# The Prognostic Role of Genes with Skewed Expression Distribution in Lung Adenocarcinoma

Yajing Chen[1], Shikui Tu[1], and Lei Xu[1,2(✉)]

[1] Center for Cognitive Machines and Computational Health, and Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai, China
{cyj907,tushikui,leixu}@sjtu.edu.cn
[2] Department of Computer Science and Engineering, The Chinese University of Hong Kong, Hong Kong SAR, China

**Abstract.** Many studies assumed gene expression to be normally distributed. However, some were found to have left-skewed distribution, while others have right-skewed distribution. Here, we investigated the gene expression distribution of five lung adenocarcinoma data sets. We assumed that samples in the tail and non-tail of a skewed distribution were drawn from different populations with different survival outcomes. To investigate this hypothesis, skewed genes were detected to build a tail indicator matrix comprising of binary values. Survival analysis revealed that patients with more skewed genes in their tails had worse survival. Hierarchical clustering of the tail indicator matrices discovered a gene set with similar tail configurations for either left or right skewed genes. The two gene sets divided patients into three groups with different survivals. In conclusion, there is a direct association between genes with skewed distribution and the prognosis of lung adenocarcinoma patients.

**Keywords:** Skewed distribution · Gene expression · RNA-sequencing · Microarray · Survival · Lung adenocarcinoma

## 1 Introduction

Gene expression profiling measures the activity of a large number of genes at a time. DNA microarray technology and RNA-sequencing are two main approaches to obtain gene expression data, though the latter is taking place of the former nowadays. RNA-sequencing uses next-generation sequencing to measure the RNA quantity in a sample, while microarray is based on hybridization of the predesigned probes and RNAs. Despite the difference in the techniques, the obtained expression values are highly correlated, which implies that the analysis approaches and the results from one might be applicable to the other [4,16,20]. Many gene expression profiling studies assume the distribution of gene expression to be Gaussian. In this case, statistical methods like t-test can be applied to detect genes with the differential expression between patient groups. However,


Y. Sun et al. (Eds.): IScIDE 2017, LNCS 10559, pp. 631–640, 2017.
DOI: 10.1007/978-3-319-67777-4_57

Thomas et al. showed that gene expression is not always normally distributed [15]. Some genes display heavy-tailed distributions [6]. The reason why these genes have non-Gaussian distribution remains unknown.

Lung cancer is the most frequent cancer in the world, covering 13% of the total cancer incidence. It can be further divided into small cell lung carcinoma (SC) and non-small cell lung carcinoma (NSCLC). Lung adenocarcinoma is a histological type of NSCLC. 40% of the lung cancers in the US are adenocarcinomas. Patients with lung cancer have a 5-year survival of 10–15% [13]. The poor prognosis of this disease urges the discovery of new reliable and effective therapeutic approaches.

Here, we focused on the genes with skewed distribution in lung adenocarcinoma, and hypothesized that the tail and non-tail of a skewed distribution indicated distinct populations that might form subtypes of the disease. We investigated the survival of patients in tail and non-tail of skewed distribution, and see if different prognostic groups were formed. The data sets used in this study included a RNA-sequencing and four microarray data sets. We computed skewness to detect genes with heavy-tailed distributions, and obtained tail indicator matrices by labelling the tails. Survival analysis showed that patients with more tails in the skewed genes had poorer overall survival, regardless of the tail direction. Hierarchical clustering of the tail indicator matrices helped discover and select genes with similar tail configurations for either left or right skewed genes. They classified patients into three groups, i.e., one with both left and right tails, one with either left or right tail, and the other with no tail. Kaplan-Meier plots showed that patients with both left and right tails had worst survival, while those with no tail had the best prognosis. Literature review on the genes in *L-list* and *R-list* demonstrated their potential roles as therapeutic targets.

In conclusion, genes of skewed distribution are correlated to the prognosis of lung adenocarcinoma patients. A high number of tails in the skewed genes predicts a lower survival rate. Patients can be divided into three prognostic groups according to the tail configuration in *L-list* and *R-list*. The genes in the *L-list* and *R-list* can provide reliable therapeutic targets for the disease.

