(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 106295246 A
(43)申请公布日 2017.01.04

(21)申请号 201610639453.8

(22)申请日 2016.08.07

(71)申请人 吉林大学
地址 130012 吉林省长春市前进大街2699号

(72)发明人 李瑛 何叶 梁艳春 张禹 苏静 韩斯禹

(74)专利代理机构 长春市四环专利事务所(普通合伙) 22103
代理人 鞠传龙

(51)Int.Cl.
G06F 19/18(2011.01)
G06F 19/20(2011.01)
G06F 19/24(2011.01)

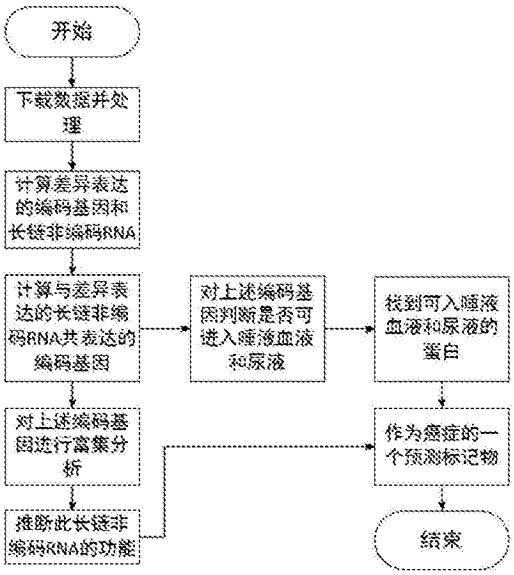
权利要求书3页 说明书8页 附图9页

(54)发明名称

找到与肿瘤相关的lncRNA并预测其功能

(57)摘要

本文是找到与肿瘤相关的lncRNA并预测其功能。我们把lncRNA在肿瘤中的差异表达作为诊断的参考，找出lncRNA与肿瘤之间的关系。第一步，从GEO数据库中下载数据，对其处理后得到外显子和部分lncRNA的表达数据。第二步，对处理好的表达数据进行差异表达分析。第三步，对差异表达的lncRNA分析与它共表达且差异的编码基因和lncRNA。第四步，将编码基因进行探针平台注释。第五步，对差异表达lncRNA进一步筛选，选出最显著差异的lncRNA。第六步，进行富集分析，得到GO BP过程和pathway。通过编码基因所涉及的生物过程来推测lncRNA的功能。第七步，对上一步得到的公共编码基因，分析是否可入血液、唾液和尿液，对可以进入的基因进行分析，这些基因和lncRNA就可以作为癌症的一个潜在的预测标记。



CN 106295246 A

1.找到与肿瘤相关的lncRNA并预测其功能，相对于蛋白编码序列以及小分子RNA，lncRNA的研究还仅仅只是处于起步阶段，目前并不能仅根据序列或者结构来推测它们的功能，它们在基因组上相对于蛋白编码基因的位置，对于推测lncRNA的功能有很大帮助；其特征是：在肿瘤细胞中，某些特定的lncRNA的表达水平会发生改变，这种表达水平的变化能够作为癌症诊断的标志物，通过找到与变化lncRNA共表达且差异的编码基因，从而推测lncRNA的功能，其特征在于，所述方法分为四个阶段进行：

第1阶段，计算差异表达基因；

第2阶段，计算相似性矩阵；

第3阶段，判断既与差异表达的lncRNA共表达又差异的编码基因是否可进入唾液、血液和尿液；

第4阶段，找出与肿瘤显著相关的lncRNA作为biomarkers并推断其功能；

其中在第1阶段中，对处理过的癌症的片子进行差异表达分析，通过fold change和p-value两个指标来进行筛选；在第2阶段中，计算上述差异表达的lncRNA与数据中的所有基因间的相似性矩阵并画出共表达网络图；在第3阶段中，判断既差异表达又与差异表达的lncRNA共表达的编码基因是否可在唾液、血液和尿液中找到对应物；在第4阶段中，将第3阶段得到的编码基因放到DAVID网站上分析它们的GO BP和pathway来进一步分析lncRNA的功能；对差异表达的lncRNA进一步筛选，筛选完留下来的lncRNA及其共表达又差异的编码基因可作为癌症的潜在标记物。

2.根据权利要求1所述的预测功能，所述第1阶段的具体流程包括：

步骤1.1：对处理过的exon数据计算fold change值和p-value值，其中fold change值用均值来计算，首先判断数据是否符合正态分布，若符合，则用tumor样本的2的幂次方的均值除以normal样本的2的幂次方的均值；若不符合，则直接用数据中tumor样本的均值除以normal样本的均值；

步骤1.2：对处理过的exon数据计算fold change值和p-value值，其中p-value值用wilcoxon秩和检验来进行计算，wilcoxon秩和检验是基于样本数据秩和，也是把数据样本分成tumor和normal两部分来计算，在wilcox检验里选择p-value这项；

步骤1.3：对步骤1.1中计算的fold change值取1.5作为阈值；对步骤1.2中计算的p-value值取0.01作为阈值，把p-value值小于0.01且fold change值大于1.5的归为上调的基因；

把p-value值小于0.01且fold change值小于1/1.5的归为下调的基因；

步骤1.4：对筛选出来的基因在原来的数据中找到对应的样本数据另存成一个差异表达的基因文件；

步骤1.5：统计癌症的各个片子的差异表达基因数并制成表格，包括：上调的编码基因数与lncRNA数、下调的编码基因数与lncRNA数。

3.根据权利要求1所述的预测功能，所述第2阶段的具体流程包括：

步骤2.1：通过阶段1所找到的差异表达的基因，求出差异表达的lncRNA与其他基因的相似性系数；

步骤2.2：用pearson方法计算相似性矩阵，其公式如下：

$$\rho_{X,Y}=\frac{\mathrm{cov}(X,Y)}{\sigma_X\sigma_Y}=\frac{E[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X\sigma_Y}$$

步骤2.3:得到一个用pearson方法求得的一个矩阵,行名为数据集的所有基因,列名为癌症中差异表达的lncRNA,其中相似系数数据均取绝对值;

步骤2.4:用spearman方法计算相似性矩阵;其公式如下:

$$\rho=\frac{\sum_i(x_i-\overline{x})(y_i-\overline{y})}{\sqrt{\sum_i(x_i-\overline{x})^2\sum_i(y_i-\overline{y})^2}}$$

步骤2.5:得到一个用spearman方法求得的一个矩阵,行名为数据集的所有基因,列名为癌症中差异表达的lncRNA,其中相似系数数据均取绝对值;

步骤2.6:将阈值取到0.7,筛选掉小于0.7的相似系数,大于0.7的相似系数认为两者之间是共表达的;

步骤2.7:通过步骤2.3和步骤2.6得到与差异表达的lncRNA共表达的编码基因和lncRNA,并通过cytoscape画出共表达网络;

步骤2.8:通过步骤2.5和步骤2.6得到与差异表达的lncRNA共表达的编码基因和lncRNA,并通过cytoscape画出共表达网络。

4.根据权利要求1所述的预测功能,所述第3阶段的具体流程包括:

步骤3.1:对步骤2.7得到的编码基因与阶段1得到的差异表达的编码基因取交集,观察既共表达又差异表达的编码基因有哪些;

步骤3.2:对步骤3.1得到的交集的编码基因通过探针注释的方法,转化为平台上的genesymbol形式;

步骤3.3:对步骤3.2得到的gene symbol形式的基因判断是否可进入唾液并在唾液中找到其对应物;

步骤3.4:对步骤3.2得到的gene symbol形式的基因判断是否可进入血液并在血液中找到其对应物;

