

ESRP1 controls biogenesis and function of a large abundant multiexon circRNA

Dawei Liu¹, B. Kate Dredge¹, Andrew G. Bert¹, Katherine A. Pillman^{1,2}, John Toubia^{1,3}, Wenting Guo⁴, Boris J.A. Dyakov^{5,6}, Melodie M. Migault¹, Vanessa M. Conn^{1,7}, Simon J. Conn^{1,7}, Philip A. Gregory¹, Anne-Claude Gingras^{5,6}, Dinshaw Patel⁸, Baixing Wu^{1,4,*} and Gregory J. Goodall^{1,2,9,*}

¹Centre for Cancer Biology, SA Pathology and University of South Australia, Adelaide, SA 5000, Australia

²School of Molecular and Biomedical Science, University of Adelaide, Adelaide, SA 5005, Australia

³ACRF Cancer Genomics Facility, Centre for Cancer Biology, SA Pathology and University of South Australia, Frome Road, Adelaide, SA 5000, Australia

⁴Guangdong Provincial Key Laboratory of Malignant Tumor Epigenetics and Gene Regulation, Guangdong-Hong Kong Joint Laboratory for RNA Medicine, RNA Biomedical Institute, Medical Research Center, Sun Yat-Sen Memorial Hospital, Sun Yat-Sen University, Guangzhou, 510120, China

⁵Lunenfeld-Tanenbaum Research Institute, Mount Sinai Hospital, Sinai Health, 600 University Ave, Toronto, ON M5G 1X5, Canada

⁶Department of Molecular Genetics, University of Toronto, Toronto, ON M5S 1A8, Canada

⁷Flinders Health and Medical Research Institute, College of Medicine & Public Health, Flinders University, Bedford Park, SA, 5042, Australia

⁸Structural Biology Program, Memorial Sloan Kettering Cancer Center, New York, NY, USA

⁹Adelaide Medical School, Faculty of Health and Medical Sciences, University of Adelaide, Adelaide, SA, Australia

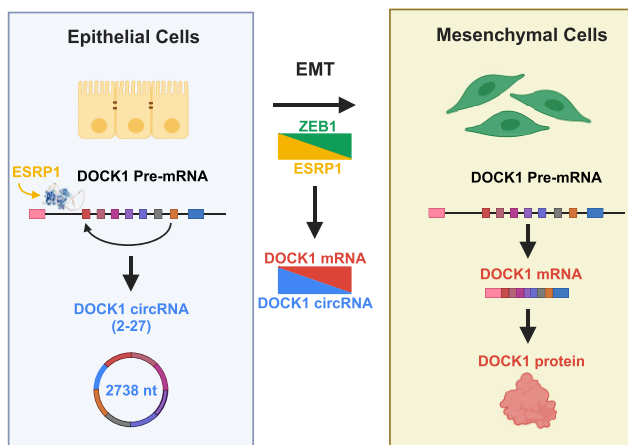
*To whom correspondence should be addressed. Tel: +61 8 83027751; Email: greg.goodall@unisa.edu.au

Correspondence may also be addressed to Baixing Wu. Tel: +86 18818264561; Email: wubx28@mail.sysu.edu.cn

Abstract

While the majority of circRNAs are formed from infrequent back-splicing of exons from protein coding genes, some can be produced at quite high level and in a regulated manner. We describe the regulation, biogenesis and function of circDOCK1(2–27), a large, abundant circular RNA that is highly regulated during epithelial-mesenchymal transition (EMT) and whose formation depends on the epithelial splicing regulator ESRP1. CircDOCK1(2–27) synthesis in epithelial cells represses cell motility both by diverting transcripts from DOCK1 mRNA production to circRNA formation and by direct inhibition of migration by the circRNA. HITS-CLIP analysis and CRISPR-mediated deletions indicate ESRP1 controls circDOCK1(2–27) biosynthesis by binding a GGU-containing repeat region in intron 1 and detaining its splicing until Pol II completes its 157 kb journey to exon 27. Proximity-dependent biotinylation (BioID) assay suggests ESRP1 may modify the RNP landscape of intron 1 in a way that disfavors communication of exon 1 with exon 2, rather than physically bridging exon 2 to exon 27. The X-ray crystal structure of RNA-bound ESRP1 qRRM2 domain reveals it binds to GGU motifs, with the guanines embedded in clamp-like aromatic pockets in the protein.

Graphical abstract



Received: April 5, 2023. Revised: October 24, 2023. Editorial Decision: November 10, 2023. Accepted: November 13, 2023

© The Author(s) 2023. Published by Oxford University Press on behalf of Nucleic Acids Research.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

Introduction

It is now well established that eukaryotic cells express many circular RNAs (circRNAs). Some circRNAs are produced from intron lariats, but the majority are entirely comprised of exons and arise from back-splicing, which links an exon to a preceding exon, or to itself in the case of single exon circRNAs (1). Two principal mechanisms cause the formation of such back-spliced circRNAs. The majority are believed to result from inverted repeats within introns, (typically Alu repeats), which base pair with each other to form a large hairpin that brings the back-spliced splice sites into proximity, with the spliceosome then catalysing the formation of the back-splice. These circRNAs are typically present at very low levels (2). Some circRNAs could be formed due to the binding of a dimeric RNA binding protein to introns flanking the back-splice junction, as observed with the alternative splicing regulators MBNL1 (3) and QKI (4). Since QKI and MBNL1 form homodimers (5,6), they can presumably bind to two distinct introns and thereby bring the flanking splice sites into proximity for backsplicing to occur.

While the extreme sensitivity of detection of circRNAs by short read sequencing has led to the identification of hundreds of thousands of different circRNAs, most of these are almost certainly accidents of mis-splicing, present at such low level that they are unlikely to be of functional consequence (2). However, a small proportion of the known circRNAs are present at much higher levels that are likely to be of functional significance. A few such circRNAs have indeed been shown to affect cell proliferation or phenotype (7), while some that do not yet have functions identified are directly regulated during cellular differentiation, consistent with a role in contributing to the cell phenotype (4).

We have chosen to investigate the mechanism of formation and the function of the DOCK1(2–27) circRNA because it is highly regulated during EMT. This unusually large circRNA of 2738 nt comprising exons 2–27 from the DOCK1 gene (with circBase designation hsa_circ_0020397) is expressed in epithelial cells but is strongly downregulated during epithelial to mesenchymal transition (EMT), while the DOCK1 mRNA is increased in level during EMT. We found that the formation of this circRNA is dependent on the splicing regulator ESRP1, which promotes circDOCK1(2–27) formation by inhibiting splicing of exon 1 to exon 2, thereby holding the intron 1 acceptor site unspliced and available while Pol II completes its 157 kb journey from exon 2 to exon 27. We found by HITS-CLIP analysis that ESRP1 binds to a GU-rich motif in a tandem repeat region in intron 1. We solved the crystal structure of ESRP1 qRRM2 bound to the repeat motif, revealing that the DOCK1 qRRM2 domain binds a GGU motif in a double clamp arrangement, with aromatic residues on either side of each G forming an aromatic pocket, while bound water molecules form hydrogen bonds to the uracil sidechain. The binding of ESRP1 to the intron 1 region is necessary and sufficient for the high-efficiency formation of the circDOCK1(2–27), so that splicing of the DOCK1 pre-mRNA is diverted to circRNA formation, limiting the production of DOCK1 mRNA and protein, while the circDOCK1(2–27) itself also directly reduces the migratory capacity of cells.

Materials and methods

Cell culture and transduction

Human cancer cell line culture conditions were described previously (8). Briefly, cell lines were cultured with 5% CO₂ at

37°C. HEK293T, MCF-7, T47D, MDA-MB-361, ZR-75-1, MDA-MB-415, MDA-MB-134-VI, Hs578T and MDA-MB-231 were cultured in DMEM (Thermo Fisher) + 10% FCS. MDA-MB-436, MDA-MB-157, CAL51 and CAL120 cells were cultured in RPMI media with 10% FCS, 20 mM HEPES and 288 µl Insulin/100 ml. SUM159PT cells were cultured in Ham's F12 media with 5% FCS, 5 µg/ml insulin and 1 µg/ml hydrocortisone, HMLE cells in HuMEC basal serum free media (Thermo Fisher) and mesHMLE cells in Weinberg media (DMEM + F12 media with 5% FCS, 4 mg/ml insulin, 20 µg/ml EGF and 1 mg/ml hydrocortisone).

siRNAs were transfected with Lipofectamine RNAiMAX (Life Technologies) at 10 nM concentration following the manufacturer's protocol. DNA plasmid transfections were performed with Lipofectamine 2000 (Life Technologies) following the manufacturer's protocol.

CRISPR knockout generation

A pair of lentiCRISPR v2 plasmids (a gift from Feng Zhang (Addgene plasmid # 52961) with designed sgRNAs flanking the target region were co-transfected with pcDNA3-GFP simultaneously into T47D and 293T cells. Three days after transfection, individual GFP positive cells were sorted into 96-well plates. PCR primers flanking the target region were designed and used to screen for individual clones with successful genomic deletion. To validate the genomic deletions by Sanger sequencing, 10 ng of purified PCR product and 10 pmol of sequencing primer were mixed and Sanger sequencing was performed by the Australian Genome Research Facility (AGRF).

Recombinant DNA constructs

For CRISPR experiments, gRNAs were cloned into lentiCRISPR v2 (Plasmid #52961, Addgene) using the BsmBI site. For pINDUCER-20 (Plasmid #44012) doxycycline inducible gene expression, cDNA was firstly cloned into pENTR2B gateway entry vector, the pENTR2B derivative was then recombined with pINDUCER-20 using LR Clonase to produce doxycycline inducible expression. Viral production and transduction were performed as previously described (8). Generally, Lentivirus was produced in a T25 flask of HEK 293T cells beginning at 50–60% confluency. Plasmids encoding the gag-pol genes, the rev gene and the VSV-g envelope gene were co-transfected with pINDUCER-20 for viral production. After viral transduction, G418 was used for cell selection.

Cell sorting

The optimal number of cells were resuspended into cell sorting buffer (5 mM EDTA, 25 mM HEPES, 2% fetal calf serum in 1× PBS). After filtering the cells through a 30 µm filter into a FACS tube, the cells were then sorted and collected by MoFlo Astrios Cell Sorter according to the manufacturer's protocol.

Western Blotting

Cells were rinsed with cold 1× PBS then lysed on ice in 1× RIPA buffer (Abcam) supplemented with protease inhibitor cocktail (Roche) and phosphatase inhibitor (PhosSTOP, Roche). After centrifugation at 13 000 rpm for 20 min at 4°C, protein concentration was quantified by Bicinchoninic Acid assay (Pierce). Samples were diluted in 1× Bolt LDS sample buffer (ThermoFisher) supplemented with 2% beta-mercaptoethanol (Merck) final concentration. After denaturation at 70°C for 10 min, 10–20 µg of proteins were separated on Bolt bis-tris gel in 1× MOPS SDS running buffer

(ThermoFisher) and transferred onto 0.45 μm nitrocellulose membrane (Advantec). Membranes were blocked in 5% skim milk in TBST buffer (20 mM Tris-HCl pH7.6, 150 mM NaCl, 0.2% Tween-20) for 1 h at room temperature then incubated with primary antibodies diluted in 3% BSA in TBST buffer. After 5 min wash with TBST buffer thrice, membranes were incubated with either HRP-conjugated (Pierce) or IRdye secondary antibodies (Licor) diluted in TBST buffer at 1:10 000. After 5 min wash with TBST buffer thrice, membranes were imaged using Chemidoc MP (Bio-Rad). Primary antibodies used in this study were: DOCK1 antibody (1:1000) from Bethyl Lab (A301-288), ESRP1 antibody (1:1000) from ThermoFisher (PA5-25833) and alpha-Tubulin (1:10000) from Abcam (ab7291). Secondary antibodies used in this study were: Goat anti-mouse IR680RD (LCR-926-68070), Goat anti-rabbit IR680RD (LCR-926-68071), Goat anti-rabbit IR800CW (LCR-926-32211), Goat anti-mouse-HRP (31430) and Goat anti-rabbit-HRP (31460).

