

Ταυτοποίηση και λειτουργικός χαρακτηρισμός των μεγάλων μη κωδικών RNA που εμπλέκονται στην ομοιόσταση και καρκινογένεση του εντέρου

Διδακτορική διατριβή

Αλέξανδρος Γαλάρας,
ΑΜΔ: Δ1818007

Identification and functional characterization of long non-coding RNAs involved in colorectal homeostasis and carcinogenesis

PhD dissemination

Alexandros Galaras,
ID: Δ1818007

Examination Committee

Antonis Giakountis^{2,3}, Assistant Professor of Molecular Biology and Genomics, Department of Biochemistry & Biotechnology, University of Thessaly

Pantelis Hatzis^{1,2,3}, Researcher B', Institute For Fundamental Biomedical Research

Nikolaos Balatsos^{2,3}, Associate Professor of Biochemistry, Department of Biochemistry & Biotechnology, University of Thessaly

Anna-Maria Psarra³, Professor of Biochemistry, Department of Biochemistry & Biotechnology, University of Thessaly

Theodora Sideri³, Assistant Professor of Biochemistry, Faculty of Medicine, University of Thessaly

Mihalis Verykokakis³, Researcher B', Institute For Fundamental Biomedical Research

Georgios Hamilos³, Professor of Clinical Microbiology and Microbial Pathogenesis, School of Medicine, University of Crete

1. Supervisor
2. Three-member Committee
3. Seven-member Committee

Acknowledgements

Completing my PhD thesis was among the greatest challenges that I have chosen to undertake throughout my life. Despite the difficult moments and the more anxious days, it remains a rewarding journey and a hard-earned conquest. Along the way, the guidance, mentorship, and collaboration of multiple individuals proved essential. I am extremely grateful that my academic journey began at the Biomedical Sciences Research Center “Alexander Fleming”, an institute that hosts remarkable scientists, but more importantly, brilliant people and dear friends. Even though a PhD thesis is regarded as a personal task, its details are full of these bright people. In this section I hope that I will acknowledge the most important among them.

First and foremost, I would like to express my gratitude for Dr. **Pantelis Hatzis**, Researcher B at BSRC “Alexander Fleming”, for his unwavering support from the very beginning. His continuous trust helped me acknowledge myself and grow into the scientist I am today. He gave me all the freedom to be creative, pursue my ideas, collaborate with amazing scientists and become independent.

I am also deeply thankful to Dr. **Mihalis Verykokakis**, who shared his passion for immunology, mentored me, and included me in his awesome projects. Many of the successes and recognition I received were possible due to his continuous support.

I am extremely grateful to Dr. **Antonis Giakountis**, Assistant Professor in Molecular Biology and Genomics. I consider myself fortunate to have met Antonis as a Postdoctoral researcher in the group and then became my PhD supervisor. His energy, enthusiasm and dedication continue to inspire me to pursue my passion for research. I am always looking forward to our fruitful discussions and I am more than proud to be one of his PhD students.

Also, probably one of the most important people in my academic path is Dr. **Panagiotis Moulos**, the “Bioinformatics Guru” at BSRC “Alexander Fleming”. Panagiotis was my first mentor in the Hatzis Lab and through his patience and unlimited solutions, guided me in bioinformatics analysis even when much of it felt daunting for me.

Next, I would like to thank the other members of the thesis committee, especially Dr. **Nikolaos Balatsos**, Associate Professor in Biochemistry for his guidance and suggestions throughout the years.

I would like to specially thank Dr. **Vasiliki Zarkou**, who was my mentor in the laboratory for four years. Apart from an amazing colleague, Vicky is an extraordinary person and a great friend, and I consider myself lucky to have worked alongside her. Vicky did not just train me on the bench and contributed to many of the experiments, but also shaped many other aspects of my life.

Moreover, I am extremely grateful for the huge help of my collaborator and friend **Marios Itsios**. I am very proud of seeing his evolution from a Master student to an amazing independent young scientist. I am looking forward to his breakthroughs!

I would like to further thank all the Master’s students that I have had the pleasure to supervise throughout the years, for their hard work, resilience, and (occasionally tested) patience. I am proud of their scientific impact and always makes me happy to witness how far they have come. Special thanks to the exceptional **Anna Vathrakokili** and **Marili Vasilogianni** for their invaluable work in the first part of the PhD thesis. I am also deeply grateful to my dear

friend **Katerina Kalandrani**, the talented **Katerina Pravi** and the resilient **Giannis Chiotakakos** for their significant contributions to the second part. Special thanks to our former lab member **Eszter Szantai** for her valuable help, and mentoring during her postdoctoral days.

I owe heartfelt thanks to all the past and present members of the Fousteri and Verykokakis Labs – **Maria Fousteri**, **Giorgos Nikolopoulos**, **Giorgos Stefos**, **Matthie Lavigne**, **Dimitris Konstantopoulos**, **Chloe Synacheri**, **Nellie Angelova**, **Smaragda Fanourgaki**, **Maria Papanastasatou**, **Ariana Ntatsopoulou**, and **Dafni Anagnostopoulou** – for our daily laughs, discussions and troubleshooting sessions. I am especially grateful to **Anastasios Liakos** and **Marianthi Gioulbasani** for our collaborations and the unforgettable moments we shared inside and outside the lab.

Next, I would like to warmly thank Eleni Katsoni and Takis Lebessis for their work at the Genomics Facility, our effortless collaboration and the countless NGS experiments they prepared for us over the years.

Special thanks go to the Panayotou Lab and the Fleming Proteomics Facility. I am particularly grateful to **George Panayotou**, **Martina Samiotaki**, **Gina Rouni**, and **George Stamatakis** for performing the Proteomics experiments and patiently training me in the analysis of these datasets. Moreover, I am deeply thankful to **Panagiotis Handris** for sharing his knowledge in mitochondrial (and not only) biology.

Last but never least, I would like to thank my **family** and **dearest friends** for constantly supporting and encouraging me to follow my dreams.

Abbreviations

AMPK – AMP-activated protein kinase
APC – Adenomatous Polyposis Coli
ASCL2 – Achaete-Scute Complex-Like 2
ASNSD1 – Asparagine Synthetase Domain Containing 1
ASOs – Antisense Oligonucleotides
ATF3 – Activating Transcription Factor 3
β-catenin – Beta-catenin
BMP – Bone Morphogenetic Protein
BRAWNIN – A 71aa microprotein encoded by C12orf73
CDK6 – Cyclin-dependent Kinase 6
CEBPA – CCAAT/Enhancer Binding Protein Alpha
ChIP-seq – Chromatin Immunoprecipitation Sequencing
CLCC1 – Chloride Channel Accessory 1
CRC – Colorectal Cancer
CRD – Cysteine-Rich Domain
CRISPR – Clustered Regularly Interspaced Short Palindromic Repeat
CRISPRa – CRISPR Activation
CRISPRi – CRISPR Interference
CAGE – Cap Analysis of Gene Expression
CYREN – A microprotein encoded by C7orf49
DACOR1 – DNMT1-Associated Colon Oncogenic RNA
DC – Destruction Complex
dCas9 – Dead Cas9
DNMT1 – DNA (Cytosine-5)-Methyltransferase 1
DKK1 – Dickkopf WNT Signaling Pathway Inhibitor 1
DKK – Dickkopf
eIF4G – Eukaryotic Translation Initiation Factor 4G
eRNA – Enhancer RNA
EMT – Epithelial-to-Mesenchymal Transition
Evi – Evenness Interrupted (alternate name for WLS)
EZH2 – Enhancer of Zeste Homolog 2
FANTOM – Functional Annotation of Mammalian Genomes
FAP – Familial Adenomatous Polyposis
FUS – Fused in Sarcoma
FZD – Frizzled
GAS5 – Growth Arrest Specific 5
GSK3 – Glycogen Synthase Kinase 3
HDAC1 – Histone Deacetylase 1
HDAC2 – Histone Deacetylase 2
HEK293T – Human Embryonic Kidney 293T cells
HIF-1α – Hypoxia-Inducible Factor 1-α
HLA – Human Leukocyte Antigen

HLA-I – Human Leukocyte Antigen Class I
HLA-II – Human Leukocyte Antigen Class II
HNF1A-AS1 – Hepatocyte Nuclear Factor 1A Antisense RNA 1
HGNC – HUGO Gene Nomenclature Committee
HOXB – HOX B
HOXA – Homeobox A
HOTTIP – HOXA Transcript at the Distal Tip
Igf2r – Insulin-like Growth Factor 2 Receptor
ISC – Intestinal Stem Cells
ISCs – Intestinal Stem Cells
IRES – Internal Ribosome Entry Sites
KRAB – Krüppel-Associated Box
Ku70/80 – Proteins involved in DNA repair
Lgr4/5 – Leucine-rich repeat-containing G-protein-coupled receptor 4/5
LED – Long Enhancer-associated RNA Dependent
LETN – LncRNA Essential for the Granular Component of the Nucleolus
lncRNA – Long Non-Coding RNA
linc-HOXA1 – Long Intergenic Noncoding RNA Associated with HOXA1
LncORF – Long Non-Coding RNA Open Reading Frame
lncPRESS1 – Long Noncoding RNA p53 Regulated by SIRT6
LncRNA-FAST – Long Non-Coding RNA FAST
lincRNA – Long Intergenic Noncoding RNA
lncRNA – Long Non-Coding RNA
LRP5/6 – LDL-Receptor-Related Protein 5 and 6
MALAT1 – Metastasis-Associated Lung Adenocarcinoma Transcript 1
MEG3 – Maternally Expressed Gene 3
miR-34a – MicroRNA-34a
miRNA – MicroRNA
MLN – Myoregulin
MYC – Myelocytomatosis Oncogene
MYU – MYC Upregulated LncRNA
NHEJ – Non-Homologous End Joining
NGS – Next-Generation Sequencing
NPM1 – Nucleophosmin
NORAD – Noncoding RNA Activated by DNA Damage
NXF1 – Nuclear Export Factor 1
PABPC1 – Poly(A) Binding Protein Cytoplasmic 1
PAF1 – PAF1 Complex Subunit 1
PCAT-1 – Prostate Cancer-Associated Transcript 1
PCP – Planar Cell Polarity
PRC1 – Polycomb Repressive Complex 1
PRC2 – Polycomb Repressive Complex 2
PFA1 – Prefoldin Complex Subunit 1
PRC2 – Polycomb Repressive Complex 2

Ribo-seq – Ribosome Profiling Sequencing
RMRP – RNA Component of Mitochondrial RNA Processing Endoribonuclease
ROSALIND – Regulatory Oxidative Stress-Activated LncRNA in Defense
RPF – Ribosome-Protected Fragment
RPS – Ribosomal Proteins
SIRT6 – Sirtuin 6
SIRT1 – Sirtuin 1
SEP – Small Encoded Protein
SNAIL – Snail Family Transcriptional Repressor 1
SLERT – Small LncRNA Enriched in the Dense Fibrillar Component of the Nucleolus
SPHK1 – Sphingosine Kinase 1
SPP1 – Secreted Phosphoprotein 1
SWI/SNF – SWItch/Sucrose Non-Fermentable
Tns1 – Tensin 1
TSS – Transcription Start Site
Tsix – Antisense LncRNA to Xist
Uph – Upperhand
uORF – Upstream Open Reading Frame
VDAC – Voltage-Dependent Anion Channel
WDR5 – WD Repeat Domain 5
WIF – Wnt Inhibitory Protein
Wls – Wntless
WNTRLINC1 – WNT-Regulated Long Non-Coding RNA 1
XCI – X Chromosome Inactivation
XIC – X-Inactivation Center
Xist – X-Inactive Specific Transcript
YAP/TAZ – Yes-associated Protein/Transcriptional Coactivator with PDZ-binding Motif
ZEB1 – Zinc Finger E-Box Binding Homeobox 1
ZEB2 – Zinc Finger E-Box Binding Homeobox 2

Table of contents

Acknowledgements	5
Abbreviations	7
Περίληψη	12
Abstract	14
1. Introduction	15
1.1 Long non-coding RNAs (LncRNAs)	15
1.1.1 Identification and characterization of lncRNAs	16
1.1.2 Classification of lncRNAs	18
1.1.3 Experimental approaches to study lncRNA Biology	19
1.1.4 Molecular mechanisms of cis-acting lncRNAs	20
1.1.5 Molecular mechanisms of trans-acting lncRNAs	23
1.1.6 The impact of lncRNAs in cancer	26
1.1.7 LINC01133	28
1.1.8 EPB41L4A-AS1	30
1.2 Intestinal homeostasis and carcinogenesis	31
1.2.1 The physiology of the intestinal epithelium	31
1.2.2 The Wnt/ β -catenin pathway	33
1.2.3 The Wnt/ β -catenin pathway in intestinal carcinogenesis	37
1.2.4 Functional interplay between the Wnt pathway and lncRNAs	39
1.3 Small open reading frames (smORFs)	41
1.3.1 Experimental discovery of novel SEPs	42
1.3.2 Classification of smORFs and SEPs	44
1.3.3 Roles of SEPs in cellular physiology and diseases	46
2. Aim of the thesis	51
3. Materials and Methods	52
Cell lines	52
Transient and stable transfection	52
Lentiviral transduction	53
CRISPR-Cas9 technologies for in cis overexpression and genome editing (KO)	53
RNA isolation followed by qPCR	54
Whole proteome extraction	55
Mitochondrial isolation	55
Western blot analysis	55
Immunofluorescence	56
Single molecule fluorescent in-situ hybridization (smFISH)	56
3' UTR mRNA Quant-seq	57
RNA sequencing analysis	57
Immunoprecipitation	58
LC-MS/MS Analysis (Proteomics)	58

Proteomics analysis	59
Correlation analysis between lnc-IGSF9 and IGSF9 expression	59
Ribosome profiling analysis	60
Flow cytometry analysis (FACs) for cell cycle analysis and TMRE staining	60
Luciferase assay	61
Colony formation assays (CFAs)	61
RNA pulldown	62
Xenographs in RAG1-/- immunocompromised mice	62
RNA pol II ChIP	63
Graphs and statistical analysis	63
4. Results	65
Part I: Lnc-IGSF9 is a novel Wnt-regulated lncRNA that promotes cell differentiation in CRC cells	65
Lnc-IGSF9 is a target of the β -catenin/TCF4 complex and is negatively regulated by the Wnt pathway.	65
Lnc-IGSF9 is a bona-fide nuclear lncRNA with no protein-coding potential	68
Lnc-IGSF9 exerts tumor suppressive action in CRC	70
Lnc-IGSF9 regulates IGSF9 expression to promote cell differentiation	75
CDX2 interacts with lnc-IGSF9 to regulate IGSF9 expression	79
Part II: LncRNA EPB41L4A-AS1 encodes a novel mitochondrial microprotein with tumor suppressive effects	83
LncRNA EPB41L4A-AS1 encodes a 13kDa mitochondrial microprotein	83
EPB-SEP overexpression reduces mitochondrial membrane potential	87
EPB-SEP exerts tumor suppressive effects	91
EPB-SEP interacts with several mitochondrial proteins	92
5. Discussion	95
Part I	95
Part II	100
Appendix I: Oligonucleotides that were used in Part I	104
Appendix II: Oligonucleotides that were used in Part II	106
Funding of the study	107
Bibliography	108

Περίληψη

Τα Μακρά Μη-Κωδικά RNA (lncRNAs) είναι μετάγραφα μήκους άνω των 200 νουκλεοτιδίων που δεν μεταφράζονται σε πρωτεΐνη. Παρουσιάζουν κοινά χαρακτηριστικά με τα αγγελιοφόρα RNA (mRNA), όπως το μάτισμα, τη πολυαδενυλίωση του 3' άκρου και την προσθήκη 5' καλύπτρας. Παρόλα αυτά, σε πολλές περιπτώσεις δεν είναι εξελικτικά συντηρημένα και εμφανίζουν ιστοειδική έκφραση. Τα lncRNAs έχουν φανεί να εμπλέκονται στη ρύθμιση πολλών βιολογικών φαινομένων και σηματοδοτικών μονοπατιών, ενώ η απορρύθμιση τους έχει συνδεθεί με νοσήματα, συμπεριλαμβανομένου του καρκίνου.

Το σηματοδοτικό μονοπάτι Wnt/β-κατενίνη αποτελεί βασικό ρυθμιστή της ομοιόστασης των βλαστοκυττάρων του εντερικού επιθηλίου και της ικανότητας του να αναγεννάται, καθώς ενορχηστρώνει την έκφραση πολλών σχετικών γονιδίων-στόχων. Η συνεχής ενεργοποίηση του μονοπατιού, εξαιτίας μεταλλάξεων που σταθεροποιούν την β-κατενίνη, αποτελεί μια από τις κυρίαρχες αιτίες του καρκίνου του παχέος εντέρου (CRC). Έχει διαπιστωθεί πως πολλά lncRNAs ρυθμίζουν το μονοπάτι Wnt, επηρεάζοντας το μοριακό του αποτύπωμα και ρυθμίζοντας τελικά την καρκινογένεση.

Παρόλο που τα lncRNAs χαρακτηρίζονται ως μη κωδικά, πρόσφατες μελέτες ανέδειξαν την παρουσία μικρών πλαισίων ανάγνωσης (smORFs) σε γενετικούς τόπους ορισμένων γονιδίων-lncRNA (lncORFs) τα οποία μεταφράζονται σε μικροπρωτεΐνες. Αυτές οι μικροπρωτεΐνες αποτελούν μια νέα κλάση βιομορίων με σημαντικό ρόλο σε πολλαπλές βιολογικές και μοριακές λειτουργίες.

Στο πρώτο μέρος της παρούσας διατριβής, ταυτοποιήσαμε το *LINC01133* (*lnc-IGSF9*) ως ένα πυρηνικό lncRNA του οποίου ο γενετικός τόπος περιέχει θέσεις πρόσδεσης από το σύμπλοκο β-κατενίνη/TCF4 και η έκφραση του ρυθμίζεται αρνητικά από το μονοπάτι Wnt. Η υπερέκφραση του μέσω CRISPR-a φανέρωσε ότι το μετάγραφο lnc-IGSF9 μειώνει τον πολλαπλασιασμό και προάγει τη διαφοροποίηση των CRC, ενώ αναστέλλει την ενεργοποίηση του μονοπατιού Wnt. Ανακαλύψαμε ότι το lnc-IGSF9 δρα *in cis*, ρυθμίζοντας την έκφραση του γειτονικού γονιδίου *IGSF9*, το οποίο κωδικοποιεί μια πρωτεΐνη κυτταρικής προσκόλλησης. Τα δύο μετάγραφα συν-εκφράζονται σε πολλούς ιστούς και είναι κατεσταλμένα σε ασθενείς με CRC. Μηχανιστικά, πειράματα ανοσοκατακρήμνισης του RNA έδειξαν ότι το *lnc-IGSF9* αλληλεπιδρά με την πρωτεΐνη CDX2 η οποία δρα ως κυρίαρχος ρυθμιστής της εντερικής διαφοροποίησης. Η υπερέκφραση της CDX2 επάγει την έκφραση του *lnc-IGSF9* και αυτό με τη σειρά του, αυξάνει τα επίπεδα έκφρασης του γονιδίου *IGSF9*. **Τα δεδομένα μας υποδεικνύουν ότι ο άξονας CDX2-lnc-IGSF9-IGSF9 αποτελεί έναν σημαντικό ρυθμιστικό μηχανισμό της εντερικής διαφοροποίησης και της βιολογίας του CRC.**

Στο δεύτερο μέρος της διατριβής εξετάσαμε την ύπαρξη κωδικών lncORFs στα καρκινικά κύτταρα. Ειδικότερα, αξιοποιήσαμε πειράματα ριβοσωμικού προφίλ (Ribo-seq) από κύτταρα HCT116 στα οποία είχε χορηγηθεί Κυκλοεξιμίδιο ή Λακτιμιδομυκίνη. Με τη χρήση των ORF-RATER και Ribosome Release Score (RRS) ταυτοποιήσαμε πολλαπλά lncORFs με δυνητική μεταφραστική δραστηριότητα.

Ανακαλύψαμε ότι το lncRNA EPB41L4A-AS1 κωδικοποιεί μια νέα μιτοχονδριακή μικροπρωτεΐνη (EPB-SEP). Παρατηρήσαμε ότι η υπερέκφραση της μειώνει το μεμβρανικό δυναμικό των μιτοχονδρίων και τη κλωνογενετικότητα των καρκινικών κυττάρων, ενώ προάγει τη στάση του κυτταρικού κύκλου στο στάδιο G2. Αντίθετα, η καταστολή του EPB41L4A-AS1 μέσω

shRNA ενισχύει τον κυτταρικό πολλαπλασιασμό. Η υπερέκφραση της EPB-SEP οδηγεί στη καταστολή πολλών μιτοχονδριακών πρωτεϊνών, καθώς και την απορρύθμιση μιτωτικών παραγόντων. Πειράματα ανοσοκατακρήμνισης έδειξαν ότι η EPB-SEP αλληλεπιδρά με αρκετές μιτοχονδριακές πρωτεΐνες κρίσιμες για τη μιτοχονδριακή ομοιόσταση. **Συνολικά, τα αποτελέσματα αυτά αναδεικνύουν την EPB-SEP ως μια νέα μιτοχονδριακή μικροπρωτεΐνη που κωδικοποιείται από το EPB41L4A-AS1 και έχει ογκοκατασταλτική δράση.**

Abstract

Long non-coding RNAs (lncRNAs) are transcripts longer than 200 nucleotides, without protein-coding potential. They are frequently spliced, polyadenylated, 5'capped, yet they lack evolutionary conservation and exhibit tissue-specific expression. lncRNAs regulate diverse biological processes and signalling pathways, and their deregulation is often linked to diseases, such as cancer.

The Wnt/ β -catenin pathway is a key regulator of intestinal stem cell maintenance and epithelial self-renewal, with the β -catenin/TCF4 complex orchestrating the expression of critical target genes. Aberrant activity of the Wnt pathway, due to mutations that stabilize β -catenin expression, is the major driver of colorectal cancer (CRC). Several lncRNAs have been identified as modulators of Wnt signaling, affecting its transcriptional repertoire and carcinogenesis.

Even though lncRNAs are typically considered non-coding, emerging studies have revealed the presence of small open reading frames (smORFs) residing within lncRNA loci (lncORFs), expressing microproteins. These microproteins represent an underappreciated class of biomolecules with critical roles in fundamental physiological and molecular processes.

In the first part of the thesis, we identified LINC01133 (lnc-IGSF9) as a nuclear lncRNA whose locus is bound by the β -catenin/TCF4 complex and its expression is negatively regulated by the Wnt pathway activity in CRC cells. CRISPR-a-mediated overexpression revealed that lnc-IGSF9 promotes cell adhesion and differentiation, reduces cell proliferation, and perturbs Wnt initiation. We discovered that lnc-IGSF9 operates *in cis*, regulating the expression of its neighboring gene *IGSF9*, which encodes a cell adhesion protein. Both transcripts are co-expressed across different tissues, and are downregulated in CRC patients, suggesting that both *lnc-IGSF9* and *IGSF9* are commonly regulated. Mechanistically, RNA pull-down followed by mass spectrometry demonstrated that lnc-IGSF9 interacts with CDX2, an intestinal-specific orchestrator of differentiation. CDX2 overexpression induces the expression of lnc-IGSF9 and subsequently, increases the IGSF9 expression levels. **Our findings suggest that the CDX2-lnc-IGSF9-IGSF9 axis is a novel regulator of intestinal differentiation.**

In the second part, we examined the existence of lncORFs in cancer cells by leveraging Ribo-seq experiments from HCT116 cells treated with Cycloheximide or Lactimidomycin. ORF-RATER and Ribosome Release Score (RRS) analyses identified multiple lncORFs with coding potential. Among the seven manually curated candidates, only **EPB41L4A-AS1** lncRNA was validated to encode a mitochondrial microprotein (EPB-SEP). Functional assays demonstrated that EPB-SEP overexpression disrupts mitochondrial membrane potential, induces G2 cell-cycle arrest, and reduces clonogenicity, while the shRNA-mediated knockdown of EPB41L4A-AS1 enhances proliferation. Proteomic profiling revealed that EPB-SEP predominantly impairs the mitoproteome and dysregulates mitosis-associated proteins. Immunoprecipitation experiments revealed that EPB-SEP interacts with several mitochondrial proteins essential for mitochondrial homeostasis. **Together, these results uncover EPB-SEP as a novel tumor-suppressive mitochondrial microprotein encoded by EPB41L4A-AS1.**

1. Introduction

1.1 Long non-coding RNAs (LncRNAs)

The completion of the Human Genome Project marked a groundbreaking effort to comprehensively characterize the human genome, revealing the complexity of our genetic material^{1,2}. The vast amounts of generated data helped us discover novel molecular mechanisms, enhance our understanding of gene regulation, and decipher the genetic and molecular causes of multiple diseases. Nevertheless, scientists encountered a surprising revelation: only about 2-3% of the genome encodes proteins, corresponding to approximately 20,000 genes, while the remaining majority is transcribed but not translated into proteins³⁻⁵.

Initially, researchers focused on protein-coding genes, regarding the non-coding part of the genome as transcriptional noise. This theory was further reinforced from the fact that non-coding transcripts generally exhibit lower conservation and expression levels, both key criteria for evaluating translational potential. On the other hand, they observed that humans share a similar number of protein-coding genes with less complex organisms, such as mice and nematodes, yet contain many more non-coding genes^{3,4,6,7}.

Indeed, early advancements in next-generation sequencing (NGS) and transcriptomics revealed that thousands of these transcripts exhibit critical functions for cellular physiology, orchestrating fundamental biological processes^{5,8}. Importantly, many of these transcripts are also implicated in human diseases such as cancer. The first non-coding RNAs (ncRNAs) that were studied were transfer RNAs (tRNAs) and ribosomal RNAs (rRNAs), both essential components of the translational machinery⁸. Afterwards, small-nucleolar RNAs (snRNAs) were discovered as vital players in the spliceosome⁸. Then, the development of NGS technologies led to the discovery of a much greater number of functional ncRNAs.

NcRNAs can be broadly classified into two groups based on their size: microRNAs (miRNAs), which are shorter than 200 nucleotides, and long non-coding RNAs (lncRNAs), which are longer than 200 nucleotides. miRNAs have been shown to target the 3' untranslated region (3' UTR) of mRNAs, leading to their degradation and suppression. They can also target other mRNA regions, such as the promoter or the 5' untranslated region (5' UTR)⁹ and they can control the translational rate of proteins¹⁰. miRNAs are vital orchestrators of multiple pathways and even serve as diagnostic and therapeutic biomarkers¹¹.

LncRNAs are transcripts longer than 200 nucleotides and do not encode proteins. Due to their diverse functions and varied localization within the cell, lncRNAs are considered a heterogeneous group of transcripts. This category includes intergenic transcripts, sense and antisense transcripts that overlap protein-coding genes, as well as transcripts arising from regulatory regions such as enhancer RNAs (eRNAs) and promoter upstream transcripts⁷. According to the HUGO Gene Nomenclature Committee (HGNC) and GENCODE, over 16,000 genes in the human genome are predicted to produce lncRNAs¹², while other studies suggest the existence of up to 100,000 distinct lncRNAs^{13,14}. Still, it remains unclear and challenging to

define which of these transcripts are functionally relevant or may be mere byproducts of transcription.

Recently, a consortium tried to address several issues regarding the vagueness of our current classification system of non-coding RNAs. Briefly, they proposed a new framework that divides the non-coding genome into 3 classes based on their size and the transcriptional process they undergo: i) small RNAs less than 50 nucleotides in length, ii) lncRNAs larger than 500 nucleotides, transcribed by RNA polymerase II (RNA pol II), and iii) a broader class encompassing Pol III transcripts (such as tRNAs or 5S RNAs), Pol II small RNAs (such as snRNAs, ~50-500 nucleotides), and Pol V transcripts in plants⁶.

1.1.1 Identification and characterization of lncRNAs

The first efforts that led to the massive identification of lncRNA genes in the mouse genome were performed by the FANTOM consortium using Cap Analysis of Gene Expression (CAGE) in full-length cDNAs⁴. CAGE enabled researchers to map the transcription start sites (TSS), estimate the transcript abundance, and gain deeper insights of global gene expression. This study also discovered that overlapping transcription can occur in both strands, which is an important feature of lncRNA biology as will be further discussed⁴. The development of NGS technology was critical for the next discoveries regarding the transcriptome. Researchers used chromatin immunoprecipitation followed by sequencing (ChIP-seq) to map genome-wide chromatin states to indirectly investigate non-coding gene expression. Specifically, they focused on histone H3 lysine 4 trimethylation (H3K4m3) and H3 lysine 36 trimethylation (K4-K36 domains) which mark active promoters and active transcriptional units accordingly¹⁵. At the same time, a separate study discovered novel lncRNAs based on their interaction to chromatin modifying complexes¹⁶. Both studies were among the first to describe the importance of the non-coding genome and the existence of lncRNAs and were further reinforced by RNA-sequencing (RNA-seq) experiments that comprehensively characterized the mammalian transcriptomes¹⁷. Next, the lack of protein-coding potential for lncRNA genes was validated based on their lower conservation scores through the use of PhyloCSF, a comparative genomics tool¹⁸. By integrating all these findings and applications, the GENCODE consortium generated the first comprehensive catalog of lncRNAs in the human genome¹⁹. This catalog continues to be enriched and updated as more RNA-seq and single-cell sequencing experiments are performed across a growing number of biological systems.

One additional aspect that was critical for annotating lncRNAs was protein-coding capacity. Since most lncRNA genes contain multiple open reading frames (ORFs) and typically include at least one 100 bp ORF by chance, several criteria were adopted to distinguish protein-coding capacity²⁰. These criteria were based on the following features: i) coding regions tend to be longer than expected by chance, ii) functional ORFs do not exhibit random codon usage, iii) higher substitution rates in silent codons are seen in protein-coding genes, iv) protein coding genes often include predicted protein domains, and v) coding genes display sequence similarities in protein databases²¹. These criteria were either used separately (for instance through the use of PhyloCSF¹⁸) or in combination to try to predict whether the ORFs residing in lncRNAs (lncORFs) might produce proteins.

Although thousands of lncRNAs have been annotated, distinguishing functional lncRNAs from transcriptional noise is challenging and requires additional experimental validation. Given that most lncRNAs are characterized by low expression levels and minimal sequence conservation scores, computational approaches that have been designed for characterizing protein-coding genes often prove adequate.

Besides several exceptions such as metastasis-associated lung adenocarcinoma transcript 1 (MALAT1) and nuclear paraspeckle assembly transcript 1 (NEAT1), both of which are expressed in high levels across all tissues, most lncRNAs exhibit tissue- or cell-specific expression²². Generally, lncRNA genes are lowly expressed due to the repressive histone marks decorating their promoters^{23,24}. In some cases, lncRNAs are only expressed during specific developmental²⁵ or differentiation²⁶ stages probably due to their precise temporal effect on cellular physiology. The restricted expression patterns and the generally low expression levels result in an underrepresentation of the lncRNA numbers in bulk RNA-sequencing approaches⁶. While single cell technologies capture cell-specific lncRNAs, they present several technical challenges for massively cataloging lncRNAs⁶. Moreover, low expression cannot be associated directly with non-functionality, as it may be a critical feature of the transcript. For instance, the low abundance of X-Inactive specific transcript (Xist), a lncRNA that orchestrates chromosome X inactivation, ensures that it will repress only the one chromosome X without affecting other chromosomes²⁷.

Most lncRNAs exhibit lower conservation scores compared to protein-coding mRNAs. However, low conservation does not necessarily imply lack of function^{28,29}. Only few lncRNAs, such as MALAT1 and Xist, are exceptions with high conservation across mammals. Indeed, traditional sequence comparison between different species often reveals low similarity among lncRNAs. Interestingly, the promoters of lncRNA loci share similar conservation scores with those of protein-coding genes. Additionally, several lncRNA genes show conserved exons, splice junctions and sequence patches^{4,19}. Higher conservation is also observed in structural modules, such as loops, spacers, or bulges, which are crucial for lncRNA function. For instance, Xist consists of two conserved functional modules^{4,19}, while the telomerase RNA template component (TERC) shows high sequence variability except a highly conserved core³⁰. Conclusively, loci that transcribe lncRNAs may contain highly conserved regions, particularly in regulatory elements (promoters, enhancers, splice junctions) or in regions that influence the secondary structure of the lncRNA.

Despite their low expression and conservation, lncRNAs constitute an important class of biomolecules that are implicated in most of the biological processes. lncRNAs share similar characteristics with mRNAs, as they are 5' capped, polyadenylated, contain multiple exons and undergo splicing^{6,7,21,31,32}. While lncRNAs are transcribed by RNA pol II, their post-transcriptional processing, export, and turnover differ from those of mRNAs⁷. Another distinctive feature is that many lncRNAs are retained in the nucleus, suggesting either a potential nuclear function or inefficient transcription and turnover³³.

Interestingly, a recent study demonstrated that lncRNAs are transcribed by RNA pol II with different CTD phosphorylation profiles than mRNAs, leading to inefficient transcription. These lncRNAs remain associated with chromatin and are subsequently targeted by the RNA exosome³³. Moreover, some nuclear lncRNAs contain U1 small nuclear RNA binding sites which serve as scaffolds for the recruitment of U1 small nuclear ribonucleoprotein (U1 snRNP) and are

tethered to chromatin³⁴. Additionally, lncRNAs are often inefficiently spliced, with longer sequences between the 3' splice site and the branch point, as well as shorter polypyrimidine tracts compared to mRNAs³⁵. Many functionally active lncRNAs also contain sequence motifs, such as repeat elements, that facilitate the recruitment of nuclear factors promoting their nuclear localization⁷.

On the other hand, lncRNAs that are not retained in the nucleus are exported to the cytoplasm through conventional mRNA export pathways. Both gene architecture and RNA sequence composition play a crucial role in determining nuclear export. Notably, NXF1, a key export factor, preferentially targets transcripts with high A/U content or those containing few exons, both features commonly found in lncRNAs³⁶. Once exported, lncRNAs are distributed throughout the cytoplasm and other organelles, with the majority localizing in polysomes (will be further discussed in next chapters)^{37,38}. Finally, transcriptomic studies have unveiled that lncRNAs can also be found in mitochondria³⁹.

1.1.2 Classification of lncRNAs

Since lncRNAs are loosely defined as a group of transcripts, several classification systems have been adopted. The two major classifications are based either on the genomic location of the lncRNA genes or on the effect they exert on DNA sequences³².

lncRNAs can be classified into different categories based on the position of their host gene relative to protein-coding genes. The major category consists of lncRNAs that are transcribed from regions that do not overlap with protein-coding genes (Figure 1A), known as long intergenic non-coding RNAs (lincRNAs). LincRNAs are generally shorter than protein-coding genes, containing fewer but longer exons, and are characterized by similar histone marks, splicing, and polyadenylation signals^{19,21,40}. lncRNAs transcribed from loci within the introns of protein-coding genes are classified as intronic lncRNAs (Figure 1B). Their unique characteristic is that they tend to act as circular RNAs derived from the excision of a spliced intron lariat^{41,42}. Few studies have shown that these intronic lncRNAs affect the transcription of their host protein-coding gene by interacting with the elongating RNA pol II machinery^{41,42}.

Sense lncRNAs are transcribed from the sense strand of protein-coding genes, overlapping a portion or the entirety of the protein-coding gene (Figure 1B). Conversely, antisense lncRNAs are transcribed from the opposite strand of protein-coding genes (Figure 1C). Bidirectional lncRNAs share the same promoter with a protein-coding gene, but are transcribed from the opposite strand (Figure 1D)⁴³. Sense lncRNAs are understudied likely due to the challenge of distinguishing their effects from those of their overlapping protein-coding genes. However, some sense lncRNAs may function both as a RNA and a protein-coding gene. Interestingly, steroid receptor RNA activator (SRA) is an example of sense lncRNA that encodes a protein and also functions as a scaffold for regulatory proteins involved in gene expression⁴⁴. Antisense lncRNAs often affect the half-life or stability of sense mRNA by directly binding to them, recruiting proteins, or through sequestration. Alternatively, they can modify the epigenetic state of the sense gene's promoter⁴⁵.

lncRNAs can be broadly classified into two categories based on their mode of action; those that act *in cis* and those that perform a series of different processes *in trans*. Cis-acting lncRNAs primarily influence local gene expression by modulating chromatin architecture or the

transcriptional activity of neighboring genes. In contrast, trans-acting lncRNAs can be found either in the nucleus or the cytoplasm, where they play crucial roles in various cellular processes^{6,31,32}.

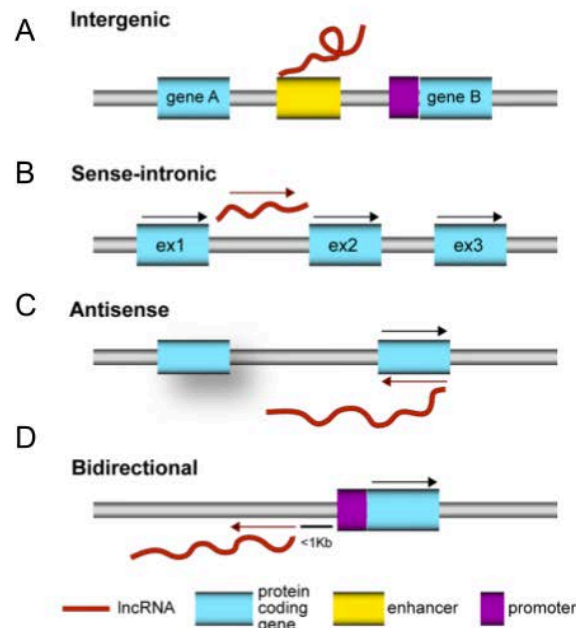


Figure 1: LncRNA classification based on their genomic location.

A. LincRNAs are transcribed from intergenic regions between protein-coding genes, with several also acting as transcriptionally active enhancers, B. Sense-intronic lncRNA are transcribed from the sense strand of a protein-coding gene, typically located within introns, C. Antisense lncRNAs are transcribed from the opposite strand of protein-coding genes, D. Bidirectional lncRNAs are transcribed from the opposite strand of a promoter that is shared with a protein-coding gene. Image adapted by Lanzafame et.al.⁴⁶

1.1.3 Experimental approaches to study lncRNA Biology

Understanding the mechanisms of function of lncRNAs is complex. For instance, despite nearly 30 years of investigation of Xist, its characterization remains incompletely understood³¹. To explore lncRNA biology, scientists have developed various experimental approaches. Firstly, it is critical to determine the localization of the lncRNA using techniques such as subcellular fractionation or fluorescent in situ hybridization (FISH). Next, functional studies typically involve loss-of-function or overexpression experiments to assess the lncRNA-mediated effects on the phenotype of cell lines or animal models⁶.

The development of Clustered Regularly Interspersed Short Palindromic Repeat (CRISPR) technologies has enabled scientists to manipulate lncRNAs precisely^{6,47}. The CRISPR/Cas9 system, comprising the Cas9 protein that introduces double-strand breaks in the DNA, and single guide RNAs (sgRNAs) that precisely target a genomic region, has become a

powerful genome editing tool. Advances in sgRNA design algorithms and the discovery of alternative Cas proteins have expanded the CRISPR applications in gene knock-out, knock-in, or transcriptional regulation.

In the case of nuclear lncRNAs, ectopic overexpression is often functionally irrelevant due to their context-dependent action. Instead, CRISPR activation (CRISPRa) systems utilize a dead Cas9 (dCas9) fused with transcriptional activators such as VP64 and p65, inducing the transcriptional activity of the endogenous lncRNA gene target⁴⁸. Compared to protein-coding genes, frameshift and INDEL mutations do not alter lncRNA function. Therefore, CRISPR interference (CRISPRi) systems were developed using a dCas9 fused with transcriptional repressors such as KRAB to block RNA pol II activity and inhibit transcription⁴⁹. Gene repression is also achieved through RNAi or antisense oligonucleotides (ASOs), but both technologies have been associated with multiple off-target effects. The CRISPR/Cas13 system that consists of a Cas13 RNase that degrades the RNA targets is similarly efficient with fewer off-target effects⁵⁰. Also, CRISPR-Cas9 mediated knock-outs (KO) are commonly utilized to completely excise the lncRNA gene or its regulatory regions (promoter or enhancer) and examine its function.

Although several cis-acting lncRNAs have been identified, distinguishing whether the transcript itself or the activity of the locus (and its corresponding regulatory elements) are responsible for observed phenotypes remains challenging. While CRISPRa and CRISPRi technologies do not permanently alter the genome, offering fewer off-target effects, the recruitment of transcriptional regulators might affect the expression levels of neighboring genes, especially in cases where the promoter resides inside another protein-coding gene. Deleting the exons of a lncRNA gene while preserving its promoter intact or inducing early transcription termination by inserting polyA sites are commonly used methods to separate the effects of lncRNA expression from sequence-dependent regulatory activity^{6,47}. Alternatively, fusing the sgRNA to the lncRNA of interest in order to guide the CRISPR/Cas9-lncRNA complex to the neighboring target gene can help determine whether the transcript is indeed regulating its expression^{51,52}.

On the other hand, more conventional methodologies are required to study trans-acting lncRNAs. The primary approaches involve loss-of-function and ectopic restoration in knock-out cells to characterize their function. Techniques such as RNA-pulldown and chromatin isolation by RNA purification (ChIRP) facilitate the identification of interactions between the lncRNA and the proteome or chromatin, respectively. Combining these approaches with information about the lncRNA's expression levels and localization can help scientists to thoroughly investigate its mechanism of function³¹.

1.1.4 Molecular mechanisms of cis-acting lncRNAs

Several mechanisms have been proposed through which cis-acting lncRNAs regulate local gene expression^{6,31}. In summary, either lncRNA transcripts or their act of transcription can affect the expression of neighboring genes, or the lncRNA locus may function as a regulatory element that orchestrates nearby gene expression in a lncRNA-independent manner^{6,31}.

Nuclear lncRNAs can affect chromatin through interacting with chromatin-modifying protein complexes¹⁶. Among these, Polycomb repressive complexes 1 & 2 (PRC1 & PRC2) are

two well-established epigenetic factors altering histone modifications to ultimately orchestrate gene expression. Specifically, PRC1 acts as a histone ubiquitin ligase for H2A Lys119, while PRC2 functions as a histone methyltransferase that is responsible for dimethylation or trimethylation of H3 Lys27 (H3K27)⁵³. Interestingly, both these factors lack DNA-binding sequences in mammals but instead interact with a plethora of RNAs^{54–56}. Notably, PRC2 interacts with thousands of lncRNAs in human¹⁶ and mouse cells⁵⁷. While these interactions have been characterized as non-specific, the depletion of selected interacting RNAs results in the derepression of PRC2-target genes¹⁶. In most cases, abundant lncRNAs, such as Xist, MEG3, and linc-HOXA1 recruit PRC2 to target loci and silence them through the deposition of the H3K27me3 repression mark (Figure 2A)⁵⁸.

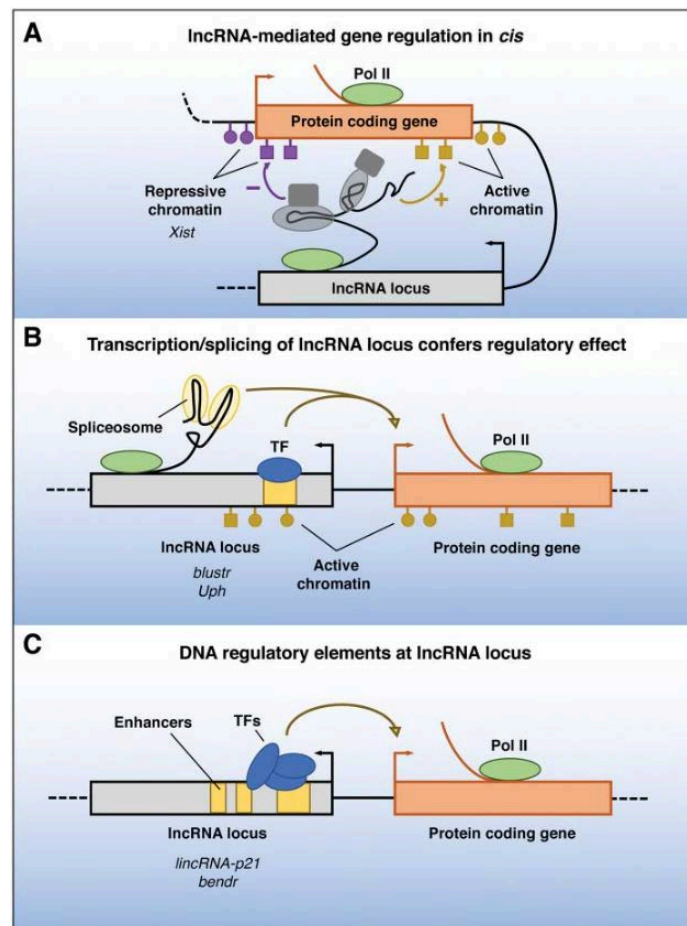


Figure 2: Mechanisms of function of cis-acting lncRNAs.

A. lncRNAs can regulate the expression of neighboring genes by recruiting regulatory proteins that modify local chromatin architecture and transcription, B. The act of transcription or splicing of a lncRNA regulates the expression of a neighboring gene in a lncRNA-independent manner, C. DNA elements within lncRNA loci can function as enhancers that regulate the expression of an adjacent protein-coding gene. Adapted by Koop & Mendell, 2018³¹.

Xist is a cis-acting lncRNA responsible for X chromosome inactivation (XCI), a process necessary for dosage compensation in female mammals. Xist is the first lncRNA thoroughly investigated, is ~17000 bp long, and consists of six conserved repeats (named A-to-F repeats). The Xist gene is located within the X-inactivation center (XIC) on the X chromosome, which also houses Tsix, its antisense lncRNA. When the chromosome X pairing occurs in the embryo both Xist and Tsix are upregulated, initiating a cascade of events leading to XCI. Briefly, Xist binds the PRC2 complex through its repeat A region (repA). Then, it binds and spreads across one of the two chromosomes X, facilitating the PRC2 recruitment and H3K27me3 deposition, ultimately rendering it inactive throughout the life of a female cell⁵⁹. This process requires the interaction between Xist and multiple protein factors essential for XCI initiation and Xist-mediated silencing⁶⁰. In contrast, Tsix is an antagonist of Xist, acting as a negative regulator of Xist expression, promoting the maintenance of the active chromosome X⁶¹.

LncRNAs also interact with other regulatory complexes, orchestrating epigenetic modifications on chromatin. For example, DNA methyltransferase DNMT1 is crucial for maintaining methylation patterns which are responsible for gene repression. DACOR1 lncRNA was found to interact with DNMT1, thereby modulating global gene expression in colon tumorigenesis through DNA methylation⁶². Similarly, a lncRNA transcribed by the *CEBPA* locus (named ecCEBPA) interacts with DNMT1, preventing the methylation of the host gene and ultimately increasing its expression levels⁶³.

Another notable example is HOTTIP, a lincRNA that orchestrates homeotic gene expression⁶⁴. Transcribed from the 5' tip of the HOXA locus that encompasses several HOXA genes that encode homeodomain transcription factors crucial for regional patterning during organismal development, HOTTIP binds to WDR5 and mediates chromatin looping to reach its target genes⁶⁵. It then recruits WDR5/MLL complexes to HOXA genes, promoting H3K4me3 modifications and maintaining active chromatin⁶⁴. In contrast, lncRNA PACER binds to the repressor p50, preventing it from binding to the COX2 promoter, ultimately inducing its expression⁶⁶. Additionally, the KHPS1 lncRNA facilitates the binding of chromatin-modifying complexes to its antisense gene *SPHK1*, through a RNA-DNA-DNA triplex formation^{67,68}. Several studies have demonstrated that lncRNAs can form triplexes that ultimately affect gene expression^{69,70}. Briefly, KHPS1 forms a triple helix upstream of the *SPHK1* enhancer, recruiting chromatin-modifying complexes that induce its transcription. In turn, the produced eRNA further induces *SPHK1* expression.

In some cases, the lncRNA itself is not functional but the act of transcription can affect gene expression (Figure 2B). Upperhand (Uph) is a lncRNA transcribed by a bidirectional promoter shared with the *Hand2* gene. Specifically, premature termination of the *Uph* transcription impaired *Hand2* transcription, while knock-down of the lncRNA transcript did not affect *Hand2* expression levels. Interestingly, premature termination reduced epigenetic modifications on the *Hand2* enhancer, subsequently preventing the binding of GATA4 transcription factor⁷¹. Similarly, *antisense Igfr2 RNA noncoding (Airn)* overlaps with the *Igf2r* gene and its transcription is responsible for imprinting the protein-coding gene in a lncRNA-independent manner⁷².

