G&I Genomics & Informatics

eISSN 2234-0742
Genomics Inform 2018;16(4):e32
https://doi.org/10.5808/GI.2018.16.4.e32

ORIGINAL ARTICLE

Ovarian Cancer Prognostic Prediction Model Using RNA Sequencing Data

Seokho Jeong[1§], Lydia Mok[2§], Se Ik Kim[3], TaeJin Ahn[4], Yong-Sang Song[3], Taesung Park[1,2*]

[1]Department of Statistics, Seoul National University, Seoul 08826, Korea, [2]Interdisciplinary Program in Bioinformatics, Seoul National University, Seoul 08826, Korea, [3]Department of Obstetrics and Gynecology, Seoul National University College of Medicine, Seoul 03080, Korea, [4]Department of Life Science, Handong Global University, Pohang 37554, Korea

Ovarian cancer is one of the leading causes of cancer-related deaths in gynecological malignancies. Over 70% of ovarian cancer cases are high-grade serous ovarian cancers and have high death rates due to their resistance to chemotherapy. Despite advances in surgical and pharmaceutical therapies, overall survival rates are not good, and making an accurate prediction of the prognosis is not easy because of the highly heterogeneous nature of ovarian cancer. To improve the patient's prognosis through proper treatment, we present a prognostic prediction model by integrating high-dimensional RNA sequencing data with their clinical data through the following steps: gene filtration, pre-screening, gene marker selection, integrated study of selected gene markers and prediction model building. These steps of the prognostic prediction model can be applied to other types of cancer besides ovarian cancer.

Keywords: ovarian neoplasms, penalized Cox regression, prediction model, RNA sequencing data

Introduction

The genetic background of patients with complex diseases, such as cancer, has been continually studied. Traditional attempts mainly focus on finding unique differentially expressed genes (DEGs) for diagnosis or survival times using microarray techniques [1, 2]. Genetic markers can be indicators of the activity state of a pathway of therapy, showing their potential as prognostic predictors for specific treatments. Ovarian cancer, especially high-grade serous ovarian cancer (HGSC), is one of the most lethal gynecological malignancies in women. Vague symptoms and a lack of robust biomarkers for detection are the main causes of difficulties in making an early diagnosis. Major factors associated with the poor prognosis and poor management of ovarian cancers include the late discovery of the disease, chemotherapy resistance, and a lack of clinical variables that are crucial for accurate prognostic predictions [3]. Thus, there is a need to identify new biomarkers that can be used to improve the treatment of ovarian cancer patients [4].

Previous studies have mainly focused on identifying novel molecular markers or subtyping through molecular markers of HGSC [5,6]. In this study, we aim to identify clinical and genetic markers of the prognosis of HGSC. Through analyzing high-dimensional genetic information, we can gain a better understanding of HGSC and its biological mechanism. For cancers with a poor prognosis, genetic information can be effective in improving the prognosis of patients based only on clinical information [3].

Next-generation sequencing (NGS) technology has made it possible to generate mRNA expression data for the tens of thousands of genes from only a few hundred samples [7]. These RNA sequencing (RNA-Seq) data have the advantage of knowing the sequence count directly, without being affected by the background noise, unlike microarray data. RNA-Seq data also contain more genes to discover than microarray data. High-throughput NGS-based tumor genome profiling has significantly advanced our understanding of the molecular variability exhibited by tumors.

However, the nature of RNA-Seq data, consisting of non-negative count data, requires a new predictive model

Received December 10, 2018; Accepted December 16, 2018; Published online December 28, 2018

***Corresponding author:** Tel: +82-2-880-9168, Fax: +82-2-880-6693, E-mail: tspark@stats.snu.ac.kr

[§]These authors contributed equally to this work as co-first authors.

that differs from that of microarray data. For RNA-Seq data, it is difficult to apply a traditional statistical approach because of its relatively small number of samples compared with the large number of genes. Therefore, we present a process for finding genes that play important roles in the outcome and specific procedures for designing and evaluating a model, based on the characteristics of RNA-Seq data.

The purpose of this study is to find gene markers that can significantly affect the prognosis, when analyzing RNA-Seq data, and to present a protocol that can improve the predictive performance using transcription data by integrating clinical information. Specifically, we present a protocol to build a prediction model for the prognosis of HGSC, including (1) selection of markers and clinical variables and (2) model construction and evaluation.

