Zhao *et al. BMC Medical Genomics* 2019, **12**(Suppl 8):183
https://doi.org/10.1186/s12920-019-0630-4

BMC Medical Genomics

RESEARCH Open Access

Identification of lung cancer gene markers through kernel maximum mean discrepancy and information entropy

Zhixun Zhao[1], Hui Peng[1], Xiaocai Zhang[1], Yi Zheng[1], Fang Chen[2], Liang Fang[3] and Jinyan Li[1*]

From 18th International Conference on Bioinformatics (InCoB 2019)
Jakarta, Indonesia. 10–12 Septemebr 2019

Abstract

Background: The early diagnosis of lung cancer has been a critical problem in clinical practice for a long time and identifying differentially expressed gene as disease marker is a promising solution. However, the most existing gene differential expression analysis (DEA) methods have two main drawbacks: First, these methods are based on fixed statistical hypotheses and not always effective; Second, these methods can not identify a certain expression level boundary when there is no obvious expression level gap between control and experiment groups.

Methods: This paper proposed a novel approach to identify marker genes and gene expression level boundary for lung cancer. By calculating a kernel maximum mean discrepancy, our method can evaluate the expression differences between normal, normal adjacent to tumor (NAT) and tumor samples. For the potential marker genes, the expression level boundaries among different groups are defined with the information entropy method.

Results: Compared with two conventional methods t-test and fold change, the top average ranked genes selected by our method can achieve better performance under all metrics in the 10-fold cross-validation. Then GO and KEGG enrichment analysis are conducted to explore the biological function of the top 100 ranked genes. At last, we choose the top 10 average ranked genes as lung cancer markers and their expression boundaries are calculated and reported.

Conclusion: The proposed approach is effective to identify gene markers for lung cancer diagnosis. It is not only more accurate than conventional DEA methods but also provides a reliable method to identify the gene expression level boundaries.

Keywords: Lung cancer, Maximum mean discrepancy, Information theory, Biomarker discovery

Background

Small-cell lung carcinoma (SCLC) and non-small-cell lung carcinoma (NSCLC) are two main types of lung cancer, comprising the majority of clinic cases [1]. As the most common cancer, lung cancer is the leading cause of cancer-related deaths all over the world [2, 3]. However, most lung cancer cases were diagnosed in a very late stage when symptoms like coughing, coughing up blood, shortness of breath and chest pains appeared. Many early-diagnosed lung cancer cases were detected by accident [3, 4]. In the clinic practice, the most widely used examinations for lung cancer are chest radiography and computed tomography(CT), but these two methods require visible and irreversible histological variants in human lung, resulting in rather low sensitivity in the early stage [5–7]. Therefore, it is a crucial issue to find more timely and accurate approaches for lung cancer early-stage diagnosis.

Due to the progress in molecular biology, some molecules which play vital roles in lung cancer

*Correspondence: Jinyan.Li@uts.edu.au
[1]Advanced Analytics Institute, Faculty of Engineering and Information Technology, University of Technology Sydney, PO Box 123, Broadway, 2007 Sydney, NSW, Australia
Full list of author information is available at the end of the article

development are possible to diagnose cancer and distinguish the specific cancer sub-types [8–10]. Researchers have explored to identify efficient biomarkers from these molecules as the indicator of the pathogenic process to improve the diagnosis sensitivity [11]. These explorations are mainly focused on genetic mutations, DNA methylation profile, miRNA synthesis profile and especially blood proteins [12–19]. Till now, panels of protein markers have been identified and intensively used in clinic applications. For example, the combinations of CEACAM, CYFRA 21-1, ProGRP, CA125, NSE (neuron-specific enolase) and NY-ESO (cancer-testis antigen) are popular lung cancer diagnosis markers [20–24]. Recently, researchers also discovered that β-chain of human haptoglobin [25], SAA (serum amyloid A) [26], APOA1 (apolipoprotein A-1) [27] and some other proteins [28] may be potential biomarkers. Despite the advances in protein marker discovery, some disadvantages of protein markers are still existing, like genetic heterogeneity of tumors, poor reproducibility of laboratory test and low concentration of the proteins [18, 29]. Recent years the next-generation sequence technologies have promoted the study of disease-related genomes. Projects like The Cancer Genome Atlas (TCGA) [30] and the Genotype-Tissue Expression (GTEx) [31] have collected a large number of sequencing experiments and provided tissue-specific gene expression data in public. As some genes have distinct expression levels between normal and tumor tissues for the reason of disease development, they are promising to diagnose lung cancer more timely and accurately.

During the past years, gene differential expression analysis (DEA) has been extensively applied in the preprocess of high-throughput profiling data collected from microarrays [32–34]. Based on statistical models, researchers developed tools to identify genes which had distinct expression levels between different experiment groups. Compared with the microarray data, the RNA-seq raw data comes with the unique feature of discrete reads which should be analyzed under an appropriate statistical hypothesis [35]. According to the statistical hypothesis, the existing RNA-seq analysis models can be categorized into Poisson model, negative binomial model, beta-binomial model, and Bayesian model [36, 37]. These models can tell whether the gene expression levels are the same between experiment groups and calculate a confidence coefficient scores (also named p-value) suggesting the magnitude of expression difference.

In cancer studies, the histologically normal tissue adjacent to tumor is usually used to compare with the tumor tissue under the assumption that they are the same with real healthy tissues This approach allows researchers to compare samples from the same patient and reduce the individual specific effects. However, recent studies have deepened our understanding about NAT tissue, indicating that NAT is not exactly equal to the real healthy tissue [38]. In NAT tissues, the specific micro-environment surrounding tumor makes the change of gene expression in various pathways that are related to disease development. In order to identify efficient and meaningful marker genes, we proposed to detect differentially expressed genes(DEGs) from real normal, NAT and tumor tissues.

Here, we present a novel approach to identify genes markers for lung cancer with kernel maximum mean discrepancy (MMD) and Information Entropy. As mentioned above, the conventional DEA methods can calculate a p-value to evaluate the expression difference based on certain statistical hypothesis, but it's hard to decide which distribution assumption is correct before calculation. Inspired by the distribution measure method of transfer learning, we use the kernal MMD to detect DEGs between tumor, NAT and normal tissues. This method can output the maximum mean discrepancy score which indicates the degree of differential expression which does not require a statistical hypothesis on data distribution. Besides, although the p-value of conventional techniques can identify DEGs, it is essential to define a threshold of expression level to distinguish different types of tissue. Commonly, Researchers would like to take the upper boundary of lower expressed tissue or lower edge of higher expressed tissue as the threshold when there is a distinct expression gap. But this kind of gap is not always existing and then the threshold is hard to define. As the gene expression level is continuous data and how to choose a definite threshold point is a tough task. Here we applied the information theory to solve this problem.

