PIBB 生物化学与生物物理进展
Progress in Biochemistry and Biophysics
2016, 43(1): 63~74
www.pibb.ac.cn

肺鳞状细胞癌癌症发展模式识别分类模型及特征基因识别 *

张 飞[1)] 王世祥[1)] 王 玲[2)] 宋 凯[1, 3)**]

([1)] 天津大学化工学院，天津 300072；[2)] 大连医科大学附属第一医院肿瘤科，大连 116011；
[3)] *University of Texas Southwestern Medical Center, Dallas, Texas* 75235, USA)

摘要 本文利用先进的生物信息学方法，首次从全基因组水平综合基因表达、甲基化水平和拷贝数变异三类数据，寻找与肺鳞状细胞癌(LUSC)发生和发展密切相关的特征基因，为进一步解释其内在机理、开发新的靶向药物和治疗手段提供更加深入的理论依据．为克服全基因组数据超高维高噪声小样本特性对机器学习算法性能的影响，防止信息饱和现象的干扰，本文创新性地组合应用 4 种特征基因筛选方法，分别从特异性、相关性、生物学功能和对肿瘤分类模型的贡献等多个方面，通过迭代降维技术递归筛选真正的特征基因．研究中，我们以 TCGA(The Cancer Genome Atlas project)数据库中的 LUSC Ⅰ～Ⅲ期病人样本为例，对其基因表达数据(GE)、基因甲基化数据 (ME) 以及拷贝数变异数据(CNV)进行分析．结果筛选出 67 个 GE 特征基因，对 3 类样本分类的平均准确率达到 86.29%，70 个 ME 特征基因，相应的分类准确率为 90.92%，31 个 CNV 特征基因，相应的分类准确率为 69.16%．KEGG(Kyoto Encyclopedia of Genes and Genomes)和 IPA(Ingenuity Pathway Analysis)对上述 3 类特征基因集在代谢通路水平和基因调控网络水平上的分析，证明了其在调控水平上的密切关系．同时也表明，识别的特征基因与 LUSC 肿瘤进展之间有着重要的直接关系，这对了解肿瘤机理以及新靶向治疗的发展非常重要．

关键词 肺鳞状细胞癌，基因表达，基因甲基化，拷贝数变异，肿瘤进展，模式识别
学科分类号 Q7，Q81 **DOI**: 10.16476/j.pibb.2015.0352

肺癌以其常年高居首位的致死率，一直是癌症研究领域的热点和难点．按照病理学或组织学来划分，肺癌分为非小细胞肺癌(non-small cell lung cancer，NSCLC)和小细胞肺癌，其中 NSCLC 占肺癌病人总数大约 85%，而鳞状细胞癌(lung squamous cell carcinoma，LUSC)则占 NSCLC 的 30%以上，在全球范围内每年 LUSC 患者的死亡数大约达到了 40 万人[1-3]．LUSC 患者通常在确诊时往往已是晚期，使得目前可用的治疗手段无法得到实施[4-5]，且患者对放疗、化疗不如小细胞未分化癌患者敏感．但同时 LUSC 生长缓慢，如果能够尽早诊断并进行手术切除治疗，患者的生存率则会明显提高．统计数据表明，LUSC Ⅰ期与Ⅱ期的患者在手术后大约为 70%拥有 5 年生存几率，然而对于 ⅢB 期与Ⅳ期的患者来说仅有不超过 5%的患者拥有 5 年生存几率．如果ⅠB-Ⅲ期的患者在手术切除后得到相应的辅助化疗则能明显提升其生存几率[6]．由此可见，对 LUSC 发展机制的研究，识别导致其恶化或者抑制其发展的关键基因，有利于延缓 LUSC 的恶化速度，为患者争取更多的治疗时间，具有非常重要的理论和临床价值．

Tseng 等[7]研究 SLIT2 在肺癌进展中的基因表达水平，结果表明 SLIT2 可以抑制肺癌进展，并认为其可能为肺癌治疗及预后的潜在“治疗靶标”．Gao 等[8]的研究结果表明， LKB1 功能性缺失可以

* 国家自然科学基金资助项目(31271351).
** 通讯联系人.
Tel: 13820186019, E-mail: ksong@tju.edu.cn
收稿日期：2015-11-06，接受日期：2015-12-03

通过细胞外基质微环境的重构导致肺癌的恶化，并发现 LOX 可以作为肺癌患者治疗的潜在靶点. Xiong 等[9]研究结果第一次公开表明 Bmi-1 基因表达水平在早期非小细胞肺癌中呈现强烈的上升而在后期表现出下降的规律，这对于非小细胞肺癌转移和进展过程的机理研究具有重要作用. 此外，Kang 等[10]进行了高分辨率阵列比较基因组杂交(CGH)研究，其中使用了 4 046 个人工细菌染色体克隆阵列，在非小细胞肺癌早期阶段的 36 个肿瘤样本中筛选出与单个基因相关的 DNA 拷贝数变异，结果表明肺癌早期阶段中 5p15.33 区域的扩增最为稳定，并且一系列在 5p15.33 临界区域的基因可作为新的肺癌早期检测和分类的生物标志物.

随着微阵列测序技术和生物信息学的飞速发展，人们能够从人类全基因组水平分析基因表达、甲基化水平变化或拷贝数变异对癌症发生和发展的影响[11]. Lau 等[12]使用基因表达数据为患者总体生存率建立了三基因(STX1A、HIF1A、CCR7)分类器(风险比为 3.8；95%CI 1.7～8.2；$P < 0.001$)，利用此 3 个基因的基因表达水平能够对非小细胞肺癌Ⅰ期与Ⅱ期患者进行分类，并辅助改善组织学对肿瘤阶段的预测能力. Sandoval 等[13]研究了非小细胞肺癌Ⅰ期患者预后中的 DNA 甲基化特征基因，结果表明，基于特定的高甲基化特征基因可以把非小细胞肺癌Ⅰ期患者区分为高风险和低风险患者，并且最佳的 DNA 甲基化生物标志物分析能大大提高对患者预后反应预测的准确度. Anthony 等[14]基于 4 种不同架构的支持向量机(SVM)多类分类的机器学习算法，对肺癌患者 TNM 分期进行分类，其中采用等级分类层次结构的标准二元 SVM 具有最好的分类效果：对 T 和 N 的总体分类准确率分别为 64.10%和 81.90%，其中对 T 的 4 个分期中各个类别的准确率分别为 52.50%、70.10%、73.10%、53.10%，对 N 的 3 个分期的准确率分别为 91.60%、61.10%和 66.70%.

以上研究分别从基因表达水平或者甲基化水平对肺癌发展机制进行了相关研究. 本文则首次从全基因组水平综合基因表达、甲基化水平和拷贝数变异 3 类数据，从微阵列数据出发，利用多类分类生物信息学模式识别技术，从 3 个方面全面寻找 LUSC 阶段进展的相关特征基因，为 LUSC 发生和发展内在机理的进一步揭示、新靶向药物和治疗手段的研发奠定理论基础.

为克服全基因组数据超高维高噪声小样本特性对机器学习算法性能的影响，防止信息饱和(即少数重要基因信息淹没于数万基因所含噪声中)现象，本文创新性地组合应用多种特征基因筛选方法，并通过迭代降维技术递归筛选真正的特征基因.

研究中，我们以 TCGA (The Cancer Genome Atlas project)数据库中的 LUSC Ⅰ～Ⅲ期病人样本为例，对其基因表达数据(GE)、基因甲基化数据(ME)以及拷贝数变异数据(CNV)进行分析，分别筛选出 67 个 GE 特征基因，对 3 类样本分类的平均准确率达到 86.29%，70 个 ME 特征基因，相应的分类准确率为 90.92%，31 个 CNV 特征基因，相应的分类准确率为 69.16%. 同时，KEGG(Kyoto Encyclopedia of Genes and Genomes) 和 IPA (Ingenuity Pathway Analysis)对上述 3 类特征基因集在代谢通路水平和基因调控网络水平上的分析证明了其在调控水平上的密切关系，从而为进一步的代谢通路分析和靶向治疗方法开发的研究提供必要的理论支持.

1 数据与方法

1.1 数据及预处理

本文中肺鳞状细胞癌相应的 GE、ME 和 CNV 数据均来源于 TCGA (the cancer genome atlas project)公共数据库(https://tcga-data.nci.nih.gov/tcga/tcgaHome2.jsp). 其中 GE 数据检测平台为 Illumina HiSeq 2000 RNA Sequencing Version 2，ME 数据检测平台为 Illumina Infinium Human Methylation 450，CNV 数据检测平台为 Affymetrix Genome-Wide Human SNP Array 6.0. 以上 3 种数据均为第三水平数据(经过 TCGA 预处理的标准化数据).

