Table S1: Breast Pre-cancer Atlas Retrospective RAHBT Cohort for LCM. Related to Figure 1 and Table 1.

	RAHBT					
	DCIS without recurrence (N=184)	**DCIS with Ipsilateral DCIS Recurrence (N=17)**	**DCIS with Ipsilateral Invasive Recurrence (N=29)**	**DCIS with Contralateral DCIS (N=19)**	**DCIS with Contralateral Invasive Disease (N=16)**	**RAHBT Total (N=265)**
Year of Diagnosis						
Median	2002	2005	2000	2002	1991	2002
Age at Diagnosis						
Median	53	57	48	57	54	53
Mean (±SD)	55.6 (±11.4)	61.1 (±12.8)	49.9 (±10.3)	55.9 (±9.9)	58.2 (±12.2)	55.5 (±11.5)
Grade						
1	51 [27.7%]	3 [17.6%]	8 [27.6%]	7 [36.8%]	4 [25.0%]	73 [27.5%]
2	65 [35.3%]	7 [41.2%]	15 [51.7%]	8 [42.1%]	7 [43.8%]	102 [38.5%]
3	65 [35.3%]	6 [35.3%]	4 [13.8%]	4 [21.1%]	5 [31.3%]	84 [31.7%]
Missing	2 [1.1%]	1 [5.9%]	2 [6.9%]	0	0	6 [2.3%]
Pathologic Tumor Size						
Median	NA	NA	NA	NA	NA	NA
Mean (±SD)	NA	NA	NA	NA	NA	NA
Marker Status						
ER(+)	123 [66.8%]	11 [64.7%]	24 [82.8%]	17 [89.5%]	14 [87.5%]	189 [71.3%]
ER(-)	61 [33.2%]	6 [35.3%]	5 [17.2%]	2 [10.5%]	2 [12.5%]	76 [28.7%]
ER(+) Dx before 2000	46 [25.0%]	2 [11.8%]	10 [34.5%]	7 [36.8%]	9 [56.2%]	74 [27.9%]
ER(+) Dx 2000 & after	77 [41.8%]]	9 [52.9%]	14 [48.3%]	10 [52.6%]	5 [31.2%]	67 [25.3%]
ER(-) Dx before 2000	29 [15.8%	3 [17.6%]	4 [13.8%]	2 [10.5%]	1 [6.3%]	87 [32.8%]
ER(-) Dx 2000 & after	32 [17.4%]	3 [17.6%]	1 [3.4%]	0	1 [6.3%]	37 [14.0%]
Treatment						
Lumpectomy w Radiation	91 [49.5%]	12 [70.6%]	18 [62.1%]	8 [42.1%]	9 [50.0%]	17 [51.7%]
Lumpectomy no Radiation	34 [18.5%]	5 [29.4%]	7 [24.1%]	1 [5.3%]	0	47 [17.7%]
Lumpectomy Radiation Unknown	3 [1.6%]	0	1 [3.4%]	1 [5.3%]	1 [6.3%]	6 [2.3%]
Mastectomy	56 [30.4%]	0	3 [10.3%]	9 [47.4%]	7 [43.8%]	75 [28.3%]

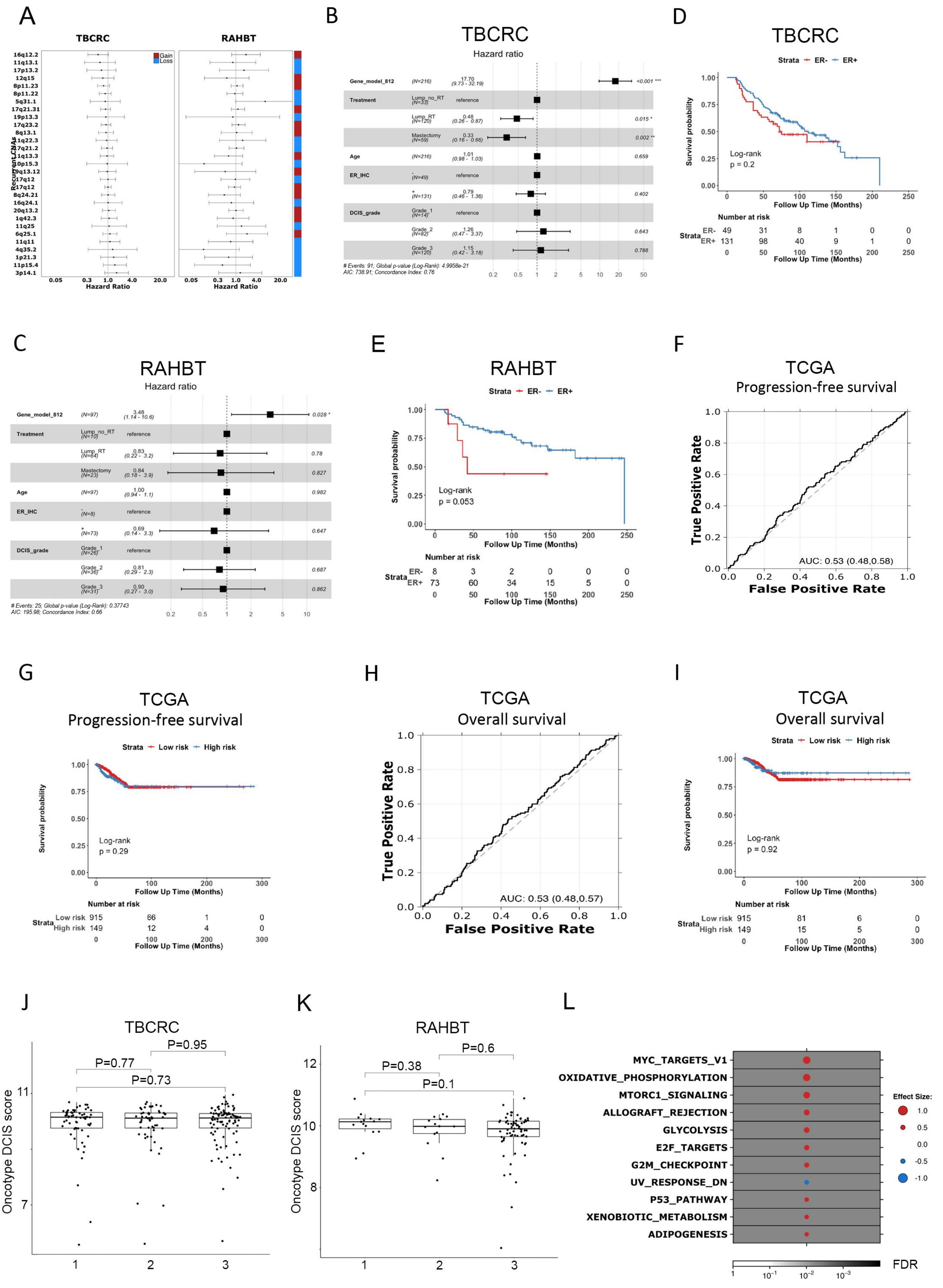
Time to Recurrence* (months)						
Median	111*	49	80	81	56	62.3
Mean (±SD)	127.1 (±84.4)	61.5 (±43.6)	93.2 (±74.2)	107.3 (±89.1)	71.3 (±56.3)	85.5 (±70.6)
Margins						
Ink on tumor	9 [4.9%]	2 [11.8%]	2 [6.9%]	3 [15.8%]	2 [12.5%]	17 [6.4%]
<2mm	24 [13.0%]	3 [17.6%]	3 [10.3%]	3 [15.8%]	3 [18.8%]	36 [13.6%]
At least 2mm	27 [14.7%]	4 23.5%]	2 [6.9%]	1 [5.3%]	1 [6.3%]	38 [14.3%]
Clear, unknown mm	81 [44.0%]	8 [47.1%]	17 [58.6%]	10 [52.6%]	4 [25.0%]	118 [44.5%]
Missing	43 [23.4%]	0	5 [17.2%]	2 [10.5%]	6 [37.5%]	56 [21.1%]
Race						
White	138 [75.0%]	12 [70.6%]	22 [75.9%]	15 [78.9%]	10 [62.5%]	197 [74.3%]
Black	45 [24.5%]	5 [29.4%]	7 [24.1%]	3 [15.8%]	6 [37.5%]	66 [24.9%]
Asian	0	0	0	0	0	0
Pacific Islander	0	0	0	1 [5.3%]	0	1 [0.4%]
Other	0	0	0	0	0	0
Unknown	1 [0.5%]	0	0	0	0	1 [0.4%]

*To end of follow-up for no recurrence

Table S2: Breast Pre-cancer Atlas Multi-scale Characterization Assays. Related to Figure 1.

