Review

Analysis of IncRNA expression profiles in non-small cell lung cancers (NSCLC) and their clinical subtypes

Jingcheng Yang\textsuperscript{a,1}, Jinyuan Lin\textsuperscript{a,1}, Tianxiao Liu\textsuperscript{a,1}, Ting Chen\textsuperscript{b}, Shangling Pan\textsuperscript{c}, Weiqiang Huang\textsuperscript{a,1,\ast}, Shikang Li\textsuperscript{a,1,\ast}

\textsuperscript{a} Department of Thoracic and Cardiovascular Surgery, The First Affiliated Hospital of Guangxi Medical University, Nanning, China
\textsuperscript{b} Department of Management Information System, College of Computer and Information Engineering, Guangxi Teachers Education University, Nanning, China
\textsuperscript{c} Department of Pathophysiology, Guangxi Medical University, Nanning, China

\textbf{Article history:}
Received 30 January 2014
Received in revised form 4 April 2014
Accepted 11 May 2014

\textbf{Keywords:}
Non-small cell lung cancer
Biomarker
IncRNA
Expression profiling
Adenocarcinoma
Squamous cell carcinoma

\textbf{ABSTRACT}

Lung cancer is one of the most common human cancers worldwide. Among all lung cancer cases, non-small cell lung cancer (NSCLC) accounts for approximately 85%. Long non-coding RNAs (lncRNAs) are non-protein-coding transcripts that have been shown to play important roles in tumourigenesis and tumor progression. To reveal novel tumor-related lncRNAs in NSCLC and their associations with clinical subtypes, we herein identified 2935 probe sets mapped to lncRNAs on Affymetrix HG-U133 Plus 2.0 array with an lncRNA classification pipeline. We found 47 lncRNAs differentially expressed between normal lung tissues and tumor samples and 19 lncRNAs differed in expression between SCC and AC, two subtypes of NSCLC, after analyses of the gene expression profiles of five datasets downloaded from the gene expression omnibus (GEO) with a leave one dataset out validation process. The different lncRNA expression profiles between NSCLC and normal tissue and between the subtypes of NSCLC may have potential implications in the pathogenesis of this cancer. lncRNAs screening may be beneficial in the diagnosis, subclassification, and the personalized treatment of NSCLC.

© 2014 Elsevier Ireland Ltd. All rights reserved.

\section{Introduction}

lncRNA (long non-coding RNA) is a RNA molecule that is longer than 200 nucleotides which is not translated into a protein [1]. Many identified lncRNAs are transcribed by RNA polymerase II (RNA pol II). They mainly locate within nucleus or cytosolic compartment [2]. lncRNAs can be further subcategorized into the following locus biotypes based on their location with respect to protein-coding genes: (1) sense, (2) antisense, (3) bidirectional, (4) intronic, and (5) intergenic [3]. They regulate gene expression through epigenetic regulation, splicing, imprinting, transcriptional regulation and subcellular transport [4]. Over the last few decades, researches have been focusing on the role of protein-coding genes in the pathogenesis of cancer [5], paying less attention to the possible effects of lncRNAs. However, in recent years emerging evidence indicate that lncRNAs are dysregulated and play important roles in tumourigenesis and tumor progression [6]. Some of the regulatory mechanisms of lncRNAs have been elucidated. For example, HOXAIR, an lncRNA locates in the HOX locus on 12q13.13, was firstly described as having a fundamental role in human breast cancer [7]. It binds to PRC2, silencing a portion of the HOXD locus, inducing H3 lysine 27 trimethylation, then remodeling the gene expression pattern of breast epithelial cells as does in embryonic fibroblasts [4,8]. lncRNA expression profiles are also altered in other types of cancers, including human prostate cancer, renal carcinoma, ovarian cancer, and human lung adenocarcinoma, raising the possibility that lncRNAs may become a promising diagnostic biomarker [9].

Lung cancer is one of the most common human cancers worldwide with considerable high morbidity and mortality [10,11]. Among all lung cancer cases, non-small-cell lung cancer (NSCLC) accounts for approximately 85% [12] whose most effective therapy is complete lung resection plus appropriate chemotherapeutic strategy [13]. Nevertheless, several studies have observed different therapeutic responses due to specific cell types of NSCLC [14].

\begin{itemize}
\item \textsuperscript{1} These authors contributed equally.
\item \textsuperscript{\ast} Corresponding authors at: Department of Thoracic and Cardiovascular Surgery, The First Affiliated Hospital of Guangxi Medical University, 6 Shuangyong Road, Nanning 530021, Guangxi Zhuang Autonomous Region, China. Tel.: +86 13397705186; fax: +86 771 5352775.
\item \textsuperscript{\ast} E-mail address: shikangli@hotmail.com (S. Li).
\end{itemize}

