(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 109616198 A
(43)申请公布日 2019.04.12

(21)申请号 201811617489.1

(22)申请日 2018.12.28

(71)申请人 陈洪亮
地址 364000 福建省龙岩市新罗区东肖镇后田村下洋尾路38号

(72)发明人 陈洪亮

(51)Int.Cl.
G16H 50/20(2018.01)
G16H 50/70(2018.01)
G16B 30/00(2019.01)
G16B 40/00(2019.01)

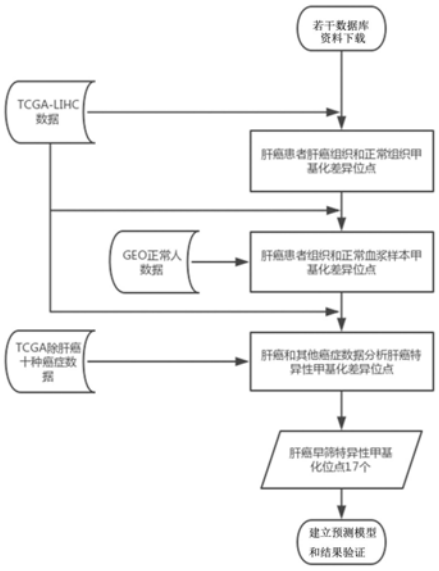
权利要求书3页 说明书8页 附图1页

(54)发明名称

仅用于肝癌单一癌种筛查的特异甲基化检测位点组合的选取方法

(57)摘要

本发明公开一种仅用于肝癌单一癌种筛查的特异甲基化检测位点组合的选取方法，其通过一系列的筛选步骤寻找针对单一肝癌的特异甲基化位点，其主要通过设计不同样本类型的对比，校准个体甲基化差异、不同时期甲基化差异、不同组织甲基化差异和不同肿瘤间甲基化差异，从而获得一组特异甲基化位点作为诊断标志物来检测肝癌；本方法筛选出来肝癌特异甲基化位点的敏感性能达到90%以上，特异性能达到97%以上，能在诊断过程中表现为只针对肝癌这个癌种进行检测。



CN 109616198 A

1.仅用于肝癌单一癌种筛查的特异甲基化检测位点组合的选取方法,其特征在于:该方法包括以下步骤:

步骤1:从若干数据库中集中肝癌和其他癌症甲基化数据,其中,数据包括正常人的样本及癌症患者的样本;

步骤2:比较肝癌患者肝癌组织和正常人的正常组织中甲基化数据,以找出肝癌和正常组织之间甲基化差异位点,并将获得的甲基化差异位点进行信息注释;

步骤3:从步骤1的数据库中分析、获得肝癌特异性甲基化差异位点,具体如下:

(a)、将步骤1中所有肝癌患者的肝癌组织甲基化数据及50个患者的正常组织的甲基化数据进行对比,并根据步骤2中筛选出的甲基化差异位点,将这些样本中差异位点数据整理成一个文件;

(b)、将步骤1 中375个肝癌患者的肝癌组织甲基化数据和正常人血浆样本甲基化数据合并后,并筛选处步骤3(a)中找出的甲基化差异位点信息整理出一个文件;

(c)、将步骤1中375个肝癌患者肝癌组织甲基化数据和数据库中的其他癌症的数据集合并,根据步骤3(b)中选出的肝癌特异性甲基化差异位点,并将整理出这些位点再所有肿瘤中的数据成一个文件;

(d)、将上述或的文件集中,并选出的位点组合,从而获得肝癌早期筛查用的17个特异性甲基化位点,作为肝癌肿瘤标志物;

步骤4:对17个特异性甲基化位点建立预测模型和结果验证,进行评估。

2.根据权利要求1中所述仅用于肝癌单一癌种筛查的特异甲基化检测位点组合的选取方法,其特征在于:所述步骤1中若干个数据库均为现有数据库,各现有数据库内下载大量肝癌和其他癌症甲基化数据,以及正常人的样本。

3.根据权利要求2中所述仅用于肝癌单一癌种筛查的特异甲基化检测位点组合的选取方法,其特征在于:各数据库中的数据均是HumanMethylation450 BeadChip(GPL13534)芯片数目,相同的数据格式才能进行对比分析,同时可以排除不同平台的偏差。

4.根据权利要求2或3中所述仅用于肝癌单一癌种筛查的特异甲基化检测位点组合的选取方法,其特征在于:所述步骤1中所述若干数据库为:

(a)、从NIH网上进入GDC的TCGA数据库,下载带TCGA-LIHC标签的肝癌DNA甲基化、基因表达数据和临床信息注释文件;

(b)、 从TCGA数据库下载其他10种癌症的DNA甲基化数据,所下载的癌症数据包括:

BLCA(409肿瘤,21正常),BRCA(774肿瘤,82正常),COAD(292肿瘤,38正常),GBM(126肿瘤,2正常) ,HNSC(523肿瘤,45正常),KIRC(316肿瘤,160正常),LUAD(455肿瘤,32正常),LUSC(365肿瘤,41正常),READ(95肿瘤,7正常)和UCEC(425肿瘤),46正常);