Migration assay

Transwell migration assays were performed as previously described (8). Generally, 2×10^5 MDA-MB-231 or mesHMLE cells were plated into Transwells (Corning, 6.5 mm, 8.0 μm pore size) in serum-free medium, then 10% FCS was added into the lower chamber to induce chemotactic migration for 4 hours.

RNA Isolation and PCR

RNA extraction, RT-PCR and qPCR were performed as previously described (8). TRIzol (ThermoFisher) was used for RNA extraction following the manufacturer's instructions. QuantiTect RT kit (Qiagen) was used for mRNA and circular RNA reverse transcription. The synthesised cDNA was then diluted 1:20 for quantitative PCR (qPCR). qPCR was performed in triplicate on a Rotor-Gene-Q series PCR machine (Qiagen) using the QuantiTect SYBR Green PCR kit (Qiagen). The sequences of primers are shown in [Supplementary Table S4](#). Rotor-Gene software was used for data analysis and GAPDH expression was used for gene expression normalization.

Standard Taq polymerase (NEB) or Phusion DNA Polymerase (ThermoFisher) were used for standard PCR for gene clones.

HITS-CLIP assay

HITS-CLIP was performed on endogenous ESRP1 from 3 biological replicates of wt HMLE cells using an ESRP1 specific antibody. We also performed HITS-CLIP using anti-FLAG antibody on 2 biological replicates each of N- and C-terminally 3XFLAG-tagged ESRP1 expressed in HMLE-i-ESRP1_FLAG cells after induction with 1 $\mu\text{g}/\text{ml}$ doxycycline for 72 h. Cells grown to 80% confluency in 150mm dishes were rinsed with ice-cold PBS and UV irradiated twice with 300 mJ/cm^2 , 254 nm, in ice-cold PBS using a Spectrolinker XL1500 (Spectro-UV). Cells were lysed in the dish with 750 μl 1 \times PXL [1 \times PBS, 0.1% SDS, 0.5% deoxycholate, 0.5% Igepal] + EDTA-free Complete protease inhibitor cocktail (PIC; Roche), collected by scraping and stored at -80°C until use. Thawed lysates were triturated using a 21G needle and DNA digested with 40 μl Turbo DNase (Ambion AM2238) at 37°C , 350 rpm for 10 min. RNA was partially digested with RNase 1 (Ambion AM2295) by adding 10 μl of 1:25 diluted RNase 1 per 1ml of

lysate at 37°C for 5 min. Lysates were centrifuged at 21 000 g for 30 min at 4°C and supernatant transferred to a fresh tube.

ESRP1-RNA complexes were immunoprecipitated using rabbit polyclonal anti-ESRP1 antibody (Proteintech Cat# 21045-1-AP); 8 μg antibody pre-bound to to 75 μl protein A Dynabeads (ThermoFisher, 10002D), or mouse monoclonal anti-FLAG M2 antibody (Sigma Cat# F1804); 18 μg antibody pre-bound to to 75 μl protein G Dynabeads (ThermoFisher, 10004D). Negative controls were performed using wt HMLE lysates immunoprecipitated with rabbit IgG (ThermoFisher Cat# 02-6102), or anti-FLAG antibodies coupled to protein A or G beads respectively. Washed beads were resuspended with 1.0 ml of prepared lysate at $\sim 1.4\text{mg}/\text{ml}$ and rotated 75 min at 4°C . Bound ESRP1-RNA complexes were washed twice each consecutively with ice cold 1 \times PXL, 5 \times PXL [5 \times PBS, 0.1% SDS, 0.5% sodium deoxycholate, 0.5% Igepal], and 1 \times PNK [50 mM Tris-Cl pH 7.5, 10 mM MgCl_2 and 0.5% Igepal]. Beads were first treated with T4 PNK (NEB, M0201L; 10 U in 80 μl reaction volume) in the absence of ATP (37°C , 850 rpm for 20 min in a thermomixer) to dephosphorylate 3' RNA ends followed by washes with 1 \times PNK, 1 \times PNK + EGTA [50 mM Tris-Cl pH 7.5, 20 mM EGTA, and 0.5% Igepal], and two washes with 1 \times PNK at 4°C .

The 3' preadenylated linker (NEBNext 3'SR adaptor for Illumina; /5rApp/AGA TCG GAA GAG CAC ACG TCT /3AmMO/) was ligated to the RNA fragments on bead using T4 RNA ligase 2 truncated KQ (NEB M0373) at 16°C , overnight with shaking. Beads were washed consecutively with ice cold 1 \times PNK, 5 \times PXL and twice with 1 \times PNK. Bound RNAs were then labelled with P32 γ -ATP using T4 PNK, 45 min at 37°C , followed by addition of 2.5 μM ATP, 5 mins at 37°C . Beads were washed twice each with ice-cold 1 \times PNK + EGTA and 1 \times PNK. The 5' RNA linker (5'-blocked and containing a 10 nt UMI (/5AmMC6/GUUCAGAGUU CUACAGUCCGACGAUCNNNNNNNNNN3')) was ligated to the RNA fragments on bead using T4 RNA ligase (NEB M0437) for 1 hr at 25°C , with intermittent shaking. Beads were washed 3 times with ice-cold 1 \times PNK + EGTA.

Protein A-anti-ESRP1 and Protein A-IgG beads were eluted with 40 μl 1 \times Bolt LDS sample buffer (ThermoFisher) + 1% β -mercaptoethanol at 70°C for 10 min. Protein G-anti-FLAG beads were eluted with 30 μl 3 \times FLAG peptide (Sigma F4799; 250 $\mu\text{g}/\text{ml}$ in PBS, 0.02% Tween-20), 20°C for 30 min with shaking. 15 μl 4 \times Bolt LDS sample buffer + 4% β -mercaptoethanol was then added to the FLAG eluates and the samples heated at 70°C for 10 min. All samples were then separated through Bolt 8% Bis-tris Plus gels (ThermoFisher) using BOLT MOPS SDS running buffer at 175 V for 1 h. Protein-RNA complexes were then transferred to nitrocellulose (Schleicher&Schuell, BA-85) by wet transfer using 1 \times Bolt transfer buffer with 10% methanol. Filters were placed on a phosphor screen and exposed using a Typhoon imager (GE).

The 100–130 kDa region of each lane was excised, corresponding to ESRP1:RNA complexes with RNAs ~ 20 –100nt + linkers, and the RNA liberated by proteinase K digestion (2 mg/mL proteinase K, 100 mM Tris-HCl pH 7.5, 50 mM NaCl, 10 mM EDTA, 0.2% SDS) at 50°C for 60 min, 1200 rpm, followed by extraction with acid phenol (ThermoFisher, AM9712) and precipitation with 1:1 isopropanol:ethanol. RNA was pelleted by centrifugation then separated on an 8% denaturing polyacrylamide gel (1:19 acrylamide, 1–TBE, 7 M urea). The wet gel was wrapped in plastic wrap and exposed to a phosphor screen and imaged using a

Typhoon. Gel slices were cut (size 75–150nt) and eluted by the ‘crush and soak’ method, followed by ethanol precipitation.

Reverse transcription was performed using a nested RT primer (IDT, AGACGTGTGCTCTTCCGA) with SuperScript IV and MnCl₂ buffer [50 mM Tris pH 8.0, 75 mM KCl, 3 mM MnCl₂] to enhance read-through at crosslink sites (9). Products were amplified for 12–16 cycles using NEBNext Ultra II Q5 mastermix (NEB cat#M0544) with a common forward primer (NEBNext SR primer for Illumina) and barcoded reverse primers for each sample (NEBNext Index primers for Illumina). PCR products were purified using 1.8 volumes of Axygen AxyPrep magnetic beads (MAG-PCR-CL), separated on an 8% acrylamide (19:1), 7 M urea TBE semi-denaturing gel, stained with SYBR Gold nucleic acid gel stain (ThermoFisher) and imaged on a ChemiDoc (BioRad). Products corresponding to an insert size of > 20 nt were excised from the gel and extracted by the ‘crush and soak’ method. Library quantity was determined by qPCR using NEBNext Library Quant kit for Illumina, pooled and sequenced on an Illumina NextSeq 500 (1 × 75 bp).

HITS-CLIP bioinformatic analyses

The eleven HITS-CLIP libraries average raw sequencing depths of 79 million, 34 million, 36 million, 57 million and 4 million reads for the endogenous ESRP1, N-terminal FLAG-tagged inducible ESRP1, C-terminal FLAG-tagged inducible ESRP1, control IgG and control FLAG replicates, respectively. FASTQ files were analysed at various stages for quality and content using FastQC v0.11.9 (<http://www.bioinformatics.babraham.ac.uk/projects/fastqc/>) and raw reads were adapter trimmed and filtered using cutadapt v2.8 (10) using an adapter sequence of AGATCGGAAGAGCACACGTCT-GAACTCCAGTCA, error rate of 0.2, and overlap of 5 and minimum length of 28. Reads derived from PCR duplication were collapsed using Unique Molecular Identifiers (UMIs) using UMI-tools (v0.5.3) (11) by first using the ‘extract’ method with default parameters to cut the 10 bp UMIs from the 3’ end of the reads. Reads were then mapped against the human reference genome (hg19) using the STAR (v2.7.2c) spliced alignment algorithm (12) with parameters `-twopassMode basic` and `-quantMode GeneCounts` and otherwise default parameters, at an alignment rate of ~82–97%. Subsequently, unique molecular identifiers (UMIs) were used to collapse PCR duplicate reads using the UMI-tools ‘dedup’ method with default parameters. To identify enriched regions of the genome, replicate samples were pooled using the Picard Tools function MergeSamFiles (<http://broadinstitute.github.io/picard/>) and quality filtered using samtools (-q 10) (13).

For the endogenous ESRP1, N-terminal FLAG-tagged inducible ESRP1 and C-terminal FLAG-tagged inducible ESRP1, peak calling was then performed separately for each strand using MACS2 peak caller (version 2.1.1) (14) using the combination of the IgG and FLAG control samples as the control. The following settings were used (`-f BAM -g hs -keep-dup all -nomodel -extsize 50 -B -call-summits -slocal 0 -llocal 0 -fe-cutoff 3 -q 0.05`) and the resulting peak files from each strand were merged. HITS-CLIP peaks and alignments were visualized and interrogated using the Integrative Genomics Viewer v2.8.0 (15). Homer (16) was used to perform de novo motif enrichment analysis (findMotifsGenome.pl parameters: `-size given -norevpp -len 5, 6, 7, 8, 9, 10`). This identified several motifs similar to the published ESRP1 motif which were

highly enriched; shown are motifs which were significantly enriched (*P*-value: <<1e-100) in the C-terminal FLAG samples.

Relative intron abundance measurement

Relative RNA sequence read numbers in each DOCK1 intron were calculated using data from (8) and are deposited to the European Nucleotide Archive database (<http://www.ebi.ac.uk/ena/data/view/PRJEB25042>) with the study accession number PRJEB25042. We first determined the number of reads mapping to each intron in a strand-specific manner, counting only reads with the read start inside the intron. The two smallest introns, intron 34 and intron 41, both less than 200 bp, were discarded as they contained too few reads to be accurately quantified. The counts for each intron were then normalised by the length of that intron, adding a pseudo-count of 1 and transforming the data by log₂. Note that for DOCK1 intron 1, the iGenomes hg19 genome assembly version used contains a large region of low complexity which is masked (represented as ‘N’s). This region was excluded from the analysis. To remove the influence of differences in DOCK1 gene expression, we subtracted the expression of the gene in that sample (defined as the median intron coverage per kb) from the values calculated above. To determine whether reads were statistically significantly overrepresented in the HMLE intron 1 relative to (a) the MesHMLE intron 1, or (b) other introns in the same cell line, we used a student’s *t*-test with multiple-testing correction using the Bonferroni method. We used two biological replicates for HMLE samples and three for MesHMLE samples.

