Identification of potential molecular mechanisms of radiation pneumonitis development in non-small-cell lung cancer treatment by data mining

L. Zhu1,2, J. Zhang2, B. Xia1,3, S. Chen1 and Y. Xu2,*

1 Department of Oncology, Key Laboratory of Clinical Cancer Pharmacology and Toxicology Research of Zhejiang Province, Hangzhou Cancer hospital, Hangzhou 310006, PR China.
2 Department of Oncology, Affiliated Hangzhou First People’s Hospital, Zhejiang University School of Medicine, No. 6 Xiaonv Road, Shangcheng District, Hangzhou 310006, PR China.
3 Department of Oncology, Jiande Second People’s Hospital, Zhejiang 311604, PR China.

Received: 8 December 2019 / Accepted: 21 April 2020

Abstract – Introduction: Radiation pneumonitis (RP) is the most significant dose-limiting toxicity in patients receiving thoracic radiotherapy. The underlying mechanisms of RP are still inconclusive. Our objective was to determine the genes and molecular pathways associated with RP using computational tools and publicly available data. Methods: RP-associated genes were determined by text mining, and the intersection of the two gene sets was selected for Gene Ontology analysis using the GeneCodis program. Protein-protein interaction network analysis was performed using STRINGdb to identify the final genes. Results: Our analysis identified 256 genes related to RP with text mining. The enriched biological process annotations resulted in 47 sets of annotations containing a total of 156 unique genes. KEGG analysis of the enriched pathways identified 24 pathways containing a total of 41 unique genes. The protein-protein interaction analysis yielded 23 genes (mostly the PI3K family). Conclusion: Gene discovery using in silico text mining and pathway analysis tools can facilitate the identification of the underlying mechanisms of RP.

Keywords: radiation pneumonitis / text mining / data mining

1 Introduction

Lung cancer remains the most common cancer and the leading cause of cancer-related deaths in China and worldwide (Ferlay et al., 2010; Chen et al., 2016). Currently, approximately 60–70% of patients receive radiation therapy during the course of disease. Radiotherapy has both radical and palliative effects and plays an important role in lung cancer (Cannon et al., 2013). Due to the inflammation of normal lung tissues, radiation pneumonitis (RP) can occur after radiotherapy and is the most significant dose-limiting toxicity, affecting 5–36% of patients (Medhora et al., 2012; Wang et al., 2012). Therefore, patients with thoracic cancer are unable to finish the entire treatment course because of RP. A previous study found that RP was an independent negative prognostic factor in non-small-cell lung cancer (NSCLC) patients (Farr et al., 2015). Currently, the underlying mechanisms of RP remain to be understood. Numerous factors, including smoking status, pulmonary function, radiation dose, and irradiated lung volume, might have an association with RP (Das et al., 2007; Tsoutsou et al., 2006; Schallenkamp et al., 2007). However, these results are not sufficiently informative about the underlying mechanisms of RP development. Increasing evidence indicates that genetic and molecular factors might play a critical role in the development of RP and clinical outcomes of NSCLC (Li et al., 2016; Du et al., 2018; Wen et al., 2018). Therefore, the discovery of genetic pathways is needed to understand the underlying mechanisms of RP.

Due to the popularity of global communication and convenience of the internet, the biomedical literature in public databases such as PubMed has grown exponentially (Hunter et al., 2006). The manual collection of relevant published information on RP is an extremely time-consuming task. Furthermore, selecting specific genetic pathways among mountains of publications requires much labor, and some publications will be missed due to carelessness. Text mining (TM) adopts many algorithms to analyze correlations and statistical patterns in unstructured text to obtain specific information. In recent years, the TM technique could simultaneously process large-scale information extraction and discover knowledge (Simpson et al., 2012) and has been
successfully applied in medical research. For instance, the TM technique was used in cancer research, toxicogenomics, and cancer drug effects/safety (Korhonen et al., 2012; Zhu et al., 2013; Harpaz et al., 2014; Lee et al., 2014; Baker et al., 2016).

In this study, we used TM, pathway analysis, and database analytical tools to identify genetic pathways potentially related to RP. We applied a proven natural language processing methodology and other computational resources (e.g., STRINGdb, GeneCodis) to perform in silico analysis (Shim et al., 2014). Biological knowledge can be categorized and classified using Gene Ontology (GO) hierarchical relationships (Andronis et al., 2011). Combined with TM, sets of genes can be further prioritized by assessing their impact on protein interaction networks. This technique is useful because it reveals novel relationships through the analysis of gene interactions that tend to cluster together in networks because of certain properties of the disease or pathological condition (Liu et al., 2014). This study aims to investigate the genetic pathways of RP by using computational methods to mine a list of high priority target genes with publicly available biological data.

