Identification of Prognostic Genes and Pathways in Lung Adenocarcinoma Using a Bayesian Approach

Yu Jiang[1,2], Yuan Huang[2,3], Yinhao Du[4], Yinjun Zhao[3], Jie Ren[4], Shuangge Ma[2,3] and Cen Wu[4]

[1]Division of Epidemiology, Biostatistics and Environmental Health, School of Public Health, University of Memphis, Memphis, TN, USA. [2]Cooperative Studies Program, VA Connecticut Healthcare System, West Haven, CT, USA. [3]Department of Biostatistics, Yale University, New Haven, CT, USA. [4]Department of Statistics, Kansas State University, Manhattan, KS, USA.

Cancer Informatics
1–7

DOI: 10.1177/1176935116684825

ABSTRACT: Lung cancer is the leading cause of cancer-associated mortality in the United States and the world. Adenocarcinoma, the most common subtype of lung cancer, is generally diagnosed at the late stage with poor prognosis. In the past, extensive effort has been devoted to elucidating lung cancer pathogenesis and pinpointing genes associated with survival outcomes. As the progression of lung cancer is a complex process that involves coordinated actions of functionally associated genes from cancer-related pathways, there is a growing interest in simultaneous identification of both prognostic pathways and important genes within those pathways. In this study, we analyse The Cancer Genome Atlas lung adenocarcinoma data using a Bayesian approach incorporating the pathway information as well as the interconnections among genes. The top 11 pathways have been found to play significant roles in lung adenocarcinoma prognosis, including pathways in mitogen-activated protein kinase signalling, cytokine-cytokine receptor interaction, and ubiquitin-mediated proteolysis. We have also located key gene signatures such as *RELB, MAP4K1,* and *UBE2C*. These results indicate that the Bayesian approach may facilitate discovery of important genes and pathways that are tightly associated with the survival of patients with lung adenocarcinoma.

KEYWORDS: Bayesian approach, cancer prognosis, TCGA, lung adenocarcinoma, pathway analysis

RECEIVED: June 22, 2016. **ACCEPTED:** November 24, 2016.

PEER REVIEW: Seven peer reviewers contributed to the peer review report. Reviewers' reports totaled 1552 words, excluding any confidential comments to the academic editor.

TYPE: Methodology

FUNDING: The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by the National Institutes of Health (CA182984, CA142774), the National Social Science Foundation of China (13CTJ001, 13&ZD148), and the VA Cooperative Studies Program of the Department of VA, Office of Research and Development.

DECLARATION OF CONFLICTING INTERESTS: The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

CORRESPONDING AUTHOR: Cen Wu, Department of Statistics, Kansas State University, Manhattan, KS 66506, USA. Email: wucen@ksu.edu

Introduction

Lung cancer is the second most common cancer among both men and women in the United States and the leading cause of cancer-related mortality all over the world. In 2016, an estimate of 224 390 new cases and 158 080 deaths from lung cancer is expected.[1] Lung adenocarcinoma, a subtype of non–small cell lung cancer, is the most common form of lung cancer in the United States. Clinical and pathologic features for prognosis have been studied extensively, including race, age at diagnosis, smoking status, tumour stage, performance status, liver metastases, and comorbidity disease.[2–4] Due to tumour molecular heterogeneity, patients with similar clinic-pathologic characteristics may experience different disease outcomes. It suggests that molecular biomarkers are important in lung cancer prognosis.[5,6] Despite the long-term research, the understanding of the molecular mechanisms for lung cancer prognosis is still quite limited. Therefore, identification and the further study of prognosis biomarkers are critical in better understanding the disease progression, predicting patient disease outcome, defining patient subpopulation, and developing therapeutic targets.

Considerable efforts have been devoted to identifying biomarkers that may be associated with lung adenocarcinoma prognosis. Single-marker strategies prevail especially in early days, where one or a small number of markers can be analysed at a time. The major genes identified include *EGFR*, *RAS*, *P53*, *BCRA1*, *RRM1*, *beta Tubulin*, and others, as summarized by Rose-James and Sreelekha[5] and Sholl.[6] These genes mostly are found to be proto-oncogenes and tumour suppressor genes and belong to mitogen-activated protein kinase (MAPK) pathway, cell cycle, DNA repair, and apoptosis pathways.[5]

Despite huge success, marginal analysis is limited, in that it ignores the joint actions of multiple genes and pathways for the progression of lung cancer. Pathway-based analysis, on the contrary, is a representative and most extensively investigated method that fully takes the advantage of the functional relatedness among genes.[7] It can accommodate the joint association among genes and is considered to have greater power in the prediction of disease outcomes and produce more reliable results, compared with single marker strategies.[7] The most popular pathway-based analysis method is gene set enrichment analysis (GSEA[8,9]). The related methods in this subject and available tools are summarized by Jin et al.[10] Lee et al[11] selects 11 pathways that are significantly associated with lung cancer using both GSEA and adaptive rank-truncated product methods in genome-wide association study. Using a combined pathway-based risk score approach, Chang et al[12] identified 15 pathways that are associated with survival in lung carcinoma, and the top 3 pathways are HMGB1/RAGE signalling, beta-adrenergic receptor regulation of extracellular signal–regulated kinase (ERK), and clathrin-coated vesicle cycle.

