Libertas Academica
FREEDOM TO RESEARCH

Open Access: Full open access to this and thousands of other papers at http://www.la-press.com.

Cancer Informatics

Supplementary Issue: Computational Advances in Cancer Informatics (A)

Revealing Biological Pathways Implicated in Lung Cancer from TCGA Gene Expression Data Using Gene Set Enrichment Analysis

Binghuang Cai and Xia Jiang

Department of Biomedical Informatics, School of Medicine, University of Pittsburgh, Pittsburgh, PA, USA.

ABSTRACT: Analyzing biological system abnormalities in cancer patients based on measures of biological entities, such as gene expression levels, is an important and challenging problem. This paper applies existing methods, Gene Set Enrichment Analysis and Signaling Pathway Impact Analysis, to pathway abnormality analysis in lung cancer using microarray gene expression data. Gene expression data from studies of Lung Squamous Cell Carcinoma (LUSC) in The Cancer Genome Atlas project, and pathway gene set data from the Kyoto Encyclopedia of Genes and Genomes were used to analyze the relationship between pathways and phenotypes. Results, in the form of pathway rankings, indicate that some pathways may behave abnormally in LUSC. For example, both the cell cycle and viral carcinogenesis pathways ranked very high in LUSC. Furthermore, some pathways that are known to be associated with cancer, such as the p53 and the PI3K-Akt signal transduction pathways, were found to rank high in LUSC. Other pathways, such as bladder cancer and thyroid cancer pathways, were also ranked high in LUSC.

KEYWORDS: gene expression, GSEA, SPIA, TCGA, LUSC, signal transduction pathway

SUPPLEMENT: Computational Advances in Cancer Informatics (A)

CITATION: Cai and Jiang. Revealing Biological Pathways Implicated in Lung Cancer from TCGA Gene Expression Data Using Gene Set Enrichment Analysis. *Cancer Informatics* 2014:13(S1) 113–121 doi: 10.4137/CIN.S13882.

RECEIVED: April 16, 2014. **RESUBMITTED:** September 5, 2014. **ACCEPTED FOR PUBLICATION:** September 9, 2014.

ACADEMIC EDITOR: JT Efird, Editor in Chief

TYPE: Original Research

FUNDING: Research reported in this publication was supported by the National Library of Medicine of the National Institutes of Health under Grant R00 LM010822 and Grant R01 LM011663. The authors confirm that the funder had no influence over the study design, content of the article, or selection of this journal.

COMPETING INTERESTS: Authors disclose no potential conflicts of interest.

CORRESPONDENCE: bic9@pitt.edu; bhcai8@gmail.com

Introduction

Cancer is a disease involving unregulated cell growth, a process in which cells divide and grow uncontrollably, forming malignant tumors that invade other parts of the body.[1] There are many different types of cancers affecting humans, such as breast cancer, lung cancer, and bladder cancer. The causes of cancer are diverse and complex. A signaling pathway is a series of actions among molecules occurring within a cell. Such pathways are important biological mechanisms in cell growth.[2] Determining how pathways and the genes therein are related to cancer is one of most essential problems investigated by cancer researchers in the past couple of decades.[3–5]

With the rapid development of information technology and medical equipment comes the ability to collect more and more data, including clinical and genomic information that can be used to improve medical knowledge and treatment.[6–9] One of the growing types of data is that obtained from DNA microarray, a collection of microscopic DNA spots attached to a solid surface.[6,7] Microarray data are used to measure the expression levels of large numbers of genes simultaneously. What biological insights can be gleaned from the data? How are the gene expression data related to disease phenotypes (especially cancer phenotypes)? How does the expression of genes in different pathways function in cancer and what pathways might be targeted for cancer drug treatment? These are all important and challenging questions that are driving current genomics research.

Techniques for analyzing gene expression data. Of the many methods employed to analyze gene expression data for insight into the biology of diseases (especially cancers),[6,7,9,10] most can be classified into two types: single-gene analysis and gene set analysis.[10,11] The single-gene analysis method is a conventional statistical analysis of the gene expression data that examines one gene at a time. The method determines the differentially expressed (DE) levels of the gene in different phenotypes and then makes adjustments to the levels for multiple gene testing. This method, however, possesses several limitations: high-ranking genes may score highly simply by chance, given the large number of hypotheses involved; significant genes may show distressingly little overlap among different studies of the same biological system; and analysis may miss important effects of sets of genes in pathways.[7,11,12] Because of the limitations of single-gene analysis, researchers have increasingly turned to the development of gene set analysis methods,[10–21] which consider a set of genes as a whole and determine its correlation with disease phenotypes based on the differing levels of the genes' expression. Different gene set analysis methods, which either find gene sets that were previously unknown or select gene sets in a known collection (such as known pathways), have been proposed for genomic data analysis.[12–24]

Gene Set Enrichment Analysis (GSEA) uses overrepresentation analysis to determine if given sets of genes are DE in different disease phenotypes and has been widely adopted to analyze data in biological experiments.[11,14] The goal of GSEA is to determine if members of a gene set tend to occur toward the top of the gene list because of the genes' correlation with the phenotypic class distinction.[12] The given gene set can be a set of genes in a pathway, a set of genes in a gene ontology category, or any user-defined set. A preliminary version of GSEA[13] was proposed and designed to detect modest but coordinate changes in the expression of groups of functionally related genes using data from muscle biopsies from diabetics and healthy controls. An improved version of GSEA,[12] designed to interpret gene expression data including that derived from leukemia and lung cancer, has been developed; results demonstrate GSEA's effectiveness for relationship analysis of gene sets and phenotypes. Bioinformatics researchers (including the original inventors themselves) have also developed different GSEA extensions for genomic data analysis.[11,12,14,22,25]

GSEA ignore the topology of the pathway and so does not account for key biological information. Signaling pathway impact analysis (SPIA)[15–20] analyzes gene expression data to identify whether a pathway is implicated by combining overrepresentation analysis with a measurement of the perturbation measured in a pathway.