2 Materials and methods

2.1 Text mining

We performed a search in PubMed (https://www.ncbi.nlm.nih.gov/) using the terms “radiation”, “pneumonitis”, and “lung cancer”. All abstracts were retrieved and input into R software (version 3.4.4). The jieba package (https://cran.r-project.org/web/packages/jiebaR/index.html) was used for word segmentation (the R script is provided in supplementary file S1). We then extracted all the unique gene hits from each result as the starting point for the next steps.

2.2 Biological process and pathway analysis

GeneCodis (http://genecodis.cnb.csic.es/) is a web-based tool for interpreting genomic data by integrating various sources of GO and functional information (Nogales-Cadenas et al., 2009; Tabas-Madrid et al., 2012). The gene list from TM was used for enrichment analysis in GeneCodis. We performed a query with genes regarding the GO biological processes (BPs) involved. We selected the most highly enriched terms closely related to RP. Then, the genes in the most highly enriched terms were analyzed in GeneCodis with annotations of the Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways.

2.3 Protein interaction network

Search Tool for the Retrieval of Interacting Genes/Proteins (STRING) (http://string-db.org/) is also a web-based tool for analyzing the protein-protein interaction network of the selected genes. The STRING database integrates TM in PubMed, experimental/biochemical evidence, co-expression networks, and database association to provide an interactive platform in which the connections, associations, and interactions between proteins can be assessed (Szklarczyk et al., 2015). All the genes selected from the pathway enrichment analysis were used as the input set. To further narrow the candidate gene field, a confidence level of $\geq 0.90$ was set within STRING to display gene interactions (Szklarczyk et al., 2015).

3 Results

Based on the data mining strategy described in Figure 1, 1099 articles were retrieved from PubMed. Then, the abstracts were input into R for TM, and a list of 256 genes were identified to be related to the search terms “lung”, “radiation” and “pneumonitis” (Supplementary Tab. S1).

The GO BP annotations revealed that the most highly enriched terms were particularly relevant to the search terms “lung”, “radiation” and “pneumonitis”, validating the utility of the TM search. The enriched BP annotations resulted in 47 sets of annotations containing a total of 156 unique genes. Table 1 shows the top 15 enriched BP annotations (Tab. 1). The three most enriched BP annotations were positive regulation of gene expression ($P = 6.80E-07$), blood coagulation ($P = 9.60E-07$), and signal transduction ($P = 1.22E-06$), containing 10, 18, and 29 genes from the query set, respectively (Tab. 1). Signal transduction is broadly defined as a BP. Hence, the higher gene representation and relatively low $P$ value make this process particularly relevant to RP. Other highly enriched BP annotations included negative regulation of apoptotic process, cell adhesion, and acute-phase response.

The analysis of enriched pathway annotations resulted in the selection of 24 pathways containing a total of 41 unique genes when combined (Tab. 2). The three most significantly enriched pathways were focal adhesion ($P = 5.22E-10$), Jak-STAT signaling pathway ($P = 8.00E-08$), and complement and coagulation cascades ($P = 2.98E-07$), containing 13, 10, and 7 genes from the query set, respectively. Additional highly enriched relevant pathways were pathways in cancer, apoptosis, non-small-cell lung cancer, and chemokine
signaling pathway (Tab. 2). STRING combined data from current database, gene fusion, gene co-expression, protein homology, and biological experiments to determine the protein-protein interaction. The analyses yielded 23 genes (Fig. 2). Of these 23 genes, PIK3CG, PIK3CB, PIK3CD, PIK3CA, and PIK3R5 were PI3K isoforms; SERPINC1 and PDGFRB were molecules upstream of PI3K pathways; CREBBP was downstream molecule of PI3K pathways.

