Cross-disease analysis identified novel common genes for both lung adenocarcinoma and lung squamous cell carcinoma

GUANGHUI ZHANG¹, WEIJIE WANG¹, WEIYANG HUANG¹, XIAOLI XIE¹, ZHIGANG LIANG² and HONGBAO CAO³

¹Department of Cardiothoracic Surgery, Ningbo Fourth Hospital, Ningbo, Zhejiang 315037; ²Department of Thoracic Surgery, Ningbo First Hospital, Ningbo, Zhejiang 315000, P.R. China; ³Statistical Genomics and Data Analysis Core, National Institutes of Health, Bethesda, MD 20852, USA

Received September 6, 2018; Accepted May 25, 2019

DOI: 10.3892/ol.2019.10678

Abstract. Lung squamous cell carcinoma (LSCC) exhibits a number of similarities with lung adenocarcinoma (LA) in patients with LSCC (P<4.0x10⁻⁴), which was a significant factor for the levels of expression of all three genes identified using statistical analysis. Subsequently, a functional pathway analysis was performed to identify a possible association between any of the significant genes identified from the mega-analysis and LSCC, followed by a co-expression analysis. A multiple linear regression (MLR) model was employed to investigate the possible influence of sample size, country of origin and study date on gene expression in patients with LSCC. Disease-gene association data analysis identified 1,178 genes involved in LA, 334 in LSCC, with a significant overlap of 187 genes (P<1.02x10⁻⁶). Mega-analysis revealed that three LA-associated genes, such as solute carrier family 2 member 1 (SLC2A1), endothelial PAS domain protein 1 (EPAS1) and cyclin-dependent kinase 4 (CDK4), were significantly associated with LSCC (P<1.60x10⁻⁶), which raises the possibility of the presence of common genetic features between these two diseases. In addition, SLC2A1, EPAS1 and CDK4 may be novel risk genes associated with LA and LSCC.

Introduction

Non-small-cell lung carcinoma (NSCLC) accounts for ~85% of all cases of lung cancer worldwide, and the most common histological subtypes of NSCLC are lung adenocarcinoma (LA) and lung squamous cell carcinoma (LSCC) (1). LA and LSCC cells originate from lung epithelial cells and differentiate into glandular and squamous phenotypes, lining the larger airways and the peripheral small airways (2,3). LSCC exhibits many similarities with LA in terms of somatic copy number alterations (4), which raises the possibility of the presence of common genetic features between these two diseases (5,6).

Investigation of disease-associated genes can improve the understanding of disease etiology and development, thereby facilitating design and development of novel preventive and treatment strategies (7,8). Cross disease-gene studies and further pathway analyses provide an opportunity to resolve overlapping associations into discrete pathways and investigate possible shared etiologies (9,10).

The aim of the present study was to identify shared risk genes and to improve the understanding of shared pathways and biological mechanisms involved in LA and LSCC using a mega-analysis of gene expression data. Considering that the range of genetic alterations in LSCC is less understood compared with LA, the present study investigated genes that were involved in LA but not with LSCC using LSCC gene expression datasets.

Materials and methods

Study design. First, a large-scale literature-based analysis of disease-associated genes was performed to identify genes...
involved in LSCC and LA. Subsequently, for each of the LA-associated genes identified, a mega-analysis was performed using LSCC gene expression data. Pathway analysis was then performed to identify possible functional pathways associated with LSCC-specific genes. Finally, a co-expression-based protein-protein interaction (PPI) analysis was performed using LSCC expression data to evaluate the pathways identified. The workflow diagram is presented in Fig. 1.

