International Journal of Statistics and Probability; Vol. 5, No. 6; November 2016
ISSN 1927-7032 E-ISSN 1927-7040
Published by Canadian Center of Science and Education

Identification of Biomarkers for Predicting the Overall Survival of Ovarian Cancer Patients: a Sparse Group Lasso Approach

Kristi Mai[1] & Qingyang Zhang[1]

[1] Department of Mathematical Sciences, University of Arkansas, Fayetteville, AR, USA

Correspondence: Qingyang Zhang, Department of Mathematical Sciences, University of Arkansas, Fayetteville, AR 72701, USA. E-mail: qz008@uark.edu

Received: September 12, 2016 Accepted: October 11, 2016 Online Published: October 26, 2016

doi:10.5539/ijsp.v5n6p57 URL: http://dx.doi.org/10.5539/ijsp.v5n6p57

Abstract

Next-generation sequencing has been routinely applied to cancer biology, making it possible for researchers to elucidate the molecular mechanisms underlying cancer initiation and progression. However, how to identify oncomarkers from massive complex genomic data poses a great challenge for both modeling and computing. In this paper, we propose a novel computational pipeline to identify genes related to the overall survival of ovarian cancer patients from the rich Cancer Genome Atlas data. Different from the existing studies, we incorporate dependence structure among genes and pathway information into the variable selection. Firstly, the dimensionality of the ovarian cancer data is reduced by a novel stepwise feature screening which mimics the hierarchy of the underlying causal network. The second step of the pipeline is to divide genes into clusters with distinct cellular functions by k-means, x-means and PAMSAM learning algorithms. In the final step, we fit a cox proportional hazard model with a sparse group lasso penalty for further variable selection. Of the 115 genes in the final list, many were reported to be associated with cancer initiation or progression in the literature. In addition, we find several gene families including the NEK family and RNF family, which are closely associated with the survival of ovarian cancer patients.

Keywords: The Cancer Genome Atlas, ovarian cancer, k-means clustering, stepwise feature selection, sparse group lasso.

1. Introduction

Ovarian cancer is one of the most malignant gynecologic cancers, ranking fifth as the cause of cancer-related deaths among women in the United States. According to American Cancer Society, about 22, 280 women will receive a new diagnosis of ovarian cancer and about 14, 240 women will die from this disease in 2016. The latest data shows that about 70% of deaths occur in patients with high-grade serous epithelial ovarian cancer. The standard treatment for these patients is usually debulking surgery, followed by platinum-taxane chemotherapy. Platinum resistant cancer recurs within six months in about 25% of patients and the overall five-year survival rate is about 31%. Approximately 13% of high-grade serous ovarian cancer can be attributed to germline mutations in *BRCA1* and *BRCA2* and a smaller percentage can be accounted for by other germline mutations (The Cancer Genome Atlas Research Network, 2011).

With the rapid advances in high-throughput sequencing technology, it is now possible to investigate a large number of genetic and epigenetic features simultaneously (The Cancer Genome Atlas Research Network, 2011; Zhang, Burdette, & Wang, 2014; Zhang & Wang, 2016; Zhang, 2015; Kumar, Breen, & Ranganathan, 2013; Konstantinopoulos, Spentzos, & Cannistra, 2008; Popovic et al., 2014). The Cancer Genome Atlas (TCGA, http://cancergenome.nih.gov/) project provides the most comprehensive genomic data resource for more than 20 cancer types and subtypes including ovarian serous cystadenocarcinoma (OV), breast invasive carcinoma (BRCA), glioblastoma multiforme (GBM), lung adenocarcinoma (LUAD) and lung squamous cell carcinoma (LUSC). For instance, TCGA ovarian cancer data contains both clinical information and molecular profile of 568 tumor samples. The clinical information includes records on age, race, survival, outcome of debulking surgery, and treatment resistance etc. The molecular profile includes copy number variation (CNV), DNA methylation, exon expression, gene expression (both microarray and RNA-seq), genotype (SNP), MicroRNA expression (microarray), protein expression, and somatic mutation. These massive complex datasets have driven enthusiasm to elucidate molecular mechanisms of cancer through computational approaches (Zhang et al., 2014; Chen et al., 2012; Xu et al., 2012; Xi et al., 2014; Matveeva et al., 2016).