Methods

The data used in this study were clinical information and RNA-Seq data from The Cancer Genome Atlas (TCGA) Genomic Data Commons (GDC) portal (https://portal.gdc.cancer.gov). The National Institutes of Health (NIH)'s TCGA webpage provides clinical information, biomedical slide images, molecular information, CpG methylation information, DNA copy number, mRNA expression, and miRNA expression. RNA-Seq data were used, because large amounts of gene expression could be explored collectively from samples in the study. We downloaded genomic alteration data and expression data in patients with high-grade serous ovarian cancer and the corresponding clinicopathological profiles from the Genomic Data Commons (https://portal.gdc.cancer.gov), Firebrowse (http://firebrowse.org), and cBioPortal for Cancer Genomics (http://www.cbioportal.org) web portals. This study complied with TCGA publication guidelines and policies (http://cancergenome.nih.gov/publications/publicationguidelines). The Institutional Review Board of Seoul National University Hospital ruled that no formal ethical approval was required in this study.

Survival time refers to the period from the diagnosis of HGSC in patients to death. Clinical variables were selected, based on their influence on the outcome. Among 365 HGSC cases, the median survival time for the samples was 43.9 months, and the censoring proportion of the data was 55%.

Pre-processing

HGSC patient data collected from the GDC portal can be separated into two categories. Since RNA-seq data from the GDC portal consist of raw count data, normalization was required to control for the sample bias and gene length bias. Two common normalization methods were applied. The first method is called relative long expression (RLE) normalization, implemented in the R package "Deseq2" [8], which is also implemented in the R package "edgeR" [9]. The RLE method utilizes the geometric mean of the read count, whereas the TMM method estimates normalizing factors after extreme expressions are removed to get more a robust estimation. A simple simulation study for DEGs has been performed for comparison [10]. Both methods were used to give more specific comparison results in the model building.

Since the clinical data had many missing values, we imputed missing clinical variables with the R package "mice," which performs chained equations to find estimates for missing values using Gipps sampling [11]. There were no tendencies in the missing values; so, the missing-at-random assumption was applied [11]. After imputation, significant clinical variables were chosen using a Cox regression model via stepwise selection methods.

Clinical data and RNA-Seq data were integrated and divided into training and test sets. To avoid unwanted effects of censoring, balanced sampling in terms of censoring status was used to make the training and test sets have the same censoring rate.

Analysis plan

In this study, we considered a prediction model and process in terms of performance and interpretability. The main steps of our analysis were as follows: (1) gene filtration, (2) pre-screening, (3) gene marker selection, and (4) integrated study of selected gene markers and the final prediction model.

Gene filtration

Gene filtration for low expression is important because it can reduce false positives [12]. Sampling noise that is added during the sequencing process can also be removed by gene filtration. In this study, a 20% zero proportion threshold was selected, as this value is commonly chosen in previous studies [13]. Gene filtration was performed only on the training set.

Pre-screening

Pre-screening of gene markers was used after gene filtration. This process is necessary for actual computation and controlling false positives. Unsupervised filtering using median absolute deviation (MAD) was used. MAD is widely used in microarray data and single-cell RNA-Seq data as a pre-screening method. This measure can be used to select more variable genes. Compared with standard deviation, MAD is a more robust measure [14]. Since the raw count data consist of sparse and non-negative numbers, MAD was

used to represent variability. The genes with MAD were selected for model building.

Another supervised pre-screening method was used, based on univariate Cox regression with adjustment of clinical variables. This Cox regression procedure was performed only for the training set to avoid overfitting. Next, genes were selected, based on individual p-values. Both methods were jointly applied to select candidate gene markers.

Gene selection

Although the pre-screening process greatly reduced the number of genes, there remained the problem of an ill-posed model, since the sample size was very small compared to the number of features. Therefore, sequential variable selection using a multiple Cox regression model was not applicable. To solve such a problem, as presented below, multiple Cox regression using penalized partial likelihood was used for variable selection and shrinkage estimation.

$$\hat{\beta} = \text{argmin}_{\beta} \frac{1}{n} \left[\sum_{s=1}^{S} - X_{i_s}^t + \log\left\{ \sum_{i \in R_s} \exp(X_{i_s}^t \beta) \right\} \right] + P_{\alpha,\lambda}(\beta),$$

where $P_{\alpha,\lambda}(\beta) = \sum_{c=1}^{C} \lambda(\delta|\beta_c| + 0.5(1-\delta)\beta_c^2) + \sum_{g=1}^{G} \lambda(\delta|\beta_g| + 0.5(1-\delta)\beta_g^2)$ is the penalty function. Note that C and G are the indicator sets of clinical and genetic variables, respectively.

