A circle RNA regulatory axis promotes lung squamous metastasis via CDR1 regulation of Golgi trafficking

Emily B. Harrison, Alessandro Porrello, Brittany M. Bowman, Adam R. Belanger, Gabriella Yacovone, Salma H. Azam, Ian A. Windham, Subrata K. Ghosh, Menglin Wang, Nick Mckenzie, Trent A. Waugh, Amanda E. D. Van Swearingen, Stephanie M. Cohen, Devon G. Allen, Tyler J. Goodwin, Teresa Mascenik, James E. Bear, Sarah Cohen, Scott H. Randell, Pierre P. Massion, Michael B. Major, Leaf Huang, Chad V. Pecot

Supplementary methods:

Selection of TCGA LUSC samples. TCGA LUSC samples used for these computational analyses were selected according to the completeness of their genomics information and after determining their subtype. Such subtypes were assigned (1) through a specific and validated LUSC expression subtype predictor (1). Overall, we included the samples having numerical data available (downloaded on January 2014) as for messenger RNA (mRNA), human methylation (HuMet), copy number variation (CNV) and micro-RNAs (miRNAs) (348 total). In particular, gene expression (RNA-Seq V2 pipeline, level 3, RNA-Seq by Expectation Maximization (RSEM)-normalized (2)), micro-RNA expression values (see the paragraph 'TCGA LUSC miRNA data: preliminary analyses at the gene level') and clinical data of lung squamous cancer (LUSC) samples were downloaded from TCGA Data Portal and the Firehose Broad GDAC (Genome Data Analysis Centers) data hub. The selected set of samples allowed us to i) leverage some results of our previous publication on LUSC (1), ii) integrate mRNA and miRNA information (see the paragraph 'Analysis of gene sets'), always using samples having both measures, and iii) exclude samples whose annotation or characterization using genomics variables was deemed insufficient (1).

Analysis of gene sets. Accounting for the level of 'sparsity' of the full gene expression matrix and for the specific type of signal of TCGA RNA-Seq data, a number of gene expression rows were trimmed from the expression spreadsheet of TCGA LUSC patients before running a gene-set based analysis. All genes that had a median expression value strictly in the lowest 1/8 of the dataset were discarded. Then, we deleted genes whose samples having a signal (here defined as a value > 0.5) were < 60% of the total. These pre-processing steps prevent the inflation of positive enrichment results. We also excluded genes without as assigned name. In addition, since the expression value of CDR1 is used for categorizing the samples into two groups (see below), CDR1 was also excluded; finally, we only used one entry per named gene, picking the gene entry with the highest median value. Then, Gene Set Enrichment Analysis (GSEA) (3) was run using this trimmed matrix of gene expression (with 17,215 genes) after splitting (inside the categorical class (cls) file) the samples into two groups: 1) 'Low CDR1, High miR-671-5p' (LCDR1HMIR6715P, 102

samples) and 2) ‘High CDR1, Low miR-671-5p’ (LCDR1HMIR6715P, 102 samples). ‘High’ and ‘Low’ were based on the median value inside a set of 348 TCGA LUSC samples previously characterized, whose data were available both at the level of miRNAs and mRNAs and whose genomic subtype was also known (4). GSEA was run using the gene symbol identifiers and the number of sample label permutations was set at 1,000. The cutoff thresholds for gene set sizes were 15 and 500, respectively at the upper and lower end, and the ‘metric’ used for ranking the genes was the signal-to-noise (S2N). The gene matrix transposed (gmt) files that were used to define gene sets to be tested for enrichment were obtained from the Molecular Signatures Database (MSigDB) (http://software.broadinstitute.org/gsea/msigdb/). The two collections of gene sets that were used, in two independent GSEA analyses, were: Hallmark v.6.2 (50 gene sets) and C2-Curated Gene Sets v.6.2 (4,762 gene sets). The enrichment score (ES) of a gene set measures the level of enrichment found in the ranked list for that gene set; for our analysis, the ranking went from genes whose S2N was highest in the HCDR1LMIR6715P ensemble (whose samples are shown in the right side of the heat map described in the paragraph ‘Analysis of RNA-Seq data: visualization of the clustered expression matrix’) to genes whose S2N was highest in the LCDR1HMIR6715P ensemble. We considered statistically significant gene sets having a false discovery rate (FDR) < 0.05 and highly significant those with FDR < 0.01; these two significance thresholds are much lower than the value originally suggested by the Authors of the GSEA method (0.25), to provide an increased degree of selectivity. Among the gene sets with the lowest FDRs, we performed further refinements according to their relationships with the findings described in this article. Due to the biological importance for cancer of the Hallmark gene set named ‘Epithelial mesenchymal transition’ (EMT), for this gene signature we performed a ‘GSEA leading edge analysis’. The leading edge of a gene set includes its genes most upregulated in the biological group of interest, up to the ES peak point (which is obtained from the enrichment plot). These genes were collected in a gmx file and their identifiers were used for data visualization. The computational evidence for the role played by genes downstream of CDH1 in HCDR1LMIR6715P is based on the statistical significance of the two C2 gene sets ONDER_CDH1_TARGETS_2_UP and ONDER_CDH1_SIGNALING_VIA_CTNNB1 (5).

Differential expression and gene ontology analysis of RNA-Seq data. With the aim to be more selective for gene ontology (GO) purposes, in addition to what was already done for selecting the GSEA genes, we performed a two-pronged differential expression analysis, based on these criteria: a) FDR of the p-values obtained after using the Wilcoxon rank sum test (between the two ensembles defined in the previous paragraph) < 0.01, for all genes that, after the computational steps of the previous paragraph, were left in the gene list; b) ratio of variation (median-based) > 1.5 or < 0.66667 (= 1/1.5), thus requesting a variation, in either direction, greater than 50%. The list of differentially expressed genes so obtained (2,123 genes total) was hierarchically clustered (see the paragraph 'Analysis of RNA-Seq data: visualization of the clustered expression matrix') and split into two sets, depending on the expression ratio between the two sample groups (i.e., 'Low CDR1, High miR-671-5p' and 'High CDR1, Low miR-671-5p'). Genes whose ratio of the medians was > 1.5 were 270; genes whose ratio of the medians was < 0.6667 were 1,853. The GO analysis that was done is based on the Expression Analysis Systematic Explorer (EASE) score (a p-value generated by an adjusted Fisher's exact test) (6) and was performed using DAVID Bioinformatics Resources (7); the selected background was 'Homo sapiens'. GO categories were considered as potentially relevant when: 1) were referred to GO biological processes (BP); 2) had two or more gene members inside the list of genes of the considered experimental group; 3) had a p-value < 0.001.

Analysis of RNA-Seq data: visualization of the clustered expression matrix. Selected genes (those that were kept in the analysis after the steps described in the paragraph 'Analysis of gene sets' and 'Differential expression and gene ontology analysis of RNA-Seq data') were log2-transformed, median centered, hierarchically clustered using the version 3.0 of Cluster (8,9) and visualized with the Java-based program TreeView (10). Because normalized gene expression values to be displayed are log-transformed, after this transformation they are assigned to Not-a-Number (NaN) values, which TreeView displays as gray rectangles. The hierarchical clustering is done with respect to the genes (matrix rows); the following expression matrix has 2,123 rows and 204 columns. The first number accounts for what is explained in the paragraphs 'Analysis of gene sets' and 'Differential expression and gene ontology analysis of RNA-Seq data'; the second number is based on the definition of the two experimental groups that we compared, which

are described in the paragraph 'Analysis of gene sets'. The final data visualization was achieved through a heat map.

Analysis of RNA-Seq data: visualization of genes of EMT. Gene belonging to the leading edge of the hallmark EMT signature (see the paragraph 'Analysis of gene sets') and a set of 11 genes also involved in the EMT according to the literature and partially overlapping with the leading edge, were clustered and displayed as described in the paragraph 'Analysis of RNA-Seq data: visualization of the clustered expression matrix'. The 204 samples of these two heat maps were kept in the same order of the general heat map.

TCGA LUSC miRNA data: preliminary analyses at the gene level. miRNA gene values (reads per million) from the two available sequencing platforms (Illumina (https://www.illumina.com/) Genome Analyzer (GA) and High-throughput Sequencing (HiSeq)) of TCGA LUSC samples characterized by genomic subtype (4) and described in the paragraph 'Selection of TCGA LUSC samples' were combined using the algorithm Combat (11). Because of the way this algorithm operates, the null rows cannot be used for this data transformation and are left out; negative numbers obtained after running Combat are kept unchanged in the utilized matrix. When we refer to miRNA gene values, they are intended after this adjustment.

Survival data analysis of TCGA LUSC samples according to miRNA levels. We extracted from the clinical annotation files, for the selected samples (see the paragraph 'Selection of TCGA LUSC samples' four types of survival data: 1) 'Days to death', 2) 'Days to last follow-up', 3) 'Days to last known alive', and 4) 'Vital status'. The processing and data cleaning followed these steps: i) 5 patients having a negative value for their 'Days to last follow-up' were disregarded (their presence was deemed incompatible with survival analyses and this conflict could not be solved otherwise based on the available data); ii) patients recorded as 'dead' received a vital status of 1 and those recorded as 'alive' had their vital status set to 0; iii) the 'Survival days' for 'dead' patients were the 'Days to death'; iv) the 'Survival days' for 'alive' patients were the maximum value between 'Days to last follow-up' and 'Days to last known alive'; v) the survival days were converted to survival years by using the following elementary formula: years = days/365. TCGA

clinical data (biotab type) also allowed us to assign two patients having a not well-defined status to the 'alive' group. For analyses at the miRNA gene level, all gene entries were pre-processed according to their standard deviation and the worst 60% was discarded (due to lower or insufficient signal quality). Later, also genes having a negative median were removed, in order to further improve the quality of our selection after the data transformation made by Combat. Then, log-rank tests were run using the survival data of these samples and categorizing them according to the levels of each of these miRNA genes, independently. For every gene of this list (N=387), besides the p-value, we calculated the a) the hazard ratio (HR), b) upper and c) lower bound of the HR 95% confidence interval (CI), and d) FDR. In particular, survival assessments were made between samples with 'High' and 'Low' levels of each miRNA, with respect to their medians, and those having a p-value < 0.05 were selected. This heuristic procedure exclusively aimed at identifying the most plausible candidates for the following experimental validation; indeed, all these genes had a FDR > 0.05. A similar approach was also followed for defining the expression level-dependent survival of the gene miR-671 and of the two isoforms of miR-671 (3p and 5p) (see the paragraph 'TCGA LUSC miRNA data: preliminary analyses at the isoform level'), with respect to their means across the 348 samples. The Kaplan-Meier (KM) curves of mir-671, mir-671-3p and mir-671-5p (each of the type 'High" vs. 'Low") and the associated log-rank p-values shown in the KM plots were created using Prism.

Selection of candidate miRs for qPCR validation. In total, 42 miRs associated with overall LUSC survival ($p<0.05$) in the TCGA by either mean or median expression were considered. Nanostring data comparing miR expression between parental SK-MES-1 and LN1 cells was used to eliminate miR candidates that had opposing directionality with what was biologically expected. For example, miRs associated with worse survival was expected to be increased in LN1 compared with the parental SK-MES-1, and conversely, miRs associated with better survival was expected to be decreased in the LN1 subclones. In order remain comprehensive during our screen, and to not exclude potentially important miRs relevant to LUSC metastasis, we included 6 miRs from our TCGA survival analyses that either were not included on the Nanostring array or had low read-counts. This selection process resulted in 12 miRs (shown in Fig. 1d) that we then analyzed by RT-qPCR to compare the parental lines (SK-MES-1 and H520) with their

respective metastatic sub-clones (LN1 and LN3). In one case (miR-99b), RT-qPCR was performed and the results of the Nanostring data and the PCR data were discordant, leading to its elimination as a candidate (not shown).

TCGA LUSC miRNA data: preliminary analyses at the isoform level. Shared miRNA isoforms (intended as entries with the same 'miRNA ID' and 'miRNA region') belonging to samples previously selected (see the paragraph 'Selection of TCGA LUSC samples') were identified in the Illumina GA and HiSeq platforms of TCGA LUSC data, converted into a matrix (miRNA isoforms x samples) format, sample-annotated, and analyzed. At this stage it was also checked that this genomics information was not shared across these two platforms for any of these samples. Similar to what was done at the gene level, there was a computational adjustment between these two platforms using the algorithm Combat (11). The miR-671-3p and miR-671-5p values that were generated after this step were those used for further analyses (see the paragraphs 'Survival data analysis of TCGA LUSC samples according to miRNA levels' and 'Selection of candidate target genes of miR-671-5p').

