

Figure S1

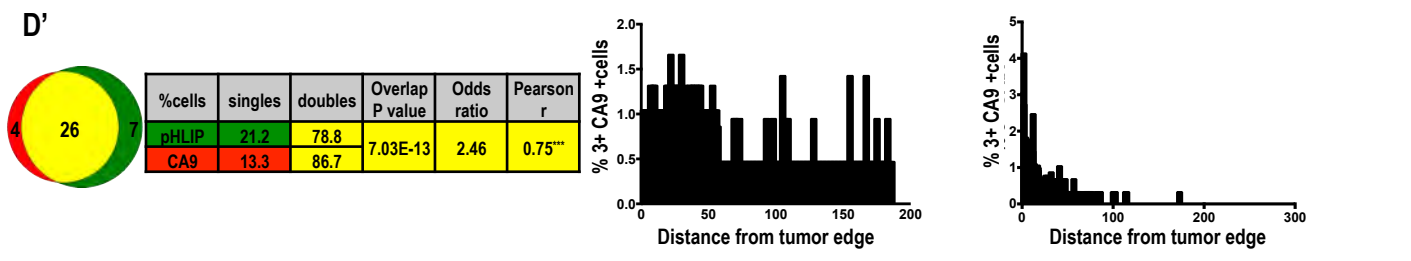
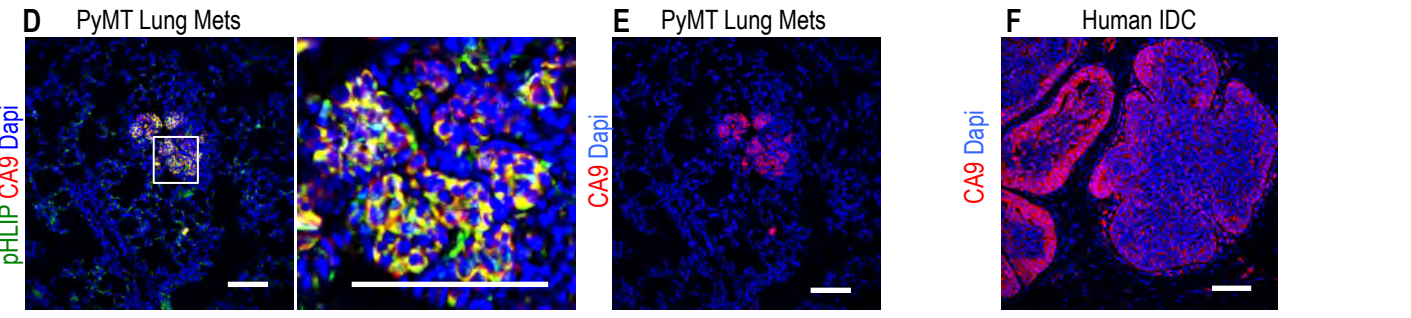
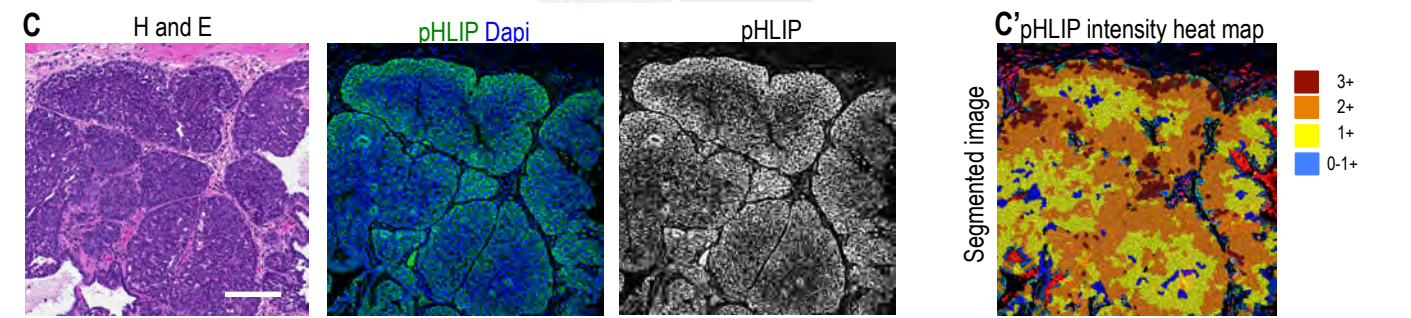
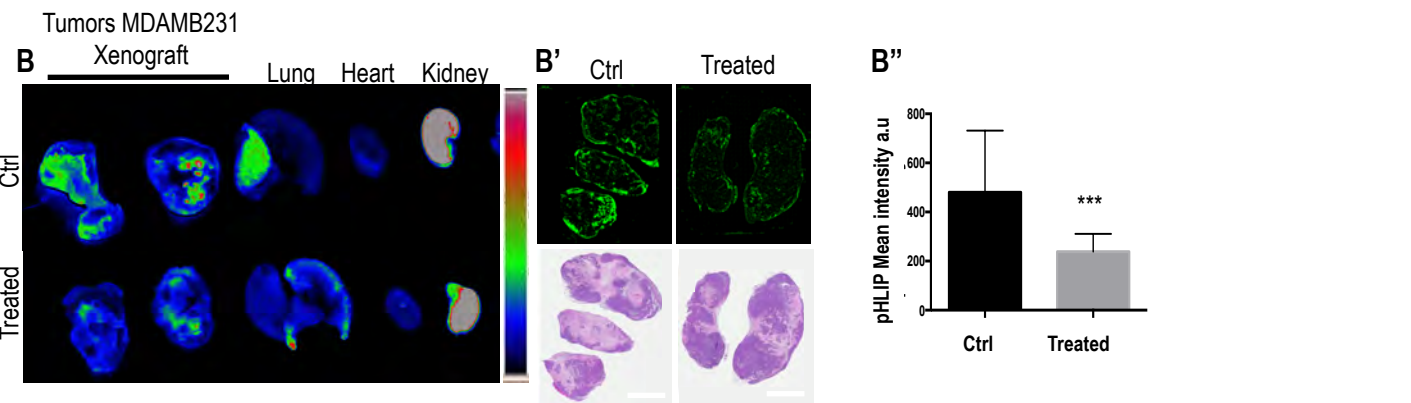
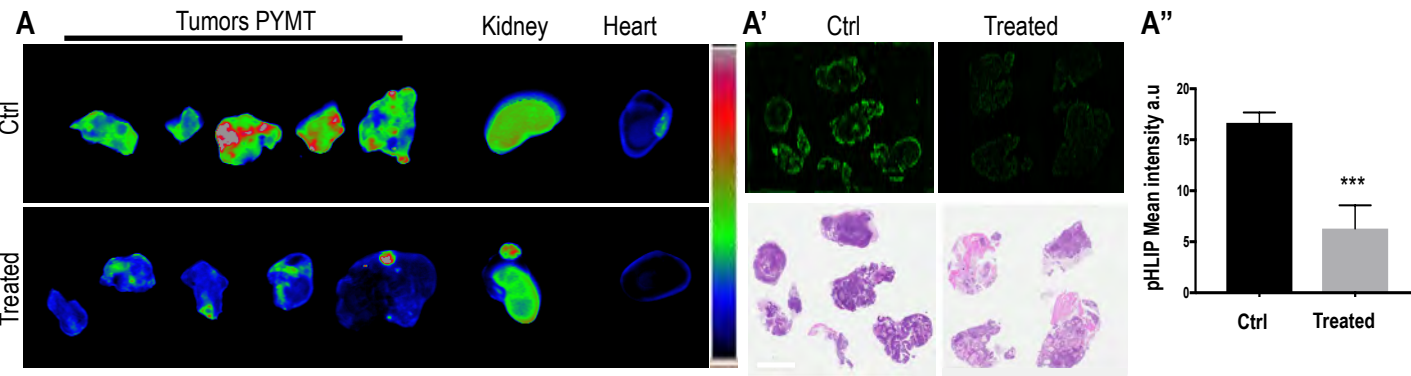


