Original Research

Analysis of genomics and immune infiltration patterns of epithelial-mesenchymal transition related to metastatic breast cancer to bone

Shuzhong Liu [a,#], An Song [b,#], Yunxiao Wu [c], Siyuan Yao [a], Muchuan Wang [a], Tong Niu [a], Chengao Gao [a], Ziquan Li [a], Xi Zhou [a], Zhen Huo [d], Bo Yang [a,*], Yong Liu [a,*], Yipeng Wang [a,*]

[a] Department of Orthopaedic Surgery, Peking Union Medical College Hospital, Peking Union Medical College and Chinese Academy of Medical Sciences, Beijing, China
[b] Department of Endocrinology (AS), Key Laboratory of Endocrinology, National Health and Family Planning Commission, Peking Union Medical College Hospital, Chinese Academy of Medical Science & Peking Union Medical College, Beijing, China
[c] School of Public Health (YW), Peking University Health Science Center, 38 Xueyuan Road, Beijing 100191, China
[d] Department of Pathology (ZH), Peking Union Medical College Hospital, Chinese Academy of Medical Science & Peking Union Medical College, Beijing, China

A R T I C L E   I N F O

A B S T R A C T

Objective: This study aimed to design a weighted co-expression network and a breast cancer (BC) prognosis evaluation system using a specific whole-genome expression profile combined with epithelial-mesenchymal transition (EMT)-related genes; thus, providing the basis and reference for assessing the prognosis risk of spreading of metastatic breast cancer (MBC) to the bone.

Methods: Four gene expression datasets of a large number of samples from GEO were downloaded and combined with the dbEMT database to screen out EMT differentially expressed genes (DEGs). Using the GSE20685 dataset as a training set, we designed a weighted co-expression network for EMT DEGs, and the hub genes most relevant to metastasis were selected. We chose eight hub genes to build prognostic assessment models to estimate the 3-, 5-, and 10-year survival rates. We evaluated the models' independent predictive abilities using univariable and multivariable Cox regression analyses. Two GEO datasets related to bone metastases from BC were downloaded and used to perform differential genetic analysis. We used CIBERSORT to distinguish 22 immune cell types based on tumor transcripts.

Results: Differential expression analysis showed a total of 304 DEGs, which were mainly related to proteoglycans in cancer, and the PI3K/Akt and the TGF-β signaling pathways, as well as mesenchyme development, focal adhesion, and cytokine binding functionally. The 50 hub genes were selected, and a survival-related linear risk assessment model consisting of eight genes (*FERMT2, ITGA5, ITGB1, MCAM, CEMIP, HGF, TGFBR1, F2RL2*) was constructed. The survival rate of patients in the high-risk group (HRG) was substantially lower than that of the low-risk group (LRG), and the 3-, 5-, and 10-year AUCs were 0.68, 0.687, and 0.672, respectively. In addition, we explored the DEGs of BC bone metastasis, and *BMP2, BMPR2*, and *GREM1* were differentially expressed in both data sets. In GSE20685, memory B cells, resting memory T cell CD4 cells, T regulatory cells (T$_{regs}$), $\gamma\delta$ T cells, monocytes, M0 macrophages, M2 macrophages, resting dendritic cells (DCs), resting mast cells, and neutrophils exhibited substantially different distribution between HRG and LRG. In GSE45255, there was a considerable difference in abundance of activated NK cells, monocytes, M0 macrophages, M2 macrophages, resting DCs, and neutrophils in HRG and LRG.

Conclusions: Based on the weighted co-expression network for breast-cancer-metastasis-related DEGs, we screened hub genes to explore a prognostic model and the immune infiltration patterns of MBC. The results of this study provided a factual basis to bioinformatically explore the molecular mechanisms of the spread of MBC to the bone and the possibility of predicting the survival of patients.

Fig. 1. flowchart of this study.

Introduction

Breast cancer (BC) constitutes the leading cause of cancer-associated fatalities in women worldwide [1–3]. Metastatic breast cancer (MBC) occurs in most patients with advanced BC, with bone as the most common site of distant metastasis [1–3]. Bone destruction often leads to bone-related complications, including pain, spinal cord compression, fractures, severe hypercalcemia, etc., which negatively impact the quality of life of the patient [1–3].

BC metastasis to bones is an extremely complex multi-stage process that requires the primary BC cells to pass through the blood/lymphatic system, survive in the bone microenvironment, and then proliferate in bone tissue [4,5]. Genomics studies have revealed that several molecular events are associated with each step of the metastasis [6,7]. However, the key pathways and interaction networks of the molecular mechanisms related to BC metastasis have not been fully elucidated.

On this basis, we constructed a weighted co-expression network and BC prognosis evaluation model using the whole-gene expression profile combined with epithelial-mesenchymal transition (EMT)-related genes to provide a basis and reference for the prognostic risk of MBC. We aimed to establish a complete protein-interaction network to reveal the molecular mechanisms of early BC metastasis. This study attempted to further explore the molecular biological mechanisms in the early stages of BC metastasis. In addition, we evaluated the interconnection between immune cells and hub genes to create a bioinformatic base for discovering possible molecular pathways and clinical predictors. We have drawn a flowchart as a guide to help readers understand the analytical process of the study (Fig. 1).

Materials and methods

GEO data download and preprocessing

We downloaded GSE20685, GSE12276, GSE16446, GSE45255, GSE2034, and GSE124647 datasets of BC and BC metastasis chip data from the GEO database. Six samples without metastasis information were removed from each of GSE16446, and GSE45255, and GSE124647, and only primary BC samples and samples with bone metastasis were retained.

The data preprocessing steps were as follows: if the data set was not log2 converted, then log2 conversion was performed; if data not quantile-normalized, then the limma package normalizeBetweenArrays method of R was used to quantile-normalize. The probe was mapped to the gene, the empty probe removed, and multiple probes relative to the same gene. We calculated the average value of gene expression. The Wilcox rank-sum test was used for difference analysis, with a P-value < 0.05 as the filtering condition.

Candidate gene collection and data expression

The differentially expressed genes (DEGs) between primary and invasive samples obtained by searching GSE20685, GSE12276, and GSE16446 and the EMT-related genes indicated by the dbEMT database were jointly investigated, and the intersection of the DEGs between primary and invasive samples of the three datasets from the GEO database and the EMT database was used [8].

Functional enrichment analysis of candidate genes

Using R's clusterProfiler software package, KEGG pathway and GO functional enrichment analyses were performed on the EMT-related DEGs between primary and invasive samples, and a p-value < 0.05 and q-value of < 0.2 were used as the filtering conditions.

Weighted co-expression network analysis

Weighted co-expression network analysis (WGCNA) is a systematic method that uses gene expression data to build a scale-free network. We used the WGCNA package of R to build a weighted co-expression network with the expression profile data of the candidate gene set obtained. We then screened the modules related to BC metastasis, extracted the genes, and selected the top 50 as hub genes based on their degree of inter-gene connectivity.

Risk prognosis model construction and model effectiveness evaluation

The GSE20685 dataset was used as the training set, using univariable Cox regression and a *p*-value of < 0.01 as the filtering condition, to select genes related to prognosis. We then used the LASSO–Cox regression model to screen out prognosis-related DEGs between primary and metastatic samples and obtain their correlation coefficients. A risk prognosis model was built based on these genes and coefficients. Based on the model, the risk score of each patient was estimated, the median score was taken as the cut-off value, and the sample was divided into a high-risk group (HRG) and low-risk group (LRG). A time-dependent ROC curve was used to predict the 3-, 5- and 10-year survival rates and the survival curves of the HRG and LRG were analyzed. We used the GSE45255 dataset to verify the prognostic risk model. Subsequently, the correlations between risk scores and other clinical prognostic factors were compared, and univariable and multivariable Cox regression analyses were used to evaluate the independent predictive ability of the prognostic model.

Immune cell abundance analysis

CIBERSORT software (http://CIBERSORT.stanford.edu/) was used to predict the proportion of 22 immune cells in all samples of the datasets. We used R's CIBERSORT package to assess the abundance of the 22 immune cells in HRG and LRG. We sorted the proportions of certain cell types for all samples to explore the clinical significance of different samples with different proportions of immune cells. The samples were divided into high-ratio and low-ratio groups for survival analysis using the median as the dividing line. Subsequently, Kaplan-Meier analysis was done on only those cases that had a CIBERSORT $p < 0.05$.

Results

Screening of DEGs

We downloaded three sets of BC chip data, GSE20685, GSE12276, and GSE16446, from GEO, which had 244 vs. 83, 19 vs. 185, and 83 vs. 24 primary cancer vs metastasis samples, respectively. The Wilcox rank-sum test was used for differential analysis. Taking a *P*-value of < 0.05 as the threshold, we obtained 2950 DEGs in GSE20685, including 1429 upregulated genes and 1521 downregulated genes; 1209 DEGs in GSE12276, including 719 upregulated genes and 490 downregulated genes; and 830 DEGs in GSE16446, including 544 upregulated genes and 286 downregulated genes (Fig. 2).

