scientific reports

Check for updates

OPEN

m^5C related-regulator-mediated methylation modification patterns and prognostic significance in breast cancer

Zhe Wang[1,8], Jinpeng Li[2,8], Fucheng Wang[3], Chen Cheng[4], Xinpei Wu[4], Wendi Guo[5], Chenquan Li[6], Yinyi Luo[4], Guangwen Zhang[4], Sanyuan Zhang[1], Jie Hou[4]✉, Wei Wang[7]✉ & Shiming Wang[4]✉

5-Methylcytosine (m^5C) is closely associated with cancer. However, the role of m^5C in breast cancer(BC) remains unclear. This study combined single-cell RNA sequencing (scRNA-Seq) and transcriptomics datasets to screen m^5C regulators associated with BC progression and analyze their clinical values. Firstly, This study elucidates the mechanisms of the m^5C landscape and the specific roles of m^5C regulators in BC patients. we found that the dysregulation of m^5C regulators with m^5Cscore play the essential role of the carcinogenesis and progression in epithelial cells and myeloid cells of BC at single cell level. External validation was conducted using an independent scRNA-Seq datasets. Then, three distinct m^5C modification patterns were identified by transcriptomics datasets. Based on the m^5C differentially expressed regulators, the m^5Cscore was constructed, and used to divide patients with BC into high and low m^5Cscore groups. Patients with a high m^5Cscore had more abundant immune cell infiltration, stronger antitumor immunity, and better prognoses. Finally, Quantitative real-time (PCR) and immunohistochemistry were used for the in vitro experimental validation, which had extensive prognostic value. In this study, we aimed to assess the expression of m^5C regulators involved in BC and investigate their correlation with the tumor microenvironment, clinicopathological characteristics, and prognosis of BC. The m^5C regulators could be used to effectively assess the cell specific regulation prognosis of patients with BC and develop more effective immunotherapy strategies.

Keywords m^5C RNA modification, Breast cancer, Epigenetics, Cell specific regulation, Immunotherapy

Breast cancer (BC) is now the most diagnosed cancer and the leading cause of cancer-related death in women[1]. The incidence of BC has risen in most of the past four decades; approximately 13% of women will be diagnosed with invasive BC and 3% will die from the disease in their lifetimes[2]. In BC, triple-negative breast cancer (TNBC) is characterized by genomic instability and a higher mutation rate, which makes it aggressive, prone to early recurrence, and associated with a poor prognosis[3,4]. Therefore, it is crucial to enhance research focused on elderly TNBC patients in order to improve the therapeutic outcomes for this special population.

Methylation of C5 cytosine (m^5C) is a dynamic and reversible process that can affect various aspects of RNA metabolism. Three different types of proteins regulate the m^5C modification: methyltransferases, demethylases, and binding proteins, which are also termed "writers", "erasers", and "readers" respectively[5]. In humans, m^5C RNA methylation is generally catalyzed by NOP2/NSUN family (*NSUN1–7*) and DNA methyltransferase member (*DNMT*, *DNMT3A*, *DNMT3B*, and *TRDMT1*), with a residue specificity[6], and some documents have reported that the demethylation process predominantly relied on the ten-eleven translocator family (*TET*) and

[1]Department of Gynecology, First Hospital of Shanxi Medical University, Taiyuan, Shanxi, China. [2]Department of Obstetrics and Gynecology, Taiyuan People's Hospital, Taiyuan, Shanxi, China. [3]Shanxi Medical University, Taiyuan, Shanxi, China. [4]Department of Breast Surgery, First Hospital of Shanxi Medical University, Taiyuan, Shanxi, China. [5]School of Medicine, Nankai University, Tianjin 300071, China. [6]Department of Breast Surgery, Second Hospital of Shanxi Medical University, Taiyuan, Shanxi, China. [7]Department of General Dentistry, Shanghai Ninth People's Hospital, College of Stomatology, National Center for Stomatology, Shanghai Key Laboratory of Stomatology, Shanghai Jiao Tong University School of Medicine, Shanghai Jiao Tong University, National Clinical Research Center for Oral Diseases, Shanghai, China. [8]These authors contributed equally: Zhe Wang and Jinpeng Li. ✉email: houjie0213@163.com; wangwei_1212d@163.com; wangshiming@sxmu.edu.cn

Alpha-Ketoglutarate-Dependent Dioxygenase AlkB Homolog 1 (*ALKBH1*)[7,8]. The aberrant levels of m^5C and its regulators are closely associated with various human diseases, including cancer[9]. The biological functions of the m5C regulator and its regulatory mechanisms are summarized in the supplemental Table S1.

The BC tumor microenvironment (TME) is a complex, dynamic entity. It plays a key role in various biological behaviors of BC, such as inhibiting apoptosis, promoting tumor progression and immune escape, and inducing immune tolerance[10]. Several studies have revealed the association between m^5C modification and TME-infiltrating immune cells, such as the deficiency of *TET2* and *TET3* in regulatory T cells (Treg cells) results in a dysregulated expression of multiple Treg-activation and phenotypic molecules, leading to deleterious effectors, and eliciting disease in healthy mice[11]. Therefore, clearing the role of m^5C related regulatory factors in immune cell infiltration in TME can help predict immune therapy responses, improve the success rate of existing immune therapies, and develop new immune therapy strategies[12].

Single-cell RNA sequencing (scRNA-seq) is a powerful tool that can provide expression profiling of human cancer at the resolution of individual cells, which allows the identification and characterization of specific subclusters that bear unique biological effects[13]. Such studies have been widely performed in breast cancer research to investigate the TME and the evolvement of tumor cells[14].

In this study, we collected 5 normal samples and 5 BC samples for single-cell RNA-seq data and used a gene expression dataset, consisting of eligible 56,619 cells to analyze the m^5C regulation landscape at the molecular level. We found that m^5Cscore is of great significance in analyzing and evaluating immune infiltration and prognosis of BC. In addition, based on the m^5C-related differentially expressed genes (DEGs), the m^5Cscore was constructed, and used to divide BC patients into high and low m^5Cscore groups. Altogether, we provided a new perspective that m^5C modification could be a potential epigenetic mechanism in BC development, which could provide a reference for reasonable diagnosis and treatment strategies.

Results

Distribution and expression profiles of m5C regulators across diverse cell types regulated BC progression and their associated signaling pathways

To understand the cellular diversity and molecular features of the breast tissue in BC patients, five normal samples and five BC samples were collected for single-cell RNA-seq data. After quality control, 56,619 cells were retained for subsequent analysis, comprising 24,330 cells from normal samples and 32,289 from BC samples. Five known cell types including 42,712 epithelial cells, 3568 lymphocytes, 3221 fibroblasts, 2840 myeloid cells and 4278 endothelial cells were identified and annotated by using classical marker genes[15,16] (Fig. 1a). To investigate the role of m^5C regulator regulation in BC, the heatmaps show the distribution of m^5C regulators at the single cell level. We found that m^5C regulators both distribution of genes either expressed and each cell population are heterogeneous in five cell types (Fig. 1b-c). To reveal the DEGs of m^5C regulators between BC and normal group, we found endothelial cells and epithelial cells show more differential genes than others (Fig. 1d). The *DNMT1* is over-expressed and *MECP2* is down regulation in myeloid cells (Fig. 1e). To identify the cell-type expression and correlation between m5C regulators and cell types, We used the upset plot to analyze the intersection of m^5C regulators and marker genes of diverse cell types. m^5C had little overlap with marker genes of each cell type in BC (Fig S1a). Survival analysis indicates that within the first 100 months, the survival rate of the low-expression group of *DNMT3B* is superior to that of the high-expression group (Fig. 1f). We used the AddModuleScore function to define m5Cscore, we found that epithelial cells and endothelial cells have higher scores, suggesting that m^5C-related regulators play function more in epithelial cells and endothelial cells (Fig. 1g). *DNMT1*and *ALYREF* were found to have broad-spectrum expression in BC samples (Fig. 1h-i). To investigate the association between the m^5C-regulation and progression of BC, we used the functional enrichment analysis based on the GSEA database, we performed the correlation between m^5C regulators by AUCell and classical pathways in BC to explore the influence of m^5C regulators on breast cancer pathways. Notably, the heatmap shows that most m^5C-related regulators exhibit high expression levels in the WNT pathway and in BRCA1 pathway (Fig. 2a, S1b-c). Next, we observed the differences in m^5C signature across various cell types using violin plots and found that in endothelial, epithelial, and myeloid cells, the signature in the tumor group were higher than those in the normal group (Fig. 2b-d). To infer biologically interpretable results, using Metascape's functional enrichment analysis capability, the several most significantly enriched ontology terms were combined to annotate the putative biological roles of the m^5C-related regulators, such as chromatin organization, mitotic cell cycle processm, RNA metabolic process, cell division and DNA damage response (Fig S1d). Beside, the cumulative frequency curve plot shows that the cumulative frequency in the tumor group is significantly higher than that in the normal group, indicating that the m^5c signature in the tumor group is higher than in the normal group (Fig S1e). To explore whether m^5C regulators were associated with pathways were significantly correlated with tumor progression. We used the Metacell algorithm (K = 40) and conducted a correlation analysis by WGCNA, WGCNA was performed to identify the biological functions of m^5C regulators in each cell types in BC. WGCNA showed that some genes were closely related to *ALYREF* expression in epithelial cells, pathway enrichment analysis to verify the biological functions of genes of *ALYREF* related module in epithelial cells, like cellular response to tumor necrosis, intrinsic apoptotic signaling pathway and DNA damage response (Fig. 2e-f); And some genes were closely related to *DNMT1* expression in myeloid cells, pathway enrichment analysis to verify the biological functions of genes of *DNMT1* related module in myeloid cells, like regulation of tumor necrosis factor production, macrophage differentiation, myeloid leukocyte differentiation and regulation of T cell activation(Figs. 2g-h); Next, we validated our above results using another data. The UMAP plots show the distribution of the different cell types in the tumor tissues (Fig. 2i). To understand the proportion of m^5C regulators in each cell types, we used the bar graph to show the percentage of m^5C regulator expression in each cell type. YTHDF2 YBX1, *DNMT1* and *ALYRF* have a high proportion of five cell types (Fig. 2j). We used the AddModuleScore function to define m^5Cscore, we found that epithelial and myeloid cells have higher scores, suggesting that m^5C-related regulators

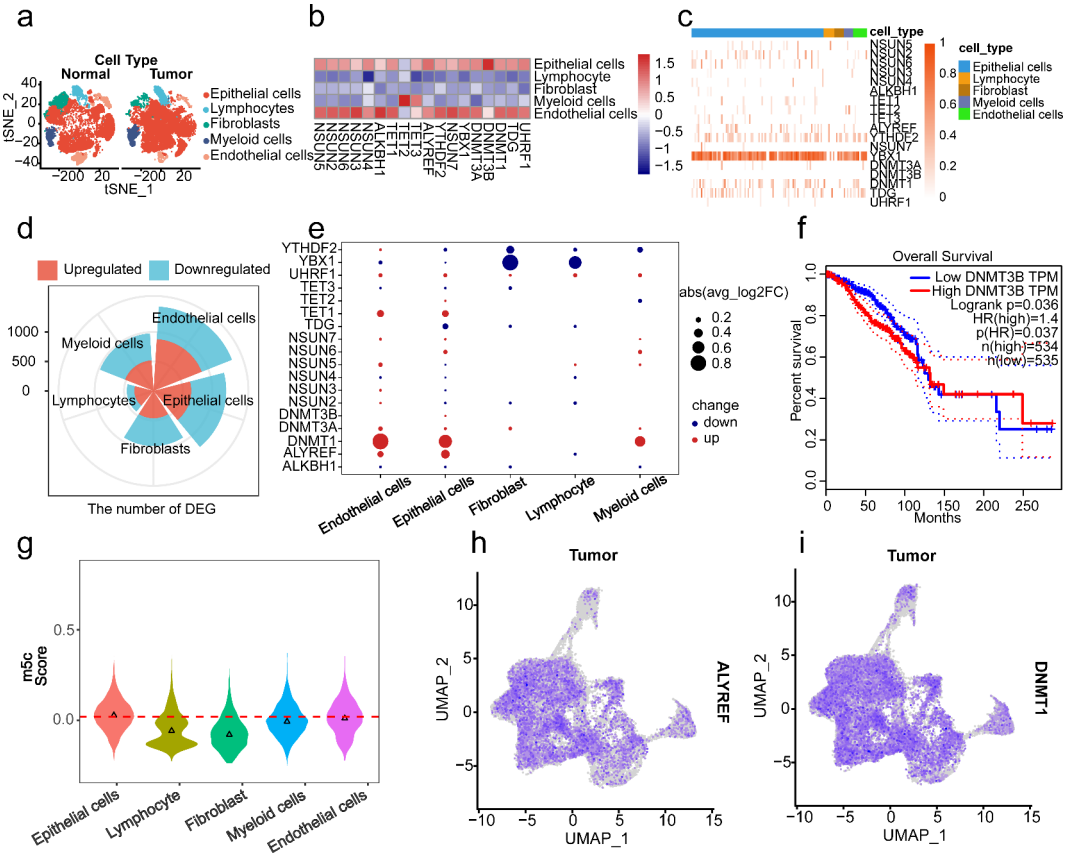

Fig. 1. Single-cell transcriptomic landscape of m^{5}C regulators regulating breast tissue key pathways in breast cancer (BC). (**a**) T-Distributed Stochastic Neighbor Embedding (TSNE) plot of normal cells and BC cells, colored by cell type. (**b**) Heatmap for differences in the expression of m^{5}C regulators in different cell types between normal and BC samples. Red, up-regulation; blue, down-regulation. (**c**) Heatmap of each m^{5}C regulators expression in each cell type. (**d**) Differential expression of genes (DEGs) in different cell types of BC patients compared with control samples. Red, up-regulation; blue, down-regulation. (**e**) Dot plot showing DEGs expression patterns of m^{5}C regulators of each cell types. Each dot represents a regulator, of which the color saturation indicates the average expression level, and the size indicates the percentage of cells expressing the regulator. (**f**) Kaplan-Meier survival analysis based on m^{5}C regulator expression. Red, high expression of m^{5}C regulator; blue, low expression of m^{5}C regulator. (**g**) Analysis of m^{5}Cscore for 5 cell types. The two T-SNE plots of m^{5}C regulators expression in BC samples. (**h**) *ALYREF*. (**i**) *DNMT1*.

