MEDICAL SCIENCE MONITOR

CLINICAL RESEARCH

e-ISSN 1643-3750
© Med Sci Monit, 2017; 23: 774-779
DOI: 10.12659/MSM.899638

Received: 2016.05.18
Accepted: 2016.06.07
Published: 2017.02.12

Personalized Identification of Differentially Expressed Modules in Osteosarcoma

Authors' Contribution:
Study Design A
Data Collection B
Statistical Analysis C
Data Interpretation D
Manuscript Preparation E
Literature Search F
Funds Collection G

AEG **Xiaozhou Liu***
BC **Chengjun Li***
BDF **Lei Zhang**
AG **Xin Shi**
AE **Sujia Wu**

Department of Orthopedics, Jinling Hospital affiliated to Nanjing University, Nanjing, Jiangsu, P.R. China

* Contributed equally

Corresponding Authors: Sujia Wu, e-mail: wusujiamed@yeah.net, Xin Shi, e-mail: shixinmed@yeah.net

Source of support: This study was funded by General Hospital of Nanjing military region, project number 2015003

Background: Osteosarcoma (OS), an aggressive malignant neoplasm, is the most common primary bone cancer mainly in adolescents and young adults. Differentially expressed modules tend to distinguish differences integrally. Identifying modules individually has been crucial for understanding OS mechanisms and applications of custom therapeutic decisions in the future.

Material/Methods: Samples came from individuals were used from control group (n=15) and OS group (n=84). Based on clique-merging, module-identification algorithm was used to identify modules from OS PPI networks. A novel approach – the individualized module aberrance score (iMAS) was performed to distinguish differences, making special use of accumulated normal samples (ANS). We performed biological process ontology to classify functionally modules. Then Support Vector Machine (SVM) was used to test distribution results of normal and OS group with screened modules.

Results: We identified 83 modules containing 2084 genes from PPI network in which 61 modules were significantly different. Cluster analysis of OS using the iMAS method identified 5 modules clusters. Specificity=1.00 and Sensitivity=1.00 proved the distribution outcomes of screened modules were mainly consistent with that of total data, which suggested the efficiency of 61 modules.

Conclusions: We conclude that a novel pipeline that identified the dysregulated modules in individuals of OS. The constructed process is expected to aid in personalized health care, which may present fruitful strategies for medical therapy.

MeSH Keywords: **Gene Regulatory Networks • Osteosarcoma • Support Vector Machines**

Abbreviations: **ANS** – accumulated normal samples; **DEG** – different expression genes; **FC** – fold change; **FCS** – functional class scoring; **FDR** – false discovery rate; **FN** – false-negatives; **FP** – false-positives; **GEO** – gene expression omnibus; **GO** – gene ontology; **iMAS** – individualized module aberrance score; **iPAS** – individualized pathway aberrance score; **LIMMA** – Linear Models for Microarray Data; **OS** – osteosarcoma; **PPI** – protein-protein interaction; **RMA** – robust multi-array average; **SD** – standard deviation; **STRING** – Retrieval of Interacting Genes; **SVM** – Support Vector Machine; **TN** – true-negatives; **TP** – true-positives

Full-text PDF: http://www.medscimonit.com/abstract/index/idArt/899638

1837 — 2 29

Indexed in: [Current Contents/Clinical Medicine] [SCI Expanded] [ISI Alerting System] [ISI Journals Master List] [Index Medicus/MEDLINE] [EMBASE/Excerpta Medica] [Chemical Abstracts/CAS] [Index Copernicus]

Background

Osteosarcoma (OS), an aggressive malignant neoplasm, is the most common primary bone cancer mainly in adolescents and young adults [1,2]. Though the survival state has been improved after the introduction of neoadjuvant chemotherapy, the 5-year survival rate of OS patients with metastases at diagnosis is less than 30% [3,4]. Thus, it is significant that identifying potential therapeutic targets in the insights of molecular mechanisms.

As the high-throughput microarray experiments have been used in many fields, detection of expression level is an efficient approach to find useful genomic alteration in OS patients. Recently, it has been reported that there are many important differently expressed genes (DEG) and pathways in OS [5–8].

