

Identification of 6 gene markers for survival prediction in osteosarcoma cases based on multi-omics analysis

Runmin Li¹, Guosheng Wang², ZhouJie Wu¹, HuaGuang Lu¹, Gen Li¹, Qi Sun¹ and Ming Cai¹ 

¹Department of Orthopaedics, Shanghai Tenth People's Hospital, Tongji University, School of Medicine, Shanghai 200072, China;

²Institute of Translational Medicine, Zhejiang University School of Medicine, Hangzhou 310029, China

Corresponding author: Ming Cai. Email: cmdoctor@tongji.edu.cn

Impact statement

In the present study, 512 prognosis-related genes, 336 copies of amplified genes, and 36 copies of deleted genes were obtained. Six characteristic genes (i.e. MYC, CHIC2, CCDC152, LYL1, GPR142, and MMP27) were obtained by Lasso feature selection and stepwise multivariate regression study. 6-gene sign was one single prognosis-related element for osteosarcoma cases, and the samples were able to be risk stratified inside the set of training, test set, and externally validating set ($P < 0.01$). Here, 6-gene sign was built to be novel prognosis-related markers for predicting osteosarcoma cases' surviving state.

Abstract

Multiple-omics sequencing information with high-throughput has laid a solid foundation to identify genes associated with cancer prognostic process. Multiomics information study is capable of revealing the cancer occurring and developing system according to several aspects. Currently, the prognosis of osteosarcoma is still poor, so a genetic marker is needed for predicting the clinically related overall survival result. First, Office of Cancer Genomics (OCG Target) provided RNASeq, copy amount variations information, and clinically related follow-up data. Genes associated with prognostic process and genes exhibiting copy amount difference were screened in the training group, and the mentioned genes were integrated for feature selection with least absolute shrinkage and selection operator (Lasso). Eventually, effective biomarkers received the screening process. Lastly, this study built and demonstrated one gene-associated prognosis mode according to the set of the

test and gene expression omnibus validation set; 512 prognosis-related genes ($P < 0.01$), 336 copies of amplified genes ($P < 0.05$), and 36 copies of deleted genes ($P < 0.05$) were obtained, and those genes of the mentioned genomic variants display close associations with tumor occurring and developing mechanisms. This study generated 10 genes for candidates through the integration of genomic variant genes as well as prognosis-related genes. Six typical genes (i.e. MYC, CHIC2, CCDC152, LYL1, GPR142, and MMP27) were obtained by Lasso feature selection and stepwise multivariate regression study, many of which are reported to show a relationship to tumor progressing process. The authors conducted Cox regression study for building 6-gene sign, i.e. one single prognosis-related element, in terms of cases carrying osteosarcoma. In addition, the samples were able to be risk stratified in the training group, test set, and externally validating set. The AUC of five-year survival according to the training group and validation set reached over 0.85, with superior predictive performance as opposed to the existing researches. Here, 6-gene sign was built to be new prognosis-related marking elements for assessing osteosarcoma cases' surviving state.

Keywords: Bioinformatics, copy amount variation, prognostic marker, gene expression omnibus, osteosarcoma

Experimental Biology and Medicine 2021; 246: 1512–1523. DOI: 10.1177/1535370221992015

Introduction

Osteosarcoma refers to a highly heterogeneous bone malignant tumor originating in interstitial tissue and usually occurring in children, adolescents, and the elderly.^{1,2} The incidence rate in 24 years is 4.4 parts per million; for the second age peak, the disease is commonly secondary, accompanying with other lesions of bones or Paget's

disease.³ On the whole, osteosarcoma receives the diagnosis under a significantly late phase since the initial sign exhibited by a malignant tumor refers to pain, generally misguided in terms of more frequent diseases (e.g. "growth pain"⁴) Late osteosarcoma detection causes 15–25% of cases showing far metastases during diagnosis.⁵ The five-year survival rate of cases without metastases detected during

the diagnosis was 70%.⁶ The main treatment in terms of osteosarcoma refers to surgery and chemotherapy combination. However, overall survival remains unoptimistic in terms of cases carrying metastatic or recurrent diseases, and it continues to be 20% in terms of the last three decades.⁷ Despite advancements for surgery technologies, targeted therapies and oncology, the infection, inconvenience, complications, and low survival rates of limb rescue operations should be urgently improved. For osteosarcoma cases, the prognosis biomarking elements should be found for helping clinicians assess clinically related results and provide reference for individualized medical treatment in an accurate manner.

Cancer occurring and developing processes are regulated by genetically and epigenetically related events,⁸ and genetic variations are critical to childhood cancer. Childhood cancer usually features the small mutation load. Molecular subdivided types, clinically related heterogeneity, and disease progression prediction under several conditions display a relationship to genomic variations. The genetic component of osteosarcoma has been studied in depth,⁹ and the increasing focus on channel analysis and genetics has provided new potential biomarkers for disease. For instance, Cheng *et al.* identified genomic copy amount variation in pediatric osteosarcoma as a predictive biomarker for therapeutic response.¹⁰ Single nucleotide polymorphism variation in NFIB affects osteosarcoma cell migration and proliferation and displays associations with certain lineage metastasis.¹¹ miR-214 is up-regulated in osteosarcoma and independently assesses progression-free and overall survival.¹² A locus at 14q32 associated with miR-544, miR-134, and miR-382 has demonstrated one inverting relationship of aggressive tumor characteristics as well as residual expression of microRNAs.¹³ However, the mentioned findings comply with a single omics level, and no multi-omics integration analysis has been conducted to identify relevant biomarkers in osteosarcoma.

In this study, to find one dependable prognostic process of gene sign with the relationship to osteosarcoma in an effective manner, this study proposed a system of pipes to identify genetic markers for osteosarcoma related from Target and GEO database in large data set to obtain the gene expression patterns in osteosarcoma cases, copy amount variation data. Six-gene signs were created by screening prognosis marking elements through the integration of genomic and transcriptomic information, and this study, with the use of internally-related test sets and externally validating sets, demonstrated their capability for predicting surviving state. The 6-gene sign was reported to impact vital biologically related procedures and channels in osteosarcoma. Furthermore, the GSEA study suggested consistent outcomes, demonstrating that the 6-gene sign is capable of predicting the prognostic risk of the osteosarcoma prognostic process' molecular system.

Materials and methods

Information collecting and processing

RNA-Seq fragments per kilobase million (FPKM) data pertaining to Target contain total 101 samples and clinically

Table 1. Clinical information statistics of three data sets.

Characteristic	TARGET training datasets (n = 76)	TARGET test set (n = 84)	GSE21257 (n = 53)
Age(years)			
≤18	59	66	35
>18	17	18	18
Survival status			
Living	49	55	30
Dead	27	29	23
Gender			
Female	33	37	19
Male	43	47	34
Race			
White	44	51	–
Other	13	13	–
Unknown	19	20	–
Metastatic			
metastatic	21	21	34
Non-metastatic	55	63	19

related follow-up data contain 92 samples. In addition, SNP chips 6.0's copy amount change information contains 88 samples which were downloaded from Office of Cancer Genomics (<https://ocg.cancer.gov/programs/target/data-matrix>). Overall, the authors used gene expression omnibus (GEO) to download 53 samples of standardized expressing state profiles and clinically related data according to the GSE21257¹⁴ dataset on 5 October 2019. In terms of TCGA RNAseq information, 84 tumor specimens containing follow-up data and simultaneous detection of copy amount variations (CNVs) were selected. Furthermore, 90% samples received the random selection to be the training group (n = 76). Moreover, the remaining samples received the selection to be the test set (n = 84). GSE21257 data set to be the externally verifying set. Table 1 lists the respective group's sample data. The workflow is shown in Figure 1.

Univariate Cox proportional hazards regression study

The authors conducted Univariate Cox proportional hazard regression study for the respective gene by referencing Jin-Cheng *et al.*¹⁵ for screening the genes displaying obvious relationships to overall survival (OS) according to the set of training information, and $P < 0.01$ was chosen to be the threshold.

Copy amount variation data analysis

Genomic Identification of Significant Targets in Cancer (GISTIC) has been extensively employed, which identifies focal and broad (probably overlapping) recurrence events. The CNV seg file data of the Target sample acted as input, and GISTIC 2.0¹⁶ software was used for identifying genes obviously amplified or deleted. The threshold of the parameter referred to one part exhibiting the amplified or deleted length over 0.1 and $P < 0.05$.

