The University Of Sheffield.

Nuclear Surveillance Pathways Play a Key Role in Regulating the Transcription Landscape of Eukaryotic Genomes

By:

Lee Garry Davidson

A thesis submitted in partial fulfilment of the requirements for the degree of

Doctor of Philosophy

The University of Sheffield

Faculty of Science

Department of Molecular Biology and Biotechnology

Submission Date

August 2018

Acknowledgements

Firstly, I would like to thank Professor Steve West for giving me the opportunity to continue working with him after all these years and allowing me to undertake this project and, whose support and advice has been much appreciated throughout.

I would also like to thank my colleagues in the lab: Francesca Carlisle, Josh Eaton, Christopher Estell, Laura Francis and Ryan Kelly. I would like to give a special mention to Laura and Steve for generating the DIS3-AID and XRN2-AID cell lines respectively, to which many of the bioinformatics analysis were based. Thanks also to Karen Moore and Audrey Farbos who run the sequencing facility at Exeter University and kindly shared their expertise and lab space during preparation of the RNA-Seq data.

I would like to give a big thank you to my family and friends back in sunny Scotland for their support and encouragement which has helped me greatly in reaching this point in my career.

Last and most importantly, I would like to show gratitude to James, whom this thesis is dedicated, for having such a huge influence on my life and the one to blame for setting me on the path to geekdom!

Cheers peeps.

Table of Contents

Table of contents I

Index of Figures VII

Index of Tables XII

Abstract XIII

Abbreviations XIV

Chapter 1: Introduction 1

1.1 Transcription by RNA Pol II 1

1.1.1 Initiation 2

1.1.2 Elongation 3

1.1.3 Termination 3

1.1.4 Gene Punctuation 5

1.2 Co-transcriptional RNA Processing 5

1.2.1 Capping at the 5' end 7

1.2.2 Splicing 7

1.2.3 Cleavage and Polyadenylation 8

1.3 Origins of Pervasive Transcription in Yeast 9

1.3.1 Chromatin Structure Promotes Pervasive Transcription 10

1.3.2 Spurious Transcription Overlapping Gene Flanks 11

1.3.3 Exosome Sensitive Cryptic Transcripts 12

1.4 Pervasive Transcription in Higher Eukaryotes 13

1.4.1 Hidden Transcription from Bidirectional Promoters 14

1.4.2 Promoter Associated Transcription 15

1.4.3 Enforcing Promoter Directionality 17

1.4.4 lncRNAs: A Diverse Catalogue of Pol II Transcripts 17

1.4.5 Enhancer RNAs 18

1.4.6 Biological Importance of lncRNAs 19

1.5 Nuclear Surveillance Pathways in Humans 20

1.5.1 Xrn2 and Dxo 20

1.5.2 The Exosome Complex 22

1.5.3 The Exosome Complex and Disease 24

1.5.4 Exosome Substrate Recognition by Mtr4 26

1.6 Functional Genomics through Exploitation of CRISPR/Cas9 Gene Engineering 27

1.7 The Auxin-Inducible Degron System in Plants 29

1.7.1 Harnessing AID in Non-Plant Cells 30

1.8 Project Aims 32

Chapter 2: Materials and Methods

2.1 Materials 34

2.1.1 Bacterial Strains 34

2.1.2 Tissue Culture 35

2.1.3 Vectors 37

2.1.4 Buffers 37

2.1.5 Molecular Biology Kits 39

2.1.6 RNA Sequencing Library Kits 39

2.2 Experimental Methods 40

2.2.1 Molecular Biology 40

2.2.2 Synthesis of Repair Template Plasmids 45

2.2.3 Western Blotting Assay 47

2.2.4 Northern Blot Analysis 48

2.2.5 Library Preparation of Nuclear RNA 49

2.2.6 Cell Biology 50

2.3 Bioinformatics Methods 54

2.3.1 Software Catalogue 54

2.3.2 RNA-Seq Read Alignment 55

2.3.3 Calculation of Genome Coverage and Depth 55

2.3.4 Differential Expression Analysis 55

2.3.5 Metagene Profiling 56

2.3.6 Read Enrichment over Genomic Elements 56

2.3.7 *De novo* Transcript Assembly 57

2.3.8 Determination of eRNA Directionality 57

2.3.9 Generation of Synthetic Intron Annotation 58

2.3.10 Histone Peak Calling from ChIP-Seq Analysis 58

Chapter 3: The Functional Role of Exosc10 in the Nucleus 60

3.1 Generating the AID Tagged EXOSC10 cell Line 62

3.1.1 Retroviral Integration of the Plant Specific TIR1 Gene 62

3.1.2 Modification of EXOSC10 by Exploiting HDR Templates 63

3.1.3 Validation of Genome Engineering by Western Blot Analysis 66

3.1.4 Identification of EXOSC10-AID Gene by Genomic DNA Screening 66

3.2 Depletion of the Exosc10-AID Protein is Rapid 67

3.3 The AID Tag Doesn't Interfere with Exosc10 Function 70

3.4 Exosc10 is Essential for Cell Viability 72

3.5 Catalytic Activity of Exosc10 is Dispensable for Cell Survival 74

3.6 Transcriptome-Wide Determination of Exosc10 Substrates 80

3.6.1 Global Differential Gene Expression 80

3.6.2 Precursor snoRNA Processing 86

3.6.3 Is snoRNA Maturation a 2-Step Exosome Process? 89

3.7 Summary 91

Chapter 4: Dis3 Prevents the Accumulation of Pervasive Transcripts 93

4.1 PROMPT Transcripts are Substrates of Dis3 94

4.1.1 PROMPT Transcription is Detectable Following Dis3 Depletion 95

4.1.2 Dis3 Stabilises Promoter Proximal Transcripts in the Coding Direction 100

4.2 Gene Expression is unaltered by Dis3 Depletion 102

4.2.1 Differential Gene Expression Analysis 102

4.2.2 False Discovery of Differential Expressed Genes 102

4.3 Dis3 Degrades RNA Derived from Premature Transcription Termination 108

4.4 Premature Termination of RNA Pol II Generates Small Dis3 RNA Substrates 113

4.5 Dis3 Downregulation Stabilises Transcripts Originating from Intergenic Sequences 115

4.5.1 Detection of Unannotated Intergenic Transcripts 115

4.5.2 Characterisation of Potential Novel eRNA Transcripts 116

4.5.3 Identification of Enhancer Sequences using Histone Modifications 120

4.5.4 Novel Intergenic Transcripts have eRNA-like Properties 123

4.6 Summary 126

Chapter 5: XRN2 Enhances Transcription Termination at Gene 3' Ends 130

5.1 Xrn2 Degrades 3' Flanking RNA downstream of the TES 131

5.2 Xrn2 is not responsible for Transcription Termination of Histone and snRNA Genes 138

5.2.1 Stabilised 3' Flanking RNA is Absent Downstream of Histone Genes 138

5.2.2 snRNA Genes are Efficiently Terminated Independently of Xrn2 Depletion 140

5.3 Xrn2 Downregulation has a Minimal Impact on Nascent RNA Expression 142

5.3.1 Differential Gene Expression Analysis 142

5.3.2 Failure to Terminate Transcription Causes Accumulation of 3' Flanking RNA over Neighbouring Genes 144

5.4 truncRNA Transcripts are not Xrn2 Substrates 146

5.5 *De novo* eRNA-like Transcripts are not Degraded by Xrn2 148

5.6 Summary 150

Chapter 6: Discussion 153

Future Work 163

References 165

Supplementary Figures 185

Appendix 197

Index of Figures

Chapter 1: Introduction

Figure 1.1: Co-transcriptional RNA Processing 6

Figure 1.2: Products of RNA Pol II transcription 16

Figure 1.3: Overview of RNA Degradation by the Nuclear Surveillance Pathway in Humans 23

Figure 1.4 Schematic of CRISPR/Cas9 Gene Engineering 28

Figure 1.5: The Auxin-Inducible Degradation Pathway in Plants 31

Chapter 2: Materials and Methods

Figure 2.1: Graphical Representation of Vector Maps used to Engineer the HCT116 Cell Lines 46

Figure 2.2: Workflow of Nuclear RNA Extraction, RNA Screening and RNA Seq Library Preparation 51

Chapter 3: The Functional Role of EXOSC10 in the Nucleus

Figure 3.1: The *Sleeping Beauty* Transposon Delivery Mechanism 61

Figure 3.2: Method of Tagging EXOSC10 with AID by HDR 64

Figure 3.3: Schematic of P2A Cleavage Generating two Distinct Proteins 65

Figure 3.4: Western Blot Screening of Positive EXOSC10-AID Colonies 68
Figure 3.5: Time Course Analysis of Exosc10-AID Depletion Efficiency 69
Figure 3.6: Determination of Exosc10-AID Protein Function 71
Figure 3.7: EXOSC10 is Essential for Colony Formation 73
Figure 3.8: Diagram of Exosc10 Protein Structure 76
Figure 3.9: Western Blot Screening of WT and D313A Exosc10 Overexpression EXOSC10-AID Cell Lines 77
Figure 3.10: Analysis of the Catalytically Inactive D313A Exosc10 Protein 79
Figure 3.11: Graphical Representation of EXOSC10-AID RNA-Seq Genomic Coverage, Depth and Mapping Efficiency 82
Figure 3.12: MA Plot of Differentially Expressed Genes in Exosc10-AID Null Cells 83
Figure 3.13: RNA-Seq Coverage Tracks of Cytochrome P450 Gene Upregulation after Exosc10-AID Depletion 85
Figure 3.14: Read Coverage Tracks of SnoRNA Genes 87
Figure 3.15: Read Coverage Tracks of SnoRNA Genes in Dis3-AID Depleted Cells 88
Figure 3.16: Comparison of *SNORA68* Processing Defects After Exosc10 and Dis3 Depletion 90

Chapter 4: Dis3 Prevents the Accumulation of Pervasive Transcripts

Figure 4.1: Western Blot Analysis of Dis3 Protein Depletion in the DIS3-AID Cell Line 97

Figure 4.2: Graphical Representation of DIS3-AID RNA-Seq Genomic Coverage, Depth and Mapping Efficiency 98

Figure 4.3: Visualisation of Upstream Stabilised PROMPT Transcripts 99

Figure 4.4: Metagene Analysis of Upstream Transcription 101

Figure 4.5: DIS3-AID Differential Gene Expression Analysis 104

Figure 4.6: False Positive Differentially Expressed Gene Visualisation 105

Figure 4.7: Adjusted Differential Gene Expression Analysis 107

Figure 4.8: Stabilisation of TSS Proximal Intron Sequences 111

Figure 4.9: Increased Coverage over Intron 1 Caused by Dis3 Depletion 112

Figure 4.10: Loss of Dis3 Stabilises Prematurely Terminated truncRNA Transcripts 114

Figure 4.11: Dis3 Degrades RNA Originating from Intergenic DNA Sequences 117

Figure 4.12: Annotated eRNA Transcripts Lie in Close Proximity to a Minority of *de novo* Intergenic Transcripts 119

Figure 4.13: ChIP-Seq Determination of Novel Intergenic Transcript Identity 122

Figure 4.14: Comparison of Novel eRNA-like and *de novo* PROMPT Transcripts 125

Chapter 5: XRN2 Enhances Transcription Termination at Gene 3' Ends

Figure 5.1: Graphical Representation of XRN2-AID RNA-Seq Genomic Coverage, Depth and Mapping Efficiency 135

Figure 5.2: Sequencing Coverage Analysis Downstream of the TES Following Xrn2 Depletion 136

Figure 5.3: Average Transcription Profiles of non-overlapping Annotated Genes 137

Figure 5.4: Histone Gene Read-through Analysis 139

Figure 5.5: Analysis of snRNA Transcription Read-though 141

Figure 5.6: Xrn2 Depleted Differential Gene Expression Analysis 143

Figure 5.7: Transcription beyond the Termination Window Overlaps Nearby Genes 145

Figure 5.8: Abortive truncRNA Abundance in Xrn2 Depleted Cells 147

Figure 5.9: eRNA-like Expression Analysis in Xrn2 Depleted Cells 149

Chapter 6: Discussion

Figure 6.1: Reviewed Model of Nuclear RNA Surveillance Pathways in Human Nuclei During early RNA Biogenesis 157

Figure 6.2: Proposed Model of snoRNA Maturation 158

Supplementary Figures

Figure S1: Genomic DNA Nested PCR Screening of EXOSC10-AID Cells 185

Figure S2: Tracks of Raw RNA-Seq Reads Depicting AID incorporation at the EXOSC10 Gene 3' End 186

Figure S3: Average Colony Size in each EXOSC10-AID Cell Line used During the Colony Formation Assays 187

Figure S4: Additional Replicate Metagene Plot of PROMPT Transcription in Exosome Depleted Cells 188

Figure S5: Total Synthetic Intron Differential Expression MA Plot 189

Figure S6: Accumulation of Dis3-sensitive truncRNAs 190

Figure S7: Additional Biological Replicate eRNA Metagene Plot 191

Figure S8: Second Biological Replicate Coverage Track Analysis of Read-through RNA after Xrn2 Depletion 192

Figure S9: Metagene Profile of Downstream Read-through 193

Figure S10: Replicate 2 Coverage Analysis of Histone Clusters in Xrn2 Depleted Cells 194

Figure S11: snRNA Read-through Analysis 195

Figure S12: Comparison of Dis3 and Xrn2 Stabilised truncRNAs 196

Index of Tables

Chapter 3: The Functional Role of EXOSC10 in the Nucleus

Table 3.1: List of Differentially Expressed Genes in Exosc10-AID Depleted Cells 84

Table 3.2: Potential Pathways Associated with Differentially Expressed Genes 84

Abstract

The bulk of the eukaryotic genome is pervasively transcribed, the largest proportion of which represents a diverse collection of non-coding RNA transcripts. Nuclear surveillance pathways play an integral role in regulating the expression of pervasive transcripts and protect the integrity of the transcriptome by engaging the activity of several nuclear exoribonucleases. In human nuclei, Xrn2 degrades RNA with 5'→3' directionality, whereas the exosome complex contains two catalytic subunits: Exosc10 and Dis3 capable of 3'→5' RNA decay. Functional studies of nuclear surveillance pathways in the past used RNA interference (RNAi) mediated protein depletion. Although informative, RNAi requires prolonged periods of gene downregulation which can often be incomplete and introduces indirect effects, further obfuscating the immediate function of nuclear exoribonucleases. The rise in popularity and efficiency of CRISPR/Cas9 mediated gene editing have stimulated a renaissance in the field of functional genomics for those wishing to apply a more direct approach within human models. Combining CRISPR/Cas9 with a post-translational degron depletion system, human cell lines can be engineered to undergo rapid, conditional and reversible downregulation of gene expression. As such, three auxin-inducible degron HCT116 cells lines were generated in this study with the aim to dissect the three major exonucleases: Xrn2, Dis3 and Exosc10. High-throughput RNA sequencing of nuclear transcriptomes in each scenario have identified distinct substrates for each exoribonuclease and new layers of gene regulation. The data presented hereafter highlights the extent of pervasive transcription and the separate nuclear surveillance pathways available within human nuclei.

Abbreviations

AID	Auxin-Inducible Degron
AUX	Auxin
bp	base pairs
BSA	Bovine Serum Albumin
cDNA	complimentary DNA
ChIP	Chromatin Immunoprecipitation
CMV	Cytomegalovirus
CPSF	Cleavage/Polyadenylation Specificity Factor
CRISPR	Clustered Regularly Interspaced Short Palindromic Repeats
CstF	Cleavage Simulation Factor
CTD	C-Terminal Domain
CUTs	Cryptic Unstable Transcripts
dH_2O	Distilled water
DMEM	Dulbecco's Modified Eagle Media
DOC	Sodium Deoxycholate
DoGs	Downstream of Gene transcripts
DSBs	Double-Stranded Breaks
D313A	Catalytically inactive EXOSC10 GAC → GCC at amino acid residue 313
eRNA	enhancer RNA
FCS	Foetal Calf Serum

gRNA	guide RNA
IAA	indole-3-acetic acid
kb	kilobase
kDa	kilodalton
lincRNA	long intergenic non-coding RNA
lncRNA	long non-coding RNA
miRNA	microRNA
ml	millilitres
mm	millimetres
mM	millimolar
mRNA	messenger RNA
NEXT	Nuclear EXosome Targeting complex
NFR	Nucleosome Free Region
nM	nanomolar
ng	nanogram
nt	nucleotides
ORF	Open Reading Frame
PABPN1	Nuclear poly(A) binding protein 1
PAM	Protospacer Adjacent Motif
padj	adjusted p-value
PAP	Poly(A) Polymerase
PIC	Pre-Initiation Complex
Pol	RNA polymerase
PROMPTs	PRoMoter uPstream Transcripts

RISC	RNA Induced Silencing Complex
RNAi	RNA interference
RNA-Seq	RNA sequencing
rpm	revolutions per minute
rRNA	ribosomal RNA
SB	Sleeping Beauty
SCF	Skp1, Cullin1 and F-box complex
snRNA	small nuclear RNA
snoRNA	small nucleolar RNA
TES	Transcription End Site
TIR1	Transport Inhibitor Response 1
TRAMP	Trf4/5-Air1/2-Mtr4 Complex
tRNA	transfer RNA
truncRNA	truncated nascent RNA
TSS	Transcription Start Site
UTR	Untranslated Region
v/v	volume/volume
WT	wild-type
w/v	weight/volume
x g	relative centrifugal force
XUTs	XRN1-sensitive Unstable Transcripts
µg	microgram
µl	microlitre
µm	micrometer

Chapter 1

Introduction

The hallmark of gene expression in eukaryotes involves transcription of DNA into RNA by RNA polymerases and for protein-coding genes, translation in the cytoplasm. RNA polymerase I and III are responsible for transcribing ribosomal RNA (rRNA) and transfer RNA (tRNA) respectively, whereas RNA polymerase II transcribes many classes of functional non-coding RNA (ncRNA) such as small nuclear RNA (snRNA), small nucleolar RNA (snoRNA), and micro RNA (miRNA), in addition to transcribing protein-coding messenger RNA (mRNA). Despite the importance of both tRNA and rRNA in gene expression, generally there has been a greater focus of the products of RNA polymerase II transcription, in part due to their capacity to encode protein. Similarly, many aspects of transcription are highly conserved between yeast and metazoans, but the general purpose of this introduction will be focused on human cell lines unless otherwise stated. This thesis focusses on transcripts produced by Pol II.

1.1 Transcription by RNA Pol II

RNA Polymerase II (hereafter referred to as Pol II) is a large multi-subunit complex composed of 12 core subunits, the largest of which, Rpb1, possesses catalytic activity. Additionally, the C-terminal domain of Rpb1 (hereafter referred to as CTD) contains numerous tandem heptad repeats of the amino acid sequence YSPTSPS, of which there are 26 in yeast and 52 in humans. The CTD can undergo dynamic and extensive post-translational modifications mediated by cyclin-dependent kinases (CDKs) at different stages of transcription and, the combination of CTD modifications coordinates the recruitment of a myriad of protein

complexes to the transcription apparatus (Buratowski 2009; Hsin & Manley 2012). Phosphorylation occurs mainly on residues Serine 2 (Ser2) and Ser5, however phosphorylation of Tyrosine 1 (Tyr1), Threonine 4 (Thr4) and Ser7 have also been observed (Heidemann *et al* 2012). Tight regulation of this 'CTD code' dissects mRNA transcription into three key stages; initiation, elongation and termination (Heidemann & Eick 2012) (**Figure 1.1**).

1.1.1 Initiation

In order for transcription to proceed several conditions must be met, chiefly the recruitment of Pol II and dozens of transcription factors (TFs) to DNA promoter regions and an open chromatin architecture. For initiation, five general TFs recognise and bind to the TATA-box domain located ~30 nucleotides (nt) upstream of the transcription start site (TSS), assembling into the pre-initiation complex (PIC). Pol II binding to the PIC follows and in turn acts as a platform for the recruitment of numerous accessory TFs that facilitate duplex template DNA unwinding (Jonkers & Lis 2015; Sainsbury *et al* 2015) and initiation of RNA synthesis.

The CTD of Pol II is unphosphorylated preceding assembly onto the promoter DNA. During transit through the promoter region, CDK7 catalyses phosphorylation of the CTD at residue Ser5 (Ser5-P) triggering early elongation, which is halted shortly thereafter at a promoter-proximal pause site located ~20-60 nt downstream of the TSS (Adelman & Lis 2012; Kwak & Lis 2013). Arresting Pol II at this early checkpoint allows the addition of a 5' cap on the nascent RNA and the phosphorylation of the CTD serine 2 (Ser2-P) residue by CDK9. In turn, Ser2-P modification reorganises transcription factors associated with Pol II and recruits numerous processing factors needed to release stalled Pol II. The doubly phosphorylated Ser2/5-P therefore marks the escape of promoter-proximal paused Pol II during the transition from early to productive elongation.

1.1.2 Elongation

After escaping the promoter, levels of Ser5-P begin to drop within a few hundred nucleotides of the TSS and remain low throughout the gene body. Conversely, Ser2-P accumulates steadily reaching a peak toward the 3' end of the gene. Ser2-P orchestrates the recruitment of several RNA processing complexes and additionally plays in integral role in transferring phosphorylation patterns into epigenetic marks, coordinated by histone chromatin modifying enzymes recruited to the CTD. This interplay provides positional awareness to chromatin remodelling complexes which dynamically rearrange the positions of nucleosomes and help to maintain a chromatin landscape conducive to productive elongation. Remodelling chromatin in such a way has the added advantage of regulating the rate of elongation and provides a mechanism of reengaging other Pol II complexes in successive rounds of transcription (Buratowski 2009).

1.1.3 Termination

Termination of transcripts produced by Pol II vary depending on the biotype of the nascent RNA. Despite their differences, non-coding RNAs such as snRNA as well as coding mRNAs require the CTD of Pol II to coordinate 3' end formation prior to termination (Hsin & Manley 2012).

In the case of protein-coding mRNAs which have been studied extensively, Ser2 hyperphosphorylation at gene 3' ends enhances the recruitment of cleavage and polyadenylation (CPA) factors to the pre-mRNA at Pol II pause sites downstream of the polyadenylation (poly[A]) site (Gromak *et al* 2006). Assembly of the cleavage/polyadenylation specificity factor (CPSF) and cleavage stimulation factor (CstF) complexes are reciprocally connected to phosphorylation of Ser2 in a manner that is independent of their catalytic activity (Davidson *et al* 2014). Additionally, poly(A) site pausing of Pol II at the 3' end has been demonstrated to facilitate the selection of alternative 3' ends in genes containing multiple

poly(A) sites (Fusby *et al* 2016). At this stage, CPSF and CstF complexes recognise and bind the consensus AAUAAA hexamer and G/U rich sequences of the poly(A) site (Proudfoot 2011), catalysing co-transcriptional cleavage of the nascent RNA by the CPSF73 endonuclease. Cleavage of the pre-mRNA releases the upstream transcript from the site of transcription and provides an entry site for polyadenylation factors to the 3' end. However, the free 5'-P associated with the downstream transcript becomes a substrate for the 5'→3' nuclear exoribonuclease Rat1 in yeast (Kim *et al* 2004), or Xrn2 in humans (West *et al* 2004) and is rapidly degraded. Coined as the "torpedo" model, Rat1/Xrn2 degradation of the downstream RNA chases the elongating Pol II complex, their collision releases Pol II from the template DNA triggering transcription termination by an unknown mechanism.

An alternative but not mutually exclusive "allosteric" termination model has also been described whereby, transcription of the poly(A) site triggers a conformational change slowing the elongation rate of the Pol II complex causing it to dissociate from the template DNA independently of cleavage (Osheim *et al* 2002). Furthermore, there is evidence that both the allosteric and torpedo termination models exist in higher eukaryotes, as depletion of Xrn2 merely delays transcription termination by a few thousand nucleotides (Fong *et al* 2015). Therefore, both termination pathways may act redundantly to prevent transcription read-through into downstream genes.

While this mechanism is applicable to the termination of polyadenylated transcripts, a subset of protein-coding transcripts produced by Pol II lack a poly(A) tail. Instead, replication-dependent histone RNA termination relies on the recognition of different *cis* elements on the nascent RNA namely, a stem-loop structure and a purine rich Histone Downstream Element (HDE) by the stem-loop binding protein (SLBP) and U7 snRNA (forming the U7 small ribonucleoprotein [snRNP]) respectively (Dominski *et al* 2007). Similar CPA factors (including CPSF and

CstF) are then directed to the 3' end of the histone RNA preceding cleavage by CPSF73 and termination of transcription.

1.1.4 Gene Punctuation

Combined, these three phases of transcription serve to punctuate genes by defining a clear start and end site. Through strict regulation of RNA Pol II activity, owing to the dynamic alteration of the CTD code, recruitment of RNA specific processing complexes and chromatin chaperones to the site of transcription can be temporally regulated at each stage of transcription, which ultimately helps to contain transcription events within the boundaries of defined gene intervals.

1.2 Co-transcriptional RNA Processing

Nascent RNA produced by Pol II must undergo extensive processing into usable mature RNA transcripts. Unsurprisingly, the proximity of the CTD to the emerging transcript and the plasticity of the post-translation modifications available function as a binding platform for numerous RNA processing factors. By controlling the traffic of processing factor engagement, the CTD intimately couples RNA processing with the kinetics of transcription (Bentley 2014). While processing steps differ for ncRNA and histone transcripts, most of the focus of co-transcriptional processing has be directed towards premature mRNA (pre-mRNA) maturation. Correct processing of pre-mRNA involves capping of the 5' end, excision of intronic sequences together with ligation of exons by splicing, and the formation of the 3' end via cleavage and polyadenylation before they are exported to the cytoplasm.

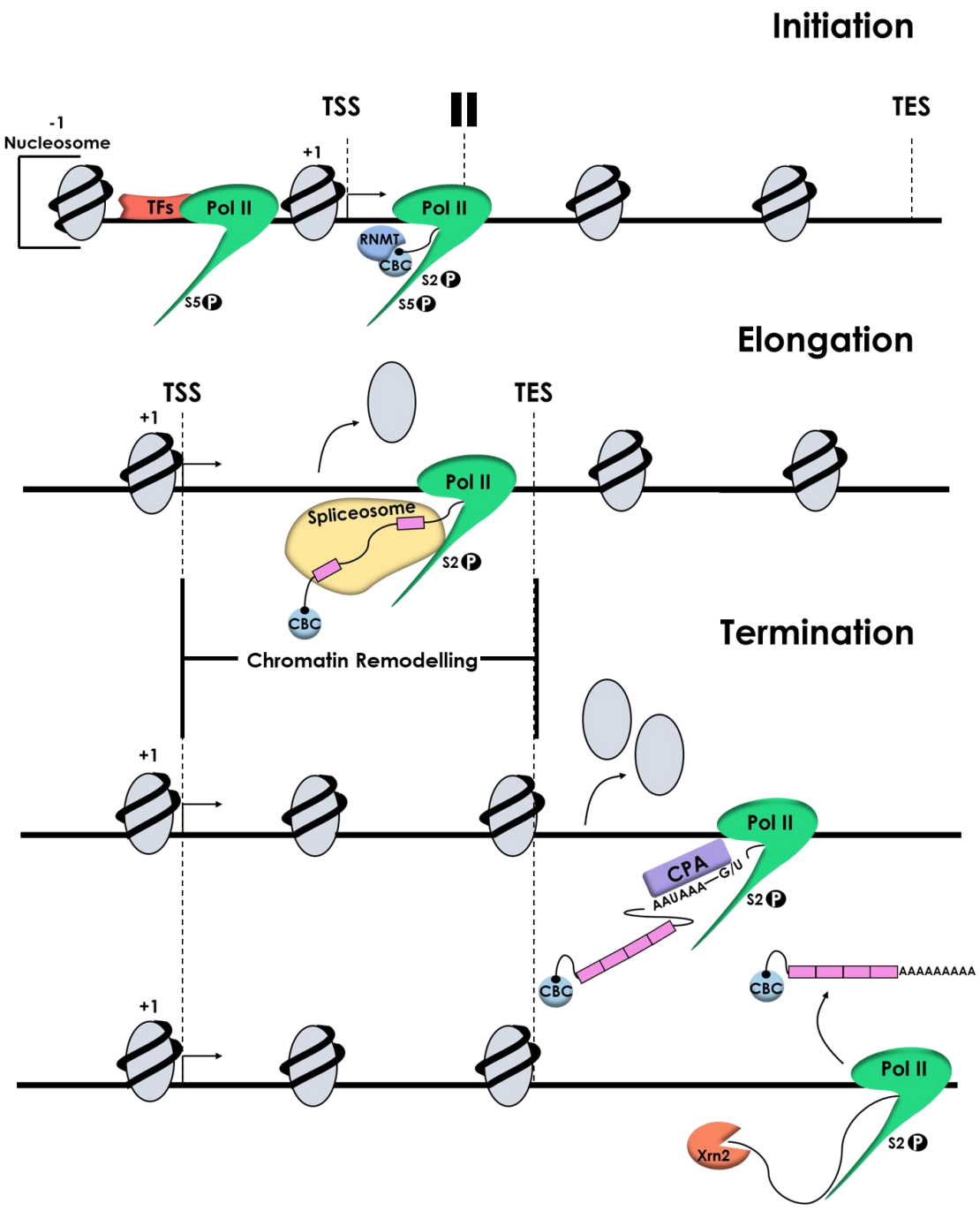


Figure 1.1: Transcription by RNA Pol II proceeds in 3 key stages: initiation, elongation and termination. During each stage, the phosphorylation status of the CTD helps to overcome proximal TSS pausing ("Pause" symbol) and coordinates the recruitment of several RNA processing factors. Additionally, the CTD code provides positional awareness to chromatin remodelling enzymes that dynamically reshuffle histones to facilitate DNA unwinding and maintain a chromatin landscape conducive to transcription elongation.

1.2.1 Capping at the 5' end

Capping takes place during the initial stages of transcription as the nascent RNA emerges from the active cleft of Pol II. A 7-methylguanosine cap is attached to the 5' end of the short 20-30 nt transcript by RNA guanine-7-methyltransferase (RNMT) in humans (Lewis & Izaurralde 1997). Assembly of the cap-binding complex (CBC) to the capped RNA protects it from 5'→3' degradation as well as facilitating mRNA export. RNMT recruitment is enhanced during Pol II stalling at promoter-proximal pause sites and has been shown to interact with either Ser2-P or Ser5-P modified CTD. The guanylyltransferase activity of RNMT however is stimulated by increasing levels of CTD Ser5-P which are typically found during transcription initiation (Ho & Shuman 1999). Pre-mRNA capping is an important quality control (QC) checkpoint that assesses the viability of the nascent RNA before committing Pol II to productive elongation. Additionally, cap formation on pre-mRNA is a reversible co-transcriptional step whereby, decapping can lead to premature transcription termination and co-transcriptional degradation of aberrant transcripts by Xrn2 (Davidson *et al* 2012).

1.2.2 Splicing

Intronic sequences are excised from pre-mRNA and exons are ligated together in a two-step transesterification reaction catalysed by the large multi-subunit spliceosome complex. Core splicing factors that make up the spliceosome are highly conserved between yeast and metazoans, however, due to the complexity of higher eukaryotic alternative splicing, the repertoire of splicing factors present in higher eukaryotes is much greater (~2-fold). For splicing to occur, five conserved snRNPs U1, U2 and the trimer U4/U5/U6 (each containing an snRNA of the same name) recognise canonical *cis* 5' and 3' splice site (ss) sequences (GU/AG respectively) flanking introns, as well as the branch point sequence (BPS) located ~20-40 nt upstream of the 3'ss. In humans, a polypyrimidine tract

located between the BPS and 3'ss is also required. Commitment to splicing following the initial binding of U snRNP molecules triggers a cascade of splicing factor recruitment to the pre-mRNA and the assembly of an active spliceosome complex. Rearrangements of the spliceosome then catalyse intron excision and exon ligation (Herzel *et al* 2017).

Binding of the U snRNPs is dependent on the transcription of 5'ss and 3'ss of the nascent RNA. This is therefore a rate-limiting step of splicing which is intrinsically linked to the elongation rate of Pol II. By altering the speed of Pol II traversal across the gene, splicing can be fine-tuned to incorporate or reject alternative exon sequences providing a mechanism of generating multiple transcript isoforms from a single gene (Bentley 2014). For example, reducing the speed of Pol II elongation in humans has the potential to incorporate non-conserved exons with weaker splice sites (deviating from consensus sequences) into the growing transcript, whereas faster Pol II elongation may result in skipping of non-consensus exons.

As previously mentioned, chromatin structure is integral to regulating elongation and it has now been demonstrated that nucleosome density is greater within exon sequences relative to intronic regions, which help to slow or pause Pol II over exons allowing time for splicing to occur (Saldi *et al* 2016). This "window of opportunity" as described by Saldi *et al*, postulates that splicing of certain exons requires a specific Pol II elongation rate in order for their incorporation into mature mRNA that is governed by dynamically reshaping the chromatin landscape.

1.2.3 Cleavage and Polyadenylation

Finally, it has already been mentioned that CPA factors recruited to transcript 3' ends catalyse cleavage of the pre-mRNA during transcription termination. Following cleavage by CPSF, the pre-mRNA 3'-OH is polyadenylated by poly(A) polymerase (PAP) which assembles (alongside

~80 polyadenylation factors in humans) with the core CPSF/CstF CPA factors co-transcriptionally (Shi *et al* 2009; Xiang *et al* 2014). Shortly after the generation of the free 3'OH, a short non-templated poly(A) tail is incorporated slowly by PAP due to its relatively low affinity to the nascent transcript. As the poly(A) tail grows, nuclear poly(A) binding protein 1 (PABPN1) coats the poly(A) tail stimulating PAP activity. Interactions between PABPN1 and CPSF bound to the upstream poly(A) site control poly(A) tail length which in humans is typically ~250 nt. After which the interaction between CPSF and PABPN1 can no longer be supported and the polyadenylation apparatus dissociates from the mRNA (Kuhn *et al* 2009).

The majority of mRNA transcripts, snRNA and long non-coding RNA (lncRNA) transcripts cleaved by CPSF73 are polyadenylated and the length of the poly(A) tail has been shown to improve the efficiency of mRNA translation initiation as well as protecting transcripts from 3'→5' degradation. More recently, alteration of poly(A) tail length has been shown to influence the half-life of certain transcripts providing a mechanism of post-transcriptionally regulating gene expression (Eckmann *et al* 2011).

1.3 Origins of Pervasive Transcription in Yeast

The composition of the eukaryote transcriptome is much more diverse than previously envisaged, and as high-throughput sequencing technologies have continued to improve, the catalogue of RNA subtypes discovered has dramatically increased. Recent genome-wide studies estimate that ~75% of eukaryote genomes (Djebali *et al* 2012) is transcribed into RNA, the clear majority of which representing ncRNA transcripts. While many biotypes of ncRNAs documented are highly conserved with well characterised functions, a plethora of novel groups of long and short unannotated ncRNA transcripts have recently emerged however, their potential function remains elusive. It is now understood that

Pol II is responsible for the biogenesis of a broad range of ncRNAs originating as both by-products of normal coding gene expression, and from unannotated intergenic DNA elements. Originally termed "junk RNA", these novel transcripts are now more commonly referred to as pervasive transcripts due to their widespread occurrence within the transcriptome (Jacquier 2009; Jensen *et al* 2013).

Despite their extensive presence within eukaryotes, only a handful of uncategorised novel ncRNAs have identifiable roles in regulating biological processes such as: X chromosome inactivation by *XIST* (Pontier & Gribnau 2011), *7SK*-mediated transcription pausing at promoters (Quaresma *et al* 2016) and as hallmarks of several cancers such as *MALAT1* (Gutschner *et al* 2013). Categorising the remaining bulk of pervasive transcripts remains a challenge due to the ever-expanding catalogue of ncRNAs discovered. Moreover, ncRNA sequences are generally less conserved and exhibit relatively lower expression levels compared to coding transcripts, collectively casting doubt on their biological relevance. Nevertheless, since their discovery there has been renewed interest in the field in part, aided by the evolution of next generation sequencing technologies and the development of new initiatives setup to annotate emerging ncRNA biotypes.

Pervasive transcription therefore remains an exciting area of research that is encouraging a redefinition of our current concept of what constitutes a functional RNA within the transcriptome.

1.3.1 Chromatin Structure Promotes Pervasive Transcription

Chromatin architecture plays an important role in regulating gene expression. Dynamic rearrangements of chromatin structure through nucleosome repositioning, or chromatin modifications have been shown to significantly impact the efficiency of transcription by actively engaged Pol complexes (Jonkers *et al* 2014; Tanny 2014). Dysfunction of the chromatin landscape through downregulation of certain key chromatin

remodelling enzymes such as SPT6, which restores chromatin structure left behind by elongating Pol II, can stimulate the expression of short transcripts from internal cryptic promoter regions within actively transcribed genes (Kaplan *et al* 2003). Similarly, Set2 associates with elongating Pol II and methylates histone H3 at lysine residue 36 (H3K36) stimulating recruitment of the histone deacetylase Rpd3. Rpd3 prevents spurious transcription initiation within ORFs by stripping transcription initiation marks (e.g. H3K27ac) from cryptic promoters incorporated in the wake of elongating Pol II (Carrozza *et al* 2005).

In addition to restricting access of polymerases to genomic DNA, gene expression is also regulated through termination of transcribing polymerase complexes at the 3' end of genes. Sen1 is a helicase enzyme involved in terminating Pol II on numerous non-coding RNAs and the formation of snoRNA 3' ends. Using chromatin immunoprecipitation in combination with microarray analysis (ChIP-chip) in *S. cerevisiae*, Steinmetz *et al* (2006) were able to map the location of Pol II across the entire genome. Comparison of WT and Sen1 mutant strains revealed that efficient Sen1-dependent termination of transcription at the 3' end of ncRNA genes prevents read-though of Pol II and the expression of downstream intergenic sequences.

1.3.2 Spurious Transcription Overlapping Gene Flanks

Even under normal conditions, high-resolution microarray mapping of *S. cerevisiae* detected low level expression of antisense transcripts originating from the opposite strand of regions flanking known genes (David *et al* 2006). In contrast to transcription read-through products, many of these novel transcripts appeared to originate from independent transcription units overlapping 5' and 3' untranslated (UTR) domains. In addition to these findings, RNA-Seq analysis designed to globally map gene boundaries detected extensive transcription of intergenic regions of the yeast genome (Nagalakshmi *et al* 2008).

1.3.3 Exosome Sensitive Cryptic Transcripts

The most significant advancement in the identification of pervasively transcribed regions however, came from the combination of whole-genome microarrays with mutant strains defective in nuclear RNA degradation pathways. Deletion of RRP6 (the yeast homologue of the 3'→5' exoribonuclease nuclear exosome complex component EXOSC10) (rrp6Δ), dramatically stabilises short, capped and polyadenylated cryptic unstable transcripts (CUTs) (Wyers *et al* 2005; Davis & Ares 2006). Normally undetectable in WT strains, CUTs were demonstrated to be widely distributed over the yeast genome at intergenic loci closely associated with the promoters of active genes as well as originating from unannotated intergenic regions. Unlike mRNAs, which are co-transcriptionally cleaved at poly(A) sites and subsequently polyadenylated by PAP, CUT transcripts are terminated via the Nrd1-Nab3-Sen1 termination pathway (Thiebaut *et al* 2006) becoming substrates of degradative polyadenylation by the TRAMP (Trf4/5-Air1/2-Mtr4 polyadenylation) complex and are subsequently degraded by the exosome (LaCava *et al* 2005; Vanacova *et al* 2005; Carneiro *et al* 2007). Furthermore, Wyers *et al* also indicated that CUTs can accumulate as non-adenylated transcripts in strains either lacking Trf4 (*trf4*Δ) or expressing a catalytically inactive (*trf4*-236) poly(A) polymerase.

In agreement with earlier outcomes, mapping the distribution of CUTs in *S. cerevisiae* revealed a strong correlation of CUT initiation within nucleosome free regions (NFRs), a common characteristic of active gene promoters. In fact, a high proportion of CUTs present in *rrp*6Δ strains were shown to share a TSS within promoter sequences of active protein-coding genes, initiating transcription in both directions within a few hundred nucleotides of the TSS (Neil *et al* 2009; Xu *et al* 2009). Neil *et al* were also able to distinguish separate PICs responsible for driving transcript expression in both sense and antisense orientations from common TSS. Equally important to CUT transcription within promoters, NFRs involved in transcription termination are also present downstream of stop codons and

are a common source of CUT expression (Xu *et al* 2009). Together NFRs flanking genes at their 5' and 3' ends accounts for most of the CUT initiation observed in the absence of RNA degradation pathways. Divergent transcription reflects an inherent attribute of eukaryote promoters, with both groups suggesting that bidirectional promoters have a positive influence on gene expression by maintaining NFRs desirable for transcription initiation.

Finally, another class of regulatory ncRNAs degraded by the cytoplasmic 5'→3' exoribonuclease Xrn1 were also annotated in yeast. Like CUTs, Xrn1-sensitive unstable transcripts (XUTs) are capped and polyadenylated transcripts expressed from antisense Pol II initiation near ORFs (van Dijk *et al* 2011). Despite their apparent instability, XUTs have been strongly implicated in gene silencing.

The considerable abundance of ncRNA transcripts identified in *S. cerevisiae* highlights a previously hidden layer of complexity contradicting the relatively small and simple yeast genome, with the current repertoire of ncRNA transcripts described significantly outnumbering protein-coding transcripts.

1.4 Pervasive Transcription in Higher Eukaryotes

Shortly following genome-wide profiling of ncRNAs in *S. cerevisiae*, large-scale initiatives such as the ENCODE project designed to identify and characterise functional DNA sequences in the human genome were successful in generating a catalogue of transcripts from the 15 cell lines analysed (Djebali *et al* 2012). The ENCODE project estimated transcription of 74.7% of the human genome and despite expression levels an order of magnitude above ncRNA, primary transcription of protein-coding exon sequences account for < 3% of the transcriptome. Analogous to yeast, human transcriptomes are composed primarily of highly conserved classes of functional ncRNAs (e.g. snRNA, snoRNA, tRNA and miRNAs), in addition to low level expression of unannotated novel ncRNA transcripts

(**Figure 1.2**) which were separated into two broad groups: short ncRNA (< 200 nt) and long ncRNA (> 200 nt).

1.4.1 Hidden Transcription from Bidirectional Promoters

Like yeast, several species of unstable short (< 50 nt) "promoter-associated short RNA" (PASR) (Kapranov *et al* 2007) and "transcription start site-associated RNA" (TSSa-RNA) (Seila *et al* 2008) are divergently transcribed by Pol II close to the TSS of metazoan protein-coding genes. Mapping transcription initiation sites of TSSa-RNAs in mouse cell lines distinguished two separate peaks of RNA enrichment in both sense (between +0 and +50 nt) and antisense (between -100 and -300 nt) positions relative to the orientation of the coding gene TSS. To further classify TSSa-RNA, Seila *et al* performed chromatin immunoprecipitation DNA sequencing (ChIP-Seq) to examine the local chromatin landscape surrounding the TSS, detecting prominent peaks of trimethylated histone 3 lysine 4 (H3K4me3), an epigenetic mark linked to transcription initiation. Contrasting the twin H3K4me3 peaks, chromatin marks associated with active Pol II elongation (H3K79me2) were found to be excluded over regions upstream of the TSS and were found to be exclusively enriched over elongation regions in the same direction as productive gene transcription. Early termination of upstream elongation explains the unilateral distribution and low abundance of transcripts frequently observed flanking genic TSSs. Supporting this notion, global run-on sequencing (GRO-Seq) assays used to map and quantify the genome-wide distribution of Pol II in human fibroblasts, characterised bidirectional transcription from distinct, engaged polymerase complexes consistent with mouse models (Core *et al* 2008). The number of bidirectional promoters identified by GRO-Seq far exceeded previous estimates, the majority of which produce functional mRNAs, indicating that bidirectionality is a common feature associated with active promoters.

1.4.2 Promoter Associated Transcription

Bidirectional promoter transcription is responsible for the production of an unstable class of lncRNA. Dissimilar to TSSa-RNAs, PRoMoter uPstream Transcripts (PROMPTs) initiate from a narrow window (~0.5 and 2.5 kilobases [kb]) further upstream from associated TSS in an inverse orientation relative to the connected downstream gene and tend to be on average several hundred nucleotides in length, making them comparable to CUTs in yeast (Preker *et al* 2008; Neil *et al* 2009).

PROMPT transcripts form one division of a broad catalogue of lncRNA whose expression is closely associated with protein-coding gene promoters and are structurally similar to coding transcripts (Preker *et al* 2011). As such, PROMPTs are processed by the same transcription machinery as mRNAs since they undergo capping and are terminated via poly(A) site recognition and cleavage. Polyadenylation of PROMPT transcript 3' ends is catalysed by PAPD5 (formerly Trf4-2), a human homologue of the yeast TRAMP complex component Trf4. Reminiscent to CUTs, PROMPT transcripts are highly unstable due to their sensitivity to the nuclear exosome complex. Downregulation of key exosome genes such as EXOSC3 (hRrp40), DIS3 (hRrp44) and the nuclear specific EXOSC10 (hRrp6) have been demonstrated to stabilise PROMPT transcripts causing an accumulation of RNA covering sequences up to ~3 kb upstream of TSS (Preker *et al* 2008; Flynn *et al* 2011; Szczepinska *et al* 2015) which, under normal conditions remain undetectable. Interestingly, the polyadenylation of PROMPTs is not a strict requirement of exosome-mediated RNA turnover (Preker *et al* 2011; Ntini *et al* 2013) indicating that PROMPT turnover does not completely reflect CUT degradation pathways available in *S. cerevisiae*.