Assay	Scale	Type of Data	Integration and validation with other assays	Analyzed in RAHBT	Analyzed in RAHBT LCM	Analyzed in TBCRC
RNA-seq (Single duct, tumor microenvironment)	Duct, organ, normal tissue	Whole transcriptome gene expression profiling per single duct	Gene expression and prediction of cell type composition (CibersortX) confirmed by MIBI (single cell)	Figure 2B, C, E, G Figure 3 Figure 5F	Figure 4B, C Figure 6A-C, E, G	Figure 2A, D, F Figure 3, Figure 4A, C Figure 5F
Low-pass whole genome DNA-seq	Duct and adjacent normal	CNV profiling per single duct	Analysis of CNV supported by RNA-seq (single duct) and MIBI (single cell)	Figure 5A-E	NA	Figure 5A-E
Multiplex IHC (MIBI)	Cell	1. Cell type 2. Proteomic analysis	Analysis of protein expression and cell type supported by RNA-seq of ducts (CibersortX)	NA	Figure 4D, E, F Figure 6D, F	NA

Figure S1

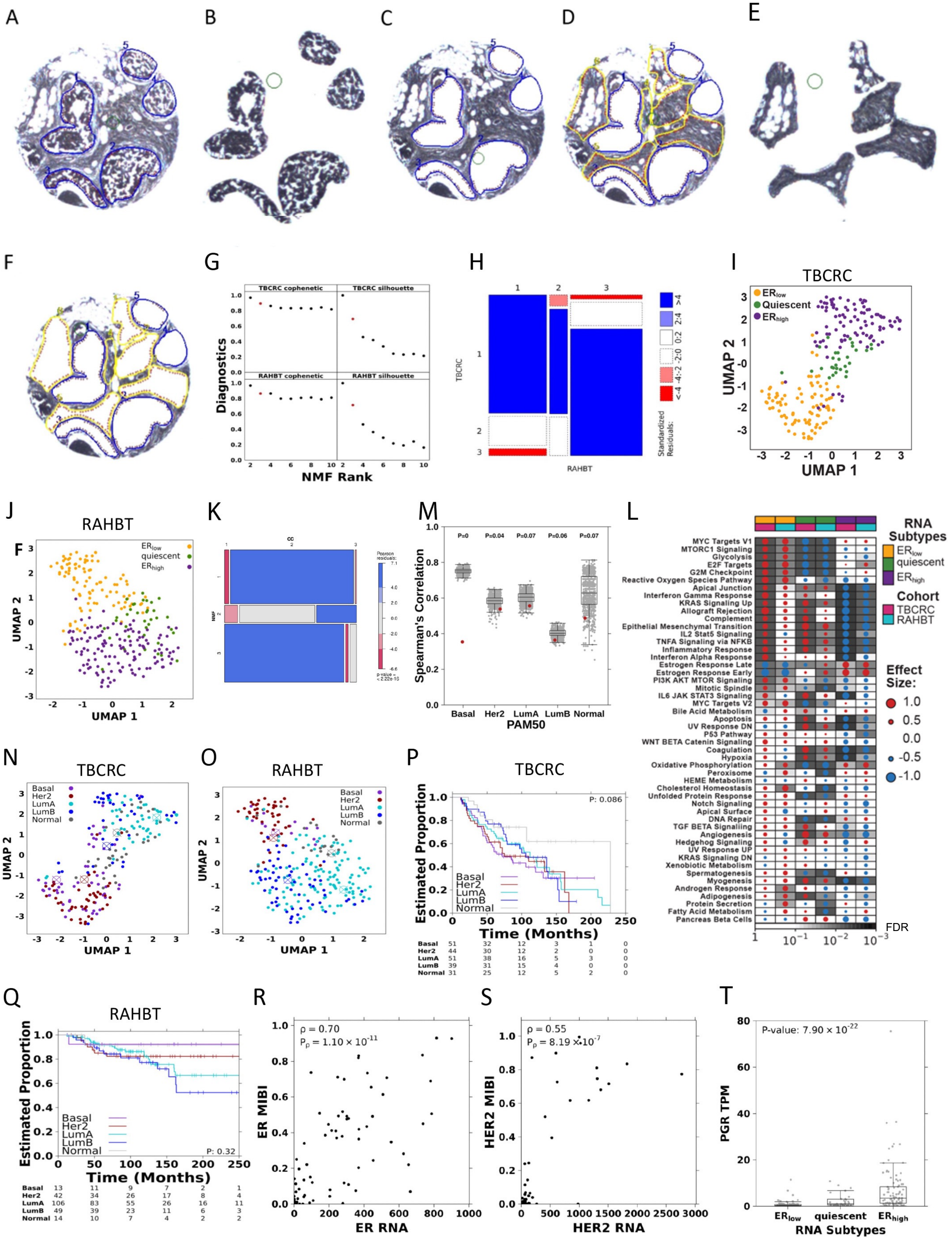


A
TBCRC
RAHBT
Gain
Loss
Recurrent CNAs
16q12.2
11q13.1
17p13.2
12q15
8p11.23
8p11.22
5q31.1
17q21.31
19p13.3
17q23.2
8q13.1
11q22.3
17q21.2
11q13.3
10p15.3
19q13.12
17q12
17q12
8q24.21
16q24.1
20q13.2
1q42.3
11q25
6q25.1
11q11
4q35.2
1p21.3
11p15.4
3p14.1
0.05 0.3 1.0 4.0 20.0
Hazard Ratio
B
TBCRC
Hazard ratio
Gene_model_812 (N=216) 17.70 (9.73 - 32.19) <0.001 ***
Treatment Lump_no_RT (N=33) reference
Lump_RT (N=120) 0.48 (0.26 - 0.87) 0.015 *
Mastectomy (N=59) 0.33 (0.16 - 0.66) 0.002 **
Age (N=216) 1.01 (0.98 - 1.03) 0.659
ER_IHC - (N=49) reference
+ (N=131) 0.79 (0.46 - 1.36) 0.402
DCIS_grade Grade_1 (N=14) reference
Grade_2 (N=82) 1.26 (0.47 - 3.37) 0.643
Grade_3 (N=120) 1.15 (0.42 - 3.18) 0.788
# Events: 91; Global p-value (Log-Rank): 4.9958e-21
AIC: 738.91; Concordance Index: 0.76
0.2 0.5 1 2 5 10 20 50
D
TBCRC
Strata ER- ER+
Survival probability
1.00 0.75 0.50 0.25 0.00
Log-rank
p = 0.2
0 50 100 150 200 250
Follow Up Time (Months)
Number at risk
Strata ER- 49 31 8 1 0 0
ER+ 131 98 40 9 1 0
C
RAHBT
Hazard ratio
Gene_model_812 (N=97) 3.48 (1.14 - 10.6) 0.028 *
Treatment Lump_no_RT (N=10) reference
Lump_RT (N=64) 0.83 (0.22 - 3.2) 0.78
Mastectomy (N=23) 0.84 (0.18 - 3.9) 0.827
Age (N=97) 1.00 (0.94 - 1.1) 0.982
ER_IHC - (N=8) reference
+ (N=73) 0.69 (0.14 - 3.3) 0.647
DCIS_grade Grade_1 (N=26) reference
Grade_2 (N=36) 0.81 (0.29 - 2.3) 0.687
Grade_3 (N=31) 0.90 (0.27 - 3.0) 0.862
# Events: 25; Global p-value (Log-Rank): 0.37743
AIC: 195.98; Concordance Index: 0.66
0.2 0.5 1 2 5 10
E
RAHBT
Strata ER- ER+
Survival probability
1.00 0.75 0.50 0.25 0.00
Log-rank
p = 0.053
0 50 100 150 200 250
Follow Up Time (Months)
Number at risk
Strata ER- 8 3 2 0 0 0
ER+ 73 60 34 15 5 0
F
TCGA
Progression-free survival
True Positive Rate
False Positive Rate
0.0 0.2 0.4 0.6 0.8 1.0
AUC: 0.53 (0.48,0.58)
G
TCGA
Progression-free survival
Strata Low risk High risk
Survival probability
1.00 0.75 0.50 0.25 0.00
Log-rank
p = 0.29
0 100 200 300
Follow Up Time (Months)
Number at risk
Strata Low risk 915 66 1 0
High risk 149 12 4 0
H
TCGA
Overall survival
True Positive Rate
False Positive Rate
0.0 0.2 0.4 0.6 0.8 1.0
AUC: 0.53 (0.48,0.57)
I
TCGA
Overall survival
Strata Low risk High risk
Survival probability
1.00 0.75 0.50 0.25 0.00
Log-rank
p = 0.92
0 100 200 300
Follow Up Time (Months)
Number at risk
Strata Low risk 915 81 6 0
High risk 149 15 5 0
J
TBCRC
Oncotype DCIS score
7 11
P=0.77
P=0.95
P=0.73
1 2 3
K
RAHBT
Oncotype DCIS score
8 10 12
P=0.38
P=0.6
P=0.1
1 2 3
L
MYC_TARGETS_V1
OXIDATIVE_PHOSPHORYLATION
MTORC1_SIGNALING
ALLOGRAFT_REJECTION
GLYCOLYSIS
E2F_TARGETS
G2M_CHECKPOINT
UV_RESPONSE_DN
P53_PATHWAY
XENOBIOTIC_METABOLISM
ADIPOGENESIS
Effect Size:
1.0 0.5 0.0 -0.5 -1.0
FDR
1 10^{-1} 10^{-2} 10^{-3}

Figure S1: Outcome analysis. Related to Figure 2.

