Published online 9 November 2009 *Nucleic Acids Research, 2010, Vol. 38, Database issue* ***D665–D669***
doi:10.1093/nar/gkp945

HLungDB: an integrated database of human lung cancer research

Lishan Wang[1], Yuanyuan Xiong[1], Yihua Sun[2], Zhaoyuan Fang[2], Li Li[2], Hongbin Ji[2,*] and Tieliu Shi[1,3,*]

[1]Center for Bioinformatics and Computational Biology, and The Institute of Biomedical Sciences, College of Life Science, East China Normal University, Shanghai 200241, [2]Laboratory of Molecular Cell Biology, Institute of Biochemistry and Cell Biology and [3]Shanghai Information Center for Life Sciences, Shanghai Institutes for Biological Sciences, Chinese Academy of Science, Shanghai 200031, China

Received August 15, 2009; Revised October 10, 2009; Accepted October 12, 2009

ABSTRACT

The human lung cancer database (HLungDB) is a database with the integration of the lung cancer-related genes, proteins and miRNAs together with the corresponding clinical information. The main purpose of this platform is to establish a network of lung cancer-related molecules and to facilitate the mechanistic study of lung carcinogenesis. The entries describing the relationships between molecules and human lung cancer in the current release were extracted manually from literatures. Currently, we have collected 2585 genes and 212 miRNA with the experimental evidences involved in the different stages of lung carcinogenesis through text mining. Furthermore, we have incorporated the results from analysis of transcription factor-binding motifs, the promoters and the SNP sites for each gene. Since epigenetic alterations also play an important role in lung carcinogenesis, genes with epigenetic regulation were also included. We hope HLungDB will enrich our knowledge about lung cancer biology and eventually lead to the development of novel therapeutic strategies. HLungDB can be freely accessed at http://www.megabionet.org/bio/hlung.

INTRODUCTION

Lung cancer, one of the most common causes of cancer-related death in both men and women, is responsible for 1.3 million deaths worldwide every year. Lung cancer can be roughly divided into two groups according to pathology: non-small cell lung cancer (NSCLC) (80.4%) and small cell lung cancer (16.8%) (1). Many factors potentially contribute to lung cancer formation, e.g. tobacco smoke, ionizing radiation and viral infection. However, the mechanisms involved in lung carcinogenesis remain largely unknown.

Similar to many other cancers, lung cancer is initiated by activation of oncogenes or inactivation of tumor suppressor genes (2). Previous studies have revealed the various causes of lung cancer at the genomic level. Mutations in the *K-ras* proto-oncogene are responsible for 10–30% of lung adenocarcinomas (3,4). The epidermal growth factor receptor (EGFR) regulates cell proliferation, apoptosis, angiogenesis and tumor invasion (3). Oncogenic mutations and amplification of EGFR are common in non-small cell lung cancer and thus provide the basis for treatment with EGFR inhibitors. In contrast, Her2/neu oncogenic mutation is less frequently observed (3). Other oncogenes involved include c-MET, NKX2-1, PIK3CA and BRAF (3). Inactivation of tumor suppressor genes plays important role in lung carcinogenesis. The *p53* tumor suppressor gene, located on chromosome 17p, is affected in 60–75% of lung cancer including both NSCLC and SCLC while Rb is more likely inactivated in SCLC (5). P16 is also frequently inactivated through the methylation of its promoter region at genomic DNA level. Another important tumor suppressor gene is LKB1, whose loss-of-function mutation/deletion is observed in ~30% lung adenocarcinomas and 20% of squamous cell carcinomas (6,7). Genetic polymorphisms are also indicated to be involved in lung carcinogenesis, e.g. interleukin-1 (8), cytochrome P450 (9), apoptosis promoters such as caspase-8 (10) and DNA repair molecules such as XRCC1 (11). People with these polymorphisms are susceptible to lung cancer

*To whom correspondence should be addressed. Tieliu Shi. Tel: +86 21 54345020; Fax: +86 21 54344016; Email:tlshi@sibs.ac.cn
Correspondence may also be addressed to Hongbin Ji. Email: hbji@sibs.ac.cn

The authors wish it to be known that, in their opinion, the first three authors should be regarded as joint First Authors.

development after exposure to carcinogens. Studies also suggest that the MDM2 309G allele is a low-penetrant risk factor for lung cancer development in Asian population (12).