\url{http://dx.doi.org/10.1016/j.lungcan.2014.05.011}

0169-5002/© 2014 Elsevier Ireland Ltd. All rights reserved.
2.1. GEO lung cancer gene expression data

To identify all relevant datasets, we searched GEO for NSCLC expression profiling studies. Studies were included in the systematic review if (i) they were gene profiling studies in patients with NSCLC; (ii) they used NSCLC tissue and normal lung tissue for comparison; (iii) they used the same platform; (iv) they contained more than three samples meeting the quality control standard in experimental and controlled group. As a result, 5 panels of NSCLC gene expression datasets were included: GSE27262, GSE19804, GSE19188, GSE30219 and GSE18842 [19–23]. While GSE27262, GSE19804, GSE19188 and GSE30219 were used in the leave one dataset out validation process, and GSE18842 served as an independent dataset to validate the gene signature derived from the meta-analysis. Of these datasets, GSE18842, GSE19188 and GSE30219 included information of SCC and AC. These datasets were selected to compare the lncRNA expression signatures among NSCLC subtypes.

2.2. lncRNA classification pipeline

To evaluate the lncRNA expressions in NSCLC gene expression data, we applied a pipeline as described by Zhang et al. [18] to identify the probe sets uniquely mapped to lncRNAs from the Affymetrix array by using the following steps. Firstly, we mapped Affymetrix HG-U133 Plus 2.0 probe set ID to the NetAffx Annotation Files (HG-U133 Plus 2.0 Annotations, CSV format, release 33, 10/30/12) (Table S1). The annotations contained the probe set ID, Ensembl gene ID, Refseq transcript ID gene symbol, gene title, gene symbol and other informative items for the specific probe set. Then the probe sets were filtered by the

<table>
<thead>
<tr>
<th>Scenario I</th>
<th>Scenario II</th>
<th>Scenario III</th>
<th>Scenario IV</th>
</tr>
</thead>
<tbody>
<tr>
<td>GSE27262</td>
<td>GSE19804</td>
<td>GSE19188</td>
<td>GSE30219</td>
</tr>
<tr>
<td>GSE19804</td>
<td>GSE19188</td>
<td>GSE27262</td>
<td>GSE30219</td>
</tr>
<tr>
<td>GSE30219</td>
<td>GSE19804</td>
<td>GSE19188</td>
<td>GSE27262</td>
</tr>
<tr>
<td>GSE19188</td>
<td>GSE27262</td>
<td>GSE19804</td>
<td>GSE30219</td>
</tr>
</tbody>
</table>

Fig. 1. Consistency-based meta-analysis process. The four datasets (GSE27262, GSE19804, GSE19188 and GSE30219) were switched around to create four scenarios. After the individual analysis of each dataset, three of the datasets were used to perform the signature and the result was then validated in an independent dataset in each scenario. The signatures from each scenario were then combined to be the final gene signatures, which were confirmed by the fifth dataset.
following criteria: (1) for the probe sets only assigned with Refseq IDs, those labeled as “NR” (NR indicates non-coding RNA in the Refseq database) were retained; (2) for the probe sets only assigned with Ensembl gene IDs, those annotated with “antisense”, “transcribed”, “lincRNA” or “non_sense-mediated_decay”, “sense_intronic”, “sense_overlapping” in Ensembl annotations were retained; (3) for the probe sets with both Ensembl gene IDs and Refseq IDs, we retained those annotated with Ensembl gene titles above and labeled as “NR” in Refseq database at the same time. Secondly, we filtered out pseudogenes, rRNAs, microRNAs and other short RNAs including tRNAs, snRNAs and snoRNAs and obtained 2935 Affymetrix probe sets with corresponding annotated lncRNA transcripts.

Supplementary Table S1 related to this article can be found, in the online version, at [http://dx.doi.org/10.1016/j.lungcan.2014.05.011](http://dx.doi.org/10.1016/j.lungcan.2014.05.011).

### 2.3. Individual data processing and analysis

After quality control assessment was done, each dataset was normalized by applying a Bioconductor package Robust Multi-Array Average (RMA) [24] through R 3.0.1. The normalized data were then analyzed with linear models for microarray data (LIMMA), a modified t-test incorporating the Benjamini–Hochberg multiple hypotheses correction technique [25]. The probe sets of which the adjusted P-value was below 0.01 and the expression level differed by ≥2-fold between two groups were defined as significantly different lncRNAs, through which we obtained five lncRNA lists from the five datasets.