Related work on discovery of aberrant pathways in cancer. The complex procedure of finding pathway abnormalities in lung cancer could have many steps involved, such as information extraction from biological data, simulation verification, biological experimental testing, and clinical trials. Among these steps, analysis based on biological data to determine the relationship of pathways (and the gene sets therein) to lung cancer is one of the most important steps and has been investigated in different ways.[21,26–31] For example, the relationship of signaling pathways and squamous cell lung carcinoma was investigated by Shi et al,[21] in which Fisher's exact test was used to identify the related pathways based on the significance level of DE genes. The study showed that over 100 signaling pathways, including the cell cycle regulation pathway and the p53 tumor-suppressor pathway, were implicated in squamous cell lung carcinoma. The study used only the selected and altered genes in the pathway to find the lung cancer–related pathways. Another study by Qian et al.[26] investigated several single-nucleotide polymorphisms (SNPs), the loci of genes that encode proteins on the DNA repair pathways (including the both excision repair [BER] pathway and the nucleotide excision repair [NER] pathway), to determine whether these SNPs are associated with non–small-cell lung cancer. The study, which considered only some of the SNPs related to the pathways, instead of all the genes involved in the pathway, showed that the NER pathway seems to have a greater influence on lung cancer than the BER pathway.

Another study on PI3K pathway activity in lung cancer development used computational and biochemical measurements to show their close relationship.[27] In the work of Toonke et al.[28], the supervised analysis of messenger RNA microarray data from human tumors identified the transforming growth factor-β signaling pathway as an important mediator of tumor invasion. As described by Ekman et al.[29], aberrant activation of the Akt/mTOR pathway is commonly observed in lung cancer, while deregulated PI3K/Akt/mTOR activity is known to contribute to the development and maintenance of lung cancer.

Most these studies were based on analysis of selected biological features (eg, SNP or the expression level of genes) rather than all the features in the pathways, and many of them focused on studying one or a few pathways rather than many pathways. More closely related to the work presented here, Neapolitan et al.[32] and Neapolitan and Jiang[33] developed a new method for learning aberrant signal pathways from data called causal analysis of signal transduction pathway aberrations (CASA) and used both CASA and SPIA to analyze the The Cancer Genome Atlas (TCGA) breast cancer data set[32] and the TCGA ovarian cancer data set.[33]

Application of GSEA and SPIA to TCGA lung cancer data. This paper applies the GSEA and SPIA methods to a well-known data set (TCGA[18] lung squamous cell carcinoma [LUSC][30] expression data set). Specifically, one of our main purposes is to mine interesting biological information from TCGA LUSC expression data related to pathway implications in lung cancer. We also hope to obtain information from the data to confirm some existing knowledge, especially pathway abnormalities in LUSC. We further hope to reveal new possible LUSC-related pathway implications for subsequent biological testing. Therefore, we investigated the implications of a rich set of pathways in lung cancer by mining the microarray gene expression data. GSEA was employed to rank the pathways based on their correlations with the phenotypes. Experiments employed the LUSC expression data from the TCGA project,[18] as well as data (especially gene set data) on 26 pathways from the Kyoto Encyclopedia of Genes and Genomes (KEGG),[33] a database resource for understanding high-level functions and utilities of the biological system, especially large-scale molecular data sets generated by genome sequencing and other high-throughput experimental technologies. Our experiments compared results in the form of pathway rankings that revealed that some pathways may be highly implicated in LUSC.

The remainder of this paper is organized in five sections: Section 2 introduces data sets and data-processing methods; Section 3 describes the experiment's approach of applying GSEA and SPIA to pathway analysis for lung cancer; in Section 4, experimental results are described, analyzed and compared; Section 5 discusses our results; and Section 6 provides final remarks.

Data Sets

This paper investigates the implications of biological pathways in LUSC using gene expression data obtained from the data portal of TCGA.[18] Pathway information was gathered from the KEGG.[34]

TCGA gene expression data. TCGA is a well-known project in cancer research that collects and analyzes high-quality tumor samples and makes the related data available to researchers. At the TCGA data portal, researchers can search, download, and analyze data from approximately 30 different tumor types. Our research explored the LUSC Level 3 gene expression data set, which encompasses 17,814 genes/features and 256 tumor samples, including 101 normal tissue samples and 155 LUSC tumor samples.