(c)、从GEO数据库下载甲基化数据集GSE69270(184名年轻芬兰人的血液),GSE54503(66配对的肿瘤和正常),GSE89852(37配对的肿瘤和正常),GSE56588(224 肿瘤,9个肝硬化,10正常)。

5.根据权利要求1中所述仅用于肝癌单一癌种筛查的特异甲基化检测位点组合的选取方法,其特征在于:所述步骤2包括以下步骤:

(a)、在步骤1中选出同时具有正常组织甲基化数据和肝癌组织甲基化数据的50个患者;

(b)、根据上述选的50个患者,将每个患者正常数据和肿瘤数据整理在一个文件,过滤掉缺失较多的位点,建立文件,该文件的行是位点名字,列是样本编号;

(c)、根据上述(b)中获得文件计算正常和肝癌组织的甲基化差异,记作p值,同时用p.adjust命令对T-test结果进行校正,记作FDR;

(d)、根据上步计算的p值和FDR,用P值小于0.05,FDR大于0.2,作为筛选条件选出符合条件的位点备用;

(e)、将(d)得出的位点,利用HumanMethylation450 BeadChip(GPL13534)对位点信息进行含位点所在基因的注释;

(f)、将注释后的差异位点所在基因与差异表达基因统计分析,找出共同基因,这些基因认为是甲基化差异导致基因表达有差异;

(h)、根据(f)的注释结果,选出位于启动子区的位点(TSS1500|TSS200)备用。

6.根据权利要求5中所述仅用于肝癌单一癌种筛查的特异甲基化检测位点组合的选取方法,其特征在于:所述差异表达基因的获得步骤如下:

步骤1):于步骤1中的数据库选出同时具有肿瘤甲基化和表达谱的41对样本;

步骤2):将上述步骤1)选出的41对样本的肝癌部位和正常部位的表达谱数据以配对方式整理成备用文件,文件行是基因名,列名是样本编号;

步骤3):将上述步骤2)的数据文件利用软件是Bioconductor package edgeR,选用基于广义线性模型的统计方法模式鉴定表达差异基因;进而计算出每个表达基因的结果,该结果作为41个人肝癌和正常组织表达差异的衡量指标;

步骤4):将上述步骤3)的计算出的每个基因结果,筛选出FDR小于0.05且绝对值log 2(fold change)大于1的位点,筛选出的被认为是具有差异表达的基因,其中表达差异包括肝癌组织比正常组织高表达,或者正常组织比肝癌组织高表达两种情况;

步骤5):上述步骤4)中选取的条件不限于FDR小于0.05且绝对值log 2(fold change)大于1,是统计中表明两组数据有显著差异的条件。

7.根据权利要求1中所述仅用于肝癌单一癌种筛查的特异甲基化检测位点组合的选取方法,其特征在于:所述步骤3中(d)的具体计算方法如下:

(1)、计算肝癌的均值和其他所有肿瘤数据的均值,将肝癌均值减肿瘤均值绝对值高于0.1的位点过滤出来作为肝癌特异性早筛位点,即(chr2:166650805、chr2:232260305、chr2:9144605、chr3:123167770、chr3:101497876、chr3:101497857、chr3:101497980、chr3:101497982、chr6:116691863、chr8:102504447、chr8:102504482、chr8:102504501、chr10:21463485、chr11:67350976、chr11:66624853、chr17:4981610、chr20:44540794);

(2)、上述(1)选出的位点组合,每个位点所在CpG岛内相邻位甲基化点具有同样,包括但不限于450k芯片中找不出的位点,因为每个CpG岛的甲基化位点具有一致性,故,所述位点组合拓展至位点所在CpG岛的组合,即(chr10:21462128-21463808、chr11:66623620-66626614、chr11:67350928-67351953、chr17:4981357-4981979、chr2:166649909-166650966、chr2:232260100-232261134、chr2:9143127-9144630、chr20:44540445-44540957、chr3:101497830-101498648、chr3:123166218-123168567、chr6:116691827-116692868、chr8:102504478-102504841)。

8.根据权利要求1中所述仅用于肝癌单一癌种筛查的特异甲基化检测位点组合的选取

方法，其特征在于：所述步骤4的为使用weka软件进行建模和结果验证，且具体步骤如下：

(a)、将TCGA-LIHC的具有正常组织甲基化数据和肝癌组织数据的50个患者的甲基化数据集作为训练集，使用weka软件，输入17个候选位点建立预测模型，所选模型为J48模型、DecisionStump模型、LMT模型、REPtree模型、RandomForest模型、NaiveBayes模型、logistic模型、MultilayerPerceptron模型；

