

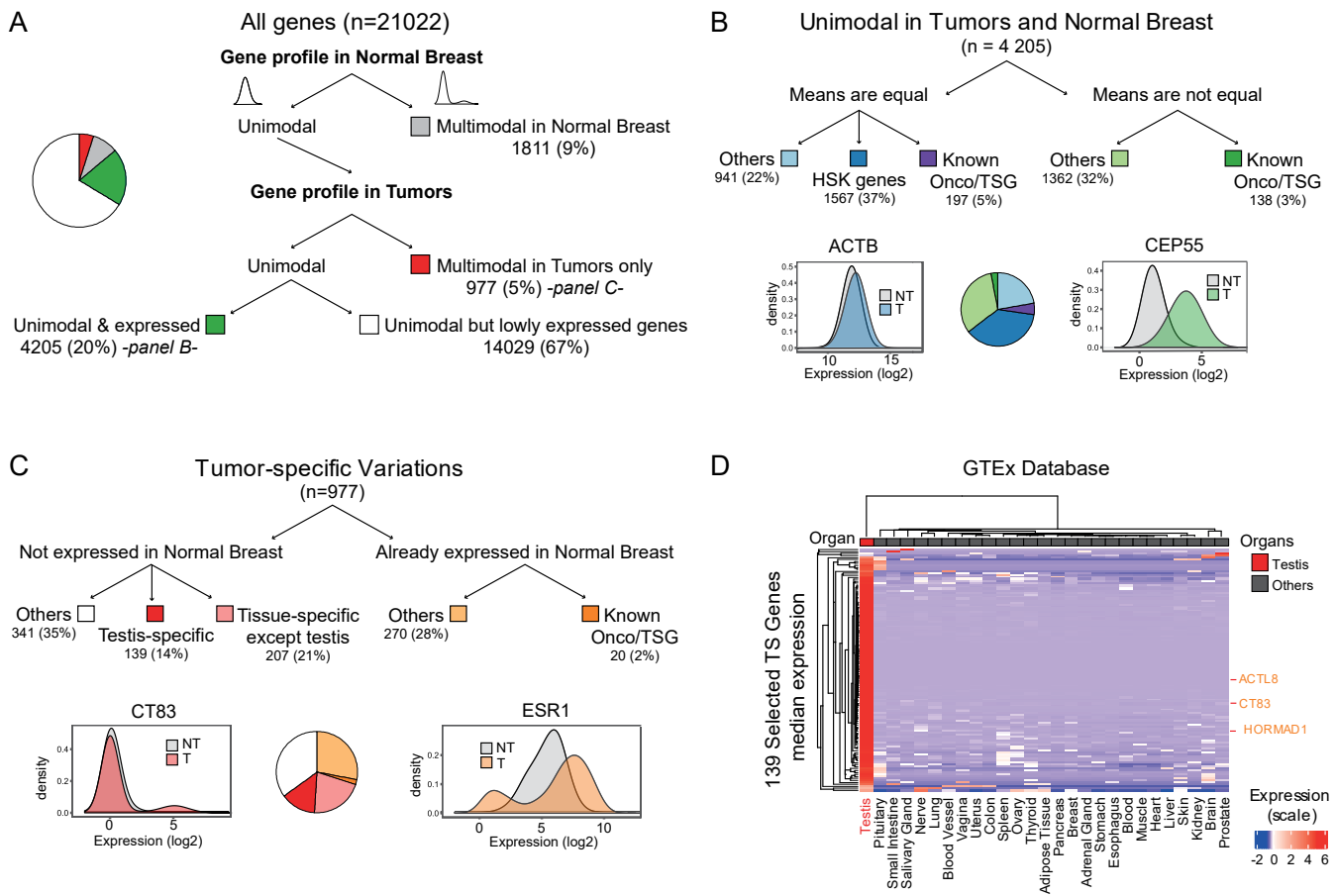


Figure S1: Further uses and validation of our novel bioinformatic screening approach

A. Classification of all genes according to our parameters: we were interested in genes with a homogeneous expression in Normal Breast (NB); ie. with a unimodal profile in Normal Breast. These genes were subsequently divided according to their expression pattern in breast tumors, with two situations of specific interest: genes also homogeneously expressed in breast tumors (panel B) and genes that are overexpressed or repressed in a subset of breast tumors (panel C).