LUSC 患者癌症分期以 TCGA 提供的病人临床信息为准，忽略分期子类(即ⅡA、ⅡB 均视为Ⅱ期样本). 因 LUSC 第Ⅳ期患者样本数过少(具有临床数据的Ⅳ期样本个数只有 3 个)，为此我们仅采用 LUSC 第Ⅰ～Ⅲ期患者的样本. 在多类分类模式识别研究中，将样本按照其临床阶段诊断结果分成 3 类. 同时为后续研究起见，删除缺少重要临床信息(年龄、性别、癌症阶段、生存时间和生存状态)的病人样本. 最终的 GE 数据有 282 个病人样本，ME 数据剩余 178 个样本，CNV 数据有 292 个样本，具体信息见表 1.

Table 1　The summary of the clinical information of TCGA-LUSC samples

	LUSC-GE	LUSC-ME	LUSC-CNV
Number	282	178	292
Age	(67.5±8.7)	(68.0±8.9)	(67.5±8.8)
Gender			
Female	70	43	74
Male	212	135	218
Cancer stage			
Stage Ⅰ	154	96	158
Stage Ⅱ	81	57	82
Stage Ⅲ	47	25	52
Vital status			
Dead	98	58	104
Alive	184	120	188

研究表明，影响基因表达水平的甲基化主要发生在该基因的启动子区域[15]. 因此本文采用所有位于基因启动子区域探针的 β 值均值作为该基因的 ME 数据. 对于 CNV 数据，采用基因所在片段(segment)CNV 的均值作为该基因的 CNV. 为提高分类结果精度和模型训练的速度，在进行特征基因识别分析之前，对上述 3 种基因组数据进行初步滤波处理，删除在所有样本中值为“NA”(Not available)或者“0”的基因. 最终在 GE 数据中共有 20 502 个基因，ME 数据中共有 20 424 个基因，CNV 数据中有 24 663 个基因. 在数据进入模式识别之前，对数据进行中心化处理.

1.2　癌症进展相关特征基因的筛选

本研究项目属于多类分类问题，是机器学习领域的主要难题之一. 微阵列数据的超高维高噪声小样本特性对基于机器学习算法的模式识别分类是另外一个挑战. 少数重要基因信息很容易淹没于全基因组数万基因的噪声中造成信息饱和现象，进一步增加了特征基因识别的难度.

为了克服这些不利因素的影响，本文创新性地组合应用以下特征基因筛选方法，并通过迭代降维递归筛选真正的特征基因.

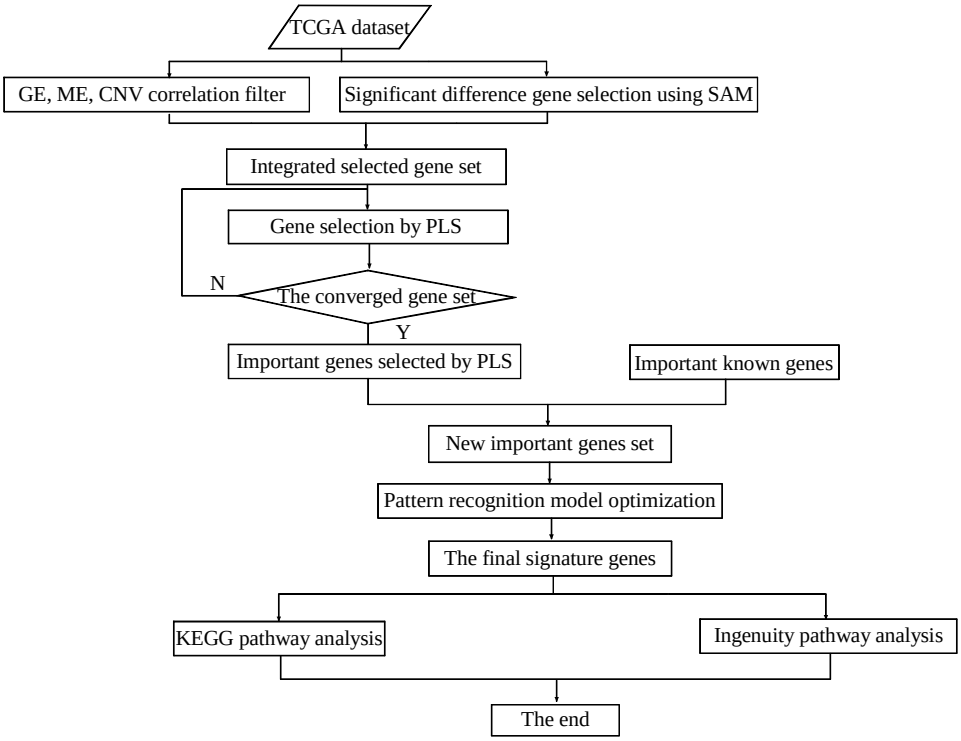

Fig. 1　Flow chart of signature genes identification for LUSC tumor progression

a. 相关性筛选. 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等，从而影响基因表达[16]. 同样，拷贝数变异(CNV)也会在一定程度上影响基因表达，最终影响整个相关代谢网络. 因此，相比之下，能够影响基因表达水平的甲

基化和拷贝数变异具有更加显著的生物学意义，对癌症的发展能起到更加关键的作用．因此，本文筛选 ME 或 CNV 与基因表达水平具有明显相关性(相关系数的绝对值大于 0.5)的基因作为候选基因集的第一部分．

b．显著性筛选(SAM)．研究表明 2/3 以上的基因其 CNV 和 GE 之间没有显著相关性[17]．并且 ME 未必会通过直接相关性影响相应基因的表达水平而影响癌症的发展．为避免筛选方法一中遗漏与基因表达水平无直接相关性的特征 ME 或 CNV 基因，在第二步采用基因差异显著性分析(SAM)对全基因组数据进行筛选．SAM 是一种常用的微阵列数据预处理方法，常用于高维基因数据的初步筛选．该方法通过多重基因特异性检验识别在不同类别中具有显著差异的基因，并通过错误发现率(FDR)算法控制多重检验的错误率[18-20]．

c．偏最小二乘算法(PLS) 筛选．SAM 虽然能够克服常规显著性检验的局限性并在一定程度上限制错误发现率，但毕竟属于单变量分析方法，无法克服噪声及变量相关性的影响，因此只适合于基因变量的初步筛选．PLS(偏最小二乘)算法则可以通过提取与原始变量线性相关的互相正交的潜变量[21]，将原始高维样本压缩至低维空间进行模式识别和分类分析，因此能够有效地克服数据中噪声和多重相关性等问题，在生物信息学领域得到越来越广泛地应用．为此本文采用 PLS 筛选最终的特征基因．

d．基于模式识别分类精度的综合筛选．只有对 LUSC 癌症进展分类模型具有最大贡献的基因才是最终的特征基因集．为此本文最终采用多重交叉验证法对模式识别分类模型进行优化，选取具有最高分类精度的最小基因集作为特征基因集．同时为突出已知重要基因(见附件表 S1)的作用，在模型优化过程中，将其与前三步筛选的候选基因集合并后进行模式识别分类模型训练．

具体流程如图 1 所示．SAM、PLS 均在 R 语言环境下运行，相应细节请参见本实验室网站(http://www.csssk.net)提供的附件支持材料．

1.3 癌症阶段分类模型优化及评价方法

本文的研究内容为模式识别分类问题中的多类分类模型，其中病人样本按癌症阶段诊断分类为Ⅰ期、Ⅱ期和Ⅲ期患者，忽略分期子类(即ⅡA、ⅡB 均视为Ⅱ期样本)，在本文中相应为Ⅰ类、Ⅱ类和Ⅲ类样本．同时采用 5 重交叉验证方法对癌症阶段模式识别分类模型进行优化．