Modified partial likelihood was also used to select clinical variables more favorably, since some clinical variables have already been reported to be significant.

$$P_{\gamma,\alpha,\lambda}(\beta) = \sum_{c=1}^{C} \lambda\gamma(\delta|\beta_c| + 0.5(1-\delta)\beta_c^2) + \sum_{g=1}^{|G|} \lambda(\delta|\beta_g| + 0.5(1-\delta)\beta_g^2)$$

Thus, the penalty factor γ is used to reduce the penalty of clinical variables.

The partial likelihood estimation process explained above is implemented in the R package "glmnet," also known as "coxnet" [15].

Prediction model building and detailed study of selected genes

Along with selected genetic markers and clinical variables we fitted a multiple Cox regression model. For further integrated study of the selected gene markers, we grouped patients into two groups in terms of fitted hazard ratios as high- and low-risk groups. Then, log rank test was performed to test for homogeneity of survival rates between the two groups. Also, a functional study for the selected gene markers using a pathway database was provided for biological understanding [16].

Results

To build and validate the prediction models, the whole dataset was divided into training and test datasets at a 2:1 ratio.

Clinical results

Descriptive statistics of the clinical information are summarized in Table 1. First, all categorical variables were represented by dummy variables. Note that each variable has less than 30% of observations missing. When we put all variables together, however, they made up more than

Table 1. Descriptive statistics for clinical information

Variable name	Variable type	Descriptive statistics	No. of missing	Missing rate (%)
Sex	Binary	Female: 365, Male: 0	-	-
Age	Count	Mean: 60.07 Standard deviation: 11.29	8	2.19
Race	Category	White: 324, non-white: 41	10	2.74
Neoplasm histologic grade	Category	G2: 42, G3: 315	8	2.19
Primary therapy Outcome success	Category	Complete response: 203 Partial response: 43 Progressive disease: 27 Stable disease: 22	70	19.18
Tumor stage	Category	II: 20, III: 288, IV: 55	2	0.55
Tumor residual disease	Category	1–10 mm: 169, 11–20 mm: 26 More than 20 mm: 69 No macroscopic disease: 63	38	10.41
Platinum status	Category	Resistant: 63 Sensitive: 151 Too early: 47	104	28.49

Table 2. Variable selection results for clinical variables

Variable name	Coefficient	Exp(coef)	se(coef)	z	Pr(>|z|)
P. Therapy outcome: complete response	−0.52334	0.592538	0.252489	−2.073	0.0382
P. Therapy outcome: progressive disease	0.57476	1.776705	0.283553	2.027	0.0427
Tumor residual: >20 mm	0.135709	1.145349	0.219691	0.618	0.5368
Tumor stage: IIC	−1.28502	0.276646	0.729785	−1.761	0.0783
Age	0.014672	1.014781	0.007917	1.853	0.0639
Platinum status: resistant	1.040858	2.831644	0.241527	4.309	1.64E-05

Table 3. Selected gene markers using modified penalized Cox regression

Method	Tuning parameter	Selected marker	Coefficient
LASSO with	Lambda: 0.1743	REN (ENSG00000143839)	-2.64×10^{-5}
Penalty factor on clinical variables	Gamma: 0.3	LEFTY1 (ENSG00000243709)	-3.51×10^{-6}
	Alpha: 1	AP1S2 (ENSG00000182287)	-5.32×10^{-5}

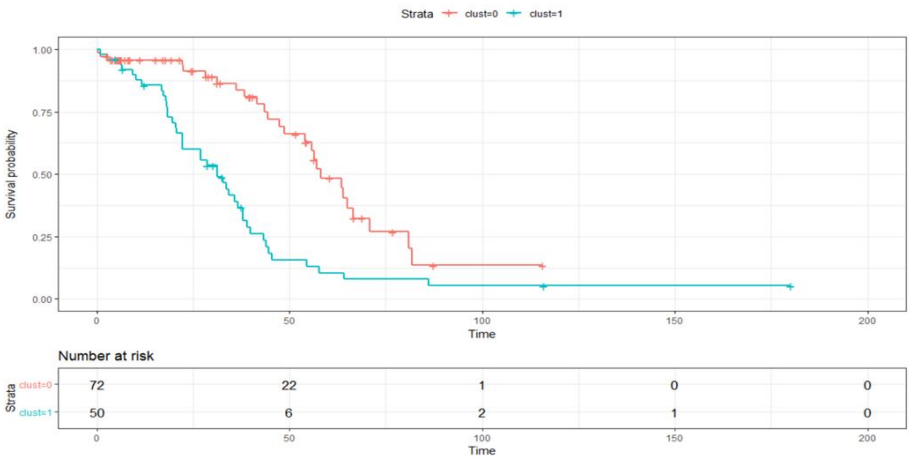

Fig. 1. Survival curve based on predicted values (3 genetic markers with 6-clinical-variable adjustment, time unit is month).