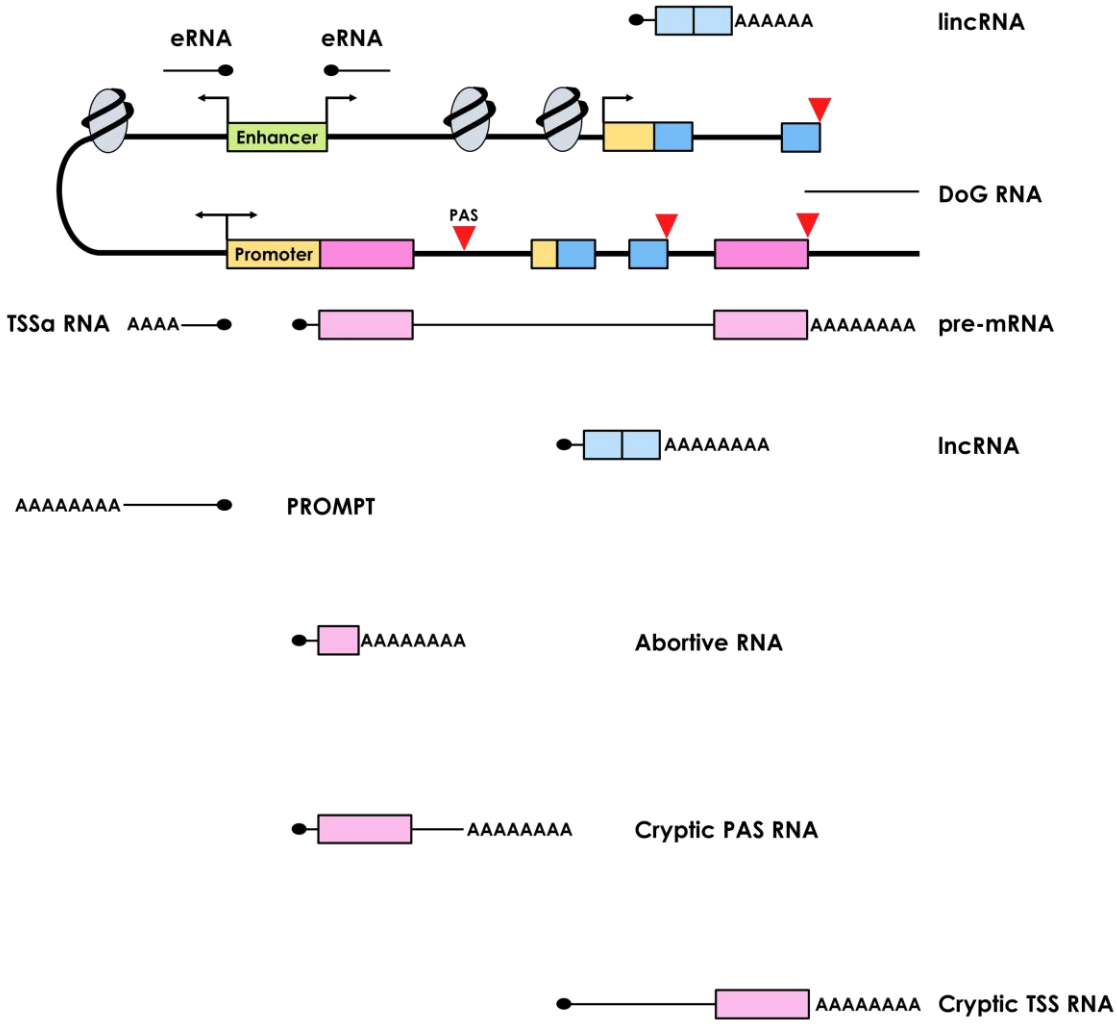


Figure 1.2: RNA pol II is responsible for the transcription of previously hidden, highly unstable groups of RNA in higher eukaryotes, many of which are the result of non-productive transcription of protein-coding genes generating transcription start-site associated RNAs (TSSa-RNA), PRoMoter uPstream Transcripts (PROMPTs) or prematurely abortive transcripts. Likewise, internal cryptic poly(A) sites (PAS) (red triangles) or TSSs can also produce shorter alternative transcript isoforms, in addition to the expression of long ncRNAs (lncRNAs) that overlap intronic sequences. Downstream of Gene RNAs (DoGs) can be produced from transcription read-through downstream of the termination site in response to osmotic shock. Transcription from intergenic loci is an additional source of enhancer RNAs (eRNAs) and long intergenic ncRNAs (lincRNAs).

1.4.3 Enforcing Promoter Directionality

Termination of PROMPT transcription at TSS-proximal poly(A) sites has been strongly implicated in regulating transcription directionality of bidirectional promoters (Ntini *et al* 2013) and transcriptionally active DNase hypersensitive sites (DHSs) (Andersson *et al* 2014). GRO-Seq profiles of human genes detected an asymmetric distribution of poly(A) sites surrounding promoters with a higher probability for Pol II to encounter a poly(A) site within ~500 nt downstream of the TSS (Ntini *et al* 2013). Equally, the likelihood of mRNA transcribing Pol II encountering a 5'ss is also much more favourable within the coding RNA sequence. Despite this, the higher abundance of poly(A) sites are silenced through recruitment of the U1 splicing factor to nearby 5'ss sequences protecting coding transcripts from premature termination (Kaida *et al* 2010; Berg *et al* 2012; Almada *et al* 2013). In contrast, PROMPT transcripts terminate shortly after transcription initiation and, in combination with exosome-mediated clearance of unwanted transcripts, a unidirectional output of productive RNA is enforced. This raises further questions as to how the exosome differentiates TSS-proximal poly(A) terminated transcripts.

1.4.4 lncRNAs: A Diverse Catalogue of Pol II Transcripts

The landscape of mammalian transcriptomes is rich in other lncRNA biotypes that are less sensitive to exosome-mediated decay, and like PROMPTs are structurally similar to mRNAs since many groups of lncRNA additionally contain exons. Splicing of lncRNA occurs through recognition of canonical splice-site signals flanking introns utilising the same splicing apparatus as mRNA, however lncRNA molecules show a remarkable propensity to containing only two exons per transcript (Derrien *et al* 2012). While intron and exon sequence are on average longer in lncRNA transcripts relative to mRNA, processed lncRNA transcripts are generally shorter due to their smaller exon number. Mapping the location of lncRNA biogenesis has been difficult in the past in part, due to their relatively low

expression rate and structural similarities to mRNA. Nevertheless, groups of lncRNAs were shown to partially or fully overlap genic sequences and have been sub-classified into separate biotypes based on their location relative to known protein-coding transcripts (Cabili *et al* 2011; Derrien *et al* 2012):

- **Exonic:** lncRNAs overlap coding gene exons transcribing RNA from the opposite strand relative to coding RNA transcription.
- **Intronic:** the lncRNA is located within an intron of a coding gene on the same strand and does not overlap exon regions.
- **Overlapping:** lncRNAs contain a protein-coding gene within an intronic region on the same strand.

To complicate annotation further, the majority of lncRNAs do not intersect protein-coding genes and instead initiate from independent loci at intergenic regions throughout the genome. By analysing changes in chromatin architecture associated with actively transcribed Pol II genes (e.g. H3K4me3 at active promoters or H3K36me3 over the transcribed gene body) in parallel with ChIP-Seq, long intergenic ncRNA (lincRNA) transcription was discovered at loci outside of annotated protein-coding genes in mouse (Guttman *et al* 2009) and human cell lines (Khalil *et al* 2009). In agreement, RNA-Seq analysis enhanced by ENCODE annotation found that a significant proportion of lincRNAs are transcribed from intergenic DNA elements commonly associated with trait inheritance and disease-associated single nucleotide polymorphisms (SNPs) (Cabili *et al* 2011; Hangauer *et al* 2013; Iyer *et al* 2015).

1.4.5 Enhancer RNAs

Finally, of particular interest is a class of lncRNA transcripts accompanying the transcription of enhancer DNA sequences which have not been described in yeast. Enhancers are distal regulatory elements that improve mRNA synthesis in a tissue specific manner (Kim *et al* 2010) that, when brought in to close proximity with active protein-coding genes remain

accessible to TFs due to their open chromatin architecture. Cap analysis of gene expression (CAGE), which can map capped RNAs with single nucleotide resolution, identified bidirectional synthesis of short (< 2 kb) capped, enhancer RNA (eRNA) (Andersson *et al* 2014) transcripts with no apparent directional bias compared to TSSs of coding genes. Moreover, there is little evidence of eRNA splicing or polyadenylation contrary to the majority of lncRNAs (Kim *et al* 2015). Analogous to PROMPTs, eRNAs are sensitive to exosome-mediated degradation irrespective of their transcript orientation (Andersson *et al* 2014).

There is still much debate about the importance of eRNAs in promoting coding-gene transcription, especially considering their low-level of expression and rapid turnover, which can easily be dismissed as by-products of nearby promiscuous Pol II transcription at accessible genomic loci. It is yet to be determined whether the products of enhancer transcription, or the act of transcribing these regions itself augments the expression of distal genes (Young *et al* 2017).

1.4.6 Biological Importance of lncRNAs

Investigation of lncRNA function has been met with considerable challenges in recent years, in part due to the large repertoire of transcripts discovered. As the complexity of the lncRNA landscape continues to rise, it is becoming increasingly important to catalogue transcripts with confirmed functions. While large scale initiatives such as GENCODE (Derrien *et al* 2012), NONCODE (Xie *et al* 2014) and FANTOM5 (FANTOM Consortium and the RIKEN PMI and CLST [DGT]) have generated comprehensive reference databases of thousands of lncRNAs, the function of less than ~10% have been comprehensively studied.

More recently, lncRNA transcripts have emerged as significant regulators of the transcriptome with broad roles in epigenetic gene silencing, as molecular scaffolds of protein complexes and paraspeckle formation or as regulators of alternative mRNA splicing (Mercer & Mattick

2013; Goff & Rinn 2015; St Laurent *et al* 2015). Additionally, many lncRNAs serve as precursors for smaller regulatory ncRNA such as miRNA and snoRNA. Importantly, the expression of numerous lncRNA transcripts overlapping disease risk loci have been exploited as early biomarkers of cancer and disease (Ulitsky & Bartel 2013; Iyer 2015). Given the importance of the handful of lncRNA characterised, there is a potential for uncovering biological functions to the vast majority of transcripts discovered.

1.5 Nuclear Surveillance Pathways in Humans

To deal with the volume of pervasive transcripts generated, multiple nuclear surveillance pathways have evolved in eukaryotes that monitor the transcriptional output of polymerases, degrading any nonsense or spurious transcripts produced (**Figure 1.3**). Importantly, various surveillance pathways can travel alongside elongating polymerases and mediate rapid co-transcriptional degradation, in addition to their post-transcriptional activity. Collectively, nuclear RNA surveillance pathways can act during the early stages of transcription preventing the accumulation of potentially deleterious transcripts.

1.5.1 Xrn2 and Dxo

The importance of the nuclear exoribonuclease Xrn2 has already been highlighted with regards to efficiently terminating Pol II complexes from the template DNA. However, the activity of Xrn2 is not restricted to the 3' end of genes, in fact Xrn2 interacts with the transcription apparatus during the early stages of transcription (Davidson *et al* 2012). A significant number of Pol II complexes prematurely abort transcription as a consequence of promoter-proximal pausing generating short nonsense capped RNAs. Since the activity of Xrn2 is largely restricted to the degradation of uncapped RNAs, abortive transcripts must undergo decapping prior to 5'→3' decay. In eukaryotes, Dcp2 is responsible for removing the cap from

the bulk of mRNA products, and while mainly cytoplasmic, Dcp2 can shuttle into the nucleus. Dcp2 interacts with both Xrn2 and transcription termination factors such as TTF2 to synchronise cap removal and degradation of abortive transcripts co-transcriptionally at the 5' end of the gene (Brannan *et al* 2012). Likewise, improperly capped or spliced transcripts are retained near the transcription start site and degraded by Xrn2. In this case, decapping of defective transcripts does not involve the activity of Dcp2 (Davidson *et al* 2012). In yeast there are two homologous proteins, Rai1 and Dxo1 that activate and partner with Rat1 (Xrn2) to remove incompletely capped transcripts prior to decay. While no Rai1 alternative has been detected in humans, a weak Dxo (Dom3Z) homologue exists. Dxo preferentially degrades partially capped transcripts as well as possessing 5'→3' exoribonuclease activity (Jiao *et al* 2013), however it is still unclear if Dxo functions in concert or independently of Xrn2.

More recently, Dxo has been shown to remove non-canonical 5' nicotinamide adenine dinucleotide (NAD) cap structures from a subset of mRNAs and ncRNAs in humans (Jiao *et al* 2017). Moreover, Jiao *et al* demonstrated that transcripts harbouring a 5' NAD cap do not undergo translation, instead the NAD cap promotes Dxo-mediated "deNADing" and degradation via the intrinsic exoribonuclease activity present within Dxo.

As previously mentioned, Xrn2 is responsible for degrading RNA downstream of poly(A) site cleaved transcripts. Downregulation of Xrn2 expression either by RNAi combined with overexpression of a catalytically inactive Xrn2 mutant, or through conditional AID protein depletion causes stabilisation of downstream RNA sequences (Fong *et al* 2015; Eaton & Davidson *et al* 2018). An added consequence of Xrn2 knockdown is the delayed termination of Pol II, which in some cases continues to elongate up to ~100 kb downstream of the typical "transcription window" potentially merging into neighbouring genes. Supporting this, ~10% of protein-coding genes accumulate downstream of gene (DoG)

sequences generated by transcription read-through beyond the poly(A) site in response to osmotic shock (Vilborg *et al* 2015).

1.5.2 The Exosome Complex

The exosome is a multi-subunit complex that is highly conserved in all eukaryotes capable of degrading a broad range of unstable or non-coding RNA substrates. It is composed of 9 core subunits (EXO-9) that form a catalytically inactive double ring-like structure and two 3'→5' exoribonucleases, Exosc10 and Dis3, the latter also possesses endonuclease activity (Mitchell *et al* 1997). Both Exosc10 and Dis3 occupy opposing positions at the entry and exit pores (respectively) of a central channel running through EXO-9. Although RNA substrates can be targeted directly to either exoribonuclease independently, the majority of RNA substrates are threaded into the central channel by Exosc10, which can widen the entry pore facilitating interaction between substrates and the active centre of Dis3 (Mitchell 2014; Kilchert *et al* 2016; Ogami *et al* 2018). Recently, characterisation of the C-terminal domain of yeast Rrp6 identified an RNA binding domain dubbed the "lasso". The lasso domain is an RNA binding platform proximal to the entry pore of the central channel that also stimulates the exo- and endonuclease activity of Dis3. Despite no apparent sequence conservation, properties of the lasso domain are apparent in human Exosc10 and other eukaryote equivalents (Wasmuth & Lima 2016).

The exosome is present in both the nucleus and cytoplasm however, Dis3 is predominately distributed in the nucleus but is restricted from the nucleolus. Two alternative cytoplasmic homologues, Dis3L1/Dis3L2 also exist in humans, the former comprises an inactive N-terminal PIN endonuclease domain which is absent in the latter (Robinson *et al* 2015). Furthermore, an additional Dis3 isoform expressing a shorter PIN domain, generated through alternative splicing of exon 2, was recently identified in human nuclei (Robinson *et al* 2018).

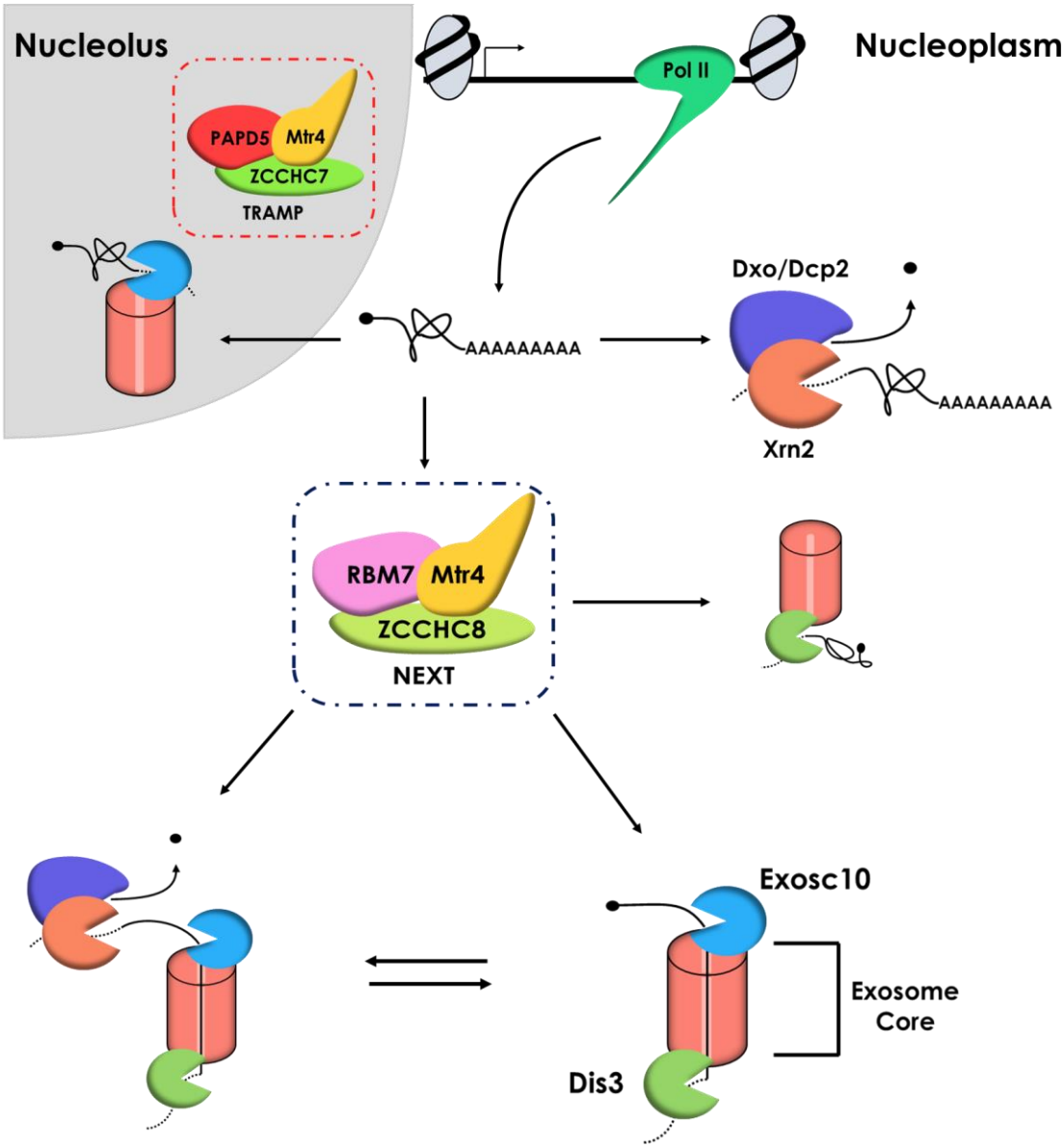


Figure 1.3: Overview of the RNA degradation routes available to the nuclear surveillance pathway in humans. Products of RNA pol II can be targeted to the exosome by the NEXT complex or by the TRAMP complex in the nucleolus for either 3'→5' trimming/maturation or degradation. Additionally, RNA transcripts can be decapped by either Dxo or Dcp2 (which can shuttle in from the cytoplasm) followed by 5'→3' mediated decay by Xrn2. Finally, the possibility of coupling between Xrn2 and exosome-mediated degradation also exists.

The distribution of Exosc10 is strictly nuclear with a particular enrichment in the nucleolus however, a small sub-fraction is also found in the nucleoplasm. Thus, three potential exosome complexes exist within the nucleus; nucleolar EXO-9/Exosc10, nucleoplasmic EXO-9/Dis3 and nucleoplasmic EXO-9/Exosc10/Dis3 (Lykke-Andersen *et al* 2011). The sub-cellular distribution of alternatively composed exosomes therefore provides a mechanism to specialise the functions of each exosome isoform in different cellular compartments (Kilchert *et al* 2016).

With regards to nuclear RNA surveillance, Dis3 has been reported as the main exoribonuclease degrading a broad range of pervasive RNA transcripts and represents the only decay pathway available for PROMPT RNA (Dziembowski *et al* 2007; Szczepinska *et al* 2015). In the more recent study, Dis3 was also proposed to degrade eRNA and snoRNAs as well as processing pre-rRNA during later nucleoplasmic maturation steps. Exosc10 on the other hand has a more prominent role in degradation and processing of small RNAs with more complex secondary structures, including pre-rRNA and snoRNAs (Januszyk *et al* 2011). Furthermore, tethering of aberrant pre-mRNAs near transcription foci by Exosc10 has also been described whereby, transcript integrity is assessed at an exosome-dependent checkpoint, potentially linking co-transcriptional exosome-mediated RNA degradation at gene 3' ends in a similar fashion to the role of Xrn2 at nascent 5' ends (Almeida *et al* 2010).

1.5.3 The Exosome Complex and Disease

It is unsurprising, given the importance of the exosome complex as part of nuclear RNA surveillance pathways in higher eukaryotes, that disruption of Exosc10 or Dis3 function has been associated with several forms of cancer and disease.

Exosc10 (also known as PM/Scl-100) was originally described as a target of an autoantibody present in patients suffering from a variety of systemic autoimmune diseases commonly associated with connective

tissues (Maes *et al* 2010). These include, polymyositis (PM), scleroderma (Scl) and PM/Scl overlap syndrome, which is the result of overlap with other connective tissue disorders. In each condition, anti-PM/Scl (formerly PM-1) antibody is directed towards the PM/Scl-100 (Exosc10) subunit of the exosome, supporting the notion of an autoimmune response directed against the exosome (Brower *et al* 2001). Despite the usefulness of this autoantibody as an early diagnostic marker, the precise clinical association between anti-Pm/Scl and PM/Scl-100 remains to be further elucidated (Mahler & Raijmakers 2007).

The cellular abundance of Dis3 is an important factor contributing to cell proliferation. Aberrant expression or the loss of Dis3 function has been associated with the progression of numerous forms of cancer, including multiple myeloma (MM), colorectal carcinoma and multiple forms of leukaemia such as chronic lymphocytic leukaemia (CLL), and acute myeloid leukaemia (AML) (Ng *et al* 2007; de Groen *et al* 2014; Robinson *et al* 2015). However, the precise role of Dis3 in cancer progression remains to be elucidated.

Interestingly, Dis3L2 (a cytoplasmic isoform that acts independently of the exosome complex) was the first example of a human disease related to dysfunction of the exosome (Morris *et al* 2013). Inactivation of Dis3L2 exoribonuclease activity leads to the progression of Perlman syndrome, a rare inherited overgrowth condition predominant in children, in which patients are at an increased risk of developing Wilms tumour. Transcriptome-wide analysis of cells expressing inactive Dis3L2 revealed a significant disruption to transcriptome homeostasis, in part caused by the accumulation of several classes of ncRNAs and stabilisation of non-functional extended snRNA transcripts (Labno *et al* 2016). The loss of Dis3L2 RNA surveillance was also shown to upregulate a small subset of mRNA transcripts that play a prominent role in the early development of *Drosophila* wing imaginal discs, causing significant wing overgrowth (Towler *et al* 2016).

RNA surveillance by Dis3 therefore has a significant impact in the control of cell proliferation, most likely as a consequence of its function in post-transcriptional gene regulation. Identifying the tissue-specific RNA substrates would therefore provide valuable insight into the progression of several forms of cancer and disease in humans, in addition to acting as potential clinical markers of disease onset.

1.5.4 Exosome Substrate Recognition by Mtr4

Substrate recognition and RNA unwinding are important early steps regulating exosome-mediated degradation, as they assist threading of the RNA into the central channel of EXO-9. Several factors can interact with the exosome regulating its activity (Fox *et al* 2016), most notable is the intimate interactions between Exosc10 and Mtr4 which co-purify at an almost equal proportion (Lubas *et al* 2011). Mtr4 is an RNA helicase that forms part of the aforementioned TRAMP complex in yeast (LaCava *et al* 2005; Vanacova *et al* 2005; Carneiro *et al* 2007) consisting of two zinc-knuckle binding proteins Air1/Air2 and a non-canonical poly(A) polymerase Trf4 (or in some cases Trf5). Close association between TRAMP and the exosome in *S. cerevisiae* recruits and facilitates RNA degradation through the addition of short poly(A) tails by a non-canonical poly(A) polymerase. Importantly, an Mtr4 homologue is present in humans and a TRAMP-like complex has also been described through identification of close orthologues. In humans ZCCHC7 and PAPD5 fulfil the roles of Air1/Air2 and Trf4/Trf5 respectively. Unlike yeast, the activity of the TRAMP-like complex is restricted to the nucleolus in humans in part due to the strict nucleolar localisation of ZCCHC7 (Lubas *et al* 2011). PAPD5 has been shown to polyadenylate snoRNA and pre-rRNA transcripts aiding exosome processing and/or degradation, consistent with the enrichment of TRAMP-like complexes in the nucleolus (Ogami *et al* 2018).

In humans, Mtr4 also associates with a second targeting complex located in the nucleoplasm. The trimeric nuclear exosome-targeting (NEXT) complex composed of Mtr4, the RNA binding protein RBM7 and zinc-knuckle protein ZCCHC8 are connected to the degradation of cryptic ncRNA transcripts such as PROMPTs and eRNAs (Kilchert *et al* 2016). RBM7 also binds mature and pre-snRNA isoforms suggesting a QC role during snRNA biogenesis (Hrossova *et al* 2015). Furthermore, NEXT complexes that associate with the ARS2-associated cap-binding complex (forming the CBCA complex) can stimulate transcription termination of PROMPT transcripts through Pol II stalling proximal to the 5' cap, bridging the interaction between the exosome and RNA Pol II, and significantly enhancing the degradation of abortive transcripts (Andersen *et al* 2013).

1.6 Functional Genomics through Exploitation of CRISPR/Cas9 Gene Engineering

Adaptation of the RNA-directed CRISPR/Cas9 technology has revitalised engineering of a wide array of metazoan genomes with a high degree of precision and ease. While CRISPR/Cas9 is a common feature among numerous strains of bacteria, most human optimised CRISPR/Cas9 technology has been adapted from *Streptococcus pyogenes*.

Bacteria can incorporate foreign viral genetic material into a clustered regularly interspaced short palindromic repeats (CRISPR) array within their own genome. The function of the CRISPR array is twofold; firstly, it acts as an immune memory of viral infections and secondly it protects the bacterium upon recurring viral invasion (Doudna & Charpentier 2014). The CRISPR cluster itself is composed of repetitive elements flanking exogenously inserted DNA known as protospacers. Each protospacer in the cluster is associated with a 3' protospacer adjacent motif (PAM).

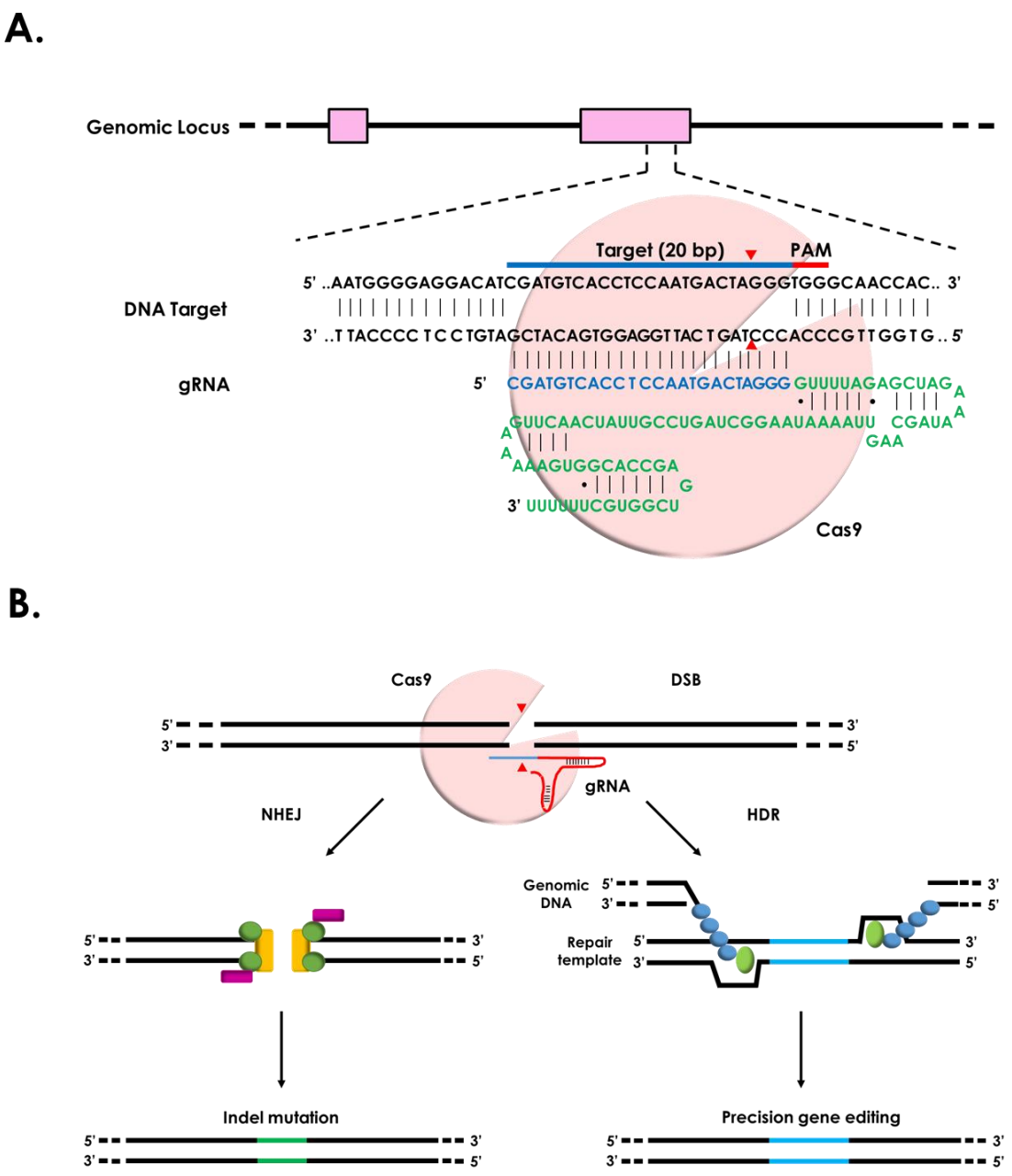


Figure 1.4: (**A**) Schematic representation of the Cas9 endonuclease (codon optimised for mammalian systems) from *S. pyogenes* which is targeted to genomic DNA using a complementary 20 nucleotide associated guide RNA (gRNA) sequence. Cas9 then cleaves both strands of the DNA (red triangle) upstream of the required PAM motif (5'-NGG). (**B**) Following cleavage, double strand DNA breaks (DSBs) are then repaired using either error prone non-homologous end joining (NHEJ) or high-fidelity homology-directed repair (HDR) pathways. This figure was adapted from Ran *et al* (2013).

The basis of CRISPR/Cas9 genome engineering involves the association of the 20 nt protospacer RNA, and the accompanying PAM sequence (more commonly referred to as a guide RNA [gRNA]) with the Cas9 nuclease forming the CRISPR/Cas9 complex. The gRNA directs CRISPR/Cas9 to a complementary DNA sequence where it catalyses DNA double-stranded breaks (DSBs) provided that a suitable PAM sequence is located 3' of the complementary DNA target site (**Figure 1.4**[**A**]). (Mali *et al* 2013). Targeting of the CRISPR/Cas9 complex to genomic loci is limited by the presence of a PAM sequence, which is unique for each Cas9 orthologue. For example, Cas9 in *S. pyogenes* requires a 5'-NGG PAM sequence, whereas 5'-NNNNGATT is required in *Neisseria meningitis* (Ran *et al* 2013; Komor *et al* 2017).

Following cleavage, DSBs are then repaired by one of two major pathways: error-prone non-homologous end joining (NHEJ) or high-fidelity homology-directed repair (HDR). Repairing DNA by NHEJ can cause the formation of insertion/deletion (indel) mutations, whereas HDR uses a repair template to ligate the DNA, although the frequency of HDR in metazoans is generally much lower in frequency compared to NHEJ (**Figure 1.4**[**B**]). In terms of gene editing, NHEJ repair is an effective and simple way of introducing random deleterious mutations into a genome and provides an effective way to study genetic variation. Alternatively, HDR is capable of introducing large alterations to the genome through the design of custom repair templates, which is often more time-consuming and labour intensive.

1.7 The Auxin-Inducible Degron System in Plants

Gene expression in eukaryotes can be regulated by controlling both the level of transcription at DNA and the abundance of mRNA. However, downregulation of either of these pathways is often slow, particularly for proteins with a long half-life. To overcome these limitations, several systems exist that downregulate gene expression by modifying proteins post-

translationally. All eukaryotes possess Skp1, Cullin1 and F-box proteins which form part of the E3 ubiquitin ligase SCF complex (**Figure 1.5**[**A**]). The SCF complex catalyses polyubiquitination of proteins by bridging interactions between the E2 ubiquitin conjugating enzyme and substrates recognised by the variable F-box protein. Substrate polyubiquitination then leads to rapid protein degradation by the proteasome complex (Gray *et al* 2001; Holland *et al* 2012).

Multiple forms of SCF complex can therefore exist in eukaryotes due to the variability of the associated F-box protein. The F-box transport inhibitor response 1 (TIR1) protein is unique among plants and recognises substrate proteins expressing an auxin-inducible degron (AID) sequence (Gray *et al* 2001; Dharmasiri *et al* 2005). Substrate recognition by the SCF-TIR1 complex is only possible in the presence of members of the plant auxin (AUX) family of hormones e.g. indole-3-acetic acid (IAA). AUX/IAA promotes the interaction between the AID tag presenting proteins and SCF-TIR1, leading to polyubiquitination and degradation of AID presenting proteins in a tightly regulated, conditional and reversible manner (**Figure 1.5**[**B**]; see also **Appendix Figure 1**).

1.7.1 Harnessing AID in Non-Plant Cells

Due to the lack of TIR1 orthologues in non-plant eukaryotes, the AID system has fast become an appealing and versatile tool used to study gene function in yeast, as well as in cells derived from mammalian tissues such as mouse, hamster, chicken and human (Nishimura *et al* 2009; Holland *et al* 2012; Morawska & Ulrich 2013). However, the success of AID protein regulation is limited by the fusion of the AID tag to an endogenous target protein and the co-expression of the TIR1 protein within desired eukaryote system. Recent advancement to CRISPR/Cas9 gene engineering however have greatly improved the integration of an AID system outside of plant cell cultures (Natsume *et al* 2016).

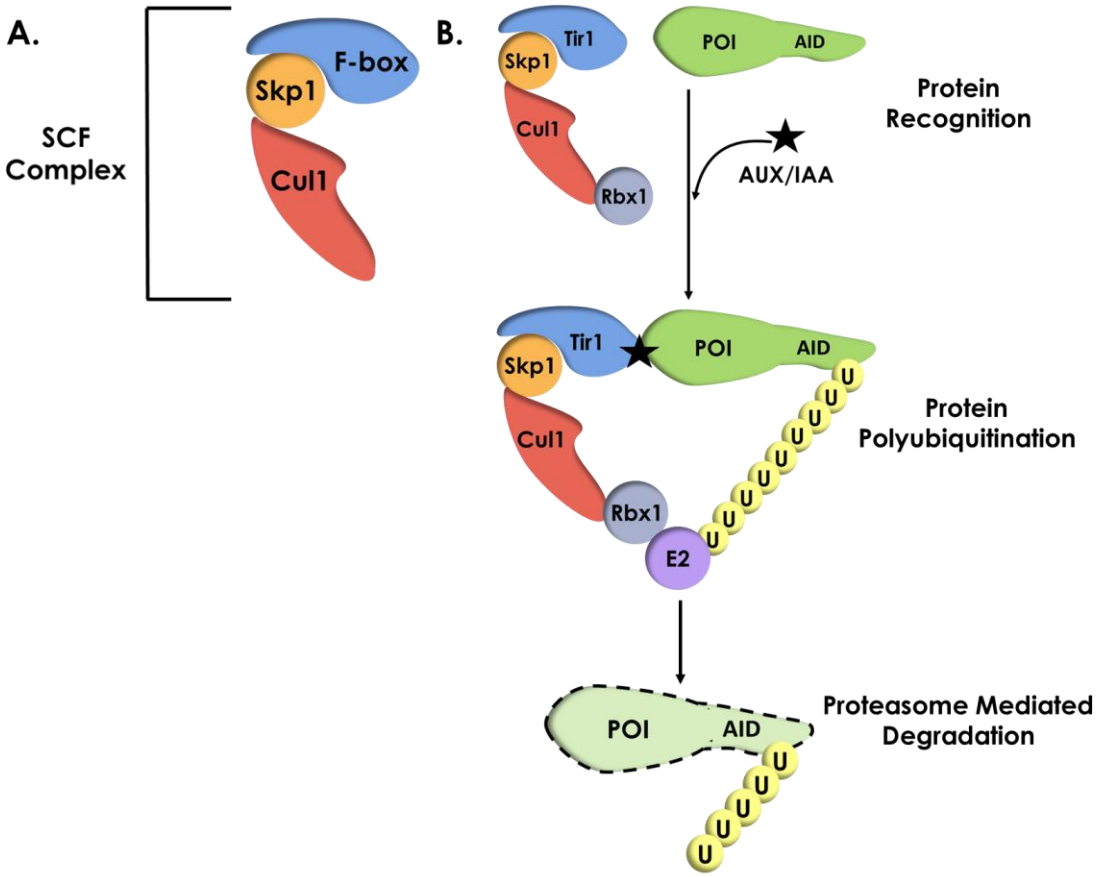


Figure 1.5: (**A**) Trimeric Skp1, Cullin1 (Cul1) and variable F-box proteins form the SCF complex. (**B**) Introduction of AUX/IAA stimulates TIR1 recognition of the AID presenting protein of interest (POI) which mediates polyubiquitination by E2 followed by proteasome mediated degradation.

1.8 Project Aims

As previously mentioned, nuclear surveillance monitors and regulates the output of all thee RNA polymerase complexes, collectively protecting the integrity of the transcriptome. In addition to this, many of these exoribonucleases have secondary functions involved in maturation processes of specific classes of RNA. Identifying the substrates of individual exoribonucleases would therefore provide valuable insight into their functional role within the nucleus.

Classically in human cell lines, the function of these nuclear exoribonucleases has been interrogated using RNAi which overcame many of the difficulties associated with functional genomic approaches. Unlike yeast, the generation of inactive gene knockout cell lines is more cumbersome in humans due to the much larger and increased complexity of the genome. Furthermore, studying gene knockout cell lines is limited to non-essential genes. Instead, RNAi hijacks the endogenous RNA induced silencing complex (RISC) system, common to most higher eukaryotes to specifically target and degrade specific gene mRNA transcripts. As an added advantage, RNAi provides a mechanism to study essential genes critical for cell proliferation (Elbashir *et al* 2001; Kim 2003; Sen & Blau 2006). Thus, the relative ease-of-use and target specificity of RNAi led to its widespread application in functional gene studies.

Regardless of its widespread application, RNAi has a number of limitations. For example, RNAi is limited to mRNA degradation which does not have any impact on the pre-existing target protein within the cell. Consequentially, it can often take several days for protein downregulation to occur depending of the stability of the target protein. As such, RNAi has the potential to accumulate non-specific effects gradually over long incubation periods, often obfuscating the interpretation of protein function. Also, knocking down expression of a target gene is often incomplete (especially with regards to RNA processing factors that have the potential to self-regulate their own expression) hampering detection

of altered genotypic effects. Lastly, RNAi relies on RNA-RNA interactions to guide mRNA destruction which can cause off-target mRNA recognition.

Despite extensive studies of yeast Rrp6, very little investigation of the role of Exosc10 in humans has been undertaken. Likewise, the exact processing activities and substrates of Exosc10 remain unknown. By taking advantage of more modern CRISPR/Cas9 mediated genome editing, I have successfully generated a human HCT116 cell line capable of expressing a 3' AID-tagged Exosc10 protein. Through a combination of rapid protein depletion coupled with transcriptome-wide high-throughput RNA-Seq analysis of nascent RNA, I aim to provide a more complete characterisation of Exosc10 function in relation to other nuclear exoribonucleases, namely, Xrn2 and Dis3, providing a more detailed insight into nuclear surveillance pathways present in humans than has been previously possible.

Chapter 2

Materials and Methods

2.1 Materials

2.1.1 Bacterial Strains

In this project all genetic recombination/molecular cloning was performed using NEB 5-alpha competent *Escherichia coli* (high efficiency) cells.

Growth Media

All growth media was autoclave sterilised and stored at room temperature prior to use.

- **Luria Bertani (LB) Broth:** 10% (w/v) Tryptone, 10% (w/v) NaCl, 5% (w/v) Yeast Extract
- **LB Agar:** As above with the addition of 2% (w/v) Agar

Antibiotic Selection

To identify positive plasmid transfected *E. coli* clones, cells were grown in the presence of selective antibiotics. The following final concentrations of antibiotics were used for positive selection:

- **Ampicillin:** 100 µg/ml
- **Kanamycin:** 50 µg/ml

2.1.2 Tissue Culture

Cell Lines

Several human colon carcinoma (HCT116) cell lines were generated using CRISPR/Cas9 genome engineering. HCT116 cells were kindly donated by Professor Stuart Wilson's lab at The University of Sheffield. No further genotyping was performed on these cell lines. Cells were periodically tested for mycoplasma contamination by visualisation of DAPI (4', 6-diamidino-2-phenylindole) stained DNA using a wide-field Olympus IX81 light microscope. These cells were selected due to their obligate diploid karyotype and the combination of modifications present in each cell line are listed below:

Name	TIR1 Expression	Description
HCT116	No	Unmodified parental cells
HCT116 TIR1	Yes	Sleeping beauty (SB)-integrated TIR1; otherwise unmodified
EXOSC10-AID	Yes	SB-integrated TIR1; homozygous 3' AID tagged EXOSC10
EXOSC10-AID Wild-Type (WT) Rescue	Yes	SB-integrated TIR1; homozygous 3' AID tagged EXOSC10; SB-integrated WT EXOSC10 cDNA
EXOSC10-AID D313A Rescue	Yes	SB-integrated TIR1; homozygous 3' AID tagged EXOSC10; SB-integrated catalytically inactive EXOSC10 cDNA
XRN2-AID	Yes	SB-integrated TIR1; homozygous 3' AID tagged XRN2
DIS3-AID	Yes	SB-integrated TIR1; homozygous 3' AID tagged DIS3

Sleeping beauty (SB) integrated genes/cDNA sequences are under the control of a constitutive ON cytomegalovirus (CMV) driven promoter; multiple copies were also randomly inserted at SB loci to ensure overexpression of protein.

Tissue Culture Media

All HCT116 culture cell lines were maintained in Dulbecco's Modified Eagle Media (DMEM) supplemented with 10% foetal calf serum (FCS) in the presence of penicillin/streptomycin (100 µg/ml). Cell lines expressing TIR1 were additionally maintained in the presence of a low concentration of blasticidin (5 µg/ml) to prevent loss of SB integration.

Selection of CRISPR/Cas9 Engineered HCT116 Cell Lines

Identification of positive homozygous HCT116 cell lines harbouring a 3' auxin-inducible degron (AID) tag on the target gene and integration of TIR1 at SB loci was achieved by selective antibiotic resistance. The final concentrations of antibiotics maintained in culture media are listed:

- **Blasticidin:** 20 µg/ml
- **Hygromycin:** 150 µg/ml
- **Neomycin:** 800 µg/ml
- **Puromycin:** 1 µg/ml

Antibodies

EXOSC10 protein was detected by western blot analysis using anti-EXOSC10 (ab95028) antibody supplied by Abcam. Endogenous MYC and recombinant TIR1-9xMYC proteins were detected with anti-c-MYC (9E10) antibody supplied by Abcam. Additional antibodies used in the DIS3-AID and XRN2-AID cell lines include: Dis3 (Bethyl A303756A), a-tubulin (Sigma T6074, AID tag (MBL: M214-3) and Xrn2 (Bethyl Laboratories, A301-101).

2.1.3 Vectors

All vectors used in the molecular cloning steps were supplied by Addgene and the table below summarises their contents:

Plasmid	Description	Reference
pX330-U6-Chimeric_BB-CBh-hSpCas9	Human codon-optimised Cas9 from *S. pyogenes;* cloning backbone for U6 promoter driven gRNA expression	Cong *et al* 2013
pUC19	Empty backbone cloning vector	Norrander *et al 1983*
pCMV(CAT) T7-SB100	SB-transposase; constitutively expressed from a CMV promoter	Mates *et al* 2009
pBABE osTIR1	Human codon-optimised TIR1; 9x myc tagged	Holland *et al* 2012
pSBbi-Pur	Empty SB-transposon; constitutive bidirectional promoter; puromycin resistance gene	Kowarz *et al* 2015
pSBbi-Blast	As above; puromycin replaced with a blasticidin resistance gene	Kowarz *et al* 2015

2.1.4 Buffers

Buffers were sterilised either by autoclave or by passage through a syringe driven Millex-GP 0.22 µm filter (Sigma) before use.