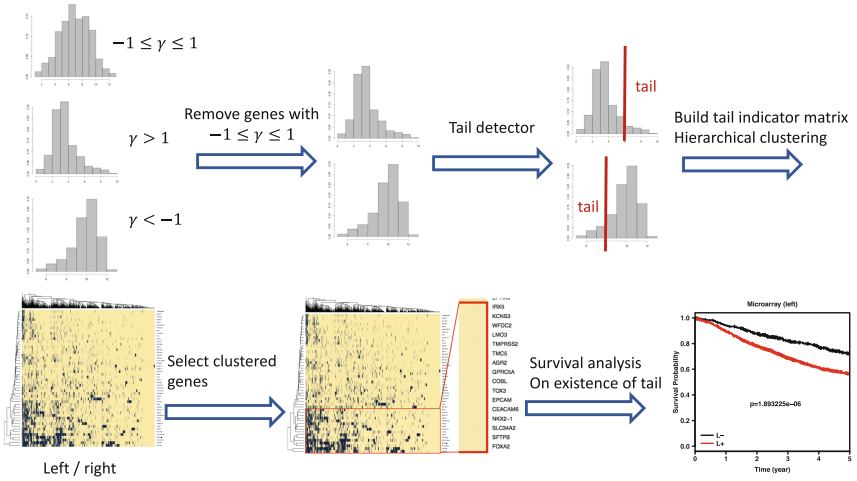
## 2 Methods

### 2.1 Analysis Procedures

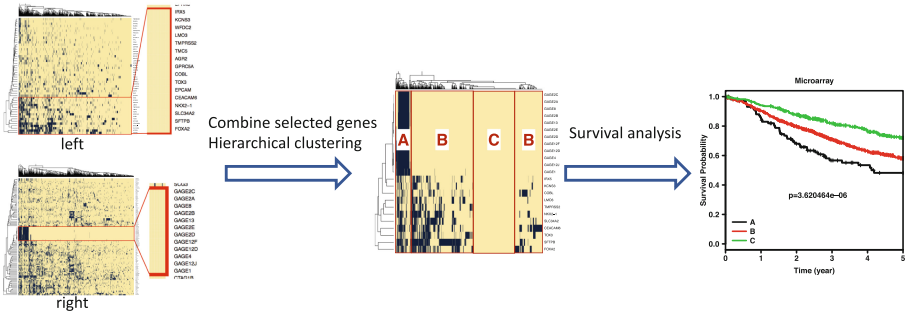
Shown in Fig. 1 are the two main analysis procedures applied in this study. Details of these procedures are explained in the following sections.

### 2.2 Tail Detector

After obtaining a list of left-tail and right-tail genes, their tails were labelled to compute tail indicator matrices (see Fig. 3(b) as an example). The tail indicator matrix comprises of binary values, with 0 as non-tail and 1 as tail. To label the tails in a right-tail gene, samples were removed one by one according to the




(a) Procedure I




(b) Procedure II

**Fig. 1.** Illustration of analysis procedures.

decreasing order of expression levels until the skewness of the remaining samples was smaller than or equal to zero. The samples removed were labelled as tail, while the remaining were labelled as non-tail. A similar procedure was applied to determine the tails for left-tail genes, though the samples were removed according to an increasing order of expression instead.

## 3 Experiments and Results

### 3.1 Data Preprocessing

The data sets we used in this study included a RNA-sequencing and four microarray data sets. The RNA-sequencing data set was generated by the Cancer Genome Atlas (TCGA) (https://cancergenome.nih.gov/) and downloaded from the UCSC Xena browser (https://xenabrowser.net/datapages/, [3]). The