In this paper, we firstly evaluate the expression level difference of 23368 genes in normal, normal adjacent tumor and tumor tissues with the kernel maximum mean discrepancy. Then the top-ranked genes selected by kernel MMD method are compared with genes selected by two conventional DEA methods, t-test and fold change. Then GO and KEGG pathway enrichment analysis are conducted to analyze the top 100 genes ranked by average MMD scores. Lastly, the top 10 genes are selected as marker genes for lung cancer and their expression boundaries between normal, NAT and tumor tissues are identified by the proposed information theory method.

Methods

Dataset

Three gene expression datasets used in this paper are collected from different tissue types in reference [38], containing the expression data of 23368 genes. Dataset 1 includes the gene expression data of 373 normal healthy samples. The raw reads file of dataset 1 is obtained from the GTEx program (phs000424.v6.p1, 18 November 2015). Dataset 2 has 59 NAT tissues, while dataset 3 has 541 lung cancer tumor tissues. Their raw feature counts

and FPKM values are original from NCBI Gene Expression Omnibus (GEO) [39]. Since the raw values are from different data sources, the RNA-sequencing raw reads files were processed and normalized with the Rsubread package and aligned to the UCSC hg19 reference genome with the same pipeline. The processed GTEx expression profiles of dataset 1 are available in GEO under an accession number GSE86354 and other two datasets are deposited as GSE62944.

Gene marker identification framework

With the above three datasets, we apply a novel approach to detect DEGs and determine the expression boundaries between normal, NAT and tumor cells as the criterion of lung cancer diagnosis.

In our method, there are mainly four steps: First, the kernel Maximum Mean Discrepancy is used to identify DEGs between two types of tissues respectively and genes are ranked by the MMD values; Second, the genes with top average MMD rankings are selected from all genes; Third, genes selected from the previous step are put into KEGG pathway analysis and GO enrichment analysis to validate the efficiency of those gene markers; Last, we define the gene expression boundaries for the top 10 marker genes with information gain theory. The whole framework of the proposed approach is illustrated in Fig. 1.

Kernel maximum mean discrepancy

The problem of comparing the probability distribution between two sample groups, also referred to as two-sample problem, widely exists in data science areas. In bioinformatics field, this problem is extensively existing in micro-array data analysis, database attribute matching, data integration from different platforms and so on. The key point of two-sample problem is how to determine if two groups of observations are from the same distribution and some statistical test methods were applied to address that in previous researches.

However, these methods have different statistical modelings based on specific assumptions of data distribution, which is commonly unknown before calculation in practical use. In some previous studies, researchers have explored to using the kernel Maximum Mean Discrepancy (MMD) method to test the distribution difference in RNA-Transcript expression and pathway differential expression and achieved better performance than traditional statistical tests [40, 41].Here, we adopt kernal MMD to identify the DEGs in lung cancer gene expression data.

Give F to be a class of functions $f : \chi \rightarrow R$. Two samples $X = \{x_1, x_2, \ldots x_m\}$ and $Y = \{y_1, y_2, \ldots y_n\}$ are drawn form two probability distribution p and q, respectively. The empirical estimation of MMD value is as following [42]:

$$MMD[F, p, q] := \sup_{f \in F} \left(E_p[f(x)] - E_q[f(y)] \right) \tag{1}$$

$$MMD[F, p, q] := \sup_{f \in F} \left(\frac{1}{m} \sum_{i=1}^{m} F(x_i) - \frac{1}{n} \sum_{i=1}^{n} F(y_i) \right) \tag{2}$$