Even if a lncRNA locus exhibits transcriptional activity, it may contain DNA elements with independent cis-regulatory activity, functioning as enhancers (Figure 2C). Interestingly, lncRNA

TSSs are more frequently classified as enhancers compared to protein-coding genes, according to chromHMM⁷³. Some enhancer lncRNAs (elncRNAs) can serve as scaffolds for recruiting regulatory proteins or mediate chromatin looping^{74,75}. Also, these lncRNAs are often essential for enhancer activity⁶ with the increased transcriptional activity correlating with augmented enhancer activity⁷³. While the existence of these loci is well-documented, their characterization remains elusive.

1.1.5 Molecular mechanisms of trans-acting lncRNAs

Alternatively, many lncRNAs operate in distance from their transcription site. They can localize in the nucleus, cytoplasm, or within cellular organelles, where they i) regulate chromatin state and distal gene expression, ii) influence nuclear structure and organization, or iii) interact with and modulate RNA and protein complexes^{7,31}. The expression levels of trans-acting lncRNAs are usually more abundant as they interact with DNA, RNA or proteins in a stoichiometric manner^{6,31}.

Trans-acting lncRNAs can affect the chromatin state of multiple genomic loci similarly to cis-acting mechanisms. One prominent example is HOTAIR, which is transcribed from the HOXC locus and regulates the expression of the distant *HOXD* locus. Specifically, HOTAIR binds to a site proximal to the *HOXD* cluster, maintaining the repressive chromatin mark H3K27me3^{76,77}. HOTAIR's function remains elusive, as some studies suggest that HOTAIR serves as a scaffold, recruiting the PRC2 to establish H3K27me3 modification^{76,77}, while others propose that it directly binds to chromatin, mediating PRC2-independent repression⁷⁸.

Similarly, Mammary Tumor Associated RNA 25 (MaTAR25), a nuclear lncRNA in mice, interacts with the PURB transcription factor to regulate the expression of its target gene *Tns1*, ultimately reorganizing the actin cytoskeleton in mammary cells⁷⁹. In contrast, lncPRESS1 is regulated by p53 and acts as a decoy by interacting with SIRT6, tethering it from chromatin, and preventing it from repressing pluripotency-related genes⁸⁰. These studies highlight the important implications of lncRNAs in recruiting or sequestering chromatin-modifying complexes and transcription factors in distant chromatin sites.

lncRNAs are also implicated in the organization and regulation of nuclear architecture. Metastasis-associated lung adenocarcinoma transcript 1 (MALAT1) is among the most abundant and well-studied transcripts in humans. This transcript, which is approximately 8.7 kb in length, and conserved across vertebrates, has been extensively studied^{7,31,81–83}. Initially, MALAT1 was characterized as a major component of nuclear speckles, a sub-nuclear compartment enriched in spliceosomal subunits, snRNPs, transcription factors, and proteins involved in pre-mRNA processing. It acts as a scaffold that facilitates the recruitment of nuclear speckles to gene targets through interactions with chromatin or nascent RNAs even though it does not participate directly in nuclear speckle formation⁸⁴. Additional studies propose that MALAT1 binds to PRC2 target genes, CTCF binding sites, and regulatory regions, thereby mediating long-range chromatin interactions and affecting the transcriptional activity of its targets⁸⁵. Beyond its role in splicing and epigenetic modifications, MALAT1 is also involved in competing endogenous RNAs⁸⁶ or as a miRNA sponge, sequestering them from their targets and modulating gene expression⁸⁷.

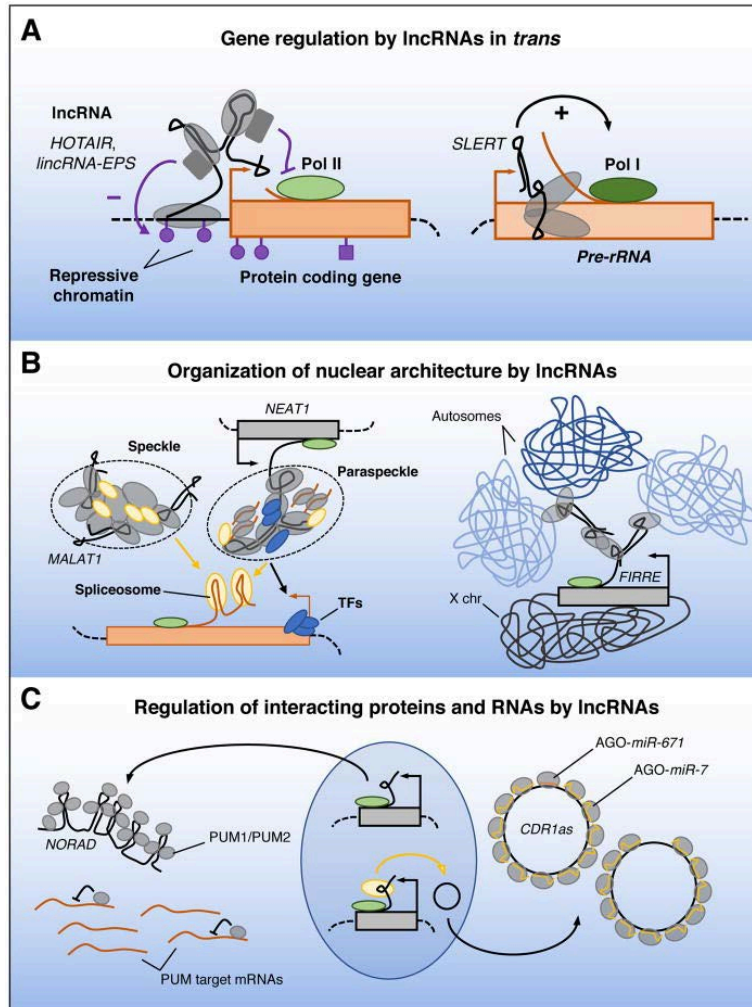


Figure 3: Mechanisms of function of trans-acting lncRNAs.

A. lncRNAs may act as scaffolds, binding to promoters or enhancers of distal genes and recruiting regulatory proteins that affect epigenetic modifications and transcriptional activity, B. Several lncRNAs are components of complex biological condensates, such as nuclear speckles and the paraspeckles, guiding them to target genes ultimately affecting splicing and gene expression, C. lncRNAs can interact with proteins and RNAs, modulating their activity. Adapted by Koop & Mendell, 2018³¹.

Nuclear enriched abundant transcript 1 (NEAT1) is another example of a lncRNA involved in nuclear organization. The *NEAT1* gene is located adjacent to the *MALAT1* gene and is evolutionary conserved. It generates two isoforms, one long and one short that share a common 5' UTR⁷. NEAT1 long is responsible for the formation and organization of the paraspeckles, a subnuclear membrane-less structure consisting of proteins related to transcriptional and post-transcriptional processes⁸⁸. Interestingly, NEAT1 long interacts with paraspeckle proteins NONO and SFPQ, through its middle domain, to ultimately assemble the phase-separated paraspeckles⁸⁹.

In addition to MALAT1 and NEAT1, two other architectural lncRNAs, SLERT and LETN, are implicated in the formation of nuclear condensates through phase separation^{90,91}. Specifically, SLERT orchestrates the biophysical properties of the dense fibrillar component of the nucleolus, where RNA pol I transcription takes place⁹⁰. Similarly, lncRNA LETN is essential for the formation of the granular component of the nucleolus by interacting with NPM1, a key nucleolar component⁹¹. All these findings underscore the critical involvement of lncRNAs in the formation of biological condensates that spatially organize the genome.

Cytoplasmic lncRNAs regulate various biological processes by physically interacting with proteins in a stoichiometric-dependent manner. For example, the noncoding RNA activated by DNA damage (NORAD) is an abundant cytoplasmic lncRNA that interacts with two RNA-binding proteins from the PUMILIO family, PUM1 and PUM2. NORAD sequesters these proteins, preventing them from binding to their target mRNAs^{92,93}. These proteins repress the stability of their targets through accelerating mRNA decay and translational blockage. By binding to PUM1 and PUM2, NORAD preserves the stability and expression of these mRNAs, playing a crucial role in maintaining cellular homeostasis^{92,93}. Intriguingly, NORAD also localizes in the nucleus where it interacts with components of the DNA-damage response, assembling a topoisomerase complex responsible for genome stability maintenance⁹⁴.

lncRNAs can also regulate miRNAs by binding to them through base complementarity, thereby modulating their abundance^{7,31}. These lncRNAs are termed competing endogenous RNAs (ceRNAs), although only a few examples have been reported. The lncRNA-PNUTS results from alternative splicing of the PNUTS pre-mRNA and contains seven binding sites for miR-205. By sequestering miR-205, lncRNA-PNUTS prevents it from targeting ZEB1 and ZEB2, two key regulators of epithelial cell maintenance⁹⁵.

Surprisingly, recent studies have revealed that several lncRNAs are also localized to subcellular organelles such as exosomes, ribosomes, or mitochondria. RNA-sequencing in extracellular vesicles has revealed the existence of several lncRNAs which are likely incorporated through interaction of lncRNAs with RBPs⁷. Exosomal lncRNAs can be secreted and facilitate communication between secreting and receiving cells. They are involved in epigenetic modifications, splicing, genomic stability, and tumor development⁹⁶. Additionally, weakly expressed lncRNAs that are transcribed by dysregulated RNA pol II are rapidly degraded by the RNA exosome³³.

Ribosomal lncRNAs remain largely understudied, primarily due to the technical challenges involved in their identification. Ribosome profiling (Ribo-seq), a widely used technique to assess the translational landscape, has shown that many lncRNAs, in addition to mRNAs, are associated with ribosomes^{37,57,97}. Despite the use of sophisticated algorithms to distinguish between bona fide protein-coding and non-coding genes, only a few ribosome-bound lncRNAs have been confirmed to produce proteins³⁷. For example, ROSALIND is a nuclear-encoded lncRNA that interacts with the mitochondrial ribosomes and protects the mitochondrial translational machinery upon oxidative stress⁹⁸. It is suggested that most ribosome-bound lncRNAs are either detected due to technical limitations of Ribo-seq or because they play regulatory roles in the translational machinery.

While the mitochondrial genome transcribes hundreds of non-coding RNAs, only a few lncRNAs localize in the mitochondria and can be transcribed either by nuclear or mitochondrial DNA⁹⁹. The three major mitochondria-encoded lncRNAs lncND5, lncND6, and lncCyt form

duplexes with their mRNA targets through base complementarity, ultimately cleaving them and controlling their expression levels¹⁰⁰. RMRP is an abundant nuclear-encoded lncRNA that localizes in the cytoplasm, nucleus, and mitochondria. It is the component of RNase MRP and is essential for mitochondrial DNA replication¹⁰¹. Accordingly, GAS5, another nuclear-encoded mitochondrial lncRNA, regulates the association of key metabolites, tuning energy homeostasis in response to nutrient stress¹⁰².

1.1.6 The impact of lncRNAs in cancer

Due to their multidimensional functions in regulating gene expression, many lncRNAs have been associated with carcinogenesis. Cancer arises from the genomic alterations that disrupt gene regulatory networks essential for cellular homeostasis. A seminal review classified the molecular mechanisms driving carcinogenesis into six key hallmarks: sustained proliferative signaling, evasion of growth suppressors, activation of invasion and metastasis, replicative immortality, induction of angiogenesis, and resistance to cell death¹⁰³. More recent discoveries have expanded these hallmarks to include reprogramming of energy metabolism, evasion from immune destruction, alterations in phenotypic plasticity, epigenetic reprogramming, polymorphic microbiomes, and cellular senescence^{103,104}. Interestingly, lncRNAs are involved in regulating or disrupting multiple of these hallmarks, ultimately leading to tumor development^{105–107}.

Dysregulation of lncRNA expression is a key indicator for their involvement in cancer. Moreover, lncRNA loci are often linked with cancerous genetic abnormalities, such as their presence in copy number variations, where amplifications or deletions of the lncRNAs directly contribute to tumorigenesis¹⁰⁷. Similarly, multiple cancer-related single nucleotide polymorphisms reside in regulatory regions, such as promoters or enhancers, affecting lncRNA expression¹⁰⁸. To study lncRNAs in cancer, scientists employ bioinformatic approaches, loss- or gain-of-function experiments, and generate animal models to recapitulate the disease physiology^{107,109}. Bioinformatic tools aim to predict and correlate lncRNA function with critical carcinogenic signaling pathways. Experimental approaches (as described in section 1.1.3) are essential to elucidate the involvement of a lncRNA to cancer cell phenotype. Mouse models are primarily used for studying lncRNAs, with knockout mice allowing the characterization of their in vivo role in cancer. Despite the convenience of using mouse models, many lncRNAs are not evolutionary conserved and cannot be easily studied in vivo.

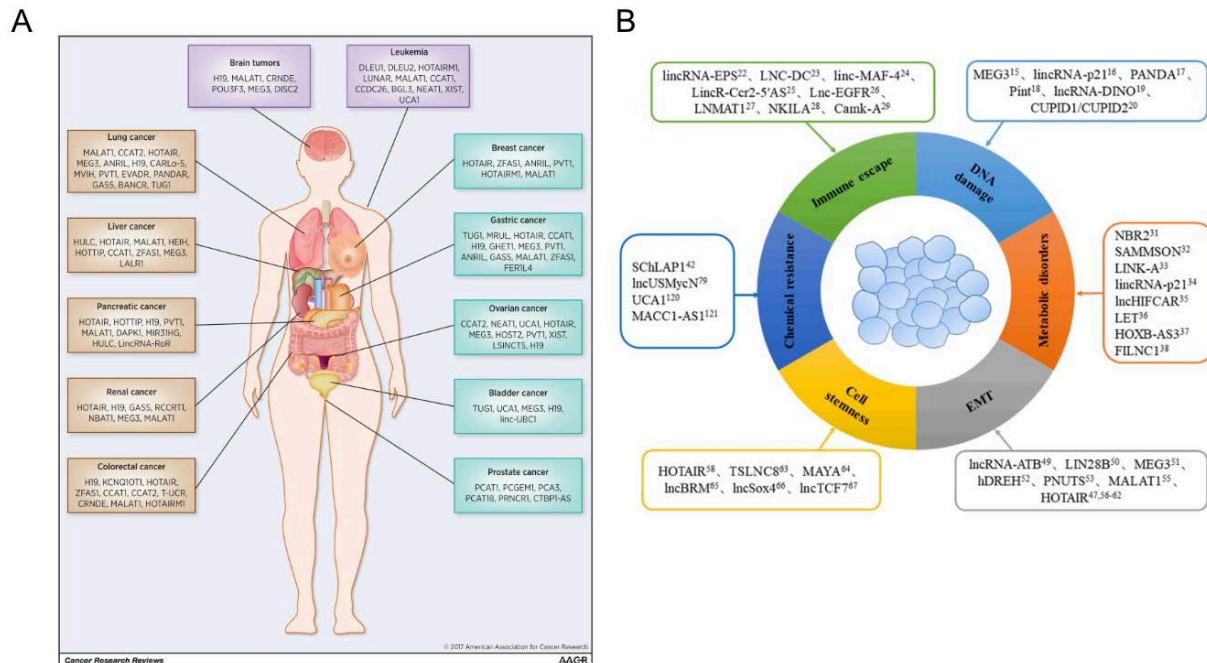


Figure 4: The involvement of lncRNAs in cancer.

A. Graphical illustration showing the lncRNAs that have been associated with different types of cancer¹¹⁰, **B.** Graph depicting the lncRNAs classified based on their involvement in different hallmarks of cancer¹¹¹.

lncRNAs exert oncogenic or cancer suppressive effects through their interplay with signalling cascades¹⁰⁵. For instance, p53, a transcription factor that orchestrates tumor suppressor in response to genotoxic stress, directly targets and activates multiple lncRNAs¹⁰⁷. For example, LINC-PINT enhances p53-dependent apoptosis through PRC2-mediated epigenetic silencing¹¹². LED is another p53-induced lncRNA that facilitates p53 function by interacting with and activating p53-related enhancers¹¹³. As expected, these lncRNAs are downregulated in tumors, highlighting their roles as putative tumor suppressors^{112,113}. Similarly, MEG3 activates the p53-dependent transcriptional program and is downregulated in various cancers, even though it is not regulated by p53 itself¹¹⁴.

Other lncRNAs, such as MALAT1, are implicated in numerous cancer types due to their diverse and critical functions. MALAT1 promotes proliferation in retinoblastoma, induces various forms of cell death in different cancers, enhances cancer metastasis through regulating epithelial-to-mesenchymal transition, orchestrates angiogenesis, and is associated with decreased tumor response and increased chemoresistance¹¹⁵. HOTAIR is a similar case, as its overexpression serves as a prognostic marker in multiple types of cancer¹¹⁰.

In contrast, tissue-specific lncRNAs can serve as highly sensitive biomarkers. For example, Prostate Cancer Antigen 3 (PCA3), a lncRNA regulated by the androgen receptor (AR), is overexpressed in nearly 95% of prostate cancer patients and affects cell survival by modulating AR-target gene expression¹¹⁶. It is currently used as a biomarker in the urine of prostate cancer patients, offering high specificity for disease prediction^{117,118}. Similarly, Prostate

Cancer-Associated ncRNA Transcript 1 (PCAT-1) is overexpressed specifically in metastatic prostate cancer. PCAT-1 represses the tumor suppressor BRCA2, increases sensitivity to PARP1 inhibitors, and promotes MYC expression, leading to enhanced cell proliferation¹¹⁰.

Due to their high specificity and critical role in carcinogenesis, lncRNAs serve as promising targets for cancer therapy. Current approaches include transcriptional or post-transcriptional inhibition through RNAi technology, antisense oligonucleotides (ASOs), and morpholinos, synthetic lncRNA introduction, and CRISPR-mediated editing of lncRNA loci¹¹⁹. However, these methods face several challenges, such as increased immunogenicity, undesirable off-target effects, and difficulties in delivering RNA to specific organs¹¹⁹. Despite their advantages, most clinical trials focus on using lncRNAs as diagnostic and monitoring tools, with few investigating them as a therapeutic agent¹²⁰. Advances in the miRNA-based therapies suggest that lncRNAs may also become putative therapeutic targets in the future.

1.1.7 LINC01133

A major part of this thesis will focus on the function of LINC01133 in CRC. *LINC01133* is located on chromosome 1q32.2 and generates four transcripts: ENST00000635112.1 (1996 bases, 3 exons), ENST00000657602.1 (1418 bases, 4 exons), ENST00000423943.2 (1405 bases, 3 exons), and ENST00000443364.6 (1266 bases, 4 exons). It is not conserved and does not encode a protein. It is expressed in a tissue-restricted manner and exhibits distinct subcellular localization patterns across cancer types.^{121,122}

Several studies have addressed the role of LINC01133, particularly in carcinogenesis, as it is dysregulated in different cancer types^{121,122}. It functions as either an oncogene or tumor suppressor, depending on the cancer type. Duplications of the *LINC01133* gene have been observed in 49 Chinese patients with hepatocellular cell carcinoma, resulting in upregulated LINC01133 levels¹²³. In cancers such as triple-negative breast cancer¹²⁴ and hepatocellular carcinoma¹²³, LINC01133 promotes the proliferation of cancer stem-like cells, and it is frequently upregulated in osteosarcoma and pancreatic cancer¹²¹. In contrast, it is downregulated in CRC, nasopharyngeal cancer, and oral esophageal squamous cell carcinoma, where it often exerts tumor-suppressive properties¹²². In colorectal cancer (CRC) particularly, LINC01133 was downregulated in 219 samples, while its knockdown promoted EMT and metastasis in CRC cells¹²⁵.

LINC01133 is implicated in diverse mechanisms depending on its subcellular localization. Several studies have revealed that LINC01133 functions as an endogenous competitor RNA or microRNA sponge. For instance, LINC01133 is abundant in the cytoplasm of gastric cells, where it interacts with miR-106-3p, leading to upregulation of APC. The LINC01133-mediated induction of APC levels inhibits Wnt/ β -catenin signaling activity and gastric cancer cell migration, suggesting that LINC01133 acts as a tumor suppressor in a Wnt-dependent manner¹²⁶. Similarly, LINC01133 sponges miR-199a-5p, blocking it from downregulating *SNAI1*, ultimately promoting epithelial-to-mesenchymal transition (EMT) in hepatocellular carcinoma cells¹²³. Multiple additional studies have focused on the role of LINC01133 on competing microRNAs, however their mechanisms remain elusive¹²¹.

Cytoplasmic LINC01133 also binds to proteins to modify their function. *Kong et al.* showed that LINC01133 is a TGF- β -regulated lncRNA that inhibits EMT in CRC by interacting

with the splicing factor SRSF6¹²⁵. In a different study, LINC01133 was found to interact with FUS to facilitate the stability of the mRNA FSP1, providing enhanced resistance to ferroptosis in pancreatic cancer cells¹²⁷. Additionally, LINC01133 interacts with Arp3 in the cytoplasm of pancreatic ductal adenocarcinoma cells to reduce the mRNA degradation of SPP1 and promote EMT¹²⁸.

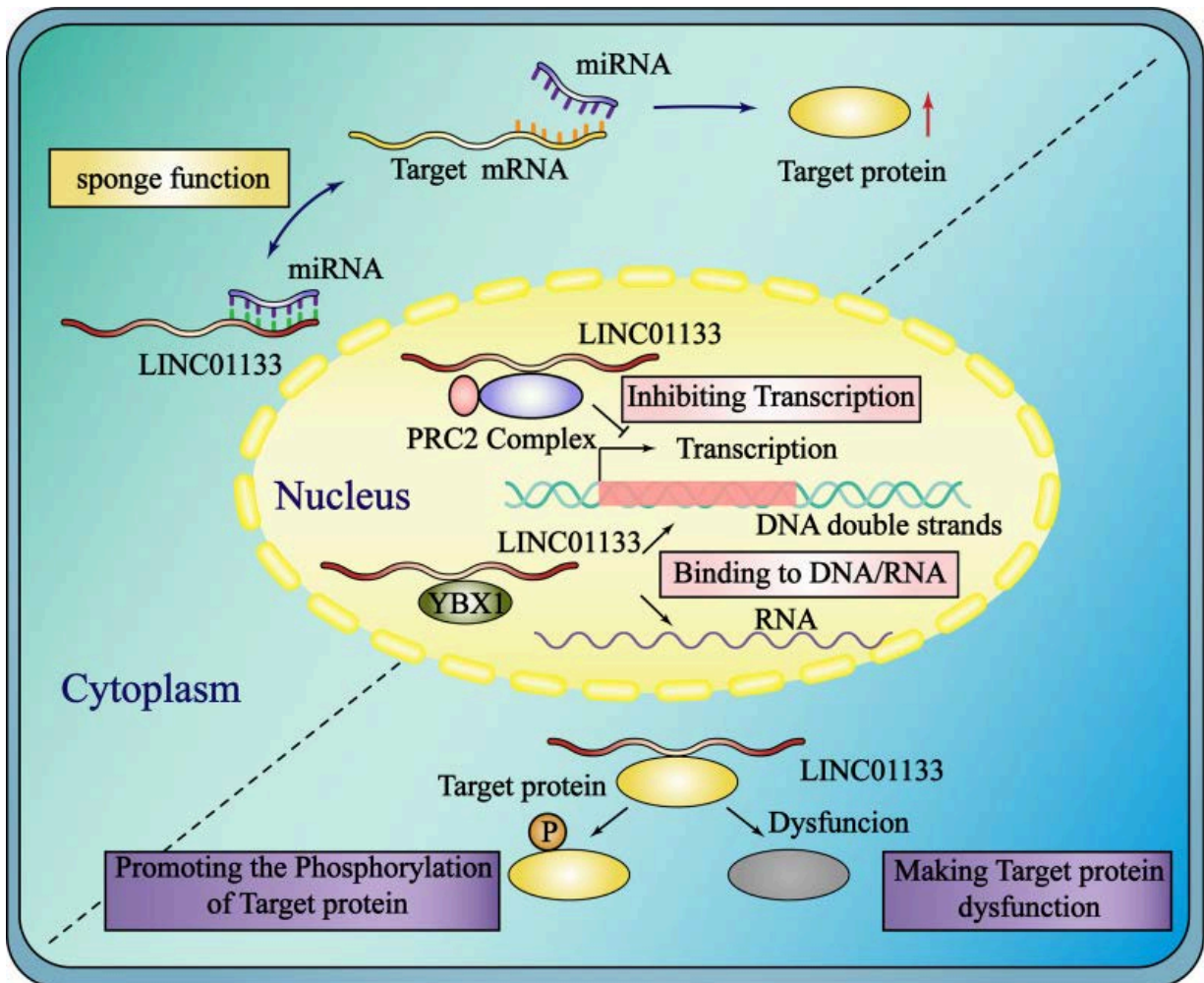


Figure 5: The multifaceted roles of LINC01133.

The functions of LINC01133 are determined by its subcellular localization and the host tissue context. In the cytoplasm, LINC01133 acts as a competing endogenous RNA (ceRNA), sequestering microRNAs (miRNAs) from their targets or binding to RNA-binding proteins (RBPs). In the nucleus, LINC01133 primarily interacts with the EZH2 subunit of the PRC2 complex to mediate transcriptional repression¹²¹.

Few studies have addressed the role of LINC01133 in the nucleus and corroborate that LINC01133 participates in epigenetic regulation through interactions with the EZH2 subunit of the PRC2 complex. Song *et. al.* showed that LINC01133 interacts with EZH2 to promote the H3K27me3 signal in the promoter of SOX4, inhibiting breast cancer invasion and metastasis¹²⁹. Additionally, LINC01133 interacts with EZH2 and recruits the PRC2 complex on the promoters of *KLF2*, *P21*, and *E-cadherin* to repress their expression in non-small cell lung cancer cells¹³⁰.

Another group showed that LINC01133 epigenetically methylates the promoter of DKK1, to promote pancreatic tumorigenesis¹³¹.

Beyond the subcellular localization of LINC01133, it has also been found enriched in the exosomes of pancreatic ductal adenocarcinoma cells. More specifically, a study reported that Periostin increases exosome secretion and upregulates LINC01133 expression through c-myc binding on the *LINC01133* promoter. Next, LINC01133 interacts with EZH2 to promote H3K27 trimethylation on the *AXIN2* gene to stabilize β -catenin and activate the Wnt/ β -catenin pathway¹³².

1.1.8 EPB41L4A-AS1

EPB41L4A-AS1 is a lncRNA located on chromosome 5 at the 5q22.2 region, spanning over 3.5 kb in length and consisting of three exons. Only a few studies have investigated EPB41L4A-AS1 function, focusing on its effect on metabolism and tumorigenesis. Notably, its genomic region is frequently deleted in human cancer, and its low expression in the cervix, liver, breast, bladder, and other cancers is associated with poor survival based on the Progenetix cohort¹³³.

Intriguingly, an early study demonstrated that the first exon of EPB41L4A-AS1 encodes TIGA1, a 120kDa small outer mitochondrial membrane protein. TIGA1 expression is induced after 3 days of starvation when cells enter the deep G0 phase, and its overexpression was found to inhibit proliferation¹³⁴. Its role has been addressed in only one other study, where they showed that TIGA1 interacts with VDAC, a key mitochondrial channel protein, ultimately inducing microtubule depolarization¹³³.

Three studies from the same group have examined the role of EPB41L4A-AS1 in cellular metabolism. Consistently, their findings reveal that EPB41L4A-AS1 perturbs the Warburg effect, a metabolic shift that allows cancer cells to produce more glucose by favoring glycolysis over the Krebs cycle. More specifically, they showed that siRNA-mediated knockdown of EPB41L4A-AS1 accelerates glycolysis and glutamine metabolism, while its overexpression has the opposite effects in HeLa and HepG2 cells. Mechanistically, nuclear EPB41L4A-AS1 interacts with HDAC2 to promote the H3K27 acetylation on *VHL* and *VDAC1* promoters, two critical regulators of glycolysis and glutaminolysis¹³³. Next, they discovered that trophoblast cells, derived from placental tissue in early recurrent miscarriage, are characterized by elevated EPB41L4A-AS1 and TIGA1 expression levels. Microarray-based transcriptomic profiling revealed that EPB41L4A-A1 represses glycolysis while activating oxidative metabolism. Particularly, the overexpression of EPB41L4A-AS1 epigenetically affected the protein levels of two crucial metabolic orchestrators by downregulating HIF-1 α while increasing VDAC1 expression in a TIGA1-independent manner¹³⁵. Accordingly, they demonstrated that EPB41L4A-AS1 is overexpressed in the liver of type 2 diabetes mellitus patients, inhibiting their glucose uptake¹³⁶. The inhibitory effects of EPB41L4A-AS1 in glycolysis were also observed in neuroblastoma patient-derived natural killer cells (NK). Neuroblastoma NK cells are classified into two subpopulations with distinct functions and responses and are characterized by impaired glycolysis upon cytokine stimulation. Interestingly, the one population has increased EPB41L4A-AS1 levels and encapsulates it into exosomes to transfer it to the second population to perturb glycolysis in neuroblastoma patients¹³⁷.

Beyond its role in metabolic rewiring, several studies have addressed the EPB41L4A-AS1 involvement in carcinogenesis. *Hegre et. al.* integrated RNA-seq and ChIP-seq experiments in synchronized HaCaT cells and distinguished EPB41L4A-AS1 as an M/G1 enriched lncRNA. Knockdown of EPB41L4A-AS1 resulted in diminished growth in HaCaT, A549, and DLD1 cell lines while causing a cell cycle shift to the G2/M phase¹³⁸. In another study, researchers analyzed a publicly available microarray dataset from a CRC cohort (GSE21510) comprising 19 CRC samples and 25 normal colorectal tissues and discovered that *EPB41L4A-AS1* is upregulated in CRC. This pattern was also validated through q-PCR in CRC biopsies and is augmented in later stages of CRC. Overexpression of EPB41L4A-AS1 resulted in increased proliferation and EMT, while its knockdown reduced cellular clonogenicity and EMT in CRC cell lines¹³⁹. On the other hand, EPB41L4A-AS1 was found to exhibit opposite effects in gastric cancer¹⁴⁰.

1.2 Intestinal homeostasis and carcinogenesis

The gastrointestinal (GI) system is a complex network that includes the esophagus, stomach, small intestine, colon, and rectum. These organs cooperate to transport and digest food and liquids facilitating nutrient assimilation and waste excretion. The intestinal epithelium is a single layer of specialized epithelial cells that lines the luminal surface of the small and large intestines. It is responsible for the exchange and absorption of nutrients and water which occurs mainly in the small intestine^{141,142}. Beyond its role in nutrient absorption, it also serves as a defense mechanism, protecting the body from harmful molecules and microorganisms. It functions as a mechanical barrier and modulates immune response through its interplay with the gut microbiota¹⁴³.

1.2.1 The physiology of the intestinal epithelium

The structural organization of the intestinal epithelium in the small intestine and the colon is designed to maximize surface area, optimizing the efficient absorption of nutrients and water. This is achieved through the finger-like projections, called villi, that protrude in the intestinal lumen. At the base of these villi are found the crypts which are mucosal invaginations that house intestinal stem cells (ISCs) and their progenitors.

Throughout digestion, the intestinal epithelium is subjected to various chemical, biological, and mechanical stresses. To maintain its integrity, it undergoes a renewal process every 4-5 days, shedding stressed cells. This remarkable self-renewal capacity makes the intestinal epithelium one of the most intensively replenished tissues in the body. The major drivers of intestinal homeostasis are the ISCs located in the crypt niche. They generate highly proliferative progenitors known as transit-amplifying cells, which undergo a series of 4-5 cell divisions before differentiating to intestinal cells while migrating upwards toward the villi. This complex process constantly supplies the villi with fully differentiated intestinal epithelial cells to achieve nutrient absorption^{141,142}. Three days after terminal differentiation, mature epithelial cells reach the tip of the villus, where they undergo spontaneous apoptosis (Figure 6)¹⁴⁴.

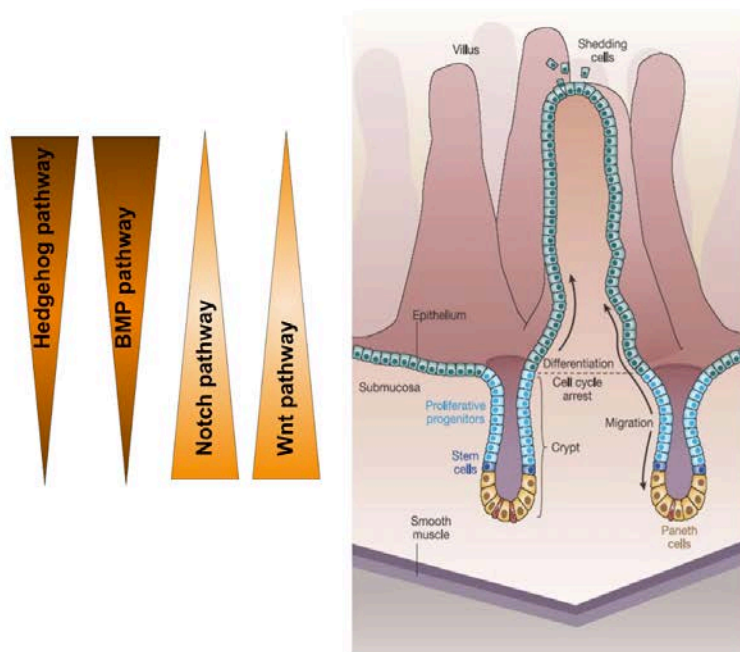


Figure 6: The cellular architecture of the intestinal epithelium.

The intestinal epithelium consists of protruding villi with crypts at their bases. ISCs (shown in blue) reside at the center of the crypts and give rise to highly proliferative transit-amplifying cells (shown in cyan). Paneth cells are positioned at the base of the crypts near the ISCs (shown in brown). Transit-amplifying cells differentiate as they migrate upwards toward the villi tips, where they undergo spontaneous apoptosis. This dynamic architecture is regulated by multiple signaling pathways: the Wnt pathway induces stemness and the Notch pathway keeps ISCs in undifferentiated state, while the BMP and Hedgehog pathway activate a differentiation program. Image adapted by Reya and Clevers 2005¹⁴⁵.

The intestinal epithelium comprises four distinct types of differentiated epithelial cells: enterocytes, goblet cells, Paneth cells, and enteroendocrine cells. Enterocytes are the primary cell type in the intestinal epithelium and are responsible for nutrient absorption and transport of water. Goblet cells secrete mucus, contributing to the formation of a mucus layer that safeguards the small intestine¹⁴⁶. Paneth cells are the only mature epithelial type that primarily resides in the crypts. They secrete antimicrobial peptides, such as defensins and lysozyme to boost immune protection in the gut, and also facilitate the differentiation of transit-amplifying cells in the crypts^{147,148}. Lastly, enteroendocrine cells are scattered throughout both crypts and villi, producing multiple gut hormones that regulate nutrient assimilation, intestinal function, and insulin secretion¹⁴⁹.

The Wnt pathway is the driving force behind ISC maintenance. Knockdown of Transcription Factor 4 (TCF4), a critical transcription factor in the Wnt pathway, leads to downregulation of target genes essential for proliferation, such as Myc¹⁵⁰. Conversely, aberrant Wnt pathway activation, through β -catenin stabilization or R-spondin induction, causes crypt hyperplasia and mis-localization of differentiated intestinal cells^{151,152}. Interestingly, while Paneth cells provide the necessary Wnt ligands to activate the Wnt pathway in ISCs, their ablation or

inhibition of Wnt production do not disrupt ISC homeostasis, suggesting the involvement of other neighboring cell populations contributing to Wnt signaling in the stem cell niche¹⁵³.

The complex architecture of the intestinal epithelium and the cell fate decision are governed by the fine-tuned interplay of multiple signaling cascades (Figure 6). The stem cell niche is orchestrated by Wnt and Notch signaling, while the villi are regulated by the BMP and Hedgehog pathways. Notch signaling is essential for keeping ISCs in an undifferentiated and proliferative state¹⁴⁴. Hyperactivation of the pathway, through the overexpression of a constitutively active Notch, disrupts the differentiation of secretory cells, leading to depletion of goblet cells and suppression of endocrine and Paneth cell differentiation^{154,155}. As cells migrate from the crypts to the villi, Wnt and Notch signaling diminish, and the BMP and Hedgehog pathways become more active. Transit-amplifying cells that are distant from the crypts are characterized by reduced Wnt activation and increased Hedgehog and BMP signaling activity initiating a differentiation program that restricts proliferation^{156,157}. Notably, Hedgehog signaling is necessary for BMP pathway activation, and its aberrant hyperactivation can lead to BMP-mediated depletion of progenitor cells¹⁵⁷. Both BMP and Hedgehog signaling are essential for suppressing the Wnt pathway, promoting differentiation¹⁴².

For many years, two primary models have explained the physiology of ISCs. The traditional model posits that ISCs reside in the +4 position relative to the crypt bottom and serve as a constant reserve for cellular regeneration. Notably, genetically damaged stem cells can be replaced by transit-amplifying progenitors which migrate back to the +4 position and reacquire stem cell properties¹⁵⁸. The model was further reinforced by another study that demonstrated that these cells are highly sensitive to radiation, suggesting that they have developed advanced DNA damage defense mechanisms¹⁵⁹.

While this theory dominated the field, research into Wnt-regulated target genes led to the discovery of critical molecular mechanisms in intestinal physiology. A pivotal breakthrough came with the identification of the stem cell marker Lgr5, which allowed the characterization and live monitoring of the stem cell niche within the crypts. *Barker et. al.* generated an eGFP knockin mice to label Lgr5+ cells, through lineage tracing experiments, and identified the rare crypt base columnar (CBC) cells¹⁶⁰. These are small undifferentiated cycling cells that reside between Paneth cells and serve as the primary stem cell population driving intestinal regeneration. In summary, Lgr5+ CBC cells play a central role in intestinal physiology and renewal, while the +4 stem cells serve as a quiescent alternative reserve for slower regeneration in response to injury and stress¹⁶¹.

1.2.2 The Wnt/ β -catenin pathway

The Wnt pathway is essential for ISC maintenance and physiology, but also orchestrates a wide range of physiological processes, including embryonic development, cell proliferation and differentiation, and tissue regeneration. It is highly conserved and initially emerged in evolution as a regulator of longitudinal embryonic axis patterning and embryonic development¹⁶². It consists of two distinct branches, the canonical (Wnt/ β -catenin) pathway and the non-canonical (β -catenin-independent) pathway. The canonical Wnt pathway is extensively studied and activates the Wnt-mediated transcriptional repertoire, while the non-canonical pathway shapes cellular spatial polarization, coordinating cellular division, migration, and

elongation. The non-canonical pathway can be further divided into the Planar Cell Polarity (PCP) pathway and the Wnt/Ca²⁺ pathway, which operate through different molecular mechanisms to orchestrate cellular polarization.

The canonical Wnt pathway initiates when Wnt ligands bind to Frizzled receptors (FZD), triggering a cascade of intracellular events, involving multiple proteins, that regulate a complex gene expression program. Briefly, the Wnt/ β -catenin pathway switches between an ON and OFF state, based on the presence or absence of Wnt ligands accordingly. In the OFF state, cytoplasmic β -catenin, the key component of the signaling cascade, is phosphorylated by the destruction complex and subsequently degraded by the proteasome. Constant degradation prevents β -catenin from entering the nucleus, keeping the Wnt-regulated transcriptional program repressed (Figure 7, Wnt OFF). Conversely, when Wnt ligands bind to the membrane receptors (ON state), the destruction complex is recruited on the receptors, enabling the stabilization of β -catenin. Next, β -catenin translocates to the nucleus, where it interacts with the TCF/LEF complex to activate the expression of multiple target genes (Figure 7, Wnt ON)^{163–165}.

The binding of Wnt ligands to FZD receptors is critical for activating the canonical Wnt pathway. Wnt ligands are secreted proteins, approximately 400 amino acids long and rich in cysteine residues, which regulate cell proliferation, polarity, and cell fate determination.¹⁶⁴. Among these, only Wnt3a and Wnt1 can activate the Wnt/ β -catenin pathway. Upon their translation, Wnt proteins are transferred to the endoplasmic reticulum (ER), where they interact with enzymes and chaperones and undergo several post-translational modifications, including glycosylation, acylation, and lipid modifications. Glycosylation commonly occurs in the N-terminus of Wnt proteins and is critical for their proper protein folding and secretion. Acylation occurs in glycosylated Wnt and is essential for intracellular processing and signaling activity. Lipid modifications are catalyzed by the enzyme PORCN in the ER, and are essential for activating the signaling cascade, as non lipid-modified Wnt proteins fail to bind to FZD receptors^{166,167}.

Wnt biogenesis and secretion have been mainly studied for Wnt3a. To become active, it undergoes two lipid modifications at cysteine 77 and serine 209. Mutations in these residues, or in *PORCN*, limit Wnt secretion, ultimately impairing Wnt activation. Once lipid-modified in the ER, Wnt3a is transferred between endosomes and the Golgi apparatus by the retromer complex¹⁶⁸. Finally, the transmembrane protein Wntless (WLS or Evi) binds to modified Wnt3a in the ER and transports it to the plasma membrane for secretion¹⁶⁹.

Wnt ligands activate Wnt/ β -catenin signaling by binding to a heterodimeric receptor complex comprising FZD receptors and LDL-receptor-related proteins 5 and 6 (LRP5/6). FZD receptors are seven-pass transmembrane proteins with an extracellular cysteine-rich domain (CRD) in the N-terminus. Structural analysis has revealed that the CRD domain serves as a scaffold containing multiple sites for interaction with the C-terminus of lipid-modified Wnts¹⁷⁰. The C-terminal side of FZD receptors interacts with another critical component of the pathway, the Dishevelled protein (Dvl). The binding of Wnt to FZD receptors induces a conformational change that leads to the dimerization of the FZD and LRP5/6 receptors. The LRP5/6 proteins act as coreceptors and play a pivotal role in activating the Wnt pathway. Their intracellular side contains multiple sites that get phosphorylated by several kinases, such as GSK3^{163–165}.

Several proteins can act as agonists of Wnt ligands to induce or enhance the canonical Wnt signaling cascade. An interesting example is R-spondin, a protein that indirectly enhances

the signaling outcome in low doses of secreted Wnt ligands^{171–173}. Specifically, R-spondin binds to the Lgr4/5 receptor, inducing it to interact with Znr3/Rnf3, a complex that normally leads Wnt receptors to degradation¹⁷⁴. Lgr4/5 prevents Znr3/Rnf3-mediated membrane Wnt receptor clearance, enabling the activation of the pathway. Notably, Lgr5 is a critical orchestrator of stemness and tissue regeneration through R-spondin-mediated signaling, and its expression levels are used to mark intestinal stem cells¹⁶⁰. Another case is Norrin which interacts with FZD4 and LRP5/6 to activate the Wnt/ β -catenin signaling pathway. While Norrin is unrelated to Wnt protein, crystallography experiments have demonstrated that its interaction with the two receptors, simulates the finger-like loop of Wnt, enabling it to mimic the action of Wnt ligands^{175,176}.

Conversely, several antagonists inhibit Wnt/ β -catenin signaling activity by affecting different components of the pathway. sFRPs (secreted Frizzled related proteins), WIF (Wnt inhibitory protein), glypicans, Tiki, and Notum directly target Wnt ligands¹⁶⁵. sFRPs and WIF contain particular motifs that allow them to bind to Wnt ligands, preventing them from activating the Wnt/ β -catenin pathway^{177,178}. Similarly, glypicans contain CRD domains that facilitate their binding to Wnt proteins, ultimately controlling their extracellular levels and distribution¹⁷⁹. Tiki and Notum are enzymes that catalyze Wnt protein inactivation^{180,181}.

Except for the antagonists that target the Wnt proteins, additional regulators can affect downstream components of the Wnt pathway. Particularly, proteins of the Dickkopf (DKK) and Sclerostin/SOST families bind to the LRP5/6 receptor, perturbing the Wnt-mediated dimerization of the FZD-LRP complex. As mentioned earlier, Znr3 and Rnf3 are responsible for the ubiquitination of the lysines in the FZD receptors, leading them to rapid lysosomal degradation^{174,182}. Finally, Shisa proteins trap the FZD proteins in the ER, preventing it from reaching the plasma membrane¹⁸³.

Moreover, several signaling pathways can modulate Wnt activity. An interesting example is the Hippo pathway, whose key transcriptional effectors YAP/TAZ play dual roles in the regulation of Wnt signaling. In the absence of Wnt ligands, they interact with the destruction complex (DC) to enhance the β -catenin degradation. However, upon activation of the Wnt pathway, YAP/TAZ are released from the DC and accumulate into the nucleus where they promote the Wnt-dependent transcriptional program¹⁸⁴. Furthermore, critical components of the Wnt pathway are also involved in the molecular mechanisms of other signaling pathways, facilitating the crosstalk of the Wnt pathway with multiple biological processes¹⁸⁵.

The phosphorylation of β -catenin is the determining factor of the Wnt pathway's output. In the Wnt-OFF state, cytosolic β -catenin is targeted for degradation by DC, which comprises Axin, GSK3, CK1 α , and APC. Axin serves as a scaffold for the recruitment of the other DC factors and β -catenin, while APC enhances the interaction between Axin and β -catenin. CK1 α and GSK3 are kinases that sequentially phosphorylate serine 45 (CK1 α), threonine 41, serine 37, and serine 33 (GSK3). Next, the phosphorylation of serine 33 and serine 37 form a degron motif, which enables the binding of β -Trcp, an E3 ubiquitin ligase that ubiquitinates β -catenin, leading it to degradation. The cooperation of all these components to achieve phosphorylation and degradation of β -catenin is tightly regulated, and mutations affecting their expression levels or function are linked to diseases, particularly cancer (Figure 7, APC mutation)^{163–165}.

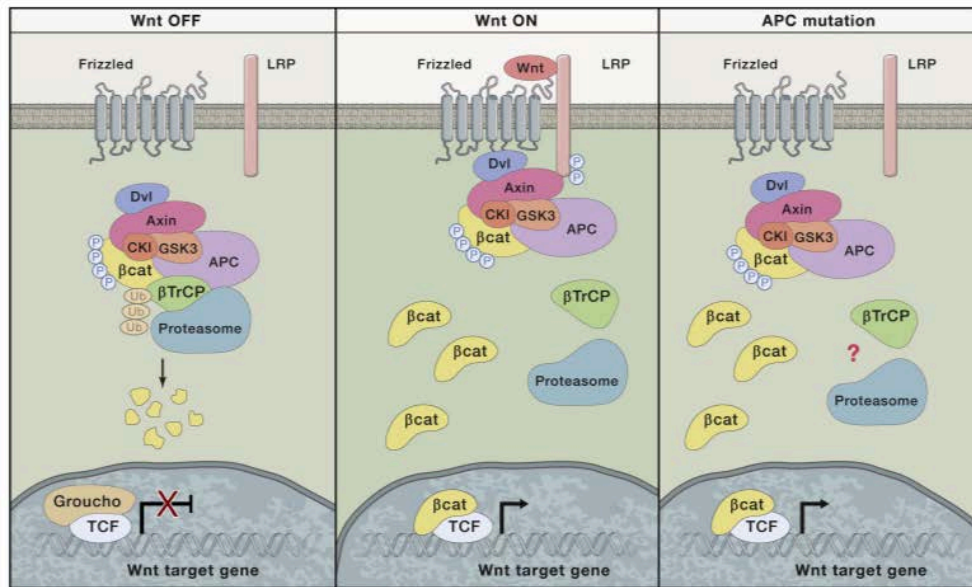


Figure 7: The Wnt/β-catenin pathway in homeostasis and disease (APC mutation).

In the “Wnt off” state, the absence of Wnt ligands allows the assembly of the DC comprising Axin, GSK3, APC, and other components. The DC complex phosphorylates β-catenin, targeting it for ubiquitination and subsequent proteasomal degradation. Also, the Groucho repressor proteins interact with the TCF transcription factors preventing them from activating gene expression. In the “Wnt ON” state, Wnt ligands bind to their receptors, provoking a conformational change in FZD/LRP5/6 complex that displaces the DC to the plasma membrane, enabling the expression of cytosolic β-catenin which in turn, translocates to the nucleus. Then, it interacts with TCF transcription factors and replaces Groucho proteins, initiating the wnt-transcriptional repertoire. Mutations in the APC gene that result in dysfunctional APC protein prevent β-catenin degradation enabling its accumulation and ultimately causing constitutive activation of the Wnt pathway¹⁶³.

Conversely, when Wnt ligands bind to the FZD receptors (Wnt-ON state), a conformational change of the FZD and LRP5/6 receptors takes place, displacing the DC to the plasma membrane. Upon the formation of the receptor complex, Dvl binds on FZD, serving as a scaffold for the additional recruitment of Axin and GSK3 to subsequently phosphorylate LRP6. These phosphorylations occur in two PPPSPxS motifs (P, proline; S, serine or threonine, x, a variable residue) of LRP6 and are the docking site for the DC complex¹⁸⁶. Phosphorylated LRP6 also acts as a direct inhibitor of GSK3, preventing it from phosphorylating β-catenin¹⁸⁷. Structural analysis of the DC during the Wnt-ON state showed that ubiquitination of phosphorylated β-catenin is blocked within the DC, leading to its saturation¹⁸⁸. When the DC is inactive, β-catenin levels dramatically increase in the cytoplasm, and it subsequently translocates into the nucleus where it interacts with the TCF protein family (Figure 7, Wnt ON).