A) Forest plot showing hazard ratios from CoxPH modeling of the 29 recurrent copy number aberrations (CNAs) association with progression. Vertical dotted line represent hazard ratio = 1. The covariate on the right indicates if the CNA is a gain (red) or loss (blue). **B-C**) Forest plot of multivariable Cox regression analysis including 812 gene classifier (high- vs. low-risk groups), treatment, age, DCIS grade, and ER status by IHC for any iBE with full follow-up in TBCRC (**B**) and RAHBT (**C**). **D-E**) Kaplan-Meier plot of time to progression (any iBE, full follow-up) stratified by clinical ER status in TBCRC (**D**) and RAHBT (**E**). P-values from log-rank tests. **F**) ROC curve of the 812 gene classifier tested towards progression-free survival in TCGA IBC samples. **G**) Kaplan-Meier plot of time to progression in TCGA IBC samples. P-value from log-rank test. **H**) ROC curve of the 812 gene classifier tested towards overall survival in TCGA IBC samples. **I**) Kaplan-Meier plot of time to death in TCGA IBC samples. P-value from log-rank test. **J-K**) Box plot of Oncotype DX DCIS score in the three different outcome groups in TBCRC (**J**) and RAHBT (**K**). 1: DCIS with DCIS recurrence. 2: DCIS with IBC recurrence. 3: DCIS with no recurrence. The score was calculated as described by Solin *et al*[18] but based on RNA-seq raw reads instead of Ct values from RT-qPCR. P-values from Wilcoxon rank sum test. Boxplots represent median, 0.25 and 0.75 quantiles with whiskers at 1.5x interquartile range. **L**) Gene Set Enrichment Analysis (GSEA) with Hallmark gene sets of the differentially expressed genes between cases with any iBE at 5 years after treatment vs the rest in TBCRC. Only significant pathways shown (FDR<0.05). Pathways sorted by effect size. Size of the dot and color represents the magnitude and direction of pathway deregulation, *i.e.,* blue indicates the pathway is downregulated while red indicates the pathway is upregulated. Background shading indicates false discovery rate (FDR). Effect size and FDR from GSEA algorithm.

Figure S2



A
B
C
D
E
F
G
TBCRC cophenetic
TBCRC silhouette
RAHBT cophenetic
RAHBT silhouette
Diagnostics
NMF Rank
H
TBCRC
RAHBT
Standardized Residuals:
>4
2:4
0:2
-2:0
-4:-2
<-4
I
TBCRC
ER$_{low}$
Quiescent
ER$_{high}$
UMAP 2
UMAP 1
J
RAHBT
ER$_{low}$
quiescent
ER$_{high}$
UMAP 2
UMAP 1
K
CC
NMF
Pearson residuals:
p-value = < 2.22e-16
M
P=0
P=0.04
P=0.07
P=0.06
P=0.07
Spearman's Correlation
Basal
Her2
LumA
LumB
Normal
PAM50
L
RNA Subtypes
ER$_{low}$
quiescent
ER$_{high}$
Cohort
TBCRC
RAHBT
Effect Size:
1.0
0.5
0.0
-0.5
-1.0
FDR
MYC Targets V1
MTORC1 Signaling
Glycolysis
E2F Targets
G2M Checkpoint
Reactive Oxygen Species Pathway
Apical Junction
Interferon Gamma Response
KRAS Signaling Up
Allograft Rejection
Complement
Epithelial Mesenchymal Transition
IL2 Stat5 Signaling
TNFA Signaling via NFKB
Inflammatory Response
Interferon Alpha Response
Estrogen Response Late
Estrogen Response Early
PI3K AKT MTOR Signaling
Mitotic Spindle
IL6 JAK STAT3 Signaling
MYC Targets V2
Bile Acid Metabolism
Apoptosis
UV Response DN
P53 Pathway
WNT BETA Catenin Signaling
Coagulation
Hypoxia
Oxidative Phosphorylation
Peroxisome
HEME Metabolism
Cholesterol Homeostasis
Unfolded Protein Response
Notch Signaling
Apical Surface
DNA Repair
TGF BETA Signaliing
Angiogenesis
Hedgehog Signaling
UV Response UP
KRAS Signaling DN
Xenobiotic Metabolism
Spermatogenesis
Myogenesis
Androgen Response
Adipogenesis
Protein Secretion
Fatty Acid Metabolism
Pancreas Beta Cells
N
TBCRC
Basal
Her2
LumA
LumB
Normal
UMAP 2
UMAP 1
O
RAHBT
Basal
Her2
LumA
LumB
Normal
UMAP 2
UMAP 1
P
TBCRC
P: 0.086
Estimated Proportion
Time (Months)
Basal 51 32 12 3 1 0
Her2 44 30 12 2 0 0
LumA 51 38 16 5 3 0
LumB 39 31 15 4 0 0
Normal 31 25 12 5 2 0
Q
RAHBT
P: 0.32
Estimated Proportion
Time (Months)
Basal 13 11 9 7 2 1
Her2 42 34 26 17 8 4
LumA 106 83 55 26 16 11
LumB 49 39 23 11 6 3
Normal 14 10 7 4 2 2
R
ρ = 0.70
P$_ρ$ = 1.10 × 10^{-11}
ER MIBI
ER RNA
S
ρ = 0.55
P$_ρ$ = 8.19 × 10^{-7}
HER2 MIBI
HER2 RNA
T
P-value: 7.90 × 10^{-22}
PGR TPM
ER$_{low}$
quiescent
ER$_{high}$
RNA Subtypes

Figure S2: LCM dissection of RAHBT epithelial and stromal samples, and subtype characterization, related to Figure 4.