In this paper, we aim to identify prognostic genes which play crucial roles in the survival of the ovarian cancer patients. Relevant works include but are not limited to McLaughlin et al. (McLaughlin et al., 2013), Nagle, Chenevixtrench, Webb, & Spurdle (Nagle, Chenevixtrench, Webb, & Spurdle, 2007), and Konstantinopoulos et al. (Konstantinopoulos et al., 2008). However, the methods used in current studies tend to be over simplistic and inaccurate, mostly the single-round independent screening based on the t test or proportional hazard model. As pointed out by many researchers, these naive methods might result in poor selection of important features by overlooking the complex dependence structure among them. For instance, the independent test may suffer from spurious correlations in high dimensional data and fail to identify important features presented in the underlying causal network. Due to several major difficulties, the association between cancer survival and different signaling pathways has been much less studied and numerous questions remain unanswered in this field. To fill this gap, we develop an efficient and general pipeline which achieves higher accuracy in the biomarker/pathway identification. The advantage of the proposed methods is twofold. First, the feature selection is conducted in account of the dependence structure among genes, resulting in a more accurate selection of biomarkers that may directly or indirectly affect cancer survival. Second, the sparse group lasso method offers a further refinement for the candidate set of biomarkers, by encouraging genes in the same pathway to be selected and balancing gene-wise and group-wise selection in the meanwhile.

The rest of paper is organized as follows. In Section 2, we briefly summarize the TCGA ovarian cancer data and elaborate the three steps in the computational pipeline: (1) stepwise feature selection for initial screening; (2) k-means clustering along with a ”elbow method” to determine the number of clusters; (3) Cox proportional hazards model with sparse group lasso penalty for further variable selection. We present and discuss the main results from the analysis in Section 3, and conclude the article in Section 4.

2. Material and Method

2.1 Data Integration and Preprocessing

Using ”data matrix” tool on the TCGA website, we extracted the level-3 microarray data containing the expression level of 17, 814 genes in 568 tumor samples, as well as the clinical information. Table 1 summarizes our data set. The overall survival time of each patient is defined as the time between diagnosis and death. The censoring indicator was set to be 1 if death event occurred and 0 otherwise. Throughout this study, we assume that censoring mechanism is independent of survival mechanism.

Table 1. Data types, platforms and sample size in the analysis

Data type	Platform	Cases
Gene expression	Agilent 244K	572 (8 organ-specific controls)
Clinical information	N/A	583

The gene expression data were normalized using a quantile normalization method by Balstad et al. (Balstad et al., 2002) to correct the bias due to non-biological causes. We applied an existing method by Hsu et al. (Hsu et al., 2012) to remove age and batch effects (three age groups are defined as < 40 y.o., $[40, 70]$ y.o., and > 70 y.o.). This method is based on a median-matching and variance-matching strategy. For example, the batch-effect-adjusted gene expression value can be obtained as follows:

$$g^*_{ijk} = M_i + (g_{ijk} - M_{ij})\frac{\hat{\sigma}_{g_i}}{\hat{\sigma}_{g_{ij}}},$$

where g_{ijk} represents the gene expression value for gene i from batch j and sample k, M_{ij} refers to the median of $g_{ij} = (g_{ij1}, ..., g_{ijn})$, M_i refers to the median of $g_i = (g_{i1}, ..., g_{iJ})$, $\hat{\sigma}_{g_i}$ and $\hat{\sigma}_{g_{ij}}$ are the sample standard deviation of g_i and g_{ij}, respectively.

2.2 Stepwise Feature Selection

A necessary and crucial step for genome-wide association study is feature screening, i.e., to filter out irrelevant or redundant features. A refined variable set helps improve computing efficiency and estimation accuracy (Zhang et al., 2014). Existing feature selection methods can be classified into either wrapper approach (Kohavi & John, 1997; Leng, Valli, & Armstrong, 2010) or filter approach (Haindl, Somol, Ververidis, & Kotropoulos, 1999; Jouve & Nicoloyannis, 2010). The filter approach using independent test for two conditions is more commonly used due to its efficiency and simplicity. However, it tends to filter out many related features in high-dimensional settings. To this end, Zhang et al. (Zhang et al., 2014) proposed a novel stepwise correlation-based selector (SCBS) to select features from TCGA data for further Bayesian network inference. Assume there is a causal chain X→Y→cancer. Though X to Y or Y to cancer has directed association, the association between X and cancer could greatly decay so that it cannot be detected by independent test. The SCBS procedure starts with detection of features strongly correlated with the phenotype and then progressively selects features that correlate with features selected in previous step. This procedure is a natural mimic of sparse network structure and is capable of identifying nodes that are indirectly associated with the phenotype. In practice, the method can be implemented as follows:

- Step 1: Calculate the Spearman’s correlation coefficients between the current variable X_i and all the other variables, denoted by ρ_{ij}, $j \neq i$. Keep k most correlated variables with X_i based on ρ_{ij} for further filtering.

