Literature Data Mining Based Enrichment Analysis on 1,925 Genes for Lung Cancer

Xinming Dong [1], McKenzie Ritter [2], Hongbao Cao [3*], Dexiang Yang [4*]

[1] Tianjin Sanatorium, Tianjin 300191, P.R. China;

[2] Unit on Statistical Genomics, National Institute of Mental Health, NIH, Bethesda, MD 20852, USA;

[3] Elsevier Inc., Biology Prod Research, Rockville, MD 20852, USA;

[4] Respiratory Department, People's Hospital of Tongling, Tongling, Anhui 244000, P.R. China.

***Corresponding Author:** Dr. Yang, Department of Respiratory, the People's Hospital of Tongling, Anhui 244000, P.R. China. Email: drr11154@rjh.com.cn. or Dr. Cao, Elsevier Inc., Biology Prod Research, Rockville, MD 20852, USA. Email: h.cao@.com.

http://mo.qingres.com

OPEN ACCESS

DOI: 10.20900/mo.20160006

Received: January 8, 2016

Accepted: February 12, 2016

Published: April 25, 2016

ABSTRACT

Background: Approximately 8 % of all lung cancer is due to inherited factors. Risk more than doubles for those with relatives who have had lung cancer. To date, numerous genetic studies have found large number of genes that are related to lung cancer. Most of the studies focused on separate activities of genes that influence disease development.

Methods: Literature data mining (LDM) of over 17,884 articles for publications from 1978 to February 2016 was conducted. The articles reported multiple types of marker-disease associations between 1,925 genes and lung cancer. A gene set enrichment analysis (GSEA) and a sub-network enrichment analysis (SNEA) was performed to discover functional profiles and validate pathogenic significances of these genes to lung cancer. Finally, a network connectivity analysis (NCA) was performed to study associations between the reported genes.

Results: The reported genes demonstrate multiple types of association with lung cancer. Results from the enrichment analysis confirm the reports and suggest that these genes play significant roles in lung cancer pathogenesis, as well as in the pathogenesis of other lung cancer-related disorders. NCA results demonstrate that these genes, especially those with high RScores, present strong functional associations with each other.

Conclusion: The results suggest that lung cancer genetic causes are

linked to a network composed of a large number of genes. LDM together with enrichment and network analysis could serve as an effective approach in finding these potential target genes.

1 INTRODUCTION

Lung cancer is a malignant tumor, characterized by uncontrolled cell growth in lung tissues. Approximately 10-15 % of the cases occur in those who have never smoked [1]. These cases are often caused by a combination of genetic factors and exposure to radon gas, asbestos, second-hand smoke, or other forms of air pollution. It is estimated that inherited factors alone account for about 8 % of all lung cancer cases [2]. It is believed that the genetic causes are a combination of multiple genes and polymorphisms on chromosomes 5, 6, and 15, which are known to affect lung cancer risk [3].

Recently, there has been a number of articles reporting nearly 2,000 genes/proteins that relate to lung cancer. Many are suggested as disease biomarkers. Most these studied separate gene/protein activities. Based on how the gene-lung cancer relations were reported, these articles can generally be classified into the following categories: 1) biomarker; 2) clinical trial; 3) genetic change; 4) quantitative change; 5) regulation; and, 6) state change.

Biomarkers refer to proteins/genes that have been identified as either prognostic for, or diagnostic of, the disease. Relatively few articles have claimed that the genes investigated in their study could serve as disease biomarkers [4-9]. The observations have been inconsistent [10]. Groen *et al.* conclude ed that PTGS2 expression in patients with advanced non-small-cell lung cancer was neither a prognostic, or predictive, marker for treatment with celecoxib [11].

For a number of reasons, including expense and ethical issues, relatively few clinical trials have been conducted to study the relationships between these genes and lung cancer [12, 13]. Many studies, including independent studies and meta-analyses, have reported a genetic change in these genes in the case of lung cancer [14-22]. Mutation changes of these genes have demonstrated sub-group sensitivities [23-26]. Sasaki *et al.* showed that EGFR mutations were found in only 63 of 575 lung cancer patients [27]. This limits the use of these genes as biomarkers for diagnosis and treatment.

Quantitative change refers to changes in the abundance, activity, and/or, expression of a gene/protein in a disease state. Most reports for this type of relationship have come from gene expression studies, in which many genes were observed to demonstrate increased activity, gene and/or expression levels in lung cancer including: EGFR, CYP1A1, ALK, ROS1, ERBB2, MET, KEAP1, VEGF, PTGS2, TERT, and TP53 [28-36]. Some genes showed decreased activity, such as: GSTM1, GSTT1, ERCC1, and KRAS [37-39]. In a manner similar to to genetic changes, observed quantitative changes demonstrate case sensitivity among lung cancer patients [40, 41].

Regulation refers to changing the activity of the target by an unknown mechanism. This type of relationship is usually equivocal in describing the mechanism of the association [42-45]. Some of the studies suggested the mechanisms of the genes for lung cancer [46-49]. In addition, several revealed functional correlations between different genes and genetic factors [50, 51].

State change refers to changes in a protein/gene post-translational modification status, or alternative splicing events, associated with a disease. Only a few papers reported gene state changes in lung cancer cases [52, 53]. These studies reveal specific protein/gene state changes that may relate to lung cancer and are important for the understanding the disease mechanism.

No systematic analysis, to our knowledge, has evaluated the quality, and strength, of these reported genes as a single functional network/group in a study of the underlying biological processes of lung cancer. This study, rather than focusing on a specific marker, or function, attempts to provide a fuller view of the genetic-map related to lung cancer.

2 METHODS AND MATERIALS

This study is structured as follows: 1) Literature data mining (LDM) to discover gene-lung cancer relations; 2) Enrichment analysis on the genes identified to validate their pathogenic significance to lung cancer; and, 3) Network connectivity analysis (NCA) to test the functional association between these reported genes.

2.1 Literature Data Mining

A literature data mining (LDM) was performed for all articles available on the Pathway Studio database

(www.pathwaystudio.com) until Februay 2016. This covers over 40 million scientific articles. It sought those that reported gene-lung cancer relations. It was conducted by employing a finely-tuned, Natural Language Processing (NLP) system of Pathway Studio software, which purports to be able to identify and extract relationship data from scientific literature. Only those publications containing a biological interaction defined by ResNet Exchange (RNEF) data format were included (http://www.gousinfo.com/AIC%20project/Pathway%20Studio/Elsevier%20RNEF-1.3.htm). Results are presented, including a complete gene name list, underlying article information, and the marker scores, which are described below.

2.2 Quality Metric Analysis

A quality metrics analysis was performed on all marker-disease relations. Analysis output includes quality score (QScore), citation score (CScore), novelty score (NScore), and report frequency score (RScore) at the article, as well as the marker level. These quality measures can be used to sort the marker list to obtain those with the most significance.

Using the RScore, the most frequently reported markers can be identified. At the article level, RScore = 1, indicates a marker-disease relation has been reported. If there is no reported relation then RScore = 0. At the marker level, the RScore is the sum of article level RScores, representing reported marker frequency.

Using the NScore, newly reported markers can be identified. Publication age is defined as the current year - publication year +1. According to different publication age thresholds n, NScores are differentiated into NScore_n, where n (years) = 1, 2... ; at article level, NScore_n = 0 when the article publication age is older than n; otherwise NScore_n > 0. At the marker level, NScore_n = 0 means the marker-disease relation has been reported more than n years prior to the date of this article.

Using the CScore, marker-disease relations that are highly cited can be identified. An The article CScore is its number of citations. The marker level CScore of a relation is the sum of the total citations of all the articles supporting the relation.

