

Computational approaches to study the RNA response to hypoxia in human cancer cells

Dissertation
zur Erlangung des Doktorgrades
der Naturwissenschaften

vorgelegt beim
Fachbereich Biowissenschaften (FB15) der
Goethe-Universität Frankfurt am Main

von
Antonella Di Liddo
aus Bari (Italien)

Frankfurt (2019)
(D 30)

Vom Fachbereich Biowissenschaften (FB15) der Goethe-Universität als Dissertation
angenommen.

Dekan: Prof. Dr. Sven Klimpel

Gutachter: Dr. Kathi Zarnack

Gutachter: Prof. Dr. Ingo Ebersberger

Datum der Disputation: 27/05/2020

Contents

Contents	ii
List of Figures	v
List of Tables	viii
Abbreviations	1
Zusammenfassung	5
Abstract	12
Preface	14
1 Introduction	15
1.1 The human transcriptome	15
1.2 The complex life of mRNA	16
1.3 The pre-mRNA splicing	17
1.4 Alternative splicing	21
1.5 Alternative splicing regulators: MBNL proteins	23
1.6 Back-splicing generates circular RNAs	25
1.7 RNA-Sequencing: a powerful tool for transcriptomics studies	31
1.8 Computational identification of circRNAs from RNA-Seq	33
1.9 Hypoxia: a hallmark of cancer	37
1.10 Aim of this thesis	42
2 Methods	43
2.1 Transcriptome analysis from RNA-Sequencing data	44
2.1.1 Processing and mapping of sequencing reads	44
2.1.2 Gene-level quantification and differential expression	45

2.1.3	Analysis of alternative splicing changes	46
2.1.4	Prediction of MBNL2 binding sites	47
2.2	Establishment and evaluation of the pipeline for circRNA detection .	48
2.3	Identification and annotation of circRNAs	52
2.3.1	Quantification of circRNA and host gene expression	53
2.3.2	Prediction of RBP binding sites	53
2.3.3	Detection of putative miRNA binding sites	54
2.4	Programs	55
2.5	Databases	60
3	Results	62
3.1	Transcriptional and post-transcriptional changes in response to hypoxia	62
3.1.1	Hypoxia strongly affects linear RNA abundance	64
3.1.2	Hypoxia alters the splicing pattern in a cell type-specific manner	68
3.2	The role of MBNL2 in hypoxia	70
3.2.1	MBNL2 modulates the transcript abundance of hypoxia re- response genes	72
3.2.2	MBNL2 controls hypoxia-dependent alternative splicing	79
3.3	Establishing a pipeline to identify circRNAs from rRNA-depleted RNA-Seq data	82
3.3.1	Testing <code>CIRCexplorer</code> and <code>find_circ</code> on rRNA-depleted RNA-Seq data	83
3.3.2	A novel combined pipeline for circRNA detection	87
3.3.3	Evaluating the performance of the pipeline with RNase R- treated RNA-Seq data	89
3.4	CircRNome profiling in cancer cells	94
3.4.1	Genomic context of circRNAs	97
3.4.2	The circRNA profile differs between cancer cells	102
3.4.3	CircRNA levels change upon hypoxic stress	106
3.4.4	Insights into the mechanisms of circRNA biogenesis	112
3.4.5	CircRNA as miRNA sponges	118
4	Discussion	121
5	Conclusions	130

Supplementary Material	131
References	141
Acknowledgements	166

List of Figures

1.1	Overview of the mRNA life cycle	16
1.2	Schematic of the pre-mRNA splicing	19
1.3	Spliceosome assembly step-by-step	20
1.4	Constitutive and alternative splicing	22
1.5	The muscleblind-like (MBNL) protein family	23
1.6	Back-splicing generates circRNAs	25
1.7	Publication rates of circRNAs.	26
1.8	Hypothetical mechanisms for circRNA biogenesis	28
1.9	Functions of circRNAs	30
1.10	Different approaches for library preparation for RNA-Seq	31
1.11	Schematic of reads spanning back-splice junctions	33
1.12	Hypoxic niches in solid tumours.	38
1.13	Hypoxia-inducible factor (HIF) regulation in normoxia and hypoxia.	39
3.1	Overview of RNA-Seq data from human cancer cells	64
3.2	Influence of hypoxia on gene expression at RNA level	65
3.3	Functional characterisation of differentially expressed genes upon hypoxia	67
3.4	Alternative splicing profile upon hypoxia	69
3.5	Changes in mRNA levels of genes annotated to the GO term "RNA splicing"	71
3.6	Cisplatin-induced cell death increases upon <i>MBNL2</i> depletion in hypoxic cancer cells	72
3.7	<i>MBNL2</i> depletion influences transcript abundance in hypoxic cancer cells	75
3.8	<i>In silico</i> prediction of MBNL2 binding sites on 3'UTRs sequences of MBNL2-regulated genes	78

3.9	MBNL2 mainly functions as negative regulator of alternative cassette exons in hypoxia	80
3.10	Investigation of convergences and discrepancies between CIRCexplorer and find_circ tools in predicting circRNAs from a single HeLa sample	85
3.11	Pipeline to identify circRNAs and analyse gene expression and splicing from rRNA-depleted RNA-Seq data	88
3.12	Evaluation of circRNA prediction with our pipeline compared to CIRCexplorer and find_circ, based on published RNase R-treated RNA-Seq datasets.	93
3.13	Identification of circRNAs in human cancer cell lines	95
3.14	Validation of abundant circRNAs by RT-PCR	96
3.15	CircRNAs are mainly derived from internal exons of protein-coding genes	98
3.16	Genomic features of back-splice sites	100
3.17	Alternative back-splicing produces multiple circRNA isoforms from a single host gene	101
3.18	CircRNA profiles in human cancer cells.	102
3.19	Relationship between circRNA and host gene expression	105
3.20	Hypoxia induces changes in circRNA levels.	107
3.21	Validation of hypoxia-regulated circRNAs by RT-qPCR.	108
3.22	Changes of circRNA levels upon hypoxia often reflect variations of their host gene level	110
3.23	CircRNA biogenesis via read-through transcription of the upstream gene	111
3.24	CircRNA biogenesis via flanking inverted repeats	112
3.25	RNA-binding proteins as regulators of circRNA formation	116
3.26	<i>HNRNPC</i> depletion affects circRNA levels	117
3.27	<i>In silico</i> prediction of miRNA binding sites on circRNA sequences.	118
3.28	CircRNAs as potential miRNA targets	120
S1	3'UTR sequences of putative MBNL2 stability targets: <i>CSNRP1</i> and <i>OSMR</i>	137
S2	3'UTR sequences of putative MBNL2 stability targets: <i>SERTAD2</i>	138
S3	3'UTR sequences of putative MBNL2 stability targets: <i>SMAD7</i>	139

S4	Investigation of convergences and discrepancies between <code>CIRCexplorer</code> and <code>find_circ</code> tools in predicting circRNAs from a single MCF-7 sample	140
----	--	-----

List of Tables

1	List of abbreviations	1
1.1	Overview of available circRNA detection tools	36
2.1	List of software and algorithms used in this study	55
3.1	Summary of RNA-Seq data from hypoxia experiments in human cancer cells.	63
3.2	Summary of RNA-Seq data from <i>MBNL2</i> knockdown experiments in hypoxic MCF-7 cells.	73
3.3	List of putative <i>MBNL2</i> stability targets downregulated upon <i>MBNL2</i> knockdown.	77
3.4	Alternative splicing events after <i>MBNL2</i> knockdown.	79
3.5	Overview of the main features of algorithms adopted in this thesis to detect circRNAs	87
3.6	Overview of RNA-Seq datasets used for the performance evaluation of the proposed pipeline	91
3.7	Number of circRNAs identified in cancer cells.	94
S1	List of hypoxia-regulated circRNAs grouped by cell line	135

Dedicated to my husband and my lovely family

Abbreviations

Table 1: List of abbreviations

%	Percentage
°C	Degree Celsius
3'SS	3' splice site
5'SS	5' splice site
A	Adenosine
A3SS	Alternative 3' splice site
A5SS	Alternative 5' splice site
AS	Alternative splicing
BAM	Binary Alignment Map
BMLS	Buchmann Institute for Molecular Life Sciences
bp	Base pairs
BPS	Branch point site
BSJ	Back-splice junction
BSS	Back-splice site
C	Cytosine
CDS	Coding sequence
CE	Cassette exon
circ	Circular
circRNA	Circular RNA
ciRNA	Circular intronic RNA
CLIP	Cross linking and immunoprecipitation
CLR	circular-to-linear ratio
CO ₂	Carbon dioxide
DEG	Differentially expressed gene
DMEM	Dulbecco's modified eagle's medium
DMSO	Dimethylsulfoxide
DNA	Desoxyribonucleic acid
ecircRNAs	Exonic circRNAs

Table 1: List of abbreviations (*continued*)

EDTA	Ethylenediaminetetraacetic acid
EIciRNA	Exonic-intronic circRNAs
EMT	Epithelial-to-Mesenchymal Transition
ER	Endoplasmic reticulum
ESE	Exonic splicing enhancer
ESS	Exonic splicing silencer
et al.	et alteri
FDR	False discovery rate
G	Guanine
GO	Gene Ontology
h	Hours
hg19	Human reference genome 19
hg38	Human reference genome 38
HIF	Hypoxia-inducible factor
HNRNP	Heterogeneous nuclear ribonucleoproteins
HRE	Hypoxia-responsive element
iCLIP	Individual nucleotide resolution cross linking and immunoprecipitation
IMB	Institute of Molecular Biology
IRES	Internal ribosome entry sites
ISE	Intronic splicing enhancer
ISS	Intronic splicing silencer
kb	kilobase
kcal	kilocalories
lncRNA	Long non-coding RNA
MBNL	Muscleblind-like
min	Minutes
miRNA	microRNAs
ml	milliliter
mRNA	Messenger RNA
MXE	Mutually exclusive exon
ncRNA	non-coding RNA
ng	nanogram
nt	Nucleotide
O ₂	Oxygen
PCG	Protein-coding gene
PCR	Polymerase chain reaction

Table 1: List of abbreviations (*continued*)

piRNA	Piwi-interacting RNA
PPT	Polypyrimidine tract
pre-mRNA	Precursor messenger RNA
PSI	Percent spliced in
RBP	RNA-binding protein
RI	Retained intron
RNA	Ribonucleic acid
RNA Pol II	RNA polymerase II
RNA-Seq	RNA sequencing
RNase R	Ribonuclease R
RPM	Reads per million
RRM	RNA recognition motif
RRMH	RNA recognition motif homolog
rRNA	Ribosomal RNA
RT	Reverse transcription
RT-qPCR	Reverse transcription-quantitative polymerase chain reaction
SAM	Sequence Alignment Map
SF1	Splicing factor 1
siCTRL	Control siRNA
siRNA	small interfering RNA
snRNA	small nuclear RNAs
snRNP	small nuclear ribonucleoprotein
SRSF	Serine-, arginine-rich splicing factor
TER	Transcription elongation rate
TPM	Transcripts per million
U2AF	U2 auxiliary factor
UPR	Unfolded protein response
UTR	Untranslated region
VEGFA	Vascular endothelial growth factor A
Y	Pyrimidine
ZnF	Zinc finger
µg	microgram
µM	micromolar
log	logarithm
log2	binary logarithm
min	minutes

Table 1: List of abbreviations (*continued*)

nM	nanomolar
nm	nanometer
r	Pearson correlation coefficient
PBS	phosphate buffered saline
μg	microgram
μl	microliter

Zusammenfassung

Hypoxie ist ein Merkmal solider Tumore und trägt zum Fortschreiten von Krebs, zur Metastasierung sowie zu einer schlechten Prognose bei. Die Anpassung an Hypoxie wird hauptsächlich durch die Aktivierung von Hypoxie-induzierbaren Faktor (HIF)-Proteinen angetrieben. HIF-Proteine sind eine Familie von Transkriptionsfaktoren, welche die Expression von mehr als hundert Genen bei reduzierter Sauerstoffversorgung regulieren. Diese sogenannten HIF-regulierten Gene spielen in zahlreichen zellulären Prozessen eine Rolle, wie zum Beispiel in Angiogenese, Proliferation und metabolischer Anpassung, welches alles wichtige Faktoren für ein Tumorwachstum sind. Obwohl Änderungen in der Transkription, die in hypoxischen Tumoren induziert werden, weitestgehend charakterisiert sind, ist bisher nicht vollständig geklärt, wie die Hypoxie zu der veränderten posttranskriptionellen Regulation in Tumoren führt. In dieser Studie habe ich das Transkriptom von drei menschlichen Zelllinien aus Lungen- (A549), Brust- (MCF-7) und Gebärmutterhalskrebs (HeLa) in normoxischen, sowie hypoxischen Bedingungen analysiert. Die Ergebnisse meiner Analysen haben zu einem verbessertem Verständnis der hypoxiegetriebenen, posttranskriptionellen Genregulation in Krebs beigetragen.

Unter Verwendung tiefer RNA-Sequenzierung von Proben mit reduziertem ribosomalem RNA-Gehalt (rRNA-*depleted* RNA-Seq), konnte ich insgesamt mehr als 10000 Gene in den drei Zelllinien identifizieren, die ihre RNA-Menge unter Hypoxie verändert haben. Hypoxie induzierte ähnliche Veränderungen der Transkriptionshäufigkeit in den drei Krebsstypen, sowie modulierte sie die Expression bekannter HIF-Zielgene, welche an krebisbedingten Prozessen wie Angiogenese und Zellmigration beteiligt sind. Darüber hinaus wurde die Glykolyse zur Energieerzeugung aktiviert, und energieverbrauchende Prozesse, wie die DNA-Replikation und Ribosomen-Biogenese moduliert, was zu einer metabolischen Anpassung geführt hat.

Neben der Anpassung der Transkription umfasst die Regulation der Genexpression eine Vielzahl von Mechanismen, die von Zellen verwendet werden, um die Produktion und Funktion spezifischer Genprodukte (Proteine oder RNAs) zu regulieren. Proteinkodierende Gene werden in Prä-messenger-RNAs (Prä-mRNA) transkribiert. mRNAs bestehen aus kodierenden Regionen, die als Exons bezeichnet werden, und dazwischenliegenden,

nicht-kodierenden Regionen, welche als Introns bezeichnet werden. Genexpression kann in nahezu allen Stadien der RNA-Prozessierung reguliert werden. Einer der wichtigsten Schritte für die Reifung der Prä-mRNA zu Messenger-RNA (mRNA) ist das RNA-Spleißen, welches durch einen makromolekularen Komplex, das Spleißosom, katalysiert wird. Während des Spleißens werden Exons miteinander verbunden und Introns entfernt, um eine reife mRNA zu erzeugen. Im Falle von Krebserkrankungen ist das Spleißen häufig beeinträchtigt und beeinflusst dadurch die Zellproliferation, das Zellüberleben, die Migration und Metastasierung. In dieser Studie zeige ich, dass Hypoxie das Spleißmuster in Krebszellen stark veränderte. Dabei waren vorrangig alternative Kassetten-Exons und die Beibehaltung von Introns (*Intron Retention*) betroffen. Im Gegensatz zu Veränderungen in der RNA-Menge verursachte Hypoxie unterschiedliche Spleißreaktionen in den drei Zelllinien, was die Zelltypspezifität alternativer Spleißprogramme unterstreicht.

Während des Spleißens ist die Aktivierung von Spleißstellen von grundlegender Bedeutung, um festzulegen, welche Exons in die mRNA aufgenommen werden sollen und um somit verschiedene Transkriptisoformen zu erzeugen. Dies wird im Allgemeinen durch eine Kombination von Spleißfaktoren vermittelt, die jeden Schritt der Spleißosomfunktion auf dynamische Weise modulieren. Folglich ist anzunehmen, dass Änderungen im alternativen Spleißen auf Variationen in der Expression und Aktivität von Spleißfaktoren zurückzuführen sind. In dieser Studie zeige ich, dass das mRNA-Niveau von Spleißfaktoren, wie zum Beispiel SR-Proteinen, in Hypoxie überwiegend reduziert waren. Eine globale Reduktion der Spleißaktivität spiegelt sich in der signifikanten Anreicherung des Ontologie-Terminus "*RNA splicing*" (GO:0008380) wider. Im Gegensatz dazu induzierte Hypoxie gezielt die Expression des *muscleblind-like protein 2* (*MBNL2*) in allen drei Zelllinien. Meine Beobachtung aus den RNA-Seq-Experimenten konnten durch Western Blot-Analysen unserer Kooperationspartner bestätigt werden, die einen deutlichen Anstieg der MBNL2-Menge nach Hypoxie belegten.

MBNL2 hat zwei paraloge Proteine im Menschen: MBNL1 und MBNL3. Es ist bekannt, dass MBNL1 und MBNL3 ähnliche Sequenzmotive wie MBNL2 erkennen und binden, sowie dass sie häufig die gleichen mRNA-Ziele wie MBNL2 anvisieren. Es ist hervorzuheben, dass in unseren Daten der Effekt von Hypoxie spezifisch für *MBNL2* war, da das mRNA- und Proteinniveau von *MBNL1* unter Sauerstoffmangel stabil blieben. Es wurde zuvor gezeigt, dass *MBNL2* je nach Krebsart entweder als Onkogen oder als Tumorsuppressorgen fungieren kann. Auf Basis dessen habe ich in dieser Studie die Rolle von *MBNL2* als Antwort auf hypoxischen Stress in Krebszellen untersucht. Unsere *MBNL2-Knockdown*-Experimente in hypoxischen Zellen bestätigten die Beteiligung von *MBNL2* in der Hypoxieanpassung von Krebszellen. Des Weiteren hat die Analyse der RNA-Seq-Daten von

MBNL2-abgereicherten Krebszellen gezeigt, dass die Hypoxieanpassung durch die Steuerung der Transkriptmenge und des alternativen Spleißens von Hypoxie-Antwort-Genen erreicht wurde. Im Gegensatz zu früheren Studien, die eine Rolle von *MBNL2* bei der Stabilisierung von mRNAs vorhergesagt hatten, deuteten unsere Daten nicht darauf hin, dass diese Funktion auf die Mehrheit der *MBNL2*-regulierten mRNAs zurückgeführt werden kann. Zusätzlich zeigten Experimente unserer Kollaborationspartnern, dass die Abreicherung von *MBNL2* die Proliferation und Migration von Krebszellen verringerte, was die Rolle von *MBNL2* als wichtigen Krebstreiber unterstreichte. Zusammenfassend konnten wir zeigen, dass Hypoxie die Genexpression auf transkriptioneller- und posttranskriptioneller Ebene beeinflusst und somit Tumorentstehung vorantreibt. Die spezifische Induktion der Expression von *MBNL2* bei niedrigem Sauerstoffgehalt fördert die hypoxische Anpassung von Krebszellen. Dies wird erreicht, indem *MBNL2* die Transkripthäufigkeit von HIF-Zielgenen kontrolliert und zum Hypoxie-abhängigen alternativen Spleißen beiträgt.

Eine neue Klasse von hauptsächlich nicht-kodierenden RNAs, die zirkulären RNAs (circRNAs), hat in den letzten Jahren mehr und mehr Beachtung erlangt. CircRNAs werden durch einen bestimmten Spleißmechanismus hergestellt, der als Zurückspleißen (*back-splicing*) bezeichnet wird. Während des Zurückspleißens wird eine 5'-Spleißstelle mit einer 3'-Spleißstelle verbunden, welche sich an einer vorhergehenden Position im Transkript befindet. Dieser Prozess erzeugt ein sehr stabiles, kovalent geschlossenes Molekül. Es wurde gezeigt, dass die Menge vieler circRNAs in Krebszellen fluktuieren kann. Dank ihrer hohen Stabilität sind sie vielversprechende Kandidaten für diagnostische Biomarker in Krebs. Darüber hinaus haben jüngste Studien circRNAs beschrieben, die in hypoxischen Endothelzellen und Magenkrebszellen dereguliert sind, wie beispielsweise die circRNA, welche aus einer kryptischen Spleißseite des *ZNF292*-Gens entsteht. Trotz der bestehenden Studien über circRNAs ist das Verständnis über die Auswirkungen von Hypoxie auf circRNAs in Krebszellen bisher limitiert.

In dieser Studie habe ich rRNA-*depleted* RNA-Seq-Daten analysiert, um die Expression von circRNAs in menschlichen Krebszellen, sowie ihre Veränderungen als Reaktion auf Hypoxie umfassend zu untersuchen. Die Identifizierung von circRNAs aus rRNA-*depleted* RNA-Seq-Daten ist anspruchsvoll, da lineare, sowie zirkuläre RNAs in den Daten enthalten sind. Da circRNAs oft mit ihrem linearen RNA-Gegenstücken überlappen, enthalten die Daten wenig diskriminative Sequenzinformationen, die eine Basis für eine zuverlässige Detektion der circRNAs schaffen. Mithilfe computergestützter Berechnungen kann man circRNAs detektieren, indem man exklusiv nach Sequenzstücken sucht, die sich über Zurückspleiß-*Junctions* erstrecken. Zusätzlich ist die quantitative Analyse von circRNAs problematisch, da sie global nur in geringen Mengen vorhanden sind. Somit ist die *de novo*-Vorhersage

von circRNAs an kryptischen Spleißstellen weiterhin keine triviale Aufgabe. Seit 2013 wurden verschiedene Algorithmen zur Vorhersage und Quantifizierung von circRNAs aus rRNA-*depleted* RNA-Seq-Daten entwickelt, die verschiedene Alleinstellungsmerkmale mit sich bringen. Die unterschiedlichen Ansätze sind jedoch oft mit einer hohen Falsch-Positiv-Rate verbunden. Folglich wurde empfohlen, die Ergebnisse mehrerer Algorithmen zu kombinieren, um einen zuverlässigen Katalog von circRNAs aus RNA-Seq-Daten zu erhalten. Um dieses Problem zu adressieren habe ich in dieser Studie eine Pipeline erstellt, die zwei verfügbare Programme für die circRNA-Identifizierung, **CIRCexplorer** und **find_circ**, kombiniert. Zudem integriert die Pipeline maßgeschneiderte Ansätze für Quantifizierungen und statistische Analysen. Die beiden kombinierten Werkzeuge ergänzen sich, da sie auf unterschiedlichen Sequenzalignier-Algorithmen (**Bowtie2** und **STAR**) und verschiedenen konzeptionellen Ansätzen beruhen. Der Ansatz von **find_circ** neigt aufgrund ungenauer Zuweisungen von Zurückspleißen zu Falsch-Positiven, wohingegen **CIRCexplorer** auf Exon-Koordinaten aus der Referenzgenom-Annotation angewiesen ist. In unserer Pipeline wurden Vorhersagen mit **CIRCexplorer** und **find_circ** vereint und Artefakten, die ich zuvor durch einen umfassenden Vergleich der beiden Programme identifizierte, wurden herausgefiltert. Anschließend wurde die Quantifizierung der circRNA-Expression basierend auf chimären Sequenzalignments aus **STAR** abgeglichen. Unter Verwendung öffentlich verfügbarer rRNA-*depleted* und RNase-behandelter RNA-Seq-Daten bewertete ich die Leistung der Pipeline hingehend der Erkennung echter Zurückspleiß-Ereignisse und verglich ich mit Ergebnissen von **CIRCexplorer** und **find_circ**. Dies zeigte, dass unsere Pipeline eine bessere Leistung im Vergleich zu **find_circ** und eine zumindest vergleichbare Leistung im Vergleich zu **CIRCexplorer** erzielte, mit dem Vorteil, dass sie die bereits wertvolle Vorhersage von **CIRCexplorer** um die von **find_circ** durchgeführte *de novo*-Vorhersage von circRNAs erweitert. Unsere Pipeline liefert einen umfassenden Katalog von präzise quantifizierten circRNAs, welcher als Ausgangspunkt für darauffolgende Analysen verwendet werden kann.

Durch die konsolidierte bioinformatische Pipeline konnte ich 12006 circRNAs in den drei analysierten Krebszelllinien identifizieren. Unter diesen befanden sich 2844 neu-identifizierte circRNAs, die zuvor noch in keinen circRNA-Datenbanken annotiert wurden. Beispiele dafür sind unter anderem circHUWE1, circSPIDR und circPICALM, welche in den untersuchten Zelllinien in großen Mengen vorhanden waren. Die Zirkularität detektierter circRNAs wurde von unseren Kollaborationspartnern experimentell über RT-PCR validiert, was die Zuverlässigkeit der Pipeline untermauerte. Des Weiteren analysierte ich die genomischen Merkmale unseres circRNA-Katalogs und konnte zeigen, dass die Mehrheit der circRNAs aus kodierenden Sequenzen (engl. *coding sequence*, CDS) von proteinkodierenden Genen stammt. Durch alternatives 3'- oder 5'-Zurückspleißen könnten theoretisch

mehrere circRNAs von einem einzigen Locus erzeugt werden. In den meisten Fällen gab es eine spezifische Isoform, die stärker als die anderen Isoformen exprimiert wurde. Unter den insgesamt 12006 identifizierten circRNAs gab es nur eine geringe Anzahl von circRNAs, die in allen drei Zelllinien vorkamen. Dies deutete darauf hin, dass jede analysierte Krebszelle eine einzigartige circRNA-Signatur aufweist.

Um Einblicke in den Mechanismus der Regulation von circRNAs zu erhalten, untersuchte ich den Zusammenhang zwischen der Expression von circRNAs und ihren jeweiligen Wirtsgenen. Generell konnten wir nur eine schwache quantitative Korrelation zwischen circRNA- und Wirtsgen-Expression beobachten. Das deutete darauf hin, dass in vielen Fällen die circRNA-Menge nicht nur die Expression des Wirtsgens widerspiegelte, sondern auch von unabhängigen Parametern, wie zum Beispiel dem unterschiedlichen Grad an Zurückspleißen oder der Stabilität der circRNA, beeinflusst wurde. Weiterhin untersuchte ich die Effizienz des Zurückspleißens, indem ich die "Prozentuale Zirkularisierung" berücksichtigte. Durch diese Berechnung war es möglich, die relative Häufigkeit von circRNAs im Vergleich zu allen anderen Isoformen, welche aus den gleichen Exons entstanden sind, abzuschätzen. Obwohl circRNAs generell seltener als ihre linearen Gegenstücke vorkamen, gab es in unserem Katalog 210 Ausnahmen, welche die Haupt-Transkriptisoform ihres Wirtsgens darstellten. Unter ihnen befand sich zum Beispiel die circRNA, die aus den Exons 2 und 3 des *ATXN7*-Gens hergestellt wird. Unterschiedliche regulatorische Prozesse können die Expression von circRNAs steuern, und das Zurückspleißen kann in der Effizienz zwischen verschiedenen Wirtsgenen stark variieren. Als nächstes ging ich auf die Frage ein, ob Hypoxie die circRNA-Menge in den Krebszelllinien modulieren kann. Unter Verwendung von DESeq2 fand ich insgesamt 64 circRNAs, die ihre Menge unter Hypoxie in den untersuchten Krebszelllinien signifikant änderten. Unter ihnen befanden sich nur sechs herunterregulierte circRNAs, was womöglich auf die intrinsische Stabilität von circRNAs zurückzuführen ist. Im Gegensatz zur einheitlichen transkriptionellen Antwort auf Hypoxie in den drei Zelllinien und in Übereinstimmung mit den divergierenden Spleißänderungen, wurden circRNAs in zelltypspezifischer Weise reguliert. Die einzigen zwei Ausnahmen waren circPLOC2 und circZNF292, welche sowohl in HeLa- als auch in MCF-7-Zellen signifikant reguliert wurden. In Übereinstimmung damit wurde in einer früheren Studie gezeigt, dass circZNF292 unter Hypoxie in Endothelzellen induziert wird. Die Regulation ausgewählter circRNAs wurde zudem von unseren Kollaborationspartnern durch RT-qPCR validiert. Um die Frage zu beantworten, ob Veränderungen der circRNA in Reaktion auf Hypoxie eine Regulation des jeweiligen Wirtsgens widerspiegeln, verglich ich die Häufigkeit der Hypoxie-regulierten circRNAs mit der Häufigkeit der linear RNAs. Ich konnte keine globale Korrelation zwischen den Expressionsmengen feststellen. Viele der Hypoxie-induzierten circRNAs stammten jedoch von hochregulierten Genen, was darauf hindeutete, dass ihre Regulation mit

einer erhöhten Transkription des Wirtsgens zusammenhängen könnte. Im Gegensatz dazu stammten andere hochregulierte circRNAs von Genen, die stabil exprimiert blieben, wie zum Beispiel circBARD1 und circRANBP17, oder sogar von herunterregulierten Genen, wie zum Beispiel circHNRNPM. Ein Vergleich von Änderungen in der Menge der circRNAs hinsichtlich ihrer *Circular-to-Linear-Ratio*, welche die *Zurückspleiß-Junctions* zu den *Junction*-Sequenzen der zugehörigen linear Spleißereignisse in Vergleich setzt, unterstützte weiter die Hypothese eines gemeinsamen Regulationsmechanismus für circRNA und mRNA auf der Ebene der Transkription. Im Gegensatz zu früheren Studien fand ich keine Hinweise auf eine *readthrough* Transkription des vorhergehenden Gens als Mechanismus der Biogenese von circRNAs. Um die molekularen Mechanismen, die der Entstehung und Regulation von circRNA zugrunde liegen, zu untersuchen, analysierte ich die molekularen Eigenschaften der circRNAs. Ähnlich wie in früheren Studien gezeigt, deuteten unsere Daten auf eine Beteiligung komplementärer RNA-Sequenzen an der circRNA-Biogenese hin, insbesondere von *Alu*-Elementen. Diese *Alu*-Elemente befinden sich in Introns, welche die zirkularisierten Exons flankieren. Über diese *cis*-wirkenden Faktoren hinaus wurde beschrieben, dass auch *trans*-wirkende Faktoren, wie MBNL-, QKI-, FUS- und SR-Proteine, eine Rolle bei der circRNA-Biogenese spielen. Durch *in silico* Vorhersage der Bindestellen von RNA-bindenden Proteinen in den Regionen vor und nach zirkularisierten Exons konnte ich HuR, PABPC4 und HNRNPC als potenzielle Akteure in der circRNA-Biogenese identifizieren. Vor allem die Metaanalyse der Daten von HNRNPC UV-Kreuzvernetzungs- und Immunpräzipitationsexperimenten (engl. *individual-nucleotide resolution UV crosslinking and immunoprecipitation*, iCLIP) zeigte, dass HNRNPC verstärkt an die Region unmittelbar vorhergehend der 3'-Zurückspleißstelle band. Dies war unabhängig davon, ob die circRNAs Hypoxie-reguliert waren oder nicht. Zu betonen ist, dass die Stärke der Bindung vorhergehend der 3'-Zurückspleißstelle im Vergleich zu linear gespleißten Exons wesentlich höher war, was unsere Hypothese der Rolle von HNRNPC in der circRNA-Biogenese bekräftigte. Dies wurde weiterhin durch einen nicht-zielgerichteten Ansatz bestätigt, bei dem drei circRNAs identifiziert wurden, die auf die *HNRNPC*-Abreicherung in HeLa-Zellen reagiert haben.

Alles in allem habe ich im Verlauf dieser Arbeit eine vergleichende Transkriptomcharakterisierung von drei menschlichen Krebszelllinien unter hypoxischen Stressbedingungen durchgeführt. Meine Ergebnisse haben gezeigt, dass MBNL2 ein wichtiger Akteur bei der Progression von hypoxischem Krebs ist, der sowohl die Transkriptionshäufigkeit als auch das Spleißen beeinflusst. Zusätzlich zu linearen RNAs habe ich das circRNA-Profil in Krebszellen, unter normoxischem sowie unter hypoxischem Stress charakterisiert. Dies hat neue Einblicke in die Regulation und Biogenese von circRNAs geliefert und Hypoxie-regulierte circRNAs identifiziert, welche als Biomarker für eine hypoxische Tumormikroumgebung

dienen könnten.

Abstract

Hypoxia is a condition in which cells are deprived of adequate oxygen supply and represents a main feature of solid tumours. Cells under hypoxic stress activate transcriptional responses driven by hypoxia-inducible factors (HIFs), which affect multiple cellular pathways, including angiogenesis, metabolic adaptation and cell proliferation. While the transcriptional changes induced in hypoxic tumours are well characterised, it is still poorly understood how hypoxia contributes to the aberrant post-transcriptional regulation observed in tumours. In this PhD thesis, I studied the RNA response to hypoxia in cancer, to provide novel insights into its regulation.

Using deep RNA-Sequencing (RNA-Seq), I investigated transcriptome changes of three human cell lines from lung, cervical and breast cancer under hypoxia, advancing our knowledge of post-transcriptional gene regulation in hypoxic cancer. I show that hypoxia induced consistent changes in transcript abundance in the three cancer types. This was coupled to divergent splicing responses, highlighting the cell type specificity of alternative splicing programs. While the mRNA levels of RNA-binding proteins were mainly reduced, hypoxia upregulated muscleblind-like protein 2 (*MBNL2*) in all three cell lines. Hypoxia control was specific for *MBNL2*, since it did not affect its paralogs *MBNL1* and *MBNL3*. Via knockdown experiments of *MBNL2* in hypoxic cells, I could show that *MBNL2* induction promotes adaptation of cancer cells to low oxygen by regulating both transcript abundance and alternative splicing of hypoxia response genes. In addition, depletion of *MBNL2* reduced the proliferation and migration of cancer cells, corroborating a function of *MBNL2* as cancer driver.

In the last few years, a novel class of RNAs has gained attention, namely circular RNAs (circRNAs), which are produced by a particular splicing mechanism, known as back-splicing. CircRNAs have been reported to change their abundance in cancer and their high stability makes them promising candidates as diagnostic biomarkers. In this study, I took advantage of deep rRNA-depleted RNA-Seq data to comprehensively investigate the expression of circRNAs in human cancer cells and their changes in response to hypoxia. To reliably identify circRNAs, I established a pipeline that integrates two available tools

for circRNA detection with custom approaches for quantification and statistical analysis. Using this pipeline, I identified 12006 circRNAs in the three cancer cell lines. Their molecular features suggest an involvement of complementary RNA sequences as well as *trans*-acting factors in circRNA biogenesis, including the splicing factor HNRNPC. Remarkably, I detected 210 circRNAs that are more abundant than their linear counterparts. Upon hypoxic stress, 64 circRNAs were differentially expressed in cancer cells, in most cases in a cell type-specific manner. In summary, in this PhD thesis, I present a comparative transcriptome profiling in human cancer cell lines. It reveals MBNL2 as an important player in hypoxic cancer progression and provides novel insights into the biogenesis and regulation of circRNAs under hypoxic stress.

Preface

The content of this thesis is based on a research collaboration between the group of Dr. Kathi Zarnack (BMLS, Frankfurt am Main) and experimental scientists in the groups of Prof. Dr. Michaela Müller-McNicoll (Goethe University, Frankfurt am Main) and Dr. Julia E. Weigand (Technical University, Darmstadt). During my PhD, I implemented the computational pipeline and performed all bioinformatics data analyses reported in this thesis. Camila de Oliveira Freitas Machado and Sandra Fischer performed hypoxia treatments, RNA-Sequencing, and validation experiments. For the sake of completeness, also experimental results are reported in the thesis. The project was conceived by Dr. Kathi Zarnack and Prof. Dr. Michaela Müller-McNicoll as regards the investigation of circular RNAs in hypoxia, and Dr. Julia E. Weigand for the study about the role of *MBNL2* in hypoxia. All bioinformatics analyses were performed under the supervision of Dr. Kathi Zarnack, with additional supervision by Dr. Stefanie Ebersberger (IMB, Mainz). Partial results of the presented work have been published in the following article:

Di Liddo, A., de Oliveira Freitas Machado, C., Fischer, S., Ebersberger, S., Heumüller, A. W., Weigand, J. E., Müller-McNicoll, M. & Zarnack, K. A combined computational pipeline to detect circular RNAs in human cancer cells under hypoxic stress. *Journal of molecular cell biology* **11**, 829–844 (2019);

or collected in the manuscript:

Fischer, S., Di Liddo, A., Taylor, K., Sobczak, K., Zarnack, K. & Weigand, J. E. Muscleblind-like 2 controls the hypoxia response of cancer cells (*in revision*).

Chapter 1

Introduction

1.1 The human transcriptome

The central dogma of molecular biology describes the expression of a gene via a two-step process, by which the information is transferred from DNA to RNA via transcription, and from RNA to protein via translation. The RNA molecule that serves as a template for protein translation is named messenger RNA (mRNA), and the full set of RNA molecules in a cell or a population of cells is defined as transcriptome. In 2012, the ENCODE project revealed that, although more than 70% of the genome is transcribed into RNA, only a small proportion of the transcriptome is finally translated into protein (2%) (Djebali *et al.*, 2012). The latest release of GENCODE annotation of the human genome reports 19975 protein-coding genes representing only one third of the 60603 total genes (<https://www.gencodegenes.org/human/stats.html>, version 31). Thus, most RNA molecules can be the final product in themselves. These RNAs are defined as non-coding RNAs (ncRNA) and include the well-characterised transfer RNAs (tRNA) and ribosomal RNAs (rRNAs), involved in the translation of mRNAs, as well as small nuclear RNAs (snRNAs), microRNAs (miRNAs), small interfering RNAs (siRNAs) and Piwi-interacting RNAs (piRNAs). Another class of ncRNAs is represented by long non-coding RNAs (lncRNAs) defined by the size of the transcript longer than 200 nucleotides (nt). In the last few years, a novel class of ncRNAs has gained widespread attention in genomics studies, namely circular RNAs (circRNAs).

1.2 The complex life of mRNA

The transcription of a protein-coding gene consists of copying the genetic information stored in a DNA segment (template DNA) into a precursor mRNA (pre-mRNA). A schematic of the pre-mRNA transcription and processing is shown in Figure 1.1.

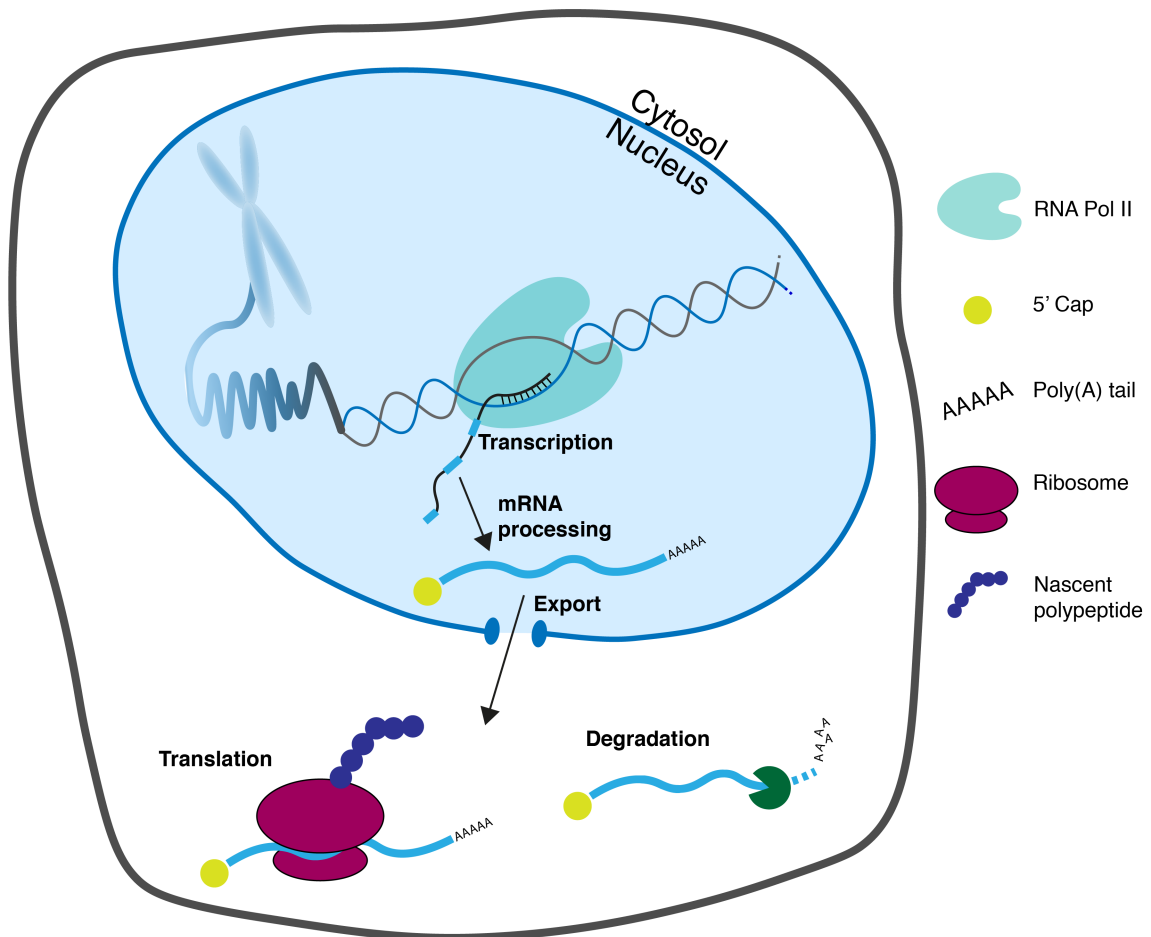


Figure 1.1: Overview of the mRNA life cycle. The genetic information encoded in genes in the DNA packed in chromosomes is transferred to RNA molecules via transcription. The resulting transcript undergoes multiple maturation steps, from capping to nuclear export and translation.

In eukaryotes, the transcription of mRNAs is catalysed by the RNA Polymerase II (RNA Pol II) and occurs in the nucleus of the cell. At this point, the pre-mRNA is composed of exons and introns. Exons constitute the sequences that are preserved in the mature mRNA, while introns are non-coding regions that are not present in the final mature mRNA. Exons may constitute the coding region of the pre-mRNA, which is translated into the final protein, or they may contain non-coding sequences that are not translated (untranslated

regions, UTRs) but serve important roles in the regulation of the mRNA life. Before getting exported to the cytosol to be translated, the pre-mRNA undergoes a series of modifications to generate a mature transcript. The first modification is the addition of a cap structure at the 5' end of the transcript, consisting of a modified guanine base that protects the RNA from degradation by exonucleases. The capping is followed by the removal of introns and the connection of exons in a process known as splicing. Finally, the pre-mRNA undergoes cleavage at the 3' end, strictly coupled to the addition of multiple adenine residues to its 3' end, in a process named polyadenylation. The pre-mRNA is then detached from the template DNA and released, with the subsequent termination of the transcription. When all processing steps are carried out, the mature mRNA is ready to be transferred to the cytosol via recruitment of the nuclear export factor 1 (NXF1), that interacts with the nuclear pore complex (Müller-McNicoll & Neugebauer, 2013). Once in the cytoplasm, the mRNA can get localised to specific regions, such as membrane compartments, and can be translated into proteins in a process catalysed by the ribosome. There can be a wide diversity of events which alter the form of the eukaryotic mRNAs, including alternative transcription start, alternative splicing, and alternative polyadenylation. These events lead to a wide diversity of transcript isoforms being produced from a single gene, which can influence the nature of the produced protein. During the process of RNA biogenesis in the nucleus, a number of factors begin to associate with the mRNA. These include RNA-binding proteins (RBPs), which bind directly to the transcript. Many of these factors then travel out to the cytoplasm together with the mRNA, constituting a code that drives the localisation, translation and degradation of the mRNA (Singh *et al.*, 2015).

1.3 The pre-mRNA splicing

The pre-mRNA splicing is a fine-tuned process in which exons are covalently joined together to generate the mature transcript (Figure 1.2A). Intron excision and exon ligation are achieved via two consecutive transesterification reactions (Figure 1.2B). This process constitutes the basis to generate protein diversity from a single gene. Splicing mostly occurs co-transcriptionally in humans (Tilgner *et al.*, 2012) and is catalysed by a large macromolecular ribonucleoprotein complex, the spliceosome. This complex is composed of five uridine-rich snRNAs (U1, U2, U4, U5 and U6) and about 200 proteins that catalyse the different steps of the RNA splicing in a dynamic manner. The spliceosome assembly is guided by the presence of specific sequences in the intron to be spliced out. These are the 5' splice site (5'SS), also known as donor site, a branch point sequence (BPS), a polypyrimidine tract (PPT), and the 3' splice site (3'SS), also known as acceptor site (Wahl *et al.*,

2009; Shi, 2017; Fica & Nagai, 2017; Yan *et al.*, 2019) (Figure 1.2A).

During the earliest stages of the spliceosome assembly, the 5'SS is recognised by U1 snRNP, while the recognition at the 3'SS is initiated by the heterodimer U2 small nuclear RNA auxiliary factor (U2AF). U2AF consists of the subunits U2AF1 and U2AF2, which bind to the AG dinucleotide at the 3'SS and the PPT immediately upstream of the 3'SS, respectively. The splicing factor 1 (SF1) binds to the BPS (Berglund *et al.*, 1997; Berglund *et al.*, 1998; Liu *et al.*, 2001) (Figure 1.3). This early arrangement is named E complex and is converted to the A complex (pre-spliceosome complex) when U2 snRNP replaces SF1 at the BPS. Next, the tri-snRNP U4, U5 and U6 joins the spliceosome, thereby forming the B complex. At this point the spliceosome undergoes a series of compositional and conformational changes that lead to the formation of an active B complex. This is able to carry out the first step of splicing, that consists of the disruption of the phosphodiester bond at the 5'SS and the link of the free end of the intron to the adenosine (A) at the BPS. Further conformational changes affect the spliceosome, with the formation of a C complex, able to carry out the second step of the splicing reaction, leading to the formation of a phosphodiester bond between the two exons and the removal of the intron as a lariat structure (Wahl *et al.*, 2009; Shi, 2017; Fica & Nagai, 2017; Yan *et al.*, 2019) (Figure 1.3). The components of this post-spliceosomal complex are then disassembled and recycled for further splicing reactions, while the ligated exons are released. During the transition between the multiple complexes, a large number of additional proteins that do not directly constitute the spliceosome contribute to its assembly and activation. These include SR proteins, heterogeneous nuclear ribonucleoproteins (HNRNPs), and others (Fu & Ares, 2014; Jeong, 2017).

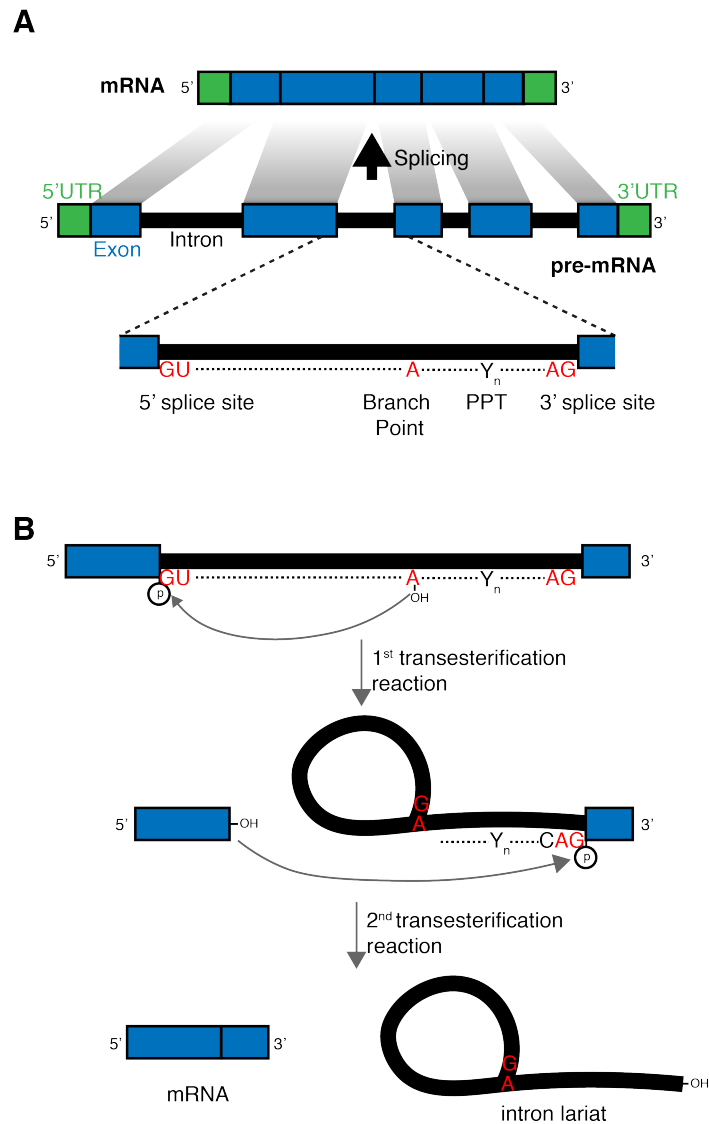


Figure 1.2: Schematic of the pre-mRNA splicing. (A) Top: Pre-mRNA is converted into mRNA via connection of exons (blue) and removal of the intervening introns (black). Untranslated regions (UTRs, green) are preserved in the mRNA molecule. Bottom: Enlarged pre-mRNA region including consensus sequences in the intron, which are essential for its excision: 5' and 3' splice sites, branch point and polypyrimidine tract (PPT).

(B) Pre-mRNA splicing consists of two sequential transesterification reactions. The first reaction involves the adenosine at the branch point and the 5' splice site. Next, the first exon is joined to the downstream exon in the second transesterification reaction, thereby excising the intron lariat previously formed.

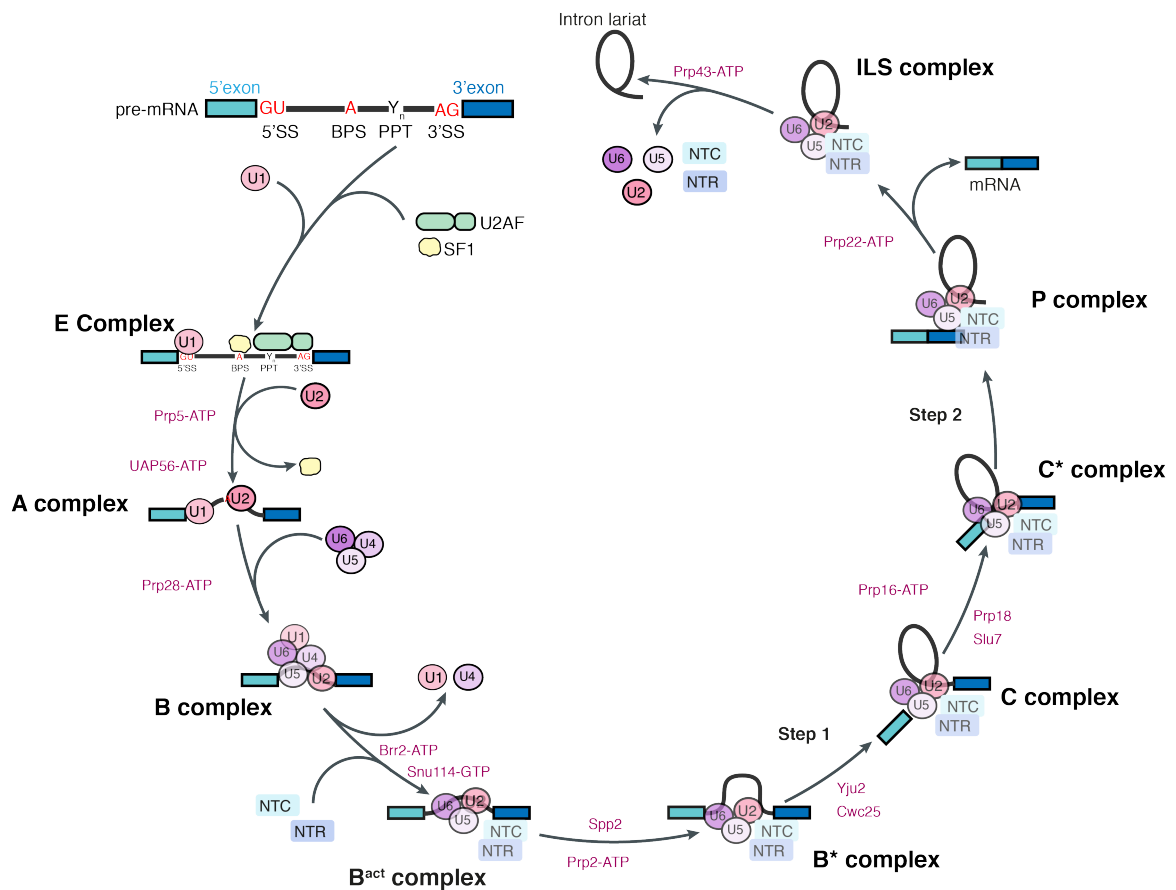


Figure 1.3: Spliceosome assembly step-by-step. Core splicing signals within introns of pre-mRNAs include a 5' splice site (5'SS), a branch point (BPS), a polypyrimidine tract (PPT), and a 3' splice site (3'SS) (Figure 1.2). The splicing signals are bound by early splicing factors, including U1 snRNP, SF1 and the U2AF heterodimer, thereby forming the E complex. Later on, U2 snRNP joins the complex, converting it into the A complex. The recruitment of the U4-U6-U5 tri-snRNP on the pre-mRNA generates the pre-catalytic B complex and is required to excise the intron. The spliceosome assembly is highly dynamic and additional factors transiently interact with the pre-mRNA during the different steps of the splicing process. E: early complex; A: pre-spliceosome complex; B: pre-catalytic spliceosome; B^{act} : activated spliceosome; B^* : catalytically activated spliceosome; C: catalytic step 1 spliceosome; C^* : catalytic step 2 spliceosome; P: post-splicing complex; ILS: intron lariat spliceosome; NTC: NineTeen Complex; NTR: NTC-related complex. Adapted from Shi, 2017.

1.4 Alternative splicing

During the splicing process, a key step is the recognition of splice sites. Constitutive exons are generally included in the mature transcript, due to the presence of a highly conserved and strong splice signal that favours their splicing (constitutive splicing, Figure 1.4). Other exons can be alternatively included in the mature transcripts depending on the distinct selection of splice sites during pre-mRNA splicing. This category of exons is defined as alternative exons. The alternative inclusion of certain exons in the mature transcript is known as alternative splicing (AS) and constitutes the major post-transcriptional process that ensures biological complexity by increasing the proteome diversity. Indeed, it has been estimated that about 95% of the human genes undergo alternative splicing (Pan *et al.*, 2008; Kornblihtt *et al.*, 2013). Even from an annotation point of view, the latest release of GENCODE annotation of the human genome reports 83666 protein-coding transcripts and 19975 protein-coding genes (<https://www.encodegenes.org/human/stats.html>, version 31), indicating that most pre-mRNAs produce multiple different mRNA isoforms. In general, alternative exons have weaker splice signals compared to constitutive exons, and their recognition and usage is influenced by the presence of *cis*-acting elements in the pre-mRNA, named exonic enhancer or silencer (ESE, ESS) when located in exons, and intronic splicing enhancers or silencer (ISE and ISS) when located in introns. These elements are bound by *trans*-acting factors defined as splicing factors that, with their activity, either enhance or inhibit the usage of alternative splice sites by influencing the recruitment of the spliceosomal machinery (Roy *et al.*, 2013; Wang *et al.*, 2015b). In addition, also the transcription rate plays an important role in alternative splicing regulation, with the splicing of alternative exons being favoured when transcriptional elongation is slowed down and the RNA polymerase pauses (Kornblihtt, 2007).