There are abundant genes and pathways related to OS in databases. Critical to implicating novel genes is the identification of core modules containing dysregulated pathways and complexes. A straightforward way was performed with identification and comparison of modules across normal and cancer tissue conditions by integrating PPI map and gene expression data [9]. Detecting deregulated pathways between disease and normal groups appears to be hotspot in recent years. The method individualized module aberrance score (iMAS) is designed to compare the expression profiles of a single patient with cohort data to detect molecular aberrances that are particular to the disease [10]. We use iMAS on account of the comparison of one OS patient with a lot of accumulated normal samples (ANS). This is a biologically intuitive guideline to interpret an individual disease that even lacks vast data, which is absolutely different from the traditional gene expression analysis. The new method covers 4 steps: data processing, gene-level statistics, iMAS and a significance test. It could capture biological and clinical information in a sensible, valid and useful way for colorectal cancer and lung cancer [10]. We used iMAS to explore modules in OS samples in order to distinguish differences from the control group.

In this study, we introduced a pipeline that identified modules from PPI network in individuals of OS. After narrowing down the number of correlated modules by Support Vector Machine (SVM), we used iMAS to distinguish differences in individuals. We hope that this will served as therapy-targeting markers and benefit individualized medical treatment of OS.

Material and Methods

Data preprocessing

The transcription profile was obtained from EMBI-EBI ArrayExpress [11]. Gene expression profiling of 99 tissues were collected from E-GEOD-33382 and E-GEOD-28974 to study the behavior of genes and their modules in individuals [5]. Samples were used from 15 controls and 84 OS patients. The platform used was Illumina human-6 v2.0 expression BeadChip (using nuIDs as identifier). Data of the gene chip was read in the array [12]. The Linear Models for Microarray Data (LIMMA) [13] was then used to preprocess data. Background adjustment and quantile data normalization were performed by robust multi-array average (RMA) [14]. To protect against outlier probes we used a robust procedure, median polish [15], to estimate model parameters. The average value of a gene symbol with multiple probes was calculated and 23 214 genes were obtained.

The human protein-protein interaction (PPI) map was collected from the Retrieval of Interacting Genes (STRING; v 9.0) [16], including 1 048 576 PPI sets. Filtering repeated ones, PPI sets were gathered in the condition of the combine score <0.8. Then we constructed a PPI subnet after getting intersection with 23214 genes and PPI network, which contained 37381 PPI sets and 6665 nodes.

All analyses were performed in the bioinformatics platform from Honghui Biotech Co. Ltd. (Jinan, China).

Identifying modules from the PPI network

Using the human subnet as a backbone, we inferred a re-weighted PPI network with expression and mutation profiles of normal and OS. Every side of the constructed PPI network was assigned with absolute value of Spearman correlation coefficient of every interaction according to gene expression data.

Based on clique-merging, a module-identification algorithm was used to identify modules from OS PPI networks [9,17,18]. We identified the set *C* of all maximal cliques of size at least *k* in the PPI network using a fast depth-first search with pruning-based algorithm (CLIQUES) by Tomita et al. (2006). Next, we calculate its *weighted interaction density score (C)* as,

$$score\ (C)=\frac{\sum_{u\in C,v\in C} w(u,v)}{|C|\cdot(|C|-1)} \tag{1}$$

We ranked these cliques in descending order of their *score (C)*. A predefined overlap-threshold t_0=0.5 was set to go through the list repeatedly. The modules were gathered by merging highly overlapping cliques.

Individualized analysis

The individualized module aberrance score (iMAS) for the personalized identification of modules was performed in OS, making special use of accumulated normal samples (ANS). ANS was obtained from the gene expression omnibus (GEO) database of NCBI (*www.ncbi.nlm.nih.gov/geo/*) [19]. Fifteen ANS served with normal samples were collected for identifying obtained modules.

The definition of expression level was developed with the multichip average [14]. For individual OS cases,

$$\text{proj}_d q_k = \left(\frac{1}{\text{n}} \sum_{j=1}^{n} qkj, \ldots, \frac{1}{n} \sum_{j=1}^{n} qkj \right) \tag{2}$$

as quantile normalization was performed after combining the single OS data. Z= (z_1, z_2,..., z_n) represents the expression state of a module, where z_i denotes the standardized expression value of *i*-th gene and the number of genes existing in the module is *n*. Module statistics of every module:

$$iMAS = \frac{\sum_i^n z_i}{n} \tag{3}$$

z_i represents the standardized gene level statistics of 1-*i* gene and the number of genes existing in the module is *n* [10].