(b)、将四个其他独立数据集(GSE54503、GSE89852、GSE56588)作为测试集，使用weka软件，使用(a)得到的模型来测试模型效果；

(c)、记录模型在得到的模型效果，包括敏感性、特异性以评估预测模型的准确性，以及选择效果最优的模型。

仅用于肝癌单一癌种筛查的特异甲基化检测位点组合的选取方法

技术领域

[0001] 本发明涉及生物信息领域，尤其是涉及一种能通过一系列步骤筛选出一组特异甲基化位点作为肝癌筛查的标志物的方法。

背景技术

[0002] 随着对肿瘤研究的深入，在癌症的诊断和治疗过程中发现组织活检技术有一定的局限性。主要表现为：肿瘤具有异质性，对于癌细胞已经发生转移的患者而言，仅仅取某个部位的肿瘤组织，并不能反映患者的整体情况，但对所有的肿瘤组织都取样检测又不切实际；某些患者自身的情况决定了他不适合做组织活检；受到手术的扰动之后，有些肿瘤有加速转移的风险；组织活检的滞后性对患者的治疗也是不利的。

[0003] 因此，对于癌症的诊断和检测技术有更高的要求，液体活检技术的出现，解决了上述的问题，也提前了癌症的诊断时间。

[0004] 作为体外诊断的一个分支，液体活检就是通过血液或者尿液等对癌症等疾病做出诊断。以血中循环肿瘤细胞或循环肿瘤DNA为检测靶标对肿瘤进行诊断和预测，就需要选择合适的肿瘤标记物，DNA甲基化与癌症的发生有着密切的关系，在许多癌症中都发现存在DNA甲基化异常的现象。DNA甲基化具有一定的稳定性，它是癌症发生中的复发事件。近年来许多研究证明，DNA的甲基化异常可以作为一种癌症诊断的生物标志物。

[0005] 近几年有不少研究希望找出一组甲基化位点 作为肿瘤诊断或者预后的标志物，然而人体甲基化位点数目太多，且甲基化位点在不同个体间有差异性，不同器官甚至不同时间都有差异性。目前，现有的研究方法大多根据现有文献选取特定基因后进一步研究(中国专利申请号为201510688727 .8，及201680012042.4的公开文献)，或者侧重甲基化实验，仅用简单统计筛选，缺乏系统严谨的步骤(中国专利申请号为201710413695 .X的公开文献)，或者一些研究单纯列出各种数学模型，没有过程，没有结果评估。

[0006] 因此，如何针对于单一癌症，即肝癌特异性甲基化位点的方法，以选出一组位点作为诊断标志物，同时评估其敏感性和特异性是本领域技术人员需要解决的重要技术问题。

发明内容

[0007] 本发明解决的问题是如何针对肝癌特异性进行甲基化位点选定，并选出一组位点作为诊断标志物，同时评估其敏感性和特异性。

[0008] 为解决上述问题，本发明提供一种仅用于肝癌单一癌种筛查的特异甲基化检测位点组合的选取方法的技术方案，该方法包括以下步骤：

步骤1：从若干数据库中集中肝癌和其他癌症甲基化数据，其中，数据包括正常人的样本及癌症患者的样本；

步骤2：比较肝癌患者肝癌组织和正常人的正常组织中甲基化数据，以找出肝癌和正常组织之间甲基化差异位点，并将获得的甲基化差异位点进行信息注释；

步骤3:从步骤1的数据库中分析、获得肝癌特异性甲基化差异位点,具体如下:

(a)、将步骤1中所有肝癌患者的肝癌组织甲基化数据及50个患者的正常组织的甲基化数据进行对比,并根据步骤2中筛选出的甲基化差异位点,将这些样本中差异位点数据整理成一个文件;

(b)、将步骤1 中375个肝癌患者的肝癌组织甲基化数据和正常人血浆样本甲基化数据合并后,并筛选处步骤3(a)中找出的甲基化差异位点信息整理出一个文件;

(c)、将步骤1中375个肝癌患者肝癌组织甲基化数据和数据库中的其他癌症的数据集合并,根据步骤3(b)中选出的肝癌特异性甲基化差异位点,并将整理出这些位点再所有肿瘤中的数据成一个文件;

(d)、将上述或的文件集中,并选出的位点组合,从而获得肝癌早期筛查用的17个特异性甲基化位点,作为肝癌肿瘤标志物;

步骤4:对17个特异性甲基化位点建立预测模型和结果验证,进行评估。

[0009] 进一步优选的:所述步骤1中若干个数据库均为现有数据库,各现有数据库内下载大量肝癌和其他癌症甲基化数据,以及正常人的样本。

[0010] 进一步优选的:各数据库中的数据均是HumanMethylation450 BeadChip(GPL13534)芯片数目,相同的数据格式才能进行对比分析,同时可以排除不同平台的偏差。

[0011] 进一步优选的:所述步骤1中所述若干数据库为:

(a)、从NIH网上进入GDC的TCGA数据库,下载带TCGA-LIHC标签的肝癌DNA甲基化、基因表达数据和临床信息注释文件;

(b)、 从TCGA数据库下载其他10种癌症的DNA甲基化数据,所下载的癌症数据包括:

BLCA(409肿瘤,21正常),BRCA(774肿瘤,82正常),COAD(292肿瘤,38正常),GBM(126肿瘤,2正常) ,HNSC(523肿瘤,45正常),KIRC(316肿瘤,160正常),LUAD(455肿瘤,32正常),LUSC(365肿瘤,41正常),READ(95肿瘤,7正常)和UCEC(425肿瘤 ,46正常);

