JBUON 2016; 21(5): 1203-1209
ISSN: 1107-0625, online ISSN: 2241-6293 • www.jbuon.com
E-mail: editorial_office@jbuon.com

ORIGINAL ARTICLE

Integrating differential pathway analysis with Monte Carlo cross-validation reveals seed cross-talks in non-small cell lung carcinoma

Meng-Li Zheng [1], Nai-Kang Zhou [2], Cheng-Hua Luo [3]

[1] Department of Chest Surgery, the 309th Hospital, PLA, Beijing, 100091, China; [2] Department of Chest Surgery, General Hospital, PLA, Beijing, 100853, China; [3] Department of General Surgery, Peking University International Hospital, Beijing, 100026, China

Summary

Purpose: *The objective of this study was to identify seed pathway cross-talks in non-small cell lung carcinoma (NSCLC), and to reveal potential pathological mechanism at molecular level systematically.*

Methods: *Differentially expressed genes (DEGs) between NSCLC and normal controls were identified using quantile-adjusted conditional maximum likelihood (QCML) method. Subsequently, differential pathways (DPs) enriched by DEGs were determined according to the Ingenuity Pathways Analysis (IPA) pathways and Fisher's exact test. A discriminating score (DS) was computed for each pair of DPs also called as cross-talk, and random forest (RF) algorithm was implemented to investigated hub cross-talks. Finally, global cross-talks with repeated times > 5 were calculated by Monte Carlo Cross-Validation (MCCV). By taking intersections between hub cross-talks and global cross-talks, we obtained seed cross-talks.*

Results: *We obtained 122 DEGs and 5 DPs between NSCLC samples and normal controls. Based on DS and RF algorithm, 5 hub cross-talks with best area under the curve (AUC) were identified, of which Agranulocyte Adhesion and Diapedesis, and IL-17A Signaling in Fibroblasts were the best with AUC=0.996. After intersected with global cross-talks, we gained 2 seed cross-talks (Agranulocyte Adhesion and Diapedesis, Granulocyte Adhesion and Diapedesis and Agranulocyte Adhesion and Diapedesis, Glutathione Redox Reactions I).*

Conclusions: *Two seed cross-talks were identified and validated by MCCV, which may give insights for revealing pathological mechanism and potential biomarkers for target therapy in NSCLC.*

Key words*: cross-talk, differential pathway, Monte Carlo cross-validation, non-small cell lung cancer, random forest classification.*

Introduction

Lung cancer is the leading cause of cancer mortality worldwide, with more than 1.3 million deaths each year [1], of which NSCLC accounts for approximately 80% [2]. At present, treatment of NSCLC is based on histopathological features and staging, however, pathologically similar tumors with comparable stage show dramatically different response to the same therapy of NSCLC [3]. Common features at the molecular level may be able to predict such outcome discrepancies among patients more reliably. Recent advances in microarray technology enable researchers to recapitulate molecular properties of NSCLC at the level of individual genes [4-6]. For instance, epidermal growth factor receptor (EGFR) has emerged as the most significant target in the treatment of NSCLC [6].

However, genes and their protein products do not only function individually, but also interact with others [7]. Similar genes and their interac-

Correspondence to: Meng-Li Zheng, MD. Department of Chest Surgery, the 309th Hospital, PLA, No. 17 on Heishanhujia Road, Haidian District, Beijing, 100091, China. Tel: +86 01066775079, Fax: +86 01066767722, E-mail: menglizheng2016@yeah.net

Received : 04/02/2016; Accepted : 18/02/2016

tions may display similar functions and participate in the same biological pathway. Moreover, a pathway does not act as independent mechanism, but correlates to other pathways, a situation referred to as cross-talk [8]. A cross-talk between two pathways can regulate interactions or express the gene overlap among them [9]. Many of the cellular signaling pathways are interconnected to maintain homeostasis in normal cells [10]. With the development of cancer, the pathway cross-talks are deeply affected [11]. Therefore, the cross-talk among pathways is a crucial step for understanding the pathological mechanisms and the synergistic effects in cancers.

The objective of this work was to identify seed pathway cross-talks in NSCLC by integrating differential pathway analysis with MCCV. Firstly, we explored DEGs across NSCLC patients and normal controls based on QCML. Secondly, DPs enriched by DEGs were identified according to the IPA pathways and Fisher's exact test. Thirdly, hub cross-talks of DPs were evaluated by combining DS with RF classification model. Finally, global cross-talks with repeated times >5 were calculated by MCCV. The intersections between hub and global cross-talks were considered to be seed cross-talks of pathways in NSCLC.

