确性。

附图说明

[0044] 图1示出了利用10个cfDNA片段特征在训练集和验证集中进行肿瘤识别的结果。

[0045] 图2示出了利用20个cfDNA片段特征在训练集和验证集中进行肿瘤识别的结果。

[0046] 图3示出了利用30个cfDNA片段特征在训练集和验证集中进行肿瘤识别的结果。

[0047] 图4示出了利用40个cfDNA片段特征在训练集和验证集中进行肿瘤识别的结果。

[0048] 图5示出了利用50个cfDNA片段特征在训练集和验证集中进行肿瘤识别的结果。

[0049] 图6示出了利用60个cfDNA片段特征在训练集和验证集中进行肿瘤识别的结果。

[0050] 图7示出了利用10个cfDNA片段特征在外部测试集中进行肿瘤识别的结果。

具体实施方式

[0051] 除非另有说明、从上下文暗示或属于现有技术的惯例,否则本申请中所有的份数和百分比都基于重量,且所用的测试和表征方法都是与本申请的提交日期同步的。在适用的情况下,本申请中涉及的任何专利、专利申请或公开的内容全部结合于此作为参考,且其等价的同族专利也引入作为参考,特别这些文献所披露的关于本领域中的技术术语等的定义。如果现有技术中披露的具体术语的定义与本申请中提供的任何定义不一致,则以本申请中提供的术语定义为准。

[0052] 本申请中的数字范围是近似值,因此除非另有说明,否则其可包括范围以外的数值。数值范围包括以1个单位增加的从下限值到上限值的所有数值,条件是在任意较低值与任意较高值之间存在至少2个单位的间隔。这些仅仅是想要表达的内容的具体示例,并且所列举的最低值与最高值之间的数值的所有可能的组合都被认为清楚记载在本申请中。

[0053] 为了使本发明所解决的技术问题、技术方案及有益效果更加清楚明白,以下结合实施例,对本发明进行进一步详细说明。

[0054] 实施例

[0055] 以下例子在此用于示范本发明的优选实施方案。本领域内的技术人员会明白,下述例子中披露的技术代表发明人发现的可以用于实施本发明的技术,因此可以视为实施本发明的优选方案。但是本领域内的技术人员根据本说明书应该明白,这里所公开的特定实施例可以做很多修改,仍然能得到相同的或者类似的结果,而非背离本发明的精神或范围。

[0056] 除非另有定义,所有在此使用的技术和科学的术语,和本发明所属领域内的技术人员所通常理解的意思相同,在此公开引用及他们引用的材料都将以引用的方式被并入。

[0057] 那些本领域内的技术人员将意识到或者通过常规试验就能了解许多这里所描述的发明的特定实施方案的许多等同技术。这些等同将被包含在权利要求书中。

[0058] 下述实施例中未作具体说明的分子生物学实验方法,均按照《分子克隆实验指南》(第四版)(J.萨姆布鲁克、M.R.格林,2017)一书中所列的具体方法进行,或者按照试剂盒和产品说明书进行。其他实验方法,如无特殊说明,均为常规方法。下述实施例中所用的仪器设备,如无特殊说明,均为实验室常规仪器设备;下述实施例中所用的试验材料,如无特殊说明,均为自常规生化试剂商店购买得到的。

[0059] 实施例1 cfDNA片段分布肿瘤标志物的识别

[0060] 1.cfDNA测序

[0061] 为了获得cfDNA片段分布肿瘤标志物,发明人获得了417个肿瘤患者(183个结直肠癌、40个肝癌、92个胃癌、68个胰腺癌、9个食管癌和25个胶质母细胞瘤)和813个正常人的血液样本。提取cfDNA并采用甲基化DNA富集测序技术(MBD-Seq,Methylated DNA Binding Domain-Sequencing)进行测序。

[0062] 2.数据预处理

[0063] a)数据清洗:使用fastp-0.20.0软件去除建库过程中引入的接头序列以及低质量碱基片段(超过40%的碱基的质量值低于Q15和超过5个N的整条片段、基于滑窗裁剪片段末端平均质量<Q20的4个碱基)。

[0064] b)数据比对:使用bowtie2-2.3.4.2软件将fastq文件的碱基序列比对到人类参考基因组hg19(GRCH37)上生成bam文件,并根据基因组坐标对bam文件进行排序,使用picard MarkDuplicates-2.18.25-SNAPSHOT对排序后的bam进行去重,最后筛选配对reads均比对到参考基因组并且MAPQ>20的读段。

[0065] c)cfDNA筛选:为了删除MBD蛋白非特异捕获的cfDNA片段,将bam文件中不包含CG碱基对的片段过滤掉。进一步保留片段长度在(60,400]的cfDNA进行后续分析。

[0066] 3.cfDNA片段分布谱

[0067] 使用R包Rsamtools分析最终处理好的bam文件,计算出每条cfDNA的片段长度。然后,分别以步长2bp、3bp、4bp、5bp……10bp的长度,将cfDNA片段长度划分为不同的片段区间(如步长2bp,则划分的片段区间为61-62bp、63-64bp……、398-400bp;如步长3bp,则划分的片段区间为61-63bp、64-66bp……396-399bp;如步长10bp,则划分的片段区间为61-70bp、71-80bp……391-400bp),每个片段区间包括的全部cfDNA片段定义为片段特征,并计算每个片段特征中的cfDNA片段数占总片段数的比例,以生成cfDNA的片段分布谱。

[0068] 4.识别cfDNA片段肿瘤标志物

[0069] 在肿瘤和健康两组样本中,对每个cfDNA片段特征进行wilcox秩和检验并使用BH校正得到校正p值,进一步计算每个片段特征区分肿瘤和健康样本的ROC曲线下面积(AUC)值。认定校正p值<0.05且AUC>0.6的片段特征在肿瘤和健康样本中是差异分布的。

[0070] 将训练集中的肿瘤样本随机平均分成两份,在健康样本中随机生成与肿瘤样本数一致的两份样本,分别混合两份肿瘤样本和健康样本,然后按照每个片段特征依次对两份样本进行排序,计算两份样本中的片段特征区分肿瘤和健康样本的优势比OR值。将以上过

程重复100次，然后计算每个片段特征100次的平均OR值，并保留平均OR值>1.5的片段特征。

[0071] 由此得到100个片段特征，其中50个片段特征的片段数量占总片段数的比例在肿瘤样本中增加，50个片段特征的片段数量比例在肿瘤样本中减少，如表1所示：

[0072] 表1 100个片段特征

[0073]

片段特征	OR 值	片段特征	OR 值
163-164bp	2.612688	339-340bp	3.117721
163-165bp	2.600275	341-342bp	3.118119
161-165bp	2.554762	343-344bp	3.112204
161-164bp	2.54652	337-339bp	3.091199
165-166bp	2.502182	340-342bp	3.11615
159-165bp	2.493026	337-340bp	3.104653
157-164bp	2.449415	341-344bp	3.114477
155-156bp	2.379277	336-340bp	3.087465
163-168bp	2.439811	341-345bp	3.109406
160-168bp	2.444516	337-342bp	3.115018
157-158bp	2.372072	337-338bp	3.076744
154-156bp	2.275175	343-345bp	3.106828
161-170bp	2.378009	341-347bp	3.104026
156-160bp	2.366692	340-348bp	3.099929
161-162bp	2.424789	334-340bp	3.052213
157-159bp	2.354162	341-348bp	3.09608
165-168bp	2.372219	345-346bp	3.102963
157-162bp	2.364465	343-348bp	3.086484
157-160bp	2.338269	341-350bp	3.09244

