Expression Profile Analysis Identifies Key Genes as Prognostic Markers for Metastasis of Osteosarcoma

Xiaoqing Guan (guanxiaoqing@bjmu.edu.cn)
Beijing Cancer Hospital https://orcid.org/0000-0001-8010-5715
Zhiyuan Guan
Peking University Third Hospital
Jiafu Ji
Peking University Cancer Hospital
Chunli Song
Peking University Third Hospital

Research article

Keywords: Biomarker, Osteosarcoma, Metastasis, Molecular classifier, Prognosis, Gene expression

Posted Date: January 14th, 2020

DOI: https://doi.org/10.21203/rs.2.20786/v1

Version of Record: A version of this preprint was published at Cancer Cell International on March 30th, 2020. See the published version at https://doi.org/10.1186/s12935-020-01179-x.

Abstract

Background : Osteosarcoma (OS) is the most common malignant tumor of bone which was featured with osteoid or immature bone produced by the malignant cells, and biomarkers are urgently needed to identify patients with this aggressive disease. Methods : We downloaded gene expression profiles from Gene Expression Omnibus (GEO) and The Therapeutically Applicable Research to Generate Effective Treatments (TARGET) datasets for OS, respectively, and performed weighted gene co-expression network analysis (WGCNA) to identify the key module. Whereafter, functional annotation and Gene Set Enrichment Analysis (GSEA) demonstrated the relationships between target genes and OS. Results : In this study, we discovered four key genes – ALOX5AP, HLA-DMB, HLA-DRA and SPINT2 as new prognostic markers and confirmed their relationship with OS metastasis in the validation set. Conclusions : Overall, our work may shed light on the roles of ALOX5AP, HLA-DMB, HLA-DRA and SPINT2, thus providing valuable clues to investigate the metastasis of OS and corroborating the potential clinical application value of the 4-gene signature to some extent.

Background

Osteosarcoma (OS) is the most common type of cancer that arises in bones and most people diagnosed with OS are under the age of 25 [1]. The incidence of OS in the general population is 2–3/million/year and the peak at the age of 15–19 is 8–11/million/year [2, 3]. OS is characterized by early metastasis, poor prognosis without treatment [4], and more than 90% of patients die from lung metastasis before multiple chemotherapy. OS is currently undergoing multidisciplinary treatment, with approximately 15–20% of patients showing signs of metastasis at diagnosis, most in the lungs. Metastasis remains the leading cause of death in patients with OS, compared with 70% of patients with localized disease, and only about 20% becoming long-term survivors.

Previous studies have investigated mutational alterations or gene factors in an attempt to identify candidate OS driver oncogenes or tumors suppressors [5–8]. So far, for patients with metastatic OS, neither prognostic factors nor optimal treatment methods have been well established. Therefore, more attention must be paid to more precise risk assessment, not only for patient consultation, but also for determining treatment options based on reliable stratified criteria. In order to detect pulmonary metastasis OS early and improve poor survivorship, it is important to further explore more effective prognostic biomarkers and therapeutic targets.

Although research on biomarkers for metastasis within OS has recently expanded [9–11], the targets after any OS diagnosis within the clinic and suitable for various sequencing platforms remain sparse. Recent development of gene chips and high-throughput sequencing technology, have enabled the identification of key genes related to tumor progression and prognosis based on big data integration and bioinformatics. Weighted gene co-expression network analysis (WGCNA) is a systematic biological method that could identify highly synergistically altered gene sets and screen out therapeutic targets or

candidate biomarkers based on the inherent characteristics of the gene sets and the correlation between gene sets and phenotypes.

Aiming at identifying and validating key genes in OS metastasis, the present study firstly identified associated module by WGCNA according to the gene expression profiles from GEO datasets and determined the differentially expressed genes between metastatic OS samples and non-metastatic samples. Subsequently, GO and KEGG pathway analyses were performed to determine the most significant pathways associated with OS metastasis. Additionally, we constructed KM-curves and receiver operating characteristic (ROC) curves and screened the key genes related to OS prognosis. Moreover, univariate and multivariate Cox regression analysis were conducted to evaluate the predictive effect of the gene signature. Finally, we validated the gene signature using an external RNA-Seq expression data obtained from TARGET. The results may reveal the prognostic value of the gene signature for OS (Fig. 1).

Methods

Data Sources and Data Preprocessing

We downloaded standardized matrix profile (*series matrix.txt) of GSE21257 (a microarray dataset) and obtained patient information is a microarray from the Gene Expression Omnibus (GEO) database (Table 1) [12]. The platform of the dataset is the GPL10295 Illumina human-6 v2.0 expression beadchip. We removed probes not mapping to the Gene symbol using platform annotation file. For different probes corresponding to the same gene, their median expression values were taken as the final gene expression value. Differentially expressed genes (DEGs) between OS samples with metastasis and those without metastasis were identified using the “limma” (linear models for microarray data) R package (False discovery rate (FDR) < 0.05 and absolute of log2fold change (FC) > 1) [13].

Table 1
Clinical features of patients in the training set and validation set.