步骤3.5:对步骤3.2得到的gene symbol形式的基因判断是否可进入尿液并在尿液中找到其对应物;

步骤3.6:将步骤3.3至步骤3.5得到的结果制成一张表格。

5.根据权利要求1所述的预测功能,所述第4阶段的具体流程包括:

步骤4.1:对步骤3.2得到的gene symbol形式的基因放到DAVID网站上进行富集分析,分析它们的GO BP和pathway;

步骤4.2:将基因所涉及的GO BP和pathway按p-value取前十个,画出直方图,观察都参与了哪些生物过程;

步骤4.3:通过步骤4.2所涉及的过程来推测差异表达的lncRNA所涉及的生物过程得到其功能;

步骤4.4:对于阶段1得到的差异表达的lncRNA,我们进一步分析更显著差异的lncRNA,用fisher exacttest方法来进行筛选,计算几个算法需要的参数,包括:

与lncRNA共表达的基因集合(n);

与tumor相关的差异表达的基因集合(x)；

求出上述两步的交集的基因集合(y)；

整个数据集的基因数目(N)；

步骤4.5:计算完上述参数后,用fisher exact test方法计算p-value值,得到每个差异表达基因所对应的p-value值的文档；

步骤4.6:对步骤4.5中的文档,筛选掉p-value值大于0.05的lncRNA；

步骤4.7:把步骤4.6中留下的lncRNA作为更显著表达的lncRNA,将癌症的各个片子所留下的lncRNA取交集,将其作为潜在的生物标记物；

步骤4.8:整理步骤4.7中取交集的lncRNA和对应的与其共表达又差异表达的编码基因,将其作为癌症的biomarkers。

找到与肿瘤相关的lncRNA并预测其功能

技术领域

[0001] 本发明涉及与肿瘤和基因，更具体的说，涉及lncRNA和肿瘤间的关系。

背景技术

[0002] 随着生物学技术的发展，癌症的诊断和治疗是人们一直在努力去攻克的难题。关于基因转录物在癌症病人中的表达数据也做了很多，但是并没有找出这些数据所蕴含的意义，因此对这些数据进行科学的分析非常重要，而lncRNA是基因很重要的一类转录产物，分析它在病人和非病人间的差异表达，将会对癌症的治疗与诊断提供重要的参考信息。最初人们认为是编码基因导致了癌症，后来发现lncRNA虽然不直接参与编码蛋白质，但是可以调控其他基因的表达，与癌症也是密切相关的。我们可以通过对lncRNA的表达数据进行分析处理，了解哪些长链非编码RNA在癌症中会起到关键作用。

[0003] ncFANs是一个重新注释的算法，可以修正被错误标注的探针集合，发现其中隐含的lncRNA的表达信息。ncFANs进行规范化处理时使用MAS5.0方法，MAS5.0方法不受读入数据集的背景影响，无论处于什么样的数据集中，特定样本的值是不变的。并且它可以将数据进行分批处理，然后合并在一起就可以了，非常方便。

[0004] 在判断差异表达的编码基因和lncRNA时，采用倍数变化和秩和检验来将不符合要求的筛选掉。倍数变化的数学表达如下公式所示：

[0005] FC=mean(tumor)/mean(normal)

[0006] p-value通过wilcoxon秩和检验来计算，这会使对于有些数据集找到的差异表达基因较多。建立编码基因与非编码基因的共表达网络时，利用pearson方法和spearman方法进行选择，最后用cytoscape画出共表达网络。这些是实现找到差异lncRNA和预测lncRNA功能的关键技术。

[0007] 国内在lncRNA方面已有了一些研究成果。2002年Okazaki Y,Furuno M,Kasukawa T(Okazaki Y,Furuno M,Kasukawa T,et al.Analysis of the mouse transcriptome based on functional annotation of 60,770full-length cDNAs[J].Nature,2002,420(6915):563-573)等在对小鼠全长互补DNA文库的大规模测序过程中发现了一类转录物，即长链非编码RNA。长链非编码RNA是一类转录本长度超过200个核苷酸的功能性RNA分子，它们缺乏编码蛋白的能力，位于细胞核或细胞质内，以RNA形式在多种层面上(如表观遗传学、转录调控及转录后调控等)调控基因的表达水平。2005年，Bentwich I,Avniel A,Karov Y等人发现微RNA类及小核仁RNA类等的众多小型RNA类都显现出了跨多物种的保守性。与之相反，大多数长链非编码RNA则保守性不强，这一点常被引用为其不具备功能的证据。但是，尽管长链非编码RNA总体来说保守性较低，但仍然可以看见许多长链非编码RNA具有较强的保守元件。1992年，Lukiw等人在衰老和人类神经系统疾病的研究中发现了长链非编码RNA丰度的变化。接下来，对肿瘤细胞和正常细胞中非编码RNA的表达分析显示了它们表达量的变化，许多相关的研究都证实了长链非编码RNA在疾病状态下的异常表达，但是对于它们在疾病的病因当中的贡献依然知之甚少。2004年，Reis等人报道了内含子反义非编码RNA在前

列腺瘤分化程度中具有很强的表达相关性。2006年,Fu等人发现原本被认为大量表达的非编码RNAMALAT1(也被称为NEAT2)会在早期非小细胞肺癌的新陈代谢过程中出现正调节,而它的过表达量与病人的存活率息息相关。2007年,Calin等人通过全基因组的分析发现在人类不同癌症状态下,许多转录的非编码RNA的超保守区域存在很大区别。2011年,Bellucci M(Bellucci M,Agostini F,Masin M,Tartaglia GG.Predicting protein associations with long noncoding RNAs.Nat Methods,2011,8(6):444-445)等人研发了catRAPID在线算法,可以用于预测RNA与蛋白质的相互作用。同年3月,Qi Liao等人研发了ncFANs在线长链非编码RNA的功能注释算法。2012年,Pascal Gellert等人研发了长链非编码RNA的详细功能注释工具Noncoder。同年,李瑛(Li,Ying,Ming Duan,and Yanchun Liang."Multi-scale RNA comparison based on RNA triple vector curve representation."BMC bioinformatics 13.1(2012):280)等人提出了一种基于RNA三矢量曲线表示的比较新颖的小波变换方法——多尺度RNA比较法。这些方法并没有明确知道lncRNA的功能。

[0008] 综上可以看出,目前研究lncRNA已经成为一种趋势,但是并没有能明确知道lncRNA与疾病的关系以及推测其功能的方法。而通过找与差异表达的lncRNA共表达的编码基因是寻找lncRNA与疾病之间关系的关键步骤,然后通过对这些共表达的编码基因分析来进一步推测lncRNA的功能。其中求解共表达相似性矩阵时,时间较长,求出的相似性矩阵较大,对实验的效率有所影响。

发明内容

[0009] 根据上面所描述,本发明的研究重点在于寻找差异表达的编码基因和lncRNA以及寻找与差异表达的lncRNA共表达的编码基因。

[0010] 本发明通过计算倍数变化值和秩和检验的p-value值来寻找肿瘤中的差异表达基因;通过pearson和spearman方法来计算相似性矩阵获得共表达网络;然后判断既共表达又差异表达的编码基因是否可入唾液、血液和尿液,从而找到癌症的潜在预测标记物并推测其功能。上述过程可分为四个阶段进行:

[0011] 第1阶段,计算差异表达基因。

[0012] 第2阶段,计算相似性矩阵。

[0013] 第3阶段,判断既与差异表达的lncRNA共表达又差异的编码基因是否可进入唾液、血液和尿液。

[0014] 第4阶段,找出与肿瘤显著相关的lncRNA作为biomarkers并推断其功能。

[0015] 进一步,所述第1阶段的具体流程包括:

[0016] 步骤1.1:对处理过的exon数据计算fold change值和p-value值,其中fold change值用均值来计算。首先判断数据是否符合正态分布,若符合,则用tumor样本的2的幂次方的均值除以normal样本的2的幂次方的均值;若不符合,则直接用数据中tumor样本的均值除以normal样本的均值。

[0017] 步骤1.2:对处理过的exon数据计算fold change值和p-value值,其中p-value值用wilcoxon秩和检验来进行计算,wilcoxon秩和检验是基于样本数据秩和,也是把数据样本分成tumor和normal两部分来计算,在wilcox检验里选择p-value这项。

[0018] 步骤1.3:对步骤1.1中计算的fold change值取1.5作为阈值;对步骤1.2中计算的

p-value值取0.01作为阈值，把p-value值小于0.01且fold change值大于1.5的归为上调的基因；把p-value值小于0.01且fold change值小于1/1.5的归为下调的基因。

[0019] 步骤1.4:对筛选出来的基因在原来的数据中找到对应的样本数据另存成一个差异表达的基因文件。

[0020] 步骤1.5:统计癌症的各个片子的差异表达基因数并制成表格，包括：上调的编码基因数与lncRNA数、下调的编码基因数与lncRNA数。

[0021] 进一步，所述第2阶段的具体流程包括：

[0022] 步骤2.1:通过阶段1所找到的差异表达的基因，求出差异表达的lncRNA与其他基因的相似性系数。

[0023] 步骤2.2:用pearson方法计算相似性矩阵。其公式如下：

[0024] $$\rho_{X,Y}=\frac{\operatorname{cov}(X,Y)}{\sigma_X\sigma_Y}=\frac{E[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X\sigma_Y}$$

[0025] 步骤2.3:得到一个用pearson方法求得的一个矩阵，行名为数据集的所有基因，列名为癌症中差异表达的lncRNA。其中相似系数数据均取绝对值。

[0026] 步骤2.4:用spearman方法计算相似性矩阵。其公式如下：

[0027] $$\rho=\frac{\sum_i(x_i-\overline{x})(y_i-\overline{y})}{\sqrt{\sum_i(x_i-\overline{x})^2\sum_i(y_i-\overline{y})^2}}$$

[0028] 步骤2.5:得到一个用spearman方法求得的一个矩阵，行名为数据集的所有基因，列名为癌症中差异表达的lncRNA。其中相似系数数据均取绝对值。

[0029] 步骤2.6:将阈值取到0.7，筛选掉小于0.7的相似系数，大于0.7的相似系数认为两者之间是共表达的。

[0030] 步骤2.7:通过步骤2.3和步骤2.6得到与差异表达的lncRNA共表达的编码基因和lncRNA，并通过cytoscape画出共表达网络。

[0031] 步骤2.8:通过步骤2.5和步骤2.6得到与差异表达的lncRNA共表达的编码基因和lncRNA，并通过cytoscape画出共表达网络。

[0032] 进一步，所述第3阶段的具体流程包括：

[0033] 步骤3.1:对步骤2.7得到的编码基因与阶段1得到的差异表达的编码基因取交集，观察既共表达又差异表达的编码基因有哪些。

[0034] 步骤3.2:对步骤3.1得到的交集的编码基因通过探针注释的方法，转化为平台上的gene symbol形式。

[0035] 步骤3.3:对步骤3.2得到的gene symbol形式的基因判断是否可进入唾液并在唾液中找到其对应物。

[0036] 步骤3.4:对步骤3.2得到的gene symbol形式的基因判断是否可进入血液并在血液中找到其对应物。

[0037] 步骤3.5:对步骤3.2得到的gene symbol形式的基因判断是否可进入尿液并在尿液中找到其对应物。

[0038] 步骤3.6:将步骤3.3至步骤3.5得到的结果制成一张表格。

[0039] 进一步,所述第4阶段的具体流程包括:

[0040] 步骤4.1:对步骤3.2得到的gene symbol形式的基因放到DAVID网站上进行富集分析。分析它们的GO BP和pathway。

[0041] 步骤4.2:将基因所涉及的GO BP和pathway按p-value取前十个,画出直方图。观察都参与了哪些生物过程。

[0042] 步骤4.3:通过步骤4.2所涉及的过程来推测差异表达的lncRNA所涉及的生物过程得到其功能。

[0043] 步骤4.4:对于阶段1得到的差异表达的lncRNA,我们进一步分析更显著差异的lncRNA。用fisher exact test方法来进行筛选。计算几个算法需要的参数,包括:

[0044] 与lncRNA共表达的基因集合(n)。

[0045] 与tumor相关的差异表达的基因集合(x)。

[0046] 求出上述两步的交集的基因集合(y)。

[0047] 整个数据集的基因数目(N)。

[0048] 步骤4.5:计算完上述参数后,用fisher exact test方法计算p-value值。得到每个差异表达基因所对应的p-value值的文档。

[0049] 步骤4.6:对步骤4.5中的文档,筛选掉p-value值大于0.05的lncRNA。

[0050] 步骤4.7:把步骤4.6中留下的lncRNA作为更显著表达的lncRNA。将癌症的各个片子所留下的lncRNA取交集,将其作为潜在的生物标记物。

[0051] 步骤4.8:整理步骤4.7中取交集的lncRNA和对应的与其共表达又差异表达的编码基因,将其作为癌症的biomarkers。

[0052] 通过本发明,可以推测出lncRNA的一些功能,本发明流程简单,操作方便。主要是计算出差异表达的基因并求相似性矩阵,这两个阶段运用的方法比较简单,便于理解,并且通过R语言中的函数调用,一些结果很容易得出。

附图说明

[0053] 图1为本发明的流程图。

[0054] 图2为上调表达的编码基因和lncRNA的韦恩图。

[0055] 图3为下调表达的编码基因和lncRNA的韦恩图。

[0056] 图4为上调表达的lncRNA的韦恩图。

[0057] 图5为下调表达的lncRNA的韦恩图。

[0058] 图6为求取差异表达基因的流程图。

[0059] 图7为与差异表达的lncRNA共表达的编码基因的共表达网络。

[0060] 图8为与差异表达的lncRNA共表达的lncRNA的共表达网络。

[0061] 图9为求共表达基因的流程图。

具体实施方式

[0062] 以下结合附图对本发明的流程进行描述,所举实例只用于解释本发明,并非用于限定本发明的范围。

[0063] 本发明是要找到与肿瘤相关的lncRNA并预测其功能,包括:第1阶段,对处理过的

癌症的片子进行差异表达分析，通过fold change和p-value两个指标来进行筛选。第2阶段，计算上述差异表达的lncRNA与数据中的所有基因间的相似性矩阵并画出共表达网络图。第3阶段，判断既差异表达又与差异表达的lncRNA共表达的编码基因是否可在唾液、血液和尿液中找到对应物。第4阶段，将第3阶段得到的编码基因放到DAVID网站上分析它们的GOBP和pathway来进一步分析lncRNA的功能。对差异表达的lncRNA进一步筛选，筛选完留下来的lncRNA及其共表达又差异的编码基因可作为癌症的潜在标记物。

