

<https://doi.org/10.1038/s42003-025-09330-y>

# Age-related aberrant alternative splicing as a prognostic tool in older breast cancer patients

Check for updates

Mengxin Chen<sup>1,7</sup>, Taiping Xiao<sup>1,7</sup>, Hao Ke<sup>2,3,7</sup>, Chun Ye<sup>1</sup>, Fan Yu<sup>1</sup>, Jianbin Su<sup>1</sup>, Shixin Yang<sup>4</sup>, Yun Feng<sup>1</sup>, Huaimeng Xu<sup>1</sup>, Shiting Fu<sup>1</sup>, Liang Lin<sup>1</sup>, Junqiang Lin<sup>1</sup>, Yao Xie<sup>1</sup>, Qianzhe Ding<sup>1</sup>, Ye Qiu<sup>5</sup>, Ruifan Yu<sup>1</sup>, Hongyu Lv<sup>1</sup>, Wulian Li<sup>1</sup>, Yuhan Zhang<sup>1</sup>, Zhenluo Ding<sup>6</sup>, Yang Zou<sup>2</sup>✉, Huozhong Yuan<sup>6</sup>✉ & Limin Zhao<sup>1</sup>✉

Breast cancer is one of the most prevalent malignancies among women. Although the impact of age on breast cancer progression is well documented, the role of aberrant alternative splicing events in older adults breast cancer patients remains poorly understood. Here, we identified that older adults breast cancer patients exhibit a higher frequency of aberrant alternative splicing events compared to younger patients using public database, a finding that was further validated by data from FUSCC cohorts. A total of 390 high-variability-specific splicing events were observed exclusively in older adults patients. The unsupervised clustering analysis revealed the existence of three distinct subtypes of older adults patients, each displaying significantly different immune cell infiltration profiles and prognostic outcomes. To identify the key regulatory factors of these splicing subtypes, we conducted AS activity score analysis and identified 68 RNA-binding proteins as potential modulators. Subsequently, a machine learning approach using SelectKBest-SVM was employed to construct a predictive model, which demonstrated optimal performance in predicting the prognosis of older adults breast cancer patients, with a high AUC and validation on an independent test set. The developed predictive model offers a promising tool for personalized treatment strategies and accurate prognostication, advancing precision medicine for older adults breast cancer patients.

Breast cancer is one of the most common malignancies worldwide, ranking second among cancers in women, just behind lung cancer<sup>1</sup>. It is well documented that age is a critical risk factor for the development of breast cancer. Both incidence and mortality rates increase markedly in individuals aged 65 years and above<sup>2,3</sup>. The ageing global population has resulted in an increased prevalence of breast cancer in older adults women, emphasizing the necessity for more effective treatment strategies for this growing population<sup>4</sup>. In fact, older adults patients often display distinct clinical and molecular characteristics compared to younger individuals, which can result in less effective treatment responses when standard therapeutic protocols are used<sup>5,6</sup>. However, the specific molecular mechanisms contributing to the

distinctive characteristics of breast cancer in older adults patients, particularly those involving age-related alterations, remain poorly understood<sup>6,7</sup>.

One key regulatory process that contributes to these age-related differences is alternative splicing (AS), a post-transcriptional mechanism that generates protein diversity by selectively including or excluding exons from the mature mRNA. AS plays an essential role in numerous cellular processes, and its dysregulation is closely associated with cancer progression<sup>8–11</sup>. Recent studies have shown that the frequency of aberrant AS events increases with age, and such alterations have been implicated in the development of various cancers, including breast cancer<sup>12–15</sup>. Nevertheless, the characteristics of AS in older adults breast cancer patients have been unclear.

<sup>1</sup>Human Aging Research Institute (HARI) and School of Life Science, Nanchang University, and Jiangxi Key Laboratory of Human Aging and Disease, Nanchang, Jiangxi, China. <sup>2</sup>Central Laboratory, Jiangxi Maternal and Child Health Hospital, Nanchang, Jiangxi, China. <sup>3</sup>The MOE Basic Research and Innovation Center for the Targeted Therapeutics of Solid Tumors, School of Basic Medicine, Jiangxi Medical College, Nanchang University, Nanchang, China. <sup>4</sup>Department of Breast Surgery, Jiangxi Maternal and Child Health Hospital, Nanchang, Jiangxi, China. <sup>5</sup>Huankui Academy, Nanchang University, Nanchang, Jiangxi, China. <sup>6</sup>Ganzhou People's Hospital, Ganzhou, China. <sup>7</sup>These authors contributed equally: Mengxin Chen, Taiping Xiao, Hao Ke. ✉e-mail: [zouyang81@163.com](mailto:zouyang81@163.com); [yhuozh@163.com](mailto:yhuozh@163.com); [zhaolimin@ncu.edu.cn](mailto:zhaolimin@ncu.edu.cn)

In this study, we demonstrate that older adults breast cancer patients exhibit a higher frequency of aberrant AS events compared to younger patients, suggesting that age-related changes in splicing contribute to the distinct clinical characteristics of older adults breast cancer. We identify 390 splicing events uniquely prevalent in older adults patients, which are associated with immune microenvironmental differences and prognosis. Furthermore, we identify 68 RNA-binding proteins that regulate these splicing events and construct a predictive machine learning model, which shows optimal accuracy in prognostic prediction for older adults patients. These findings provide a foundation for using AS and RBPs as potential biomarkers and therapeutic targets in personalized breast cancer treatment.

## Results

### Increased abnormal AS events in breast cancer with aging

In this study, we downloaded clinical data and corresponding gene expression profiles of BRCA patients from the TCGA database. We collected 342 cancer samples and 31 adjacent normal samples from older adults breast cancer patients. To compare the number of AS abnormal events, we also collected 186 cancer samples and 31 adjacent normal samples from younger breast cancer patients. We downloaded the PSI matrix of AS data for BRCA patients from TCGASpliceSeq and ultimately organized a gene expression matrix for 31,529 genes and a PSI matrix for 13,479 AS events of 5685 genes. To examine whether there were differences in splicing patterns between cancer and normal samples that could facilitate the subsequent search for AS abnormal splicing events, we used t-SNE dimensionality reduction clustering visualization with PSI values associated with each AS event (Fig. 1A). The t-SNE visualization results indicated apparent differences in splicing patterns between normal and cancer samples. To minimize analytical interference, six samples with discrete value (three aged tissues and three young tissues) were excluded from subsequent analyses (see Discussion section). After screening, we finally obtained 340 cancer samples ( $n = 340$ ) and 30 adjacent normal samples ( $n = 30$ ) in the older adults group; in the young group, we obtained 184 cancer samples ( $n = 184$ ) and 30 adjacent normal samples ( $n = 30$ ) (Supplementary Data 1). We then used t-SNE for visualization again and confirmed that after the removal of the abnormal samples, the t-SNE clustering results more clearly distinguished between normal and cancer samples (Fig. 1B, C). We used t-tests to identify AS abnormal splicing events in both the older adults and young groups (Supplementary Data 2–3). The results showed that although there was no apparent difference in the proportion of AS abnormal splicing event types between the two groups (Fig. 1D), the older adults group had a higher number of abnormal splicing events overall (Fig. 1E). To test the accuracy of the differential analysis and explore its biological relevance, we confirmed the intersection of differential splicing events from both groups (Fig. 1F), of which 643 intersection differential splicing events involved 561 genes (Fig. 1G) and were significantly enriched in pathways related to cell-substrate junction (Fisher's exact test,  $P < 0.05$ ) and focal adhesion (Fisher's exact test,  $P < 0.05$ ), which are directly or closely linked to tumor progression (Fig. 1H). Among the 643 intersection differential splicing events, there are many variable splicing events that have been experimentally confirmed and proven to be closely related to the occurrence and progression of cancer, such as the variable splicing of GNAS<sup>16</sup>, which was found and confirmed to be closely related to the occurrence of breast cancer in previous reports. These results indicate the commonality of the intersection abnormal splicing events in driving the occurrence and progression of breast cancer in the two groups. The visualization of t-SNE dimensionality reduction clustering of all differential splicing events (abnormal splicing events in the older adults and young breast cancer patients, Supplementary Data 2 and 3) in the two groups (Fig. S2A). To investigate age-related splicing alterations in breast cancer, we analyzed AS events in older adults patients. A heatmap visualizing older adults-specific PSI values revealed different splicing patterns between tumor and adjacent normal tissues in the older adults cohort (Fig. S2B). To further investigate whether these older adults-specific splicing events could distinguish tumor from normal tissue within the older adults group, we performed t-SNE analysis. The result demonstrated clear

segregation between tumor and normal tissue samples in the older adults patients based on their older adults-specific splicing profiles (Fig. S2C). Similarly, we assessed the mRNA expression levels of these older adults-specific splicing events using t-SNE. In contrast to the splicing profiles, gene expression patterns showed unfavorable separation between tumor and normal tissues in the older adults patients (Fig. S2C).

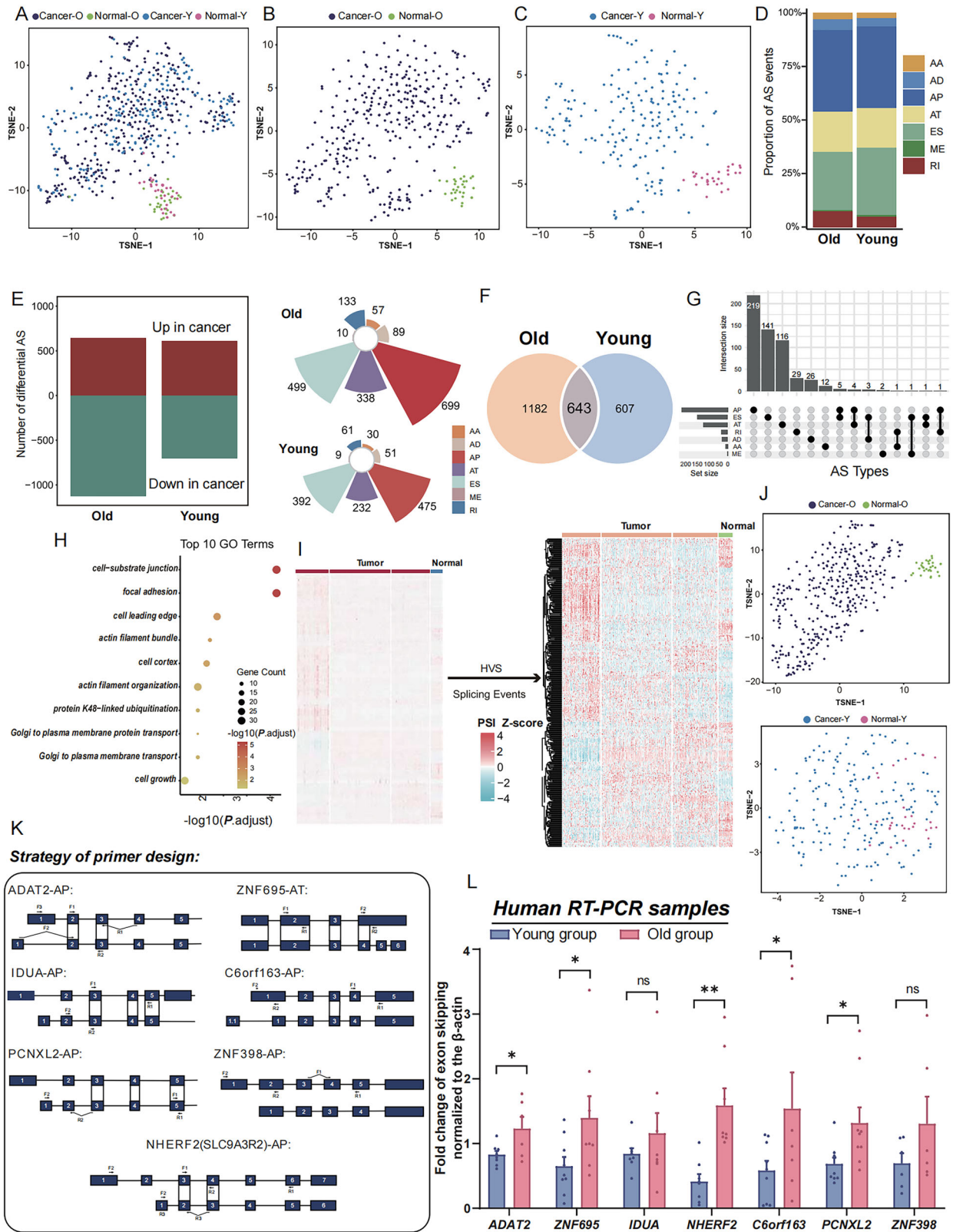
To validate the observation that older adults breast cancer patients exhibit a higher frequency of aberrant splicing events, we analyzed the data using an independent tool, TCGA\_SpIAdder. Consistent with our SpliceSeq findings, TCGA\_SpIAdder analysis revealed more aberrant splicing events in older patients compared to younger patients (Fig. S4A, S4B). Moreover, in FUSCC cohort, we independently validated findings using rMATS, MISO, and SUPPA2 tools. Analysis with rMATS (Fig. S4C, upper), MISO (Fig. S4C, lower), and SUPPA2 (Fig. S4D, S4E) confirmed an increased prevalence of aberrant AS events in older patients.

Exploring high-variability-specific (HVS) splicing events and the splicing factors that drive them in this population could better represent the older adults and hold important clinical relevance. Splicing events with high variance have a substantial impact on the overall splicing landscape. By using a threshold of  $\text{std\_psi} > 0.15$  to identify high-variance splicing events, we identified 390 such events (Fig. 1I and Supplementary Data 4). These events were visualized using t-SNE clustering (Fig. 1J), which demonstrated that they could distinguish two distinct types of samples within the older adults group but lacked similar representativeness in the young group. A heatmap of PSI values further confirmed that these splicing events are more representative in older adults breast cancer patients. Among the 390 splicing events with high variability, most showed significantly different PSI levels compared to normal tissues (Wilcoxon test,  $P < 0.05$ ) (Fig. 1I and Fig. S3B). These events involved 326 genes (Supplementary Data 5) and covered a variety of splicing event types (Figure S3A).