[0074]

151-160bp	2.28437	345-348bp	3.069282
152-158bp	2.253274	346-348bp	3.045124
160-162bp	2.372103	333-340bp	3.03351
153-156bp	2.198956	347-348bp	3.030067
151-159bp	2.260956	346-350bp	3.0281
166-168bp	2.299412	335-336bp	3.002344
148-150bp	2.241652	331-340bp	2.995438
149-150bp	2.223731	334-336bp	2.977327
149-156bp	2.145107	349-350bp	2.995169
159-160bp	2.251132	349-351bp	2.99234
151-156bp	2.124366	331-339bp	2.976961
167-168bp	2.196728	349-352bp	2.99892
147-148bp	2.180979	333-336bp	2.954215
146-150bp	2.175948	348-354bp	2.977332
165-172bp	2.144438	351-352bp	2.98334
166-170bp	2.156023	349-354bp	2.967413
151-155bp	2.111366	349-356bp	2.946788
153-154bp	2.107802	352-354bp	2.937396
149-152bp	2.129839	351-355bp	2.93858
145-150bp	2.119438	333-334bp	2.913146
145-151bp	2.124849	331-336bp	2.911482
166-172bp	2.109298	349-357bp	2.93595
145-148bp	2.031472	353-354bp	2.904138
151-153bp	2.068739	331-335bp	2.871848
151-152bp	2.070665	353-356bp	2.88991
169-170bp	2.008504	351-360bp	2.876282
145-147bp	1.955188	355-356bp	2.875354
169-171bp	1.996834	355-357bp	2.868673
142-150bp	1.962088	331-333bp	2.815735
169-172bp	1.962926	355-360bp	2.859608
141-150bp	1.915405	357-358bp	2.851846

[0075] 由表1可知,在肿瘤样本中增加的片段特征中,大小集中在131-172bp,在肿瘤样本中减少的片段特征中,大小集中在331-360bp。

[0076] 实施例2不同片段特征判断肿瘤的效能

[0077] 利用上述50个在肿瘤样本中增加的片段特征和50个在肿瘤样本中减少的片段特征,在训练集中计算每个特征在肿瘤样本中相对于正常对照样本增加或减少的比例,据此标准判断样本属于肿瘤或正常,并在测试集中进行验证。其各自区分肿瘤样本和正常样本的效能如下表2和表3所示:

[0078] 表2 50个在肿瘤样本中增加的片段特征的判断结果

[0079]

片段特征 Features	训练集/Training set				测试集/Testing set			癌症样本增加比例（至少）
	灵敏度 Sensitivity	特异度 Specificity	AUC	Cutoff	灵敏度 Sensitivity	特异度 Specificity	AUC	
163-164bp	0.672	0.805	0.776	0.044	0.661	0.79	0.767	6%
163-165bp	0.689	0.788	0.775	0.07	0.653	0.802	0.767	7%
161-165bp	0.676	0.793	0.770	0.106	0.645	0.802	0.761	6%
161-164bp	0.659	0.802	0.768	0.08	0.686	0.753	0.76	6%
165-166bp	0.618	0.847	0.766	0.053	0.562	0.877	0.756	8%
159-165bp	0.655	0.795	0.765	0.137	0.686	0.753	0.758	6%
157-164bp	0.598	0.847	0.763	0.14	0.587	0.852	0.758	8%
155-156bp	0.564	0.868	0.763	0.022	0.603	0.802	0.746	8%
163-168bp	0.639	0.819	0.763	0.151	0.636	0.798	0.750	7%
160-168bp	0.625	0.823	0.762	0.203	0.628	0.819	0.752	7%
157-158bp	0.578	0.856	0.759	0.027	0.595	0.819	0.750	9%
154-156bp	0.574	0.856	0.757	0.033	0.62	0.761	0.739	11%
161-170bp	0.642	0.812	0.757	0.237	0.496	0.914	0.741	6%
156-160bp	0.598	0.830	0.756	0.070	0.570	0.823	0.747	7%
161-162bp	0.578	0.849	0.755	0.036	0.562	0.860	0.746	6%
157-159bp	0.598	0.825	0.755	0.042	0.579	0.819	0.747	7%
165-168bp	0.611	0.84	0.755	0.107	0.537	0.885	0.742	8%
157-162bp	0.578	0.846	0.754	0.095	0.628	0.782	0.748	7%
157-160bp	0.584	0.837	0.752	0.059	0.570	0.823	0.745	8%
151-160bp	0.662	0.77	0.751	0.119	0.603	0.765	0.735	6%
152-158bp	0.645	0.777	0.751	0.078	0.579	0.786	0.732	6%
160-162bp	0.578	0.849	0.751	0.053	0.537	0.864	0.742	7%
153-156bp	0.581	0.842	0.750	0.043	0.620	0.749	0.729	9%
151-159bp	0.639	0.779	0.750	0.103	0.562	0.807	0.732	7%
166-168bp	0.588	0.851	0.748	0.081	0.521	0.885	0.733	8%
148-150bp	0.618	0.800	0.747	0.023	0.653	0.733	0.727	8%
149-150bp	0.649	0.761	0.746	0.016	0.653	0.733	0.726	9%
149-156bp	0.655	0.756	0.743	0.076	0.645	0.724	0.720	6%
159-160bp	0.564	0.851	0.743	0.032	0.628	0.761	0.737	7%
151-156bp	0.591	0.816	0.742	0.062	0.636	0.728	0.719	8%
167-168bp	0.601	0.835	0.741	0.054	0.620	0.757	0.722	7%
147-148bp	0.608	0.804	0.740	0.014	0.653	0.720	0.720	10%
146-150bp	0.645	0.760	0.738	0.036	0.628	0.733	0.717	7%
165-172bp	0.601	0.842	0.738	0.204	0.620	0.770	0.718	7%
166-170bp	0.611	0.826	0.738	0.131	0.570	0.807	0.718	6%
151-155bp	0.645	0.760	0.736	0.050	0.636	0.716	0.711	7%
153-154bp	0.645	0.763	0.735	0.020	0.645	0.708	0.708	5%

[0080]

149-152bp	0.632	0.761	0.734	0.035	0.678	0.675	0.711	7%
145-150bp	0.611	0.784	0.732	0.043	0.661	0.671	0.709	8%
145-151bp	0.645	0.751	0.732	0.051	0.653	0.683	0.708	6%
166-172bp	0.581	0.861	0.731	0.178	0.612	0.778	0.709	7%
145-148bp	0.584	0.798	0.723	0.027	0.653	0.654	0.699	8%
151-153bp	0.649	0.732	0.722	0.029	0.653	0.658	0.695	6%
151-152bp	0.649	0.726	0.721	0.019	0.595	0.720	0.695	6%
169-170bp	0.581	0.833	0.715	0.051	0.537	0.831	0.690	6%
145-147bp	0.557	0.812	0.715	0.020	0.421	0.868	0.688	9%
169-171bp	0.598	0.816	0.713	0.074	0.521	0.844	0.688	5%
142-150bp	0.591	0.774	0.713	0.061	0.661	0.626	0.685	8%
169-172bp	0.581	0.830	0.711	0.097	0.603	0.765	0.685	6%
141-150bp	0.591	0.770	0.707	0.067	0.653	0.626	0.697	8%

