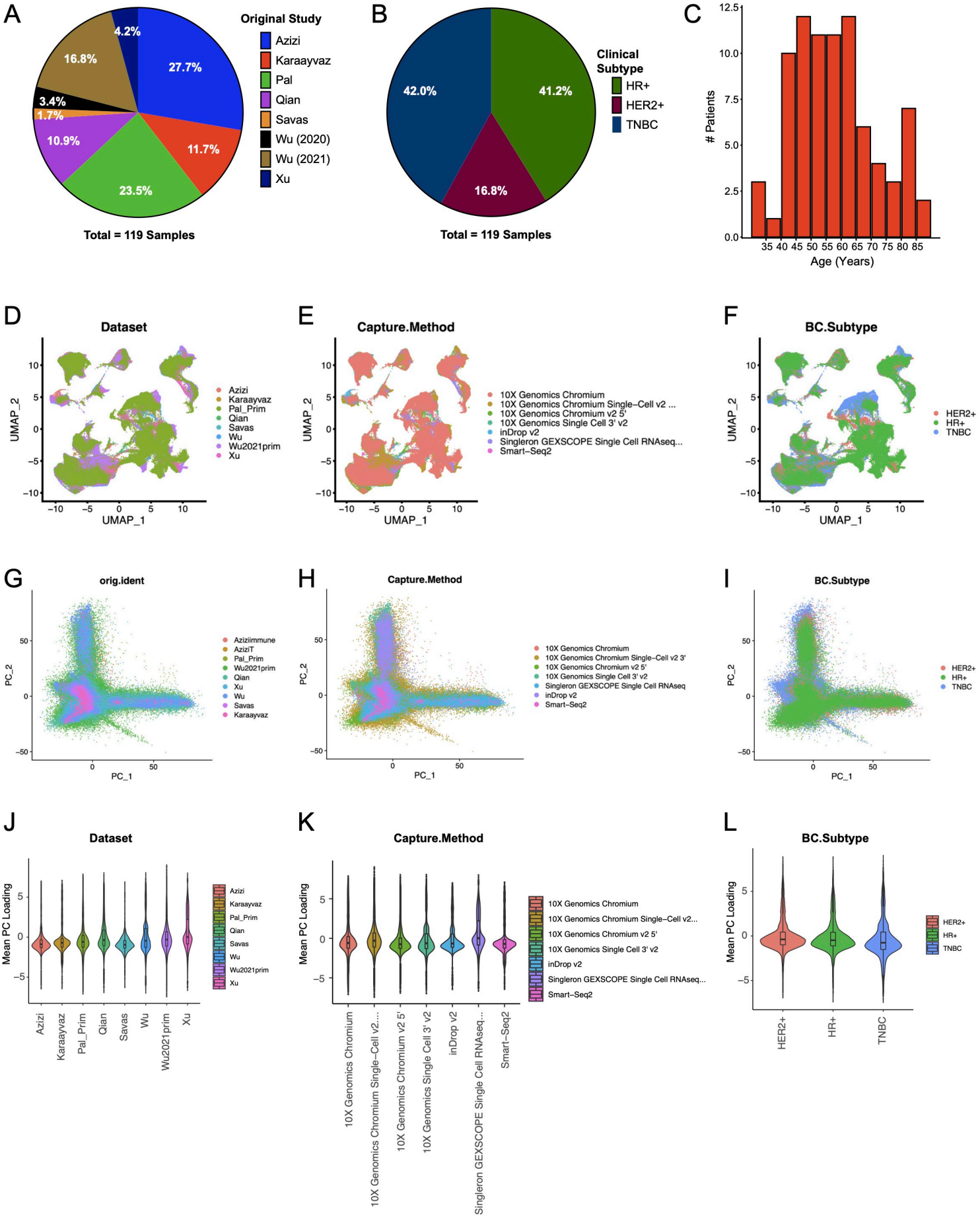
Cell Reports Medicine, Volume *5*

Supplemental information

A comprehensive single-cell breast tumor atlas defines epithelial and immune heterogeneity and interactions predicting anti-PD-1 therapy response

Lily Xu, Kaitlyn Saunders, Shao-Po Huang, Hildur Knutsdottir, Kenneth Martinez-Algarin, Isabella Terrazas, Kenian Chen, Heather M. McArthur, Julia Maués, Christine Hodgdon, Sangeetha M. Reddy, Evanthia T. Roussos Torres, Lin Xu, and Isaac S. Chan