(c)、从GEO数据库下载甲基化数据集GSE69270(184名年轻芬兰人的血液),GSE54503(66配对肿瘤和正常),GSE89852(37配对肿瘤和正常),GSE56588 (224 肿瘤,9个肝硬化,10个正常)。

[0012] 进一步优选的:所述步骤2包括以下步骤:

(a)、在步骤1中选出同时具有正常组织甲基化数据和肝癌组织甲基化数据的50个患者;

(b)、根据上述选的50个患者,将每个患者正常数据和肿瘤数据整理在一个文件,过滤掉缺失较多的位点,建立一文件,该文件的行是位点名字,列是样本编号;

(c)、根据上述(b)中获得文件计算正常和肝癌组织的甲基化差异,记作p值,同时用p.adjust命令对T-test结果进行校正,记作FDR;

(d)、根据上步计算的p值和FDR,用P值小于0.05,FDR大于0.2,作为筛选条件选出符合条件的位点备用;

(e)、将(d)得出的位点,利用HumanMethylation450 BeadChip(GPL13534)对位点信息进行含位点所在基因的注释;

(f)、将注释后的差异位点所在基因与差异表达基因统计分析,找出共同基因,这些基因认为是甲基化差异导致基因表达有差异;

(h)、根据(f)的注释结果,选出位于启动子区的位点(TSS1500|TSS200)备用。

[0013] 进一步优选的:所述差异表达基因的获得步骤如下:

步骤1):于步骤1中的数据库选出同时具有HCC甲基化和表达谱的41对样本;

步骤2):将上述步骤1)选出的41对样本的肝癌部位和正常部位的表达谱数据以配对方式整理成备用文件,文件行是基因名,列名是样本编号;

步骤3):将上述步骤2)的数据文件利用软件是Bioconductor package edgeR,选用基于广义线性模型的统计方法模式鉴定表达差异基因;进而计算出每个表达基因的结果,该结果作为41个人肝癌和正常组织表达差异的衡量指标;

步骤4):将上述步骤3)的计算出的每个基因结果,筛选出FDR小于0.05且绝对值log 2(fold change)大于1的位点,筛选出的被认为是具有差异表达的基因,其中表达差异包括肝癌组织比正常组织高表达,或者正常组织比肝癌组织高表达两种情况;

步骤5):上述步骤4)中选取的条件不限于FDR小于0.05且绝对值log 2(fold change)大于1,是统计中表明两组数据有显著差异的条件。

[0014] 进一步优选的:所述步骤3中(d)的具体计算方法如下:

(1)、计算肝癌的均值和其他所有肿瘤数据的均值,将肝癌均值减肿瘤均值绝对值高于0.1的位点过滤出来作为肝癌特异性早筛位点,即(chr2:166650805、chr2:232260305、chr2:9144605、chr3:123167770、chr3:101497876、chr3:101497857、chr3:101497980、chr3:101497982、chr6:116691863、chr8:102504447、chr8:102504482、chr8:102504501、chr10:21463485、chr11:67350976、chr11:66624853、chr17:4981610、chr20:44540794);

(2)、上述(1)选出的位点组合,每个位点所在CpG岛内相邻位甲基化点具有同样,包括但不限于450k芯片中找不出的位点,因为每个CpG岛的甲基化位点具有一致性,故,所述位点组合拓展至位点所在CpG岛的组合,即(chr10:21462128-21463808、chr11:66623620-66626614、chr11:67350928-67351953、chr17:4981357-4981979、chr2:166649909-166650966、chr2:232260100-232261134、chr2:9143127-9144630、chr20:44540445-44540957、chr3:101497830-101498648、chr3:123166218-123168567、chr6:116691827-116692868、chr8:102504478-102504841)。

[0015] 进一步优选的:所述步骤4的为使用weka软件进行建模和结果验证,且具体步骤如下:

(a)、将TCGA-LIHC的具有正常组织甲基化数据和肝癌组织数据的50个患者的甲基化数据集作为训练集,使用weka软件,输入17个候选位点建立预测模型,所选模型为J48模型、DecisionStump模型、LMT模型、REPtree模型、RandomForest模型、NaiveBayes模型、logistic模型、MultilayerPerceptron模型;

(b)、将四个其他独立数据集(GSE54503、GSE89852、GSE56588)作为测试集,使用weka软件,使用(a)得到的模型来测试模型效果;

(c)、记录模型在得到的模型效果,包括敏感性、特异性以评估预测模型的准确性,以及选择效果最优的模型。

[0016] 与现有技术相比,本发明具有以下优点:

本发明通过一系列的筛选步骤寻找处针对单一肝癌的特异甲基化位点,其主要通过设计不同样本类型的对比,校准个体甲基化差异,不同时期甲基化差异,不同组织甲基化差异

和不同肿瘤间甲基化差异，从而获得一组特异甲基化位点作为诊断标志物来检测肝癌；本方法筛选出来肝癌特异甲基化位点的敏感性能达到90%以上，特异性能达到97%以上，能在诊断过程中表现为只针对肝癌这个癌种进行检测。