The QScore is a composite index considering three factors of an article-reported relation: 1) the number of citations; 2) the publication age, and, 3) the RScore. The range of the QScore of an article is (0, 1), and is inversely related to publication age and positively related to its citation number. If an article is recently published with a high citation number, its QScore will be close to 1, and if the article is older with a low number of citations, its QScore approaches 0. The marker level QScore is the sum of the QScores of all the articles supporting the marker.

Both article, and marker, level scores are designed at the relation level to evaluate article/s significance to the relation. If multiple marker-disease relations have been reported by one article, the article will have scores for each of the relations.

2.3 Gene Set Enrichment Analysis

A gene set/pathway enrichment analysis (GSEA) and a sub-network enrichment analysis (SNEA) on five groups to better understand the underlying functional profile and validate the pathogenic significance of the reported genes. It included: 1) entire gene list (1,925 genes); and 2) 4-subgroups selected according to highest quality matrix scores (150 genes per group). Pathway Studio (www.pathwaystudio.com) performed a network connectivity analysis on subsets.

GSEA (also functional enrichment analysis) is a method for analyzing biological high-throughput experiments, which identifies classes of genes, or proteins, that are over-represented in a large set of genes, or proteins. These gene sets could be known-biochemical pathways, or otherwise functionally-related genes. The method uses statistical approaches to identify significantly enriched, or depleted, gene groups to retrieve a functional profile of the input gene set which should provide a better understanding of the underlying biological processes. With this method, one does not consider the perturbation of single genes but instead, entire (functionally-related) gene sets. This approach is more robust. Single gene falsely perturb more readily than entire pathways.

A sub-network enrichment analysis (SNEA) was performed which was implemented by Pathway Studio using master casual networks (database) containing more than 6.5 million relationships derived from more than 4 million full-text articles and 25 million PubMed abstracts. These networks are generated by a Natural Language Processing (NLP) text-mining system to extract relationship data from scientific literature. It is an alternative to the manual curation processes used by IPA (http://www.ingenuity.com/products/ipa). The ability to quickly update the terminologies and linguistics rules used by NLP systems purports to ensure that new terms

are captured soon after entering regular use in the literature.

This extensive database of interaction data provides high levels of confidence when interpreting experimentally-derived genetic data against the background of previously published results. (http://help.pathwaystudio.com/fileadmin/standalone/pathway_studio/help_ps_10.0/index.html?analyze_experiment.htm)

3 RESULTS

3.1 Summary of LDM Results

In this study, an LDM was conducted on 17,884 articles that reported 1,925 genes associated with lung cancer. Using the reported category of gene-lung cancer relations, these 17,884 articles can generally be clustered into 6 different groups: 1) biomarker (0.62 %); 2) clinical trial (0.16 %); 3) genetic change (53.91 %); 4) quantitative change (22.51 %); 5) regulation (21.75 %); and, 6) state change (1.05 %).

The publication date distribution of these 17,884 articles appears in in Fig. 1, where it is shown that this study covers literature data from 1976 to 2016. The articles have an average publication age of only 6.2 years, indicating that most of the articles were published relatively recently. Publication age as defined as the current year - publication date + 1. Recently, especially after 2010, the number of publications has rapidly grown. The analysis shows that article publication date distributions underlying each of the 1,925 genes are similar to those in Fig. 1.

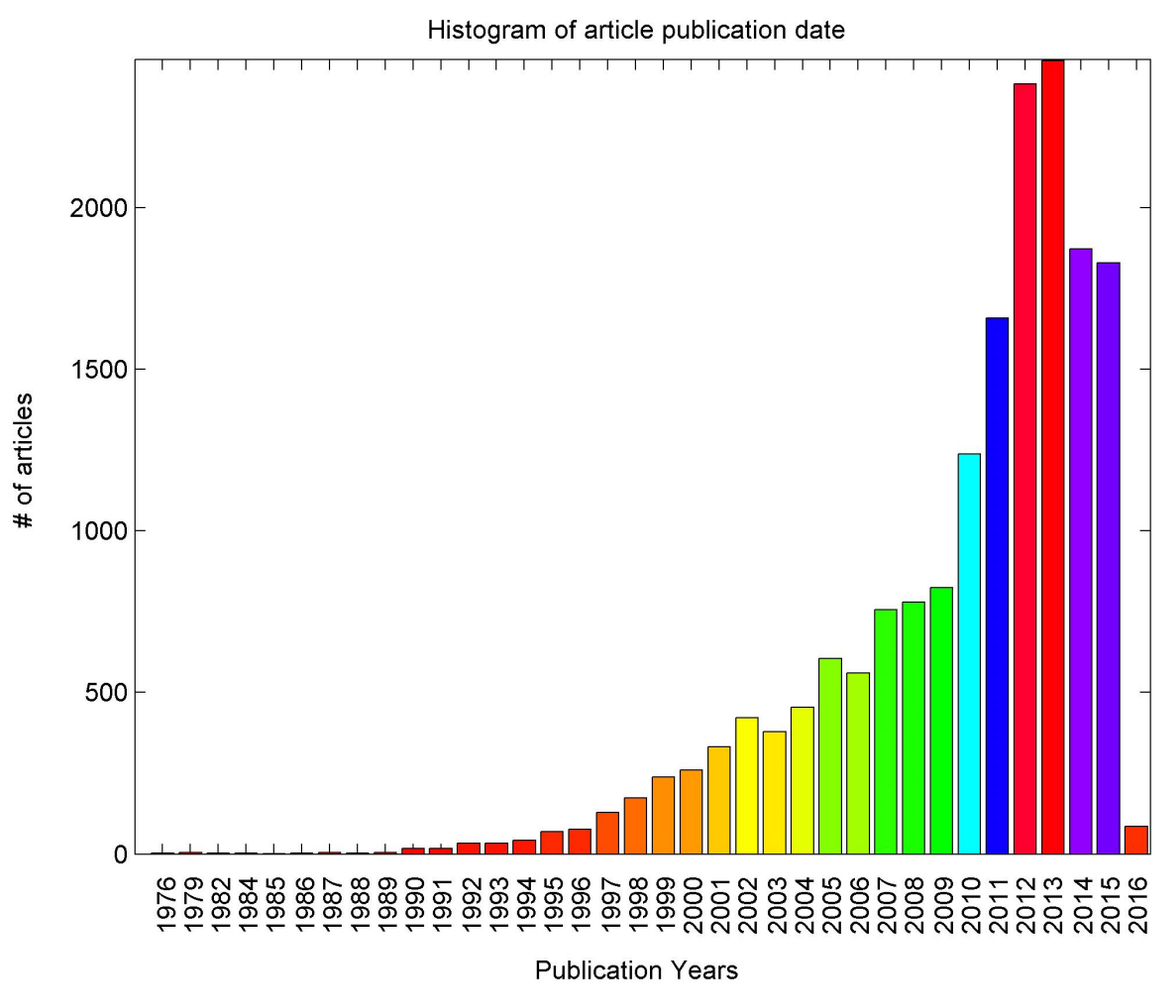


Fig. 1 Histogram of publications reporting marker-disease relationships between lung cancer and 1,925 genes

3.2 Marker Ranking

Fig. 2 shows the marker-wise score values for the 1,925 genes. The x-axis represents the index of markers ranked by QScore and the y-axis contains the CScore, QScores, NScore, and RScore normalized by their maximum values, respectively.