To further explore these findings, we conducted univariate Cox regression analysis to select the top ten splicing event genes with potential prognostic relevance, involving thirteen splicing events (Fig. S3C and Supplementary Table 1). Additionally, we experimentally validated the specificity of these splicing events in the older adults breast cancer population. We conducted experimental confirmation of the specificity of some of these splicing events in the older adults population. We designed specific primers targeting these splicing events (Fig. 1K, see Methods for details) and analyzed samples from the Ganzhou cohort, dividing them into young and older adults groups for RT-qPCR experiments. The results demonstrated that these splicing events occurred at significantly higher frequencies in the older adults group (Student's  $t$  test,  $P < 0.05$ ), confirming their specific presence in this population (Fig. 1L). Based on univariate Cox regression analysis and experimental validation results, we found several aberrant splicing events with high incidence in the older adults population, which may be closely associated with their prognosis. These findings provide important insights into the biological relevance of older adults-specific abnormal splicing events and lay a foundation for further exploration of their role in the prognosis of the older adults population.

### Unsupervised clustering analysis revealed three different splicing patterns in older adults breast cancer patients according to HVS splicing events in the older adults

In older adults breast cancer patients, more abnormal AS event disorder occurs, and different patients show varying degrees of splicing pattern disorder. According to the similarity of splicing patterns, older adults breast cancer patients can be further divided into multiple subtypes. To more accurately explore and analyze these subtypes, we applied non-negative matrix factorization (NMF) clustering, based on the percentage splicing index (PSI) values of the high-variability specific (HVS) splicing events, to cluster analysis of older adults BRCA patients. Through multiple repetitions of clustering, the optimal number of clusters was determined to be  $K = 3$  (Fig. 2A), and finally, 340 older adults BRCA patients were divided into three AS splicing patterns, named AS Cluster1 ( $n = 85$ ), AS Cluster2 ( $n = 156$ ), and AS Cluster3 ( $n = 99$ ) (Fig. 2B). T-SNE (t-distribution



stochastic neighbor embedding) results further verified the validity of this classification (Fig. 2C). Subsequently, we conducted survival analysis on the three types of patients to explore the differences in their clinical prognosis. The three splicing patterns showed significant survival differences (log-rank test,  $P < 0.05$ ) in overall survival (OS) (Fig. 2D). Among them, AS Cluster1 showed a clear survival advantage, while AS Cluster3 had dismal

survival prognosis. In disease-free survival (DFS) analysis, AS Cluster3 also showed dismal survival prognosis (log-rank test,  $P < 0.05$ ) (Fig. S5A). This finding prompted us to delve deeper into the molecular differences among the three splicing patterns. We further compared the differences among the three clusters in terms of breast cancer IHC classification and PAM50 typing (Fig. 2E). The results indicated that patients in Cluster3 were more likely to

**Fig. 1 | Comparison of Abnormal Splicing Events in Older Adults and Young Breast Cancer Patients.** **A** The positions of splicing models for different sample groups in two-dimensional space after t-SNE dimensionality reduction. Deep blue: older adults breast cancer patient samples; green: older adults normal samples; light blue: young breast cancer patient samples; pink: young normal samples. **B** Positioning of older adults breast cancer patient samples and older adults normal samples in two-dimensional space after t-SNE dimensionality reduction, excluding abnormal samples. **C** Positioning of young breast cancer patient samples and young normal samples in two-dimensional space after t-SNE dimensionality reduction, excluding abnormal samples. **D** Bar graph showing the proportions of seven types of abnormal splicing events in the older adults and young groups. The seven AS events analyzed are: AA Alternate Acceptor site, AD Alternate Donor site, AP Alternate Promoter, AT Alternate Terminator, ES Exon Skip, ME Mutually Exclusive Exons, and RI Retained Intron. **E** Bar graph comparing the number of abnormal splicing events between the older adults cancer group and the young cancer group (right). Red indicates upregulated AS; green indicates downregulated AS. The radar map specifically shows the specific number of the corresponding splicing events (left).

**F** Venn diagram showing the number of common abnormal splicing events between the older adults cancer group and the young cancer group. **G** UpSet plot displaying the types of splicing events involved in the intersecting splicing events between the older adults and young cancer groups (points indicate which splicing events are involved, and the bars above represent the intersecting genes for those splicing events). **H** Dot plot showing the top ten pathways from GO enrichment analysis of genes involved in the intersecting splicing events between the older adults and young cancer groups. **I** Heatmap of the PSI matrix that for high-variability specific (HVS) splicing events in older older adults. Data include  $n = 340$  tumor samples and  $n = 30$  normal samples in the older adults group. **J** t-SNE plot of HVS splicing events in the older adults group (left). t-SNE plot of HVS splicing events in the young group (right). **K** The strategy used for the primer design. **L** qPCR experimental results of alternative splicing-related genes between the young and older adults groups. Data are presented as mean  $\pm$  SD,  $n = 9$  biologically independent samples per group. Statistical significance was determined using a two-tailed unpaired Student's t-test. ns, not significant ( $P > 0.05$ ); \* $P < 0.05$ ; \*\* $P < 0.01$ .

exhibit the most malignant molecular subtype—triple-negative breast cancer (TNBC). This finding is consistent with the previous prognosis analysis, which suggested that patients in Cluster3 have the most dismal prognosis in clinical settings. In addition, in TNM staging analysis, the proportion of N2/N3 in Cluster3 patients in lymph node (Node) staging was significantly higher (Fisher's exact test,  $P < 0.05$ ) than in the other two clusters, indicating that these patients may be more prone to local and distant lymph node metastasis (Fig. S5B). However, there were no significant differences (Fisher's exact test,  $P > 0.05$ ) among the three groups in tumor primary site (T stage) and distant metastasis (M stage). To provide a more detailed of the characteristics distinguishing the different clusters, we further analyzed the tumor microenvironment and genomic instability among the clusters. This is further reflected in the hypoxia scores, with Cluster3 exhibiting significantly higher hypoxia levels across all three scoring systems: Buffa, Winter, and Ragnum, indicating a more hypoxic tumor microenvironment (Fig. 2F). Moreover, Cluster3 exhibited significantly higher (Kruskal-Wallis test,  $P < 0.05$ ) Aneuploidy scores (Fig. S5C), indicative of strong genomic instability. Further supporting this, higher Fraction Genome Altered scores (Kruskal-Wallis test,  $P < 0.05$ ) suggest that Cluster3 is characterized by a genomic landscape dominated by chromosomal instability (Fig. S5C). Moreover, we analyzed the H&E staining results in the TCGA cohort to explore the tumor-infiltrating lymphocyte (TIL) levels in the three types of patients<sup>17</sup>. The results showed that the TIL levels in Cluster3 patients were significantly higher (Kruskal-Wallis test,  $P < 0.05$ ) than in other groups (Fig. 2G, H), we also further show the representative H&E staining and its statistical data (Fig. 2H). While the TIL levels and the proportion of high-TILs areas in Cluster1 were lower. Studies have shown that an increase in TIL proportion is usually associated with prolonged survival in HER2-positive or TNBC patients; however, in HER2-negative ductal breast cancer patients, a high TIL proportion is a negative prognostic factor, indicating that different subtypes of breast cancer may have different immune response patterns<sup>18</sup>. Although patients in Cluster3 have a high TIL level and a higher proportion of TNBC subtype patients, theoretically, their survival should be extended, but the actual results are not the case. These findings prompted us to further explore the differences in the tumor immune microenvironment among the three clusters to reveal potential molecular mechanisms.

### Three different splicing patterns exhibit distinct immune landscapes and molecular features

We used the ESTIMATE method to estimate the overall levels of immune cell and stromal cell infiltration in older adults breast cancer samples (Fig. 3A–C). Notably, compared to AS Cluster1 and AS Cluster2, AS Cluster3 showed increased immune scores and increased stromal scores (Kruskal-Wallis test,  $P < 0.05$ ), but lower tumor purity, indicating a higher level of tumor-infiltrating immune cells in the Cluster3 subtype. Cluster1 had higher immune scores than Cluster2, while Cluster2 had higher stromal

scores than Cluster1. The higher immune score in Cluster1 suggests a stronger immune response in its tumor microenvironment, which may be associated with more favorable survival prognosis. Cluster2, with a higher stromal score and the highest tumor purity among the patients, indicates a rich stromal component in the tumor microenvironment, with tumor cells dominating the microenvironment, which may promote tumor invasiveness and metastasis, leading to more dismal prognosis. To compare the differences in TME cell infiltration between clusters, we comprehensively assessed the TME infiltration patterns in older adults breast cancer patients. As shown in Fig. 3D, cells, such as Memory B cells and Monocytes exhibited higher levels of infiltration in AS Cluster1 (Kruskal-Wallis test,  $P < 0.05$ ), which may lead to stronger immune surveillance and anti-tumor responses, thus improving prognosis. Although AS Cluster3 was in a state of high immune infiltration, they may be functionally exhausted and unable to effectively combat the tumor, resulting in a more dismal prognosis<sup>19,20</sup>. We further compared the differences in immune-related gene set scores among the three splicing subtypes in terms of immune checkpoints, antigen processing mechanisms, CD8 T effector features and EMT marker gene sets, including EMT1, EMT2, and EMT3, were highest in Cluster3, suggesting its tumor cells have stronger invasive and metastatic abilities (Fig. 3E). In summary, the AS Cluster3 subtype with high immune infiltration may be more prone to metastasis. HLA genes play a key role in antigen presentation and immune recognition, and their expression can influence the immune response in the tumor microenvironment. Therefore, we compared the expression of HLA family genes among the three clusters. Cluster3 showed higher expression of multiple HLA genes (Kruskal-Wallis test,  $P < 0.05$ ), indicating active antigen presentation, although high HLA gene expression is typically associated with a stronger immune response, in Cluster3, combined with the high expression of immune checkpoint genes, it may suggest that the tumor activates the HLA pathway while also evading immune surveillance through other means<sup>21</sup> (Fig. 3F). Finally, we used various deconvolution algorithm methods, such as Quantiseq, Timer, MCP\_counter, and Xcell to calculate the levels of immune cell infiltration in breast cancer samples using RNA-seq data<sup>22–25</sup>, further demonstrating the high infiltration level of Cluster3 (Fig. S6A). Additionally, there were significant differences (Kruskal-Wallis test,  $P < 0.05$ ) in mRNA expression of immune checkpoints between clusters (Fig. S6B). To reveal possible T cell exhaustion in Cluster3, we examined the expression of exhaustion-related genes, such as LAG3, TIGIT, LAYN, PDCD1, HAVCR2, and CXCL13 in the three splicing subtypes<sup>26,27</sup>, and these marker genes were all highly expressed (Kruskal-Wallis test,  $P < 0.05$ ) in Cluster3 (Figs. 3G and S6C–S6E), further suggesting the possibility of T cell exhaustion in this cluster.

### Mutational characteristics of splicing pattern clusters

Subsequently, to further characterize and discuss the mechanisms of the three splicing patterns, we performed gene set variation analysis (GSVA). We first compared the pathways upregulated in each of the three clusters: in

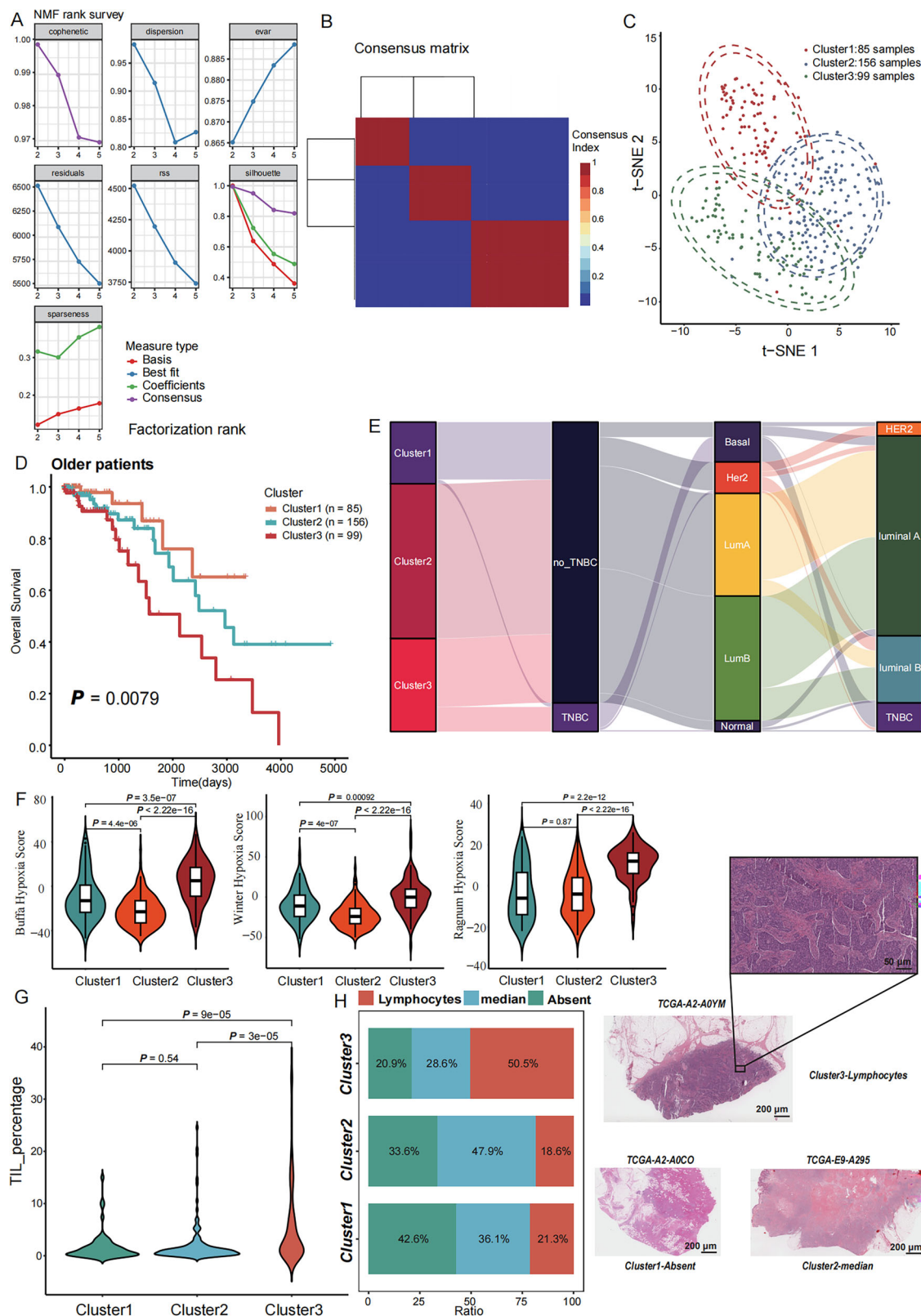


Fig. 3H, Cluster1 showed higher enrichment in signaling pathways, such as INFLAMMATORY RESPONSE, E2F TARGETS, and MTORC1 SIGNALING; Cluster2 showed relatively neutral or slightly downregulated pathways, but the upregulation of its ESTROGEN RESPONSE LATE pathway indicated increased sensitivity to estrogen signaling, which may be related to tumor proliferation, metabolic activation, and sensitivity to

hormone therapy. Cluster3 showed higher enrichment in pathways related to metabolism, cell proliferation, and oxidative phosphorylation. Next, we compared the pathways downregulated in each cluster (Fig. 3I). In Cluster3, we particularly noted immune-related signaling pathways, such as INTERFERON GAMMA RESPONSE and INTERFERON ALPHA RESPONSE, which are closely related to immune responses. The