本文采用总体预测准确率(ACC)以及Ⅰ期，Ⅱ期和Ⅲ期各自识别的准确率($ACC1$，$ACC2$，$ACC3$)来评价分类模型的分类结果．ACC，$ACC1$，$ACC2$ 与 $ACC3$ 的定义如下：

$$ACC=\frac{TN1+TN2+TN3}{N1+N2+N3} \tag{1}$$

$$ACC1=\frac{TN1}{N1} \tag{2}$$

$$ACC2=\frac{TN2}{N2} \tag{3}$$

$$ACC3=\frac{TN3}{N3} \tag{4}$$

其中：TNi(i=1，2，3)为 i 期预测样本中被正确识别的样本个数；Ni(i=1，2，3)为 i 期预测样本的总数．

2 结果与讨论

2.1 GE、ME 和 CNV 数据特征基因分类结果比较

本文的研究目标是识别对 LUSC 癌症进展至关重要的特征基因，为进一步揭示肺鳞状细胞癌发生和发展的机理奠定基础．为此所识别的特征基因必须具有足够高的癌症阶段分类能力，相应的模式识别模型必须具有足够高的分类精度，才能有效证明特征基因的代表性：即仅分别采用 GE、ME 和 CNV 特征基因相应的表达水平、甲基化水平或拷贝数变异，即可对病人样本按照癌症阶段进行分类(本研究中分为癌症Ⅰ期、Ⅱ期和Ⅲ期样本三类)．本研究所识别的 GE、ME 和 CNV 特征基因，相应模式识别分类模型的分类结果如表 2 所示．其中包含了 PLS 分类器对测试样本的总体分类准确率和各个阶段类别的正确识别率．三种分类模型的总体分类准确率分别为：86.3%、90.9%和 69.2%．

Table 2 Performance of stage classification model for LUSC tumor progression

Data	Signature gene number	$ACC1$	$ACC2$	$ACC3$	ACC
GE	67	91.59%	85.13%	72.83%	86.29%
ME	70	94.95%	86.92%	83.60%	90.92%
CNV	31	86.16%	54.00%	36.83%	69.16%

从 GE、ME 和 CNV 的生物学机理进行分析可知：

a．人类基因组基因的表达水平可以受到除疾病以外包括环境、个体样本健康状况、情绪等多种因素的影响[22]．因此基因表达水平数据中所包含的其他因素引起的噪声信息也最多．

b．虽然表观遗传学变异性随时间变化的规律尚未明确，但已有的研究表明，相当一部分基因甲基化水平在非常长的一段时间(11～20 年)内基本保持稳定[23]，甚至有些甲基化模式具有遗传性[24]．由此可见 ME 数据是本文所研究的三类数据中较为稳定，信噪比(信息噪声含量)相应较高的一类信息数据.

c．从数据所包含的信息变异量而言，GE (log2)的变化范围为 0～17.976， ME 的变化范围为 0%～99.286%，而 CNV 值变化范围为–5.5871～4.7908，(TCGA 中第三水平拷贝数变异经过 log2 处理，即 $\log_2$(CNV/2))为三类数据中所含变异信息最少的一类数据．且 CNV 信息正向变异(拷贝数大于 2)的范围为 0～4.7908，而负向变异(拷贝数小于 2)的范围为–5.5871～0．由此可见，ME 信息含量最高最稳定，GE 次之但噪声最高，CNV 信息含量最少．

从上述分析可以看出，理论上讲，ME模型应该具有最佳分类效果，GE 次之，CNV 最低． 表 2 所示结果充分证实了这一点．ME 模型在三组数据中取得了最好的分类效果：其总体准确率高达 90.92%，并且Ⅰ期、Ⅱ期和Ⅲ期三个类别之间的准确率相差最小(即模型的稳定性最高)，相差仅为 7.32%．ME 对三类样本各自的分类准确率为 94.95%、86.92%和 83.60%，均高于 80%．充分证明了其最优的分类精度．

而 GE 模型介于两者之间，总体分类精度为 86.29%，对三类样本各自的分类准确率为 91.59%，85.13%和 72.83%．其中对于Ⅲ期样本的分类精度略低于 75%，不是特别理想．较 ME 的分类能力差些．

相比之下，CNV 模型的分类结果最差，其精度仅为 69.2%．此精度虽然很低，但仍高于三类分类模型的平均分类精度 33.3%的 2 倍以上．而且 CNV 模型的特征基因个数仅为 31 个，更有利于进一步的实验验证．

本文的分类结果还与已有的研究成果进行了对比，Showe 等[25]采用基因表达数据与 SVM 算法识别出 29 个区分 NSCLC 与 NHC 的特征基因，同时利用 29 个特征基因对 NSCLC 不同癌症阶段和 NHC 两类样本之间进行分类，分类结果表明非小细胞肺癌Ⅰ期与 NHC 之间的总体分类准确率为 82.26%，Ⅱ期与Ⅲ期与 NHC 之间总体分类准确率依次为 81.49%和 86%．这些分类结果均低于本文 GE 模型的总体分类准确率 86.29%，并且本文为三类分类问题在模式识别问题的难度远远高于其他两类问题，由此突出本文算法的可靠性与优越性．

2.2 LUSC 阶段进展相关特征基因

如前所述，微阵列数据的超高维小样本特性是影响特征基因识别的主要原因之一．本文从 GE 数据 20 502 个基因，ME 数据 20 424 个基因，CNV 数据 24 663 个基因中采用多重迭代筛选，最终分别识别出 67 个 GE 特征基因，70 个 ME 特征基因以及 31 个 CNV 特征基因．相应特征基因详细信息见附件表 S2～S4 所示．

在这三组特征基因中，GE 特征基因和 ME 特征基因有 2 个公共基因，HIF1A 和 STAT5B，ME 特征基因和 CNV 特征基因有 1 个公共基因，DLC1．这 3 个基因都是非常重要的肺癌相关基因．

a．HIF1A (hypoxia inducible factor 1)基因编码 α 亚基的转录因子低氧诱导因子 1，由一个 α 和一个 β 亚基组成的异二聚体．它在胚胎血管、肿瘤血管生成和缺血性疾病病理生理学中起重要作用．研究证明，由于低氧或遗传交替导致 HIF1A 的下调或者过度表达，对大量的癌症生物学以及其他一些病理生理学具有重要影响，特别是在血管化和血管生成、能量代谢、细胞存活和肿瘤侵入等方面．临床上 HIF1A 在非小细胞肺癌中表现出的高表达现象与肿瘤进展密切相关，能够通过其在引发血管生成和调节细胞代谢克服低氧的作用促进肿瘤生长和转移，因此是放射治疗，化疗和死亡率上升预测和预后标志物基因之一[12, 26–27]．

b．STAT5B (signal transducer and activator of transcription 5B)由该基因编码的蛋白质是转录因子 STAT 家族的一个成员．已经证明它参与多种生物学过程，如 T 细胞受体信号传导、细胞凋亡、成人乳腺发育和肝基因表达的性别差异[28]．

c．DLC1(deleted in liver cancer 1)位于染色体 8p22-p21.3 区域，DLC1 基因在几个实体癌，如非小细胞肺癌、鼻咽癌、乳腺癌等中通常表现为下调或者删失[29]．当细胞处于压力之下，它作为一种肿瘤抑制基因可以抑制细胞的生长和增殖以及诱导凋亡．DLC1 还参与局部黏附的形成，所以 DLC1 的

删失导致细胞黏附的减小和细胞转移潜力的增加. 此外 Cao 等[30]的研究表明，tensin-DLC1-RhoA信号传导轴在肿瘤发生和肿瘤转移中起重要作用，并且可以用于探索癌症干预.

GE 特征基因中已知重要基因为：HIF1A、STAT5B，其功能分析如上所述. 它们在三类癌症阶段患者样本的基因表达水平蜂群图及箱线图如图 2 所示. 由图 2 可见，HIF1A 在Ⅰ、Ⅱ和Ⅱ、Ⅲ期样本的 *P* 值均小于阈值 0.05，说明其表达量在上述三组之间明显不同，且由中值可以看出，HIF1A 的表达量先是升高，进入Ⅲ期后反而有减小的趋势，值得深入研究. STAT5B 则有所不同，仅在Ⅱ期和Ⅲ期样本中其表达水平具有显著差异. 但从模式识别模型的系数可以看出，STAT5B 的贡献率在 2 万多个候选基因中排名第 67，由此可见单纯采用显著性差异分析很有可能漏掉此类特征基因，相反，也证明了本文所采用的多种筛选方法的必要性和优越性.