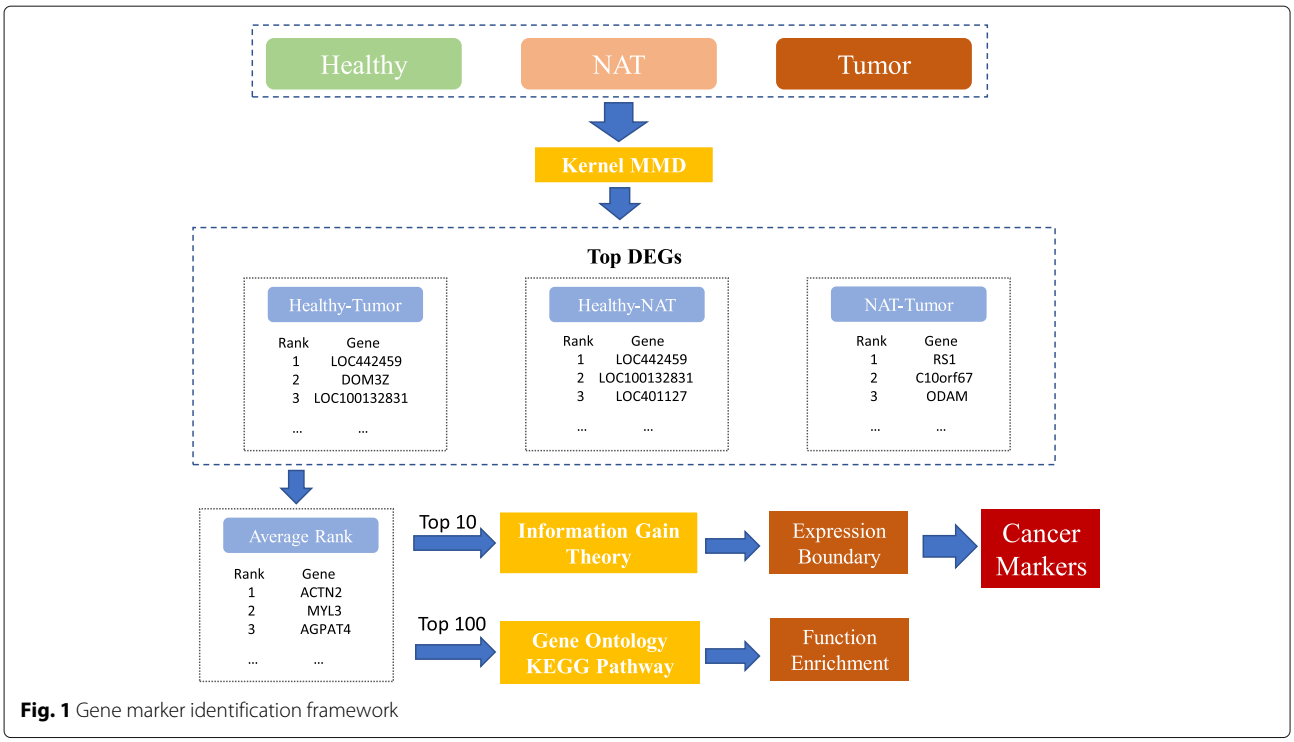

Fig. 1 Gene marker identification framework

As the definition above, if the function class F is rich enough, the value of MMD will be zero if and only if $p=q$. But a too rich F will lead to that MMD differs from zero for most finite sample estimates. Thus some restrictions ought to be placed on the function class. One trade-off way is to set F as the unit ball in a universal reproducing kernel Hilbert space H, defined on the compact metric space χ. Since H is a complete inner product space of functions $f : \chi \rightarrow R$, the function mapping $f \rightarrow f(x)$ can be expressed as an inner product via $f(x) = \langle f, \phi(x) \rangle_H$,where $\phi : \chi \rightarrow H$ is the feature space map from x to H. Then MMD can be rewritten as:

$$\begin{aligned} MMD[F,p,q] &= \sup_{\|f\|_H \leq 1} E_p[f(x)] - E_q[f(y)] \\ &= \sup_{\|f\|_H \leq 1} E_p\left[\langle f, \phi(x) \rangle_H\right] - E_q\left[\langle f, \phi(y) \rangle_H\right] \\ &= \sup_{\|f\|_H \leq 1} \langle \mu_p - \mu_q, f \rangle_H \\ &= \|\mu_p - \mu_q\|_H \end{aligned} \tag{3}$$

Then we can calculate like the following function:

$$\begin{aligned} MMD^2 &= \langle \mu_p - \mu_q, \mu_p - \mu_q \rangle_H \\ &= \langle \mu_p, \mu_p \rangle_H + \langle \mu_q, \mu_q \rangle_H - 2\langle \mu_p, \mu_q \rangle_H \\ &= E_p \langle \phi(x), \phi(x') \rangle_H + E_p \langle \phi(y), \phi(y') \rangle_H \\ &\quad - 2E_{p,q} \langle \phi(x), \phi(y) \rangle_H \end{aligned} \tag{4}$$

As the inner product can be replaced by Gaussian kernel $k(x,x') = exp(-\|x - x'\|^2 / (2\sigma^2))$,the value of MMD^2 can be figured out as:

$$\begin{aligned} MMD^2 &= \frac{1}{m(m-1)} \sum_{i \neq j}^{m} k(x_i, x_j) - \frac{2}{mn} \sum_{i,j=1}^{m,n} k(x_i, y_j) \\ &\quad + \frac{1}{n(n-1)} \sum_{i \neq j}^{m} k(y_i, y_j) \end{aligned} \tag{5}$$

In our method, the minimum variance unbiased estimate of MMD value is obtain according to the above functions based on Shogun package in python [43]. The computational complexity of MMD method is $O(n^2)$. The MMD score can evaluate the gene expression difference between different sample types, while a higher MMD score means greater gene expression level difference.

Boundary discovery method

As a biomarker, there should be an expression threshold for the marker gene as the indicator for disease diagnosis. If the gene expression level is proved to be different in normal and tumor tissues, it is necessary to define a threshold of expression level as the boundary. When the gene expression level has a distinct gap between normal and tumor samples, the threshold is commonly the lower or upper boundary of this gap. However, the expression level does not have that kind of obvious gap all the time, thus how to define a reliable boundary is challenging in these cases.

Here we propose to identify the threshold with information theory which has been widely used in decision tree algorithms for classification problems. According to the information theory, the change of information entropy which is also named information gain can evaluate the classification efficiency of a threshold point. If there is the expression data of a gene from m normal samples and n tumor samples in dataset D, p_m and p_n refer to the proportions of normal and tumor samples in all samples, then the original entropy of D is defined as:

$$Ent(D) = -\sum_{k=m,n} p_k \log_2 p_k \tag{6}$$

In the boundary identification, all samples are reclassified by the gene expression level with a split point of x and D^v denotes the new dataset re-classed by x. Then the information gain of this split point can be computed as:

$$Gain(D,x) = Ent(D) - \sum_{v=1}^{2} \frac{|D^v|}{|D|} Ent\left(D^v\right) \tag{7}$$

Different from discrete data, the expression level is continuous and it is inappropriate to use the expression level values in samples as the split points. Besides, as the distribution of the expression level is also unknown, we cannot use the probability function to calculate the entropy. To address this problem, we propose to deal with continuous data like discrete data: First, the expression level values are sorted from small to large and the middle points between two expression level values are taken as the split points; Second, we calculate the information gain of the split points respectively and choose the point that has the highest information gain as the boundary. The algorithm of expression boundary identification with information theory is illustrated in Algorithm S1 in Additional file 2.

GO and KEGG enrichment analysis

The GO enrichment analysis is the major gene-annotation analysis method based on the Gene Ontology resource, describing the gene function at a molecular level. The Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway enrichment analysis has been widely used to model and simulate the molecular interactions and reaction networks in system biology. In this paper, these two methods

are applied to figure out the molecular functions of identified potential marker genes and validate whether these genes are related to lung cancer. Here the enrichment analysis methods are both implemented based on the R package called ClusterProfiler developed by Guangchuang Yu's team [44]. The GO terms and enriched pathways are all filtered with the p-value <0.05.

Conventional DEA method and machine learning evaluation

In this work, two conventional differentially expressed gene analysis methods, t-test and fold change, are compared with the proposed kernel MMD. The t-test is completed based on a python package called 'Scipy' [45]. The fold change is calculated as below:

$$FoldChange = \left| log_2(\frac{E1}{E2}) \right| \tag{8}$$

Where $E1$ and $E2$ are the average of gene expression level in two different issue types. The p-value, fold change value and MMD score are calculated for every single gene in our datasets. Then genes are ranked with the same strategy and top ranking genes are regarded as potential markers. Here a 10-fold cross validation based on the random forest classifier is applied to evaluate the efficiency of these top genes under four frequently used metrics: recall, F1-score, accuracy and Matthews correlation coefficient (MCC).