DNA/RNA Buffers

- **Total RNA Extraction:** TRI Reagent solution (Sigma)
- **Hypotonic Lysis Buffer (HLB):** 10 mM Tris-HCl (pH 5.5), 10 mM NaCl, 2.5 mM $MgCl_2$, 0.5% (v/v) NP40

SDS/PAGE Western Buffers

- **RIPA Buffer:** 150 mM NaCl, 1% (v/v) NP40, 0.5% (w/v) sodium deoxycholate (DOC), 0.1% (w/v) sodium dodecyl sulphate (SDS), 50 mM Tris-HCl (pH 7.4)
- **4x SDS-PAGE Loading Buffer:** 8% SDS, 40% Glycerol, 0.25 M Tris-HCl (pH 6.8), 0.006% bromophenol blue (before use warm to 50°C, take 0.5ml and add 50 µl β-mercaptoethanol)
- **4x SDS-PAGE Resolving Gel Buffer:** 1.5 M Tris-HCl (pH 8.8), 0.4% (w/v) SDS
- **4x SDS-PAGE Stacking Gel Buffer:** 0.5 M Tris-HCl (pH 6.8), 0.4% (w/v) SDS
- **Running Buffer:** 192 mM glycine, 25 mM Tris, 0.1% (w/v) SDS
- **Transfer Buffer:** 192 mM glycine, 25 mM Tris, 20% (v/v) methanol
- **Enhanced Chemi-Luminescence (ECL) Solution 1:** 100 mM Tris-HCl (pH 8.5), 2.5 mM Luminol, 400 µM p-Coumaric Acid
- **ECL Solution 2:** 100 mM Tris-HCl (pH 8.5), 5.3 mM Hydrogen Peroxide

Northern Blot Buffers

- **TBE (5x):** 1.1 M Tris, 900 mM Boric acid, 25 mM EDTA (pH 8.3)
- **RNA Loading Buffer:** 80% formamide, 10 mg/ml bromophenol blue, 10 mg/ml xylene cyanol, 10 mM EDTA (pH 8)
- **DNA Loading Buffer:** 80% formamide, 10 mg/ml bromophenol blue, 10 mg/ml xylene cyanol, 10 mM EDTA (pH 8), 20 mM NaOH
- **Denaturing Urea-PAGE Gel:** Gel density based on ratio of urea, acrylamide 19:1, 1x TBE, ammonium persulphate (APS), TEMED
- **SSPE (1x):** 150 mM NaCl, 9 mM NaH_2PO_4, 1 mM EDTA (pH to 7.4 with NaOH)
- **Denhardt's Reagent:** 0.04% ficoll, 0.04 polyvinylpyrrolidone, 0.04% bovine serum albumin (BSA)
- **Hybridisation Buffer:** 6x (v/v) SSPE, 5x (v/v) Denhardt's Reagent, 0.2% (w/v) SDS

Miscellaneous Buffers

- **qRT-PCR Master Mix:** Agilent Brilliant III Ultra-Fast SYBR Green QPCR Master Mix
- **PBS (1x):** 137 mM NaCl, 2.7 mM KCl, 10 mM Na_2HPO_4, 1.8 mM KH_2PO_4 (pH to 7.4 with HCl). Add 0.05% (v/v) Tween 20 to make **PBST**
- **Oligo Annealing Buffer (2x):** 20 mM Tris (pH 7.5), 100 mM NaCl, 1 mM EDTA (pH 8)
- **Orange G Stock (50x):** 0.125 g/ml orange G, 0.5 g/ml bromophenol blue, 0.5 g/ml xylene cyanol
- **Orange G Loading Dye:** 25 % (w/v) ficoll, 1x orange G stock, 10 mM EDTA (pH 8)
- **Trypsin PBS-EDTA:** 500 ml PBS (1x), 1 mM EDTA, 0.25% Trypsin

2.1.5 Molecular Biology Kits

- **Plasmid extraction from transformed *E. coli*:** Qiagen QIAprep Spin Miniprep Kit
- **Radiolabelled oligo/probe purification:** Qiagen QIAquick Nucleotide Removal Kit

2.1.6 RNA Sequencing Library Kits

- **Ribosomal RNA Depletion:** Illumina Ribo-Zero Gold rRNA Removal Kit
- **RNA-Seq Library Generation:** Illumina TruSeq Stranded Total RNA Library Prep Kit
- **RNA Purification:** Beckman Coulter Agencourt RNAClean XP Beads
- **DNA Purification:** Beckman Coulter Agencourt AMPure XP Beads
- **QC Analysis of RNA and DNA:** Agilent ScreenTape RNA; High Sensitivity RNA; D1000 Assay for Tapestation

2.2 Experimental Methods

2.2.1 Molecular Biology

DNA Extraction Using Phenol/Chloroform

DNA PCR templates were purified by addition of an equal volume of DNA phenol/chloroform (pH 8) (Sigma), homogenised by vigorous shaking and centrifuged at 13,000 rpm for 5 minutes. Upper aqueous phase was transferred to a fresh tube and washed in an equal volume of isopropanol (absolute). Supernatant was discarded; washed DNA pellets in 70% ethanol, 10% (v/v) 3 M sodium acetate and centrifuged at 13,000 rpm for 10 minutes. Isolated pellets were air dried and resuspended in dH_2O.

Transformation of Competent E. coli

Competent NEB 5-alpha *E. coli* were thawed to room temperature slowly on ice and added to 10-20 ng of plasmid DNA. Mixtures were incubated on ice for 10 minutes and then heat-shocked at 42°C for 90 seconds. Cells were allowed to recover in 300 µl Super Optimal broth with Catabolite repression (SOC) media for 30-60 minutes at 37°C. Spread 70 µl of the total cell volume onto an LB agar plate supplemented with the appropriate antibiotic.

Extraction of Plasmid DNA from Transformed E. coli

For mini-preps, transformed bacteria was inoculated into 5 ml of LB broth supplemented with the suitable antibiotic and grown at 37°C with shaking overnight. Plasmid DNA was isolated using a Qiagen QIAprep spin miniprep kit according to the manufacturer's protocol.

Genomic DNA Extraction from Culture Cells

Positive CRISPR/Cas9 engineered culture cells were screened by extracting genomic DNA using QuickExtract DNA extraction solution (Cambio). Cells grown in 60 mm culture dishes were harvested in ice cold 1x PBS and spun at 500 x g for 5 minutes. Supernatant was discarded and pellets were resuspended in 20-50 µl (based on DNA pellet size) of QuickExtract, vortex mixed and incubated at 65°C for 6 minutes. Sampled were vortex mixed again and incubated at 98°C for 2 minutes to denature QuickExtract. DNA was stored at -20°C; for PCR 1 µl of DNA was used as template.

PCR

High fidelity PCR reactions of 25 µl were set up using Q5 DNA polymerase (NEB) as follows: 10 ng of DNA template, 500 nM forward primer, 500 nM reverse primer, 200 µM dNTPs, 1x Q5 reaction buffer, 1x high GC enhancer buffer and 0.5-1 U Q5 DNA polymerase. Thermocycler settings were typically set up with 27 cycles using an annealing temperature between 60-67°C for 30 seconds and an extension temperature of 72°C for 30 seconds/kb.

For plasmid DNA templates, PCR reactions were subsequently treated with Dpn1 (NEB) for 1 hour at 37°C to digest methylated plasmid DNA before ligation and transformation steps.

For positive colony screening of transformed competent cells, PCR reactions using Taq polymerase were set up in 25 µl reactions as: variable template DNA < 500 ng, 200 nM forward primer, 200 nM reverse primer, 200 µM dNTPs, 1x standard reaction buffer and 1.25 U Taq DNA polymerase. Typically, 30-32 cycles were used with an annealing temperature between 50-65°C for 30 seconds and an extension temperature of 68°C for 1 minute/kb.

Restriction Digestion of Vectors

Unless otherwise specified, restriction digests were performed using Cutsmart buffer (NEB) as per the manufacturer's instructions. Typically, reactions were performed at 37°C for 1 hour.

Ligation of Linearized Vectors

Linearized and restriction digested vectors were ligated using 100 ng of DNA in 20 µl reactions at 16°C for 1 hour to overnight using T4 DNA ligase (NEB) following the manufacturer's protocol. Half of the reaction mix was then transformed into competent *E. coli*.

Gibson Assembly of Vectors

Vectors were amplified as multiple fragments and assembled using Gibson assembly (NEB). Plasmid cassettes were typically amplified using divergent PCR to open the DNA backbone creating blunt ends. Vectors were then dephosphorylated using Alkaline Phosphatase, Calf Intestinal (CIP) (NEB) for 1 hour at 37°C, after which DNA was extracted using phenol/chloroform and ethanol purified.

Either PCR amplicons or synthesised DNA oligos were used as insert fragments. For small inserts such as gRNAs, complementary DNA oligos were synthesised with 5' and 3' extended arms homologous to the vector backbone blunt ends. Oligo annealing buffer (1x) was added to each reaction; primer oligos were melted at 98°C for 10 minutes and hybridised by slowly cooling to room temperature to form a dsDNA insert with 3' and 5' overhanging homology arms. Likewise, large PCR amplicons, were amplified using forward and reverse primers designed with extended 5' and 3' arms (respectively) that share sequence complementarity with the vector.

After DNA phenol/chloroform extraction and ethanol purification of both vector and inserts, Gibson reactions (10 µl) were set up using 100 ng

of vector with a 3-6 fold excess of insert (based on relative size) combined with 1x Gibson reaction master mix solution (NEB). Reactions were heated to 50°C for 15 minutes and cooled to room temperature before transformation into competent *E. coli*.

Total RNA Extraction

Total RNA was extracted from cells grown on 60 mm culture plates; cells were harvested by resuspension in 1 ml TRI Reagent (Sigma) and incubated for 5 minutes at room temperature before transferring to an Eppendorf tube. 200 µl of chloroform (absolute) was added, contents were homogenised by vigorous shaking and centrifuged in a table top centrifuge at 13,000 rpm for 15 minutes. The top aqueous layer was then transferred to a fresh tube and washed in an equal volume of isopropanol (absolute); centrifuged for 10 minutes at 13,000 rpm. Supernatant was discarded and the RNA pellet was washed in 70% ethanol, 10% (v/v) 3 M sodium acetate before a final centrifugation step at 13,000 rpm for 10 minutes. Supernatant was then discarded, the RNA pellet was then air dried and resuspended in distilled water (dH_2O).

Nuclear RNA Extraction

Harvested cells grown in a 100 mm culture dish were collected in 5 ml of ice cold 1x PBS and spun at 500 x g for 5 minutes. Cell pellets were resuspended in 4 ml hypotonic lysis buffer (HLB) (see recipe on **p37**) and incubated on ice for 5 minutes. A 1 ml underlay of HLB supplemented with 10% sucrose was then applied; samples were then spun again at 500 x g for 5 minutes to pellet whole nuclei. Supernatant was then drained, and nuclear pellets were washed a second time (to remove any trace cytoplasmic material) in 5 ml HLB before pelleting whole nuclei at 500 x g for 5 minutes. Nuclear RNA was then extracted from isolated nuclei with TRI Reagent (Sigma) using the same protocol as mentioned above.

Removal of Genomic DNA

Total and nuclear RNA was treated with 4 U of Turbo DNase (ThermoFisher) for 1 hour at 37°C in the presence of 1 U/µl RNase Inhibitor Murine (NEB) according to the manufacturer's guidelines. Following this, an equal volume of RNA phenol/chloroform (pH 4.3) (Sigma) was added and the solution was homogenised by vigorous shaking before centrifugation at 13,000 rpm for 5 minutes. The top aqueous phase was transferred to a fresh tube and washed in 70% ethanol, 10% (v/v) 3 M sodium acetate; centrifuging at 13,000 rpm for 10 minutes. Supernatant was removed and the pellet was air dried before resuspension in dH_2O and storage at -20°C.

Reverse Transcription

Purified, genomic DNA depleted, total and/or nuclear RNA was first quantified on a nanoDrop 2000 spectrophotometer (ThermoFisher) before reverse transcription into cDNA using Protoscript II (NEB). 1 µg of RNA mixed with 0.4 µg random hexamers (Bioline) was primed by heating to 70°C for 5 minutes and snap quenched on ice. Samples were combined with reverse transcriptase (RT) mix composed of 500 µM dNTPs, 1x reaction buffer, 10 nM DTT and 10 U Protoscript II RT (NEB); incubated at 25°C for 5 minutes, 42°C for 1 hour and heat denatured at 70°C for 15 minutes.

q-PCR

Typically, q-PCR was performed using 20-50 ng of cDNA per reaction. For each reaction 100 nM forward primer and 100 nM reverse primer was mixed with 4 µl of 2x Brilliant III SYBR green master mix (Agilent) to a total volume of 8 µl. Each sample/primer reaction was run in triplicate on a Qiagen Rotor-Gene Q for 45 cycles detecting short amplicons < 150-200 nt in length using the following settings:

- Melt at 95°C for 10 seconds
- Anneal and elongate at 60°C for 10 seconds; acquire on green

Fold enrichment was calculated using in-built comparative quantitation analysis relative to a control sample. In most scenarios, base level expression across samples was measured by comparison of a housekeeping gene such as GAPDH, U6 or MYC.

Agarose Gel DNA Electrophoresis

Identification of DNA PCR amplicons was achieved through association with its size when separated by gel electrophoresis. Typically, 1% (w/v) agarose was dissolved in 1x TBE (providing a separation resolution between 0.5-10 kb in size) by heating in a microwave for 3-5 minutes. Solution was cooled to ~50°C before adding ethidium bromide (to a final concentration of 10 µg/ml), cast in a mould and set by cooling to room temperature. DNA was mixed with 1/5th (v/v) orange G loading buffer and loaded on the gel alongside an appropriate DNA ladder; gels were run in 0.5x TBE at a constant 150 V until dye front reached the end of the gel. Bands of DNA were visualised by fluorescence emitted by exposure to UV light using a Gel Doc XR+ system (Bio-Rad).

2.2.2 Synthesis of Repair Template Plasmids

Generation of EXOSC10-AID Repair Template Vector

Vector maps of plasmids used to engineer HCT116 cells are represented in **Figure 2.1**. Plasmids were externally sequenced by eurofins following each recombination step to confirm sequence editing. All CRISPR/Cas9 repair templates were Gibson assembled into the empty pUC19 backbone. EXOSC10 homology arms of ~400 nt in length flanking the poly(A) site were synthesised by Integrated DNA Technologies (IDT) and ligated into pUC19 (**Figure 2.1**[**B**]). This vector was then linearized between the penultimate codon and stop codon to accommodate insertion of a pre-synthesised (IDT) AID-P2A tag sequence and either a hygromycin or neomycin resistant IDT synthesised gene using Gibson assembly.

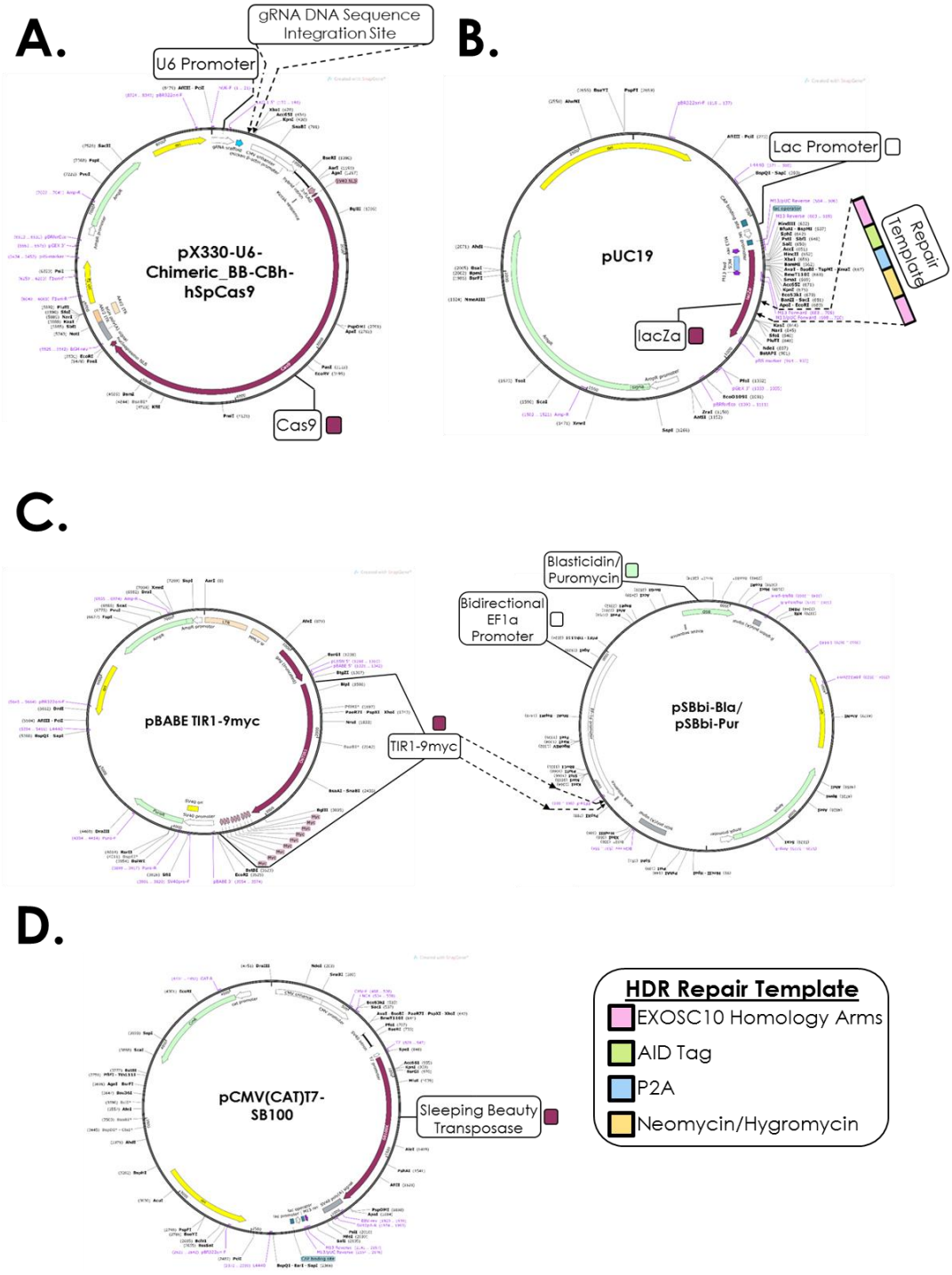


Figure 2.1: Vector maps were obtained from https://www.addgene.org and visualised using SnapGene Viewer. (**A**) Cas9 expression vector with U6 expression scaffold for duplex gRNA DNA sequence integration and expression. (**B**) Integration site of the HDR repair template driven by a lac promoter within pUC19. (**C**) Human codon optimised TIR1-9myc sequence was transferred from the pBABE vector to an empty backbone sleeping beauty transposon vector. (**D**) Sleeping Beauty transposase containing vector for expression in human systems.

Generation of XRN2-AID and DIS3-AID Repair Template Vectors

The XRN2-AID and DIS3-AID repair template vectors were made by Professor Steve West and Laura Francis (respectively) using the same approach as above.

Insertion of gRNA into Cas9 Expression Plasmid

IDT synthesised gRNA oligos were designed using the online Benchling software tools (retrieved from https://benchling.com), annealed and inserted into pX330-U6-Chimeric_BB-CBh-hSpCas9 using Gibson assembly (NEB) (**Figure 2.1**[**A**]).

Generation of SB-Transposon Vectors

Human codon optimised osTIR1-9xmyc (TIR1) was PCR isolated from the pBABE osTIR1 and inserted into the pSBbi-Blast empty vector using SfiI restriction sites (**Figure 2.1**[**C**]). For EXOSC10 rescue gene analysis, a synthetic WT EXOSC10 cDNA sequence was synthesised by Dharmacon, amplified out of the vector using PCR primers carrying homology arms to the pSBbi-Pur vector and assembled using Gibson assembly (NEB) replacing the TIR1 sequence. Divergent PCR was then performed to introduce a single point mutation of residue 313 (GAC → GCC; Aspartate → Alanine) described by Januszyk *et al* 2011, generating a catalytically inactive D313A EXOSC10 mutant.

2.2.3 Western Blotting Assay

Protein was isolated from cells cultured in 60 mm plates by re-suspending in 4°C RIPA buffer and incubated for 20 minutes on ice. Samples were spun at 13,000 rpm on a bench top centrifuge collecting the final supernatant. Protein lysate was mixed with 4x SDS loading buffer and separated on a 10% SDS-PAGE gel in 1x running buffer using the Mini-PROTEAN system (Bio-

Rad). Gels were transferred on to a nitrocellulose membrane (GE Healthcare) in 1x transfer buffer using a Trans-Blot Turbo Transfer System (Bio-Rad).

Membranes were blocked for 1 hour in 5% skim milk PBST and then incubated for 1 hour in 2% skim milk PBST with primary antibody. Membranes were rinsed in PBST for 10 minutes and incubated with secondary antibody in 2% skim milk PBST for 1 hour at a concentration of 1:10,000. Two final 10 minute room temperature washes were performed in PBST before ECL detection and imaging on a Gel Doc XR+ system (Bio-Rad).

2.2.4 Northern Blot Analysis

Loaded 5 µg of nuclear extracted, genomic DNA depleted RNA on a 12% denaturing urea-PAGE gel and run in 0.5x TBE at 200 V until the dye front reached the end of the gel. Separated RNA was then transferred in 0.5x TBE on to a Hybond-N+ nylon membrane (GE Healthcare) at 10 V for 16 hours. Transferred membranes were then dried and UV crosslinked (2 x 1200 µjoules/cm2) before incubation in hybridisation buffer at 37°C for 1 hour.

For 5.8S processing analysis, DNA oligo probes were designed to target the premature 3' 40 nucleotide extended isoform or the mature 5.8S rRNA. Each probe was 5'[γ-^{32}P]ATP radiolabelled for 1 hour at 37°C using T4 PNK (NEB) and cleaned using a Qiagen QIAquick nucleotide removal kit (as per the manufacturer's guidelines) to remove unincorporated 5'[γ-^{32}P]ATP. Probes were then added to the hybridisation buffer and incubated at 42°C overnight. Membranes were then rinsed in hybridisation buffer 3 times for 1 minute. A final fourth wash was then performed at 42°C for 15 minutes before drying and developing on a Phosphor screen. Image was developed on a GE Typhoon FLA 7000 (GE Healthcare). Developed images were then quantitated and analysed using the ImageJ suite (Schindelin *et al* 2012).

Importantly, membranes were initially hybridised with the extended 3' 5.8S rRNA probe (since it is present at a very low abundance) and after image development, washed 3 times in hybridisation buffer before re-probing with the mature 5.8S probe. In doing so, the quantity of RNA loaded was preserved between each sample and probe used.

2.2.5 Library Preparation of Nuclear RNA

I prepared all of the RNA-Seq libraries which were sequenced at the Exeter Sequencing Service within Exeter University. Each library was prepared using 1 µg of genomic DNA depleted, nuclear RNA. To ensure an unbiased analysis of the transcriptome in each cell line no additional measures were taken to enrich for small, highly structured RNAs such as tRNAs, snRNAs or snoRNAs. Before library preparation the RNA integrity of each sample was determined using the Tapestation apparatus (Agilent) (**Figure 2.2**). RIN^e is a measure of RNA integrity generated by evaluating the ratio of 28S to 18S rRNA and assigning a score from 1-10, with 10 being the highest. Samples with sufficiently high RIN^e scores typically > 7 (displayed in green) were selected for further processing. Next, rRNA was removed using Ribo-Zero Gold rRNA removal kit (Illumina) according to the user manual; RNA was purified using RNAClean XP Beads (Beckman Coulter). Depletion of rRNA from the samples was screened using a high-sensitivity RNA screen tape for Tapestation (Agilent). Following rRNA depletion, sample RIN^e scores were either absent or low indicating successful rRNA depletion. Additionally, no RNA was detected from the EXOSC10-AID plus IAA sample at this stage, possibly due to falling below the quantitative range (500-10,000pg/µl) of the Tapestation apparatus. Despite this the sample was still processed. Libraries were then prepared from the rRNA depleted samples using TruSeq Stranded Total RNA Library Prep Kit (Illumina) according to the manual and purified using Ampure XP beads (Beckman Coulter). Resulting cDNA libraries were then screened for fragment size and concentration by Tapestation D1000 (Agilent). Libraries passing QC were then pooled and sequenced using Hiseq2500 (Illumina) culminating in

approximately 25-60 million 50 bp single-end reads per sample. Information regarding the depth sequencing, mapping efficiency and genomic coverage can be found within **Figures 3.11**, **4.2** and **5.1**.

Each of the three cell lines used to generate the RNA-Seq libraries were processed following 60 minutes of either IAA or ethanol (solvent) treatment to maintain consistency between analyses. Three biological replicates were sequence for EXOSC10-AID cells and two biological replicates for both DIS3-AID and XRN2-AID cell lines.

2.2.6 Cell Biology

Mammalian Tissue Culture

As mentioned previously cell lines were maintained in DMEM supplemented with 10% FCS and penicillin/streptomycin (100 µg/ml). Cells were grown as a fixed monolayer in an incubator set at 37°C with 5% CO_2. On average, cells were passaged 2-3 times per week depending on the cell line. Media was removed and cells were washed in trypsin PBS-EDTA, incubated for 5 minutes at 37°C then deactivated by resuspension in DMEM, transferring a small fraction to fresh flasks containing an appropriate volume of DMEM.

Conditional auxin induced depletion was achieved by auxin/IAA (Sigma) addition to culture medium at 500 µM for 60 minutes (unless otherwise stated). Cells used as an untreated control were instead treated with an equivalent volume of ethanol (solvent).

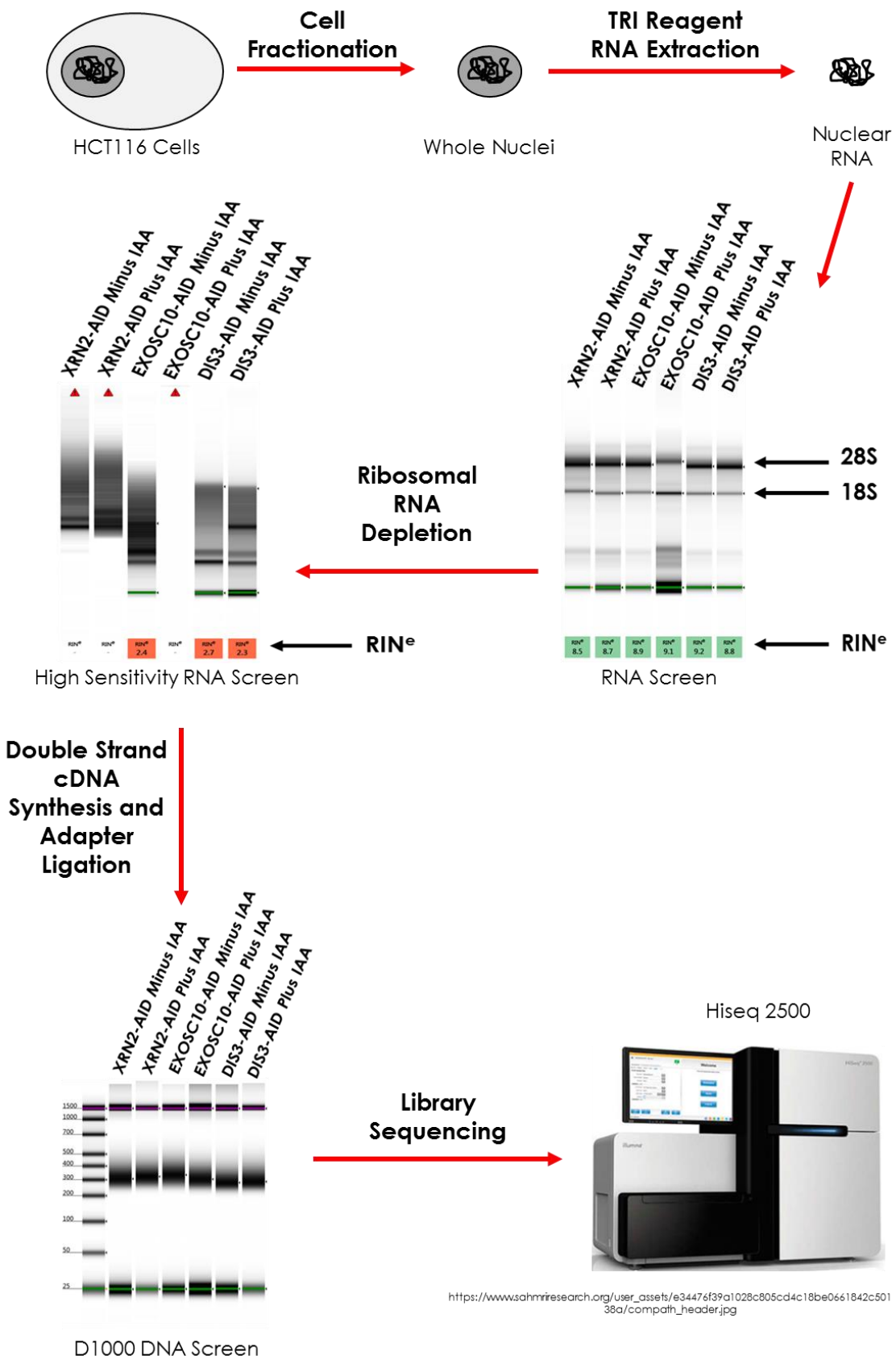


Figure 2.2: Work flow of nuclear RNA extraction, RNA screening steps and RNA-Seq library preparation. Only one replicate of each cell line used in this study was shown for clarity.

Lasting Storage of Cultured Cell Lines

Long-term storage of cells was achieved by passaging fully confluent 750 mm^2 culture flasks (as above); cells were resuspended in 10 ml of DMEM and centrifuged at 500 x g for 5 minutes. Solution was drained from collected cells, homogenised in DMEM supplemented with 10 % DMSO and transferred to a cryotube for storage at -80°C.

To retrieve cells from -80°C storage, frozen stocks were thawed slowly to room temperature, mixed with 9 ml of DMEM media and centrifuged at 500 x g for 5 minutes. The solution was drained from collected cells which were then seeded into a fresh 750 mm^2 culture flask containing an appropriate volume of DMEM.

Plasmid Transfection

For CRISPR/Cas9 integration of the AID tag at EXOSC10 3' ends, 1 µg of repair template vector bearing either hygromycin or neomycin resistance and 1 µg of guide Cas9 plasmid was transfected into HCT116 using JetPrime reagent according to the manufacturer's instructions. For TIR1 and EXOSC10 rescue cDNA integration, 200 ng of SB-transposon and 20 ng of transposase plasmids were transfected using JetPrime.

Generation of EXOSC10-AID HCT116 TIR1 Cell Line

TIR1 was first integrated into the parental HCT116 cell line using the SB transposon system (Hackett *et al* 2010; Skipper *et al* 2013; Hou *et al* 2015; Kowarz *et al* 2015). After 24 hours, cells were passaged into 100 mm culture dishes and selected with 20 µg/ml blasticidin for 48 hours using the entire population for subsequent CRISPR/Cas9 engineering.

Transfection of the TIR1 expressing HCT116 cells was implemented in 6-well culture plates seeded at a density of ~20% 16 hours prior to the introduction of the chimeric gRNA Cas9 vector and two HDR template plasmids (containing neomycin and hygromycin resistant genes).

Following transfection, cells were incubated for 48 hours before passaging into a 100 mm plate in the presence of 800 µg/ml neomycin and 150 µg/ml hygromycin. After ~10-14 days, single colonies were picked, transferred to 24-well plates and screened via genomic DNA sequencing and/or PCR analysis and western blotting for homozygous integration of EXOSC10-AID.

Finally, WT and D313A EXOSC10 rescue cDNA were incorporated into EXOSC10-AID TIR1 expressing cells using the same protocol, albeit selecting cells using 1 µg/ml puromycin for 48 hours instead.

Colony Formation Assay

Approximately 200 cells from each cell line were seeded into 100 mm cell culture plates and grown in the presence of either 500 µM auxin (IAA) or ethanol (solvent) for a period of 10 days. In this instance, growth media and IAA were replaced every 2-3 days. After 10 days plates were washed twice in cold (4°C) 1x PBS, emerged colonies were then fixed in ice cold methanol (absolute) for 10 minutes and stained using 0.5% (w/v) crystal violet + 25% (v/v) methanol for 10 minutes. Excess crystal violet was washed from plates using dH_2O before air drying and imaging. Stained colonies were counted using the ImageJ particle analyser function (Schindelin *et al* 2012). Genuine colonies were defined as existing at a density ranging between 50-8000 pixels with a circularity rating between 0.75-1 (1 = perfect circle).

2.3 Bioinformatics Methods

2.3.1 Software Catalogue

Name	Version	Description	Reference
BamTools	2.4	Tools for handling genome alignment (BAM) files	Barnett *et al* 2011
BEDTools	2.2.6	Flexible tools for genome arithmetic	Quinlan & Hall 2010
BEDOPS	2.4.34	Fast highly scalable and easily-parallelizable genome analysis toolkit	Neph *et al* 2012
CutAdapt	1.15	Removes adapter sequences from high-throughput sequencing reads	Martin 2011
Deeptools	3.0.2	Tools developed for analysis of high-throughput sequencing data	Ramirez *et al* 2014; Ramirez *et al* 2016
DESeq2 *	1.18.1	Differential gene expression analysis based on the negative binomial distribution	Love *et al* 2014
FastQC	0.11.5	High-Throughput Sequence QC reporting	Andrews 2010
FeatureCounts	1.5.2	Ultrafast and accurate read summarization program	Liao *et al* 2013; Liao *et al* 2014
GenomicRanges *	1.30.2	Representation and manipulation of genomic intervals and variables defined along a genome	Lawrence *et al* 2013
ggplot2 *	2.2.1	Create elegant data visualisations using the grammar of graphics	Wickham 2009
HISAT2	2.1.0	Graph-based alignment of next generation	Kim *et al* 2015
IGV	2.4.6	Visualization tool for interactive exploration	Robinson *et al* 2011; Thorvaldsdottir *et al*
MACS2	2.1.0	Finds Peaks of Enrichment in ChIP-Seq Data	Zhang *et al* 2008
R	3.4.4	R is a free software environment for statistical computing and graphics	http://www.R-project.org
Rtracklayer *	1.38.3	R interface to genome annotation files and the UCSC genome	Lawrence *et al* 2009
SAMtools	1.4.1	Tools for alignments in the SAM	Li *et al* 2009; Li 2011
SortMeRNA	2.1.0	Fast and accurate filtering of ribosomal RNAs in metatranscriptomic data	Kopylova *et al* 2012
StringTie	1.3.3b	De novo transcript assembly and quantification for RNA-Seq	Pertea *et al* 2016
Trim_Galore!	0.4.4	Quality/Adapter/RRBS-Trimming of sequencing reads (powered by CutAdapt)	Krueger 2012

Software denoted with * is derived from a Bioconductor package within the R environment.

2.3.2 RNA-Seq Read Alignment

Raw single-end 50 bp reads were screened for sequencing quality using FastQC (Andrews 2010), adapter sequences were removed using Trim Galore! (Krueger 2012), trimmed reads shorter than 20 bp were discarded. All bioinformatic analyses were carried out using the Ensembl GRCh38.p10 and GRCh38.90) human gene annotations. Before alignment, trimmed reads were passed through the SortMeRNA pipeline (Kopylova *et al* 2012) to remove trace rRNA matching the in-built 18S and 28S human databases. Reads were then mapped to GRCh38 using HISAT2 (Kim *et al* 2015) with default parameters supplemented with known splice sites. Unmapped and low MAPQ reads (< 20) were then discarded from the final alignment file using SAMtools (Li *et al* 2009; Li 2011).

2.3.3 Calculation of Genome Coverage and Depth

Genome coverage was calculated using the following formula:

$$genome\ coverage = \frac{(total\ mapped\ reads\ x\ read\ length\ [bp])}{genome\ length\ (bp)}$$

The size of the human GRCh38 genome used in the alignment can be found at: https://www.ensembl.org/Homo_sapiens/Info/Annotation and the number of mapped reads were obtained from the summary output option of the HISAT2 alignment software (Kim *et al* 2015). Sequencing depth was determined using the SAMtools depth function (Li *et al* 2009; Li 201) and calculated as an average depth per base.

2.3.4 Differential Expression Analysis

The number of reads per gene or per transcript (as described for each experiment) were counted using featureCounts (Liao *et al* 2013; Liao *et al* 2014). Differential expression was called using DESeq2 (Love *et al* 2014).

Statistically significant upregulated genes or transcripts were defined as fold change ≥ 2, padj < 0.05.

2.3.5 Metagene Profiling

Metaplots were calculated at the gene level by counting the number of reads aligned to genes using featureCounts (Liao *et al* 2013; Liao *et al* 2014), removing any genes with low expression levels (< 50 reads per gene). An extended transcriptional window was then applied to each gene to include a 3 kb region 5' of the TSS and a 7 kb region 3' of the TES. Genes which overlapped as a result of these extended windows were detected using BEDTools merge (Quinlan & Hall 2010) and discarded to prevent double counting of mapped reads.

Metagene profiles of these filtered genes were then generated from RPKM normalised reads using the deeptools suite (Ramírez *et al* 2014; Ramírez *et al* 2016) with further graphical processing performed in the R environment (http://www.R-project.org). Normalised coverage plots (RPKM) were visualised using the Integrated Genome Viewer (IGV) suite (Robinson *et al* 2011; Thorvaldsdóttir *et al* 2013).

2.3.6 Read Enrichment over Genomic Elements

Per base read coverage was calculated over defined genomic intervals using the SAMtools depth function (Li *et al* 2009; Li 2011). Exons were merged into a single synthetic bed interval to overcome differential exon usage caused by transcript isoforms. Likewise, a custom intron annotation file was produced and merged into synthetic intervals. Coverage was calculated as:

$$coverage = \frac{(read\ count\ x\ read\ length)}{total\ interval\ length\ (bp)}$$

The final coverage results were then normalised to account for differences in library size for each alignment file.

2.3.7 De novo Transcript Assembly

Dis3 RNA-Seq libraries were pooled and the RNA transcriptome was *de novo* assembled using the StringTie suite (Pertea *et al* 2016) for each library with default parameters guided by current GRCh38 reference annotation. Known annotated genes were dropped leaving only novel *de novo* transcripts. Assembled transcriptomes were merged into a single consensus annotation, reads were then counted per transcript using featureCounts (Liao *et al* 2013; Liao *et al* 2014) and differential expression was called using DESeq2 (Love *et al* 2014).

Upregulated *de novo* transcripts (≥ 2-fold, padj < 0.05) that did not align to known gene intervals were extracted and categorised into PROMPT and eRNA transcripts based on their relative distance to the nearest annotated gene. Transcripts within 3 kb of known genes were designated as PROMPTs, whereas transcripts greater than 3 kb from the nearest annotated gene were categorised as eRNAs. *De novo* eRNA designated transcripts were then filtered against all annotated human eRNA transcripts defined by the FANTOM5 database (Lizio *et al* 2015), producing a final list of novel unannotated eRNAs.

2.3.8 Determination of eRNA Directionality

DIS3 mapped reads were split into sense and antisense reads using SAMtools (Li *et al* 2009; Li 2011) and independently counted over non-stranded eRNA gene annotation using featureCounts (Liao *et al* 2013; Liao *et al* 2014). The sum reads of each strand was used to calculate transcription directionality for each condition using the following formula described by Szczepińska *et al* (2015):

$$directionality = \frac{sense\ reads - antisense\ reads}{sense\ reads + antisense\ reads}$$

Bidirectional eRNA transcription was defined as having a range between -0.5 and 0.5, sense directionality greater than 0.5 and antisense directionality less than -0.5.

2.3.9 Generation of Synthetic Intron Annotation

A custom intron annotation file was produced by merging all exon intervals derived from each transcript isoform to generate a synthetic transcript representative of every gene. Synthetic exons were then subtracted from gene interval producing intron intervals with inherited gene information. Synthetic introns were then counted and numbered according to their strand orientation i.e. sense introns numbered ascending, antisense introns descending, finally merging into a single annotation file.

2.3.10 Histone Peak Calling from ChIP-Seq Analysis

ChIP-Seq data was generated by ENCODE from immunoprecipitation (IP) of acetylated histone 3 lysine 27 (H3K27ac) (GEO: GSE31755), monomethylated histone 3 lysine 4 (H3K4me1) (GEO: GSE31755), trimethylated histone 3 lysine 4 (H3K4me3) (GEO: GSE35583) and an input control sample (GEO: GSE31755) in unmodified HCT116 cells.

The raw single-end ChIP-Seq reads were pre-processed to remove adapter sequences and low quality reads similar to the RNA-Seq pre-processing steps detailed above. Reads were next mapped to GRCh38 using HISAT2 (Kim *et al* 2015) with no splice site detection set as a parameter before filtering mapped reads with MAPQ values > 20.

Aligned reads were then converted from BAM file format into BED and sequencing duplicates (i.e. reads that perfectly matched chromosome location and strand) were removed leaving only a single copy of each aligned read. Reads were collapsed into a coverage BEDGRAPH file for further peak calling by MACS2 (Zhang *et al* 2008). A

background ChIP-Seq signal was first calculated from the input control sample then, each histone modification was compared against background signal after normalisation of sequencing depth, generating a set of peaks for each epigenetic mark. Identified peaks were then passed through a Poisson test to call peaks with a q-value cut-off < 0.05 before producing coverage files of peak enrichment. Finally, the enrichment of H3K4me1 and H3K4me3 were directly compared (taking in to account the differences in sequencing depth) and visualised as a log2 ratio using the bigwigCompare function within the deeptools suite (Ramírez *et al* 2014; Ramírez *et al* 2016).

Chapter 3

The Functional Role of Exosc10 in the Nucleus

The human Exosc10 and functional yeast homologue Rrp6, form part of the nuclear exosome complex which is responsible for the processing and degradation of a wide catalogue of RNA transcripts. As previously mentioned, Exosc10/Rrp6 is a 3'→5' exoribonuclease that is active independently as well as when in complex with the exosome. While the exosome is more commonly associated with RNA degradation, Exosc10/Rrp6 has been shown to play an important role regarding maturation and degradation of numerous small, highly structured precursor RNA transcripts such as snoRNAs and pre-rRNA (Januszyk *et al* 2011). Additionally, Exosc10/Rrp6 also has a secondary role when bound to the exosome as it facilitates threading of RNA substrates into the central channel of the exosome complex where they are subsequently degraded by Dis3, a second exosome-bound 3'→5' exoribonuclease (Mitchell 2014; Kilchert *et al* 2016; Ogami *et al* 2018).

Due to the ease of creating gene knockout and temperature sensitive mutant *S. cerevisiae* cell lines (compared to higher eukaryotes) the focus of Rrp6 function has largely been interrogated within yeast, and much less is known about the activity of Exosc10 within humans. While many aspects of human Exosc10 activity is in agreement with Rrp6 in yeast, RNAi based knockdown approaches are very disparate compared to the classical functional genomics methodologies undertaken in *S. cerevisiae*. To gain a better understanding of the role of Exosc10 in humans, I sought to study the immediate impact of Exosc10 loss by using a more direct protein-based functional genomics approach.

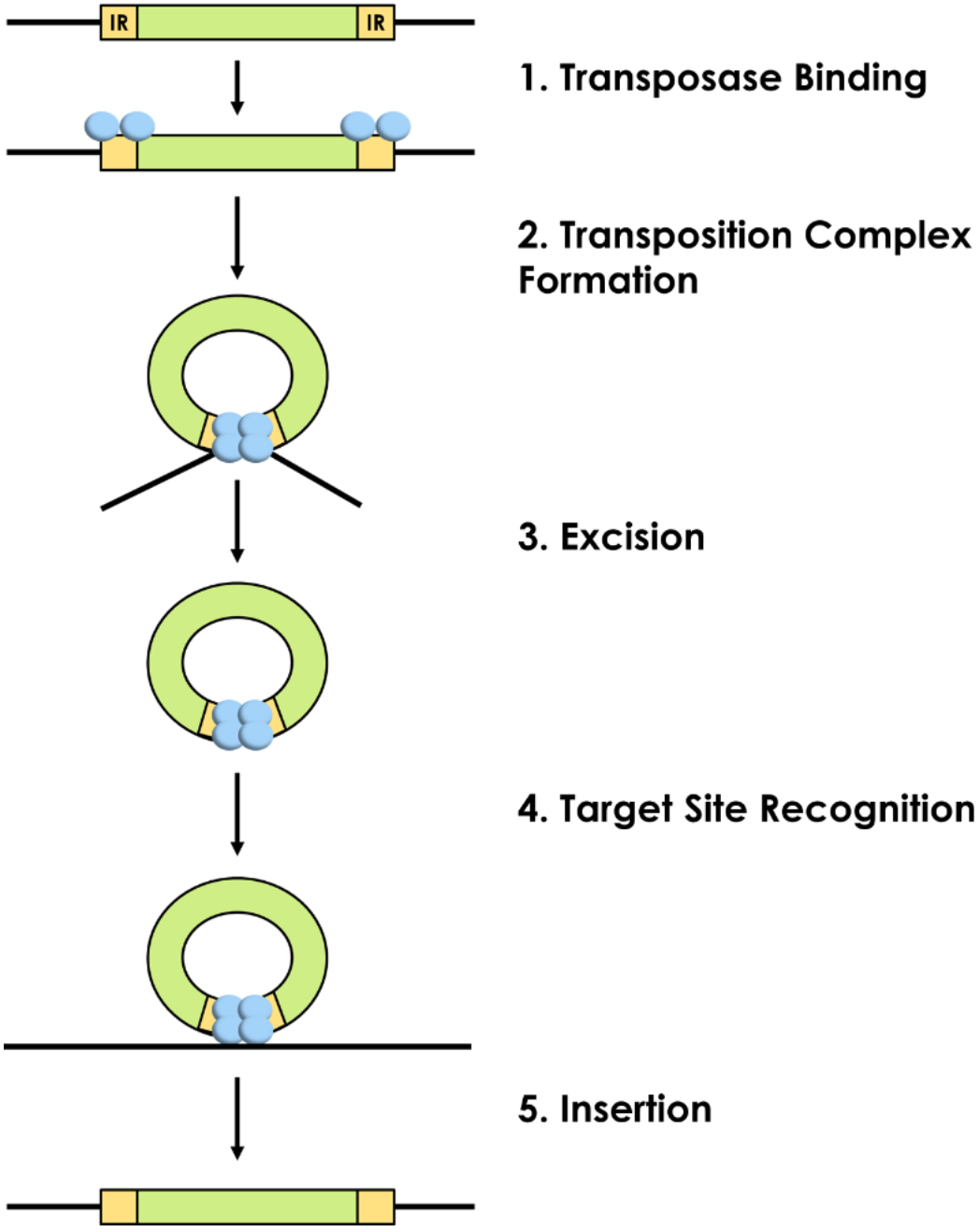


Figure 3.1: Step-by-step mechanism of the *Sleeping Beauty* transposon system adapted from Skipper *et al* (2013). Inverted repeat (IR) sequences (coloured yellow) flank the DNA sequence designated for transfer (green) and contain binding sites that are recognised by 4 transposase (blue) enzymes which carry out gene transfer into the host genome.