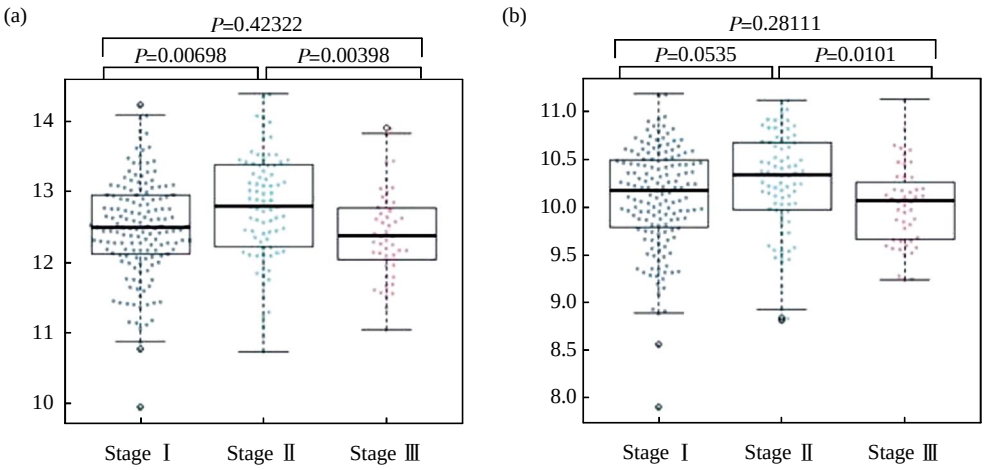

Fig. 2　Beeswarm of GE signature genes in three cancer stage

(a), (b) Corresponding to HIF1A and STAT5B GE signature genes.

ME 数据识别 70 个特征基因中有 2 个高甲基化基因：RPL13AP3 和 WBP11P1，其基因的甲基化水平均高于 80%，与基因表达量的相关性分别为–0.01 和 0.15，由此证实了对自身基因表达水平无直接影响的甲基化也有可能通过代谢或者调控网络等影响其他基因，进而对整个癌症网络产生影响. ME 特征基因中还包含 12 个低甲基化基因：LDHB、CBR1、ZNF607、IGF2BP3、GDF6、SIM2、ADHFE1、DAP、HIF1A、XRCC1、TUBB3 和 STAT5B，它们的甲基化水平均低于 20%. 其中 LDHB、ZNF607 和 ADHFE1 与相应的基因表达量相关系数小于–0.5，表现出较强的负相关性.

近年来，一些研究工作证实了在 70 个 ME 特征基因中有 9 个基因与癌进展相关. 其在三个不同阶段的箱线图如图 3所示，已知重要 ME 特征基因与基因表达数据的相关性分析如图 4 所示. 其中 3 个已知的重要基因 HIF1A、STAT5B 和 DLC1 已经详细进行了讨论，其余 6 个已知的重要基因分别是：

a. STX1A. STX1A 在小细胞肺癌中表现出欠表达，并且在结肠癌和直肠癌中表现的更加突出[12]. 从图 3 可以看出其在癌症Ⅲ期病人样本中 ME 水平有所升高.

b. CCR7. CCR7 的作用在各种癌细胞，如非小细胞肺癌、胃癌和食道癌中都有所表现. 研究表明，低氧 -HIF-1α, 2α-CCR7-ERK1/2 通路可以调节肺癌细胞在低氧条件下的迁移和侵袭，并促进肺癌转移[31]. 由图 4 可知 CCR7 的甲基化水平与基因表达呈现较强的负相关性，说明甲基化水平的升高会抑制相应的基因表达水平，从而影响肿瘤的进展.

c. BRCA1. BRCA1 编码一个维持基因组稳定性作用的核磷蛋白，并且还起到抑制肿瘤的作用. 它在 DNA 修复中起着中心作用，BRCA1 基因过表达与 NSCLC 患者不良预后紧密相关[32].

图 4 可见 BRCA1 的甲基化水平与基因表达呈现负相关性，因此可以看出其甲基化水平的升高，导致其表达水平的降低，使得其抑制肿瘤发展的作用减小，最终导致肿瘤的恶化.

d. DAP. DAP 激酶在非小细胞肺癌患者早期到晚期的疾病进展中发挥重要作用[33].

e. XRCC1. XRCC1 在非小细胞肺癌中表现出过表达[34]，并且在非小细胞肺癌淋巴结转移后达到一个更高的水平[35].

f. TUBB3. TUBB3 低表达水平对于紫杉醇和长春碱的反应是肿瘤的一个良好指标，而高表达的 TUBB3 与紫杉醇和长春碱的抗性相关，且 TUBB3 与肿瘤分化和病理分期显著相关，其基因水平在低分化肿瘤和预后较差的晚期非小细胞肺癌患者中显著偏高[36].

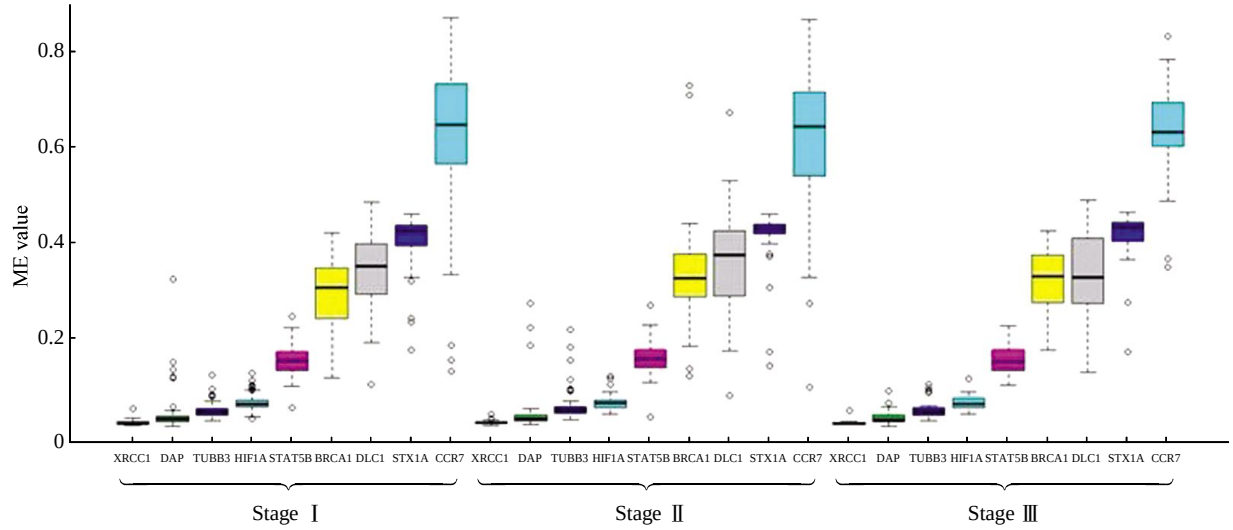

Fig. 3　Boxplot of important known ME signature genes in three cancer stages

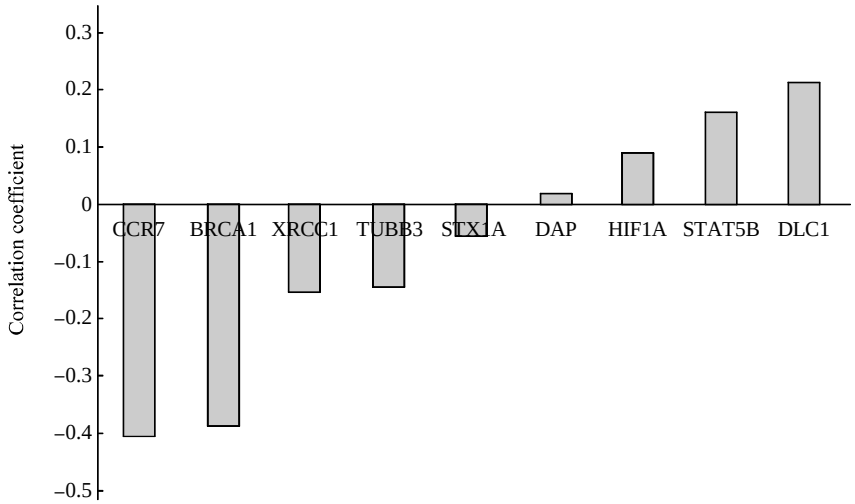

Fig. 4　The correlation coefficients between ME and GE data for ME important known signature genes

在 CNV 数据识别的 31 个 LUSC 肿瘤进展生物标志物中包含了已知的重要基因 DLC1. DLC1 是 Okayama 等[26]研究发现的对早期肺腺癌术后患者临床管理决策有重要辅助作用的 4 个重要基因之一. 其他 CNV 特征基因中，USP14 对应的 CNV 数据与 GE 数据的相关系数最大，其中 CNV 与 GE 具有最强烈的正相关性，如图 5 所示. 由此可见其拷贝数的变异可能是导致其表达水平变化的原因之一，进而对癌症进展产生影响.