Figure S1: A- B) pHLIP localization to the acidic tissues *in vivo* is pH-dependent. A) Intensity heat map for Cy7-pHLIP localization in tumor and organs collected from MMTV-PyMT subject to tumor neutralization experiment. Cy7-pHLIP signal intensity is lower in tumors collected from the bicarbonate water-treated group that have neutralized tumor microenvironment, n=7 mice per group. A'-A'') pHLIP mean intensity per area measured from tumor cross sections collected from mice in the bicarbonate treated and control groups shows significant reduction. B) Same as A, but for NSCID Xenograft MDAMB231 mice. n=4 per group. C) pHLIP membrane localization in cross sections of the tumor, H and E serial section shows tumor and adjacent stroma. C') Cell-based segmentation and intensity profiling shows degree of positivity, cells with highest pHLIP retention at the membrane (3+ population) are located closer to the tumor-stroma interface. D-D') pHLIP membrane retention and CA9 expression correlate in metastatic lesions from mice. E) Spatial distribution of CA9-positive cells in lung metastases from PyMT mice indicates lower extracellular pH closer to the rim of the metastatic lesion, a pattern similar to the primary tumor. F) CA9 expression is highest at the tumor-stroma interface in human Invasive ductal carcinoma. CA9 expression, similar to pHLIP labeling, marks cells within the areas of the tumor that are acidic; these areas are frequently at the tumor-stroma interface. Scale bars C-F=100um, A-B=4mm.