Selection of candidate target genes of miR-671-5p. The Target Scan database, Release 7.1, was used as a guide to define possible matches between the isoform miR-671-5p and its candidate target genes. This database is split into three parts: conserved miRNA families, non-conserved (and confidently annotated) miRNA families and predicted targets. These database files were pre-processed in order to exclude entries that were not referred to humans. Also, since miR-671-5p belongs to the non-conserved miRNA family, we exclusively used this database's part for our analysis. Genes were considered when a) were linked to miR-671-5p through Target Scan; b) fulfilled the computational requirements (biological significance + signal quality) that brought to their selection in a previous publication from our group, which utilized TCGA LUSC RNA-Seq data (4). Point 'b' aimed to orientate the validation effort toward genes that, on average, had better chances to be biologically relevant while partially capping the total number of candidates. This approach led to a set of 574 genes, on which correlation coefficients and linear regressions (both between the gene g and miR-671-5p) were calculated. Both the Pearson correlation coefficients and the linear regressions used the adjusted values of miR-671-5p and log-transformed values of RNA-Seq data. At this

stage, for visualization purposes and preliminary assessments, we selected genes that had a FDR of the p-value obtained after performing a linear regression between miR-671-5p and them below 0.001 (Figure S5A). They were a total of 224. Then, we performed individual and joint survival assessments for these genes. For individual assessments, samples are categorized, according to the levels of a gene g, below (<) or above (≥) median(x). For joint survival purposes, samples are categorized in one of these three groups: a) samples that, at the same time, have an expression value of miR-671-5p above median(miR-671-5p) and of the gene g below median(g); b) samples that, at the same time, have an expression value of miR-671-5p below median(miR-671-5p) and of the gene g above median(g); c) samples that do not meet neither the requirement 'a' nor 'b'. In joint survival analysis assessments, a log-rank test between samples meeting 'a' and samples meeting 'b' was run, for each of the 574 genes, independently. At the end, we selected genes: 1) having a joint survival hazard ratio (HR) > 2; 2) having a joint log-rank p-value < 0.01; 3) being statistically significant, individually and using the median as separation threshold between samples, with a log-rank p-value < 0.05; 4) having an individual HR > 1.5, in the same comparison; 5) having a correlation coefficient with miR-671-5p < -0.25 (a value picked accounting for the considerable data noise of this dataset); 6) whose FDR of the p-value obtained after performing a linear regression between miR-671-5p and them was < 0.00001; 7) whose variation, after splitting the samples in those below and above the miR-671-5p median across the 348 samples, and calculating the ratio of expression in these two groups, was > 25% (as a measure of effective gene variation after this categorization based on miR-671-5p. levels). In all these cases, HR > 1 means that when the gene levels, alone or jointly with miR-671-5p (see before), are above the median, the group survival rate is worse, and vice versa. This selection narrowed the number of relevant genes negatively correlated with miR-671-5p down to 13. These genes were considered as potential target genes of miR-671-5p, based on computational evidence, to be experimentally validated.

Additional survival analyses of TCGA LUSC samples using RNA-Seq data. The same categorization described in the paragraph 'Selection of candidate target genes of miR-671-5p' was also used to collect samples a) having expression above median both for CDR1 and miR-671-5p and b) having expression below median both for CDR1 and miR-671-5p. These two groups were statistically compared through a

log-rank-test. Finally, a direct comparison, in terms of survival, was also performed between samples having a CDR1 level above the median vs. samples having a CDR1 level below the median, based on normalized TCGA LUSC RNA-Seq values.

Gene expression analyses from TCGA. Expression of has-miR-671-3p, has-miR-671-5p, EGFR, SIGMA1R, ESR1, PGR, and AR were accessed via Oncolnc.org (12) using the gene names as search terms. Anonymized patient data were downloaded and plotted using GraphPad Prism.

Analysis of Nanostring data. Total RNA was extracted using the Quick RNA MiniPrep Zymo Research Kit (Genesee Scientific) and RNA quality confirmed by NanoDrop. For each sample 250 ng of total RNA in 3 µL of H_2O was prepared with a HSA miRNA V2 Assay Kit (Catalogue #150325) miRNA Sample Prep kit according to manufacturer's directions. Hybridized samples were processed using the Nanostring nCounter system. Count data from the Nanostring platform were processed through nSolver Analysis Software version 2.0 (https://www.nanostring.com/), with these settings: 1) subtraction of the maximum background signal of the analyzed lane (with final transformation into 0 if this result is negative); 2) code set normalization based on the top 100 genes, in terms of expression across the samples of this experiment (3,13) normalization factor based on the geometric mean. These normalized values were used for the analysis. Specifically, miRNA entries were selected when i) corresponded to endogenous genes, ii) showed a variation greater than 25% between the two samples that were the focus of our analysis (namely, SK-MES Parental and SK-MES-LN1) and iii) had normalized counts greater than 150 in at least one of the two compared samples. The genes identified through this procedure were hierarchically clustered and visualized as described in the paragraph 'Analysis of RNA-Seq data: visualization of the clustered expression matrix'.

Lentivirus packaging and infection. Lentiviral particles for miR-671 and control miR overexpression were purchased from Biosettia (mir-LV464 and mir-LV000). Lentiviral vectors for CDR1as were custom made by GeneCopoeia in the psi-LVRH1H backbone, shR sequences where previously published by Memczak et al. (14) and are included in Supplementary. Lentiviral vectors for CDR1 knockdown and overexpression were purchased from GeneCopoeia: CDR1 shR#1 (HSH000455-31 LVRH1H), CDR1

shR#2 (HSH00455-34 LVRH1H), CDR1 ORF (EX-Z3225-Lv152-GS, EX-Z3225-Lv105). OgNLuc vector was a kind gift from Dr. Antonio Amelio (Lineberger Comprehensive Cancer Center; UNC Chapel Hill, NC). The CDR1 ORF was gateway cloned into a custom gateway lentiviral vector (pHAGE-CMV-FLAG-DEST) from GC-Z3225-CF-GS (GeneCopoeia) for mass spectrometry experiments. Backbone and entry plasmids generously provided by the Protein Expression Laboratory at the Frederick National Laboratory for Cancer Research (Fredericksburg, MD). Lentivirus was produced by transfecting human embryonic kidney cells (293 T) with the lentiviral vector, packaging plasmid (psPAX2) and envelope plasmid (pMD.2G). Media was changed the next day, and 2 days later viral supernatant was collected and filtered to remove cellular debris. Cells were infected with lentiviral particles for 24h in the presence of 8 µg/mL polybrene and were then selected with growth medium containing 200 µg/mL hygromycin (for shR and ORF lentiviruses for each respective cell line) or 2 µg/mL puromycin (for miR-671, miR-control and OgNLuc lentivirus).

Proliferation assays. H520, H520-LN3, SK-MES-1 or SK-MES-LN1 cells were seeded at a density of 20,000 cells per well in 6-well plates in triplicate and counted on a hemocytometer using a Trypan Blue counterstain. To compare proliferation of stably transduced cell lines, cells were plated at either 5,000 or 25,000 cells per well in 96 well plates in quadruplicate. At indicated time points 10% alamarBlue was added and incubated for 1-2 hrs in a 5% CO2/95% air at 37 °C incubator. Fluorescence was measured at 530 nm excitation and 590 nm emission on a Synergy2 fluorescent plate reader (BioTek).

Trans-well migration assays. Trans-well inserts with an 8 µm pore size (Corning) were coated with either 0.1% gelatin on the upper surface for migration or with type I collagen on the lower surface for haptotaxis. After coating, 100k cells in serum free media were added to the upper chamber. The lower chamber contained 10% FBS as a chemoattractant. For brefeldin A treatments, 1 µg/mL brefeldin A or 1% DMSO was added to both the upper and lower chambers. After 4-6 hours, cells were fixed with methanol and stained with methylene blue after careful removal of all non-migrated cells from the upper

chamber. Cells per high powered field were determined using a pipeline in CellProfiler 3.0 software **(15)**. Five high powered fields were analyzed per well.

Scratch assay. Cells were plated on glass coverslips coated with 50 µg/mL type I collagen. Once the cells formed a monolayer, a 200 µL pipette tip was used to remove a thin band of cells creating a "scratch." Cells were rinsed and maintained in media containing 1% FBS for 12 hours. Cells were fixed with 4% PFA and stained with Golgi marker GM130 as detailed in the "Immunocytochemistry" section. Imaging was performed using a Leica DMi8 inverted microscope. For each group, 200x images were obtained of 6 random fields along the scratch edge. The number of cells in each field that entered the scratch was counted manually. The orientation of the Golgi in cells at the scratch border was scored as towards or away. Cells with a Golgi orientation parallel to the scratch were excluded.

Cell surface protein isolation. Cell surface proteins were isolated using a Pierce cell surface isolation kit (Thermo) according to manufacturer's directions. Briefly, to isolate cell surface proteins, cells were treated with cell impermeable Sulfo-NHS-SS-Biotin. Control cells were treated with vehicle only. Biotinylated cell surface proteins were isolated with an avidin resin. Proteins not bound to the resin were collected in the flow through and biotinylated proteins were eluted with SDS-PAGE buffer and 50 mM DTT. Protein fractions were evaluated by Western blotting as detailed in the "Western blotting" section.

Western blotting. Cells were lysed in RIPA buffer (Thermo) containing complete protease inhibitor cocktail (Roche) and Halt phosphatase inhibitor cocktail (Thermo). Total protein was mixed with Laemmli buffer and 5% 2-Mercaptoethanol denatured at 95°C for 5 min and loaded onto 10% SDS-PAGE gels, after which protein was transferred to nitrocellulose membranes (BioRad). Membranes were blocked in 5% non-fat dried milk in tris-buffered saline-tween 20 (TBS-T) for one hour at room temperature prior to probing with primary antibodies overnight at 4 °C or 1 hour at room temperature. Primary antibodies included anti-Jak1 (clone 6G4, #3344S), and anti-VCP (#2648) from Cell Signaling Technology; anti-vinculin (clone hVIN-1, #V9131), anti-FLAG (#F3165), and anti-AP1G1 (#HPA041224) from Sigma; anti-SIG-1R (clone B-5, #sc-137075) from Santa Cruz Biotechnology; anti-GM130 (clone 35, 610823) from BD; anti-EGFR

(ab2430) from Abcam; and anti-CDR1 (NBP2-57758) from Novus Biologicals. After probing with primary antibodies, membranes were washed three times in TBS-T and then probed with the appropriate horseradish peroxidase-conjugated secondary antibodies (anti-mouse (#115-035-003) or anti-rabbit (#111-035-003) from Jackson ImmunoResearch). Then, the membranes were washed three times in TBS-T and developed using Clarity Western ECL substrate (BioRad) or SuperSignal West Fempto (Thermo). Membranes were visualized using a BioRad ChemiDoc MP system (BioRad).

Quantitative real-time PCR. Total RNA from cell lysates was extracted using the Quick RNA MiniPrep Zymo Research Kit (Genesee Scientific). For mRNA and circRNA analysis cDNA was synthesized using an iScript cDNA Synthesis Kit (Bio-Rad) as per the manufacturer's instructions, except for strand specific analysis of CDR1, which was performed using a SuperScript First-Strand cDNA kit (Invitrogen) with gene specific primers (Supplementary Table 1). Analysis of RNA levels was determined by a StepOnePlus Real-Time PCR System (Applied Biosystems). A list of gene specific primers used for RT-qPCR is included in the extended data (Supplementary Table 1). RT-qPCR was performed with 1-2.5 μL cDNA, 1 μL each of 20 μM forward and reverse primers, and 12.5 μL of PowerUp SYBR Green Master Mix (Thermo) for a total volume of 25 μL. TaqMan Assays (Applied Biosystems) were used for miR-129b (Assay ID: 000449), miR-452 (Assay ID: 002329), miR-671-3p (Assay ID:002322), miR-542 (Assay ID: 002428), miR-1301 (Assay ID:002827), miR-181-d (Assay ID: 001099), miR-421 (Assay ID: 002700), miR-505 (Assay ID: 002089), miR-374 (Assay ID: 002125), miR-500 (Assay ID: 001046), miR-340 (Assay ID: 002258), miR-671-5p (Assay ID:197646_mat) and snRNA U6 (Assay ID: 001973). For both TaqMan and SYBR PCR, each cycle consisted of 15 s of denaturation at 95 °C and 1 min of annealing and extension at 60 °C (40 cycles). Reactions were run in duplicate or triplicate. Fold change was calculated using the $2^{-\Delta\Delta CT}$ method. The rRNA 18S alone or in combination with GAPDH and TBP was used for normalization of mRNA and long non-coding RNA as indicated in figure legends and snRNA U6 was used for miR normalization.

