Integration analysis based on survial associated co-expression gene modules for predicing neuroblastom patients survival times

Yatong Han,[1] Xiufen Ye,[1*] Jun Cheng,[3] Siyuan Zhang,[1] Jie Zhang,[2*] Kun Huang[2*]

[1]Department of Automation, Harbin Engineering University, Harbin, China.
[2]Department of Biomedical Informatics, The Ohio State University, Columbus, Ohio 43210, USA.
[3]Guangdong Province Key Laboratory of Medical Image Processing, Southern Medical University, Guangzhou, China.
*Corresponding Authors.

Abstract

In this paper, we provide a workflow to improve survival prognosis for neuroblastom patients. With a step of gene co-expression network/module (GCN) mining in microarray and RNA-seq data, we extracted the molecular features from each module and summarized them into eigengenes. Then we adopted the lasso-regularized Cox proportional hazards model to select the most informative eigengen features in terms of association to the risk of matastasis. Nine eigengenes were selected which show strong association with patient survival prognosis. All of the nine modules also have highly enriched biological functions or cytoband locations. Three of them are unique modules to RNA-seq data, which complement the modules from microarray in terms of survival prognosis. We then merged all eigengenes from the nine modules and used an integrative method called Similarity Network Fusion to test the prognostic power of these eigengenes for prognosis. The prognostic accuracies are significantly improved as compared to use all eigengens, and two subgroups of patients with very poor survival rate were identified.

Keyword: neuroblastom patients prognosis; gene co-expression network mining; integrative cluster

1 Introduction

Neuroblastoma (NB) is one of the most common cancers in children. Those prognosed as high-risk (HR) subtype usually have poor prognostic outcome [1]. Better survival prediction with these HR patients will help doctors adjust their treatment plans, thus improve the patients chances of survival. With the abundant high-throughput transcriptomic data, a better prognosis method may benefit from an integrative approach which incorporates molecular and clinical data may extract high correlative molecular features, and identify them as potential biomarkers for patient survival prognosis. However, There are two major problems to solve for integrative approach(1) the relatively small number of samples compared to the large number of measurements; (2) complementary nature of information provide by different types of data [2]. In this paper, we provide an effective workflow to solve these problems. For complementary nature in NB transcriptomic data, a study has compared RNA-seq and Agilent microarray gene expression profiles for clinical endpoint prediction of 498 pediatric patients, and found the two technological platforms do not significantly affect performances of the models [3]. However, instead of examining the large number of genes genome-wide, which contains noise and poses a problem on the statistical power of prognosis, we reduce the data dimensionality by mining GCN first. We mined densely connected gene co-expressed modules, then summarize each module into an "eigengene" using the protocol described in [4]. To distinguish this study from another study we did on NB, which was focus on efficiently integration of the transcriptomic data and clinical data using concensus clustering, in this paper, we probed into details for these eigengenes and their biological function, and identify which of them can be used as potential biomarkers to improve statistical power for NB patient survival prognosis. Therefore, after the eigengene construction and analysis, we built a lasso-regularized Cox proportional hazards (lasso-Cox) model to compute the risk index for each patient in HR group with each eigengene, and identify the highly correlated ones [5]. Finally, we use an integrative method called Similarity Network Fusion (SNF) to merge these eigengenes and test the power of their prognostic power as potential biomarkers [2].

2 Materials and methods

2.1 Dataset and preprocessing

Dataset are obtained from Neuroblastoma Data Integration Challenge of CAMDA 2017 (http://camda.info), including RNAseq and Agilent microarray gene expression profiles for clinical endpoint prediction of 239 children patients in high risk group. RNA-seq data contains 60778 probes, microarray data contains 45198 probes, 9583 common probes were selected in both RNA-seq and mircoarray data for further analysis and data integration.

2.2 Gene co-expression analysis and summarization

We applied our recently developed weighted network mining algorithm local maximum Quasi-Clique Merging (lmQCM)[6] for GCN mining. This algorithm is a greedy approach and allows genes to be shared among multiple clusters, agreeing with the

fact genes often participate in multiple biological processes. In addition, lmQCM has been shown to find smaller co-expressed gene clusters that are often associated structural mutations such as copy number variations in cancers [6]. The adjacency (weight) matrix was constructed using Spearman Correlation Coefficient (SCC) for every pair of gene studied, as SCC can accommodate the large non-linear range of RNA-seq data better than Pearson Correlation Coefficient. Four parameters in lmQCM algorithm need initialization, they are γ, α, t, and β. Among them, γ is the most important one. It determines the initia-tion of a new cluster by setting the weight threshold for the first edge of the cluster as a sub-module. In our GCN analysis, we transform the absolute values of the Spearman correlation coefficients between a pair of expression profiles of genes into weights using a normalization procedure adopted from spectral clustering, which have been shown to be effective in previous studies [6]. Based on previous work [6], we chose γ =0.80, t = 1, $\alpha = 1$, and $\beta = 0.4$, which yielded 38 co-expressed gene clusters from mircoarray and 24 co-expressed gene clusters from RNA- seq.