A major limitation of the current pathway-based methods is that most of them only conduct pathway-level selection. Because not all genes are key drivers in the pathway, pinpointing important genes is also of great interest, in addition to pathways. We term the identification at both pathway level and gene level as bilevel selection in this article. Furthermore, the lung adenocarcinoma data analysed in many existing studies, such as Chang et al (2013), Lu et al,[13] and Li et al,[4] have been generated a decade or even 2 decades ago. Rigorous statistical analysis of most up-to-date, high-quality lung adenocarcinoma data is thus in pressing need.

In this study, we apply the Bayesian approach developed in Stingo et al[14] to The Cancer Genome Atlas (TCGA) lung adenocarcinoma data to identify prognostic pathways and the important genes involved in the relevant biological processes. Our work may complement existing studies and be warranted in the following aspects. First, it conducts a Bayesian analysis of the lung adenocarcinoma data. The pathway and gene-gene interaction information has been incorporated in the Bayesian framework. Posterior inclusion probabilities of pathways and genes are available after Markov chain Monte Carlo (MCMC) converges, which provide us a solid ranking criterion according to importance. To the best of our knowledge, such analysis has not yet been carried out for lung adenocarcinoma. Second, we have performed a timely analysis on the TCGA lung adenocarcinoma data. The Cancer Genome Atlas, organized and conducted by National Institutes of Health and the related participated research institute, is to 'generate comprehensive, multidimensional maps of the key genomic changes' in cancer. On the same subjects, multiple types of omics changes, such as messenger RNA (mRNA) gene expression, microRNA (miRNA), copy number alteration, and DNA methylation, have been profiled and made available in TCGA. For lung adenocarcinoma, the TCGA data have recently been collected with high quality and subsequently published by National Cancer Institute, making it possible to conduct analysis and more accurately describe its prognosis.

Method

TCGA lung adenocarcinoma data

The Cancer Genome Atlas is one of the largest cancer genomic studies providing comprehensive genomic characterization for a cohort of cancer and normal samples. We are interested in analysing the TCGA lung adenocarcinoma data with gene expression measurements. The data collection forms for patient enrolment and follow-up are available at http://www.nationwidechildrens.org/tcga-clinical-data-forms-standard#j-l. Details about the TCGA genomics measurements, data collection, and preliminary analysis can be found in TCGA (2014).[47] Both survival and gene expression data are downloaded from the Cancer Genomics Data Server using CGDS-R package on March 24, 2016.[15]

By the time of downloading the data, the total number of samples with gene expression measured by Illumina Hiseq RNAseq V2 platform is 517, whereas 32 samples have been analysed by Affymetrix microarray. To avoid bias, we only include samples measured by RNAseq in the analysis. The downloaded data set is 'luad_tcga_rna_seq_v2_mrna_median_Zscores' with robust z scores as the main entries. The scores have already been lowess normalized, log transformed, and median centred. The total number of genes is 18 908.

To study the prognostic marker for patients with earlier stage of lung adenocarcinoma, a total of 388 qualified samples are identified. Details of the patient information are available in Supplementary Table 1. In this study, overall survival is our prognosis outcome of interest. This primary endpoint is defined as the time from diagnosis of lung adenocarcinoma to either last known date alive or death. Patients who are known to be alive have been censored at the time of last contact. Among the total 388 samples, 118 died during follow-up. The median survival time is calculated to be 40.3 months, with 95% confidence interval (33.8-50.0 months).

To determine the pathway membership of all the genes, we first downloaded the Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways from Molecular Signature Database (MSigDB) (http://software.broadinstitute.org/gsea/msigdb/collections.jsp#C2). The total number of pathways is 186, with sizes ranging from 10 to 389 and a median of 53. The 2 pathways that only have 10 genes are taurine and hypotaurine metabolism pathway, and limonene and pinene degradation pathway. The largest pathway is olfactory transduction.

After matching with the TCGA lung cancer genomics data sets, the total number of unique genes is 4994. Two genes that have identical measures for the sample are excluded, and they are '*OR6K2*' and '*FGF6*'. There are a total of 4992 genes involved in the study.

Bayesian modelling

The goal of the analysis is to simultaneously identify important pathways and genes within those pathways for cancer survival. To make this article self-contained, we briefly summarize the Bayesian approach in a study by Stingo et al[14] below.

Denote T as the survival time, C as the censoring time, and $\delta = I(T \leqslant C)$ as the censoring indicator. Then, we observe $(Y = \min(T, C), \delta = I(T \leqslant C))$ under right censoring. Consider the gene expression measurements and denote $X = (X_1, \ldots, X_p)$ as the $n \times p$ gene expression matrix for n i.i.d subjects. The accelerated failure time (AFT) model has been adopted for the survival outcome, and the data augmentation approach in Tanner and Wong (1987)[16] has been taken to impute the censored outcomes. In particular, for the ith subject, let

$$\begin{cases} Z_i = \log(Y_i), & \text{if } \delta_i = 1 \\ Z_i > \log(Y_i), & \text{if } \delta_i = 0 \end{cases} \quad (1)$$

and the latent vector $Z = (Z_1, \ldots, Z_n)'$.

