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

Xinming Dong¹, McKenzie Ritter², Hongbao Cao³*, Dexiang Yang⁴*

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.

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 of these studies separate genes or 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. concluded that PTGS2 expression in patients with advanced non-small-cell lung cancer was neither a prognostic, or diagnostic marker, or function, attempts to provide a fuller view of the genetic-map related to lung cancer.

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 [11, 12]. Many studies, including independent studies and meta-analyses, have reported a genetic change in these genes in the case of lung cancer [13-20]. 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.pathwaysstudio.com) until February 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 article's 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.pathwaysstudio.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 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.

![Histogram of article publication date](image-url)
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.

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, CScore, 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

<table>
<thead>
<tr>
<th>Genes By RScore</th>
<th>Genes By NScore</th>
</tr>
</thead>
<tbody>
<tr>
<td>EGFR; KRAS; TP53; CYP1A1; XRCC1; BRAF; ALK; GSTT1; GSTM1; GSTD1; ERBB2; ERCC2; MET; PTGS2; PIK3CA; CDK2A; VEGFA; ROS1; CYP2E1; KEAP1; ERCC1; EPHX1; MPO; FHIT; MTHFR; OGG1; TERT; CYP1A2; CYP1B1; CYP2D6; RASSF1; STK11; ACE; CYP2A6; TGFB1; MIR2; PTEN; XPC; AKT1; NFE2L2; XRCC3; APEX1; CHRNA5; CHRNA3; FGFR1; APC; IL6; RARB; MMP2; MLH1; CCND1; BCL2; ABCB1; FAS; BIRC5; MIR17HG; NTR; HRAS; MYC; SOD2; CTNNB1; EGFR; NQO1; TNFSF10; AHR; IL10; IL17A; TNF; MDM2; MYCL; PXN; ERBB3; IGF1; MIR31; SMAD2; CLPTM1L; ERBB4; IL1B; MMP1; BRCA1; CCNE1; IL8; NXX2-1; SMARCA4; SPP1; AGER; FOXM1; HGF; IGFBP3; POUSF1; SULT1A1; BAG6; ELANE; PBK; RET; MIR155; MMP9; MSH2; SERPINA1; TUSC2; BAX; EPHA2; NOTCH1; PPARG; TP73; WDFC2; ATM; RAF1; SMAD4; STAT3; CTDK1A; ENO2; IGFBP2; MIR34A; NBN; NCOA6; RB1; SHOX2; EFEMP1; GPX1; IL2; MTO; TGFB2; WISP1; XPA; AXIN2; BM1; BRCA2; CASP7; CYP2A13; CYP2C9; DRR; IGF1R; MIR196A2; NME1; RBM5; CHEK2; FASLG; LEPI; RBL2; SIRT1; CAND1; EPAG3; KRT19</td>
<td></td>
</tr>
<tr>
<td>MAPK8; MIR423; SIRT2; VTI1A; NPTX1; CCL20; PTPN1; IL9; PAK1; UCA1; MIR30A; MIR5010; STK33; MIR224; AFAP1-A51; CANX; CXX16; ELF3; LRG1; MEG3; MIR137; MOY6; PTPR; RAP2B; RICTOR; RIPK3; TP1; TNKS2; TPST1; TPTE; FBXL5; ASH1L; KCNN4; CA2; MIR512-1; ATG2B; KCHN5; KCHN8; MIR29C; MLK1; PRDM14; SERPINB10; ADH1B; AKT1S1; EPS15; GMAP6; GRK6; LCN2; MIR19A; MIR3662; MIR944; PLAG26; PPLOD2; ABC1; AGO1; ATG4A; ATP2A3; ATP2C1; BIRC3; BTG2; C4A; C9; CCR5; CEBPB; CIZ1; CLASP1; CRISPLD2; CRM1; CXC6; DCKL1; E2F8; E2J2; EIF2S1; ELN; ENPTD5; EPC1; EPAG4; ERP63; EPAG8; ERAG; FANC; FAT4; FBLN5; FLNB; FOXO6; G6PD; GLO1; GNRH1; HERC4; HOTTIP; IL33; IL5; ITIH5; IWS1; KDM6B; KLK6; LMO4; LYNX1; MALT1; MAP3K3; MAP3K5; MIR9; 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; NDN1; NF1; PAP; PDC2; PRSS3; PTPR; PTPRN2; RBPI; RPS15A; SESN2; SF3B1; SMAD2; SMYD3; SNPH; SRSF7; TCF7; T ffi1; TNC; TNFAIP6L2; TRA2; TRPM2; USP36; VIP; XCR1</td>
<td></td>
</tr>
</tbody>
</table>