**Fig. 2 | Specific Differential Splicing Events in the Older Adults Lead to Three Splicing Patterns.** **A** Performance metrics of the NMF algorithm under different ranks of decomposition. Each subplot represents a performance metric, with the x-axis indicating the decomposition rank and colors representing different measurement types, determining the optimal k value through correlation coefficients, dispersion coefficients, and silhouette coefficients. **B** Optimal threshold for patient classification determined using the NMF R package. Older adults BRCA patients were divided into three AS splicing patterns. AS Cluster1 ( $n = 85$ ), AS Cluster2 ( $n = 156$ ), and AS Cluster3 ( $n = 99$ ). **C** Visualization of sample clustering results based on the t-SNE algorithm to verify the reliability of the non-negative matrix factorization analysis results. Each point represents a sample; color represents different clustering results, and dashed ellipses represent the boundaries of each cluster. **D** Kaplan-Meier survival curves for different subgroups of older adults cancer samples. The x-axis represents time (in days), and the y-axis represents overall survival. Different colored curves represent different patient subgroups (Clusters 1–3), with the number of samples in each group indicated in the legend.  $P$ -values indicate the statistical significance of differences in overall survival among different subgroups (Log-rank test). **E** Sankey diagram illustrating the relationship between the three clusters and the molecular subtypes of breast cancer. The left bar graph

represents the sample grouping based on non-negative matrix factorization analysis (Clusters 1–3), followed by the molecular subtypes of breast cancer (TNBC subtype, PAM50 subtype, and IHC identified subtype). The connecting bands indicate the correspondence of samples between the two classification systems, with the width of the bands representing sample quantity. **F** Comparison of hypoxia scores across three scoring systems (Buffa, Winter, and Ragnum) in different clusters. Cluster3 exhibits significantly higher hypoxia levels compared to Cluster1 and Cluster2. Horizontal lines and numbers indicate  $P$ -values for pairwise comparisons between subgroups, as determined by the Kruskal-Wallis test ( $*P < 0.05$ ,  $**P < 0.01$ ,  $***P < 0.001$ ,  $****P < 0.0001$ ). The box plots show the median (center line) and the 25–75th percentiles (box limits). **G** Violin plot showing differences in TIL (tumor-infiltrating lymphocyte) levels among the three AS splicing patterns. The width of the violin shape represents the frequency of occurrence of that percentage, and the black point in the middle represents the median. The horizontal line and numbers at the top indicate the  $P$ -values for pairwise comparisons between subgroups. **H** Proportions of TIL levels in the three AS splicing patterns (left) and representative pathological HE-stained images (right). Absent: low TILs (immune cells occupy 0–10% of tumor stroma); median: moderate TILs (11–59%); and Lymphocytes: high TILs ( $\geq 60\%$ ). Scale bar: 200  $\mu\text{m}$  (50  $\mu\text{m}$  in magnified images).

downregulation of interferon signaling pathways may imply a poorer response to viral and anti-tumor immunity in Cluster3. Additionally, the downregulation of the IL6\_JAK\_STAT3\_SIGNALING pathway suggests possible immune suppression or a reduction in inflammatory responses in Cluster3. Through GSEA analysis, we found that Cluster1 exhibited strong activity in antigen processing and presentation pathways, suggesting higher immune activity, which may contribute to a more favorable prognosis. The enrichment of the HIF-1 signaling pathway in Cluster2 suggests enhanced tumor cell adaptation to hypoxic conditions, further promoting tumor survival and spread. In Cluster3, the enrichment of the ECM receptor interaction pathway also suggests potential invasiveness and migratory ability of tumor cells. In addition, we conducted GO enrichment analysis for the differentially expressed genes in each subtype compared to the other two subtypes (Fig. S6F–S6H), further validating the differences in pathway mechanisms mentioned above. In summary, we propose to name Cluster1 the Immune Enhancement type, Cluster2 the Hormone Activation type, and Cluster3, which exhibits enhanced metabolism and invasiveness, the Metabolic Invasion type. We next performed the GSEA enrichment analysis to further demonstrate the characteristic pathways of each cluster (Fig. 3J).

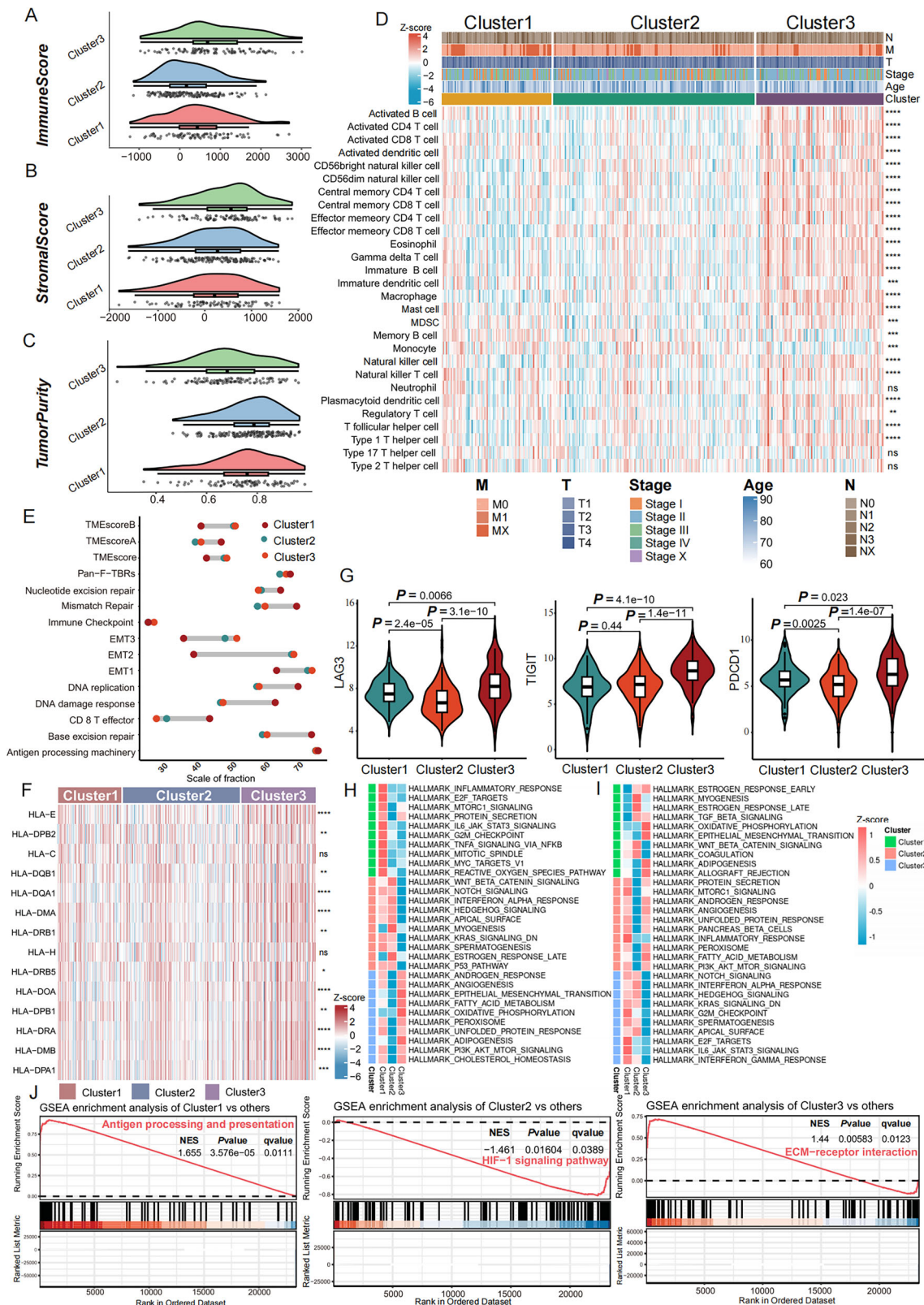
Subsequently, we analyzed somatic copy number variation (SCNVs) data from the TCGA database to comprehensively analyze the differences in somatic mutations caused by the three splicing patterns. Waterfall plots showed the differences in mutation patterns among the three splicing subtypes (Fig. 4A–C). The top 20 mutated genes showed substantial overlap across the three subtypes, indicating shared oncogenic mechanisms. Cluster3 stood out with a notably higher mutation rate (91.58%) compared to the other two groups, particularly in TP53 (Fisher's exact test, odds ratio = 4.960344,  $P < 0.05$ ). This pattern suggests that TP53 mutations may drive enhanced genomic instability in Cluster3 tumor cells. High mutation rates of TP53 are typically associated with more aggressive tumors and more dismal prognosis, indicating that the response to treatment in Cluster3 may be more complex<sup>28</sup>. Further analysis using GISTIC 2.0 revealed complex copy number variations among the three different splicing types (Clusters 1–3). In Figs. 4D, E, both the gistic score and frequency showed marked differences in chromosome 8 and chromosome 16 in Cluster3. In the comparison of tumor mutation burden (TMB), Cluster3 had a significantly higher mutation burden (Kruskal-Wallis test,  $P < 0.05$ ) than the other two clusters (Fig. 4F), further supporting the idea that tumors in Cluster3 have higher genomic instability. These results consistently indicate that tumors in Cluster3 show higher genomic instability, which may lead to a higher mutation burden and chromosomal copy number variations. This genomic instability may be an important characteristic of the altered splicing pattern phenotype and is closely related to tumor invasiveness and heterogeneity<sup>29</sup>.

### Identify RNA-binding proteins that may act as HVS splicing events regulators

AS events are primarily regulated by RNA-Binding Protein (RBPs). RBPs typically bind to precursor mRNA and regulate RNA splicing by influencing the selection of exons and splice sites. Further exploration of the RBP-AS regulatory network will help understand the molecular mechanisms underlying the development of BRCA and provide potential targets for precision therapy. Therefore, in this study, exploring the regulatory network between RBPs and AS events is of great importance. To determine the molecular regulators associated with the 390 high-variability specific (HVS) splicing events, we first determined the extent of perturbation of these HVS splicing events by comparing the PSI (splicing percentage index) for each splicing event with the normal sample (Fig. 5A). We calculated two activity scores for these splicing events: an activation activity score (SA) and a suppression activity score (SR), ultimately identifying 201 over-activated and 189 over-suppressed splicing events. We then manually collected 2217 RBPs as candidate regulators of splicing events<sup>30–33</sup> (Supplementary Data 6). Next, we determined the potential RBP regulators of AS events in different cancer types based on the correlation between the expression of candidate regulators and AS activity scores (Fig. 5B). In total, we found 1552 RBPs positively correlated with the activation score and 1450 RBPs positively correlated with the suppression score in over-activated splicing events (Fig. 5C, D, and Supplementary Data 7–8). Given the adverse prognostic nature of Cluster3 in our earlier analyses, we focused on splicing factors abnormally expressed in Cluster3. Ultimately, we identified 68 differentially expressed splicing factors (Wilcoxon test,  $P < 0.05$ ) in Cluster3 (Fig. 5E, F). Nine of these splicing factors were activating RNA-binding proteins (RBPs) (Fig. 5E), and 59 were inhibitory RBPs (Fig. 5F). Each of these factors was predicted to regulate at least 30 of the 390 HVSs (Supplementary Data 7 and 8), based on previously identified molecular regulators associated with HVSs (Fig. 5A–D). Representative examples illustrated the relationship between HVS splicing events and specific splicing factors (Fig. 5G).

### Constructing a machine learning classifier model based on RBPs as regulators to predict the prognosis of older adults breast cancer patients

We believe that these potential splicing factors may indeed be able to predict splicing pattern subtypes and prognosis in the older adults population. We attempted to construct a classifier that could distinguish between Cluster3 and Non-Cluster3. Using these potential splicing factors, we built a machine learning model, and after feature selection and classifier construction (Fig. 6A), we systematically evaluated 50 combinations of feature selection methods and machine learning algorithms. Model performance was



assessed using metrics including Area Under the Curve (AUC) (Fig. 6B, S7A–D) and Matthews Correlation Coefficient (MCC) (Fig. 6C). Thirteen models demonstrated strong performance, with AUC > 0.75 and MCC > 0.4. These included SelectKBest-LR, SelectKBest-GP, SelectKBest-KNN, SelectKBest-SVM, RFE\_RF-GP, RFE\_RF-DA, RFE\_RF-SVM, SelectFromModel\_LR-KNN, SelectFromModel\_RF-LR, SelectFromModel\_RF-

GP, SelectFromModel\_RF-DA, SelectFromModel\_RF-DT, and SelectFromModel\_RF-RF. To validate our findings externally, we considered the METABRIC dataset, a comprehensive, publicly available breast cancer resource containing clinical and molecular data from over 2000 patients. Results indicated that only the SelectKBest-SVM model demonstrated satisfactory performance (Fig. 6D, Supplementary Table 2). Further

**Fig. 3 | Different TME and Signaling Pathway Characteristics of the Three AS RNA Processing Patterns.** A–C Differences in immune scores, stromal scores, and tumor purity among the AS clusters (AS Cluster1:  $n = 85$ ; AS Cluster2:  $n = 156$ ; AS Cluster3:  $n = 99$ ) were assessed using the ESTIMATE algorithm. Violin plots indicate the probability density distribution of the data, and box plots summarize statistical features (median, quartiles, range, and outliers). Statistical significance was determined using the Kruskal–Wallis test. D Heatmap displaying the infiltration levels of 28 immune cell types among the three AS clusters determined by the ssGSEA algorithm. Asterisks indicate  $P$ -values (Kruskal–Wallis test,  $*P < 0.05$ ,  $**P < 0.01$ ,  $***P < 0.001$ ,  $****P < 0.0001$ ). E Dot plot showing differences in multiple functional scores among the three subgroups (Cluster1, Cluster2, and Cluster3); different colors represent different subgroups, with horizontal line segments connecting points for the same score across different subgroups for easier comparison. F Heatmap displaying mRNA expression levels of multiple HLA genes in AS Cluster1, AS Cluster2, and AS Cluster3. Red indicates high expression, and blue indicates low expression. Asterisks on the right indicate statistical significance of gene expression differences between different clusters ( $*P < 0.05$ ,  $**P < 0.01$ ,

$***P < 0.001$ ,  $****P < 0.0001$ ). G Box plots showing mRNA expression of Tumor marker genes (LAG3, TIGIT, PDCD1) in AS Cluster1, AS Cluster2, and AS Cluster3, with horizontal lines and numbers indicating  $P$ -values for pairwise comparisons between subgroups (Kruskal–Wallis test). The box plots show the median (center line) and the 25–75th percentiles (box limits). H Heatmap displaying differences in activation states of multiple HALLMARK biological processes among the three AS clusters as revealed by GSVA enrichment analysis, with red representing activation and blue representing suppression. I Heatmap showing differences in suppression states of biological processes among the three AS clusters as revealed by GSVA enrichment analysis, with red representing activation and blue representing suppression. J GSEA enrichment analysis of specific biological processes for the three AS splicing patterns. The upper curve plot shows enrichment of three significantly enriched gene sets in the ordered gene list, with peaks representing the most significantly enriched positions for each gene set. The middle bar graph shows the position of each gene in the ordered list and its associated gene set. The right table lists the three significantly enriched gene sets along with their normalized enrichment scores (NES),  $P$ -values, and adjusted  $P$ -values ( $Q$ -values).

validation using independent datasets (GSE42568 and GSE20685) consistently revealed that the SelectKBest-SVM model predicted a significantly poorer prognosis for patients in Cluster3 (log-rank  $P < 0.05$ ) (Fig. 6E, F, Supplementary Table 2). In contrast, other models with similar cross-validation metrics failed to consistently predict prognosis robustly. The poor prognostic signature associated with Cluster3 was further validated in an independent GSE25066 external dataset (Fig. S7E). Finally, we confirmed SelectKBest-SVM as the optimal classifier.