Results

In the first part, we present the genes ranking with kernel MMD score and analysis the gene expression difference between different issue types. Then the top ranked genes are reported and compared with those genes identified by conventional t-test and fold change methods. The third part shows the results of GO and KEGG pathway analysis of the top ranked genes. At last, we choose the top ten genes of average ranking as marker gene and identify the expression boundaries of these gene markers with information gain theory.

Gene differential expression between different tissue types

For the three mentioned datasets, kernel MMD values are calculated on each two of them respectively to discover DEGs. For every single gene, we calculate three MMD values which are from Normal-NAT, Normal-Tumor and NAT-Tumor groups. The MMD scores indicate the difference of expression levels among three types of samples. The top 10 ranked genes in each group are shown in Table 1. As illustrated in the table, the top MMD scores in Normal-Tumor group are over 200, which are much higher than the other two groups. The Normal-NAT group has comparable MMD scores with NAT-Tumor group. It is clear that gene expression level difference in normal-tumor group is much greater than other two groups.

In addition, the NAT samples have different expression profiles from not only tumor samples but also the real healthy samples. Since the NAT samples are always considered as healthy samples in the state-of-art researches, we test the top 10 ranked genes selected by NAT-Tumor group, Normal-Tumor group and their average ranking to explore the influence of regarding NAT as real normal samples. To evaluate the effectiveness of selected genes, the expression data of the top genes above is applied to classify tumor samples from other samples via 10-fold cross-validation. The results of the 10-fold cross validation are reported in Table 2.

As shown in Table 2, the selected genes from each group can classify tumor samples from other samples. However, the performance of the three groups of genes varies greatly. When considering normal samples and NAT samples together, the top average ranked genes have the best scores under all metrics with an accuracy of 0.9907. The highest F1 score of 0.9914 implies that these genes also have a better classification balance. The results show that

Table 1 Top ranking differentially expressing genes between each two tissue types (NAT: Normal Adjacent Tumor)

Ranking	Normal-NAT	MMD scores	Normal-Tumor	MMD scores	NAT-Tumor	MMD scores
1	LOC442459	81.56	LOC442459	300.89	RS1	67.06
2	DOM3Z	70.85	LOC100132831	293.86	C10orf67	58.85
3	LOC100132831	68.89	LOC401127	288.92	ODAM	57.90
4	LOC401127	67.45	PIN1P1	265.11	LOC100128164	57.16
5	CSNK1A1P1	67.02	CSNK1A1P1	264.75	SH3GL3	56.96
6	MKRN9P	66.54	WNT2B	248.53	JPH4	56.68
7	TPI1P2	65.14	LOC100287632	247.69	SGCG	56.56
8	CYP2D7P1	64.72	CSNK1A1L	247.45	GYPE	55.70
9	CSNK1A1L	63.69	LOC100507373	244.54	LOC643650	53.05
10	PIN1P1	62.24	AOC4	240.66	IHH	52.79

Table 2 Cross Validation Performance of Top Ten genes from different groups (NAT: Normal Adjacent Tumor)

Group	Recall	F1	Accuracy	MCC
Normal-Tumor	0.9857	0.9540	0.9476	0.8659
NAT-Tumor	0.9534	0.9670	0.9640	0.9279
Average	0.9885	0.9914	0.9907	0.9816

the real normal samples and NAT samples are not exactly the same. Researchers should take both of them into consideration in cancer study rather than simply replacing real normal samples with NAT samples. The detailed results of differential expressing gene identification conducted by MMD and other two conventional methods are listed in Additional file 1.

Identify marker genes for lung cancer development

In this work, two conventional DEA methods t-test and fold change are compared with our approach. T-test and fold change methods are both applied to identify DEGs between different tissue types. The p-value of t-test and fold change value are calculated to evaluate the gene expression difference. Since the ability to detect tumor samples is more significant in clinical application, the top 10 genes of average rankings from Normal-Tumor group and NAT-tumor group selected by t-test and fold change are compared with the genes selected by our method. Another 10-fold cross validation is conducted and the results are reported in Table 3.

As shown in Table 3, the proposed kernel MMD method outperforms other two conventional methods under all metrics with the recall of 0.9885, F1 score of 0.9914, accuracy of 0.9907 and MCC of 0.9816. The fold change method has the worst performance and the selected genes by fold change method are not efficient enough to classify tumors from other samples. The t-test has a comparable result with MMD method. Since the t-test and fold change methods have been widely used, the kernel MMD method is promising to improve the differential gene analysis efficiency in practical use.

From Table 1, we can see there are some overlapping genes like LOC442459, LOC100132831, LOC401127, CSNK1A1P1, CSNK1A1L and PIN1P1 in Normal-NAT group and Normal-Tumor group. These genes can distinguish normal samples from not only NAT samples, but also tumor samples. Inspired the previous part, the average ranking of all groups can help to identify more significant genes. Thus, the gene average ranking of the three groups is calculated and top genes of average ranking are chosen to be potential marker genes to diagnose lung cancer. In Fig. 2, expression levels in normal, NAT and tumor samples of the top 4 genes of average ranking are presented. From the figure, the four genes exactly have distinct expression levels in different types of tissues.

Table 3 Cross Validation Performance of top ten genes selected by different DEA methods

Method	Recall	F1	Accuracy	MCC
Fold Change	0.7044	0.7992	0.8048	0.6382
T-test	0.9796	0.9815	0.9794	0.9582
Kernel MMD	0.9885	0.9914	0.9907	0.9816

GO and KEGG pathway enrichment

From the average ranking gene list, we choose the top 100 genes to conduct the GO and KEGG pathway enrichment analysis. In the GO enrichment analysis, we select 'Biology Process' as the enrichment target, and there are 12 GO terms with p-value <1.0e-04 and count ≥5. As shown in Table 4, the top two terms, 'GO:0051480' and 'GO:0007204', are both related to the regulation factors of cytosolic calcium ion concentration while term No.5 and No.6 are also involved in cellular calcium ion homeostasis. The influence of calcium ion channels on lung cancer has been studied for a long time [46–48], and the cellular calcium ion level change has been explored in lung cancer development [48]. It is suggested that these calcium ion regulation related genes are significant in lung cancer.

The results of KEGG pathway enrichment analysis are illustrated in Fig. 3. There are 20 pathways with a p-value below 0.05 and count number over 2. The adrenergic signaling pathway and the cGMP-PKG signaling pathway are the most significant pathways. Currently, the role of adrenergic signaling pathways plays in lung cancer development have not been fully studied. However, the β-adrenergic signaling have been found to be a possible novel cancer therapy in tumor cells [49]. Besides, some researches have made some explorations about that [50]. The second top significant pathway is the cGMP-PKG signaling pathway which mediates the action of cellular ion concentration and sensitivity, influencing cell proliferation. The regulation relationship between cGMP-PKG signaling pathway and lung cancer has been studied in [51]. The results of GO and KEGG pathway enrichment analysis show that the top gene selected by MMD method is indeed highly related to lung cancer.