Although lung cancer research data have accumulated dramatically during the past several years, to our knowledge, there is no database specifically focusing on lung cancer molecular biology yet available. OMIM contains information on all known Mendelian disorders and focuses on the relationship between phenotype and genotype (13). MethyCancer is developed to study the interplay of DNA methylation, gene expression and cancer. It contains both highly integrated data of DNA methylation, cancer-related genes, mutation and cancer information from public resources, and the CpG Island (CGI) clones derived from the large-scale sequencing projects (14). MiR2Disease aims at providing a comprehensive resource of microRNA misregulation in various human diseases (15). EGFR Mutation Database has a convenient compilation of somatic EGFR mutations in NSCLC and associated epidemiological and methodological data, including response to the tyrosine kinase inhibitors Gefitinib and Erlotinib (16). These databases focus on cancer pathogenesis from different angles with a little touch of lung cancer. Thus, it is beneficial to establish a lung cancer-related database or platform involving genes/proteins/miRNAs.

High-throughput techniques applied in the lung cancer research have generated a mass of data and provided important resources for us to potentially explore the molecular mechanisms and identify lung cancer-related molecules. The integration of information generated by small-scale studies and using high-throughput technology could provide a unique resource to facilitate the systematic study of the lung carcinogenesis process. To this end, we collected lung cancer-related molecules and other detailed information for database construction through text mining in combination with bioinformatics analysis. This repository and maintenance system specially designed for lung cancer information can no doubt facilitate future lung cancer investigations.

Overall, HLungDB enables the exploration of relevant information for human lung cancer-related molecules from multiple angles, making it a unique resource for human lung cancer and will serve as a useful platform for those interested in lung cancer biology.

DATA COLLECTION AND CONTENT

As aforementioned, initial entries describing the relationship between genes and human lung cancer are collected manually. The gene–lung cancer relationship documented in the current release were collected through searching the PubMed database with a list of keywords, such as 'lung cancer gene', 'pulmonary cancer gene', 'pulmonary adenocarcinoma gene', etc. After we obtained the literature with the keywords above, we read through and interpreted each paper by collecting the important information, including the type of gene alteration, the clinical correlation and/or significance of the gene alternation with lung cancer, the lung cancer subtype, the potential mechanism of gene regulation and the experimental methods involved.

Each entry in the database contains detailed information on a lung cancer–gene relationship, including a basic description of the gene, the expression pattern of gene (up- or down-regulated) in the lung cancer patient, the experimentally validated regulatory information (transcription factors, their binding motif and the promoter) and protein–protein interaction (PPI) network etc.

Gene expression profiling data for lung cancer patient samples were also retrieved from GEO. The differentially expressed genes were selected if the change between lung cancer samples and normal control is larger than 2-fold. To make the results more reliable, we only selected those genes differentially expressed from at least three patients in a dataset and displayed them on our web site.

In the current release of HLungDB, 2585 genes were selected for their relationships with lung carcinogenesis. A total of 271 lung cancer samples from six expression profiling datasets were analyzed to get the gene expression pattern (17–22). For the lung cancer-related SNPs, we searched PubMed with key words, namely 'SNP' and 'lung cancer'. Then, we collected the SNPs proven to be correlated with lung cancer from those returned papers. In total, 424 SNPs, no matter whether they could be mapped to a gene or not, were added into the database. Additionally, 360 transcription factors with 1160 binding motifs and 253 lung cancer-related genes with detailed epigenetic information were also placed into the database.