The many studies employing TCGA LUSC data include more than 80 recent publications.[18,30,35–38] For instance, Liu et al.[35] analyzed the TCGA LUSC data to identify both common and unique mutation spectra and pathway activation, which used whole-exome sequencing technology rather than the GSEA method. Győrffy et al.[36] developed a real-time meta-analysis tool for the TCGA microarray data sets to identify biomarkers related to survival, which was an analysis based on signal gene rather than gene set in our paper. Deng et al.[37] used the TCGA LUSC data to determine the prognostic value of BCAR1 expression and its associations with clinical–demographical characteristics, which is different from the purpose of this paper to mine information of pathway implication. Barrett et al.[38] analyzed the TCGA LUSC transcriptome data to identify potential therapeutic strategies for squamous cell lung carcinoma, which used TCGA LUSC RNA-seq data rather than gene expression data in this paper. Jiang et al.[39] used GSEA to indicate enriched genes in order to evaluate the messenger RNA expression of insulin receptor isoform A and insulin receptor isoform B, which used the TCGA LUSC RNA-seq data rather than the gene expression data in our paper. Ylipää et al.[40] analyzed the TCGA LUSC gene expression data for the similarities with other types of cancers based on a GSEA-inspiring method of computing the pathway aberration profile for each tumor sample, which was to study the similarities of different cancers rather than to analyze the pathway implication in LUSC in this paper. Above all, to our knowledge, the TCGA LUSC gene expression data, together with the whole set of genes in the pathway, might not have been used to analyze the pathway implications in LUSC via GSEA.

KEGG pathway gene set data. In this research, 26 *Homo sapiens* pathways were selected from pathways previously known to be related to different types of cancers (eg, bladder cancer, chronic myeloid leukemia, colorectal cancer, lung cancer, pancreatic cancer, skin cancer, and thyroid cancer), as well as from noncancer-associated pathways. Pathways previously identified as being related to LUSC (eg, the cell cycle pathway,[30] the p53 tumor-suppressor pathway,[21] and the mTOR pathway[29]) were also included. The map data and gene sets of the 26 pathways are from KEGG.[34] The selected 26 pathways and the number of genes on the pathways are listed in Table 1.

Data preprocessing method. The LUSC Level 3 gene expression data, downloaded from TCGA's portal, were collected by UNC (University of North Carolina at Chapel Hill) using the AgilentG45502A_07 platform. The downloaded data consist of many individual files, including one for each tissue sample. We extracted the relevant information from these individual files and generated a single file that contains gene expression profiles for all the tissue samples. In this regenerated file, a row contains gene expression information for a particular gene in all samples and a column contains gene expression information for a tissue sample. The first column contains the gene names and the first row contains the TCGA sample IDs.

Because there are missing values in the downloaded data, imputation was conducted to handle the missing value problem based on the known data. We used a mean imputation program (available at http://www.bioconductor.org/packages/release/bioc/html/impute.html) to compute the missing values.

Methods

This section describes our application of GSEA and SPIA to performing pathway abnormality analysis using gene expression data.

GSEA. GSEA was developed to determine DE levels of a predefined gene set in two different phenotypes. Genome-wide expression profiles from two-class samples were used to rank all genes in the data set, and the ranking list was then used to calculate enrichment score (ES) and *P* value. The procedure included obtaining the gene-ranking list, calculating an ES, estimating

Table 1. List of pathways and the number of genes in each pathway.

NO.	NAME OF PATHWAY	NUMBER OF GENES
1	Bladder cancer	38
2	Cell cycle	124
3	Chronic myeloid leukemia	73
4	Colorectal cancer	62
5	Complement and coagulation cascades	68
6	ErbB signaling pathway	88
7	Glioma	65
8	Hedgehog signaling pathway	51
9	Melanoma	71
10	mTOR signaling pathway	60
11	Non–small-cell lung cancer	56
12	Notch signaling pathway	48
13	p53 signaling pathway	68
14	Pancreatic cancer	66
15	PI3K-Akt signaling pathway	344
16	Protein processing in endoplasmic reticulum	167
17	Ras signaling pathway	226
18	Salivary secretion	90
19	Small-cell lung cancer	86
20	Transforming growth factor-beta signaling pathway	80
21	Thyroid cancer	29
22	Type I diabetes mellitus	47
23	Type II diabetes mellitus	48
24	Viral carcinogenesis	207
25	Viral myocarditis	74
26	Wnt signaling pathway	141

the significance level of the ES, and correcting the significance level for multiple gene sets.[11–13] Details of the steps of the GSEA procedure appear in.[11–13] GSEA has been implemented in JAVA and R, and different versions of GSEA packages can be downloaded at www.broadinstitute.org/gsea/index.jsp.

We used the javaGSEA desktop application with graphical user interface from the GSEA Web site (www.broadinstitute.org/gsea/index.jsp). Different special types of data files are required for gene set analysis. In our analysis, we used the following three types of file formats for gene expression data, phenotype data, and gene set data.