**LA-and LSCC-associated gene data.** LA-and LSCC-associated gene data were acquired from the Pathway Studio (version 12.1.0.9; www.pathwaystudio.com) (11) mammalian database, which is a group of real-time updated literature knowledge databases, including curated signaling pathways, cellular processes, megabolic pathways, ontologies, annotations, molecular interactions and functional associations (http://pathwaystudio.gousinfo.com/ResNetDatabase.html). Association data were extracted from >41,000,000 references, including PubMed (https://www.ncbi.nlm.nih.gov/pubmed) abstracts and full-text articles. The Pathway database employs an automated natural language processing-based information extraction system, MedScan, with a precision >91% (12). Association data within the database are supported with one or more reference. The Pathway Studio ResNet Database is the largest literature database (13). These data were organized into a genetic dataset termed ‘LA_LSCC’, which is available at the Bioinformatics Database (http://database.gousinfo.com). The downloadable excel spreadsheet containing the dataset is available at http://gousinfo.com/database/Data_Genetic/LA_LSCC.xlsx. The full list of genes associated with LA and/or LSCC are presented in the groups ‘LA-alone genes’, ‘LSCC-alone genes’ and ‘Common genes’. In addition, the references for every disease-gene association are presented in the groups ‘Ref for LA-alone genes’ for LA-specific genes, ‘Ref for LSCC-alone genes’ for LSCC-specific genes and ‘Ref for common genes’ for shared genes. Information regarding the titles of the references and the sentences where the disease-gene associations were identified are presented in the ‘LA_LSCC’ dataset.

**Gene expression data selected for mega-analysis.** Following the initial search with ‘lung squamous cell carcinoma’, 158 microarray expression datasets were identified on gene expression omnibus (https://www.ncbi.nlm.nih.gov/geo/) (14,15). Subsequently, the following criteria were applied: i) The organism used in the study was *Homo sapiens*; ii) the data type was microarray expression profiling; and iii) the studies were limited to comparison between LSCC and healthy controls. A total of 12 datasets satisfied the inclusion criteria for the mega-analysis. However, one dataset (GSE27489 (16); www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE27489) was excluded from further investigation as each gene in this dataset demonstrated a small variation in expression level, which may lead to biased results in the mega-analysis. The 11 included datasets are listed in Table 1 (17-27).

**Mega-analysis models.** The log2 fold-change (LFC) of the gene expression level was used to indicate the effect size. Both fixed-effect and random-effects models were employed to investigate and compare the effect size (28). The heterogeneity of the mega-analysis was analyzed to study the variance within and between different studies. In the case that the total variance (Q) was equal to or smaller than the expected between-study variance (df), the within-study variance percentage (ISq) =100% x (Q-df)/Q was set at 0 and a fixed-effect model was selected for the mega-analysis. Otherwise, a random-effects model was selected. Q-p represents the probability that the total variance was only due to within-study variance. Significantly associated genes from this mega-analysis were identified using the following criteria: i) P<1x10^-7; and ii) LFC >1. When a gene exhibited a ILFC >1 in the mega-analysis, the change in the expression level of the gene was ≥2-fold or <0.5-fold. The current study presented all the mega-analysis results identified in the ‘Mega-analysis’ group in the ‘L.A_LSCC’ dataset; however, only genes with a ILFC >1 were further discussed. All analyses were performed using Matlab (version R2017a; https://www.mathworks.com/products/matlab.html).

**Multiple linear regression analysis.** A multiple linear regression (MLR) model was employed to investigate the possible influence of sample size, country of origin and study date on the gene expression in LSCC. P-values and 95% CIs were reported for each of these factors.

**Pathway analysis.** To test the functional profile of the common genes associated with LA and LSCC, a Gene Set Enrichment Analysis (GSEA) was conducted using Pathway Studio (version 12.1.0.9; www.pathwaystudio.com) against Gene Ontology (GO; http://geneontology.org) and Pathway Studio Ontology (version 12.1.0.9; www.pathwaystudio.com). In addition, functional pathway analysis was performed to investigate potential biological associations between the identified risk genes and LSCC. The analysis was performed using the ‘Shortest Path’ module of Pathway Studio in order to identify various ‘entities’, including complexes, proteins and functional classes, that were associated to both the genes and LSCC. The reference information included the types of associations, the number of underlying supporting references and the sentences where these associations had been identified and described.