Different types of AS events might contribute to the diversity of transcript isoforms from a single gene (Figure 1.4). These include cassette exon (CE) events, also known as exon skipping, in which an internal exon of the pre-mRNA can alternatively be spliced in or skipped to generate two different transcript isoforms. More rarely, AS can affect also two consecutive exons that are alternatively included in the mature transcript. This kind of event is defined as mutually exclusive exons (MXE) (Pohl *et al.*, 2013). Different transcript isoforms can be generated also via alternative selection of 5' or 3' splice sites, that may lead to subtle changes in the coding sequence (Wang *et al.*, 2015b). Finally, AS can also affect intronic regions, when an intron is either retained or removed from the mature transcript (retained intron). Intron retention has been observed for more than half of all human introns (Braunschweig *et al.*, 2014; Jacob & Smith, 2017). For a long time, the importance of AS has not been fully appreciated. The development of high-throughput

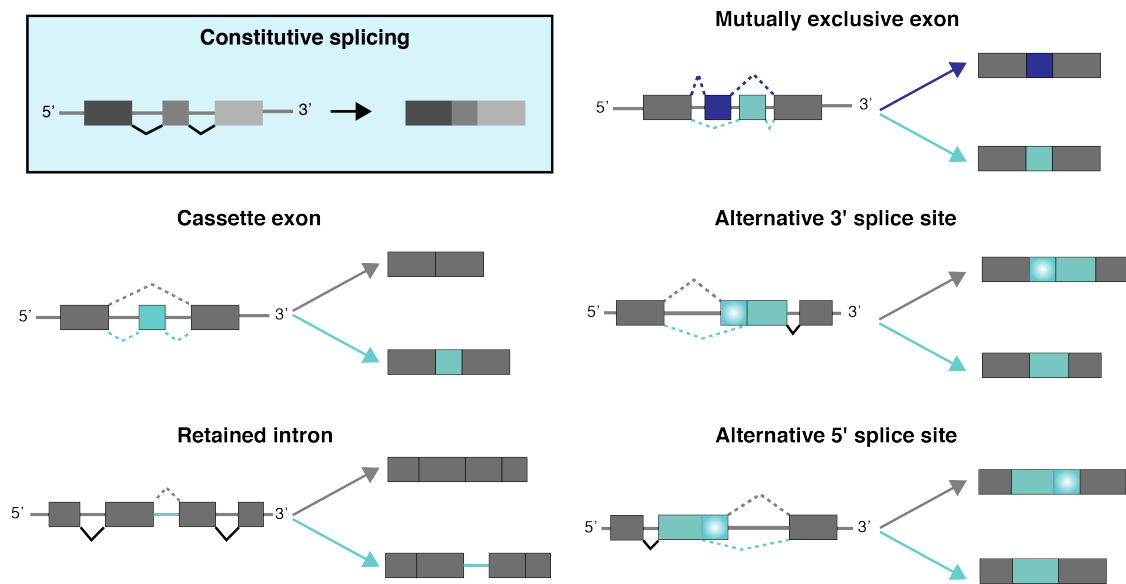


Figure 1.4: Constitutive and alternative splicing. Different types of alternative splicing are illustrated, together with their products, in comparison to constitutive splicing. Modified from Wang *et al.*, 2015b.

technologies and computational tools able to profile the AS pattern largely extended the current knowledge of the impact of AS on different biological contexts. Alterations in the AS process have been associated with multiple human hereditary diseases as well as with different forms of cancer (Oltean & Bates, 2014; Ciepły & Carstens, 2015; Urbanski *et al.*, 2018; Yang *et al.*, 2019; Coomer *et al.*, 2019). Studying the link between cancer biology and splicing regulation is fundamental to understand the influence on the disease and to develop novel anti-cancer therapeutic approaches (Coltri *et al.*, 2019).

1.5 Alternative splicing regulators: MBNL proteins

As mentioned in the previous sections, the splicing process is regulated by a multitude of RBPs that localise at specific binding sites on the pre-mRNA and influence the splicing of alternative exons, thus functioning as splicing factors. Splicing factors define the specific set of mRNA isoforms and their encoded proteins that characterise a certain tissue or developmental stage. Splicing factors include SR proteins, containing serine/arginine-rich motifs, and heterogeneous nuclear ribonucleoproteins (HNRNPs) (Fu & Ares, 2014). Classically, SR proteins are thought as activators of AS, while HNRNPs are seen as silencers. However, recent studies showed a more complex scenario, in which both SR proteins and HNRNPs can enhance or inhibit splicing. Their action depends on the context, whether they bind to alternative exons or introns (Dvinge *et al.*, 2016; Fu & Ares, 2014). Many other RBPs are involved in splicing regulation, often regulating cell- and tissue-specific splicing events, including the CCUG-BP and ETR-3-like factors (CELFs), RBFOX proteins, NOVA proteins, ELAVL1 (ELAV Like RNA Binding Protein 1, also known as HuR), T cell-restricted intracellular antigen 1 (TIA1) and TIA1-like (TIAL1) (Dvinge *et al.*, 2016; Fu & Ares, 2014). Also the muscleblind-like (MBNL) protein family belongs to this group of *trans*-acting factors that regulate alternative splicing.

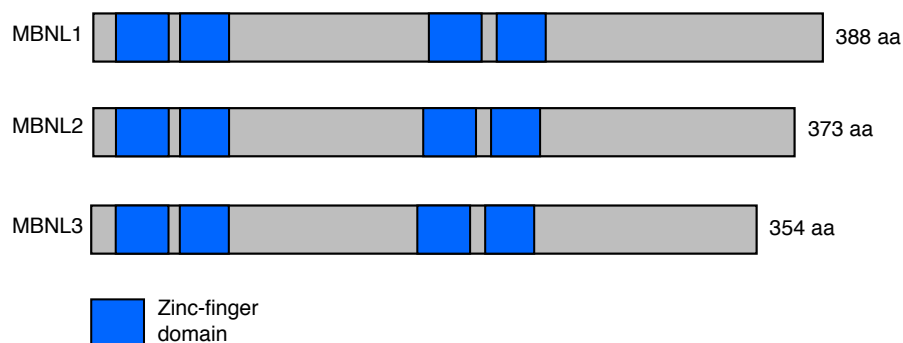


Figure 1.5: The muscleblind-like (MBNL) protein family. Schematic representation of MBNL proteins and their domain structure, as described in the UniProt database (<https://www.uniprot.org/>; UniProt entries Q9NR56, Q5VZF2, and Q9NUK0 for MBNL1, MBNL2 and MBNL3, respectively). All three paralogs present two pairs of zinc finger domains, which are able to bind YGCY motifs with different affinities and specificities.

In human, the MBNL protein family consists of three paralogs, MBNL1, MBNL2 and MBNL3, which show a different expression pattern across tissues and developmental stages (Konieczny *et al.*, 2014). *MBNL1* and *MBNL2* are both ubiquitously expressed, with *MBNL1* primarily exerting its function in most tissues. Only in brain, *MBNL2* is the

predominant expressed paralog in respect to *MBNL1* (Konieczny *et al.*, 2014). In contrast, *MBNL3* expression has been shown to be limited to muscle cells during differentiation and regeneration, and it reaches high levels in placenta (Fardaei *et al.*, 2002; Squillace *et al.*, 2002; Lee *et al.*, 2010; Poulos *et al.*, 2013). All the three members of the family contain four zinc finger (ZnF) domains of the type C3H1. These are organised in two pairs, which are separated by a linker region of about 80 residues (Taylor *et al.*, 2018). The ZnF domains serve to recognise and bind specific sequences on mRNA molecules, consisting of two or more clustered 5'-YGCY-3' motifs (Lambert *et al.*, 2014). The ability of MBNL proteins to bind RNA does not depend exclusively on the primary sequence of the binding site, but also on the flanking nucleotides (Park *et al.*, 2017), as well as the structural context of the RNA targets (deLorimier *et al.*, 2017; Taylor *et al.*, 2018). In general, AS is controlled by a differential distribution of MBNL paralogs and the affinity of these proteins for specific RNA-binding regions. Similar to other splicing factors, MBNLs activate or repress mRNA alternative splicing depending on their binding location (Charizanis *et al.*, 2012; Wang *et al.*, 2012). Among the three MBNL paralogs, MBNL1 is considered the most potent splicing regulator (Sznajder *et al.*, 2016). In addition to alternative splicing, MBNL proteins are involved in the regulation of several other steps in the mRNA life, including mRNA localisation (Adereth *et al.*, 2005; Wang *et al.*, 2012), stability (Masuda *et al.*, 2012), local translation, degradation, as well as alternative polyadenylation (Batra *et al.*, 2014). Finally, MBNLs have been reported to influence miRNA biogenesis (Rau *et al.*, 2011). In particular, MBNL1 has been implicated in the negative regulation of mRNA stability (Masuda *et al.*, 2012; Wang *et al.*, 2015a). In contrast, it has been suggested that MBNL2 might enhance the stability of mRNAs encoding extracellular matrix components (Du *et al.*, 2010), and a recent publication predicted that MBNL2 might function as mRNA-stabilising factor (Perron *et al.*, 2018). Finally, all three MBNL paralogs are involved in autoregulatory feedback loops, and often can compensate each other in their function (Konieczny *et al.*, 2018).

MBNL proteins are strongly associated to diseases, such as myotonic dystrophy, in which their availability is reduced due to the expansion of CUG and CCUG repeats, which causes the sequestration of MBNL proteins (deLorimier *et al.*, 2017; Du *et al.*, 2010; Miller *et al.*, 2000). It is known that post-transcriptional regulation, mediated by RBPs, is often globally deregulated in cancer (Oltean & Bates, 2014; Escobar-Hoyos *et al.*, 2019; Khabar, 2017), and aberrant RBP activities have been identified as cancer drivers (Anczuków & Krainer, 2016; Pereira *et al.*, 2017). While the importance of MBNL proteins in myotonic dystrophy has been largely confirmed, only recently MBNL proteins have been shown to play a role in tumour progression, acting either as oncogene or tumour suppressor. For instance, *MBNL1* was described as tumour suppressor gene in breast cancer metastasis

(Fish *et al.*, 2016). In contrast, *MBNL1* was also shown to contribute to carcinogenesis in colorectal cancer by interfering with the recruitment of DICER1 to miRNAs (Tang *et al.*, 2015). These contrasting functions of *MBNL1* in different cancers derive from different levels of specific isoforms (Tabaglio *et al.*, 2018). Similarly, *MBNL2* has been reported to function either as tumour suppressor in hepatocarcinogenesis (Lee *et al.*, 2016) or as oncogene in clear cell renal cell carcinoma (Perron *et al.*, 2018). Finally, *MBNL3* was found to promote hepatocellular carcinoma (Yuan *et al.*, 2017). Despite these first evidences of a role of MBNL proteins in cancer, the molecular mechanism by which they influence cancer progression and their impact at transcriptome level requires additional investigation.

1.6 Back-splicing generates circular RNAs

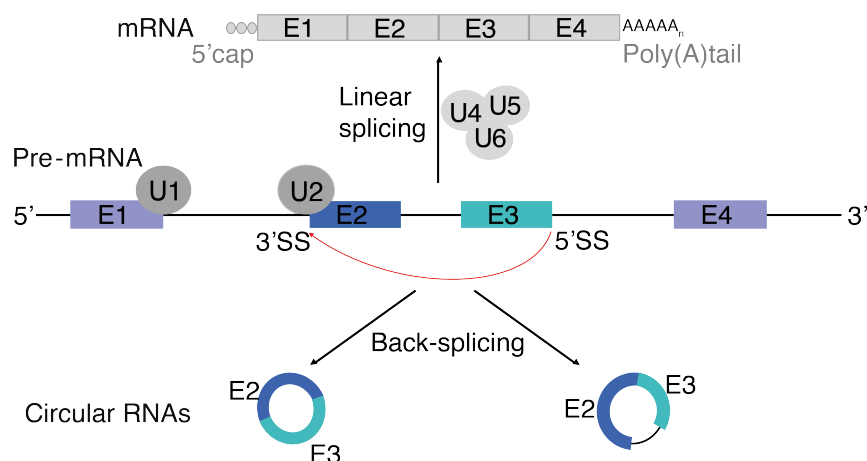


Figure 1.6: Back-splicing generates circRNAs. The majority of circRNAs originate from pre-mRNA. In linear splicing, a 5'SS is covalently joined to a downstream 3'SS to generate the messenger RNA. The mechanism by which a 5'SS is joined to a 3'SS located upstream with respect to the transcription direction is named back-splicing and produces circRNAs. Introns can be either removed or retained in the final circular transcript.

In addition to linear splicing, exons in the pre-mRNA can undergo another form of alternative splicing, when a 5' splice site is covalently joined to a 3' splice site that is located upstream in the pre-mRNA. This process is known as back-splicing or head-to-tail splicing, and leads to the formation of circular RNAs (circRNAs) (Figure 1.6). The first evidence of circRNAs dates back to 1976, when Sanger *et al.* discovered them as viroids in RNA viruses (Sanger *et al.*, 1976). Three years later, the first circRNA in eukaryotic cells was detected by electron microscopy (Hsu & Coca-Prados, 1979), followed by the observation of a circRNA from the hepatitis δ virus (Kos *et al.*, 1986). In the next years, circRNAs

were identified from the *DCC* gene and described as scramble exons (Nigro *et al.*, 1991), and from the sex-determining region Y (*Sry*) in mouse testis (Capel *et al.*, 1993; Dubin *et al.*, 1995). Despite these first observations of circular RNA molecules, for many years circRNAs have been considered the product of splicing errors (Cocquerelle *et al.*, 1993). Thanks to major progress in high-throughput technologies, for the first time in 2012 it

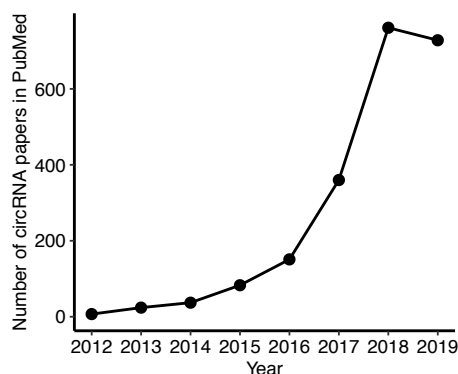


Figure 1.7: Publication rates of circRNAs. Shown is the number of publications referring to circRNAs reported in PubMed as updated to 11/07/2019. The terms "circular RNA" OR circRNA OR circRNAs OR "circular RNAs" were queried. For the sake of simplicity, only PubMed records from 2012 are shown.

has been possible to detect thousands of circRNAs from RNA-Sequencing experiments of patients with acute lymphoblastic leukemia, as well as of normal and cancer cell lines (Salzman *et al.*, 2012). Only one year later, two other studies used bioinformatics approaches to detect circRNAs in eukaryotic cells (Memczak *et al.*, 2013; Jeck *et al.*, 2013). Among the identified thousands of circRNAs, CDR1as/ciRs-7 emerged as miRNA sponge of miR-7 (Hansen *et al.*, 2013; Memczak *et al.*, 2013). From this point, an exponential number of circRNA-related studies have been published, accumulating to a total of 2365 papers reported in PubMed to date (<https://www.ncbi.nlm.nih.gov/pubmed/>), of which 1489 (63%) were published in 2018-2019 (Figure 1.7).

CircRNAs lack 5' and 3' free ends, making them highly stable when compared to linear RNAs, because they are resistant to the action of endogenous exonucleases. Indeed, it has been estimated that the half-life of a circRNA ranges up to 48 hours (h) (Jeck *et al.*, 2013), against 4-9 h for an mRNA (Schwanhäusser *et al.*, 2011). In addition, due to their circular conformation, circRNAs do not undergo polyadenylation. Different types of circRNAs have been discovered. The majority of circRNAs originate from protein-coding genes. Those containing exclusively exonic sequences are referred to as exonic circRNAs (ecircRNAs) and can include a single or multiple exons. CircRNAs can also contain both exons and introns; these are known as EIciRNA, and are often retained in the nucleus (Li *et al.*, 2015).

In contrast, ecircRNAs show cytoplasmic localisation (Zhang *et al.*, 2014). Additional types of circRNAs include intronic circRNAs or ciRNAs (Zhang *et al.*, 2013), formed from intron lariats, as well as circRNAs from untranslated regions (UTRs), non-coding loci and intergenic regions (Memczak *et al.*, 2013; Guo *et al.*, 2014; Zheng *et al.*, 2016). CircRNAs from the interior regions of exons, introns, and intergenic transcripts in human and mouse have also recently been described and referred to as interior circRNAs or i-circRNAs (Liu *et al.*, 2019). Thus, circRNAs constitute a large heterogeneous class of RNAs with the common feature of being covalently closed molecules. CircRNAs have been detected in various organisms, from Archaea to mammals (Danan *et al.*, 2011; Salzman *et al.*, 2012; Wang *et al.*, 2014), with many circRNAs being evolutionarily conserved across species (Pamudurti *et al.*, 2017). CircRNAs were reported to be generally lowly abundant (Jeck *et al.*, 2013; Guo *et al.*, 2014; Salzman *et al.*, 2013). However, several circRNAs have been shown to be more expressed than their linear counterparts, especially in neuronal tissues (Salzman *et al.*, 2013; Jeck *et al.*, 2013; Memczak *et al.*, 2013). In humans, circRNAs have been detected in most tissues and cell types, and multiple transcriptome-wide studies reported cell-, tissue- and developmental stage-specific expression patterns for circRNAs (Salzman *et al.*, 2013; Memczak *et al.*, 2013; Rybak-Wolf *et al.*, 2015; Kristensen *et al.*, 2017b). The tissue and developmental stage specificity, together with the conservation and the high expression levels of certain circRNAs, point to a functional relevance of such circular molecules. However, how the regulation of circRNAs in different tissues, developmental stages or cellular conditions is achieved, is not fully understood.

Biogenesis of circRNAs The back-splicing process involves the canonical spliceosomal machinery (Ashwal-Fluss *et al.*, 2014; Starke *et al.*, 2015; Wang & Wang, 2015). The steady-state levels of circRNAs were shown to be increased upon inhibition of spliceosome components at the expense of linear RNAs from the same pre-mRNA (Liang *et al.*, 2017a). Moreover, inhibiting the RNA Pol II termination led to increased circRNA levels, most likely due to read-through transcription that continues into the downstream genes, resulting in transcripts that undergo back-splicing (Liang *et al.*, 2017a). Furthermore, it was found that the average transcription elongation rate (TER) of RNA Pol II for nascent circRNA-producing genes is higher than for non-circRNA genes (Zhang *et al.*, 2016b). As for alternative splicing (Braunschweig *et al.*, 2013), a relatively modest variation of the TER had an appreciable effect on circRNA formation. A large fraction of nascent circRNAs was detected only after the completion of transcription of their host pre-mRNAs (Zhang *et al.*, 2016b), indicating that back-splicing largely occurs post-transcriptionally (Li *et al.*, 2018b).

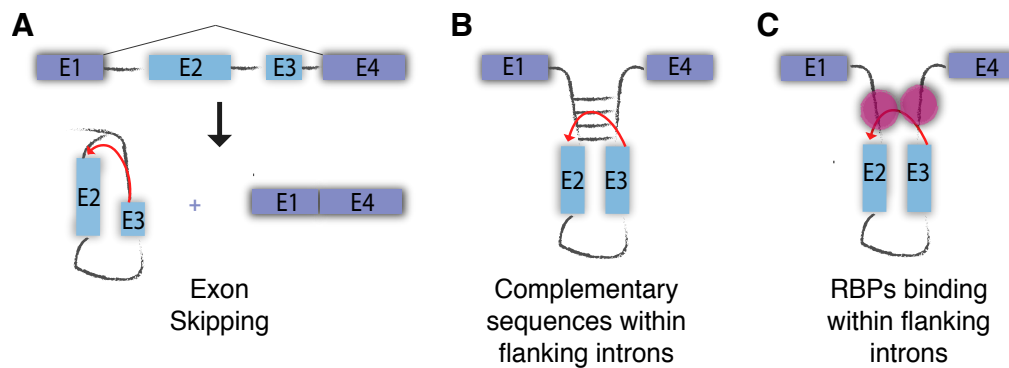


Figure 1.8: Hypothetical mechanisms for circRNA biogenesis. Three mechanisms have been proposed for generating circRNAs. **(A)** Lariat-driven: Exon skipping leads to the formation of an mRNA as well as a lariat structure from the skipped exons. The lariat is then circularised in a second splicing reaction. **(B)** Intron-pairing-driven: Complementary regions in introns flanking the circRNAs pair and move back-splice sites in close proximity, enhancing the circularisation. **(C)** RBP-mediated: RBPs bind both flanking introns and bring the back-splice sites closer by interacting with each other, as shown for the splicing factor Quaking (QKI) (Conn *et al.*, 2015). Adapted from Ebbesen *et al.*, 2016.

Currently, there are three hypothetical mechanisms suggested for the biogenesis of circRNAs: lariat-driven, intron pairing-driven and RBP-mediated back-splicing (Figure 1.8). According to the lariat-driven model, during pre-mRNA transcription an exon skipping event might generate an exon-containing lariat structure that is further processed to remove the intron and close the circular transcript. In parallel, an mRNA lacking the skipped and circularised exons is produced (Jeck *et al.*, 2013) (Figure 1.8A). The best characterised mechanism requires the presence of complementary inverted repeats, often *Alu* element retrotransposons (Zhang *et al.*, 2014), in introns flanking the circularised exons. Pairing of these sequences moves the back-splice sites in close proximity, favouring back-splicing (Jeck *et al.*, 2013) (Figure 1.8B). Another mechanism involves RBPs that bind both flanking introns and bring the back-splice sites closer by interacting with each other. This is the case for the splicing factor Quaking (QKI) (Conn *et al.*, 2015) (Figure 1.8C). Additional RBPs have been shown to regulate the circRNA formation, including MBL/MBNL1 (Ashwal-Fluss *et al.*, 2014), as well as HNRNPs and SR proteins in *Drosophila melanogaster* (Kramer *et al.*, 2015), and human (Fei *et al.*, 2017). The splicing factor FUS has been shown to regulate circRNA formation by binding introns flanking back-splicing junctions in mouse motor neurons (Errichelli *et al.*, 2017). Moreover, the intron-driven mechanism of circRNA biogenesis is subjected to inhibition by A-to-I editing operated by the enzyme ADAR1 (Ivanov *et al.*, 2015; Rybak-Wolf *et al.*, 2015). Similarly, the nuclear RNA helicase DHX9 suppresses the formation of circRNAs by disrupting RNA pairs that flank circularised exons (Aktas *et al.*, 2017). Finally, similarly to alternative

splicing, the biogenesis of circRNAs is likely to be influenced by a combination of *cis*-acting and *trans*-acting splicing factors that can either positively or negatively interfere with the back-splicing process.

Functions of circRNAs Although thousands of circRNAs have been detected in various organisms, the molecular function of the majority of circRNAs remains unknown. A number of circRNAs have been reported to act as miRNA sponge, presenting multiple binding sites for miRNAs, with CDR1as/ciRs-7 being the best characterised, with more than 60 conserved binding sites for miRNA-7 (Hansen *et al.*, 2013; Memczak *et al.*, 2013; Wang *et al.*, 2016; Zheng *et al.*, 2016; Han *et al.*, 2017b; Han *et al.*, 2017a). By sequestering miRNAs and reducing their activity, circRNAs indirectly upregulate the expression of miRNA target genes. A recent study reported a regulatory network through sponge function in the mammalian brain that involves the lncRNA *Cyrano*, ciRS-7 and two microRNAs miR-671 and miR-7 (Kleaveland *et al.*, 2018). CircRNAs have been shown to sequester not only miRNAs but also RBPs. For instance, circMbl acts as a sponge for RNA-binding protein MBL encoded by the same gene, thereby regulating the protein expression in a feedback loop. It also competes with the linear splicing of pre-mRNA and affects the formation of linear RNA to regulate the expression of related genes (Ashwal-Fluss *et al.*, 2014). EicircRNAs, such as EicIPAIP2 and EicEIF3J, are retained in the nucleus and regulate the transcription of their own host gene by interacting with RNA Pol II (Li *et al.*, 2015). Additionally, it has been suggested that circRNAs can act as protein scaffolds, as for circFOXO3 that represses the cell cycle progression by forming a complex with p21 and CDK2 (Du *et al.*, 2016). Although initially considered exclusively non-coding RNAs, circRNAs have recently been reported to contain internal ribosome entry sites (IRES) and to be translated into small peptides (Legnini *et al.*, 2017; Pamudurti *et al.*, 2017; Yang *et al.*, 2017).

CircRNAs have been found to be deregulated in several human tumours, including lung, cervical and breast cancer. They were efficiently used to distinguish tumours from adjacent normal tissue (Geng *et al.*, 2018; Kristensen *et al.*, 2017a). In addition, several studies have reported that circRNAs can regulate cellular stress (Fischer & Leung, 2017). Boeckel and colleagues identified circRNAs that are regulated in human endothelial cells upon hypoxic stress, and showed that circZNF292 promotes angiogenesis (Boeckel *et al.*, 2015). circZNF292 was also reported to modulate cell proliferation and tube formation in human glioma (Yang *et al.*, 2016). Similarly, a circRNA from the *DENND4C* gene was found to be induced in human breast cancer cells (MCF-7) in hypoxic conditions (Liang *et al.*, 2017b). Nevertheless, the influence of hypoxia on the circRNA repertoire in cancer cells remains to be fully explored.

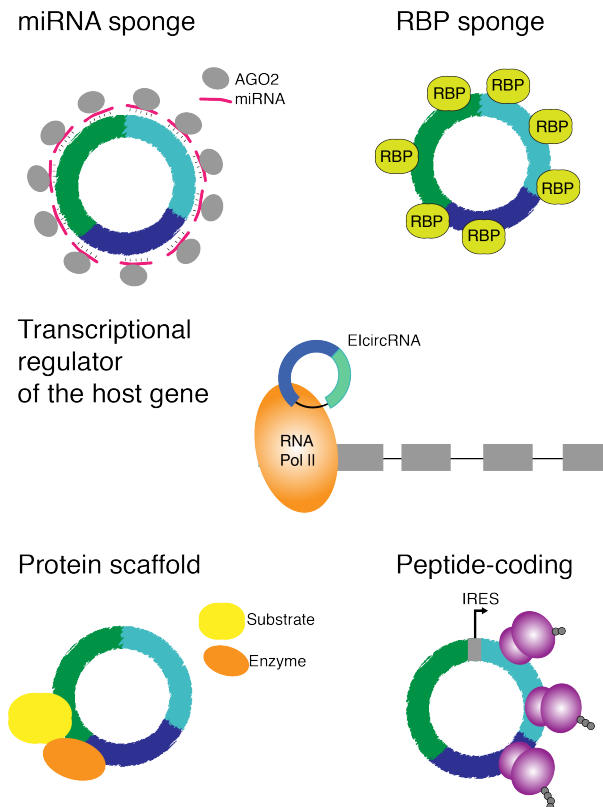


Figure 1.9: Functions of circRNAs. CircRNAs can function as miRNA sponges, thereby indirectly impairing the degradation or promoting the translation of miRNA targets. CircRNAs can be bound by RBPs that recognise binding sites on their sequence, thereby regulating the availability of RBPs. EIcircRNAs have been shown to associate with RNA Pol II and enhance the transcription of their host genes. CircRNAs can act as scaffold for proteins, mediating their interaction. CircRNAs with IRES elements and AUG sites can function as template for translation of peptides. Adapted from Kristensen *et al.*, 2019.

1.7 RNA-Sequencing: a powerful tool for transcriptomics studies

To date, the most suited techniques for transcriptomics studies are microarray and RNA-Seq. While microarray assays have been set up for genome-wide studies on circRNAs (Zhang *et al.*, 2018; Tang *et al.*, 2018; Su *et al.*, 2016; Sand *et al.*, 2016; Zhang *et al.*, 2017; Liu *et al.*, 2017), currently RNA-Seq represents the method of choice for circRNA profiling. In contrast to microarray technology, RNA-Seq does not require prior knowledge of existing circRNAs, allowing *de novo* identification of circRNAs. Moreover, the currently available microarrays do not provide any information about the levels of the linear counterpart of circRNAs (Kristensen *et al.*, 2019).

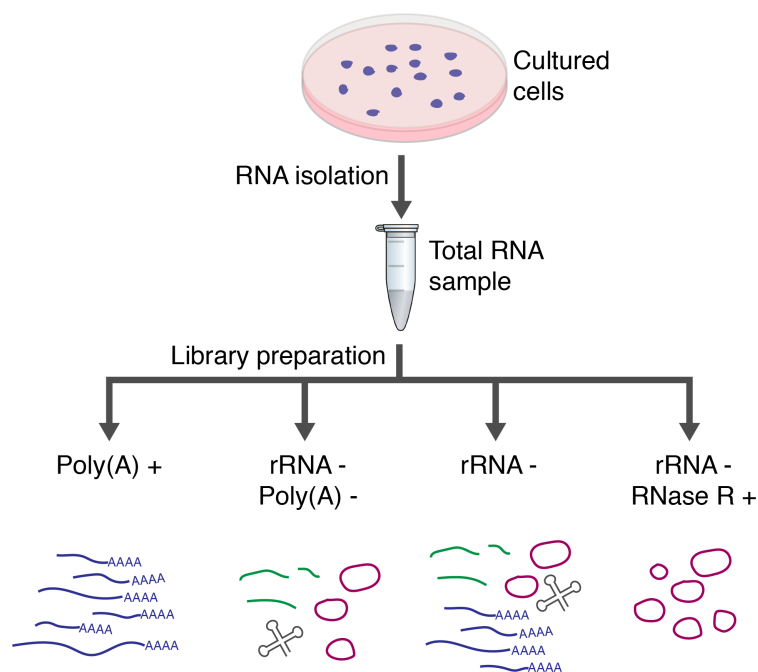


Figure 1.10: Different approaches for library preparation for RNA-Seq. Setting up poly(A)-selected RNA-Seq libraries, only mRNAs and some lncRNAs are purified. In contrast, when depleting both rRNA and polyadenylated transcripts in poly(A)-depleted RNA-Seq libraries, all non-coding RNAs lacking a poly(A) tail are purified, including circRNAs. rRNA-depleted total RNA samples retain both circular transcripts and linear RNAs, independently of the presence of a poly(A) tail. CircRNAs are the primary RNA molecule in libraries depleted of the rRNA and treated with RNase R to remove linear RNAs. Poly(A)+: poly(A)-selected RNA-Seq or mRNA-Seq; rRNA-/Poly(A)-: poly(A)-depleted RNA-Seq; rRNA-: rRNA-depleted total RNA-Seq; rRNA-/RNase R+: RNase R-treated RNA-Seq or CircleSeq. Adapted from Szabo & Salzman, 2016.

A typical protocol for RNA-Seq library preparation uses oligo-dT beads capable of capturing exclusively polyadenylated RNAs, thereby generating poly(A)-selected RNA-Seq libraries. This approach is frequently employed for mRNA transcriptome profiling experiments, therefore it is also known as mRNA-Seq. CircRNAs lack a poly(A) tail, making them undetectable in poly(A)-selected RNA-Seq libraries (Figure 1.10). Indeed, circRNAs can be detected only from RNA-Seq libraries preserving the non-polyadenylated pool of RNAs. In addition, it is recommended to remove the ribosomal RNA (rRNA) molecules, that constitute the majority of the total RNA pool. For instance, circRNAs can be detected in poly(A)-depleted RNA-Seq libraries (Salzman *et al.*, 2013) generated upon depletion of rRNA molecules and removal of polyadenylated transcripts. In this way, all RNA species lacking a poly(A) tail are preserved, not only circRNAs (Figure 1.10). With this approach, circRNAs will still represent a minor part of the sequenced RNAs.

A specific protocol of library preparation to enrich for circRNAs has been developed, known as CircleSeq or CircSeq (Jeck *et al.*, 2013). It takes advantage of the property of circRNAs to be resistant to the digestion by exonucleases. After rRNA depletion, RNA samples are treated with ribonuclease R (RNase R), a highly processive 3' → 5' exonuclease that selectively digests linear RNAs which contain at least seven unstructured nucleotides at their 3' end (Vincent & Deutscher, 2006; Szabo & Salzman, 2016) (Figure 1.10).

Finally, an alternative approach consists in removing the highly abundant rRNA prior to RNA sequencing, generating rRNA-depleted total RNA-Seq data (from now referred to as rRNA-depleted RNA-Seq). This approach has the advantage that it simultaneously provides expression information for both coding and non-rRNA non-coding RNA (Figure 1.10). Currently, the rRNA-depleted RNA-Seq approach remains the most widely used method for transcriptomics studies involving both linear and circRNAs, since it allows the analysis of different RNA species and their comparison from the same library, thereby overcoming the problem of variability due to different preparation protocols.

Very recently, a more complex protocol named RPAD (RNase R treatment followed by Polyadenylation and poly(A)+ RNA Depletion) has been described. It combines poly(A) depletion with RNase R treatment, followed by a polyadenylation step and removal of the resulting polyadenylated transcripts, with the scope of efficiently removing non-polyadenylated and highly-structured RNAs (Pandey *et al.*, 2019). Ideally, all described approaches would benefit from high sequencing depth, paired-end protocols as well as long reads for a reliable detection of back-splice reads (Szabo & Salzman, 2016; Kristensen *et al.*, 2019).

1.8 Computational identification of circRNAs from RNA-Seq

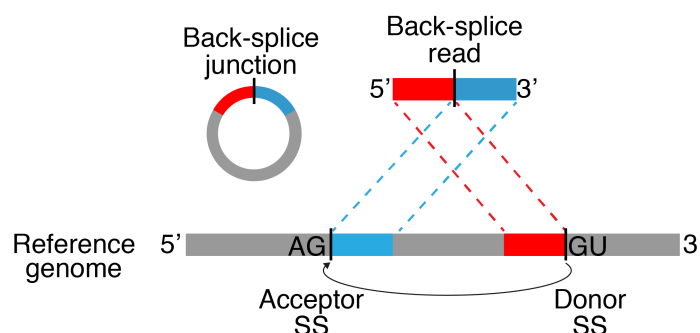


Figure 1.11: Schematic of reads spanning back-splice junctions. The ends of reads spanning the back-splice junction align discontinuously and in a reversed orientation to the reference genome.

Computationally, circRNAs can be detected from RNA-Seq data based on sequencing reads that span back-splice junctions (BSJs). These are chimeric reads, which align to two distinct portions of the genome. For circRNAs, the ends of these reads align in reversed orientation to the reference genome with respect to transcription (Figure 1.11). Depending on the type of library, some limitations have to be considered when searching for circRNAs. For instance, the largely adopted rRNA-depleted RNA-Seq protocol cannot provide information about the internal exon-intron organisation of a circRNA. In fact, reads mapping to exons or other genomic features that are shared between a circRNA and a linear RNA cannot be uniquely assigned to any of these molecules. Moreover, considering that circRNAs are in general lowly abundant, relying only on back-splice reads makes quantitative analyses still challenging. On the other hand, having in the same library both linear and circRNAs allows a direct comparison for instance of linear and back-splicing or circRNA and host gene levels, avoiding variability generated by the usage of different protocols for library preparation.

At the time of this study, several computational tools were already available for the detection of circRNAs from RNA-Seq data and many other tools have been published over the last years (reviewed in Szabo & Salzman, 2016; Zeng *et al.*, 2017; Gao & Zhao, 2018; Jakobi & Dieterich, 2019). In general, based on the strategy adopted, computational tools for circRNA detection can be divided into two large categories: fragmentation-based or pseudo reference-based (Table 1.1). The fragmentation-based strategy focuses on reads that do not map linearly to the genome and splits them into two fragments that are aligned separately to the reference genome. These are defined as chimeric alignments, and

correspond to back-splicing events when they are aligned in reversed orientation to the reference genome. This approach potentially allows *de novo* detection of circRNAs and can be completely independent from an existing genome annotation. This category comprises most of the available circRNA tools, including `find_circ` and `CIRCexplorer` (Table 1.1). `find_circ` is based on custom Python scripts that analyse the data generated by aligning sequencing reads to the reference genome with `Bowtie2`. Unmapped reads from `Bowtie2` (Langmead & Salzberg, 2012) are used to extract anchor sequences from both ends of the read, that will be aligned to genome independently. When anchors map in reverse orientation, the read fragments are extended to define the location of the break-point. A series of filters are then applied to obtain a list of potential circRNAs (Memczak *et al.*, 2013). A similar approach is used by `CIRCexplorer`, that in its original version relies on a combination of `TopHat` and `TopHat-Fusion` algorithms to extract back-splicing events (Zhang *et al.*, 2014). In later versions, it can also be used in combination to the splice-aware aligner `STAR` (Dobin *et al.*, 2012) to parse its chimeric alignments and annotate back-splicing events. Different from `find_circ`, `CIRCexplorer` requires an annotated genome to call back-splicing events, thus it is restricted to exonic circRNAs and ciRNAs from annotated splice sites, with the advantage that it will likely get rid of some of the noise in the data.

In contrast to the fragmentation-based strategy, the pseudo reference-based strategy strictly requires a reference genome annotation to generate putative events on which reads are tested for alignment. Although these tools do not allow the discovery of novel circRNAs from unannotated junctions, they likely provide a more reliable list of back-splicing events. Moreover, alignment to a pre-defined set of BSJs would increase the sensitivity of the algorithm in detecting circRNAs, as these BSJs are unlikely to be missed. The pseudo reference-based category includes `NCLscan`, `KNIFE` and `PTESfinder` (Table 1.1).

Finally, also machine learning-based approaches have been proposed to predict circRNAs, and in the last year, deep learning has been applied to predict circRNAs from specific features with `DeepCIRcode` and `circDeep` (Table 1.1).

Recent studies evaluated the performance of multiple circRNA tools on rRNA-depleted RNA-Seq samples, by relying on RNase R-treated RNA-Seq samples as a source to detect real circRNAs. Interestingly, these studies agreed on the fact that the outcome of circRNA detection tools is only partially consistent and their performance can vary considerably (Hansen *et al.*, 2016, Zeng *et al.*, 2017, Hansen, 2018). Thus, the choice of a specific algorithm is critical for the downstream analysis and for deriving conclusions about circRNA biology. Benchmarking based on RNase R-treated RNA-Seq data revealed `CIRCexplorer` as one of the outperforming tools (Hansen *et al.*, 2016), although it has the limitation that it does not allow the detection of novel circRNA species due to the strict dependence

on the genome annotation. On the other hand, `find_circ` did not perform as good as `CIRCexplorer`, but is capable to identify circRNAs originating from unannotated junctions, thus providing a broader view of the circRNA landscape. It has been suggested that combining the prediction of multiple tools for circRNA detection would lead to a more reliable catalogue of circRNAs from RNA-Seq data (Hansen *et al.*, 2016; Zeng *et al.*, 2017; Hansen, 2018). Considering only circRNAs detectable with multiple tools would remove potential false positives derived from the usage of a specific algorithm (Hansen *et al.*, 2016; Zeng *et al.*, 2017; Hansen, 2018). However, this approach might discard *bona fide* circRNAs, which are detectable thanks to specific features of a certain algorithm. In summary, multiple tools are available to detect circRNAs from RNA-Seq data, and it is advisable to use different tools for circRNA identification, depending on the type of dataset to be analysed and the experimental question to address.

Table 1.1: Overview of available circRNA detection tools

Tool	Reference	Language	Aligner	Strategy
MapSplice2	Wang <i>et al.</i> , 2010	Python	Bowtie	Fragmentation-based
find_circ	Memczak <i>et al.</i> , 2013	Python	Bowtie2	Fragmentation-based
CIRCfinder (only ciRNA)	Zhang <i>et al.</i> , 2013	Python	Bowtie (TopHat)	Fragmentation-based
Segemehl	Hoffmann <i>et al.</i> , 2014	C	segemehl	Fragmentation-based
circRNA_finder	Westholm <i>et al.</i> , 2014	Perl	STAR	Fragmentation-based
CIRCexplorer	Zhang <i>et al.</i> , 2014	Python	Bowtie (TopHat-Fusion), STAR	Fragmentation-based
CIRI	Gao <i>et al.</i> , 2015	Perl	BWA	Fragmentation-based
NCLscan	Chuang <i>et al.</i> , 2016	Python	BWA	Pseudo reference-based
Acfs	You <i>et al.</i> , 2015; You & Conrad, 2016	Perl	BWA-MEM	Fragmentation-based
KNIFE	Szabo <i>et al.</i> , 2015	Perl	Bowtie, Bowtie2	Pseudo reference-based
DCC	Cheng <i>et al.</i> , 2015	Python	STAR	Fragmentation-based
UROBORUS	Song <i>et al.</i> , 2016	Perl	Bowtie, Bowtie2, TopHat	Fragmentation-based
CIRCexplorer2	Zhang <i>et al.</i> , 2016a	Python	Bowtie (TopHat-Fusion), STAR, MapSplice, BWA, segemehl	Fragmentation-based
PTESfinder	Izuogu <i>et al.</i> , 2016	Shell, Java	Bowtie, Bowtie2	Pseudo reference-based
PredcircRNA	Pan & Xiong, 2015	-	-	Machine learning
PredcircRNATool	Liu <i>et al.</i> , 2016	-	-	Machine learning
DeepCIRcode	Wang & Wang, 2019	-	-	Deep learning
circDeep	Chaabane <i>et al.</i> , 2019	-	-	Deep learning

1.9 Hypoxia: a hallmark of cancer

The Nobel Prize in Physiology or Medicine 2019 has been awarded to three scientists, William G. Kaelin Jr., Peter J. Ratcliffe, and Gregg L. Semenza, for their research on how cells detect oxygen levels and react to hypoxia (<https://www.nobelprize.org/all-2019-nobel-prizes/>). Hypoxia is defined as a condition in which tissues are deprived of an adequate amount of oxygen, that is required for their normal metabolic functions. From a physiological point of view, hypoxia is achieved at increased altitude (Simonson *et al.*, 2010; Sarkar *et al.*, 2003), during muscle exercises, in mammalian embryogenesis (Semenza, 2012), and in case of wound healing (Hong *et al.*, 2014). In addition, hypoxia has been associated to multiple pathological conditions, including cardiovascular diseases and cancer (Abe *et al.*, 2017; Schito & Semenza, 2016). In particular, hypoxia is considered a common hallmark of solid tumours, in which the high metabolic activity and the rapid proliferation of cancer cells increase the oxygen demand (Hanahan & Weinberg, 2011). In fact, because of the tumour development and progression, blood vessels have restricted access to internal cancer cells. This limits the supply of nutrients and oxygen to cells, resulting in the formation of hypoxic regions (Semenza, 2014; Schito & Semenza, 2016). Tumour hypoxia is often translated into a more aggressive phenotype derived from increased growth and metastasis. Cells in the hypoxic regions have the capability to survive radio-, chemo- and immunotherapy, invade, form metastases, and evade the immune system, highlighting the importance of developing efficient hypoxia-targeted therapies (Graham & Unger, 2018; Schito & Semenza, 2016).

The hypoxia-inducible factor pathway Oxygen sensing and adaptation to hypoxia are primarily driven by hypoxia-inducible factors (HIFs), which constitute a family of transcription factors that activate the expression of hundreds of genes to sustain proliferation, produce energy, undertake biosynthesis and evade apoptosis. The HIF protein family includes HIF1, HIF2 and HIF3, with the first two being the best studied. The activity of HIFs is finely regulated by the available levels of oxygen. In normal oxygen (normoxic) conditions, their alpha subunits, HIF1 α , HIF2 α (also known as Endothelial PAS Domain Protein 1, EPAS1) and HIF3 α , are translated and hydroxylated by prolyl hydroxylases (PHDs, also known as EGLNs) at specific proline residues (Ivan *et al.*, 2001; Jaakkola *et al.*, 2001). PHDs use molecular oxygen O₂ as a co-substrate (Dengler *et al.*, 2013; Yang *et al.*, 2014). The hydroxylated proline residues of HIF α are recognised by the von Hippel-Lindau (VHL) protein, an E3 ubiquitin ligase, which targets them to degradation by the proteasomal machinery (Gossage *et al.*, 2015). At low oxygen levels, PHDs are no longer able to hydroxylate HIF α due to the scarcity of their co-substrate. This stabilises

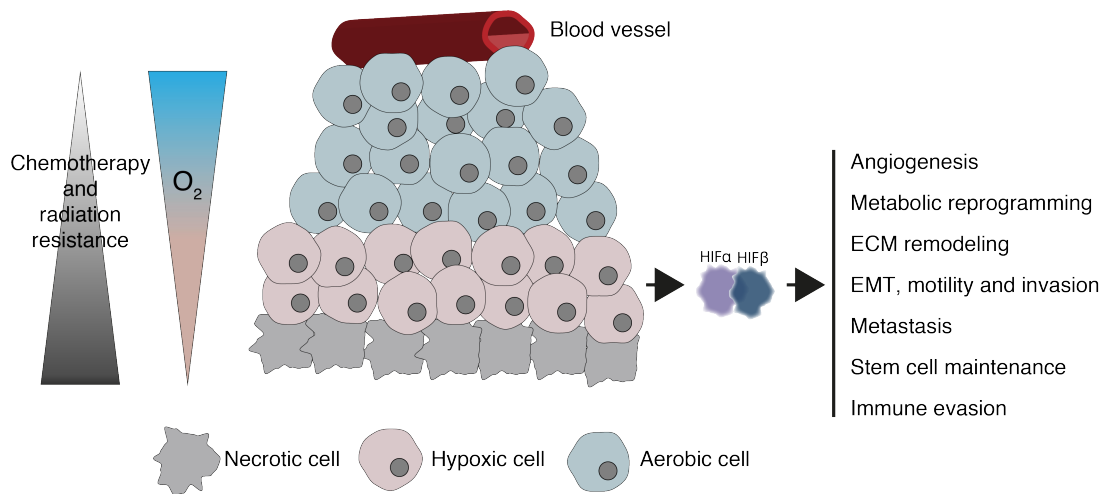


Figure 1.12: Hypoxic niches in solid tumours. In solid tumours, cells in close proximity to blood vessels receive the adequate amount of oxygen for their aerobic metabolism. Cells become increasingly hypoxic with increased distance from blood vessels. Hypoxic cells express high levels of HIFs, which activate target genes that impact on every critical aspect of cancer progression. HIF: hypoxia-inducible factor; ECM: extracellular matrix; EMT: epithelial-to-mesenchymal transition. Adapted from Schito & Semenza, 2016 and Al Tameemi *et al.*, 2019.

the alpha subunits, which can form heterodimers with the constitutively present HIF1 β (also known as Aryl Hydrocarbon Receptor Nuclear Translocator - ARNT) (Choudhry & Harris, 2018). Once dimerised, the complex is able to recognise specific sequences, known as hypoxia-responsive elements (HREs), located in the promoter of the target gene (Salceda & Caro, 1997; Kaelin & Ratcliffe, 2008; Masoud & Li, 2015; Choudhry & Harris, 2018). For the activation of the expression of HIF target genes, the stabilised HIF1 heterodimer couples to the general co-activators p300/CBP (CREB binding protein), thereby forming an active transcription factor which initiates the hypoxic response (Wei *et al.*, 2018). By this mechanism, HIFs induce the transcription of hundreds of genes involved in diverse cellular processes, including the formation of blood vessels from pre-existing vessels (angiogenesis) to improve oxygen delivery to cells. This is achieved by increasing the expression and secreting specific growth factors upon HIF1/2 α activation, such as the vascular endothelial growth factors (VEGFs), which stimulate the sprouting and proliferation of endothelial cells to form new blood vessels (Weis & Cheresh, 2011; Krock *et al.*, 2011). In particular, *VEGFA* expression and induction in hypoxic cells has been reported to be pivotal for tumour and ischemic tissue survival (Shibuya, 2011). With the formation of new blood vessels, cells tend to detach from the tumour, move into lymphatic or blood vessels and reach other tissues. Here, cancer cells leave the vessel, converting the tumour into a metastatic state. All steps that lead to the formation of metastases involve HIF

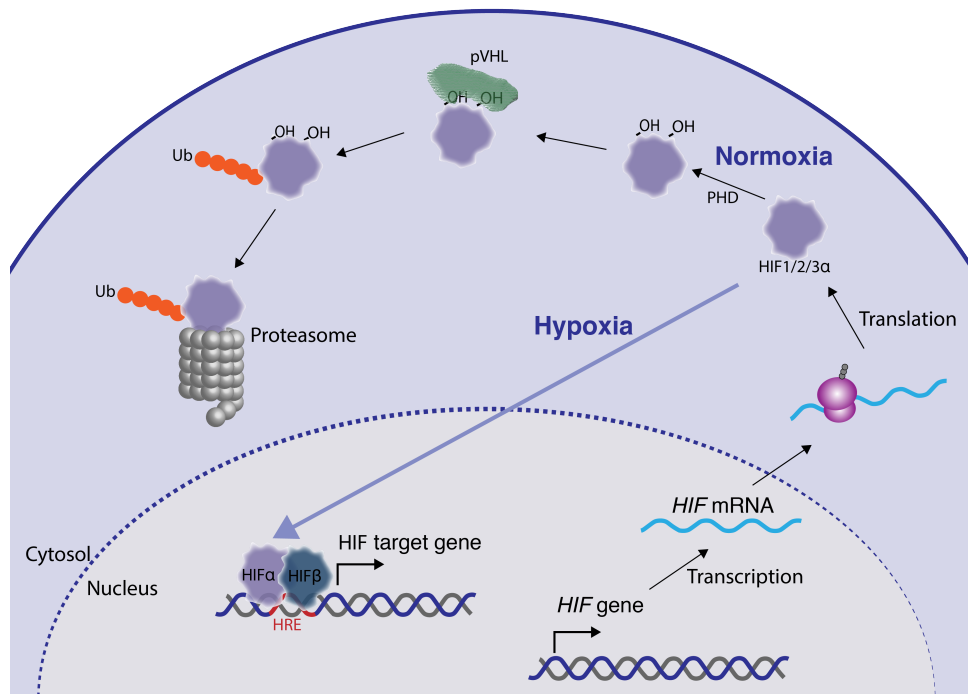


Figure 1.13: Hypoxia-inducible factor (HIF) regulation in normoxia and hypoxia. In the presence of normal oxygen conditions, HIF α is hydroxylated (OH) on proline residues by prolyl-4-hydroxylases (PHDs). The von Hippel–Lindau protein (pVHL) recognises the hydroxylated HIF and marks it with ubiquitin (Ub). This directs HIF α towards degradation by the proteasome system. Conversely, in hypoxic conditions, HIF α is stabilised and translocated into the nucleus. Here, HIF α dimerise with HIF1 β and activates the transcription of HIF target genes. Adapted from Lee *et al.*, 2019.

target genes (Schito & Semenza, 2016). In addition, HIF1 α target genes have been shown to regulate cell proliferation, migration and survival, by affecting the apoptosis pathway (Mori *et al.*, 2016; Greijer, 2004), and to be involved in the metabolic adaptation (Singh *et al.*, 2017). Additional processes that are modulated in hypoxia upon activation of HIF α are the remodelling of the extracellular matrix and the endothelial-to-mesenchymal transition (EMT) (Gilkes *et al.*, 2014; Platel *et al.*, 2019).

Despite a major role of the HIF signalling pathway in the response to hypoxia, additional transcription factors are activated at low oxygen concentration. These include the nuclear factor- κ B (NF- κ B) and many others. NF- κ B is activated in hypoxia in a HIF-independent manner, via proteasomal degradation of I κ b, that makes NF- κ B available (Lee *et al.*, 2019). In addition, the unfolded protein response (UPR) pathway is activated in hypoxia to react to the endoplasmic reticulum (ER) stress. In hypoxia, this is induced by the accumulation

of mis- and unfolded proteins. By activating the UPR pathway, cells attenuate the protein synthesis and re-establish the proper folding and processing of proteins at the ER (Chipurupalli *et al.*, 2019). In summary, the hypoxic cell adaptation is primarily achieved by the activation of HIF proteins, which activate the expression of numerous genes to promote angiogenesis, metabolic remodelling, cell proliferation and metastasis. In this way, the hypoxia pathway participates or even initiates tumour progression.

Alternative splicing and back-splicing in hypoxia Beyond the transcriptional response to hypoxia, it is becoming evident that alternative splicing plays a pivotal role in the adaptation to hypoxic stress. Recent studies explored splicing changes in hypoxic cells from different cancers, including hepatocellular carcinoma, breast, head and neck, and prostate cancer (Sena *et al.*, 2014; Han *et al.*, 2017c; Brady *et al.*, 2017; Bowler *et al.*, 2018). They reported widespread alteration of the splicing pattern in response to hypoxia, with hundreds to thousands of differential AS events (Han *et al.*, 2017c; Brady *et al.*, 2017; Bowler *et al.*, 2018). Interestingly, many of these AS events affected genes of which the overall transcription did not change in hypoxia, highlighting that AS adds an additional layer of gene regulation in the response to hypoxia (Sena *et al.*, 2014). Despite this large number of AS events changed in hypoxia, so far only few specific cases have been further investigated to understand their underlying mechanism of regulation. For instance, the HIF target *VEGFA* is regulated not only at transcriptional level, but it also undergoes AS, which leads to the production of multiple transcript isoforms. Interestingly, the encoded proteins can play contrasting roles in hypoxia, acting either as pro- or anti-angiogenesis factors (Kikuchi *et al.*, 2014; Guyot & Pagès, 2015). In addition, Brady and colleagues reported an increased intron retention event that affected the gene *EIF2B5* in hypoxic head and neck cancer cells (Brady *et al.*, 2017). *EIF2B5* is a key regulator of mRNA translation. The retention of this intron caused the production of a truncated protein that could not exert its function, leading to global inhibition of translation and increased cell survival (Brady *et al.*, 2017). A recent study reported that hypoxia primarily affects intron retention events in human breast cancer cells (Han *et al.*, 2017c). Despite the evidences that hypoxia alters alternative splicing, still little is known about the underlying mechanism of regulation by splicing factors. Only recently, it has been reported that the expression of multiple splice factors and splice factor kinases increases in hypoxic cancer, including the Cdc-like splice factor kinases CLK1 and CLK3 (Bowler *et al.*, 2018). As mentioned before, in addition to perturbation of the AS pattern, also circRNAs were found to be regulated upon hypoxic stress, including circZNF292 and circDENND4C (Boeckel *et al.*, 2015; Liang *et al.*, 2017b).

