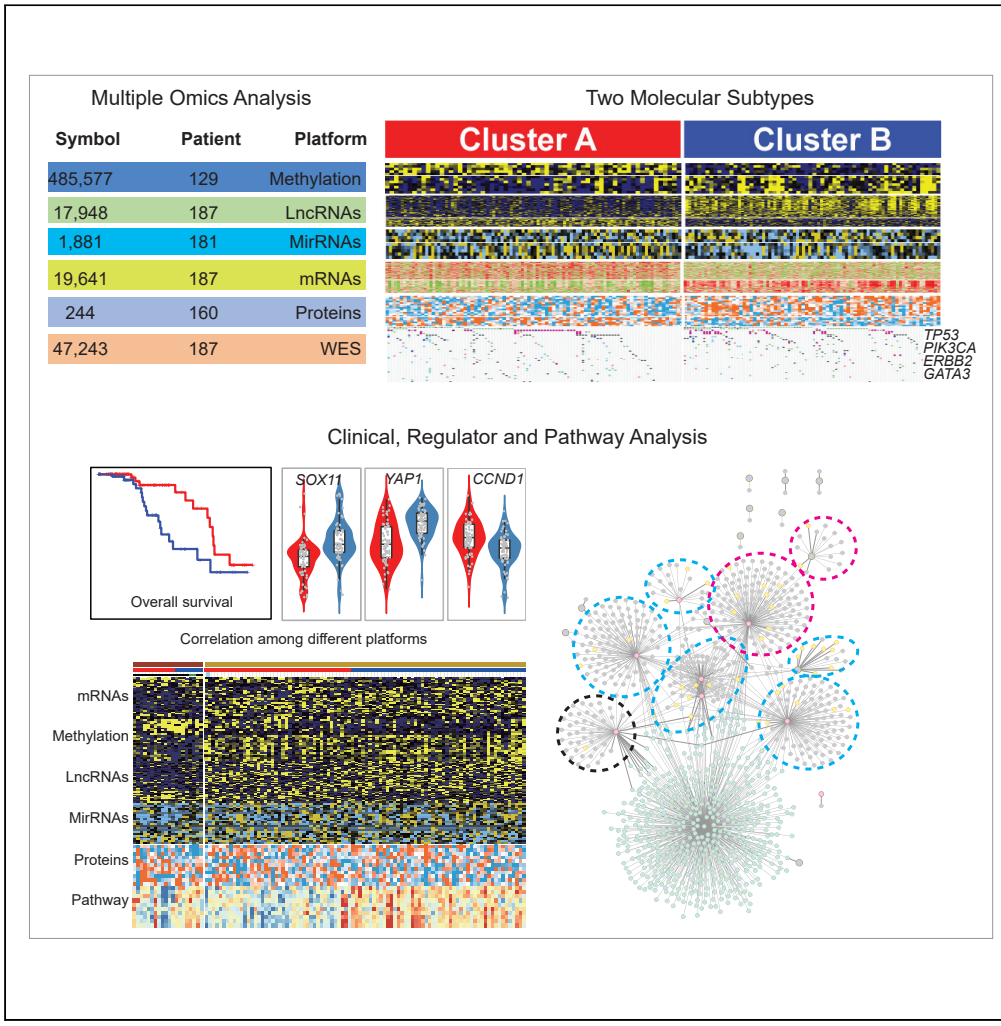
iScience

CellPress
OPEN ACCESS

Article

Integrative analysis identifies two molecular and clinical subsets in Luminal B breast cancer



Huina Wang, Bo Liu, Junqi Long, ..., Zhidong Liu, Shu Wang, Shuangtao Zhao

liuzhidong@bjxkyy.cn (Z.L.)
shuwang@pkuph.edu.cn (S.W.)
zst-1981@163.com (S.Z.)

Highlights

Multiplatform analysis of Luminal B breast cancer identified two subgroups

Poor-prognosis cluster is characterized by a hypermethylation of key genes

Whole-exome sequencing analysis identified significantly mutated genes

Cohorts with poor prognosis have unique molecular characteristics

Wang et al., iScience 26, 107466
September 15, 2023
https://doi.org/10.1016/j.isci.2023.107466

iScience

CellPress
OPEN ACCESS

Article

Integrative analysis identifies two molecular and clinical subsets in Luminal B breast cancer

Huina Wang,[1,6] Bo Liu,[2,6] Junqi Long,[1] Jiangyong Yu,[3] Xinchan Ji,[1] Jinmeng Li,[1] Nian Zhu,[1] Xujie Zhuang,[1] Lujia Li,[1] Yuhaoran Chen,[1] Zhidong Liu,[4,*] Shu Wang,[5,*] and Shuangtao Zhao[4,7,*]

SUMMARY

Comprehensive multiplatform analysis of Luminal B breast cancer (LBBC) specimens identifies two molecularly distinct, clinically relevant subtypes: Cluster A associated with cell cycle and metabolic signaling and Cluster B with predominant epithelial mesenchymal transition (EMT) and immune response pathways. Whole-exome sequencing identified significantly mutated genes including *TP53*, *PIK3CA*, *ERBB2*, and *GATA3* with recurrent somatic mutations. Alterations in DNA methylation or transcriptomic regulation in genes (*FN1*, *ESR1*, *CCND1*, and *YAP1*) result in tumor microenvironment reprogramming. Integrated analysis revealed enriched biological pathways and unexplored druggable targets (cancer-testis antigens, metabolic enzymes, kinases, and transcription regulators). A systematic comparison between mRNA and protein displayed emerging expression patterns of key therapeutic targets (CD274, YAP1, AKT1, and CDH1). A potential ceRNA network was developed with a significantly different prognosis between the two subtypes. This integrated analysis reveals a complex molecular landscape of LBBC and provides the utility of targets and signaling pathways for precision medicine.

INTRODUCTION

Breast cancer is the most common malignant tumor in women that poses a serious threat to women's life and health. According to the World Health Organization (WHO), the incidence of breast cancer reached 2.06 million in 2020 worldwide, replacing lung cancer as the "world's leading cancer".[1] And China has the largest number of cases, with more than 410,000 breast cancer patients currently. In the past 10 years, the incidence of breast cancer in China has increased by 3–4% per year on average. Therefore, exploring the pathogenesis of breast cancer and developing precise screening and treatment methods can have a significant impact on improving women's health.

Breast cancer is normally divided into four molecular subtypes, including HER2 positive, triple negative, Luminal A, and Luminal B, based on immunohistochemistry. Among them, Luminal breast cancer is defined as a hormone receptor (ER and/or PR) positive breast cancer, which can be divided into Luminal A and Luminal B according to the expression of HER2 and Ki67. Luminal A is HER2 negative breast cancer with Ki67 < 14%, while the rest are Luminal B subtype. As reviewed by Metzger-Filho, O et al.,[2] Luminal B breast cancer (LBBC) accounts for a higher proportion (40%) of all breast cancer subtypes, with complex clinicopathological features, such as large mass, higher chance of lymph node involvement, low grade of histological differentiation, and relative insensitivity to endocrine therapy and, therefore, with worse prognosis compared to the Luminal A subtype.[3] Furthermore, even within the LBBC category, prognosis varies greatly due to the high grade of disease heterogeneity. For example, some LBBC patients portend a similar prognosis with HER2 positive and triple negative breast cancers.[4] Fortunately compared with other types, LBBC has higher specificity in clinical treatment. The expert consensus in the 12th St. Gallen International Breast Cancer Meeting (2011) emphasized that endocrine therapy combined with chemotherapy could be considered to treat patients with a high expression of Ki67 in LBBC. The selection of treatment regimens depends on various factors such as the patient's hormone receptor expression level, risk factors, and other relevant clinical factors. The high expression of HER2 and Ki67 in LBBC and the insensitivity to endocrine therapy

[1]School of Software Engineering, Faculty of Information Technology, Beijing University of Technology, Beijing 100124, China

[2]School of Mathematical and Computational Sciences, Massey University, Palmerston North 4472, New Zealand

[3]Department of Medical Oncology, Beijing Hospital, National Center of Gerontology, Institute of Geriatric Medicine, Chinese Academy of Medical Sciences, Beijing 100730, China

[4]Department of Thoracic Surgery, Beijing Tuberculosis and Thoracic Tumor Research Institute/Beijing Chest Hospital, Capital Medical University, Beijing 101149, China

[5]Breast Disease Center, Peking University People's Hospital, Peking University, Beijing 100044, China

[6]These authors contributed equally

[7]Lead contact

*Correspondence: liuzhidong@bjxkyy.cn (Z.L.), shuwang@pkuph.edu.cn (S.W.), zst-1981@163.com (S.Z.)

https://doi.org/10.1016/j.isci.2023.107466

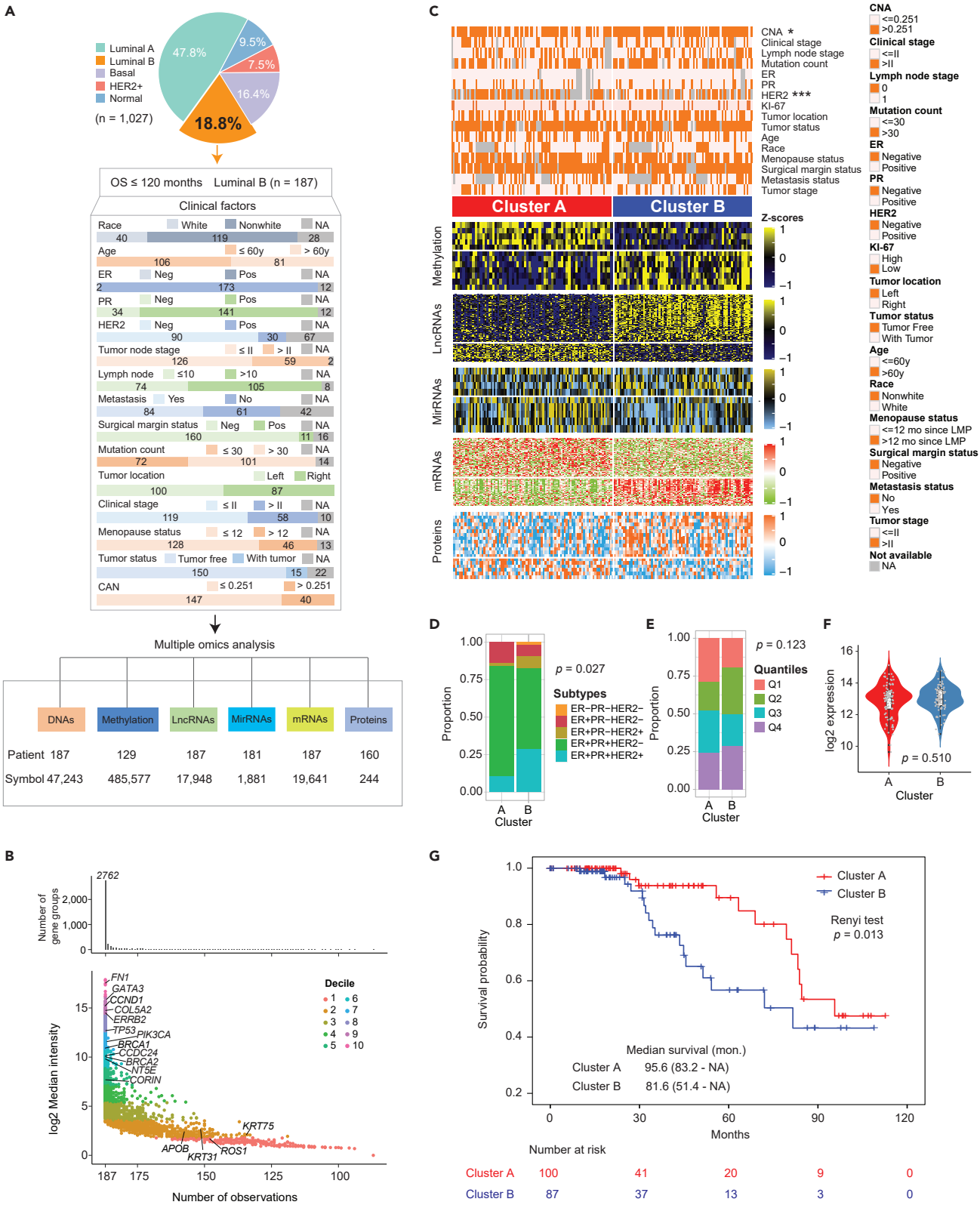

A
Luminal A
Luminal B
Basal
HER2+
Normal
(n = 1,027)
47.8%
9.5%
7.5%
16.4%
18.8%
OS ≤ 120 months Luminal B (n = 187)
Clinical factors
Race White Nonwhite NA
40 119 28
Age ≤ 60y > 60y
106 81
ER Neg Pos NA
2 173 12
PR Neg Pos NA
34 141 12
HER2 Neg Pos NA
90 30 67
Tumor node stage ≤ II > II NA
126 59 2
Lymph node ≤10 >10 NA
74 105 8
Metastasis Yes No NA
84 61 42
Surgical margin status Neg Pos NA
160 11 16
Mutation count ≤ 30 > 30 NA
72 101 14
Tumor location Left Right
100 87
Clinical stage ≤ II > II NA
119 58 10
Menopause status ≤ 12 > 12 NA
128 46 13
Tumor status Tumor free With tumor NA
150 15 22
CAN ≤ 0.251 > 0.251
147 40
Multiple omics analysis
DNAs
Methylation
LncRNAs
MirRNAs
mRNAs
Proteins
Patient 187 129 187 181 187 160
Symbol 47,243 485,577 17,948 1,881 19,641 244
B
2762
Number of gene groups
2,000
1,000
0
log2 Median intensity
FN1
GATA3
CCND1
COL5A2
ERRB2
TP53
PIK3CA
BRCA1
CCDC24
BRCA2
NT5E
CORIN
KRT75
APOB
KRT31
ROS1
Decile
1 2 3 4 5 6 7 8 9 10
15
10
5
0
187 175 150 125 100
Number of observations
C
CNA *
Clinical stage
Lymph node stage
Mutation count
ER
PR
HER2 ***
KI-67
Tumor location
Tumor status
Age
Race
Menopause status
Surgical margin status
Metastasis status
Tumor stage
Cluster A
Cluster B
Methylation
LncRNAs
MirRNAs
mRNAs
Proteins
Z-scores
1
0
−1
CNA
<=0.251
>0.251
Clinical stage
<=II
>II
Lymph node stage
0
1
Mutation count
<=30
>30
ER
Negative
Positive
PR
Negative
Positive
HER2
Negative
Positive
KI-67
High
Low
Tumor location
Left
Right
Tumor status
Tumor Free
With Tumor
Age
<=60y
>60y
Race
Nonwhite
White
Menopause status
<=12 mo since LMP
>12 mo since LMP
Surgical margin status
Negative
Positive
Metastasis status
No
Yes
Tumor stage
<=II
>II
Not available
NA
D
Proportion
1.00
0.75
0.50
0.25
0.00
p = 0.027
Subtypes
ER−PR−HER2−
ER+PR−HER2−
ER+PR−HER2+
ER+PR+HER2−
ER+PR+HER2+
A B
Cluster
E
p = 0.123
Quantiles
Q1
Q2
Q3
Q4
F
log2 expression
16
14
12
10
p = 0.510
G
Survival probability
1.0
0.8
0.6
0.4
0.2
Cluster A
Cluster B
Renyi test
p = 0.013
Median survival (mon.)
Cluster A 95.6 (83.2 - NA)
Cluster B 81.6 (51.4 - NA)
0 30 60 90 120
Months
Number at risk
Cluster A 100 41 20 9 0
Cluster B 87 37 13 3 0

Figure 1. Proteogenomic landscape of TCGA LBBC

(A) Study overview showing the clinical characteristics of LBBC patients enrolled in this cohort (n = 187) and the number of samples with whole-genome sequencing, DNA methylation data, RNA-seq including lncRNAs, miRNAs and mRNAs, and reverse phase protein array (RPPA) data (including phosphorylation data). The data were tested for batch effect correction and missing values were processed by removing genes that were missing more than 70% of all samples.