Protein purification

The cDNA for human ESRP1 (amino acids 1–681, UniProt ID: Q6NXG1) was purchased from Shanghai Generay Biotech Co., Ltd, China. The target fragment of qRRM2 (312–430) was amplified by PCR reaction and cloned into a modified pET28a-SUMO vector, the recombinant vector was then transferred into *Escherichia coli* BL21 (DE3) competent cell for protein expression. The plasmids of ESRP1 mutants were obtained by overlap PCR using the wild-type ESRP1 plasmid as the template. Sequences of wildtype and all mutant plasmids were confirmed by DNA sequencing. The frozen recombinant strains were cultivated at 37°C in LB medium supplemented with 50 µg/mL kanamycin. The protein expressions were induced at OD₆₀₀ of 0.6–0.8 by adding of isopropyl β-D-1-thiogalactopyranoside (IPTG) at a final concentration of 0.2 mM. The cultures were incubated at 18°C for an additional 16 h to allow the accumulation of expressed proteins. Cells were harvested by centrifugation, resuspended in buffer 1 (20 mM Tris-HCl pH 8.0, 500 mM NaCl, 25 mM imidazole pH 8.0), and lysed under high pressure. Cell extracts were centrifuged at 18000 rpm for 1h at 4°C. Supernatants were applied to a HisTrap™ HP column pre-equilibrated with buffer 1, and the target proteins were eluted from the column using buffer 2 (20 mM Tris pH 8.0, 500 mM NaCl, 500 mM imidazole pH 8.0) with a gradient. The recombinant protein was dialyzed against buffer S (20 mM Tris pH 8.0, 500 mM NaCl) for 3 hours, Ulp1 protease was added to remove the His-Sumo tag. The mixture was applied to HisTrap™ HP column again and the fractions containing the target protein were pooled, concentrated and loaded onto a HiLoad™ 16/600 Superdex 75 column pre-equilibrated using gel filtration buffer (10 mM Tris pH 8.0, 100 mM NaCl). Purities of the proteins were

analyzed using SDS-PAGE gel and the samples were stored at -80°C until use. Protein concentration was determined using UV observation at A280.

Crystallization and data collection

Crystals were grown using the sitting drop vapor diffusion method at 20°C with the drop composed of $0.5\ \mu\text{l}$ of protein-RNA sample and $0.5\ \mu\text{l}$ of crystallization solution. The qRRM2-RNA complex crystals suitable for X-ray diffraction were grown in reservoir solution consisting of $0.2\ \text{M}$ Sodium malonate pH 7.0, 20% w/v Polyethylene glycol 3350. Crystals were cryoprotected using their mother liquor supplemented with 25% glycerol and snap-frozen in liquid nitrogen. X-ray diffraction data were collected on beamline BL19U1, BL17U1 and BL18U1 at the Shanghai Synchrotron Radiation Facility (SSRF). Data processing was carried out using the HKL3000 programs or XDS (17,18). The data collection and processing statistics are summarized in [Supplementary Table S3](#).

Structure determination and refinement

The qRRM2-RNA complex structure was determined by molecular replacement (MR) method using the Phaser program embedded in the CCP4i suite (19), the structure of the protein (PDB code: 2RVJ) was used as the search model (20). The resulting model was refined against the diffraction data using the REFMAC5 program (21) in CCP4i and the model building was performed using the program COOT (22). The 2Fo-Fc and Fo-Fc electron density maps were regularly calculated and used as guides for model building. The final refinement of structure was done using the phenix.refine program (23) of PHENIX. Ramachandran analysis showed that all the protein residues are located in the most favored or allowed region. The detailed structural refinement statistics are summarized in [Supplementary Table S3](#). All structure figures were prepared with PyMOL (DeLano Scientific).

Isothermal titration calorimetry

ITC titrations were performed at 25°C using a MicroCal PEAQ instrument with titration buffer composed of $100\ \text{mM}$ NaCl and $10\ \text{mM}$ Tris-HCl pH 8.0. The concentrations of proteins were determined spectrophotometrically. The RNA substrates were diluted in the reaction buffer. The ITC experiments involved 20 or 25 injections of protein into RNA. The sample cell was loaded with $300\ \mu\text{L}$ of RNA at $20\ \mu\text{M}$ and the syringe with $80\ \text{ml}$ of protein at $300\ \mu\text{M}$. Curve fitting to a single binding site model was performed using the ITC data analysis module of Origin 7.0 (MicroCal) provided by the manufacturer. ΔG° of protein-RNA binding was computed as $R\ln(1/K_d)$, where R , T and K_d are the gas constant, temperature and dissociation constant, respectively.

BioID samples

The ESPR1 Open Reading Frame (corresponding to NP_060167.2, from the ORFeome 8.1 collection in pDONR223) was cloned using GatewayTM enzymes in the pDEST-pcDNA5-C-term-FLAG-BirA* vector, adding a FLAG and BirA* fusion to the C-terminus of ESPR1. This construct was transfected (along negative controls consisting of the parental BirA*-FLAG tag) into Flp-In T-REx 293 cells (Invitrogen), and pools of stable transfectants

were selected (internal reference C2982) and processed for BioID (alongside an additional negative control consisting of non-transfected cells), essentially as in (24) with minor modifications. Harvested cells were lysed in modified RIPA buffer [$50\ \text{mM}$ Tris (pH 7.5), $150\ \text{mM}$ NaCl, $1.5\ \text{mM}$ MgCl_2 , $1\ \text{mM}$ EGTA, 0.1% SDS, 1% IGEPAL CA-630] with freshly added sodium deoxycholate (0.4%) and protease inhibitors (Sigma-Aldrich P8340) at $400\ \mu\text{l}/0.1\ \text{g}$ cells, sonicated, and treated benzonase ($375\ \text{U}$ per sample) for $15\ \text{min}$ at 4°C . SDS was adjusted to a final concentration of 0.4% (incubation $15\ \text{min}$, 4°C). Samples were centrifuged ($16\ 000 \times g$ for $20\ \text{min}$) and the cleared lysates were incubated with pre-washed streptavidin Sepharose beads (GE 17-5113-01; $30\ \mu\text{l}$ bed volume) for $3\ \text{h}$, 4°C . Beads were washed once with wash buffer [2% SDS, $50\ \text{mM}$ Tris (pH 7.5)], $2\times$ with modified RIPA buffer with 0.4% SDS, and $3\times$ with ABC buffer [$50\ \text{mM}$ ammonium bicarbonate (pH 8.5)]. Beads were pelleted, supernatant removed, and on-bead trypsin digest of peptides was performed in two-steps ($1\ \mu\text{g}$ trypsin overnight; $0.5\ \mu\text{g}$ trypsin for $2\ \text{h}$; both at 37°C). Supernatants were transferred to new tubes (beads were rinsed twice in water and supernatants combined). Freshly made 50% formic acid was added to samples to a final concentration of 2% prior to drying by vacuum centrifugation and storage at -80°C .

Mass spectrometry acquisition and analysis

Each sample ($6\ \mu\text{L}$ in 2% formic acid; corresponding to $1/6\text{th}$ of a $15\ \text{cm}$ tissue culture dish) was directly loaded at $800\ \text{nL}/\text{min}$ onto an equilibrated HPLC column (pulled and packed in-house). The peptides were eluted from the column over a $90\ \text{min}$ gradient generated by a Eksigent ekspertTM nanoLC 425 (Eksigent, Dublin CA) nano-pump and analysed on a TripleTOF 6600 instrument (AB SCIEX, Concord, Ontario, Canada). The Data Dependent Acquisition method consisted of one $250\ \text{ms}$ MS1 TOF survey scan from 400 to $1800\ \text{Da}$ followed by ten $100\ \text{ms}$ MS2 candidate ion scans from 100 to $1800\ \text{Da}$ in high sensitivity mode. Only ions with a charge of $2+$ to $5+$ that exceeded a threshold of $300\ \text{cps}$ were selected for MS2, and former precursors were excluded for $7\ \text{s}$ after one occurrence. Data were stored, searched and analyzed using ProHits laboratory information management system (25). Within ProHits, WIFF files were converted to an MGF format using the WIFF2MGF converter and to an mzML format using ProteoWizard (V3.0.10702) and the AB SCIEX MS Data Converter (V1.3 beta). The data was then searched using Mascot (V2.3.02) (26) and Comet (V2016.01 rev.2) (27). The spectra were searched with the human and adenovirus sequences in the RefSeq database (version 57, 30 January 2013) acquired from NCBI, supplemented with 'common contaminants' from the Max Planck Institute (<http://maxquant.org/contaminants.zip>) and the Global Proteome Machine (GPM; <ftp://ftp.thegpm.org/fasta/cRAP/crap.fasta>), forward and reverse sequences (labeled 'gil9999' or 'DECOY'), sequence tags (BirA, GST, mCherry and GFP) and streptavidin, for a total of 72481 entries. Database parameters were set to search for tryptic cleavages, allowing up to 2 missed cleavages sites per peptide with a mass tolerance of $35\ \text{ppm}$ for precursors with charges of $2+$ to $4+$ and a tolerance of $0.15\ \text{amu}$ for fragment ions. Variable modifications were selected for deamidated asparagine and glutamine and oxidized methionine. Results from each search engine were analyzed through TPP (the Trans-Proteomic Pipeline,

v.4.7 POLAR VORTEX rev 1) via the iProphet pipeline (28), and only proteins with a 95% FDR iProphet filter were considered further. SAINTExpress version 3.6.1 was used as a statistical tool to calculate the probability of potential protein-protein associations compared to background contaminants using default parameters, and control compression set to 2 (effectively compressing the four negative controls to two virtual controls) (29). SAINT scores with a Bayesian false discovery rate (BFDR) $\leq 1\%$ were considered high-confidence protein interactions. All non-human protein interactors (did not start with 'NP' in Prey column) were removed from the SAINT analysis, except for BirA_R118G_H0QFJ5. Visualization and additional analysis of the data was through ProHits-viz.org (30) and the humancellmap.org (31) resources, using default options. Enrichment was performed using g:profiler (32) with default options.

Statistical analyses

Unless otherwise stated, data are shown as the mean of three replicate experiments \pm SEM, with statistically significant *t*-tests shown as * $P < 0.05$, ** $P < 0.01$ or *** $P < 0.001$.

Results

Formation of the DOCK1(2–27) circRNA is regulated by TGF- β and downregulated in mesenchymal tumours

Motivated by the observation in our previous study of regulated changes in abundance of circRNAs during EMT that a circRNA arising from the DOCK1 gene appears to be regulated in the opposite direction from QKI-regulated circRNAs (4), we performed qPCR to measure circDOCK1(2–27) and DOCK1 mRNA in RNA extracted from human breast epithelial cells (HMLE cells) before and after prolonged treatment with TGF- β (producing mesenchymal 'mesHMLE' cells). This confirmed that the large circRNA from the DOCK1 gene, comprised of exons 2–27, which we call circDOCK1(2–27), was strongly downregulated following the TGF- β treatment (Figures 1A and S1A, B). Interestingly, the DOCK1 mRNA did not decrease, but instead was increased, suggesting that the TGF- β was repressing biogenesis of the circRNA rather than repressing transcription of the DOCK1 gene. To confirm this regulation of the circDOCK1(2–27) by TGF- β , we monitored its expression in HMLE cells during a time course of treatment with TGF- β and found the circRNA level progressively reduced over the 21 day period, while the DOCK1 mRNA level progressively increased (Figure 1B), similar to the mesenchymal gene ZEB1 (Supplementary Figure S1D). To check whether the epithelial-specific expression of the circDOCK1(2–27) circRNA is a general feature of breast cancers, we measured its expression in a panel of breast cancer cell lines, and also calculated its expression in epithelial versus mesenchymal tumours of diverse lineage, which confirmed its highly epithelial-specific expression (Figure 1C–E and Supplementary Figure S1C, E).