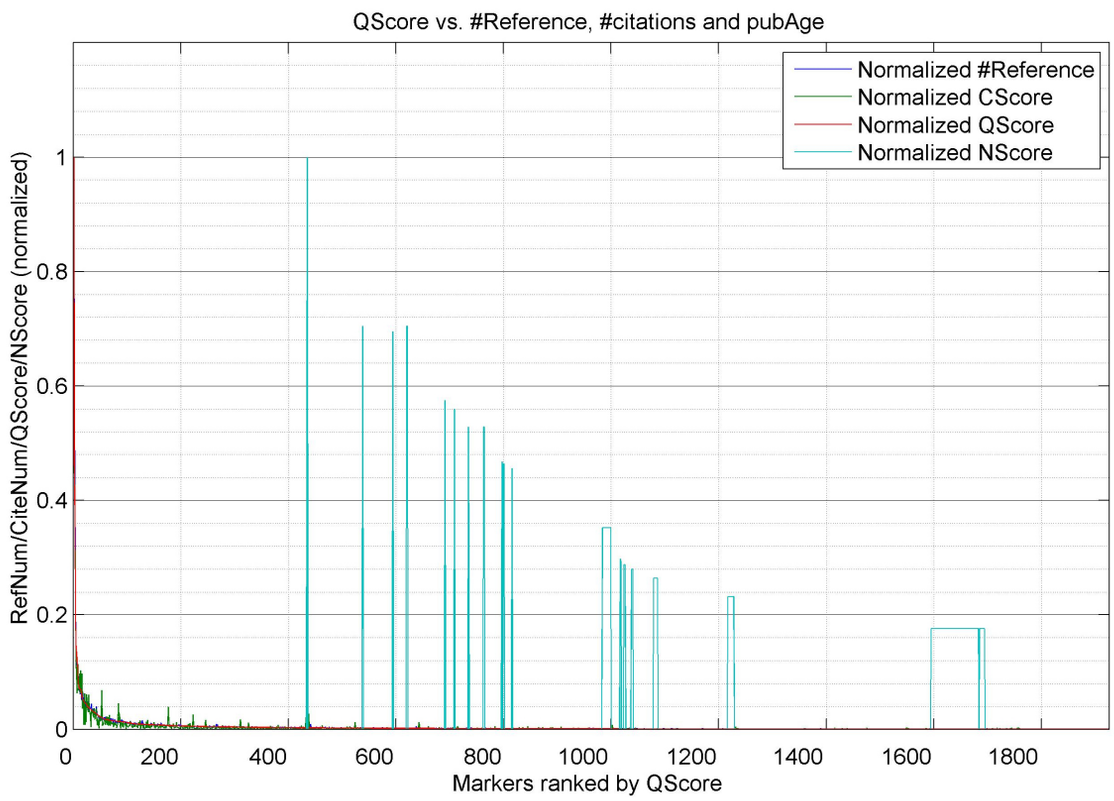


Fig. 2 Plot of CScore, QScore, NScore, and RScore. Each measure was normalized using the corresponding maximum value. The NScore presented in this figure is NScore_2. NScore will be zero if the corresponding gene gets supports from published prior to 2014.

Using the 4 scores, some genes were identified as frequently reported with large numbers of articles to support them, such as EGFR (2,141 articles), KRAS (1,085 articles), and TP53 (1,003 articles). These genes have the highest RScore. Some genes recently reported since 2014 have a high NScore, such as MAPK8 (NScore_2: 5.7), MIR423 (NScore_2: 4.0), and SIRT2 (NScore_2: 4.0). These genes usually have fewer supporting articles which yields a low RScore (Supplementary Material 1). Genes with high report frequencies (RScore) do not necessarily have a higher number of citations (CScore). This may be caused by many factors such as the total number, and publication age, of underlying articles. The QScore was used to balance these factors.

Of these 1,925 genes, 150 were reported during 2015 and 2016, with an NScore_2 > 0 (Fig. 1). These 150 genes appear in Table 1. The full results are in Supplementary Material 1. For comparison purposes, in Table 1 the highest 150 RScore genes (most frequently reported) are shown. There are large overlap percentages among the top genes selected with the RScore, CSore, and QScore, (e.g., overlap > 75 % for the top 150 genes). Redundancy is reduced by presenting only the 150 highest CScore and QScore genes (Supplementary Material 1).

Table 1. Top 150 genes with reported associations to lung cancer ranked by different scores

Genes By RScore	EGFR; KRAS; TP53; CYP1A1; XRCC1; BRAF; ALK; GSTT1; GSTM1; GSTP1; ERBB2; ERCC2; MET; PTGS2; PIK3CA; CDKN2A; VEGFA; ROS1; CYP2E1; KEAP1; ERCC1; EPHX1; MPO; FHIT; MTHFR; OGG1; TERT; CYP1A2; CYP1B1; CYP2D6; RASSF1; STK11; ACE; CYP2A6; TGFB1; MIR21; PTEN; XPC; AKT1; NFE2L2; XRCC3; APEX1; CHRNA5; CHRNA3; FGFR1; APC; IL6; RARB; MMP2; MLH1; CCND1; BCL2; ABCB1; FAS; BIRC5; MIR17HG; MGMT; NRAS; HNRNPA2B1; VDR; CDH1; CRP; DNMT3B; HRAS; MYC; SOD2; CTNNB1; EGF; NQO1; TNFSF10; AHR; IL10; IL17A; TNF; MDM2; MYCL; PXN; ERBB3; IGF1; MIR31; SMAD2; CLPTM1L; ERBB4; IL1B; MMP1; BRCA1; CCNE1; IL8; NKX2-1; SMARCA4; SPP1; AGER; FOXM1; HGF; IGFBP3; POU5F1; SULT1A1; BAG6; ELANE; PBK; RET; MIR155; MMP9; MSH2; SERPINA1; TUSC2; BAX; EPHA2; NOTCH1; PPARG; TP73; WFDC2; ATM; RAF1; SMAD4; STAT3; CDKN1A; ENO2; IGFBP2; MIR34A; NBN; NCOA6; RB1; SHOX2; EFEMP1; GPX1; IL2; MTOR; TGFBR2; WISP1; XPA; AXIN2; BMI1; BRCA2; CASP7; CYP2A13; CYP2C9; DDR2; IGF1R; MIR196A2; NME1; RBM5; CHEK2; FASLG; LEP; RBL2; SIRT1; CADM1; EPHA3; KRT19
Genes By NScore	MAPK8; MIR423; SIRT2; VTI1A; NPTX1; CCL20; PTPN1; IL9; PAK1; UCA1; MIR30A; MIR5100; STK33; MIR224; AFAP1-AS1; CANX; CXCL16; ELF3; LRG1; MEG3; MIR137; MYO6; PTPRF; RAP2B; RICTOR; RIPK3; TJP1; TNKS2; TPST1; TPTE; FBXL5; ASH1L; KCNN4; CA2; MIR512-1; ATG2B; KCNH5; KCNH8; MIR29C; MKL1; PRDM14; SERPINB10; ADH1B; AKT1S1; EPS15; GIMAP6; GRK6; LCN2; MIR19A; MIR3662; MIR944; PLA2G6; PLOD2; ABCE1; AGO1; ATG4A; ATP2A3; ATP2C1; BIRC3; BTG2; C4A; C9; CCR5; CEBPB; CIZ1; CLASP1; CRISPLD2; CRMP1; CXCR6; DCLK1; E2F8; EI24; EIF2S1; ELN; ENTPD5; EPC1; EPHA4; EPHB3; FAM46A; FANCD2; FANCF; FAT4; FBLN5; FLNB; FOXO6; G6PD; GLO1; GNRH1; HERC4; HOTTIP; IL33; IL5; ITIH5; IWS1; KDM6B; KLK6; LMO4; LYNX1; MALT1; MAP3K3; MAP3K5; MAPK9; MIR103A1; MIR106A; MIR10A; MIR124-1; MIR1244-1; MIR125B1; MIR129-2; MIR135A1; MIR1469; MIR153-1; MIR191; MIR3152; MIR4293; MIR448; MIR4513; MIR4520A; MIR520H; MIR5579; MIR5689; MIR608; MIRLET7BHG; NAMPT; NANOS3; NDUFA13; NIPBL; NR1I3; PARVA; PLD2; PRSS3; PTPRH; PTPRN2; RBBP7; RPS15A; SESN2; SF3B1; SMPD2; SMYD3; SNPH; SRSF7; TCF7; TGIF1; TNC; TNFAIP8L2; TRAF2; TRPM2; USP36; VIP; XCR1

The NScore here is NScore_2; Any marker that has been reported before 2015 will have an NScore of 0. This study found 150 genes to be newly reported in 2015 and 2016.