We applied this classifier to our Ganzhou cohort to predict the subtypes to which patients might belong. First we confirmed by H&E staining results that older breast cancer patients predicted as Cluster3 subtype indeed had higher TIL levels (Fig. 6G). In addition, by comparing the abnormal splicing events in the older adults group and the abnormal splicing events in the younger group in the Ganzhou cohort, Fig. 6H demonstrated representative examples with split-read counts and exon quantification data. ADAM15 exhibited increased the proportion of exon exclusion in Cluster3, while CCDC88B displayed decreased the proportion of exon exclusion. These directional changes were independently validated using MISO (Fig. S8A), confirming their consistency with rMATS outputs (Fig. 6H).

To provide an isoform-level interpretation of AS changes, we employed IsoformSwitchAnalyzeR to identify differential isoform switching between Cluster3 and Non-Cluster3 groups (Fig. S8B). KEGG enrichment analysis further revealed that genes undergoing significant isoform switching were associated with tumor-related pathways, including cell adhesion and p53 signaling (Fig. S8C). Significant isoform switches were observed in key genes, including MGP and ARHGDI1. In Cluster3, MGP exhibited a pronounced shift from the full-length transcript ENST00000228938.5 (encoding the functional Gla domain) to the truncated isoform ENST00000545199.5, which lacks exons 1, 2, and 4 (Fig. S8D). Similarly, ARHGDI1 showed decreased expression of the functional transcript ENST00000541078.7 and increased expression of the NMD-sensitive isoform ENST00000582520.1 (Fig. S8E). We further analyzed differential exon usage in RNA-seq data using DEXSeq, which revealed significant exon usage differences for the genes MGP (ENSG00000111341) and ARHGDI1 (ENSG00000141522) between Cluster3 and Non-Cluster3 (Fig. S8F).

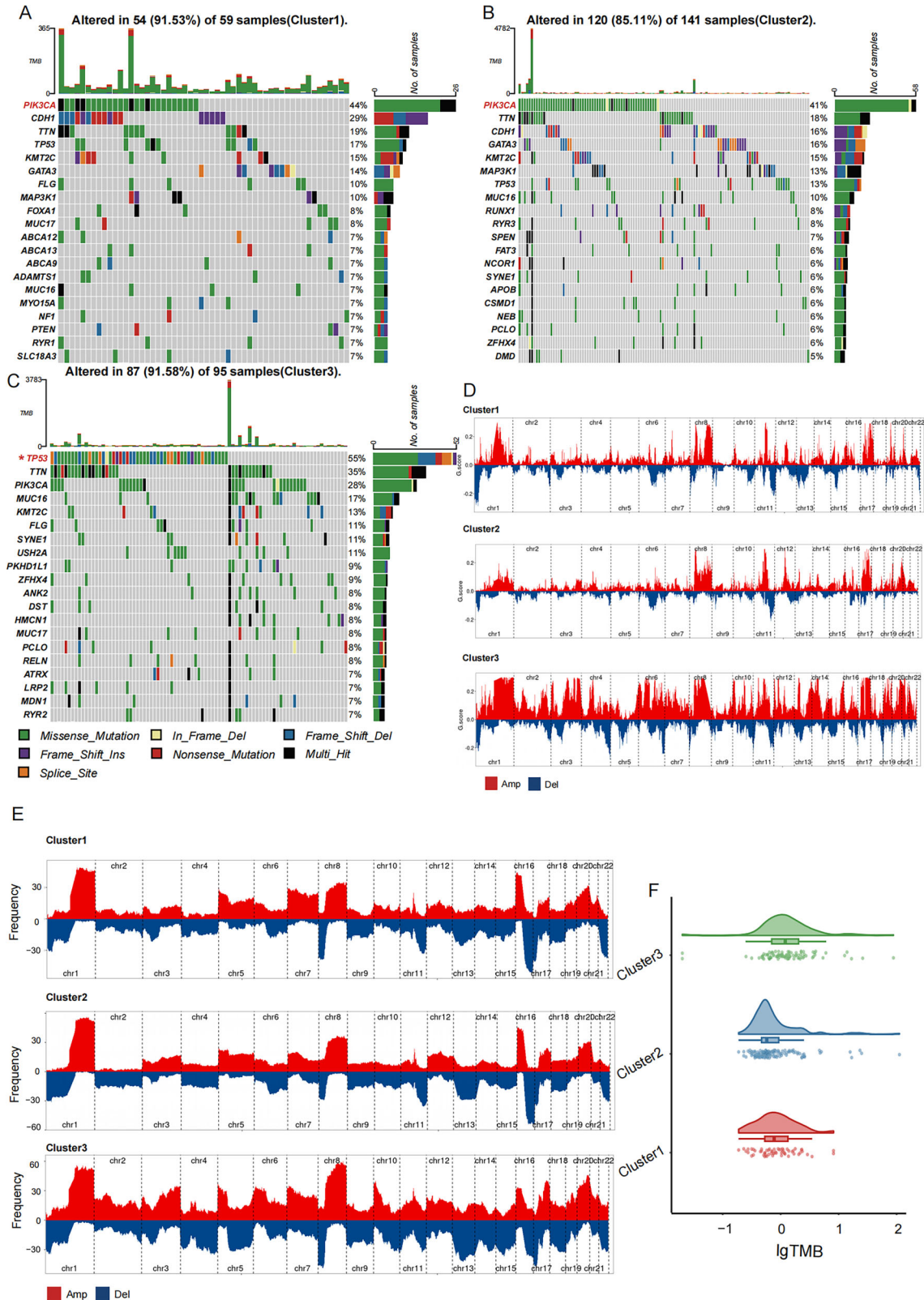
We also performed preliminary qPCR experiments at the protein level on the 10 RBPs that were eventually used for model construction, such as SNRNP70 low expression of activating RBPs in patients judged to be Cluster3, and other RBPs also showed the accuracy of the previous analysis on the overall trend (Fig. 6I).

To explore the application of our developed classifier in single-cell data, we collected single-cell sequencing data from breast cancer patients over the age of 65 and conducted integrated analysis<sup>34–36</sup>. Through cell annotation, we ultimately identified various cell types (Fig. 6J and S9A). We next compared the difference in the proportion of cell types predicted as Cluster3 and those predicted as Non-Cluster3, where Cluster3 exhibited more T cell proportions

(Fig. 6K and S9B). Since we had previously discussed the immune-related characteristics of Cluster3 in detail, we next focused on the T cell subpopulations. After annotating the T cell subpopulations, we determined several T cell subgroups, including Activated T cell, Cytotoxic T cell, Exhausted T cell, Naive, Proliferating, Th1, Treg (Fig. 6L and S9C). Then, we applied the classifier to determine the subtype affiliation of these cells (Fig. S9D). We analyzed the proportions of different T cell types across all samples in both groups. The results revealed a higher proportion of exhausted T cells in Cluster3 (15.8% VS 6.3%, chi-square tests,  $P < 2.2e-16$ ) compared to the Non-Cluster3 group (Fig. 6M). This finding aligns with our previous results, which showed increased expression of immune exhaustion-related genes in Cluster3 from RNA-seq data (Figs. 3G and S6C–S6E).

### Clinical recommended drugs for different splicing patterns

To explore drugs suitable for patients judged as Cluster3 by the machine learning model and to enhance the clinical value of the classifier, we collected drug sensitivity and response data from three pharmacosensitivity databases. We used the OncoPredict algorithm to convert gene expression data from 340 older adults breast cancer patients into a drug sensitivity matrix. In the GDSC (Fig. 7A), patients judged as Cluster3 showed higher sensitivity to 262 drugs, including AZD939 (mTOR inhibitor), XAV939 (tankyrase inhibitor), AZD2440 (receptor tyrosine kinase inhibitor). In the CCLE (Fig. 7B), Cluster3 patients were more sensitive to 13 drugs, with representative drugs including ABT2 (Bcl-2 inhibitor), BMS-2 (IKK inhibitor), Navitoclax (Bcl-1 inhibitor), TAF1\_5496 (TAF9 inhibitor). In the CTRP (Fig. 7C), Cluster3 patients were more sensitive to 360 drugs, including ABT2, BMS-2, Navitoclax, TAF5496 1, I-BRD9, and other targeted drugs. We took the intersection of drugs predicted to be helpful in treating Cluster3 patients from these databases. Ultimately, Paclitaxel (microtubule inhibitor), Erlotinib (EGFR inhibitor), and Saracatinib (Src family tyrosine kinase inhibitor) were predicted as highly sensitive drugs for Cluster3 patients in all three databases (Fig. 7D). To validate the results, we compared the expression levels of Saracatinib drug targets in each Cluster, confirming the high expression (Kruskal–Wallis test,  $P < 0.05$ ) of Saracatinib drug targets in Cluster3 (Fig. 7E). We believe these drugs may be helpful in treating patients judged as Cluster3 by the machine learning model. Additionally, in searching for possible resistance drugs (Figure S10A–S10C), although no common intersection was confirmed across the three datasets, drugs, such as Selumetinib (MEK inhibitor), Nutlin (p53-MDM2 inhibitor), Sorafenib (Raf kinase inhibitor), Panobinostat (HDAC inhibitor), and Tanespimycin (Hsp90 inhibitor) may have potential resistance capabilities related to Cluster3 patients. In summary, through intersection analysis of the three pharmacosensitivity databases, Paclitaxel, Erlotinib, and Saracatinib were predicted as highly sensitive drugs for Cluster3 patients, indicating that these drugs may provide effective treatment options for older adults breast cancer patients judged as Cluster3 by



the machine learning model, and offer suggestions for subsequent personalized treatment plans.

**Discussion**

This study systematically identifies and compares abnormal AS events in older adults and young breast cancer patients, confirming a largely increase

in abnormal AS events among older adults breast cancer patients. This phenomenon may be closely related to factors, such as the accumulation of gene mutations and epigenetic changes that occur with aging<sup>37,38</sup>. Our research reveals that the frequency of abnormal AS events increases with age, implicating aberrant splicing as a potential association between age and cancer occurrence and development<sup>39</sup>. Overall, our study highlights the

**Fig. 4 | Different Somatic Mutation Characteristics of the Three AS RNA Processing Patterns.** A–C Waterfall plots displaying somatic mutation profiles for the three AS clusters (AS Cluster1:  $n = 85$ ; AS Cluster2:  $n = 156$ ; AS Cluster3:  $n = 99$ ). Top horizontal bar plots show mutation burden per sample, with gene mutation frequencies indicated at right. Right vertical bar plots illustrate the proportion of different mutation types (missense, nonsense, frameshift, etc.) for each gene. C TP53 mutation frequency showed significant association with cluster subtype (Fisher's exact test,  $P < 0.05$ ), indicated by asterisk. Color keys for mutation types are provided below each plot. D Genomic copy number variation (CNV) map showing the

GISTIC scores for samples with copy number events in each AS cluster subtype, with the y-axis representing G-scores. Red regions indicate amplification, and blue regions indicate deletion. E Genomic copy number variation (CNV) analysis displaying the percentage of samples with copy number events in each AS cluster subtype. Red regions indicate amplification, and blue regions indicate deletion. F Differences in tumor mutational burden among the three AS splicing patterns, with violin plots indicating the probability density distribution of the data and box plots summarizing statistical features (median, quartiles, range, and outliers). Each gray point represents a sample's lgTMB value (Kruskal-Wallis test).

importance of age-specific AS in breast cancer progression. By integrating subtypes, immune profiling, RBP regulators, and isoform-level analyses, we provide a valuable resource for the field of AS for cancer research.

To define age groups, we selected thresholds of <45 years (young) and >65 years (older adults) based on clinical relevance and biological distinctions. This approach aligns with established criteria, where  $\leq 45$  years defines young-onset breast cancer and  $\geq 65$  years characterizes the older adults subgroup, as supported by prior literature<sup>40,41</sup>.

Through unsupervised clustering analysis of highly variable specific splicing events in older adults breast cancer patients, we identified three splicing pattern subtypes and further validated their association with differences in patient prognosis. Notably, the Cluster3 subtype exhibited poorer prognosis. Analysis of immune infiltration revealed a higher presence of lymphocytes, particularly in Cluster3. This observation is supported by both the analysis and the results from hematoxylin and eosin (H&E) staining (Figs. 2H, 6G). Further deconvolution algorithm of RNA-seq data confirmed an enrichment of activated and regulatory T cells within Cluster3 (Fig. 3D). Gene expression analysis also showed elevated expression of exhaustion-related genes in Cluster3 samples (Fig. 3F, G). These findings suggest that Cluster3 exhibits high immune infiltration, particularly highlighting a substantial coexistence of exhausted T cells, which may contribute to its adverse prognosis<sup>20</sup>. Single-cell sequencing data further corroborates a higher proportion of activated and exhausted T cells within Cluster3. This finding further emphasizes the critical role of the microenvironment in older adults breast cancer and offers a potential strategy for improving the prognosis of Cluster3 subtype patients through immune regulation.