- Step 2: Calculate the p-value of correlation coefficient for each of the k variables from step 1, select the variable if the p-value is significant under Benjamini-Hochberg (BH) procedure with FDR≤ 0.05.
- Step 3: Repeat step 1 and 2 until p variables are selected.

In practice, the total number of selected variables p is subject to the scale of the model to build. For the TCGA data, we run the SCBS for 4 rounds, in order to select more than 500 but less than 1000 genes. The computing time of SCBS is sublinear to p. The choice of k is essential for SCBS which partially depends on the network density. Based on an extensive simulation study, a k of 4 or 5 is recommended by Zhang et al. (Zhang et al., 2014) to attain moderate complexity or sparsity of the model. We set $k = 4$ in our analysis and obtained a set of 603 genes.

2.3 K-means Clustering

The unsupervised k-means clustering is applied to cluster the 603 selected genes based on a correlation metric defined as follows:

$$||\mathbf{g}_i, \mathbf{g}_j||_\rho = 1 - |\rho_{\mathbf{rg}_i,\mathbf{rg}_j}|,$$

where $\mathbf{g}_i$ and $\mathbf{g}_j$ represent expression level of gene i and gene j, $i, j = 1, 2, ..., p$, $\mathbf{g}_i$ is n-dimensional vector where n is number of samples. The Spearman's correlation between gene i and gene j is denoted by $\rho_{\mathbf{rg}_i,\mathbf{rg}_j}$, where $\mathbf{rg}_i$ and $\mathbf{rg}_j$ represent the ranks of $\mathbf{g}_i$ and $\mathbf{g}_j$. An immediate consequence by this definition is $0 \leq ||\mathbf{g}_i, \mathbf{g}_j||_\rho \leq 1$, where the two equalities hold when $\rho_{\mathbf{rg}_i,\mathbf{rg}_j} = 0$ and $|\rho_{\mathbf{rg}_i,\mathbf{rg}_j}| = 1$, respectively. The k-means clustering algorithm aims to partition p variables into K clusters $\mathbf{C} = (C_1, C_2, ..., C_K)$, where K is a predefined number of clusters. Its objective is to find:

$$\arg\min_{\mathbf{C}} \sum_{k=1}^{K} \sum_{\mathbf{g} \in C_k} ||\mathbf{g} - \boldsymbol{\mu}_k||_\rho.$$

An "elbow method" was used for the choice of optimal number of clusters. Figure 1a shows the percentage of variance explained by the clusters against the number of clusters. At K=4 or 5, the marginal gain began to drop substantially, giving an angle in the graph. The number of clusters was chosen at the "elbow" K=4. The multi-dimensional Scaling (MDS) plot is shown in Figure 1b where clusters were highlighted by different colors. The x-means clustering (Pelleg & Moore, 2000), an alternative and variation of k-means, and PAMSAM algorithm were also applied to our data set. However, in terms of clustering, three methods did not give a significant difference.

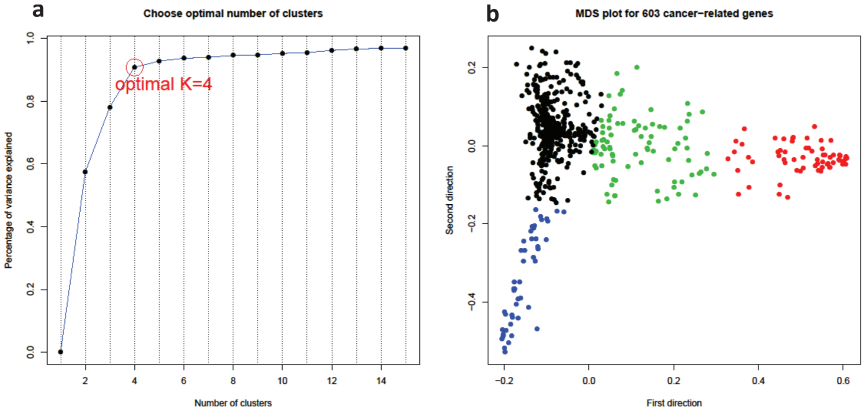

Figure 1. Four gene clusters. (a) The proportion of variance that can be explained by clustering (y-axis) against the number of clusters (x-axis) based on different values of k (k=1,2,...,15) by k-means clustering method. From this plot, the most likely number of clusters is four. (b) Multidimensional scaling (MDS) plots based on correlation dissimilarity metric among 603 genes, where genes in different clusters were highlighted by different colors.