Methylation analysis

We downloaded the methylation dataset of Illumina Infinium Humanation450 Microarray platform and deleted

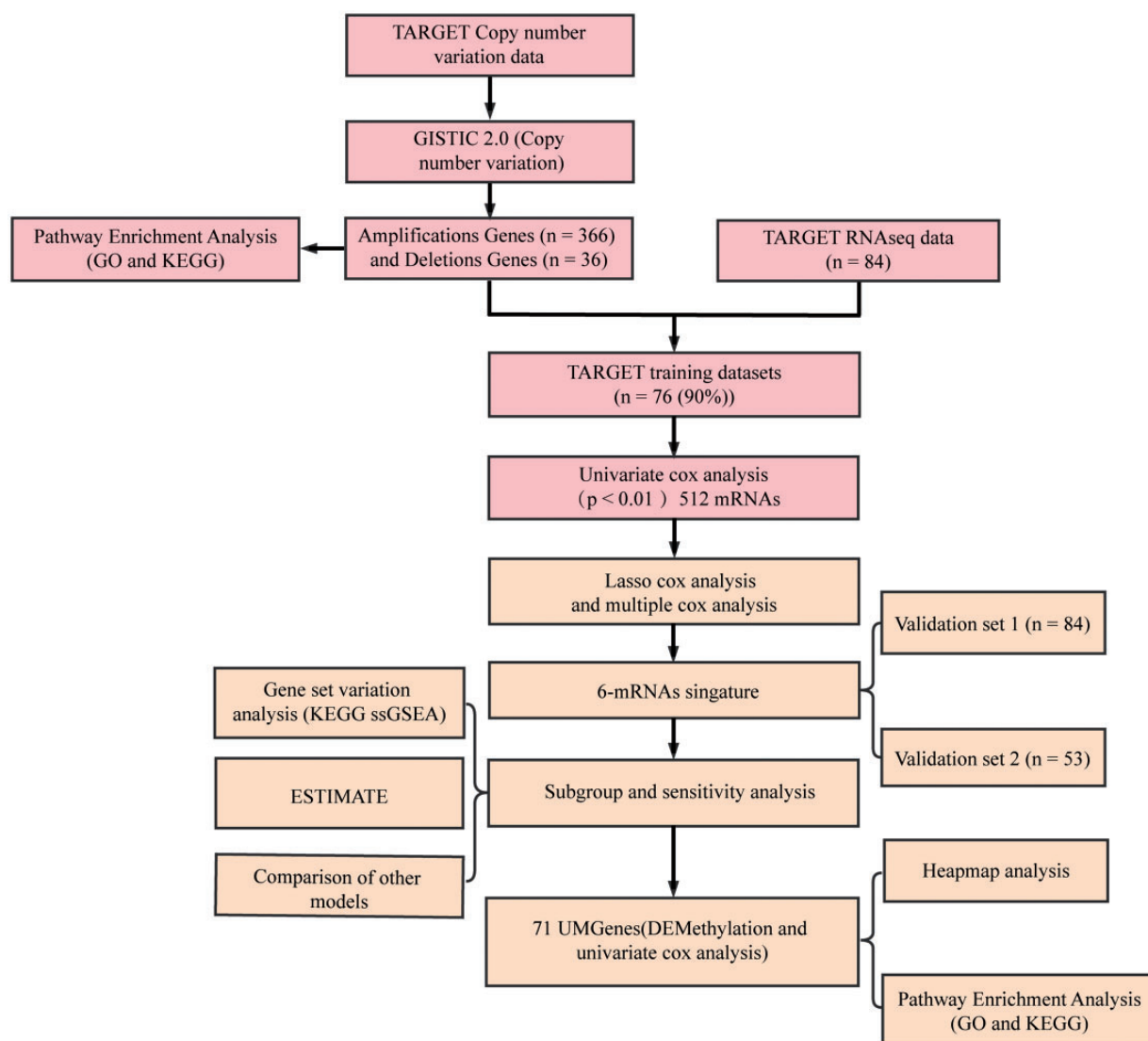


Figure 1. Work flow chart. (A color version of this figure is available in the online journal.)

the methylation sites with NA samples greater than 30%. Meanwhile, according to a previous study,¹⁷ a locality of CpG that is cross-reactive in the genome was stripped. A total of 206,351 methylation sites were obtained with the use of the KNN method pertaining to R software package impute. Furthermore, the authors employed the *t*-test for analyzing the difference in the distribution of the methylation average β value of the low-risk and high-risk groups.

Constructing process for prognosis-related gene sign

Least absolute shrinkage and selection operator (Lasso) refers to one commonly used approach to mode based on regression, exhibiting considerable probable prognosis characteristics, since it is capable of automatically selecting features, through which signatures exhibiting overall high prognosis performance are caused.¹⁸ The LASSO approach receives the extension into the Cox mode to conduct surviving state study and is well employed to build sparse signatures in terms of surviving state prognostic process under numerous use areas (e.g. oncology).^{19–21} To be

specific, the genes exhibiting significant relationships to the prognosis of the cases and the genes that were amplified or deleted were selected, and the characteristics of the prognosis genes were further selected using Lasso regression. The authors carried out further stepwise multiple-variate Cox regression study, and the risks below were constructed

$$RiskScore = Zscore \left(\sum_{k=1}^n Exp_k * e^{HR_k} \right)$$

where e^{HR_k} represents the regression coefficient estimated of genes in the multivariate Cox regression study; Exp_k denotes the expression value of prognostic genes; n denotes prognostic genes' number.

Functional enrichment analyses

Gene ontology (GO) and Kyoto encyclopedia of genes and genomes (KEGG) channel enrichment study was

performed by employing the R package clusterprofiler²² for genes to identify over-represented GO terms in three types (i.e. cellular component, molecular function, and biological processes), as well as KEGG channel. In terms of the analyses, an FDR < .05 exhibited statistical significance.

Gene set enrichment analyses (GSEA)²³ were conducted by employing the JAVA program (<http://software.broadinstitute.org/gsea/downloads.jsp>) with the MSigDB²⁴ C2 Canonical channels gene set collection, containing 1320 gene sets. Gene sets with a false discovery rate (FDR) lower than 0.05 after performing 1000 permutations were considered significantly enriched.

Comparison with existing modes

Lastly, two prognosis-related risk modes (i.e. 8-gene sign²⁵ and 4-pseudogene sign²⁶) were selected for comparison with our 6-gene mode by reviewing the literature. To make the modes comparable, this study determined the risk score for the respective sample in TCGA based on the identical method to assess the ROC and KM survival curves for the respective mode based on the corresponding genes in the two modes.

Statistical analysis

ESTIMATE_{Score}, Immune_{Score}, and Stromal_{Score} were analyzed using ESTIMATE²⁷ package. Kaplan–Meier (KM) curves were plotted when 0 acted as a cutoff in each data set to compare the risk of survival between the high-risk group and the low-risk group. Multivariate Cox regression study was performed to test whether gene markers are independent prognosis-related elements. Significance was defined as $P < 0.05$. The C index and the restricted mean survival (RMS) curve were estimated with the R package survRM2, while the C index²⁸ was compared by using the R package compareC. All the mentioned were performed in R 3.4.3.

Results

Identification of the total set of genes associated with patient survival

The relationship between the cases' OS and gene expressions was developed by conducting the univariate regression study in the training group samples, and 512 genes with P value less than 0.01 were identified, in which 357 genes with $HR > 1$ and 155 genes with $HR < 1$, and the top20 gene is listed in Table 2. The workflow is illustrated in Figure 1.

Gene set for identifying genomic variation

Specific to the CNV data in the training group, GISTIC 2.0 was adopted to identify genes with significant amplification or deletion, with a total of 20 regions identified for significant amplification and 22 deletion fragments (Figure 2(a)). The mentioned fragments involved numerous genes displaying close relationships to tumor (e.g. MYC significantly amplified in segment 8q24.21 (q value = 2.31E-16), PTGFR significantly amplified in segment 1p31.1 (q value = 0.00063), and LYL1 significantly

Table 2. Top20 gene with OS.

Gene	P value	HR	Low 95%CI	High 95%CI
COL13A1	4.97E-07	1.018	1.011	1.025
SLC8A3	1.31E-06	1.024	1.014	1.034
RHBDL2	1.54E-06	1.013	1.008	1.018
CGREF1	1.55E-06	1.013	1.008	1.019
DNAI1	2.56E-06	1.555	1.293	1.868
SMPD3	2.88E-06	1.021	1.012	1.030
CORT	3.49E-06	1.044	1.025	1.063
AOC3	6.96E-06	1.043	1.024	1.063
COL22A1	1.52E-05	1.008	1.004	1.012
KIF25	1.75E-05	1.130	1.069	1.195
GALNT14	2.10E-05	1.026	1.014	1.037
FKBP11	2.12E-05	1.010	1.005	1.015
GRAMD1B	2.51E-05	1.050	1.026	1.074
TBRG1	2.88E-05	1.037	1.020	1.055
RIPPLY2	3.12E-05	1.452	1.218	1.731
CHMP4C	5.16E-05	1.038	1.020	1.058
PROSER2	5.28E-05	1.097	1.049	1.147
DLX2	5.48E-05	1.096	1.048	1.146
BMP8A	6.17E-05	1.055	1.028	1.083
SGMS2	6.64E-05	1.014	1.007	1.021

amplified in segment 19p13.2 (q value = 0.014)), a total of 336 genes. Significant deletions in the genome (e.g. CDKN2A) were significantly absent on the 9p21.3 segment (q value = 6.78E-07), and DLG2 was obviously absent on the 11q14.1 segment (q value = 6.84E-06). In addition, RB1 was significantly absent on the 13q14.2 segment (q value = 0.00012) and contained a total of 36 genes. The first 10 regions where the copy amount variation was high are presented in Figure 2(b).