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/group sets that were enriched with a $p$-value$<$1.4E-015 appears in Supplementary Material 2. There were 114 pathways/group 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

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

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 treatment 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

<table>
<thead>
<tr>
<th>Gene Set Seed</th>
<th>Total # of Neighbors</th>
<th>Overlap</th>
<th>( p )-value</th>
<th>Jaccard similarity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Infection</td>
<td>2507</td>
<td>857</td>
<td>&lt;1E-323</td>
<td>0.24</td>
</tr>
<tr>
<td>Neoplasms</td>
<td>5096</td>
<td>1559</td>
<td>&lt;1E-323</td>
<td>0.29</td>
</tr>
<tr>
<td>Melanoma</td>
<td>1235</td>
<td>641</td>
<td>&lt;1E-323</td>
<td>0.26</td>
</tr>
<tr>
<td>Neoplasm Metastasis</td>
<td>1554</td>
<td>807</td>
<td>&lt;1E-323</td>
<td>0.31</td>
</tr>
<tr>
<td>Cancer</td>
<td>4222</td>
<td>1446</td>
<td>&lt;1E-323</td>
<td>0.31</td>
</tr>
<tr>
<td>Breast Neoplasms</td>
<td>1067</td>
<td>572</td>
<td>&lt;1E-323</td>
<td>0.24</td>
</tr>
<tr>
<td>Breast Cancer</td>
<td>2756</td>
<td>1172</td>
<td>&lt;1E-323</td>
<td>0.34</td>
</tr>
<tr>
<td>Glioma</td>
<td>1059</td>
<td>556</td>
<td>&lt;1E-323</td>
<td>0.24</td>
</tr>
<tr>
<td>Prostate Cancer</td>
<td>1760</td>
<td>826</td>
<td>&lt;1E-323</td>
<td>0.3</td>
</tr>
<tr>
<td>Colon Cancer</td>
<td>1189</td>
<td>628</td>
<td>&lt;1E-323</td>
<td>0.26</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Pathway/gene set Name</th>
<th>GO ID</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>The first 10 pathways/gene sets enriched by top 150 genes with highest NScores</td>
<td>Innate immune response</td>
<td>0002226;</td>
</tr>
<tr>
<td></td>
<td>Cytosol</td>
<td>0005829;</td>
</tr>
<tr>
<td></td>
<td>miRNAs</td>
<td>Pathway Studio Ontology</td>
</tr>
<tr>
<td></td>
<td>Inflammatory response</td>
<td>0006954;</td>
</tr>
<tr>
<td></td>
<td>Programmed necrotic cell death</td>
<td>0097300;</td>
</tr>
<tr>
<td></td>
<td>INSR phosphatase</td>
<td>Pathway Studio Ontology</td>
</tr>
<tr>
<td></td>
<td>Actin cytoskeleton reorganization</td>
<td>0031532;</td>
</tr>
<tr>
<td></td>
<td>JNK/MAPK Signaling</td>
<td>Pathway Studio Ontology</td>
</tr>
<tr>
<td></td>
<td>JNK</td>
<td>Pathway Studio Ontology</td>
</tr>
<tr>
<td></td>
<td>Chemotaxis</td>
<td>0006935;</td>
</tr>
<tr>
<td>The first 10 pathways/gene sets enriched by top 150 genes with highest RScore</td>
<td>Response to drug</td>
<td>0042493;</td>
</tr>
<tr>
<td></td>
<td>Response to organic cyclic compound</td>
<td>0014070;</td>
</tr>
<tr>
<td></td>
<td>Response to estradiol</td>
<td>0032355;</td>
</tr>
<tr>
<td></td>
<td>Aging</td>
<td>0016280;</td>
</tr>
<tr>
<td></td>
<td>Negative regulation of cell proliferation</td>