Module statistics from OS and normal sample tissues were tested in pairs with Wilcoxon test [20]. Since the test might induce false-positive results, we adjusted the raw P-values into false discovery rate (FDR) to circumvent the problem [21]. The FDR<0.05 and |log fold change (FC)|>1 were used as the cut-off criteria. The adjusted P<0.01 was set to screen differentially expressed modules.

Biological process analysis

Gene ontology (GO) analysis has been used frequently in functional studies of large-scale genomic data [22,23]. To functionally classify modules, we performed biological process ontology using Bingo of Cytoscape version 3.2.0 [24], which is able to reveal enriched GO terms. A *P*-value less than 0.01 was considered statistically significant.

Distribution outcomes in SVM

Support Vector Machine (SVM) was supervised computational methods used for classification and regression tasks that originated from statistical learning theory [25]. This measure is defined as the total number of good classifications over the total number of available examples. SVM is widely used in computational biology due to its high accuracy, ability to deal with high-dimensional databases, and strong flexibility in modeling diverse sources of data [26]. We used C-classification to test the consistency of distribution results between OS modules and expression data. The expressing data of GeneChips were randomly grouped in experimental and test groups by the proportional 6: 4. SVM model was performed in 5 times fold cross-validation method. The classified test points can be divided into 4 categories: true-positives (*TP*), true-negatives (*TN*), false-positives (*FP*), and false-negatives (*FN*).

$$\text{Accuracy} = \frac{\text{TP+TN}}{\text{TP+FP+TN+FN}} \tag{4}$$

Sensitivity and specificity describe how well the classifier discriminates the positive and the negative classes, respectively [27].

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP+FN}} \tag{5}$$

$$\text{Sensitivity} = \frac{\text{TN}}{\text{FP+TN}} \tag{6}$$

Results

Modules identification and GO analysis

We identified 83 modules containing 2084 genes from PPI network used clique-merging. Among them, 61 modules were significantly different with Wilcoxon test (P<0.01). In the Go analysis, 1568 biological processes were obtained from 61 modules. Their relation with p<1.0E-31 is shown in Figure 1.

Differentially expressed modules in individual

Cluster analysis of OS using the iMAS technology on Beer's data distinguished 5 modules clusters (shown by 1–5 in Figure 2). Sample clusters represent histopathological differentiation status. On the ordinate, 15 normal samples are clustered together with red color, and blue color represents the differentiation status of OS. It claimed that impartial clustering-based iMAS was highly sensitive to gather crucial correlation with OS disease. From module clusters of Figure 2, M1, M2, and M5 are distinguished obviously with the control group. Module cluster M3 and M4 are relatively weak in distinguishing the differences between OS samples and ANS. Basically, in most OS samples, modules can detect differences from ANS. Therefore, iMAS is clinically useful in individualized medical treatment of OS.

Distribution outcomes in SVM

We used SVM to test distribution results of normal and OS group with screened modules. The classifier mainly judged the test set true-positive and true-negative exactly. Accuracy=100 indicates the C-classification was precise enough to evaluate modules. Specificity=1.00 and Sensitivity=1.00 prove the distribution outcomes of screened modules were mainly consistent with that of total data, which suggested the efficiency of 61 modules. It proved modules could take the place of vast numbers of genes to distinguish differences in distribution results of OS patients and normal samples.