**Co-expression analysis.** For each pair of the genes and proteins identified in the aforementioned pathway analysis, another mega-analysis was performed to investigate their co-expression using the 11 LSCC expression datasets. The Fisher's Z-value (FisherZ) of Pearson's correlation was used to determine the effect size, and the following equation was used to calculate it: FisherZ = 0.5 x log [(1+ Correlation)-(1- Correlation)]. The purpose of this analysis was to validate the associations identified in the pathway analysis. The present study used the following criteria for the selection of a non-random meaningful association: i) An absolute value of FisherZ >0.3; and ii) P<0.05. The detailed FisherZ values and P-values are presented in the ‘Co-expression’ analysis.

**Results**

**LA and LSCC genes.** LA-and LSCC-associated gene analyses revealed 1,178 genes associated with LA, supported by 7,355
references, and 334 genes associated with LSCC, supported by 838 references. The full list of these genes and the associated references are presented in the ‘LA_LSCC’ dataset. A significant overlap of 187 genes, which are presented in the ‘Common genes’ group, was identified for both LA and LSCC (right tail Fisher’s Exact test; P=1.02x10⁻¹⁶¹). This accounted for 55.99% of all the LSCC-associated genes and 15.87% of all the LA-associated genes.

To test the functional profile of the 187 common genes associated with both LA and LSCC, a GSEA was conducted using Pathway Studio against the GO and Pathway Studio Ontology. In total, nine pathways/gene sets (73 unique genes) associated with protein kinase, three pathways/gene sets (71 unique genes) associated with cell growth proliferation, two pathways/gene sets (nine unique genes) associated with cell apoptosis and one pathway/gene set (ten unique genes) associated with transcription factors were significantly enriched. The full list of the 39 pathways/gene sets enriched with P<1.7x10⁻⁵ (with 144 out of 187 unique genes) are presented in the ‘Common pathways’ group contained in the ‘LA_LSCC’ dataset. The majority of these pathways were involved in LA and LSCC, indicating a shared genetic basis for these diseases.

Three novel common genes in LA and LSCC. Although an overlap was identified between LA- and LSCC-associated genes, the majority of the LA-specific genes (991 genes, 84.13%) were not implicated in LSCC. A systematic mega-analysis was performed to collectively assess differential expressed mRNAs and determine whether previously investigated LA-associated genes were also linked to LSCC. Notably, certain datasets do not contain the three genes and therefore will not be included in the current study. However, the LFCs of the genes were estimated from the majority of the 11 studies (>9 studies). The associations between the LA-specific genes with 11 LSCC gene expression datasets (Table I) were evaluated. A total of three genes, including solute carrier family 2 member 1 (SLC2A1), endothelial PAS domain protein 1 (EPAS1) and cyclin-dependent kinase 4 (CDK4), passed the significant criteria (P<1x10⁻⁷ and |LFC| >1) and are presented in Table II. The detailed results are presented in the ‘Mega-analysis’ group in the ‘LA_LSCC’ dataset.

The effect sizes, 95% CIs and weights of different studies for the three identified genes (SLC2A1, EPAS1 and CDK4) are presented in Fig. 2. EPAS1 and CDK4 exhibited significant variances between studies (ISq >0% and Q-test P<0.1). Therefore, the random-effects model was selected for their mega-analysis. By contrast, no significant between-study variance was observed for SLC2A1 (Q-test P>0.4), and the fixed-effect model was selected for SLC2A1 (Fig. 2). Notably, multiple line regression analyses demonstrated that the country of origin was a significant factor that influenced the LFC of all three genes in the case of LSCC (P<0.004; Table II).

Figure 1. Workflow diagram of the cross LA-LSCC analysis. LA, lung adenocarcinoma; LSCC, lung squamous cell carcinoma; PPI, protein-protein interaction.

Table I. Datasets used for lung squamous cell carcinoma-gene association mega-analysis.