3.3 Enrichment Analysis

GSEA and SNEA results for 3 different groups are presented: all 1,925 genes, and the 2 gene groups listed in Table 1. The results for the top 150 genes with the highest CScores, and QScores, appear in Supplementary Material 2 and 3.

3.3.1 Enrichment Analysis on All 1,925 Genes

The entire list of 198 pathways/gene sets that were enriched with a *p*-value<1.4E-015 appears in Supplementary Material 2. There were 114 pathways/gene sets enriched with *p*-values < 1E-20; 32 are enriched with *p*-values < 1E-40; and, 7 are enriched with *p*-values < 1E-70. In Table 2, the top 20 pathways/groups enriched by all the 1,925 genes, with *p*-values < 1e-047 appear.

Table 2. Molecular function pathways/groups enriched by 1,925 genes reported

	Hit type	GO ID	# of Entities	Overlap	p-value	Jaccard similarity
Cytoplasm	Cellular component	0005737	6831	900	9.87E-88	0.12
Cytosol	Cellular component	0005829	3173	539	4.68E-82	0.12
Response to drug	Biological process	0042493	509	192	3.65E-81	0.09
Negative regulation of apoptotic process	Biological process	0006916	650	210	1.32E-74	0.09
Positive regulation of transcription from RNA polymerase II promoter	Biological process	0010552	1041	271	1.4E-73	0.1
Extracellular space	Cellular component	0005615	1557	335	2.23E-73	0.11
Positive regulation of cell proliferation	Biological process	0008284	568	192	8.82E-72	0.09
Response to hypoxia	Biological process	0001666	259	128	2.21E-70	0.06
Apoptotic process	Biological process	0006917	790	224	9.89E-68	0.09
Negative regulation of cell proliferation	Biological process	0008285	471	168	1.26E-66	0.08
Nucleoplasm	Cellular component	0005654	2669	452	3.34E-66	0.11
Response to lipopolysaccharide	Biological process	0032496	252	120	3.78E-64	0.06
Response to organic cyclic compound	Biological process	0014070	253	119	7.36E-63	0.06
Positive regulation of transcription, DNA-templated	Biological process	0045941	623	186	8.14E-60	0.08
Nucleus	Cellular component	0005634	6877	832	1.23E-58	0.11
Positive regulation of apoptotic process	Biological process	0043065	393	140	2.19E-55	0.07
Aging	Biological process	0016280	254	106	9.39E-50	0.05
Response to estradiol	Biological process	0032355	175	88	9.93E-50	0.05
Negative regulation of transcription from RNA polymerase II promoter	Biological process	0000122	799	198	1.65E-49	0.08
Signal transduction	Biological process	0007165	1843	329	1.01E-47	0.1

In these significantly enriched pathways, there were: 11 pathways/gene sets identified as relating to cell apoptosis; 15 identified as relating to cell growth and proliferation; 7 identified as relating to protein phosphorylation; 9 identified as relating to protein kinase; and one identified as relating to the immune system. All these pathways/gene sets are related to the cancer development. In addition: one pathway/gene set was identified as relating to aging; 5 pathways/gene sets were identified as relating to the neural system; and 5 pathways/gene sets were identified as relating to drug effects.

Cancer is a disease of cell/tissue growth regulation failure. A normal cell transforming into a cancer cell indicates that the genes regulating cell growth and differentiation have been altered [54]. The GSEA performed for this study showed that 26 pathways/gene sets related to cell apoptosis, cell growth, and cell proliferation, and were significantly enriched with by the 1,925 genes reported.

Specifically, there were 11 pathways/gene sets related to cell apoptosis (*p*-value: [1.3e-074, 1e-016]): negative regulation of apoptotic process (GO: 0006916; *p*-value= 1.3e-074, overlap: 210); apoptotic process (GO: 0006917; *p*-value = 9.9e-068, overlap: 224); positive regulation of apoptotic process (GO: 0043065; *p*-value = 2.2e-055, overlap: 140); negative regulation of neuron apoptotic process (GO: 0043524; *p*-value = 8.4e-025, overlap: 59); intrinsic apoptotic signaling pathway (GO: 0008629; *p*-value = 3.9e-022, overlap: 35); positive regulation of neuron apoptotic process (GO: 0043525; *p*-value = 2.7e-021, overlap: 34); negative regulation of cysteine-type endopeptidase activity involved in apoptotic process (GO: 0001719; *p*-value=4.9e-020, overlap: 37); regulation of apoptotic process (GO: 0042981; *p*-value = 2.4e-019, overlap: 72); activation of cysteine-type endopeptidase activity involved in apoptotic process (GO: 0006919; *p*-value = 1.4e-018, overlap: 39); intrinsic apoptotic signaling pathway in response to DNA damage (GO: 0008630; *p*-value = 5.4e-018, overlap: 31); apoptotic signaling pathway (GO: 0097190; *p*-value = 1e-016, overlap: 43).

There were 11 pathways/gene sets related to cell apoptosis (*p*-value: [1.3e-074,1e-016]): (1) negative regulation of apoptotic process (GO: 0006916; *p*-value = 1.3e-074, overlap: 210); (2) poptotic process (GO: 0006917; *p*-value = 9.9e-068, overlap: 224); (3) positive regulation of apoptotic process (GO: 0043065; *p*-value = 2.2e-055, overlap: 140); (4) negative regulation of neuron apoptotic process (GO: 0043524; *p*-value = 8.4e-025, overlap: 59); (5) intrinsic apoptotic signaling pathway (GO: 0008629; *p*-value = 3.9e-022, overlap: 35); (6) positive regulation of neuron apoptotic process (GO: 0043525; *p*-value = 2.7e-021, overlap: 34); (7) negative regulation of cysteine-type endopeptidase activity involved in apoptotic process (GO: 0001719; *p*-value = 4.9e-020, overlap: 37); (8) regulation of apoptotic process (GO: 0042981; *p*-value=2.4e-019, overlap: 72); (9) activation of cysteine-type endopeptidase activity involved in apoptotic process (GO: 0006919; *p*-value = 1.4e-018, overlap: 39); (10) intrinsic apoptotic signaling pathway in response to DNA damage (GO: 0008630; *p*-value=5.4e-018, overlap: 31); and, and (11) apoptotic signaling pathway (GO: 0097190; *p*-value = 1e-016, overlap: 43).

In addition, there were 15 pathways/gene sets related to cell growth and proliferation (*p*-value: [8.8e-072,1e-015]): positive regulation of cell proliferation (GO: 0008284; *p*-value = 8.8e-072, overlap: 192); negative regulation of cell proliferation (GO: 0008285; *p*-value = 1.3e-066, overlap: 168); cell proliferation (GO: 0008283; *p*-value = 3.8e-044, overlap: 131); regulation of cell proliferation (GO: 0042127; *p*-value = 2.1e-041, overlap: 94); epidermal growth factor receptor signaling pathway (GO: 0007173; *p*-value = 6.8e-030, overlap: 73); positive regulation of smooth muscle cell proliferation (GO: 0048661; *p*-value = 4.2e-026, overlap: 40); fibroblast growth factor receptor signaling pathway (GO: 0008543; *p*-value = 1.7e-025, overlap: 61); vascular endothelial growth factor receptor signaling pathway (GO: 0048010; *p*-value = 3.6e-025, overlap: 49); negative regulation of cell growth (GO: 0030308; *p*-value = 1.5e-023, overlap: 54); transforming growth factor beta receptor signaling pathway (GO: 0007179; *p*-value = 4.1e-019, overlap: 49); positive regulation of fibroblast proliferation (GO: 0048146; *p*-value = 5.6e-018, overlap: 30); positive regulation of epithelial cell proliferation (GO: 0050679; *p*-value = 3.3e-017, overlap: 33); growth factor activity (GO: 0008083; *p*-value = 2.3e-016, overlap: 52); negative regulation of epithelial cell proliferation (GO: 0050680; *p*-value = 6.4e-016, overlap: 31); positive regulation of endothelial cell proliferation (GO: 0001938; *p*-value = 1e-015, overlap: 31). The immune system is another cancer-related factor [55]. This study identified one related gene set: the innate immune response (GO: 0045087; *p*-value = 1.9e-046, overlap: 192).