Expression boundary identification

Although the conventional methods can detect the differential expressed gene, they can only manually define the expression boundary when there is a distinct expression level gap. After selecting lung cancer marker genes, we identify the expression boundaries between normal,NAT and tumor with the mentioned information theory method. Here the top 10 genes in MMD average ranking list are chosen as the lung cancer marker genes and the expression boundaries of them are illustrated in

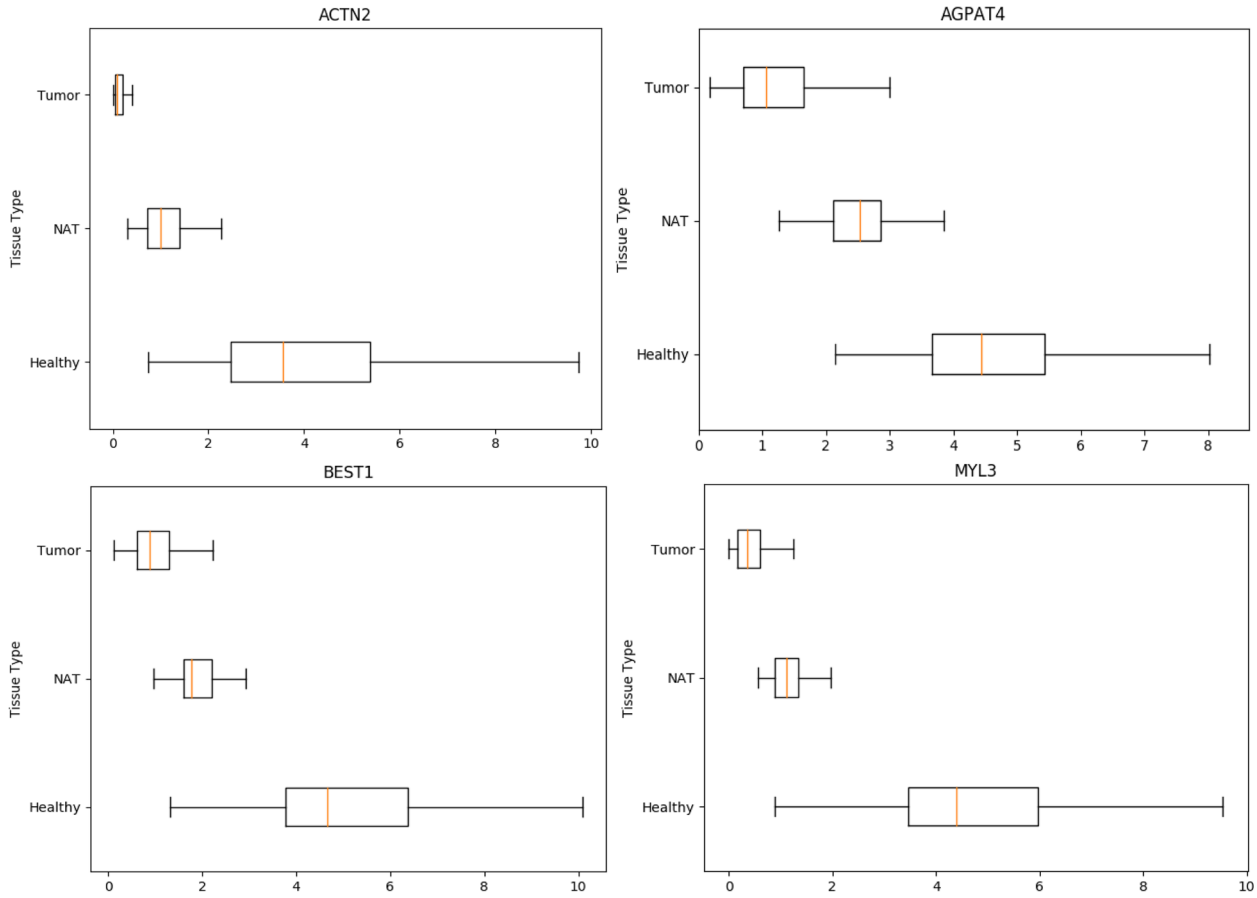

Fig. 2 Box-plot of Gene Expression levels in three tissue types.The X-axis is the FPKM expression level; the Y-axis is the tissue type

Table 5. The identified expression boundaries of all genes are reported in Additional file 1.

As shown in Table 5, the ten gene markers have a distinct expression range in normal, NAT and tumor samples, which can be an indicator of lung cancer development. Additionally, in practical clinic application, the boundary between tumor and other tissues is the most significant for disease diagnosis.The boundary between normal samples and NAT samples also implied that there would be some gene expression changes in the disease

Table 4 Go Function analysis for the top ranking genes(*p*-value $<$1.0e-04 and count $\geq$5)

No.	GOBPID	*p*-Value	Count	Term
1	GO:0051480	7.6032e-07	10	regulation of cytosolic calcium ion concentration
2	GO:0007204	3.0453e-06	9	positive regulation of cytosolic calcium ion concentration
3	GO:0019229	4.4969e-06	5	regulation of vasoconstriction
4	GO:0007200	6.6689e-06	6	phospholipase C-activating G-protein coupled receptor signaling pathway
5	GO:0006874	7.5060e-06	10	cellular calcium ion homeostasis
6	GO:0055074	9.4074e-06	10	calcium ion homeostasis
7	GO:0042310	1.4462e-05	5	vasoconstriction
8	GO:0072503	1.5632e-05	10	cellular divalent inorganic cation homeostasis
9	GO:0072507	2.1785e-05	10	divalent inorganic cation homeostasis
10	GO:0097756	2.3563e-05	5	negative regulation of blood vessel diameter
11	GO:0007189	6.5898e-05	5	adenylate cyclase-activating G-protein coupled receptor signaling pathway
12	GO:0019932	7.4403e-05	8	second-messenger-mediated signaling