Table 1. Top 11 pathways selected in lung adenocarcinoma.

NAME OF THE PATHWAYS	NO. OF GENES	POSTERIOR PROBABILITY
Purine metabolism	154	1
MAPK signalling pathway	260	1
Cytokine-cytokine receptor interaction	254	1
Ubiquitin-mediated proteolysis	134	1
Lysosome	118	1
Aminoacyl tRNA biosynthesis	41	.99879
Neuroactive ligand-receptor interaction	268	.995925
ABC transporters	44	.982115
Spliceosome	122	.985495
Endocytosis	172	.98706
Peroxisome	78	.982835

Abbreviation: MAPK, mitogen-activated protein kinase; tRNA, transfer RNA.

To select both prognostic pathways and important genes within the pathways at the same time, 2 binary indicator vectors, a $G\times 1$ vector α and a $p\times 1$ vector β for the inclusion of pathways and individual genes, respectively, have been created. We assume the following AFT model:

$$Z = a + \sum_{g=1}^{G_\alpha} U_{g(\beta)} b_{g(\beta)} + \varepsilon, \varepsilon \sim N\left(0, \sigma^2 I\right)$$

where $G_\alpha = \sum_{g=1}^{G} \alpha_g$ is the number of chosen pathways, and $U_{g(\beta)}$ is the first latent principal component analysis (PCA) component produced from the expressions of selected genes in pathway g. The effect of the PCA latent component from pathway g on the response is measured by regression coefficient b_g. Prior distributions have been assigned to the 2 indicator vectors, respectively. The prior distribution takes into consideration not only the pathway membership of genes but also gene-gene interactions using a Markov random field (MRF). We need to obtain the pathway and gene-gene interaction information to fully characterize the prior distributions. For all the p genes in our study, a $G\times p$ matrix T with binary entries has been generated to indicate the pathway memberships, where entry $t_{gj} (1\leqslant g \leqslant G, 1\leqslant j \leqslant p)$ is 1 if gene j is in pathway g, and 0 otherwise.

Different from constructing gene dependence structure using known biological information from KEGG as suggested in the original paper, we build the prior for MRF via gene networks. In a gene network, a node is corresponding to the expression of a gene. Two nodes are connected if the 2 gene expressions are associated statistically or biologically. The most important element for constructing a network is the adjacency matrix which quantifies the strength of connection between any 2 nodes.[17] Let $R = (r_{ij}, 1\leqslant i, j \leqslant p)$ be the adjacency matrix and c_{ij} be the Pearson correlation coefficient between nodes (gene expressions) i and j. We propose $r_{ij} = c_{ij}^d (c_{ij} > c)$ with $d=5$ to measure the connection intensity. Such measure retains the strong correlations while downweighing the weak ones. Furthermore, it guarantees that r_{ij} and c_{ij} have the same sign. c is the cut-off computed from Fisher transformation. The gene network is weighted and sparse. Note that in Stingo et al,[14] the adjacency matrix R has been constructed using the biological information. That is, the matrix R describes whether there is a direct link between 2 genes with a 0/1 based on the information extracted from KEGG database. We expect that generating network from the data-driven perspective is equally applicable. We also acknowledge that there are multiple ways to generate the adjacency matrix R statistically. As our purpose is not to compare different network measures, we focus on this particular network structure in this article.

Gene and pathway identification

The aforementioned data-driven approach has been adopted to specify the MRF prior. We run 2 MCMC chains in the data analysis, each chain with a total number of 150 000 iterations, where 50 000 are burn-ins. The starting numbers of included pathways are 20 and 5, respectively. The first principal component for each pathway has been used in the analysis, and the 2 chains are combined for posterior inference. We set the hyperparameter of the MRF as −3.5 to control the sparsity of the model. The hyperprior that controls the smoothness of the distribution of gene selection is set as .04. The marginal posterior inclusion probability for pathway and the conditional posterior gene inclusion probability are calculated accordingly.

To evaluate the performance of the methods used in this study, we compare the proposed method with 2 alternative approaches: (1) pathway and gene selection without