Functional analysis of genomic variant genes

For the analysis of the function of the genomic variant gene, a total of 372 amplified and deleted genes were identified by integrating copy amount variation. GO biological process and KEGG functional enrichment study were performed on the 372 genes. The results of the KEGG enrichment study revealed important channels for developing various cancers related to cell cycle, PI3K-Akt signaling channel, prostate cancer, glioma, bladder cancer, non-small cell lung cancer (Figure 3(a)). The KEGG enrichment channel found that the mentioned significantly enriched channels are highly aggregated (Figure 3(b)), suggesting that considerable copy amount variant genes were shared between the mentioned channels. In the biological process category, protein processing, extracellular structure organization, extracellular matrix organization, and fibroblast proliferation were primarily enriched (Figure 3(c)). The mentioned terms were mostly related to protein processing and display close associations with the development of cancer. In short, the genes of the mentioned genomic variants display close associations with tumors.

Identification of a six-gene sign for osteosarcoma survival

First, genes associated with genomic variation and prognosis were integrated, and 10 genes were obtained by

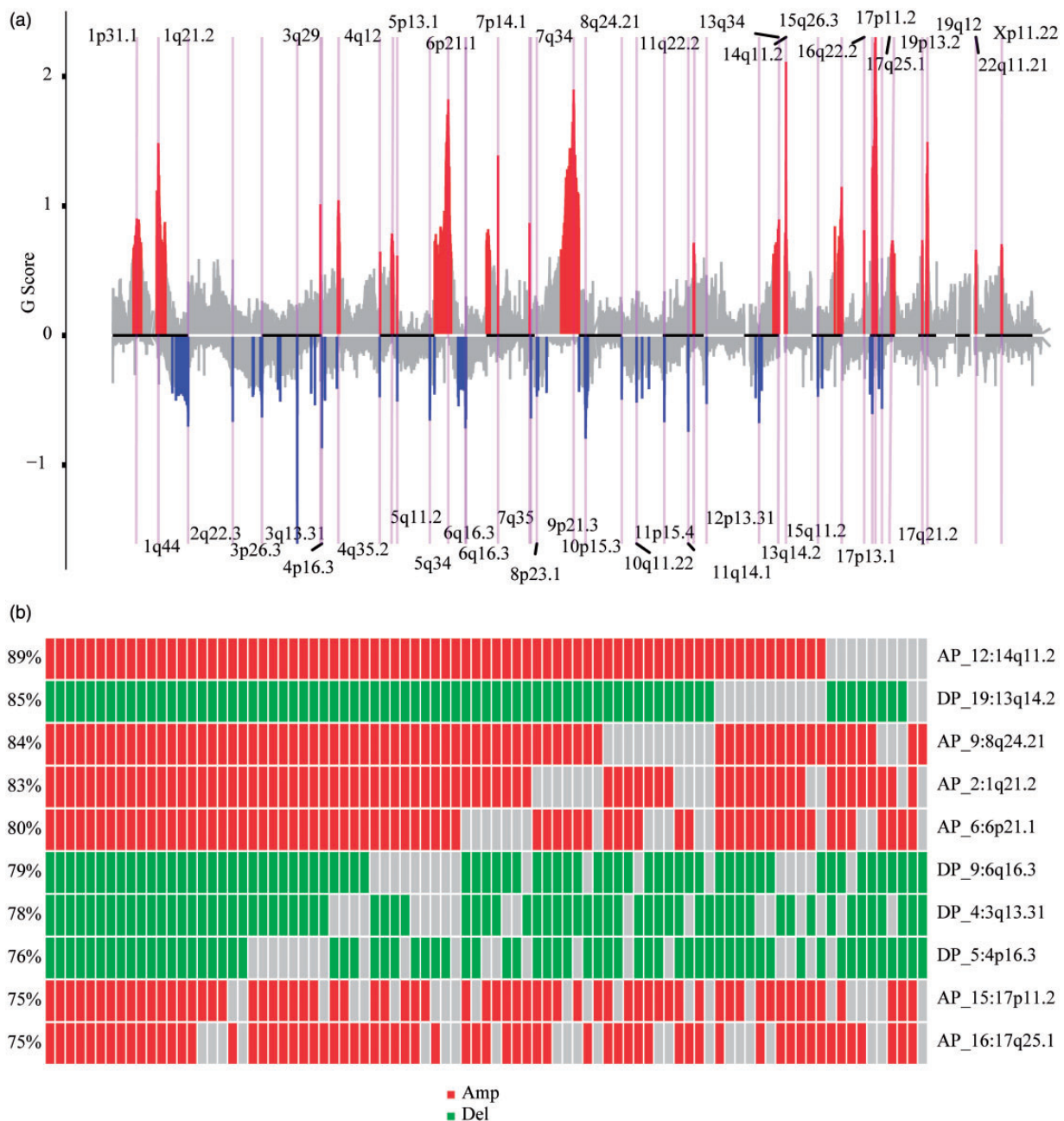


Figure 2. Gene set for identifying genomic variation. (a) Fragments under obvious amplification and significant deletion inside the osteosarcoma genome. The ordinate represents the copy amount change gene score on the respective chromosome, while the abscissa denotes the number of chromosome. Red denotes amplification of copy amount, and blue refers to copy amount deletion. Gray indicates that the copy amount change is not significant, and the top and bottom of the graph represent the chromosomal location. (b) The heat map of CNV in the osteosarcoma genome, the left side refers to the percent change in copy amount, the abscissa denotes the sample, the right side is the copy amount deletion or amplification and the position on the chromosome, red represents the amplification of copy amount, and green denotes the deletion of copy amount. (A color version of this figure is available in the online journal.)

selecting the intersection of the two groups as candidate genes. The track of the respective single variate was analyzed, and with the increase in the lambda, the number of independent coefficients was found to tend to zero (Figure 4(a)). Mode construction was performed by 10-fold cross-validation, and the confidence interval based on the respective lambda was analyzed, and the mode was optimal when $\lambda = 0.0490$ (Figure 4(b)). Nine genes at the time of $\lambda = 0.0490$ were selected as the target gene. Further, a 6-gene sign was established using a stepwise multivariate Cox regression study with a

minimum AIC value ($AIC = 180.40$). The mode is

$$\begin{aligned} \text{Risk Score}_6 = & z\text{-score}(0.4441 * \exp^{\text{MYC}} \\ & - 0.5052 * \exp^{\text{CHIC2}} - 0.5059 * \exp^{\text{CCDC152}} \\ & - 0.5202 * \exp^{\text{LYL1}} + 0.6294 * \exp^{\text{GPR142}} \\ & + 1.5615 * \exp^{\text{MMP27}}) \end{aligned}$$

The risk score of the respective sample was determined, and this study identified the relationship between the expression of the six genes and the risk score. The great