During the Wnt-OFF state, TCF transcription factors interact with Groucho proteins to promote histone deacetylation and chromatin compaction, leading to gene repression. Upon activation of the Wnt/β-catenin pathway, nuclear β-catenin displaces Groucho and interacts with

the TCFs to induce the Wnt-dependent transcriptional repertoire. There are four TCF proteins in mammals: TCF1, LEF1 (TCF7), TCF3, and TCF4, all acting as transcription factors regulating gene expression. Specifically, they bind with high affinity to the sequence CCTTTGWW (W represents either T or A), which is called the Wnt responsive element (WRE) and is found in multiple β -catenin/TCF4 target genes, often located at large distances from their TSS¹⁸⁹.

TCF1 and TCF4 can act either as activators or repressors, based on their interacting partner (β -catenin and Groucho accordingly), LEF1 primarily acts as activator, while TCF3 is mainly a repressor¹⁶⁴. In a recent study, they performed CRISPR-Cas9-mediated genetic engineering to knock-out all TCF genes in HEK293T cells and investigate the TCF-independent role of nuclear β -catenin. Interestingly, they demonstrated that even though β -catenin lost multiple binding sites, it could still interact with chromosomes and regulate the expression of multiple target genes without the involvement of TCFs¹⁹⁰.

Additional regulatory mechanisms control TCF/ β -catenin-dependent transcription. Notably, *TCF-1* and *LEF-1* genes can produce the dominant negative (dn) isoforms dnTCF-1 and dnLEF-1 respectively, using alternative promoters. These protein isoforms lack the β -catenin-binding domain and act as inhibitors of Wnt signalling^{191,192}. For instance, dnTCF-1 is expressed in normal colon cells and inhibits the activity of TCF4, while its decrease is associated with colorectal carcinogenesis¹⁹³. Furthermore, several proteins that interact with β -catenin lead to the disassembly of the β -catenin/TCF complex¹⁶⁴. For example, Chibby and ICAT are two nuclear proteins that bind to β -catenin to disrupt its interaction with TCF proteins and export it outside the nucleus^{194,195}. Finally, post-translational modifications of TCF proteins alter their binding affinity to either β -catenin or Groucho, ultimately enhancing or diminishing the Wnt-mediated transcriptional activity.

The β -catenin/TCF complexes bind and affect the expression of a plethora of genes involved in ISC maintenance. As mentioned, *Lgr5* was initially identified as a Wnt signaling target and has since been established as a key marker for ISCs within the intestinal crypts^{150,160}. Similarly, the same group identified *ASCL2* as a direct target of β -catenin/TCF4, encoding a transcription factor that governs intestinal stem cell fate. Overexpression of *ASCL2* results in crypt hyperplasia and the development of ectopic crypts in the villi¹⁹⁶. *ASCL2* interacts with β -catenin/TCF4 to synergistically induce the expression of Wnt targets. Furthermore, Wnt/R-spondin activates an autoregulatory loop that enables *ASCL2* to regulate its own locus¹⁹⁷.

An interesting target is *Axin2* whose encoded protein acts as a scaffold for the DC, and negatively regulates the Wnt pathway, thereby creating a negative feedback loop that keeps Wnt pathway activity under control¹⁹⁸. Another critical target is the oncogene *MYC*, which contains multiple promoter proximal WREs for β -catenin/TCF4 binding. Wnt activation through APC inactivating mutations induces *MYC* expression, while the overexpression of dnTCF4 resulted in *MYC* repression^{199,200}.

1.2.3 The Wnt/ β -catenin pathway in intestinal carcinogenesis

CRC is the second most deadly cancer worldwide, with a higher incidence observed in developed countries and in men. Both genetic predisposition and environmental factors, including lifestyle and diet, contribute to CRC pathogenesis. Also, individuals with ulcerative

colitis and Crohn's disease face an elevated risk of CRC development. Hence, CRC cases can be broadly classified as sporadic, hereditary, or colitis-associated²⁰¹.

Early screening remains the most effective measure for CRC prevention, as it typically takes approximately 10 years for an adenoma to progress to carcinoma. Early detection facilitates the removal of precancerous or non-metastasized polyps, ultimately improving patient outcomes²⁰². Current screening methods, including colonoscopies, sigmoidoscopies, colonography, and fecal immunochemistry can only detect around 40% of early cases^{201,203}. The most common therapeutic strategies are surgical excision of the polyps and chemotherapeutic treatment, however, in many cases, it recurs or metastasizes. Therefore, understanding the molecular mechanisms underlying CRC development and progression will help us gain insights into identifying novel biomarkers for early prediction and exploring new therapeutic avenues²⁰⁴.

CRC initiation involves the accumulation of genetic and epigenetic alterations within the ISCs in the crypt niche. Aberrant activity of the Wnt pathway leads to uncontrolled proliferation of the stem cells which serve as the cancer-initiating cells, continuously fueling tumor growth. These cells divide asymmetrically, generating differentiated epithelial cells that constitute the bulk of the tumor²⁰⁵. Polyps arising from the proliferation of non-cancerous cells are initially benign, but persistent growth and prolonged inflammation can drive dysplasia and subsequently carcinogenesis (Figure 8). Then, the adenocarcinoma can become invasive and spread to other organs through the blood and lymphatic systems, with the liver being the primary organ of metastasis²⁰¹.

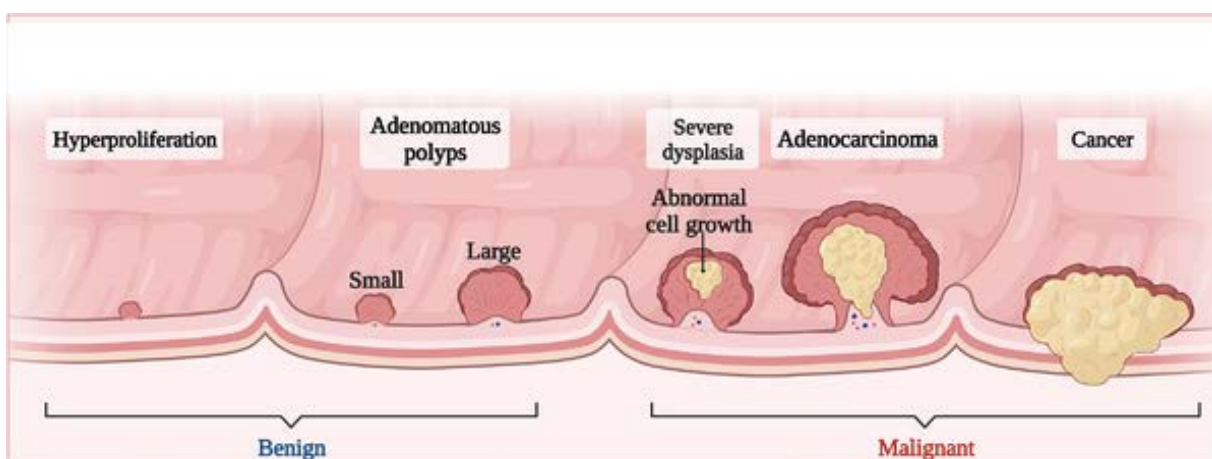


Figure 8: CRC development and progression.

Hyperproliferation of ISCs due to genetic defects leads to the formation of adenomatous polyps. Continuous inflammation drives dysplasia resulting in cancer development. Image adapted by Hanitrariimalala et. al.

Due to its critical role in ISC maintenance, the Wnt pathway is a pivotal driver of CRC^{163–165}. Mutations that affect the expression levels or disrupt the function of key components and regulators of the Wnt pathway lead to its aberrant activation, ultimately promoting carcinogenesis^{206–208}. Somatic mutations in the tumor suppressor gene *APC* are the most frequent in sporadic CRC, occurring in more than 80% of cases. Additionally, loss of heterozygosity of this genomic region is observed in 30-40% of CRC cases²⁰⁹. Germline

mutations in *APC* cause familial adenomatous polyposis (FAP), a hereditary disorder that predisposes individuals to CRC development²¹⁰. *APC* is an essential component of the DC, so mutations that disrupt its binding domains promote the stabilization of β -catenin and the uncontrolled Wnt pathway activation²¹¹. Moreover, loss of *APC* function is associated with extensive chromosomal instability in CRC²¹². Interestingly, it was found that different *APC* genotypes lead to variable expression levels of *APC* which correlate with distinct Wnt/ β -catenin activity²¹³.

Beyond *APC*, missense and truncating mutations of *Rnf43* are frequently detected in CRC. Loss-of-function mutations in *Rnf43* prevent β -catenin ubiquitination by E3 ubiquitin ligases, thereby preventing its degradation and maintaining persistent Wnt pathway activation²⁰⁶. Additionally, mutations in exon 3 of *CTNNB1*, disrupt the phosphorylation sites recognized by GSK3 kinase, allowing the constant activation of the pathway²¹⁴. Similarly, *Axin2* mutations result in β -catenin stabilization, inducing CRC development through defective mismatch repair²¹⁵.

Dysregulation of the Wnt pathway is also implicated in the development of other types of tumors such as cholangiocarcinoma, leukemia, melanoma, and breast cancer²⁰⁷. Moreover, It is associated with various non-cancerous diseases, including lung inflammatory disorders, cardiovascular diseases, neurodegenerative conditions, and bone diseases¹⁶⁵. Currently, several therapeutic interventions are being developed based on our knowledge of the Wnt pathway, and are even tested in clinical trials. Specifically, these include antibodies targeting Wnt receptors, synthetic peptides mimicking Wnt ligands, small-molecule inhibitors targeting critical components of the pathway (such as Porc, GSK3, CK1, and Dvl), and agents that block β -catenin/TCF interactions²⁰⁶.

1.2.4 Functional interplay between the Wnt pathway and lncRNAs

lncRNAs have emerged as critical regulators of the Wnt pathway activity. They can affect Wnt signaling by modulating the expression levels of Wnt pathway components or targets from the transcriptional to the post-translational level^{216,217}. Notably, several lncRNAs stabilize β -catenin, enhancing Wnt pathway activity¹⁶⁵. The characterization of the β -catenin/TCF4 binding sites in CRC cells revealed that many target genes reside within the non-coding part of the genome, suggesting a previously underappreciated layer of regulation within the Wnt pathway¹⁸⁹.

Several lncRNAs affect the Wnt pathway by interacting with transcriptional regulators. CCAT2 is a lncRNA, transcribed from a region adjacent to the *MYC* gene, which physically interacts with TCF4, enhancing its transcriptional activity. The CCAT locus is also regulated by TCF4, suggesting the existence of a positive feedback loop in the CCAT-mediated Wnt induction (Figure 9A)²¹⁸. Similarly, ASBEL is a β -catenin-activated lncRNA that interacts with TCF3 to promote the Wnt-regulated transcriptional program. Moreover, ASBEL recruits TCF3 on the TSS of the tumor suppressor *ATF3*, thereby reducing its expression. On the other hand, *ATF3* negatively regulates ASBEL, forming a negative feedback loop (Figure 9A)²¹⁹. Accordingly, RBM5-AS1 is a nuclear lncRNA that binds β -catenin and enhances its interaction with TCF4, ultimately increasing the β -catenin/TCF4 occupancy on target sites. Elevated levels of RBM5-AS1 result in increased Wnt signaling activity and proliferation in ISCs (Figure 9A)²²⁰. An

alternative example is CCAL which promotes the ubiquitination of AP-2a, a negative regulator of the Wnt pathway, ultimately leading to its proteasomal degradation. Thus, CCAL indirectly activates the transcriptional activity of Wnt targets and is upregulated in CRC (Figure 9A)²²¹.

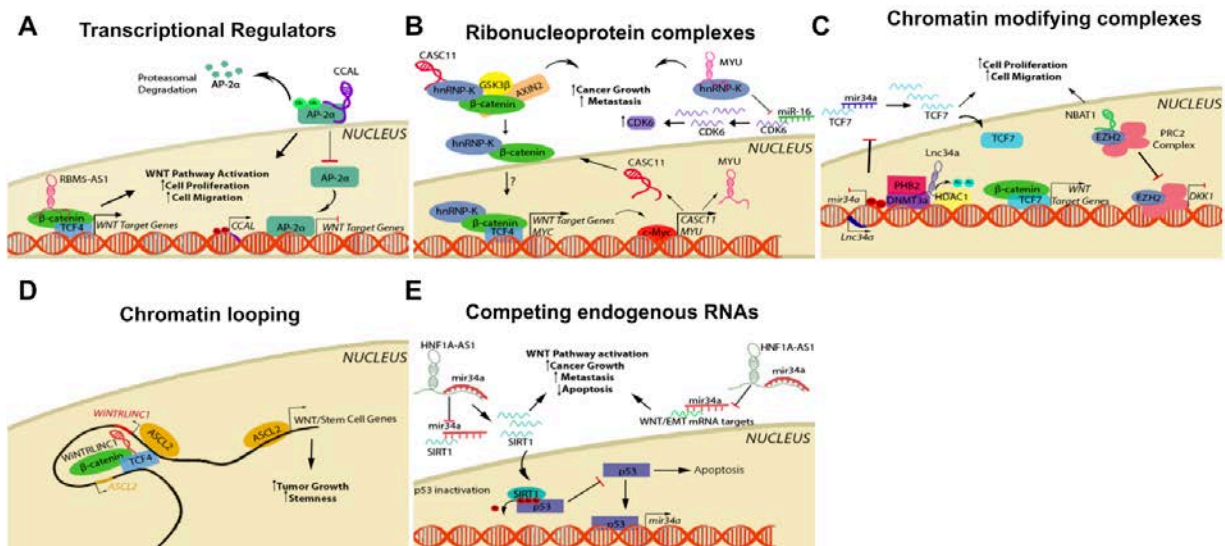


Figure 9: Functional interplay between lncRNAs and the Wnt pathway.

A. Many lncRNAs interact with transcription factors, either recruiting them to their target genes (e.g., RBM-AS1) or sequestering them away (e.g., CCAL). B. Other lncRNAs bind to regulatory proteins and act as scaffolds for the assembly of ribonucleoprotein complexes, ultimately stabilizing their targets (e.g., CASC11, MYU). C. Several lncRNAs affect the recruitment of chromatin remodelers to activate (e.g., NBAT1) or inhibit (e.g., lnc34a) the expression of genes implicated in Wnt signaling. D. WNTRLINC1 forms an intra-chromosomal loop to regulate ASCL2 expression by forming a complex with β -catenin/TCF4 and ASCL2 itself. E. Cytoplasmic lncRNAs affect the Wnt-mediated transcriptional repertoire by sequestering miRNAs from their target mRNAs (e.g., HNF1A-AS1). Image adapted by Zarkou et. al. 2018.²¹⁶

Other lncRNAs participate in ribonucleoprotein complexes to modulate Wnt pathway activity. Notably, CASC11 is a cytoplasmic lncRNA that stabilizes heterogeneous ribonucleoprotein K (hnRNP-K). The hnRNP-K protein interacts with β -catenin and components of the DC, including Axin2 and GSK3, preventing them from leading β -catenin to degradation. Interestingly, CASC11 is bound by the Wnt target MYC, which activates its expression, completing a positive feed-forward loop that sustains Wnt signaling active (Figure 9B)²²². MYC also regulates the expression of another lncRNA target, MYU, which stabilizes CDK6 expression by interacting with hnRNP-K and protecting it from microRNA-mediated degradation. CDK6 is a cyclin-dependent kinase, so its stabilization by the Wnt/c-Myc/MYU axis facilitates the G1-S transition and promotes tumorigenesis in CRC cells (Figure 9B)²²³.

LncRNAs may also modulate Wnt activity by recruiting chromatin remodelers on crucial genes. For example, lnc34a recruits DNMT3a/PHB2 and HDAC1 on the *mir-34a* promoter,

inducing methylation and deacetylation to repress its expression. Mir-34a is a microRNA that targets TCF7 mRNA, negatively regulating the Wnt-mediated outcome. Therefore, high levels of lnc34a lead to decreased mir-34a expression and induced Wnt activity in CRC (Figure 9C)²²⁴. Similarly, lncTCF7 acts as a positive regulator of the Wnt pathway and promotes the proliferation of liver cancer stem cells. Specifically, lncTCF7 recruits the SWI/SNF complex to the *TCF7* promoter, increasing its transcriptional activity (Figure 9C)²²⁵. Conversely, NBAT1 interacts with the EZH2 catalytic subunit of the PRC2 complex to prevent it from methylating the *DKK1* promoter. *DKK1* encodes a Wnt signaling inhibitor that targets LRP6²²⁶. Hence, NBAT1 induces the expression of DKK1 to negatively regulate Wnt pathway activity and suppress breast cancer metastasis (Figure 9C)²²⁷.

An intriguing case is *WNTRLINC1*, whose gene is a direct target of the β -catenin/TCF4 complex. The *WNTRLINC1* locus resides in the vicinity of the *ASCL2* gene (16 kb distance) which expresses a critical transcription factor for ISC physiology (see section 1.2.2). *WNTRLINC1* is essential for the formation of an intra-chromosomal loop that bridges the *WNTRLINC1* promoter with the *ASCL2* downstream enhancer. The looping facilitates the recruitment of RNA pol II, along with β -catenin/TCF4 and mediator complexes to the regulatory regions of both genes. At the same time, *ASCL2* binds to the *WNTRLINC1* promoter, upregulating its expression. The *WNTRLINC1*/*ASCL2* axis promotes Wnt signaling activity, acting as a regulatory axis for ISCs maintenance (Figure 9D)²²⁸.

Cytoplasmic lncRNAs can also modulate microRNA function, often acting as competing endogenous RNAs that prevent microRNAs from reducing the expression levels of their targets. For instance, HNF1A-AS1 acts as a decoy of miR-34a sequestering it from binding to conserved regions within the 3' UTR of essential Wnt signaling components, including Wnt ligands, LRP6, LEF1, AXIN2, and β -catenin²²⁹. Moreover, HNF1A-AS1 promotes SIRT1-mediated deacetylation of the tumor suppressor *p53*, ultimately inducing the expression of the Wnt targets MYC and CCND1 to enhance CRC metastasis (Figure 9E)²³⁰.

Other cytoplasmic lncRNAs interact directly with Wnt pathway components to modulate β -catenin activity. lncRNA-FAST binds to the ubiquitin ligase β -TrCP and inhibits β -catenin ubiquitination, enabling it to stabilize and accumulate within the nucleus²³¹. Another example is NEAT1 which interacts with DDX5, facilitating the assembly of a β -catenin/DDX5 complex that activates the Wnt signaling, ultimately promoting CRC progression²³².

1.3 Small open reading frames (smORFs)

Eukaryotic genomes harbor millions of smORFs ranging from 10 to 100 codons in length. They are widely distributed across the genome, most residing in intergenic regions, while thousands are located within lncRNA genes²³³. Despite their large numbers, they remain unannotated by conventional computational annotation systems, leaving them “hidden” from the research community.

This is primarily the result of the limitations of current computation methods used to annotate protein-coding genes, which heavily rely on evolutionary conservation through sequence similarity. Due to their small size, smORFs tend to exhibit lower conservation scores compared to protein-coding genes. A study demonstrated that smORFs that are actively

translated tend to be highly conserved in animals²³⁴, but recent discoveries suggest that they can be the result of *de novo* gene birth in humans²³⁵. An arbitrary threshold has been established, wherein only open reading frames (ORFs) longer than 100 codons are considered as canonical protein-coding genes^{233,236}. Additionally, traditional annotation methods solely use AUG as the initiation codon to predict canonical protein-coding genes, despite the existence of multiple cases of non-cognate start codons (CUG, GUG, UUG) initiating translation^{237,238}.

Advances in Ribo-seq and mass spectrometry (MS) have enabled the detection of novel smORF-encoded peptides (SEPs), expanding our understanding of the proteome. Translation is pervasive in non-protein-coding genes, yet it remains largely uncharacterized²³⁸. Nevertheless, several studies have revealed that hundreds of smORFs are translated into SEPs^{239–241}. It is estimated that nearly 1.2% of smORFs are transcribed with only one-third being ultimately translated, suggesting that there are hundreds of microproteins whose existence is currently ignored²³³.

1.3.1 Experimental discovery of novel SEPs

Ribo-seq has emerged as a critical method for providing evidence that a smORF is putatively translated. It is a powerful technique that is used for the monitoring of protein synthesis, the detection of translation initiation sites, and the discovery of novel translation events²⁴². In Ribo-seq, cells are treated with antibiotics such as cycloheximide or lactimidomycin to arrest translational elongation, ultimately halting ribosome movement on the RNAs. The process then involves the deep sequencing of RNA fragments, typically around 28–32 nucleotides in length, which are protected by the 80S ribosomes, enabling the assessment of global translational activity. The RNA fragments enclosed by the ribosomes are protected by nuclease treatment and are used as template for deep sequencing. Therefore, the number of aligned reads can be used as an indirect measurement of ribosome intensity on RNAs or protein synthesis levels, and also allow the specification of the exact positions that are bound by the ribosomes (Figure 10A)^{97,242,243}.

Bioinformatic analysis of Ribo-seq datasets is crucial for discovering novel translatable smORFs. The initial steps of alignment and read count estimation are similar to RNA-seq workflows. However, the interpretation of ribosome footprints and the detection of translational events requires the application of more sophisticated tools²⁴⁴. These tools incorporate parameters such as ribosome occupancy, and trinucleotide periodicity. For example, *translation efficiency (TE)* measures ribosome occupancy compared to transcriptome abundance²⁴⁵, while *ORFscore* estimates the enrichment of reads on one frame compared to a uniform signal distribution over the three frames²⁴⁰. The *Ribosome Release Score (RRS score)* estimates the ratio of ribosome-protected fragments on the coding sequence to those in the 3' UTR, normalized by their corresponding lengths and mRNA coverage. The *RRS score* effectively distinguishes between protein-coding and non-coding genes based on the observation that ribosomes are rapidly released at stop codons of bona fide translated mRNAs³⁷.

More advanced methods employ machine learning approaches to detect translation events²⁴⁶. *ORF-RATER (ORF Regression Algorithm for Translational Evaluation of RPFs)* employs a linear regression model to analyze aggregated ribosome profiling patterns at start and stop codons, with a positive regression coefficient indicating translational activity of an ORF.

Then the model is statistically validated using a random forest classifier that was trained from known protein-coding ORFs, ultimately generating a score ranging from 0 to 1, with highest scores suggesting a greater likelihood of translation²⁴⁷. Moreover, the *Fragment Length Organization Similarity Score (FLOSS)* takes advantage of the characteristic length of ribosome footprints to distinguish true translational signals from non-specific contaminants. The *FLOSS score* is estimated for each ORF by comparing the similarity of its fragment length distribution to a consensus reference distribution of expected ribosome-protected fragments²⁴⁸.

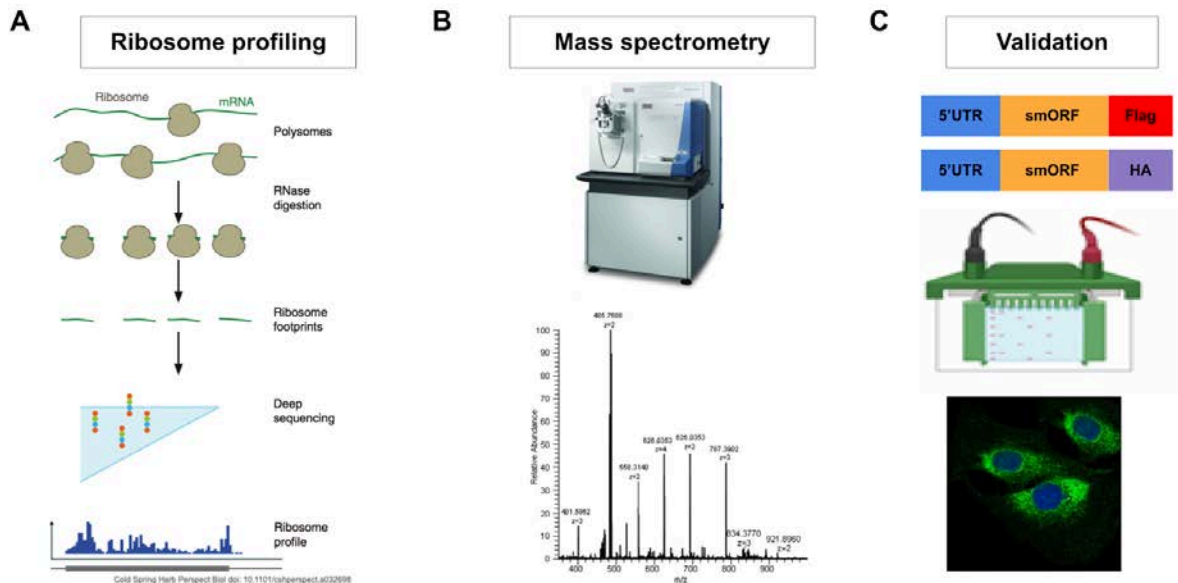


Figure 10: Workflow for the identification of novel translatable smORFs.

A. Schematic representation of a typical ribosome profiling experiment pipeline for identifying translatable smORFs. B. Mass spectrometry analysis detects spectra corresponding to predicted SEP sequences, utilizing a custom proteome library derived from Ribo-seq data. C. Validation of smORF translation through the incorporation of FLAG or HA epitopes into the smORF sequence, followed by confirmation using traditional methods such as Western blotting or immunofluorescence. Image generated by the combination of different sources (ohmx.bio¹, [Broad Institute](https://www.broadinstitute.org)²)

MS is an essential complementary approach to use for detecting novel SEPs. Recent advances in proteomics have increased the sensitivity of MS, enabling the detection of even low-abundance proteins. Although MS-based identification of peptides that match a predicted SEP offers strong evidence of their existence, the technology faces several limitations for the detection of SEPs, with only few successfully validated through MS (Figure 10B)^{249,250}. The Human Proteome Organization/Human Proteome Project (HUPO/HPP) requires at least two non-overlapping unique peptides of length greater than 9 amino acids to prove that a protein is

¹ <https://ohmx.bio/translatomics-ribosome-profiling-service/ribosome-profiling>

² <https://www.broadinstitute.org/technology-areas/what-mass-spectrometry>

expressed based solely on MS-based approaches²⁵¹. Because many SEPs remain unannotated or their translation initiates from non-canonical start codons, scientists generate custom protein libraries derived from Ribo-seq data to use as input for MS analysis. Due to their small size, SEPs generate fewer unique tryptic peptides. Also, proteins shorter than 80 amino acids are more prone to digestion by endogenous proteases²⁵², reducing the number of peptides that can be trypsinized. Additionally, several SEPs identified by Ribo-seq technologies are rapidly degraded after their synthesis, so MS approaches cannot capture them²⁵³. To address these, *Khitun & Slavoff* suggest a proteogenomic pipeline which includes the enrichment of small proteins through gel-size selection and improved MS bioinformatic workflow for increasing the sensitivity of SEPs detection²⁵⁴.

While Ribo-seq and MS are robust methods for discovering novel translatable smORFs, experimental validation is further required to confirm the translational activity. The most common strategy involves the ectopic overexpression of the putative smORF, tagged with epitopes, such as FLAG or HA, followed by immunostaining or Western Blotting (Figure 10C). These constructs typically include the 5' UTR which is critical for translation initiation, as it contains RNA secondary structures, Internal Ribosome Entry sites (IRES), or motifs for RNA-binding proteins²⁵⁵. However, this method is labor-intensive and must be conducted for each SEP individually.

Recent advances in CRISPR screenings have enabled the massive investigation of multiple smORFs at the same time, significantly accelerating smORF identification²⁵⁶. Specifically, CRISPR loss-of-function assays use libraries of sgRNAs to target different smORF-derived transcripts for knock-out, enabling the interrogation of the effects on cell viability. Observing smORF-dependent effects on cell proliferation increases the likelihood that a smORF is translated, allowing researchers to prioritize specific candidates for further study. The emergence of single-cell CRISPR screens promises new insights into smORF discovery²⁵⁵, though their application in this context is limited yet²³⁸.

1.3.2 Classification of smORFs and SEPs

Despite the technical challenges in discovering novel microproteins encoded by non-coding RNAs, numerous studies have demonstrated the existence of multiple smORFs that are actively translated into SEPs. Notably, *Prensner et. al.* employed CRISPR screening to interrogate the translation activity of 550 Ribo-seq-derived smORFs. They discovered that 257 smORFs are translated, while the ectopic overexpression of 401 smORFs resulted in transcriptomic changes in HEK293 cells²⁵⁶. A recent global consortium aimed to create a comprehensive catalog of potentially translated smORFs, focusing on 7,264 smORFs identified through Ribo-seq data. The study was limited to smORFs that initiate translation from canonical ATG start codons and are at least 16 codons in length. Several of these smORFs were also validated to encode functional proteins and are now included in the GENCODE database²⁴¹. The same consortium upgraded this catalog using a pan-proteome analysis that integrates proteomic and immunopeptidomic data from 95520 experiments. This large-scale effort revealed that at least 25% of the cataloged smORFs produce microproteins at some stage. Alternative approaches to catalog smORFs include the development of databases that are based on published Ribo-seq results, such as SmProt²⁵⁷ and OpenProt²⁵⁸.

SmORFs can be classified into diverse classes based on their genomic location, each with distinct features and functions. The most prevalent class of smORFs are intergenic ORFs which are DNA segments between an ATG start codon and a stop codon, without overlap with protein-coding genes, residing throughout the genome (Figure 11). These are random genomic sequences that are generally not translated, though a few might play regulatory roles^{233,259}.

Transcribed smORFs	ORF class	RNA type	Median size (codons)	Translation ¹⁵	Conservation	Coding features	Function
	Intergenic ORFs	None	22	None	None ^{6,8}	Non-canonical AA	None
	uORFs	5' UTRs of mRNAs	22	Low	None ^{8,30}	• Nonrandom AA • No domains	• Non-coding • Translation regulation
	lncORFs	lncRNAs	24	Low	None ^{8,10}	• Nonrandom AA • No domains	Non-coding or coding
	Short CDSs	Short mRNAs	79	High	Class	• Positively charged AA • Transmembrane α -helices	• Coding • Regulators of canonical proteins
	Short isoforms	Spliced mRNAs	79	High	Kingdom	• Canonical AA • Protein domain loss	• Coding • Small interfering peptides
	Canonical ORFs	mRNAs	491	High	Kingdom	• Canonical AA • Multiple protein domains	• Coding ⁴² • Structural, enzymatic, regulatory

Figure 11: Classification and characteristics of smORFs in fruit flies and mammals.

Intergenic smORFs are scattered throughout the genome between genes. These smORFs are not transcribed and do not encode proteins. Upstream ORFs (uORFs) are located in the 5' UTR of the transcripts before the CDS and primarily regulate translation efficiency of the downstream ORF. lncORFs represent smORFs hidden in genes that have been annotated as lncRNAs. Short CDSs are smORFs located within protein-coding genes, producing mRNAs but encoding sequences shorter than 100 codons. Short isoforms are alternative isoforms of canonical CDSs, also limited to fewer than 100 codons. Canonical ORFs, in contrast, are the standard ORFs exceeding 100 codons in length. These are predominantly transcribed and translated into functional proteins²³³.

The second most prominent class is upstream ORFs (uORFs) which are located in the 5'UTR of coding DNA sequences (CDS). uORFs can terminate before the CDS (Figure 11) or partially overlap with it in an alternate reading frame (uoORFs)²⁵⁹. Notably, they are found in about half of human and mouse genes, with many affecting the translation of their downstream CDS²⁶⁰. uORFs typically reduce the translatability of the downstream CDS, either through their own translational activity or the peptides they produce. Studies have shown that uORF translation can interfere with the scanning mechanism of the ribosomes during the translation initiation of the canonical ORF, ultimately reducing its translation efficiency. In some cases, this disruption has been linked to disease manifestation^{261,262}. When ribosomes initiate translation from the uORF's ATG and encounter a stop codon, reinitiation at the downstream ATG is insufficient²⁶³. However, this effect is mitigated in cases where the initiation site of the uORF is

suboptimal, leading to leaky scanning that allows more efficient translation of the downstream CDS²⁶⁴. Alternatively, ribosome occupancy on the uORF might prevent the recruitment of ribosomes in the downstream ORF through ribosome stalling or ribosome dissociation²⁵⁹. Moreover, the termination site of the uORF can induce nonsense-mediated decay in the mRNA, ultimately reducing its transcriptional levels²⁶⁵.

LncORFs, a class of smORFs residing within lncRNA genes, represent the third most abundant group of small open reading frames (Figure 11). Ribosome profiling experiments tend to detect multiple ribosome-bound lncRNAs, prompting the research community to delve more deeply into their coding potential. *Guttman et. al.* introduced the *RRS score* to evaluate the translational potential of the lncRNA transcriptome. Their findings revealed that most ribosome-bound lncRNAs are not actively translated, but may result from technical limitations of Ribo-seq technology or serve regulatory roles on the translational machinery³⁷. Nonetheless, several transcripts initially classified as lncRNAs have been found to encode microproteins. Another study exhaustively interrogated the translation of lncORFs in *Drosophila Melanogaster*, discovering more than 200 lncORF-encoded microproteins. Moreover, they observed that robustly translated lncRNAs tend to associate with more ribosomes, facilitated by features such as cistronicity, Kozak sequences, and increased evolutionary conservation²⁶⁶.

More underrepresented categories include short coding sequences (short CDSs), short isoforms, and downstream ORFs (dORFs). Short CDSs are smORFs located in *bona fide* protein-coding sequences, exhibiting typical mRNA features but are smaller than 100 codons (Figure 11). Short isoforms are alternative transcripts or isoforms produced by smORFs in different frames from the CDSs (Figure 11)^{259,266}. dORFs are found in the 3'UTR where they enhance the translation of the corresponding CDSs, with the effect depending more on their numbers than their protein product²⁶⁷. The only dORF that was identified to produce a functional microprotein resides downstream of the *ABCB5* gene and it is implicated in inflammatory response²⁶⁸.

Although the classification of smORFs typically depends on their genomic location, SEPs can also be categorized based on the MS-based approach used for their detection. The common strategies that have been described refer to microproteins mainly discovered by proteomics after tryptic digestion. However, Immunopeptidomics has emerged as a pivotal MS-based approach for characterizing antigen presentation, including the discovery of novel SEPs. Immunopeptidomics have gained increased attention, as HLA-I and HLA-II presented peptides exhibit antigenic properties, serving as putative therapeutic targets, particularly against cancer. These peptides are naturally produced by the proteasome of antigen-presenting cells. Recent studies have demonstrated that non-canonical proteins are more frequently discovered in HLA-I immunopeptidome compared to traditional tryptic whole-cell extracts²⁶⁹. This enrichment is likely due to the plethora of HLA-binding motifs that present diverse peptides, the increased stability of the HLA-peptide complexes, which allows the detection of rapidly degraded microproteins, and the HLA-based representation of low-abundance microproteins²⁴⁶.

1.3.3 Roles of SEPs in cellular physiology and diseases

Several recent studies have highlighted the critical role of SEPs in essential biological processes. SEPs have been identified across various subcellular compartments, where they

regulate transcriptional, translational, post-translational processes, DNA replication and repair, as well as metabolism and regeneration²⁷⁰. Notably, SEPs have been linked to the pathogenesis of diverse diseases, including various cancers, diabetes, and inflammatory disorders^{270–272}. Particularly, SEPs have emerged as novel players in regulating different types of cancers, either promoting or suppressing carcinogenesis²⁷¹.

Early pioneering studies in the field investigated novel lncRNA-encoded SEPs in muscle physiology. Myoregulin (MLN) is a transmembrane microprotein produced by *LINC00948*, and interacts with SERCA, ultimately orchestrating Ca²⁺ uptake in the sarcoplasmic reticulum of muscle cells. Deletion of MLN in mice, improved exercise performance, highlighting its critical role in skeletal muscle physiology (Figure 12)²⁷³. Accordingly, the same group discovered another lncRNA encoding dwarf open reading frame (DWORF), a microprotein that localizes to the sarcoplasmic membrane enhancing SERCA activity by sequestering its inhibitors. Overexpression of DWORF in mouse cardiomyocytes accelerates calcium clearance during contraction and relaxation, while its suppression slows calcium decay²⁷⁴. An alternative example is Minion, which is a conserved 84aa microprotein that is encoded by *GM7325* in mice, and is an essential orchestrator of muscle regeneration. Specifically, it induces cellular fusion through rapid cytoskeletal reorganization by interacting with Myomaker²⁷⁵.

Recent studies have shown that SEPs also play crucial roles in the nucleus, particularly in regulating transcription and DNA repair. Linc-PINT is a well-studied lncRNA induced by p53 that is functionally implicated in multiple pathways and diseases²⁷⁶. The *Linc-PINT* gene produces over 100 spliced isoforms, including a circular isoform, whose exon 2 of CircPIN is translated into the PINT87aa microprotein. PINT87aa interacts with PAF1, a component of the PAF complex, to inhibit transcriptional elongation, exerting tumor suppressive effects in brain cancer tissues in a circRNA-independent manner (Figure 12)²⁷⁷. Another notable example is CYREN, one of the few MS-discovered microproteins, which is one of the three protein isoforms encoded by *C7orf49*. CYREN interacts with Ku70/80 to stabilize DNA breaks and promote non-homologous end joining (NHEJ) repair²⁷⁸. Interestingly, a more recent study demonstrated that CYREN operates in a cell cycle-dependent manner, promoting NHEJ during the G1 phase, but inhibiting it during the S and G2 phases²⁷⁹.

Recent findings have revealed that two lncRNAs encode microproteins involved in splicing regulation. HOXB-AS3 is translated into a 53-aa SEP that interacts with the RGG motif of the RNA-binding protein hnRNP A1, preventing its binding to exon 9 of *PKM*. Thus, the peptide regulates the splicing pattern of *PKM*, increasing the expression of *PKM1* while decreasing *PKM2* levels, both of which are important modulators of aerobic glycolysis and lactate production. The study demonstrated that through this regulatory axis, HOXB-AS3 microprotein, rather than its encoding lncRNA, drives metabolic reprogramming in colon cancer cells, suppressing their growth²⁸⁰. Interestingly, another study found that HOXB-AS3 microprotein binds to IGF2BP2 and stabilizes the mRNA stability of c-Myc mRNA, promoting cell proliferation in oral squamous cell carcinoma cells (Figure 12)²⁸¹. Splicing Regulatory Small Protein (SRSP) is another example of lncRNA-derived SEP involved in the regulation of splicing. Both the LOC90024 lncRNA and its encoded product, SRSP, are upregulated in CRC patients and are associated with poor prognosis. Meng *et. al.* demonstrated that SRSP promotes CRC tumorigenesis in a lncRNA-independent manner by modulating the splicing of *Sp4* transcript. More specifically, SRSP interacts with the splicing factor SRSF3, enhancing its

binding to exon 3 of Sp4, leading to the inclusion of this exon in the mature transcript. The resulting long isoform of Sp4 promotes CRC tumorigenesis, in contrast to the shorter isoform²⁸².

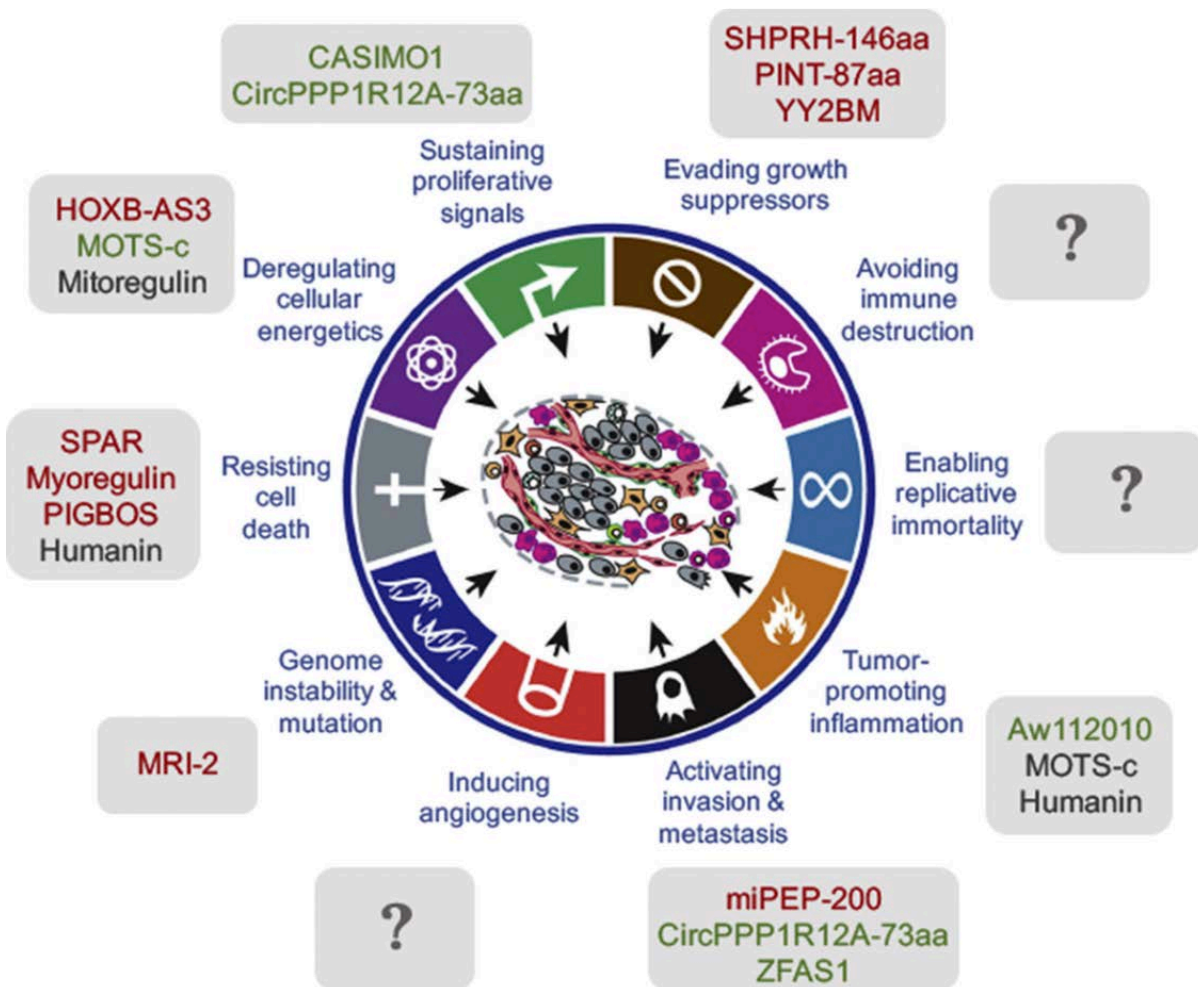


Figure 12: The Role of SEPs in the Hallmarks of Cancer

A graphical representation of the hallmarks of cancer, highlighting the SEPs associated with each biological hallmark according to their functional roles. SEPs depicted in green are involved in promoting oncogenesis, while those shown in red act to suppress or inhibit carcinogenesis²⁷¹.

SEPs significantly influence the endoplasmic reticulum (ER) physiology by modulating key molecular processes, such as protein synthesis and stress response. One such example is ASH1L-AS1, a lncRNA abundantly expressed in acute myeloid leukemia (AML) patients, which encodes the microprotein APPLE. It localizes in the ER and interacts with PABPC1, a critical orchestrator of translational activity rate. Specifically, it strengthens the interaction between PABPC1 and eIF4G, subsequently promoting mRNA circularization and translation. Elevated APPLE levels provoke aberrant translational activity, contributing to AML progression²⁸³. Additionally, PIGBOS is an antisense lncRNA that encodes a conserved microprotein that localizes to the outer mitochondrial membrane, particularly at ER-mitochondria contact sites.

The PIGBOS microprotein interacts with the ER component CLCC1 to modulate the unfolded protein response pathway, a crucial mechanism that prevents the accumulation of unfolded proteins in the ER. Thus, PIGBOS protects the cells from ER-induced apoptosis due to prolonged ER stress (Figure 12)²⁸⁴.

Another interesting case is the non-annotated P-body dissociating polypeptide (NoBody). NoBody interacts with proteins that are involved in mRNA decapping in the P-bodies. NoBody plays a crucial role in regulating P-body numbers and physiology, influencing the fate of decapped mRNAs by directing them to nonsense-mediated decay²⁸⁵. In a separate study, the lncRNA Aw112010 was found to express a SEP essential for mucosal immunity during bacterial infection and colitis. Specifically, the Aw112010 microprotein, but not its encoding lncRNA, regulates IL-12P40 modulating the macrophage response to infection (Figure 12)²⁸⁶.

Recent studies have identified numerous novel SEPs that localize to mitochondria. Interestingly, one study demonstrated that the inner mitochondrial membrane (IMM) is particularly enriched in both canonical and non-canonical microproteins, highlighting it as a hub for SEPs²⁸⁷. By integrating prediction tools and proteomics they successfully identified 16 novel mitochondrial SEPs. Among these, BRAWNIN, a 71aa microprotein encoded by *C12orf73*, emerged as a critical orchestrator of oxidative phosphorylation. Specifically, it is positively regulated by the AMPK pathway and plays a key role in the respiratory chain complex III assembly. Depletion of BRAWNIN affects multiple metabolic pathways and ATP production, underscoring its crucial role in cellular bioenergetics²⁸⁷. The same group further demonstrated that the respiratory chain complex III biogenesis occurs in a stepwise fashion, through the formation of three subcomplexes, and is mediated by BRAWNIN's interaction with other mitochondrial proteins and SEPs²⁸⁸.

Similarly, in a large-scale effort to characterize the human heart's transcriptome, *Van Heesch et. al.* identified several lncRNA-encoded microproteins, many of which localize to mitochondria or their expression positively correlates with the levels of oxidative phosphorylation subunits²⁸⁹. Notably, LINC00116, a lncRNA enriched in skeletal muscle and heart, encodes mitoregulin, a SEP residing in the IMM. Mitoregulin binds to cardiolipin and is crucial for the respiratory chain complexes I and V assembly. Overexpression or knock-down of mitoregulin leads to altered mitochondrial membrane potential, Ca^{2+} retention, and perturbed ROS accumulation (Figure 12)²⁹⁰. Another example is LINC00467, a CRC-related lncRNA that encodes the mitochondrial microprotein ATP synthase-associated peptide (ASAP). ASAP interacts with ATP5A and ATP5C to regulate ATP production, inducing CRC tumorigenesis²⁹¹.

While most studies have shown that translation in uORFs mainly affects the translational activity of downstream CDS, there are a few exceptions where uORFs generate a functional microprotein^{292,293}. Notably, researchers demonstrated that the uORFs residing in the 5'UTR of the *PRL-1* and *PRL-2* gene regulate the translation of the downstream CDS in a magnesium-sensitive manner²⁹². A different study demonstrated that the uORF in the *ASNSD1* gene produces a 96aa microprotein (ASNSD1-uORF) whose expression levels are elevated in MYC-driven childhood medulloblastoma. They discovered that ASNSD1-uORF interacts with multiple components of the prefoldin complex, enhancing cancer cell survival in childhood medulloblastoma²⁹³.

Even though several studies have explored the translational potential of lncRNAs, only a few have examined the interplay between the lncRNA and its encoded microproteins. Many

SEP-related studies introduce frameshift mutation or early stop codons into lncRNA sequences to distinguish whether observed effects are lncRNA-mediated^{281,282}. However, these studies often focus on individual effects or specific tissues, overlooking the potential effects of a given lncRNA in a different system. Conversely, research centered on lncRNAs typically does not consider potential translation, due to the challenges in detecting pervasive translation. For example, HOXB-AS3 encodes a microprotein involved in splicing in CRC tissues, while earlier studies demonstrated that this lncRNA regulates ribosomal RNA transcription in acute myeloid leukemia, ignoring its microprotein production²⁹⁴. Similarly, LINC-PINT has been extensively studied in physiological processes and diseases, but its encoded microprotein has been largely overlooked²⁷⁶. Intriguingly, the LINC00961 locus transcribes a lncRNA that mediates inhibitory effects on angiogenesis, while its translated SEP, SPARR, promotes endothelial network formation (Figure 12). The study further demonstrated that the lncRNA and its encoded microprotein are involved in separate mechanisms to ultimately perform opposite effects in endothelial physiology. On the other hand, there are cases such as UPPERHAND, whose transcription regulates the expression of the neighboring gene *Hand2 in cis*⁷¹, while in the cytosol was found to encode a 134aa ER-localized microprotein²⁸⁹.

