LCAL62 supplementary methods

Table of Contents

Genome assembly and annotation

The human reference genome assembly version GRCh37/hg19 and the corresponding gene annotations were used in microarray and RNA-Seq analysis. Gene annotations were combined from Gencode v19 [1], RefSeq downloaded from the UCSC Genome Browser [2], the Broad lncRNA catalog [3], and the miTranscriptome lncRNA catalog [4]. Redundant transcripts were removed and overlapping transcripts were assigned to the same gene. Transcript sequences were generated from the annotation GTF file using BedTools v2.25.0 getfasta [5].

Microarray probe set re-annotation and data acquisition

Microarray probe sequences were downloaded from NCBI GEO [6] and were subsequently aligned to the transcript sequences generated from our annotation file using SeqMap tool [7] allowing one mismatch on the reverse strand. Only probe sets consisting of probes that were uniquely aligned to transcripts from the same genes were retained. Log transformed pre-normalized probe intensity expression matrices were downloaded from NCBI GEO and used for further downstream analysis. The probe sets used for *LCAL62* / *FOXA2* are 228004_at / 40284_at (HG U133 plus 2) and merck2.AI344189_at/merck.NM_153675_at (Rosetta/Merck).

Microarray differential expression analysis

Student t-test was used to calculate p-value and assess if a lncRNA is differentially expressed between metastatic and primary tumors in the Bittner cohort. Fold change was calculated as the ratio of the mean metastasis expression and mean primary tumor expression. LncRNAs with a p-value ≤ 0.05 were considered differentially expressed.

TCGA RNA-Seq data analysis

Pre-aligned RNA-Seq bam files were downloaded from TCGA data portal and imported into the McDonnell Genome Institute Genome Modeling System [8]. Lung adenocarcinoma clinical data (tab format) were also downloaded from TCGA data portal and expression subtype was retrieved from the original publication [9]. FeatureCounts v1.5.0 [10] was used to generate fragment counts for individual transcripts requiring a mapping quality score ≥ 1. FPKM expression was subsequently calculated manually. To obtain the best resolution and reduce the number of tests to perform, for each gene, the transcript with highest FPKM among isoforms were selected to represent a gene locus, similar to our previous approach [11,12]. EdgeR 3.8.6 [13] was used to perform a TMM normalization on raw count data and differential expression analysis between matched primary tumors and normal tissues.

NanoString nCounter probe design and data analysis

Probes were designed by nanoString Technology. Each transcript was targeted by a 100 base probe that uniquely matched with the transcript sequence. Digital transcript counts from nanoString nCounter assays were normalized according to nanoString nCounter data analysis manual using both positive spike-in controls and house-keeping genes. First, the geometric mean of the positive spike-in controls were used to calculate the scale factors for normalization between technical experiments. Next, the geometric mean of the house-keeping genes were used to calculate the normalization factors for normalization between and within biological samples.

Gene coexpression analysis

Spearman correlation coefficient of expression between *LCAL62* and other genes was calculated using the log-transformed values of FPKM (RNA-Seq), normalized probe intensity (microarray), or normalized nanoString nCounter counts.

Survival analysis

Survival analysis was performed using the R survival package v2.37-7 [14] stratifying by tumor stages. Kaplan-Meier curves were plotted using the R survplot package [15]. To reduce non-cancer related death (eg. deaths from surgery complication), patients that died within two months after surgery were removed from the analysis. Patients were stratified into high and low expression using the median expression within a cohort as a cutoff. When cohorts were combined, patient grouping from individual cohorts was used for survival analysis.

Guilt-by-association gene set enrichment analysis

Protein coding genes were ranked by Spearman correlation coefficient of expression with *LCAL62* determined from the TCGA lung adenocarcinoma cohort. To identify gene signatures associated with *LCAL62* expression, gene set enrichment analysis (GSEA) was performed on the ranked list of protein coding genes with 10,000 gene set permutations [16]. Significantly enriched gene sets were determined by false discovery rate (FDR) adjusted P <= 0.001 and normalized enrichment score >= 2.5. Only gene set pairs (ie. up-regulated gene sets and down-regulated gene sets curated from the same study) with either set of the pair that has at least 50 genes were retained. Those enriched gene sets were subsequently manually curated

and assigned to different cancer functional categories.