A) Marked DCIS epithelium (blue) prior to dissection. **B**) Dissected DCIS epithelium on cap. **C**) Remaining tissue on slide after dissection. **D**) Marked stroma (yellow) adjacent to dissected DCIS epithelium (blue, panel A-C) prior to dissection. **E**) Dissected stroma on cap. **F**) Remaining tissue on slide after dissection of DCIS epithelium and adjacent stroma. All images taken at 2X magnification. **G**) NMF diagnostic scatterplots show cophenetic and silhouette values with increasing numbers of clusters in TBCRC and RAHBT. **H**) Mosaic plot showing concordance of de novo clustering in RAHBT *vs* clusters determined from centroids identified in TBCRC. Blue indicates an enrichment while red indicates a depletion. **I-J)** UMAP projection of DCIS transcriptome colored by *de novo* RNA clusters in TBCRC (**I**) and RAHBT (**J**). **K)** Mosaic plot showing concordance between clusters obtained by NMF and consensus clustering (CC) of TBCRC cohort (85.6% concordance). **L)** GSEA Hallmark pathway analysis of each cluster vs rest for TBCRC and RAHBT LCM in full . Size of the dot and color represents the magnitude and direction of pathway deregulation, *i.e.,* blue indicates the pathway is downregulated while red indicates the pathway is upregulated. Background shading indicates FDR. Effect size and FDR from GSEA algorithm. **M**) Boxplot shows median Spearman ρ of DCIS and IBC samples with PAM50 centroids. IBC samples were randomly downsampled 1,000 times to match the DCIS cohort size. Grey dots present median Spearman ρ of downsampled cohort. The red dot represents the median Spearman ρ of the DCIS cohort. P-values were calculated as 1- the proportion of downsampled IBC cohorts with median Spearman ρ greater than the DCIS cohort. Boxplot represents median, 0.25 and 0.75 quantiles with whiskers at 1.5x interquartile range. **N-O)** UMAP projection of DCIS transcriptome in TBCRC (**N**) and RAHBT LCM (**O**) colored by PAM50 subtype. Large circles represent the PAM50 subtype centroids. **P-Q**) Kaplan-Meier plots of time to progression in PAM50 subtypes in TBCRC (**P**) and RAHBT (**Q**). P-values from log-rank test. **R-S**) Correlation between mRNA abundance and MIBI protein levels of *ESR1*/ER (**R**) and *ERBB2*/HER2 (**S**). Correlations coefficients and P-values from Spearman's correlation. **T**) *PGR* mRNA abundance in the three DCIS subtypes. P-value from Kruskal-Wallis test. Boxplot represents median, 0.25 and 0.75 quantiles with whiskers at 1.5x interquartile range.

Table S4: Associations between recurrent CNAs and sequencing coverage or cohort. Related to Figure 5.

	Association with coverage			**Association with cohort**		
Recurrent CNAs	**Difference in medians**	**P-values**	**FDR**	**Difference in medians**	**P-value**	**FDR**
Amp_1q42.3	1550611	0.321	0.416	0.033	0.757	0.955
Del_17p13.2	2800768	0.931	0.931	0.058	0.444	0.804
Amp_17q12	2800768	0.651	0.697	0.044	0.587	0.895
Amp_17q23.2	1737324	0.758	0.784	0.044	0.954	0.996
Del_16q24.1	1217342	0.349	0.431	0.222	0.102	0.591
Amp_8q24.21	2532147	0.882	0.897	-0.071	0.348	0.745
Del_11q25	-6290327	0.618	0.674	-0.196	0.644	0.903
Amp_8q13.1	-8252690	0.139	0.206	0	0.654	0.903
Amp_20q13.2	7046467	0.132	0.201	0	0.436	0.804
Del_11q22.3	13158102	0.032	0.058	0	0.314	0.745
Amp_17q21.31	-2459037	0.180	0.259	0	0.996	0.996
Amp_8p11.23	20778292	0.036	0.062	0	0.957	0.996
Del_3p14.1	2684015	0.729	0.767	0	0.163	0.599
Del_8p11.22	1648757	0.518	0.585	0	0.343	0.745
Amp_11q13.3	13932379	0.423	0.486	0	0.757	0.955
Del_17q12	-8452507	0.353	0.431	0	0.575	0.895
Del_17q21.2	10296570	0.207	0.281	0	0.993	0.996
Amp_19q13.12	10494689	0.341	0.431	0	0.338	0.745
Amp_12q15	-12560166	0.050	0.081	0	0.029	0.233
Del_11p15.4	3708923	0.304	0.403	0	0.275	0.745
Del_5q31.1	8548178	0.388	0.455	0	0.827	0.996
Del_11q11	-166682	0.610	0.674	0	0.531	0.895
Del_1p21.3	18621923	0.044	0.073	0	0.002	**0.028**
Del_11q13.1	10494689	0.061	0.095	0	0.165	0.599
Amp_16q12.2	-7615355	0.367	0.439	0	0.032	0.233
Del_4q35.2	-11531034	0.200	0.277	0	0.157	0.599
Amp_6q25.1	-7615355	0.182	0.259	0	0.360	0.745
Del_10p15.3	-21959310	0.027	0.050	0	0.000	**0.005**
Del_19p13.3	-30105171	0.042	0.072	0	0.904	0.996