附图说明

[0017] 图1是本发明实施例中流程框图。

具体实施方式

[0018] 现有的研究方法大多根据现有文献选取特定基因后进一步研究(中国专利申请号为201510688727 .8，及201680012042.4的公开文献)，或者侧重甲基化实验，仅用简单统计筛选，缺乏系统严谨的步骤(中国专利申请号为201710413695 .X的公开文献)，或者一些研究单纯列出各种数学模型，没有过程，没有结果评估。

[0019] 发明人针对上述技术问题，经过对原因的分析，不断研究发现一种仅用于肝癌单一癌种筛查的特异甲基化检测位点组合的选取方法的技术方案，该方法包括以下步骤：

步骤1：从若干数据库中集中肝癌和其他癌症甲基化数据，其中，数据包括正常人的样本及癌症患者的样本；

步骤2：比较肝癌患者肝癌组织和正常人的正常组织中甲基化数据，以找出肝癌和正常组织之间甲基化差异位点，并将获得的甲基化差异位点进行信息注释；

步骤3：从步骤1的数据库中分析、获得肝癌特异性甲基化差异位点，具体如下：

(a)、将步骤1中所有肝癌患者的肝癌组织甲基化数据及50个患者的正常组织的甲基化数据进行对比，并根据步骤2中筛选出的甲基化差异位点，将这些样本中差异位点数据整理成一个文件；

(b)、将步骤1 中375个肝癌患者的肝癌组织甲基化数据和正常人血浆样本甲基化数据合并后，并筛选处步骤3(a)中找出的甲基化差异位点信息整理出一个文件；

(c)、将步骤1中375个肝癌患者肝癌组织甲基化数据和数据库中的其他癌症的数据集合并，根据步骤3(b)中选出的肝癌特异性甲基化差异位点，并将整理出这些位点再所有肿瘤中的数据成一个文件；

(d)、将上述或的文件集中，并选出的位点组合，从而获得肝癌早期筛查用的17个特异性甲基化位点，作为肝癌肿瘤标志物；

步骤4：对17个特异性甲基化位点建立预测模型和结果验证，进行评估。

[0020] 进一步优选的：所述步骤1中若干个数据库均为现有数据库，各现有数据库内下载大量肝癌和其他癌症甲基化数据，以及正常人的样本。

[0021] 在上述技术方案中，本发明仅针对肝癌单个癌症并通过设计不同样本类型的对比，校准个体甲基化差异，不同时期甲基化差异，不同组织甲基化差异和不同肿瘤间甲基化差异，进而筛选一组特异甲基化位点作为诊断标志物来检测肝癌，所述的检测方法敏感性能达到90%，特异性能达到97%，能在诊断过程中表现为只针对肝癌这个癌种进行检测。

[0022] 下面结合附图对本发明的具体实施方式做详细的说明。

[0023] 实施例：

如图1所示，一种仅用于肝癌单一癌种筛查的特异甲基化检测位点组合的选取方法的

技术方案，该方法包括以下步骤：

步骤1：从若干数据库中集中肝癌和其他癌症甲基化数据，其中，数据包括正常人的样本及癌症患者的样本；

具体如下：

所述步骤1中若干个数据库均为现有数据库，各现有数据库内下载大量肝癌和其他癌症甲基化数据，以及正常人的样本；各数据库中的数据均是HumanMethylation450 BeadChip(GPL13534)芯片数目，相同的数据格式才能进行对比分析，同时可以排除不同平台的偏差；

所述步骤1中所述若干数据库为：

(a)、从NIH网上进入GDC的TCGA数据库，下载带TCGA-LIHC标签的肝癌DNA甲基化、基因表达数据和临床信息注释文件；

(b)、从TCGA数据库下载其他10种癌症的DNA甲基化数据，所下载的癌症数据包括：

BLCA(409肿瘤，21正常)，BRCA(774肿瘤，82正常)，COAD(292肿瘤，38正常)，GBM(126肿瘤，2正常) ，HNSC(523肿瘤，45正常)，KIRC(316肿瘤，160正常)，LUAD(455肿瘤，32正常)，LUSC(365肿瘤，41正常)，READ(95肿瘤，7正常)和UCEC(425肿瘤 ，46正常)；

(c)、从GEO数据库下载甲基化数据集GSE69270(184名年轻芬兰人的血液)，GSE54503(66配对肿瘤和正常)，GSE89852(37配对肿瘤和正常)，GSE56588(224 肿瘤，9个肝硬化，10个正常)。

[0024] 步骤2：比较肝癌患者肝癌组织和正常人的正常组织中甲基化数据，以找出肝癌和正常组织之间甲基化差异位点，并将获得的甲基化差异位点进行信息注释；这一步过配对比较的方式比较肝癌和癌旁的甲基化情况，找出肝癌的特异性位点，能保证差异性位点不是肝和其他器官的差异，校准组织差异性：