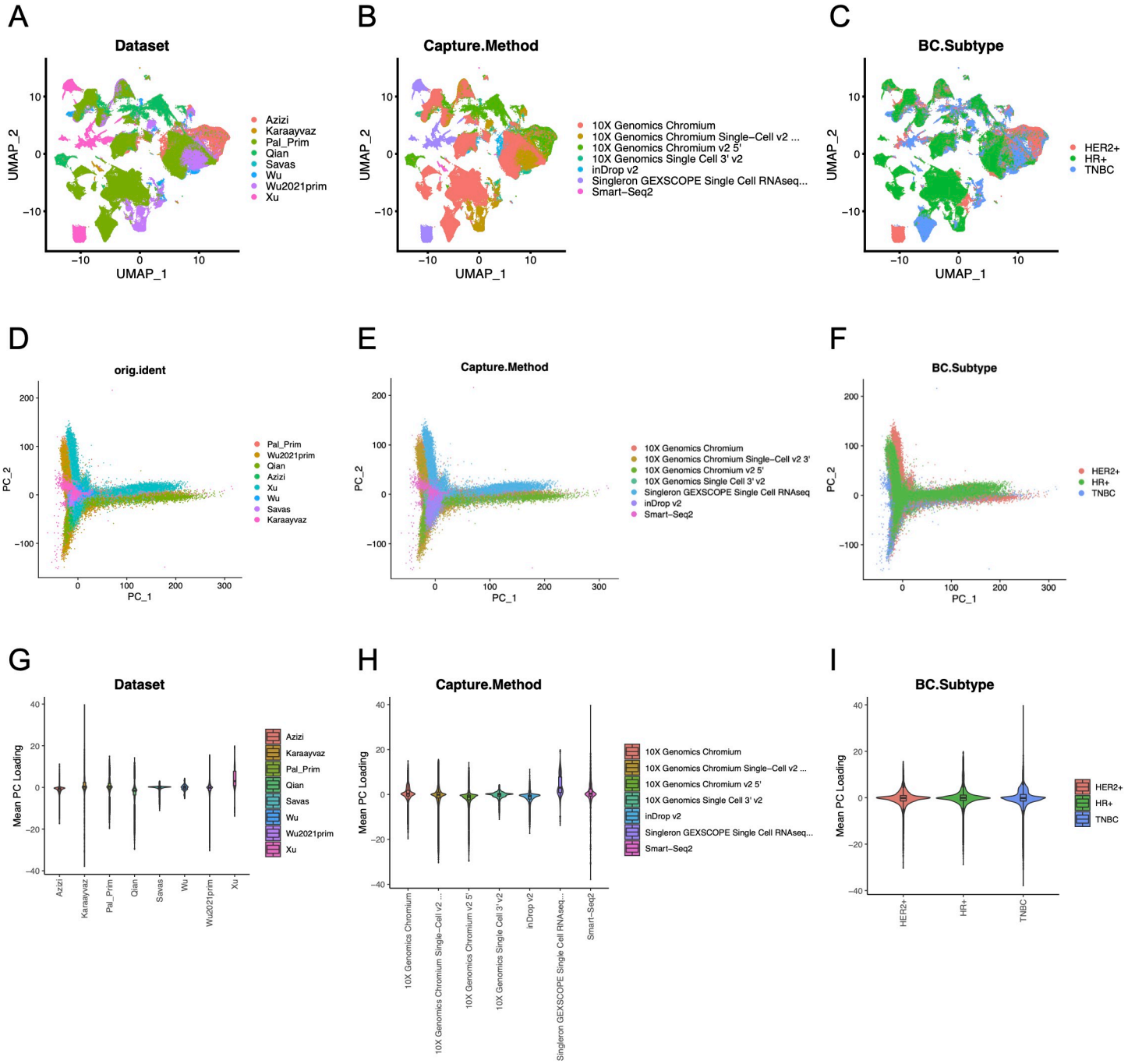
Supplemental Figure 1



Supplemental Figure 1. Metadata of integrated dataset and visualization of integrated dataset following batch correction, Related to Figure 1 and STAR Methods.

A. Pie chart of composition of integrated scRNA-seq data by original study.
B. Pie chart of composition of integrated scRNA-seq data by clinical subtype. The proportion of clinical subtypes within this integrated dataset is close to real-life clinical subtype distributions.
C. Bar plot showing number of patients per age group. Most of the original datasets stayed within a sole age group, whereas the integrated dataset includes a much broader age range.
D. UMAP visualization of integrated dataset following batch correction, grouped by source dataset.
E. UMAP visualization of integrated dataset following batch correction, grouped by capture technology.
F. UMAP visualization of integrated dataset following batch correction, grouped by clinical subtype. This shows lineage drives clustering of non-epithelial populations, while epithelial populations cluster by clinical subtype. This matches the observed subtype clustering seen in other datasets.
G. PCA plot of first 2 PCs for all cells in the integrated dataset following batch correction, labeled by original source dataset. No cluster is driven by a single study, thus confirming there is no batch effect due to different studies.
H. PCA plot of first 2 PCs for all cells in the integrated dataset following batch correction, labeled by technology. No cluster is driven by a single technology, thus confirming there is no batch effect due to differing technologies.
I. PCA plot of first 2 PCs for all cells in the integrated dataset following batch correction, labeled by clinical subtype.
J. Violin plots of mean PC loadings across top 20 PCs for the integrated dataset following batch correction, stratified by source dataset.
K. Violin plots of mean PC loadings across top 20 PCs for the integrated dataset following batch correction, stratified by capture technology.
L. Violin plots of mean PC loadings across top 20 PCs for the integrated dataset following batch correction, stratified by clinical subtype.

