Observational Study

Medicine®

OPEN

DNA methylation patterns–based subtype distinction and identification of soft tissue sarcoma prognosis

Kai Li[a], Zhengyuan Wu[a], Jun Yao, MD, PhD[b,c], Jingyuan Fan[a], Qingjun Wei, MD, PhD[a,*]

Abstract

Soft tissue sarcomas (STSs) are heterogeneous at the clinical with a variable tendency of aggressive behavior. In this study, we constructed a specific DNA methylation-based classification to identify the distinct prognosis-subtypes of STSs based on the DNA methylation spectrum from the TCGA database. Eventually, samples were clustered into 4 subgroups, and their survival curves were distinct from each other. Meanwhile, the samples in each subgroup reflected differentially in several clinical features. Gene Ontology (GO) and the Kyoto Encyclopedia of Genes and Genomes (KEGG) analysis was also conducted on the genes of the corresponding promoter regions of the above-described specific methylation sites, revealing that these genes were mainly concentrated in certain cancer-associated biological functions and pathways. In addition, we calculated the differences among clustered methylation sites and performed the specific methylation sites with LASSO algorithm. The selection operator algorithm was employed to derive a risk signature model, and a prognostic signature based on these methylation sites performed well for risk stratification in STSs patients. At last, a nomogram consisted of clinical features and risk score was developed for the survival prediction. This study declares that DNA methylation-based STSs subtype classification is highly relevant for future development of personalized therapy as it identifies the prediction value of patient prognosis.

Abbreviations: BP = biological process, CC = cellular component, CDF = cumulative distribution function, FC = fold change, FDR = false discovery rate, GO = Gene Ontology, KEGG = Kyoto Encyclopedia of Genes and Genomes, KNN = k-nearest neighbor, LASSO = least absolute shrinkage and selection operator, MF = molecular function, OS = overall survival, PI3K = phosphatidylinositol 3–kinase, RNA-Seq = RNA-sequencing, STSs = soft tissue sarcomas, TCGA = The Cancer Genome Atlas, UCSC Xena = University of California Santa Cruz Xena.

Keywords: DNA methylation, soft tissue sarcomas, molecular subtype, prognosis

1. Introduction

Soft tissue sarcomas (STSs), which arise predominantly from the embryonic mesoderm, are a set of malignancies that account for 0.73% to 0.81% of all and 6% of pediatric cancers.[1,2] In accordance with the heterogeneity in histopathological features, clinical manifestations, and molecular signature, approximately 50 different histological subtypes have been discovered within STSs patients.[3] Meanwhile, STSs are commonly presenting as a symptomless mass in almost every part of the human body, including the retroperitoneum, viscera, and extremities,[2,4] with approximately 50% of 5-year overall survival (OS).[5] Although STSs have some common morphologic features, its proper diagnosis and treatment are still challenging for pathologists and physicians due to its extremely variable biology and genetics and low incidence. Up to now, the most optimal management of STSs is still surgical resection, although its less successful in the advanced STS in accompany with the high rate of local recurrence.[6,7] Consequently, molecular characteristics of these heterogeneous tumors are warranted to be further explored and new classification systems should be elucidated for providing more potential prognostic factors in the clinic.

Previously, several genomic and transcriptome data have focused on exploring effective diagnostic or prognostic markers in STSs, including alternative splicing,[8] copy number variation,[9] and genes expression.[10] As one of the core elements in tumorigenesis progression, DNA methylation occurs early and frequently in regulating a variety of genomic functions.[11] DNA methylation is a posttranslational modification process, which are selectively occurred on the cytosines of 5"-CpG-3" to generates 5-methyldeoxycytidine. The aberrant of DNA methylation, especially in CpG-rich regions (CpG islands), has been found closely related with physio-pathologic mechanisms underlying an array of human diseases.[12] Moreover, CpG islands was frequently detected in the promoter regions of the structural transcription gene,[13] and abnormal CpG island

Editor: Peeyush Goel.

The authors declare that they have no competing interests.

The datasets generated during and/or analyzed during the current study are available from the corresponding author on reasonable request.

[a] Department of Orthopedics Trauma and Hand Surgery, [b] Department of Bone and Joint Surgery, The First Affiliated Hospital of Guangxi Medical University, [c] Guangxi Collaborative Innovation Center for Biomedicine, Guangxi Medical University, Nanning, China.

[*] Correspondence: Qingjun Wei, Department of Orthopedics Trauma and Hand Surgery, The First Affiliated Hospital of Guangxi Medical University, Nanning 530021, China (e-mail: weiqingjungxnn@163.com).

How to cite this article: Li K, Wu Z, Yao J, Fan J, Wei Q. DNA methylation patterns–based subtype distinction and identification of soft tissue sarcoma prognosis. Medicine 2021;100:5(e23787).

Received: 2 July 2020 / Received in final form: 13 October 2020 / Accepted: 13 November 2020

http://dx.doi.org/10.1097/MD.0000000000023787

hypermethylation of various tumor suppressor genes and hypomethylation of oncogenes play vital role in carcinogenesis.[14,15] At present, as a promising molecular marker of STSs, abnormal DNA methylation appears in early detection, prognosis prediction, molecular classification.[16,17] Meanwhile, multiple biological studies have also clarified that a series of methylations in gene promoter sequences is correlated with the prognosis and progression of STSs patients.[18–20] Nevertheless, the prognostic value of these aberrantly methylation sites in STSs subtypes and the complex role of DNA methylation in distinct gene regions are still largely unclarified and require further validation in the prognostic role of DNA methylation in STSs.

Therefore, in this study, we addressed a classification method by identifying specific prognosis-subtypes which were based on DNA methylation profiles of STS from TCGA database, which may help to identify new markers to accurately subdivide STSs patients. Moreover, our classification system provides a more accurate prediction of clinical behavior and identifies higher risk assessment accuracy for clinicians on personalized treatments

2. Material and methods

2.1. Date pre-processing and initial screening of DNA methylation sites in STSs

The latest DNA methylation data were downloaded from the Illumina Infinium HumanMethylation450 Bead-Chip array of TCGA Genomic Data Commons application programming interface (https://portal.gdc.cancer.gov)[12] on April 6th, 2020, which containing a total of 269 STSs samples and 485,577 CpG sites. The corresponding clinical and prognostic parameters were acquired from the University of California Santa Cruz Xena (UCSC Xena; http://xena.ucsc.edu/) [21] on April 5th, 2020, for further methylation profile matching, ultimately, 258 STSs samples were selected for methylation analysis. Approval by the Ethics Committee was not necessary because all data were collected from publicly available databases (TCGA and UCSC). Afterward, the RNA-sequencing (RNA-Seq) date included 265 samples were also collected from the TCGA public database (accessed April 5th, 2020) which was normalized by the DESeq package in R platform.[22] Acquiring and usage of all data in the current study are following the publication guidelines of TCGA (https://cancergenome.nih.gov/publications/publicationguidelines).

Moreover, the CpG sites with missing data over 70% from all samples were all removed. The k-nearest neighbor (KNN) imputation method in the sva R package was utilized to estimate the missing values, with the further removal of the unstable genomic methylation sites which contained CpG sites in sex chromosomes and single nucleotide polymorphisms.[23] Subsequently, DNA methylation with strongly genetic modulation effects was selected for the following exploration based on the annotation between CpG sites and gene promoter regions,[24] and the promoter region was defined as the 2 kb upstream to 0.5 kb in the transcription start site. Yielding 206,636 methylation sites were obtained for further analysis. Our workflow for bioinformatics analysis of publicly available datasets is illustrated in Figure 1.

2.2. COX proportional risk regression models regarding methylation sites

Univariate COX proportional risk regression models were constructed with the survival data and every methylation site by utilizing the survival coxph function R package with the significant threshold was set as $P < .001$.[25] Subsequently, the obtained methylation sites were introduced into further multivariate Cox proportional regression model analysis, where age, gender, tissue or organ of origin, and histological type were selected as the covariates from the downloaded STSs clinical data, with $P < .001$ set as the significance threshold. Finally, the CpG sites, which were identified as independent prognostic factors, were chosen as the classification features.