3.1 Generating the AID Tagged EXOSC10 Cell Line

The first stage in studying the functional role of Exosc10 in humans was to develop a conditional knockdown cell line that could be rapidly and reversibly induced, therefore bypassing many of the complications and ambiguity of RNAi based knockdown. For this, the CRISPR/Cas9 gene editing tool-set was utilised to specifically modify the gene sequences of both EXOSC10 alleles and incorporate an AID tag sequence at their 3' end. To ensure that all copies of the endogenous EXOSC10 were modified with the 3' AID tag, the HCT116 cell line (derived from human colon carcinoma cultured cells), was selected for manipulation since they have a well-established strict diploid karyotype.

3.1.1 Retroviral Integration of the Plant Specific TIR1 Gene

Since the auxin degron system is present exclusively in plants, recognition of AID tagged proteins requires the co-expression of the plant specific F-box protein, TIR1. In order to consistently express a sufficient abundance of the TIR1 protein, the TIR1 gene was stably integrated into transcriptionally active, "empty" loci within the host HCT116 genome using a transposon-based delivery system (**Figure 3.1**).

DNA transposons employ a "cut-and-paste" mechanism to integrate DNA sequences into a host genome without the requirement of viral vectors or machinery. Instead, cargo DNA sequences flanked by inverted repeat (IR) sequences can be shuffled directly from non-viral vectors into the host genome through the activity of four transposase enzymes (Hackett *et al* 2010). Several transposon delivery systems exist, many of which have been altered to improve their efficiency and target recognition sequence within the host genome as well as their cargo DNA capacity. The synthetic *Sleeping Beauty* (SB) transposon delivery mechanism was selected for TIR1 gene integration as it can accommodate the transfer of relatively large (~10 kb) DNA sequences. Additionally, the SB transposase is hyperactive and has been shown to

provide long-term integration and expression of transgenes within host vertebrate genomes (Mates *et al* 2009; Hou *et al* 2015; Kowarz *et al* 2015).

The TIR1 cDNA sequence was placed under the control of a constitutively ON CMV promoter sequence prior to integration between flanking IR sequences of the transposon vector. In addition, the transposon vector also carries a blasticidin resistance gene which was also transferred into the HCT116 genome. Following drug resistance selection, a heterogeneous population of HCT116 TIR1 expressing cells were then cultivated for further genomic manipulation.

3.1.2 Modification of EXOSC10 by Exploiting HDR Templates

Next, two EXOSC10 repair templates were designed to take advantage of the HDR repair pathway. For this, the AID degron tag, self-cleaving peptide sequence P2A (Kim *et al* 2011; Kreidenweiss *et al* 2013) and drug resistance selection marker (in this case both neomycin and hygromycin were used for each allele) were sandwiched between flanking sequences which share sequence homology with the 3' end of the EXOSC10 gene (**Figure 3.2**). Following double-stranded DNA cleavage by Cas9, which was directed to the EXOSC10 gene by a designed gRNA sequence, the DNA sequence flanked by the homology was then integrated into the HCT116 genome by HDR. One benefit of using the HCT116 cell line in this study includes the high efficiency and ease of plasmid transfection achievable, but more importantly, HCT116 cells have an obligate diploid karyotype, therefore selecting homozygous tagged EXOSC10 populations of cells is relatively simple though integration of the two drug resistant markers mentioned earlier.

By separating the 3'end of the EXOSC10-AID sequence and drug resistance marker with the P2A sequence, two distinct proteins can be expressed from a single endogenous promoter and mRNA, ensuring that positively selected colonies derive from full CRISPR/Cas9 integration and not from partial integration of the selection marker alone (**Figure 3.3**).

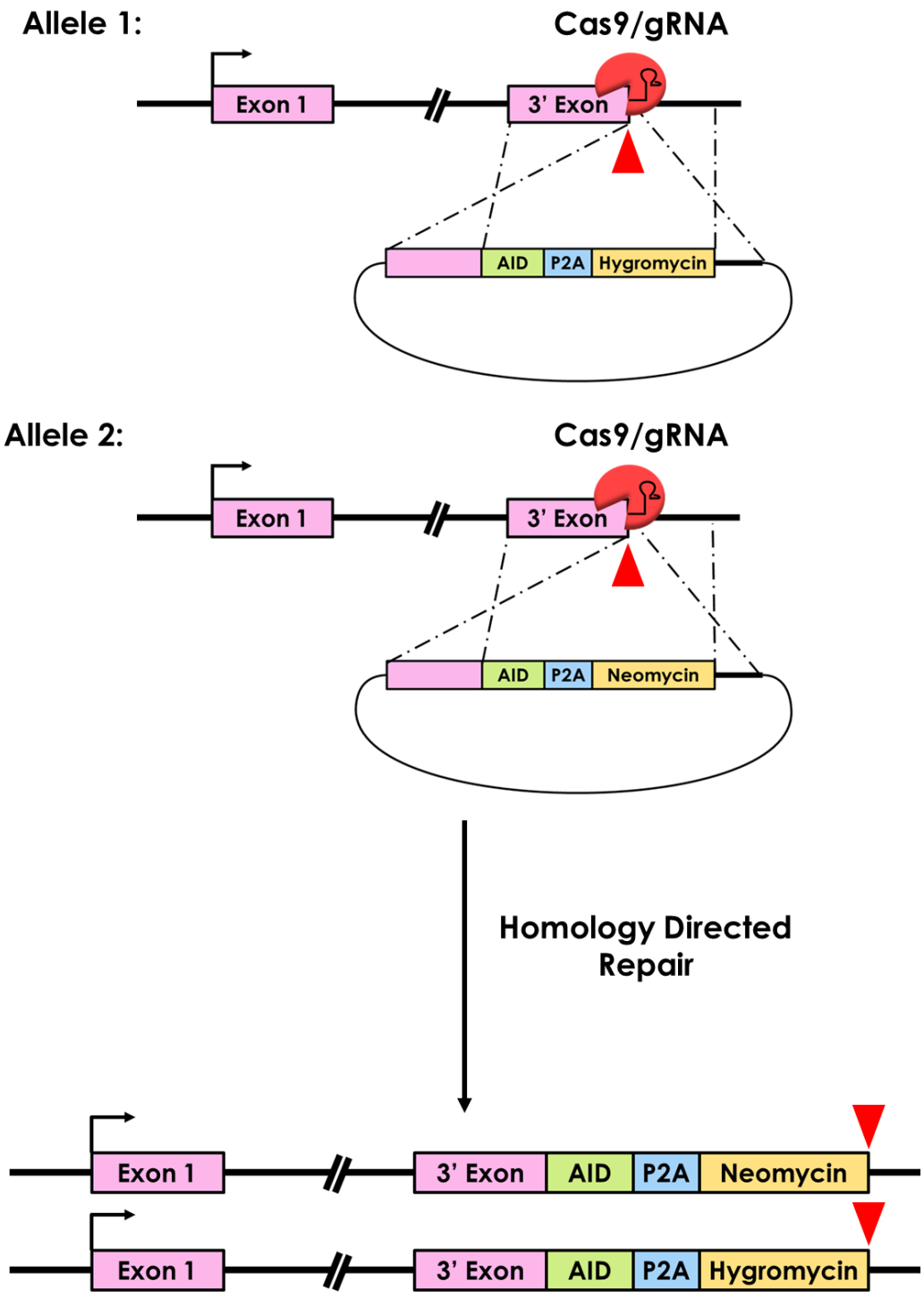


Figure 3.2: Method of tagging both alleles of the EXOSC10 gene using CRISPR/Cas9 genome engineering. Cas9 cleavage is directed by the gRNA between the penultimate codon and stop codon. Homologous sequences (dashed lines) were used to repair the cleaved DNA. After repair the endogenous PAS site (indicated by the red triangle) and 3' UTR shifted downstream of the selection marker.

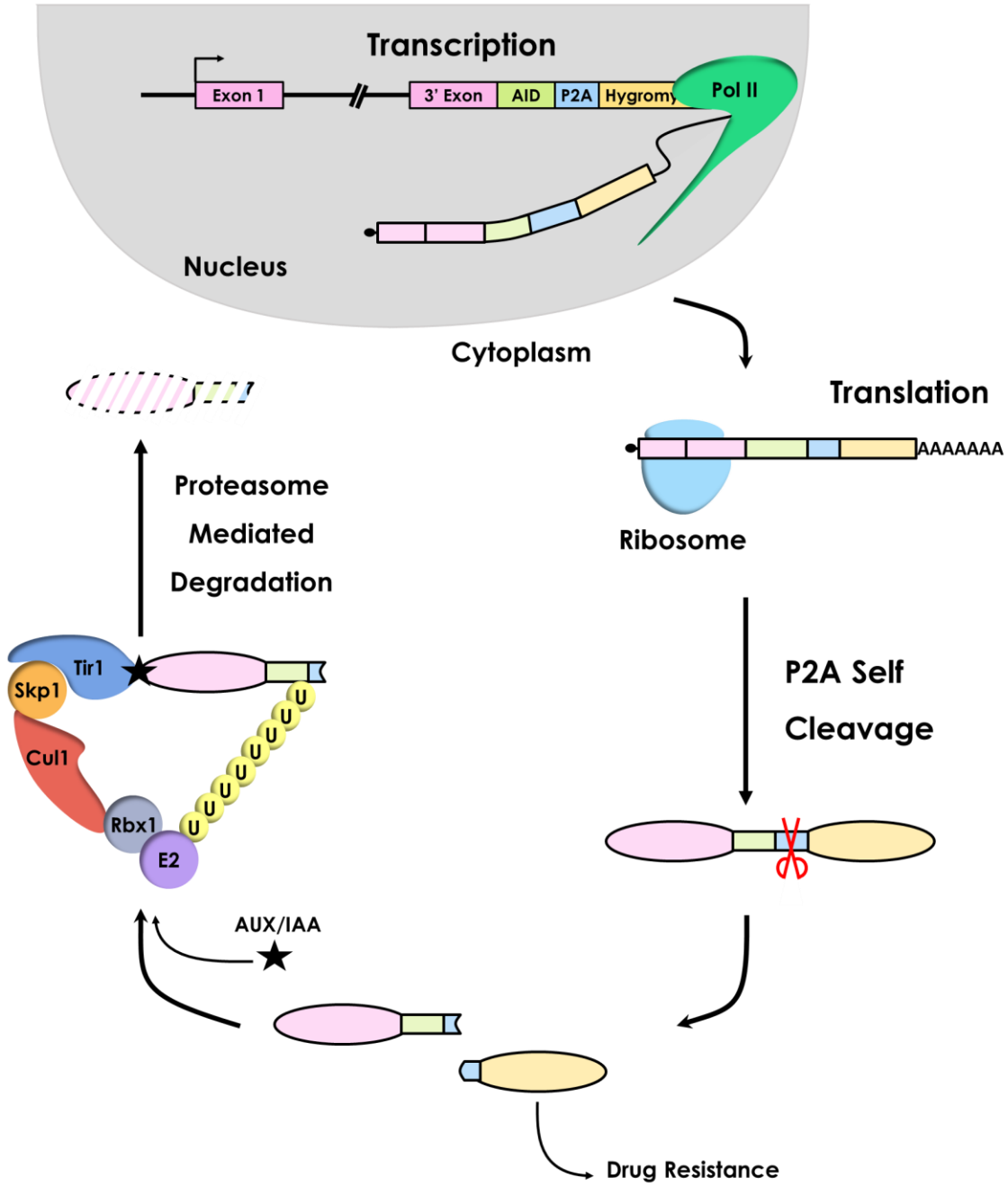


Figure 3.3: Following incorporation of the AID tag and selection marker to the 3' end of the EXOSC10 gene, a single recombinant mRNA transcript is transcribed from the endogenous promoter and terminated using endogenous poly(A) site signals. This mRNA later becomes translated by ribosomes in the cytoplasm. Shortly after the completion of protein synthesis, the P2A peptide self cleaves releasing the mature Exosc10-AID and drug resistant proteins. Introduction of auxin/IAA stimulates TIR1 recognition of the AID tag, leading to polyubiquitination and proteasome mediated degradation of Exosc10.

3.1.3 Validation of Genome Engineering by Western Blot Analysis

Following antibiotic selection, positive CRISPR/Cas9 modified cell lines grown from single colonies were screened by western blot assay to identify populations expressing the Exosc10-AID tagged protein. An anti-EXOSC10 antibody that recognises an internal amino-acid sequence was used so that both endogenous and tagged isoforms could be detected.

In the unmodified parent TIR1 cell line, only the 100 kDa endogenous Exosc10 isoform was detected, however, in all 3 selected colonies screened, a larger Exosc10 specific protein band of ~130 kDa was detected bearing a size consistent with the inclusion of the 3' AID tag (**Figure 3.4**). The lack of detectable endogenous Exosc10 protein combined with the survival of each colony in the presence of both selective drugs indicates that each cell line screened represents a homozygous EXOSC10-AID modified population. Additionally, expression of the TIR1 protein in all cell lines was achieved through detection of the incorporated 9xMyc tags fused to the C-terminal domain using an anti-MYC antibody.

Colony number 1 was selected for all subsequent analyses as the expression of Exosc10-AID protein was comparative to endogenous Exosc10 in the unmodified parent TIR1 cell line.

3.1.4 Identification of EXOSC10-AID Gene by Genomic DNA Screening

Genomic DNA isolated from positive EXOSC10-AID cells was then screened using a nested end-point PCR approach. Despite designing several primers flanking the proposed integration site of the AID tag at the 3' end of the EXOSC10 gene, no clear PCR amplicon could be detected possibly due to the high GC content of the template DNA (**Supplementary Figure S1**). This was similar to the issues experienced when designing the

homologous arms intended for the HDR templates, which were eventually synthetically synthesised. As an alternative genomic DNA screen, RNA-Seq reads that overlap the EXOSC10 3' end were shown to abruptly terminate before reaching the 3' UTR region, creating a sequencing gap indicative of the inserted AID tag sequence not present in the reference genome (**Supplementary Figure S2**).

3.2 Depletion of the Exosc10-AID Protein is Rapid

One of the biggest advantages of the auxin degron system is the speed with which AID presenting proteins are degraded by the proteasome. Therefore, the next logical step was to determine the rate of protein depletion following auxin induction.

To test this, the abundance of Exosc10-AID protein was measured by western blot assay after cells were treated with auxin (hereafter referred to as IAA) over a range of intervals between 0 and 60 minutes. The level of Exosc10-AID protein decreased steadily over the course of an hour, culminating with the almost complete absence of protein at 60 minutes following the introduction of IAA (**Figure 3.5**). To confirm that depletion of Exosc10-AID protein is specific to the presence of a complete auxin-inducible degron system, parent TIR1 cells expressing unmodified endogenous Exosc10 were also treated with IAA for 0 or 60 minutes. For parent TIR1 cells, no protein depletion was observed indicating that protein degradation requires AID tag inclusion at the 3' end of the Exosc10 protein and the expression of the TIR1 F-box protein.

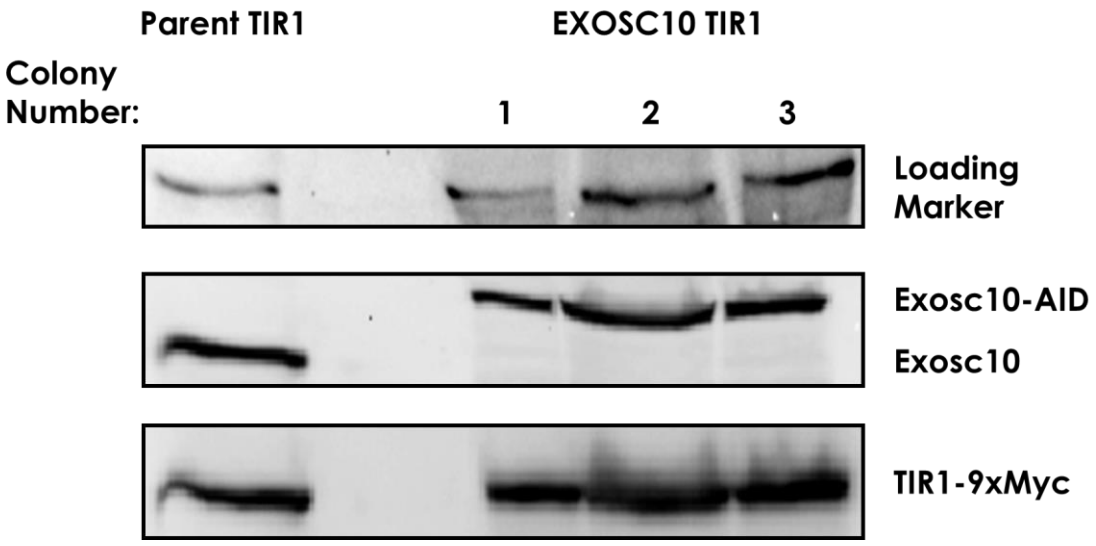


Figure 3.4: Screening of 3 cell lines derived from single colonies surviving double antibiotic resistance selection. Anti-EXOSC10 antibody was used to detect both the endogenous and 3' AID modified Exosc10 proteins. TIR1 protein expression was detected using anti-Myc recognition of the 9xMyc tags fused to the 3' end of TIR1. A non-specific band detected with anti-EXOSC10 antibody was used as a loading marker.

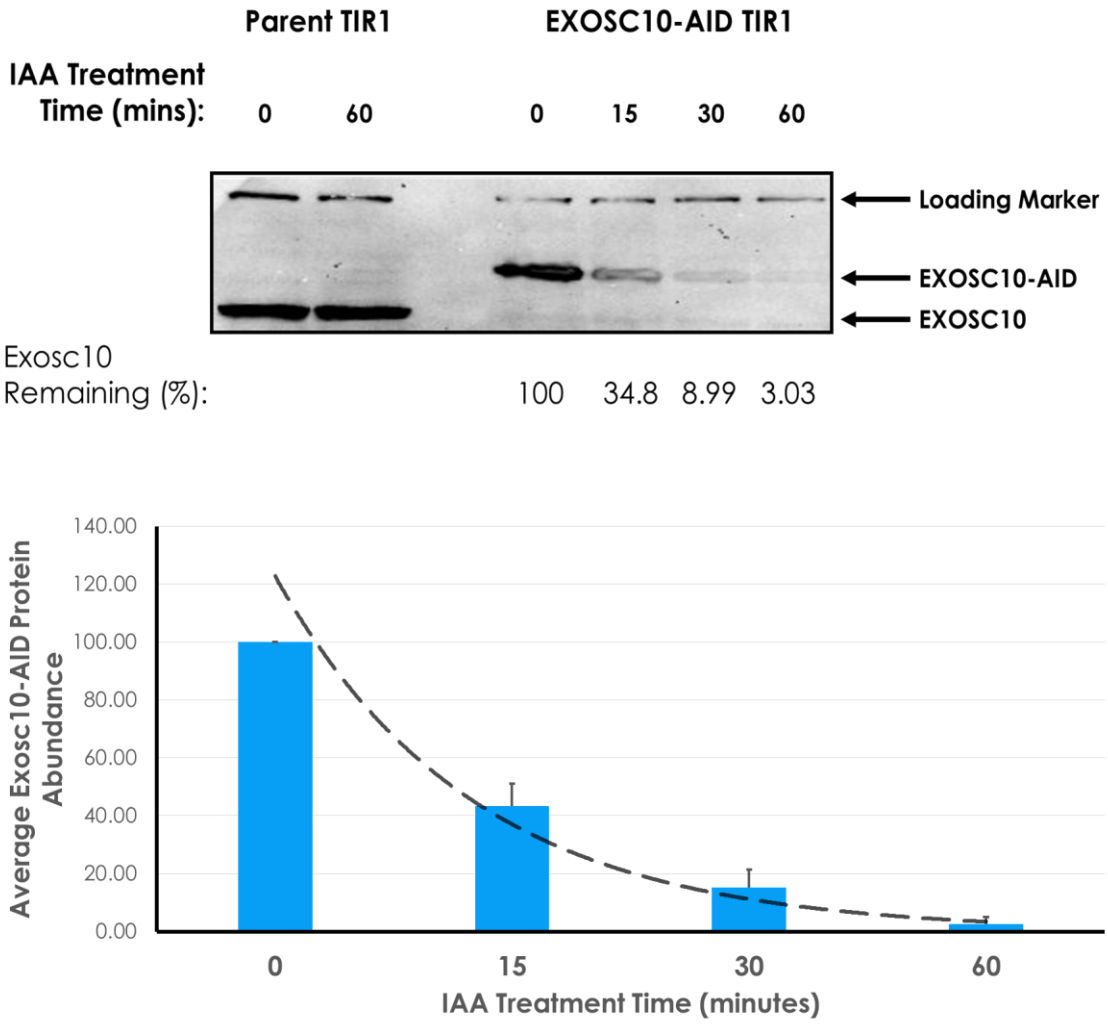


Figure 3.5: The EXOSC10-AID TIR1 positive cell line was treated with 500 µM IAA for 0, 15, 30 or 60 minutes prior to protein extraction. As a control, parent TIR1 cells were treated with the same concentration of IAA for 0 or 60 minutes. A non-specific protein detected with the anti-EXOSC10 antibody was used as a loading marker. Protein abundance was calculated using ImageJ. Bar plot was generated from 3 biological replicates (error bars = standard deviation). An exponential line of best fit was also added between each time point.

3.3 The AID Tag Doesn't Interfere with Exosc10 Function

Fusion of the AID tag to the 3' end of the Exosc10 protein has the potential to interfere with substrate recognition and reduce or abolish its catalytic activity. To address these concerns, the function of the Exosc10-AID protein was investigated.

In yeast, Rrp6 has been shown to be essential for the processing of precursor 5.8S rRNA into mature rRNA through trimming of a 30 nt sequence from the 3' end of the transcript (Briggs *et al* 1998). An analogous processing step exists in humans, whereby Exosc10 trims a 40 nt 3' extended sequence from the precursor 5.8S rRNA (hereafter referred to as 5.8S+40).

To determine if the Exosc10-AID protein fulfils this role in the modified cell line, the ratio of unprocessed 5.8S+40 to mature 5.8S rRNA transcript was calculated by northern blotting analysis. Under normal (uninduced) conditions, the abundance of 5.8S+40 rRNA transcript is maintained at a similar level to the parent TIR1 cell line expressing the unmodified endogenous Exosc10 protein (**Figure 3.6**). However, after 60 minutes of Exosc10-AID depletion the ratio of 5.8S rRNA is significantly altered due to accumulation (~6-fold on average) of the precursor 5.8S+40 transcript. Therefore under normal conditions, the fusion of the AID tag to the Exosc10 3' end does not negatively impact catalytic function of Exosc10 in this cell line. Importantly, this experiment confirms that depletion of the Exosc10-AID protein in the presence of IAA causes a significant precursor rRNA processing defect similar to previous findings in budding yeast systems.

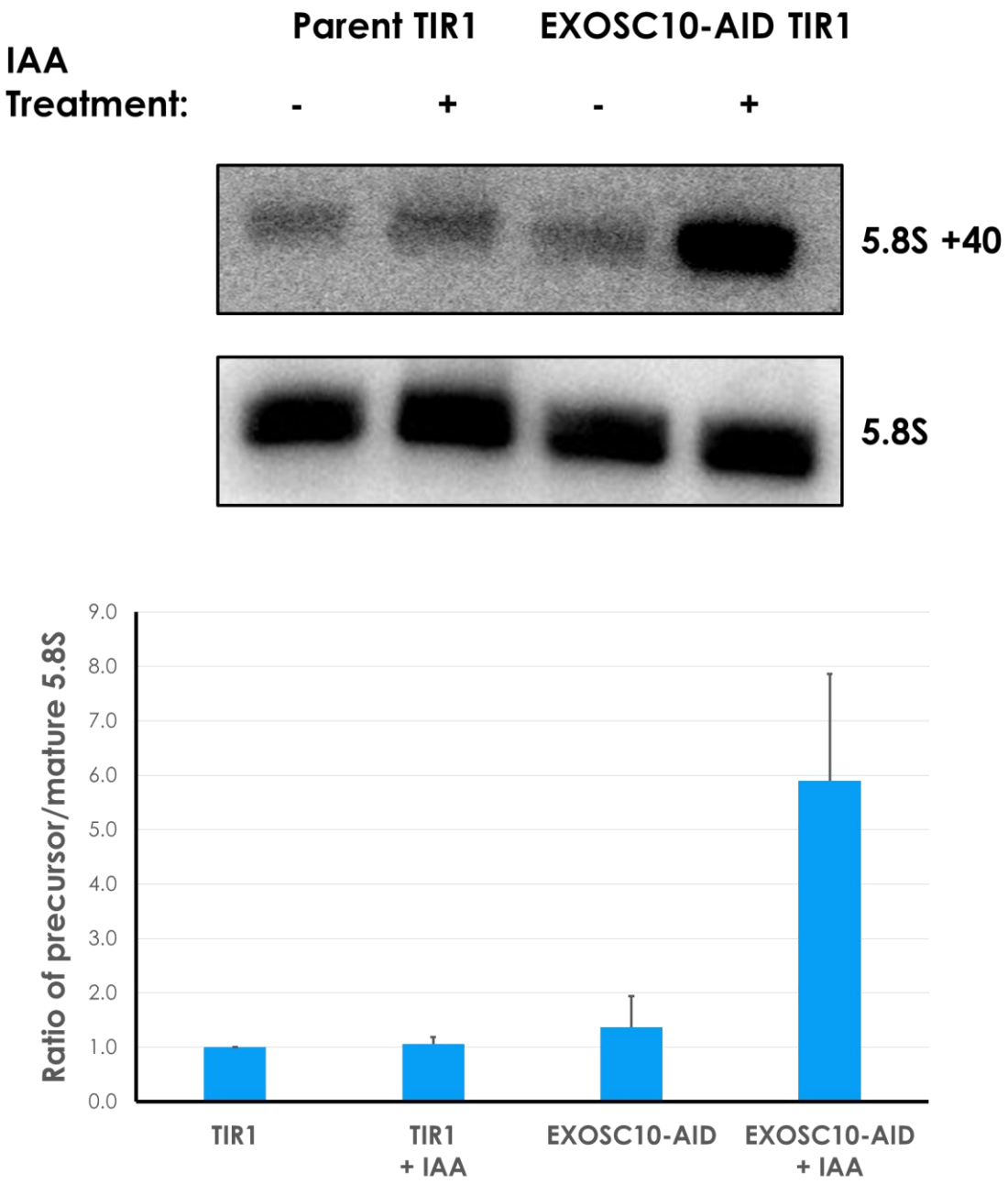


Figure 3.6: Northern blot analysis of precursor extended 5.8S rRNA (+40) and mature 5.8S (top). For analysis, 5 µg of nuclear RNA extracted from cell lines treated with either 500 µM IAA of ethanol (solvent) for 60 minutes was loaded on a 12% denaturing PAGE-Urea gel. Ratio of precursor/mature 5.8S was calculated from particle density analysis using the ImageJ suite. Bottom bar plot represents the average of 3 replicates, (error bars = standard deviation).

3.4 Exosc10 is Essential for Cell Viability

In yeast, the Rrp6 protein is the only member of the exosome complex that is not essential for cell viability (Januszyk *et al* 2011) instead, deletion of the RRP6 gene confers a slow growth phenotype at 30°C and loss of growth at 37°C, indicating that Rrp6 is only essential for cell viability at elevated temperatures (Briggs *et al* 1998). In higher eukaryotes such as *Drosophila melanogaster* however, Rrp6 homologues were shown to be essential for S2 cell progression through mitosis implying a secondary function of Rrp6 within cells that require nuclear envelope breakdown during mitotic spindle assembly (Graham *et al* 2009; Kiss & Andrulis 2010).

To address these observations in human cells, a colony formation assay was performed in the HCT116 EXOSC10-AID cell line. A small number of cells were first seeded onto culture plates and grown for an extended period of time in the presence or absence of IAA. After around 10 days of severe and constitutive Exosc10-AID protein downregulation, an almost complete loss of cell viability was observed compared to an untreated counterpart (**Figure 3.7**). Supporting the notion that the incorporation of the of the AID tag to the 3' end of Exosc10 does not interfere with its function, colonies formed from untreated EXOSC10-AID cells are similar in size compared to unmodified parent TIR1 cells (**Supplementary Figure S3**[**A**]), indicating that AID inclusion does not significantly impact growth rate in these cell lines.

Depletion of Exosc10-AID protein was observed solely in cells expressing a complete auxin-inducible degron system and not in the parent TIR1 cell line which recovered an almost equal number of colonies after 10 days. More importantly, cell death can be attributed exclusively to the loss of Exosc10 since no adverse growth defects were observed from the prolonged presence of IAA in the growth media. Similar to *D. melanogaster* S2 cells, Exosc10 is essential for cell viability in humans. This may be due to its proposed secondary function during mitosis, however this is yet to be determined.

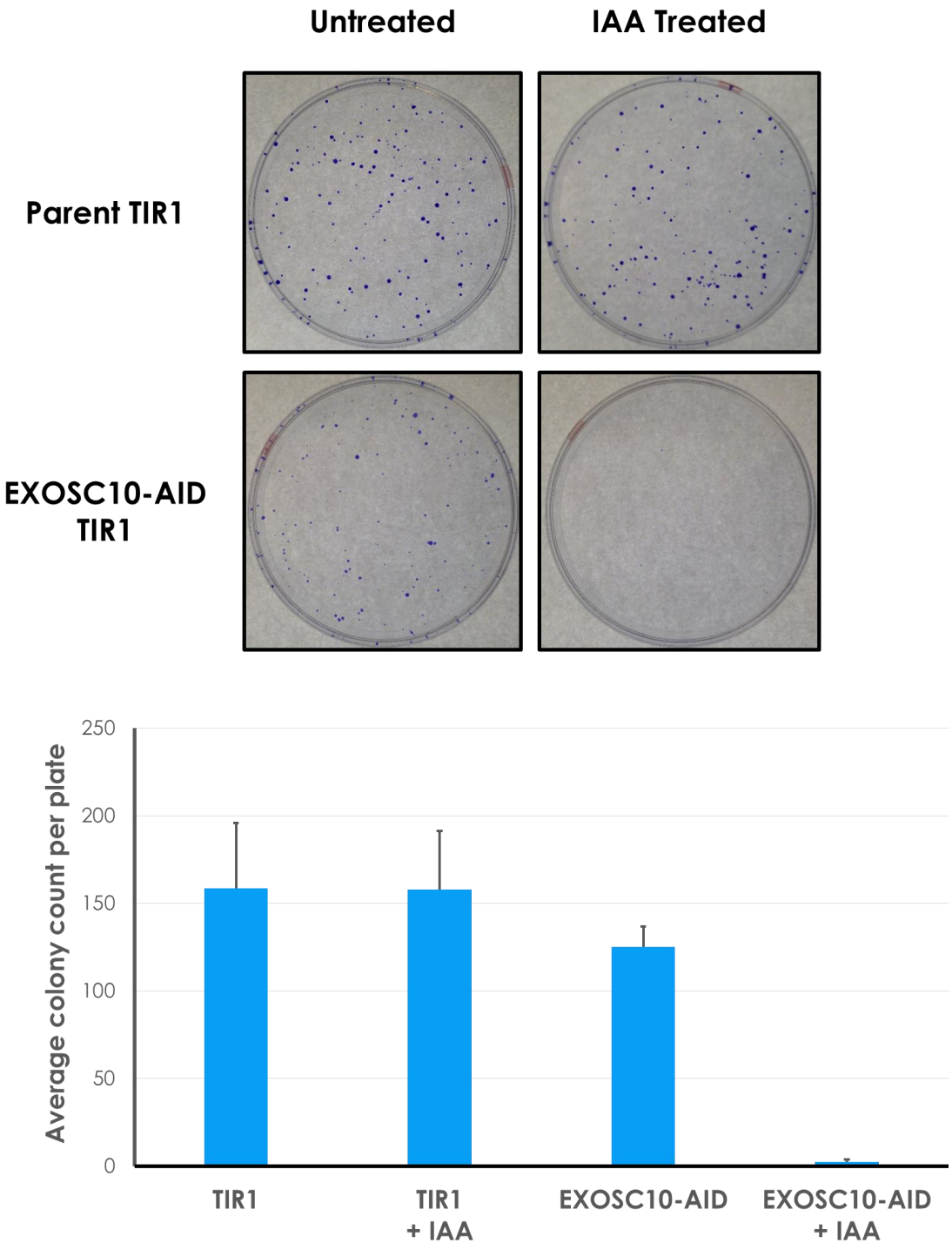


Figure 3.7: Colony formation assay of approximately 200 cells seeded into 100 mm plates and grown in the presence of either 500 µM IAA or ethanol (solvent) for 10 days. Colonies grown were fixed, stained and counted using ImageJ and the average of 3 replicates was charted (error bars = standard deviation).

3.5 Catalytic Activity of Exosc10 is Dispensable for Cell Survival

Although the absence of Exosc10-AID protein resulted in cell death, it was not clear whether this was caused by the loss of Exosc10 catalytic activity or the physical depletion of the protein itself. This was important to address because Exosc10 can also interact with a broad range of RNA substrates that require unfolding and threading into the central channel of the exosome prior to meeting the active centre of Dis3 (Mitchell 2014; Kilchert *et al* 2016; Ogami *et al* 2018). Moreover, the C-terminal domain of yeast Rrp6 has been proposed to make contact with Dis3 in order to enhance its exoribonuclease activity (Wasmuth & Lima 2016). To confirm that Exosc10 protein depletion is attributable to cell death, the structural and enzymatic functions of Exosc10 must be separated.

Exosc10 is a DEDD-Y ribonuclease enzyme that forms part of the DEDD nuclease superfamily. The active centre of the exonuclease domain is highly conserved between yeast Rrp6 and human Exosc10 and is comprised of 4 key DEDD residues that act in concert with a fifth conserved tyrosine residue at position 436 (**Figure 3.8**[**A**]). Mutation of any single residue that form part of the DEDD-Y active site has been shown to completely abolish the enzymatic activity of Rrp6 in yeast (Januszyk *et al* 2011). For this analysis, the first key aspartate at residue 313 was chosen and mutated into alanine (henceforth referred to as D313A) via single nucleotide mutagenesis (**Figure 3.8**[**B**]).

Using the HCT116 EXOSC10-AID TIR1 cell line as a parent, the WT or catalytically inactive D313A EXOSC10 cDNA sequences were next integrated into the genome under the control of a constitutive CMV promoter using the SB transposon system. After selection (for puromycin resistance), both the WT rescue and D313A cell lines were screened by western blot analysis to confirm that both of these proteins were expressed respectively as well as to ensure that SB integration of Exosc10 cDNA did not interfere with IAA induced degradation of the Exosc10-AID protein. In

both cell lines generated, WT and D313A Exosc10 proteins are unaffected by IAA mediated depletion and can functionally replace the Exosc10-AID protein after 1 hour of depletion (**Figure 3.9**).

Analysis of colonies formed after 10 days of Exosc10-AID protein depletion revealed that the overexpressed WT Exosc10 protein successfully rescued cell proliferation with an almost equal number of colonies recovered compared with an untreated counterpart (**Figure 3.10** [**A**]). This additionally demonstrates that cell death on treatment of EXOSC10-AID cells with auxin is due to loss of Exosc10 function rather than any other unanticipated event. Surprisingly, the catalytically inactive D313A Exosc10 protein partially rescued cell viability with ~40 colonies recovered (~50% survival rate compared to untreated D313A cells), a significant boost in survivability compared to Exosc10 deficient cells (**Figure 3.10**[**A**]). The colonies recovered from D313A Exosc10 overexpression however, appeared to be slightly smaller on average in comparison to the WT Exosc10 rescue cell line however, similar to the unmodified parent and EXOSC10-AID cell lines assayed previously, this does not appear to be statistically relevant given the spread of colony size observed from the overlapping error bars observed between each sample population (**Supplementary Figure S3**[**B**]). Therefore, unlike yeast Rrp6, depletion or inactivation of Exosc10 in humans does not confer an obvious slow growth phenotype. While the catalytic activity of Exosc10 appears to be dispensable for cell viability, cell proliferation is nevertheless significantly impeded. While the reasons for this are not clear from this assay, one possible explanation may be attributed to the loss of interactions between Exosc10 and the mitotic spindle as suggested by Graham *et al* (2009). Furthermore, replacing Exosc10 with the inactive D313A mutant as part of the exosome partially restores viability arguing against the possibility that cell death is caused by failed substrate targeting to Dis3. In any case, the effects of an inactive Exosc10 appears to correlate with both yeast and *D. melanogaster* S2 observations.

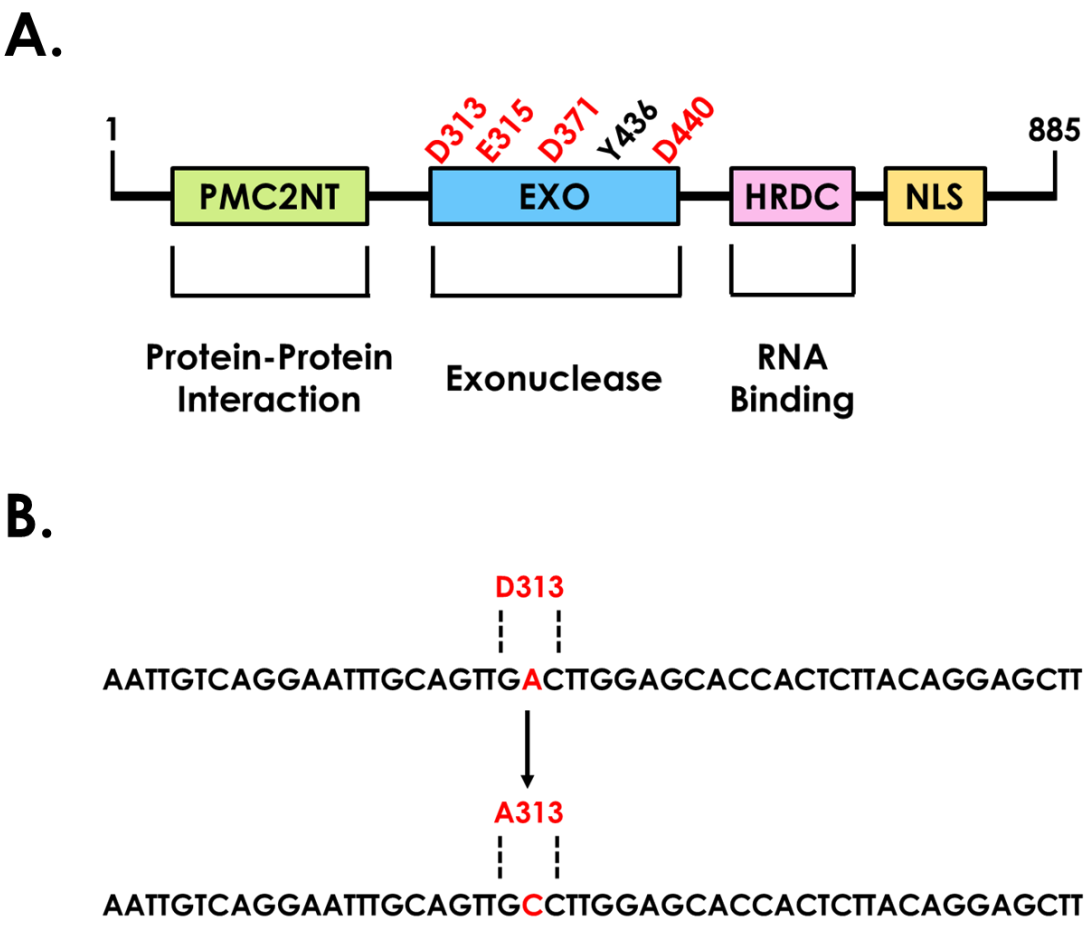


Figure 3.8: Molecular structure of the human EXOSC10 protein. Catalysis is dependent on 4 key active-site DEDD-Y residues within the exonuclease domain (highlighted in red). An additional conserved tyrosine residue (black) is also present in Exosc10. Two nuclear localisation signals (NLS) are represented by a single box at the gene 3' end (**A**). Partial sequence of the exonuclease domain, conversion of an aspartate at residue 313 to alanine by single nucleotide mutagenesis is highlighted in red (**B**).

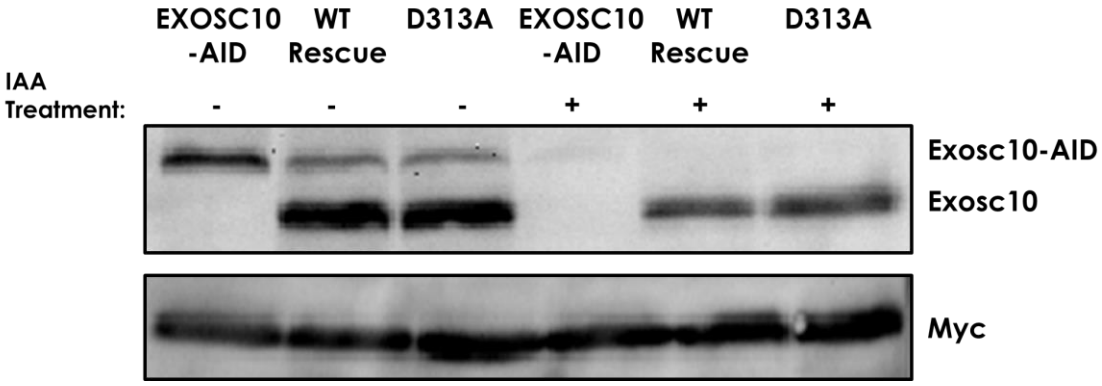


Figure 3.9: WT and D313A mutant Exosc10 overexpression protein screen. The EXOSC10-AID TIR1 parent cell line of which these lines were generated was used as a control. Each cell line was treated with either 500 μM IAA or ethanol (solvent) for 60 minutes prior to protein extraction. Anti-EXOSC10 antibody was used to detect both Exosc10-AID and the overexpressed rescue proteins whereas, an anti-Myc antibody was used as a loading control.

To confirm that the WT Exosc10 protein efficiently and functionally replaces the Exosc10-AID depleted protein, 5.8s rRNA processing was next considered. A northern blot assay was set up to measure the ratio of 5.8S+40 precursor RNA relative to mature 5.8S with the addition of the WT and D313A overexpression cell lines. As before, depletion of Exosc10-AID stimulated an accumulation of precursor 5.8S transcript (**Figure 3.10**[**B**]). Co-expression of WT Exosc10 effectively restored the ratio of precursor/mature 5.8S to a similar level as the unmodified parent TIR1 cell line, unlike the inactive D313A Exosc10 protein which accumulates a greater abundance of 5.8S+40 rRNA. Interestingly, cells expressing D313A Exosc10 show defective 5.8S processing even when the endogenously expressed Exosc10-AID protein is present, with a similar level of 5.8S+40 detected regardless of IAA treatment. This is probably because D313A is constitutively expressed and so outcompetes Exosc10-AID in some cases, blocking its activity. Given the importance of rRNA involvement in the translation of mRNA into protein, one would expect that mitosis, which requires regulated translation of numerous proteins at key points during the cell cycle to successfully coordinate cell division, would be severely disrupted in cells overexpressing the inert D313A Exosc10 mutant. However, it appears that the residual low levels of Exosc10-AID protein likely maintains a sufficient level of mature rRNA in order for cell proliferation to continue.