2.4 Cox Proportional Hazard Model with Sparse Group Lasso Penalty

The last step of the pipeline conducts pathway level selection of prognostic genes. A natural way is to fit a regression model with group lasso penalty where each group represents a pathway. The group lasso, however, only works for large number of groups and gives a sparse set of groups. We therefore turned to a sparse group lasso (SGL), which generates a solution balancing both between-group and within-group sparsity. A Cox proportional hazards model and sparse group lasso regularization were then pieced together for further variable selection.

Let p be the number of genes, $\mathbf{X} = (\mathbf{X}_1, \mathbf{X}_2, ..., \mathbf{X}_n)^T$ be the $n \times p$ data matrix, where $\mathbf{X}_i = (X_{i1}, X_{i2}, ..., X_{ip})^T$. Let $\mathbf{Y} = (Y_1, Y_2, ..., Y_n)^T$ denote an n-dimensional vector which corresponds to failure/censor times. Let $\boldsymbol{\beta} = (\beta_1, \beta_2, ..., \beta_p)^T$ be the vector of coefficients, and $\boldsymbol{\delta} = (\delta_1, \delta_2, ..., \delta_n)$ be the censoring indices, where $\delta_i = 1$ indicates event (death) occurred for subject i and $\delta_i = 0$ indicates censoring. The Cox proportional hazards model can be written as follows:

$$\log \frac{\lambda(t|\mathbf{X}_i)}{\lambda_0(t)} = \mathbf{X}_i^T \boldsymbol{\beta}.$$

where $\lambda_0(t)$ stands for the baseline hazard function. The loglikelihood can be written as follows:

$$\ell(\boldsymbol{\beta}) = \frac{1}{n}\{\log(\sum_{i:\delta_i=1}(\sum_{j:Y_j \geq Y_i} \exp(\mathbf{X}_j^T \boldsymbol{\beta}) - \mathbf{X}_j^T \boldsymbol{\beta}))\}$$

With the assumption of sparsity, the parameters $\boldsymbol{\beta}$ can be estimated through a SGL penalized likelihood:

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \frac{1}{n}\{\log(\sum_{i:\delta_i=1}(\sum_{j:Y_j \geq Y_i} \exp(\mathbf{X}_j^T \boldsymbol{\beta}) - \mathbf{X}_j^T \boldsymbol{\beta}))\} + (1-\alpha)\lambda \sum_{k=1}^{K} \sqrt{p_k} \|\boldsymbol{\beta}^k\|_2 + \alpha\lambda\|\boldsymbol{\beta}\|_1,$$

where $\|\cdot\|_1$ and $\|\cdot\|_2$ denote ℓ_1-norm and ℓ_2-norm respectively, and p_k represents the size of group k and $\boldsymbol{\beta}^k$ represents the coefficients of genes in group k. The SGL fit is simply a combination of the lasso and group lasso penalties ($\alpha = 0$ gives the group lasso fit, $\alpha = 1$ gives the lasso fit). In practice, one should choose α before the parameter estimation. In our problem, we expect a strong overall sparsity but encourage grouping, therefore a $\alpha = 0.8$ was used. Here the choice of α is different from the choice of λ, which can be determined by data-driven method. In practice, the mixing rate α need to be predefined depending on the expected overall sparsity and group sparsity. Given two tuning parameters α and λ, a routine blockwise coordinate descent (BCD) approach can solve the optimization problem and we implemented the BCD algorithm using R package *SGL* (Simon, Friedman, Hastie, & Tibshirani, 2013). A sequence of ten candidate λ's with $\lambda_{\min} = 0.05\lambda_{\max}$ in the regularization path was used, as suggested by R package *SGL.*

In the lasso-type problems, the common method for selecting the tuning parameter λ is cross-validation. However, it tends to yields a large number of false positives in the sparse network problem, as pointed out by Fu and Zhou in their seminal paper (Fu & Zhou, 2013). Fu and Zhou proposed an "elbow method" that outperforms the cross-validation method, where the optimal tuning parameter corresponds to the change point at which an increase of λ does not yield a substantial decrease of log-likelihood. In our Cox model with SGL regularization, the optimal lambda selected by this rule is $\lambda = 0.000492$ as shown in Figure 2 and 115 genes were identified in the final list.

