**附件：**

**肺腺癌吸烟相关甲基化模式识别分类模型及特征基因的识别研究**

王世祥1 张飞1 王玲2 宋凯1,3[[1]](#footnote-1)\*

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（德克萨斯大学西南医学中心，达拉斯75235）

表S1 已知重要基因列表[[1-16](#_ENREF_1)]

**Tab.S1 The list of known important genes**

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 编号 | 基因 | 位置 | 编号 | 基因 | 位置 | 编号 | 基因 | 位置 | 编号 | 基因 | 位置 |
| 1 | GSTM1 | 1p13.3 | 7 | EGFR | 7p12 | 13 | MGMT | 10q26 | 19 | MEK1 | 15q22.1-q22.33 |
| 2 | LCK | 1p34.3 | 8 | MET | 7q31 | 14 | KRAS | 12p12.1 | 20 | CHRNA5 | 15q24 |
| 3 | ALK | 2p23 | 9 | AKR1B10 | 7q33 | 15 | ERBB3 | 12q13 | 21 | MMD | 17q |
| 4 | STAT1 | 2q32.2 | 10 | BRAF | 7q34 | 16 | MDM2 | 12q14.3-q15 | 22 | XRCC1 | 19q13.2 |
| 5 | FHIT | 3p14.2 | 11 | NAT2 | 8p22 | 17 | DUSP6 | 12q22-q23 | 23 | ERCC2 | 19q13.3 |
| 6 | ROS-1 | 6q22 | 12 | RET | 10q11.2 | 18 | OLFM4 | 13q21.1 | 24 | RASSF2 | 20p13 |

表**S2.** 箱线图基因顺序列表及**P**值

**Tab.S2 The gene list of boxplot and the corresponding P-value**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 编号 | 基因 | p值 | 编号 | 基因 | p值 | 编号 | 基因 | p值 |
| 1 | MDM2 | 0.3401 | 17 | GSTM1 | 0.1142 | 33 | FGF11 | 0.0091 |
| 2 | DUSP6 | 0.5021 | 18 | PXMP4 | 0.0008 | 34 | OLFM4 | 0.0001 |
| 3 | ZFP28 | 0.0071 | 19 | MET | 0.6416 | 35 | LCK | 0.3018 |
| 4 | GLDC | 0.0255 | 20 | CPXM1 | 0.0017 | 36 | LEAP2 | 0.0133 |
| 5 | MMP25 | 0.0005 | 21 | CHRNA5 | 0.9607 | 37 | TCP11 | 0.0110 |
| 6 | KRAS | 0.0031 | 22 | SRGN | 4.3E-06 | 38 | HTR2B | 0.0044 |
| 7 | STAT1 | 0.0508 | 23 | MMD | 0.7780 | 39 | AKR1B10 | 0.0948 |
| 8 | ALK | 0.4229 | 24 | SULT4A1 | 0.5290 | 40 | GNLY | 0.0087 |
| 9 | EGFR | 0.5468 | 25 | RNASE6 | 0.0004 | 41 | CALML3 | 0.0027 |
| 10 | CYBA | 0.0784 | 26 | ANKRD45 | 0.0029 | 42 | GTSF1 | 3.0E-05 |
| 11 | ZNF572 | 0.0555 | 27 | SPAG6 | 0.0027 | 43 | PAX8 | 0.0158 |
| 12 | CD40 | 0.0002 | 28 | UCHL1 | 0.0170 | 44 | C1orf64 | 0.0012 |
| 13 | RET | 0.6444 | 29 | PCDHB11 | 0.0117 | 45 | GPR152 | 0.0003 |
| 14 | B4GALNT4 | 0.6116 | 30 | PPYR1 | 0.0017 | 46 | CA6 | 0.0008 |
| 15 | WBSCR17 | 5.3E-05 | 31 | FHIT | 0.0029 | 47 | C7orf45 | 0.0963 |
| 16 | ERBB3 | 0.1569 | 32 | MGMT | 0.1840 | 48 | NAT2 | 0.4835 |