At the protein level, we identified 7 pathways/gene sets that were related to protein phosphorylation and 9 pathways/gene sets related to protein kinase: protein phosphorylation (GO: 0006468; *p*-value = 3.9e-045, overlap: 173); positive regulation of protein phosphorylation (GO: 0001934; *p*-value = 3.3e-032, overlap: 69); phosphorylation (GO: 0016310; *p*-value = 3.3e-030, overlap: 147); peptidyl-tyrosine phosphorylation

(GO: 0018108; *p*-value = 3.4e-028, overlap: 60); protein autophosphorylation (GO: 0046777; *p*-value = 3.5e-027, overlap: 68); positive regulation of peptidyl-serine phosphorylation (GO: 0033138; *p*-value=8.8e-017, overlap: 33); positive regulation of peptidyl-tyrosine phosphorylation (GO: 0050731; *p*-value = 1.2e-016, overlap: 37); protein kinase binding (GO: 0019901; *p*-value=2.9e-037, overlap: 125); protein kinase activity (GO: 0050222; *p*-value = 9e-037, overlap: 142); positive regulation of protein kinase B signaling (GO: 0 051897; *p*-value = 2.4e-032, overlap: 51); kinase activity (GO: 0016301; *p*-value = 1.5e-030, overlap: 150); protein tyrosine kinase activity (GO: 0004718; *p*-value = 8.7e-030, overlap: 59); transmembrane receptor protein tyrosine kinase activity (GO: 0004714; *p*-value = 4.1e-025, overlap: 36); transmembrane receptor protein tyrosine kinase signaling pathway (GO: 0007169; *p*-value = 8.6e-021, overlap: 48); positive regulation of I-kappaB kinase-NF-kappaB signaling (GO: 0043123; *p*-value = 2.9e-018, overlap: 52); positive regulation of MAP kinase activity (GO: 0043406; *p*-value = 6.2e-016, overlap: 29).

A protein kinase is a kinase enzyme that modifies other proteins by chemically adding phosphate groups to them (phosphorylation). Phosphorylation usually results in a functional change of the target protein (substrate) by changing enzyme activity, cellular location, or the association with other proteins.

Deregulated kinase activity is a frequent cause of cancer, and medications to inhibit specific kinases are being developed for cancer treatment [56].

Five enriched pathways/gene sets are related to neural system (*p*-value: [1e-047, 5.1e-016]) and another 5 to drug response (*p*-value: [3.7e-081,2e-031]). A single gene set related to aging (GO: 0016280), which was also significantly enriched (*p*-value = 9.4e-050, overlap: 106). Although these pathways/gene sets may directly relate to lung cancer, enrichment helps to understand any underlying biological processes of the disease to the benefit of treatement and medication development. More significantly enriched pathways appear in Supplementary Material 2.

Pathway Studio was used to perform a SNEA to identify pathogenic significances of the reported genes to other disorders may relate to lung cancer. The complete list of results appear in Supplementary Material 3. Table 3 is the disease-related, sub-networks enriched with a *p*-value < 1E-323.

Table 3. Sub-networks enriched by the 1,925 genes

Gene Set Seed	Total # of Neighbors	Overlap	*p*-value	Jaccard similarity
Infection	2507	857	<1E-323	0.24
Neoplasms	5096	1559	<1E-323	0.29
Melanoma	1235	641	<1E-323	0.26
Neoplasm Metastasis	1554	807	<1E-323	0.31
Cancer	4222	1446	<1E-323	0.31
Breast Neoplasms	1067	572	<1E-323	0.24
Breast Cancer	2756	1172	<1E-323	0.34
Glioma	1059	556	<1E-323	0.24
Prostate Cancer	1760	826	<1E-323	0.3
Colon Cancer	1189	628	<1E-323	0.26

Table 3 shows that many of these reported lung cancer-related genes were also identified in other cancers and have a large overlap (Jaccard similarity > 0.24).

3.3.2 Enrichment Analysis of Top 150 Genes With Highest Scores

QScore, CScore, and RScore are strongly related, while the NScore is not so strongly related. Here their differences, in terms of GSEA and SNEA results, are compared. Considering the similarity of the groups selected by QScore, CScore, RScore, only the results for the NScore group and the RScore group are provided (Table 4 and Table 5), and only the full results for QScore and CScore groups appear in Supplementary Material 2 and 3.

Table 4. Pathways/groups enriched by 150 genes with the highest NScore and RScore

	Pathway/gene set Name	GO ID	*p*-value
The first 10 pathways/ gene sets enriched by top 150 genes with highest NScores	Innate immune response	0002226;	5.09E-08
	Cytosol	0005829;	1.05E-07
	miRNAs	Pathway Studio Ontology	1.24E-07
	Inflammatory response	0006954;	2.07E-06
	Programmed necrotic cell death	0097300;	3.08E-06
	INSR phosphatase	Pathway Studio Ontology	6.53E-06
	Actin cytoskeleton reorganization	0031532;	9.18E-06
	JNK/MAPK Signaling	Pathway Studio Ontology	1.31E-05
	JNK	Pathway Studio Ontology	1.95E-05
	Chemotaxis	0006935;	3.73E-05
The first 10 pathways/ gene sets enriched by top 150 genes with highest RScore	Response to drug	0042493;	2.13E-48
	Response to organic cyclic compound	0014070;	1.76E-31
	Response to estradiol	0032355;	5.22E-29
	Aging	0016280;	3.16E-27
	Negative regulation of cell proliferation	0008285;	3.25E-25
	Response to organic substance	0010033;	4.75E-23
	Positive regulation of cell proliferation	0008284;	1.41E-22
	Tumor Suppressors	Pathway Studio Ontology	7.26E-22
	Negative regulation of apoptotic process	0006916;	9.75E-22
	Response to hypoxia	0001666;	1.26E-21

NScore is used as NScore_2, a non-zero-value of which represents that the gene is reported in 2015 or 2016. One hundred and fifty genes are reported to have non-zero NScore_2 values. The genes with the top NScores and those with the top RScores enrich different groups of pathways, with different *p*-values (NScore group: 5.09E-08~3.73E-05; RScore group: 2.13E-48 ~ 1.26E-21), indicating that the recently-reported genes are functionally different from the frequently-reported ones (Table 4).

Eight out of the 10 pathways/gene sets enriched by the RScore group (Table 4) were also enriched by the overall 1,925 genes that rank in the top 20 (Table 2). Similarly, the cytosol group (GO: 0005829) was enriched by both overall genes and the NScore group, although with much weaker significance (4.68E-82 vs. 1.05E-07), indicating that many more genes with similar functions have already been discovered.

The SNEA analysis tested disease sub-networks that had been enriched by the two groups of genes. Complete results appear in Supplementary Material 3. Table 5 shows the top 10 disease related sub-networks enriched by the two groups of genes.