### 2.4. Identification of differentially expressed lncRNAs between normal lung tissue and tumor samples

To investigate the distinctive lncRNAs between normal lung tissue and tumor samples, a workflow (Fig. 1) of consistency-based meta-analysis approach was designed. This approach used a leave one dataset out validation process and the result was then validated in an independent dataset (GSE18842). Firstly, the meta-analysis was replicated four times, with each time leaving out one of the four datasets (GSE27262, GSE19804, GSE19188 and GSE30219) and performing the analysis using the remaining three datasets. In each scenario, when aggregating the distinctive lncRNAs in the three test datasets, the lncRNAs differentially expressed in at least two datasets were selected to be the gene signature of this scenario. They were then validated in the fourth dataset. Both principal component analysis (PCA) and hierarchical cluster analysis (HCA) done by the package pheatmap and pcaMethods were used to visually inspect the leave one dataset out cross-validation results [26,27]. Secondly, we picked out the lncRNAs differentially expressed in at least two scenarios to be the final gene signature. An independent dataset (GSE18842) was used for validation. Both PCA and HCA were used to visually inspect the result.

Moreover, we performed hierarchical cluster analyses to evaluate the degree of concordance between the results of the five individual datasets. We constructed overall rank matrix based on rank matrices obtained from these datasets for up-regulated and down-regulated lncRNA lists, such as different regions of people, sample size and tumor histology. In this matrix, value of 0.5 was performed by blank, which means that this lncRNA was not in the distinctive lncRNA list of that dataset. The up-regulated lncRNAs were performed by value >0.5; while value <0.5 means that this lncRNA is down-regulated in that dataset. In cluster analysis, Euclidean with complete method was used.

### 2.5. Identification of differentially expressed lncRNAs in NSCLC subtypes

lncRNAs differentially expressed (with adjusted P-value ≤0.01 and fold-change ≥2) between AC and SCC in at least two of the three datasets (GSE18842, GSE19188, GSE30219) were considered as the gene signature of NSCLC subtypes. Both PCA and HCA were used to visually inspect the result.

### 3. Results

#### 3.1. Datasets characteristics

Our systematic review included five cohorts of NSCLC gene expression data from GEO: GSE27262, GSE19804, GSE19188, GSE30219 and GSE18842. All these datasets contained lung cancer tissue and normal lung tissue. GSE19188 contained 137 samples, including 27 SCC, 45 AC, and 65 non-tumoral lung tissue controls. GSE30219 contained 160 samples, including 61 SCC, 85 AC, and 14 non-tumoral lung tissue controls. GSE18842 contained 91 samples, including 32 SCC and 14 AC, as well as 45 non-tumoral lung tissue controls. Table 1 summarizes the details of these five datasets. All the datasets were used to identify a set of lncRNA expression signatures in lung cancer. GSE18842, GSE19188 and GSE30219 acted as the test dataset to select distinctive lncRNA expressions between AC and SCC.

#### 3.2. lncRNA expression profiles on Affymetrix HG-U133 Plus 2.0 arrays

According to the NetAffx annotation of probe sets and the Refseq and Ensembl annotations of lncRNAs, we identified 2935 probe sets corresponding to lncRNA genes on the Affymetrix HG-U133 Plus 2.0 arrays (Table S2). Among them, 1002 probe sets mapping to lncRNA genes were annotated by both the RefSeq and Ensembl database; 500 probe sets were annotated only by the RefSeq database; 1423 probe sets were annotated only by the Ensembl database. Those probe sets with controversial annotations in the two databases were culled.

Supplementary Table S2 related to this article can be found, in the online version, at [http://dx.doi.org/10.1016/j.lungcan.2014.05.011](http://dx.doi.org/10.1016/j.lungcan.2014.05.011).

#### 3.3. Distinctive lncRNA expressions between normal lung tissue and tumor samples

By using a leave one dataset out validation process, we identified four groups of lncRNAs from the four scenarios, the numbers of which are shown in Table 2. For each scenario, the gene signatures