play function more in epithelial and myeloid cells (Fig. 2k). WGCNA showed that some genes were closely related to *ALYREF* expression in epithelial cells, pathway enrichment analysis to verify the biological functions of genes of *ALYREF* related module in epithelial cells, like epidermis development, ERBB signaling pathway and canonical Wnt signaling pathway (Fig. 2l-m); And some genes were closely related to *DNMT1* expression in myeloid cells, pathway enrichment analysis to verify the biological functions of genes of *DNMT1* related module in myeloid cells, like endothelial cell development positive regulation of canonical Wnt signaling, positive regulation of mast cells and macrophages (Fig. 2n-o); Some genes were closely related to *DNMT3A* expression in endothelial cells, pathway enrichment analysis to verify the biological functions of genes of *DNMT3A* related module in endothelial cells, like chromatin remodeling, regulation of DNA damage response, endothelial cell development and positive regulation of canonical Wnt signaling (Figs S1f-g);

Overall, these results showed that the up-regulation of m^{5}C regulators play the essential role of the carcinogenesis and progression in epithelial cells and myeloid cells of BC.

Evaluation of m5C methylation modification patterns based on 17 m5C-related regulators

To ascertain the influence of m^{5}C methylation in BC, we obtained the gene expression data and full clinical annotations of 1089 BC patients from the TCGA database for analysis. We found compare to the normal groups, the expression in the mRNA levels of m^{5}C regulators were more higher in the tumor groups (Fig. 3a). Survival analysis showed that m^{5}C regulatory genes were closely related to prognosis (Fig S2). To further study the interaction between m^{5}C regulators, we depicted the comprehensive landscape of m^{5}C regulator interactions using the m^{5}C regulator network (Fig. 3b). These results illustrate that the dysregulation of m^{5}C RNA methylation regulators leads to different m^{5}C modification patterns, playing a vital role in the occurrence and development of BC. Based on this result, we used unsupervised clustering and principal component analysis

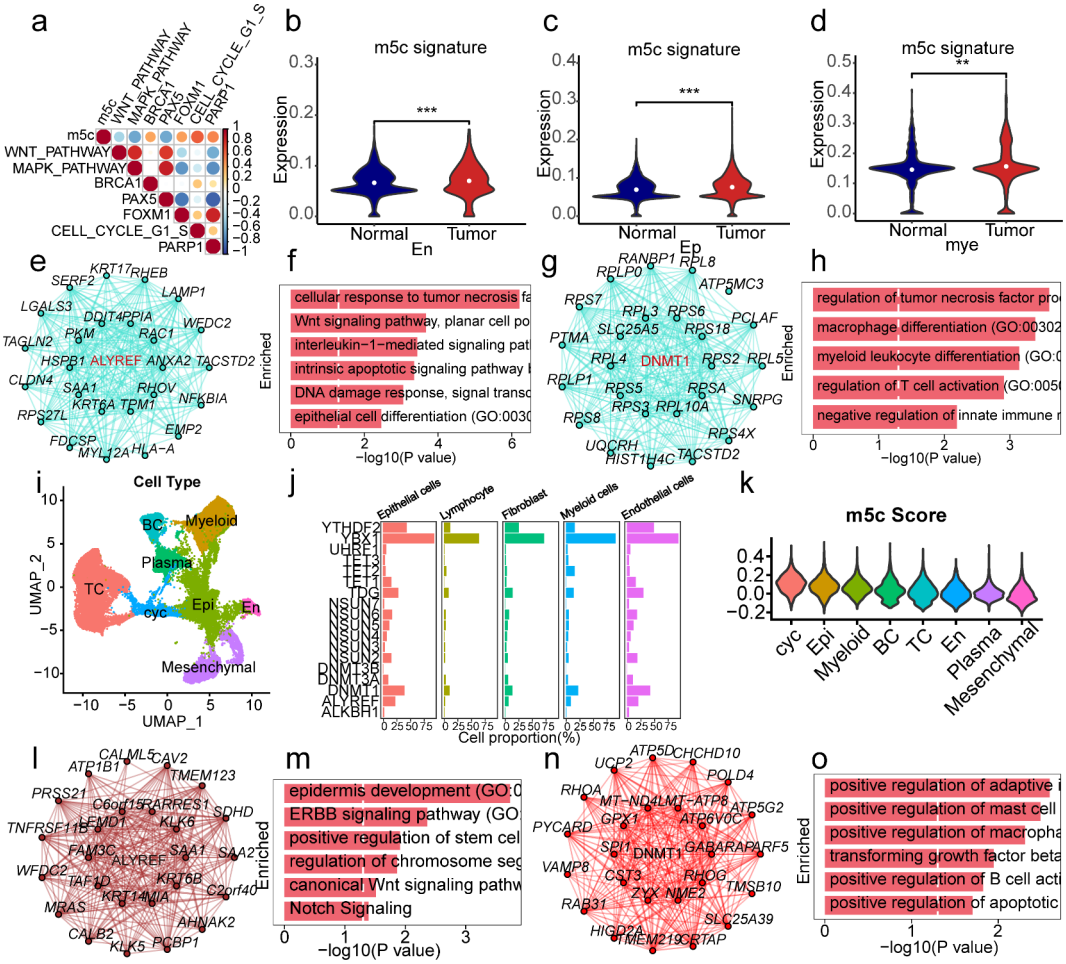

Fig. 2. Regulation of tumor-related pathways by the m5C regulators. (**a**) Correlation analysis was used to analyze the association between the m^{5}C regulators and tumor related pathways. Red, positive correlation; blue, negative correlation. (**b**–**d**) Differences in m^{5}C signature between tumor and normal groups in different cell types, including endothelial cells (En), epithelial cells (Ep), and myeloid cells (mye). (**e**) Networks of WGCNA module which included *ALYREF* in myeloid cells. (**f**) Functional enrichment of module which included *ALYREF* in myeloid cells. (**g**) Networks of WGCNA module which included *DNMT1* in myeloid cells. (**h**) Functional enrichment of module which included *DNMT1* in myeloid cells. (**i**) UMAP plots showing the expression of different cell types in other BC tissues. (**j**) Proportion of m^{5}C regulators expression in different cell types. (**k**) Analysis of m^{5}C score for several cell types. (**l**) Networks of WGCNA module which included *ALYREF* in myeloid cells. (**m**) Functional enrichment of module which included *ALYREF* in myeloid cells. (**n**) Networks of WGCNA module which included *DNMT1* in myeloid cells. (**o**) Functional enrichment of module which included *DNMT1* in myeloid cells.

to classify 1089 patients with BC based on their distinct modification patterns. We identified three different patterns, termed m^{5}C cluster A, m^{5}C cluster B, and m^{5}C cluster C, which were displayed by 567, 334, and 188 patients, respectively (Fig. 3c-d). Survival analysis indicated that the m^{5}C clusters were significantly related to prognosis in patients with BC, and patients in m^{5}C cluster B had the greatest survival advantage (Fig. 3e). To further explore the characteristics of m^{5}C modification patterns, we conducted unsupervised clustering of 14 m^{5}C regulators in the TCGA cohort, including demographic and clinical data such as age, sex, Tumor Node Metastasis (TNM) classification, clinical stage, and survival status. This unsupervised cluster analysis also identified three significantly different patterns of m^{5}C modification. The heatmap not only revealed the characteristics of different clinical traits in the three m^{5}C modification patterns but also their correlation with the expression of m^{5}C regulators. There was a significant difference in m^{5}C-related gene transcriptional profiles among the three m^{5}C modification patterns; most m^{5}C regulators were downregulated in m^{5}C cluster B (Fig. 3f). Based on these results, we performed GSVA to further compare the differences in enriched pathways among the three m^{5}C clusters. As shown in the heatmap, m^{5}C clusters A and C were significantly enriched in multiple pathways, such as mismatch repair, homologous recombination, nucleotide excision repair, spliceosome, and cell cycle, and the enrichment scores of these pathways were higher in m^{5}C cluster C than in m^{5}C cluster A. m^{5}C cluster B was enriched in the arachidonic acid metabolism pathway (Fig. 3g-i).

