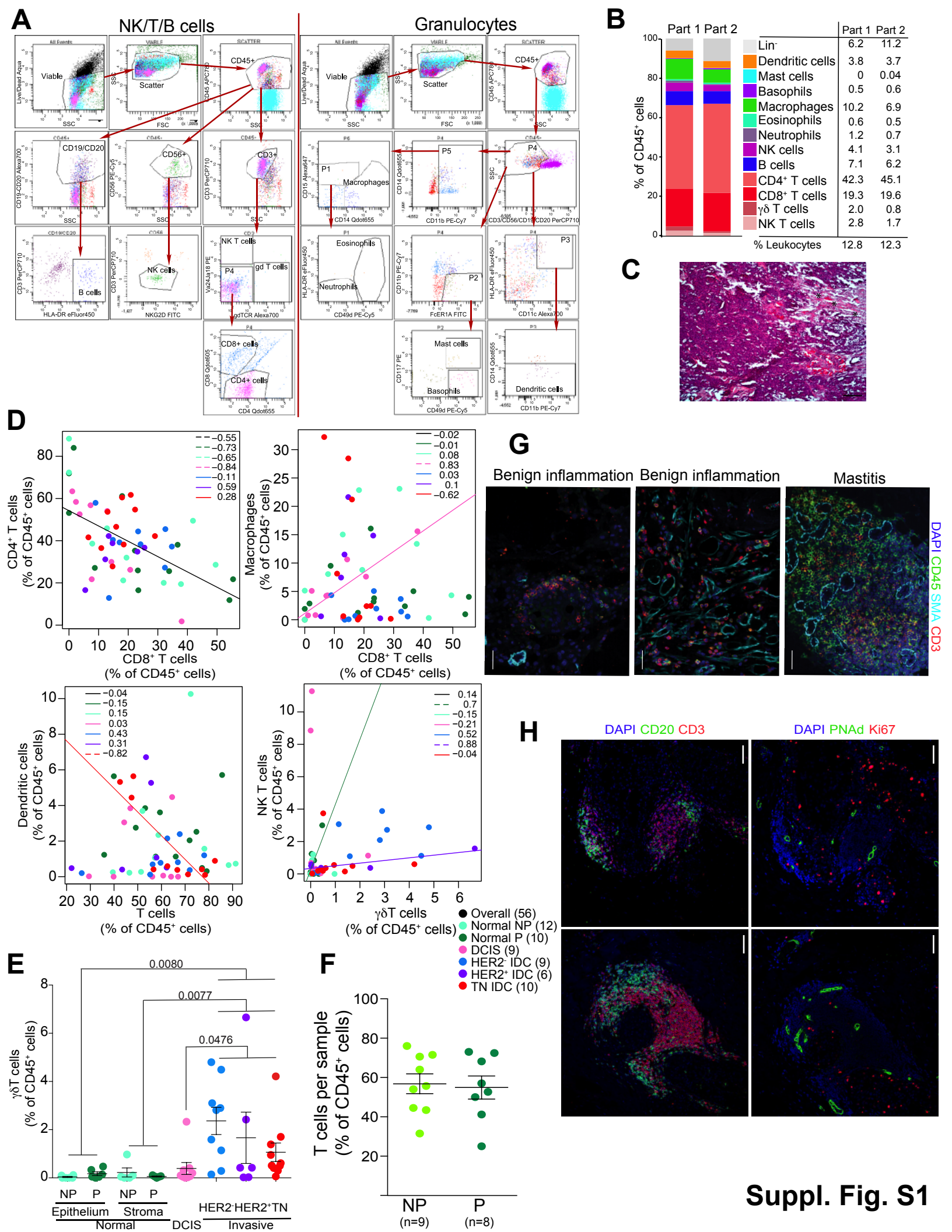
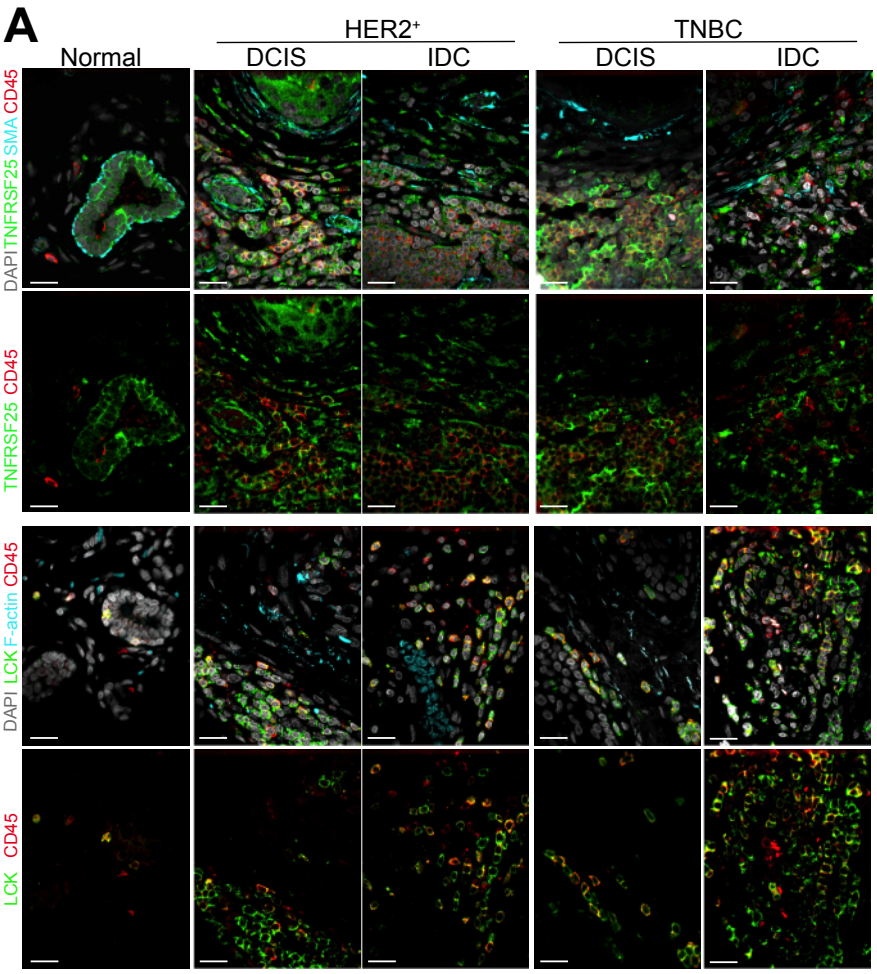
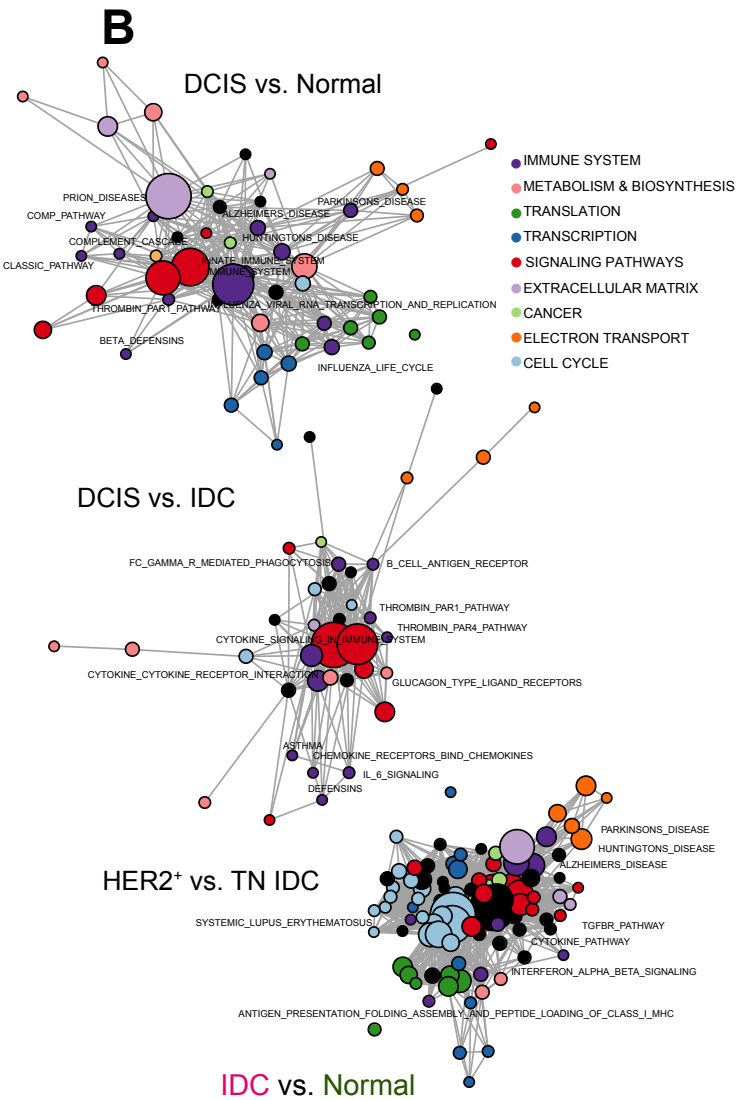
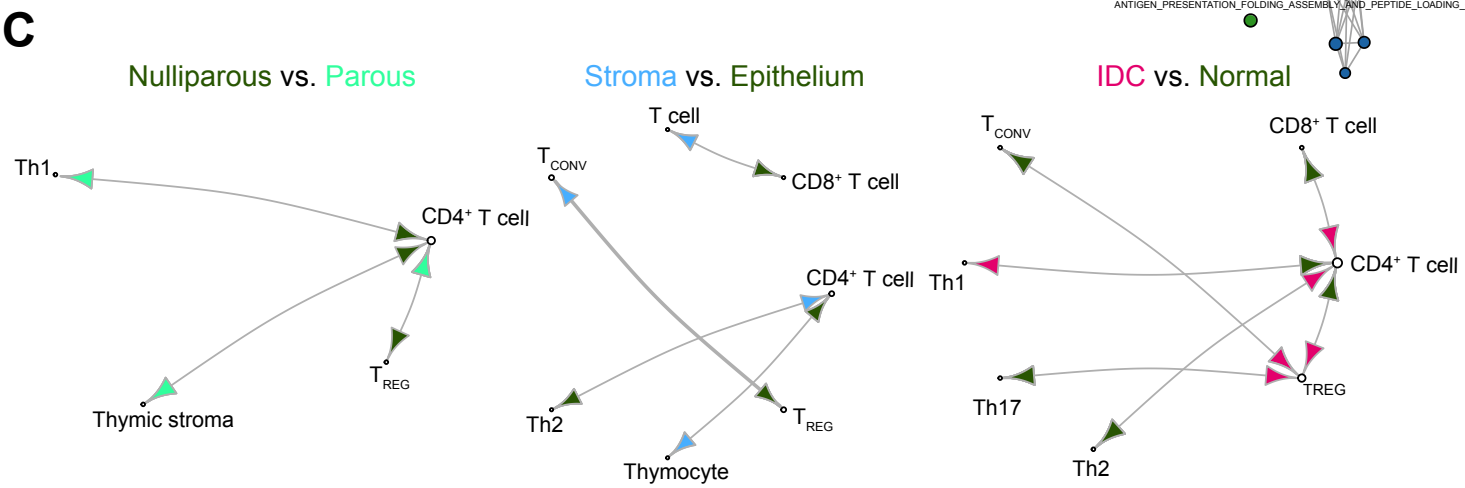