EMT-related DEG set

The DEGs and the EMT-related genes obtained from the EMT database were jointly investigated, and the intersections of the differential genes obtained from the three sets of GEO data and the EMT database were analyzed to obtain 304 genes (Fig. 2). We took the expression data of these 304 genes in GSE20685 and used them for the WGCNA.

Functional enrichment analysis of EMT-related DEGs (KEGG pathway and GO analysis)

Using R's clusterProfiler software package, KEGG pathway and GO functional enrichment analyses were performed on the EMT-related DEGs (Fig. 2). These genes were mainly found to be related to proteoglycans in cancer, the PI3K/Akt and TGF-β signaling pathways, as well as mesenchyme development, focal adhesion, and cytokine binding.

Construction of protein interaction network of differential genes

The DEGs from three datasets (GSE20685, GSE12276, and GSE16446) were uploaded to the string protein database (https://string-db.org/), which was used for protein interaction analysis. The species was set as Homo sapiens, and the minimum interaction threshold was set as 0.4 (the medium confidence), and other parameters remained at the default settings. Among them, 288 genes interact with each other, and a total of 3121 edges and 288 nodes are obtained. The top 10 genes were *GAPDH, VEGFA, FN1, CDH1, STAT3, Notch1, CD44, ERBB2, ESR1, ITGB1*. See Supplementary file 1 and 2 for more details.

Candidate gene set weighted co-expression network construction and hub gene screening

We designed a weighted co-expression network for the candidate gene set using the WGCNA software package of R. Research has shown that the co-expression network follows a scale-free network, i.e., the log (k) of a node with a connection (k) is inversely related to the log of the probability of the node (P [k]), and the correlation coefficient > 0.85. To ensure that the network was scale-free, we chose the optimal $\beta = 6$ (Fig. 3A, "**β": power value, "point": a set of soft-thresholding power**). In the next step, the expression matrix was converted into an adjacency matrix; then, the adjacency matrix was converted into a topological matrix. Based on topological overlap measure (TOM), we used the average-linkage hierarchical clustering method to cluster genes according to a hybrid dynamic shear tree standard and set the minimum number of genes for each gene network module to 30. After using dynamic shearing to determine the gene modules, we calculated the eigengenes of each module once and then performed cluster analysis on the modules to merge the modules that were closer to the new module and a set height of 0.25. A total of two modules were obtained (Fig. 3C). We calculated Pearson's correlation coefficient of the ME (module eigengene) of each module and the sample characteristics (primary vs. metastatic) (the higher the module, the more important it was). The row in Fig. 3B represents the feature vector gene of each module, and the column represents the sample classification information. From red to blue, the correlation coefficients decrease from high to low. The numbers in each small box indicate the correlation coefficients of the gene modules and corresponding features (BC metastasis and primary BC), and the numbers in parentheses indicate the P-value. From Fig. 3B and Fig. 3C, we can conclude that there is one module marked with turquoise in the cluster dendrogram with a total of 72 genes most relevant to the metastasis. The top 50 genes were selected as hub genes according to the degree of connection (Supplementary file 3).

Prognostic factor screening and model construction

For the 50 hub genes obtained from the previous step, we first performed a univariable Cox regression analysis, selected 10 genes related to prognosis according to *P*-values < 0.01 (Fig. 4A), and then used the minimum absolute contraction and selection operator (Lasso)-Cox penalty regression model to select genes related to prognosis. We selected the Lambda.min as the critical point according to the parameter Lambda value (Fig. 4B) and, finally, selected a survival-related linear risk assessment model consisting of eight genes (Fig. 4C). The risk score model was Risk Score = (0.341711734325434*FERMT2) + (0.047360511272754*ITGA5) + (0.0274687909333705*ITGB1) + (0.193677760524873*MCAM) + (0.0779279290392332*CEMIP) + (−0.278104212976493*HGF) + (0.0732296389625441*TGF1) + (−0.123123692280744*F2RL2).

Prognostic risk model evaluation

Using the model to calculate the risk score of each patient in the GSE20685 dataset, taking the median of all patients' risk score as the cutoff value, the sample was divided into HRG and LRG. There were 163 samples in the HRG and 164 samples in the LRG. Using R's survival package for survival analysis, the survival curves (Fig. 5C) showed that
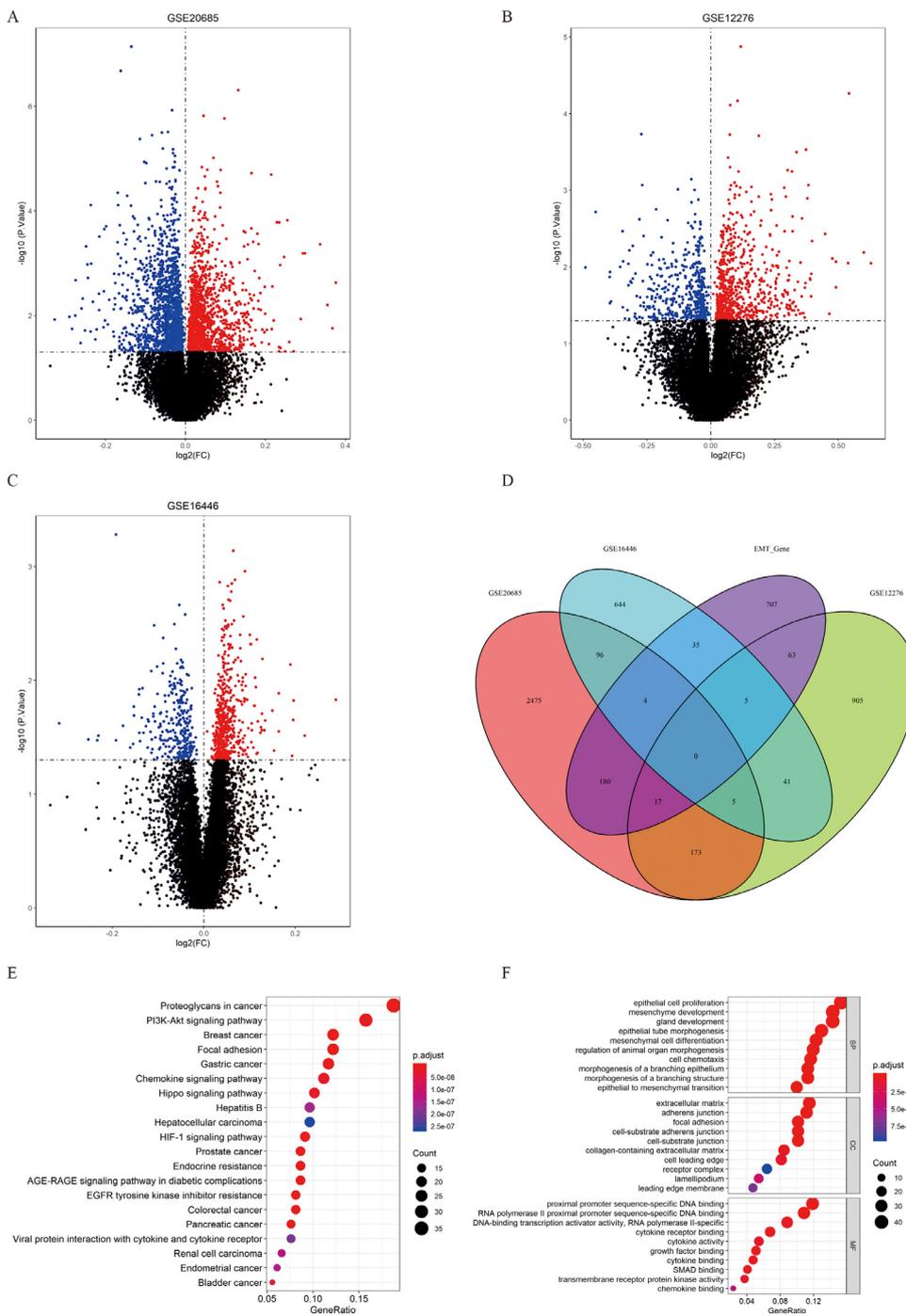
Fig. 2. (A-C) Differentially expressed genes (Volcano plot) of breast cancer in situ and breast cancer with metastasis from the analysis of three datasets in the GEO database (Red: upregulated; Blue: downregulated; Black: no significant change). **(D)** Venn diagram of differentially expressed genes and EMT-related genes from GEO. **(E)** KEGG signal pathway enrichment analysis diagram of differentially expressed EMT-related genes. **(F)** GO enrichment analysis diagram of differentially expressed EMT-related genes.

the overall survival rate of patients in the HRG was low, and the difference between the two groups was statistically significant, indicating that the model can predict survival well. At the same time, the ROC curve was drawn using the survivalROC package of R. The AUCs of the 3-, 5-, and 10-year survival rates were 0.68, 0.687, and 0.672, respectively, indicating that the model had good predictive ability (Fig. 5A). There were significant differences in survival time between the HRG and LRG, and there was no significant difference in age or TNM (Tumor, Node, Metastases) stage (Fig. 5E).