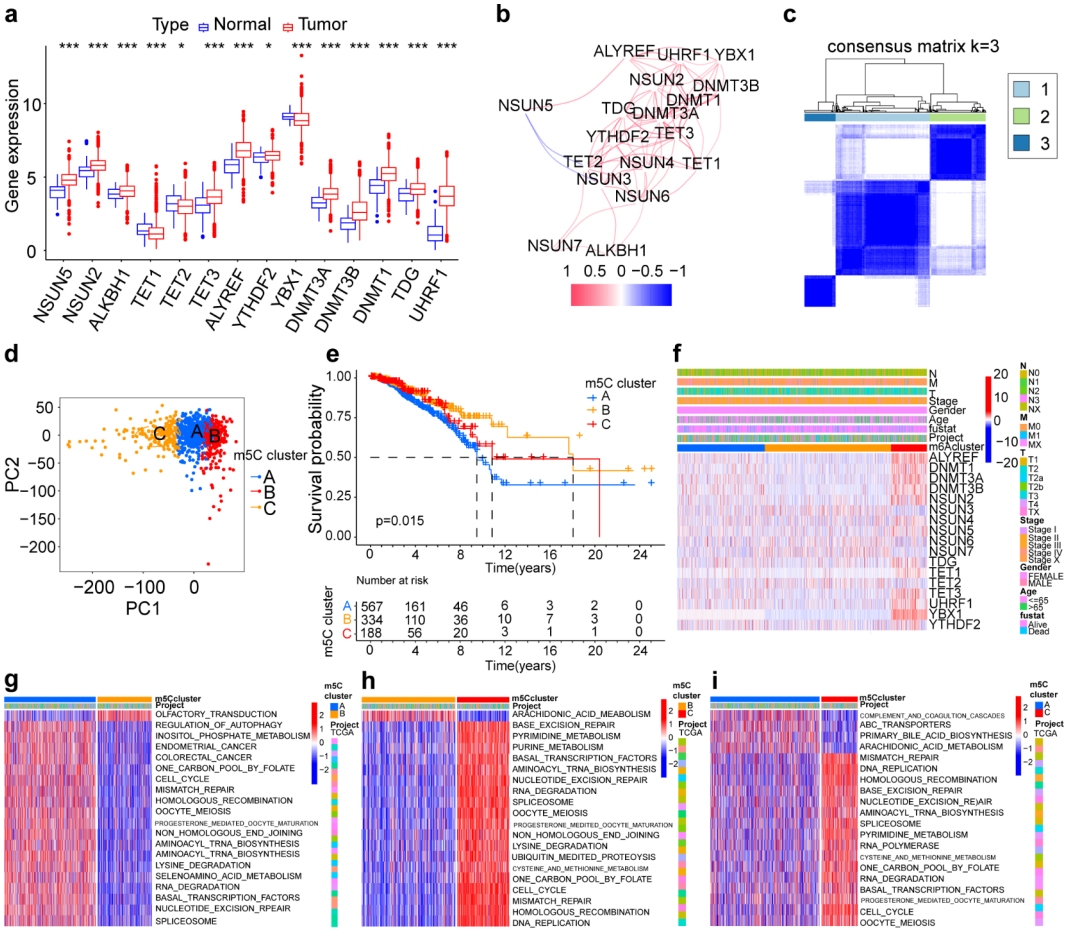

Fig. 3. Evaluation of m^{5}C methylation modification patterns. (**a**) Differential expression of m^{5}C regulators between breast cancer and normal breast tissues. Blue, normal tissue; red, tumor tissue. The lines in the boxes represent the median value, the bottoms and tops of the boxes represent the interquartile range, and the dots represent outliers. $^{***}P<0.001$, $^{**}P<0.01$, $^{*}P<0.05$. Differences among the three modification patterns were tested by one-way ANOVA. (**b**) The interaction between m^{5}C regulators in breast cancer. The lines connecting the m^{5}C regulators represent the interaction between them. Blue, negative correlation; red, positive correlation. (**c**) Three different m^{5}C modification subtypes were identified by unsupervised clustering based on m^{5}C regulators (m^{5}C cluster A, B, and C). (**d**) PCA derived from the m^{5}C clusters showed a difference between the three clusters. Blue, m^{5}C gene cluster A; yellow, m^{5}C gene cluster B; and red, m^{5}C gene cluster C. (**e**) Survival analysis based on the three m^{5}C clusters in 1089 patients with breast cancer in the TCGA-BRCA cohort ($P=0.015$, log-rank test). Blue, 567 patients in m^{5}C cluster A; yellow, 334 patients in m^{5}C cluster B; and red, 188 patients in m^{5}C cluster C. (**f**) Unsupervised clustering of 17 m^{5}C regulators in the TCGA-BRCA cohort identified a significant difference in the expression of regulators among the three modification patterns. The m^{5}C clusters, TCGA project, age, sex, TNM classification, clinical stage, and survival status were used as patient annotations. Red, high expression of regulators; blue, low expression of regulators. (**g**–**i**) GSVA enrichment analysis showing the activation states of biological pathways in distinct m^{5}C modification patterns. Red, activated pathways; blue, inhibited pathways. (**g**) m^{5}C cluster A compared with m^{5}C cluster B; (**h**) m^{5}C cluster A compared with m^{5}C cluster C; (**i**) m^{5}C cluster B compared with m^{5}C cluster C.

Construction of m5C gene signatures and functional annotation

To further investigate the potential biological behavior associated with each m^{5}C modification pattern, we identified 2312 m^{5}C phenotype-related DEGs (The gene list is presented in Table S2) among the m^{5}C cluster groups and conducted unsupervised clustering analyses based on these genes (Fig. 4a). Consistent with the clustering grouping of the m^{5}C modification pattern, the unsupervised clustering analysis and principal component analysis revealed three distinct m^{5}C modification genomic phenotypes. We named these three clusters m^{5}C gene cluster A, m^{5}C gene cluster B, and m^{5}C gene cluster C, which contained 326, 621, and 142 patients, respectively (Fig. 4b-c). Further survival analysis to investigate the correlation between the different m^{5}C genomic phenotypes and the prognosis of patients with BC revealed significant differences in prognosis among the three m^{5}C gene cluster groups. Patients in gene cluster B had an advantageous prognosis, whereas patients in gene cluster A had a poor prognosis (Fig. 4d). These results demonstrated that three m^{5}C methylation modification patterns are present in BC and are closely related to clinicopathological characteristics. To explore

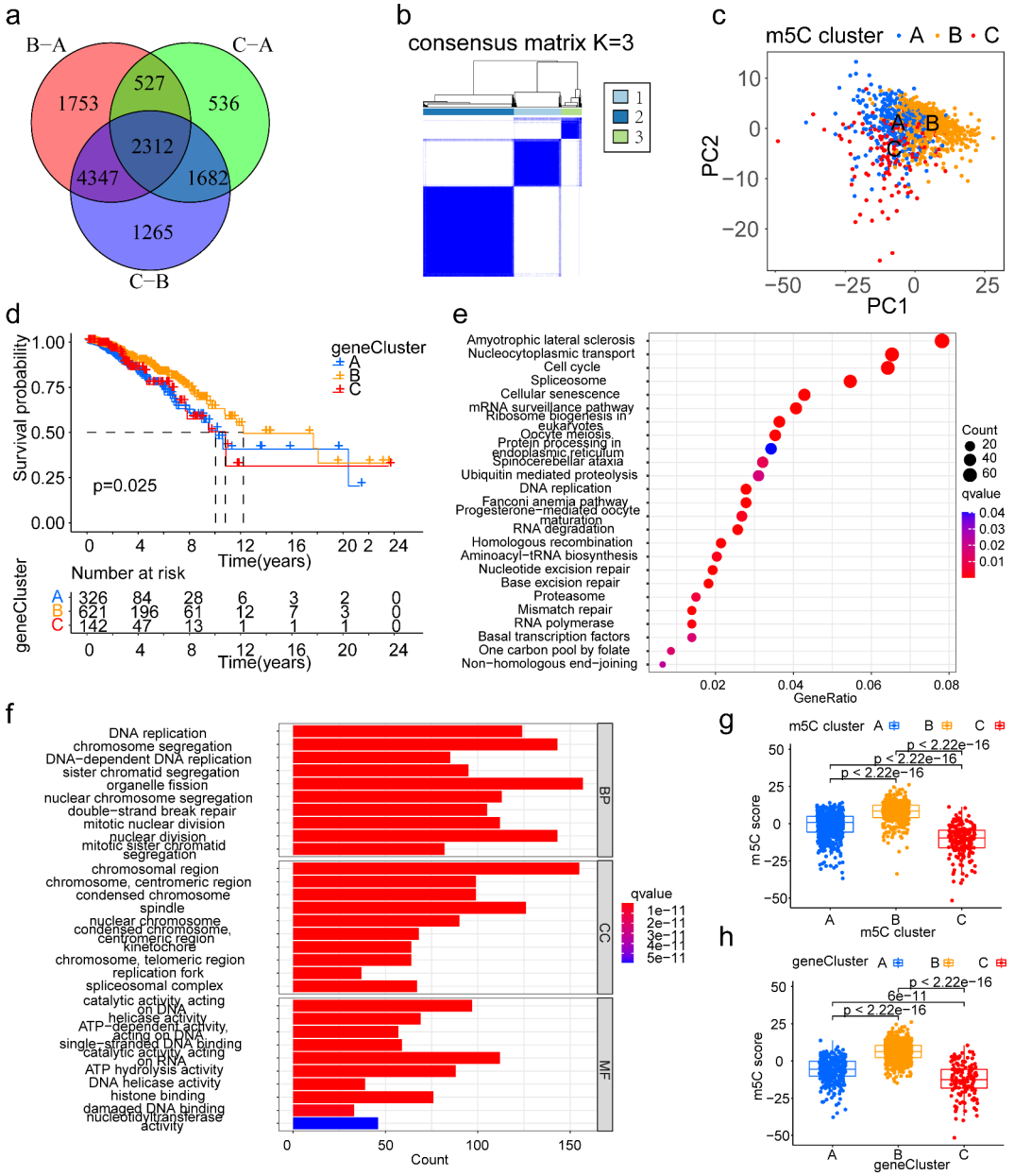

Fig. 4. Construction of m^{5}C gene signatures and functional annotation. (**a**) Overlapping m^{5}C phenotype-related DEGs in the three m^{5}C clusters. (**b**) Three different genomic subtypes identified by unsupervised clustering based on the overlapping m^{5}C phenotype-related DEGs. (**c**) PCA derived from the three m^{5}C gene clusters showed a difference between gene clusters. Blue, m^{5}C gene cluster A; yellow, m^{5}C gene cluster B; and red, m^{5}C gene cluster C. (**d**) Survival analysis based on the three m^{5}C gene clusters in 1089 patients from the TCGA-BRCA cohort ($P=0.025$, log-rank test). Blue, 326 patients with m^{5}C gene cluster A; yellow, 621 patients with m^{5}C gene cluster B; and red, 142 patients with m^{5}C gene cluster C. (**e**) GO functional enrichment analysis of 2312 overlapping m^{5}C phenotype-related DEGs. (**f**) KEGG pathway enrichment analysis of 2312 overlapping m^{5}C phenotype-related DEGs. (**g**) Differences in the m^{5}Cscore among the three m^{5}C clusters in the TCGA-BRCA cohort ($P<0.001$, Kruskal–Wallis test). Blue, m^{5}C cluster A; yellow, m^{5}C cluster B; and red, m^{5}C cluster C. (**h**) Differences in the m^{5}Cscore among the three m^{5}C gene clusters in the TCGA-BRCA cohort ($P<0.001$, Kruskal–Wallis test). Blue, m^{5}C gene cluster A; yellow, m^{5}C gene cluster B; and red, m^{5}C gene cluster C.

the relationship between biological behaviors and m^{5}C methylation modification in BC, we performed gene ontology functional enrichment analysis and Kyoto Encyclopedia of Genes and Genomes pathway enrichment analysis of 2312 m^{5}C phenotype-related DEGs (Fig. 4e-f). Gene ontology functional enrichment analysis showed that the m^{5}C-related genes were enriched in DNA replication, organelle fission, chromosomal region, chromosome centromeric region, replication fork, and helicase activity.

These analyses were based on the entire cohort. To further explore the heterogeneity and complexity of m^5C methylation modifications, we constructed a set of scoring models based on these phenotype-related genes to quantify the m^5C modification pattern of individual patients with BC and to predict treatment responses and prognoses, which we termed the m^5Cscore. The Kruskal–Wallis test revealed an association not only between the m^5C clusters and the m^5Cscore but also between the m^5C gene clusters and the m^5Cscore. Differential expression analysis of the m^5Cscore in m^5C clusters indicated that patients in m^5C cluster B had a significantly higher m^5Cscore than patients in other clusters, and patients in m^5C cluster C had the lowest median score (Fig. 4g). We also conducted differential expression analysis of m^5Cscore in the m^5C gene clusters and found that patients in m^5C gene cluster B had the highest median score and those in m^5C gene cluster C had the lowest median score (Fig. 4h). Based on the correlation of the m^5Cscore with different m^5C methylation modification patterns and m^5C modification genomic phenotypes, we determined the optimal cutoff value and divided our patients with BC into high and low m^5Cscore groups.

The m5Cscore activates immune infiltration

To investigate the role of m^5C methylation in immune cell infiltration in the TME, we first compared immune cell characteristics among different m^5C clusters. Among the three m^5C clusters, there was no difference in the infiltration of activated dendritic cells, gamma delta T cells, immature B cells, or Treg cells; however, the three types of m^5C clusters were significantly correlated with infiltration of the other 19 types of immune cells. m^5C cluster B was remarkably rich in innate immune cell infiltration, including activated B cells, activated $CD8^+$ T cells, eosinophils, myeloid-derived suppressor cells, macrophages, mast cells, monocytes, NK cells, and neutrophils (Fig S3a). Similarly, the infiltration of most immune cell populations was significantly different among the three m^5C gene clusters, and immune cells, such as activated B cells, activated $CD8^+$ T cells, eosinophils, macrophages, mast cells, monocytes, NK cells, and neutrophils, were enriched in m^5C gene cluster B (Fig S3b). Crosstalk among m^5C regulators may create different m^5C modification patterns and different m^5C modification genomic phenotypes, thereby playing a critical role in the formation of different cell-infiltrating characteristics in BC. Therefore, we analyzed the immune cell and immune function scores between the high and low m^5Cscore groups, noting remarkable differences. The high m^5Cscore group generally had higher immune cell scores, including for B cells, $CD8^+$ T cells, cytolytic activity, HLA, infiltrating dendritic cells (iDCs), mast cells, NK cells, T cell co-stimulation, T helper cells, tumor-infiltrating lymphocytes, and type II interferon (IFN) response (Fig S3c).

To investigate the relationship between the m^5Cscore and different infiltrating immune cell characteristics, we analyzed the correlation between the m^5Cscore and classical infiltrating immune cell populations. The m^5Cscore was positively correlated with eosinophils, mast cells, NK cells, and plasmacytoid dendritic cells and negatively correlated with activated $CD4^+$ T cells (Fig S3d). These results indicate the m^5Cscore can not only better evaluate the m^5C modification patterns of individual tumors, but also further evaluate the immune infiltration characteristics.