### Table 1

<table>
<thead>
<tr>
<th>Reference</th>
<th>Year</th>
<th>No. of control</th>
<th>No. of tumor</th>
<th>Platform</th>
</tr>
</thead>
<tbody>
<tr>
<td>GSE27262</td>
<td>2012</td>
<td>25</td>
<td>25</td>
<td>Affymetrix</td>
</tr>
<tr>
<td>GSE19804</td>
<td>2010</td>
<td>60</td>
<td>60</td>
<td>HG-U133 Plus 2.0</td>
</tr>
<tr>
<td>GSE19188</td>
<td>2010</td>
<td>65</td>
<td>72</td>
<td></td>
</tr>
<tr>
<td>GSE30219</td>
<td>2013</td>
<td>14</td>
<td>146</td>
<td></td>
</tr>
<tr>
<td>GSE18842</td>
<td>2011</td>
<td>45</td>
<td>46</td>
<td></td>
</tr>
</tbody>
</table>
aggregated from any three datasets were validated in the remaining dataset. Hierarchical clustering of all samples in the remaining dataset using the gene signature obtained above revealed clear distinctions between normal lung and tumor samples (Figs. S1A, 2A, 3A, 4A). Further analysis using PCA could also distinguish the normal lung and tumor samples. In scenario I, all samples were perfectly distributed into lung cancer and normal tissues by PC1 = 6 (Fig. S1B); in scenario II, most samples were correctly distributed by PC1: PC1 < 0 in 55 out of 60 tumor samples, and PC1 > 0 in 57 out of 60 normal lung tissue samples (Fig. S2B); in scenario III, 68 out of 72 tumor samples and 64 out of 65 normal lung tissue samples were distributed by PC1 = 0 (Fig. S3B); in scenario IV, all samples were correctly distributed by PC1 = 0 (Fig. S4B). After aggregating the four scenarios of gene signatures, we identified 47 lncRNAs (Table S3). Of them, 14 lncRNAs were up-regulated and 33 were down-regulated. 25 lncRNAs with great consistency were summarized in Table S4. HCA and PCA of all samples in the independent dataset (GSE18842) using the final list of gene signature indicated that all samples could be perfectly distributed (Fig. 2).

Supplementary Figs. S1–S4 and Tables S3, S4 related to this article can be found, in the online version, at http://dx.doi.org/10.1016/j.lungcan.2014.05.011.

After assessing the degree of concordance between the results of these five datasets, we found datasets GSE18842, GSE19188 and GSE30219 shared high similarity, and GSE19804 and GSE27262 were in high concordance (Fig. 3).

### 3.4. Distinctive lncRNA expression patterns between SCC and AC

We also compared the lncRNA expression signatures between AC and SCC in our study by comparing the two subtypes of NSCLC in the three datasets, i.e., GSE18842 AC (n = 14) versus SCC (n = 32), GSE19188 AC (n = 45) versus SCC (n = 27), and GSE30219 AC (n = 85) versus SCC (n = 61). 19 lncRNAs with different expression were identified (Table S5). The results of HCA and PCA showed that these unique lncRNAs could discern the majority of AC and SCC (Figs. S5–S7).

Supplementary Figs. S5–S7 and Table S5 related to this article can be found, in the online version, at http://dx.doi.org/10.1016/j.lungcan.2014.05.011.

### 4. Discussion

In the past few decades, molecular signatures have been indicated to play important roles in the molecular mechanism, diagnosis and treatment of lung cancer [9,28,4,29,13]. Recently, accumulating evidence indicates that lncRNAs, as a kind of non-coding RNAs, are associated with a diverse range of functions in cell biology [30]. But the lncRNA expression profiles in lung cancer have not been systematically explored. To further clarify the expression patterns of lncRNA in lung cancer and its subtypes, we investigated the lncRNA expression signatures in lung cancer and its subtypes, hoping to shed more light on its molecular regulatory mechanisms which may be critical in its personalized treatment.

Affymetrix HG-U133 Plus 2.0 array series is one of the most commonly used commercial microarrays in human cancer profiling [31,32], from which abundant gene expression data of lung cancer are available. However, how to extract the expression data of lncRNAs becomes a new challenge. In the present study, we applied an lncRNA classification pipeline to achieve this goal. This method is of low cost and feasible and it has been applied by several researchers [18,30].

To the best of our knowledge, this is the first study that correlates lncRNA gene expression to histological differentiation of human lung cancer.