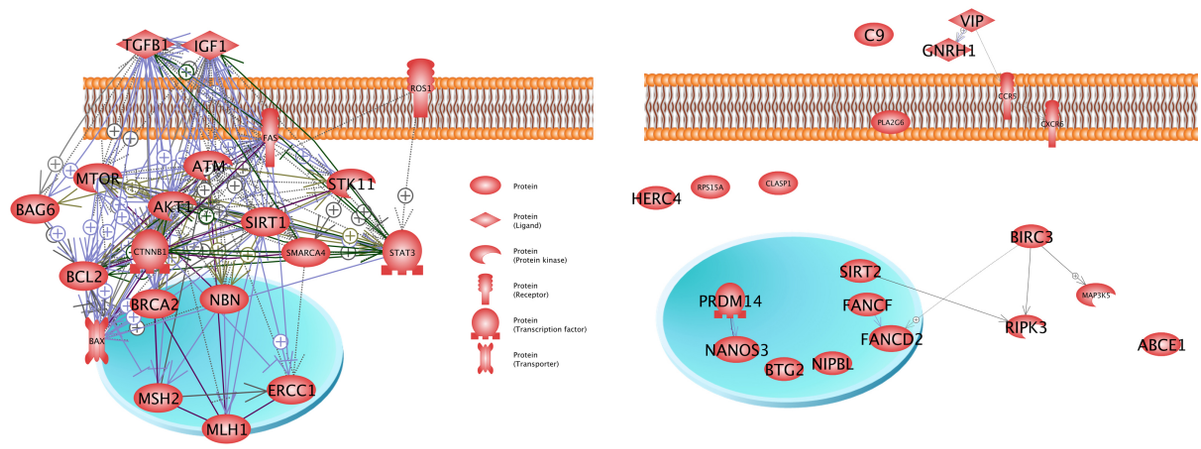
Table 5. SNEA results by 150 genes with the highest NScore and RScore

	Gene Set Seed	Overlap	*p*-value	Jaccard similarity
The first 10 pathways/gene sets enriched by top 150 genes with highest NScores	Neoplasms	98	1.18E-32	0.02
	Breast Cancer	74	1.36E-29	0.02
	Cancer	87	5.55E-29	0.02
	Carcinoma, Non-Small-Cell Lung	47	2.55E-22	0.03
	Colorectal Cancer	54	1.14E-20	0.02
	Carcinoma, Hepatocellular	55	2.38E-20	0.02
	Neoplasm Metastasis	48	4.78E-20	0.03
	Stomach Cancer	45	8.08E-18	0.02
	Infection	55	1.43E-17	0.02
The first 10 pathways/gene sets enriched by top 150 genes with highest RScore	Carcinoma, Non-Small-Cell Lung	142	1.3E-162	0.09
	Adenocarcinoma	130	6.1E-159	0.12
	Carcinoma, Squamous Cell	128	3.1E-148	0.11
	Urinary Bladder Cancer	120	4.5E-147	0.13
	Head and Neck Cancer	96	4E-146	0.25
	lung adenocarcinoma	111	6.1E-141	0.15
	Esophageal Cancer	102	7.3E-141	0.2
	Carcinogenesis	135	4.3E-140	0.08
	Smoking	119	7.6E-140	0.12

Both groups enriched some cancer/neoplasms related sub-networks. *p*-values enrichment by the RScore group is much more significant than those by the NScore group.

3.4 Connectivity Analysis

A network connectivity analysis (NCA) on the top 150 genes with the highest RScores and NScores (from Table 1) was performed to generate functional networks. Results show that for the RScore group, there are more than 5,000 relationships among those 150 genes, all having many literature supports. Fig. 3 (a) presents a network built using 20 genes, randomly selected from these 150 genes, where it can be seen that the genes are functionally connected to each other and form a complex network. In contrast, the 20 genes randomly selected from the 150-NScore group demonstrate only a few connections (Fig. 3 (b)). An NCA analysis shows only 290 relationships among the 98 genes for the whole 150-NScore group. This observation is consistent with GSEA and SNEA, suggesting that genes with a high NScore are not as functionally close to each other as those within the RScore group.



(a) By RScore group (b) By NScore group

Fig. 3 Connectivity networks built using 20 genes from the different groups. Pathway Studio generates the networks. The 20 genes were randomly selected from the 150 genes with highest RScores for (a) and NScores for (b), respectively.

4 DISCUSSION AND CONCLUSION

In this study, an LDM was performed on 17,884 articles (1976 through February 2016) which reported 1,925 genes associated with lung cancer. Supplementary Materials 1 provides the entire gene list and related parameters. A GSEA and an SNEA was performed to study the functional profile and pathogenic significance of the genes to lung cancer. An NCA was done to study functional associations between the top genes ranked by different scores. This approach is different from genetic studies that use raw data to report novel discoveries. It is a literature-based summarization and validation of already reported marker-diseases relations.

This study has several limitations. The literature data from the 17,884 articles studied were extracted from the Pathway Studio database. The Pathway Studio database is composed of over 40 million articles, it is still possible that some articles studying gene-lung cancer associations were not included.

Additionally, the 4 quality scores, RScore, NScore, CScore and QScore were proposed as quality measures of LDR identified marker-disease relations, feasible to rank the markers/relations according to different needs/significance.

However, although related to, they are not biological significance measures of the markers to the disease. Therefore, these measures cannot replace genetic statistical studies like GWAS, meta-analysis, and enrichment analysis.

As an automatic data mining approach, the Natural Language Processing (NLP) technique used for LDM is necessary and effective, when dealing with millions of articles. Any automatic LDM methods may produce false positives. This study is intended to lay groundwork for further studies. Towards this purpose, we provide, in Supplementary material 1, the detailed information for all 17,884 articles studied. This includes the sentences where a specific relation has been located.

The results from this up-to-date LDM reveal that these 1,925 genes have multiple types of associations with lung cancer. Enrichment analysis suggests that they play significant roles in lung cancer pathogenesis, as well as many other lung cancer-related disorder pathogeneses. NCA results demonstrate that these genes, especially high RScore genes, present strong functional associations with each other. The results suggest that these genes may operate as a functional biomarker network influencing lung cancer development.

Lung cancers are a complex diseases whose genetic causes are linked to a network composed of a large number of genes. LDM together with GSEA, SNEA, and NCA could serve as an effective approach for identifying these potential target genes.

CONFLICT OF INTERESTS

The authors declare no conflict of interests.

REFERENCES

1. Alavanja MC. Biologic damage resulting from exposure to tobacco smoke and from radon: implication for preventive interventions. Oncogene. 2002; 21: 7365-7375.
2. Alemany R, Ruan S, Kataoka M, Koch PE, Mukhopadhyay T, Cristiano RJ, Roth JA, Zhang WW. Growth inhibitory effect of anti-K-ras adenovirus on lung cancer cells. Cancer Gene Ther. 1996; 3(5): 296-301.
3. Antonia SJ, Mirza N, Fricke I., Chiappori A, Thompson P, Williams N, Bepler G, Simon G, Janssen W, Lee J H, Menander K, Chada S, Gabrilovic DI. Combi- nation of p53 cancer vaccine with chemotherapy in patients with extensive stage small cell lung cancer. Clin Cancer Res. 2006; 12: 878-887.
4. Aras G, Kanmaz D, Urer N, Purisa S, Kadakal F, Yentürk E, Tuncay E. Immunohistochemical expression of telomerase in patients with non-small cell lung cancer: prediction of metastasis and prognostic significance. Anticancer Res. 2013; 33(6): 2643-2650.
5. Cathcart MC, Gately K, Cummins R, Drakeford C, W Kay E, O'Byrne KJ, Pidgeon GP. Thromboxane synthase expression and correlation with VEGF and angiogenesis in non-small cell lung cancer. Biochim Biophys Acta. 2014; 1842(5): 747-755.
6. Cheng CY, Lee MC, Cheng CY, Hu CC, Yang HJ, Lee MC, Kao ES. Inhibitory effects of scutellarein on proliferation of human lung cancer A549 cells through ERK and NFkB mediated by the EGFR pathway. Chinese J Physiol. 2014; 57(4): 182-187.
7. Costa DB, Shaw AT, Ou S-HI, Martin Shreeve S, Selaru P, Wilner KD, Solomon BJ, Riely GJ, Schnell P, Ahn MJ, Zhou C, Polli A, Crinò L, Wiltshire R, Ross Camidge D. Clinical experience with crizotinib in patients with advanced ALK-rearranged non-small-cell lung cancer and brain metastases. J Clin Oncol. 2015; 33(17): 1881-1888.
8. Croce CM. Oncogenes and cancer. N Engl J Med. 2008; 358(5): 502-511.
9. Do K, Zlott J, Collins J, Chen AP, Doroshow JH, Kummar S, Wilsker D, Ji J, Kinders RJ, Freshwater T. Phase I Study of single-agent AZD1775 (MK-1775), a wee1 kinase inhibitor, in patients with refractory solid tumors. J Clin Oncol. 2015; 33(30): 3409-3415.
10. Finn OJ. Immuno-oncology: Understanding the function and dysfunction of the immune system in cancer. Ann Oncol. 2012; 23 Suppl 8: viii 6-9.

