Gut Online First, published on February 10, 2018 as 10.1136/gutjnl-2017-314353

ORIGINAL ARTICLE

Comprehensive characterisation of compartment-specific long non-coding RNAs associated with pancreatic ductal adenocarcinoma

Luis Arnes,[1,2] Zhaoqi Liu,[1,2] Jiguang Wang,[2,3] Hans Carlo Maurer,[4] Irina Sagalovskiy,[1,2] Marta Sanchez-Martin,[5] Nikhil Bommakanti,[1,2] Diana C Garofalo,[6] Dina A Balderes,[6] Lori Sussel,[6,7] Kenneth P Olive,[4,8,9] Raul Rabadan[1,2]

► Additional material is published online only. To view please visit the journal online (http://dx.doi.org/10.1136/gutjnl-2017-314353).

For numbered affiliations see end of article.

Correspondence to
Dr Luis Arnes, Department of Systems Biology, Columbia University Medical Center, New York City, NY 10032, USA; la2384@columbia.edu, Dr Kenneth P Olive, Department of Medicine, Division of Digestive and Liver Diseases and Department of Pathology and Cell Biolog, Herbert Irving Comprehensive Cancer Center, Columbia University Medical Center, New York City, New York, USA; kenolive@columbia.edu and Dr Raul Rabadan, Department of Systems Biology, Columbia University Medical Center, New York City, New York, USA; rr2579@cumc.columbia.edu

LA, ZL and JW contributed equally.

Received 19 April 2017
Revised 22 December 2017
Accepted 1 January 2018

To cite: Arnes L, Liu Z, Wang J, *et al*. *Gut* Epub ahead of print: [*please include* Day Month Year]. doi:10.1136/gutjnl-2017-314353

ABSTRACT

Objective Pancreatic ductal adenocarcinoma (PDA) is a highly metastatic disease with limited therapeutic options. Genome and transcriptome analyses have identified signalling pathways and cancer driver genes with implications in patient stratification and targeted therapy. However, these analyses were performed in bulk samples and focused on coding genes, which represent a small fraction of the genome.

Design We developed a computational framework to reconstruct the non-coding transcriptome from cross-sectional RNA-Seq, integrating somatic copy number alterations (SCNA), common germline variants associated to PDA risk and clinical outcome. We validated the results in an independent cohort of paired epithelial and stromal RNA-Seq derived from laser capture microdissected human pancreatic tumours, allowing us to annotate the compartment specificity of their expression. We employed systems and experimental biology approaches to interrogate the function of epithelial long non-coding RNAs (lncRNAs) associated with genetic traits and clinical outcome in PDA.

Results We generated a catalogue of PDA-associated lncRNAs. We showed that lncRNAs define molecular subtypes with biological and clinical significance. We identified lncRNAs in genomic regions with SCNA and single nucleotide polymorphisms associated with lifetime risk of PDA and associated with clinical outcome using genomic and clinical data in PDA. Systems biology and experimental functional analysis of two epithelial lncRNAs (*LINC00673* and *FAM83H-AS1*) suggest they regulate the transcriptional profile of pancreatic tumour samples and PDA cell lines.

Conclusions Our findings indicate that lncRNAs are associated with genetic marks of pancreatic cancer risk, contribute to the transcriptional regulation of neoplastic cells and provide an important resource to design functional studies of lncRNAs in PDA.

INTRODUCTION

Pancreatic ductal adenocarcinoma (PDA) is the third-leading cause of cancer mortality in western countries, and is projected to become the second-leading cause by 2030, surpassed only by lung cancer.[1] The median overall survival of patients with PDA is less than 6 months, and only 8% of patients survive more than 5 years.[2] PDA is characterised by high penetrance mutations in four genes (*KRAS*, *TP53*, *CDKN2A* and *SMAD4*). Large-scale sequencing efforts also identified a host of low-frequency alterations in coding genes, with an average of 60 total per patient. However, while these efforts led to novel molecular classification approaches for the disease, it was learnt that only a low fraction of PDAs harbour 'actionable' mutations.[3–7]

The introduction of unbiased expression profiling technologies like RNA-Seq has resulted in a greater appreciation for the extent of transcription arising from non-coding regions.[8] In particular, transcriptomic analyses across several human tissues and cell lines have identified roughly 10 000 long non-coding RNAs (lncRNAs), which are defined as transcripts longer than 200 nucleotides that lack coding potential.[9 10] To date, only a few cancer-associated lncRNAs have been extensively characterised[11]; however, there is increasing evidence suggesting that these transcripts play a role in tumorigenesis: (1) genome-wide expression analysis showed that lncRNAs are deregulated in cancer and associated with tumour progression[10]; (2) analysis of copy number somatic variations identified lncRNA drivers of tumorigenesis[12 13]; and (3) 93% of the germline variants associated with risk susceptibility identified by genome-wide association studies (GWAS) map to chromatin-modified non-coding regions with the features of lncRNA loci.[14] Indeed, GWAS in PDA identified risk alleles near lncRNAs.[15] In light of these data, we hypothesised that lncRNAs are an integral part of the signalling pathways that regulate the initiation and progression of PDA, and we sought to develop a means to identify lncRNAs that play a role in this challenging disease.

Here, we present Non-coding RNA Identification (NORI), an open source computational tool, to identify lncRNAs using next-generation sequencing. We applied NORI to human PDA and identified lncRNAs that we further annotated with information pertinent to somatic recurrent genomic alterations, cancer-enriched germline variants and clinical prognosis. In addition, we determined epithelial or stromal expression by integrating a cohort of RNA-Seq samples obtained from a large collection of laser-captured microdissected (LCM)

Significance of this study

What is already known on this subject?

- Pancreatic ductal adenocarcinoma (PDA) is one of the most aggressive malignancies, exhibiting only limited and transient responses to current treatments. 'Targetable' alterations in protein coding genes are uncommon in PDA.
- A large fraction of recurrent somatic copy number alterations and single nucleotide polymorphism associated with lifetime risk of cancer are devoid of coding gene drivers of tumorigenesis.
- Long non-coding RNAs (lncRNAs) are emerging as essential players in the biology and progression of a variety of tumours as active regulators of gene expression (non-coding oncogenes or tumour suppressors) and/or passive readouts of tumour progression or clinical prognosis (biomarkers).

What are the new findings?

- We used an integrative analysis of genomic and clinical data from PDA tumour samples to define a catalogue of lncRNAs associated with genetic traits of pancreatic cancer and associated with clinical outcome.
- The identified set of lncRNAs was independently validated in a cohort of paired epithelial and stromal RNA-Seq profiles derived from laser capture microdissected human pancreatic tumours, allowing us to annotate the compartment specificity of their expression.
- LncRNAs segregate tumour samples into subgroups distinguished by differentiation status and associated with clinical prognosis in PDA.
- Using this approach, we identified *FAM83H-AS1* and *LINC00673* in recurrent amplified genomic regions and associated with clinical outcome in PDA.
- We found that loss of *LINC00673* regulates the epithelial differentiation state in PDA cells, increases migratory capacity in vitro and in vivo, and results in loss of epithelial and gain of mesenchymal markers, both in vitro and in tumour samples. This finding is further reflected in poor clinical outcome in low *LINC00673* tumours.

How might it impact on clinical practice in the foreseeable future?

- We expect that the collection of PDA-associated lncRNAs will aid in the design of targeted therapies and may contribute to the development of improved diagnostic tools for PDA. The recent clinical approval of the first antisense therapy for human disease provides a viable, practical approach for leveraging this new understanding of cancer biology.

PDA samples. Functional analysis of two epithelial-enriched PDA-associated lncRNAs validates the significance of the integrative analysis.

MATERIALS AND METHODS

Patients and samples

This study used RNA-Seq data from 147 samples deposited at The Cancer Genome Atlas (TCGA) annotated as pancreas-adenocarcinoma ductal type, based on their neoplasm histologic type. Binary alignment map (BAM) files were downloaded from Cancer Genomics Hub (https://cghub.ucsc.edu/) in February 2016. Detailed clinical information of the samples was downloaded from Cbioportal (http://www.cbioportal.org/study?id=paad_tcga#clinical).

LCM-RNA-Seq profiles were acquired from PDA specimens obtained from patients who underwent surgical resection at the Pancreas Center at Columbia University Medical Center (CUMC). Prior to surgery, all patients had given surgical informed consent, which was approved by the Columbia University Institutional Review Board. Immediately after surgical removal, the specimens were cryopreserved, sectioned and microscopically evaluated by the Columbia University Tumor Bank (IRB protocol AAAB2667). Suitable samples were transferred into optimal cutting temperature compound (OCT) (Tissue Tek) and snap frozen in a 2-methylbutane dry ice slurry. The tissue blocks were stored at −80°C until further processing. H&E stained sections of frozen PDA samples from the Tumor Bank were initially screened to confirm diagnosis and overall sample RNA quality was assessed by the Pancreas Center supported by Next Generation Tumor Banking programme using gel electrophoresis, with samples exhibiting high RNA quality used for subsequent analyses.

Non-coding RNA Identification

We developed NORI to identify lncRNAs from RNA-Seq data. We first used cufflinks[16] to reconstruct the transcriptome profiles on 109 TCGA Pancreatic Adenocarcinoma (PAAD) cohort. Then, individual transcriptome from each sample within the TCGA was merged into a single gene transfer format (GTF) file. Our NORI pipeline extracts a list of lncRNAs from the GTF file by removing transcripts if any of the following criteria were met: (A) they were overlapped with genes annotated in Ensembl and not annotated as 'lincRNA', 'non_coding', 'antisense', '3prime_overlapping_ncrna', 'processed_transcript', 'miRNA', 'misc_RNA', 'polymorphic_pseudogene', 'processed_pseudogene' or 'pseudogene'; (B) overlapped with RefSeq genes (h19) annotated as protein coding, where the RefSeq ID began with 'NM'; (C) overlapped with pseudogenes from Pseudogene.org; (D) monoexonic or less than 200bp in length; (E) predicted to have protein coding potential by the Coding Potential Assessment Tool (CPAT[17]; coding probability >0.364); (F) gene-level maximum reads per kilobase of transcript per million mapped reads (RPKM) was less than $0.05 * n$, where n is the number of samples. NORI is available for download at https://github.com/RabadanLab and accepts a number of optional arguments; see the vignette (browseVignettes ('NORI')) for more details.