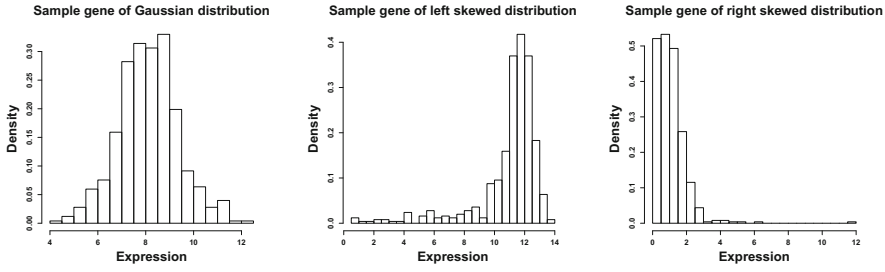
four microarray data sets were obtained from National Center for Biotechnology Information (NCBI) (https://www.ncbi.nlm.nih.gov/, [10]). The probes of the same gene in microarray data sets were first merged by computing the mean expression. Only the genes shared by all the five data sets ($n = 12164$) were kept for further analysis. The samples without survival data were also removed. The details of the data are listed in Table 1.

**Table 1.** Information of the five data sets.

| Data set | Type | #Sample | Ref |
|---|---|---|---|
| TCGA | RNA-seq | 503 | [8] |
| GSE31210 | Microarray | 226 | [9] |
| GSE50081 | Microarray | 130 | [1] |
| GSE68465 | Microarray | 442 | [12] |
| GSE72094 | Microarray | 398 | [11] |

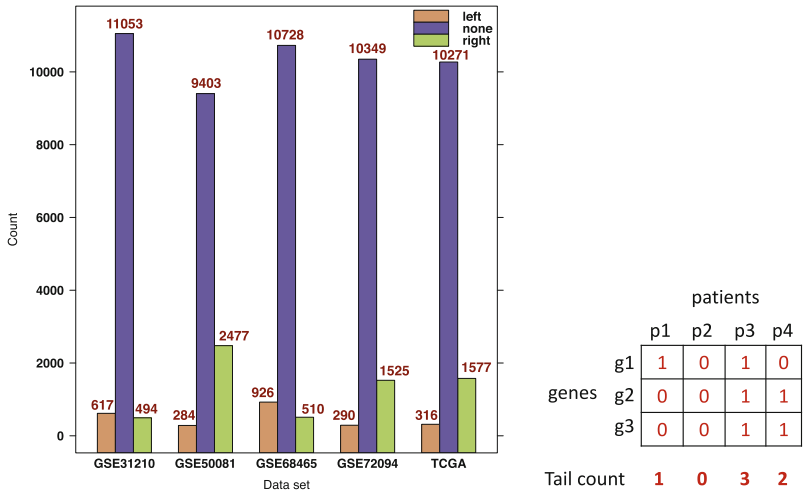
### 3.2 Tail Counts Have a Negative Correlation with Survival

To investigate the relation between skewed distribution and patient survival, we first selected the genes whose expression have heavy left or right tails. Skewness $\gamma$ for each gene was computed by the Pearson's moment coefficient of skewness in all the five data sets. Those with $\gamma > 1$ or $\gamma < -1$ were detected as skewed genes. Figure 2 shows three example genes whose expression distributions are normal, left-skewed (left-tail) and right-skewed (right-tail). The number of *left-tail* and *right-tail* genes in the five data sets are shown in Fig. 3(a). Though RNA-seq expression profiles share similarities with microarray expression profiles, it would be more reasonable to analyze the data generated by these two techniques separately. First, we integrated the resulting tail genes of the four microarray data sets. Figure 3(a) shows that the number of right-tail genes is much larger than the left-tail genes in the microarray data sets in general. Therefore, to




(a) Normal (b) Left-skewed (c) Right-skewed

**Fig. 2.** Histogram of three example gene distributions.




(a) Gene count (b) Tail count

**Fig. 3.** Barplots for skewed genes in the five data sets and the illustration for counting tails.

obtain two balanced sets of reliable left- and right-tail genes, we selected the right-tail genes ($n = 129$) that were detected in all four microarray data sets, and the left-tail genes ($n = 209$) that were detected in more than one data set.