Figure S3

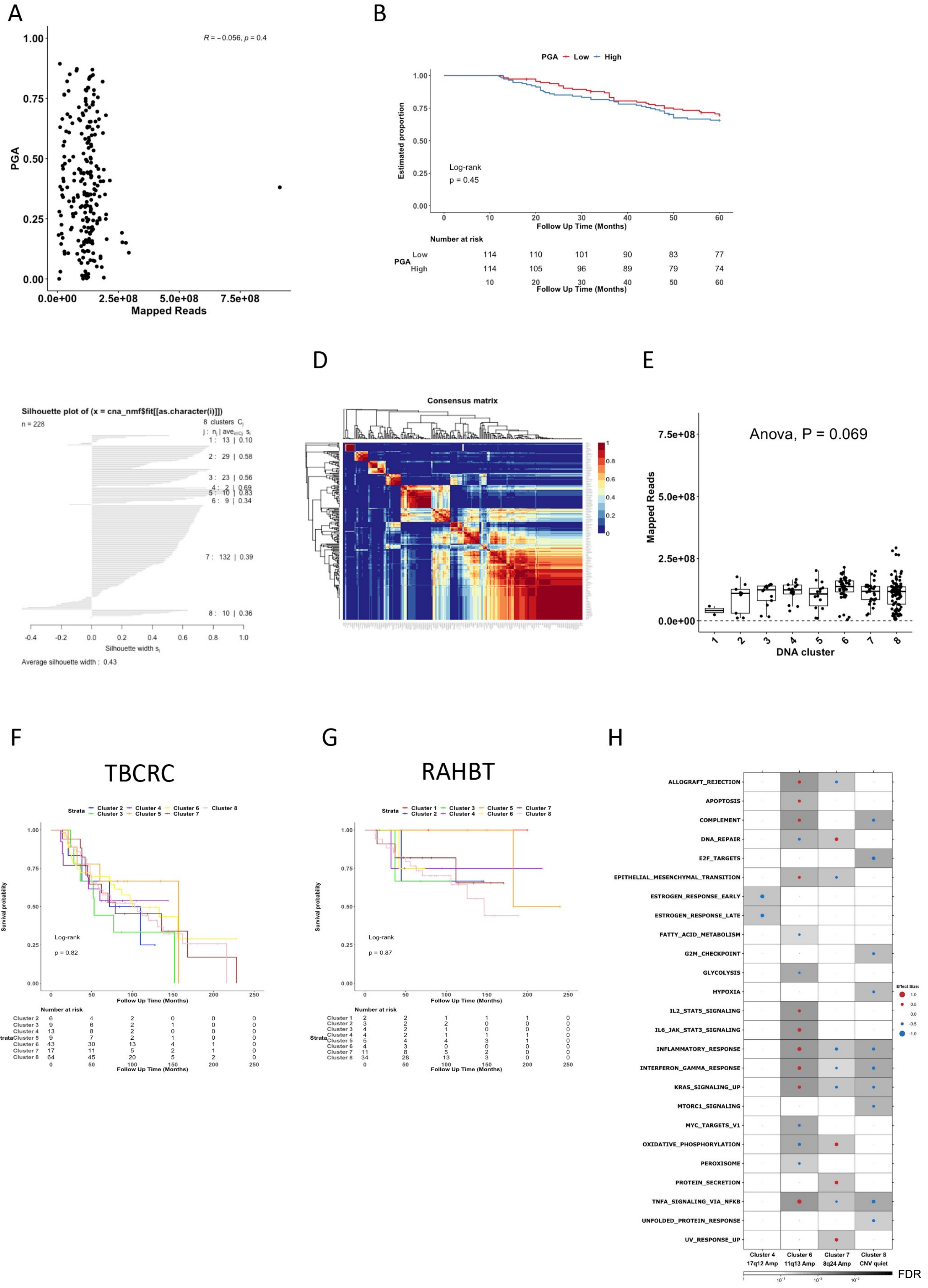

A
R = −0.056, p = 0.4
PGA
1.00
0.75
0.50
0.25
0.00
0.0e+00
2.5e+08
5.0e+08
7.5e+08
Mapped Reads
B
PGA Low High
Estimated proportion
1.00
0.75
0.50
0.25
0.00
Log-rank
p = 0.45
0 10 20 30 40 50 60
Follow Up Time (Months)
Number at risk
PGA
Low 114 110 101 90 83 77
High 114 105 96 89 79 74
10 20 30 40 50 60
Follow Up Time (Months)
C
Silhouette plot of (x = cna_nmf$fit[[as.character(i)]])
n = 228
8 clusters C_j
j : n_j | $ave_{i \in Cj}$ s_i
1 : 13 | 0.10
2 : 29 | 0.58
3 : 23 | 0.56
4 : 2 | 0.69
5 : 10 | 0.83
6 : 9 | 0.34
7 : 132 | 0.39
8 : 10 | 0.36
-0.4 -0.2 0.0 0.2 0.4 0.6 0.8 1.0
Silhouette width s_i
Average silhouette width : 0.43
D
Consensus matrix
1
0.8
0.6
0.4
0.2
0
E
Anova, P = 0.069
Mapped Reads
7.5e+08
5.0e+08
2.5e+08
0.0e+00
1 2 3 4 5 6 7 8
DNA cluster
F
TBCRC
Strata Cluster 2 Cluster 3 Cluster 4 Cluster 5 Cluster 6 Cluster 7 Cluster 8
Survival probability
1.00
0.75
0.50
0.25
0.00
Log-rank
p = 0.82
0 50 100 150 200 250
Follow Up Time (Months)
Number at risk
Strata
Cluster 2 6 4 2 0 0 0
Cluster 3 9 6 2 1 0 0
Cluster 4 13 8 2 0 0 0
Cluster 5 9 7 2 1 0 0
Cluster 6 43 30 13 4 1 0
Cluster 7 17 11 5 2 1 0
Cluster 8 64 45 20 5 2 0
0 50 100 150 200 250
Follow Up Time (Months)
G
RAHBT
Strata Cluster 1 Cluster 2 Cluster 3 Cluster 4 Cluster 5 Cluster 6 Cluster 7 Cluster 8
Survival probability
1.00
0.75
0.50
0.25
0.00
Log-rank
p = 0.87
0 50 100 150 200 250
Follow Up Time (Months)
Number at risk
Strata
Cluster 1 2 2 1 1 1 0
Cluster 2 3 2 2 0 0 0
Cluster 3 4 2 1 0 0 0
Cluster 4 4 2 1 1 1 0
Cluster 5 5 4 4 3 1 0
Cluster 6 4 3 0 0 0 0
Cluster 7 11 8 5 2 0 0
Cluster 8 34 28 13 3 0 0
0 50 100 150 200 250
Follow Up Time (Months)
H
ALLOGRAFT_REJECTION
APOPTOSIS
COMPLEMENT
DNA_REPAIR
E2F_TARGETS
EPITHELIAL_MESENCHYMAL_TRANSITION
ESTROGEN_RESPONSE_EARLY
ESTROGEN_RESPONSE_LATE
FATTY_ACID_METABOLISM
G2M_CHECKPOINT
GLYCOLYSIS
HYPOXIA
IL2_STAT5_SIGNALING
IL6_JAK_STAT3_SIGNALING
INFLAMMATORY_RESPONSE
INTERFERON_GAMMA_RESPONSE
KRAS_SIGNALING_UP
MTORC1_SIGNALING
MYC_TARGETS_V1
OXIDATIVE_PHOSPHORYLATION
PEROXISOME
PROTEIN_SECRETION
TNFA_SIGNALING_VIA_NFKB
UNFOLDED_PROTEIN_RESPONSE
UV_RESPONSE_UP
Cluster 4 17q12 Amp
Cluster 6 11q13 Amp
Cluster 7 8q24 Amp
Cluster 8 CNV quiet
Effect Size:
1.0
0.5
0.0
-0.5
-1.0
FDR
1
10^{-1}
10^{-2}
10^{-3}

Figure S3: Characterizing the CNA landscape of DCIS, related to Figure 5.

A) Correlation between PGA and mapped reads. Correlation coefficient and P-value from Pearson Correlation. **B**) Kaplan-Meier plot of time to progression (any iBE, full follow-up) stratified by PGA (median dichotomized). P-value from log-rank test. **C**) Silhouette plot from NMF unsupervised clustering of the CNA landscape of DCIS in TBCRC and RAHBT combined. **D**) Consensus matrix from NMF unsupervised clustering of the CNA landscape of DCIS in TBCRC and RAHBT combined. **E**) Box plots of mapped reads in the eight DNA clusters. P-value from ANOVA. Boxplot represents median, 0.25 and 0.75 quantiles with whiskers at 1.5x interquartile range. **F-G**) Kaplan-Meier plot of time to progression stratified by the eight CNA clusters in TBCRC (**F**) and RAHBT (**G**). P-values from log-rank tests. **H**) GSEA Hallmark pathway analysis of DE genes by DNA cluster in matched RNA samples (each cluster vs rest) for TBCRC and RAHBT in full. Size of the dot and color represents the magnitude and direction of pathway deregulation, *i.e.,* blue indicates the pathway is downregulated while red indicates the pathway is upregulated. Background shading indicates FDR. Effect size and FDR from GSEA algorithm. Clusters with no significant pathway enrichment or depletion not included in plot.