2.3. Consensus clustering of prognosis molecular subtypes and clinical characteristic analyses

The K-means clustering algorithm in the Concensus Cluster Plus R packet was utilized for consensus clustering to determine STSs subgroups.[26] In this study, 80% of the STSs samples were carried out 100 times by adopting the resampling method, and Euclidean distance and κ-means algorithm was calculated to measure the similarity in samples distance and determine the stability and reliability of classification results, respectively. The optimal cluster number was then determined by using the cumulative distribution function (CDF) and the delta area plot which should be with relative high consistency, low variation coefficient, and no obviously increased area under the CDF curve, and the cluster outcomes were applied for the following clinical characteristics analysis. The pheatmap R package was generated to construct the corresponding consensus cluster heatmap which was plotted based on the age, gender, tissue or organ of origin, and histological type of each sample. Afterward, the overall survival analysis of STSs subgroups was constructing by Kaplan–Meier method and log-rank test with R Bioconductor survival package, and Chi-Squared test was used to comprehensively determine the associations between DNA methylation clusters and clinical characteristics.

2.4. Identification of methylation sites annotated gene expression and pathway enrichment analysis

DNA methylation at the promoter site can modulate gene expression. In order to investigate the association of previously obtained CpG sites with gene expression in the classified subgroups, methylation sites were subjected to genomic annotations to determine the corresponding genes, and then, these genes were identified for plotting the expression profile heat map. Meanwhile, to explore the biological terms of specific gene lists, GO and KEGG was performed for enrichment analysis based on the Cluster Profiler R package,[27] where a $P < .05$ was set as the cut-off criteria. And all 3 aspects of GO analysis were included for providing gene function, including, biological process (BP), cellular component (CC), and molecular function (MF).

2.5. Screening of intragroup-specific methylation sites, and prognosis prediction model construction

To identify the differences among the clustered methylation sites, Kolmogorov-Smirnov tests were applied for each CpG site to determine its different distribution in the methylation level,[28] and the thresholds set at log2(fold change (FC)) > 1 and false discovery rate (FDR) < 0.05 were indicated as the significant difference. Furthermore, the heatmap corresponding to the differential frequency of every CpG site in each subgroup was further detected by ComplexHeatmap R package. In this

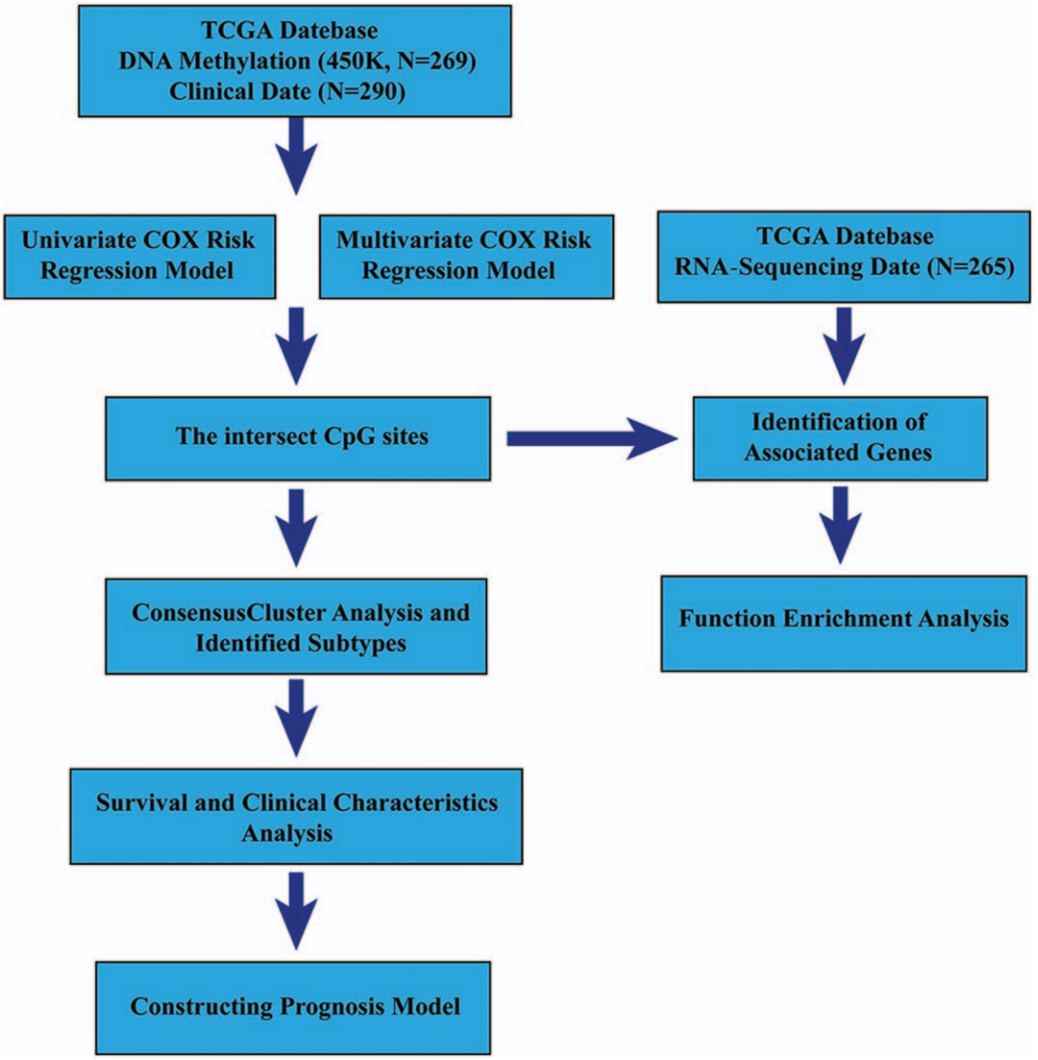

Figure 1. Flowchart describing the schematic overview of the study design.

exploration, the specific methylation sites in Cluster1 with a great number of specific CpG sites and the best prognosis were selected for the prognostic model construction. From this, the least absolute shrinkage and selection operator (LASSO) algorithm was developed to construct a potential risk signature, finally, 13 methylation sites were contained in the model. The formula of risk score was calculated as follows: Risk score $= \sum_{i=1}^{n} Coef_i * x_i$, where $Coef_i$ is the coefficient, and x_i is the z-score-transformed relative expression value of each CpG sites.

Based on the LASSO model, patients were categorized into high or low risk group, and the overall survival analysis and clinical relevance of STSs samples at high and low risk were also generated as we described before. Furthermore, the survivalROC packet in R Program was calculated to verify the stability and reliability of the model,[29] and Univariate and multivariate Cox regression analyses were counted to validate the predictive accuracy of the clinical characteristics and risk score.

2.6. Construction a predictive nomogram

To provide an application tool for predicting STSs clinical outcome, the nomogram incorporated with age, gender, tissue or organ of origin, histological type, and risk score was plotted by using R package.[30]

3. Results

3.1. Characteristics of DNA methylation sites based on the prognosis results of STSs

After conducting univariate Cox proportional hazards regression model to each methylation site, 2693 CpG sites were identified as significant survival correlated methylation sites ($P < .001$). Furthermore, these significant CpG sites were then introduced into the multivariate Cox proportional risk regression models. As a result, 1445 intersected independent prognostic CpG sites between 2 analysis were chosen for the further prognosis analysis, and top 20 CpG sites were revealed in Table 1.

3.2. Consensus clustering of characteristic DNA methylation sites of STSs identified prognosis subtypes

To obtain the optimal cluster subtypes, the obtained 1445 CpGs sites of 258 STSs samples were employed for the Consensus

Table 1

The top 20 most significantly different methylation sites regarding prognosis.