B. Refinement of the characterization of genes homogeneously expressed both in Normal Breast and in breast tumors: when means were significantly different between Normal Breast and tumors, these genes could be used as tumor markers. Some of these genes are known as overexpressed oncogenes or repressed tumor suppressor genes; a significant portion (1362 genes) have no currently known role in cancer but could play a role in breast tumor development. HSK: housekeeping gene; TSG: Tumor Suppressor Gene. Onco: oncogene.

C. Refinement of the characterization for tumor-specific variables genes: approximately 70% of them are repressed in Normal Breast and abnormally activated in breast tumors; amongst these genes are known tissue-specific genes (including testis-specific genes). The remaining 30% are overexpressed or repressed genes in some breast tumors, including known subtype-specific oncogenes like ESR1 and other genes that could be used as a marker of specific tumor subgroups.

D. Heatmap showing the mean expression values (Z-score) for the 139 selected C/T genes in various human adult tissues based on RNA-seq data from GTEx.

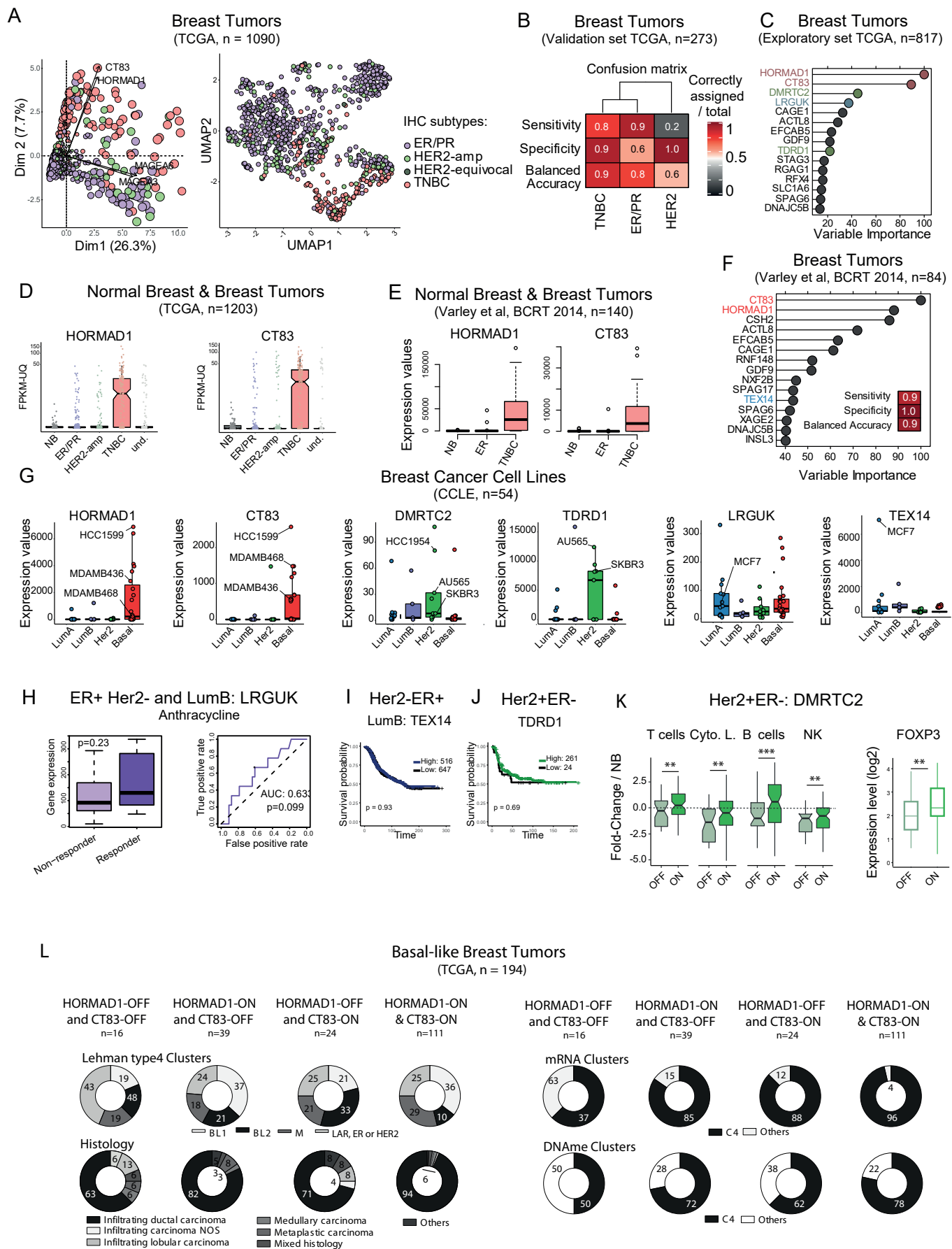


Figure S2: Further associations between expression of the 6 discriminating Cancer/Testis genes, tumor properties and survival

A. Multidimensional analysis of TCGA breast tumor and healthy samples based on expression of the 139 selected C/T genes. Each dot represents a sample, the color code corresponds to immunohistochemistry (IHC) classification (based on ER/PR/HER2 expression). Left: Principal Component Analysis, dot sizes are proportional to the quality of representation in PC1/PC2 space. The C/T genes best correlated to PC1/PC2 are represented. Right: Uniform Manifold Approximation and Projection.