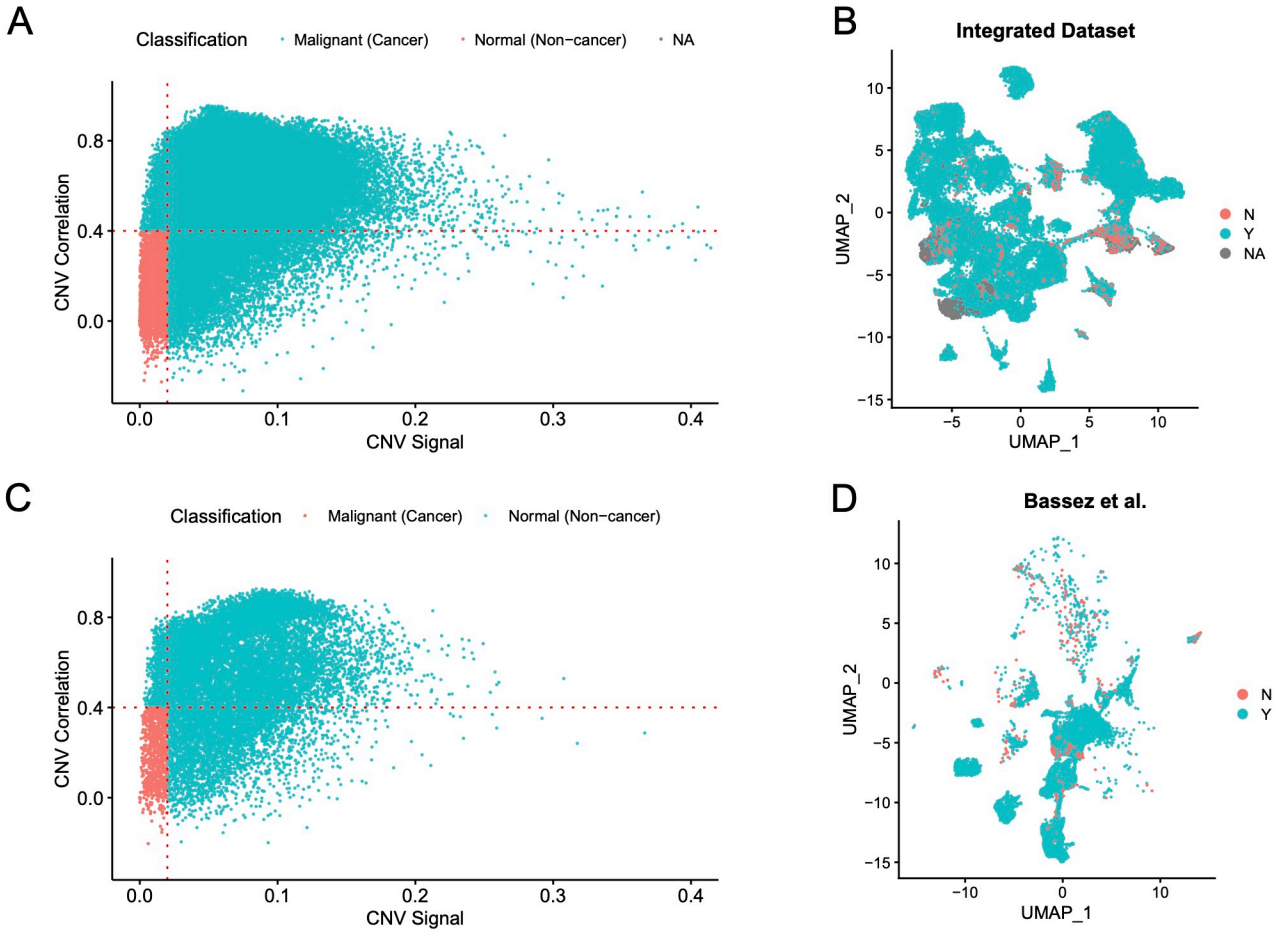
Supplemental Figure 2



Supplemental Figure 2. Visualization of combined original source datasets prior to batch correction, Related to Figure 1 and STAR Methods.

A. UMAP visualization of combined original source datasets prior to batch correction, grouped by source dataset.
B. UMAP visualization of combined original source datasets prior to batch correction, grouped by capture technology.
C. UMAP visualization of combined original source datasets prior to batch correction, grouped by clinical subtype.
D. PCA plot of first 2 PCs for combined original source datasets prior to batch correction, labeled by source dataset.
E. PCA plot of first 2 PCs for combined original source datasets prior to batch correction, labeled by capture technology.
F. PCA plot of first 2 PCs for combined original source datasets prior to batch correction, labeled by clinical subtype.
G. Violin plots of mean PC loadings across top 20 PCs for combined original source datasets prior to batch correction, stratified by source dataset.
H. Violin plots of mean PC loadings across top 20 PCs for combined original source datasets prior to batch correction, stratified by capture technology.
I. Violin plots of mean PC loadings across top 20 PCs for combined original source datasets prior to batch correction, stratified by clinical subtype.