	Training Set		p	Validation Set		p
	Metastasis	Non-Metastasis		Metastasis	Non-Metastasis	
Age			0.41			0.56
Median	16	18		14.05	14.37	
Gender			0.11			0.25
Female	9	10		12	25	
Male	25	9		9	38	
Grade			0.33	-	-	-
1	11	2				
2	9	7				
3	7	6				
4	3	2				

The OS RNA sequencing (RNA-seq) expression data and the corresponding clinical follow-up data were obtained from the publicly available website of the National Cancer Institute TARGET Data Matrix (https://ocg.cancer.gov/programs/target/data-matrix). To meet the requirement for data analysis, we excluded the samples with incomplete information, then 84 OS expression data were remained. Genes that have average expression (TPM > 1) between samples were deemed as expressed. The expression value was processed as log2 (TPM + 1) for subsequent analysis.

Constructing Dynamic Weighted Gene Co-Expression Network

We chose the 3000 most-varying genes for network construction and module detection. Specifically, the median absolute deviation (MAD) was used as a robust measure of variability. The network was built based on the protocols of R package WGCNA [14, 15]. We firstly clustered the samples to detect outliers. It appeared there was one outlier and we removed it by hand (Supplementary Fig. 1A). Use PickSoftThreshold function to select $\beta = 7$ (scale-free $R2 = 0.89$) to build an adjacency matrix to make our gene distribution conform to scale-free networks based on connectivity for training set (Supplementary Fig. 1B and 1C). Next, we used a blockwiseModules function to build a gene co-expression network in one step and a dynamic tree-cutting algorithm detected the modules. Set the minimum number of genes per network to 30 and the cutting height to 0.25. Meanwhile we calculated module eigengene of each module by measuring the first principal component of a specific module, which represented the overall level of gene expression within this module. Then, according to the correlation between the clinical traits and the

module eigengene and the p-value to mine the modules related to the traits, we selected the module with the highest Pearson correlation coefficient for metastasis into subsequent analysis. Finally, to find hub genes for a given module, gene significance (GS, the absolute value of the correlation between the gene and the trait) and module membership (MM, the correlation of the module eigengene and the gene expression profile) were evaluated. Based on criteria of MM > 0.8 and GS > 0.2, hub genes in the blue module were screended. Cytoscape version 3.7.2 was used for network visualization. The above analysis is implemented using the R package “WGCNA”.

Functional Annotation and Gene Set Enrichment Analysis (GSEA)

R package clusterProfiler was used to conduct Gene Ontology (GO) [16] and Kyoto Encyclopedia of Genes and Genomes (KEGG) biological pathway over representation analysis for interesting module genes [17]. GO terms and KEGG pathways with adjust $p < 0.05$ were considered statistically significant pathways. The enrichment analysis was implemented in command line of GSEA [18, 19]. An expression dataset and phenotype labels in the GSE21257 dataset were used to conduct GSEA analysis according to metastasis status (metastasis vs. non-metastasis). The data was then interrogated against reactome gene sets (1499 gene-sets) from MSigDB version 6.2 [18, 20, 21]. We set the cut-off criteria as gene set size > 15, Number of enriched gene sets that are significant, as indicated by a FDR of less than 25%.

Cox-regression based survival analysis

Univariate cox regression analysis was firstly performed to screen survival related genes. Furthermore, the receiver operating characteristic (ROC) analysis was performed to evaluate the predicting efficiency of the gene risk score and the area under curve (AUC) was calculated. The genes with p-value < 0.05 as well as AUC > 0.85 were screened as candidate genes for next analysis. These candidate genes were further selected for predictive signature construction. Risk scores were calculated and included in multivariate regression analysis in a Cox proportional hazard regression model for survival analysis. The Kaplan–Meier curve was used to visualize the survival probability for each group and p-value was calculated by the log-rank test. The survival analysis was implemented with package survival and survminer. The ROC analysis was performed using pROC package.

Statistical Analysis

Our study used a Wilcoxon rank sum test to compare continuous data between two groups. a chi-square test or Fisher's exact test to test the difference between categorical variables. A P-value < 0.05 or a adjusted P-value < 0.05 was considered statistically significant. The Kaplan–Meier method and log-rank test was used to evaluate the correlation between gene expression and overall survival. The WGCNA method was analyzed by Pearson correlation analysis. All of these processes were conducted by R software (version 3.5.1 (x64)).

Results

Identification of key modules associated with OS metastasis

After data preprocessing and quality evaluation, an expression matrix with 3000 most varying genes and 52 OS samples with clinical information in GSE21257 was used for gene co-expression network construction (Fig. 2A). After merging similar modules, we were able to identify a total of six modules and each module was designated by distinct colors to distinguish between modules (Fig. 2B). The number of genes in each module were presented in Fig. 2C. Genes in grey module were removed in the further analysis. Supplementary Fig. 1D allows us to visualize the interaction relationship of 5 modules. The representaion showed a high-scale independence degree between any two modules even between genes within each gene module. Furthermore, eigengenes of all modules were calculated and clustered based on their correlation. The plot can be found in Supplementary Fig. 1E and 1F. It is clear from this plot that the 5 modules were mainly divided into two clusters, which were consistent with the result of eigengene network heatmap. Next, relevance of all with all traits were assessed and results were presented in Fig. 2D. The highest correlation observed was for the blue module with metastasis (correlation coefficient values, -0.51; P-values, 1e-04). In addition, the turquoise module was also found to be significantly related to metastasis (correlation coefficient values, 0.36; P-values, 0.009). Overall, we focused on the 560 genes in the blue modules in subsequent analysis.