The intersection of the genes generated by the RNA-seq data and microarray data were computed, resulting in 50 left-tail and 100 right-tail genes. Subsequently, the tail labelling approach was used to compute tail indicator matrices for the shared left-tail ($n = 50$) and right-tail ($n = 100$) genes in the five data sets. The four matrices computed from the microarray data were merged. The total number of tails in the selected left or right tail genes was counted for each patient (see Fig. 3(b) for details). The results of Cox regression showed that a higher number of tail counts leads to a lower survival rate in both RNA-seq and microarray data (shown in Table 2).

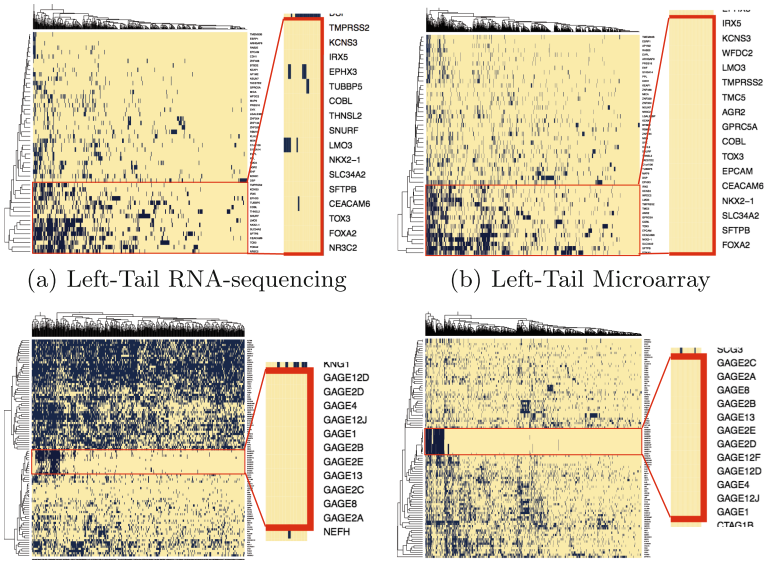
**Table 2.** Results from Cox regression for survival versus tail counts. **Left** means left-tailed; **Right** means right-tailed. **p** means p-value obtained from likelihood-ratio test. **HR** means hazard ratio.

| Data set | Left (p) | Left (HR) | Right (p) | Right (HR) |
|---|---|---|---|---|
| RNA-seq | 0.000178 | 1.0595 | 0.347 | 1.00645 |
| Microarray | 1.23e-05 | 1.04205 | 0.000107 | 1.01822 |

### 3.3 Hierarchical Clustering Reveals Similar Tail Configurations in Genes and Patients

Hierarchical clustering was performed to analyze the tail indicator matrices for RNA-seq and microarray respectively. As shown in Fig. 4, their clustering results were similar. The red frames indicate the genes with similar tail configurations. For left-tail genes, the intersection of the clustered genes in the red frames for RNA-seq and microarray tail indicator matrices form a list of 16 genes, denoted as *L-list*. For right-tail genes, genes in the GAGE family were grouped together in both matrices, denoted as *R-list*. Subsequently, we investigated the survival of patients with different tail configurations in the *L-list* and *R-list* (shown in Fig. 5). Patients with no tails (black) had a longer survival than those with tails (red). The difference is significant for both *L-list* and *R-list* microarray tail indicator matrices, and is also significant for the *L-list* RNA-seq tail indicator matrix. Though the separation of the K-M curves for *R-list* RNA-seq matrix is not significant, the gap between the two curves is large. To conclude, the patients with tails in the *L-list* or *R-list* had worse prognosis than those with no tails.