[0064] 一种较为具体的实施例为，本发明从胃癌入手，找到与胃癌显著相关的lncRNA并预测其功能。首先在GEO网站上下载exon array的数据，并用ncFANs处理成需要用的表达谱文件，该数据包括18921个编码基因和1392个lncRNA的表达值。通过计算胃癌中三套片子数据的fold change值和p-value值，选取fold change大于1.5且p-value小于0.01的作为上调表达的；选取fold change小于1/1.5且p-value小于0.01的作为下调表达的。将筛选出来的差异表达基因进行整理。接下来建立共表达网络，对选择出来的差异表达的lncRNA与数据中的所有基因计算相似性矩阵，选取大于0.7的作为与其共表达的基因，并用cytoscape画出共表达网络图。然后把既与差异表达的lncRNA共表达又差异表达的编码基因通过探针注释映射成genesymbol的形式放到DAVID网站上做富集分析推测lncRNA的功能，并判断这些基因是否可进入唾液、血液和尿液，最后再对lncRNA进行筛选，选出更显著差异的lncRNA。将其与可进入体液的基因作为胃癌的生物标记物。

[0065] 以下结合附图来予以说明。图1示出了本发明的基本流程图。具体包括：

[0066] 第1阶段，求出胃癌三套片子的差异表达基因。

[0067] 第2阶段，分别求三套片子的与差异表达的lncRNA共表达的编码基因并每套片子画出lncRNA与lncRNA共表达的网络图以及lncRNA与编码基因共表达的网络图。

[0068] 第3阶段，分别判断胃癌三套片子中既差异表达又与差异表达的lncRNA共表达的编码基因是否可在唾液、血液和尿液中找到对应物并制成一张图表。

[0069] 第4阶段，将第3阶段得到的编码基因通过探针注释找到对应的gene symbl放到DAVID网站上分析它们的GO BP和pathway来进一步分析lncRNA的功能。对差异表达的lncRNA进一步筛选，筛选完留下来的lncRNA及其共表达又差异的编码基因可作为癌症的潜在标记物。

[0070] 首先，说明本发明的第1阶段，差异表达基因的确定。

[0071] 图2示出了胃癌三套片子差异表达的上调的基因数的韦恩图。

[0072] 图3示出了胃癌三套片子差异表达的下调的基因数的韦恩图。

[0073] 图4示出了胃癌三套片子差异表达的上调的lncRNA数目的韦恩图。

[0074] 图5示出了胃癌三套片子差异表达的下调的lncRNA数目的韦恩图。

[0075] 由求差异表达基因的方法可知，需要先选取两个衡量指标：倍数变化和p-value值。分别对这两个指标进行阈值的选取，留下表达较为差异的基因。可根据下面两个公式来计算倍数变化值和p-value值：

[0076] 若数据不符合正态分布，则FC＝mean(tumor)/mean(normal)

[0077] 若数据符合正态分布，则FC＝mean(2^tumor)/mean(2^normal)

[0078] p-value直接在调用wilcox.test函数。

[0079] 由此对胃癌的每套片子的数据都计算每一行的fold change值和p-value值，再进

行筛选，就能确定差异表达的编码基因和lncRNA并统计它们的数目。此过程只需要计算这两个指标，操作简单。具体流程如图6所示，包括如下步骤：

[0080] 步骤1.1:计算差异表达的基因要计算两个参数，一个是fold change;一个是p-value。

[0081] 步骤1.2:计算fold change值用tumor样本的均值除以normal样本的均值。计算p-value值用wilcoxon秩和检验直接调用函数即可。

[0082] 步骤1.3:对这两个指标进行筛选。选择fold change值大于1.5且p-value值小于0.01的作为上调表达基因;选择fold change值小于1/1.5且p-value值小于0.01的作为下调表达基因。

[0083] 步骤1.4:统计胃癌三套片子的上调和下调表达基因数。

[0084] 完成差异表达基因的选取后，本发明进入第2阶段。分别求胃癌三套片子的与差异表达的lncRNA共表达的编码基因，通过探针注释的方法将他们转换成可用的gene symbol的形式。并对每套片子画出lncRNA与lncRNA共表达的网络图以及lncRNA与编码基因对应的gene symbol形式的共表达的网络图。

[0085] 图7示出了胃癌中一套片子与差异表达的lncRNA共表达的编码基因对应的gene symbol形式的共表达网络。

[0086] 图8示出了胃癌中一套片子与差异表达的lncRNA共表达的lncRNA的共表达网络。

[0087] 对共表达基因的选定的具体流程如图9所示，包括如下步骤：

[0088] 步骤2.1:用pearson(spearman方法与这个过程一样)方法去计算相似性矩阵。得到的相似性矩阵行名是数据中的所有基因，列名是这套关于胃癌的片子的差异表达的lncRNA。

[0089] 步骤2.2:对相似性矩阵进行筛选，选取大于0.7所对应的基因留下，将不符合的筛选掉。

[0090] 步骤2.3:对留下的可看做与差异表达lncRNA共表达的。对与差异表达的lncRNA共表达的编码基因通过探针注释的方法映射成gene symbol的形式。

[0091] 步骤2.4:对上一步得到的gene symbol形式与差异表达的lncRNA用cytoscape画出共表达网络图。

[0092] 步骤2.5:对步骤2.3得到的与差异表达的lncRNA共表达的lncRNA和这些差异表达的lncRNA用cytoscape画出共表达网络图。

[0093] 在进行阈值的选取过程中，要根据实验数据的情况而定，如果选取的基因过多，可适当放大阈值;如果选取的基因过少，可适当缩小阈值。在这里，我们也应用的spearman方法，比较这两种方法哪种更适用。

[0094] 下面，具体对上述步骤中涉及的方法予以详细的说明。

[0095] 1、pearson相关系数

[0096] 皮尔逊积矩相关系数被广泛用来度量两个变量A和B之间的相关(线性相关)性，其值介于-1和+1之间。设两个变量X和Y，它们之间的皮尔逊相关系数一般会被定义为它俩之间的协方差与标准差的商，其公式如下：

[0097] $$\rho_{X,Y} = \frac{\mathrm{cov}(X,Y)}{\sigma_X \sigma_Y} = \frac{E[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X \sigma_Y}$$

[0098] 其中,σ代表了样本的标准差。

[0099] 2、spearman相关系数

[0100] 斯皮尔曼等级相关系数与Pearson相关系数类似,它也是用来衡量两个变量或两组变量之间的相关性,它的取值范围也是在+1和-1之间(闭区间)。

[0101] 设两组变量X和Y各有n个样本,先将它们的原始数据Xi以及Yi都转换成等级数据xi,yi,然后相关系数ρ根据下面这个公式计算:

[0102] $$\rho = \frac{\sum_i (x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_i (x_i - \overline{x})^2 \sum_i (y_i - \overline{y})^2}}$$

[0103] 完成共表达分析之后,本发明进入第3阶段。分别判断胃癌这三套片子中既差异表达又与差异表达的lncRNA共表达的编码基因是否可在唾液、血液和尿液中找到对应物。

[0104] 具体的步骤如下:

[0105] 步骤3.1:将第2阶段中得到的与差异表达lncRNA共表达的编码基因与第1阶段得到的在胃癌中差异表达的编码基因取个交集。

[0106] 步骤3.2:把步骤3.1得到的交集转换成gene symbol的形式。

[0107] 步骤3.3:对步骤3.2中的gene symbol形式进行分析,看是否能在唾液中找到对应物并记录下来。

[0108] 步骤3.4::对步骤3.2中的gene symbol形式进行分析,看是否能在血液中找到对应物并记录下来。

[0109] 步骤3.5::对步骤3.2中的gene symbol形式进行分析,看是否能在尿液中找到对应物并记录下来。

[0110] 步骤3.6:将它们制成一张表格,并判断是否有公共的部分,即找到既能进入唾液又能进入血液和尿液的基因。

[0111] 完成既差异又共表达的编码基因是否可进入体液之后,本发明进入第4阶段。将第3阶段得到的编码基因通过探针注释找到对应的gene symbl放到DAVID网站上分析它们的GOBP和pathway来进一步分析lncRNA的功能。对差异表达的lncRNA进一步筛选,筛选完留下来的lncRNA及其共表达又差异的编码基因可作为癌症的潜在标记物。