Functional annotation and GSEA

We conducted GO function and KEGG pathway analyses to examine the potential functional significance of the genes within blue module. Biological process of GO analysis showed that blue module was mainly enriched with cell migration, cell proliferation, cell cycle and immune response related pathway (Fig. 3A). Figure 3B presented the top 10 statistically significant observations of KEGG. The significant pathways included cytokine-cytokine receptor interaction, chemokine signaling pathway, toll-like receptor signaling pathway, cell differentiation, antigen processing and presentation and metabolism related pathway. In order to further understand the biological function of genes in blue module, GSEA was utilized to perform a pathway enrichment analysis and find enrichment of pathways defined by reactome. Then we found 2 gene sets (cell cycle and cell cycle mitotic) were significantly upregulated in phenotype Metastasis in reactome gene sets (Fig. 3C). The detailed results are available in Fig. 3D and 3E.

Detection of hub genes based on the training set

As described in Materials and Methods, we analyzed the blue module and plotted module membership (MM) against gene significance (GS) in Fig. 4A. All the DEGs were showed in Fig. 4B. After overlapping genes found by WGCNA and DEGs, we obtained 29 genes recognized as candidate genes. Among them, ALOX5AP, HLA-DMB, HLA-DRA and SPINT2 were negatively associated with the overall survival of OS patients (Figs. 4C). Moreover, the expression levels of these 4 genes were significantly higher in OS patients with metastasis, compared with non-metastasis patients (Fig. 4D). In addition, the diagnostic

performance of these 4 genes was evaluated by ROC curves. The AUC showed that ALOX5AP, HLA-DMB, HLA-DRA and SPINT2 indicated excellent diagnostic efficiency for patients with metastasis and those with non-metastasis (Fig. 4E). Figure 4F showed that ALOX5AP, HLA-DMB, HLA-DRA and SPINT2 were highly connected in the network and demonstrated that the 4 genes play an important role in the development of OS.

Evaluation and Validation of 4-gene signature for survival prediction

To investigate whether the 4-gene signature could provide an accurate prediction of overall survival in OS patients, the 4-gene signature risk score were calculated for each patient in the training set according to the expression of these 4 genes for OS prediction. Then patients were divided into high- and low-risk groups using the median risk score as the cutoff. As expected, risk model might be a diagnostic marker for OS with an AUC of 0.861 (Fig. 5A) and patients with high-risk scores had a poor prognosis than those with low-risk scores ($p = 0.0088$) (Fig. 5B). As such, the 4-gene signature was validated using OS data from TARGET, and we achieved consistent results. Kaplan-Meier curves revealed that the high-risk scores of 4-gene signature were significantly associated with shorter overall survival time of OS patients ($p = 0.043$) (Fig. 5C), which were similar to those observed in the training series. In order to further evaluate whether the expression levels of these four genes can provide good prognostic value, a multivariate Cox regression analysis was performed. The results can be seen in Table 2. It was evident that risk scores calculated from these four gene signature remained a strong independent prognostic factor for patients with OS ($p = 0.02$).

Table 2
Multivariate analysis adjusted for age, gender, grade, and risk score based on 4 genes signature in the training set.

	HR	p
Risk Score (Low vs. High)	0.34	0.02
Age	1.00	0.88
Gender (Male vs. Female)	1.52	0.38
Grade (3&4 vs. Unknown)	0.22	0.10
Grade (1&2 vs. Unknown)	0.61	0.56

Discussion

OS is the most common primary malignant tumor of bone, and the susceptible population is adolescents. Its prognosis is very poor, and early metastases often occur. 20% of patients died of tumor metastasis or

unresectable tumors, and the remaining 80% of patients had small metastases at the time of diagnosis. Many patients develop lung metastases within 1 year, and the 5-year survival rate is only 15%. Like many other malignancies, its etiology remains unknown. Recently, a large number of new diagnostic techniques and effective chemotherapy methods have been developed, and the current 5-year survival rate has risen to 55%-70%. Preoperative adjuvant chemotherapy followed by radical resection is still the most effective treatment. If surgical resection is not possible, radiotherapy may be beneficial for controlling local tumors. As a generality, metastasis is the most adverse factors at diagnosis among known prognostic factors [22]. There are large differences in survival between patients with metastatic OS (10–20%) and non-metastatic OS (50–78%) [26, 27]. Moreover, metastatic OS are still very difficult to control and there are few effective therapeutic targets. Kinase targets, immune checkpoint inhibitors and cell surface marker GD2 have been actively investigated in multiple current clinical trials, but are inadequately evaluated. Therefore, further studies on early diagnosis or prediction of metastasis are warranted.

In our study, multiple bioinformatics analysis tools were used to identify 4 key genes related to metastasis and prognosis of OS patients, thus we constructed a risk score model which may benefit the treatment and prognosis evaluation of OS.

Using GO, KEGG and GSEA, we annotated the function of genes in the key module most related with metastasis, and clarified the underlying mechanism of metastasis in OS. Our results revealed that these genes were found to be enriched in cell cycle, cell proliferation, cell migration and immune response. Some researchers had demonstrated the functional link between cell cycle disorder and cancer cell invasion and metastasis [13, 23, 24]. Several small pilot studies have reported that expression of molecules of tumor cell immune response, particularly HLA class II, can induce anti-tumor T cell responses, which may affect tumor progression and survival time of patients [25–27]. Hence, we suggested that genes in blue module probably involved in the development and metastasis of OS through cell cycle pathway and immune response pathway.