40.55% of observations with at least one variable missing. To include cases with missing variables, imputation was performed using the "mice" package.

Variable selection of clinical variables was performed for Cox regression by using stepwise selection procedures. The results are summarized in Table 2. A total of 6 clinical variables were chosen out of 26 variables. Among the 6 variables, platinum status and primal therapy outcome were significant for prognosis at a 5% significance level.

Gene marker selection

After applying a 20% zero proportion threshold for filtration, a total of 42,747 genes were selected. For ease of interpretation, only protein-coding genes and microRNA genes, which numbered up to 18,415 genes, were used. We performed two pre-screening processes. The first one used MAD.

Genes with MAD under the 10% quantile were filtered out to control false positives. The second filtering was performed using p-values of the gene markers. After that, genes with p-values larger than 0.3 were removed. After the pre-screening, 6,161 genes remained. We then fit the modified penalized Cox regression model as explained in the Methods section. Among the many values, 0.2 was chosen as the gamma parameter to be effective for both marker selection and prediction in this study. Finally, three genes—*REN*, *LEFTY1*, and *AP1S2*—were selected, along with 6 clinical variables. These selected markers are summarized in Table 3.

Integrated study of selected gene markers

Along with the selected genetic markers and clinical variables, we fitted a multiple Cox regression model. The genetic markers chosen for the final model were *REN*, *LEFTY1*, and *AP1S2*. From the Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway database [16], we found out that these genes belong to the following pathways, respectively: KEGG RENIN ANGIOTENSIN SYSTEM, KEGG TGF BETA SIGNALING PATHWAY, and KEGG

LYSOSOME. The false discovery rate q-values from the univariate Cox regression with a single gene were 0.693, 0.687, and 0.138, respectively. These q-values mean that those genetic markers cannot be selected through the false discovery rate procedure with a level of 0.1 via univariate Cox regression.

For further integrated study of the selected gene markers, we grouped patients into two groups—high- and low-risk—using the fitted hazard ratios. As shown in Fig. 1, the two groups were well separated, with a log-rank test p-value of 8.02×10^{-6}. We could see that our model predicts the prognosis of patients quite well. The three selected genes have been reported to have an association with ovarian cancer in previous studies. The renin-angiotensin system works as an angiogenic factor through type 1 angiotensin receptors. These angiogenic factors have a correlation with ovarian cancer patient survival [17]. Lefty1 is an activator of the *TCEA3* gene, the expression of which is related to cell death in ovarian cancer [18]. Lastly, *AP1S2* has been reported as a prognostic marker of ovarian cancer, and its expression level is differentially changed in drug-resistant cell lines [19, 20].

Prediction model results

Two measures of assessment were considered. One was Harrell's concordance index, and the other was the 2-year time-dependent area under the receiver operating characteristic curve (AUC) [21, 22]. For purposes of comparison, we considered three models: a multiple Cox model with only clinical variables (M1), penalized Cox regression (M2), and our proposed method (M3). Table 4 shows the C-index and time-dependent AUC of each model. Among the three models, penalized Cox regression M2 showed even worse results than M1 alone with the clinical variables, whereas our method, M3, showed the best performance.

The time-dependent receiver operating characteristic curves for the training sets and test sets are shown in Fig. 2. In most of the range of Fig. 2, the curve of M3 is higher than the other curves on in the upper left corner, showing better performance with regard to prognostic prediction.