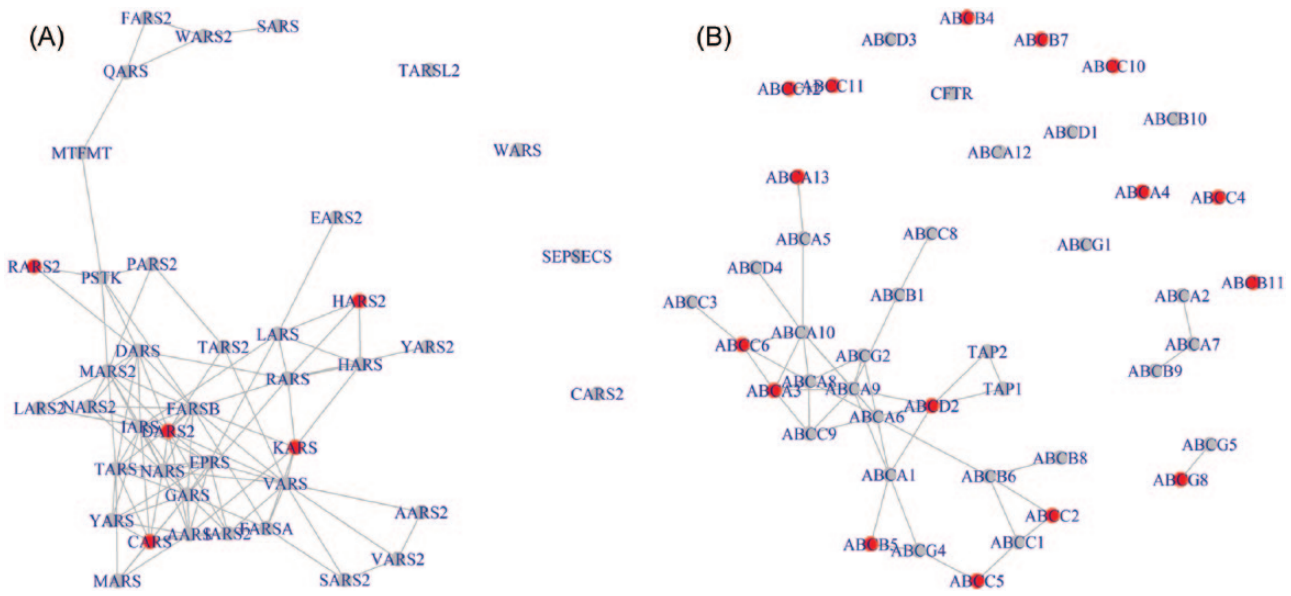

Figure 1. (A) The genes (dots, the identified genes are in red) and network of amino-acyl transfer RNA biosynthesis pathway and (B) the genes (dots, the identified genes are in red) and network of ABC transporters pathway.

incorporating network information, which is equivalent to using 0 as prior for MRF, and (2) pathway and gene selection incorporating prior information based on KEGG pathway introduced by Stingo et al.[14] We randomly split the whole data set into training data set and testing data set using a 3:1 ratio. The training data set (n = 291) is analysed using the 3 methods. The testing data set is used to compute the predicted survival time and mean squared error (MSE). All the analysis is conducted in MATLAB, using code modified from http://www.stat.rice.edu/~marina/software.html.

Results

We obtain the marginal posterior probability for all the pathways. The number of visited pathways by MCMC samplers is shown to be around 95, and the number of genes is around 190. The 2 chains show high agreement (Supplementary Figure 1), and the correlation between marginal posterior probabilities of pathway selections is .93. The results from the 2 chains are combined and summarized in Supplementary Figure 2. Among the 186 pathways, there are 11 pathways with the posterior marginal probability larger than .98. The selected pathways are presented in Table 1. Conditional on the selected pathways, we further evaluate the gene selections with a marginal probability cut-off .98. The conditional posterior probabilities for each gene are shown in Supplementary Figure 3. We choose 2 pathways, the amino-acyl transfer RNA (tRNA) biosynthesis pathway and the ABC transporters pathway, from the 11 representative examples and show their gene networks in Figure 1A and B, respectively. The top-ranked genes in terms of conditional posterior probabilities are marked in red. Using training and testing data sets, the predictive MSE is 6.98 using network-constructed MRF, 8.27 without prior information

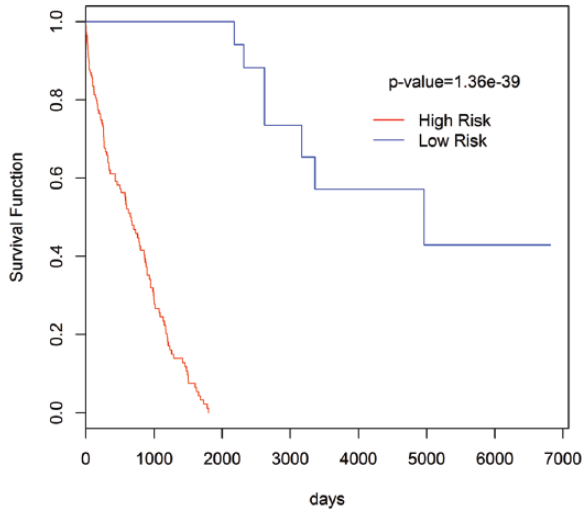

Figure 2. Survival curves for the low-risk (red lines) and high-risk (blue lines) samples using the proposed methods. The *P*-values are from the log-rank test.

(MRF = 0), and 12.68 using KEGG pathway information as prior for MRF, which indicates that the proposed method outperforms the alternatives of prediction.