Accumulating evidence has indicated that miRNAs play an important role in lung cancer pathology. Previous experiments, both with high-throughput and small-scale methods, have identified many miRNAs differentially expressed in lung cancer and/or confirmed to be related to lung cancer. Hence, miRNA data are an important resource for lung cancer research. Therefore, we selected lung cancer-related miRNAs with experimental information from the literature. For those miRNAs with identified targets, the targets along with the experiment methods used are also provided in the platform. Currently, there are 212 lung cancer-related miRNAs included in the HLungDB.

Next, we built the HLungDB database by integrating the data we collected with information from other resources (Figure 1), which makes our database a one-stop and knowledgeable platform for the lung cancer research community.

DATA ACCESS

HLungDB provides a search engine to query detailed information on each gene–lung cancer relationship documented in the database. Query keywords, including gene/protein symbol or its synonym, are all allowed. The information flow is roughly described in Figure 2.

After submission of the symbol or the alias of a gene, gene centered information will be displayed in a new page, including symbol, alias, description, protein–protein interactions, expression alterations based on the

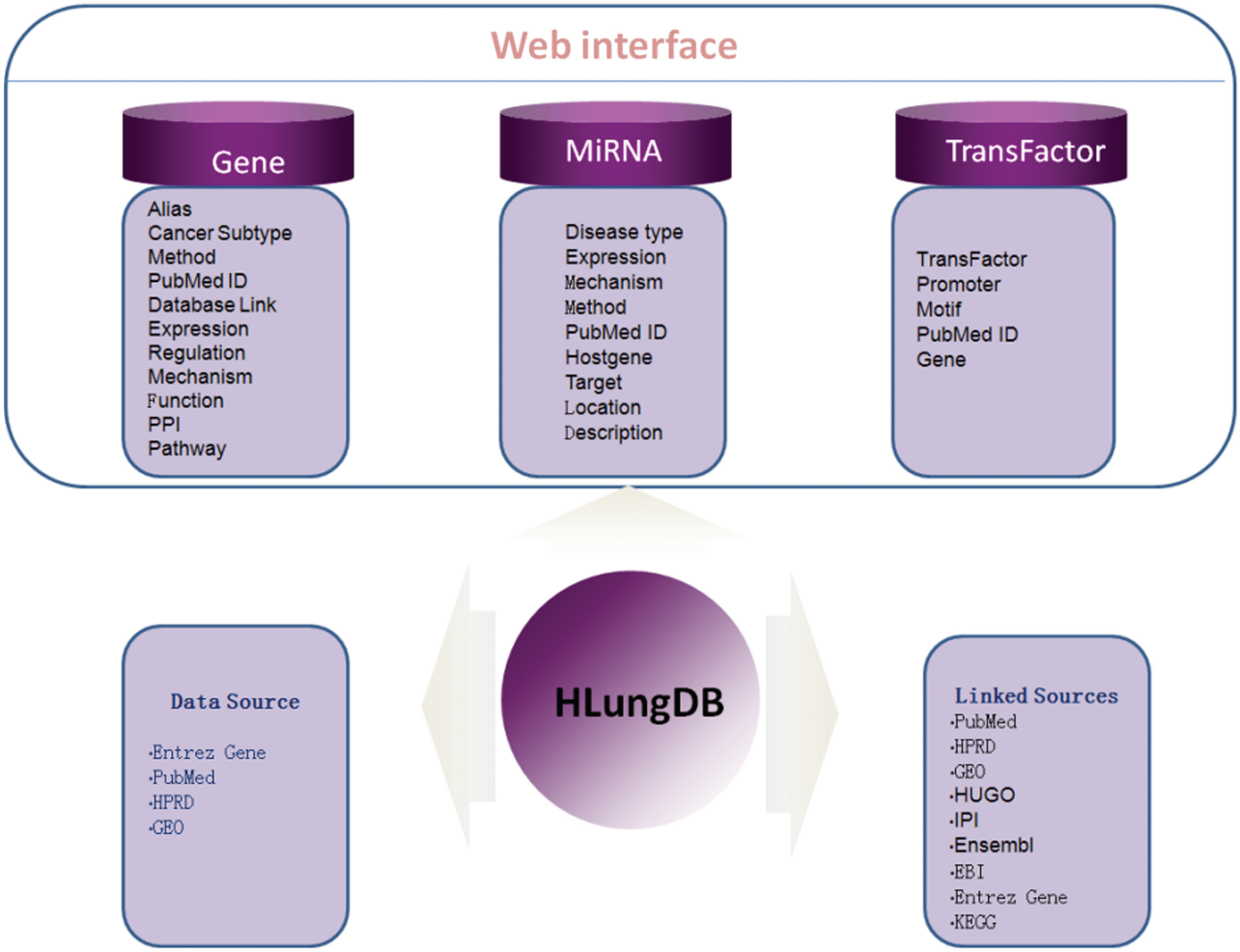

Figure 1. The database structure of HLungDB.

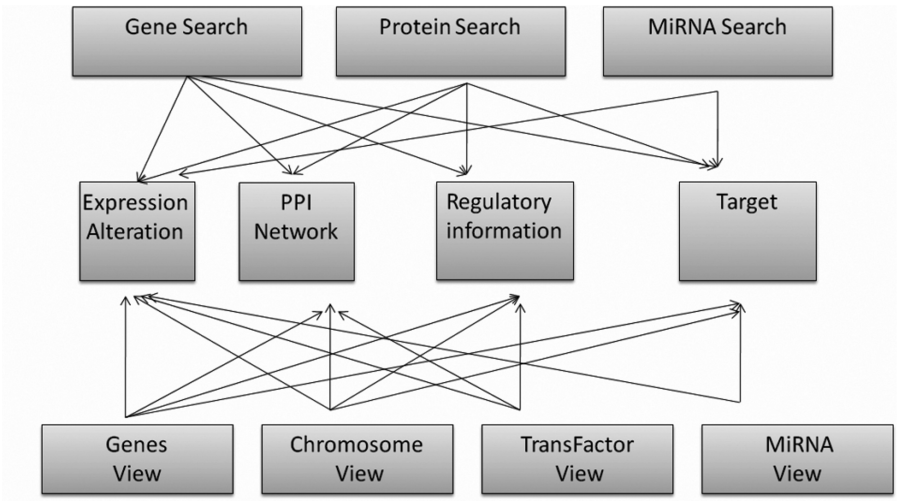

Figure 2. The flowchart of query in the HLungDB database.