Formation of the DOCK1(2–27) circRNA is dependent on ESRP1

Given its epithelial-specific expression, we hypothesised that the formation of circDOCK1(2–27) may be regulated by an epithelial-specific splicing factor and so we asked whether its level is affected by depletion of either ESRP1 or ESRP2,

which are well-known epithelial-specific splicing regulators (33). Depletion of ESRP1 by either of two independent siRNAs (Supplementary Figure S2A) drastically reduced the level of circDOCK1(2–27) in HMLE cells (Figure 2A), whereas depletion of ESRP2, which is present at 40% the level of ESRP1 (Supplementary Figure S2), had no effect on circDOCK1(2–27) (Figure 2A). To assess whether introducing ESRP1 into mesenchymal cells was sufficient to drive the formation of the circRNA, we constructed dox-inducible ESRP1 lentivirus and expressed ESRP1 in mesenchymal MDA-MB-231 and mesHMLE cells. ESRP1 induction caused a large increase in the level of circDOCK1(2–27) in each of these cell lines (Figure 2B). Consistent with these observations, the decline of ESRP1 in TGF- β -treated cells is matched by a decline in circDOCK1(2–27) level (Supplementary Figure S2C). Moreover, circDOCK1(2–27) levels are strongly correlated with the level of ESRP1 in breast cancer cell lines and in cancers in general (Figure 2C and D), and circDOCK1(2–27) is more strongly correlated with ESRP1 than with any other transcript in cancers in general (Figure 2E and Supplementary Table S1). Together these data indicate that ESRP1 potentially regulates the formation of circDOCK1(2–27).

DOCK1 circRNA formation competes with linear splicing, reducing DOCK1 mRNA and protein and affecting cell migration

The apparently reciprocal relationship between DOCK1 mRNA and circRNA in epithelial versus mesenchymal cells (Figure 1) raises the possibility that channelling of pre-mRNA into the circRNA form in epithelial cells contributes to reducing DOCK1 mRNA and protein levels. To assess this, we first examined whether depletion or overexpression of ESRP1 affects DOCK1 levels. Indeed, knockdown of ESRP1 increased the level of DOCK1 mRNA and protein in epithelial cells (Figure 3A), while enforced expression of ESRP1 in mesenchymal cells decreased the level of DOCK1 mRNA and protein (Figure 3B), indicating that the formation of circDOCK1 competes substantially with the production of DOCK1 mRNA. Consequently, the DOCK1 protein is more abundant in mesenchymal cells than in epithelial cells (Supplementary Figure S3A).

Since DOCK1 is well known to be a promoter of cell migration (34), a reduction in DOCK1 protein level is likely to reduce the migratory capacity of cells. With this in mind, we were also interested in determining whether the DOCK1 circRNA itself can have a direct effect on cell migration, in addition to its indirect effect via modulation of DOCK1 protein, so we assessed the effect of artificial overexpression of the circDOCK1(2–27) on cell migration. To do this, we constructed a circDOCK1(2–27) expression vector by incorporating the cDNA for the circRNA flanked by splice sites and inverted repeats in the intron regions to promote circularisation (Figure 3C). We confirmed that the expression vector gives rise to predominantly circRNA by performing qPCR assessment of linear and circRNA forms of the transcript (Supplementary Figure S3B, C). Enforced expression of the circDOCK1(2–27) in mesenchymal cells at levels comparable to those in epithelial cells reduced the cell migration rate (Figure 3D and Supplementary Figure S3D). Thus the formation of circDOCK1(2–27) has a two-fold effect on the migratory capacity of cells, acting both directly to limit migration, and indirectly by reducing the level of DOCK1 protein production.

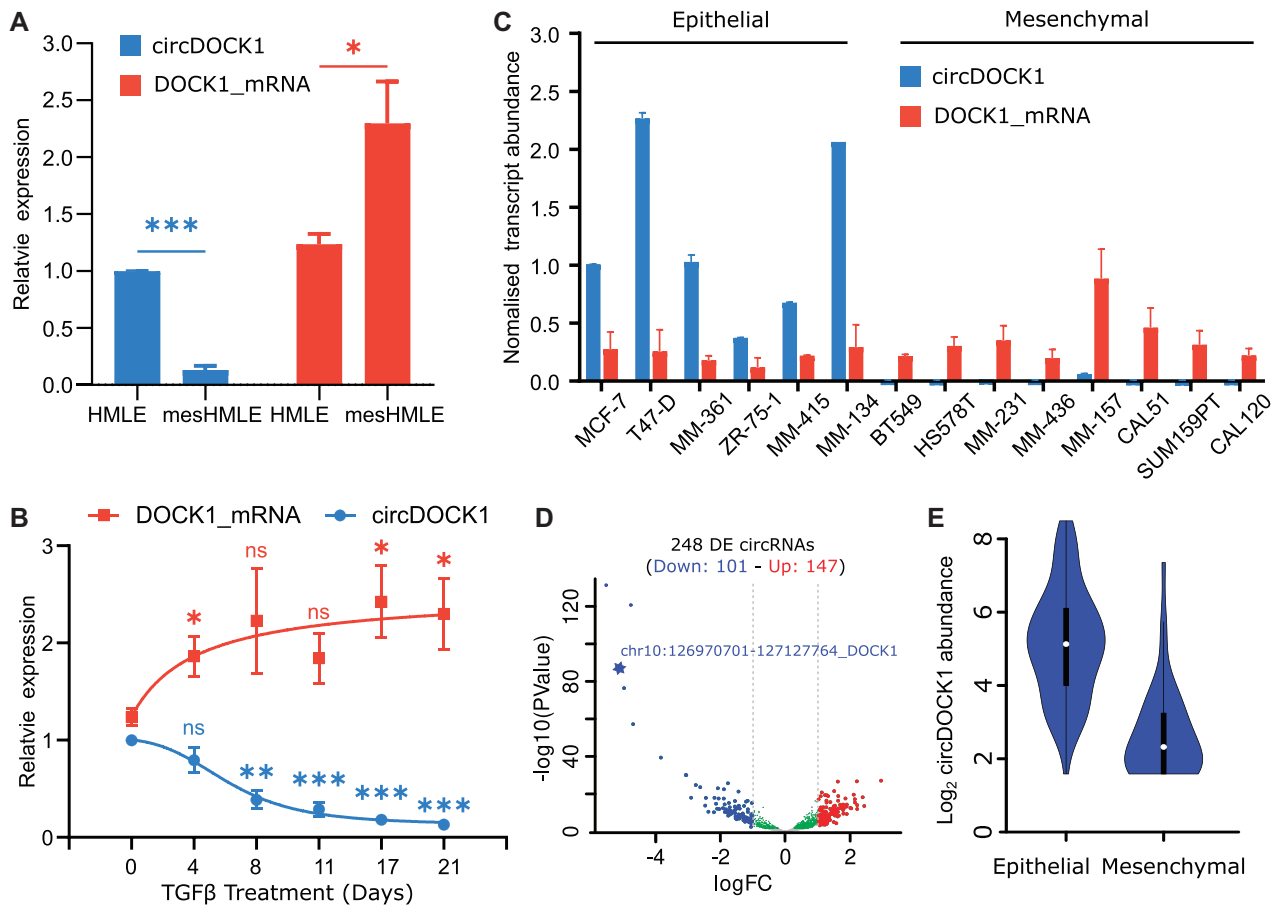


Figure 1. circDOCK1(2–27) is epithelial-specific and regulated during EMT. **(A)** circDOCK1(2–27) and DOCK1 mRNA quantitation from qPCR of RNA from HMLE cells before and after prolonged treatment with TGF- β . Mean \pm SEM. $n = 3$ biological replicates, each performed in technical triplicate. * $P < 0.05$, *** $P < 0.001$, Student's t -test, two-tailed. The qPCR Ct values are given in [Supplementary Figure S1A](#). **(B)** Time course of circDOCK1(2–27) and DOCK1 mRNA levels in TGF- β -treated HMLE cells measured by qPCR. **(C)** circDOCK1(2–27) and DOCK1 mRNA levels measured by qPCR in epithelial and mesenchymal breast cancer cell lines. The data are all expressed relative to the level of the circRNA in MCF7 cells. **(D)** Differential expression analysis of 4095 high confidence circRNAs in epithelial versus mesenchymal tumors using data from MiOncoCirc (54). Tumor samples were in-silico classified into epithelial or mesenchymal groups based on the parental gene expression of eight EMT marker genes (epithelial markers; CDH1, ESRP1, ESRP2 and CLDN7—mesenchymal markers; CDH2, VIM, ZEB1 and ZEB2) ([Supplementary Figure S1E](#)). **(E)** CircDOCK1 expression in the 191 epithelial and 159 mesenchymal primary tumor samples classified as in **(D)**.

HITS-CLIP analysis and crystal structure determination show ESRP1 binds to GGU sequences within the DOCK1 intron 1

To assess whether ESRP1 is directly involved in the biogenesis of the circDOCK1(2–27), we asked whether ESRP1 binds to DOCK1 pre-mRNA by performing HITS-CLIP analysis in HMLE cells. We observed a large peak of ESRP1 binding in intron 1, 23 kb upstream of the 3' splice site (Figure 4A). Motif search analysis of ESRP1 binding peaks across the transcriptome produced a preferred motif of UGGUGGUGG, although several other G-rich motifs were also abundant (Figure 4B). The large binding peak within DOCK1 intron 1 covered a 1.2 kb region that contains 31 copies of this motif, as well as a nearby broad set of peaks spanning 4 kb with 21 additional copies of the UGGUGGUGG motif (Figures 4A and [Supplementary Figure S4](#)). The other motifs from the global motif search were absent or very low in number in these intron 1 regions.