#### 4.1. Distinctive lncRNA expressions between normal lung tissues and tumor samples

In the current study, we analyzed 47 lncRNAs that differentiated (with at least 2-fold difference) between normal lung tissues
and NSCLC tumor samples in the five datasets. When we use PCA to visually inspect the result, we noted that although the normal lung tissue samples and tumor samples can be well distinguished, neither of them could be perfectly clustered together. The possible reason could be as follows: (1) the differences between samples were remarkable; (2) the lncRNA signature identified by us cannot stably distinguish between normal lung tissues and tumor samples. Among the 47 lncRNAs, some have been proved to be related to lung cancer. For example, IGKC, which was significantly up-regulated and in great consistency of the five datasets, has been reported to be expressed in tumor-infiltrating plasma cells [33] and associated with longer survival in NSCLC and have specific impact on the adenocarcinoma subgroup [34]. As known, immune response plays pivotal roles in all stages of carcinogenesis which may promote or inhibit tumor progression [35]. In light of our meta-results, IGKC could be a B-cell related gene signature, making it an appealing target for cancerous immunotherapy. Thus, the precise roles of IGKC signature warrant further study. However, the majority of the differentially expressed lncRNAs such as BRE-AS1, FOXF1-AS1 and KDM5B-AS1 identified in our study, so far, have not been implicated in lung cancer or other cancer. Whether these lncRNAs play important roles in the development and progression of lung cancer deserves further clarification. Interestingly, while searching for NSCLC-related lncRNAs, we found some intensely investigated lung cancer-related lncRNAs were not part of our lncRNA gene signature. For example, HOTAIR, which was reported to be highly expressed in NSCLC with advanced stage, lymph node metastasis and short disease free interval [36], did not reach statistical significance in our meta-analyses. Through checking the specific information of patients in the five datasets except GSE19188 and GSE30219 that did not provide related information, we found the majority of samples we adopted were of stage I or II without lymph node or distant metastasis. We thus speculate that HOTAIR may less highly correlate with the early stage of NSCLC, which also needs further investigations. In addition, H19, which showed modest expression alteration in our study, was also noted to serve as a tumor suppressor gene involved in the pathogenesis of lung cancer [37]. This could probably indicate that H19 is distinctively expressed in different types of lung cancer and there are distinct growth mechanisms between NSCLC and other subtypes, which provide a direction for our future study.

4.2. lncRNA expression difference between AC and SCC

We analyzed 19 lncRNAs that differed (with at least 2-fold difference) between squamous cell carcinoma and adenocarcinoma. This may indicate that the expression patterns of lncRNA in SCC and AC are different. Through searching for studies describing the differentially expressed lncRNAs, we found none of them have been reported to have positive relationships between SCC and AC. And while searching for related studies, we found several NSCLC subtype-related lncRNAs showed relatively modest expression change in our study. For example, IGF2, located on chromosome 11p15.5, and MEST on chromosome 7q32, were noted to be correlated with the histologic classification of adenocarcinomas [38]. Whether they can be the gene signature between NSCLC is yet to be determined.

4.3. Clinical implications

As known, identification of biomarkers associated with disease progression is extremely meaningful. Cancer is initiated by a series of cumulative genetic and epigenetic changes that affect normal cellular identity, growth and differentiation. Nowadays, more and more studies indicate that lncRNAs are promising markers of cancer and are useful in tumor diagnosis, classification, prognosis and therapeutic evaluation. In the current analysis, we identified 47 lncRNAs which expressed differentially between NSCLC and non-tumor samples, and 19 lncRNAs which differed between SCC and AC, two subtypes of NSCLC. Further investigations are warranted to clarify the mechanisms lie behind these differences. It is of great interest to screen for these lncRNAs, especially for which discern SCC and AC, in the personalized therapy of NSCLC.

4.4. Limitations

There were some limitations in our study. First of all, the probable mechanism of action of lncRNAs has not been delineated due to its heterogeneity [39]. Secondly, although Affymetrix HG-U133 Plus 2.0 array series is one of the most commonly used commercial microarrays in human cancer profiling, it does not represent all of the possible lncRNAs presented [18]. Thirdly, the sample size of each dataset is relatively small, which may reduce the accuracy of
the results. Furthermore, the genetic profiles in different subgroups of the lung cancer patients in terms of age, gender and stages of lung cancer were not taken into account because the supplementary files of some datasets we adopted did not provide related information. In addition, the datasets we used were class-imbalanced, e.g., in GSE30219, the number of normal tissues was significantly fewer than the lung cancer samples, which may decrease the accuracy on other minority class samples [40]. Thus, further large-scale studies are needed to confirm our findings.

Conflict of interest statement

The authors declare that there is no conflict of interest.

Funding

Research Project of Guangxi Colleges and Universities (201203YB035), Guangxi Key Project of Science and Technology (1355005-4-8).

Acknowledgements

We thank Hui Chen and Shanshan Zeng for technical assistance and constructive suggestion.

References

学霸图书馆

www.xuebalib.com

本文献由“学霸图书馆-文献云下载”收集自网络，仅供学习交流使用。

学霸图书馆（www.xuebalib.com）是一个“整合众多图书馆数据库资源，提供一站式文献检索和下载服务”的24小时在线不限IP图书馆。

图书馆致力于便利、促进学习与科研，提供最强文献下载服务。

图书馆导航:

图书馆首页  文献云下载  图书馆入口  外文数据库大全  疑难文献辅助工具