A
NK/T/B cells
Granulocytes
Viable
Scatter
CD45+
CD19/CD20
CD56+
CD3+
B cells
NK cells
NK T cells
gd T cells
CD8+ cells
CD4+ cells
Macrophages
Eosinophils
Neutrophils
Mast cells
Basophils
Dendritic cells
B
Part 1 Part 2
% of CD45+ cells
Lin-
Dendritic cells
Mast cells
Basophils
Macrophages
Eosinophils
Neutrophils
NK cells
B cells
CD4+ T cells
CD8+ T cells
γδ T cells
NK T cells
% Leukocytes
C
D
CD4+ T cells (% of CD45+ cells)
CD8+ T cells (% of CD45+ cells)
Macrophages (% of CD45+ cells)
Dendritic cells (% of CD45+ cells)
T cells (% of CD45+ cells)
NK T cells (% of CD45+ cells)
γδT cells (% of CD45+ cells)
Overall (56)
Normal NP (12)
Normal P (10)
DCIS (9)
HER2- IDC (9)
HER2+ IDC (6)
TN IDC (10)
E
γδT cells (% of CD45+ cells)
0.0080
0.0077
0.0476
NP
P
Epithelium
Stroma
Normal
DCIS
HER2-
HER2+
TN
Invasive
F
T cells per sample (% of CD45+ cells)
NP (n=9)
P (n=8)
G
Benign inflammation
Benign inflammation
Mastitis
DAPI CD45 SMA CD3
H
DAPI CD20 CD3
DAPI PNAd Ki67

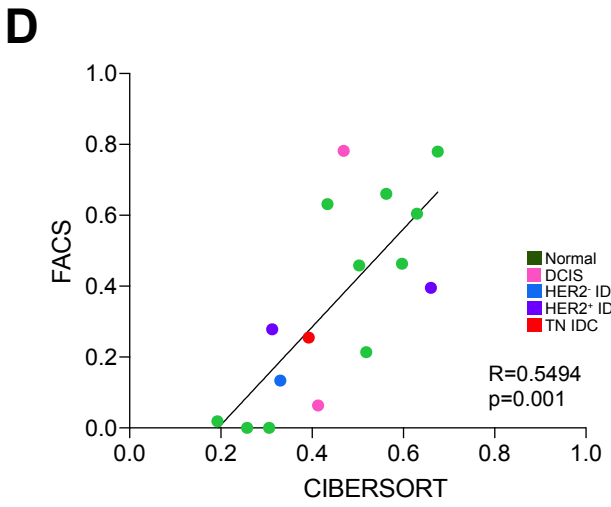
Suppl. Fig. S1

Supplementary Figure S1. Leukocyte frequencies in breast tissues. **A**, Gating strategy for the identification of lymphocyte (left) and myeloid-lineage (right) populations. Arrows indicate directionality of subgates. Identified populations are marked in bold. **Supplementary Table S8** lists all antibodies used for FACS. **B**, Summary of polychromatic FACS analyses of leukocyte populations in two independent parts of the same tumor sample. Leukocyte abundance is shown as percentage of total viable cells, whereas the relative fractions of specific cell populations are shown as a percent of total $CD45^+$ cells. **C**, Hematoxylin-eosin staining of the tumor sample analyzed in panel B. Scale bar 100 μm. Stars mark leukocytes. **D**, Correlation plots to depict associations between the frequencies of $CD8^+$ T cells and $CD4^+$ T cells, T cells and dendritic cells, $CD8^+$ T cells and macrophages, and γδ T cells and NK T cells. Combined epithelium and stroma data are shown in light and dark green for nulliparous (Normal NP) and parous (Normal P) samples, respectively. All other colors reflect tissue types as indicated. Correlation coefficients, r, are shown for each individual group as well as for the combined data (overall). Trend lines are indicated only for significant (<0.05) correlations, are corresponding r are marked with a dotted line, n – number of samples in group. **E**, Relative fraction of γδT cells quantified based on FACS. Results are shown as percent of total $CD45^+$ cells. **F**, Frequency of $CD3^+$ T cells in normal breast tissue samples quantified based on immunofluorescence images. Results are shown as percent of total $CD45^+$ cells. NP- nulliparous and P - parous, respectively. **G**, Immunofluorescence analysis of $CD3^+$ T cells in mastitis and benign inflammation samples. **H**, Examples of potential tertiary lymphoid structures in DCIS. Scale bar, 150 μm.