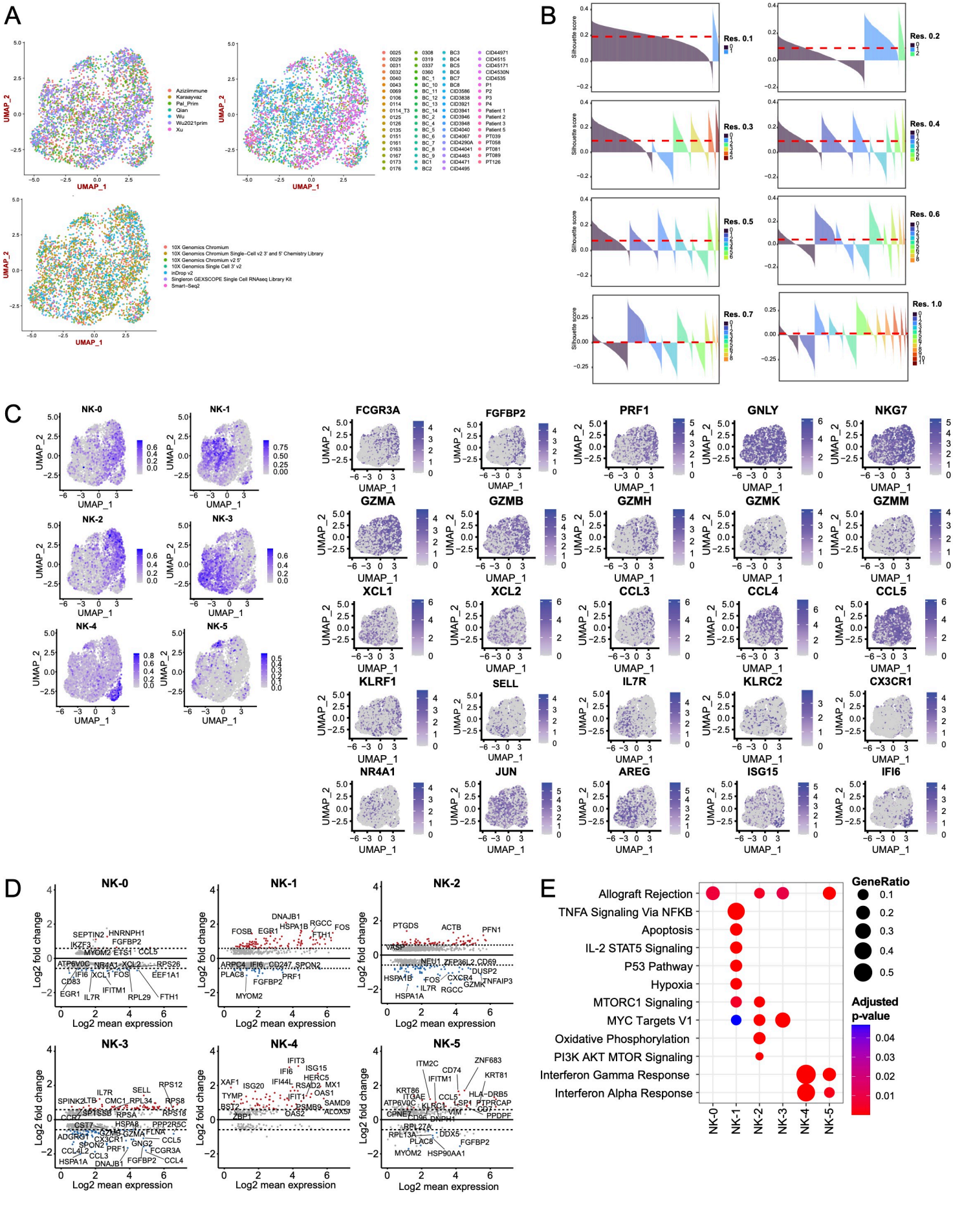
Supplemental Figure 3



Supplemental Figure 3. Classification of epithelial cells as cancer versus normal using CNV profile analysis, Related to Figure 1 and STAR Methods.

A. Scatter plot showing classification of epithelial cells in the integrated dataset as cancer (malignant) versus normal (non-malignant) on inferCNV signal (x-axis) and CNV correlation (y-axis). Thresholds shown in red dashed lines. CNV signal reflects the extend of CNVs, while CNV correlation reflects the similarity between the cellular CNV pattern and that of other cells from the same tumor. Cells assigned as cancer (malignant) are shown in blue, while the rest are shown in red.
B. UMAP visualization of all epithelial cells in the integrated dataset, grouped by classification as cancer (malignant) versus normal (non-malignant). Cancer cells are shown in blue, while normal cells are shown in red. Unassigned cells are shown as NAs and are colored grey.
C. Scatter plot showing classification of epithelial cells in the Bassez et al. dataset as cancer (malignant) versus normal (non-malignant) on inferCNV signal (x-axis) and CNV correlation (y-axis).
D. UMAP visualization of all epithelial cells in the Bassez et al. dataset, grouped by classification as cancer (malignant) versus normal (non-malignant). Cancer cells are shown in blue, while normal cells are shown in red.

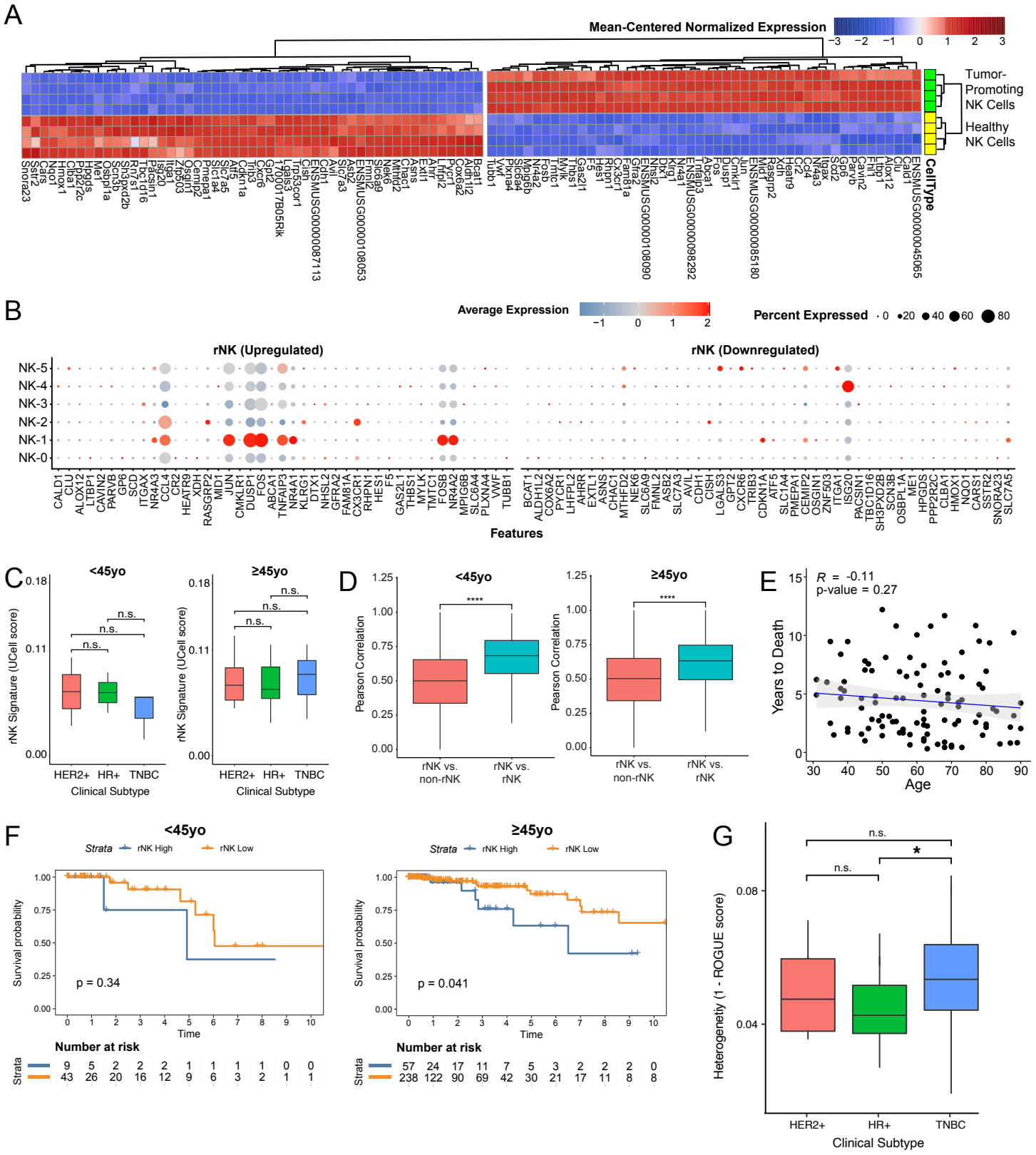
Supplemental Figure 4



Supplemental Figure 4. Unsupervised clustering of NK cells and analysis of NK cell subsets, Related to Figure 1.