[0112] 具体的步骤如下:

[0113] 步骤4.1:把第3阶段得到的gene symbol形式的基因放到DAVIA网站上,能得到两张表格。一个是GO BP表格、一个是pathway表格。

[0114] 步骤4.2:对步骤4.1的两张表格进行分析,取前十个,来研究它们涉及哪些过程。对三套片子共同的过程进行分析,分析哪些过程可能与癌症有关并导致癌症的发生,从而也推断lncRNA可能参与的过程。

[0115] 步骤4.3:对胃癌中三套片子的差异表达的lncRNA进一步筛选。用fisher exact test方法来筛选出更为显著的lncRNA。

[0116] 步骤4.4:对筛选留下来的lncRNA,判断三套片子的交集lncRNA。

[0117] 步骤4.5:对共同的lncRNA,去找它所对应的既共表达又差异的编码基因。将三套片子所对应的这些基因再寻找公共的编码基因。

[0118] 步骤4.6:将筛选留下的公共lncRNA和所对应的公共的编码基因作为胃癌的biomarkers。下面,具体对上述步骤中涉及的方法予以详细的说明。

[0119] 1、fisher exact test方法

[0120] 当样本数比较小时,fisher exact test是一个用列连表进行分析的统计显著性检验。fisher检验是建立在超几何分布的基础上,它对应的列连表以及公式如下:

[0121]

	与 tumor 差异表达的 gene 数目	不与 tumor 差异表达的 gene 数目	总数
与 lncRNA 共表达的 gene 数目	a	b	a+b
不与 lncRNA 共表达的 gene 数目	c	d	c+d
总数	a+c	b+d	n

[0122] $$p=\binom{a+b}{a}\binom{c+d}{c}\Big/\binom{n}{a+c}=\frac{(a+b)!(c+d)!(a+c)!(b+d)!}{n!a!b!c!d!}$$

[0123] 在具体分析完各个步骤之后,用实验结果来证明本发明的有效性。

[0124] 实验结果发现,在胃癌的三套片子中,经过fisher exact test精确检验得到的lncRNA有一个公共的lncRNA,OTTHUMG00000016663_noncode,并在最后找到一个公共的可进入血液和唾液的基因SOSTDC1。因此我们可以将此lncRNA和此基因作为胃癌的潜在标记物。

[0125] 同时,我们在PubMed of NCBI website上找到一篇文章是说SOSTDC1与胃癌是有关的,这更能论证我们实验的正确性。

[0126] 本发明提出的这种寻找与胃癌相关的lncRNA和编码基因的方法,简单实用,得出的结果也是比较令人满意,本发明将继续应用在其他癌症上,相信也会得到比较满意的结果。

[0127] 以上所述为本发明的较好的实施例,并不用以限制本发明,凡在本发明的精神和原则之内,所作的任何修改、等同替换、改进等,均应包含在本发明的保护范围之内。

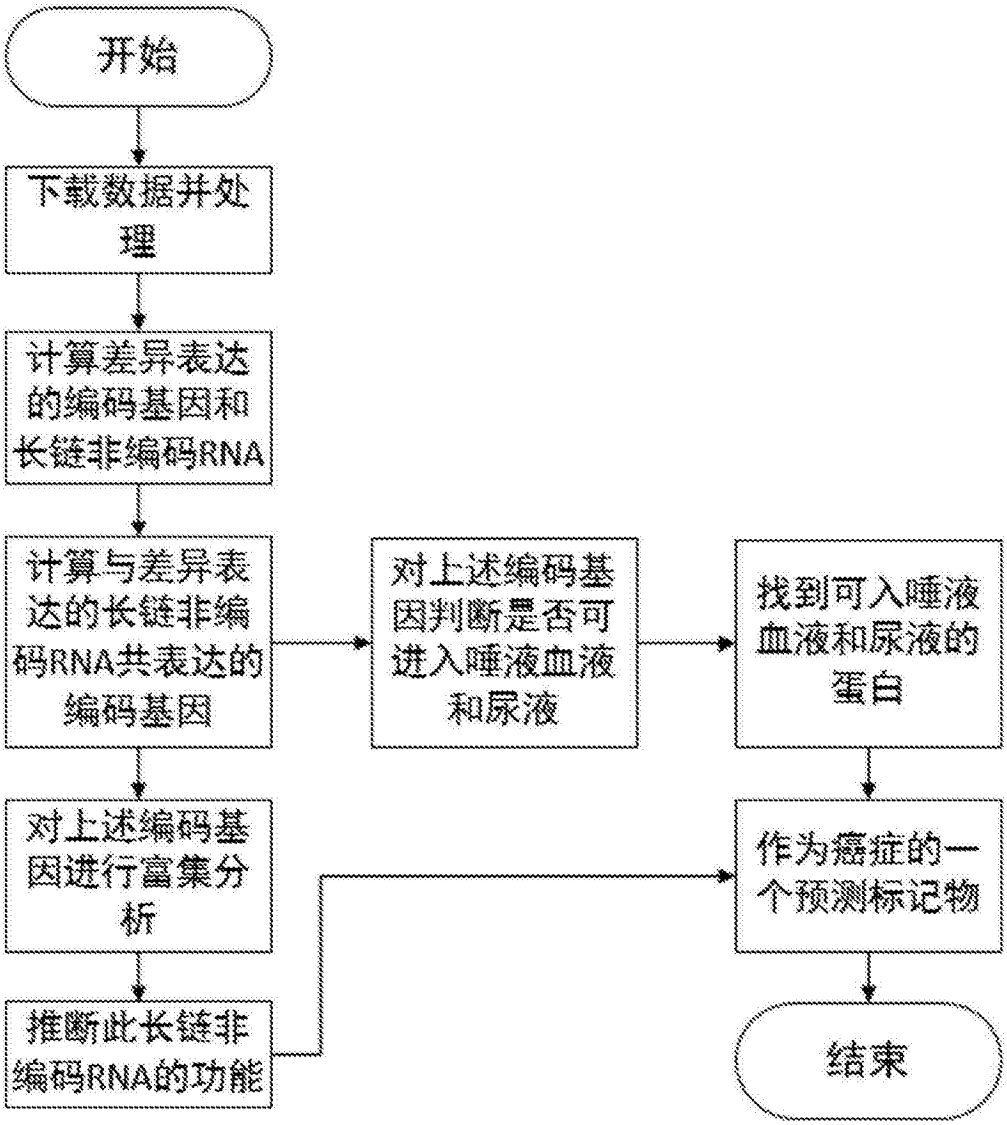

开始
下载数据并处理
计算差异表达的编码基因和长链非编码RNA
计算与差异表达的长链非编码RNA共表达的编码基因
对上述编码基因进行富集分析
推断此长链非编码RNA的功能
对上述编码基因判断是否可进入唾液血液和尿液
找到可入唾液血液和尿液的蛋白
作为癌症的一个预测标记物
结束