(B) Distribution of mRNA. The median value of mRNA expression in all samples was used to measure the intensity of mRNAs and the intensity interval of all mRNAs was divided into 10 equal parts, so that the intensity of all genes could be mapped to 1–10. The top bar graph shows the total counts of mRNAs contained in the different numbers of samples, which can be seen to contain 2,762 genes quantified in the 187 LBBC samples. The scatterplot at the bottom shows the distribution of genes of different intensities and marks the names of genes that are associated with cancer or are biologically important.

(C) Unsupervised hierarchical clustering of 187 samples (k = 2) using the top 1,000 variable genes. Clinical covariatare shown in the heatmap above. The DEGs were performed within RNA expression values for either mRNA (n = 537), miRNA (n = 9), or lncRNA (n = 153), within protein expression values (n = 19) and DNA methylation values (n = 26,284) in the whole LBBC cohort, and the names of cancer-related genes were labeled on the right. Fisher's exact test: * <0.05, ** <0.01, *** <0.001.

(D) The different subtypes(i.e., ER-PR-HER2-, ER + PR-HER2-, ER + PR-HER2+, ER + PR + HER2-, and ER + PR + HER2+) status between cluster A and B. chi-square test, p = 0.027.

(E) Four quantiles (i.e., Q1, Q2, Q3 and Q4) of Ki-67 mRNA expression between cluster A and B. chi-square test, p = 0.123.

(F) Violin plot shows Ki-67 mRNA expression between Clusters A and B. *t* test, p = 0.510.

(G) Differences in patient overall survival between the two Luminal B subtypes (log rank p value). See also Figure S1 and Tables S1, S2, S3, and S4.

were the reasons for the lower survival rate and poor prognosis compared with Luminal A breast cancer. Blows et al. analyzed 12 studies involving 10,159 LBBC patients with poor prognosis within 5 years to diagnosis.[5] Generally, LBBC has complex clinical characteristics and poor therapeutic effects, therefore, it is very important to use biomarkers to refine the molecular classification of LBBC and explore an untried way to achieve a precise treatment and improve its prediction accuracy.

In order to address these challenges, in this paper, the LBBC patients are classified into Cluster A (cell cycle-enriched group) and Cluster B (EMT and immune-related group) according to gene expression data to maximize the differences in prognosis within these two groups, and then deep excavation is performed based on tumor multi-omics data in two clusters. The integrative proteogenomic analysis revealed innovative therapeutic targets in signaling proteins, metabolic enzymes, kinases, and cancer testis antigens for LBBC treatment. This study provided a rich source characterizing proteogenomics of LBBC, and further informed strategies to target LBBC vulnerabilities.

RESULTS

The multiple omics classification of LBBC

To understand the workflow of biological information in LBBC, we obtained 187 LBBC samples from 1,027 breast cancer samples in the TCGA database,[6] in which the subtype was defined based on the PAM50 classification system (see Figure 1A). Tables S1, S2, and S3 provide a summary of the clinical and pathological characteristics enrolled in this study. A total of 182 (97%) patients were treated with surgery, with a median follow-up survival of 23.92 (95%CI: 20.24–32.03) months (see Table S1). LBBC patients in the TCGA database with initial diagnoses underwent array-based copy number aberration (CNA) profiling and whole-genome sequencing analysis to detect genomic alterations, RNA sequencing (RNA-seq) analysis to detect the expression of lncRNAs, miRNAs, and mRNAs, DNA methylation analysis to evaluate epigenome, and reverse phase protein array (RPPA) analysis to quantify the proteins expression.

We processed the TCGA data to address missing values by removing genes with missing expression in more than 70% of the samples. This resulted in 16,875 mRNAs out of 19,641. We then identified minimal batch effects through PCA analysis (see Figure S1A). Of these, 4,248 DEGs were filtered between normal and tumor for the subsequent analysis (see Figure S1B). A total of 2,762 gene groups were detected and quantified in all 187 samples (see Figure 1B, top), including those corresponding to classic breast cancer-associated genes like the cell cycle gene *CCND1*, targeted therapy gene *ERBB2*, familial inherited genes *BRCA1/2* and the luminal cell transcriptional program specified gene *GATA3*, for which both germline polymorphisms and somatic SNVs are associated with patient prognosis. We divided the 4,248 gene groups into deciles based on their median abundance (see Figure 1B, bottom). As expected, high-abundance mRNAs were observed in a larger fraction of patients, and most breast cancer deriver genes were detected in over 70% of the LBBC samples, including *APOB*, *KRT31/75* and *ROS1*.

To define molecular subgroups of LBBC, we first selected the top 1000 genes with the highest standard deviation and mean value of log2 expression from 4,248 DEGs (see Figure S1C and Table S4). We identified

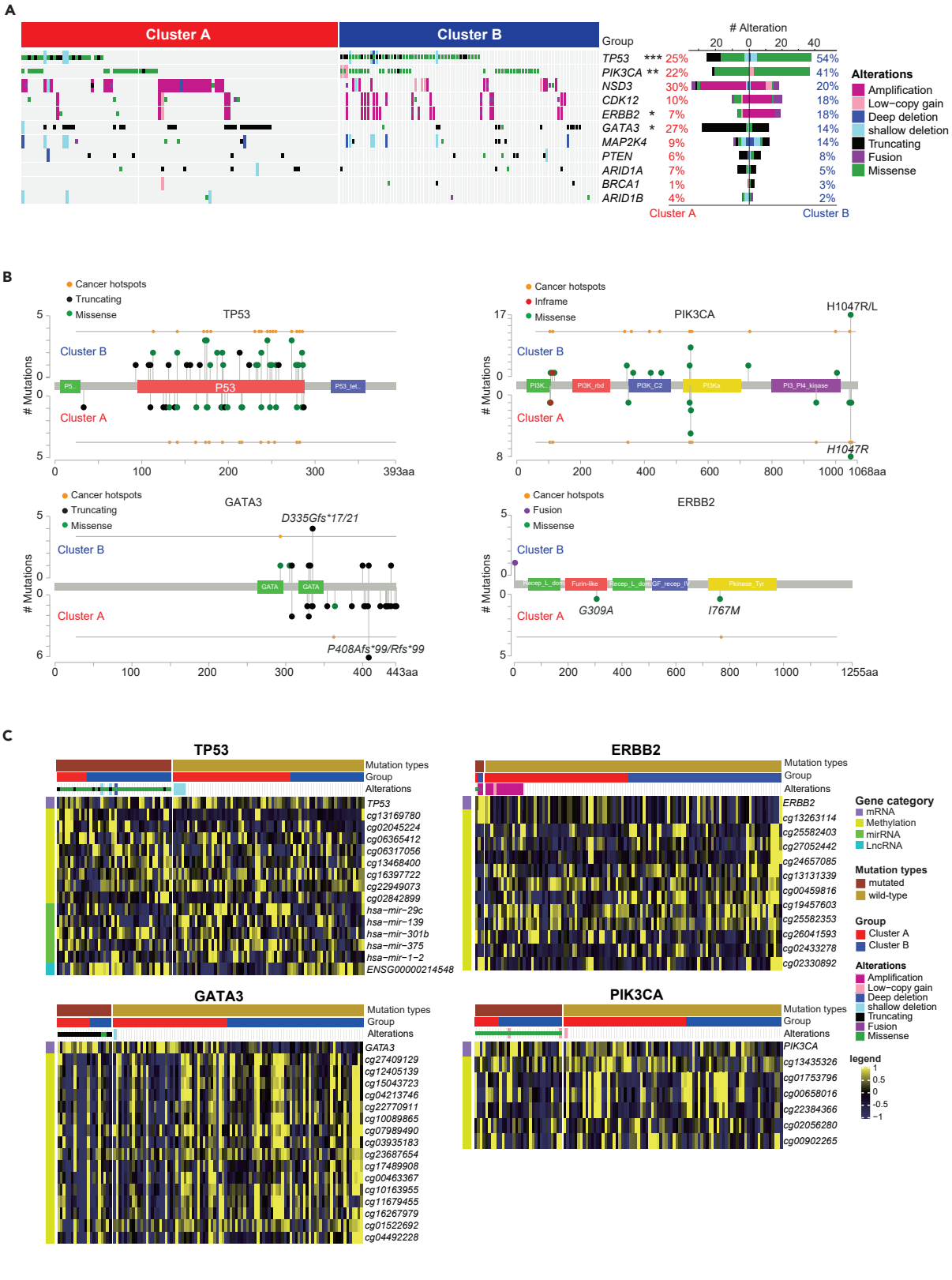

A
Cluster A
Cluster B
Group
Alteration
20 0 20 40
TP53 *** 25% 54%
PIK3CA ** 22% 41%
NSD3 30% 20%
CDK12 10% 18%
ERBB2 * 7% 18%
GATA3 * 27% 14%
MAP2K4 9% 14%
PTEN 6% 8%
ARID1A 7% 5%
BRCA1 1% 3%
ARID1B 4% 2%
Cluster A
Cluster B
Alterations
Amplification
Low−copy gain
Deep deletion
shallow deletion
Truncating
Fusion
Missense
B
Cancer hotspots
Truncating
Missense
TP53
Mutations
Cluster B
Cluster A
P53
393aa
Cancer hotspots
Inframe
Missense
PIK3CA
H1047R/L
H1047R
1068aa
Cancer hotspots
Truncating
Missense
GATA3
D335Gfs*17/21
P408Afs*99/Rfs*99
443aa
Cancer hotspots
Fusion
Missense
ERBB2
G309A
I767M
1255aa
C
TP53
ERBB2
GATA3
PIK3CA
Mutation types
Group
Alterations
TP53
cg13169780
cg02045224
cg06365412
cg06317056
cg13468400
cg16397722
cg22949073
cg02842899
hsa−mir−29c
hsa−mir−139
hsa−mir−301b
hsa−mir−375
hsa−mir−1−2
ENSG00000214548
ERBB2
cg13263114
cg25582403
cg27052442
cg24657085
cg13131339
cg00459816
cg19457603
cg25582353
cg26041593
cg02433278
cg02330892
GATA3
cg27409129
cg12405139
cg15043723
cg04213746
cg22770911
cg10089865
cg07989490
cg03935183
cg23687654
cg17489908
cg00463367
cg10163955
cg11679455
cg16267979
cg01522692
cg04492228
PIK3CA
cg13435326
cg01753796
cg00658016
cg22384366
cg02056280
cg00902265
Gene category
mRNA
Methylation
mirRNA
LncRNA
Mutation types
mutated
wild-type
Group
Cluster A
Cluster B
Alterations
Amplification
Low−copy gain
Deep deletion
shallow deletion
Truncating
Fusion
Missense
legend
1
0.5
0
-0.5
−1

Figure 2. The Somatic mutations and chromosomal instability between the two subtypes of TCGA LBBC patients

(A) Somatic genomic alterations identified in two subtypes of LBBC patients. The bottom panel showed somatic mutations and gene-level copy number alterations by patients (column) and by genes (row). The middle track showed the two clusters of LBBC patients (n = 187, cluster A n = 100, cluster B n = 87). The bar plot on the right indicated the alteration rates between the two subtypes. *p* value was calculated by Fisher's exact test, * <0.05, ** <0.01, *** <0.001.

(B) Location of somatic mutations of *TP53*, *PIK3CA*, *GATA3* and *ERBB2* between the two subtypes within the protein sequence in TCGA dataset.

(C) Multi-omics data heatmap of TP53, PIK3CA, GATA3 and ERBB2 between the mutated and wild-type samples. The samples (n = 121, cluster A n = 54, cluster B n = 58) were divided into mutated and wild-type groups based on whether the samples had Truncating, Missense and Fusion mutations in these four genes. See also Figure S2 and Table S5.

two patient subtypes (Clusters A and B) using unsupervised clustering and visualized the clustering results using PCA analysis (see Figure S1D). Based on the molecular profiles of these two subtypes across multi-omics platforms (Proteins, mRNAs, MirRNAs, LncRNAs, Methylation), we identified DEGs between the two subtypes, and integrated the expression data across multi-omics platforms for each subtype (see Figures 1C and S1E–S1G). We further examined the clinical relevance of multiomics-based classification which were indicative of tumor heterogeneity. The number of patients with CNA $\leq$0.2525 was significantly higher in cluster A (70% vs. 30%, Fisher's exact test p = 0.021), but patients with HER2+ were notably more in Cluster B compared with Cluster A (77% vs. 23%, Fisher's exact test p < 0.001, see Figure 1C and Table S3). The three-gene subtypes were barely different between Clusters A and B (Chi-square test p = 0.027, see Figure 1D). To access Ki-67 mRNA expression, we divided it into four quantiles (i.e., Q1, Q2, Q3 and Q4) based on the overall distribution of Ki-67 expression, which did not reach the statistical significance between the two subgroups (Chi-square test p = 0.123, see Figure 1E). The expression of Ki-67 has no significant difference between the two subgroups (*t* test p = 0.510, see Figure 1F). We also discovered a significant trend of shortened survival in patients of Cluster B (HR = 2.132, 95%CI: 1.014–4.484; Renyi test p = 0.013, see Figure 1G). The 5-year OS rates in Cluster B was 57% (95%CI: 41%–78%), which was greatly lower than the 90% of Cluster A (95%CI: 80%–100%). All these results indicated that the comprehensive multi-omics analysis of 187 LBBCs in TCGA dataset identified two molecularly distinct, clinically relevant subtypes.

