Biomolecules and Biomedicine
ISSN: 2831-0896 (Print) | ISSN: 2831-090X (Online)
Journal Impact Factor® (2025): 2.8
CiteScore® (2025): 4.7
www.biomolbiomed.com | blog.bjbms.org

SUPPLEMENTAL DATA

TFPI-high myofibroblast states and a meta-program-related five-gene prognostic signature in breast cancer

Yu Huang[1#], Xianmin Wu[1#], Xiangnan Li[1], Xueting Zhu[2], Xiaoyun Huang[1], An Huang[1], Haitao Liu[1], Yan Yang[1], Guanghui Zhou[1], Yining Zhang[1], Jingpei Long[3*], Qingjing Feng[1*]

[1]Department of Breast Surgery, Yiwu Maternal and Child Health Hospital, Yiwu, China;

[2]Department of Pathology, Yiwu Maternal and Child Health Hospital, Yiwu, China;

[3]Department of Breast Surgery, The Affiliated Women's Hospital, School of Medicine，Zhejiang University, Hangzhou, China.

*Correspondence to **Qingjing Feng**: ywrxwk@163.com and **Jingpei Long**: longjp06@163.com.

[#]Equally contributed to this work: Yu Huang and Xianmin Wu.

Full article is available at the following link: TFPI-high myofibroblast states and a meta-program-related five-gene prognostic signature in breast cancer

Table S1. Clinical and demographic information for the institutional cohort

Patient ID	Age (years)	Gender	Molecular subtype	Histological type	Pathological stage	Lymph node status	Tumor size (mm)	ER status	PR status	HER2 status
CR-001	52	F	Luminal A	IDC	IIA	Negative	18	Positive	Positive	Negative
CR-002	58	F	Luminal B	IDC	IIB	Positive	28	Positive	Positive	Negative
CR-003	45	F	HER2-enriched	IDC	IIIA	Positive	35	Negative	Negative	Positive
CR-004	62	F	Triple-Negative	IDC	IIIB	Positive	42	Negative	Negative	Negative
CR-005	55	F	Luminal A	ILC	IIA	Negative	22	Positive	Positive	Negative
CR-006	48	F	Luminal B	IDC	IIB	Positive	32	Positive	Negative	Negative
CK-001	51	F	Normal	Normal	-	-	-	-	-	-
CK-002	56	F	Normal	Normal	-	-	-	-	-	-
CK-003	49	F	Normal	Normal	-	-	-	-	-	-
CK-004	60	F	Normal	Normal	-	-	-	-	-	-
CK-005	54	F	Normal	Normal	-	-	-	-	-	-
CK-006	47	F	Normal	Normal	-	-	-	-	-	-

Detailed clinical and demographic information was collected for all 12 patients in the institutional cohort. This included 6 primary breast cancer tissues (CR group) and 6 adjacent non-tumor tissues (CK group). All patients underwent surgical resection and had not received any prior neoadjuvant or preoperative antitumor therapies. **Abbreviations**: CK, adjacent non-tumor tissue group; CR, primary breast cancer tissue group;

ER, estrogen receptor; F, female; HER2, human epidermal growth factor receptor 2; ID, identification; IDC, invasive ductal carcinoma; ILC, invasive lobular carcinoma; mm, millimeter; PR, progesterone receptor.

Table S2. TCGA cohort stratification and baseline clinical characteristics

Characteristic	**Training set (*n*=779)**	**Testing set (*n*=334)**	***p* value**
Age (years), median (IQR)	58 (48-67)	59 (49-68)	0.412
Pathological Stage, *n* (%)			0.287
Stage I	142 (18.2%)	68 (20.4%)	
Stage II	456 (58.5%)	189 (56.6%)	
Stage III	181 (23.2%)	77 (23.1%)	
Molecular Subtype, *n* (%)			0.156
Luminal A	312 (40.1%)	143 (42.8%)	
Luminal B	198 (25.4%)	91 (27.2%)	
HER2-enriched	89 (11.4%)	32 (9.6%)	
Triple-Negative	180 (23.1%)	68 (20.4%)	
Overall Survival Events, *n* (%)	156 (20.0%)	68 (20.4%)	0.891
Progression-Free Survival Events, *n* (%)	189 (24.3%)	81 (24.3%)	1

The TCGA-BRCA cohort (n = 1,113 tumor samples) was randomly divided into training (n = 779, 70%) and testing (n = 334, 30%) sets. This stratification was based on survival status to ensure balanced outcome distributions. Baseline clinical characteristics, including age, stage, and molecular subtype, were compared between the two sets. Wilcoxon rank-sum tests were used for continuous variables, and Chi-square tests for categorical variables. No statistically significant differences were found (all $p > 0.05$), confirming successful randomization. **Abbreviations**:

HER2, human epidermal growth factor receptor 2; IQR, interquartile range; TCGA, The Cancer Genome Atlas; TCGA-BRCA, The Cancer Genome Atlas Breast Invasive Carcinoma.