Immunocytochemistry. Cells were grown on glass coverslips and were fixed for 15 minutes with 4% PFA and permeabilized with .25% Tween 20 for 15 minutes. For brefeldin A treatments, 1 μg/mL Brefeldin A was added to cells for 5 hours before fixation. Cell proteins were blocked in 2% BSA and 0.25% Tween in

PBS for 1 hour at room temperature. Coverslips were then incubated with the primary antibodies anti-CDR1 (Novus, NBP2-57758, 1:500) and anti-COPA (Santa Cruz, H-4, sc-398099, 1:100) or anti-Adaptin γ (BD, Clone 88, 610385, 1:100), or anti-GM130 (BD, Clone 35, 610823, 1:500) diluted in blocking buffer overnight at 4°C. Cells were washed and incubated with appropriate secondary antibodies, goat anti-rabbit (Alexa fluor 594) and goat anti-mouse (Alexa fluor 488) diluted 1:500 in blocking buffer for 1 hour at room temperature, protected from a light source. Hoescht (1:10000) was used for nuclear staining. Coverslips were mounted with prolong gold (Invitrogen). Confocal microscopy was performed with a Zeiss LSM 710 and 3D deconvolution was performed with AutoQuant X3 (Bitplane) and FIJI software (16) was used for image processing. For colocalization analysis, background subtraction was performed on a single slice of a z-stack and ROIs were created around individual cells, then Pearson's correlation coefficients were calculated using the Coloc2 plugin in FIJI.

Proximity ligation assay. Cells were plated on glass coverslips coated with 50 µg/mL type I collagen and fixed 48h later using 4% PFA. Proximity ligation assays (PLAs) were performed according to the manufacturer's instructions (MilliporeSigma). Briefly, cells were permeabilized with .25% Tween 20 for 15 minutes and blocked with 2% BSA and 0.25% Tween in PBS for 1 hour at room temperature. Coverslips were then incubated with the primary antibodies anti-CDR1 (Novus, NBP2-57758, 1:500) and anti-COPA (Santa Cruz, H-4, sc-398099, 1:500), or anti-Adaptin γ (BD, Clone 88, 610385, 1:500) diluted in blocking buffer overnight at 4°C. Control coverslips were incubated with no antibodies or single antibodies. Coverslips were then washed and incubated with plus and minus PLA probes followed by ligation and amplification. Coverslips were mounted in media containing DAPI and images were acquired with Leica DMi8 inverted microscope using a 63x oil objective.

Live cell imaging. Glass bottom dishes were coated with Rat Tail I Collagen at 10ng/ml and allowed to incubate for 1 hour at room temperature. After washing, 75,000 cells were plated followed by 2 ml of complete media (MEM, 10% FBS, NEAA, Sodium Pyruvate, Pen Strep). Cells were placed in the Olympus VivaView FL Incubator for 1 hour at 37°C. Images were acquired every 5 minutes for 16 hours and tracked

using the Manual Tracking FIJI software plugin (https://imagej.nih.gov/ij/plugins/track/track.html). Velocity and distance were analyzed using the Chemotaxis FIJI software plugin (https://ibidi.com/chemotaxis-analysis/171-chemotaxis-and-migration-tool.html).

Retention using selective hooks assay. SK-MES-1 EV or CDR1 cells were seeded in an 8-well Cellvis chambered cover glass coated with 50 µg/mL type I collagen (20,000 cells per well). Cells were transfected with the RUSH construct Ii-Str_ssSBP-EGFP and/or the Golgi marker mApple-SiT using Lipofectamine 2000 according to the manufacturer's protocol. mApple-Sit-N-15 was a gift from Michael Davidson (Addgene plasmid #54948;http://n2t.net/addgene:54948;RRID: Addgene_54948) and Ii-Str_ssSBP-EGFP was a gift from Franck Perez (Addgene plasmid #65277;http://n2t.net/addgene:65277;RRID: Addgene_65277). After 4 hours, the transfection media was replaced with 200 µL antibiotic-free and phenol red-free MEM for live cell imaging. Imaging was conducted on a Zeiss Laser Scanning Microscope 800 using a 63x/1.4 NA objective lens at 37 °C and 5% CO2, 20-24 hours post-transfection. 5-6 fields containing 1-2 transfected cells each were selected in ZEN Blue software using the tiles feature. D-Biotin diluted in antibiotic-free and phenol red-free MEM was added to each well to a final concentration of 40 µM. Imaging of cells began 5 minutes after the addition of biotin to the well, and continued for up to 1 hour at 1 minute intervals. The Golgi marker mApple-SiT was used to mask the Golgi apparatus, and the mean gray value of the GFP signal in the Golgi was measured for each time point using NIH ImageJ software.

Immunostaining. CDR1 staining was performed in formalin-fixed, paraffin embedded tumor sections (4 µm thickness). After deparaffinization, rehydration and citrate antigen retrieval, 3% H_2O_2 was used to block the endogenous peroxidase activity for 10 min, avidin/biotin blocking was performed with a Vector Labs blocking kit. Protein blocking of non-specific epitopes was done using 10% normal goat serum + 0.3% Triton-X for 30 min. Slides were incubated with primary antibody anti-CDR1 (NBP2-57758, Novus Biologicals) in 3% normal goat serum + 0.3% Triton-X. After washing with PBS, biotinylated anti-rabbit IgG (Biocare Medical) was added followed by incubation with Avidin-Biotin Complexes (Vector Labs) and visualized with 3,3′-diaminobenzidine chromogen (Vector labs) and

counterstained with Gill's hematoloxylin #3. Slides were dehydrated and coverslips were mounted with Permount (Fisher). Images were obtained with a Leica DMi8 inverted microscope. For CDR1 staining we examined 5–10 random fields at 200 × magnification for each group and analyzed using CellProfiler 2.0 software **(15)** to quantify the number of positively staining cancer cells per high-powered field (200 × magnification); on average between 5-15 high powered fields were used to quantify CDR1 positive cells. For air liquid interface (ALI) culture staining, cultures were fixed with 4% PFA and permeabilized with 0.2% Triton X 100 and blocked with 1% BSA, 1% fish gelatin, 0.1% Triton X-100, and 5% normal goat serum. ALI cultures were incubated with primary antibodies, rat anti-tubulin (MAB 1864, Millipore), mouse anti-Mucin5AC (45M1, Thermo), in blocking buffer overnight. After washing, ALI cultures were incubated with appropriate secondary antibodies (Jackson ImmunoResearch) at 1:1,000 dilution and stained with phalloidin and Hoechst 33342. Images were obtained with an Olympus FV1000 confocal microscope.

In situ hybridization. Staining and analysis of CDR1as ISH was done in collaboration with UNC's Tissue Pathology Laboratory. For each TMA, two consecutive 5 µm thick sections were stained. On the first slide, HS-CDR1as RNA was detected using the RNAscope® 2.5 LS Probe Hs-CDR1as-C1 and Hs-PPIB-C2 was detecting using the Hs-PPIB-C2 Probe (catalog #s 532658 and 532658, respectively; ACD Biotechne, Newark, CA) in a Bond RX autostainer (Leica Biosystems, Buffalo Grove, IL) following the manufacturer's directions. Probes were visualized using TSA-Cy5 (HS-CDR1as) and TSA-Cy3 (Hs-PPIB-C2) and nuclei were counterstained with DAPI, all part of the ACD detection kit. On the second slide for each TMA, an RNAscope Negative control Probe (RNAscope® 2.5 LS Negative Control Probe_dapB; catalog # 312038) was applied as described above, and detected using both TSA-Cy5 and TSA-Cy3. All stained TMA slides were scanned in the Aperio Versa 200 scanner (Leica Biosystems) at an apparent magnification of 20X. Images were uploaded to the eSlideManager database (Aperio; eSlideManager version 12.3.3.7075) at the Translational Pathology Laboratory at UNC. In order to remove auto-fluorescent and non-specific secondary antibody signal from analyzed regions, tissue cores from negative control probe TMA slides were digitally aligned with cores on consecutive sections stained for Hs-CDR1as and Hs-PPIB probes (VIS TissuealignTM Module; VIS version 2018.4.5.4643;

Visiopharm, Hoersholm, Denmark). After alignment, the VIS Image Analysis module was used to generate Regions of Interest (ROIs) on the negative probe slides that were used for analysis on the aligned positive probe slides. These ROIs were mostly devoid of non-specific fluorescent signal in both the Cy3 and Cy5 channels. As a further filtering step, size and shape restrictions were used for Cy3 and Cy5 signal on positive probe slides; regions of fluorescent signal with an area larger than 14 μm^2 were removed from analysis because they tended to be non-specific in nature. Finally, the same ROIs and filtering steps were applied to non-aligned copies of each negative probe image and any remaining areas of non-specific Cy3 and Cy5 signal were quantified. The area (μm^2) of the remaining non-specific signal was used to normalize signal for each tissue core on the positive probe slides as follows:

Area Cy3 (positive probe) – Area Cy3 (negative probe) = Total area normalized Cy3 signal

Area Cy5 (positive probe) – Area Cy5 (negative probe) = Total area normalized Cy5 signal

The ratio of normalized Cy5 (Hs-CDR1as-C1) area: normalized Cy3 (Hs-PPIB-C2) area was then determined for each core.

Immunohistochemistry (IHC) in clinical samples. IHC was performed with a rabbit monoclonal antibody to CDR1 (Novus, Centennial, CO, NBP2-57758). IHC was carried out in the Bond Autostainer (Leica Microsystems Inc.; Norwell MA). Slides were dewaxed in Bond Dewax solution (AR9222) and hydrated in Bond Wash solution (AR9590). Antigen retrieval was performed for 20 min at 100ºC in Bond-Epitope Retrieval solution 1, pH-6.0 (AR9961). Slides were incubated with primary antibody (1:300) for 30 min. Antibody detection was performed using the Bond Intense R detection system (DS9263) with Novolink Polymer (Leica; RE7260-K). Stained slides were dehydrated and coverslipped. Positive and negative controls (no primary antibody) were included during the run. Stained slides were digitally scanned at 20x magnification using the Aperio ScanScope-XT (Aperio Technologies, Vista, CA) and were uploaded to the Aperio eSlideManager database (Leica Biosystems Inc; eSlideManager version 12.4.3.5008) at the Translational Pathology Laboratory at UNC. TMA slide images were digitally segmented into individual cores using Aperio TMA lab (Leica Biosystems Inc.). For whole tissue sections, tumor regions were manually annotated on images. All images were analyzed using the Aperio Cyto v2 algorithm. The number

and percentage of cells with light (1+), medium (2+) and strong (3+) nuclei and cytoplasmic staining was determined.