[0081] 表3 50个在肿瘤样本中增加的片段特征的判断结果

[0082]

片段特征 Features	训练集/Training set				测试集/Testing set			癌症样本减少比例（至少）
	灵敏度 Sensitivity	特异度 Specificity	AUC	Cutoff	灵敏度 Sensitivity	特异度 Specificity	AUC	
339-340bp	0.679	0.851	0.825	0.003	0.711	0.835	0.820	16%
341-342bp	0.736	0.798	0.824	0.003	0.769	0.794	0.821	16%
343-344bp	0.747,	0.786	0.824	0.003	0.785	0.782	0.821	15%
337-339bp	0.682	0.846	0.824	0.004	0.760	0.774	0.818	26%
340-342bp	0.736	0.800	0.825	0.004	0.760	0.794	0.821	25%
337-340bp	0.679	0.853	0.824	0.006	0.760	0.778	0.819	17%
341-344bp	0.733	0.802	0.824	0.006	0.769	0.798	0.821	16%
336-340bp	0.679	0.849	0.824	0.007	0.752	0.782	0.818	23%
341-345bp	0.743	0.789	0.824	0.007	0.769	0.798	0.820	21%
337-342bp	0.679	0.853	0.825	0.009	0.769	0.774	0.820	17%
337-338bp	0.679	0.849	0.823	0.003	0.760	0.770	0.818	17%
343-345bp	0.743,	0.788	0.823	0.004	0.785	0.782	0.820	24%
341-347bp	0.686	0.842	0.823	0.010	0.785	0.778	0.819	19%
340-348bp	0.689	0.839	0.823	0.013	0.785	0.778	0.819	18%
334-340bp	0.662	0.868	0.823	0.010	0.744	0.786	0.817	22%
341-348bp	0.686	0.842	0.823	0.011	0.785	0.778	0.819	21%
345-346bp	0.699	0.830	0.823	0.003	0.769	0.794	0.819	14%
343-348bp	0.723	0.807	0.823	0.009	0.769	0.798	0.819	14%
341-350bp	0.642	0.888	0.823	0.013	0.769	0.798	0.819	25%
345-348bp	0.723	0.811	0.822	0.006	0.719	0.844	0.819	13%
346-348bp	0.723	0.807	0.822	0.004	0.719	0.844	0.819	22%
333-340bp	0.666	0.863	0.822	0.012	0.727	0.798	0.816	19%
347-348bp	0.645	0.886	0.822	0.003	0.719	0.844	0.819	12%
346-350bp	0.689	0.839	0.821	0.007	0.719	0.852	0.819	17%

[0083]

335-336bp	0.662	0.865	0.821	0.003	0.736	0.790	0.814	20%
331-340bp	0.666	0.861	0.820	0.015	0.736	0.794	0.813	19%
334-336bp	0.693	0.837	0.820	0.005	0.736	0.790	0.813	11%
349-350bp	0.652	0.879	0.820	0.003	0.719	0.848	0.819	10%
349-351bp	0.652	0.879	0.820	0.004	0.719	0.856	0.819	20%
331-339bp	0.666	0.861	0.820	0.014	0.736	0.794	0.812	17%
349-352bp	0.652	0.879	0.819	0.005	0.719	0.856	0.819	25%
333-336bp	0.652	0.875	0.819	0.006	0.736	0.794	0.812	20%
348-354bp	0.652	0.877	0.819	0.009	0.736	0.840	0.819	22%
351-352bp	0.662	0.868	0.819	0.002	0.736	0.848	0.819	39%
349-354bp	0.662	0.867	0.818	0.007	0.736	0.844	0.819	29%
349-356bp	0.662	0.870	0.817	0.010	0.736	0.848	0.818	23%
352-354bp	0.662	0.872	0.817	0.004	0.736	0.852	0.818	18%
351-355bp	0.662	0.872	0.817	0.006	0.736	0.852	0.818	26%
333-334bp	0.652	0.872	0.817	0.003	0.636	0.881	0.810	21%
331-336bp	0.649	0.875	0.817	0.009	0.727	0.79	0.810	21%
349-357bp	0.662	0.870	0.817	0.011	0.736	0.852	0.818	24%
353-354bp	0.662	0.870	0.816	0.002	0.736	0.852	0.818	38%
331-335bp	0.669	0.851	0.816	0.008	0.636	0.881	0.809	16%
353-356bp	0.720	0.807	0.815	0.005	0.719	0.864	0.817	21%
351-360bp	0.676	0.851	0.814	0.012	0.727	0.848	0.815	24%
355-356bp	0.716	0.807	0.814	0.003	0.727	0.852	0.816	4%
355-357bp	0.703	0.821	0.814	0.004	0.727	0.848	0.815	15%
331-333bp	0.662	0.858	0.813	0.005	0.653	0.860	0.805	13%
355-360bp	0.682	0.842	0.812	0.007	0.719	0.852	0.813	24%
357-358bp	0.682	0.842	0.812	0.002	0.719	0.848	0.812	35%

[0084] 由此可见，上述100个片段特征可以作为识别肿瘤的标志物。通过判断其在样本中的比例，来判断样本是否属于肿瘤样本。

[0085] 实施例3不同片段特征组合进行肿瘤识别

[0086] 1．10个标志物

[0087] 按单个特征的AUC值排序，分别取前5个肿瘤样本中增加的片段特征(T5)和前5个肿瘤样本中减少的片段特征(N5)进行组合。

[0088] 其中，

[0089] T5包括：163-164bp、163-165bp、161-165bp、161-164bp、165-166bp

[0090] N5包括：339-340bp、341-342bp、343-344bp、337-339bp、340-342bp

[0091] 针对每个样本，计算一个得分score10＝sum(T5)/sum(N5)，然后在训练集中计算该得分在肿瘤样本中相对于正常对照样本增加的比例，据此标准判断样本属于肿瘤或正常，并在测试集中进行验证。

[0092] 2．20个标志物

[0093] 按单个特征的AUC值排序，分别取前10个肿瘤样本中增加的片段特征(T10)和前10

个肿瘤样本中减少的片段特征(N10)进行组合。

[0094] 其中，

[0095] T10包括:163-164bp、163-165bp、161-165bp、161-164bp、165-166bp、159-165bp、157-164bp、155-156bp、163-168bp、160-168bp

[0096] N10包括:339-340bp、341-342bp、343-344bp、337-339bp、340-342bp、337-340bp、341-344bp、336-340bp、341-345bp、337-342bp

[0097] 针对每个样本，计算一个得分score20＝sum(T10)/sum(N10)，然后在训练集中计算该得分在肿瘤样本中相对于正常对照样本增加的比例，据此标准判断样本属于肿瘤或正常，并在测试集中进行验证。