Given the crucial role of RNA-binding proteins (RBP) in regulating AS events, changes in RBP activity may lead to tumor progression, metastasis, or relate to prognosis<sup>42</sup>. We analyzed and screened various potential RBP regulatory factors, experimentally validating the representation of these RBPs in older adults breast cancer patients. Among these RBPs are potential RNA-Binding Protein (RBP), which may serve as targets to aid in the treatment of older adults breast cancer patients in the future. We also constructed a machine learning classifier model based on SelectKBest-SVM to predict patient prognosis. Our model demonstrated excellent predictive performance in both internal and external validation sets, showcasing its potential in predicting the prognosis of older adults breast cancer patients. By utilizing the machine learning classifier model built with these RBP regulatory factors, we aim to better predict the prognosis of older adults breast cancer patients in clinical settings and assist physicians in selecting more precise treatment strategies.

We focused on older adults breast cancer patients and conducted an in-depth analysis of the characteristics and mechanisms of AS events that are more prevalent in this demographic. In previous studies, some researchers developed prognostic models for breast cancer based on other genomic features<sup>43,44</sup>. The highlight of this study lies in further revealing the role of RBPs in older adults breast cancer patients and establish a prognostic prediction model specifically for this population, based on RBP-related splicing events. By integrating these RBPs into a machine learning model for prognosis prediction, this approach has broad clinical application prospects as a potential auxiliary prognostic tool for breast cancer. This methodology could contribute to the development of more personalized treatment approaches for older adults breast cancer patients.

Our isoform-level analysis (Fig. S8B–S8F) revealed biologically consequential splicing switches with direct implications for breast cancer progression. The MGP isoform shift toward ENST00000545199.5 in poor-prognosis Cluster3 results in loss of the critical Gla domain (aa 57–82), thereby disrupting MGP's ability to inhibit metastasis via BMP-2 binding and calcification blockade<sup>42–44</sup>. Thus, the shift towards the Gla-domain-deficient ENST00000545199.5 isoform in poor-prognosis Cluster3 represents a functional inactivation of MGP's metastasis-suppressive activity, contributing to the adverse outcomes in this cluster. Concurrently, the ARHGDI1 switch toward an NMD-sensitive isoform (ENST00000582520.1) likely diminishes tumor-suppressor activity, consistent with ARHGDI1's role in inhibiting oncogenesis<sup>45,46</sup>. These isoform transitions—affecting protein domain integrity and stability—provide a functional basis for the adverse outcomes in Cluster3. This isoform-level perspective provides a clearer biological context for the observed AS changes, including potential clinical relevance by demonstrating how AS impacts older adults breast cancer progression in a biologically meaningful way.

While our analysis identified dysregulated splicing patterns in older adults breast cancer, we acknowledge the potential influence of splicing factor mutations (e.g., SF3B1, SRSF2, U2AF1, PRPF8, and RBM10). Only 12/340 patients (3.5%) harbored such mutations (Supplementary Table 3), distributed across all three molecular clusters (Cluster1:  $n = 2$ ; Cluster2:  $n = 5$ ; Cluster3:  $n = 5$ ). Critically, Fisher's exact test revealed no significant association between mutation status and cluster subtypes ( $P = 0.644$ ).

Of course, our model and explorations do have certain limitations. In the future, we will further increase the amount of data to improve the robustness of the model while exploring other regulatory factors of AS events, such as the microenvironment and immune responses, to gain a more comprehensive understanding of the pathological mechanisms in older adults breast cancer patients. Additionally, further functional experiments will help verify the direct regulatory role of RBPs on AS events, thereby deepening our understanding of cancer occurrence and development.

While Fig. 5E, F confirms marked dysregulation of activating/inhibitory RBPs distinguishing Cluster3 from Non-Cluster3 (Cluster1 & Cluster2), we observed subgroup heterogeneity within Non-Cluster3 samples. Re-analysis confirmed this pattern is neither an annotation artifact nor attributable to merging clusters 1 and 2—subgroups persisted even when these clusters were analyzed separately (Fig. 5E, F). This intra-group heterogeneity likely reflects unmodelled biological variation that future studies should explore.

This study employed a descriptive, non-interventional approach using computational models to predict drug sensitivity in older adults breast cancer patients. While the results offer valuable insights into survival prognosis and potentially clinically relevant drug recommendations, the conclusions should be interpreted cautiously given the lack of direct experimental manipulation. The absence of intervention prevents definitive establishment of causal relationships. Further research, including interventional studies directly testing the proposed mechanisms, is needed to validate these predictions.

## Methods

### Data acquisition and processing from public databases

TCGA SpliceSeq (<http://bioinformatics.mdanderson.org/TCGASpliceSeq>) provides AS event data for 33 different tumor types (including available



**Fig. 5 | Identification of Potential RNA-Binding Protein (RBP).** **A** Flowchart for constructing a variable splicing (AS) activity disturbance map. (1) Genome-wide AS profile: Collect RNA sequencing data from normal and cancer samples, calculate the splicing percentage (PSI) values for each AS event in each sample, and construct a genome-wide AS profile (left side of the heatmap). Blue indicates lower PSI values, while orange indicates higher PSI values. (2) AS disturbance profiles in cancer samples: Normalize the PSI values of cancer samples against the mean PSI values of normal samples, and calculate  $\Delta$ PSI values (formula shown in the figure). **B** List of RBPs filtered based on the correlation between the expression of candidate splicing event regulators and AS activity scores (active activity score (SA) and repressive activity score (SR)) ( $R > 0.3$ ). **C** Circle plot showing the correlation between over-active variable splicing events and RBPs. **D** Circle plot showing the correlation

between overly suppressed variable splicing events and RBPs. **E** Heatmap displaying the expression of nine active RBPs potentially regulating over 30 activated splicing events in Cluster3 ( $n = 99$ ) versus Non-Cluster3 (Cluster1,  $n = 85$  + Cluster2,  $n = 156$ ; total  $n = 241$ ), with asterisks on the right indicating statistical significance of gene expression differences between different clusters ( $*P < 0.05$ ,  $**P < 0.01$ ,  $***P < 0.001$ ,  $****P < 0.0001$ ). **F** Heatmap displaying the expression of 59 suppressive RBPs potentially regulating over 30 suppressed splicing events in Cluster3 ( $n = 99$ ) versus Non-Cluster3 (Cluster1,  $n = 85$  + Cluster2,  $n = 156$ ; total  $n = 241$ ), with asterisks on the right indicating statistical significance of gene expression differences between different clusters ( $*P < 0.05$ ,  $**P < 0.01$ ,  $***P < 0.001$ ,  $****P < 0.0001$ ). **(G)** PPI network diagram of representative RBPs and splicing events.

Corresponding bulk RNA-sequencing data, quantified as STAR-counts, and associated clinical information were procured from the GDC TCGA portal (<https://portal.gdc.cancer.gov/>). After integration, we obtained gene expression data for 60,660 genes across 1231 samples, with clinical information available for 1097 patients. For organizing the STAR-counts files, we first removed duplicate genes to ensure each gene appeared only once in the expression matrix. We then filled in missing values and eliminated low-expressed genes, specifically those that were not expressed in more than half of the samples, to reduce noise in subsequent analyses. Ultimately, we selected cancer and adjacent samples from patients aged over 65 and under 45 as the older adults breast cancer data and young breast cancer data, yielding a total of 585 eligible samples. Specifically, the older adults breast cancer dataset included 342 tumor samples and 31 adjacent samples; the young breast cancer dataset included 186 tumor samples and 31 adjacent samples. Additionally, we retrieved key genomic features, such as Aneuploidy Score<sup>47</sup>, Buffa Hypoxia Score<sup>48</sup>, Ragnum Hypoxia Score<sup>49</sup>, Winter Hypoxia Score<sup>50</sup>, MSI MANTIS Score, MSIsensor Score, TMB, Fraction Genome Altered<sup>51–53</sup>. These metrics were obtained from cBioPortal<sup>51</sup>, and were used to analyze genomic alterations and their potential impact on breast cancer progression and patient prognosis.

**TMB Calculation:** TMB was calculated by cBioPortal<sup>51,52</sup> using a variant calling pipeline that processes somatic mutations, including both single-nucleotide variants (SNVs) and small insertions/deletions (Indels). The mutations were identified from tumor samples, with paired normal samples to distinguish somatic mutations. The mutation count is normalized per megabase (mutations/megabase) to account for differences in sequencing depth.

**MSI Calculation:** MSI status was determined based on the analysis of microsatellite markers using cBioPortal's standardized pipeline. MSI-high (MSI-H), MSI-low (MSI-L), and microsatellite stable (MSS) classifications were assigned based on the instability observed at multiple microsatellite loci. MSI-H tumors show instability at two or more markers, while MSS tumors show no instability.

**Variant Calling Pipeline and Filtering:** cBioPortal uses well-established variant calling tools, including GATK and Mutect2, to identify somatic mutations. The data are subject to several filters, including quality control checks for sequencing depth, the exclusion of germline mutations, and the removal of low-confidence variants.

### External dataset validation for machine learning classifier

The external data validation sets for machine learning classification in this study were mainly obtained from The Gene Expression Omnibus (GEO) and Molecular Taxonomy of Breast Cancer International Consortium (METABRIC). Specifically, we downloaded GSE42568<sup>54</sup>, GSE20685<sup>55</sup>, GSE25066<sup>56</sup> datasets from GEO database (<https://www.ncbi.nlm.nih.gov/geo/>). Additionally, we acquired transcriptome data and relevant clinical information for breast cancer patients from the METABRIC database (<https://www.mercuriolab.umassmed.edu/metabric>) for validation purposes<sup>57</sup>.

### Patient cohorts

**FUSCC Patient Cohort:** The FUSCC cohort included 465 TNBC patients (100% female, mean age =  $53 \pm 11$  years)<sup>58</sup>. Breast cancer specimens and

corresponding adjacent tissues were obtained from patients undergoing mastectomy at Fudan University Shanghai Cancer Center (Shanghai, China). All experiments were conducted in accordance with the Declaration of Helsinki. A total of 360 patients' breast cancer RNA-seq data were obtained, and all tissue samples were obtained following approval from the ethics committee of Fudan University Shanghai Cancer Center. The ethics committee reference number was 1807188-16. All patients provided written informed consent to participate in this study.

### Breast cancer sample collection

Breast cancer specimens and corresponding adjacent tissues were obtained from patients undergoing mastectomy or biopsy at GanZhou People's Hospital (Ganzhou, China), which included 29 BRCA patients (100% female, mean age =  $53 \pm 11$  years). All experiments were conducted in accordance with the Declaration of Helsinki. RNA-seq data was available, and all tissue samples were obtained following approval from the ethics committee of GanZhou People's Hospital (TY-ZKY2024-064-01).

### RNA extraction

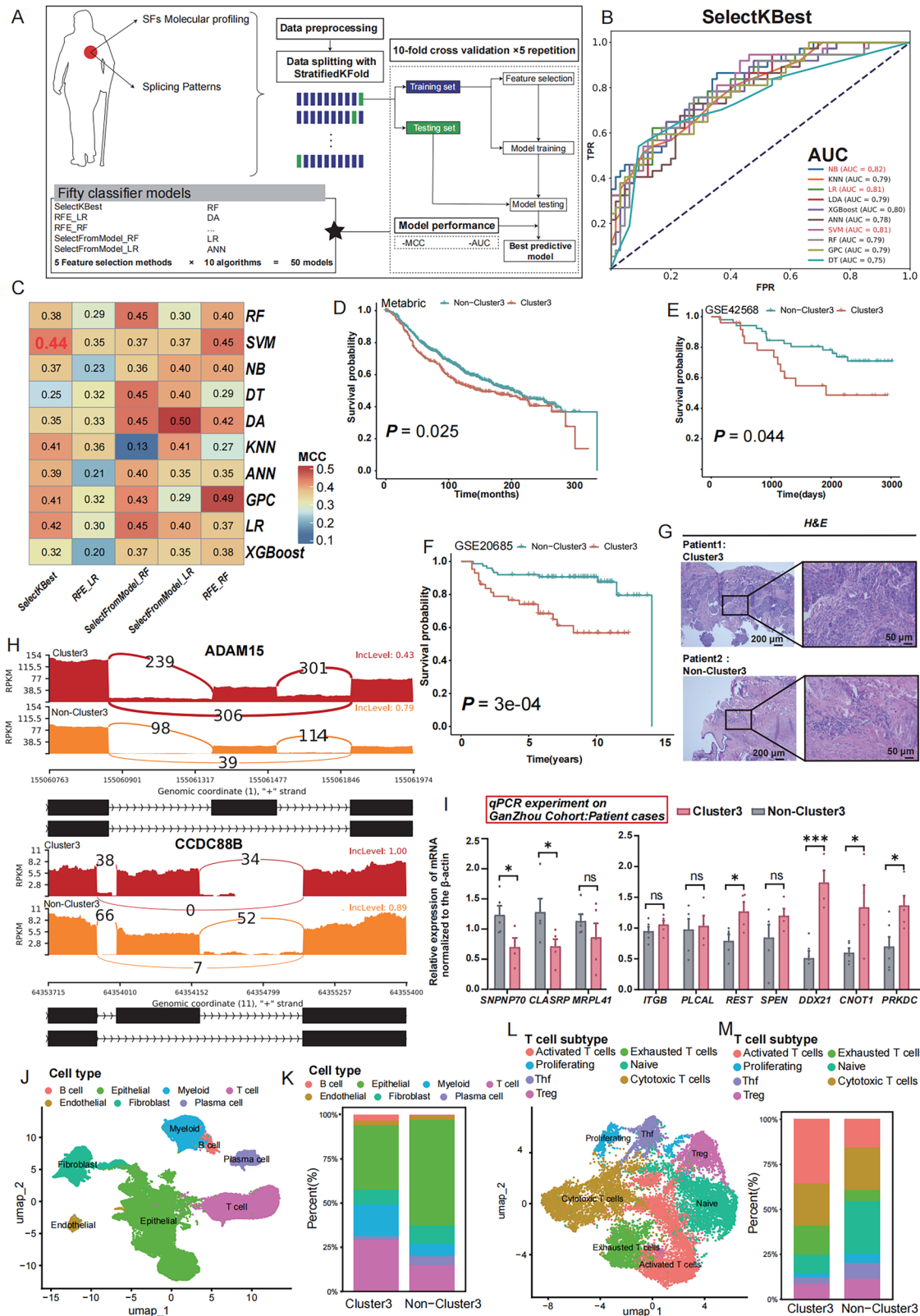
Paired tumor and adjacent normal samples were obtained from patients diagnosed with BRCA who underwent surgery at GanZhou People's Hospital. Total RNA was extracted from tumor tissues using TRIzol reagent (TaKaRa), and then converted to cDNA using the PrimeScript RT reagent kit (TaKaRa, containing gDNA Eraser).