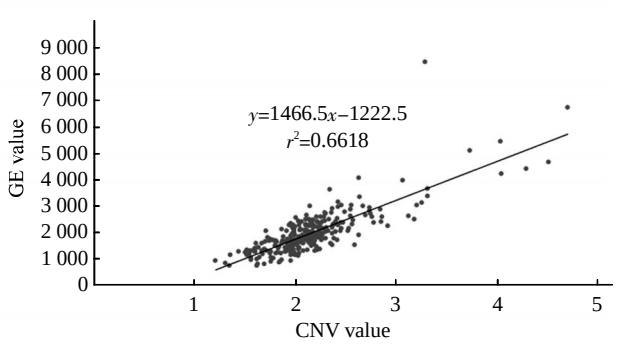

Fig. 5　The scatterplot of USP14 CNV signature gene in CNV and GE data

2.3　特征基因的 KEGG 通路分析

本文采用 DAVID 生物信息学资源 6.7NIAID/NIH (http://david.abcc.ncifcrf.gov/tools.jsp) 功能注释工具对特征基因集进行 KEGG 通路分析．为研究调控网络或代谢水平上的相互联系，对识别的 67 个 GE 特征基因、70 个 ME 特征基因和 31 个 CNV 特征基因集共 165 个特征基因进行综合分析．

KEGG 分析的结果表明 LUSC 中的 165 个特征基因与趋化因子信号通路以及 TGF-β 信号通路显著相关($P < 0.05$)，具体信息如表 3 所示．趋化因子信号通路具有能够使受体二聚化和 TK 通路激活的功能[37]．近年来的研究认为 TGF-β 的信号传导通路既是一个肿瘤抑制途径又是肿瘤进展和侵入的启动子．Derynck 等[38]在其研究中评估了 TGF-β 在肿瘤发展中的作用，并核对 TGF-β 在致癌中的正面和负面影响．Markowitz 等[39]在其研究中指出 TGF-β 及其完整的信号通路是新发现的人类肿瘤抑制基因家族的成员．这些研究既表明 TGF-β 信号通路和肿瘤进展有着重要的关系，同时也说明本文识别的特征基因集的重要性和可信性．

Table 3　Enrichment analysis of KEGG pathways for the 165 signature genes

KEGG_ID	Term	*P*-value	Gene symbols
hsa04062	Chemokine signaling pathway	0.018	CCR7, DOCK2, VAV3, STAT5B, ADRBK2, CCL7
hsa04350	TGF-beta signaling pathway	0.034	NOG, GDF6, NODAL, BMP5

2.4　特征基因的 IPA 分析

利用 IPA (ingenuity pathway analysis) 对所有特征基因进行调控网络分析，得到 18 个基因调控网络与特征基因具有直接关系，其中包括 6 个主要的基因调控网络，如图 6 所示，其详细信息见表 4．从表 4 中可以看出调控网络与细胞生长和增殖、细胞形态、癌症以及其他重要的功能和疾病密切相关．IPA 分析同时表明了三类特征基因在基因调控网络水平上的密切相关性．如：对于基因调控网络 1，FLRT3、MX2、PMAIP1 和 TDRD1 属于 GE 特征基因；CLEC7A、FES、HOXB4 和 ZNF483 属于 ME 特征基因；ACO2、CSNK1D、SGTB 和 USP14 属于 CNV 特征基因；DLC1 则同时属于 ME 和 CNV 特征基因．

通常认为在调控网络中拥有 5 个或以上直接连接的基因为“Hub”基因．本研究中识别的特征基因中共有 5 个“Hub”基因：LDLR、BRCA1、HIF1A、STAT5B 和 TNFSF11．HIF1A、STAT5B 和 BRCA1 的基因功能前面已经做了详细讨论，这里主要讨论 LDLR 和 TNFSF11 的基因功能．

a．LDLR (low density lipoprotein receptor)．该基因位于染色体 19p13.1-13.3 区域，跨越 45 kb 由 18 个外显子与 17 个内含子组成，其编码 839 个氨基酸的成熟蛋白[40]．属于一种内吞性受体，系低密度脂蛋白受体基因家族中的一员．由于其能够与多种结构及功能各异的配体相互作用，不仅可以对血脂的动态平衡及纤溶功能的稳定进行调节，而且能参与多种生长因子、细胞激酶生物学效应的发挥．

b．TNFSF11 (tumor necrosis factor (ligand) superfamily，member 11)．该基因也叫作 RANKL，它编码肿瘤坏死因子(TNF)细胞因子家族的一个成员[41]．研究表明，RANKL 的表达水平可使充分的微环境条件影响癌症细胞迁移(即慢性淋巴细胞性白血病(CLL)和多发性骨髓瘤)[42]．

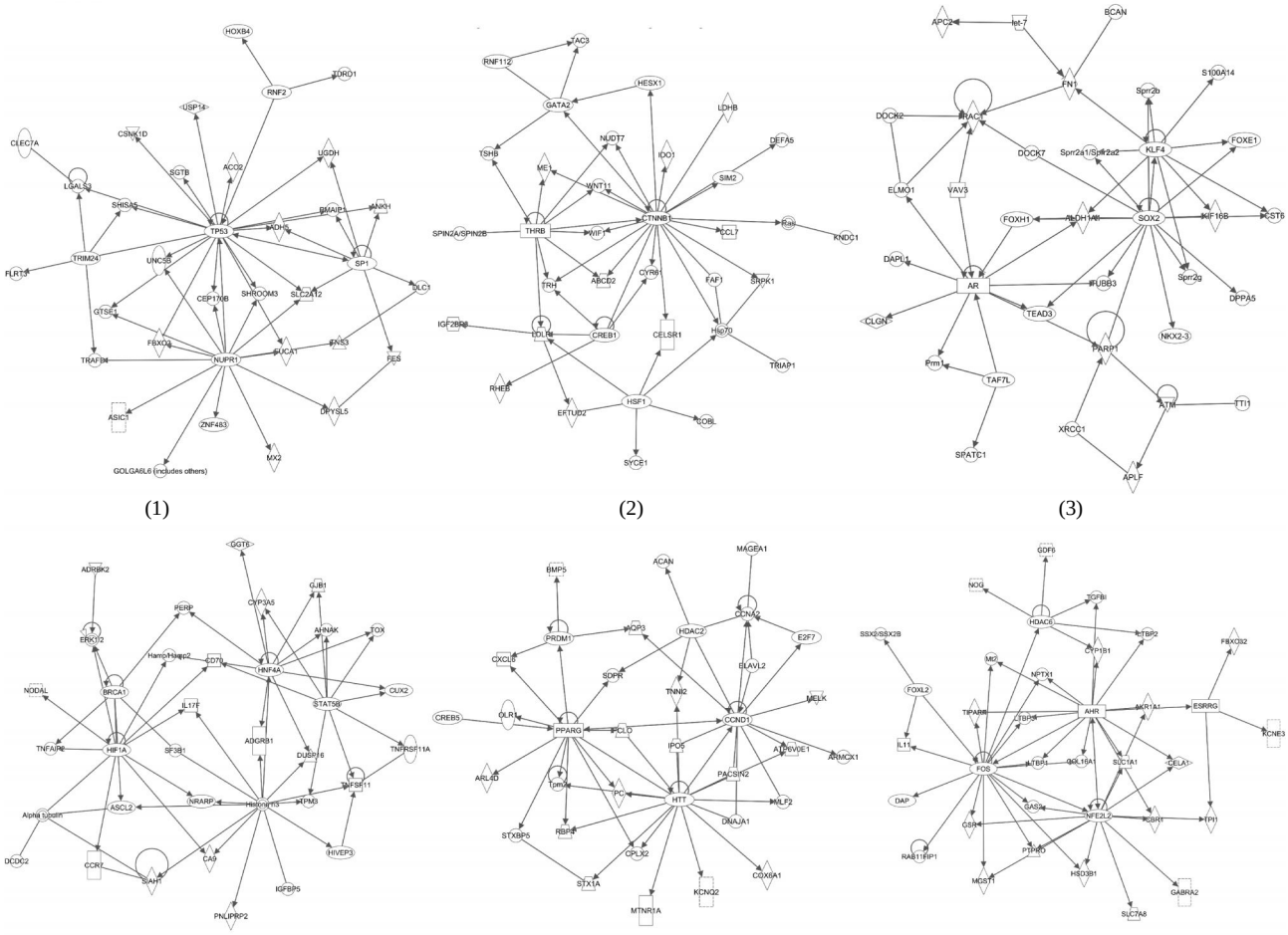

Fig. 6 The main gene regulatory networks of signature genes by IPA

Table 4 The summary of the IPA genetic network analysis of 165 signature genes in LUSC