A. UMAP visualization of all NK cells in the integrated dataset, grouped by source dataset. UMAP visualization of all epithelial cells in the integrated dataset, grouped by patient. UMAP visualization of all epithelial cells in the integrated dataset, grouped by capture technology.
B. Silhouette scores for clustering of NK cells at various resolutions (0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 1.0). Mean silhouette score is shown as a red dashed line. Maximum mean silhouette score was observed at resolution 0.1 (2 clusters), and second highest mean silhouette score was observed at resolution 0.3 (6 clusters).
C. Feature plots showing expression of NK subset markers across all NK cells in our integrated dataset. Feature plots showing expression of functional NK cell genes across all NK cells in our integrated dataset.
D. MA plots showing differentially expressed genes between individual NK cell subsets and all other NK cell subset types (Bonferroni adjusted p-value < 0.05).
E. Gene set enrichment of the differentially expressed genes by each NK cell subset. Significantly enriched gene sets from the MSigDB HALLMARK collection are shown (Benjamini-Hochberg adjusted p-value < 0.05).

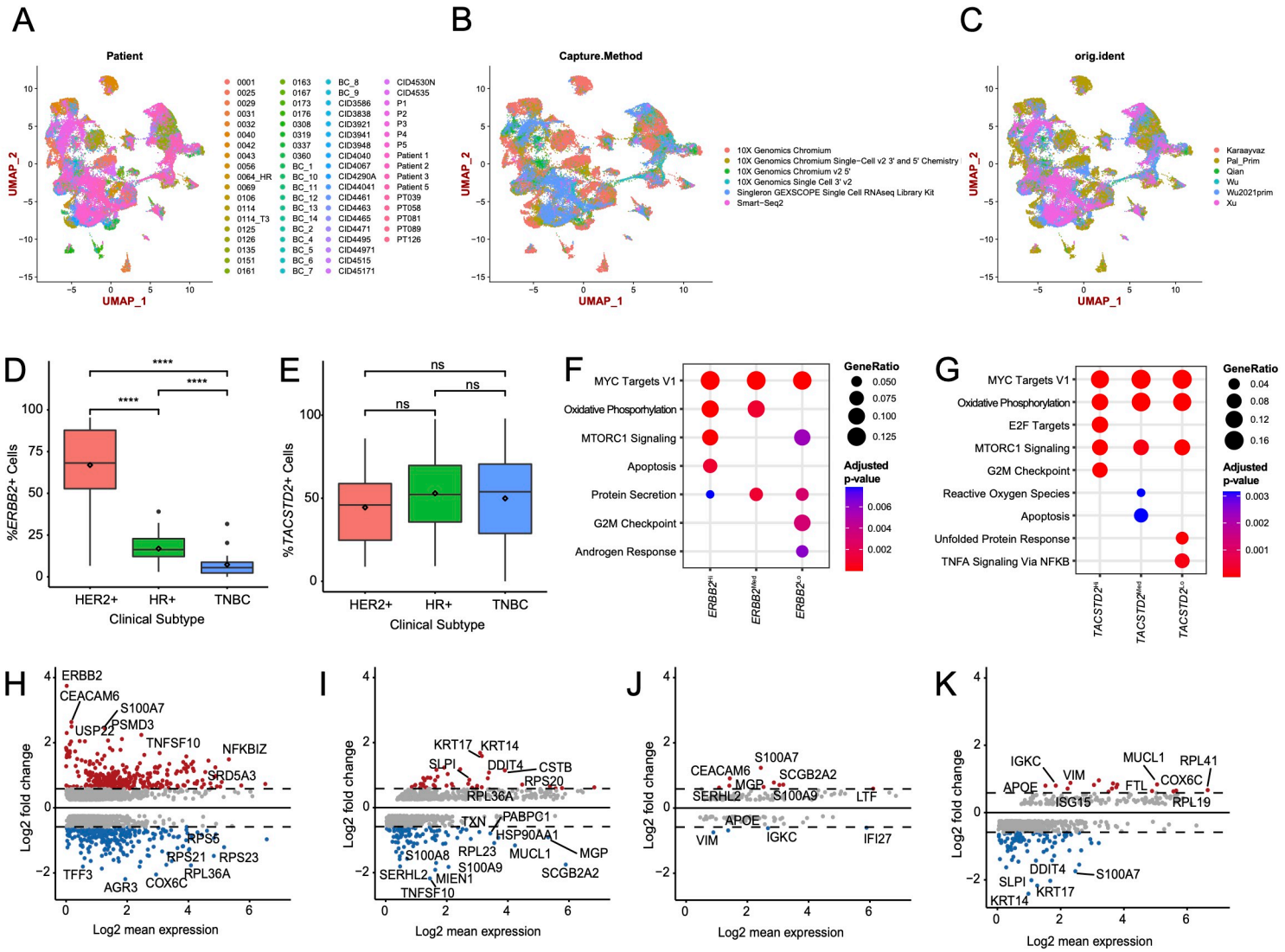
Supplemental Figure 5



Supplemental Figure 5. rNK signature development and analysis, Related to Figure 1.