A
Normal
HER2+
DCIS
IDC
TNBC
DCIS
IDC
DAPI TNFRSF25 SMA CD45
TNFRSF25 CD45
DAPI LCK F-actin CD45
LCK CD45

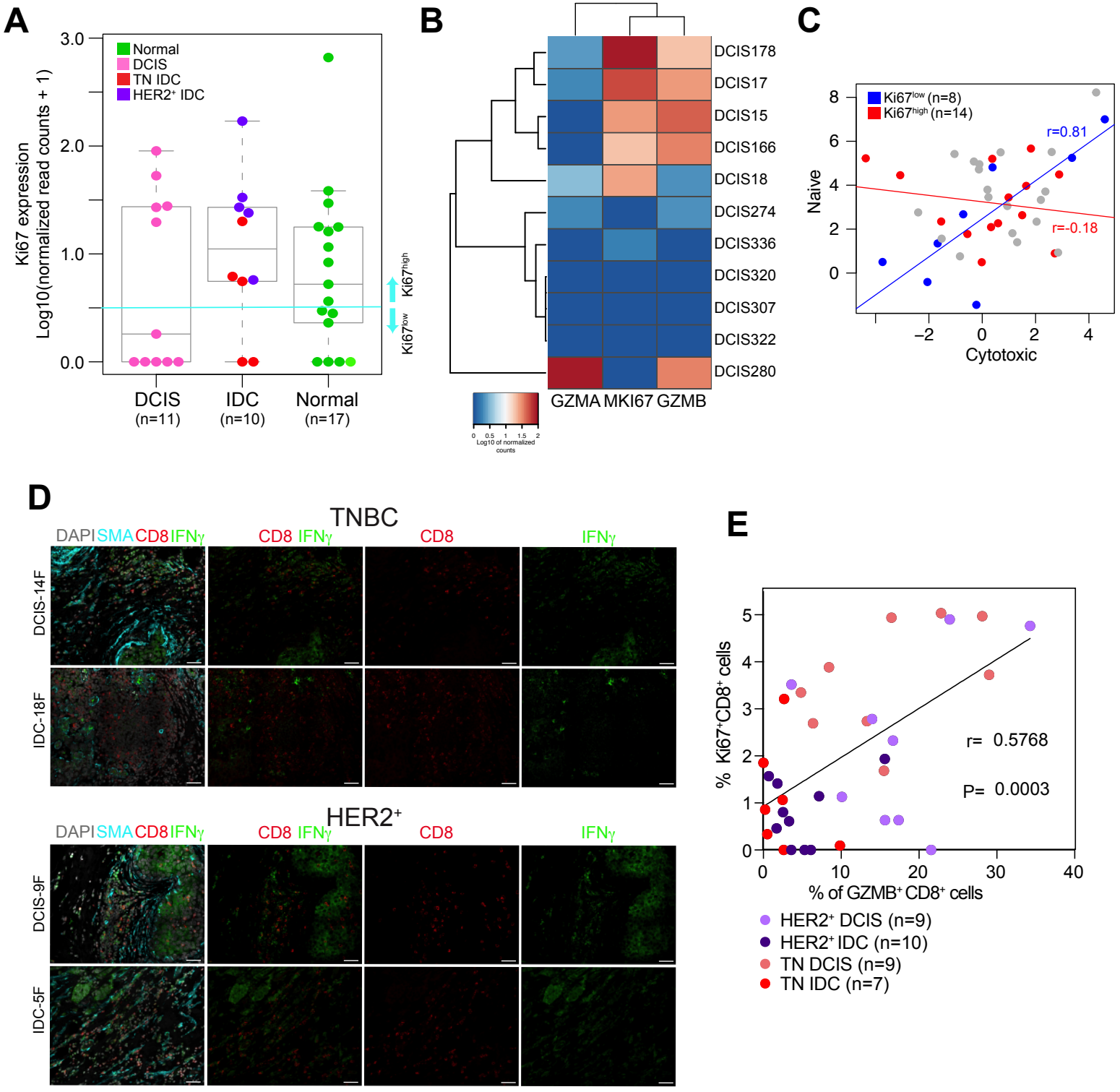

B
DCIS vs. Normal
IMMUNE SYSTEM
METABOLISM & BIOSYNTHESIS
TRANSLATION
TRANSCRIPTION
SIGNALING PATHWAYS
EXTRACELLULAR MATRIX
CANCER
ELECTRON TRANSPORT
CELL CYCLE
PRION_DISEASES
COMP_PATHWAY
COMPLEMENT_CASCADE
CLASSIC_PATHWAY
ALZHEIMERS_DISEASE
PARKINSONS_DISEASE
HUNTINGTONS_DISEASE
INNATE_IMMUNE_SYSTEM
ADAPTIVE_IMMUNE_SYSTEM
THROMBIN_PART_PATHWAY
INFLUENZA_VIRAL_RNA_TRANSCRIPTION_AND_REPLICATION
BETA_DEFENSINS
INFLUENZA_LIFE_CYCLE
DCIS vs. IDC
FC_GAMMA_R_MEDIATED_PHAGOCYTOSIS
B_CELL_ANTIGEN_RECEPTOR
THROMBIN_PAR1_PATHWAY
THROMBIN_PAR4_PATHWAY
CYTOKINE_SIGNALING_IN_IMMUNE_SYSTEM
CYTOKINE_CYTOKINE_RECEPTOR_INTERACTION
GLUCAGON_TYPE_LIGAND_RECEPTORS
ASTHMA
CHEMOKINE_RECEPTORS_BIND_CHEMOKINES
IL_6_SIGNALING
DEFENSINS
HER2+ vs. TN IDC
PARKINSONS_DISEASE
HUNTINGTONS_DISEASE
ALZHEIMERS_DISEASE
SYSTEMIC_LUPUS_ERYTHEMATOSUS
TGFBR_PATHWAY
CYTOKINE_PATHWAY
INTERFERON_ALPHA_BETA_SIGNALING
ANTIGEN_PRESENTATION_FOLDING_ASSEMBLY_AND_PEPTIDE_LOADING_OF_CLASS_I_MHC


C
Nulliparous vs. Parous
Th1
CD4+ T cell
TREG
Thymic stroma
Stroma vs. Epithelium
T cell
TCONV
CD8+ T cell
CD4+ T cell
Th2
TREG
Thymocyte
IDC vs. Normal
TCONV
CD8+ T cell
CD4+ T cell
Th1
TREG
Th17
Th2


D
FACS
CIBERSORT
1.0
0.8
0.6
0.4
0.2
0.0
Normal
DCIS
HER2- IDC
HER2+ IDC
TN IDC
R=0.5494
p=0.001

Suppl. Fig. S2

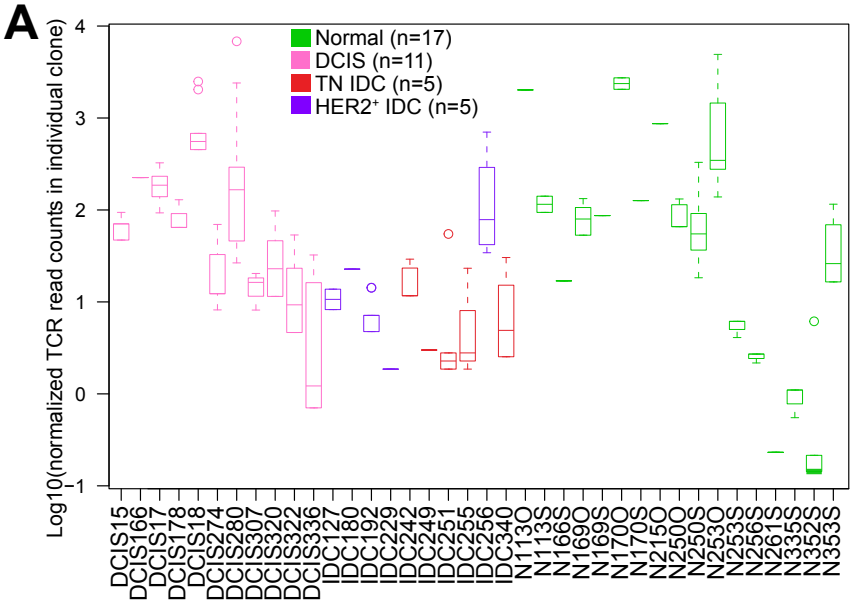
Supplementary Figure S2. Genes expressed in T cells. **A**, Immunofluorescence analysis of LCK and TNFRSF25 expression. Scale bar 50 μm. **B**, Network of significantly enriched canonical gene sets in the indicated comparisons. Node size is reflective of the number of genes in the set. **C**, Network diagrams of enriched cell type-specific gene sets from using ImmuneSigDB. **D**, Plot depicting correlation between frequency of $CD8^+$ T cells assessed by FACS or CIBERSORT analysis of $CD3^+$ T cell RNA-seq data.

