

Identification of diagnostic biomarkers and immune infiltration analysis in prostate cancer bone metastasis based on EMT-NM- related genes

Ke Chen

Department of orthopedics, Huangshan No.96601 Hospital

Jiayang Yue

The First Clinical Medical College of Nanjing University of Chinese Medicine

Zhaogui Ding

Department of orthopedics, Huangshan No.96601 Hospital

Min Cheng

Department of orthopedics, Huangshan No.96601 Hospital

Ning Yao

School of Basic Medical Sciences, Xinjiang Medical University

Tiansheng Ye

Department of orthopedics, Huangshan No.96601 Hospital

Jia Sha

Department of orthopedics, Nanjing Jinling Hospital, Affiliated Hospital of Medical School, Nanjing University

Zhanqiang Hua

hzq5175@163.com

Department of orthopedics, Huangshan No.96601 Hospital

Research Article

Keywords: prostate cancer¹, bone metastatic², differentially expressed genes³, diagnostic value⁴, immune infiltration⁵

Posted Date: October 6th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10915534/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Identification of diagnostic biomarkers and immune infiltration analysis in prostate cancer bone metastasis based on EMT-NM-related genes

Ke Chen^{1,#}, Jiayang Yue^{2,#}, Zhaogui Ding^{1,#}, Min Cheng¹, Ning Yao³, Tiansheng Ye^{1,*}, Jia Sha^{4,*}, Zhanqiang Hua^{1,*}

¹ Department of orthopedics, Huangshan No.96601 Hospital, Anhui, China.

² The First Clinical Medical College of Nanjing University of Chinese Medicine, Nanjing, China.

³ School of Basic Medical Sciences, Xinjiang Medical University, Xinjiang, China.

⁴ Department of orthopedics, Nanjing Jinling Hospital, Affiliated Hospital of Medical School, Nanjing University, Nanjing, China.

These authors contributed equally to this work and should be considered as co-first authors.

* Correspondence:

Zhanqiang Hua, E-mail: hzq5175@163.com

Jia Sha, E-mail: sjfmmu@163.com

Tiansheng Ye, E-mail: 2981869267@qq.com

Keywords: prostate cancer¹, bone metastatic², differentially expressed genes³, diagnostic value⁴, immune infiltrations⁵.

Abstract: Prostate cancer (PCa) is one of the most common malignancies in men and frequently develops bone metastasis, which is associated with high morbidity and mortality. Biomarkers are essential for the early diagnosis, treatment guidance, and prognostic evaluation of bone metastatic prostate cancer (BMPCa). Therefore, this study aimed to identify epithelial – mesenchymal transition (EMT)- and nucleotide metabolism (NM)-related genes with diagnostic potential in BMPCa and to characterize their associated immune microenvironment features. Transcriptomic profiles of primary PCa and BMPCa samples from GEO datasets were analyzed. Differentially expressed genes (DEGs) associated with EMT and nucleotide metabolism were screened, and multiple machine learning algorithms, including LASSO regression, Boruta, and random forest, were applied to identify diagnostic biomarkers. Subsequently, a biomarker-based nomogram was constructed, and immune infiltration analysis was performed to investigate the tumor immune microenvironment. Among seven candidate genes enriched in nucleotide metabolism, *S100A8* and *PLOD2* were identified as robust diagnostic biomarkers for BMPCa, with AUC values of 0.82 and 0.83 in the training cohort, respectively. The constructed nomogram exhibited excellent diagnostic performance, achieving an AUC of 0.89, and demonstrated potential clinical utility. Furthermore, *S100A8* expression showed a strong positive correlation with myeloid-derived suppressor cell (MDSC) infiltration ($r = 0.803$). Collectively, *S100A8* and *PLOD2* represent promising biomarkers for evaluating the risk of bone metastasis in PCa. The association between *S100A8* expression and MDSC infiltration provides new insights into the immune microenvironment of BMPCa and highlights potential targets for early detection and precision therapeutic intervention.

1 Introduction

Prostate cancer (PCa) is one of the most common malignancies of the male genitourinary system worldwide [1; 2]. Although patients with early-stage prostate cancer can achieve favorable outcomes through surgery and endocrine therapy, advanced PCa is prone to distant metastasis, with bone being the most common metastatic site[3]. The 5-year survival rate for patients with localized prostate cancer approaches 99%, whereas it declines dramatically to approximately 25% in patients with bone metastatic prostate cancer[4]. Epidemiological data indicate that 81% of newly diagnosed PCa cases in the United States are localized disease, whereas this proportion is only 33% in China[5]. Among patients with advanced PCa, approximately 80%–90% will eventually develop bone metastases[6]. The development of bone metastases in patients with PCa is associated with a significantly reduced life expectancy. Moreover, skeletal-related events caused by bone metastases, such as bone pain and pathological fractures, can severely impair patients' quality of life. Currently, there is still a lack of highly sensitive and specific molecular biomarkers for the early diagnosis of bone metastatic prostate cancer (BMPCa) in clinical practice, resulting in many patients missing the optimal window for intervention. Therefore, the early identification of potential biomarkers is of great significance for the development of novel therapeutic agents and the improvement of patient prognosis.

Epithelial-mesenchymal transition (EMT) is a key process by which malignant tumors acquire migratory and invasive capabilities. EMT is a central biological mechanism by which cancer cells acquire migratory and invasive properties. EMT is typically characterized by the downregulation of epithelial markers, such as E-cadherin, and the upregulation of mesenchymal markers, including N-cadherin and Vimentin. These molecular alterations lead to the loss of cell polarity, weakened intercellular adhesion, and enhanced motility, ultimately promoting the formation of distant metastatic lesions^[7; 8]. Previous study has demonstrated aberrant expression of EMT-related marker proteins, including E-cadherin and Vimentin, in recurrent or metastatic thyroid carcinoma tissues, suggesting that thyroid cancer metastasis is closely associated with EMT^[9]. Notably, EMT-related markers are also widely expressed in prostate cancer tissues. During prostate cancer progression, tumor cells can undergo EMT, characterized by a morphological transition from a cuboidal epithelial phenotype to a spindle-shaped mesenchymal phenotype^[10]. The molecular regulatory network governing EMT plays a central role in prostate cancer metastasis. Zhou et al. demonstrated that KIF1A promotes the EMT process in prostate cancer by interacting with OGT to regulate protein expression^[11]. Ge et al. demonstrated that ASPN enhances the stemness and EMT of prostate cancer cells by activating the Wnt/ β -catenin signaling pathway, increasing β -catenin expression and promoting its nuclear accumulation, thereby facilitating prostate cancer metastasis^[12]. During tumor progression, EMT endows cancer cells with the ability to breach the basement membrane, enter the circulation, and colonize distant organs. Study has shown that EMT plays an indispensable driving role in the migration and invasion of prostate cancer cells toward the bone microenvironment^[13]. Meanwhile, metabolic reprogramming, a hallmark of cancer, provides the material basis for the rapid proliferation and metastasis of cancer cells. Aberrantly activated nucleotide metabolism (NM) not only supports the extensive synthesis of DNA and RNA required for tumor cell proliferation[14], but also regulates the tumor micro-environment through nucleotide-mediated signaling[15]. However, although both EMT and NM play critical roles in tumor metastasis, studies integrating these two processes remain limited.

Immune cell infiltration, a key driving force in the remodeling of the tumor immune microenvironment (TIME), plays a decisive role in the progression and distant metastasis of malignant tumors[16]. The establishment of prostate cancer bone metastases is a highly complex process that depends not only on the intrinsic invasive and proliferative capacities of tumor cells but also on the dynamic interactions between cancer cells and various immune cell populations within the host bone

marrow microenvironment[17]. Studies have shown that aberrant alterations in immune cell infiltration within bone metastatic lesions promote tumor immune evasion, thereby creating a microenvironment conducive to cancer cell colonization and metastatic establishment[18; 19]. Therefore, it holds significant clinical importance for elucidating the regulatory mechanisms of the microenvironment and identifying potential targets for combined immunotherapy.

The present study aimed to identify specific diagnostic biomarkers for BMPCa by integrating the analysis of EMT and NM-related genes. Prostate cancer and bone metastasis cohorts from the GEO public database were analyzed in combination with multiple machine learning approaches. Furthermore, we constructed a clinical predictive nomogram and systematically evaluated the immune cell infiltration landscape within bone metastatic lesions. We also comprehensively investigated the phenotypic associations between the identified hub biomarkers and key immune cell populations. Our findings may provide a novel theoretical basis for the early diagnosis and precision microenvironment-targeted therapy of bone metastatic prostate cancer.