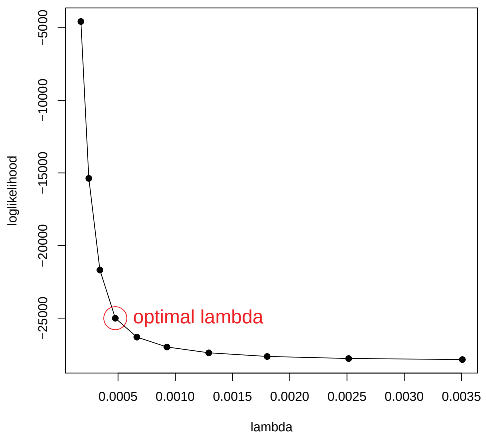

Figure 2. Selection of tuning parameter for the penalty term is sparse group lasso regression. The log-likelihood (y-axis) against tuning parameter (x-axis) for the sparse group lasso penalty, where the optimal λ is circled.

3. Results and Discussion

3.1 Gene Clusters

Using the k-means approach, the set of 603 genes from initial screening were further clustered into four subgroups. Gene functions in each cluster were investigated. Interestingly, we found that genes within the same cluster tend to have similar/related cellular functions. For instance, cluster 1 (black dots in Figure 1b), the core cluster containing 407 genes including *CENPJ* and *CDK5RAP2*, is functionally related to cell cycle, spindle formation, and mitosis etc. Cluster 2 (blue dots in Figure 1b) contains 48 genes including *MYOG* and *CDK5R2*, mostly related to protein binding and transmembrane activity etc. Cluster 3 (green dots in Figure 1b), containing 85 genes including *COL5A2* and *COL8A2*, corresponds to the pathways related to collagen biosynthesis and enzymes modification etc. Cluster 4 (red dots in Figure 1b), containing 63 genes including *CD48* and *CD53*, is related to immune response and T-cell and B-cell development. This finding indicates that certain cellular pathways/functions may play crucial roles in the progression of the serous ovarian cancer, which may provide new clues for the cancer prevention and treatment.

Table 2. List of 115 identified prognostic genes and corresponding coefficients in the Cox model.

Gene	Group	$\hat{\beta}$	Gene	Group	$\hat{\beta}$	Gene	Group	$\hat{\beta}$
ADCK1	1	-0.150	LOC389458	1	-0.139	TIPRL	1	0.118
ADH4	1	0.139	MAP9	1	0.255	TOP2B	1	0.162
ADORA2A	1	0.107	MBL2	1	-0.096	TTBK2	1	-0.094
ARHGEF12	1	-0.134	MGC27348	1	0.140	VAMP4	1	0.201
ATP4B	1	-0.188	MRGPRX4	1	-0.227	VPS29	1	0.094
BOLA3	1	0.111	MRPS22	1	0.199	WDFY3	1	0.120
C1orf75	1	0.107	NARF	1	0.204	WTAP	1	-0.213
C4orf27	1	0.127	NDUFB4	1	-0.165	YSK4	1	0.129
C6orf115	1	0.090	NEK1	1	0.175	YY1AP1	1	-0.082
CACNA1S	1	-0.084	NEK2	1	-0.091	ZFHX2	1	-0.094
CDK5RAP2	1	-0.096	NEK9	1	-0.107	ZNF167	1	-0.080
CNGB1	1	-0.124	NLRX1	1	0.090	ZNF197	1	-0.164
COX17	1	0.109	NRAS	1	-0.136	ZNF621	1	-0.117
CPA2	1	-0.085	NSL1	1	0.119	ZNF782	1	-0.146
CRIPT	1	0.120	OR7D4	1	0.123	CTRB2	2	-0.103
CRY1	1	0.170	OR9Q1	1	0.176	DRD3	2	-0.144
DEDD2	1	-0.133	OS9	1	0.153	LCE3A	2	-0.092
DLG3	1	-0.117	PALB2	1	-0.169	LMAN1L	2	-0.105
DNAH7	1	-0.087	PLEKHH1	1	0.191	OR2G2	2	0.102
DNAI1	1	-0.118	POMP	1	-0.094	TAAR8	2	0.208
DNAJC19	1	0.164	PPP2R2B	1	0.128	CLEC4A	3	0.105
DNAJC5	1	0.087	RNF12	1	-0.104	CTSS	3	0.086
DNASE1	1	-0.105	RNF181	1	0.093	EBI3	3	0.099
ELA2A	1	0.154	RNF20	1	0.175	LY86	3	0.080
EPB41	1	-0.082	RNF31	1	-0.144	RNASE6	3	0.096
EWSR1	1	-0.197	RNF7	1	0.248	SRGN	3	0.099
EXOSC8	1	0.152	RPL21	1	-0.100	ABCG5	4	0.137
FAM86B1	1	-0.083	SCYL1BP1	1	0.178	AP1B1	4	-0.224
GAS2L2	1	-0.231	SEBOX	1	0.093	CD248	4	-0.080
GHRH	1	-0.108	SEC22B	1	0.083	CIDEA	4	-0.131
GRM6	1	-0.090	SFT2D1	1	0.087	COL8A2	4	-0.109
GSTA3	1	0.192	SLC8A2	1	0.168	FABP4	4	-0.098
HEXDC	1	-0.097	SNRPG	1	-0.143	GPBAR1	4	-0.137
HMOX2	1	-0.127	SPN	1	0.228	GRN	4	-0.087
JUB	1	-0.095	SRrp35	1	0.135	OAS1	4	0.082
KIAA0323	1	-0.108	STAT2	1	0.124	OASL	4	0.091
KIF27	1	-0.099	TBPL1	1	0.094	TIMP4	4	0.201
KIF4B	1	-0.097	KLHL22	1	-0.100	ZNF660	4	-0.120