Table S3. Comprehensive marker gene panel for cell-type identification

Cell type	Primary markers	Secondary markers	Exclusion markers
Epithelial cells	EPCAM, KRT8, KRT18	CDH1, OCLN	KRT5, ACTA2
Myoepithelial cells	KRT5, ACTA2, TP63	MYL9, TAGLN	KRT8, EPCAM
Fibroblasts	COL1A1, COL3A1, FN1	PDGFRA, DCN	ACTA2, MYH11
Myofibroblasts	ACTA2, TAGLN, MYL9	COL1A1, COL3A1, FN1	MYH11, PDGFRB
Vascular smooth muscle	MYH11, ACTA2	TAGLN, MYL9	COL1A1, PDGFRB
Pericytes	PDGFRB, CSPG4	RGS5, ANPEP	ACTA2, MYH11
Endothelial cells	PECAM1, CDH5, VWF	ESAM, PLVAP	CD45, EPCAM
T cells	CD3D, CD3E, CD3G	TIGIT, XCL1, CCL5	CD19, CD14
Macrophages	CD14, CD68, CD163	CCL18, C1QA, C1QB	CD3D, CD19
Plasma cells	IGHA1, IGLC3, JCHAIN	CD27, CD38	CD3D, CD14

A comprehensive marker gene panel was utilized for robust identification and validation of the eight major cell populations within the GSE161529 single-cell RNA sequencing dataset. Primary markers facilitated initial cell-type assignment, secondary markers provided validation, and exclusion markers prevented misclassification. Myofibroblasts were distinctly differentiated from fibroblasts through the co-expression of contractile markers (ACTA2, TAGLN, MYL9) and extracellular matrix-related genes (COL1A1, COL3A1, FN1). **Abbreviations**: ACTA2, actin alpha 2, smooth muscle; ANPEP, alanyl aminopeptidase, membrane; C1QA, complement C1q A chain; C1QB, complement C1q B chain;

CCL5, C-C motif chemokine ligand 5; CCL18, C-C motif chemokine ligand 18; CD3D, CD3 delta subunit of T-cell receptor complex; CD3E, CD3 epsilon subunit of T-cell receptor complex; CD3G, CD3 gamma subunit of T-cell receptor complex; CD14, CD14 molecule; CD19, CD19 molecule; CD27, CD27 molecule; CD38, CD38 molecule; CD45, cluster of differentiation 45; CD68, CD68 molecule; CD163, CD163 molecule; CDH1, cadherin 1; CDH5, cadherin 5; COL1A1, collagen type I alpha 1 chain; COL3A1, collagen type III alpha 1 chain; CSPG4, chondroitin sulfate proteoglycan 4; DCN, decorin; EPCAM, epithelial cell adhesion molecule; ESAM, endothelial cell adhesion molecule; FN1, fibronectin 1; GSE, Gene Expression Omnibus Series; IGHA1, immunoglobulin heavy constant alpha 1; IGLC3, immunoglobulin lambda constant 3; JCHAIN, joining chain of multimeric immunoglobulin A and immunoglobulin M; KRT5, keratin 5; KRT8, keratin 8; KRT18, keratin 18; MYH11, myosin heavy chain 11; MYL9, myosin light chain 9; OCLN, occludin; PDGFRA, platelet-derived growth factor receptor alpha; PDGFRB, platelet-derived growth factor receptor beta; PECAM1, platelet and endothelial cell adhesion molecule 1; PLVAP, plasmalemma vesicle-associated protein; RGS5, regulator of G protein signaling 5; RNA, ribonucleic acid; TAGLN, transgelin; TIGIT, T-cell immunoreceptor with immunoglobulin and immunoreceptor tyrosine-based inhibitory motif domains; TP63, tumor protein p63; VWF, von Willebrand factor; XCL1, X-C motif chemokine ligand 1.

Table S4. Somatic mutation analysis stratified by molecular subtype

Gene	Luminal A	Luminal B	HER2-enriched	Triple-negative	*p* value (TFPI-high vs. TFPI-low)
TP53					
TFPI-high	22/156 (14.1%)	45/98 (45.9%)	28/45 (62.2%)	78/89 (87.6%)	0.0412
TFPI-low	18/156 (11.5%)	52/100 (52.0%)	26/44 (59.1%)	82/91 (90.1%)	

CDH1					
TFPI-high	34/156 (21.8%)	12/98 (12.2%)	3/45 (6.7%)	2/89 (2.2%)	0.0012
TFPI-low	18/156 (11.5%)	8/100 (8.0%)	2/44 (4.5%)	1/91 (1.1%)	
KMT2C					
TFPI-high	28/156 (17.9%)	15/98 (15.3%)	4/45 (8.9%)	3/89 (3.4%)	0.0057
TFPI-low	16/156 (10.3%)	10/100 (10.0%)	3/44 (6.8%)	2/91 (2.2%)	

Somatic mutation frequencies were stratified by TFPI expression level and molecular subtype. This analysis reveals that the association between TFPI expression and specific mutations (TP53, CDH1, KMT2C) is not fully explained by molecular subtype differences, though subtype-specific patterns are evident. For instance, TP53 mutations are more common in triple-negative and HER2-enriched tumors, irrespective of TFPI status. Conversely, CDH1 mutations show the strongest prevalence in luminal subtypes with high TFPI expression. **Abbreviations**: CDH1, cadherin 1; HER2, human epidermal growth factor receptor 2; KMT2C, lysine methyltransferase 2C; TFPI, tissue factor pathway inhibitor; TP53, tumor protein p53.

Table S5. Multivariable Cox regression analysis of the five-gene signature in the TCGA training cohort

Gene	Coefficient	Hazard ratio (HR)	HR 95% CI (Low)	HR 95% CI (high)	*p* value
TCN1	-0.122771848	0.884465432	0.703725202	1.111625813	0.292506728
FOXJ1	-0.334867115	0.71543315	0.561031553	0.912327639	0.006940145
PIGR	-0.18544733	0.830732589	0.659036872	1.047159369	0.11644486
SLAIN1	-0.374189322	0.687846675	0.525081793	0.901065424	0.006604248

TFPI	0.193693812	1.213724598	1.017156745	1.448279636	0.031660419

The table presents the coefficient, HR, 95% CI for the HR, and the corresponding p-value for each of the five signature genes (TCN1, FOXJ1, PIGR, SLAIN1, and TFPI). *p* values were derived from the multivariable Cox proportional hazards model. **Abbreviations**: CI, confidence interval; FOXJ1, forkhead box J1; HR, hazard ratio; PIGR, polymeric immunoglobulin receptor; SLAIN1, SLAIN motif family member 1; TCGA, The Cancer Genome Atlas; TCN1, transcobalamin 1; TFPI, tissue factor pathway inhibitor.