2. Aim of the thesis

This research thesis aims to deepen our understanding of the function of the non-coding genome in human biology and carcinogenesis, with a particular focus on long non-coding RNAs (lncRNAs), which are pivotal regulators of essential biological processes. Although lncRNAs are traditionally regarded as non-protein-coding transcripts, emerging studies have revealed that lncRNAs harbor smORFs that encode microproteins. Our goal is to identify lncRNAs associated with CRC and uncover novel cancer-related microproteins encoded by lncORFs.

The thesis is divided into two distinct parts, each addressing these different aspects of lncRNA biology. The first part explores lncRNAs regulated by the Wnt/ β -catenin pathway, a key driver of CRC development. Specifically, we investigate the role of the Wnt-regulated lncRNA lnc-IGSF9 (LINC01133) in CRC physiology. In the second part, we leverage ribosome profiling datasets to identify novel microproteins encoded by lncORFs. Notably, we demonstrate that the EPB41L4A-AS1 lncRNA encodes a mitochondrial protein, and we further explore its potential functional implications in cancer.

The specific aims of our research are:

- i) To identify and characterize novel lncRNAs that are regulated by the Wnt signaling pathway.
- ii) To understand their involvement in CRC progression.
- iii) To investigate the molecular mechanisms underlying the function of these lncRNAs.
- iv) To discover novel microproteins that are translated by lncRNAs.
- v) To characterize the function of these microproteins and examine their impact in carcinogenesis.

3. Materials and Methods

Cell lines

For the first study we used Ls174T, Caco-2, SW480, DLD1, HCT116, and HT-29 cell lines, which are immortalized epithelial cells derived from human colorectal cancers harboring mutations to maintain constitutively active Wnt signaling. For the second study we also utilized the HEK293T cell line which consists of immortalized human embryonic kidney cells. For viral production we used HEK293-LentiX cells, which are engineered HEK293T cells that support high viral protein production.

To identify novel WNT-regulated lncRNAs, derivative cell lines of Ls174T were used that stably express shRNAs targeting β -catenin mRNA (Lsbcac cells). Treatment of these cell lines with 1 μ g/mL doxycycline suppressed the β -catenin/TCF4 transcriptional program, inactivating the Wnt signaling pathway.

All cell lines were tested for mycoplasma contamination prior to experiments using a PCR-based Mycoplasma Detection Kit (abm). Cells were cultured in Dulbecco's Modified Eagle Medium (DMEM; Gibco, Thermo Scientific) supplemented with 10% (v/v) fetal bovine serum (FBS; Gibco, Thermo Scientific), 1% (v/v) penicillin-streptomycin (Thermo Scientific), gentamicin (50 μ g/mL), and amphotericin B (0.25 μ g/mL). Cultures were maintained in flasks or dishes in a humidified incubator at 37°C with 5% CO₂.

Transient and stable transfection

Transient transfections were performed using Polyethylenimine (PEI; PolySciences Inc.), JetPrime (PolyPlus), or Lipofectamine 3000 (Invitrogen) following the manufacturers' protocols. PEI was used for applications regarding Western Blots, immunofluorescence, and qPCR analyses. JetPrime was employed for lentiviral production, while Lipofectamine 3000 was utilized for immunoprecipitation and phenotypic assays. For all transient transfections, a forward transfection strategy was used to minimize cellular toxicity, whereby cells were plated on the first day and transfected on the following day.

Reverse transfections were conducted solely to engineer cell lines for stable overexpression. 10⁶ cells were plated and transfected at the same time with 5 μ g of linearized recombinant plasmids, and positive cell clones were selected after treatment with appropriate antibiotics.

For the *in trans* overexpression of *lnc-IGSF9*, the corresponding mRNA sequence was cloned into the pcDNA4/TO plasmid vector (Invitrogen), under the control of a CMV promoter containing TET operator sequences. Ls174T cells expressing the TET repressor were transfected with pcDNA4.TO-*lnc-IGSF9*. Positive clones were selected by culturing cells in medium containing Blasticidin (10 μ g/mL) and Zeocin (500 μ g/mL). Overexpression of *lnc-IGSF9* was induced by treating the selected clones with doxycycline (1 μ g/mL) for 72 hours.

To evaluate the protein-coding potential of selected lncORFs, their sequences were cloned into the pcDNA4/TO plasmid vector (Invitrogen). Each lncORF comprised the 5'UTR, the lncORF sequence, and a FLAG epitope in the C' terminus. Transient transfections were

performed in HEK293T and HCT116 cells, and analyses were conducted 48–72 hours post-transfection.

Lentiviral transduction

Lentiviral transduction was used to overexpress i) proteins that are involved in lncRNA regulation, ii) sgRNAs that induce the CRISPR-a system, or iii) shRNAs that target the lnc-IGSF9, IGSF9, EPB41L4A-AS1 transcripts for knockdown. For lentivirus production, 4×10^6 HEK293-LentiX cells were seeded into 10 cm² culture dishes pre-coated with 50 µg/mL collagen (Collagen Type I, Rat Tail, Corning). Cells were incubated overnight and the following day, they were co-transfected with a DNA mix containing i) 3.75 µg of a reference plasmid encoding the sequence of interest (e.g., shRNA, sgRNA, or a sequence encoding a protein/lncRNA of interest), ii) 1.32 µg of pMD2.G plasmid (Addgene #12259) encoding the lentiviral envelope proteins, iii) 2.43 µg of psPAX plasmid (Addgene #12260) encoding packaging proteins. Transient transfection was performed using JetPrime reagent (PolyPlus) at a 1:2 ratio (1 µg DNA: 2 µL JetPrime reagent). Eight hours post-transfection, the culture medium was replaced with fresh medium, and cells were incubated for an additional three days. On day 5, the culture medium containing the produced lentiviruses was collected into Falcon tubes and centrifuged at $500 \times g$, 4°C, for 5 minutes to remove cellular debris. The supernatant was filtered through 0.45 µm filters and either used immediately for transductions or stored at -80°C for future use.

When titration was required, 10^5 HEK293T or Ls174T cells were seeded into 12-well culture plates and transduced with serial dilutions of the produced lentivirus. The following day, cells were washed thoroughly with $1 \times$ PBS and analyzed by flow cytometry (FACS) to determine the percentage of cells expressing green fluorescent protein (GFP+) or blue fluorescent protein (BFP+), encoded by the reference plasmid. The viral multiplicity of infection (MOI) was calculated by correlating the percentage of infected cells (determined by FACS) to the viral copy number per cell using the protocol provided by Collecta Inc³.

For lentiviral transduction, 3×10^5 cells per well were seeded into 6-well plates and transduced with either 300 µL of shRNA-overexpressing lentivirus, 300 µL of protein-overexpressing lentivirus, or 500 µL of sgRNA-overexpressing lentivirus (CRISPR/Cas9 system). All transductions were performed in the presence of 8 µg/mL Polybrene to maximize infection efficiency.

For protein overexpression, we cloned the coding DNA sequences in EF.PGK.GFP vector (Addgene #17618), while shRNAs were cloned in the PLB vector (addgene #11619). Viral transduction efficiency was evaluated through visual inspection of GFP-positive cells in a fluorescent microscope.

CRISPR-Cas9 technologies for in cis overexpression and genome editing (KO)

We utilized advances in CRISPR-Cas9 technology to induce endogenous transcription of lnc-IGSF9, and perform knock-out experiments in the promoter region of lnc-IGSF9 gene.

³<https://manuals.collecta.com/lentiviral-construct-packaging-and-transduction/v10a/en/topic/lentiviral-titer-calculation>

To overexpress *Inc-IGSF9* *in cis* at its native locus, we developed a clonal Ls174T cell line stably expressing dCas9 fused to the transcriptional activator VP64. This cell line, named LdC10 (Ls174T cells overexpressing dCas9-VP64 clone 10), also expressed transcriptional activators HSF1 and p65, fused to the bacterial protein MS2. Lentiviral infection of these cells with vectors expressing sgRNAs targeting the promoter of *Inc-IGSF9* resulted in the *in cis* overexpression of *Inc-IGSF9*, as the cooperative recruitment of the dCas9-VP64-sgRNA-MS2-HSF1/p65 complex to its promoter region induced the transcriptional activity of *Inc-IGSF9*.

The design of sgRNAs for activating the transcription of *Inc-IGSF9*, as well as control sgRNAs, was performed using the ZiFiT Targeter tool (parameters: U6 promoter, sgRNA size = 18 nt). To facilitate gene overexpression through CRISPR/Cas9, we developed a new lentiviral vector derived from pU6-sgRNA EF1Alpha-puro-T2A-BFP (Addgene #60955). In this new vector, the sgRNA cloning cassette was replaced with a modified cassette flanked by two MS2-recognized bacterial sequences. These sequences recruit HSF1 and p65, enabling cooperative gene activation.

Silencing of *Inc-IGSF9* expression was achieved by deleting most of its promoter. For this purpose, we used the pX333 expression vector (Addgene #64073), which is specifically designed to express both Cas9 and two sgRNAs targeting regions flanking the desired deletion site. To generate a stable cell line lacking the selected promoter region of ***Inc-IGSF9***, Ls174T cells were transfected (Lipofectamine 3000, Invitrogen) with 1.8 µg of the pX333 vector containing the two sgRNAs and 0.2 µg of pcDNA6TR plasmid, which encodes Blasticidin resistance. After 24 hours, cells were cultured in a medium containing Blasticidin (10 µg/mL) for 72 hours to select successfully transfected cells. Cells with the desired promoter deletion were identified through genomic DNA isolation and genotyping.

All sgRNA sequences used for the silencing or overexpression of *Inc-IGSF9* and the EPB-ORF knock-in are provided in Appendix I.

RNA isolation followed by qPCR

Total RNA was isolated from cultured cells using the TRIzol G Reagent (ITW Reagents), was purified by phenol/chloroform extraction and precipitated using ethanol. For cDNA synthesis, 1 µg of total RNA was reverse transcribed in 20 µL reactions using 500 ng of oligodT primers (NEB) and the MMLV reverse transcriptase (Invitrogen). The resulting cDNA was diluted 1:20 with ddH₂O before use.

Quantitative PCR (qPCR) was performed using 4 µL of diluted cDNA as a template, 1 µM of each primer specific to the gene of interest, and 2x SYBR Green PCR master mix (Applied Biosystems). The qPCR reactions were conducted on either a LightCycler96 system (Roche) or a QuantStudio 5 (Applied Biosystems). All experiments were performed in at least duplicate to ensure reproducibility. The qPCR cycling protocol used was as follows:

- Initial denaturation: 95 °C for 5 minutes
- Amplification: 45 cycles of
 - 95 °C for 20 seconds
 - 58 °C for 20 seconds
 - 72 °C for 30 seconds

- Final extension: 72 °C for 5 minutes

Gene expression quantification was carried out using the relative expression method, normalized to housekeeping genes such as GAPDH and RPLP1, which regulate basic cellular functions. The sequences of all primers used in the experiments are listed in Appendix I and Appendix II .

Whole proteome extraction

7×10^5 of LS174, HEK293T or HCT116 cells were seeded in 100mm dishes for whole proteome extraction. After 24 hours, cells were transfected with 10 µg of DNA per plate. Seventy-two hours post-transfection, cells were rinsed with ice-cold 1X PBS, scraped from the dishes, transferred into tubes, and gently centrifuged to pellet the cells. Approximately 10 volumes of RIPA buffer (50 mM Tris-HCl, pH 8.0; 150 mM NaCl; 1% Nonidet P-40; 0.5% sodium deoxycholate) supplemented with 1x protease inhibitor cocktail (PIC, Roche) was added to the cell pellet in each tube. The suspension was gently rotated at 4°C for 20 minutes to ensure lysis. The lysates were then centrifuged at high speed for 20 minutes at 4°C to remove cellular debris. The resulting supernatant, containing total proteins, was transferred to fresh tubes.

Mitochondrial isolation

Cells were harvested from three 150-mm plates by trypsinization and centrifugation. The resulting cell pellet was washed with ice-cold 1× PBS. The washed pellet was resuspended in 1–1.5 mL of lysis buffer (250 mM mannitol, 5 mM HEPES pH 7.4, 5 mM EGTA) supplemented with protease and phosphatase inhibitors. The suspension was transferred to a Teflon Dounce homogenizer and homogenized gently with 80 strokes on ice. The homogenate was centrifuged twice at $600 \times g$ for 5 minutes at 4°C to remove unbroken cells, nuclei, and cellular debris. A 50 µL aliquot of the clear lysate was reserved as the total crude fraction (T). The remaining lysate was centrifuged at $10,500 \times g$ for 15 minutes at 4°C to isolate the crude mitochondrial fraction (M), which was collected as the pellet and resuspended in 100 µL of lysis buffer. The supernatant from this step was further centrifuged at $100,000 \times g$ for 1 hour at 4°C to separate the endoplasmic reticulum (ER) fraction (pellet) from the cytosolic fraction (C, supernatant). The ER pellet was resuspended in 100–200 µL of lysis buffer. Protein concentrations of the total crude (T), cytosolic (C), ER, and mitochondrial (M) fractions were determined using Bradford assay. Equal amounts of protein (10–20 µg per sample) were loaded onto SDS-polyacrylamide gels for subsequent analysis.

Western blot analysis

Protein lysates were prepared by boiling appropriate volumes of lysates in 1x Laemmli sample buffer (60 mM Tris-HCl, pH 6.8; 2% SDS; 10% glycerol; 0.01% bromophenol blue; 5% β-mercaptoethanol) at 95°C for 5 minutes. Proteins were then separated by SDS-PAGE on a 15% polyacrylamide gel in a 1x running buffer.

Protein transfer was conducted onto a 0.45 µm nitrocellulose membrane using a cold 1x transfer buffer and the Bio-Rad wet transfer apparatus, running at 400 mA for 1 hour and 15

minutes at 4°C. Only for EPB-SEP, protein transfer was performed onto a 0.2 µm PVDF membrane in a dry setup, running at 10mA for 20 minutes. Then, the membrane was blocked with 5% non-fat dry milk in 1x Tris-buffered saline (TBS) containing 0.01% Tween-20 (TBS-T) for 1 hour at room temperature.

The blocked membrane was incubated overnight at 4°C with primary antibodies diluted 1:2000 in blocking buffer: mouse anti-FLAG antibody (M2, Sigma) and mouse anti-GFP antibody (Santa Cruz, sc-101536). The antibodies against mitochondrial markers Mitofilin (Proteintech, 10179, 1:5000), MCU1 (D2Z3B Cell signaling 1:2500) were also used. The following day, the membrane was washed three times with 1X TBS-T and incubated for 1 hour at room temperature with HRP-conjugated the appropriate secondary antibody (Southern Biotech) diluted 1:5000 in blocking buffer. After additional washes with TBS-T, chemiluminescent detection was performed using Immobilon Crescendo Western HRP substrate (Millipore). Images were acquired and analyzed using the Bio-Rad ChemiDoc XRS system with Image Lab 6.1 software for densitometry.

Immunofluorescence

10⁵ cells were plated and transfected with 1µg DNA the following day. After 3 days the cells were fixed with methanol or 4% PFA in cases that we retained endogenous GFP expression. Permeabilization was performed by incubating the cells with 350 µL of 0.1% Triton X-100 in PBS containing 1% bovine serum albumin (BSA) at room temperature (RT) for 10 minutes on a shaker. Blocking was performed by incubating cells with 350 µL of PBS containing 1% BSA at RT for 30–60 minutes. Cells were incubated with 30 µL of primary antibody diluted in PBS containing 0.1% Triton X-100 and 1% BSA for 1 hour at RT. Anti-FLAG (Sigma) antibody was diluted in 1:400 ratio, anti-HSP60 (D8F1 Cell signaling) in 1: 1000, and anti-PNPT1 diluted in 1:;200. Cells were then incubated with 30 µL of secondary antibody diluted in PBS containing 0.1% Triton X-100 and 1% BSA for 1 hour at RT. For this thesis, Alexa Fluor 488, 555, 594 were used in 1:500 dilution. Nuclear staining was performed by incubating the cells with 500 µL of 1:5000 diluted DAPI (1mg/ml) in PBS for 5 minutes in the dark and mounted with Mowiol mounting medium. Between each step, cells were washed three more times with PBS, followed by a 15-minute PBS wash on a shaker. Images were acquired in different z-stacks with the use of a Zeiss Leica microscope with Leica LAS X software. ImageJ (Fiji) was used to process the resulting images.

Single molecule fluorescent in-situ hybridization (smFISH)

The subcellular localization of lnc-IGSF9 was determined using single-molecule fluorescence in situ hybridization (sm-FISH), performed with the QuantiGene ViewRNA Assay kit (Affymetrix) according to the manufacturer's instructions. This method enabled simultaneous visualization of DNA, GAPDH mRNA, and the lncRNA of interest. Ls174T cells expressing a scrambled shRNA and a shRNA targeting β-catenin were seeded onto coverslips pre-coated with 0.01% poly-D-lysine in 24-well culture plates. Cells were cultured in the presence or absence of 1 µg/mL doxycycline for 72 hours. After the 72-hour treatment, cells were fixed in

4% formaldehyde for 30 minutes at room temperature and permeabilized by incubation in QC buffer for 5 minutes. Cells were washed with 1x PBS prior to further processing. Next, the cells were incubated with QS protease (1:4000 dilution in 1x PBS) for 10 minutes at room temperature, followed by another 1x PBS wash. Hybridization was carried out using a mixture of probes (1:100 dilution in QF buffer) targeting GAPDH mRNA or the mRNA of Inc-IGSF9. Hybridization was performed for 3 hours at 40°C. Signal amplification was carried out according to the manufacturer's instructions. Following amplification, the cell nuclei were stained with DAPI (1:100 dilution in 1x PBS) for 5 minutes. Coverslips were mounted cell-side down onto glass slides using Prolong Gold Antifade Reagent. Probes were excited at specific wavelengths depending on the linked fluorophore: Type 1 probes (e.g., GAPDH) were excited at 550 nm and visualized using Cy3 filters. Type 6 probes (e.g., WINTRLINC1 or Inc-IGSF9) were excited at 650 nm and visualized using Cy5 filters. Images were pseudo-colored for analysis, with red corresponding to Type 1 probes and green to Type 6 probes. The resulting images revealed the subcellular localization of the targeted lncRNAs.

3' UTR mRNA Quant-seq

For transcriptome profiling we utilized 3' mRNA QuantSeq protocol²⁹⁵ which generates sequenced reads that cover the 3' end of polyadenylated RNAs. Thus, the QuantSeq protocol is simpler and requires lower sequencing depth, enabling faster analysis. Since each gene is represented by a single fragment in the resulting libraries, gene expression quantification is straightforward, relying on the number of reads mapped to each fragment without requiring complex normalization procedures.

Library preparation for this protocol was performed using the QuantSeq 3' mRNA-Seq Library Prep Kit (Lexogen), following the manufacturer's instructions. Briefly, the process began with the binding of oligodT primers to the poly-A tails located at the 3' ends of mRNAs, enabling the synthesis of first-strand cDNA through reverse transcription. The RNA template was then removed, followed by the synthesis of the second strand using random primers binding along the length of the initial cDNA. The resulting double-stranded cDNA was precipitated using magnetic beads and amplified via PCR.

Subsequent next-generation sequencing was performed using the Ion Proton™ System, adhering to the manufacturer's guidelines. This streamlined approach ensures efficient and accurate generation of sequencing libraries for downstream analysis.

RNA sequencing analysis

Fastq files were mapped to the human genome build hg19 using a two-step alignment process. In the first round, reads were mapped with TopHat2²⁹⁶ using the default parameters and then, the unmapped reads were mapped with bowtie2²⁹⁷ using the --local and --very-sensitive parameters. The resulting bam files were used to estimate the number of raw read counts through metaseqR²⁹⁸, and were normalized with the DESEQ²⁹⁹ algorithm as previously described. Genes with zero reads across all RNA-seq samples were excluded. Further, genes were filtered out based on the following criteria: i) gene length < 500 bp, ii) mean

reads per 100 bp < 25% of the total normalized reads per 100 bp, iii) total read counts below the median of all normalized read counts. Differential gene expression was conducted with the DESeq algorithm. Genes exhibiting a pvalue < 0.05 and a fold change greater than 0.58 or less than -0.58 on the log2 scale (corresponding to a 1.5-fold change in the natural scale) were considered differentially expressed. These analyses were used for the Inc-IGSF9 *in cis* overexpression and the shRNA-mediated knock-down of β -catenin in LS174 cells (Lsbcat).

Additional RNA-seq datasets were meta-analyzed to examine the expression levels of Inc-IGSF9. For the Michels et. al. dataset³⁰⁰ we performed the same initial steps of analysis but differential expression analysis was conducted with DESeq2³⁰¹. Due to intra-patient variability we performed paired analysis for the following comparisons: i) organoids treated with WRN compared to wild-type organoids, and ii) organoids with APC knock-out compared to wild-type organoids. The same thresholds were applied accordingly.

Accordingly, raw RNA-seq reads from 434 colorectal cancer (CRC) patients and 40 adjacent normal tissue samples from the TCGA³⁰² cohort were analyzed.

Immunoprecipitation

2×10^6 HEK293T cells were seeded, and the following day transfected with 10 μ g pcDNA4.TO (negative control), pcDNA4.TO-EPB-SEP-FLAG, and pcDNA4.TO-FLAG-EPB-SEP. After 3 days, cells were washed with 1X PBS and collected by scraping. The cell suspension was centrifuged at 1200 rpm for 5 minutes at 4°C, and the supernatant was discarded. The cell pellet was lysed by adding 240 μ L of RIPA lysis buffer (50 mM Tris-HCl, pH 8.0; 150 mM NaCl; 1% Nonidet P-40; 0.5% sodium deoxycholate) and transferred to a 1.5 mL Eppendorf tube. Samples were rotated for 30 minutes at 4°C to ensure complete lysis. The lysate was centrifuged at 14,000 rpm for 20 minutes at 4°C, and the supernatant was carefully collected and transferred to a new Eppendorf tube.

For the immunoprecipitation, protein G agarose beads (Millipore 16-266) were used after preclearing them with 1X PBS followed by RIPA Buffer. A total of 1mg of the extracted lysate was added to the precleared beads and incubated 4°C for 30 min under rotation. The precleared lysate was collected and added 5 μ g of anti-FLAG (Sigma M2) antibody for 1h under rotation at 4°C. The lysate-antibody complex was transferred to freshly pre-cleared beads and rotated for 3h at 4°C.

Beads were then washed three times with 700 μ L of ice-cold RIPA buffer, centrifuging at 6000 rpm for 2 minutes at 4°C after each wash and discarding the supernatant. To dissociate the immunocomplexes from the beads, 100 μ L of Proteomics lysis Buffer (4% SDS, 100mM HCl PH 7.6, 0.1 M DTT) was added to the beads. The mixture was boiled at 90–100°C for 5–10 minutes, and then centrifuged at 10,000 x g to pellet the beads. The resulting supernatant was transferred to a fresh tube for downstream applications such as Western blotting or mass spectrometry.

LC-MS/MS Analysis (Proteomics)

The eluates from the immunoprecipitation reactions were processed according to the Single-Pot Solid-Phase-enhanced Sample Preparation (Sp3) method of Hughes³⁰³, without acidification and including protein reduction in 100mM DTT and alkylation in 100 mM Iodoacetamide. Digestion was carried out for continuous shaking at 1200rpm at 37°C using 0.5 µg Trypsin/LysC mixture in a 100mM triethylammonium bicarbonate (TEAB) buffer. Next day, the magnetic beads were removed and the peptidic samples were further purified by Sp3 peptide cleanup and evaporated to dryness in a vacuum centrifuge. The dried samples were solubilized in Buffer A, sonicated for 5 minutes and the peptide concentration was determined by measuring the absorbance at 280 nm using a nanodrop.

Samples were run on a liquid chromatography tandem mass spectrometry (LC-MS/MS) setup consisting of a Dionex Ultimate 3000RSLC online with a Thermo Q Exactive HF-X Orbitrap mass spectrometer. Peptidic samples were directly injected and separated on an 25 cm-long analytical C18 column (PepSep, 1.9µm3 beads, 75 µm ID) using an one-hour long run, starting with a gradient of 7% Buffer B (0.1% Formic acid in 80% Acetonitrile) to 35% for 40 min and followed by an increase to 45% in 5 min and a second increase to 99% in 0.5min and then kept constant for equilibration for 14.5min.

A full MS was acquired in profile mode using a Q Exactive HF-X Hybrid Quadrupole-Orbitrap mass spectrometer, operating in the scan range of 375-1400 m/z using 120K resolving power with an AGC of 3x 10⁶ and max IT of 60ms followed by data independent analysis using 8 Th windows (39 loop counts) with 15K resolving power with an AGC of 3x 10⁵ and max IT of 22ms and a normalized collision energy (NCE) of 26. Each biological replica was analyzed in three technical replicas on the system.

Proteomics analysis

Orbitrap raw data was analysed in DIA-NN 1.9.2 (Data-Independent Acquisition by Neural Networks)³⁰⁴ through searching against the reviewed Canonical Human Uniprot database supplemented with the FASTA file of our construct in the library free mode of the software, allowing up to two tryptic missed cleavages. A spectral library was created from the DIA runs and used to reanalyse them. DIA-NN default settings have been used with oxidation of methionine residues and acetylation of the protein N-termini set as variable modifications and carbamidomethylation of cysteine residues as fixed modification. N-terminal methionine excision was also enabled. The match between runs (MBR) feature was used for all analyses and the output (precursor) was filtered at 1% FDR and finally the protein inference was performed on the level of genes using only proteotypic peptides. The generated results were processed statistically and visualized in the Perseus v1.6.15.0 software³⁰⁵.

Correlation analysis between Inc-IGSF9 and IGSF9 expression

Correlation analyses between Inc-IGSF9 and IGSF9 expression were conducted using Pearson's correlation coefficient. Three independent and publicly available datasets were employed for this analysis: (i) CRC cell lines, (ii) normal human tissues, and (iii) clinical CRC patient samples.

For CRC cell lines, correlation analysis of RNA-seq reads (normalized to transcripts per million, TPM) was performed using data from the Sanger Genomics of Drug Sensitivity in Cancer Project (GDSC) RNA-seq cancer cell line gene expression data accessed via the Expression Atlas database. For normal human tissues, correlation analysis utilized TPM values derived from RNA-seq data provided by the GTEx portal³⁰⁶. Pearson's correlation coefficient (R) was calculated separately for each tissue type. For clinical CRC patient samples, correlation analysis was based on TCGA RNA-seq data³⁰², as previously described.

Ribosome profiling analysis

We meta-analyzed publicly available ribosome profiling experiments from HCT116 cells treated with cycloheximide and lactimidomycin³⁰⁷. Adapters AGATCGGAAGAGCACAC and CTGTAGGCACCATCAAT were removed using fastq-clipper. Reads mapping to snRNA, rRNA, and tRNA were filtered out, and the remaining reads were aligned to the human genome (hg38 build) using TopHat2²⁹⁶.

Putatively translated open reading frames (ORFs) were identified using *ORF-RATER* with default parameters, targeting all ORFs initiating with NTG²⁴⁷. Gold-standard ORFs were defined as those with a minimum length of 100 codons. ORFs with an ORF-RATER score greater than 0.8 were considered translated. To further assess translation efficiency, we calculated the Ribosome Release Score (RRS)³⁷ for all high-confidence ORFs using custom scripts. Visual validation of Ribo-seq signals was performed using UCSC Genome Browser, with *bigWig* files generated from downsampled datasets normalized to the smallest library size.

LncORFs were extracted based on their genomic positions within genes annotated as lncRNAs. Only lncORFs identified in both cycloheximide and lactimidomycin datasets were retained for further analysis. Candidate lncORFs were manually curated by examining their Ribo-seq signal and considering their RRS scores.

Flow cytometry analysis (FACs) for cell cycle analysis and TMRE staining

For cell cycle analysis, 7×10^6 HEK293T cells were seeded in 100mm plate and transfected with 5 μ g pcDNA4.TO-EPB-T2A-GFP the following day. 10^6 HEK293T cells were collected and centrifuged at 3000 rpm for 5 min at 4°C. The supernatant was discarded, and the pellet was washed with 1 mL cold 1X PBS, followed by centrifugation under the same conditions. For fixation, the supernatant was removed, and cells were resuspended in 500 μ L cold 1X PBS. To retain GFP expression and distinguish GFP+ from GFP- cells we included an additional fixation step with paraformaldehyde (PFA). Specifically, an equal volume (500 μ L) of 2% PFA was added, and cells were incubated for 1 hour at 4°C. Next, cells were centrifuged (3000 rpm, 5 min, 4°C), washed with 1 mL cold PBS, and centrifuged again. The pellet was resuspended dropwise in 1 mL of 70% ethanol on a vortex mixer and incubated for 30 min at 4°C. Cells were centrifuged (3000 rpm, 20 min, 4°C) and washed with 1 mL cold PBS. The centrifugation and washing steps were repeated, ensuring minimal cell loss during supernatant removal. Cells were stained with 250 μ L of RNase/PI solution (prepared as 50 μ L RNase A, 100 μ g/mL; and 200 μ L propidium iodide (PI), 50 μ g/mL) and incubated at 37°C for 30 min. Samples

were analyzed by flow cytometry within 1 hour with BD FACS Canto II. G1,S, and G2 peaks were visually selected and the frequencies of cell populations within each stage were estimated with Flowjo.

For TMRE staining, 7×10^6 HEK293T cells were seeded in 100mm plate and transfected with $5 \mu\text{g}$ pcDNA4.TO-EPB-T2A-GFP the following day. After 2 days, cells were treated with 10nM TMRE (ThermoFischer) for 30 minutes in the dark and collected for FACS analysis in BD Canto II.

For sorting experiments to perform proteomic analysis, 2×10^6 HEK293T cells were seeded in 150mm plate and transfected with $10 \mu\text{g}$ pcDNA4.TO-EPB-T2A-GFP the following day. After 3 days, cells were collected for sorting in the FACS Aria. At least 2×10^5 sorted cells were collected for each biological replicate and population, followed by centrifugation (3000 rpm, 5 min, 4°C) and removal of supernatant. Proteomics lysis Buffer (4% SDS, 100Mm HCl PH 7.6, 0.1 M DTT) was added to the cell pellets.

Luciferase assay

The sequences of the *Inc-IGSF9* gene promoter and enhancer were amplified with PCR using genomic DNA isolated from Ls174T cells as a template and gene-specific primers (sequences provided in Supplementary Table I). PCR products were separated on 1% agarose gels, extracted, and cloned into the pGEM-T-Easy vector (Promega) using T4 DNA ligase (NEB). Recombinant plasmids were digested with the *EcoRI* enzyme (Minotech) to isolate the cloned sequences, which were then treated with Klenow DNA polymerase (NEB) to generate blunt ends. Subsequently, the sequences were subcloned into the pGL4.10 [luc2, promega] vector (for promoter sequences) or the pGL4.10-ADML [luc2, promega] vector (for enhancer sequences).

For luciferase reporter assays, 1.5×10^4 Ls174T cells were seeded into 24-well plates and transduced with $100 \mu\text{L}$ viral particles containing scrambled shRNA or shRNA targeting β -catenin, in the presence of $8 \mu\text{g/mL}$ polybrene. The following day, cells were washed with $1 \times$ PBS to remove viral particles, and the culture medium was refreshed with DMEM prior to transfection with the reporter plasmid. The following day, cells were transiently transfected with 50 ng of the reporter plasmid using Polyethylenimine (PEI, Polysciences Inc.) Luciferase activity was measured as described, reflecting regulatory activity of the cloned sequences.

Colony formation assays (CFAs)

We performed CFAs to evaluate the clonogenicity of LDC10 cells that overexpress *Inc-IGSF9* in cis, and HEK293T and HCT116 cells with overexpression or knock-down of EPB-SEP. A total of 2×10^3 LDC10 cells were seeded into 6-well culture plates and transduced with $50 \mu\text{L}$ of lentiviral particles in the presence of $8 \mu\text{g/mL}$ polybrene. The following day, cells were washed with $1 \times$ PBS to remove residual virus, and fresh culture medium was added. Cells were incubated for approximately two weeks and stained with Coomassie Blue.

For HEK293T and HCT116 cells, 1×10^3 cells were seeded and transduced with shRNAs targeting EPB-SEP. Additionally, 1×10^3 GFP-positive (GFP⁺) and GFP-negative (GFP⁻)

HEK293T cells were collected via fluorescence-activated cell sorting (FACS) after transfection with either pcDNA4.TO-EPB-T2A-GFP or pcDNA4.TO-GFP (used as a negative control).

RNA pulldown

Dynabeads M-280 Streptavidin (10003D, Thermo Fisher Scientific) were washed once with 1 mL 1×BW buffer (2× stock: 10 mM Tris-HCl, pH 7.5, 1 mM EDTA, 2 M NaCl) with rotation, followed by magnetic separation for 2 min. Beads were subsequently incubated twice with 500 µL Solution A (0.1 M NaOH, 0.05 M NaCl; both DEPC-treated) for 5 min each, with magnetic separation between washes, and finally resuspended in 0.5 mL of Solution B (0.1 M NaCl, DEPC-treated) for 5 min.

Synthetic single-stranded Inc-IGSF9 (500 pmol) was mixed with an equimolar complementary biotinylated oligonucleotide (Integrated DNA Technologies) in hybridization buffer (10 mM Tris-HCl, pH 7.5, 50 mM KCl) to a total volume of 200–400 µL. The mixture was denatured at 95°C for 3 min and then slowly cooled to room temperature. The beads were resuspended in an equal volume of 2× BW buffer and incubated with the RNA–oligonucleotide mixture, rotating at room temperature for 30 min. Beads were collected using a magnet for 3 min, washed three times with 500 µL of 1× BW buffer, and the RNA–oligonucleotide complexes were eluted by boiling in 0.1% SDS for 5 min.

LDC10 cells were harvested after three days of CRISPR-a-mediated Inc-IGSF9 overexpression by scraping (as described before). The cell pellet was resuspended in Buffer A (25mM Hepes pH 7.9, 1.5mM MgCl₂, 10mM KCl, 0.1% NP40 and protease inhibitors), centrifuged in 3000 rpm for 5 minutes, and resuspended in RIPA buffer (50mM Tris-HCl pH 7.5, 1% NP40, 0.25% Na-Deoxycholate, 150mM NaCl, 1mM EDTA pH 8.0, 10% Glycerol). After 30 minutes rotation and centrifugation at 14000 rpm for 15 minutes, the supernatant was precleared with Dynabeads after 1 hour rotation and washed twice with RIPA buffer. The precleared extracts were incubated with Dynabeads–RNA complexes for 1 h at 4°C in the presence of RNase inhibitor, washed four times with RIPA buffer and eluted using the Wash buffer supplemented with 1M NaCl after boiling for 5 min in 0.1% SDS.

Xenographs in RAG1^{-/-} immunocompromised mice

To evaluate the role of Inc-IGSF9 in the tumorigenic capacity of Ls174T cells in vivo, xenograft experiments were conducted in immunocompromised RAG1^{-/-} mice. Cell preparations for injection were performed as follows: Four T175 flasks, each containing 10×10⁶ LDC10 cells, were infected with 10 mL of lentiviral particles expressing either Inc-IGSF9-inducing sgRNA or control sgRNA, in a total culture volume of 20 mL, supplemented with 8 µg/mL polybrene. The following day, cells were washed twice with 1× PBS to remove residual virus, and the culture medium was refreshed. After 48 hours, cells from both conditions were harvested, counted, and pelleted by centrifugation at 1500 rpm for 5 min at 4°C. Subsequently, 20×10⁶ cells per condition were resuspended in 2250 µL DMEM and mixed with 750 µL Matrigel (final volume = 3 mL).

A total of 10 male RAG1^{-/-} C57BL/6 mice, three months old, were used in the experiment. Five mice were injected subcutaneously with 250 µL of the control cell/Matrigel suspension on both the left and right flanks, while the other five mice received 250 µL of the Inc-IGSF9-overexpressing cell/Matrigel suspension in the same manner. After three weeks, mice were euthanized, tumors were excised, and tumor dimensions were measured using a manual caliper. Tumor volume was calculated using the formula:

$$\text{Tumor volume} = (\text{width})^2 \times \text{length} / 2$$

Tumors derived from both control and Inc-IGSF9-overexpressing cells were homogenized in TRI Reagent, and total RNA was extracted for RT-qPCR analysis to assess the expression levels of genes of interest. Primer sequences used in these experiments are listed in Supplementary Table I

RNA pol II ChIP

Cells were cultured in 150-mm dishes to 40–50% confluence. For each condition, chromatin was crosslinked with 1.1% formaldehyde in PBS at room temperature for 30 minutes (10 minutes for histones) and quenched by adding 1 M glycine to a final concentration of 0.125 M. Cells were washed twice with ice-cold PBS, scraped, and collected by centrifugation at 1200 rpm for 5 minutes at 4°C. Pellets were sequentially lysed in Buffer A (10 mM Tris-HCl, pH 8.0, 10 mM NaCl, 0.2% NP-40) and Buffer B (10 mM Tris-HCl, pH 8.0, 200 mM NaCl, 1 mM EDTA, 0.5 mM EGTA), each with protease inhibitors, followed by rotation at 4°C for 10 minutes and centrifugation. Nuclei were resuspended in Buffer C (10 mM Tris-HCl, pH 8.0, 100 mM NaCl, 1 mM EDTA, 0.5 mM EGTA, 0.1% Na-deoxycholate, 0.5% N-lauroylsarcosine) and sonicated (Covaris) to shear chromatin into ~500 bp fragments. Debris was removed by centrifugation at 14,000 rpm for 10 minutes at 4°C, and the supernatant was collected. The final chromatin was adjusted to 0.15% SDS using ChIP Incubation Buffer (20 mM Tris-HCl, pH 8.0, 150 mM NaCl, 2 mM EDTA, 0.15% SDS, 0.1% BSA).

Protein G beads were blocked by incubating with ChIP Incubation Buffer containing 0.15% SDS and 0.1% BSA at 4°C for 2 hours with rotation. Immunoprecipitation was performed overnight at 4°C with antibody against RNA pol II (abcam 5095), protease inhibitors, and ~50 µL of packed, blocked beads per reaction.

Beads were washed sequentially with four cold buffers for six times in total: twice with Buffer 1 (20 mM Tris-HCl, pH 8.0, 150 mM NaCl, 2 mM EDTA, 1% Triton X-100, 0.1% SDS), once with Buffer 2 (20 mM Tris-HCl, pH 8.0, 500 mM NaCl, 2 mM EDTA, 1% Triton X-100, 0.1% SDS), once with Buffer 3 (10 mM Tris-HCl, pH 8.0, 250 mM LiCl, 1% NP-40, 1% Na-deoxycholate, 1 mM EDTA), and twice with Buffer 4 (10 mM Tris-HCl, pH 8.0, 1 mM EDTA). Each wash involved rotation for 5 minutes at 4°C followed by centrifugation at 6000 rpm for 2 minutes.

Chromatin was eluted in Elution Buffer (1% SDS, 0.1 M NaHCO₃) at room temperature for 30 minutes with rotation. Crosslinks were reversed overnight at 65°C by adding 5 M NaCl to a final concentration of 200 mM. Input chromatin was processed similarly. DNA was extracted by phenol-chloroform extraction, followed by precipitation with 3 M sodium acetate (final 0.3 M),

10 μ L glycogen (1 μ g/ μ L), and 2.5 $\times$ volume of ethanol at -80°C for 30 minutes. Pellets were washed with 70% ethanol, air-dried, and resuspended in water for downstream qPCR analysis.

Graphs and statistical analysis

GraphPad Prism 8 software was used for statistical analyses and graphs generation. Significances between two groups were analyzed using two-tailed unpaired Student's t-tests, while statistical differences between more than two groups were conducted with ANOVA. Plots regarding RNA-seq and proteomic experiments were generated with Bioconductor and ggplot2.

4. Results

Part I: *lnc-IGSF9* is a novel Wnt-regulated lncRNA that promotes cell differentiation in CRC cells

***lnc-IGSF9* is a target of the β -catenin/TCF4 complex and is negatively regulated by the Wnt pathway.**

The discovery of *WNTRLINC1* as a critical target of the β -catenin/TCF4 complex unveiled an intriguing mechanism driven by a lncRNA, which promotes intestinal stemness and tumorigenesis²²⁸. This finding prompted us to investigate whether other lncRNAs are regulated by the Wnt pathway and contribute to CRC pathogenesis. To explore this, we utilized LS174 cells, a CRC cell line harboring APC mutations that lead to constitutively nuclear accumulation of β -catenin in the nucleus and sustained activation of the Wnt pathway. We analyzed ChIP-seq data using antibodies against β -catenin and TCF4 to map their genome-wide binding sites. In parallel, ChIP-seq against RNA pol II was conducted in LS174 cells engineered to stably overexpress an shRNA against β -catenin (*Lsbcat*) in a doxycycline-dependent manner¹⁸⁹. Thus, doxycycline treatment results in robust suppression of the Wnt-driven transcriptional program.

We observed that the β -catenin/TCF4 complex binds to a putative intronic enhancer of *linc01133*, located near the *IGSF9* gene, which we named *lnc-IGSF9* (Figure 13A). Interestingly, knockdown of β -catenin led to increased RNA polymerase II occupancy at the *lnc-IGSF9* locus (Figure 13A). To further characterize the Wnt-dependent transcriptional landscape, we conducted RNA-seq in *Lsbcat* cells treated with doxycycline or control cells. Notably, *lnc-IGSF9* expression was significantly upregulated upon doxycycline treatment (Figure 13B), which was subsequently validated by qPCR (Figure 13C). This effect was consistent across independent experiments using lentiviral transduction of an shRNA against β -catenin in HT29 and DLD1 cells (Figure 13D). Moreover, luciferase assays revealed that β -catenin knockdown increased the activity of the *lnc-IGSF9* promoter, but not its enhancer, suggesting that Wnt signaling regulates *lnc-IGSF9* transcription through promoter-dependent mechanisms, at least in episomal assays (Figure 13E). Our results show that suppression of the Wnt pathway activates *lnc-IGSF9* expression.

Next, we examined whether Wnt pathway activation leads to opposite effects on *lnc-IGSF9* expression. Specifically, we meta-analyzed publicly available RNA-seq datasets from patient-derived organoids that were either treated with WRN/R-spondin medium (WRN) to extrinsically activate the Wnt pathway or were genetically engineered to carry APC mutations (APCKO) that stabilize β -catenin expression, activating the pathway in an intrinsic manner³⁰⁰. As expected, *lnc-IGSF9* was significantly downregulated in both conditions of Wnt activation in two out of three donors (Figures 14A and 14B). To further validate it, we treated HT29 cells, which are characterized by mild or absent Wnt activity, with CHIR99021, a GSK3 inhibitor that prevents β -catenin degradation. Indeed, CHIR99021 treatment resulted in a pronounced

reduction of *Inc-IGSF9* expression (Figure 14C). Collectively, our results suggest that *Inc-IGSF9* represents a rare case of a gene directly negatively regulated by the Wnt/ β -catenin signaling pathway.

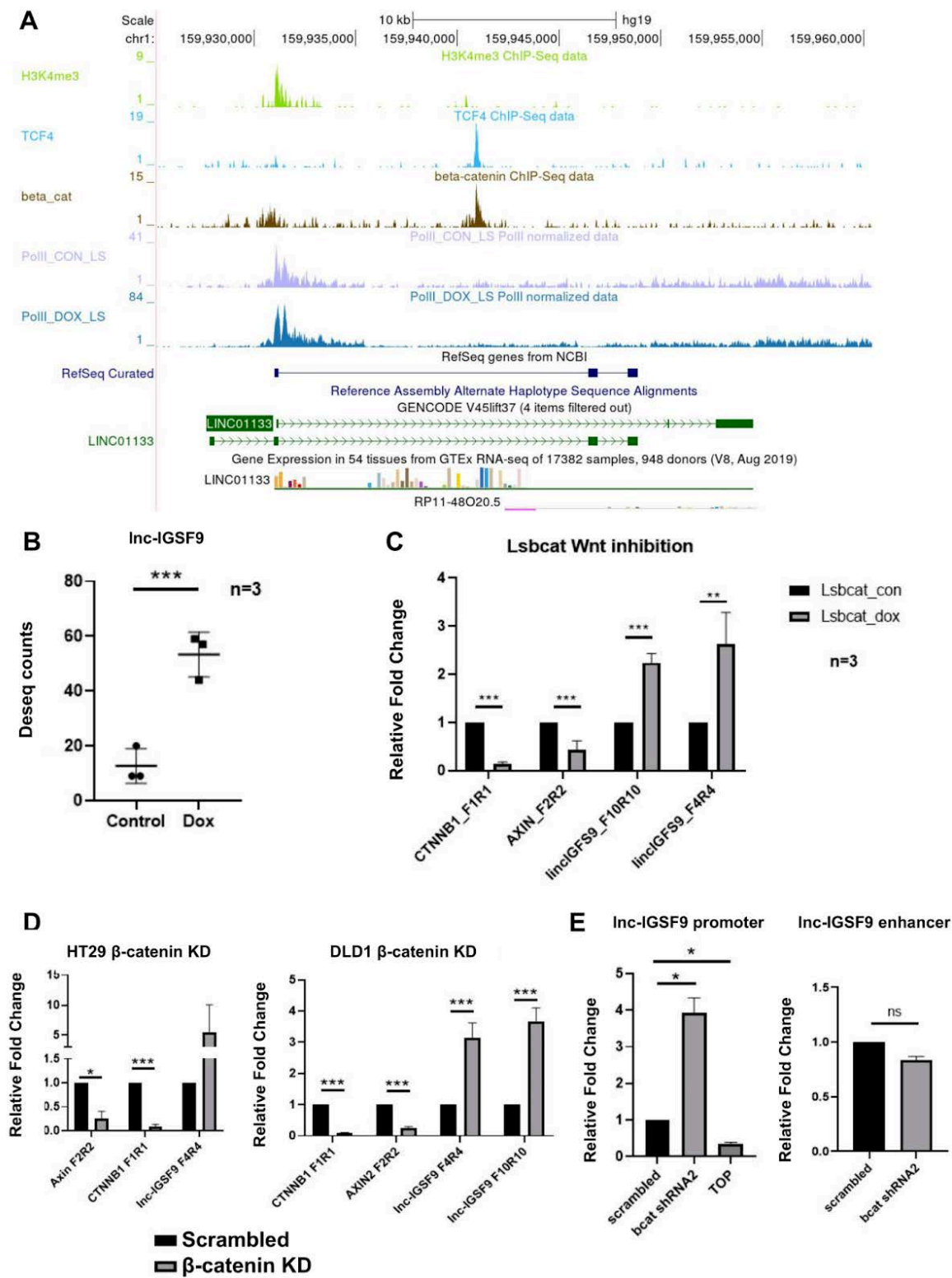


Figure 13: The β -catenin/TCF4 complex binds on an intronic enhancer of *Inc-IGSF9*, negatively regulating its expression in LS174 cells.

A. Snapshot illustrating the ChIP-seq reads in the locus of *Inc-IGSF9* (LINC01133). Antibodies against H3K4me3, TCF4, and β -catenin were used to mark active promoters and map β -catenin/TCF4 binding sites accordingly, while antibodies against RNA pol II were used to assess transcriptional activity. B. *Inc-IGSF9* deseq counts derived from RNA-seq experiments in Lsbcats cells with or without the overexpression of shRNAs against β -catenin. C. qPCR analysis of Lsbcats cells after 72h doxycycline treatment. D. qPCR analysis of *Inc-IGSF9* expression in HT29 and DLD1 cells transduced with shRNA targeting β -catenin. CTNNB1 and AXIN2 serve as control genes that are positively regulated by the Wnt pathway. E. Luciferase assays demonstrating the transcription activity of *Inc-IGSF9* promoter and putative enhancer upon lentiviral-mediated knock-down of β -catenin.