4 Discussion

RP and subsequent radiation pulmonary fibrosis are the two main dose-limiting factors in thoracic radiation that can have severe implications for patients’ quality of life, especially when combined with tyrosine kinase inhibitors, immunotherapy, or chemotherapy. To explore the detailed mechanism of RP, the present study identified a list of 256 genes related to the search

<table>
<thead>
<tr>
<th>Process</th>
<th>Genes in query set</th>
<th>Total genes in genome</th>
<th>FDR corrected hypergeometric (P) value</th>
<th>Genes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Positive regulation of gene expression (BP)</td>
<td>10</td>
<td>104</td>
<td>6.80E-07</td>
<td>SLC6A4, CD44, TWIST1, PLAUR, AVPR2, NCOA3, MYLK2, CSF2, CALR, KIT</td>
</tr>
<tr>
<td>Blood coagulation (BP)</td>
<td>18</td>
<td>457</td>
<td>9.60E-07</td>
<td>THPO, PPBP, L1CAM, PROS1, IFN1A, PIK3CA, SERPINC1, SERPINB2, PLAUR, TP53, F11R, F8, PIK3CG, PIK3CB, TIMP1, BCAR1, SERPINE1, PIK3R5</td>
</tr>
<tr>
<td>Signal transduction (BP)</td>
<td>29</td>
<td>1176</td>
<td>1.22E-06</td>
<td>RASGRF1, NR2E3, KCNIP3, OPTN, NR1P1, APOL3, COL4A3BP, PIK3CA, NR1I2, ARTN, PLAUR, ADCY3, ARL2BP, ESR1, GJA1, ARFRP1, HINT1, PIK3CB, OPN1LW, KIT, LTBR, PIK3CD, CXXC5, BCAR1, ANGPTL4, PAG1, CREBBP, IMPA1, PDGFRB</td>
</tr>
<tr>
<td>Transport (BP)</td>
<td>19</td>
<td>604</td>
<td>5.10E-06</td>
<td>RBP5, NCBP1, DBI, CLCN2, SLC25A47, APOL3, COL4A3BP, SERPINA5, CACNA1B, SLC25A33, RBM8A, GJA1, SLC6A8, CNGB1, MTCH1, SLC25A32, SLC12A3, PARP4, MTCH2</td>
</tr>
<tr>
<td>Axon guidance (BP)</td>
<td>14</td>
<td>307</td>
<td>5.25E-06</td>
<td>L1CAM, COL9A1, NR1P1, FARP2, SCN8A, ARTN, COL5A2, PTPRC, PTTPRA, CACNA1F, SEMA3A, COL9A2, MYL6, COL5A1</td>
</tr>
<tr>
<td>Cell adhesion (BP)</td>
<td>18</td>
<td>556</td>
<td>6.55E-06</td>
<td>L1CAM, COL9A1, AMBP, CD44, DSC3, NR1P1, DSC2, RPSA, F11R, HABP2, F8, CTNNBD1, PIK3CB, COL11A2, LAMC2, BCAR1, COL5A1, COMP</td>
</tr>
<tr>
<td>Platelet activation (BP)</td>
<td>11</td>
<td>234</td>
<td>7.60E-05</td>
<td>THPO, PPBP, SAA1, PROS1, PIK3CA, F8, PIK3CG, PIK3CB, TIMP1, SERPINE1, PIK3R5</td>
</tr>
<tr>
<td>Cell junction assembly (BP)</td>
<td>7</td>
<td>90</td>
<td>0.000247</td>
<td>CD44, PLAUR, TP53, LEP, PIK3CG, CFLAR, CSF2, TIMP1, ANGPTL4, SERPINE1, JTB</td>
</tr>
<tr>
<td>Negative regulation of apoptotic process (BP)</td>
<td>11</td>
<td>272</td>
<td>0.000257</td>
<td>NR0B1, NR2E3, NR1I2, PPARA, PTGER3</td>
</tr>
<tr>
<td>Intracellular receptor mediated signaling pathway (BP)</td>
<td>5</td>
<td>32</td>
<td>0.000277</td>
<td>SERPINA5, SERPINC1, SERPINB2, PLAUR, SERPINB3</td>
</tr>
<tr>
<td>Negative regulation of transcription from RNA polymerase II promoter (BP)</td>
<td>13</td>
<td>416</td>
<td>0.000341</td>
<td>SAA1, F8, REG3G, REG3A, SAA4</td>
</tr>
<tr>
<td>Regulation of proteolysis (BP)</td>
<td>5</td>
<td>37</td>
<td>0.000359</td>
<td>PROS1, SERPINA5, SERPINC1, SERPINB2, SERPINE1, SERPINB3</td>
</tr>
<tr>
<td>Acute-phase response (BP)</td>
<td>5</td>
<td>37</td>
<td>0.000359</td>
<td>SERPINA5, SERPINC1, SERPINB2, PLAUR, SERPINB3</td>
</tr>
<tr>
<td>Negative regulation of endopeptidase activity (BP)</td>
<td>6</td>
<td>67</td>
<td>0.00039</td>
<td>SAA1, F8, REG3G, REG3A, SAA4</td>
</tr>
<tr>
<td>Cell migration (BP)</td>
<td>7</td>
<td>105</td>
<td>0.000432</td>
<td>NR1P1, JUP, ATN1, BCAR1, COL5A1, PARP9, PDGFRB</td>
</tr>
</tbody>
</table>