CpGs	HR	Lower 95% CI	Upper 95% CI	*P*
cg09347923	1.17E+04	8.79E+02	1.57E+05	1.37E-12
cg00579036	4.92E+02	7.91E+01	3.06E+03	3.02E-11
cg07058109	1.12E+03	1.36E+02	9.29E+03	7.15E-11
cg16174121	1.58E+02	3.44E+01	7.27E+02	7.93E-11
cg19357499	2.01E+01	7.89E+00	5.12E+01	3.20E-10
cg19112957	1.56E+02	3.21E+01	7.57E+02	3.70E-10
cg08508337	9.85E+00	4.81E+00	2.02E+01	3.85E-10
cg07691531	2.61E+01	9.34E+00	7.28E+01	4.80E-10
cg11491074	1.04E+05	2.59E+03	4.18E+06	8.79E-10
cg22982767	6.23E+01	1.64E+01	2.36E+02	1.25E-09
cg26132723	1.42E-03	1.70E-04	1.18E-02	1.34E-09
cg22271305	1.21E+02	2.56E+01	5.72E+02	1.40E-09
cg12213680	2.11E-02	6.01E-03	7.42E-02	1.77E-09
cg10662943	1.77E-02	4.64E-03	6.77E-02	3.71E-09
cg19912470	3.00E-03	4.32E-04	2.09E-02	4.35E-09
cg17133388	9.94E+05	9.75E+03	1.01E+08	4.84E-09
cg19853703	4.85E+01	1.32E+01	1.79E+02	5.13E-09
cg08588180	1.13E+04	4.75E+02	2.69E+05	7.83E-09
cg04217927	1.13E+02	2.25E+01	5.68E+02	9.47E-09
cg12818699	4.38E-02	1.48E-02	1.30E-01	1.56E-08

CI = confidence interval; HR = hazard ratio.

clustering, and the intercluster variation coefficient and average cluster consistency were both calculated for the number of each cluster. As shown in Figure 2A and B, the area under the CDF curve tended to be stable after 4 clusters. Meanwhile, according to the result of consistent clustering (Fig. 3A), Cluster 4 had a stable clustering result with small variation coefficient, and there were no extremely small sample sizes in Cluster 4. Therefore, k = 6 was regarded as the optimal cluster number for the subsequent analysis.

Heatmap annotated by labels of clinical characters was generated corresponded to the DNA methylation classification (Fig. 3B). It was discovered that most CpG sites had high abundance in each sample, besides, the DNA methylation profiles of the 4 categories were obviously different, while Cluster 3 has the lowest methylation level and Cluster 1 has the highest methylation level.

3.3. Predictive value of DNA methylation clustering for STSs clinical characteristics

Kaplan–Meier and log-rank tests showed that the patients in 4 different molecular subtypes showed a significant diversity OS

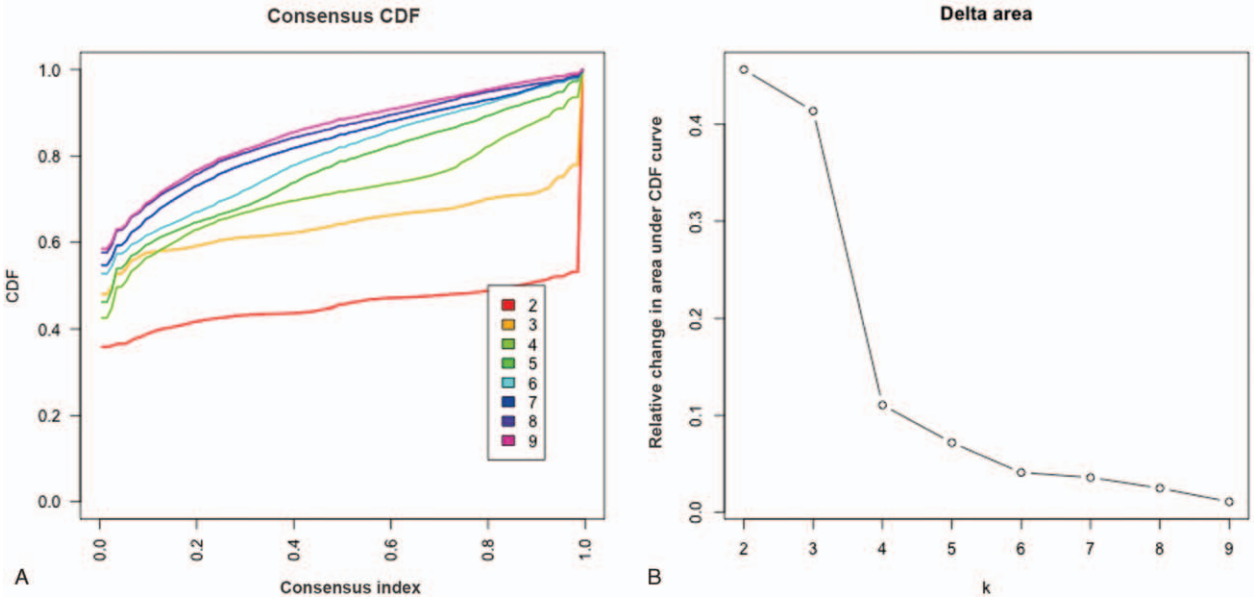

Figure 2. Consensus clustering of DNA methylation-based prognostic subgroups. (A) The consensus among clusters for each category number k. (B) Delta area curve of consensus clustering, which indicates the relative change in the area under the cumulative distribution function (CDF) curve for each category number k compared with k-1. The horizontal axis represents the category number k and the vertical axis represents the relative change in the area under the CDF curve.

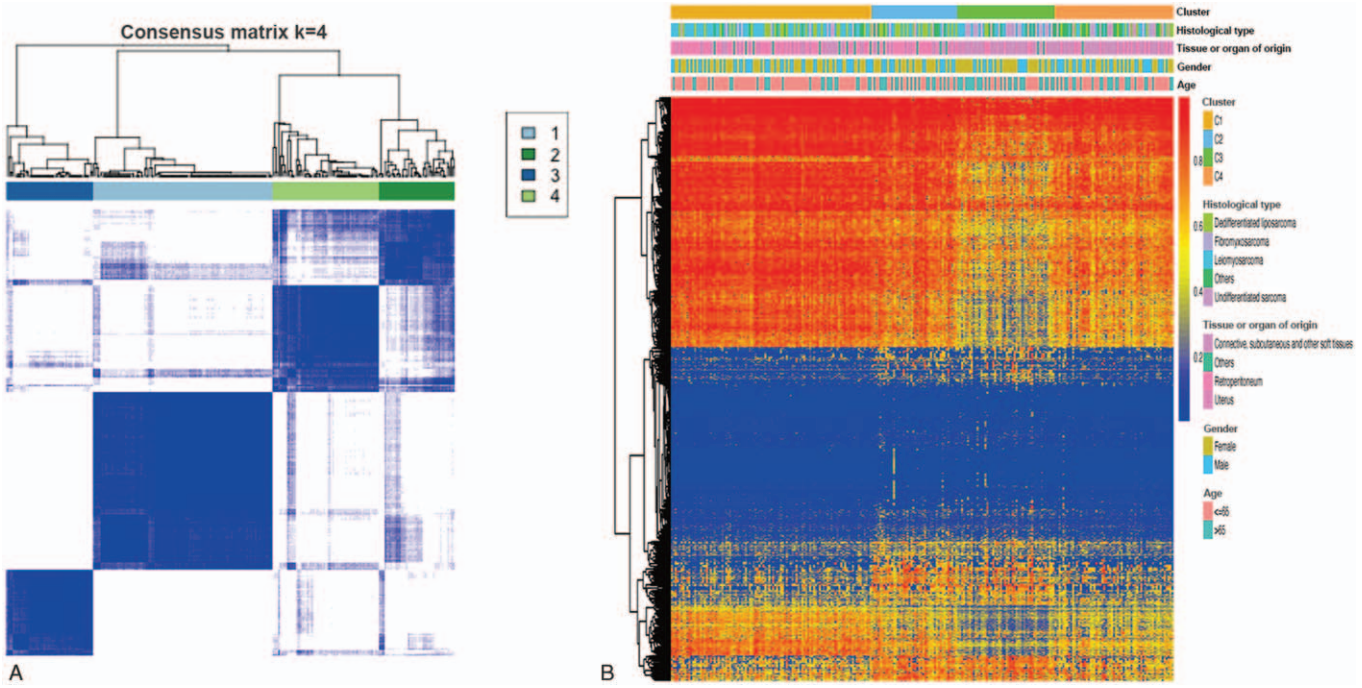

Figure 3. Cluster Analysis of 4 subtypes with the corresponding heat map. (A) The heat map corresponding to the consensus matrix for 4 molecular subtypes obtained by applying consensus clustering. (B) The heatmap corresponding to the dendrogram in the figure A, which was generated using the pheatmap function in R with DNA methylation classification, age, gender, histological type, and tissue or organ of original as the annotations.