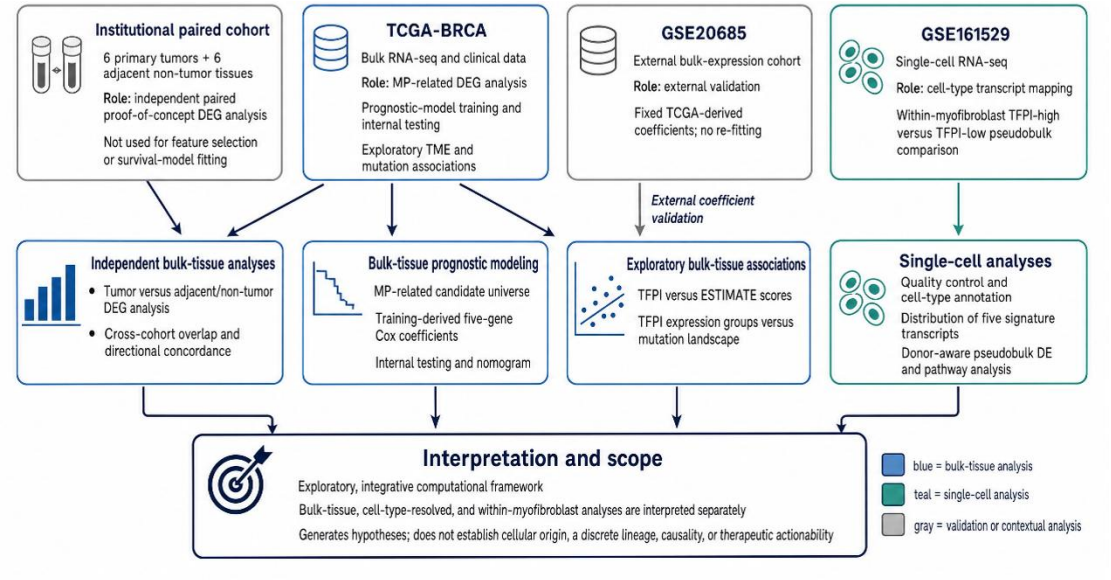


Figure S1. Dataset-specific analytical workflow and interpretive scope of the integrative bulk-tissue and single-cell study. The institutional paired cohort, comprising six primary tumours and six matched adjacent non-tumour tissues, was used exclusively for an independent paired proof-of-concept MP-related DEG analysis and did not contribute to prognostic feature selection, coefficient estimation, or survival-model fitting. Independent bulk-tissue analyses included separate tumour-versus-adjacent/non-tumour DEG analyses and assessment of cross-cohort overlap and directional concordance. TCGA-BRCA served as the primary bulk-tissue cohort for MP-related DEG analysis, training and internal testing of the five-gene Cox model, nomogram construction, and exploratory associations of bulk *TFPI* expression with ESTIMATE-derived TME scores and somatic mutation profiles. GSE20685 was used for external coefficient validation by applying the fixed TCGA-BRCA training-derived coefficients without re-fitting. GSE161529 was used for scRNA-seq QC and cell-type annotation, mapping of the five signature transcripts, and donor-aware pseudobulk comparison of operational TFPI-high and TFPI-low myofibroblast states followed by pathway analyses. Arrows indicate the progression from dataset-specific inputs to the corresponding analytical modules and final interpretive framework. Blue, teal, and grey denote bulk-tissue analyses, single-cell analyses, and validation or contextual analyses, respectively. Bulk-tissue, cell-type-resolved, and within-myofibroblast findings were interpreted at their respective biological scales. The integrated workflow is exploratory and hypothesis-generating and does not establish the cellular source of the bulk signal, a discrete TFPI-defined lineage, causality, or

therapeutic actionability. **Abbreviations**: DEG, differentially expressed gene; ESTIMATE, Estimation of STromal and Immune cells in MAlignant Tumour tissues using Expression data; MP, meta-program; QC, quality control; scRNA-seq, single-cell ribonucleic acid sequencing; TCGA-BRCA, The Cancer Genome Atlas Breast Invasive Carcinoma; TFPI, tissue factor pathway inhibitor; TFPI-high, upper-quartile TFPI-expressing myofibroblast state; TFPI-low, lower-quartile TFPI-expressing myofibroblast state; TME, tumour microenvironment.