3.2 Prognostic Gene Identification

Using the Cox model with sparse group lasso penalty, we obtain a final list of 115 prognostic genes (in Table 2), of which many were reported to be involved in cancer initiation and progression. To name a few, gene *CTSS* is closely related to gastric cancer and silencing *CTSS* expression suppressed the migration and invasion of gastric cancer cells (Yang et al., 2010). Gene *CD248* can facilitate tumor growth via its cytoplasmic domain and multiple pathways regulated by the cytoplasmic domain of *CD248* highlight its potential as a therapeutic target to treat cancer (Maia et al., 2011). Gene

DRD3 is a dopamine receptor, whose expression can change as stress factors associated with breast cancer (Pornour et al., 2014). Gene *CDK5RAP2* is required for spindle checkpoint function and is a common target in paclitaxel and doxorubicin. Cancer cells cultured in the presence of paclitaxel or doxorubicin exhibit a dramatic decrease in *CDK5RAP2* levels (Zhang et al., 2009).

We also identify several subgroups (families) of genes whose associations with cancer have been reported. For instance, three genes in the NEK family, *NEK1*, *NEK2* and *NEK9*, were identified in our final list. Mutations of NEK family members have also been identified as drivers behind the development of ciliopathies and cancer. Recent emergence of comprehensive cancer genomes is highlighting certain members of the NEK family as targets of frequent mutations (Moniz, Dutt, Haider, & Stambolic, 2011). We also identified five genes in the RNF family: *RNF12, RNF181, RNF20, RNF31,RNF7*. A recent study reported that the RNF family such as *RNF20* drives histone *H2B* monoubiquitylation and modulates inflammation and inflammation-associated cancer in mice and humans (Tarcic et al., 2016).

The predictive power of our 115-gene signature was illustrated using Kaplan-Meier curves in Figure 3, where the samples were equally divided into two groups based on the hazard risk. The moderate separation of two groups demonstrates the effectiveness of our method.

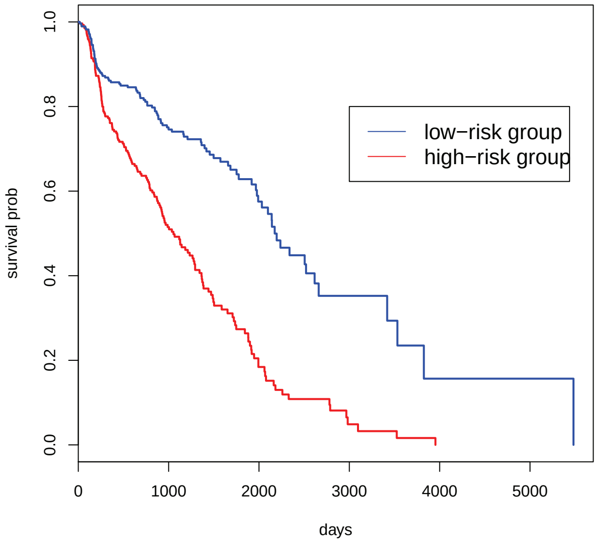

Figure 3. Kaplan-Meier curves. Survival probability against time (in days) of different groups by the hazard risk based on the 115-gene signature. The red and blue lines are based on high-risk group and low-risk group, respectively.