2 Materials and Methods

2.1 Data sources

The prostate cancer GSE32269 transcriptome dataset was downloaded from the GEO database (<https://www.ncbi.nlm.nih.gov>) as the training set (22 PCa and 29 BMPCa samples, microarray platform GPL96). The GSE77930 dataset was downloaded as the validation set (22 PCa and 20 BMPCa samples, microarray platform GPL15659). The platform annotation files were used to map probes to gene symbols. For multiple probes mapping to the same gene, the average expression value was calculated. A total of 5,596 epithelial-mesenchymal transition (EMT)-related genes and 605 nucleotide metabolism (NM)-related genes were retrieved from the GeneCards database (<https://www.genecards.org/>). The keywords “epithelial-mesenchymal transition” and “nucleotide metabolism” were used for searching the GeneCards database to obtain the corresponding gene sets. The retrieved EMT-related and NM-related genes are provided in Supplementary Table S1 and Table S2.

2.2 Identification of differentially expressed genes (DEGs)

Differential expression analysis was performed to identify genes exhibiting significant differences in expression levels between different sample groups. To identify differentially expressed genes (DEGs) between the PCa and BMPCa groups in the training cohort, differential expression analysis was performed using the "limma" package (v3.64.0; <https://bioinf.wehi.edu.au/limma/>). Non-coding RNAs were excluded from the analysis. Genes with an adjusted $P < 0.05$ and an absolute $\log_2$ fold change ($|\log_2FC|$) > 1.5 were considered as DEGs.

2.3 Candidate gene prediction

The R package "VennDiagram" was used to generate Venn diagrams, and the overlapping genes among DEGs, EMT-related genes, and NM-related genes were identified. These intersecting genes were subsequently regarded as candidate genes for further analyses.

2.4 GO and KEGG enrichment analysis of candidate genes

Gene Ontology (GO) consists of three categories that describe gene functions from different perspectives, including molecular function (MF), cellular component (CC), and biological process

(BP). Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway analysis was performed using the KEGG database to annotate the identified genes and determine the major metabolic and signaling pathways in which these genes are involved. Gene Ontology (GO) and KEGG enrichment analyses of the candidate genes were performed using the R package "clusterProfiler". A P value < 0.05 was considered indicative of statistically significant enrichment.

2.5 Machine learning

LASSO regression analysis, Boruta feature selection, and random forest (RF) analysis were performed on the candidate genes using all samples in the training cohort. LASSO regression was performed using the candidate genes to construct a machine learning model, with the training cohort serving as the input dataset. To reduce the dimensionality of the features, the "glmnet" package in R (version 4.1-8; <https://cran.r-project.org/web/packages/glmnet/index.html>) was used to perform least absolute shrinkage and selection operator (LASSO) logistic regression analysis. The parameters were set as follows: family = "binomial" and nfolds = 5. Strongly associated features were selected, and coefficient profiles as well as cross-validation error curves were generated to identify candidate feature genes. In this study, the Boruta algorithm was implemented using the R package "Boruta" (version 8.0.0; <https://cran.r-project.org/web/packages/Boruta>) to screen candidate feature genes. Random forest (RF) analysis was performed using the R package "randomForest" (version 4.7-1.2; <https://www.stat.berkeley.edu/breiman/RandomForests/>). The optimal model was constructed by iteratively evaluating different numbers of trees (ntree) and automatically selecting the optimal tree number. Feature importance was assessed based on the mean decrease in accuracy (MeanDecreaseAccuracy). The ranked feature importance plot was subsequently generated, and the top five genes with the highest importance scores were retained. Finally, the candidate feature genes identified by the three machine learning algorithms were intersected to obtain the final feature genes.

2.6 Expression analysis

Expression data for the feature genes were extracted from the training and validation cohorts. The Wilcoxon rank-sum test was used to calculate and visualize the expression levels of the feature genes between the PCa and BMPCa groups. Genes exhibiting significant differential expression ($P < 0.05$) with consistent expression trends in both the training and validation cohorts were selected as candidate biomarkers.

2.7 ROC verification of candidate biomarkers

Receiver operating characteristic (ROC) curves of the candidate biomarkers were generated using the R package "pROC" (version 1.18.5; <https://cran.r-project.org/web/packages/pROC/index.html>), and the area under the curve (AUC) was calculated. Candidate biomarkers with an AUC > 0.7 were further selected to identify the final biomarkers.

2.8 Construction and validation of the flowchart

A nomogram was constructed in the training cohort based on the identified biomarkers and their expression levels using the R package "rms" (version 8.0-0; <https://cran.r-project.org/web/packages/rms/index.html>). ROC curves were used to evaluate the ability of the model to discriminate between different outcome categories. ROC curves illustrate the trade-off between the true positive rate (sensitivity) and the false positive rate (1-specificity) across different decision thresholds. ROC curve analysis was performed using the R package "pROC" (version 1.18.5; <https://xrobin.github.io/pROC/>) to evaluate the discriminatory ability of the model. The area under the

165 curve (AUC) and its corresponding 95% confidence interval (CI) were calculated to quantify the
166 overall discriminative performance of the model. An AUC value closer to 1.0 indicates superior
167 discriminatory ability.

168 Calibration curves were used to evaluate the accuracy of the model in predicting the probability
169 of clinical outcomes at the individual level. The calibration curve reflects the agreement between the
170 predicted probabilities and the observed outcomes. Calibration analysis was performed using the R
171 package "regplot" (version 1.1;
172 <https://mirrors.sustech.edu.cn/CRAN/web/packages/regplot/index.html>) to assess the predictive
173 accuracy of the model, and the results of the Hosmer–Lemeshow goodness-of-fit test were visualized.

174 Decision curve analysis (DCA) was performed to evaluate the clinical utility of the nomogram.
175 The R package "ggDCA" (version 1.2; <https://rpkg.net/package/ggDCA>) was used to generate DCA
176 curves. A positive net benefit across threshold probabilities ranging from 0 to 1 indicated that the
177 constructed model was clinically useful and effective. The predictive performance of the nomogram
178 was comprehensively assessed using ROC curve analysis, calibration curves, and DCA.

179 2.9 GSEA analysis

180 Based on the Molecular Signatures Database (MSigDB), the C2: Curated gene sets (KEGG
181 pathways) were obtained using the R package "msigdb" and used as the reference gene sets for GSEA
182 analysis. The C2 gene set collection of Homo sapiens was downloaded, and KEGG-related gene sets
183 were extracted according to gene set names containing "KEGG". Using all samples in the training
184 cohort, the R package "psych" was employed to calculate Pearson correlation coefficients between
185 each biomarker and all other genes. Genes significantly correlated with each biomarker ($P < 0.05$) were
186 selected to construct biomarker-associated gene sets. Single-gene gene set enrichment analysis (GSEA)
187 was performed for each biomarker using the R package "clusterProfiler" to investigate the associations
188 between biomarkers and signaling pathways. Pathways with $P < 0.05$ were considered statistically
189 significant.

190 2.10 Immune infiltration analysis

191 To explore the immune infiltration patterns in the PCa and BMPCa groups, single-sample gene
192 set enrichment analysis (ssGSEA) implemented in the gsva function of the R package "GSVA" (version
193 2.2.0; <https://github.com/rcastelo/GSVA>) was applied to the training cohort to assess the immune
194 activity of each sample. The Wilcoxon rank-sum test was used to identify immune cells that differed
195 significantly between the two groups ($P < 0.05$). Correlation analysis between the identified biomarkers
196 and differentially infiltrated immune cells was performed using the R package "psych", and the R
197 package "pheatmap" (version 1.0.12; <https://cran.r-project.org/web/packages/pheatmap/index.html>)
198 was used to visualize the correlation results in a heatmap.

199 3 Results

200 3.1 Identification of DEGs and candidate genes

201 Based on the training dataset, differential expression analysis was performed using the "Limma"
202 package ($p < 0.05$ and $|\log_2FC| > 1.5$), and a total of 201 DEGs were identified, including 106
203 upregulated and 95 downregulated genes in the BMPCa group. To comprehensively characterize the
204 distribution of DEGs, a volcano plot was generated using the "ggVolcano" R package to visualize the
205 differentially expressed genes. In the volcano plot, the top five upregulated and top five downregulated

genes ranked by $|\log_2FC|$ were annotated. To further visualize the results, heatmaps were generated using the “ComplexHeatmap” R package, illustrating the expression patterns of the top five DEGs ranked by $|\log_2FC|$ in the BMPCa and PCa groups (Figure 1A–B). Based on 201 DEGs between the BMPCa and PCa groups, 5,596 EMT-related genes, and 605 NM-related genes, a Venn diagram was generated using the “Venn Diagram” R package to identify their intersections. A total of seven intersecting genes were obtained as candidate genes (Figure 1C).