A. Heatmap showing z-scores for the variance-stabilized transformed expression of differentially expressed genes between healthy NK cells and tumor-promoting NK cells from previous study.
B. Bubble heatmap showing expression of upregulated and downregulated human rNK orthologs for each major NK cell subset.
C. Boxplot showing the expression level of the rNK signature by clinical subtype, stratified by age. No significant difference was found between subtypes (Kruskal-Wallis $p > 0.05$).
D. Boxplot showing the Pearson correlations of rNK signature gene expression in reprogrammed NK (rNK) cells compared to non-rNK cells versus rNK cells compared to rNK cells, stratified by age. Pearson correlations between rNK cells and rNK cells are higher than those between rNK cells and non-rNK cells for both age strata (two-sided Wilcoxon test, ****p-value < 0.0001).
E. Scatterplot showing the Pearson correlation between age at initial diagnosis and survival across TCGA samples (p-value > 0.01).
F. Kaplan-Meier plots evaluating the influence of rNK cell gene signature expression on survival outcomes in TCGA patients with relatively high fraction of NK cells, stratified by age. For patients ≥45yo, high rNK cell gene signature expression is associated with worse survival outcomes (log-rank test, p-value < 0.05).
G. Boxplot showing heterogeneity calculated as 1 – ROGUE score for NK cells in each sample by breast cancer clinical subtype (*p-value < 0.05).

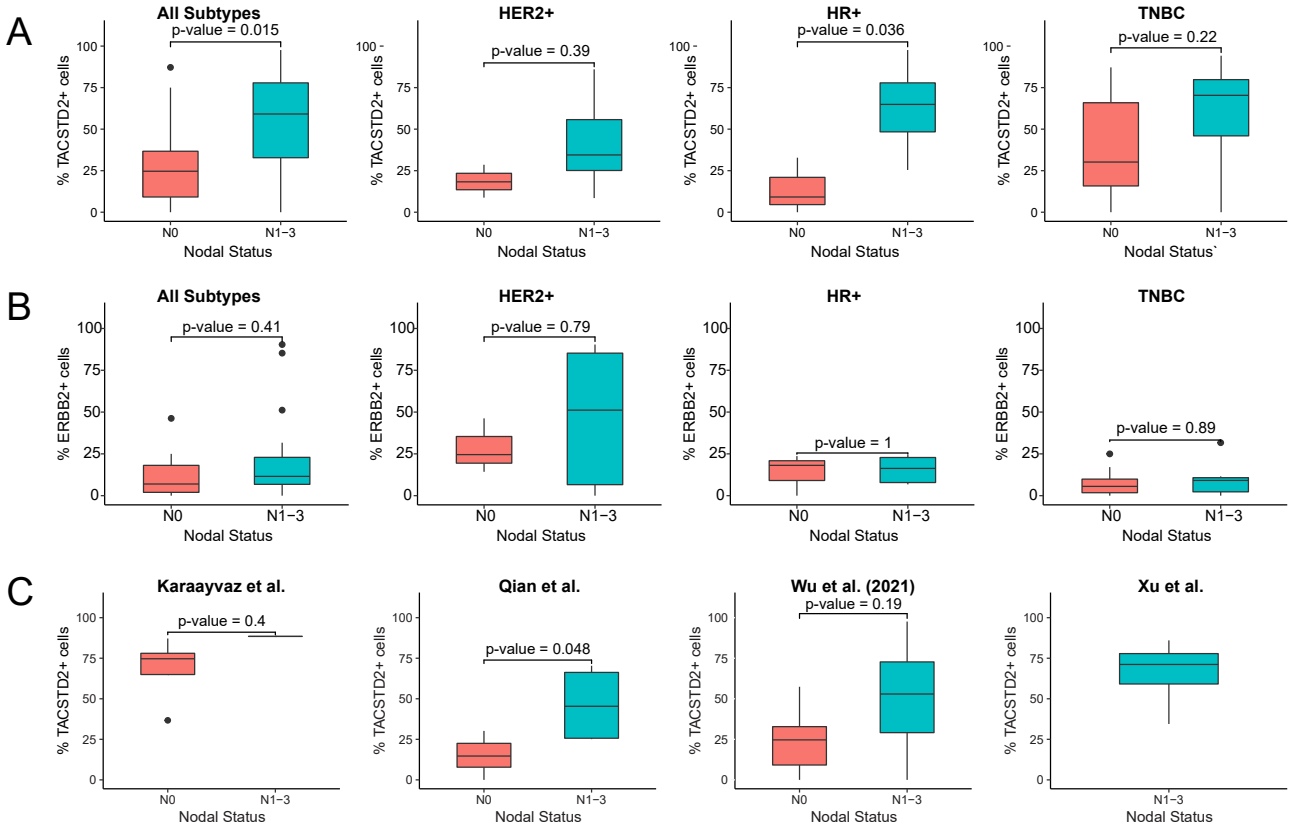
Supplemental Figure 6



Supplemental Figure 6. Differential gene expression and gene set enrichment analyses for each *ERBB2* and *TACSTD2* population, Related to Figure 2.