Lipid protamine hyaluronic acid nanoparticles. Condensation of small RNAs within the protamine and hyaluronic nanocomplex followed by liposomal encapsulation and decoration with polyethylene glycol was performed as previously described (17). In brief, DOTAP/cholesterol liposomes were prepared as follows: DOTAP (NOF Corporation) and cholesterol (Sigma) were both dissolved in chloroform at a concentration of 20 mM and mixed by 1:1 mole ratio. The solvent was removed under vacuum evaporator. The lipid film formed and was then hydrated with isovolumetric distilled water to form cationic DOTAP/cholesterol liposomes (10 mM), which were mixed by ultrasonication in an ice bath and sequentially extruded through polycarbonate membranes (400 nm 10 times, 200 nm 10 times and 100 nm 10 times) (Millipore). LPH cores were prepared by adding 100 µL of 5% glucose solution A (containing the 13 ug of protamine (Sigma)) to 100 µL of 5% glucose solution B (containing the 12.5 µg of miRNA and 12.5 ug of HA (HA1M-1, Lifecore), then mixed well and incubated at room temperature for 10 min. 30 µL of cationic DOTAP/cholesterol liposomes were added and incubated at room temperature for 10 min to ensure lipid coating. The lipid-coated nanoparticles were PEGylated using a post-insertional approach by adding 10 µL DSPE-PEG (NOF Corporation) (10 mg/mL) and 10 µL DSPE-PEG-anisamide (10 mg/mL) and incubating at 60 ºC for 15 min. For *in vitro* uptake of LPH NPs, 20 µM Cy5.5 labeled RNA encapsulated in LPH-NPs with or without targeting ligand AE-AA was added to complete culture media and incubated at 37°C in 5% CO_2/95% air for 2 hours then analyzed by flow cytometry. Data are represented as mean fluorescent intensity normalized to untreated cells. For biodistribution, kinetic, and therapeutic experiments, mice were administered 12.5 µg of RNA in 250 µL, approximately 0.5 mg/kg, IV. For therapeutic experiments, mice were randomized and then adjusted so that mean baseline IVIS signal was consistent between groups. Beginning 20 days after orthotopic injection of cancer cells, mice were injected with NPs containing scr-miR or miR-671-5p mimic 3 times per week for a total of 8 injections. *In vivo* grade RNA was purchased from Dharmacon: Scr-Cy5.5, sense (5’ AAUUCUCCGAACGUGUCACGUCy5.5 3’) and antisense (5' ACGUGACACGUUCGGAGAAUU 3'), miR-671-5p mimic, sense (5'

AGGAAGCCCUGGAGGGGCUGGAG 3') and antisense (5' CUCCAGCCCCUCCAGGGCUUCCU 3'), Scr miR, sense (5' GAGGCGAGGGCGAGGCGGUACA 3') and antisense (5' GAGGCGAGGGCGAGGCGGUACA 3').

Characterization of LPH NPs. Particle size was measured by dynamic light scattering with a Malvern ZS90 Zetasizer in water. Transmission electron micrographs were acquired with a JEOL 100CX II transmission electron microscope. For staining LPH NPs were applied to a 300-mesh carbon-coated copper grid (Ted Pella Inc.) and allowed to settle for 5 min. Grids were then stained with 5 µL of 1% uranyl acetate for 10 sec, then quickly dried. Images were acquired at an accelerating voltage of 100kV.

Harvesting protein, immunoprecipitation, and peptide generation. HEK293T (4x15cm plates) or SK-MES1-LN1 (8x15cm plates) parental cells or cells stably expressing Flag-HcRED or Flag-CDR1, were washed thrice in 1X cold PBS and then scrapped in 1X cold PBS. Collected cells were centrifuged and resuspended in 0.1% NP40 lysis buffer (0.1% NP40, 150mM KCl, 2mM EDTA, 50mM TRIS-HCl pH7.4, 10% glycerol) containing Benzonase (Sigma E1404), protease and phosphatase inhibitors (Halt Thermo). Lysates were passed through a 26.5-gauge syringe needle five times. Lysates were cleared through centrifugation and quantified by BCA. Prior to the IP, EZview Red Anti-FLAG M2 beads (Sigma F2426) were washed three times in 0.1% NP40 buffer. The IP was performed overnight (16-18hr) at 4^{o}C with 10mg (HEK293T) or 25mg (SK-MES1-LN1) of protein by rotating gently with the washed FLAG beads. Samples were washed four times in 0.1% NP40 lysis buffer before being subjected to a modified FASP digestion and on bead trypsinization. Trypsinization was achieved by incubating beads with excess trypsin (Promega V5111) at 37^{o}C overnight (16-18 hours). Subsequently, eluted peptides were desalted using a Pierce C-18 spin column (Thermo 89870), followed by an ethyl acetate cleanup step.

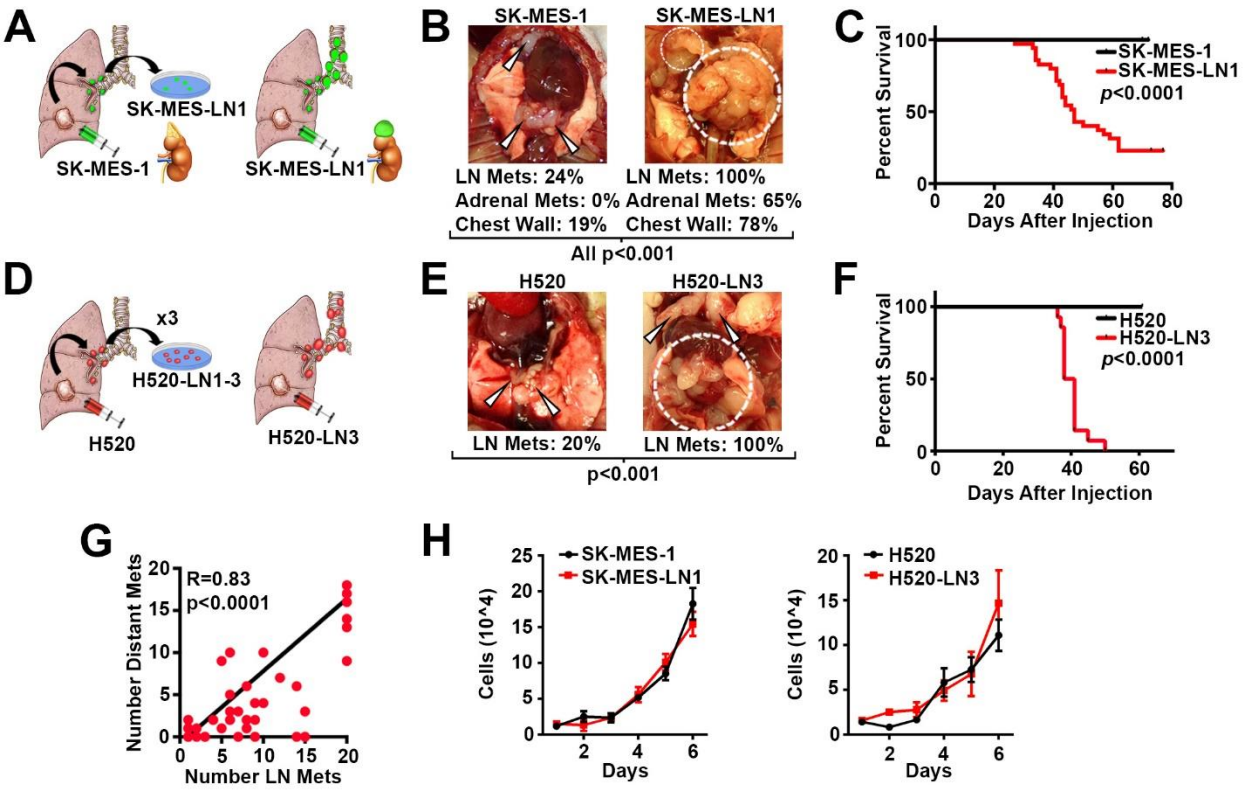
Mass spectrometry. To separate peptides, reverse-phase nano-HPLC was performed by a nanoACQUITY UPLC system (Waters Corporation). Peptides were trapped on a 2 cm column (Pepmap 100, 3µM particle size, 100 Å pore size), and separated on a 25cm EASYspray analytical column (75µM ID, 2.0µm C18 particle size, 100 Å pore size) at 45^{o}C. The mobile phases were 0.1% formic acid in water (buffer A) and 0.1% formic acid in acetonitrile (buffer B). A 180-minute gradient of 2-25% buffer B was used with a flow

rate of 300nl/min. Mass spectral analysis was performed by a Orbitrap Fusion Lumos mass spectrometer (Thermo Scientific). The ion source was operated at 2.4kV and the ion transfer tube was set to 300ºC. Full MS scans (350-2000 m/z) were analyzed in the Orbitrap at a resolution of 120,000 and 1e6 AGC target. The MS2 spectra were collected using a 1.6 m/z isolation width and were analyzed either by the Orbitrap or the linear ion trap depending on peak charge and intensity using a 3s TopSpeed CHOPIN method (18)**.** Orbitrap MS2 scans were acquired at 7500 resolution, with a 5e4 AGC, and 22 ms maximum injection time after HCD fragmentation with a normalized energy of 30%. Rapid linear ion trap MS2 scans were acquired using an 4e3 AGC, 250ms maximum injection time after CID 30 fragmentation. Precursor ions were chosen based on intensity thresholds (>1e3) from the full scan as well as on charge states (2-7) with a 30-s dynamic exclusion window. Polysiloxane 371.10124 was used as the lock mass. Raw mass spectrometry data were searched against the Swiss-Prot human sequence database (released 2/2017) (19) using MaxQuant version 1.6.2.3 (20,21)**.** The parameters for the search were as follows: specific tryptic digestion with up to two missed cleavages, static carbamidomethyl cysteine modification, variable protein N-terminal acetylation and methionine oxidation, Label Free Quantification (LFQ) and match between runs were enabled.

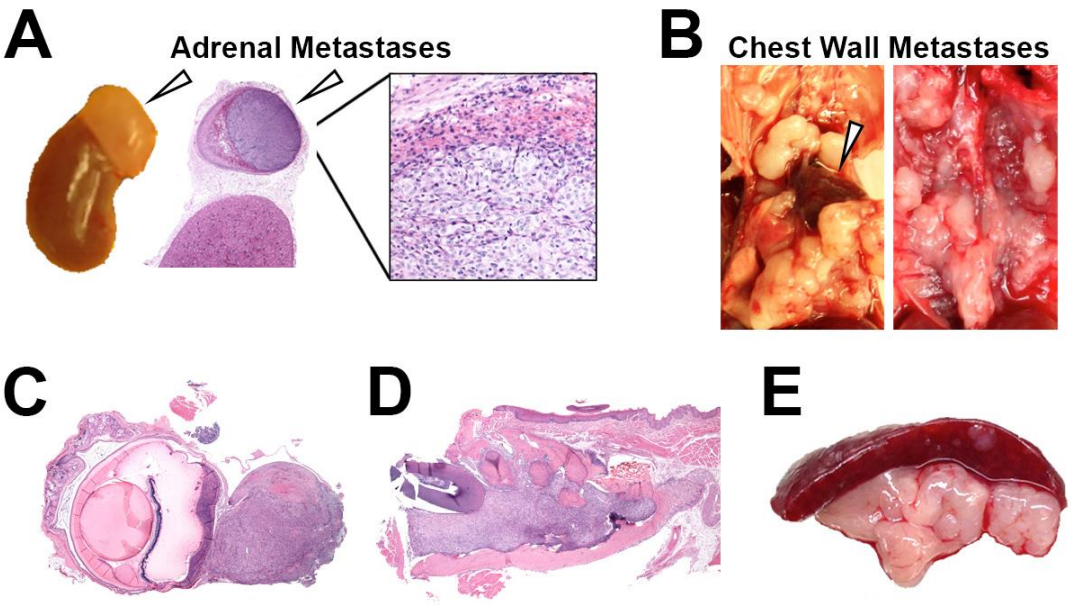
Data filtering and bioinformatics analysis of proteomics data. Data filtering and visualizations were accomplished using Perseus version 1.5.6.0 (22)**.** Protein identifications were filtered for a FDR of 1%, and potential contaminants and decoys were removed. LFQ intensities were log2-transformed and missing values were imputed from a normal distribution using a down-shift of 1.8 and a distribution width of 0.3. Average LFQ values from three replicates and corresponding p-values were calculated using a two-tailed t-test and the FDR was determined by the permutation test in Perseus. To score candidate protein-protein interactions, SAINTq version 0.0.4 (23) using LFQ values was used and then filtered for a 5% FDR. The resulting data was imported to Cytoscape Version 3.6.1 to illustrate the interaction network. The mass spectrometry proteomics data have been deposited to the ProteomeXchange Consortium via the PRIDE[19] partner repository with the dataset identifier PXD012286.