After screening and filtering, we obtained four genes that may predict OS metastasis and have prognostic effects, and were evaluated in a cox regression model, indicating that it is an independent prognostic factor. The 4 key genes consist of ALOX5AP, HLA-DMB, HLA-DRA and SPINT2. ALOX5AP, also called 5-LO-activating protein (FLAP), which plays an important role in Synthesis of leukotriene and associates with prognosis of primary neuroblastoma patients [28] and esophageal squamous cell carcinoma patients [29]. HLA-DMB, one of the HLA class II beta chain paralogues, is expressed in antigen-presenting cells. Previous studies have confirmed that mRNA and protein levels of HLA-DMB are highly expressed in tumor samples from patients with advanced serous ovarian cancer with a large number of tumor-penetrating CD8 T lymphocytes, which can significantly prolong the median survival time [30]. HLA-DRA, a component of MHC II, alpha chain paralogues, also are expressed in antigen presenting cells. Both at transcription and protein levels, reduced expression of HLA-DRA has been shown to predict poor overall survival and progression-free survival in diffusive large B-cell lymphoma [31]. Moreover, enrichment analysis revealed up-regulation of immune response gene sets, including antigen presentation (HLA-DMB and HLA-DRA). SPINT2 gene expression was down-regulated, altering dysregulation of the HGF/MET

signaling pathway, which contributes to cancer development and progression [32–34]. Whether these genes play the same role in the development and metastasis of OS deserves further study.

However, there were still some limitations in our work. Firstly, there are relatively small numbers of patients in two datasets obtained from publicly available database. In order to verify the stability and accuracy of the risk prediction model, more expression data and corresponding clinical information need to be collected, especially independent cohorts from multiple centers to further evaluate the applicability of the model. Secondly, our analysis is completely based on bioinformatics analysis, we need to accumulate more comprehensive experimental evidence, including in vivo and in vitro experiments Secondly, our analysis was entirely based on bioinformatics analysis to clarify the effect and possible molecular mechanisms of 4 genes on OS.

Conclusions

In summary, we found 4 genes that may play a key role in OS metastasis and prognosis, and further constructed a risk score model, which may provide new clues for the prediction of OS metastasis and establish foundation to reveal prognostic markers and treatment targets for OS patients.

Abbreviations

osteosarcoma(OS)

Gene Expression Omnibus (GEO)

Therapeutically Applicable Research to Generate Effective Treatments (TARGET)

weighted gene co-expression network analysis (WGCNA)

Gene Set Enrichment Analysis (GSEA)

receiver operating characteristic (ROC)

differentially expressed genes (DEGs)

false discovery rate (FDR)

fold change (FC)

RNA sequencing (RNA-seq)

median absolute deviation (MAD)

gene significance (GS)

module membership (MM)

Gene Ontology (GO)

Kyoto Encyclopedia of Genes and Genomes (KEGG)

area under curve (AUC)

Declarations

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable.

Availability of data and materials

The study data is available at GEO (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE21257) and TARGET (https://ocg.cancer.gov/programs/target/data-matrix and choose “OS”).

Competing interests

The authors report no conflicts of interest in this work.

Funding

This work was supported by grants from the National Natural Science Foundation of China (Grant Number 81672133).

Authors' contributions

X.G. and Z.G. designed and performed analysis. X.G. drafted the manuscript. C.S. and J.J. offered critical revision of the manuscript. All authors reviewed and approved of the final draft of the manuscript.

Acknowledgements

Not applicable

References

1. Arndt, V., et al., *Up-to-date monitoring of childhood cancer long-term survival in Europe: tumours of the sympathetic nervous system, retinoblastoma, renal and bone tumours, and soft tissue sarcomas.* Ann Oncol, 2007. **18**(10): p. 1722-33.