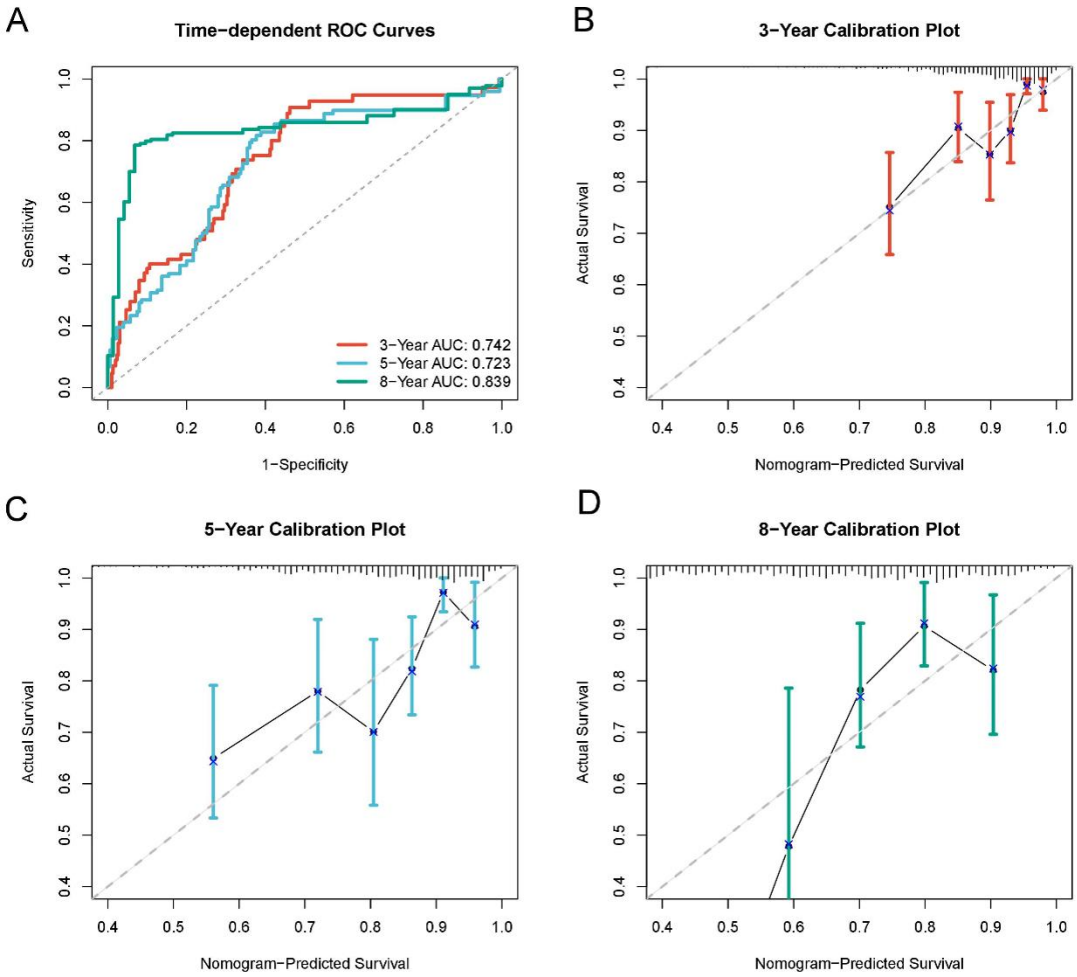


Figure S2. Time-dependent discrimination and calibration of the five-gene prognostic score in the TCGA-BRCA training cohort. (A) Time-dependent ROC curves evaluating the five-gene risk score for predicting 3-, 5-, and 8-year OS in the TCGA-BRCA training cohort (n = 779). The corresponding AUCs were 0.742, 0.723, and 0.839, respectively. The dashed diagonal line represents the no-discrimination reference. **(B–D)** Calibration plots comparing model-predicted and observed 3-year

(B), 5-year **(C)**, and 8-year **(D)** OS probabilities. The dashed 45° line represents perfect agreement between predicted and observed probabilities, points connected by solid lines represent calibration estimates, vertical error bars indicate 95% CIs, and rug marks along the upper margin show the distribution of predicted probabilities. These analyses evaluate discrimination and calibration within the training cohort and do not constitute independent validation. **Abbreviations**: AUC, area under the curve; CI, confidence interval; OS, overall survival; ROC, receiver operating characteristic; TCGA-BRCA, The Cancer Genome Atlas Breast Invasive Carcinoma.