microarray data and regulatory information if the gene has been confirmed to be related to lung cancer in our database. To see more details about how the gene is related to lung cancer, the user can click on the gene symbol link and a new page will appear to display evidence of the genes relationship to lung cancer. 'Clinical Significance' indicates the effect of the gene alteration on the lung cancer in the point of clinical view that is collected from the literature; 'Function' describes the gene's role in lung cancer extracted from the published papers; 'Gene Regulation' presents the regulatory relationship of the gene with other genes, while 'Expression Alteration' shows the analysis results of lung cancer related microarray datasets, in which the user can see how many patients show gene upregulation and/or downregulation.

The PPI link leads to a new page that shows the proteins interacting with the query protein, the 'Show PPI Network' link will display the selected protein–protein interaction network based on experimental evidence mostly from the HPRD system (23). In the PPI network section, user-friendly interfaces have made all the features of HLungDB PPI easily accessible and also provide direct view for the user to explore the relationship among the proteins.

The 'Regulatory Information' links the user to the names of the transcription factors confirmed to regulate the gene. 'See Details' links the user to a new page that displays the binding site motifs of those transcription factors with the supporting PubMed ID. The 'Show Promoter of Gene' link will display the promoter sequence(s) of the selected gene. A gene with an unknown transcription factor will only show its promoter sequence(s).

Alternatively, the user can query our system with the protein symbol, and a new summary page will provide a brief description of the protein, the PPI, the links to other related resources and the PPI network. Users can navigate each item in detail by clicking the related links.