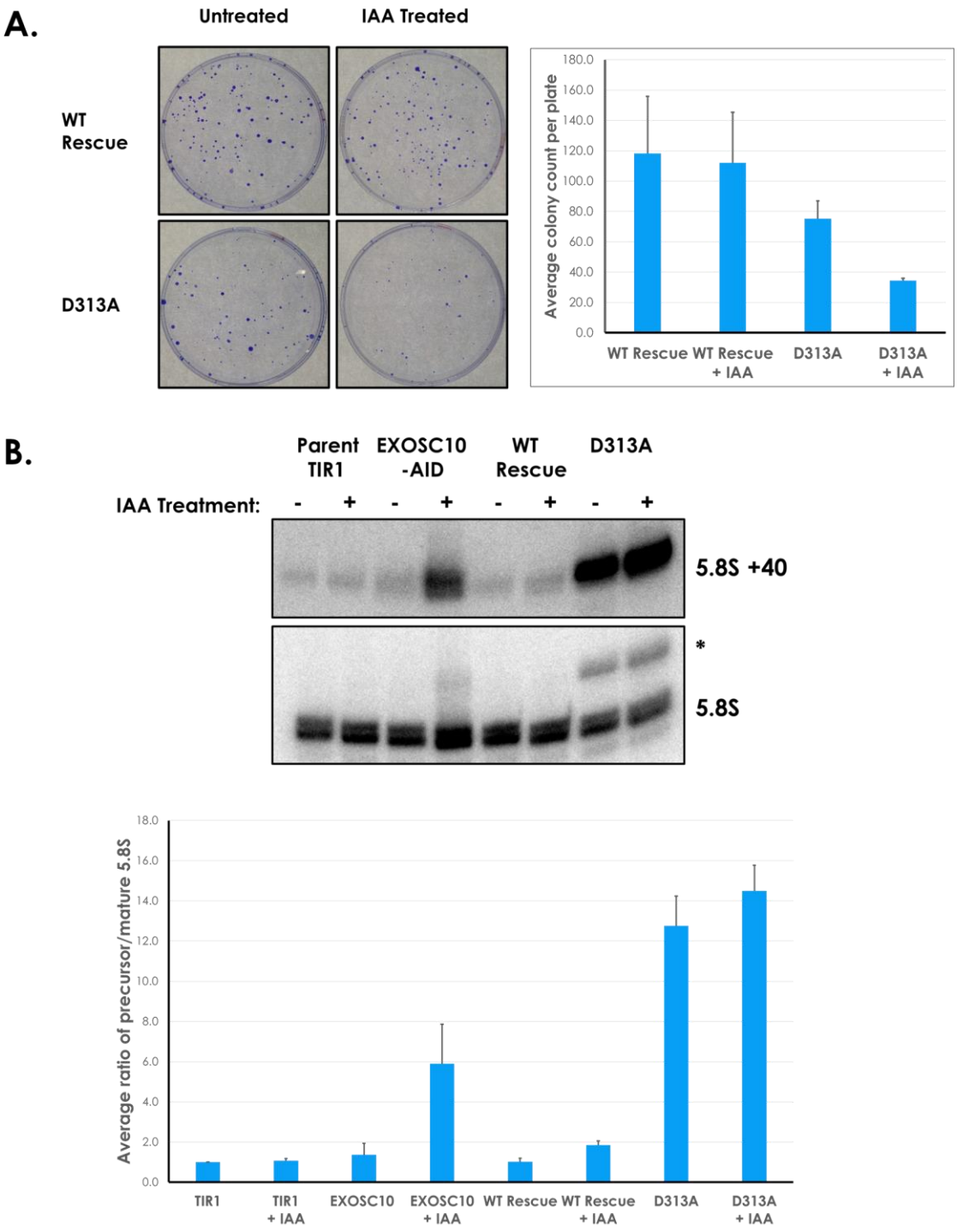


Figure 3.10: Colony formation assay for the WT rescue and D313A mutant overexpressed Exosc10-AID depleted cell lines after 10 days (**A**). Northern blot analysis of 5.8S processing including the WT and D313A overexpression cell lines (**B**). Precursor 5.8S rRNA was also detected with the mature 5.8S probe (*). In both experiments 500 µM IAA or ethanol (solvent) was added to the growth media. Graphs were plotted from 3 biological replicates (error bars = standard deviation).

3.6 Transcriptome-Wide Determination of Exosc10 Substrates

Compared with RNAi, the auxin-inducible degron system has the potential to elucidate the immediate effects of protein depletion, and hence gene downregulation. For an RNA processing enzyme like Exosc10, this can provide valuable information about the immediate RNA substrates that are processed and/or degraded by Exosc10 as well as allowing measurements of RNA abundance over much shorter intervals. Thus, the obvious potential overlapping activities and substrates that exist between the Exosc10 and Dis3 exosome components could finally be disconnected. To investigate the full impact of Exosc10 loss within humans, a transcriptome-wide RNA-Seq analysis of nascent RNA was undertaken.

3.6.1 Global Differential Gene Expression

Nuclear RNA was isolated from the EXOSC10-AID cell line following a 60 minute depletion of the Exosc10-AID protein, as this would maximise the detection of short-term fluctuations of RNA isoforms within the transcriptome. Following read alignment and filtering, basic statistical analysis of each RNA-Seq library was performed to determine the percent of genome coverage, average sequencing depth (per base) and read mapping efficiency by HISAT2 alignment software (**Figure 3.11**). Each library covered between ~40-70% of the genome with an average depth between ~1.5-2.5. This is consistent with the fact that each library represents the transcriptome where not all genomic elements are transcribed. Overall the unique mapping efficiency was consistently greater than 80% of reads sequenced.

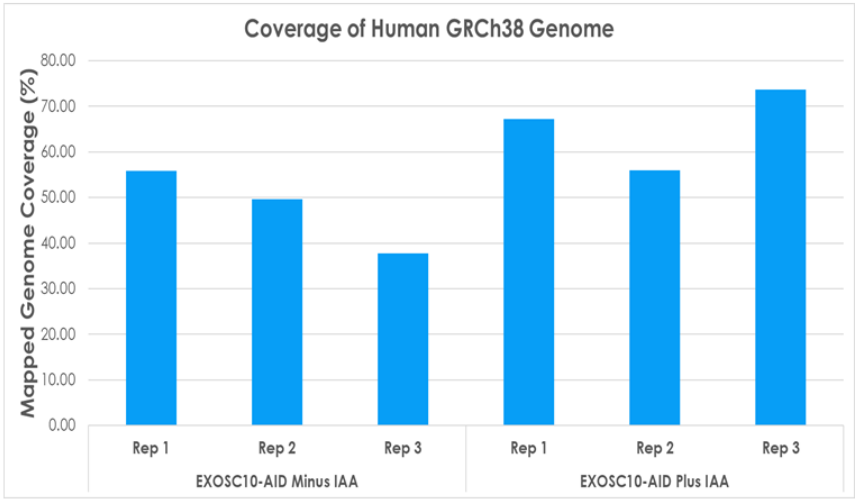
Once the reads were correctly aligned and filtered, the number of reads associated with each Ensembl annotated gene were counted. Importantly, the decision to measure expression at the gene level instead of at the transcript level was largely due to the role Exosc10 fulfils in RNA

degradation and 3' end processing which would provide much clearer interpretations compared to studying alternative transcript isoforms. Counted gene intervals were then passed to DESeq2, a program designed to identify differences in gene expression between samples.

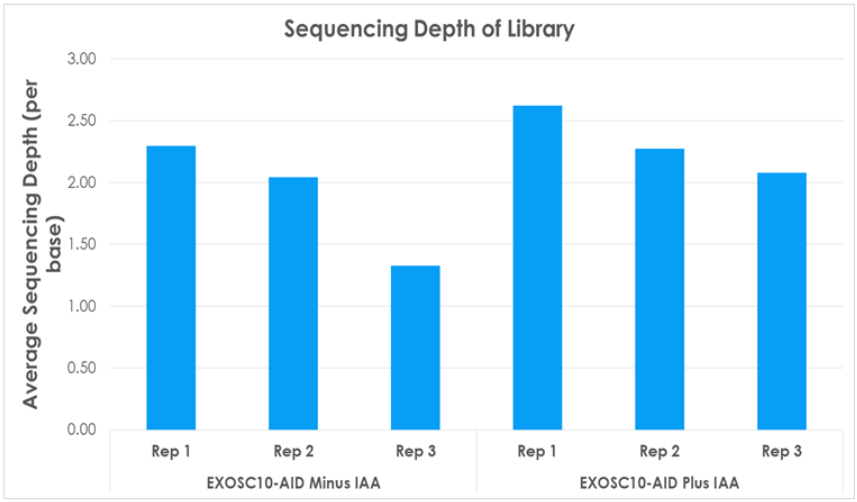
Unexpectedly, only a handful of differentially expressed genes were detected by DESeq2 (**Figure 3.12**). Each of the 7 genes discovered displayed a significant degree of upregulation in RNA abundance as a consequence of Exosc10-AID protein depletion (**Table 3.1**), and in some cases do not appear to be expressed under normal conditions (e.g. *CYP1A1*), becoming hyperactively expressed following EXOSC10 downregulation (**Figure 3.13**). Functional Gene ontology analysis (GO) revealed that for the most part, these genes are involved in the metabolism of uremic toxins (**Table 3.2**), chiefly the IAA present in the growth media (Sallee *et al* 2014) and, therefore represents stimulation of a metabolic pathway instead of RNA stabilisation through Exosc10 loss. Even so, this small number of indirect targets of IAA is much lower than the number of predicted indirect effects typically associated with RNAi (Qui *et al* 2005; Smith *et al* 2017).

Short-term depletion of Exosc10 did not cause any detectable transcriptome-wide alteration of nascent RNA abundance however, the library used to generate this data was depleted of rRNA transcripts prior to sequencing. From these observations so far, one can assume that the predominant role of Exosc10 in humans is almost exclusively concerned with rRNA processing (and perhaps turnover). This is consistent with the observed enrichment of Exosc10 within nucleoli where the bulk of pre-rRNA processing occurs (Lykke-Andersen *et al* 2011).

A.



B.



C.

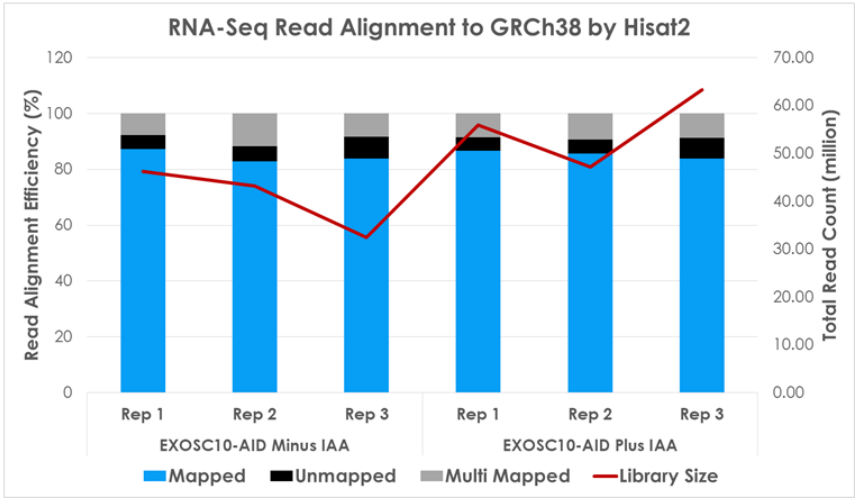


Figure 3.11: Graphical representation of (**A**) genome coverage, (**B**) average sequencing depth (per base) and (**C**) HISAT2 mapping efficiency for each replicate (Rep) of the EXOSC10-AID RNA-Seq libraries used in this study.

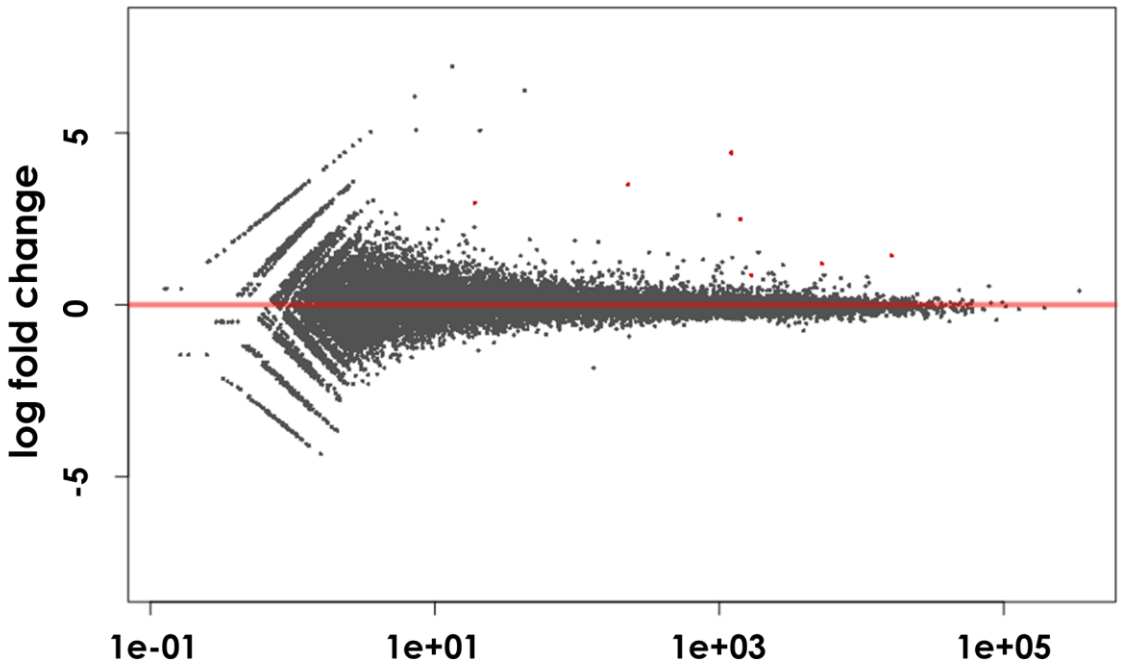


Figure 3.12: MA plot comparison of untreated and Exosc10-AID depleted differentially expressed genes (n = 58,302). Red coloured points represent genes with significant differential expression detected by the DESeq2 algorithm. Plot and analysis were generated from 3 biological replicates.

Table 3.1: Full list of differentially expressed genes upregulated in response to Exosc10 depletion (padj < 0.05).

Gene ID	Gene Name	log2FoldChange
CYP1A1	Cytochrome P450 1A1	4.42
CYP1B1	Cytochrome P450 1B1	3.49
CIITA	MHC Class II Transactivator	2.97
AHRR	Aryl Hydrocarbon Receptor	2.49
ALDH1A3	Aldehyde Dehydrogenase Family 1; Member A3	1.42
TIPARP	TCDD-inducible poly [ADP-ribose] polymerase	1.20
LEKR1	Leucine-, Glutamate- and Lysine-Rich Protein 1	0.87

Table 3.2: Pathways associated with the differentially expressed genes represented in **Table 3.1** including their false discovery rate (FDR) values. Gene Ontology (GO) analysis performed using the Panther online tool-set (http://pantherdb.org).

GO Biological Process	FDR
Retinoic Acid Biosynthetic Process	1.33E-02
Omega-Hydroxylase P450 Pathway	1.02E-02
Retinal Metabolic Process	1.41E-02
Epoxygenase P450 Pathway	2.54E-02
Toxin Metabolic Process	2.39E-02
Estrogen Metabolic Process	2.52E-02
Reactive Oxygen Species Biosynthetic Process	2.39E-02
Retinol Metabolic Process	3.53E-02
Xenobiotic Metabolic Process	1.28E-02
Cellular Response to Organic Cyclic Compound	1.37E-02

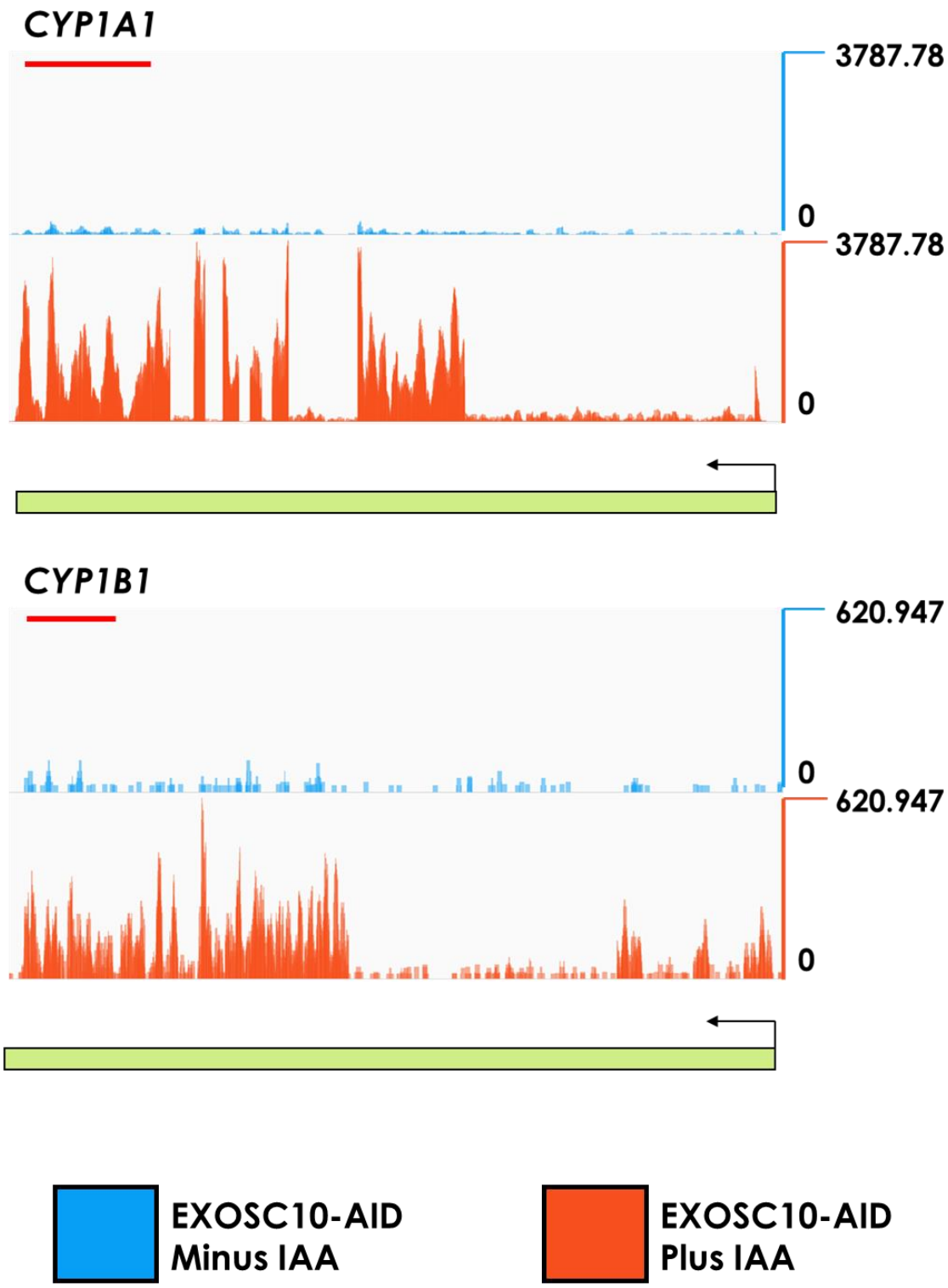


Figure 3.13: IGV produced read coverage tracks (normalised by RPKM) of two cytochrome P450 genes upregulated after depletion of Exosc10-AID protein. Each track represents 3 merged biological replicates. A red scale bars = 1 kb in length, applicable to both coverage tracks was applied to each gene image respectively.

3.6.2 Precursor snoRNA Processing

Similar to rRNA transcripts, snoRNA precursors also undergo 3' end trimming during maturation; a processing step that has been shown to involve either the activity of Rrp6 in yeast (Callahan & Butler 2008; Mitchell 2014) or Dis3 in humans (Szczepinska *et al* 2015). To address the discrepancy found in previous studies, the RNA-Seq library was next used to ascertain the involvement of Exosc10 in snoRNA processing.

Unlike rRNAs, reads mapped to snoRNA transcripts were particularly abundant within the data set, however due to their short transcript length (median length = 120 nt) metagene profiles of every snoRNA transcript expressed within HCT116 could not be generated with any degree of clarity. Furthermore, 50 nt RNA-Seq reads were used to form this analysis, reducing the resolution available to detect the short 3' extended pre-snoRNA isoforms. Instead, normalised sequence read coverage over a handful of snoRNA transcripts was used to visualise extension of snoRNA 3' ends. In each snoRNA transcript analysed, reads were shown to accumulate up to 30 nt downstream of the snoRNA transcript end site in cells lacking Exosc10, consistent with failed 3' trimming of pre-snoRNA transcripts (**Figure 3.14**). The length of this extension is similar to that for the 5.8S rRNA indicating that this may be a signature of Exosc10 loss.

This processing defect cannot be definitively attributed to the loss of Exosc10 protein function, as comparison of another auxin-inducible degron cell line generated for the Dis3 exosome protein displays a similar accumulation of reads 3' of TES (**Figure 3.15**). Interestingly, the impact of Exosc10 or Dis3 depletion has a varying degree of effect for each individual snoRNA transcript. For example, in the absence of Exosc10, reads immediately downstream of *SCARNA10*, *SNORD53* and *SNORA48* transcript 3' end sites show a greater enrichment compared with Dis3 depletion, however *SNORA21* has a much greater processing defect when Dis3 is depleted. While overlapping snoRNA substrate recognition appears to exist in this data, it is possible that Exosc10 and Dis3 target different subclasses of snoRNA transcript.

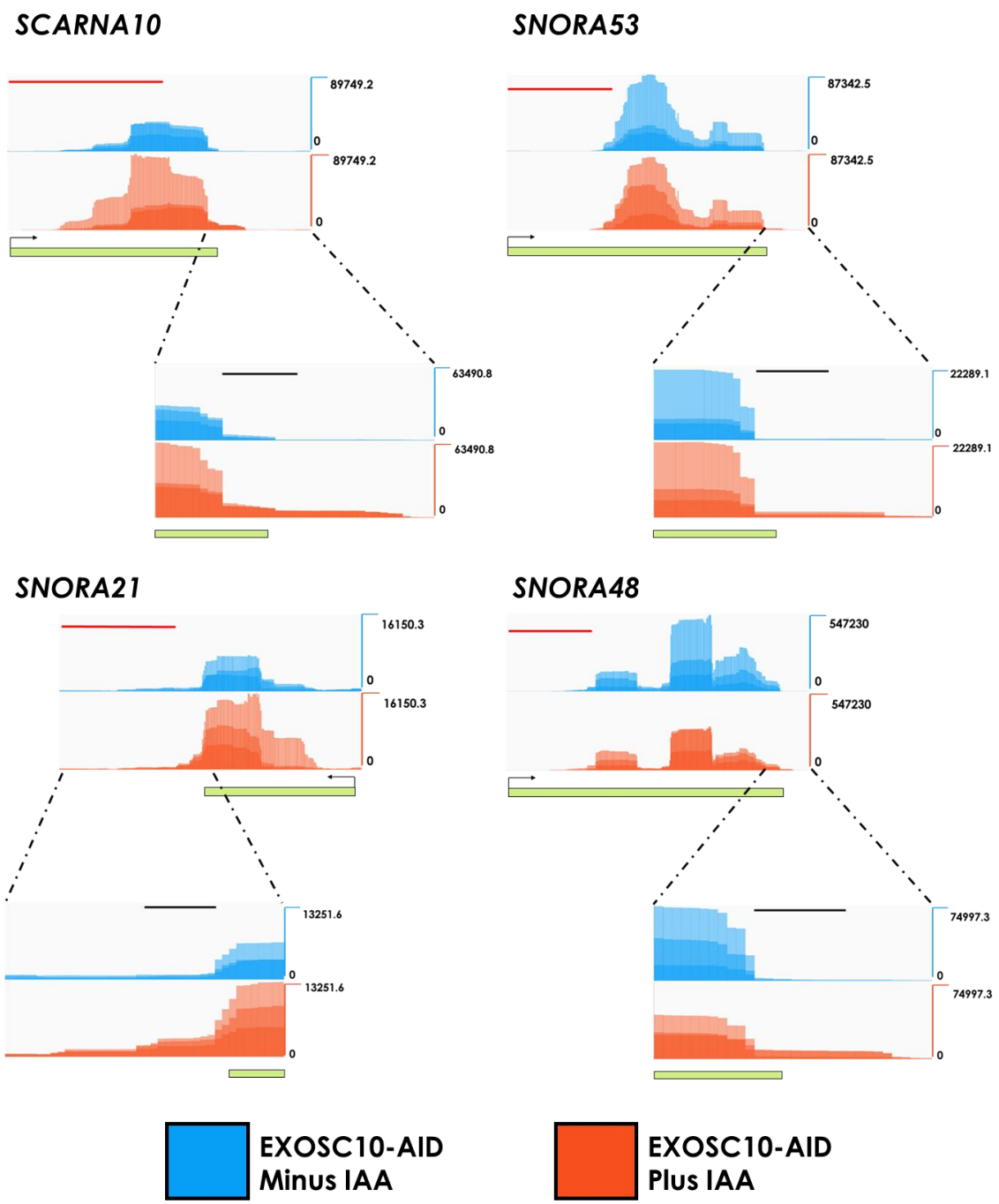


Figure 3.14: RPKM normalised read coverage plots of 4 snoRNA transcripts expressed in control and Exosc10-AID depleted cell lines, including an enhanced image of reads over the 3' end of each gene. Red scale bars = 100 nt, were applied to both coverage tracks for each snoRNA transcript. For coverage over the 3' flanking region, black scale bars = 10 nt were applied to both tracks for each snoRNA image. Images constitute 3 biological RNA-Seq replicates merged into a single coverage track.

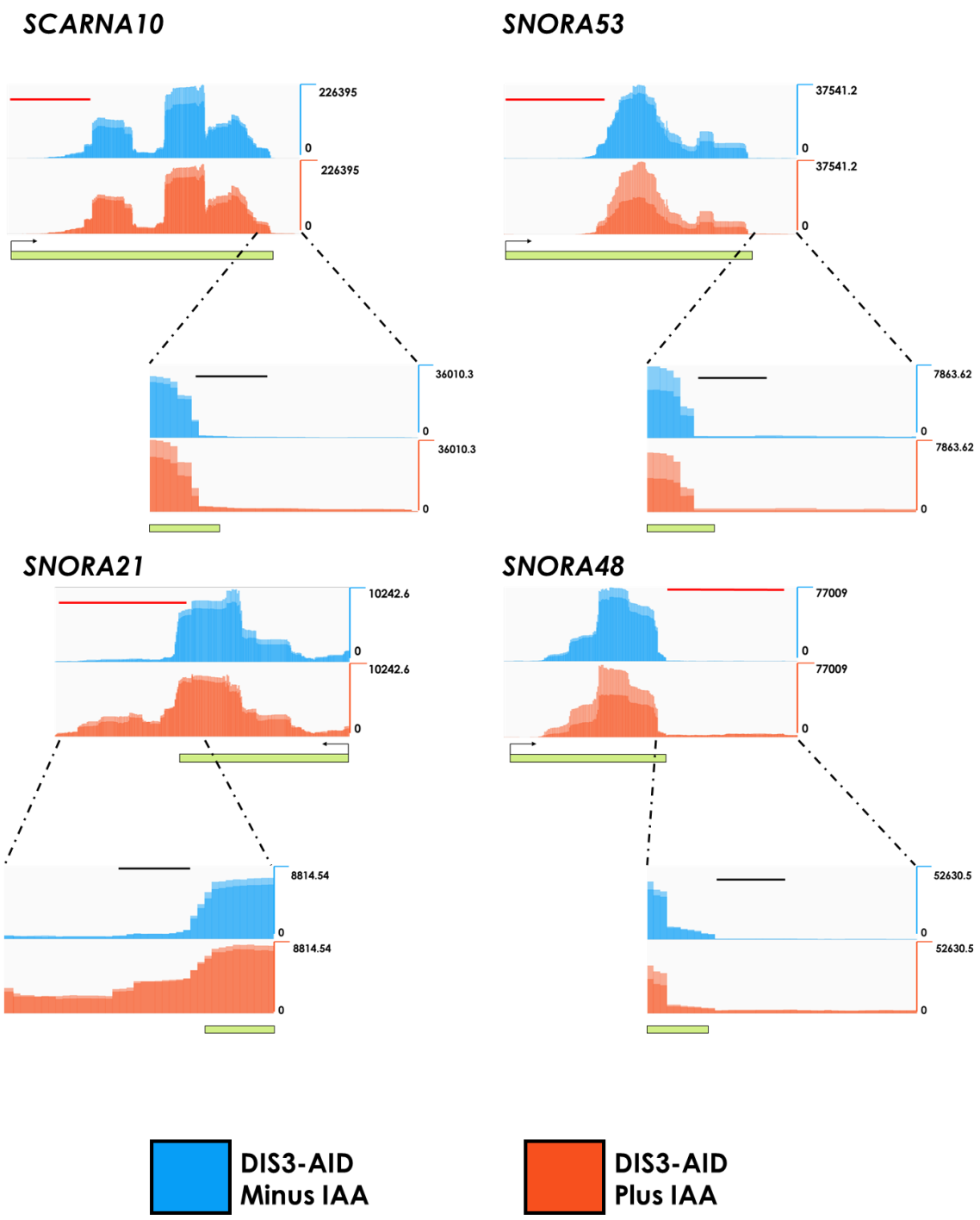


Figure 3.15: Control and Dis3-AID depleted, RPKM normalised read coverage plots of 4 snoRNA transcripts. The 3' downstream region was also included as an enhanced image for each gene. Red scale bars = 100 nt, were applied to both coverage tracks for each snoRNA transcript. For coverage over the 3' flanking region, black scale bars = 10 nt were applied to both tracks for each snoRNA image. Images constitute 2 biological RNA-Seq replicates merged into a single coverage track.

3.6.3 Is snoRNA Maturation a 2-Step Exosome Process?

Despite a clear stabilisation of reads downstream of snoRNA 3' ends in cells depleted of either Exosc10 or Dis3, there is a clear distinction concerning the length of snoRNA 3' extension. For example, in the absence of Exosc10 a short 10-30 nt 3' extension appears, whereas after Dis3 knockdown snoRNA transcript 3' ends can be extended up to ~100 nt (**Figure 3.14**; **Figure 3.15**). This disparity is even more apparent when both conditionally depleted exosome components are directly compared (**Figure 3.16**).

One possible explanation of the two distinct patterns observed over *SNORA68* may be due to snoRNA 3' end processing occurring as a 2-step mechanism involving both exoribonucleases. In this hypothesis, snoRNA transcripts are transcribed with a relatively long 3' terminal extension that is terminated by endonuclease cleavage at ~100 nt downstream of the TES. Although the exact mechanism of human snoRNA termination has not yet been fully described, termination may proceed in a similar fashion to structurally related snRNA transcripts involving the large multi-subunit integrator complex (O'Reilly *et al* 2014). This data supports the possibility that snoRNA transcripts are cleaved by an as yet unknown endonuclease, releasing the snoRNA from intron regions, therefore bypassing the need for intron debranching and trimming of flanking intron sequences. Following cleavage, Dis3 then trims the free snoRNA 3' end up to ~30 nt downstream of the TES before the precursor snoRNA is then transferred to Exosc10 which completes trimming forming the mature 3' end.

This theory is illustrated in the coverage plots shown in **Figure 3.16**[**B**] where, in the absence of Dis3, no precursor *SNORA68*+20 is detected owing to the lack of initial snoRNA trimming, instead the *SNORA68* 3' end is much longer. Conversely, initial 3' trimming by Dis3 still occurs in Exosc10 depleted cells however, removal of the final 20-30 nt sequence cannot proceed leading to an accumulation of the precursor *SNORA68*+20 isoform.

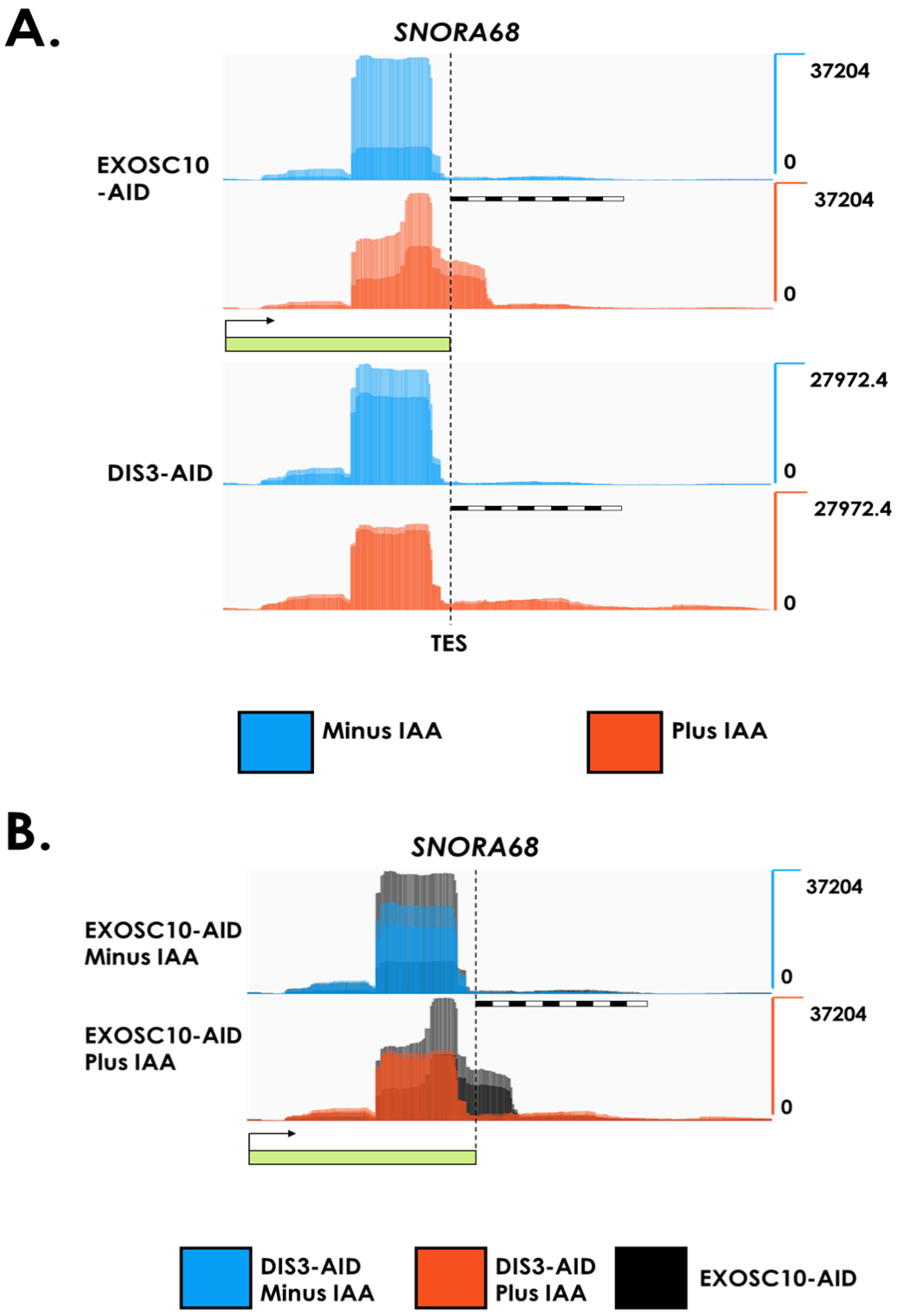


Figure 3.16: RPKM normalised coverage tracks aligning to the *SNORA68* transcripts in cells with either Exosc10 or Dis3 protein depletion (**A**). Tracks represented in (**A**) were overlaid for direct comparison on the same track (**B**). For each track 2 biological replicates were used. Scale bars = 100 nt, split into 10 nt segments apply to each coverage track within each image.

3.7 Summary

In this chapter, I have shown that CRISPR/Cas9 mediated fusion of an AID tag to the 3' end of the EXOSC10 is a relatively simple, reproducible and direct approach, providing a near complete and rapid rate of protein depletion (**Figure 3.5**) by taking advantage of proteasome mediated degradation pathways already present within humans. Furthermore, the ability to induce AID-mediated protein degradation has the added advantage of being reversible through removal of AUX/IAA from growth media and more importantly, has the potential to study the effects of immediate protein loss over short periods of time at a resolution unachievable by RNAi. Despite the detection of some indirect effects of AID-mediated protein depletion (which are easily explainable), expression of the plant specific TIR1 F-box protein and the fusion of the AID tag has very few deleterious effects within the HCT116 culture cells used in this study. As an extra precaution, smaller mini-AID tags can also substitute the AID tag used in this study improving protein stability which could be a potential concern for the analysis of other protein targets (Natsume *et al* 2016).

Contrary to yeast Rrp6 and many of the RNAi based knockdown methods utilised previously in human models, Exosc10 is an essential protein in humans that is required to maintain cell viability (**Figure 3.7**) however, the requirement of its enzymatic function within the cell is less certain. Surprisingly, over-expression of a catalytically inert mutant Exosc10 protein severely disrupts the processing of 5.8S rRNA (and perhaps numerous other rRNA transcripts) with greater potency than Exosc10 depletion alone. Furthermore, precursor rRNA processing defects are also visible in cells irrespective of AID-mediated Exosc10 degradation (**Figure 3.10**[**B**]). Despite the overexpression of a catalytically inert Exosc10, cells remained viable possibly indicating that a potential alternative structural role of Exosc10 either independently or as part of the exosome, is important for cell proliferation in humans.

Apart from its involvement in rRNA processing, few apparent RNA substrates of Exosc10 were detected from transcriptome-wide differential expression analysis (**Figure 3.12**). The exoribonuclease activity of Exosc10 must therefore have a greater involvement in the processing and degradation of small structured RNA substrates in humans, as evidenced by the reduced 3' trimming of a variety of snoRNA substrates. This is consistent with structural analysis of the Exosc10 active site which is much larger than yeast Rrp6 homologues and can accommodate interactions with RNA transcripts with more complex secondary structures (Januszyk *et al* 2011). While trimming of the small 10-30 nt extended isoform appears to exclusively require the activity of Exosc10, an additional role for Dis3 during snoRNA processing may also be present (**Figure 3.16**), however further experimental analysis is required to validate these findings since many of these small RNAs were not specifically enriched during library preparation and, as such have a varying degree of abundance within the library. The possibility of a 2-step snoRNA processing mechanism would also explain the confusion of exoribonuclease involvement during early snoRNA maturation observed from previous RNAi based approaches. Furthermore, two distinct classes of snoRNA transcript exist; HCA/A box or C/D box containing transcripts (Reichow *et al* 2007; Jorjani *et al* 2016), however so far only the former showed the greatest processing defect in the absence of Exosc10 protein.

Collectively, these observations emphasise the importance of Exosc10 as part of nuclear RNA surveillance, and in particular its involvement in rRNA and snoRNA processing. Furthermore, the specific RNA substrates detected in this data set remains consistent with the proposed nucleolar localisation of Exosc10 observed previously in metazoans (Januszyk *et al* 2011; Lykke-Andersen *et al* 2011). In the next chapter, the activity of the exoribonuclease Dis3 will be explored.

Chapter 4

Dis3 Prevents the Accumulation of Pervasive Transcripts

To thoroughly understand the role of the nuclear exosome complex in nuclear surveillance, attention was shifted to the Dis3 ribonuclease. As previously described, Dis3 is associated with the "bottom" of the EXO-9 double ring structure opposite to Exosc10, occupying a position close to the exit channel of the central channel (**Figure 1.3**). Similar to Exosc10, Dis3 is a 3'→5' exoribonuclease that possesses additional endonuclease activity and is capable of functioning in isolation. As a monomer, Dis3 degradation of RNA substrates requires binding to an unstructured 3' sequence of ~7-12 nt in length however, when in complex with EXO-9, Dis3 degradation requires ~30-35 nt of unstructured 3' RNA (Mitchell 2014; Kilchert *et al* 2016). Unwinding and threading of RNA substrates either by the Mtr4 helicase or Exosc10 threading, into the central channel therefore plays a critical role in regulating Dis3 activity.

In humans, genome-wide analysis has connected the degradation of a broad range of RNA substrates to Dis3 activity which can be selectively targeted to the exosome by the NEXT complex (Lubas *et al* 2011; Kilchert *et al* 2016). Dis3 dysfunction has also been shown to cause an accumulation of transcripts originating from non-protein coding regions of the genome such as enhancer sequences, and stimulates PROMPT RNA transcription by relieving biased promoter directionality (Preker *et al* 2008; Flynn *et al* 2011; Szczepinska *et al* 2015). Dis3 is regarded as the major processive exoribonuclease subunit of the nuclear exosome. I next wanted to investigate the immediate effects of Dis3 depletion compared to Exosc10 using the AID degron system as this provided the

best approach to uncouple their functional interactions, providing a mechanism to investigate only direct, immediate RNA substrates targeted by each exoribonuclease.

4.1 PROMPT Transcripts are Substrates of Dis3

Like the EXOSC10-AID cell line, a DIS3-AID cell line was designed and produced by Laura Francis. Initially the Dis3-AID protein was undetectable by western blot assay despite detecting unmodified Dis3 protein in parental HCT116 TIR1 expressing cells (**Figure 4.1**[**A**]). This is likely due to occlusion of antibody recognition at the C-terminal domain of Dis3. Switching to an AID tag specific antibody however, recognised Dis3-AID at the intended size only in DIS3-AID modified cells which is readily depleted within 60 minutes of IAA treatment (**Figure 4.1**[**B**]). To maintain consistency with the EXOSC10-AID RNA-Seq analysis (and to reduce the accumulation of indirect effects) DIS3-AID RNA-Seq libraries were generated from nascent nuclear RNA after 60 minutes of protein depletion following introduction of IAA.

Similar to the previous EXOSC10-AID analysis, each DIS3-AID library was checked to determine the level of genome coverage, depth and mapping efficiency (**Figure 4.2**). Genomic coverage and the average sequencing depth (per base) was lower in the DIS3-AID libraries (~30% and 1-1.5 respectively) compared to EXOSC10-AID (**Figure 3.11**), likely due to fewer sequenced reads in the libraries (which were ~50% smaller that EXOSC10-AID libraries) however, unique mapping efficiency was consistently greater than 80%. To verify loss of Dis3 function, I initially looked at PROMPT RNA transcription to determine if a comparison can be made to previous results observed by Szczepinska *et al*.

4.1.1 PROMPT Transcription is Detectable Following Dis3 Depletion

PROMPT transcripts originate as by-products of bidirectional promoter transcription and therefore transcribe on the opposite strand to their associated coding gene. To visualise PROMPT RNAs, the aligned reads were first separated by strand before calculating their normalised coverage between each sample. Following Dis3 depletion, a marked increase in PROMPT RNA transcription was observed upstream of the TSS site of coding genes (read coverage aligned over rose coloured bars shown in **Figure 4.3**). Interestingly, the stimulation of PROMPT transcription did not cause any observable downregulation of the associated protein-coding gene. It is likely that longer periods of Dis3 depletion (> 60 minutes) would cause downregulation of coding gene transcription, since degradation of PROMPTs have been shown to enhance transcription of associated coding genes (Ntini *et al* 2013). PROMPT transcription was undetectable in Exosc10 depleted cells, where their expression level remained similar to the untreated DIS3-AID control cell line, indicating that Dis3 alone is responsible for the turnover of PROMPT RNA. This result emphasises the benefits of rapid AID-degron mediated protein depletion, since previous RNAi bases studies originally show that both Dis3 and Exosc10 redundantly contribute to PROMPT RNA degradation (Preker *et al* 2008). This example therefore highlights the potential indirect or redundant RNA degradation pathways that may become activated as a consequence of gradual, prolonged gene downregulation via RNAi, which obfuscates the detection of *bona fide* exoribonuclease specific substrates.

To determine how widespread PROMPT transcription was in Dis3 depleted cell lines, the aligned reads were assembled into *de novo* transcripts, guided by known gene annotation to prevent the production of chimeric transcripts. Although PROMPT transcripts have a characterised length between ~200-600 nt (Preker *et al* 2011), transcription of PROMPT RNA has been demonstrated to originate up to ~3 kb from the

downstream TSS (Preker *et al* 2008; Flynn *et al* 2011; Szczepinska *et al* 2015). For this reason, the *de novo* transcripts were defined as PROMPT RNA if they originated < 3 kb from the nearest TSS and were upregulated as a consequence of Dis3 knockdown (≥ 2-fold). By these criteria I detected 1092 potential PROMPT transcripts, however due to the limitations of the software and the evidence that all eukaryotic promoters are inherently bidirectional, this number is likely a significant underestimation of the overall level of PROMPT RNA output in humans.

