Cancer Treatment Reviews 39 (2013) 77–88

Contents lists available at SciVerse ScienceDirect

Cancer Treatment Reviews

journal homepage: www.elsevierhealth.com/journals/ctrv

Laboratory-Clinic Interface

Genetic network and gene set enrichment analysis to identify biomarkers related to cigarette smoking and lung cancer

Xiaocong Fang[a], Michael Netzer[b], Christian Baumgartner[b], Chunxue Bai[a], Xiangdong Wang[a,c,*]

[a] Department of Pulmonary Medicine, Zhongshan Hospital, Fudan University, No.180, Fenglin Road, Shanghai 200032, China
[b] Research Group for Clinical Bioinformatics, Institute of Electrical and Biomedical Engineering, UMIT, 6060 Hall in Tirol, Austria
[c] Biomedical Research Center, Zhongshan Hospital, Fudan University, Shanghai, China

ARTICLE INFO

Article history:
Received 24 April 2012
Received in revised form 3 June 2012
Accepted 6 June 2012

Keywords:
Lung cancer
Genes
Networks
Biomarkers
Smoking
Treatment

ABSTRACT

Objectives: Cigarette smoking is the most demonstrated risk factor for the development of lung cancer, while the related genetic mechanisms are still unclear.
Methods: The preprocessed microarray expression dataset was downloaded from Gene Expression Omnibus database. Samples were classified according to the disease state, stage and smoking state. A new computational strategy was applied for the identification and biological interpretation of new candidate genes in lung cancer and smoking by coupling a network-based approach with gene set enrichment analysis.
Measurements: Network analysis was performed by pair-wise comparison according to the disease states (tumor or normal), smoking states (current smokers or nonsmokers or former smokers), or the disease stage (stages I–IV). The most activated metabolic pathways were identified by gene set enrichment analysis.
Results: Panels of top ranked gene candidates in smoking or cancer development were identified, including genes involved in cell proliferation and drug metabolism like cytochrome P450 and WW domain containing transcription regulator 1. Semaphorin 5A and protein phosphatase 1F are the common genes represented as major hubs in both the smoking and cancer related network. Six pathways, e.g. cell cycle, DNA replication, RNA transport, protein processing in endoplasmic reticulum, vascular smooth muscle contraction and endocytosis were commonly involved in smoking and lung cancer when comparing the top ten selected pathways.
Conclusion: New approach of bioinformatics for biomarker identification and validation can probe into deep genetic relationships between cigarette smoking and lung cancer. Our studies indicate that disease-specific network biomarkers, interaction between genes/proteins, or cross-talking of pathways provide more specific values for the development of precision therapies for lung.

Introduction

It has become a major public health problem as a leading course of cancer death in men and increasingly in women, for which over 1 million people died every year in the world.[1] Cigarette smoking has been considered as one of the most demonstrated risk factors for the development of lung cancer where genetic variants are also considered to play an essential role. Cigarette smoking could induce dysregulation of genes or mRNAs, associated with lung cancer, carcinogenesis, and a great number of genes or pathways as potential candidates of targeted therapies for lung cancer.[2,3] However, the major challenge is the great variation between studies and the inconsistency of selected candidates which need further validation.

Bioinformatics is the application of omics science, information technology, mathematics and statistics in the field of biomarker

Abbreviations: NSCLC, non-small cell lung cancer; GSEA, gene set enrichment analysis; IG, information gain; WWTR1, WW domain containing transcription regulator 1; SEMA5A, sema domain, seven thrombospondin repeats (type 1 and type 1-like), transmembrane domain (TM) and short cytoplasmic domain, (semaphorin) 5A; Ppm1f, protein phosphatase 1F; Agtr1, angiotensin II receptor, type 1; Ramp2, receptor (G protein-coupled) activity modifying protein 2; GRK5, G protein-coupled receptor kinase 5; arap3, ArfGAP with RhoGAP domain-ankyrin repeat and PH domain 3; CRYAB, crystallin, alpha B; UBE2I, ubiquitin-conjugating enzyme E2I; ednrb, endothelin receptor B; LEPR, leptin receptor; Eif2α, eukaryotic initiation factor 2, α subunit; srpk1, SFRS protein kinase 1; MICAL, microtubule associated monoxygenase, calponin and LIM domain; nthl1, nth endonuclease III-like 1 (*E. coli*).

* Corresponding author.
E-mail addresses: fangxiaocong@gmail.com (X. Fang), michael.netzer@umit.at (M. Netzer), christian.baumgartner@umit.at (C. Baumgartner), bai.chunxue@zs-hospital.sh.cn (C. Bai), xiangdong.wang@telia.com (X. Wang).

0305-7372/$ - see front matter © 2012 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.ctrv.2012.06.001

discovery. There has been a great increase in genomics and other molecular research to produce a tremendous amount of information related to molecular biology.[4] Various powerful data mining and statistics methods have been propagated to handle the huge amount of data. Biomarker networks or dynamic biomarker networks were proposed as new way of biomarker identification and validation to emphasize the relationship between genes, proteins or metabolites, taking into account the kinetic characteristics.[5,6]

Numerous candidate biomarkers were reported to be associated with the development or prognosis of lung cancer using bioinformatics approaches.[7–9] Targeted therapy for lung cancer patients with epidermal growth factor receptor (EGFR) mutations was considered as an example of the promising development toward personalized therapy for lung cancer patients.[10] Public expression data with bioinformatics approaches was applied to identify useful biomarkers for lung cancer, which could be further validated in clinical trials.[11] Genetic relationships between cigarette smoking and lung cancer, e.g. the significance of differentially expressed genes resulted from cigarette smoking and roles in the development of lung cancer, should be further clarified. Cigarette smoking and lung cancer may share the same pathophysiological pathways, where the expression profile of genes may change along with the disease progress. The present studies propose a network-based approach to identify genes related to lung cancer and cigarette smoking, including the differentially expressed genes in patients with different smoking histories at different disease stages. The approach with gene set enrichment analysis (GSEA) was coupled to search for the activated pathways, which are involved in the pathophysiological processes between cigarette smoking and lung cancer, to identify a panel of expected and unexpected genes and associated pathways in lung cancer.