Statistical analysis for experimental data and tissue microarrays. Between 5 and 10 mice were assigned per treatment group; this sample size gave approximately 80% power to detect a 50% reduction

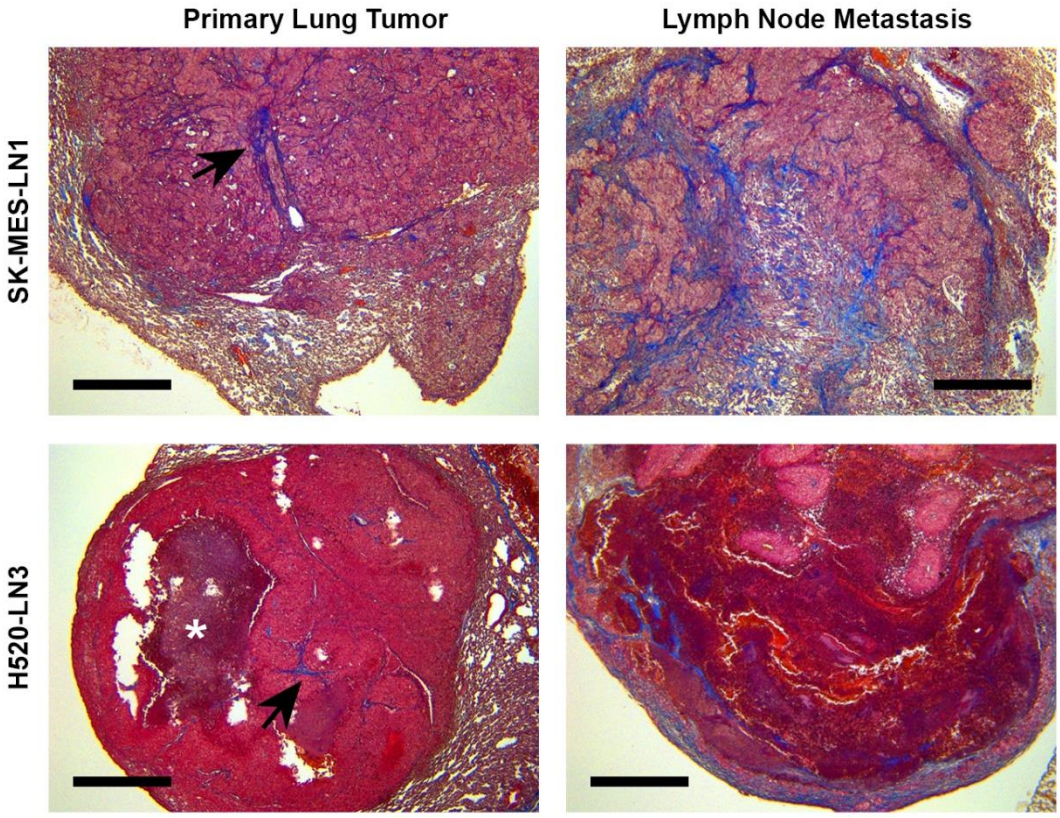
in tumor weight with 95% confidence. Results for each group were compared using Student's *t*-test (for comparisons of two groups) and analysis of variance (ANOVA) (for multiple group comparisons). The multiple hypothesis testing correction of these statistical results was made using the FDR[20]. Survival analyses were performed using the log-rank test. A *p*-value less than 0.05 was deemed statistically significant. All statistical tests for in vitro and in vivo experiments and the log-rank test for the tissue microarray were performed using GraphPad Prism 7 (GraphPad Software, Inc., San Diego, CA).



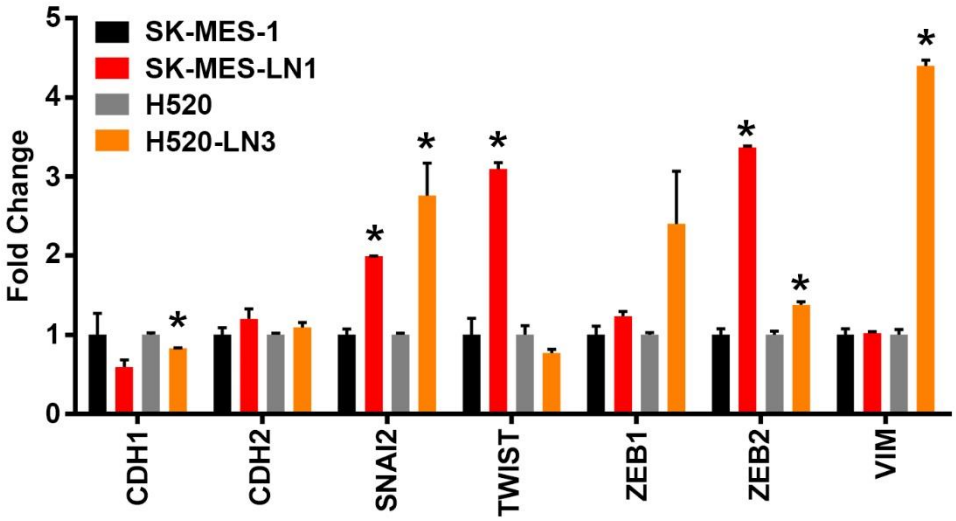
Supplementary Figure 1. Generation of highly metastatic LUSC models. A, Development of the SK-MES-LN1 (LN1) mouse model through iterative *in vivo* passaging. **B**, Gross disease after orthotopic injection of SK-MES-1 or LN1 cells. Arrows and dotted circles represent lymph node metastases. Below, frequency of metastases. **C**, Survival plots of mice following orthotopic injection of SK-MES-1 (n=31) or LN1 (n=35). **D**, Schematic of H520-LN3 (LN3) model development. **E**, Gross disease burden after orthotopic injection of H520 or LN3. Arrows and dotted circles represent lymph node metastases. Below, frequency of metastases. **F**, Survival plots of mice following orthotopic injection of H520 (n=25) or LN3 (n=15). **G**, Correlation between LN and distant metastases in SK-MES-1, LN1, H520, and LN3 orthotopic lung cancer models with significance calculated by Spearman's two-sided t-test (n=83). **H**, Proliferation curves of parental and metastatic sub-clones. Shown are the mean and SEM for two independent experiments performed in triplicate.



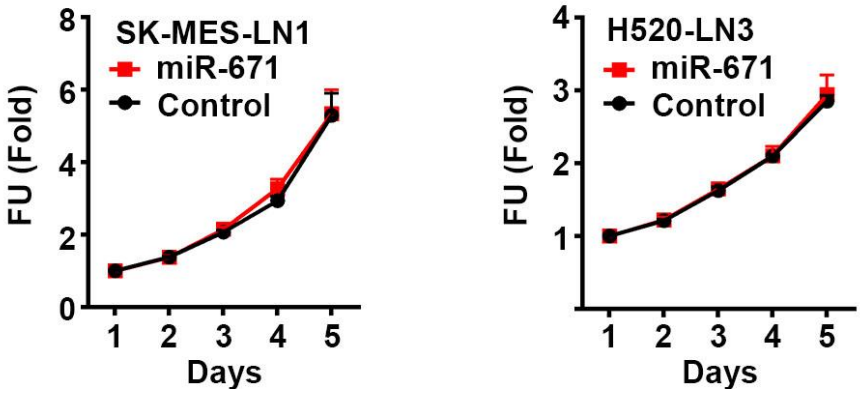
Supplementary Figure 2. Distant metastases in metastatic SK-MES-LN1 model. A, Gross view (left) and H&E (right) of a representative adrenal metastasis (white arrows). **B**, Representative distant metastases located in entire chest cavity (*left*, white arrow points at the heart) and chest wall (*right*), **C**, ocular nerve, **D**, submandibular region, and **E**, pancreas.



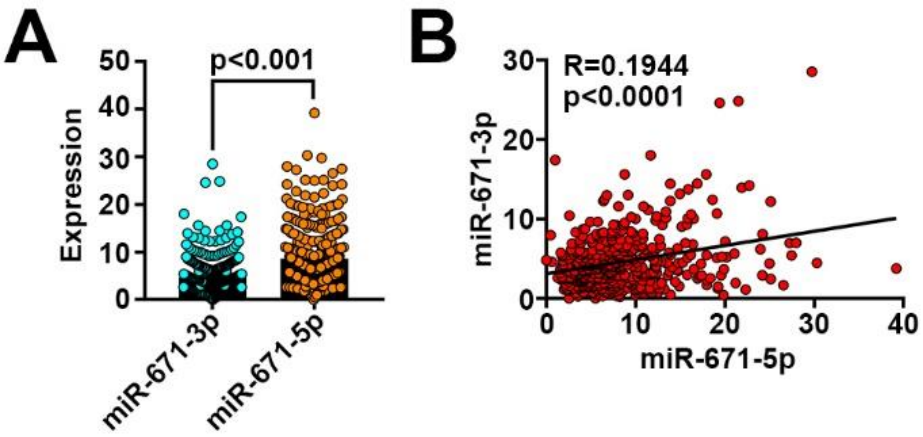
Supplementary Figure 3. Histology of primary lung tumors and lymph node metastases in metastatic LUSC sub-clones. White asterisk indicates central necrosis; arrows indicate intra-tumoral collagen. Tissues were stained with Mason's trichrome, where collagen is stained blue. Scale bars represent 500 μm.



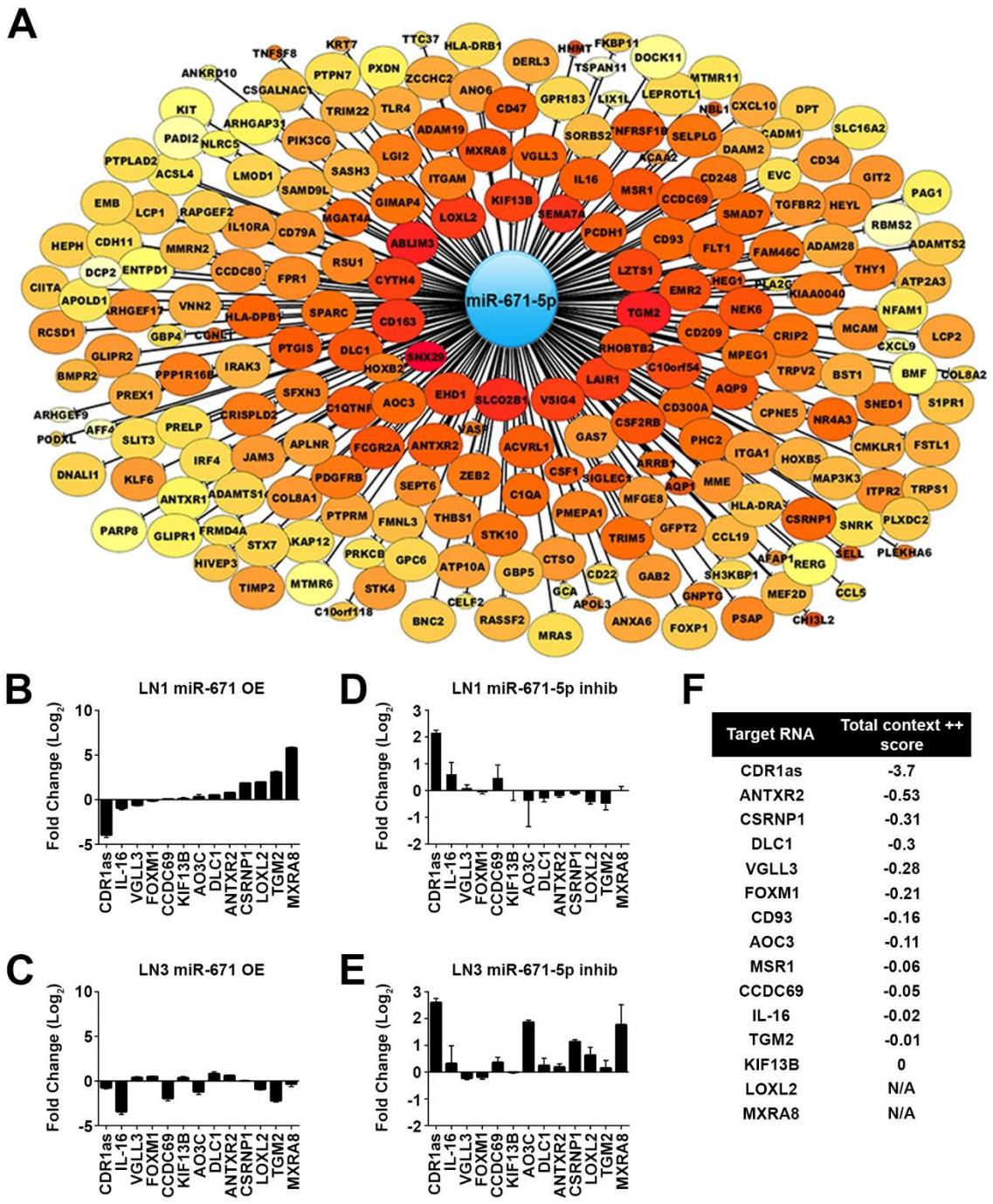
Supplementary Figure 4. Profiling of EMT markers in parental and metastatic sub-clones. Expression of key EMT genes profiled by qPCR normalized to parental expression. Error bars represent mean and SEM of three technical replicates. * $p<0.05$ by Student's t-test corrected for multiple comparisons using the FDR.