Considering the role of immune cell infiltration in tumor occurrence and development and its prognostic impact, we conducted survival analyses to assess the value of the m^5Cscore in predicting patient outcomes. The Kaplan–Meier curves showed that patients with a high m^5Cscore had significantly increased survival than those with a low m^5Cscore (Fig S3e). Subsequently, we applied the constructed signature to two independent BC cohorts from the Gene Expression Omnibus database (GSE7390 and GSE103091) to assess the stability of the m^5C gene signature. The m^5Cscore of each patient in the testing dataset was acquired, and the optimal cutoff point was identified. Thus, the 305 patients with BC in the testing dataset were also divided into two groups: 138 in the high m^5Cscore group and 167 in the low m^5Cscore group. Survival analysis yielded a similar result to that obtained in the TCGA cohort; patients with a high m^5Cscore had a significantly better prognosis than those with a low m^5Cscore (Fig S3f). Next, we analyzed the correlation between survival and the ssGSEA scores of classical immune cells. As indicated by the heatmap, there was no significant correlation between survival status and the ssGSEA score of immune cells other than that of plasma cells (Fig S3g). These results suggest that although a single type of immune cell is not significantly associated with prognosis, the different characteristics of immune cell infiltration formed by the interaction between multiple different immune cell types can have a crucial impact on prognosis. We further analyzed the correlation between the m^5Cscore and the infiltration abundances of immune cells. We found that the m^5Cscore was positively correlated with the abundance of activated NK cells (Fig S3h). Subsequently, we measured mRNA expression to further explore the relationship between the m^5Cscore and the stem-like properties of BC tumor cells (Fig S3i). The m^5Cscore was negatively associated with expression of stem cell mRNAs in a statistically significant manner, indicating that a higher m^5Cscore is closely correlated with lower tumor stem cell activity and a higher degree of tumor differentiation.

We next examined the expression profiles of immune checkpoint genes and the correlation between these genes and the m^5Cscore. Although most immune checkpoint genes had low expression in the three m^5C clusters, the statistical difference among clusters was significant, and the same was true for the m^5C gene clusters. The expression of CD44 was lower in m^5C cluster B than in the other m^5C clusters, and among the m^5C gene clusters, the expression of VTCN1 and TNFRSF18 were higher in m^5C gene cluster B than in the other m^5C gene clusters (fig S4a-b). Correlation analysis of the m^5Cscore and immune checkpoint genes showed that most correlations between them were statistically significant, and the m^5Cscore was positively associated with TNFRSF14, TNFRSF4, CD27, TMIGD2, TNFRSF25, CD40LG, and CD200 and negatively associated with CD80 and ICOSLG (Fig S4c). The analysis of HLA gene expression indicated that the expression of HLA genes, including HLA-E, HLA-C, HLA-J, HLA-DMA, HLA-DRB1, HLA-DOA, HLA-DPB1, and HLA-DRB6, was significantly different among the m^5C clusters, and HLA-E, HLA-C, HLA-DRB1, and HLA-DPB1 were highly expressed in m^5C cluster B, whereas HLA-DRA was highly expressed in m^5C cluster A. Similarly, the expression of most HLA genes significantly differed among the m^5C gene clusters, and HLA-E, HLA-C, HLA-A, HLA-DRB1, HLA-DRB5, and HLA-DPB1 were markedly higher in m^5C gene cluster B than in the other gene

clusters (Fig S4d-e). We also noted a positive association between the m^{5}Cscore and the expression of HLA-DPB2, HLA-C, HLA-J, HLA-DQB1, HLA-DQB2, HLA-DMA, HLA-DRB1, HLA-H, HLA-DRB5, HLA-DPB1, HLA-DRB6, HLA-L, HLA-DMB, and HLA-DPA1 (Fig S4f). In addition, the expression of interleukin (IL)-4 and IL-33 were significantly different among the m^{5}C clusters, and the expression of IL-4, TSLP, and IL-33 were significantly different among the m^{5}C gene clusters. IL-33 levels were significantly higher in m^{5}C cluster B and m^{5}C gene cluster B than in the other clusters and gene clusters, respectively (Fig S4g-h), and the expression of IL-5, TSLP, and IL-33 were positively correlated with the m^{5}Cscore in a statistically significant manner (Fig S4i). Furthermore, we performed GSVA enrichment analysis to compare differences in the activation states of immune functions and immune cells between distinct m^{5}Cscore groups. As shown in the heatmap, the high m^{5}Cscore group had significant enrichment in multiple immune pathways, such as mast cells and type II IFN response (Fig S4j).

To explore the correlation between the m^{5}Cscore and the proportion of immune and stromal cells in the TME and further examine the differences in survival between the two m^{5}Cscore groups, we analyzed the stromal, immune, and ESTIMATE scores (Fig S4k). Patients with a high m^{5}Cscore had significantly higher stromal, immune, and ESTIMATE scores than patients with a low m^{5}Cscore. Therefore, compared with patients with BC that had a low m^{5}Cscore, those with a high m^{5}Cscore had tumors with more abundant immune and stromal components and had stronger immune function and better prognosis.

The negative correlation between the m5Cscore and response to pharmacotherapy

ICB, as represented by PD-1 and CTLA-4 inhibitors, has caused a breakthrough in tumor immunotherapy. On this basis, we used TIDE to predict the therapeutic effect of ICB based on pretreatment tumor profiles and to establish an indirect connection between the m^{5}Cscore and the immune response. The TIDE score was significantly higher in the high m^{5}Cscore group than in the low m^{5}Cscore group, indicating that tumors in patients in the high m^{5}Cscore group were more likely to induce immune escape and that these patients would have a lower therapeutic response to ICB (Fig. 5a). Consistent with this result, prognosis and survival analysis indicated patients with a high TIDE score had a distinctly better prognosis than those with a low TIDE score (Fig. 5b). Survival analysis based on both the TIDE score and the m^{5}Cscore showed that patients with a low TIDE score and a low m^{5}Cscore had the worst prognosis, whereas patients with a high m^{5}Cscore and a low TIDE score had the best prognosis (Fig. 5c). Moreover, we also found that the low m^{5}Cscore group had a significantly higher response rate to ICB, and the Area Under the Curve (AUC) value illustrated that the TIDE model has acceptable performance in predicting the therapeutic response to ICB in patients with BC (Fig. 5D). We further analyzed targeted immune dysfunction and exclusion. Consistent with the TIDE score distribution, patients with a high m^{5}Cscore were more likely to have immune dysfunction (Fig. 5e). Survival analysis combining immune dysfunction with the m^{5}Cscore showed that patients with a high m^{5}Cscore and high immune dysfunction had increased survival (Fig. 5f). Moreover, we also found similar results in the analyses of immune exclusion (Fig. 5g-h). Therefore, regardless of the TIDE, immune dysfunction, and immune exclusion scores, patients in the high m^{5}Cscore group consistently had increased survival than those in the low m^{5}Cscore group, indicating the value of m^{5}Cscore in predicting the therapeutic response to ICB.

The efficacy of doxorubicin-based chemotherapy as a first-line therapy after BC surgery has been widely demonstrated. However, researchers continue to investigate novel drugs for BC. Considering the differences in survival in the different m^{5}Cscore groups, we analyzed the ability of the m^{5}Cscore to predict responses to several different novel chemotherapeutic drugs. The results revealed higher estimated half maximal inhibitory concentrations for chemotherapeutic drugs in the low m^{5}Cscore group, and the m^{5}Cscore was negatively correlated with the therapeutic effect of these drugs (Fig. 5i-t, S5), indicating that patients with BC that have a low m^{5}Cscore have better therapeutic response to chemotherapeutic drugs than patients in the high m^{5}Cscore group.

The positive correlation between the good clinicopathological characteristics and the m5Cscore

Based on these results, we performed survival analysis to explore the distribution of survival status between patients with high and low m^{5}Cscores. Patients who were still alive had a significantly higher m^{5}Cscore than those who died (Fig. 6a), and the low m^{5}Cscore group had a decreased survival (Fig. 6b). patients with N0 and N2 disease had a lower m^{5}Cscore than those with N1 and N3 disease (Fig. 6c). We also assessed the correlation between the m^{5}Cscore and human epidermal growth factor receptor 2 (HER2) status, TNM stage, and clinical stage (Figs S6A-D). We then analyzed the correlation of the m^{5}Cscore with clinicopathological characteristics. The m^{5}Cscore differed based on T, N, and clinical stage (Figs S6 E-F); Patients in the low m^{5}Cscore group were more likely to have HER2 + disease, whereas those with a high m^{5}Cscore were more likely to have HER2- disease (Fig. 6d).

Because of these findings, we used the Sankey diagram to show the flow of m^{5}Cscore fraction construction and the changes in individual patient attributes, such as age, TNM classification, clinical stage, and survival status, based on the m^{5}Cscore (Figs S6I-L). m^{5}C cluster and m^{5}C gene cluster were both associated with the m^{5}Cscore. Patients in m^{5}C cluster B and m^{5}C cluster A tended to be classified into m^{5}C gene cluster B, and almost all patients in m^{5}C gene cluster B were classified into the high m^{5}Cscore group, which had better prognosis than the low m^{5}Cscore group. Conversely, m^{5}C gene cluster C was associated with a low m^{5}Cscore and poor prognosis (Fig. 6e). The high and low m^{5}Cscore groups had similar proportions of patients < 65 and ≥ 65 years (Fig. 6f). To further assess the prognostic value of the m^{5}Cscore in different subgroups, we performed Kaplan–Meier analyses (Fig. 6g-l and S6g-h). We found that the m^{5}Cscore exhibited prognostic power in various subgroups; among women, those younger than 65 years, and those with M0, N1, T2, or stage III disease. Additionally, the high m^{5}Cscore group had a better prognosis than the low m^{5}Cscore group.