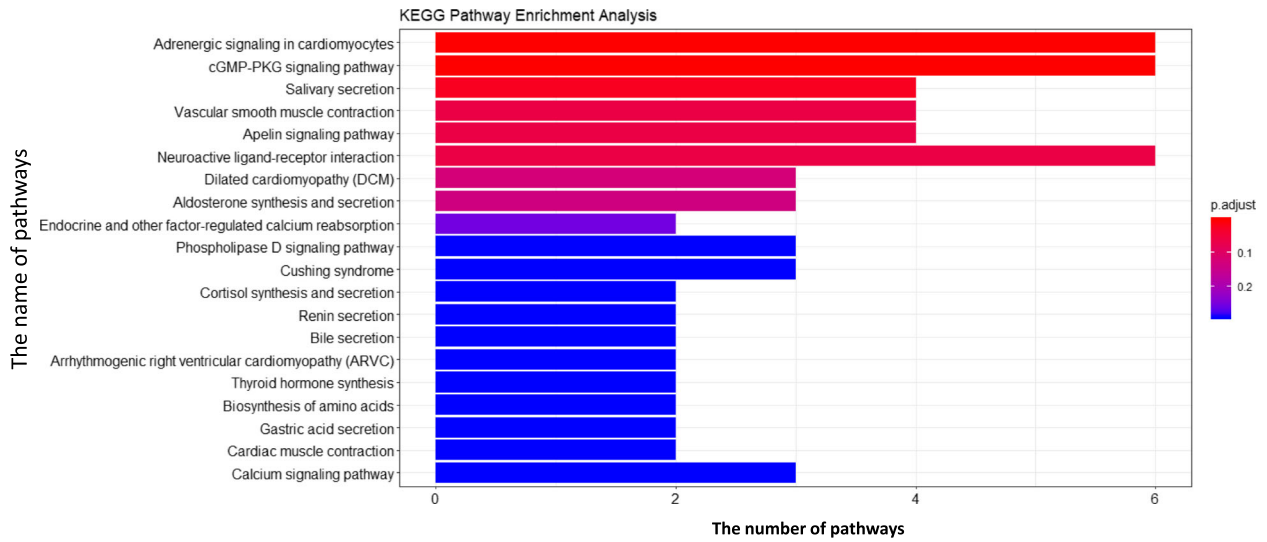

Fig. 3 KEGG pathway enrichment analysis for top ranking genes

development and the NAT samples may serve to detect cell carcinogenesis, which can help to understand the lung cancer mechanisms.

Discussion

Since the early-diagnosis of lung cancer has been a long-term critical problem in clinical practice, researchers have explored various types of biomarkers, like genetic mutations, blood proteins. Here, this paper proposed a novel method to identify genes markers for lung cancer. There are two main problems in efficient gene markers identification: first, how to evaluate the gene expression difference; second, how to find the reliable expression boundary between tumor and other samples. The most existing DEA methods were built to solve the first problem, but they can only give out a p-value to assess the differential expressing gene without defining the expression boundary. The ln of this research is to address both of the problems in biomarker identifications.

The gene markers are given out based on the existing lung cancer dataset. We think there are two limitations in our work. First, a larger dataset can help to obtain more accurate results; Second, a threshold of MMD value to define the differentially expressed gene can be defined with a large dataset, while here we just take the top ranked genes as potential marker genes.

Table 5 Expression Boundary of Lung Cancer Biomarkers(e : FPKM expression level)

Gene Name	Normal	Normal Adjacent Tumor	Tumor
ACTN2	$e \geq 3.5247$	$0.7146 < e < 3.5247$	$e \leq 0.7146$
MYL3	$e \geq 5.3223$	$4.9211 < e < 5.3223$	$e \leq 4.9211$
AGPAT4	$e \geq 4.3722$	$2.3052 < e < 4.3722$	$e \leq 2.3052$
BEST1	$e \geq 3.7487$	$1.6216 < e < 3.7487$	$e \leq 1.6216$
TWIST2	$e \geq 4.2450$	$1.2030 < e < 4.2450$	$e \leq 1.2030$
LINC00472	$e \geq 3.4721$	$0.8045 < e < 3.4721$	$e \leq 0.8045$
MYO7B	$e \geq 4.1723$	$0.7450 < e < 4.1723$	$e \leq 0.7450$
CCNF	$e \leq 16.4506$	$16.4506 < e < 20.5656$	$e \geq 20.5656$
NECAB1	$0.9961 < e < 4.7770$	$e \geq 4.7770$	$e \leq 0.9961$
NOTCH4	$e \geq 4.6829$	$1.9808 < e < 4.6829$	$e \leq 1.9808$

Conclusion

In this paper, we not only proposed a more efficient method, kernel MMD, to evaluate the expression changes, but also provide a information theory based algorithm to identify the gene expression boundary. The experiment results show our method can select more significant genes than traditional methods and give out the expression boundary of the marker gene. Through the GO and KEGG pathway enrichment analysis, the function of marker genes in lung cancer is studied, and these marker genes are indeed related to lung cancer development. In the future, we will collect more gene expression data related to lung cancer and calculate more accurate results. In addition, we will explore the application of our method on biomarker discovery for other diseases.

Supplementary information

Supplementary information accompanies this paper at https://doi.org/10.1186/s12920-019-0630-4.

Additional file 1: The results of DEA and boundary identification. In Additional file 1, the results of MMD, t-test and fold change analysis between Normal-NAT, Normal-Tumor and NAT-Tumor are reported. The mmd score, p-value and fold change score for every single gene are all presented. Besides, the boundary identification results are also included in this file.

Additional file 2: The supplementary files. In Additional file 2, the gene differential expression boundary identification algorithm is included.

Abbreviations
DEA: Differentially expressed analysis; GO: Gene ontology; KEGG: Kyoto encyclopedia of genes and genomes; MMD: Maximum mean discrepancy; NAT: Normal adjacent to the tumor

Acknowledgements
This research work was partially supported by the China Scholarship Council (CSC).

About this supplement
This article has been published as part of *BMC Medical Genomics, Volume 12 Supplement 8, 2019: 18th International Conference on Bioinformatics*. The full contents of the supplement are available at https://bmcmedgenomics.biomedcentral.com/articles/supplements/volume-12-supplement-8.

Authors' contributions
ZZ: Collected data, proposed predictor model, designed experiment and wrote the paper. HP: Designed experiment, reviewed and edited paper. XZ: Reviewed and edited the paper. YZ: Data process and model optimization. LF: reviewed and edited the paper. FC: reviewed and edited the paper. JL: Designed experiment, reviewed and edited paper. All authors have read and approved the manuscript.