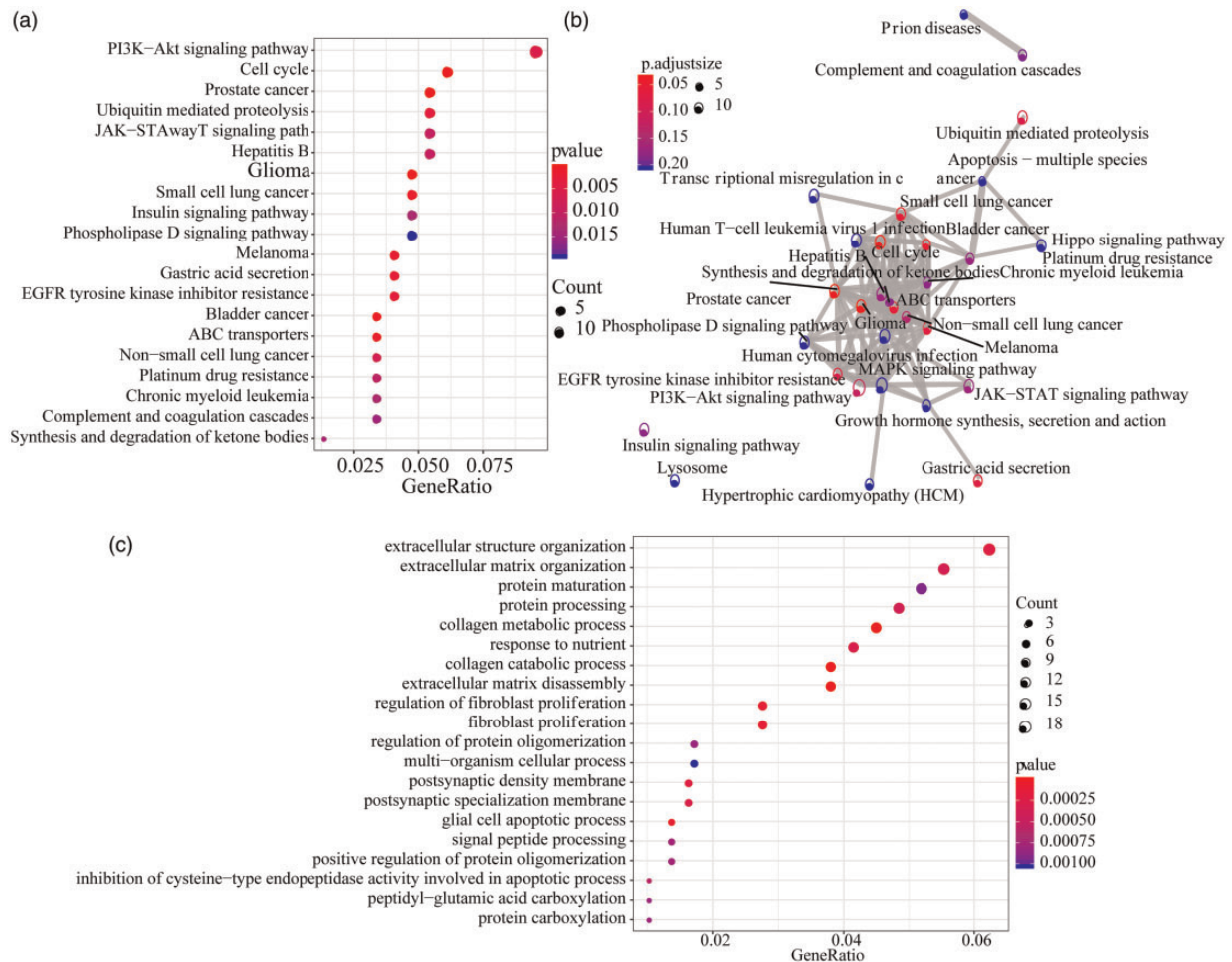


Figure 3. Functional analysis of genomic variant genes. (a) 372 biological processes involving the occurrence of copy amount variant genes (KEGG). (b) KEGG enrichment function integration network. (c) 372 copy amount variant genes were involved in the GO term. The color (red to blue) denotes the P data significance, the bluer denotes the lower P data, and the dot size denotes the amount pertaining to genes receiving the enrichment in the channel, the larger the amount, the greater the dot. (A color version of this figure is available in the online journal.)

expressing states of MYC, MMP27, and CPR142 genes displayed associations with high risk, acting as risk factors, and CHIC2, CCDC152, and LYL1 were significantly expressed, displaying associations with low risk, and they acted as protection factors (Figure 4(c)). In the training group, the average AUC of 6-gene sign reached 0.9 for one year, three years, and five years (Figure 4(d)). The samples received the grouping process by complying with the risk score threshold (cutoff = 0), and the prognostic processes of the groups with high and low risks were significantly different (Figure 4(e)). Furthermore, this study repeated 1000 random samples and applied the mode here to the mentioned 1000 random samples to analyze the significance distribution of the respective random sample (Figure S1(a)). It was therefore observed that they all showed a significant difference in prognosis. Accordingly, the TCGA sample suggested good reproducibility. The expression diversifications of six genes in the high-low risk group (Figure S1(b)) showed that there are significant diversifications in the expression of MYC, CHIC2, CCDC152, LYL1, and MMP27. Furthermore, this study added the correlation of the expressing state of six genes and immunity (Figure S1(c)). On the whole, MYC

showed significantly negative correlations with immunity score, while LYL1 was significantly and positively related to immunity score.

Robustness of 6-gene sign mode

To verify the robustness of the 6-gene sign mode, this study determined the risk score of the respective sample in the test set, and the correlation of the expressing states pertaining to the mentioned six genes and the risk score complied with the training group as well (Figure 5(a)). The average ROC of the same mode was greater than 0.86 (Figure 5(b)). When the samples fell to two groups in accordance with the training group's threshold, the OS of low-risk group was significantly better than the high-risk group (Figure 5(c)). To verify the classification performance of the 6-gene sign mode in the data of different data platforms, the data of GEO platform acted as an external data set to calculate the risk score of each sample. It was observed that the correlation of the expressing states of the mentioned six genes and risk score complied with the training group as well (Figure 5(d)), and ROC analysis showed that the five-year AUC reached 0.74, which was close to the training group

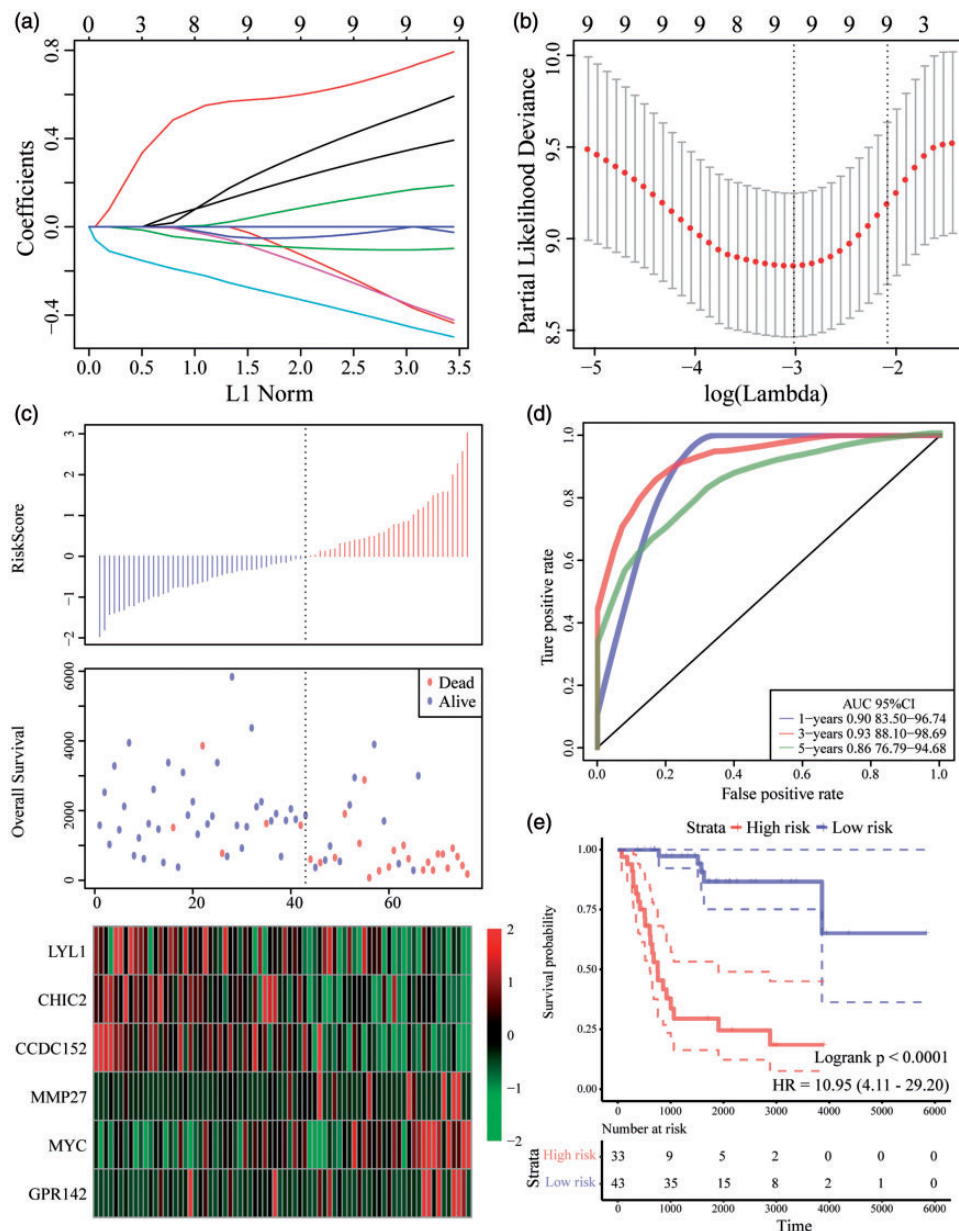


Figure 4. Identifying process for one six-gene sign in terms of osteosarcoma surviving state. (a) The track of the respective single variate, the horizontal axis denotes the log data pertaining to the single variate lambda, and the vertical axis represents the coefficient of the single variate. (b) The confidence interval is based on the respective lambda. (c) Expressions, surviving periods and surviving states, and risk scores pertaining to six genes inside the training group. (d) ROC curve and AUC of 6-gene sign. (e) KM surviving curve distributing process pertaining to 6-gene sign in the training group. (A color version of this figure is available in the online journal.)