ID	Molecules in network	Score	Top diseases and functions
1	ACO2[&], ADH5, ANKH, ASIC1, CEP170B, CLEC7A[#], CSNK1D[&], DLC1[#&], DPYSL5, FBXO3, FES[#], FLRT3[*], FUCA1, GOLGA6L6(includes others), GTSE1, HOXB4[#], LGALS3, MX2[*], NUPR1, PMAIP1[*], RNF2, SGTB[&], SHISA5, SHROOM3, SLC2A12, SP1, TDRD1[*], TNS3, TP53, TRAFD1, TRIM24, UGDH, UNC5B, USP14[&], ZNF483[#]	24	Cellular growth and proliferation, cell morphology, hematological system development and function
2	ABCD2, CCL7[#], CELSR1, COBL[*], CREB1, CTNNB1, CYR61, DEFA5[#], EFTUD2, FAF1, GATA2, HESX1, HSF1, Hsp70, IDO1[*], IGF2BP3[#], KNDC1[*], LDHB[#], LDLR[#], ME1, NUDT7, Ras, RHEB[&], RNF112, SIM2[#], SPIN2A/SPIN2B, SRPK1, SYCE1[*], TAC3[*], THRB, TRH, TRIAP1[&], TSHB, WIF1, WNT11	22	Amino acid metabolism, small molecule biochemistry, embryonic development
3	ALDH1A1[#], APC2[*], APLF, AR, ATM, BCAN[#], CLGN[*], CST6, DAPL1[#], DOCK2[#], DOCK7, DPPA5[#], ELMO1, FN1, FOXE1, FOXH1, KIF16B, KLF4, let-7, NKX2-3[#], PARP1, Prm1, RAC1, S100A14[#], SOX2, SPATC1[*], Sprr2a1/Sprr2a2, Sprr2b, Sprr2g, TAF7L, TEAD3, TTI1[&], TUBB3[#], VAV3[*], XRCC1[#]	22	Cancer, gastrointestinal disease, organismal injury and abnormalities
4	ADGRB1, ADRBK2[&], AHNAK, Alphatubulin, ASCL2[*], BRCA1[#], CA9, CCR7[#], CD70, CUX2, CYP3A5, DCDC2[*], DUSP16, ERK1/2, GGT6[*], GJB1, Hamp/Hamp2, HIF1A[*#], Histoneh3, HIVEP3, HNF4A, IGFBP5[*], IL17F, NODAL[*], NRARP, PERP, PNLIPRP2[#], SF3B1, SIAH1, STAT5B[*#], TNFAIP2[#], TNFRSF11A, TNFSF11[*], TOX, TPM3	19	Digestive system development and function, lymphoid tissue structure and development, organ morphology
5	ACAN[#], AQP3, ARL4D[&], ARMCX1[#], ATP6V0E1, BMP5[*], CCNA2, CCND1, COX6A1[&], CPLX2, CREB5[*], CXCL6, DNAJA1, E2F7, ELAVL2, HDAC2, HTT, IPO5, KCNQ2[#], MAGEA1[#], MELK[*], MLF2, MTNR1A[#], OLR1[#], PACSIN2, PC, PPARG, PRDM1, RBP4, SDPR, STX1A[#], STXBP5, TNNI2, Tpm2	18	Hereditary disorder, neurological disease, psychological disorders
6	AHR, AKR1A1, CBR1[#], CELA1, COL16A1, CYP1B1, DAP[#], ESRRG, FBXO32[*], FOS, FOXL2, GABRA2[&], GAS2, GDF6[#], GSR, HDAC6, HSD3B1, IL11, KCNE3[#], LTBP1, LTBP2, LTBP3, MGST1, Mt2, NFE2L2, NOG[&], NPTX1, PTPRO, RAB11FIP1[*], SLC1A1, SLC7A8[*], SSX2/SSX2B, TGFBI, TIPARP, TPI1	14	Ophthalmic disease, organismal injury and abnormalities, developmental disorder

[*]GE signature genes; [#]ME signature genes; [&]CNV signature genes.

附件 有关微阵列显著性分析(SAM)与偏最小二乘法(PLS)详细介绍以及表 S1～S4 见本文网络版附录(http://www.pibb.ac.cn) 或宋凯实验室网站(http://www.csssk.net)提供的附件支持材料.

参 考 文 献

[1] Perez-Moreno P, Brambilla E, Thomas R, *et al*. Squamous cell carcinoma of the Lung: molecular subtypes and therapeutic opportunities. Clinical Cancer Research, 2012, **18**(9): 2443−2451

[2] Couraud S, Zalcman G, Milleron B, *et al*. Lung cancer in never smokers - A review. Eur J Cancer, 2012, **48**(9): 1299−1311

[3] Kenfield S A, Wei E K, Stampfer M J, *et al*. Comparison of aspects of smoking among the four histological types of lung cancer. Tob Control, 2008, **17**(3): 198−204

[4] Bach P B, Kelley M J, Tate R C, *et al*. Screening for lung cancer - A review of the current literature. Chest, 2003, **123**(1): 72s−82s

[5] Yu D P, Li J, Han Y, *et al*. Gene expression profiles of ERCC1, TYMS, RRM1, TUBB3 and EGFR in tumor tissue from non-small cell lung cancer patients. Chinese Medical Journal, 2014, **127**(8): 1464−1468

[6] El-Sherif A, Luketich J D, Landreneau R J, *et al*. New therapeutic approaches for early stage non-small cell lung cancer. Surgical Oncology-Oxford, 2005, **14**(1): 27−32

[7] Tseng R C, Lee S H, Hsu H S, *et al*. SLIT2 Attenuation during lung cancer progression deregulates beta-catenin and E-cadherin and associates with poor prognosis. Cancer Research, 2010, **70** (2): 543−551

[8] Gao Y J, Xiao Q A, Ma H M, *et al*. LKB1 inhibits lung cancer progression through lysyl oxidase and extracellular matrix remodeling. Proc Natl Acad Sci USA, 2010, **107**(44): 18892−18897

[9] Xiong D, Ye Y L, Fu Y J, *et al*. Bmi-1 expression modulates non-small cell lung cancer progression. Cancer Biology & Therapy, 2015, **16**(5): 756−763

[10] Kang J U, Koo S H, Kwon K C, *et al*. Gain at chromosomal region 5p15.33, containing TERT, is the most frequent genetic event in early stages of non-small cell lung cancer. Cancer Genetics and Cytogenetics, 2008, **182**(1): 1−11

[11] Bhattacharjee A, Richards W G, Staunton J, *et al*. Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses. Proc Natl Acad Sci USA, 2001, **98**(24): 13790−13795

[12] Lau S K, Boutros P C, Pintilie M, *et al*. Three-gene prognostic classifier for early-stage non-small-cell lung cancer. J Clin Oncol, 2007, **25**(35): 5562−5569

[13] Sandoval J, Mendez-Gonzalez J, Nadal E, *et al*. A prognostic DNA methylation signature for stage Ⅰ non-small-cell lung cancer. J Clin Oncol, 2013, **31**(32): 4140

[14] Nguyen A, Moore D, Mccowan I, *et al*. Multi-class classification of cancer stages from free-text histology reports using support vector machines; proceedings of the Engineering in Medicine and Biology Society, 2007 EMBS 2007 29th Annual International Conference of the IEEE, F, 2007 [C]. IEEE

[15] Jones P A. The DNA methylation paradox. Trends in Genetics, 1999, **15**(1): 34−37

[16] Phillips T. The role of methylation in gene expression. Nature Education, 2008, **1**(1): 116

[17] Henrichsen C N, Chaignat E, Reymond A. Copy number variants, diseases and gene expression. Human Molecular Genetics, 2009, **18**: R1−R8

[18] George G, Raj V C. Review on feature selection techniques and the impact of SVM for cancer classification using gene expression profile. arXiv preprint arXiv:1109.1062, 2011

[19] Tusher V G, Tibshirani R, Chu G. Significance analysis of microarrays applied to the ionizing radiation response. Proc Natl Acad Sci USA, 2001, **98**(9): 5116−5121