Further studies are required to expand the current knowledge of the impact of transcriptional and post-transcriptional processes in the hypoxia adaptation. Given the influence of hypoxia on tumour progression and resistance to current therapies, this would facilitate the discovery of potential biomarkers and therapeutic approaches for cancer.

1.10 Aim of this thesis

Different studies have demonstrated the importance of transcriptional changes in the adaptation of solid tumours to hypoxia. On the other hand, the impact of post-transcriptional processes in response to hypoxic stress has only recently gained attention and further research is required to understand how oxygen availability influences RNA regulation in cancer.

This project intended to extensively characterise the reaction of solid tumours to hypoxia at RNA level and its regulation, exploiting bioinformatics approaches. For this purpose, the first aim of this thesis was using high-throughput RNA-Sequencing to investigate alterations in gene expression and splicing in cancer cells under hypoxic stress. Secondly, I aimed to identify splicing factors of which the expression is altered in hypoxic cancer, which might play a role in cancer progression. Lastly, since the influence of hypoxia on the circRNA repertoire in cancer cells remains still poorly explored, the third objective of the thesis was the profiling of circular RNAs in cancer cells and the investigation of their regulation under hypoxic conditions from RNA-Seq data.

Chapter 2

Methods

This chapter includes methods used for the bioinformatics analyses as well as a brief description of programs and databases used for this study. A description of the experimental methods adopted by Camila de Oliveira Freitas Machado and Sandra Fischer is provided in Supplementary Material, as reported in Di Liddo *et al.*, 2019, and Fischer *et al.*, *in revision*.

Here, Section 2.1 reports methods for the transcriptome-wide analysis of RNA-Seq data. Section 2.2 describes methods adopted to establish and evaluate the performance of the pipeline for circRNA identification. Section 2.3 provides a description of methods adopted for the identification and characterisation of circRNAs in cancer cells, in normoxic or hypoxic conditions. Finally, Sections 2.4 and 2.5 describe computational tools and databases used in the study, respectively.

2.1 Transcriptome analysis from RNA-Sequencing data

In this study, four different RNA-Seq datasets were generated and analysed: three rRNA-depleted RNA-Seq datasets from hypoxic treatment of HeLa, A549 and MCF-7 cells and one poly(A)-selected RNA-Seq dataset from *MBNL2* knockdown experiments in hypoxic MCF-7 cells. Here, a detailed description of methods for the transcriptome-wide analysis of these data is provided.

2.1.1 Processing and mapping of sequencing reads

First, the overall quality of sequencing reads was checked using FastQC (<https://www.bioinformatics.babraham.ac.uk/projects/fastqc/>) in combination to MultiQC (Ewels *et al.*, 2016) to aggregate reports. To improve the read mapping rates, Flexbar (Dodt *et al.*, 2012) was used to filter RNA-Seq reads, demanding a minimum quality of 20 (Phred score) and applying a threshold of at least 20 nt to the read length. When necessary, Flexbar was additionally used to remove adapter contaminations. An example Flexbar call is reported:

```
flexbar -r <filename.fastq.gz>
        --zip-output GZ
        --adapters <adapter.fasta>
        --format i1.8
        --pre-trim-phred 20
        --min-read-length 20
```

To evaluate the rRNA-depletion efficiency in total RNA-Seq data, reads were mapped to human rRNA sequences using Bowtie2:

```
bowtie2 -x <human_rRNA_index_prefix>
        --phred33
        --sensitive
        --seedlen=22
        -U file.fastq.gz
        -S /dev/null 2> rRNA_bowtie.log
```

Single-end reads were aligned to the human genome (version hg38/GRCh38) using the splice-aware alignment software STAR (Dobin *et al.*, 2012), after generating genome index

files, with the parameter `--sjdbOverhang` set to the maximum read length reduced by 1. GENCODE release 24 was used as reference annotation. Maximum two mismatches were allowed and only uniquely mapped reads were collected for downstream analyses. In addition, for the subsequent detection of circRNAs and comparison to linear splicing reads, parameters to detect chimeric alignments were also defined. For *MBNL2* knockdown data, parameters for chimeric alignments in STAR were omitted due to the nature of the data (poly(A)-selected) that does not allow for the detection of circular transcripts.

A detailed list of parameters set for flexbar and STAR is reported here:

```
STAR --runMode alignReads
      --readFilesIn <path to fastq file>
      --readFilesCommand zcat
      --genomeDir <path to genome index>
      --outFilterMultimapNmax 1
      --outFilterMismatchNmax 2
      --outSAMtype BAM SortedByCoordinate
      --alignSJDBoverhangMin 15
      --alignSJoverhangMin 15
      --chimSegmentMin 15
      --chimScoreMin 15
      --chimScoreSeparation 10
      --chimJunctionOverhangMin 15
```

2.1.2 Gene-level quantification and differential expression

The quantification of gene expression at RNA level was performed with `htseq-count` script of the HTSeq library (Anders *et al.*, 2015), counting reads overlapping to the exonic component of genes, using default parameters. BAM files from STAR were used as input, together to GENCODE genome annotation in GTF format.

To estimate the relative expression of a gene, a custom R script was used to normalise read counts per gene to "transcripts per million" (TPM), dividing by sequencing depth and gene length, since longer genes generate a higher number of reads. One advantage of TPM is that the sum of all TPMs in each sample is the same (10^6), allowing an easier comparison of the proportion of reads mapped to a specific gene in each sample.

Differential expression analysis was performed in R with DESeq2 (Love *et al.*, 2014), using raw counts as input. Significantly regulated genes were defined by setting an adjusted *P*-value threshold of 0.05. For *MBNL2* knockdown experiments and comparison to hypoxia

data, differentially expressed genes were additionally filtered for a fold change ≥ 1.5 and at least 1 TPM in any sample.

Gene Ontology (GO) enrichment analysis was performed using the over-representation test implemented in the `enrichGO` function of the `clusterProfiler` package (Yu *et al.*, 2012) in the R statistical software environment. Enrichment was tested against the union of all genes that were tested in the DESeq2 analysis of any cell line. *P*-value and *q*-value cutoffs were set to 0.05. "Biological Process" and "Molecular Function" categories were explored, as well as Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways with the `enrichKEGG` function of the same package.

2.1.3 Analysis of alternative splicing changes

For the alternative splicing profile, reads were further filtered with `Flexbar` to have all a fixed length (75 bp for HeLa and *MBNL2* knockdown, 71 for A549 and MCF-7 data), as required by `rMATS`. The resulting reads were mapped to the human genome with `STAR`, setting the parameters `--outFilterMultimapNmax 1 --outFilterMismatchNmax 2` as above, and `--alignEndsType EndToEnd` to switch off soft-clipping at both ends of the read. Then, BAM files containing aligned reads were used as input for `rMATS`, considering hypoxic replicates as sample 1 and normoxic replicates as sample 2 to estimate the ΔPSI and $\text{PSI}_{\text{hypoxia}} - \text{PSI}_{\text{normoxia}}$, as with the following code:

```
python RNASEq-MATS.py \  
-b1 <sample1_rep1.bam>,<sample1_rep2.bam> \  
-b2 <sample2_rep1.bam>,<sample2_rep2.bam> \  
-gtf gencode.v24.primary_assembly.annotation.gtf \  
-bi /path/to/genome/dir -o /path/to/output/dir \  
-t single -len <input read length> \  
-libType fr-firststrand \  
-c 0.0001 \  
-novelSS <1 or 0>
```

For hypoxia datasets, the parameter `-novelSS` was set to 1 to enable the detection of unannotated splice sites for eventual comparison to back-splicing events generated from unannotated junctions. Alternative splicing (AS) events were identified based exclusively on reads overlapping the splice junctions. Significantly changed AS events were defined based on a false discovery rate (FDR) cutoff of 0.05. In addition, we required: an absolute $\Delta\text{PSI} \geq 10\%$; a minimum inclusion level of 10% in either sample 1 or sample 2; at least an average junction count (over replicates) of 10/5 reads in inclusion/skipping junctions.

For *MBNL2* knockdown experiments and comparison to hypoxia data, also a threshold of 1 TPM in any sample was applied.

2.1.4 Prediction of MBNL2 binding sites

Putative MBNL2 binding sites were predicted by scanning 3'UTR sequences for clustered 5'-YGCY-3' motifs (5'-YGCYGCY-3' and 5'-YGCYN₀₋₃YGCY-3'). The search was restricted to genes expressed in *MBNL2* knockdown experiments (TPM > 1 in any sample) and tested for differential expression with DESeq2. First, the 3'UTR annotation was retrieved from GENCODE release 24 basic annotation, considering protein-coding transcripts with support level < 4 from genes with support level < 3, to exclude automatic annotation. 3'UTRs shorter than 10 nt were excluded. When multiple 3'UTRs were annotated for a single gene, we considered only the longest one. This led to 11575 expressed genes with annotated 3'UTRs to investigate. Those genes were further grouped into regulated when the fold change was ≥ 1.5 and the adjusted *P*-value < 0.05, and unchanged when the fold change was ≥ 1.3 and the adjusted *P*-value > 0.5.

2.2 Establishment and evaluation of the pipeline for circRNA detection

In order to investigate strengths and weaknesses of `CIRCexplorer` and `find_circ` for circRNA detection, RNA-Seq data of replicate 1 from normoxia samples of HeLa and MCF-7 cells were analysed separately.

CIRCexplorer First, the reference human annotation was filtered for gene level 1 or 2, to remove automatically annotated genes. Then, the script `gtfToGenePred` (http://hgdownload.cse.ucsc.edu/admin/exe/macOSX.x86_64/gtfToGenePred) was used to obtain a reference annotation in the format required by `CIRCexplorer` (Gene Predictions and RefSeq Genes with Gene Names). Cleaned reads were processed and mapped with `STAR` as described above. `Chimeric.out.junction` files were collected and used as a input for `CIRCexplorer`. The tool consists of two main python scripts: first, `star_parse.py` parses chimeric junction tables searching for mates mapping on the same chromosome and strand in reverse orientation, generating fusion junction tables; next, `CIRCexplorer.py` uses fusion junction tables in combination to the reference genome sequences (file in FASTA format) and the annotation file previously produced, to define and annotate back-splicing events. Here, an example code is reported:

```
# parse STAR chimeric junction table
star\_parse.py Chimeric.out.junction Fusion.junction

# detect and annotate circRNAs
CIRCexplorer.py -j Fusion.junction
                 -g GRCh38.genome.fa
                 -r gencode.ref.txt
                 -o circ.txt
```

To get the final list of circRNAs from `CIRCexplorer`, back-splicing events were required to be supported by ≥ 2 total reads.

Find_circ In order to detect circRNAs with `find_circ`, the same cleaned reads were mapped to the human reference genome with `Bowtie2`, as described in Memczak *et al.*, 2013. Next, unmapped reads from output BAM files were collected and 20 nt long anchors were extracted from both ends of the reads. Anchors were mapped against the human reference genome and `find_circ.py` script was used to detect circRNAs.

```
# map reads to the human genome hg38 and sort alignments
bowtie2 --very-sensitive \
        --phred33 \
        --score-min=C,-15,0 \
        -q \
        -x <path/to/hum_genome/index> \
        -U <file.fastq.gz> \
2> bowtie.log | samtools view -hbuS - | samtools sort - mapped

# collect unmapped reads
BAM=mapped.bam
samtools view -hf 4 $BAM | \
samtools view -Sb - > unmapped.bam

# split reads to obtain 20 nucleotides anchors
# from both ends of the read
~/find_circ_v1.2/unmapped2anchors.py unmapped.bam | \
gzip > anchors.qfa.gz

# map anchors to the human genome
# and detect circRNAs with find_circ.py
ANCHORS=anchors.qfa.gz
bowtie2 --score-min=C,-15,0 \
        --reorder \
        -q \
        -x <path/to/hum_genome/index> \
        -U $ANCHORS 2> circ_bowtie.log | \
find_circ.py --genome=GRCh38.genome.fa \
--name=<sample name> \
--stats=samplename_stats.log \
# default parameters
--anchor=20 \
--min_uniq_qual=2 \
--margin=2 \
--maxdist=2 \
--bam=anchor_alignments.bam > splice_sites.bed
```

The `splice_site.bed` file includes both back-splicing and linear splicing events and was

further filtered for circRNAs as in Memczak *et al.*, 2013, demanding unambiguous breakpoints, unique alignments, a maximum distance of 100 kb, and a GU/AG splice site signal, which is present in more than 98% of all intron sequences that are removed by the spliceosome (Burset *et al.*, 2000). Finally, to define a circRNA as detected, a filter based on unique reads was applied (≥ 2).

Our pipeline Based on the usage of `CIRCexplorer` and `find_circ` in combination, we established a pipeline for circRNA detection as shown in the schematic in Figure 3.11. Cleaned reads from samples of a single dataset were merged and mapped to the genome with `STAR` and `Bowtie2`. The `Chimeric.out.junction` table was used as a input to detect circRNAs with `CIRCexplorer` and unmapped reads were extracted from `Bowtie2`'s BAM files and used to detect circRNAs with `find_circ`. The output from `CIRCexplorer` and `find_circ` was unified and sample-by-sample quantification of circRNA expression was performed based on chimeric alignments from `STAR`, discriminating total back-splice reads from unique back-splice reads with a custom script in R. Differently from `find_circ`, we discriminated unique reads based on the mapping position rather than the read sequence, thus increasing the stringency in detecting PCR artefacts. CircRNAs were then required to have GU/AG or GC/AG as splice-site signal (the first and the second top used splice-site signals, respectively; (Burset *et al.*, 2000), a maximum distance between back-splice sites ≤ 100 kb or to span a single annotated gene. To define a circRNA as present in a given dataset, we demanded a minimum of two unique reads supporting the back-splice junction in at least one sample of the dataset.

In order to evaluate the performance of our pipeline against the individual circRNA detection tools, RNA-Seq data from Hs68 cells (Jeck *et al.*, 2013) and RNA-Seq data from HeLa cells (Gao *et al.*, 2015; Gao *et al.*, 2016) were downloaded from SRA with the accession number reported in Table 3.6, and used to detect circRNAs with our pipeline, or `find_circ` and `CIRCexplorer` separately. `Find_circ`, was tested with two different filtering approaches for expression, one based on total back-splice reads, the other based on unique back-splice reads (≥ 2 in any sample of the dataset). `CIRCexplorer` was tested exclusively on circRNAs supported by ≥ 2 total back-splice reads in any sample of the dataset, since it does not output counts of unique reads.

For each tool/pipeline and dataset, back-splicing events were first detected in RNase R-untreated samples (RNaseR-), then compared to matched RNase R-treated (RNaseR+) samples for validation, with the assumption that the treatment with RNase R leads the detection of genuine circRNAs. Total back-splice reads were normalised by sequencing depth and a ratio between RNaseR- and RNaseR+ samples was computed to estimate the

eventual enrichment in RNase R+ samples in terms of fold change. When replicates were available, the mean of normalised counts between replicates was used to calculate the fold change. Based on the fold change value, circRNAs were grouped into "RNase R-resistant" and "RNase R-sensitive", when an increase or decrease of the total read counts in RNase R+ samples was observed, respectively. RNase R-sensitive circRNAs were further classified into "RNase R-depleted" when the circRNA was either undetectable in the RNase R+ samples or at least 5-fold decreased, and "RNase R-reduced" when a decrease up to 5 fold was observed. RNase R-resistant circRNAs were further divided into "RNase R-enriched" for those circRNAs with at least a 5-fold increase and "RNase R-unaffected" when their level remained stable or increased up to 5-fold in the RNase R+ samples (Zeng *et al.*, 2017; Hansen, 2018).

2.3 Identification and annotation of circRNAs

CircRNAs were identified and quantified from hypoxia and normoxia RNA-Seq data with the pipeline described above, demanding minimum two unique back-splice reads in any sample for each dataset. CircRNAs were annotated through a custom script in R, by comparing the genomic coordinates of circRNAs to those of genomic features described in GENCODE reference annotation (release 24, basic annotation). Automatically annotated genes (gene level = 3) were discarded for circRNA annotation. More specifically, to annotate circRNAs and define their internal exon/intron boundaries, we relied as much as possible on canonical splice variants annotated in GENCODE, as downloaded from the UCSC Genome Browser (knownCanonical). When back-splice sites did not match with any annotated junction of these canonical transcripts, we searched for the remainder transcripts with coincident splice sites, thus defining a best parental transcript to rank exons. To define the internal structure of circRNAs, we conservatively assumed that all exons annotated within back-splice sites were spliced in. When a circRNA did not overlap any annotated exons, we defined it as intronic/intergenic, depending on the genomic locus. The genome annotation might include multiple genes annotated at the same genomic locus, making it difficult to define a unique host gene. In such cases, circRNAs were labeled as "ambiguous" annotation. Finally, back-splicing events spanning multiple non-overlapping genes ($n = 70$) were excluded. For downstream analyses, the circRNA catalogue ("full set") was further filtered, demanding at least five total reads in any two samples, thereby defining a "high-confidence set" of circRNAs.

Overlap of identified circRNAs with public databases CircRNAs reported in circBase (Glažar *et al.*, 2014) were downloaded from <http://www.circbase.org/cgi-bin/downloads.cgi> as updated to 29/07/2017, containing a total of 140790 circRNAs. CircRNAs collected in circRNADb (Chen *et al.*, 2016) were downloaded from <http://reprod.njmu.edu.cn/circrnadb/resources.php>, including 32914 circRNAs. Both databases reported genomic coordinates from the human genome version hg19. For comparison to our catalogue of circRNAs, hg19 coordinates were converted to version hg38 by using the `liftOver` utility from UCSC implemented in the R package `rtracklayer` and the `hg19ToHg38.over.chain` file downloaded from UCSC.

Molecular characterisation of circRNAs and flanking regions The software `MaxEntScan` (Yeo & Burge, 2004) was used to predict the strength of back-splice and linear splice sites and compare these to 2000 randomly picked splice sites annotated in GENCODE for protein-coding transcripts with transcript support level ranging from 1 to

3.

The genomic coordinates from the human genome version hg19 of *Alu* repeats based on RepeatMasker annotation (www.repeatmasker.org) were downloaded from UCSC Genome Browser. They were compared to regions up- and downstream to circRNAs in a 500 bp window. For each hypoxia-regulated circRNA ($n = 64$) a pairwise local alignment of sequences in a 500-bp window up- and downstream of the back-splice sites was performed, using the R package `Biostrings` and setting parameters `gapOpening=10` and `gapExtension=4`. A χ^2 test was performed to compare the presence of *Alu* repeats between regulated and all circRNAs.

2.3.1 Quantification of circRNA and host gene expression

For a relative estimate of circRNA expression, total row counts per circRNA were normalised to "reads per million" by dividing by the total number of counted linear reads in the sample divided by one million. Similarly, for host genes TPM values were computed to estimate the expression value of linear RNAs, allowing the normalisation both by length of the gene and library size. Alternatively, TPM were also computed by excluding from host genes those exons that were internal to circRNAs in our data, thus avoiding a bias due to the circRNA associated to the specific host gene.

The back-splicing rate was estimated by considering the number of total back-splice reads supporting a certain circRNA and the number of reads supporting the linear junctions from the same splice sites involved in the circularisation (linear junction reads). Back-splice reads were divided by the average counts between the two linear junctions to calculate a "circular-to-linear ratio" (CLR) value. Alternatively, a "percent circularised" metric was computed by dividing the number of back-splice reads by the sum of back-splice reads and linear junction reads and multiplying by 100 (Figure 3.19).

For differential expression testing of circRNAs between normoxic and hypoxic conditions, raw back-splice read counts were combined to raw read counts per gene and used as input for DESeq2 to improve library size estimation, normalisation and statistical power along the DESeq2 algorithm. Significantly regulated circRNAs were defined based on an adjusted *P*-value threshold of 0.1.

2.3.2 Prediction of RBP binding sites

To search for putative binding sites of RBPs in regions flanking hypoxia-regulated circRNAs ($n = 64$) and unchanged circRNAs from the high-confidence set ($n = 2141$), the tool FIMO

(Grant *et al.*, 2011) was used with default parameters (`--thresh 0.0001`) to scan sequences 1,000 bp up- and downstream of back-splice sites and search for known RBP motifs from *in vitro* binding assays (Ray *et al.*, 2013) available in MEME database. FIMO prediction was further filtered for binding sites with q -value < 0.05 .

To confirm the binding of HNRNPC to regions up- and downstream to back-splice sites, publicly available HNRNPC iCLIP data in HeLa cells were investigated (Zarnack *et al.*, 2013). iCLIP tracks (BAM files) were merged and PureCLIP (Krakau *et al.*, 2017) was used for peak calling with the parameter `-ld` for higher precision in estimating emission probabilities. As by default for PureCLIP, crosslink sites were merged if located at a distance < 9 nt. The positioning of HNRNPC binding to regions flanking circRNAs was explored based on the average read coverage at each position. We considered a window of 1,000 nt up-/downstream of the back-splice sites in addition to the first and last 50 nt of the circRNA. CircRNAs of the high-confidence set expressed in HeLa ($n = 1133$) were investigated, dividing them into hypoxia-regulated and non-regulated circRNAs. For comparison, HNRNPC binding was also investigated in linear exons from expressed protein-coding genes in HeLa that do not undergo back-splicing. Linear exons were randomly picked from transcripts composed by more than two exons, excluding first and last exons to allow the investigation of introns on both sides of the exon.

2.3.3 Detection of putative miRNA binding sites

Potential interactions between circRNAs and miRNAs were predicted using `miRanda` with the parameter `-strict` to demand a strict alignment in the seed region, and requiring a match score ≥ 150 . The sequences and annotation of a high-confidence set of 542 miRNAs in humans (`high_conf_mature.fa.gz`) were downloaded from the miRBase database (<http://www.mirbase.org/>). The prediction was performed on a subset of 9754 circRNAs (only high-confidence circRNAs are shown in the figures), for which we could assign a parental transcript, thus a defined internal sequence, as described above. We manually added to this set of circRNAs the previously published intronic isoform of circZNF292 (hsa_circ_0004383), excluding the cryptic intron located between exons 1A and 2 (see Boeckel *et al.*, 2015 for the sequence). As a control, the sequence of the circRNA CDR1as/ciRS-7 was included (Hansen *et al.*, 2013; Memczak *et al.*, 2013). The frequency of miRNA binding sites within the circRNA sequences was compared to 10000 CDS and 3'UTR sequences, which were randomly selected from the GENCODE annotation. Finally, the number of detected target sites in a circRNA, CDS or 3'UTR was normalised by the size of the region to get an estimate of the density of miRNA binding sites on circRNA sequences.

2.4 Programs

Table 2.1 provides a list of software and algorithms used in this study, followed by a brief description of the tools.

Table 2.1: List of software and algorithms used in this study

Program	Version	Reference
FastQC	0.11.4	http://www.bioinformatics.babraham.ac.uk/projects/fastqc/
MultiQC	1.2	Ewels <i>et al.</i> , 2016
Flexbar	2.5	Dodt <i>et al.</i> , 2012
Bowtie2	2.2.6	Langmead & Salzberg, 2012
STAR	2.4.5a	Dobin <i>et al.</i> , 2012
IGV	2.4.6	Robinson <i>et al.</i> , 2011
HTSeq	0.6.1	Anders <i>et al.</i> , 2015
rMATS	3.2.5	Shen <i>et al.</i> , 2014
find_circ	1.2	Memczak <i>et al.</i> , 2013
CIRCexplorer	1.1.7 (2016-1-28)	Zhang <i>et al.</i> , 2014
FIMO	5.0.2	Grant <i>et al.</i> , 2011
MiRanda	3.3a	Enright <i>et al.</i> , 2003
Biostrings	2.46.0	Pagès <i>et al.</i> , 2017
DESeq2	1.18.1	Love <i>et al.</i> , 2014
ClusterProfiler	3.6.0	Yu <i>et al.</i> , 2012
SAMtools	1.2	Li <i>et al.</i> , 2009 http://samtools.sourceforge.net
MaxEntScan	-	(Yeo & Burge, 2004) http://genes.mit.edu/burgelab/maxent/Xmaxentscan_scoreseq.html
PureCLIP	1.1.2	(Krakau <i>et al.</i> , 2017)

FastQC FastQC (<http://www.bioinformatics.babraham.ac.uk/projects/fastqc/>) is a program for the quality check of sequencing reads, submitted either in fastq, BAM or SAM format. It provides quality scores across all reads as well as estimates of the GC content, duplication rate and presence of adapter contaminations, all collected in HTML reports. FastQC was used for quality check of sequenced reads, before and after trimming.

MultiQC MultiQC (Ewels *et al.*, 2016) was used to aggregate results from FastQC analyses across multiple samples.

Flexbar Flexbar is a computational tool for the pre-processing of high-throughput sequencing reads, providing functions such as demultiplexing, adapter and barcode removal, trimming based on quality scores or read length (Dodt *et al.*, 2012). Data need to be provided in FASTA or fastq format. Flexbar was used to filter and trim sequencing reads in order to increase the read mapping rates.

Bowtie2 Bowtie2 is an alignment program designed to map high-throughput sequencing reads to a reference genome by exploiting a full-text minute (FM) index to increase the speed and memory efficiency of the algorithm (Langmead & Salzberg, 2012). It supports the alignment of gapped sequencing reads as well as paired-end alignment modes. In this study, Bowtie2 was used for the alignment of the RNA-Seq libraries to the human reference genome, in order to subsequently extract unmapped reads for circRNA detection with `find_circ`. In addition, Bowtie2 was adopted to align the sequencing reads to the human rRNA sequences, to thereby estimate the efficiency of the rRNA-depletion step in the RNA-Seq library preparation.

STAR The Spliced Transcripts Alignment to a Reference (STAR) was originally designed to deal with spliced reads, typical of RNA-Seq data, originating from splicing junctions (Dobin *et al.*, 2012). It also allows the detection of chimeric alignments, in which two distinct fragments of the same read align in a non-linear order to the genome, including head-to-tail arrangements. Here, STAR was used for mapping RNA-Seq reads to the human genome and selected STAR output files were used for downstream analyses, i.e. `Chimeric.out.junction`, `Chimeric.out.sam`, `SJ.out.tab`, and `Aligned.sortedByCoord.out.bam`.

SAMtools SAMtools (Li *et al.*, 2009) provides a collection of tools for the manipulation of alignments files in the Sequence Alignment/Map format (SAM). SAMtools was used in this study for SAM/BAM file sorting and indexing, as well as to extract unmapped reads from the Bowtie2 alignment output.

IGV In this study, the Integrative Genomics Viewer (IGV, Robinson *et al.*, 2011) was used to visualise the RNA-Seq alignments to the human genome as well as binding sites from iCLIP data.

HTSeq HTSeq (Anders *et al.*, 2015) is a Python library including multiple tools for the processing of high-throughput sequencing data. The `htseq-count` script of the package was used to count reads that overlap exons of annotated genes, thus to estimate the expression level of each gene.

rMATS The replicate-MATS (rMATS, <http://rnaseq-mats.sourceforge.net/>, Shen *et al.*, 2014) program was used in this study to identify alternative splicing events from hypoxia and knockdown RNA-Seq data, relying on reads aligned to splice junctions by STAR and applying the multivariate analysis of transcript splicing method (MATS).

find_circ The algorithm `find_circ` (Memczak *et al.*, 2013) was one of the first computational tools available to identify circRNAs from RNA-Seq data. It is written in Python and uses a split-alignment-based approaches in which unmapped reads, obtained from a first alignment with Bowtie2, are split into two segments. These are mapped to the reference genome in order to detect back-splicing events when they align in a reverse orientation. Here, `find_circ` was used to identify circRNAs after investigating its main features, implementing its usage in our combined pipeline for circRNA detection.

CIRCexplorer CIRCexplorer is a program written in Python for the identification and annotation of exonic circRNAs (ecircRNAs) as well as intronic circRNAs (ciRNA) (Zhang *et al.*, 2014). It was initially designed to parse the information obtained from the combination of TopHat and TopHat-Fusion algorithms about transcripts representing fusion products, and extract from these back-splicing events. From version 1.1.0 it additionally supports the splice-aware aligner STAR, retrieving back-splice junctions from STAR chimeric alignments. CIRCexplorer strictly relies on gene annotation to define back-splicing events. CIRCexplorer was used in this study to first investigate its features; next it was integrated in our pipeline for circRNA detection in combination with STAR.

FIMO Find Individual Motif Occurrences (FIMO) is part of the MEME Suite (<http://meme-suite.org/>) and was implemented for the identification of provided motifs within query biological sequences (Grant *et al.*, 2011). FIMO scans a set of sequences for individual matches to each of the motifs that are provided. In this study, the command line version of the program was used to predict binding sites of RBPs in the regions surrounding back-splice junctions.

MiRanda MiRanda (<http://www.microrna.org/microrna/home.do>, Enright *et al.*, 2003) is an algorithm developed to predict miRNA targets in genomic sequences provided in FASTA format (reference sequences). It performs a dynamic programming local alignment between miRNA and reference sequences, searching for complementarity and assigning a score. Next, it estimates the stability of the alignments in terms of free energy. Here, MiRanda was used to scan the putative internal circRNA sequences in order to predict miRNA binding sites.

R scripts and packages Within this study, custom scripts were written in R version 3.4.3 (R Core Team, 2017) to process and analyse the RNA-Seq data, including normalisation of expression values, quantification of circRNA expression from STAR chimeric alignments, genomic annotation of circRNAs, statistical testing, generation of input files for external tools, as well as parsing of output files from other tools. Custom scripts were used to summarise data and make figures with `ggplot2` v3.2.0, `ggpubr` v0.1.7 and `ComplexHeatmap` v1.17.1.

`Biostrings` is a Bioconductor package for the manipulation of biological sequences, either from proteins or DNA and RNA, used in this study to retrieve sequences from genomic coordinates and for the alignment of circRNA flanks.

`DESeq2` (Love *et al.*, 2014) is a Bioconductor package developed for differential expression analysis starting from count-based expression data. DESeq2 analysis runs through three steps: first raw counts are normalised by a size factor computed with a "median ratio method"; next the dispersion of data is estimated; finally, data are fitted to a negative binomial generalised linear model (GLM) and Wald statistics are computed. DESeq2 was used in this study to reveal expression changes upon hypoxia at gene level and for circRNAs. In addition, it was used for differential gene expression analysis after *MBNL2* depletion in hypoxic MCF-7 cells.

ClusterProfiler (Yu *et al.*, 2012) is a Bioconductor package for the functional enrichment analysis and visualisation (Gene Ontology, GO, and KEGG). This package was used in this study for the functional characterisation of genes. In particular, the functions `enrichGO` and `compareCluster` were used for overrepresentation testing, to contrast enriched GO terms between different cell lines or direction of regulation.

MaxEntScan The Perl scripts of **MaxEntScan** (Yeo & Burge, 2004) were designed to estimate the strength of a splice site from a 9 bp long sequence (3 bases in the exon and 6 in the downstream intron) for the 5' splice site, and a 23 bp long sequence (20 bases in the intron upstream to the exon and 3 in the exon) for the 3' splice site. The **MaxEntScan** algorithm is based on the Maximum Entropy Principle and models the input splice site sequences accounting for non-adjacent and adjacent dependencies between positions. The lower the score obtained with **MaxEntScan**, the weaker the splice site. **MaxEntScan** was used here to estimate the strength of 3' and 5' back-splice sites, as well as the matching linear splice sites.

PureCLIP **PureCLIP** is a program designed to perform peak calling from single-nucleotide CLIP-seq data (i.e. iCLIP or eCLIP data), thus to identify protein-RNA interactions (Krakau *et al.*, 2017). **PureCLIP** is based on a Hidden Markov Model (HMM) approach, that discriminates four different states for each position: *enriched* or *non-enriched*, depending on whether the specific position is enriched or not in pulled-down RNA fragments; *crosslinked* or *non-crosslinked*, if the position constitutes a crosslink (CL) site or not, based on the truncation position of the read. CL sites are defined when both *enriched* and *crosslink* states co-occur. When an input control is available, **PureCLIP** can additionally incorporate such data to take into account RNA transcript abundances. Finally, **PureCLIP** can incorporate in the model also information about specific motifs known to be preferentially UV-crosslinked (CL-motifs), to avoid a crosslinking sequence bias.

Adobe Illustrator Adobe Illustrator CC 2018 was used to combine and improve the layout of figures, being careful in avoiding manipulation of data.

2.5 Databases

PubMed – NCBI PubMed (<https://www.ncbi.nlm.nih.gov/pubmed/>) is a free database developed by the National Center for Biotechnology Information (NCBI) for collecting biomedical literature. It accounts millions of records that often link to full-text articles. PubMed was used in this thesis for literature review. Pubmed offers a built-in tool to retrieve records for specific terms by year. This was used for the generation of the graphic in Figure 1.7.

GENCODE GENCODE is a project aimed to the comprehensive identification and annotation of gene features in human and mouse (Frankish *et al.*, 2019). Data from this project are available at <https://www.encodegenes.org/>. In this study, the FASTA sequence and the GENCODE annotation (release 24) of the human genome (GRCh38.p5) were downloaded and used as a reference.

circBase circBase is a publicly available repository of circRNAs (<http://circbase.org/>, Glažar *et al.*, 2014), collecting circRNA annotation from multiple genome-wide studies on different species, tissues or cell lines. As of July 2017, circBase contained information about 140790 circRNAs in human, in addition to circRNAs from other organisms such as *Caenorhabditis elegans*, *Drosophila melanogaster*, and *Mus musculus*. Human circRNA annotation was downloaded from circBase and used in this study to compare our catalogue of circRNAs, after conversion of the circBase hg19 genomic coordinates to version hg38 of the human genome.

circRNADb circRNADb (Chen *et al.*, 2016) is an online database for human circRNAs including exclusively exonic circRNAs, accounting 32914 entries, and available at <http://202.195.183.4:8000/circrnadb/circRNADb.php>. For each circRNA, in addition to the genomic annotation, circRNADb reports information about the protein-coding potential as well as the literature reference. The genomic coordinates of the 32914 human circRNAs were downloaded, converted to version hg38 of the human genome, and used for comparison to our catalogue of circRNAs.

miRBase miRBase (microRNA database) is a free database collecting published miRNA sequences and their respective annotation, available at <http://www.mirbase.org/>. In addition to the nucleotide sequences, miRBase entries report both information about the mature miRNA and the corresponding precursor miRNA (pre-miRNA), together with literature references. In this study, the nucleotide sequences of mature miRNAs defined as high-confidence were downloaded from miRBase (`high_conf_mature.fa.gz`, release 21, Kozomara & Griffiths-Jones, 2013), and used to predict miRNA binding sites on circRNA sequences.

Chapter 3

Results

3.1 Transcriptional and post-transcriptional changes in response to hypoxia

The aim of this thesis was the characterisation of the transcriptional and post-transcriptional response to hypoxia in human cancer cells. To investigate the extent at which hypoxia alters the gene expression in human cancer cells, I analysed rRNA-depleted RNA-Sequencing (RNA-Seq) data of three human cancer cell lines - from lung adenocarcinoma (A549), cervical carcinoma (HeLa) and breast adenocarcinoma (MCF-7), kept in normoxic or hypoxic conditions. To induce hypoxic stress, HeLa cells were incubated at 0.2% O₂ for 24 h, while MCF-7 and A549 cells were incubated at 0.5% O₂ for 48 h. Hypoxic cells were compared to normoxic control cultures, kept at 21% O₂ for an equal time. For each cell line, two or three biological replicates were prepared for the hypoxic and normoxic condition. Hypoxia treatment as well as library preparation and RNA-Seq of HeLa cells were carried out by the group of Prof. Dr. Michaela Müller-McNicoll (Goethe University Frankfurt am Main), while experiments in A549 and MCF-7 cells were performed by the group of Dr. Julia E. Weigand (Technical University Darmstadt).

An overview of the RNA-Seq data and alignment statistics is provided in Table 3.1. The total number of sequencing reads across the different libraries ranged from 60 million reads for HeLa normoxia replicate 1 (N1) up to 144 million reads for A549 normoxia replicate 2 (N2). Reads from A549, HeLa and MCF-7 cells in normoxic and hypoxic conditions were mapped to the human reference genome, after checking the ribosomal RNA content and pre-processing them for a better alignment to the

Table 3.1: Summary of RNA-Seq data from hypoxia experiments in human cancer cells. RNA-Seq was performed from A549, HeLa and MCF-7 cells. The total number of sequenced reads is reported for each replicate and condition, as well as the percentage of sequenced reads mapping to rRNA. The number of processed reads and their percentage that uniquely mapped to the genome, together with the percentage of chimeric reads is also shown.

Cell line	Sample	Total reads	rRNA reads (%)	Processed reads	Mapped reads (%)	Chimeric reads (%)
A549	N1	118103929	8.07	117993957	78.91	0.21
A549	N2	143641614	8.51	143504453	77.65	0.20
A549	H1	105410274	5.34	105296954	79.46	0.21
A549	H2	121800191	4.34	121681536	81.06	0.21
HeLa	N1	60631903	4.07	60595974	85.77	0.31
HeLa	N2	64227967	3.72	64188177	86.69	0.30
HeLa	N3	62772278	4.07	62733339	85.20	0.30
HeLa	H1	64106501	4.22	64067445	84.81	0.32
HeLa	H2	66800563	3.58	66759182	85.27	0.31
MCF-7	N1	113507866	2.32	113397461	81.51	0.21
MCF-7	N2	113938974	2.13	113825152	82.01	0.24
MCF-7	H1	113450829	2.66	113340824	81.27	0.25
MCF-7	H2	125385855	9.28	125263956	73.68	0.26

genome. The percentage of reads mapping to rRNA sequences was higher for A549 normoxia replicate 1 and 2 (N1 and N2) and MCF-7 hypoxia replicate 2 (H2). This may explain the lower alignment rate observed for those samples with higher rRNA content (Figure 3.1A).

The next step in a standard workflow for the analysis of gene expression from RNA-Seq data is the quantification of RNA levels. To capture levels of mature transcripts, the amount of reads mapping to the exonic component of genes was estimated. In order to assess the overall similarity between samples, their distance was computed and visualised (Figure 3.1B). As expected from the experimental setting, replicates clustered together, with a clear separation of conditions (normoxia and hypoxia) and cell lines. A higher similarity between samples from A549 and MCF-7 cells was observed, compared to HeLa samples. This might reflect the different exposure time to low oxygen (48 h versus 24 h), but also that HeLa samples were produced in a different laboratory and sequenced separately.

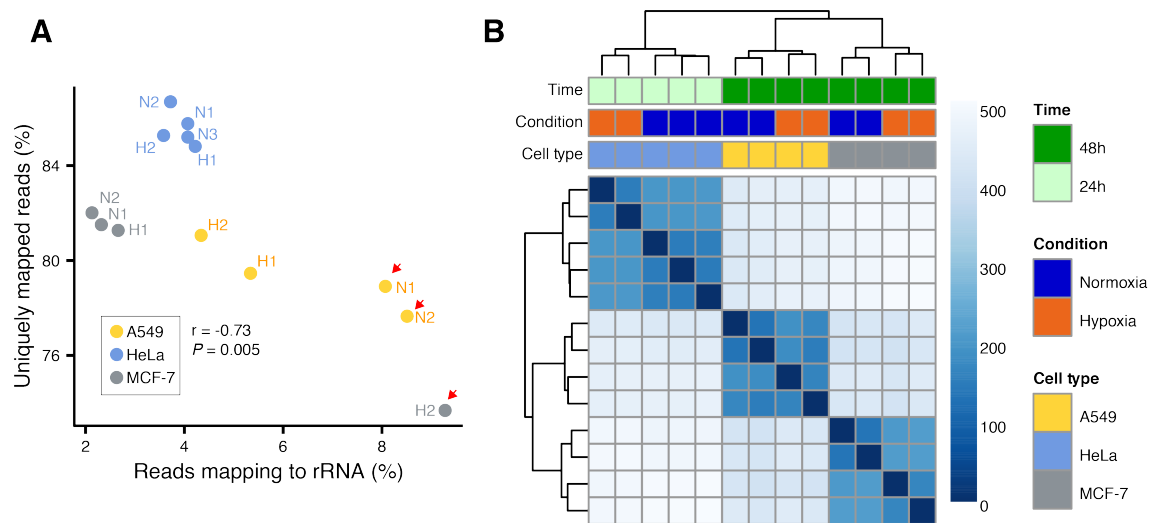


Figure 3.1: Overview of RNA-Seq data from human cancer cells. (A) Relationship between rRNA content and alignment rate to the reference genome. Samples with the highest rRNA content are indicated by red arrows. (B) Heatmap of the sample-to-sample Euclidean distance obtained from the count data from the three cell lines. To avoid that the distance measure is dominated by a few highly variable genes, the variance-stabilising transformation was applied to the count data.

3.1.1 Hypoxia strongly affects linear RNA abundance

Next, raw read counts per gene were used for differential expression testing between normoxic and hypoxic conditions with DESeq2 (Love *et al.*, 2014). Protein-coding genes as well as other linear RNAs annotated in GENCODE (version 24) were considered in this analysis, consisting of three independent tests to compare normoxic and hypoxic samples, one for each cell line. Widespread changes of RNA levels were observed, with 4749, 7962, and 5504 genes changing their level upon hypoxia in A549, HeLa and MCF-7 cells, respectively, summing up to a total of 11876 genes that were regulated in at least one cell line (adjusted P -value < 0.05). Hypoxia affected mainly protein-coding genes (10089, 85%), followed by lincRNAs (564, 5%) and antisense genes (561, 5%). A general trend towards downregulation was observed across the three cell types. However, when a more stringent cutoff was applied on the fold change ($|\log_2(\text{Foldchange})| \geq 1$) to capture the largest changes, the tendency shifted towards a general up-regulation for A549 and HeLa cells and an equal proportion of down/up-regulation for MCF-7 cells (Figure 3.2A). This indicates that there is a stronger induction of the gene expression, probably due to the activation of the HIF proteins. When comparing the 11876 differentially

expressed genes among cell types, a high percentage of them (42%) was regulated in at least two of the analysed cells and 1372 (12%) genes in all cell lines, suggesting a consistent response to hypoxia across cancer cells (Figure 3.2B).

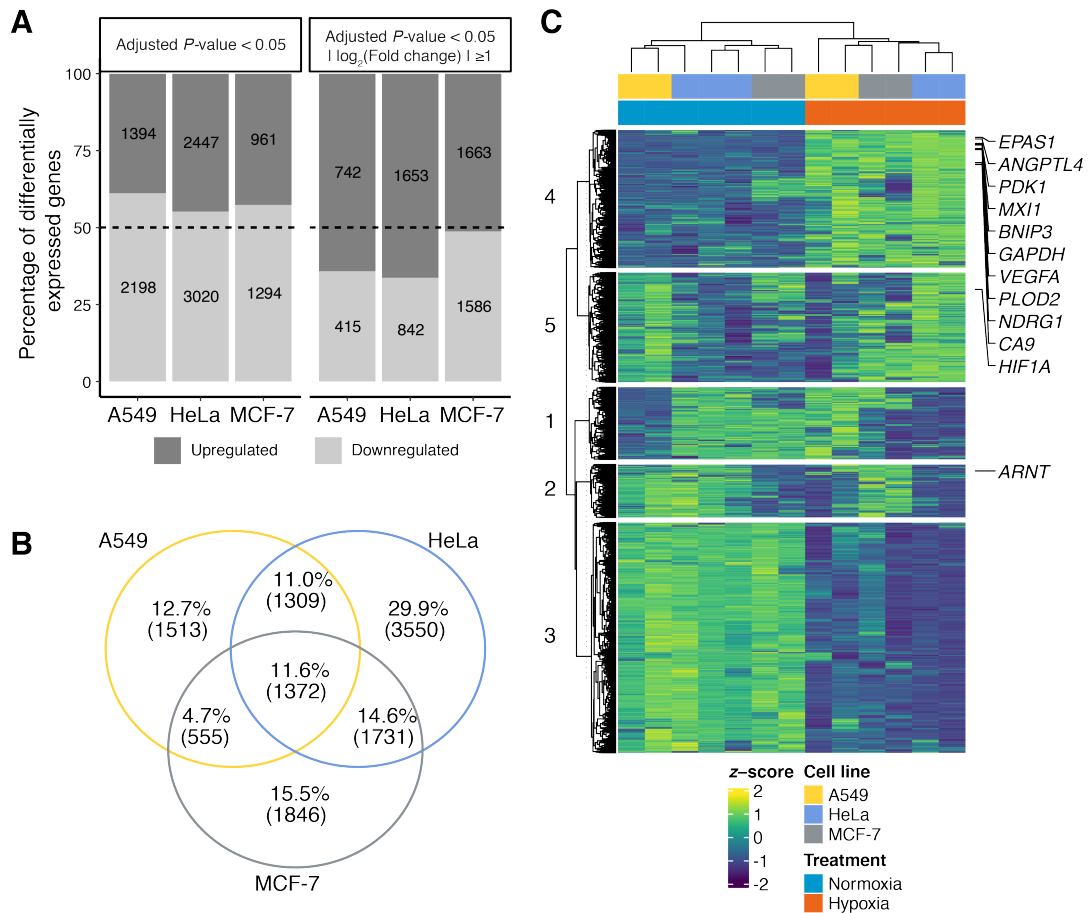


Figure 3.2: Influence of hypoxia on gene expression at RNA level. (A) Differentially expressed genes between normoxia and hypoxia in A549, HeLa and MCF-7 cell lines (adjusted P -value < 0.05). The effect of a further filter on fold change is shown. (B) Venn diagram showing the overlap of hypoxia-regulated genes between cell lines. 4976 circRNAs change their level in at least two cell lines and 1372 genes are regulated in all three cell lines. In brackets, percentages of genes over the 11876 genes significantly regulated across the three cell lines. (C) Heatmap showing the expression levels as z -scores of 11876 genes significantly regulated across the three cell lines. z -scores were computed from log-transformed "transcript per million" values (TPM). Rows were ordered by hierarchical clustering (based on Euclidean distance), and split into five groups based on k-means clustering. Several genes commonly regulated under hypoxia are labelled.

In order to capture the different patterns of regulation upon hypoxia, k-means clustering of differentially expressed genes was performed. This allowed to discriminate

those genes with consistent downregulation in all three cell lines in cluster 3. Moreover, cluster 4 grouped those genes which were induced in all three cell lines, including many well-known hypoxia-responsive genes, such as *CA9*, *ANGPTL4*, *NDRG1*, *PDK1*, *BNIP3*, *PLOD2*, and *VEGFA* (Figure 3.2C). The hypoxia-inducible factor-1 alpha (HIF1 α) governs the initial adaptation to hypoxia and its encoding gene was induced only in HeLa, while its levels decreased in MCF-7 and A549 cells, most likely reflecting the different exposure times to low oxygen adopted for HeLa cells. On the other hand, *HIF2a* (or Endothelial PAS domain protein 1, *EPAS1*) is known to be expressed at a prolonged exposure to oxygen (chronic hypoxia, > 24h). Indeed, it was upregulated in all three cell types, suggesting that HeLa cells were at a transition point between acute and chronic hypoxia. *ARNT* (*HIF1 β*) was only slightly induced in MCF-7 cells ($\log_2(\text{Foldchange}) = 0.59$), and remained stable in the other two cell lines.

To gain functional insights into the transcriptional response to hypoxia, Gene Ontology (GO) enrichment analysis of differentially expressed genes in the three cell lines was performed (Figure 3.3). As expected, hypoxia-induced genes were associated to the response to decreased oxygen levels, as well metabolic adaptation, angiogenesis and cell migration. In HeLa and MCF-7 cells, also GO terms related to the apoptotic signalling pathways were among the top enriched. On the other hand, downregulated genes were related to ribosome biogenesis, DNA replication and aerobic metabolism, reflecting a shift towards lower energy consumption in response to hypoxia. In addition, downregulated genes were involved in RNA splicing, suggesting possible alterations of the splicing pattern.

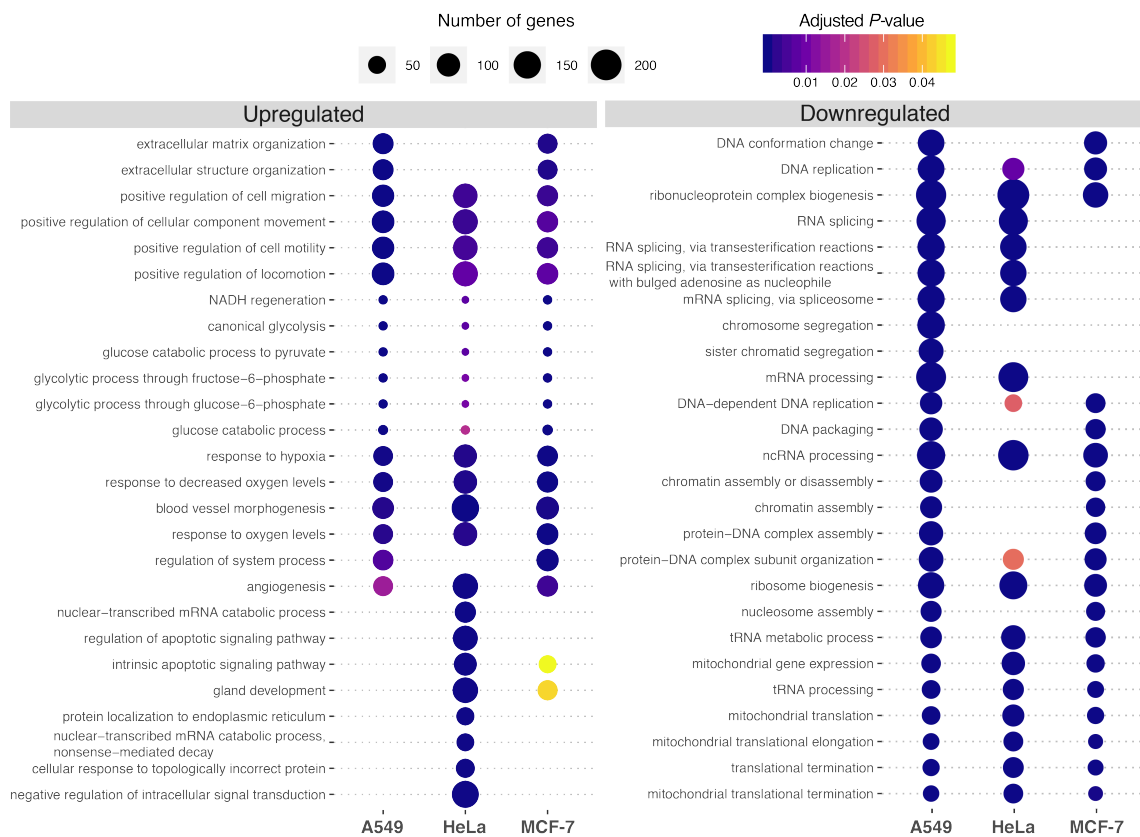


Figure 3.3: Functional characterisation of differentially expressed genes upon hypoxia. Results from Gene Ontology (Biological Process) enrichment analysis by hypergeometric testing for hypoxia-regulated protein-coding genes are shown (adjusted P -value/ q -value < 0.05). The 10 most significant GO terms per group are compared.

3.1.2 Hypoxia alters the splicing pattern in a cell type-specific manner

An additional layer of regulation of gene expression is alternative splicing, which from a single gene produces multiple transcript isoforms, adding further variability to the final proteome of a cell. In order to investigate the alternative splicing pattern in cancer cells and how it changes upon hypoxia, we applied replicate multivariate analysis of transcript splicing (rMATS; Shen *et al.*, 2014) to our RNA-Seq data. The following types of alternative splicing events were explored: "cassette exon" (CE), "retained intron" (RI), "mutually exclusive exon" (MXE), "alternative 3' splice site" (A3SS) and "alternative 5' splice site" (A5SS) usage, as illustrated in Figure 3.4A. Accompanying the transcriptional response, we observed a global change in splicing, detecting in total 9701 significant differential AS events upon hypoxia from 4715 genes (absolute difference in percent spliced-in, $|\Delta\text{PSI}| \geq 10\%$; false discovery rate, $\text{FDR} < 5\%$). More specifically, 2441, 3473 and 4482 alternatively spliced loci changing their levels were detected in A549, HeLa and MCF-7 cells, respectively. Among the detected AS events, three types were specifically enriched in the analysed cell lines, namely the alternative inclusion of cassette exons, mutually exclusive exons and intron retention events (P -value < 0.00001 , 2-sample test for equality of proportions). The predominant directionality of change varied across cell types (Figure 3.4B,C). In contrast to the convergent transcriptional response, however, the splicing changes upon hypoxia were divergent, both in terms of regulated events and directionality, with only 37 significantly regulated AS events overlapping between all three cell lines (Figure 3.4D). To further assess the relationship between hypoxia-regulated genes and AS changes, we compared genes that undergo differential AS to the differentially expressed genes upon hypoxia. 41-66% genes subjected to splicing changes were not regulated at RNA level, suggesting that alternative splicing is a process often regulated independently from transcription in hypoxia. In summary, the three human cancer cell lines react to hypoxia similarly in terms of gene expression variation but differently in terms of alternative splicing, which adds an additional layer of regulation and confers cell type specificity to the hypoxic adaptation.