Comparison in somatic mutation between cluster A and B of LBBC

In order to explore the somatic mutation profiles, we performed whole exome sequencing (WES) analysis and identified somatic DNA alterations in 187 LBBC samples, including truncating, missense, fusion, amplification, low-copy gain, deep deletion and shallow deletion. It was observed that *TP53* was the most frequently altered gene (25% vs. 54%) between Clusters A and B in this study followed by *PIK3CA* (22% vs. 41%), *NSD3* (30% vs. 20%), *CDK12* (10% vs. 18%), *ERBB2* (7% vs. 18%) and *GATA3* (27% vs. 14%) (Figure 2A and Table S5). Compared with Cluster A, patients in Cluster B had significantly more either missense or truncating mutations in *TP53* (p = 7.845e-05) and *PIK3CA* (p = 0.007), and amplification *ERBB2* (p = 0.025), but less truncating mutations in *GATA3* (p = 0.031, Fisher's exact test), suggesting potential association with significantly poor prognosis of patients in Cluster B. We also observed recurrent mutations in several genes previously reported as altered in breast cancer, including mutations in other known oncogenes, chromatin modification and DNA damage repair genes, such as *KMT2C* (9% vs. 13%), *RB1* (4% vs. 11%), *APOB* (2% vs. 8%), *PTEN* (6% vs. 8%), *BRCA2* (9% vs. 6%), *ARID1A* (7% vs. 5%), *CDH1* (6% vs. 5%), and so on, although it did not reach the statistical significance between Clusters A and B. Similarly, the cancer hotspots and domains of specific proteins (TP53, PIK3CA, GATA3 and ERBB2 mutants) were shown in Figure 2B. Then we examined the copy number variation and identified that deletions in *ARID1A* and *PTEN* were predominantly observed in patients of Cluster A, but *TP53* and *MAP2K4* in Cluster B (Figure S2A). A low-copy gain of *PIK3CA* was observed in Cluster B, while a low-copy gain of BRCA1 was observed in Cluster A. The burden of copy number gains was dramatically increased in patients of Cluster A (Wilcoxon test p = 0.048, see Figure S2B), but the losses were not markedly altered between the two subgroups (Wilcoxon test p = 0.420). The increased burden of copy number changes was unlikely due to differences in tumor cells between the two subgroups of samples. To gain a clear understanding of the mutation status of *TP53*, *PIK3CA*, *GATA3*, and *ERBB2*, we divided the samples into mutant and wild-type groups and generated multi-omics data heatmaps for these four genes (see Figure 2C). Generally, these results indicated the intratumor heterogeneity of patients with LBBC between Clusters A and B.

Differences analysis in DNA methylation between clusters A and B of LBBC

To understand the differences in mean methylation values between Clusters A (n = 65) and B (n = 64) in 129 patients, we performed distribution analysis of differentially methylated regions (DMRs) using ChAMP. The

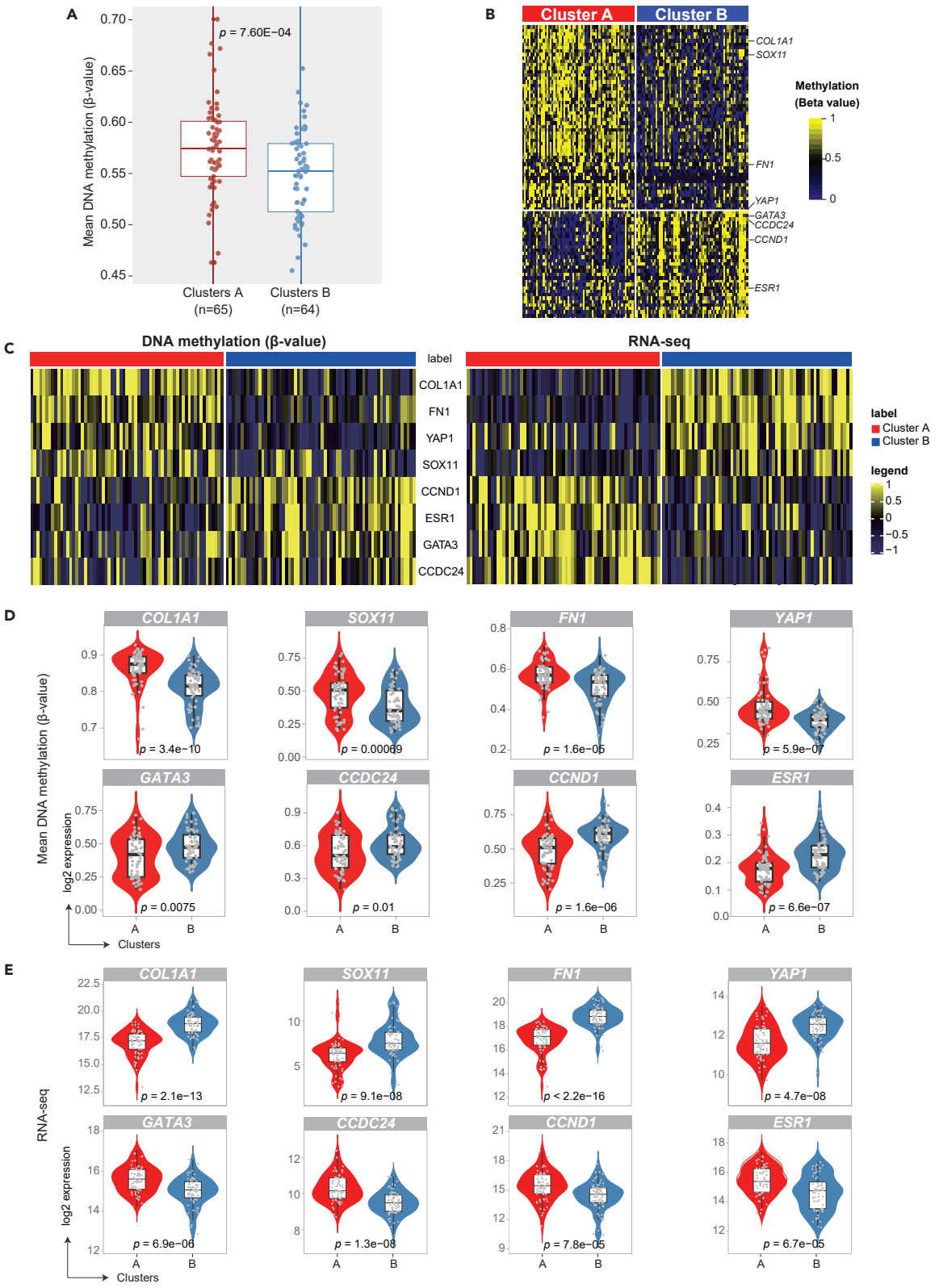

A
Mean DNA methylation (β-value)
p = 7.60E−04
Clusters A (n=65)
Clusters B (n=64)
B
Cluster A
Cluster B
COL1A1
SOX11
FN1
YAP1
GATA3
CCDC24
CCND1
ESR1
Methylation (Beta value)
C
DNA methylation (β-value)
RNA-seq
label
Cluster A
Cluster B
legend
D
Mean DNA methylation (β-value)
log2 expression
Clusters
p = 3.4e−10
p = 0.00069
p = 1.6e−05
p = 5.9e−07
p = 0.0075
p = 0.01
p = 1.6e−06
p = 6.6e−07
E
RNA-seq
p = 2.1e−13
p = 9.1e−08
p < 2.2e−16
p = 4.7e−08
p = 6.9e−06
p = 1.3e−08
p = 7.8e−05
p = 6.7e−05

Figure 3. DNA methylation analysis between Clusters A and B of TCGA LBBC

(A) Boxplot showed the different methylation value between Clusters A (n = 65) and B (n = 64) in 26,284 methylation sites. The points in the violin plot depict the samples in each subtype.

(B) Supervised clustering of 85 key genes across 128 samples between two subtypes. The symbols of cancer-related DEGs were labeled on the right.

(C) Heatmap of methylation and RNA-seq expression of 8 genes between the two subtypes (cluster A (red) and B (blue)), with cancer-related gene symbols marked in the middle.

(D and E) The violin plot showed the different methylation and RNA-seq value of 8 genes between cluster A (red) and B (blue). The boxplot (black) in the violin represents the interquartile range (IQR) and median value. The points in the violin plot depicts the samples in each subtype. See also Figure S3.

analysis revealed significant hypermethylation in patients of Cluster A compared to those of Cluster B (Wilcoxon test p = 7.60e-04, Figure 3A). Unsupervised clustering was performed with DNA methylation data including 85 significant genes from 26,284 DMRs in 129 high-purity samples between Clusters A and B. There were more extensive DNA hypermethylation regions of some oncogenes (*COL1A1*, *SOX11*, *FN1*, and *YAP1*) in Cluster A, and other oncogenes (*GATA3*, *CCDC24*, *CCND1*, and *ESR1*) in Cluster B (Figures 3B–3D), suggesting that these genes contained the notably highest number of DMSs in a certain chromosome position, even though the DNA methylation were consistently weaker across CpG sites (Figure S3). Consistent with these results, patients from Cluster A had significantly lower mRNA expression for *COL1A1*, *SOX11*, *FN1*, *and YAP1* genes, but higher expression for *GATA3*, *CCDC24*, *CCND1* and *ESR1* genes (Wilcoxon test p < 0.0001, Figures 3C–3E).

The integrated analysis of the DNA methylation and mRNA expression data revealed that these 8 genes were silenced by DNA methylation, which had been implicated in the development of breast cancer as well as previously reported to be altered in other cancers.

Dysregulated signaling pathways and potential druggable genes between two subgroups of LBBC

To gain an understanding of the differences of biological function between two clusters, we performed pathway enrichment analysis to identify the dysregulated molecular processes in the genomics data. The top 10 pathways in the Hallmark and KEGG dataset were identified in Clusters A and B, respectively. The Cluster A was predominantly composed of cell cycle and metabolic reprogramming pathways (Figures 4A and S4B), such as E2F targets, G2M checkpoint, cell cycle, and glycolysis pathways. With similar approaches, we discovered that Cluster B was mainly oncogenic and cancer immune response signaling (Figures 4A and S4A), including epithelial mesenchymal transition (EMT), TNFA signaling via NFKB, cytokine and cytokine receptor interaction, and IL6-JAK-STAT3 signaling. Then we applied two deconvolution approaches: MCP-counter to produce absolute abundance scores of 8 major immune cells, endothelial cells and fibroblasts, and CIBERSORT to evaluate the relative cellular fraction of 22 immune cell types. Among 10 cell types identified by MCP-counter, the abundance scores of fibroblasts, endothelial cells, cytotoxic lymphocytes and myeloid dendritic cells were significantly higher in Cluster B (Wilcoxon test p < 0.007, Figure 4B), as well as the relative cellular fraction of macrophage M1 produced by CIBERSORT (Wilcoxon test p < 3.831e-04, Figure 4C). Together, our results indicated that tumor cells might have reprogrammed the immune-related response in TME to facilitate the progression of patients in Cluster B.

Next, we selected 39 tumor-specific, highly abundant and significantly enriched genes through a stepwise filtering process, which were annotated as functionally important in cancer development (Figure 4D), including *PLAC1*, *BRDT*, *CABYR*, *CTNNA2*, and *TEX101* as known cancer testis antigens, *FN1*, *CDH2*, *CDH11*, *PDGFRA*, *COL3A1*, and *LAMA3* as emerging and attractive targets involved in EMT, three check points molecules *TNFRSF18*, *TNFSF4*, and *IDO1*, *CCND1* correlated with cell cycle, *COX6C* and *MRPS30* appearance in oxidative phosphorylation (OXPHOS), *DCN*, *COL5A1*, *VCAN*, *NT5E*, *TFF3*, *LCT*, and *CACNA1H* for glycolysis metabolism, and *SERPINF1* and *EPGN* for angiogenesis. Analysis based on canonical markers for tumor-specific DEGs (*YAP1*, *FN1*, and *ESR1*) also supported the two-type clustering with significantly different prognosis. In order to explore differently expressed proteins (DEPs) between two clusters, we performed DEPs analysis and identified 19 significant DEPs in 160 LBBC patients (Figures 4E, S4C, and S4D). They included the metabolic enzymes (MYOSINIIA, FIBRONECTIN, DJ1, and RAB11) involved in tumor growth,[7] a ligand-dependent nuclear receptor ERALPHA with a good response to anti-estrogen therapy,[8] a transcription regulator YAP_pS127 associated with poor prognosis of breast cancer by promoting tumor cell growth,[9] some kinases (MAPK_pT202Y204, MEK1_pS217S221, AKT_pS473, SRC_pY416, ARAF_pS299, BAP1C4, and P27) regulating the hallmarks of cancer, e.g., tumor growth, survival and