(Fig. 4A). Among them, Cluster 1 had the best survival rate, whereas Clusters 2 and 3 had a dismal prognosis, indicating that lower DNA methylation levels may be connected with the poorer OS in STSs patients. Indeed, these 4 molecular subtypes were also found significantly related to some clinical characteristics of each sample. As shown in Figure 4B, Cluster 1 was highly enriched with age $<=65$, and more older patients were enriched in Cluster 2 and 3. Meanwhile, Figure 4D indicated Cluster 1 had more leiomyosarcoma and dedifferentiated liposarcoma, Cluster 3 had more fibromyxosarcoma and undifferentiated sarcoma. Figure 4E implicated more sarcoma in Clusters 1 and 2 were original sited in retroperitoneum, and sarcoma in Clusters 3 and 4 were mainly enriched in connective, subcutaneous and other soft tissues. Figure 4C demonstrated no difference in gender among these 4 subgroups. These results indicate that these DNA methylation profiles could serve as prognosis markers to better understand the OS of STSs patients, more importantly, they may also demonstrate as clinical biomarkers for some clinical features prediction of SKSs patients (including age, histological types, and original sites).

3.4. Identification of CpGs corresponded genes and pathway enrichment analysis

The 1445 methylation sites in each subtype were subsequently subjected to genomic annotations and 1268 corresponded genes were identified, meanwhile, the expression profile of thus CpG sites associated genes were also extracted for plotting a heat map. As shown in Figure 4F, each subgroup displayed a variation in the gene expression levels, which indicating a partial consistency between the methylation modifications and their corresponding genes expression.

The enrichment analysis of corresponding gene functions by GO and KEGG enrichment were the conventional means to investigate the molecular mechanism of these pathogenesis methylation sites in STSs. The KEGG enrichment suggested that these genes were enriched in 16 KEGG pathways, especially belonged to Olfactory transduction (Fig. 5D). GO enrichment analysis also draw the conclusion that these genes were mainly related to Olfactory-related biological processes, like olfactory receptor activity, sensory perception of smell, and cellular component of synaptic and postsynaptic membrane (Fig. 5A-C). In addition, these genes were also closely correlated with tumors progression, such as Cell adhesion process, apoptosis, SMAD pathway, phosphatidylinositol 3–kinase (PI3K) pathway, intermediate filament cytoskeleton formation, and so on. These results indicated that the CpG sites in this study were related with Olfactory transduction and tumors progression of STSs.

3.5. Identification of specific DNA methylation sites and prognosis model construction

To find the specific methylation sites, the differences of the 1,445 CpG sites in STSs subgroups were further compared, resulting in 232 CpGs were identified as cluster specific methylation sites. The heat map declared in Figure 6A revealed that the most of the specific methylation sites were located in Clusters 1 and 3, and the most of which in Cluster 1 were hypomethylated sites as compared with other subgroups (Fig. 6B). In addition, result of Cluster 1 was connected with the best prognosis among all clusters, so the specific CpG sites in Cluster 1 were selected for the further construction. However, due to a large number of CpG sites in Cluster 1 (81 CpG sites) was not good for clinical prediction, thus, the LASSO algorithm was employed for the

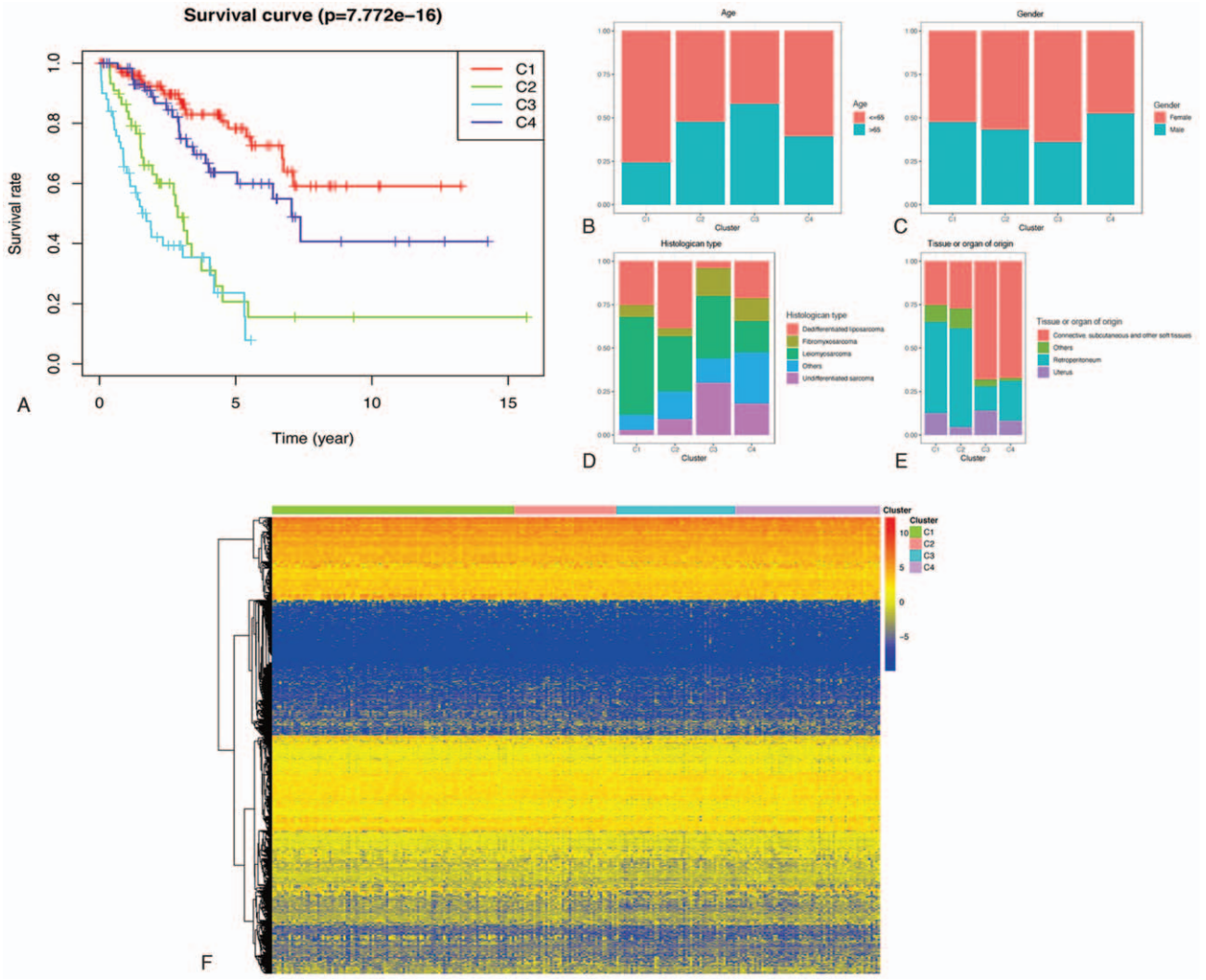

Figure 4. Characterization of different features between each DNA methylation cluster. (A) The survival curves of each sub-clusters indicate the prognostic differences among STSs patients. The distribution of age, (B)gender, (C)histological type, (D)and tissue or organ of original € in each DNA methylation subgroups. (F) The heatmap of methylation sites annotated genes distribution in 4 DNA methylation clusters.

specific CpG site range shrinkage (Fig. 7A and B). Consequently, 13 CpG sites were selected for the calculation of risk score, and all STSs patients were separated into low- and high-risk groups. As Survival analysis indicated in Figure 7D, the OS of STSs samples was gradually decreased with an increased risk score, which declared that this predicting model was significantly correlated with the prognosis of STSs patients. And Figure 7D also discovered that most of the methylation levels were increased in high-risk subgroup, except methylation sites cg15094605 and cg27321439.