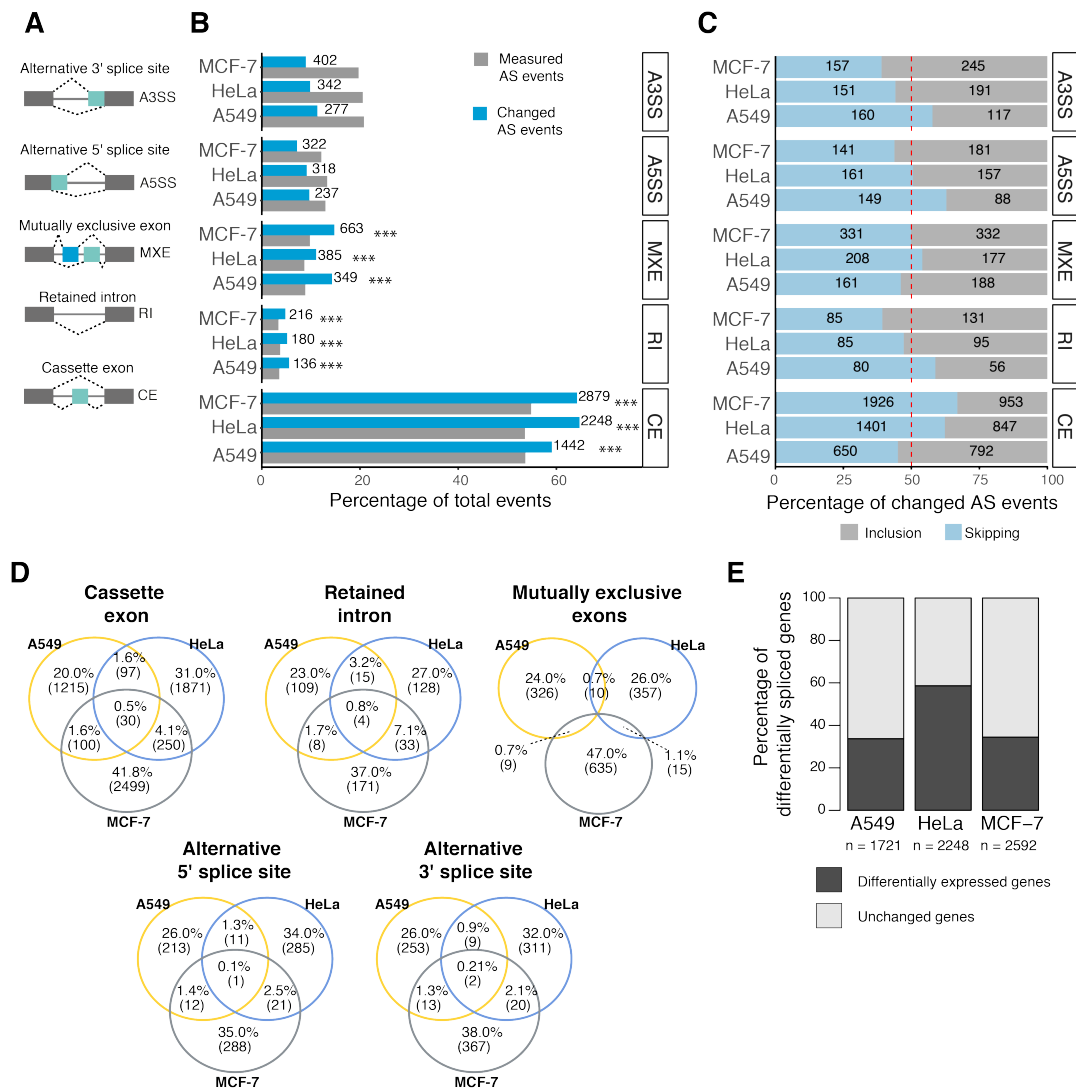


Figure 3.4: Alternative splicing profile upon hypoxia. (A) Scheme of the different types of alternative splicing events detectable with rMATS. (B) Percentage of total measured and significantly changed alternative splicing events in the three cell types ($\Delta\text{PSI} \geq 10\%$, $\text{FDR} < 5\%$; P -value < 0.00001 , test for equality of proportions). (C) Percentage and number of inclusion and skipping AS events in the three cell lines. Dashed red line: 50%, equal proportion of directionality. For A5SS and A3SS, the inclusion isoform is the longest exon isoform. For MXE, ΔPSI the inclusion isoform is the first exon. (D) Hypoxia-deregulated alternative splicing is highly different between A549, HeLa and MCF-7 cells. Venn diagram comparing AS events between cell lines. (E) The regulation of alternative splicing of many genes is independent RNA abundance. Bar plot shows the percentage of differentially spliced genes that do not change their global RNA level (light grey) in hypoxia, in comparison to those that additionally undergo expression changes at RNA level (dark grey).

3.2 The role of MBNL2 in hypoxia

Hypoxia both affected the mRNA abundance of splicing-related genes and altered the splicing pattern in A549, HeLa and MCF-7 cells. Prompted by this, we explored in more detail the expression changes of genes annotated to the GO term "RNA splicing" (GO:0008380). 314 out of 428 genes annotated in this list were regulated in at least one of the three cell lines, with 57 genes significantly changing in all three cell lines (Figure 3.5A). These included core spliceosomal proteins and members of the SR protein family. SR proteins are essential RNA-binding proteins able to influence each step of the mRNA life. Among genes encoding SR proteins, *SRSF1*, *SRSF2*, *SRSF6*, *SRSF7* and *SRSF8* were all consistently downregulated upon hypoxia at mRNA level. *SRSF6*, was previously shown to influence the splicing of *VEGFA*, a well-known HIF target in the response to hypoxia. In particular, *SRSF6* promotes the splicing of the anti-angiogenic isoform *VEGFA165b* (Peiris-Pagès, 2012). *SRSF6* showed strongly reduced mRNA levels, being almost halved in HeLa and MCF-7 cells (\log_2 -transformed fold change = -0.92 and -0.93, respectively; adjusted P -value < 0.05). Only few splicing-related genes were consistently induced by low oxygen levels in the three cell types, namely *MBNL2*, *DHX32*, *NOL3*, *AHNAK*, and *CLK1*. *MBNL2* is a well-known splicing factor, and a member of the muscleblind-like protein family, together with *MBNL1* and *MBNL3*. *MBNL2* was recently shown to be abundant in clear cell renal cell carcinoma (Perron *et al.*, 2018) and hepatocellular carcinoma (Lee *et al.*, 2016), acting as oncogene or tumour suppressor gene, respectively, pointing out a possible role of *MBNL2* in cancer cells' adaptation to hypoxia. Noteworthy, only *MBNL2* was regulated upon hypoxia, while *MBNL1* remained stable and showed high abundance already in normoxic conditions. *MBNL3* levels were low both in normoxic and hypoxic conditions (Figure 3.5B). Changes in expression of *MBNL2* were confirmed by reverse transcription-quantitative polymerase chain reaction (RT-qPCR) and Western blot experiments in A549 and MCF-7 cells (Figure 3.5C,D).

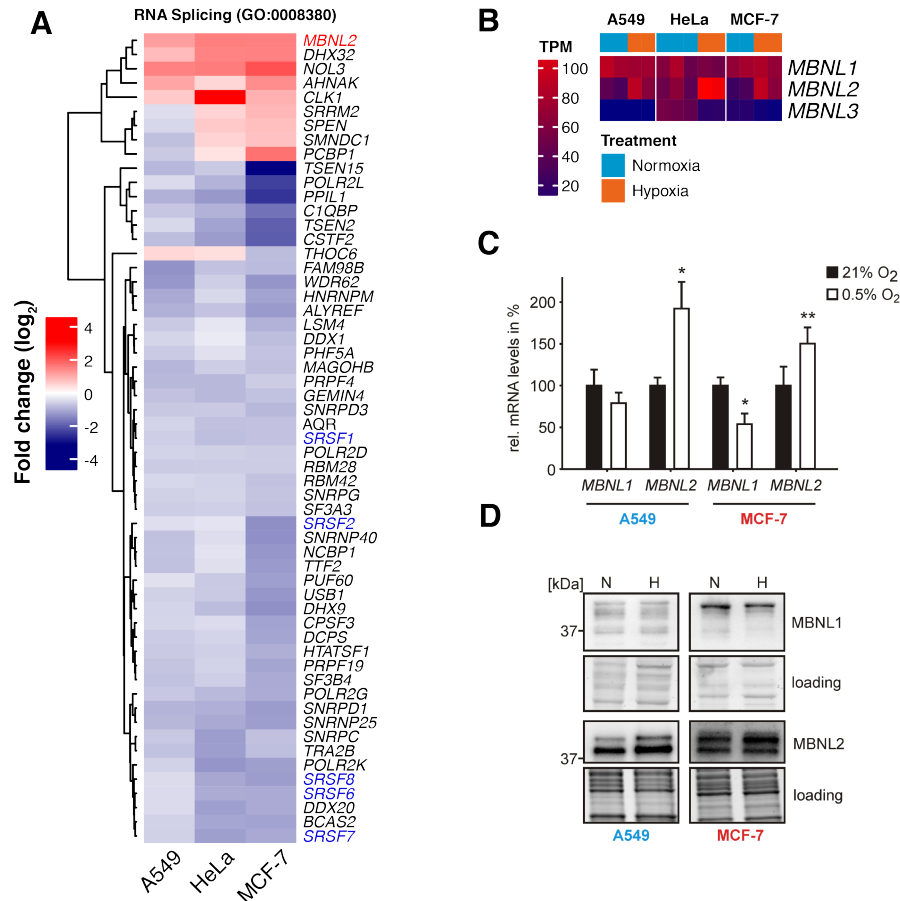


Figure 3.5: Changes in mRNA levels of genes annotated to the GO term "RNA splicing" (GO:0008380). (A) Heatmap representing fold changes of RNA splicing genes co-regulated in A549, HeLa and MCF-7 cells. Members of the SR protein family are labelled in blue. MBNL2 is among the few upregulated RBPs upon hypoxia (labelled in red). (B) Comparison of expression levels as "transcripts per million" (TPM) of MBNL proteins. *MBNL2* is consistently upregulated in the three cell lines, while *MBNL1* levels are abundant and stable upon hypoxia and *MBNL3* shows generally low levels. (C) RT-qPCR validating *MBNL2* induction upon hypoxia at mRNA level in A549 and MCF-7 cells. Values were normalised to the *RPLP0* gene ($n = 4$). 21% O₂ = normoxia; 0.5% O₂ = hypoxia. * P -value < 0.05 , ** P -value < 0.01 . (D) Western blot of MBNL1 and MBNL2 in normoxic (N) and hypoxic (H) conditions. Anti-MBNL1 and anti-MBNL2 antibodies were used to visualise the respective protein levels. Total lane protein is shown as loading control ($n = 2$). RT-qPCR and Western blot experiments were performed by Sandra Fischer, Technical University Darmstadt.

3.2.1 MBNL2 modulates the transcript abundance of hypoxia response genes

In collaboration with the Weigand group (Technical University Darmstadt), *MBNL2* was selected to elucidate its role in hypoxia. With this aim, hypoxic samples from MCF-7 and A549 cells were treated with a short interfering RNA (siRNA) targeting *MBNL2* (siMBNL2) and a control siRNA (siCTRL). The efficiency of the knock-down was confirmed by Western blot (Figure 3.6A). Interestingly, the treatment of hypoxic cancer cells with cisplatin, a commonly used chemotherapeutic drug (Dasari & Bernard Tchounwou, 2014), caused an increase of the cell death rate exclusively upon *MBNL2* knockdown (Figure 3.6B). Moreover, *MBNL2* knockdown decreased the migration rate of hypoxic A549 cells (unpublished data; Fischer *et al.*, *in revision*). Altogether, these experiments suggest that the induction of *MBNL2* observed in hypoxia might contribute to the adaptation of cancer cells to hypoxia.

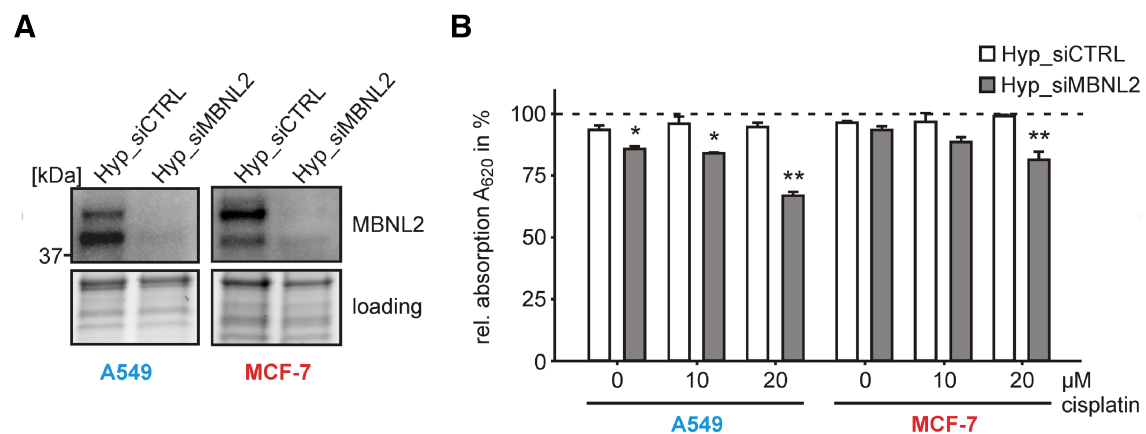


Figure 3.6: Cisplatin-induced cell death increases upon *MBNL2* depletion in hypoxic cancer cells. (A) Western blot confirming the depletion at protein level of MBNL2 in hypoxic cells upon treatment with siMBNL2 ($n = 3$). Total lane protein is shown as loading control. (B) Cell viability assay of hypoxic cancer cells upon *MBNL2* knockdown and treatment with cisplatin for 24 h. The relative absorption was normalised to normoxic control cells treated with the same cisplatin concentration ($n = 3$). * P -value < 0.05 , ** P -value < 0.01 . Experiments performed by Sandra Fischer, Technical University Darmstadt.

In order to elucidate the molecular mechanism by which MBNL2 influences the adaptation of cancer cells to hypoxia at transcriptome-wide level, I analysed RNA-Seq data of hypoxic MCF-7 cells with siMBNL2 or siCTRL treatment. Two replicates per condition were prepared and the extracted RNA was sequenced after poly(A)

Table 3.2: Summary of RNA-Seq data from *MBNL2* knockdown experiments in hypoxic MCF-7 cells. RNA-Seq was performed from A549, HeLa and MCF-7 cells. The total number of sequenced reads is reported for each replicate and condition. In addition, the number of processed reads and the percentage of reads uniquely mapped to the genome is shown.

Cell line	Condition	Treatment	Total reads	Processed reads	Mapped reads (%)
MCF-7	Hypoxia	siCTRL	82850336	82776484	91
MCF-7	Hypoxia	siCTRL	83766468	83694572	91
MCF-7	Hypoxia	siMBNL2	77005724	76941479	91
MCF-7	Hypoxia	siMBNL2	91045087	90964015	91

selection. An overview of the RNA-Seq data and alignment statistics is provided in Table 3.2.

Considering the dual function of *MBNL2* in regulating stability and alternative splicing of its RNA targets, we examined both alterations in expression at RNA level, as well as in the splicing pattern. Next, we compared the outcome of these analyses to our previous results from the normoxia/hypoxia RNA-Seq experiments (from now referred to as "hypoxia experiments"). As for the hypoxia experiments, after a quality check and pre-processing of sequencing reads, they were aligned to the human reference genome. The total number of reads across the different libraries ranged from 77 up to 91 millions. Once trimmed, an alignment rate higher than 90% was obtained for all samples considering only uniquely mapped reads, indicating a general high quality of the data (Table 3.2). Gene levels were estimated as described above for the hypoxia experiments' data and differential expression testing was performed with DESeq2. In total, 5580 genes significantly changed their level upon *MBNL2* knockdown (adjusted P -value < 0.05). Here, a further filter on the list of differentially expressed genes was applied, demanding TPM > 1 in any single sample and an absolute fold change higher than 1.5, thus restricting the number of regulated genes to 4529. This resulted in an equal proportion of up- and downregulated genes (2137 and 2392, respectively). The efficiency of the treatment with siRNA targeting *MBNL2* was confirmed by RNA-Seq, that identified *MBNL2* as the top downregulated gene (\log_2 -transformed fold change = -3.18; adjusted P -value = 7.48E-63). Importantly, the levels of *MBNL1* and *MBNL3* were not affected (\log_2 -transformed fold change = 0.02 and -0.07, respectively), indicating the specificity of the knock-down treatment. Notably, functional inspection of the regulated genes revealed for downregulated genes the enrichment of the GO term "response to hypoxia", as well

as additional biological processes directly attributable to the hypoxia adaptation, including glucose metabolism and translation (Figure 3.7A). No enriched GO term could be found for upregulated genes.

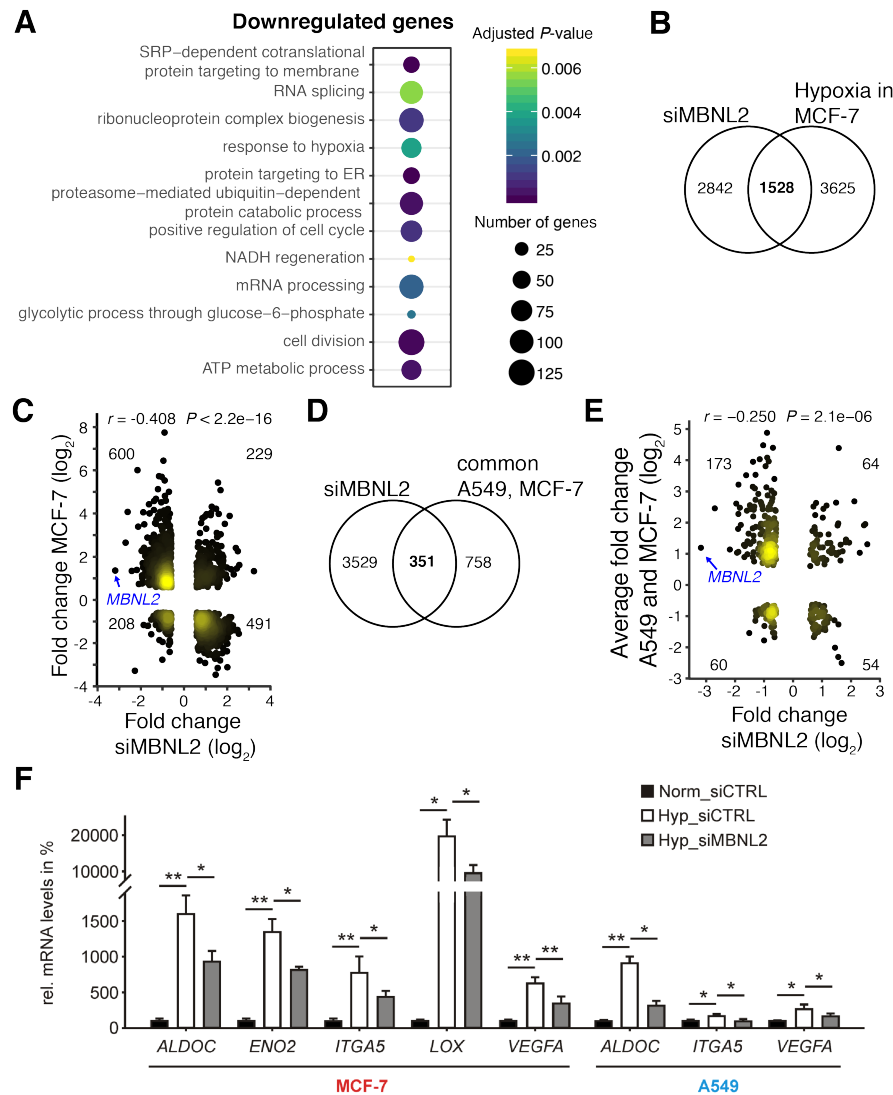


Figure 3.7: $MBNL2$ depletion influences transcript abundance in hypoxic cancer cells. (A) Gene Ontology (GO) analysis of differentially expressed genes (DEGs) after $MBNL2$ knockdown in hypoxic cells. Selected enriched GO terms are shown P -value < 0.05 , q -value < 0.05 . (B) Comparison of significantly changed genes after $MBNL2$ depletion and in response to hypoxia in MCF-7 cells (adjusted P -value < 0.05 ; absolute fold change > 1.5 ; TPM > 1 in any single sample). For $MBNL2$ knockdown (siMBNL2), only DEGs expressed in the hypoxia experiments in MCF-7 cells were considered in the comparison (TPM > 1 in any single sample; $n = 4370$). (C) Scatter plot of fold changes of DEGs regulated both after $MBNL2$ knockdown and upon hypoxia in MCF-7 cells ($n = 1528$). Number of genes is given in each quadrant. r : Pearson correlation. (D) Venn diagram of DEGs after $MBNL2$ depletion and shared between MCF-7 and A549 in hypoxia (adjusted P -value < 0.05 ; absolute fold change > 1.5). For $MBNL2$ knockdown (siMBNL2), only DEGs expressed in the hypoxia experiments in MCF-7 and A549 cells were considered in the comparison (TPM > 1 in any single sample; $n = 3880$). (E) Scatter plot of fold changes of genes changing abundance in both MCF-7 and A549 cells in response to hypoxia and in $MBNL2$ -depleted hypoxic MCF-7 cells. Visualisation as in (C). (F) RT-qPCR confirms the modulatory role of $MBNL2$ on the induction of HIF target genes upon hypoxia. Values were normalised to $RPLP0$ ($n = 3-6$). (* P -value < 0.05 , ** P -value < 0.01 , two-tailed Student's t -test). RT-qPCR performed by Sandra Fischer.

To gain insights into the role of MBNL2 in the transcriptional adaptation to hypoxic stress, the expression analysis of RNA-Seq data from *MBNL2* knockdown and hypoxia experiments in MCF-7 cells were integrated. For this purpose, also the differentially expressed genes (DEGs) from hypoxia experiments in MCF-7 cells were similarly filtered (absolute fold change higher than 1.5, TPM > 1 in any single sample), leading to 5153 DEGs. In addition, for comparison between expression changes in hypoxia and upon *MBNL2* knockdown, DEGs upon *MBNL2* knockdown that were not expressed in MCF-7 hypoxic experiments (TPM < 1 in all samples) were not considered for downstream analyses. In total, 30% of the hypoxia-regulated genes changed their RNA abundance also upon *MBNL2* knockdown, with most of them (71%) showing opposite regulation in the two experiments, consistent with MBNL2's upregulation in hypoxia (Figure 3.7B,C). The strong overlap together with the anticorrelation of shared changes are good indicators of a role of MBNL2 in hypoxia. This led us to speculate that the regulation of these genes in the hypoxic condition might be dependent on MBNL2 function. Similarly, DEGs upon *MBNL2* knockdown were compared to genes consistently regulated in hypoxic A549 and MCF-7 cells (n= 1109), showing an overlap of 351 genes (32%, Figure 3.7D). This suggests that the role of MBNL2 in hypoxia response might be independent of the cell type. Again, these shared genes were mostly regulated in the opposite direction compared to *MBNL2* knockdown (65%, Figure 3.7E). They included known HIF targets such as *BNIP3*, *ANKRD37* and *CA9*, the genes involved in promoting angiogenesis, namely *VEGFA*, *ADM1* and *ANGPTL4* and the genes encoding the glycolytic enzymes ALDOA, ALDOC and GAPDH (Chi *et al.*, 2006, Benita *et al.*, 2009, Lendahl *et al.*, 2009, Sena *et al.*, 2014, Semenza, 2012). In order to verify that MBNL2 affects the hypoxia-dependent induction of selected HIF targets, RT-qPCR was performed in MCF-7 and A549 cells. The experiments confirmed the modulatory function of MBNL2 on the RNA levels of *ALDOC*, *ENO2*, *ITGA5*, *LOX* and *VEGFA* in hypoxic MCF-7 cells and *ALDOC*, *ITGA5* and *VEGFA* in hypoxic A549 cells (Figure 3.7F).

Perron and coauthors predicted that MBNL2 might influence the stability of mRNAs, mainly acting as a stabilising factor, as for *VEGFA* (Perron *et al.*, 2018). They defined a list of 130 genes that are stabilised by MBNL2 binding to their 3'UTRs (Ray *et al.*, 2013, Perron *et al.*, 2018). We inspected these genes to verify whether their RNA levels decreased upon *MBNL2* knockdown, and tested whether they were also induced upon hypoxia. 81 of the 130 genes were expressed in *MBNL2*

Table 3.3: List of putative MBNL2 stability targets (Ray *et al.*, 2013; Perron *et al.*, 2018), which are downregulated upon *MBNL2* knockdown. The fold change values upon *MBNL2* knockdown (siMBNL2) and upon hypoxia treatment of MCF-7 cells (hypoxia) are reported if significant (adjusted *P*-value < 0.05).

Gene name	Fold change siMBNL2 (\log_2)	Fold change hypoxia MCF-7 (\log_2)
<i>SMAD7</i>	-0.85	0.64
<i>CSRNP1</i>	-1.68	0.80
<i>SERTAD2</i>	-1.41	0.88
<i>OSMR</i>	-0.90	1.06
<i>LONRF2</i>	-0.59	-
<i>MGST3</i>	-0.60	-
<i>SPOPL</i>	-0.63	-
<i>PNPLA8</i>	-0.73	-
<i>TBC1D20</i>	-1.00	-
<i>HPCAL1</i>	-1.01	-

knockdown experiments, and only 10 genes were downregulated (Table 3.3). Among them, *SMAD7*, *OSMR*, *SERTAD2* and *CSRNP1* were additionally upregulated upon hypoxia in MCF-7 cells (Table 3.3). For these four genes, putative MBNL2 binding sites consisting of two clustered 5'-YGCY-3' motifs (Lambert *et al.*, 2014) could be identified in their 3'UTR sequences (Figures S1, S2, S3). Additional experiments are required to establish whether MBNL2 effectively binds these motifs and to test its effect on mRNA decay and translation.

To further explore the function of MBNL2 as mRNA stabilising factor, the prediction of putative MBNL2 binding sites on 3'UTRs was extended to all genes regulated upon *MBNL2* knockdown, by scanning the 3'UTR sequences for clustered 5'-YGCY-3' motifs (5'-YGCYGCY-3' and 5'-YGCYN₀₋₃YGCY-3'). No substantial difference was found between genes with predicted MBNL2 binding sites or not in terms of expression changes (Figure 3.8A). Similarly, deregulated genes upon *MBNL2* knockdown did not show higher density of predicted MBNL2 binding sites (Figure 3.8B). It has been reported that the binding of MBNL proteins depends on the number of YGCY repeats, with a preference for multiple GC with variable spacing (Lambert *et al.*, 2014; Taylor *et al.*, 2018). In order to test this preference, we used the size of predicted binding sites as a proxy of clustered 5'-YGCY-3' motifs. The size of

predicted binding sites was not significantly higher for deregulated genes compared to genes that do not change upon *MBNL2* knockdown (Figure 3.8C). In contrast to previous publications that hypothesised an mRNA-stabilising role of MBNL2, our results did not reveal any evidence of this function via 3'UTR binding.

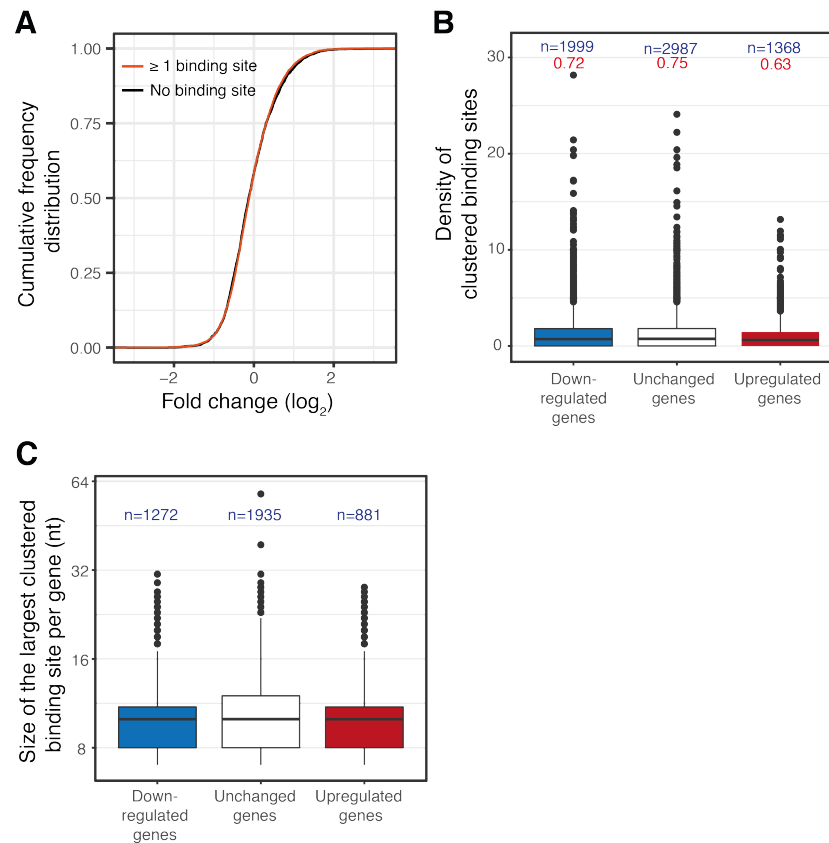


Figure 3.8: *In silico* prediction of MBNL2 binding sites on 3'UTRs sequences of MBNL2-regulated genes. (A) Cumulative frequency distribution of expression changes after *MBNL2* knockdown in hypoxic MCF-7 cells. Genes with at least one predicted MBNL2 binding site (cluster of multiple YGCY motifs) are compared to genes lacking MBNL2 binding sites. (B) Density of non-overlapping MBNL2 binding sites (number of binding sites divided by 3'UTR length in kb). Adjacent MBNL2 binding sites within a 3 bp distance were collapsed and counted as one. Up- and downregulated genes are compared to unchanged genes after *MBNL2* knockdown, defined by fold change < 1.3 and adjusted *P*-value > 0.5 . Labels in blue: number of observations; in red: median density. (C) Maximum width of MBNL2 binding sites per gene. Adjacent MBNL2 binding sites within a 3 bp distance were collapsed. The width of the resulting collapsed binding sites was used as a proxy of the number of YGCY repeats. When multiple collapsed binding sites were still present on the 3'UTR of a single gene, the largest collapsed binding site was considered in the graphic (maximum width). Genes with a single or no YGCY motif in their 3'UTR sequence were excluded. Labels in blue: number of observations.

3.2.2 MBNL2 controls hypoxia-dependent alternative splicing

MBNL2 is mainly known for its function as regulator of RNA splicing. We profiled the alternative splicing pattern upon *MBNL2* knockdown, finding 2074 significantly changed AS events, including the alternative inclusion of cassette exons, mutually exclusive exons, retained intron, and alternative 3' or 5' splice site selection events, again with a prevalence of CE events (absolute $\Delta\text{PSI} \geq 10\%$, $\text{FDR} < 5\%$, $\text{TPM} > 1$ in any sample; Table 3.4). AS events upon *MBNL2* knockdown were then compared to 4411 AS events detected in hypoxic MCF-7 cells (absolute $\Delta\text{PSI} \geq 10\%$, $\text{FDR} < 5\%$, $\text{TPM} > 1$ in any sample). In contrast to the effect of hypoxia on MCF-7 cells, which caused more inclusion of cassette exons (Figure 3.4C), cassette exons were preferentially skipped after *MBNL2* depletion (Table 3.4). In total, 393 AS events were significantly regulated in both experiments, which represented the 9% of the events upon hypoxia treatment and the 19% of the events due to *MBNL2* knockdown (Figure 3.9A). Among these shared AS events, 307 (78%) were cassette exons events. Similarly to changes at RNA level, the majority of AS events shared between the two experiments (89%) were regulated in opposite direction, underlying a potential role of MBNL2 in modulating splicing in the hypoxia adaptation (Figure 3.9B). The alternative inclusion of exon 12 from the *PIGN* gene (RefSeq transcript NM_176787) was selected for validation by RT-qPCR. The lower inclusion rate of exon 12 observed upon hypoxia was restored upon treatment with siMBNL2, indicating a function as negative regulator of the alternative cassette exon for MBNL2 (Figure 3.4C). In line with the cell type-specific alteration of the splicing patterns observed in A549, HeLa and MCF-7, only 24 of the 393 events shared between *MBNL2* knockdown and hypoxia in MCF-7 cells were also detected in A549 cells (22 CE, 1 MXE, and 1 RI events) (Figure 3.4D). In addition, cassette exons of

Table 3.4: Alternative splicing events after *MBNL2* knockdown. CE: cassette exon; MXE: mutually exclusive exons; RI: retained intron; A5SS and A3SS: alternative 5' and 3' splice site. For A5SS and A3SS, ΔPSI indicates the difference of inclusion levels for the longest exon isoform. For MXE, ΔPSI indicates the difference of inclusion levels for the first exon.

	CE	MXE	RI	A5SS	A3SS
Number of events	1497	201	140	120	116
$\Delta\text{PSI} \geq 10\%$	1093	99	22	40	33
$\Delta\text{PSI} \leq -10\%$	404	102	118	80	83

EXOC7, *MORF4L2*, *ESYT2* and *NOL8* were similarly regulated in hypoxic HeLa cells ($\Delta\text{PSI} = -24\%$, -12% , -19% , and 23% , respectively; $\text{FDR} < 5\%$). It would be interesting to verify whether *MBNL2* depletion affects splicing of this genes also in hypoxic A549 and HeLa cells.

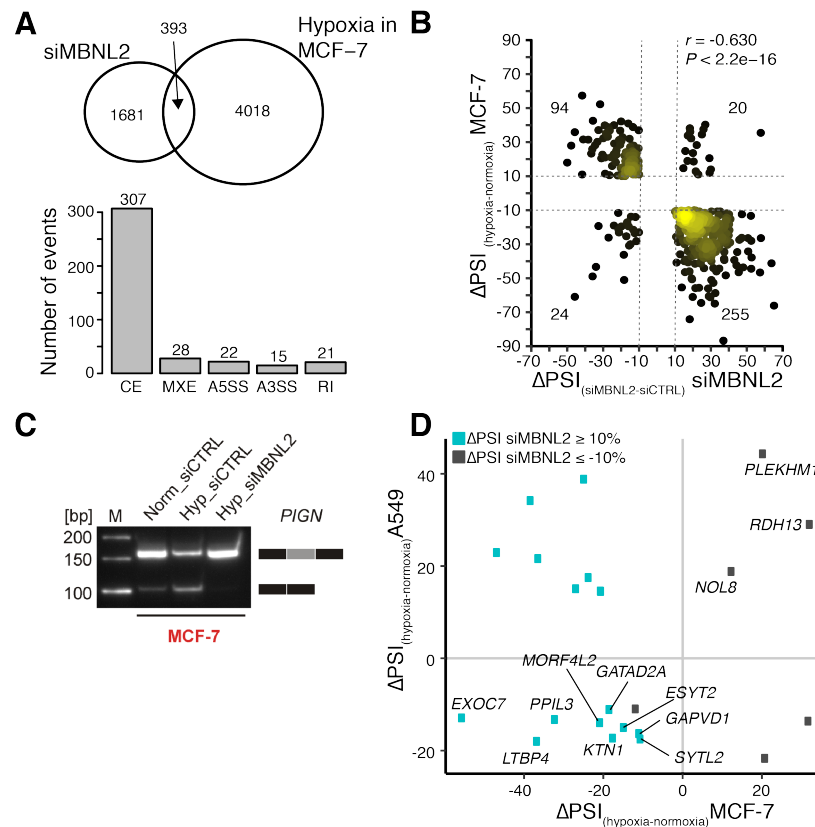


Figure 3.9: MBNL2 mainly functions as negative regulator of alternative cassette exons in hypoxia. (A) Top: Venn diagram comparing alternative splicing (AS) events upon hypoxia and after *MBNL2* knockdown (absolute $\Delta\text{PSI} \geq 10\%$, $\text{FDR} < 5\%$, $\text{TPM} > 1$ in any sample). Bottom: Shared AS events ($n = 393$) between hypoxia and *MBNL2* knockdown in MCF-7 cells, divided by type. (C) Scatter plot of ΔPSI values of the 393 significant AS events shared between hypoxia and *MBNL2* knockdown in MCF-7 cells. (D) RT-qPCR confirms the role of *MBNL2* in regulating the cassette exon of *PIGN* upon hypoxia (exon 12, RefSeq transcript NM_176787). (E) Scatter plot of ΔPSI values of 22 cassette exon events shared between hypoxic A549 cells, hypoxic MCF-7 cells and *MBNL2* knockdown experiments.

In summary, our results revealed that *MBNL2* is upregulated at low oxygen and contributes to the response to hypoxia in cancer cells. It regulates the mRNA levels of known HIF target genes involved in key processes for cancer progression, including angiogenesis, metastasis and metabolic reprogramming. In addition, *MBNL2* acts as

a splicing factor, modulating a large fraction of alternative splicing events occurring in hypoxia, with a tendency to promote the skipping of cassette exons at low oxygen.

3.3 Establishing a pipeline to identify circRNAs from rRNA-depleted RNA-Seq data

CircRNAs represent a class of ncRNAs produced by a particular splicing mechanism, named back-splicing or head-to-tail splicing, by which a 5' splice site (5'SS) is covalently joined to a 3' splice site (3'SS) located upstream in the transcript. CircRNAs have been reported to be altered in cancer (Geng *et al.*, 2018) and recent studies revealed their regulation in hypoxic endothelial cells (Boeckel *et al.*, 2015) as well as hypoxic gastric cancer cells (Li *et al.*, 2018a). These observations, together with the widespread changes affecting RNA levels and alternative splicing patterns in hypoxic cancer cells revealed by our analyses, motivated us to further investigate the expression profile of circRNAs in cancer cells and their changes upon hypoxia. In this study, the first step towards the characterisation of the circRNA profile in cancer cells was the establishment of a pipeline to detect back-splicing events from the rRNA-depleted RNA-Seq data provided by the Müller-McNicoll Group (Goethe University Frankfurt am Main), and the Weigand Group (Technical University Darmstadt).

With the advances in high-throughput technologies and the discovery of the pervasive expression of circRNAs across the tree of life, several computational algorithms have been proposed to detect back-splicing events from RNA-Seq data (Table 1.1 and reviewed in Szabo & Salzman, 2016, Gao & Zhao, 2018, Jakobi & Dieterich, 2019). Recent studies evaluated the performance of several circRNA tools on rRNA-depleted RNA-Seq samples, by relying on RNase R-treated RNA-Seq samples as a source to detect real circRNAs. They agreed on the fact that the outcome of circRNA detection tools is only partially consistent and their performance can vary considerably. They further suggested to combine the prediction of such tools to obtain a more reliable catalogue of circRNAs from RNA-Seq data (Hansen *et al.*, 2016, Zeng *et al.*, 2017, Hansen, 2018). Consequently, we decided to combine outcomes from two different algorithms, `CIRCexplorer` (Zhang *et al.*, 2014) and `find_circ` (Memczak *et al.*, 2013), to comprehensively describe the circRNA repertoire of different cancer cell lines. When circRNA prediction tools were compared, `CIRCexplorer` was described as one of the outperforming tools (Hansen *et al.*, 2016, Zeng *et al.*, 2017). However, a major disadvantage of `CIRCexplorer` might be that it strongly relies on the gene annotation to detect circRNAs from back-spliced exons and intron

lariats. For instance, it cannot detect relevant circRNAs such as circZNF292, which originates from a cryptic black-splice site located in an intronic region, known to get induced upon hypoxia in endothelial cells (Boeckel *et al.*, 2015). In addition, its predictions would strongly depend on the choice of the gene annotation. Although `find_circ` showed worse performance in previous comparisons (Hansen *et al.*, 2016, Zeng *et al.*, 2017), it would be suitable to complement `CIRCexplorer` prediction, since it allows a *de novo* prediction of back-splicing events independently of prior knowledge of exon-intron annotation, thus expanding the spectrum of circRNA types in the final catalogue. In addition, it has the advantage that it is fast and has low RAM requirements (Hansen *et al.*, 2016). `CIRCexplorer` was initially designed to parse fusion junction information from mapping results of `TopHat` (`Bowtie2`) and `TopHat-Fusion` algorithms, to identify and annotate those reads potentially attributable to circRNAs. In later versions, it supports also the alignment software `STAR`. To avoid a bias related to the usage of a single read mapper (`Bowtie2`) for the initial prediction of circRNAs, we decided to test and use `CIRCexplorer` in combination with the alignment software `STAR`. `STAR` was originally designed to deal with reads that map to non-contiguous regions of the reference genome (spliced reads) and allows the detection of chimeric alignments, i.e. discontinuous arrangements in which the two aligned fragments of the read are in a non-linear order (Dobin *et al.*, 2012). These chimeric alignments are parsed by `CIRCexplorer` to retrieve back-splicing events.

3.3.1 Testing `CIRCexplorer` and `find_circ` on rRNA-depleted RNA-Seq data

In order to build up a pipeline based on `CIRCexplorer` and `find_circ`, I first explored their strengths and weaknesses independently, predicting circRNAs from a single sample of our RNA-Seq data (normoxia replicate 1, HeLa cells, Figure 3.10A). `find_circ` was used as described in Memczak *et al.*, 2013, demanding unique alignments, unambiguous breakpoints, a maximum genomic distance between the back-splice sites of 100 kb, and GU/AG splice site sequences, leading to the prediction of 3752 back-splicing events. Differently from `CIRCexplorer`, the `find_circ` algorithm outputs two different measures of back-splice reads supporting circRNAs: "`n_reads`" indicates the total number of reads supporting the back-splice junction,

and "n_uniq" refers to the number of reads supporting the back-splice junction that differ in their sequence. Here, I refer to "n_reads" as total reads and "n_uniq" as unique reads. In the original paper (Memczak *et al.*, 2013), candidate circRNAs were further filtered based on unique reads, in order to consider only those circRNA reads arising from independent reverse transcription events, thus avoiding putative PCR duplicates. Here, the initial prediction of `find_circ` was further filtered requiring a minimum of two unique reads. This significantly reduced the amount of candidate circRNAs to 1118. Even applying this further filter, only 37% of the detected circRNAs could also be predicted in the other normoxia replicates, suggesting either a certain level of noise in `find_circ` predictions or in the experiment itself. `CIRCexplorer` initially predicted a total of 3343 circRNAs; the number decreased to 2269 circRNAs when a cutoff of 2 total back-splice reads was applied, including 76 intronic circRNAs (ciRNAs). The reproducibility of `CIRCexplorer` was higher compared to `find_circ`, with 68% circRNAs detectable also in the other normoxia replicates with the same tool, indicating that the low reproducibility of the `find_circ` outcome is rather attributable to a certain level of noise in the `find_circ` prediction.

Comparing the 2269 circRNAs detected by `CIRCexplorer` to the 1118 circRNAs found with `find_circ`, 903 circRNAs were consistently detected by both tools (Figure 3.10A). This observation was in line with previous studies that showed a modest overlap between predictions of five different circRNA tools (Hansen *et al.*, 2016). For the 903 circRNAs in common, the number of total reads reported by the two tools was highly correlated, although not always identical (Pearson correlation coefficient = 0.93) (Figure 3.10B). 285 circRNAs were detected only by `find_circ` and 1,366 circRNAs were detected only by `CIRCexplorer` (Figure 3.10A). A deeper investigation of back-splicing events detected exclusively by `CIRCexplorer` or `find_circ` revealed that those events are supported by less total reads if compared to the 903 circRNAs predicted by both tools, although several circRNAs were particularly abundant, suggesting a potential biological relevance (Figure 3.10C).

At this point, the main features that might influence discrepancies of the outcomes were investigated. Approximately two-thirds (67%) of the 285 circRNAs detected by `find_circ` but not `CIRCexplorer` were detected by `STAR` and present in its chimeric junction output. 133 circRNAs in this group were filtered out by `CIRCexplorer` because they originated from unannotated junctions. For the remaining 58 circRNAs, the reason of the discrepancy remains unclear. Finally, 94 (33%) circRNAs were

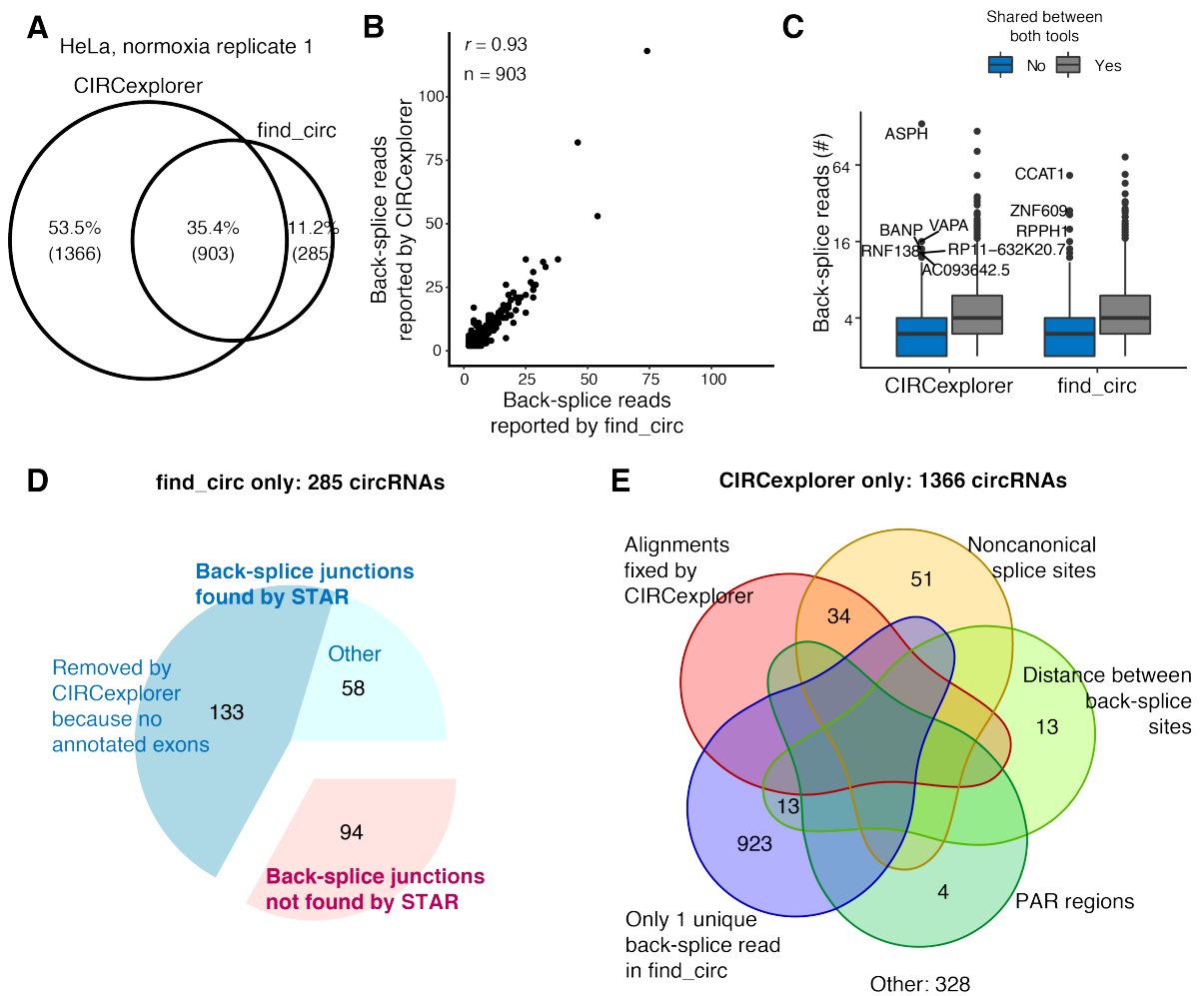


Figure 3.10: Investigation of convergences and discrepancies between CIRCexplorer and find_circ tools in predicting circRNAs from a single HeLa sample (normoxia, replicate 1). (A) Venn diagram depicting the overlap between the predictions of the CIRCexplorer and the find_circ algorithms. (B) Scatter plot comparing total back-splice reads estimated by the two algorithms for the 903 circRNAs in common. r : Pearson correlation. (C) Box plot shows the quantification of circRNAs by CIRCexplorer and find_circ for circRNAs in common or from only a single algorithm. CircRNAs detected only by a single tool are in general less abundant, with notable exceptions (labelled). (D) Characterisation of the 285 circRNAs predicted exclusively by find_circ, in terms of genomic aligner (STAR vs. Bowtie2) and gene annotation. (E) Characterisation of the 1366 circRNAs predicted exclusively by CIRCexplorer, in terms of splice-site signal, genomic size, supporting back-splice reads, alignment adjustment by CIRCexplorer and gene annotation. 69% circRNAs are supported by a single unique back-splice read in find_circ measurements, highlighting the importance of filtering on unique back-splice reads to exclude PCR artefacts. For 328 circRNAs, the reason of the discrepancy remained unclear.

not present in the chimeric junction output of **STAR**, therefore predictable only by **find_circ** from **Bowtie2** alignments (Figure 3.10D).

Of the 1366 circRNAs detected exclusively by **CIRCexplorer**, multiple features were the cause of the inconsistency with the outcome of **find_circ** (Figure 3.10E). For instance, 24 circRNAs were initially detected by the **find_circ** algorithm, but filtered out due to a genomic distance between back-splice sites higher than 100 kb. 85 circRNAs showed a splice site signal different from GU/AG, that is not allowed by **find_circ**. Just a small number of circRNAs (n=34) derived from an adjustment of **STAR** read alignments intrinsically performed by **CIRCexplorer**. Interestingly, four circRNAs were reported to originate from chromosome Y, although HeLa cells were originally derived from cervical cancer of a female patient. Further investigation revealed that **CIRCexplorer** assigned them to pseudoautosomal regions (PAR), homologous sequences between chromosome X and Y. The filter applied on the number of back-splice reads was the main underlying reason of the discrepancy between **CIRCexplorer** and **find_circ**, since 946 out of 1366 circRNAs (69%) were initially predicted by **find_circ**, but filtered out due to the presence of a single unique back-splice read, thus considered as putative PCR artefacts. Indeed, the overlap between **CIRCexplorer** and **find_circ** was larger when **find_circ** circRNAs were filtered based on total reads (≥ 2 reads, 2513 circRNAs), with 1803 circRNA predicted by both tools.

To avoid dependence on the specific dataset, I also investigated the circRNA prediction on a sample from MCF-7 cells (normoxia, replicate 1), obtaining consistent results (Supplementary Figure S4). Indeed, a total of 3056 and 4890 circRNAs were identified with **find_circ** and **CIRCexplorer**, respectively, as described above. 2381 (48%) circRNAs were predicted by both tools and highly correlated in terms of total back-splice reads (Pearson correlation coefficient = 0.98) (Figure S4A,B). The overall amount of supporting reads was lower for circRNAs found only by a single algorithm, with several exceptions (Figure S4C). Of the 675 circRNAs detected exclusively by **find_circ**, 170 (25%) circRNAs were not detected by **STAR** but only from unmapped reads obtained from **Bowtie2**. 295 circRNAs were included in **STAR** chimeric alignments but not reported in the **CIRCexplorer** output, because they originated from unannotated exons (Figure S4D). Again, the main reason justifying the discrepancy between **CIRCexplorer** and **find_circ** was the filter that **find_circ** applies on unique back-splice reads, with 1863 out of 2509 (74%) of the **CIRCexplorer**-only circRNAs filtered out by **find_circ** because they were

supported by less than 2 unique reads (Figure S4E).

3.3.2 A novel combined pipeline for circRNA detection

These results offered the basis to design a pipeline that combines the strengths of `CIRCexplorer` and `find_circ` to obtain a comprehensive catalogue of circRNAs in human cancer cells (Table 3.5). Figure 3.11 shows a scheme of the pipeline.

Table 3.5: Overview of the main features of algorithms adopted in this thesis to detect circRNAs. BSSs: Back-splice sites.

Tool	<code>find_circ</code>	<code>CIRCexplorer</code>	Our pipeline
Alignment algorithm	<code>Bowtie2</code>	<code>STAR</code>	Both
Splice site motifs	GU/AG	Any	GU/AG or GC/AG
Splice sites	Annotated + <i>de novo</i>	Annotated	Annotated, <i>de novo</i>
Genomic distance of BSSs	≤ 100 kb	Within single gene	≤ 100 kb
Expression filter	2 unique reads (sequence-based)	2 total reads	2 unique reads (coordinate-based)

The workflow starts with a quality check of sequencing reads followed by filtering and trimming of reads, when necessary. Next, samples of a single dataset are merged together and reads are independently mapped to the reference genome using `Bowtie2` (Langmead & Salzberg, 2012) and the splice-aware aligner `STAR` (Dobin *et al.*, 2012) setting parameters that allow to output chimeric alignments. Unmapped reads are extracted from the `Bowtie2` output and used as input for `find_circ`, followed by standard filtering (unambiguous breakpoint, unique alignments, maximum genomic distance between BSSs of 100 kb, GU/AG splice site signal). The chimeric junction table from `STAR` is used as input for `CIRCexplorer`. At this point, no expression filter is applied. Differently to what it was suggested in previous publications to increase reliability of the prediction (Hansen *et al.*, 2016, Hansen, 2018), we decided to consider not only the intersection between `CIRCexplorer` and `find_circ` outcomes, but rather to unify the predictions of both tools and filter out the detection artefacts revealed by the comparison described above. The underlying reason is that we did not want to leave out abundant circRNAs that are detected by a single algorithm, such as the hypoxia-induced circZNF292 (Boeckel *et al.*, 2015), originating from

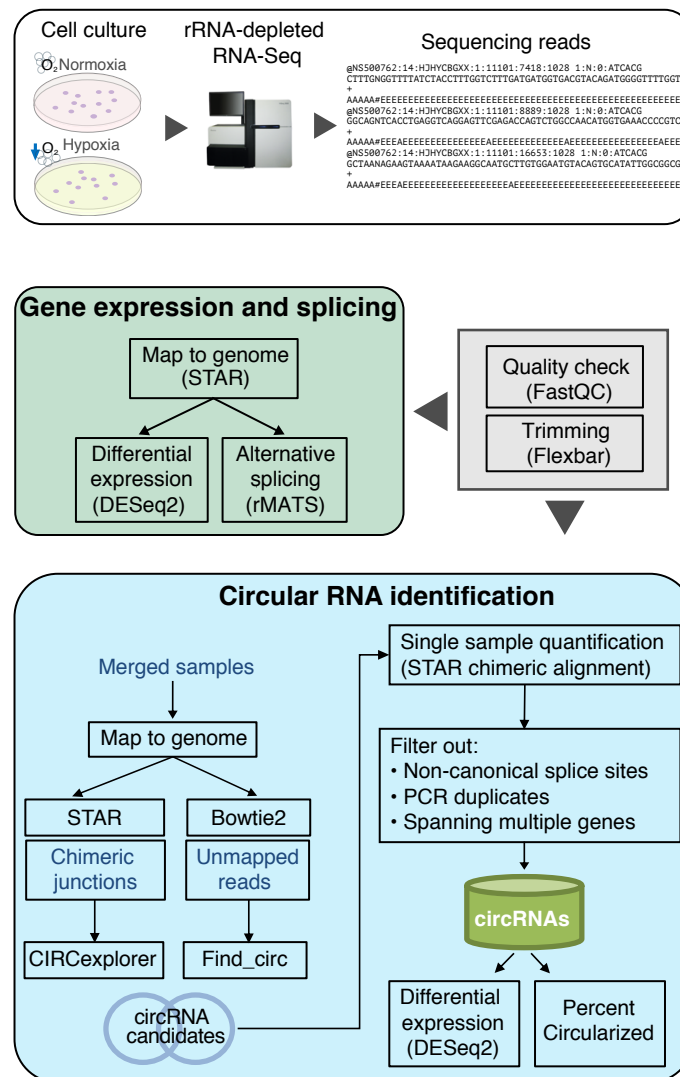


Figure 3.11: Pipeline to identify circRNAs and analyse gene expression and splicing from rRNA-depleted RNA-Seq data.

the covalent bond between the 5' splice site of exon 4 and a cryptic 3' splice site located within the first intron of *ZNF292*. Thus, the full lists of candidate circRNAs detected by either CIRCexplorer or find_circ were merged and used as annotation to quantify circRNA expression. This was done by counting the number of unique and total reads supporting back-splice junctions from the STAR chimeric alignments. This was performed for each single sample of the dataset with a custom script in R that parses STAR chimeric alignments. Importantly, differently from find_circ, unique back-splice reads were defined based on the mapping position rather than the read sequence itself, increasing the stringency in detecting PCR artefacts. Finally,

the pipeline filtered out: (i) candidates with a genomic distance between BSSs higher than 100 kb or spanning non-overlapping annotated genes; (ii) circRNAs with splice-site signal different to GU/AG (canonical signal) or GC/AG, known to be the most represented non-canonical splice site (Burset *et al.*, 2000); (iii) circRNAs supported by less than 2 unique reads. The final catalogue of obtained circRNAs, together with the circRNA quantification, was then used for downstream analyses, including differential expression testing with DESeq2 and a "circular-to-linear ratio" (or "percent circularised") for comparison to the expression of host genes. Due to the usage of STAR to obtain both chimeric alignments and uniquely mapped reads, the circRNA prediction and quantification complements with the gene expression and alternative splicing analyses described above.