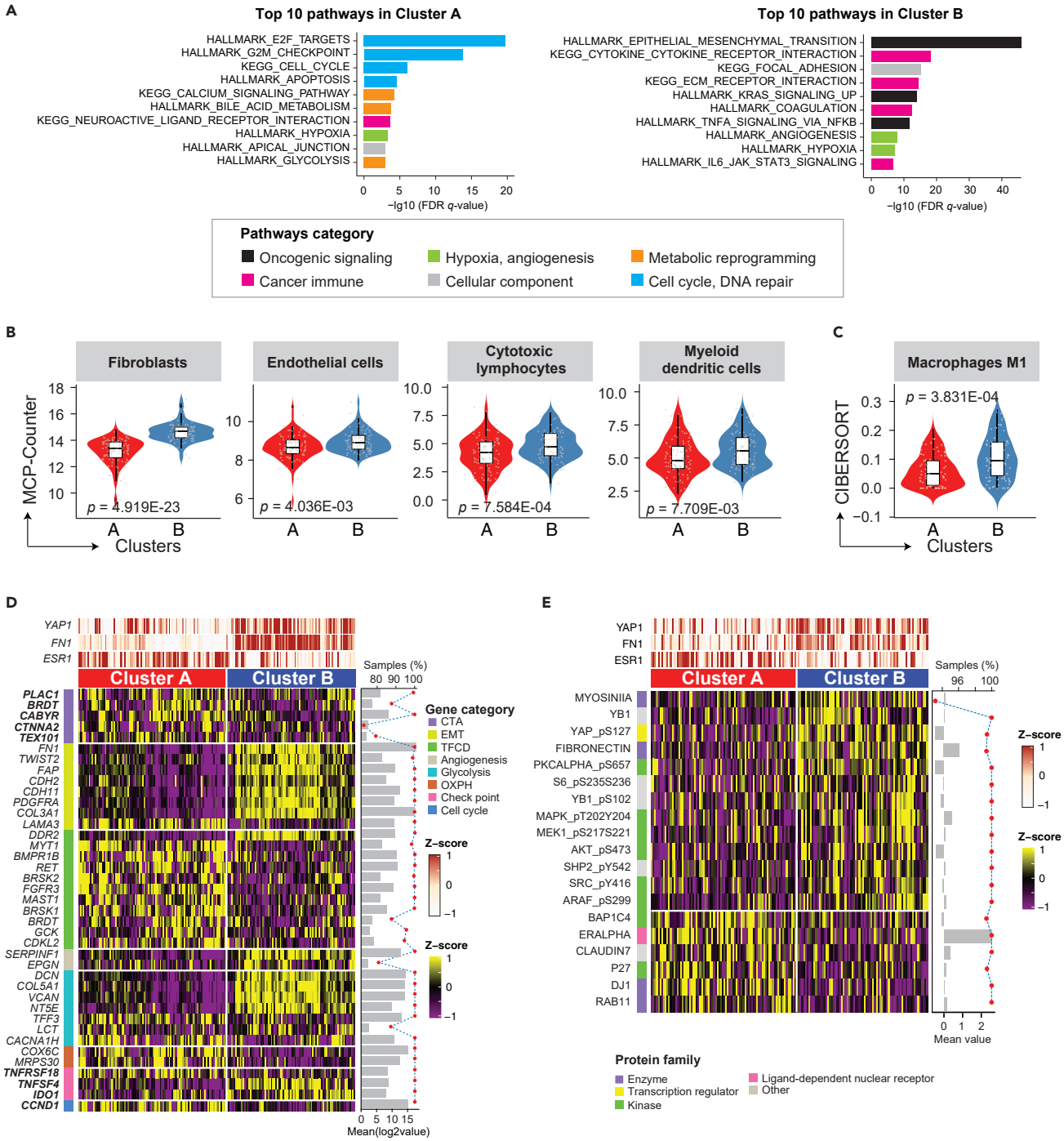


Figure 4. Enriched cancer-related pathways, immune profiles and pathway-specific molecules between two subtypes in patients with TCGA LBBC

(A) Pathway enrichment analysis identified top 10 biological pathways enriched in Clusters A and B by the cancer hallmark and KEGG gene sets in the Molecular Signature Database (MSigDB). The selected pathways were colored by their biological functions. FDR *q*-value, the p value are adjusted by the false discovery rate (FDR). A *q*-value threshold of 0.05 (5% FDR) was selected.

(B) Absolute abundance of fibroblasts, endothelial cells, cytotoxic lymphocytes and myeloid dendritic cells inferred by MCP-counter between two subtypes.

(C) Relative fraction of macrophages M1 inferred by CIBERSORT method between two subtypes.

(D) A heatmap of 39 significantly expressed genes with known functions was shown between two clusters. The tumor specific markers (e.g., *YAP1*, *FN1*, *EGFR*, *GATA3*, *ERBB2* and *ESR1*) were labeled on the top tracks for sample classification. Subgroup classification and gene names were annotated on the left with the corresponding track color coded by the functional category. Right histogram shows the fraction of samples (n = 187) with detected gene expression (y

Figure 4. *Continued*

axis on the top) for each listed gene. The red dots indicated the mean expression of each gene averaged across 187 samples (y axis at the bottom, Log2 transformed).

(E) A heatmap of 19 significantly expressed proteins (including phosphorylation data) with known functions was shown between two clusters. The tumor specific markers (e.g., YAP1, FN1, MAPK1/3, MAP2K1/2 and ESR1) were labeled on the top tracks for sample classification. Subgroup classification and protein names were annotated on the left with the corresponding track color coded by the functional category. Right histogram shows the fraction of samples (n = 160) with detected protein expression (y axis on the top) for each listed protein. The red dots indicated the mean expression of each protein averaged across 160 samples (y axis at the bottom, Log2 transformed). See also Figure S4.

invasiveness of tumor cells.[10–16] In line with the results above, the tumor-specific DEPs (YAP1, FN1, and ESR1) were with similarly altered expression between two clusters.

Genomic profiling correlated with somatic mutations of *TP53, PIK3CA, ERBB2,* and *GATA3*

To investigate the genomic features of patients with frequently somatic mutations in *TP53*, *PIK3CA*, *ERBB2*, and *GATA3*, we divided the samples into mutated and wild-type groups based on whether somatic mutations occurred. The characteristics of the two groups on other datasets were plotted and the characteristic pathways between the mutated and wild-type groups were obtained by calculating the ssGSEA (single sample Gene Set Enrichment Analysis) score for each tumor sample (Figure 5). The mutated and wild-type groups differed in the expression of DNA methylation, lncRNA, miRNA, mRNA, and protein. Interestingly, genes of biological significance between the mutated and wild-type groups included two druggable genes in clinical trials (*ERBB2* and *SOX11*), a gene associated with breast cancer metastasis (*SLC39A6*), the cell cycle-related gene *CCNE1*, the tumor invasion-related gene *MMP1*, and other cancer-related genes, such as *CCND1*, *YAP1*, *COX6C*, *APOB*, *KRT81*, and *CDH2*. The analysis of ssGSEA revealed that the mutated groups within *TP53* and *PIK3CA* received higher scores in the immune pathway, indicating that mutations in these genes resulted in alterations in genes associated with immunity. Additionally, it can be seen that Cluster B scored higher in the immune pathway than Cluster A. These results suggest that Cluster B has a different immune status than Cluster A, resulting in different survival differences between the two groups.

An innovative lncRNA-miRNA-mRNA competing endogenous RNA network associated with the clustering of LBBC

To explore the differentially expressed lncRNAs (DELs), we performed DELs analysis in 17,948 lncRNAs and identified 1,521 DELs between breast cancer and normal controls. We then filtered 12 significant DELs between Clusters A and B based on the above results. Similarly, we found 9 miRNAs and 20 mRNAs with significantly different expression which were predicted to be targeted by these 12 significant DELs (Spearman $|r| \geq 0.3$, q-value <0.01; Figures 6A and 6B). And 6 out of 9 significant miRNAs were predicted to target 16 significant mRNAs by miRbase between Clusters A and B (Spearman $|r| \geq 0.3$, q-value <0.01; Figure 6C). We further performed pathways enrichment analysis to explore the dysregulated molecular processes informed by 114 mRNAs with strong correlation with the significant DELs between two subgroups. In line with the results above, Cluster A was predominantly composed of cell cycle and DNA repair signaling, but EMT and immune response pathways for Cluster B (Figure 6D).

Ultimately, lncRNA-miRNA-mRNA competing endogenous RNA (ceRNA) network was constructed based on the DEGs results including 12 DELs, 8 miRNAs and 594 mRNAs (see Figure 6E). Moreover, a total of 532 mRNAs with significantly different expression were predicted to be targeted by these 8 miRNAs, and 44 mRNAs for these 12 DELs between two subgroups. Through parsing the co-expression network into different hub-based subnetworks, we observed 8 lncRNA/miRNA centered subnetworks with signaling pathway enrichment, which also revealed the predominant pathways composed of cell cycle, EMT and immune response. Generally, the results suggested that each component in the ceRNA network was remarkably related to the prognosis of LBBC patients between the two clusters.

The 20-gene signature could classify LBBC patients into two subgroups

To validate the two classifications of LBBC, we performed Lasso Cox regression analysis in 537 significant DEGs between Clusters A and B and revealed 37 genes without multivariate collinearity (see Figure S5A). Of which 20 important genes were obtained based on the mean decrease accuracy and mean decrease Gini score using random forest algorithm (see Figure S5B). And this 20-gene signature could stratify the LBBC patients into two subgroups (Cluster A and B) with significant prognosis in TCGA data (see Figure 7B).

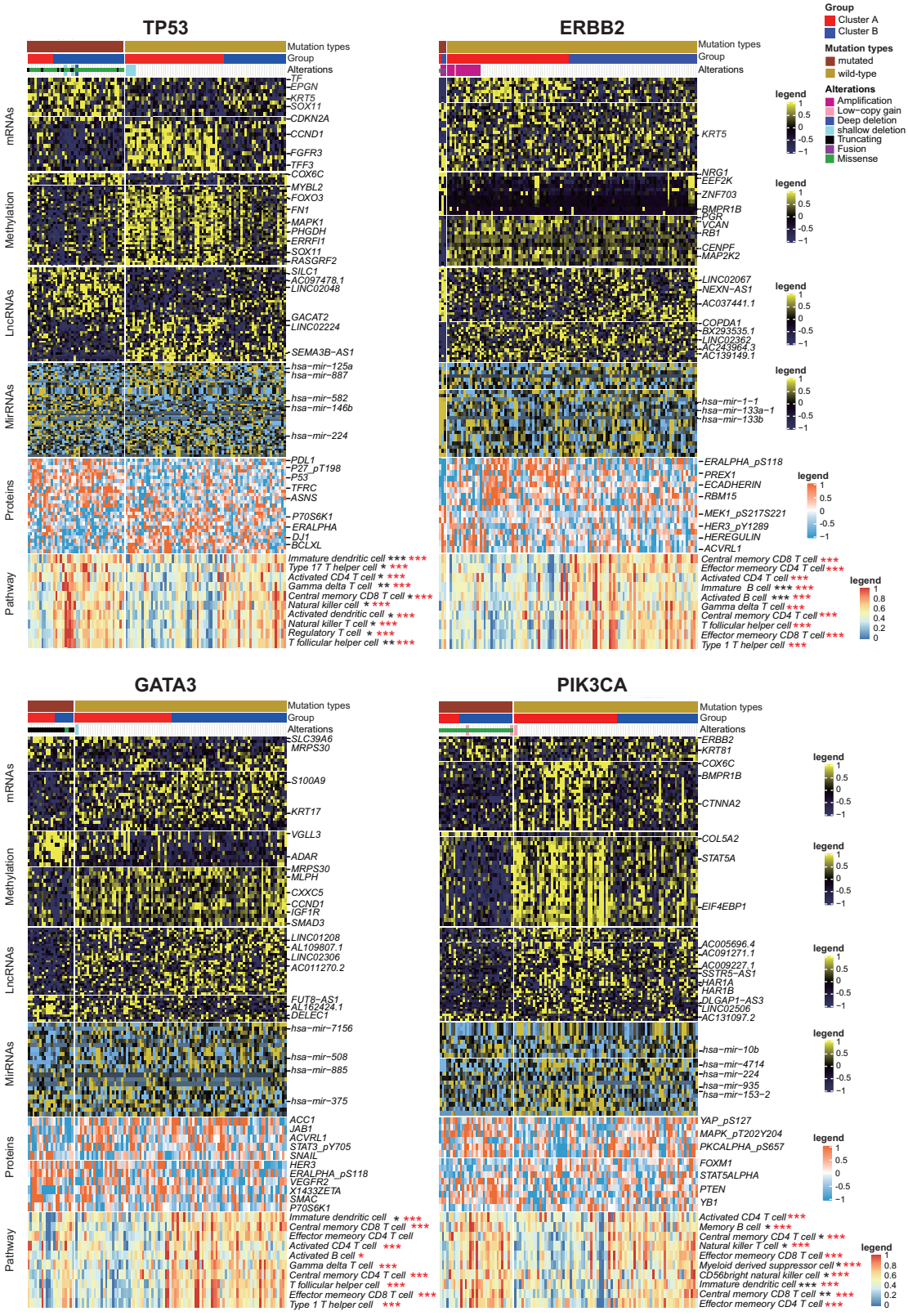

TP53
ERBB2
GATA3
PIK3CA
Mutation types
Group
Alterations
mRNAs
Methylation
LncRNAs
MirRNAs
Proteins
Pathway
Cluster A
Cluster B
mutated
wild-type
Amplification
Low-copy gain
Deep deletion
shallow deletion
Truncating
Fusion
Missense
legend

Figure 5. Distinct expression features and signaling pathways correlated with aberrant expression of TP53, PIK3CA, ERBB2 and GATA3 in TCGA LBBC patients

Heatmap showing DNA methylation, LncRNAs, MirRNAs, mRNAs, protein (including phosphorylation data) and pathway expression of TP53, PIK3CA, ERBB2 and GATA3 between mutated and wild-type. The subtype to which the samples belonged is labeled at the top of the figure, and important cancer-related genes are listed on the right. In addition, ssGSEA was used to calculate differences in immune pathways for TP53, PIK3CA, ERBB2 and GATA3, and the pathways that differed between mutated and wild-type are indicated by "*" and the pathways that differed between Cluster A and Cluster B are indicated by "*". p value was calculated by T-test, * <0.05, ** <0.01, *** <0.001.

Among these 20 genes, three genes including *CACNA2D2*, *CCDC24*, and *RAB3A* were significantly upregulated in Cluster A, and the other 17 signature genes were upregulated in Cluster B (DeSeq2 algorithm $p < 0.05$).