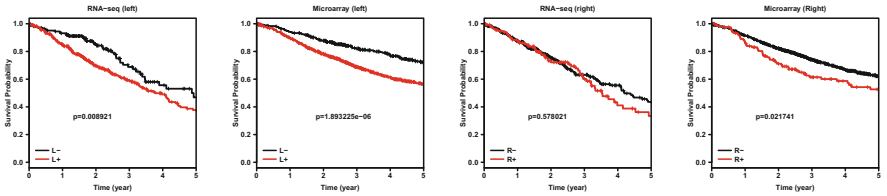
*L-list* and *R-list* were combined together for analysis to determine whether similarities exist between both sets of data. Figure 6(a), (b) display the hierarchical clustering results for their RNA-seq and microarray tail indicator matrices. Patients were clustered into three distinct groups, where one showed a large number of tails in both *L-list* and *R-list* (A), one showed a high number of tails




(a) Left-Tail RNA-sequencing (b) Left-Tail Microarray

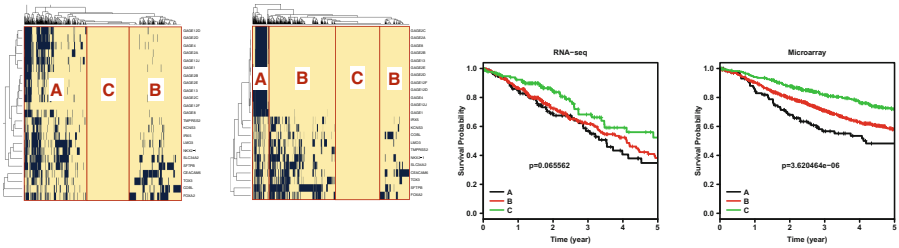
(c) Right-Tail RNA-sequencing (d) Right-Tail Microarray

**Fig. 4.** Hierarchical clustering of left-tail and right-tail indicator matrices, where yellow represents non-tail and blue represents left-tail or right-tail. The red frames were added manually to indicate the clustered genes. (Color figure online)




(a) (b) (c) (d)

**Fig. 5.** Kaplan-Meier plots for different tail configurations. *L-list* in RNA-seq, $p = 0.0089$ (a). *L-list* in microarray, $p = 1.89e-06$ (b). *R-list* in RNA-seq $p = 0,57$ (c). *R-list* in microarray, $p = 0.022$ (d). $L+$ represents patients with at least one tail in *L-list*; $L-$ represents patients with no tail in *L-list*. $R+$ represents patients with at least one tail in *R-list*; $R-$ represents patients with no tail in *R-list*. The p-values in the plots were computed by log-rank test. (Color figure online)




(a) RNA-seq (b) Microarray (c) RNA-seq (d) Microarray

**Fig. 6.** Hierarchical clustering for *L-list* and *R-list* (a,b). Kaplan Meier plots for different tail configuration of genes in *L-list* and *R-list* (c,d). **A** represents patient group with tails in *L-list* and *R-list*. **B** represents patient groups with tails in either *L-list* or *R-list*. **C** represents patient groups with no tails in the two list. The p-values in the plots ($p = 0.066$ for RNA-seq, $p = 3.62e-06$ for microarray) were computed by log-rank test. (Color figure online)

in merely *L-list* (B), and the other showed almost no tails (C). Based on the clustering results, we divided the patients into three groups by their tail configurations, i.e., one with tails in *L-list* and *R-list* (A), one with tails in either list (B), and the other with no tails (C). Figure 6(c), (d) show the survival curves for the three groups. Group A (black) has the worst survival, group B (red) has intermediate survival, and group C (green) has the best prognosis. Though only the microarray data show significant difference with $p < 0.05$, there is clear separation of the three survival curves for both RNA-seq and microarray data. These results indicated that *L-list* and *R-list* can serve as prognostic markers for lung adenocarcinoma.

### 3.4 Functional Interpretation of Genes in *L-list* and *R-list*

Literature review was conducted on the genes in the *L-list* and *R-list*. Some genes were found to be associated with lung adenocarcinoma or non-small cell lung cancer (NSCLC). First, the genes in the *L-list* were investigated, and results are listed below.