4. Conclusion

In this paper, we developed a flexible three-step computational pipeline for identifying prognostic biomarkers related to the overall survival of serous ovarian cancer patients using the rich TCGA data set. This pipeline facilitates the pathway level analysis of the biomarkers associated with cancer survival. The proposed methods are computationally efficient and can be generally applied to many large-scale genomic cancer data sets including the TCGA data. We applied this pipeline to TCGA ovarian cancer data and identified a list of 115 genes, as well as several gene families including the NEK family and RNF family, which may greatly affect the overall survival of ovarian cancer patients. Some of these findings are well supported by literature.

Acknowledgements

Support has been provided in part by the Arkansas Biosciences Institute, the major research component of the Arkansas Tobacco Settlement Proceeds Act of 2000.

References

Bolstad, B., Irizarry, R., Astrand, M., & Speed, T. (2002). A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics, 19*(2). http://dx.doi.org/10.1093/bioinformatics/btg202

Chen, L., Xuan, J., Gu, J., Wang, Y., Zhang, Z., Wang, T., & Shih, L. (2012). Integrative network analysis to identify aberrant pathway networks in ovarian cancer. *Pacific Symposium Biocomputing, 31*. http://dx.doi.org/10.1142/978981436

6496-0004

Fu, F., & Zhou, Q. (2013). Learning Sparse Causal Gaussian Networks With Experimental Intervention: Regularization and Coordinate Descent. *Journal of American Statistical Association, 108*(501), 288-300. http://dx.doi.org/10.1080/01621459.2012.754359

Friedman, N., Linial, M., Nachman, I., & Pe'er, D. (2000). Using Bayesian networks to analyze expression data. *Journal of Computational Biology, 7* (3), 601-20. http://dx.doi.org/10.1145/332306.332355

Haindl, M., Somol, P., Ververidis, D., & Kotropoulos, C. (1999). Feature Selection Based on Mutual Correlation. Technical Report

Hsu, F., Serpedin, E., Hsiao, T., Bishop, A., Dougherty, E., & Chen, Y. (2012). Reducing confounding and suppression effects in TCGA data: an integrated analysis of chemotherapy response in ovarian cancer. *BMC Genomics, 13*. http://dx.doi.org/10.1186/1471-2164-13-s6-s13

Jouve, P. E., & Nicoloyannis, N. (2010). A Filter Feature Selection. Technical Report

Kohavi, R., & John, G. H. (1997). Wrappers for feature subset selection. *Artificial Intelligence, 97*, 273-324.

Kumar, G., Breen, E. J., & Ranganathan, S. (2013). Identification of ovarian cancer associated genes using an integrated approach in a Boolean framework. *BMC System Biology, 7* (12). http://dx.doi.org/10.1186/1752-0509-7-12

Konstantinopoulos, P. A., Spentzos, D., & Cannistra, S. A. (2008). Gene-expression profiling in epithelial ovarian cancer. *Nature Clinical Practice Oncology, 5*, 577-87. http://dx.doi.org/10.2174/138920206777304641

Leng, J., Valli, C., & Armstrong, L. (2010). A Wrapper-based Feature Selection for Analysis of Large Data. Technical Report

Moniz, L., Dutt, P., Haider, N., & Stambolic, V. (2011). Nek family of kinases in cell cycle, checkpoint control and cancer. *Cell Division, 6*(18). http://dx.doi.org/10.1186/1747-1028-6-18

Matveeva, E., Maiorano, J., Zhang, Q., Wang, J. P., & Fondufe-Mittendorf, Y. (2016). Involvement of PARP1 in the regulation of alternative splicing. *Cell Discovery, 2*(15046). http://dx.doi.org/10.1038/celldisc.2015.46