2. Collins, M., et al., *Benefits and adverse events in younger versus older patients receiving neoadjuvant chemotherapy for osteosarcoma: findings from a meta-analysis.* J Clin Oncol, 2013. **31**(18): p. 2303-12.
3. Ritter, J. and S.S. Bielack, *Osteosarcoma.* Ann Oncol, 2010. **21 Suppl 7**: p. vii320-5.
4. Mirabello, L., R.J. Troisi, and S.A. Savage, *International osteosarcoma incidence patterns in children and adolescents, middle ages and elderly persons.* Int J Cancer, 2009. **125**(1): p. 229-34.
5. Bousquet, M., et al., *Whole-exome sequencing in osteosarcoma reveals important heterogeneity of genetic alterations.* Ann Oncol, 2016. **27**(4): p. 738-44.
6. Kansara, M., et al., *Translational biology of osteosarcoma.* Nature Reviews Cancer, 2014. **14**(11): p. 722-735.
7. Subbiah, V., et al., *Alpha Particle Radium 223 Dichloride in High-risk Osteosarcoma: A Phase I Dose Escalation Trial.* Clin Cancer Res, 2019. **25**(13): p. 3802-3810.
8. Martinez-Velez, N., et al., *The Oncolytic Adenovirus VCN-01 as Therapeutic Approach Against Pediatric Osteosarcoma.* Clin Cancer Res, 2016. **22**(9): p. 2217-25.
9. Tian, H., D. Guan, and J. Li, *Identifying osteosarcoma metastasis associated genes by weighted gene co-expression network analysis (WGCNA).* Medicine (Baltimore), 2018. **97**(24): p. e10781.
10. Scott, M.C., et al., *Comparative Transcriptome Analysis Quantifies Immune Cell Transcript Levels, Metastatic Progression, and Survival in Osteosarcoma.* Cancer Res, 2018. **78**(2): p. 326-337.
11. Cortini, M., S. Avnet, and N. Baldini, *Mesenchymal stroma: Role in osteosarcoma progression.* Cancer Lett, 2017. **405**: p. 90-99.
12. Barrett, T., et al., *NCBI GEO: archive for functional genomics data sets—update.* Nucleic Acids Research, 2012. **41**(D1): p. D991-D995.
13. Ritchie, M.E., et al., *limma powers differential expression analyses for RNA-sequencing and microarray studies.* Nucleic Acids Research, 2015. **43**(7): p. e47-e47.
14. Langfelder, P. and S. Horvath, *WGCNA: an R package for weighted correlation network analysis.* BMC Bioinformatics, 2008. **9**(1): p. 559.
15. Langfelder, P. and S. Horvath, *Fast R Functions for Robust Correlations and Hierarchical Clustering.* 2012, 2012. **46**(11): p. 17 %J Journal of Statistical Software.
16. *The Gene Ontology Resource: 20 years and still GOing strong.* Nucleic Acids Res, 2019. **47**(D1): p. D330-d338.
17. *clusterProfiler: an R Package for Comparing Biological Themes Among Gene Clusters.* 2012. **16**(5): p. 284-287.
18. Subramanian, A., et al., *Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles.* 2005. **102**(43): p. 15545-15550.
19. Mootha, V.K., et al., *PGC-1α-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes.* Nature Genetics, 2003. **34**(3): p. 267-273.

20. Fabregat, A., et al., *The Reactome Pathway Knowledgebase.* Nucleic Acids Res, 2018. **46**(D1): p. D649-d655.
21. Liberzon, A., et al., *The Molecular Signatures Database Hallmark Gene Set Collection.* Cell Systems, 2015. **1**(6): p. 417-425.
22. Smeland, S., et al., *Survival and prognosis with osteosarcoma: outcomes in more than 2000 patients in the EURAMOS-1 (European and American Osteosarcoma Study) cohort.* Eur J Cancer, 2019. **109**: p. 36-50.
23. Kohrman, A.Q. and D.Q. Matus, *Divide or Conquer: Cell Cycle Regulation of Invasive Behavior.* Trends Cell Biol, 2017. **27**(1): p. 12-25.
24. Otto, T. and P. Sicinski, *Cell cycle proteins as promising targets in cancer therapy.* Nat Rev Cancer, 2017. **17**(2): p. 93-115.
25. Oldford, S.A., et al., *Tumor cell expression of HLA-DM associates with a Th1 profile and predicts improved survival in breast carcinoma patients.* Int Immunol, 2006. **18**(11): p. 1591-602.
26. Gonzalez, H., C. Hagerling, and Z. Werb, *Roles of the immune system in cancer: from tumor initiation to metastatic progression.* Genes Dev, 2018. **32**(19-20): p. 1267-1284.
27. Janssen, L.M.E., et al., *The immune system in cancer metastasis: friend or foe?* J Immunother Cancer, 2017. **5**(1): p. 79.
28. Sveinbjornsson, B., et al., *Expression of enzymes and receptors of the leukotriene pathway in human neuroblastoma promotes tumor survival and provides a target for therapy.* Faseb j, 2008. **22**(10): p. 3525-36.
29. Wu, B., et al., *The arachidonic acid metabolism protein-protein interaction network and its expression pattern in esophageal diseases.* Am J Transl Res, 2018. **10**(3): p. 907-924.
30. Callahan, M.J., et al., *Increased HLA-DMB expression in the tumor epithelium is associated with increased CTL infiltration and improved prognosis in advanced-stage serous ovarian cancer.* Clin Cancer Res, 2008. **14**(23): p. 7667-73.
31. Brown, P.J., et al., *FOXP1 suppresses immune response signatures and MHC class II expression in activated B-cell-like diffuse large B-cell lymphomas.* Leukemia, 2016. **30**(3): p. 605-16.
32. Kongkham, P.N., et al., *An epigenetic genome-wide screen identifies SPINT2 as a novel tumor suppressor gene in pediatric medulloblastoma.* Cancer Res, 2008. **68**(23): p. 9945-53.
33. Roversi, F.M., S.T. Olalla Saad, and J.A. Machado-Neto, *Serine peptidase inhibitor Kunitz type 2 (SPINT2) in cancer development and progression.* Biomed Pharmacother, 2018. **101**: p. 278-286.
34. Yue, D., et al., *Epigenetic inactivation of SPINT2 is associated with tumor suppressive function in esophageal squamous cell carcinoma.* Exp Cell Res, 2014. **322**(1): p. 149-58.