Data selection and recruitment

The preprocessed microarray expression dataset from the study conducted by Landi et al.[12] available at Gene Expression Omnibus online database (GDS3257) was used. The platform ID is GPL96. The data set comprises 107 samples, including 58 tumor tissues and 49 normal tissues from never smokers, non-smokers and former smokers. These samples were also classified according to the disease stages and gender (Table 1). There are 22,283 genes with defined Affymetrix_3PRIME_IVT_ID for each sample.

Coupled computational approach

Genetic network analysis

A new network-based biomarker discovery approach, as described previously,[6,13] was applied for the search for highly discriminatory genes according to their connectivity strength within the network. Inferring of the network for n features results in $\frac{n\cdot(n-1)}{2}$comparisons. The network for $n = 22{,}283$ genes in reasonable time was calculated as a filtering step (Step 1).

Overall, the modality includes two steps:

Step 1: Feature selection: In order to reduce the number of features (genes) by excluding the genes that show no or minimal discriminatory ability, the information gain (IG),[14] an entropy-based method, was applied for feature extraction. Reasons for choosing IG included (i) the ability to easily identify features (genes) with no discriminatory ability (i.e., IG = 0), (ii) the applicability to multiple class problems, and (iii) a low calculation time. The IG score for each gene in the dataset was calculated with respect to the classes (i.e., disease state, disease stage and smoking state). The IG score for a feature f is defined as follows[15]:

$$\mathrm{IG} = \mathrm{H(Y)} - \mathrm{H(Y|X)}, \text{where}$$

$$\mathrm{H(Y)} = -\sum_{y\in \mathrm{Y}} p(y)\ \log_2(p(y)), \text{and}$$

$$\mathrm{H(Y|X)} = -\sum_{x\in \mathrm{X}} p(x) \sum_{y\in \mathrm{Y}} p(y|x)\ \log_2(p(y|x)),$$

where H(Y) denotes the entropy for Y (class variable) and H(Y|X) is the entropy of Y after observing X. A threshold of IG = 0 was set to remove all features (genes) with no discriminatory ability.

In the present study the network analysis was performed by pair-wise comparison according to the disease states (tumor or normal), smoking states (current smokers or nonsmokers or former smokers), or the disease stages (stages I–IV).

Step 2: Inferring the network: On the reduced gene sets the networks were inferred including three steps:

(i) calculating all ratios R between genes G which represent gene g interactions, where $\mathrm{r}_{ij} = \left|\log_2\left(\frac{g_i}{g_j}\right)\right|$ with $i > j$, $\mathrm{r} \in \mathrm{R}$,
(ii) computing IG scores s_{ij}; $\mathrm{s} \in \mathrm{S}$ on the logarithmic ratios R, and
(iii) constructing a graph G with: $\mathrm{G}_{ij} = \left\{\frac{1 \ \ \mathrm{if}|s_{ij}|>\tau}{0 \ \ \mathrm{else}}\right.$ for $i, j \in g_1, g_2, \ldots, g_n$.

After constructing the network, the genes are ranked according to their degrees (i.e., number of edges of each gene) in the network.

GSEA

Briefly, GSEA searched for groups of genes that shared common biological function, chromosomal location, or regulation.[16] This analysis contained two steps:

Step 1: Feature selection by a non-specific filtering to remove those genes with high variants according to the disease state or the smoking history, respectively.

Step 2: (i) Selection of the activated metabolic pathways with more than 10 genes involved in the pathophysiological processes of cigarette smoking or lung cancer, respectively, according to Kyoto Encyclopedia of Genes and Genomes (KEGG) database; (ii) calculation of the T value for every gene and the mean T value for all genes involved in every specific pathway; (iii) ranking the pathways according to the mean value of T. Top ranked metabolic pathways were compared with regard to pathophysiological processes of cigarette smoking and lung cancer. Overlaid pathways were indicated.

Network analysis

Networks of different groups, i.e., disease state (tumor/normal), smoking history (current smoker/former smoker/nonsmoker), and disease stage (I/II/III/IV), respectively, were compared for lung cancer and cigarette smoking to identify and validate common genes by reviewing the online databases, including PubMed, Database for Annotation, Visualization and Integrated Discovery, and Kyoto Encyclopedia of Genes and Genomes database. R[17] was used to perform network analysis and GSEA. The overall work flow for the present study was described in Fig. 1.

Genetic network

For the identification of relevant gene markers, the IG score for each gene between the tumor tissues and normal tissues (T/NT), as well as in the current smokers, former smokers and never smokers (C/F/N) or current smokers and never smokers (C/N) was calculated. Out of 22,283 genes in the dataset, over 9365 genes were different between tumor and normal patient having a IG score >0,