Expression data set preparation

For a given molecular feature list, for example, known coding genes/identified lncRNAs from NORI, we used featureCounts from package 'Subread' to call read counts from .bam files (TCGA/CUMC). Genes with low read depths across the cohort are removed. Then, read counts are transformed into RPKM values, followed by log2 transformation, and quantile normalised on sample level.

Molecular subtyping

To identify molecular subtypes with biological relevance, we select a subset of lncRNAs generated by NORI, based on expression correlation with PDA driver mutations. Mutation allele frequency of KRAS, TP53, CDKN2A and SMAD4 was downloaded from Cbioportal (http://www.cbioportal.org/study?id=paad_tcga#mutations). The association between lncRNA expression and mutation allele frequency was assessed by Spearman correlation. We only kept the transcripts from NORI with Spearman q-value <0.001 on at least one of the four drivers, resulting in 652 lncRNAs as input features for clustering. Then, non-negative matrix factorisation (NMF) consensus clustering from GenePattern[18] was employed to identify stable sample clusters on 147 pancreas-adenocarcinoma ductal samples. Detailed clustering parameters are: predefined clusters *k* from 2 to 7, num clusterings 20, max num iterations 2000, error function Euclidean, stop convergence 40 and stop frequency 10. The final number of clusters *k* was selected with the highest cophenetic coefficient.

Additionally, we determine whether the molecular subtyping of lncRNA transcription is an indirect reflection of the transcription of neighbour coding genes. We removed antisense lncRNAs and run-offs of coding genes by filtering lncRNAs that share an overlap with 10 kb up/downstream of a coding gene. NMF was performed as described above for the resulting 354 lncRNAs.

Differential expression analysis

DESeq2 was applied to call differential expressed genes between predefined groups with interests.[19] To generate the input matrix required by DESeq2, featureCounts was used to call raw read counts for both lncRNA and coding genes (GRCh37.75.gtf) on the TCGA cohort.[20]

To investigate the biological content for each subtype defined by NMF clustering, one versus rest comparisons was performed using DESeq2 to identify subtype-specific expressed genes. We also made an additional analysis by comparing lncRNA clusters 1 and 2.

DESeq2 was used to distinguish lncRNAs enriched in the epithelium (n=66) and in the stroma (n=65) using the RNA-Seq from the CUMC cohort.

Overlap of lncRNAs with chromatin annotations of lncRNAs

For chromatin modifications we selected DNAse1 hypersensitivity, H3K27ac, H3K4me1 and H3K4me3 profiles tested on PANC-1 cell line (ENCODE). MACS2 outputted narrowPeak file was downloaded from UCSC Genome Browser for H3K27ac, H3K04me1 and H3K04me3. DNAse1 narrowPeak file was downloaded from Gene Expression Omnibus (GEO GSM736519). Overlapping of chromatin modifications or DNAse1 hypersensitivity with lncRNAs was assessed by calculating the ratio of overlapped lncRNA sequence with respect to the total length that we defined as percentage of overlapped region (POR). Controls were generated by keeping the same length and chromosome distributions as the tested lncRNAs. Significance was calculated by performing 10 000 permutation tests. P value is given by n/10 000, where n is the number of permutations that the control gave a larger POR than lncRNAs.

Next, we analysed the chromatin modifications at the transcriptional start site (TSS) of lncRNAs and coding genes. For each lncRNA/coding gene, the reads depth on each base at the TSS±1 kb was counted from the ChIP-Seq data sets (DNAse1, H3K27ac, H3K4me1 and H3K4me3) from PANC1 cells available at ENCODE. Raw counts are log transformed and averaged for total number of lncRNAs and coding genes, respectively.

Concordant lncRNA expression and somatic copy number alteration status

Somatic copy number alteration (SCNA) segment scores from TCGA PDA samples were downloaded from cBioportal (http://www.cbioportal.org/). The correlation between segment scores and lncRNA expression was determined by Spearman correlation on patient level. Next, to determine the significance, we compared the lncRNA expression-SCNA correlations with random controls. To generate the controls, we randomly selected 85 lncRNAs and arbitrarily matched their expression with the original SCNA segment scores.

Survival analysis

Overall survival and disease-specific survival records of TCGA PAAD cohort were downloaded from Cbioportal (http://www.cbioportal.org/study?id=paad_tcga#clinical). Survival analysis was conducted using the R package 'survival' and 'survcomp'. To demonstrate the clinical relevance of NMF clusters, patient survival in different subtypes was compared using Kaplan-Meier survival curves together with the log-rank test. To evaluate the prognostic power of an lncRNA on a specific sample set, we first split the cohort into two equal-sized subsets based on the median expressing value of the lncRNA, and then compare the survival differences between the two using the log-rank test.

Regulatory network

The pancreatic cancer regulatory network was reverse engineered by ARACNe-AP[21] from the TCGA PAAD cohort. The RNA-Seq level 3 data were downloaded from TCGA data portal, raw counts were normalised to account for different library sizes after filtering out genes with less than one fragment per million mapped fragments in at least 20% of the samples, and the variance was stabilised by fitting the dispersion to a negative binomial distribution as implemented in the DESeq2 R package (Bioconductor). ARACNe was run with standard settings (using data processing inequality (DPI), with 100 bootstrap iterations using all gene symbols mapping to a set of 1813 transcription factors that includes genes annotated in the Gene Ontology (GO) molecular function database as GO:0003700 ('transcription factor activity'), GO:0004677 ('DNA binding'), GO:0030528 ('transcription regulator activity') or as GO:0004677/GO:0045449 ('regulation of transcription')). In addition to these coding gene products, we added the 453 lncRNAs to the list of transcriptional regulators. Thresholds for the tolerated DPI and mutual information P value were set to 0 and 10–8, respectively.

Cell lines and transfection

PANC1, BxPC3, MiaPaCa2 and Aspc1 were passaged and maintained following standard techniques in 5% CO_2% and 95% air cultured following manufacturer instructions (American Type Culture Collection ATCC). PANC1 cells that constitutively express the firefly luciferase gene (PANC1/Luc1) were generated by lentiviral transduction of a bicistronic expression vector with dtomato and luciferase (Addgene plasmid number 48688).

Cells were transfected with 5 nM siRNA targeting *FAM83H-AS1* or *LINC00673* and a non-targeting control (Silencer select, Ambion) using Lipofectamine 3000 following manufacturer's instructions (Life Technologies). siRNA sequences are listed in online supplementary table 1.

For *SOX9* overexpression, *SOX9* cDNA sequence (Origene, NM_000346) was cloned into pcDNA3. Cells were transfected with *SOX9* and with an empty pcDNA3 vector as control. Transfection was done using Lipofectamine 3000 following manufacturer's instructions (Life Technologies). Cell lines were purchased and verified by ATCC, maintained at low passage and tested for mycoplasma.

Splenic injection and live imaging

PANC1/Luc was transiently transfected 48 hours prior surgery with either control or targeted siRNA by using Lipofectamine 3000 (Invitrogen). Cells were harvested by trypsinisation and a single cell suspension of 2×10^6 cells in 100 µL was prepared in phosphate buffered saline (PBS) and kept on ice before injection. Cells were injected into the spleen of 6–7 weeks of age NU/J male mice (002019, The Jackson Laboratories), keeping syringe inside spleen for 5 min after injection to allow tumour cells to drain. Tumour growth was monitored by bioluminescence imaging using the In Vivo Imaging System (IVIS) whole body imaging system. Luciferin substrate was given by intraperitoneal injection 10 min before imaging. Bioluminescence flux was quantified using live imaging software (Perkin Elmer). All procedures were approved by the Ethics Committee of Columbia University.

Cell cycle analysis

PANC1 cells were fixed in 70% ethanol 48 hours after transfection with the indicated siRNA and were stained with propidium iodide (final concentration 1 µg/mL). Cells were sorted by Fluorescence-activated cell sorting (FACS) and cell cycle distribution was analysed with FlowJo.

Western blot

Cell were lysed and proteins extracted following the whole-cell extract from adherent cells protocol provided with the Nuclear Extract Kit (Active Motif). Primary antibodies are listed in online supplementary table 1.

RNA extraction and qRT-PCR analysis

Total RNA was isolated from cultured cell lines using the RNeasy Mini Kit (Qiagen). RLT buffer was supplemented with 2-mercaptoethanol (Sigma-Aldrich) and DNase treatment was performed for 20 min using the RNase-Free DNase set (Qiagen). About 1 µg of total RNA was reverse transcribed into cDNA using random hexamers with SuperScript III First-Strand Synthesis kit (Life Technologies). About 20 ng of cDNA was used in the qRT-PCR reaction with iQ SYBR Green supermix (Bio-Rad) and custom-designed primers. All experiments were calculated as a function of gene expression relative to either control *TBP* expression or *GAPDH*. qPCR data were expressed as mean fold change ($2^{-\Delta\Delta CT}$). Primers are listed in online supplementary table 1.

RNA-Seq in PANC1 and ASPC1

About 200 ng of total RNA from culture cell lines was subjected to transcriptome analysis. Total RNA was converted into cDNA libraries (TruSeq RNA Sample Prep Kit v2, Illumina) using poly-A pull down for mRNA enrichment. Sequencing was performed to a depth of 30 million pairs. Differential expression between replicates was assayed using DESeq2 (R package). All samples had RNA integrity number (RIN) values higher than 9.0 as determined with Agilent Bioanalyzer 2100. Complete RNA-Seq data are available through GEO Express GSE96931.

Gene set enrichment analysis

We used gene set enrichment analysis (GSEA) software with defined gene sets (MSigDB, v6.1) to investigate molecular profiles enriched before and after targeting the expression of *LINC00673* and *FAM83H-AS1*. Overlap with SOX9 target genes was performed using the gene set CATTGTYY_SOX9_B1 from the Broad Institute. Phenotype labels were created for siCTRL and siRNA. siRNA1 and siRNA2 versus siCTRL were assessed independently. Genes with Fragments Per Kilobase of transcript per Million mapped reads (FPKM)=0 across all samples studied were removed from the gct file for GSEA analysis. Parameters used were *Collapse data: false*, *Permutation type: gene_set*, *1000 permutations*, *Chip platform: gene symbol.*

Gene set variation analysis

We used the R implementation of single sample GSEA: gene set variation analysis with default parameters 1. The input expression matrix was filtered for the most variable 75% of the genes. For annotation of epithelial subtypes, we tested a set of gene sets shown previously to be discriminating between molecular subtypes of PDA.[3 7] Differential enrichment analysis of these gene sets between the different siRNA treatments was carried out using the limma R package 4 and an false discovery rate FDR<0.1 was considered significant.