Users can also check whether a miRNA is related to lung cancer with the miRNA symbol. The results page will display the manually collected details for the related miRNA, including the disease type the miRNA is related to, the alterations in expression of the miRNA, the mechanism of the miRNA in lung cancer, the experiment methods used to confirm the mechanism, the targets of the miRNA if any with PubMed ID and the description of the miRNA involved in lung cancer.

HLungDB provides two ways to view all lung cancer-related genes. The first approach is to query the database via visualized chromosome browser through 'Chromosome' listed on the first page. The user then clicks 'Chromosome' on the top of this page, and a chromosome map will return. In the Chromosome page, the user can view lung cancer-related genes by Chromosome ID. With the second approach, 'Browse' on the first page of HLungDB allows users to see all the genes confirmed to be related to lung cancer. The genes in this list are sorted by alphabetical order. Using these two approaches, users can easily retrieve all genes that are related to lung cancer.

Another way to view lung cancer-related genes is provided in the pathway view. On the pathway list, users can check those lung cancer genes by clicking on the pathway name and view the network about this pathway through the 'Pathway Network' entrance. User can also click on the marginal node on the network to expand the network. For more detailed usage of the network, users can read the annotation on the pathway network page.

Users can view lung cancer-related information in our database through browsing SNP, transcription factor and methylation lists. The 'SNP View' provides the user with lung cancer-related SNP obtained from PubMed by searching 'SNP' and 'lung cancer'. The 'TransFactor View' presents transcription factors related to lung cancer with other detailed information. The 'Methylation View' displays genes with epigenetic alterations observed in lung cancer.

In addition, convenient links are provided to other databases. HLungDB has been developed with crosslink to other relevant external resources. It includes the National Center for Biotechnology Information, a repository for published gene information, and PubMed, US National Library of Medicine, that includes over 18 million citations from MEDLINE and other life science journals. HPRD, HUGO, IPI, EBI and KEGG are also linked to HLungDB.

DISCUSSION

In order to provide a central resource for biologists in the lung cancer research community, we developed HLungDB, a database system aimed at providing a comprehensive resource of gene information and their relationships to lung cancer.

The goal of the lung cancer database project was to construct a large-scale platform for lung cancer that would contribute to basic research and clinical research in the future. In the past 2 years, large amounts of data have been collected for this project. Information on lung cancer data was obtained from the PubMed and GEO databases. Genes, miRNAs, gene promoters, transcription factors, transcription factor-binding sites and the SNPs related to lung cancer have been collected and integrated into this system. Clinical information related to gene expression profile data was also extracted from GEO. We have systematically extracted information from published lung cancer-related studies. The database currently contains 2585 full-text entries describing lung cancer and genes. They have been integrated in such a way that investigators can rapidly query whether a gene or protein is found in human lung cancer, and other detailed lung cancer-related information about this gene. User-friendly query interfaces have made all the features of HLungDB easily accessible.

HLungDB provides a comprehensive resource for human lung cancer research. We believe that HLungDB will be particularly interesting to the life science community and will greatly facilitate cancer biologists' mission of unraveling the pathogenesis of lung cancer.

FUTURE DIRECTIONS

We are working to increase the quality and quantity of data and to supply additional database function. We plan to adopt two strategies to achieve these goals. First, text-mining tools will be adopted to improve our data collection. We will use text-mining tools to help us prescreen PubMed abstracts regularly that potentially describe the lung cancer–gene relationships. Second, since many proteins in the signaling transduction pathways are involved in the lung cancer development and progression, our next step is to identify those signal transduction pathways that have significant changes and display their components with identified alteration in lung cancer in a network view. At the same time, we will also collect the downstream genes for each altered signaling pathway in lung cancer and further characterize the relationship between them to ultimately fulfill the goal of identifying new potentially relevant lung cancer genes and new mechanisms.