Using the proposed method and the 2 alternative approaches, we compute each subject's posterior predicted survival time and dichotomize it to create 2 risk groups: high-risk group and low-risk group. The difference in the survival of the 2 risk groups has been compared by log-rank test. Figure 2 shows the survival curves of 2 risk groups predicted by the proposed model. The 2 risk groups have a very different survival function. The survival of high-risk group has a much lower survival probability compared with low-risk group. The log-rank statistic is 272, and the *P*-value is1.39e−39. The log-rank statistics obtained by the other 2 methods are both 269.

There are 5 pathways chosen with a posterior marginal probability of 1. They are MAPK signalling pathway, purine metabolism, cytokine-cytokine receptor interaction, ubiquitin-mediated proteolysis, and lysosome. It is not surprising that the MAPK pathway has been identified critical in the survival of lung cancer. The MAPK pathway consists of the family of serine/threonine protein kinase that links extracellular signal to fundamental cellular processes, such as cell proliferation and differentiation, stress response, and apoptosis.[18] In this pathway, the top-ranked genes are *RELB*, *MAP4K1*, *MAP3K2*, *RASGRF2*, *NTF3*, *ATF4*, *MAP4K4*, *NFATC2 PLA2G2D*, and *PLA2G10*. Among these genes, *RELB* is an NF-κB family member and has been found to suppress cigarette smoke–induced *COX-2* through miR-146a.[19] MAP4Ks, a family of MAPK, play important roles in cell transformation, adhesion, migration, and invasion.[20] MAP4K1 has been reported to stimulate NF-κB signalling, which in turn inhibits the process of apoptosis. Higher relative copy number of MAP4K1 has been related to the increased risks of death in colon cancer patients treated with oxaliplatin-based chemotherapy.[21] A study in stage II pancreas cancer has discovered that overexpression of MAP4K4 is associated with poorer survival.[22] MAP4K has also been shown to be a prognostic maker in patients with hepatocellular carcinoma.[23] Among the most selected 10 genes, 2 genes, *PLA2G2D* and *PLA10*, are from phospholipase A2 family. The major function of PLA2 is to hydrolyse glycerol phospholipids and produce lysophospholipids and free fatty acids. The PLA2 proteins are important in inflammation and immune response. Elevated serum group PLA2 has been observed in patients with advanced cancer.[24] Higher expression of group IIa secretory phospholipase A2 has been found to be positively associated with metastasis in patients with lung adenocarcinoma and shorter survival time.[25]

Another pathway with marginal posterior probability 1 is cytokine-cytokine receptor pathway, which is critical in the regulation of immune response. The top-ranked genes according to inclusion probabilities are as follows: *EDA, IL18RAP*, *EPOR*, *TNFRSF19*, *TNFRSF11A*, *BMP7*, *CXCL9*, *IL12RB2*, *IL4R*, and *IL-19. EDA*, *ectodysplasin A*, *TNFRSF19*, and *TNFRSF11A* all belong to tumour necrosis factor (TNF) receptor family proteins. These receptors can bind to various TNF receptor-associated factor (TRAF) family proteins, through which they will induce the activation of NF-κB and MAPK8/c-Jun N-terminal kinase (JNK) pathway. Overexpression of TNF receptor proteins has been shown to positively correlate with bone metastasis and poor survival in various cancers.[26,27] Interleukins are cytokines that involve in cell immunity. IL18RAP encodes a protein receptor for IL-18 and plays a key role in IL-18 signalling. Higher levels of inflammatory cytokine IL-18 in the serum of patients with hepatocellular carcinoma is positively associated with worse survival.[28] IL-19 expression has been reported to be prognostic in breast cancer, and it is related to increased mitotic figures, advanced tumour stage, higher metastasis, and poor survival.[29]

Ubiquitin-mediated proteolysis pathway also has marginal posterior probability 1. Modification of ubiquitin in a number of protein targets within cells is indispensable to a diversity of biological processes, including regulation of gene expression, DNA repair, assembly of ribosomes, and programmed cell death.[30] According to inclusion probabilities, the top genes are *KLHL13*, *HERC1*, *UBE2C*, *UBE3B*, *UBE2QL1*, *CUL7*, *SAE1*, *FZR1*, *MID1*, and *RNF7*. Increasing lines of evidence have shown that the HERC family proteins are the essential components of a broad spectrum of cellular functions. Among these are cell growth, DNA damage repair, immune response, and neurodevelopment.[31] Ubiquitin includes 3 classes of enzymes, ubiquitin-activating enzymes (UBE1), ubiquitin-conjugating enzymes (UBE2), and ubiquitin-protein ligases (UBE3). UBE2C encodes a protein in the E2 ubiquitin-conjugating enzyme family, which plays critical roles in the destruction of cyclins in mitosis. UBE2C-transfected lung cancer cells have a higher percentage in S phase and increased cell proliferation and enhanced cell invasion.[32] It is suspected that the regulation of UBE2C on cell growth and apoptosis is coupled with ERK pathway.[33] The gene expression of *UBE2C* is found to increase in lung cancer tissues[34] and has been identified as one of the prognostic markers in lung adenocarcinoma by Li et al.[4] *UBE3B* is a member of ubiquitin-protein ligases, which functions in transferring ubiquitin from ubiquitin-conjugating enzyme to target substrates. *UBE3* is indispensable in cell proliferation, apoptosis, and DNA repair.[35]