Figure S2

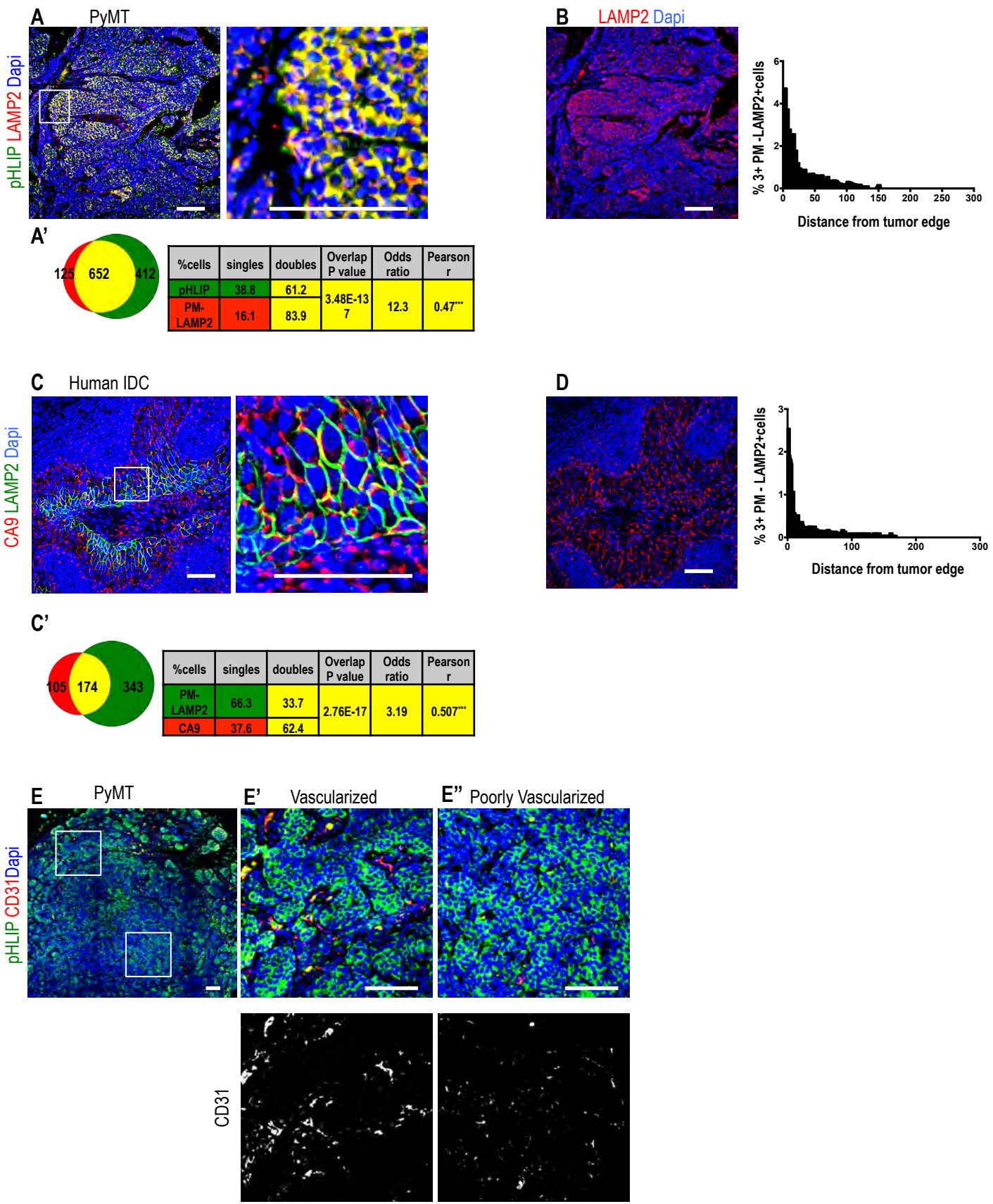


Figure S2: A) PM-LAMP2 overlaps significantly with cells with retained pHLIP at their membrane. A') Cell based co-localization and overlap analysis number of single and double positivity and percentages summarized in venn diagram and table respectively. B-D) The spatial distribution of PM-LAMP2 in mouse B and human D. Most cells with highest membrane positivity for LAMP2 (3+) localize closer to tumor-stroma interface. C) Most CA9 positive cells in human invasive ductal carcinoma tumor sections overlap plasma membrane LAMP2. C') cell based co-localization analysis indicate significant overlap between CA9 expression and LAMP2 membrane signal. E) Immunofluorescence labeling for CD31 relative to pHLIP membrane pHLIP membrane localization is not restricted to vascularized areas localization. pHLIP peptide is well diffused in poorly vascularized areas. E') Highly vascularized area of the tumor as indicated with CD31 labeling, cells located distant from the vasculature have retained pHLIP at the membrane. E'') Poorly vascularized –perfused area of the tumor, pHLIP retained in cells in areas with little or no proximity to CD31 labeled endothelial cells. Images from at least 4 mice (5 areas each) were subjected to cell-based segmentation and intensity profiling in all experiments. Pearson p values were quantified using two-tailed Student's t test. (Statistical analysis Student's t test: ***= < 0.001 , *= < 0.05). Scale bars=100um.

Figure S3

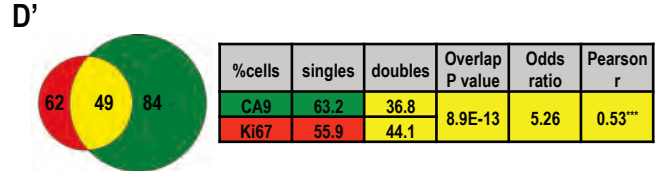
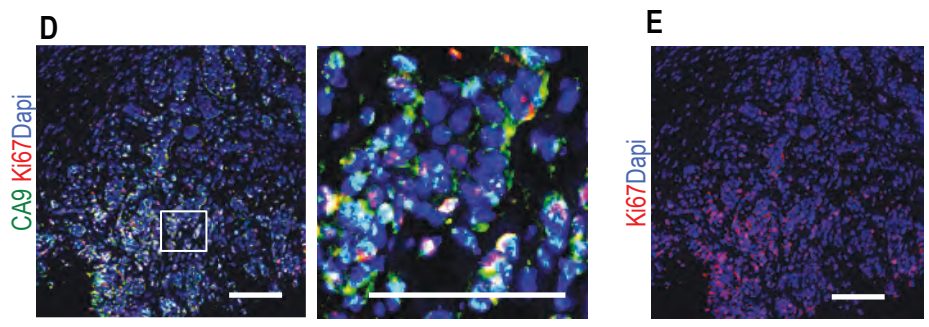
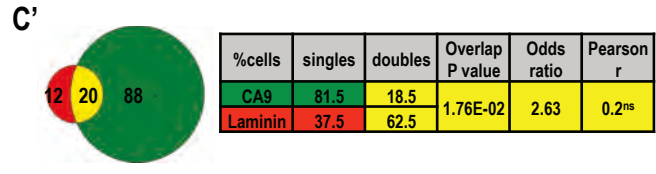
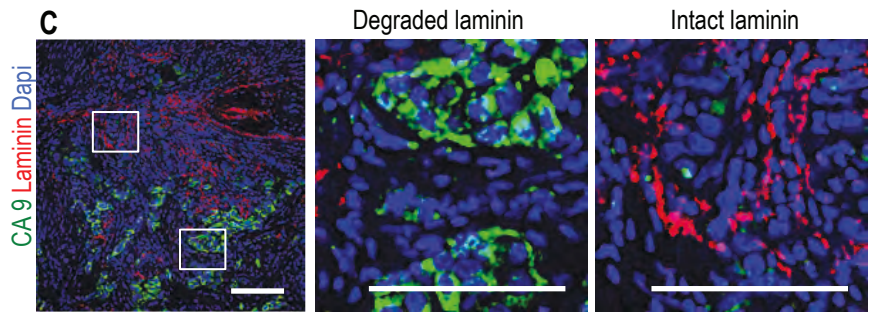
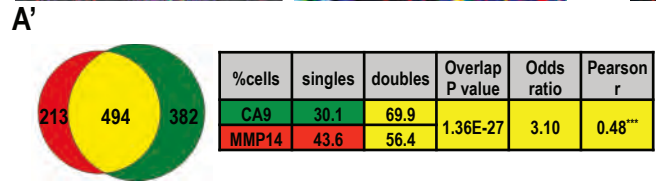
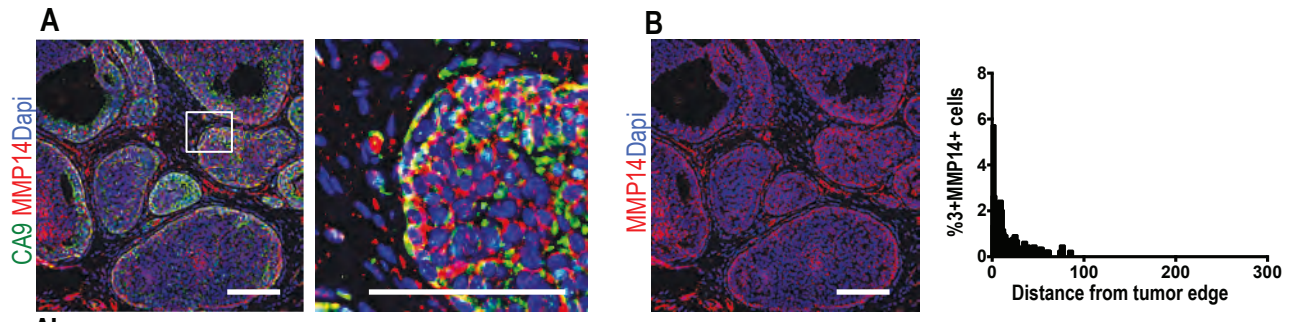
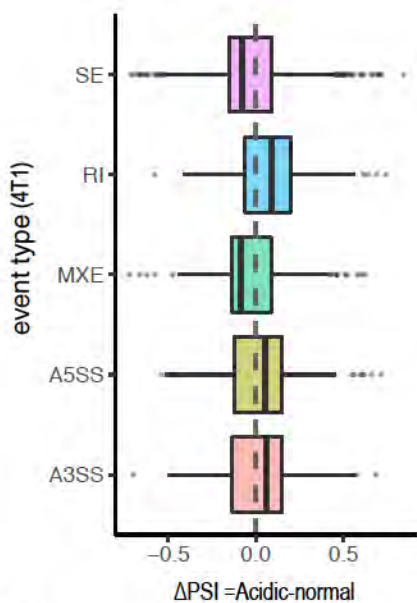
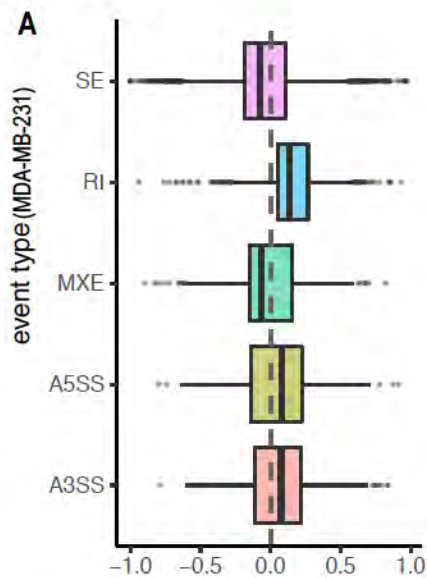