具体如下：实现步骤2需要做一个前提，该前提就是确认差异表达基因， 表达谱差异基因准备，大量研究表明，DNA甲基化能引起染色质结构、DNA构象、DNA稳定性及DNA与蛋白质相互作用方式的改变，从而控制基因表达；肿瘤研究中表观遗传研究较多的就是通过甲基化导致基因沉默，基因启动子区域的改变，导致基因表达的减少，是目前主流的看法；因此，本方法中有考虑通过表达差异来筛选位点，不过本步骤不是必须步骤。具体实施如下：

步骤1)：于步骤1中的数据库选出同时具有肿瘤甲基化和表达谱的41对样本；

步骤2)：将上述步骤1)选出的41对样本的肝癌部位和正常部位的表达谱数据以配对方式整理成备用文件，文件行是基因名，列名是样本编号；

步骤3)：将上述步骤2)的数据文件利用软件是Bioconductor package edgeR，选用基于广义线性模型的统计方法模式鉴定表达差异基因；进而计算出每个表达基因的结果，该结果作为41个人肝癌和正常组织表达差异的衡量指标；

步骤4)：将上述步骤3)的计算出的每个基因结果，筛选出FDR小于0.05且绝对值log 2(fold change)大于1的位点，筛选出的被认为是具有差异表达的基因，其中表达差异包括肝癌组织比正常组织高表达，或者正常组织比肝癌组织高表达两种情况；

步骤5)：上述步骤4)中选取的条件不限于FDR小于0.05且绝对值log 2(fold change)大于1，是统计中表明两组数据有显著差异的条件；

综上，所述步骤2包括以下步骤：

(a)、在步骤1中选出同时具有正常组织甲基化数据和肝癌组织甲基化数据的50个患者；

(b)、根据上述选的50个患者，将每个患者正常数据和肿瘤数据整理在一个文件，过滤掉缺失较多的位点，建立一文件，该文件的行是位点名字，列是样本编号；

(c)、根据上述(b)中获得文件计算正常和肝癌组织的甲基化差异，记作p值，同时用p.adjust命令对T-test结果进行校正，记作FDR；

(d)、根据上步计算的p值和FDR，用P值小于0.05，FDR大于0.2，作为筛选条件选出符合条件的位点备用；

(e)、将(d)得出的位点，利用HumanMethylation450 BeadChip(GPL13534)对位点信息进行含位点所在基因的注释；

(f)、将注释后的差异位点所在基因与差异表达基因统计分析，找出共同基因，这些基因认为是甲基化差异导致基因表达有差异；

(h)、根据(f)的注释结果，选出位于启动子区的位点(TSS1500|TSS200)备用；选出位于启动子区的位点(TSS1500|TSS200)备用，表观遗传研究较多的就是通过甲基化导致基因沉默，基因启动子区域的改变，导致基因表达的减少，因此选择启动子区域具有生物学意义，当然并不是所有表观遗传改变都是通过启动子发挥作用，因此本步骤可选，但优选方案是选取所有位置有差异的甲基化位点。

[0025] 步骤3：从步骤1的数据库中分析、获得肝癌特异性甲基化差异位点，具体如下：

(a)、将步骤1中所有肝癌患者的肝癌组织甲基化数据及50个患者的正常组织的甲基化数据进行对比，并根据步骤2中筛选出的甲基化差异位点，将这些样本中差异位点数据整理成一个文件；这一步目的是为了在跟多样本中验证找到的位点是在肝癌患者中都具有的，排除有些位点是在部分人群或者人种中才有，其具体实施办法参考步骤2；

(b)、将步骤1 中375个肝癌患者的肝癌组织甲基化数据和正常人血浆样本甲基化数据合并后，并筛选处步骤3(a)中找出的甲基化差异位点信息整理出一个文件；这一步是为了做保证筛选的位点能用于液体活检，通过比较这些位点和在肝癌和血液中的数据，过滤掉那些在血浆中差异不大的点；如果液体活检是通过其他体液，只需将数据换成具体体液实验数据，原理不变；如果是组织活检的标记物，此步骤可选；具体实施办法参考步骤2；

(c)、将步骤1中375个肝癌患者肝癌组织甲基化数据和数据库中的其他癌症的数据集合并，根据步骤3(b)中选出的肝癌特异性甲基化差异位点，并将整理出这些位点再所有肿瘤中的数据成一个文件；这一步的目的是想找出肝癌特异性位点，因为不同肿瘤发生机制有很多起因相同，所以肝癌也会有很多标志物适用于其他肿瘤，过滤掉在和其他肿瘤差异不大的点就剩下肝癌特异性标记物；具体实施办法参考步骤2；