2.3 Unique module analysis

We used Jaccard index less than 0.05 and Fisher exact test p value greater than 0.05 as the metrics to determine the uniqueness of a co-expression modules (Supplementary table-1). In order to evaluate the degree of correlation of genes within each module, we introduced the term Correlation Index using SCC matrix. Correlation Index (C)of a module is formulated as:

$$C = \sum_{i=1}^{N} \sum_{j=1}^{N} (W_{ij} - I_{ij})^2$$

where C is Correlation Index and W is the correlation matrix. C is computed for GCN modules from both microarray and RNA-seq. a P value is also computed for each C using the randomly selected genes for the same number 1000 times to obtain an average Correlation Index (C*). It is formulated as below:

$$P - value = \frac{C_R}{\sum_{i=1}^{1000} C^*}$$

2.4 Lasso-regularized Cox proportional hazards model for feature selection

We built a lasso-regularized Cox proportional hazards (lasso-Cox) model to compute the risk index of each patient, using the eigengenes generated from GCN. Lasso penalty (i.e. L1 penalty) generates sparsity and outputs an informative subset of features. To help select the parameters, we used a two-level cross validation (CV) strategy—first leave-one-out CV then 10-fold CV to select the best regularization parameter. Regularized Cox proportional hazards model was built on the training set using the selected parameter to compute the risk indices of all patients. After that, patients were split into low-risk and high-risk groups according to the median of risk indices of the training examples. At last, we tested if these two groups have distinct survival outcome using Kaplan-Meier estimator and log-rank test, where p less than 0.05 was consider significant. Since our initial goal is screening for all possible survival-associated features, we did not apply multiple test compensation control such as FDR. The lasso-Cox model was learned on the selected survival-associated features. Cox proportional hazards regression model was fitted, and 95% confidence intervals were computed to determine the prognostic values of our lasso-Cox risk indices and clinical stage.

2.5 Similarity Network Fusion(SNF)

We applied SNF approach to integrate five microarray eigengenes with four RNA-seq eigengenes which were shown to be highly correlated to survival by Lasso-Cox model. SNF construct networks of sample for each available data type and then fusing these into one network that represents the full spectrum of underlying data. There are three parameter in SNF: k is number of neighbors, μ is a hyperparameter, and T is number of Iterations. We setting k is 30, μ is 0.8 and T is 20.

2.6 Gene ontology enrichment analysis

The online gene ontology enrichment tool ToppGene (http://toppgene.cchmc.org) developed by Cincinnati Children's Hospital Medical Center was used for all of the module functional enrichment analysis.

3 Results

3.1 Compare Co-expression modules between microarray and RNA-seq data

After applying lmQCM, 38 co-expression modules from microarray and 24 from RNA-seq modules were identified. In order to determine if data format affects the correlation as well as modules identified, a comparison was performed between each pair of modules from microarray and RNA-seq Among them, 17 co-expression modules from microarray and 10 from RNA-seq are unique to its own data type (Supplementary Table-1), and several of them are enriched with biological processes, molecular functions or specific pathways related to cancer physiology or to neurological functions (Supplementary table-2,3). By computing the correlation indices of these unique modules, we discovered that most of the unique GCNs from the microarray data are not highly correlated in RNA-seq data (Figure 1(b)), whereas the unique GCNs in the RNA-seq data are weakly correlated in microarray data (Figure 1(a)).

3.2 Identify survival-associated eigengenes

We tested each eigengene from all of the microarray and RNA-seq GCN for the statistical significance of differentiate overall survival between low and high-risk groups by Kaplan-Meier estimator. Log-rank test results show 14 eigengenes were significantly related to prognosis (p less than 0.05). The log-rank test results of all survival-related variables are listed in Table 1,2,3.