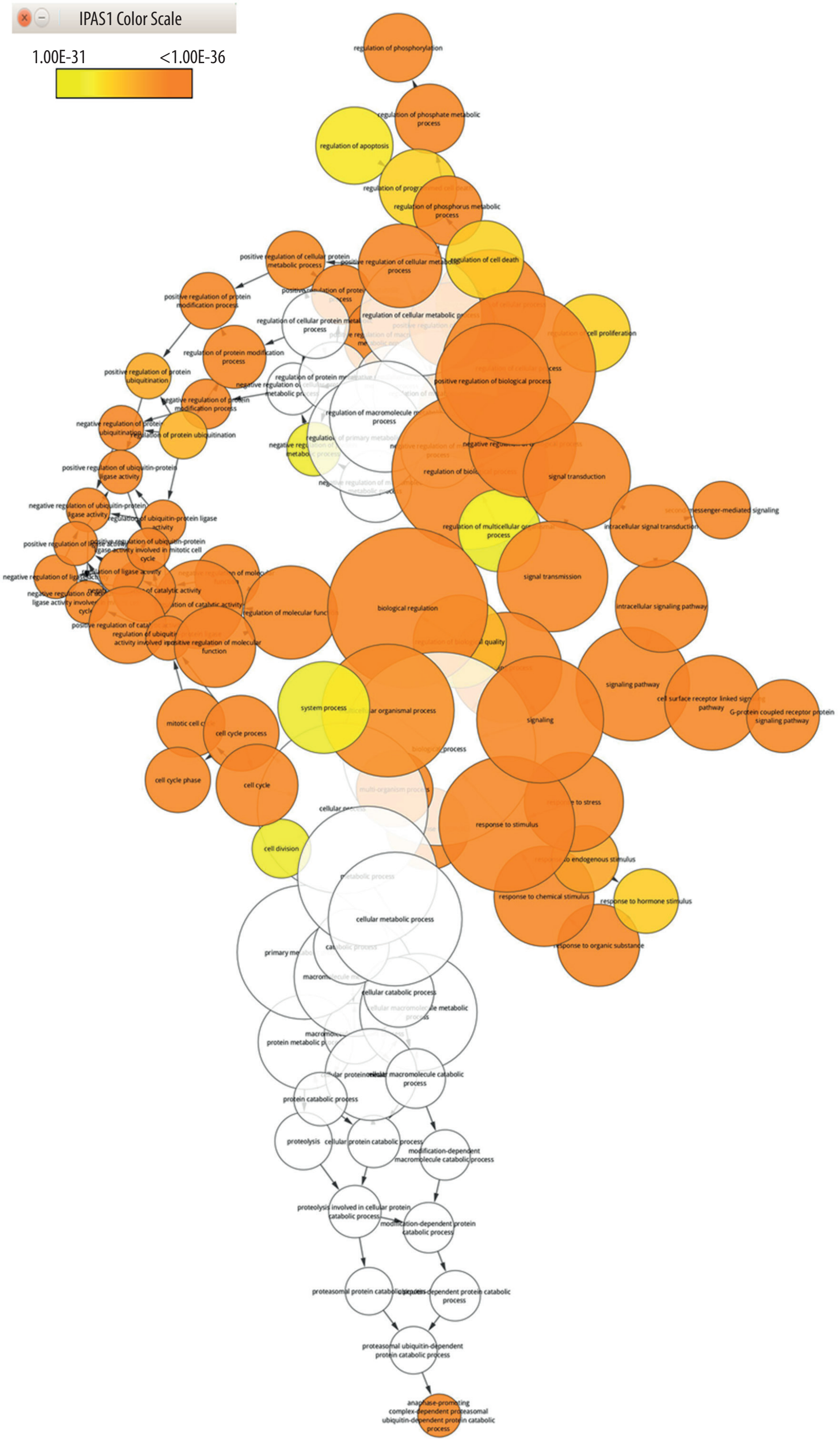
IPAS1 Color Scale
1.00E-31
<1.00E-36
regulation of phosphorylation
regulation of phosphate metabolic process
regulation of apoptosis
regulation of programmed cell death
regulation of phosphorus metabolic process
positive regulation of cellular protein metabolic process
positive regulation of cellular metabolic process
regulation of cell death
positive regulation of protein modification process
positive regulation of protein metabolic process
regulation of cellular protein metabolic process
regulation of cellular metabolic process
positive regulation of macromolecule metabolic process
regulation of cellular process
regulation of cell proliferation
regulation of protein modification process
positive regulation of protein ubiquitination
regulation of protein metabolic process
negative regulation of cellular protein metabolic process
positive regulation of biological process
negative regulation of protein modification process
regulation of macromolecule metabolic process
negative regulation of protein ubiquitination
regulation of protein ubiquitination
regulation of primary metabolic process
negative regulation of metabolic process
negative regulation of macromolecule metabolic process
negative regulation of biological process
regulation of biological process
signal transduction
positive regulation of ubiquitin-protein ligase activity
negative regulation of ubiquitin-protein ligase activity
regulation of ubiquitin-protein ligase activity
second messenger-mediated signaling
intracellular signal transduction
regulation of multicellular organismal process
positive regulation of ligase activity
positive regulation of ubiquitin-protein ligase activity involved in mitotic cell cycle
regulation of ligase activity
negative regulation of ligase activity