Figure S4

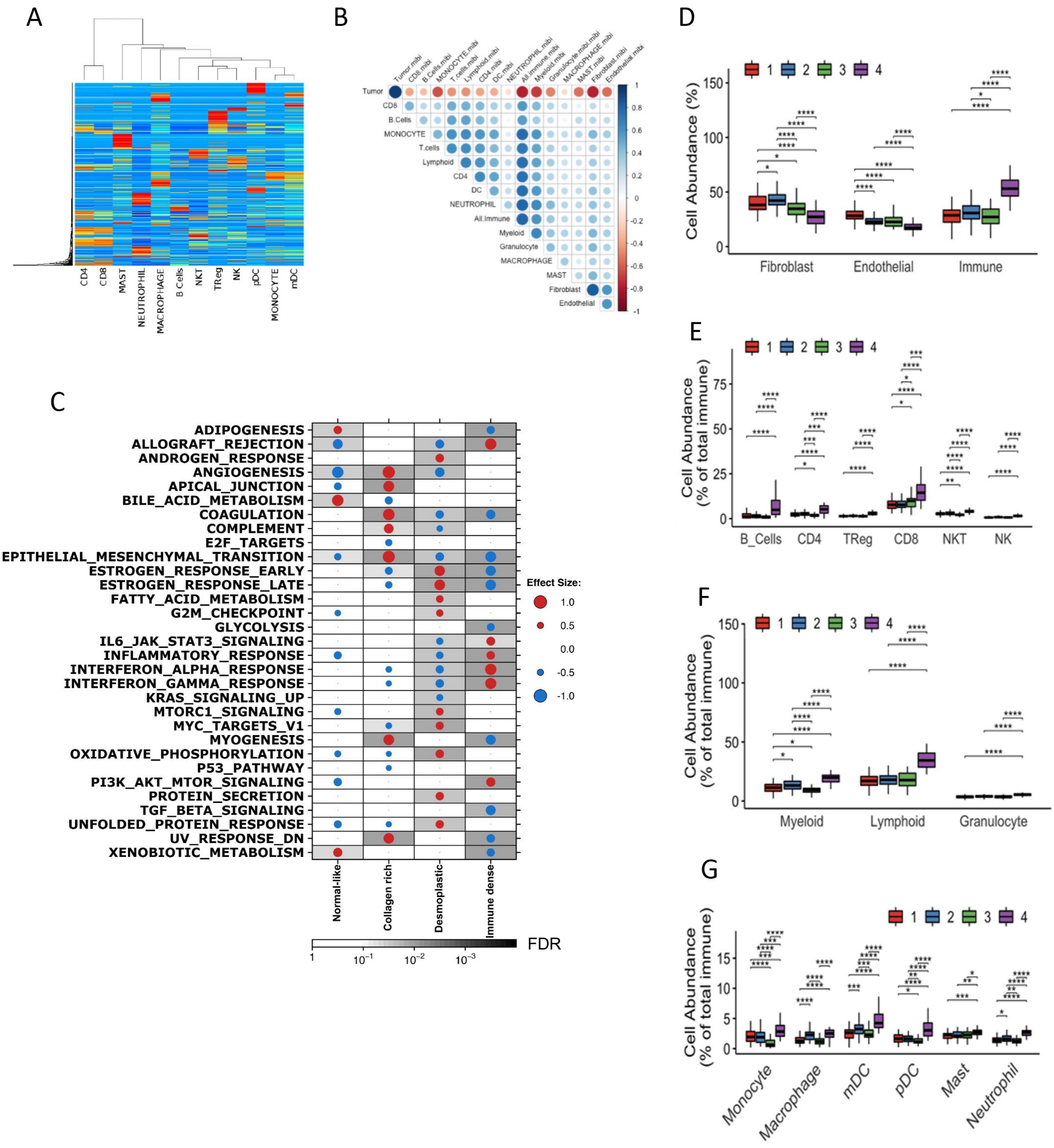

A
CD4
CD8
MAST
NEUTROPHIL
MACROPHAGE
B Cells
NKT
TReg
NK
pDC
MONOCYTE
mDC
B
Tumor.mibi
CD8.mibi
B.Cells.mibi
MONOCYTE.mibi
T.cells.mibi
Lymphoid.mibi
CD4.mibi
DC.mibi
NEUTROPHIL.mibi
All.immune.mibi
Myeloid.mibi
Granulocyte.mibi
MACROPHAGE.mibi
MAST.mibi
Fibroblast.mibi
Endothelial.mibi
Tumor
CD8
B.Cells
MONOCYTE
T.cells
Lymphoid
CD4
DC
NEUTROPHIL
All.Immune
Myeloid
Granulocyte
MACROPHAGE
MAST
Fibroblast
Endothelial
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1
C
ADIPOGENESIS
ALLOGRAFT_REJECTION
ANDROGEN_RESPONSE
ANGIOGENESIS
APICAL_JUNCTION
BILE_ACID_METABOLISM
COAGULATION
COMPLEMENT
E2F_TARGETS
EPITHELIAL_MESENCHYMAL_TRANSITION
ESTROGEN_RESPONSE_EARLY
ESTROGEN_RESPONSE_LATE
FATTY_ACID_METABOLISM
G2M_CHECKPOINT
GLYCOLYSIS
IL6_JAK_STAT3_SIGNALING
INFLAMMATORY_RESPONSE
INTERFERON_ALPHA_RESPONSE
INTERFERON_GAMMA_RESPONSE
KRAS_SIGNALING_UP
MTORC1_SIGNALING
MYC_TARGETS_V1
MYOGENESIS
OXIDATIVE_PHOSPHORYLATION
P53_PATHWAY
PI3K_AKT_MTOR_SIGNALING
PROTEIN_SECRETION
TGF_BETA_SIGNALING
UNFOLDED_PROTEIN_RESPONSE
UV_RESPONSE_DN
XENOBIOTIC_METABOLISM
Normal-like
Collagen rich
Desmoplastic
Immune dense
Effect Size:
1.0
0.5
0.0
-0.5
-1.0
FDR
1
10^{-1}
10^{-2}
10^{-3}
D
1 2 3 4
Cell Abundance (%)
150
100
50
0
Fibroblast
Endothelial
Immune
E
1 2 3 4
Cell Abundance (% of total immune)
75
50
25
0
B_Cells
CD4
TReg
CD8
NKT
NK
F
1 2 3 4
Cell Abundance (% of total immune)
150
100
50
0
Myeloid
Lymphoid
Granulocyte
G
1 2 3 4
Cell Abundance (% of total immune)
15
10
5
0
Monocyte
Macrophage
mDC
pDC
Mast
Neutrophil

Figure S4: Analysis of the tumor microenvironment, related to Figure 6.