Methods

Data

In this paper, microarray dataset with accessing number E-GEOD-19188 was downloaded from ArrayExpress database for NSCLC related studies. E-GEOD-19188 was composed of 91 NSCLC samples and 65 normal samples, and presented on A-AFFY-44 - Affymetrix GeneChip Human Genome U133 Plus 2.0 Platform. By removing duplicated probes and converting probes into gene symbols, a total of 20544 genes were obtained in the data. Subsequently, the data were normalized through quantiles based algorithm [12], whose goal was to make the distribution of symbol intensities for each array in a set of arrays the same. We selected genes which had mean higher than the 0.25×quantile mean across all samples for further exploitation.

DEGs

To determine whether a gene was differentially expressed across NSCLC and normal controls, we utilized the edgeR package from Bioconductor [13] based on QCML method [14]. It dispersed the parameter of the negative binomial (NB) distribution and compared its performance among different conditions [15]. For gene **g** in sample **i**, its NB model was distributed as follows:

$$Y_{gi} \sim NB(T_i S_{gj}, \varphi_g)$$

where $\mathbf{T_i}$ was the total number of reads, was $\boldsymbol{\varphi_g}$ the dispersion, $\mathbf{S_{gj}}$ was the relative abundance of gene **g** in the experimental group **j** to which sample **i** belonged; $\boldsymbol{\varphi_g}$ represents the coefficient of variation of biological variation between the samples; the mean of NB parameterization was $\boldsymbol{\mu_{gi}} = \mathbf{T_i S_{gj}}$ and variance was $\boldsymbol{\mu_{gi}}(1+\boldsymbol{\mu_{gi}\varphi_g})$. For differential expression analysis, the parameters of interest are $\mathbf{S_{gj}}$. The P was corrected by Benjamini-Hochberg (BH) procedure for multiple testing corrections [16]. Only genes that met to the thresholds of false discovery rate (FDR) < 0.01 and |logFoldChange| > 2 were considered as DEGs between NSCLC and normal controls.

DPs

Pathway enrichment analysis for DEGs was performed on the basis of 589 biological pathways derived from the IPA tool [17]. The first step was mapping DEGs into biological pathways to make them more confident, and we obtained the pathways responsible for coordinating DEGs activities. Next, Fisher's exact test was applied between DEGs and genes of IPA pathways, which is a statistical significance test used in the analysis of contingency Tables [18]. Hence we obtained pathways enriched with FDR < 0.01 corrected by BH test [16], and denoted them as DPs in NSCLC with respect to normal controls. The P was calculated as follows:

$$P = 1 - \sum_{k=0}^{a-1} \frac{\binom{A}{k}\binom{B-A}{b-k}}{\binom{B}{b}}$$

of which **B** was the number of total genes, **A** was the amount of genes in one gene set, **k** was the gene number of one gene set in the gene lists, **k** = **a**-1; **b** was the gene number of one gene list in the total genes.

DP cross-talks

In order to explore the cross-talk of two DPs, we implemented the DS, which indicates the relationships between pairs of pathways, with a larger value indicating relatively higher difference of activity between pathways [11]. The DS is mainly dependent on the comparison of gene expression levels between each pair of DPs. For any pair of DPs (**u**, **v**) the following formula was used:

$$DS(u,v) = \frac{|M_u - M_v|}{D_u + D_v}$$

Where $\mathbf{M_u}$ and $\mathbf{D_u}$ represent mean and standard deviation of expression levels of genes in a pathway **u**; $\mathbf{M_v}$ and $\mathbf{D_v}$ represent mean and standard deviation of expression levels of genes in a pathway **v**.

Hub cross-talks

A RF classification model was employed to evaluate the performance of DP cross-talks across NSCLC and normal controls [19]. The algorithm included three parts: drawing $\mathbf{N}_{tree}$ bootstrap samples from the data was the first part ($\mathbf{N}_{tree}$ = 500). Next, a regression tree was grown from each of the bootstrap samples. And last, new data were predicted by aggregating the predictions of the $\mathbf{N}_{tree}$ trees. For the purpose of accessing the classification results, the area under the receiver operating characteristics curve (AUC) was engaged by 10-fold cross-validation method, due to its consideration of the nature of the incorrect predictions than accuracy [20]. Classification was applied on DP cross-talks based on DS for each sample, and we defined top 5 cross-talks in descending order of AUC as hub cross-talks.