1. TXT gene expression data file: The expression data CSV file is converted to tab-delimited TXT file format, and a column of gene description is added as the second column in the data set. The added column is used to describe each gene. If there is no description for a gene, the value can be simply set to "na."
2. CLS phenotype data file: The phenotype of each tissue sample is formatted in the CLS phenotype file. The first row of the CLS is the total number of samples and total number of phenotypes, separated by a space and terminated by the constant 1. The second row is the visible names of the phenotypes, such as case and control. The third row is the phonotype of each sample, separated by a space. For the LUSC gene set analysis, we have two phenotypes, ie, control-normal sample and case-tumor sample. There are 101 normal control samples and 155 tumor case samples. The control sample is set to 0 and the case sample is set to 1 in the third row of the CLS file.
3. GMT gene set data file: All gene sets are combined in one tab-delimited GMT file. Each row represents one gene set, with the name of the gene set in the first column, the description of the gene set in the second, and genes in the gene set in the subsequent columns (one gene per column). Our analysis generated 26 rows, with the names of the 26 pathways (listed in the second column of Table 1) placed in the first columns. The list of genes in each pathway is from the KEGG Web site.[34]

Finally, we imported the data into the GSEA application and set the following parameters. The parameter "Number of permutations" is set as default (ie, 1000, based on the GSEA theory discussed in[12]). The parameter "Collapse data set to gene symbols" is set to "false," as we use gene symbols in gene expression data and do not need to do any mapping from probe ID to gene. The parameter "Enrichment statistic" is set to "weighted," as we use the improved version of GSEA (see[12]). The parameter "Metric for ranking genes" is set to "Singal2Noise," while "Gene list of sorting mode" is set to "abs." Other parameters are set to their default values.

The ES, normalized ES (NES), *P* value, and false discovery rate (FDR) *Q* value were obtained from the GSEA output reports, which were then used to rank the gene sets. Definitions of these output variables can be found in.[12]

SPIA. SPIA combines the evidence obtained from data on differential expression of genes with measurement of the actual perturbation on a given pathway under a given condition. SPIA calculates a global pathway significance *P* value combining the DE and perturbation *P* values.[15] The main steps of SPIA include calculating DE probability, perturbation probability, and global probability.[15] For detailed theoretical analysis and procedures, please refer to Tarca et al.[15] SPIA has been implemented as a standard R library and can be downloaded at http://www.bioconductor.org/packages/release/bioc/html/SPIA.html.

We used the SPIA R Library to analyze the 26 pathways in LUSC by following these procedures:

1. Reprocessing the gene expression data: The first row of sample ID was deleted from the expression CSV file

generated in Section 2, according to the file-formatting requirements of SPIA.

2. Getting the list of DE genes: Linear model fitting and empirical Bayes statistics were used in our experiments. We used "limma," an R package employed for the analysis of gene expression data arising from microarray technologies.[41] The functions "lmFit" (implementation of linear model fitting) and "eBayes" (implementation of empirical Bayes statistics) are used to estimate the fold changes and standard errors by fitting a linear model for each gene. The list of DE genes is obtained by using the function "topTable" in the "limma" package. Note that phenotypes of all samples (ie, case and control) are reformatted as a design matrix in the "limma" R package and input into "lmFit." For details and to download the "limma" package, please visit http://www.bioconductor.org/packages/release/bioc/html/limma.html.
3. Getting the pathway topology: The topologies of 26 pathways were downloaded from the KEGG Web site in the file format XML.[34] The XML files were formatted using KGML (KEGG Markup Language) to represent the pathway maps.
4. Running SPIA: The list of all genes, the list of all DE genes, and the 26 pathway KGML files were input into the SPIA function in its R library.

The numbers of DE genes (N_{DE}), DE probability (P_N), perturbation probability (P_P), global probability (P_G), and global probability of FDR (P_{GFDR}) were obtained from the SPIA output reports and then used to rank the pathways.

All experiments were run on a high-performance computer that has two AMD Opteron™ 4280 2.80-GHz 8-core processors, 128 GB of memory, and a Windows Server 2008 R2 Enterprise operating system.

Results

The results of using GSEA and SPIA to discover implicated pathways in LUSC are shown in Tables 2–4 and Figure 1.

Table 2 shows results from GSEA including ES, NES, *P* values, and FDR *Q* values. The ranking of each pathway is derived from the NES values. The cell cycle pathway is ranked first, with an extremely small *P* value (rounded to 0). The next three top-ranked pathways, ie, the p53 signaling pathway,

Table 2. Results from GSEA (column NES is the ranking metric).