Clonogenic assay

Human pancreatic cancer cell lines were transfected with targeting siRNAs and siCTRL in triplicate. Forty-eight hours after transfection, cells were detached and $5-10\times10^3$ cells were seeded in triplicate in six well plates and incubated to allow colony formation for 10–14 days. Colonies were visualised by staining with crystal violet and quantified with image J (http://rsb.info.nih.gov/ij/).

Migration assay

Cells pretransfected 48 hours before with siCTRL or siRNA-1 were seeded (50×10^3) onto transwell membrane inserts in 2% fetal bovine serum (FBS) culture media (5 µm pore, Corning). Regular media was added to the lower chamber and cells were incubated for 12 hours at 37°C. After incubation, cells that migrated across the membrane were fixed and stained with crystal violet. For each membrane, the same nine randomly distributed fields were counted. The data represent the mean of three independent experiments performed in triplicate.

Immunofluorescence

Cells were platted onto coverslips and fixed with 4% paraformaldehyde for 5 min and 100% methanol for additional 5 min. After fixation, coverslips were kept at 4°C submerged in PBS. For immunofluorescence, coverslips containing fixed cells were blocked in 5% serum in 1×PBS+0.5% Triton-X for 30 min. Primary antibodies were diluted in blocking buffer and incubated overnight at 4°C. The primary antibody used was rabbit α-Vimentin (1:500; ab92547, Abcam). Coverslips were incubated with appropriate secondary antibodies conjugated to DyLight-488 (Jackson Immunoresearch). DAPI (1:1000; Invitrogen) was applied for 30 min following secondary antibody incubation.

Code availability

The NORI framework is fully available for academic use on Github (https://github.com/RabadanLab/NORI).

RESULTS

Identification of lncRNAs in PDA

We developed NORI, a computational approach to annotate and characterise the non-coding transcriptome using next-generation sequencing data (figure 1A). We applied NORI to reconstruct the non-coding transcriptome of 109 PDA human samples deposited at TCGA. We identified 3433 lncRNAs including several described as drivers of tumour progression or as functional regulators of organ development such as *UCA1*,[22] *DEANR1*,[23] *PVT1*,[13] *NBR2*,[24] *MALAT1*[25] and *NEAT1* (online supplementary table 2).[26] These data confirm the capability of NORI to

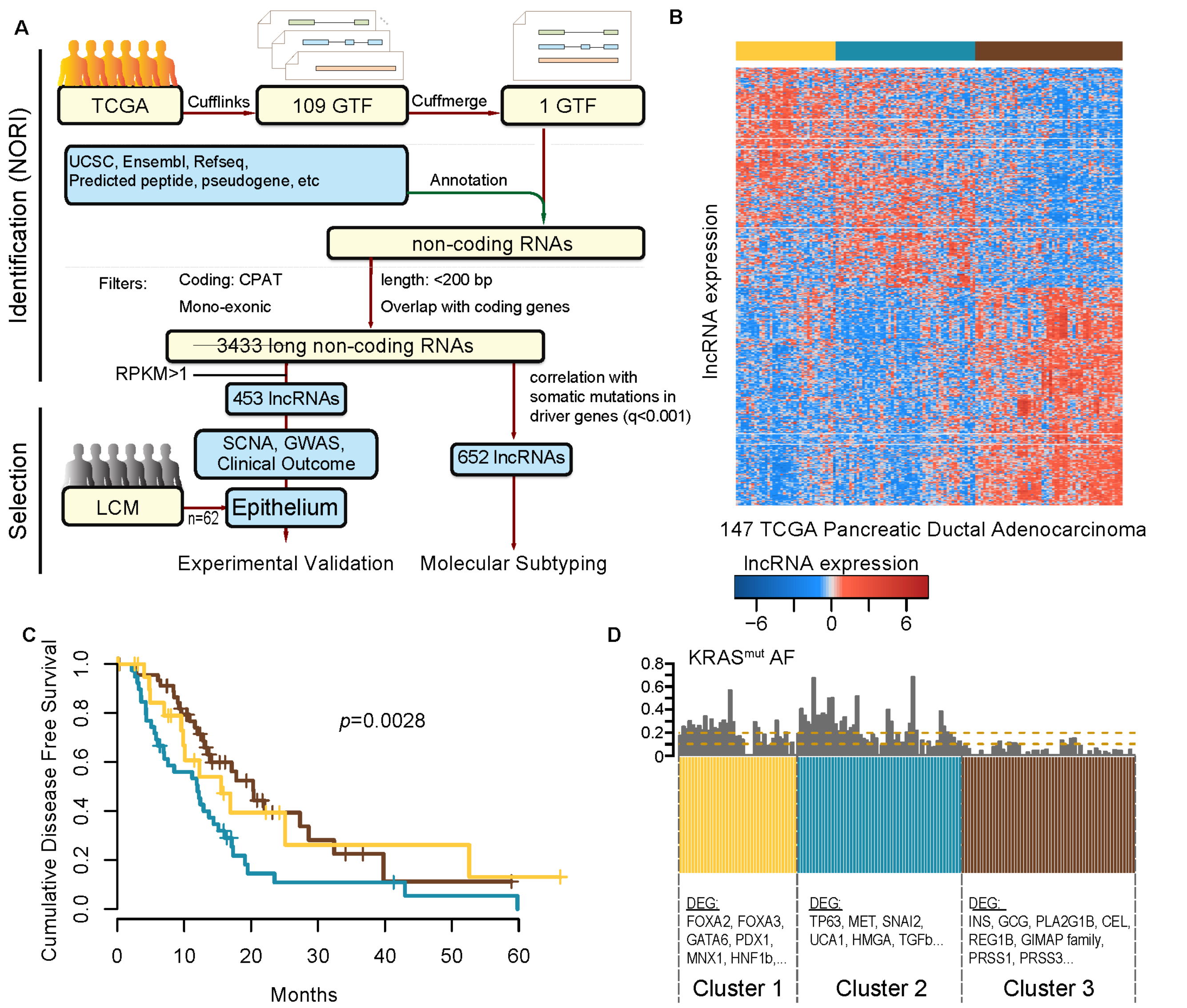

Figure 1 Identification of lncRNAs and molecular subtyping of pancreatic ductal adenocarcinoma (PDA). (A) Schematic representation of the computational analysis. NORI identified 3433 lncRNAs expressed in PDA using RNA-Seq from a cohort of 109 tumours from TCGA. The output of NORI was a subset into abundant lncRNAs (RPKM>1) prioritised for experimental validation, and lncRNAs whose expression correlates ($q<0.001$) with the allele frequency of PDA driver genes for the identification of molecular subtypes in PDA by non-negative matrix factorisation (NMF). Abundant lncRNAs were annotated with the genomic distance to recurrent SCNA and/or single nucleotide polymorphisms (SNP) associated with PDA risk and with the expression correlation with clinical outcome. In addition, an independent cohort of LCM PDA samples (n=66 epithelium, 65 stroma) was analysed to validate expression of lncRNAs in PDA and to select epithelial lncRNAs for functional analysis. (B) NMF using the expression of lncRNAs identified three molecular subtypes in the TCGA cohort (n=147). (C) Kaplan-Meier disease-free survival estimations for the individual subtypes. (D) Differential gene expression analysis between molecular subtypes. Relevant genes are shown (see online supplementary table 4 for full list). Each TCGA sample is colour coded according to the molecular subtype. $KRAS^{mut}$AF is depicted as an independent estimation of tumour cellularity of each sample. AF, allele frequency; CPAT, Coding Potential Assessment Tool; GTF, gene transfer format; GWAS, genome-wide association studies; LCM, laser-captured microdissected; lncRNAs, long non-coding RNAs; NORI, Non-coding RNA Identification; RPKM, reads per kilobase of transcript per million mapped reads; SCNA, somatic copy number alterations; TCGA, The Cancer Genome Atlas; UCSC, University of California Santa Cruz.

reconstruct non-coding transcriptomes, and constitute an initial set of PDA-associated lncRNAs as the basis for further analysis.

Molecular subtyping defined by the expression of lncRNAs

Global analyses of coding transcripts in PDA have been used to dissect molecular subtypes of the disease.[3 6 7] We investigated whether lncRNAs might prove useful for the classification of molecular subtypes independent of coding genes. Because expression profiles from TCGA samples represent a mixture of both malignant epithelial cells and stromal cells, we used a computational approach to focus our classification effort on lncRNAs whose expression correlates with the allele frequency of the major drivers in PDA (*KRAS*, *TP53*, *CDKN2A* and *SMAD4*), which are mutated in 97% of the TCGA samples (see the Materials and methods section, online supplementary table 3). We used the resulting 652 genes to define molecular subtypes, applying NMF to 147 PDA samples from the TCGA. This analysis revealed the presence of three molecular subtypes, with a cophenetic coefficient of 0.9931 (figure 1B, online supplementary figure 1). To determine whether molecular subtyping was an indirect effect of the transcription of neighbour coding genes, we removed lncRNAs that were within 10 kb of coding genes and still obtained very similar clustering ($P=7.01\ 10^{-50} \chi^2$ test, online supplementary figure 2). Clusters 1 and 2 were associated with elevated mutant *KRAS* allele frequencies, while tumours in cluster 3 were associated with low frequencies, perhaps indicating that this group emerged because of the inclusion of higher amounts of stroma and/or infiltrated normal tissues. Analysis of outcomes data available through the TCGA indicated an association of tumours from cluster 2 with reduced disease-free survival relative to those in clusters 1 and 3 (log-rank test, P=0.0028) (figure 1C).