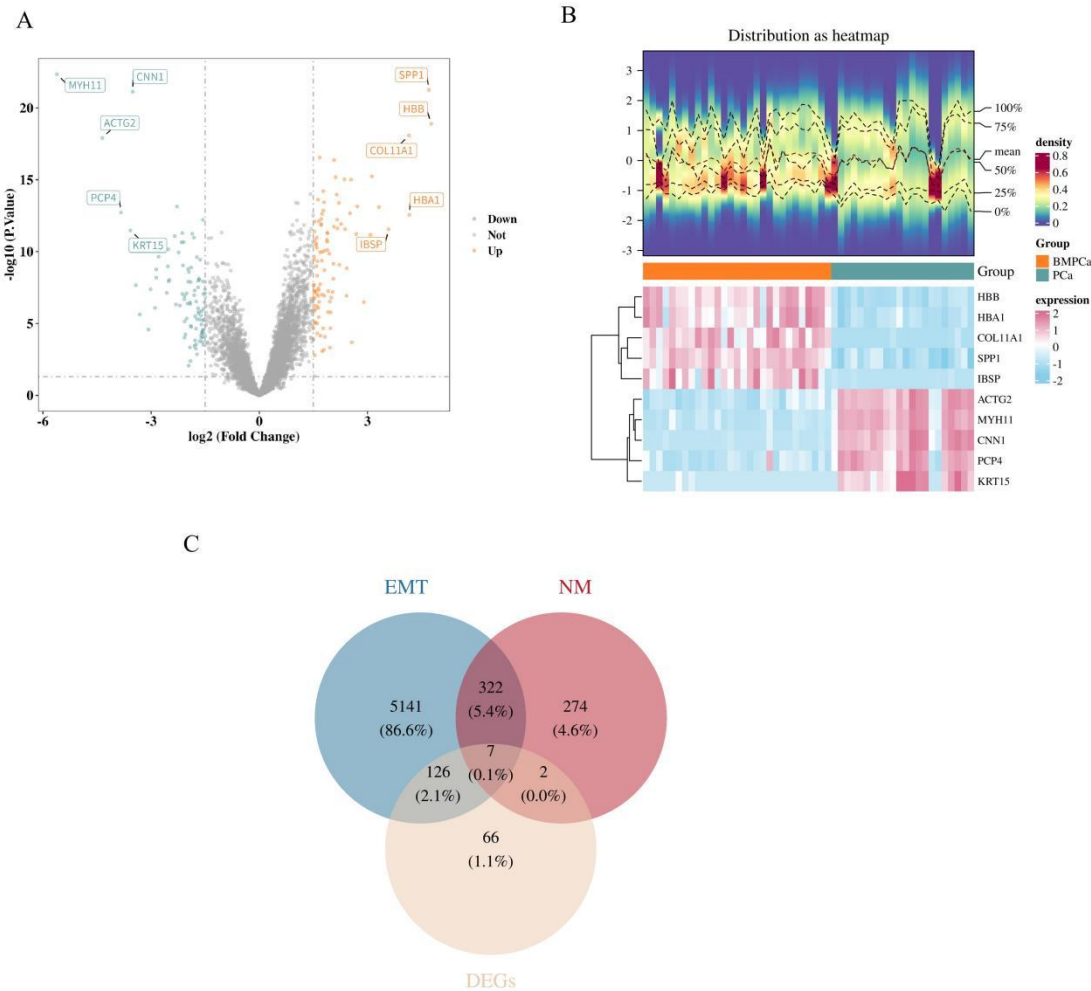


Figure 1. Analysis of DEGs and candidate genes. (A) Volcano plot of differentially expressed genes; (B) Heatmap of gene expression levels of DEGs; (C) Venn diagram identifying candidate genes. Note: (B) The upper panel displays the expression density heatmap of the top 10 genes with the largest fold change across samples, while the lower panel shows the gene expression heatmap. The colors indicate the gene expression levels in the samples (red represents high expression, and blue represents low expression). (C) Different color blocks represent distinct gene categories, and the numbers indicate the gene counts.

3.2 GO and KEGG enrichment analysis of candidate genes

The seven obtained candidate genes were subjected to Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) enrichment analyses using the “clusterProfiler” R package to investigate their associated biological functions. GO enrichment analysis was performed with a significance threshold of $p < 0.05$, yielding 174 biological process (BP) terms, 10 cellular

component (CC) terms, and 30 molecular function (MF) terms. As shown in the figure, the top five enriched pathways with the highest number of associated genes in each category are presented (Figure 2A).

Candidate genes in the top 5 pathways of the BP: deoxyribonucleotide biosynthetic process, 2'-deoxyribonucleotide biosynthetic process, deoxyribose phosphate biosynthetic process, 2'-deoxyribonucleotide metabolic process, deoxyribonucleotide metabolic process. Candidate genes in the TOP5 pathway of CC: nuclear membrane, collagen-containing extracellular matrix, nuclear envelope, NLS-dependent protein nuclear import complex, nucleocytoplasmic transport complex. Candidate genes in the TOP5 pathways of MF: carboxylic acid binding, organic acid binding, vitamin binding, iron ion binding, microtubule binding.

KEGG enrichment analysis was performed with a significance threshold of $p < 0.05$, identifying 10 enriched pathways in total. The pathways were ranked in descending order based on the number of enriched genes, and the top five pathways are shown in the figure, namely pyrimidine metabolism, nucleotide metabolism, antifolate resistance, folate transport and metabolism, and one carbon pool by folate (Figure 2B).

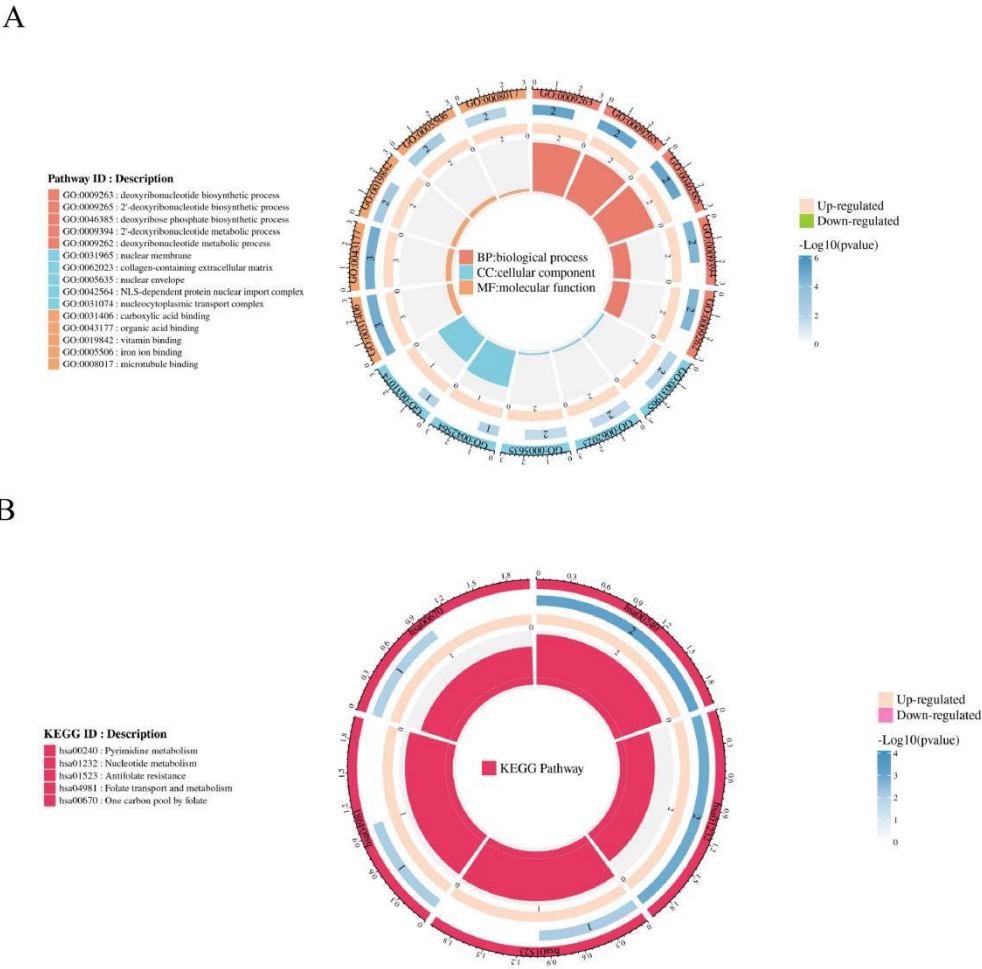


Figure 2. Enrichment analysis of the candidate genes. (A) GO enrichment analysis. (B) KEGG enrichment analysis. Note: (A) In the second ring, the numbers above the blue band indicate the number of enriched genes. In the third ring, pink and green represent upregulated and downregulated genes,

respectively. In the innermost ring, the height of the bars corresponds to the enrichment score. (B) In the outermost ring, red represents the top 5 pathways with the most enriched genes. In the second ring, the numbers above the blue band denote the number of enriched genes. In the innermost ring, taller bars indicate higher enrichment scores.