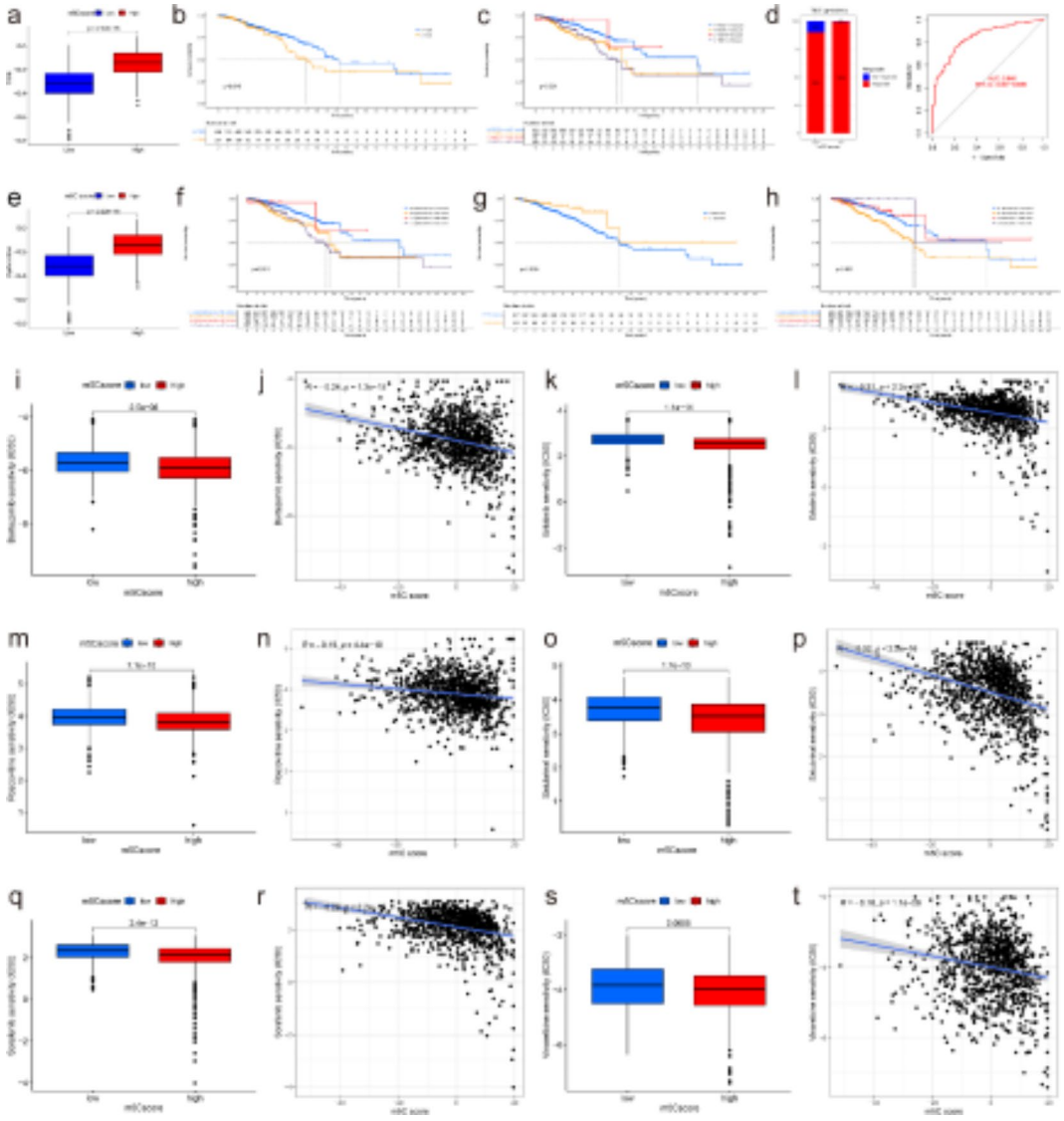

Fig. 5. Association between the m^{5}Cscore and response to pharmacotherapy. (**a**) The relative distribution of TIDE scores was compared between the low and high m^{5}Cscore groups. The lines in the boxes represent the median value, the bottoms and tops of the boxes represent the interquartile range, and the dots represent outliers. Blue, low m^{5}Cscore group; red, high m^{5}Cscore group. (**b**) Survival analysis for 858 patients with a high TIDE score and 231 with a low TIDE score ($P < 0.05$, log-rank test). (**c**) Survival curves of TIDE scores combined with m^{5}C scores ($P < 0.05$, log-rank test). (**d**) Comparisons of the proportions of non-responders and responders to ICB between the low and high m^{5}Cscore groups; the ROC curves of the TIDE score model in patients with BC (AUC: 0.846, 95% CI: 0.801–0.888). Blue, non-responder groups; red, responder groups. (**e**) Relative distribution of immune dysfunction scores between the low and high m^{5}Cscore groups ($P < 0.001$). Blue, low m^{5}Cscore group; red, high m^{5}Cscore group. (**f**) Survival analysis stratified by both m^{5}Cscore and immune dysfunction scores ($P = 0.001$, log-rank test). (**g**) survival analysis for 977 patients with a high immune exclusion and 112 with a low immune exclusion ($P < 0.05$, log-rank test). (**h**) Survival analyses stratified by both the m^{5}Cscore and immune exclusion ($P < 0.001$, log-rank test). (**i**–**t**) Predicted response of patients to six chemotherapeutic drugs based on the m^{5}Cscore. (**I**, **j**) bortezomib; (**k**, **l**) erlotinib; (**m**, **n**) roscovitine; (**o**, **p**) salubrinal; (**q**, **r**) sorafenib; (**s**, **t**) vinorelbine. Blue, low m^{5}Cscore group; red, high m^{5}Cscore group.

The m5C related genes expressions are generally increased in BC tissue

To further validate the expression of m^{5}C related genes in BC tissues, we utilized the transcriptomic information of BC patients and normal breast patients from public databases, founding the expression levels of *DNMT1*, *DNMT3B*, *TET3*, and *UHRF1* in the tumor group were higher than that in normal group (Fig. 7a-d). Next, mmunohistochemistry (IHC) for three normal mammary gland tissues and three paracancerous tissues was also performed using different m^{5}C related gene antibodies. The results showed that compare to the paracancerous tissues, *ALYREF*, *DNMT1* and *DNMT3a* showed stronger expression on the cytoplasm and nucleus in tumor tissues (Fig. 7e-m and S7a-i). The bar chart showed that in the tumor tissues, the percentage contribution of

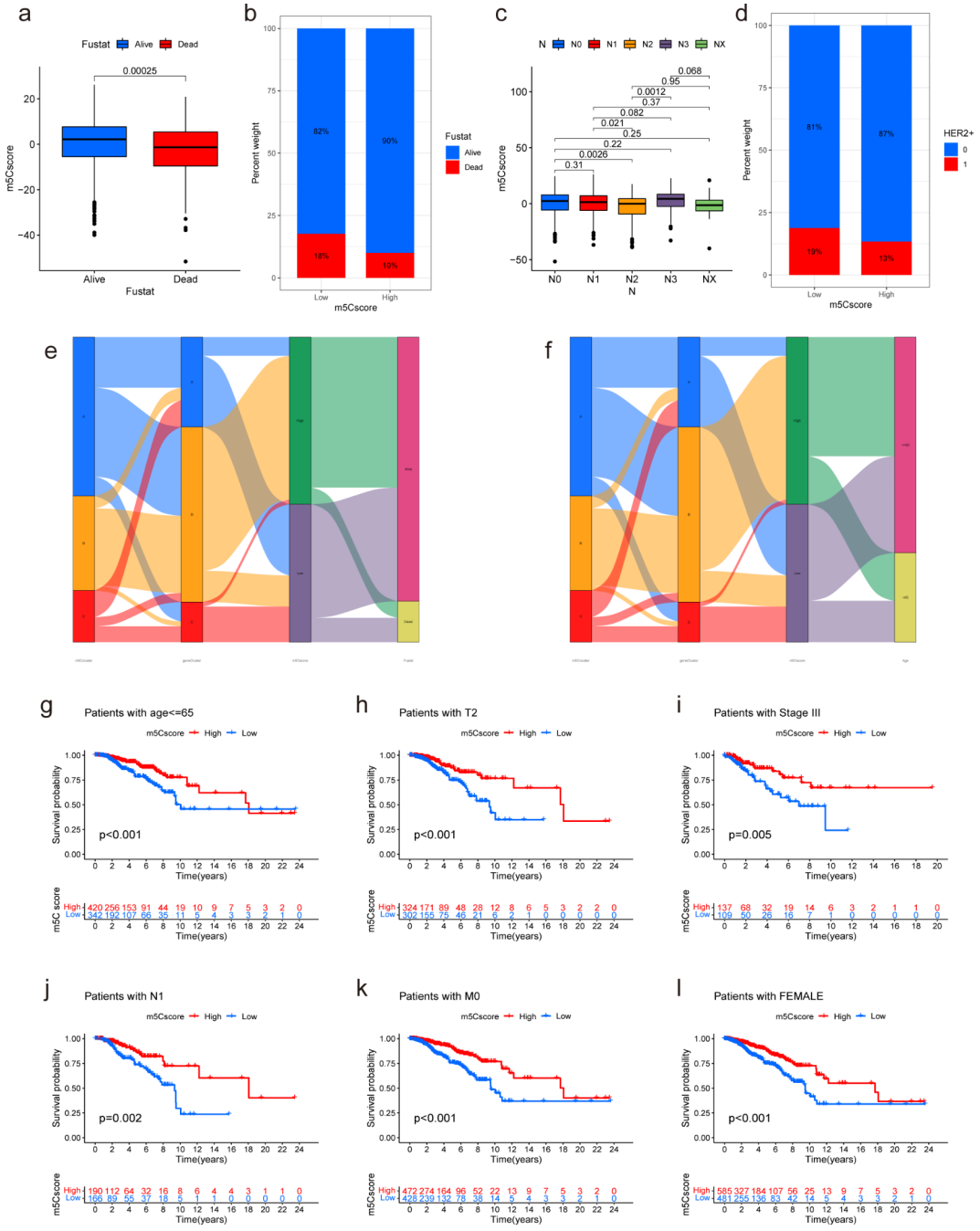

Fig. 6. Correlation between clinicopathological characteristics and the m^{5}Cscore. (**a**) m^{5}Cscore based on survival status ($P < 0.001$). The lines in the boxes represent the median value, the bottoms and tops of the boxes represent the interquartile range, and the dots represent outliers. Blue, living patients; red, deceased patients. (**b**) The proportions of living and dead patients with BC in the low and high m^{5}Cscore groups. In the low m^{5}Cscore group, 82% of patients were alive and 18% were dead, and in the high m^{5}Cscore group, 90% of patients were alive and 10% were dead. Blue, living patients; red, deceased patients. (**c**) The m^{5}Cscore based on N stage. The Kruskal–Wallis test was used to compare the statistical difference between five N stage groups. The lines in the boxes represent the median value, the bottoms and tops of the boxes represent the interquartile range, and the dots represent outliers. Blue, N_0 stage group; red, N_1 stage group; yellow, N_2 stage group; purple, N_3 stage group; green, N_x stage group. (**d**) HER2 expression status in the low and high m^{5}Cscore groups. In the low m^{5}Cscore group, 81% and 19% of patients had HER2 + and HER2- disease, respectively; in the high m^{5}Cscore group, 87% and 13% of patients had HER2+ and HER2− disease, respectively. Blue, patients with HER2− disease; red, patients with HER2+ disease. (**e**) Sankey diagram showing the flow of m^{5}C cluster, m^{5}C gene cluster, m^{5}Cscore, and survival status. (**f**) Sankey diagram showing the flow of m^{5}C cluster, m^{5}C gene cluster, m^{5}Cscore, and age. (**g**–**l**) Kaplan–Meier survival analysis based on the m^{5}Cscore in subgroups with different clinical characteristics. Red, high m^{5}Cscore group; blue, low m^{5}Cscore group. (**g**) patients ≤ 65 years; (**h**) patients with T_2 stage; (**i**) patients with stage III disease; (j) patients with N_1 disease; (**k**) patients with M_0 disease; (**l**) Women.

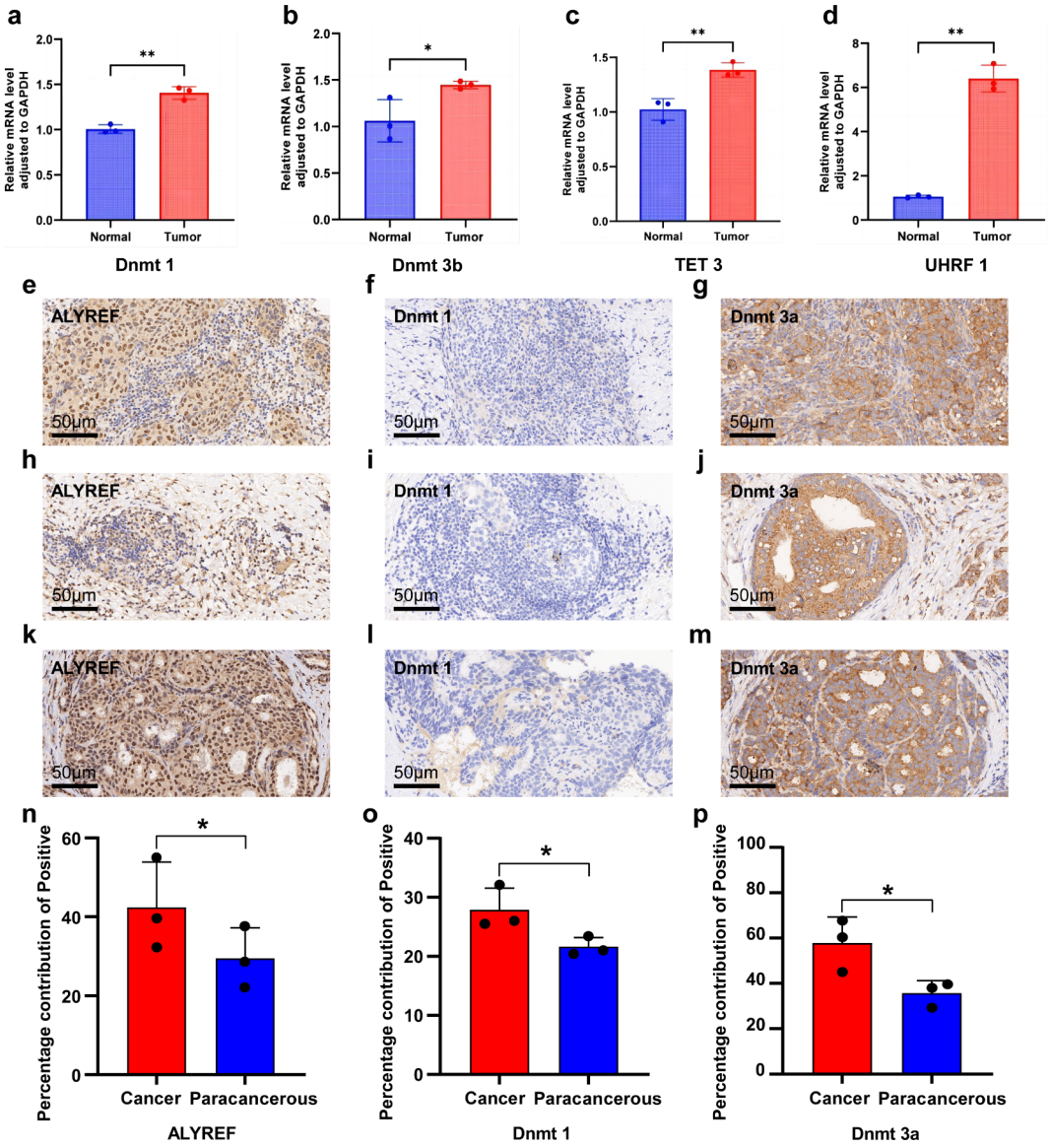

Fig. 7. The expression of m5C-related genes was verified using RT-qPCR and IHC. (**a**–**d**) Differential expression of m^{5}C-related genes in normal and tumor groups. (**e**–**m**) IHC of the *ALYREF*, *DNMT1* and *DNMT3a* in tumor tissue. (**n**–**p**) Percentage of positive staining for m^{5}C-related genes between the paracancerous and tumor groups.

positive for m^{5}C related gene were higher than that in the paracancerous group (Fig. 7n-p). These results showed that m^{5}C related genes expression levels gradually increased from normal tissues to tumor tissues, which further verified that m^{5}C related genes can act as a pro-oncogenic gene to regulate the occurrence and development of BC.