[0098] 3．30个标志物

[0099] 按单个特征的AUC值排序，分别取前15个肿瘤样本中增加的片段特征(T15)和前15个肿瘤样本中减少的片段特征(N15)进行组合。

[0100] 其中，

[0101] T15包括:163-164bp、163-165bp、161-165bp、161-164bp、165-166bp、159-165bp、157-164bp、155-156bp、163-168bp、160-168bp、157-158bp、154-156bp、161-170bp、156-160bp、161-162bp

[0102] N15包括:339-340bp、341-342bp、343-344bp、337-339bp、340-342bp、337-340bp、341-344bp、336-340bp、341-345bp、337-342bp、337-338bp、343-345bp、341-347bp、340-348bp、334-340bp

[0103] 针对每个样本，计算一个得分score30＝sum(T15)/sum(N15)，然后在训练集中计算该得分在肿瘤样本中相对于正常对照样本增加的比例，据此标准判断样本属于肿瘤或正常，并在测试集中进行验证。

[0104] 4．40个标志物

[0105] 按单个特征的AUC值排序，分别取前20个肿瘤样本中增加的片段特征(T20)和前20个肿瘤样本中减少的片段特征(N20)进行组合。

[0106] 其中，

[0107] T20包括:163-164bp、163-165bp、161-165bp、161-164bp、165-166bp、159-165bp、157-164bp、155-156bp、163-168bp、160-168bp、157-158bp、154-156bp、161-170bp、156-160bp、161-162bp、157-159bp、165-168bp、157-162bp、157-160bp、151-160bp

[0108] N20包括:339-340bp、341-342bp、343-344bp、337-339bp、340-342bp、337-340bp、341-344bp、336-340bp、341-345bp、337-342bp、337-338bp、343-345bp、341-347bp、340-348bp、334-340bp、341-348bp、345-346bp、343-348bp、341-350bp、345-348bp

[0109] 针对每个样本，计算一个得分score40＝sum(T20)/sum(N20)，然后在训练集中计算该得分在肿瘤样本中相对于正常对照样本增加的比例，据此标准判断样本属于肿瘤或正常，并在测试集中进行验证。

[0110] 5．50个标志物

[0111] 按单个特征的AUC值排序，分别取前25个肿瘤样本中增加的片段特征(T25)和前25个肿瘤样本中减少的片段特征(N25)进行组合。

[0112] 其中，

[0113] T25包括:163-164bp、163-165bp、161-165bp、161-164bp、165-166bp、159-165bp、157-164bp、155-156bp、163-168bp、160-168bp、157-158bp、154-156bp、161-170bp、156-160bp、161-162bp、157-159bp、165-168bp、157-162bp、157-160bp、151-160bp、152-158bp、160-162bp、153-156bp、151-159bp、166-168bp

[0114] N25包括:339-340bp、341-342bp、343-344bp、337-339bp、340-342bp、337-340bp、341-344bp、336-340bp、341-345bp、337-342bp、337-338bp、343-345bp、341-347bp、340-348bp、334-340bp、341-348bp、345-346bp、343-348bp、341-350bp、345-348bp、346-348bp、333-340bp、347-348bp、346-350bp、335-336bp

[0115] 针对每个样本,计算一个得分score50＝sum(T25)/sum(N25),然后在训练集中计算该得分在肿瘤样本中相对于正常对照样本增加的比例,据此标准判断样本属于肿瘤或正常,并在测试集中进行验证。

[0116] 6.60个标志物

[0117] 按单个特征的AUC值排序,分别取前30个肿瘤样本中增加的片段特征(T30)和前30个肿瘤样本中减少的片段特征(N30)进行组合。

[0118] 其中,

[0119] T30包括:163-164bp、163-165bp、161-165bp、161-164bp、165-166bp、159-165bp、157-164bp、155-156bp、163-168bp、160-168bp、157-158bp、154-156bp、161-170bp、156-160bp、161-162bp、157-159bp、165-168bp、157-162bp、157-160bp、151-160bp、152-158bp、160-162bp、153-156bp、151-159bp、166-168bp、148-150bp、149-150bp、149-156bp、159-160bp、151-156bp

[0120] N30包括:339-340bp、341-342bp、343-344bp、337-339bp、340-342bp、337-340bp、341-344bp、336-340bp、341-345bp、337-342bp、337-338bp、343-345bp、341-347bp、340-348bp、334-340bp、341-348bp、345-346bp、343-348bp、341-350bp、345-348bp、346-348bp、333-340bp、347-348bp、346-350bp、335-336bp、331-340bp、334-336bp、349-350bp、349-351bp、331-339bp

[0121] 针对每个样本,计算一个得分score60＝sum(T30)/sum(N30),然后在训练集中计算该得分在肿瘤样本中相对于正常对照样本增加的比例,据此标准判断样本属于肿瘤或正常,并在测试集中进行验证。

[0122] 7.不同标志物组合的判断结果

[0123] 根据上述6种标志物组合的得分,在训练集和测试集中的判断结果如图1～6及表4所示:

[0124] 表4不同标志物组合的肿瘤识别结果

[0125]

得分	训练集/Training set				测试集/Testing set			癌症样本增加比例（至少）
	灵敏度 Sensitivity	特异度 Specificity	AUC	Cutoff	灵敏度 Sensitivity	特异度 Specificity	AUC	
score10	0.703	0.816	0.818	19.784	0.744	0.823	0.818	11%
score20	0.753	0.765	0.817	17.591	0.744	0.819	0.816	5%
score30	0.750	0.768	0.816	14.019	0.744	0.819	0.816	5%
score40	0.770	0.747	0.815	12.229	0.744	0.823	0.815	3%
score50	0.770	0.747	0.815	12.064	0.744	0.815	0.814	3%
score60	0.753	0.763	0.815	10.748	0.760	0.794	0.813	4%

[0126] 由上表可知，利用10个片段特征即可以很好地识别出肿瘤样本，进一步增加片段特征没有使得识别效果更好，反面有一定程度的降低，表明利用10个片段特征具有较好的肿瘤识别效果，可能通过计算得分预测受试者是否患有肿瘤或者是否具有患肿瘤的风险。

[0127] 实施例4 10个片段特征组成的标志物组合在外部测试集中的验证

[0128] 为了进一步验证上述10个片段特征作为预测肿瘤的标志物的性能，发明人使用外部测试集(external data)进行进一步验证，结果如图7所示。

[0129] 从图7中可以看出，利用10个片段特征得到的得分可以明显区别肿瘤样本和正常样本，具体地，在肿瘤样本中得分显著高于正常样本的得分，ROC曲线AUC达到0.827。

[0130] 在本发明提及的所有文献都在本申请中引用作为参考，就如同每一篇文献被单独引用作为参考那样。此外应理解，在阅读了本发明的上述讲授内容之后，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。