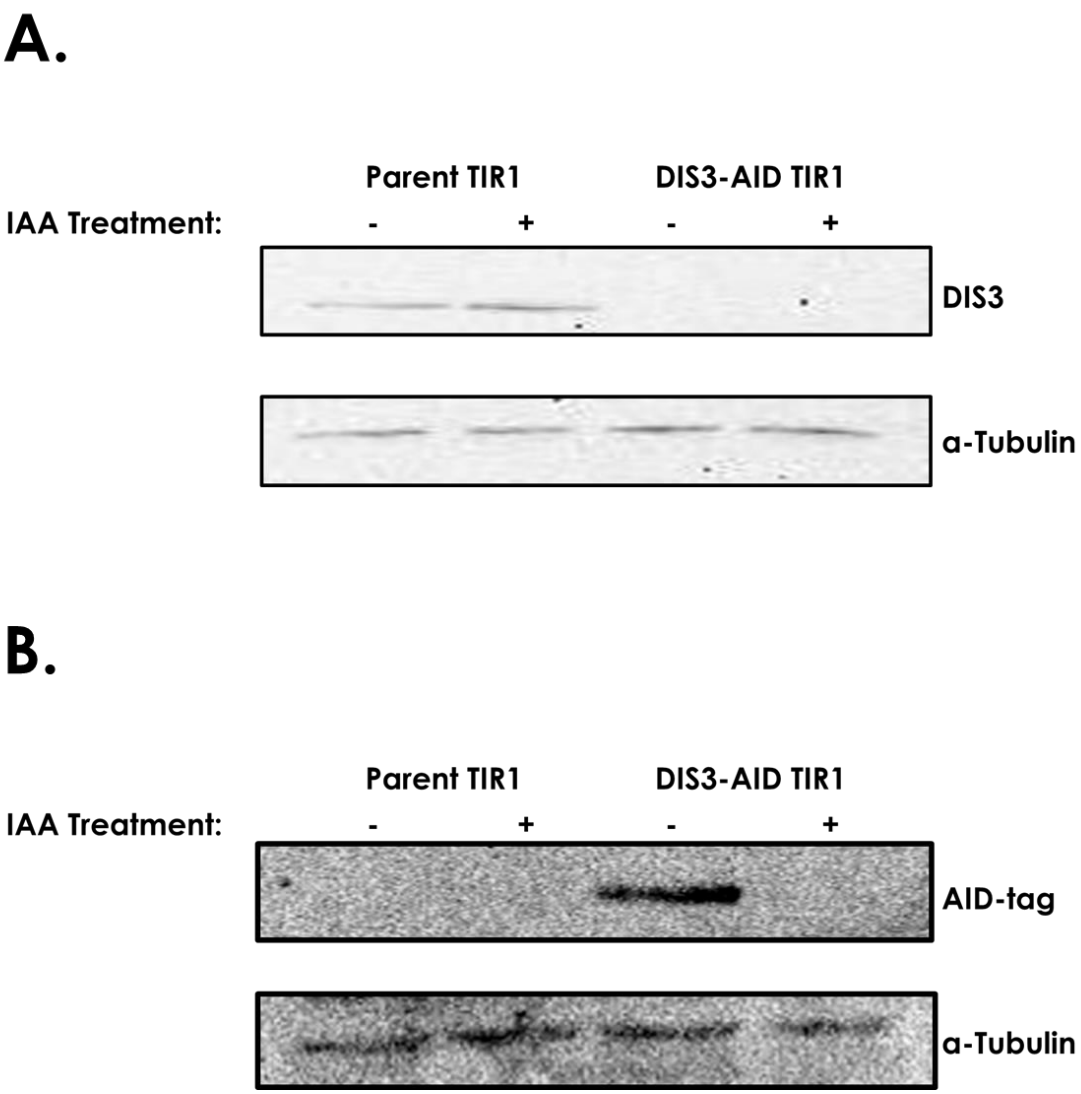
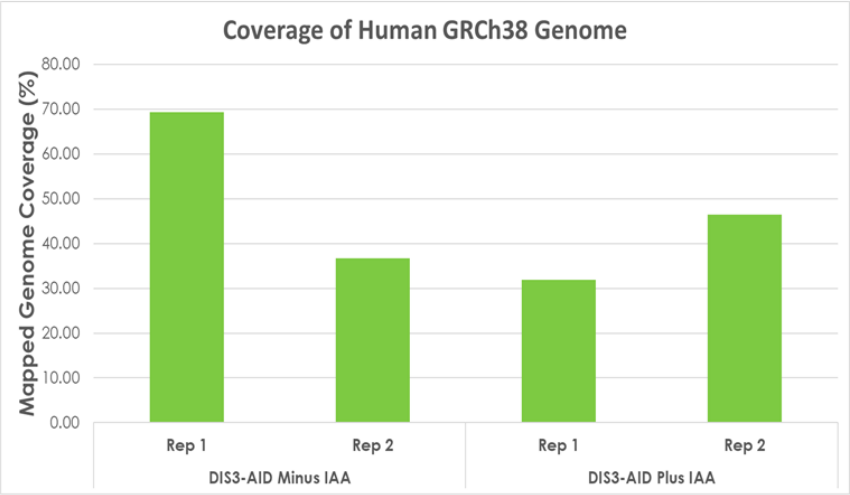
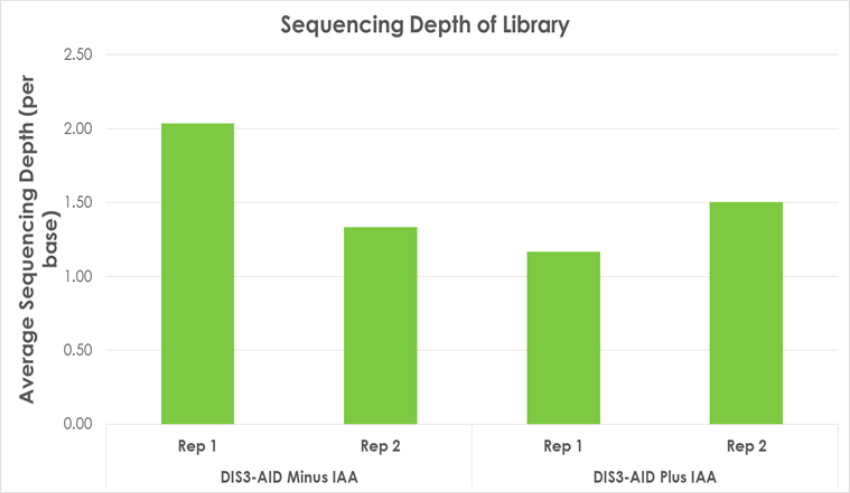


Figure 4.1: Western blot screening of parental HCT116 TIR1 and DIS3-AID tagged cell lines using antibodies directed against Dis3 (**A**) or the anti-AID tag (**B**) proteins. Each cell line was treated with either 500 μM IAA or ethanol (solvent) for 60 minutes prior to protein extraction. Both western blots assays were performed by Laura Francis.

A.



B.



C.

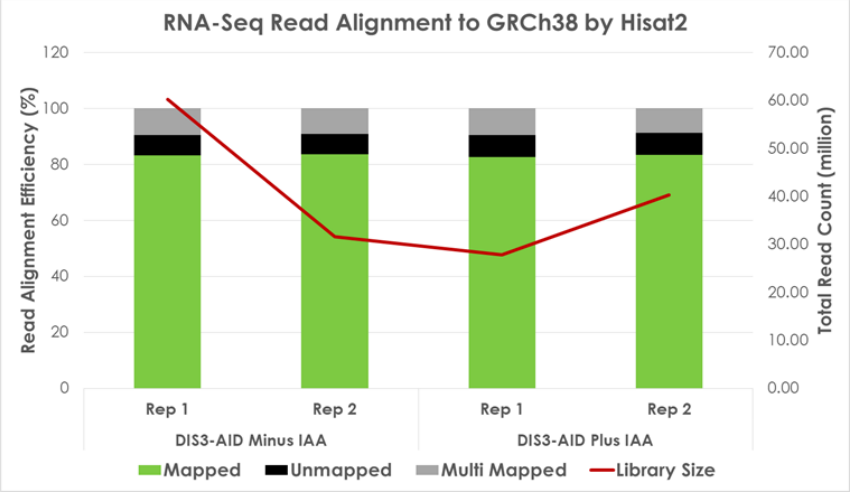


Figure 4.2: Graphical representation of (**A**) genome coverage, (**B**) per base sequencing depth and (**C**) HISAT2 mapping efficiency for each replicate (Rep) of the DIS3-AID RNA-Seq libraries used in this study.

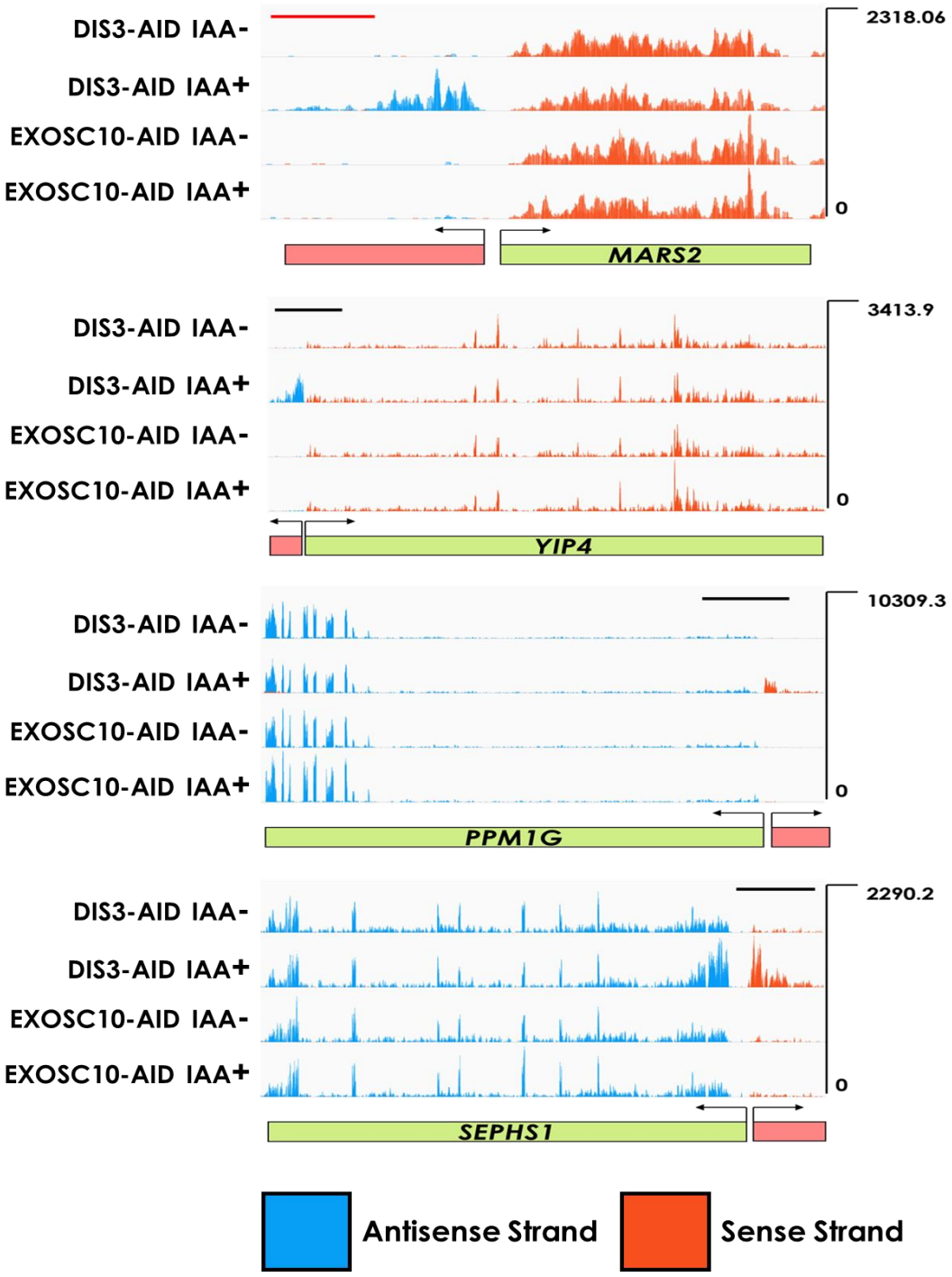


Figure 4.3: RPKM normalised read coverage tracks of split stranded reads in DIS3-AID and EXOSC10-AID +/- IAA treated cell lines. Each coverage track represents the average read count of 2 biological replicates. The red scale bar = 1 kb, is applicable to every coverage track over *MARS2*. Black scale bars = 5 kb, apply to all 4 coverage tracks within each of the remaining gene images respectively. PROMPT transcripts are represented by rose coloured boxes adjacent to green genes.

4.1.2 Dis3 Stabilises Promoter Proximal Transcripts in the Coding Direction

Although visualisation of PROMPT transcription did not detect any downregulation within the associated gene body, I next decided to investigate possible alterations in transcription in response to Dis3 downregulation. To do this, all annotated genes were extracted from the annotation and filtered for expression by removing any low or unexpressed genes. To measure the transcription profile of both PROMPT and the associated gene, an inclusion window was incorporated around each gene, in effect shifting the TSS 3 kb upstream and the TES 7 kb downstream from their original positions (only 3 kb downstream of the TES was shown for clarity). Any genes that overlapped as a consequence of the inclusion window application were dropped, reducing the pool of expressed genes for testing to 4701. The purpose of the exclusion window was to ensure that any transcripts overlapping more than one gene were discounted to minimise false-positive discovery of Dis3-dependent changes. From this, a metagene plot representing the average transcription profile over every expressed non-overlapping gene was produced.

Depletion of Dis3 has very little impact on the transcription of the gene body and no influence downstream of the TES (**Figure 4.4**). Similar to the visualisation of individual genes, PROMPT transcription is significantly more prevalent in the absence of Dis3 (compare orange line to others). In agreement with Preker *et al* (2011), PROMPT expression peaks between ~0.5-1 kb upstream of the TSS before gradually dissipating at ~3 kb upstream of the TSS, indicating that the majority of PROMPT transcripts are relatively short and undergo termination proximal to the TSS. As expected, no PROMPT transcription was detected following Exosc10 depletion. The magnitude of PROMPT upregulation observed in this system is more than I had typically observed previously, which again highlights the benefits of the AID system in cases such as this.

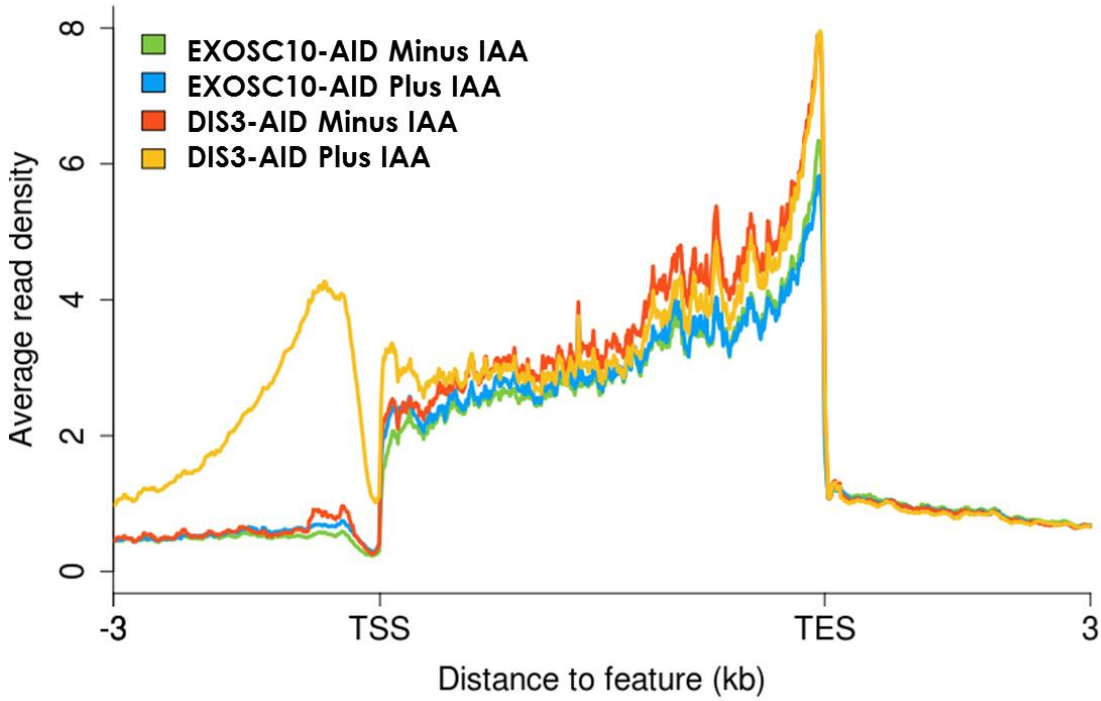


Figure 4.4: Metagene read coverage profile comparison of non-overlapping expressed genes with a 3 kb inclusion window flanking the TSS and TES with a gene body scaled to 5 kb (n = 4701). Profile represents 1 biological replicate; an additional replicate is represented in **Supplementary Figure S4**.

4.2 Gene Expression is unaltered by Dis3 Depletion

As mentioned in the previous chapter, gene expression remains unaltered after Exosc10 downregulation. Unlike Exosc10, Dis3 can recognise and degrade a much broader range of RNA substrates including both coding and non-coding transcripts. Differential gene expression was next used to identify stabilised nascent RNA transcripts over all annotated protein-coding and non-coding genes.

4.2.1 Differential Gene Expression Analysis

Using the same gene annotation list (composed of ~58,000 Ensembl annotated genes) as the earlier Exosc10 differential expression analysis, ~3200 genes were shown to be significantly upregulated in Dis3 impaired cells (**Figure 4.5**[**A**]). Each of these upregulated genes were then categorised based on their transcript biotype to determine if any particular group of RNA is more susceptible to Dis3 degradation. Dis3 appears to degrade protein-coding mRNA, antisense and non-coding RNAs such as lincRNA indiscriminately, as demonstrated by the almost equal proportions of transcripts stabilised after Dis3 depletion (**Figure 4.5**[**B**]). Whether accumulation of these transcripts is due to reduced RNA turnover or degradation of aberrant nascent transcript isoforms by Dis3 will be investigated in the next section.

4.2.2 False Discovery of Differential Expressed Genes

To verify the upregulated transcripts detected from the differential expression analysis, genes were ranked based of their relative log2 fold-change and a handful of the top differentially expressed genes were then visualised. While it is true that the DESeq2 software was able to identify significantly upregulated transcripts based on their associated gene intervals, many of these candidates represent false positives within the data set.

On closer inspection these upregulated transcripts are attributed to internal cryptic promoter transcription within intron sequences, similar to observations made by Kaplan *et al* (2003) in *S. cerevisiae* (**Figure 4.6**[**A**]), spurious transcription of proximal intergenic regions (**Figure 4.6**[**B**]), or from PROMPT transcription originating from nearby genes within clustered loci (**Figure 4.6**[**C**]). In every case presented in **Figure 4.6**, there is almost no expression of these genes in the untreated control cell line, arguing against the possibility of alternative expression in the wake of Dis3 downregulation. Furthermore, it appears that although Dis3 does not affect the expression of nascent RNA *per se*, it is critical for the maintenance of a coherent transcription landscape by preventing expression of cryptic promoter products and ensuring that clustered genes are correctly punctuated, reducing the likelihood of generating nonsense or missense transcripts. This observation might explain why a previous study suggested function for Dis3 in the turnover of some mRNAs (Dziembowski *et al* 2007).

How many of the original differentially expressed genes detected are caused by such transcription dysfunction? In order to address this, a new list of genes was generated for the differential expression analysis - one that will take into account nearby annotated genes that may express a PROMPT transcript upon Dis3 downregulation. Comparable in approach to the metagene analysis gene list generated, the total annotated list of genes were filtered to remove closely spaced, potentially clustered genes that were located ≤ 3 kb from the nearest TSS and/or TES. In effect, this reduced the list of genes for analysis to ~12,000. While this approach was valid for the removal of overlapping PROMPT expression, cryptic internal transcription still remained a possible source of false positive hits.

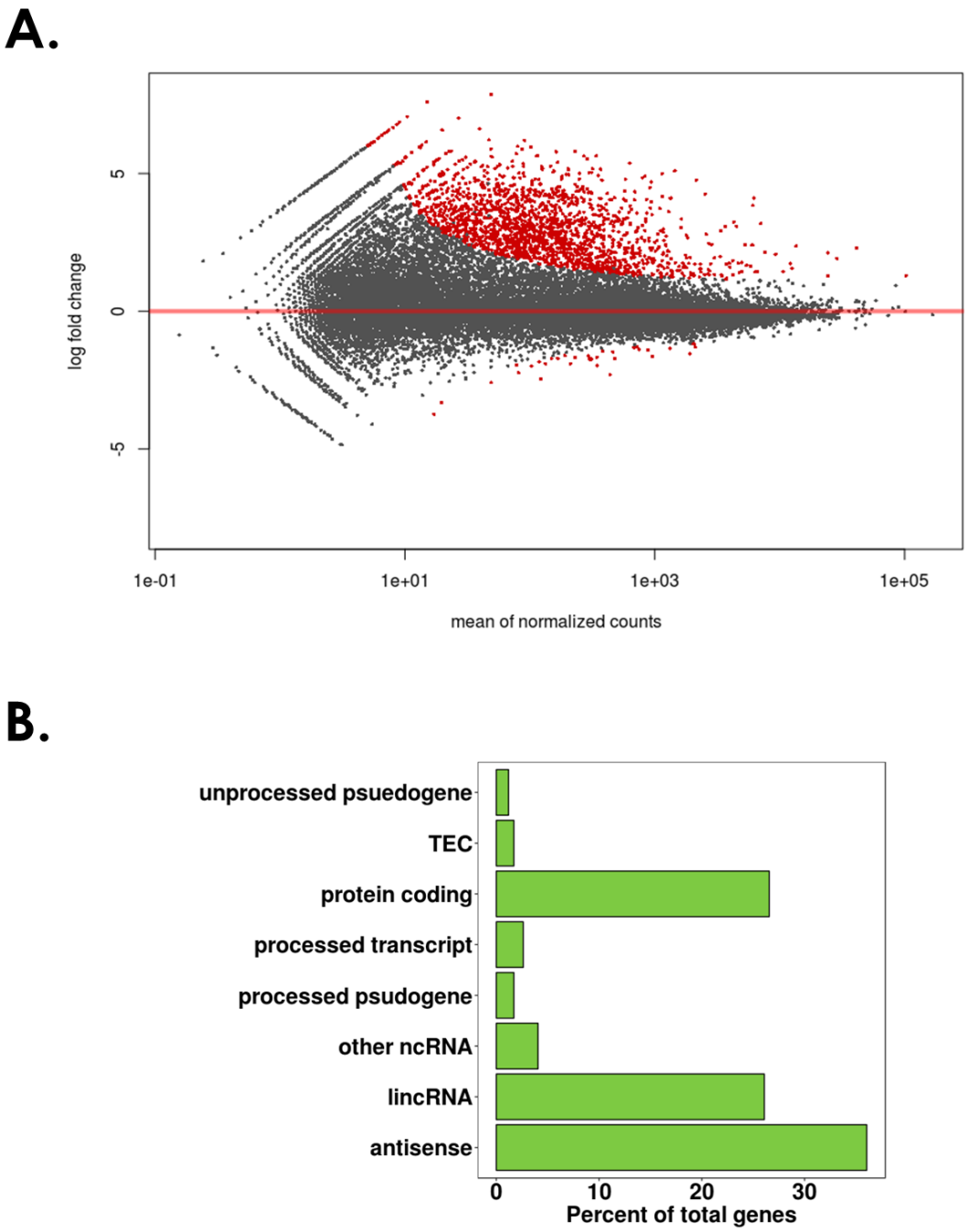


Figure 4.5: (**A**) MA plot representation of differentially expressed genes in Dis3-AID protein depleted cells compared with an untreated control (n = 58,302). Red coloured points represent genes with significant differential expression according to DESeq2. (**B**) Categorisation of upregulated genes (≥ 2-fold, padj < 0.05) in Dis3 depleted cells based on transcript biotype (n = 3279). Analysis was generated from 2 biological replicates. TEC = To be Experimentally Confirmed, are non-spliced transcripts with poly(A) features.

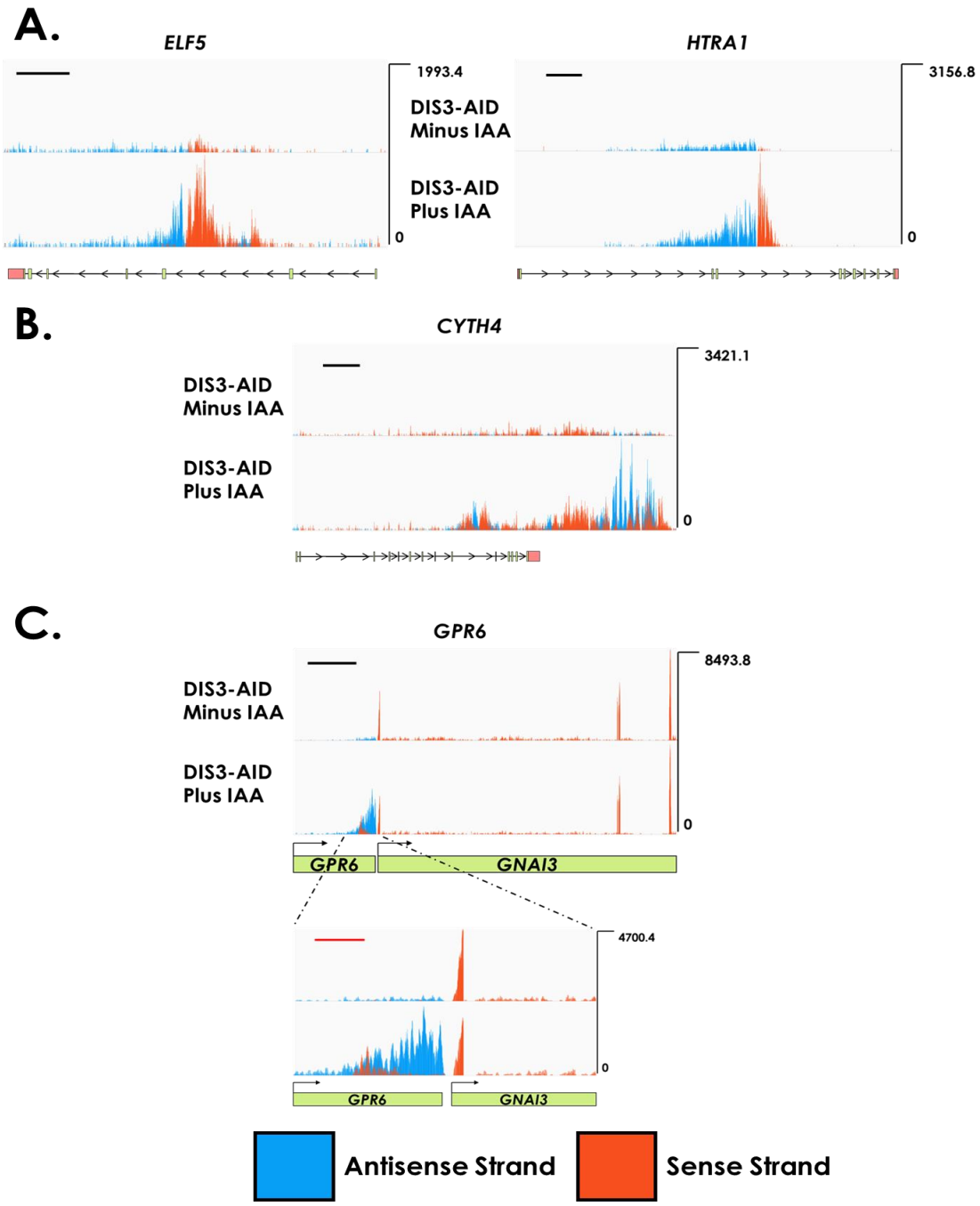


Figure 4.6: False positive upregulated differentially expressed gene coverage tracks derived from detection of divergent intronic transcription (**A**), spurious transcription of nearby open chromatin (**B**) or from PROMPT transcript stabilisation of downstream genes (**C**). Coverage was normalised by RPKM using 2 biological replicates. Black scale bars = 5 kb, were applied to both coverage tracks within each gene image respectively. The red scale bar shown over GPR6 = 1 kb and applies to both coverage tracks. Rose coloured boxes represent UTR regions.

Subsequent differential gene expression analysis was only able to detect 266 upregulated (≥ 2-fold) genes after confidence filtering (**Figure 4.7**[**A**]). Again, these genes were categorised according to their transcript biotype however, while mRNA transcripts still form a large proportion of these stabilised transcripts (~20%), lincRNAs are much more dominant in the data set representing ~50% of the 266 transcripts identified (**Figure 4.7**[**B**]). This is consistent with the notion that lincRNAs are generally located outside of annotated protein-coding gene clusters at distant intergenic loci (Guttman *et al* 2009; Khalil *et al* 2009). Furthermore, far fewer antisense transcripts are present in this dataset, indicative of their removal during initial filtering of the gene list. Finally, both differential upregulated gene data sets were compared to determine how many of the original genes could be considered as true Dis3 upregulated genes. Almost 97% of the genes detected from the non-overlapping differential expression analysis are present within the original dataset, demonstrating the high level of probable false positive Dis3 substrates detected from the analysis (**Figure 4.7**[**C**]).

Although the immediate loss of Dis3 does not radically alter the abundance of nascent coding or non-coding RNA within 1 hour, it does however play a more direct role in the turn-over of numerous spurious RNAs deriving from what appear to be cryptic transcription events. Transcription initiation from internal cryptic start sites has been described previously in budding yeast whereby, loss of transcription elongation factors such as Spt6 failed to restore normal chromatin structure left in the wake of transcribing Pol II complexes, leaving a chromatin landscape permissive for transcription initiation (Kaplan *et al* 2003). Rapid AID-mediated Dis3 depletion therefore provides evidence of cryptic initiation in human cells, where chromatin architecture has not been directly altered. From these initial findings, I speculate that prolonged Dis3 depletion, in combination with downregulation of the human Spt6 homologue, Supt6h, would most likely increase the abundance of many of these internal unstable pervasive transcripts.

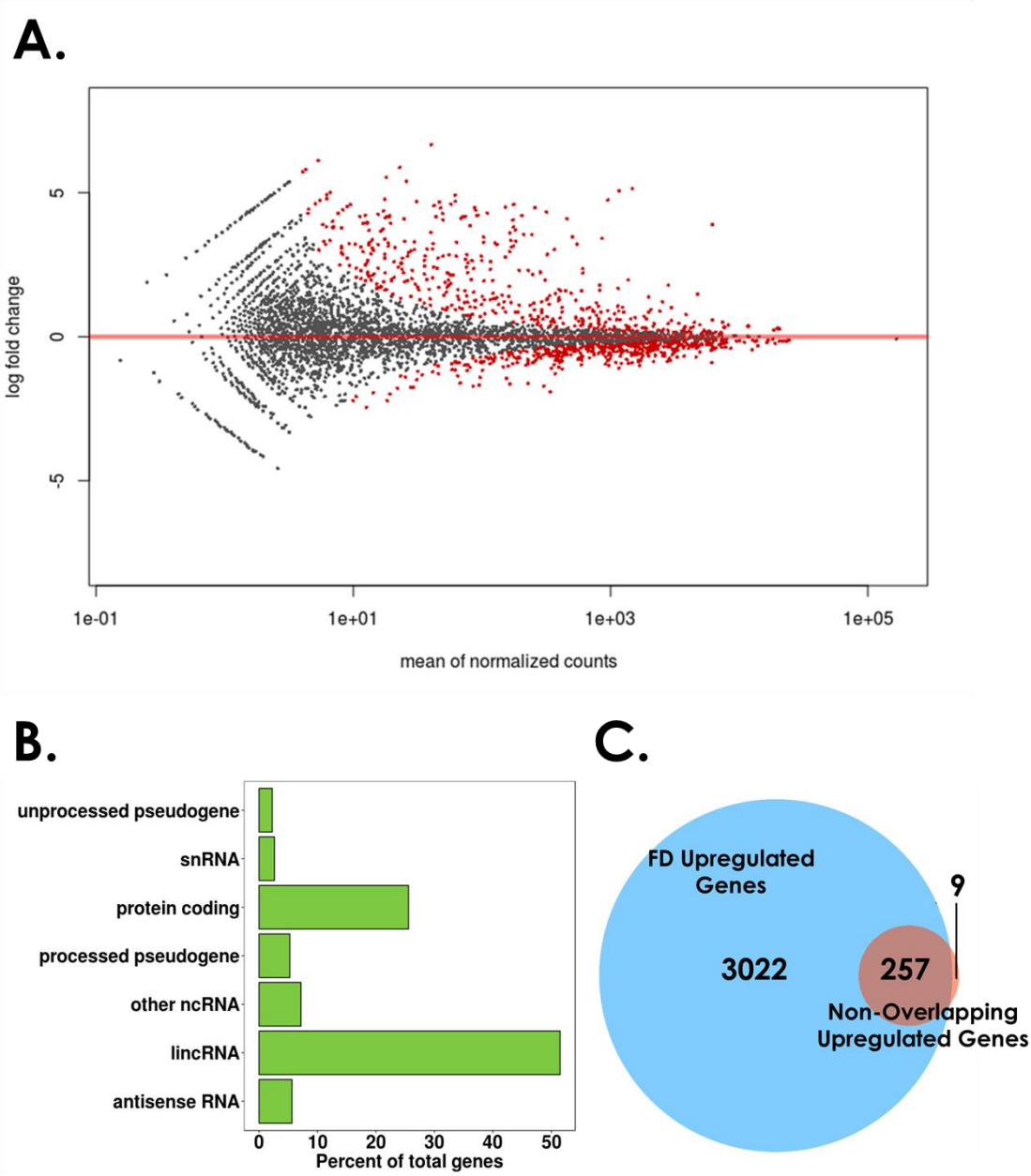


Figure 4.7: An inclusion window of 3 kb flanking each gene was included before filtering to remove overlapping gene intervals. Differential expression analysis of 12,327 non-overlapping genes are displayed as a MA plot (**A**), red pixels represent significantly altered expression levels prior to filtering. Categorisation of 266 upregulated genes (≥ 2-fold, padj < 0.05) based on transcript biotype (**B**). Comparison of false discovered (FD) and non-overlapping differentially expressed genes (**C**).

4.3 Dis3 Degrades RNA Derived from Premature Transcription Termination

Continuing with the realisation that downregulation of Dis3 expression facilitates the accumulation of cryptic unstable RNA molecules from intronic sequences, I next examined the abundance of intragenic sequence elements such as introns and exons, as a mechanism of understanding the global shift in transcription in response to Dis3 depletion. Moreover, given the strong 5' bias observed in the earlier metagene transcription profile (**Figure 4.4**), I also wanted to determine if the proximity of intron and exon sequences relative to the TSS is a key factor that could explain the 5' expression bias.

To measure the abundance of exonic and intronic RNA, the depth of sequencing coverage was achieved by counting aligned reads over custom synthetic intervals. Synthetic exonic intervals for example, were generated by collapsing every transcript isoform associated with each gene into a single synthetic transcript representative of every combination of exon sequence. These synthetic transcripts were then subtracted from the gene interval producing synthetic introns. Gene names and IDs were attached to each of the synthetic exons and introns which were also numbered based on their position relative to the TSS thus, providing a method of tracing the heritage of each intragenic interval back to their parent gene.

Downregulation of Dis3 protein expression causes a marked increase (~1.5-fold) in read coverage over the first intron sequence of every annotated gene compared to untreated counterpart cell lines, represented as a ratio (**Figure 4.8**[**A**]). Interestingly, calculation of sequencing depth over the final intron region did not detect the same stabilisation, indicating that RNA originating from the first intron sequence is more sensitive to Dis3-mediated degradation. Additionally, I did not observe any change in sequencing coverage aligned to the first or last exon region as a consequence of Dis3 downregulation. This is perhaps due

to exons being more stable compared to introns and therefore would not undergo any significant change in expression/stability within the 60 minutes of Dis3 downregulation allocated in this study.

Following the discovery that intron 1 RNA sequences are preferentially degraded by Dis3, I next wanted to determine if subsequent downstream intron regions are also stabilised. However, due to the heterogeneous composition of genes (i.e. gene length, intron length or intron number), I decided to investigate the global change in sequencing coverage by comparing the total annotated synthetic introns directly against the total synthetic exon regions. Depletion of Dis3 causes a significant (~4-fold) stabilisation of intron RNA sequences (**Figure 4.8**[**B**]), indicating that RNA derived from introns are sensitive to Dis3-mediated degradation. Collectively, these initial results indicate that nascent intron RNA sequences commonly associated TSS-proximal introns are substrates of Dis3 mediated degradation. Finally, Exosc10 was included to confirm that stabilisation of RNA over the intronic intervals are exclusive to Dis3 downregulation.

Since stabilisation of intronic regions may not be common to all genes, I decided to look more closely at a subset of genes with significant upregulation of intronic RNA reads in order to determine the pattern of transcription. Initially, differential expression was performed on every synthetic intron (**Supplementary Figure S5**), detecting 9252 upregulated introns (≥ 2-fold) originating from 6159 genes. Of these genes the 7 false positive IAA responsive genes determined by the Exosc10 differential gene analysis were removed. Additionally, any gene containing an internal annotated gene such as a snoRNA was then dropped from the analysis leaving 4356 genes which were then used to produce transcription metagene profiles across their entire gene body. In the absence of Dis3, a robust accumulation of RNA associated with the 5' end of the gene is seen (**Figure 4.8**[**C**]). As the transcription profile progresses towards the TES, the level of RNA coverage gradually decreases becoming comparable to untreated cells. Supporting this, separate metagene profiles produced

from the first and last intron interval from each of these genes highlight the stabilisation of RNA originating from introns proximal to the TSS with almost no stabilisation of the final intron (**Figure 4.9**). The gradual decline of read coverage over the gene body is consistent with the notion that the majority of these stabilised transcripts have arisen through premature transcription termination downstream of the TSS. However, I cannot rule out the possibility that in some cases, upregulated intron 1 regions are caused by overlapping stabilisation of nearby PROMPT transcripts originating from upstream divergent genes, which may allow spurious transcription by maintaining a NFR. However, if overlapping PROMPT RNAs were a common cause of intron 1 accumulation however, PROMPTs derived from downstream tandem genes should also potentially overlap the TES region however, no apparent change in coverage over the TES site was detected in either metagene plot represented in **Figure 4.8**(**C**) or **Figure 4.9**(**B**).

This data emphasises the importance of Dis3 during the early stages of transcription, indicating the possibility of exosome recruitment to the PIC prior to transcription initiation, providing a mechanism to rapidly degrade pervasive or abortive transcripts, perhaps even co-transcriptionally. As these promoter proximal transcripts accumulate so quickly within just 60 mins of auxin treatment, it can be speculated that premature termination occurs very frequently on a large number of genes.

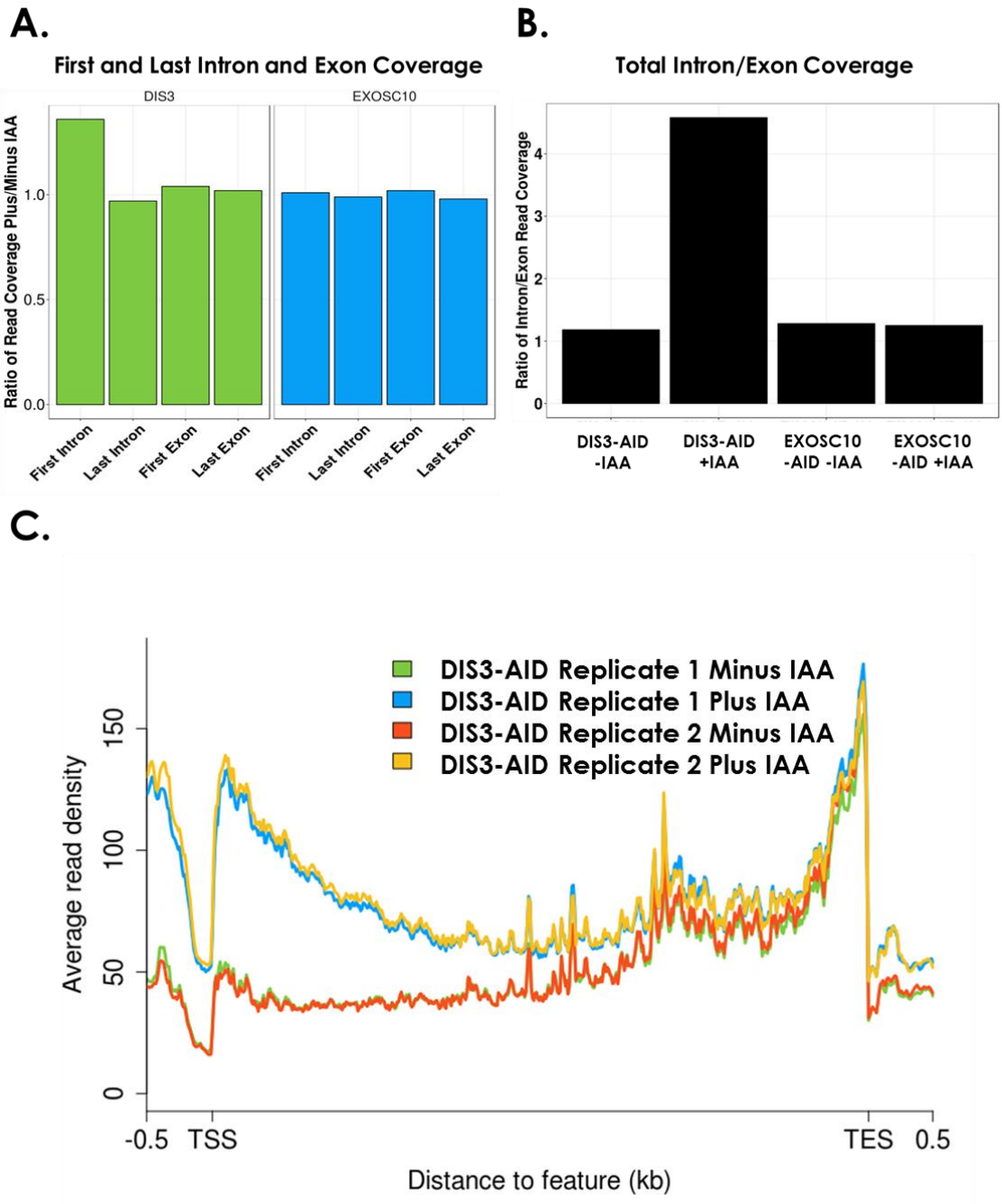


Figure 4.8: (**A**) Analysis of RNA-Seq read coverage over the every first and last intron and exon after 60 minutes of exosome protein depletion, represented as a ratio of plus/minus IAA treatment. (**B**) Ratio of read coverage over every annotated synthetic intron and exon. (**C**) Coverage analysis of 4356 non-overlapping genes detected from intronic differential expression analysis, gene body scaled to 5 kb.

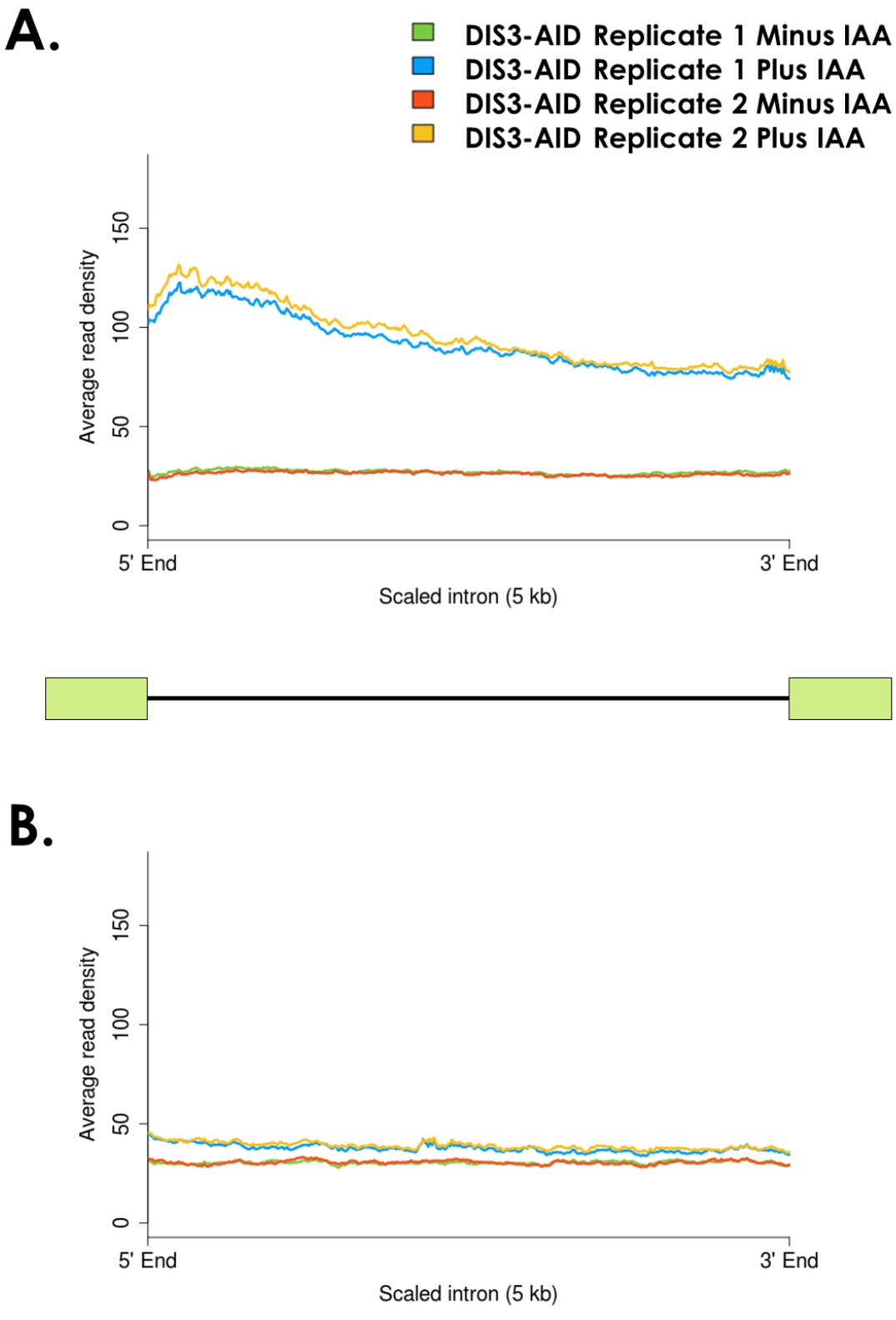


Figure 4.9: Metagene coverage of the (**A**) first intron (n = 4356) and (**B**) last intron (n = 3905) of 2 biological replicates normalised by RPKM. Note that 451 introns were dropped from **B** since these genes only contained a single intron.

4.4 Premature Termination of RNA Pol II Generates Small Dis3 RNA Substrates

Promoter-proximal pausing is an important rate limiting step during the early initiation stage of transcription. Transcription arrest at this early stage facilitates capping of the nascent RNA 5' end as well as acting as a QC checkpoint before the commitment of Pol II to progressive transcription elongation (Adelman & Lis 2012; Kwak & Lis 2013). In metazoans, Pol II transcription is typically arrested between ~20-60 nt downstream of the TSS; the mechanism of this pausing has been linked to strict placement of the +1 nucleosome immediately after the TSS (Jimeno-Gonzalez *et al* 2015). Normally, nucleosomes are depleted from promoter regions as a consequence of chromatin remodelling, or the presence of CpG islands (Deaton & Bird 2011; Flynn *et al* 2011), however the +1 nucleosome has a fixed position between the promoter and TSS sequence. Progression of Pol II into processive elongation requires transcription of the 146 nt of DNA wound ~1.7 turns around the eight histone proteins that comprise the +1 nucleosome (Annunziato 2008), which acts as a barrier to elongation.