3.3 Survival- associated feature selection using lasso-regularized Cox proportional hazard model

We further filtered for the features highly correlated to survival among the 14 eigengens features. We built a lasso-regularized Cox proportional hazard model to select the most informative features and calculate a risk index for each patient. The results show that the log-rank test p values are 1.71e-10 for nine identified features and 3.88e-5 for clinical stage respectively. Among them, five are the survival-associated eigengenes from microarrary data (M2, M7, M10, M36, and M37), and four from RNA-seq (R2, R7, R17, R21). Especially, R7, R17, R21 are from RNA-seq only modules, these modules are not present in microarray data. Most of the nine modules are highly enriched with biological functions: M2 (127 genes) and R2 (268 genes) are highly enriched with cell cycle genes (contain 39 and 64 cell cycle genes each, Bonferroni-corrected-p-values 1.05E-70 and 3.88E-78 respectively). Although they highly overlap each other, the additional 141 genes in R2 contribute more information for the prognosis with SNF in the analysis below. M10 and M37 are highly enriched with immune response genes, M7 is highly enriched with extracellular matrix organization genes (p value 3.01E-12). All of these agree with the previous pan-cancer study that the top three most common GCN in cancer are cell cycle, immune response and extracellular matrix organization genes [4]. M36 contains no enrihced molecular function or biological process, but five of the genes are co-localized on the same cytoband, which could indicate a structural variant in NB patients. R17 and R21 are enriched with RNA polymerase II transcription regulatory genes (Supplementary table-3).

3.4 Prognostic prediction based on Integrative analysis

We tested prognostic power of these selected eigengenes combined as biomarkers. This was carried out in two steps: First, we tested GCNs for prognosis from microarray and RNA-seq separately, and compared the prognosis results between above selected eigengenes with all of the eigengenes in one data type. We used spectral clustering to classify the NB patients by the five eigengenes vs by all eigengenes from microarray, and the four eigengenes vs. all eigengenes from RNA-seq modules respectively. The results shown the selected nine eigengenes can greatly improve spectral clustering results: in microarray data type, the p value is reduced from 0.00147 to 2,26e-7 (Figture 2b and Figure 2d); in RNA-seq data type, the p value is reduced from 0.0241 to 2.58e-7 (Figure 2c and Figure 2e). Second, We chose SNF to integrate above nine eigengenes. The result show that using these nine eigengenes performed better in risk prognosis than using all 62 eigengenes from both data types. The log-rank test p value is reduced to 3.84e-11 as compared to 3.46e-6 (Figure 2c and Figure 2d) .The prognosis is also better than using clinical stage (p value 3.88e-05 Figure 2a). More importantly, the prognosis using the nine eigengenes are able to further stratify the high risk patients. One additional subgroup of patients with extremely poor survival were identified. The survival rate of the worst group is less than 40% within the first 50 months (Figure 2(h))

4 Conclusion

In this study, we first compared GCNs mined from microarray and RNA-seq data. We discovered that each data format contains unique GCNs, which are enriched with clear biologic functions.. By multivariate Cox regression analysis, we identified nine survival-associated eigengenes features from microarray data (5 eigengenes) and RNA-seq data (4 eigengenes). To test the power of the combination of these nine eigengenes as prognostic biomakers, we use spectral clustering as well as SNF for survival prognosis, these nine eigengenes significantly improved the survival prognosis by several magnitude in terms of log-rank test p-value, as compared to using all of the modules, modules from one data type, as well as to the clinical stage information. This indicates that the integration of both data types provide more survival-associated information, which not only help achieve a more accurate survival prognosis, but further identifies one subgroup of patients with very poor survival among high risk patients.

References

[1] Stefano Moretti Sara Stigliani, Simona Coco. High Genomic Instability Predicts Survival in Metastatic High-Risk Neuroblastoma. *Neoplasia*, 14(9):823–832, 2012.