3.3 Characteristics gene

The seven candidate genes identified above were used to construct a LASSO logistic regression model based on the training dataset to select strongly correlated features. As a result, two plots were generated, including a lambda selection curve and a LASSO logistic regression coefficient penalty plot. As shown in the figure, the optimal lambda value (0.009385354) was selected at the minimum error, resulting in the identification of six candidate feature genes (Figure 3A–B). In addition, the Boruta algorithm was applied using the “Boruta” R package to construct the model, which identified seven candidate feature genes (Figure 3C). The RF model was constructed using the “randomForest” R package to rank gene feature importance, and the top five genes were identified and visualized (Figure 3D). By intersecting the results obtained from the three machine learning algorithms, five final feature genes were identified, including *S100A8*, *CENPF*, *PLOD2*, *TYMS*, and *KPNA2* (Figure 3E).

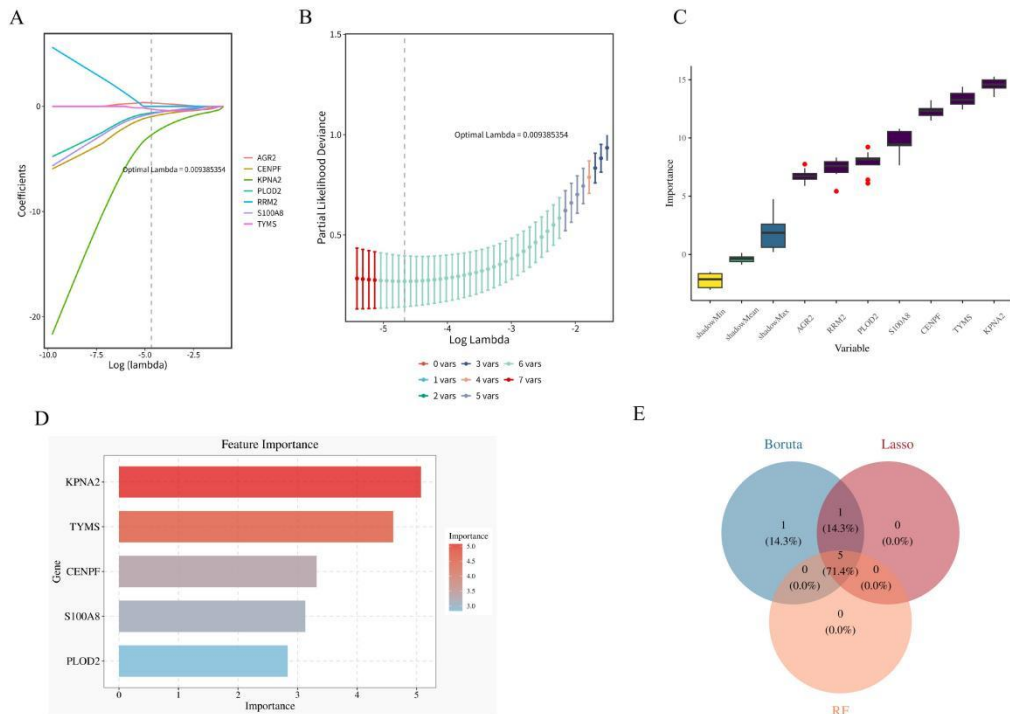


Figure 3. Selection of feature genes using machine learning. (A) LASSO coefficient profiles. (B) Tuning parameter (lambda) selection in the LASSO model. (C) Feature importance boxplot generated by the Boruta algorithm. (D) Top 5 important genes identified by the Random Forest (RF) algorithm. (E) Venn diagram of the candidate feature genes. Note: (A) Each curve represents the coefficient trajectory of an independent variable. The y-axis indicates the coefficient value, and the x-axis represents the number of non-zero coefficients in the model. (B) The red dots represent the mean squared error (MSE) and its standard error, where a smaller MSE indicates better model performance. The numbers at the top indicate the number of independent variables remaining in the model. The first vertical dashed line represents the minimum MSE.

3.4 Expression analysis

To determine the expression patterns of the feature genes in BMPCa and PCa samples, violin plots were generated to visualize the results. The results showed that only *S100A8* and *PLOD2* were significantly overexpressed in the BMPCa group in both the training and validation datasets ($p < 0.05$). Therefore, these two feature genes were selected as candidate biomarkers for subsequent analyses in this study (Figure 4).

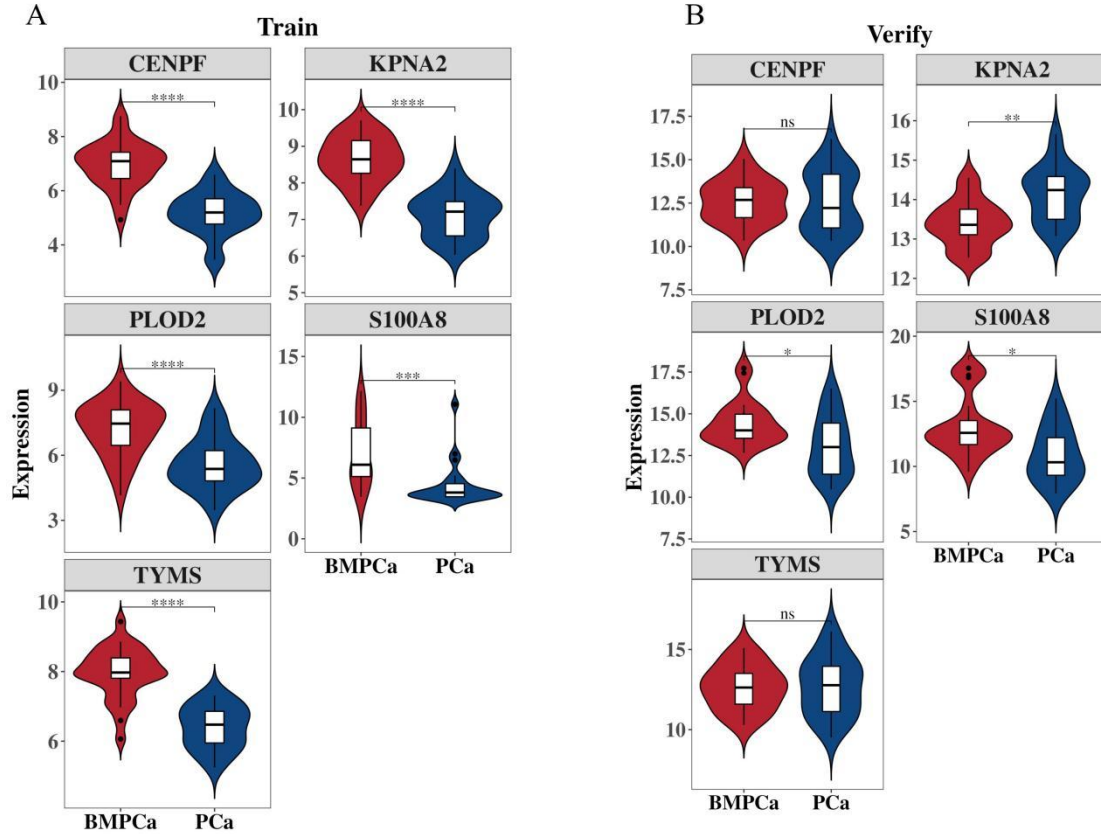


Figure 4. Violin plots of feature gene expression levels. (A) Training set. (B) Validation set. ns, not significant; *, $P < 0.05$; **, $P < 0.01$; ***, $P < 0.001$; ****, $P < 0.0001$.

3.5 ROC verification of candidate biomarkers

ROC analysis of the candidate biomarkers was performed using the R package “pROC”, and the area under the curve (AUC) was calculated. In the training cohort, the AUC values for *S100A8* and *PLOD2* were 0.82 and 0.83, respectively (Figure 5A–B). In the validation cohort, the AUC values for *S100A8* and *PLOD2* were 0.76 and 0.72, respectively (Figure 5C–D). Candidate biomarkers with an $AUC \geq 0.7$, namely *S100A8* and *PLOD2*, were selected as the final biomarkers.