图1

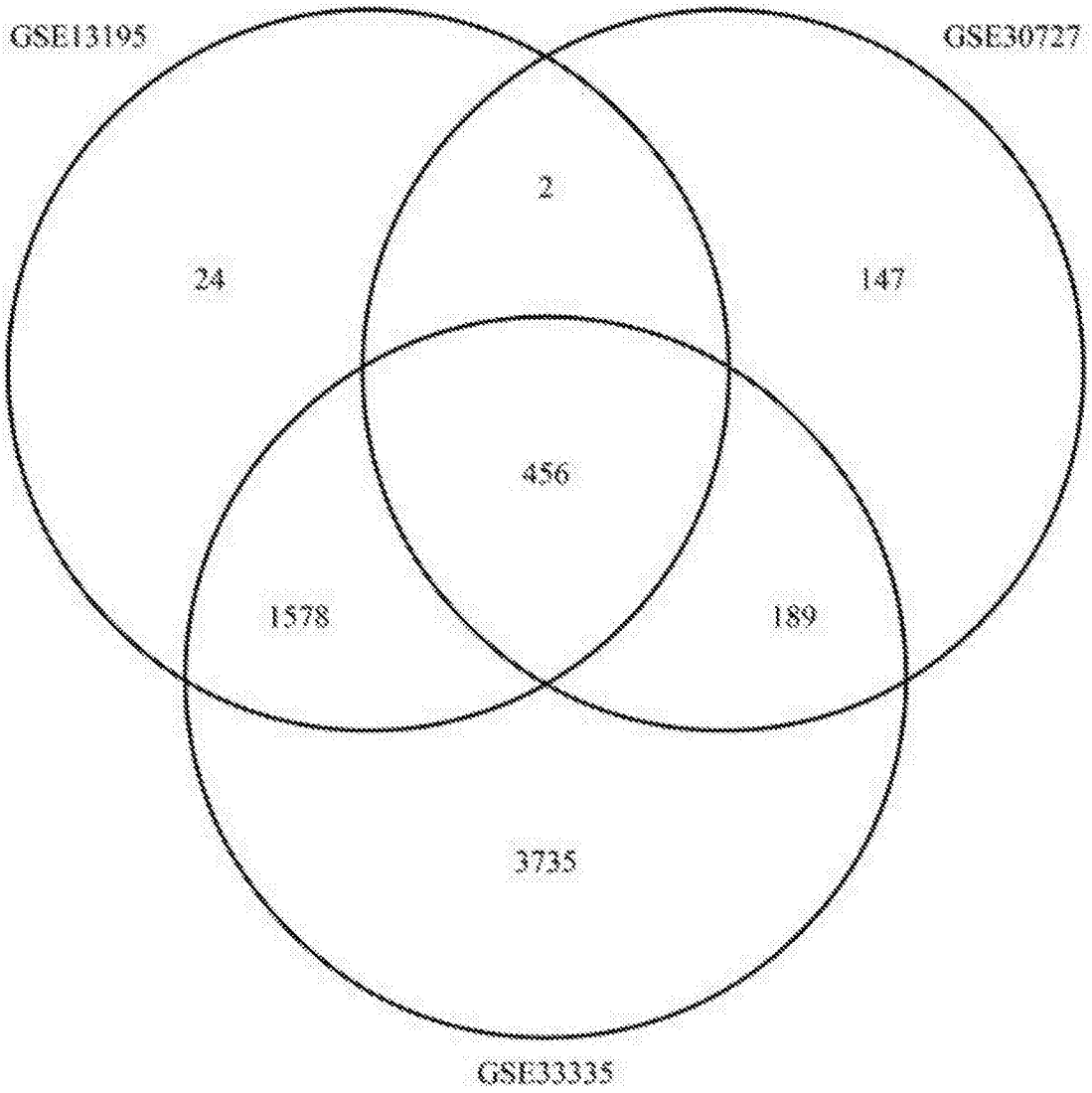


图2

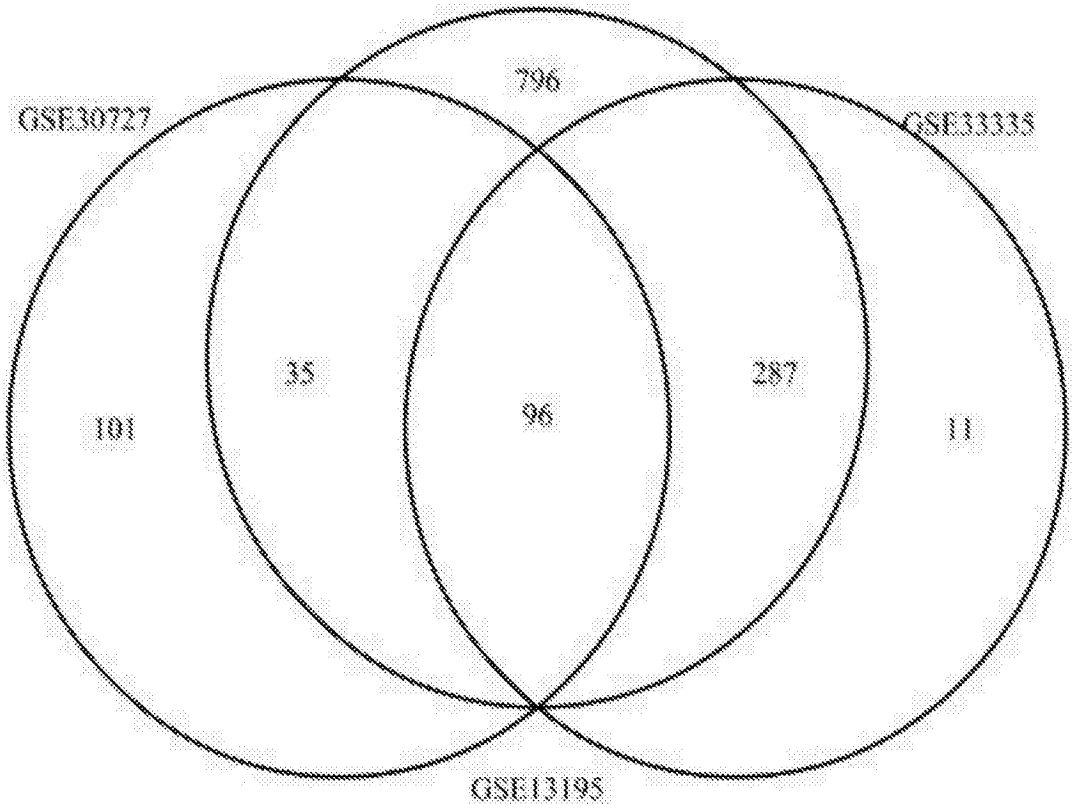


图3

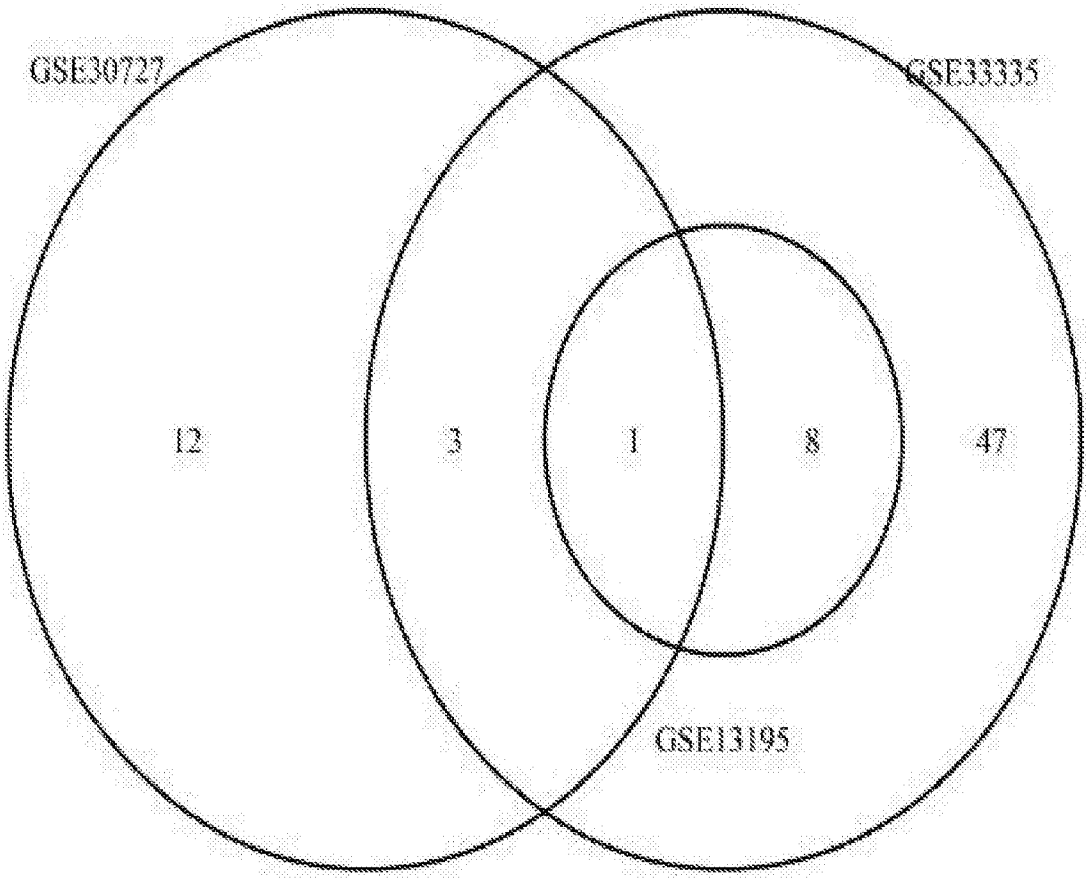


图4

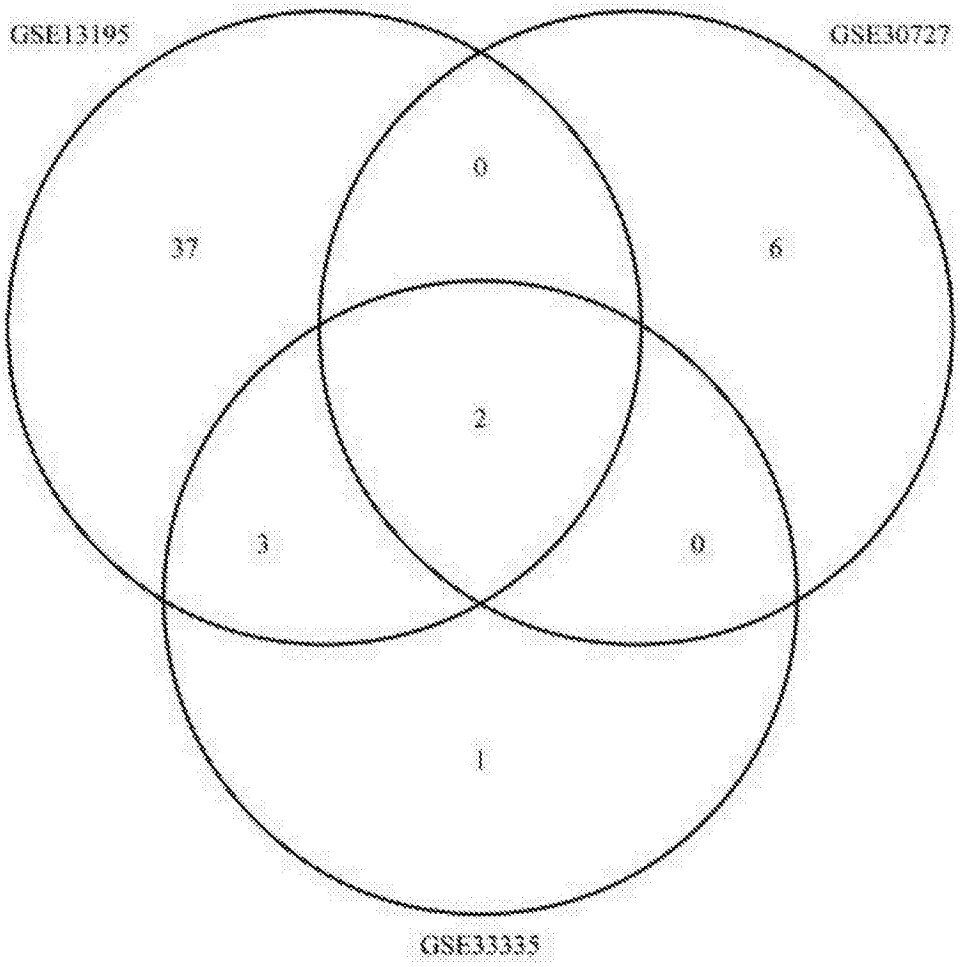