Purine metabolism pathway is also among the pathways with marginal posterior probability 1. The most frequently selected genes in this pathway are *NUDT9*, *PDE7B*, *NUDT2*, *PDE6A*, *IMPDH1*, *AK5*, and *AMPD2*. This is a metabolic pathway for the synthesis and breakdown of purines in many organisms. Weber[36] has reported the connection between transformation and progression in cancer cells and the imbalance in the enzymic pattern of purine metabolism. In particular, PDE7B is a new therapeutic target in glioblastoma (GBM). The overexpression of PDE7B leads to the growth of a stem-like cell subpopulation in vitro and facilitates the tumour expansion in an in vivo intracranial GBM model.[37]

The fifth pathway that has marginal posterior probability of 1 is the lysosome pathway, with the top genes *GM2A*, *SUMF1*, *NPC2*, *CTSA*, *CD63*, *ASAH1*, *ACP2*, *ABCA2*, *SLC11A2*, and *GGA1*. This pathway mainly includes arsenal of degradative enzymes. GM2A encodes protein ganglioside GM2 activator, which works together with lysosomal enzyme beta-hexosaminidase to catalyse the degradation of the ganglioside GM2 in breast cancer cell lines.[38] The major function of Niemann-Pick type C2 (NPC2) is to regulate the transport of cholesterol through the lysosomal system. A recent study shows that NPC2 can be secreted by early-stage lung tumours. In return, it interferes with the tumour microenvironment.[39] Overall, more and more attention has been paid to this pathway, and it is becoming an area of interest in oncology.[40]

Aminoacyl tRNA pathway biosynthesis has a marginal posterior probability of .999, with the top genes *DARS2*, *HARS2*, *KARS*, *RARS2*, *CARS*, *VARS*, *TARS2*, *PARS2*, *AARS2*, and *TARS*. Neuroactive ligand-receptor interaction has a marginal posterior probability of .996, and *NPBWR1*, *DRD3*, *CYSLTR1*, *MAS1*, *NMUR1*, *AGTR2*, *LPAR3*, *CHRNA10*, *AGTR1*, and *F2RL3* are the top genes in the pathway. Another pathway selected is endocytosis, with selected genes *TSG101, CHMP2B, PDCD6IP, STAM2, RAB11FIP3, CHMP2A, EPN2, VPS37C*, and *PSD*. The last 2 pathways are ABC transporters and spliceosome. We provide a list of explanations for all the aforementioned gene symbols in Supplementary List 1.

Discussion

In this study, we conduct analysis on the recently published TCGA lung adenocarcinoma with gene expression measurements. The main conclusion is that incorporation of the pathway and gene correlation information significantly improves the prediction precision. We provide a data-driven alternative to build the prior for gene dependence, which yields satisfactory results. In the data analysis, we identify 11 pathways that are important for lung adenocarcinoma prognosis. Among the selected pathways, MAPK pathway has been studied extensively in many types of cancers in TCGA. Subsets of MAPK pathways, such as signalling through ERK and JNK, have also been reported in other pathway analyses.[11] The identified genes and pathways are worth further investigation as biomarkers for lung adenocarcinoma progression.

The Bayesian approach adopted in this article achieves bilevel selection on the pathway and gene levels, or group and individual levels in a more general sense, by ranking according to posterior probabilities. Bilevel selection has been investigated intensively under the frequentist penalization framework. Breheny and Huang[41] developed a composite group minimax concave penalty (MCP) penalty which applies an outer MCP penalty to the sum of a group of inner MCP penalties to achieve selection at the individual level and group level simultaneously. Friedman et al[42] proposed the sparse group lasso criterion by adding the lasso penalty to the group lasso. The sparsity at and within group can consequently be attained at the same time. Multiple studies trail behind the innovative works, as reviewed in Huang et al[43] and seen in papers thereafter. Existing studies on penalized bilevel variable selection perform well when covariates in the same group are strongly correlated, whereas the correlations are moderate or even low when they belong to different groups. However, such situation is not common in practice as genes not in the same pathway may also exhibit high correlations. In the literature, multiple penalization approaches have been developed to identify important genes while incorporating the interconnections among genes, namely, gene-gene interactions or gene networks, as shown in Li and Li[44] and the following papers. As integrating these gene correlation information in the simultaneous selection of pathways and genes has not been fully examined in the penalization framework, we turn to the Bayesian formulation.

Our analysis can be potentially improved in the following aspects. First, as multiple types of omics measurements in addition to gene expressions are available in TCGA for lung adenocarcinoma, integrative analysis can be conducted. It is worthwhile to examine whether inclusion of one or several more types of omics features, such as methylation or copy number alteration, will lead to better prediction or identification results. Second, as contamination of prognosis data is common in genetics studies,[45] it is necessary to develop robust models to select important pathways and markers. We can robustify the adopted approach by assigning heavy-tailed error distributions to the AFT model, like Sha et al.[46] Furthermore, as the Bayesian method is computationally intensive, it is urgent to develop penalized bilevel variable selections taking the gene-gene interaction into consideration.