### qPCR

The cDNA was then used for qPCR to detect the AS and gene expression of interested transcripts. qPCR was performed using SYBR Green PCR Master Mix on an Analytik Jena qTOWER apparatus. Transcript-specific primers were utilized (see Supplementary Table 4).

### RNA-seq analysis

RNA integrity and purity were assessed before library preparation. Following the manufacturer's instructions, sequencing libraries were created using the NEBNext Ultra RNA Library Prep Kit for Illumina, and sequenced using the Illumina NovaSeq 6000 platform with 150 bp paired-end reads with a depth of 20 million reads per sample. The SAM file was obtained by aligning the clean reads for each sample to the human reference genome (human: GRCh38) using HISAT2 (version 2.1.0) software with default parameters<sup>59</sup>, achieving an average mapping rate of 94%. The SAM file was then converted into a BAM file using SAMtools (version 1.4) with the 'view -bS' parameters. This BAM file was further sorted using SAMtools with the 'sort' parameter. Then featurecounts was used to quantify gene expression. Differentially variable splicing (AS) events among the defined groups were identified using rMATS (version 4.0)<sup>60</sup> and MISO (version 0.5.4)<sup>61</sup>. The primary function of rMATS is to identify and classify differential AS events by statistically comparing RNA-seq profiles from two or more experimental conditions. MISO employs a statistical model to estimate the expression levels of alternatively spliced exons and isoforms, with a key feature being its ability to assess the confidence of these estimates. For paired-end RNA-seq data, the model's performance is greatly enhanced by incorporating the



mRNA fragment length distribution, leading to more accurate quantification of AS.

For the FUSCC cohort, we quantified gene expression using salmon (version 1.4.0) and analyzed AS events to generate the PSI matrix using SUPPA2 (version 2.3), a tool known for rapid and accurate differential splicing analysis<sup>62</sup>.

For the TCGA cohort, we analyzed differential AS by using SpliceSeq<sup>63</sup> and SplAdder<sup>64</sup>. TCGA SpliceSeq is a web-based resource that provides a quick, user-friendly, highly visual resource for exploring the AS patterns of TCGA tumors. SplAdder is a graph-based AS toolbox that can integrate any number of RNA-Seq alignments and a given annotation file to enhance shared annotations based on RNA-Seq evidence. The shared enhanced

**Fig. 6 | Construction and Validation of the SelectKBest-SVM Classifier.** **A** Flow chart of machine learning classifier construction. **B** AUC values of ten models selected using SelectKBest combined with different machine learning algorithms in the test set. **C** Heatmap of MCC values for the five feature selection methods and ten models. **D–F** Kaplan–Meier survival curves showing survival rates of Cluster3 versus Non-Cluster3 using the constructed classifier in the TCGA cohort and other external datasets, with  $P$ -values  $< 0.05$  indicating statistical significance. **G** H&E-stained pathological slides were collected from eligible older adults breast cancer patients in the Ganzhou cohort, and a SelectKBest-SVM classifier was applied to distinguish between subtypes and display the results. Scale bar: 200  $\mu\text{m}$  (50  $\mu\text{m}$  in magnified images). **H** The sashimi plot shows the differential splicing events of ADAM15 and CCDC88B between Cluster3 (red) and Non-Cluster3 (yellow) patients in the Ganzhou older adults breast cancer patient cohort, analyzed using rMATS. The exon structure is shown at the bottom left, while the visualization of aligned exon reads per sample is shown at the top left, represented by RPKM. Curves connecting different splicing types indicate the read counts mapped to that region, with different colors representing samples from each group. The right side shows the expression values of

the corresponding variable splicing events in the samples. All RNA-seq data were aligned to the human reference genome (GRCh38). **I** qPCR experiments at the protein level on the ten RBPs that were eventually used for model construction. Data are presented as mean  $\pm$  SD,  $n = 6$  biologically independent samples per group. Statistical significance was determined using a two-tailed unpaired Student's  $t$ -test. ns, not significant ( $P > 0.05$ ); \* $P < 0.05$ ; \*\* $P < 0.01$ . **J** UMAP dimensionality reduction visualization showing cell clustering results of integrated single-cell RNA-seq data from breast cancer patients aged 65 and older. UMAP plot showing clustering of six cell populations obtained through marker annotation, including B cells, endothelial cells, epithelial cells, fibroblasts, myeloid cells, and T cells. **K** UMAP plot Show the type proportion of cells predicted to be Cluster3. **L** UMAP dimensionality reduction visualization showing the situation of seven groups of cells annotated through T cell markers, with different colors representing different cell subpopulations. **M** Bar graph showing the proportions of cells predicted by the classifier to be Non-Cluster3, Cluster3 within each T cell subpopulation, with rows representing different T cell subtypes and columns indicating enrichment scores; higher enrichment scores indicate a higher likelihood of the cell being in that subtype.

annotation graph can then be used to identify, quantify, and confirm AS events based on RNA-Seq data.

### H&E staining

Breast cancer tissue was fixed in formalin and embedded in paraffin. Tissue blocks were then sectioned (5  $\mu\text{m}$  thickness) and stained with H&E. For immunostaining, sections were dehydrated using graded alcohol and subjected to antigen retrieval by boiling in 10 mM sodium citrate for 20 min, referring to previous research<sup>65</sup>. A microscope was used to observe the distribution of immune infiltration in patients at 5 $\times$  and 20 $\times$  magnification.

### t-SNE clustering analysis and redefining RNA processing pattern subtypes

Clustering analysis of splicing data was performed using  $t$ -distributed stochastic neighbor embedding ( $t$ -SNE)<sup>66</sup>. The Rtsne package (version 0.17) were used for  $t$ -SNE analysis of 13,479 splicing events with default settings for major parameters, and ggplot2 (version 3.5.1) was utilized for visualization. For NMF clustering, we utilized the NMF package (version 0.27) to analyze 390 highly variable splicing events, classifying TCGA older adults breast cancer samples into three distinct splicing pattern subtypes. The NMF was performed using the Brunet algorithm and parameters were set as: maxIter = 500, rank = 3, nrun = 50, seed = 21<sup>67</sup>. Consensusmap method was used to evaluate cluster stability based on specified rank<sup>68</sup>.

### Immune cell infiltration levels in cancer patients

The mRNA expression of HLA family genes and immune checkpoints was quantified in each BRCA sample. The ESTIMATE algorithm was employed to estimate the stroma and immune cell content in malignant tumors<sup>69</sup>, calculating stromal and immune scores to infer the proportions of infiltrating stromal and immune cells in BRCA tissues. Related tumor immune and microenvironment scores were calculated using methods previously described<sup>70</sup>.

The ssGSEA algorithm was used to quantify the relative abundance of 28 tumor-infiltrating immune cell types in BRCA samples<sup>71,72</sup>. Enrichment scores were calculated to represent the relative abundance of tumor-infiltrating immune cells in BRCA specimens. These scores reflect the proportion of each immune cell type in the tumor microenvironment.

According to the previous study<sup>17</sup>, TILs of TCGA tumor tissues were quantified from H&E images through computational staining using a convolutional neural network trained to classify patches of images.

### RNA-binding protein (RBP) activity analysis

To identify potential RBPs that may regulate AS events in cancer, we proposed a three-step method integrating RBP expression and AS analysis<sup>73</sup>. First, we constructed a disturbance AS score for each splicing event  $i$ . The

disturbance level for a specific AS event was calculated as follows:

$$\Delta\text{PSI}_i = \text{PSI}_i - \frac{\sum_{K=1}^N \text{PSI}_{ik}}{N}$$

where  $N$  is the number of normal samples in specific cancer. Subsequently, we identified RBPs with transcriptome disturbances in cancer through  $t$ -tests. Assuming a correlation between the expression levels of these RBPs and their activity scores, we further calculated the Spearman correlation coefficient (SCC) between RBP expression levels and activity/inhibition scores, filtering for RBPs with a  $P$ -value less than 0.1 for further analysis. We employed the Benjamini-Hochberg (BH) method for multiple testing correction, identifying RBPs with adjusted  $P$ -values less than 0.05. Ultimately, RBPs associated with activity/inhibition scores were defined as active or inhibitory RBP regulatory factors.

### Survival analysis

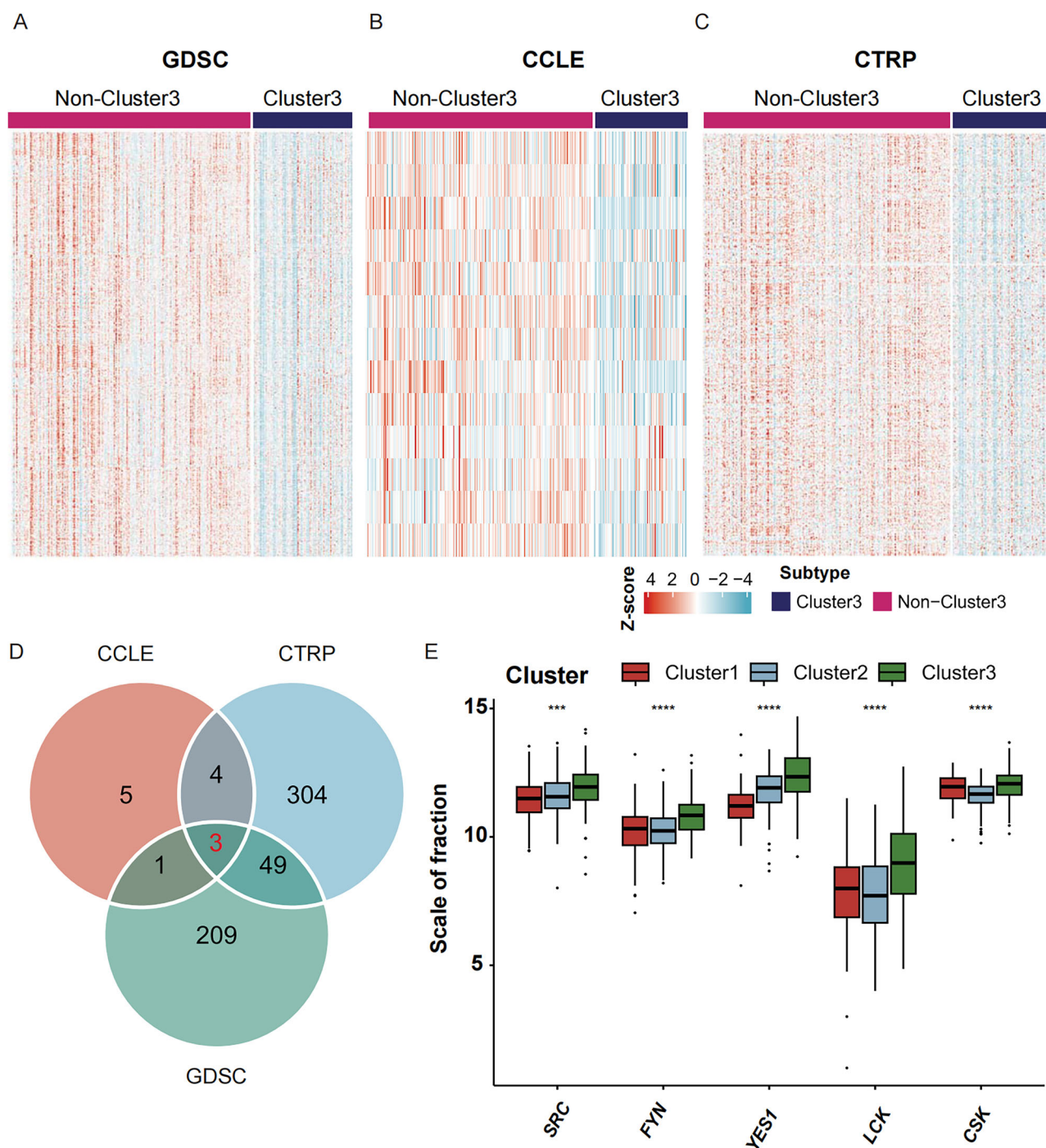
Kaplan–Meier survival curve analysis is a nonparametric statistical method used to estimate the survival function and show the change in survival probability over time. By grouping the survival time and calculating the survival probability within each group, the Kaplan–Meier curve shows the survival rate at different time points in a step-like decline, thus intuitively reflecting the survival status of the sample at different time points.

Survival analysis was conducted on patients with survival information in the collected cohorts. Univariate Cox analysis was used to assess the relationship between PSI values of AS events and OS in BRCA patients ( $P < 0.05$ ). The survfit function was used to fit the survival curve to explore the differences in survival time among the patients in each group, survival function “survdiff” of the R package survival (version 3.7-0) was used to analyze clinical information, and the difference in survival was determined using the log-rank test.

### Identification of abnormal as events and differentially expressed genes in breast cancer patients

Abnormal AS events in breast cancer patients were identified using the  $t$ -test. The  $P$ -values were corrected using the Benjamini-Hochberg (BH) method. AS events with  $|\text{Fold change}| \geq 1.2$  and adjusted  $P$ -values  $< 0.05$  were classified as abnormal. Differentially expressed genes were determined using the limma package (version 3.54.2) with a false discovery rate (FDR)  $\leq 0.05$  and  $|\text{Fold change}| \geq 2$ .