ACKNOWLEDGEMENTS

We are thankful for M. Tyler Hougland, Peng Li, Tian Xiao, Chao Zheng, Yan Feng, Rong Fang, Yijun Gao, Yujuan Jin, Zuoyun Wang, Xiankun Han, Junhua Zhang, Xiaolei Ye, Bin Gao, Hongling Huang, Fei Li, Ye Wang for technical supports.

FUNDING

State Key Program of Basic Research of China (Grant 2007CB108800, 2009CB918402, 2010CB912102); National High Technology Research and Development Program of China (863 project) (Grant No. 2006AA02Z313); National Natural Science Foundation of China (Grant 30870575, 30740084 and 30871284); Chinese Academy of Sciences (2008KIP101); Science and Technology Commission of Shanghai Municipality (06DZ22923, 08PJ14105). H.J. is a scholar of the Hundred Talents Program of the Chinese Academy of Sciences. Funding for open access charge: National Natural Science Foundation of China and the State Key Program of Basic Research of China.

Conflict of interest statement. None declared.

REFERENCES

1. Travis,W.D., Travis,L.B. and Devesa,S.S. (1995) Lung cancer. *Cancer*, **75**, 191–202.
2. Fong,K.M., Sekido,Y., Gazdar,A.F. and Minna,J.D. (2003) Lung cancer. 9: Molecular biology of lung cancer: clinical implications. *Thorax*, **58**, 892–900.
3. Herbst,R.S., Heymach,J.V. and Lippman,S.M. (2008) Lung cancer. *N. Engl. J. Med.*, **359**, 1367–1380.
4. Aviel-Ronen,S., Blackhall,F.H., Shepherd,F.A. and Tsao,M.S. (2006) K-ras mutations in non-small-cell lung carcinoma: a review. *Clin. Lung Cancer*, **8**, 30–38.
5. Devereux,T.R., Taylor,J.A. and Barrett,J.C. (1996) Molecular mechanisms of lung cancer. Interaction of environmental and genetic factors. Giles F. Filley Lecture. *Chest*, **109**, 14S–19S.
6. Ji,H., Ramsey,M.R., Hayes,D.N., Fan,C., McNamara,K., Kozlowski,P., Torrice,C., Wu,M.C., Shimamura,T., Perera,S.A. *et al.* (2007) LKB1 modulates lung cancer differentiation and metastasis. *Nature*, **448**, 807–810.
7. Ding,L., Getz,G., Wheeler,D.A., Mardis,E.R., McLellan,M.D., Cibulskis,K., Sougnez,C., Greulich,H., Muzny,D.M., Morgan,M.B. *et al.* (2008) Somatic mutations affect key pathways in lung adenocarcinoma. *Nature*, **455**, 1069–1075.
8. Engels,E.A., Wu,X., Gu,J., Dong,Q., Liu,J. and Spitz,M.R. (2007) Systematic evaluation of genetic variants in the inflammation pathway and risk of lung cancer. *Cancer Res.*, **67**, 6520–6527.
9. Wenzlaff,A.S., Cote,M.L., Bock,C.H., Land,S.J., Santer,S.K., Schwartz,D.R. and Schwartz,A.G. (2005) CYP1A1 and CYP1B1 polymorphisms and risk of lung cancer among never smokers: a population-based study. *Carcinogenesis*, **26**, 2207–2212.
10. Son,J.W., Kang,H.K., Chae,M.H., Choi,J.E., Park,J.M., Lee,W.K., Kim,C.H., Kim,D.S., Kam,S., Kang,Y.M. *et al.* (2006) Polymorphisms in the caspase-8 gene and the risk of lung cancer. *Cancer Genet. Cytogenet.*, **169**, 121–127.
11. Yin,J., Vogel,U., Ma,Y., Qi,R., Sun,Z. and Wang,H. (2007) The DNA repair gene XRCC1 and genetic susceptibility of lung cancer in a northeastern Chinese population. *Lung Cancer*, **56**, 153–160.