We acknowledge that more extensive bioinformatics and functional studies are needed to fully understand the identified results. The approach can also be applied to other cancer types from TCGA or different databases. Those investigations will be postponed to future studies.

Author Contributions

YJ, CW, and SM conceived and designed the study. YJ and CW analysed the data. YJ and CW wrote the first draft of the manuscript. YH, YD, YZ, and JR contributed to the writing of the manuscript. YJ, CW, YH, YZ, JR, YD, and SM agree with manuscript results and conclusions. YJ, CW, and SM jointly developed the structure and arguments for the paper. YH, YD, YZ, JR, and SM made critical revisions and approved the final version. All authors reviewed and approved the final manuscript.

Disclosures and Ethics

As a requirement of publication, author(s) have provided to the publisher signed confirmation of compliance with legal and ethical obligations including but not limited to the following: authorship and contributorship, conflicts of interest, privacy and confidentiality, and (where applicable) protection of human and animal research subjects. The authors have read and confirmed their agreement with the ICMJE authorship and conflict of interest criteria. The authors have also confirmed that this article is unique and not under consideration or published in any other publication, and that they have permission from rights holders to reproduce any copyrighted material. Any disclosures are made in this section. The external blind peer reviewers report no conflicts of interest.

REFERENCES

1. American Cancer Society. *Cancer Facts & Figures 2016*. Atlanta, GA: American Cancer Society; 2016.
2. Tas F, Ciftci R, Kilic L, Karabulut S. Age is a prognostic factor affecting survival in lung cancer patients. *Oncol Lett*. 2013;6:1507–1513.