To confirm this emerging classification, we selected two additional independent datasets (METABRIC/Nature2012, n = 263; and GSE96058, n = 656) for validation. The detailed clinical and pathological characteristics of the patients were displayed in Figure 7A, Tables S2, S6, and S7, which were consistent with those in the TCGA dataset. We divided the LBBC patients into two similar subgroups (Clusters A and B) with significant prognosis in each independent cohort, and the distribution of these 20 signature genes between the two clusters was consistent with that in the TCGA dataset (Figures 7B and S5C). For Ki-67 protein expression, there was no significant difference between Clusters A and B in the mean (Mann-Whitney U test $p = 0.130$) and quartile range (Fisher's exact test $p = 0.946$) of expression, and this situation was the same for Ki-67 gene expression (Fisher's exact test $p = 0.130$, Figure 7C). In accordance with the survival analysis in TCGA, patients with LBBC in the METABRIC/Nature2012 dataset in Cluster B (n = 130) had a significantly higher risk (HR = 1.551, 95%CI: 1.014–2.035, Renyi test $p = 0.001$) than those in Cluster A (n = 133), while the 5-year overall survival (OS) rates in Cluster B were lower than those in Cluster A, i.e., 49% (95%CI: 41%–58%) compared to 64% (95%CI: 56%–73%), respectively (see Figure 7D upper). The 20-gene signature-based classification of another cohort in GSE96058 also produced similar results (Figure 7D lower). The HR (Cluster B vs. Cluster A) in this cohort was 1.703 (95%CI: 1.100–2.637). Furthermore, in the GSE96058 cohort, the 5-year overall survival (OS) rates in Cluster B were significantly worse compared to Cluster A, with rates of 79% (95%CI: 73%–86%) and 88% (95%CI: 84%–92%), respectively (Renyi test $p = 0.017$). Furthermore, to provide additional validation for the two subgroups, we analyzed the E-MTAB-6703, GSE20685, GSE54275, and GSE2109 datasets. Based on the heatmap (see Figure S6) and detailed clinical analysis of the patients (see Table S8), it can be concluded that the expression features of Clusters A and B in the four supplemental validation sets are consistent with the expression features in the training set. Therefore, the 20-gene signature was able to classify LBBC patients into two groups with significant prognostic differences. In summary, the 20-gene signature was able to classify LBBC patients into two groups with significantly different prognoses.

DISCUSSION

To refine the molecular classification of LBBC using biomarkers and explore innovative approaches to achieve precise treatment, we identified two LBBC subgroups based on comprehensive genomic data. One subgroup (Cluster A) was enriched in cell cycle related genes and had a favorable prognosis. The other subgroup (Cluster B) was mainly enriched in EMT and immune response-related genes.

Somatic mutations and copy number alterations (CNAs) are key to distinguishing clusters A and B. We discovered four significantly altered genes between the two subgroups; *TP53*, *PIK3CA*, *ERBB2*, and *GATA3*. 25% of patients in Cluster A had at least one of the following *TP53* mutations: truncation, missense, and shallow deletion. 54% of cluster B patients had similar alterations, which lined up with the previous study report[17] and were more likely to be aggressive.[18,19] We also confirmed that the *PIK3CA* oncogene is the second most frequently mutated gene in breast cancer after the *TP53* suppressor gene.[20] Mutations were observed in 22% of patients in Cluster A, and in 41% of patients in Cluster B. Given the role of PI3K reported by Nixon, M. J et al.[21] in supporting proliferation, survival, and hormone receptor pathway activity, it is not surprising that patients in Cluster A would have cellular mechanisms to maintain cell cycle dominance. Unfortunately, clinical trials have not yet demonstrated meaningful activity for single-agent PI3K inhibitors. According to the report by Mukohara, T,[22] *PIK3CA* mutations could coexist with other PI3K-enhancing mechanisms including *ERBB2* amplification and PTEN protein loss. *ERBB2* is amplified and/or overexpressed in 15–30% of invasive breast carcinomas reported by Iqbal, N.[23] However, the amplification and/or overexpression of *ERBB2* is only 7% in Cluster A patients, while it is 18% in Cluster B patients.

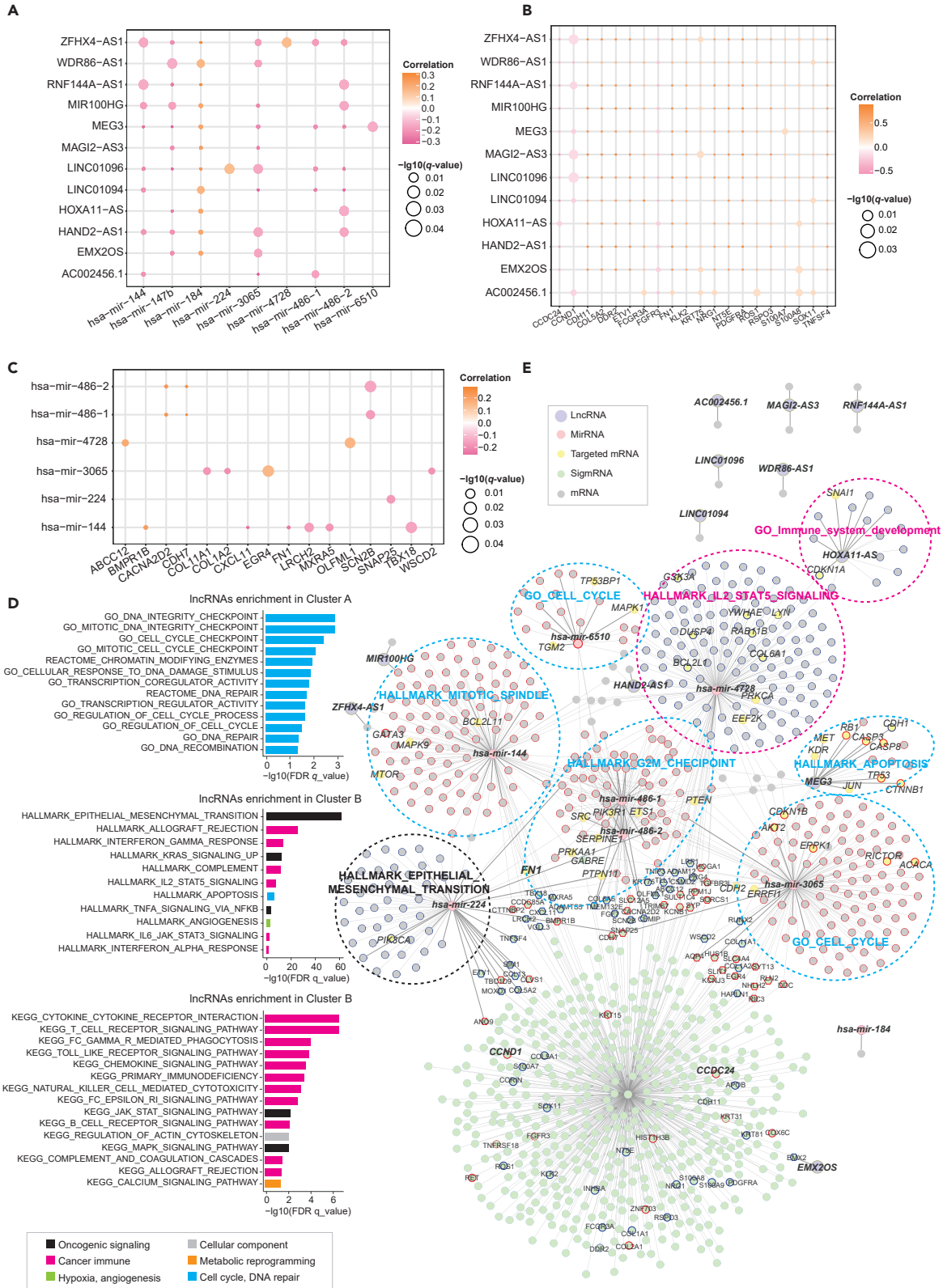

A
ZFHX4−AS1
WDR86−AS1
RNF144A−AS1
MIR100HG
MEG3
MAGI2−AS3
LINC01096
LINC01094
HOXA11−AS
HAND2−AS1
EMX2OS
AC002456.1
hsa−mir−144
hsa−mir−147b
hsa−mir−184
hsa−mir−224
hsa−mir−3065
hsa−mir−4728
hsa−mir−486−1
hsa−mir−486−2
hsa−mir−6510
Correlation
−lg10(q−value)
B
C
hsa−mir−486−2
hsa−mir−486−1
hsa−mir−4728
hsa−mir−3065
hsa−mir−224
hsa−mir−144
ABCC12
BMPR1B
CACNA2D2
CDH7
COL11A1
COL1A2
CXCL11
EGR4
FN1
LRCH2
MXRA5
OLFML1
SCN2B
SNAP25
TBX18
WSCD2
D
lncRNAs enrichment in Cluster A
GO_DNA_INTEGRITY_CHECKPOINT
GO_MITOTIC_DNA_INTEGRITY_CHECKPOINT
GO_CELL_CYCLE_CHECKPOINT
GO_MITOTIC_CELL_CYCLE_CHECKPOINT
REACTOME_CHROMATIN_MODIFYING_ENZYMES
GO_CELLULAR_RESPONSE_TO_DNA_DAMAGE_STIMULUS
GO_TRANSCRIPTION_COREGULATOR_ACTIVITY
REACTOME_DNA_REPAIR
GO_TRANSCRIPTION_REGULATOR_ACTIVITY
GO_REGULATION_OF_CELL_CYCLE_PROCESS
GO_REGULATION_OF_CELL_CYCLE
GO_DNA_REPAIR
GO_DNA_RECOMBINATION
−lg10(FDR q_value)
lncRNAs enrichment in Cluster B
HALLMARK_EPITHELIAL_MESENCHYMAL_TRANSITION
HALLMARK_ALLOGRAFT_REJECTION
HALLMARK_INTERFERON_GAMMA_RESPONSE
HALLMARK_KRAS_SIGNALING_UP
HALLMARK_COMPLEMENT
HALLMARK_IL2_STAT5_SIGNALING
HALLMARK_APOPTOSIS
HALLMARK_TNFA_SIGNALING_VIA_NFKB
HALLMARK_ANGIOGENESIS
HALLMARK_IL6_JAK_STAT3_SIGNALING
HALLMARK_INTERFERON_ALPHA_RESPONSE
−lg10(FDR q_value)
lncRNAs enrichment in Cluster B
KEGG_CYTOKINE_CYTOKINE_RECEPTOR_INTERACTION
KEGG_T_CELL_RECEPTOR_SIGNALING_PATHWAY
KEGG_FC_GAMMA_R_MEDIATED_PHAGOCYTOSIS
KEGG_TOLL_LIKE_RECEPTOR_SIGNALING_PATHWAY
KEGG_CHEMOKINE_SIGNALING_PATHWAY
KEGG_PRIMARY_IMMUNODEFICIENCY
KEGG_NATURAL_KILLER_CELL_MEDIATED_CYTOTOXICITY
KEGG_FC_EPSILON_RI_SIGNALING_PATHWAY
KEGG_JAK_STAT_SIGNALING_PATHWAY
KEGG_B_CELL_RECEPTOR_SIGNALING_PATHWAY
KEGG_REGULATION_OF_ACTIN_CYTOSKELETON
KEGG_MAPK_SIGNALING_PATHWAY
KEGG_COMPLEMENT_AND_COAGULATION_CASCADES
KEGG_ALLOGRAFT_REJECTION
KEGG_CALCIUM_SIGNALING_PATHWAY
−lg10(FDR q_value)
Oncogenic signaling
Cancer immune
Hypoxia, angiogenesis
Cellular component
Metabolic reprogramming
Cell cycle, DNA repair
E
LncRNA
MirRNA
Targeted mRNA
SigmRNA
mRNA
GO_CELL_CYCLE
HALLMARK_IL2_STAT5_SIGNALING
HALLMARK_MITOTIC_SPINDLE
HALLMARK_G2M_CHECPOINT
HALLMARK_APOPTOSIS
HALLMARK_EPITHELIAL_MESENCHYMAL_TRANSITION
GO_Immune_system_development
GO_CELL_CYCLE
hsa-mir-6510
hsa-mir-4728
hsa-mir-144
hsa-mir-486-1
hsa-mir-486-2
hsa-mir-224
hsa-mir-3065
hsa-mir-184
FN1
CCND1
CCDC24
EMX2OS

Figure 6. Competitive endogenous RNA (ceRNAs) associated with subtypes of TCGA LBBC patients

(A–C) Summary of correlations (A) between lncRNAs and miRNAs, (B) between lncRNAs and mRNAs, (C) between miRNAs and mRNAs based on the mean expression values (log2). The node size was associated with the significance of correlation coefficient between genes.

(D) Enrichment analysis of selected processes higher significantly in each subtype. First, lncRNAs that differed both between tumor and normal samples and between cluster A and B were screened using DESeq2. Second, based on the lncRNAs obtained, mRNAs that were strongly correlated (spearman r > 0.5) with the lncRNAs highly expressed in cluster A were selected as input for the enrichment analysis of lncRNAs in clusterA. Similarly, the enrichment analysis of Cluster B was performed by taking the mRNAs with strong correlation (spearman r > 0.7) with the highly expressed lncRNAs of Cluster B. Enriched pathways were colored as indicated. FDR *q*-value, the p value adjusted by the false discovery rate (FDR).

(E) The ceRNAs network displayed differentially activated pathway features between Clusters A and B. A total of 12 lncRNAs, 8 miRNAs and 594 mRNAs (42/594 genes were significant between two subtypes in the protein expression) were enrolled into this network. The edge stands for the correlation between genes. Node size and color reflect different RNA types (light purple: lncRNA; light pink: miRNA; light yellow: targeted mRNA; light green: significant mRNA between Clusters A and B; light gray: mRNA; red-edged: Cluster A; blue-edged: Cluster B). The largest interconnected regulatory subnetworks of differentially activated ceRNAs were displayed, with network hubs showing cancer-related or biologically functional pathways. The enriched pathways were labeled with different colors (light blue: Cell cycle and apoptosis; black: oncogenic signaling; pink: cancer immune pathways).

This study showed that patients with locally advanced breast cancer (LBBC) in Cluster B with more ERBB2 amplification underwent crosstalk between epithelial-mesenchymal transition (EMT) and immune response, which led to restructuring of the extracellular matrix (ECM) and immune landscape to support tumor proliferation, progression, and metastasis, and this observation was supported by Singh, S et al.[24] *GATA3* mutations were observed in 27% of LBBC patients, present in 14% of cases in Cluster B, and in Cluster A with predominantly truncating mutations. According to the report of Takaku, M et al.,[25] *GATA3* suppresses the expression of factors critical to EMT and metastasis and has been identified as an important negative regulator of tumor characteristics correlating with poor prognosis. These data highlight the potential value of clinical application for these somatic alterations even in the absence of a clear family history of cancer.