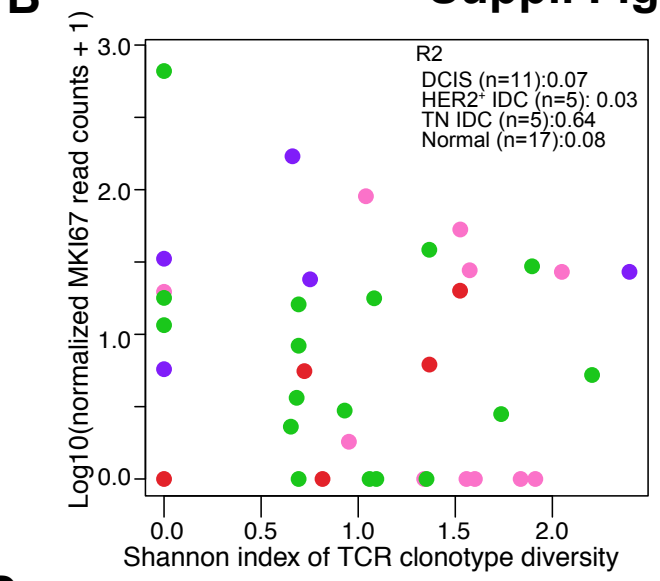

A
Normal
DCIS
TN IDC
HER2+ IDC
Ki67 expression
Log10(normalized read counts + 1)
3.0
2.0
1.0
0.0
Ki67high
Ki67low
DCIS
(n=11)
IDC
(n=10)
Normal
(n=17)
B
DCIS178
DCIS17
DCIS15
DCIS166
DCIS18
DCIS274
DCIS336
DCIS320
DCIS307
DCIS322
DCIS280
GZMA MKI67 GZMB
Log10 of normalized counts
C
Ki67low (n=8)
Ki67high (n=14)
r=0.81
r=-0.18
Naive
Cytotoxic
D
TNBC
DAPI SMA CD8 IFNγ
CD8 IFNγ
CD8
IFNγ
DCIS-14F
IDC-18F
HER2+
DCIS-9F
IDC-5F
E
% Ki67+CD8+ cells
% of GZMB+ CD8+ cells
r= 0.5768
P= 0.0003
HER2+ DCIS (n=9)
HER2+ IDC (n=10)
TN DCIS (n=9)
TN IDC (n=7)

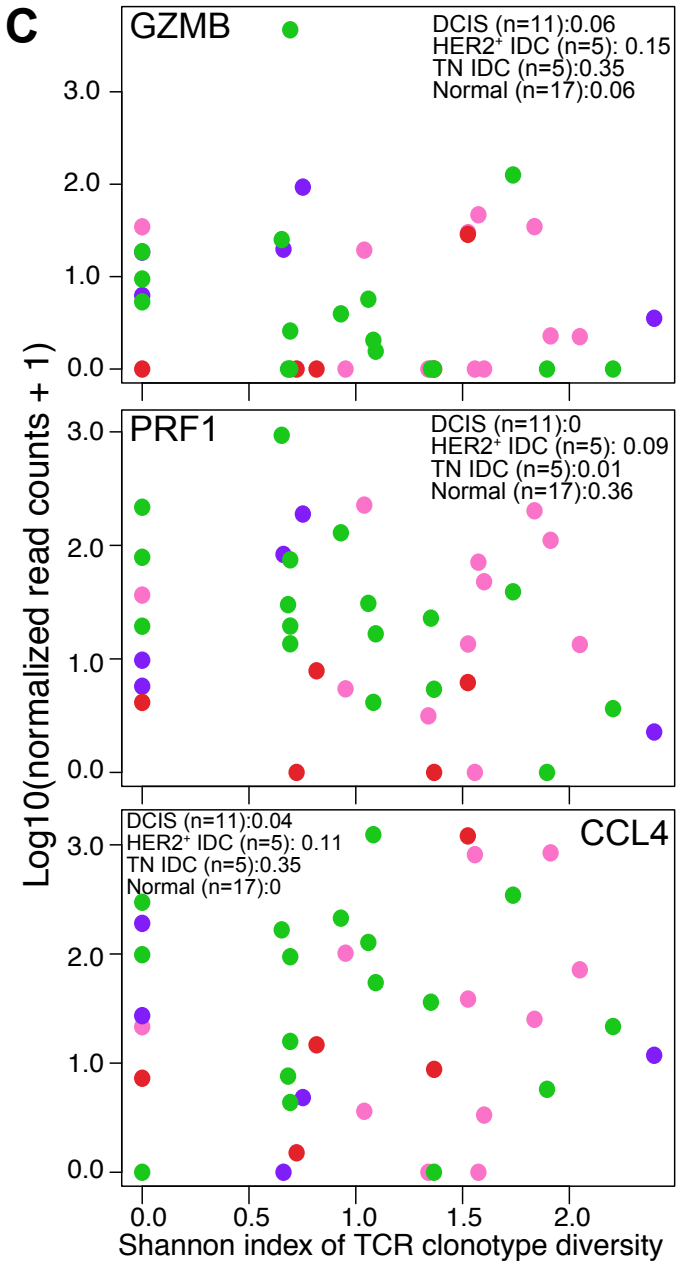
Suppl. Fig. S3

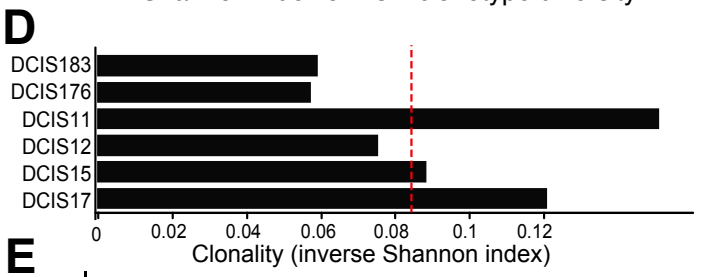
Supplementary Figure S3. Activity status of T cells. **A**, MKI67 expression in T cells based on RNA-seq data. Low and high $Ki67^{low}$ and $Ki67^{high}$ samples were defined by a cut-off of 0.5. **B**, Heatmap depicting the clustering of DCIS samples based on mRNA levels of MKI67, GZMA, and GZMB. **C**, Plot depicting correlation between naïve and cytotoxic T cell gene expression signatures in $MKI67^{high}$ (red) and $MKI67^{low}$ (blue) T cells. **D**, Immunofluorescence analysis of $IFN\gamma^{+}$ $CD8^{+}$ T cells in TNBC and $HER2^{+}$ breast tumors. Images are a montage of nine fields captured from one area of the tissue. Scale bar 50 μm. **E**, Plot depicting correlation between proportion of $GZMB^{+}$ and $Ki67^{+}$ $CD8^{+}$ cells in $HER2^{+}$ DCIS and IDC, and TN DCIS and IDC based on immunofluorescence analysis.