<table>
<thead>
<tr>
<th>Study name</th>
<th>Dataset GEO ID</th>
<th>Control (n)</th>
<th>Case (n)</th>
<th>Study age (years)</th>
<th>Country</th>
<th>(Refs.)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nazarov et al, 2017</td>
<td>GSE84784</td>
<td>9</td>
<td>9</td>
<td>2</td>
<td>Luxembourg</td>
<td>(17)</td>
</tr>
<tr>
<td>Tong et al, 2016</td>
<td>GSE67061</td>
<td>8</td>
<td>69</td>
<td>3</td>
<td>China</td>
<td>-</td>
</tr>
<tr>
<td>Mascaux et al, 2014</td>
<td>GSE33479</td>
<td>27</td>
<td>14</td>
<td>5</td>
<td>USA</td>
<td>-</td>
</tr>
<tr>
<td>Rousseaux et al, 2014</td>
<td>GSE30219</td>
<td>14</td>
<td>61</td>
<td>5</td>
<td>France</td>
<td>(18)</td>
</tr>
<tr>
<td>Girard et al, 2012</td>
<td>GSE32036</td>
<td>59</td>
<td>12</td>
<td>7</td>
<td>USA</td>
<td>(19,20)</td>
</tr>
<tr>
<td>Philipson et al, 2010</td>
<td>GSE19188</td>
<td>65</td>
<td>27</td>
<td>9</td>
<td>Netherlands</td>
<td>(21)</td>
</tr>
<tr>
<td>Boelens et al, 2009</td>
<td>GSE12472</td>
<td>28</td>
<td>35</td>
<td>10</td>
<td>Netherlands</td>
<td>(22)</td>
</tr>
<tr>
<td>Ishikawa et al, 2009</td>
<td>GSE2088</td>
<td>30</td>
<td>48</td>
<td>10</td>
<td>Japan</td>
<td>(23)</td>
</tr>
<tr>
<td>Boelens et al, 2008</td>
<td>GSE12428</td>
<td>28</td>
<td>34</td>
<td>11</td>
<td>Netherlands</td>
<td>(24)</td>
</tr>
<tr>
<td>Rosskopf et al, 2006</td>
<td>GSE6044</td>
<td>5</td>
<td>14</td>
<td>13</td>
<td>Germany</td>
<td>(25)</td>
</tr>
<tr>
<td>Takeuchi et al, 2009</td>
<td>GSE11969</td>
<td>5</td>
<td>35</td>
<td>10</td>
<td>Japan</td>
<td>(26,27)</td>
</tr>
</tbody>
</table>

GEO, gene expression omnibus; -, unavailable. Study age =current year -publication year+1.
Functional pathway analysis. According to the approach used to identify the genes associated with LSCC, SLC2A1, EPAS1, and CDK4 exhibited no direct link with LSCC. However, functional pathway analysis revealed multiple potential pathways through which these three genes may serve roles in the pathology of LSCC (Fig. 3). Each edge in Fig. 3 was supported by ≥1 references, and details of these associations are presented in the ‘LSCC-3Genes_potential pathways’ group in the ‘LA_LSCC’ dataset.

Table II. Statistically significant genes identified from the mega-analysis of lung squamous cell carcinoma.

<table>
<thead>
<tr>
<th>Gene name</th>
<th>Random effects model</th>
<th>Datasets included (n)</th>
<th>Mega-analysis results</th>
<th>MLR analysis results (P-values)</th>
</tr>
</thead>
<tbody>
<tr>
<td>SLC2A1</td>
<td>0</td>
<td>11</td>
<td>LFC 1.63  SD 0.25  P-value 4.31x10^{-11}</td>
<td>Sample size 0.59  Population region 4x10^{-3}  Study age 0.69</td>
</tr>
<tr>
<td>EPAS1</td>
<td>1</td>
<td>9</td>
<td>LFC -1.50  SD 0.26  P-value 1.27x10^{-8}</td>
<td>Sample size 0.49  Population region 2x10^{-5}  Study age 0.09</td>
</tr>
<tr>
<td>CDK4</td>
<td>1</td>
<td>11</td>
<td>LFC 1.02  SD 0.18  P-value 1.60x10^{-8}</td>
<td>Sample size 0.87  Population region 6x10^{-5}  Study age 0.98</td>
</tr>
</tbody>
</table>

SLC2A1, solute carrier family 2 member 1; EPAS1, endothelial PAS domain protein 1; CDK4, cyclin-dependent kinase 4; LFC, log-fold change; SD, standard deviation. D Study age = current year - publication year + 1.