3.3.3 Evaluating the performance of the pipeline with RNase R-treated RNA-Seq data

To evaluate the overall performance of our combined pipeline compared to the usage of CIRCexplorer and find_circ independently, the algorithms were tested on publicly available RNA-Seq data. These RNA-Seq data were produced from the sequencing of total RNA, depleted of the rRNA and with or without RNase R treatment to enrich for circRNA reads. An overview of the datasets used in this analysis is presented in Table 3.6. The RNA-Seq raw data were downloaded from the NCBI's Sequence Reads Archive (SRA) using the SRA accession numbers reported in Table 3.6. Three different datasets were analysed, two from cervical cancer (HeLa cells) (Gao *et al.*, 2015, Gao *et al.*, 2016), and one from human fibroblasts (Hs68 cells, Jeck *et al.*, 2013). Datasets consisted of paired-end reads sequenced on an Illumina platform, with a read length of 101 bp for HeLa data and 100 bp for Hs68 data. The total number of reads was considerably different among datasets, ranging from 26.6 million for the HeLa SRR1636985 sample to 412.7 million reads for the Hs68 SRR444975 sample. After cleaning reads based on the quality, they were mapped to the human reference genome (GRCh38/hg38) with STAR. Interestingly, the percentage of reads mapping to a single genomic locus was relatively low for all RNase R-treated samples, independent of the library, corresponding to a high percentage of multimapping reads, likely arising from rRNA, as highlighted from an independent mapping to human rRNA sequencing performed with Bowtie2 (Table 3.6). For the

Hs68 dataset, also the rRNA-depleted RNA-Seq samples contained a large amount of reads from rRNAs.

Similar to previous studies that compared computational tools for the detection of circRNAs (Hansen *et al.*, 2016, Wang *et al.*, 2017, Zeng *et al.*, 2017, Hansen, 2018), we detected back-splicing events in the rRNA-depleted RNA-Seq data and compared the supporting read counts to the ones obtained from RNase R-treated samples. We assumed that RNase R-treated data allow to detect real circRNAs, therefore to estimate the amount of true positives and false positives. Although variability might be generated by the biochemical treatment with RNase R and some real circRNAs like the human CDR1as, circCAMSAP1, circMAN1A2, and circNCX1 are sensitive to RNase R treatment (Szabo & Salzman, 2016), RNase R-treated datasets remain one of the most widely used and accepted method to validate the circularity of this class of ncRNAs in genome-wide studies.

For each library, first a list of candidate circRNAs from the rRNA-depleted RNA samples was obtained. CircRNAs were detected applying our pipeline to the rRNA-depleted RNA samples for each specific dataset, as described above. A minimum of 2 unique reads in at least one of the rRNA-depleted RNA samples was required to call the circRNA as detected by our pipeline. Our prediction was compared to `CIRCexplorer` and `find_circ` separately. `CIRCexplorer` was used in combination with the aligner `STAR`, filtering on total back-splice reads (≥ 2) since the tool does not provide any estimate of the unique reads. `find_circ` performance was evaluated either filtering on total reads, similar to `CIRCexplorer`, or on unique reads, similar to our pipeline, since both estimates are reported by the tool. Our pipeline predicted 1350 circRNAs from the HeLa sample SRR3476958, 2685 circRNAs from the HeLa samples SRR1637089 and SRR1637090, and up to 3786 circRNAs from the Hs68 samples SRR444975 and SRR444655. The higher number of circRNAs detected in SRR444975 and SRR444655 might be explained by the deeper sequencing of the library and the slightly lower content in rRNA (Table 3.6 and Figure 3.12A). Compared to our pipeline, `find_circ` predicted more circRNAs with either settings, while the amount of circRNAs predicted with `CIRCexplorer` was comparable for all tested datasets (Figure 3.12A). Considering that for both `CIRCexplorer` and our pipeline, back-splice reads are quantified from `STAR` chimeric alignments, it is likely that the higher number of predicted circRNAs is due to the lower stringency applied to back-splice counts, since it does not differentiate between unique and total reads.

The underlying reason for the larger amount of circRNAs predicted by `find_circ` compared to `CIRCexplorer` might partially reside in the types of circRNAs the tools are able to detect, *de novo* or from annotated exon boundaries, respectively. Venn diagrams in Figure 3.12B illustrate the overlap between the different tools under investigation for each of the analysed libraries. Most circRNAs predicted with our pipeline originate from `CIRCexplorer`, although `find_circ` predicted a much higher number of back-splicing events.

Table 3.6: Overview of RNA-Seq datasets used for the performance evaluation of the proposed pipeline

Cell type	SRA accession number	Library preparation	Total reads	rRNA (%)	Uniquely mapped reads (%)	Chimeric reads (%)	Reference
HeLa	SRR3476956	rRNA-/RNase R-treated	100,236,474	71.56	19.91	0.56	Gao et al. (2016)
HeLa	SRR3476958	rRNA-	51,780,130	1.04	83.05	0.64	Gao et al. (2016)
HeLa	SRR1636985	rRNA-/RNase R-treated	26,619,490	71.66	19.61	0.50	Gao et al. (2015)
HeLa	SRR1636986	rRNA-/RNase R-treated	47,011,426	71.09	19.53	0.48	Gao et al. (2015)
HeLa	SRR1637089	rRNA-	89,866,900	1.04	82.53	0.62	Gao et al. (2015)
HeLa	SRR1637090	rRNA-	71,370,620	0.69	87.56	0.57	Gao et al. (2015)
Hs68	SRR444974	rRNA-/RNase R-treated	316,611,710	71.10	15.54	0.52	Jeck et al. (2013)
Hs68	SRR445016	rRNA-/RNase R-treated	399,844,972	56.09	26.59	1.25	Jeck et al. (2013)
Hs68	SRR444655	rRNA-	314,106,316	55.29	16.21	0.23	Jeck et al. (2013)
Hs68	SRR444975	rRNA-	412,725,466	66.72	26.64	0.49	Jeck et al. (2013)

In order to compare the performance of our pipeline to `CIRCexplorer` and `find_circ`, read counts of circRNAs were normalised by sequencing depth (total number of sequenced reads) and a fold change was computed as the ratio between normalised back-splice counts in RNase R-treated RNA-Seq samples over rRNA-depleted RNA-Seq samples (RNase R / rRNA-depleted). When replicates were available, the mean of normalised counts between replicates was used to calculate the fold change. Similar to Zeng *et al.*, 2017 and Hansen, 2018, the resulting fold-enrichment values were then used to classify the candidate circRNAs into "RNase R-sensitive", and "RNase R-resistant", depending on whether a reduction or increase of back-splice counts in RNase R-treated samples was observed, respectively, compared to rRNA-depleted RNA samples. RNase R-sensitive circRNAs were further divided into "RNase R-depleted" when the circRNA is either undetectable in the RNase R-treated samples or at least 5-fold decreased, and "RNase R-reduced" when a decrease up to 5-fold was observed. RNase R-resistant circRNAs were further divided into "RNase R-enriched" for those circRNAs with at least a 5-fold increase and "RNase R-unaffected" when their level remains stable or increases up to 5-fold in the RNase R-treated samples.

When evaluating the performance of the tools through the fold-enrichment values,

`find_circ` reported the largest proportion of RNase R-sensitive circRNAs over all the datasets (34-65%), with only a slight improvement when unique reads were used for the detection (30-60%). `CIRCexplorer` performed much better, predicting only 20-46% RNase R-sensitive circRNAs, together with a higher rate of RNase R-enriched circRNAs (6-56%). The amount of circRNAs identified in SRR1637089 and SRR1637090 (HeLa cells) and SRR444975 and SRR444655 (Hs68 cells) with `find_circ` and `CIRCexplorer`, as well as the percentage of RNase R-depleted circRNAs, was comparable to previous studies in which the same cutoff on back-splice reads was applied (≥ 2), although we averaged between replicates instead of combining them into a single sample (Zeng *et al.*, 2017). Our pipeline showed higher "precision" when compared to `find_circ` and was at least comparable to `CIRCexplorer`, if not better for the HeLa samples SRR1637089 and SRR1637090, detecting 20-29% RNase R-sensitive circRNAs and 6-58% RNase R-enriched circRNAs. Altogether, these results suggest that our pipeline extends the already valuable circRNA prediction of `CIRCexplorer` to circRNAs originated from unannotated junctions of the genome and detectable only with `find_circ`, while keeping high precision.

In summary, strengths and weaknesses of `CIRCexplorer` and `find_circ` were investigated on two samples from independent experiments, revealing the dependence on a gene annotation and the filter on unique/total reads as major sources of disagreement between the two tools. Based on these findings, we established a pipeline that combines the initial circRNA predictions by both tools and then filters out inconsistencies and detection artefacts of either algorithms. The pipeline harmonises the quantification estimates by recounting back-splice reads for all circRNAs from the `STAR` chimeric alignments. The evaluation of the performance of our pipeline using RNase R-treated RNA-Seq samples as a source of genuine circRNAs, led us to the conclusion that this pipeline is well suited to obtain a comprehensive and reliable catalogue of circRNAs from rRNA-depleted RNA-Seq data.

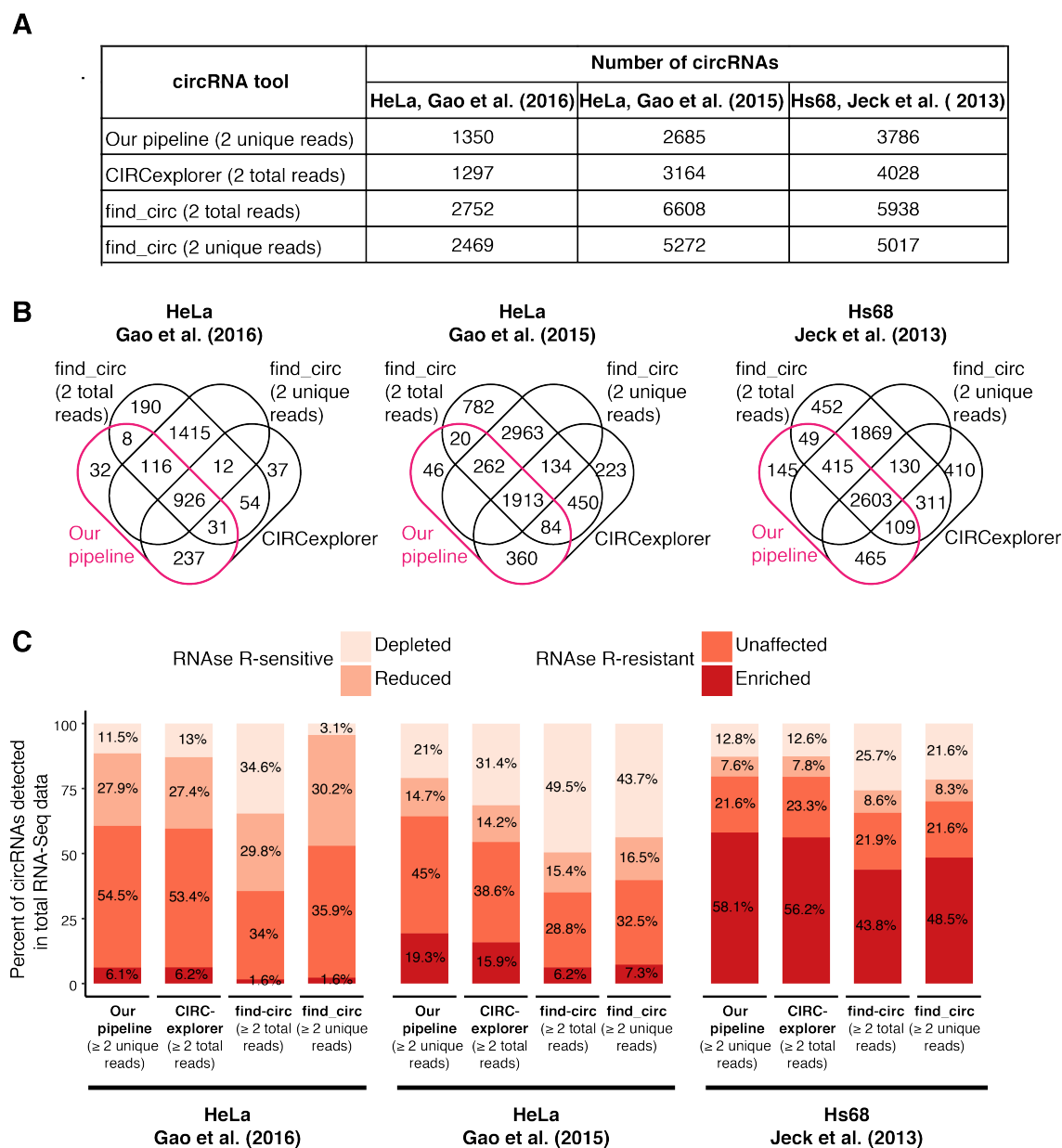


Figure 3.12: Evaluation of circRNA prediction with our pipeline compared to CIRCexplorer and find_circ, based on published RNase R-treated RNA-Seq datasets. Candidate circRNAs were classified into RNase R-sensitive and RNase R-resistant based on the fold-enrichment between normalised back-splice counts of rRNA-depleted RNA-Seq and RNase R-treated RNA-Seq samples. **(A)** Summary of circRNAs predicted with our pipeline, CIRCexplorer and find_circ from total RNA-Seq data. **(B)** Venn diagrams showing the overlap of the prediction from our pipeline to CIRCexplorer and find_circ predictions for each RNA-Seq dataset. Only rRNA-depleted RNA-Seq data were used for this estimate. **(C)** Percentages of RNase R-sensitive and RNase R-resistant circRNAs for each RNA-Seq dataset. RNase R-sensitive circRNAs are further divided into RNase R-depleted if decreased by at least 5-fold and RNase R-reduced if reduced up to 5-fold. Similarly, RNase R-resistant circRNAs were further divided into RNase R-unaffected if increased up to 5-fold and RNase R-enriched if increased by more than 5-fold.

3.4 CircRNome profiling in cancer cells

Once I established the pipeline for the detection of circRNAs, I applied it to our normoxia and hypoxia RNA-Seq data from A549, HeLa and MCF-7 cells, detecting in total 12006 circRNAs across the three cell lines (Table 3.7, Figure 3.13A).

Table 3.7: Number of circRNAs identified in cancer cells.

Sample	Back-splice reads per million mapped	Number of circRNAs
A549		4599
N1	144	2870
N2	129	2998
H1	186	1999
H2	141	2406
HeLa		3926
N1	134	1508
N2	116	1489
N3	155	1728
H1	154	1746
H2	169	1973
MCF-7		7527
N1	242	4194
N2	282	4275
H1	325	4612
H2	382	5429
All		12006

Even though the sequencing depth of MCF-7 and A549 RNA-Seq data was very similar (Table 3.1), the number of circRNAs detected in MCF-7 cells (n=7527) was considerably higher compared to A549 cells (n=4599). This might indicate genuine differences in the abundance of circRNAs as well as experimental variations, for instance originating from the variable efficiency of the rRNA depletion during library preparation. Indeed, the rRNA content in MCF-7 was in general lower compared to A549 cells, except for hypoxia, replicate 2 (Table 3.1). HeLa data could not be

considered in this comparison due the reduced sequencing depth compared to MCF-7 and A549 data, that resulted in a lower number of detected circRNAs ($n = 3926$) (Table 3.7).

I compared our catalogue to the 140790 circRNAs deposited in circBase (Glažar *et al.*, 2014) and the 32914 circRNAs collected in circRNADb (Chen *et al.*, 2016), finding that 2844 (24%) of circRNAs detected in this study had not been reported previously (Figure 3.13B). For instance, our pipeline predicted novel circRNAs from the genes *HUWE1*, *SPIDR* and *PICALM*, which were present in multiple cell lines and supported by more than 20 back-splice reads. As expected from previous studies (Memczak *et al.*, 2013, Salzman *et al.*, 2013, Guo *et al.*, 2014, Zhang *et al.*, 2014), the majority of circRNAs were supported by less than five back-splice reads, suggesting that they might be by-products of the splicing process and have no relevant biological function (Figure 3.13C). The low abundance might also be due to loss of material during library preparation and/or sequencing. Despite a generally low abundance of circRNAs, the pipeline identified many abundant circRNAs, with 1392 circRNAs being supported by a minimum of 10 back-splice reads in at least one replicate, over the three cell lines. The top expressed circRNAs were circCYP24A1 (circBase ID hsa_circ_0060927), circASPH (hsa_circ_0084615) and circATXN7 (hsa_circ_0007761) in A549, HeLa, MCF-7 cells, respectively. These circRNAs captured more than 150 back-splice reads in a single replicate.

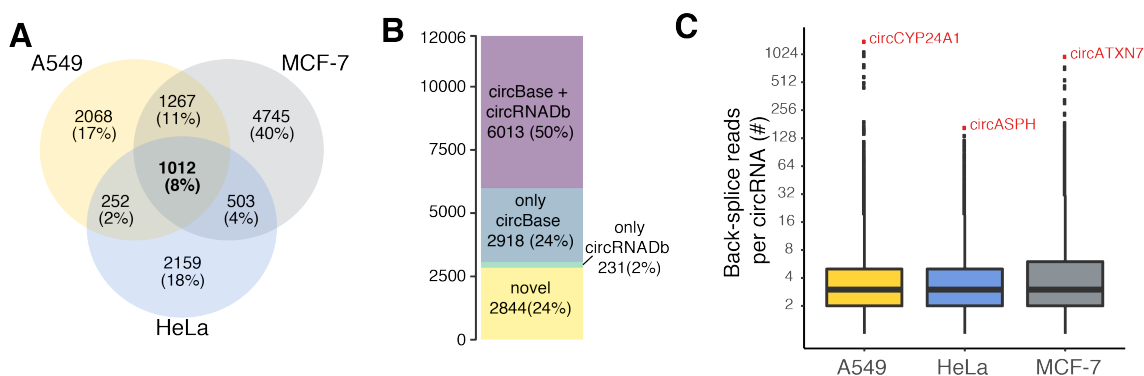


Figure 3.13: Identification of circRNAs in human cancer cell lines. (A) Venn diagram depicting the overlap of circRNA predictions in A549, HeLa and MCF-7 cells. 1012 circRNAs are expressed in all three cell types. (B) Comparison of the 12006 circRNAs in our catalogue to circBase and circRNADb annotations. (C) Boxplot showing the distribution of total back-splice reads from each cell line. Labelled in red, the top expressed circRNAs for each cell line (circCYP24A1 in A549, circASPH in HeLa and circATXN7 in MCF-7 cells).

In collaboration with Camila de Oliveira Freitas Machado (Müller-McNicoll Group), we selected ten abundant circRNA candidates to validate the prediction by RT-PCR. Divergent pairs of PCR primers flanking the back-splice junctions were designed, in order to amplify a PCR product exclusively from the circular transcript but not from the corresponding linear transcript, as outlined in Figure 3.14A. Two different approaches were used to validate that the amplified PCR products are circRNAs. One of the main features of circRNAs is that they lack a poly(A) tail, thus they should be amplified only in the non-polyadenylated fraction (poly(A)-), when this is separated from the polyadenylated fraction (poly(A)+) of the total RNA. Indeed, we verified the presence of amplification products only in the poly(A)- fraction using divergent primers for all selected circRNAs (Figure 3.14B,C). In addition, RNase R exonuclease can only digest linear RNA molecules, not affecting circRNAs. The ten candidate circRNAs were resistant to RNase R-treatment, while the control linear RNA was digested (*PLOD2*, Figure 3.14B,C). In line with the test on RNase R-treated RNA-Seq data, these experiments further confirmed that the pipeline is suitable to detect genuine circular transcripts.

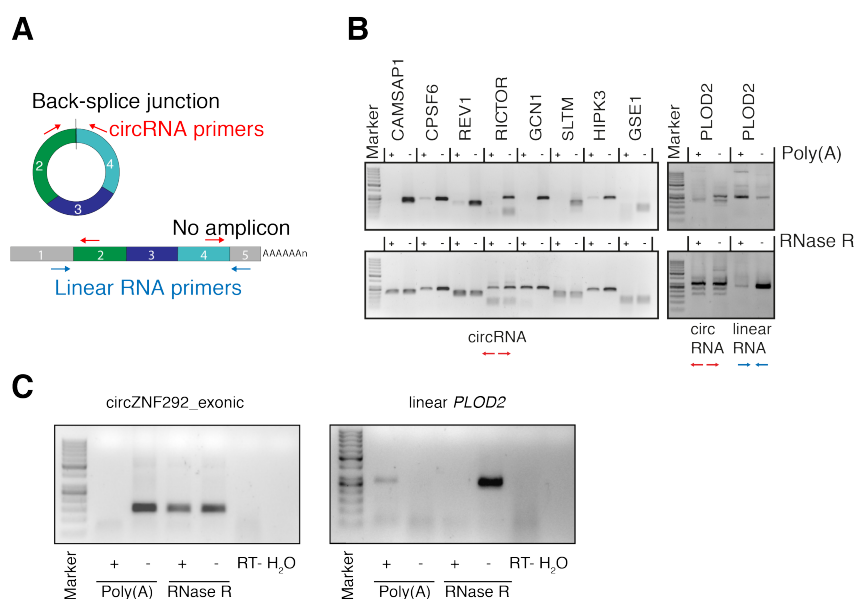


Figure 3.14: Validation of abundant circRNAs by RT-PCR. (A) Scheme depicting how we designed divergent primers to amplify only the circRNA but not the linear counterpart. (B) Gel showing PCR products obtained using pairs of divergent primers to amplify circRNAs upon selection of non-polyadenylated RNAs or RNase R treatment. Primers to amplify the linear *PLOD2* were used as a control. (C) Validation of the circularity of the circRNA generated from exons 2 to 5 of *ZNF292* gene. Similarly to (B), primers to amplify the linear *PLOD2* were used as a control.

In summary, our combined pipeline yielded a comprehensive catalogue of twelve thousand circRNA candidates in three human cancer cell lines under normoxic or hypoxic conditions.

3.4.1 Genomic context of circRNAs

To further characterise our candidate circRNAs, we investigated their genomic origin. The vast majority of circRNAs in our catalogue (95%) originated from protein-coding genes (PCGs). Several circRNAs were also detected in intergenic (2%) and other genomic regions, indicating the heterogeneity of this class of ncRNAs (Figure 3.15A). A deeper investigation of circRNA-producing genes revealed that 4252 (21%) of all PCGs annotated in GENCODE (version 24, $n = 19940$) hosted at least one back-splice event across the cancer cell lines under investigation. CircRNA-producing PCGs were significantly longer than average PCGs, with a median length of 102 kb compared to 30 kb (P -value $< 2.2e-16$, Wilcoxon rank sum test, Figure 3.15B). Consistently, they contained on average 21 non-overlapping exons, more than the average 12 exons composing annotated PCGs. Of note, in order to mimic circRNA host genes, in this comparison we considered only annotated PCGs containing at least three exons. This suggests that the longest genes, composed of a higher number of exons, are more likely to undergo back-splicing. If not coupled with RNase R-treated or poly(A)-selected sequencing libraries, rRNA-depleted RNA-Seq alone is not sufficient to infer the internal structure of circRNAs in terms of exon and intron content. Despite this limitation, from a conservative perspective, when all exons annotated between the back-splice sites were assumed to be part of the circRNA, circRNAs contained up to 40 annotated exons, with a median of four exons per circRNA and 4% single-exon circRNAs (Figure 3.15C).

The median distance between back-splice sites for multi-exon circRNAs hosted by PCGs ($n = 9201$) was 9869 bp and the median length of exons undergoing back-splicing was comparable to all annotated exons (125-127 bp compared to 143 bp, Figure 3.15D,E). Single-exon circRNAs hosted by PCGs ($n = 475$) had a median size of 390 bp, resulting to be generated from unusually long exons, even when compared to the sizes of annotated single-exon genes (median length = 188 bp, Figure 3.15D,E). As reported in Zhang *et al.*, 2014, introns flanking the back-splice sites were found to be unusually long (median 6091-7207 bp) when compared to the average annotated introns (median 1684 bp, Figure 3.15F).

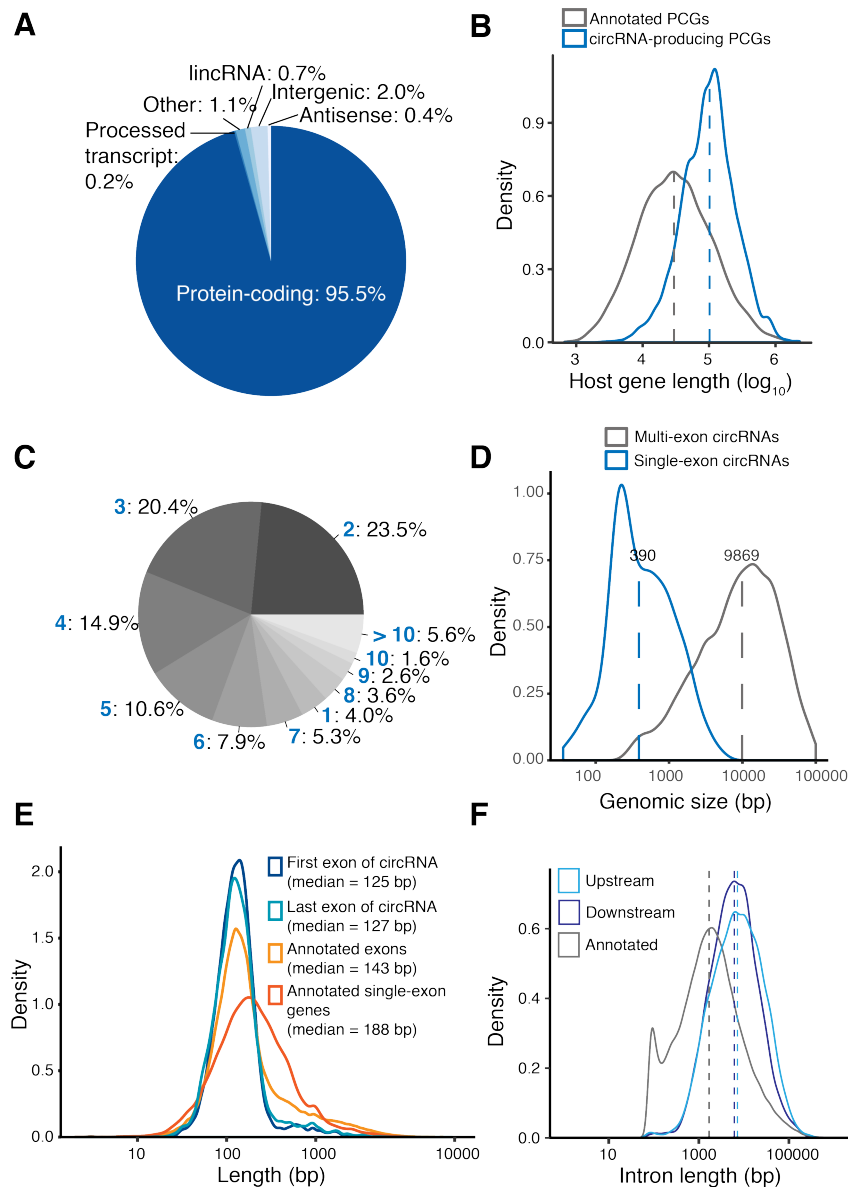


Figure 3.15: CircRNAs are mainly derived from internal exons of protein-coding genes. (A) Pie chart depicting the genomic origin of circRNAs. (B) Distribution of the genomic size of circRNA-producing protein-coding genes (PCGs) genes compared to GENCODE-annotated PCGs. circRNA-producing genes are significantly longer than average PCGs (P -value $< 2.2e-16$; Wilcoxon rank sum test). Dashed lines indicate medians. (C) Pie chart shows the fraction of circRNAs grouped by number of internal annotated exons (blue) within back-splice sites. Most circRNAs contain up to five exons. (D) Density distribution of the genomic size (distance between back-splice sites) of multi- and single-exon circRNAs. Dashed lines indicate medians. (E) Density plot comparing the length distribution of circularised exons (i.e. first and last exons of circRNAs) to all GENCODE-annotated internal exons, distinguishing between exons from multi-exon and single-exon genes. Back-splicing exons are not longer than average exons. (F) Density plot comparing the length of introns flanking back-splice sites to GENCODE-annotated introns. Dashed lines indicate medians.

Along the same line, we investigated the genomic origin of the donor (5'BSS) and acceptor (3'BSS) back-splice sites of circRNAs produced from PCGs, finding that 91% circRNAs had at least one of the back-splice sites residing in the coding sequences (CDS, Figure 3.16A). In the nascent pre-mRNA, back-splicing preferentially occurred at the first genuine 3' splice site, corresponding to the second exon of the pre-mRNA, while no specific exon position was favoured for the donor splice site (Figure 3.16B). Surprisingly, when estimating the splice site strengths of the acceptor and donor back-splice sites with **MaxEntScan** (Yeo & Burge, 2004), we observed that the splice site strength at the donor back-splice site was significantly higher compared to the flanking 5' splice site as well as randomly selected 5' splice sites that do not undergo back-splicing (Figure 3.16C). This suggested that the decision of back- versus linear splicing might depend on the spliceosome assembly at the donor splice site. In contrast, the selection of the second exon for the acceptor splice site was not found to be dependent on the splice site strength.

Although rRNA-depleted RNA-Seq data do not allow to discriminate alternative splicing events occurring within back-splice sites of a circRNA, based exclusively on back-splice junctions, our data revealed that more than half of the host genes undergo alternative back-splicing, producing multiple circRNAs (Figure 3.17A). For instance, the gene *ZNF292* produced four different circRNA isoforms, including hsa_circ_000383, which was previously reported in Boeckel *et al.*, 2015 to be generated from back-splicing of a splice site in the first intron of the gene, and hsa_circ_0004058, which is generated from annotated splice sites. Alternative back-splicing events could generate from 2 up to 29 distinct circRNAs from the gene *BRIP1* and 36 circRNAs from *TRIM37*.

Comparing the relative abundance of circRNA isoforms produced from a certain gene, the majority of circRNA-producing genes produced few predominant circRNAs that exceeded the expected frequency based on equal proportions, often with a strong prevalence of one single circRNA isoform over the others (Figure 3.17B). The distinct circRNAs derived from alternative selection of the acceptor or donor back-splice site in equal measure (Figure 3.17C). For instance, the gene *BARD1* harboured 14 distinct circRNAs, either via alternative 5' back-splicing or alternative 3' back-splicing (Figure 3.17C). As stated above, it is likely that a deeper knowledge of alternative splicing events that occur internally to back-splice junctions would further expand the circRNA repertoire in cancer cells used in this study.

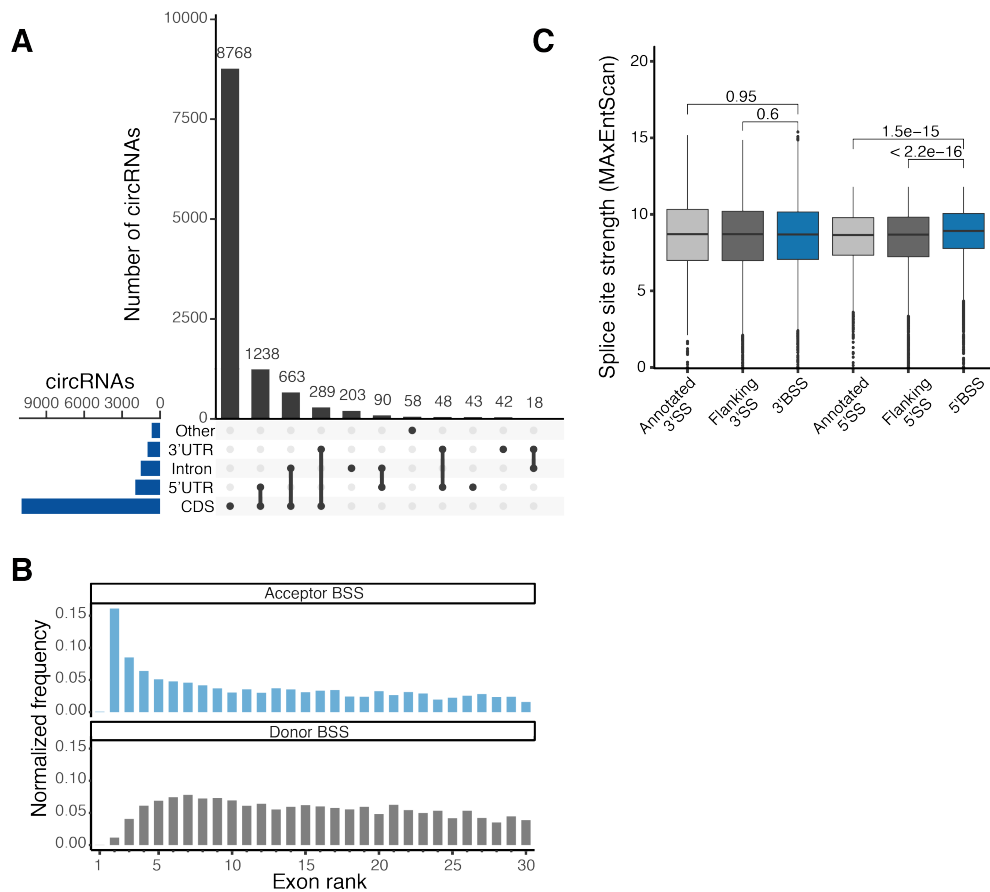


Figure 3.16: Genomic features of back-splice sites. (A) Genomic origin of 3' and 5' back-splice sites of circRNAs produced from protein-coding genes. circRNAs are mainly produced from coding sequences (CDS) of PCGs. (B) Distribution of exon ranks involved in circRNA formation as acceptor (top) or donor splice site (bottom). The normalized frequency was calculated dividing the frequency of a specific exon rank as acceptor/donor splice site by the number of genes containing at least this number of exons + 1 (according to GENCODE version 24 annotation). Only circRNAs produced from PCGs and with both back-splice sites residing in the same annotated transcript were investigated (n = 9676). (C) Strength of 5' back-splice sites (MaxEntScan score) is significantly higher than at flanking 5' linear splice sites and 2000 randomly selected 5' linear splice sites (GENCODE version 24). Same circRNA selection as in (B), further excluding circRNAs involving first/last exons of annotated transcripts (n=9664 circRNAs).

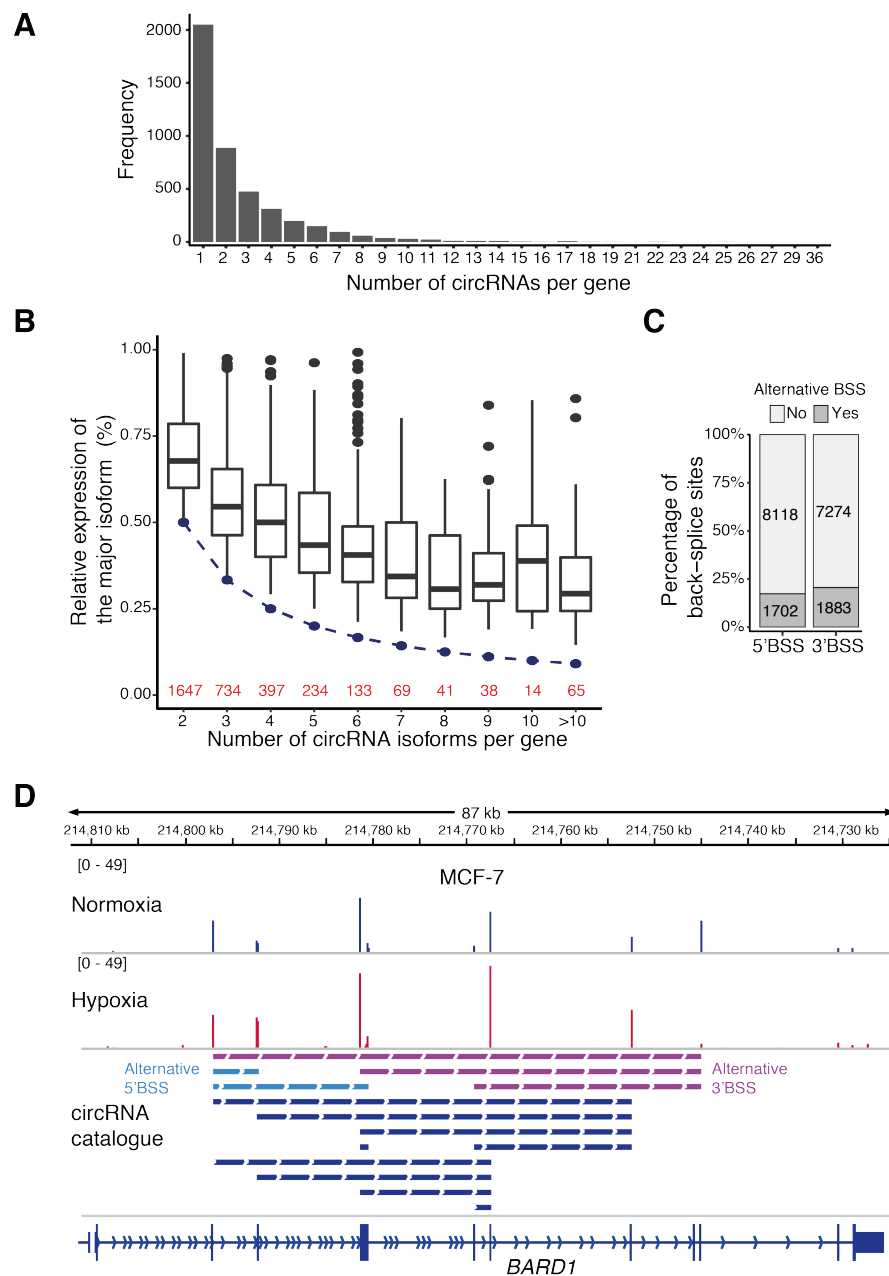


Figure 3.17: Alternative back-splicing produces multiple circRNA isoforms from a single host gene. (A) Bar chart illustrates the number of distinct circRNA isoforms per host gene. The frequency of alternative back-splicing is likely to be underestimated, as this estimate is solely based on back-splice junctions and does not account for possible internal alternative splicing events. (B) Most host genes produce few predominant circRNA isoforms. Boxplot compares the relative abundance of circRNA isoforms produced from a given gene to the expected frequency based on equal proportions. Genes were stratified by the number of associated circRNA isoforms, grouping genes with ≥ 10 circRNA isoforms. Blue line and dots indicate the expected relative abundance, as computed from total circRNA isoforms from a certain gene ($1/\text{number of circRNA isoforms}$). (C) Alternative back-splicing affects more than 20% of back-splice sites. (D) Examples of alternative 3' back-splicing and alternative 5' back-splicing from the *BARD1* gene, which is located on the minus strand. In total, 14 different circRNA isoforms are produced from *BARD1* across A549, HeLa and MCF-7 cells. Genome browser view shows chimeric alignments (back-splice reads) from RNA-Seq of MCF-7 cells under normoxic and hypoxic conditions.

3.4.2 The circRNA profile differs between cancer cells

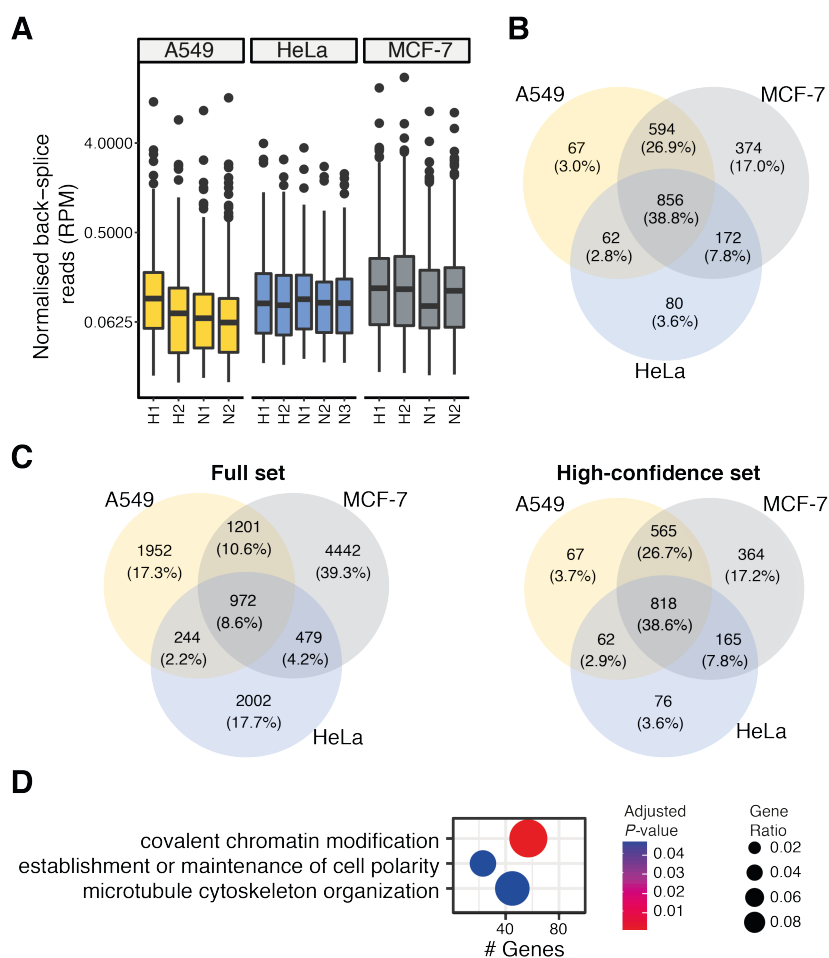


Figure 3.18: CircRNA profiles in human cancer cells. (A) Distribution of normalised back-splice read counts (RPM, reads per million) for the 1012 circRNAs expressed in all three cell lines (Figure 3.13A). MCF-7 cells express circRNAs supported by more reads when compared to A549 and HeLa cells. (B) Venn diagram comparing high-confidence circRNAs (supported by a minimum of five back-splice reads in any two samples) across A549, HeLa and MCF-7 cells. Most circRNAs are expressed in at least two cell types (76%). (C) Venn diagram comparing full set and high-confidence circRNAs, when a further filter on minimum expression of the host gene is applied (TPM ≥ 5 in any single sample of the specific library). (D) Gene Ontology (GO) analysis of 690 host genes harbouring common circRNAs between A549, HeLa and MCF-7 cells.

Previous transcriptome-wide studies reported that the expression of circRNAs is variable among tissues and cell types (Salzman *et al.*, 2013). Indeed, when we compared the 12006 candidate circRNAs across the different cell types, 25% of the total circRNAs were detected in at least two cell types and only 8% of them ($n =$

1012) were expressed in all three cell lines (Figure 3.13A). A deeper investigation on the number of supporting back-splice reads, normalised to reads per million (RPM) to avoid a bias due to the different sequencing depth, revealed that those 1012 circRNAs in common were on average more abundant in MCF-7 compared to A549 and HeLa cells (Figure 3.18A). These results suggest a physiological difference between cell types, with MCF-7 expressing a higher number of circRNAs (Table 3.7), together with more abundant circRNAs when shared with the other cell lines. Since a high percentage of circRNAs were supported by a small number of reads, we further filtered the circRNA catalogue applying a cutoff of five total reads in any two samples, removing the 82% of them and restricting the list to 2205 circRNAs. Throughout this thesis, I will use the terms "full set" and "high-confidence set" to refer to the entire catalogue of 12006 circRNAs and the filtered list of 2205 circRNAs, respectively.

When I performed a similar comparison across cell types but on the high-confidence set, I observed that the percentage of circRNAs shared by at least two cell lines increased to 76%, and 856 (39%) circRNAs were commonly expressed in A549, HeLa and MCF-7 cells (Figure 3.18B). In addition, to understand whether the cell type specificity was attributable to differences in the expression of host genes rather than to genuine back-splicing events, we applied an additional filter on the expression level of host gene (transcript per million, $\text{TPM} \geq 5$). Even increasing the stringency, the proportion of cell type-specific circRNAs was confirmed, both for the full and high-confidence sets (Figure 3.18C). To gain insights into the functional role of circRNAs shared among the three cell lines, I performed Gene Ontology (GO) enrichment analysis of the 690 genes hosting them. This showed the overrepresentation of terms 'covalent chromatin modification', 'establishment or maintenance of cell polarity' and 'microtubule skeleton association' (P -value/ q -value < 0.05), that point to housekeeping roles (Figure 3.18D). Next, we examined the relationship between circRNAs and their respective host gene (mRNA) abundance. We found only a weak correlation (Figure 3.19A,B), both when estimating the gene expression by including or excluding exons between back-splice sites of circRNAs. This indicates that circRNA abundance does not always reflect the host gene expression, with additional factors that might influence their steady-state level, such as circRNA stability or varying efficiency of back-splicing. To verify whether there is a competition between circRNA biogenesis and linear splicing at the same genomic locus, we relied on linear and back-splice junction reads in the RNA-Seq data to calculate

the 'percent circularised' metric, as shown in Figure 3.19C. This metric measures the relative abundance of a circRNA in comparison to all isoforms containing the same exon. In agreement with previous studies, most circRNAs were less abundant than the linear counterpart, with notable exceptions (Figure 3.19D). In total, 210 circRNAs represented the major transcript isoform of their host genes in at least one cell type. Among them, a back-splicing event occurred between exons 3 and 4 of the ataxin 7 (*ATXN7*) gene, yielding a circular transcript that was more abundant than its linear counterpart (Figure 3.19D,E).

Taken together, these results suggested that different regulatory processes might direct the expression of circular and linear transcript isoforms, and that the efficiency of the back-splicing process strongly varies between host genes.

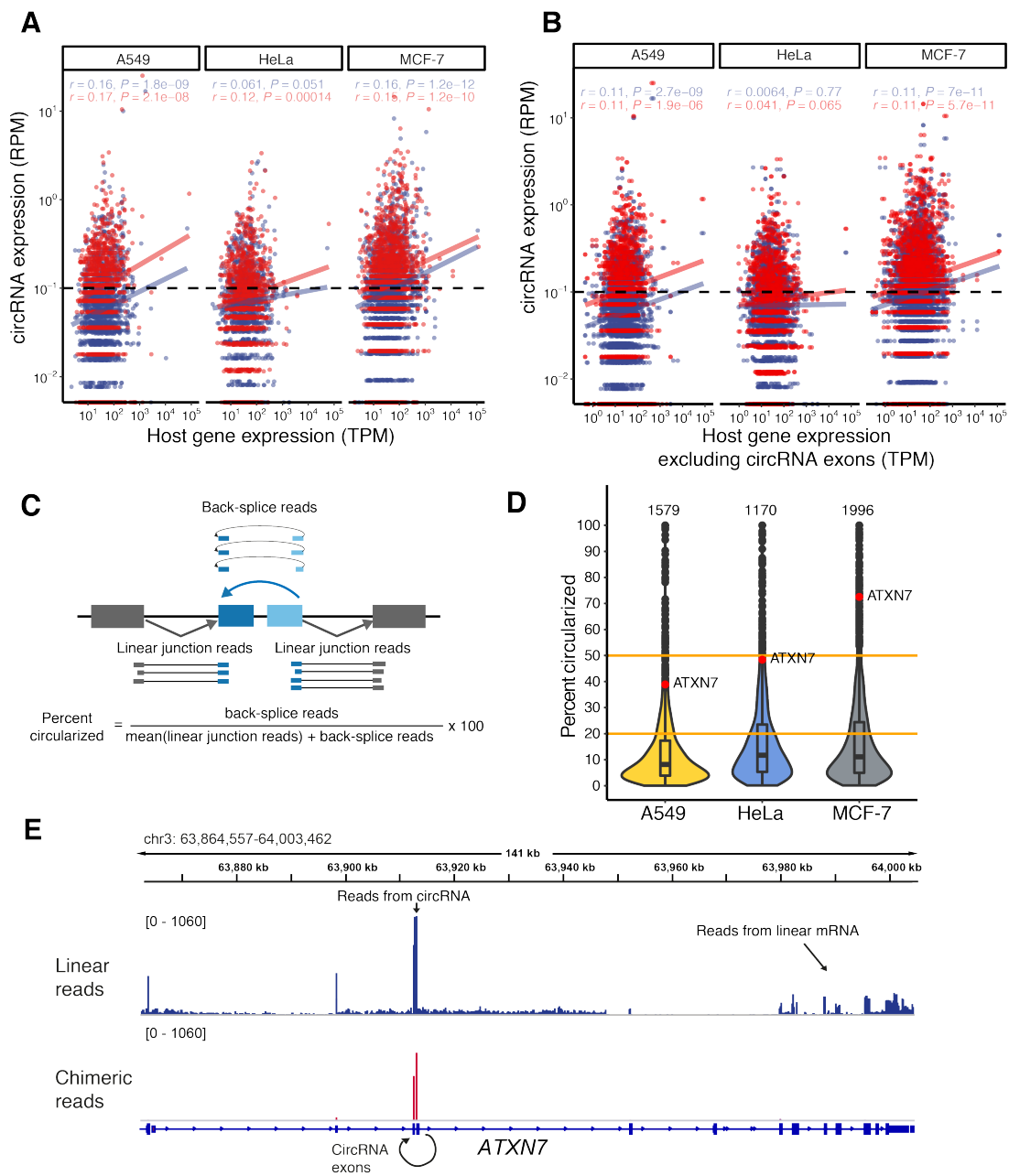


Figure 3.19: Relationship between circRNA and host gene expression. (A,B) Scatter plots compare the expression of circRNAs (back-splice RPM) and their host gene (TPM). For each cell line, the mean expression across replicates is shown, in hypoxic (blue) and normoxic (red) conditions. Linear regression lines, Pearson correlation coefficients and associated P -values are shown. In (B), exons annotated between back-splice sites were excluded for the quantification of the host gene expression. (C) Schematic representation of how the "percent circularised" metric was computed. (D) 210 circRNAs are more abundant than their linear counterparts, as for circATXN7 (hsa_circ_0007761) in MCF-7 cells. Violin plot shows distribution of "percent circularised" values for circRNAs from the three cell lines (mean per cell line across all replicates and conditions). Orange lines indicate 20% and 50% relative abundance of circRNAs. (E) Genome browser view of *ATXN7* gene showing RNA-Seq data from MCF-7 cells under normoxic conditions. Chimeric alignments (bottom) indicate back-splicing of exon 4 to exon 3 to generate circATXN7. The high level of circATXN7 is reflected in a peak in the coverage of linearly mapped reads (top), which corresponds to internal regions of the circRNA, while the other exons of the linear transcript show less coverage.

3.4.3 CircRNA levels change upon hypoxic stress

Recent studies showed that circRNAs change their abundance when the oxygen levels decrease in human endothelial and mesenchymal stem cells (Boeckel *et al.*, 2015; Sun *et al.*, 2017), as well as in mouse lung tissues (Wang *et al.*, 2018). Despite their known alteration upon hypoxia and the relevance of oxygen levels in cancer progression, no study addressing the influence of hypoxia on circRNAs in cancer is available. Our results so far revealed widespread changes of the alternative splicing pattern in the hypoxic cancer cells, together with an extensive production of circular transcripts via back-splicing in the three cell lines.

Next, we investigated whether hypoxia affects back-splicing and circRNA abundance. The overall amount of circRNAs did not change significantly between normoxic or hypoxic conditions (Table 3.7). To identify circRNAs that significantly changed expression upon hypoxia, we used the statistical model implemented in DESeq2 (Love *et al.*, 2014). Back-splice reads represent only a minor proportion of the sequenced reads in rRNA-depleted RNA-Seq data. Thus, for a better estimate of the library size for normalisation and to increase the statistical power, we performed a combined DESeq2 analysis of circular and linear RNAs. In total, 64 circRNAs significantly changed their levels upon hypoxia across the three cell lines, more specifically 6 circRNAs in A549 cells, 22 circRNAs in HeLa and 38 circRNAs in MCF-7 cells (adjusted P -value < 0.1) (Figure 3.20A and Supplementary Table S1). We observed a prevalence of upregulated circRNAs, with only 8 circRNAs being downregulated, all in A549 cells. This is likely due to the high stability that characterises circular transcripts and leads to their accumulation over time, making it difficult to reveal their reduction in steady-state RNA-Seq data. Consistently with the cell type-specific variation of the splicing pattern, the circRNA response to hypoxia was different across cell lines, with only two circRNAs being upregulated both in HeLa and MCF-7 cells, hosted by the *PLOD2* and *ZNF292* genes (Figure 3.20B,C).