negative regulation of molecular function
regulation of catalytic activity
negative regulation of ubiquitin-protein ligase activity involved in mitotic cell cycle
regulation of catalytic activity
positive regulation of catalytic activity
regulation of ubiquitin-protein ligase activity involved in mitotic cell cycle
positive regulation of molecular function
regulation of molecular function
biological regulation
signal transmission
intracellular signaling pathway
regulation of biological quality
signaling process
signaling pathway
cell surface receptor linked signaling pathway
G-protein coupled receptor protein signaling pathway
system process
multicellular organismal process
signaling
mitotic cell cycle
cell cycle process
biological_process
cell cycle phase
cell cycle
multi-organism process
response to stress
cellular process
response to stimulus
cell division
response to endogenous stimulus
metabolic process
response to chemical stimulus
response to hormone stimulus
cellular metabolic process
response to organic substance
primary metabolic process
catabolic process
macromolecule metabolic process
cellular catabolic process
cellular macromolecule metabolic process
macromolecule catabolic process
protein metabolic process
cellular proteinmetabolic process
cellular macromolecule catabolic process
protein catabolic process
proteolysis
cellular protein catabolic process
modification-dependent macromolecule catabolic process
proteolysis involved in cellular protein catabolic process
modification-dependent protein catabolic process
proteasomal protein catabolic process
ubiquitin-dependent protein catabolic process
proteasomal ubiquitin-dependent protein catabolic process
anaphase-promoting complex-dependent proteasomal ubiquitin-dependent protein catabolic process