- *FOXA2* is a transcription factor that is involved in lung development. The loss of FOXA2, CDX2 and NKX2-1 can activate the metastatic process of lung adenocarcinoma [5].
- The non-detectable status of *SFTPB* is related to high risk of lung cancer [14].
- *SLC34A2* is an important gene during the fetal lung development, which was suppressed in the lung adenocarcinoma cell line A549. Its up-regulation inhibits cell invasion, tumor growth and metastasis ability [17].
- *NKX2-1* inhibits tumor differentiation and metastatic potential in vivo. The loss of this gene enhances tumor seeding and metastatic proclivity [19].
- *LMO3* is activated by the amplification of NKX2-1. It is an important downstream effector from NKX2-1 in enhancing proliferation and survival of NKX2-1-amplified lung adenocarcinoma cell lines [18].

The genes in the *R-list* are all from the GAGE family. They are cancer/testis antigens (CTA) which are expressed in some tumors and not expressed in normal lung tissue except for the testis. The frequency of the CTA expression is higher in patients of higher stages in NSCLC, indicating its role as a poor prognostic marker [2]. Therefore, patients with more tails in *L-list* and *R-list* had a higher probability of tumor metastasis and invasion of lung carcinoma, leading to poor prognosis.

## 4 Conclusion

Genes with skewed distribution in expression data were investigated and a correlation with survival was discovered. Patients with more tails in the left or right-tail genes had worse survival. Furthermore, RNA-seq and microarray data shared many siimlar tail configurations in some left and right tail genes, which were denoted as *L-list* and *R-list*. These genes helped to classify the patients into three prognostic groups, indicating three possible subtypes of the disease. Kaplan-Meier plots displayed that the patient group with tails in both *L-list* and *R-list* suffer from poorer prognosis than those with tail in either list. Those with no tail had the best survival among the three groups. Literature survey revealed that some genes in the *L-list* and *R-list* were reported to be related to lung adenocarcinoma. Patients with more tails in the *L-list* and *R-list* were more likely to have tumor metastasis, which explained why more tail counts predicted worse survival. The genes in *L-list* and *R-list* might provide potential therapeutic targets for lung adenocarcinoma. We think that the analysis procedures illustrated here can also serve as a biomarker detector and survival predictor for other diseases.

**Acknowledgments.** This work was supported by the Zhi-Yuan chair professorship start-up grant (WF220103010) from Shanghai Jiao Tong University.