There were notable differences in genome-wide methylation between Cluster A (n = 65) and B (n = 64). One of the most frequent epigenetic alterationswas *YAP1* with hypomethylation in Cluster B, which was mapped to chromosome region 11q22 amplicon[26] and considered a specific transcriptional activator and a leading effector of the Hippo tumor suppressor pathway that was potently pro-metastatic in breast cancer, and this observation is supported by Lamar, J. M et al.[27] And *FN1*, *SOX11*, and *COL1A1* had a similar methylation trend in Cluster B compared with Cluster A, which was involved in cell adhesion and migration processes such as metastasis and host defense, with several reports supporting this observation.[28] *GATA3* could promote a transcriptional program specifying luminal cell identity in the normal development reported by Takaku, M et al.[25] However, *GATA3* with hypomethylation in Cluster A could activate and function downstream of *BRCA1* to suppress EMT in breast cancer.[29] Typically, DNA methylation is a risk factor for breast cancer. Our genome-wide study confirmed the diverse methylation status between the two subgroups with significantly different prognoses. Some noteworthy points include the selection of CpG-containing regions based on observed differences in methylation levels in different locations.

Integrated pathway enrichment analysis revealed diverse biological pathways and untried druggable targets (cancer testis antigens, enzymes, kinases and TFs). Cell cycle and metabolic signaling pathways were mainly in Cluster A, whereas EMT and immune response predominated in Cluster B. Thus, Cluster A was more representative of tumor cell proliferation with higher expressions of cancer markers associated with cell cycle, including *CCND1*, *E2F1* and *SMAD4*. The tumor cells in Cluster B might benefit from alternative therapies targeted by cancer testis antigens *PLAC1*, *BRDT*, *CABRY*, *CTNNA2,* and *TEX101*, and EMT markers *FN1*, *TWIST2*, *FAP*, *CDH2*, *CDH11*, *FDGFRA*, *COL3A1*, and *LAMA3*. Combining previous studies,[24,30–37] we found that the signaling pathway between EMT and immune response may be cross-talking in Cluster B patients. Furthermore, M1 macrophages have been shown to enhance the metastatic potential of cancer cells through NF-kB activation, which may explain their higher CIBERSORT fraction in Cluster B, as reported by Cho, U et al.[38] A systematic comparison of mRNA and protein expression patterns revealed untried expression patterns of key therapeutic targets, notably low mRNA and high protein expression of *CD274* (PDL1), and high mRNA and low protein expression of *YAP1* (YAP_pS127), *AKT1* (AKT_pT308), and *CDH1* (E-cadherin). Taken together, these results highlight the diverse characteristics of the TME that assist tumor cells in their proliferation, invasion, and metastasis, and help distinguish between Clusters A and B.

Proteogenomic analysis of *TP53*, *PIK3CA*, *ERBB2* and *GATA3* revealed a poor correlation between genomic and proteomic data as previously discovered in other studies.[39] We analyzed the distribution

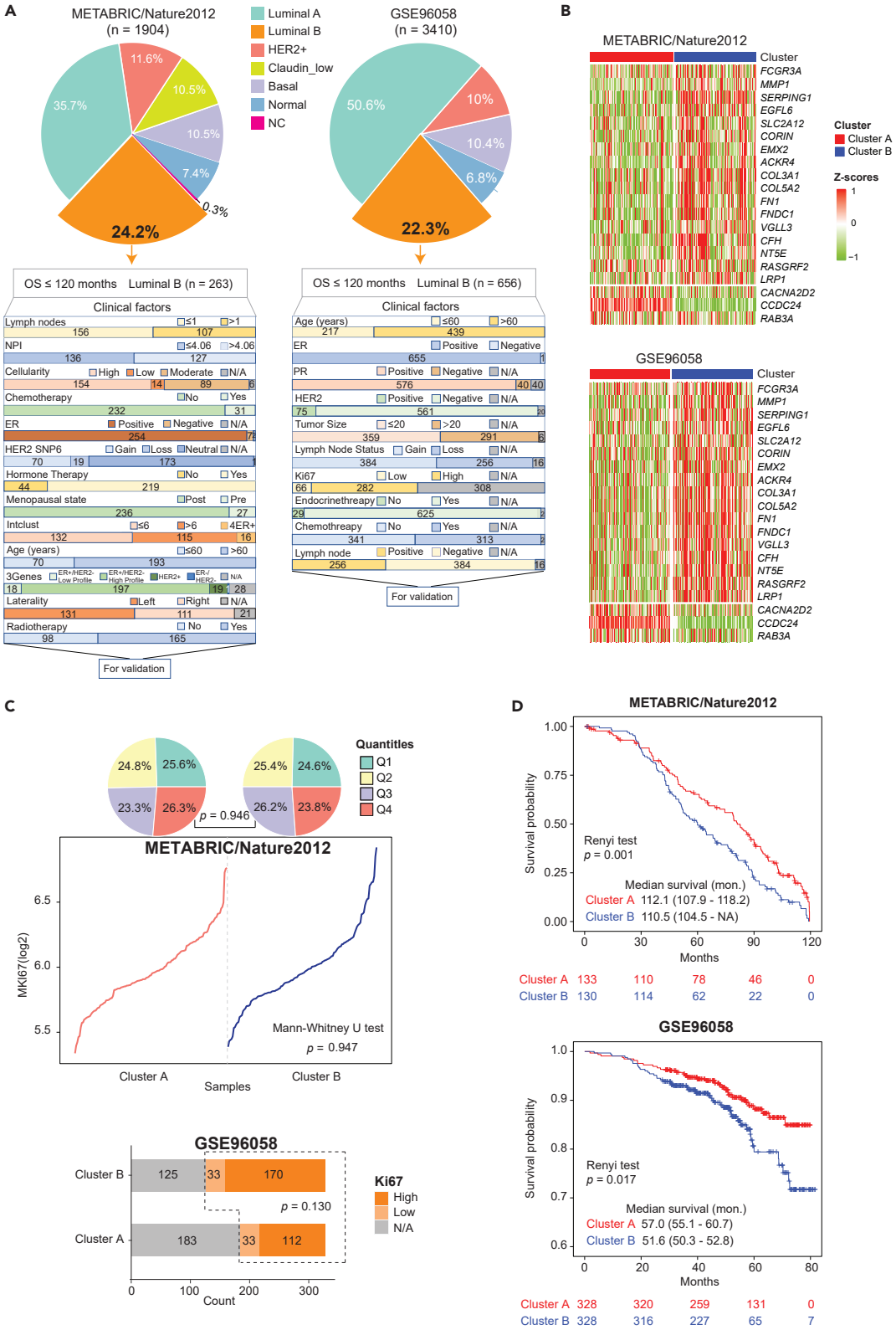

A
METABRIC/Nature2012
(n = 1904)
GSE96058
(n = 3410)
Luminal A
Luminal B
HER2+
Claudin_low
Basal
Normal
NC
35.7%
11.6%
10.5%
10.5%
7.4%
0.3%
24.2%
50.6%
10%
10.4%
6.8%
22.3%
OS ≤ 120 months Luminal B (n = 263)
OS ≤ 120 months Luminal B (n = 656)
Clinical factors
Lymph nodes ≤1 >1
NPI ≤4.06 >4.06
Cellularity High Low Moderate N/A
Chemotherapy No Yes
ER Positive Negative N/A
HER2 SNP6 Gain Loss Neutral N/A
Hormone Therapy No Yes
Menopausal state Post Pre
Intclust ≤6 >6 4ER+
Age (years) ≤60 >60
3Genes ER+/HER2- Low Profile ER+/HER2- High Profile HER2+ ER-/HER2- N/A
Laterality Left Right N/A
Radiotherapy No Yes
Age (years) ≤60 >60
ER Positive Negative
PR Positive Negative N/A
HER2 Positive Negative N/A
Tumor Size ≤20 >20 N/A
Lymph Node Status Gain Loss N/A
Ki67 Low High N/A
Endocrinethreapy No Yes N/A
Chemothreapy No Yes N/A
Lymph node Positive Negative N/A
For validation
For validation
B
METABRIC/Nature2012
GSE96058
Cluster
Cluster A
Cluster B
Z-scores
FCGR3A
MMP1
SERPING1
EGFL6
SLC2A12
CORIN
EMX2
ACKR4
COL3A1
COL5A2
FN1
FNDC1
VGLL3
CFH
NT5E
RASGRF2
LRP1
CACNA2D2
CCDC24
RAB3A
C
Quantiles
Q1
Q2
Q3
Q4
24.8%
25.6%
23.3%
26.3%
p = 0.946
25.4%
24.6%
26.2%
23.8%
METABRIC/Nature2012
MKI67(log2)
Mann-Whitney U test
p = 0.947
Cluster A
Samples
Cluster B
GSE96058
Cluster B
125
33
170
p = 0.130
Cluster A
183
33
112
Count
Ki67
High
Low
N/A
D
METABRIC/Nature2012
Survival probability
Renyi test
p = 0.001
Median survival (mon.)
Cluster A 112.1 (107.9 - 118.2)
Cluster B 110.5 (104.5 - NA)
Months
Cluster A 133 110 78 46 0
Cluster B 130 114 62 22 0
GSE96058
Survival probability
Renyi test
p = 0.017
Median survival (mon.)
Cluster A 57.0 (55.1 - 60.7)
Cluster B 51.6 (50.3 - 52.8)
Months
Cluster A 328 320 259 131 0
Cluster B 328 316 227 65 7

Figure 7. Validation of clustering in another 2 independent dataset of LBBC

(A) The overview of study showing the clinical characteristics of LBBC patients in another two independent cohorts with mRNA data: METABRIC/Nature 2012 (n = 263) and GSE96058 (n = 656).

(B) Unsupervised hierarchical clustering was performed with the same signature (20 DEGs) in METABRIC/Nature 2012 and GSE96058 by RNA-seq analysis.

(C) Scatterplot (top) showed the Ki-67 mRNA expression in each sample between Clusters A and B in METABRIC/Nature 2012 dataset, p value was calculated with Mann-Whitney U test. Pie plots indicated the quantities distribution of Ki67 expression value between two subtypes, p value was calculated with Fisher's exact test. Bar plot (bottom) showed the immunohistochemical status of Ki-67 in LBBC patients from GSE96058, p value was calculated with fisher exact test between Ki67 high and low group.

(D) Kaplan-Meier survival analysis for LBBC between the two subtypes in two testing sets. See also Figure S5, Tables S2, S6, and S7.

of Clusters A and B, as well as the frequency of mutations in *TP53*, *PIK3CA*, *ERBB2*, and *GATA3* in both mutated and wild-type samples, which were divided based on the presence or absence of mutations in each gene. Several cancer-related isogenes *SOX11*, *CCNE1*, *MMP1*, *SLC39A6*, *YAP1*, and *ERBB2* differed significantly between mutated and wild-type. Furthermore, the ssGSEA analysis results for these four genes in the mutated and wild-type (*TP53*, *PIK3CA*, *ERBB2*, and *GATA3*) showed that Cluster B obtained a higher ssGSEA score, suggesting that it could provide more therapeutic targets.

In summary, our comprehensive analysis of multiple molecular profiling platforms revealed the complex molecular landscape of LBBC, and the analysis of differences between two subtypes provided useful targets and signaling pathways for achieving more precise therapy of LBBC.

Limitations of the study

Our study is not without limitations. First, as breast cancer is a heterogeneous disease for which LBBC accounts for up to 30%, the small number of patients included may reduce the power of our study. Second, although our study results provide fresh perspectives into the pathogenesis of LBBC, further molecular biology experiments are needed to validate these findings. Finally, due to the incomplete clinical/demographic data of the samples included in the experimental process, it was not possible to make a more accurate judgment on the clinical differences between the subtypes defined in this paper, such as whether there are differences in ethnicity between Cluster A and Cluster B.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS
  - Ethics approval and consent to participate
  - Patient cohort and sample collection
- METHOD DETAILS
  - Identification of differently expressed genes (DEGs) and enriched signaling pathways
- QUANTIFICATION AND STATISTICAL ANALYSIS
  - Samples classification and validation based on mRNA data
  - Mutation signature analysis
  - DNA copy number analysis
  - DNA methylation analysis
  - Deconvolution of the cellular composition with LBBC samples
  - Analysis of reverse phase protein array (RPPA)
  - Construction of competitive endogenous RNA (ceRNAs) network
  - Statistical analysis

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.isci.2023.107466.

ACKNOWLEDGMENTS

This study was supported by National Natural Science Foundation of China (Grant Number: 62076015).

AUTHOR CONTRIBUTIONS

S.T.Z., S.W., and Z.D.L. conceived and jointly supervised this study. B.L. provided guidance on research methodology and paper organization. H.N.W. was responsible for the organization and completion of the experiments. The others assisted with samples collection and manuscript revision.

DECLARATION OF INTERESTS

The authors declare that they have no competing interests.

Received: September 28, 2022
Revised: January 30, 2023
Accepted: July 21, 2023
Published: July 26, 2023

REFERENCES

1. Sung, H., Ferlay, J., Siegel, R.L., Laversanne, M., Soerjomataram, I., Jemal, A., and Bray, F. (2021). Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA. Cancer J. Clin. *71*, 209–249. https://doi.org/10.3322/caac.21660.

2. Metzger-Filho, O., Sun, Z., Viale, G., Price, K.N., Crivellari, D., Snyder, R.D., Gelber, R.D., Castiglione-Gertsch, M., Coates, A.S., Goldhirsch, A., and Cardoso, F. (2013). Patterns of recurrence and outcome according to breast cancer subtypes in lymph node–negative disease: results from International Breast Cancer Study Group Trials VIII and IX. J. Clin. Oncol. *31*, 3083–3090. https://doi.org/10.1200/JCO.2012.46.1574.

3. Kos, T., Aksoy, S., Sendur, M.A.N., Arik, Z., Civelek, B., Kandemir, N., Ozdemir, N.Y., Zengin, N., and Altundag, K. (2013). Variations in tumor marker levels in metastatic breast cancer patients according to tumor subtypes. J BUON *18*, 608–613.

4. He, Z.-Y., Wu, S.-G., Yang, Q., Sun, J.-Y., Li, F.-Y., Lin, Q., and Lin, H.-X. (2015). Breast cancer subtype is associated with axillary lymph node metastasis: a retrospective cohort study. Medicine *94*, e2213. https://doi.org/10.1097/MD.0000000000002213.

5. Blows, F.M., Driver, K.E., Schmidt, M.K., Broeks, A., Van Leeuwen, F.E., Wesseling, J., Cheang, M.C., Gelmon, K., Nielsen, T.O., Blomqvist, C., et al. (2010). Subtyping of breast cancer by immunohistochemistry to investigate a relationship between subtype and short and long term survival: a collaborative analysis of data for 10,159 cases from 12 studies. PLoS Med. *7*, e1000279. https://doi.org/10.1371/journal.pmed.1000279.