Supplemental Figure Legend

Supplementary Figure 1. Network construction and module detection

(A) Clustering dendrogram of samples based on their Euclidean distance. (B) Analysis of the scale-free fit index and mean connectivity for various soft-thresholding powers (β). Panels illustrate the scale-free fit index (y-axis) as a function of the soft-thresholding power (x-axis). Solid red horizontal lines are guides of the index at 0.9. At the power = 7, the index curve flattened out upon reaching the higher value in all groups. Effects of power values on the scale independence of genes co-expression modules for OS. (C) Effects of power values on the average connectivity of genes co-expression modules for OS. The panel displays the mean connectivity (degree, y-axis) as a function of the soft-thresholding power (x-axis). (D) Analysis of relationship between pairwise gene co-expression modules. Different colors of horizontal axis and vertical axis represent different modules. The brightness of yellow in the middle represents the degree of connectivity of different modules. There was no significant difference in interactions among different modules, indicating a high-scale independence degree among these modules. The modules in the horizontal and vertical axes were marked with different colors. The degree of the yellow brightness indicated the relevance. The overall relationship between the different modules was small, indicating that the modules had a high degree of independence. (E and F) The modules produced in the clustering analysis were summarized module eigengene dendrogram (E) and eigengene network heatmap (F). The eigengenes were mainly clustered into two clusters, containing 2 modules (modules green and blue) and 3 modules (modules brown, turquoise and yellow), respectively.

Figures

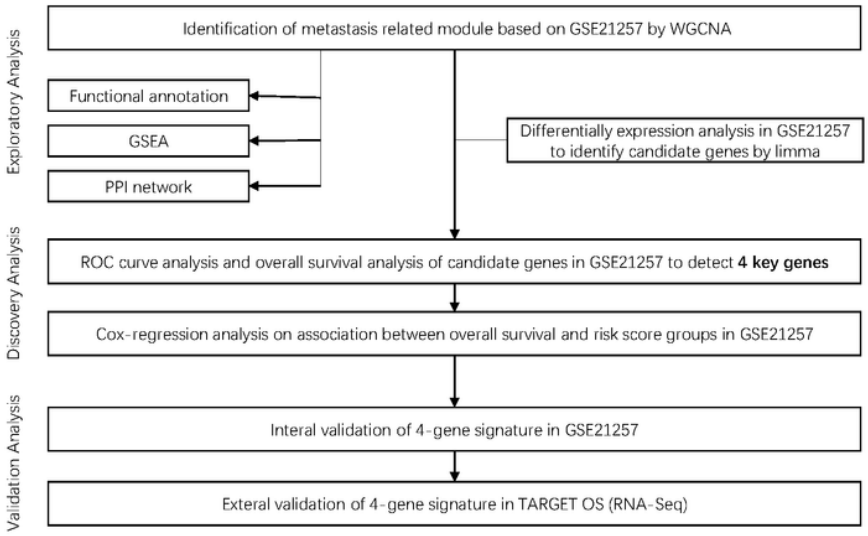

Figure 1

Flow chart of study design.

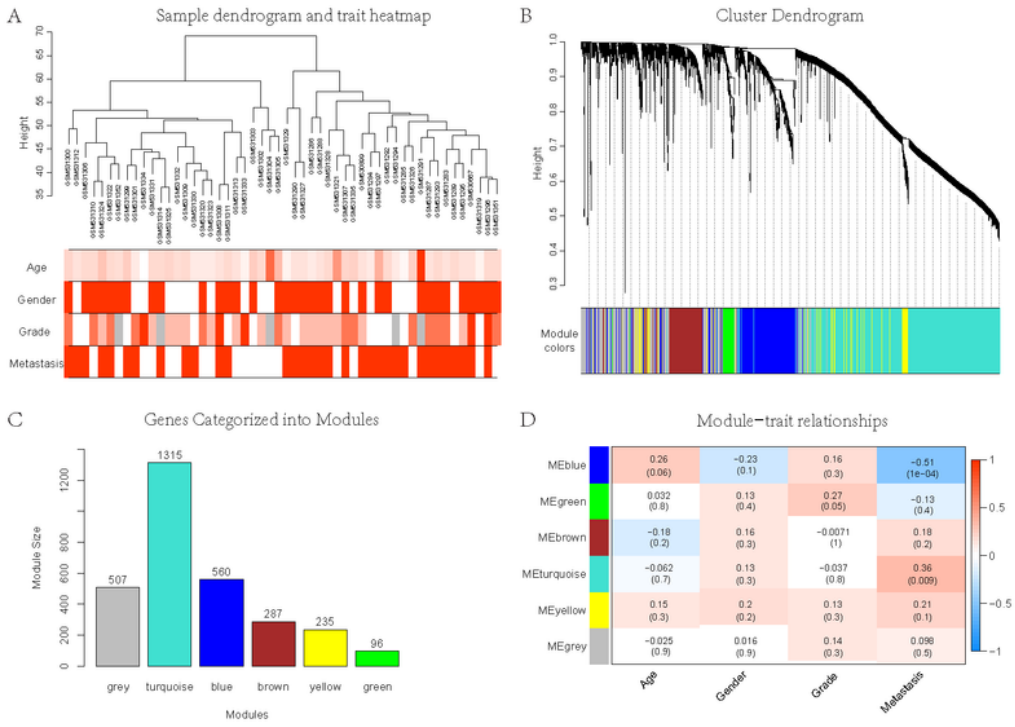

Figure 2

Construction and identification of modules associated with the clinical traits. (A) Clustering dendrogram of OS samples and the clinical traits. For age and grade, white means a low value, red a high value, and

grey a missing entry; for gender and metastasis, white means female or non-metastasis, red means male or metastasis. (B) Hierarchical clustering based on the dynamic tree, each branch above represented a gene, and each color below represented a gene co-expression module. Grey module color is a reserved one for genes that are not part of any module. (C) Number of genes in different gene co-expression modules. Note that genes in the grey module were identified as not co-expressed. (D) Heatmap of the correlation between module eigengenes and clinical traits. Each row corresponds to a module eigengene, column to a trait. Each cell contains the corresponding correlation and p-value. The table is color-coded by correlation according to the color legend. The blue module was significantly correlated with metastasis.