(Figure 5(e)). The OS of low-risk group was noticeably greater over that of the high-risk group (Figure 5(f)). In conclusion, the 6-gene sign mode we selected has a good prognostic function in both internal and external data sets.

Clinical independence of six-genes signature mode

For identifying the independent property exhibited by the 6-gene sign mode for clinically related use, this study employed single-variate and multivariate Cox regression for the analysis of relevant HR, 95%CI of HR, and P value in clinical information carried by TARGET data set and GSE21257 data. The authors conducted a systematical analysis on the data set and clinical information recorded by TARGET data set and GSE21257 cases, including age,

gender, and metastasis, as well as grouping information of our 6-gene sign (Table 3). According to the TARGET data set, as revealed from univariate Cox regression study, high-risk group and metastasis were significantly associated with survival, and the corresponding multivariable Cox regression study reported that only the high-risk group (HR = 7.969, 95% CI = 3.220–19.725, $P = 7.17 \times 10^{-6}$) and metastatic (HR = 5.432, 95% CI = 2.488–11.861, $P = 2.16 \times 10^{-5}$) displayed significant correlations with survival. In GSE21257, as indicated by univariate Cox regression and multivariate analyses, high-risk group (HR = 3.340, 95% CI = 1.273–8.767, $P = 0.0143$) displayed significant correlations with survival. The mentioned results indicate that 6-gene sign is one prognosis-related

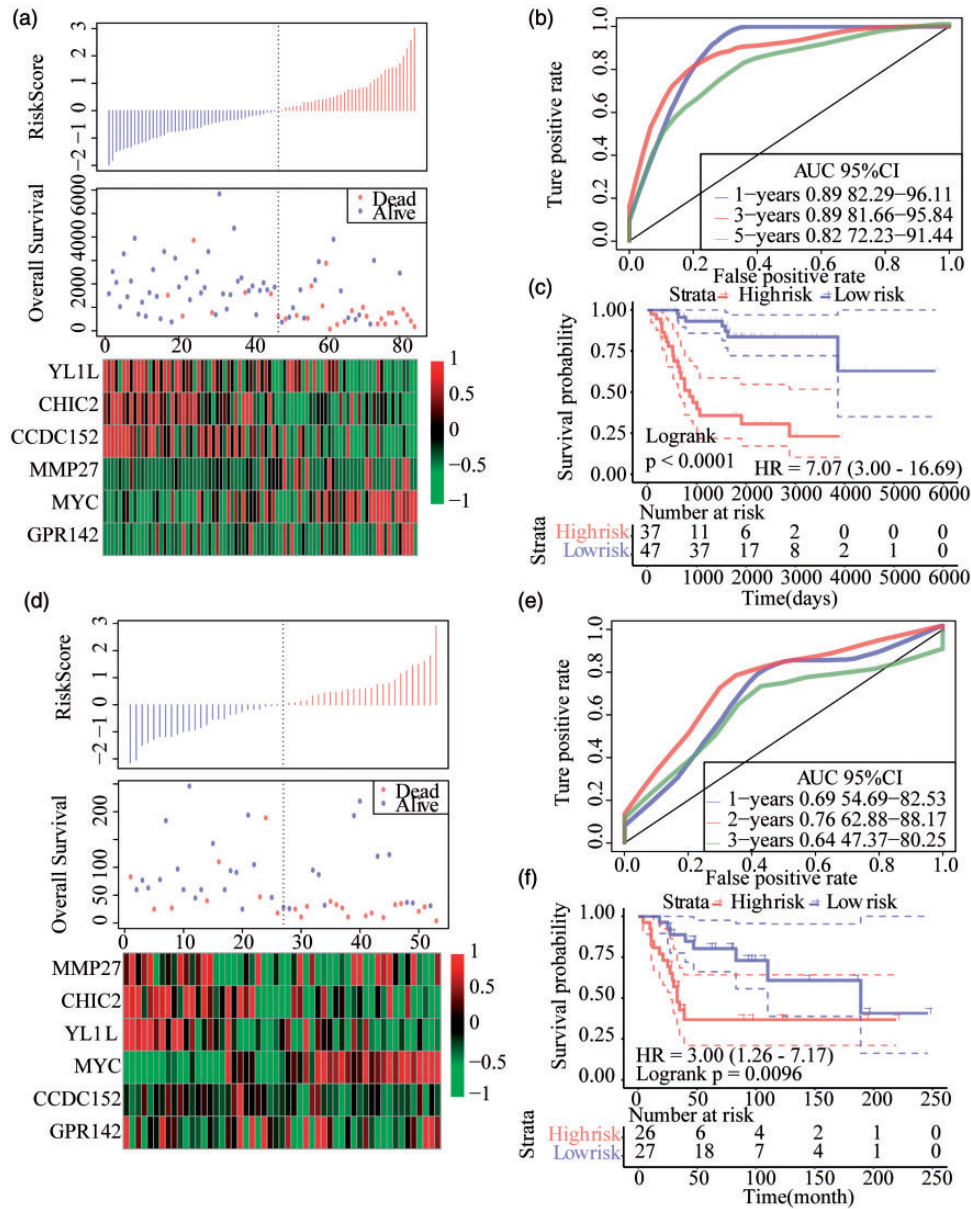


Figure 5. Robustness of 6-gene sign mode. (a) Expressions, surviving periods and surviving states, and risk scores pertaining to six genes inside the test data set. (b) ROC curve and AUC of 6-gene sign in test data set. (c) KM surviving curve distributing process pertaining to 6-gene sign in the test set. (d) Expressions, surviving periods and surviving states, and risk scores pertaining to six genes inside GSE21257 data set. (e) ROC curve and AUC of 6-gene sign in GSE21257 data set. (f) KM surviving curve distributing process pertaining to 6-gene sign in GSE21257 data set. (A color version of this figure is available in the online journal.)

Table 3. Univariate and multivariate Cox regression analysis identified clinical factors and clinical independence associated with prognosis.

Variables	Univariate analysis			Multivariable analysis		
	HR	95%CI of HR	P value	HR	95%CI of HR	P value
TARGET datasets						
6-gene risk score						
Risk score (high/low)	7.07	2.997–16.694	7.99e-06	7.969	3.220–19.725	7.17e-06
Age	0.99	0.912–1.075	0.81	1.040	0.944–1.145	0.427
Gender (male/female)	0.76	0.364–1.602	0.47	1.080	0.497–2.345	0.846
Metastatic vs. non-metastatic	4.74	2.271–9.895	3.42E-05	5.432	2.488–11.861	2.16E-05
GSE21257 validation datasets						
6-gene risk score						
Risk score (high/low)	3.003	1.258–7.171	0.013	3.340	1.273–8.767	0.0143
Age	1.0007	0.997–1.003	0.603	1.002	0.999–1.005	0.283
Gender (male/female)	1.403	0.588–3.348	0.445	0.997	0.390–2.552	0.995