3.6. The great predictive accuracy of constructed model for STSs patients

Overall, the predictive accuracy and stability was also analyzed to determine the function of the prognostic prediction model. As shown in Figure 7C, there was a significant inverse correlation between the OS and risk valuation, indicating the remarkably better prognosis of low-risk samples as compared with the higher one. And the area under the ROC curve all reached over 0.79 revealed the prediction model was quietly precised to predict 1-, 3-, and 5-year survival rates for STSs patients (Fig. 8B). Furthermore, the heatmap revealed the expression of the 13 specific CpG sites in each group (Fig. 8A), and there were significant differences between the 2 prognosis model groups with respect to age, tissue or organ of original, and histological type. The nomogram is another application tool for predicting clinical outcomes. In this exploration, the nomogram of risk score and others clinical characters showed that risk score will contribute to the most points in the model, compared to other traditional clinical parameters (Fig. 8E). These results indicated that this prognostic predictor showed great promise for predicting STSs outcomes and clinical features.

To determine whether the risk signature was the independent prognostic factor, the univariate and multivariate Cox regression analyses were also performed in the next exploration (Fig. 8C and D). In the univariate analysis, the risk score and age were both showed obviously connected with the STSs prognosis, and similar

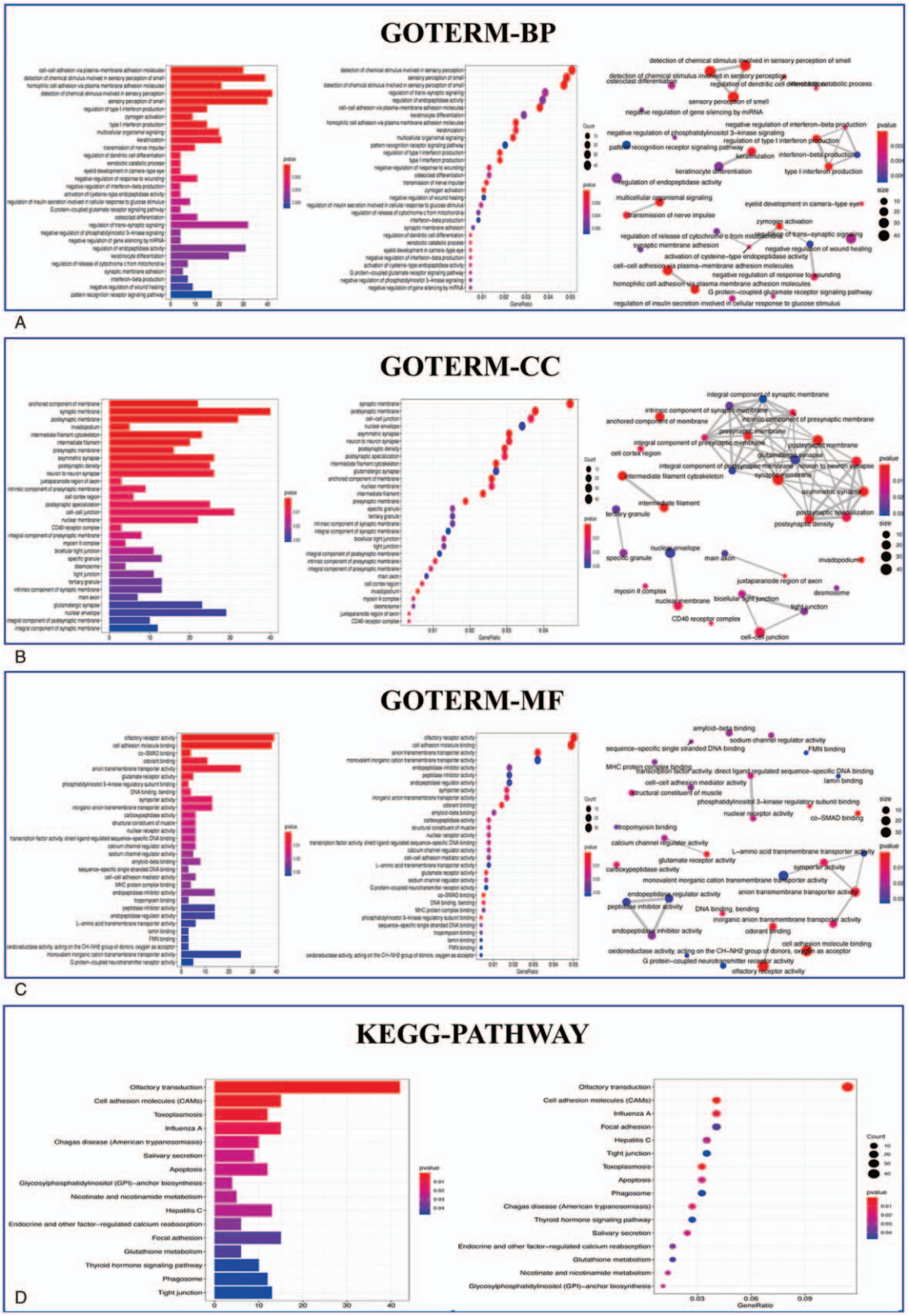

Figure 5. Functional enrichment analysis of methylation sites annotated genes. (A-C) Top 30 classes of GO enrichment terms in biological process (BP), cellular component (CC), and molecular function (MF). (D) KEGG enrichment analysis of CpG sites. In each bubble plot, the size of the dot represents the number of enriched genes.

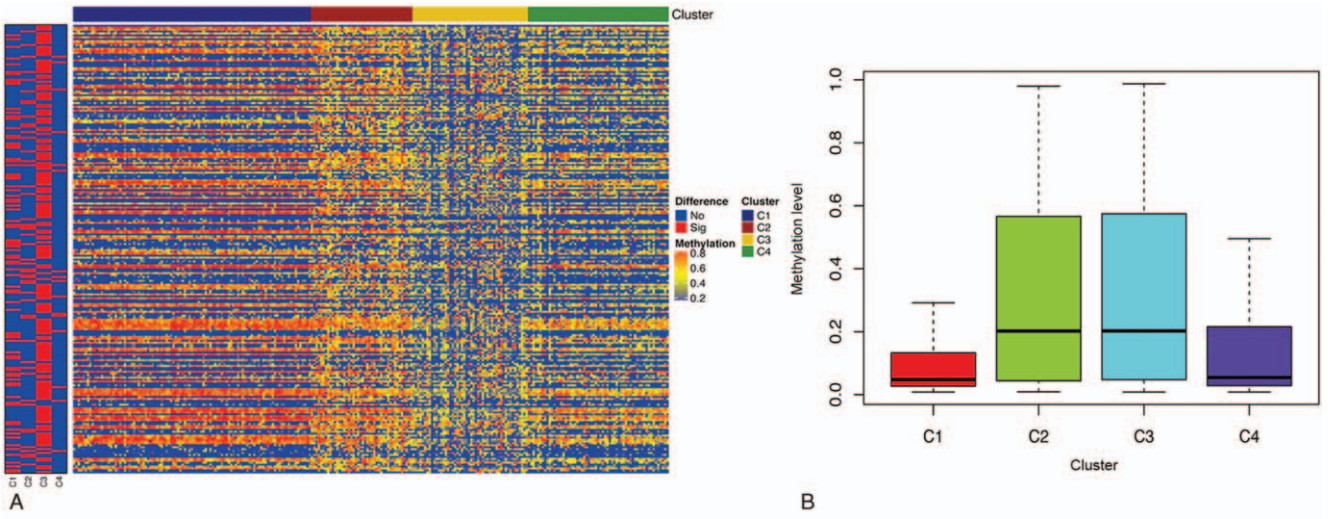

Figure 6. The specific CpG sites for each DNA methylation cluster. (A) The distribution of specific methylation sites in each DNA methylation prognostic subtype. (B) The boxplot based on the specific methylation sites in Cluster 1 for comparison the methylation level in each subgroup.