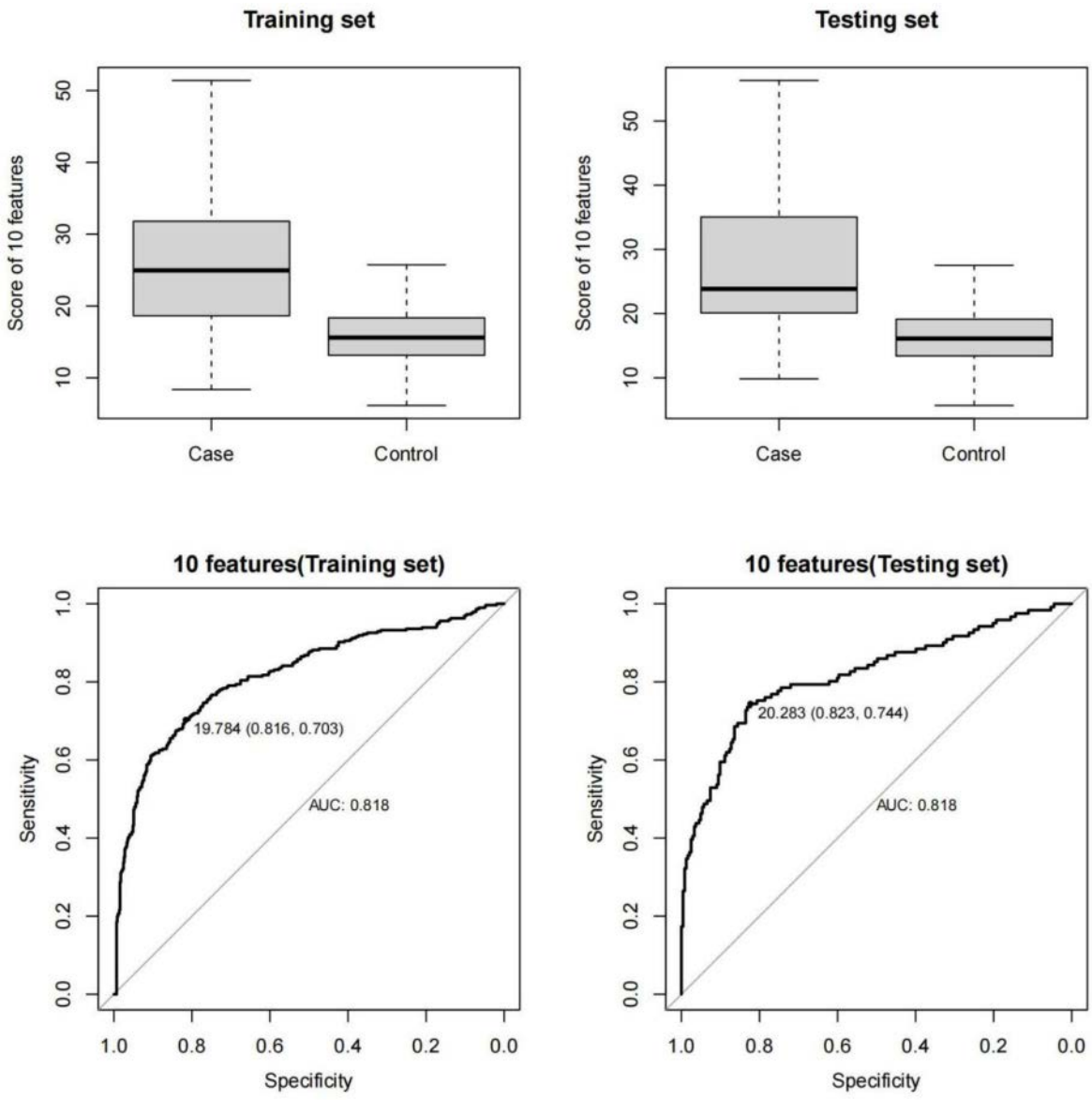

Training set
Score of 10 features
10
20
30
40
50
Case
Control
Testing set
Score of 10 features
10
20
30
40
50
Case
Control
10 features(Training set)
Sensitivity
0.0
0.2
0.4
0.6
0.8
1.0
19.784 (0.816, 0.703)
AUC: 0.818
1.0
0.8
0.6
0.4
0.2
0.0
Specificity
10 features(Testing set)
Sensitivity
0.0
0.2
0.4
0.6
0.8
1.0
20.283 (0.823, 0.744)
AUC: 0.818
1.0
0.8
0.6
0.4
0.2
0.0
Specificity