Discussion

In this study, we presented the process of building prediction models, based on clinical variables and high-dimensional RNA-Seq data for HSGC patients provided by

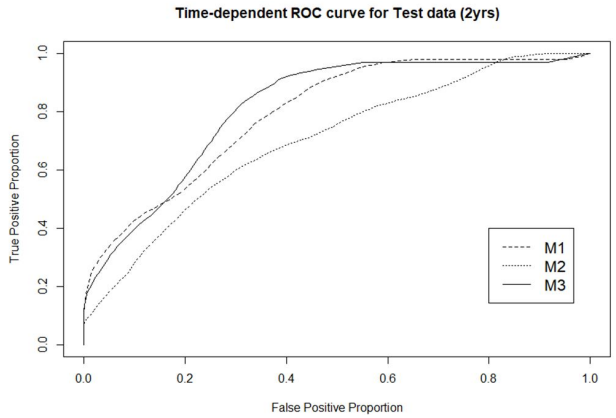

Fig. 2. Time-dependent receiver operating characteristic (ROC) curves of the three models for the test dataset.

Table 4. Comparison of prediction with 3 methods: M1, M2, and M3

Method	Training C-index	Training AUC (2 years)	AIC	Test C-index	Test AUC (2 years)
M1	0.715	0.797	1,129.668	0.687	0.802
M2	0.763	0.804	-	0.642	0.700
M3	0.731	0.771	1,105.292	0.6953	0.813

AUC, area under the receiver operating characteristic curve; AIC, Akaike's information criterion.

Table 5. Model coefficients and inferences for proposed model

	coef	exp(coef)	se(coef)	z	Pr(>|z|)
P. Therapy outcome: complete response	−6.00E-01	5.49E-01	2.48E-01	−2.417	0.0157
P. Therapy outcome: progressive disease	5.18E-01	1.68E+00	2.87E-01	1.805	0.0711
Tumor residual: >20 mm	7.14E-02	1.07E+00	2.21E-01	0.324	0.7463
Tumor stage: IIC	−1.22E+00	2.95E-01	7.31E-01	−1.672	0.0946
Age	1.93E-02	1.02E+00	7.89E-03	2.441	0.0147
Platinum status, resistant	1.05E+00	2.87E+00	2.34E-01	4.505	6.65E-06
REN	−3.81E-04	1.00E+00	2.36E-04	−1.615	0.1063
LEFTY1	−1.22E-04	1.00E+00	7.35E-05	−1.657	0.0975
AP1S2	−9.93E-04	9.99E-01	6.75E-04	−1.471	0.1414

the TCGA GDC portal. The specific steps were as follows: gene filtration, pre-screening, gene marker selection, and integrated study of selected gene markers with the final prediction model.

To put more emphasis on the clinical variables, we applied modified penalized Cox regression. We first selected clinical variables that had major impacts on the survival outcome of HGSC, and then, we found genes using these clinical variables with higher weights than genes in the penalized Cox regression. As shown in Table 5, six clinical variables and three genes were chosen as the markers for prognostic prediction. While these three genes were reported to be associated with HGSC, no significant results were found in the single-gene analysis.

When analyzing data, a single marker alone may not be sufficient to explain the outcome. For example, the *RENIN* gene itself cannot distinguish survival patterns of ovarian cancer patients well. However, a model that includes additional gene markers can contribute to improving the predictability by further explaining the areas that cannot be explained by clinical variables [20].

Furthermore, it is important to keep in mind that the issue of false positives can arise, since the analysis is conducted with a relatively small number of samples compared to the number of genes. Therefore, it is important to determine which genes to use in the Cox model by considering the appropriate gene filtering criteria.

We proposed procedures for building prediction models for predicting survival outcomes by integrating RNA-Seq data and clinical information for HGSC. The approach in this study is general, in the sense that it may not be highly dependent on the characteristics of the cancer type or other types of biomarkers. Therefore, this predictive model development process could be easily applied to other types of cancer.

In the future, we hope to apply our approach to other types of genomic data, such as DNA methylation and copy number alterations. In addition, we want to build our integrative prediction model to predict survival times more accurately for other cancer patients.

ORCID: Seokho Jeong: https://orcid.org/0000-0002-1864-9962; Lydia Mok: https://orcid.org/0000-0002-4029-5793; Se Ik Kim: https://orcid.org/0000-0002-9790-6735; TaeJin Ahn: https://orcid.org/0000-0001-5165-2744; Yong-Sang Song: https://orcid.org/0000-0001-7115-4021; Taesung Park: https://orcid.org/0000-0002-8294-590X

Authors' contribution

Conceptualization: TA, YSS, TP
Data curation: SJ, LM, SIK
Formal analysis: SJ, LM, TA
Funding acquisition: YSS, TP
Methodology: SJ, LM, TA, TP
Writing – original draft: SJ, LM, TP
Writing – review & editing: TA, YSS, TP

Conflicts of Interest

No potential conflict of interest relevant to this article was reported.