McLaughlin, J., Rosen, B., Moody, J., Pal, T., Fan, I., Shaw, R., Narod, S. (2013). Long-term ovarian cancer survival associated with mutation in BRCA1 or BRCA2. *Journal of National Cancer Institute, 105*(2). http://dx.doi.org/10.1093/jnci/djs494

Nagle, C., ChenevixTrench, G., Webb, P., & Spurdle, A. (2007). Ovarian cancer survival and polymorphisms in hormone and DNA repair pathway genes. *Cancer Letters, 251*(1). http://dx.doi.org/10.1016/j.canlet.2006.11.011

Pelleg, D., & Moore, A. (2000). X-means: Extending K-means with Efficient Estimation of the Number of Clusters. *ICML Proceedings of the Seventeenth International Conference on Machine Learning*. http://dx.doi.org/10.1007/3-540-44491-2-3

Popovic, R., Martinez-Garcia, E., Giannopoulou, E. G., Zhang, Q., Zhang, Q., Ezponda, T., & Licht, J. (2014). Histone Methyltransferase MMSET/NSD2 Alters EZH2 Binding and Reprograms the Myeloma Epigenome through Global and Focal Changes in H3K36 and H3K27 Methylation. *PLoS Genetics, 10*(9). http://dx.doi.org/10.1371/journal.pgen.1004566

Pornour, M., Ahangari, G., Hejazi, S., Ahmadkhaniha, H., & Akbari, M. (2014). Dopamine receptor gene (DRD1-DRD5) expression changes as stress factors associated with breast cancer. *Asian Pacific Journal of Cancer Prevention, 15*(23). http://dx.doi.org/10.7314/apjcp.2014.15.23.10339

Simon, N., Friedman, J., Hastie, T., & Tibshirani, R. (2013). A Sparse-Group Lasso. *Journal of Computational and Graphical Statistics, 22*. http://dx.doi.org/10.1080/10618600.2012.681250

Tarcic, O., Paterals, I., Cooks, T., Shema, E., Kanterman, J., & Oren, M. (2016). RNF20 Links Histone H2B Ubiquitylation with Inflammation and Inflammation-Associated Cancer. *Cell Reports, 14*(6). http://dx.doi.org/10.1016/j.celrep.2016.01.020

The Cancer Genome Atlas Research Network. (2011). Integrated genomic analyses of ovarian carcinoma. *Nature, 474*. http://dx.doi.org/10.1158/2159-8290.cd-rw042711-14

Xu, Y., Zhang, J., Yuan, Y., Mitra, R., Muller, P., & Ji, Y. (2012). A Bayesian graphical model for integrative analysis of TCGA data. *2012 IEEE International Workshop on Genomic Signal Processing and Statistics, 31*. http://dx.doi.org/10.1017/cbo9781107706484.010

Xi, L., Brogaard, K., Zhang, Q., Lindsay, B., Widom, J., & Wang, J. P. (2014). A locally convoluted cluster model for nucleosome positioning signals in chemical maps. *Journal of the American Statistical Association, 109*(505). http://dx.doi.org/10.1080/01621459.2013.862169

Yang, Y., Lim, S., Choong, L., Lee, H., Chen, Y., Chong, P., & Lim, Y. (2010). Cathepsin S mediates gastric cancer cell migration and invasion via a putative network of metastasis-associated proteins. *Journal of Proteome Research, 9* (9). http://dx.doi.org/10.1021/pr100492x

Zhang, X., Liu, D., Lv, S., Wang, H., Zhong, X., Liu, B., & Xu, X. (2009). CDK5RAP2 is required for spindle checkpoint function. *Cell Cycle, 8*(8) http://dx.doi.org/10.4161/cc.8.8.8205

Zhang, Q., Burdette, J. E., & Wang, J. P. (2014). Integrative network analysis of TCGA data for ovarian cancer. *BMC Systems Biology, 8* (1338), 1-18. http://dx.doi.org/10.1186/s12918-014-0136-9

Zhang, Q., & Wang, J. P. (2016). A Bayesian network approach for modeling mixed features in TCGA ovarian cancer data. *Handbook of Mathematical Methods in Cancer Biology, 1.*

Zhang, Q. (2015). Learning Sparse Bayesian Network with Mixed Variables and its Application to Cancer Systems Biology. Unpublished PhD Dissertation, Northwestern University.

Copyrights

Copyright for this article is retained by the author(s), with first publication rights granted to the journal.

This is an open-access article distributed under the terms and conditions of the Creative Commons Attribution license (http://creativecommons.org/licenses/by/4.0/).