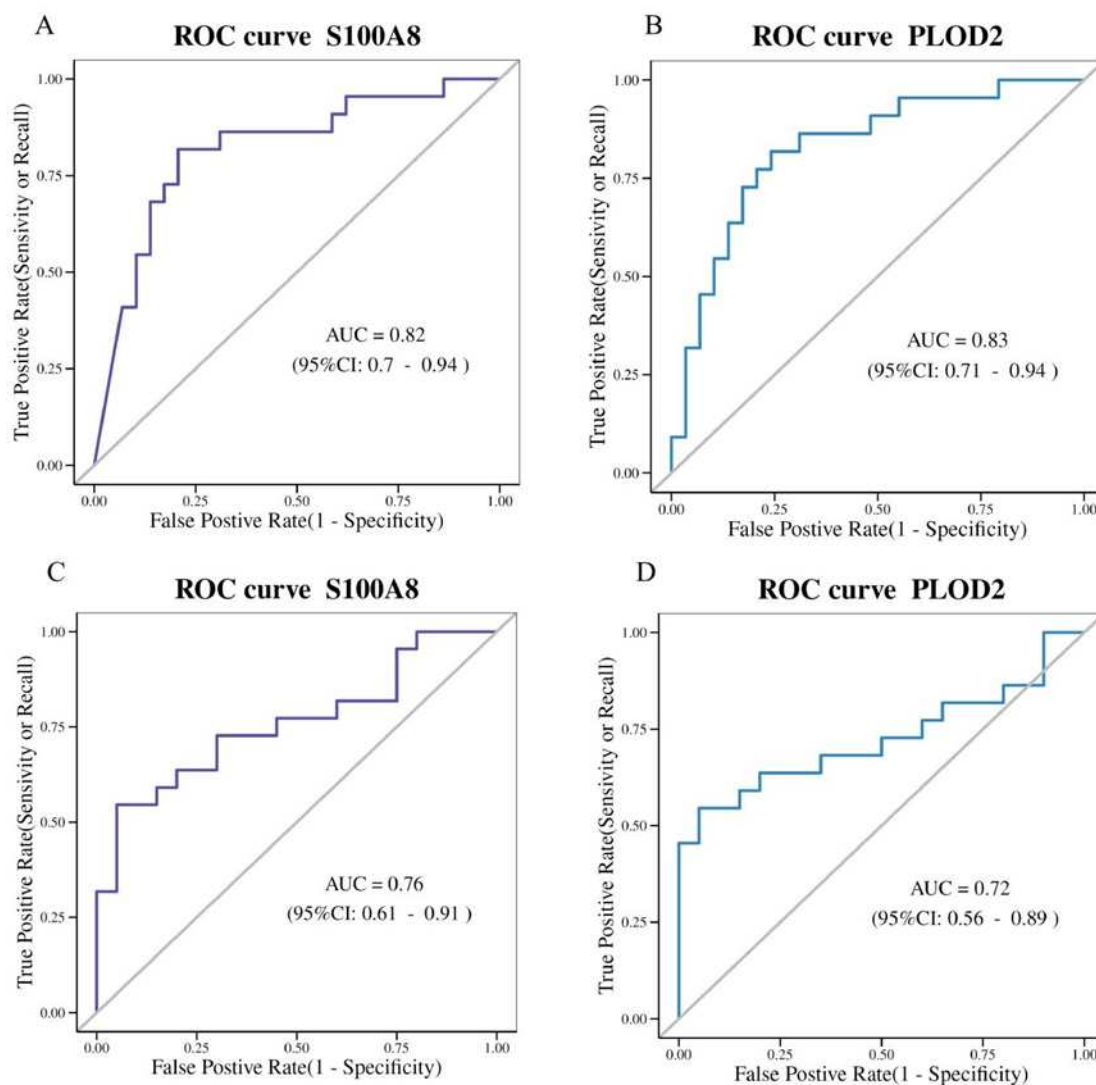


Figure 5. ROC curves in the training and validation sets. (A) ROC curve for the *S100A8* gene in the training set. (B) ROC curve for the *PLOD2* gene in the training set. (C) ROC curve for the *S100A8* gene in the validation set. (D) ROC curve for the *PLOD2* gene in the validation set. Note: AUC represents the area under the ROC curve.

3.6 Construction and validation of the flowchart

To predict the probability of bone metastasis in prostate cancer, a nomogram model was constructed using the R package “rms” in the training cohort. The expression levels of the two biomarkers were integrated, and the occurrence of BMPCa was defined as the outcome event for all samples in the training set. In the figure, Pr represents the probability of bone metastasis in each sample (Figure 6A). ROC curves were plotted based on the training cohort, and the predictive performance of the nomogram model for BMPCa samples was evaluated using the AUC. As shown, the AUC was 0.89, indicating that the nomogram exhibited high diagnostic value (Figure 6B). Furthermore, calibration curves were generated using the R package “rms” to evaluate the predictive performance of the nomogram model. The Hosmer–Lemeshow test showed $p > 0.05$, indicating good agreement between predicted and observed outcomes. In addition, the mean absolute error ($MAE = 0.064 < 0.1$) demonstrated high accuracy of the nomogram in predicting BMPCa (Figure 6C). In this study, Decision Curve Analysis (DCA) was performed in the training cohort based on the biomarkers. As shown in the

figure, the DCA curve demonstrated that the nomogram provided favorable net clinical benefit (Figure 6D).

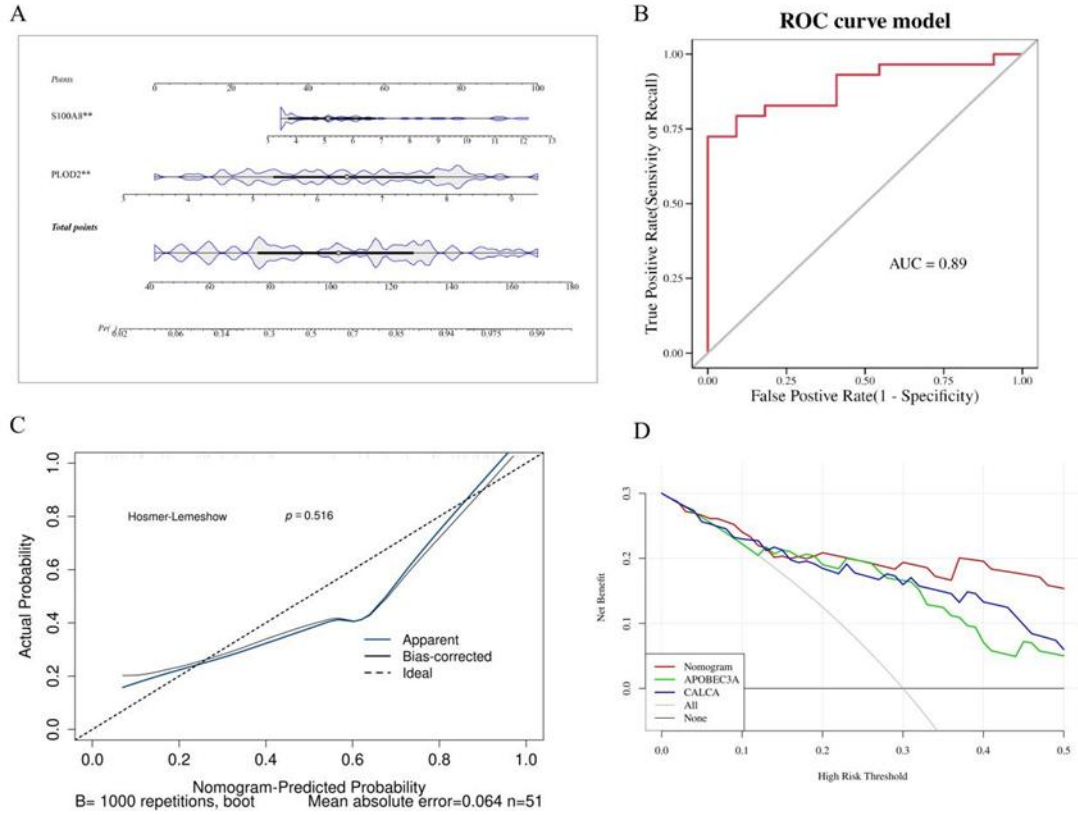


Figure 6. Construction and validation of the nomogram. (A) The biomarker-based nomogram. (B) ROC curves. (C) Calibration curves. (D) DCA curves.

3.7 GSEA analysis of biomarkers

To further investigate the signaling pathways and underlying biological mechanisms associated with each biomarker, Gene Set Enrichment Analysis (GSEA) was performed in the training set. Among these, the *S100A8* and *PLOD2* genes were significantly enriched in 65 and 27 pathways, respectively. *PLOD2* gene top 5 signaling pathways: Ribosome, Ecm Receptor Interaction, Cell Cycle, Dna Replication, Focal Adhesion (Figure 7A). *S100A8* gene top 5 signaling pathways: Ribosome, Oxidative Phosphorylation, Parkinsons Disease, Natural Killer Cell Mediated Cytotoxicity, Cytokine Cytokine Receptor Interaction (Figure 7B).

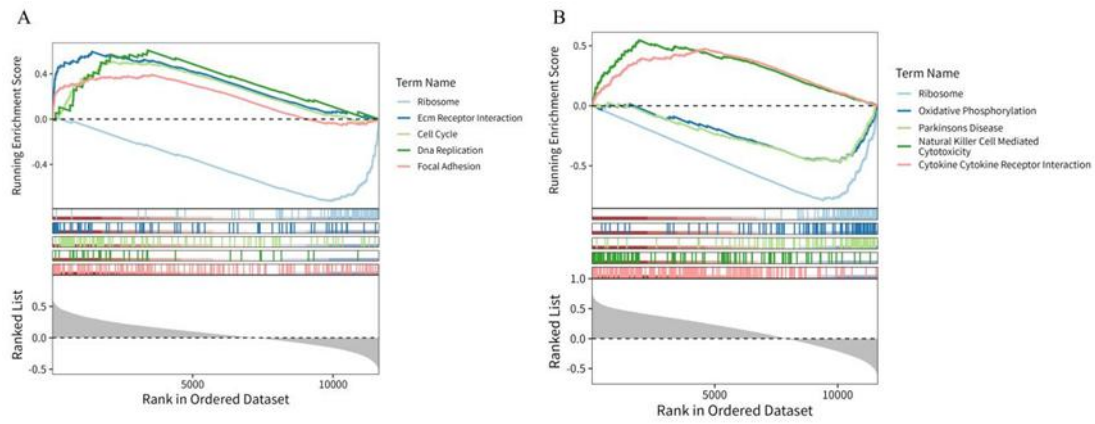


Figure 7. GSEA analysis. (A) GSEA of the *PLOD2* gene; (B) GSEA of the *S100A8* gene.