results were also discovered in the multivariate analysis, thus indicating that the risk score from specific CpG sites could be identified as an independently prognosis feature in STSs.

4. Discussion

STSs is a rare group of malignancies with 50 histological subtypes and performs differing in behavior, biology, and sensitivity to treatment.[31] However, therapies for each subtype remains similar in situ STSs that surgical resection is the main method and supplemented with radiotherapy.[2] The application of appropriate biomarkers is critical in tumor biology for prediction or risk stratification. For example, genotyping methods based on genomics have been widely used to classify tumors, thus conducting the clinical trials. Recent studies have confirmed that DNA methylation provide insights into various tumor early diagnosis, molecular classification, and precise treatment.[12,28,32] Meanwhile, aberrant DNA methylation has been regarded as one of the hallmarks of cancer tissues,[33,34] alterations in DNA methylation also play a virtual role in the progression and development of STSs.[18,19] In addition, S. Peter Wu et all had confirmed that methylation-based classifier could be used to provide diagnostic assistance in bone sarcoma.[17] Therefore, we carried out this discovery to indicate the potential application of DNA methylation in STSs epigenomes classification. The TCGA database is a publicly available resource that contains more than 30 large cohorts of human tumors with a comprehensive multidimensional analysis,[35] these large sample sizes are absolutely the basis for us to provide an in-depth understanding of the etiology of STSs. In this study, the whole genome DNA methylation sites corresponding to 269 STSs samples were also obtained from TCGA database and methylation sites in the gene promoter regions were first applied to select the prognosis associated CpG sites. Four specific prognosis subgroups, classified by 1445 intersected independent prognostic CpG sites, were developed to present a molecular stratification for individual tumors, which has the significance of making therapeutic decisions and exploring the biological mechanisms involved in the progression of RCC.

STSs is not only a heterogeneous tumor, but also has been found presenting in almost every part of the human body, and patients with different histological types are significantly distinct in clinical outcomes. Thus, analysis combined with the comprehensive clinical characteristics, such as age, gender, tissue or organ of origin, and histological, might effectively improve the accuracy of prognosis in STSs patients.[36] In this study, the distribution of the disease-specific OS in these 4 distinct prognostic subtypes of STSs was seemed to be predicted, as well as original sites, histological classification, age, and gender distribution. As the results suggested, the survival curves of these 4 specific subgroups were distinct from each other. Meanwhile, our classification scheme also provided an accurate diagnosis for individual tumors. For example, while patients were assigned to Cluster 1, they were found to have high chances to be diagnosed as leiomyosarcoma primary from retroperitoneum and connected with better prognosis, and these results might prompt clinicians to re-evaluate the treatment for STSs patients. Conclusion, our classifications of 4 subtypes based on the DNA methylation sites can classify STSs more accurately and guide clinicians in terms of clinical diagnosing, treating strategies and prognostic judgment of different STSs patients.

Previous researches have reported that CpG island methylation was shown to promote carcinogenesis by disrupting the function of tumor suppressor genes or oncogenes.[37] Furthermore, it is also verified that the hypomethylation of hub genes was significantly correlated to tumor proliferation and metastasis.[38] In terms of tumor progression, methylation has also been identified significantly associated with multiple biological processes and signaling pathways in cancers, including tumor stem cell growth,[39] self-renewal,[40] ultraviolet-induced DNA damage response,[41] focal adhesion pathway,[28] and so on. So, for comprehensively evaluating the mechanism of DNA methylation sites in STSs progression, the enrichment analysis of CpG sites corresponded genes was also applied in this study. Based on the gene expression profiles and CpG sites on 4 subgroups of STSs, we found that the levels of gene expression and DNA methylation were consistent, which revealed that methylation sites might affected the pathogenesis

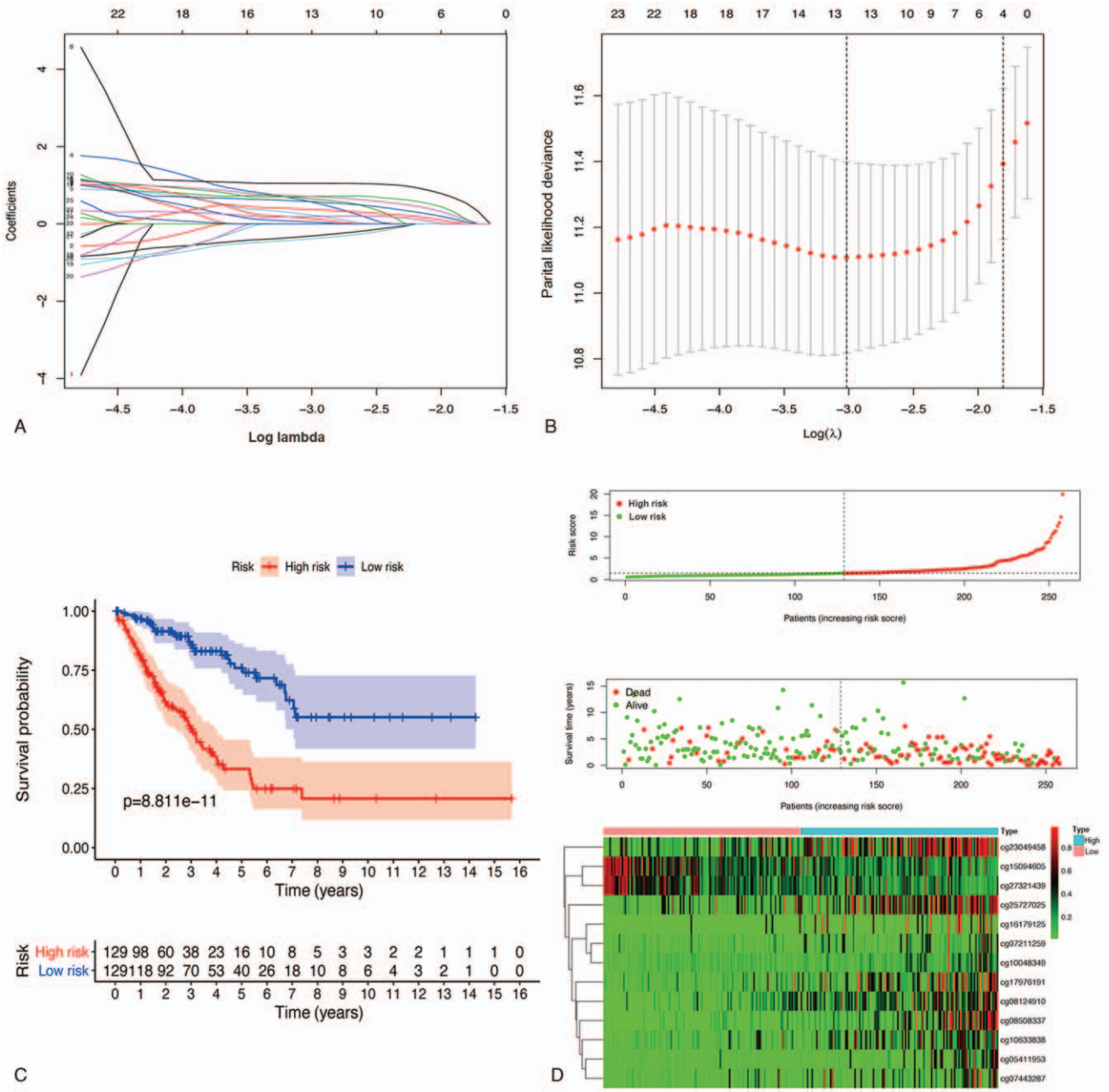

Figure 7. Construction of the prognosis prediction model by LASSO. (A) The changing trajectory of each independent variable. (B) Confidence intervals for each lambda. (C) The Kaplan–Meier overall survival curves were assigned to compare the prognostic difference based on the risk score. (D) The relationship between the methylation profile, overall survival, and risk scores.