[20] Zhang C Y, Girard L, Das A, *et al*. Nonlinear quantitative radiation sensitivity prediction model based on NCI-60 cancer cell lines. Sci World J, 2014: 903602−903602

[21] Abdi H. Partial least square regression (PLS regression)// Encyclopedia for Research Methods for the Social Sciences, Thpusards Oaks(CA): Sage, 2003: 792−795

[22] Gibson G. The environmental contribution to gene expression profiles. Nat Rev Genet, 2008, **9**(8): 575−581

[23] Talens R P, Boomsma D I, Tobi E W, *et al*. Variation, patterns, and temporal stability of DNA methylation: considerations for epigenetic epidemiology. Faseb Journal, 2010, **24**(9): 3135−3144

[24] Bird A. DNA methylation patterns and epigenetic memory. Gene Dev, 2002, **16**(1): 6−21

[25] Showe M K, Vachani A, Kossenkov A V, *et al*. Gene expression profiles in peripheral blood mononuclear cells can distinguish patients with non-small cell lung cancer from patients with nonmalignant lung disease. Cancer Research, 2009, **69** (24): 9202−9210

[26] Okayama H, Schetter A J, Ishigame T, *et al*. The expression of four genes as a prognostic classifier for stage Ⅰ lung adenocarcinoma in 12 independent cohorts. Cancer Epidem Biomar, 2014, **23** (12): 2884−2894

[27] Bos R, Van Der Groep P, Greijer A E, *et al*. Levels of hypoxia-inducible factor-1alpha independently predict prognosis in patients with lymph node negative breast carcinoma. Cancer, 2003, **97**(6): 1573−1581

[28] Pastuszak-Lewandoska D, Domanska D, Czarnecka K H, *et al*. Expression of STAT5, COX-2 and PIAS3 in correlation with NSCLC histhopathological features. PloS One, 2014, **9**(8): e104265

[29] Liao Y C, Lo S H. Deleted in liver cancer-1 (DLC-1): a tumor suppressor not just for liver. The International Journal of Biochemistry & Cell Biology, 2008, **40**(5): 843−847

[30] Cao X, Voss C, Zhao B, *et al*. Differential regulation of the activity of deleted in liver cancer 1 (DLC1) by tensins controls cell migration and transformation. Proc Natl Acad Sci USA, 2012, **109**(5): 1455−1460

[31] Li Y, Qiu X, Zhang S, *et al*. Hypoxia induced CCR7 expression *via* HIF-1alpha and HIF-2alpha correlates with migration and invasion

in lung cancer cells. Cancer Biology & Therapy, 2009, **8**(4): 322−330

[32] Boukovinas I, Papadaki C, Mendez P, *et al*. Tumor BRCA1, RRM1 and RRM2 mRNA expression levels and clinical response to first-line gemcitabine plus docetaxel in non-small-cell lung cancer patients. PloS One, 2008, **3**(11): e3695

[33] Kim D H, Nelson H H, Wiencke J K, *et al*. Promoter methylation of DAP-kinase: association with advanced stage in non-small cell lung cancer. Oncogene, 2001, **20**(14): 1765−1770

[34] Kang C H, Jang B G, Kim D W, *et al*. The prognostic significance of ERCC1, BRCA1, XRCC1, and betaIII-tubulin expression in patients with non-small cell lung cancer treated by platinum- and taxane-based neoadjuvant chemotherapy and surgical resection. Lung Cancer, 2010, **68**(3): 478−483

[35] Kang C H, Jang B G, Kim D-W, *et al*. Differences in the expression profiles of excision repair crosscomplementation group 1, X-Ray repair crosscomplementation group 1, and βIII-tubulin between primary non-small cell lung cancer and metastatic lymph nodes and the significance in mid-term survival. J Thorac Oncol, 2009, **4**(11): 1307−1312

[36] Yu D, Li J, Han Y, *et al*. Gene expression profiles of ERCC1, TYMS, RRM1, TUBB3 and EGFR in tumor tissue from non-small cell lung cancer patients. Chinese Medical Journal, 2014, **127**(8): 1464−1468

[37] Mellado M, Rodriguez-Frade J M, Manes S, *et al*. Chemokine signaling and functional responses: the role of receptor dimerization and TK pathway activation. Annu Rev Immunol, 2001, **19**: 397−421

[38] Derynck R, Akhurst R J, Balmain A. TGF-beta signaling in tumor suppression and cancer progression. Nature Genetics, 2001, **29**(2): 117−129

[39] Markowitz S D, Roberts A B. Tumor suppressor activity of the TGF-beta pathway in human cancers. Cytokine & Growth Factor Reviews, 1996, **7**(1): 93−102

[40] Sudhof T C, Goldstein J L, Brown M S, *et al*. The LDL receptor gene: a mosaic of exons shared with different proteins. Science, 1985, **228**(4701): 815−822

[41] Hanada R, Hanada T, Sigl V, *et al*. RANKL/RANK-beyond bones. Journal of Molecular Medicine, 2011, **89**(7): 647−656

[42] Schmiedel B J, Scheible C A, Nuebling T, *et al*. RANKL expression, function, and therapeutic targeting in multiple myeloma and chronic lymphocytic leukemia. Cancer Research, 2013, **73**(2): 683−694

Pattern Recognition of The Lung Squamous Cell Carcinoma Tumor Progression Classification Model and Signature Genes Identification*

ZHANG Fei[1)], WANG Shi-Xiang[1)], WANG Ling[2)], SONG Kai[1, 3)**]

([1)] *School of Chemical Engineering and Technology, Tianjin University, Tianjin* 300072, China;
[2)] *The First Affiliated Hospital Oncology of Dalian Medical University, Dalian* 116011, China;
[3)] *University of Texas Southwestern Medical Center, Dallas, Texas* 75235, USA)

Abstract To identify signature genes for the tumor progression of lung squamous cell carcinoma, which provides a deeper theoretical basis for further explanation of its inherent mechanism, new targeted drugs and treatments development. The pattern recognition method was used to analysis the genome-wide mRNA gene expression (GE) values, methylation values (ME), and copy number variation (CNV) data. To overcome the disadvantages inherent in the genome-wide data such as ultrahigh-dimensional-small-size, high-noise and multi-correlation among genes, and to overcome the predominate influence of the whole genome to the dozens of signature genes, a new iterative multiple variable selection strategy was used to identify signature genes step by step. The importance of genes was comprehensively evaluated by their significant difference with SAM (significant analysis of microarray), statistical analysis using PLS (partial least squares), known biological functions and contributions to the classification model. 67 GE signature genes, 70 ME signature genes and 31 CNV signature genes were identified from the LUSC stage Ⅰ ~ Ⅲ patient samples in TCGA (The Cancer Genome Atlas project) database. The corresponding accuracies from 5 fold cross-validation are: 86.29%, 90.92 % and 69.16% respectively. The genetic network analysis and pathway analysis using KEGG (Kyoto Encyclopedia of Genes and Genomes) and IPA (Ingenuity Pathway Analysis) indicated the highly related relationship among these three kinds of genes. They also indicated the immediate relationship between our signature genes and the progression of LUSC which is very important to the understanding of its mechanism and to the development of new targeted therapy.

Key words lung squamous cell carcinoma, genome-wide mRNA gene expression, gene methylation, copy number variation, tumor progression, pattern recognition
DOI: 10.16476/j.pibb.2015.0352

*This work was supported by a grant from The National Natural Science Foundation of China (31271351).
**Corresponding author.
Tel: 86-13820186019, E-mail: ksong@tju.edu.cn
Received: November 6, 2015 Accepted: December 3, 2015

附 录

1 微阵列显著性分析 (SAM)

SAM 是一种广泛应用于高维微阵列数据候选基因初步筛选的统计方法[1-3]. SAM 通过计算统计量来衡量基因数据变化在统计学意义上的显著性，该得分用于测量 X 变量和响应变量(类别)之间的关系的强度．基因的得分值超过一个阈值就被假定为与类别成员显著相关，并且该阈值是可以调节，并通过 FDR(false discovery rate)控制多重显著性检验的错误发现率[1,4].

2 偏最小二乘法 (PLS)

偏最小二乘法(PLS)是一种非常有效的针对高维小样本数据的模式识别算法．PLS 算法通过提取与原始变量线性相关的互相正交的潜变量，将原始高维样本压缩至低维空间进行模式识别和回归分析，因此能够有效地克服数据中噪声和多重相关性等问题．PLS 算法已经广泛应用于许多领域，特别是生物信息学中的高维微阵列数据的分类问题[5]．在本文中 PLS 发挥着删除模式分类中与分类无关的基因，以及在基因筛选以后对挑选的基因进行模式分类的功能.