Model verification was performed using the GSE45255 dataset. The median risk score of all patients in the GSE20685 dataset was also used as the cut-off value, and the samples were divided into HRG and LRG. There were 52 samples in the HRG and 81 samples in the LRG. The survival curves of HRG and LRG (Fig. 5D) indicate that the overall survival

rate of patients in the HRG was low, and the difference between the two groups was statistically significant, indicating that the model can accurately distinguish between the patients in HRG and LRG. The AUCs of 3-, 5-, and 10-year survival rates were 0.675, 0.636, and 0.725, respectively, indicating that the model has good predictive ability (Fig. 5B).

Differences in immune cell infiltration between HRG and LRG

We used the CIBERSORT package of R (an analysis tool that uses gene expression data to assess the abundance of 22 immune cells in a sample) for HRG and LRG of the GSE20685 and GSE45255 datasets. In GSE20685, the abundance of memory B cells, resting memory CD4 T cells, T cells regulatory (T$_{regs}$), $\gamma\delta$ T cells, monocytes, M0 macrophages, M2 macrophages, resting DCs, resting mast cells, and neutrophils were

A

Scale independence

B

Module-trait relationships
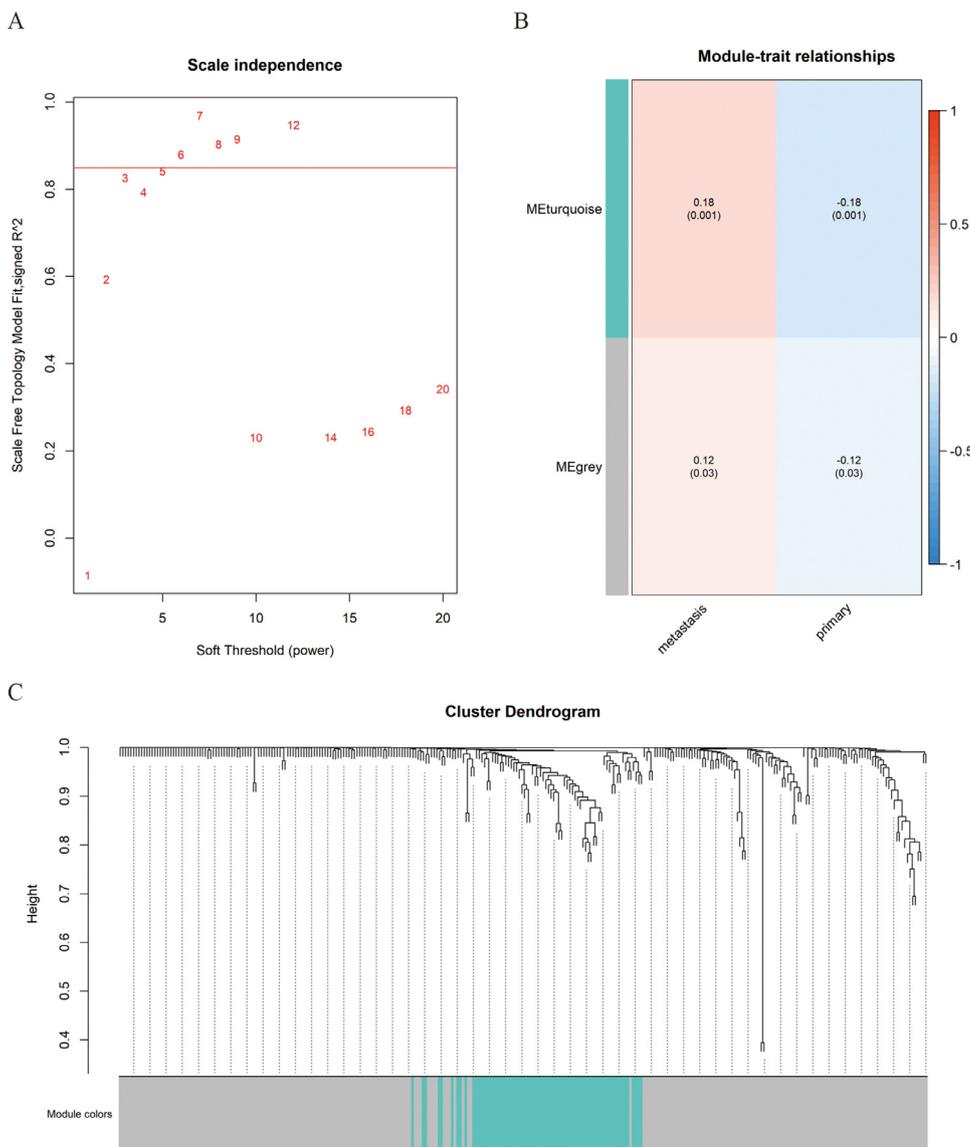
Fig. 3. The results of weighted co-expression network analysis. (A) The optimal β value result graph. (B) The Pearson's correlation coefficient for each module's ME and sample features (primary vs. metastatic). (C) The clustering result of the modules shows the one marked as turquoise in the cluster dendrogram with a total of 72 genes most relevant to the metastasis ME: module eigengene. The "β" stands for power value, "point" stands for a set of soft-thresholding power.

C

Cluster Dendrogram

substantially differentially expressed between HRG and LRG (Fig. 6A). In GSE45255, the abundance of activated NK cells, monocytes, M0 macrophages, M2 macrophages, resting dendritic cells, and neutrophils was substantially different between HRG and LRG (Fig. 6B).

Risk score prognostic performance and correlation with other clinical prognostic factors

Risk score analysis of the clinical characteristics of age and TNM stage revealed that HRG and LRG were considerably different in the stage I and stage II, and stage I and stage III groups, but were not substantially different among the other groups (Fig. 7A, B). The prognosis model had higher prediction accuracy than age and was basically consistent with the prognostic efficacy of TNM staging (Fig. 7C). Comparing the differences in survival between HRG and LRG of each TNM stage, the survival rates of HRG and LRG of TNM stage II were substantially different (Fig. 7F-I).

Univariate Cox regression analysis showed that TNM staging and the prognosis models had prognostic value, while age had no association with prognosis (Fig. 7D). The age, TNM stage, and prognosis models were included in the multivariate Cox regression analysis. The results

showed that the TNM stage and prognosis models were independent predictors of prognosis (Fig. 7E).

Screening of DEGs in bone metastasis and enrichment analysis of KEGG signaling pathway