of STSs through modulating the expression of corresponded genes. Furthermore, the functional analysis discovered that these annotated genes were enriched in several biological processes and signaling pathways, like Cell adhesion process, apoptosis, SMAD pathway, phosphatidylinositol 3–kinase (PI3K) pathway, intermediate filament cytoskeleton formation, and others. All of these biological processes and signaling pathway has been declared significantly correlated with tumorigenesis and progression.[42–44] These explores could provide clues to emphasize the relationship between these specific methylation sites and STSs biological processes and signaling pathways.

Moreover, in this study, we identified whether these specific methylation sites could be used at the prognostic level and followed by the construction of a STSs prognostic prediction model. Eventually, we focused on the differential CpG sites among 4 clusters and developed a novel 13 methylation sites for prognostication. The constructed risk model was determined with a robust prognostic value and demonstrated to be an independent prognostic factor for STSs. In addition, the risk score was also confirmed with a major advantage of its biological implications for predicting STSs intrinsic histological subtypes and its original sites. A similar scenario was also observed in the nomogram analysis that risk signature played a virtual role in

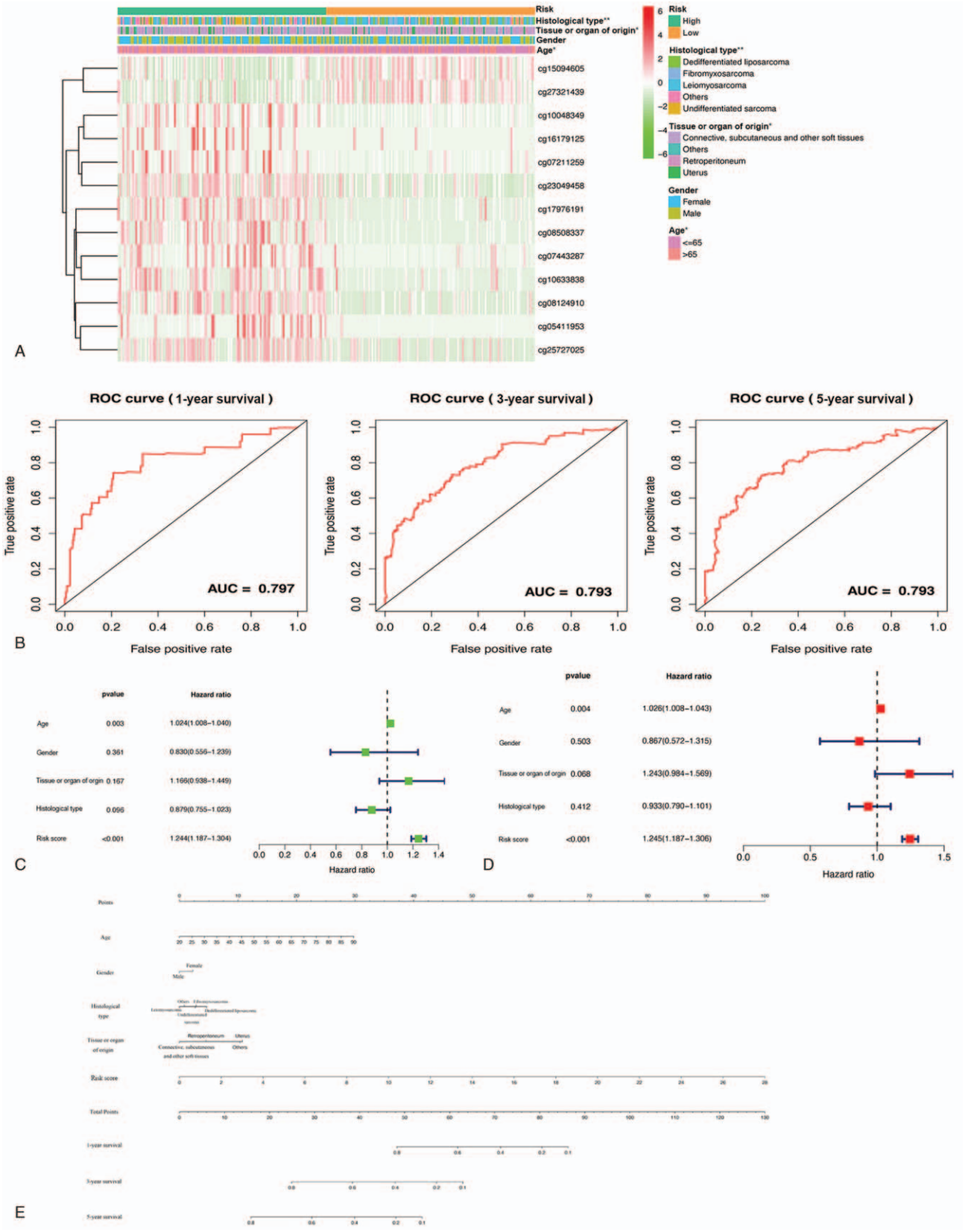

Figure 8. Verifying the prognostic value of prediction model for patients with STSs. (A) The heatmap shows the distribution of clinicopathological features and specific methylation sites expression level in the low- and high-risk groups. *$P < .05$ and **$P < .01$. (B) ROC curves showed the predictive efficiency of the risk signature on the 1-, 3-, and 5-year survival rate. (C) Univariate and multivariate (D) Cox regression analysis of the relationship between clinicopathological features (including the risk score) and overall survival of STSs patients. (E) Nomogram of risk score and other clinical factors for STSs 1-, 3-, and 5-year event prediction.

predicting the OS of STSs, which may be caused by the intensive correlation between the risk signature and STSs pathogenesis.

5. Conclusion

In conclusion, our research identified a new classification of STSs into 4 different prognosis subgroups based on the DNA methylation data. This classification will help to provide more accurate subdivision of STSs and facilitate clinicians to choose a more individualized treatment. Furthermore, the specific CpG sites and corresponding genes in each epigenetic subtype can be used as biomarkers for early diagnosis, precise prognosis prediction, and biological processes and signaling pathways exploration. Most importantly, our study provides a framework to construct a novel classification of molecular subtypes associated with specific tumors.

Author contributions

Conceptualization: Qingjun Wei.
Data curation: Kai Li.
Formal analysis: Kai Li, Zhengyuan Wu.
Methodology: Kai Li.
Project administration: Qingjun Wei.
Writing – original draft: Jun Yao, Jingyuan Fan.
Writing – review & editing: Kai Li, Qingjun Wei.