Discussion

BC is the most common malignant disease in women, and owing to the lack of cost-effective therapies, it has become one of the most severe disease burdens globally. Triple-negative breast cancer is the type with the highest malignancy and the worst prognosis. It is challenging to predict the treatment effect and prognosis of different methods. The estrogen receptor, progesterone receptor, and HER2 expression patterns in different subtypes of BC represent a predictive method for the therapeutic guidance of BC. However, the existing classification models based on these molecules cannot accurately reflect tumor heterogeneity or evaluate the prognosis of BC. Therefore, a comprehensive understanding of the molecular mechanisms of BC development and progression is urgently needed to further explore more effective therapeutic targets and prognostic biomarkers.

The m^{5}C modification is one of the most important RNA modifications in eukaryotes, and it plays an indispensable role in posttranscriptional regulation, which is closely related to tumor formation, maintenance and progression[5,17]. m^{5}C modification is involved in bladder cancer progression by modulating mRNA stability[18], and

some studies have confirmed that m^5C is involved in the progression of hepatocellular carcinoma[19]. Furthermore, recent studies have indicated that m^5C modification is associated with the infiltration of multiple immune cells, including $CD8^+$ T cells and neutrophils, regulating their behavior[20]. In addition, RNA modification regulators have the potential to act as biomarkers for the diagnosis of cancer and in prognostic monitoring[21,22]. For example, high expression of the m^5C writer *NSUN1* has been identified as a prognostic marker for non-small cell lung cancer[23]. However, as most previous studies have focused on a single m^5C regulatory factor, for example, In oral squamous cell carcinoma, inhibition of DNMT1 expression by inhibitors increases tumor-infiltrating T cells and subsequently blocks tumor growth[24]. The characteristics of immune infiltration mediated by multiple m^5C regulatory factors are unclear, and comprehensive analysis of the prognostic value and functional annotation of m^5C regulators in BC are still lacking. Hence, identifying the function of m^5C modification patterns in immune cell infiltration is fundamental for improving our understanding of the interaction between m^5C methylation and the antitumor immune response and facilitating the advancement of personalized treatments for patients with BC.

In this study, we performed single-cell RNA sequencing from 5 normal and 5 BC samples and successfully characterized samples into 5 cell clusters: epithelial cells, fibroblasts, endothelial cells, lymphocytes and myeloid cells.

We have found that m^5C-related regulators can modulate the expression of tumor-associated pathways, thereby affecting the formation and development of BC. For example, m^5C can inhibit the PAX5 pathway, reducing the production of chemokines, which in turn inhibits the infiltration of T cells in the TME and weakens the ability to restrict tumor cell metastasis[25]. m^5C can also activate the FOXM1 pathway. In tumor cells, the overexpression of FOXM1 may lead to errors during the G2/M phase of mitosis, increasing chromosomal instability and thereby accelerating the proliferation of tumor cells[26]. Meanwhile, the up-regulation of m^5C regulators also play the essential role of the carcinogenesis and progression in myeloid cells and epithelial cells of BC. In myeloid cells, the biological function of *DNMT1* related module has been demonstrated to be enriched in neutrophil activation and T cell proliferation. Consistent with the previous study that targeting *DNMT1* in breast tumors can upregulate major histocompatibility class-I mediated antigen presentation and tip the balance at equilibrium to elicit a $CD8^+$ T cell response which promotes tumor regression and anti-tumor immunity[27]. In endothelial cells, *DNMT3A*'s related module's biological function has been verified correlated with ferroptosis which is a new form of programmed cell death caused by the accumulation of lipid-based reactive oxygen species, and is closely related to immune response of BC through by PPAR signaling pathway and IL-17 signaling pathway[28]; In the epithelial cells, *ALYREF* related module's biological functions was verified enriched in Wnt signaling pathway and DNA damage response, when Wnt ligands bind to Frizzled receptors on the cell surface, they can lead to the stabilization and accumulation of β -catenin and affect cell proliferation, differentiation[29]. When wnt is dysregulated, it can cause uncontrolled cell proliferation and tumor formation.

At the same time, we identified three distinct patterns of m^5C modification, termed m^5C cluster A, m^5C cluster B, and m^5C cluster C, based on the expression of 17 regulatory factors associated with m^5C modification. These three patterns were associated with significantly different immune infiltration characteristics, functional characteristics, and prognoses. Furthermore, we demonstrated that the differentially expressed mRNAs between distinct m^5C modification patterns were important m^5C-related signature genes that were significantly associated with m^5C- and immune-related biological pathways in BC. Subsequently, similar to the clustering results of m^5C modification phenotypes, we also identified three genomic subtypes based on m^5C-related signature genes, which were also significantly correlated with immune cell infiltration and BC prognoses, demonstrating that m^5C modification is of great significance in shaping the TME landscape. Based on these results and considering the heterogeneity of m^5C modification, we established scoring systems to evaluate the m^5C modification pattern of individual patients with BC, that is, the m^5C gene signature and functional annotation, which we called the m^5Cscore, to further quantify the m^5C modification patterns of individual tumors. We investigated not only the association between the m^5Cscore and immune infiltration characteristics but also the association between the m^5Cscore and clinicopathological characteristics. We also predicted the pharmacotherapy response based on the m^5Cscore.

The current study investigated the functions and pathways of 2312 m^5C phenotype-related DEGs. *NSUN2* is an important methyltransferase for m^5C modification in tRNA, abundant noncoding RNAs, and a small subset of mRNAs and can promote cell growth by regulating cyclin-dependent kinase 1 expression in a cell cycle-dependent manner[30–32]. Moreover, IGF2BP3, a newly reported reader of RNA methylation, is associated with DNA replication, and knockdown of IGF2BP3 significantly represses cell proliferation and the percentage of cells in S phase[33]. We showed that m^5C phenotype-related DEGs were highly enriched in cancer-related functions and pathways, such as DNA replication, spliceosome, and cell cycle signaling. In addition, there were three distinct m^5C methylation modification patterns in BC, and both the expression of m^5C-related genes and their enriched pathways were different. Significantly, m^5C cluster B, in which most m^5C regulators were downregulated, lacked enrichment of classical cancer-related pathways, such as DNA replication and the cell cycle, and was associated with better prognosis than other clusters. Analysis of the m^5C gene clusters led to similar results. In m^5C gene cluster B, most m^5C-related genes were downregulated, and patients exhibited better survival. These results indicate that m^5C regulators play an important role in the occurrence, development, and prognosis of BC, have a prognostic predictive value, and could be novel prognostic indicators for patients with BC.

We also established the m^5Cscore, a scoring system that could individually quantify the m^5C modification pattern in patients with BC, and investigated its association with immune infiltration characteristics. Immune cell infiltration has become a new research focus because immune cells are a major component of the TME, and many studies have reported the critical role of the TME in tumor progression, response to therapeutics, and prognosis of BC. $CD8^+$ T cells are important immune cells in the antitumor response. Numerous studies have reported that $CD8^+$ T cells can directly mediate tumor lysis in vitro, and their increased abundance is

closely associated with better survival outcomes in patients with BC[34,35]. B cells are also important immune cells in the antitumor response and have been reported to generate humoral immune responses and promote effective antitumor immunity at the BC tumor site, thereby improving the clinical outcomes[36]. NK cells also play an important role in antitumor immunity. These cells can autonomously kill target cells and serve as the main innate immune effector cells against cancer[37–39]. Our analyses showed that m^5C cluster B and m^5C gene cluster B had abundant immune cell infiltration. More importantly, the m^5Cscore was significantly correlated with the infiltration of most immune cells and was positively associated with the infiltration of $CD8^+$ T cells, B cells, and NK cells. In addition, we analyzed the association between classical immune genes, such as immune checkpoint molecules, HLA family genes, and IL family genes, and the m^5Cscore. Our data revealed that the m^5Cscore was significantly correlated with the expression of multiple immune genes, including TNFRSF14, HLA-DPB1, IL-5 and IL-33, and their expression among m^5C clusters and m^5C gene clusters were significantly different. Because the roles of these immune genes in BC have not been reported, these results provide new directions for further research. Furthermore, previous evidence demonstrated that activated IFN signaling plays an important role in antitumor immunity by modulating immune surveillance, which could regulate either tumor cells to exert direct antitumor effects or immune cells to exert indirect antitumor effects[40,41]. We found a significant positive association between the m^5Cscore and type II IFN response, which could be one of the reasons for the better prognosis of patients with a high m^5Cscore. These findings suggest that patients with BC exhibiting a higher m^5Cscore have more abundant antitumor immune cell infiltration, higher expression of immune genes, stronger antitumor immunity, and a better prognosis. Overall, the m^5Cscore is a potential tool for comprehensive assessment of individual tumor m^5C modification patterns and further determination of TME infiltration patterns.

We further investigated the relationship between the m^5Cscore and pharmacotherapy response. The m^5Cscore was significantly correlated with predictors of the immune response, such as the TIDE score, indicating that m^5C modification impacts the therapeutic effect of immunotherapy and that it can be used to improve personalized treatment of patients with BC. Additionally, a higher m^5Cscore was significantly correlated to higher TIDE, higher immune dysfunction scores, and immune exclusion scores; immune dysfunction and immune exclusion may have been the key reasons for a higher TIDE score. Patients with a higher TIDE score tend to have greater immune dysfunction and immune exclusion and decreased ability to kill cancer cells, which could explain why they have a worse response rate to ICB[42]. Conversely, although patients with a low m^5Cscore had lower TIDE scores and a worse prognosis, they were predicted to be more likely to benefit from ICB treatment. Therefore, patients with a low m^5Cscore could have prolonged survival through ICB therapy. Moreover, the m^5Cscore was also able to predict the response to pharmacotherapy. These results suggest the m^5Cscore could be used to develop individualized treatment plans for patients with BC. Multiple anti-PD-1/PD-L1 antibodies such as pembrolizumab have entered clinical trials for BC and have been shown to induce a durable clinical response in certain patients with metastatic BC[43]. However, our results need to be verified in future clinical trials of immunotherapy.

Our study also revealed that the m^5Cscore based on m^5C modifications was closely associated with the main clinical characteristics of BC, including pathologic stage, HER2 expression, and survival status, and could be used to assess the clinical characteristics and predict prognosis in patients with BC. To complement the theoretical and computational analyses conducted in our study, we reinforced our findings with empirical validation through RT-qPCR and IHC. The RT-qPCR results clearly indicated that the expression levels of the m^5C-related genes were significantly increased in BC as compared to the corresponding normal tissue controls. This certain amount of molecular evidence provides strong confirmation of our bioinformatics prediction, confirming the differential expression of m^5C related genes in cancerous tissues and healthy mammary gland tissues. Moreover, the application of IHC once again confirmed our molecular findings. IHC staining showed a significant increase in the positive rate of m^5C-related gene expression in tumor tissues, and these qualitative data, combined with quantitative RT-qPCR results, help to strengthen the validity of our theoretical study.

In summary, we used single-cell and bulk-RNA sequencing data to identify the m^5Cscore can act as an independent prognostic biomarker in clinical practice for predicting patient survival and may be used to comprehensively evaluate m^5C modification patterns and their corresponding immune cell infiltration characteristics within individual patients with BC and assess their clinicopathological features. More importantly, we could also predict the efficacy of pharmacotherapy and the patients' clinical response to ICB through the m^5Cscore, which can guide more effective clinical practice.

Materials and methods

Single-cell RNA sequencing

Study design and data collection

Single-cell RNA sequence (scRNA-seq) data from 5 breast cancer (BC) patients and 5 control patient samples in the Gene Expression Omnibus (GEO) (GSE161529) dataset were collected to analyze the landscape of m^5C regulators. scRNA-seq data from 5 breast cancer (BC) patients and 5 control patient samples in the Gene Expression Omnibus (GEO) (GSM5354517) dataset were collected to validated the landscape of m^5C regulators. Full data was downloaded in the GEO database (www.ncbi.nlm.nih.gov/geo), the clinical characteristics of single-cell samples are listed in Table S3. All data generated or analyzed during this study are freely available in previous publications or the public domain.

Analysis of scRNA-seq data

Gene expression data for both control and BC tissues were analyzed using R software (version 4.2.2). The data was converted into Seurat objects using the Seurat R package (version 4.3.0.1). Only cells expressing 200 to 7500 genes (including mitochondrial content less than 15%) were retained. The expression data for control and breast

tissues were integrated using the "FindIntegrationAnchors" and "IntegrateData" functions in the Seurat package. The count data were normalized after quality control. To reduce the computational burden and noise in the data, principal component analysis (PCA) was used for initial dimensionality reduction. K nearest neighbour graphs were constructed using the FindNeighbors function based on the Euclidean distance in the PCA space, whereas cells were clustered using the Louvain algorithm. The annotated information for each cell in the dataset reported in a previous article was visualized using t-Distributed Stochastic Neighbor Embedding (TSNE).