In collaboration with Camila de Oliveira Freitas Machado (Müller-McNicoll Group) and Sandra Fischer (Weigand Group) we selected 7-8 of the significantly deregulated circRNA in HeLa and MCF-7 cells for further validation by RT-qPCR. As a control, we also tested additional circRNAs that did not change according to the RNA-Seq

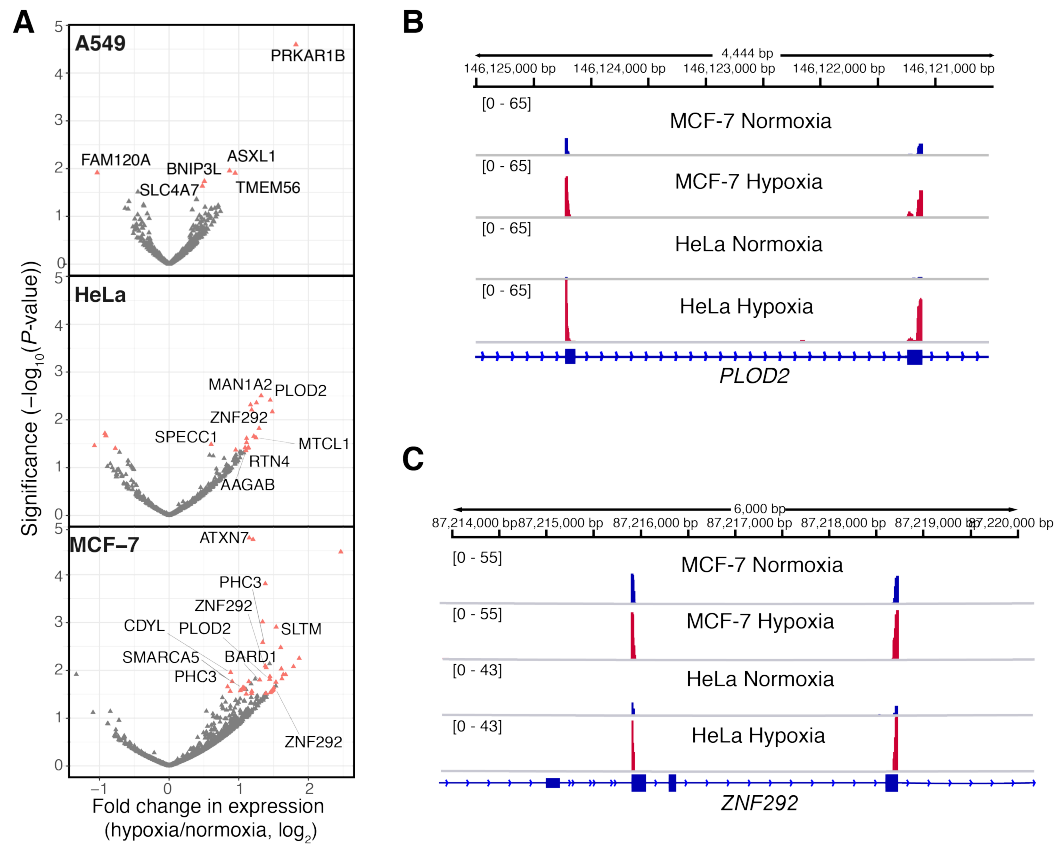


Figure 3.20: Hypoxia induces changes in circRNA levels. (A) 64 circRNAs significantly change their level upon hypoxia. Volcano plots reports \log_2 -transformed moderated fold changes in expression (hypoxia over normoxia) of circRNAs in the three cancer cell lines against associated P -values ($-\log_{10}$). Red: Differentially expressed circRNAs (adjusted P -value < 0.1). Only high-confidence circRNAs with \geq five reads in any two samples of a single cell line were tested for differential expression. (B,C) Genome browser views of exons 2-3 of the *PLOD2* gene (B), and exons 2-5 of the *ZNF292* gene (C), which generate circRNAs that are consistently upregulated under hypoxia in MCF-7 and HeLa cells. Chimeric alignments (back-splice reads) from RNA-Seq data for MCF-7 and HeLa cells under normoxic and hypoxic conditions are shown. *PLOD2* is located on the minus strand.

data analysis. Indeed, RT-qPCR data confirmed the regulation of circMAN1A2, circMTCL1, circRTN4, circPLOD2, circSPECC1 and the exonic circZNF292 in HeLa cells. In addition, the circZNF292 isoform from a cryptic splice site, which was already described in Boeckel *et al.*, 2015 to be hypoxia-regulated in endothelial cells, was significantly changed (Figure 3.21A). The regulation of circATXN7, circPHC3, circPLOD2, circSLTM, circSMARCA5, and the exonic circZNF292 was confirmed by RT-qPCR in MCF-7 cells (Figure 3.21B). Although the circZNF292 isoform from

a cryptic splice site did not reach significance here, it clearly showed a trend to up-regulation in MCF-7 cells.

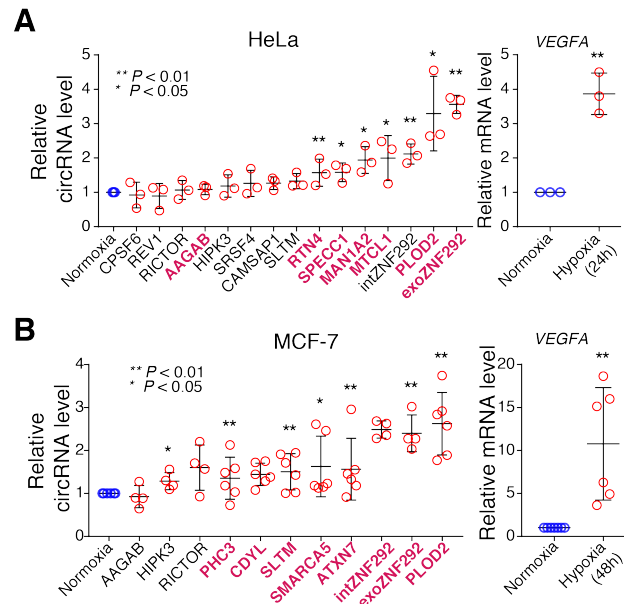


Figure 3.21: Validation of hypoxia-regulated circRNAs by RT-qPCR. (A,B) Expression changes of hypoxia-regulated (magenta) and control (regular) circRNAs in HeLa (A) and MCF-7 (B) cells upon hypoxic treatment for 24 h and 48 h, respectively. The graphics show relative circRNA levels (over normoxia) based on quantitative PCR (RT-qPCR). CircRNA levels were normalised to U6 snRNA for HeLa and *P0* for MCF-7 cells. Mean and standard deviation of the mean are shown. Red circles indicate the single replicate measurements. In HeLa cells, all seven expected circRNAs were significantly upregulated ($n = 3$, * $P < 0.05$, ** $P < 0.01$). In MCF-7 cells, six of the seven expected circRNAs were significantly upregulated ($n \geq 3$, * $P < 0.05$, ** $P < 0.01$), together with circHIPK3 which was not found as significantly regulated in the RNA-Seq data. The circZNF292 isoform from a cryptic splice site, which was already described in Boeckel *et al.*, 2015 to be hypoxia-regulated in endothelial cells, was additionally included in the experiment, although its regulation did not reach significance in our RNA-Seq data analysis. Upregulation of *VEGFA* mRNA was used as control for hypoxia response.

To address the question whether circRNA changes upon hypoxia reflect a regulation of the respective host gene, we compared their levels in RNA-Seq data, finding no global correlation between their expression. Still, the majority of the hypoxia-induced circRNAs originated from upregulated linear mRNA, suggesting a dependence of the circRNA abundance on the general expression of the host gene (Figure 3.22A). However, we found notable exceptions such as circHNRNPM, which was significantly upregulated under hypoxia in MCF-7 cells, while the level of the respective mRNA decreased. Similarly, circBARD1 and circRANBP17 were both upregulated

in MCF-7 cells, with the respective linear mRNAs remaining stable. Similar to the "percent circularised" metric, I estimated a "circular-to-linear ratio" (CLR) value from reads supporting back-splicing events and reads originated from linear splicing at the same splice sites (Figure 3.19C). A comparison between replicates and conditions in the single cell lines revealed that the back-splicing rate remained generally constant between replicates, although it strongly varied between circRNAs (Figure 3.22B). Specifically for hypoxia-regulated circRNAs, their change at low oxygen levels often reflected only little variation of the back-splicing rate (Figure 3.22D, orange points), still supporting the hypothesis of a common mechanism of regulation for circRNAs and mRNAs at transcriptional level.

Recently, Liang *et al.*, 2017a proposed a mechanism by which circRNAs are generated by read-through transcription of the gene located upstream in the genome. Supporting this hypothesis, we observed that back-splicing tends to occur at the first genuine splice site in the pre-mRNA (Figure 3.16B). However, when we compared the expression of circRNAs and their upstream genes both for the high-confidence set and hypoxia-regulated circRNAs, we did not observe any direct correlation that might confirm this mechanism (Figure 3.23).

In summary, from our catalogue of circRNAs in cancer cells, we identified 64 circRNAs that significantly changed in response to hypoxic stress in A549, HeLa and MCF-7 cells, often in parallel to their host gene.

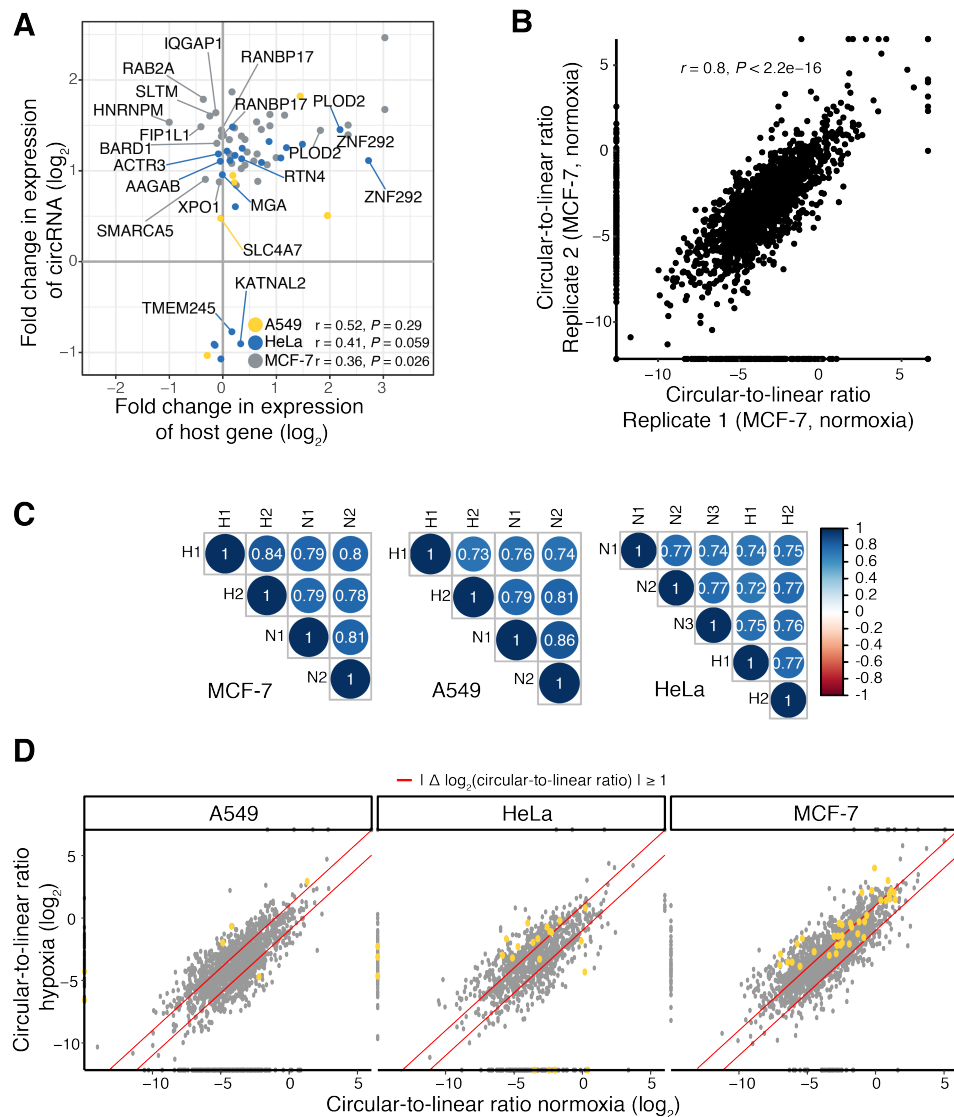


Figure 3.22: Changes of circRNA levels upon hypoxia often reflect variations of their host gene level. (A) Many upregulated circRNAs originate from upregulated host genes. Scatter plot compares \log_{12} -transformed moderated fold changes in expression (hypoxia over normoxia, taken from DESeq2) of 64 hypoxia-regulated circRNAs and their host genes in the three cancer cell lines. For each cell line, Pearson correlation coefficients and associated P -values are reported. (B) Back-splicing rates are consistent between replicates. Scatter plot compares circular-to-linear ratios (CLRs) of all high-confidence circRNAs between two replicate samples with MCF-7 cells under normoxic conditions. Pearson correlation coefficient and associated P -values are given above. (C) Matrix reporting pairwise Pearson correlation coefficients between all samples for each cell line. (D) Most circRNAs do not change in back-splicing rate between conditions. Scatter plots compare circular-to-linear ratios (CLR) of all high-confidence circRNAs in the three cell lines under hypoxic versus normoxic conditions. Red lines mark 2-fold change in CLR between conditions. The hypoxia-regulated circRNAs in each cell line are highlighted in orange.

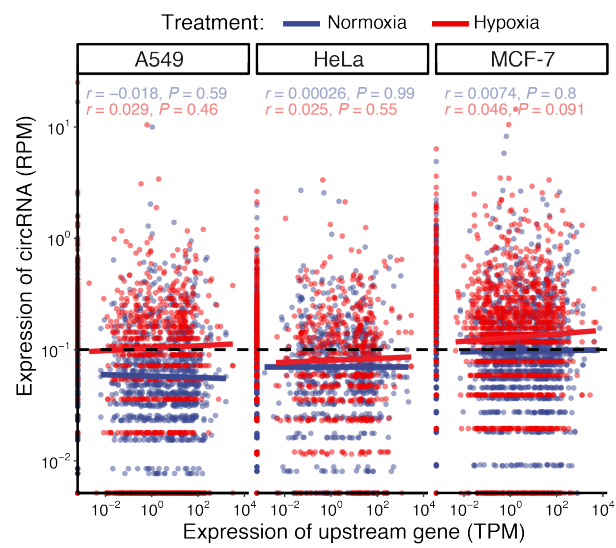


Figure 3.23: Circular RNA biogenesis via read-through transcription of the upstream gene. Scatter plot comparing the expression of circRNAs (in back-splice reads per million, RPM) to the expression of the gene encoded upstream of the circRNA host gene in the genome (in transcripts per million, TPM). Mean expression across replicates is shown for each cell line under hypoxic (blue) and normoxic (red) conditions. Linear regression lines and Pearson correlation coefficients with associated P -values are shown.

3.4.4 Insights into the mechanisms of circRNA biogenesis

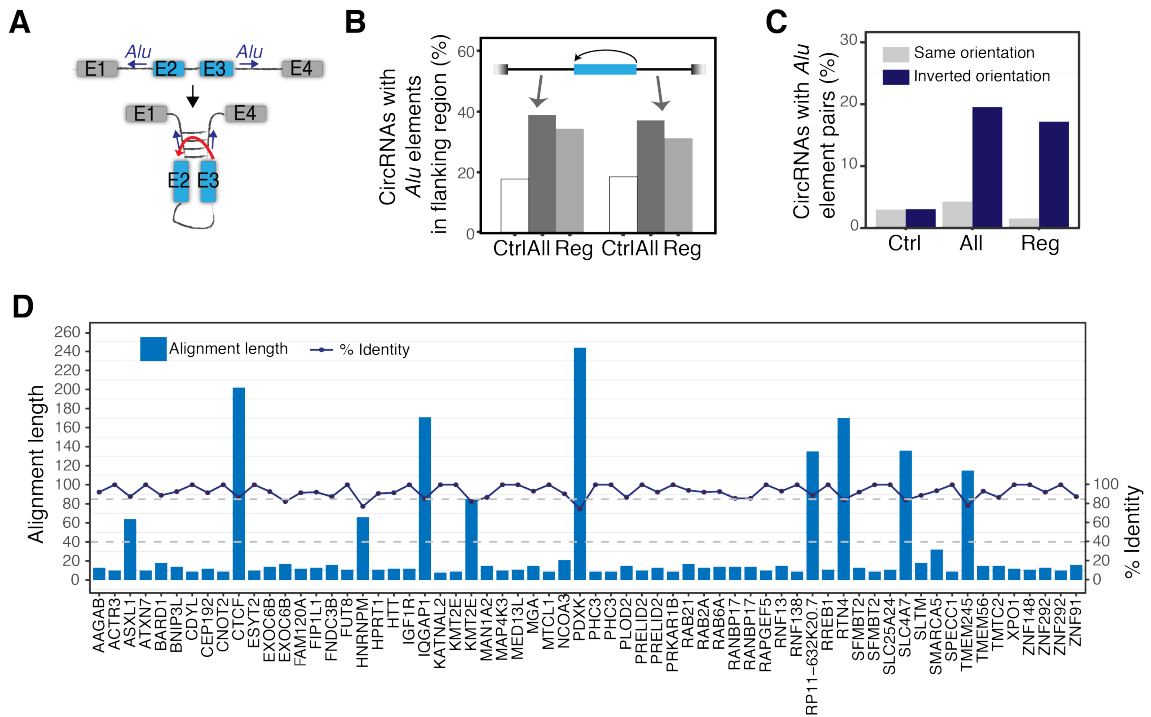


Figure 3.24: CircRNA biogenesis via flanking inverted repeats. (A) Pairing of inverted *Alu* elements in flanking introns can promote circRNA formation (Jeck *et al.*, 2013). (B) CircRNA-flanking introns are enriched in *Alu* elements. Barchart shows the percentage of introns with *Alu* elements in a 500-bp window next to splice sites of annotated internal exons (Ctrl) compared to back-splice sites of all (All) and hypoxia-regulated (Reg) circRNAs. (C) CircRNA-flanking *Alu* element pairs are more frequently in inverted orientation. Barchart shows the percentage of *Alu* element pairs in the same or inverted orientation. (C) Some of the 64 the hypoxia-regulated circRNAs harbour large complementary sequences in their flanking introns. Barchart depicts length and identity of longest local alignment (left and right scale, respectively) from pairwise alignments of flanking introns (500-bp window). The upper and lower dashed lines mark 85% sequence identity and alignment length of 40 nt, respectively. Mutation experiments demonstrated that inverted repeats of 30-40 nt are sufficient to promote back-splicing (Liang & Wilusz, 2014).

The RNA circularisation was reported to be enhanced by the presence of inverted repeats in regions flanking back-spliced exons that bring back-splice sites in close proximity. These repeats often belong to the short interspersed nuclear elements (SINEs), in particular to the *Alu* element family (Figure 3.24A). Indeed, circRNAs in our high-confidence set showed an enrichment of *Alu* elements in the flanking regions when compared to random annotated exons that do not undergo circularisation, in line with the larger size of flanking introns (Figure 3.24B). When *Alu*

elements were present both up- and downstream to circRNAs, they were often in inverted orientation. No significant difference for the hypoxia-regulated circRNAs was observed compared to unchanged circRNAs (χ^2 test, P -value > 0.1 , Figure 3.24C). Pairwise local alignment between flanking regions of the hypoxia-regulated circRNAs confirmed the presence of highly complementary sequences in inverted orientation for 7 out of 64 circRNAs, in all cases reflecting the presence of *Alu* repeats (Figure 3.24D). Altogether, these results support the hypothesis that inverted *Alu* repeats might play a role in circRNA biogenesis, although not for all back-splicing events.

Alu elements are constitutively present in the genome and not all circRNAs harbour *Alu* pairs in their flanking sequences. Thus, other factors are expected to contribute to the regulation of circRNA biogenesis. Exon skipping during pre-mRNA maturation has been proposed as a source of intron lariats, which potentially can undergo back-splicing (Kelly *et al.*, 2015; Khan *et al.*, 2016). A preliminary comparison of back-splicing events to cassette exon events predicted by rMATS in the three cancer cells did not confirm this mechanism. Moreover, we could not detect linear transcripts which skipped the circularised exons for any of the hypoxia-regulated circRNAs. This suggested that the mechanism via intron lariat formation did not play a prominent role in this scenario, or that the skipped transcripts are not stable and get quickly degraded.

Another possible mechanism of circRNA biogenesis involves RNA-binding proteins (RBPs) that localise to circRNA-flanking introns, and, similar to inverted repeats, move back-splice sites close enough to enhance circularisation. Indeed, previous studies have reported different RBPs acting as regulators of the back-splicing process, including known splicing factors like MBNL, QKI and FUS (Ashwal-Fluss *et al.*, 2014; Conn *et al.*, 2015; Errichelli *et al.*, 2017) and SR proteins (Kramer *et al.*, 2015). In order to identify potential RBPs involved in the regulation of circRNAs in cancer cells and upon hypoxia, potential RBP binding sites in the regions up- and downstream to circularised exons were initially predicted with FIMO (Grant *et al.*, 2011). The high-confidence set ($n = 2205$) was investigated, dividing circRNAs in hypoxia-regulated ($n = 64$) and unchanged circRNAs ($n = 2,141$). The *in silico* prediction revealed the potential binding of multiple RBPs, including HNRNPC, HuR (ELAVL1) and PABPC4. This was observed indistinctly for the hypoxia-regulated and the unchanged circRNAs, indicating a general role in the circRNA biogenesis,

not necessarily linked to hypoxia (Figure 3.25A). Each of these RBPs had at least one putative binding site in the region up- and/or downstream to more than 60% of the tested circRNAs. In many cases, binding sites for a specific RBP were predicted in both circRNA flanks (Figure 3.25B), supporting the model of the formation of RBP pairs to enhance circularisation. HNRNPCL1, a paralog of HNRNPC, was also predicted to have binding sites in circRNA flanks. However, a deeper investigation revealed that the HNRNPCL1 binding motif used to scan the nucleotide sequences was almost identical to HNRNPC binding motif (ATTTTTT) making the prediction redundant.

HuR is a ubiquitously expressed RBP known to influence various steps of the post-transcriptional life of the mRNA, in particular its splicing and stability. It binds to uridine tracts in introns and 3'UTRs (Lebedeva *et al.*, 2011; Mukherjee *et al.*, 2011). HuR binding sites were predicted in flanking regions of 84% of the high-confidence circRNAs (Figure 3.25A). This is in good agreement with a recent study in which Abdelmohsen and coauthors reported the binding of HuR to multiple circRNAs in HeLa cells, and proposed a link between circPABPN1, PABPN1 and HuR (Abdelmohsen *et al.*, 2017).

The splicing factor HNRNPC was previously reported to influence the recognition of the 3' splice site (König *et al.*, 2010; Zarnack *et al.*, 2013). To address the question whether HNRNPC might play a role in regulating not only linear but also back-splicing, a meta-analysis of previously published HNRNPC iCLIP data (Zarnack *et al.*, 2013) was performed. This revealed an enrichment of HNRNPC binding to the region immediately upstream to the 3' back-splice site, usually corresponding to the polypyrimidine tract (Figure 3.25C). For instance, HNRNPC showed substantial binding adjacent to the back-splice sites of circSMARCA5 (Figure 3.25D). Notably, the average coverage at the polypyrimidine tract was higher for circRNAs compared to randomly selected linear exons. In agreement with the FIMO prediction, no difference was found between the hypoxia-regulated and the unchanged circRNAs, again suggesting a mechanism generally attributable to the circRNA class. Consistently, the *HNRNPC* gene (mRNA) was only slightly regulated in A549 and HeLa cells (\log_2 -transformed = -0.5 and -0.4, respectively, FDR < 0.05), and stable in MCF-7 cells upon hypoxia, based on RNA-Seq data. The HNRNPC influence on back-splicing was further confirmed by knockdown experiments of *HNRNPC* with two independent siRNAs, followed by RT-qPCR on a panel of 25 circRNAs chosen

independently from the computational analyses. These experiments revealed the deregulation of three circRNA (Figure 3.26A,B,C). Interestingly, both down- (for circCDYL) and upregulation (for circRARS and circSMARCA5) could be observed, in line with the dual role of several splicing factors in enhancing or suppressing splicing.

Altogether, these results confirm the enrichment of complementary repeats in introns directly flanking the circularised exons and introduce HNRNPC as a putative regulator of circRNA biogenesis by binding in close proximity to the 3' back-splice site.

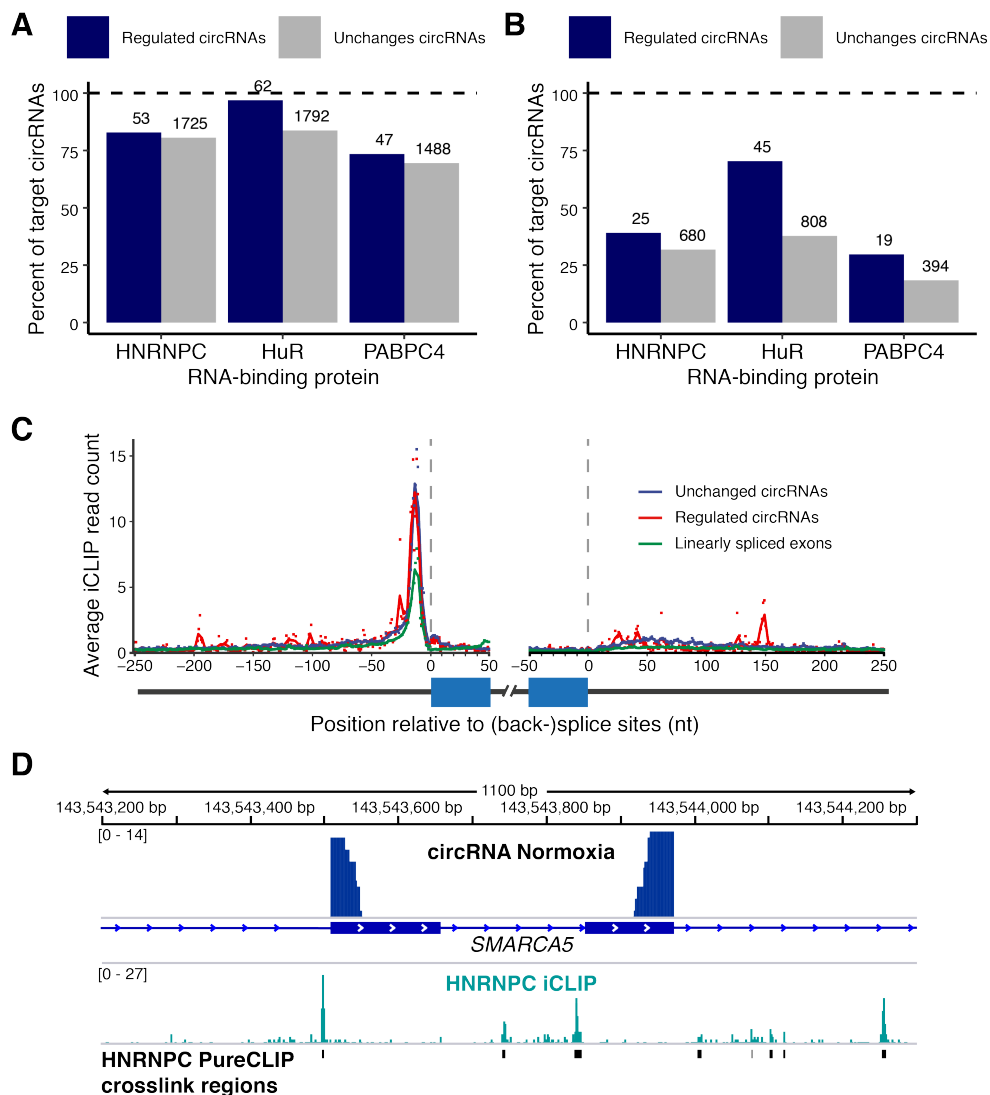


Figure 3.25: RNA-binding proteins as regulators of circRNA formation.

(A,B) *In silico* predictions indicate individual RBP binding sites (A) and flanking binding site pairs (B) at a large fraction of the hypoxia-regulated and unchanged circRNAs (high-confidence set) for HNRNPC, HuR and PABPC4. Barchart shows the number of circRNAs with predicted binding sites of a given RBP. Dashed lines indicate 100%. Above each bar, the number of circRNAs in each category is reported. Predictions for HNRNPC1 were removed since the motif is almost identical to HNRNPC. (C) HNRNPC shows more binding at back-splice sites compared to linearly spliced exons. Metaprofile of HNRNPC binding from iCLIP data in a 300-nt window around back-splice sites (250 nt intron and 50 nt into the circularised exons). High-confidence circRNAs expressed in HeLa ($n = 1133$) were divided into hypoxia-regulated and non-regulated circRNAs and compared to linear exons from expressed PCGs that do not undergo circularisation ($n = 4853$). For each position, dots indicate the mean coverage in each set. Lines were smoothed with locally weighted polynomial regression (loess, span = 0.05). (D) HNRNPC binds upstream of the 3' back-splice site of circSMARCA5. Genome browser view of the *SMARCA5* gene, including RNA-Seq data (chimeric alignments) from HeLa cells in normoxic conditions and HNRNPC iCLIP data from HeLa cells. Binding sites predicted with PureCLIP are shown in black.

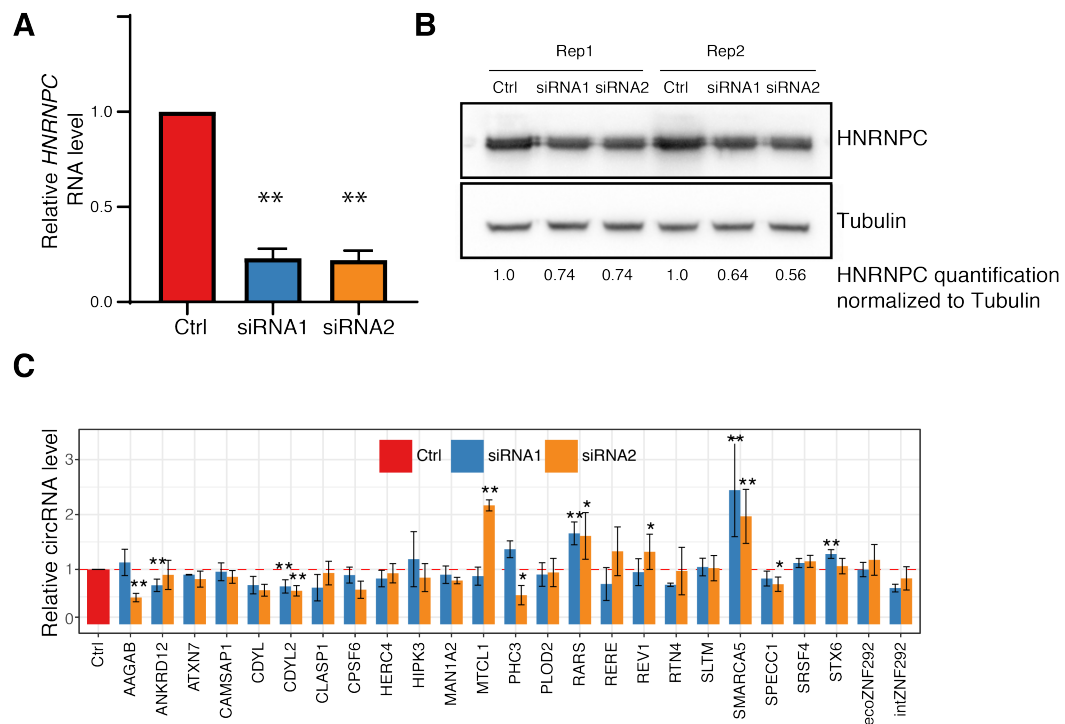


Figure 3.26: *HNRNPC* depletion affects circRNA levels. (A) RT-qPCR shows efficient depletion of *HNRNPC* RNA in HeLa cells upon knockdown with two independent siRNAs (siRNA1 and siRNA2). *HNRNPC* levels were normalised to U6 snRNA and related to control levels (Ctrl) with unspecific siRNA (n = 3, ** $P < 0.01$). (B) *HNRNPC* is modestly downregulated at the protein level, as revealed by Western blot experiment. (C) RT-qPCR estimated expression changes of a panel of 25 circRNAs upon *HNRNPC* depletion with two independent siRNAs in HeLa cells. Barchart shows the relative circRNA levels normalised to U6 snRNA. circCDYL2, circRARS and circSMARCA5 were significantly regulated upon *HNRNPC* depletion (n = 3, * $P < 0.05$, ** $P < 0.01$). Data are shown as mean \pm SD. Performed by Camila de Oliveira Freitas Machado.

3.4.5 CircRNA as miRNA sponges

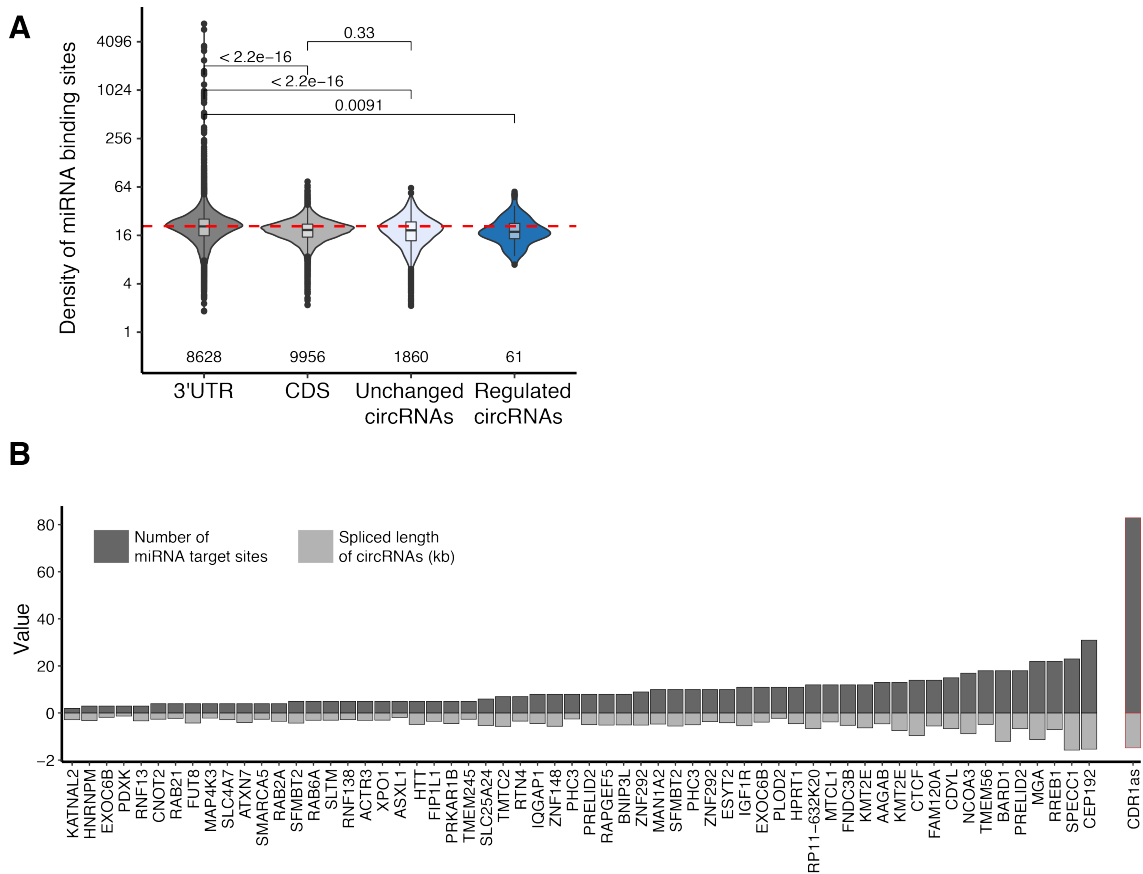


Figure 3.27: *In silico* prediction of miRNA binding sites on circRNA sequences. (A) CircRNAs are not enriched for miRNA target sites. Violin plot shows distribution of predicted miRNA target sites per region for circRNAs from the high-confidence set with assigned parental transcript, as well as 10000 randomly selected coding sequences (CDS) and 3'UTR sequences (GENCODE version 24). *P*-values were computed with two-sample Wilcoxon test. (B) Compared to CDR1as/ciRS-7, which is known to function as miRNA sponge, none of the hypoxia-regulated circRNAs harbours an excess of miRNA target sites. Barchart contrasts the number of predicted miRNA target sites to the predicted circRNA length.

Despite the widespread detection on back-splicing events in mammals, the function of the majority of circRNAs remains unclear. Among the most characterised functions of circRNAs, they have been shown to act as competing endogenous RNAs (ceRNA) presenting multiple binding sites for the same or distinct miRNAs (Hansen *et al.*, 2013; Memczak *et al.*, 2013; Wang *et al.*, 2016; Zheng *et al.*, 2016).

In order to verify whether circRNAs in our catalogue also exert this function, an *in silico* prediction of miRNA binding sites on circRNA sequences was performed

using the miRanda algorithm (Enright *et al.*, 2003). The density of miRNA binding sites on circRNA sequences from the high-confidence set was compared to randomly selected annotated 3'UTRs as well as CDS regions, revealing that circRNAs are not generally enriched for miRNA binding sites, similar to CDS regions (Figure 3.27A). Despite this, 724 circRNAs exceeded the median density of 3'UTRs (20.7 miRNA binding sites/kb) with 34 of them harbouring more than 50 potential miRNA binding sites, including the well-studied CDR1as/ciRS-7. This number might still be overestimated by the prediction of overlapping binding sites for distinct miRNAs. The number of putative miRNA binding sites on hypoxia-regulated circRNAs generally reflected the spliced length of the circRNA and never reached a level comparable to CDR1as/ciRS-7 (Figure 3.27B).

Although no evidence of miRNA sponge function among the hypoxia-regulated circRNAs was found, miRNA binding sites on several hypoxia-regulated circRNAs showed almost perfect complementarity and free-energy lower than -25 Kcal/mol (Figure 3.28), similar to hsa-miR-671-5p binding site on CDR1as sequence. miR-671-5p known to trigger the cleavage of CDR1as in an Ago2-dependent manner (Hansen *et al.*, 2011). For instance, circPLOC2 harboured a binding site for hsa-miR-197-3p with more than 90% identity, suggesting that the circRNA may still be a target of the specific miRNA rather than sequestering it. Taken together, this analysis indicated that a function as miRNA sponge cannot be attributed to the majority of circRNAs.

```

hsa-miR-671-5p vs chrX:140783175-140784659(+)|CDR1as|hsa_circ_0001946
Score: 193.000000 Q:2 to 22 R:1286 to 1308 Align Len (20) (90.00%) (95.00%)

Query: 3' gaGUCGGGGAGGUCCGAAGGa 5'
      |||||:|||||||
Ref: 5' ttCCAGCATCTCCAGGGCTTCCa 3'

Energy: -43.799999 kCal/Mol
-----
hsa-miR-197-3p vs chr3:146121112-146124229(-)|PLOC2|hsa_circ_0122319
Score: 166.000000 Q:2 to 17 R:108 to 128 Align Len (15) (93.33%) (93.33%)

Query: 3' cgaccACCUCUCCACCACUu 5'
      ||||| |||||
Ref: 5' gaagaaTGGAG-AGGTGGTgAT 3'

Energy: -25.799999 kCal/Mol
-----
hsa-miR-340-3p vs chr18:8718424-8720496(+)|MTCL1|hsa_circ_0000825
Score: 177.000000 Q:2 to 18 R:158 to 179 Align Len (16) (87.50%) (100.00%)

Query: 3' cgauaUUUCAUUGACUCUGCCu 5'
      |||||:|||||||
Ref: 5' gcctgAAAGTGGCTGAGACGg 3'

Energy: -31.770000 kCal/Mol
-----
hsa-miR-7-5p vs chrX:140783175-140784659(+)|CDR1as|hsa_circ_0001946
Score: 174.000000 Q:2 to 20 R:208 to 231 Align Len (19) (78.95%) (94.74%)

Query: 3' uguuGUU-UUAGUAUCAGAAGGu 5'
      ||: |||: ||: |||||
Ref: 5' cttcCAGCAATTACTGGTCTTCCa 3'

Energy: -28.270000 kCal/Mol
-----
hsa-miR-616-3p vs chr20:32366384-32369123(+)|ASXL1|hsa_circ_0001136
Score: 175.000000 Q:2 to 20 R:12 to 33 Align Len (18) (77.78%) (94.44%)

Query: 3' gacGAGUUUGGAGGUUACUGa 5'
      |||: ||: |||||
Ref: 5' ctactCGGATGCTCCAATGACa 3'

Energy: -28.480000 kCal/Mol
-----
hsa-miR-1249-5p vs chr3:172247533-172251541(+)|FNDC3B|hsa_circ_0006156
Score: 173.000000 Q:2 to 18 R:88 to 111 Align Len (16) (87.50%) (93.75%)

Query: 3' uugaaccGGUGAGAGGAGGAGGa 5'
      |||||: |||||
Ref: 5' tctcaaCCCATCATCTCCCTCCc 3'

Energy: -32.310001 kCal/Mol
-----
hsa-miR-510-5p vs chr3:172247533-172251541(+)|FNDC3B|hsa_circ_0006156
Score: 167.000000 Q:2 to 16 R:38 to 59 Align Len (14) (92.86%) (92.86%)

Query: 3' cacuaacGGUGAGAGGACUCAu 5'
      ||| |||||
Ref: 5' tcacaccCCAGTCTCCTGAGTg 3'

Energy: -26.650000 kCal/Mol
-----
hsa-miR-1908-3p vs chr3:63912588-63913225(+)|ATXN7|hsa_circ_0007761
Score: 173.000000 Q:2 to 20 R:69 to 88 Align Len (18) (88.89%) (88.89%)

Query: 3' gcCCCGCCUCGGCCGCGGc 5'
      ||||| || |||||
Ref: 5' gcGGGCGGAG-CAGCGGCGc 3'

Energy: -40.070000 kCal/Mol

```

Figure 3.28: CircRNAs as potential miRNA targets. Examples of circRNA harbouring miRNA binding sites with high identity and free energy < -25 kcal/mol.

Chapter 4

Discussion

Hypoxic regions are typically generated during the growth of solid tumours and hypoxia is strongly associated to cancer progression, formation of metastases and poor patient outcome. Hypoxic cells are resistant to available therapies against cancer and a better understanding of the hypoxia response in cancer would contribute to the development of more efficient cancer therapies and the identification of biomarkers. Recently, some studies shed light on the contribution of post-transcriptional regulation to hypoxia adaptation, in addition to the well-studied transcriptional regulation by hypoxia-inducible factors (Sena *et al.*, 2014; Han *et al.*, 2017c; Brady *et al.*, 2017; Bowler *et al.*, 2018). Finally, few studies suggested that back-splicing, which leads to the formation of circular RNAs, can also be affected by hypoxia (Boeckel *et al.*, 2015; Liang *et al.*, 2017b), although the impact of hypoxia on the circRNA repertoire in cancer cells remains to be fully investigated. In this study, I comprehensively characterised the RNA response to hypoxia, exploiting deep RNA-Sequencing in human cell lines from lung, cervical and breast cancers. The hypoxic adaptation occurred at different layers, including RNA expression, splicing and back-splicing. In particular, this study revealed: i) strong variations of gene expression and alternative splicing patterns in cancer cells under hypoxic stress; ii) the induction of the splicing factor MBNL2 upon hypoxia, which contributed to the hypoxia adaptation modulating both splicing and transcript abundance; and iii) the expression of thousands of circRNAs in cancer cells, including several circRNAs that changed their levels in response to hypoxia.

Hypoxia alters gene expression and alternative splicing Earlier studies on hypoxia adaptation mainly focused on the profiling of transcriptional changes induced by the HIF proteins, causing widespread alterations of the transcriptome at low oxygen (Mole *et al.*, 2009; Harris *et al.*, 2015; Schito & Semenza, 2016). Here, I confirm the global expression changes in hypoxia, detecting more than 10000 genes with altered mRNA levels across three different types of human cancer cell lines. My results point to a highly consistent response to hypoxia across different cell types, with many of the shared upregulated genes belonging to the HIF signalling pathway, including *CA9*, *VEGFA*, *GAPDH* and *PLOD2*. In addition, our data show that more than half of the hypoxia-regulated genes are downregulated. Functional characterisation of the differentially expressed genes confirmed the activation of angiogenesis and cell migration, two fundamental processes for the tumour growth (Liao & Johnson, 2007; Lv *et al.*, 2017), as well as the metabolic response to hypoxia. This consisted of an increased rate of glycolysis to support the high energy demand of tumour cells, coupled to an impaired mitochondrial gene expression (Al Tameemi *et al.*, 2019). In parallel, ribosome biogenesis and tRNA processing, necessary for translation (Liu *et al.*, 2006; Uniacke *et al.*, 2012; Chee *et al.*, 2019), as well as DNA replication (Young *et al.*, 1988), were inhibited. These are all energy-consuming processes. Our findings are in line with a recent study comparing 16 datasets of breast cancer cell lines under hypoxic stress, which showed that hypoxia-downregulated genes are involved in processes such as ribosome biogenesis, mitochondrial translation, as well as RNA splicing (Abu-Jamous *et al.*, 2017).

Regulation at the level of alternative splicing is emerging as a key feature of cancer, leading to cell proliferation, survival, migration and metastasis (El Marabti & Younis, 2018; Song *et al.*, 2017). Recent studies showed that a large number of transcripts are differentially spliced upon hypoxic stress in cancer cells (Sena *et al.*, 2014; Brady *et al.*, 2017; Han *et al.*, 2017c; Bowler *et al.*, 2018). In this study, I expanded the set of hypoxia-regulated alternative splicing events, finding 9701 differential alternative splicing events across three cancer cell types in hypoxia. When comparing AS events, I found only slight overlap between cell types. Similarly, Han and coauthors found that splicing events validated in previous studies in different cell lines did not occur in their experiments (Han *et al.*, 2017c). Moreover, differential alternative splicing affected a large portion of genes that did not change their global mRNA level. Thus, AS adds a complementary regulatory layer of complexity to the hypoxic adaptation, generating splice isoforms which encode proteins with

distinct functions. Altogether, these findings indicate that hypoxia-regulated alternative splicing is mostly restricted to a specific cell type, thus conferring specificity to the hypoxia adaptation. Transcripts alternatively spliced in all three cancer cell lines in the same direction are likely to play a critical role. In contrast to a previous study (Han *et al.*, 2017c), I identified cassette exons as the prevalent differential AS event in hypoxia, making up 62% of all AS events, while intron retention represented only the 5%. Despite this discrepancy, intron retention was enriched over the total splicing events measured, together with cassette exons and mutually exclusive exons. It has to be noted that differences in the prevalence of one type of AS events might arise from the diverse algorithms used for the differential splicing analysis. They might be more or less sensitive towards a specific event type, for instance depending on whether a genome annotation is used as a reference to find splicing junctions. On the other hand, my data are in line with Sena *et al.*, 2014 and Brady *et al.*, 2017, that similarly found cassette exons as the most abundant event. In summary, my transcriptome-wide analysis revealed a shared regulation of RNA levels across different cell types upon hypoxia, coupled to cell type-specific changes in alternative splicing.

MBNL2 is induced in hypoxia Alternative splicing is finely modulated by a combination of splicing factors and their regulators, hence variations in the alternative splicing pattern in hypoxia are likely to derive from changes in the expression and activity of these *trans*-acting factors. In addition to the widespread changes in alternative splicing, my data showed that mRNA processing and RNA splicing are biological processes downregulated in the hypoxia adaptation. Previous studies have shown a general decrease of splicing-related proteins in cancer cells in response to stress, including chemotherapy, radiation and hypoxia (Anufrieva *et al.*, 2018). Moreover, SR proteins were reported with altered activity at low oxygen in prostate cancer cells (Bowler *et al.*, 2018). Despite a general downregulation of splicing factors such as SR proteins, our study revealed a consistent increase of MBNL2 in distinct cancer cell lines in response to hypoxia. Only recently, MBNL2 has been associated to cancer progression. For instance, it has been shown that *MBNL2* can function as tumour suppressor gene in hepatocarcinogenesis (Lee *et al.*, 2016). On the other hand, *MBNL2* was also reported to act as oncogene in clear cell renal cell carcinoma (ccRCC) (Perron *et al.*, 2018). Supporting the oncogenic function in ccRCC, *MBNL2* depletion in renal cancer cell lines led to decreased colony-forming ability as

well as activation of the caspase signalling for programmed cell death. Interestingly, the Von Hippel-Lindau (VHL) factor, which mediates HIF α degradation by the proteasome (Gossage *et al.*, 2015), is inactivated in ccRCCs (Meléndez-Rodríguez *et al.*, 2018). Consequently, the HIF pathway remains constitutively active in ccRCCs and might cause the upregulation of *MBNL2*. Moreover, upregulation of *MBNL2* was observed under intermittent and chronic hypoxia in murine breast cancer cells (Chen *et al.*, 2018).

In addition, we observed a physiological impact of *MBNL2* depletion in hypoxic cancer cells treated with the chemotherapeutic cisplatin, in terms of increased cell death and reduced migration (data not shown; Fischer *et al.*, *in revision*), which suggested that *MBNL2* might contribute to the hypoxia adaptation of cancer cells. Altogether, this prompted us to further investigate the role of *MBNL2* in the response to hypoxia in cancer.

In this study, I show that *MBNL2* exerts its function as splicing regulator affecting the alternative splicing pattern of hypoxic cells. In particular, my data showed reversion of many of the hypoxia-dependent AS events upon *MBNL2* knockdown, with a preference for cassette exons, which tended to be skipped in hypoxia and more included in *MBNL2*-depleted hypoxic cells. This indicates a predominant repressive function of *MBNL2* in the splicing response to hypoxia. Our data suggest that the activity of *MBNL2* on splicing in hypoxia is cell type-specific, since only a small fraction of *MBNL2*-regulated events in breast cancer were also shared in lung cancer cells. The mechanism by which MBNL proteins exactly control alternative splicing is still unclear. As for other splicing factors, the positioning of MBNLs with respect to the cassette exons defines whether MBNLs activates or represses splicing (Wang *et al.*, 2012; Charizanis *et al.*, 2012). Based on CLIP experiments in mouse, these studies reported that *Mbnl2* binding upstream or within cassette exons promoted exon skipping, while the binding of *Mbnl2* to the downstream intron led to exon inclusion. It remains to be clarified whether the AS events observed in our data are directed by the different positioning of *MBNL2*. CLIP experiments in human cancer cell lines in normoxic and hypoxic conditions would contribute to assess the binding pattern of *MBNL2* at the alternatively spliced exon loci. Altogether, my findings add *MBNL2* to the factors that are responsible of the global effects of hypoxia on splicing.

In addition to changes in alternative splicing patterns, my study revealed that

MBNL2 reverts hypoxia-dependent changes in transcript abundance. In particular, I found that MBNL2 contributes to the hypoxic adaptation in cancer cells by controlling the mRNA levels of HIF target genes. In contrast to a previous study (Perron *et al.*, 2018), which predicted a role for MBNL2 as mRNA-stabilising factor, my results do not confirm this function for the majority of the predicted MBNL2 targets. Nevertheless, I identified four of the MBNL2 stability targets (*SMAD7*, *CSRNP1*, *SERTAD2*, and *OSMR*), which were induced upon hypoxia and downregulated upon *MBNL2* depletion. Among them, *SMAD7* is often deregulated in tumours, including colorectal, β -cell lymphoma, melanoma, and breast cancer (Slattery *et al.*, 2010; Huse *et al.*, 2012; Javelaud *et al.*, 2007; Salot & Gude, 2013). *SMAD7* negatively controls the transforming growth factor-beta (TGF β)-activated signalling pathway, which contributes to tumour growth, invasion, and formation of metastasis (Luo *et al.*, 2014; Syed, 2016). It has been shown that *SMAD7* is induced under hypoxic stress in a HIF and VHL-dependent manner, thereby activating invasion (Heikkinen *et al.*, 2010). The activation of *SMAD7* might be mediated by the hypoxia-induced MBNL2 via binding to its 3'UTR. Further experiments are required to validate the binding of MBNL2 to the *SMAD7* mRNA and investigate whether this affects its stability and translation.

To my knowledge, while the mRNA-destabilising activity of MBNL1 has been confirmed (Masuda *et al.*, 2012), the stabilising function of MBNL2 has not been experimentally proven yet. My prediction of MBNL2 binding sites in 3'UTRs of hypoxia-regulated genes does not confirm this as a general function in hypoxia. The mechanism by which MBNL2 influences the abundance of hypoxia-regulated genes remains still unclear. One important step to understand this would be to extend the knowledge of the binding preferences of MBNL2. Currently, most information is available only for the paralog MBNL1, for which CLIP data in human have been produced (Fish *et al.*, 2016). MBNL2 recognises a similar motif (Sznajder *et al.*, 2016), but its binding *in vivo* is still likely to be not identical. This was supported experimentally in a minigene binding assays and pulldown experiments of MBNL1 and MBNL2 performed by our collaborators (data not shown; Fischer *et al.*, *in revision*).

In summary, my study revealed a consistent increase of *MBNL2* in the different cell types in response to hypoxia. MBNL2 induction promoted hypoxia adaptation of cancer cells by controlling transcript levels of hypoxia response genes and alternative splicing.

CircRNA profiling in cancer and in response to hypoxia CircRNAs represent a recently re-discovered class of RNA molecules, which have gained increasing attention for their peculiar RNA biology and their high potential as biomarkers in cancer (Kristensen *et al.*, 2017a). For a long time, circRNAs have been overlooked in classical transcriptome profiling studies, which were based on the usage of poly(A)+ RNA-Sequencing. This protocol discards circRNAs, since they typically lack a poly(A) tail. Moreover, due to the large overlap of circRNAs with their linear RNA counterparts, they offer little discriminative sequence information for their reliable detection from RNA-Seq data. Finally, a *de novo* prediction of back-splicing events at cryptic splice sites remains still challenging (Szabo & Salzman, 2016). Starting from 2013 (Memczak *et al.*, 2013), several algorithms have been developed and made available for the detection and quantification of circRNAs from rRNA-depleted RNA-Seq data. Each of these algorithms offers advantages, although this is often associated to a high level of false positives (Szabo & Salzman, 2016). In order to increase accuracy in circRNA detection, it has been suggested to combine outcomes from multiple tools for circRNA prediction (Hansen, 2018).

With the objective of extending our transcriptome-wide analysis in cancer cells to circRNAs, we established a computation pipeline for a reliable detection of circRNAs. Our pipeline is based on two widely used algorithms `find_circ` and `CIRCexplorer` (Memczak *et al.*, 2013; Zhang *et al.*, 2014), which complement each other, as they rely on different algorithms for the alignment of sequencing reads (`Bowtie2` and `STAR`). In addition, previous studies indicated `CIRCexplorer` as one of the outperforming tools for circRNA detection (Hansen *et al.*, 2016; Zeng *et al.*, 2017), although it has the disadvantage that it limits its predictions to exon coordinates from reference annotation. On the other hand, despite the high rate of false positives predicted by `find_circ` because of inaccurate assignment of back-splice junctions, `find_circ` allows *de novo* prediction of back-splicing events independently of prior knowledge of exon annotation. In order to gain most in terms of sensitivity and specificity, our pipeline combines the output from both tools, followed by a rigorous filtering to overcome the tool-specific weaknesses. Finally, with the scope of obtaining consistent quantitative estimates, reads supporting back-splice junctions are recounted for all predicted circRNAs, based on chimeric alignments detected with the splice-aware alignment algorithm `STAR`. We tested the performance of our pipeline in comparison to the usage of `find_circ` and `CIRCexplorer` independently, using RNase R-treated RNA-Seq data as a source of genuine circRNAs.

We show that our pipeline performed better than `find_circ`, in the sense that it retrieved a lower fraction of false positive back-splicing events. Also compared to `CIRCexplorer`, our pipeline performed at least equally, with the advantage of additionally detecting circRNAs from cryptic splice sites. Thus, our pipeline outputs a comprehensive list of accurately quantified circRNAs that can be used for downstream investigations.