RANK	NAME OF PATHWAY	ES	NES	*P* VALUE	FDR *Q* VALUE
1	Cell cycle	0.561028	1.742564	0.000000	0.003000
2	p53 signaling pathway	0.535046	1.599214	0.001000	0.011000
3	Bladder cancer	0.516461	1.478080	0.012012	0.040013
4	Thyroid cancer	0.477446	1.317630	0.040120	0.215370
5	Type I diabetes mellitus	0.488248	1.313840	0.082996	0.178696
6	Viral carcinogenesis	0.408460	1.301336	0.024000	0.172586
7	Type II diabetes mellitus	0.448148	1.286026	0.064000	0.172811
8	Notch signaling pathway	0.430467	1.243207	0.082000	0.234539
9	ErbB signaling pathway	0.410243	1.241607	0.063000	0.211702
10	Wnt signaling pathway	0.373434	1.173343	0.090000	0.366271
11	mTOR signaling pathway	0.388213	1.142176	0.213000	0.443047
12	Transforming growth factor-beta signaling pathway	0.373428	1.120824	0.200000	0.482590
13	Melanoma	0.371462	1.107658	0.252000	0.491501
14	Colorectal cancer	0.375154	1.106625	0.235000	0.459683
15	Salivary secretion	0.360758	1.089830	0.268000	0.486541
16	Protein processing in endoplasmic reticulum	0.347418	1.084794	0.249000	0.471890
17	Hedgehog signaling pathway	0.363243	1.064658	0.352000	0.506981
18	Glioma	0.350125	1.035410	0.398000	0.578538
19	Non–small-cell lung cancer	0.347019	1.023491	0.420000	0.586171
20	Pancreatic cancer	0.336510	0.993367	0.522000	0.656374
21	Ras signaling pathway	0.313613	0.991516	0.518000	0.631262
22	PI3K-Akt signaling pathway	0.298414	0.962014	0.624000	0.687766
23	Complement and coagulation cascades	0.319877	0.960393	0.537000	0.662823
24	Chronic myeloid leukemia	0.319337	0.956266	0.603000	0.645878
25	Viral myocarditis	0.291712	0.854560	0.797000	0.849612
26	Small-cell lung cancer	0.264050	0.802581	0.899000	0.889872

Table 3. Results from SPIA (column P_G is the ranking metric).

RANK	NAME OF PATHWAY	N_{DE}	P_N	P_P	P_G	P_{GFDR}
1	Viral carcinogenesis	159	0.002331	0.992	0.016347	0.267259
2	Melanoma	55	0.559829	0.010	0.034627	0.267259
3	Protein processing in endoplasmic reticulum	126	0.225201	0.027	0.037107	0.267259
4	PI3K-Akt signaling pathway	259	0.212714	0.040	0.049066	0.267259
5	Cell cycle	96	0.079648	0.113	0.051396	0.267259
6	Transforming growth factor-beta signaling pathway	66	0.074211	0.618	0.187216	0.631745
7	Bladder cancer	30	0.440452	0.109	0.193783	0.631745
8	Type I diabetes mellitus	18	0.626067	0.077	0.194383	0.631745
9	Thyroid cancer	24	0.383301	0.154	0.226063	0.653071
10	Viral myocarditis	39	0.611114	0.113	0.253631	0.659441
11	Wnt signaling pathway	106	0.607008	0.153	0.313586	0.741203
12	Small-cell lung cancer	66	0.553092	0.207	0.362621	0.753911
13	Colorectal cancer	48	0.648092	0.187	0.376956	0.753911
14	p53 signaling pathway	54	0.158231	0.942	0.432769	0.788239
15	Ras signaling pathway	172	0.424384	0.379	0.454753	0.788239
16	Non–small-cell lung cancer	44	0.467202	0.408	0.506565	0.789732
17	Notch signaling pathway	37	0.343326	0.598	0.530363	0.789732
18	Salivary secretion	69	0.249502	0.865	0.546737	0.789732
19	ErbB signaling pathway	63	0.865496	0.313	0.624697	0.818210
20	Chronic myeloid leukemia	57	0.510093	0.599	0.667818	0.818210
21	mTOR signaling pathway	40	0.830403	0.381	0.680478	0.818210
22	Hedgehog signaling pathway	41	0.447103	0.731	0.692331	0.818210
23	Glioma	45	0.931147	0.487	0.812085	0.917039
24	Pancreatic cancer	52	0.545290	0.925	0.849599	0.917039
25	Complement and coagulation cascades	46	0.966263	0.639	0.915154	0.917039
26	Type II diabetes mellitus	34	0.750456	0.828	0.917039	0.917039

bladder cancer pathway, and thyroid cancer pathway, also exhibit small *P* values (<0.05). The small *P* values suggest that their rankings are not the result of random chance, but are significantly linked to their relationship with LUSC.

Table 3 shows results from SPIA, including the number of DE genes, the DE probability, accumulated perturbation, perturbation probability, global probability, and the global probability of FDR. The rankings are based on the values of global probability. The top 5 pathways, ie, the viral carcinogenesis, melanoma, protein processing in endoplasmic reticulum, PI3K-Akt signaling, and cell cycle pathways, exhibit significantly small global probabilities (around 0.05 or smaller) and the same global probability of FDR. Because these five pathways are significantly high ranked, rather than ranked highly by chance in views of small *P* values (ie, global probabilities), the rankings could indicate that these pathways are more closely implicated in LUSC than the other 21 pathways.

Table 4 and Figure 1 provide a pathway ranking comparison of GSEA with SPIA. Here, we can see that the cell cycle pathway (the second pathway in Table 1 and Fig. 1) and the viral carcinogenesis pathway (the 24th pathway in Table 1 and Fig. 1) are ranked by both GSEA and SPIA at the very top. The three pathways ranked highest based on our GSEA results, the bladder cancer, thyroid cancer, and type I diabetes mellitus pathways, are also ranked among the SPIA results' top 10. This finding increases our confidence that these three pathways may be implicated in LUSC. We note that some pathways are ranked differently by GSEA as compared with SPIA: for example, p53, ranked number 2 according to GSEA, comes in at 13, according to SPIA. The type II diabetes mellitus pathway is ranked number 7 by GSEA but is ranked last by SPIA. Protein processing in the endoplasmic reticulum pathway is ranked only number 16 by GSEA but is ranked third by SPIA. GSEA ranks the PI3K-Akt signaling pathway at number 22, whereas SPIA puts it at fourth. According to the methodologies of GSEA and SPIA, the consideration of pathway ontologies in SPIA may be an important reason for these differences.