3.8 Immune infiltration analysis

To perform immune infiltration analysis, we utilized the single-sample Gene Set Enrichment Analysis (ssGSEA) algorithm to calculate the infiltration levels of 28 distinct immune cell types within the training set. A Wilcoxon test was performed to compare immune cell infiltration between the BMPCa and PCa groups in the training cohort, identifying 14 immune cell types with significant differences. The results were subsequently visualized using heatmaps and boxplots generated via the 'ggplot2' R package (Figure 8A-B). To further explore the relationships among the differentially infiltrated immune cells, Spearman correlation analysis was conducted ($|\text{cor}| > 0.3$, $p < 0.05$). A heatmap was then generated using the 'corrplot' R package to visualize the correlations among these 14 immune cell types. Additionally, Spearman correlation analysis was performed between the biomarkers and the differentially infiltrated immune cells. The strongest positive correlation was observed between MDSCs and the *S100A8* gene ($r = 0.803$), indicating a highly significant association (Figure 8C-D).

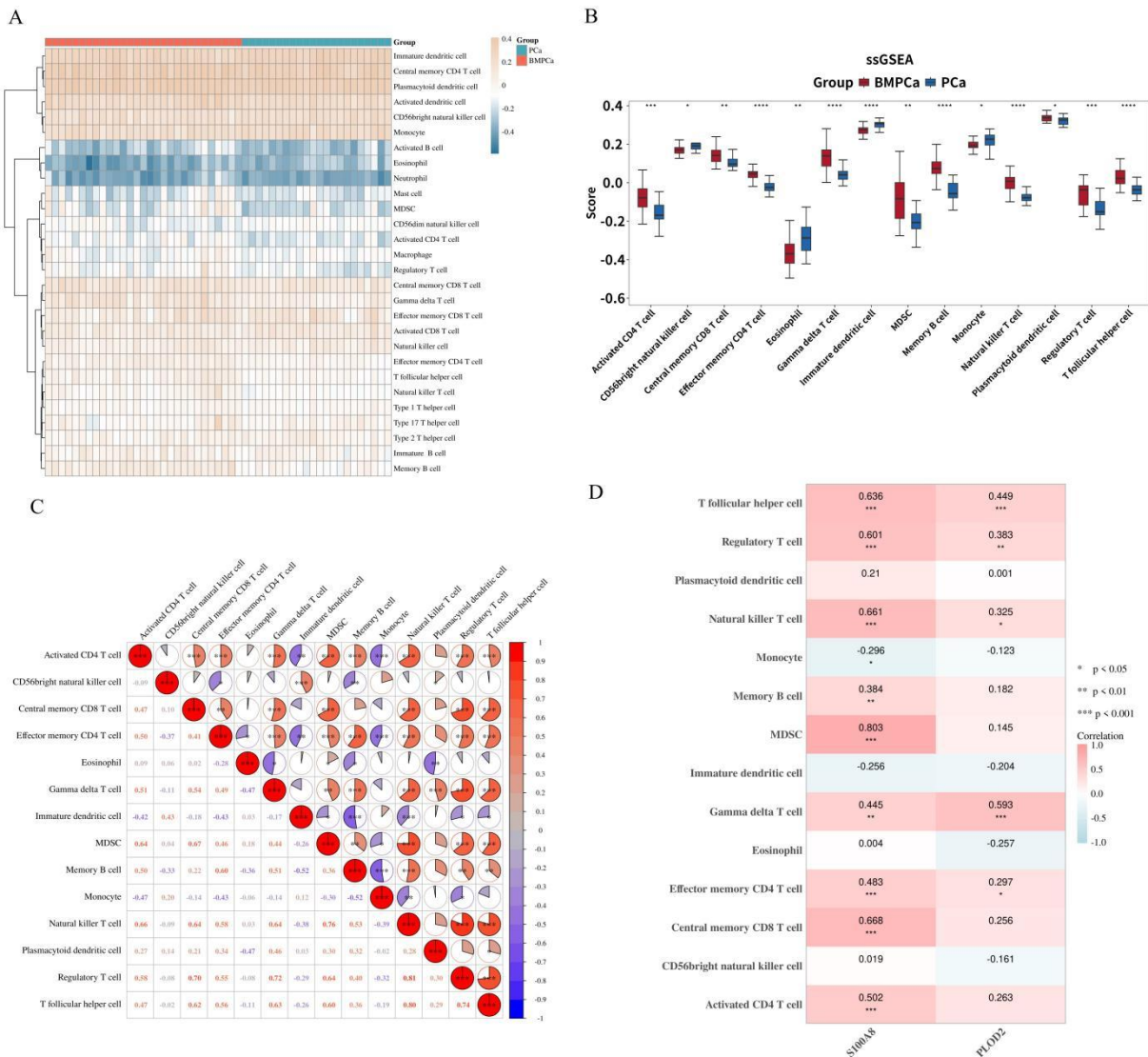


Figure 8. Immune cell infiltration analysis. (A) Immune cell infiltration levels between different groups. (B) Differences in immune cell infiltration between the groups. (C) Correlation analysis among

the differentially infiltrated immune cells. (D) Correlation analysis between the biomarkers and differentially infiltrated immune cells. Note: (C) This panel displays the correlations among the differentially infiltrated immune cells. The color scale from red to blue indicates positive to negative correlations, respectively. The size of the red area in the pie charts represents the strength of the positive correlation. *, $P < 0.05$; **, $P < 0.01$; ***, $P < 0.001$; ****, $P < 0.0001$.

4. Discussion

Bone metastasis frequently occurs in advanced PCa and is a major contributor to poor prognosis and high mortality among affected patients[4]. Therefore, identifying highly specific molecular biomarkers for bone metastasis remains a major challenge for clinical intervention. EMT confers invasive and metastatic phenotypes on tumor cells, while reprogramming of NM supplies the necessary energy and biosynthetic substrates for sustained proliferation and metastatic colonization of cancer cells. Although previous studies have investigated the role of individual mechanisms in PCa progression, the combined diagnostic value of EMT and NM-related genes in BMPCa, as well as their associated immune microenvironment characteristics, remains insufficiently elucidated. In this study, we integrated these two gene sets and applied multiple machine learning algorithms to precisely identify two differentially expressed core biomarkers (*S100A8* and *PLOD2*). We further investigated their potential roles in clinical prediction, immune infiltration, and epigenetic regulatory networks.

After LASSO, Boruta, and random forest (RF) analyses, *S100A8* and *PLOD2* were identified as the most predictive biomarkers and were significantly upregulated in BMPCa samples. *S100A8* is an important calcium-binding protein and pro-inflammatory factor[20; 21]. Previous study has shown that *S100A8* can induce EMT in tumor cells by activating inflammatory signaling cascades, thereby promoting an invasive phenotype[22]. In this study, GSEA analysis revealed that *S100A8* was significantly enriched in the ribosome, oxidative phosphorylation, Parkinson's disease, natural killer cell-mediated cytotoxicity, and cytokine–cytokine receptor interaction pathways. These enrichment results are highly consistent with recent reports on the mechanisms of tumor metastasis. Aberrant activation of ribosome biogenesis has been confirmed as a key factor driving tumor metastasis, promoting adaptive translation in cancer cells, and contributing to therapeutic resistance[23]. Meanwhile, oxidative phosphorylation (OXPHOS)-driven metabolic reprogramming provides essential energy support for tumor cell invasion, including EMT, as well as for colonization within the bone microenvironment[24; 25]. At the level of the immune microenvironment, effective activity of natural killer (NK) cells plays a decisive role in controlling prostate cancer metastasis and improving patient prognosis[26; 27]. *S100A8* is a key target involved in the formation of the pre-metastatic niche[28]. These findings strongly suggest that high expression of *S100A8* may promote the progression of BMPCa through multiple mechanisms, including enhancement of ribosomal translation, reprogramming of oxidative phosphorylation metabolism in tumor cells, and coordinated regulation of NK cell-related immune evasion, thereby collectively driving disease progression in a multidimensional manner.