**算法介绍**

* 基因表达差异显著性分析（SAM）

差异表达基因筛选的关键是控制假阳性，同时又能保持较高的筛检效率，在数据挖掘方面有着很好的应用．在应用基因表达差异显著性分析（SAM）筛选变量时，它会在基因表达差异性的基础上分配给每个基因一个得分，通过估计假发现率（FDR），达到控制多重检验错误率的目的，使得得分高于阈值的基因被挑选出来[[17-19](#_ENREF_17)]．SAM通过基因特异性t检验识别在统计上显著差异的基因．FDR定义如下：

(1)

其中，*V*表示m个检验中错误拒绝的个数，*S*表示m个检验中正确检验的个数，*R*表示m个检验中拒绝原假设的个数．控制FDR不仅提高了检验的功效，同时也改进了传统的多重假设检验过程过于保守的缺陷．

* 偏最小二乘（PLS）

PLS算法是一种新型的多元统计数据分析方法，它通过提取与原始变量线性相关的互相正交的潜变量[[20](#_ENREF_20)]，将原始高维样本压缩至互相正交的低维空间进行模式识别和回归分析，有效地克服了数据噪声等问题，在生物信息学领域应用得越来越广泛．其最简单的形式，只用一个线性模型来描述独立变量*Y*与预测变量组*X*之间的关系：

*Y*= *b0* + *b1X1* + *b2X2* + ... + *bpXp*(2)

作为一个多元线性回归方法，偏最小二乘的主要目的是要建立一个线性模型：

*Y*= *XB* + *E* (3)

其中，*B*是回归系数，*E*是残差矩阵．利用提取潜变量的方法，对*X*，*Y*矩阵线性分解可以建立如下模型：

*Y = UQT* + *F* (4)

*X* = *TPT* + *E* (5)

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．

最终可得PLS回归系数为：

*B*=*W*(*PTW*)-1*QT*=*XTU*(*TTXXTU*)-1*TTY* (6)其中，*W*是*X*的权重矩阵．即可对模型进行回归．

表S3 特征基因IPA分析汇总

**Tab.S3 The IPA analysis of signature genes**

|  |  |  |  |
| --- | --- | --- | --- |
| ID | Molecules in Network | Score | Top Diseases and Functions |
| 1 | AHSP, Akt, **ALK**, ASF1B, BTG3, **CD40**, **CHRNA5**, **DUSP6**, DUSP16, **EGFR**, **ERBB3**, ERK, ERK1/2, HAT1, HDAC10, Histone h3, Histone h4, Hsp90, HTRA1, JMY, **KRAS**, **LCK**, LRIG1, Mapk, **MDM2**, **MET**, **MGMT**, PI3K (complex), PLC gamma, RBM39, **RET**, **STAT1**, STAT5a/b, UCHL1, VRK2 | 28 | Cancer, Organismal Injury and Abnormalities, Cellular Development |
| 2 | ABCD2, ACOT2, **AKR1B10**, AQP8, arginase, Ces, COL1A1, **CYBA**, CYP27A1, Cyp2c54 (includes others), CYP39A1, Cyp4a14, DIO1, fructose 1,6 biphosphatase, **GSTM1**, GSTM2, GSTM4, Gstt1, GSTT2/GSTT2B, **LEAP2**, MGLL, MGST3, **MMD**, **MMP25**, **NAT2**, NFE2L2, NR1I3, **PAX8**, PPARA, RORC, Slco1a4, Sult1a1, SULT1C3, TNFRSF1A, UGT2B28 | 14 | Drug Metabolism, Protein Synthesis, Glutathione Depletion In Liver |

注：粗体所示为本文所识别的特征基因

**Note：The bold genes are the signature genes**

**参考文献**

[1] Nakajima T, Elovaara E, Anttila S, *et al*. Expression and polymorphism of glutathione S-transferase in human lungs: risk factors in smoking-related lung cancer[J]. Carcinogenesis, 1995, 16(4): 707-711.

[2] Liu Yang, Lan Qing, Siegfried JM, *et al*. Aberrant promoter methylation of p16 and MGMT genes in lung tumors from smoking and never-smoking lung cancer patients[J]. Neoplasia, 2006, 8(1): 46-51.