(d)、将上述或的文件集中，并选出的位点组合，从而获得肝癌早期筛查用的17个特异性甲基化位点，作为肝癌肿瘤标志物；

具体的说：所述步骤3中(d)的具体计算方法如下：

(1)、计算肝癌的均值和其他所有肿瘤数据的均值，将肝癌均值减肿瘤均值绝对值高于0.1的位点过滤出来作为肝癌特异性早筛位点，即(chr2:166650805、chr2:232260305、chr2:9144605、chr3:123167770、chr3:101497876、chr3:101497857、chr3:101497980、chr3:101497982、chr6:116691863、chr8:102504447、chr8:102504482、chr8:102504501、

chr10:21463485、chr11:67350976、chr11:66624853、chr17:4981610、chr20:44540794)；

(2)、上述(1)选出的位点组合，每个位点所在CpG岛内相邻位甲基化点具有同样，包括但不限于450k芯片中找不出的位点，因为每个CpG岛的甲基化位点具有一致性，故，所述位点组合拓展至位点所在CpG岛的组合，即(chr10:21462128-21463808、chr11:66623620-66626614、chr11:67350928-67351953、chr17:4981357-4981979、chr2:166649909-166650966、chr2:232260100-232261134、chr2:9143127-9144630、chr20:44540445-44540957、chr3:101497830-101498648、chr3:123166218-123168567、chr6:116691827-116692868、chr8:102504478-102504841)。

[0026] 步骤4:对17个特异性甲基化位点建立预测模型和结果验证，进行评估；

具体的说:对所得的生物标志物建立预测模型和结果验证，将找出的17个甲基化位点组合作为肿瘤标志物，需要对其的进行评估，主要通过肝癌和正常样本混合评价其检出率，敏感性和特异性，进而所述步骤4的具体步骤如下：

(a)、将TCGA-LIHC的具有正常组织甲基化数据和肝癌组织数据的50个患者的甲基化数据集作为训练集，使用weka软件，输入17个候选位点建立预测模型，所选模型为J48模型、DecisionStump模型、LMT模型、REPtree模型、RandomForest模型、NaiveBayes模型、logistic模型、MultilayerPerceptron模型；

(b)、将四个其他独立数据集(GSE54503、GSE89852、GSE56588)作为测试集，使用weka软件，使用(a)得到的模型来测试模型效果；

(c)、记录模型在得到的模型效果，包括敏感性、特异性以评估预测模型的准确性，以及选择效果最优的模型。

[0027] 最优模型效果：

数据集	敏感性	特异性
GSE54503	95%	98%
GSE56588	90%	100%
GSe89852	95%	97%

以下为本技术方案及本实施例中，具体专有名词说明：

1、表达谱:指通过构建处于某一特定状态下的细胞或组织的非偏性cDNA文库，大规模cDNA测序，收集cDNA序列片段、定性、定量分析其mRNA群体组成，从而描绘该特定细胞或组织在特定状态下的基因表达种类和丰度信息，这样编制成的数据表就称为基因表达谱。

[0028] 2、甲基化:本专利提到的甲基化都指DNA甲基化(DNA methylation)为DNA化学修饰的一种形式，能够在不改变DNA序列的前提下，改变遗传表现。所谓DNA甲基化是指在DNA甲基化转移酶的作用下，在基因组CpG二核苷酸的胞嘧啶5'碳位共价键结合一个甲基基团。DNA甲基化能引起染色质结构、DNA构象、DNA稳定性及DNA与蛋白质相互作用方式的改变，从而控制基因表达。

[0029] 3、肿瘤:肝细胞癌(hepatocellular carcinoma,HCC)是一种高死亡率的原发性肝癌。它是一种全球范围最常见的恶性肿瘤。

[0030] 4、illunima HumanMethylation450K BeadChip:用于DNA甲基化分析的一种芯片，可以检测出DNA甲基化程度，由illumina公司研发生产。

[0031] 7、CpG岛:CpG岛(CpG islands)是指DNA上一个区域，此区域含有大量相联的胞嘧

啶(C)、鸟嘌呤(G),以及使两者相连的磷酸酯键(p)。

[0032] 8、敏感性:又称真阳性率,指诊断方法对疾病的敏感程度或识别能力。敏感性越高,漏诊概率越低,计算公式为:TP/TP+FN=诊出患病人数/诊出患病人数+漏诊人数。

[0033] 9、特异性:又称真阴性率,指诊断方法对疾病的误诊率,特异性越高,误诊率越低。计算公式为:=TN/TN+FP=诊出非患病人数/诊出非患病人数+误诊人数。

[0034] 10、BLCA:Bladder Urothelial Carcinoma,膀胱尿路上皮癌。

[0035] 11、BRCA:Breast invasive carcinoma,乳腺侵袭性导管癌。

[0036] 12、COAD:Colon adenocarcinoma,结直肠腺癌。

[0037] 13、GBM:Glioblastoma multiforme,胶质母细胞瘤 。

[0038] 14、HNSC:Head and Neck squamous cell carcinoma,头颈部鳞状细胞癌 。

[0039] 15、KIRC:Kidney renal clear cell carcinoma,透明细胞肾癌。

[0040] 16、LUAD:Lung adenocarcinoma,肺腺癌。

[0041] 17、LUSC:Lung squamous cell carcinoma,鳞状细胞肺癌。

[0042] 18、READ:Rectum adenocarcinoma,直肠腺癌。

[0043] 19、UCEC:Uterine Corpus Endometrial Carcinoma,子宫内膜癌

20、FDR:FDR(*false discovery rate*),是统计学中常见的一个名词,翻译为伪发现率,其意义为是 错误拒绝(*拒绝真的(原)假设*)的个数占所有被拒绝的原假设个数的比例的期望值。