Figure S3: Acidic front is invasive and proliferative in human IDC. A-A') The expression of MMP14 (antibody specific to active domain) relative to CA9 in human IDC. A') Cell based overlap and Pearson correlation coefficient for MMP14 and CA9 intensity per cell indicates significant correlation. Most MMP14 expressing cells overlap CA9 positive cells. B) MMP14 distribution relative to tumor dimension, cells expressing highest levels of MMP14 frequently located at the tumor-stroma interface. C) Most CA9 positive and laminin positive cells are frequently excluded. Laminin is degraded on cells that are positive for CA9 however cellular areas that are negative for CA9 have maintained linear laminin structure (magnified insets). C') Overlap and correlation analysis, Pearson correlation coefficient indicates absence of correlation between CA9 expression and Laminin. Cell based segmentation analysis shows only a small degree of overlap. D-D') Human tumor staining shows Ki67 and CA9 partially overlap. 44% of Ki67 positive cells also express CA9. E) Ki67 positive cells were enriched at the tumor-stroma interface a pattern reminiscent of low pH areas marked by CA9. Image segmentation and analysis for human samples was performed on 3 tumors from 3 different patients and at least 5 random field each. Pearson p values were quantified using two-tailed Student's t test. (Statistical analysis Student's t test: *** $p < 0.001$, * $p < 0.05$)

Figure S4



B

Gene Name	Exon Usage	Functional consequence	Protein class	Species
CD44	Exon19 - short tail	Drug resistance	Transmembrane glycoprotein	Human
ENAH-Mena	Exon 4-INV	Invasion-metastasis	Actin binding protein	Mouse-Human
DOCK7	Exon-23	Not known	Guanyl nucleotide exchange factor	Mouse-Human
DLG1	Exon-6	Not known	Trans membrane adaptor protein	Mouse-Human

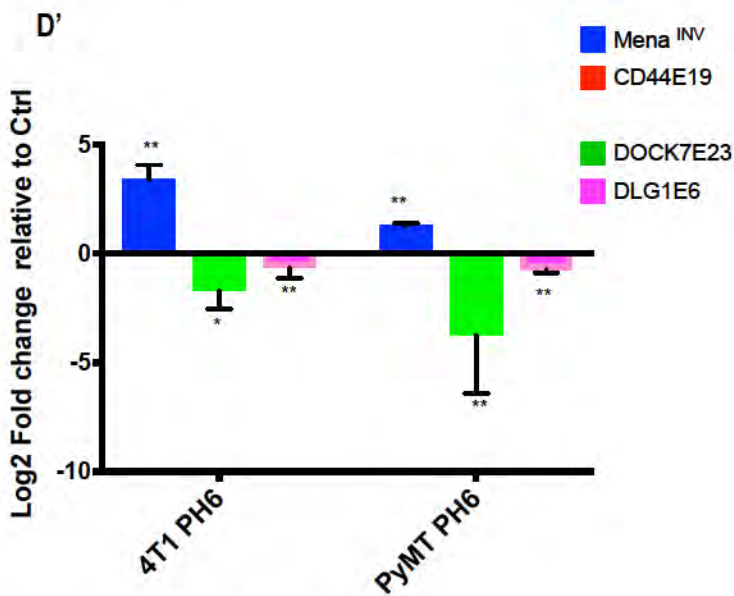
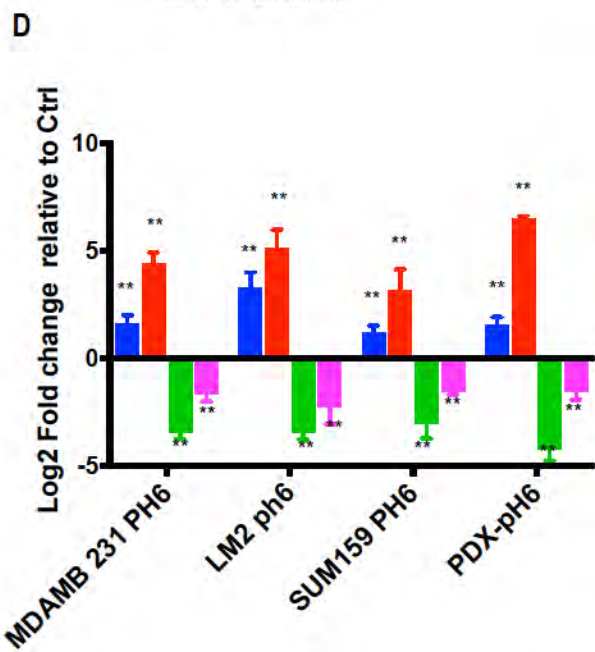
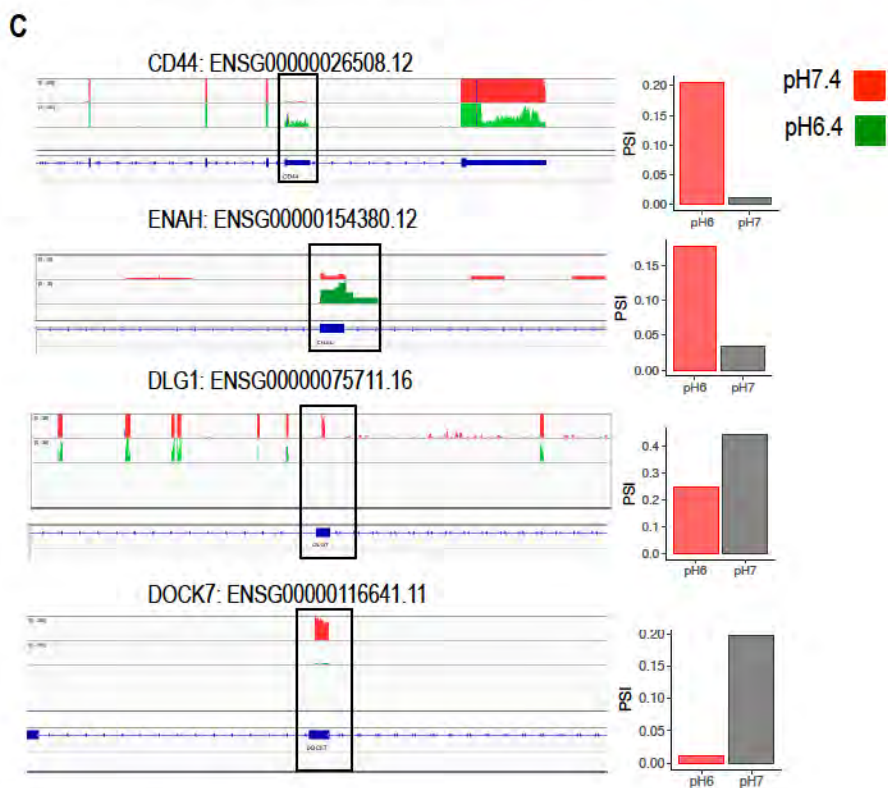
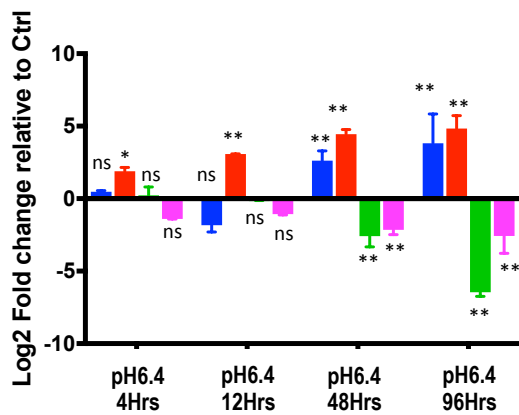