Differential gene expression analysis

The difference in m^5C regulators expression between cells was analyzed using the "FindMarkers" function (Wilcoxon rank-sum test) in the Seurat package (version 4.2.0). Significant expression was based on $|Log2FC| > 0.25$ at statistical significance of $p < 0.05$.

Construction of metacell maps

The MetaCell method useed the K-nn graph algorithm to divide the scRNA-seq dataset into unconnected and uniform cell groups (metacells) for epithelial and myeloid cells, respectively. Based on the gene count matrix, feature genes with scaling variance (variance/mean of the down-sampled matrices) exceeding 0.08 were selected, and the similarity between cells was calculated using Pearson correlation. Based on the inter-cell similarity matrices, the equilibrium K-nn similarity maps of two different cells are constructed with K as the parameter (the number of neighbors per cell is limited to K, $K = 40$). Perform the resampling process (resampling 75% of the cells per iteration, 500 iterations) and construct co-clustering graph (minimum cluster size was 50). A graphic of metacells (and the cells belonging to them) were projected onto a two-dimensional space to explore the similarities between cells and metacells.

Single-cell consensus weighted gene co-expression network analysis

We constructed metacells in which the software applied a bootstrapped aggregation process to the single-nucleus transcriptome. After the computation, cells of the same cell type and within the same sample will retain the new metacell for high-dimensional weighted gene coexpression network analysis (hdWGCNA). Modules were defined according gene expression in the metacell. The first principal component of the module, called the module eigengene, to correlate with diagnosis and other variables. Hub genes were defined using intra-modular connectivity (kME) parameters. Gene-set enrichment analysis was done using EnrichR.

Functional enrichment analysis of differential expression of genes (DEGs)

To explore functions and pathways of DEGs, the DEGs were loaded into the cluster profile package for GO(Gene Ontology) enrichment analysis. The adjusted p value of < 0.05.

Definition of m5C-related genes score

AddModuleScore and AUCell score are used to evaluate the strength of various cellular phenotypes or biological processes based on gene sets corresponding to each signature and gene expression data. For m^5C scores, m^5C regulators were identified between epithelial cells, lymphocytes, fibroblasts, endothelial and myeloid cells.

Bulk-RNA sequencing

Data sources and preprocessing

We used publicly available gene expression data and full clinical annotation of patients with BC in The Cancer Genome Atlas (TCGA) database. Patients with BC that lacked survival information were excluded; 1089 eligible tumor samples from the TCGA-Breast Invasive Carcinoma (BRCA) cohort were included for further analysis. RNA sequencing data (FPKM value) were downloaded from the Genomic Data Commons (https://portal.gdc.cancer.gov/), and the R package TCGAbiolinks, which is a software package developed for Genomic Data Commons data analysis, was used for integrative analysis[44]. Somatic mutation data were also acquired from TCGA and analyzed using R (version 4.1.2) and R Bioconductor. For an external testing dataset, we identified datasets related to BC in the Gene Expression Omnibus database (GSE7390 and GSE103091 cohorts) that were applied to validate the prognostic value of m^5C modification signature.

Unsupervised clustering analysis for 17 m5C regulators

We selected 17 m^5C methylation regulators that had expression data in TCGA. To classify patients for further analysis, we applied unsupervised cluster analysis to identify different m^5C modification patterns based on the expression of 17 m^5C regulators. The number and stability of clusters were determined by the consistent clustering algorithm. The "ConsensusClusterPlus" package was used for clustering[45]. In order to ensure the stability of the classification, we performed 1000-times cycle computation.

Gene set variation analysis and functional annotation

We performed gene set variation analysis (GSVA) using the "GSVA" R package to investigate differences in the enriched biological processes between different m^5C modification patterns. GSVA, a non-parametric and unsupervised method, is commonly employed to estimate variations in pathway and biological process activity using RNA-Seq data[46]. The "c2.cp.kegg.v6.2.symbols" gene sets were used to run the GSVA; these gene sets were downloaded from the Molecular Signatures Database (https://www.gsea-msigdb.org/gsea/msigdb). A corrected P-value of less than 0.005 was considered statistically significant. The "clusterProfiler" R package was used to annotate the functions of m^5C-related genes, with a critical false discovery rate (FDR) of less than 0.05.

Comparison of cell infiltration abundance in tumor microenvironment based on m5C patterns
To compare the infiltration of immune cells among samples with different m^5C modification patterns, we used single-sample gene set enrichment analysis (ssGSEA) to quantify the relative abundance of each cell type in the breast TME and applied the gene set obtained from the study by Charorntong for ssGSEA[47]. The relative abundance of each immune cell in each sample was assessed using the enrichment scores calculated by ssGSEA. We also calculated the tumor purity, stromal, immune, and ESTIMATE scores in each sample using the ESTIMATE algorithm to evaluate the tumor component[48]. The "limma" package in R was used to analyze the scores between different subgroups.

DEGs identification of different m5C modification modes
According to the expression of 17 m^5C regulatory factors, the tumor samples were divided into three different m^5C modification modes. The empirical Bayesian method was used to identify DEGs in three groups[49]. The significance criterion of deg was adjusted to $P<0.05$,and $|\log2(FC)| > 1.0$. Gene Ontology and Kyoto Encyclopedia of Genes and Genomes were used to enrich pathways associated with DEGs.

Generation of an m5C gene signature
Owing to the heterogeneity and complexity of m^5C modifications, we constructed a scoring systems to quantify the m^5C modification pattern of individual patients with BC. This gave rise to the m^5C gene signature, which we termed the m^5Cscore. We extracted overlapping DEGs in different m^5C clusters from all BC samples and analyzed the extracted DEGs using unsupervised clustering. The number and stability of gene clusters were determined by consensus clustering algorithm. The univariate Cox regression model was used to analyze the relationship between prognosis and overlapping DEG, and m^5C gene markers were constructed by principal component analysis of genes with significant prognostic differences. Principal component 1 and principal component 2 were selected as signature scores. Finally, we defined the m^5C score using a method similar to the Genome Grading Index[50,51], as follows:

$$m5Cscore = \sum(\mathrm{PC1i} + \mathrm{PC2i})$$

where i is the expression of m^5C phenotype-related genes.

We evaluated the constructed m^5C modification signature in both the training (TCGA-BRCA cohort) and testing (GSE7390 and GSE103091 cohorts) datasets as described above.

Assessing the response to immunotherapy and drug sensitivity
Tumor Immune Dysfunction and Exclusion (TIDE) is a computational method that models tumor immune evasion[41]. We used TIDE to predict the response to immune checkpoint blockade (ICB) in different m^5Cscore groups (http://tide.dfci.harvard.edu/). The R package "limma" was used to calculate TIDE scores in different m^5Cscore groups, and the R package "pRRophetic" was used to examine the expression profile of the TCGA-BRCA cohort for a drug sensitivity analysis[52].

RT-qPCR
Patients were selected from BC screening of eligible women who were residents in Shanxi province (Table S4). Invasive breast cancer was diagnosed by histopathology, while the control group consisted of normal breast cells. Three cases each of BC tissue and normal breast tissue were included in the experiment. Participants were excluded based on the following criteria: (1) lactating women, (2) patients with a history of mastectomy, (3) patients with a history of treatment for breast lesions, (4) patients with other malignant tumors, and (5) patients with blood and digestive system diseases. There was no statistically significant difference in age between the two groups of patients ($P>0.05$). All patients with BC underwent histopathology. For BC tissue, histopathology was performed by a pathologist with more than two years of experience. To limit the deviation of routine pathological diagnosis for BC, forceps were used to remove two additional pieces of tissue (approximately 5 mm each). Then, the tissue was washed with physiological saline and placed in a solution containing RNA preservation and tissue fixation. The samples used for RT qPCR were stored overnight in a 4 ℃ refrigerator and then transferred to a -20 ℃ refrigerator for storage at room temperature for hematoxylin and eosin samples.

Total RNA was extracted from normal and malignant breast tissue using TRIzol reagent (Invitrogen) according to the manufacturer's instructions. The primers were synthesized by Sangon Biotech (Shanghai, China). The housekeeping gene GADPH was used as an internal control. The primers used are listed in Table S5.

All reactions were conducted on Roche LightCycler 96 PCR Machine (Roche, Mannheim, Germany) using the following cycling parameters: step 1: denaturation at 94 °C for 30 s; step 2: 40 cycles of 94 °C for 5 s and 60 °C for 30 s. Gene expression was calculated using the ΔΔCt method. All data represent the average of three replicates.

Statistical analysis

One-way ANOVA and the Kruskal–Wallis test were used to compare differences among three or more groups[53]. Spearman and distance correlation analyses were used to calculate the correlation coefficient between the expression of m^5C regulators and infiltrating immune cells. The "survminer" R software package was used to determine the optimal cutoff point of the m^5Cscore for predicting prognosis, and patients were divided into high and low m^5Cscore groups. Survival curves were generated using the Kaplan–Meier method, and log-rank tests were used to identify the significance of differences. A univariate Cox regression model was used to calculate hazard ratios for m^5C regulators and m^5C phenotype-related genes. Multivariate Cox regression analysis was used to evaluate independent prognostic factors. Waterfall plots representing the mutation landscapes of the high and low m^5Cscore groups were created using the waterfall function of the R software "maftools" package[54].

All tests were bilateral, and $P < 0.05$ was considered statistically significant. and the Benjamini-Hochberg method was applied to control the false discovery rate (FDR) for multiple hypothesis testing[55].

Data availability

Publicly available datasets were analyzed in this study. The data for this study can be found in the Cancer Genome Atlas database (https://portal.gdc.cancer.gov/), the Gene Expression Omnibus database (https://www.ncbi.nlm.nih.gov/geo/). And the Immunohistochemical samples' data from the Human Protein Atlas (https://www.proteinatlas.org/).

Received: 19 April 2024; Accepted: 22 October 2024
Published online: 11 November 2024