GSE2034 and GSE124647 BCE chip datasets were downloaded from GEO. There were 217 vs. 69 and 19 vs. 11 primary cancer vs. bone metastasis samples. The Wilcox rank-sum test was used to assess the differences. When a *p*-value of < 0.05 was taken as the threshold, GSE2034 contained 2363 DEGs, including 1424 upregulated genes and 939 downregulated genes, while GSE124647 consisted of 1706 DEGs, including 888 upregulated genes and 818 downregulated genes. There were 368 DEGs shared between GSE2034 and GSE124647. Using R's clusterProfiler software package, KEGG Pathway enrichment analysis was performed on the DEGs in the GSE2034 and GSE124647 datasets. The enrichment results of the first 20 signal pathways are shown in Fig. 8A-B; the enriched signaling pathways mainly included Th17 cell differentiation, the PI3K/Akt signaling pathway, and the TGF-β signaling pathway. At the same time, we used Ingenuity Pathway Analysis (IPA) software to focus on the BMP signaling pathway in GSE2034 and GSE124647.
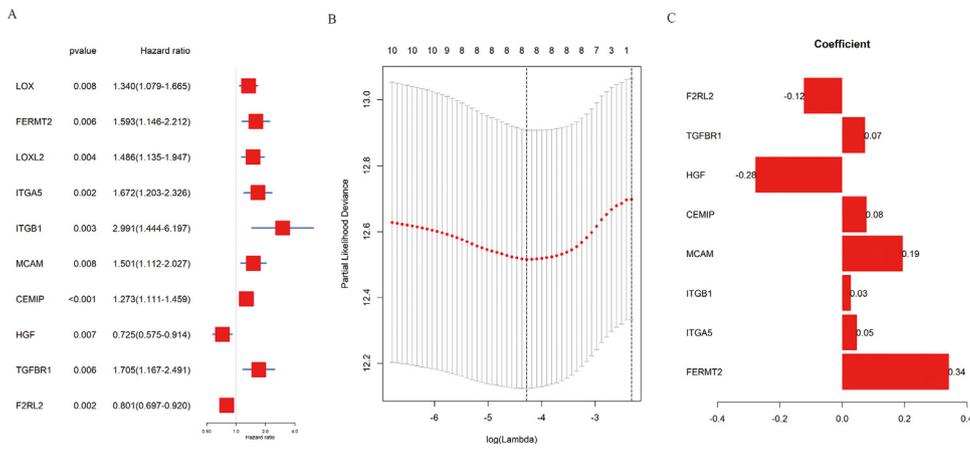
Fig. 4. Prognostic factor screening and model construction. **(A)** Forest map of the results of univariable Cox regression analysis, screening out 10 genes related to prognosis. **(B)** The determination of Lambda coefficient by Lasso regression. **(C)** Regression coefficients (coefficient values) of the eight genes constructed by the model.
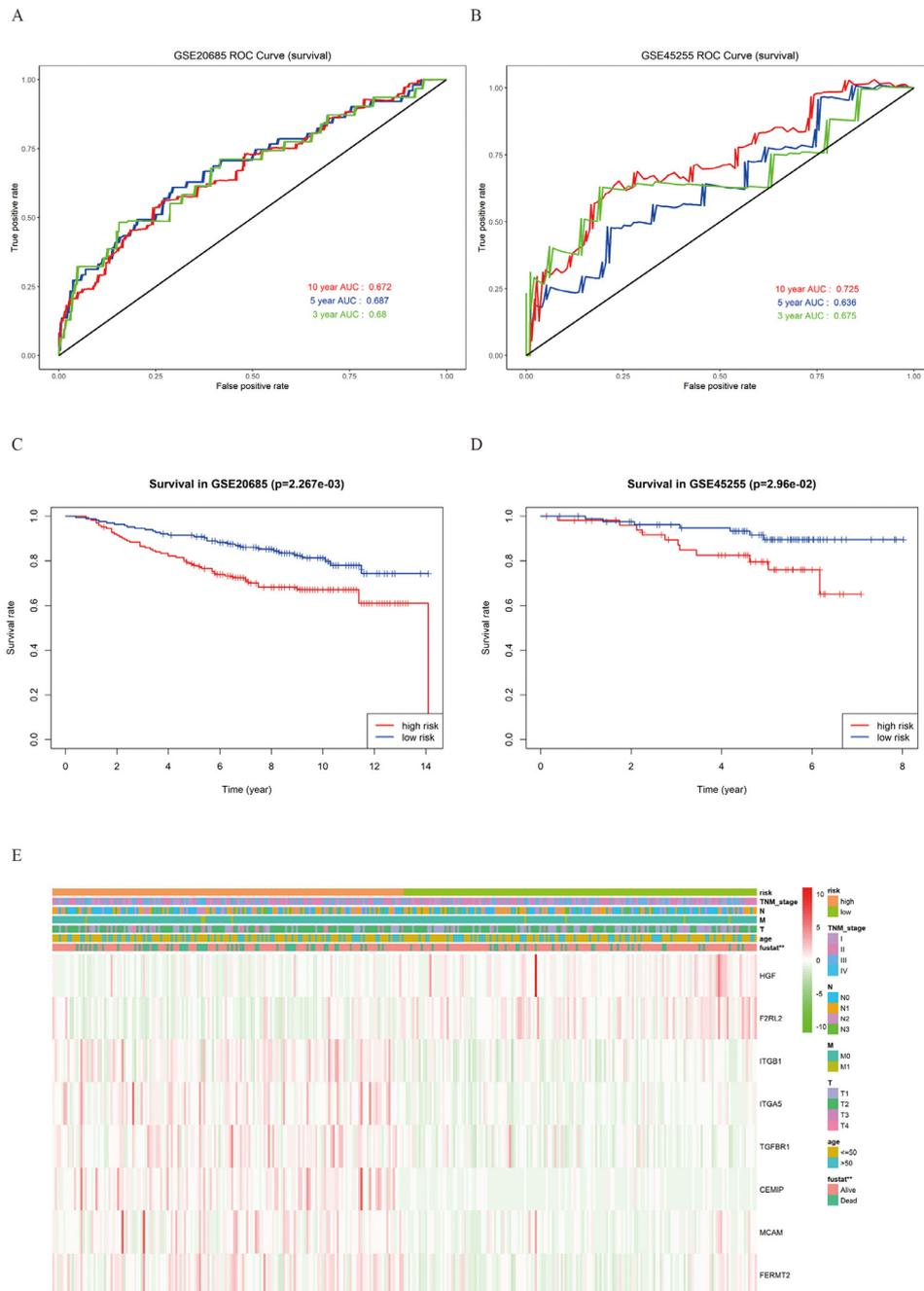
Fig. 5. Prognostic risk model evaluation chart. **(A)** GSE20685 ROC curve. **(B)** GSE45255 ROC curve. **(C)** Survival curve of GSE20685. **(D)** Survival curve of the verification dataset of GSE45255. **(E)** Modeled gene expression heat map of the high- and low-risk groups and the analysis of the differences in different clinical indicators between the high- and low-risk groups (** indicates $P < 0.01$).
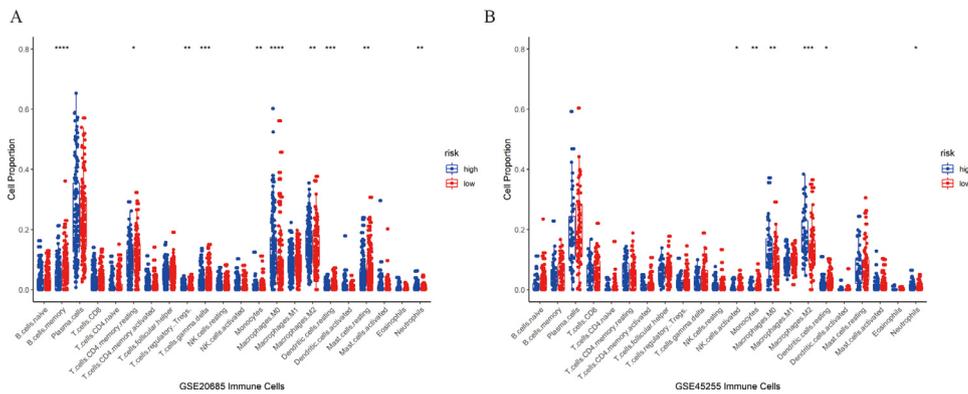
A
B

Table 1
BMP-related genes.

Symbol	Entrez Gene Name	Location
BMP8B	bone morphogenetic protein 8b	Extracellular Space
BMP2	bone morphogenetic protein 2	Extracellular Space
BMP3	bone morphogenetic protein 3	Extracellular Space
BMP4	bone morphogenetic protein 4	Extracellular Space
BMP5	bone morphogenetic protein 5	Extracellular Space
BMP6	bone morphogenetic protein 6	Extracellular Space
BMP7	bone morphogenetic protein 7	Extracellular Space
GDF2	growth differentiation factor 2	Extracellular Space
BMP10	bone morphogenetic protein 10	Extracellular Space
BMP15	bone morphogenetic protein 15	Extracellular Space
BMP2K	BMP2 inducible kinase	Nucleus
BMP8A	bone morphogenetic protein 8a	Extracellular Space
BMPR2	bone morphogenetic protein receptor type 2	Plasma Membrane
GDF10	growth differentiation factor 10	Extracellular Space
GDF7	growth differentiation factor 7	Extracellular Space
GDF11	growth differentiation factor 11	Extracellular Space
GDF5	growth differentiation factor 5	Extracellular Space
GDF6	growth differentiation factor 6	Extracellular Space
DAND5	DAN domain BMP antagonist family member 5	Extracellular Space
BMPER	BMP binding endothelial regulator	Extracellular Space
BAMBI	BMP and activin membrane bound inhibitor	Plasma Membrane
CER1	cerberus 1, DAN family BMP antagonist	Extracellular Space
GREM1	gremlin 1, DAN family BMP antagonist	Extracellular Space
GREM2	gremlin 2, DAN family BMP antagonist	Extracellular Space
HJV	hemojuvelin BMP co-receptor	Plasma Membrane
MICOS10	NBL1, DAN family BMP antagonist	Nucleus
KCP	kielin cysteine rich BMP regulator	Extracellular Space
CRIM1	cysteine rich transmembrane BMP regulator 1	Extracellular Space
RGMA	repulsive guidance molecule BMP co-receptor a	Plasma Membrane
RGMB	repulsive guidance molecule BMP co-receptor b	Plasma Membrane
TWSG1	twisted gastrulation BMP signaling modulator 1	Extracellular Space
NBL1	NBL1, DAN family BMP antagonist	Nucleus

GSE124647 was also enriched in the BMP signaling pathway, while the pathway was inhibited in GSE2034 (Fig. 8D).

Using IPA software to search for genes related to the BMP family, a total of 32 genes were obtained (Table 1). In the GSE2034 dataset, there were four differentially expressed BMP genes, including *BMPR2, BMP6, BMP2,* and *GREM1*. In the GSE124647 dataset, there were five differentially expressed BMP genes, including *BMPR2, BMP2, GREM1, BMP7,* and *BMP8A*. In both data sets, *BMPR2, BMP2,* and *GREM1* (Fig. 8C) were simultaneously differentially expressed.

Three differentially expressed BMP genes in GSE2034 were EMT-related genes, including *BMP6, BMP2,* and *GREM1*; three differentially expressed BMP genes in GSE124647 were EMT-related genes, including *BMP2, GREM1,* and *BMP7*.