12. Tomoda,K., Ohkoshi,T., Hirota,K., Sonavane,G.S., Nakajima,T., Terada,H., Komuro,M., Kitazato,K. and Makino,K. (2009) Preparation and properties of inhalable nanocomposite particles for treatment of lung cancer. *Colloids Surf. B: Biointerfaces*, **71**, 177–182.
13. Amberger,J., Bocchini,C.A., Scott,A.F. and Hamosh,A. (2009) McKusick's Online Mendelian Inheritance in Man (OMIM). *Nucleic Acids Res.*, **37**, D793–D796.
14. He,X., Chang,S., Zhang,J., Zhao,Q., Xiang,H., Kusonmano,K., Yang,L., Sun,Z.S., Yang,H. and Wang,J. (2008) MethyCancer: the database of human DNA methylation and cancer. *Nucleic Acids Res.*, **36**, D836–D841.
15. Jiang,Q., Wang,Y., Hao,Y., Juan,L., Teng,M., Zhang,X., Li,M., Wang,G. and Liu,Y. (2009) miR2Disease: a manually curated database for microRNA deregulation in human disease. *Nucleic Acids Res.*, **37**, D98–D104.
16. Gu,D., Scaringe,W.A., Li,K., Saldivar,J.S., Hill,K.A., Chen,Z., Gonzalez,K.D. and Sommer,S.S. (2007) Database of somatic mutations in EGFR with analyses revealing indel hotspots but no smoking-associated signature. *Hum. Mutat.*, **28**, 760–770.
17. Landi,M.T., Dracheva,T., Rotunno,M., Figueroa,J.D., Liu,H., Dasgupta,A., Mann,F.E., Fukuoka,J., Hames,M., Bergen,A.W. *et al.* (2008) Gene expression signature of cigarette smoking and its role in lung adenocarcinoma development and survival. *PLoS ONE*, **3**, e1651.
18. Stearman,R.S., Dwyer-Nield,L., Zerbe,L., Blaine,S.A., Chan,Z., Bunn,P.A. Jr, Johnson,G.L., Hirsch,F.R., Merrick,D.T., Franklin,W.A. *et al.* (2005) Analysis of orthologous gene expression between human pulmonary adenocarcinoma and a carcinogen-induced murine model. *Am. J. Pathol.*, **167**, 1763–1775.
19. Wachi,S., Yoneda,K. and Wu,R. (2005) Interactome–transcriptome analysis reveals the high centrality of genes differentially expressed in lung cancer tissues. *Bioinformatics*, **21**, 4205–4208.
20. Rohrbeck,A., Neukirchen,J., Rosskopf,M., Pardillos,G.G., Geddert,H., Schwalen,A., Gabbert,H.E., von Haeseler,A., Pitschke,G., Schott,M. *et al.* (2008) Gene expression profiling for molecular distinction and characterization of laser captured primary lung cancers. *J. Transl. Med.*, **6**, 69.
21. Wrage,M., Ruosaari,S., Eijk,P.P., Kaifi,J.T., Hollmen,J., Yekebas,E.F., Izbicki,J.R., Brakenhoff,R.H., Streichert,T., Riethdorf,S. *et al.* (2009) Genomic profiles associated with early micrometastasis in lung cancer: relevance of 4q deletion. *Clin. Cancer Res.*, **15**, 1566–1574.
22. Spira,A., Beane,J.E., Shah,V., Steiling,K., Liu,G., Schembri,F., Gilman,S., Dumas,Y.M., Calner,P., Sebastiani,P. *et al.* (2007) Airway epithelial gene expression in the diagnostic evaluation of smokers with suspect lung cancer. *Nat. Med.*, **13**, 361–366.
23. Keshava Prasad,T.S., Goel,R., Kandasamy,K., Keerthikumar,S., Kumar,S., Mathivanan,S., Telikicherla,D., Raju,R., Shafreen,B., Venugopal,A. *et al.* (2009) Human Protein Reference Database—2009 update. *Nucleic Acids Res.*, **37**, D767–D772.