B. Confusion matrix for breast tumor samples in the validation cohort (randomly selected 25% samples from the TCGA breast tumors) of the IHC tumor subtypes prediction obtained with the best Random Forest model. This model was established after a 500-tree training on the discovery cohort (the remaining 75%), based on the expression level of the 139 C/T genes.

C. Top 15 most important variables in the best Random Forest model for IHC tumor subtype prediction.

D. Expression levels for the 2 basal-specific C/T genes in the TCGA breast cohort, according to IHC tumor subtype.

E. Expression levels for the 2 basal-specific C/T genes in the Varley et al. breast cohort, according to IHC tumor subtype.

F. Top 15 most important variables in the best Random Forest model for IHC tumor subtype prediction in the Varley et al. cohort.

G. Expression levels for six subtype-specific C/T genes in breast cancer cell lines from the Cancer Cell Line Encyclopedia, according to PAM50 tumor subtype. Some commonly used cell lines are highlighted.

H. Left: Expression value for the luminal-specific C/T gene LRGUK in luminal B tumors, according to the clinical evaluation of tumor response to chemotherapy. Right: ROC curve evaluating the potential of LRGUK as a predictive biomarker of anthracycline chemotherapy response of ER+ Her2- Luminal B tumors.

I. Relapse-free survival curve for Her2-ER+ Luminal B breast cancer patients, as a function of TEX14 expression.

J. Same as in I for TDRD1 expression in Her2+ER- tumors.

K. Immune infiltration of Her2-positive breast tumors that express (ON) or do not express (OFF) DMRTC2, inferred from whole tumor RNA-seq data using MCPcounter. Fold-Changes were computed against Normal Breast (NB). Right: Expression level of the immune suppressive factor FOXP3 in the same tumors. P-value < 0.01: ** ; P-value < 0.001: ***

L. Anatomopathological classification (left) and IntClust classification (right) of basal-like breast tumors from TCGA, according to their expression status of HORMAD1 and CT83 genes.