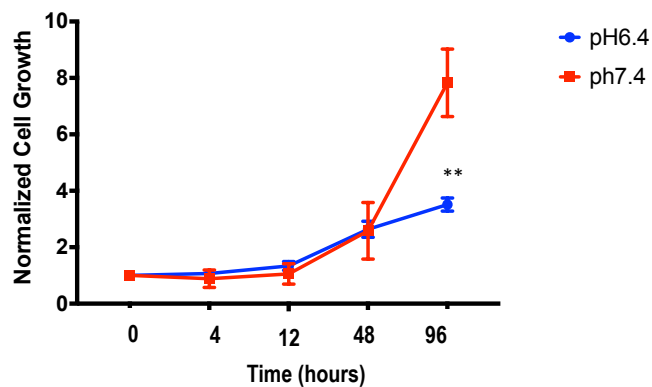
Figure S4: Validation of alternatively spliced candidates associated with cytoskeletal regulation. A) Comparison of the distributions of splicing changes by events type. Δ PSI values (PSI Acidic-PSI normal) for different types of splicing events in normal versus low pH conditions in human and mouse. Retained introns (RI) show highest magnitude of changes in low pH conditions. B) Table summarizes functional relevance of the included and excluded exons for selected genes; CD44 among the candidates is human specific. C) Integrative genomic viewer (IGV) view of the sequencing read counts over exon of interest for selected genes in low pH and control conditions. Bar graph indicates PSI values (percent spliced) in low pH (pH6.4) and normal pH (pH7.4). D) qPCR validation of the candidate splicing events in a panel of human and mouse breast mammary carcinoma cell lines. Log_2 fold change of inclusion/exclusion ratio for the tested exons in a panel of human carcinoma cell lines. D') Log_2 fold change of candidate pH-induced splicing events in mouse carcinoma cell lines. N=4 experiments 3 technical replicates, Students' t-test $p < 0.05$ *, $p < 0.01$ **.