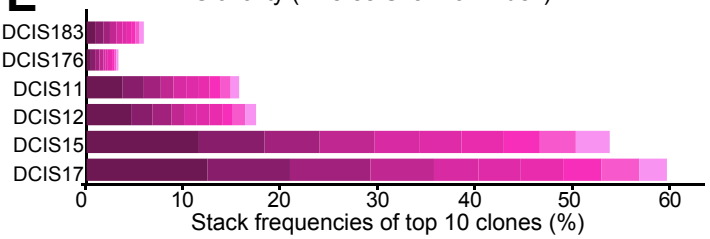
Suppl. Fig. S4

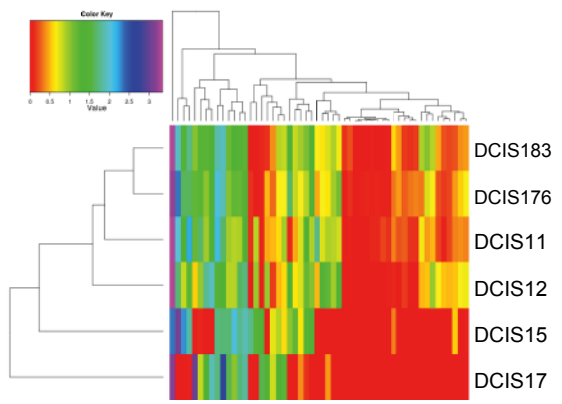


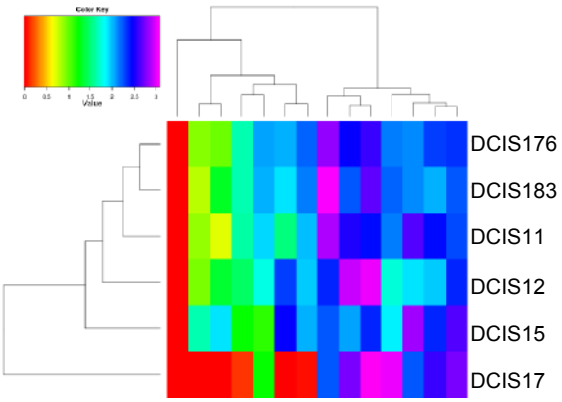










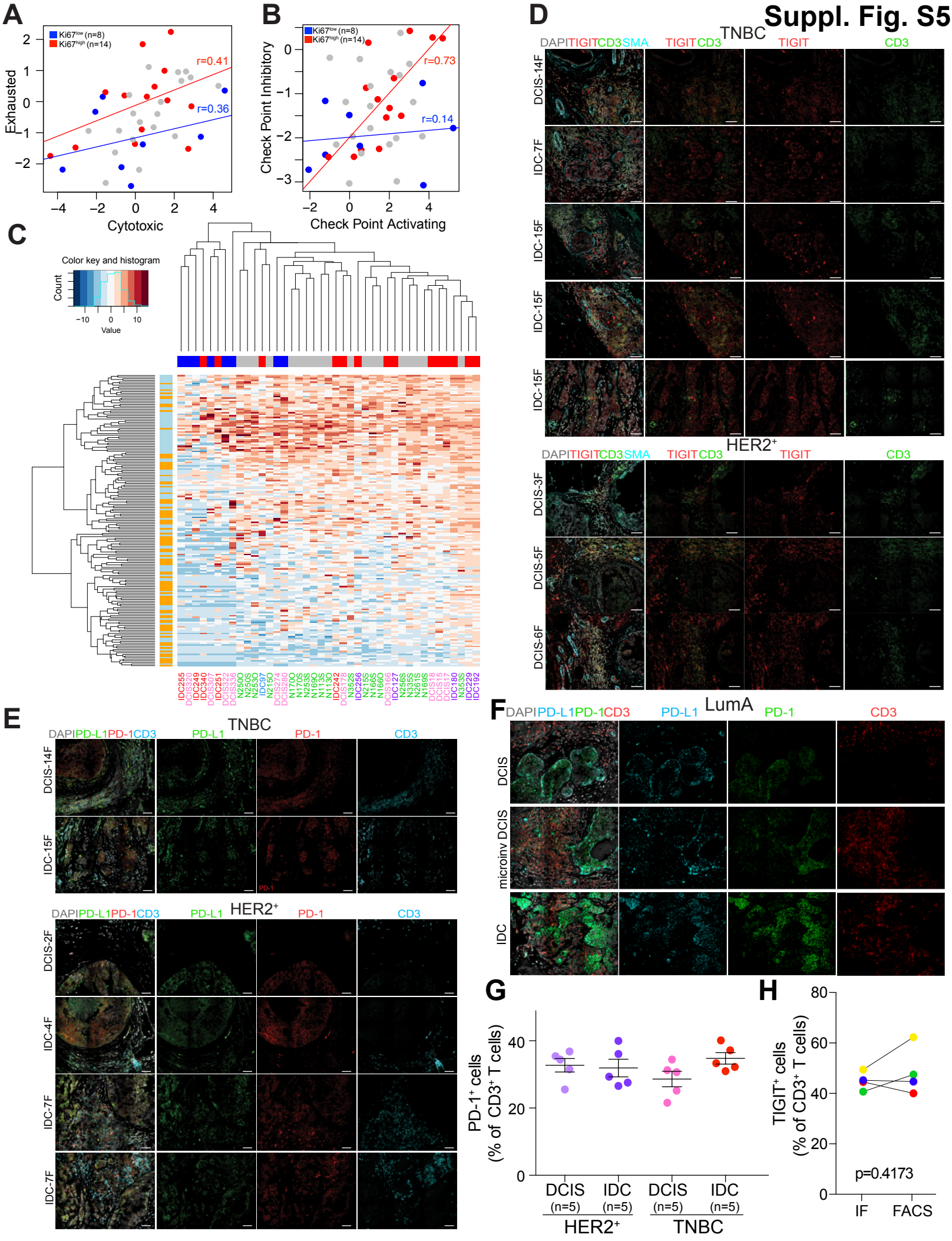


H

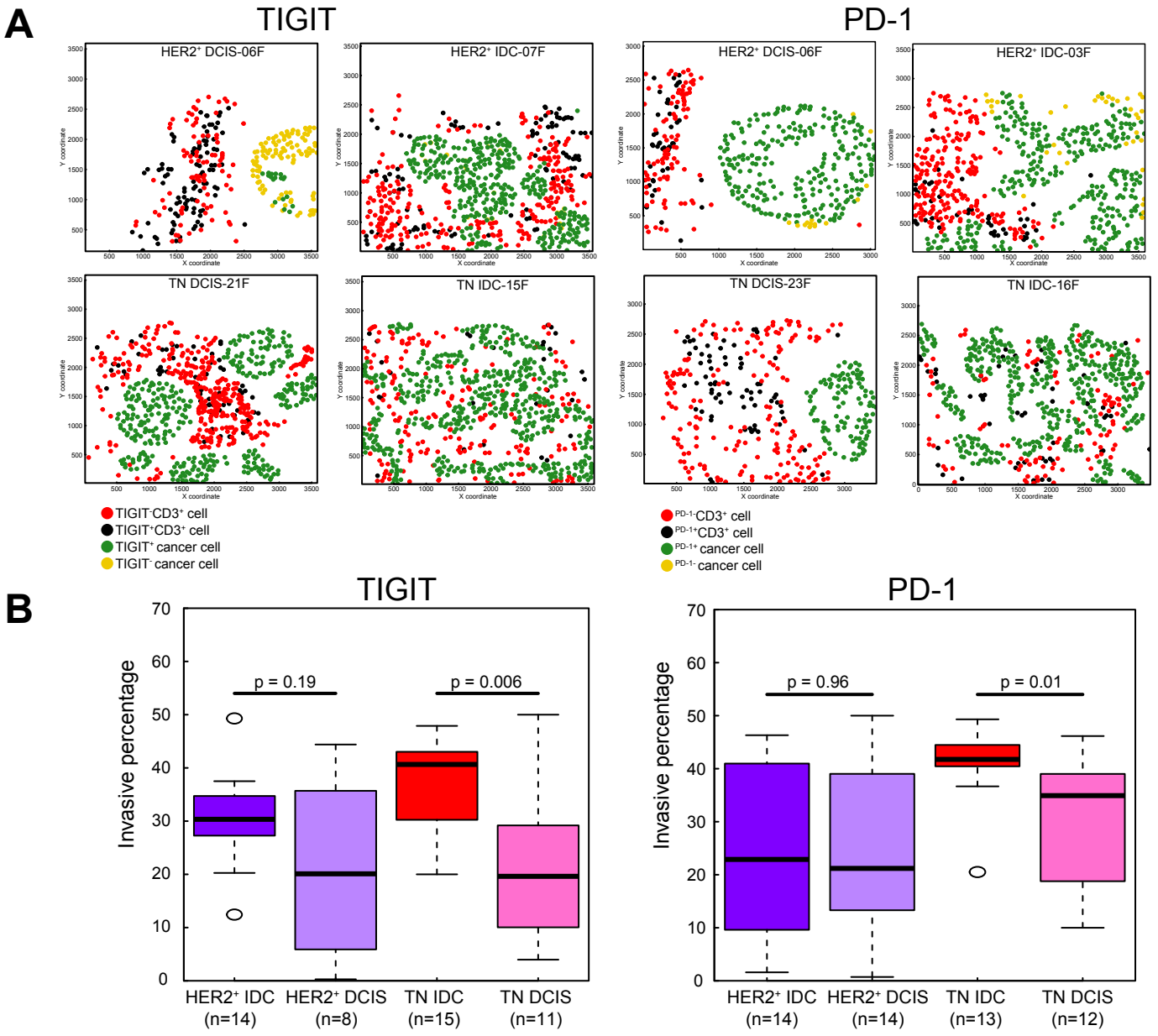
AA sequence	Clone frequency (%)						Epitope	Epitope sequence	HLA restriction	Epitope ID	Protein
	DCIS11	DCIS12	DCIS15	DCIS17	DCIS176	DCIS183					
CASSLGQAYEQYF	0.000	0.000	0.000	0.000	0.100	0.044	EBV-FL9	FLRGRAYGL	HLA-B*08	16878	EBNA-3
CSALIGADPYGYTF	0.000	0.000	0.000	2.770	0.039	0.020	UK	UK	UK	UK	UK
CASSYWGLAGDTQYF	0.000	0.000	0.000	0.112	0.002	0.001	UK	UK	UK	UK	UK
CASSESGYGYTF	0.042	0.000	4.454	0.000	0.040	0.000	UK	UK	UK	UK	UK

Supplementary Figure S4. TCR clonality in breast tissues. **A**, TCR clonotypes in each sample. **B**, Correlation plot between MKI67 expression (y-axis) and Shannon index of TCR clonotype diversity. R2 values and number of samples in each group are indicated. **C**, Correlation plot between the expression of the indicated T cell activation markers (y-axis) and Shannon index of TCR clonotype diversity. R2 values and number of samples in each group are indicated. **D**, Measurement of T cell clonality (inverse of normalized Shannon index) in each DCIS sample. Clonality varies from 0-1 with 0 being a flat distribution and 1 being an entirely oligoclonal sample. Red dash line represents median of clonality. **E**, The frequencies of the top 10 clones in each sample. **F-G**, Heatmap of normalized frequencies of TCRv (**F**) and j (**G**) genes in each sample. **H**, Four common TCR clones in the top 100 clones in each sample based on amino acid sequences. Numbers are percentage of total reads. Clone frequency in each DCIS, predicted epitope, epitope sequence, HLA restriction, epitope ID, and protein are indicated. UK – unknown. Shading intensity correlates with relative frequency of the clone.

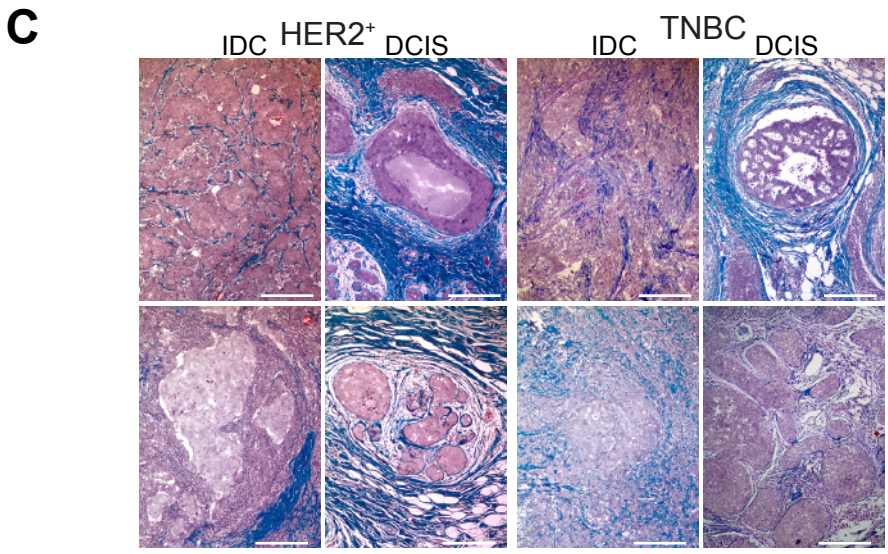
Suppl. Fig. S5


A
Ki67low (n=8)
Ki67high (n=14)
r=0.41
r=0.36