Cell culture

A549 and HOP62 cells were a kind gift from Dr. Brian VanTine (Washington University). Cells were grown in RPM1-1640 (Invitrogen, Carsbad, CA) with 10% fetal bovine serum (Invitrogen) and 1% penicillin/streptomycin (Invitrogen).

Epithelial-mesenchymal transition (EMT) in cell culture

StemXVivo EMT Inducing Media Supplement (100X) (R&D Systems, Minneapolis, MN) was added at 1X in complete media for five days according to the manufacturer's directions. Media was replaced every three days.

RNA isolation and cDNA synthesis

Total RNA was isolated with the RNeasy Mini Kit (Qiagen, Valencia, CA) with DNase 1 treatment following the manufacturer's instructions. cDNA was synthesized from total RNA using High Capacity cDNA Reverse Transcription Kit with random hexamers (Invitrogen).

Quantitative Real Time PCR

siRNA knockdown was confirmed with quantitative RT-PCR using PowerSyBr Green (Invitrogen). The comparative CT (ΔΔCT) method was used with values first normalized to the housekeeping gene, *RPL32,* and then to the scrambled control sample. All primers were obtained from Integrated DNA Technologies (Coralville, IA). Primers used were:

- *LCAL62* Forward (5'-ACATTTGGTAGCCCGTGGA-3')

- *LCAL62* Reverse (5'-TCTTCCCCGGAGAACTAGCA-3')
- *FOXA2* Forward (5'-GGTGTCTGAGGAGTCGGAGAG-3')
- *FOXA2* Reverse (5'-TTACAGTTCAGACCCGGAACG-3'),
- *RPL32* Forward (5'-AGGCATTGACAACAGGGTTC-3'),
- *RPL32* Reverse (5'-GTTGCACATCAGCAGCACTT-3').

Primer efficiency between 90 -110% was determined for each primer candidate.

Western immunoblotting

Antibodies against E-cadherin (ab15148) and N-cadherin (ab12221) were obtained from Abcam (Cambridge, MA) and used at 1:1000, the antibody against actin (8H10D10) was obtained from Cell Signaling (Danvers, MA) and used at 1:10,000, and Goat anti-Mouse or Goat anti-Rabbit peroxidase conjugated secondary antibodies were obtained from ThermoFisher Scientific (Waltham, MA). Protein samples were prepared with ice-cold lysis buffer (50 mM Tris, pH 8.0, 1% Triton X-100, 131 mM NaCl, 10 mM NaF, 1 mM Na_3VO_4, 10 nM Na4P2O_7 and 1X Halt™ Protease Inhibitor Cocktail (ThermoFisher Scientific) and protein concentration determined with the DC Protein Assay (BioRad, Hercules, CA). Samples were diluted in loading buffer and boiled for 10 min at 70^{o}C. An equal concentration of proteins was loaded and resolved by SDS-polyacrylamide gel electrophoresis in 4-12% Bolt Bis-Tris precast gels (Invitrgoen) and transferred to a nitrocellulose membrane (BioRad). Proteins were detected with specific antibodies and visualized on the ChemDoc MP Imaging System (BioRad) using secondary antibodies and Clarity Western ECL Substrate (ThermoScientific).