图5

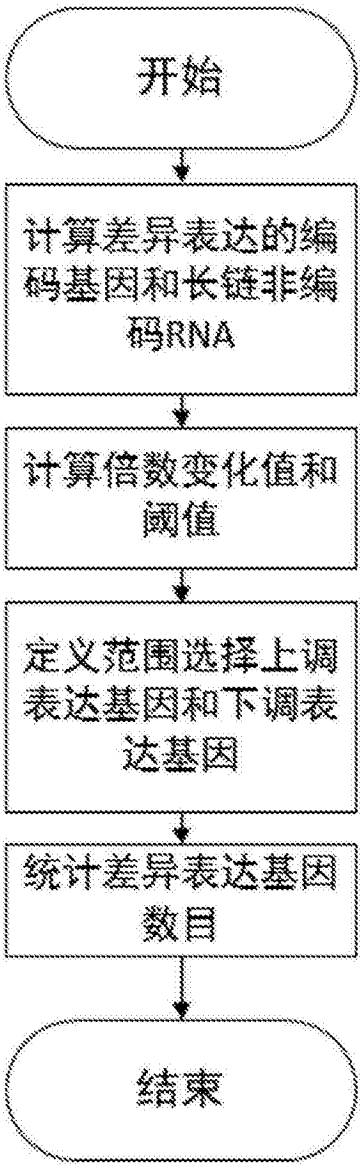

开始
计算差异表达的编码基因和长链非编码RNA
计算倍数变化值和阈值
定义范围选择上调表达基因和下调表达基因
统计差异表达基因数目
结束

图6

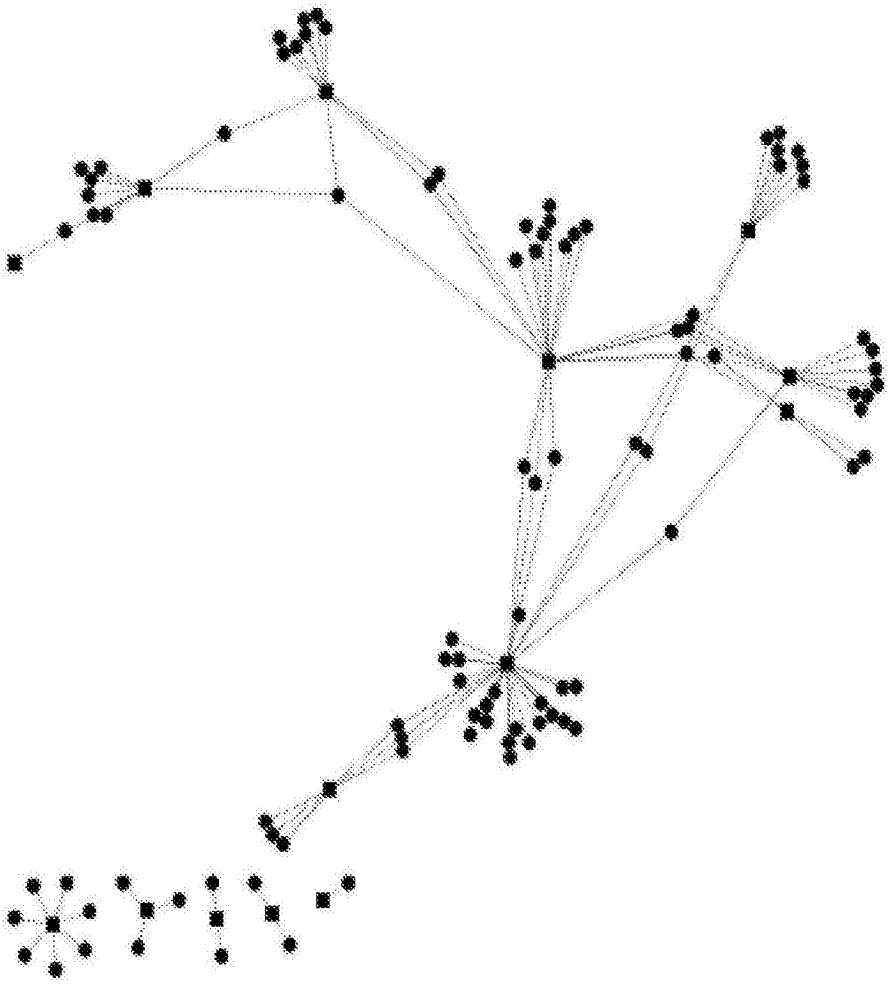


图7