Applying this pipeline to our RNA-Seq data, 12006 circRNAs were identified in three human cell lines from cervical, lung, and breast cancer patients. Among them, about one quarter had not been reported before, including some highly abundant circRNAs. We found that a large fraction circRNAs were expressed exclusively in one cell line, indicating a unique circRNA signature for the analysed cancer cells (Xia *et al.*, 2017). This is in agreement with a recent study on 51 breast cancer patients, in which more than 1000 circRNAs were shown to be deregulated in tumours, but not neighbouring tissue (Lü *et al.*, 2017). Although the biological relevance of most circRNAs is still debated and some of the circRNAs detected in the three cancer cell lines may still represent splicing by-products and reflect the expression of their host gene, their abundance could provide important signatures for cancer. Together with their high expression in cancer cells, an increasing number of circRNAs was reported to stimulate oncogenic mechanisms. This is the case of circGFRA1, which was previously found to reach high expression levels in triple-negative breast cancer patients and to be associated with poor prognosis (He *et al.*, 2017). I found that circGFRA1 (hsa_circ_0005239) was highly abundant in MCF-7 cells, although absent from the other two cancer cell lines. These differences were reflected in the expression of the host gene, which was not expressed in HeLa and lowly abundant in A549 (average TPM in normoxia = 5.3), while it was highly expressed in MCF-7 (average TPM in normoxia = 856.6). In addition, it was reported that circHIPK3 and circPIP5K1A are associated with tumour progression and malignancy (Geng *et al.*, 2018). These circRNAs were both expressed in at least one of the cancer cell lines used in our study.

CircRNAs can be highly expressed and they are characterised by a high stability given by their covalently closed structure. Due to these features, circRNAs constitute promising targets for the diagnosis, prognosis, and therapy of cancer (Lü *et al.*, 2017; Kristensen *et al.*, 2017a). In particular, circRNAs may serve as robust indicators of a

hypoxic tumour microenvironment in case of advanced cancer progression. Furthermore, many circRNAs have been shown to be involved in oncological pathways and correlate with poor clinical outcome (Verduci *et al.*, 2019). Recent studies focused on the development of strategies to express proteins from circRNA (Wesselhoeft *et al.*, 2018) and to exploit circRNAs for miRNA sequestration (Jost *et al.*, 2018), with the scope of opening new therapeutic possibilities in the future. In this study, we report that 64 circRNAs significantly changed their level under hypoxic conditions in the studied cancer cell lines (Table S1). This regulation was mainly cell type-specific. However, we found a consistent and robust induction of circZNF292 isoforms. circZNF292 in its intronic variant (hsa_circ_0004383) was found to be upregulated upon hypoxia in endothelial cells (Boeckel *et al.*, 2015) and to correlate with cell proliferation and tube formation in glioblastoma (Yang *et al.*, 2016). The function of the exonic circRNF292 isoform (hsa_circ_0004058) isoform remains still unknown and requires further investigation. Similarly, circPLOD2 (hsa_circ_0122319), which is generated from exons 2-3 of the respective gene, was reported to be abundant in glioblastoma (Song *et al.*, 2016). circPLOD2 was also detected in our study and was consistently regulated in HeLa and MCF-7 cells. Moreover, my analyses revealed a 3-5-fold upregulation of two circPRELID2 isoforms in hypoxic MCF-7 cells, which were generated via alternative selection of the 5' back-splice site. A recent study reported one of these circRNAs (hsa_circ_0006528) being highly abundant in breast cancer cells that are resistant to chemotherapy treatment, suggesting that it might be used as therapeutic target or prognostic indicator for therapy response (Gao *et al.*, 2017). My analyses revealed only six circRNAs as downregulated under hypoxia. This is likely due to their higher stability in respect to linear RNAs, which leads to their accumulation and can cover changes in response to short-term stimuli. However, downregulation can also occur due to specific mechanisms of circRNA degradation, as for CDR1as/ciRS-7, which is sliced by miR-671 in a complex regulatory feedback loops (Hansen *et al.*, 2011; Kleaveland *et al.*, 2018). Moreover, my results revealed the downregulation of a circRNA produced from the *FAM120A* gene (hsa_circ_0001875) in A549 cells. This was in contrast to a recent study reporting an increased expression in A549 cells upon hypoxia. On the other hand, they also reported another isoform of circFAM120A (hsa_circ_0008193) that was downregulated (Cheng *et al.*, 2019). The two isoforms are produced via alternative selection of the 5' back-splice site. This discrepancy might result from the different hypoxic conditions used in their study, with more oxygen (1% O₂) and for a short

time (4 h) (Cheng *et al.*, 2019). Thus, it is possible that acute and chronic hypoxia preferentially influence the expression of one or the other circFAM120A isoforms.

The molecular characteristics of circRNAs investigated in this study suggested an involvement of complementary RNA regions as well as *trans*-acting factors in back-splicing regulation. Indeed, multiple RNA-binding proteins have been previously shown to affect circRNA biogenesis, including MBNL, QKI, FUS, and SR proteins (Ashwal-Fluss *et al.*, 2014; Kramer *et al.*, 2015; Conn *et al.*, 2015; Errichelli *et al.*, 2017). Our data suggest HNRNPC as a novel *trans*-acting factor involved in circRNA formation. With an untargeted approach, we identified three circRNAs, which significantly changed their levels upon *HNRNPC* knockdown in HeLa cells. Although we cannot define whether these changes are linked to changes in transcript abundance, different mechanisms might explain how HNRNPC acts on back-splicing. Previous studies reported that HNRNPC affects U2AF2 binding at genuine and cryptic 3' splice sites, to avoid exon inclusion and preserve splicing fidelity (Zarnack *et al.*, 2013). We could speculate that HNRNPC similarly interferes with back-splicing, thereby affecting circRNA levels. In addition, HNRNPC was previously reported to bind *Alu* retrotransposons in nascent transcripts (Zarnack *et al.*, 2013), and *Alu* elements pairs are frequently present in introns flanking circularised exons and drive RNA circularisation (Chen, 2016). Thus, HNRNPC may also regulate back-splicing via binding to *Alu* elements. This hypothesis is further supported by *HNRNPC* depletion experiments in MCF-7 cells, in which an increased formation of double-stranded RNA regions was observed. These regions were highly enriched in *Alu* elements (Wu *et al.*, 2018). More research is required to clarify whether, similar to linear alternative splicing, the positioning of HNRNPC in respect to back-splice sites may direct the regulation of the different circRNAs. Finally, HNRNPC activity on back-splicing may be influenced by additional regulatory elements.

In summary, we identified and characterised thousands of circRNAs in three human cancer cell lines in normoxic and hypoxic conditions, expanding the knowledge about the expression and regulation of circRNAs in response to hypoxia.

Chapter 5

Conclusions

To conclude, we performed a comparative transcriptome profiling of three human cancer cell lines under hypoxic stress. Our research highlighted that the hypoxic adaptation occurs at different layers, from RNA expression, to splicing and back-splicing. Our results revealed MBNL2 as a novel RNA-binding protein involved in the hypoxia adaptation, together with several circRNAs that respond to low oxygen. In addition, based on our results, we propose HNRNPC as a new regulator of back-splicing. Further studies are required to elucidate the mechanisms by which MBNL2 influences hypoxia adaptation, and to expand the knowledge about the expression and putative functions of circRNAs in human physiology and disease. These findings might have important implications for the development of new biomarkers and therapeutic approaches for cancer in the future.

Supplementary Material

In this appendix, experimental methods used by Camila de Oliveira Freitas Machado for HeLa cells and Sandra Fischer for MCF-7 and A549 cells are described, based on material included in Di Liddo *et al.*, 2019, and Fischer *et al.*, *in revision*.

Cell culture and treatments HeLa cells were cultured in 10 cm-wide plates in high glucose (4.5 g/l) DMEM medium (Sigma Aldrich) supplemented with 10% FBS, 100 U/ml penicillin and 100 µg/ml streptomycin (Pen Strep, Thermo Fisher Scientific). Cells were plated and grown in a normal incubator until they reached 60% confluency (21% O₂, 5% CO₂, 37°C), and then either kept in the normal incubator (normoxic conditions) or transferred to a hypoxia chamber (0.2% O₂, 5% CO₂, 37°C) for 24 h. A549 (DSMZ no. ACC-107) and MCF-7 cells (DSMZ no. ACC-115) were cultured in T75 flasks in DMEM (Sigma-Aldrich) or RPMI-1640 medium (Sigma-Aldrich), respectively, and supplemented with 10% FBS, 1 mM sodium pyruvate and Pen Strep (all from Thermo Fisher Scientific). For hypoxia treatment, 100,000 A549 or 200,000 MCF-7 cells were seeded in 12-well plates. 24 h after seeding, cells were exposed to hypoxia (0.5% O₂, 5% CO₂, 37°C). RNA samples were prepared 48 h later. For *MBNL2* knockdown, 1x10⁵ and 2x10⁵ A549 and MCF-7 cells, respectively, were transfected using Lipofectamine RNAiMAX (Thermo Fisher Scientific). For subsequent RNA isolation, cells were transfected in a 12-well format with 200 pmol of an siRNA targeting *MBNL2* (siMBNL2: 5'-CACCGUAACCGUUUGUAUG[dT][dT]-3') (Paul *et al.*, 2006) or a non-silencing control siRNA (siCTRL: 5'-UUCUCCGAACGUGUCACGU[dT][dT]-3').

RNA preparation and sequencing For RNA-Seq, total RNA was isolated using the miRNeasy Mini kit (Qiagen), including the optional on-column DNA digestion with the RNase-Free DNase Set (Qiagen). After isolation, 500 ng RNA were quality checked on a 1% agarose gel. rRNA was depleted using the RiboZero kit (Zymo).

Libraries were prepared and sequenced on a Illumina NextSeq sequencer with High-Output (75-nt single-end reads) obtaining ca. 100 Mio reads per sample. For HeLa cells, two and three biological replicates were prepared for the hypoxic and normoxic condition, respectively. For A549 and MCF-7 cells, two biological replicates were prepared for each cell line and condition.

RNA preparation and RT-(q)PCR HeLa and MCF-7 cells were cultured and exposed to hypoxia as described above. After hypoxia treatment, cells were harvested and re-suspended in Trizol for RNA extraction, followed by DNase treatment (Turbo DNase, Invitrogen). For validation of circularity, two approaches were taken, based on (i) polyA(+) RNA separation and (ii) RNase R treatment. All validation experiments were performed with a representative sample of HeLa cells under normoxic conditions. PolyA(+) RNA separation was performed using Oligo d(T)25 magnetic beads (New England Biolabs) following the manufacturer's protocol with small modifications. Briefly, 10 μ g total RNA were incubated with 50 μ l beads. The supernatant was collected and saved as polyA(-) fraction. To achieve higher purity, the polyA(-) fraction was incubated a second time with fresh beads, and the protocol repeated. For the bead-bound polyA(+) fractions, protocol and washes were continued as recommended by the manufacturer. After washing, eluted polyA(+) RNA and polyA(-) RNA suspension were precipitated overnight by adding 100% ethanol and 0.3 M sodium acetate. For the RNase R treatment, 10 μ g total RNA were incubated at 37 °C for 40 min, with or without 10 units of RNase R (Epicentre), followed by 3 min incubation at 95 °C for RNase R inactivation. The reaction was performed in 20 μ l. After treatment, 100% ethanol and 3M sodium acetate were added for precipitation. After treatment and precipitation, RNA was recovered and cDNAs were synthesized by RT-PCR using SuperScript III Reverse Transcriptase (Life Technologies), dNTPs and random hexamers (dNTP Mix and Hexanucleotide Mix, Sigma-Aldrich), following the SuperScript III protocol recommended by the manufacturer.

The presence of the circRNAs specifically in the polyA(-) fraction and the RNase R-treated samples was confirmed using convergent primers flanking the back-splice junctions (primers were designed using SnapGene and ordered at Sigma-Aldrich) by semiquantitative PCR. Primers against linear *PLOD2* mRNA were used as control. The PCR reaction was performed with Phusion Polymerase (New England Biolabs),

10 mM dNTPs (dNTP Mix, Sigma-Aldrich), 10 nM forward and reverse primers, and 1:1 DMSO per Phusion volume. After preparing the master mix, 1 μ l cDNA was added to the reaction, and the PCR was performed in the following conditions: 98 °C for 2 min, 34 cycles of 98 °C for 30 s, 55-60 °C (depending on the primer) for 30 s, 72 °C for 30 s, and final extension at 72 °C for 5 min. PCR products were visualised using 2% agarose gel electrophoresis (VWR Maxi or Midi Electrophoresis System). The 2% agarose gels were pre-stained with RedSafe (HiSS Diagnostics).

For validation of differentially expressed circRNAs under hypoxia, RNA was prepared from hypoxic and normoxic HeLa and MCF-7 cells, and reverse transcription was performed from 2 μ g total RNA as described above. Differential expression was validated by quantitative PCR (qPCR) using 1x final concentration of 2X ORA qPCR Green ROX L Mix (highQu GmbH), 500-2000 nM forward and reverse primers (depending on the primer) and 1 μ l of 1:8 dilution of cDNA. Primers targeting U6 for HeLa and P) for MCF-7 cells were included in each experiment, and their quantification cycle number (C_q value) posteriorly used for normalisation. The qPCR was performed in a PikoReal 96 Real-Time PCR System (Thermo Fisher Scientific) using the following program: 95 °C for 2 min, 30 cycles of [95 °C for 20 s, 60 °C for 20 s, 72 °C for 30 s] and final extension at 72 °C for 5 min, followed by a step-wise melting curve (60-95 °C). The same primers were used for semiquantitative PCR and qPCR.

Western blot for validation experiment For western blot analyses, A549 and MCF-7 cells were lysed in lysis buffer (137 mM NaCl, 10% glycerol, 20 mM Tris-HCl pH 8.0, 2 mM EDTA pH 8.0, 1% Igepal, 5 μ l protease inhibitor cocktail [Sigma-Aldrich]) for 20 min on ice. After centrifugation (15 min at 17,000g, 4 °C) the protein content of the samples was determined in three technical replicates according to the Bradford method. 10 μ g protein were loaded onto precast gels and blotted onto PVDF membranes (both Bio-Rad). Primary antibody targeting HNRNPC (sc-32308) was used. Horseradish peroxidase-conjugated anti-rabbit IgG (Jackson ImmunoResearch) was used as secondary antibody. Blots were developed with the ECL system (Bio-Rad). Images were detected using the ChemiDoc Imaging System (Bio-Rad).

Cell viability assay A549 and MCF-7 cells were transfected as described above and incubated under normoxic or hypoxic conditions (48 h, 0.5% O₂). 24 h later, cisplatin was added at final concentrations of 10 or 20 μM. After another 24 h under normoxia/hypoxia the cells were fixed with 0.5% formaldehyde in PBS and stained with 0.5% crystal violet in PBS. After three washing steps with PBS, the cells were incubated with 33% acetic acid. Samples were transferred to a 96-well plate. Absorption at 570 nm was measured in a TECAN infinite M 200 Pro plate reader. Absorption was normalized to normoxic control cells with the respective cisplatin concentration.

***HNRNPC* knockdown** siRNA transfection for *HNRNPC* knockdown was performed using previously described siRNAs (Zarnack *et al.*, 2013): Stealth Select RNAi siRNAs HSS179304 and HSS179305 as well as control siRNA Stealth RNAi siRNA Negative Control. For knockdown experiments, HeLa cells were cultured under normal conditions and seeded into 6 cm dishes 24 h prior to siRNA transfection. A final concentration of 20 nM of each siRNA was transfected into HeLa cells using jetPRIME® DNA and siRNA transfection reagent (VWR) following the manufacturer's protocol. Knockdown was performed for 48 h, and cells were subsequently harvested for RNA extraction.

Supplementary Tables

Table S1: List of hypoxia-regulated circRNAs grouped by cell line. Downregulated circRNAs are coloured in blue.

Genomic coordinates (hg38)	circBase ID	Host gene	Fold change (log ₂)	Adjusted <i>P</i> -value
A549				
chr20:32366384-32369123(+)	hsa_circ_0001136	<i>ASXL1</i>	0.87	0.05605
chr8:26391243-26408376(+)	hsa_circ_0002131	<i>BNIP3L</i>	0.51	0.08313
chr9:93471141-93498886(+)	hsa_circ_0001875	<i>FAM120A</i>	-1.03	0.06043
chr7:579256-607452(-)	hsa_circ_0079040	<i>PRKAR1B</i>	1.82	0.00034
chr3:27437388-27448797(-)	hsa_circ_0006215	<i>SLC4A7</i>	0.48	0.09883
chr1:95143891-95173889(+)	hsa_circ_0005720	<i>TMEM56</i>	0.95	0.06134
HeLa				
chr15:67231814-67236820(-)	hsa_circ_0000620	<i>AAGAB</i>	1.10	0.09562
chr2:113939959-113942359(+)	hsa_circ_0008712	<i>ACTR3</i>	1.19	0.01725
chr18:12999421-13030608(+)	hsa_circ_0107922	<i>CEP192</i>	1.22	0.05285
chr12:70278132-70311017(+)	hsa_circ_0007127	<i>CNOT2</i>	1.09	0.08488
chr16:67610824-67612121(+)	hsa_circ_0002122	<i>CTCF</i>	-0.91	0.05149
chr3:172247533-172251541(+)	hsa_circ_0006156	<i>FNDC3B</i>	1.29	0.03804
chr14:65561337-65561766(+)	hsa_circ_0003028	<i>FUT8</i>	-1.07	0.07800
chrX:134473359-134493590(+)	hsa_circ_0004549	<i>HPRT1</i>	1.48	0.01873
chr15:98707562-98708107(+)	hsa_circ_0005035	<i>IGF1R</i>	1.32	0.00939
chr18:46946057-46946923(+)	hsa_circ_0108513	<i>KATNAL2</i>	-0.90	0.05198
chr1:117402186-117414831(+)	hsa_circ_0000117	<i>MAN1A2</i>	1.17	0.01390
chr2:39331917-39337581(-)	hsa_circ_0054211	<i>MAP4K3</i>	-0.92	0.04698
chr15:41668828-41669958(+)	hsa_circ_0000591	<i>MGA</i>	0.96	0.09318
chr18:8718424-8720496(+)	hsa_circ_0000825	<i>MTCL1</i>	1.25	0.05616
chr3:146121112-146124229(-)	hsa_circ_0122319	<i>PLOD2</i>	1.45	0.01140
chr3:149846011-149872154(+)	hsa_circ_0067716	<i>RNF13</i>	1.11	0.06862
chr6:7176655-7189322(+)	hsa_circ_0001573	<i>RREB1</i>	1.14	0.08457
chr2:54982515-54987698(-)	hsa_circ_0001006	<i>RTN4</i>	1.13	0.08674
chr10:7220411-7276989(-)	hsa_circ_0017627	<i>SFMBT2</i>	1.25	0.01273
chr17:20204333-20205912(+)	hsa_circ_0000745	<i>SPECC1</i>	0.60	0.07391
chr9:109050283-109050692(-)	hsa_circ_0087905	<i>TMEM245</i>	-0.77	0.08711
chr6:87215903-87218731(+)	hsa_circ_0004058	<i>ZNF292</i>	1.11	0.05761
MCF-7				
chr3:63912588-63913225(+)	hsa_circ_0007761	<i>ATXN7</i>	1.15	0.00019
chr2:214767482-214781509(-)	hsa_circ_0001098	<i>BARD1</i>	1.30	0.05875
chr6:4891713-4892379(+)	hsa_circ_0008285	<i>CDYL</i>	0.88	0.04392

Table S1: List of hypoxia-regulated circRNAs grouped by cell line (*continued*)

Genomic coordinates (hg38)	circBase ID	Host gene	Fold change (log ₂)	Adjusted <i>P</i> value
chr7:158788004-158799072(-)	hsa_circ_0001777	<i>ESYT2</i>	1.07	0.08355
chr2:72731007-72733118(-)	hsa_circ_0001030	<i>EXOC6B</i>	1.19	0.09038
chr2:72718103-72733118(-)	hsa_circ_0009043	<i>EXOC6B</i>	1.34	0.00605
chr4:53414615-53428183(+)	hsa_circ_0007476	<i>FIP1L1</i>	1.48	0.08668
chr19:8455405-8463686(+)	hsa_circ_0006382	<i>HNRNPM</i>	1.53	0.06392
chr4:3086939-3107423(+)	hsa_circ_0001392	<i>HTT</i>	1.02	0.08818
chr15:90439332-90443478(+)	hsa_circ_0000651	<i>IQGAP1</i>	1.64	0.04792
chr7:105073619-105077433(+)	hsa_circ_0001736	<i>KMT2E</i>	1.50	0.08476
chr7:105073619-105078963(+)	hsa_circ_0081819	<i>KMT2E</i>	1.62	0.05624
chr12:116230533-116237705(-)	hsa_circ_0000443	<i>MED13L</i>	0.84	0.07626
chr20:47623911-47633636(+)	hsa_circ_0001165	<i>NCOA3</i>	1.21	0.00021
chr21:43746079-43749080(+)	hsa_circ_0008021	<i>PDXK</i>	1.87	0.02553
chr3:170136419-170149244(-)	hsa_circ_0001359	<i>PHC3</i>	1.06	0.07819
chr3:170145423-170149244(-)	hsa_circ_0001360	<i>PHC3</i>	1.34	0.01380
chr3:146121112-146124229(-)	hsa_circ_0122319	<i>PLOD2</i>	1.45	0.05722
chr5:145764931-145826200(-)	hsa_circ_0008132	<i>PRELID2</i>	1.67	0.04809
chr5:145817894-145826200(-)	hsa_circ_0006528	<i>PRELID2</i>	2.47	0.00035
chr12:71769800-71774022(+)	hsa_circ_0099178	<i>RAB21</i>	1.39	0.09774
chr8:60572046-60591969(+)	hsa_circ_0007581	<i>RAB2A</i>	1.79	0.03522
chr11:73707420-73718718(-)	hsa_circ_0000339	<i>RAB6A</i>	1.11	0.09985
chr5:171183199-171205612(+)	hsa_circ_0002713	<i>RANBP17</i>	1.45	0.05161
chr5:171183195-171205612(+)	hsa_circ_0003718	<i>RANBP17</i>	1.38	0.03423
chr7:22291175-22318037(-)	hsa_circ_0001681	<i>RAPGEF5</i>	1.45	0.09440
chr18:32111754-32113860(+)	hsa_circ_0005729	<i>RNF138</i>	1.08	0.08110
chr15:32526813-32533368(-)		<i>RP11-632K20.7</i>	1.47	0.09038
chr10:7276892-7285954(-)	hsa_circ_0000211	<i>SFMBT2</i>	1.54	0.00747
chr1:108148279-108161293(-)	hsa_circ_0004270	<i>SLC25A24</i>	1.38	0.00129
chr15:58912563-58916999(-)	hsa_circ_0000605	<i>SLTM</i>	1.60	0.01687
chr4:143543509-143543972(+)	hsa_circ_0001445	<i>SMARCA5</i>	0.91	0.06248
chr12:82857010-82857580(+)	hsa_circ_0002886	<i>TMTC2</i>	1.61	0.03837
chr2:61522611-61533903(-)	hsa_circ_0001017	<i>XPO1</i>	0.88	0.09038
chr3:125313308-125331238(-)	hsa_circ_0001333	<i>ZNF148</i>	1.14	0.06266
chr6:87210451-87218731(+)	hsa_circ_0004383	<i>ZNF292</i>	1.50	0.07943
chr6:87215903-87218731(+)	hsa_circ_0004058	<i>ZNF292</i>	1.40	0.03662
chr19:23358430-23362725(-)	hsa_circ_0109315	<i>ZNF91</i>	1.18	0.09776

Supplementary Figures

CSNRP1**3' UTR coordinates: chr3:39141855-39143054 (-)**

5' -

GGACCAGGAGUUCUUUCCAGCCCAAGAGACCUGUUGCUGCUUUCUUGUAAUUAUGGGGCUCCAGAGUCUGCGUAACA
 GUCUCCACUGGGCUGGCUACCCACAGGUGCCAUUGGCACACUCCUGGUUUCAAACAUAUCUGGAUUUUUUUUUUUU
 UUAACUUUUCUGUGUGAAGAGAGGACUGGGGGGAGGGGCUUCCUUUACAGUCGCCGGCCCCACACCCACAGCUUUC
 UCUUUAUCUCCACACGUGAGCCUGGAAGAGAGAAAUAUGGGCUCUCUGGAGCUUUGGAGACCACUUUUCGGUUCUUGC
 GUGAUGUUCUUAAGCCAAAGACGGUGAGACAGGGCUGAAUAUAGGUGGCUUCUGCCACUCUGAGCCUAGACCAGUUGGUG
 GCUAAAUCACUGGACUGGAAGACUAUAAUUUAUUUUAUUAUUAUUUUGGAGAUUGAGGAGGCUUUGGUUGCACUUUUU
 GCUUGGUGGGUAAUCCAGGGUGGGGUGGGCACAGGCCUUAAGAGCCCUUUUUGCCUUGUAGUCUACACCUUGCUCUG
UUGGGCUUUGGUGACUAAGGUGGUAUUUGAGCUCUGUUAUCUAGUCUUGGUCUCCUAGAUUGGCUUGUGGGCA
 GGUGCGGCCAAGGACUAUCUAGGCGGGGAGAGCCUGGGUGAACAGCUGUACCAAACUCCUUUGCCACCCU
 GCCCCUCCACUUCUGCCUCUUGUCCAUUCCUUCCUUCCAAAGGCCACAGCCUUUUAUCCAGGCCAGGGAUUGAGGA
 GGGGAAGGAGAAACAGGAAGCCAGAGAGGGCAAAGGCCUACUCGGGGCCGAACCAUGCCAGACUAUUUUCAG
 GGCUUUCUGGGCACUGCACUUCAGCGUGGCCACUGCACAUGCCCUGAGGCCAGUUGGCGAGGGGUGGCUUCUGAGGGUUU
 UUAUACCCUUUGUUGCUAAUGUUAAUUUUGCAUCAUAAUUUUAUUAUUGCCUGAGUGUCAGAAUAUAAUUUAUUC
 UUUCUCUCUGUGCCAAAGACCGACGGCUCUGGGCCUGCUCUUGCCCAGGAGGCCUUGCCAGCCUGUGUGCUUGG
 GGAACACCUUGUACUGAGCUUACAGGUACCAUAAAGAGGCUUUUUUUUA-3'

OSMR**3' UTR coordinates: chr5:38933445-38935641 (+)**

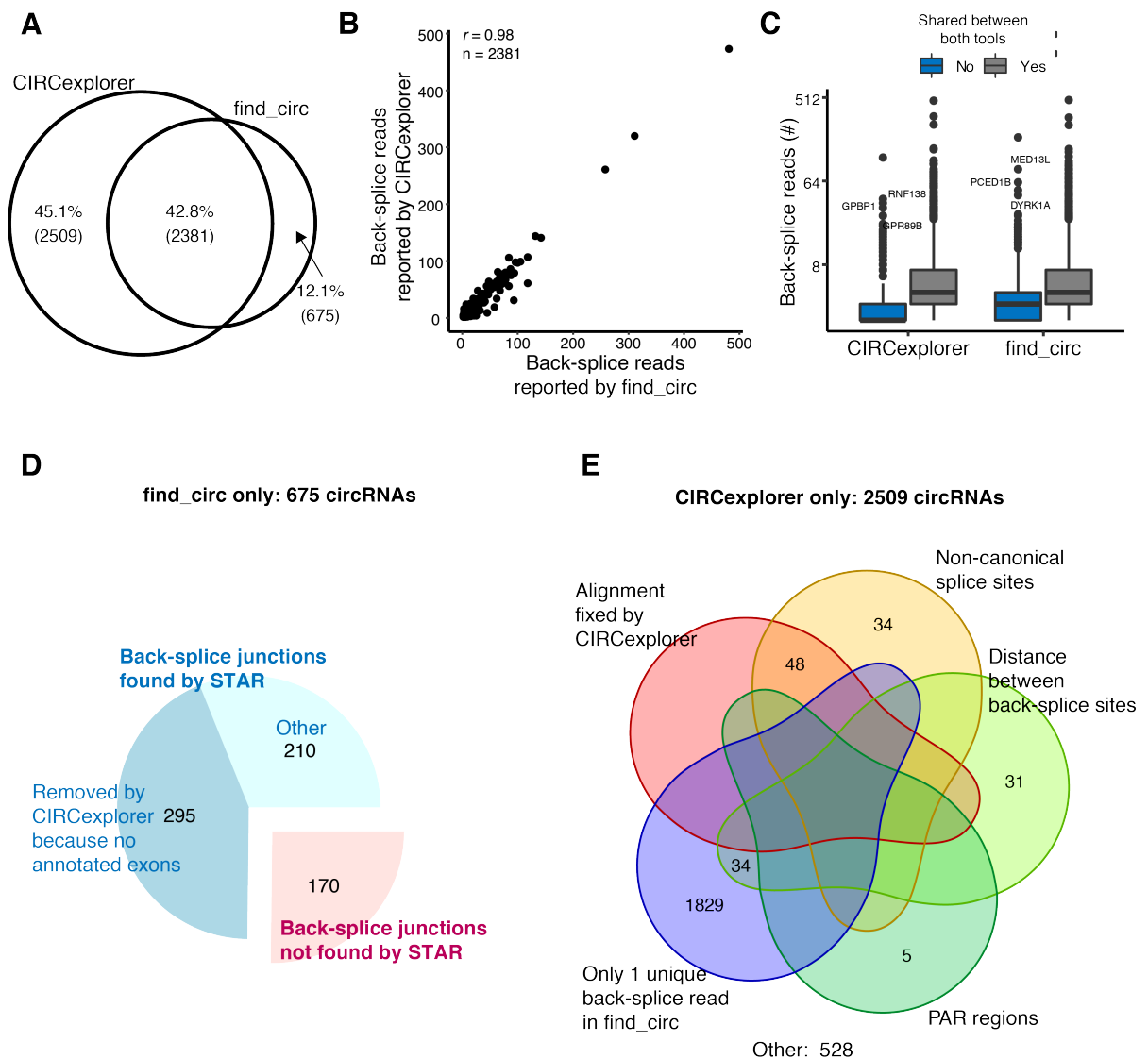
5' -

CCAGCAUGCCGAUUUCAUACC UU AUGCUACAGACAUUAAGAAGAGCAGAGCUGGCACCCUGUCAUCACAGUGGGCUUGG
 UCUUAAUCCAGUACGAUUUGCAGGUCUGGUUUUAUUAAGACCACUACAGUCUGGCUAGGUUAAAGGCCAGAGGCUAUGGA
 ACUUAACACUCCAUUGGAGCAGCUGGCCUUAAGAGCGGACAGAUUUGGAGCAGUUCUUCUGUUGUUUUUUUUUU
 GGUCUACUUUAAGAACAGGAGACUUGAGCUUGACCUAAGGAUAUGCAUUAACACUACAGACUCCACUCAGUACUUAU
 AGGGUGGCUUGUUGCCUAAGAAGUUCAGUUUUUACUGAGGAUAUUUUUCAAUAACAGCAUUUAUAUUAUGAAGGCUUUU
 AAAGGCCACAGGAGACUAUUAUAGCAUAGAUUGUCAAAUGUAAAUAUUAUGAGCGGUUUUAUAAAAAACUCACAGGUG
 UUGAGGCCAAAACAGAUUUAGACUUAUCUUGAAGGUAAGAAUCUUAUAGUUCACUGACACAGUAAAAUUAUCUUGGG
 UGGGGCGGGGGCAUAGCUUAUUCUUAUAUUAUAAAUGUGUAGUAUUAACAAGAUUUCCACAUAUUUCUGUCAAGC
 UACUACAGUGAAAGAUUGGGAUUGGCAAGUAACUUCUGACUUAUCUGUAGUUGUACUUCUGUCCAUAGACAUAGUAUU
 GCCAUCAUUUUUGAUGACUACUCCAGAAUAUAAAAGGAAAGCUUAUACAUAAUUAUCCAGUCACAGUUUUUGGUUCU
 UCUUUCAAGAACUAUUAUAAAUGACCUGUUUACUUAAGCAUCUUGGACUCUGCAUAGGUGUUGCUGGGUCAAAGUAAC
 UUCUAGUCACAUUAUUAUUAUUAUUGUAAAUAUUAUUAUUAUUAUUAUUAUUAUUAUUAUUAUUAUUAUUAUUAU
 AGUUUAUUAAGUUGGAAAUUCUGUUGGCUUGGAGCAGCUUUGUCUCCUUGAACCAAUAUUAUCCAAAACAAUUAUUA
 CAAAGCACCGUUAACAACUGGUAUUUAUUAUUAUUAUUAUUAUUAUUAUUAUUAUUAUUAUUAUUAUUAUUAUUAU
 UAGAUAACAACAACAGUUCUUCUGCCCAAGCCUCCAGAGCACCUGGACUCCAGGUGCAUCCACACUAGCUGACUAGUU
 UUAUUAUUUUUAU
 UUAUUAUUUUUUUGAUGAUAUUUGGUUUUAUUAUUAUUAUUAUUAUUAUUAUUAUUAUUAUUAUUAUUAUUAUUAU
 CGCCUCCAGGGUCAAAGAAUUCUUGCCUACGCCUCUGAGUAGCUGGGAAUUAAGGCAUCCACCAACCCAGGUA
 UUUUUA
 UUGGCCUCCAAAGUUGCUGGGAUUAACAGGCGUGAGCCACUUGCCUAGCCGUCACAUUAUUAUUAUUAUUAUUAUUA
 AUUUUUGUUCUUCUUGCUGCCGUCAUGGUGGAAUUGGCUUGCUAUUGCUAUGCUUUGGUGCCAAUGCCUUUGC
 ACUGGCAUUAACAACUUAUGAAGAGAAACAAGUAGCCACACCUCAAUAUUAUUGGCUUGUACAACAACUGCCUUAUA
 CUACACAACCAUUAU
 UAACAGUUGUUAAGACUUAAGGCCAGUUAUUAUUAUUAUUAUUAUUAUUAUUAUUAUUAUUAUUAUUAUUAUUAU
 UUAAGAAUUGUUAU
 UCACGAAAUCUUAAGUUAUUUUUGCUUAUUAUUAUUAUUAUUAUUAUUAUUAUUAUUAUUAUUAUUAUUAUUAUUAU
 CAUCAUCA-3'

Supplementary Figure S1: 3'UTR sequences of putative MBNL2 stability targets: *CSNRP1* and *OSMR*. Putative binding sites are highlighted in yellow (clustered 5'-YGCY-3' motifs; Lambert *et al.*, 2014).

SMAD7
3' UTR coordinates: chr18:48919853-48921371 (-)
5' -
CCGC GUCGGAGGGGACAGAGCGUGAGCUGAGCAGGCCACACUUAACAUCUUUGCUUCUAAUAUUUUCCUCUGAGUGCU
UGCUUUUCAUGCAAACUCUUUGGUCGUUUUUUUUUUUUUUGUUGGUUGGUUUUCUUUCUCGUCUCGUUUUGUUGUUCUGUU
UUUUUUUCUCUUUGAGAAUAGCUUAUGAAAAGAAUUUGGGGGUUUUUUUGGAAGAAGGGGCAGGUAUGAUCGGCAGGA
CACCUGAUGAGGAGGGGAAAGCAGAAUCCAAGCACCAACACAGUUAUGAAGGGGGGGGUAUCUUUACUUG
UCAGGAGUGUGUGAGUGUGAGUGUGGCGUGUGUGCACGGUGUGCAGGAGCGGCAGUAGGGGAGACAACGUCUCUU
UGUUUUUGUCUUUUGGAUGUCCCGCAGAGAGGUUUGCAGUCCAAAGCGGUGUCUUCUUGCCUUUGGACACGCUCA
GUGGGGACAGGGCAGUACUUGGCAAGCUGGCGGCGGGGUCAGCAGUCCAGGAGCAGGACGGCUCUGUCCAGCCUGGG
AAAGCCCUUGCCUCUCUCCUUAAGGACACGGGCUCCAGGCUUUGAGAGCGAGCCUGCUAUGUGCCGAA
CCAGAACC AAUUUUUCAUCCUUGUCUUUUCCUUCUGCCAGCCUUGCCAUUUGUAGCGUCUUUUUUUUGGCAUCU
GCUCUGAUCUCCUGAGAUUGGCUUCCAAAGGCGUGCCGGGCGAGCCCCUACAGUUAUUGCUCACCCAGUCCUCUCC
CCUCAGCCUCUCCCUGCCUGCCUGGUGACAUCAAGUUUUUCCGGACUUAGAAAACAGCUCAGCACUGCCUGCUCCAU
CCUGUGUUUAGCUCUGCUUUUAGGCCAGCAAGCGGGGAGUCCUUGGGAGGACAUUCUAGCAGUCCUCCUCCUCAA
GAGGAUUUGGUCCGUAUAACCCAAAGGUACCAUCCUAGGUGACACCUAACUCUUUUAUUUUAUUUUAUUUUAUUUUAUU
CUCGUAUGAUACUUCGACACUGUUCUUAAGCUAAUGAGCAUGUUUAGACUUUAACAUAAGCUUUUUUUAACUACAAGGU
UUAAUAGAACAGAGAGCAUUCUAUUGGAAUUUAGCAUUGUAGUGCUUUUGAGAGAGAAAGGACUCCUGAAAAAAACCU
GAGAUUUUUAAGAAAAAAUUGUUAUUUAUGUUUAUUAUAAUUAUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUA
UCAUUUUUUUUAUUUGCAUUGUUAUAAACAAGAAAAUAAAGAAUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUA
CAAAGCCAAAUUAAAAAGAAACACAAAGAUUGGUGUUUUUUUCUUAUGGGGUAUACCCUAGCUGAAUUGUUUUUAUU
GGAGUUUAUGUCCAUUAACGAUUUUUUUUUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUAUUUA

Supplementary Figure S3: 3'UTR sequences of putative MBNL2 stability targets: *SMAD7*. Putative binding sites are highlighted in yellow (clustered 5'-YGCY-3' motifs; Lambert *et al.*, 2014).



Supplementary Figure S4: Investigation of convergences and discrepancies between CIRCexplorer and find_circ tools in predicting circRNAs from a single MCF-7 sample (normoxia, replicate 1), similar to Figure 3.10. (A) Venn diagram depicting the overlap between the predictions of CIRCexplorer and find_circ algorithms. **(B)** Scatter plot comparing total back-splice reads estimated by the two algorithms for the 2381 circRNAs in common. r : Pearson correlation. **(C)** Box plot shows the quantification of circRNAs by CIRCexplorer and find_circ for circRNAs in common or from only a single algorithm. circRNAs detected only by a single tool are in general less abundant, with notable exceptions (labelled). **(D)** Characterisation of the 675 circRNAs predicted exclusively by find_circ, in terms of genomic aligner (STAR vs. Bowtie2) and gene annotation. **(E)** Characterisation of the 2509 circRNAs predicted exclusively by CIRCexplorer, in terms of splice-site signal, genomic size, supporting back-splice reads, alignment adjustment by CIRCexplorer and gene annotation. 74% circRNAs are supported by a single unique back-splice read in find_circ measurements, highlighting the importance of filtering on unique back-splice reads to exclude PCR artefacts. For 528 circRNA the reason of the discrepancy remained unclear.

References

1. Abdelmohsen, K., Panda, A. C., Munk, R., Grammatikakis, I., Dudekula, D. B., De, S., Kim, J., Noh, J. H., Kim, K. M., Martindale, J. L. & Gorospe, M. Identification of HuR target circular RNAs uncovers suppression of PABPN1 translation by CircPABPN1. *RNA Biology* **14**, 361–369 (2017).
2. Abe, H., Semba, H. & Takeda, N. The Roles of Hypoxia Signaling in the Pathogenesis of Cardiovascular Diseases. *Journal of Atherosclerosis and Thrombosis* **24**, 884–894 (2017).
3. Abu-Jamous, B., Buffa, F. M., Harris, A. L. & Nandi, A. K. In vitro down-regulated hypoxia transcriptome is associated with poor prognosis in breast cancer. *Molecular Cancer* **16**, 105–19 (2017).
4. Adereth, Y., Dammai, V., Kose, N., Li, R. & Hsu, T. RNA-dependent integrin α 3 protein localization regulated by the Muscleblind-like protein MLP1. *Nature Cell Biology* **7**, 1240–1247 (2005).
5. Aktas, T., Avşar İlk, bibinitperiodI., Maticzka, D., Bhardwaj, V., Pessoa Rodrigues, C., Mittler, G., Manke, T., Backofen, R. & Akhtar, A. DHX9 suppresses RNA processing defects originating from the Alu invasion of the human genome. *Nature* **544**, 115–119 (2017).
6. Al Tameemi, W., Dale, T. P., Al-Jumaily, R. M. K. & Forsyth, N. R. Hypoxia-Modified Cancer Cell Metabolism. *Frontiers in cell and developmental biology* **7**, 4 (2019).
7. Anczuków, O. & Krainer, A. R. Splicing-factor alterations in cancers. *RNA* **22**, 1285–1301 (2016).
8. Anders, S., Pyl, P. T. & Huber, W. HTSeq—a Python framework to work with high-throughput sequencing data. *Bioinformatics* **31**, 166–169 (2015).
9. Anufrieva, K. S., Shender, V. O., Arapidi, G. P., Pavlyukov, M. S., Shakhparonov, M. I., Shnaider, P. V., Butenko, I. O., Lagarkova, M. A. & Govorun,

- V. M. Therapy-induced stress response is associated with downregulation of pre-mRNA splicing in cancer cells. *Genome Medicine* **10**, 49 (2018).
10. Ashwal-Fluss, R., Meyer, M., Pamudurti, N. R., Ivanov, A., Bartok, O., Hanan, M., Evantal, N., Memczak, S., Rajewsky, N. & Kadener, S. circRNA Biogenesis Competes with Pre-mRNA Splicing. *Molecular Cell* **56**, 55–66 (2014).
 11. Batra, R., Charizanis, K., Manchanda, M., Mohan, A., Li, M., Finn, D. J., Goodwin, M., Zhang, C., Sobczak, K., Thornton, C. A. & Swanson, M. S. Loss of MBNL leads to disruption of developmentally regulated alternative polyadenylation in RNA-mediated disease. *Molecular Cell* **56**, 311–322 (2014).
 12. Benita, Y., Kikuchi, H., Smith, A. D., Zhang, M. Q., Chung, D. C. & Xavier, R. J. An integrative genomics approach identifies Hypoxia Inducible Factor-1 (HIF-1)-target genes that form the core response to hypoxia. *Nucleic Acids Research* **37**, 4587–4602 (2009).
 13. Berglund, J. A., Abovich, N & Rosbash, M. A cooperative interaction between U2AF65 and mBBP/SF1 facilitates branchpoint region recognition. *Genes & Development* **12**, 858–867 (1998).
 14. Berglund, J. A., Chua, K, Abovich, N, Reed, R & Rosbash, M. The splicing factor BBP interacts specifically with the pre-mRNA branchpoint sequence UACUAAC. *Cell* **89**, 781–787 (1997).
 15. Boeckel, J.-N., Jaé, N., Heumüller, A. W., Chen, W., Boon, R. A., Stellos, K., Zeiher, A. M., John, D., Uchida, S. & Dimmeler, S. Identification and Characterization of Hypoxia-Regulated Endothelial Circular RNA. *Circulation research* **117**, 884–890 (2015).
 16. Bowler, E., Porazinski, S., Uzor, S., Thibault, P., Durand, M., Lapointe, E., Rouschop, K. M. A., Hancock, J., Wilson, I. & Lodomery, M. Hypoxia leads to significant changes in alternative splicing and elevated expression of CLK splice factor kinases in PC3 prostate cancer cells. *BMC Cancer* **18**, 355–11 (2018).
 17. Brady, L. K., Wang, H., Radens, C. M., Bi, Y., Radovich, M., Maity, A., Ivan, C., Ivan, M., Barash, Y. & Koumenis, C. Transcriptome analysis of hypoxic cancer cells uncovers intron retention in EIF2B5 as a mechanism to inhibit translation. *PLoS biology* **15**, e2002623–29 (2017).

18. Braunschweig, U., Gueroussov, S., Plocik, A. M., Graveley, B. R. & Blencowe, B. J. Dynamic integration of splicing within gene regulatory pathways. *Cell* **152**, 1252–1269 (2013).
19. Braunschweig, U., Barbosa-Morais, N. L., Pan, Q., Nachman, E. N., Alipanahi, B., Gonatopoulos-Pournatzis, T., Frey, B., Irimia, M. & Blencowe, B. J. Widespread intron retention in mammals functionally tunes transcriptomes. *Genome Research* **24**, 1774–1786 (2014).
20. Burset, M, Seledtsov, I. A. & Solovyev, V. V. Analysis of canonical and non-canonical splice sites in mammalian genomes. *Nucleic Acids Research* **28**, 4364–4375 (2000).
21. Capel, B, Swain, A, Nicolis, S, Hacker, A, Walter, M, Koopman, P, Goodfellow, P & Lovell-Badge, R. Circular transcripts of the testis-determining gene Sry in adult mouse testis. *Cell* **73**, 1019–1030 (1993).
22. Chaabane, M., Williams, R. M., Stephens, A. T. & Park, J. W. circDeep: Deep learning approach for circular RNA classification from other long non-coding RNA. *Bioinformatics* **10**, e0141287. (2019).
23. Charizanis, K., Lee, K.-Y., Batra, R., Goodwin, M., Zhang, C., Yuan, Y., Shiue, L., Cline, M., Scotti, M. M., Xia, G., Kumar, A., Ashizawa, T., Clark, H. B., Kimura, T., Takahashi, M. P., Fujimura, H., Jinnai, K., Yoshikawa, H., Gomes-Pereira, M., Gourdon, G., Sakai, N., Nishino, S., Foster, T. C., Ares Jr, M., Darnell, R. B. & Swanson, M. S. Muscleblind-like 2-Mediated Alternative Splicing in the Developing Brain and Dysregulation in Myotonic Dystrophy. *Neuron* **75**, 437–450 (2012).
24. Chee, N. T., Lohse, I. & Brothers, S. P. mRNA-to-protein translation in hypoxia. *Molecular Cancer* **18**, 49–13 (2019).
25. Chen, A., Sceneay, J., Götde, N., Kinwel, T., Ham, S., Thompson, E. W., Humbert, P. O. & Möller, A. Intermittent hypoxia induces a metastatic phenotype in breast cancer. *Oncogene*, 1–12 (2018).
26. Chen, L.-L. The biogenesis and emerging roles of circular RNAs. *Nature Reviews Molecular Cell Biology* **17**, 205–211 (2016).
27. Chen, X., Han, P., Zhou, T., Guo, X., Song, X. & Li, Y. circRNADb: A comprehensive database for human circular RNAs with protein-coding annotations. *Scientific Reports*, 1–6 (2016).
28. Cheng, J., Metge, F. & Dieterich, C. Specific identification and quantification of circular RNAs from sequencing data. *Bioinformatics*, 1–3 (2015).

29. Cheng, X., Qiu, J., Wang, S., Yang, Y., Guo, M., Wang, D., Luo, Q. & Xu, L. Comprehensive circular RNA profiling identifies CircFAM120A as a new biomarker of hypoxic lung adenocarcinoma. *Annals of Translational Medicine* **7**, 442 (2019).
30. Chi, J.-T., Wang, Z., Nuyten, D. S. A., Rodriguez, E. H., Schaner, M. E., Salim, A., Wang, Y., Kristensen, G. B., Helland, A., Børresen-Dale, A.-L., Giaccia, A., Longaker, M. T., Hastie, T., Yang, G. P., van de Vijver, M. J. & Brown, P. O. Gene expression programs in response to hypoxia: cell type specificity and prognostic significance in human cancers. *PLoS medicine* **3**, e47 (2006).
31. Chipurupalli, S., Kannan, E., Tergaonkar, V., D'Andrea, R. & Robinson, N. Hypoxia Induced ER Stress Response as an Adaptive Mechanism in Cancer. *International Journal of Molecular Sciences* **20**, 749–17 (2019).
32. Choudhry, H. & Harris, A. L. Advances in Hypoxia-Inducible Factor Biology. *Cell Metabolism* **27**, 281–298 (2018).
33. Chuang, T.-J., Wu, C.-S., Chen, C.-Y., Hung, L.-Y., Chiang, T.-W. & Yang, M.-Y. NCLscan: accurate identification of non-co-linear transcripts (fusion, trans-splicing and circular RNA) with a good balance between sensitivity and precision. *Nucleic Acids Research* **44**, e29–e29 (2016).
34. Cieply, B. & Carstens, R. P. Functional roles of alternative splicing factors in human disease. *Wiley Interdisciplinary Reviews: RNA* **6**, 311–326 (2015).
35. Cocquerelle, C, Mascrez, B, Héтуin, D & Bailleul, B. Mis-splicing yields circular RNA molecules. *The FASEB Journal* **7**, 155–160 (1993).
36. Coltri, P. P., dos Santos, M. G. P. & da Silva, G. H. G. Splicing and cancer: Challenges and opportunities. *Wiley Interdisciplinary Reviews: RNA* **219**, e1527–20 (2019).
37. Conn, S. J., Pillman, K. A., Toubia, J., Conn, V. M., Salmanidis, M., Phillips, C. A., Roslan, S., Schreiber, A. W., Gregory, P. A. & Goodall, G. J. The RNA binding protein quaking regulates formation of circRNAs. *Cell* **160**, 1125–1134 (2015).
38. Coomer, A. O., Black, F., Greystoke, A., Munkley, J. & Elliott, D. J. Alternative splicing in lung cancer. *Biochimica et biophysica acta. Gene regulatory mechanisms* **1862**, 194388 (2019).

39. Danan, M., Schwartz, S., Edelheit, S. & Sorek, R. Transcriptome-wide discovery of circular RNAs in Archaea. *Nucleic Acids Research* **40**, 3131–3142 (2011).
40. Dasari, S. & Bernard Tchounwou, P. Cisplatin in cancer therapy: Molecular mechanisms of action. *European Journal of Pharmacology* **740**, 364–378 (2014).
41. deLorimier, E., Hinman, M. N., Copperman, J., Datta, K., Guenza, M. & Berglund, J. A. Pseudouridine Modification Inhibits Muscblind-like 1 (MBNL1) Binding to CCUG Repeats and Minimally Structured RNA through Reduced RNA Flexibility. *The Journal of biological chemistry* **292**, 4350–4357 (2017).
42. Dengler, V. L., Galbraith, M. D. & Espinosa, J. M. Transcriptional regulation by hypoxia inducible factors. *Critical reviews in biochemistry and molecular biology* **49**, 1–15 (2013).
43. Di Liddo, A., de Oliveira Freitas Machado, C., Fischer, S., Ebersberger, S., Heumüller, A. W., Weigand, J. E., Müller-McNicoll, M. & Zarnack, K. A combined computational pipeline to detect circular RNAs in human cancer cells under hypoxic stress. *Journal of molecular cell biology* **11**, 829–844 (2019).
44. Djebali, S., Davis, C. A., Merkel, A., Dobin, A., Lassmann, T., Mortazavi, A., Tanzer, A., Lagarde, J., Lin, W., Schlesinger, F., Xue, C., Marinov, G. K., Khatun, J., Williams, B. A., Zaleski, C., Rozowsky, J., Röder, M., Kokocinski, F., Abdelhamid, R. F., Alioto, T., Antoshechkin, I., Baer, M. T., Bar, N. S., Batut, P., Bell, K., Bell, I., Chakraborty, S., Chen, X., Chrast, J., Curado, J., Derrien, T., Drenkow, J., Dumais, E., Dumais, J., Dutttagupta, R., Falconnet, E., Fastuca, M., Fejes-Toth, K., Ferreira, P., Foissac, S., Fullwood, M. J., Gao, H., Gonzalez, D., Gordon, A., Gunawardena, H., Howald, C., Jha, S., Johnson, R., Kapranov, P., King, B., Kingswood, C., Luo, O. J., Park, E., Persaud, K., Preall, J. B., Ribeca, P., Risk, B., Robyr, D., Sammeth, M., Schaffer, L., See, L.-H., Shahab, A., Skancke, J., Suzuki, A. M., Takahashi, H., Tilgner, H., Trout, D., Walters, N., Wang, H., Wrobel, J., Yu, Y., Ruan, X., Hayashizaki, Y., Harrow, J., Gerstein, M., Hubbard, T., Reymond, A., Antonarakis, S. E., Hannon, G., Giddings, M. C., Ruan, Y., Wold, B., Carninci, P., Guigó, R. & Gingeras, T. R. Landscape of transcription in human cells. *Nature* **489**, 101–108 (2012).

45. Dobin, A, Davis, C. A., Schlesinger, F, Drenkow, J, Zaleski, C, Jha, S, Batut, P, Chaisson, M & Gingeras, T. R. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2012).
46. Dodt, M., Roehr, J. T., Ahmed, R. & Dieterich, C. FLEXBAR-Flexible Barcode and Adapter Processing for Next-Generation Sequencing Platforms. *Biology* **1**, 895–905 (2012).
47. Du, H., Cline, M. S., Osborne, R. J., Tuttle, D. L., Clark, T. A., Donohue, J. P., Hall, M. P., Shiue, L., Swanson, M. S., Thornton, C. A. & Ares, M. Aberrant alternative splicing and extracellular matrix gene expression in mouse models of myotonic dystrophy. *Nature Structural & Molecular Biology* **17**, 187–193 (2010).
48. Du, W. W., Yang, W., Liu, E., Yang, Z., Dhaliwal, P. & Yang, B. B. Foxo3 circular RNA retards cell cycle progression via forming ternary complexes with p21 and CDK2. *Nucleic Acids Research* **44**, 2846–2858 (2016).
49. Dubin, R. A., Kazmi, M. A. & Ostrer, H. Inverted repeats are necessary for circularization of the mouse testis Sry transcript. *Gene* **167**, 245–248 (1995).
50. Dvinge, H., Kim, E., Abdel-Wahab, O. & Bradley, R. K. RNA splicing factors as oncoproteins and tumour suppressors. *Nature Reviews Cancer* **16**, 413–430 (2016).
51. Ebbesen, K. K., Kjems, J. & Hansen, T. B. Circular RNAs: Identification, biogenesis and function. *Biochimica et biophysica acta* **1859**, 163–168 (2016).
52. El Marabti, E. & Younis, I. The Cancer Spliceome: Reprogramming of Alternative Splicing in Cancer. *Frontiers in Molecular Biosciences* **5**, 80 (2018).
53. Enright, A. J., John, B., Gaul, U., Tuschl, T., Sander, C. & Marks, D. S. MicroRNA targets in Drosophila. *Genome Biology* **5**, R1 (2003).
54. Errichelli, L., Dini Modigliani, S., Laneve, P., Colantoni, A., Legnini, I., Caputo, D., Rosa, A., De Santis, R., Scarfò, R., Peruzzi, G., Lu, L., Caffarelli, E., Shneider, N. A., Morlando, M. & Bozzoni, I. FUS affects circular RNA expression in murine embryonic stem cell-derived motor neurons. *Nature Communications* **8**, 14741 (2017).
55. Escobar-Hoyos, L., Knorr, K. & Abdel-Wahab, O. Aberrant RNA Splicing in Cancer. *Annual Review of Cancer Biology* **3**, 167–185 (2019).
56. Ewels, P., Magnusson, M., Lundin, S. & Källner, M. MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics* **32**, 3047–3048 (2016).