<td>0008285;</td>
</tr>
<tr>
<td></td>
<td>Response to organic substance</td>
<td>0010033;</td>
</tr>
<tr>
<td></td>
<td>Positive regulation of cell proliferation</td>
<td>0008284;</td>
</tr>
<tr>
<td></td>
<td>Tumor Suppressors</td>
<td>Pathway Studio Ontology</td>
</tr>
<tr>
<td></td>
<td>Negative regulation of apoptotic process</td>
<td>0006916;</td>
</tr>
<tr>
<td></td>
<td>Response to hypoxia</td>
<td>0001666;</td>
</tr>
</tbody>
</table>
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>
<thead>
<tr>
<th>Gene Set Seed</th>
<th>Overlap</th>
<th>( p )-value</th>
<th>Jaccard similarity</th>
</tr>
</thead>
<tbody>
<tr>
<td>The first 10 pathways/gene sets</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>enriched by top 150 genes with highest NScores</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Neoplasms</td>
<td>98</td>
<td>1.18E-32</td>
<td>0.02</td>
</tr>
<tr>
<td>Breast Cancer</td>
<td>74</td>
<td>1.36E-29</td>
<td>0.02</td>
</tr>
<tr>
<td>Cancer</td>
<td>87</td>
<td>5.55E-29</td>
<td>0.02</td>
</tr>
<tr>
<td>Carcinoma, Non-Small-Cell Lung</td>
<td>47</td>
<td>2.55E-22</td>
<td>0.03</td>
</tr>
<tr>
<td>Colorectal Cancer</td>
<td>54</td>
<td>1.14E-20</td>
<td>0.02</td>
</tr>
<tr>
<td>Carcinoma, Hepatocellular</td>
<td>55</td>
<td>2.38E-20</td>
<td>0.02</td>
</tr>
<tr>
<td>Neoplasm Metastasis</td>
<td>48</td>
<td>4.78E-20</td>
<td>0.03</td>
</tr>
<tr>
<td>Stomach Cancer</td>
<td>45</td>
<td>8.08E-18</td>
<td>0.02</td>
</tr>
<tr>
<td>Infection</td>
<td>55</td>
<td>1.43E-17</td>
<td>0.02</td>
</tr>
<tr>
<td>The first 10 pathways/gene sets</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>enriched by top 150 genes with highest RScore</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Carcinoma, Non-Small-Cell Lung</td>
<td>142</td>
<td>1.3E-162</td>
<td>0.09</td>
</tr>
<tr>
<td>Adenocarcinoma</td>
<td>130</td>
<td>6.1E-159</td>
<td>0.12</td>
</tr>
<tr>
<td>Carcinoma, Squamous Cell</td>
<td>128</td>
<td>3.1E-148</td>
<td>0.11</td>
</tr>
<tr>
<td>Urinary Bladder Cancer</td>
<td>120</td>
<td>4.5E-147</td>
<td>0.13</td>
</tr>
<tr>
<td>Head and Neck Cancer</td>
<td>96</td>
<td>4E-146</td>
<td>0.25</td>
</tr>
<tr>
<td>Lung adenocarcinoma</td>
<td>111</td>
<td>6.1E-141</td>
<td>0.15</td>
</tr>
<tr>
<td>Esophageal Cancer</td>
<td>102</td>
<td>7.3E-141</td>
<td>0.2</td>
</tr>
<tr>
<td>Carcinogenesis</td>
<td>135</td>
<td>4.3E-140</td>
<td>0.08</td>
</tr>
<tr>
<td>Smoking</td>
<td>119</td>
<td>7.6E-140</td>
<td>0.12</td>
</tr>
</tbody>
</table>

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.

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-disease 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


28. Natukula K, Jamil K, Pingali UR, Attili VS, Madireddy UR. The codon 399 Arg/Gln XRCC1 polymorphism is associated with lung cancer...