Figure S5

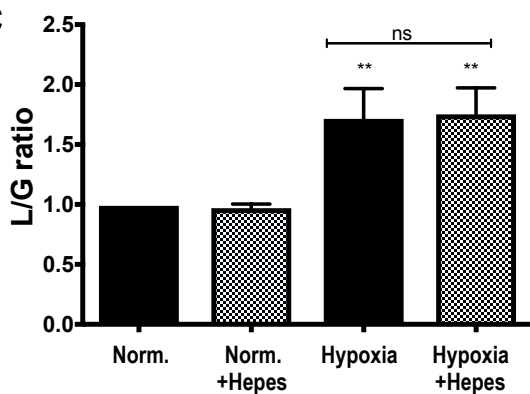
A



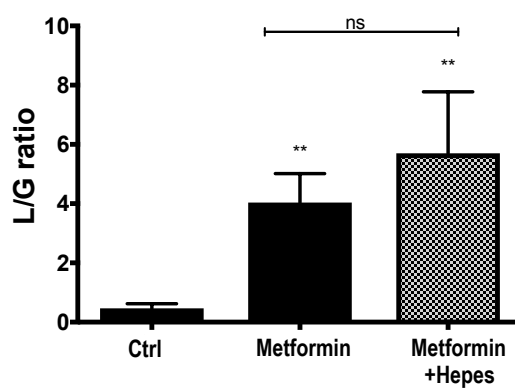
B



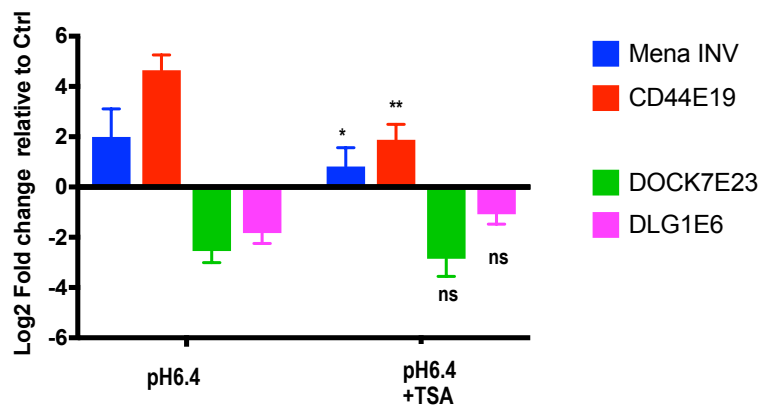
C



D



E



F

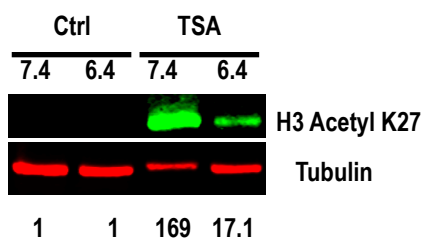


Figure S5: A) The expression of selected candidate splicing events relative to time cultured in low pH. CD44E19 is significantly included at 4hrs and increases in a time dependent manner whereas, Mena^{INV} inclusion and DOCK7 E23 , DLG1E6 exclusions only occurs at 48 hour onwards. Buffering of medium by HEPES does not alter the lactate to glucose ratio. B) Cell growth measured over time in pH7.4 and pH6.4 cell number in each time point is normalized to time 0. C- D) Medium from cells under (C) long-term hypoxia (10days) or (D) metformin addition has significantly higher lactate to glucose ratio. Neutralization of the pH of the medium by addition of HEPES does not affect this ratio. E-F) The inclusion of Mena^{INV} and CD44E19 is sensitive to histone acetylation levels. Addition of 500nm Trichostatin A (TSA) to culture media under low pH significantly reduced the inclusion of Mena^{INV} and CD44E19 exons but does not affect DLG1E6 and DOCK7E23 skipping. F) Western blot indicating increases in H3 acetylation in response to TSA under pH7.4 and pH6.4 conditions. Values indicate normalized fold change in acetylated H3 signal.

Figure S6

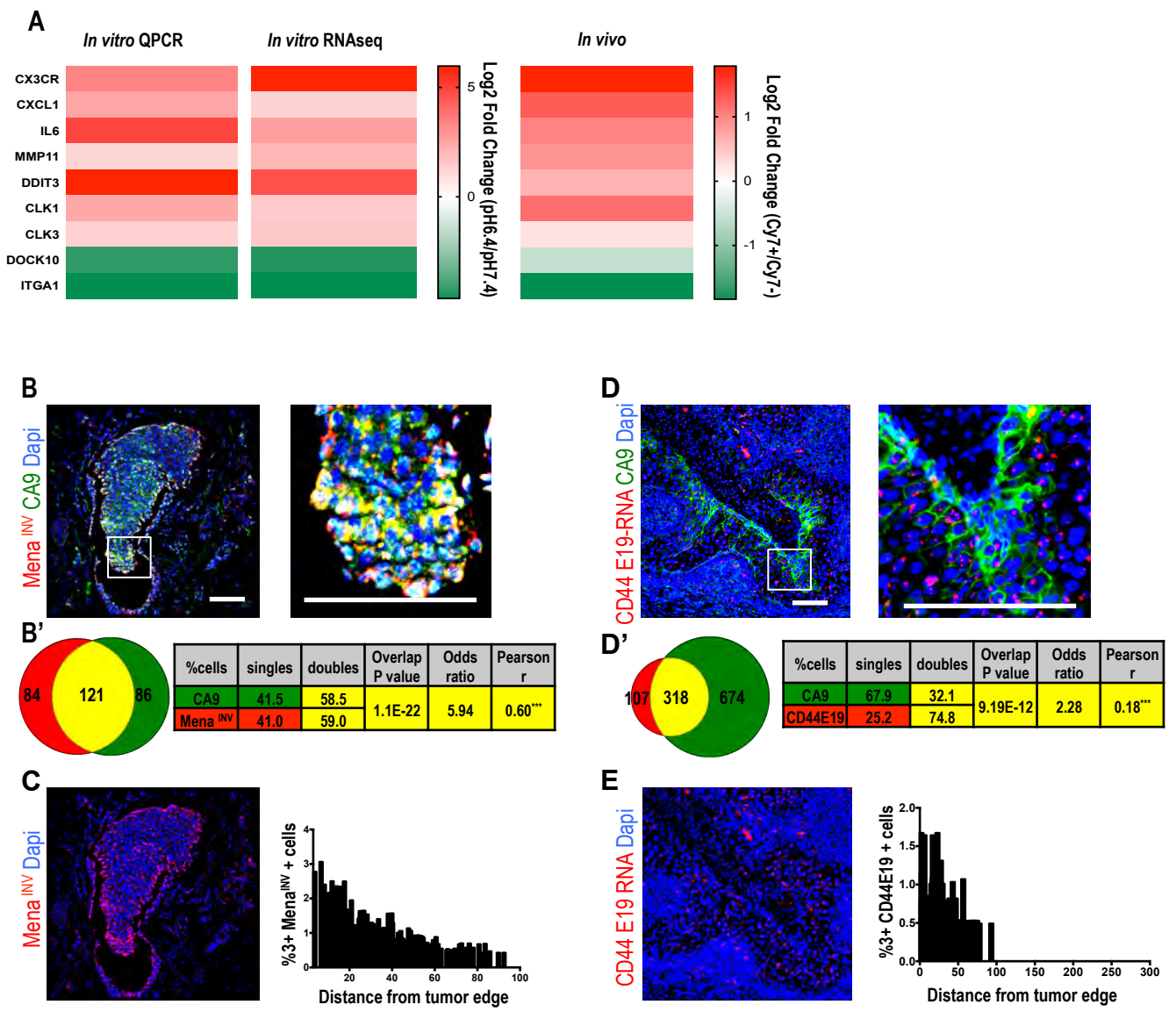


Figure S6: Validation of pH-responsive transcriptomic changes *in vivo*. A) Heat map comparing \log_2 fold changes of candidate genes from pH signature from the RNAseq dataset that were validated by qPCR *in vitro* in 4T1 cells and *in vivo* in cells collected by FACS sorting approach. B) CA9 expression overlaps the expression of the Mena^{INV} isoform in human invasive ductal carcinoma (IDC). B') Cell based co-localization and Pearson correlation coefficient indicates high degree of overlap. C) Mena^{INV} distribution in human tumor samples relative to tumor dimensions shows the highest expressing cells located proximal to the tumor-stroma interface. D) RNAscope exon-specific mRNA detection relative to CA9 expression in human IDC tumors. D') Cell based co-localization and Pearson correlation indicates significant overlap. Most CD44E19 RNA-positive cells are also expressing CA9 protein. E) CD44E19 expressing cell distribution in human IDC tissue samples. Highest expressing cells are located closer to the tumor-stroma interface.