Procollagen-lysine, 2-oxoglutarate 5-dioxygenase 2 (*PLOD2*) is a catalytic enzyme primarily responsible for hydroxylating lysine residues in the telopeptide regions of procollagen[29]. This process is essential for stable collagen cross-linking and extracellular matrix (ECM) remodeling. In recent years, it has emerged as a highly studied target in the regulation of the tumor microenvironment and in tumor cell invasion and metastasis[30; 31]. GSEA results showed that *PLOD2* was primarily associated with ribosome, ECM–receptor interaction, cell cycle, DNA replication, and focal adhesion pathways. These pathways are known to exhibit profound synergistic effects in tumor progression. For example, the ECM–receptor interaction and focal adhesion pathways serve as central hubs for cellular

sensing and response to mechanical signals from the microenvironment[32]. Focal adhesions act as cellular mechanosensors, mediating bidirectional mechanical and biochemical signaling between cells and the ECM through molecules such as integrins[32; 33]. This interaction can activate kinase cascades such as FAK and Src, thereby regulating cytoskeletal dynamic remodeling as well as ECM degradation and redeposition[32]. Meanwhile, ECM stiffness-mediated alterations in mechanical signaling can further activate downstream transcriptional pathways, thereby directly interfering with cell cycle regulation and DNA replication, and enhancing tumor cell proliferative capacity and invasive plasticity[34]. Therefore, targeting *PLOD2* and its associated cell adhesion and mechanotransduction pathways may represent a potential therapeutic strategy for inhibiting tumor invasion and distant dissemination[33; 35].

In terms of clinical translational value, the nomogram constructed based on *S100A8* and *PLOD2* demonstrated favorable diagnostic performance. ROC curve analysis in the training cohort showed high accuracy (AUC = 0.89), while calibration curves (MAE = 0.064) and DCA further confirmed the model's strong predictive concordance and significant clinical net benefit. This suggests that the two-gene-based predictive signature may serve as a valuable complement to conventional pathological diagnosis, facilitating early screening and intervention for patients at high risk of bone metastasis.

The formation of the bone metastatic microenvironment is accompanied by complex immune cell remodeling. MDSCs are key immunosuppressive cells within the tumor microenvironment[36]. MDSC-induced PD-1–PD-L1+ B-cell subset in breast tumor microenvironment possesses immunosuppressive properties[37]. Research has shown that MDSC facilitate distant metastasis of malignancies by shielding circulating tumor cells (CTC) from immune surveillance[38]. Combining TIGIT blockade with MDSC inhibition hinders breast cancer bone metastasis by activating antitumor immunity[39]. The immune infiltration analysis conducted in this study identified 14 significantly different immune cell populations between the PCa and BMPCa groups. Among these, the infiltration level of myeloid-derived suppressor cells (MDSCs) exhibited a strong positive correlation with the expression of *S100A8* (cor = 0.803). This finding suggests that elevated *S100A8* expression may act as a chemotactic signal to recruit large numbers of myeloid-derived suppressor cells (MDSCs) to bone metastatic lesions, thereby suppressing T-cell function and facilitating immune evasion by tumor cells.

Although this study provides promising combined biomarkers for the early diagnosis of BMPCa, several limitations remain from the perspective of clinical translation. First, the sample size included in this study was relatively limited, and potential population heterogeneity may have existed. The generalizability and diagnostic performance of the nomogram model in real-world clinical cohorts still require rigorous external validation through multicenter, large-scale prospective studies. Second, due to the lack of long-term follow-up data, this study was unable to evaluate the independent prognostic value of *S100A8* and *PLOD2* in predicting long-term clinical outcomes, such as progression-free survival and overall survival.

5. Conclusion

By integrating bioinformatics analyses with multiple machine learning algorithms, this study identified and validated *S100A8* and *PLOD2* as core diagnostic biomarkers for BMPCa. The nomogram prediction model constructed based on EMT- and NM-related genes demonstrated excellent clinical diagnostic performance and substantial clinical net benefit. Furthermore, immune infiltration analysis revealed a strong positive correlation between *S100A8* expression levels and MDSC infiltration, suggesting that *S100A8* may play a critical role in driving the remodeling of the immune

421 microenvironment during bone metastasis. This study provides new insights and reliable targets for the
422 early screening and precise targeted intervention of BMPCa.

423 **Author Contributions**

424 Ke Chen and Jiayang Yue contributed to manuscript writing, review, and editing, and were
425 responsible for figure preparation. Zhaogui Ding contributed to public database retrieval and data
426 acquisition. Min Cheng performed the data analysis. Ning Yao and Tiansheng Ye contributed to the
427 data analysis. Jia Sha and Zhanqiang Hua contributed to study conception and supervision, reviewed
428 the manuscript, and provided technical support.

429 **Funding**

430 This research received no external funding.

431 **Data Availability Statement**

432 The datasets analyzed in this study were obtained from the Gene Expression Omnibus (GEO)
433 database. The training dataset (GSE32269) is available at
434 <https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE32269> and the validation dataset
435 (GSE77930) is available at <https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE77930>. The
436 epithelial-mesenchymal transition (EMT)-related genes and nucleotide metabolism (NM)-related
437 genes were obtained from the GeneCards database (<https://www.genecards.org/>) using the keywords
438 “epithelial-mesenchymal transition” and “nucleotide metabolism”, and the corresponding gene lists are
439 provided in Supplementary Table. The gene sets used for Gene Set Enrichment Analysis (GSEA) were
440 obtained from the Molecular Signatures Database (MSigDB) using the R package msigdb, including
441 C2: Curated gene sets (KEGG pathways) (<https://www.gsea-msigdb.org/gsea/msigdb/>). No new raw
442 data were generated in this study. The processed data and analysis codes are available from the
443 corresponding author upon reasonable request.

444 **Acknowledgments**

445 We thank the GEO database for providing publicly available transcriptomic datasets. We also
446 acknowledge the developers and contributors of the R packages used in this study for making their
447 software and source code publicly available.

448 **Ethics declarations**

449 **Ethics approval and consent to participate**

450 Not applicable

451 **Consent for publication**

452 Not applicable

453 **Competing interests**

454 The authors declare no competing interests.

455 **Reference**

- [1] L. Filippi, and A. Chiaravalloti, Prostate Cancer: From Molecular Imaging to Immunological and Target Therapies. *Biomedicines* 11 (2023).
- [2] L.N. Garrido Castillo, A. Mejean, P. Vielh, J. Anract, A. Decina, B. Nalpas, N. Benali-Furet, I. Desitter, and P. Paterlini-Bréchet, Predictive Value of Circulating Tumor Cells Detected by ISET(®) in Patients with Non-Metastatic Prostate Cancer Undergoing Radical Prostatectomy. *Life (Basel)* 12 (2022).
- [3] R.L. Siegel, K.D. Miller, N.S. Wagle, and A. Jemal, Cancer statistics, 2023. *CA: a cancer journal for clinicians* 73 (2023) 17-48.
- [4] R.B. Berish, A.N. Ali, P.G. Telmer, J.A. Ronald, and H.S. Leong, Translational models of prostate cancer bone metastasis. *Nature Reviews Urology* 15 (2018) 403-421.
- [5] C. Xia, X. Dong, H. Li, M. Cao, D. Sun, S. He, F. Yang, X. Yan, S. Zhang, and N. Li, Cancer statistics in China and United States, 2022: profiles, trends, and determinants. *Chinese medical journal* 135 (2022) 584-590.
- [6] S. Lee, T.R. Mendoza, D.N. Burner, M.T. Muldong, C.C. Wu, C. Arreola-Villanueva, A. Zuniga, O. Greenburg, W.Y. Zhu, and J. Murtadha, Novel dormancy mechanism of castration resistance in bone metastatic prostate cancer organoids. *International journal of molecular sciences* 23 (2022) 3203.
- [7] H. Allgayer, S. Mahapatra, B. Mishra, B. Swain, S. Saha, S. Khanra, K. Kumari, V.K. Panda, D. Malhotra, and N.S. Patil, Epithelial-to-mesenchymal transition (EMT) and cancer metastasis: the status quo of methods and experimental models 2025. *Molecular cancer* 24 (2025) 167.
- [8] A. Dong, and C. Blanpain, Identification, functional insights and therapeutic targeting of EMT tumour states. *Nature Reviews Cancer* 26 (2026) 8-26.
- [9] H. Vosgha, A. Ariana, R.A. Smith, and A.K.-y. Lam, miR-205 targets angiogenesis and EMT concurrently in anaplastic thyroid carcinoma. *Endocrine-related cancer* 25 (2018) 323-337.
- [10] V. Odero-Marrah, O. Hawsawi, V. Henderson, and J. Sweeney, Epithelial-mesenchymal transition (EMT) and prostate cancer. *Cell & molecular biology of prostate cancer: updates, insights and new frontiers* (2018) 101-110.
- [11] Q. Zhou, M. Yang, J. Fu, X. Sun, J. Wang, H. Zhang, J. Hu, and B. Han, KIF1A promotes neuroendocrine differentiation in prostate cancer by regulating the OGT-mediated O-GlcNAcylation. *Cell death & disease* 15 (2024) 796.
- [12] S. Ge, J. Cen, X. Liu, Y. Hong, Y. Tang, Y. Yu, H. Li, T. Xie, C. Wang, M. Cai, Y. Qiu, X. Zeng, T. Peng, Q. Li, Q. Li, X. Wu, X.L. Song, and S.C. Zhao, TGF β -activated Asporin interacts with STMN1 to promote prostate cancer docetaxel chemoresistance and metastasis by upregulating the Wnt/ β -catenin signaling pathway. *Drug Resist Updat* 81 (2025) 101227.
- [13] B. Zhang, Y. Li, Q. Wu, L. Xie, B. Barwick, C. Fu, X. Li, D. Wu, S. Xia, and J. Chen, Acetylation of KLF5 maintains EMT and tumorigenicity to cause chemoresistant bone metastasis in prostate cancer. *Nature communications* 12 (2021) 1714.
- [14] X. Feng, M. Tang, M. Dede, D. Su, G. Pei, D. Jiang, C. Wang, Z. Chen, M. Li, L. Nie, Y. Xiong, S. Li, J.M. Park, H. Zhang, M. Huang, K. Szymonowicz, Z. Zhao, T. Hart, and J. Chen, Genome-wide CRISPR screens using isogenic cells reveal vulnerabilities conferred by loss of tumor suppressors. *Sci Adv* 8 (2022) eabm6638.
- [15] M. Cai, H. Liu, C. Shao, T. Li, J. Jin, Y. Liang, J. Wang, J. Cao, B. Yang, Q. He, X. Shao, and M. Ying, Metabolomics and metabolites in cancer diagnosis and treatment. *Mol Biomed* 6 (2025) 109.
- [16] X. Lei, Y. Lei, J.K. Li, W.X. Du, R.G. Li, J. Yang, J. Li, F. Li, and H.B. Tan, Immune cells within the tumor microenvironment: Biological functions and roles in cancer immunotherapy. *Cancer Lett* 470 (2020) 126-133.
- [17] L. Xiang, and D.M. Gilkes, The Contribution of the Immune System in Bone Metastasis Pathogenesis. *International journal of molecular sciences* 20 (2019).