图1

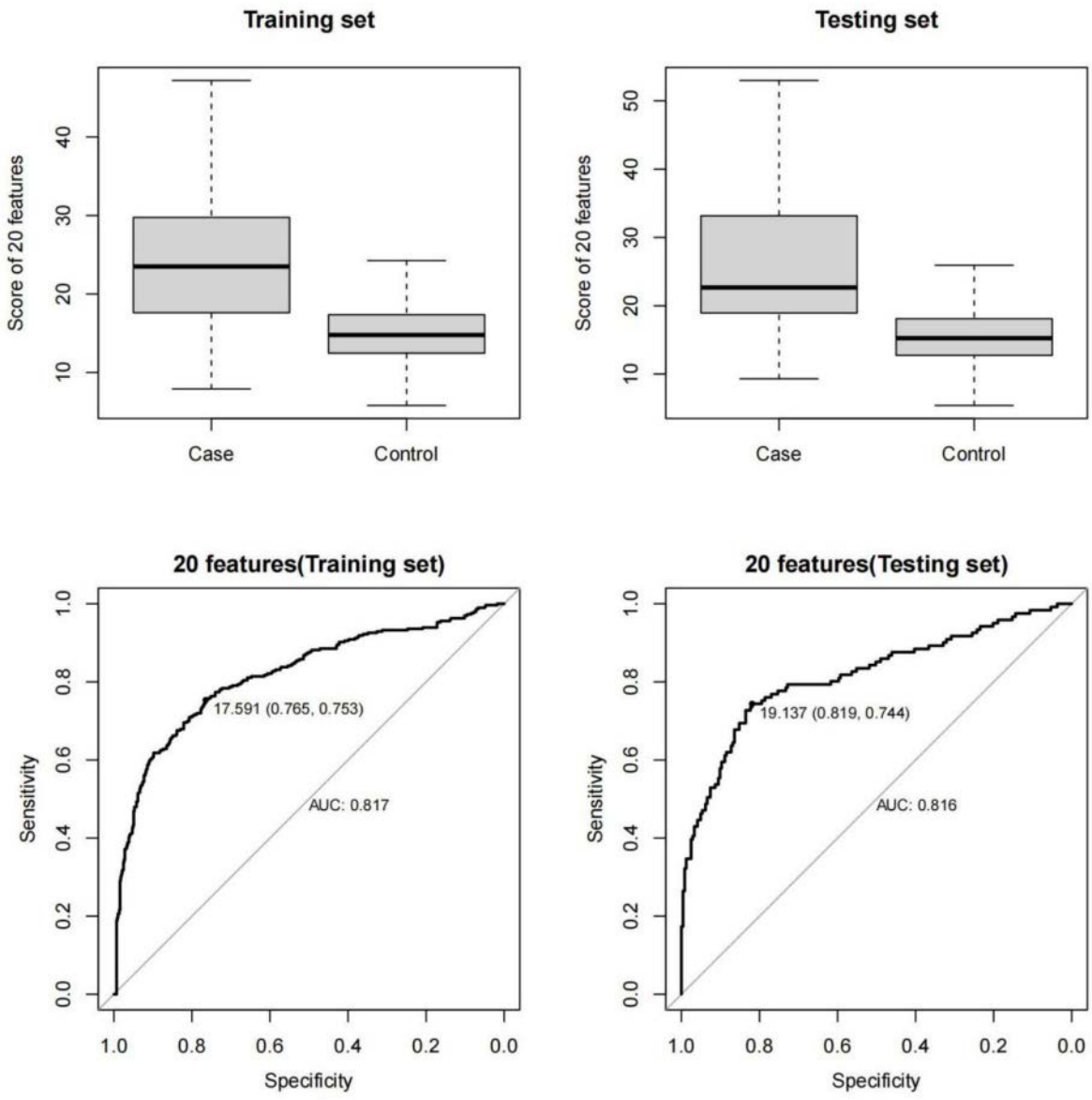

Training set
Score of 20 features
10
20
30
40
Case
Control
Testing set
Score of 20 features
10
20
30
40
50
Case
Control
20 features(Training set)
Sensitivity
0.0
0.2
0.4
0.6
0.8
1.0
17.591 (0.765, 0.753)
AUC: 0.817
1.0
0.8
0.6
0.4
0.2
0.0
Specificity
20 features(Testing set)
Sensitivity
0.0
0.2
0.4
0.6
0.8
1.0
19.137 (0.819, 0.744)
AUC: 0.816
1.0
0.8
0.6
0.4
0.2
0.0
Specificity

图2

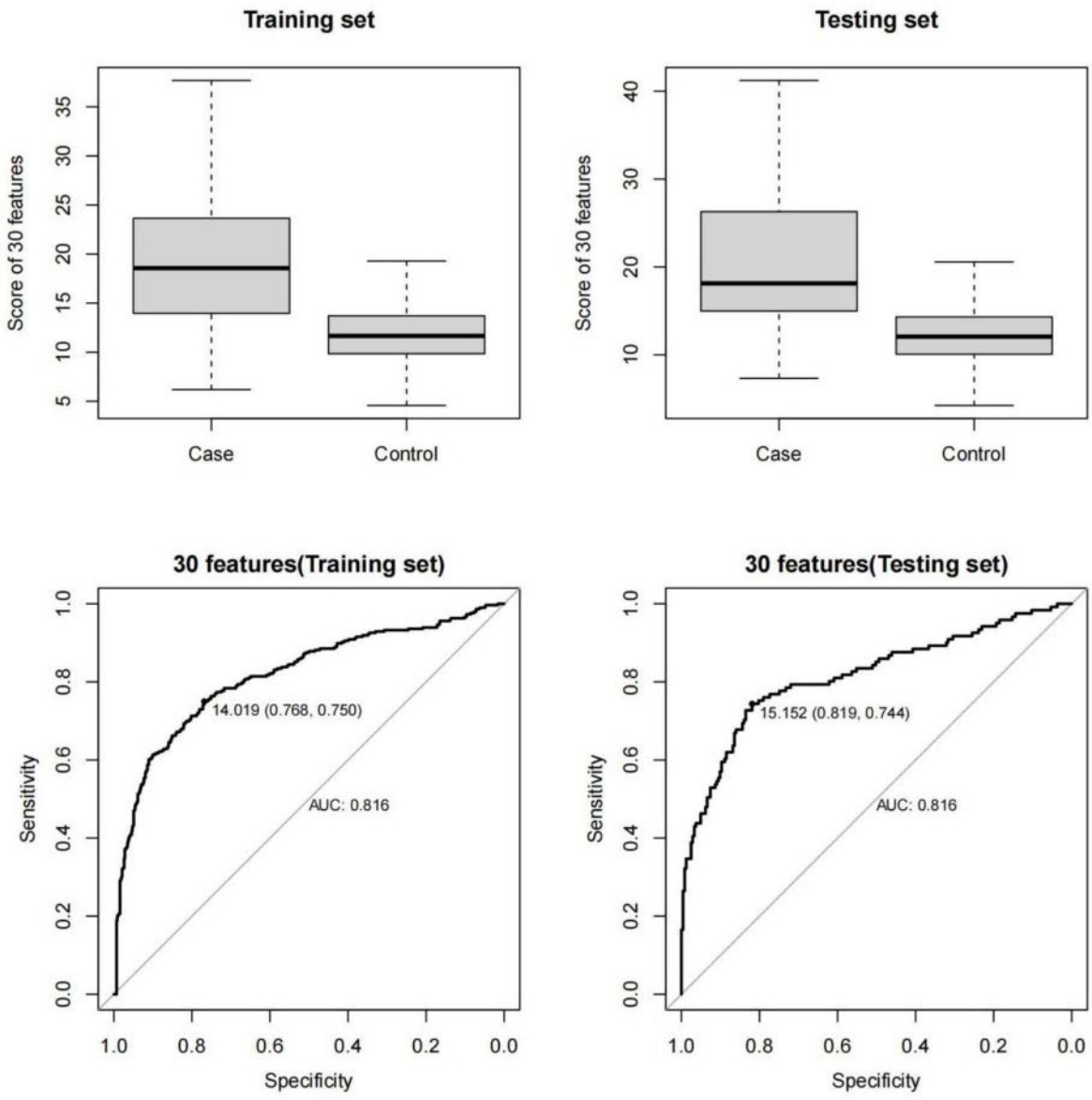

Training set
Score of 30 features
5
10
15
20
25
30
35
Case
Control
Testing set
Score of 30 features
10
20
30
40
Case
Control
30 features(Training set)
Sensitivity
0.0
0.2
0.4
0.6
0.8
1.0
14.019 (0.768, 0.750)
AUC: 0.816
1.0
0.8
0.6
0.4
0.2
0.0
Specificity
30 features(Testing set)
Sensitivity
0.0
0.2
0.4
0.6
0.8
1.0
15.152 (0.819, 0.744)
AUC: 0.816
1.0
0.8
0.6
0.4
0.2
0.0
Specificity