假设数据 $X \in R_{m\times p}$ 和 $X \in R_{m\times 3}$(此处 X 为 GE、ME 或 CNV 数据，Y 为相应肿瘤阶段模式识别标号)存在如下线性关系：

$$Y=XB+V \tag{1}$$

其中，B 是回归系数，V 是残差矩阵，m 是样本数量，p 是变量数量.基于潜变量提取思想，对 X，Y 矩阵线性分解建立如下模型：

$$Y=UQ^T+F \tag{2}$$

$$X=TP^T+E \tag{3}$$

其中，U，T 分别是从 X，Y 得到的潜变量矩阵，Q，P 分别是对应的载荷矩阵，F，E 分别是对应的残差矩阵．U 和 T 之间的关系：

$$u_i = b_i t_i + r_i (i=1, 2, \cdots, h) \tag{4}$$

其中，u_i 和 t_i 分别是 X 和 Y 的第 i 个潜变量，b_i 是最小偏差 r_i 确定的系数，h 是最优潜变量个数．最终 PLS 回归系数为：

$$B=W(P^TW)^{-1}Q^T=X^TU(T^TXX^TU)^{-1}T^TY \tag{5}$$

其中，W 是 X 的权重矩阵.

3 基因列表

Table S1 The important known genes

Gene symbol	Location	Gene symbol	Location	Gene symbol	Location
STX1A	7q11.23	*CCR7*	17q21.2	*HIF1A*	14q23.2
BRCA1	17q21.31	*DLC1*	8p22	*XPO1*	2p15
STAT5B	17q21.2	*DAP*	5p15.2	*XRCC1*	19q13.31
TUBB3	16q24.3				

Table S2 67 GE signature genes

Gene symbol	Location	Gene symbol	Location	Gene symbol	Location
LOC90784	2p11.2	*TMEM88B*	1p36.33	*C9orf172*	9q34.3
TAC3	12q13.3	*TNFSF11*	13q14.11	*LAIR2*	19q13.42
WFDC10B	20q13.12	*FLJ40292*	–	*PMAIP1*	18q21.32
LRRC4B	19q13.33	*C9orf50*	9q34.11	*MESTIT1*	7q32.2
CHCHD10	22q11.23	*C4orf31*	4q27	*CTCFL*	20q13.31
RAB11FIP1	8p11.23	*FLJ30679*	16q24.1	*TCP11*	6p21.31
VAV3	1p13.3	*BMP5*	6p12.1	*C9orf84*	9q31.3
SLC7A8	14q11.2	*DFNB59*	2q31.2	*C7orf40*	7p13
FBXO32	8q24.13	*NAP1L5*	4q22.1	*ZNF90*	19p12
HAAO	2p21	*CLGN*	4q31.1	*MYH15*	3q13.13
C1orf168	1p32.2	*SLC10A5*	8q21.13	*APC2*	19p13.3
ABCA9	17q24.2	*C9orf140*	9q34.3	*MX2*	21q22.3
IDO1	8p11.21	*C2orf74*	2p15	*RIC3*	11p15.4
DCDC2	6p22.3	*HIF1A*	14q23.2	*GGT6*	17p13.2
CREB5	7p15.1	*NODAL*	10q22.1	*KNDC1*	10q26.3
ENTHD1	22q13.1	*C19orf46*	19q13.12	*ASCL2*	11p15.5
IGFBP5	2q35	*SSX2*	Xp11.22	*SYCE1*	10q26.3
PBXIP1	1q21.3	*COBL*	7p12.1	*ZNF14*	19p13.11
IL1RL2	2q12.1	*SPATC1*	8q24.3	*MELK*	9p13.2
ACCN2	12q13.12	*PCDP1*	2q14.2	*FLRT3*	20p12.1
RPS26P11	Xq13.1	*TDRD1*	10q25.3	*CDNF*	10p13
ATP11C	Xq27.1	*ADAM32*	8p11.22	*STAT5B*	17q21.2
C8orf46	8q13.1				

Table S3 70 ME signature genes

Gene symbol	Location	Gene symbol	Location	Gene symbol	Location
LDHB	12p12.1	*WBP11P1*	18q12.1	*ZNF671*	19q13.43
CBR1	21q22.12	*HOXB4*	17q21.32	*ACAN*	15q26.1
OR2B11	1q44	*MTNR1A*	4q35.2	*ADHFE1*	8q13.1
CLEC7A	12p13.2	*PNLIPRP2*	10q25.3	*KCNQ2*	20q13.33
SPRN	10q26.3	*DOCK2*	5q35.1	*S100A14*	1q21.3
DAPL1	2q24.1	*KCNE3*	11q13.4	*LOC648691*	22q11.22
SNORD116-29	15q11.2	*ARMCX1*	Xq22.1	*BRCA1*	17q21.31
C17orf46	17q21.31	*ABCA10*	17q24.3	*TEX13A*	Xq22.3
KLHL34	Xp22.12	*ALDH1A1*	9q21.13	*CSN1S2B*	4q13.3
DEFA5	8p23.1	*PRSS50*	3p21.31	*LOC646813*	11p11.12
BCAN	1q23.1	*PRKY*	Yp11.2	*STX1A*	7q11.23
NXPH1	7p21.3	*ZNF844*	19p13.2	*ZAN*	7q22.1
DPPA5	6q13	*TNFAIP2*	14q32.32	*FES*	15q26.1
RPL13AP3	14q22.3	*GDF6*	8q22.1	*SDR42E1*	16q23.3
ZNF607	19q13.12	*ZNF483*	9q31.3	*MAGEA1*	Xq28
LDLR	19p13.2	*NKX2-3*	10q24.2	*CCR7*	17q21.2
MIR888	Xq27.3	*SLFN12*	17q12	*DAP*	5p15.2
FSD1	19p13.3	*SPIN2A*	Xp11.21	*DLC1*	8p22
TPPP3	16q22.1	*SIM2*	21q22.13	*CYP2B7P1*	19q13.2
IGF2BP3	7p15.3	*FGF11*	17p13.1	*HIF1A*	14q23.2
LOC647121	-	*OR1J4*	9q33.2	*XRCC1*	19q13.31
ZIK1	19q13.43	*CCL7*	17q12	*TUBB3*	16q24.3
LOC284788	20p11.21	*OLR1*	12p13.2	*STAT5B*	17q21.2
PLEKHG4	16q22.1				

Table S4 31 CNV signature genes

Gene symbol	Location	Gene symbol	Location	Gene symbol	Location
SH3BGR	21q22.2	*TRIAP1*	12q24.31	*NOG*	17q22
ADRBK2	22q12.1	*GATC*	12q24.31	*SGTB*	5q12.3
ACO2	22q13.2	*LOC646214*	15q11.2	*TRAPPC13*	5q12.3
CRISP3	6p12.3	*FAM27C*	9q13	*LINC01387*	18p11.31
SLC30A8	8q24.11	*TSG1*	6q16.1	*LOC101927150*	-
CSNK1D	17q25.3	*LOC101927637*	12q24.32	*PIGU*	20q11.22
LINC00911	14q31.3	*USP14*	18p11.32	*TTI1*	20q11.23
ARL4D	17q21.31	*GABRA2*	4p12	*ADARB2*	10p15.3
GOLGA6L6	15q11.2	*MIR4273*	3p12.3	*DLC1*	8p22
LOC101927616	12q24.32	*RHEB*	7q36.1	*CCZ1B*	7p22.1
COX6A1	12q24.31				

参 考 文 献

[1] Tusher V G, Tibshirani R, Chu G. Significance analysis of microarrays applied to the ionizing radiation response (vol 98, pg 5116, 2001). Proc Natl Acad Sci USA, 2001, **98**(18): 10515-10515

[2] Zhang C Y, Girard L, Das A, *et al*. Nonlinear quantitative radiation sensitivity prediction model based on NCI-60 cancer cell lines. Sci World J, 2014

[3] Damle M T, Kshirsagar M. Role of permutations in significance analysis of microarray and clustering of significant microarray gene list. International Journal of Computer Science Issues(IJCSI), 2012, **9**(2):

[4] Benjamini Y. Discovering the false discovery rate. J R Stat Soc B. 2010, **72**: 405-416

[5] Song K, Zhang Z, Tong T P, *et al*. Classifier assessment and feature selection for recognizing short coding sequences of human genes. J Comput Biol, 2012, **19**(3): 251-260