[3] Sanborn JZ, Salama SR, Grifford M, *et al*. Double minute chromosomes in glioblastoma multiforme are revealed by precise reconstruction of oncogenic amplicons[J]. Cancer Res, 2013, 73(19): 6036-6045.

[4] Sozzi G, Pastorino U, Moiraghi L, *et al*. Loss of FHIT function in lung cancer and preinvasive bronchial lesions[J]. Cancer Research, 1998, 58(22): 5032-5037.

[5] Liu G, Wheatley-Price P, Zhou Wei, *et al*. Genetic polymorphisms of MDM2, cumulative cigarette smoking and nonsmall cell lung cancer risk[J]. International Journal of Cancer, 2008, 122(4): 915-918.

[6] Ma PC, Jagadeeswaran R, Jagadeesh S, *et al*. Functional expression and mutations of c-Met and its therapeutic inhibition with SU11274 and small interfering RNA in non–small cell lung cancer[J]. Cancer Research, 2005, 65(4) :1479-1488.

[7] Korpanty GJ, Graham DM, Vincent MD, *et al*. Biomarkers that currently affect clinical practice in lung cancer: EGFR, ALK, MET, ROS-1, and KRAS[J]. Frontiers in Oncology, 2014, 4(1): 204-211.

[8] Kang M, Lee E, Yoon S, *et al*. AKR1B10 is associated with smoking and smoking-related non-small-cell lung cancer[J]. Journal of International Medical Research, 2011, 39(1): 78-85.

[9] Sun Guangyuan, Liu Bing, He Jin, *et al*. Expression of EGFR is closely related to reduced 3-year survival rate in Chinese female NSCLC[J]. Medical Science Monitor, 2015, 21(1): 2225-2231.

[10] Bublil EM, Yarden Y. The EGF receptor family: spearheading a merger of signaling and therapeutics[J]. Curr Opin Cell Biol, 2007, 19(2): 124-134.

[11] Koromilas AE, Sexl V. The tumor suppressor function of STAT1 in breast cancer[J]. Jak-Stat, 2013, 2(2): 1-5.

[12] Soda M, Choi YL, Enomoto M, *et al*. Identification of the transforming EML4–ALK fusion gene in non-small-cell lung cancer[J]. Nature, 2007, 448(7153): 561-566.

[13] Osawa Y, Osawa KK, Miyaishi A, *et al*. NAT2 and CYP1A2 polymorphisms and lung cancer risk in relation to smoking status[J]. Asian Pacific Journal of Cancer Prevention, 2007, 8(1): 103-108.

[14] Takeuchi K, Soda M, Togashi Y, *et al*. RET, ROS1 and ALK fusions in lung cancer[J]. Nature Medicine, 2012, 18(3): 378-381.

[15] Chen H-Y, Yu S-L, Chen C-H, *et al*. A five-gene signature and clinical outcome in non–small-cell lung cancer[J]. New England Journal of Medicine, 2007, 356(1): 11-20.

[16] Spitz MR, Amos CI, Dong Q, *et al*. The CHRNA5-A3 region on chromosome 15q24-25.1 is a risk factor both for nicotine dependence and for lung cancer[J]. Journal of the National Cancer Institute, 2008, 100(21): 1552-1556.

[17] George G, Raj VC. Review on feature selection techniques and the impact of SVM for cancer classification using gene expression profile[J]. International Journal of Computer Science & Engineering Survey, 2011, 2(3): 42-55.

[18] Tusher VG, Tibshirani R, Chu G. Significance analysis of microarrays applied to the ionizing radiation response[J]. Proceedings of the National Academy of Sciences, 2001, 98(9): 5116-5121.

[19] Zhang Chunying, Girard L, Das A, *et al*. Nonlinear quantitative radiation sensitivity prediction model based on NCI-60 cancer cell lines[J]. The Scientific World Journal, 2014, 2014(5): 602-612.

[20] Abdi H. Partial least square regression (PLS regression)[J]. Encyclopedia for Research Methods for the Social Sciences, 2003, 6(4): 792-795.

1. \* 基金项目：国家自然科学基金面上项目（No. 31271351）

   \* 通讯作者（Corresponding author）， E-mail：[ksong@tju.edu.cn](mailto:ksong@tju.edu.cn) [↑](#footnote-ref-1)