Increasing the resolution of the transcription profiles presented in **Figure 4.4** to show single nucleotide coverage around the TSS, detected a short-enriched peak over the first ~150 nt, similar in size as the proposed average first exon length in humans (Bieberstein *et al* 2012), immediately downstream of the TSS common to both treated and untreated DIS3-AID (**Figure 4.10**). While the distance of this peak is consistent with the proposed length of DNA wrapped around the +1 nucleosome, without mapping nucleosome position genome-wide, I can only speculate that this enriched peak coincides with the +1 nucleosome. This short species of truncated nascent RNAs (truncRNA), stabilised in Dis3 downregulated cells has not been previously discovered in RNAi knockdown studies, and represents a potential new Dis3 substrate. Given the proposed association between promoter-proximal pausing and 5' mRNA capping, I cannot rule out that some truncRNAs may arise from failed capping.

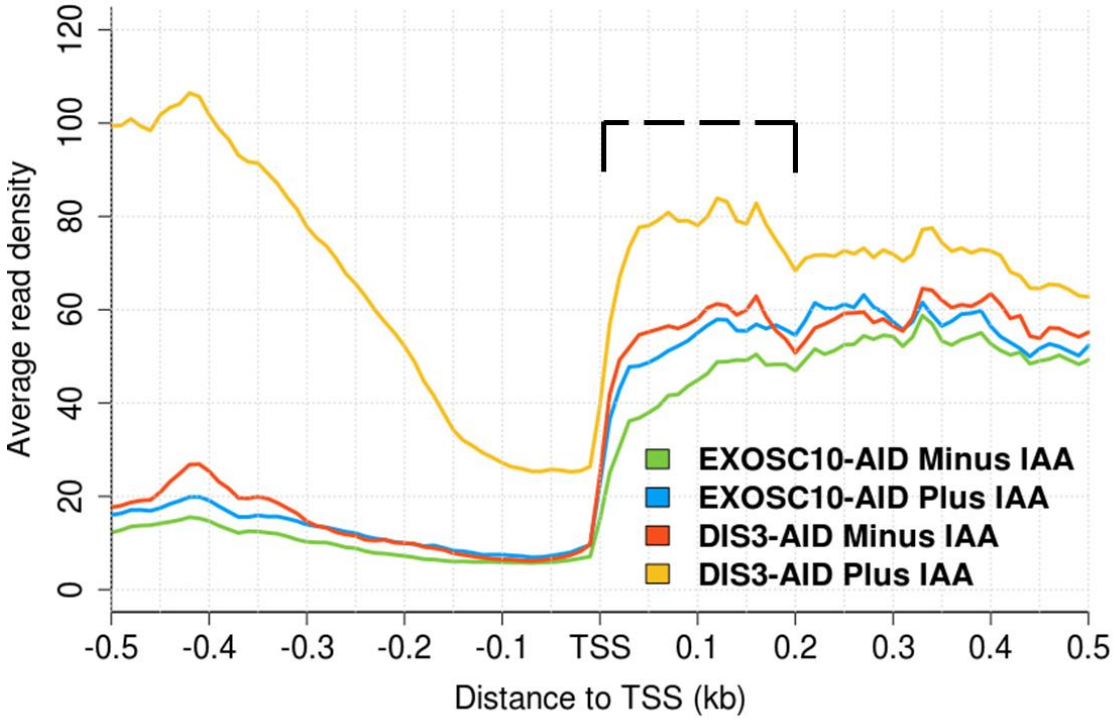


Figure 4.10: Single nucleotide resolution metagene coverage plot analysis centred on the TSS of 4701 genes in EXOSC10-AID and DIS3-AID cell lines. An additional biological replicate is presented in **Supplementary Figure S6**.

4.5 Dis3 Downregulation Stabilises Transcripts Originating from Intergenic Sequences

In addition to PROMPT RNA stabilisation, Dis3 dysfunctional has also been attributed with the accumulation of transcripts originating from unannotated intergenic regions of the genome (Szczepinska *et al* 2015). Transcripts produced from distal loci can arise from spurious transcription of open chromatin by Pol II, or from certain enhancer sequences that, when brought close to the active TSS of an expressed gene, undergoes opportunistic transcription (into so-called eRNAs) by nearby Pol II complexes (Kim *et al* 2010; Kim *et al* 2015). Under normal conditions eRNAs are generally short transcripts (< 2 kb) that arise from bidirectional promoters and exist at very low levels within the cell due to their rapid degradation by the exosome (Andersson *et al* 2014). During the *de novo* transcriptome assembly, I was also able to detect a high degree of potential eRNA transcripts derived from uncharacterised intergenic DNA sequences, which warranted further investigation.

4.5.1 Detection of Unannotated Intergenic Transcripts

As mentioned earlier during the list of *de novo* assembled transcript output (generated by the StringTie assembler) was sub-divided into PROMPT RNAs based on their proximity to known annotated genes (< 3 kb). Using the remaining transcripts I was able to identify 960 novel transcripts aligned to distal intergenic regions of the genome > 3 kb from the nearest gene. Interestingly, visualising these novel transcripts demonstrated that each *de novo* transcript created by the software instead consists of two distinct transcripts that appear to be transcribed in opposing orientations and strand from a single bidirectional promoter-like region, indicated by the dip in coverage between enriched peaks from the approximate start site of each transcript (**Figure 4.11**[**A**]). While the length of this coverage dip differs in length within each transcript, there is a clear separation of sense and antisense oriented transcripts consistent with the notion that the -1

and +1 nucleosomes are present demarcating a promoter boundary (Andersson *et al* 2014). Therefore, the transcripts detected by *de novo* transcript assembly can be viewed as transcription intervals similar to enhancer sequences where bidirectional transcription occurs.

The average expression of these novel transcription intervals increases by ~10-fold in response to Dis3 depletion, indicating that under normal conditions, the exosome quickly removes these pervasive transcripts from the transcriptome (**Figure 4.11**[**B**]). Moreover, Exosc10 does not appear to be involved in degradation of this species of RNA signifying the possibility that these transcripts are either relatively unstructured or are unwound and targeted to Dis3 independent of threading through the central channel of EXO-9, perhaps by Mtr4 or the NEXT complex.

To determine the global extent of intergenic transcription events caused by reduced Dis3 expression, read coverage calculated over known genes was compared against intergenic regions of the genome. Transcription of RNA derived from intergenic loci is increased by ~1.7 fold in the absence of Dis3, suggesting that open intergenic chromatin is an abundant source of pervasive transcripts that quickly accumulates when Dis3 activity is impaired (**Figure 4.11**[**C**]).

4.5.2 Characterisation of Potential Novel eRNA Transcripts

Because of the similarities observed between the novel transcription intervals assembled and enhancers, I decided to further characterise these transcripts in order to determine whether they originate from enhancer DNA sequences or are instead the result of spurious transcription from open chromatin loci. For this reason, the novel transcripts were first compared against a database of known, expressed human enhancer sequences, generated and curated by the FANTOM Consortium (Andersson *et al* 2014; The FANTOM Consortium *et al* 2014) to determine if any of these novel transcripts coincide with annotated eRNAs.

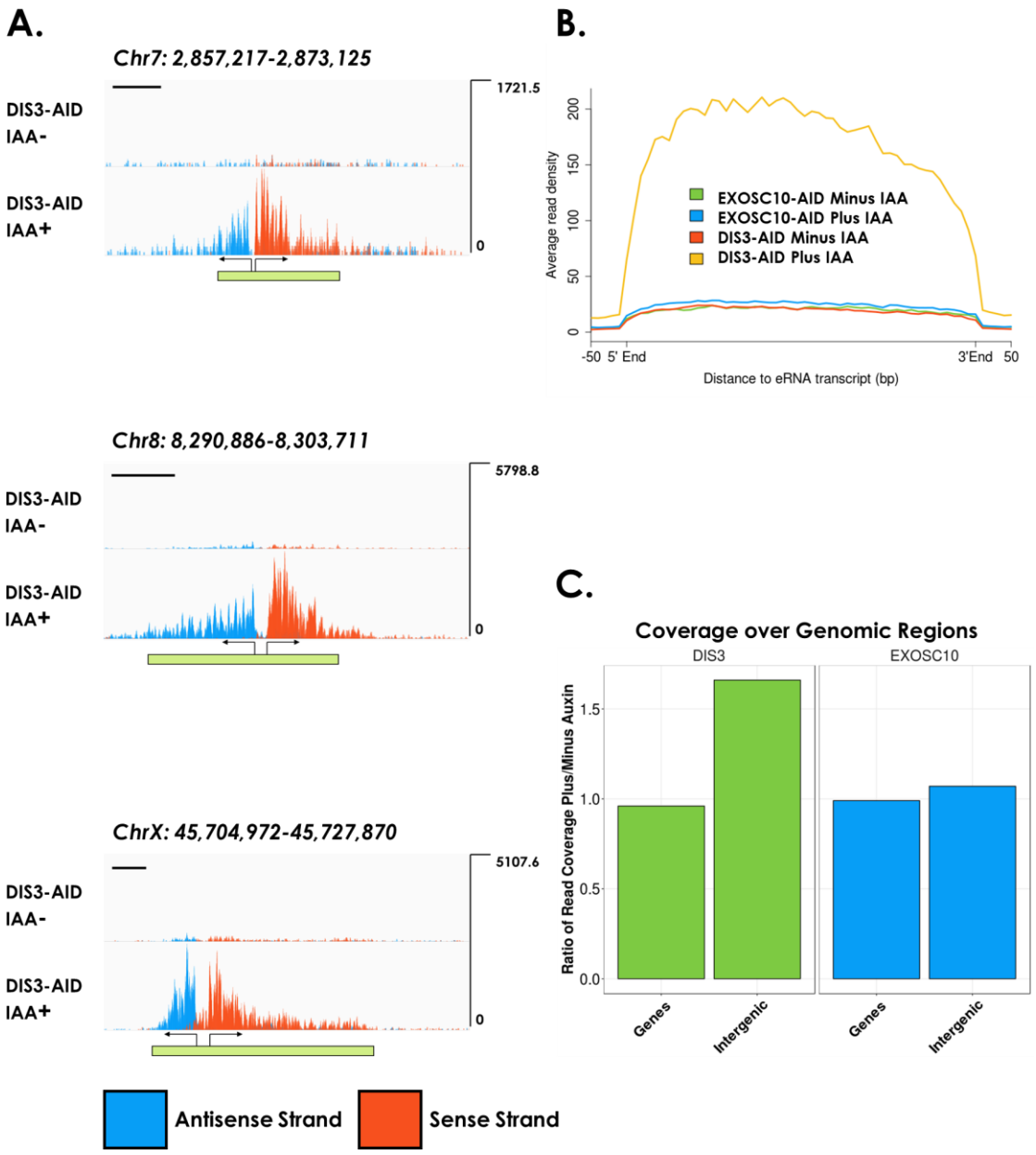


Figure 4.11: (**A**) RPKM normalised coverage tracks of *de novo* transcripts detected over intronic intervals in 2 biological replicates. Scale bars = 2 kb, and apply to both coverage tracks within each gene image. (**B**) Metagene expression plot of the same transcripts (n = 960) from a single biological replicate (replicate 2 shown in **Supplementary Figure S7**). (**C**) Ratio of sequencing coverage between treated/untreated cells over genes and intergenic regions within the whole genome normalised to library size.

The expression of specific enhancers is essential in the determination of cell-type specificity with only a subset of enhancers being active within any given cell-type. However, for the HCT116 cell line used in this study, no dataset specific for colon tissue cells was available within the FANTOM5 collection. Instead I opted to use the full human annotation dataset, currently composed of 32,693 eRNA transcripts as a reference.

The genomic positions of known FANTOM5 eRNA transcripts were compared against the positions of *de novo* intervals to compare any overlap in genomic location, however no overlap was detected for any of the 960 *de novo* generated transcripts. One possible reason for this may be attributed to the parameters of the StringTie algorithm during transcript assembly since it relies heavily on the sequencing coverage obtained from raw aligned reads. Thus, a high degree of variability can be observed between different cell-types and conditions, hence why these *de novo* transcripts only represent a close approximation to potential transcripts. Furthermore, the majority of eRNAs were previously identified by FANTOM5 in cells without prior exosome depletion and, due to rapid Dis3 turnover, would have potentially escaped detection. An alternative approach was instead devised to determine if any of the FANTOM5 eRNAs reside within close proximity (< 5 kb) to the list of *de novo* transcripts. Interestingly, known annotated eRNA transcripts were detected in close proximity to ~14% of the *de novo* transcripts (**Figure 4.12**). In some cases, several eRNAs were shown to be clustered both upstream and downstream of these novel transcription intervals, implying that a small proportion of the assembled transcripts potentially represent enhancer-like sequences.

Despite failing to detect FANTOM5 eRNA transcripts within the vicinity of the majority of assembled novel intergenic transcripts, it remains plausible that the rapid and direct depletion of Dis3, mediated by the AID degron system, has uncovered a group of uncharacterised divergent RNA transcripts derived from cryptic intergenic promoters. In order to determine if these transcripts are undiscovered eRNAs further validation must be performed.

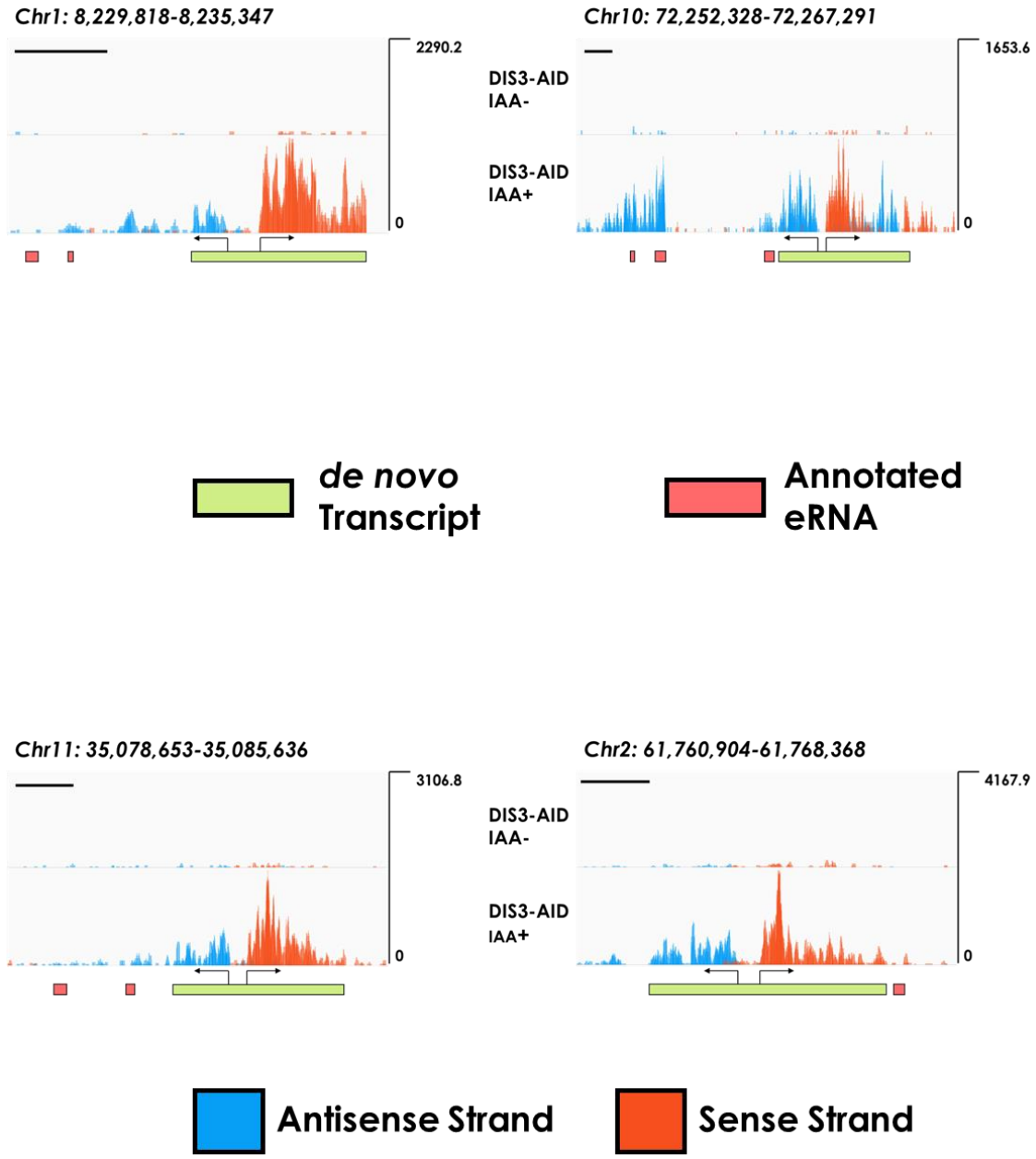


Figure 4.12: Visualisation of FANTOM5 annotated eRNAs (red) located proximal to *de novo* assembled transcripts intervals (green). Two biological replicates normalised by RPKM are shown per coverage track. Scale bars = 1 kb were applied to both coverage track within each gene respectively.

4.5.3 Identification of Enhancer Sequences using Histone Modifications

Typically, most enhancer sequences are discovered based on the identification of a signature chromatin profile, namely high levels of monomethylated histone H3 lysine 4 (H3K4me1) which facilitates binding of the transcriptional co-activator p300/CBC and, low levels of a promoter-specific modification H3K4me3 (Kim *et al* 2010; Kim *et al* 2015). Additionally, co-enrichment of H3K27ac with high levels of H3K4me1 signifies that the enhancer is transcriptionally active instead of being poised (Andersson *et al* 2014). ChIP-Seq data generated by the ENCODE project was next used to determine if the list of *de novo* transcripts are eRNAs by comparing each of the 3 histone modifications mentioned above in conjunction with the Dis3 nascent transcriptome analysis.

For each ChIP-Seq data set, DNA fragments were enriched through antibody mediated pull-down of H3K27ac, H3K4me1 and H3K4me3 chromatin modifications and then subsequently sequenced. After subtraction of background DNA sequencing noise (present in an input sample generated without antibody enrichment), the locations of each epigenetic modification were then determined by calling peaks of sequence reads.

Starting with enhancer-like *de novo* transcripts detected proximal to known FANTOM5 eRNAs, I discovered two distinct peaks of H3K27ac that overlap the approximate TSS of both divergent transcripts (**Figure 4.13**[**A**]). The lack of RNA sequencing depth between both transcripts is consistent with the dip in H3K27 acetylation, highlighting the presence of a shared cryptic bidirectional promoter. I next decided to directly compare both H3K4me1 and H3K4me3 ChIP-Seq libraries by plotting the log2 ratio of called peaks (H3K4me1/H3K4me3). Consistent with previously published data, the list of enhancer-like regions have low levels of H3K4me3 near the promoter and accumulated peaks of H3K4me1 co-

occurring with the stabilised transcripts, characterising them as enhancers domains (Kim *et al* 2010).

In contrast to enhancer DNA regions, protein-coding genes are known to have enrichment of H3K4me3 peaks around promoters and low levels of H3K4me1. To determine if this pattern of histone modifications are consistent within this dataset, I chose 2 highly expressed protein-coding genes: *MARS2* and *SEPHS1* which, like the enhancer-like regions, exhibits strong bidirectional promoter transcription in the absence of Dis3 (**Figure 4.3**). Both protein-coding genes remain consistent with previously published findings, where the equilibrium is shifted to highlight a much greater enrichment of H3K4me3 compared to its monomethylated counterpart (**Figure 4.13**[**B**]). Moreover, the general level of H3K4me1 remains low over the protein-coding promoter region and throughout the gene body. The methylation status of H3K4 therefore, appears to play an important role in differentiating protein-coding genes from enhancer-like transcripts, despite the presence of a bidirectional promoter at both intervals. Additionally, the double H3K27ac peaks were also observed around the promoter region indicating that this modification appears to be a common property of all promoters regardless of transcript biotype regulated.

Lastly, the histone modifications of the *de novo* transcripts lacking nearby known FANTOM5 eRNA transcripts were investigated. Consistent with known enhancer domains, peaks of H3K4me1 were found to be enriched over the list of intergenic transcripts at a much higher level relative to H3K4me3, which was restricted to the promoter region and in some cases, was virtually undetectable (**Figure 4.13**[**C**]). Likewise, the twin peaks of H3K27ac surrounding the proposed promoter region is consistent with the notion that the majority of *de novo* assembled transcripts are likely derived from previously undiscovered active enhancer domains.

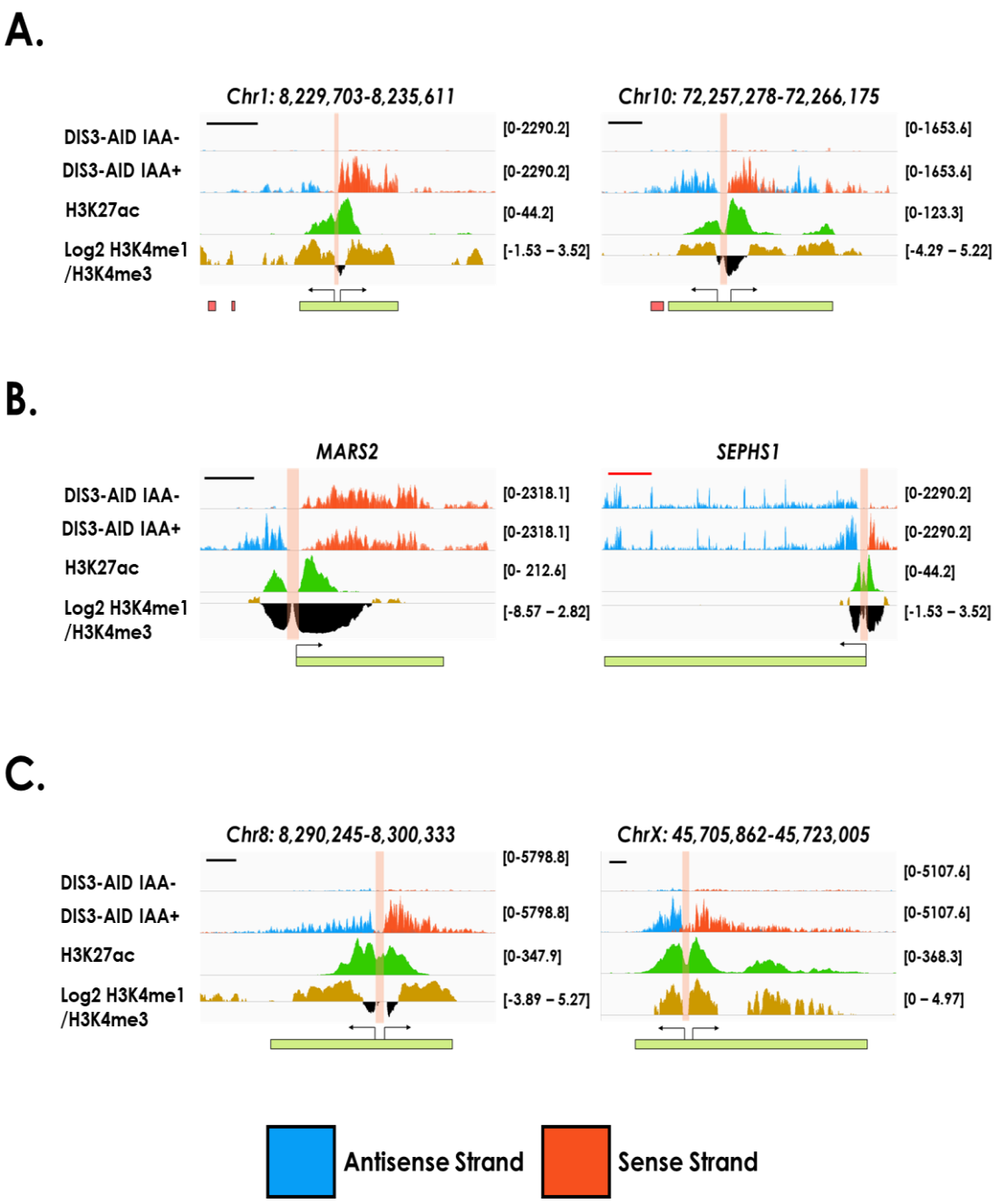


Figure 4.13: Comparison of RPKM normalised RNA-Seq and histone ChIP-Seq coverage tracks over enhancer-like DNA sequences (**A**), protein-coding genes (**B**) and novel intergenic intervals (**C**). RNA-Seq and H3K27ac tracks normalised on a linear scale, log2 scale used for H3K4me1/H3K4me3 ratio. Each track represents the average of 2 biological replicates. Black scale bars = 1 kb, applicable to every coverage track within each gene image respectively. For *SEPHS1*, the red scale bar = 5 kb are applicable to all 4 coverage tracks shown.

4.5.4 Novel Intergenic Transcripts have eRNA-like Properties

To conclude this investigation into determining the identity the *de novo* assembled transcripts, I next used the average transcript orientation in each interval to measure the overall promoter directionality present within this dataset since eRNA domains are transcribed bidirectionally.

In each of the assembled intergenic intervals, mapped reads aligned to sense and antisense transcripts were counted and the average promoter direction (calculated as ratio) was determined in both untreated and Dis3 depleted conditions, since low level expression of nascent RNA over was also detected in cells expressing Dis3. On average, ~30% of promoters within the enhancer-like intervals transcribe RNA bidirectionally (**Figure 4.14**[**A**]). Interestingly, the majority of intergenic promoters predominantly transcribe in the sense orientation and as a consequence of Dis3 depletion, sense direction transcription increases by ~10%. Furthermore, while antisense transcription from these promoters is generally lower, an almost equivalent reduction in antisense transcription (~8%) is observed following Dis3 downregulation. It is unclear why knockdown of Dis3 would alter promoter directionality since transcripts generated in both orientations should be equally sensitive to exosome-mediated degradation. However, it appears that transcripts in one direction are be more susceptible to Dis3-mediated degradation. This is similar to what has been described in protein-coding genes, where termination of antisense PROMPTs at TSS-proximal poly(A) sites enhance transcription in the sense direction (Ntini *et al* 2013). Defining the 3' ends of each eRNA-like transcript originating from both directions using mammalian native elongating transcript sequencing (mNET-Seq), a technique that can reveal the genome-wide position of Pol II at single-nucleotide resolution via immunoprecipitation (IP) and sequencing of RNA released from its active site (Nojima *et al*. 2015), could provide valuable insight into the directionality, termination mechanism and susceptibility to Dis3-mediated decay for each eRNA-like transcript annotated.

Lastly, I calculated the average length of the enhancer-like transcripts since unlike lincRNAs, eRNAs tend to be much shorter (< 2kb) in length (Kim *et al* 2004). On average, the enhancer-like transcripts are ~500 bp in length and generally do not exceed 1 kb, which are slightly shorter than previous predictions (Andersson *et al* 2014). By comparison, the assembled PROMPT RNAs have an average length of ~1 kb (**Figure 4.14**[**B**]), which is slightly longer that previously described (Preker *et al* 2011). A likely reason for the discrepancy in RNA length between this study and previously published data can be explained by the *de novo* transcriptome assembly process. Transcript assembly relies on the presence of nearby or overlapping mapped reads to generate transcripts, therefore low background level aligned reads can contribute to *de novo* transcript assembly. This would also explain why some PROMPTs reach an apparent length of ~3 kb. To circumvent this in the future, sequencing an RNA population enriched for capped RNA (via immunoprecipitation of capped RNAs using an antibody recognising the 7-methylguanosine cap) in Dis3 depleted cells would accurately determine the lengths of each PROMPT and eRNAs detected from this study.

Collectively, I present strong evidence that, through a combination of chromatin landscape mapping, identification of proximal FANTOM5 eRNAs and determination of promoter directionality that, the majority of intergenic *de novo* transcripts are uncharacterised eRNAs or eRNA-like transcripts. However, the requirements of the StringTie algorithm during *de novo* assembly, namely the need of sufficient sequencing depth, is a potential source of error to which many discrepancies can be attributed.

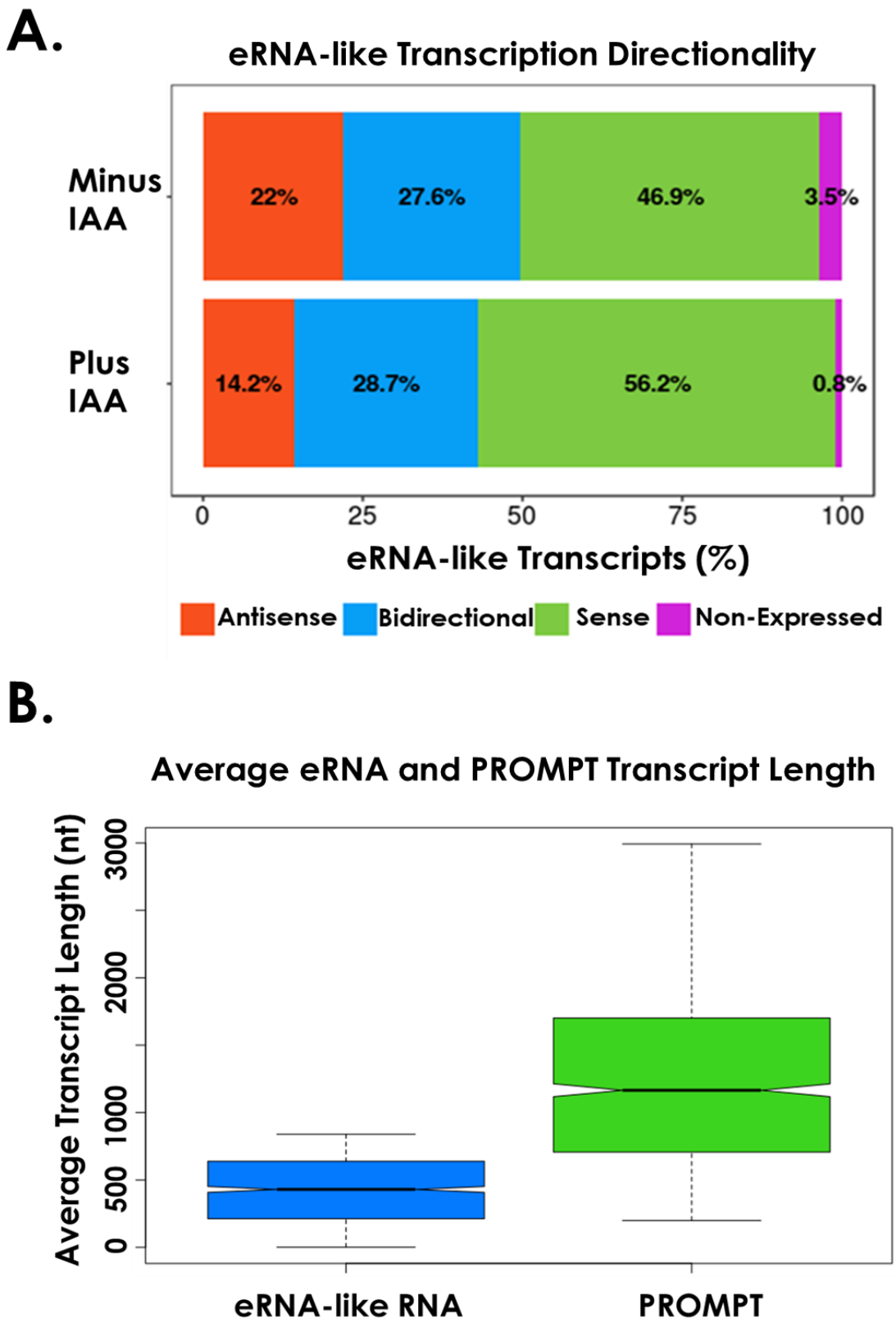


Figure 4.14: (**A**) Calculation of enhancer-like RNA transcription directionality (n = 960). (**B**) Comparison of the average *de novo* enhancer-like RNA and PROMPT transcript lengths (n = 960 and 1092 respectively).

4.6 Summary

In this chapter, I have shown that Dis3 is the most processive exoribonuclease subunit of the exosome, responsible for the degradation of a broad range of nascent RNA transcripts in the nucleoplasm. In as little as 60 minutes of Dis3 depletion using the AID degron system, I detected an increased level of transcription originating from both promoter proximal sequences of protein-coding genes and unannotated intergenic regions of the genome.

The most prominent species of RNA stabilised as a consequence of Dis3 dysfunction are divergently transcribed PROMPTs originating from bidirectional promoters (**Figure 4.3**). Consistent with previously findings, PROMPT RNAs detected in these data are transcribed from protein-coding promoters in reverse orientation on the opposing strand (Preker *et al* 2008; Flynn *et al* 2011; Szczepinska *et al* 2015). Interestingly, PROMPT transcription was shown to gradually decrease upstream of the TSS of protein-coding genes, indicating that the window of PROMPT transcription from nearby promoters is finite and eventually transcription terminates within ~3 kb upstream (**Figure 4.4**). Despite the asymmetrical distribution of poly(A) sites which are less frequent over PROMPT sequences (Ntini *et al* 2013), termination of PROMPTs could still be achieved through conventional cleavage at poly(A) sites providing a free 3' end for rapid Dis3-mediated degradation. Moreover, I detected that the 1092 *de novo* assembled PROMPT transcripts have an average, length of ~1 kb, significantly longer than previously characterised (Preker *et al* 2011). Given the evidence that all eukaryotic promoters are inherently bidirectional, and taking into account tissue specific gene expression, the list of nascent assembled PROMPT transcripts characterised in this study are likely to be a significant underestimation of the abundance of unstable divergent promoter RNAs that exist within humans.

The majority of upregulated genes detected through differential gene expression were false positives caused by cryptic intronic

transcription, spurious transcription from nearby open chromatin and stabilisation of PROMPTs that overlap nearby genes. PROMPT transcript overlap contributed to a large proportion of upregulated genes detected, since removal of closely spaced genes (< 3 kb from the TSS and TES) discarded ~92% of the false positive hits (**Figure 4.7**). PROMPT RNA read-through therefore has the potential to interfere with the expression of nearby genes by transcriptional interference. I was unable to show this within this dataset and from the example presented in **Figure 4.6**, since the overlapping genes detected were not expressed. However, I hypothesise that prolonged Dis3 downregulation would have a greater potential to interfere with gene expression globally, either through depletion of transcription factors from the pool of active Pol II complexes or from disruption of promoter definition. Therefore, Dis3 appears to play an important role in maintaining strict promoter directionality by rapidly clearing unwanted by-products of divergent promoters, which in turn enhances transcription of RNA downstream of the TSS, ultimately preserving correct punctuation at the 5' end of the gene.

During this investigation into the stabilisation of promoter associated transcripts, I was able to detect a significantly augmented level of coverage localised to the 5' end of a subset of genes in the coding direction. Upon closer inspection I discovered that a considerable proportion of genes exhibit an accumulation of mapped reads aligned to the first intron of the transcript following Dis3 knockdown. I observed that as the calculated coverage progressed towards the 3' end of the gene, the sequencing coverage gradually declined eventually reaching a level complimentary to the untreated cell line (**Figure 4.8**). Consistently, I did not detect any significant accumulation of reads over the terminal intron within these genes (**Figure 4.9**). Moreover, in some cases I also observed a smaller increase of reads over introns downstream of intron 1, particularly in shorter genes, indicating that this effect is not solely localised to intron 1 but the 5' end of the gene. Given the gradual decline of RNA-Seq reads, I propose that RNA transcribed from these genes undergo premature

termination of transcription prior to reaching the TES and are subsequently degraded by Dis3. Since under normal conditions Dis3 would be present, and can rapidly act on the newly formed 3' end of these transcripts, this data implies that Dis3 potentially degrades these abortive transcripts co-transcriptionally through close association with the elongating Pol II complex. The susceptibility of these prematurely terminated transcripts to Dis3-mediated decay coupled with their proximity to the TSS is similar to antisense PROMPT RNA expression. Therefore, the open chromatin structure found around the promoter region of several genes is likely to be a major source of spurious transcription initiation that requires post-transcriptional gene downregulation through RNA degradation by Dis3.

Similar to abortive transcription, I also detected a peak of RNA stabilisation within ~150 nt of the TSS consistent with the length of DNA wrapped around the +1 nucleosome (**Figure 4.10**). This short peak of coverage was shown to be modestly stabilised as a consequence of Dis3 depletion, but was also mildly detectable in untreated cells. Additionally, the length of these short stabilised truncRNAs approximately coincides with the site of promoter-proximal pausing, a common processing step where 5' capping takes place. I therefore surmise that short truncRNAs are the result of abortive transcripts that fail to be released from this early transcription QC checkpoint. Furthermore, the degree of truncRNAs accumulation observed in this analysis is obscured by the stability of the first exon, since stabilisation of the first 150 nt coincides with the average (128-350 nt) first exon length (Bieberstein *et al* 2012). As exons are generally more stable within mRNA transcripts compared to introns or PROMPTs, they will remain relatively stable over the brief period of Dis3 depletion, reflecting much smaller effects. This would also explain why the previous calculation of sequencing depth of mapped exonic RNA reads did not detect any significant change in stability (**Figure 4.8**).

Finally, I show that a large quantity of Dis3 substrates are derived from cryptic promoters originating from unannotated intergenic regions of the genome (**Figure 4.11**). Through a combination of *de novo* transcript

assembly, identification of the chromatin landscape, determination of promoter directionality and proximity analysis of nearby known eRNA transcripts, I ascertained that the bulk of these transcripts are likely to be derived from enhancer-like gene intervals, many of which have not been previously discovered. While I can speculate from initial characterisations that these are enhancer-like domains, further validation similar to the FANTOM curated datasets would need to be performed to confirm that these transcripts are *bona fide* eRNAs.

Collectively, this analysis highlights the impact of Dis3 activity particularly during the early stages of transcription in addition to its role in preventing the accumulation of spurious transcripts derived from dysfunction of promoter directionality and unmasking of cryptic or intergenic promoter sequences. In the next chapter I will investigate the role of the 5'→3' exoribonuclease Xrn2 as part of the nuclear surveillance pathway.

Chapter 5

XRN2 Enhances Transcription Termination at Gene 3' Ends

So far I have only considered the involvement of the exosome complex as part of nuclear RNA surveillance, however another key exoribonuclease, Xrn2, is also present within the nucleus, and unlike the exosome, Xrn2 degrades uncapped RNA with 5'→3' directionality. Xrn2 is recruited during the early stages of transcription and can "travel" alongside actively engaged Pol II complexes, facilitating rapid co-transcriptional degradation of nascent RNA by-products of failed transcription or RNA processing (Brannan *et al* 2012; Davidson *et al* 2012). More importantly, Xrn2 has been shown to play a significant role in orchestrating the termination of transcribing Pol II complexes downstream of the cleavage and poly(A) site (CPA) by the torpedo mechanism (Kim *et al* 2004; West *et al* 2004). Following cleavage at the poly(A) site catalysed by the endonuclease CPSF73, Xrn2 then rapidly degrades the 3' flanking RNA, effectively chasing down the elongating Pol II complex. Transcription is then terminated by an undefined mechanism that leads to the dissociation of the transcription complex from the template DNA (Proudfoot 2011).

Turning to the AID degron system again, I decided to continue the characterisation of nuclear surveillance pathways by investigating the immediate impact of transcription termination in the wake of rapid Xrn2 downregulation, and to determine if the RNA composition of the transcriptome is altered as a result.

5.1 Xrn2 Degrades 3' Flanking RNA downstream of the TES

Understanding the function of Xrn2 during transcription termination has become a controversial topic following the recent publication of several contradictory results. This most notable includes the findings provided by Nojima *et al* (2015), who using mNET-Seq to map the 3' ends of nascent RNA transcripts, did not detect any transcriptional read-though, or termination defects downstream of the TES site following RNAi-mediated Xrn2 knockdown. However, shortly after these findings were published, the role of Xrn2 during Pol II termination was re-examined by Fong *et al* (2015), this time combining RNAi knockdown with the co-expression of a catalytically inactive Xrn2 mutant. In this scenario using ChIP-Seq to determine the position of Pol II across the genome, Fong *et al* detected a significant accumulation of Pol II occupancy up to 5 kb downstream of the TES, indicating that transcription termination is significantly delayed as a consequence of Xrn2 loss. Although transcription termination still occurs much farther downstream, these findings support the notion that Xrn2 plays a crucial role in the efficient termination of Pol II transcription in eukaryotes (Kim *et al* 2004; West *et al* 2004).

Collectively, both studies highlight the limitations of RNAi as a method to study gene function in metazoans, since despite near complete depletion of Xrn2 protein after 60-72 hours (Nojima *et al* 2015), residual Xrn2 can still fulfil its intended function within the cell, unless a dominant-negative mutant is co-expressed. However, since the AID-degron system directly degrades the target protein of interest, many of these limitations can be circumvented, providing a much more reliable and robust analysis of gene function.

To investigate the function of Xrn2 during transcription termination in an XRN2-AID degron cell line, produced by Professor Steve West. Similar to both EXOSC10-AID and DIS3-AID cells, Xrn2 protein is depleted within 60 minutes of IAA introduction to the growth media (**Appendix Figure 1**).

Nascent nuclear RNA was therefore sequenced after 60 minutes of Xrn2 depletion. Again, basic statistical analysis was performed on both replicate XRN2-AID RNA-Seq libraries to measure the percent coverage of the genome, average sequencing depth and read mapping efficiency (**Figure 5.1**). Each library covered ~40% of the genome consistent with the notion that these libraries represent the nascent transcriptome and have and average depth of ~1-2 per base. Consistent with previous libraries, >80% of sequence reads were uniquely mapped to the GRCh38 genome. Defects in transcription termination at the 3' end of the genes were next determined through detection of read-though RNA stabilised downstream of the TES.

Consistent with previously published RNAi based approaches in combination with Xrn2 mutant overexpression (Fong *et al* 2015), I discovered that mapped reads aligned to intergenic regions flanking the 3' end of the gene are significantly stabilised as a consequence of Xrn2 depletion (**Figure 5.2**). In addition to the detection of stabilised RNA proximal to the TES, increased RNA expression was also shown to occur from intergenic sequences much farther downstream of the gene 3' end in agreement with the observation that Xrn2 downregulation impairs termination of the transcribing Pol II complex. Moreover, the distance that Pol II complexes escaping termination are able to traverse is highly variable with some genes showing only a modest read-through distance of several kilobases, whereas other genes such as *TBL1XR1* continue to transcribe up to 100 kb into the 3' flanking intergenic sequence. Interestingly, I detected transcription read-though in both protein-coding and non-coding genes such as miRNA, indicating that torpedo is not restricted to a single gene biotype. As a comparison, Exosc10 downregulation did not cause any observable defects in transcription termination, indicating that the 3' flanking RNA sequence is uncapped and degraded co-transcriptionally from the 5'→3', most likely due to actively engaged Pol II complexes shielding the 3' end of the growing nascent 3' flanking RNA from exosome-mediated degradation.

To study the extent of Pol II read-through in the 3' flanking region globally, I decided to generate a metagene plot based on the expression of RNA across the gene body and into the flanking intergenic sequences. To remain consistent with the earlier exosome analysis during my characterisation of Dis3, I used the same list of expressed non-overlapping genes that included an inclusion window of 3 kb upstream of the TSS and 7 kb downstream of the TES (n = 4701). In stark contrast to Dis3 depletion (**Figure 4.4**), the effects of Xrn2 knockdown were largely restricted to the 3' flanking region downstream of the TES, where RNA was shown to be stabilised by ~2-fold relative to untreated control cells (**Figure 5.3**[**A**]). RNA expression from 3' flanking regions gradually decline as the distance increased and continued to fall to background expression levels observed in the untreated cells. This indicates that some alternative termination process still occurs in the absence of Xrn2 activity which facilitates release of engaged Pol II complexes from the template DNA. In addition to the accumulation of 3' flanking RNA, the metagene analysis was also able to detect a modest, but noticeable reduction in RNA expression within the gene body, which can be explained by a reduction in initiation-ready Pol II as a result of a profound defects in transcriptional termination.

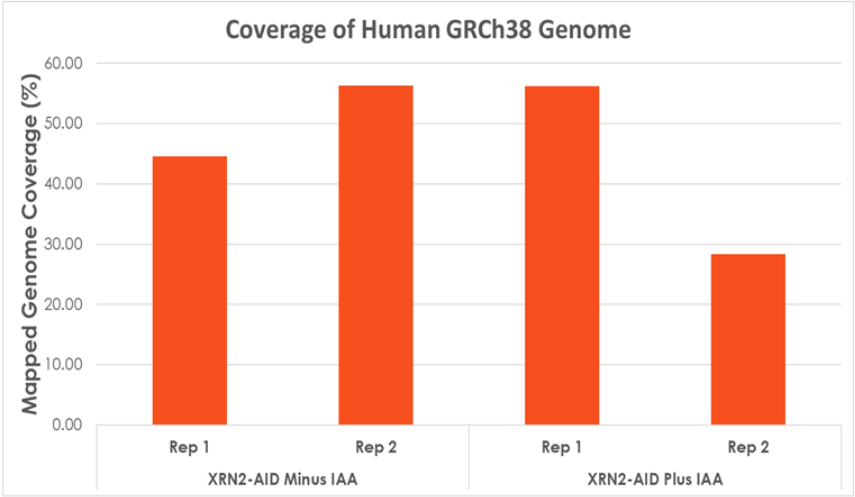
To confirm that the increase in 3' flanking RNA occurs as a consequence of transcription read-though by Pol II and not from failure of cleavage at the poly(A) site, I enhanced the metagene expression analysis around the TES (**Figure 5.3**[**B**]). I observed a dip in mapped read enrichment as a result of poly(A) site cleavage over the TES. Since the expression level of RNA over this cleavage site is similar in all samples I can therefore surmise that cleavage at the poly(A) site is unaffected by Xrn2 depletion. This was also verified by the analysis of individual transcripts (Eaton & Davidson *et al* 2018).