6. The Cancer Genome Atlas homepage. http://cancergenome.nih.gov/abouttcga.

7. Finley, L.W.S. (2019). Metabolic signal curbs cancer-cell migration. Nature *571*, 39–40. https://doi.org/10.1038/d41586-019-01934-9.

8. Reis-Filho, J.S., Drury, S., Lambros, M.B., Marchio, C., Johnson, N., Natrajan, R., Salter, J., Levey, P., Fletcher, O., Peto, J., et al. (2008). ESR1 gene amplification in breast cancer: a common phenomenon? Nat. Genet. *40*, 809–810. author reply 810-812. https://doi.org/10.1038/ng0708-809a.

9. Guo, L., Chen, Y., Luo, J., Zheng, J., and Shao, G. (2019). YAP 1 overexpression is associated with poor prognosis of breast cancer patients and induces breast cancer cell growth by inhibiting PTEN. FEBS Open Bio *9*, 437–445. https://doi.org/10.1002/2211-5463.12597.

10. Chiarle, R., Pagano, M., and Inghirami, G. (2001). The cyclin dependent kinase inhibitor p27 and its prognostic role in breast cancer. Breast Cancer Res. *3*, 91–94. https://doi.org/10.1186/bcr277.

11. Hinz, N., and Jücker, M. (2019). Distinct functions of AKT isoforms in breast cancer: a comprehensive review. Cell Commun. Signal. *17*, 1–29. https://doi.org/10.1186/s12964-019-0450-3.

12. Huang, C.Y., Chou, Y.H., Hsieh, N.T., Chen, H.H., and Lee, M.F. (2012). MED28 regulates MEK1-dependent cellular migration in human breast cancer cells. J. Cell. Physiol. *227*, 3820–3827. https://doi.org/10.1002/jcp.24093.

13. Irby, R.B., and Yeatman, T.J. (2000). Role of Src expression and activation in human cancer. Oncogene *19*, 5636–5642. https://doi.org/10.1038/sj.onc.1203912.

14. Lin, W., Tong, C., Zhang, W., Cen, W., Wang, Y., Li, J., Zhu, Z., Yu, J., and Lu, B. (2020). Silencing ARAF Suppresses the Malignant Phenotypes of Gallbladder Cancer Cells. BioMed Res. Int. *2020*, 3235786. https://doi.org/10.1155/2020/3235786.

15. Santen, R.J., Song, R.X., McPherson, R., Kumar, R., Adam, L., Jeng, M.-H., and Yue, W. (2002). The role of mitogen-activated protein (MAP) kinase in breast cancer. J. Steroid Biochem. Mol. Biol. *80*, 239–256. https://doi.org/10.1016/s0960-0760(01)00189-3.

16. Shahriyari, L., Abdel-Rahman, M., and Cebulla, C. (2019). BAP1 expression is prognostic in breast and uveal melanoma but not colon cancer and is highly positively correlated with RBM15B and USP19. PLoS One *14*, e0211507. https://doi.org/10.1371/journal.pone.0211507.

17. Varna, M., Bousquet, G., Plassa, L.-F., Bertheau, P., and Janin, A. (2011). TP53 status and response to treatment in breast cancers. J. Biomed. Biotechnol. *2011*, 284584. https://doi.org/10.1155/2011/284584.

18. Langerød, A., Zhao, H., Borgan, Ø., Nesland, J.M., Bukholm, I.R., Ikdahl, T., Kåresen, R., Børresen-Dale, A.-L., and Jeffrey, S.S. (2007). TP53mutation status and gene expression profiles are powerful prognostic markers of breast cancer. Breast Cancer Res. *9*, 1–16. https://doi.org/10.1186/bcr1675.

19. Wang, Y., Helland, Å., Holm, R., Skomedal, H., Abeler, V.M., Danielsen, H.E., Tropé, C.G., Børresen-Dale, A.L., and Kristensen, G.B. (2004). TP53 mutations in early-stage ovarian carcinoma, relation to long-term survival. Br. J. Cancer *90*, 678–685. https://doi.org/10.1038/sj.bjc.6601537.

20. Cizkova, M., Susini, A., Vacher, S., Cizeron-Clairac, G., Andrieu, C., Driouch, K., Fourme, E., Lidereau, R., and Bièche, I. (2012). PIK3CA mutation impact on survival in breast cancer patients and in ERα, PR and ERBB2-based subgroups. Breast Cancer Res. *14*, R28–R29. https://doi.org/10.1186/bcr3113.

21. Nixon, M.J., Formisano, L., Mayer, I.A., Estrada, M.V., González-Ericsson, P.I., Isakoff, S.J., Forero-Torres, A., Won, H., Sanders, M.E., Solit, D.B., et al. (2019). PIK3CA and MAP3K1 alterations imply luminal A status and are associated with clinical benefit from pan-PI3K inhibitor buparlisib and letrozole in ER+ metastatic breast cancer. NPJ Breast Cancer *5*, 31–39. https://doi.org/10.1038/s41523-019-0126-6.

22. Mukohara, T. (2015). PI3K mutations in breast cancer: prognostic and therapeutic implications. Breast Cancer *7*, 111–123. https://doi.org/10.2147/BCTT.S60696.

23. Iqbal, N., and Iqbal, N. (2014). Human epidermal growth factor receptor 2 (HER2) in cancers: overexpression and therapeutic implications. Mol. Biol. Int. *2014*, 852748. https://doi.org/10.1155/2014/852748.

24. Singh, S., and Chakrabarti, R. (2019). Consequences of EMT-driven changes in the immune microenvironment of breast cancer and therapeutic response of cancer cells. J. Clin. Med. *8*, 642. https://doi.org/10.3390/jcm8050642.

25. Takaku, M., Grimm, S.A., and Wade, P.A. (2015). GATA3 in breast cancer: tumor suppressor or oncogene? Gene Expr. *16*, 163–168. https://doi.org/10.3727/105221615X14399878166113.

26. Overholtzer, M., Zhang, J., Smolen, G.A., Muir, B., Li, W., Sgroi, D.C., Deng, C.-X., Brugge, J.S., and Haber, D.A. (2006). Transforming properties of YAP, a candidate oncogene on the chromosome 11q22 amplicon. Proc. Natl. Acad. Sci. USA *103*, 12405–12410. https://doi.org/10.1073/pnas.0605579103.

27. Lamar, J.M., Stern, P., Liu, H., Schindler, J.W., Jiang, Z.-G., and Hynes, R.O. (2012). The Hippo pathway target, YAP, promotes metastasis through its TEAD-interaction domain. Proc. Natl. Acad. Sci. USA *109*, E2441–E2450. https://doi.org/10.1073/pnas.1212021109.

28. Liu, J., Shen, J.-X., Wu, H.-T., Li, X.-L., Wen, X.-F., Du, C.-W., and Zhang, G.-J. (2018). Collagen 1A1 (COL1A1) promotes metastasis of breast cancer and is a potential therapeutic target. Discov. Med. *25*, 211–223.

29. Bai, F., Zhang, L.-H., Liu, X., Wang, C., Zheng, C., Sun, J., Li, M., Zhu, W.-G., and Pei, X.-H. (2021). GATA3 functions downstream of BRCA1 to suppress EMT in breast cancer. Theranostics *11*, 8218–8233. https://doi.org/10.7150/thno.59280.

30. Caja, L., Dituri, F., Mancarella, S., Caballero-Diaz, D., Moustakas, A., Giannelli, G., and Fabregat, I. (2018). TGF-β and the Tissue Microenvironment: Relevance in Fibrosis and Cancer. Int. J. Mol. Sci. *19*, 1294. https://doi.org/10.3390/ijms19051294.

31. Chattopadhyay, I., Ambati, R., and Gundamaraju, R. (2021). Exploring the crosstalk between inflammation and epithelial-mesenchymal transition in cancer. Mediators Inflamm. *2021*, 9918379. https://doi.org/10.1155/2021/9918379.

32. Chockley, P.J., Chen, J., Chen, G., Beer, D.G., Standiford, T.J., and Keshamouni, V.G. (2018). Epithelial-mesenchymal transition leads to NK cell–mediated metastasis-specific immunosurveillance in lung cancer. J. Clin. Invest. *128*, 1384–1396. https://doi.org/10.1172/JCI97611.

33. Dumont, N., Liu, B., DeFilippis, R.A., Chang, H., Rabban, J.T., Karnezis, A.N., Tjoe, J.A., Marx, J., Parvin, B., and Tlsty, T.D. (2013). Breast fibroblasts modulate early dissemination, tumorigenesis, and metastasis through alteration of extracellular matrix characteristics. Neoplasia *15*, 249–262. https://doi.org/10.1593/neo.121950.

34. Gao, M.-Q., Kim, B.G., Kang, S., Choi, Y.P., Park, H., Kang, K.S., and Cho, N.H. (2010). Stromal fibroblasts from the interface zone of human breast carcinomas induce an epithelial–mesenchymal transition-like state in breast cancer cells *in vitro*. J. Cell Sci. *123*, 3507–3514. https://doi.org/10.1242/jcs.072900.

35. Sigurdsson, V., Hilmarsdottir, B., Sigmundsdottir, H., Fridriksdottir, A.J.R., Ringnér, M., Villadsen, R., Borg, A., Agnarsson, B.A., Petersen, O.W., Magnusson, M.K., and Gudjonsson, T. (2011). Endothelial induced EMT in breast epithelial cells with stem cell properties. PLoS One *6*, e23833. https://doi.org/10.1371/journal.pone.0023833.

36. Soon, P.S.H., Kim, E., Pon, C.K., Gill, A.J., Moore, K., Spillane, A.J., Benn, D.E., and Baxter, R.C. (2013). Breast cancer-associated fibroblasts induce epithelial-to-mesenchymal transition in breast cancer cells. Endocr. Relat. Cancer *20*, 1–12. https://doi.org/10.1530/ERC-12-0227.

37. Yang, F., Wei, Y., Cai, Z., Yu, L., Jiang, L., Zhang, C., Yan, H., Wang, Q., Cao, X., Liang, T., and Wang, J. (2015). Activated cytotoxic lymphocytes promote tumor progression by increasing the ability of 3LL tumor cells to mediate MDSC chemoattraction via Fas signaling. Cell. Mol. Immunol. *12*, 66–76. https://doi.org/10.1038/cmi.2014.21.

38. Cho, U., Kim, B., Kim, S., Han, Y., and Song, Y.S. (2018). Pro-inflammatory M1 Macrophage enhances metastatic potential of ovarian cancer cells through NF-κB activation. Mol. Carcinog. *57*, 235–242. https://doi.org/10.1002/mc.22750.

39. Gry, M., Rimini, R., Strömberg, S., Asplund, A., Pontén, F., Uhlén, M., and Nilsson, P. (2009). Correlations between RNA and protein expression profiles in 23 human cell lines. BMC Genom. *10*, 1–14. https://doi.org/10.1186/1471-2164-10-365.

40. Curtis, C., Shah, S.P., Chin, S.-F., Turashvili, G., Rueda, O.M., Dunning, M.J., Speed, D., Lynch, A.G., Samarajiwa, S., Yuan, Y., et al. (2012). The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. Nature *486*, 346–352. https://doi.org/10.1038/nature10983.

41. Barrett, T., Suzek, T.O., Troup, D.B., Wilhite, S.E., Ngau, W.-C., Ledoux, P., Rudnev, D., Lash, A.E., Fujibuchi, W., and Edgar, R. (2005). NCBI GEO: mining millions of expression profiles—database and tools. Nucleic Acids Res. *33*, D562–D566. https://doi.org/10.1093/nar/gki022.

42. Love, M.I., Anders, S., Kim, V., and Huber, W. (2015). RNA-Seq workflow: gene-level exploratory analysis and differential expression. F1000Res. *4*, 1070. https://doi.org/10.12688/f1000research.7035.1.

43. Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S., and Mesirov, J.P. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc. Natl. Acad. Sci. USA *102*, 15545–15550. https://doi.org/10.1073/pnas.0506580102.

44. Liberzon, A., Birger, C., Thorvaldsdóttir, H., Ghandi, M., Mesirov, J.P., and Tamayo, P. (2015). The molecular signatures database hallmark gene set collection. Cell Syst. *1*, 417–425. https://doi.org/10.1016/j.cels.2015.12.004.

45. Chang, M.T., Asthana, S., Gao, S.P., Lee, B.H., Chapman, J.S., Kandoth, C., Gao, J., Socci, N.D., Solit, D.B., Olshen, A.B., et al. (2016). Identifying recurrent mutations in cancer reveals widespread lineage diversity and mutational specificity. Nat. Biotechnol. *34*, 155–163. https://doi.org/10.1038/nbt.3391.

46. Robinson, J.T., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E.S., Getz, G., and Mesirov, J.P. (2011). Integrative genomics viewer. Nat. Biotechnol. *29*, 24–26. https://doi.org/10.1038/nbt.1754.

47. Morris, T.J., Butcher, L.M., Feber, A., Teschendorff, A.E., Chakravarthy, A.R., Wojdacz, T.K., and Beck, S. (2014). ChAMP: 450k chip analysis methylation pipeline. Bioinformatics *30*, 428–430. https://doi.org/10.1093/bioinformatics/btt684.

48. Du, P., and Bourgon, R. (2013). methyAnalysis: An R Package for DNA Methylation Data Analysis and Visualization. R Package Version 1.1.8.0.

49. Becht, E., Giraldo, N.A., Lacroix, L., Buttard, B., Elarouci, N., Petitprez, F., Selves, J., Laurent-Puig, P., Sautès-Fridman, C., Fridman, W.H., and de Reyniès, A. (2016). Estimating the population abundance of tissue-infiltrating immune and stromal cell populations using gene expression. Genome Biol. *17*, 218–220. https://doi.org/10.1186/s13059-016-1070-5.

50. Newman, A.M., Liu, C.L., Green, M.R., Gentles, A.J., Feng, W., Xu, Y., Hoang, C.D., Diehn, M., and Alizadeh, A.A. (2015). Robust enumeration of cell subsets from tissue expression profiles. Nat. Methods *12*, 453–457. https://doi.org/10.1038/nmeth.3337.

51. Ritchie, M.E., Phipson, B., Wu, D., Hu, Y., Law, C.W., Shi, W., and Smyth, G.K. (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. *43*, e47. https://doi.org/10.1093/nar/gkv007.