Global cross-talks

In this paper, to evaluate the activities and functions of hub cross-talks in NSCLC samples, the MCCV method was adopted [21]. The total **n** samples (**X**) were randomly split into two sets, the first part (calibration set), denoted as $\mathbf{S}_c$, contained $\mathbf{n}_c$ samples for fitting the models. The other part (validation set), denoted as $\mathbf{S}_v$, included $\mathbf{n}_v$ samples for validating the model:

$$MCCV_{n_v} = \frac{1}{Rn_v}\sum_{k=1}^{R} \|E\|^2$$

of which **E** represents the squared prediction error, **R** stands for the procedure repeated times (**R**=50). Theoretically, the fewer samples used in model calibration, the more repeat times were needed. For each bootstrap, the DEGs, DPs, cross-talks and their DS values were carried out. The DP cross-talks with repeated times>5 were statistically counted, and denoted as global cross-talks. The more repeated times might imply the more significant of this cross-talk was.

Results

DEGs

After eliminating genes of lower than the 0.25 ⊠ quantile mean across all samples, we obtained 15408 genes in the data and detected DEGs between NSCLC patients and normal controls based on them. Using QCML method, a total of 122 DEGs were identified under the thresholds of FDR < 0.01 and |logFoldChange| > 2.

DPs

By taking intersections among DEGs and IPA pathways, we gained common genes and DEGs enriched pathways. For these pathways, we used Fisher's exact test, and obtained 5 DPs of FDR < 0.01 in total, as shown in Table 1. The DPs were Agranulocyte Adhesion and Diapedesis (FDR=4.330E-04), IL-17A Signaling in Fibroblasts (FDR=1.108E-03), Granulocyte Adhesion and Diapedesis (FDR=2.232E-03), Glutathione Redox Reactions I (FDR=4.341E-03) and Differential Regulation of Cytokine Production in Macrophages and T Helper Cells by IL-17A and IL-17F (FDR=4.865 E-03).

Hub cross-talks

In the current study, a DS was computed by comparing the gene expression levels of each cross-talk formed by DPs, and then RF classification was applied on cross-talks utilizing DS for each sample. Additionally, the 5 DPs composed 10 cross-talks at random. When setting the criterion to top 5 of AUC in descending order, a total of 5 hub cross-talks were identified (Table 2). The best one was Agranulocyte Adhesion and Diapedesis, IL-17A Signaling in Fibroblasts with AUC=0.996, indicating that this cross-talk had a good performance in classifying NSCLC samples and normal samples. Interestingly, we found 4 of 5 cross-talks were comprised of Agranulocyte Adhesion

Table 1. Differential pathways with FDR < 0.01

Differential pathways	*FDR*	*Number of genes in pathway*	*Number of DEGs*
Agranulocyte Adhesion and Diapedesis	4.330E-04	173	6
IL-17A Signaling in Fibroblasts	1.108E-03	35	3
Granulocyte Adhesion and Diapedesis	2.232E-03	163	5
Glutathione Redox Reactions I	4.341E-03	17	2
Differential Regulation of Cytokine Production in macrophages and T helper cells by IL-17A and IL17F	4.865 E-03	18	2

DEGs: differentially expressed genes, FDR: false discovery rate

Table 2. Hub cross-talks with AUC value for random forest classification

ID	*Cross-talks*	*AUC*
1	a: Agranulocyte Adhesion and Diapedesis b: IL-17A Signaling in Fibroblasts	0.996
2	a: Agranulocyte Adhesion and Diapedesis b: Granulocyte Adhesion and Diapedesis	0.986
3	a: Agranulocyte Adhesion and Diapedesis b: Glutathione Redox Reactions I	0.981
4	a: Agranulocyte Adhesion and Diapedesis b: Differential Regulation of Cytokine Production in macrophages and T helper cells by IL-17A and IL-17F	0.975
5	a: IL-17A Signaling in Fibroblasts b: Granulocyte Adhesion and Diapedesis	0.974

and Diapedesis, which suggested that this DP played significant role in the NSCLC progression. There also were two cross-talks with AUC >0.980 (Agranulocyte Adhesion and Diapedesis, Granulocyte Adhesion and Diapedesis) with AUC=0.986 and Agranulocyte Adhesion and Diapedesis, Glutathione Redox Reactions I with AUC = 0.981.