From the 390 HVS splicing events (Fig. 1I, S3A–B; Supplementary Data 5), we prioritized events meeting two criteria: Significant differential PSI versus normal tissue (Wilcoxon test  $P < 0.05$ ); Prognostic potential (top 10 genes by univariate Cox regression, representing 13 events). CD44 and NFATC4 events were excluded due to multi-exon spanning structures



**Fig. 7 | Classifier Prediction of Drug Sensitivity for Cluster3 and Non-Cluster3.** A–C Heatmap displaying drug sensitivity of Cluster3 and Non-Cluster3 patients based on the OncoPredict algorithm, converting gene expression of 340 older adults breast cancer patients into a drug sensitivity matrix according to different drug databases. **D** Venn diagram showing three drugs specifically sensitive to Cluster3

identified from the selected drug databases: Paclitaxel, Erlotinib, and Saracatinib (AZD0530). **E** Box plots showing the expression of Saracatinib drug target genes in each Cluster. Asterisks indicate *P*-values (Kruskal-Wallis test, \**P* < 0.05, \*\**P* < 0.01, \*\*\**P* < 0.001, \*\*\*\**P* < 0.0001). The box plots show the median (center line) and the 25th to 75th percentiles (box limits).

preventing specific primer design. SMARCA2 was excluded after failed PCR amplification. Final validation focused on seven genes for further analysis.

**GO and KEGG enrichment analysis**

Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) enrichment analyses were conducted using the R package clusterProfiler (version 4.6.2)<sup>74</sup>. The clusterProfiler package includes data from these two databases, allowing for the analysis of a given gene cluster and the visualization of gene functions. These analyses provided pathways

potentially associated with the differentially expressed genes, with a focus on biological processes (BP) in GO. The Benjamini corrected *P*-value < 0.05 of annotations was considered statistically significant. The *P*-value calculations are based on Fisher-exact test.

**Gene set enrichment analysis (GSEA)**

To further characterize and discuss the molecular biological features of the three clusters, we performed gene enrichment analysis on the gene expression matrices of these three clusters using the R package msigdb

(version 7.5.1) to obtain gene set enrichment scores for 50 pathways. We analyzed differences among the three clusters using limma and selected pathways with  $P < 0.05$ , visualizing the results using the heatmap function of the R package heatmap (version 1.0.12).

### Gene set variation analysis (GSVA)

Gene Set Variation Analysis (GSVA) was performed as a non-parametric and unsupervised method to estimate pathway activity changes in older adults breast cancer samples, using the c2.cp.kegg.v6.2 gene set retrieved from the Molecular Signatures Database (MSigDB)<sup>75</sup>. Significant pathways were filtered based on the criteria:  $\log_2|\text{fold-change}| > 0.2$  and adjusted  $P < 0.05$ .

### Somatic copy number variation (SCNV) analysis

SCNV data were downloaded from the TCGA GISTIC 2.0 portal (<http://software.broadinstitute.org/cancer/software/gistic>). We filtered for focal amplifications/deletions with a q-value  $< 0.25$  and minimum locus width of 1 Mb. For each AS-associated gene, we evaluated its copy number status using GISTIC-derived segment data, excluding genes with neutral copy number ( $\log_2$  ratio between -0.2 and 0.2).

### Constructing the machine learning prediction model

Using the TCGA database's 340 older adults breast cancer patients data for model construction, the entire model framework was built in Python (version 3.12.4) utilizing the following libraries: scikit-learn (version 1.5.1), numpy (version 1.16.4), scipy (version 1.3), and pandas (version 0.24.2). The data was split into training and testing sets in a 7:3 ratio. Feature selection was performed using SelectKBest, RFE\_LR, RFE\_RF, SelectFromModel\_RF and SelectFromModel\_LR. Subsequently, ten models were constructed based on the selected features: Random Forest (RF), Discriminant Analysis (DA), Gaussian Process (GPC), K-Nearest Neighbors (KNN), Logistic Regression (LR), Artificial Neural Network (ANN), Naïve Bayes (NB), eXtreme Gradient Boosting (XGBoost), Decision Tree (DT), Support Vector Machine (SVM). The optimal parameters were determined using 10-fold cross-validation and grid search, and the model's performance was evaluated using metrics, such as Area Under the Curve (AUC) and Matthews Correlation Coefficient (MCC). Thirteen models demonstrated satisfactory performance, with an area under AUC  $> 0.75$  and MCC  $> 0.4$ . Subsequently, the performance of the selected models were validated using independent external datasets.

### Analysis of isoform switching

To assess isoform switching between Cluster3 and Non-Cluster3, we employed the DEXSeq method, which identifies differentially utilized exonic bins across conditions. DEXSeq is a statistical method that uses generalized linear models to test for differential exon usage in RNA-seq data, offering reliable control of false discoveries by accounting for biological variation<sup>76</sup>. The magnitude of isoform switching was quantified using the difference in isoform fraction (dIF), computed as  $\text{IF}_{\text{treat}} - \text{IF}_{\text{control}}$ . This dIF metric serves as an effect size measure, analogous to fold change in traditional gene/isoform expression analyses. Statistically significant switching events were defined as those with  $|\text{dIF}| > 0.1$  (a conventional threshold) and FDR-adjusted  $P$ -value  $< 0.05$ . For gene-level differential expression, we applied thresholds of  $|\log_2$  fold change  $| > 2$  and  $Q$  value  $< 0.05$ .

Subsequently, we predicted the open reading frame (ORF) and nonsense-mediated decay (NMD) susceptibility of the identified isoforms by extracting transcript sequences from the reference genome based on their genomic coordinates. The nucleotide sequences of the ORFs were then translated into their corresponding amino acid sequences to infer potential protein-coding implications. The isoform switching analyses were performed using the IsoformSwitchAnalyzerR (version 1.14.1)<sup>77</sup>.

### Single-cell transcriptome analysis

Single-cell transcriptome data were processed for quality control and filtering using the R package Seurat (version 4.1.1). Low-quality cells were

filtered out based on the following criteria: fewer than 400 genes per cell, fewer than three cells per gene, over 25% mitochondrial gene content, or over 3% blood cell gene content. The R package DoubletFinder (version 4.1.1) was used to identify and remove doublets. The data were normalized using the SCTransform function, and dimensionality reduction was conducted using RunPCA, selecting the top 20 principal components for harmony integration. Clustering analysis was performed with a resolution parameter of 0.7, followed by UMAP for visualization. Cell types were determined based on the top differentially expressed genes (DEGs) and previously reported biologically relevant genes. B cells were identified by the expression of CD19, CD19A, and MS4A1; Endothelial cells were defined by high expression of PECAM1 and VWF; Epithelial cells predominantly expressed EPCAM, KRT19, and PROM1; Myeloid cells were characterized by high expression of CD68 and CD163; Fibroblast cells were identified by COL1A1, COL1A2, and DCN; Plasma cells expressed IGHG1, MZB1, and SDC1; T cells were defined by CD3D, CD3E, CD8A, and CD4A. The annotation of T cell subpopulations was determined based on previous literature<sup>78</sup>.

To predict cluster classification of single-cell sequencing samples, we performed similarity analysis on pseudobulk gene counts derived from single-cell gene expression data. Specifically, the sum of the expression of each gene of all cells belonging to this patient in the single-cell data of each patient is calculated, and then the mean operation is performed. Finally, the value calculated by each gene is taken as the pseudo-bulk RNA data of this patient<sup>79,80</sup>. Subsequently, the data for each patient is log-transformed and input into the previously established model to predict whether the patient's data belongs to Cluster3. Finally, the distribution of different cell types is observed to facilitate the comparison of differences between the two types at the single-cell level.

### Drug response prediction

Alterations in the cancer genome strongly influence the clinical response to therapy and are in many cases effective biomarkers of response to drug therapy. Data on tumor cell sensitivity to drugs and gene expression were obtained from the Cancer Cell Line Encyclopedia (CCLE)<sup>81</sup>, Genomics of Drug Sensitivity in Cancer (GDSC)<sup>82</sup>, and the Cancer Therapeutics Response Portal v2 (CTRPv2)<sup>83</sup>. We collected expression profile data for human cancer cell lines from CCLE. The half-maximal inhibitory concentration (IC50) was used to assess drug sensitivity. Information on drug responses in human cancer cell lines was queried from the CTRP (<https://www.cancerrxgene.org/>) and GDSC (<https://www.cancerrxgene.org/>) databases. The sensitivity of tumor samples to various drugs was predicted from gene expression profiles using the R package oncoPredict (version 1.2)<sup>84</sup>. The tumor cell drug sensitivity and response data obtained from GDSC served as training data to establish a ridge regression model based on gene expression and drug sensitivity in the GDSC database, which was then applied to other gene expression datasets to predict patient responses to clinical drug treatments. The IC50 values for each drug in TCGA samples were calculated using the calcPhenotype function from the R package to predict the sensitivity of TCGA samples to specific drugs.

### Statistics and reproducibility

Statistical analyses were conducted using R software (version 4.2.0). Comparisons between two groups in bioinformatics data were performed using the Wilcoxon test, while multiple groups were assessed using the Kruskal-Wallis test. For categorical variables, Chi-square tests or Fisher's exact tests were used as appropriate. For survival curve analyses, survival curves were plotted using the survminer package (version 0.4.9), followed by log-rank tests to compare survival curves. Asterisks indicate statistical significance as follows: \*  $P < 0.05$ ; \*\*  $P < 0.01$ ; \*\*\*  $P < 0.001$ .

The TCGA-BRCA cohort served as the discovery dataset for identifying age-related abnormal splicing events and defining molecular subtypes. After age-based stratification, we obtained 340 cancer samples and 30 adjacent normal samples in the older adult group ( $\geq 65$  years), and 184

cancer samples and 30 adjacent normal samples in the young group (< 45 years).

External validation of the constructed classifier was performed using multiple independent datasets with available clinical information for patients aged  $\geq 65$  years, including GSE42568 ( $n = 77$ ), GSE20685 ( $n = 118$ ), GSE25066 ( $n = 244$ ) from the GEO database, and the METABRIC cohort ( $n = 1033$ ).

Additional RNA-seq and experimental validation were conducted using 360 samples from the FUSCC cohort and 33 older adult breast cancer samples from the Ganzhou cohort.

Data visualization and additional analyses were performed using the complexheatmap (version 2.22.0), upsetR (version 0.4.9), oncoprint (version 0.4.9), cor.test (version 0.4.9), and venn.diagram (version 0.4.9) packages.

### Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

### Data availability

The datasets presented in this study can be found in various online repositories. All raw RNA seq data generated from this study can be accessed in the National Genomics Data Center (NGDC) under accession numbers HRA006220 and HRA009553. The datasets generated in this study have been deposited in public repositories. All raw RNA-seq data are available under controlled access at the National Genomics Data Center (NGDC) under accession numbers HRA006220 and HRA009553. The scRNA-seq data used in the published dataset<sup>34–36</sup> are available at NCBI GEO under accession number GSE161529 and GSE210347. Additional data related to this paper can be requested from the authors. All source data were provided as Supplementary Data 9.

### Code availability

The codes used to perform the analysis are available on GitHub: [https://github.com/HARI-Zhaolab/Age-Related\\_Aberrant\\_Splicing](https://github.com/HARI-Zhaolab/Age-Related_Aberrant_Splicing) and have been deposited in Zenodo with the DOI 10.5281/zenodo.17460917<sup>85</sup>. The specific computational tools and resources utilized in our analyses include: Seurat<sup>86</sup> version 4.1.1; DoubletFinder<sup>87</sup> version 4.1.1; scikit-learn<sup>88</sup> version 1.5.1. All analyses were conducted in the R and Python statistical environments.

Received: 30 January 2025; Accepted: 25 November 2025;

Published online: 16 December 2025

### References

- Bray, F. et al. Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J. Clin.* **74**, 229–263 (2024).
- Benz, C. C. Impact of aging on the biology of breast cancer. *Crit. Rev. Oncol. Hematol.* **66**, 65–74 (2008).
- Giaquinto, A. N. et al. Breast cancer statistics 2024. *CA Cancer J. Clin.* **74**, 477–495 (2024).
- Bray, F., McCarron, P. & Parkin, D. M. The changing global patterns of female breast cancer incidence and mortality. *Breast Cancer Res.* **6**, 229–239 (2004).
- Rule, J. et al. The impact of patient age on clinical decision-making in oncology. *J. Geriatr. Oncol.* **3**, S24–S25 (2012).
- Smith, B. D. et al. Improvement in breast cancer outcomes over time: are older women missing out? *J. Clin. Oncol.* **29**, 4647–4653 (2011).
- Scher, K. S. & Hurria, A. Under-representation of older adults in cancer registration trials: known problem, little progress. *J. Clin. Oncol.* **30**, 2036–2038 (2012).
- Tazi, J., Bakkour, N. & Stamm, S. Alternative splicing and disease. *Biochim Biophys. Acta* **1792**, 14–26 (2009).
- Koch, L. Alternative splicing: a thermometer controlling gene expression. *Nat. Rev. Genet.* **18**, 515 (2017).
- Liu, Q., Fang, L. & Wu, C. Alternative splicing and isoforms: from mechanisms to diseases. *Genes* **13**, 401 (2022).
- Chen, L.-L. The expanding regulatory mechanisms and cellular functions of circular RNAs. *Nat. Rev. Mol. Cell Biol.* **21**, 475–490 (2020).
- Rodríguez, S. A. et al. Global genome splicing analysis reveals an increased number of alternatively spliced genes with aging. *Aging Cell* **15**, 267–278 (2016).
- Angarola, B. L. & Anczuków, O. Splicing alterations in healthy aging and disease. *Wiley Interdiscip. Rev. RNA* **12**, e1643 (2021).
- Anczuków, O. et al. SRSF1-regulated alternative splicing in breast cancer. *Mol. Cell* **60**, 105–117 (2015).
- Hatami, R. et al. KLF6-SV1 drives breast cancer metastasis and is associated with poor survival. *Sci. Transl. Med.* **5**, 169ra12 (2013).
- Achour, C., Bhattarai, D. P., Groza, P., Román, Á.-C. & Aguilo, F. METTL3 regulates breast cancer-associated alternative splicing switches. *Oncogene* **42**, 911–925 (2023).
- Saltz, J. et al. Spatial organization and molecular correlation of tumor-infiltrating lymphocytes using deep learning on pathology images. *Cell Rep.* **23**, 181–193 (2018).
- Denkert, C. et al. Tumour-infiltrating lymphocytes and prognosis in different subtypes of breast cancer: a pooled analysis of 3771 patients treated with neoadjuvant therapy. *Lancet Oncol.* **19**, 40–50 (2018).
- Bates, G. J. et al. Quantification of regulatory T cells enables the identification of high-risk breast cancer patients and those at risk of late relapse. *J. Clin. Oncol.* **24**, 5373–5380 (2006).
- Liu, Y.-H., Jin, H.-Q. & Liu, H.-P. Identification of T-cell exhaustion-related gene signature for predicting prognosis in glioblastoma multiforme. *J. Cell Mol. Med.* **27**, 3503–3513 (2023).
- Morad, G., Helmink, B. A., Sharma, P. & Wargo, J. A. Hallmarks of response, resistance, and toxicity to immune checkpoint blockade. *Cell* **184**, 5309–5337 (2021).
- Li, B. et al. Comprehensive analyses of tumor immunity: implications for cancer immunotherapy. *Genome Biol.* **17**, 174 (2016).
- Becht, E. et al. Estimating the population abundance of tissue-infiltrating immune and stromal cell populations using gene expression. *Genome Biol.* **17**, 218 (2016).
- Finotello, F. et al. Molecular and pharmacological modulators of the tumor immune contexture revealed by deconvolution of RNA-seq data. *Genome Med.* **11**, 34 (2019).
- Aran, D., Hu, Z. & Butte, A. J. xCell: digitally portraying the tissue cellular heterogeneity landscape. *Genome Biol.* **18**, 220 (2017).
- van der Leun, A. M., Thommen, D. S. & Schumacher, T. N. CD8+T cell states in human cancer: insights from single-cell analysis. *Nat. Rev. Cancer* **20**, 218–232 (2020).
- Rigopoulos, C., Georgakopoulos-Soares, I. & Zaravinos, A. A multi-omics analysis of an exhausted T Cells' molecular signature in pan-cancer. *J. Pers. Med.* **14**, 765 (2024).
- Baslan, T. et al. Ordered and deterministic cancer genome evolution after p53 loss. *Nature* **608**, 795–802 (2022).
- Shkreta, L. & Chabot, B. The RNA Splicing Response to DNA Damage. *Biomolecules* **5**, 2935–2977 (2015).
- Liao, Y. et al. The cardiomyocyte RNA-binding proteome: links to intermediary metabolism and heart disease. *Cell Rep.* **16**, 1456–1469 (2016).
- Sebestyén, E. et al. Large-scale analysis of genome and transcriptome alterations in multiple tumors unveils novel cancer-relevant splicing networks. *Genome Res.* **26**, 732–744 (2016).
- Kechavarzi, B. & Janga, S. C. Dissecting the expression landscape of RNA-binding proteins in human cancers. *Genome Biol.* **15**, R14 (2014).
- Sundararaman, B. et al. Resources for the comprehensive discovery of functional RNA elements. *Mol. Cell* **61**, 903–913 (2016).
- Luo, H. et al. Pan-cancer single-cell analysis reveals the heterogeneity and plasticity of cancer-associated fibroblasts in the tumor microenvironment. *Nat. Commun.* **13**, 6619 (2022).