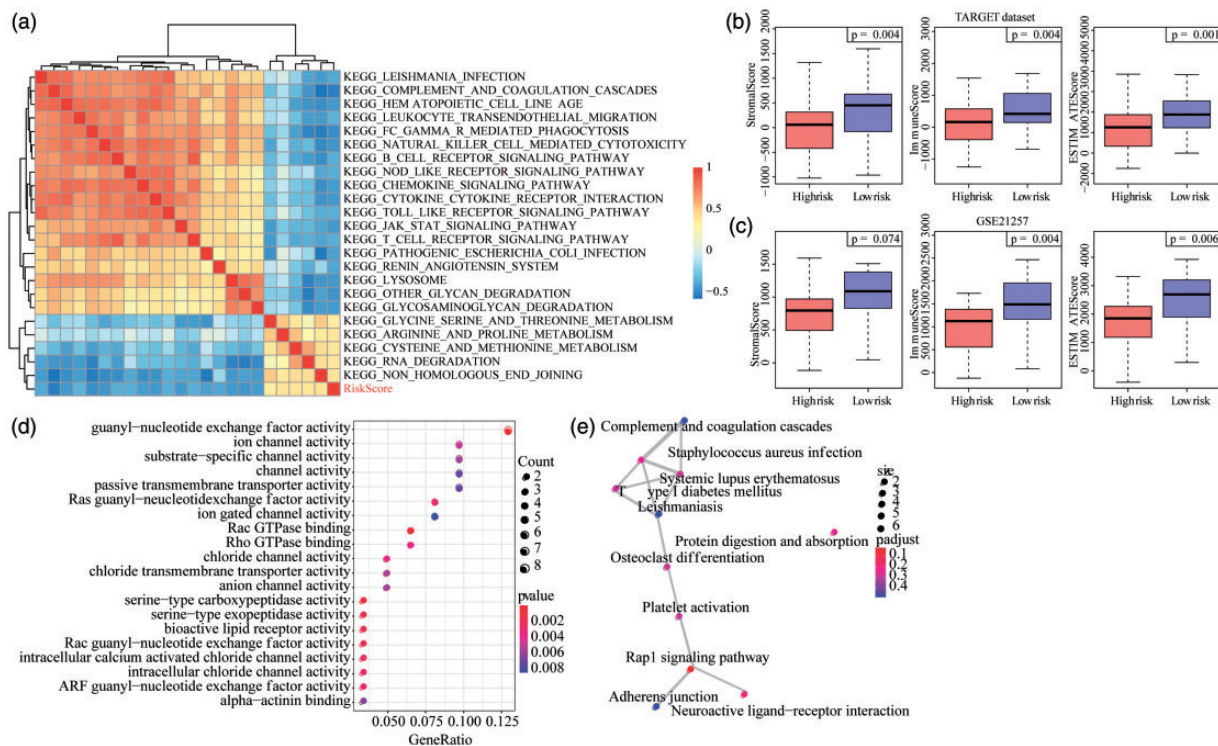


Figure 6. Analysis of functional diversifications of samples with high and low risk. (a) Correlation coefficients clustering of KEGG channels with RiskScore over 0.35 and RiskScore. (b) ESTIMATEScore, ImmuneScore, and StromalScore distribution inside high- and low-risk group of the group for training of TARGET. (c) Distribution of ESTIMATEScore, ImmuneScore, and StromalScore inside the groups with high and low risk of GSE21257 set. (d) GO enrichment results of prognosis-related genes with different methylation of 71 promoters. The horizontal axis represents gene percentage, and the vertical axis denotes enrichment GO term. The circle size represents the amount pertaining to genes receiving the enrichment to the channel. The larger the circle, the more genes receiving the enrichment to the channel. (e) KEGG enrichment function integration network, in which the color (red to blue) denotes P data significance, the bluer represents the lower P data, and the dot size denotes the amount pertaining to genes receiving the enrichment into the channel. (A color version of this figure is available in the online journal.)

indicating element not relying on other clinically related elements and exhibits good predicting performance for clinically related use significance.

Analysis of functional diversifications of samples with high and low risk

For observing the association of various samples' risk scores with biologically related functions, GSEA analyzed and calculated the scores of the respective sample on different functions. Further, this study obtained the relationship between the mentioned functions and risk score, and biological channels showing a relationship over 0.35 were selected. To be specific, 23 KEGG channels with significant negative correlation are primarily related to B_cell_receptor_signaling_channel, T_cell_receptor_signaling_channel, cytokine_cytokine_interaction, and other immune channels. There were five KEGG channels exhibiting significant positive correlations, primarily related to metabolic channels (Figure 6(a)). Tumor microenvironment plays a crucial role in the process of tumor metastasis. Here, we calculated ESTIMATEScore, ImmuneScore, and StromalScore of each sample respectively, and analyzed the diversifications of three immune microenvironment scores in the high- and low-risk groups. To be specific, there were significant diversifications in immune and basal score in the high- and low-risk samples in the group for training TARGET (Figure 6 (b)). The same phenomenon was observed in the GEO

dataset (Figure 6(c)). This suggests that poor prognosis in the high-risk group may be related to immunosuppression. Methylation diversifications inside the groups with high and low risk were analyzed and 4465 different methylation sites were identified, including 3676 high methylation site and 789 low methylation sites. A total of 71 prognostic-related genes were identified by analyzing the genes downstream of the mentioned methylation site. The mentioned genes are primarily enriched in the Rap1 signaling channel, osteoclast differentiation, and other osteoclast differentiation channels and molecular functions related to several enzyme activities (Figure 6(d) and (e)).

6-gene sign mode has superior performance over existing signatures

The authors drew the comparison of existing studies and identified two reported robust prognosis-related risk modes, 8-gene sign²⁵ and 4-pseudogene sign.²⁶ To make the modes comparable, the authors obtained the risk score of the respective osteosarcoma sample inside the training group by employing the identical approach according to the relevant genes under the two modes. The ROC of the respective mode was assessed, and the samples fell to high- and low-risk groups according to the optimal threshold; in addition, this study obtained the difference in OS prognosis between the two groups of samples. The three-year AUC of 8-gene sign is 0.75, $P = 0.00024$

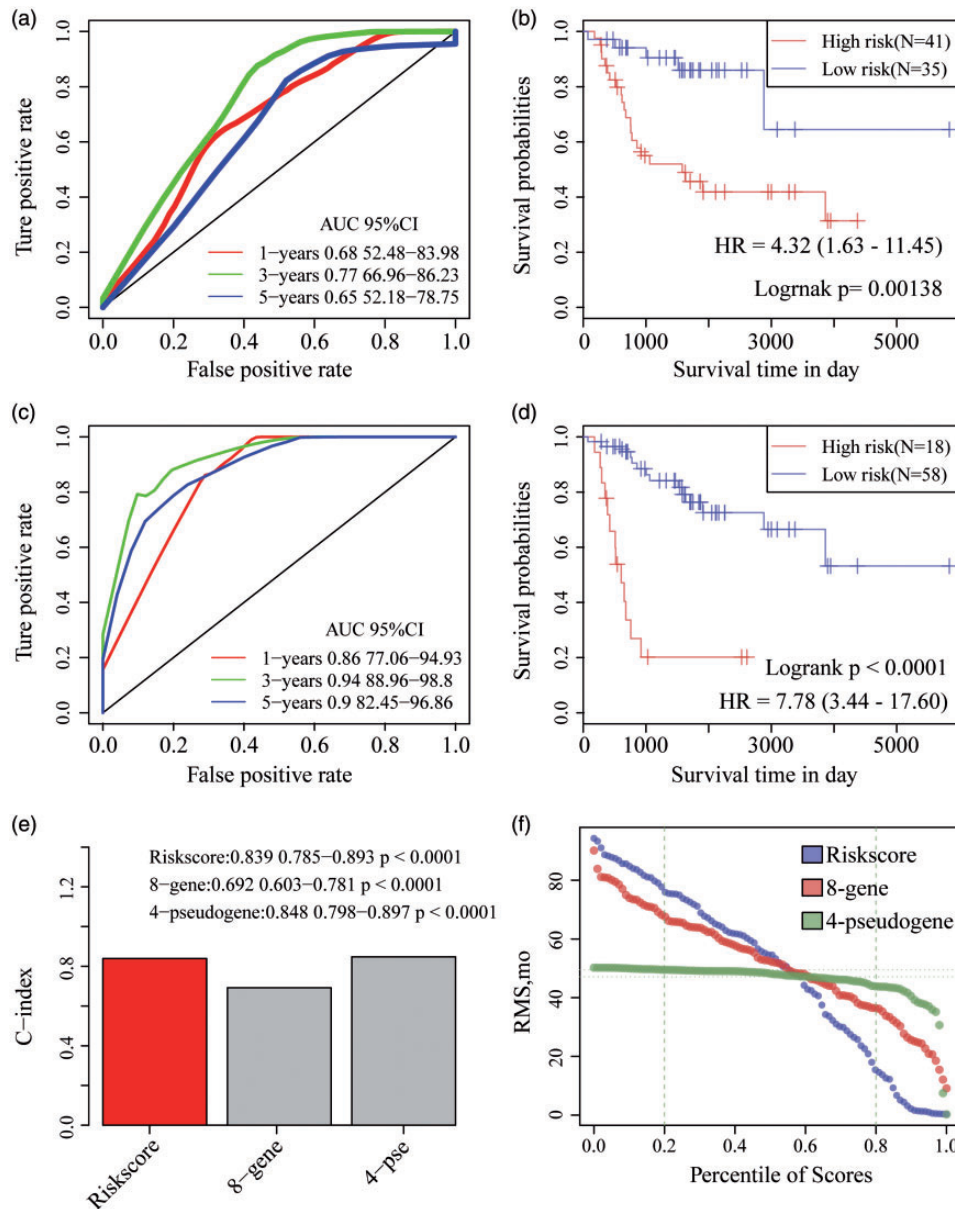


Figure 7. 6-gene sign mode has superior performance over existing signatures. (a) AUC curve of 8-gene sign in the group for training of TARGET. (b) KM curve of 8-gene sign in the group for training of TARGET. (c) AUC curve of 4-pseudogene sign in the group for training of TARGET. (d) KM curve of 4-pseudogene sign in the group for training of TARGET. (e) C-index of three prognosis-related risk modes. (f) RMS (limited mean surviving state) curve of three prognosis-related risk modes. The dash line denotes the RMS time (months) representing 20% and 80% percentile scores, separately. (A color version of this figure is available in the online journal.)