Exhausted
Cytotoxic
B
Ki67low (n=8)
Ki67high (n=14)
r=0.73
r=0.14
Check Point Inhibitory
Check Point Activating
C
Color key and histogram
Count
Value
D
TNBC
DAPITIGITCD3SMA
TIGITCD3
TIGIT
CD3
DCIS-14F
IDC-7F
IDC-15F
IDC-15F
IDC-15F
HER2+
DAPITIGITCD3SMA
TIGITCD3
TIGIT
CD3
DCIS-3F
DCIS-5F
DCIS-6F
E
TNBC
DAPIPD-L1PD-1CD3
PD-L1
PD-1
CD3
DCIS-14F
IDC-15F
HER2+
DAPIPD-L1PD-1CD3
PD-L1
PD-1
CD3
DCIS-2F
IDC-4F
IDC-7F
IDC-7F
F
LumA
DAPIPD-L1PD-1CD3
PD-L1
PD-1
CD3
DCIS
microinv DCIS
IDC
G
PD-1+ cells (% of CD3+ T cells)
DCIS (n=5)
IDC (n=5)
DCIS (n=5)
IDC (n=5)
HER2+
TNBC
H
TIGIT+ cells (% of CD3+ T cells)
p=0.4173
IF
FACS

Supplementary Figure S5. Activity status of T cells. **A-B** Plot depicting correlation between exhausted and cytotoxic T cell signatures (**A**) and between the expression of inhibitory and activating immune checkpoint proteins (**B**) in $MKI67^{+}$ (red) and $MKI67^{-}$ (blue) cells. **C**, Heatmap depicting clustering of T cells based on expression of activation-dysfunction-related genes (**Supplementary Table S5**). Red and blue marks $MKI67^{high}$ and $MKI67^{low}$ T cells in tumors, respectively, whereas gray corresponds to normal samples. Orange are activation related genes and pale blue are dysfunction related. **D**, Immunofluorescence analysis of TIGIT, CD3, and SMA in triple negative (TNBC) and $HER2^{+}$ DCIS and IDC. **E**, Immunofluorescence analysis of PD1, PDL1, CD3, and SMA in triple negative (TNBC) and $HER2^{+}$ DCIS and IDC. **F**, Immunofluorescence analysis of PD1, PDL1, and CD3 in a luminal A, IDC with adjacent DCIS. Throughout, images are a montage of nine fields captured from one area of the tissue. Scale bar 50 μm. **G**, Graphs depicting the frequencies of $PD1^{+}CD3^{+}$ T cells in multiple regions of five samples per group. Error bars, S.E.M. **H**, Frequency of $TIGIT^{+}$ T cells determined by FACS or immunofluorescence (IF) in the same cases. Frequencies are not significantly different between the two methods.