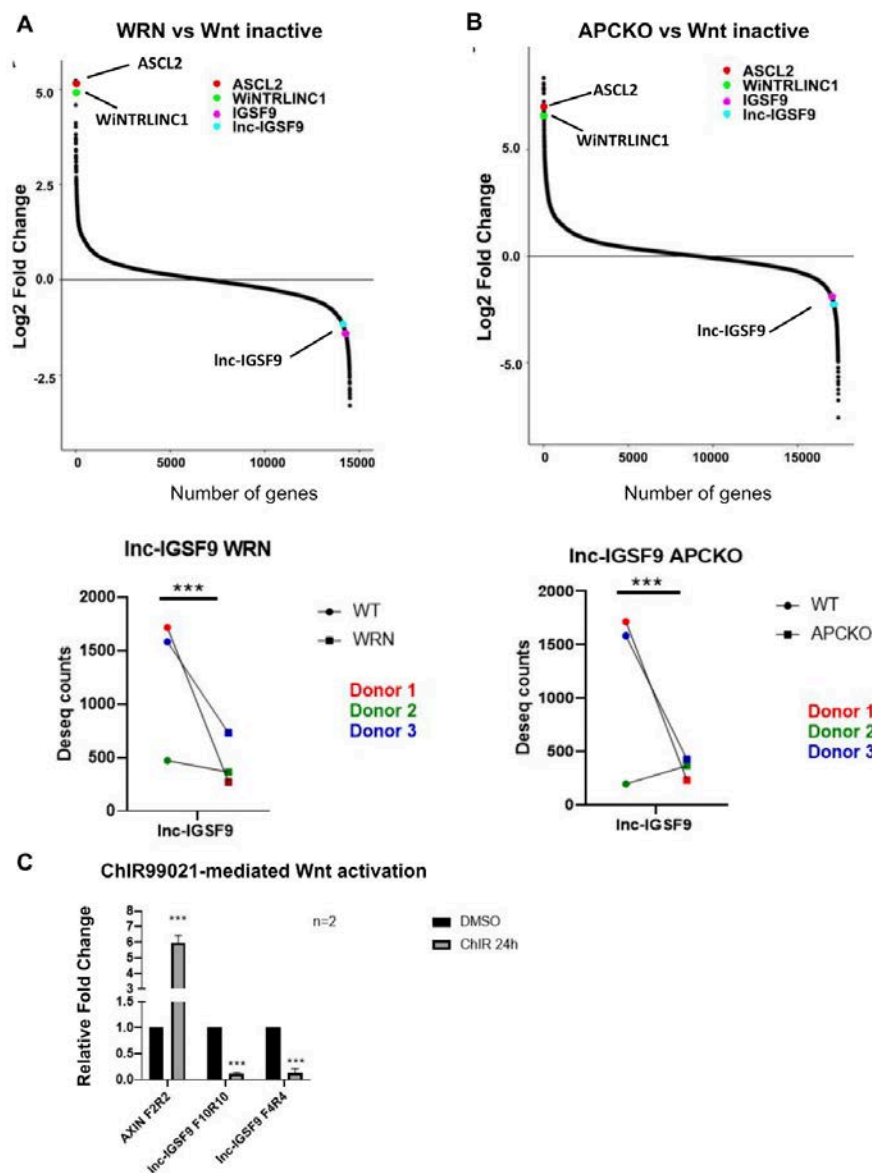


Figure 14: Wnt pathway activation represses *lncIGSF9* expression in CRC patient-derived organoids and cell lines.

A. Scatter plot depicting the regulation pattern of all genes, ranked based on fold change, in CRC patient-derived organoids that were treated or not with WRN/R-spondin medium to extrinsically activate the Wnt pathway. B. Scatter plot depicting the regulation pattern of all genes, ranked based on fold change, in CRC patient-derived organoids that were genetically modified to carry APC mutations to intrinsically activate the Wnt pathway. *lnc-IGSF9* and *IGSF9* are colored with purple and light blue, respectively, while *ASCL2* and *WNT1* serve as controls of positively regulated genes upon Wnt activation²⁵, and are colored with red and green respectively. Below each scatter plot, the means of *deseq* counts for *lnc-IGSF9* is presented in points and the line showcases the pattern of regulation, while the color of the line corresponds to a different donor. C. qPCR analysis for *lnc-IGSF9* and *AXIN2* (serving as positive control) in HT29 cells treated with ChIR.

***Lnc-IGSF9* is a bona-fide nuclear lncRNA with no protein-coding potential**

Next, we evaluated the translation potential and intra-cellular localization of *lnc-IGSF9*. PhyloCSF¹⁸ and InterProScan5³⁰⁸ analysis revealed that the transcript exhibits low protein coding potential, with scores comparable to those of well-established lncRNAs (Figure 15A & 15B) and markedly lower than *bona-fide* protein coding genes. *In vitro* transcription/translation assays followed by coomassie staining and autoradiography confirmed that *lnc-IGSF9* does not appear to encode any protein product (Figure 15C). Also, visual interrogation of ribosome profiling experiments from the colorectal cancer cell line HCT116 did not provide evidence of ribosomal binding to *lnc-IGSF9* (data not provided).

Subcellular fractionation followed by qPCR revealed that *lnc-IGSF9* is predominantly associated with chromatin in LS174 cells, in contrast with the cytoplasmic localization that has been reported in other tissues^{123,309}. FISH experiments using probes that target *lnc-IGSF9* confirmed its localization within the nucleus (Figure 15E). Our results corroborate that *lnc-IGSF9* is a nuclear lncRNA that does not encode any protein.

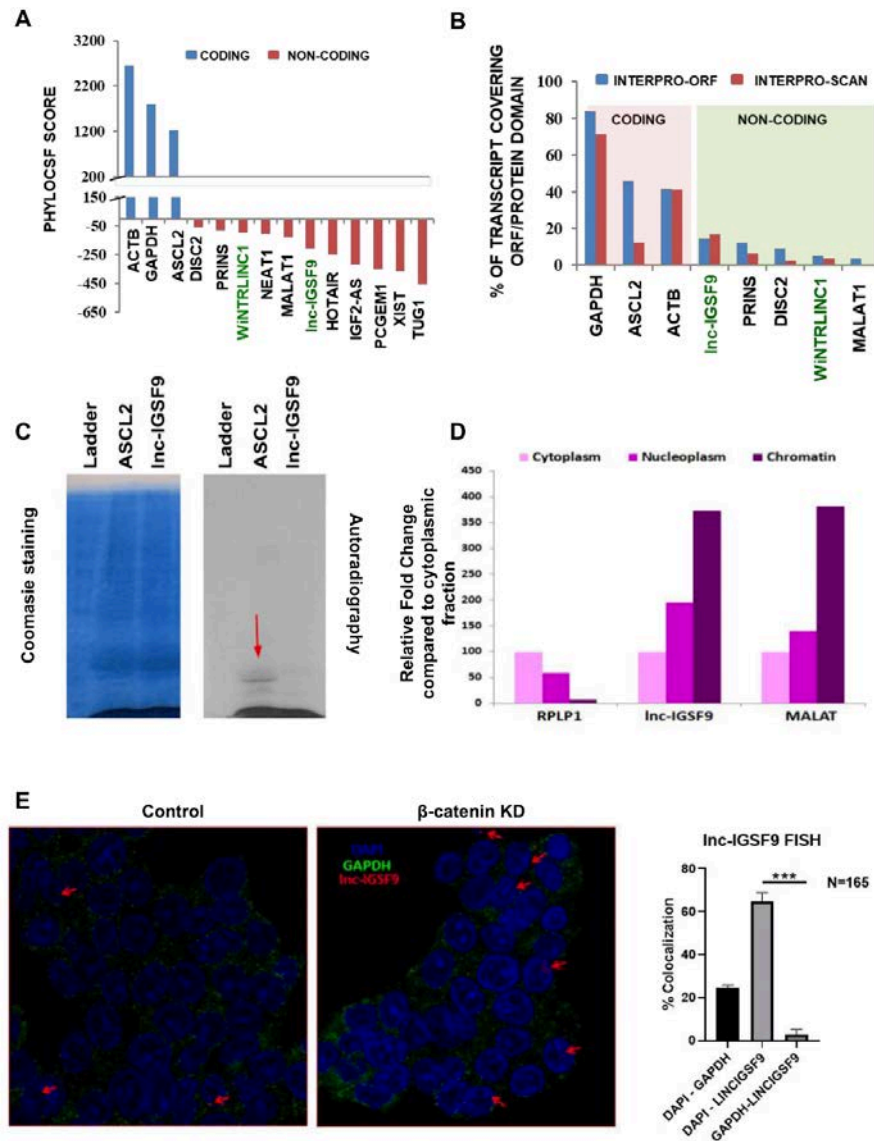


Figure 15: *Inc-IGSF9* is a bona-fide nuclear *lncRNA* with no transcriptional activity.

A. PhyloCSF analysis of the coding potential of *Inc-IGSF9* compared to that of coding (*ACTB*, *GAPDH*, *ASCL2*) and non-coding (*DISC2*, *PRINS*, *WINTRLINC1*, *NEAT1*, *MALAT1*, *HOTAIR*, *IGF2-AS*, *PCGEM1*, *XIST*, *TUG1*) transcripts. B. InterProScan ORF and InterProScan protein domain analysis assessing the coding potential of *Inc-IGSF9* compared to that of coding (*ACTB*, *GAPDH*, *ASCL2*) and non-coding (*DISC2*, *PRINS*, *NEAT1*, *WINTRLINC1*, *MALAT1*) transcripts. C. Coomassie staining and autoradiography of in vitro transcription/translation products of *Inc-IGSF9*, with *ASCL2* serving as positive control. D. Subcellular fractionation of LS174 cells followed by qPCR to analyze the distribution of *Inc-IGSF9* across cytoplasmic, nucleoplasmic, and chromatin fractions. Results are shown as fold change relative to the cytoplasmic fraction, with *RPLP1* serving as a cytoplasmic marker and *MALAT1* as a chromatin marker. E. FISH experiments showing the localization of *Inc-IGSF9* (red staining) in LS174 cells treated with or without shRNAs against β -catenin. *GAPDH* served as cytoplasmic marker, while nuclei were stained with DAPI. The percentage of colocalization between DAPI-GAPDH, DAPI-*Inc-IGSF9*, and *GAPDH-Inc-IGSF9* was quantified in 165 cells.

***Lnc-IGSF9* exerts tumor suppressive action in CRC**

To investigate the role of *Lnc-IGSF9* in CRC physiology, we overexpressed *Lnc-IGSF9* in LS174 cells. Using the Tet Repressor system, we established LS174 cells that inducibly overexpress the *Lnc-IGSF9* transcript as described before. We observed robust overexpression of *Lnc-IGSF9* after doxycycline treatment (Figure 16A & 16B), however no significant differences in CRC clonogenicity were detected (Figure 16C). Notably, overexpression predominantly resulted in increased levels of *Lnc-IGSF9* in the cytoplasm, where it is not normally expressed according to our previous findings (Figure 16B). In contrast, the increase of *Lnc-IGSF9* expression levels in the chromatin fraction was minimal.

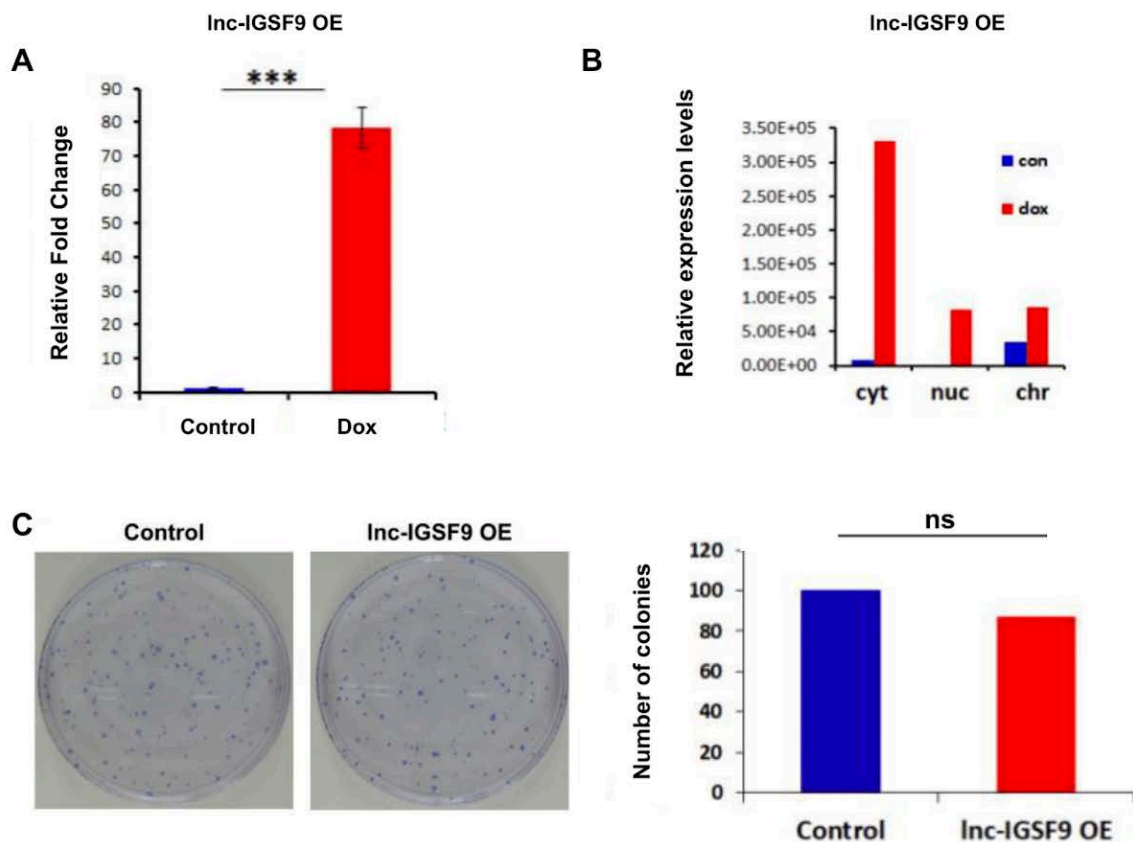


Figure 16: In trans overexpression of *Lnc-IGSF9* does not affect the clonogenicity of CRC cells.

A. qPCR analysis showing *Lnc-IGSF9* expression levels in LS174 cells stably overexpressing *pcDNA4.TO-Lnc-IGSF9* in a doxycycline-dependent manner. B. Subcellular fractionation followed by qPCR for *Lnc-IGSF9* distribution in LS174 cells that overexpress *pcDNA4.TO-Lnc-IGSF9*, with or without doxycycline treatment. C. CFAs of LS174 cells stably overexpressing *pcDNA4.TO-Lnc-IGSF9*, with or without doxycycline treatment.

Given that nuclear lncRNAs often regulate gene expression *in cis*, we hypothesized that the artificial cytoplasmic accumulation of *Lnc-IGSF9* might be irrelevant to CRC physiology. Therefore, we employed the CRISPR-a system, using two distinct sgRNAs targeting the

lnc-IGSF9 promoter compared to a non-targeting guide (NTG) that serves as negative control. These sgRNAs guide the dCas9-VP64-p65-HSF1 complex to the *lnc-IGSF9* promoter to induce transcriptional activation at the endogenous *lnc-IGSF9* locus (Figure 17A).

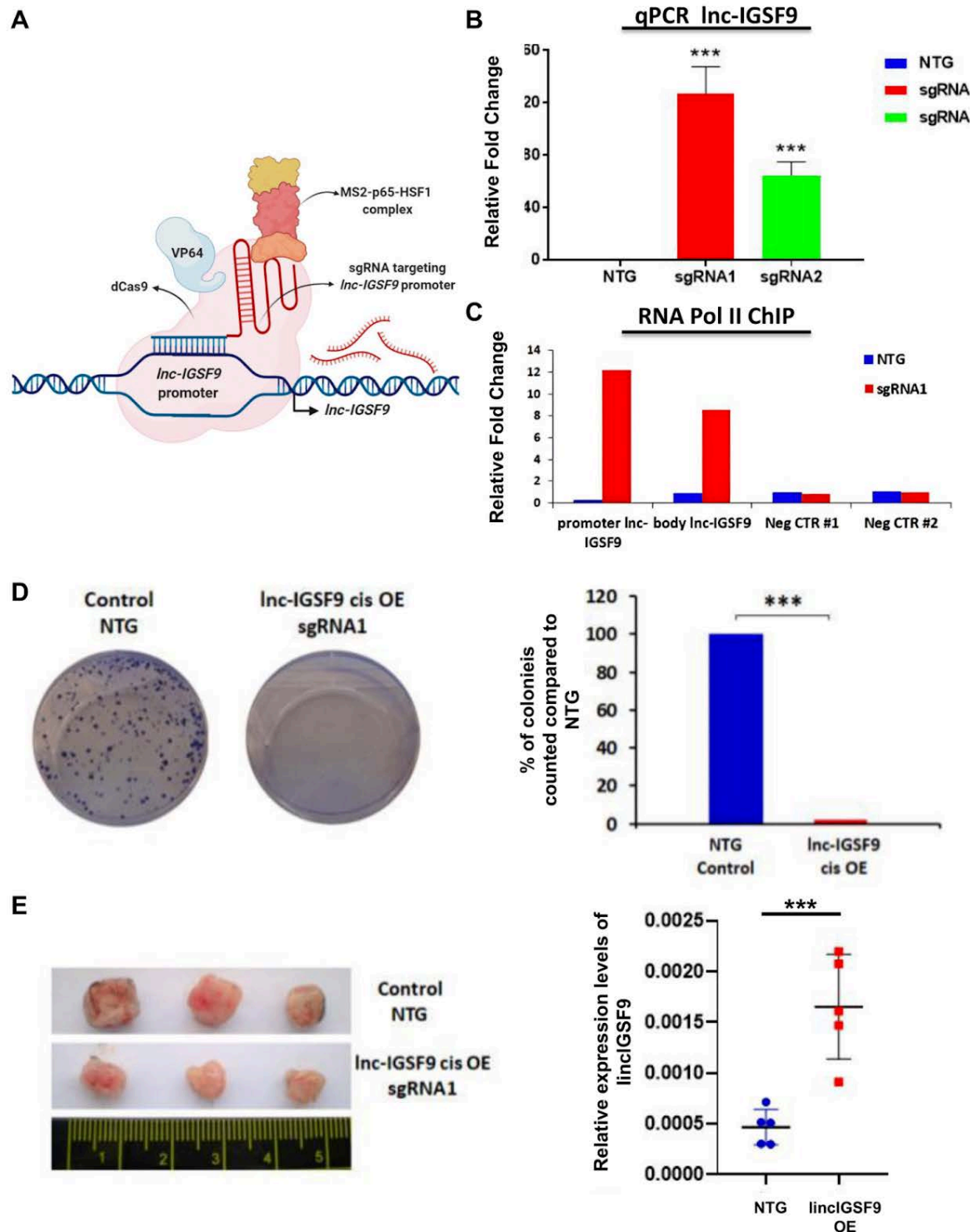


Figure 17: In cis overexpression of lincIGSF9 reduces CRC cell clonogenicity and proliferation.

A. Graphical representation of the CRISPR-a system that induces the transcriptional activity of *lnc-IGSF9*. A sgRNA guides the dCas9 fused to the transactivators VP64, p65, and HSF1 to the *lnc-IGSF9* promoter, augmenting the endogenous transcription of the locus. B. qPCR analysis showing *lnc-IGSF9* expression levels in LS174 cells after CRISPR-a activation using two distinct sgRNAs. C. ChIP followed by qPCR against RNA pol II to evaluate its occupancy at the *lnc-IGSF9* promoter and gene body after in cis overexpression with sgRNA1. Negative controls (Neg CTR #1 and CTR #2) are two random sequences. D. CFAs of LS174 cells overexpressing sgRNA1 or a non-targeting guide (NTG) through the CRISPR-a system. E. Estimation of the size of the tumors that developed following the subcutaneous injection of control cells or cells overexpressing *lnc-IGSF9* in cis in RAG1^{-/-} immunocompromised mice, as well as the expression levels of *lnc-IGSF9* in these tumors. The animals were sacrificed 3 weeks after the introduction of the cancer cells.

Indeed, *in cis* overexpression of *lnc-IGSF9* using two separate sgRNAs resulted in robust overexpression of *lnc-IGSF9* (Figure 17B). To confirm that the overexpression was driven by increased transcriptional activity of the endogenous gene, we performed ChIP against RNA pol II followed by qPCR. Our results demonstrated elevated RNA pol II occupancy at both the *lnc-IGSF9* promoter and gene body, suggesting that *lnc-IGSF9* is overexpressed from its own locus (Figure 17C). Unlike ectopic overexpression, the CRISPR-a-mediated *in cis* overexpression of *lnc-IGSF9* led to decreased cell clonogenicity. Furthermore, subcutaneous injection of NTG- and sgRNA-expressing cells into immunocompromised mice demonstrated that tumors derived from cells overexpressing *lnc-IGSF9* were smaller. Our findings indicate that *lnc-IGSF9* functions as a tumor suppressor, likely through regulating genes in close proximity to its genomic locus.

To elucidate the role of *lnc-IGSF9* in CRC cells, we employed 3'-mRNA-seq (Quant-seq) on LS174 cells expressing either of two sgRNAs targeting the *lnc-IGSF9* locus or the NTG. Transcriptome profiling confirmed a robust increase in *lnc-IGSF9* expression levels (Figure 18A), accompanied by the significant differential expression of hundreds of genes (Figure 18B). Specifically, 928 genes were significantly upregulated and 721 genes were downregulated in LS174 cells expressing sgRNA1 (Figure 18B). Similarly, in LS174 cells expressing sgRNA2, 1257 genes were upregulated, while 1095 genes were downregulated (data not shown). Notably, 871 were differentially expressed in both conditions, with 857 following a consistent pattern of de-regulation (Figure 18C). Functional enrichment analysis of these commonly deregulated genes revealed enrichment in pathways related to extracellular matrix remodeling, cell-matrix interactions, MET signaling, and lipid and cholesterol metabolism (Figure 18D), suggesting that *lnc-IGSF9* plays a role in processes affecting cell adhesion and signaling.

To further explore the function of *lnc-IGSF9*, we compared its deregulated gene set to genes differentially expressed upon β -catenin knockdown. We found 238 overlapping genes (Figure 19A), indicating that *lnc-IGSF9* likely affects a subset of the Wnt-dependent transcriptional repertoire. While pathview analysis revealed that *lnc-IGSF9* has a modest effect on Wnt pathway components, it likely represses the Wnt-mediated activation of the FRD/LRP complex by downregulating the expression of both Wnt activators and FZD genes (Figure 19B). At the same time, *lnc-IGSF9* overexpression results in a dramatic increase of DKK1 (Figure 18B & 19B) which inhibits the Wnt pathway by sequestering the Wnt ligands. Although *lnc-IGSF9*

suppresses Wnt activation by modulating the expression of Wnt ligands, FZD, and DKK1, the LS174 cell line harbors mutations in β -catenin that maintain the Wnt pathway constitutively active in an intrinsic manner, rendering it tolerant to perturbations occurring upstream in the Wnt signaling cascade.

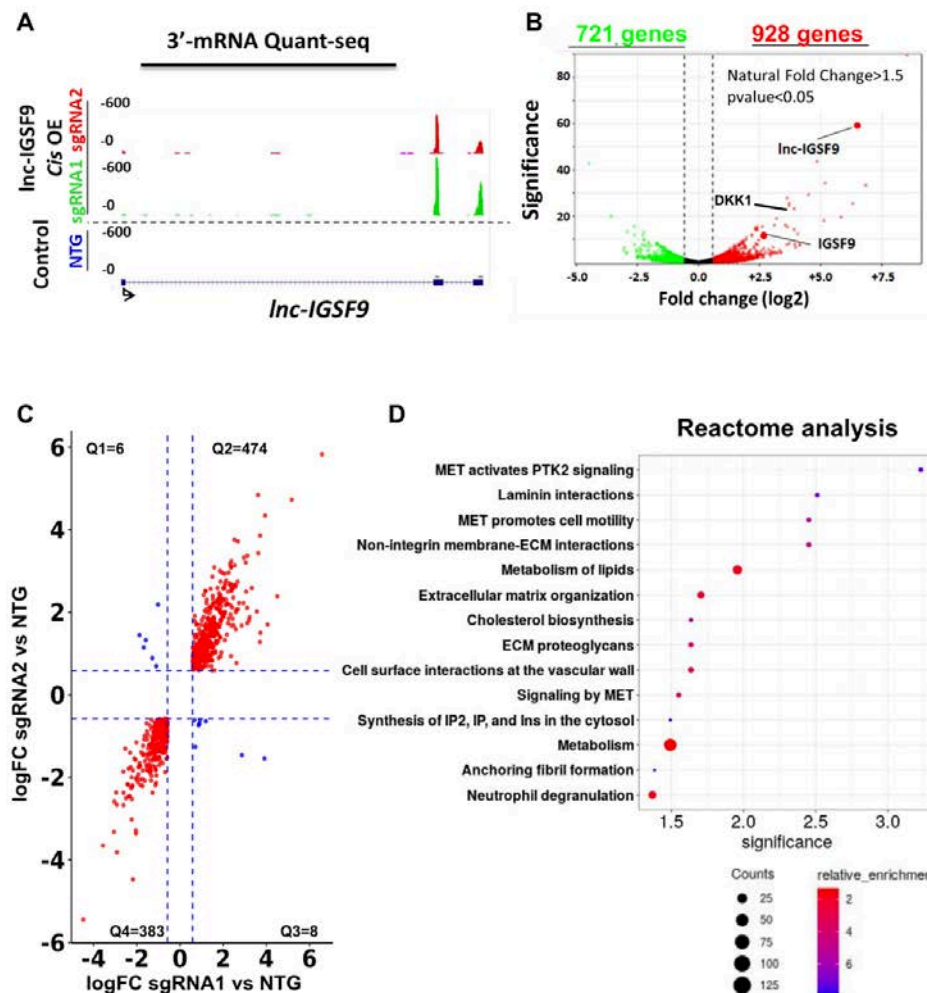


Figure 18: *In cis* overexpression of *Inc-IGSF9* de-regulates cell adhesion and differentiation-related transcripts.

A. Snapshot showing *Inc-IGSF9* expression levels in LS174 cells transduced with NTG (control), sgRNA1, and sgRNA2. **B.** Volcano plot depicting the differentially expressed genes in LS174 cells overexpressing *Inc-IGSF9* in cis (sgRNA1) compared to NTG-expressing cells. **C.** Scatter plot showing the deregulation patterns of genes commonly deregulated in cells overexpressing *Inc-IGSF9* with sgRNA1 or sgRNA2. **D.** Dotplot demonstrating the results of the gene ontology analysis of genes commonly deregulated in comparisons of sgRNA2 vs NTG and sgRNA1 vs NTG using the reactome database via GeneCodis³¹⁰. Dot size reflects the number of genes included in each enriched pathway, while the coloring represents the relative enrichment to the total number of genes associated with each GO term.

Lnc-IGSF9 regulates IGSF9 expression to promote cell differentiation

Based on our previous phenotypic observations, we speculated that *lnc-IGSF9* operates *in cis* to affect local gene expression. Therefore, we examined the set of genes deregulated by sgRNA-mediated *lnc-IGSF9* overexpression that were located within a 1 Mb genomic window surrounding the *lnc-IGSF9* locus. Among these, only *IGSF9* emerged as significantly affected by *lnc-IGSF9* overexpression (Figure 20A) which was further validated through qPCR (Figure 20B). *IGSF9* is located 16kb upstream of the *lnc-IGSF9* locus and encodes a 1179aa protein of the immunoglobulin superfamily. IGSF9 is a cell adhesion molecule which has been implicated in axon guidance, synapse differentiation, and EMT in cancer^{311–313}.

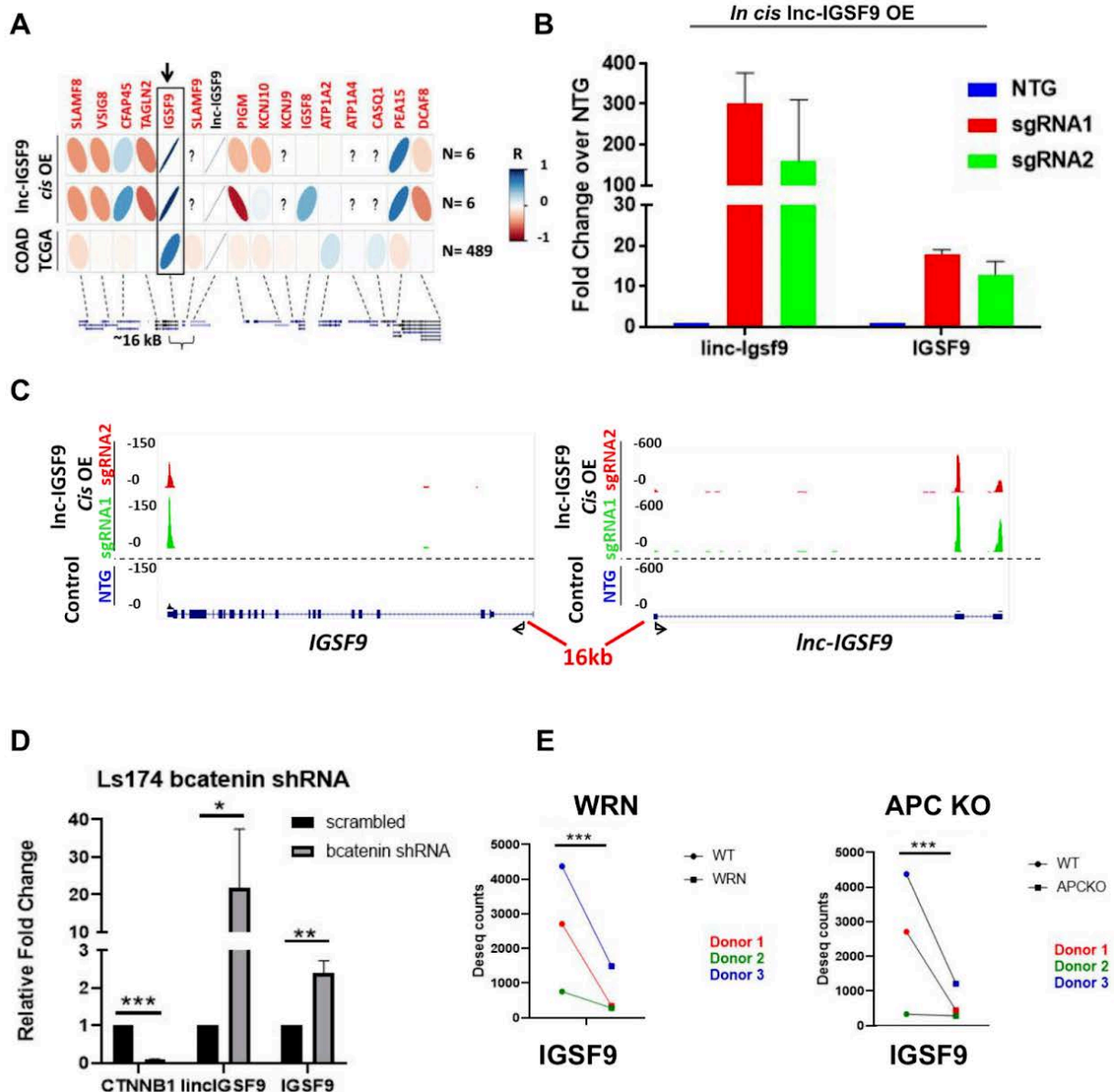


Figure 20: In cis overexpression of *Inc-IGSF9* leads to the upregulation of the neighboring gene *IGSF9*.

A. Correlogram depicting the correlation between *Inc-IGSF9* expression levels and the expression of genes residing within 1Mb around the *Inc-IGSF9* locus after in cis overexpression of *Inc-IGSF9* or in the COAD cohort (TCGA). High correlation is depicted with narrow, blue ellipses, oriented upward from left to right, with a small diameter. In contrast, when the correlation coefficient is low, the ellipses are wider, red in color, and oriented upward from right to left. B. qPCR analysis showing the expression levels of *Inc-IGSF9* and *IGSF9* in LS174 cells transduced with NTG (control), sgRNA1, and sgRNA2. C. Snapshot showing the expression levels of both *Inc-IGSF9* and *IGSF9* in LS174 cells transduced with NTG (control), sgRNA1, and sgRNA2. Their relative genomic distance is ~16kb. D. qPCR analysis showing the expression levels of *CTNNB1*, *Inc-IGSF9*, and *IGSF9* in LS174 cells after shRNA-mediated β -catenin knock-down. E. Mean DESEQ counts in RNA-seq datasets derived from CRC patient-derived organoids that were treated or not with WRN/R-spondin medium to extrinsically activate the Wnt pathway or were genetically modified to carry APC mutations to intrinsically activate the Wnt pathway³⁰⁰.

Next, we investigated *IGSF9* expression levels in CRC cells following perturbation of the Wnt pathway. Interestingly, *IGSF9* was significantly upregulated after shRNA-mediated β -catenin knock-down (Figure 20D) despite evidently lacking direct β -catenin/TCF4 binding sites at its locus (data not shown). Conversely, both intrinsic and extrinsic activation of the Wnt pathway in intestinal organoids resulted in downregulation of *IGSF9* (Figure 20E). Together, these findings suggest that *IGSF9* is negatively regulated by the Wnt pathway, similarly to *Inc-IGSF9*.

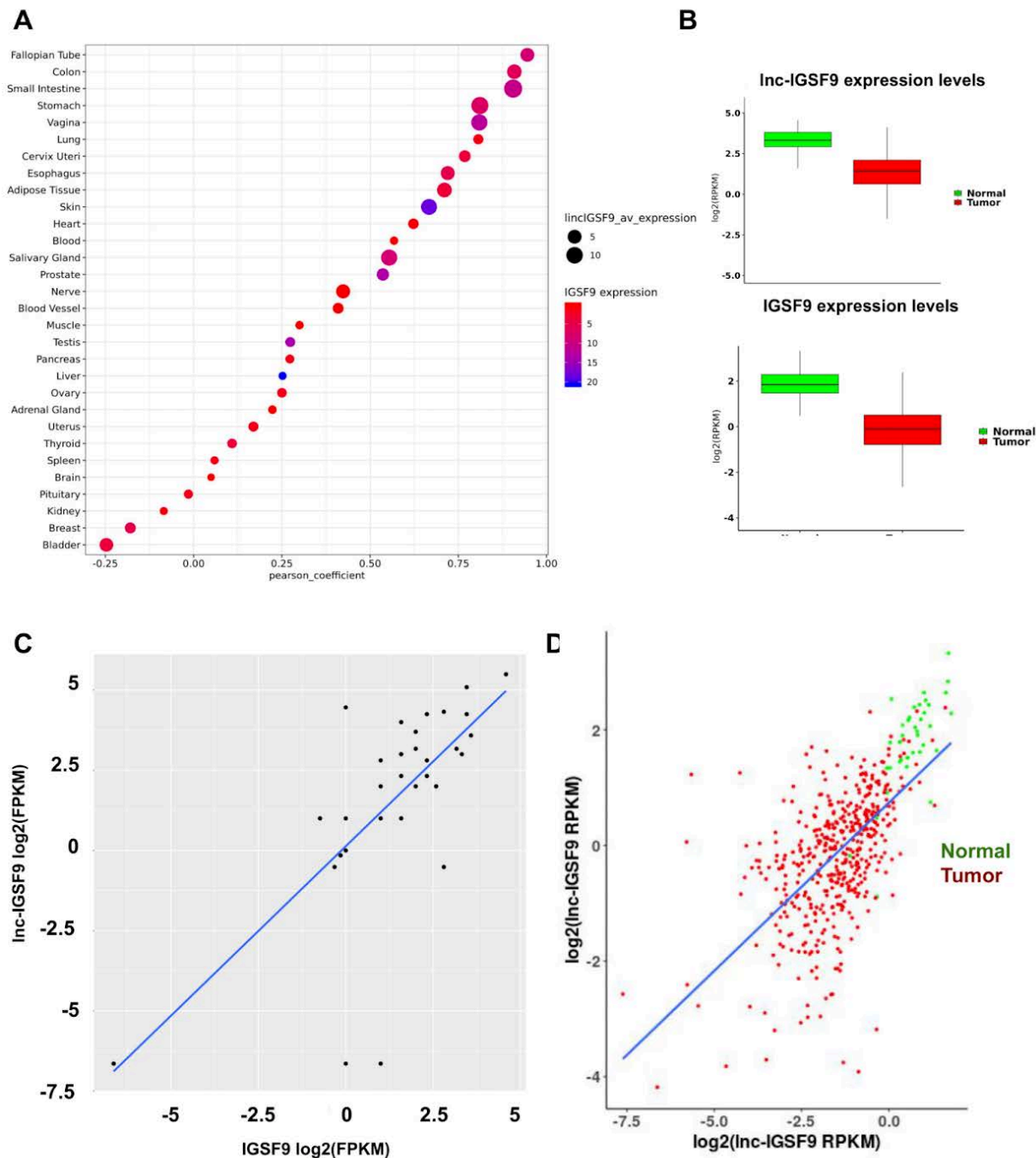


Figure 21: *Lnc-IGSF9* and *IGSF9* expression levels are correlated in multiple tissues and are downregulated in CRC.

A. Dotplot showing the correlation levels between *lnc-IGSF9* and *IGSF9* expression levels across all tissues, through the GTEX consortium. *B.* Boxplots depicting the expression patterns of *lnc-IGSF9* and *IGSF9* between tumor samples and adjacent normal tissues. *C.* Scatter plot demonstrating the correlation between *lnc-IGSF9* and *IGSF9* expression levels in 34 CRC cell lines based on RNA-seq data derived from the EBI expression atlas. *D.* Scatter plot displaying the correlation between *lnc-IGSF9* and *IGSF9* expression levels in CRC tumor and normal samples based on RNA-seq data from TCGA.

Furthermore, we interrogated the correlation between *lnc-IGSF9* and IGSF9 expression levels across different biological systems. Firstly, we analyzed publicly available RNA-seq datasets from the GTEx consortium, which includes data from all human tissues. Notably, both genes exhibited high correlation scores in most tissues (Figure 21A), with the colon, small intestine, and the stomach showing particularly strong correlation scores (Figure 21A). In tissues with weaker correlation scores, expression of either gene was frequently absent.

We then extended our analysis to cancerous samples and found that both *lnc-IGSF9* and IGSF9 were downregulated in CRC samples compared to normal adjacent tissue (Figure 21B), consistent with previous reports for *lnc-IGSF9* downregulation in CRC within distinct patient cohorts^{125,314}. A strong positive correlation score between *lnc-IGSF9* and IGSF9 was also observed in RNA-seq datasets from 34 CRC cell lines (Figure 21C) and TCGA COAD samples (Figure 21D). These findings support the notion that *lnc-IGSF9* and IGSF9 are co-regulated in intestinal tissues and CRC.

At the cellular level, a meta-analysis of single-cell RNA-seq data from intestinal epithelium revealed that both genes are minimally expressed in intestinal stem cells (ISCs), where Wnt signaling is active (Figure 22A). In contrast, their expression was higher in differentiated cells, with both genes predominantly expressed in enterocytes (Figure 22A). Using the CACO2 cell line as a model for intestinal differentiation, we confirmed that both *lnc-IGSF9* and IGSF9 are upregulated during differentiation (Figure 22B).

Transcriptional activation using CRISPR-a systems can induce the expression of neighboring genes by activating regulatory elements within the lncRNA locus. Therefore, we explored whether IGSF9 upregulation is mediated by overexpression of the *lnc-IGSF9* transcript or by activation of the *lnc-IGSF9* locus itself. Given that both *lnc-IGSF9* and IGSF9 are downregulated in CRC, with low expression levels in CRC lines, we performed co-transduction experiments employing shRNAs targeting either *lnc-IGSF9* or IGSF9 in conjunction with the activation of the *lnc-IGSF9* promoter.

We observed that IGSF9 expression was repressed in LS174 cells following treatment with both sgRNA1 and an shRNA targeting *lnc-IGSF9* (Figure 23A), suggesting that the *lnc-IGSF9* transcript is required for IGSF9 upregulation. In contrast, knockdown of IGSF9 did not affect *lnc-IGSF9* expression (Figure 23C). CFAs further revealed that *lnc-IGSF9* knockdown reversed the reduced clonogenicity associated with sgRNA1 transduction (Figure 23C), while IGSF9 knockdown reversed the phenotype induced by *lnc-IGSF9* activation (Figure 23D). These findings establish *lnc-IGSF9* as a pro-differentiation factor, acting through the regulation of IGSF9.

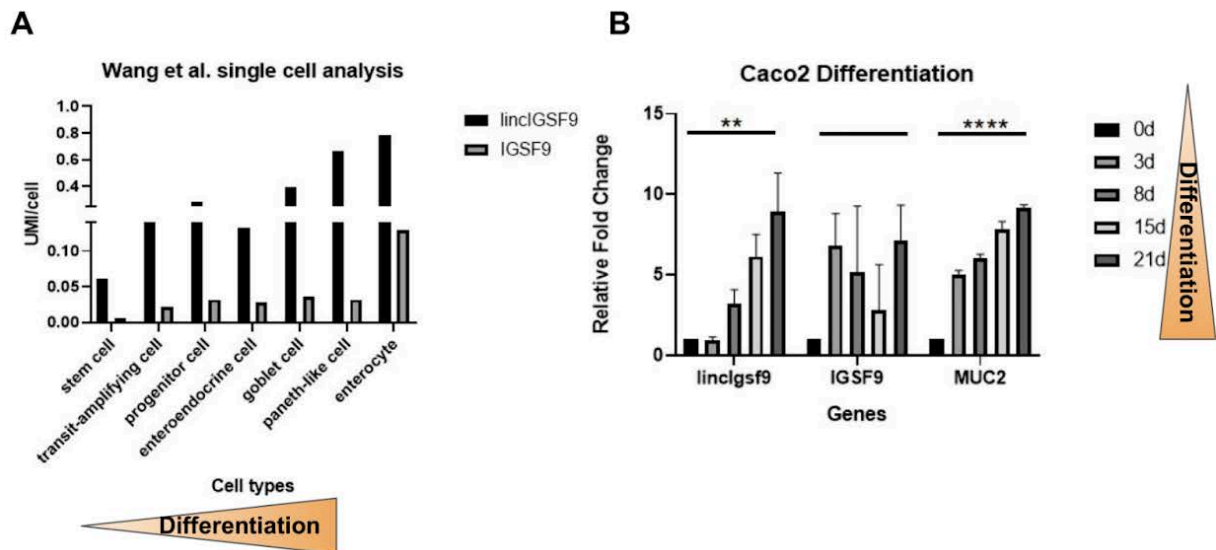


Figure 22: *Inc-IGSF9* and *IGSF9* are upregulated during intestinal cell differentiation.
A. Bar charts depicting the *Inc-IGSF9* and *IGSF9* expression levels (UMI/cell) across various intestinal epithelial cell types, derived from single cell RNA-seq³¹⁵. B. qPCR analysis of *Inc-IGSF9* and *IGSF9* expression levels in CACO2 cells during differentiation at 0, 3, 8, 15, and 21 days after reaching confluency. MUC2 serves as a positive control of differentiation.

CDX2 interacts with *Inc-IGSF9* to regulate *IGSF9* expression

To elucidate the mechanism through which *Inc-IGSF9* regulates *IGSF9*, we performed RNA-pull down followed by mass spectrometry in LS174 cells. We identified 13 proteins to interact with *Inc-IGSF9* (Figure 24A), all of which were nuclear. These include proteins implicated in chromatin architecture organization (SATB1, SATB2, SMARCA1 and CHD4), RNA processing and gene expression regulation (SAFB1, SAFB2, and HNRNPUL2) or epigenetic regulation (HDAC2). Notably, CDX2 caught our attention as a tissue-specific transcription factor that orchestrates intestinal differentiation through gene expression regulation³¹⁶.

CDX2 overexpression in LS174 cells led to significant upregulation of both *Inc-IGSF9* and *IGSF9* (Figure 24B). Meta-analysis of ChIP-seq data performed with an anti-CDX2 antibody revealed two binding sites for CDX2 on the *Inc-IGSF9* promoter (Figure 24C), while a weaker binding site was detected on the *IGSF9* promoter (data not shown).

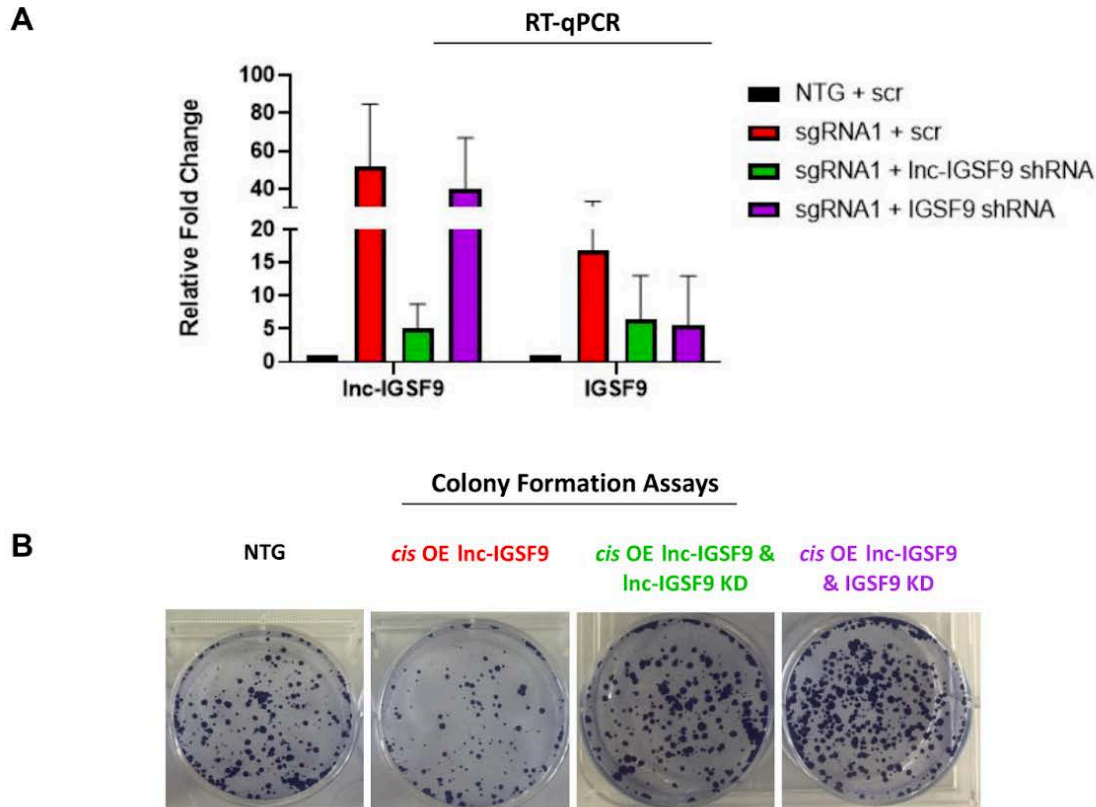


Figure 23: The *Inc-IGSF9* lncRNA, and not its genomic locus, regulates IGSF9 to reduce CRC clonogenicity.

A. qPCR analysis of *Inc-IGSF9* and *IGSF9* expression levels in LS174 cells transduced with NTG & scrambled, sgRNA1 & scrambled, sgRNA1 & shRNA targeting *Inc-IGSF9*, and sgRNA1 & shRNA targeting *IGSF9*. *B.* CFAs of LS174 cells that were transduced with NTG & scrambled, sgRNA1 & scrambled, sgRNA1 & shRNA targeting *Inc-IGSF9*, and sgRNA1 & shRNA targeting *IGSF9*.

To determine whether *Inc-IGSF9* transcript is necessary for the CDX2-mediated IGSF9 upregulation, we used CRISPR-Cas9 to delete the *Inc-IGSF9* promoter, which contains one of the two CDX2 binding sites (Figure 24C). This resulted in dramatic reduction of *Inc-IGSF9* expression levels in two cell lines with homozygous deletion of the *Inc-IGSF9* promoter (Figure 24D). The reduction of IGSF9 expression was not statistically significant, most likely because the basal IGSF9 levels in these cell lines are very low due to constitutive Wnt activity. Nevertheless, we discovered that CDX2 failed to induce IGSF9 expression in the absence of *Inc-IGSF9* (Figure 24D), suggesting that CDX2 regulates IGSF9 expression through *Inc-IGSF9*.