We also calculated the correlation of ranking numbers between GSEA and SPIA and determined that the correlation coefficient of all 26 pathways is 0.2855; this finding demonstrates that the GSEA and SPIA results share some

Table 4. Comparisons of rankings from GSEA and SPIA.

NAME OF PATHWAY	GSEA	SPIA
Cell cycle	1	5
p53 signaling pathway	2	14
Bladder cancer	3	7
Thyroid cancer	4	9
Type I diabetes mellitus	5	8
Viral carcinogenesis	6	1
Type II diabetes mellitus	7	26
Notch signaling pathway	8	17
ErbB signaling pathway	9	19
Wnt signaling pathway	10	11
mTOR signaling pathway	11	21
Transforming growth factor-beta signaling pathway	12	6
Melanoma	13	2
Colorectal cancer	14	13
Salivary secretion	15	18
Protein processing in endoplasmic reticulum	16	3
Hedgehog signaling pathway	17	22
Glioma	18	23
Non–small-cell lung cancer	19	16
Pancreatic cancer	20	24
Ras signaling pathway	21	15
PI3K-Akt signaling pathway	22	4
Complement and coagulation cascades	23	25
Chronic myeloid leukemia	24	20
Viral myocarditis	25	10
Small-cell lung cancer	26	12

similar rankings. Moreover, the correlation coefficient of GSEA and SPIA rankings of the top 10 pathways is 0.4532 (bigger than the one for 26 pathways), suggesting that the rankings of the top 10 pathways by GSEA and SPIA have more common pathways as compared with the ones for all 26 pathways. That is, the pathways ranked among the top 10 by GSEA and SPIA have more in common than the pathways ranked lower by both. Thus, pathways ranked at the top by SPIA could also be ranked at the top by GSEA, strengthening our confidence that pathways ranked at the top in GSEA (eg, the cell cycle pathway and the viral carcinogenesis pathway) could be implicated in LUSC.

Above all, GSEA proved useful in extracting information from TCGA LUSC gene expression data and showing implications of pathways and LUSC, compared to SPIA.

Discussion

The results described above reveal that the ranking scores are significantly higher for some pathways than for others in LUSC. It is noteworthy that some high-scoring abnormal pathways in LUSC have never been reported anywhere else based on our best knowledge. A discussion of the results follows.

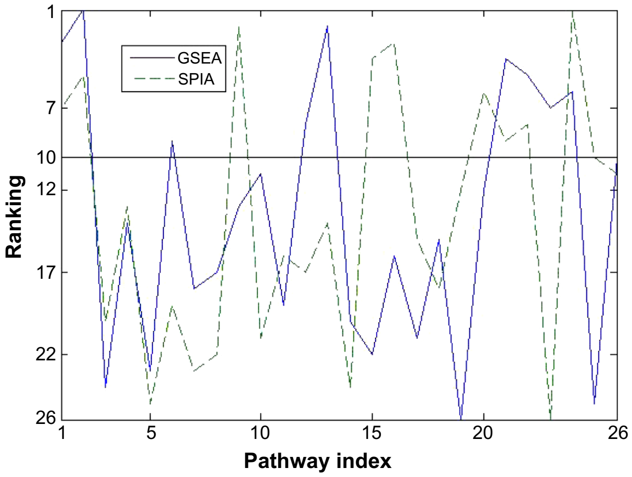

Figure 1. Comparison of pathway rankings from GSEA and SPIA (X-axis labels are the indices of pathways, which are in the first column of Table 1).

First, according to the results from GSEA and SPIA, using TCGA gene expression data reveals that the cell cycle pathway is closely related to LUSC. The cell cycle is the process leading a cell's division and duplication/replication, which is accomplished through a reproducible sequence of events: DNA replication (S phase) and mitosis (M phase), separated temporally by gaps known as G1 and G2 phases.[34] Cell cycle regulation disorder or DNA demand may lead to uncontrollable cell growth and the forming of malignant tumors. In this sense, we can see that cell cycle regulation disorder could be related to the development of cancer. Moreover, the literature reveals that cell cycle pathway has a close relationship with cancer.[10,21] For example, in the work of Shi et al.[21], the cell cycle pathway was ranked first among the altered signaling pathways linked with LUSC, based on the analysis of three different microarray gene expression data sets and using statistical methods that differed from GSEA and SPIA. This supports our result that the top-ranking pathway, namely, the cell cycle pathway, is implicated in LUSC.

Second, the viral carcinogenesis pathway is also ranked at the top by GSEA as well as SPIA, indicating a high-level association with LUSC. According to KEGG,[34] there is a strong association between viruses and the development of human malignancies. Specifically, through the expression of many potent oncoproteins, tumor viruses promote an aberrant cell proliferation via modulating cellular cell-signaling pathways and escape from cellular defense system.[34] Human tumor virus oncoproteins can also disrupt pathways that are necessary for the maintenance of the integrity of the host cellular genome.[34] Viruses that encode such activities can contribute to the initiation, as well as progression, of human cancers.[31,34] In our study, the high ranking of the viral carcinogenesis pathway indicates the pathway has a role in LUSC development.