3. Sørensen JB, Badsberg JH, Olsen J. Prognostic factors in inoperable adenocarcinoma of the lung: a multivariate regression analysis of 259 patients. *Cancer Res*. 1989;49:5748–5754.
4. Li Y, Tang H, Sun Z, et al. Network-based approach identified cell cycle genes as predictor of overall survival in lung adenocarcinoma patients. *Lung Cancer*. 2013;80:91–98.
5. Rose-James A, Sreelekha TT. Molecular markers with predictive and prognostic relevance in lung cancer. *Lung Cancer Int*. 2012;2012:Article ID 729532 (12 pp.).
6. Sholl LM. Biomarkers in lung adenocarcinoma: a decade of progress. *Arch Pathol Lab Med*. 2015;139:469–480.
7. Ramanan VK, Shen L, Moore JH, Saykin AJ. Pathway analysis of genomic data: concepts, methods, and prospects for future development. *Trends Genet*. 2012;28:323–332.
8. Mootha VK, Lindgren CM, Eriksson KF, et al. PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nat Genet*. 2003;34:267–273.
9. 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 U S A*. 2005;102:15545–15550.
10. Jin L, Zuo XY, Su WY, et al. Pathway-based analysis tools for complex diseases: a review. *Genomics Proteomics Bioinformatics*. 2014;12:210–220.
11. Lee D, Lee GK, Yoon KA, Lee JS. Pathway-based analysis using genome-wide association data from a Korean non-small cell lung cancer study. *PLoS ONE*. 2013;8:e65396.
12. Chang YH, Chen CM, Chen HY, Yang PC. Pathway-based gene signatures predicting clinical outcome of lung adenocarcinoma. *Sci Rep*. 2015;5:10979.
13. Lu TP, Chuang EY, Chen JJ. Identification of reproducible gene expression signatures in lung adenocarcinoma. *BMC Bioinformat*. 2013;14:371.
14. Stingo FC, Chen YA, Tadesse MG, Vannucci M. Incorporating biological information into linear models: a Bayesian approach to the selection of pathways and genes. *Ann Appl Stat*. 2011;5:1978–2002.
15. Gao J, Aksoy BA, Dogrusoz U, et al. Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. *Sci Signal*. 2013;6:pl1.
16. Tanner MA, Wong WH. The calculation of posterior distributions by data augmentation. *JASA*. 1987;82:528–540.
17. Zhang B, Horvath S. A general framework for weighted gene co-expression network analysis. *Stat Appl Genet Mol Biol*. 2005;4:1128.
18. Dhillon AS, Hagan S, Rath O, Kolch W. MAP kinase signalling pathways in cancer. *Oncogene*. 2007;26:3279–3290.
19. Zago M, Rico SA, Hecht E, et al. The NF-κB family member RelB regulates microRNA miR-146a to suppress cigarette smoke-induced COX-2 protein expression in lung fibroblasts. *Toxicol Lett*. 2014;226:107–116.
20. Chuang HC, Wang X, Tan TH. MAP4K family kinases in immunity and inflammation. *Adv Immunol*. 2016;129:277–314.
21. Lin M, Zhang Y, Li A, et al. High-throughput RNAi screening of human kinases identifies predictors of clinical outcome in colorectal cancer patients treated with oxaliplatin. *Oncotarget*. 2015;6:16774–16785.
22. Liang JJ, Wang H, Rashid A, et al. Expression of MAP4K4 is associated with worse prognosis in patients with stage II pancreatic ductal adenocarcinoma. *Clin Cancer Res*. 2008;14:7043–7049.
23. Liu AW, Cai J, Zhao XL, et al. ShRNA-targeted MAP4K4 inhibits hepatocellular carcinoma growth. *Clin Cancer Res*. 2011;17:710–720.
24. Yamashita S, Ogawa M, Sakamoto K, Abe T, Arakawa H, Yamashita J. Elevation of serum group II phospholipase A2 levels in patients with advanced cancer. *Clin Chim Acta*. 1994;228:91–99.
25. Wang M, Hao FY, Wang JG, Xiao W. Group IIa secretory phospholipase A2 (sPLA2IIa) and progression in patients with lung cancer. *Eur Rev Med Pharmacol Sci*. 2014;18:2648–2654.
26. Cui X, Peng H, Jin J, et al. RANK overexpression as a novel esophageal cancer marker: validated immunohistochemical analysis of two different ethnicities. *Int J Clin Exp Pathol*. 2015;8:2249–2258.
27. Beuselinck B, Jean-Baptiste J, Couchy G, et al. RANK/OPG ratio of expression in primary clear-cell renal cell carcinoma is associated with bone metastasis and prognosis in patients treated with anti-VEGFR-TKIs. *Br J Cancer*. 2015;113:1313–1322.
28. Markowitz GJ, Yang P, Fu J, et al. Inflammation-dependent IL18 signaling restricts hepatocellular carcinoma growth by enhancing the accumulation and activity of tumor-infiltrating lymphocytes. *Cancer Res*. 2016;76:2394–2405.
29. Chen YY, Li CF, Yeh CH, Chang MS, Hsing CH. Interleukin-19 in breast cancer. *Clin Dev Immunol*. 2013;294–320.
30. Ciechanover A, Orian A, Schwartz AL. The ubiquitin-mediated proteolytic pathway: mode of action and clinical implications. *J Cell Biochem Suppl*. 2000;77:40–51.
31. Sánchez-Tena S, Cubillos-Rojas M, Schneider T, Rosa JL. Functional and pathological relevance of HERC family proteins: a decade later. *Cell Mol Life Sci*. 2016;73:1955–1968.
32. Tang XK, Wang KJ, Tang YK, Chen L. Effects of ubiquitin-conjugating enzyme 2C on invasion, proliferation and cell cycling of lung cancer cells. *Asian Pac J Cancer Prev*. 2014;15:3005–3009.
33. Zhang Z, Liu P, Wang J, et al. Ubiquitin-conjugating enzyme E2C regulates apoptosis-dependent tumor progression of non-small cell lung cancer via ERK pathway. *Med Oncol*. 2015;32:149.
34. Kadara H, Lacroix L, Behrens C, et al. Identification of gene signatures and molecular markers for human lung cancer prognosis using an in vitro lung carcinogenesis system. *Cancer Prev Res (Phila)*. 2009;2:702–711.
35. Snoek BC, De Wilt LH, Jansen G, Peters GJ. Role of E3 ubiquitin ligases in lung cancer. *World J Clin Oncol*. 2013;4:58–69.
36. Weber G. Enzymes of purine metabolism in cancer. *Clin Biochem*. 1983;16:57–63.
37. Brooks MD, Jackson E, Warrington NM, et al. PDE7B is a novel, prognostically significant mediator of glioblastoma growth whose expression is regulated by endothelial cells. *PLoS ONE*. 2014;9:e107397.
38. Shin J, Kim G, Lee JW, et al. Identification of ganglioside GM2 activator playing a role in cancer cell migration through proteomic analysis of breast cancer secretomes. *Cancer Sci*. 2016;107:828–835.
39. Kamata T, Jin H, Giblett S, et al. The cholesterol-binding protein NPC2 restrains recruitment of stromal macrophage-lineage cells to early-stage lung tumours. *EMBO Mol Med*. 2015;7:1119–1137.
40. Kirkegaard T, Jäättelä M. Lysosomal involvement in cell death and cancer. *Biochim Biophys Acta*. 2009;1793:746–754.
41. Breheny P, Huang J. Penalized methods for bi-level variable selection. *Stat Interface*. 2009;2:369–380.
42. Friedman J, Hastie T, Tibshirani R. A note on the group lasso and sparse group lasso. https://arxiv.org/abs/1001.0736. Published 2010.
43. Huang J, Patrick B, Ma S. A selective review of group selection in high-dimensional models. *Statist Sci*. 2012;27:481–499.
44. Li C, Li H. Network-constrained regularization and variable selection for analysis of genomic data. *Bioinformatics*. 2008;24:1175–1182.
45. Wu C, Ma S. A selective review of robust variable selection with applications in bioinformatics. *Brief Bioinform*. 2015;16:873–883.
46. Sha N, Tadesse MG, Vannucci M. Bayesian variable selection for the analysis of microarray data with censored outcomes. *Bioinformatics*. 2006;22:2262–2268.
47. Cancer Genome Atlas Research Network. Comprehensive molecular profiling of lung adenocarcinoma. *Nature*. 2014;511:543–550.