Comparison of risk score between primary cancer and bone metastasis

The prognostic model was used to calculate the risk score values of all samples in the GSE2034 and GSE124647 datasets. The Wilcox rank-sum test found that the risk score differences between primary cancer

and bone MBC samples of the two data sets were not substantially different (Fig. 9).

Difference in immune cell infiltration in GSE2034 and GSE124647 between HRG and LRG

The CIBERSORT package of R was used to evaluate the abundance of 22 immune cells in the GSE2034 and GSE124647 datasets. The abundances of resting NK cells, M0 macrophages, and M2 macrophages between HRG and LRG were substantially different in GSE2034 (Fig. 10).

Co-expression analysis of immune cells and key genes

Based on the model constructed above, GSE20685 and GSE45255 were divided into HRG and LRG, respectively. Pearson's correlation coefficient demonstrated the correlation between the heat map of EMT-related DEGs and 22 types of lymphocytes in BC samples (Fig. 11). Figs. 11A-C show the correlation between genes and immune cell abundance, genes in HRG and immune cell abundance, genes in LRG and
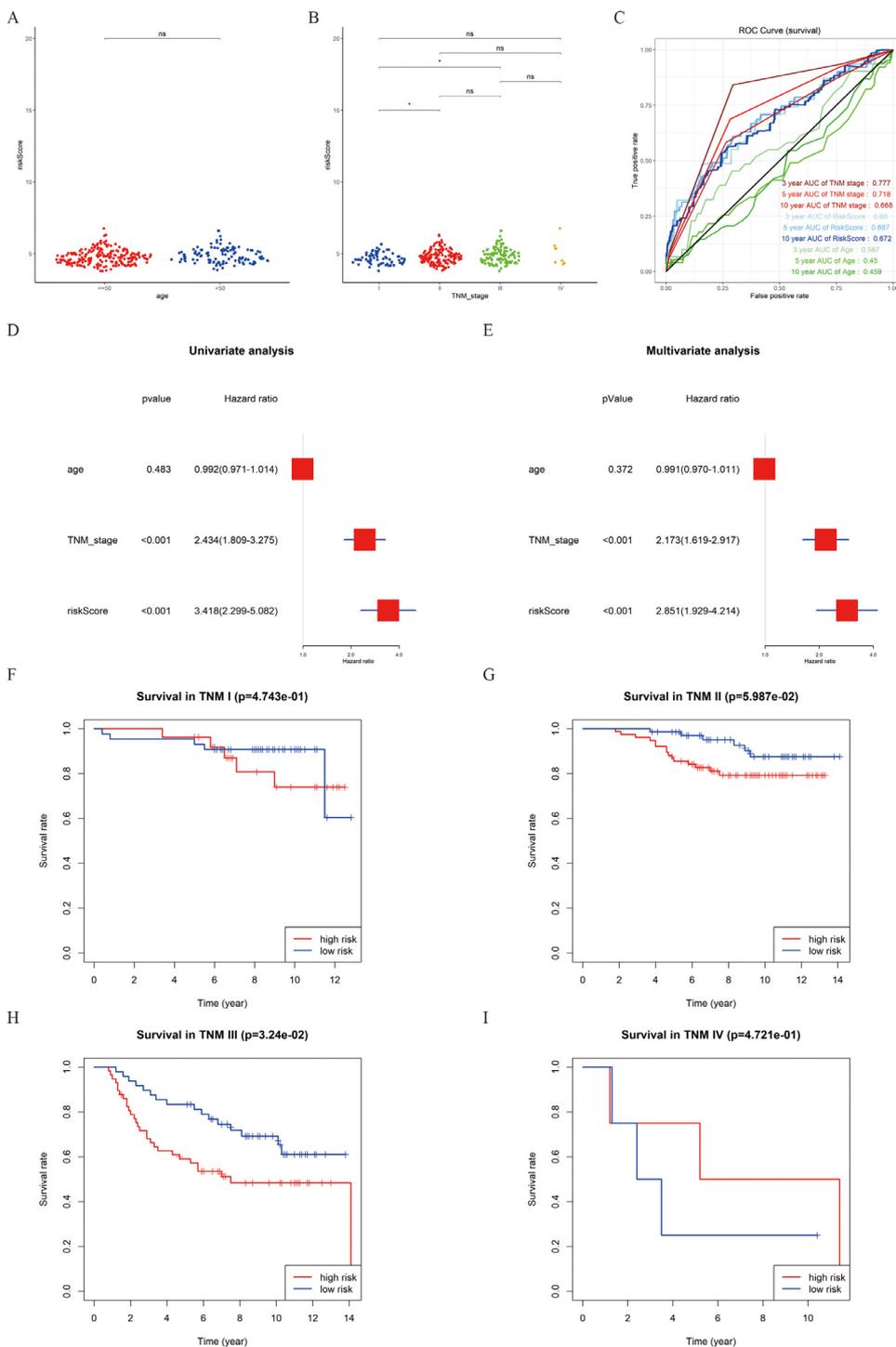
Fig. 7. Risk score prognostic performance and correlation with other clinical prognostic factors. **(A)** Comparison of age group risk score differences. **(B)** Comparison of risk score differences between TNM stages. **(C)** Risk scoring model, age, and TNM stage ROC curve; **(D)** Univariable Cox analysis for age, TNM stage, and risk assessment model. **(E)** Multivariable Cox analysis for age, TNM stage, and risk assessment model. **(F)** Survival curve of the high- and low-risk groups in TNM stage I. **(G)** Survival curve of high- and low-risk groups TNM stage II. **(H)** Survival curve of high- and low-risk groups in TNM stage III. **(I)** Survival curve of high- and low-risk groups in TNM stage IV.

immune cell abundance in GSE20685. Figs. 11D-F show the correlation between genes and immune cell abundance, genes in HRG and immune cell abundance, genes in LRG and immune cell abundance in GSE45255. Here we specifically analyzed the correlation between plasma cells, Naïve CD4+ T cells, and the EMT-related genes and found that plasma cells and eight genes, Naïve CD4+ T cells and eight genes were negatively correlated.

Discussion

BC constitutes one of the most common malignant tumors in women and is the primary cause of cancer death in women [1–3]. Among pa-

tients with MBC, bone metastasis occurred as the first metastatic site in about 83% of patients [1–3]. The median survival time of patients with BC with bone metastasis only is 24–40 months, which is longer than the median survival time for patients with metastasis in all other tissue types [3,9]; however, when bone metastasis occurs, the patient may present with severe bone pain, pathological fractures, nerve compression symptoms, hypercalcemia, or bone-related events (skeletal-related events), which seriously affect the quality of life of patients with advanced BC.