Acknowledgments

This research was supported by a grant of the Korea Health Technology R&D Project through the Korea Health Industry Development Institute (KHIDI), funded by the Ministry of Health & Welfare, Republic of Korea (grant number: HI16C2037).

References

1. Welsh JB, Zarrinkar PP, Sapinoso LM, Kern SG, Behling CA, Monk BJ, *et al*. Analysis of gene expression profiles in normal and neoplastic ovarian tissue samples identifies candidate molecular markers of epithelial ovarian cancer. *Proc Natl Acad Sci U S A* 2001;98:1176-1181.
2. Kristiansen G, Denkert C, Schluns K, Dahl E, Pilarsky C, Hauptmann S. CD24 is expressed in ovarian cancer and is a new independent prognostic marker of patient survival. *Am J Pathol* 2002;161:1215-1221.
3. Au KK, Josahkian JA, Francis JA, Squire JA, Koti M. Current state of biomarkers in ovarian cancer prognosis. *Future Oncol* 2015;11:3187-3195.
4. Nowsheen S, Aziz K, Panayiotidis MI, Georgakilas AG. Molecular markers for cancer prognosis and treatment: have we struck gold? *Cancer Lett* 2012;327:142-152.
5. Tothill RW, Tinker AV, George J, Brown R, Fox SB, Lade S, *et al*. Novel molecular subtypes of serous and endometrioid ovarian cancer linked to clinical outcome. *Clin Cancer Res* 2008;14:5198-5208.
6. Yoshihara K, Tajima A, Yahata T, Kodama S, Fujiwara H, Suzuki M, *et al*. Gene expression profile for predicting survival in advanced-stage serous ovarian cancer across two independent datasets. *PLoS One* 2010;5:e9615.
7. Ozsolak F, Milos PM. RNA sequencing: advances, challenges and opportunities. *Nat Rev Genet* 2011;12:87-98.
8. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* 2014;15:550.
9. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* 2010;26:139-140.
10. Maza E. In Papyro Comparison of TMM (edgeR), RLE (DESeq2), and MRN normalization methods for a simple

two-conditions-without-replicates RNA-Seq experimental design. *Front Genet* 2016;7:164.
11. van Buuren S, Groothuis-Oudshoorn K. mice: multivariate imputation by chained equations in R. *J Stat Softw* 2011; 45:1-67.
12. Bourgon R, Gentleman R, Huber W. Independent filtering increases detection power for high-throughput experiments. *Proc Natl Acad Sci U S A* 2010;107:9546-9551.
13. Grimes T, Walker AR, Datta S, Datta S. Predicting survival times for neuroblastoma patients using RNA-seq expression profiles. *Biol Direct* 2018;13:11.
14. Leys C, Ley C, Klein O, Bernard P, Licata L. Detecting outliers: do not use standard deviation around the mean, use absolute deviation around the median. *J Exp Soc Psychol* 2013;49: 764-766.
15. Friedman J, Hastie T, Tibshirani R, Simon N, Narasimhan B, Qian J. Package 'glmnet': Lasso and elastic-net regularized generalized linear models. R package version, 1.4 [software]. The Comprehensive R Archive Network; 2009.
16. Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res* 2000;28:27-30.
17. Ino K, Shibata K, Kajiyama H, Yamamoto E, Nagasaka T, Nawa A, *et al*. Angiotensin II type 1 receptor expression in ovarian cancer and its correlation with tumour angiogenesis and patient survival. *Br J Cancer* 2006;94:552-560.
18. Cha Y, Kim DK, Hyun J, Kim SJ, Park KS. TCEA3 binds to TGF-beta receptor I and induces Smad-independent, JNK-dependent apoptosis in ovarian cancer cells. *Cell Signal* 2013;25: 1245-1251.
19. Nam S, Long X, Kwon C, Kim S, Nephew KP. An integrative analysis of cellular contexts, miRNAs and mRNAs reveals network clusters associated with antiestrogen-resistant breast cancer cells. *BMC Genomics* 2012;13:732.
20. Pontén F, Jirström K, Uhlen M. The Human Protein Atlas: a tool for pathology. *J Pathol* 2008;216:387-393.
21. Harrell FE Jr, Lee KL, Mark DB. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Stat Med* 1996;15:361-387.
22. Heagerty PJ, Lumley T, Pepe MS. Time-dependent ROC curves for censored survival data and a diagnostic marker. *Biometrics* 2000;56:337-344.