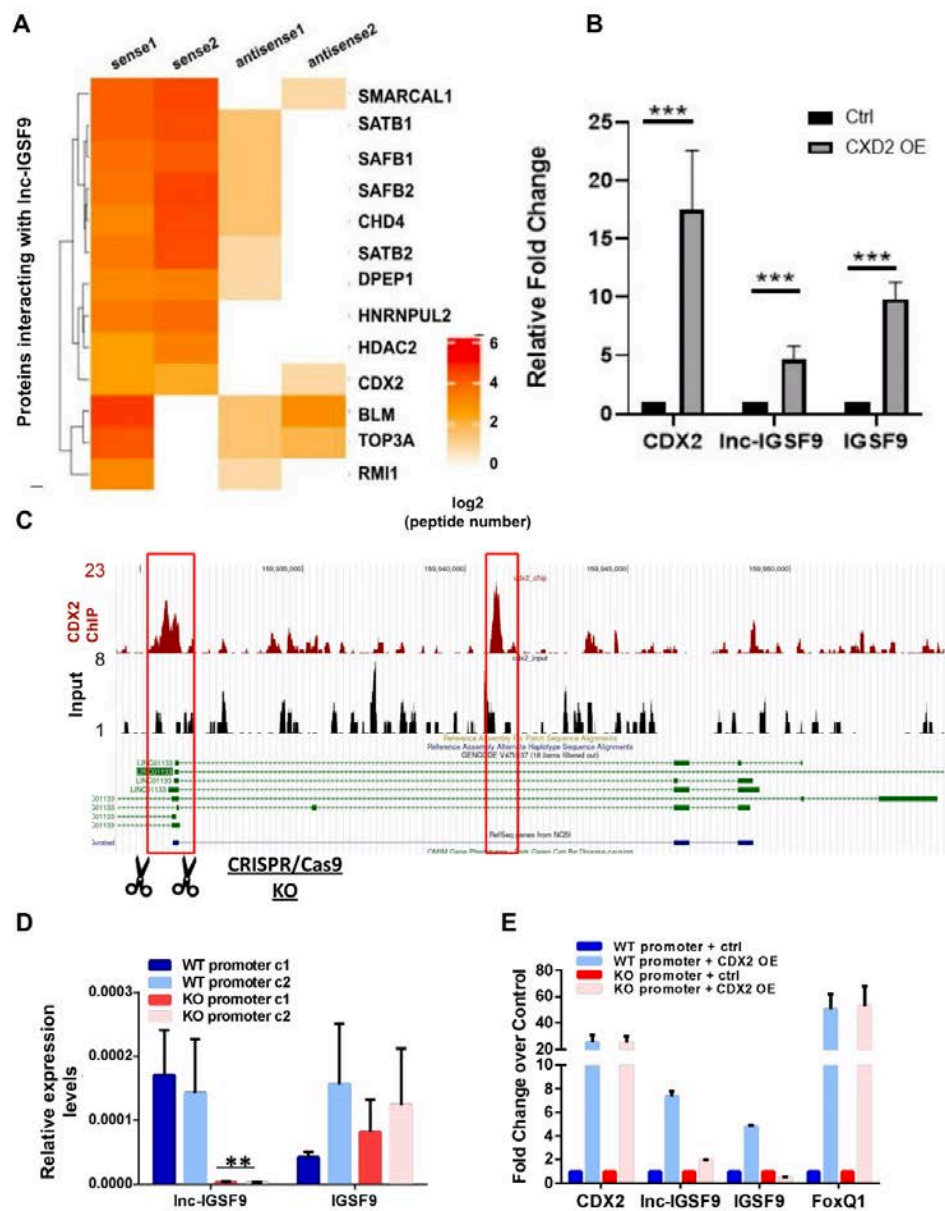


Figure 24: CDX2 binds on the Inc-IGSF9 locus to activate the expression of Inc-IGSF9 and IGSF9.

A. Heatmap illustrating proteins interacting with Inc-IGSF9, identified through RNA pull-down followed by mass spectrometry. Peptide counts of proteins binding to the sense Inc-IGSF9 sequence are compared to the antisense sequence, which serves as a control. **B.** qPCR analysis showing Inc-IGSF9 and IGSF9 expression levels in LS174 cells upon CDX2 overexpression. **C.** Genomic snapshot illustrating the binding sites of CDX2 on the Inc-IGSF9 locus. Binding sites are marked with red boxes. **D.** qPCR analysis of Inc-IGSF9 and IGSF9 expression levels in wild-type (WT) LS174 cells compared to cells with homozygous knockout of the Inc-IGSF9 promoter (KO). **E.** qPCR analysis of Inc-IGSF9 and IGSF9 expression levels in WT and homozygous Inc-IGSF9 promoter knockout cell clones (c1 & c2) following CDX2 overexpression. Expression levels in WT and knockout cells are normalized separately to their respective controls.

Part II: LncRNA EPB41L4A-AS1 encodes a novel mitochondrial microprotein with tumor suppressive effects

LncRNA EPB41L4A-AS1 encodes a 13kDa mitochondrial microprotein

Emerging studies demonstrate that numerous lncRNAs contain open reading frames (ORFs) with protein-coding potential, which have critical implications for human biology. To investigate the presence of translatable lncORFs in CRC, we meta-analyzed ribosome profiling experiments from HCT116 cells treated with ribosome stalling drugs cycloheximide and lactimidomycin. We leveraged the ORF-RATER²⁴⁷ algorithm to discover novel ribosome-occupied small ORFs (smORFs) contained within lncRNAs (lncORFs) in HCT116 cells (data not shown). Specifically, we detected 241 lncORFs in the cycloheximide dataset, 274 in the lactimidomycin dataset, and 96 that were common to both conditions (Figure 25A).

The distribution of lncORFs based on codon length and initiation codon revealed that the majority of lncORFs were between 20-30 codons in length (Figure 25B), and initiated with ATG (Figure 25C). The second most common start codon used for lncORF initiation was CTG (Figure 25C), consistent with previous reports²³³. We focused on the overlapping lncORFs, considering them as more robust candidates for potential lncRNA translation.

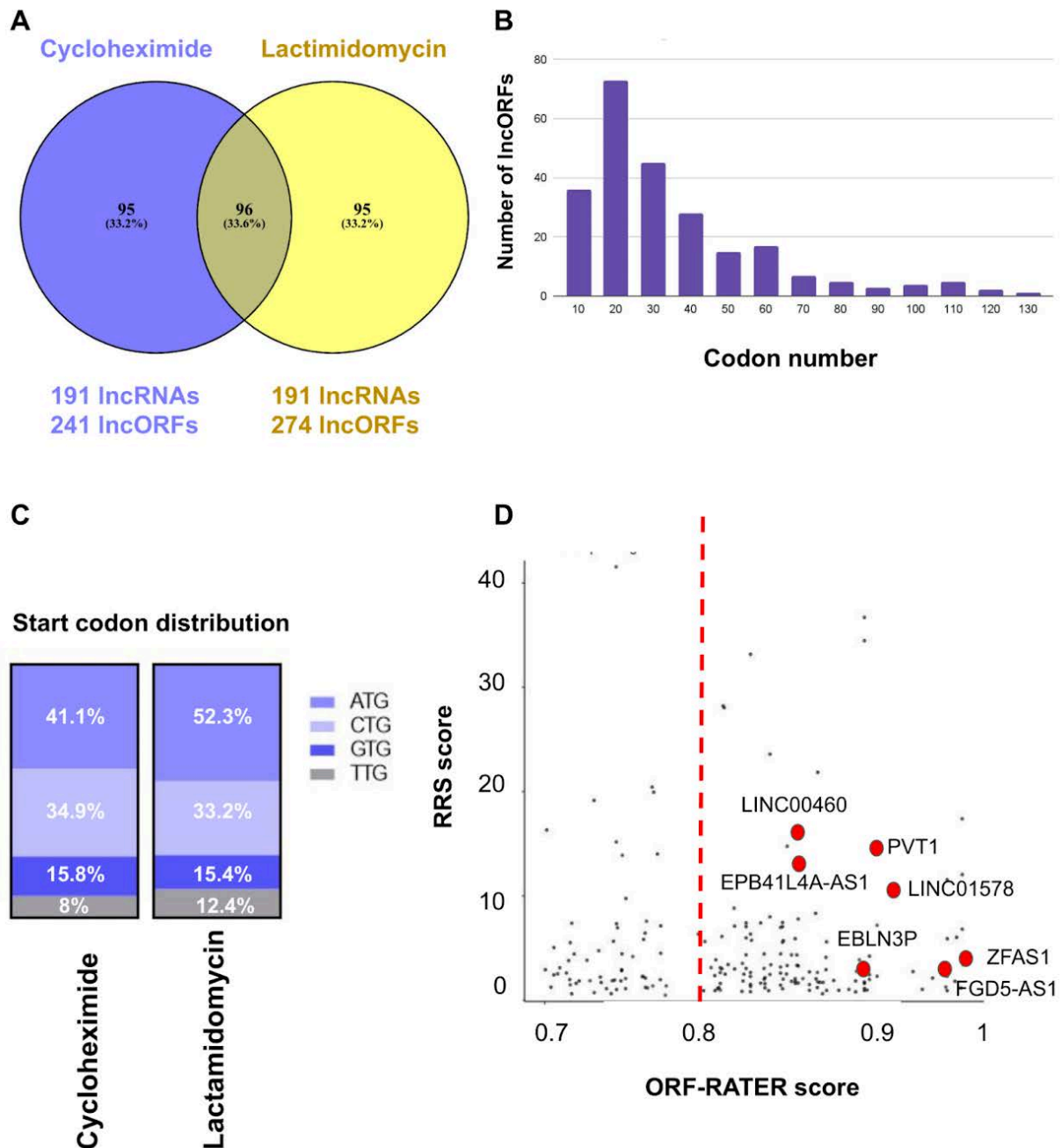


Figure 25: Ribosome profiling reveals the potential translation of multiple IncORFs in HCT116.

A. Venn diagrams showing the overlap of identified IncORFs from ribosome profiling of HCT116 cells treated with cycloheximide and lactimidomycin. B. Bar charts illustrating the distribution of IncORF sizes, categorized by codon count, in ribosome profiling of HCT116 cells treated with lactimidomycin. C. Bar plot displaying the percentage frequency of IncORFs with canonical (ATG) and non-cognate start codons. D. Scatter plot representing the distribution of IncORFs based on ORF-RATER and RRS scores. Red dots indicate IncORFs selected through manual curation.

Subsequent proteomic analysis of HCT116 cells and interrogation of publicly available proteome datasets did not validate the presence of the predicted SEPs, likely due to the limitations of mass spectrometry technology^{233,246}. To increase the likelihood of successfully predicting IncORFs that encode microproteins and refine candidate selection, we estimated the RRS³⁷ score of each IncORF. By taking into account both ORF-RATER scores and RRSs, we manually curated their ribo-seq signal to finally proceed with a list of 7 high confidence candidates (Figure 25D).

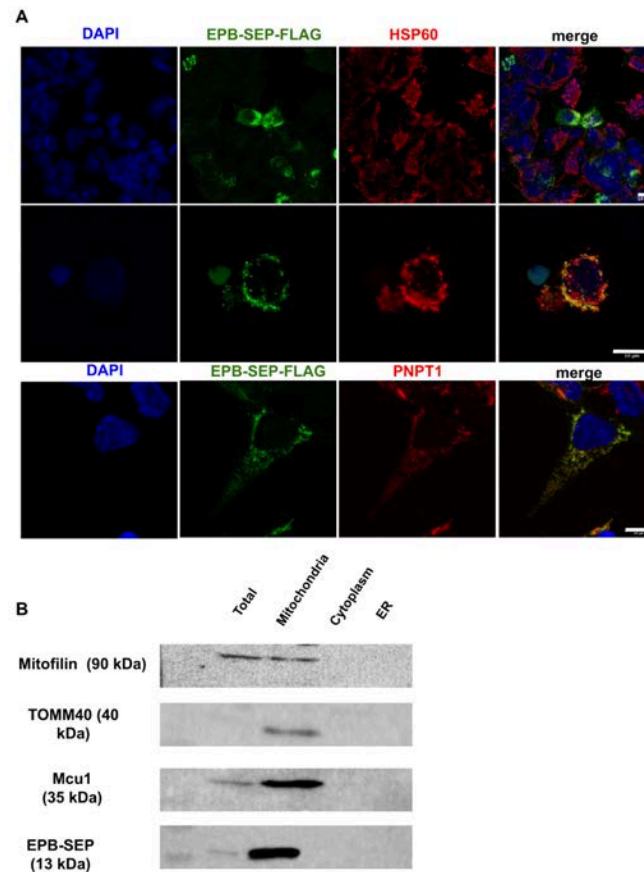


Figure 26: EPB41L4A-AS1 encodes the mitochondrial microprotein EPB-SEP.

A. Immunostaining of HEK293T cells overexpressing EPB-SEP-FLAG (green) with the mitochondrial markers HSP60 (upper panel, red) & PNPT1 (lower panel, red). B. Western blot analysis of protein extracts from subcellular fractions (mitochondrial, cytoplasmic & endoplasmic reticulum) with Mitofilin, TOMM40, MCU1 serving as mitochondrial markers.

To examine whether the predicted IncORFs are translated into SEPs, we cloned the IncORFs along with their respective 5'UTRs, fused them to a FLAG epitope at the C' terminus, within the pcDNA4.TO expression vector. Recombinant constructs were overexpressed in HEK293T and HCT116 cells, and protein expression was confirmed through immunofluorescence and Western blotting, using an anti-FLAG antibody. Strikingly, we

discovered that the EPB41L4A-AS1 lncRNA encodes a microprotein, which we named EPB-SEP (Figure 26A). Co-staining experiments of EPB-SEP and different subcellular organelle markers revealed co-localization with the mitochondrial markers HSP60 and PNPT1 (Figure 26A). Furthermore, cell fractionation followed by Western Blotting confirmed that EPB-SEP is predominantly localized in the mitochondrial fraction, while it is not detected in the cytoplasmic or the endoplasmic reticulum fractions (Figure 26B).

The identified lncORF is located within the first exon of EPB41L4A-AS1 and is initiated by a non-cognate CTG start codon (Figure 27A). The lncORF is not evolutionary conserved and is restricted to primates, suggesting that it likely arose through de novo gene birth²³⁵. The predicted protein is 107aa long and structural modeling through AlphaFold3³¹⁷ failed to identify any domains with high confidence (Figure 27A). Western blot analysis confirmed the size of the microprotein as ~13 kDa (Figure 26B). Additionally, mass spectrometry analysis of HEK293T cells overexpressing EPB-SEP-FLAG identified multiple predicted peptide sequences (Figure 27B).

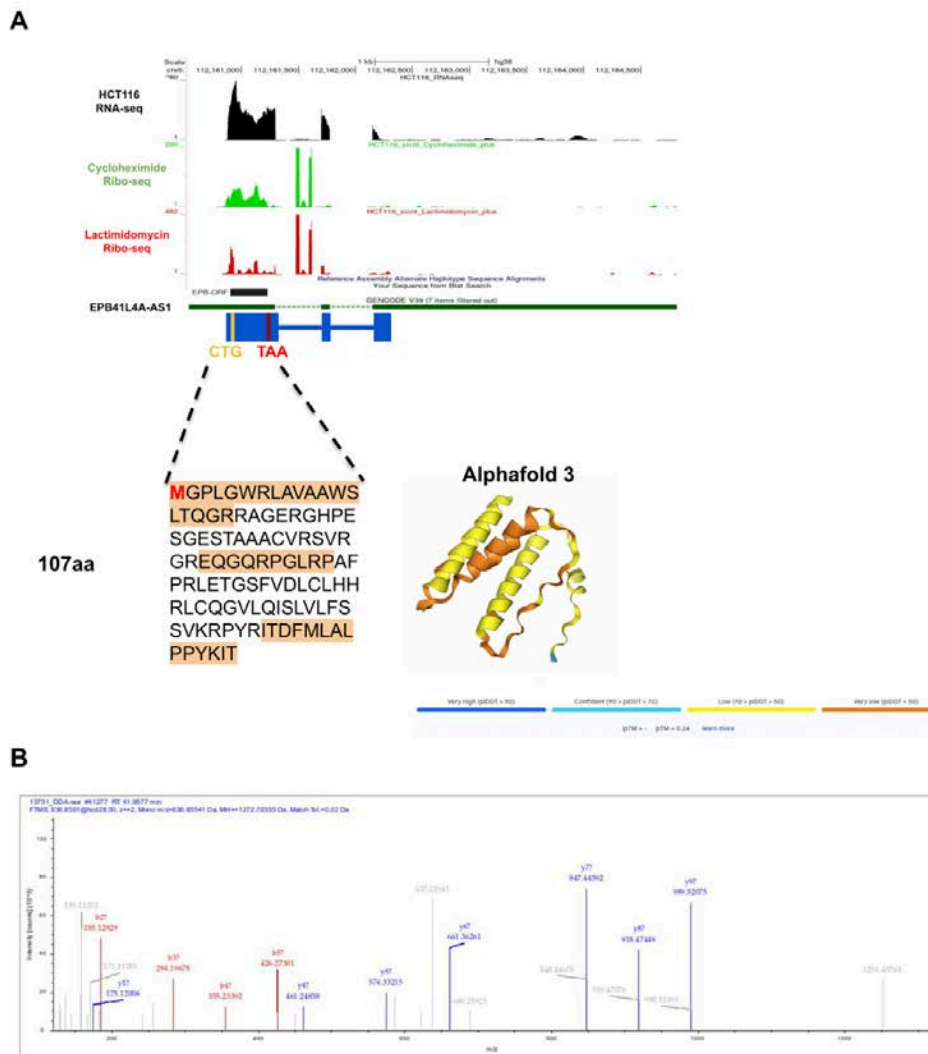


Figure 27: A CTG-initiating IncORF located within EPB41L4A-AS1 is translated to give rise to the 107aa long EPB-SEP microprotein.

A. Snapshot of Ribosome profiling data on the predicted CTG-initiating IncORF. The resulting 107-amino-acid (aa) EPB-SEP microprotein lacks high-confidence structural domains as predicted by AlphaFold3. Colors of the protein structure correspond to different levels of confidence (blue: very high, cyan: confident, yellow: low, orange: very low). B. Mass spectrometry spectra confirm the presence of EPB-SEP peptides in HEK293T cells overexpressing the IncORF, confirming its translational potential. B-ions (represented in red) correspond to fragments starting from the N-terminus, while y-ions (represented in blue) correspond to fragments starting from the C-terminus.

EPB-SEP overexpression reduces mitochondrial membrane potential

To isolate cells overexpressing and translating EPB-SEP, we generated a construct encoding a fusion protein of EPB-SEP and GFP (Figure 28A). Immunostaining of cells overexpressing EPB-SEP-GFP with mitotracker Red revealed an altered localization of the fusion protein in the mitochondrial periphery (Figure 28B). We speculate that the addition of GFP altered the biochemical properties of EPB-SEP. To overcome this, we generated a cassette containing the 5'UTR-EPB-ORF-FLAG linked to the GFP ORF through a T2A sequence (EPB-T2A-GFP). The T2A sequence impairs the formation of a peptide bond at the end of its sequence, enabling ribosome skipping and allowing independent translation of the downstream ORF³¹⁸. So, cells transfected with the EPB-T2A-GFP cassette express a bicistronic transcript, resulting in the production of EPB-SEP-FLAG and GFP separately (Figures 28C & 28D).

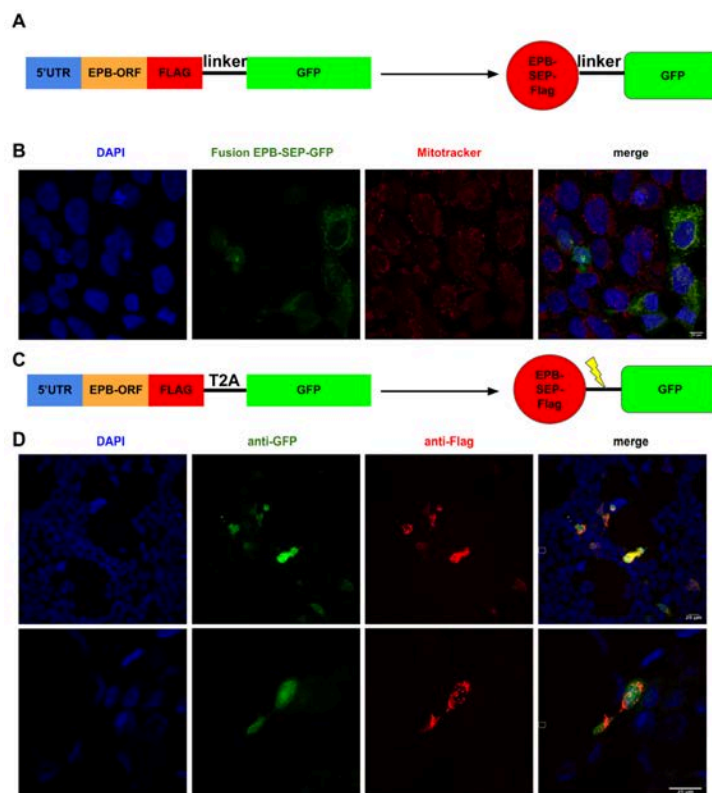


Figure 28: Generation and Characterization of EPB-SEP-GFP and EPB-T2A-GFP Constructs to Study EPB-SEP Localization and Translation.

A. Graphical representation of the EPB-SEP-GFP fusion cassette. The 5'UTR coupled to the EPB-ORF and FLAG epitope was fused with the GFP sequence through a linker. **B.** Immunostaining experiments showing the localization of EPB-SEP-GFP in HEK293T cells, after staining with mitotracker Red which marks active mitochondria. **C.** Schematic of the EPB-T2A-GFP cassette. The 5'UTR and EPB-ORF-FLAG are linked to GFP through a T2A sequence enabling the separate translation of EPB-SEP and GFP by the same bicistronic transcript. **D.** Immunostaining experiments using antibodies against the GFP protein and the FLAG epitope to demonstrate the co-translation of EPB-SEP and GFP.

Given the mitochondrial localization of EPB-SEP, we examined its function in mitochondrial physiology. We used mitotracker Red staining, which stains active mitochondria as it accumulates based on negative mitochondrial potential and becomes oxidized by interacting with thiols residing in mitochondrial proteins³¹⁹. Interestingly, we observed that HEK293T and HCT116 cells overexpressing EPB-SEP-FLAG were characterized by reduced mitotracker staining (Figures 29A & 29B). This was further validated by FACs analysis using the TMRE dye, a more sensitive method to quantify mitochondrial membrane potential, which confirmed the significant reduction in mitochondrial membrane potential in EPB-T2A-GFP+ cells (Figure 29C). Immunostaining also demonstrated the reduced TMRE staining in EPB-T2A-GFP+ cells compared to cells overexpressing pcDNA4.TO-GFP (Figure 29D).

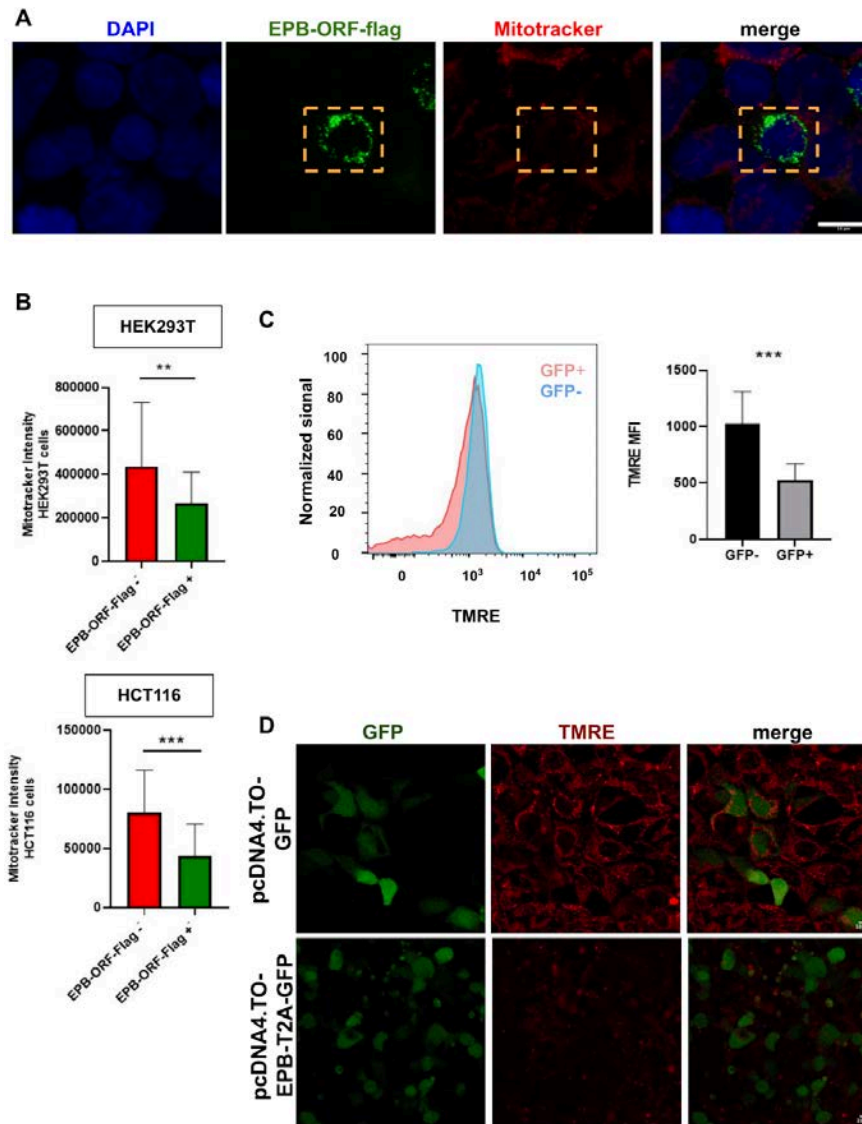


Figure 29: EPB-ORF overexpression reduces mitochondrial membrane potential.

A. Immunostaining experiments in HEK293T cells that overexpress EPB-SEP-FLAG, using an antibody against the FLAG epitope and Mitotracker Red to mark active mitochondria. **B** Quantification of Mitotracker Red signal intensity in HEK293T and HCT116 cells expressing EPB-SEP-FLAG compared to control cells. **C.** Flow cytometry histograms showing TMRE staining in HEK293T cells that overexpress EPB-ORF-FLAG-T2A-GFP with barplots comparing TMRE mean fluorescence intensities (MFIs) in GFP⁺ versus GFP⁻ cell populations. **D.** Representative image of TMRE staining in SKNSH cells overexpressing either pcDNA4.TO-GFP or pcDNA4.TO-EPB-T2A-GFP.

To decipher how EPB-SEP affects mitochondrial physiology, we performed proteomic profiling of HEK293T cells that overexpress EPB-T2A-GFP compared to control cells transfected with the empty vector. We found 40 upregulated proteins and 147 downregulated proteins (Figure 30A). Pathway analysis revealed that most proteins are involved in mitochondrial regulation and mitosis (Figure 30B), highlighting the predominant enrichment of

mitochondrial-related proteins (Figure 30C). To further explore this observation, we interrogated the expression levels of all mitochondrial proteins in our list of differentially expressed proteins, using the MitoCarta database³²⁰. Strikingly, 31 mitochondrial proteins were found downregulated in EPB-T2A-GFP+ cells while only one (PAM16) was upregulated, suggesting that the overexpression of EPB-SEP impairs the mitoproteome (Figure 30D).

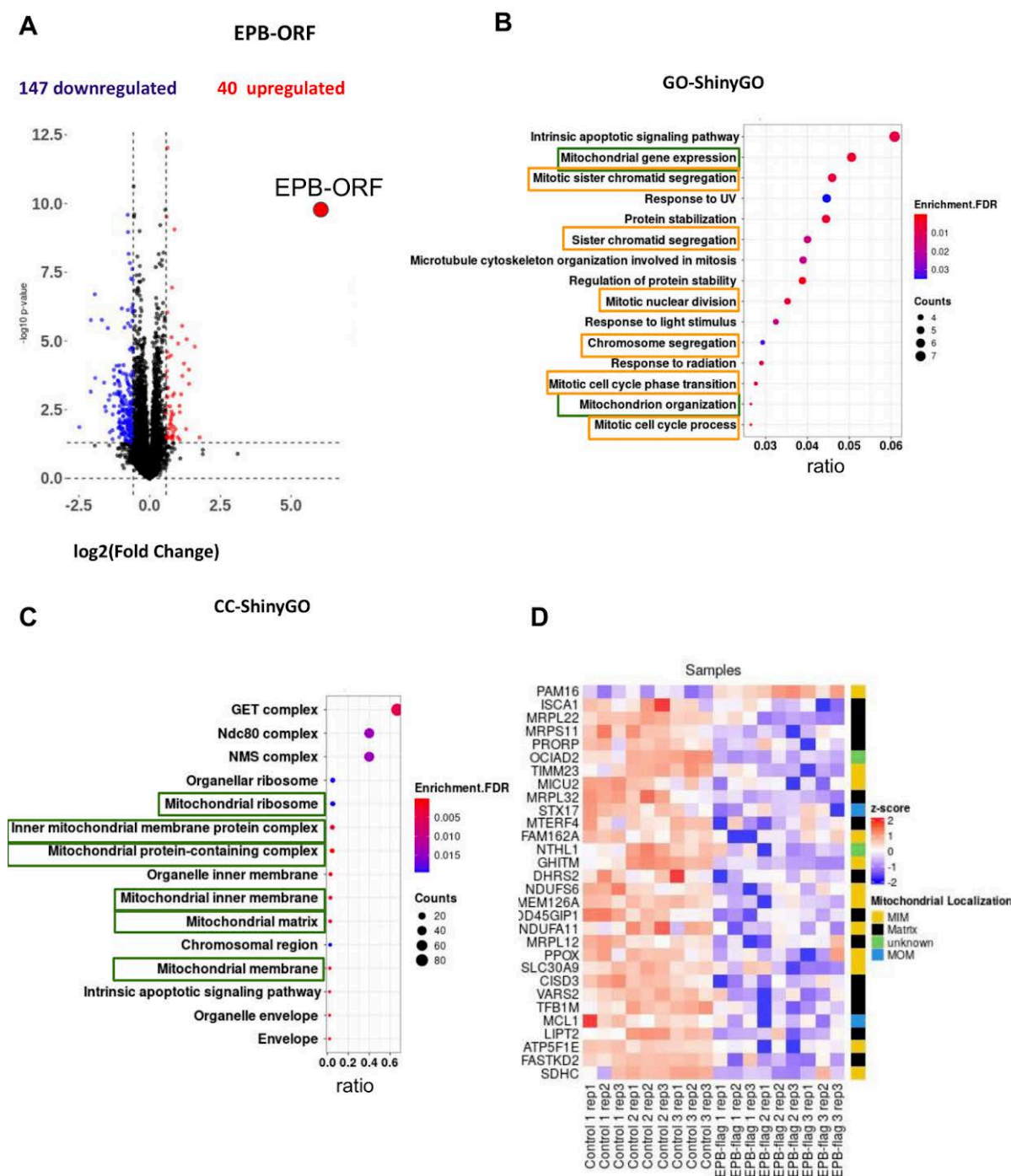


Figure 30: EPB-ORF overexpression impairs the mitoproteome and downregulates mitosis-associated proteins.

A. Volcano plot depicting the proteins differentially expressed between EPB-T2A-GFP+ and EPB-T2A-GFP- HEK293T cells. B. Gene ontology (GO) analysis of downregulated proteins using GeneCodis, with mitochondrial-related pathways highlighted in green and mitosis-related pathways in orange. C. GO analysis of cellular compartments, showing downregulated proteins enriched in mitochondrial compartments (green). D. Heatmap showing the mitochondrial proteome deregulated upon EPB-ORF OE. Every protein was also annotated based on its sub-mitochondrial localization: inner mitochondrial membrane (IMM, yellow), Matrix (black), unknown (green), mitochondrial outer membrane (MOM, blue). Data was retrieved from the MitoCarta database³²⁰.

EPB-SEP exerts tumor suppressive effects

Gene ontology analysis demonstrated that beyond mitochondrial impairment, EPB-SEP overexpression also led to the downregulation of mitotic proteins. CFAs revealed that HEK293T cells overexpressing either the EPB-SEP protein or EPB41L4A-AS1 lncRNA exhibited reduced clonogenicity compared to control cells expressing the empty vector (Figure 31A). Similarly, EPB-T2A-GFP+ cells were characterized by diminished clonogenicity compared to EPB-T2A-GFP- cells (Figure 31B). In contrast, knock-down of EPB41L4A-AS1 using two different shRNAs resulted in a significant decrease of EPB41L4A-AS1 expression levels (Figure 31C) and increased cell clonogenicity (Figure 31D). Together, our results suggest that EPB-SEP impairs cancer cell proliferation.

To further understand how EPB-SEP suppresses carcinogenesis, we conducted cell cycle analysis with propidium iodide (PI). We observed that EPB-T2A-GFP+ cells exhibited a marked arrest in the Synthesis and G2 phase, indicating disruption of normal cell division (Figure 32A). FACs analysis for Annexin V staining showed no differences in apoptotic rates between EPB-T2A-GFP+ and EPB-T2A-GFP- cells (Figure 32B), implying that reduced viability occurs through apoptosis-independent mechanisms.

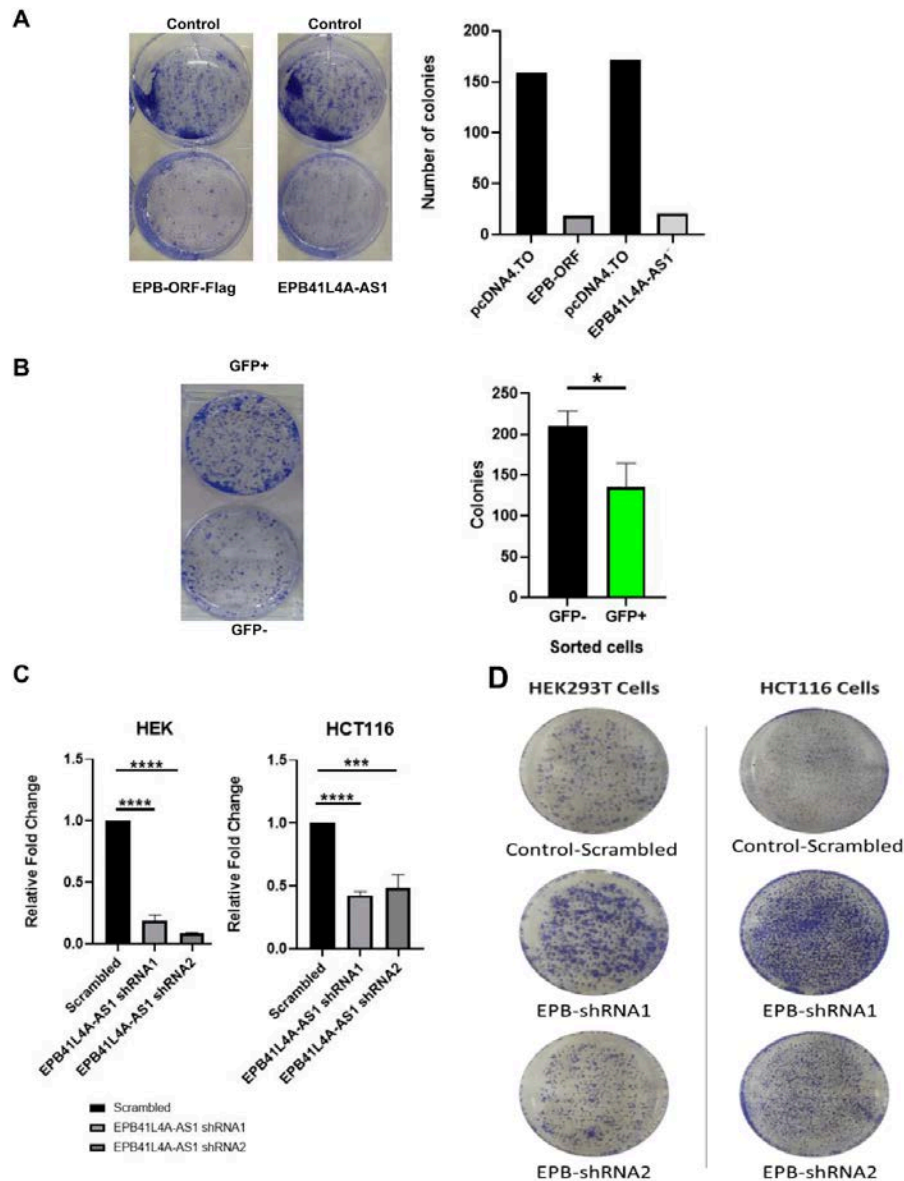


Figure 31: EPB-ORF functions as a tumor suppressor microprotein.

A) CFAs of HEK293T cells that overexpress either EPB-ORF (protein) or EPB41L4A-AS1 (lncRNA) compared to cells expressing the empty vector (pcDNA4.TO). B) CFAs of sorted HEK293T cells that overexpress EPB-T2A-GFP compared to EPB-T2A-GFP- cells. C) qPCR analysis of EPB41L4A-AS1 expression in HEK293T and HCT116 cells following shRNA-mediated knockdown, using two primer sets (EPB41L4A-AS1 F1R1 and EPB-ORF F1R1). (EPB41L4A-AS1 F1R1 & EPB-ORF F1R1). D) CFAs of HEK293T and HCT116 cells expressing shRNAs against EPB41L4A-AS1.

EPB-SEP interacts with several mitochondrial proteins

To investigate the mechanism of function of EPB-SEP, we performed FLAG immunoprecipitation followed by mass spectrometry in HEK293T cells expressing

EPB-SEP-FLAG (Figure 33A). More specifically, we performed IP in HEK293T cells that overexpressed the EPB-SEP microprotein fused to the FLAG epitope either in the N- or the C-terminus. By combining the proteomics results of the two conditions, we identified 13 common protein interactors (Figure 33B), five of which are mitochondrial proteins (Figure 33C). Notably, HAX1 and CLPB have already been shown to interact with each other in the intermembrane space to regulate mitochondrial proteostasis³²¹. Ongoing studies include reverse immunoprecipitation targeting these interactors and proximity ligation assays, followed by mass spectrometry, to further characterize these interactions and their functional implications.

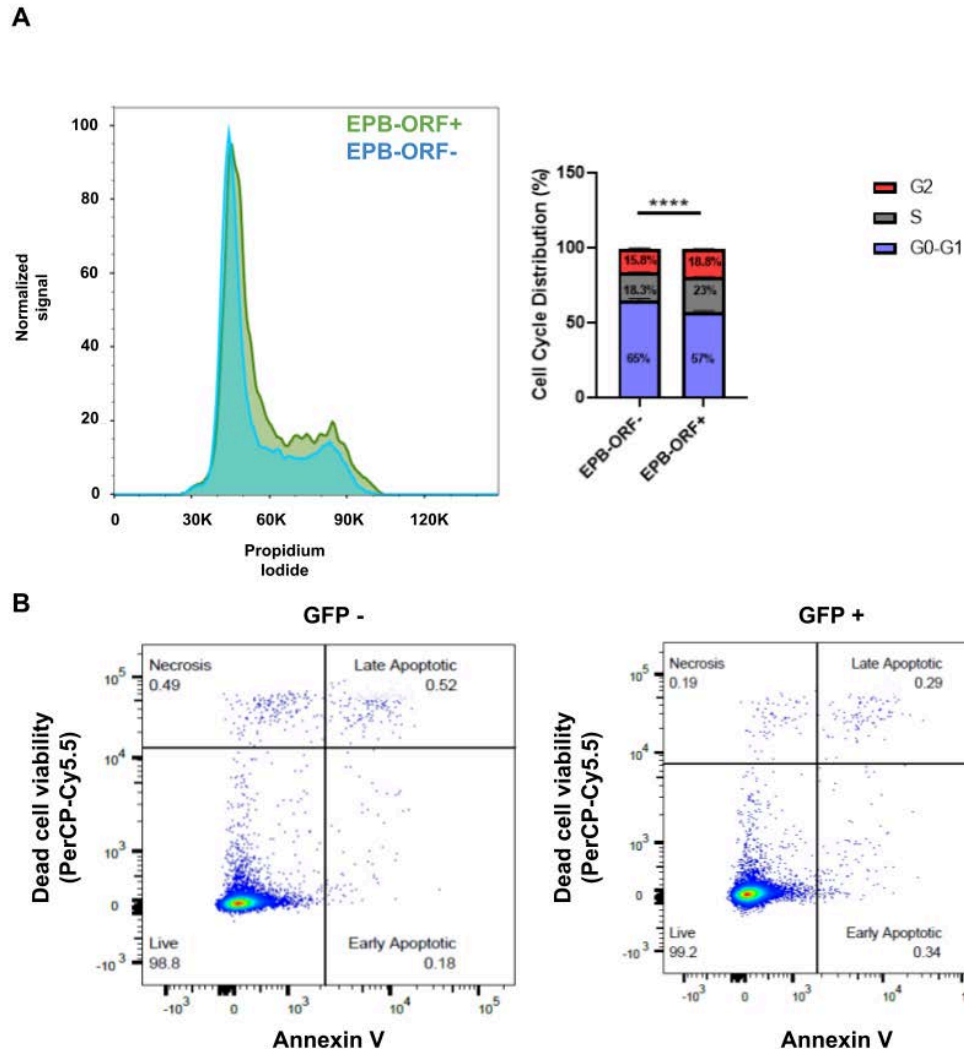


Figure 32: EPB-ORF overexpression leads to G2 arrest without affecting apoptosis.

A) Flow cytometry histograms showing propidium iodide (PI) staining of HEK293T cells overexpressing EPB-T2A-GFP versus GFP- cells. Bar plots depict the percentage of cells in each cell cycle phase, with G1 (red), S (gray), and G2 (black) stages. B) Flow cytometry charts of Annexin V staining for apoptosis analysis of HEK293T cells overexpressing EPB-T2A-GFP versus GFP- cells. Quadrants represent necrotic, late apoptotic, early apoptotic, and live cell populations.

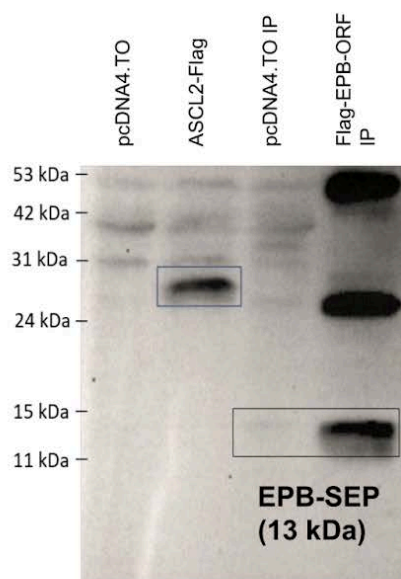
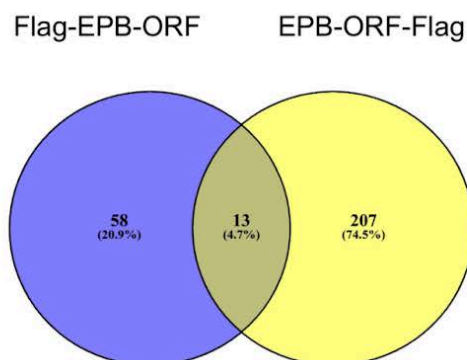
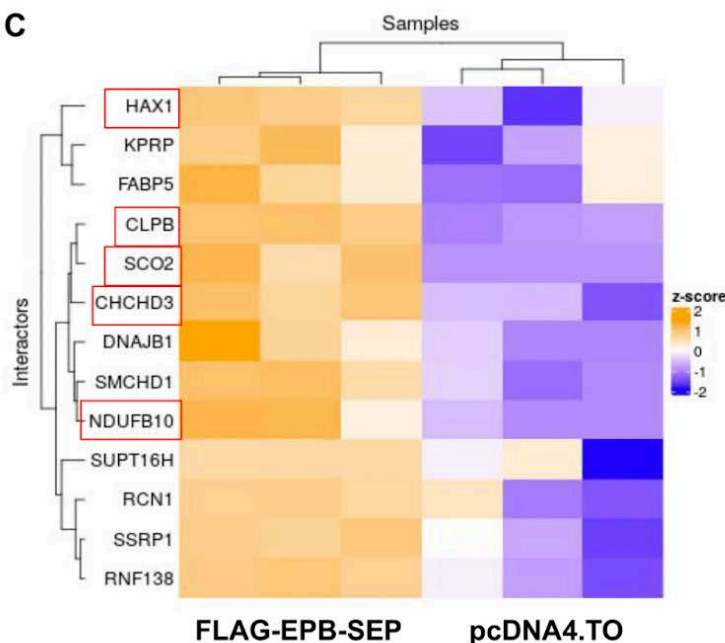
A**B****C**

Figure 33: IP-MS reveals that EPB-ORF interacts with several mitochondrial proteins.

A) Western blot analysis of EPB-SEP levels in total protein extracts from EPB-SEP-FLAG overexpressing HEK293T cells subjected to IP with FLAG antibody. Extracts from HEK293T cells that overexpress ASCL2-FLAG served as positive control, while cells overexpressing the empty vector (pcDNA4.TO) served as negative control for background of the IP. B) Venn diagrams summarizing protein interactants identified in HEK293T cells overexpressing EPB-SEP-FLAG with FLAG at either the N- or C-terminus. C) Heatmap showing the z-scores of the intensity derived from MS for the 13 common protein interactants.

5. Discussion

Part I

The Wnt pathway is the major driver of CRC, with genetic perturbations that affect key players of the pathway leading to the continuous accumulation of β -catenin in the cytosol and uncontrollable proliferation of stem cells in the intestinal epithelium^{145,163}. While several lncRNAs participate in complexes which involve components or effectors of the Wnt pathway²¹⁶, only *WNTRLINC1* has been characterized as a direct β -catenin/TCF4 target with implications in the Wnt-dependent transcriptional repertoire. Its transcription is activated by the β -catenin/TCF4 complex, and *WNTRLINC1* interacts with this complex to form an intra-chromosomal loop, connecting its promoter with the 3' regulatory region of the adjacent gene *ASCL2*²²⁸, which encodes a master regulator of intestinal stem cell homeostasis¹⁹⁶. Notably, *WNTRLINC1* recruits the β -catenin/TCF4 complex to the *ASCL2* gene, and in turn, *ASCL2* activates *WNTRLINC1* expression²²⁸.

Building on these findings, we investigated the role in CRC of other lncRNAs regulated by the Wnt pathway. To achieve this, we conducted i) transcriptomic profiling of LS174 cells with repressed Wnt pathway activity to delineate the Wnt-mediated transcriptional repertoire, and ii) analyzed ChIP-seq data to map genome-wide β -catenin/TCF4 binding sites. Our approach revealed that *lnc-IGSF9* (*LINC01133*) is a direct Wnt target negatively regulated by its activity. More specifically, the β -catenin/TCF4 complex binds to an intronic enhancer of *lnc-IGSF9*, and the expression of the gene increases upon β -catenin knock-down. Conversely, Wnt signaling activation in intestinal organoids, either through treatment with WRN medium or through genetic deletion of *APC* resulted in decreased *lnc-IGSF9* expression. Furthermore, treatment of CACO2 cells with Chir, a GSK3 inhibitor, increased *lnc-IGSF9* expression, confirming its negative regulation by Wnt signaling. **Thus, while the β -catenin/TCF4 complex typically promotes the expression of target genes, *lnc-IGSF9* represents an unusual case of apparent Wnt-mediated negative transcriptional regulation, even though the details of this mechanism have not been fully elucidated.**

To investigate the function of *lnc-IGSF9* in CRC, we conducted subcellular fractionation and FISH experiments in LS174 cells. Both methods revealed that *lnc-IGSF9* is predominantly localized within the nuclear chromatin fraction. While multiple studies have linked *lnc-IGSF9* with cytoplasmic roles as a ceRNA that sequesters microRNAs¹²¹, evidence supporting its cytoplasmic localization remains limited^{123,309}. In contrast, other studies have demonstrated its nuclear localization in breast¹²⁹, pancreatic¹³¹, and colorectal cancer¹²⁵.

Nuclear lncRNAs often regulate local gene expression *in cis* or interact with protein complexes to influence chromatin architecture and the epigenetic state^{6,31}. Indeed, we found that while ectopic overexpression of *lnc-IGSF9* did result in transcript accumulation within the nucleus and the cytoplasm, this had no effect on CRC proliferation. On the other hand,

CRISPR-a-mediated *in cis* overexpression of *Inc-IGSF9* inhibited CRC proliferation and reduced tumor size in xenograft models. Our findings corroborate that *Inc-IGSF9* primarily operates within the nuclear compartment, potentially affecting gene expression locally.

Therefore, we performed transcriptome profiling of CRC cells that overexpress *Inc-IGSF9 in cis* and discovered that it led to the deregulation of genes involved in processes related to cell adhesion and mobility, suggesting that *Inc-IGSF9* serves as an orchestrator of intestinal cell differentiation. Also, this finding aligns with the observation that *Inc-IGSF9* is repressed upon activation of the stemness-promoting Wnt signaling. However, we did not detect major differences in the Wnt-effector repertoire. Notably, we observed the downregulation of the FZD receptors and the striking upregulation of the Wnt inhibitor DKK1, indicating that *Inc-IGSF9* may negatively regulate the initiation of the Wnt signaling cascade. This effect does not extend downstream in the pathway, likely because the experiments were performed in LS174 cells, which exhibit constitutive β -catenin nuclear accumulation and maintain an active Wnt pathway independently of upstream regulation. Surprisingly, DKK1 emerged as one of the top upregulated genes, despite previous reports indicating that *Inc-IGSF9* binds to the DKK1 promoter in pancreatic cancer cells, inducing its trimethylation and subsequent repression. Furthermore, another study demonstrated that *Inc-IGSF9* interacts with EZH2, silencing the DC components AXIN2 and GSK to promote Wnt pathway activity in pancreatic cells¹³². This discrepancy underscores the potential tissue-dependency of the roles and mechanisms of action of *Inc-IGSF9*.