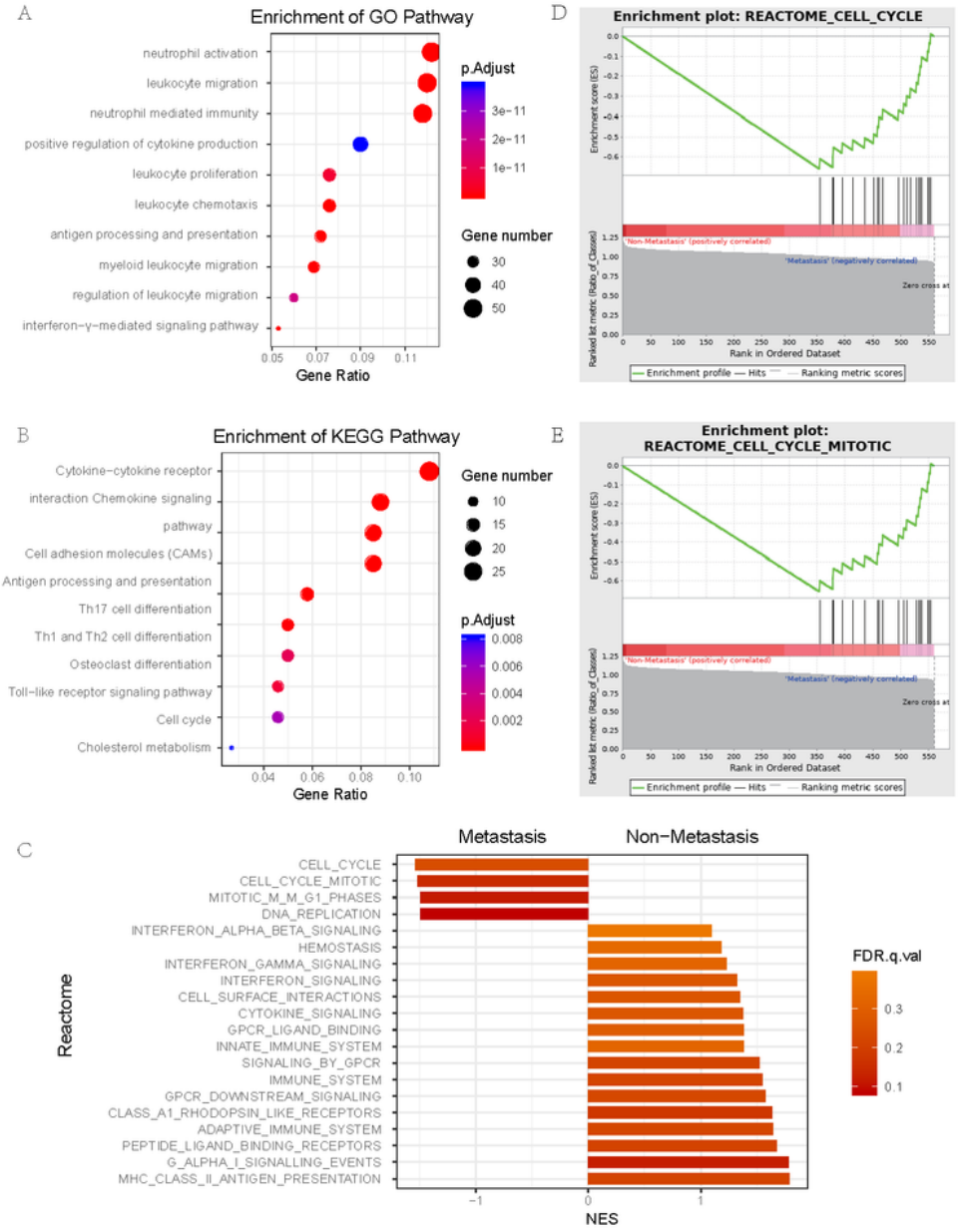

Figure 3

Functional enrichment analysis of blue module. (A) GO analysis of all genes in blue module. (B) KEGG pathway analysis of all genes in blue module. (C GSEA in reactome gene sets. (D) An enrichment plot for cell cycle and cell cycle mitotic gene set.