Because the global motif search of the ESRP1 HITS-CLIP peaks produced multiple candidate binding motifs, we sought to clarify the RNA sequence that ESRP1 binds by X-ray

crystallography of the protein–RNA complex. The full-length ESRP1 is predicted to contain an RNase H-like domain and three quasi-RNA-recognition motifs (qRRMs) similar to those in hnRNP F (35) (Figure 5A and [Supplementary Figure S5A](#)). Crystals that diffracted to 1.65Å with space group $P3_12_1$ were obtained for the ESRP1 qRRM2 domain complexed to a 12-mer RNA with sequence UGGUGGUGGUGG. The qRRM2 comprises four β -stands flanked by two α -helices at each side of the β -sheet (Figure 5B). These helices cover the β -sheet region, making the qRRM domain different from canonical RRM domains, in which the RNA-binding region is located at the β -sheet region, whereas the qRRM2 of ESRP1 binds to the RNA substrate through loop regions (Figure 5B). The structure shows contacts with just three nucleotides of the RNA, with the sequence GGU binding within a double clamp arrangement (Figure 5B–E). The three nucleotides are accommodated into a positively charged region (Figure 5C), with two loop regions important for nucleotide binding, one of which is between $\beta 1$ and $\alpha 2$ (Loop12 in Figure 5D) and another one between $\alpha 3$ and $\beta 4$ (Loop34). The three nucleotide bases face down to Loop12 and Loop34, and the phosphate

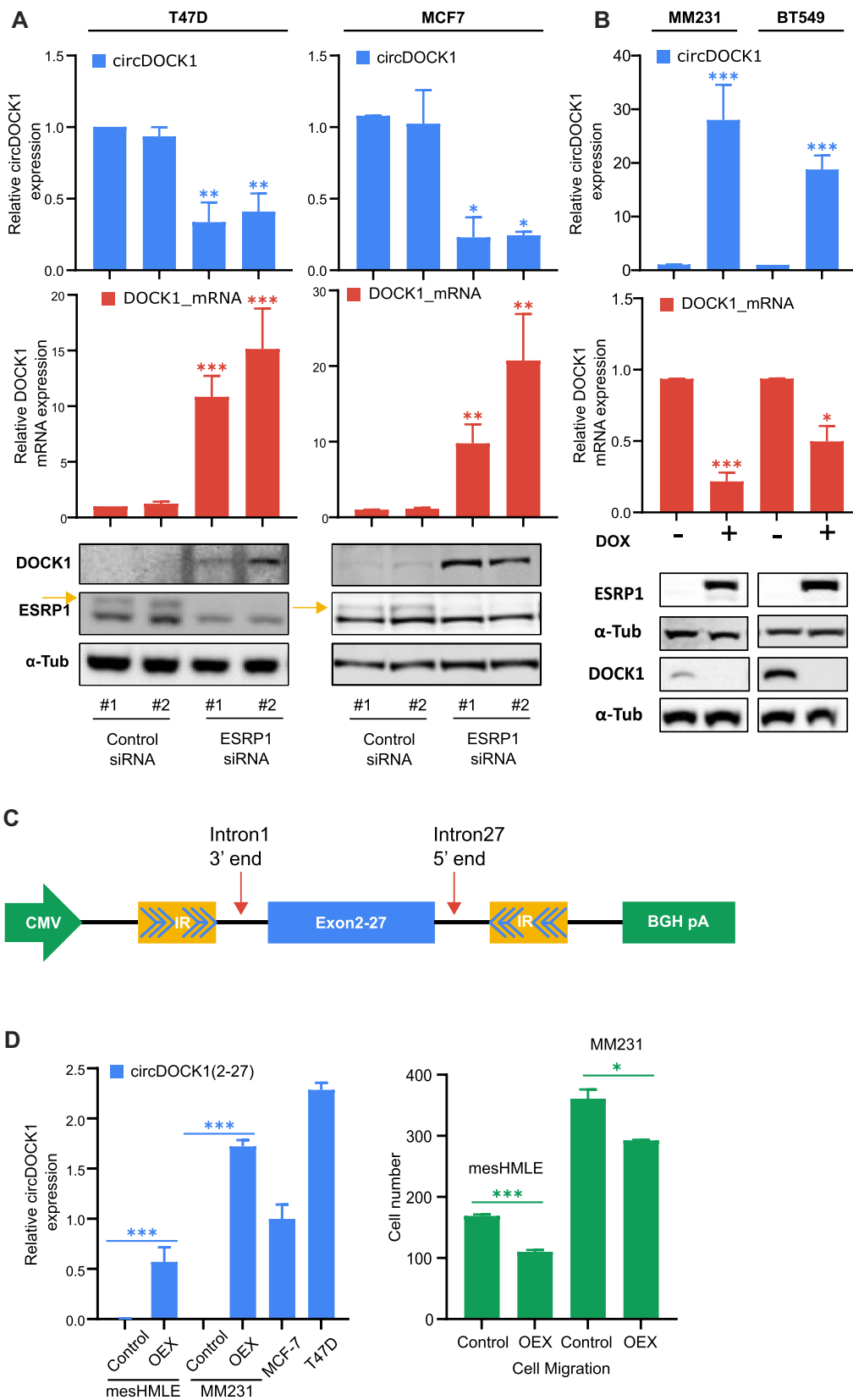


Figure 3. Effect of ESRP1 on DOCK1 expression and function. **(A)** qPCR of circDOCK1(2–27) and DOCK1 mRNA and immunoblot of DOCK1 and ESRP1 protein in epithelial cells transfected with control siRNAs or siRNAs to ESRP1. **(B)** qPCR of circDOCK1(2–27) and DOCK1 mRNA and immunoblot of DOCK1 and ESRP1 protein in mesenchymal cells with and without doxycycline induction of ESRP1. **(C)** Schematic of construct for expressing circDOCK1(2–27) in transfected cells, showing inverted repeat regions (IR) inserted into the DOCK1 intron regions. The downstream half of the inverted repeat was made by inverted insertion of a segment from intron 1. **(D)** Effect of circDOCK1(2–27) expression, measured by qPCR (left panel) on migration of mesenchymal cells measured by Transwell assay (right panel).

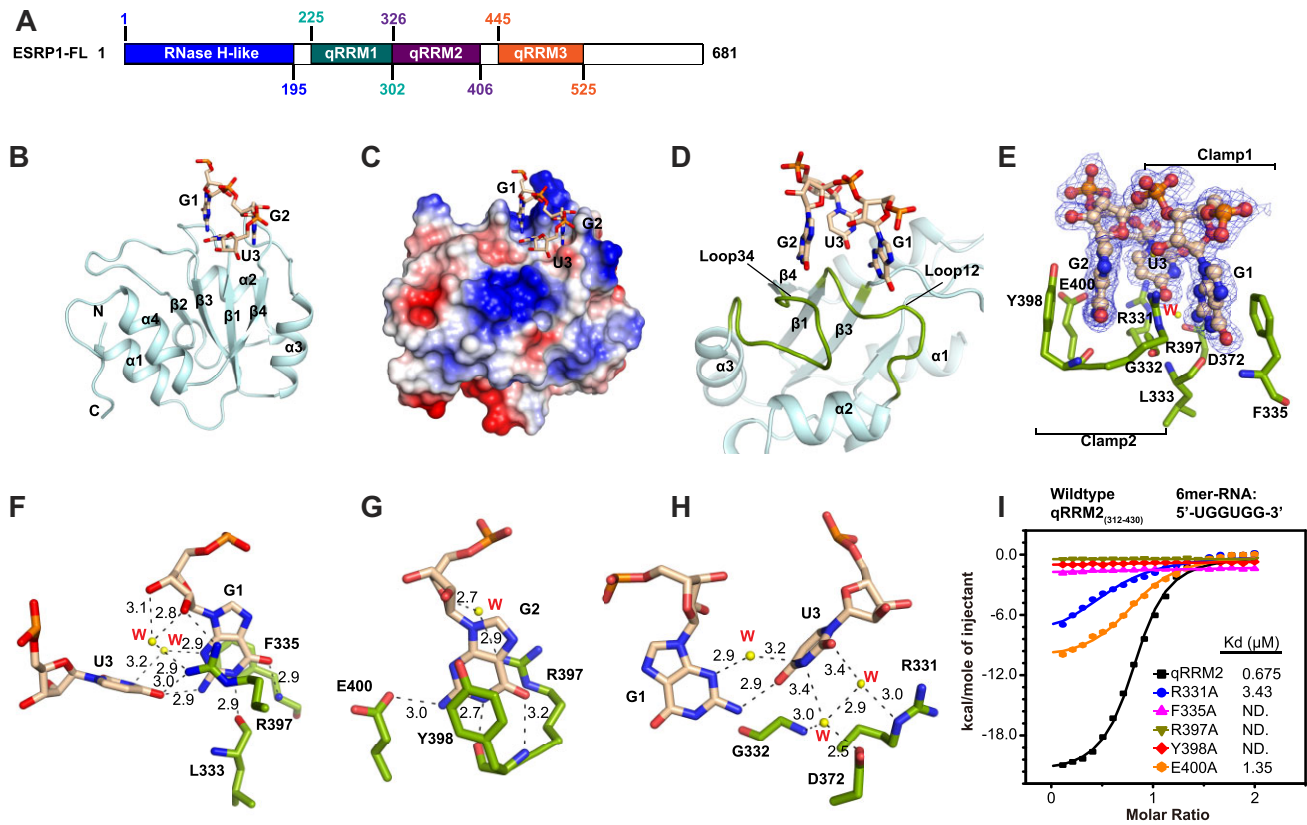


Figure 5. Crystal structure of the ESRP1 RRM2-RNA complex. **(A)** Domain organization of ESRP1. **(B)** The overall structure of ESRP1 qRRM2 with RNA. The qRRM2 is shown as ribbon and the RNA strand is shown as stick form. **(C)** Surface representation of the ESRP1 qRRM2-RNA complex with positive charged regions in blue and negative in red. **(D)** The RNA binding region in qRRM2 is coloured by green and indicated as Loop12 and Loop34. **(E)** Overall interactions between qRRM2 and RNA. The $F_o - F_c$ electron density contour level is 1.5 σ at 1.65 Å resolution. **(F-H)** Detailed interactions between qRRM2 and the 5'-GGU-3' motif. **(I)** Isothermal titration calorimetry (ITC) of wildtype qRRM2 and its mutants with 6mer-RNA substrate (5'-UGGUGG-3').

to the relative level in mesHMLE cells, with $P < 0.01$ (Figure 6C), consistent with a reduced rate of intron 1 splicing in the epithelial cells (which we call 'intron detention'). To further confirm this, we devised a qPCR assay for detained intron 1, amplifying from within intron 1 to exon 3, thereby requiring the 7 kb intron 2 to be spliced out for the PCR to be productive (Figure 6D). We compared the level of this splicing intermediate (intron 1 present, intron 2 removed) to the level of DOCK1 RNA with intron 1 removed, in the presence and absence of ESRP1. Depletion of ESRP1 in HMLE cells caused a large decrease in the level of detained intron 1 and increased the level of spliced RNA (Figure 6E). In the complementary experiment, introducing ESRP1 into mesenchymal cells resulted in an increased level of detained intron 1 and a decrease in spliced DOCK1 RNA with intron 1 removed (Figure 6F). These experiments demonstrate that ESRP1 detains the splicing of DOCK1 intron 1.

Potential mechanisms of DOCK1 circRNA formation

We next asked whether the role of ESRP1 is simply to detain intron 1, or whether it has additional roles in promoting circRNA formation. If intron 1 detention is alone sufficient to promote the formation of the circDOCK1(2–27), then artificially detaining intron 1 splicing should promote circRNA formation even in the absence of ESRP1. To prevent intron 1 splicing we used CRISPR/Cas9 to remove the intron 1 5' splice site, but leaving the bulk of the intron and its 3' splice site in-

tact (Figure 7A). This was done in HEK293 cells, which express minimal ESRP1 (<0.1% of the level expressed in HMLE cells as determined by qPCR). In two independent HEK293 clones with the 5' splice site deleted, we observed that the level of detained intron 1 was increased, as was the level of circDOCK1(2–27) (Figure 7B). This result is consistent with the role of ESRP1 being simply to detain intron 1 while transcription proceeds to exon 27, creating the possibility of back-splicing of exon 27 to exon 2.