Based on our phenotypic observations, we investigated whether *Inc-IGSF9* affects the expression of genes residing in the vicinity of its locus. We discovered that *in cis* overexpression of *Inc-IGSF9* led to the upregulation of its neighbouring gene *IGSF9*, while no other gene within a 1Mb genomic window displayed altered expression levels. Moreover, *IGSF9* also appeared to be negatively regulated by the Wnt pathway, similar to *Inc-IGSF9*. By combining RNA-seq datasets from GTEx, TCGA (COAD), and cancer cell lines, we discovered that *Inc-IGSF9* expression positively correlates with *IGSF9 expression* in most human tissues and COAD. Furthermore, both *Inc-IGSF9* and *IGSF9* were found to be significantly downregulated in CRC tissue compared to adjacent normal tissue. Our *in vitro* experiments and clinical observations corroborate that *Inc-IGSF9* and *IGSF9* follow a consistent pattern of regulation across most human tissues, are negatively regulated by the Wnt pathway, and are downregulated in CRC.

Additional studies have also pointed to the prognostic role of *Inc-IGSF9* as a tumor suppressor gene in CRC in independent cohorts. Zhang et. al. demonstrated that *Inc-IGSF9* expression was downregulated in 187-CRC patients cohort. Notably, lower *Inc-IGSF9* expression was associated with lymph node metastases, and distant metastases³¹⁴. In another study, *Inc-IGSF9* was downregulated in 219 CRC tissues compared to matched normal tissues, with only 11 samples exhibiting higher *Inc-IGSF9* expression. *In trans* overexpression of *Inc-IGSF9* repressed migration and invasion of HCT116 cells by inhibiting EMT¹²⁵.

IGSF9 protein is a member of the Immunoglobulin superfamily and appears to function as a cell adhesion molecule and only a few studies have investigated its function³¹¹. It comprises an extracellular region (737 amino acids), a transmembrane region (23 amino acids), and a 419-amino acid long cytoplasmic domain. It is predominantly expressed in the brain, so early studies focused on the role of mammalian *IGSF9* in dendrite growth and branching, and inhibitory synapse maturation in *Drosophila*^{322,323}. Moreover, *IGSF9* expression is upregulated in

gallbladder cancer, ovarian cancer, and endometrial cancer, while is downregulated in breast cancer, melanoma, and CRC³¹². Particularly in breast cancer, *IGSF9* was found to be regulated by p53 and suppressed EMT, inhibiting the migration and invasion of cancer cells by interacting with focal adhesion kinase (FAK), ultimately disrupting FAK/AKT signalling activity³¹². *IGSF9* could operate in a similar manner in CRC, but that remains unknown. A different study demonstrated that tumor cells expressing *IGSF9* suppress T-cell proliferation. More specifically, *IGSF9* binds to T cells through its extracellular domain, inhibiting their proliferation and activation³²⁴.

Considering that *lnc-IGSF9* induces intestinal differentiation and *IGSF9* is a cell adhesion molecule, we investigated their expression in the context of differentiation. Indeed, meta-analysis of single-cell RNA sequencing of intestinal epithelial cells revealed that both *lnc-IGSF9* and *IGSF9* are predominantly expressed in differentiated cells, particularly enterocytes, while their expression is minimal in ISCs. Furthermore, differentiation of CACO2 cells resulted in a significant upregulation of both *lnc-IGSF9* and *IGSF9* after 8 days. Collectively, our findings suggest that *lnc-IGSF9* and *IGSF9* promote intestinal cell differentiation.

Next, we aimed to determine whether the expression of *lnc-IGSF9* itself or the transcriptional activation of the *lnc-IGSF9* locus is responsible for regulating *IGSF9*. Through co-transduction experiments, we activated the *lnc-IGSF9* locus using CRISPR-a while simultaneously knocking down the *lnc-IGSF9* transcript. We discovered that the *lnc-IGSF9* transcript is essential for *IGSF9* upregulation. Accordingly, CFAs demonstrated that knocking down either *lnc-IGSF9* or *IGSF9*, after activating the *lnc-IGSF9* locus, rescued the observed phenotype. Therefore, we conclude that the *lnc-IGSF9* transcript regulates *IGSF9* expression to ultimately suppress CRC clonogenicity.

To further elucidate how *lnc-IGSF9* regulates the expression of *IGSF9*, we focused on its protein interactants. Immunoprecipitation followed by mass spectrometry identified several interacting nuclear proteins involved in chromatin remodeling or gene expression regulation. Among these, we focused on CDX2, an intestine-specific master regulator of intestinal differentiation which plays a crucial role in determining embryonic intestinal identity³¹⁶. CDX2 is expressed in both progenitor and differentiated cells exclusively within the adult intestines³²⁵. Ectopic overexpression of Cdx2 in mouse stomach or esophageal cells confers intestinal properties³²⁶, while its ablation disrupts intestine development, resulting in pyloric stomach-like features³²⁷. CDX2 operates as a transcription factor that primarily binds to differentiation-related genes, inducing their transcriptional activity and ultimately promoting intestinal differentiation³²⁸. Notably, CDX2 overexpression reduces CRC viability, colony formation, and migration, while low expression levels of CDX2 are associated with increased lymph node metastasis, higher tumor grade and stage, and poorer prognosis in CRC³²⁸.

The functional interaction between CDX2 and Wnt signaling activity is complex. CDX2 has been shown to negatively affect Wnt pathway activity in CRC cells by binding to the promoter of GSK-3 β and upstream enhancer of *AXIN2*, thereby upregulating their expression and facilitating β -catenin degradation³²⁹. Conversely, another study demonstrated that CDX2 can synergistically cooperate with the β -catenin/TCF4 complex to regulate Claudin-1 promoter activity³³⁰.

Our findings revealed that CDX2 binds to the promoter and intronic enhancer of *Inc-IGSF9*, although no robust binding site on the *IGSF9* promoter was identified. Overexpression of CDX2 resulted in significant upregulation of both *Inc-IGSF9* and *IGSF9*. To determine whether CDX2-mediated regulation of *IGSF9* requires *Inc-IGSF9* expression, we generated LS174 cell lines with homozygous deletion of the *Inc-IGSF9* promoter. Our results demonstrated that CDX2 requires *Inc-IGSF9* expression to activate *IGSF9* expression.

Even though we established that CDX2 upregulates *Inc-IGSF9* to regulate *IGSF9* expression and promote intestinal differentiation, the precise mechanism underlying this regulation remains unclear. Future studies employing chromatin conformation capture methodologies are necessary to characterize the chromatin architecture around the *Inc-IGSF9* and *IGSF9* genomic loci. These experiments could reveal contact sites that may explain whether intra-chromosomal bridges allow CDX2 to regulate the *IGSF9* locus. Furthermore, the dynamics between the β -catenin/TCF4 complex and CDX2 on the *Inc-IGSF9* locus should be addressed more extensively. For instance, we observed that β -catenin knockdown led to increased transcriptional activity of the *Inc-IGSF9* promoter, which may be mediated by enhanced CDX2 activation.

It is important to note that our experiments were conducted in cellular models characterized by persistent and high Wnt activity, which represses the expression of both *Inc-IGSF9* and *IGSF9*. Given that *Inc-IGSF9* is not conserved, we cannot utilize traditional mouse models to address its function in a normal, non-cancerous cellular setting. On the other hand, human intestinal organoids would serve as an optimal model to investigate the physiological roles of both *Inc-IGSF9* and *IGSF9*. These organoid models would also allow us to explore the effects of knocking down *Inc-IGSF9* in the context of its robust expression, supplementing our previous *in cis* overexpression experiments.

Conclusively, our study reveals that *Inc-IGSF9* is a direct target of the β -catenin/TCF4 complex and is negatively regulated by its activity. The produced lncRNA is localized in the nucleus of CRC cells, where it operates *in cis* to regulate the expression of its neighboring gene, *IGSF9*. Both *Inc-IGSF9* and *IGSF9* exhibit a consistent pattern of coregulation across multiple tissues, are regulated accordingly by Wnt activation/inhibition, and are downregulated in CRC tissue. Notably, their expression increases during intestinal differentiation, and *in cis* overexpression of *Inc-IGSF9* alters the expression levels of cell adhesion molecules. *Inc-IGSF9* expression is essential for regulating *IGSF9*, which in turn, contributes to the *Inc-IGSF9*-mediated reduction in CRC proliferation. Finally, CDX2, a master regulator of intestinal differentiation, binds to both the promoter and enhancer of *Inc-IGSF9*, inducing its expression. Moreover, it interacts with the *Inc-IGSF9* transcript and activates the expression of *IGSF9*. Our findings illuminate the dynamic interplay between the Wnt-mediated oncogenic program and the CDX2-related differentiation program, highlighting the role of *Inc-IGSF9* as a critical pro-differentiation gene that regulates *IGSF9* expression and intestinal homeostasis.

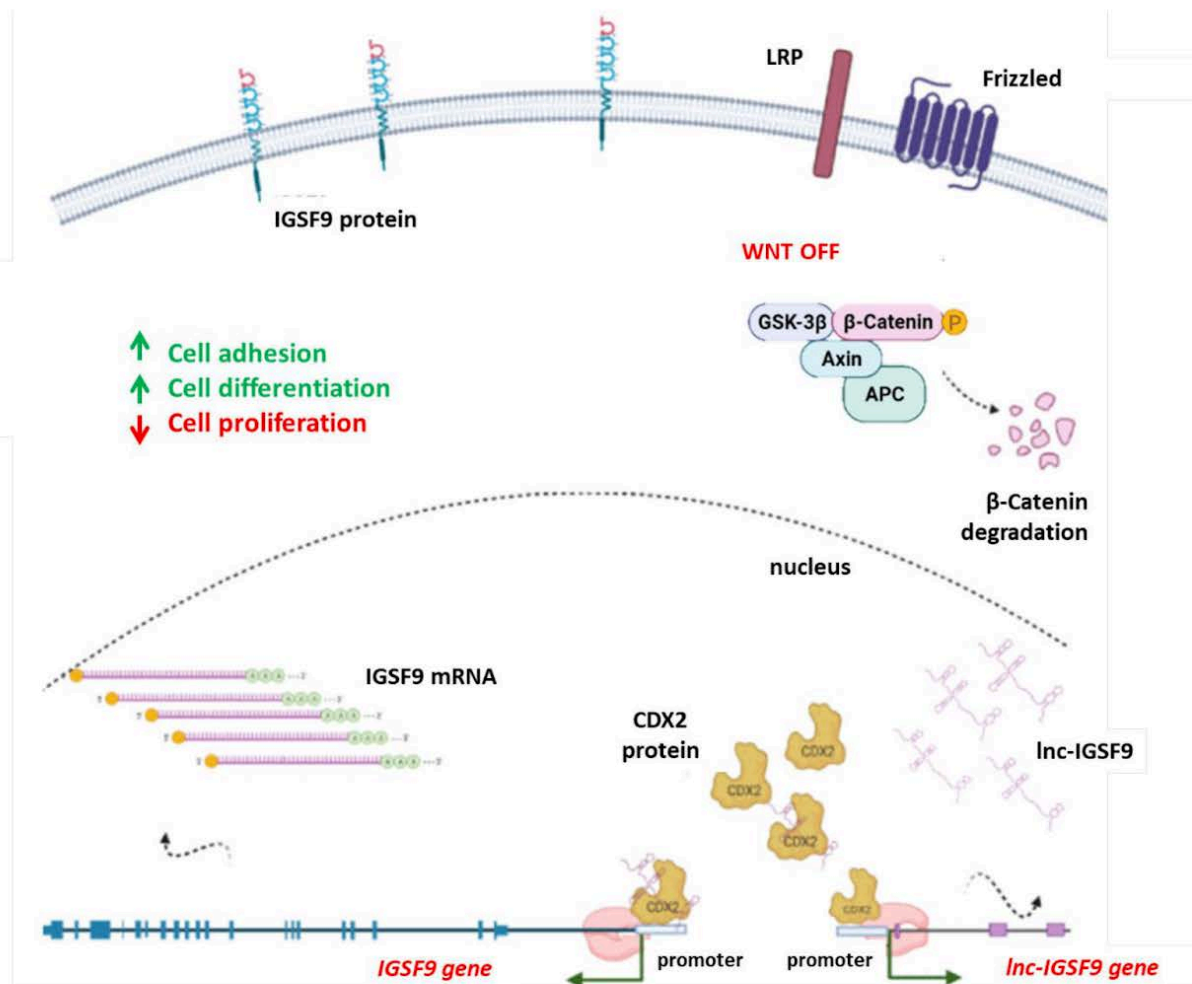


Figure 34: Graphical illustration displaying the molecular mechanism of *Inc-IGSF9* in CRC cells.

Upon Wnt pathway suppression, *Inc-IGSF9* is upregulated and regulates the expression of its neighboring gene *IGSF9*. CDX2 binds to the *Inc-IGSF9* promoter, activating its expression and interacting with the IncRNA to ultimately modulate *IGSF9* expression. The CDX2-*Inc-IGSF9*-*IGSF9* axis drives a transcriptional program that promotes intestinal differentiation and cell adhesion while suppressing carcinogenesis.

Part II

Recent advances in ribosome profiling, mass-spectrometry, and bioinformatics have revealed the existence of protein-coding smORFs within genes previously annotated as non-coding²⁴⁶. In this study, we leveraged Ribo-seq datasets from HCT116 cells to identify putative SEPs in CRC cells. Even though our approach revealed a limited number of candidates with translational potential, we successfully identified EPB-SEP, a novel microprotein encoded by the EPB41L4A-AS1 lncRNA. Our findings demonstrate that EPB-SEP localizes to mitochondria, where it plays a pivotal role in regulating mitochondrial membrane potential and modulating the proliferative capacity of cancer cells.

Our Ribo-seq analysis identified nearly 100 potentially translated lncORFs in HCT116 cells. The codon length and start codon distribution of these lncORFs were consistent with expected patterns²³³, underscoring the robustness of our analysis. While ORF-RATER is optimal for discovering sites of translation of high confidence, it is prone to a higher false negative rate²⁴⁷. We reasoned that focusing only on prominent candidates would limit our study in a few translatable lncORFs. Moreover, ORF-RATER analysis includes even non-AUG initiating smORFs, which still remain an underrepresented class of proteins. Future efforts using alternative ribosome-stalling drugs, such as harringtonine, complementary approaches like translation complex profiling³³¹, and additional bioinformatics tools like FLOSS²⁴⁸, or RiboTaper³³², could expand the list of identified smORFs. lncORFs identified by the ORF-RATER algorithm which exhibited high RRS scores were selected for manual curation. Among the selected candidates, we also included ZFAS1, a lncRNA previously shown to encode a microprotein³³³.

We demonstrated that the EPB41L4A-AS1 lncORF, initiating with a CUG codon, encodes a 17 kDa microprotein, EPB-SEP. The protein sequence does not contain any predicted protein domains, even when analyzed by AlphaFold3, which appears to overcome previous challenges posed by the small size of microproteins³¹⁷. The lncORF sequence exhibits low evolutionary conservation which has been observed in other functional smORFs that are the result of de novo gene birth²³⁵. Despite the absence of predicted protein domains or evolutionary conservation, which often suggests low translational potential, overexpression of the lncORF with its 5' UTR and a FLAG epitope confirmed EPB-SEP as a mitochondrial microprotein. Previous studies have shown that mitochondria are organelles with a significant enrichment with microproteins²⁸⁷ and that multiple smORFs produce functional mitochondrial microproteins essential for cellular physiology²⁸⁹. While we demonstrated the colocalization of EPB-SEP with mitochondrial markers through confocal microscopy, further experiments, such as through the use of proteinase K assays and sodium carbonate extraction, may help us elucidate the submitochondrial localization of EPB-SEP³³⁴.

Overexpression of EPB-SEP significantly reduced mitochondrial membrane potential, as shown by Mitotracker and TMRE staining. Additionally, proteomic analysis revealed the downregulation of ~40 mitochondrial proteins, suggesting that EPB-SEP disrupts the mitoproteome and perturbs mitochondrial physiology. Further experiments are needed to address key questions in cellular bioenergetics, including ATP production, reactive oxygen species accumulation, and calcium retention, to better understand EPB-SEP's function in cellular metabolism.

Our proteomic profiling revealed that EPB-SEP overexpression led to the downregulation of mitotic checkpoint regulators (e.g., BUB1, CCNB2), kinetochore components (e.g., NUF2, SPC25), and spindle assembly factors (e.g., NUSAP1, PIMREG). This disruption of key mitotic proteins likely underlies the observed cell cycle arrest and the reduced clonogenic potential of EPB-SEP-overexpressing cells. The overexpression of EPB-SEP or EPB41L4A-AS1 lncRNA significantly diminished cell clonogenicity, whereas knockdown of the EPB41L4A-AS1 increased cell clonogenicity in HEK293T and HCT116 cells. Even though EPB-SEP overexpression led to S- and G2-phase arrest, these cells did not undergo apoptosis, suggesting the involvement of alternative cell death pathways driven by the mitotic and metabolic stress.

Given that EPB-SEP localizes to the mitochondria, the most parsimonious scenario is that EPB-SEP overexpression induces mitochondrial dysfunction, ultimately leading to reduced cancer cell proliferation. Several microproteins have been previously proposed as critical players in regulating mitochondrial physiology and oxidative phosphorylation. For instance, the lncRNA-encoded microprotein BRAWNIN is an essential component for the CIII assembly, and its depletion leads to impaired ATP production²⁸⁷. In contrast with EPB-SEP, Mtlm overexpression leads to augmented mitochondrial membrane potential, respiration rates, and Ca²⁺ retention capacity and reduced ROS production²⁹⁰.

Our immunoprecipitation experiments further support the direct implications of the EPB-SEP in mitochondrial physiology. We identified 13 interacting proteins among which 5 are mitochondrial proteins. The other proteins were either predominantly cytoplasmic (DNAJB1, FABP5, KPRP) or nuclear (SMCHD1, SUPT16H, RCN1, SSRP1, RNF138) and their interaction with EPB-SEP could be attributed to EPB-SEP molecules located in these compartments as a byproduct of increased overexpression or spurious association with EPB-SEP because of the *in vitro*, whole-cell character of our immunoprecipitation experiments. The 5 mitochondrial proteins are implicated in different aspects of mitochondrial physiology and localize in different sub-mitochondrial compartments. For instance, NDUFB10 is a component of the CI complex of the electron transport chain³³⁵, SCO2 is necessary for cytochrome c oxidase assembly³³⁶, and CHCHD3 is a component of the MICOS complex which is critical for the maintenance of cristae junctions³³⁷. Interestingly, the anti-apoptotic protein HAX1 and CLPB have already been shown to interact to maintain mitochondrial proteostasis³²¹, pointing to the potential role of EPB-SEP in affecting the function of this complex. Verification through reverse immunoprecipitation of pulling down the interacting partners, is necessary to confirm these interactions and their physiological relevance.

Our experiments heavily rely on overexpression systems, which may not fully capture the physiological context of EPB-SEP function. To address this limitation, we are developing tools to achieve endogenous tagging of the lncORF with the FLAG epitope, using CRISPR-Cas9^{D10A} nickase-based approach³³⁸ or prime editing³³⁹. These genetic knock-in approaches will allow us to investigate the expression and function of endogenous EPB-SEP.

Many smORFs predicted to encode proteins, based on Ribo-seq data, initiate with a non-cognate start codon, with CUG being the second most efficient³⁴⁰. However, their translation rates are typically lower. Translation of non-AUG smORFs takes place either upstream of CDS as a result of leaky scanning or in response to particular stress conditions, such as heat shock, and during development²³⁷. While EPB41L4A-AS1 was found associated with the ribosomes

under normal conditions, the translation of the endogenous EPB-SEP microprotein may require stress-induced activation. Therefore, treating FLAG knock-in cell lines with various stress conditions could be essential for inducing and characterizing endogenous EPB-SEP expression.

In summary, the second part of my thesis focuses on a rare subset of lncRNAs that contain 'hidden' smORFs with protein-coding capacity. We discovered a novel microprotein with important implications in mitochondrial physiology and exerting tumor suppressive features in cancer cells. Our findings illuminate the existence of these underappreciated microproteins, which could serve as a novel pool of mitochondrial regulators and cancer biomarkers.

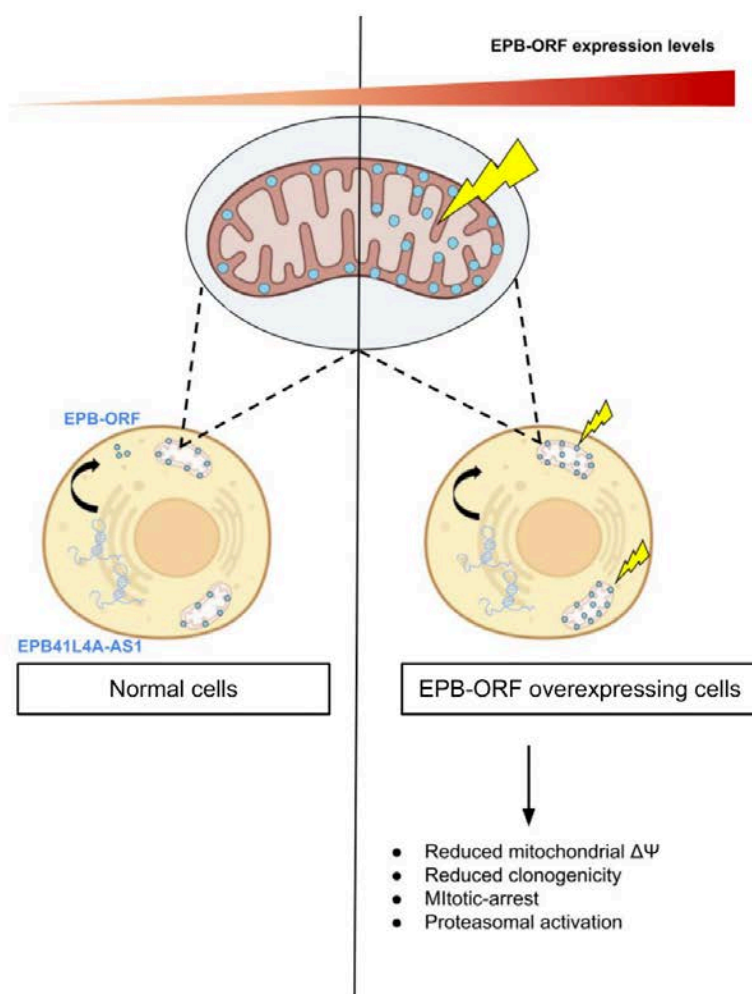


Figure 35: Graphical illustration depicting the function of EPB-ORF in cancer cells.

The EPB41L4A-AS1 lncRNA is bound by the ribosomes and produces EPB-ORF. Overexpression of EPB-ORF results in the production of a mitochondrial microprotein which impairs mitochondrial membrane potential through altering the levels of multiple mitochondrial proteins. Cells that overexpress EPB-ORF are characterized by reduced clonogenicity, and perturbations in the cell cycle.

Appendix I: Oligonucleotides that were used in Part I

Primer Name	Forward	Reverse	Application
RPLP1	AAGCAGCCGGTGTAATG TTGAGC	CATTGCAGATGAGGCTCCC AATGT	qPCR
GAPDH	ACATCAAGAAGGTGGTGA AGCAGG	TGTCGCTGTTGAAGTCAGA GGAGA	qPCR
Inc-IGSF9-F4R4	GAATGGTTGGGAGGAGGT AAAG	CTGGGCTCAAGGAATCTGA ATAG	qPCR
Inc-IGSF9-F10R10	GGTGGAGAGAATGGAGGA AAG	CCAGATCCAAAGGGAAGCT AA	qPCR
IGSF9-F1R1	CCGAGAGATGAATGTGGA TGG	GGAGAACGAAGGAAAGATG AGG	qPCR
IGSF9-F3R3	CTGGCTTTGATGGTGGTTA TCT	ACCCAGTCATGGTGCATTC	qPCR
LGR5	TCTGAATGGTGCCTCACA AATA	GACGGTTTGAGGAAGAGAT GAG	qPCR
AXIN2	CAGCAGAGGGACAGGAAT C	CAGTTTCTTTGGCTCTTTGT G	qPCR
CTNNB1	TGCAGTTCGCCTTCACTAT GGACT	GATTTGCGGGACAAAGGGC AAGAT	qPCR
MALAT1	CAGCAGTTCGTGGTGAAG ATA	GTTAGCGCTCCTTCCTTCTT	qPCR
CDX2_cloning	TACGTGAGCTACCTCCTG GACAAGGACGTG	GATCTATCACTGGGTGACG GTGGGGTTTAGCACC	PCR for cloning
CDX2	TCGGCAGCCAAGTGAAA	GATGGTGATGTAGCGACTG TAG	qPCR
FOXQ1	CCCAGGCTTCGTCTTATTT CT	GTGGAAAGGTTCCCTGA	qPCR
Inc-IGSF9 promoter	CCAGTTGTCCACAACGGT AAT	GAGGACGTCTTTCTGGGAG ATA	ChIP-qPCR
Inc-IGSF9 body	GAATGGTTGGGAGGAGGT AAAG	CTGGGCTCAAGGAATCTGA ATAG	ChIP-qPCR
Negative control-1	TGTTCTACCTTTGTTACCT GAG	CCACCACTTGAATGCTTCC	ChIP-qPCR
Negative control-2	AAGTTTGACAAGTTCAAG CACCTG	TGGCACCATGCTTCTTTAAG TC	ChIP-qPCR
Scrambled shRNA	TTAGCGACTAAACACATCA ATTCAAGAGATTG ATGTGTTTAGTCGCTATTTT	TCGAGAAAAAATAGCGACTA AACACATCAATCTCT TGAATTGATGTGTTTAGTCG	shRNA sequences

	TTC	CTAA	
CTNNB1 shRNA	TAGAAATAGTTGAAGGTTG TTTCAAGAGAACA ACCTTCAACTATTTCTTTT TTC	TCGAGAAAAAAGAAATAGT TGAAGGTTGTTCTCT TGAAACAACCTTCAACTATT TCTA	shRNA sequences
Inc-IGSF9 shRNA	TGGGAGGAGGTAAAGAGT AGTTCAAGAGACTA CTCTTTACCTCCTCCCTTT TTTC	TCGAGAAAAAAGGGAGGAG GTAAAGAGTAGTCTCT TGAAC TACTCTTACCTCCT CCCA	shRNA sequences
IGSF9 shRNA	TCGTGGAAGCTCCGAGGA AATTCAAGAGATTT CCTCGGAGCTTCCACGTT TTTTC	TCGAGAAAAAACGTGGAAG CTCCGAGGAAATCTCT TGAATTCCTCGGAGCTTC CACGA	shRNA sequences
Non targeting guide (NTG)	TTGGGACCAGGATGGGCA CCACCCGTTTAAGA GC	TTAGCTCTTAAACGGGTGG TGCCCATCCTGGTCCCA ACAAG	CRISPR-a
Inc-IGSF9 sgRNA1	TTGGGGAAAATGAGACAA GGAAGTTTAAGAGC	TTAGCTCTTAAACTTCCTTG TCTCATTTTCCCAACA AG	CRISPR-a
Inc-IGSF9 sgRNA2	TTGGGGAGGAAAGCCAAA GAAAGTTTAAGAGC	TTAGCTCTTAAACTTTCTTT GGCTTTCCTCCCAACA AG	CRISPR-a
Inc-IGSF9 promoter KO sgRNA1	CACCGGCCCCATGAAGTT TGCACCT	AAACAGGTGCAAACCTTCAT GGGGCC	CRISPR-mediated KO of Inc-IGSF9 promoter
Inc-IGSF9 promoter KO sgRNA2	CACCGAGGCTGTTACATC AGAATTG	AAACCAATTCTGATGTAACA GCCTC	CRISPR-mediated KO of Inc-IGSF9 promoter
Inc-IGSF9 promoter KO screening	CTTGCGATAGTTTGCTGAG AATG	CAGTACCATGGCTCTGCCG TG	PCR for cloning

Appendix II: Oligonucleotides that were used in Part II

Primer Name	Forward	Reverse	Application
PVT1_ORF	GAGGCCCGTCTCGCGC	CTATCACTTGTGTCATCGT CTTTGTAGTCCAGGGACCG CCAACATCC	Cloning of smORFs
LINC01578_ORF	ATGGCCGGAGAGGCAGCA	CTATCACTTGTGTCATCGT CTTTGTAGTCGCCGCGGCA GCCCCACTCC	Cloning of smORFs
EPB41L4A-AS1_ORF	CCTTTCTTCCTTTTGGTGC GAG	CTATTACTTGTGTCATCGT CTTTGTAGTCAGTTATTTTAT ACGGGGGC	Cloning of smORFs
EBLN3P_ORF	GTGTTGTCCCGGAAGTGC CTTC	CTATCACTTGTGTCATCGT CTTTGTAGTCGAGGTATCCA TGGGTTCT	Cloning of smORFs
DANCR_ORF	GCCCTTGCCAGAGTCTT CCCG	CTATCACTTGTGTCATCGT CTTTGTAGTCAGGTCGGCT GGGTT	Cloning of smORFs
LINC01234_ORF	ACATTTTCCTGGAGCGGC TGGG	CTATTACTTGTGTCATCGT CTTTGTAGTCCCTTTCAAAA CCCCTCT	Cloning of smORFs
FGD5_ORF	CCCAATACAACCCCCACTT C	CTACTACTTGTGTCATCGT CTTTGTAGTCAGCGAGAAG GCAGGG	Cloning of smORFs
LINC00460_ORF	CTAGACTCAGAGGGATGA ACGCCAG	CTATCACTTGTGTCATCGT CTTTGTAGTCTAGGTCTCG GTCCCCCAG	Cloning of smORFs
PVT1	GAATAACGGGCTCCCAGA TT	CCTGAGTCTCAAGATGCAG TAG	qPCR
LINC01578	GATGGCACACTAACCCAG ATAA	TTTCAAGAGGTCATACGCAT ACA	qPCR
EPB41L4A-AS1_In cRNA	GGATGTCCTTGGTGAGGA TTT	ACAGACTCCAGCAGGTAAT TATG	qPCR
EBLN3P	GGATACCTCTGAGCCTCTA TCT	CTGTTCTATGTCCTCCACTT CC	qPCR
ZFAS1	AGCGGGTACAGAATGGAT TT	GTCAGGAGATCGAAGGTTG TAG	qPCR
DANCR	CTGCATTCTGAACCGTTA TCT	GGGTGTAATCCACGTTTCT CAT	qPCR
LINC01234	CTTTCTGCTGGGCTACTG ATAA	CATCTCTTGGCTCTGCTCTT C	qPCR

FGD5-AS1	GGGCATCTAATTCCAGCTA CTT	GGGCGGATATGTGGCTTAA A	qPCR
LINC00460	GTGGATGAGAACGAAGGT TACG	CTTTCCCACGCTCAGTCTTT	qPCR
EPB41L4A-AS1_ORF	GCCTGTGTCCGAAGTGTC	CCTGGCATAGTCGATGATG TAAA	qPCR
RPLP1	AAGCAGCCGGTGTAATG TTGAGC	CATTGCAGATGAGGCTCCC AATGT	qPCR
GAPDH	ACATCAAGAAGGTGGTGA AGCAGG	TGTCGCTGTTGAAGTCAGA GGAGA	qPCR

Funding of the study

We acknowledge support of this work by the project “The Greek Research Infrastructure for Personalised Medicine (pMED-GR)” (MIS 5002802) which is implemented under the Action “Reinforcement of the Research and Innovation Infrastructure”, funded by the Operational Programme Competitiveness, Entrepreneurship and Innovation (NSRF 2014-2020) and co-financed by Greece and the European Union (European Regional Development Fund).

The Fleming Genomics Facility is supported by project SingleOut (HFRI-FM17C3-3780) funded by the Hellenic Foundation for Research and Innovation (H.F.R.I.) under the “1st Call for H.F.R.I. Research Projects to support Faculty members and Researchers and the procurement of high-cost research equipment” and project MIS 6004752 funded by the Regional Operational Programme 'ATTICA' (NSRF 2021-2027) and co-financed by Greece and the European Union (European Regional Development Fund).

This PhD thesis was supported by the Hellenic Foundation for Research and Innovation (HFRI) under the 3rd Call for HFRI PhD Fellowships (No 5691).

Bibliography

1. Initial sequencing and analysis of the human genome | Nature.
<https://www.nature.com/articles/35057062>.
2. The complete sequence of a human genome | Science.
<https://www.science.org/doi/10.1126/science.abj6987>.
3. Kapranov, P. *et al.* Large-scale transcriptional activity in chromosomes 21 and 22. *Science* **296**, 916–919 (2002).
4. Carninci, P. *et al.* The Transcriptional Landscape of the Mammalian Genome. *Science* **309**, 1559–1563 (2005).
5. Slack, F. J. & Chinnaiyan, A. M. The Role of Non-coding RNAs in Oncology. *Cell* **179**, 1033–1055 (2019).
6. Mattick, J. S. *et al.* Long non-coding RNAs: definitions, functions, challenges and recommendations. *Nat. Rev. Mol. Cell Biol.* **24**, 430–447 (2023).
7. Statello, L., Guo, C.-J., Chen, L.-L. & Huarte, M. Gene regulation by long non-coding RNAs and its biological functions. *Nat. Rev. Mol. Cell Biol.* **22**, 96–118 (2021).
8. Cech, T. R. & Steitz, J. A. The Noncoding RNA Revolution—Trashing Old Rules to Forge New Ones. *Cell* **157**, 77–94 (2014).
9. Pairing beyond the Seed Supports MicroRNA Targeting Specificity - PubMed.
<https://pubmed.ncbi.nlm.nih.gov/27720646/>.
10. Oliveto, S., Mancino, M., Manfrini, N. & Biffo, S. Role of microRNAs in translation regulation and cancer. *World J. Biol. Chem.* **8**, 45–56 (2017).
11. Wang, J., Chen, J. & Sen, S. microRNA as Biomarkers and Diagnostics. *J. Cell. Physiol.* **231**, 25–30 (2016).
12. GENCODE 2021 | Nucleic Acids Research | Oxford Academic.
<https://academic.oup.com/nar/article/49/D1/D916/6018430?login=true>.

13. Uszczynska-Ratajczak, B., Lagarde, J., Frankish, A., Guigó, R. & Johnson, R. Towards a complete map of the human long non-coding RNA transcriptome. *Nat. Rev. Genet.* **19**, 535–548 (2018).
14. NONCODEV5: a comprehensive annotation database for long non-coding RNAs | Nucleic Acids Research | Oxford Academic.
<https://academic.oup.com/nar/article/46/D1/D308/4616876>.
15. Guttman, M. *et al.* Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. *Nature* **458**, 223–227 (2009).
16. Khalil, A. M. *et al.* Many human large intergenic noncoding RNAs associate with chromatin-modifying complexes and affect gene expression. *Proc. Natl. Acad. Sci. U. S. A.* **106**, 11667–11672 (2009).
17. Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L. & Wold, B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat. Methods* **5**, 621–628 (2008).
18. Lin, M. F., Jungreis, I. & Kellis, M. PhyloCSF: a comparative genomics method to distinguish protein coding and non-coding regions. *Bioinformatics* **27**, i275–i282 (2011).
19. Derrien, T. *et al.* The GENCODE v7 catalog of human long noncoding RNAs: Analysis of their gene structure, evolution, and expression. *Genome Res.* **22**, 1775–1789 (2012).
20. Differentiating Protein-Coding and Noncoding RNA: Challenges and Ambiguities | PLOS Computational Biology.
<https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1000176>.
21. Ulitsky, I. & Bartel, D. P. lincRNAs: Genomics, Evolution, and Mechanisms. *Cell* **154**, 26–46 (2013).
22. Gloss, B. S. & Dinger, M. E. The specificity of long noncoding RNA expression. *Biochim. Biophys. Acta* **1859**, 16–22 (2016).
23. Melé, M. *et al.* Chromatin environment, transcriptional regulation, and splicing distinguish lincRNAs and mRNAs. *Genome Res.* **27**, 27–37 (2017).

24. Lagarde, J. *et al.* High-throughput annotation of full-length long noncoding RNAs with capture long-read sequencing. *Nat. Genet.* **49**, 1731–1740 (2017).
25. Kim, D. H. *et al.* Single-cell transcriptome analysis reveals dynamic changes in lncRNA expression during reprogramming. *Cell Stem Cell* **16**, 88–101 (2015).
26. Sarropoulos, I., Marin, R., Cardoso-Moreira, M. & Kaessmann, H. Developmental dynamics of lncRNAs across mammalian organs and species. *Nature* **571**, 510–514 (2019).
27. Jachowicz, J. W. *et al.* Xist spatially amplifies SHARP/SPEN recruitment to balance chromosome-wide silencing and specificity to the X chromosome. *Nat. Struct. Mol. Biol.* **29**, 239–249 (2022).
28. Johnsson, P., Lipovich, L., Grandér, D. & Morris, K. V. Evolutionary conservation of long noncoding RNAs; sequence, structure, function. *Biochim. Biophys. Acta* **1840**, 1063–1071 (2014).
29. Poliseno, L. *et al.* A coding-independent function of gene and pseudogene mRNAs regulates tumour biology. *Nature* **465**, 1033–1038 (2010).
30. Chen, J. L., Blasco, M. A. & Greider, C. W. Secondary structure of vertebrate telomerase RNA. *Cell* **100**, 503–514 (2000).
31. Kopp, F. & Mendell, J. T. Functional Classification and Experimental Dissection of Long Noncoding RNAs. *Cell* **172**, 393–407 (2018).
32. Ma, L., Bajic, V. B. & Zhang, Z. On the classification of long non-coding RNAs. *RNA Biol.* **10**, 924–933 (2013).
33. Schlackow, M. *et al.* Distinctive Patterns of Transcription and RNA Processing for Human lincRNAs. *Mol. Cell* **65**, 25–38 (2017).
34. Yin, Y. *et al.* U1 snRNP regulates chromatin retention of noncoding RNAs. *Nature* **580**, 147–150 (2020).
35. Rosenberg, A. B., Patwardhan, R. P., Shendure, J. & Seelig, G. Learning the Sequence Determinants of Alternative Splicing from Millions of Random Sequences. *Cell* **163**, 698–711

(2015).

36. Zuckerman, B., Ron, M., Mikl, M., Segal, E. & Ulitsky, I. Gene Architecture and Sequence Composition Underpin Selective Dependency of Nuclear Export of Long RNAs on NXF1 and the TREX Complex. *Mol. Cell* **79**, 251–267.e6 (2020).
37. Guttman, M., Russell, P., Ingolia, N. T., Weissman, J. S. & Lander, E. S. Ribosome profiling provides evidence that large non-coding RNAs do not encode proteins. *Cell* **154**, 240–251 (2013).
38. Carlevaro-Fita, J., Rahim, A., Guigó, R., Vardy, L. A. & Johnson, R. Cytoplasmic long noncoding RNAs are frequently bound to and degraded at ribosomes in human cells. *RNA* **22**, 867–882 (2016).
39. Mercer, T. R. *et al.* The human mitochondrial transcriptome. *Cell* **146**, 645–658 (2011).
40. Pauli, A. *et al.* Systematic identification of long noncoding RNAs expressed during zebrafish embryogenesis. *Genome Res.* **22**, 577–591 (2012).
41. Wu, H., Yang, L. & Chen, L.-L. The Diversity of Long Noncoding RNAs and Their Generation. *Trends Genet.* **33**, 540–552 (2017).
42. Zhang, Y. *et al.* Circular Intronic Long Noncoding RNAs. *Mol. Cell* **51**, 792–806 (2013).
43. Balas, M. M. & Johnson, A. M. Exploring the mechanisms behind long noncoding RNAs and cancer. *Non-Coding RNA Res.* **3**, 108–117 (2018).
44. Leygue, E. Steroid receptor RNA activator (SRA1): unusual bifaceted gene products with suspected relevance to breast cancer. *Nucl. Recept. Signal.* **5**, e006 (2007).
45. Zhou, M., Guo, X., Wang, M. & Qin, R. The patterns of antisense long non-coding RNAs regulating corresponding sense genes in human cancers. *J. Cancer* **12**, 1499–1506 (2021).
46. Lanzafame, M., Bianco, G., Terracciano, L. M., Ng, C. K. Y. & Piscuoglio, S. The Role of Long Non-Coding RNAs in Hepatocarcinogenesis. *Int. J. Mol. Sci.* **19**, 682 (2018).
47. S. Zibitt, M., Hartford, C. C. R. & Lal, A. Interrogating lncRNA functions via CRISPR/Cas systems. *RNA Biol.* **18**, 2097–2106.

48. Tanenbaum, M. E., Gilbert, L. A., Qi, L. S., Weissman, J. S. & Vale, R. D. A protein-tagging system for signal amplification in gene expression and fluorescence imaging. *Cell* **159**, 635–646 (2014).
49. Gilbert, L. A. *et al.* Genome-Scale CRISPR-Mediated Control of Gene Repression and Activation. *Cell* **159**, 647–661 (2014).
50. Abudayyeh, O. O. *et al.* RNA targeting with CRISPR–Cas13. *Nature* **550**, 280–284 (2017).
51. Luo, S. *et al.* Divergent lncRNAs Regulate Gene Expression and Lineage Differentiation in Pluripotent Cells. *Cell Stem Cell* **18**, 637–652 (2016).
52. Multiplexable, locus-specific targeting of long RNAs with CRISPR-Display | Nature Methods. <https://www.nature.com/articles/nmeth.3433>.
53. Piunti, A. & Shilatifard, A. The roles of Polycomb repressive complexes in mammalian development and cancer. *Nat. Rev. Mol. Cell Biol.* **22**, 326–345 (2021).
54. Trotman, J. B., Bracer, K. C. A., Cherney, R. E., Murvin, M. M. & Calabrese, J. M. The control of polycomb repressive complexes by long noncoding RNAs. *WIREs RNA* **12**, e1657 (2021).
55. Davidovich, C. & Cech, T. R. The recruitment of chromatin modifiers by long noncoding RNAs: lessons from PRC2. *RNA N. Y. N* **21**, 2007–2022 (2015).
56. Beltran, M. *et al.* The interaction of PRC2 with RNA or chromatin is mutually antagonistic. *Genome Res.* **26**, 896–907 (2016).
57. Zhao, J. *et al.* Genome-wide identification of polycomb-associated RNAs by RIP-seq. *Mol. Cell* **40**, 939–953 (2010).
58. Fiorenzano, A., Pascale, E., Patriarca, E. J., Minchiotti, G. & Fico, A. LncRNAs and PRC2: Coupled Partners in Embryonic Stem Cells. *Epigenomes* **3**, 14 (2019).
59. Wang, W. *et al.* Biological Function of Long Non-coding RNA (LncRNA) Xist. *Front. Cell Dev. Biol.* **9**, 645647 (2021).
60. da Rocha, S. T. & Heard, E. Novel players in X inactivation: insights into Xist-mediated gene

- silencing and chromosome conformation. *Nat. Struct. Mol. Biol.* **24**, 197–204 (2017).
61. Loos, F. *et al.* Xist and Tsix Transcription Dynamics Is Regulated by the X-to-Autosome Ratio and Semistable Transcriptional States. *Mol. Cell. Biol.* **36**, 2656–2667 (2016).
 62. Merry, C. R. *et al.* DNMT1-associated long non-coding RNAs regulate global gene expression and DNA methylation in colon cancer. *Hum. Mol. Genet.* **24**, 6240–6253 (2015).
 63. Di Ruscio, A. *et al.* DNMT1-interacting RNAs block gene-specific DNA methylation. *Nature* **503**, 371–376 (2013).
 64. Wang, K. C. *et al.* A long noncoding RNA maintains active chromatin to coordinate homeotic gene expression. *Nature* **472**, 120–124 (2011).
 65. Mallo, M., Wellik, D. M. & Deschamps, J. Hox Genes and Regional Patterning of the Vertebrate Body Plan. *Dev. Biol.* **344**, 7–15 (2010).
 66. Krawczyk, M. & Emerson, B. M. p50-associated COX-2 extragenic RNA (PACER) activates COX-2 gene expression by occluding repressive NF-κB complexes. *eLife* **3**, e01776 (2014).
 67. Blank-Giwojna, A., Postepska-Igielska, A. & Grummt, I. lncRNA *KHPS1* Activates a Poised Enhancer by Triplex-Dependent Recruitment of Epigenomic Regulators. *Cell Rep.* **26**, 2904-2915.e4 (2019).
 68. Postepska-Igielska, A. *et al.* LncRNA *Khps1* Regulates Expression of the Proto-oncogene *SPHK1* via Triplex-Mediated Changes in Chromatin Structure. *Mol. Cell* **60**, 626–636 (2015).
 69. Detection of RNA–DNA binding sites in long noncoding RNAs | Nucleic Acids Research | Oxford Academic. <https://academic.oup.com/nar/article/47/6/e32/5304312>.
 70. Schmitz, K.-M., Mayer, C., Postepska, A. & Grummt, I. Interaction of noncoding RNA with the rDNA promoter mediates recruitment of DNMT3b and silencing of rRNA genes. *Genes Dev.* **24**, 2264–2269 (2010).
 71. Transcription of the non-coding RNA upperhand controls Hand2 expression and heart development - PubMed. <https://pubmed.ncbi.nlm.nih.gov/27783597/>.

72. Latos, P. A. *et al.* Airn transcriptional overlap, but not its lncRNA products, induces imprinted Igf2r silencing. *Science* **338**, 1469–1472 (2012).
73. Chromatin-state discovery and genome annotation with ChromHMM | Nature Protocols.
<https://www.nature.com/articles/nprot.2017.124>.
74. Kim, Y. J., Xie, P., Cao, L., Zhang, M. Q. & Kim, T. H. Global transcriptional activity dynamics reveal functional enhancer RNAs. *Genome Res.* **28**, 1799–1811 (2018).
75. Functional roles of enhancer RNAs for oestrogen-dependent transcriptional activation | Nature. <https://www.nature.com/articles/nature12210>.
76. Rinn, J. L. *et al.* Functional Demarcation of Active and Silent Chromatin Domains in Human HOX Loci by Non-Coding RNAs. *Cell* **129**, 1311–1323 (2007).
77. Chu, C., Qu, K., Zhong, F. L., Artandi, S. E. & Chang, H. Y. Genomic maps of long noncoding RNA occupancy reveal principles of RNA-chromatin interactions. *Mol. Cell* **44**, 667–678 (2011).
78. PRC2 is dispensable for HOTAIR-mediated transcriptional repression | The EMBO Journal.
<https://www.embopress.org/doi/full/10.15252/emboj.201695335>.
79. Chang, K.-C. *et al.* MaTAR25 lncRNA regulates the Tensin1 gene to impact breast cancer progression. *Nat. Commun.* **11**, 6438 (2020).
80. Jain, A. K. *et al.* LncPRESS1 Is a p53-Regulated lncRNA that Safeguards Pluripotency by Disrupting SIRT6-Mediated De-acetylation of Histone H3K56. *Mol. Cell* **64**, 967–981 (2016).
81. Ji, P. *et al.* MALAT-1, a novel noncoding RNA, and thymosin beta4 predict metastasis and survival in early-stage non-small cell lung cancer. *Oncogene* **22**, 8031–8041 (2003).
82. Landscape of transcription in human cells | Nature.
<https://www.nature.com/articles/nature11233>.
83. Arun, G., Aggarwal, D. & Spector, D. L. MALAT1 Long Non-Coding RNA: Functional Implications. *Non-Coding RNA* **6**, 22 (2020).
84. Zhang, B. *et al.* The lncRNA Malat1 is dispensable for mouse development but its

- transcription plays a cis-regulatory role in the adult. *Cell Rep.* **2**, 111–123 (2012).
85. A Comprehensive Characterization of the Function of LincRNAs in Transcriptional Regulation Through Long-Range Chromatin Interactions | Scientific Reports.
<https://www.nature.com/articles/srep36572>.
 86. MALAT1 functions as a competing endogenous RNA to mediate Rac1 expression by sequestering miR-101b in liver fibrosis - PubMed.
<https://pubmed.ncbi.nlm.nih.gov/26697839/>.
 87. Hsa-miR-125b suppresses bladder cancer development by down-regulating oncogene SIRT7 and oncogenic long noncoding RNA MALAT1 - PubMed.
<https://pubmed.ncbi.nlm.nih.gov/24512851/>.
 88. Clemson, C. M. *et al.* An architectural role for a nuclear noncoding RNA: NEAT1 RNA is essential for the structure of paraspeckles. *Mol. Cell* **33**, 717–726 (2009).
 89. Yamazaki, T. *et al.* Functional Domains of NEAT1 Architectural lncRNA Induce Paraspeckle Assembly through Phase Separation. *Mol. Cell* **70**, 1038-1053.e7 (2018).
 90. Wu, M. *et al.* lncRNA SLERT controls phase separation of FC/DFCs to facilitate Pol I transcription. *Science* **373**, 547–555 (2021).
 91. Wang, X. *et al.* Mutual dependency between lncRNA LETN and protein NPM1 in controlling the nucleolar structure and functions sustaining cell proliferation. *Cell Res.* **31**, 664–683 (2021).
 92. Lee, S. *et al.* Noncoding RNA NORAD Regulates Genomic Stability by Sequestering PUMILIO Proteins. *Cell* **164**, 69–80 (2016).
 93. A conserved abundant cytoplasmic long noncoding RNA modulates repression by Pumilio proteins in human cells | Nature Communications.
<https://www.nature.com/articles/ncomms12209>.
 94. Munschauer, M. *et al.* The NORAD lncRNA assembles a topoisomerase complex critical for genome stability. *Nature* **561**, 132–136 (2018).