Third, the p53 signaling pathway, well known for its relationship to cancers,[21,30,31] was ranked by GSEA in the second place. Another important pathway, the PI3K-Akt signaling pathway, which has been found to have an association with lung cancer,[30,31] was also ranked near the top (in fourth place) by SPIA. The established association of these pathways with cancers, along with our findings, suggests a close relationship of these pathways to LUSC.

Fourth, two other pathways, the bladder cancer and thyroid cancer pathways, were also ranked near the top by GSEA. As their names suggest, these pathways are known to be associated with their own cancer types (ie, bladder cancer and thyroid cancer). Our study found that these pathways could also be implicated in lung cancer, particularly in LUSC. This result suggests that different types of cancers could share similar biological mechanisms.

Fifth, our study also highly ranked some pathways no one believed to be implicated in cancer. For example, type I diabetes mellitus and type II diabetes mellitus pathways were ranked number 5 and number 7 by GSEA, respectively. Although these pathways were previously known as related to their own diseases (eg, diabetes), the fact that our results highly ranked pathways already known to be related to LUSC increases our confidence that these highly ranked pathways may also be implicated in the development of LUSC. The discovery of these possible pathway implications suggests future biological experiments and clinical trials for finding new LUSC-related pathways.

Finally, in our study, GSEA and SPIA did not highly rank some pathways previously known to be related to cancer, such as the mTOR signaling pathway and the notch signaling pathway.[21,29–31] These results may be due to particularities of the data set (ie, the TCGA LUSC gene expression data set) and cancer type (ie, LUSC) or to the possibility that signals in the gene expression data may not be strong enough to uncover relationships. Future research might focus on the investigation of different cancer data for the analysis of different cancer types, as well as employing more pathways in gene set analysis to find more underlying relationships of pathways to cancers.

Conclusions

We applied GSEA and SPIA to microarray gene expression data to investigate pathway implications in lung cancer. In particular, we applied them to the analysis of the relationship of LUSC and 26 pathways from the KEGG, using the TCGA LUSC gene expression data. The results demonstrated that some pathways could be related to lung cancer. For example, the cell cycle pathway and the viral carcinogenesis pathway are highly implicated in LUSC; the p53 signaling pathway and the PI3K-Akt signaling pathway also have connections to LUSC; and pathways of other cancer types (eg, bladder and thyroid cancer pathways) also appear linked to the development of lung cancer. Future research could involve the investigation of other cancers, the consideration of a greater number of pathways for analysis, and the utilization of other gene expression data.

Acknowledgments

We thank Diyang Xue for running experiments using SPIA.

Author Contributions

Conceived and designed the experiments: XJ. Analyzed data, conducted experiments and developed related processing programs: BC. Wrote the first draft of the manuscript: BC. Contributed to the analyses of the results and writing of the manuscript: XJ. Agreed with manuscript results and conclusions: BC, XJ. Jointly developed the structure and arguments for the paper: BC, XJ. Made critical revisions and approved final version: XJ. Both authors reviewed and approved the final manuscript.

REFERENCES

1. Hesketh R. *Introduction to Cancer Biology*. Cambridge: Cambridge University Press; 2013.
2. National Human Genome Research Institute. Available from: http://www.genome.gov/27530687. 2012.
3. Vivanco I, Sawyers CL. The phosphatidylinositol 3-kinase-AKT pathway in human cancer. *Nat Rev Cancer*. 2002;2:489–501.
4. Waugh DJJ, Wilson C. The interleukin-8 pathway in cancer. *Clin Cancer Res*. 2008;14:6735–41.
5. Schutte M, Hruban RH, Geradts J, et al. Abrogation of the Rb/p16 tumor-suppressive pathway in virtually all pancreatic carcinomas. *Cancer Res*. 1997;57:3126–30.
6. Schena M, Shalon D, Davis RW, Brown PO. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science*. 1995;270(5235):467–70.
7. Pham TD, Wells C, Crane DI. Analysis of microarray gene expression data. *CBIO*. 2006;1(1):37–53.
8. Cai B, Jiang X. A novel artificial neural network method for biomedical prediction based on matrix pseudo-inversion. *J Biomed Inform*. 2014;48:114–21.
9. Jiang X, Barmada MM, Visweswaran S. Identifying genetic interactions in genome-wide data using Bayesian networks. *Genet Epidemiol*. 2010;34(6):575–81.
10. Burton M, Thomassen M, Tan Q, Kruse TA. Prediction of breast cancer metastasis by gene expression profiles: a comparison of metagenes and single genes. *Cancer Inform*. 2012;11:193–217.
11. Shi J, Walker MG. Gene set enrichment analysis (GSEA) for interpreting gene expression profiles. *CBIO*. 2007;2(2):133–7.
12. Subramanian A, Tamayo P, Mootha VK, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci USA*. 2005;102(43):15545–50.
13. Mootha VK, Lindgren CM, Eriksson KF, et al. PGC-1α-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nat Genet*. 2003;34(3):267–73.
14. Hua J, Bittner ML, Dougherty ER. Evaluating gene set enrichment analysis via a hybrid data model. *Cancer Inform*. 2014;13(S1):1–16.
15. Tarca AL, Draghici S, Khatri P, et al. A novel signaling pathway impact analysis. *Bioinformatics*. 2009;25(1):75–82.
16. Chen F, Guan Q, Nie ZY, Jin LJ. Gene expression profile and functional analysis of Alzheimer's disease. *Am J Alzheimers Dis Other Demen*. 2013;28(7):693–701.
17. Judeh T, Johnson C, Kumar A, Zhu D. TEAK: topology enrichment analysis framework for detecting activated biological subpathways. *Nucleic Acids Res*. 2013;41(3):1425–37.
18. National Human Genome Research Institute. Available from: http://cancergenome.nih.gov/. 2014.
19. Benz SC. *Sample-Specific Cancer Pathway Analysis Using PARADIGM*. PhD [dissertation]. Santa Cruz: University of California; 2012.
20. Martini P, Sales G, Massa MS, Chiogna M, Romualdi C. Along signal paths: an empirical gene set approach exploiting pathway topology. *Nucleic Acids Res*. 2013;41(1):e19.