Global cross-talks

We divided total samples (N=156) into two sets according to the ratio of 3:2, and kept the 94 to build a calibration set (29 normal samples and 55 NSCLC samples) and 62 to construct a validation set (34 normal samples and 36 NSCLC samples). The MCCV process was repeated multiple times (50 bootstraps), generating (at random) new training and test partitions each time. Figure 1 displays the 22 cross-talks with repeated times > 5, also called as global cross-talks. We discovered that the cross-talk Agranulocyte Adhesion and Diapedesis, Inhibition of Matrix Metalloproteases and Granulocyte Adhesion and Diapedesis, Inhibition of Matrix Metalloproteases possessed the most repeated times of 22; the next were IL-17A Signaling in Fibroblasts, Inhibition of Matrix Metalloproteases, Agranulocyte Adhesion and Diapedesis, HIF1 Signaling and Granulocyte Adhesion and Diapedesis, HIF1 Signaling, that had 17 repeated times. What was more, among the global cross-talks, two also presented in hub cross-talks, and were Agranulocyte Adhesion and Diapedesis, Granulocyte Adhesion and Diapedesis and Agranulocyte Adhesion and Diapedesis, Glutathione Redox Reactions I. We might infer that the two common cross-talks were more significant than others in the progression of NSCLC, and denoted them as seed cross-talks.

Discussion

The presence and amount of different pathways influences have not been completely studied although this scenario is intuitive; most importantly, the precise available methodology able to quantify the amount of such cross-talk for pairs of pathway is rare [22,23]. In this work, we applied a method by integrating DEGs, DPs, DS, RF classification with MCCV. It has been demonstrated that the method is even more interesting from a biological point of view, and thus we employed to identify seed cross-talks in NSCLC. This finding could gain an insight into revealing the pathological mechanism of NSCLC [11].

We obtained 2 seed cross-talks of pathways in NSCLC: “Agranulocyte Adhesion and Diapedesis, Granulocyte Adhesion and Diapedesis” and “Agranulocyte Adhesion and Diapedesis, Glutathione Redox Reactions I”. Both Agranulocyte and Granulocyte belong to haemocyte groups and are characterized by the presence of granules in their cytoplasm [24]. In the report by Cui and Willingham [25], separation of white blood cells into granulocyte and agranulocyte types had close relationship with cancer-killing-activity (CKA) in their white blood cells. In addition, adhesion is a fundamental feature of multicellular organisms inhibiting growth of tumor cells [26] . Specialized leukocytes (agranulocytes and granulocytes) adhere to and pass through the endothelium of the blood vessels and the underlying matrix during inflammation [27]. Diapedesis is a process that adhering leukocytes crawl to an intercellular junction of the endothelium and then transmigrate to or even through the intercellular matrix [28]. Zhang et al. had revealed that agranulocyte/granulocyte adhesion and diapedesis pathways likely contribute to immunopathogenesis, including mi-