(Figure 7(a)). OS prognosis was also significantly different (Figure 7(b)). The three-year AUC of 4-pseudogene sign is 0.93, $P < 0.0001$ (Figure 7(c)). OS prognosis was also significantly different (Figure 7(d)). Compared with the 6-gene sign, the predictive performance of the 8-gene sign is obviously weak, while the 4-pseudogene sign has similar predictive performance. The concordance index (C-index) of the mentioned two modes and our mode were calculated, and found that the C-index of the 6-gene sign mode and the 4-pseudogene sign mode are above 0.839, while the C-index of the 8-gene mode is 0.69 (Figure 7(e)). The mentioned data show that the overall performance of the 6-gene sign mode is excellent. The authors adopted RMS time for assessing the predicting influences exerted by the three modes in a range of points in time. According to the RMS curve, the

three modes intersect at nearly 50 months, and our mode has the best prediction effect in short-term survival (Figure 7(f)).

Discussion

Osteosarcoma is characterized by a variety of cytogenetic variations, abnormalities in a variety of different channels, and substantial changes between cells.²⁹ The etiology of osteosarcoma is unclear, and the only prognostic markers are the absence of metastasis at diagnosis and the extent of tumor necrosis as a measure of response to neoadjuvant chemotherapy.³⁰ Accordingly, screening prognosis-related molecular marking elements comprehensively reflecting tumor biologically related features critically impacts

individualized preventing and treating processes of osteosarcoma cases. The authors investigated the expressing state profiles of 137 osteosarcoma samples from the Target data set and GEO and identified OS-associated robust six gene signs, with no dependence on clinically related elements.

Currently gene signs are available for clinical use, e.g. Oncotype DX by expression of 21 genes for disease recurrence score,^{31–33} Coloprint, an 18 gene expression trait in colon cancer.^{34–36} The mentioned results have shown that the use of gene expression profiling to screen novel prognostic markers in cancer has become the most promising method for high-throughput molecular identification. Zhang *et al.*²⁵ employed weighted gene relationship network study and least absolute shrinkage and selection operating element Cox regression identified an 8-gene sign in the gene expression profile. However, it has not been applied in clinical practice. Accordingly, more signatures are required for selection and validation in a larger study cohort. The proposed 6-gene sign has a high AUC and a small number of genes, thereby expediting clinical transformation.

In the 6-gene sign here, MYC, MMP27, and CPR142 were risk factors, and CHIC2, CCDC152, and LYL1 were protective factors. They all had genomic copy amount abnormalities, and MYC was reported as a cell cycle-related gene, and overexpression of MYC weakened the clock and in turn facilitated cell proliferation,³⁷ displaying relationships to cell migration, invasion, and epithelial mesenchymal transformation of osteosarcoma.^{38,39} The other five genes have not been reported and may be novel markers for osteosarcoma. The functional analysis suggested that abnormalities in the high-risk group were related to B_cell_receptor_signaling_channel, T_cell_receptor_signaling_channel, cytokine_cytokine_receptor_interaction, and other immune channels. Notable diversifications existed in the immune microenvironment of the groups with high and low risk as well, with significant immunosuppression in the high-risk group. As therefore demonstrated, poor prognosis may be related to immune abnormalities in the tumor's immune microenvironment. Besides, methylation analysis showed abnormal methylation of gene promoter in Rap1 signaling channel, osteoclast differentiation, and other osteoclast differentiation channels. The mentioned results indicate that the occurrence and development of osteosarcoma are systematic and multi-group coordinated. In the present study, multi-group combined analysis was conducted initially to identify the 6-gene mode to be one new prognosis-related marking element in terms of osteosarcoma, showing promising clinical applications and probably providing one diagnosing target for clinical cases.

Though bioinformatics techniques were employed to identify probable candidate genes in term of tumor prognostic process in a significant sample, several limitations of this study are noteworthy. First, the sample had insufficient clinical follow-up information, so elements (e.g. the appearance of a patient's other healthy status) were not considered to distinguish prognostic biomarkers. Second, the outcomes achieved from bioinformatics study alone are insufficient and require an experimentally verifying process for

confirming the mentioned outcomes. Accordingly, in-depth genetically and experimentally related researches using a greater sample size and experimentally validating process should be conducted.

Conclusions

To sum up, the present study built one 6-gene sign prognosis stratifying mechanism, exhibiting an effective AUC in the group of training, validating process group, and externally independent verifying group, and is not determined by clinically related characteristics. As opposed to the clinically related characteristics, gene classifying element is capable of improving surviving risk prediction. Accordingly, the classifier may be used to be one molecular diagnosis-related testing process for assessing the prognosis-related risk facing osteosarcoma cases.

AUTHORS' CONTRIBUTIONS

The authors here overall engaged in the design, interpretation of the studies, and analysis of the data and review of the manuscript; RML, MC conceived and designed the research. RmL, GL, HGL acquired data; GSW and ZJW contributed to statistical analysis, analysis, and interpretation of data; RML drafted the manuscript; MC did revision of manuscript for important intellectual content.


DECLARATION OF CONFLICTING INTERESTS

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

FUNDING

This work was supported by the Shanghai Science and Technology Commission (19411963100).

ORCID iD

Ming Cai  <https://orcid.org/0000-0002-0020-0585>

SUPPLEMENTAL MATERIAL

Supplemental material for this article is available online.

REFERENCES

1. Simpson S, Dunning MD, de Brot S, Grau-Roma L, Mongan NP, Rutland CS. Comparative review of human and canine osteosarcoma: morphology, epidemiology, prognosis, treatment and genetics. *Acta Vet Scand* 2017;**59**:71
2. Parkin DM, Ferlay J, Curado MP, Bray F, Edwards B, Shin HR, Forman D. Fifty years of cancer incidence: CI5 I-IX. *Int J Cancer* 2010;**127**:2918–27
3. Moore DD, Luu HH. Osteosarcoma. *Cancer Treat Res* 2014;**162**:65–92
4. McCarville MB. The child with bone pain: malignancies and mimickers. *Cancer Imaging* 2009;**9**:S115–21
5. Bielack SS, Kempf-Bielack B, Delling G, Exner GU, Flege S, Helmke K, Kotz R, Salzer-Kuntschik M, Werner M, Winkelmann W, Zoubek A, Jurgens H, Winkler K. Prognostic factors in high-grade osteosarcoma of the extremities or trunk: an analysis of 1,702 patients treated on neoadjuvant cooperative osteosarcoma study group protocols. *J Clin Oncol* 2002;**20**:776–90