About 95% of BC cells are of epithelial origin, and their metastasis and invasion are the main causes of cancer death [3,10]. With in-depth studies of BC, people gradually realized the importance of EMT in the
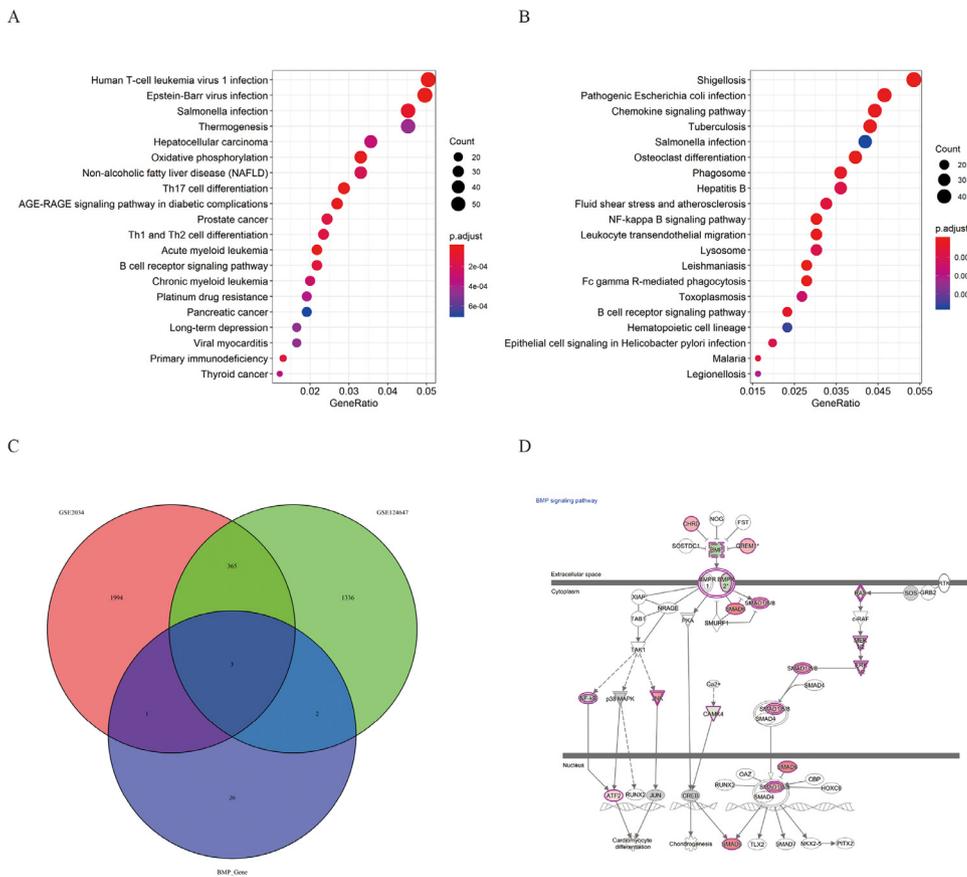
Fig. 8. (A) The enrichment of GSE2034 differentially expressed gene signaling pathways. **(B)** Enrichment of GSE124647 differentially expressed gene signaling pathways. **(C)** GSE2034 differentially expressed genes, GSE124647 differentially expressed genes, and Venn diagram of BMP-related gene family. **(D)** BMP signaling pathway was inhibited in GSE2034.
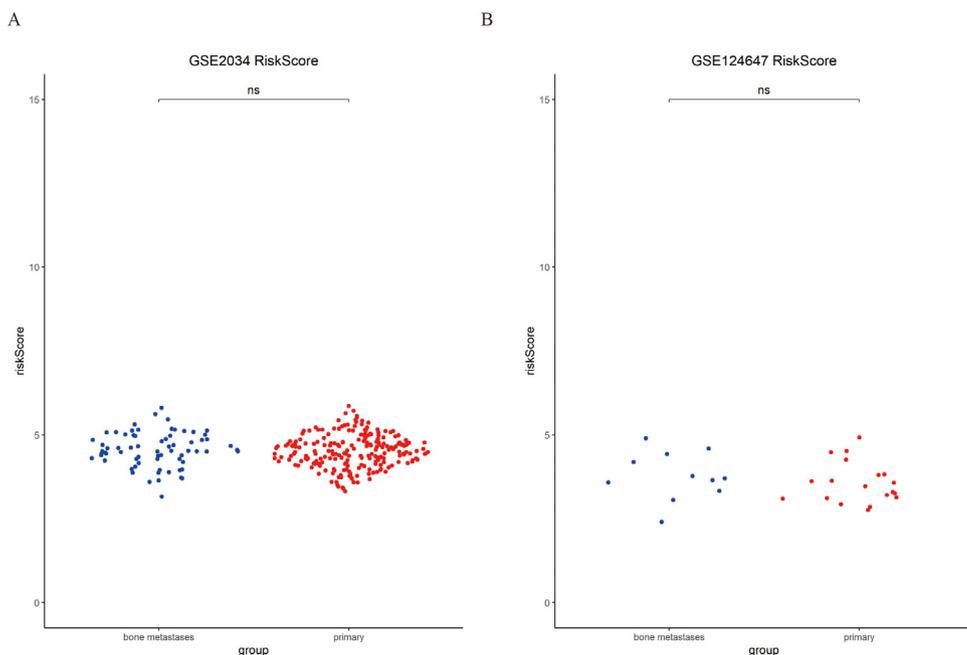
Fig. 9. (A) Comparative result of risk score between GSE2034 primary cancer and bone metastatic breast cancer samples. **(B)** Comparative result of risk score between GSE124647 primary cancer and bone metastasis breast cancer samples.

occurrence and metastasis [6,7,10–12]. Tumor cells often need to lose their epithelial-like phenotypic characteristics, such as polarity and intercellular contact via EMT to invade normal tissues and metastasize [6,7,10–14]. The metastasis of BC epithelial cells, the development of breast ductal carcinomas in situ into invasive BC, the poor efficacy of BC treatment, and the resistance of patients to chemotherapy drugs are all

closely related to the EMT process [6,7,10–14]. In this study, we used bioinformatics analysis methods to construct a full-genome expression profile of BC metastasis samples combined with an EMT-related gene weighted co-expression network and BC prognosis evaluation model to provide the basis and reference for calculating prognostic risk values for BC patients. Several studies have confirmed that genes, such as *BMP-2,*
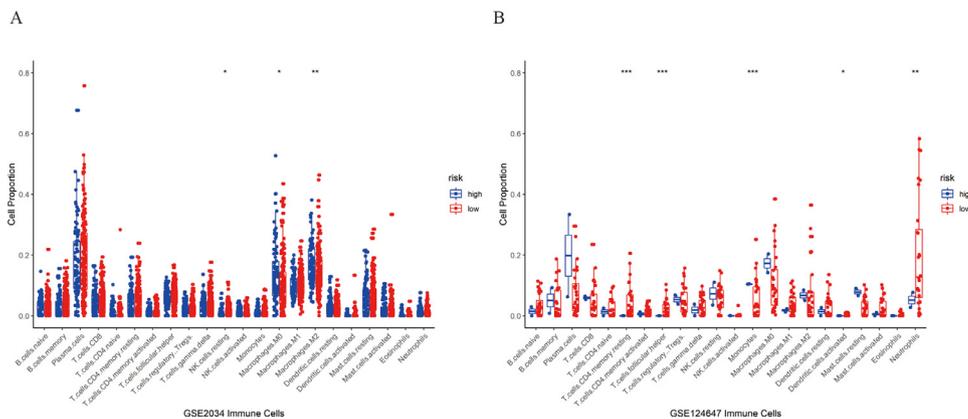
Fig. 10. The abundance of 22 immune cells of high- and low-risk groups in GSE2034 **(A)** and GSE124647 **(B)** datasets.
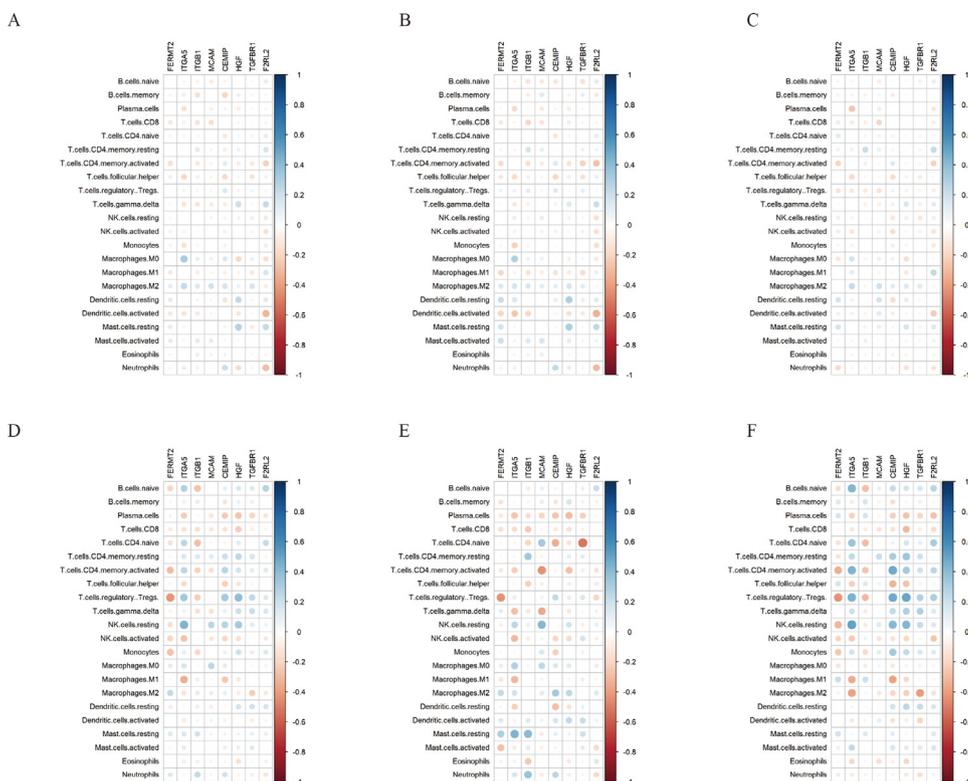
Fig. 11. (A) Correlation between genes and immune cell abundance in GSE20685. **(B)** Correlation between genes in high-risk group and immune cell abundance in GSE20685. **(C)** Correlation between genes in low-risk genes in group and immune cell abundance in GSE20685. **(D)** Correlation between genes and immune cell abundance in GSE45255. **(E)** Correlation between genes in high-risk group and immune cell abundance in GSE45255. **(F)** Correlation between genes in low-risk group and immune cell abundance in GSE45255.

BMPR2, GREM1, CEMIP, HGF, ITGA5, ITGB1, MCAM, and *TGFBR1*, are closely related to the EMT process in BC cells, which is also consistent with our research results [6,7,10–17].