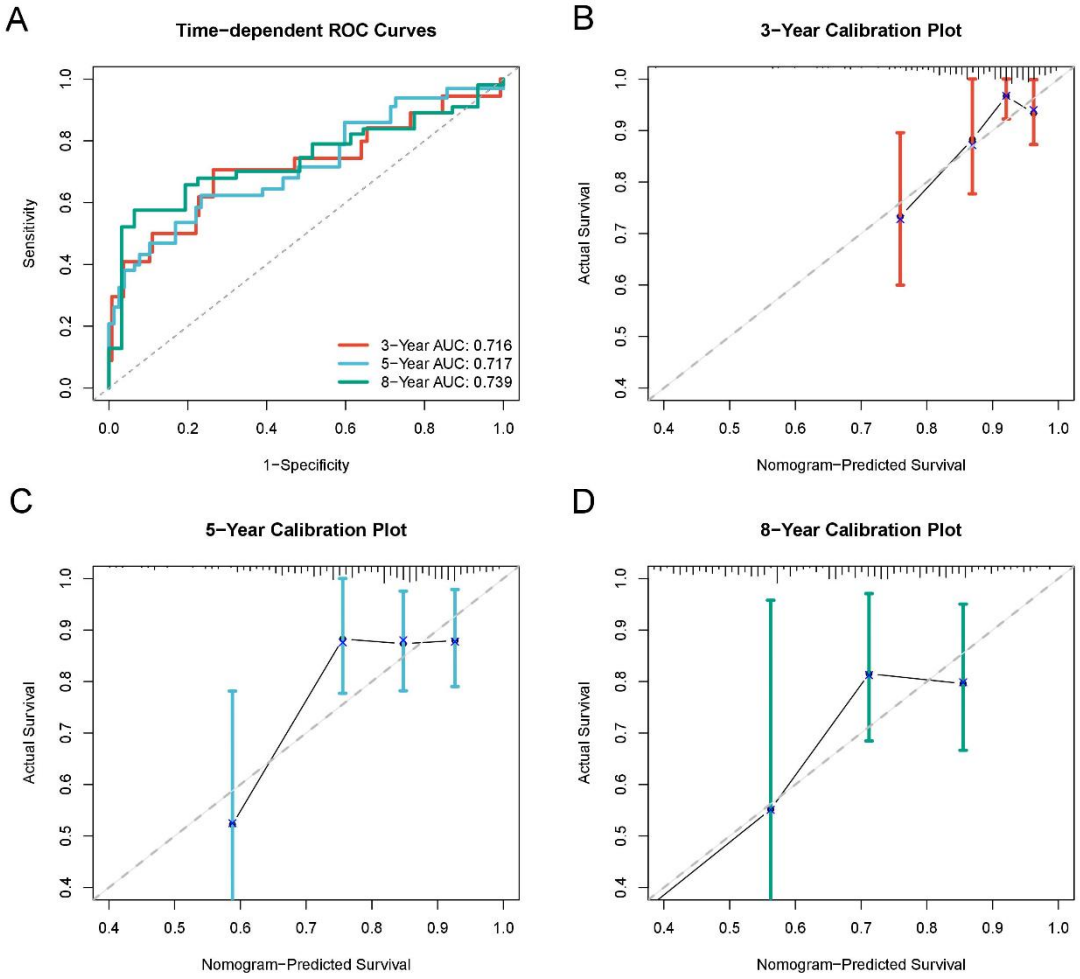


Figure S3. Time-dependent discrimination of the five-gene risk score and calibration of the integrated nomogram in the TCGA-BRCA testing cohort. (A) Time-dependent ROC curves evaluating the five-gene risk score for predicting 3-, 5-, and 8-year OS in the TCGA-BRCA testing cohort (n = 334). The corresponding AUCs were 0.716, 0.717, and 0.739, respectively. The five-gene score was calculated using the coefficients derived in the TCGA-BRCA training cohort without re-estimation. The dashed diagonal line represents the no-discrimination reference. **(B–D)**

Calibration plots comparing nomogram-predicted and observed 3-year **(B)**, 5-year **(C)**, and 8-year **(D)** OS probabilities. The dashed 45° line represents perfect agreement between predicted and observed probabilities, points connected by solid lines represent calibration estimates, vertical error bars indicate 95% CIs, and rug marks along the upper margin show the distribution of predicted probabilities. These analyses evaluate model performance in the internal testing cohort and do not constitute independent external validation. **Abbreviations**: AUC, area under the curve; CI, confidence interval; OS, overall survival; ROC, receiver operating characteristic; TCGA-BRCA, The Cancer Genome Atlas Breast Invasive Carcinoma.

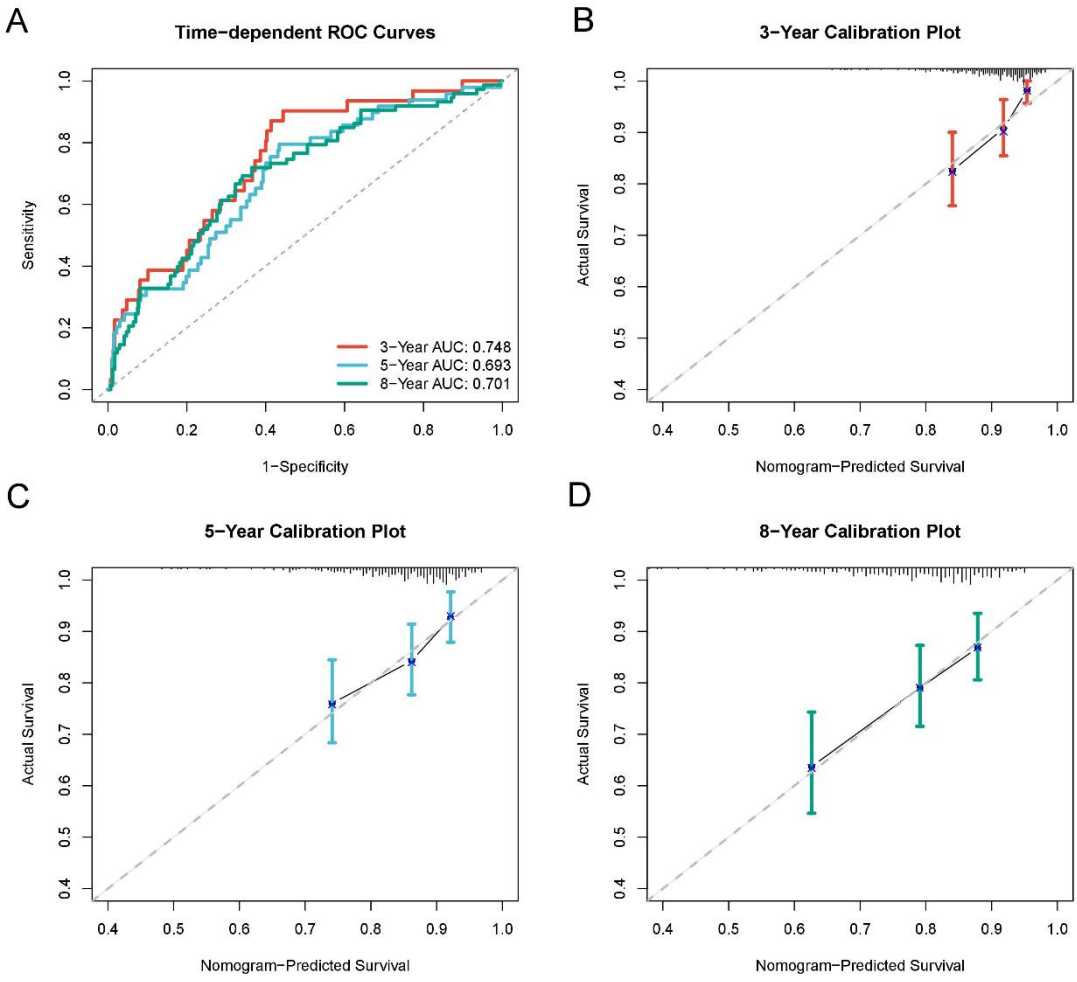


Figure S4. Time-dependent discrimination of the five-gene risk score and calibration of the integrated nomogram in the external GSE20685 cohort. (A) Time-dependent ROC curves evaluating the TCGA-BRCA training-derived five-gene