Figure 1. Relation network of modules in biology process with P<1.0E-31. Depths of color represents the value of P. The size of circles represents degree of the relation.

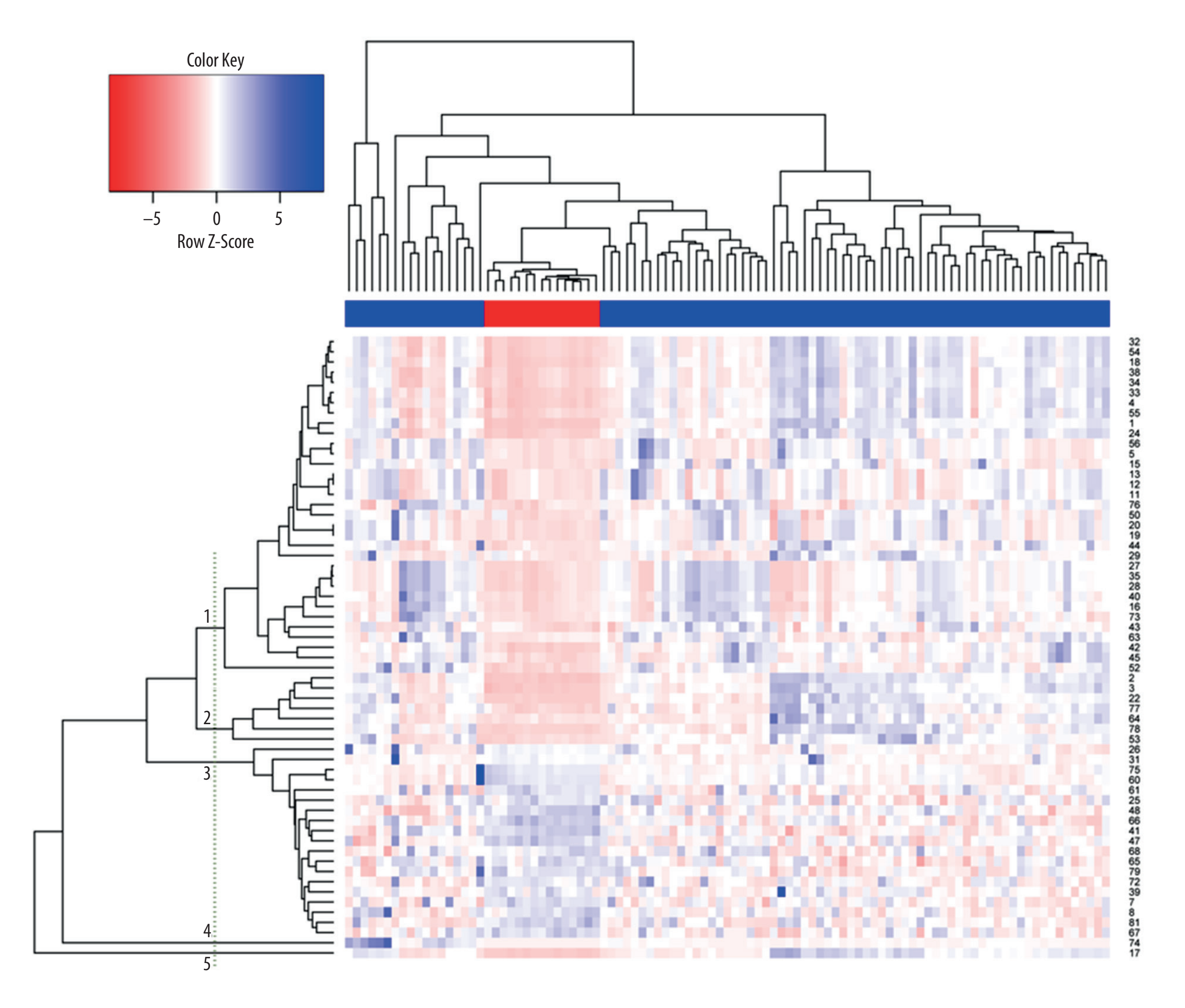

Figure 2. Clustered iPAS of OS dataset. Modules (n=83) were clustered with the abscissa and samples (n=99) were clustered with the ordinate.

Discussion

Dealing with clinical data can be problematic since the available data is usually high-dimensional and heterogeneous, which means vast numbers of features and different types of data [28]. The high accuracy and flexibility of SVM are suitable to the distribution problem proposed in this paper.

We used SVM to assess the efficiency of modules and found their distribution outcomes were mainly consistent with the original data, with high accuracy of 100. This indicates the classification analysis is helpful in precise diagnosis and prognosis of diseases. Therefore, screened modules could take the place of vast numbers of genes to perform individualize therapy in OS. However, the distribution outcomes of 61 modules have not been verified in abundance. We suggest that screened modules need to be checked in new studies and receive further modification.

Module analysis has become a common choice for extracting and explaining the underlying biology for high-throughput molecular measurements. Identification of differentially expressed modules in a single patient is significant to aid in personalized medicine, which may present fruitful strategies for OS therapy. Existing module analytical methods are unsuited to distinguish individual aberrance in pathways and modules. Therefore, we employed the iMAS to analyze the personalized identification of modules, taking advantage of vast amounts of normal data.

A key innovation of the method is the iMAS using ANS in OS. Ahn et al. [10] proved that the Average Z equation could efficiently reveal noticeable aberrance in expression profiles and clinical significance, which sufficed to confirm the best averaged validation rate and distinguish a known survival-relevant pathway statistically. Furthermore, ANS data is expected to be available in more fields of medicine along with the rapid advances of high-throughput databases.

In this study, the iMAS was used to calculate module statistics of every module and Average Z method was selected for modification of existing module analysis methods. There were 61 modules with P<0.01 after adjustment of FDR, which were screened to be the most significant modules. The majority of them were clustered in M1, M2, and M5 (Figure 2). Therefore, these screened modules were efficient in distinguishing differences in individual OS samples. It can provide broader carcinogenic insight in personalized medicine [29].

Conclusions

Based on our results, we present a novel pipeline that identified the dysregulated modules in individuals with OS. Modules can be markers in identification of OS. iMAS provides a sensitive measure for clinical features of patients and can be useful in analysis of individual medical treatment in OS. The constructed process is expected to aid in personalized health care, which may present fruitful strategies for medical therapy.

Conflict of interest

None.

References:

1. Bielack S1, Carrle D, Casali PG; ESMO Guidelines Working Group: Osteosarcoma: ESMO clinical recommendations for diagnosis, treatment and follow-up. Ann Oncol, 2009; 20(Suppl. 4): 137–39
2. Ottaviani G, Jaffe N: The epidemiology of osteosarcoma. Cancer Treat Res, 2009; 152: 3–13
3. Bacci G, Mercuri M, Briccoli A et al: Osteogenic sarcoma of the extremity with detectable lung metastases at presentation. Results of treatment of 23 patients with chemotherapy followed by simultaneous resection of primary and metastatic lesions. Cancer, 1997; 79: 245–54
4. Marina N, Gebhardt M, Teot L, Gorlick R: Biology and therapeutic advances for pediatric osteosarcoma. Oncologist, 2004; 9: 422–41
5. Kuijjer ML, Rydbeck H, Kresse SH et al: Identification of osteosarcoma driver genes by integrative analysis of copy number and gene expression data. Genes Chromosomes Cancer, 2012; 51: 696–706
6. Kresse SH, Rydbeck H, Skarn M et al: Integrative analysis reveals relationships of genetic and epigenetic alterations in osteosarcoma. PLoS One, 2012; 7: e48262
7. Both J, Wu T, Bras J et al: Identification of novel candidate oncogenes in chromosome region 17p11.2-p12 in human osteosarcoma. PLoS One, 2012; 7: e30907
8. Ying M, Liu G, Shimada H et al: Human osteosarcoma CD49f(–)CD133(+) cells: Impaired in osteogenic fate while gain of tumorigenicity. Oncogene, 2013; 32: 4252–63
9. Srihari S, Ragan MA: Systematic tracking of dysregulated modules identifies novel genes in cancer. Bioinformatics, 2013; 29: 1553–61
10. Ahn T, Lee E, Huh N, Park T: Personalized identification of altered pathways in cancer using accumulated normal tissue data. Bioinformatics, 2014; 30: i422–29
11. Parkinson H, Kapushesky M, Shojatalab M et al: ArrayExpress – a public database of microarray experiments and gene expression profiles. Nucleic Acids Res, 2007; 35: D747–50
12. Gautier L, Cope L, Bolstad BM, Irizarry RA: affy – analysis of Affymetrix GeneChip data at the probe level. Bioinformatics, 2004; 20: 307–15
13. Diboun I, Wernisch L, Orengo CA, Koltzenburg M: Microarray analysis after RNA amplification can detect pronounced differences in gene expression using limma. BMC Genomics, 2006; 7: 252
14. Irizarry RA, Hobbs B, Collin F et al: Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostatistics, 2003; 4: 249–64
15. Sasik R, Calvo E, Corbeil J: Statistical analysis of high-density oligonucleotide arrays: A multiplicative noise model. Bioinformatics, 2002; 18: 1633–40
16. Franceschini A, Szklarczyk D, Frankild S et al: STRING v9.1: Protein-protein interaction networks, with increased coverage and integration. Nucleic Acids Res, 2013; 41: D808–15
17. Liu G, Wong L, Chua HN: Complex discovery from weighted PPI networks. Bioinformatics, 2009; 25: 1891–97
18. Srihari S, Leong HW: A survey of computational methods for protein complex prediction from protein interaction networks. J Bioinform Comput Biol, 2013; 11: 1230002
19. Barrett T, Troup DB, Wilhite SE et al: NCBI GEO: Archive for functional genomics data sets – 10 years on. Nucleic Acids Res, 2011; 39: D1005–10
20. Rosner B, Glynn RJ, Lee ML: Incorporation of clustering effects for the Wilcoxon rank sum test: A large-sample approach. Biometrics, 2003; 59: 1089–98
21. Klipper-Aurbach Y, Wasserman M, Braunspiegel-Weintrob N et al: Mathematical formulae for the prediction of the residual beta cell function during the first two years of disease in children and adolescents with insulin-dependent diabetes mellitus. Med Hypotheses, 1995; 45: 486–90
22. Ashburner M, Ball CA, Blake JA et al: Gene ontology: Tool for the unification of biology. The Gene Ontology Consortium. Nat Genet, 2000; 25: 25–29
23. Huang da W, Sherman BT, Lempicki RA: Bioinformatics enrichment tools: Paths toward the comprehensive functional analysis of large gene lists. Nucleic Acids Res, 2009; 37: 1–13
24. Maere S, Heymans K, Kuiper M: BiNGO: A Cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks. Bioinformatics, 2005; 21: 3448–49
25. Vapnik VN: An overview of statistical learning theory. IEEE Trans Neural Netw, 1999; 10: 988–99
26. Ben-Hur A, Ong CS, Sonnenburg S et al: Support vector machines and kernels for computational biology. PLoS Comput Biol, 2008; 4: e1000173
27. Cismondi F, Celi LA, Fialho AS et al: Reducing unnecessary lab testing in the ICU with artificial intelligence. Int J Med Inform, 2013; 82: 345–58
28. Papaioannou VE, Chouvarda IG, Maglaveras NK et al: Temperature multiscale entropy analysis: a promising marker for early prediction of mortality in septic patients. Physiol Meas, 2013; 34: 1449–66
29. Slattery ML, Herrick JS, Mullany LE et al: Improved survival among colon cancer patients with increased differentially expressed pathways. BMC Med, 2015; 13: 75