Bone morphogenetic proteins (BMPs) are important members of the TGF-β superfamily. At present, more than 20 BMPs have been discovered, and their biological functions are very extensive, including promoting bone and cartilage production, regulating cell proliferation and differentiation, and regulating the growth and development of various organs [18–23]. Studies have found that BMP-2 can activate Smad-independent pathways in tumor cells, including in the cancers of the breast, stomach, colon, and pancreas; for example, the PI3K/Akt signaling pathway to promote tumor cell invasion and metastasis [24]. However, the relationship between BMP-2 gene expression and BC metastasis to bones needs to be further explored. Researchers have detected BMP-2 mRNA in many BC cell lines and found that it has different levels of expression in some primary BC tissues [20]. In addition, rhBMP-2 induces EMT in three BC cell lines (MCF-7, MDA-MB-231, and mouse BC cell line 4T1) and enhances cell migration and invasion ability both in vivo and in vitro [19,22]. Aberrant expression of BMP-2 has been observed

in various tumors and is known to be closely related to EMT induction and tumor invasion [18–20,22,23].

In a further in-depth study, we conducted a differential analysis of the GEO datasets GSE2034 and GSE124647 in relation to bone metastasis of BC and applied KEGG pathway analysis to the differential genes, which were found to be mainly related to Th17 cell differentiation, the PI3K/Akt and the TGF-β signaling pathways. In GSE2034, the BMP signaling pathway was in an inhibitory state. In both data sets, *BMPR2, BMP2*, and *GREM1* were differentially expressed. This further confirms the importance of BMP-2 in the process of BC metastasis to bones. Our research also found that multiple Hub genes (including *MCAM, HGF, TGFBR1, BMPR2*) closely related to BC metastasis are biologically linked to BMP-2, which confirms that BMP-2 and its pathway might be involved in BC metastasis or bone metastasis. However, the occurrence and development of tumors are both influenced by multiple factors. BMP-2 has a complex and extensive effect on tumors, and different types of tumors show different responses. There have been in-depth studies of BMP-2 in tumor tissues, and the functions of BMPs have been further explored.

The BMP2-mediated decrease in E-cadherin levels and increased levels of Snai1 and vimentin further indicate that BMP2-activated signaling is related to EMT-mediated morphological changes in these cells [21–25]. Based on the results of in vivo and in vitro experiments, Huang et al. [14] established the EMT process of the induction of BC stem cells through Rb and CD44 signaling pathways by rhBMP-2, by which it promotes the complete mechanism of BC metastasis. Zhang et al. [19] showed that BMP-2 induced EMT in BC stem cells through Rb and CD44 signaling pathways [14,19]. They showed that both PI3K/Akt and Smad signaling regulated rhBMP-2-mediated RB and CD44 expression. In vitro and in vivo results highlighted the important roles of BMP-2, Rb, and CD44 in BC metastasis, which could be used for diagnosing and treating advanced BC [14,19]. Also, BMP-2 is known to be closely related to BC calcification in orthotopic breast tumors [21–25]. The process of BC calcification requires high expression levels of BMP-2 and the biological behavior of BC cells. However, high concentrations of BMP-2 inhibit the proliferation, migration, and invasion of BC cells [21–25].

Current research on BC has found a relationship between EMT-related signaling pathways and TGF-β, NF-κB, Notch, Wnt, PI3K/AKT, MAPK signaling pathways [5,26]. The signaling pathways involved in BC metastasis are extremely complex, and there are close interactions among the BMP-2 pathway, TGF-β pathway, and PI3K/Akt signaling pathway [27]. TGF-β1 can phosphorylate Akt/PKB and activate the Akt signaling pathway, which, in turn, participates in TGF-β signaling pathway-mediated EMT process of breast epithelial cells. Thus, inhibition of PI3K/Akt signaling pathway concomitantly facilitates the inhibition of TGFβ−1-induced Smad2 phosphorylation and the promotion of BC cell migration [28].

Currently, a key issue regarding immunotherapy for BC patients is the limited understanding of the complexity of the tumors, tumor heterogeneity, and immune escape mechanisms. In addition, there is a lack of precise biomarkers to evaluate the efficacy of tumor immunotherapy, and the development of new immunotherapy targets and prognostic markers is necessary [29]. Wei et al. [30] found that BMP-2-induced osteogenesis might regulate the local bone immune environment. The active role of BMP-2 in regulating the immune response makes it possible to use the immunomodulatory properties of BMP-2 to regulate the bone immune environment to promote bone regeneration [30]. Thus, here, we compared the abundance of 22 immune cells in a HRG and LRG. In GSE20685, the abundances of memory B cell, resting memory T cells CD4, Tregs, gamma delta T cells, monocytes, M0 macrophages, M2 macrophages, resting DCs, resting mast cells, and neutrophils were substantially different between HRG and LRG. In GSE45255, the abundances of activated NK cells, monocytes, M0 macrophages, M2 macrophages, resting dendritic cells, and neutrophils were substantially different in HRG and LRG. In a study of the characteristics of immune infiltration in BC bone metastasis samples, we compared the abundance of 22 immune cells in HRG and LRG of the GSE2034 dataset. We found that the abundances of resting NK cells, M0 macrophages, and M2 macrophages were substantially different between HRG and LRG. In addition, many international studies have shown that *BMP2, BMPR2, GREM1, HGF, MCAM, TGFBR1,* and other genes might promote distant metastasis and the local progression of BC [29–35].

Gremlin1 (*GREM1*) is a member of the bone morphogenetic protein (BMP) antagonist family [31]. It binds with BMP-2, BMP-4, BMP-7, and other members of the BMP family to exert its antagonistic role; thus, inhibiting the binding ability of these ligands to receptors [31]. Recent studies have confirmed that *GREM1* regulates the feedback signals from epithelial cells to mesenchymal cells [36,37]. *GREM1* plays an independent role in promoting angiogenesis and tumor formation, which may regulate the proliferation and invasion of tumor cells by acting on the BMP signaling pathway and plays a vital role in many tumor biological processes. *GREM1* is a bone morphogenetic protein antagonist, which plays a specific role in the EMT process [6,31,36–38]. The EMT phenomenon induced by *GREM1* is vital for organ fibrosis and tumor pro-

gression. It has been reported that *GREM1* regulates the occurrence of EMT and promotes cell proliferation and migration by activating TGF-β/Smad pathway [16,39]. The knockout of the *GREM1* gene inhibits the invasion and EMT process of glioma cells. However, the role of *GREM1* in EMT of human BC cells and the exact effect of *GREM1* on EMT is unclear. The exact mechanism of *GREM1* activation, its relationship with BC bone metastasis, and its role in BC angiogenesis needs further investigation.

This study had several limitations. First, all data was collected from Western countries, so caution should be exercised when applying this conclusion to patients in Asian countries. The inadequate availability of public information limited the comprehensive analysis of clinical and pathological parameters, which might lead to bias in analysis. Thus, we tried to minimize bias by assessing the gene and protein expression of key biomarkers at the cellular and tissue levels. The results show that the key biomarkers in our model have undergone significant changes in BC metastasis or bone metastasis. However, despite these limitations, our study first applied bioinformatics methods to explore changes in the key genes related to BC metastasis and predict the prognosis of BC patients based on tumor-infiltrating immune cells of BC metastasis or bone metastasis.

Conclusions

Based on the construction of a weighted co-expression network for EMT-related DEGs in MBC, we screened hub genes to explore a prognostic model and the immune infiltration patterns of MBC. The results of this study provide a solid basis to bioinformatically explore the molecular mechanisms of BC metastasis and enable the prediction of patient survival. We also found that *BMP2, BMPR2*, and *GREM1* were differentially expressed in both data sets of BC bone metastasis. Based on the results of bioinformatics analysis in this study, we speculated that BMP-2 might regulate the immune infiltration process in BC tissues through the PI3K/Akt signaling pathway, thereby affecting the prognosis of cancer.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Data availability statement

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

Authors' contributions

Conceptualization, Shuzhong Liu, An Song, Yunxiao Wu, Bo Yang, Yong Liu and Yipeng Wang; Formal analysis, Shuzhong Liu, An Song, Yunxiao Wu, Muchuan Wang, Tong Niu, Siyuan Yao, Chengao Gao, Ziquan Li, Xi Zhou, Zhen Huo, Siyuan Yao, Bo Yang, and Yong Liu; Funding acquisition, Shuzhong Liu and Yipeng Wang; Methodology, Shuzhong Liu, Siyuan Yao, Bo Yang, and Yipeng Wang; Software, Xi Zhou, Siyuan Yao and Yong Liu; Supervision, Bo Yang, Yong Liu and Yipeng Wang;

Writing – original draft, Shuzhong Liu and An Song; Writing – review & editing, Shuzhong Liu, Bo Yang, and Yipeng Wang.

Credit authorship contribution statement

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.tranon.2020.100993.

References

[1] C.E. DeSantis, J. Ma, M.M. Gaudet, et al., Breast cancer statistics, 2019, CA Cancer J. Clin. 69 (6) (2019) 438–451.

[2] R.L. Siegel, K.D. Miller, A Jemal, Cancer statistics, 2019, CA Cancer J. Clin. 69 (1) (2019) 7–34.