risk score for predicting 3-, 5-, and 8-year OS in the external GSE20685 cohort (n = 327). The fixed coefficients derived in the TCGA-BRCA training cohort were applied without re-estimation. The AUCs at 3, 5, and 8 years were 0.748, 0.693, and 0.701, respectively. The dashed diagonal line represents the no-discrimination reference. **(B–D)** Calibration plots comparing nomogram-predicted and observed 3-year **(B)**, 5-year **(C)**, and 8-year **(D)** OS probabilities in GSE20685. The dashed 45° line represents perfect agreement between predicted and observed probabilities, points connected by solid lines represent calibration estimates, vertical error bars indicate 95% CIs, and rug marks along the upper margin show the distribution of predicted probabilities. These analyses evaluate external discrimination and calibration but do not, by themselves, establish superiority over a clinical-only model or clinical utility. **Abbreviations**: AUC, area under the curve; CI, confidence interval; OS, overall survival; ROC, receiver operating characteristic; TCGA-BRCA, The Cancer Genome Atlas Breast Invasive Carcinoma.

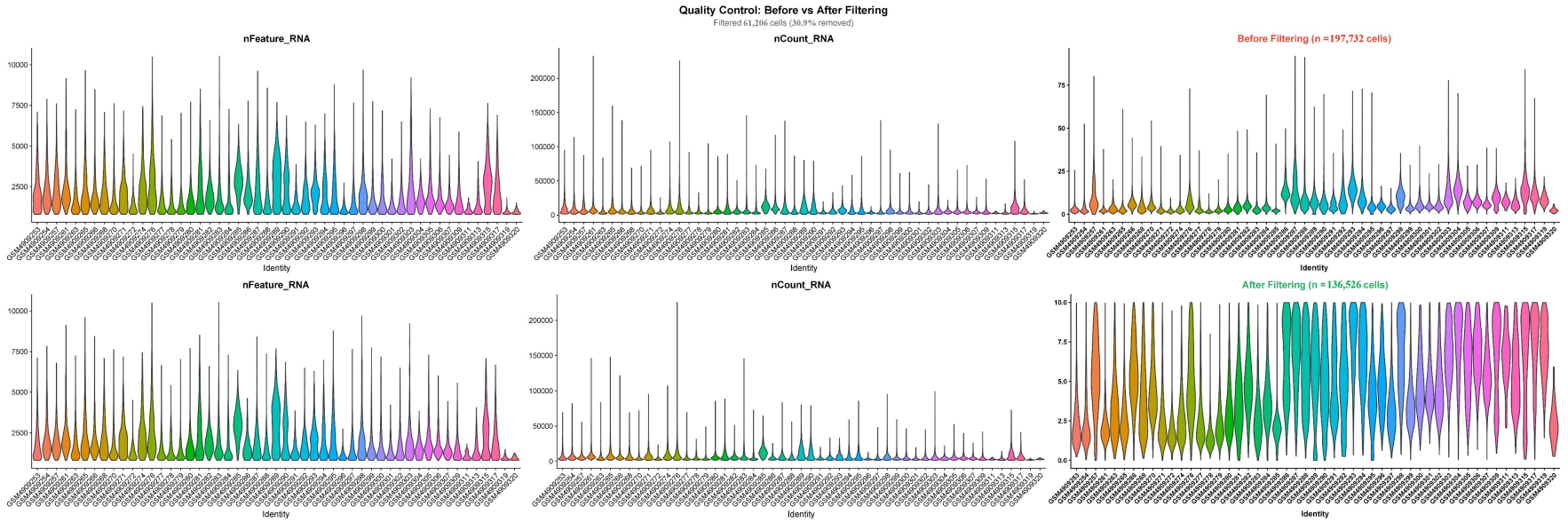


Figure S5. Sample-level scRNA-seq QC metrics before and after cell filtering in GSE161529. Violin plots show the sample-level distributions of nFeature_RNA, nCount_RNA, and percent.mt before filtering (top row; *n* = 197,732 cells/barcodes) and after filtering (bottom row; *n* = 136,526 cells). The displayed filtering step removed 61,206 cells/barcodes (≈30.9%), corresponding to the overall QC progression reported in the manuscript (197,732 → 191,906 → 185,080 after removal of 6,826 doublets → 136,526 final cells). Each violin represents an individual sample, and its width reflects the relative density of cells across the corresponding metric. QC thresholds included >800 detected genes per cell, <10% mitochondrial reads, and <40% ribosomal-gene expression. The plots document filtering-associated changes in cellular complexity, library size, and mitochondrial content across samples. **Abbreviations**: GSE, Gene Expression Omnibus Series; nCount_RNA, total RNA UMI count per cell; nFeature_RNA, number of genes detected per cell;

percent.mt, percentage of counts assigned to mitochondrial genes; QC, quality control; RNA, ribonucleic acid; scRNA-seq, single-cell RNA sequencing; UMI, unique molecular identifier.