35. Pang, L. et al. Single-cell integrative analysis reveals consensus cancer cell states and clinical relevance in breast cancer. *Sci. Data* **11**, 289 (2024).
36. Pal, B. et al. A single-cell RNA expression atlas of normal, preneoplastic and tumorigenic states in the human breast. *EMBO J.* **40**, e107333 (2021).
37. Bentley, D. L. Coupling mRNA processing with transcription in time and space. *Nat. Rev. Genet.* **15**, 163–175 (2014).
38. Heintz, C. et al. Splicing factor 1 modulates dietary restriction and TORC1 pathway longevity in *C. elegans*. *Nature* **541**, 102–106 (2017).
39. Debès, C. et al. Ageing-associated changes in transcriptional elongation influence longevity. *Nature* **616**, 814–821 (2023).
40. Anders, C. K. et al. Breast carcinomas arising at a young age: unique biology or a surrogate for aggressive intrinsic subtypes? *J. Clin. Oncol.* **29**, e18–e20 (2011).
41. Anders, C. K. et al. Young age at diagnosis correlates with worse prognosis and defines a subset of breast cancers with shared patterns of gene expression. *J. Clin. Oncol.* **26**, 3324–3330 (2008).
42. Conaway, J. W. & Conaway, R. C. Transcription elongation and human disease. *Annu. Rev. Biochem.* **68**, 301–319 (1999).
43. Akcakanat, A., Zhang, L., Tsavachidis, S. & Meric-Bernstam, F. The rapamycin-regulated gene expression signature determines prognosis for breast cancer. *Mol. Cancer* **8**, 75 (2009).
44. Takahashi, S. et al. Prediction of breast cancer prognosis by gene expression profile of TP53 status. *Cancer Sci.* **99**, 324–332 (2008).
45. Liang, L. et al. Loss of ARHGDI1 expression is associated with poor prognosis in HCC and promotes invasion and metastasis of HCC cells. *Int. J. Oncol.* **45**, 659–666 (2014).
46. Lu, W. et al. Downregulation of ARHGDI1 contributes to human glioma progression through activation of Rho GTPase signaling pathway. *Tumour Biol.* **37**, 15783–15793 (2016).
47. Sdeor, E., Okada, H., Saad, R., Ben-Yishay, T. & Ben-David, U. Aneuploidy as a driver of human cancer. *Nat. Genet.* **56**, 2014–2026 (2024).
48. Buffa, F. M., Harris, A. L., West, C. M. & Miller, C. J. Large meta-analysis of multiple cancers reveals a common, compact and highly prognostic hypoxia metagene. *Br. J. Cancer* **102**, 428–435 (2010).
49. Ragnum, H. B. et al. The tumour hypoxia marker pimonidazole reflects a transcriptional programme associated with aggressive prostate cancer. *Br. J. Cancer* **112**, 382–390 (2015).
50. Winter, S. C. et al. Relation of a hypoxia metagene derived from head and neck cancer to prognosis of multiple cancers. *Cancer Res.* **67**, 3441–3449 (2007).
51. Cerami, E. et al. The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. *Cancer Discov.* **2**, 401–404 (2012).
52. de Bruijn, I. et al. Analysis and visualization of longitudinal genomic and clinical data from the AACR project GENIE biopharma collaborative in cBioPortal. *Cancer Res.* **83**, 3861–3867 (2023).
53. Gao, J. et al. Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. *Sci. Signal* **6**, pi1 (2013).
54. Clarke, C. et al. Correlating transcriptional networks to breast cancer survival: a large-scale coexpression analysis. *Carcinogenesis* **34**, 2300–2308 (2013).
55. Kao, K.-J., Chang, K.-M., Hsu, H.-C. & Huang, A. T. Correlation of microarray-based breast cancer molecular subtypes and clinical outcomes: implications for treatment optimization. *BMC Cancer* **11**, 143 (2011).
56. Hatzis, C. et al. A genomic predictor of response and survival following taxane-anthracycline chemotherapy for invasive breast cancer. *JAMA* **305**, 1873–1881 (2011).
57. Curtis, C. et al. The genomic and transcriptomic architecture of 2000 breast tumours reveals novel subgroups. *Nature* **486**, 346–352 (2012).
58. Jiang, Y.-Z. et al. Genomic and transcriptomic landscape of triple-negative breast cancers: subtypes and treatment strategies. *Cancer Cell* **35**, 428–440 (2019).
59. Raney, B. J. et al. The UCSC genome browser database: 2024 update. *Nucleic Acids Res.* **52**, D1082–D1088 (2024).
60. Shen, S. et al. rMATS: robust and flexible detection of differential alternative splicing from replicate RNA-Seq data. *Proc. Natl. Acad. Sci. USA* **111**, E5593–E5601 (2014).
61. Katz, Y., Wang, E. T., Airoidi, E. M. & Burge, C. B. Analysis and design of RNA sequencing experiments for identifying isoform regulation. *Nat. Methods* **7**, 1009–1015 (2010).
62. Trincado, J. L. et al. SUPPA2: fast, accurate, and uncertainty-aware differential splicing analysis across multiple conditions. *Genome Biol.* **19**, 40 (2018).
63. Ryan, M. et al. TCGASpliceSeq a compendium of alternative mRNA splicing in cancer. *Nucleic Acids Res.* **44**, D1018–D1022 (2016).
64. Markolin, P., Rättsch, G. & Kahles, A. Identification, quantification, and testing of alternative splicing events from RNA-Seq data using SplAdder. *Methods Mol. Biol.* **2493**, 167–193 (2022).
65. Zhao, L. et al. TDP-43 facilitates milk lipid secretion by post-transcriptional regulation of Btn1a1 and Xdh. *Nat. Commun.* **11**, 341 (2020).
66. Maaten, L. van der. Accelerating t-SNE using tree-based algorithms. *J. Mach. Learn. Res.* **15**, 3221–3245 (2014).
67. Gaujoux, R. & Seoighe, C. A flexible R package for nonnegative matrix factorization. *BMC Bioinforma.* **11**, 367 (2010).
68. Brunet, J.-P., Tamayo, P., Golub, T. R. & Mesirov, J. P. Metagenes and molecular pattern discovery using matrix factorization. *Proc. Natl. Acad. Sci. USA* **101**, 4164–4169 (2004).
69. Yoshihara, K. et al. Inferring tumour purity and stromal and immune cell admixture from expression data. *Nat. Commun.* **4**, 2612 (2013).
70. Wu, X. et al. Molecular classification of geriatric breast cancer displays distinct senescent subgroups of prognostic significance. *Mol. Ther. Nucleic Acids* **35**, 102309 (2024).
71. Şenbabaoğlu, Y. et al. Tumor immune microenvironment characterization in clear cell renal cell carcinoma identifies prognostic and immunotherapeutically relevant messenger RNA signatures. *Genome Biol.* **17**, 231 (2016).
72. Shen, Y., Peng, X. & Shen, C. Identification and validation of immune-related lncRNA prognostic signature for breast cancer. *Genomics* **112**, 2640–2646 (2020).
73. Li, J. et al. Alternative splicing perturbation landscape identifies RNA binding proteins as potential therapeutic targets in cancer. *Mol. Ther. Nucleic Acids* **24**, 792–806 (2021).
74. Wu, T. et al. clusterProfiler 4.0: a universal enrichment tool for interpreting omics data. *Innovation* **2**, 100141 (2021).
75. Subramanian, A. et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. USA* **102**, 15545–15550 (2005).
76. Anders, S., Reyes, A. & Huber, W. Detecting differential usage of exons from RNA-seq data. *Genome Res.* **22**, 2008–2017 (2012).
77. Vitting-Seerup, K. & Sandelin, A. IsoformSwitchAnalyzeR: analysis of changes in genome-wide patterns of alternative splicing and its functional consequences. *Bioinformatics* **35**, 4469–4471 (2019).
78. Ding, S. et al. Single-cell atlas reveals a distinct immune profile fostered by T cell-B cell crosstalk in triple negative breast cancer. *Cancer Commun.* **43**, 661–684 (2023).
79. Chaffin, M. et al. Single-nucleus profiling of human dilated and hypertrophic cardiomyopathy. *Nature* **608**, 174–180 (2022).
80. Liu, Z. et al. Spatiotemporal single-cell roadmap of human skin wound healing. *Cell Stem Cell* **32**, 479–498 (2025).
81. Pharmacogenomic agreement between two cancer cell line data sets. *Nature* **528**, 84–87 (2015).
82. Yang, W. et al. Genomics of drug sensitivity in cancer (GDSC): a resource for therapeutic biomarker discovery in cancer cells. *Nucleic Acids Res.* **41**, D955–D961 (2013).

83. Rees, M. G. et al. Correlating chemical sensitivity and basal gene expression reveals mechanism of action. *Nat. Chem. Biol.* **12**, 109–116 (2016).
84. Maeser, D., Gruener, R. F. & Huang, R. S. OncoPredict: an R package for predicting in vivo or cancer patient drug response and biomarkers from cell line screening data. *Brief Bioinform* **22**, bbab260 (2021).
85. CHEN, X. M. HARI-Zhaolab/Age-Related\_Aberrant\_Splicing: Age-Related\_Aberrant\_Splicing. Zenodo <https://doi.org/10.5281/zenodo.17460917> (2025).
86. Hao, Y. et al. Integrated analysis of multimodal single-cell data. *Cell* **184**, 3573–3587 (2021).
87. Stoeckius, M. et al. Cell Hashing with barcoded antibodies enables multiplexing and doublet detection for single cell genomics. *Genome Biol.* **19**, 224 (2018).
88. Pedregosa, F. et al. Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).

## Acknowledgements

We thank Drs. Ding Ma (Department of Breast Surgery, Fudan University Shanghai Cancer Center), Mengyuan Li (Department of Gynecology and Obstetrics, Chongqing General Hospital), and Zhenying Hu (School of Life Science, Nanchang University) for constructive suggestions. We thank the patients and researchers who participated in the TCGA, METABRIC, and GEO studies and provided publicly available data. We thank ChatGPT for providing us with valuable assistance in language polishing. This work was supported by the National Natural Science Foundation of China (82260488, 32360164 and 32200679), “Double Thousand Plan” of Jiangxi Province (jxsq2023101075), the Talent Project of Jiangxi (jxsq2023201015 and gpyc20250226), Jiangxi Provincial Natural Science Foundation (20224BAB205014, 20232BAB206116, 20224BAB216071, 20232ACB215001, 20224ACB206006 and 20242BAB25351), Beijing Science And Technology Innovation Medical Development Foundation (KC2023-JX-0082-06), Nanchang Medical College Innovation Team for Science and Technology Innovation Team for Environment and Health (NYTD202415), and Jiangxi Provincial Young Elite Scientists Sponsorship Program of Science (2025QT12).

## Author contributions

M.C. and H.K. designed the study and wrote the manuscript. L.Z. and Y.Z. designed, reviewed and revised the manuscript. M.C. and T.X. performed the major data analysis and interpreted the data. H.Y. performed clinical sample collection and data curation. S.Y. and Z.D. provided expertise on patients’ diagnosis for classification review. C.Y., F.Y., J.S., Y.F., H.X., S.F., L.L., J.L., Q.D., Y.Q., R.Y., H.L., W.L., Y.X., and Y.Z. provided experimental

technical assistance. C.Y., J.S., F.Y., H.X., Y.F., L.L., J.L., and S.F. offered critical feedback and helped shape the study and manuscript. Y.X., Q.D., R.Y., Y.Q., H.L., Y.Z., and W.L. helped perform the analysis with constructive discussions. All authors read and approved the final manuscript.

## Competing interests

The authors declare no competing interests.

## Additional information

**Supplementary information** The online version contains supplementary material available at <https://doi.org/10.1038/s42003-025-09330-y>.

**Correspondence** and requests for materials should be addressed to Yang Zou, Huozhong Yuan or Limin Zhao.

**Peer review information** *Communications Biology* thanks Erick Muciño-Olmos and the other, anonymous, reviewers for their contribution to the peer review of this work. Primary Handling Editors: Aylin Bircan and George Inglis. A peer review file is available.

**Reprints and permissions information** is available at <http://www.nature.com/reprints>

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

**Open Access** This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article’s Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2025