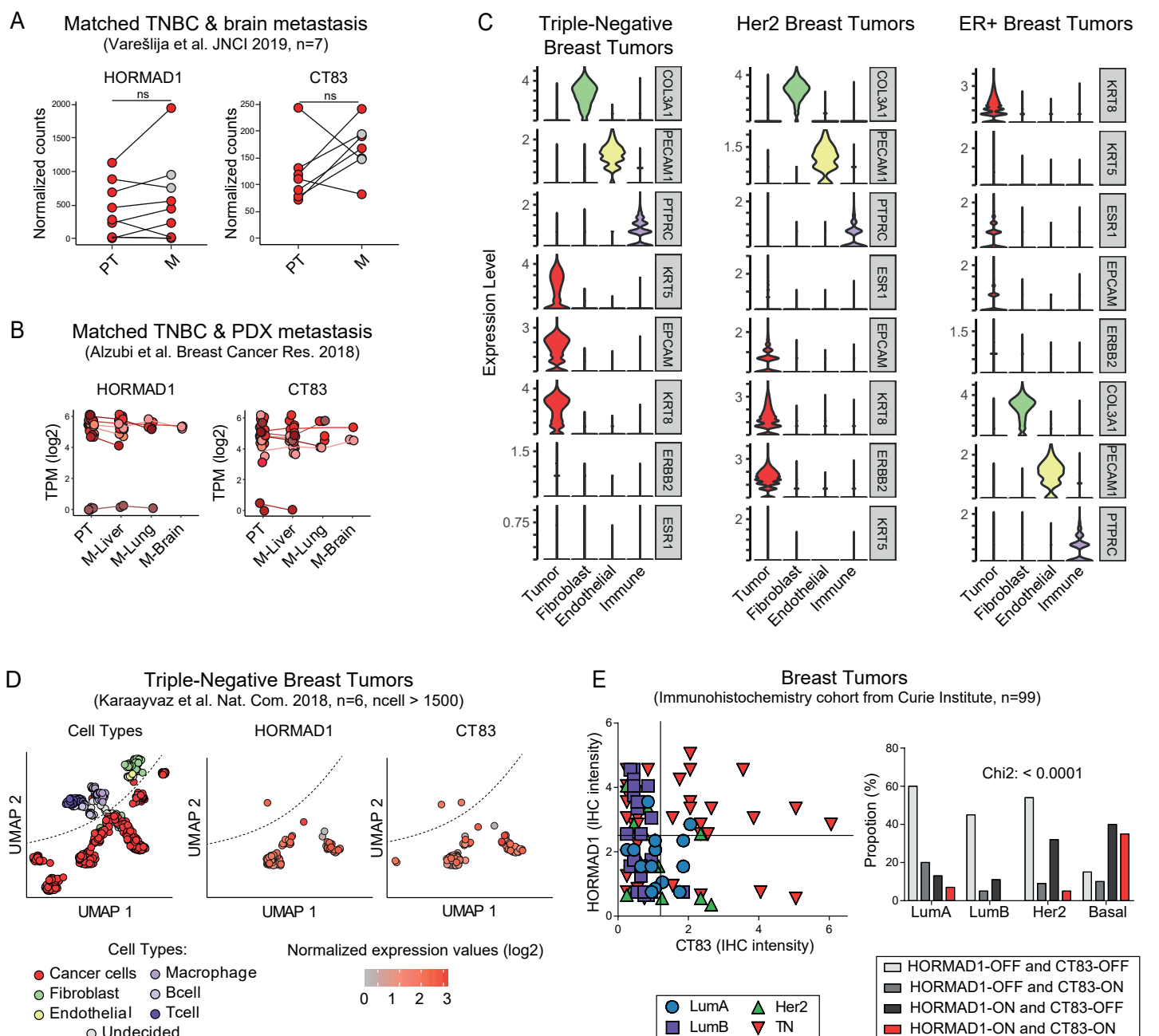


Figure S3: Further characterization of HORMAD and CT83 expression during tumor progression; validation of scRNAseq cluster identities.

A. Normalized expression levels of HORMAD1 and CT83 in matched primary tumors (PT) and metastasis (M).

B. Normalized expression levels of HORMAD1 and CT83 in matched primary tumors (PT) and multiple metastasis (M-) according to the seeding site.

C. ViolinPlot showing the expression of marker genes for each cell type cluster, in scRNA-seq data from different subtypes of breast tumors.

D. UMAP representation of a scRNA-seq study on 6 triple-negative breast tumors (GSE118390). Each dot is either a tumor cell or a cell from the tumor microenvironment. Epithelial clusters of tumor cells are highlighted, HORMAD1 and CT83 expression is shown.