References

1. Sung, H. et al. Global Cancer statistics 2020: GLOBOCAN estimates of incidence and Mortality Worldwide for 36 cancers in 185 countries. *CA Cancer J. Clin.* **71**, 209–249. https://doi.org/10.3322/caac.21660 (2021).
2. Giaquinto, A. N. et al. Breast Cancer Statistics, 2022. *CA Cancer J. Clin.* **72**, 524–541. https://doi.org/10.3322/caac.21754 (2022).
3. Kudelova, E. et al. Genetic Heterogeneity, Tumor Microenvironment and Immunotherapy in Triple-Negative Breast Cancer. *Int. J. Mol. Sci.* **23** (23), 14937. https://doi.org/10.3390/ijms232314937 (2022).
4. Wen, Q. E. et al. Recent Advances in Immunotherapy for Breast Cancer: A Review. *Breast Cancer (Dove Med. Press)*. **16**, 497–516. https://doi.org/10.2147/BCTT.S482504 (2024).
5. Trixl, L. & Lusser, A. The dynamic RNA modification 5-methylcytosine and its emerging role as an epitranscriptomic mark. *Wiley Interdiscip Rev. RNA*. **10**, e1510. https://doi.org/10.1002/wrna.1510 (2019).
6. Motorin, Y., Lyko, F. & Helm, M. 5-methylcytosine in RNA: detection, enzymatic formation and biological functions. *Nucleic Acids Res.* **38**, 1415–1430. https://doi.org/10.1093/nar/gkp1117 (2010).
7. Fu, L. et al. Tet-mediated formation of 5-hydroxymethylcytosine in RNA. *J. Am. Chem. Soc.* **136**, 11582–11585. https://doi.org/10.1021/ja505305z (2014).
8. Chen, Y. S., Yang, W. L., Zhao, Y. L. & Yang, Y. G. Dynamic transcriptomic m(5) C and its regulatory role in RNA processing. *Wiley Interdiscip Rev. RNA*. **12**, e1639. https://doi.org/10.1002/wrna.1639 (2021).
9. Gao, Y. et al. NOP2/Sun RNA methyltransferase 2 promotes tumor progression via its interacting partner RPL6 in gallbladder carcinoma. *Cancer Sci.* **110**, 3510–3519. https://doi.org/10.1111/cas.14190 (2019).
10. Dzobo, K. Taking a Full Snapshot of Cancer Biology: Deciphering the Tumor Microenvironment for Effective Cancer Therapy in the Oncology Clinic. *Omics*. **24**, 175–179. https://doi.org/10.1089/omi.2020.0019 (2020).
11. Yue, X., Lio, C. J., Samaniego-Castruita, D., Li, X. & Rao, A. Loss of TET2 and TET3 in regulatory T cells unleashes effector function. *Nat. Commun.* **10**, 2011. https://doi.org/10.1038/s41467-019-09541-y (2019).
12. Salemme, V., Centonze, G., Cavallo, F., Defilippi, P. & Conti, L. The Crosstalk Between Tumor Cells and the Immune Microenvironment in Breast Cancer: Implications for Immunotherapy. *Front. Oncol.* **11**, 610303. https://doi.org/10.3389/fonc.2021.610303 (2021).
13. Lawson, D. A., Kessenbrock, K., Davis, R. T., Pervolarakis, N. & Werb, Z. Tumour heterogeneity and metastasis at single-cell resolution. *Nat. Cell. Biol.* **20**, 1349–1360. https://doi.org/10.1038/s41556-018-0236-7 (2018).
14. Zhang, Y. et al. Single-cell analyses reveal key immune cell subsets associated with response to PD-L1 blockade in triple-negative breast cancer. *Cancer Cell.* **39**, 1578–1593e1578. https://doi.org/10.1016/j.ccell.2021.09.010 (2021).
15. Pal, B. et al. A single-cell RNA expression atlas of normal, preneoplastic and tumorigenic states in the human breast. *Embo j.* **40**, e107333. https://doi.org/10.15252/embj.2020107333 (2021).
16. Chen, Y., Pal, B., Lindeman, G. J., Visvader, J. E. & Smyth, G. K. R code and downstream analysis objects for the scRNA-seq atlas of normal and tumorigenic human breast tissue. *Sci. Data.* **9**, 96. https://doi.org/10.1038/s41597-022-01236-2 (2022).
17. Nombela, P., Miguel-López, B. & Blanco, S. The role of m(6)A, m(5)C and Ψ RNA modifications in cancer: Novel therapeutic opportunities. *Mol. Cancer*. **20**, 18. https://doi.org/10.1186/s12943-020-01263-w (2021).
18. Chen, X. et al. 5-methylcytosine promotes pathogenesis of bladder cancer through stabilizing mRNAs. *Nat. Cell. Biol.* **21**, 978–990. https://doi.org/10.1038/s41556-019-0361-y (2019).
19. Sun, Z. et al. Aberrant NSUN2-mediated m(5)C modification of H19 lncRNA is associated with poor differentiation of hepatocellular carcinoma. *Oncogene*. **39**, 6906–6919. https://doi.org/10.1038/s41388-020-01475-w (2020).
20. Guo, G. et al. Disease Activity-Associated Alteration of mRNA m(5) C Methylation in CD4(+) T Cells of Systemic Lupus Erythematosus. *Front. Cell. Dev. Biol.* **8**, 430. https://doi.org/10.3389/fcell.2020.00430 (2020).
21. Zeng, H. et al. Construction and Analysis of a Colorectal Cancer Prognostic Model Based on N6-Methyladenosine-Related lncRNAs. *Front. Cell. Dev. Biol.* **9**, 698388. https://doi.org/10.3389/fcell.2021.698388 (2021).
22. Zhang, B. et al. m(6)A regulator-mediated methylation modification patterns and tumor microenvironment infiltration characterization in gastric cancer. *Mol. Cancer*. **19**, 53. https://doi.org/10.1186/s12943-020-01170-0 (2020).
23. Liu, T. et al. 5-methylcytosine RNA methylation regulators affect prognosis and tumor microenvironment in lung adenocarcinoma. *Ann. Transl Med.* **10**, 259. https://doi.org/10.21037/atm-22-500 (2022).
24. Yang, S. C. et al. Inhibition of DNMT1 potentiates antitumor immunity in oral squamous cell carcinoma. *Int. Immunopharmacol.* **111**, 109113. https://doi.org/10.1016/j.intimp.2022.109113 (2022).
25. Wang, L. et al. PAX5 Haploinsufficiency Induces Low T Cell Infiltration in the Cancer Microenvironment via Reduced Chemokines. *Curr. Mol. Med.* **22** (9), 826–834. https://doi.org/10.2174/1566524021666211206094046 (2022).
26. Fischer, M., Schade, A. E., Branigan, T. B., Müller, G. A. & DeCaprio, J. A. Coordinating gene expression during the cell cycle. *Trends Biochem. Sci.* **47** (12), 1009–1022. https://doi.org/10.1016/j.tibs.2022.06.007 (2022).
27. Luo, N. et al. DNA methyltransferase inhibition upregulates MHC-I to potentiate cytotoxic T lymphocyte responses in breast cancer. *Nat. Commun.* **9**, 248. https://doi.org/10.1038/s41467-017-02630-w (2018).
28. Cheng, T. C. et al. Identification of a novel five ferroptosis-related gene signature as a promising prognostic model for breast cancer. *J. Cancer Res. Clin. Oncol.* **149**, 16779–16795. https://doi.org/10.1007/s00432-023-05423-5 (2023).
29. Wang, Z., Li, Z. & Ji, H. Direct targeting of β-catenin in the Wnt signaling pathway: Current progress and perspectives. *Med. Res. Rev.* **41** (4), 2109–2129. https://doi.org/10.1002/med.21787 (2021). Epub 2021 Jan 21.
30. Hussain, S. et al. NSun2-mediated cytosine-5 methylation of vault noncoding RNA determines its processing into regulatory small RNAs. *Cell. Rep.* **4**, 255–261. https://doi.org/10.1016/j.celrep.2013.06.029 (2013).
31. Blanco, S. et al. Aberrant methylation of tRNAs links cellular stress to neuro-developmental disorders. *Embo j.* **33**, 2020–2039. https://doi.org/10.15252/embj.201489282 (2014).
32. Tang, Y. et al. m6A-Atlas: a comprehensive knowledgebase for unraveling the N6-methyladenosine (m6A) epitranscriptome. *Nucleic Acids Res.* **49**, D134–d143. https://doi.org/10.1093/nar/gkaa692 (2021).
33. Yang, Z. et al. RNA N6-methyladenosine reader IGF2BP3 regulates cell cycle and angiogenesis in colon cancer. *J. Exp. Clin. Cancer Res.* **39**, 203. https://doi.org/10.1186/s13046-020-01714-8 (2020).

34. Oshi, M. et al. CD8 T Cell Score as a Prognostic Biomarker for Triple Negative Breast Cancer. *Int. J. Mol. Sci.* https://doi.org/10.3390/ijms21186968 (2020).
35. Baker, K. et al. Prognostic significance of CD8 + T lymphocytes in breast cancer depends upon both oestrogen receptor status and histological grade. *Histopathology*. **58**, 1107–1116. https://doi.org/10.1111/j.1365-2559.2011.03846.x (2011).
36. Garaud, S. et al. Tumor-infiltrating B cells signal functional humoral immune responses in breast cancer. JCI Insight 4. (2019). https://doi.org/10.1172/jci.insight.129641
37. Wu, S. Y., Fu, T., Jiang, Y. Z. & Shao, Z. M. Natural killer cells in cancer biology and therapy. *Mol. Cancer*. **19**, 120. https://doi.org/10.1186/s12943-020-01238-x (2020).
38. Rezaeifard, S., Talei, A., Shariat, M. & Erfani, N. Tumor infiltrating NK cell (TINK) subsets and functional molecules in patients with breast cancer. *Mol. Immunol.* **136**, 161–167. https://doi.org/10.1016/j.molimm.2021.03.003 (2021).
39. Jin, H. et al. Natural killer cells inhibit breast cancer cell invasion through downregulation of urokinase–type plasminogen activator. *Oncol. Rep.* **45**, 299–308. https://doi.org/10.3892/or.2020.7840 (2021).
40. Goodman, M. L. et al. Progesterone Receptor Attenuates STAT1-Mediated IFN Signaling in Breast Cancer. *J. Immunol.* **202**, 3076–3086. https://doi.org/10.4049/jimmunol.1801152 (2019).
41. Parker, B. S., Rautela, J. & Hertzog, P. J. Antitumour actions of interferons: implications for cancer therapy. *Nat. Rev. Cancer*. **16**, 131–144. https://doi.org/10.1038/nrc.2016.14 (2016).
42. Jiang, P. et al. Signatures of T cell dysfunction and exclusion predict cancer immunotherapy response. *Nat. Med.* **24**, 1550–1558. https://doi.org/10.1038/s41591-018-0136-1 (2018).
43. Emens, L. A. Breast Cancer Immunotherapy: Facts and Hopes. *Clin. Cancer Res.* **24**, 511–520. https://doi.org/10.1158/1078-0432.Ccr-16-3001 (2018).
44. Colaprico, A. et al. TCGAbiolinks: an R/Bioconductor package for integrative analysis of TCGA data. *Nucleic Acids Res.* **44**, e71. https://doi.org/10.1093/nar/gkv1507 (2016).
45. Wilkerson, M. D. & Hayes, D. N. ConsensusClusterPlus: a class discovery tool with confidence assessments and item tracking. *Bioinformatics*. **26**, 1572–1573. https://doi.org/10.1093/bioinformatics/btq170 (2010).
46. Hänzelmann, S., Castelo, R. & Guinney, J. GSVA: gene set variation analysis for microarray and RNA-seq data. *BMC Bioinform.* **14**, 7. https://doi.org/10.1186/1471-2105-14-7 (2013).
47. Charoentong, P. et al. Pan-cancer Immunogenomic Analyses Reveal Genotype-Immunophenotype Relationships and Predictors of Response to Checkpoint Blockade. *Cell. Rep.* **18**, 248–262. https://doi.org/10.1016/j.celrep.2016.12.019 (2017).
48. Yoshihara, K. et al. Inferring tumour purity and stromal and immune cell admixture from expression data. *Nat. Commun.* **4**, 2612. https://doi.org/10.1038/ncomms3612 (2013).
49. Ritchie, M. E. et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* **43**, e47. https://doi.org/10.1093/nar/gkv007 (2015).
50. Sotiriou, C. et al. Gene expression profiling in breast cancer: understanding the molecular basis of histologic grade to improve prognosis. *J. Natl. Cancer Inst.* **98**, 262–272. https://doi.org/10.1093/jnci/djj052 (2006).
51. Zeng, D. et al. Tumor Microenvironment Characterization in Gastric Cancer Identifies Prognostic and Immunotherapeutically Relevant Gene Signatures. *Cancer Immunol. Res.* **7**, 737–750. https://doi.org/10.1158/2326-6066.Cir-18-0436 (2019).
52. Geeleher, P., Cox, N. & Huang, R. S. pRRophetic: an R package for prediction of clinical chemotherapeutic response from tumor gene expression levels. *PLoS One*. **9**, e107468. https://doi.org/10.1371/journal.pone.0107468 (2014).
53. Hazra, A. & Gogtay, N. Biostatistics Series Module 3: Comparing Groups: Numerical Variables. *Indian J. Dermatol.* **61**, 251–260. https://doi.org/10.4103/0019-5154.182416 (2016).
54. Mayakonda, A., Lin, D. C., Assenov, Y., Plass, C. & Koeffler, H. P. Maftools: efficient and comprehensive analysis of somatic variants in cancer. *Genome Res.* **28**, 1747–1756. https://doi.org/10.1101/gr.239244.118 (2018).
55. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15** (12), 550. https://doi.org/10.1186/s13059-014-0550-8 (2014).

Acknowledgements

We would like to thank S.-M.W. laboratory from Shanxi Medical University for their discussion. We appreciate the contributions of our colleagues who participated in the research and wrote this article.

Author contributions

Zhe Wang and Jinpeng Li were involved in the design and coordination of the study, data analysis, and interpretation of results and is a major contributor in writing the manuscript. Jie Hou, Wei Wang and Shiming Wang were in charge of all the study procedures and helped perform the analysis with constructive discussions. All others, including Fucheng Wang, Chen Cheng, Xinpei Wu, Wendi Guo, Chenquan Li, Yinyi Luo, Guangwen Zhang, and Sanyuan Zhang participated in the study procedures and critically revised the content of the manuscript. All authors contributed to the article and approved the submitted version.

Funding

This work was supported by the China Postdoctoral Science Foundation under Grant 2021M691995; Shanxi Provincial Key Research and Development Project under Grant 201803D31111; Natural Science Foundation of Shanxi Province under Grant 202103021224394; Natural Science Foundation of Shanxi Province under Grant 202203021221250; The Research Special Fund for Key Clinical Specialty Construction Projects in Shanxi Province under Grant 2024-ZZ-043.

Declarations

Ethics approval and consent to participate

The studies involving human participants were reviewed and approved by Ethics Committee of the First Hospital of Shanxi Medical University, the ethics number is K-K0109. The patients provided their written informed consent to participate in this study.

Consent for publication

The patients involved have obtained ethical approval and written informed consent for the publication of any potentially identifiable images or data included in this article.

Competing interests
The authors declare no competing interests.

Statement
All methods were performed in accordance with the relevant guidelines and regulations.

Additional information
Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-024-77389-4.

Correspondence and requests for materials should be addressed to J.H., W.W. or S.W.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.