Funding
This research is funded by Australia Research Council(Grant Number: DP 180100120). The publication cost of this supplement was funded by Faculty of Engineering and Information Technology, University of Technology Sydney.

Availability of data and materials
The datasets generated and/or analyzed during the current study are available in NCBI Gene Expression Om-nibus (https://www.ncbi.nlm.nih.gov/geo/) under the accession number GSE86354 and GSE62944. The source code of our method is freely available at https://github.com/Zhixun-Zhao/GeneMarker.

Ethics approval and consent to participate
Not applicable. All utilized data sets are publicly and freely available which do not require any ethics approval and consent to participate.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Author details
[1]Advanced Analytics Institute, Faculty of Engineering and Information Technology, University of Technology Sydney, PO Box 123, Broadway, 2007 Sydney, NSW, Australia. [2]Faculty of Engineering and Information Technology, University of Technology Sydney, PO Box 123, Broadway, 2007 Sydney, NSW, Australia. [3]School of Computer, National University of Defense Technology, 410073, Changsha, China.

Received: 29 October 2019 Accepted: 18 November 2019
Published: 20 December 2019

References
1. Schnabel P, Junker K. Pulmonary neuroendocrine tumors in the new who 2015 classification: Start of breaking new grounds?. Der Pathologe. 2015;36(3):283–92.
2. Parkin DM. Global cancer statistics in the year 2000. Lancet Oncol. 2001;2(9):533–43.
3. Minna JD, Roth JA, Gazdar AF. Focus on lung cancer. Cancer Cell. 2002;1(1):49–52.
4. Jemal A, Siegel R, Ward E, Murray T, Xu J, Smigal C, Thun MJ. Cancer Statistics, 2006. CA Cancer J Clin. 2006;56(2):106–30.
5. Fontana RS, Sanderson DR, Taylor WF, Woolner LB, Miller WE, Muhm JR, Uhlenhopp MA. Early lung cancer detection: results of the initial (prevalence) radiologic and cytologic screening in the mayo clinic study. Am Rev Respir Dis. 1984;130(4):561–5.
6. Frost JK, Ball Jr WC, Levin ML, Tockman MS, Baker RR, Carter D, Eggleston JC, Erozan YS, Gupta PK, Khouri NF, et al. Early lung cancer detection: results of the initial (prevalence) radiologic and cytologic screening in the johns hopkins study. Am Rev Respir Dis. 1984;130(4):549–54.
7. Hussain A, Khatri M, Casali G, Batchelor T, West D. 194 follow up after lung cancer surgery: plain chest x ray does not increase diagnostic accuracy. Lung Cancer. 2014;83:72.
8. Capelozzi VL. Role of immunohistochemistry in the diagnosis of lung cancer. J Bras Pneumol. 2009;35(4):375–82.
9. Marshall HM, Bowman RV, Yang IA, Fong KM, Berg CD. Screening for lung cancer with low-dose computed tomography: a review of current status. J Thorac Dis. 2013;5(Suppl 5):524.
10. Vazquez MF, Koizumi JH, Henschke CI, Yankelevitz DF. Reliability of cytologic diagnosis of early lung cancer. Cancer Cytopathol Interdisc Int J Am Cancer Soc. 2007;111(4):252–8.
11. Jantus-Lewintre E, Usó M, Sanmartín E, Camps C. Update on biomarkers for the detection of lung cancer. Lung Cancer Targets Ther. 2012;3:21.
12. Rabinowits G, Gerçel-Taylor C, Day JM, Taylor DD, Kloecker GH. Exosomal microrna: a diagnostic marker for lung cancer. Clin Lung Cancer. 2009;10(1):42–6.
13. Mitas M, Hoover L, Silvestri G, Reed C, Green M, Turrisi AT, Sherman C, Mikhitarian K, Cole DJ, Block MI, et al. Lunx is a superior molecular marker for detection of non-small lung cell cancer in peripheral blood. J Mol Diagn. 2003;5(4):237–42.
14. Andre F, Schartz NE, Movassagh M, Flament C, Pautier P, Morice P, Pomel C, Lhomme C, Escudier B, Le Chevalier T, et al. Malignant effusions and immunogenic tumour-derived exosomes. Lancet. 2002;360(9329):295–305.
15. Montani F, Marzi MJ, Dezi F, Dama E, Carletti RM, Bonizzi G, Bertolotti R, Bellomi M, Rampinelli C, Maisonneuve P, et al. Mir-test: a blood test for lung cancer early detection. JNCI: J Natl Cancer Inst. 2015;107(6):.
16. Nagrath S, Sequist LV, Maheswaran S, Bell DW, Irimia D, Ulkus L, Smith MR, Kwak EL, Digumarthy S, Muzikansky A, et al. Isolation of rare circulating tumour cells in cancer patients by microchip technology. Nature. 2007;450(7173):1235.
17. Sozzi G, Boeri M, Rossi M, Verri C, Suatoni P, Bravi F, Roz L, Conte D, Grassi M, Sverzellati N, et al. Clinical utility of a plasma-based mirna signature classifier within computed tomography lung cancer screening: a correlative mild trial study. J Clin Oncol. 2014;32(8):768.
18. Sozzi G, Conte D, Leon M, Cirincione R, Roz L, Ratcliffe C, Roz E, Cirenei N, Bellomi M, Pelosi G, et al. Quantification of free circulating dna as a diagnostic marker in lung cancer. J Clin Oncol. 2003;21(21):3902–8.
19. Valenti R, Huber V, Filipazzi P, Pilla L, Sovena G, Villa A, Corbelli A, Fais S, Parmiani G, Rivoltini L. Human tumor-released microvesicles promote the differentiation of myeloid cells with transforming growth factor-β–mediated suppressive activity on t lymphocytes. Cancer Res. 2006;66(18):9290–8.
20. Doseeva V, Colpitts T, Gao G, Woodcock J, Knezevic V. Performance of a multiplexed dual analyte immunoassay for the early detection of non-small cell lung cancer. J Trans Med. 2015;13(1):55.
21. Goetsch CM. Genetic tumor profiling and genetically targeted cancer therapy, Vol. 27; 2011. p. 34–44. Elsevier.
22. Mizuguchi S, Nishiyama N, Iwata T, Nishida T, Izumi N, Tsukioka T, Inoue K, Uenishi T, Wakasa K, Suehiro S. Serum sialyl lewisx and cytokeratin 19 fragment as predictive factors for recurrence in patients with stage i non-small cell lung cancer. Lung Cancer. 2007;58(3):369–75.
23. Pujol J-L, Grenier J, Daurès J-P, Daver A, Pujol H, Michel F-B. Serum fragment of cytokeratin subunit 19 measured by cyfra 21-1 immunoradiometric assay as a marker of lung cancer. Cancer Res. 1993;53(1):61–6.
24. Okada M, Nishio W, Sakamoto T, Uchino K, Yuki T, Nakagawa A, Tsubota N. Effect of histologic type and smoking status on interpretation of serum carcinoembryonic antigen value in non–small cell lung carcinoma. Annals Thorac Surg. 2004;78(3):1004–9.
25. Kang S-M, Sung H-J, Ahn J-M, Park J-Y, Lee S-Y, Park C-S, Cho J-Y. The haptoglobin β chain as a supportive biomarker for human lung cancers. Mol BioSyst. 2011;7(4):1167–75.
26. Sung H-J, Cho J-Y. Biomarkers for the lung cancer diagnosis and their advances in proteomics. BMB Rep. 2008;41(9):615–25.