A. UMAP visualization of all epithelial cells in the integrated dataset, grouped by patient. Consistent with other tumor type and breast tumor datasets, epithelial cells appear to cluster by patient.
B. UMAP visualization of all epithelial cells in the integrated dataset, grouped by capture technology.
C. UMAP visualization of all epithelial cells in the integrated dataset, grouped by source dataset.
D. Boxplot showing % *ERBB2*+ cells by clinical subtype across samples in the integrated dataset. As anticipated, % *ERBB2*+ cells were significantly enriched in HER2+ samples compared to HR+ and TNBC samples (Kruskal-Wallis $p < 0.05$, with post-hoc Dunn test p-values shown).
E. Scatterplot showing the Pearson correlation between HER2+ protein expression and *ERBB2* mRNA expression across TCGA samples ($p < 0.0001$).
F. Gene set enrichment of the differentially expressed genes by *ERBB2*^{Hi}, *ERBB2*^{Med}, and *ERBB2*^{Lo} cells. Significantly enriched gene sets from the MSigDB HALLMARK collection are shown (Benjamini-Hochberg adjusted p-value < 0.05).
G. Gene set enrichment of the differentially expressed genes by *TACSTD2*^{Hi}, *TACSTD2*^{Med}, and *TACSTD2*^{Lo} cells. Significantly enriched gene sets from the MSigDB HALLMARK collection are shown (Benjamini-Hochberg adjusted p-value < 0.05).
H. MA plot showing differentially expressed genes between *ERBB2*^{Hi} vs. *ERBB2*^{Med} and *ERBB2*^{Lo} cells (Bonferroni adjusted p-value < 0.05).
I. MA plot showing differentially expressed genes between *ERBB2*^{Lo} vs. *ERBB2*^{Hi} and *ERBB2*^{Med} cells (Bonferroni adjusted p-value < 0.05).
J. MA plot showing differentially expressed genes between *TACSTD2*^{Med} vs. *TACSTD2*^{Hi} and *TACSTD2*^{Lo} cells (Bonferroni adjusted p-value < 0.05).
K. MA plot showing differentially expressed genes between *TACSTD2*^{Lo} vs. *TACSTD2*^{Hi} and *TACSTD2*^{Med} cells (Bonferroni adjusted p-value < 0.05).

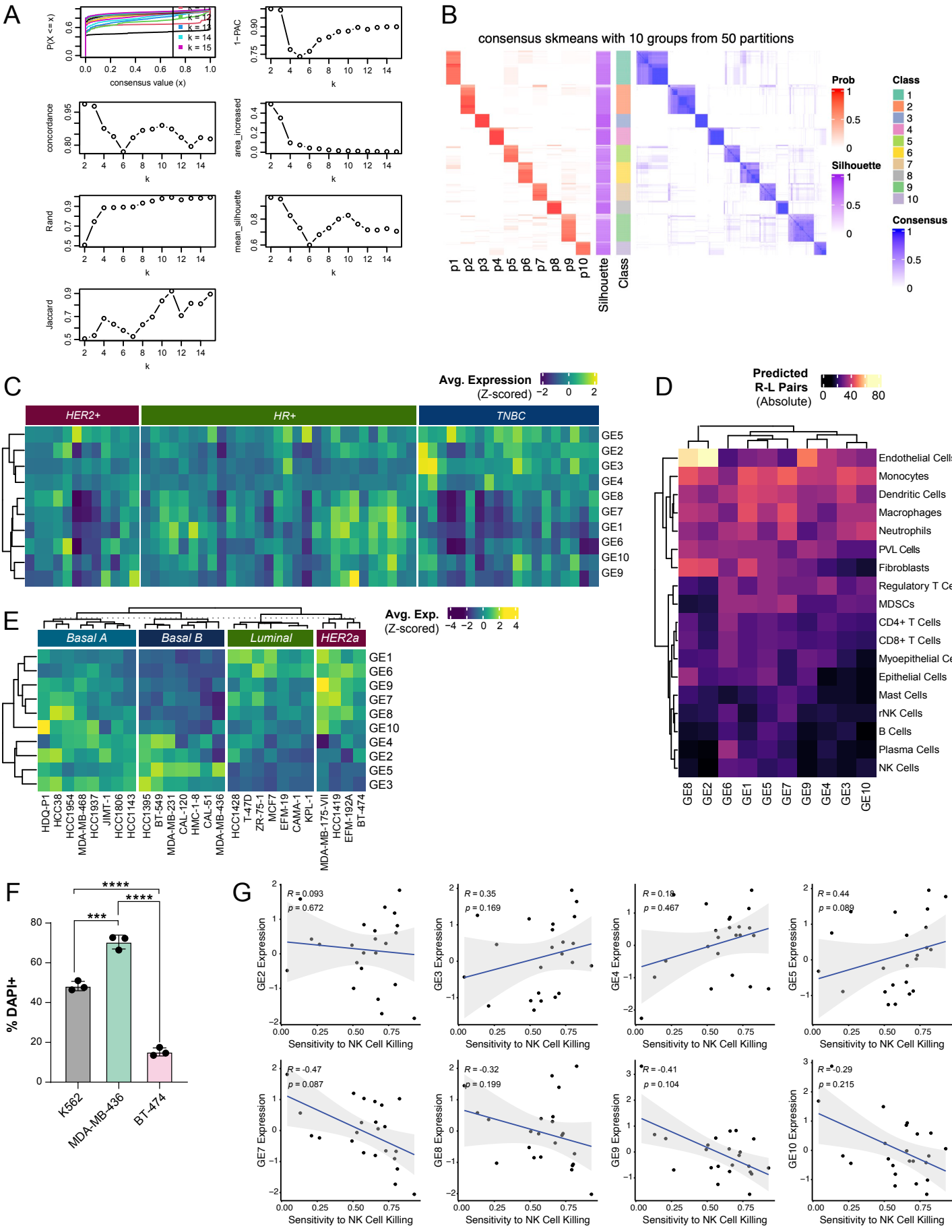
Supplemental Figure 7



Supplemental Figure 7. Analysis of clinical features and associations across samples in the integrated dataset, Related to Figure 2.

A. Boxplots showing the proportion of *TACSTD2*-expressing cells per sample by nodal status, split by clinical subtype (two-sided Wilcoxon test p-value as shown).
B. Boxplots showing the proportion of *ERBB2*-expressing cells per sample by nodal status, split by clinical subtype (two-sided Wilcoxon test p-value as shown).
C. Boxplots showing the proportion of *TACSTD2*-expressing cells per sample by nodal status, split by original source dataset (two-sided Wilcoxon test p-value as shown). The combined result was not a statistically significant finding, though it does trend toward significance (Fisher's combined probability test, $X = 11.227$, $p = 0.08$).