To understand the biological significance, we performed differential gene expression analysis (figure 1D, online supplementary table 4). In cluster 1, we observed enrichment of transcription factors necessary in pancreas development, including *FOXA2*, *FOXA3*, *GATA6*, *GATA4*, *PDX1*, *MNX1*, *HNF1b*, *HNF4g* and *HNF4a*.[3] Many of these genes, as well as others involved in lineage specification, are enriched in the previously identified 'Classical' molecular subtype, which was associated with improved overall survival relative to the 'Basal-like' subtype. Notably, cluster 2 included several genes found in the Basal-like subtype and associated with epithelial to mesenchymal transition (EMT), including *TP63*, *CAV1*, *SNAI2*, *MET*, *HMGA2* and *TGFβ*. Finally, in cluster 3, we observed genes related to digestive processes (*PLA2G1B*, *PRSS1*, *PRSS3*), endocrine function (*INS*, *GCG*, *SST*) and the immune system (*CD48*, *CCR2*, GIMAP proteins), suggesting that this subtype is defined by the contributions of non-neoplastic cell types. Together, these findings demonstrate that lncRNAs reflect the heterogeneity of biological processes in PDA with relevance to clinical outcomes.

Annotation of lncRNAs in PDA

We next sought to prioritise the 3433 lncRNAs expressed in PDA. First, we filtered out low-abundance transcripts, yielding 453 lncRNAs with a mean expression >1 RPKM. Next, we investigated the association of lncRNAs with relevant genomic and clinical features of the disease, including: (1) proximity to PDA-associated coding genes[27]; (2) location within SCNA in PDA; (3) proximity to germline variants identified in GWAS; and (4) correlation with clinical outcome data (figure 2A, online supplementary table 2). As expected, analysis of chromatin modifications at these loci found significant enrichment of accessible chromatin regions related to transcriptional activity (figure 2B, online supplementary figure 3; P<0.001, permutation test). Nonetheless, this analysis allowed us to discriminate between *bona fide* lncRNAs and spurious transcripts or sequencing artefacts.

Somatic copy number alterations

SCNAs are frequently the subject of clonal selection during tumour progression. However, many SCNAs lack known coding tumour drivers,[28 29] perhaps suggesting that unknown non-coding drivers may be located within these genomic regions. Analysis of SNP data by GISTIC2 in the TCGA cohort revealed 56 recurrent SCNAs, 23 amplifications and 33 deletions, including known events in PDA such as amplification of *GATA6*, *KRAS* and *MYC*; and deletions of *CDKN2A* and *SMAD4*. We examined the overlap of SCNA genomic locations in PDA with our candidate lncRNAs and observed that 85 of 453 lncRNAs were located within the 56 SCNAs identified in the same cohort of patients (online supplementary table 2). Significantly higher lncRNA expression-SCNA correlations were found in the 85 paired lncRNAs than in random controls (figure 2C). Among those lncRNAs, we detected the expression of *PVT1*,[13] *LINC-PINT*[30] and the antisense lncRNA for *GATA6*. Given the paucity of established cancer-associated lncRNAs, the identification of several such genes in PDA-associated SCNAs serves as conceptual validation for this approach.

Germline variants associated with PDA

GWAS studies have identified 14 genetic loci associated with increased or reduced lifetime risk of PDA.[15] Interestingly, four of these loci map to genomic regions known to contain functional lncRNAs (*PVT1*, *LINC-PINT*, *PDX1-AS1* and *LINC00673*) while several others mapped to unannotated non-coding regions.[31] We performed a similar analysis and identified five candidate lncRNAs within loci that harboured somatic SNP variants associated with increased risk of PDA (online supplementary table 2). We detected expression of *PVT1*, *LINC-PINT* and *LINC00673*. In addition, we identified a novel lncRNA located on chromosome 9q34.2 near the *ABO* gene that is associated with PDA risk; and *LINC01829*, located on chromosome 2p14 upstream of *ETAA1*, a recently characterised protein involved in DNA damage signalling.[32] Furthermore, analysis of chromatin features and topological-associated domains (TAD) in PANC1 cells showed chromatin interactions with distant SNPs, suggesting a cis-regulation (figure 2D, online supplementary figure 4). Overall, we identified expression of five PDA lncRNAs that are near SNPs associated with increased risk of pancreatic cancer.

Association with clinical outcome

We also examined the association between the expression of each lncRNA and overall or disease-free survival in the TCGA cohort. We observed that 23 (5.3%) and 36 (7.9%) of 453 lncRNAs were significantly (P<0.05) associated with overall survival and disease-free survival, respectively (online supplementary table 2). This is similar to the fraction of coding genes that are associated with survival and progression differences in PDA (5.7% and 4.1%, respectively), further supporting a possibility of a functional role for lncRNAs in driving PDA.

In summary, we have generated a large-scale resource describing lncRNAs expressed in PDA annotated with information on their expression level, association with genomic landmarks and association with clinical outcomes. We expect this resource will serve as an aid for functional studies on the contribution of lncRNAs to the progression of pancreatic cancer.

Independent validation of PDA lncRNAs using compartment-specific expression data

Previous studies using both correlative and functional analyses have established that lncRNAs can act as regulators of oncogenic pathways in malignant cells.[33 34] However, bulk tumour RNA-Seq contains a mixture of transcripts originating from neoplastic epithelial cells and non-neoplastic stromal cells. To explore the cellular origins of the candidate lncRNAs, we performed RNA-Seq on LCM samples from 65 patients with human PDA who underwent resection at the Pancreas Center

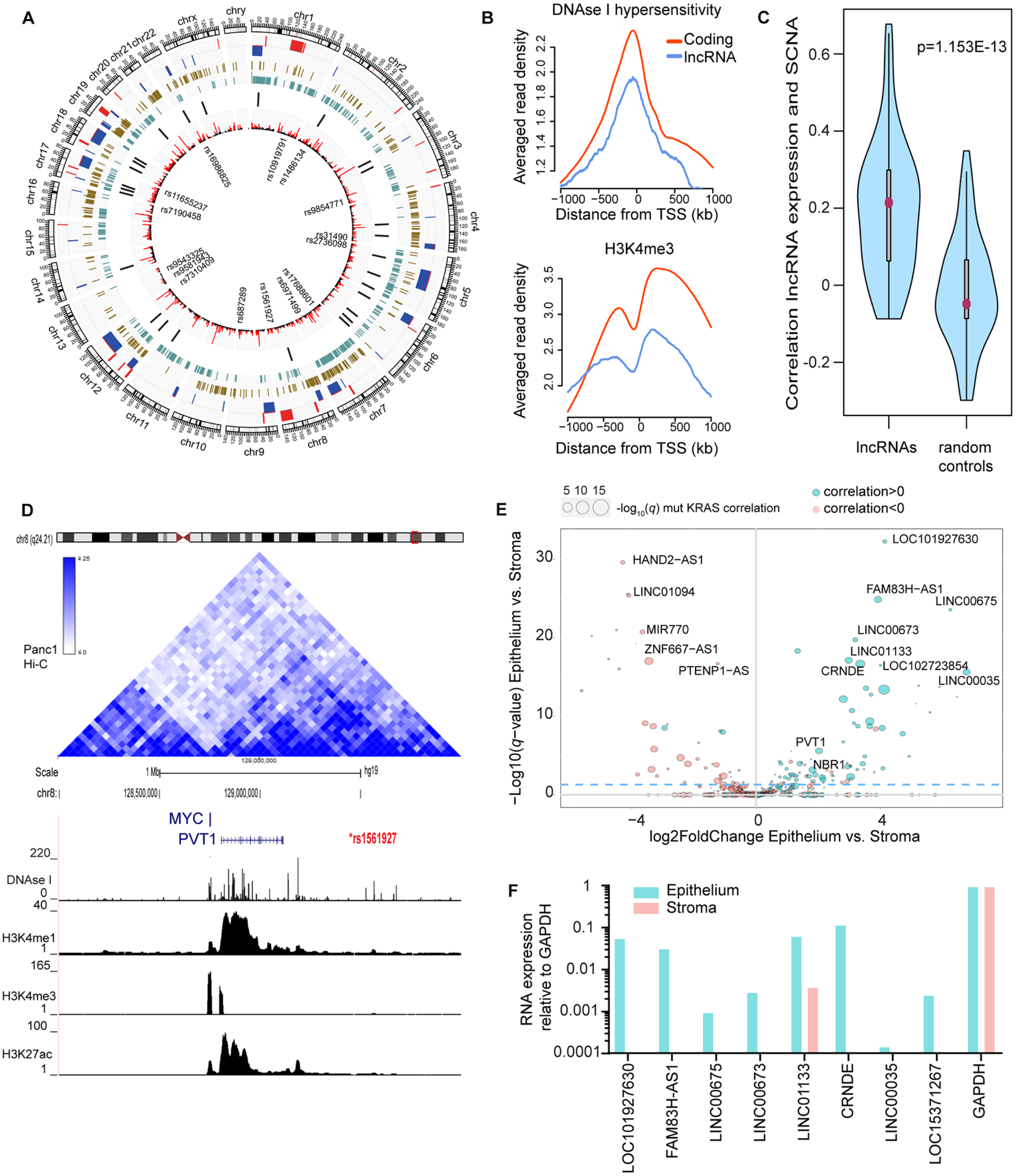

Figure 2 Annotation of lncRNA with genomic threats of pancreatic cancer and identification of epithelial or stromal expression. (A) Circos plot depicting location of lncRNAs respective to genomic marks associated with pancreatic ductal adenocarcinoma (PDA). From inner to outer: single nucleotide polymorphisms (SNP) associated to lifetime risk of PDA; lncRNAs identified by Non-coding RNA Identification (NORI) (red: expression >1 reads per kilobase of transcript per million mapped reads (RPKM)); location of PDA-associated cancer genes described in online supplementary table 2; DNase I hypersensitivity and H3K4me3 in PANC1 cells; recurrent SCNA in the The Cancer Genome Atlas (TCGA) cohort, amplifications (red) and deletions (blue). The outermost ring shows the chromosomes in clockwise order with sex chromosome at the end. Full annotation of lncRNAs is provided in online supplementary table 2. (B) Averaged reads density of DNAseI signal (upper) and H3K4me3 (lower) along the TSS region of ±1 kb, summarised for lncRNA and coding genes, respectively. Reads depth are log transformed and averaged on each base. (C) Comparison of expression-SCNA correlations on 85 lncRNAs with random controls. P value is calculated from Wilcoxon rank-sum test. (D) University of California Santa Cruz (UCSC) snapshot of the *PVT1* locus, location of the SNP associated with lifetime risk of PDA (red) and topological associated domains (TAD) in PANC1 cells indicative of higher order of genome organisation. The genomic regions are overlapped with DNAseI hypersensitivity and epigenetic marks of active transcription in PANC1 cells (ENCODE data). For clarity, only *PVT1* and *MYC* are shown. (E) Scatter plot showing distribution of lncRNAs according to epithelial and stromal expression as determined by laser-captured microdissected (LCM) RNA-Seq data (n=131). In addition, as an independent metric for expression in neoplastic epithelium, the size of each circle represents the correlation of lncRNA expression with the allele frequency of KRAS mutation. (F) Validation of epithelial enrichment for the top epithelial lncRNAs. Analysis performed in a pool of epithelial and stromal samples from the Columbia University Medical Center (CUMC) cohort. n=3 technical replicates. Only the 8 out of 10 candidates that were validated are shown. Expression relative to GAPDH. GAPDH, glyceraldehyde 3-phosphate dehydrogenase; lncRNAs, long non-coding RNAs; SCNA, somatic copy number alterations; TSS, transcriptional start site.