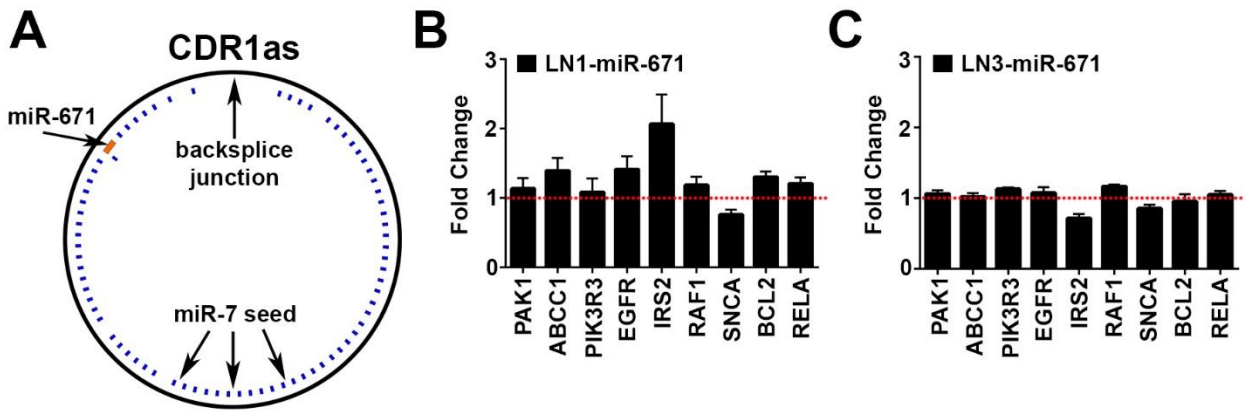
Supplementary Figure 5. miR-671 has no effect on proliferation of LUSC cell lines. Proliferation of miR-671 or control overexpressing SK-MES-LN1 and H520-LN3 cells as measured by alamar blue assay. Shown are the mean and SEM of two independent experiments performed in quadruplicate.



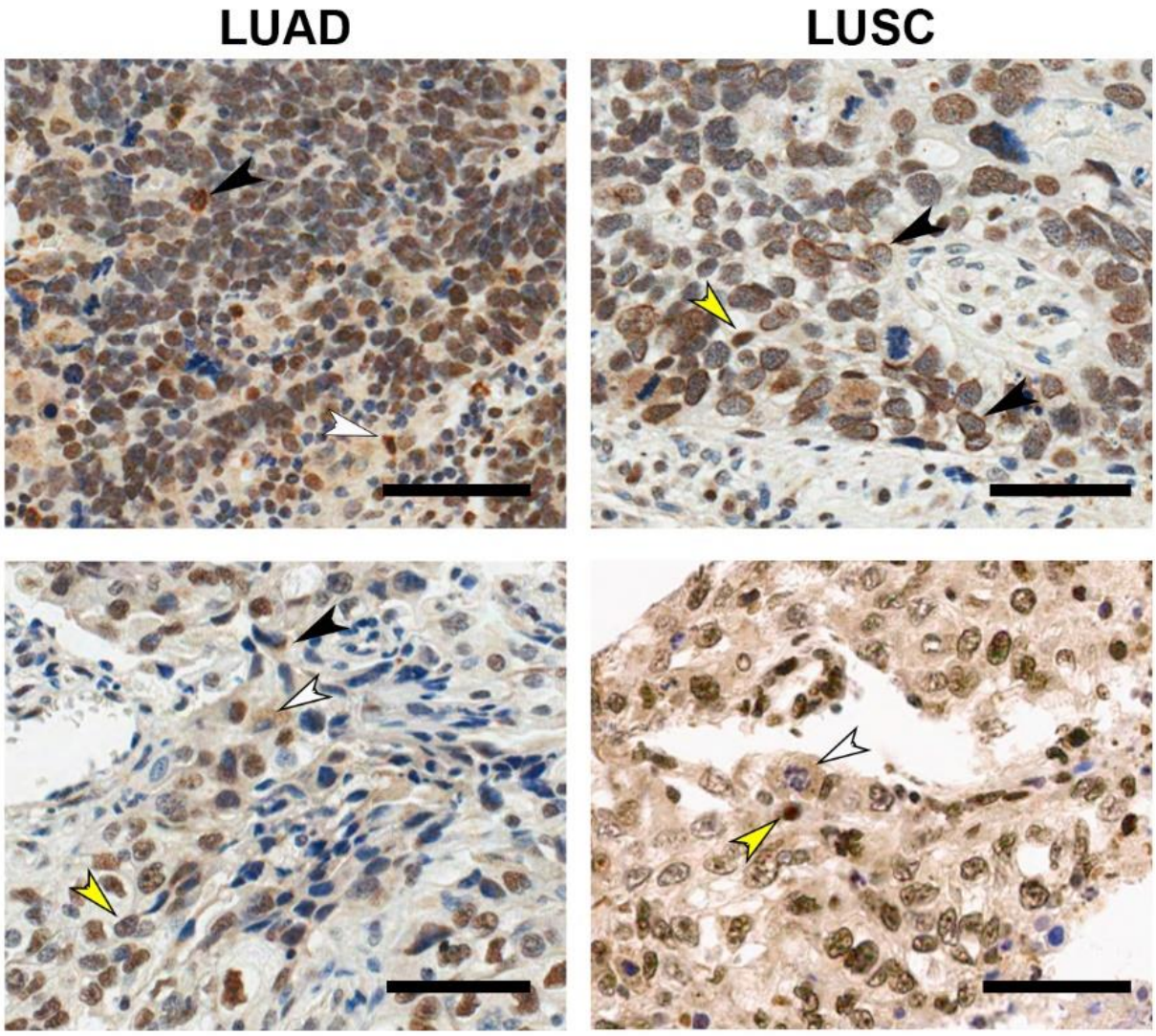
Supplementary Figure 6. Expression of miR-671 isoforms in LUSC patients. A, Expression of miR-671-3p and -5p isoforms from TCGA RNAseq data. Significance determined by two-sided Student's t-test (n=465). **B**, Correlation between miR-671-3p and -5p expression with significance calculated by Spearman's two-sided t-test (n=465). Data extracted from Oncolnc.org.



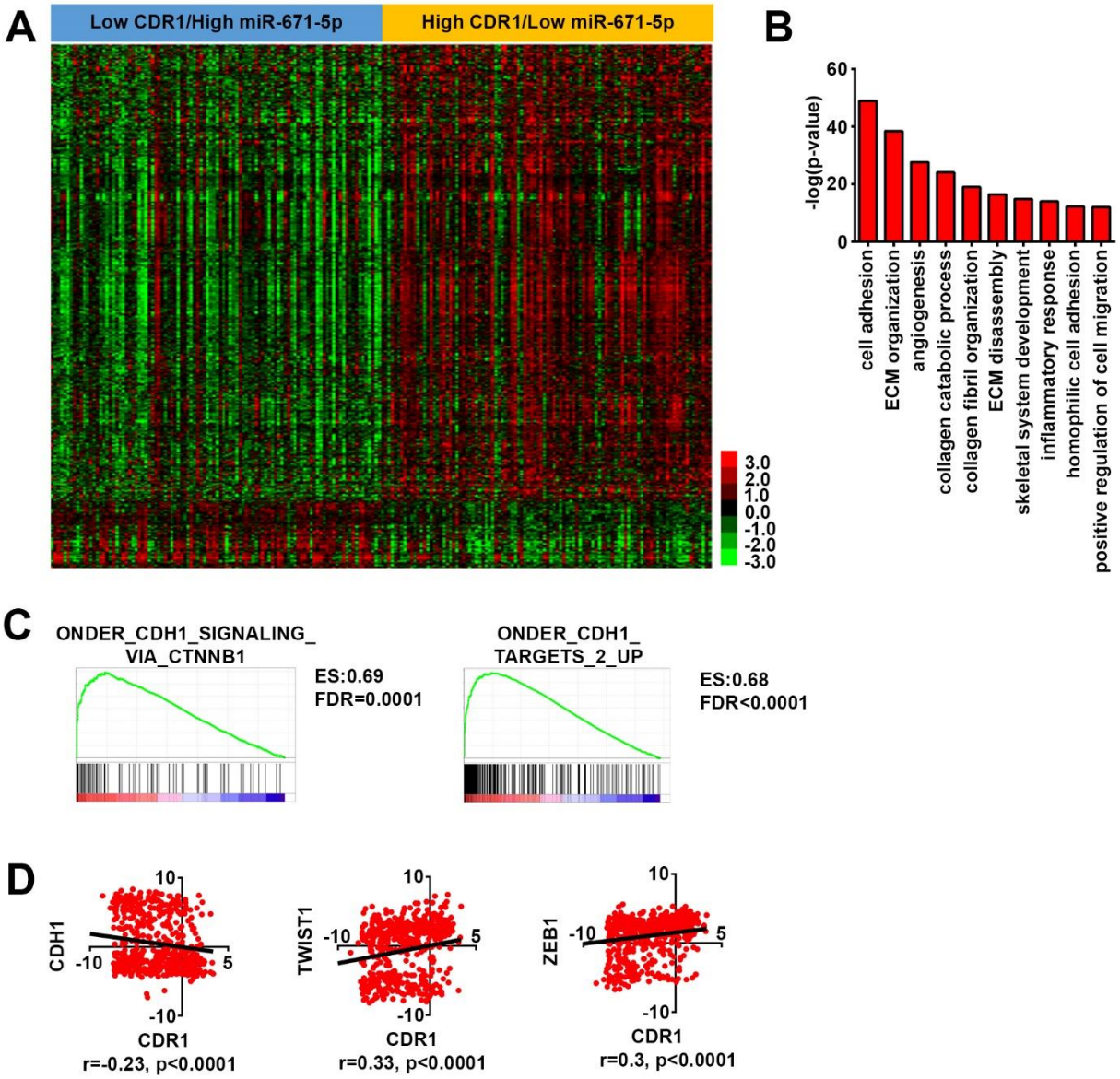
Supplementary Figure 7. miR-671-5p target prediction. A, Cytoscape plot showing a set of candidate miR-671-5p targets based on linear regressions analyses of the LUSC TCGA dataset. This plot shows genes whose FDR < 0.001. Node colors indicate degree of significant hazard ratio for overall survival (Yellow: least significant; Red: most significant). **B** thru **E**, RT-qPCR for target genes in **B**, LN1 or **C**, LN3 cells overexpressing (OE) miR-671, and **D**, LN1 or **E**, LN3 cells stably transduced with miR-671-5p inhibitor. Shown are the mean and SEM of three technical replicates. **F**, Context++ scores from Targetscan for potential target genes.



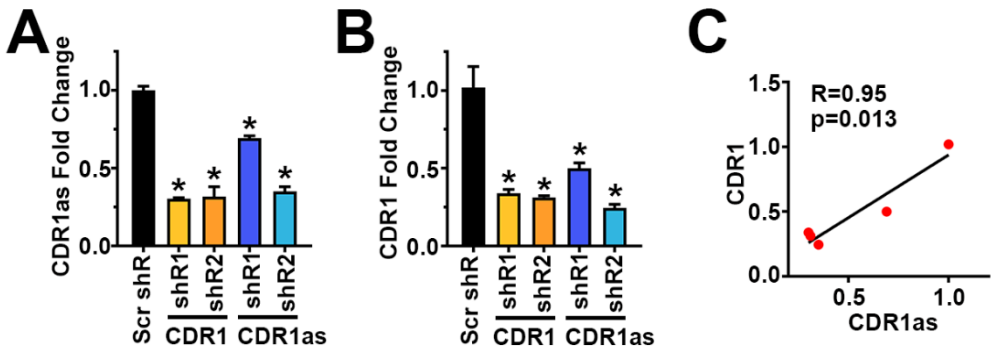
Supplementary Figure 8. Validated, cancer-relevant miR-7 targets are not altered by miR-671 overexpression in LUSC cells. A, Depiction of miR-671 and miR-7 binding sites on the circular RNA CDR1as. **B** and **C**, qPCR profiling of validated miR-7 target genes with relevance to cancer in **B**, LN1 and **C**, LN3 cells overexpressing miR-671. Values indicate the fold change relative to miR-control expressing cells; the dotted red line indicates expression level in miR-control cells. Shown are the mean and SEM of three independent samples. No statistical significance was determined by Student's t-test corrected for multiple comparisons using the FDR.