A) Heatmap showing signature matrix created using CibersortX (CSx) with 12 different immune cell types. **B**) Protein validation of CSx signature matrix by MIBI. Correlogram showing MIBI-based vs CSx-estimated cell types in RAHBT LCM samples. Correlation and statistics from Pearson's correlation. White background: P>0.05. **C**) GSEA Hallmark pathway analysis of DE genes in each stromal cluster vs the rest in RAHBT LCM stromal samples in full. Size of the dot and color represents the magnitude and direction of pathway deregulation, *i.e.,* blue indicates the pathway is downregulated while red indicates the pathway is upregulated. Background shading indicates FDR. Effect size and FDR from GSEA algorithm. **D**) Percentage of fibroblasts, endothelial and total immune cells present in each stromal cluster estimated by CSx. **E**) Abundance of total myeloid, lymphoid and granulocyte cells, represented as percentage of total immune cells. **F-G**) Abundance of 12 immune cell types (percentage of total immune cells) by stromal clusters. Box plots (**D-G**): Red: Normal-like. Blue: Collage rich. Green: Desmoplastic. Purple: Immune dense. *: FDR <0.05; **: FDR < 0.01; ***: FDR<0.001; ****: FDR < 0.0001; ns: FDR >0.05 (Wilcoxon rank sum test). Boxplots represent median, 0.25 and 0.75 quantiles with whiskers at 1.5x interquartile range.

Figure S5

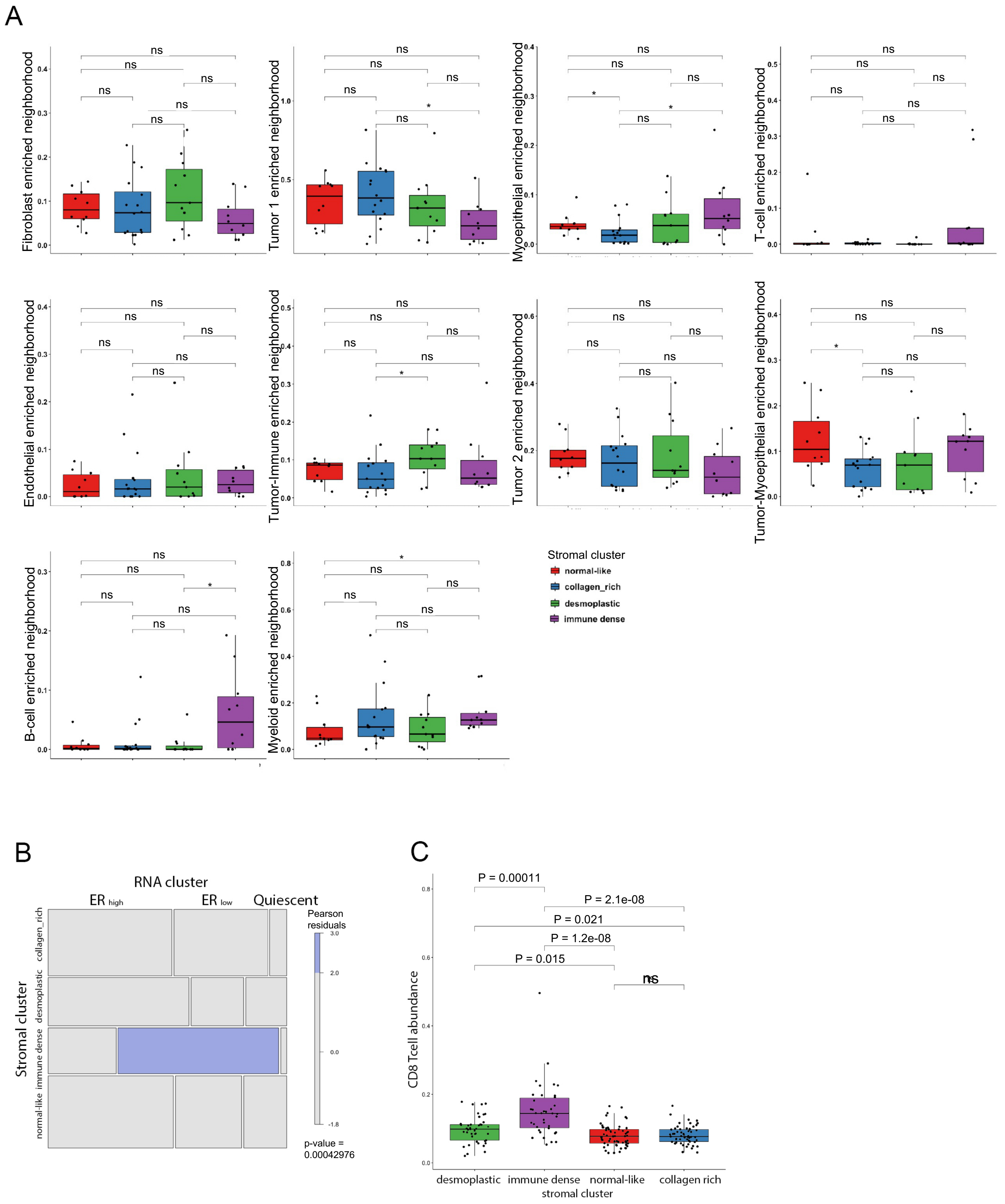

A
Fibroblast enriched neighborhood
Tumor 1 enriched neighborhood
Myoepithelial enriched neighborhood
T-cell enriched neighborhood
Endothelial enriched neighborhood
Tumor-Immune enriched neighborhood
Tumor 2 enriched neighborhood
Tumor-Myoepithelial enriched neighborhood
B-cell enriched neighborhood
Myeloid enriched neighborhood
ns
*
Stromal cluster
🟥 normal-like
🟦 collagen_rich
🟩 desmoplastic
🟪 immune dense
B
RNA cluster
ER high
ER low
Quiescent
Stromal cluster
collagen_rich
desmoplastic
immune dense
normal-like
Pearson residuals
3.0
2.0
0.0
-1.8
p-value = 0.00042976
C
P = 0.00011
P = 2.1e-08
P = 0.021
P = 1.2e-08
P = 0.015
ns
CD8 Tcell abundance
0.0
0.2
0.4
0.6
0.8
desmoplastic
immune dense
normal-like
collagen rich
stromal cluster

Figure S5: Analysis of the four stromal clusters using CSx and MIBI, related to Figure 6.

A) Cell neighborhood frequencies from MIBI by stromal clusters. *: FDR <0.05. ns: FDR >0.05 (Wilcoxon rank sum test). **B**) Mosaic plot showing correlation between the stromal clusters and the RNA 3-cluster subtypes in matched epithelial samples. **C**) Box plots showing the CD8 T-cell abundance by CSx in the four stromal clusters P-values from Wilcoxon rank sum test. **A, C**): Boxplots represent median, 0.25 and 0.75 quantiles with whiskers at 1.5x interquartile range.