E. Quantification of the IHC signal in breast tumor samples. Subtypes are indicated by the color/ shape code; proportion of positive tumors for HORMAD1 and CT83 according to tumor subtypes are summarized as barplot (right).

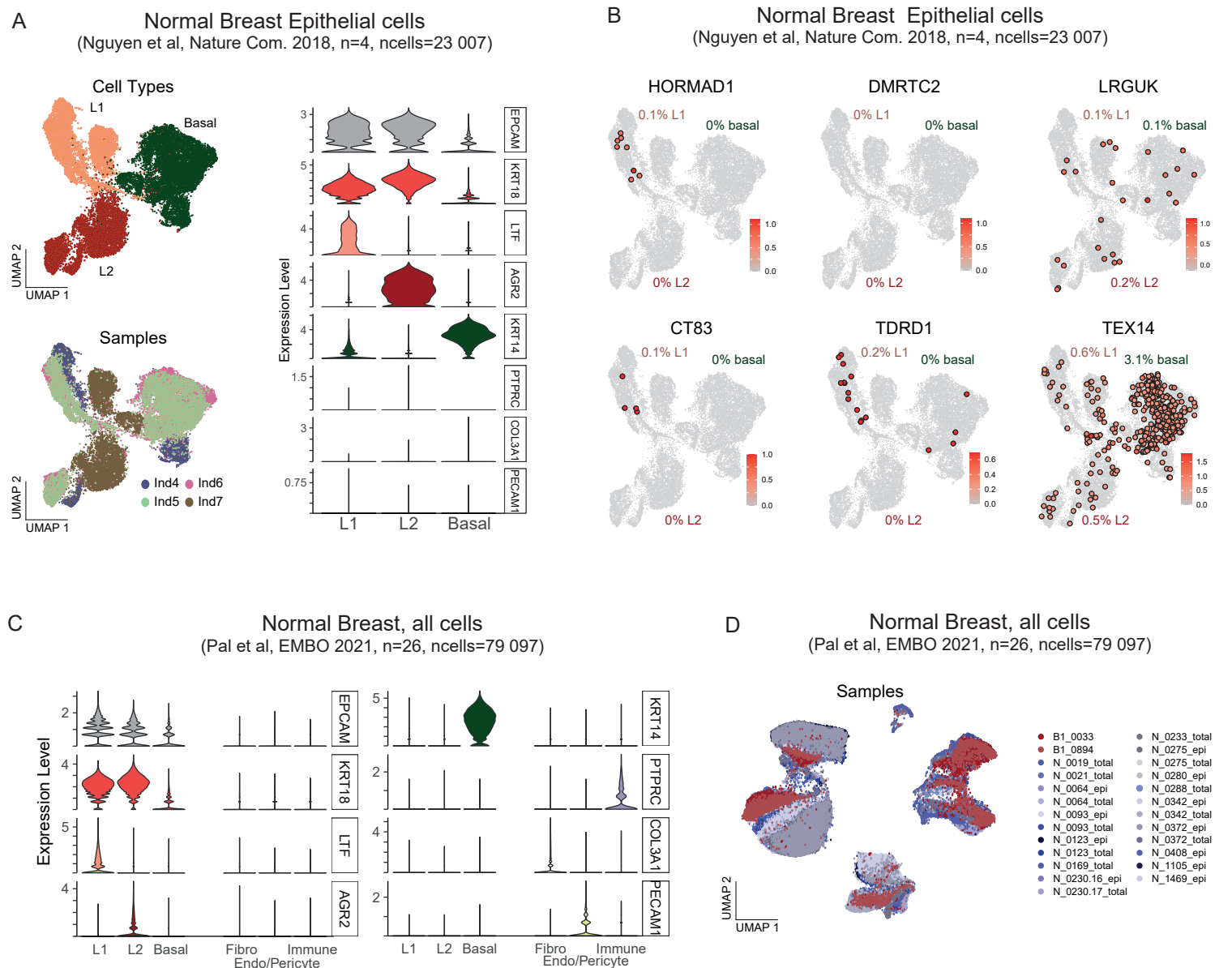


Figure S4: Markers of cell identity in scRNAseq from healthy mammary glands

A. UMAP representation of a scRNA-seq study on 4 healthy mammary glands (GSE113197) after an enrichment of epithelial cells by FACS. Top panel: Cell types. Bottom panel: sample of origin, Ind=Individual. Right: ViolinPlot showing expression values of marker genes for each identified cluster.