[3] F. Bray, J. Ferlay, I. Soerjomataram, R.L. Siegel, L.A. Torre, A Jemal, Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries, CA Cancer J. Clin. 68 (6) (2018) 394–424.

[4] M.Y. Wu, C.J. Li, G.T. Yiang, et al., Molecular regulation of bone metastasis pathogenesis, Cell Physiol. Biochem. 46 (4) (2018) 1423–1438.

[5] C. Foroni, M. Broggini, D. Generali, G Damia, Epithelial-mesenchymal transition and breast cancer: role, molecular mechanisms and clinical impact, Cancer Treat. Rev. 38 (6) (2012) 689–697.

[6] L. Ren, W. Mo, L. Wang, X Wang, Matrine suppresses breast cancer metastasis by targeting ITGB1 and inhibiting epithelial-to-mesenchymal transition, Exp. Ther. Med. 19 (1) (2020) 367–374.

[7] T.S. Gerashchenko, M.V. Zavyalova, E.V. Denisov, et al., Intratumoral morphological heterogeneity of breast cancer as an indicator of the metastatic potential and tumor chemosensitivity, Acta Nature 9 (1) (2017) 56–67.

[8] M. Zhao, L. Kong, Y. Liu, H Qu, dbEMT: an epithelial-mesenchymal transition associated gene resource, Sci Rep 5 (2015) 11459.

[9] S.J. Lee, S Park, H.K. Ahn, et al., Implications of bone-only metastases in breast cancer: favorable preference with excellent outcomes of hormone receptor positive breast cancer, Cancer Res. Treat. 43 (2) (2011) 89–95.

[10] C.M. Su, W.H. Lee, A.T. Wu, et al., Pterostilbene inhibits triple-negative breast cancer metastasis via inducing microRNA-205 expression and negatively modulates epithelial-to-mesenchymal transition, J. Nutr. Biochem. 26 (6) (2015) 675–685.

[11] P. Bendinelli, P. Maroni, V. Dall'Olio, E. Matteucci, M.A Desiderio, Bone metastasis phenotype and growth undergo regulation by micro-environment stimuli: efficacy of early therapy with HGF or TGFbeta1-type i receptor blockade, Int. J. Mol. Sci. 20 (10) (2019).

[12] Liang YK Zeng, Y.S. Xiao, et al., Inhibition of Notch1 reverses EMT and chemoresistance to cisplatin via direct downregulation of MCAM in triple-negative breast cancer cells, Int. J. Cancer 147 (2) (2020) 490–504.

[13] N.J. Sung, N.H. Kim, N.Y. Bae, H.S. Jo, S.A Park, DHA inhibits Gremlin-1-induced epithelial-to-mesenchymal transition via ERK suppression in human breast cancer cells, Biosci. Rep. 40 (3) (2020).

[14] P. Huang, A. Chen, W. He, et al., BMP-2 induces EMT and breast cancer stemness through Rb and CD44, Cell Death Discov. 3 (2017) 17039.

[15] F. Fico, M. Bousquenaud, C. Ruegg, A Santamaria-Martinez, Breast cancer stem cells with tumor- versus metastasis-initiating capacities are modulated by TGFBR1 inhibition, Stem Cell Reports 13 (1) (2019) 1–9.

[16] D. Hong, T. Liu, W. Huang, et al., Gremlin1 delivered by mesenchymal stromal cells promoted epithelial-mesenchymal transition in human esophageal squamous cell carcinoma, Cell Physiol. Biochem. 47 (5) (2018) 1785–1799.

[17] L. Ren, W. Mo, L. Wang, X Wang, Matrine suppresses breast cancer metastasis by targeting ITGB1 and inhibiting epithelial-to-mesenchymal transition, Exp. Ther. Med. 19 (1) (2020) 367–374.

[18] M. Scimeca, R. Giocondo, M. Montanaro, et al., BMP-2 variants in breast epithelial to mesenchymal transition and microcalcifications origin, Cells 9 (6) (2020).

[19] G. Zhang, P. Huang, A. Chen, et al., How BMP-2 induces EMT and breast cancer stemness through Rb and CD44? Cell Death Dis. 9 (2) (2018) 20.

[20] M. Colleoni, A. O'Neill, A. Goldhirsch, et al., Identifying breast cancer patients at high risk for bone metastases, J. Clin. Oncol. 18 (23) (2000) 3925–3935.

[21] S.F. Arnold, E. Tims, B.E Mcgrath, Identification of bone morphogenetic proteins and their receptors in human breast cancer cell lines: importance of BMP2, Cytokine 11 (12) (1999) 1031–1037.

[22] A. Chen, D. Wang, X. Liu, S. He, Z. Yu, J Wang, Inhibitory effect of BMP-2 on the proliferation of breast cancer cells, Mol. Med. Rep. 6 (3) (2012) 615–620.

[23] C.C. Tan, G.X. Li, L.D. Tan, et al., Breast cancer cells obtain an osteomimetic feature via epithelial-mesenchymal transition that have undergone BMP2/RUNX2 signaling pathway induction, Oncotarget 7 (48) (2016) 79688–79705.

[24] M.H. Kang, J.S. Kim, J.E. Seo, S.C. Oh, Y.A Yoo, BMP2 accelerates the motility and invasiveness of gastric cancer cells via activation of the phosphatidylinositol 3-kinase (PI3K)/Akt pathway, Exp. Cell Res. 316 (1) (2010) 24–37.

[25] M.H. Kang, H.N. Kang, J.L. Kim, J.S. Kim, S.C. Oh, Y.A Yoo, Inhibition of PI3 kinase/Akt pathway is required for BMP2-induced EMT and invasion, Oncol. Rep. 22 (3) (2009) 525–534.

[26] D.M. Gonzalez, D. Medici, Signaling mechanisms of the epithelial-mesenchymal transition, Sci. Signal 7 (344) (2014) re8.

[27] W. Xu, Z. Yang, N Lu, A new role for the PI3K/Akt signaling pathway in the epithelial-mesenchymal transition, Cell Adh. Migr. 9 (4) (2015) 317–324.

[28] A.V. Bakin, A.K. Tomlinson, N.A. Bhowmick, H.L. Moses, C.L Arteaga, Phosphatidylinositol 3-kinase function is required for transforming growth factor beta-mediated epithelial to mesenchymal transition and cell migration, J. Biol. Chem. 275 (47) (2000) 36803–36810.

[29] C.J. Block, G. Dyson, I.J. Campeanu, D. Watza, M. Ratnam, G Wu, A stroma-corrected ZEB1 transcriptional signature is inversely associated with antitumor immune activity in breast cancer, Sci Rep 9 (1) (2019) 17807.

[30] F. Wei, Y. Zhou, J. Wang, C. Liu, Y Xiao, The immunomodulatory role of BMP-2 on macrophages to accelerate osteogenesis, Tissue Eng. Part A 24 (7–8) (2018) 584–594.

[31] C. Lavoz, M. Alique, R. Rodrigues-Diez, et al., Gremlin regulates renal inflammation via the vascular endothelial growth factor receptor 2 pathway, J. Pathol. 236 (4) (2015) 407–420.

[32] N. Glodde, T. Bald, D. van den Boorn-Konijnenberg, et al., Reactive neutrophil responses dependent on the receptor tyrosine kinase c-met limit cancer immunotherapy, Immunity 47 (4) (2017) 789–802 .e9.

[33] M. Boutet, L. Gauthier, M. Leclerc, et al., TGFbeta signaling intersects with CD103 integrin signaling to promote T-lymphocyte accumulation and antitumor activity in the lung tumor microenvironment, Cancer Res. 76 (7) (2016) 1757–1769.

[34] C. Li, S. Jiao, G. Wang, et al., The immune adaptor ADAP regulates reciprocal TGF–beta1-integrin crosstalk to protect from influenza virus infection, PLoS Pathog 11 (4) (2015) e1004824.

[35] T. Dokoshi, L.J. Zhang, F. Li, et al., Hyaluronan degradation by Cemip regulates host defense against staphylococcus aureus skin infection, Cell Rep 30 (1) (2020) 61–68.

[36] A. Droguett, P. Krall, M.E. Burgos, et al., Tubular overexpression of gremlin induces renal damage susceptibility in mice, PLoS One 9 (7) (2014) e101879.

[37] S. Mezzano, A. Droguett, M.E. Burgos, et al., Expression of gremlin, a bone morphogenetic protein antagonist, in glomerular crescents of pauci-immune glomerulonephritis, Nephrol. Dial Transpl. 22 (7) (2007) 1882–1890.

[38] N.J. Sung, N.H. Kim, N.Y. Bae, et al., DHA inhibits Gremlin-1-induced epithelial–to-mesenchymal transition via ERK suppression in human breast cancer cells, Biosci Rep 40 (3) (2020).

[39] Y. Liu, Y. Li, R Hou, et al., Knockdown GREM1 suppresses cell growth, angiogenesis, and epithelial-mesenchymal transition in colon cancer, J. Cell Biochem. 120 (4) (2019) 5583–5596.