Supplementary Methods:

Illumina sequencing:

RNA samples from mouse and human cell lines in duplicate were subject to RNA-sequencing. KAPA Hyper Prep Kit (Kapa Biosystems) was used for library preparation. Short read sequencing was performed on NextSeq500 platform to a depth of more than 40 million reads per sample using 75bp paired-end method and Nextseq150nt kit. Raw and processed RNA-seq data: Gene expression omnibus, NCBI, accession number: GSE119646.

Bioinformatics analysis:

We confirmed RNA-seq data quality by running an in-house computational pipeline that estimated sequencing error rate, read complexity, contamination, read alignment percentages to exons, UTRs, introns, and intergenic regions, rRNA contamination, and sense to antisense read ratios. Hierarchical clustering and heatmap creation were carried out using TIBCO Spotfire 7.6.1.17 based on $\log_2(\text{fpkm}+1)$ values of the expressed coding genes. We mapped sequencing reads to either the GRCh38/hg38 reference genome for human samples or the GRCm38/mm10 reference genome for mouse samples using STAR version 2.5.3a(1). Gene-level read counts were calculated using RSEM version 1.3.0 in a strand specific manner and exon level counting for alternative splicing analyses was performed using HTSeq packaged within DEXSeq (2). We performed differential gene expression analysis of sample read counts between the low and normal pH conditions with DESeq2 version 1.10.1(3). For alternative splicing analysis, we used the DEXSeq package (version 1.22.0) (4). Filtering criteria to define significance was set to base mean greater than 10, $\log_2\text{FoldChange}$ greater than 0.5 to identify up-regulation and less than -0.5 to identify down-regulation, and padj smaller than 0.05. Overlapping between human and mice samples was based on gene symbols after matching mouse genes symbols to human gene symbols. The significance of overlap was tested using hypergeometric statistics.

Event-centric splicing analyses were performed using rMATS. Events were filtered for having ≥ 20 junction-spanning reads supporting inclusion and/or exclusion. Exons with $\text{abs}(\Delta\text{PSI}) \geq 0.05$ and $\text{FDR} < 0.05$ were considered significantly changing and exons with $\text{abs}(\Delta\text{PSI}) \leq 0.02$ and $\text{FDR} > 0.5$ were used as control exons.

Cell culture

MDA-MB231 cell line was purchased from ATCC in 2016, LM2, the highly metastatic derivative of MDA-MB231, was obtained from J. Massague (5) in 2016 and cultured in DMEM and 10% FBS, 1% penicillin-streptomycin-glutamine. SUM159 were obtained from Joan Brugge's lab in 2016 at Harvard Medical

School and cultured according to ATCC guidelines. The patient-derived xenograft (PDX) cell line- EL12-58 was established from a previously untreated triple-negative breast cancer sample; the cells were a gift from Dr. Daniel G Stover in 2016. The culture was established using a protocol adapted/optimized from (6). (In short: 2/3 DMEM, 1/3 HAMS F12, 0.5ug Hydrocortisone/0.5ug EGF, 5mg Insulin, Cholera Toxin 0.5uM, Fungizone, 5uM ROCK inhibitor Y-27632). 4T1 cell lines were purchased from ATCC in 2015 and cultured in DMEM, 5% FBS+5% FCS+1% non-essential amino acids, 1%penicillin streptomycin glutamine. John Lamar at MIT generated PyMT cell lines. The cells were derived from MMTV-PyMT tumors and metastatic lesions, and selected to enrich for highly metastatic variants by repeated *in vivo* sorting for metastatic potential. These were cultured in 50% DMEM and 50% F12, 2% calf serum, 0.01% BSA, 10 ng/ml EGF, 10ug/ml insulin ,1%penn/strep ,1% glutamine.

Growth rate

Cell count was determined over time and up to 96 hours for MDAMB 231 cells expressing nuclear RFP that are cultured in pH7.4 and pH6.4 in IncuCyte®. Growth rate was measured using normalized cell number (cell number at t=t1/ cell number at t=t0).

References:

1. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics*. 2013;29:15–21.
2. Li B, Dewey CN. RSEM: Accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics*. 2011;12.
3. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol*. 2014;15:1–21.
4. Anders S, Reyes A, Huber W. Detecting differential usage of exons from RNA-seq data. *Genome Res*. 2012;22:2008–17.
5. Minn AJ, Gupta GP, Siegel PM, Bos PD, Shu W, Giri DD, et al. Genes that mediate breast cancer metastasis to lung. *Nature*. Nature Publishing Group; 2005;436:518.
6. Palechor-Ceron N, Supryniewicz FA, Upadhyay G, Dakic A, Minas T, Simic V, et al. Radiation Induces Diffusible Feeder Cell Factor(s) That Cooperate with ROCK Inhibitor to Conditionally Reprogram and Immortalize Epithelial Cells. *Am J Pathol*. American Society for Investigative Pathology; 2013;183:1862–70.

Supplementary Tables:

Supplementary Table 1:

Table S1: List of all antibodies.

Antibody	Cat#	Company
Laminin	Ab11575	Abcam
MMP9	Ab38898	Abcam
MMP14 (active domain)	MAB3328	Millipore
FitC-Cyanine	NB-1207628	Novus Biologicals
Piminidazole	MAB1	Hpi, hypoxypobe
LDHA	NBP1-48336	Novus Biologicals
Ki 67 (SP6)	Ab16667	Abcam
CA9	Ab15086	Abcam
HDAC1	Ab7028	Abcam
Anti-Histone H3 (acetyl K27)	ab4729	Abcam
LAMP2	Ab203224	Abcam

Supplementary Table 2:

Table S2: q-PCR primer sequences.