B. Same as in A, showing normalized expression of breast-cancer-specific C/T genes. Positive cells are emphasized, the percentage of positive cells in each cluster is indicated.

C. ViolinPlot showing expression values of marker genes for each identified cluster from scRNA-seq of healthy breast samples (GSE161529).

D. UMAP representation of a scRNA-seq study from 26 healthy mammary glands (GSE161529) showing samples of origin. Two prophylactic samples from BRCA1-mutated patients are highlighted in red.

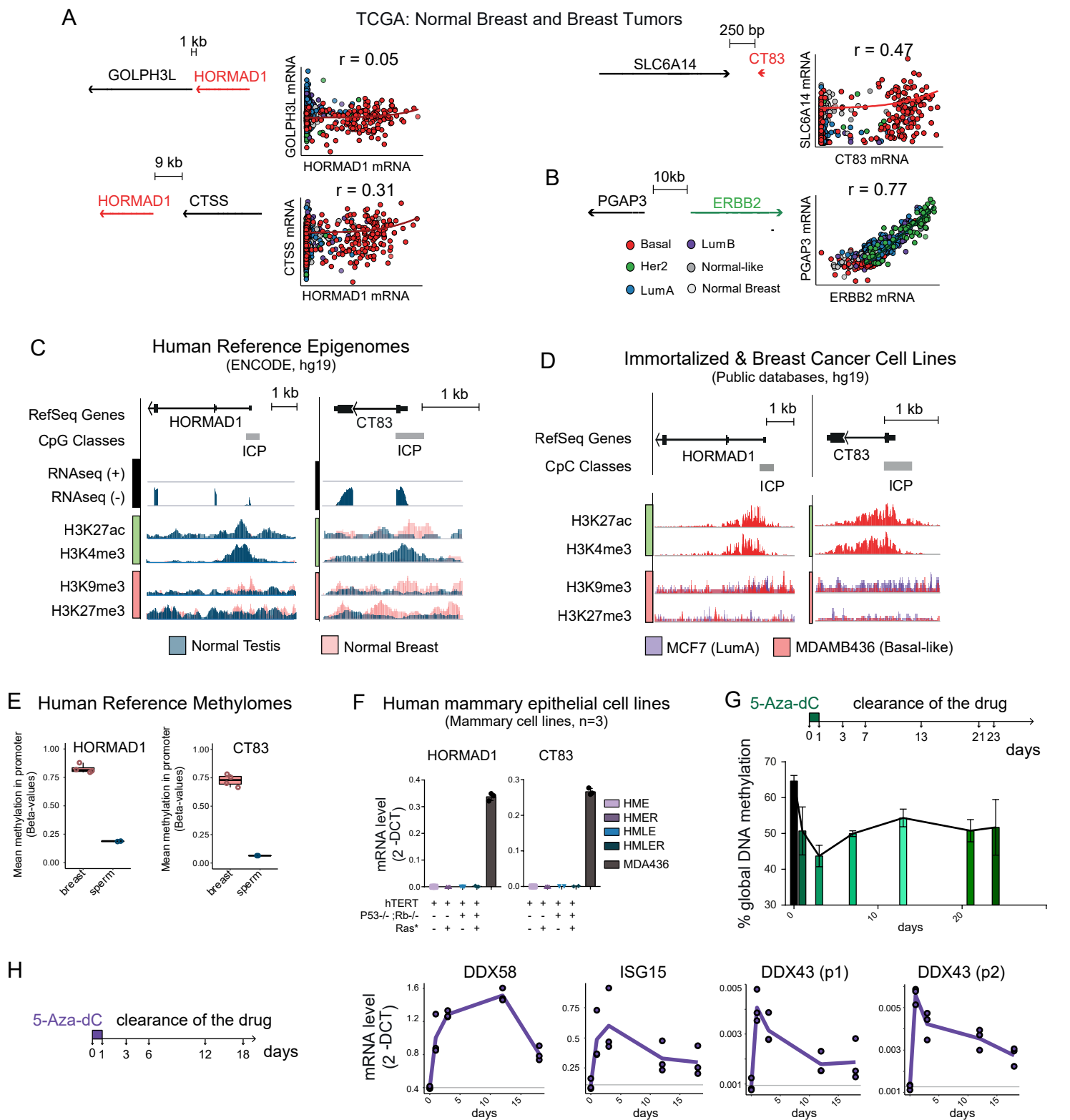


Figure S5: Genetic and epigenetic landscapes of the HORMAD1 and CT83 genes

A. Biplot showing the correlation between HORMAD1 or CT83 expression and expression of the closest neighboring genes. Genomic distances are depicted. The color code corresponds to PAM50 tumor subtypes.

B. Biplot showing the correlation between ERBB2 gene expression (coding for the frequently amplified gene Her2) and the closest neighboring gene, which is frequently co-amplified. Genomic distances are depicted. The color code corresponds to PAM50 tumor subtypes.

C. IGV representation of transcriptomic and histone modification landscapes at HORMAD1 and CT83 loci in healthy Testis and Breast samples (ENCODE).

D. IGV representation of histone modifications landscapes at HORMAD1 and CT83 loci in breast cell lines: MCF7 cells do not express HORMAD1 or CT83, whereas MDA-MB436 cells express both.

E. Mean CpG methylation of CpG probes found in HORMAD1 (left) and CT83 (right) promoters in healthy human breast and sperm samples.

F. RT-qPCR analysis of HORMAD and CT83 expression in human mammary epithelial cell lines after various genetic events and in a basal-like breast cancer cell line (MDA-MB436).

G. Global DNA methylation assessed by LUMA in HMLE cells after a 24-hour treatment with 1uM of 5-aza-dC followed by a 3-week recovery in absence of the drug.