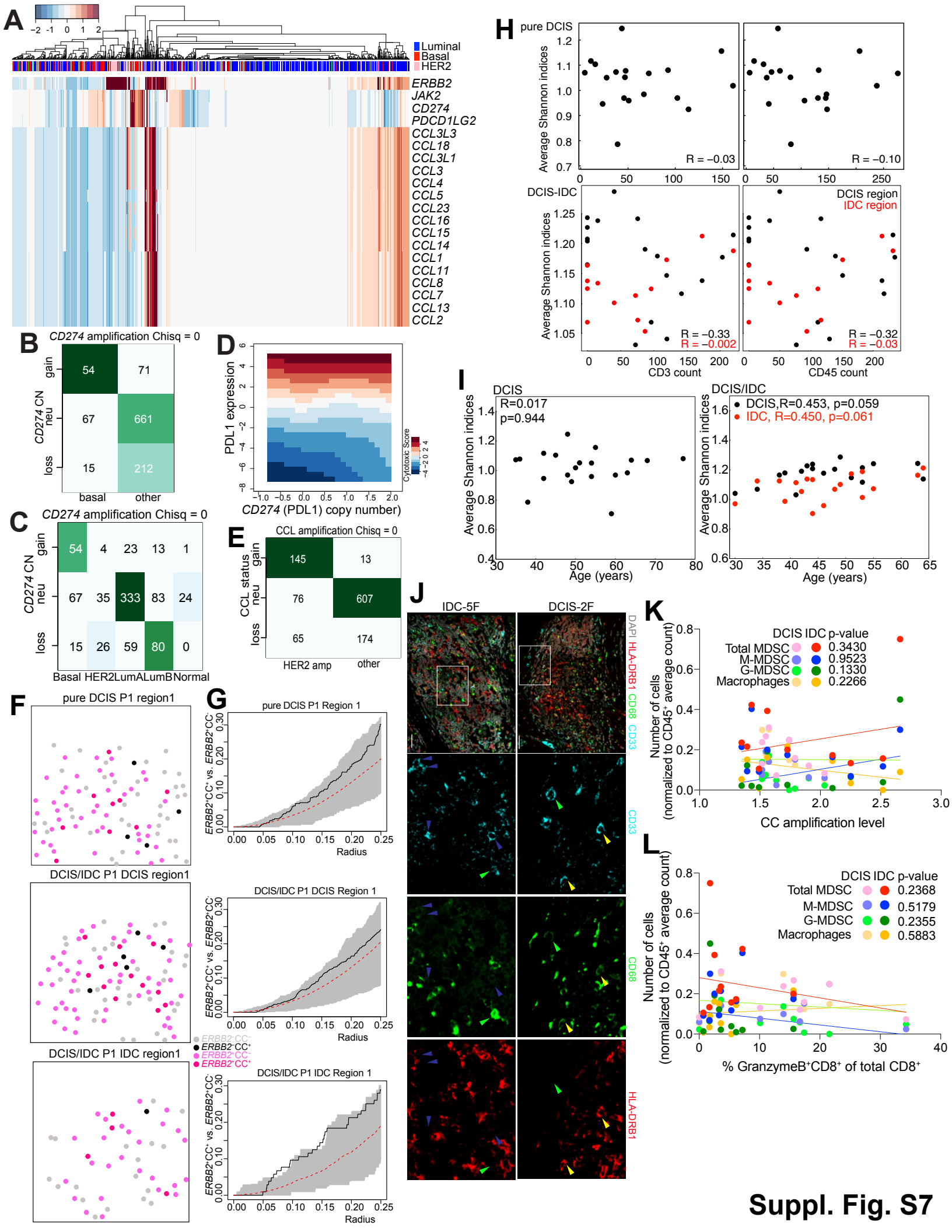

A
TIGIT
PD-1
HER2+ DCIS-06F
HER2+ IDC-07F
HER2+ DCIS-06F
HER2+ IDC-03F
TN DCIS-21F
TN IDC-15F
TN DCIS-23F
TN IDC-16F
X coordinate
Y coordinate
TIGIT-CD3+ cell
TIGIT+CD3+ cell
TIGIT+ cancer cell
TIGIT- cancer cell
PD-1-CD3+ cell
PD-1+CD3+ cell
PD-1+ cancer cell
PD-1- cancer cell

B
TIGIT
PD-1
Invasive percentage
p = 0.19
p = 0.006
p = 0.96
p = 0.01
HER2+ IDC (n=14)
HER2+ DCIS (n=8)
TN IDC (n=15)
TN DCIS (n=11)
HER2+ IDC (n=14)
HER2+ DCIS (n=14)
TN IDC (n=13)
TN DCIS (n=12)


C
IDC
HER2+
DCIS
IDC
TNBC
DCIS

Suppl. Fig. S6

Supplementary Figure S6. Tumor topology. **A**, Two-dimensional coordinate spatial distribution of T cell populations and tumor cell populations. **B**, Boxplot depicting invasive percentage scores between tumor cells and T cells (defined in **Methods** section) stratified by tissue types ($HER2^+$ IDC; $HER2^+$ DCIS; TN IDC; TN DCIS): left panel: samples with TIGIT expression being measured; right panel PD1 expression level being measured. P-values are calculated using two-sided T test. **C**, Trichrome-stained images of DCIS and IDC. Scale bar 100 μm.


A
Luminal
Basal
HER2
ERBB2
JAK2
CD274
PDCD1LG2
CCL3L3
CCL18
CCL3L1
CCL3
CCL4
CCL5
CCL23
CCL16
CCL15
CCL14
CCL1
CCL11
CCL8
CCL7
CCL13
CCL2
B
CD274 amplification Chisq = 0
CD274 CN
gain
neu
loss
basal
other
C
CD274 amplification Chisq = 0
Basal HER2 LumA LumB Normal
D
PDL1 expression
CD274 (PDL1) copy number)
Cytotoxic Score
E
CCL amplification Chisq = 0
CCL status
HER2 amp
F
pure DCIS P1 region1
DCIS/IDC P1 DCIS region1
DCIS/IDC P1 IDC region1
G
pure DCIS P1 Region 1
DCIS/IDC P1 DCIS Region 1
DCIS/IDC P1 IDC Region 1
Radius
H
pure DCIS
DCIS-IDC
Average Shannon indices
R = −0.03
R = −0.10
R = −0.33
R = −0.002
R = −0.32
R = −0.03
DCIS region
IDC region
CD3 count
CD45 count
I
DCIS
R=0.017
p=0.944
DCIS/IDC
DCIS,R=0.453, p=0.059
IDC, R=0.450, p=0.061
Age (years)
J
IDC-5F
DCIS-2F
DAPI HLA-DRB1 CD68 CD33
CD33
CD68
HLA-DRB1
K
DCIS IDC p-value
Total MDSC 0.3430
M-MDSC 0.9523
G-MDSC 0.1330
Macrophages 0.2266
Number of cells (normalized to CD45+ average count)
CC amplification level
L
DCIS IDC p-value
Total MDSC 0.2368
M-MDSC 0.5179
G-MDSC 0.2355
Macrophages 0.5883
% GranzymeB+CD8+ of total CD8+