图3

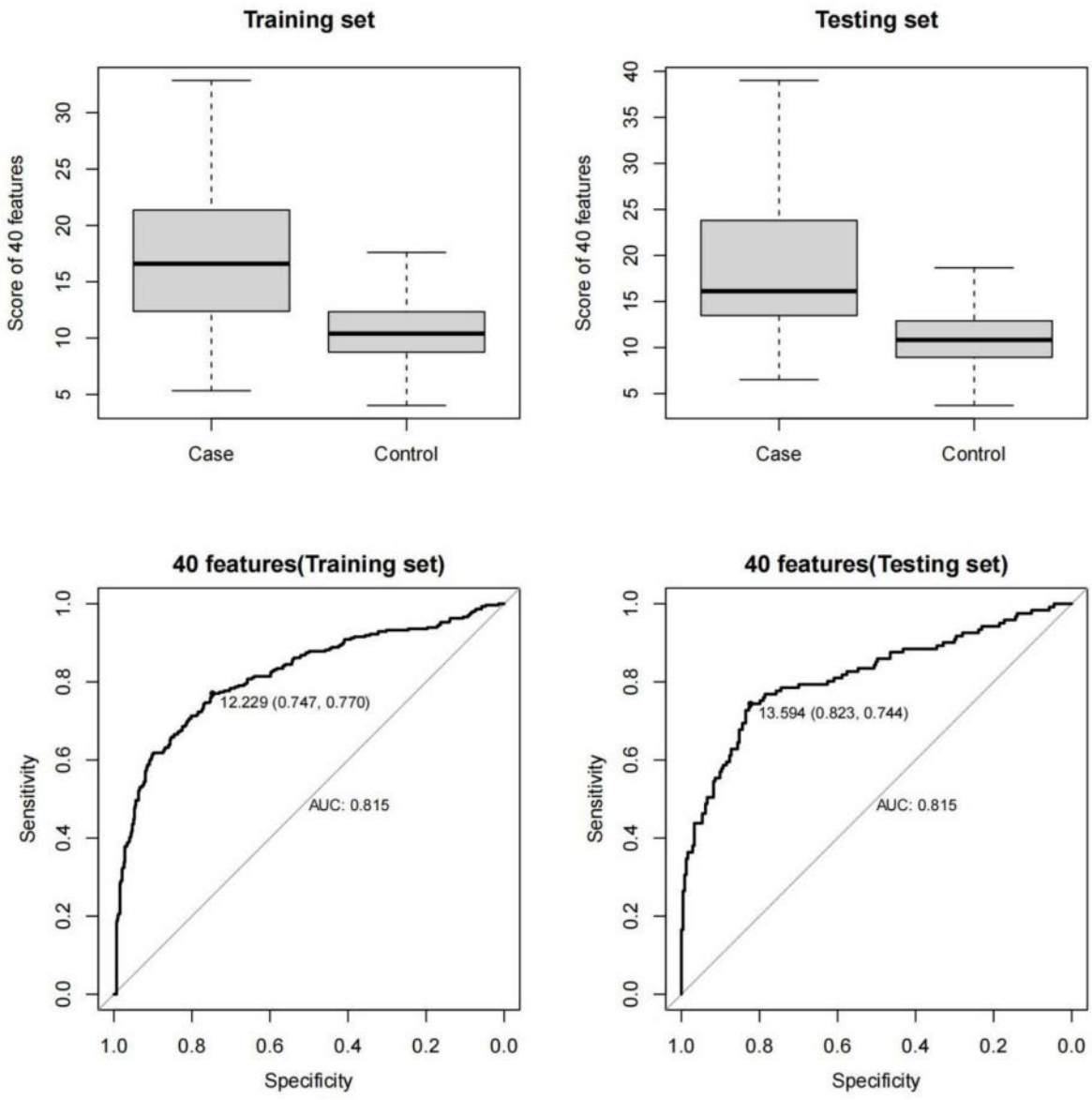

Training set
Score of 40 features
Case
Control
Testing set
Score of 40 features
Case
Control
40 features(Training set)
Sensitivity
Specificity
12.229 (0.747, 0.770)
AUC: 0.815
40 features(Testing set)
Sensitivity
Specificity
13.594 (0.823, 0.744)
AUC: 0.815

图4

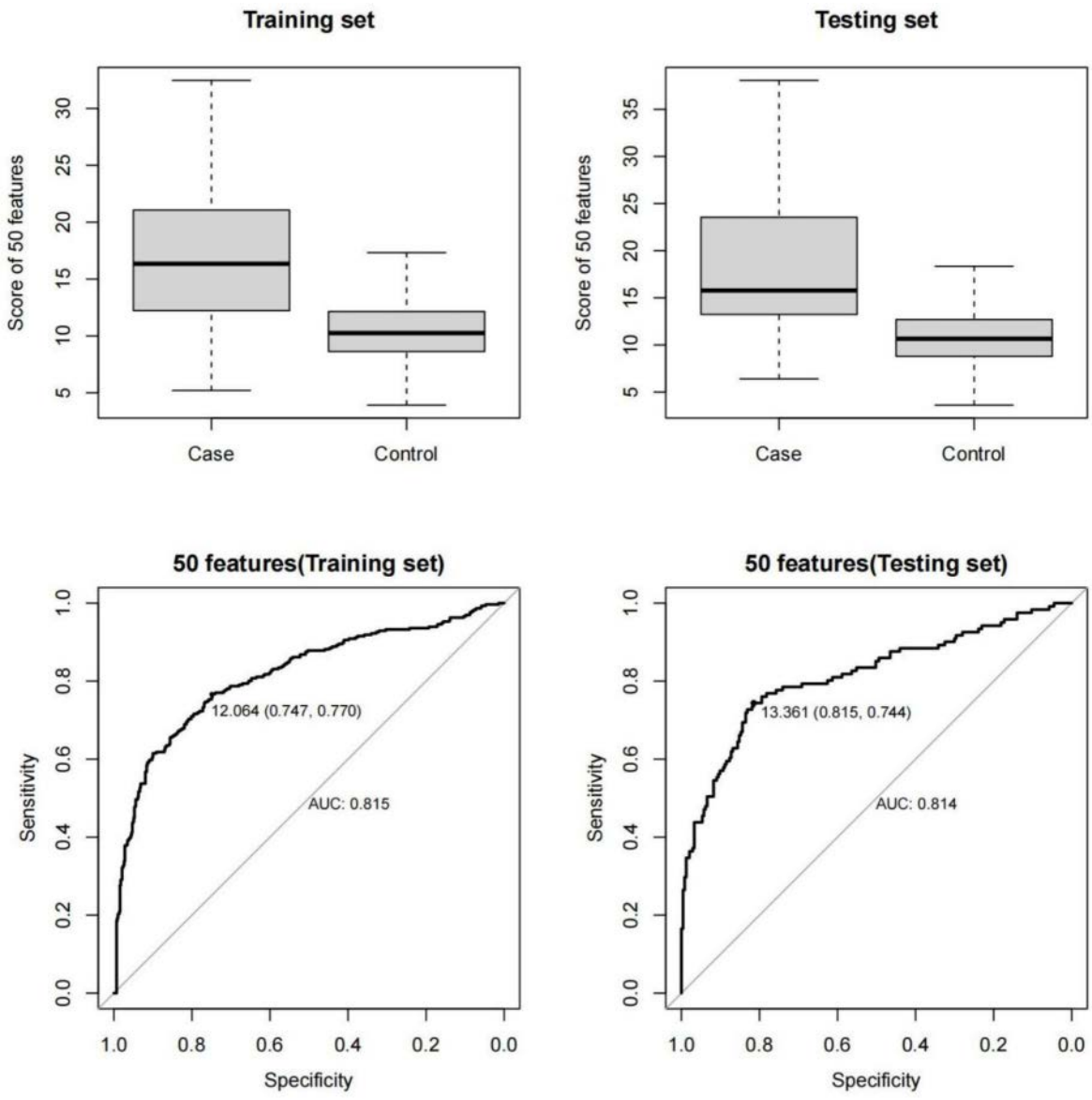

Training set
Score of 50 features
Case
Control
Testing set
Score of 50 features
Case
Control
50 features(Training set)
Sensitivity
Specificity
12.064 (0.747, 0.770)
AUC: 0.815
50 features(Testing set)
Sensitivity
Specificity
13.361 (0.815, 0.744)
AUC: 0.814