H. mRNA levels of the indicated C/T genes, after a treatment with 5-aza-dC followed by a recovery of 18 days.

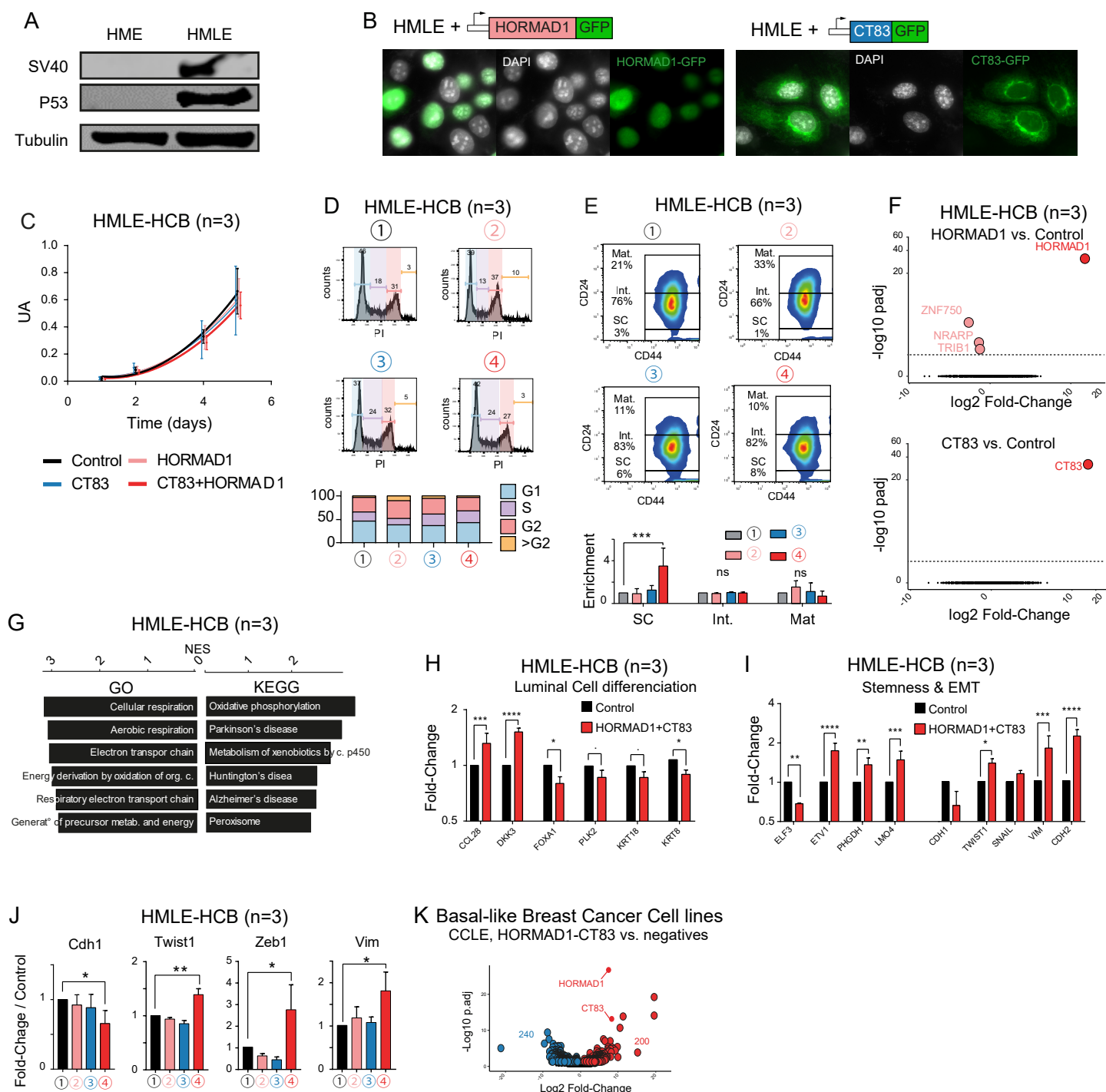


Figure S6: Characterization of HORMAD1 and CT83-expressing cultured breast cells

A. Western Blot showing the expression of transgenic constructs in the HME and HMLE cell lines. B. Immunofluorescence images of HORMAD1-GFP or CT83 in lentivirus-infected cells. Cells were sorted 48h after the infection based on the detection of the GFP signal, and immediately seeded for staining. C. Growth curves of HMLE-derived cell lines, measured by indirect detection of metabolically active cells using the MTS assay (left, mean $\pm$ SD, n=3 independent experiments). D. Cell-cycle analysis of HMLE-derived cell lines. Data are represented as mean $\pm$ SD (n=3 independent experiments); a representative FACS plot is shown. E. Flow cytometry analyses showing the percentage of Mature cells (mat, CD44+CD24^{high}), Intermediate cells (int, CD44+CD24^{low}) and Stem cells (SC, CD44+CD24^{low}) in HMLE-derived cell lines. Data are represented as mean $\pm$ SD (n=3 independent experiments); a representative FACS plot is shown. F. VolcanoPlot summarizing the differential gene expression analysis in HORMAD1- or CT83- overexpressing HMLE cells versus control HMLE cells. G. Barplot displaying the top 6 pathways differentially activated in HORMAD1 and CT83-positive HMLE-derived cells compared to the control condition, from MSigDB c5_GO and c2_KEGG annotations. X-axis corresponds to $-\log_{10}$ adjusted q-values. H. RT-qPCR analysis of genes related to luminal differentiation in HMLE overexpressing HORMAD1 and CT83 versus control. Data are represented as mean $\pm$ SD (n=3 independent experiments). I. RT-qPCR analysis of genes related to stemness and EMT in HMLE overexpressing HORMAD1 and CT83 versus control. Data are represented as mean $\pm$ SD (n=3 independent experiments). J. RT-qPCR analysis of markers of EMT and regulators of this pathway in HMLE overexpressing HORMAD1 and CT83 versus control. Data are represented as mean $\pm$ SD (n=3 independent experiments). K. Volcano plot of the distribution of differentially expressed genes (p-value adjusted < 0.01) in basal-like breast cancer cell lines (CCLE) according to HORMAD1 and CT83 expression status.