at the CUMC (He *et al*, in revision). Notably, we detected 80% of the 453 candidate PDA lncRNAs in the CUMC cohort, providing an independent validation set for our earlier analysis. We annotated each lncRNA with its relative enrichment in the stroma or epithelium of PDA samples (online supplementary table 2). We identified 138 compartment-enriched lncRNAs ($q<0.05$, DESeq2) of which 94 are enriched in the epithelium and 44 in the stroma (figure 2E, online supplementary table 5). The expression of the top 10 epithelial and stromal candidates as detected by RNA-Seq is depicted in online supplementary figure 5. We validated the expression of 8 of the 10 epithelial candidates by orthogonal methods in a pool of six random epithelial and stromal samples from the CUMC cohort (figure 2F). Consistent with previous clustering results, we identified two molecular subtypes in the CUMC cohort that were functionally characterised by gene sets associated with differentiation state (online supplementary table 6). Overall, these analyses validate the expression of selected lncRNAs in two independent PDA cohorts and confirm our non-coding transcriptome reconstruction approach. Together, these data provide a rich resource of validated PDA-associated lncRNAs annotated with information on their compartment of origin.

Functional validation of epithelial lncRNAs

Next, we sought to study the functional roles of top candidate PDA-associated lncRNAs. We focused our analysis on lncRNAs enriched in the epithelium as potential regulators of molecular pathways altered in malignant cells. In addition, we selected lncRNAs located in genomic regions that have been associated with focal amplification or deletions as potential drivers of tumour progression. We became interested in *FAM83H-AS1* and *LINC00673*.

FAM83H-AS1 is the antisense of *FAM83H*, a recently described coding gene required for the organisation of the keratin cytoskeleton in epithelial cells.[35] The two genes share a promoter region but transcribed in opposite directions (figure 3A). *FAM83H-AS1* is located on chromosome 8 on a genomic region frequently amplified in PDA (8q23.3-8q24.3). We found that *FAM83H-AS1* expression correlates significantly with amplification ($r=0.67$, $P=6.5\times10^{-21}$, online supplementary figure 6). RefSeq gene annotation and ENCODE data indicate that it has four exons and that it is located in an actively transcribed chromatin region in PANC1 (figure 3A, online supplementary figure 7). Importantly, high expression of *FAM83H-AS1* showed a borderline association with poor clinical outcome ($P=0.056$, figure 3B) and *FAM83H-AS1* expression across a panel of cancer cell lines showed elevated expression in PDA lines (figure 3C). We explored the function of *FAM83H-AS1* by targeting the gene with two siRNAs in Aspc1 cells, each of which resulted in >60% knock-down efficiency (figure 3D). RNA-Seq analysis with transient knock-down of *FAM83H-AS1* identified 1309 and 2721 differentially expressed genes with siRNA1 and siRNA2, respectively, and principal component analysis (PCA) clustered samples according to the siRNA (figure 3E). There was significant overlap in the sets of genes differentially expressed in response to the two siRNAs (Fisher's exact test, $P<2.2\times10^{-16}$), as expected (figure 3F). GSEA of common dysregulated gene sets suggests a less aggressive phenotype when *FAM83H-AS1* is downregulated, consistent with our observation of a worse prognosis for pancreatic tumours expressing high levels of

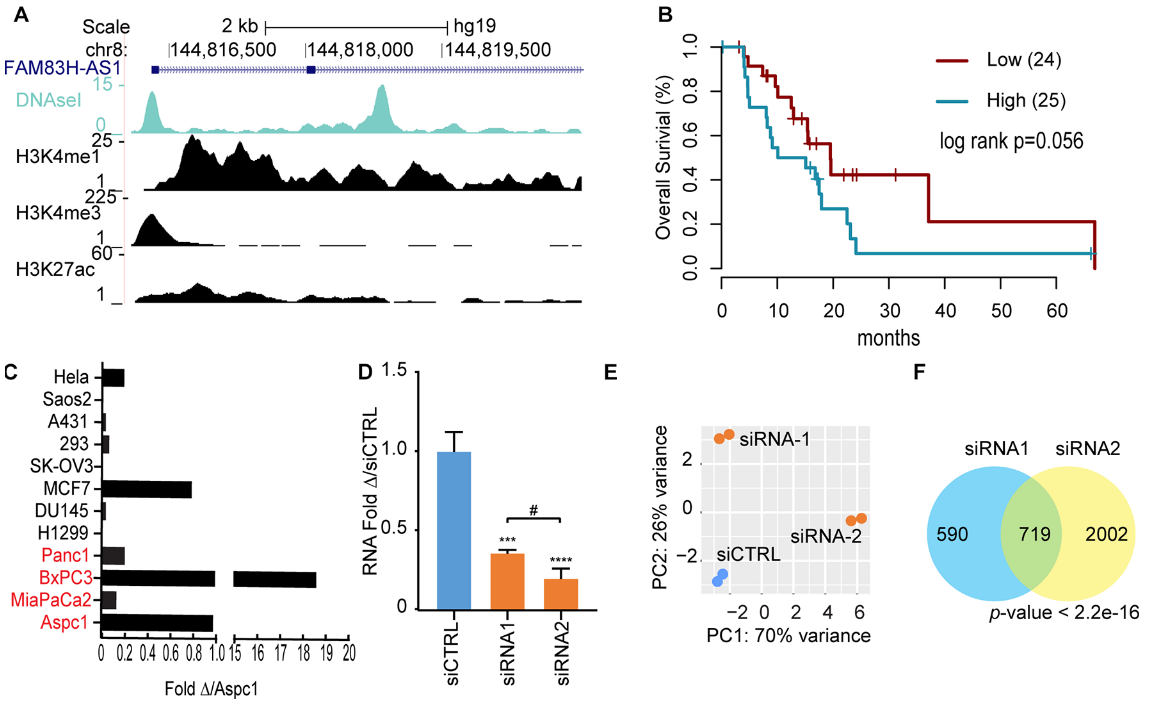

Figure 3 *FAM83H-AS1* regulates the transcriptome profile of Aspc1 cells. (A) University of California Santa Cruz (UCSC) snapshot of the *FAM83H-AS1* transcriptional start site (TSS) depicting DNase I hypersensitivity and chromatin modifications in PANC1 cells (ENCODE). (B) Kaplan-Meier overall survival estimations for samples with high and low expressions of *FAM83H-AS1*. Only samples with KRAS^{mut}allele frequency (AF)>0.2 were considered. The two groups were defined by partitioning the samples into two equal-sized sets using the median value of *FAM83H-AS1* expression. (C) *FAM83H-AS1* expression across a panel of cell lines. Normalised with glyceraldehyde 3-phosphate dehydrogenase (*GAPDH*) and relative to the expression in Aapc1. Pancreatic cancer cell lines depicted in red. (D) *FAM83H-AS1* RNA expression after transient transfection of Aspc1 cells with two different siRNAs. (E) Cluster of RNA-Seq samples by principle component analysis. (F) Overlap of dysregulated genes (Padj<0.05) with individual siRNAs. Fisher's exact test.

FAM83H-AS1 (online supplementary figure 8A, online supplementary table 7).

As a complementary approach to understand the functions of this lncRNA, we used an information theory-based systems biology technique called regulatory network analysis. Briefly, a regulatory network delineates interactions between regulatory genes (ie, genes whose activity alters RNA transcript abundance) and their target genes. We used the ARACNe algorithm[36] to reconstruct de novo a regulatory network from TCGA expression data in an unbiased manner (online supplementary table 8), providing a list of inferred target genes for known transcription factors as well as the 453 lncRNAs defined above. The network comprised over 300 000 total interactions between 1813 total regulatory genes and 453 lncRNAs. In particular, ARACNe inferred 146 potential target genes for *FAM83H-AS1*, including 78 positive (activated) targets and 68 negative (inhibited) targets. Computed overlap with gene sets (MSigDB v6.1) found that genes inferred to be negatively regulated by *FAM83H-AS1* were associated with more benign processes while positive inferred targets of *FAM83H-AS1* were associated with more malignant processes (online supplementary figure 8B, online supplementary table 8). Together, these data are consistent with a role for *FAM83H-AS1* in promoting tumour progression.

Another top candidate identified in our analysis is *LINC00673*, a transcript that is among the most epithelial-enriched PDA-associated lncRNAs. *LINC00673* is located in a recurrent, focally amplified region in PDA and is linked to a PDA-associated SNP. It is located approximately 275 kb telomeric of *SOX9* (online supplementary figure 7). Notably, *SOX9* is a well-known transcription factor expressed in multipotent pancreatic progenitor cells and is required for the neoplastic transformation of PanIN lesions in a mouse model of PDA.[37] ChIP-Seq analysis from ENCODE in PANC1 cells showed enrichment of chromatin modifications associated with active transcription at the promoter of *LINC00673* in PANC1 cells (figure 4A). In addition, we detected that *LINC00673* expression correlates with SCNA ($r=0.39$, $P=1.3\times10^{-6}$, online supplementary figure 6) and high expression of *LINC00673* in TCGA PDA samples is significantly associated ($P=0.050$) with better survival (figure 4B). Expression analysis across various cell lines showed widespread *LINC00673* expression in PDA cells (figure 4C).