Gene	Forward	Reverse	Species
CD44	TGAATATAACCTGCCGCTTTG	GCTTTCTCCATCTGGGCCAT	H
CD44-E19	GCAGTCAACAGTCGAAGAAGTTGAAG	CCATGTGTCAGTTCTAGCGAGG	H
Mena	GAA CAA AAA GAG GAC AAA GGT GA	TGC CAT TCA TTG TAT TTG TTC TTT	H
Mena	GGG CAG AAA GAT TCA AGA CC	GCG AAG ACA TTG GCA TCC	M
Mena INV	AGA GGA TGC CAA TGT CTT CG	TTA GTG CTG TCC TGG GTA GC	H
Mena INV	AGA GGA TGC CAA TGT CTT CG	TTA GTG CTG TCC TGC GTA GC	M
DOCK7	GTGGCAGCCGAAGTTAGGAAG	GCACTGTGGTGTGATGGGATA	H
Dock7 E23	GGGACTCCCACGTCACCAG	GACTGGGGTTGGATCCCCAT	H
DLG1	GCAGGAGGTACGGACAACC	ATTGACCCGCAATCTTCCATC	H
DLG1-E6	ATGAAGTGATAGGTCCAGAATTGGTTCAT	ACAGGGATCACAGGGACAGTG	H
DLG1	AGTGACGAAGTCGGAGTGATT	GTCAGGGATCTCCCCTTTATCT	M
DLG1 E6	ATGAGGTGCTAGGTCCAGAACT	GGACGACAGTACTCTCAGCAGG	M
DOCK7	CCATCTGGAAGCGCCTTTG	ACGATGATCTCTAGCGTGTCT	M
DOCK7E23	CTCCCACATCACCAGATGATGAAG	ACTTGGACTGGGGTTGGATCC	M
CLK1	ACAAGACATTATAGAGCACCGGA	GTGGTCCAAGAATCCTTTCCATC	M
CLK3	TGTCACAAACGCCGTACCAG	ACGATCTCATATCGCTCTTGGA	M
DDIT3	CTGGAAGCCTGGTATGAGGAT	CAGGGTCAAGAGTAGTGAAGGT	M
ITGa1	CCTTCCCTCGGATGTGAGTCA	AAGTTCTCCCCGTATGGTAAGA	M
DOCK10	TTTTGTGAGCGTGCCACTTTA	CTTCCGATTGTCTTGGTGTACC	M
CX3CR1	GAGTATGACGATTCTGCTGAGG	CAGACCGAACGTGAAGACGAG	M
IL6	TAGTCCTTCTACCCCAATTTCC	TTGGTCTTAGCCACTCCTTC	M
MMP11	CCGGAGAGTCACCGTCATC	GCAGGACTAGGGACCCAATG	M
CXCL1	CTGGGATTCACCTCAAGAACATC	CAGGGTCAAGGCAAGCCTC	M
GAPDH	CAT GTT CCA GTA TGA CTC CAC TC	GGC CTC ACC CCA TTT GAT GT	M
GAPDH	CAT GAG AAG TAT GAC AAC AGC CT	AGT CCT TCC ACG ATA CCA AAG T	H

Supplementary Table 3:

MDAMB231- SPMA motif hits results:

motif ID	RBPs	adjusted R squared	polynomial degree	Consistency Score (CS) p-value
783_7972035	ELAVL2	0.76	1	0.000001
M112_0.6	ELAVL1, ELAVL3	0.709	1	0.000001
M012_0.6	CPEB3, CPEB2	0.74	1	0.000001
M124_0.6	ELAVL3	0.733	1	0.000001
M031_0.6	ELAVL1, ELAVL3	0.749	1	0.000001
951_12324455	ZFP36	0.646	1	0.000001
M120_0.6	CPEB3	0.589	1	0.000006
M227_0.6	PTBP1, PTBP2, ROD1	0.673	1	0.00025
LC1	ELAVL1	0.771	1	0.000001
M075_0.6	TIA1	0.694	1	0.000002
784_7972035	ELAVL2	0.777	1	0.000014
M149_0.6	CPEB3, CPEB4	0.764	1	0.000002
M158_0.6	HNRNPCL1	0.775	1	0.000001
M025_0.6	HNRNPC	0.775	1	0.000001
M150_0.6	RALY	0.717	1	0.000001
M201_0.6	SF3B4	0.762	1	0.000001
782_8497264	ELAVL2	0.743	1	0.000021
M232_0.6	ELAVL1, ELAVL3	0.758	1	0.000001
M156_0.6	TIA1	0.763	1	0.000001
M127_0.6	ELAVL1, ELAVL3	0.703	1	0.000001
M077_0.6	U2AF2	0.73	1	0.000002
147_19457263	KHDRBS3	0.636	1	0.000017
M176_0.6	KHDRBS2	0.768	1	0.000001
M001_0.6	A1CF	0.475	1	0.002261
M229_0.6	CELF3	0.475	1	0.004264
M108_0.6	ELAVL1, ELAVL3	0.698	1	0.000001
M160_0.6	KHDRBS1	0.75	1	0.000001
M157_0.6	CELF3, BRUNOL5	0.395	1	0.105509
M079_0.6	CELF3	0.503	1	0.000272
M051_0.6	RBM41	0.378	1	0.0007
1171_19561594	PTBP1	0.649	1	0.001153

4t1- SPMA motif hits results:

motif ID	RBPs	adjusted R squared	polynomial degree	Consistency Score (CS) p-value
LC1	ELAVL1	0.827	1	0.000017
M031_0.6	ELAVL1, ELAVL3	0.73	1	0.000079
M176_0.6	KHDRBS2	0.662	1	0.000189
M112_0.6	ELAVL1, ELAVL3	0.714	1	0.002543
783_7972035	ELAVL2	0.692	1	0.005896
784_7972035	ELAVL2	0.631	1	0.008048
M236_0.6	HNRNPR, SYNCRIP	0	0	0.008048
M232_0.6	ELAVL1, ELAVL3	0.462	1	0.008048
782_8497264	ELAVL2	0.708	1	0.011619
M156_0.6	TIA1	0.446	1	0.016212
M075_0.6	TIA1	0.42	1	0.016212
M149_0.6	CPEB3, CPEB4	0.487	1	0.030619
M025_0.6	HNRNPC	0.467	1	0.017497
M158_0.6	HNRNPCL1	0.467	1	0.019996
M201_0.6	SF3B4	0.438	1	0.040825
M150_0.6	RALY	0.381	1	0.034993
M077_0.6	U2AF2	0.373	1	0.049727

Table S3: Spectrum motif analysis (SPMA) identified motif hits for MDA-MB-231 and 4t1 cell lines differential gene expression dataset sorted according to Log₂ fold change in expression. Linear relationship between fold change in transcript expression and enrichment of putative RBP binding sites, are measured by adjusted R² values and local consistency scores (CS) p-values.