11. Groen HJ, Sietsma H, Vincent A, Hochstenbag MM, van Putten JW, van den Berg A, Dalesio O, Biesma B, Smit HJ, Termeer A, Hiltermann TJ, van den Borne BE, Schramel FM. Randomized, placebo-controlled phase III study of docetaxel plus carboplatin with celecoxib and cyclooxygenase-2 expression as a biomarker for patients with advanced non-small-cell lung cancer: The NVALT-4 study. J Clin Oncol. 2011; 29(32): 4320-4326.
12. Gruber K, Kohlhäufl M, Friedel G, Ott G, Kalla C. A novel, highly sensitive ALK antibody 1A4 facilitates effective screening for ALK rearrangements in lung adenocarcinomas by standard immunohistochemistry. J Thoracic Oncol. 2015; 10(4): 713-716.
13. Guo S, Li X, Gao M, Kong H, Li Y, Gu M, Dong X, Niu W. Synergistic association of PTGS2 and CYP2E1 genetic polymorphisms with lung cancer risk in northeastern Chinese. PLoS One. 2012; 7(6): e39814.
14. Hanada N, Takahata T, Zhou Q, Ye X, Sun R, Itoh J, Ishiguro A, Saijo Y, Hanada N, Fukuda S, Kijima H, Mimura J, Itoh K, Sun R. Methylation of the KEAP1 gene promoter region in human colorectal cancer. BMC Cancer. 2012; 12: 66.
15. Hara H, Yamashita K, Shinada J, Yoshimura H, Kameya T. Clinicopathologic significance of telomerase activity and hTERT mRNA expression in non-small cell lung cancer. Lung cancer. 2001; 34(2): 219-226.
16. He M, Xu S, Wang X. Telomerase activity in lung cancer and adjacent peritumoral tissues determined by TRAP-SYBR green assay. Chinese J Prev Med. 2001; 35(5): 301-304.
17. Jiang XY, Chang FH, Bai TY, Lv XL, Wang MJ. Susceptibility of lung cancer with polymorphisms of CYP1A1, GSTM1, GSTM3, GSTT1 and GSTP1 genotypes in the population of inner Mongolia Region. Asian Pacific J Cancer Prev. 2014; 15(13): 5207-5214.
18. Larsen JE, Minna D. Molecular biology of lung cancer: clinical Implications. Clin Chest Med. 2011; 32 (4): 703-740.
19. Lee JY, Myung SK, Song YS. Prognostic role of cyclooxygenase-2 in epithelial ovarian cancer: a meta-analysis of observational studies. Gynecol Oncol. 2013; 129(3): 613-619.
20. Li W, Li K, Zhao L, Zou H. DNA repair pathway genes and lung cancer susceptibility: a meta-analysis. Gene. 2013; 538(2): 361-365.
21. Liang H, Ju Z, Verhaak RGW, Mills GB, Cheung LWT, Li J, Yu S, Stemke-Hale K, Lu Y, Gu C, Guo W, Dyer MD, Zhang F, Hennessy BT, Dogruluk T, Scott KL, Liu X, Liu C-G, Scherer SE, Carter H, Karchin R, Westin SN, Lu KH, Broaddus RR, Hennessy BT. Whole-exome sequencing combined with functional genomics reveals novel candidate driver cancer genes in endometrial cancer. Genome Res. 2012; 22(11): 2120-2129.
22. Liu L, Shao X, Gao W, Bai J, Wang R, Huang P, Yin Y, Liu P, Shu Y. The role of human epidermal growth factor receptor 2 as a prognostic factor in lung cancer: a meta-analysis of published data. J Thoracic Oncol. 2010; 5(12): 1922-1932.
23. Liu X, Li Z, Zhang Z, Zhang W, Li W, Xiao Z, Liu H, Jiao H, Wang Y, Li G. Meta-analysis of GSTM1 null genotype and lung cancer risk in Asians. Med Sci Monit. 2014; 20: 1239-1245.
24. Lopez-Chavez A, Thomas A, Rajan A, Raffeld M, Morrow B, Kelly R, Carter CA, Guha U, Killian K, Lau CC, Abdullaev Z, Xi L, Pack S, Meltzer PS, Corless CL, Sandler A, Beadling C, Warrick A, Liewehr DJ, Steinberg SM, Berman A, Doyle A, Szabo E, Wang Y, Giaccone G. Molecular profiling and targeted therapy for advanced thoracic malignancies a biomarker-derived, multiarm, multihistology phase ii basket trial. J Clin Oncol. 2015; 33(9): 1000-1007.
25. Malhotra A, Nair P, Dhawan DK. Study to evaluate molecular mechanics behind synergistic chemo-preventive effects of curcumin and resveratrol during lung carcinogenesis. PLoS One. 2014; 9(4): e93820.
26. Meric-Bernstam F, Brusco L, Shaw K, Horombe C, Kopetz S, Davies MA, Routbort M, Piha-Paul SA, Janku F, Ueno N, Hong D, De Groot J, Ravi V, Li Y, Luthra R, Patel K, Broaddus R, Mendelsohn J, Mills GB. Feasibility of large-scale genomic testing to facilitate enrollment onto genomically matched clinical trials. J Clin Oncol. 2015; 33(25): 2753-2762.
27. Narayanan R, Yepuru M, Coss CC, Wu Z, Bauler MN, Barrett CM, Mohler ML, Wang Y, Kim J, Snyder LM, He Y, Miller DD, Dalton JT, Levy N. Discovery and preclinical characterization of novel small molecule TRK and ROS1 tyrosine kinase inhibitors for the treatment of cancer and inflammation. PLoS One. 2013; 8(12): e83380.
28. Natukula K, Jamil K, Pingali UR, Attili VS, Madireddy UR. The codon 399 Arg/Gln XRCC1 polymorphism is associated with lung cancer

in Indians. Asian Pacific J Cancer Prev. 2013; 14(9): 5275-5279.