6. Chen Y, Gokavarapu S, Shen Q, Liu F, Cao W, Ling Y, Ji T. Chemotherapy in head and neck osteosarcoma: adjuvant chemotherapy improves overall survival. *Oral Oncol* 2017;**73**:124–31
7. Kansara M, Teng MW, Smyth MJ, Thomas DM. Translational biology of osteosarcoma. *Nat Rev Cancer* 2014;**14**:722–35
8. Hanahan D, Weinberg RA. Hallmarks of cancer: the next generation. *Cell* 2011;**144**:646–74
9. Bishop MW, Janeway KA, Gorlick R. Future directions in the treatment of osteosarcoma. *Curr Opin Pediatr* 2016;**28**:26–33
10. Cheng L, Pandya PH, Liu E, Chandra P, Wang L, Murray ME, Carter J, Ferguson M, Saadatzaheh MR, Bijangi-Visheshsaraei K, Marshall M, Li L, Pollok KE, Renbarger JL. Integration of genomic copy number variations and chemotherapy-response biomarkers in pediatric sarcoma. *BMC Med Genomics* 2019;**12**:23
11. Mirabello L, Koster R, Moriarity BS, Spector LG, Meltzer PS, Gary J, Machiela MJ, Pankratz N, Panagiotou OA, Largaespada D, Wang Z, Gastier-Foster JM, Gorlick R, Khanna C, de Toledo SR, Petrilli AS, Patino-Garcia A, Sierrasesumaga L, Lecanda F, Andrulis IL, Wunder JS, Gokgoz N, Serra M, Hattinger C, Picci P, Scotlandi K, Flanagan AM, Tirabosco R, Amary MF, Halai D, Ballinger ML, Thomas DM, Davis S, Barkauskas DA, Marina N, Helman L, Otto GM, Becklin KL, Wolf NK, Weg MT, Tucker M, Wacholder S, Fraumeni JF, Jr., Caporaso NE, Boland JF, Hicks BD, Vogt A, Burdett L, Yeager M, Hoover RN, Chanock SJ, Savage SA. A genome-wide scan identifies variants in NF1B associated with metastasis in patients with osteosarcoma. *Cancer Discov* 2015;**5**:920–31
12. Cai H, Miao M, Wang Z. miR-214-3p promotes the proliferation, migration and invasion of osteosarcoma cells by targeting CADM1. *Oncol Lett* 2018;**16**:2620–8
13. Kelly AD, Haibe-Kains B, Janeway KA, Hill KE, Howe E, Goldsmith J, Kurek K, Perez-Atayde AR, Francoeur N, Fan JB, April C, Schneider H, Gebhardt MC, Culhane A, Quackenbush J, Spentzos D. MicroRNA paraffin-based studies in osteosarcoma reveal reproducible independent prognostic profiles at 14q32. *Genome Med* 2013;**5**:2
14. Buddingh EP, Kuijjer ML, Duim RA, Burger H, Agelopoulos K, Myklebost O, Serra M, Mertens F, Hogendoorn PC, Lankester AC, Cleton-Jansen AM. Tumor-infiltrating macrophages are associated with metastasis suppression in high-grade osteosarcoma: a rationale for treatment with macrophage activating agents. *Clin Cancer Res* 2011;**17**:2110–9
15. Guo JC, Wu Y, Chen Y, Pan F, Wu ZY, Zhang JS, Wu JY, Xu XE, Zhao JM, Li EM, Zhao Y, Xu LY. Protein-coding genes combined with long non-coding RNA as a novel transcriptome molecular staging model to predict the survival of patients with esophageal squamous cell carcinoma. *Cancer Commun* 2018;**38**:4
16. Mermel CH, Schumacher SE, Hill B, Meyerson ML, Beroukhi R, Getz G. GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. *Genome Biol* 2011;**12**:R41
17. Chen YA, Lemire M, Choufani S, Butcher DT, Grafodatskaya D, Zanke BW, Gallinger S, Hudson TJ, Weksberg R. Discovery of cross-reactive probes and polymorphic CpGs in the illumina Infinium HumanMethylation450 microarray. *Epigenetics* 2013;**8**:203–9
18. Kostareli E, Hielscher T, Zucknick M, Baboci L, Wichmann G, Holzinger D, Mucke O, Pawlita M, Del Mistro A, Boscolo-Rizzo P, Da Mosto MC, Tirelli G, Plinkert P, Dietz A, Plass C, Weichenhan D, Hess J. Gene promoter methylation signature predicts survival of head and neck squamous cell carcinoma patients. *Epigenetics* 2016;**11**:61–73
19. Zhang JX, Song W, Chen ZH, Wei JH, Liao YJ, Lei J, Hu M, Chen GZ, Liao B, Lu J, Zhao HW, Chen W, He YL, Wang HY, Xie D, Luo JH. Prognostic and predictive value of a microRNA signature in stage II colon cancer: a microRNA expression analysis. *Lancet Oncol* 2013;**14**:1295–306
20. Papaemmanuil E, Gerstung M, Malcovati L, Tauro S, Gundem G, Van Loo P, Yoon CJ, Ellis P, Wedge DC, Pellagatti A, Shlien A, Groves MJ, Forbes SA, Raine K, Hinton J, Mudie LJ, McLaren S, Hardy C, Latimer C, Della Porta MG, O'Meara S, Ambaglio I, Galli A, Butler AP, Walldin G, Teague JW, Quek L, Sternberg A, Gambacorti-Passerini C, Cross NC, Green AR, Boultonwood J, Vyas P, Hellstrom-Lindberg E, Bowen D, Cazzola M, Stratton MR, Campbell PJ, Chronic myeloid disorders working group of the international cancer genome C. Clinical and biological implications of driver mutations in myelodysplastic syndromes. *Blood* 2013;**122**:3616–27; quiz 99
21. Yuan Y, Van Allen EM, Omberg L, Wagle N, Amin-Mansour A, Sokolov A, Byers LA, Xu Y, Hess KR, Diao L, Han L, Huang X, Lawrence MS, Weinstein JN, Stuart JM, Mills GB, Garraway LA, Margolin AA, Getz G, Liang H. Assessing the clinical utility of cancer genomic and proteomic data across tumor types. *Nat Biotechnol* 2014;**32**:644–52
22. Yu G, Wang LG, Han Y, He QY. clusterProfiler: an R package for comparing biological themes among gene clusters. *Omic* 2012;**16**:284–7
23. Subramanian A, Kuehn H, Gould J, Tamayo P, Mesirov JP. GSEA-P: a desktop application for gene set enrichment analysis. *Bioinformatics* 2007;**23**:3251–3
24. Liberzon A, Subramanian A, Pinchback R, Thorvaldsdottir H, Tamayo P, Mesirov JP. Molecular signatures database (MSigDB) 3.0. *Bioinformatics* 2011;**27**:1739–40
25. Zhang H, Guo L, Zhang Z, Sun Y, Kang H, Song C, Liu H, Lei Z, Wang J, Mi B, Xu Q, Guan H, Li F. Co-expression network analysis identified gene signatures in osteosarcoma as a predictive tool for lung metastasis and survival. *J Cancer* 2019;**10**:3706–16
26. Liu F, Xing L, Zhang X, Zhang X. A four-pseudogene classifier identified by machine learning serves as a novel prognostic marker for survival of osteosarcoma. *Genes* 2019;**10**:414
27. Chakraborty H, Hossain A. R package to estimate intracluster correlation coefficient with confidence interval for binary data. *Comput Methods Programs Biomed* 2018;**155**:85–92
28. Kang L, Chen W, Petrick NA, Gallas BD. Comparing two correlated C indices with right-censored survival outcome: a one-shot nonparametric approach. *Stat Med* 2015;**34**:685–703
29. Bridge JA, Nelson M, McComb E, McGuire MH, Rosenthal H, Vergara G, Maale GE, Spanier S, Neff JR. Cytogenetic findings in 73 osteosarcoma specimens and a review of the literature. *Cancer Genet Cytogenet* 1997;**95**:74–87
30. Wang LL. Biology of osteogenic sarcoma. *Cancer J* 2005;**11**:294–305
31. Siow ZR, De Boer RH, Lindeman GJ, Mann GB. Spotlight on the utility of the oncotype DX((R)) breast cancer assay. *Int J Womens Health* 2018;**10**:89–100
32. Bhutiani N, Egger ME, Ajkay N, Scoggins CR, Martin RC, 2nd, McMasters KM. Multigene signature panels and breast cancer therapy: patterns of use and impact on clinical decision making. *J Am Coll Surg* 2018;**226**:406–12.e1
33. Wang SY, Dang W, Richman I, Mougalian SS, Evans SB, Gross CP. Cost-effectiveness analyses of the 21-gene assay in breast cancer: systematic review and critical appraisal. *J Clin Oncol* 2018;**36**:1619–27
34. Kopetz S, Taberner J, Rosenberg R, Jiang ZQ, Moreno V, Bachleitner-Hofmann T, Lanza G, Stork-Sloots L, Maru D, Simon I, Capella G, Salazar R. Genomic classifier ColoPrint predicts recurrence in stage II colorectal cancer patients more accurately than clinical factors. *Oncologist* 2015;**20**:127–33
35. Tan IB, Tan P. Genetics: an 18-gene signature (ColoPrint((R))) for colon cancer prognosis. *Nat Rev Clin Oncol* 2011;**8**:131–3
36. Maak M, Simon I, Nitsche U, Roepman P, Snel M, Glas AM, Schuster T, Keller G, Zeestraten E, Goossens I, Janssen KP, Friess H, Rosenberg R. Independent validation of a prognostic genomic signature (ColoPrint) for patients with stage II colon cancer. *Ann Surg* 2013;**257**:1053–8
37. Shostak A, Ruppert B, Ha N, Bruns P, Toprak UH, Project IM-S, Eils R, Schlesner M, Diernfellner A, Brunner M. MYC/MIZ1-dependent gene repression inversely coordinates the circadian clock with cell cycle and proliferation. *Nat Commun* 2016;**7**:11807
38. Tang Y, Ji F. lncRNA HOTTIP facilitates osteosarcoma cell migration, invasion and epithelial-mesenchymal transition by forming a positive feedback loop with c-Myc. *Oncol Lett* 2019;**18**:1649–56
39. Chen D, Zhao Z, Huang Z, Chen DC, Zhu XX, Wang YZ, Yan YW, Tang S, Madhavan S, Ni W, Huang ZP, Li W, Ji W, Shen H, Lin S, Jiang YZ. Super enhancer inhibitors suppress MYC driven transcriptional amplification and tumor progression in osteosarcoma. *Bone Res* 2018;**6**:11