95. Grelet, S. *et al.* A regulated PNUTS mRNA to lncRNA splice switch mediates EMT and tumour progression. *Nat. Cell Biol.* **19**, 1105–1115 (2017).
96. Fatima, F. & Nawaz, M. Vesiculated Long Non-Coding RNAs: Offshore Packages Deciphering Trans-Regulation between Cells, Cancer Progression and Resistance to Therapies. *Non-Coding RNA* **3**, 10 (2017).
97. Ingolia, N. T., Brar, G. A., Rouskin, S., McGeachy, A. M. & Weissman, J. S. The ribosome profiling strategy for monitoring translation in vivo by deep sequencing of ribosome-protected mRNA fragments. *Nat. Protoc.* **7**, 1534–1550 (2012).
98. Katopodi, V. *et al.* The long non-coding RNA ROSALIND protects the mitochondrial translational machinery from oxidative damage. *Cell Death Differ.* 1–19 (2024) doi:10.1038/s41418-024-01377-4.
99. Ren, B., Guan, M.-X., Zhou, T., Cai, X. & Shan, G. Emerging functions of mitochondria-encoded noncoding RNAs. *Trends Genet.* **39**, 125–139 (2023).
100. Rackham, O. *et al.* Long noncoding RNAs are generated from the mitochondrial genome and regulated by nuclear-encoded proteins. *RNA N. Y. N* **17**, 2085–2093 (2011).
101. Noh, J. H. *et al.* HuR and GRSF1 modulate the nuclear export and mitochondrial localization of the lncRNA RMRP. *Genes Dev.* **30**, 1224–1239 (2016).
102. Sang, L. *et al.* Mitochondrial long non-coding RNA GAS5 tunes TCA metabolism in response to nutrient stress. *Nat. Metab.* **3**, 90–106 (2021).
103. Hanahan, D. & Weinberg, R. A. Hallmarks of Cancer: The Next Generation. *Cell* **144**, 646–674 (2011).
104. Hanahan, D. Hallmarks of Cancer: New Dimensions. *Cancer Discov.* **12**, 31–46 (2022).
105. Zhao, S., Zhang, X., Chen, S. & Zhang, S. Long noncoding RNAs: fine-tuners hidden in the cancer signaling network. *Cell Death Discov.* **7**, 1–10 (2021).
106. Gutschner, T. & Diederichs, S. The hallmarks of cancer: a long non-coding RNA point of view. *RNA Biol.* **9**, 703–719 (2012).

107. Huarte, M. The emerging role of lncRNAs in cancer. *Nat. Med.* **21**, 1253–1261 (2015).
108. Gong, J., Liu, W., Zhang, J., Miao, X. & Guo, A.-Y. lncRNASNP: a database of SNPs in lncRNAs and their potential functions in human and mouse. *Nucleic Acids Res.* **43**, D181–D186 (2015).
109. Ahmad, M., Weiswald, L.-B., Poulain, L., Denoyelle, C. & Meryet-Figuere, M. Involvement of lncRNAs in cancer cells migration, invasion and metastasis: cytoskeleton and ECM crosstalk. *J. Exp. Clin. Cancer Res.* **42**, 173 (2023).
110. Bhan, A., Soleimani, M. & Mandal, S. S. Long non-coding RNA (lncRNA) and cancer: a new paradigm. *Cancer Res.* **77**, 3965–3981 (2017).
111. Jiang, M.-C., Ni, J.-J., Cui, W.-Y., Wang, B.-Y. & Zhuo, W. Emerging roles of lncRNA in cancer and therapeutic opportunities. *Am. J. Cancer Res.* **9**, 1354–1366 (2019).
112. Marín-Béjar, O. *et al.* Pint lincRNA connects the p53 pathway with epigenetic silencing by the Polycomb repressive complex 2. *Genome Biol.* **14**, R104 (2013).
113. Léveillé, N. *et al.* Genome-wide profiling of p53-regulated enhancer RNAs uncovers a subset of enhancers controlled by a lncRNA. *Nat. Commun.* **6**, 6520 (2015).
114. Activation of p53 by MEG3 non-coding RNA - PubMed.
<https://pubmed.ncbi.nlm.nih.gov/17569660/>.
115. Li, Z.-X. *et al.* MALAT1: a potential biomarker in cancer. *Cancer Manag. Res.* **10**, 6757–6768 (2018).
116. Lemos, A. E. G., Matos, A. da R., Ferreira, L. B. & Gimba, E. R. P. The long non-coding RNA PCA3 : an update of its functions and clinical applications as a biomarker in prostate cancer. *Oncotarget* **10**, 6589–6603 (2019).
117. Chevli, K. K. *et al.* Urinary PCA3 as a predictor of prostate cancer in a cohort of 3,073 men undergoing initial prostate biopsy. *J. Urol.* **191**, 1743–1748 (2014).
118. Hessels, D. & Schalken, J. A. The use of PCA3 in the diagnosis of prostate cancer. *Nat. Rev. Urol.* **6**, 255–261 (2009).

119. Winkle, M., El-Daly, S. M., Fabbri, M. & Calin, G. A. Noncoding RNA therapeutics — challenges and potential solutions. *Nat. Rev. Drug Discov.* **20**, 629–651 (2021).
120. Ito, M., Miyata, Y. & Okada, M. Current clinical trials with non-coding RNA-based therapeutics in malignant diseases: A systematic review. *Transl. Oncol.* **31**, 101634 (2023).
121. Jiang, S. *et al.* New Sights Into Long Non-Coding RNA LINC01133 in Cancer. *Front. Oncol.* **12**, (2022).
122. Ghafouri-Fard, S., Khoshbakht, T., Hussien, B. M., Taheri, M. & Mokhtari, M. A review on the role of LINC01133 in cancers. *Cancer Cell Int.* **22**, 270 (2022).
123. Yin, D. *et al.* LINC01133 promotes hepatocellular carcinoma progression by sponging miR-199a-5p and activating annexin A2. *Clin. Transl. Med.* **11**, e409 (2021).
124. Tu, Z., Schmöllerl, J., Cuiffo, B. G. & Karnoub, A. E. Microenvironmental Regulation of Long Noncoding RNA LINC01133 Promotes Cancer Stem Cell-Like Phenotypic Traits in Triple-Negative Breast Cancers. *Stem Cells Dayt. Ohio* **37**, 1281–1292 (2019).
125. Kong, J. *et al.* Long non-coding RNA LINC01133 inhibits epithelial–mesenchymal transition and metastasis in colorectal cancer by interacting with SRSF6. *Cancer Lett.* **380**, 476–484 (2016).
126. Yang, X.-Z. *et al.* LINC01133 as ceRNA inhibits gastric cancer progression by sponging miR-106a-3p to regulate APC expression and the Wnt/ β -catenin pathway. *Mol. Cancer* **17**, 126 (2018).
127. Wang, S., Chen, J., Li, P. & Chen, Y. LINC01133 can induce acquired ferroptosis resistance by enhancing the FSP1 mRNA stability through forming the LINC01133-FUS-FSP1 complex. *Cell Death Dis.* **14**, 767 (2023).
128. Yang, Y. *et al.* LINC01133 promotes pancreatic ductal adenocarcinoma epithelial–mesenchymal transition mediated by SPP1 through binding to Arp3. *Cell Death Dis.* **15**, 492 (2024).
129. Song, Z. *et al.* LINC01133 inhibits breast cancer invasion and metastasis by negatively

- regulating SOX4 expression through EZH2. *J. Cell. Mol. Med.* **23**, 7554 (2019).
130. Zang, C. *et al.* Long non-coding RNA LINC01133 represses KLF2, P21 and E-cadherin transcription through binding with EZH2, LSD1 in non small cell lung cancer. *Oncotarget* **7**, 11696–11707 (2016).
 131. Weng, Y.-C., Ma, J., Zhang, J. & Wang, J.-C. Long non-coding RNA LINC01133 silencing exerts antioncogenic effect in pancreatic cancer through the methylation of DKK1 promoter and the activation of Wnt signaling pathway. *Cancer Biol. Ther.* **20**, 368 (2018).
 132. Tumor-derived exosomal long noncoding RNA LINC01133, regulated by Periostin, contributes to pancreatic ductal adenocarcinoma epithelial-mesenchymal transition through the Wnt/ β -catenin pathway by silencing AXIN2 | Oncogene.
<https://www.nature.com/articles/s41388-021-01762-0>.
 133. Liao, M. *et al.* LncRNA EPB41L4A-AS1 regulates glycolysis and glutaminolysis by mediating nucleolar translocation of HDAC2. *EBioMedicine* **41**, 200–213 (2019).
 134. Yabuta, N. *et al.* Isolation and characterization of the TIGA genes, whose transcripts are induced by growth arrest. *Nucleic Acids Res.* **34**, 4878–4892 (2006).
 135. Zhu, Y. *et al.* Overexpression of lncRNA EPB41L4A-AS1 Induces Metabolic Reprogramming in Trophoblast Cells and Placenta Tissue of Miscarriage. *Mol. Ther. - Nucleic Acids* **18**, 518–532 (2019).
 136. Liao, W. *et al.* Persistent high glucose induced EPB41L4A-AS1 inhibits glucose uptake via GCN5 mediating crotonylation and acetylation of histones and non-histones. *Clin. Transl. Med.* **12**, e699 (2022).
 137. Chai, W. *et al.* Elevated exosome-transferrable lncRNA EPB41L4A-AS1 in CD56^{bright} NK cells is responsible for the impaired NK function in neuroblastoma patients by suppressing cell glycolysis. *Clin. Immunol.* **250**, 109322 (2023).
 138. Hegre, S. A. *et al.* Joint changes in RNA, RNA polymerase II, and promoter activity through the cell cycle identify non-coding RNAs involved in proliferation. *Sci. Rep.* **11**, 18952

- (2021).
139. Bin, J. *et al.* Long noncoding RNA EPB41L4A-AS1 functions as an oncogene by regulating the Rho/ROCK pathway in colorectal cancer. *J. Cell. Physiol.* **236**, 523–535 (2021).
140. Ma, J. *et al.* Downregulation of lncRNA EPB41L4A-AS1 promotes gastric cancer cell proliferation, migration and invasion. *BMC Gastroenterol.* **24**, 136 (2024).
141. Flanagan, D. J., Austin, C. R., Vincan, E. & Phesse, T. J. Wnt Signalling in Gastrointestinal Epithelial Stem Cells. *Genes* **9**, 178 (2018).
142. Krausova, M. & Korinek, V. Wnt signaling in adult intestinal stem cells and cancer. *Cell. Signal.* **26**, 570–579 (2014).
143. Okumura, R. & Takeda, K. Roles of intestinal epithelial cells in the maintenance of gut homeostasis. *Exp. Mol. Med.* **49**, e338–e338 (2017).
144. Flier, L. G. van der & Clevers, H. Stem Cells, Self-Renewal, and Differentiation in the Intestinal Epithelium. *Annu. Rev. Physiol.* **71**, 241–260 (2009).
145. Reya, T. & Clevers, H. Wnt signalling in stem cells and cancer. *Nature* **434**, 843–850 (2005).
146. Gustafsson, J. K. & Johansson, M. E. V. The role of goblet cells and mucus in intestinal homeostasis. *Nat. Rev. Gastroenterol. Hepatol.* **19**, 785–803 (2022).
147. Lueschow, S. R. & McElroy, S. J. The Paneth Cell: The Curator and Defender of the Immature Small Intestine. *Front. Immunol.* **11**, (2020).
148. Gassler, N. Paneth cells in intestinal physiology and pathophysiology. *World J. Gastrointest. Pathophysiol.* **8**, 150 (2017).
149. Gribble, F. M. & Reimann, F. Function and mechanisms of enteroendocrine cells and gut hormones in metabolism. *Nat. Rev. Endocrinol.* **15**, 226–237 (2019).
150. van de Wetering, M. *et al.* The beta-catenin/TCF-4 complex imposes a crypt progenitor phenotype on colorectal cancer cells. *Cell* **111**, 241–250 (2002).

151. Sansom, O. J. *et al.* Loss of Apc in vivo immediately perturbs Wnt signaling, differentiation, and migration. *Genes Dev.* **18**, 1385–1390 (2004).
152. Kim, K.-A. *et al.* Mitogenic Influence of Human R-Spondin1 on the Intestinal Epithelium. *Science* **309**, 1256–1259 (2005).
153. Durand, A. *et al.* Functional intestinal stem cells after Paneth cell ablation induced by the loss of transcription factor Math1 (Atoh1). *Proc. Natl. Acad. Sci. U. S. A.* **109**, 8965–8970 (2012).
154. Fre, S. *et al.* Notch signals control the fate of immature progenitor cells in the intestine. *Nature* **435**, 964–968 (2005).
155. Stanger, B. Z., Datar, R., Murtaugh, L. C. & Melton, D. A. Direct regulation of intestinal fate by Notch. *Proc. Natl. Acad. Sci. U. S. A.* **102**, 12443–12448 (2005).
156. He, X. C. *et al.* BMP signaling inhibits intestinal stem cell self-renewal through suppression of Wnt– β -catenin signaling. *Nat. Genet.* **36**, 1117–1121 (2004).
157. van Dop, W. A. *et al.* Depletion of the colonic epithelial precursor cell compartment upon conditional activation of the hedgehog pathway. *Gastroenterology* **136**, 2195–2203.e1–7 (2009).
158. Potten, C. S., Kovacs, L. & Hamilton, E. Continuous labelling studies on mouse skin and intestine. *Cell Tissue Kinet.* **7**, 271–283 (1974).
159. Potten, C. S. Extreme sensitivity of some intestinal crypt cells to X and gamma irradiation. *Nature* **269**, 518–521 (1977).
160. Barker, N. *et al.* Identification of stem cells in small intestine and colon by marker gene Lgr5. *Nature* **449**, 1003–1007 (2007).
161. Rees, W. D., Tandun, R., Yau, E., Zachos, N. C. & Steiner, T. S. Regenerative Intestinal Stem Cells Induced by Acute and Chronic Injury: The Saving Grace of the Epithelium? *Front. Cell Dev. Biol.* **8**, 583919 (2020).
162. Schubert, M. & Holland, L. Z. The Wnt Gene Family and the Evolutionary Conservation

- of Wnt Expression. in *Madame Curie Bioscience Database [Internet]* (Landes Bioscience, 2013).
163. Nusse, R. & Clevers, H. Wnt/ β -Catenin Signaling, Disease, and Emerging Therapeutic Modalities. *Cell* **169**, 985–999 (2017).
164. MacDonald, B. T., Tamai, K. & He, X. Wnt/beta-catenin signaling: components, mechanisms, and diseases. *Dev. Cell* **17**, 9–26 (2009).
165. Liu, J. *et al.* Wnt/ β -catenin signalling: function, biological mechanisms, and therapeutic opportunities. *Signal Transduct. Target. Ther.* **7**, 1–23 (2022).
166. Willert, K. & Nusse, R. Wnt Proteins. *Cold Spring Harb. Perspect. Biol.* **4**, a007864 (2012).
167. Willert, K. *et al.* Wnt proteins are lipid-modified and can act as stem cell growth factors. *Nature* **423**, 448–452 (2003).
168. Hausmann, G., Bänziger, C. & Basler, K. Helping Wingless take flight: how WNT proteins are secreted. *Nat. Rev. Mol. Cell Biol.* **8**, 331–336 (2007).
169. J, Y. *et al.* WLS retrograde transport to the endoplasmic reticulum during Wnt secretion. *Dev. Cell* **29**, (2014).
170. Janda, C. Y., Waghray, D., Levin, A. M., Thomas, C. & Garcia, K. C. Structural basis of Wnt recognition by Frizzled. *Science* **337**, 59–64 (2012).
171. Kazanskaya, O. *et al.* R-Spondin2 is a secreted activator of Wnt/beta-catenin signaling and is required for *Xenopus* myogenesis. *Dev. Cell* **7**, 525–534 (2004).
172. R-spondins function as ligands of the orphan receptors LGR4 and LGR5 to regulate Wnt/ β -catenin signaling | PNAS. <https://www.pnas.org/doi/10.1073/pnas.1106083108>.
173. W, de L. *et al.* Lgr5 homologues associate with Wnt receptors and mediate R-spondin signalling. *Nature* **476**, (2011).
174. Hao, H.-X. *et al.* ZNRF3 promotes Wnt receptor turnover in an R-spondin-sensitive manner. *Nature* **485**, 195–200 (2012).

175. Chang, T.-H. *et al.* Structure and functional properties of Norrin mimic Wnt for signalling with Frizzled4, Lrp5/6, and proteoglycan. *eLife* **4**, e06554 (2015).
176. Ke, J. *et al.* Structure and function of Norrin in assembly and activation of a Frizzled 4–Lrp5/6 complex. *Genes Dev.* **27**, 2305–2319 (2013).
177. Chong, J. M., Uren, A., Rubin, J. S. & Speicher, D. W. Disulfide bond assignments of secreted Frizzled-related protein-1 provide insights about Frizzled homology and netrin modules. *J. Biol. Chem.* **277**, 5134–5144 (2002).
178. Malinauskas, T., Aricescu, A. R., Lu, W., Siebold, C. & Jones, E. Y. Modular mechanism of Wnt signaling inhibition by Wnt inhibitory factor 1. *Nat. Struct. Mol. Biol.* **18**, 886–893 (2011).
179. A Frizzled-Like Cysteine-Rich Domain in Glypican-3 Mediates Wnt Binding and Regulates Hepatocellular Carcinoma Tumor Growth in Mice - PubMed.
<https://pubmed.ncbi.nlm.nih.gov/30963603/>.
180. Zhang, X. *et al.* Characterization of Tiki, a New Family of Wnt-specific Metalloproteases. *J. Biol. Chem.* **291**, 2435–2443 (2016).
181. Kakugawa, S. *et al.* Notum deacylates Wnt proteins to suppress signalling activity. *Nature* **519**, 187–192 (2015).
182. Koo, B.-K. *et al.* Tumour suppressor RNF43 is a stem-cell E3 ligase that induces endocytosis of Wnt receptors. *Nature* **488**, 665–669 (2012).
183. Yamamoto, A., Nagano, T., Takehara, S., Hibi, M. & Aizawa, S. Shisa promotes head formation through the inhibition of receptor protein maturation for the caudalizing factors, Wnt and FGF. *Cell* **120**, 223–235 (2005).
184. Azzolin, L. *et al.* YAP/TAZ incorporation in the β -catenin destruction complex orchestrates the Wnt response. *Cell* **158**, 157–170 (2014).
185. Morris, S.-A. L. & Huang, S. Crosstalk of the Wnt/ β -catenin pathway with other pathways in cancer cells. *Genes Dis.* **3**, 41 (2016).

186. Tamai, K. *et al.* A mechanism for Wnt coreceptor activation. *Mol. Cell* **13**, 149–156 (2004).
187. Structural basis of GSK-3 inhibition by N-terminal phosphorylation and by the Wnt receptor LRP6 - PubMed. <https://pubmed.ncbi.nlm.nih.gov/24642411/>.
188. Wnt signaling through inhibition of β -catenin degradation in an intact Axin1 complex - PubMed. <https://pubmed.ncbi.nlm.nih.gov/22682247/>.
189. Hatzis, P. *et al.* Genome-wide pattern of TCF7L2/TCF4 chromatin occupancy in colorectal cancer cells. *Mol. Cell. Biol.* **28**, 2732–2744 (2008).
190. Doumpas, N. *et al.* TCF/LEF dependent and independent transcriptional regulation of Wnt/ β -catenin target genes. *EMBO J.* **38**, e98873 (2018).
191. Atcha, F. A. *et al.* A Unique DNA Binding Domain Converts T-Cell Factors into Strong Wnt Effectors. *Mol. Cell. Biol.* **27**, 8352 (2007).
192. Arce, L., Yokoyama, N. N. & Waterman, M. L. Diversity of LEF/TCF action in development and disease. *Oncogene* **25**, 7492–7504 (2006).
193. Najdi, R. *et al.* A Wnt-kinase network alters nuclear localization of TCF-1 in colon cancer. *Oncogene* **28**, 4133 (2009).
194. Li, F.-Q., Mofunanya, A., Harris, K. & Takemaru, K.-I. Chibby cooperates with 14-3-3 to regulate beta-catenin subcellular distribution and signaling activity. *J. Cell Biol.* **181**, 1141–1154 (2008).
195. Tago, K. *et al.* Inhibition of Wnt signaling by ICAT, a novel β -catenin-interacting protein. *Genes Dev.* **14**, 1741 (2000).
196. Van Der Flier, L. G. *et al.* Transcription Factor Achaete Scute-Like 2 Controls Intestinal Stem Cell Fate. *Cell* **136**, 903–912 (2009).
197. Schuijers, J. *et al.* Ascl2 Acts as an R-spondin/Wnt-Responsive Switch to Control Stemness in Intestinal Crypts. *Cell Stem Cell* **16**, 158–170 (2015).
198. Jho, E. *et al.* Wnt/ β -Catenin/Tcf Signaling Induces the Transcription of Axin2, a Negative

- Regulator of the Signaling Pathway. *Mol. Cell. Biol.* **22**, 1172 (2002).
199. He, T. C. *et al.* Identification of c-MYC as a target of the APC pathway. *Science* **281**, 1509–1512 (1998).
 200. Rennoll, S. & Yochum, G. Regulation of MYC gene expression by aberrant Wnt/ β -catenin signaling in colorectal cancer. *World J. Biol. Chem.* **6**, 290 (2015).
 201. Hossain, M. S. *et al.* Colorectal Cancer: A Review of Carcinogenesis, Global Epidemiology, Current Challenges, Risk Factors, Preventive and Treatment Strategies. *Cancers* **14**, 1732 (2022).
 202. Winawer, S. J. *et al.* Colorectal cancer screening: Clinical guidelines and rationale. *Gastroenterology* **112**, 594–642 (1997).
 203. Rawla, P., Sunkara, T. & Barsouk, A. Epidemiology of colorectal cancer: incidence, mortality, survival, and risk factors. *Przegląd Gastroenterol.* **14**, 89 (2019).
 204. Hervieu, C., Christou, N., Battu, S. & Mathonnet, M. The Role of Cancer Stem Cells in Colorectal Cancer: From the Basics to Novel Clinical Trials. *Cancers* **13**, 1092 (2021).
 205. Munro, M. J., Wickremesekera, S. K., Peng, L., Tan, S. T. & Itinteang, T. Cancer stem cells in colorectal cancer: a review. *J. Clin. Pathol.* **71**, 110–116 (2018).
 206. Yu, F. *et al.* Wnt/ β -catenin signaling in cancers and targeted therapies. *Signal Transduct. Target. Ther.* **6**, 1–24 (2021).
 207. Zhan, T., Rindtorff, N. & Boutros, M. Wnt signaling in cancer. *Oncogene* **36**, 1461–1473 (2017).
 208. Zhang, Y. & Wang, X. Targeting the Wnt/ β -catenin signaling pathway in cancer. *J. Hematol. Oncol.* **13**, 165 (2020).
 209. Rowan, A. J. *et al.* APC mutations in sporadic colorectal tumors: A mutational ‘hotspot’ and interdependence of the ‘two hits’. *Proc. Natl. Acad. Sci. U. S. A.* **97**, 3352–3357 (2000).
 210. Kinzler, K. W. & Vogelstein, B. Lessons from Hereditary Colorectal Cancer. *Cell* **87**, 159–170 (1996).

211. Zhang, L. & Shay, J. W. Multiple Roles of APC and its Therapeutic Implications in Colorectal Cancer. *JNCI J. Natl. Cancer Inst.* **109**, djw332 (2017).
212. Aoki, K. *et al.* Chromosomal instability by β -catenin/TCF transcription in APC or β -catenin mutant cells. *Oncogene* **26**, 3511–3520 (2007).
213. Christie, M. *et al.* Different APC genotypes in proximal and distal sporadic colorectal cancers suggest distinct WNT/ β -catenin signalling thresholds for tumourigenesis. *Oncogene* **32**, 4675–4682 (2013).
214. Clements, W. M. *et al.* β -Catenin Mutation Is a Frequent Cause of Wnt Pathway Activation in Gastric Cancer1. *Cancer Res.* **62**, 3503–3506 (2002).
215. Liu, W. *et al.* Mutations in AXIN2 cause colorectal cancer with defective mismatch repair by activating β -catenin/TCF signalling. *Nat. Genet.* **26**, 146–147 (2000).
216. Zarkou, V., Galaras, A., Giakountis, A. & Hatzis, P. Crosstalk mechanisms between the WNT signaling pathway and long non-coding RNAs. *Non-Coding RNA Res.* **3**, 42–53 (2018).
217. Astaneh, M. & Ghafouri-Fard, S. Long non-coding RNAs as regulators of Wnt/ β catenin pathway. *Gene Rep.* **16**, 100404 (2019).
218. Ling, H. *et al.* CCAT2, a novel noncoding RNA mapping to 8q24, underlies metastatic progression and chromosomal instability in colon cancer. *Genome Res.* **23**, 1446–1461 (2013).
219. ASBEL–TCF3 complex is required for the tumorigenicity of colorectal cancer cells | PNAS. <https://www.pnas.org/doi/10.1073/pnas.1605938113>.
220. Di Cecilia, S. *et al.* RBM5-AS1 Is Critical for Self-Renewal of Colon Cancer Stem-like Cells. *Cancer Res.* **76**, 5615–5627 (2016).
221. Ma, Y. *et al.* Long non-coding RNA CCAL regulates colorectal cancer progression by activating Wnt/ β -catenin signalling pathway via suppression of activator protein 2 α . *Gut* **65**, 1494–1504 (2016).

222. Zhang, Z. *et al.* Long non-coding RNA CASC11 interacts with hnRNP-K and activates the WNT/ β -catenin pathway to promote growth and metastasis in colorectal cancer. *Cancer Lett.* **376**, 62–73 (2016).
223. Kawasaki, Y. *et al.* MYU, a Target lncRNA for Wnt/c-Myc Signaling, Mediates Induction of CDK6 to Promote Cell Cycle Progression. *Cell Rep.* **16**, 2554–2564 (2016).
224. A long non-coding RNA targets microRNA miR-34a to regulate colon cancer stem cell asymmetric division - PubMed. <https://pubmed.ncbi.nlm.nih.gov/27077950/>.
225. Wang, Y. *et al.* The long noncoding RNA lncTCF7 promotes self-renewal of human liver cancer stem cells through activation of Wnt signaling. *Cell Stem Cell* **16**, 413–425 (2015).
226. Semenov, M. V., Zhang, X. & He, X. DKK1 Antagonizes Wnt Signaling without Promotion of LRP6 Internalization and Degradation. *J. Biol. Chem.* **283**, 21427 (2008).
227. Hu, P. *et al.* NBAT1 suppresses breast cancer metastasis by regulating DKK1 via PRC2. *Oncotarget* **6**, 32410–32425 (2015).
228. Giakountis, A. *et al.* A Positive Regulatory Loop between a Wnt-Regulated Non-coding RNA and ASCL2 Controls Intestinal Stem Cell Fate. *Cell Rep.* **15**, 2588–2596 (2016).
229. Cha, Y. H. *et al.* MiRNA-34 intrinsically links p53 tumor suppressor and Wnt signaling. *Cell Cycle Georget. Tex* **11**, 1273–1281 (2012).
230. Fang, C. *et al.* Long non-coding RNA HNF1A-AS1 mediated repression of miR-34a/SIRT1/p53 feedback loop promotes the metastatic progression of colon cancer by functioning as a competing endogenous RNA. *Cancer Lett.* **410**, 50–62 (2017).
231. Guo, C.-J. *et al.* Distinct Processing of lncRNAs Contributes to Non-conserved Functions in Stem Cells. *Cell* **181**, 621-636.e22 (2020).
232. Zhang, M. *et al.* The lncRNA NEAT1 activates Wnt/ β -catenin signaling and promotes colorectal cancer progression via interacting with DDX5. *J. Hematol. Oncol. J Hematol Oncol* **11**, 113 (2018).
233. Couso, J.-P. & Patraquim, P. Classification and function of small open reading frames.

- Nat. Rev. Mol. Cell Biol.* **18**, 575–589 (2017).
234. Mackowiak, S. D. *et al.* Extensive identification and analysis of conserved small ORFs in animals. *Genome Biol.* **16**, 179 (2015).
235. Vakirlis, N., Vance, Z., Duggan, K. M. & McLysaght, A. De novo birth of functional microproteins in the human lineage. *Cell Rep.* **41**, 111808 (2022).
236. Martinez, T. F. *et al.* Accurate annotation of human protein-coding small open reading frames. *Nat. Chem. Biol.* **16**, 458–468 (2020).
237. Andreev, D. E. *et al.* Non-AUG translation initiation in mammals. *Genome Biol.* **23**, 111 (2022).
238. Chen, J. *et al.* Pervasive functional translation of non-canonical human open reading frames. *Science* **367**, 1140–1146 (2020).
239. Aspden, J. L. *et al.* Extensive translation of small Open Reading Frames revealed by Poly-Ribo-Seq. *eLife* **3**, e03528 (2014).
240. Bazzini, A. A. *et al.* Identification of small ORFs in vertebrates using ribosome footprinting and evolutionary conservation. *EMBO J.* **33**, 981–993 (2014).
241. Mudge, J. M. *et al.* Standardized annotation of translated open reading frames. *Nat. Biotechnol.* **40**, 994–999 (2022).
242. Ingolia, N. T. Ribosome profiling: new views of translation, from single codons to genome scale. *Nat. Rev. Genet.* **15**, 205–213 (2014).
243. Ingolia, N. T., Ghaemmighami, S., Newman, J. R. S. & Weissman, J. S. Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. *Science* **324**, 218–223 (2009).
244. Calviello, L. & Ohler, U. Beyond Read-Counts: Ribo-seq Data Analysis to Understand the Functions of the Transcriptome. *Trends Genet.* **33**, 728–744 (2017).
245. Ruiz-Orera, J., Messeguer, X., Subirana, J. A. & Alba, M. M. Long non-coding RNAs as a source of new peptides. *eLife* **3**, e03523 (2014).

246. Prensner, J. R. *et al.* What Can Ribo-Seq and Proteomics Tell Us about the Non-Canonical Proteome? <http://biorxiv.org/lookup/doi/10.1101/2023.05.16.541049> (2023) doi:10.1101/2023.05.16.541049.
247. Fields, A. P. *et al.* A regression-based analysis of ribosome-profiling data reveals a conserved complexity to mammalian translation. *Mol. Cell* **60**, 816–827 (2015).
248. Ingolia, N. T. *et al.* Ribosome Profiling Reveals Pervasive Translation Outside of Annotated Protein-Coding Genes. *Cell Rep.* **8**, 1365–1379 (2014).
249. Verheggen, K. *et al.* Noncoding after All: Biases in Proteomics Data Do Not Explain Observed Absence of lncRNA Translation Products. *J. Proteome Res.* **16**, 2508–2515 (2017).
250. Bogaert, A. *et al.* Limited Evidence for Protein Products of Noncoding Transcripts in the HEK293T Cellular Cytosol. *Mol. Cell. Proteomics MCP* **21**, 100264 (2022).
251. Omenn, G. S., Lane, L., Lundberg, E. K., Overall, C. M. & Deutsch, E. W. Progress on the HUPO Draft Human Proteome: 2017 Metrics of the Human Proteome Project. *J. Proteome Res.* **16**, 4281–4287 (2017).
252. Loose, C. R., Langer, R. S. & Stephanopoulos, G. N. Optimization of protein fusion partner length for maximizing in vitro translation of peptides. *Biotechnol. Prog.* **23**, 444–451 (2007).
253. Carvunis, A.-R. *et al.* Proto-genes and de novo gene birth. *Nature* **487**, 370–374 (2012).
254. Khitun, A. & Slavoff, S. A. Proteomic detection and validation of translated small open reading frames. *Curr. Protoc. Chem. Biol.* **11**, e77 (2019).
255. Valdivia-Francia, F. & Sandoel, A. No country for old methods: New tools for studying microproteins. *iScience* **27**, 108972 (2024).
256. Prensner, J. R. *et al.* Noncanonical open reading frames encode functional proteins essential for cancer cell survival. *Nat. Biotechnol.* **39**, 697–704 (2021).
257. Li, Y. *et al.* SmProt: A Reliable Repository with Comprehensive Annotation of Small

- Proteins Identified from Ribosome Profiling. *Genomics Proteomics Bioinformatics* **19**, 602–610 (2021).
258. Brunet, M. A. *et al.* OpenProt: a more comprehensive guide to explore eukaryotic coding potential and proteomes. *Nucleic Acids Res.* **47**, D403–D410 (2019).
259. Kute, P. M., Soukariéh, O., Tjeldnes, H., Trégouët, D.-A. & Valen, E. Small Open Reading Frames, How to Find Them and Determine Their Function. *Front. Genet.* **12**, (2022).
260. Calvo, S. E., Pagliarini, D. J. & Mootha, V. K. Upstream open reading frames cause widespread reduction of protein expression and are polymorphic among humans. *Proc. Natl. Acad. Sci. U. S. A.* **106**, 7507–7512 (2009).
261. Silva, J., Fernandes, R. & Romão, L. Translational Regulation by Upstream Open Reading Frames and Human Diseases. *Adv. Exp. Med. Biol.* **1157**, 99–116 (2019).
262. Hinnebusch, A. G., Ivanov, I. P. & Sonenberg, N. Translational control by 5'-untranslated regions of eukaryotic mRNAs. *Science* **352**, 1413–1416 (2016).
263. Young, S. K., Willy, J. A., Wu, C., Sachs, M. S. & Wek, R. C. Ribosome Reinitiation Directs Gene-specific Translation and Regulates the Integrated Stress Response. *J. Biol. Chem.* **290**, 28257–28271 (2015).
264. Palam, L. R., Baird, T. D. & Wek, R. C. Phosphorylation of eIF2 facilitates ribosomal bypass of an inhibitory upstream ORF to enhance CHOP translation. *J. Biol. Chem.* **286**, 10939–10949 (2011).
265. Arribere, J. A. & Gilbert, W. V. Roles for transcript leaders in translation and mRNA decay revealed by transcript leader sequencing. *Genome Res.* **23**, 977–987 (2013).
266. Patraquim, P., Magny, E. G., Pueyo, J. I., Platero, A. I. & Couso, J. P. Translation and natural selection of micropeptides from long non-canonical RNAs. *Nat. Commun.* **13**, 6515 (2022).
267. Wu, Q. *et al.* Translation of small downstream ORFs enhances translation of canonical

- main open reading frames. *EMBO J.* **39**, e104763 (2020).
268. Chong, C. *et al.* Integrated proteogenomic deep sequencing and analytics accurately identify non-canonical peptides in tumor immunopeptidomes. *Nat. Commun.* **11**, 1293 (2020).
269. Ruiz Cuevas, M. V. *et al.* Most non-canonical proteins uniquely populate the proteome or immunopeptidome. *Cell Rep.* **34**, 108815 (2021).
270. Jiang, M., Lou, H. & Hou, W. Microproteins: from behind the scenes to the spotlight. *Genome Instab. Dis.* **2**, 225–239 (2021).
271. Merino-Valverde, I., Greco, E. & Abad, M. The microproteome of cancer: From invisibility to relevance. *Exp. Cell Res.* **392**, 111997 (2020).
272. Wright, B. W., Yi, Z., Weissman, J. S. & Chen, J. The dark proteome: translation from noncanonical open reading frames. *Trends Cell Biol.* **32**, 243–258 (2022).
273. A Micropeptide Encoded by a Putative Long Non-coding RNA Regulates Muscle Performance - PMC. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4356254/>.
274. Nelson, B. R. *et al.* A peptide encoded by a transcript annotated as long noncoding RNA enhances SERCA activity in muscle. *Science* **351**, 271–275 (2016).
275. Zhang, Q. *et al.* The microprotein Minion controls cell fusion and muscle formation. *Nat. Commun.* **8**, 15664 (2017).
276. Bukhari, I. *et al.* PINTology: A short history of the lncRNA LINC-PINT in different diseases. *WIREs RNA* **13**, e1705 (2022).
277. Zhang, M. *et al.* A peptide encoded by circular form of LINC-PINT suppresses oncogenic transcriptional elongation in glioblastoma. *Nat. Commun.* **9**, 4475 (2018).
278. Slavoff, S. A., Heo, J., Budnik, B. A., Hanakahi, L. A. & Saghatelian, A. A Human Short Open Reading Frame (sORF)-encoded Polypeptide That Stimulates DNA End Joining. *J. Biol. Chem.* **289**, 10950–10957 (2014).
279. Arnoult, N. *et al.* Regulation of DNA repair pathway choice in S and G2 phases by the

- NHEJ inhibitor CYREN. *Nature* **549**, 548–552 (2017).
280. Huang, J.-Z. *et al.* A Peptide Encoded by a Putative lncRNA HOXB-AS3 Suppresses Colon Cancer Growth. *Mol. Cell* **68**, 171–184.e6 (2017).
281. Leng, F. *et al.* A micro-peptide encoded by HOXB-AS3 promotes the proliferation and viability of oral squamous cell carcinoma cell lines by directly binding with IGF2BP2 to stabilize c-Myc. *Oncol. Lett.* **22**, 697 (2021).
282. Meng, N. *et al.* Small Protein Hidden in lncRNA LOC90024 Promotes “Cancerous” RNA Splicing and Tumorigenesis. *Adv. Sci.* **7**, (2020).
283. Sun, L. *et al.* The oncomicropeptide APPLE promotes hematopoietic malignancy by enhancing translation initiation. *Mol. Cell* **81**, 4493–4508.e9 (2021).
284. Chu, Q. *et al.* Regulation of the ER stress response by a mitochondrial microprotein. *Nat. Commun.* **10**, 4883 (2019).
285. D’Lima, N. G. *et al.* A human microprotein that interacts with the mRNA decapping complex. *Nat. Chem. Biol.* **13**, 174–180 (2017).
286. Jackson, R. *et al.* The translation of non-canonical open reading frames controls mucosal immunity. *Nature* **564**, 434–438 (2018).
287. Zhang, S. *et al.* Mitochondrial peptide BRAWNIN is essential for vertebrate respiratory complex III assembly. *Nat. Commun.* **11**, 1312 (2020).
288. Liang, C. *et al.* Mitochondrial microproteins link metabolic cues to respiratory chain biogenesis. *Cell Rep.* **40**, 111204 (2022).
289. Heesch, S. van *et al.* The Translational Landscape of the Human Heart. *Cell* **178**, 242–260.e29 (2019).
290. Stein, C. S. *et al.* Mitoregulin: A lncRNA-Encoded Microprotein that Supports Mitochondrial Supercomplexes and Respiratory Efficiency. *Cell Rep.* **23**, 3710–3720.e8 (2018).
291. Ge, Q. *et al.* Micropeptide ASAP encoded by LINC00467 promotes colorectal cancer

- progression by directly modulating ATP synthase activity. *J. Clin. Invest.* **131**, e152911.
292. Hardy, S. *et al.* Magnesium-sensitive upstream ORF controls PRL phosphatase expression to mediate energy metabolism. *Proc. Natl. Acad. Sci.* **116**, 2925–2934 (2019).
 293. Hofman, D. A. *et al.* *Translation of Non-Canonical Open Reading Frames as a Cancer Cell Survival Mechanism in Childhood Medulloblastoma.*
<http://biorxiv.org/lookup/doi/10.1101/2023.05.04.539399> (2023)
doi:10.1101/2023.05.04.539399.
 294. Papaioannou, D. *et al.* The long non-coding RNA HOXB-AS3 regulates ribosomal RNA transcription in NPM1-mutated acute myeloid leukemia. *Nat. Commun.* **10**, 5351 (2019).
 295. Moll, P., Ante, M., Seitz, A. & Reda, T. QuantSeq 3' mRNA sequencing for RNA quantification. *Nat. Methods* **11**, i–iii (2014).
 296. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions | Genome Biology | Full Text.
<https://genomebiology.biomedcentral.com/articles/10.1186/gb-2013-14-4-r36>.
 297. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
 298. Moulos, P. & Hatzis, P. Systematic integration of RNA-Seq statistical algorithms for accurate detection of differential gene expression patterns. *Nucleic Acids Res.* **43**, e25 (2015).
 299. Differential expression analysis for sequence count data | Genome Biology | Full Text.
<https://genomebiology.biomedcentral.com/articles/10.1186/gb-2010-11-10-r106>.
 300. Michels, B. E. *et al.* Human colon organoids reveal distinct physiologic and oncogenic Wnt responses. *J. Exp. Med.* **216**, 704–720 (2019).
 301. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).
 302. Muzny, D. M. *et al.* Comprehensive molecular characterization of human colon and rectal

- cancer. *Nature* **487**, 330–337 (2012).
303. Hughes, C. S. *et al.* Single-pot, solid-phase-enhanced sample preparation for proteomics experiments. *Nat. Protoc.* **14**, 68–85 (2019).
 304. Demichev, V., Messner, C. B., Vernardis, S. I., Lilley, K. S. & Ralser, M. DIA-NN: neural networks and interference correction enable deep proteome coverage in high throughput. *Nat. Methods* **17**, 41–44 (2020).
 305. Tyanova, S. *et al.* The Perseus computational platform for comprehensive analysis of (prote)omics data. *Nat. Methods* **13**, 731–740 (2016).
 306. Lonsdale, J. *et al.* The Genotype-Tissue Expression (GTEx) project. *Nat. Genet.* **45**, 580–585 (2013).
 307. Fijalkowska, D. *et al.* eIF1 modulates the recognition of suboptimal translation initiation sites and steers gene expression via uORFs. *Nucleic Acids Res.* **45**, 7997–8013 (2017).
 308. Jones, P. *et al.* InterProScan 5: genome-scale protein function classification. *Bioinforma. Oxf. Engl.* **30**, 1236–1240 (2014).
 309. Feng, Y., Qu, L., Wang, X. & Liu, C. LINC01133 promotes the progression of cervical cancer by sponging miR-4784 to up-regulate AHDC1. *Cancer Biol. Ther.* **20**, 1453–1461 (2019).
 310. Tabas-Madrid, D., Nogales-Cadenas, R. & Pascual-Montano, A. GeneCodis3: a non-redundant and modular enrichment analysis tool for functional genomics. *Nucleic Acids Res.* **40**, W478–W483 (2012).
 311. Hansen, M. & Walmod, P. S. IGSF9 family proteins. *Neurochem. Res.* **38**, 1236–1251 (2013).
 312. Li, Y. *et al.* Immunoglobulin superfamily 9 (IGSF9) is trans-activated by p53, inhibits breast cancer metastasis via FAK. *Oncogene* **41**, 4658–4672 (2022).
 313. Luan, H. *et al.* IGSF9 promotes tumor invasion and metastasis through GSK-3 β / β -catenin mediated EMT in lung cancer. *Neoplasia N. Y. N* **58**, 101067 (2024).

314. Downregulation of LINC01133 in CRC patients.
315. Cells of the human intestinal tract mapped across space and time | Nature.
<https://www.nature.com/articles/s41586-021-03852-1>.
316. Gao, N., White, P. & Kaestner, K. H. Establishment of Intestinal Identity and Epithelial-Mesenchymal Signaling by Cdx2. *Dev. Cell* **16**, 588–599 (2009).
317. Abramson, J. *et al.* Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature* **630**, 493–500 (2024).
318. Wang, X. & Marchisio, M. A. Synthetic polycistronic sequences in eukaryotes. *Synth. Syst. Biotechnol.* **6**, 254–261 (2021).
319. Zhitomirsky, B., Farber, H. & Assaraf, Y. G. LysoTracker and MitoTracker Red are transport substrates of P-glycoprotein: implications for anticancer drug design evading multidrug resistance. *J. Cell. Mol. Med.* **22**, 2131–2141 (2018).
320. Rath, S. *et al.* MitoCarta3.0: an updated mitochondrial proteome now with sub-organelle localization and pathway annotations. *Nucleic Acids Res.* **49**, D1541–D1547 (2021).
321. Fan, Y. *et al.* HAX1-dependent control of mitochondrial proteostasis governs neutrophil granulocyte differentiation. *J. Clin. Invest.* **132**, e153153.
322. Dendrite branching and self-avoidance are controlled by Turtle, a conserved IgSF protein in Drosophila - PubMed. <https://pubmed.ncbi.nlm.nih.gov/19783736/>.
323. Mishra, A. *et al.* Genetic Evidence for the Adhesion Protein IgSF9/Dasm1 to Regulate Inhibitory Synapse Development Independent of its Intracellular Domain. *J. Neurosci.* **34**, 4187–4199 (2014).
324. Liu, Y. *et al.* Targeting the Immunoglobulin IGSF9 Enhances Antitumor T-cell Activity and Sensitivity to Anti-PD-1 Immunotherapy. *Cancer Res.* **83**, 3385–3399 (2023).
325. Silberg, D. G., Swain, G. P., Suh, E. R. & Traber, P. G. Cdx1 and Cdx2 expression during intestinal development. *Gastroenterology* **119**, 961–971 (2000).
326. Silberg, D. G. *et al.* Cdx2 ectopic expression induces gastric intestinal metaplasia in

- transgenic mice. *Gastroenterology* **122**, 689–696 (2002).
327. Grainger, S., Savory, J. G. A. & Lohnes, D. Cdx2 regulates patterning of the intestinal epithelium. *Dev. Biol.* **339**, 155–165 (2010).
 328. Verzi, M. P., Shin, H., Ho, L.-L., Liu, X. S. & Shivdasani, R. A. Essential and Redundant Functions of Caudal Family Proteins in Activating Adult Intestinal Genes ∇ . *Mol. Cell. Biol.* **31**, 2026–2039 (2011).
 329. Yu, J. *et al.* CDX2 inhibits the proliferation and tumor formation of colon cancer cells by suppressing Wnt/ β -catenin signaling via transactivation of GSK-3 β and Axin2 expression. *Cell Death Dis.* **10**, 1–14 (2019).
 330. Bhat, A. A. *et al.* Caudal Homeobox Protein Cdx-2 Cooperates with Wnt Pathway to Regulate Claudin-1 Expression in Colon Cancer Cells. *PLOS ONE* **7**, e37174 (2012).
 331. Shirokikh, N. E., Archer, S. K., Beilharz, T. H., Powell, D. & Preiss, T. Translation complex profile sequencing to study the in vivo dynamics of mRNA–ribosome interactions during translation initiation, elongation and termination. *Nat. Protoc.* **12**, 697–731 (2017).
 332. Calviello, L. *et al.* Detecting actively translated open reading frames in ribosome profiling data. *Nat. Methods* **13**, 165–170 (2016).
 333. Guo, Z.-W. *et al.* Translated Long Non-Coding Ribonucleic Acid ZFAS1 Promotes Cancer Cell Migration by Elevating Reactive Oxygen Species Production in Hepatocellular Carcinoma. *Front. Genet.* **10**, (2019).
 334. Zhou, D. *et al.* Protocol for mitochondrial isolation and sub-cellular localization assay for mitochondrial proteins. *STAR Protoc.* **4**, 102088 (2023).
 335. Arroum, T. *et al.* Loss of respiratory complex I subunit NDUFB10 affects complex I assembly and supercomplex formation. *Biol. Chem.* **404**, 399–415 (2023).
 336. Pacheu-Grau, D. *et al.* Cooperation between COA6 and SCO2 in COX2 maturation during cytochrome c oxidase assembly links two mitochondrial cardiomyopathies. *Cell Metab.* **21**, 823–833 (2015).

337. Darshi, M. *et al.* ChChd3, an inner mitochondrial membrane protein, is essential for maintaining crista integrity and mitochondrial function. *J. Biol. Chem.* **286**, 2918–2932 (2011).
338. Chiang, T.-W. W., le Sage, C., Larrieu, D., Demir, M. & Jackson, S. P. CRISPR-Cas9D10A nickase-based genotypic and phenotypic screening to enhance genome editing. *Sci. Rep.* **6**, 24356 (2016).
339. Anzalone, A. V. *et al.* Search-and-replace genome editing without double-strand breaks or donor DNA. *Nature* **576**, 149–157 (2019).
340. Free energy landscape of RNA binding dynamics in start codon recognition by eukaryotic ribosomal pre-initiation complex | PLOS Computational Biology.
<https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1009068>.