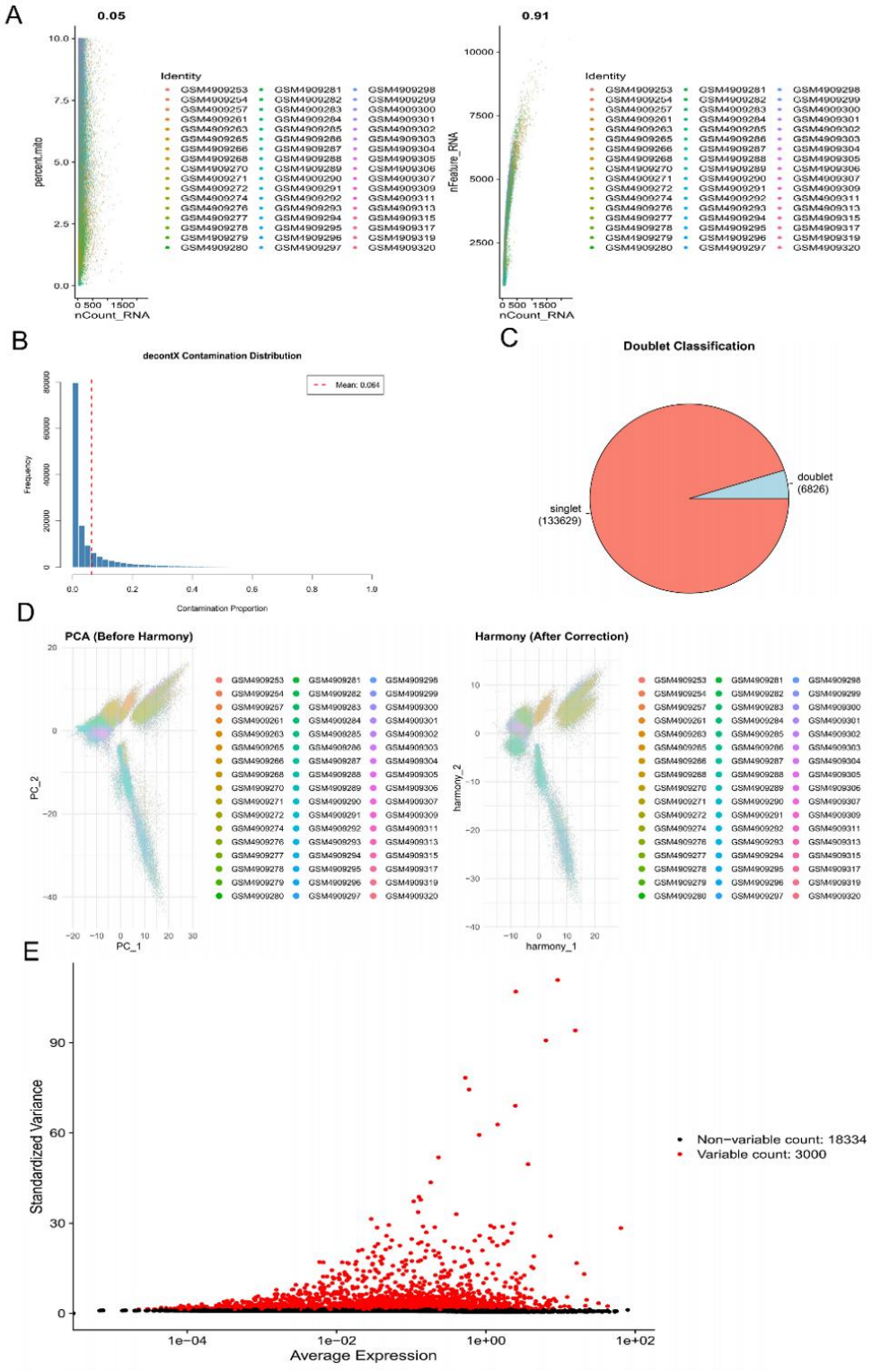


Figure S6. scRNA-seq preprocessing diagnostics, doublet classification, Harmony integration, and HVG selection in GSE161529. (A) Cell-level scatterplots of nCount_RNA versus percent.mito (left) and nCount_RNA versus nFeature_RNA (right), coloured by sample. Pearson's *r* values are shown above the respective plots (*r* = 0.05 and *r* = 0.91). **(B)** Histogram showing the distribution of

ambient RNA contamination proportions estimated using decontX. The red dashed line indicates the mean contamination proportion (0.064). decontX was used to quantify background RNA contamination and not as an independent cell-exclusion criterion. **(C)** Pie chart showing the singlet and doublet classifications generated using scDblFinder. Cells with a doublet probability >0.5 were classified as doublets. The current panel displays 133,629 singlets and 6,826 doublets. **(D)** Cell embeddings before and after Harmony integration, shown as the PCA embedding before correction (left) and Harmony-corrected coordinates (right), with cells coloured by sample. Each point represents an individual cell. **(E)** Mean–variance plot used for HVG selection by VST. Red points indicate the 3,000 HVGs selected for subsequent dimensionality reduction and graph-based clustering, whereas black points indicate 18,334 non-HVGs. Collectively, these panels document preprocessing and batch-integration diagnostics and do not constitute independent biological analyses. **Abbreviations**: GSE, Gene Expression Omnibus Series; HVG, highly variable gene; nCount_RNA, total RNA UMI count per cell; nFeature_RNA, number of genes detected per cell; PCA, principal component analysis; percent.mito, percentage of UMI counts assigned to mitochondrial genes; RNA, ribonucleic acid; scRNA-seq, single-cell RNA sequencing; UMI, unique molecular identifier; VST, variance-stabilizing transformation.