Figure 2. Effect size, 95% CI and weights for the genes SLC2A1, EPAS1, and CDK4. Results from a mega-analysis using (A) a fix-effect model for SLC2A1, and a random-effects model for (B) EPAS1 and (C) CDK4. SLC2A1, solute carrier family 2 member 1; EPAS1, endothelial PAS domain protein 1; CDK4, cyclin-dependent kinase 4.

To confirm the associations presented in Fig. 3, a co-expression PPI analysis was conducted with the purpose of validating the associations between CD4K4, EPAS1 and SLC2A1, and the 13 other genes presented in Fig. 3. The majority of the entities presented in Fig. 3 also exhibited significant associations in the co-expression analysis (Fig. 4), supporting the pathway analysis results. Co-expression analysis results are presented in the ‘Co-expression’ group in the ‘LA_LSCC’ dataset.
Figure 3. Potential pathways associating SLC2A1, EPAS1, and CDK4 to lung squamous cell carcinoma. Network was generated using Pathway Studio. Each association (edge) has ≥1 supporting reference. SLC2A1, solute carrier family 2 member 1; EPAS1, endothelial PAS domain protein 1; CDK4, cyclin-dependent kinase 4; PTGS2, prostaglandin-endoperoxide synthase 2; CDH1, cadherin 1; NOS2, nitric oxide synthase 2; GLI1, GLI family member zinc finger 1; FGFR, fibroblast growth factor receptor; BCL2L1, BCL2-like 1; FOXM1, forkhead box M1; WNT5A, Wnt family member 5A; BIRC5, baculoviral IAP repeat containing 5; MIR21, microRNA-21; GPC3, glypican 3; GSK3B, glycogen synthase kinase 3β.

Figure 4. Co-expression analysis. Each edge represents a significant association between the corresponding two entities (P<0.05). Positive associations are highlighted in green and negative associations are highlighted in red. Nodes in red indicate decreased expression and green nodes indicate increased expression. SLC2A1, solute carrier family 2 member 1; EPAS1, endothelial PAS domain protein 1; CDK4, cyclin-dependent kinase 4; PTGS2, prostaglandin-endoperoxide synthase 2; CDH1, cadherin 1; NOS2, nitric oxide synthase 2; GLI1, GLI family member zinc finger 1; BCL2L1, BCL2-like 1; FOXM1, forkhead box M1; CD274, cluster of differentiation 274; WNT5A, Wnt family member 5A; BIRC5, baculoviral IAP repeat containing 5; MIR21, microRNA-21; GPC3, glypican 3.
Discussion

The cross-analysis of different lung cancer phenotypes may facilitate the development of novel strategies and approaches for the treatment of lung cancer. In the present study, LA-specific genes were systematically mega-analyzed with LSCC differential expression data and three genes, including SLC2A1, EPAS1 and CDK4, were identified as potential risk genes for LSCC. Importantly, whether these associations between the genes and LSCC indicate causality requires further investigation.

It is a major concern that a disease-gene association derived from experiment-based literature is heavily dependent on the quality and access of the text data, and the efficiency of the mining algorithms. Candidate disease-gene analysis is more appropriate for monogenic diseases because the association between genotype and phenotype is clearer (41,42). In lung cancer, a complex disease, the etiology can be attributed to tobacco smoking, sex, ethnicity, age, diet, obesity, infections and numerous genes that work in combination to elicit the disease phenotype (29-31). It has also been observed that when individually investigated, the genes potentially responsible for the disease may not result in disease in certain patients (32-35).

In this context, cross-disease analysis based on mega-analysis can overcome the limitations of sample size and identify more reliable and robust common genes between LA and LSCC through the quantitative combination and assessment of multiple studies (36,37). In the present study, disease-gene association data were retrieved from the Pathway Studio database and mega-analysis was performed to detect their significance in terms of gene expression levels. All of these analyses can provide a more reliable and robust result.