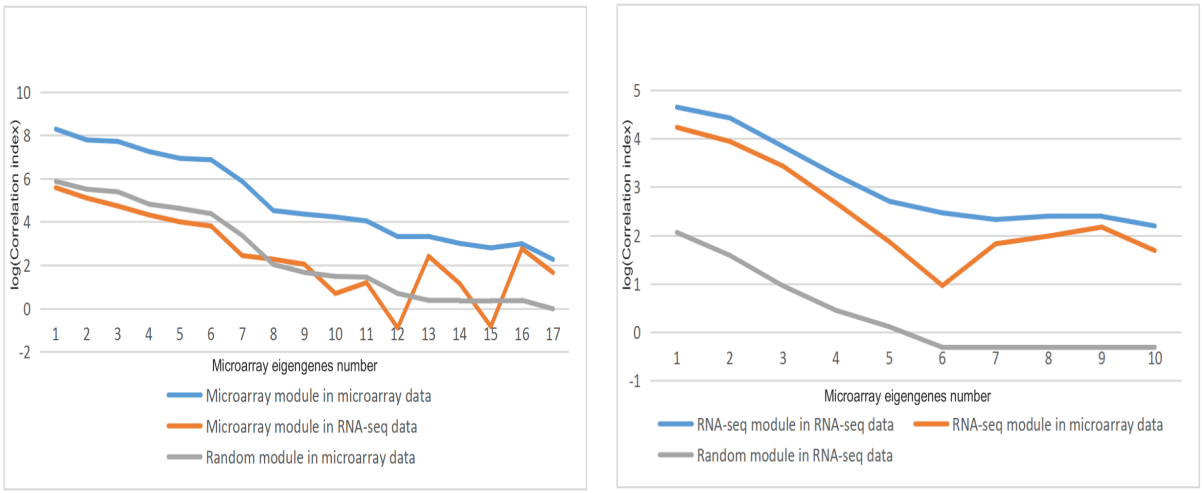
[2] Bo Wang, Aziz M Mezlini, Feyyaz Demir, Marc Fiume, Zhuowen Tu, Michael Brudno, Benjamin Haibe-Kains, and Anna Goldenberg. Similarity network fusion for aggregating data types on a genomic scale. *Nature methods*, 11(3):333–337, 2014.

[3] Falk Hertwig Jean Thierry-Mieg Wenwei Zhang etc Wenqian Zhang, Ying Yu. Comparison of RNA-seq and microarray-based models for clinical endpoint prediction. *Genome Biology*, 16(131), 2015.

[4] Jie Zhang, Kewei Lu, Yang Xiang, Muhtadi Islam, Shweta Kotian, Zeina Kais, Cindy Lee, Mansi Arora, Hui wen Liu, Jeffrey D. Parvin, and Kun Huang. Weighted Frequent Gene Co-expression Network Mining to Identify Genes Involved in Genome Stability. *PLoS Computational Biology*, 8(8), 2012.

[5] Hastie T Tibshirani R Simon N, Friedman J. Regularization paths for Cox's proportional hazards model via coordinate descent. *J Stat Softw*, 39:1–13, 2012.

[6] Huang K. Zhang J. Normalized lmQCM:an Algorithm for Detecting Weak Quasi-clique Modules in Weighted Graph with Application in Functional Gene Cluster Discovery in Cancer. *Cancer Inform*, 1(8), 2016.



(a) (b)

Figure 1: (a) Correlation index with each unique microarray module genes in microarray, RNA-seq data, and equal number random genes in microarray data. (b)Correlation index with each unique RNA-seq module genes in RNA-seq data, microarray, and equal number random genes in RNA-seq data.

	R7	R9	R13	R15	R17
P-value	0.0087	0.0101	0.0111	0.0095	0.0058
	R20	R21	R22	R23	R24
P-value	0.0039	0.0079	0.0093	0.0119	0.0081

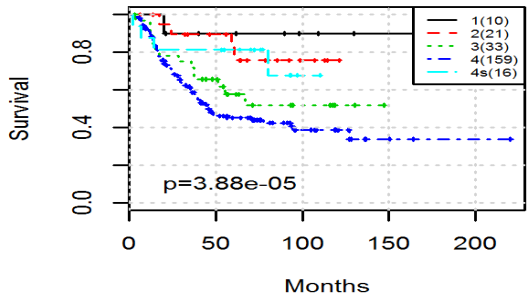
Table 1: P-value of Correlation Index with 10 unique modules in RNA-seq data

	M3	M4	M5	M8	M9
P-value	7.4766e-04	6.6568e-04	5.1788e-04	6.078e-04	5.377e-04
	M11	M13	M19	M20	M21
P-value	6.5468e-04	3.9515e-04	0.0013	0.0015	4.5433e-04
	M22	M28	M30	M31	M32
P-value	7.7567e-04	2.0277e-04	0.0077	0.0022	3.0295e-04
	M34	M38			
P-value	0.0109	0.0053			