Figure S6

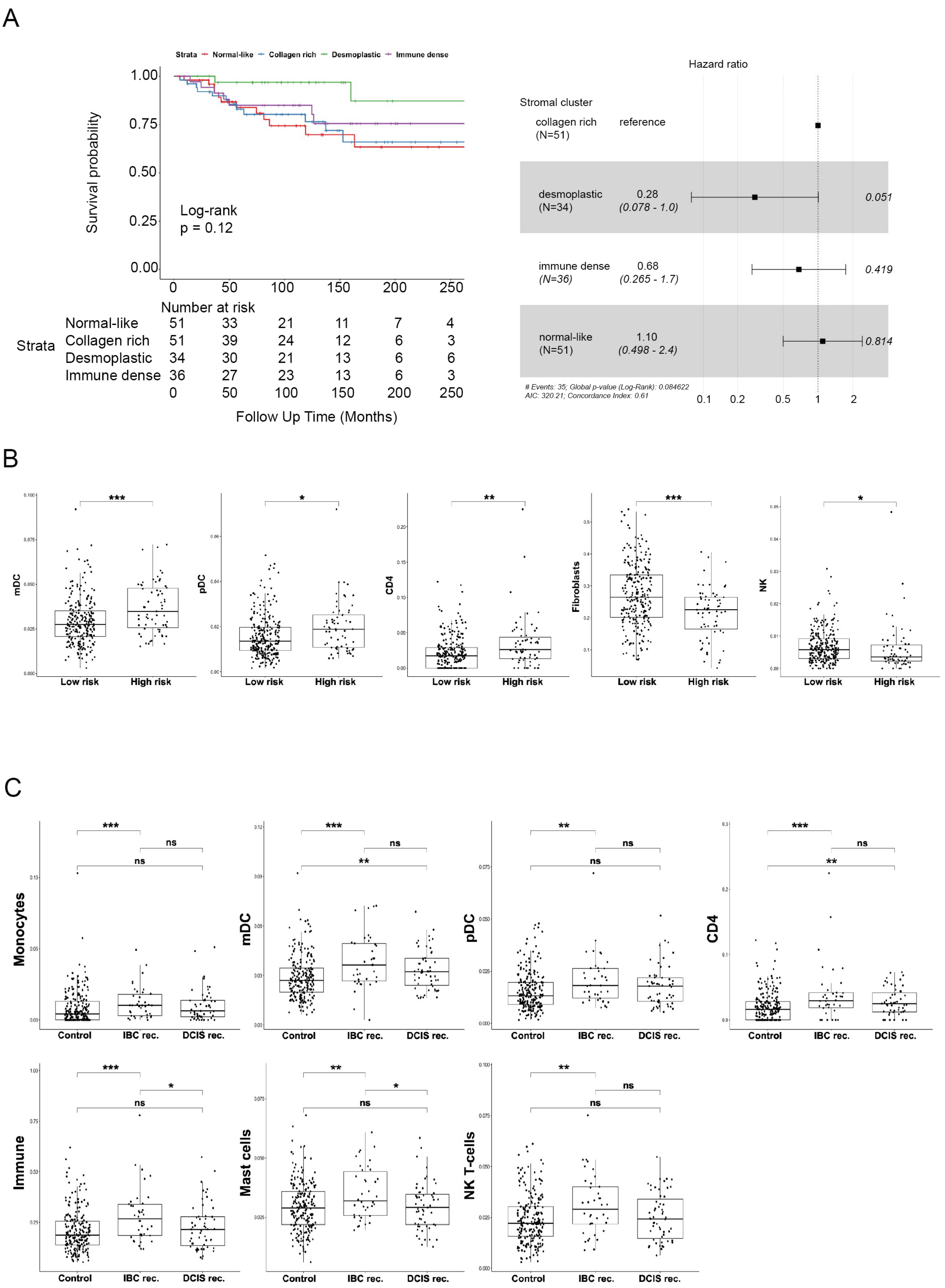

A
Strata — Normal-like — Collagen rich — Desmoplastic — Immune dense
Survival probability
1.00
0.75
0.50
0.25
0.00
Log-rank
p = 0.12
0
50
100
150
200
250
Number at risk
Strata
Normal-like 51 33 21 11 7 4
Collagen rich 51 39 24 12 6 3
Desmoplastic 34 30 21 13 6 6
Immune dense 36 27 23 13 6 3
0 50 100 150 200 250
Follow Up Time (Months)
Hazard ratio
Stromal cluster
collagen rich (N=51) reference
desmoplastic (N=34) 0.28 (0.078 - 1.0) 0.051
immune dense (N=36) 0.68 (0.265 - 1.7) 0.419
normal-like (N=51) 1.10 (0.498 - 2.4) 0.814
# Events: 35; Global p-value (Log-Rank): 0.084622
AIC: 320.21; Concordance Index: 0.61
0.1
0.2
0.5
1
2
B
mDC
0.100
0.075
0.050
0.025
0.000
***
Low risk
High risk
pDC
0.06
0.04
0.02
0.00
*
Low risk
High risk
CD4
0.20
0.15
0.10
0.05
0.00
**
Low risk
High risk
Fibroblasts
0.5
0.4
0.3
0.2
0.1
***
Low risk
High risk
NK
0.05
0.04
0.03
0.02
0.01
0.00
*
Low risk
High risk
C
Monocytes
0.10
0.05
0.00
***
ns
ns
Control
IBC rec.
DCIS rec.
mDC
0.12
0.09
0.06
0.03
0.00
***
ns
**
Control
IBC rec.
DCIS rec.
pDC
0.075
0.050
0.025
0.000
**
ns
ns
Control
IBC rec.
DCIS rec.
CD4
0.3
0.2
0.1
0.0
***
ns
**
Control
IBC rec.
DCIS rec.
Immune
1.00
0.75
0.50
0.25
***
*
ns
Control
IBC rec.
DCIS rec.
Mast cells
0.075
0.050
0.025
**
*
ns
Control
IBC rec.
DCIS rec.
NK T-cells
0.100
0.075
0.050
0.025
0.000
**
ns
ns
Control
IBC rec.
DCIS rec.

Figure S6: Outcome associations in the four stromal clusters, related to Figure 6.

A) Kaplan-Meier and forest plot of time to progression (any iBE, full follow-up time) stratified by stromal clusters in RAHBT LCM. Kaplan-Meier P-value from log-rank test. Forest plot P-values and hazard ratios from Cox multivariable analysis. **B**) CSx-inferred cell type distribution between 812 gene classifier risk groups (TBCRC and RAHBT combined). Only cell types with FDR<0.05 are shown. **C**) CSx-inferred cell type distribution between cases with IBC iBEs, DCIS iBEs, and controls (TBCRC and RAHBT combined). Only cell types with FDR<0.05 are shown. **B, C**) * FDR < 0.05. ** FDR ≤ 0.1. *** FDR ≤ 0.001; ns: FDR >0.05 (Wilcoxon rank sum test). mDC: myeloid dendritic cells. pDC: plasmacytoid dendritic cells. NK: Natural killer cells. Boxplots represent median, 0.25 and 0.75 quantiles with whiskers at 1.5x interquartile range.

Table S5: Univariate Cox regression analysis of CSx cell type abundance towards progression (any iBE) in RAHBT LCM. Related to Figure 6.

	coef	**Exp (coef)**	**Se (coef)**	**z**	**P-value**	**FDR**
mDC	25.36863	1.041E+11	6.383095	3.974347	7.06E-05	**0.00101071**
CD4	10.88261	53242.31	2.838826	3.833489	0.00012634	**0.00101071**
pDC	28.35189	2.06E+12	8.856115	3.201391	0.00136766	**0.00554251**
Immune	2.605316	13.5355	0.8147652	3.197628	0.00138563	**0.00554251**
NKT	19.62875	334702207	8.055231	2.436771	0.01481907	**0.04595117**
Fibroblast	-2.624797	0.07245444	1.102057	-2.381727	0.01723169	**0.04595117**
Macrophage	15.06461	3487218	6.808633	2.212576	0.02692691	0.06154722
Mast	19.44246	277813316	9.526299	2.040925	0.0412583	0.0825166
Monocyte	11.71033	121823.5	7.007859	1.671028	0.09471613	0.16838423
Endothelial	-3.857939	0.02111147	2.48589	-1.551935	0.1206778	0.19308448
CD8	5.864071	352.1549	4.018355	1.459321	0.1444767	0.21014793
Neutrophil	20.47718	781856966	14.86406	1.377631	0.1683173	0.22442307
T Reg	11.2979	80651.73	10.19587	1.108086	0.2678248	0.32963052
B Cells	7.016958	1115.388	7.655518	0.9165882	0.3593585	0.41069543
NK	7.482623	1776.895	18.85228	0.396908	0.6914353	0.73753099
Epithelial	-0.1520237	0.8589679	0.6357247	-0.2391345	0.8110013	0.8110013

mDC: myeloid dendritic cells. pDC: plasmacytoid dendritic cells. NKT: Natural killer T cells.
NK: Natural killer cells.