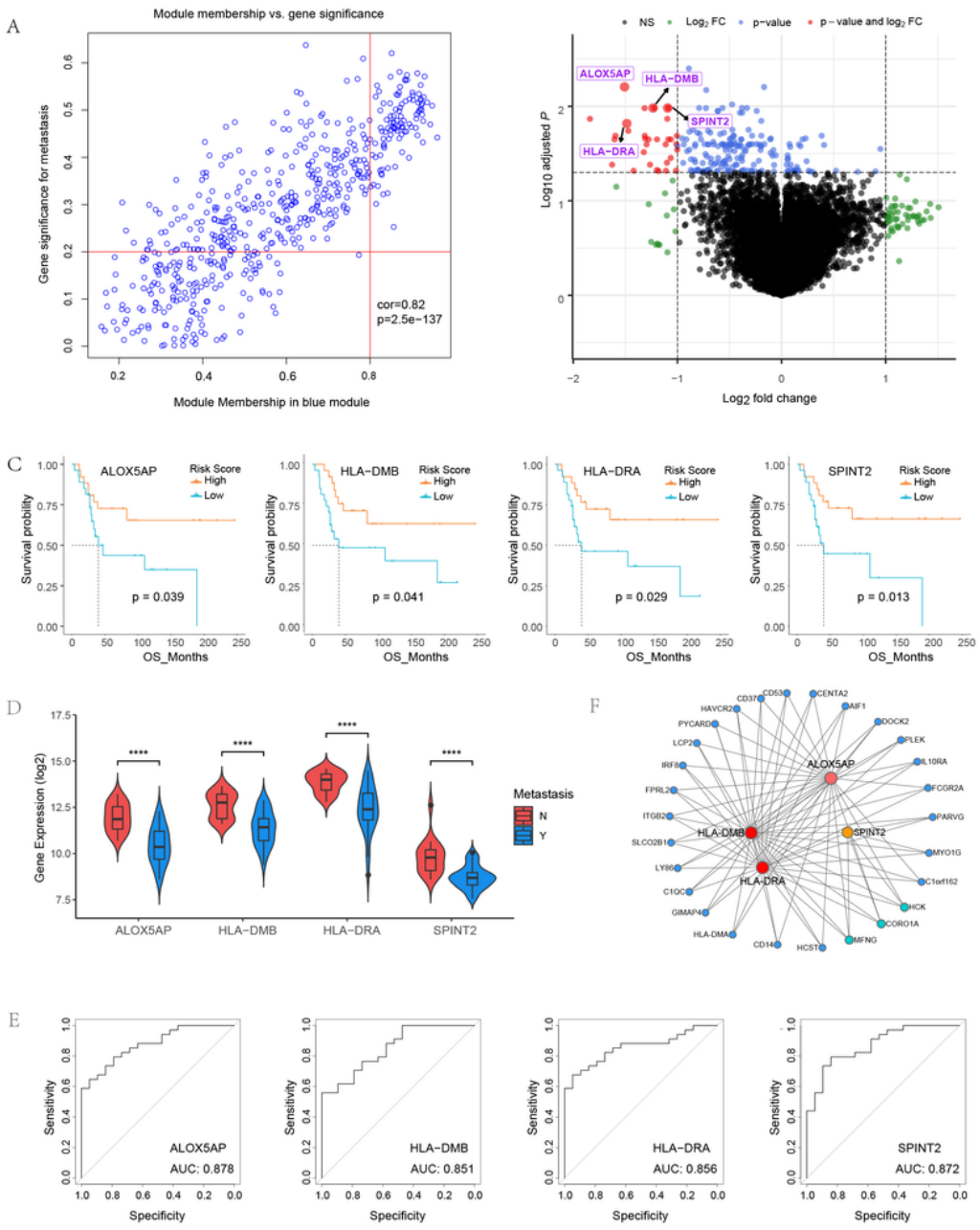

Figure 4

Identification of key genes based on training set. (A) A scatterplot of Gene Significance (GS) for weight vs. Module Membership (MM) in the blue module. There is a highly significant correlation between GS and MM in this module. (B) Volcano plot of significance of gene expression difference between metastasis and non-Metastasis patients. A gene is considered significantly differentially expressed if its |log(FC)| > 1 and p-value < 0.05. (C) Overall survival analysis of 4 key genes. Expression levels of

ALOX5AP, HLA-DMB, HLA-DRA and SPINT2 are significantly related to the overall survival of patients with OS ($P < 0.05$). (D) Boxplot of significance of gene expression levels of 4 key genes. ALOX5AP, HLA-DMB, HLA-DRA and SPINT2 are significantly downregulated in metastasis OS compared with non-metastasis OS. The **** represents $P < 0.0001$. (E) ROC curves analysis of 4 key genes diagnosis. ROC curves and AUC statistics are used to evaluate the capacity to discriminate OS with or without metastasis with excellent specificity and sensitivity. (F) The network illustrates the relationship of 4 key genes and the 36 most frequently altered neighbor genes. The 4 key genes are presented in red and orange depending on the gene importance defined as the degree of connectivity. The other genes are represented in blue and green.

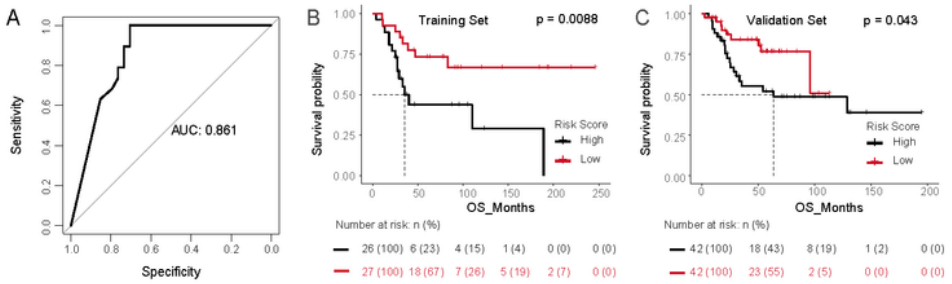

Figure 5

Evaluation and validation of the 4-gene signature risk model of OS. (A) The ROC curves are shown for risk score model in training set. (B and C) Kaplan-Meier analysis for the overall survival of OS patients in training set (B) and validation set (C).

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- suppFig1.pdf