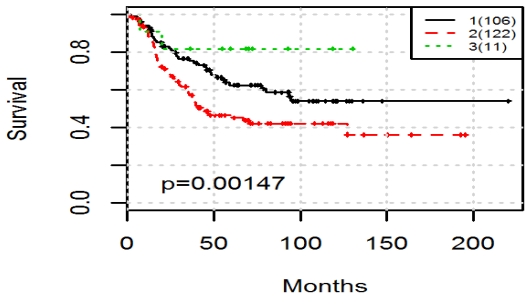
Table 2: P-value of Correlation Index with 17 unique modules in microarray data

	M2	M7	M10	M15	M30	M36	M37
P-value	5.68e-07	0.0178	0.0239	0.0142	0.0138	0.00144	0.00103
	R2	R3	R7	R8	R17	R18	R21
P-value	1.7e-08	0.00276	0.000118	0.0467	0.00314	0.0039	0.0105

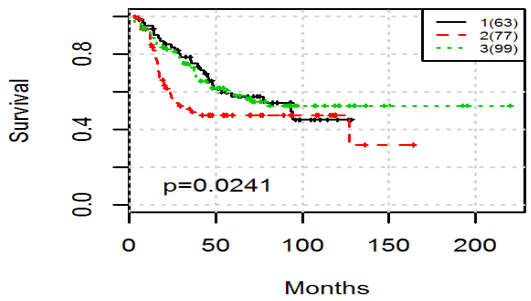
Table 3: Log-rank test of survival-associated 14 eigengens



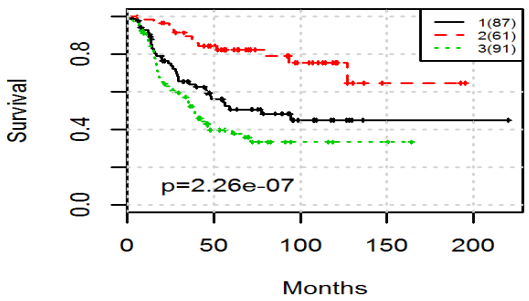
(a) Clinical Stage



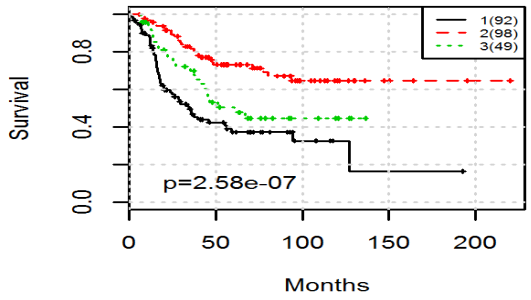
(b) Spectral clustering use all 38 Microarray eigengenes



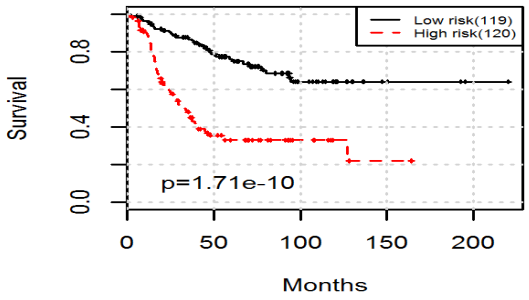
(c) Spectral clustering use all 24 RNA-seq eigengenes



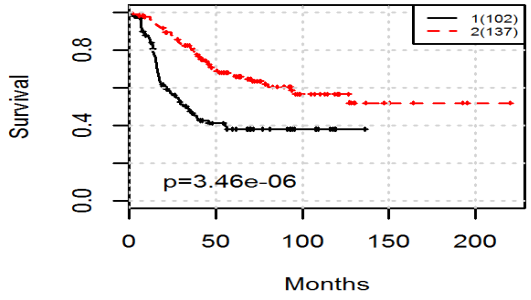
(d) Spectral clustering use 5 high survival associated microarray eigengenes



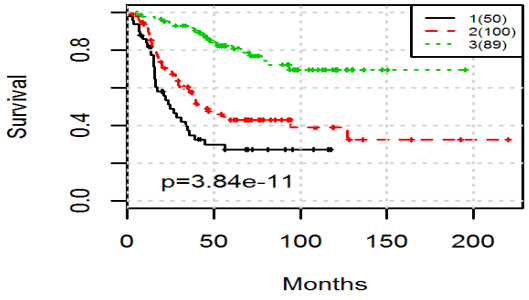
(e) Spectral clustering use 4 high survival associated RNA-seq eigengenes



(f) Lasso-Cox



(g) Similiarity Network Fusion



(h) SNF integrate 9 high survival associated eigengens

Figure 2: Compare between multiple predicting survival outcome