We next considered how ESRP1 binding detains intron 1. The ESRP1 binding sites are located at a distance (>15 kb) from the 3' splice site (Figure 4A), but a possibility is that the extensive region of ESRP1 binding in intron 1 might act as a local sink for snRNPs or other essential splicing factors, thereby depleting the local concentration at the splice sites. Another possibility is that ESRP1 might competitively bind to a key spliceosome component to block its productive interactions during splicing. To gain insight into these possibilities, we performed a BioID analysis with tagged ESRP1 and searched the list of proximal proteins for evidence of such interactions with spliceosome components or other splicing factors. The list of high-confidence proximal interactors was analysed using g:Profiler, revealing strong enrichment for gene ontology molecular function (GO MF) term RNA binding (GO:0003723; $P_{\text{adj}} = 5.17 \times 10^{-78}$), though this included both cellular components (GO CC) of the cytoplasmic ribonucleoprotein granule (GO:0036464; $P_{\text{adj}} = 4.43 \times 10^{-40}$) and significant but lower enrichment of nuclear

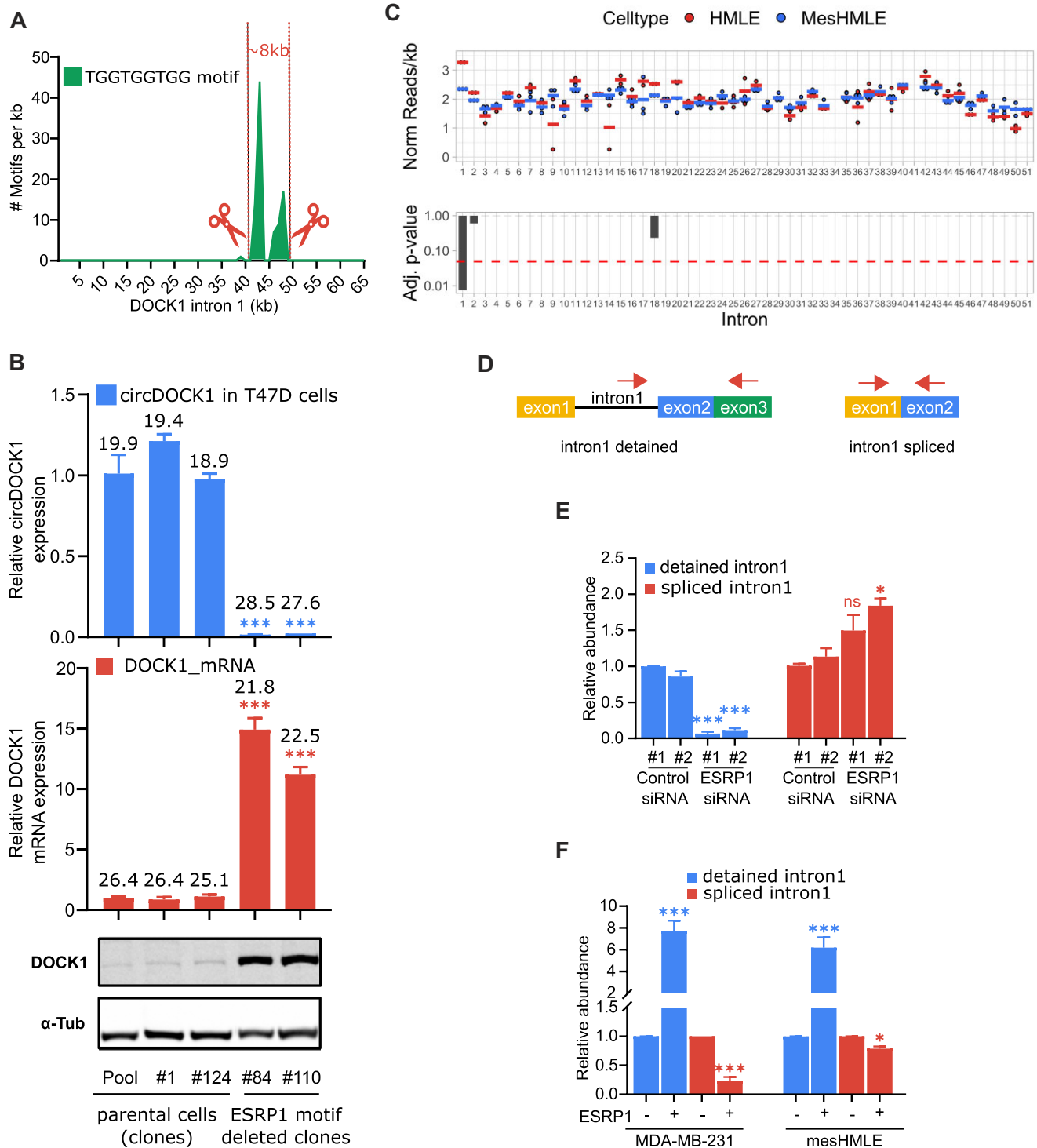


Figure 6. ESRP1 retains DOCK1 Intron 1 to facilitate circDOCK1 back splicing. **(A)** Plot of the occurrence of UGGUGGUGG motifs within intron 1, with the sites for CRISPR/Cas9-mediated deletion indicated. **(B)** The relative level of circDOCK1(2–27), DOCK1 mRNA and DOCK1 protein in parental T47D cells and in clones that have the 8kb region encompassing the ESRP1 binding peaks deleted. **(C)** Comparison of RNA sequence read numbers in each DOCK1 intron in HMLE and mesHMLE cells. The upper panel shows the average normalised read coverage for each intron from three RNA seq experiments, while the lower panel shows the significance level of the difference between coverage in HMLE versus mesHMLE cells. Only intron 1 has a large and significant difference between the cell lines. **(D)** Schematic diagram of qPCR primers detecting detected intron 1 RNA, with intron 2 spliced out but intron 1 remaining unspliced (left schematic) and DOCK1 RNA with intron 1 spliced out (right schematic). **(E)** Effect of ESRP1 depletion on intron 1 retention. The relative level of detected intron 1 (left panel) and spliced DOCK1 RNA (right panel) is shown for HMLE cells transfected with each of two control siRNAs or two ESRP1 siRNAs. **(F)** Effect of doxycycline inducible expression of ESRP1 in MDA-MB-231 and mesHMLE cells. MDA-MB-231 and mesHMLE cells stably transduced with pInducer lentivirus carrying dox-inducible ESRP1 were induced for 3 days and the change in level of detected intron 1 and spliced DOCK1 RNA measured by qPCR.

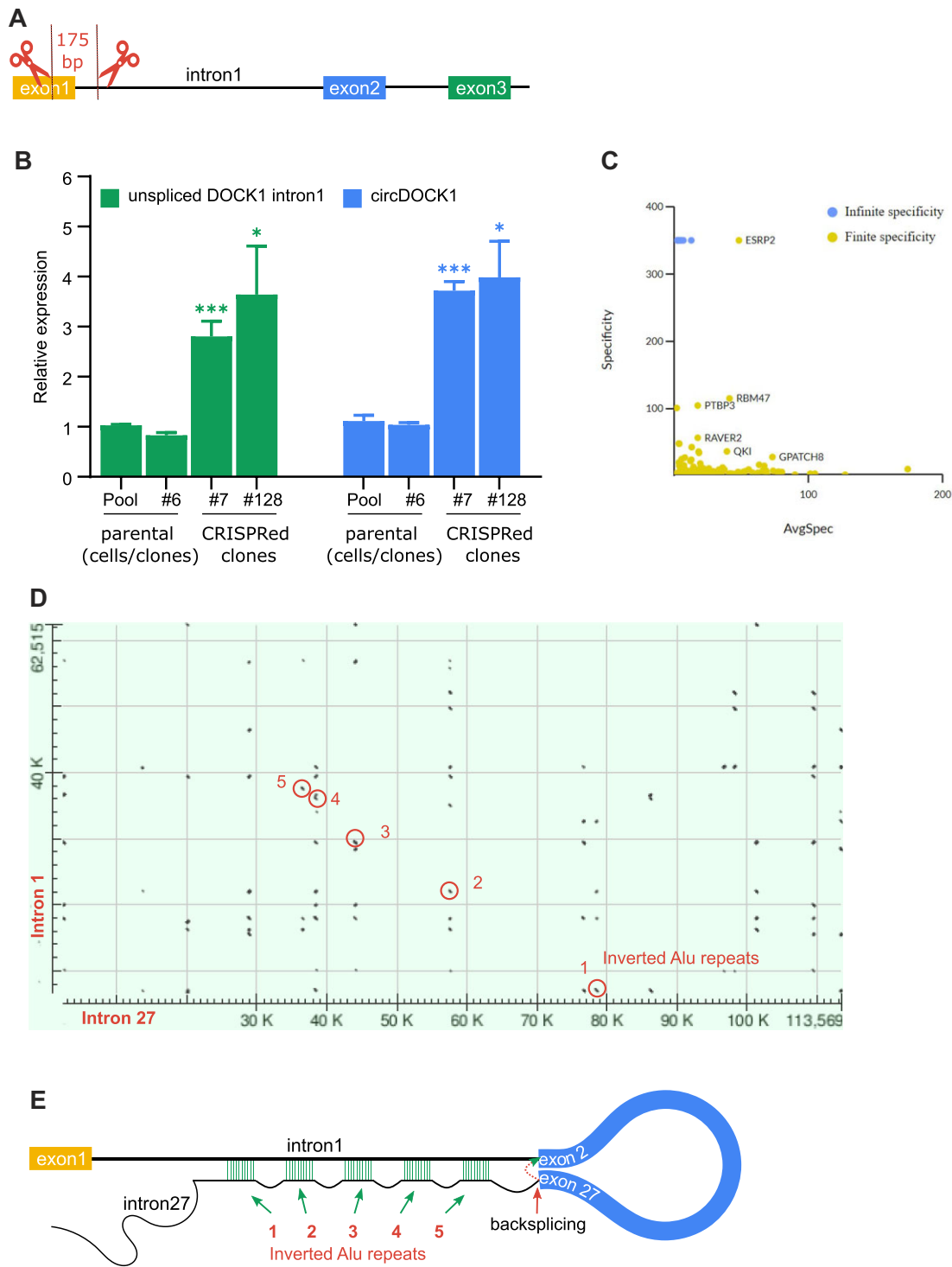


Figure 7. Intron 1 detention promotes circDOCK1(2–27) formation. **(A)** Schematic diagram of CRISPR/Cas9 cut sites removing the intron 1 5' splice site. **(B)** The relative level of detained intron 1 (green) and circDOCK1(2–27) (blue) measured by qPCR in HEK293T cells and in two independent clones that had the intron 1 splice site deleted. **(C)** Scatterplot showing the most specifically enriched proximal interactions for ESRP1 when compared against all baits in the humancellmap.org. **(D)** Alignment of intron 1 sequence with intron 27, generated from NCBI blastn. Selected Alu inverted repeats that might give rise to base pairing interactions are circled. **(E)** Schematic of potential base pairing between Alu repeats in DOCK1 intron 1 and intron 27.

components including the spliceosomal complex (GO:0005681; $P_{\text{adj}} = 2.87 \times 10^{-12}$). KEGG and REACTOME pathway analysis both recovered spliceosome/mRNA splicing as top enriched categories (KEGG:03040; $P_{\text{adj}} = 2.66 \times 10^{-7}$; REAC:R-HSA-72163; $P_{\text{adj}} = 1.33 \times 10^{-17}$), though this was dominated by Heterogeneous nuclear ribonucleoproteins (HNRNPs) rather than core spliceosome components, with a few exceptions. Interpretation of the ESRP1 BioID data in the context of the humancellmap.org (a large project that aims to systematically use BioID data to reveal subcellular organization (31)), revealed the closest similarity in prey recovery profiles with HNRNPA1 (Jaccard distance 0.733; 52 common partners/195 total partners across the two baits). The prey recovery profile similarity to HNRNPA1 was higher than for spliceosome component DHX8 (Jaccard 0.958; 11 common partners/261; scatterplots in [Supplementary Figure S7](#)), further confirming these observations. In agreement with this observation, the most specifically enriched proximal interactions with ESRP1 – compared to all baits in the humancellmap.org (31) – were with alternative splicing regulators, rather than core spliceosome components (Figure 7C, [Supplementary Table S2](#)). Though it is not possible to exclude artifacts associated with tagging ESRP1 with the BioID enzyme, together, these data suggest that the function of ESRP1 may not be to disrupt the core spliceosome function directly, but rather to modify the RNP landscape of intron 1 in a way that disfavors communication of exon 2 with exon 1.