* \(P\) value less than 0.05.
terms “lung”, “radiation” and “pneumonitis” by TM. Further analyses identified 23 genes representing a network following gene set enrichment analysis. Of these 23 genes, PI3Ks constituted the most important part. Five genes (PIK3CG, PIK3CB, PIK3CD, PIK3CA, and PIK3R5) were PI3K isoforms; three genes (SERPINC1, PDGFRB, and CREBBP) were molecules upstream or downstream of PI3K pathways.

PI3Ks participate in various BPs by phosphorylating the 3-hydroxyl group on cellular membrane phosphoinositides and were demonstrated to participate in the pathogenesis of the inflammatory response to injury by regulating several determinant events in inflammation (Hawkins et al., 2015). Zhang et al. showed apoptosis in irradiated normal lung tissue correlated with inhibition of downstream PI3K
signaling levels (Zhang et al., 2012). Radiation can induce a robust inflammatory response that contributes to the development of lung fibrosis and the resultant late morbidity from lung radiation (Rubin et al., 1995). Several studies showed PI3K pathways play important roles in pulmonary fibrosis (Miyoshi et al., 2013; Yan et al., 2014). Tsoyi et al. found syndecan-2 attenuates radiation-induced pulmonary fibrosis in mice by down-regulating PI3K signaling (Tsoyi et al., 2018). In addition, in the pathogenesis of RP, lung epithelial cells transdifferentiate into fibroblast-like cells by epithelial-to-mesenchymal transition (EMT) (Yarnold et al., 2010). The PI3K/AKT signaling pathway is a critical mediator of EMT (Barber et al., 2015). Tang et al. found that genetic variations in PIK3CA were significantly associated with the occurrence of severe RP (HR = 0.132, 95% CI: 0.042–0.416, P = 0.001) (Tang et al., 2016). Several studies have found the correlation between radiation-induced injury and PI3K pathway, while these correlations were commonly weak and indirect. Further well-designed study was still needed to explore the value of PI3Ks in the development of RP.

This study combined TM and bioinformatic approaches to explore the molecular mechanisms of RP. The lack of validation by further experiments is a major drawback of this study. It also has certain limitations that need to be addressed. Some key genes or pathways were commonly described in the abstracts, even though they were negative, and these genes obtained from PubMed might be irrelevant to RP. Some potential genes only in the full text but not in the abstract might be missed. In addition, bioinformatics stems from the current databases, which may have limited information on genes regarding their function or role in a pathway. Additionally, not all gene interactions have been fully elucidated.

5 Conclusion

We presented a method to discover the genes/pathways relevant to RP. This method may be used routinely at intervals as databases and analytical tools evolve and improve. In the present analysis, we identified 23 genes, of which the PI3K family might have a correlation with RP. In fact, our group has established a RP model in mice. We proposed some of PI3K gene family would change along with development of RP and lung tissue at different time point after ionizing radiation would be obtained to determine expression of PI3K signaling proteins. Of course, further experimental results are needed to validate the hypothesis.

Acknowledgments. This work was supported by the Science and Technology Project of Hangzhou Bureau (2018A20 and 2018A33), Scientific research funds of Zhejiang province health department (2017KY532), Social development project of Hangzhou Municipal Science and Technology Commission (20170533B95), National Nature and Science Foundation of
China (81803042) and Natural Science Foundation of Zhejiang Province (LQ17H160003 and LQ20H160020). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Conflicts of interest. The authors declare that they have no conflicts of interest in relation to this article.

Supplementary Material

Supplementary File S1.
Supplementary Tab. S1.

The Supplementary Material is available at https://www.radioprotection.org/10.1051/radiopro/2020048/olm.

References