图5

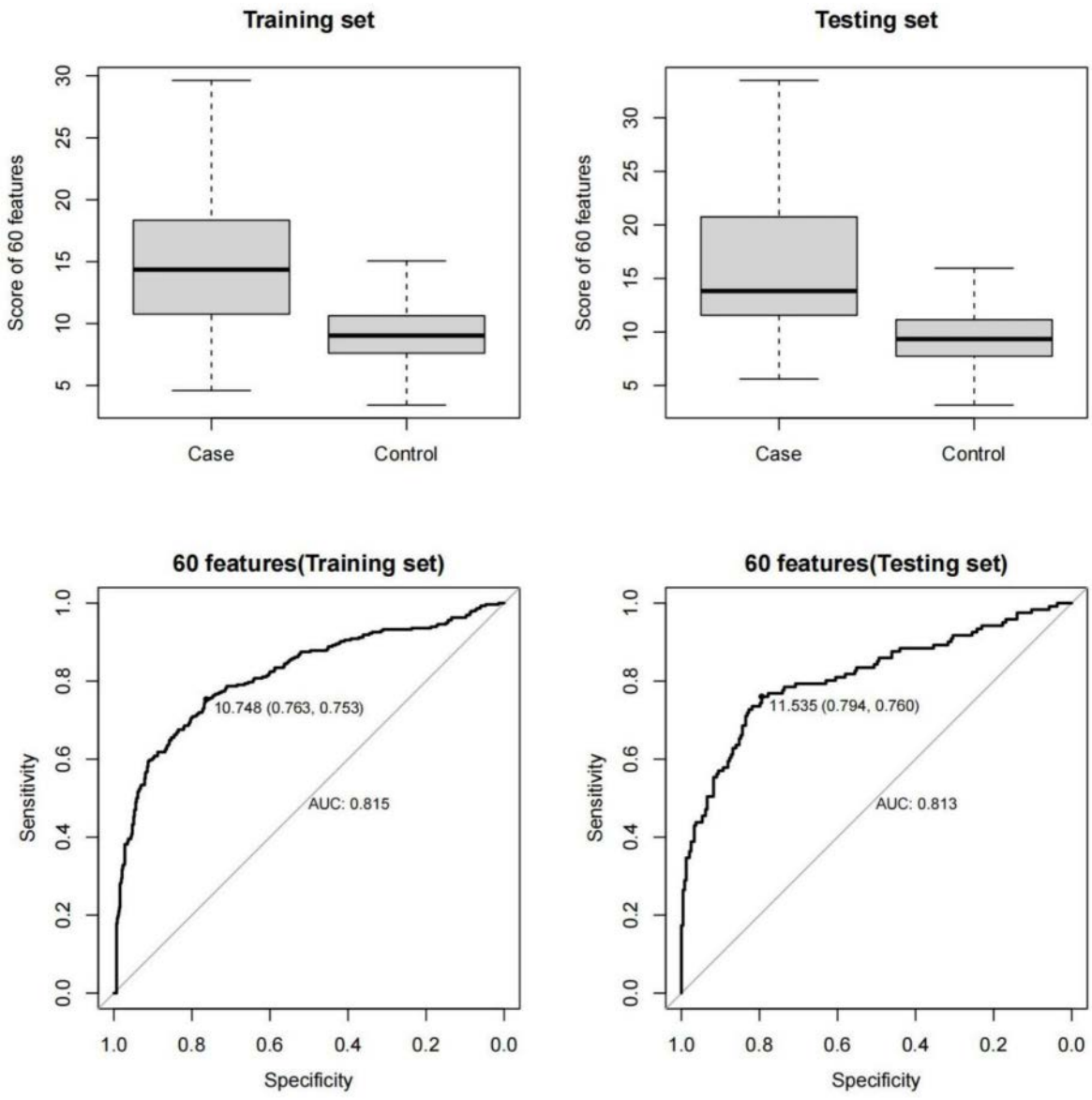

Training set
Score of 60 features
Case
Control
Testing set
Score of 60 features
Case
Control
60 features(Training set)
Sensitivity
Specificity
10.748 (0.763, 0.753)
AUC: 0.815
60 features(Testing set)
Sensitivity
Specificity
11.535 (0.794, 0.760)
AUC: 0.813

图6

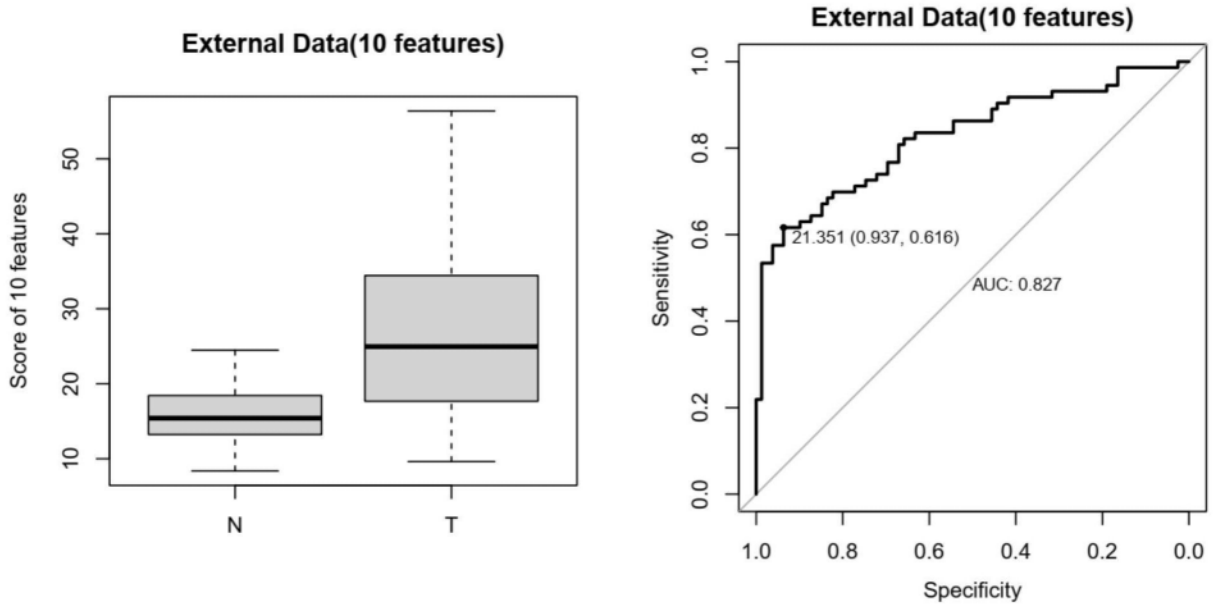

External Data(10 features)
Score of 10 features
10
20
30
40
50
N
T
External Data(10 features)
Sensitivity
0.0
0.2
0.4
0.6
0.8
1.0
21.351 (0.937, 0.616)
AUC: 0.827
1.0
0.8
0.6
0.4
0.2
0.0
Specificity

图7