Transwell migration and invasion assays

Seventy-two hours after transfection cells were plated at 200,000 cells/well in a modified Boyden chamber assay onto a transwell 8.0µm permeable membrane supports (Corning, Corning, NY) in 24-well plates overnight. A serum gradient was established with cells plated in serum-free media and complete media (10% FBS) added to the bottom of the plate. Cells were then washed, 4% paraformaldehyde (Electron Microscopy Sciences, Hatfield, PA) fixed, and 4',6-Diamidino-2-Phenylindole, Dihydrochloride (DAPI) (Sigma) (1µg/uL) stained. Cells on top of the membrane were swabbed with a cotton swab. DAPI-stained cells were imaged and quantified with Image J software. Four to seven images were taken per transwell membrane at 20X magnification on an Olympus IX51 microscope. For invasion assays, transwell membranes were coated with 200µg/mL Matrigel matrix (Corning) for at least two hours before applying cells to the membrane.

siRNA and shRNA knockdown experiments

Silencer select siRNA oligonucleotides were synthesized by Invitrogen. The following siRNA sequences were used for knockdown of *LCAL62* : *LCAL62* siRNA 1 (5'-GCAGGGUACUUAUUAACCA-3'), *LCAL62* siRNA 2 (5'-GAUUACACCUAUAAUCAGA-3'), *FOXA2* siRNA (5'-UGAACGGCAUGAACACGUA-3'), or a control (a scrambled-matched %GC oligonucleotide synthesized by Invitrogen). Cells were transfected with a final concentration of 6.25 nM of siRNA with RNAimax Lipofectamine (Invitrogen) following the manufacturer's instructions. Stable *LCAL62* knockdown cells lines were made using a retrovirus vector system following Oligoengine's protocol. The following shRNA targeting *LCAL62* was used: shRNA LCAL62 (5'-GGGCCTATAAGCAAACAGT-3').

RNA Immunoprecipitation

Nuclear lysates were collected according to the NE-PER Nuclear and Cytoplasmic Extraction Reagents protocol (ThermoFisher Scientific). RNA immunoprecipitation (RIP) experiments were performed according to Millipore's Magna RIP RNA-Binding Protein Immunoprecipitation Kit protocol with the following changes. Five micrograms of SMAD2/3 (AF3797, R&D Systems, Minneapolis, MN), control Mouse IgG (Millipore, Billerica, MA), control Rabbit IgG (Cell Signaling, Danvers, MA), or positive control SNRNP70 (Millipore) antibody were used. Antibody was incubated with protein G Dynabeads (Invitrogen) prior to addition of 600 µg of nuclear lysate for three hours rotating at 4°C. All washes were done with RIP Wash Buffer (50mM Tris-HCL (pH 7.4), 150mM NaCl, 1mM $MgCl_2$, 1% NP40, 0.005% sodium deoxycholate, 0.05% SDS, 1mM EDTA). Then RNA was extracted following the Millipore protocol and analyzed by quantitative RT-PCR.

Chromatin immunoprecipitation

5e6 cells were crosslinked for 5 minutes with 37% formaldehyde with methanol (Sigma) with rotation at room temperature. The reaction was quenched with 125mM glycine for 5 minutes with rotation. The cells were lysed in 130uL of SDS Lysis Buffer (1% SDS, 10mM EDTA, 50mM Tris HCl pH8, protease inhibitors [PI]). The cells were sheared on a Covaris sonicator with the following parameters: 10% duty cycle, 200 cycles/boost, 140W for 80 secs. Five micrograms of SMAD2/3 (AF3797, R&D Systems, Minneapolis, MN) or control Goat IgG (Millipore, Billerica, MA) antibody were used. Antibody was incubated with sonicated cells overnight with rotation at 4°C. Antibody and cells were diluted to 1mL with ChIP Dilution Buffer (0.01% SDS, 1.10% Triton X-100, 1.2nM EDTA, 16.7mM Tris HCl pH8, 167mM NaCl, PI) Protein G Dynabeads

(Invitrogen) were added for 1h at 10uL of beads per 1ug of antibody). Beads were washed for 5 minutes with rotation at 4°C with the following wash buffers: Low Salt Wash (0.1% SDS, 1% Triton X-100, 2mM EDTA, 20mM Tris HCl, 150mM NaCl, PI); High Salt Wash (0.1% SDS, 1% Triton X-100, 2mM EDTA, 20mM Tris HCl pH8, 500mM NaCl, PI); LiCl Wash (0.25M LiCl, 1% NP-40, 1% sodium deoxycholate, 1mM EDTA, 10mM Tris-HCl pH8, PI); TE Wash (10mM Tris-HCl pH8, 0.1mM EDTA). DNA was eluted with Elution Buffer (1% SDS, 0.1M sodium bicarbonate) with 15 minutes rotation at RT followed by a Proteinase K for 15 minutes at 60°C on a thermomixer. DNA was then phenol chloroform extracted. Primers used were:

- *FOXA2* Primer 1 Forward (5'-CTGGTCGAGCCCCCTTTC-3')
- *FOXA2* Primer 1 Reverse (5'-AGAGGGTGGTTTCCTCCCAG-3')
- *FOXA2* Primer 2 Forward (5'-GTGTCTGAGGAGTCGGAGAGC-3')
- *FOXA2* Primer 2 Reverse (5'-GTTACAGTTCAGACCCGGAACG-3').

Visualization

R software version 3.1.2 [17] and Microsoft Excel were used for plotting graphs. The UCSC Genome Browser was used to visualize transcript structures [2]. Cytoscape 3.1.2 [18] and the Enrichment Map plugin [19] were used to visualize GSEA results.

References

1. Harrow J, Frankish A, Gonzalez JM et al. GENCODE: the reference human genome annotation for The ENCODE Project. Genome Res. 2012; 22(9):1760–1774.

2. Speir ML, Zweig AS, Rosenbloom KR et al. The UCSC Genome Browser database: 2016 update. Nucleic Acids Res. 2016; 44(D1):D717-725.

3. Cabili MN, Trapnell C, Goff L et al. Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. Genes Dev. 2011; 25(18):1915–1927.

4. Iyer MK, Niknafs YS, Malik R et al. The landscape of long noncoding RNAs in the human transcriptome. Nat. Genet. 2015; 47(3):199–208.

5. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinforma. Oxf. Engl. 2010; 26(6):841–842.

6. Barrett T, Troup DB, Wilhite SE et al. NCBI GEO: archive for functional genomics data sets--10 years on. Nucleic Acids Res. 2011; 39(Database issue):D1005-1010.

7. Jiang H, Wong WH. SeqMap: mapping massive amount of oligonucleotides to the genome. Bioinforma. Oxf. Engl. 2008; 24(20):2395–2396.

8. Griffith M, Griffith OL, Smith SM et al. Genome Modeling System: A Knowledge Management Platform for Genomics. PLoS Comput. Biol. 2015; 11(7):e1004274.

9. Cancer Genome Atlas Research Network. Comprehensive molecular profiling of lung adenocarcinoma. Nature 2014; 511(7511):543–550.

10. Liao Y, Smyth GK, Shi W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinforma. Oxf. Engl. 2014; 30(7):923–930.

11. White NM, Cabanski CR, Silva-Fisher JM et al. Transcriptome sequencing reveals altered long intergenic non-coding RNAs in lung cancer. Genome Biol. 2014; 15(8):429.

12. Cabanski CR, White NM, Dang HX et al. Pan-cancer transcriptome analysis reveals long noncoding RNAs with conserved function. RNA Biol. 2015; 12(6):628–642.

13. Robinson MD, Oshlack A. A scaling normalization method for differential expression analysis of RNA-seq data. Genome Biol. 2010; 11(3):R25.

14. Therneau T. A Package for Survival Analysis in S. R package version 2.37-7, 2014.

15. Eklund A. survplot: Plot survival curves with number-at-risk. R package version 0.0.7, 2014.

16. Subramanian A, Tamayo P, Mootha VK et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc. Natl. Acad. Sci. U. S. A. 2005; 102(43):15545–15550.

17. R Core Team. R: A Language and Environment for Statistical Computing, Vienna, Austria: R Foundation for Statistical Computing, 2012.

18. Smoot ME, Ono K, Ruscheinski J et al. Cytoscape 2.8: new features for data integration and network visualization. Bioinforma. Oxf. Engl. 2011; 27(3):431–432.

19. Merico D, Isserlin R, Stueker O et al. Enrichment map: a network-based method for gene-set enrichment visualization and interpretation. PloS One 2010; 5(11):e13984.