- [18] C. Yin, M. Wang, Y. Wang, Q. Lin, K. Lin, H. Du, C. Lang, Y. Dai, and X. Peng, BHLHE22 drives the immunosuppressive bone tumor microenvironment and associated bone metastasis in prostate cancer. *J Immunother Cancer* 11 (2023).
- [19] S. Chen, J. Lei, H. Mou, W. Zhang, L. Jin, S. Lu, E. Yinwang, Y. Xue, Z. Shao, T. Chen, F. Wang, S. Zhao, X. Chai, Z. Wang, J. Zhang, Z. Zhang, Z. Ye, and B. Li, Multiple influence of immune cells in the bone metastatic cancer microenvironment on tumors. *Frontiers in immunology* 15 (2024) 1335366.
- [20] X. Zhong, F. Xie, L. Chen, Z. Liu, and Q. Wang, S100A8 and S100A9 promote endothelial cell activation through the RAGE-mediated mammalian target of rapamycin complex 2 pathway. *Mol Med Rep* 22 (2020) 5293-5303.
- [21] Z. Li, J. Wang, X. Zhang, P. Liu, X. Zhang, J. Wang, X. Zheng, L. Wei, Q. Peng, C. Liu, Q. Yan, S. Shen, X. Li, and J. Ma, Proinflammatory S100A8 Induces PD-L1 Expression in Macrophages, Mediating Tumor Immune Escape. *Journal of immunology (Baltimore, Md. : 1950)* 204 (2020) 2589-2599.
- [22] Z. Xu, C. Cheng, R. Kong, Y. Liu, S. Wang, Y. Ma, and X. Xing, S100A8 and S100A9, both transcriptionally regulated by PU.1, promote epithelial-mesenchymal transformation (EMT) and invasive growth of dermal keratinocytes during scar formation post burn. *Aging* 13 (2021) 15523-15537.
- [23] S.P. Hwang, and C. Denicourt, The impact of ribosome biogenesis in cancer: from proliferation to metastasis. *NAR Cancer* 6 (2024) zcae017.
- [24] Y. Hu, W. Xu, H. Zeng, Z. He, X. Lu, D. Zuo, G. Qin, and W. Chen, OXPHOS-dependent metabolic reprogramming prompts metastatic potential of breast cancer cells under osteogenic differentiation. *Br J Cancer* 123 (2020) 1644-1655.
- [25] Y. Huang, Y. Du, Y. Zheng, C. Wen, H. Zou, J. Huang, H. Zhou, H. Zhao, and L. Wu, Ct-OATP1B3 promotes high-grade serous ovarian cancer metastasis by regulation of fatty acid beta-oxidation and oxidative phosphorylation. *Cell death & disease* 13 (2022) 556.
- [26] C. Pasero, G. Gravis, S. Granjeaud, M. Guerin, J. Thomassin-Piana, P. Rocchi, N. Salem, J. Walz, A. Moretta, and D. Olive, Highly effective NK cells are associated with good prognosis in patients with metastatic prostate cancer. *Oncotarget* 6 (2015) 14360-73.
- [27] A.H. Zedan, L. Nederby, L.M. Volmer, C.V. Madsen, B.E. Sørensen, and T.F. Hansen, Natural killer cell activity in metastatic castration resistant prostate cancer patients treated with enzalutamide. *Scientific reports* 13 (2023) 17144.
- [28] A. Deguchi, and Y. Maru, S100A8 as a potential therapeutic target for cancer metastasis. *Cancer Sci* 116 (2025) 322-328.
- [29] H. Du, M. Pang, X. Hou, S. Yuan, and L. Sun, PLOD2 in cancer research. *Biomedicine & pharmacotherapy = Biomedecine & pharmacotherapie* 90 (2017) 670-676.
- [30] K. Li, Y. Niu, K. Li, C. Zhong, Z. Qiu, Y. Yuan, Y. Shi, Z. Lin, Z. Huang, D. Zuo, Y. Yuan, and B. Li, Dysregulation of PLOD2 Promotes Tumor Metastasis and Invasion in Hepatocellular Carcinoma. *J Clin Transl Hepatol* 11 (2023) 1094-1105.
- [31] Z. Lin, F. Wang, R. Yin, S. Li, Y. Bai, B. Zhang, C. Sui, H. Cao, D. Su, L. Xu, and H. Wang, Single-cell RNA sequencing and immune microenvironment analysis reveal PLOD2-driven malignant transformation in cervical cancer. *Frontiers in immunology* 15 (2024) 1522655.
- [32] Z. Liu, X. Zhang, T. Ben, M. Li, Y. Jin, T. Wang, and Y. Song, Focal adhesion in the tumour metastasis: from molecular mechanisms to therapeutic targets. *Biomark Res* 13 (2025) 38.
- [33] D. Burčík, J. Macko, N. Podrojková, J. Demeterová, M. Stano, and A. Oriňak, Role of Cell Adhesion in Cancer Metastasis Formation: A Review. *ACS Omega* 10 (2025) 5193-5213.
- [34] R. Chen, R. Zhang, F. Ke, X. Guo, F. Zeng, and Q. Liu, Mechanisms of breast cancer metastasis: the role of extracellular matrix. *Mol Cell Biochem* 480 (2025) 2771-2796.

- [35] G. Frame, H. Leong, R. Haas, X. Huang, J. Wright, U. Emmenegger, M. Downes, P.C. Boutros, T. Kislinger, and S.K. Liu, Targeting PLOD2 suppresses invasion and metastatic potential in radiorecurrent prostate cancer. *BJC Rep* 2 (2024) 60.
- [36] S. He, L. Zheng, and C. Qi, Myeloid-derived suppressor cells (MDSCs) in the tumor microenvironment and their targeting in cancer therapy. *Molecular cancer* 24 (2025) 5.
- [37] M. Shen, J. Wang, W. Yu, C. Zhang, M. Liu, K. Wang, L. Yang, F. Wei, S.E. Wang, Q. Sun, and X. Ren, A novel MDSC-induced PD-1(-)PD-L1(+) B-cell subset in breast tumor microenvironment possesses immuno-suppressive properties. *Oncoimmunology* 7 (2018) e1413520.
- [38] Q. Liu, Q. Liao, and Y. Zhao, Myeloid-derived suppressor cells (MDSC) facilitate distant metastasis of malignancies by shielding circulating tumor cells (CTC) from immune surveillance. *Med Hypotheses* 87 (2016) 34-9.
- [39] L. Monteran, N. Ershaid, Y. Scharff, Y. Zoabi, T. Sanalla, Y. Ding, A. Pavlovsky, Y. Zait, M. Langer, T. Caller, A. Eldar-Boock, C. Avivi, A. Sonnenblick, R. Satchi-Fainaro, I. Barshack, N. Shomron, X.H. Zhang, and N. Erez, Combining TIGIT Blockade with MDSC Inhibition Hinders Breast Cancer Bone Metastasis by Activating Antitumor Immunity. *Cancer Discov* 14 (2024) 1252-1275.

Supplementary Files

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

- [SupplementaryTable.xlsx](#)