Suppl. Fig. S7

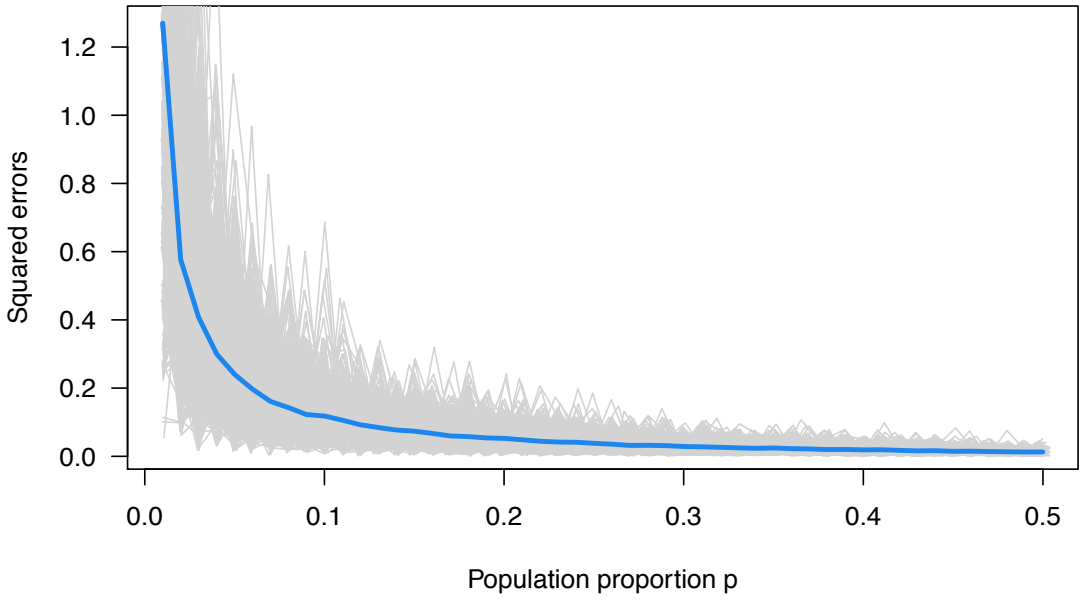
Supplementary Figure S7. Immune-related amplicons in breast cancer. **A**, Copy number (log2 ratio) for *ERBB2*, and 17q12 and 9p24 amplicon genes in the TCGA cohort. **B-C**, Contingency table of the frequency of gain of the CD274 (PDL1) locus in basal-like patients compared to all other patients in the TCGA cohort. **D**, Level plot illustrating the contribution of PDL1 gene expression and *CD274* copy number to cytotoxic $CD8^+$ T cell gene expression score. **E**, Contingency table of amplification of 17q12 (CC) locus in $HER2^+$ patients compared to all other patients. **F**, Two-dimensional coordinate spatial distribution of the indicated four cell types in pure DCIS, and in DCIS and IDC regions of DCIS/IDC. **G**, Graphs depicting that the topologic distribution of most of the real samples do not deviate from the complete spatial randomness significantly, manifested by the Ripley's K function of the data (black curve) lying within the 95% confidence band of the K function of the simulated datasets (grey area). **H**, Average Shannon index for 17q12 amplicon over all the samples of each patient vs. $CD3^+$ and $CD45^+$ cell counts in pure DCIS and in DCIS/IDC (red dots: DCIS, black dots: IDC). No significant association between CD3 or CD45 cell counts and average Shannon index is detected. **I**, Average Shannon index based on *ERBB2* and CC amplification for all the cells in each region across sampled regions in each patient vs. patient's age. **J**, Immunofluorescence analysis of myeloid-derived suppressor cells (MDSCs) and macrophages. Arrows indicate cell types as follows: macrophage – yellow, monocytic MDSC – blue, granulocytic MDSC – green. Scale bar 50 μm. **K-L**, Plot depicting correlation between number of MDSCs in DCIS and IDC and CC copy number (**K**) and frequencies of $GZMB^+CD8^+$ T cells (**L**). M-MDSC: monocytic MDSCs and G-MDSC: granulocytic MDSC.

Supplementary Methods

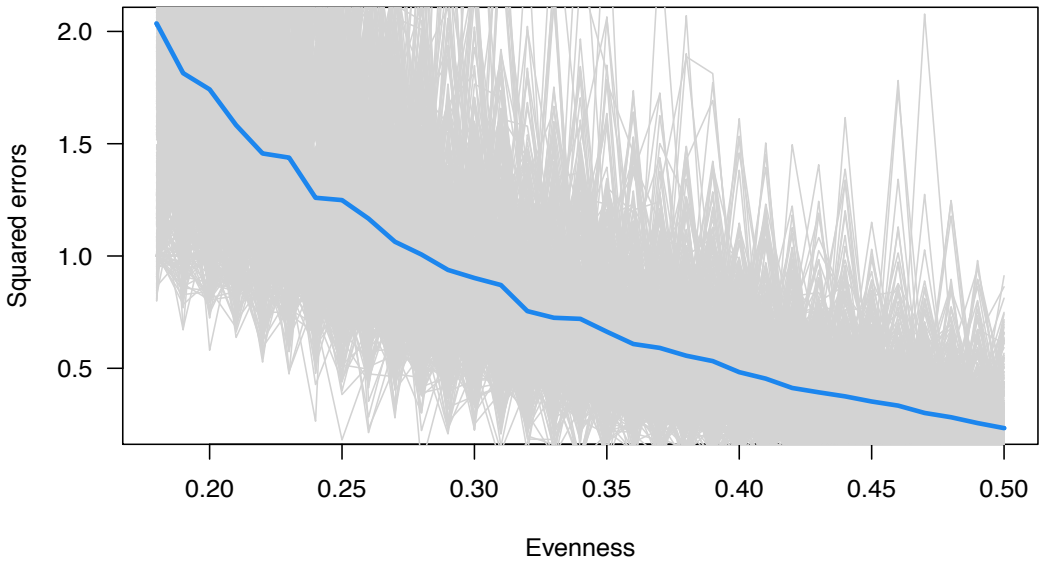
Attempt at addressing the sampling error problem when estimating cell frequencies based of staining of tissue sections vs. whole tumor analysis (e.g., FACS):

We design the following spatial-based simulations to demonstrate the difficulty level of sampling a region from a 1cm × 1cm × 1cm squared cube that is representative of the spatial distribution of different cell types. For simplicity, we assume there are only two types of cells (Type 0 and Type 1), e.g. Tregs and all other Tcells. We further assume that there exist n = 10,000 cells in total in each squared cube. To mimic the real-life experiments, we then investigate the normalized squared distance between the population proportion of Type 1 cells and the sampled proportions of Type 1 cells by taking five different 4u × 1cm × 1cm slides. The positions of the five sampled slides are chosen randomly along one of the three dimensions. We consider a range of parameters by (1) changing the population proportion of Type 1 cells and (2) changing the spatial distribution of Type 1 cells from uniformly distributed to clustered in a focal region. The simulations are generated as follows:

1. The positions of the cells not marked with type yet are distributed as a three-dimensional uniform distribution Uniform ([0,1] × [0,1] × [0,1]) .
2. We first look at how the magnitude of the proportion of Type 1 cells influences the sampling error. Conditioning on each position, we assume that this proportion is always p, regardless of the position of the cell. The figure below shows that as the proportion increases the error becomes smaller. The grey lines are the squared distance for each 500 iterations of the simulation and the blue line is the average across 500 iterations. The squared distance is normalized by the true population proportion.



3. Then we fix the proportion of Type 1 cells to be 0.05, and change the “evenness” of the cell type distribution within the cube. We define the “evenness” as how tight the Type 1 cells are distributed around the center of the cube for simplicity. For instance, when the “evenness = 0.2”, this means that all Type 1 cells are located within 0.2cm × 0.2cm × 0.2cm cubic neighborhood around the center (0.5, 0.5, 0.5) of the cube. The larger this value is, the more “even” the Type 1 cells are distributed in the original cube. Figure 2 below shows that the estimation error decreases as the evenness increases, as we expected. The grey lines are the squared distance for each 500 iterations of the simulation and the blue line is the average across 500 iterations.



The predictions of this model perfectly fit with our data where we saw poor correlation for Tregs that are <10% frequency and good correlation for $CD4^+$ and $CD8^+$ cells that are more common (~40-50% of all T cells). This simulation also highlights that all studies that quantify rare cell types based on immunostaining of slides have a high error rate and should be taken with caution.