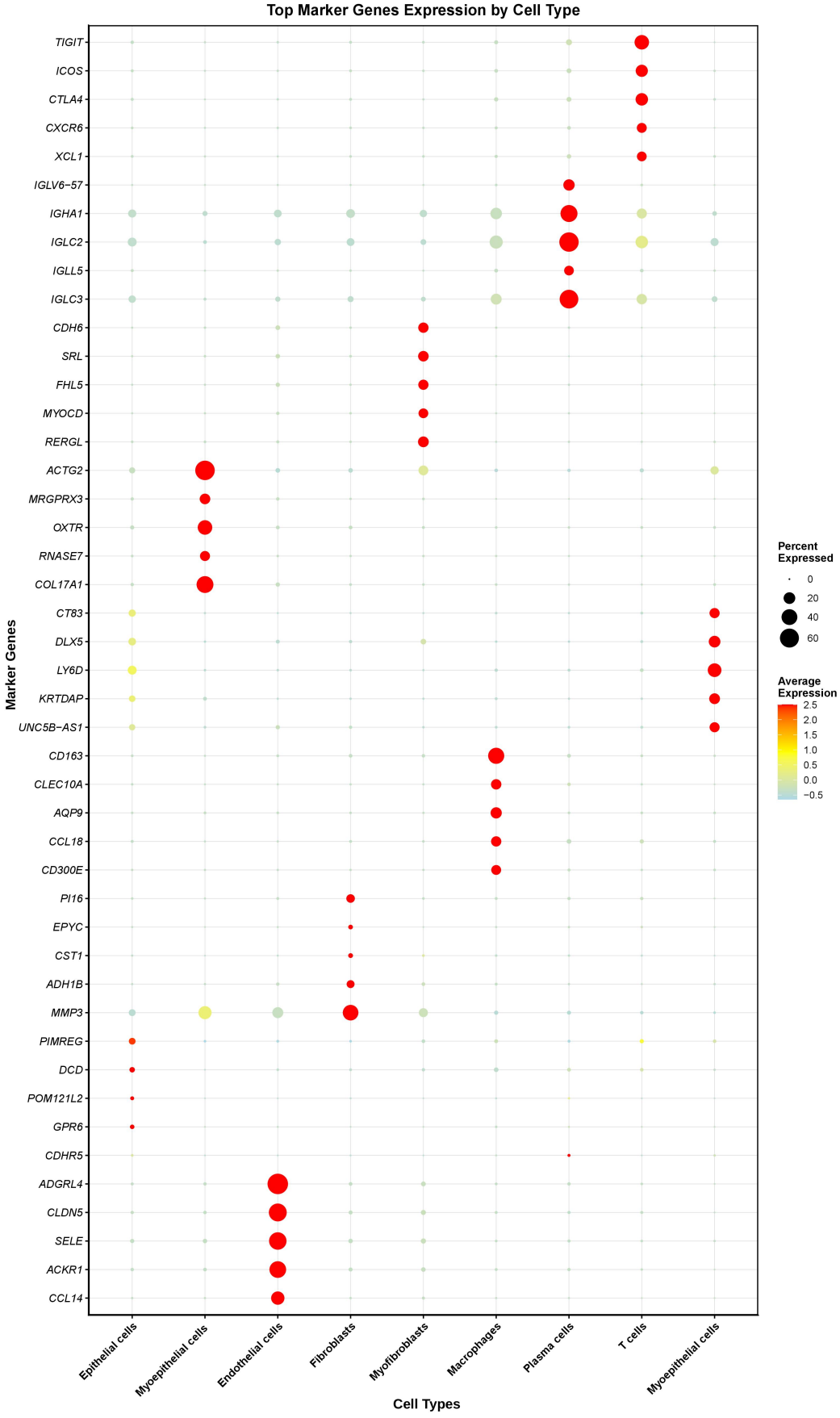

Top Marker Genes Expression by Cell Type
TIGIT
ICOS
CTLA4
CXCR6
XCL1
IGLV6−57
IGHA1
IGLC2
IGLL5
IGLC3
CDH6
SRL
FHL5
MYOCD
RERGL
ACTG2
MRGPRX3
OXTR
RNASE7
COL17A1
CT83
DLX5
LY6D
KRTDAP
UNC5B−AS1
CD163
CLEC10A
AQP9
CCL18
CD300E
PI16
EPYC
CST1
ADH1B
MMP3
PIMREG
DCD
POM121L2
GPR6
CDHR5
ADGRL4
CLDN5
SELE
ACKR1
CCL14
Marker Genes
Epithelial cells
Myoepithelial cells
Endothelial cells
Fibroblasts
Myofibroblasts
Macrophages
Plasma cells
T cells
Myoepithelial cells
Cell Types
Percent Expressed
0
20
40
60
Average Expression
2.5
2.0
1.5
1.0
0.5
0.0
−0.5

Figure S7. Marker-gene expression patterns supporting manual cell-type annotations in GSE161529. (A) Dot plot showing the expression of selected marker genes across the manually annotated cell populations in the integrated GSE161529 scRNA-seq dataset. The x-axis represents cell types and the y-axis represents marker genes. Dot size indicates the percentage of cells within each population with detectable expression, whereas colour represents scaled average expression. Distinct marker-expression patterns across epithelial cells, myoepithelial cells, endothelial cells, fibroblasts, myofibroblasts, macrophages, plasma cells, and T cells support the assigned transcriptional annotations. This visualization is descriptive and does not independently establish annotation accuracy, stable lineage relationships, or functional cell states. **Abbreviations**: GSE, Gene Expression Omnibus Series; scRNA-seq, single-cell RNA sequencing.