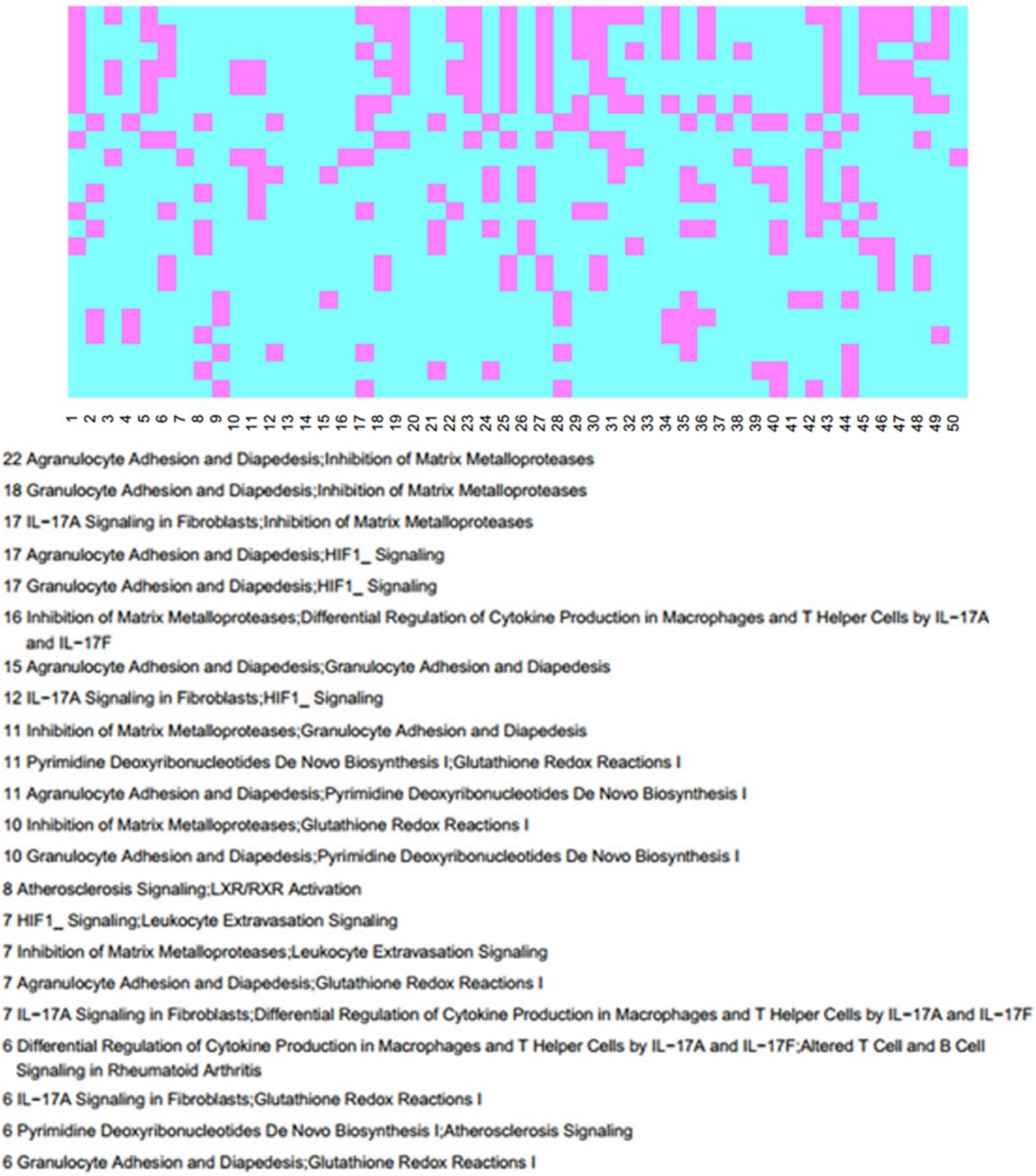

Figure 1. Heatmap for global cross-talks with repeated times > 5.

gration of leukocytes and tumor pathology [29]. Therefore we inferred that the cross-talk (Agranulocyte Adhesion and Diapedesis, Granulocyte Adhesion and Diapedesis) was tightly related to cancer. Meanwhile, this is the first time to propose their correlation with NSCLC.

Glutathione Redox Reactions I was another significant DP between NSCLC and normal controls. Glutathione plays an important role in a multitude of cellular processes, including cell differentiation, proliferation and apoptosis, and disturbances in glutathione homeostasis are involved in the etiology and progression of many human diseases, including cancer [30]. Glutathione Redox

Reactions, the major determinant of the cellular redox status, represented a promising therapeutic strategy for overcoming cancer cell progression and chemoresistance [31]. It had been suggested that the glutathione redox status decreases in the blood of tumor cells [32], and might interact with agranulocyte adhesion and diapedesis in the progression of NSCLC.

In conclusion, we have identified 2 seed cross-talks and validated them by MCCV, which may give insights for revealing pathological mechanism and potential biomarkers for target therapy in NSCLC.

Conflict of interests

The authors declare no confict of interests.