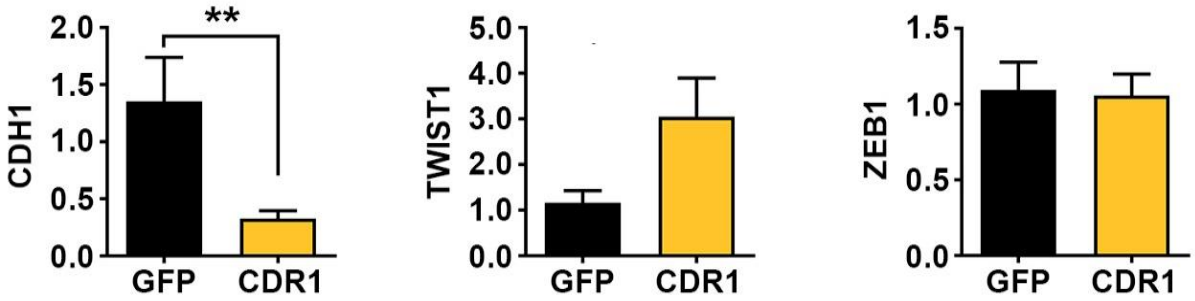
Supplementary Figure 9. Localization of CDR1 in lung tumors. Representative IHC for CDR1 in NSCLC tissue microarrays. Arrows indicate nuclear (yellow), cytoplasmic (white), and perinuclear staining (black). Scale bar is 60 µm.



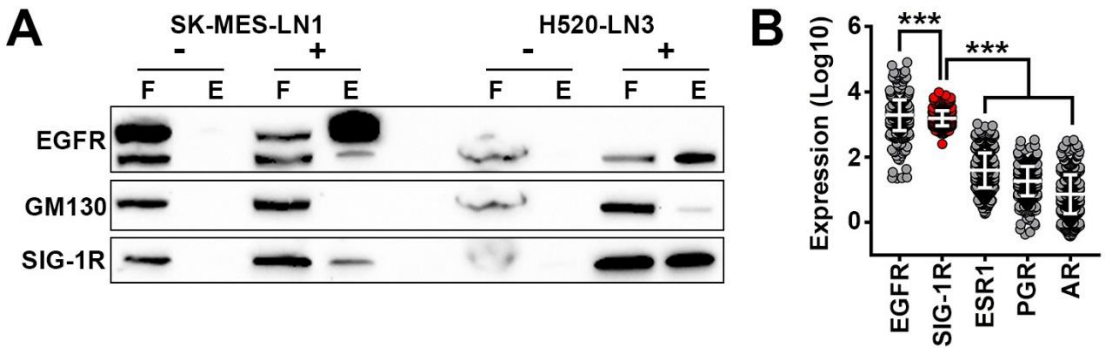
Supplementary Figure 10. TCGA analysis of High CDR1-Low miR-671-5p tumors. A, Heat map of differentially expressed genes between High CDR1-Low miR-671-5p and Low CDR1-High miR-671-5p. **B**, Gene ontology terms for genes upregulated in High CDR-Low miR-671-7p. **C**, Enrichment plots of selected gene sets enriched in High CDR1-Low miR-671-5p. **D**, Correlation of CDR1 expression with EMT markers in >1,000 cancer cell lines using data from the Cancer Cell Line Encyclopedia, with significance calculated by Spearman's two-sided t-test.



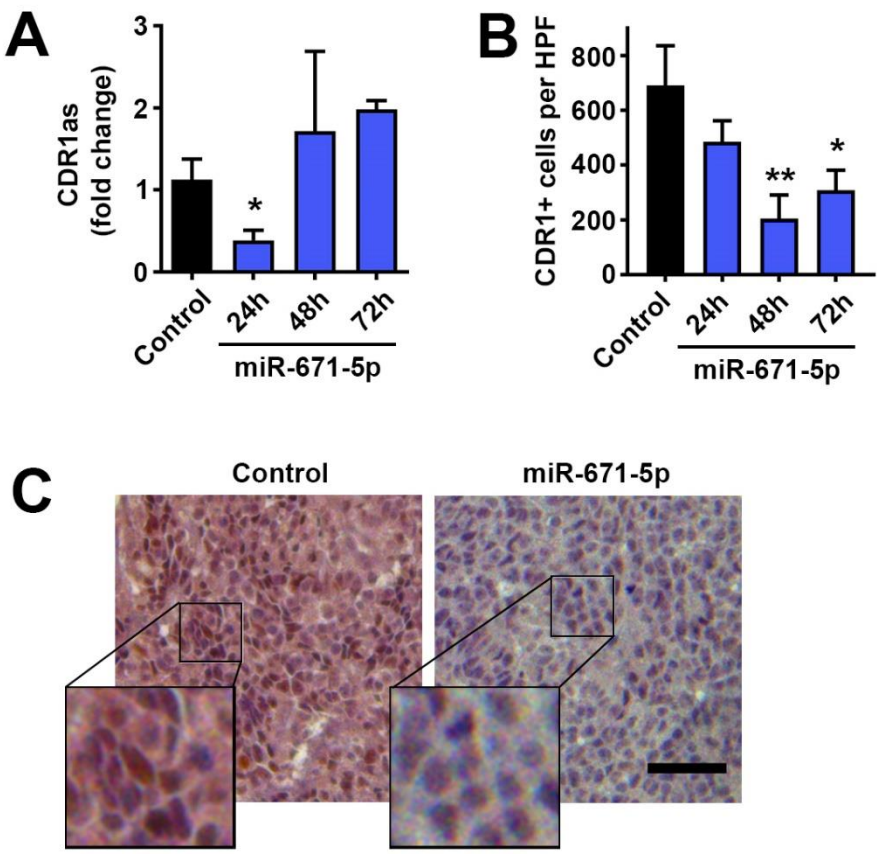
Supplementary Figure 11. Validation of CDR1as and CDR1 knockdown by shRNAs. A and **B**, CDR1as and CDR1 shRNAs were transfected into Hek293T cells; **A**, CDR1as and **B**, CDR1 knock down were assayed by RT-qPCR 48 hours after transfection using divergent and strand specific primers, respectively. Shown are the mean and SEM of a representative experiment (n=3), expression was normalized to 18S, GAPDH, and TBP housekeeping genes. **C**, Correlation between CDR1as and CDR1 expression 48h after shRNA transfection with significance calculated by Pearson's two-sided t-test.



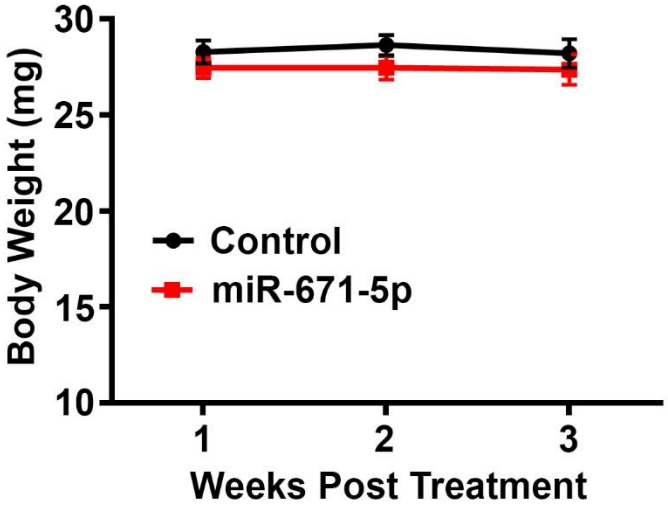
Supplementary Figure 12. Expression of epithelial to mesenchymal (EMT) markers in CDR1 overexpressing tumors. Gene expression was measured in total RNA from SK-MES-1 GFP (n=7) or CDR1 (n=10) LN metastases by qPCR. Expression was normalized to 18S, GAPDH, and TBP housekeeping genes. Shown are the mean ± SEM. ** $p<0.01$, by Student's t-test.



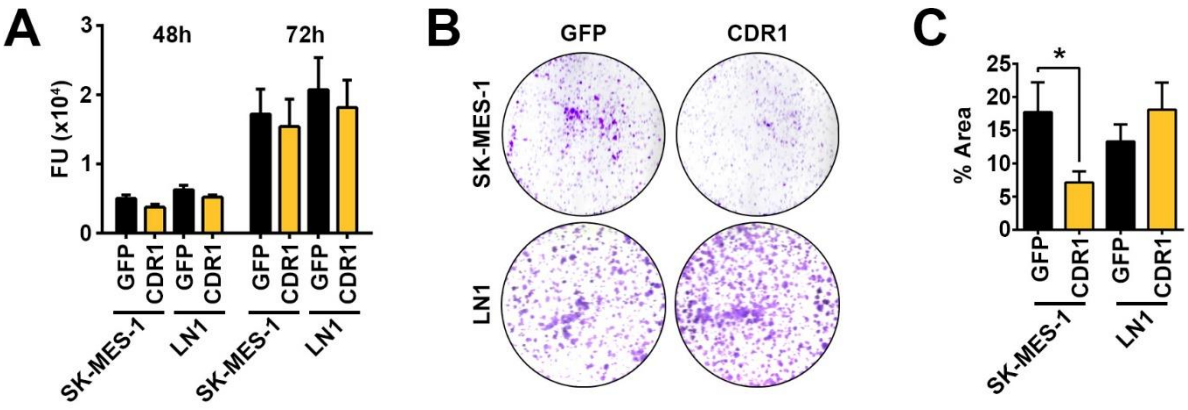
Supplementary Figure 13. Expression of SIG-1R in lung cancer. A, SK-MES-LN1 and H520-LN3 cells were either not treated (-) or treated (+) with a cell impermeable Sulfo-NHS-SS-Biotin reagent. Cell surface proteins were isolated using an avidin column. Flow through (F) and Eluate (E) were analyzed by Western blot for cell surface proteins, EGFR and SIG-1R, and intracellular protein, GM130. B, Expression of various receptors in LUSC patients from TCGA. Data extracted from Oncolnc.org. Significance determined by ANOVA corrected for multiple testing using the FDR, *** $p<0.0001$, n=488. Shown are the mean ± SD.



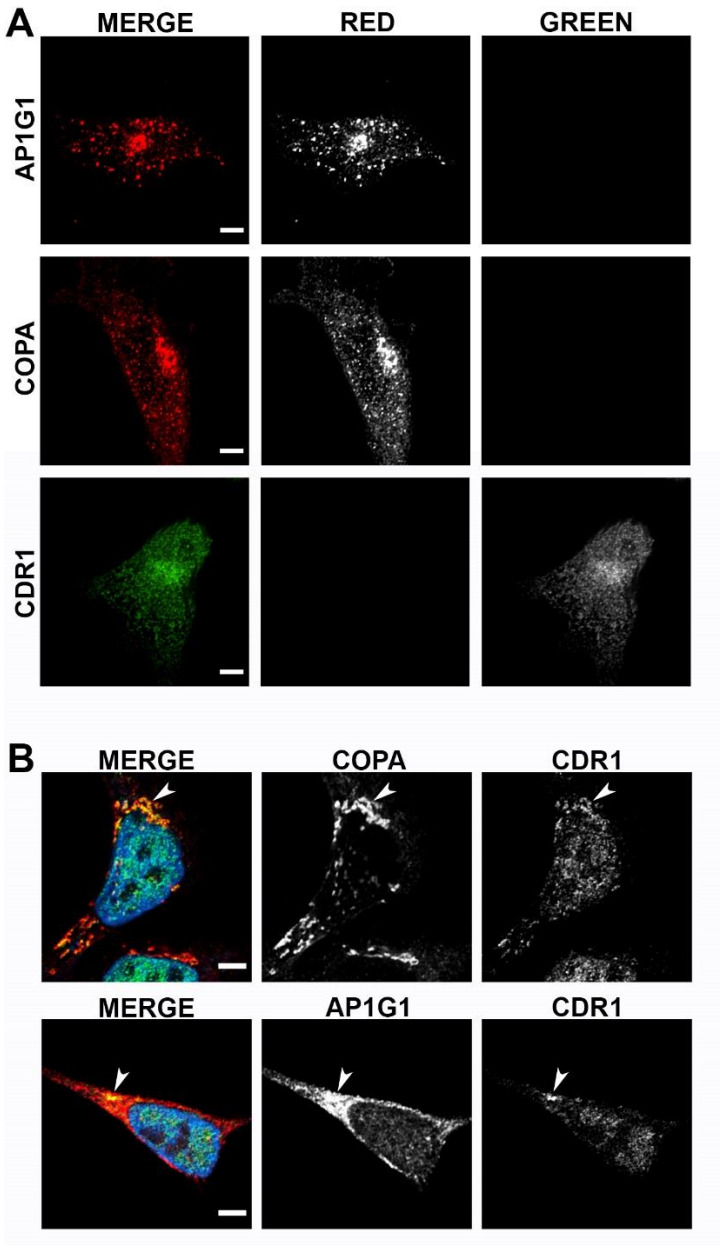
Supplementary Figure 14. Kinetics of CDR1as and CDR1 inhibition by miR-671-5p NPs. A, Expression of CDR1as after administration of control miR (n=5) or miR-671-5p NPs in LN tumors as measured by qPCR with divergent primers at 24 (n=3), 48 (n=3) or 72 (n=2) hours after IV administration. **B**, quantification of CDR1 positive cells by IHC per high powered field after injection of miR control or miR-671-5p NPs 24, 48, or 72 hours after IV administration. * $p<0.05$, ** $p<0.001$ by one-tailed Student's t-test. **C**, Representative IHC staining for CDR1 48h after IV administration of control miR or miR-671-5p NPs; scale bar is 50 µm.