29. Ohno M, Darwish WS, Ikenaka Y, Miki W, Ishizuka M. Astaxanthin can alter CYP1A-dependent activities via two different mechanisms: Induction of protein expression and inhibition of NADPH P450 reductase dependent electron transfer. Food Chem Toxicol. 2011; 49(6): 1285-1291.
30. Ohta Y, Tomita Y, Oda M, Watanabe S, Murakami S, Watanabe Y. Tumor angiogenesis and recurrence in stage I non-small cell lung cancer. Ann Thoracic Surg. 1999; 68(3): 1034-1038.
31. Ozcan MF, Dizdar O, Dincer N, Balci S, Guler G, Gok B, Pektas G, Seker MM, Aksoy S, Arslan C, Yalcin S, Balbay MD. Low ERCC1 expression is associated with prolonged survival in patients with bladder cancer receiving platinum-based neoadjuvant chemotherapy. Urol Oncol. 2012; 31(8): 1709-1715.
32. Padda SK, Burt BM, Trakul N, Wakelee HA. Early-stage non-small cell lung cancer: surgery, stereotactic radiosurgery, and individualized adjuvant therapy. Sem Oncol. 2014; 41(1): 40-56.
33. Patek CE, Arends MJ, Wallace WAH, Luo F, Hagan S, Brownstein DG, Rose L, Devenney PS, Walker M, Plowman SJ, Berry RL, Kolch W, Sansom OJ, Harrison DJ, Hooper ML. Mutationally activated K- ras 4A and 4B both mediate lung carcinogenesis. Exp Cell Res. 2008; 314(5): 1105-1114.
34. Ramalingam SS, Owonikoko TK, Shtivelband M, Soo RA, Barrios CH, Segalla JGM, Pereira JR, Makhson A, Gorbunova VA, Pittman KB, Kolman P, Srkalovic G, Belani CP, Axelrod R, Qin Q, Qian J, McKeegan EM, Devanarayan V, McKee MD, Ricker JL, Carlso. Randomized phase II study of carboplatin and paclitaxel with either linifanib or placebo for advanced nonsquamous non-small-cell lung cancer. J Clin Oncol. 2015; 33(5): 433-442.
35. Remon J, Morán T, Majem M, Reguart N, Dalmau E, Márquez-Medina D, Lianes P. Acquired resistance to epidermal growth factor receptor tyrosine kinase inhibitors in EGFR-mutant non-small cell lung cancer: A new era begins. Cancer Treat Rev. 2014; 40(1): 93-101.
36. Roskoski R Jr. ErbB/HER protein-tyrosine kinases: structures and small molecule inhibitors. Pharmacol Res. 2014; 87: 42-59.
37. Sadiq AA, Salgia R. MET as a possible target for non-small-cell lung cancer. J Clin Oncol. 2013; 31(8): 1089-1096.
38. Sasaki H, Endo K, Takada M, Kawahara M, Kitahara N, Tanaka H, Okumura M, Matsumura A, Iuchi K, Kawaguchi T, Yukiue H, Kobayashi Y, Yano M, Fujii Y. L858R EGFR mutation status correlated with clinico-pathological features of Japanese lung cancer. Lung Cancer. 2006; 54(1): 103-108.
39. Sasaki H, Okuda K, Shimizu S, Takada M, Kawahara M, Kitahara N, Okumura M, Matsumura A, Iuchi K, Kawaguchi T, Kubo A, Kawano O, Yukiue H, Yano M, Fujii Y. EGFR R497K polymorphism is a favorable prognostic factor for advanced lung cancer. J Cancer Res Clin Oncol. 2009; 135(2): 313-318.
40. Schildhaus HU, Schultheis AM, Rüschoff J, Binot E, Merkelbach-Bruse S, Fassunke J, Schulte W, Ko YD, Schlesinger A, Bos M, Gardizi M, Engel-Riedel W, Brockmann M, Serke M, Gerigk U, Hekmat K, Frank KF, Reiser M, Schulz H, Krüger S, Stoelben E. MET amplification status in therapy-naïve adeno- and squamous cell carcinomas of the lung. Clin Cancer Res. 2015; 21(4): 907-915.
41. Schreiner CA. Review of mechanistic studies relevant to the potential carcinogenicity of asphalts. Regul Toxicol Pharmacol. 2011; 59(2): 270-284.
42. Schumacker PT. Reactive oxygen species in cancer: A dance with the devil. Cancer Cell. 2015; 27(2): 156-157.
43. Song N, Liu B, Wu J, Zhang R, Duan L, He W, Zhang C. Vascular endothelial growth factor (VEGF) -2578C/A and -460C/T gene polymorphisms and lung cancer risk: a meta-analysis involving 11 case-control studies. Tumor Biol. 2014; 35(1): 859-870.
44. Thun MJ, Hannan LM, Adams-Campbell LL, Boffetta P, Buring JE, Feskanich D, Flanders WD, Jee SH, Katanoda K, Kolonel LN, Lee I, Marugame T, Palmer JR, Riboli E, Sobue T, Tang EA, Wilkens LR, Samet JM. Lung cancer occurrence in never-smokers: An analysis of 13 cohorts and 22 cancer registry studies. PLoS Med. 2008; 5(9): e185.
45. Ting CY, Wang HE, Liu YC, Ting CY, Yu CC, Liu HC, Liu YC, Chiang IT, Chiang IT. Curcumin triggers DNA damage and inhibits expression of DNA repair proteins in human lung cancer cells. Anticancer Res. 2015; 35(7): 3867-3874.

46. Wang J-Y, Cai Y, X-ray repair cross-complementing group 1 codon 399 polymorphism and lung cancer risk: An updated meta-analysis. Tumor Biol. 2014; 35(1): 411-418.

47. Wang R, An J, Ji F, Jiao H, Sun H, Zhou D. Hypermethylation of the Keap1 gene in human lung cancer cell lines and lung cancer tissues. Biochem Biophys Res Commun. 2008; 373(1): 151-154.

48. Wang Z, Qiao Q, Chen M, Li X, Wang Z, Liu C, Xie Z. miR-625 down-regulation promotes proliferation and invasion in esophageal cancer by targeting Sox2. FEBS Lett. 2014; 588(6): 915-921.

49. Wislez M, Barlesi F, Besse B, Mazières J, Merle P, Cadranel J, Audigier-Valette C, Moro-Sibilot D, Gautier-Felizot L, Goupil F, Renault A, Quoix E, Souquet PJ, Madroszyck A, Corre R, Pérol D, Morin F, Zalcman G, Soria JC. Customized adjuvant phase II trial in patients with non-small-cell lung cancer: IFCT-0801 TASTE. J Clin Oncol. 2014; 32(12): 1256-1261.

50. Xiao P, Chen J, Zhou F, Lu C, Yang Q, Tao G, Tao Y, Chen J. Methylation of P16 in exhaled breath condensate for diagnosis of non-small cell lung cancer. Lung cancer. 2013; 83(1): 56-60.

51. Yamashita T, Kamada H, Kanasaki S, Maeda Y, Nagano K, Abe Y, Inoue M, Yoshioka Y, Tsutsumi Y, Katayama S, Tsunoda S-I, Yamashita T, Kanasaki S, Maeda Y, Yoshioka Y, Tsutsumi Y, Kamada H, Yoshioka Y, Tsutsumi Y, Tsunoda S-I, Inoue M, Katayama S. Epidermal growth factor receptor localized to exosome membranes as a possible biomarker for lung cancer diagnosis. Pharmazie. 2013; 68(12): 969-973.

52. Yang IA, Holloway JW, Fong KM. Genetic susceptibility to lung cancer and co-morbidities. J Thor Dis. 2013; 5: S454-S462.

53. Yin Z, Cui Z, Ren Y, Zhang H, Yan Y, Zhao Y, Ma R, Wang Q, He Q, Zhou B. Genetic polymorphisms of TERT and CLPTM1L, cooking oil fume exposure, and risk of lung cancer: a case-control study in a Chinese non-smoking female population. Med Oncol. 2014; 31(8): 114.

54. Yoo SS, Jin C, Jung DK, Choi YY, Choi JE, Lee WK, Lee SY, Lee J, Cha SI, Kim CH, Seok Y, Lee E, Park JY. Putative functional variants of XRCC1 identified by RegulomeDB were not associated with lung cancer risk in a Korean population. Cancer Genet. 2015; 208(1-2): 19-24.

55. Zhang J, Yang PL, Gray NS. Targeting cancer with small molecule kinase inhibitors. Nat Rev Cancer. 2009; 9(1): 28-39.

56. Zhou HF, Feng X, Zheng BS, Qian J, He W. A meta-Analysis of the relationship between glutathione S-transferase T1 null/presence gene polymorphism and the risk of lung cancer including 31802 subjects. Mol Biol Rep. 2013; 40(10): 5713-5721.