52. Li, J.-H., Liu, S., Zhou, H., Qu, L.-H., and Yang, J.-H. (2014). starBase v2. 0: decoding miRNA-ceRNA, miRNA-ncRNA and protein–RNA interaction networks from

large-scale CLIP-Seq data. Nucleic Acids Res. *42*, D92–D97. https://doi.org/10.1093/nar/gkt1248.

53. Cheng, L., Wang, P., Tian, R., Wang, S., Guo, Q., Luo, M., Zhou, W., Liu, G., Jiang, H., and Jiang, Q. (2019). LncRNA2Target v2. 0: a comprehensive database for target genes of lncRNAs in human and mouse. Nucleic Acids Res. *47*, D140–D144. https://doi.org/10.1093/nar/gky1051.

54. Zhao, H., Shi, J., Zhang, Y., Xie, A., Yu, L., Zhang, C., Lei, J., Xu, H., Leng, Z., Li, T., et al. (2020). LncTarD: A manually-curated database of experimentally-supported functional lncRNA–target regulations in human diseases. Nucleic Acids Res. *48*, D118–D126. https://doi.org/10.1093/nar/gkz985.

55. Bao, Z., Yang, Z., Huang, Z., Zhou, Y., Cui, Q., and Dong, D. (2019). LncRNADisease 2.0: an updated database of long non-coding RNA-associated diseases. Nucleic Acids Res. *47*, D1034–D1037. https://doi.org/10.1093/nar/gky905.

56. Paraskevopoulou, M.D., Vlachos, I.S., Karagkouni, D., Georgakilas, G., Kanellos, I., Vergoulis, T., Zagganas, K., Tsanakas, P., Floros, E., Dalamagas, T., and Hatzigeorgiou, A.G. (2016). DIANA-LncBase v2: indexing microRNA targets on non-coding transcripts. Nucleic Acids Res. *44*, D231–D238. https://doi.org/10.1093/nar/gkv1270.

57. Kozomara, A., Birgaoanu, M., and Griffiths-Jones, S. (2019). miRBase: from microRNA sequences to function. Nucleic Acids Res. *47*, D155–D162. https://doi.org/10.1093/nar/gky1141.

58. Shannon, P., Markiel, A., Ozier, O., Baliga, N.S., Wang, J.T., Ramage, D., Amin, N., Schwikowski, B., and Ideker, T. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res. *13*, 2498–2504. https://doi.org/10.1101/gr.1239303.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
RNA-seq, miRNA, lncRNA, protein expression values, DNA methylation values and clinical data for the TCGA cohort	TCGA	https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga
METABRIC/Nature 2012	cBioPortal	https://www.cbioportal.org/
E-MTAB-6703	ArrayExpress	https://www.ebi.ac.uk/arrayexpress/
GSE96058	Gene Expression Omnibus (GEO)	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE96058
GSE20685	Gene Expression Omnibus (GEO)	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE20685
GSE54275	Gene Expression Omnibus (GEO)	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE54275
GSE2109	Gene Expression Omnibus (GEO)	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE2109
TCPA	The Cancer Proteome Atlas (TCPA)	https://www.tcpaportal.org/tcpa/
HPA	The Human Protein Atlas (HPA)	https://www.proteinatlas.org/
starBase	starBase	http://starbase.sysu.edu.cn/
LncRNA2Traget v2.0	LncRNA2Traget v2.0	http://www.lncrna2target.org/
LncTarD	LncTarD	https://lnctard.bio-database.com/
LncBase_Predicted_v2	LncBase_Predicted_v2	http://carolina.imis.athena-innovation.gr/diana_tools/web/index.php?r=lncbasev2%2Findex-predicted
miRbase	miRbase	http://www.mirbase.org/
Software and algorithms		
R version 4.0.2	R project	https://www.r-project.org/
DESeq2 version 1.36.0	Github	https://github.com/mikelove/DESeq2
GSEA	Gene Set Enrichment Analysis (GSEA)	https://www.gsea-msigdb.org/gsea/index.jsp
Survival version 3.4.0	Github	https://github.com/therneau/survival
ChAMP version 2.26.0	Github	https://github.com/swsoyee/ChAMP
MethyAnalysis version 1.8.0	Github	https://github.com/yuanjinzhang/methyAnalysis
Glmnet version 4.0	Github	https://github.com/cran/glmnet
Renyi version	Github	https://github.com/daijiang/renyi
CIBERSORT	CIBERSORT	https://cibersort.stanford.edu/
MCP-counter	MCP-counter	https://github.com/ebecht/MCPcounter
Limma version 3.52.2	Github	https://github.com/Bioconductor-mirror/limma
Cytoscape version 3.7.2	Cytoscape	https://cytoscape.org/download.html

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources should be directed to and will be fulfilled by the Lead Contact Shuangtao Zhao (zst-1981@163.com).

Materials availability

All unique/stable reagents generated in this study could be found in the supplemental information.

Data and code availability

All the data in the current study are available from the public datasets listed in the key resources table.

The code used in this study are available at https://github.com/nayangmeihao/Luminal-B-study.git.

Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS

Ethics approval and consent to participate

Ethical approval was obtained from the institutional review committee of Beijing Tuberculosis and Thoracic Tumor Research Institute/Beijing Chest Hospital of Capital Medical University. And The design and performance of the study are in accordance with the Declaration of Helsinki. Signed informed consent was obtained from all participants before inclusion, allowing analysis of tumor tissue, blood samples and clinical data.

Patient cohort and sample collection

The raw datasets were downloaded from The Cancer Genome Atlas (TCGA, https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga), METABRIC/Nature 2012 from cBioPortal[40] for Cancer Genomics (https://www.cbioportal.org/), E-MTAB-6703 form ArrayExpress (https://www.ebi.ac.uk/arrayexpress/) and GSE96058, GSE20685, GSE54275 and GSE2109 from Gene Expression Omnibus (GEO, https://www.ncbi.nlm.nih.gov/geo/) database.[41] A group of 929 patients and 98 normal samples from TCGA were treated as training set and another 5,314 patients from METABRIC/Nature 2012 (n=1,904), E-MTAB-6703 (n=1,176), GSE96058 (n=3,410), GSE54275 (n=243) GSE20685 (n=327) and GSE2109 (n = 354) were treated as validation sets. The multi-omics data types included in TCGA for LBBC patients with overall survival (OS) times $\leq$120 months had: whole exome DNA sequences (n=187), DNA methylation arrays (n=129), RNA sequences (n=187), miRNA sequences (n=181), lncRNA sequences (n=187) and reverse phase protein arrays (RPPA, n=160). And similar LBBC patients were selected to assay RNA sequencing in METABRIC/Nature 2012 (n=263) and GSE96058 (n=656). Based on data from all breast cancer patients, researchers have published articles between October 2012 and March 2018, which were selected based on available GEP results and clinical data. All the diagnoses were confirmed on the basis of WHO classification criteria. These three datasets were composed of 655 patients with surgery, 1,202 patients with chemotherapy, 395 patients with radiotherapy, and 694 patients with hormone therapy, respectively. Clinical characteristics at presentation in the validation sets were similar with that in the training data set in terms of age (> 60 in 64%, p = 1.14E-10) and ER status (98% with positive, p = 1.41E-04). All the clinical information was summarized in the supplemental information (see Tables S1, S2, S3, S6, S7, and S8). In addition, validation on the E-MTAB-6703 (n = 361), GSE20685 (n = 86), GSE54275 (n = 74) and GSE2109 (n = 105) datasets for clusters A and B using the same analysis as the training set obtained expression results similar to those from the training set (see Figures S6A–S6D).

METHOD DETAILS

Identification of differently expressed genes (DEGs) and enriched signaling pathways

The HTSeq raw counts including mRNA, miRNA and lncRNA were downloaded from TCGA dataset and then processed by DESeq2[42] software to identify DEGs between Clusters A and B. A cut-off gene expression was defined as fold change among $\geq$ 2 or $\leq$ -2 and an FDR q-value as < 0.05 to select the most significant DEGs. A ranked list of genes was obtained based on DESeq2 FDR q-values for all coding genes and processed by Gene Set Enrichment Analysis (GSEA)[43] against the curated gene sets from Molecular Signature Database (MSigDB)[44] to filter the significantly enriched signaling pathways. The most significantly enriched signaling pathways were selected based on a cut-off value of FDR q-value as $\leq$ 0.05.

QUANTIFICATION AND STATISTICAL ANALYSIS

Samples classification and validation based on mRNA data

We processed the TCGA data for missing values, removing genes that were missing expression in more than 70% of the samples, and finally obtained 16,875 from 19,641 mRNAs (Figure S1A), and then differential analysis were performed between normal (n=98) and tumor (n=1,072) yielded 4,248 DEGs (Figure S1B). The top 1,000 with the highest standard deviation and mean value out of 4,248 DEGs were selected as variable genes (Figure S1C) for explore the best clusters and two subtypes (Clusters A and B) were obtained by hierarchical clustering using the normalised values of the function normTransform in the DESeq2 package, which was visualised by PCA analysis (Figure S1D). Then a signature including 20 genes was identified as clustering biomarkers from Lasso regression and Random Forest analysis, based on which subgroups were divided based on the mRNA expression value in TCGA dataset and validated in two independent datasets (METABRIC/Nature 2012 and GSE96058).

Mutation signature analysis

All the somatic mutations data was downloaded directly from cBioPortal for Cancer Genomics (https://www.cbioportal.org/). And then the profiles of mutational signatures were displayed with Oncoprint plot and then compared between Clusters A and B. The mutated amino acid was identified as a recurrent hotspot (statistically significant) in a population-scale cohort of tumor samples with various cancer types using methods partially from Chang *et* al.[45]

DNA copy number analysis

The DNA copy number files were also fetched from TCGA (https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga), and then they were loaded into IGV[46] for visualization. "CNTools" (v1.24.0) R package was applied to identify copy number gains (log2 copy ratios > 0.3) or losses (log2 copy ratios < -0.3) at the genes level. The total number of genes with copy number gains or losses per sample was defined by the burden of copy number gain or loss. And the fraction of changed genome was identified as the proportion of the genome with copy number gains or losses against the total length of genome with copy number profiling.

DNA methylation analysis

The DNA methylation raw data was obtained from TCGA (https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga) including 129 LBBC patients. And then 26,284 significantly different methylation sites (DMSs) were discovered between Clusters A (n=65) and B (n=64) among 485,577 DNA methylation sites using "ChAMP (The Chip Analysis Methylation Pipeline)"[47] R package. Among them, caner-related genes were selected based on the mutation signature obtained above. To display the methylation sites and the CpG island near these sites on the chromosome, the R package "methyAnalysis"[48] was applied to the biologically important genes between Clusters A and B of LBBC samples.

Deconvolution of the cellular composition with LBBC samples

Two deconvolution approaches were adopted to evaluate the immune infiltration of LBBC samples. Among them, the MCP-counter[49] was applied to producing the absolute abundance scores for 8 major immune cell types (neutrophils, myeloid dendritic cells, monocytic lineage cells, B lymphocytes, NK cells, $CD3^+$ T cells, $CD8^+$ T cells and cytotoxic lymphocytes), fibroblasts and endothelial cells; the CIBERSORT algorithm[50] was utilized to assess the relative cellular fraction of 22 immune cell types. The log2-transformed HTSeq counts matrix were used as the input data for both algorithms, and the LM22 leukocyte genes signature used as input for the CIBERSORT analysis. The deconvolution profiles were performed with hierarchical clustering method and compared across two clusters and also between $HER2^-$ and $HER2^+$ groups.

Analysis of reverse phase protein array (RPPA)

The RPPA data (level 4), which includes information on 244 proteins (including phosphorylation data), was obtained from The Cancer Proteome Atlas (TCPA, https://www.tcpaportal.org/tcpa/) for all breast cancer samples. Protein expression data in normal breast tissues was downloaded from The Human Protein Atlas (HPA, https://www.proteinatlas.org/) and proteins with median (n=9,677) or high (n=2,888) values were defined as a control to select cancer specific proteins. A total of 175 proteins detected among ≥ 70%

LBBC samples were enrolled into calculating the different expression proteins (DEPs) between Clusters A (n=84) and B (n=76) using the R package "limma".[51] Then only 19 proteins were discovered across the two subtypes. Group enriched proteins were filtered by applying cut-off of the absolute expression fold change as $\geq$ 2 and FDR q-value as < 0.05.

Construction of competitive endogenous RNA (ceRNAs) network

To construct the ceRNA network, DEGs analysis was applied to selecting the candidate genes (12 lncRNAs, 8 miRNAs, 574 mRNAs) between Clusters A and B of LBBC patients in TCGA. These databases (starBase,[52] LncRNA2Traget v2.0,[53] LncTarD[54] and lncRNADisease_v2.0[55]) were selected to predict the interactions between lncRNAs and mRNAs. And LncBase_Predicted_v2[56] database was used to explore the interactions between lncRNAs and miRNAs. Additionally, miRbase[57] database was adopted to validate the interactions (scores $\geq$ 80) between miRNAs and mRNAs. Finally, ceRNA network was evaluated by computing the Spearman's correlation coefficients belonging to each ceRNA network, and the final ceRNA network including lncRNAs, targeted miRNAs and targeted mRNAs was visualized by Cytoscape software 3.7.2.[58]

Statistical analysis

The statistical methods used in this study were performed in the R statistical environment (v4.0.2) in addition to the algorithms mentioned above. Shapiro-Wilk test was used to evaluate the normal distribution before DEGs analysis. The statistical significance of differences observed between clusters was determined by the *t* test for normal distribution data and Wilcoxon test for non-normal distribution data when comparing continuous variables, and the Fisher's Exact test when comparing frequencies of clinical factors. And the Benjamini-Hochberg algorithm was conducted to compute a false discovery rate (FDR) adjusted *p*-value (or *q*-value) in order to control FDR and correct *p*-values from multiple testing. The log-rank tests or Renyi tests (crossed survival curves) were applied to univariate survival analysis in the Kaplan-Meier plots between two clusters. LASSO regression model from "glmnet" package investigated the significant signature associated with survival between two subtypes. The Random Forest algorithm evaluated the importance of the DEGs or DEPs between two groups associated with survival or classification. Spearman's correlation coefficient was calculated to evaluate the association between two continuous variables. Hypothesis testing was performed in a two-sided manner, with *p*-value or adj. *p*-value (if applicable) < 0.05 considered to be statistically significant.