57. Fardaei, M., Rogers, M. T., Thorpe, H. M., Larkin, K., Hamshire, M. G., Harper, P. S. & Brook, J. D. Three proteins, MBNL, MBLL and MBXL, co-localize in vivo with nuclear foci of expanded-repeat transcripts in DM1 and DM2 cells. *Human molecular genetics* **11**, 805–814 (2002).
58. Fei, T., Chen, Y., Xiao, T., Li, W., Cato, L., Zhang, P., Cotter, M. B., Bowden, M., Lis, R. T., Zhao, S. G., Wu, Q., Feng, F. Y., Loda, M., He, H. H., Liu, X. S. & Brown, M. Genome-wide CRISPR screen identifies HNRNPL as a prostate cancer dependency regulating RNA splicing. *Proceedings of the National Academy of Sciences of the United States of America* **114**, E5207–E5215 (2017).
59. Fica, S. M. & Nagai, K. Cryo-electron microscopy snapshots of the spliceosome: structural insights into a dynamic ribonucleoprotein machine. *Nature Structural & Molecular Biology* **24**, 791–799 (2017).
60. Fischer, J. W. & Leung, A. K. L. CircRNAs: a regulator of cellular stress. *Critical reviews in biochemistry and molecular biology* **52**, 1–17 (2017).
61. Fischer, S., Di Liddo, A., Taylor, K., Sobczak, K., Zarnack, K. & Weigand, J. E. Muscleblind-like 2 controls the hypoxia response of cancer cells (*in revision*).
62. Fish, L., Pencheva, N., Goodarzi, H., Tran, H., Yoshida, M. & Tavazoie, S. F. Muscleblind-like 1 suppresses breast cancer metastatic colonization and stabilizes metastasis suppressor transcripts. *Genes & Development* **30**, 386–398 (2016).
63. Frankish, A., Diekhans, M., Ferreira, A.-M., Johnson, R., Jungreis, I., Loveland, J., Mudge, J. M., Sisu, C., Wright, J., Armstrong, J., Barnes, I., Berry, A., Bignell, A., Carbonell Sala, S., Chrast, J., Cunningham, F., Di Domenico, T., Donaldson, S., Fiddes, I. T., García Girón, C., Gonzalez, J. M., Grego, T., Hardy, M., Hourlier, T., Hunt, T., Izuogu, O. G., Lagarde, J., Martin, F. J., Martínez, L., Mohanan, S., Muir, P., Navarro, F. C. P., Parker, A., Pei, B., Pozo, F., Ruffier, M., Schmitt, B. M., Stapleton, E., Suner, M.-M., Sycheva, I., Uszczyńska-Ratajczak, B., Xu, J., Yates, A., Zerbino, D., Zhang, Y., Aken, B., Choudhary, J. S., Gerstein, M., Guigó, R., Hubbard, T. J. P., Kellis, M., Paten, B., Reymond, A., Tress, M. L. & Flicek, P. GENCODE reference annotation for the human and mouse genomes. *Nucleic Acids Research* **47**, D766–D773 (2019).

64. Fu, X.-D. & Ares, M. Context-dependent control of alternative splicing by RNA-binding proteins. *Nature Reviews Genetics* **15**, 689–701 (2014).
65. Gao, D., Zhang, X., Liu, B., Meng, D., Fang, K., Guo, Z. & Li, L. Screening circular RNA related to chemotherapeutic resistance in breast cancer. *Epigenomics* **9**, 1175–1188 (2017).
66. Gao, Y. & Zhao, F. Computational Strategies for Exploring Circular RNAs. *Trends in Genetics* **34**, 389–400 (2018).
67. Gao, Y., Wang, J. & Zhao, F. CIRI: an efficient and unbiased algorithm for de novo circular RNA identification. *Genome Biology* **16**, 4 (2015).
68. Gao, Y., Wang, J., Zheng, Y., Zhang, J., Chen, S. & Zhao, F. Comprehensive identification of internal structure and alternative splicing events in circular RNAs. *Nature Communications* **7**, 1–13 (2016).
69. Geng, Y., Jiang, J. & Wu, C. Function and clinical significance of circRNAs in solid tumors. *Journal of hematology & oncology* **11**, 98–20 (2018).
70. Gilkes, D. M., Semenza, G. L. & Wirtz, D. Hypoxia and the extracellular matrix: drivers of tumour metastasis. *Nature Reviews Cancer* **14**, 430–439 (2014).
71. Glažar, P., Papavasileiou, P. & Rajewsky, N. circBase: a database for circular RNAs. *RNA* **20**, 1666–1670 (2014).
72. Gossage, L., Eisen, T. & Maher, E. R. VHL, the story of a tumour suppressor gene. *Nature Reviews Cancer* **15**, 55–64 (2015).
73. Graham, K. & Unger, E. Overcoming tumor hypoxia as a barrier to radiotherapy, chemotherapy and immunotherapy in cancer treatment. *International journal of nanomedicine* **13**, 6049–6058 (2018).
74. Grant, C. E., Bailey, T. L. & Noble, W. S. FIMO: scanning for occurrences of a given motif. *Bioinformatics* **27**, 1017–1018 (2011).
75. Greijer, A. E. The role of hypoxia inducible factor 1 (HIF-1) in hypoxia induced apoptosis. *Journal of Clinical Pathology* **57**, 1009–1014 (2004).
76. Guo, J. U., Agarwal, V., Guo, H. & Bartel, D. P. Expanded identification and characterization of mammalian circular RNAs. *Genome Biology* **15**, 409 (2014).
77. Guyot, M. & Pagès, G. in *VEGF Signaling: Methods and Protocols* (ed Fiedler, L.) 3–23 (Springer New York, New York, NY, 2015). ISBN: 978-1-4939-2917-7. doi:10.1007/978-1-4939-2917-7_1. <https://doi.org/10.1007/978-1-4939-2917-7_1>.

78. Han, C., Seebacher, N. A., Hornicek, F. J., Kan, Q. & Duan, Z. Regulation of microRNAs function by circular RNAs in human cancer. *Oncotarget* **8**, 64622–64637 (2017).
79. Han, D., Li, J., Wang, H., Su, X., Hou, J., Gu, Y., Qian, C., Lin, Y., Liu, X., Huang, M., Li, N., Zhou, W., Yu, Y. & Cao, X. Circular RNA circMTO1 acts as the sponge of microRNA-9 to suppress hepatocellular carcinoma progression. *Hepatology* **66**, 1151–1164 (2017).
80. Han, J., Li, J., Ho, J. C., Chia, G. S., Kato, H., Jha, S., Yang, H., Poellinger, L. & Lee, K. L. Hypoxia is a Key Driver of Alternative Splicing in Human Breast Cancer Cells. *Scientific Reports* **7**, 4108 (2017).
81. Hanahan, D. & Weinberg, R. A. Hallmarks of cancer: the next generation. *Cell* **144**, 646–674 (2011).
82. Hansen, T. B. Improved circRNA Identification by Combining Prediction Algorithms. *Frontiers in cell and developmental biology* **6**, 1094–9 (2018).
83. Hansen, T. B., Venø, M. T., Damgaard, C. K. & Kjems, J. Comparison of circular RNA prediction tools. *Nucleic Acids Research* **44**, e58–e58 (2016).
84. Hansen, T. B., Wiklund, E. D., Bramsen, J. B., Villadsen, S. B., Statham, A. L., Clark, S. J. & Kjems, J. o. r. miRNA-dependent gene silencing involving Ago2-mediated cleavage of a circular antisense RNA. *The EMBO Journal* **30**, 4414–4422 (2011).
85. Hansen, T. B., Jensen, T. I., Clausen, B. H., Bramsen, J. B., Finsen, B., Damgaard, C. K. & Kjems, J. Natural RNA circles function as efficient microRNA sponges. *Nature* **495**, 384–388 (2013).
86. Harris, B. H. L., Barberis, A, West, C. M. L. & Buffa, F. M. Gene Expression Signatures as Biomarkers of Tumour Hypoxia. *Clinical oncology* **27**, 547–560 (2015).
87. He, R., Liu, P., Xie, X., Zhou, Y., Liao, Q., Xiong, W., Li, X., Li, G., Zeng, Z. & Tang, H. circGFRA1 and GFRA1 act as ceRNAs in triple negative breast cancer by regulating miR-34a. *Journal of experimental & clinical cancer research* **36**, 145 (2017).
88. Heikkinen, P. T., Nummela, M., Jokilehto, T., Grenman, R., Kähäri, V.-M. & Jaakkola, P. M. Hypoxic conversion of SMAD7 function from an inhibitor into a promoter of cell invasion. *Cancer Research* **70**, 5984–5993 (2010).
89. Hoffmann, S., Otto, C., Doose, G., Tanzer, A., Langenberger, D., Christ, S., Kunz, M., Holdt, L. M., Teupser, D., Hackermüller, J. & Stadler, P. F. A

- multi-split mapping algorithm for circular RNA, splicing, trans-splicing and fusion detection. *Genome Biology* **15**, 1–11 (2014).
90. Hong, W. X., Hu, M. S., Esquivel, M., Liang, G. Y., Rennert, R. C., McArdle, A., Paik, K. J., Duscher, D., Gurtner, G. C., Lorenz, H. P. & Longaker, M. T. The Role of Hypoxia-Inducible Factor in Wound Healing. *Advances in wound care* **3**, 390–399 (2014).
 91. Hsu, M. T. & Coca-Prados, M. Electron microscopic evidence for the circular form of RNA in the cytoplasm of eukaryotic cells. *Nature* **280**, 339–340 (1979).
 92. Huse, K., Bakkebo, M., Wälchli, S., Oksvold, M. P., Hilden, V. I., Forfang, L., Bredahl, M. L., Liestøl, K., Alizadeh, A. A., Smeland, E. B. & Myklebust, J. H. Role of Smad Proteins in Resistance to BMP-Induced Growth Inhibition in B-Cell Lymphoma. *PLOS ONE* **7**, e46117 (2012).
 93. Ivan, M., Kondo, K., Yang, H., Kim, W., Valiando, J., Ohh, M., Salic, A., Asara, J. M., Lane, W. S. & Kaelin, W. G. HIF α targeted for VHL-mediated destruction by proline hydroxylation: implications for O₂ sensing. *Science* **292**, 464–468 (2001).
 94. Ivanov, A., Memczak, S., Wyler, E., Torti, F., Porath, H. T., Orejuela, M. R., Piechotta, M., Levanon, E. Y., Landthaler, M., Dieterich, C. & Rajewsky, N. Analysis of intron sequences reveals hallmarks of circular RNA biogenesis in animals. *Cell Reports* **10**, 170–177 (2015).
 95. Izuogu, O. G., Alhasan, A. A., Alafghani, H. M., Santibanez-Koref, M., Elliott, D. J., Elliot, D. J. & Jackson, M. S. PTESFinder: a computational method to identify post-transcriptional exon shuffling (PTES) events. *BMC bioinformatics* **17**, 31 (2016).
 96. Jaakkola, P., Mole, D. R., Tian, Y. M., Wilson, M. I., Gielbert, J., Gaskell, S. J., von Kriegsheim, A., Hebestreit, H. F., Mukherji, M., Schofield, C. J., Maxwell, P. H., Pugh, C. W. & Ratcliffe, P. J. Targeting of HIF- α to the von Hippel-Lindau ubiquitylation complex by O₂-regulated prolyl hydroxylation. *Science* **292**, 468–472 (2001).
 97. Jacob, A. G. & Smith, C. W. J. Intron retention as a component of regulated gene expression programs. *Human Genetics* **136**, 1043–1057 (2017).
 98. Jakobi, T. & Dieterich, C. Computational approaches for circular RNA analysis. *Wiley Interdisciplinary Reviews: RNA* **22**, e1528–14 (2019).
 99. Javelaud, D., Mohammad, K. S., McKenna, C. R., Fournier, P., Luciani, F., Niewolna, M., André, J., Delmas, V., Larue, L., Guise, T. A. & Mauviel, A.

- A. Stable overexpression of Smad7 in human melanoma cells impairs bone metastasis. *Cancer Research* **67**, 2317–2324 (2007).
100. Jeck, W. R., Sorrentino, J. A., Wang, K., Slevin, M. K., Burd, C. E., Liu, J., Marzluff, W. F. & Sharpless, N. E. Circular RNAs are abundant, conserved, and associated with ALU repeats. *RNA* **19**, 141–157 (2013).
101. Jeong, S. SR Proteins: Binders, Regulators, and Connectors of RNA. *Molecules and Cells* **40**, 1–9 (2017).
102. Jost, I., Shalamova, L. A., Gerresheim, G. K., Niepmann, M., Bindereif, A. & Rossbach, O. Functional sequestration of microRNA-122 from Hepatitis C Virus by circular RNA sponges. *RNA Biology* **15**, 1032–1039 (2018).
103. Kaelin, W. G. & Ratcliffe, P. J. Oxygen sensing by metazoans: the central role of the HIF hydroxylase pathway. *Molecular Cell* **30**, 393–402 (2008).
104. Kelly, S., Greenman, C., Cook, P. R. & Papantonis, A. Exon Skipping Is Correlated with Exon Circularization. *Journal of Molecular Biology* **427**, 2414–2417 (2015).
105. Khabar, K. S. A. Hallmarks of cancer and AU-rich elements. *Wiley Interdisciplinary Reviews: RNA* **8**, e1368 (2017).
106. Khan, M. A. F., Reckman, Y. J., Aufiero, S., van den Hoogenhof, M. M. G., van der Made, I., Beqqali, A., Koolbergen, D. R., Rasmussen, T. B., van der Velden, J., Creemers, E. E. & Pinto, Y. M. RBM20 Regulates Circular RNA Production From the Titin Gene. *Circulation research* **119**, 996–1003 (2016).
107. Kikuchi, R., Nakamura, K., MacLauchlan, S., Ngo, D. T.-M., Shimizu, I., Fuster, J. J., Katanasaka, Y., Yoshida, S., Qiu, Y., Yamaguchi, T. P., Matsushita, T., Murohara, T., Gokce, N., Bates, D. O., Hamburg, N. M. & Walsh, K. An antiangiogenic isoform of VEGF-A contributes to impaired vascularization in peripheral artery disease. *Nature Medicine* **20**, 1464–1471 (2014).
108. Kleaveland, B., Shi, C. Y., Stefano, J. & Bartel, D. P. A Network of Noncoding Regulatory RNAs Acts in the Mammalian Brain. *Cell* **174**, 350–362.e17 (2018).
109. Konieczny, P., Stepniak-Konieczna, E. & Sobczak, K. MBNL expression in autoregulatory feedback loops. *RNA Biology* **15**, 1–8 (2018).
110. Konieczny, P., Stepniak-Konieczna, E. & Sobczak, K. MBNL proteins and their target RNAs, interaction and splicing regulation. *Nucleic Acids Research* **42**, 10873–10887 (2014).

111. König, J., Zarnack, K., Rot, G., Curk, T., Kayikci, M., Zupan, B., Turner, D. J., Luscombe, N. M. & Ule, J. iCLIP reveals the function of hnRNP particles in splicing at individual nucleotide resolution. *Nature Structural & Molecular Biology* **17**, 909–915 (2010).
112. Kornblihtt, A. R. Coupling transcription and alternative splicing. *Advances in experimental medicine and biology* **623**, 175–189 (2007).
113. Kornblihtt, A. R., Schor, I. E., Alló, M., Dujardin, G., Petrillo, E. & Muñoz, M. J. Alternative splicing: a pivotal step between eukaryotic transcription and translation. *Nature Reviews Molecular Cell Biology* **14**, 153–165 (2013).
114. Kos, A., Dijkema, R., Arnberg, A. C., van der Meide, P. H. & Schellekens, H. The hepatitis delta (δ) virus possesses a circular RNA. *Nature* **323**, 558–560 (1986).
115. Kozomara, A. & Griffiths-Jones, S. miRBase: annotating high confidence microRNAs using deep sequencing data. *Nucleic Acids Research* **42**, D68–D73 (2013).
116. Krakau, S., Richard, H. & Marsico, A. PureCLIP: capturing target-specific protein-RNA interaction footprints from single-nucleotide CLIP-seq data. *Genome Biology* **18**, 240–17 (2017).
117. Kramer, M. C., Liang, D., Tatomer, D. C., Gold, B., March, Z. M., Cherry, S. & Wilusz, J. E. Combinatorial control of *Drosophila* circular RNA expression by intronic repeats, hnRNPs, and SR proteins. *Genes & Development* **29**, 2168–2182 (2015).
118. Kristensen, L. S., Hansen, T. B., Venø, M. T. & Kjems, J. Circular RNAs in cancer: opportunities and challenges in the field. *Oncogene* **7**, 155 (2017).
119. Kristensen, L. S., Andersen, M. S., Stagsted, L. V. W., Ebbesen, K. K., Hansen, T. B. & Kjems, J. x. r. The biogenesis, biology and characterization of circular RNAs. *Nature Reviews Genetics*, 1–17 (2019).
120. Kristensen, L. S., Okholm, T. L. H., Venø, M. T. & Kjems, J. Circular RNAs are abundantly expressed and upregulated during human epidermal stem cell differentiation. *RNA Biology* **15**, 280–291 (2017).
121. Krock, B. L., Skuli, N. & Simon, M. C. Hypoxia-induced angiogenesis: good and evil. *Genes & cancer* **2**, 1117–1133 (2011).

122. Lambert, N., Robertson, A., Jangi, M., McGeary, S., Sharp, P. A. & Burge, C. B. RNA Bind-n-Seq: Quantitative Assessment of the Sequence and Structural Binding Specificity of RNA Binding Proteins. *Molecular Cell* **54**, 887–900 (2014).
123. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nature Methods* **9**, 357–359 (2012).
124. Lebedeva, S., Jens, M., Theil, K., Schwanhäusser, B., Selbach, M., Landthaler, M. & Rajewsky, N. Transcriptome-wide analysis of regulatory interactions of the RNA-binding protein HuR. *Molecular Cell* **43**, 340–352 (2011).
125. Lee, J. W., Ko, J., Ju, C. & Eltzschig, H. K. Hypoxia signaling in human diseases and therapeutic targets. *Experimental & Molecular Medicine* **51**, 68–13 (2019).
126. Lee, K.-S., Cao, Y., Witwicka, H. E., Tom, S., Tapscott, S. J. & Wang, E. H. RNA-binding protein Muscleblind-like 3 (MBNL3) disrupts myocyte enhancer factor 2 (Mef2) β -exon splicing. *The Journal of biological chemistry* **285**, 33779–33787 (2010).
127. Lee, Y.-H., Jhuang, Y.-L., Chen, Y.-L., Jeng, Y.-M. & Yuan, R.-H. Paradoxical overexpression of MBNL2 in hepatocellular carcinoma inhibits tumor growth and invasion. *Oncotarget* **7**, 65589–65601 (2016).
128. Legnini, I., Di Timoteo, G., Rossi, F., Morlando, M., Briganti, F., Sthandier, O., Fatica, A., Santini, T., Andronache, A., Wade, M., Laneve, P., Rajewsky, N. & Bozzoni, I. Circ-ZNF609 Is a Circular RNA that Can Be Translated and Functions in Myogenesis. *Molecular Cell* **66**, 22–37.e9 (2017).
129. Lendahl, U., Lee, K. L., Yang, H. & Poellinger, L. Generating specificity and diversity in the transcriptional response to hypoxia. *Nature Reviews Genetics* **10**, 821–832 (2009).
130. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R. & 1000 Genome Project Data Processing Subgroup. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
131. Li, J., Wang, X., Lu, W., Xiao, Y., Yu, Y., Wang, X., Xu, C. & Shen, B. Comprehensive analysis of differentially expressed non-coding RNAs and mRNAs in gastric cancer cells under hypoxic conditions. *American Journal of Translational Research* **10**, 1022–1035 (2018).

132. Li, X., Yang, L. & Chen, L.-L. The Biogenesis, Functions, and Challenges of Circular RNAs. *Molecular Cell* **71**, 428–442 (2018).
133. Li, Z., Huang, C., Bao, C., Chen, L., Lin, M., Wang, X., Zhong, G., Yu, B., Hu, W., Dai, L., Zhu, P., Chang, Z., Wu, Q., Zhao, Y., Jia, Y., Xu, P., Liu, H. & Shan, G. Exon-intron circular RNAs regulate transcription in the nucleus. *Nature Structural & Molecular Biology* **22**, 256–264 (2015).
134. Liang, D. & Wilusz, J. E. Short intronic repeat sequences facilitate circular RNA production. *Genes & Development* **28**, 2233–2247 (2014).
135. Liang, D., Tatomer, D. C., Luo, Z., Wu, H., Yang, L., Chen, L.-L., Cherry, S. & Wilusz, J. E. The Output of Protein-Coding Genes Shifts to Circular RNAs When the Pre-mRNA Processing Machinery Is Limiting. *Molecular Cell* **68**, 940–954.e3 (2017).
136. Liang, G., Liu, Z., Tan, L., Su, A. N., Jiang, W. G. & Gong, C. HIF1 α -associated circDENND4C Promotes Proliferation of Breast Cancer Cells in Hypoxic Environment. *Anticancer research* **37**, 4337–4343 (2017).
137. Liao, D. & Johnson, R. S. Hypoxia: a key regulator of angiogenesis in cancer. *Cancer metastasis reviews* **26**, 281–290 (2007).
138. Liu, C., Zhang, C., Yang, J., Geng, X., Du, H., Ji, X. & Zhao, H. Screening circular RNA expression patterns following focal cerebral ischemia in mice. *Oncotarget* **8**, 86535–86547 (2017).
139. Liu, L., Cash, T. P., Jones, R. G., Keith, B., Thompson, C. B. & Simon, M. C. Hypoxia-induced energy stress regulates mRNA translation and cell growth. *Molecular Cell* **21**, 521–531 (2006).
140. Liu, X., Hu, Z., Zhou, J., Tian, C., Tian, G., He, M., Gao, L., Chen, L., Li, T., Peng, H. & Zhang, W. Interior circular RNA. *RNA Biology* **0**, 1 (2019).
141. Liu, Z., Luyten, I., Bottomley, M. J., Messias, A. C., Houngininou-Molango, S, Sprangers, R, Zanier, K, Krämer, A & Sattler, M. Structural basis for recognition of the intron branch site RNA by splicing factor 1. *Science* **294**, 1098–1102 (2001).
142. Liu, Z., Han, J., Lv, H., Liu, J. & Liu, R. Computational identification of circular RNAs based on conformational and thermodynamic properties in the flanking introns. *Computational Biology and Chemistry* **61**, 221–225 (2016).
143. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology* **15**, 550 (2014).

144. Lü, L., Sun, J., Shi, P., Kong, W., Xu, K., He, B., Zhang, S. & Wang, J. Identification of circular RNAs as a promising new class of diagnostic biomarkers for human breast cancer. *Oncotarget* **8**, 44096–44107 (2017).
145. Luo, L., Li, N., Lv, N. & Huang, D. SMAD7: a timer of tumor progression targeting TGF- β signaling. *Tumour biology : the journal of the International Society for Oncodevelopmental Biology and Medicine* **35**, 8379–8385 (2014).
146. Lv, X., Li, J., Zhang, C., Hu, T., Li, S., He, S., Yan, H., Tan, Y., Lei, M., Wen, M. & Zuo, J. The role of hypoxia-inducible factors in tumor angiogenesis and cell metabolism. *Genes & Diseases* **4**, 19–24 (2017).
147. Masoud, G. N. & Li, W. HIF-1 α pathway: role, regulation and intervention for cancer therapy. *Acta pharmaceutica Sinica. B* **5**, 378–389 (2015).
148. Masuda, A., Andersen, H. S., Doktor, T. K., Okamoto, T., Ito, M., Andresen, B. S. & Ohno, K. CUGBP1 and MBNL1 preferentially bind to 3' UTRs and facilitate mRNA decay. *Scientific Reports* **2**, 75–10 (2012).
149. Meléndez-Rodríguez, F., Roche, O., Sanchez-Prieto, R. & Aragonés, J. Hypoxia-Inducible Factor 2-Dependent Pathways Driving Von Hippel–Lindau-Deficient Renal Cancer. *Frontiers in Oncology* **8**, 263 (2018).
150. Memczak, S., Jens, M., Elefsinioti, A., Torti, F., Krueger, J., Rybak, A., Maier, L., Mackowiak, S. D., Gregersen, L. H., Munschauer, M., Loewer, A., Ziebold, U., Landthaler, M., Kocks, C., le Noble, F. & Rajewsky, N. Circular RNAs are a large class of animal RNAs with regulatory potency. *Nature* **495**, 333–338 (2013).
151. Miller, J. W., Urbinati, C. R., Teng-Umuay, P., Stenberg, M. G., Byrne, B. J., Thornton, C. A. & Swanson, M. S. Recruitment of human muscleblind proteins to (CUG)(n) expansions associated with myotonic dystrophy. *The EMBO Journal* **19**, 4439–4448 (2000).
152. Mole, D. R., Blancher, C., Copley, R. R., Pollard, P. J., Gleadle, J. M., Ragoussis, J. & Ratcliffe, P. J. Genome-wide association of hypoxia-inducible factor (HIF)-1 α and HIF-2 α DNA binding with expression profiling of hypoxia-inducible transcripts. *The Journal of biological chemistry* **284**, 16767–16775 (2009).
153. Mori, H., Yao, Y., Learman, B. S., Kurozumi, K., Ishida, J., Ramakrishnan, S. K., Overmyer, K. A., Xue, X., Cawthorn, W. P., Reid, M. A., Taylor, M., Ning, X., Shah, Y. M. & MacDougald, O. A. Induction of WNT11 by hypoxia

- and hypoxia-inducible factor-1 α regulates cell proliferation, migration and invasion. *Scientific Reports* **6**, 21520 (2016).
154. Mukherjee, N., Corcoran, D. L., Nusbaum, J. D., Reid, D. W., Georgiev, S., Hafner, M., Ascano, M., Tuschl, T., Ohler, U. & Keene, J. D. Integrative regulatory mapping indicates that the RNA-binding protein HuR couples pre-mRNA processing and mRNA stability. *Molecular Cell* **43**, 327–339 (2011).
 155. Müller-McNicoll, M. & Neugebauer, K. M. How cells get the message: dynamic assembly and function of mRNA-protein complexes. *Nature Reviews Genetics* **14**, 275–287 (2013).
 156. Nigro, J. M., Cho, K. R., Fearon, E. R., Kern, S. E., Ruppert, J. M., Oliner, J. D., Kinzler, K. W. & Vogelstein, B. Scrambled exons. *Cell* **64**, 607–613 (1991).
 157. Oltean, S & Bates, D. O. Hallmarks of alternative splicing in cancer. *Oncogene* **33**, 5311–5318 (2014).
 158. Pagès, H., Aboyoun, P., Gentleman, R. & DebRoy, S. *Biostrings: Efficient manipulation of biological strings* R package version 2.46.0 (2017).
 159. Pamudurti, N. R., Bartok, O., Jens, M., Ashwal-Fluss, R., Stottmeister, C., Ruhe, L., Hanan, M., Wyler, E., Perez-Hernandez, D., Ramberger, E., Shenzis, S., Samson, M., Dittmar, G., Landthaler, M., Chekulaeva, M., Rajewsky, N. & Kadener, S. Translation of CircRNAs. *Molecular Cell* **66**, 9–21.e7 (2017).
 160. Pan, Q., Shai, O., Lee, L. J., Frey, B. J. & Blencowe, B. J. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nature Genetics* **40**, 1413–1415 (2008).
 161. Pan, X. & Xiong, K. PredcircRNA: computational classification of circular RNA from other long non-coding RNA using hybrid features. *Molecular BioSystems* **11**, 2219–2226 (2015).
 162. Pandey, P. R., Rout, P. K., Das, A., Gorospe, M. & Panda, A. C. RPAD (RNase R treatment, polyadenylation, and poly(A)+ RNA depletion) method to isolate highly pure circular RNA. *Methods* **155**, 41–48 (2019).
 163. Park, S., Phukan, P. D., Zeeb, M., Martinez-Yamout, M. A., Dyson, H. J. & Wright, P. E. Structural Basis for Interaction of the Tandem Zinc Finger Domains of Human Muscleblind with Cognate RNA from Human Cardiac Troponin T. *Biochemistry* **56**, 4154–4168 (2017).
 164. Paul, S., Dansithong, W., Kim, D., Rossi, J., Webster, N. J. G., Comai, L. & Reddy, S. Interaction of muscleblind, CUG-BP1 and hnRNP H proteins

- in DM1-associated aberrant IR splicing. *The EMBO Journal* **25**, 4271–4283 (2006).
165. Peiris-Pagès, M. The role of VEGF 165b in pathophysiology. *Cell adhesion & migration* **6**, 561–568 (2012).
166. Pereira, B., Billaud, M. & Almeida, R. RNA-Binding Proteins in Cancer: Old Players and New Actors. *Trends in cancer* **3**, 506–528 (2017).
167. Perron, G., Jandaghi, P., Solanki, S., Safisamghabadi, M., Storoz, C., Karimzadeh, M., Papadakis, A. I., Arseneault, M., Scelo, G., Banks, R. E., Tost, J., Lathrop, M., Tanguay, S., Brazma, A., Huang, S., Brimo, F., Najafabadi, H. S. & Riazalhosseini, Y. A General Framework for Interrogation of mRNA Stability Programs Identifies RNA-Binding Proteins that Govern Cancer Transcriptomes. *Cell Reports* **23**, 1639–1650 (2018).
168. Platel, V., Faure, S., Corre, I. & Clere, N. Endothelial-to-Mesenchymal Transition (EndoMT): Roles in Tumorigenesis, Metastatic Extravasation and Therapy Resistance. *Journal of oncology* **2019**, 8361945–13 (2019).
169. Pohl, M., Bortfeldt, R. H., Grützmann, K. & Schuster, S. Alternative splicing of mutually exclusive exons—a review. *Bio Systems* **114**, 31–38 (2013).
170. Poulos, M. G., Batra, R., Li, M., Yuan, Y., Zhang, C., Darnell, R. B. & Swanson, M. S. Progressive impairment of muscle regeneration in muscleblind-like 3 isoform knockout mice. *Human molecular genetics* **22**, 3547–3558 (2013).
171. R Core Team. *R: A language and environment for statistical computing* R Foundation for Statistical Computing (Vienna, Austria, 2017). <<https://www.R-project.org/>>.
172. Rau, F., Freyermuth, F., Fugier, C., Villemain, J.-P., Fischer, M.-C., Jost, B., Dembele, D., Gourdon, G., Nicole, A., Duboc, D., Wahbi, K., Day, J. W., Fujimura, H., Takahashi, M. P., Auboeuf, D., Dreumont, N., Furling, D. & Charlet-Berguerand, N. Misregulation of miR-1 processing is associated with heart defects in myotonic dystrophy. *Nature Structural & Molecular Biology* **18**, 840–845 (2011).
173. Ray, D., Kazan, H., Cook, K. B., Weirauch, M. T., Najafabadi, H. S., Li, X., Gueroussov, S., Albu, M., Zheng, H., Yang, A., Na, H., Irimia, M., Matzat, L. H., Dale, R. K., Smith, S. A., Yarosh, C. A., Kelly, S. M., Nabet, B., Mecnas, D., Li, W., Laishram, R. S., Qiao, M., Lipshitz, H. D., Piano, F., Corbett, A. H., Carstens, R. P., Frey, B. J., Anderson, R. A., Lynch, K. W., Penalva, L. O. F., Lei, E. P., Fraser, A. G., Blencowe, B. J., Morris, Q. D.

- & Hughes, T. R. A compendium of RNA-binding motifs for decoding gene regulation. *Nature* **499**, 172–177 (2013).
174. Robinson, J. T., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E. S., Getz, G. & Mesirov, J. P. Integrative genomics viewer. *Nature Biotechnology* **29**, 24–26 (2011).
175. Roy, B., Haupt, L. M. & Griffiths, L. R. Review: Alternative Splicing (AS) of Genes As An Approach for Generating Protein Complexity. *Current genomics* **14**, 182–194 (2013).
176. Rybak-Wolf, A., Stottmeister, C., Glažar, P., Jens, M., Pino, N., Giusti, S., Hanan, M., Behm, M., Bartok, O., Ashwal-Fluss, R., Herzog, M., Schreyer, L., Papavasileiou, P., Ivanov, A., Öhman, M., Refojo, D., Kadener, S. & Rajewsky, N. Circular RNAs in the Mammalian Brain Are Highly Abundant, Conserved, and Dynamically Expressed. *Molecular Cell* **58**, 870–885 (2015).
177. Salceda, S & Caro, J. Hypoxia-inducible factor 1alpha (HIF-1alpha) protein is rapidly degraded by the ubiquitin-proteasome system under normoxic conditions. Its stabilization by hypoxia depends on redox-induced changes. *Journal of Biological Chemistry* **272**, 22642–22647 (1997).
178. Salot, S. & Gude, R. MTA1-mediated transcriptional repression of SMAD7 in breast cancer cell lines. *European journal of cancer (Oxford, England : 1990)* **49**, 492–499 (2013).
179. Salzman, J., Chen, R. E., Olsen, M. N., Wang, P. L. & Brown, P. O. Cell-Type Specific Features of Circular RNA Expression. *PLOS Genetics* **9**, e1003777 (2013).
180. Salzman, J., Gawad, C., Wang, P. L., Lacayo, N. & Brown, P. O. Circular RNAs Are the Predominant Transcript Isoform from Hundreds of Human Genes in Diverse Cell Types. *PLOS ONE* **7**, e30733 (2012).
181. Sand, M., Bechara, F. G., Gambichler, T., Sand, D., Bromba, M., Hahn, S. A., Stockfleth, E. & Hessam, S. Circular RNA expression in cutaneous squamous cell carcinoma. *Journal of dermatological science* **83**, 210–218 (2016).
182. Sanger, H. L., Klotz, G., Riesner, D., Gross, H. J. & Kleinschmidt, A. K. Viroids are single-stranded covalently closed circular RNA molecules existing as highly base-paired rod-like structures. *Proceedings of the National Academy of Sciences* **73**, 3852–3856 (1976).

183. Sarkar, S, Banerjee, P. K. & Selvamurthy, W. High altitude hypoxia: an intricate interplay of oxygen responsive macroevents and micromolecules. *Molecular and cellular biochemistry* **253**, 287–305 (2003).
184. Schito, L. & Semenza, G. L. Hypoxia-Inducible Factors: Master Regulators of Cancer Progression. *Trends in cancer* **2**, 758–770 (2016).
185. Schwanhäusser, B., Busse, D., Li, N., Dittmar, G., Schuchhardt, J., Wolf, J., Chen, W. & Selbach, M. Global quantification of mammalian gene expression control. *Nature* **473**, 337–342 (2011).
186. Semenza, G. L. Hypoxia-inducible factors in physiology and medicine. *Cell* **148**, 399–408 (2012).
187. Semenza, G. L. Oxygen sensing, hypoxia-inducible factors, and disease pathophysiology. *Annual review of pathology* **9**, 47–71 (2014).
188. Sena, J. A., Wang, L., Heasley, L. E. & Hu, C.-J. Hypoxia regulates alternative splicing of HIF and non-HIF target genes. *Molecular cancer research* **12**, 1233–1243 (2014).
189. Shen, S., Park, J. W., Lu, Z.-x., Lin, L., Henry, M. D., Wu, Y. N., Zhou, Q. & Xing, Y. rMATS: Robust and flexible detection of differential alternative splicing from replicate RNA-Seq data. *Proceedings of the National Academy of Sciences* **111**, E5593–E5601 (2014).
190. Shi, Y. Mechanistic insights into precursor messenger RNA splicing by the spliceosome. *Nature Reviews Molecular Cell Biology* **18**, 655–670 (2017).
191. Shibuya, M. Vascular Endothelial Growth Factor (VEGF) and Its Receptor (VEGFR) Signaling in Angiogenesis: A Crucial Target for Anti- and Pro-Angiogenic Therapies. *Genes & cancer* **2**, 1097–1105 (2011).
192. Simonson, T. S., Yang, Y., Huff, C. D., Yun, H., Qin, G., Witherspoon, D. J., Bai, Z., Lorenzo, F. R., Xing, J., Jorde, L. B., Prchal, J. T. & Ge, R. Genetic evidence for high-altitude adaptation in Tibet. *Science* **329**, 72–75 (2010).
193. Singh, D., Arora, R., Kaur, P., Singh, B., Mannan, R. & Arora, S. Overexpression of hypoxia-inducible factor and metabolic pathways: possible targets of cancer. *Cell & bioscience* **7**, 62 (2017).
194. Singh, G., Pratt, G., Yeo, G. W. & Moore, M. J. The Clothes Make the mRNA: Past and Present Trends in mRNP Fashion. *Annual Review of Biochemistry* **84**, 325–354 (2015).
195. Slattery, M. L., Herrick, J., Curtin, K., Samowitz, W., Wolff, R. K., Caan, B. J., Duggan, D., Potter, J. D. & Peters, U. Increased risk of colon cancer

- associated with a genetic polymorphism of SMAD7. *Cancer Research* **70**, 1479–1485 (2010).
196. Song, X., Zhang, N., Han, P., Moon, B.-S., Lai, R. K., Wang, K. & Lu, W. Circular RNA profile in gliomas revealed by identification tool UROBORUS. *Nucleic Acids Research* **44**, e87–e87 (2016).
197. Song, X., Zeng, Z., Wei, H. & Wang, Z. Alternative splicing in cancers: From aberrant regulation to new therapeutics. *Seminars in cell & developmental biology*, 1–25 (2017).
198. Squillace, R. M., Chenault, D. M. & Wang, E. H. Inhibition of Muscle Differentiation by the Novel Muscleblind-Related Protein CHCR. *Developmental Biology* **250**, 218–230 (2002).
199. Starke, S., Jost, I., Rossbach, O., Schneider, T., Schreiner, S., Hung, L.-H. & Bindereif, A. Exon Circularization Requires Canonical Splice Signals. *Cell Reports* **10**, 103–111 (2015).
200. Su, H., Lin, F., Deng, X., Shen, L., Fang, Y., Fei, Z., Zhao, L., Zhang, X., Pan, H., Xie, D., Jin, X. & Xie, C. Profiling and bioinformatics analyses reveal differential circular RNA expression in radioresistant esophageal cancer cells. *Journal of translational medicine* **14**, 225 (2016).
201. Sun, Q., Hao, Q. & Prasanth, K. V. Nuclear Long Noncoding RNAs: Key Regulators of Gene Expression. *Trends in Genetics*, 1–16 (2017).
202. Syed, V. TGF- β Signaling in Cancer. *Journal of cellular biochemistry* **117**, 1279–1287 (2016).
203. Szabo, L. & Salzman, J. Detecting circular RNAs: bioinformatic and experimental challenges. *Nature Reviews Genetics* **17**, 679–692 (2016).
204. Szabo, L., Morey, R., Palpant, N. J., Wang, P. L., Afari, N., Jiang, C., Parast, M. M., Murry, C. E., Laurent, L. C. & Salzman, J. Statistically based splicing detection reveals neural enrichment and tissue-specific induction of circular RNA during human fetal development. *Genome Biology* **16**, 126 (2015).
205. Sznajder, Ł. J., Michalak, M., Taylor, K., Cywoniuk, P., Kabza, M., Wojtkowiak-Szlachcic, A., Matłoka, M., Konieczny, P. & Sobczak, K. Mechanistic determinants of MBNL activity. *Nucleic Acids Research* **17**, gkw915–17 (2016).
206. Tabaglio, T., Low, D. H., Teo, W. K. L., Goy, P. A., Cywoniuk, P., Wollmann, H., Ho, J., Tan, D., Aw, J., Pavesi, A., Sobczak, K., Wee, D. K. B. & Guccione,

- E. MBNL1 alternative splicing isoforms play opposing roles in cancer. *Life science alliance* **1**, e201800157 (2018).
207. Tang, R., Qi, Q., Wu, R., Zhou, X., Wu, D., Zhou, H., Mao, Y., Li, R., Liu, C., Wang, L., Chen, W., Hua, D., Zhang, H. & Wang, W. The polymorphic terminal-loop of pre-miR-1307 binding with MBNL1 contributes to colorectal carcinogenesis via interference with Dicer1 recruitment. *Carcinogenesis* **36**, 867–875 (2015).
208. Tang, W., Fu, K., Sun, H., Rong, D., Wang, H. & Cao, H. CircRNA microarray profiling identifies a novel circulating biomarker for detection of gastric cancer. *Molecular Cancer* **17**, 1–6 (2018).
209. Taylor, K., Sznajder, Ł. J., Cywoniuk, P., Thomas, J. D., Swanson, M. S. & Sobczak, K. MBNL splicing activity depends on RNA binding site structural context. *Nucleic Acids Research* **46**, 9119–9133 (2018).
210. Tilgner, H., Knowles, D. G., Johnson, R., Davis, C. A., Chakraborty, S., Djebali, S., Curado, J., Snyder, M., Gingeras, T. R. & Guigó, R. Deep sequencing of subcellular RNA fractions shows splicing to be predominantly co-transcriptional in the human genome but inefficient for lncRNAs. *Genome Research* **22**, 1616–1625 (2012).
211. Uniacke, J., Holterman, C. E., Lachance, G., Franovic, A., Jacob, M. D., Fabian, M. R., Payette, J., Holcik, M., Pause, A. & Lee, S. An oxygen-regulated switch in the protein synthesis machinery. *Nature* **486**, 126–129 (2012).
212. Urbanski, L. M., Leclair, N. & Anczuków, O. Alternative-splicing defects in cancer: Splicing regulators and their downstream targets, guiding the way to novel cancer therapeutics. *Wiley Interdisciplinary Reviews: RNA* **9**, e1476 (2018).
213. Verduci, L., Strano, S., Yarden, Y. & Blandino, G. The circRNA-microRNA code: emerging implications for cancer diagnosis and treatment. *Molecular oncology* **13**, 669–680 (2019).
214. Vincent, H. A. & Deutscher, M. P. Substrate recognition and catalysis by the exoribonuclease RNase R. *Journal of Biological Chemistry* **281**, 29769–29775 (2006).
215. Wahl, M. C., Will, C. L. & Lührmann, R. The Spliceosome: Design Principles of a Dynamic RNP Machine. *Cell* **136**, 701–718 (2009).

216. Wang, E. T., Ward, A. J., Cherone, J. M., Giudice, J., Wang, T. T., Treacy, D. J., Lambert, N. J., Freese, P., Saxena, T., Cooper, T. A. & Burge, C. B. Antagonistic regulation of mRNA expression and splicing by CELF and MBNL proteins. *Genome Research* **25**, 858–871 (2015).
217. Wang, E. T., Cody, N. A. L., Jog, S., Biancolella, M., Wang, T. T., Treacy, D. J., Luo, S., Schroth, G. P., Housman, D. E., Reddy, S., Lécuyer, E. & Burge, C. B. Transcriptome-wide regulation of pre-mRNA splicing and mRNA localization by muscleblind proteins. *Cell* **150**, 710–724 (2012).
218. Wang, J., Zhu, M. C., Kalionis, B., Wu, J. Z., Wang, L. L., Ge, H. Y., Chen, C. C., Tang, X. D., Song, Y. L., He, H. & Xia, S. J. Characteristics of circular RNA expression in lung tissues from mice with hypoxia-induced pulmonary hypertension. *International journal of molecular medicine* **42**, 1353–1366 (2018).
219. Wang, J., Liu, K., Liu, Y., Lv, Q., Zhang, F. & Wang, H. Evaluating the bias of circRNA predictions from total RNA-Seq data. *Oncotarget* **8**, 110914–110921 (2017).
220. Wang, J. & Wang, L. Deep Learning of the Back-splicing Code for Circular RNA Formation. *Bioinformatics* **33**, 831 (2019).
221. Wang, K., Singh, D., Zeng, Z., Coleman, S. J., Huang, Y., Savich, G. L., He, X., Mieczkowski, P., Grimm, S. A., Perou, C. M., MacLeod, J. N., Chiang, D. Y., Prins, J. F. & Liu, J. MapSplice: accurate mapping of RNA-seq reads for splice junction discovery. *Nucleic Acids Research* **38**, e178–e178 (2010).
222. Wang, K., Long, B., Liu, F., Wang, J.-X., Liu, C.-Y., Zhao, B., Zhou, L.-Y., Sun, T., Wang, M., Yu, T., Gong, Y., Liu, J., Dong, Y.-H., Li, N. & Li, P.-F. A circular RNA protects the heart from pathological hypertrophy and heart failure by targeting miR-223. *European heart journal* **37**, 2602–2611 (2016).
223. Wang, P. L., Bao, Y., Yee, M.-C., Barrett, S. P., Hogan, G. J., Olsen, M. N., Dinneny, J. R., Brown, P. O. & Salzman, J. Circular RNA is expressed across the eukaryotic tree of life. *PLOS ONE* **9**, e90859 (2014).
224. Wang, Y., Liu, J., Huang, B. O., Xu, Y.-M., Li, J., Huang, L.-F., Lin, J., Zhang, J., Min, Q.-H., Yang, W.-M. & Wang, X.-Z. Mechanism of alternative splicing and its regulation. *Biomedical reports* **3**, 152–158 (2015).
225. Wang, Y. & Wang, Z. Efficient backsplicing produces translatable circular mRNAs. *RNA* **21**, 172–179 (2015).

226. Wei, J., Yang, Y., Lu, M., Lei, Y., Xu, L., Jiang, Z., Xu, X., Guo, X., Zhang, X., Sun, H. & You, Q. Recent Advances in the Discovery of HIF-1 α -p300/CBP Inhibitors as Anti-Cancer Agents. *Mini reviews in medicinal chemistry* **18**, 296–309 (2018).
227. Weis, S. M. & Chersesh, D. A. Tumor angiogenesis: molecular pathways and therapeutic targets. *Nature Medicine* **17**, 1359–1370 (2011).
228. Wesselhoeft, R. A., Kowalski, P. S. & Anderson, D. G. Engineering circular RNA for potent and stable translation in eukaryotic cells. *Nature Communications*, 1–10 (2018).
229. Westholm, J. O., Miura, P., Olson, S., Shenker, S., Joseph, B., Sanfilippo, P., Celniker, S. E., Graveley, B. R. & Lai, E. C. Genome-wide analysis of drosophila circular RNAs reveals their structural and sequence properties and age-dependent neural accumulation. *Cell Reports* **9**, 1966–1980 (2014).
230. Wu, Y., Zhao, W., Liu, Y., Tan, X., Li, X., Zou, Q., Xiao, Z., Xu, H., Wang, Y. & Yang, X. Function of HNRNPC in breast cancer cells by controlling the dsRNA-induced interferon response. *The EMBO Journal* **37**, e14–19 (2018).
231. Xia, S., Feng, J., Lei, L., Hu, J., Xia, L., Wang, J., Xiang, Y., Liu, L., Zhong, S., Han, L. & He, C. Comprehensive characterization of tissue-specific circular RNAs in the human and mouse genomes. *Briefings in Bioinformatics* **18**, 984–992 (2017).
232. Yan, C., Wan, R. & Shi, Y. Molecular Mechanisms of pre-mRNA Splicing through Structural Biology of the Spliceosome. *Cold Spring Harbor Perspectives in Biology* **11** (2019).
233. Yang, M., Su, H., Soga, T., Kranc, K. R. & Pollard, P. J. Prolyl hydroxylase domain enzymes: important regulators of cancer metabolism. *Hypoxia* **2**, 127–142 (2014).
234. Yang, P., Qiu, Z., Jiang, Y., Dong, L., Yang, W., Gu, C., Li, G. & Zhu, Y. Silencing of cZNF292 circular RNA suppresses human glioma tube formation via the Wnt/ β -catenin signaling pathway. *Oncotarget* **7**, 63449–63455 (2016).
235. Yang, Q., Zhao, J., Zhang, W., Chen, D. & Wang, Y. Aberrant alternative splicing in breast cancer. *Journal of molecular cell biology* **11**, 920–929 (2019).
236. Yang, Y., Fan, X., Mao, M., Song, X., Wu, P., Zhang, Y., Jin, Y., Yang, Y., Chen, L.-L., Wang, Y., Wong, C. C., Xiao, X. & Wang, Z. Extensive translation of circular RNAs driven by N6-methyladenosine. *Cell research* **27**, 626–641 (2017).

237. Yeo, G. & Burge, C. B. Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. *Journal of computational biology* **11**, 377–394 (2004).
238. You, X. & Conrad, T. O. Acfs: accurate circRNA identification and quantification from RNA-Seq data. *Scientific Reports* **6**, 38820–11 (2016).
239. You, X., Vlatkovic, I., Babic, A., Will, T., Epstein, I., Tushev, G., Akbalik, G., Wang, M., Glock, C., Quedenau, C., Wang, X., Hou, J., Liu, H., Sun, W., Sambandan, S., Chen, T., Schuman, E. M. & Chen, W. Neural circular RNAs are derived from synaptic genes and regulated by development and plasticity. *Nature neuroscience* **18**, 603–610 (2015).
240. Young, S. D., Marshall, R. S. & Hill, R. P. Hypoxia induces DNA overreplication and enhances metastatic potential of murine tumor cells. *Proceedings of the National Academy of Sciences* **85**, 9533–9537 (1988).
241. Yu, G., Wang, L.-G., Han, Y. & He, Q.-Y. clusterProfiler: an R package for comparing biological themes among gene clusters. *OMICS: A Journal of Integrative Biology* **16**, 284–287 (2012).
242. Yuan, J.-h., Liu, X.-n., Wang, T.-t., Pan, W., Tao, Q.-f., Zhou, W.-p., Wang, F. & Sun, S.-h. The MBNL3 splicing factor promotes hepatocellular carcinoma by increasing PXN expression through the alternative splicing of lncRNA-PXN-AS1. *Nature Cell Biology* **19**, 820–832 (2017).
243. Zarnack, K., König, J., Tajnik, M., Martincorena, I., Eustermann, S., Stévant, I., Reyes, A., Anders, S., Luscombe, N. M. & Ule, J. Direct competition between hnRNP C and U2AF65 protects the transcriptome from the exonization of Alu elements. *Cell* **152**, 453–466 (2013).
244. Zeng, X., Lin, W., Guo, M. & Zou, Q. A comprehensive overview and evaluation of circular RNA detection tools. *PLOS Computational Biology* **13**, e1005420–21 (2017).
245. Zhang, S., Zeng, X., Ding, T., Guo, L., Li, Y., Ou, S. & Yuan, H. Microarray profile of circular RNAs identifies hsa_circ_0014130 as a new circular RNA biomarker in non-small cell lung cancer. *Scientific Reports* **8**, 2878–11 (2018).
246. Zhang, X.-O., Wang, H.-B., Zhang, Y., Lu, X., Chen, L.-L. & Yang, L. Complementary sequence-mediated exon circularization. *Cell* **159**, 134–147 (2014).

247. Zhang, X.-O., Dong, R., Zhang, Y., Zhang, J.-L., Luo, Z., Zhang, J., Chen, L.-L. & Yang, L. Diverse alternative back-splicing and alternative splicing landscape of circular RNAs. *Genome Research* **26**, 1277–1287 (2016).
248. Zhang, Y., Li, J., Yu, J., Liu, H., Shen, Z., Ye, G., Mou, T., Qi, X. & Li, G. Circular RNAs signature predicts the early recurrence of stage III gastric cancer after radical surgery. *Oncotarget* **8**, 22936–22943 (2017).
249. Zhang, Y., Zhang, X.-O., Chen, T., Xiang, J.-F., Yin, Q.-F., Xing, Y.-H., Zhu, S., Yang, L. & Chen, L.-L. Circular intronic long noncoding RNAs. *Molecular Cell* **51**, 792–806 (2013).
250. Zhang, Y., Xue, W., Li, X., Zhang, J., Chen, S., Zhang, J.-L., Yang, L. & Chen, L.-L. The Biogenesis of Nascent Circular RNAs. *Cell Reports* **15**, 611–624 (2016).
251. Zheng, Q., Bao, C., Guo, W., Li, S., Chen, J., Chen, B., Luo, Y., Lyu, D., Li, Y., Shi, G., Liang, L., Gu, J., He, X. & Huang, S. Circular RNA profiling reveals an abundant circHIPK3 that regulates cell growth by sponging multiple miRNAs. *Nature Communications* **7**, 11215–13 (2016).

Acknowledgements

At the end of this PhD experience, I would like to thank my supervisor, Dr. Kathi Zarnack, for giving me the opportunity to work in her group and on exciting bioinformatics projects. Thanks for your guidance during my PhD, for believing in my capabilities from the beginning and for your continuous support. You really inspired me with your passion for science.

I would like to thank Prof. Dr. Michaela Müller-McNicoll for her co-supervision, for all time spent in discussions and for her kind support during these years. Thanks to Dr. Julia Weigand and Sandra Fischer for letting me collaborate on the MBNL2 project. I really enjoyed all our meetings and discussions. Thank you for all the members of my Thesis Advisory Committee, for their important advices for the project. Thanks to Dr. Stefanie Ebersberger for her initial supervision and help during my PhD.

A warm thank goes to Camila, for her valuable collaboration on the circRNA project. It was so nice working with you. We are a great team! I am very grateful to Cornelia for helping me with the challenging *Zusammenfassung*, and for her kindness and friendship.

Thanks to Samarth, who was there from the beginning, for very kind assistance along all steps of my PhD. Thanks to all the members of the Zarnack group, for the nice time spent together in the office, for the fruitful discussions, and for sharing frustrations, stress, and nice laughs in this period.

Importantly, I would like to thank my parents, my brother and my sister. No matter how far away we are, you never let me feel alone. Thank you for your love, support, care and for being always proud of me.

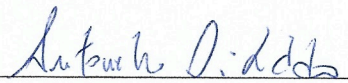
Last but not least, I am very grateful to Nello, who convinced me to move to Germany and start this new adventure together. This was the best decision ever! Thank you for your love and support throughout my PhD and the entire life, and for always giving me the strength to persevere.

Thanks!

Erklärung

Ich erkläre hiermit, dass ich mich bisher keiner Doktorprüfung im mathematisch - naturwissenschaftlichem Bereich unterzogen habe.

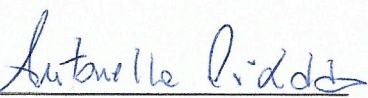
Frankfurt am Main, den 26/08/2020


Antonella Di Liddo

Versicherung

Ich versichere hiermit, dass die vorgelegte Doktorarbeit über "Computational approaches to study the RNA response to hypoxia in human cancer cells" selbständig und ohne unzulässige fremde Hilfe verfasst, andere als die in ihr angegebene Literatur nicht benutzt und, dass ich alle ganz oder annähernd übernommenen Textstellen, sowie verwendete Grafiken, Tabellen und Auswertungsprogramme gekennzeichnet habe. Außerdem versichere ich, dass die vorgelegte elektronische mit der schriftlichen Version der Doktorarbeit übereinstimmt.

Frankfurt am Main, den 26/08/2020


Antonella Di Liddo