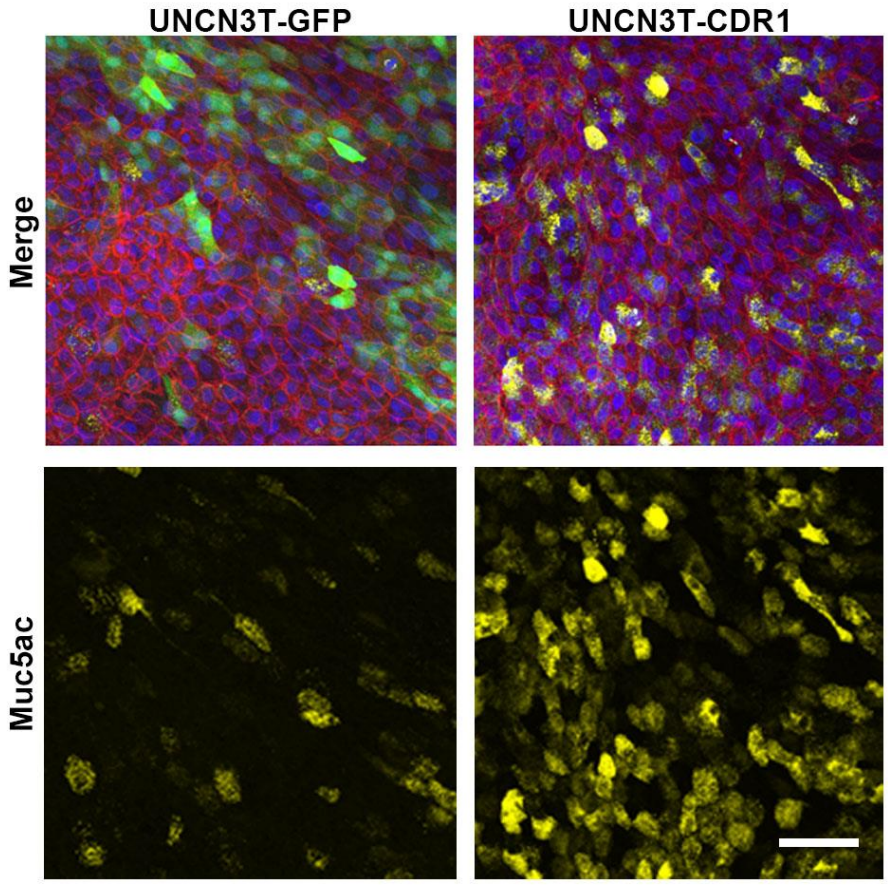
Supplementary Figure 15. Mouse weights during nanoparticle treatment. Weight of mice treated with Control (n=11) or miR-671-5p (n=10) 3x per week for 3 weeks. Shown are the mean ± SEM.



Supplementary Figure 16. CDR1 overexpression and proliferation. A, Proliferation of SK-MES-1 and LN1 stably transduced with GFP or CDR1 as measured by alamar blue assay; shown are the mean ± SEM of 3 independent experiments performed in quadruplicate. **B**, Representative images and **C**, quantification of colony forming assays ; shown are the mean ± SEM of 3 independent experiments performed in triplicate. * $p<0.05$ by Student's t-test.



Supplementary Figure 17. Coimmunostaining for CDR1 and vesicular coat proteins COPA and AP1G1. A, Crosstalk between red and green channels in SK-MES-1-CDR1 cells stained singly for AP1G1, COPA, or CDR1. **B**, Immunocytochemistry for CDR1 (green) and vesicular coat proteins (red) in HCC2814 expressing CDR1 at endogenous levels. Arrows show CDR1 and COPA or AP1G1 localized to perinuclear regions. Scale bars represent 5 μm. Shown is a single slice of a z-stack.



Supplementary Figure 18. Mucin production in air liquid interface cultures (ALIs). Multi-color imaging of UNCN3T human bronchial epithelial cells (HBECs) in ALI cultures. Mucin MUC5AC (yellow), nuclei (blue), alpha tubulin (white), filamentous actin (red), GFP (green). Scale bar is 50 µm.

Supplementary Video 1. Time lapse imaging of RUSH assay in SK-MES-1-EV cells transfected with Ii-Str_ssSBP-EGFP (green) and Golgi marker mApple-SiT (magenta) beginning 5 min after biotin addition.

Supplementary Video 2. Time lapse imaging of RUSH assay in SK-MES-1-CDR1 cells transfected with Ii-Str_ssSBP-EGFP (green) and Golgi marker mApple-SiT (magenta) beginning 5 min after biotin addition.

Supplementary Table 1 (separate file).
Sequences of shRs and RT-qPCR primers.

Supplementary Table 2 (separate file).
miRs individually associated with survival in LUSC TCGA based on median expression of miR genes. FDR values were calculated considering the whole set of genes tested (N=387).

Supplementary Table 3 (separate file).
Expression of miRs in SK-MES-1 and LN1 measured by Nanostring with normalized counts greater than 5 in at least one of the two compared samples.

Supplementary Table 4 (separate file).
Candidate target genes for miR-671-5p, values based on median expression of miR-671-5p and candidate genes.

Supplementary Table 5 (separate file).
Genes differentially expressed in LUSC TCGA in High CDR1-Low miR-671-5p tumors compared to Low CDR1-High miR-671 tumors.

Supplementary Table 6 (separate file).
Gene ontology terms associated with genes elevated in High CDR1-Low miR-671-5p tumors compared to Low CDR1-High miR-671 tumors in LUSC TCGA.

Supplementary Table 7.

EMT gene	r	95% CI	P value
CDH1	-0.2467	-0.3035 to -0.1882	<0.0001
SNAI1	0.1842	0.1242 to 0.2429	<0.0001
SNAI2	0.3131	0.2567 to 0.3675	<0.0001
TWIST1	0.304	0.2472 to 0.3587	<0.0001
TWIST2	0.2759	0.2181 to 0.3317	<0.0001
ZEB1	0.2871	0.2298 to 0.3425	<0.0001
ZEB2	0.271	0.2131 to 0.3270	<0.0001

Correlation of CDR1 expression with EMT genes in the Cancer Cell Line Encyclopedia (1,019 cell lines). All are Two-side Pearson correlations.

Supplementary Table 8.

Cell line	CDR1 and COPA	CDR1 and AP1G1
SK-MES-1-CDR1	0.70±0.05	0.29±0.13
H520-CDR1	0.33±0.16	0.37±0.03
HCC2814	0.35±0.10	0.45±0.06

Colocalization of CDR1 and vesicular proteins, mean and SEM of Pearson's correlation coefficient (n=3).

References:

1. Wilkerson MD, Yin X, Hoadley KA, Liu Y, Hayward MC, Cabanski CR, *et al.* Lung squamous cell carcinoma mRNA expression subtypes are reproducible, clinically important, and correspond to normal cell types. Clin Cancer Res **2010**;16:4864-75
2. Li B, Dewey CN. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics **2011**;12:323
3. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, *et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. P Natl Acad Sci USA **2005**;102:15545-50
4. Porrello A, Leslie PL, Harrison EB, Gorentla BK, Kattula S, Ghosh SK, *et al.* Factor XIIIA—expressing inflammatory monocytes promote lung squamous cancer through fibrin cross-linking. Nat Commun **2018**;9:1988
5. Onder TT, Gupta PB, Mani SA, Yang J, Lander ES, Weinberg RA. Loss of E-cadherin promotes metastasis via multiple downstream transcriptional pathways. Cancer Res **2008**;68:3645-54
6. Hosack DA, Dennis G, Sherman BT, Lane HC, Lempicki RA. Identifying biological themes within lists of genes with EASE. Genome Biol **2003**;4:R70
7. Huang DW, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc **2008**;4:44
8. De Hoon MJ, Imoto S, Nolan J, Miyano S. Open source clustering software. Bioinformatics **2004**;20:1453-4
9. Eisen MB, Spellman PT, Brown PO, Botstein D. Cluster analysis and display of genome-wide expression patterns. P Natl Acad Sci USA **1998**;95:14863-8
10. Saldanha AJ. Java Treeview—extensible visualization of microarray data. Bioinformatics **2004**;20:3246-8
11. Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics **2007**;8:118-27
12. Anaya J. OncoLnc: linking TCGA survival data to mRNAs, miRNAs, and lncRNAs. PeerJ Comput Sci **2016**;2:e67
13. Kumar P, Dezso Z, MacKenzie C, Oestreicher J, Agoulnik S, Byrne M, *et al.* Circulating miRNA biomarkers for Alzheimer's disease. PloS one **2013**;8:e69807
14. Memczak S, Jens M, Elefsinioti A, Torti F, Krueger J, Rybak A, *et al.* Circular RNAs are a large class of animal RNAs with regulatory potency. Nature **2013**;495:333
15. Lamprecht MR, Sabatini DM, Carpenter AE. CellProfiler™: free, versatile software for automated biological image analysis. Biotechniques **2007**;42:71-5
16. Schindelin J, Arganda-Carreras I, Frise E, Kaynig V, Longair M, Pietzsch T, *et al.* Fiji: an open-source platform for biological-image analysis. Nat Methods **2012**;9:676
17. Chono S, Li S-D, Conwell CC, Huang L. An efficient and low immunostimulatory nanoparticle formulation for systemic siRNA delivery to the tumor. J Control Release **2008**;131:64-9
18. Davis S, Charles PD, He L, Mowlds P, Kessler BM, Fischer R. Expanding Proteome Coverage with CHarge Ordered Parallel Ion aNalysis (CHOPIN) Combined with Broad Specificity Proteolysis. J Proteome Res **2017**;16:1288-99
19. Apweiler R, Bairoch A, Wu CH, Barker WC, Boeckmann B, Ferro S, *et al.* UniProt: the universal protein knowledgebase. Nucleic Acids Res **2004**;32:D115-D9
20. Cox J, Mann M. MaxQuant enables high peptide identification rates, individualized ppb-range mass accuracies and proteome-wide protein quantification. Nat Biotechnol **2008**;26:1367
21. Tyanova S, Temu T, Carlson A, Sinitcyn P, Mann M, Cox J. Visualization of LC-MS/MS proteomics data in MaxQuant. Proteomics **2015**;15:1453-6
22. Tyanova S, Temu T, Sinitcyn P, Carlson A, Hein MY, Geiger T, *et al.* The Perseus computational platform for comprehensive analysis of (prote) omics data. Nat Methods **2016**;13:731

23. Teo G, Koh H, Fermin D, Lambert JP, Knight JD, Gingras AC, *et al.* SAINTq: Scoring protein-protein interactions in affinity purification–mass spectrometry experiments with fragment or peptide intensity data. Proteomics **2016**;16:2238-45