We next considered how the coupling of exon 27 to exon 2 might be achieved. The two principal mechanisms described to date for promoting backsplicing are base pairing between inverted Alu repeats in the introns flanking the back-spliced exons (37,38), potentially aided by binding of the splicing factor SFPQ (39), and looping mediated by a dimeric RNA binding protein (4). Although ESRP1 is not known to be dimeric, it is conceivable that individual qRRM domains (of which there are three) can bind to separated GGU motifs to promote looping. However, only two very small peaks of ESRP1 binding within exon 28 were detected in the HITS-CLIP analysis and these peaks were less than 1% of the size of the intron 1 peak. Because this does not seem to indicate a strong propensity for ESRP1-mediated looping of intron 27 to intron 1, we searched for inverted Alu repeats that may cause such looping. Using the NCBI BLASTn tool to align the intron 1 sequence to the intron 27 sequence revealed multiple examples of inverted Alu repeats (Figure 7D). Thus the backsplicing of exon 27 to exon 2 is likely to be augmented by base pairing between intron 1 and intron 27 (Figures 7E and [Supplementary Figure S8](#)). The fact that ENCODE eCLIP data (40) indicate binding of SFPQ throughout both intron 1 and intron 27 of DOCK1 in HepG2 cells is consistent with this mechanism. The cooperation of multiple RNA-binding proteins and splicing factors with inverted intronic repeats to regulate circRNA formation has been previously shown to occur in control of production of a circRNA from the *Drosophila* laccase2 gene (41), further supporting the concept of combined roles of RNA binding proteins and hairpin formation in regulating the formation of some circRNAs.

Discussion

EMT confers motility on cells to allow tissue remodelling during embryogenesis, but can be recapitulated in part (called

epithelial plasticity) by cancer cells to promote tissue invasion and metastasis (42). EMT is a highly coordinated process with many contributing regulators and effectors that act on cytoskeletal components to reconfigure the cytoarchitecture and enable cell motility, with DOCK1 protein being part of this crucial regulatory network. We show here that alternative splicing of the DOCK1 transcript to generate the circDOCK1(2–27) circular RNA is highly regulated in EMT and contributes to the regulation of the migratory capacity of cancer cells.

ESRP1 has been shown to contribute to enforcing epithelial phenotype (43–45), but has not been previously linked to regulation of DOCK1. However, the regulation and functions we have ascribed to circDOCK1(2–27) align well with the known roles of both ESRP1 and DOCK1. The DOCK1 protein is a guanine exchange factor (GEF), which in conjunction with its binding partner ELMO1 activates the GTPase Rac1, leading to cytoskeletal rearrangements that promote cell membrane spreading and cell migration (34,46). ESRP1 has been shown to control alternative splicing of multiple genes during EMT, many of which contribute to alterations in the actin cytoskeleton and cell motility (43,47,48). The effects we see of ESRP1 on both DOCK1 and circDOCK1(2–27) expression contribute to effects on cell motility, and provide a further example of the coordinated, multicomponent control of the cytoskeleton and motility that is evident in EMT. Since these are crucial capacities of cells that contribute to cancer metastasis, it would be interesting to assess the ability of circDOCK1(2–27) to suppress carcinoma metastasis.

Our crystallography data confirm that the ESRP1 qRRM2 recognises GGU rather than UGG, which is consistent with the results of previous SELEX and HITS-CLIP studies (44,49), although those studies, like our HITS-CLIP study, tended to identify the GGU sequence as being within a slightly longer GU-rich context. The very high degree of sequence similarity between the different qRRM domains of ESRP1 suggests to us that the RNA contacts will be very similar for all domains, but structural studies of intact, or multi-domain regions of the ESRP1 protein may clarify whether the additional flanking bases contribute to binding, and also whether multiple domains can bind to multiple adjacent GGU motifs to enhance affinity and/or affect functional effects of the RNA-bound ESRP1.

We note that while we did not see any effect of ESRP2 on circDOCK1(2–27), it is possible that any effect of depletion of ESRP2 was masked by the more abundant ESRP1 in HMLE cells. Given the high degree of sequence similarity between ESRP1 and ESRP2, we expect they would have similar RNA-binding profiles and that ESRP2 could regulate circDOCK1(2–27) formation in cells that have more abundant ESRP2. We speculate that the proximity labelling of ESRP2 by the ESRP1 bait in the BioID experiment we performed in HEK293 cells may be due to the two family members binding to adjacent sites, such as the reiterated sites we observed by HITS-CLIP analysis in the DOCK1 intron 1.

Whereas most circRNAs are expressed at very low levels such that despite their long half-lives they are much less abundant than their cognate mRNAs, we found that the backsplicing to produce the circDOCK1(2–27) circRNA is unusually efficient in cells that express ESRP1. The circDOCK1(2–27) ranks as the second most abundant circRNA in HMLE cells (behind circHIPK3), and among the top 10 circRNAs in expression relative to that of the cognate Mrna (4). The mode

and degree of regulation of circDOCK1(2–27) is also unusual for a circRNA in that it is strongly reciprocal to the expression of mRNA from the host gene. This suggests its function is antithetically related to that of the cognate mRNA and this indeed appears to be very much the case since we find evidence of two simultaneous mechanisms that oppose function of the host gene in epithelial cells. Firstly, the formation of the circRNA reduces DOCK1 protein expression from the host gene by diverting transcripts from mRNA production to circRNA production. The efficiency with which this occurs in epithelial cells is remarkable given the large separation in sequence distance between the two exons that are ligated to form the RNA circle, with exon 2 separated from exon 27 by 157 kb. Also remarkable is the large number of exons in circDOCK1(2–27) (all 26 of which we have found are retained in the circRNA); most circRNAs have fewer than 5 exons and 99% of circRNAs are comprised of fewer than 12 exons (50,51). Secondly, the circRNA per se suppresses cell migration, as evidenced by the effect of its ectopic expression in inhibiting migration of mesenchymal cells, which has been previously shown in MDA-MB-231 breast cancer cells (52) and we show here in both mesHMLE and MDA-MB-231 cells.

Deep sequencing has revealed that many genes can produce multiple circRNAs with different backspliced ligations resulting in different exons incorporated, and the DOCK1 gene is typical in this regard. A number of previous publications have reported effects of ‘circDOCK1’ in various cancers, however those reports relate to different circRNAs and none of the reports, apart from our previous report on circDOCK1 in breast cancer cells (52) are on circDOCK1(2–27), but instead describe circRNAs with different circBase identifiers that are much less abundant than circDOCK1(2–27) (which has the circBase ID hsa_circ_0020397). To avoid possible ambiguity or confusion regarding circRNA identity, we suggest that any report focusing on the properties or functions of a specific circRNA should include a definitive description of the exons involved in backsplice formation, as recently proposed (53). Moreover, since many detected circRNAs are only expressed at an exceedingly low level, we propose a quantitative measure of the circRNA abundance be reported, to support the likelihood it is present in cells at a level commensurate with the proposed function.

Data availability

HITS-CLIP RNA-seq data were deposited to the NCBI GEO repository under accession number GSE226538. The mass spectrometry data was deposited to ProteomeXchange through partner MassIVE (massive.ucsd.edu) and assigned identifiers PXD040537 and MSV000091390.

Supplementary data

Supplementary Data are available at NAR Online.

Acknowledgements

Author contributions: D.L. designed and performed biological experiments, M.M. performed Western Blotting, B.K.D. and A.B. performed the HITS-CLIP experiments, K.P. and J.T. performed bioinformatic analysis, W.G. and B.W. performed the X-ray crystallography, with guidance from D.P., B.J.A.D. and A.-C.G. performed and analysed the BioID experiments, V.C.

and S.C. provided initial data and reagents, P.G. contributed to conceptualisation, G.G. conceived and supervised the study and wrote the manuscript with input and editing from all authors.

Funding

National Health and Medical Research Council of Australia [GNT1089167, GNT1118170 to G.G.]; National Breast Cancer Foundation [IN-16-072 to G.G.]; National Natural Science Foundation of China [31900435 to B.W.]; Guangdong Science and Technology Department [2020B1212060018, 2020B1212030004 to B.W.]; work on the Gingras lab was supported by the National Science and Engineering Research Council of Canada (NSERC) [RGPIN-2019-06297]; K.A.P. was supported by a Royal Adelaide Hospital Florey Fellowship; B.J.A.D. was supported by an NSERC studentship; A.-C.G. is the Canada Research Chair in Functional Proteomics. Funding for open access charge: Centre for Cancer Biology, University of South Australia.

Conflict of interest statement

None declared.

References

- Vicens,Q. and Westhof,E. (2014) Biogenesis of circular RNAs. *Cell*, **159**, 13–14.
- Xu,C. and Zhang,J. (2021) Mammalian circular RNAs result largely from splicing errors. *Cell Rep.*, **36**, 109439.
- Ashwal-Fluss,R., Meyer,M., Pamudurti,N.R., Ivanov,A., Bartok,O., Hanan,M., Evantal,N., Memczak,S., Rajewsky,N. and Kadener,S. (2014) circRNA biogenesis competes with pre-mRNA splicing. *Mol. Cell*, **56**, 55–66.
- Conn,S.J., Pillman,K.A., Toubia,J., Conn,V.M., Salmanidis,M., Phillips,C.A., Roslan,S., Schreiber,A.W., Gregory,P.A. and Goodall,G.J. (2015) The RNA binding protein quaking regulates formation of circRNAs. *Cell*, **160**, 1125–1134.
- Teplova,M., Hafner,M., Teplov,D., Essig,K., Tuschl,T. and Patel,D.J. (2013) Structure-function studies of STAR family quaking proteins bound to their in vivo RNA target sites. *Genes Dev.*, **27**, 928–940.
- Tran,H., Gourrier,N., Lemercier-Neuillet,C., Dhaenens,C.M., Vautrin,A., Fernandez-Gomez,F.J., Arandel,L., Carpentier,C., Obriot,H., Eddarkaoui,S., *et al.* (2011) Analysis of exonic regions involved in nuclear localization, splicing activity, and dimerization of muscleblind-like-1 isoforms. *J. Biol. Chem.*, **286**, 16435–16446.
- Chen,S., Huang,V., Xu,X., Livingstone,J., Soares,F., Jeon,J., Zeng,Y., Hua,J.T., Petricca,J., Guo,H., *et al.* (2019) Widespread and functional RNA circularization in localized prostate cancer. *Cell*, **176**, 831–843.
- Pillman,K.A., Phillips,C.A., Roslan,S., Toubia,J., Dredge,B.K., Bert,A.G., Lumb,R., Neumann,D.P., Li,X., Conn,S.J., *et al.* (2018) miR-200/375 control epithelial plasticity-associated alternative splicing by repressing the RNA-binding protein Quaking. *EMBO J.*, **37**, e99016.
- Van Nostrand,E.L., Shishkin,A.a., Pratt,G.a., Nguyen,T.B. and Yeo,G.W. (2017) Variation in single-nucleotide sensitivity of eCLIP derived from reverse transcription conditions. *Methods*, **126**, 29–37.
- Martin,M. (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal*, **17**, 10.
- Smith,T., Heger,A. and Sudbery,I. (2017) UMI-tools: modeling sequencing errors in Unique Molecular Identifiers to improve quantification accuracy. *Genome Res.*, **27**, 491–499.