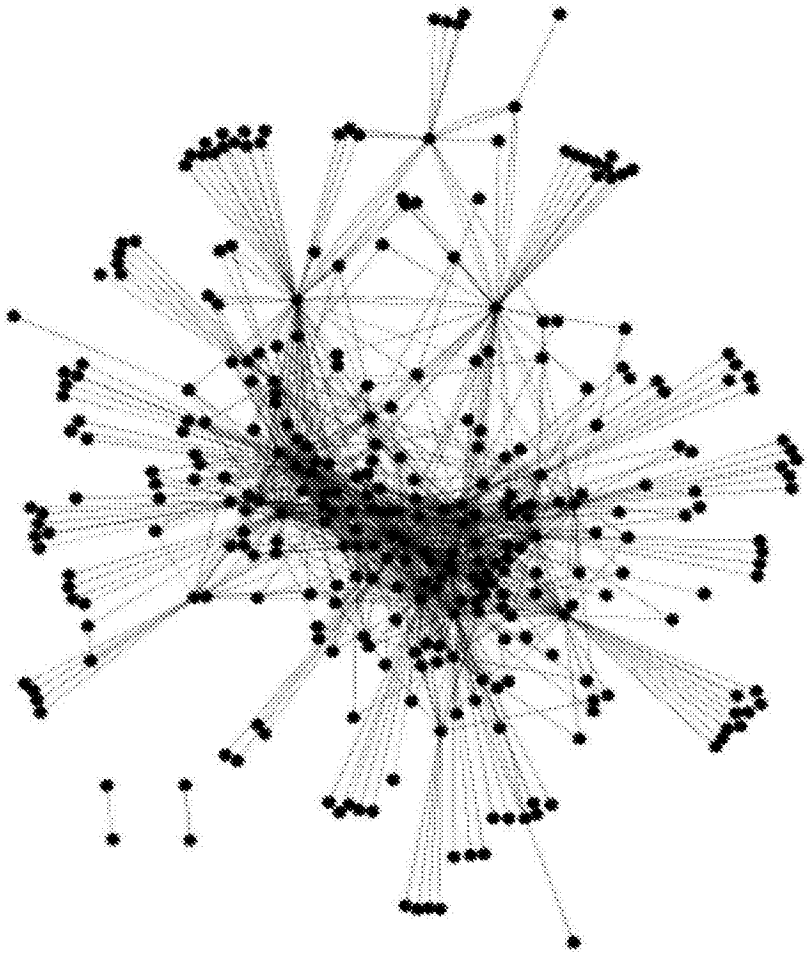


图8

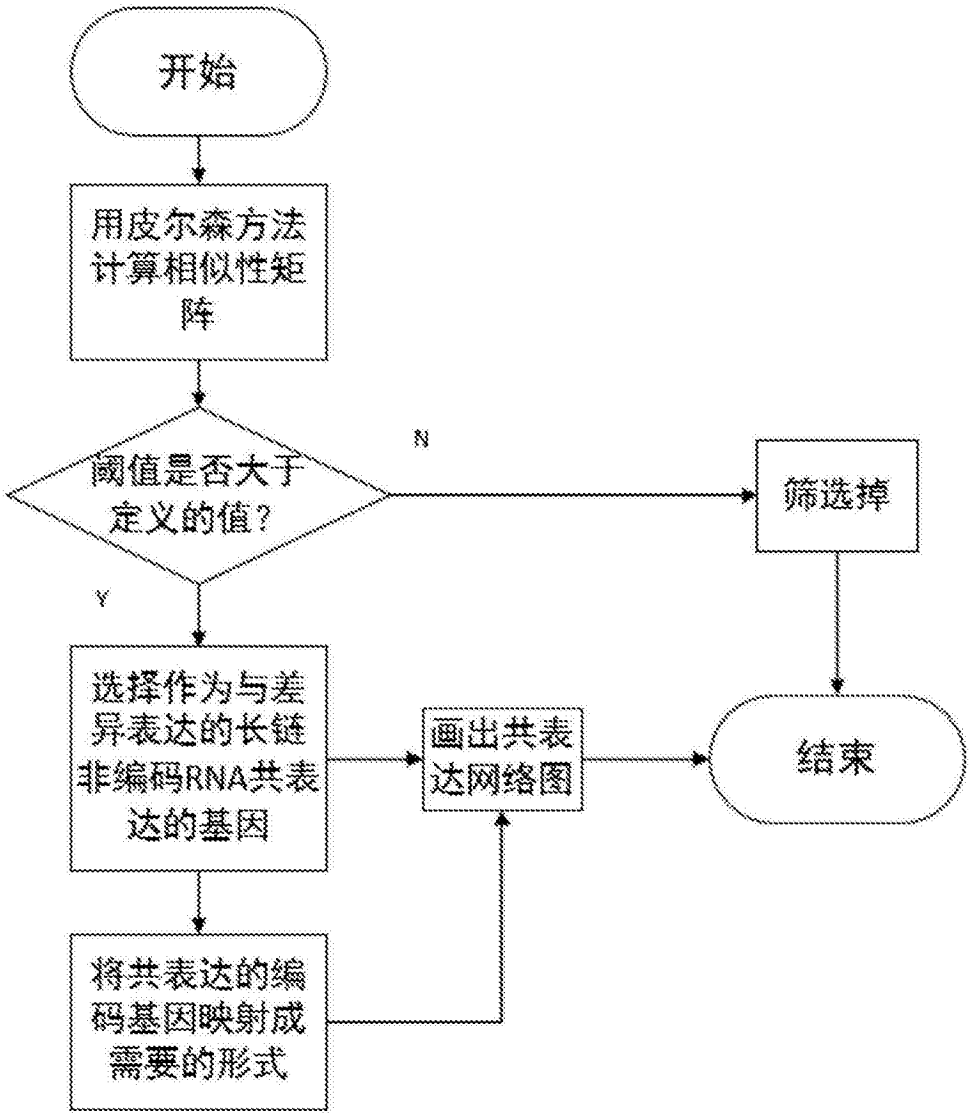

开始
用皮尔森方法计算相似性矩阵
阈值是否大于定义的值？
N
筛选掉
Y
选择作为与差异表达的长链非编码RNA共表达的基因
画出共表达网络图
结束
将共表达的编码基因映射成需要的形式

图9