To test the function of *LINC00673*, we silenced its expression in PANC1 cells through transient transfection with two different siRNAs. Each siRNA efficiently downregulated the expression of *LINC00673*, although to different levels (figure 4D). To gain insight into the mechanism of action of *LINC00673*, we profiled gene expression in PANC1 cells following *LINC00673* silencing. The resulting RNA-Seq profiles clustered by treatment group by PCA (figure 4E). Notably, there was a significant overlap in the sets of genes dysregulated by the two siRNAs (Fisher's exact test, $P<2.2\times10^{-16}$) with a larger total number of differentially expressed in cells treated with the more effective siRNA (figure 4F), consistent with a dose-dependent effect. Critically, GSEA performed on the differentially expressed genes from each siRNA revealed many overlapping processes, particularly pathways related to EMT in downregulated genes (online supplementary figure 9A, online supplementary table 9). Next, we applied regulatory network analysis to identify candidate target genes for *LINC00673*. We inferred 123 potential targets of *LINC00673*, including 91 positive and 32 negative targets.

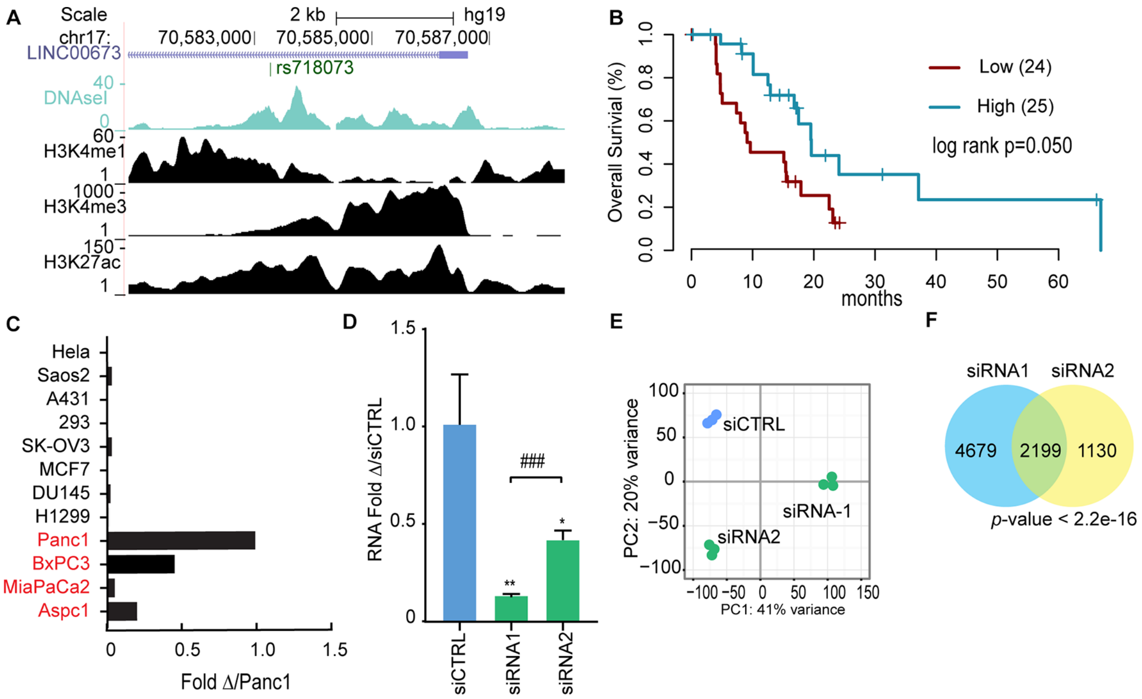

Figure 4 *LINC00673* regulates the transcription profile of pancreatic cancer cells and is necessary to maintain epithelial features. (A) University of California Santa Cruz (UCSC) snapshot of the *LINC00673* locus as described for *FAM83H-AS1* in figure 3. (B) Kaplan-Meier overall survival estimations for tumour samples with high and low expressions of *LINC00673*. Only samples with $KRAS^{mut}$allele frequency (AF)>0.2 were considered. The two groups were defined by partitioning the samples into two equal-sized sets using the median value of *LINC00673* expression. (C) *LINC00673* expression across a panel of pancreatic ductal adenocarcinoma (PDA) cell lines. Normalised with glyceraldehyde 3-phosphate dehydrogenase (GAPDH) and relative to the expression in PANC1. PDA cell lines depicted in red. (D) *LINC00673* RNA expression after transient transfection of PANC1 cells with two different siRNAs. (E–G) RNA-Seq was performed in PANC1 cells transiently transfected with two different siRNAs and a non-targeting control. (E) Principal component analysis. (F) Overlap of dysregulated genes (Padj<0.05) with both siRNAs. Fisher's exact test.

Overlap of candidate targets with gene sets from MSigDB demonstrated enrichment for gene sets related to maintenance of epithelial properties and downregulation of metastasis (positive) and downregulation of cell-to-cell communication (negative) (online supplementary figure 9B, online supplementary table 8), consistent with the results from cell lines.

Functional analysis of cells transiently depleted of *LINC00673* by siRNA shows impaired colony formation, and this effect cannot be explained by defects in cell cycle (online supplementary figure 10). Interestingly, the less efficient siRNA2 produced an intermediate phenotype, again supporting a dose dependency in the function of *LINC00673* (figure 5A). These results were reproduced in two additional human PDA cell lines, MiaPaca2 and BxPC3 (online supplementary figure 11). Furthermore, we observed that knock-down of *LINC00673* resulted in increased cell motility in several PDA cell lines (figure 5B, online supplementary figure 12), consistent with the effects of EMT. In addition, we assessed the metastatic potential of PDA cells after LINC00673 downregulation by injecting PANC1/Luc cells pretreated with siCTRL or siRNA1 in the spleen of nude mice. We determined that PANC1 cells were more efficient in producing metastatic lesions when *LINC00673* was downregulated (figure 5C). Overall, these data suggest that *LINC00673* regulates in vitro and in vivo the metastatic potential of PDA cell lines.

Examining the genes dysregulated by *LINC00673* silencing, we found that the most strongly upregulated gene was *MET* (figure 5D), a receptor tyrosine kinase involved in motility, migration and invasion in PDA cell lines.[38] We confirmed this finding by qRT-PCR and observed that upregulation of *MET* was associated with downregulation of epithelial markers such as *FOXA1* and *CDH1* (figure 5D). In addition, loss of *LINC00673* induced a mesenchymal phenotype evidenced by gain of vimentin expression (figure 5E,F, online supplementary figure 13). We were unable to confirm a role of *LINC00673* in the stimulation of extracellular signal-regulated kinase signalling pathway in *PDA* cell lines, as previously suggested (online supplementary figure 14).[39] Genetic downregulation of MET did not prevent the increased migratory capacity of PANC1 cells, suggesting that a global transcriptome switch or loss of differentiation status mediates this aggressive behaviour (online supplementary figure 15). Consistently, we observed enrichment of squamous (log2 fold change 0.86, Padj=0.03) and quasimesenchymal subtype (QM, log2 fold change 0.70, Padj=0.07) molecular classifiers after downregulation of *LINC00673* with siRNA1 in PANC1 cells (figure 5G). No significant enrichment was found with the less efficient siRNA2, overall suggesting that *LINC00673* is required to maintain epithelial differentiation and prevent expression of mesenchymal markers.

Since *SOX9* is located in close proximity to *LINC00673,* we hypothesised that *SOX9* could influence *LINC00673* function. We did not detect changes in *SOX9* protein following *LINC00673* knock-down (online supplementary figure 14); however, we observed significant dysregulation of SOX9 target genes, particularly the classical target and known mediator of pancreas differentiation, *FOXA1* (figure 5H).[38] Overexpression of *SOX9* resulted in a significant upregulation of *FOXA1* that was partially abrogated by *LINC00673* downregulation (figure 5I, online supplementary figure 16). These data suggest that *LINC00673* participates in the functional regulation of SOX9. Supporting this hypothesis, analysis of the potential SOX9 target genes in PDA identified overlap with gene sets related to the maintenance of epithelial features such as cell-cell junction, apical junction complex and epithelium development (online supplementary table 8), consistent with the downregulation of *FOXA1* and *CDH1* mediated by *LINC00673*.

Overall, we provided experimental and clinical evidence suggesting that loss of *LINC00673* induced a loss of epithelial differentiation in PDA cells, and this is reflected in poor clinical outcome in low *LINC00673* tumours, increased migratory capacity in vitro and in vivo, and loss of epithelial and gain of mesenchymal markers in vitro and in tumour samples.

DISCUSSION

Our understanding of the role of lncRNAs is rapidly evolving from spurious expression of 'junk DNA' to the current understanding that many such genes play a direct functional role in biology. As a whole, a consensus is beginning to emerge that lncRNAs act as critical modulators of cellular regulatory states. In particular, many lncRNAs contribute to the process of lineage specification, a critical need in the evolution of complex organisms with hundreds or thousands of discrete cell types.

Cancer is also intimately linked to cellular differentiation; indeed, loss of differentiation is one of the cardinal features of the disease, both at the point of initiation and during tumour progression. Individual examples of lncRNAs with functional roles, such as *XIST*[40] and *HOTAIR,*[41] serve as important proofs of principle for the potential contributions of lncRNAs. Likewise, analysis of recurrent SCNA identified *FAL1*, an oncogene frequently amplified in ovarian cancer[42]; *PVT1*, coamplified with *MYC*[13]; *SAMMSON*, frequently amplified in melanoma and required for mitochondrial function.[12] However, overall such efforts have been limited to small numbers of lncRNAs,[43] to small numbers of samples,[44] or by the constraints of array-based technologies.[45] In general, global analysis of lncRNAs has been hindered by a paucity of tools for their identification, annotation and prioritisation. We believe the open-source NORI tool and other computational approaches used here provide a useful framework for investigating lncRNAs in cancer using publicly available RNA-Seq data such as that available through TCGA.