21. Shi I, Sadraei NH, Duan ZH, Shi T. Aberrant signaling pathways in squamous cell lung carcinoma. *Cancer Inform*. 2011;10:273–85.
22. Hung JH, Yang TH, Hu Z, Weng Z, DeLisi C. Gene set enrichment analysis: performance evaluation and usage guidelines. *Brief Bioinform*. 2011; 13(3):281–91.
23. Moriya Y, Itoh M, Okuda S, Yoshizawa AC, Kanehisa M. KAAS: an automatic genome annotation and pathway reconstruction server. *Nucleic Acids Res*. 2007;35:W182–5.
24. Gilchrist A, Au CE, Hiding J, et al. Quantitative proteomics analysis of the secretory pathway. *Cell*. 2006;127(6):1265–81.
25. Oron AP, Jiang Z, Gentleman R. Gene set enrichment analysis using linear models and diagnostics. *Bioinformatics*. 2008;24(22):2586–91.
26. Qian B, Zhang H, Zhang L, Zhou X, Yu H, Chen K. Association of genetic polymorphisms in DNA repair pathway genes with non-small cell lung cancer risk. *Lung Cancer*. 2011;73:138–46.
27. Gustafson AM, Soldi R, Anderlind C, et al. Airway PI3K pathway activation is an early and reversible event in lung cancer development. *Sci Transl Med*. 2010;2(26):26ra25.
28. Toonke RL, Borczuk AC, Powell CA. TGF-ß signaling pathway in lung adenocarcinoma invasion. *J Thoracic Oncol*. 2010;5(2):153–7.
29. Ekman S, Wynes MW, Hirsch FR. The mTOR pathway in lung cancer and implications for therapy and biomarker analysis. *J Thoracic Oncol*. 2012;7(6):947–53.
30. The Cancer Genome Atlas Research Network. Comprehensive genomic characterization of squamous cell lung cancers. *Nature*. 2012;489:519–25.
31. Cooper WA, Lam DCL, O'Toole SA, Minna JD. Molecular biology of lung cancer. *J Thorac Dis*. 2013;5(S5):S479–90.
32. Neapolitan R, Xue D, Jiang X. Modeling the altered expression levels of genes on signaling pathways in tumors as causal Bayesian networks. *Cancer Inform*. 2013;13:77–84.
33. Neapolitan R, Jiang X. Inferring aberrant signal transduction pathways in ovarian cancer from TCGA data. *Cancer Informatics*. 2014;13(s1):29–36.
34. Kanehisa Laboratories. Available from: http://www.genome.jp/kegg/. 2014.
35. Liu P, Morrison C, Wang L, et al. Identification of somatic mutations in non-small cell lung carcinomas using whole-exome sequencing. *Carcinogenesis*. 2012;33(7):1270–6.
36. Győrffy B, Surowiak P, Budczies J, Lánczky A. Online survival analysis software to assess the prognostic value of biomarkers using transcriptomic data in non-small-cell lung cancer. *PLoS One*. 2013;8(12):e82241.
37. Deng B, Sun Z, Jason W, Yang P. Increased BCAR1 predicts poor outcomes of non-small cell lung cancer in multiple-center patients. *Ann Surg Oncol*. 2013;20:S701–8.
38. Barrett CL, Schwab RB, Jung H, et al. Transcriptome sequencing of tumor subpopulations reveals a spectrum of therapeutic options for squamous cell lung cancer. *PLoS One*. 2013;8(3):e58714.
39. Jiang L, Zhu W, Streicher K, et al. Increased IR-A/IR-B ratio in non-small cell lung cancers associates with lower epithelial-mesenchymal transition signature and longer survival in squamous cell lung carcinoma. *BMC Cancer*. 2014;14:131.
40. Ylipää A, Yli-Harja O, Zhang W, Nykter M. Characterization of aberrant pathways across human cancers. *BMC Syst Biol*. 2013;7(S1):S1.
41. Smyth GK. Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. *Stat Appl Genet Mol Biol*. 2004;3(1):Article3.