References

[1] Hoefkens F, Dehandschutter C, Somville J, et al. Soft tissue sarcoma of the extremities: pending questions on surgery and radiotherapy. Radiat Oncol 2016;11:136.
[2] von Mehren M, Randall RL, Benjamin RS, et al. Soft tissue sarcoma, version 2.2018, NCCN clinical practice guidelines in oncology. J Natl Compr Canc Netw 2018;16:536–63.
[3] Casali PG, Abecassis N, Aro HT, et al. Soft tissue and visceral sarcomas: ESMO-EURACAN Clinical Practice Guidelines for diagnosis, treatment and follow-up. Ann Oncol 2018;29(Suppl 4):iv51–67.
[4] Honoré C, Méeus P, Stoeckle E, et al. Soft tissue sarcoma in France in 2015: epidemiology, classification and organization of clinical care. J Visc Surg 2015;152:223–30.
[5] Kotilingam D, Lev DC, Lazar AJ, et al. Staging soft tissue sarcoma: evolution and change. CA Cancer J Clin 2006;56:282–91. quiz 314-285.
[6] Chinese expert consensus on diagnosis and treatment of soft tissue sarcomas (Version 2015). Zhonghua Zhong Liu Za Zhi. 2016;38 (4):310-320.
[7] Scheer M, Dantonello T, Hallmen E, et al. Synovial sarcoma recurrence in children and young adults. Ann Surg Oncol 2016;23(Suppl 5):618–26.
[8] Yang X, Huang WT, He RQ, et al. Determining the prognostic significance of alternative splicing events in soft tissue sarcoma using data from The Cancer Genome Atlas. J Transl Med 2019;17:283.
[9] Xie J, Lin D, Lee DH, et al. Copy number analysis identifies tumor suppressive lncRNAs in human osteosarcoma. Int J Oncol 2017;50:863–72.
[10] Zhu Z, Jin Z, Zhang H, et al. Integrative clustering reveals a novel subtype of soft tissue sarcoma with poor prognosis. Front Genet 2020;11:69.
[11] Koch A, Joosten SC, Feng Z, et al. Analysis of DNA methylation in cancer: location revisited. Nat Rev Clin Oncol 2018;15:459–66.
[12] Chen W, Zhuang J, Wang PP, et al. DNA methylation-based classification and identification of renal cell carcinoma prognosis-subgroups. Cancer Cell Int 2019;19:185.
[13] Lai HC, Lin YW, Huang TH, et al. Identification of novel DNA methylation markers in cervical cancer. Int J Cancer 2008;123:161–7.
[14] Voisin S, Eynon N, Yan X, et al. Exercise training and DNA methylation in humans. Acta Physiol (Oxf) 2015;213:39–59.
[15] Ferreira HJ, Esteller M. CpG islands in cancer: heads, tails, and sides. Methods Mol Biol 2018;1766:49–80.
[16] Tian W, Li Y, Zhang J, et al. Combined analysis of DNA methylation and gene expression profiles of osteosarcoma identified several prognosis signatures. Gene 2018;650:7–14.
[17] Wu SP, Cooper BT, Bu F, et al. DNA methylation-based classifier for accurate molecular diagnosis of bone sarcomas. JCO Precis Oncol 2017.
[18] Peille AL, Brouste V, Kauffmann A, et al. Prognostic value of PLAGL1-specific CpG site methylation in soft-tissue sarcomas. PLoS One 2013;8: e80741.
[19] Knösel T, Altendorf-Hofmann A, Lindner L, et al. Loss of p16(INK4a) is associated with reduced patient survival in soft tissue tumours, and indicates a senescence barrier. J Clin Pathol 2014;67:592–8.
[20] Xie Y, Zong P, Wang W, et al. Hypermethylation of potential tumor suppressor miR-34b/c is correlated with late clinical stage in patients with soft tissue sarcomas. Exp Mol Pathol 2015;98:446–54.
[21] Liu J, Li R, Liao X, et al. Comprehensive investigation of the clinical significance and molecular mechanisms of plasmacytoma variant translocation 1 in sarcoma using genome-wide RNA sequencing data. J Cancer 2019;10:4961–77.
[22] Anders S, Huber W. Differential expression analysis for sequence count data. Genome Biol 2010;11:R106.
[23] Zhang S, Li X, Zong M, et al. Efficient kNN classification with different numbers of nearest neighbors. IEEE Trans Neural Netw Learn Syst 2018;29:1774–85.
[24] Chen X, Zhao C, Zhao Z, et al. Specific glioma prognostic subtype distinctions based on DNA methylation patterns. Front Genet 2019;10:786.
[25] Zhang Y, Li H, Zhang W, et al. LASSO-based Cox-PH model identifies an 11-lncRNA signature for prognosis prediction in gastric cancer. Mol Med Rep 2018;18:5579–93.
[26] Wilkerson MD, Hayes DN. Consensus Cluster Plus: a class discovery tool with confidence assessments and item tracking. Bioinformatics 2010;26:1572–3.
[27] Yu G, Wang LG, Han Y, et al. Cluster Profiler: an R package for comparing biological themes among gene clusters. OMICS 2012;16:284–7.
[28] Li C, Ke J, Liu J, et al. DNA methylation data-based molecular subtype classification related to the prognosis of patients with cervical cancer. J Cell Biochem 2020;121:2713–24.
[29] Heagerty PJ, Zheng Y. Survival model predictive accuracy and ROC curves. Biometrics 2005;61:92–105.
[30] Gu HY, Zhang C, Guo J, et al. Risk score based on expression of five novel genes predicts survival in soft tissue sarcoma. Aging (Albany NY) 2020;12:3807–27.
[31] Sleijfer S, Gelderblom H. Current clinical trials for advanced osteosarcoma and soft tissue sarcoma. Curr Opin Oncol 2014;26:434–9.
[32] Ma H, Zhao C, Zhao Z, et al. Specific glioblastoma multiforme prognostic-subtype distinctions based on DNA methylation patterns. Cancer Gene Ther 2020;27:702–14.
[33] Witt H, Gramatzki D, Hentschel B, et al. DNA methylation-based classification of ependymomas in adulthood: implications for diagnosis and treatment. Neuro Oncol 2018;20:1616–24.
[34] Klutstein M, Nejman D, Greenfield R, et al. DNA methylation in cancer and aging. Cancer Res 2016;76:3446–50.
[35] Tsuboi M, Kondo K, Masuda K, et al. Prognostic significance of GAD1 overexpression in patients with resected lung adenocarcinoma. Cancer Med 2019;8:4189–99.
[36] Huang R, Meng T, Chen R, et al. The construction and analysis of tumor-infiltrating immune cell and ceRNA networks in recurrent soft tissue sarcoma. Aging (Albany NY) 2019;11:10116–43.
[37] Zhang Y, Zhang J. Identification of functionally methylated regions based on discriminant analysis through integrating methylation and gene expression data. Mol Biosyst 2015;11:1786–93.
[38] Hua S, Ji Z, Quan Y, et al. Identification of hub genes in hepatocellular carcinoma using integrated bioinformatic analysis. Aging (Albany NY) 2020;12:5439–68.
[39] Wan Q, Tang J, Han Y, et al. Co-expression modules construction by WGCNA and identify potential prognostic markers of uveal melanoma. Exp Eye Res 2018;166:13–20.
[40] Cui Q, Shi H, Ye P, et al. m(6)A RNA methylation regulates the self-renewal and tumorigenesis of glioblastoma stem cells. Cell Rep 2017; 18:2622–34.
[41] Xiang Y, Laurent B, Hsu CH, et al. RNA m(6)A methylation regulates the ultraviolet-induced DNA damage response. Nature 2017;543:573–6.
[42] Papp B, Launay S, Gélébart P, et al. Endoplasmic reticulum calcium pumps and tumor cell differentiation. Int J Mol Sci 2020;21:
[43] Conti A, Espina V, Chiechi A, et al. Mapping protein signal pathway interaction in sarcoma bone metastasis: linkage between rank, metalloproteinases turnover and growth factor signaling pathways. Clin Exp Metastasis 2014;31:15–24.
[44] Omary MB. IF-pathies": a broad spectrum of intermediate filament-associated diseases. J Clin Invest 2009;119:1756–62.