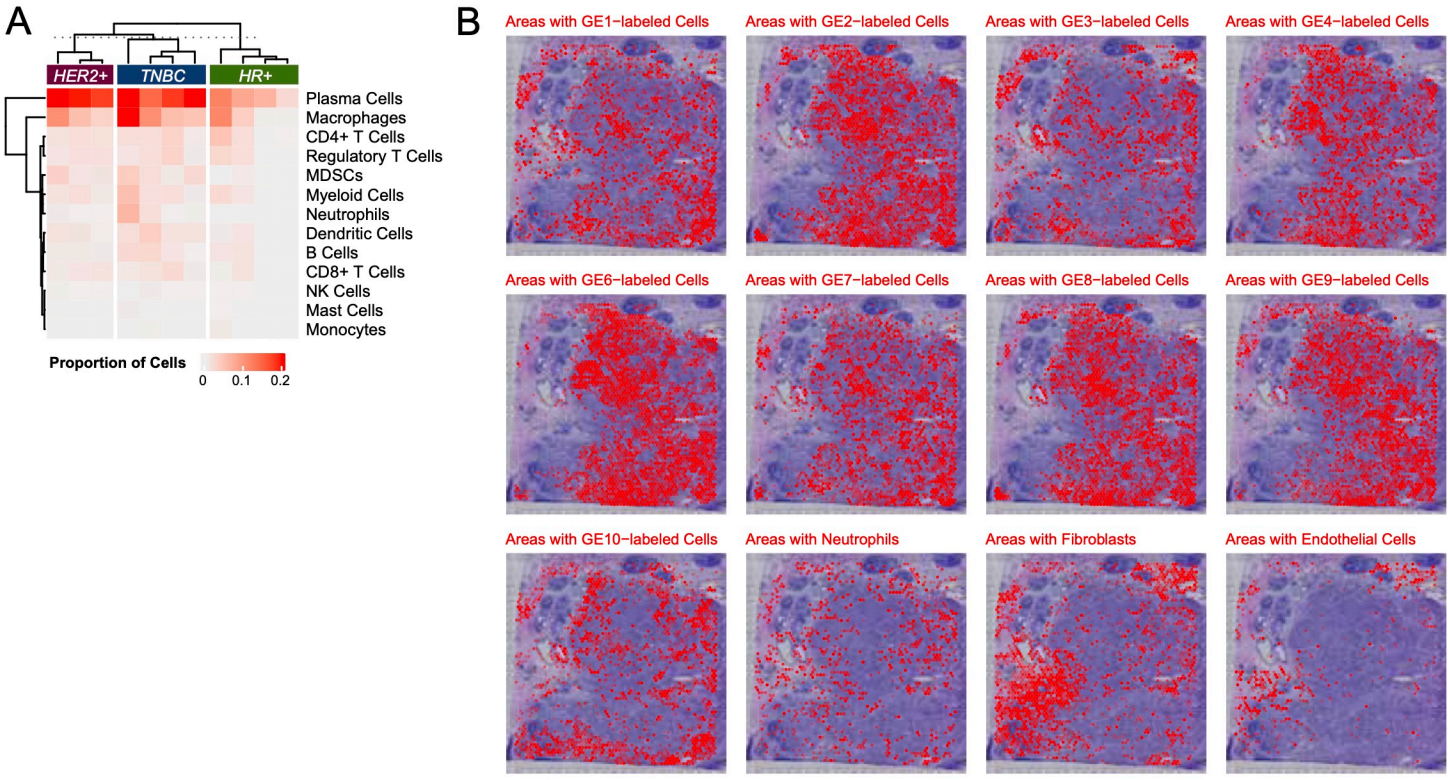
Supplemental Figure 8



Supplemental Figure 8. Generation of the 10 gene elements of cancer epithelial cell heterogeneity and exploration in breast cancer cell lines, Related to Figure 3.

A. Metrics used to select the number of clusters (10) for consensus clustering of signatures of cancer epithelial cell ITTH.
B. Spherical k-means (skmeans) consensus clustering of the Jaccard similarities between signatures of cancer epithelial cell ITTH, showing the probability (p1-p10) of each generated signature of being assigned to one of 10 classes. Silhouette scores are shown for each class or GE.
C. Heatmap of average z-scored expression of each of the 10 GEs across cancer epithelial cells in each sample in our integrated dataset.
D. Heatmap of the absolute number of curated predicted receptor-ligand pairs between cancer epithelial cells by GE and interacting immune and stromal cells.
E. Heatmap of average z-scored expression of each of the 10 GEs across human breast cancer cell lines. Cell lines are annotated by molecular subtype (luminal, basal A, basal B, HER2-amplified).
F. Cytotoxicity of NK-92 cells against BT-474, MDA-MB-436, and K562 cell lines, assessed by % DAPI+ cells at 24 hr timepoint. BT-474 highly expressed NK-resistance GEs (GE1 and GE6), while MDA-MB-436 has low expression of NK-resistance GEs. Cytotoxicity was significantly reduced for the BT-474 cell line compared to the MDA-MB-436 cell line (3 biological replications; Benjamini-Hochberg adjusted, ***p-value < 0.001, ****p-value < 0.0001).
G. Scatterplots showing Spearman correlations of expression of GEs with limited predicted interactions with NK cells (all but GE1 and GE6) and sensitivity to NK cell killing across human breast cancer cell lines (Benjamini-Hochberg adjusted p-values > 0.05).

Supplemental Figure 9



Supplemental Figure 9. Predicted GE-immune interactions and spatial analysis of the 10 gene elements, Related to Figure 4.

A. Heatmap showing the proportion of spatial tumor sample spots within a sample that contain each of the GEs and immune or stromal cell populations.
B. For a representative sample, UCell signature scores of each GE overlaid onto spatial tumor sample spots with >10% presence of cancer epithelial cells.