References

1. Vansteenkiste J, De Ruysscher D, Eberhardt W et al. Early and locally advanced non-small-cell lung cancer (NSCLC): ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up. Ann Oncol 2013, Suppl 6, Oct 24.
2. Jemal A, Bray F, Center MM et al. Global cancer statistics. CA: Cancer J Clin 2011;61:69-90.
3. Hou J, Aerts J, Den Hamer B et al. Gene expression-based classification of non-small cell lung carcinomas and survival prediction. PLoS One 2010;5:e10312.
4. Kobayashi K, Nishioka M, Kohno T et al. Identification of genes whose expression is upregulated in lung adenocarcinoma cells in comparison with type II alveolar cells and bronchiolar epithelial cells in vivo. Oncogene 2004;23:3089-3096.
5. Jones MH, Virtanen C, Honjoh D et al. Two prognostically significant subtypes of high-grade lung neuroendocrine tumours independent of small-cell and large-cell neuroendocrine carcinomas identified by gene expression profiles. Lancet 2004;363:775-781.
6. Sequist LV, Bell DW, Lynch TJ et al. Molecular predictors of response to epidermal growth factor receptor antagonists in non–small-cell lung cancer. J Clin Oncol 2007;25:587-595.
7. Vinayagam A, Zirin J, Roesel C et al. Integrating protein-protein interaction networks with phenotypes reveals signs of interactions. Nat Methods 2014;11:94-99.
8. Jin D, Lee H. A Computational Approach to Identifying Gene-microRNA Modules in Cancer. PLoS Comp Biol 2015;11:e1004042.
9. Aksamitiene E, Kiyatkin AB, Kholodenko BN. Cross-talk between mitogenic Ras/MAPK and survival PI3K/Akt pathways: a fine balance. Biochem Soc Trans 2012;40:139-146.
10. Bernards R. A missing link in genotype-directed cancer therapy. Cell 2012;151:465-468.
11. Colaprico A, Cava C, Bertoli G et al. Integrative analysis with Monte Carlo cross-validation reveals miRNAs regulating pathways cross-talk in aggressive breast cancer. BioMed Res Int 2015;2015:831314.
12. Bolstad BM, Irizarry RA, Astrand M et al. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. Bioinformatics 2003;19:185-193.
13. Robinson MD, Smyth GK. Moderated statistical tests for assessing differences in tag abundance. Bioinformatics 2007;23:2881-2887.
14. Robinson MD, Smyth GK. Small-sample estimation of negative binomial dispersion, with applications to SAGE data. Biostatistics 2008;9:321-332.
15. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics 2010;26:139-140.
16. Bogdan M, Ghosh JK, Tokdar ST. A comparison of the Benjamini-Hochberg procedure with some Bayesian rules for multiple testing. In: Beyond parametrics in interdisciplinary research: Festschrift in honor of Professor Pranab K. Sen Institute of Mathematical Statistics, 2008, pp 211-230.
17. Jiménez-Marín Á, Collado-Romero M, Ramirez-Boo M et al. Biological pathway analysis by ArrayUnlock and ingenuity pathway analysis. In: BMC Proceedings BioMed Central Ltd, p S6, 2009.
18. Routledge R (Ed). Fisher's exact test. Encyclopedia of Biostatistics. John Wiley & Sons, Ltd, 2005.
19. Liaw A, Wiener M. Classification and regression by randomForest. R News 2002;2:18-22.
20. Huang J, Ling CX. Using AUC and accuracy in evaluating learning algorithms. Knowledge and Data Engineering, IEEE Transactions on 2005;17:299-310.
21. Xu Q-S, Liang Y-Z. Monte Carlo cross validation. Chemometrics Intelligent Lab Syst 2001;56:1-11.
22. Donato M, Xu Z, Tomoiaga A et al. Analysis and correction of crosstalk effects in pathway analysis. Genome Res 2013;23:1885-1893.
23. Yang H, Cheng C, Zhang W. Average rank-based score to measure deregulation of molecular pathway gene sets. PLoS One 2011;6:e27579.
24. Russell ES, Bernstein SE. Blood and blood formation. Biol Lab Mouse 1966: 520:351-372.
25. Cui Z, Willingham MC. Methods and compositions for the treatment of cancer. Google Patents, 2011.
26. Klahan S, Wu M-S, Hsi E et al. Computational analysis of mRNA expression profiles identifies the ITG fami-

ly and PIK3R3 as crucial genes for regulating triple negative breast cancer cell migration. BioMed Res Int 2014;2014 :536591.

27. Jan L, Chalany J. Clozapine for Treating Pharmacoresistant Schizophrenia among Elders. J Clin Diagn Res 2014;2:101.
28. Carman CV, Springer TA. A transmigratory cup in leukocyte diapedesis both through individual vascular endothelial cells and between them. J Cell Biol 2004;167:377-388.
29. Zhang X, O'Connell C, Nagarajan U et al. Co-expression network analysis reveals immune response pathways important for pathogenesis of Chlamydial pelvic inflammatory disease . J Immunol 2014;192:184-185.
30. Traverso N, Ricciarelli R, Nitti M et al. Role of glutathione in cancer progression and chemoresistance. Oxid Med Cell Longevity 2013;2013:972913.
31. Schafer FQ, Buettner GR. Redox environment of the cell as viewed through the redox state of the glutathione disulfide/glutathione couple. Free Radical Biol Med 2001;30:1191-1212.
32. Mantovani G, Madeddu C, Macciò A (Eds). Antioxidants, Anorexia/Cachexia, and Oxidative Stress in Patients with Advanced-Stage Cancer. In: Oxidative Stress in Cancer Biology and Therapy. Springer, 2012, pp 373-385.