In this study, we presented an analysis of lncRNAs expressed in PDA that we validated in two independent cohorts of PDA samples. We used NORI to reconstruct the global expression of lncRNAs and then applied a series of computational criteria to probe their association with features of pancreatic malignancy. For example, localisation of candidate PDA lncRNAs to SCNA is of interest because many such regions selected during tumour evolution are devoid of known coding oncogenes or tumour suppressor.[28 29] Likewise, germline variants associated with risk susceptibility identified by GWAS very frequently map to non-coding regions. Global gene expression analysis shows that lncRNAs in trait-associated loci are expressed in cell types relevant to the trait, again suggesting a role of lncRNAs in disease.[46] By applying multiple such criteria, we prioritised those lncRNA candidates with the highest likelihood of playing a functional role in PDA.

It was therefore notable to us that many of the resulting PDA-associated lncRNAs identified were related at some level to pancreatic lineage specification. In particular, one of our top candidates, *LINC00673*, is located next to *SOX9* and regulates the expression of several SOX9 target genes. Although we did not detect changes in SOX9 expression mediated by *LINC00673*, lncRNAs have been found to interact with related protein SOX2 to regulate downstream targets.[47] We also detected lncRNAs located in proximity to *GATA6* and *FOXA2*, both important transcription factors involved in pancreas development. These

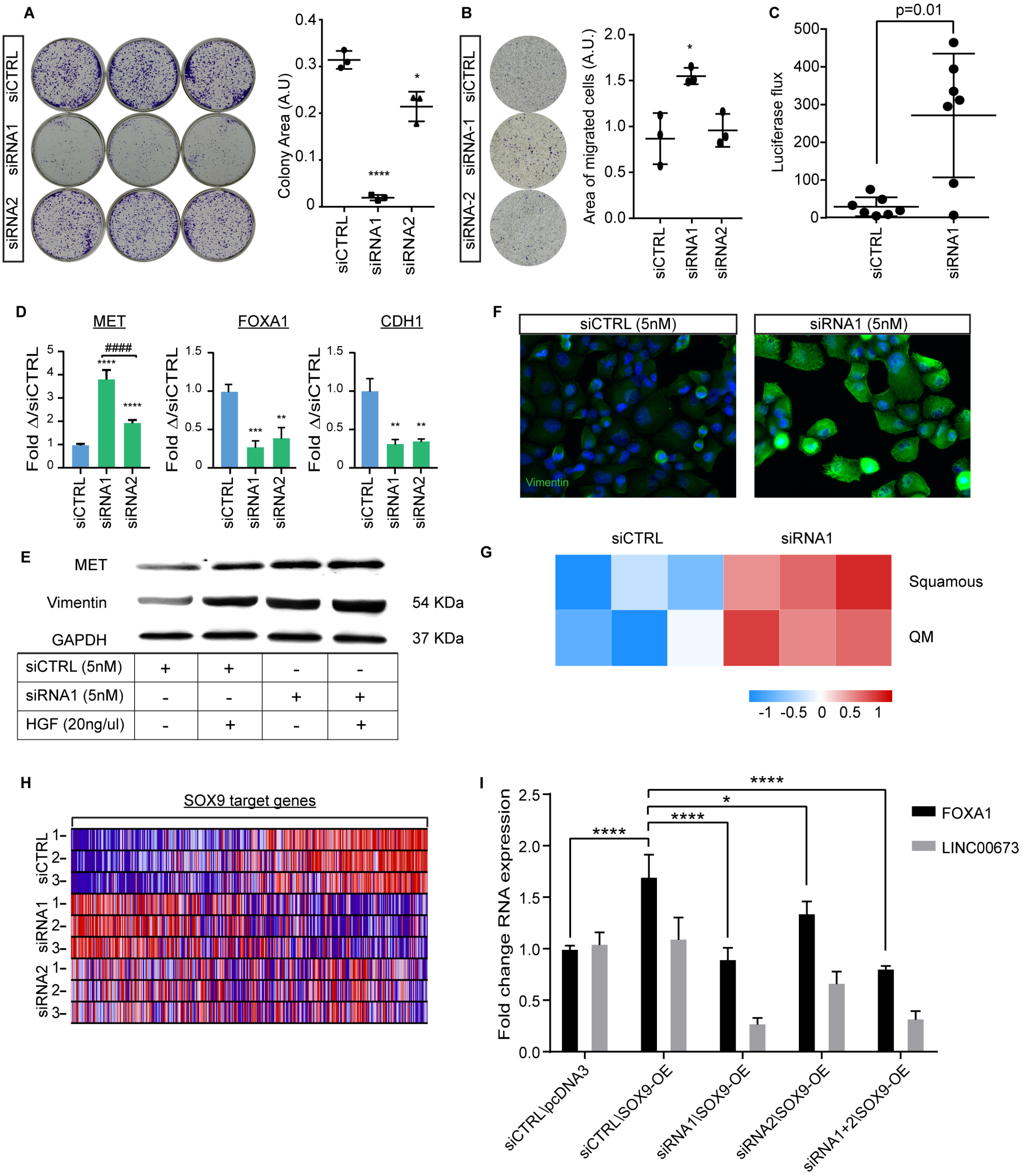

Figure 5 Transient inhibition of *LINC00673* leads to loss of epithelial markers and EMT. (A) PANC1 colony formation assay performed with the indicated siRNA and visualised with crystal violet. n=3 independent experiments with two different siRNAs. Student's t-test. (B) PANC1 migration assay in 5 µm transwell membranes. n=4 independent experiments with two different siRNAs. Student's t-test. (C) Higher metastatic burden in nude mice after splenic injections of PANC1/Luc cells transfected with siRNA1 targeting *LINC00673* for 48 hours prior to surgery. P=0.017 Mann-Whitney U test. (D) *MET*, *FOXA1* and *CDH1* mRNA expression in PANC1 cells treated with two different siRNAs against *LINC00673*. n=3. Student's t-test. (E) Western blot of MET and vimentin after transient knock-down of *LINC00673*. HGF treatment (20 ng/µL) included as positive control. Representative blot of at least three independent experiments. (F) Immunofluorescence analysis of vimentin expression in PANC1 cells transfected with siRNA1. Representative images of at least three independent experiments. (G) Molecular subtyping using Bailey and Collisson classifiers of PANC1 cells before and after *LINC00673* knock-down. (H) RNA expression after knock-down of *LINC00673* of genes containing SOX9 binding sites at the promoter. (I) *FOXA1* mRNA expression in PANC1 cells overexpressing *SOX9* and *LINC00673* knock-down. Error bars represent ±SD. EMT, epithelial to mesenchymal transition; GAPDH, glyceraldehyde 3-phosphate dehydrogenase; QM, quasimesenchymal.

findings are consistent with the paradigm that loss of developmental genes affecting the terminal differentiation of epithelial cells contributes to tumour progression, both generally[48–50] and specifically in the pancreas.[51] Furthermore, our data are consistent with previous reports showing that lineage determinants, such as GATA6 and PDX1, are amplified or reactivated at the initiation of PDA (and thus behave as oncogenes); however, their expression may be lost during tumour progression, perhaps contributing to the subsequent loss of epithelial character.[48 52] Together with early evidence suggesting that reprogramming of chromatin domains is associated with the acquisition of metastatic clones,[53] our data support a model in which lncRNAs play a role in the metastatic progression of PDA.

Our focus on lncRNAs expressed in the neoplastic epithelium was greatly facilitated by a unique RNA-Seq data set derived from LCM human pancreatic tumours, which also served as a means for independent validation of candidate lncRNA expression. However, we do expect that lncRNAs will also play an important role in the biology of stromal cells. For instance, among the top enriched stromal lncRNAs, we identified *HAND2-AS1*, a promoter-associated lncRNA, that has been recently shown in a knockout mouse model to regulate the expression of *HAND2*,[54] which is involved in paracrine stroma-to-epithelium signalling in the uterus.[55] Further investigation of these stroma-enriched lncRNAs is warranted.

LncRNAs are emerging as essential players in the biology and progression of cancer as active regulators of coding gene expression. Critically, the recent Food and Drug Administration approval of the first antisense therapy provides a viable, practical approach for leveraging this new understanding of cancer biology. Nusinersin (Spinraza, Biogen) is an antisense oligonucleotide directed against the splice junction of *SMN2*, a paralogue of the *SMN1* that is mutated in spinomuscular atrophy. This and other clever strategies may be employed to modulate lncRNA function in vivo, providing a whole new tool set to control and reverse the regulatory states that drive malignancy.

Author affiliations
[1]Department of Systems Biology, Columbia University Medical Center, New York City, New York, USA
[2]Department of Biomedical Informatics, Columbia University Medical Center, New York City, New York, USA
[3]Division of Life Science and Department of Chemical and Biological Engineering, Hong Kong University of Science and Technology, Hong Kong, China
[4]Department of Medicine, Division of Digestive and Liver Diseases, Columbia University Medical Center, New York City, New York, USA
[5]Institute for Cancer Genetics, Columbia University Medical Center, New York City, New York, USA
[6]Department of Genetics and Development, Columbia University Medical Center, New York City, New York, USA
[7]Barbara Davis Center, University of Colorado, Denver, Colorado, USA
[8]Department of Pathology and Cell Biology, Columbia University Medical Center, New York City, New York, USA
[9]Herbert Irving Comprehensive Cancer Center, Columbia University Medical Center, New York City, New York, USA

Acknowledgements We thank Steve Sastra and Christopher Damocci for excellent assistance with the xenograft model.

Contributors Conceptualisation: LA. Computational analysis: ZL and JW. Software: NB. Investigation: LA, ZL, HCM, IS, MSM, DCG and DAB. Resources: LA, KPO and RR. Visualisation: LA and ZL. Funding acquisition: LA. Project oversight and management: LA, KPO and RR. LA wrote the manuscript with feedback from KPO and RR. All authors discussed the results and commented on the manuscript.

Funding This work was funded by the IRIS (LA) and CaST (LA) programmes at Columbia University, the Juvenile Diabetes Research Foundation APF-2014-197-A-N and FAC-2018-541-A-N (LA) and NIH R21CA188059 (LS). These studies used the resources of the Herbert Irving Comprehensive Cancer Center (Center Grant P30CA013696) and the Diabetes and Endocrinology Research Center (Center Grant 5P30DK063608). RR and ZL were funded by the NIH U54 CA193313. NB was funded by an NIH-NHLBI T35 training grant. JW was funded byN_HKUST606/17 and C6002-17G

Competing interests None declared.

Ethics approval Columbia University Institutional Review Board.

Provenance and peer review Not commissioned; externally peer reviewed.