## References

1. Der, S.D., Sykes, J., Pintilie, M., Zhu, C.Q., Strumpf, D., Liu, N., Jurisica, I., Shepherd, F.A., Tsao, M.S.: Validation of a histology-independent prognostic gene signature for early-stage, non-small-cell lung cancer including stage IA patients. J. Thorac. Oncol. **9**(1), 59–64 (2014)
2. Gjerstorff, M.F., Pøhl, M., Olsen, K.E., Ditzel, H.J.: Analysis of GAGE, NY-ESO-1 and SP17 cancer/testis antigen expression in early stage non-small cell lung carcinoma. BMC Cancer **13**(1), 466 (2013)
3. Goldman, M., Craft, B., Swatloski, T., Cline, M., Morozova, O., Diekhans, M., Haussler, D., Zhu, J.: The UCSC cancer genomics browser: update 2015. Nucleic Acids Res. **43**, D812–D817 (2014)
4. Guo, Y., Sheng, Q., Li, J., Ye, F., Samuels, D.C., Shyr, Y.: Large scale comparison of gene expression levels by microarrays and RNAseq using TCGA data. PLoS one **8**(8), e71462 (2013)
5. Li, C.M.C., Gocheva, V., Oudin, M.J., Bhutkar, A., Wang, S.Y., Date, S.R., Ng, S.R., Whittaker, C.A., Bronson, R.T., Snyder, E.L., et al.: Foxa2 and Cdx2 cooperate with NKX2-1 to inhibit lung adenocarcinoma metastasis. Genes devel. **29**(17), 1850–1862 (2015)
6. Marko, N.F., Weil, R.J.: Non-gaussian distributions affect identification of expression patterns, functional annotation, and prospective classification in human cancer genomes. PLoS one **7**(10), e46935 (2012)
7. Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., Leisch, F., Chang, C.C., Lin, C.C., Meyer, M.D.: Package e1071 (2017)
8. Network, C.G.A.R., et al.: Comprehensive molecular profiling of lung adenocarcinoma. Nature **511**(7511), 543–550 (2014)
9. Okayama, H., Kohno, T., Ishii, Y., Shimada, Y., Shiraishi, K., Iwakawa, R., Furuta, K., Tsuta, K., Shibata, T., Yamamoto, S., et al.: Identification of genes upregulated in ALK-positive and EGFR/KRAS/ALK-negative lung adenocarcinomas. Cancer Res. **72**(1), 100–111 (2012)
10. Sayers, E.W., Barrett, T., Benson, D.A., Bolton, E., Bryant, S.H., Canese, K., Chetvernin, V., Church, D.M., DiCuccio, M., Federhen, S., et al.: Database resources of the national center for biotechnology information. Nucleic Acids Res. **39**(suppl 1), D38–D51 (2011)
11. Schabath, M.B., Welsh, E.A., Fulp, W.J., Chen, L., Teer, J.K., Thompson, Z.J., Engel, B.E., Xie, M., Berglund, A.E., Creelan, B.C., et al.: Differential association of STK11 and TP53 with KRAS mutation-associated gene expression, proliferation and immune surveillance in lung adenocarcinoma. Oncogene **35**, 3209 (2015)
12. Shedden, K., Taylor, J.M., Enkemann, S.A., Tsao, M.S., Yeatman, T.J., Gerald, W.L., Eschrich, S., Jurisica, I., Giordano, T.J., Misek, D.E., et al.: Gene expression-based survival prediction in lung adenocarcinoma: a multi-site, blinded validation study. Nat. Med. **14**(8), 822–827 (2008)
13. Stewart, B., Wild, C.P., et al.: World cancer report 2014 (2014)
14. Taguchi, A., Hanash, S., Rundle, A., McKeague, I.W., Tang, D., Darakjy, S., Gaziano, J.M., Sesso, H.D., Perera, F.: Circulating pro-surfactant protein B as a risk biomarker for lung cancer. Cancer Epidemiol. Prev. Biomark. **22**(10), 1756–1761 (2013)

15. Thomas, R., de la Torre, L., Chang, X., Mehrotra, S.: Validation and characterization of DNA microarray gene expression data distribution and associated moments. BMC Bioinform. **11**(1), 576 (2010)
16. Trost, B., Moir, C.A., Gillespie, Z.E., Kusalik, A., Mitchell, J.A., Eskiw, C.H.: Concordance between RNA-sequencing data and DNA microarray data in transcriptome analysis of proliferative and quiescent fibroblasts. Roy. Soc. Open Sci. **2**(9), 150402 (2015)
17. Wang, Y., Yang, W., Pu, Q., Yang, Y., Ye, S., Ma, Q., Ren, J., Cao, Z., Zhong, G., Zhang, X., et al.: The effects and mechanisms of SLC34A2 in tumorigenesis and progression of human non-small cell lung cancer. J. Biomed. Sci. **22**(1), 52 (2015)
18. Watanabe, H., Francis, J.M., Woo, M.S., Etemad, B., Lin, W., Fries, D.F., Peng, S., Snyder, E.L., Tata, P.R., Izzo, F., et al.: Integrated cistromic and expression analysis of amplified NKX2-1 in lung adenocarcinoma identifies LMO3 as a functional transcriptional target. Genes Dev. **27**(2), 197–210 (2013)
19. Winslow, M.M., Dayton, T.L., Verhaak, R.G., Kim-Kiselak, C., Snyder, E.L., Feldser, D.M., Hubbard, D.D., DuPage, M.J., Whittaker, C.A., Hoersch, S., et al.: Suppression of lung adenocarcinoma progression by NKX2-1. Nature **473**(7345), 101–104 (2011)
20. Zhao, S., Fung-Leung, W.P., Bittner, A., Ngo, K., Liu, X.: Comparison of RNA-Seq and microarray in transcriptome profiling of activated T cells. PloS one **9**(1), e78644 (2014)