12. Dobin,A., Davis,C.A., Schlesinger,F., Drenkow,J., Zaleski,C., Jha,S., Batut,P., Chaisson,M. and Gingeras,T.R. (2013) STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, **29**, 15–21.
13. Li,H., Handsaker,B., Wysoker,A., Fennell,T., Ruan,J., Homer,N., Marth,G., Abecasis,G. and Durbin,R. (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.
14. Zhang,Y., Liu,T., Meyer,C.A., Eeckhoutte,J., Johnson,D.S., Bernstein,B.E., Nussbaum,C., Myers,R.M., Brown,M., Li,W., *et al.* (2008) Model-based analysis of ChIP-Seq (MACS). *Genome Biol.*, **9**, R137.
15. Robinson,J.T., Thorvaldsdóttir,H., Winckler,W., Guttman,M., Lander,E.S., Getz,G. and Mesirov,J.P. (2011) Integrative genomics viewer. *Nat. Biotechnol.*, **29**, 24–26.
16. Heinz,S., Benner,C., Spann,N., Bertolino,E., Lin,Y.C., Laslo,P., Cheng,J.X., Murre,C., Singh,H. and Glass,C.K. (2010) Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol. Cell*, **38**, 576–589.
17. Minor,W., Cymborowski,M., Otwinowski,Z. and Chruszcz,M. (2006) HKL-3000: the integration of data reduction and structure solution - From diffraction images to an initial model in minutes. *Acta Crystallogr. Sect. D Biol. Crystallogr.*, **62**, 859–866.
18. Kabsch,W. (2010) research papers XDS research papers. *Acta Crystallogr. Sect. D Biol. Crystallogr.*, **66**, 125–132.
19. Potterton,E., Briggs,P., Turkenburg,M. and Dodson,E. (2003) A graphical user interface to the CCP4 program suite. *Acta Crystallogr. Sect. D Biol. Crystallogr.*, **59**, 1131–1137.
20. McCoy,A.J., Grosse-Kunstleve,R.W., Adams,P.D., Winn,M.D., Storoni,L.C. and Read,R.J. (2007) Phaser crystallographic software. *J. Appl. Crystallogr.*, **40**, 658–674.
21. Murshudov,G.N., Skubák,P., Lebedev,A.a., Pannu,N.S., Steiner,R.a., Nicholls,R.a., Winn,M.D., Long,F. and Vagin,A.a. (2011) REFMAC5 for the refinement of macromolecular crystal structures. *Acta Crystallogr. Sect. D Biol. Crystallogr.*, **67**, 355–367.
22. Emsley,P. and Cowtan,K. (2004) Coot: model-building tools for molecular graphics. *Acta Crystallogr. Sect. D Biol. Crystallogr.*, **60**, 2126–2132.
23. Afonine,P.V., Grosse-Kunstleve,R.W., Echols,N., Headd,J.J., Moriarty,N.W., Mustyakimov,M., Terwilliger,T.C., Urzhumtsev,A., Zwart,P.H. and Adams,P.D. (2012) Towards automated crystallographic structure refinement with phenix.refine. *Acta Crystallogr. Sect. D Biol. Crystallogr.*, **68**, 352–367.
24. Lambert,J.P., Tucholska,M., Go,C., Knight,J.D.R. and Gingras,A.C. (2015) Proximity biotinylation and affinity purification are complementary approaches for the interactome mapping of chromatin-associated protein complexes. *J. Proteomics*, **118**, 81–94.
25. Liu,G., Knight,J.D.R., Zhang,J.P., Tsou,C.C., Wang,J., Lambert,J.P., Larsen,B., Tyers,M., Raught,B., Bandeira,N., *et al.* (2016) Data independent acquisition analysis in ProHits 4.0. *J. Proteomics*, **149**, 64–68.
26. Perkins,D.N., Pappin,D.J.C., Creasy,D.M. and Cottrell,J.S. (1999) Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis*, **20**, 3551–3567.
27. Eng,J.K., Jahan,T.a. and Hoopmann,M.R. (2013) Comet: an open-source MS/MS sequence database search tool. *Proteomics*, **13**, 22–24.
28. Shteynberg,D., Deutsch,E.W., Lam,H., Eng,J.K., Sun,Z., Tasman,N., Mendoza,L., Moritz,R.L., Aebersold,R. and Nesvizhskii,A.I. (2011) iProphet: multi-level integrative analysis of shotgun proteomic data improves peptide and protein identification rates and error estimates. *Mol. Cell. Proteomics*, **10**, M111.007690.
29. Teo,G., Liu,G., Zhang,J., Nesvizhskii,A.I., Gingras,A.C. and Choi,H. (2014) SAINTexpress: improvements and additional features in Significance Analysis of INteractome software. *J. Proteomics*, **100**, 37–43.
30. Knight,J.D.R., Liu,G., Zhang,J.P., Pasculescu,A., Choi,H. and Gingras,A.C. (2015) A web-tool for visualizing quantitative protein-protein interaction data. *Proteomics*, **15**, 1432–1436.
31. Go,C.D., Knight,J.D.R., Rajasekharan,A., Rathod,B., Hesketh,G.G., Abe,K.T., Youn,J.Y., Samavarchi-Tehrani,P., Zhang,H., Zhu,L.Y., *et al.* (2021) A proximity-dependent biotinylation map of a human cell. *Nature*, **595**, 120–124.
32. Raudvere,U., Kolberg,L., Kuzmin,I., Arak,T., Adler,P., Peterson,H. and Vilo,J. (2019) G:profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update). *Nucleic Acids Res.*, **47**, W191–W198.
33. Warzecha,C.C., Sato,T.K., Nabet,B., Hogenesch,J.B. and Carstens,R.P. (2009) ESRP1 and ESRP2 are epithelial cell-type-specific regulators of FGFR2 splicing. *Mol. Cell*, **33**, 591–601.
34. Koubek,E.J. and Santy,L.C. (2022) Actin Up: an overview of the Rac GEF Dock1/Dock180 and its role in cytoskeleton rearrangement. *Cells*, **11**, 3565.
35. Dominguez,C., Fiset,J.F., Chabot,B. and Allain,F.H.T. (2010) Structural basis of G-tract recognition and encaging by hnRNP F quasi-RRMs. *Nat. Struct. Mol. Biol.*, **17**, 853–861.
36. Herzel,L., Ottoz,D.S.M., Alpert,T. and Neugebauer,K.M. (2017) Splicing and transcription touch base: co-transcriptional spliceosome assembly and function. *Nat. Rev. Mol. Cell Biol.*, **18**, 637–650.
37. Jeck,W., Sorrentino,J., Wang,K. and Slevin,M. (2013) Circular RNAs are abundant, conserved, and associated with ALU repeats. *RNA*, **19**, 141–157.
38. Zhang,X.-O., Wang,H.-B., Zhang,Y., Lu,X., Chen,L.-L. and Yang,L. (2014) Complementary sequence-mediated exon circularization. *Cell*, **159**, 134–147.
39. Stagsted,L.V.W., O’leary,E.T., Ebbesen,K.K. and Hansen,T.B. (2021) The rna-binding protein sfpq preserves long-intron splicing and regulates circrna biogenesis in mammals. *eLife*, **10**, e63088.
40. Dunham,I., Kundaje,A., Aldred,S.F., Collins,P.J., Davis,C.A., Doyle,F., Epstein,C.B., Frietze,S., Harrow,J., Kaul,R., *et al.* (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature*, **489**, 57–74.
41. Kramer,M.C., Liang,D., Tatomer,D.C., Gold,B., March,Z.M., Cherry,S. and Wilusz,J.E. (2015) Combinatorial control of Drosophila circular RNA expression by intronic repeats, hnRNPs, and SR proteins. *Genes Dev.*, **29**, 2168–2182.
42. Yang,J., Antin,P., Berx,G., Blanpain,C., Brabletz,T., Bronner,M., Campbell,K., Cano,A., Casanova,J., Christofori,G., *et al.* (2020) Guidelines and definitions for research on epithelial-mesenchymal transition. *Nat. Rev. Mol. Cell Biol.*, **20**, 341–352.
43. Warzecha,C.C., Jiang,P., Amirikian,K., Dittmar,K.a., Lu,H., Shen,S., Guo,W., Xing,Y. and Carstens,R.P. (2010) An ESRP-regulated splicing programme is abrogated during the epithelial-mesenchymal transition. *EMBO J.*, **29**, 3286–3300.
44. Dittmar,K.a., Jiang,P., Park,J.W., Amirikian,K., Wan,J., Shen,S., Xing,Y. and Carstens,R.P. (2012) Genome-wide determination of a broad ESRP-regulated posttranscriptional Network by high-throughput sequencing. *Mol. Cell. Biol.*, **32**, 1468–1482.
45. Lee,S.K., Cieply,B., Yang,Y., Peart,N., Glaser,C., Chan,P. and Carstens,R.P. (2018) Esrp1-regulated splicing of Arhgef11 isoforms is required for epithelial tight junction integrity. *Cell Rep.*, **25**, 2417–2430.
46. Gumieny,T.L., Brugnera,E., Tosello-Trampont,A.C., Kinchen,J.M., Haney,L.B., Nishiwaki,K., Walk,S.F., Nemergut,M.E., Macara,I.G., Francis,R., *et al.* (2001) CED-12/ELMO, a novel member of the CrkII/Dock180/Rac pathway, is required for phagocytosis and cell migration. *Cell*, **107**, 27–41.
47. Bebe,T.W., Park,J.W., Sheridan,K.I., Warzecha,C.C., Cieply,B.W., Rohacek,A.M., Xing,Y. and Carstens,R.P. (2015) The splicing regulators Esrp1 and Esrp2 direct an epithelial splicing program essential for mammalian development. *eLife*, **4**, e08954.

48. Ishii,H., Saitoh,M., Sakamoto,K., Kondo,T., Katoh,R., Tanaka,S., Motizuki,M., Masuyama,K. and Miyazawa,K. (2014) Epithelial splicing regulatory proteins 1 (ESRP1) and 2 (ESRP2) suppress cancer cell motility via different mechanisms. *J. Biol. Chem.*, **289**, 27386–27399.
49. Peart,N.J., Hwang,J.Y., Quesnel-Vallières,M., Sears,M.J., Yang,Y., Stoilov,P., Barash,Y., Park,J.W., Lynch,K.W. and Carstens,R.P. (2022) The global Protein-RNA interaction map of ESRP1 defines a post-transcriptional program that is essential for epithelial cell function. *iScience*, **25**, 105205.
50. Ragan,C., Goodall,G.J., Shirokikh,N.E. and Preiss,T. (2019) Insights into the biogenesis and potential functions of exonic circular RNA. *Sci. Rep.*, **9**, 2048.
51. Xin,R., Gao,Y., Gao,Y., Wang,R., Kadash-Edmondson,K.E., Liu,B., Wang,Y., Lin,L. and Xing,Y. (2021) isoCirc catalogs full-length circular RNA isoforms in human transcriptomes. *Nat. Commun.*, **12**, 266.
52. Kurosaki,M., Terao,M., Liu,D., Zanetti,A., Guarrera,L., Bolis,M., Gianni',M., Paroni,G., Goodall,G.J. and Garattini,E. (2021) A dock1 gene-derived circular rna is highly expressed in luminal mammary tumours and is involved in the epithelial differentiation, growth, and motility of breast cancer cells. *Cancers (Basel)*, **13**, 5325.
53. Chen,L., Bindereif,A., Bozzoni,I., Chang,H.Y., Matera,A.G., Gorospe,M., Hansen,T.B., Kjems,J., Ma,X., Pek,J.W., *et al.* (2023) A guide to naming eukaryotic circular RNAs. *Nat. Cell Biol.*, **25**, 1–5.
54. Vo,J.N., Cieslik,M., Zhang,Y., Shukla,S., Xiao,L., Zhang,Y., Wu,Y.M., Dhanasekaran,S.M., Engelke,C.G., Cao,X., *et al.* (2019) The landscape of circular RNA in cancer. *Cell*, **176**, 869–881.