Data sharing statement All data used in this manuscript will be available through GEO Express GSE96931 after publication.

REFERENCES

1 Rahib L, Smith BD, Aizenberg R, *et al*. Projecting cancer incidence and deaths to 2030: the unexpected burden of thyroid, liver, and pancreas cancers in the United States. *Cancer Res* 2014;74:2913–21.
2 Siegel RL, Miller KD, Jemal A, *et al*. Cancer statistics, 2017. *CA Cancer J Clin* 2017;67:7–30.
3 Bailey P, Chang DK, Nones K, *et al*. Genomic analyses identify molecular subtypes of pancreatic cancer. *Nature* 2016;531:47–52.
4 Jones S, Zhang X, Parsons DW, *et al*. Core signaling pathways in human pancreatic cancers revealed by global genomic analyses. *Science* 2008;321:1801–6.
5 Witkiewicz AK, McMillan EA, Balaji U, *et al*. Whole-exome sequencing of pancreatic cancer defines genetic diversity and therapeutic targets. *Nat Commun* 2015;6:6744.
6 Moffitt RA, Marayati R, Flate EL, *et al*. Virtual microdissection identifies distinct tumor- and stroma-specific subtypes of pancreatic ductal adenocarcinoma. *Nat Genet* 2015;47:1168–78.
7 Collisson EA, Sadanandam A, Olson P, *et al*. Subtypes of pancreatic ductal adenocarcinoma and their differing responses to therapy. *Nat Med* 2011;17:500–3.
8 Birney E, Stamatoyannopoulos JA, Dutta A, *et al*. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* 2007;447:799–816.
9 Djebali S, Davis CA, Merkel A, *et al*. Landscape of transcription in human cells. *Nature* 2012;489:101–8.
10 Iyer MK, Niknafs YS, Malik R, *et al*. The landscape of long noncoding RNAs in the human transcriptome. *Nat Genet* 2015;47:199–208.
11 Schmitt AM, Chang HY. Long noncoding RNAs in cancer pathways. *Cancer Cell* 2016;29:452–63.
12 Leucci E, Vendramin R, Spinazzi M, *et al*. Melanoma addiction to the long non-coding RNA SAMMSON. *Nature* 2016;531:518–22.
13 Tseng YY, Moriarity BS, Gong W, *et al*. PVT1 dependence in cancer with MYC copy-number increase. *Nature* 2014;512:82–6.
14 Maurano MT, Humbert R, Rynes E, *et al*. Systematic localization of common disease-associated variation in regulatory DNA. *Science* 2012;337:1190–5.
15 Childs EJ, Mocci E, Campa D, *et al*. Common variation at 2p13.3, 3q29, 7p13 and 17q25.1 associated with susceptibility to pancreatic cancer. *Nat Genet* 2015;47:911–6.
16 Trapnell C, Williams BA, Pertea G, *et al*. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotechnol* 2010;28:511–5.
17 Wang L, Park HJ, Dasari S, *et al*. CPAT: Coding-Potential Assessment Tool using an alignment-free logistic regression model. *Nucleic Acids Res* 2013;41:e74.
18 Reich M, Liefeld T, Gould J, *et al*. GenePattern 2.0. *Nat Genet* 2006;38:500–1.
19 Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* 2014;15:550.
20 Liao Y, Smyth GK, Shi W. The Subread aligner: fast, accurate and scalable read mapping by seed-and-vote. *Nucleic Acids Res* 2013;41:e108.
21 Lachmann A, Giorgi FM, Lopez G, *et al*. ARACNe-AP: gene network reverse engineering through adaptive partitioning inference of mutual information. *Bioinformatics* 2016;32:2233–5.
22 Huang J, Zhou N, Watabe K, *et al*. Long non-coding RNA UCA1 promotes breast tumor growth by suppression of p27 (Kip1). *Cell Death Dis* 2014;5:e1008.
23 Jiang W, Liu Y, Liu R, *et al*. The lncRNA DEANR1 facilitates human endoderm differentiation by activating FOXA2 expression. *Cell Rep* 2015;11:137–48.
24 Liu X, Xiao ZD, Han L, *et al*. LncRNA NBR2 engages a metabolic checkpoint by regulating AMPK under energy stress. *Nat Cell Biol* 2016;18:431–42.
25 Ji P, Diederichs S, Wang W, *et al*. MALAT-1, a novel noncoding RNA, and thymosin beta4 predict metastasis and survival in early-stage non-small cell lung cancer. *Oncogene* 2003;22:8031–41.
26 Adriaens C, Standaert L, Barra J, *et al*. p53 induces formation of NEAT1 lncRNA-containing paraspeckles that modulate replication stress response and chemosensitivity. *Nat Med* 2016;22:861–8.
27 Makohon-Moore AP, Zhang M, Reiter JG, *et al*. Limited heterogeneity of known driver gene mutations among the metastases of individual patients with pancreatic cancer. *Nat Genet* 2017;49:358–66.

28 Beroukhim R, Mermel CH, Porter D, *et al*. The landscape of somatic copy-number alteration across human cancers. *Nature* 2010;463:899–905.

29 Zack TI, Schumacher SE, Carter SL, *et al*. Pan-cancer patterns of somatic copy number alteration. *Nat Genet* 2013;45:1134–40.

30 Marín-Béjar O, Marchese FP, Athie A, *et al*. Pint lincRNA connects the p53 pathway with epigenetic silencing by the Polycomb repressive complex 2. *Genome Biol* 2013;14:R104.

31 Amundadottir LT. Pancreatic Cancer Genetics. *Int J Biol Sci* 2016;12:314–25.

32 Lee YC, Zhou Q, Chen J, *et al*. RPA-Binding Protein ETAA1 Is an ATR Activator Involved in DNA Replication stress response. *Curr Biol* 2016;26:3257–68.

33 Huarte M. The emerging role of lncRNAs in cancer. *Nat Med* 2015;21:1253–61.

34 Peng QL, Zhang YM, Yang HB, *et al*. Transcriptomic profiling of long non-coding RNAs in dermatomyositis by microarray analysis. *Sci Rep* 2016;6:32818.

35 Kuga T, Sasaki M, Mikami T, *et al*. FAM83H and casein kinase I regulate the organization of the keratin cytoskeleton and formation of desmosomes. *Sci Rep* 2016;6:26557.

36 Basso K, Margolin AA, Stolovitzky G, *et al*. Reverse engineering of regulatory networks in human B cells. *Nat Genet* 2005;37:382–90.

37 Kopp JL, von Figura G, Mayes E, *et al*. Identification of Sox9-dependent acinar-to-ductal reprogramming as the principal mechanism for initiation of pancreatic ductal adenocarcinoma. *Cancer Cell* 2012;22:737–50.

38 Song Y, Washington MK, Crawford HC. Loss of FOXA1/2 is essential for the epithelial-to-mesenchymal transition in pancreatic cancer. *Cancer Res* 2010;70:2115–25.

39 Zheng J, Huang X, Tan W, *et al*. Pancreatic cancer risk variant in LINC00673 creates a miR-1231 binding site and interferes with PTPN11 degradation. *Nat Genet* 2016;48:747–57.

40 Zhao J, Sun BK, Erwin JA, *et al*. Polycomb proteins targeted by a short repeat RNA to the mouse X chromosome. *Science* 2008;322:750–6.

41 Rinn JL, Kertesz M, Wang JK, *et al*. Functional demarcation of active and silent chromatin domains in human HOX loci by noncoding RNAs. *Cell* 2007;129:1311–23.

42 Hu X, Feng Y, Zhang D, *et al*. A functional genomic approach identifies FAL1 as an oncogenic long noncoding RNA that associates with BMI1 and represses p21 expression in cancer. *Cancer Cell* 2014;26:344–57.

43 Tahira AC, Kubrusly MS, Faria MF, *et al*. Long noncoding intronic RNAs are differentially expressed in primary and metastatic pancreatic cancer. *Mol Cancer* 2011;10:141.

44 Wang Y, Li Z, Zheng S, *et al*. Expression profile of long non-coding RNAs in pancreatic cancer and their clinical significance as biomarkers. *Oncotarget* 2015;6:35684–98.

45 Fu XL, Liu DJ, Yan TT, *et al*. Analysis of long non-coding RNA expression profiles in pancreatic ductal adenocarcinoma. *Sci Rep* 2016;6:33535.

46 Hon CC, Ramilowski JA, Harshbarger J, *et al*. An atlas of human long non-coding RNAs with accurate 5′ ends. *Nature* 2017;543:199–204.

47 Ng SY, Bogu GK, Soh BS, *et al*. The long noncoding RNA RMST interacts with SOX2 to regulate neurogenesis. *Mol Cell* 2013;51:349–59.

48 Martinelli P, Carrillo-de Santa Pau E, Cox T, *et al*. GATA6 regulates EMT and tumour dissemination, and is a marker of response to adjuvant chemotherapy in pancreatic cancer. *Gut* 2017;66.

49 Krah NM, De La O JP, Swift GH, *et al*. The acinar differentiation determinant PTF1A inhibits initiation of pancreatic ductal adenocarcinoma. *Elife* 2015;4.

50 von Figura G, Morris JP, Wright CV, *et al*. Nr5a2 maintains acinar cell differentiation and constrains oncogenic Kras-mediated pancreatic neoplastic initiation. *Gut* 2014;63:656–64.

51 Morris JP, Wang SC, Hebrok M. KRAS, Hedgehog, Wnt and the twisted developmental biology of pancreatic ductal adenocarcinoma. *Nat Rev Cancer* 2010;10:683–95.

52 Roy N, Takeuchi KK, Ruggeri JM, *et al*. PDX1 dynamically regulates pancreatic ductal adenocarcinoma initiation and maintenance. *Genes Dev* 2016;30:2669–83.

53 McDonald OG, Li X, Saunders T, *et al*. Epigenomic reprogramming during pancreatic cancer progression links anabolic glucose metabolism to distant metastasis. *Nat Genet* 2017;49:367–76.

54 Anderson KM, Anderson DM, McAnally JR, *et al*. Transcription of the non-coding RNA upperhand controls Hand2 expression and heart development. *Nature* 2016;539:433–6.

55 Li Q, Kannan A, DeMayo FJ, *et al*. The antiproliferative action of progesterone in uterine epithelium is mediated by Hand2. *Science* 2011;331:912–6.