The present study used MLR analysis to demonstrate that lung cancer outcome varies among different populations and ethnicities. In addition, the present study identified that the country of origin may be associated with the expression levels of SLC2A1, EPAS1 and CDK4 in the case of LSCC. It is therefore necessary to assess the generalizability of the present results in different ethnic groups. Socioeconomic and cultural differences among different racial groups may account for some degree of the current disparities and a personalized molecular approach may help to resolve such problems (38-41).

The current literature-based functional pathway analysis revealed several possible pathways that link the three novel genes identified to LSCC. For example, CDK4, a member of the serine/threonine protein kinase family, may contribute to the development of LSCC via a CDK4-forkhead box M1 (FOXM1)-LSCC pathway. It has been reported that CDK4 activity can increase the transcriptional activity of FOXM1 without phosphorylating FOXM1 (42), while the expression of FOXM1 has been suggested to contribute to the development or progression of LSCC (43). A previous study also suggested that CDK4 can stimulate the BRCA1 promoter in an E2F transcription factor 1-dependent manner, regulating cell cycle, DNA replication and cell proliferation processes (44). BRCA1 serves an important role in LSCC via cell cycle and DNA replication signaling pathways (44), which indicates a potential CDK4-BRCA1-LSCC pathway.

EPAS1 can bind to and inhibit the expression of the calcium-dependent cell adhesion molecule cadherin 1 (CDH1) (45,46). CDH1 has been reported to serve a dual role in the maintenance of the LSCC phenotype (47). These previous studies indicate that the EPAS1-CDH1-LSCC pathway may serve a complex role in LSCC development. EPAS1 regulates the production of prostaglandin-endoperoxide synthase (PTGS) (48), which has been indicated to promote the carcinogenesis of LSCC, suggesting a potential EPAS1-PTGS-LSCC pathway.

SLC2A1 is a major glucose transporter responsible for constitutive or basal glucose uptake, which can bind with glypican 3 (GPC3) to decrease glucose transport activity (49) and transport quercetin to balance the glucose efflux (50). Mechanisms associated with glucose efflux are also identified in the pathological process of LSCC (51,52), suggesting the existence of SLC2A1-GPC3-LSCC and SLC2A1-quercetin-LSCC pathways.

Co-expression analysis revealed that the majority of the identified genes were associated with each other in terms of expression. The majority of the literature-based pathway identified was validated by the expression data-based associations found. However, a certain number of the associations identified in the present study may not be consistent with the present co-expression analysis. For example, SLC2A1 inhibits the expression of baculoviral IAP repeat containing 5 (BIRC5) in the pathway analysis; however, SLC2A1 and BIRC5 exhibit positive co-expression in the co-expression analysis, which indicates the presence of a more complex genetic network including more regulators.

The present cross-disease analysis between LA and LSCC suggested common genes may contribute to disease comorbidities and trait manifestations. The novel common genes identified may facilitate the development of novel strategies targeting shared mechanisms across diseases. However, the conclusion of the current study was only based on a statistical analysis of previous experimental data and a literature-based pathway study. Therefore, further biological experiments, including gene-knockout or knockdown experiments, are required to validate the associations between the three genes identified and LSCC.

In conclusion, cross-disease analysis could provide a powerful tool to investigate new targets and reveal common biological mechanisms. Genes associated with LA require further analysis to identify their association with LSCC. SLC2A1, EPAS1 and CDK4 genes identified in the present study may be novel common risk genes associated with both LA and LSCC.

Acknowledgements

Not applicable.

Funding

No funding was received.

Availability of data and materials

The datasets generated and/or analyzed during the current study are available in the http://database.gousinfo.com repository, http://gousinfo.com/database/Data_Genetic/LA_LSCC.xlsx.
Authors' contributions

GZ and ZL developed the study design, supervised the whole the study process and prepared the manuscript. WW and HC contributed to data analysis and manuscript drafting and revision. WH and XX contributed to data collection and manuscript drafting and revision. All authors approve the final version of the manuscript.

Ethics approval and consent to participate

Not applicable.

Patient consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

References


This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International (CC BY-NC-ND 4.0) License.