27. Maciel CM, Junqueira M, Paschoal MEM, Kawamura MT, Duarte RLM, Carvalho MdGdC, Domont GB. Differential proteomic serum pattern of low molecular weight proteins expressed by adenocarcinoma lung cancer patients. J Exp Ther Oncol. 2005;5(1):.
28. Indovina P, Marcelli E, Maranta P, Tarro G. Lung cancer proteomics: recent advances in biomarker discovery. Int J Proteomics. 2011;2011:.
29. Zamay T, Zamay G, Kolovskaya O, Zukov R, Petrova M, Gargaun A, Berezovski M, Kichkailo A. Current and prospective protein biomarkers of lung cancer. Cancers. 2017;9(11):155.
30. Tomczak K, Czerwińska P, Wiznerowicz M. The cancer genome atlas (tcga): an immeasurable source of knowledge. Contemp Oncol. 2015;19(1A):68.
31. Lonsdale J, Thomas J, Salvatore M, Phillips R, Lo E, Shad S, Hasz R, Walters G, Garcia F, Young N, et al. The genotype-tissue expression (gtex) project. Nat Genet. 2013;45(6):580.
32. Wang Z, Gerstein M, Snyder M. Rna-seq: a revolutionary tool for transcriptomics. Nat Rev Genet. 2009;10(1):57.
33. Marioni JC, Mason CE, Mane SM, Stephens M, Gilad Y. Rna-seq: an assessment of technical reproducibility and comparison with gene expression arrays. Genome Res. 2008;18(9):1509–17.
34. Bullard JH, Purdom E, Hansen KD, Dudoit S. Evaluation of statistical methods for normalization and differential expression in mrna-seq experiments. BMC Bioinformatics. 2010;11(1):94.
35. Soneson C, Delorenzi M. A comparison of methods for differential expression analysis of rna-seq data. BMC Bioinformatics. 2013;14(1):91.
36. Seyednasrollah F, Laiho A, Elo LL. Comparison of software packages for detecting differential expression in rna-seq studies. Brief Bioinform. 2013;16(1):59–70.
37. Rapaport F, Khanin R, Liang Y, Pirun M, Krek A, Zumbo P, Mason CE, Socci ND, Betel D. Comprehensive evaluation of differential gene expression analysis methods for rna-seq data. Genome Biol. 2013;14(9):3158.
38. Aran D, Camarda R, Odegaard J, Paik H, Oskotsky B, Krings G, Goga A, Sirota M, Butte AJ. Comprehensive analysis of normal adjacent to tumor transcriptomes. Nat Commun. 2017;8(1):1077.
39. Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, Marshall KA, Phillippy KH, Sherman PM, Holko M, et al. Ncbi geo: archive for functional genomics data sets—update. Nucleic Acids Res. 2012;41(D1):991–5.
40. Stegle O, Drewe P, Bohnert R, Borgwardt K, Rätsch G. Statistical tests for detecting differential RNA-transcript expression from read counts. Nat Precedings. 2010. https://doi.org/10.1038/npre.2010.4437.1.
41. Vegas E, Oller JM, Reverter F. Inferring differentially expressed pathways using kernel maximum mean discrepancy-based test. BMC Bioinformatics. 2016;17(5):205.
42. Gretton A, Borgwardt K, Rasch M, Schölkopf B, Smola AJ. A kernel method for the two-sample-problem: Advances in Neural Information Processing Systems; 2007, pp. 513–20.
43. Sonnenburg S, Henschel S, Widmer C, Behr J, Zien A, Bona Fd, Binder A, Gehl C, Franc V, et al. The shogun machine learning toolbox. J Mach Learn Res. 2010;11(Jun):1799–802.
44. Yu G, Wang L-G, Han Y, He Q-Y. clusterprofiler: an r package for comparing biological themes among gene clusters. Omics J Integr Biol. 2012;16(5):284–7.
45. Bressert E. SciPy and NumPy: an Overview for Developers. Sebastopol: O'Reilly Media, Inc.; 2012.
46. Moody TW, Murphy A, Mahmoud S, Fiskum G. Bombesin-like peptides elevate cytosolic calcium in small cell lung cancer cells. Biochem Biophys Res Commun. 1987;147(1):189–95.
47. Moody TW, Staley J, Zia F, Coy DH, Jensen RT. Neuromedin b binds with high affinity, elevates cytosolic calcium and stimulates the growth of small-cell lung cancer cell lines. J Pharmacol Exp Ther. 1992;263(1):311–7.
48. Arbabian A, Brouland J-P, Apáti Á, Pászty K, Hegedűs L, Enyedi Á, Chomienne C, Papp B. Modulation of endoplasmic reticulum calcium pump expression during lung cancer cell differentiation. FEBS J. 2013;280(21):5408–18.
49. Schuller HM. Beta-adrenergic signaling, a novel target for cancer therapy?. Oncotarget. 2010;1(7):466.
50. Schuller HM, Cekanova M. Nnk-induced hamster lung adenocarcinomas over-express β2-adrenergic and egfr signaling pathways. Lung Cancer. 2005;49(1):35–45.
51. Wong JC, Bathina M, Fiscus RR. Cyclic gmp/protein kinase g type-iα (pkg-iα) signaling pathway promotes creb phosphorylation and maintains higher c-iap1, livin, survivin, and mcl-1 expression and the inhibition of pkg-iα kinase activity synergizes with cisplatin in non-small cell lung cancer cells. J Cell Biochem. 2012;113(11):3587–98.

Publisher's Note