[0044] 21、weka:Weka的全名是怀卡托智能分析环境(Waikato Environment for Knowledge Analysis),是一款免费的,非商业化(与之对应的是SPSS公司商业数据挖掘产品--Clementine)的,基于JAVA环境下开源的机器学习(machine learning)以及数据挖掘(data mining)软件。

[0045] 需要说明的是:

1、筛选位点的4步顺序可以调换;

2、步骤1中数据库可以用其他数据集,或者本数据集的子集;

3、保护位点以在人类参考基因组(hg19)位置方式给出,其他展示方式包括不同人类版本基因组位置;位点前后序列,位点再其他数据库的命名(如450k芯片),位点所在CpG名字。

[0046] 本发明虽然以较佳实施例公开如上,但其并不是用来限定本发明,任何本领域技术人员在不脱离本发明的精神和范围内,都可以做出可能的变动和修改,因此本发明的保护范围应当以本发明权利要求所界定的范围为准。

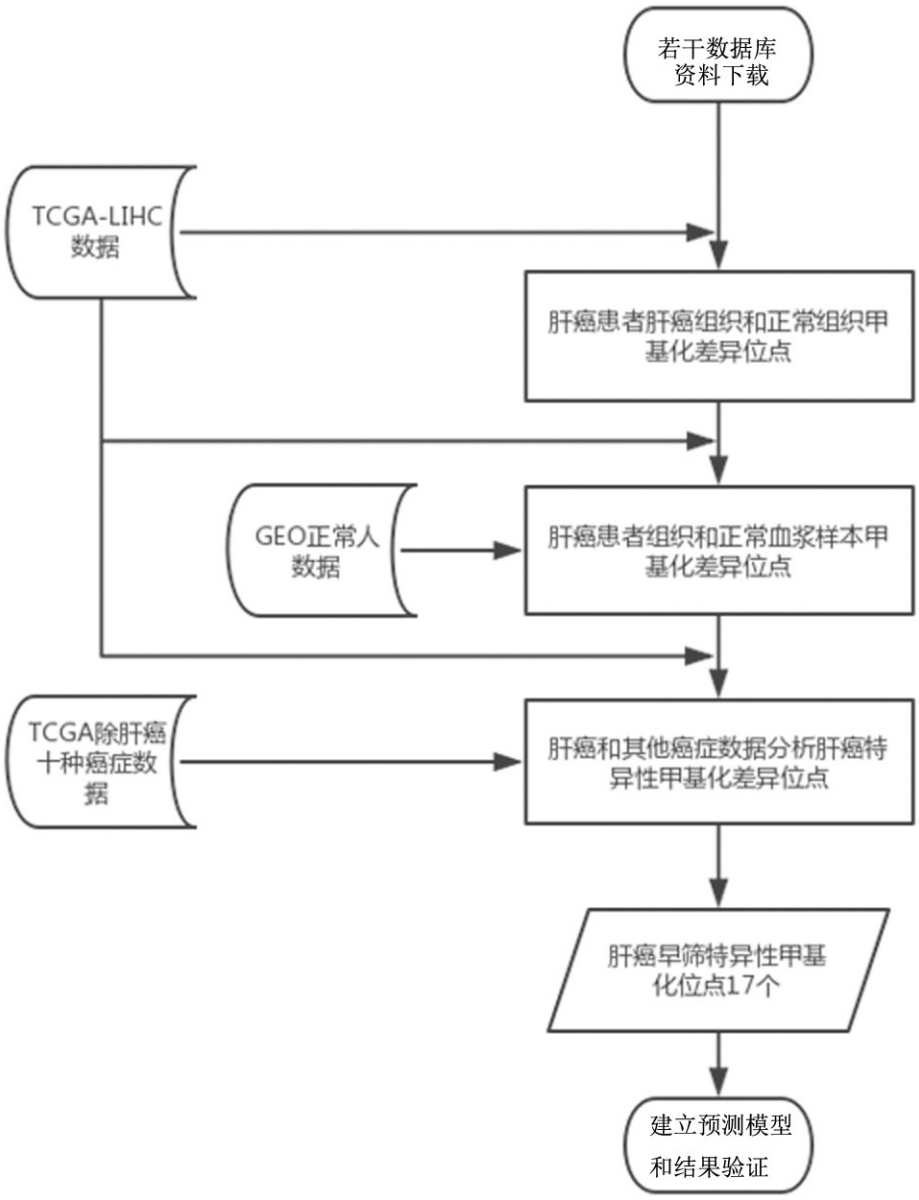

若干数据库
资料下载
TCGA-LIHC
数据
肝癌患者肝癌组织和正常组织甲
基化差异位点
GEO正常人
数据
肝癌患者组织和正常血浆样本甲
基化差异位点
TCGA除肝癌
十种癌症数
据
肝癌和其他癌症数据分析肝癌特
异性甲基化差异位点
肝癌早筛特异性甲基
化位点17个
建立预测模型
和结果验证

图1