Lastly, calculation of sequencing depth over genomic elements detected a ~1.5-fold increase in read coverage over intergenic DNA sequences comparable in abundance to Dis3 knockdown, but not within

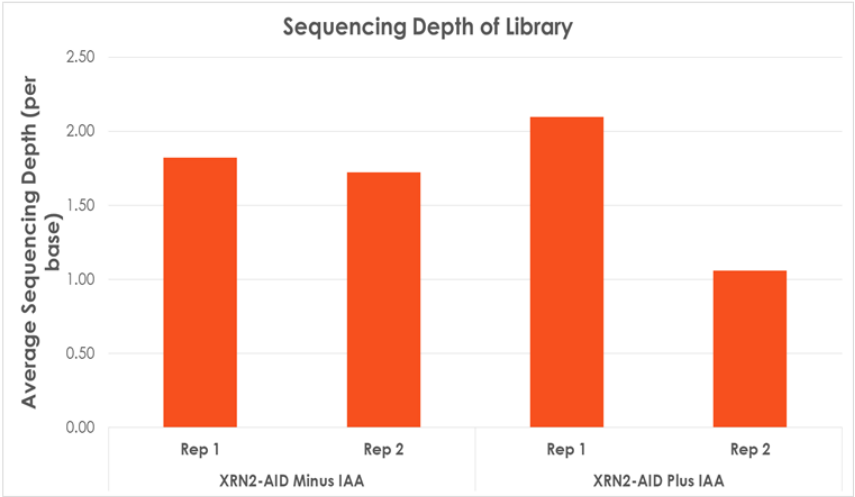
annotated gene intervals (**Figure 4.11**[**C**]; **Figure 5.3**[**C**]). This likely reflects increased 3' flanking RNA downstream of the TES site caused by Xrn2 loss.

In summary, transcription termination by torpedo consequently occurs within a predefined "termination window" dictated by the distance required for Xrn2 to successfully catch up to and collide with the elongating Pol II complex. Reduced expression of Xrn2 causes Pol II escape extending the window by several kilobases before eventually terminating by an alternative termination mechanism. The size of this window varies between genes as a result of parameters that are not yet apparent.

A.



B.



C.

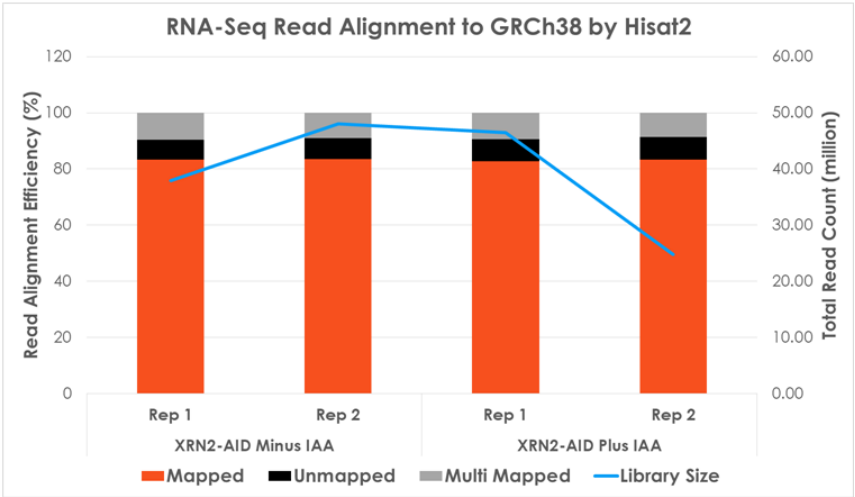
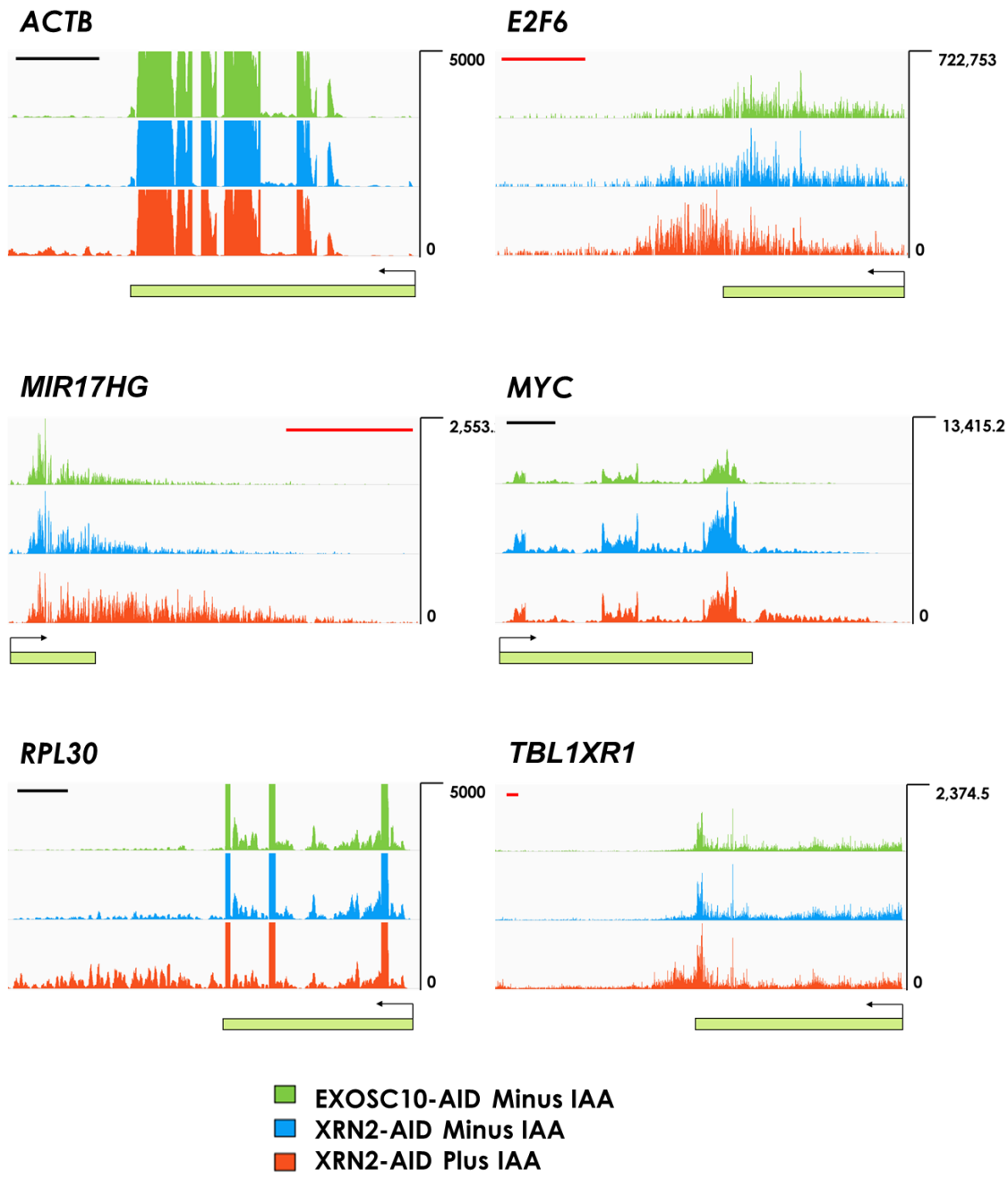


Figure 5.1: Graphical representation of (**A**) genome coverage, (**B**) per base sequencing depth and (**C**) HISAT2 mapping efficiency for each replicate (Rep) of the XRN2-AID RNA-Seq libraries used in this study.



Eaton & Davidson *et al* 2018

Figure 5.2: Single replicate analysis of RPKM normalised coverage tracks in Xrn2 depleted cells showing the extent of read-though beyond the TES. Black scale bars = 1 kb, apply to every coverage track within *ACTB*, *MYC* and *RPL30* gene images respectively. Likewise, red scale bars = 10 kb were applied to each track in *E2F6*, *MIR17HG* and *TBLXR1* genes. Additional biological replicate found in **Supplemental Figure S8**.

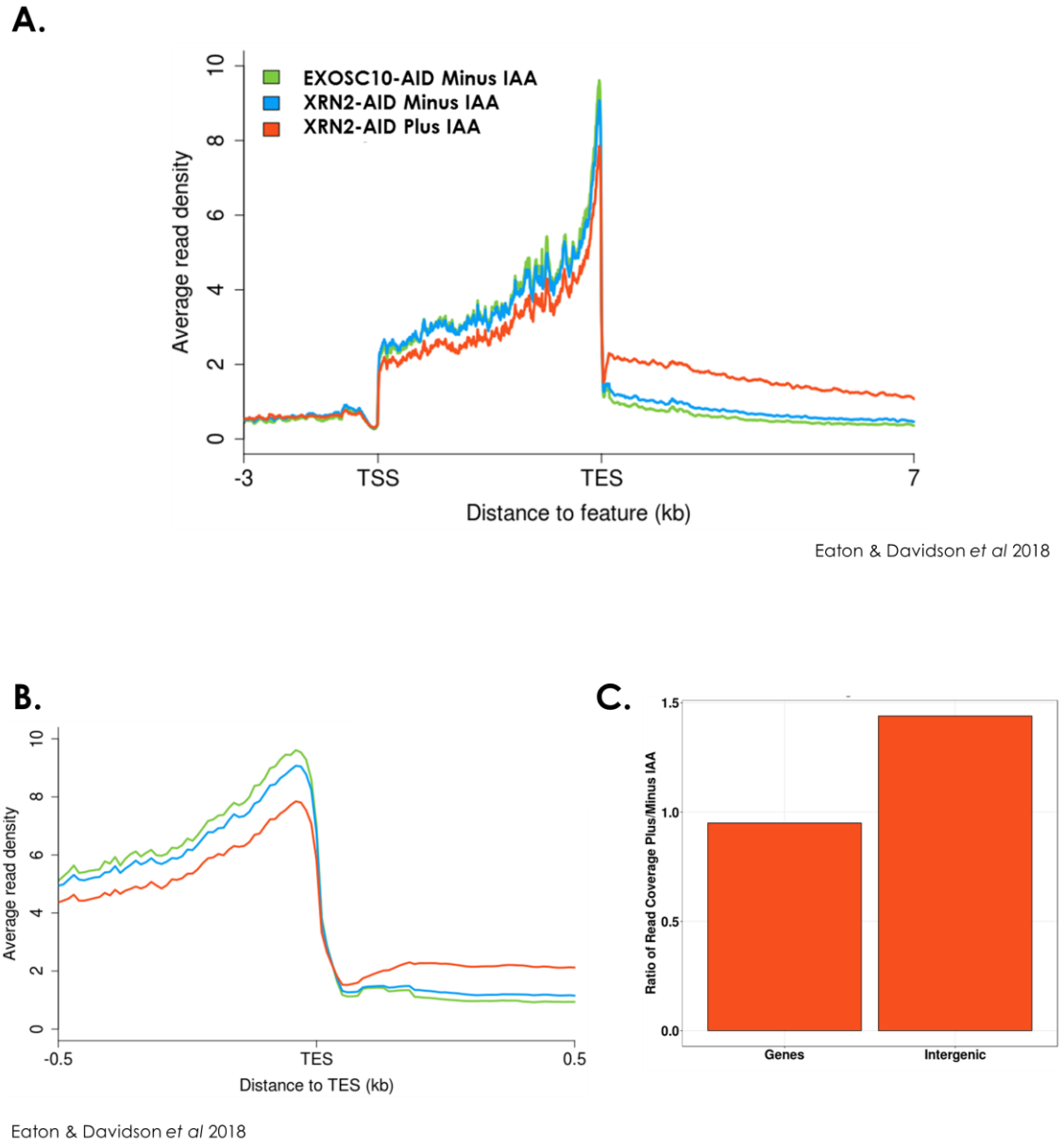


Figure 5.3: (**A**) Metagene coverage plot of normalised mapped reads of 4701 non-overlapping genes. The gene body was scaled to 5 kb and included a region 3 kb upstream of the TSS and 7 kb downstream of the TES respectively. (**B**) Increased resolution of the TES site from (**A**). (**C**) Calculated mapped RNA-Seq read coverage over gene containing and intergenic regions of the genome, normalised to library size. Additional replicate found in **Supplemental Figure S9**.

5.2 Xrn2 is not responsible for Transcription Termination of Histone and snRNA Genes

The majority of 3' flanking RNA stabilised in the initial analysis derived from intergenic downstream regions of protein-coding and lncRNA genes, each of which possess or are likely to possess a poly(A) site at the 3' end of the gene. I next decided to investigate if termination by torpedo is a characteristic of genes that are processed by cleavage and polyadenylation.

5.2.1 Stabilised 3' Flanking RNAs are Absent Downstream of Histone Genes

Despite coding protein, the termination of histone genes relies on the presence of alternative *cis* acting elements namely, a stem-loop structure and HDE sequence instead of consensus poly(A) motifs. In common with poly(A) site-containing mRNAs, nascent histone mRNAs are still cleaved by CPSF73 providing an entry site for Xrn2 mediated degradation of the 3' flanking sequence by torpedo (Dominski *et al* 2005; Dominski *et al* 2007).

Figure 5.4 displays the normalised coverage tracks of several histone genes located within clustered genomic loci. These histone genes were selected due to their high level of expression within HCT116 and close proximity to each other within the genomic locus, providing a means to visualise termination efficiency over several histone genes simultaneously. In each histone gene analysed, I was unable to detect any significant accumulation of RNA downstream of the TES indicating that Xrn2 is not required for the degradation of downstream 3' flanking RNA, despite the potential presence of a free 5'-P on the nascent transcript. Due to their small size and genomic position within clustered histone gene loci, metagene transcription profiles could not be generated, therefore each histone gene was investigated individually.

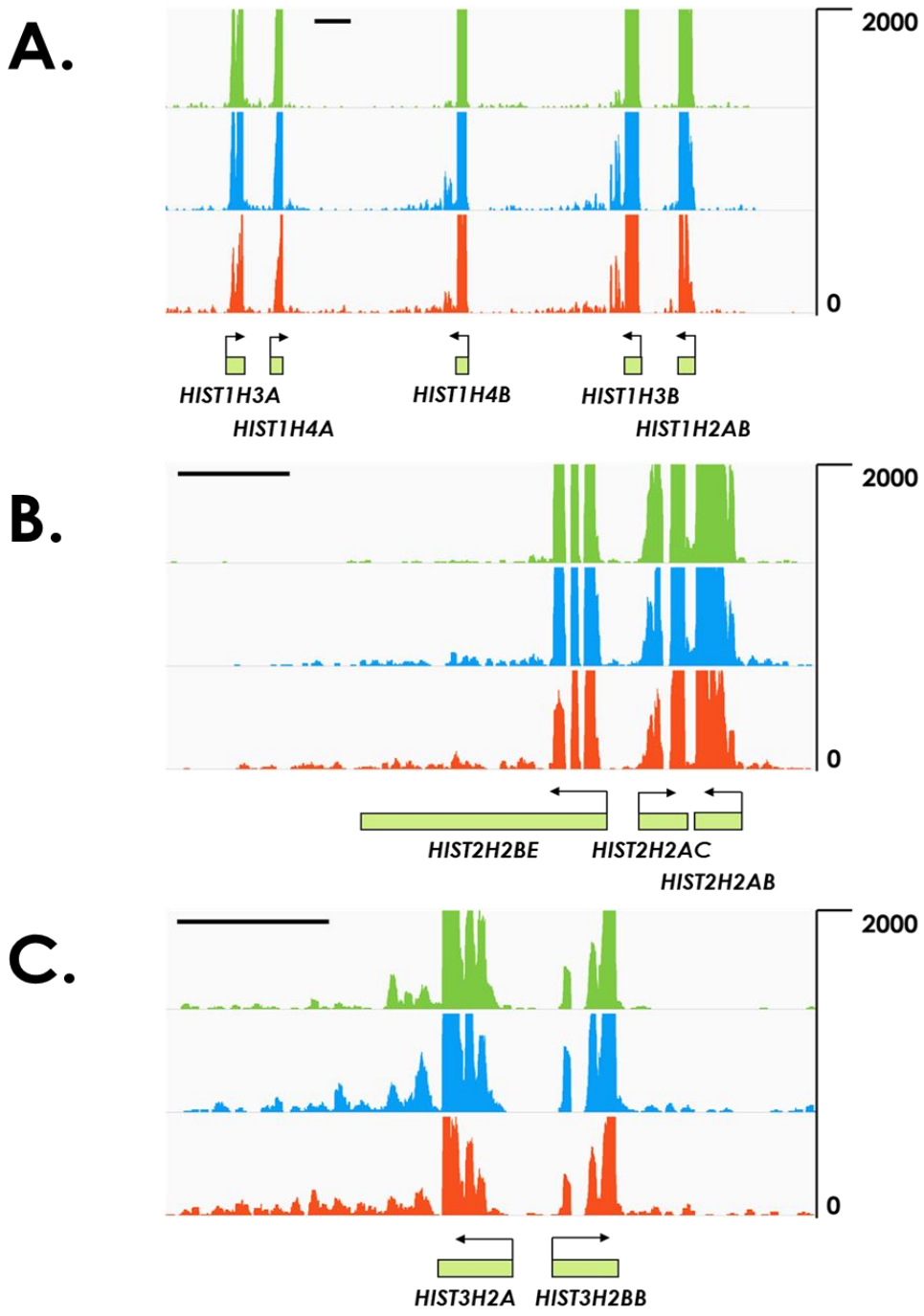


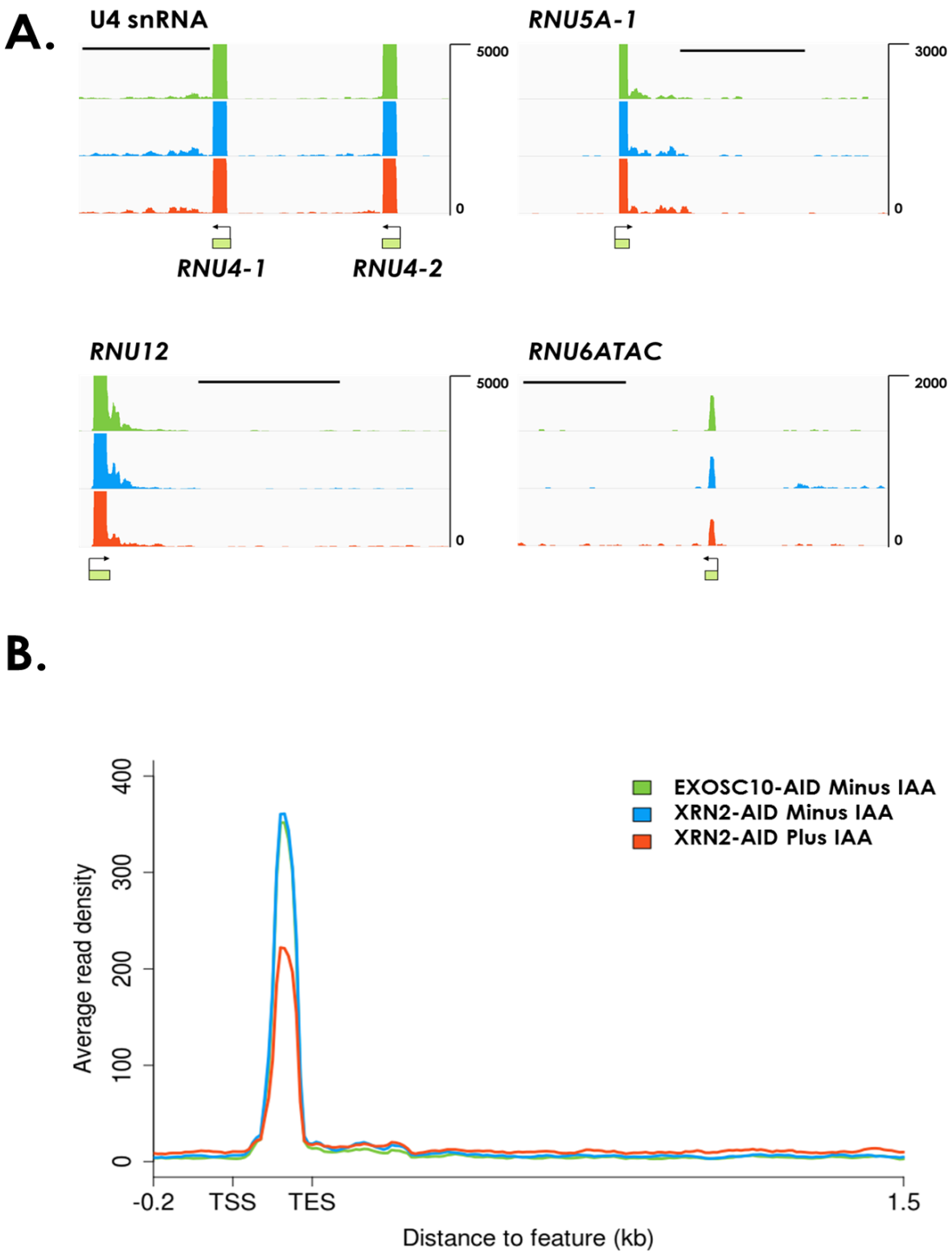
Figure 5.4: Normalised sequencing coverage tracks of histone genes found within histone clusters 1 (**A**), 2 (**B**) and 3 (**C**). Scale bars = 1 kb apply to all 3 coverage tracks within each gene image respectively. Additional biological replicate found in **Supplemental Figure S10**.

5.2.2 snRNA Genes are Efficiently Terminated Independently of Xrn2 Depletion

Nascent snRNAs also undergo 3' cleavage, but utilise the endonuclease, Int11, a member of the Integrator complex (Albrecht & Wagner 2012). The function of Ints11 is therefore analogous to the function of CPSF73 within the CPA complex, generating a 5'P which is a potential substrate for Xrn2.

Similar to histone genes, read-though RNA downstream of snRNA genes was unaltered in the wake of Xrn2 depletion (**Figure 5.5**[**A**]). To confirm this for every annotated snRNA gene within this library, I produced an average expression profile of annotated snRNA genes (and potential variants catalogued by Ensembl). An inclusion window was also wrapped around each snRNA interval to include 200 bp upstream of the TSS and 1.5 kb downstream of the TES sites. Any overlapping genes following the inclusion of this window were subsequently dropped from the metagene analysis. In agreement with the coverage tracks of RNA expression, I failed to detect any accumulation of 3' flanking product after downregulation of Xrn2 (**Figure 5.5**[**B**]). During preparation of nuclear RNA for sequencing, I did not specifically enrich for small RNAs such as snRNA, which may account for the large variance in snRNA expression between the 2 biological replicate metagene plots (**Figure 5.5; Supplemental Figure S11**). However, it is worth noting that sequencing of Pol II associated RNA in the presence and absence of Xrn2-AID confirmed the observation that Xrn2 has little involvement in snRNA termination (Eaton & Davidson *et al* 2018).

Thus, although Xrn2 is generally required for termination on poly(A) site containing protein-coding genes, transcription termination via the torpedo mechanism is not applicable to all genes. While the data presented here cannot confirm that termination is not defective over histone and snRNA genes alone, the lack of 3' flanking transcription read-through coincides with the absence of Pol II occupancy downstream of the TES during mNET-Seq profiling (Eaton & Davidson *et al* 2018), implying that Xrn2 is not involved in the termination of these genes.



Eaton & Davidson *et al* 2018

Figure 5.5: (**A**) RPKM normalised coverage tracks of read density over snRNA genes. Scale bars = 1 kb apply to a 3 coverage tracks within each gene image respectively. (**B**) Transcription profile of non-overlapping snRNA genes with a 1.5 kb extended region flanking the TES (n = 707). Additional biological replicate found in **Supplemental Figure S11**.

5.3 Xrn2 Downregulation has a Minimal Impact on Gene Expression

While I have so far focused primarily on the effects of Xrn2 during the late stages of transcription, Xrn2 can potentially degrade a broad range of transcripts, providing that they lack a mature cap structure at the 5' end of the RNA (or undergo endo- cleavage to produce a 5'-P). I reasoned that in the absence of Xrn2, short uncapped transcripts arising from failed processing during the early stages of transcription might become stabilised and detectable within my RNA-Seq library, which I sought to identify using differential gene expression analysis.

5.3.1 Differential Gene Expression Analysis

The full list of annotated genes (~58,000 Ensembl annotated genes) used in both the Exosc10 and Dis3 knockdown analysis were again used to measure gene expression, this time in cells lacking Xrn2. From the analysis of this annotation set I was only able to detect 566 (not including IAA upregulated genes) significantly upregulated genes (**Figure 5.6**[**A**]). Interestingly, global transcript abundance in the wake Xrn2 depletion is almost intermediate in effect between Exosc10 and Dis3 gene expression analysis (**Figure 3.12; Figure 4.5**), the latter of which displays a much wider dispersion pattern. Comparatively, Xrn2 has a minimal impact on the expression of known annotates genes.

Characterisation of these upregulated genes also revealed that a high proportion (~60%) of transcripts are protein-coding (**Figure 5.6**[**B**]), implying that unlike Dis3, Xrn2 has a much more narrow RNA substrate specificity. However, as I will discuss in the next section, a significant number of differentially upregulated genes detected in this analysis are false positive.

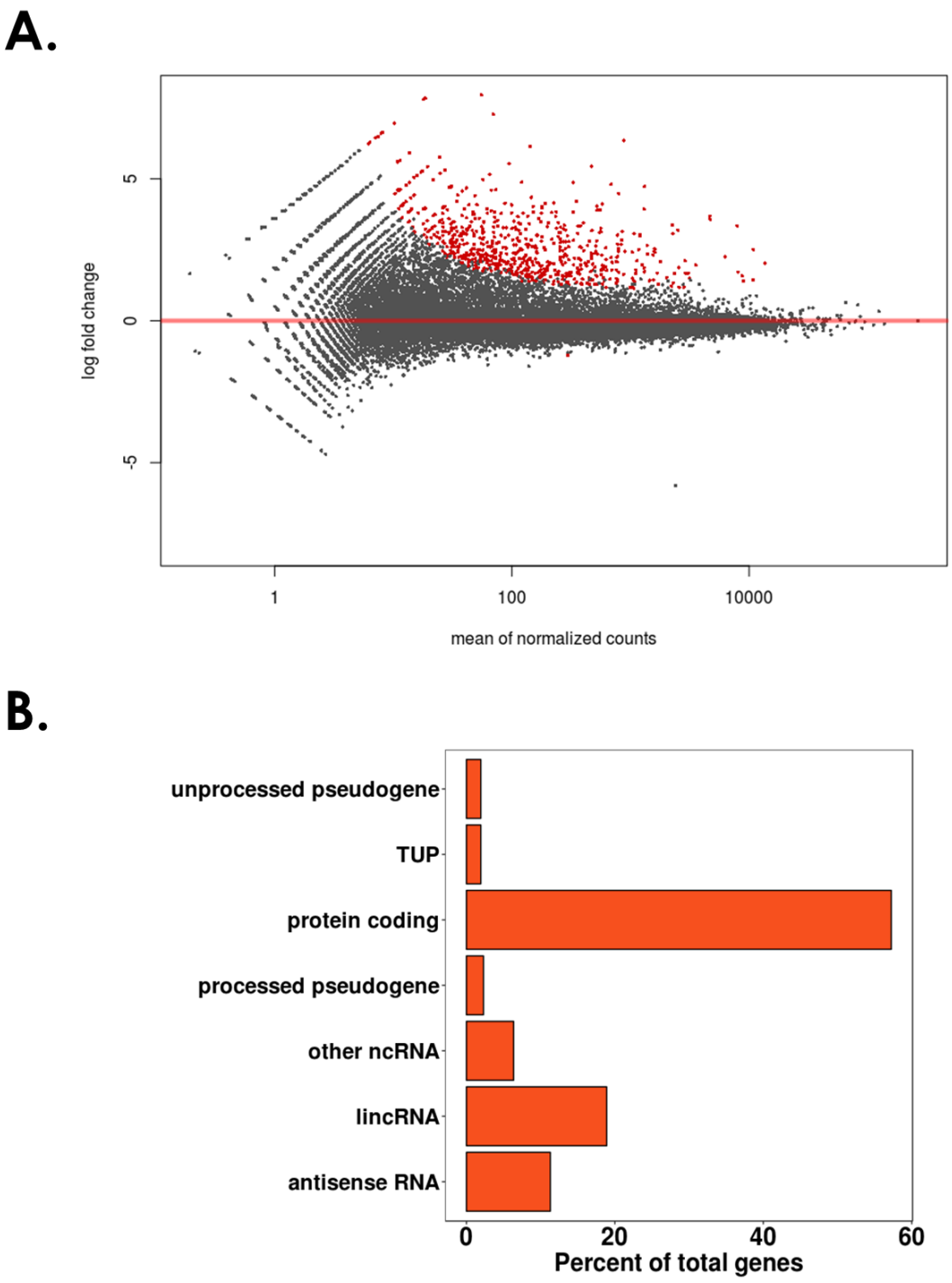


Figure 5.6: (**A**) Differential gene expression of 58,302 annotated genes represented as the log2-fold change relative to untreated XRN2-AID cells. Significantly upregulated genes are shown in red. (**B**) Classification of 566 upregulated genes (≥ 2-fold; padj < 0.05) based on their transcript biotype (TUP = transcribed unprocessed pseudogene).

5.3.2 Failure to Terminate Transcription Causes Accumulation of 3' Flanking RNA over Neighbouring Genes

Due to the high degree of false positive upregulated genes detected in the previous Dis3 differential expression analysis, coupled with the significant stabilisation of RNA from intergenic regions, I decided to validate the upregulated genes identified by visualising their relative expression levels.

Similar to the Dis3 analysis where stabilisation of PROMPT transcripts overlap nearby gene intervals, verification of the list of upregulated genes determined that the majority of DESeq2 "hits" detected by the software actually represent false positives. Visualisation of the normalised sequence coverage of each strand individually, I discovered that under normal conditions genes within the upregulated list are expressed at a very low, almost background level. Following Xrn2 downregulation however, RNA derived from a nearby upstream gene appears to be stabilised over the intergenic space between both genes eventually overlapping the neighbouring downstream ORF (**Figure 5.7**). In each of the 3 examples shown in **Figure 5.7**, RNA-Seq reads mapped only to the antisense strand, consistent with the orientation of both genes within the loci was shown to accumulate, indicating that the stabilised 3' flanking RNA likely represents read-though transcription beyond the TES.

I therefore conclude that, like for Dis3, Xrn2 loss does not alter gene expression directly, at least not within the context of this nuclear RNA enriched transcriptome analysis, but instead the dysfunction of gene punctuation at gene 3' ends inadvertently causes the expression of nearby downstream genes through transcriptional read-through. This analysis further highlights the importance of Xrn2's role at enforcing strict termination of Pol II transcription within the designated "termination window", which is particularly important for genes that are grouped in clustered loci throughout the genome.

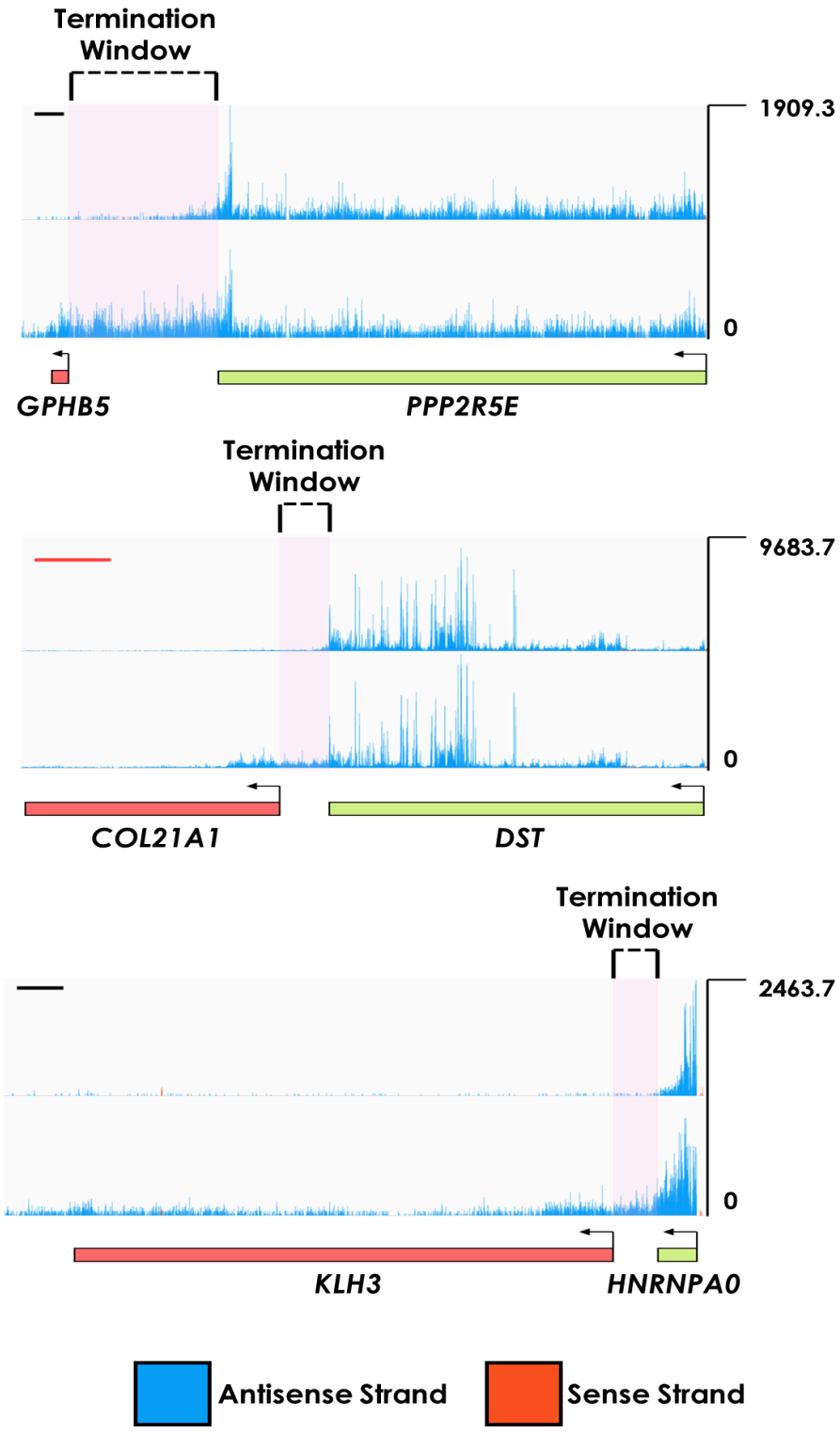


Figure 5.7: RPKM normalised coverage tracks of upregulated genes (red) arising from read-through beyond the termination window of nearby genes (green). Black scale bars = 10 kb, apply to both coverage tracks represented in *PPP2R5E* and *HNRNPA0* genes. For *DST,* the red scale bar = 100 kb and applies to both tracks. Tracks represent 2 biological replicates.

5.4 truncRNA Transcripts are not Xrn2 Substrates

Earlier I mentioned that short truncRNA transcripts detected over the +1 nucleosome in Dis3 downregulated cells may be the result of improper 5' capping, which are then released from Pol II by abortive transcription termination at promoter-proximal QC checkpoints. As such, uncapped truncRNAs would be potential substrates of Xrn2 and therefore stabilised in its absence.

I decided to directly compare both DIS3-AID and XRN2-AID RNA-Seq libraries after conditional knockdown of each protein by analysing the single nucleotide sequence coverage over the TSS of the same non-overlapping genes used in previous metagene analysis. While a peak of short stabilised truncRNA was detected in the Dis3 cell line, treated with auxin, I did not detect an equivalent peak of enriched mapped reads over the same region as a consequence of Xrn2 depletion (**Figure 5.8**). It is very likely that truncRNAs undergo 5' capping and are therefore protected from Xrn2-mediated degradation, which would explain why they are not stabilised following Xrn2 depletion. However, I cannot fully confirm truncRNAs are capped from this analysis alone. This would require further investigation using techniques like cap analysis of gene expression (CAGE), which can map the 5' end of capped RNAs at single nucleotide resolutions (Kodzius *et al* 2006).

One possible explanation regarding the emergence of truncRNA transcripts can be assumed from the inherent difficulty transcribing the initial 200 nt of DNA wound around the +1 nucleosome. Strict placement of the +1 nucleosome may act as a physical barrier that reduces the likelihood of Pol II progression into processive elongation (Chiu *et al* 2018). Thus, the majority of engaged Pol II complexes would ultimately dissociate from the template strand due to inefficient transcription around the +1 nucleosome. If this were true then truncRNAs would still undergo 5' capping and explain why Dis3, a 3'→5' exoribonuclease would be the only enzyme capable of orchestrating their decay.

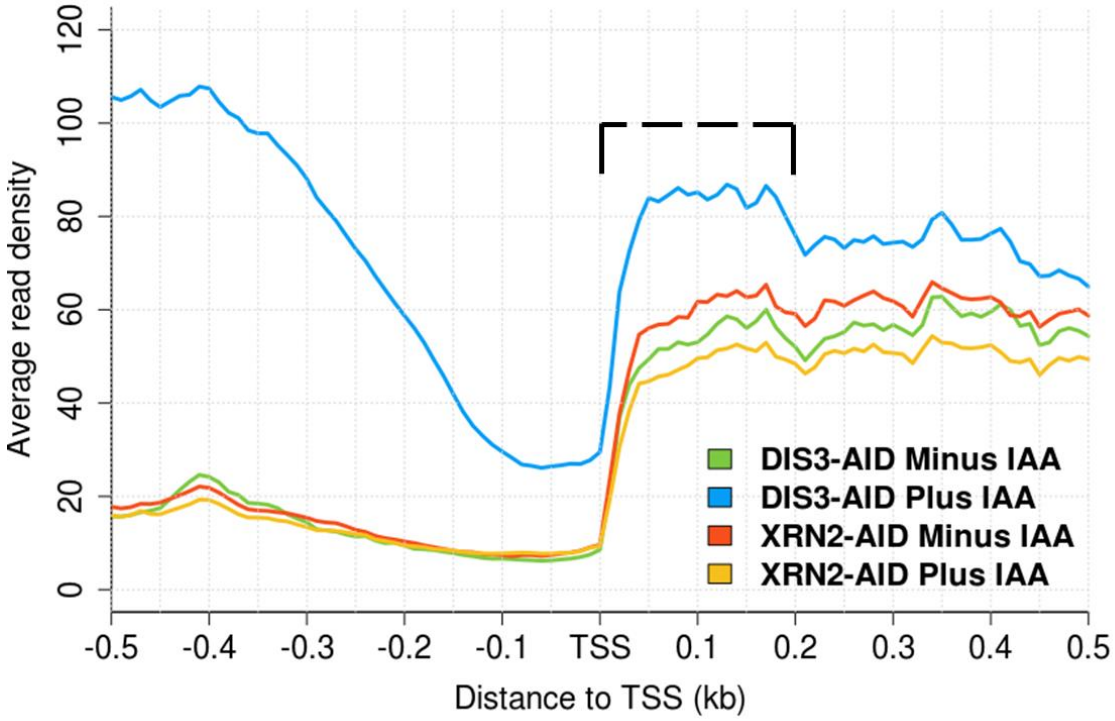


Figure 5.8: Single nucleotide coverage comparison of truncRNA centred on the TSS of 4701 genes in Dis3 and Xrn2 depleted cells. A second biological replicate is presented in **Supplemental Figure S12.**

5.5 *De novo* eRNA-like Transcripts are not Degraded by Xrn2

In the previous chapter I dedicated a significant section to the characterisation of transcripts expressed from intergenic regions of the genome that were discovered once Dis3 was depleted from the nucleus. In the end I designated these transcripts as eRNA-like since they displayed similar epigenetic patterns, promoter orientation and proximity to annotated eRNAs. An additional trait of eRNA transcripts is the inclusion of a 5' cap modification (Andersson *et al* 2014), which would prevent 5'→3' degradation by Xrn2. Using the XRN2-AID degron cell line I decided to test the eRNA-like transcripts for capping potential as a final measure to confirm the identity of these *de novo* intergenic RNAs.

Analysis of sequencing coverage over three eRNA-like intervals which exerted strong bidirectional transcription following Dis3 depletion, I only detected background levels of RNA expression even after downregulation of Xrn2 (**Figure 5.9**[**A**]). In spite of the very low levels of expression of each interval examined, the bidirectionality of the promoter within each interval was still apparent, indicating that RNA transcribed from these promoter sequences occurs at very low levels under normal conditions. Lastly, I summarised the average RNA expression over every eRNA-like interval detected by the *de novo* transcript assembly software, and determined that Xrn2 does not stabilise RNA transcribed from the total intergenic eRNA-like dataset (**Figure 5.9**[**B**]).

In agreement with published data, these eRNA-like transcripts are likely capped at the 5' end preventing their degradation by Xrn2 (Andersson *et al* 2014). Since no evidence of splicing or polyadenylation has been observed during eRNA processing (Kim *et al* 2015), turnover by Dis3 remains the most viable pathway available for their removal from the transcriptome.

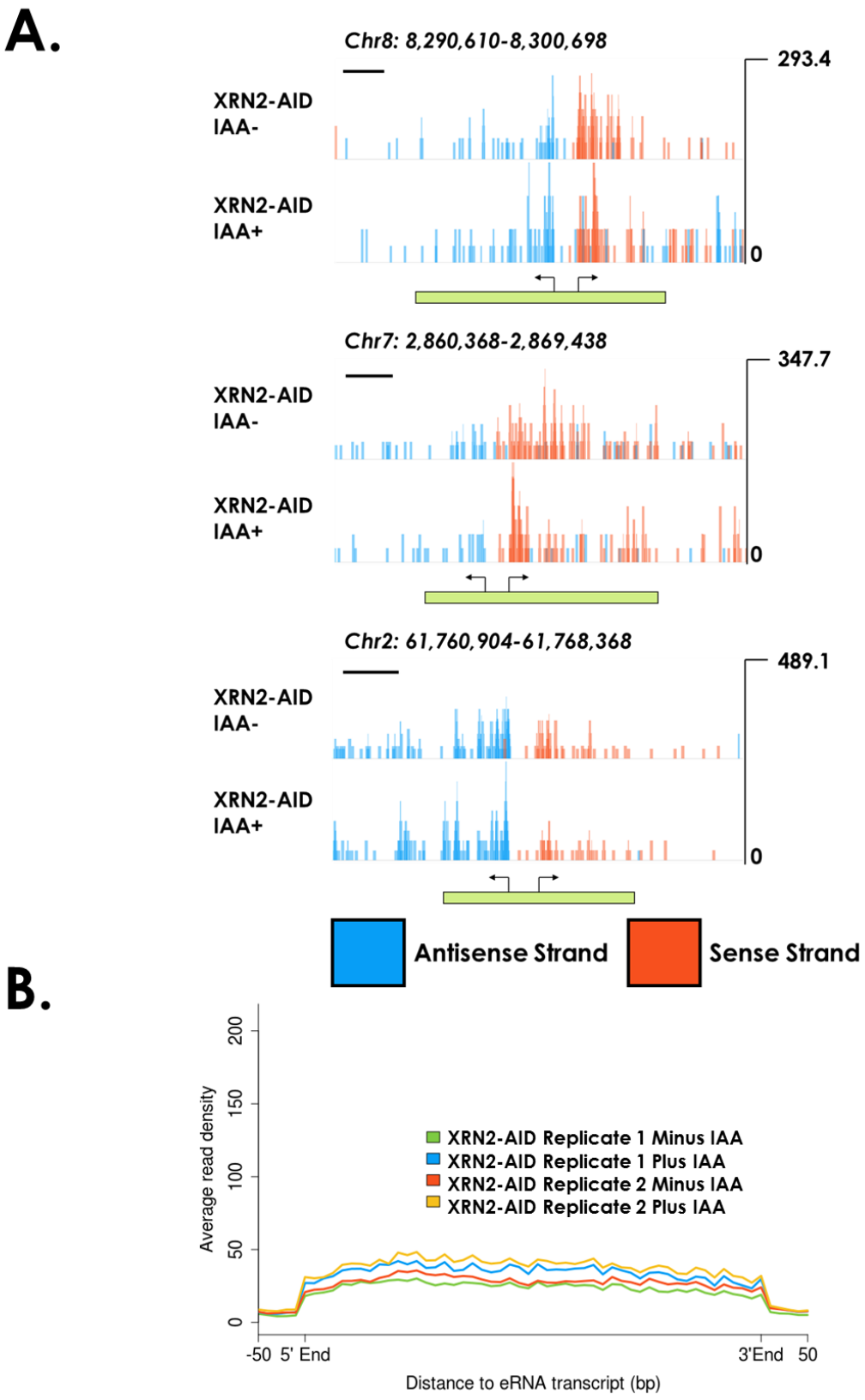


Figure 5.9: (**A**) RPKM normalised coverage tracks of 2 biological replicates over the *de novo* eRNA-like intergenic intervals. Scale bars = 1 kb apply to both coverage tracks within each image respectively. (**B**) Average sequencing coverage over every (n= 960) enhancer sequence identified.

5.6 Summary

In this chapter, I have demonstrated that Xrn2 is required for termination of Pol II transcription by torpedo through the rapid degradation of uncapped 3' flanking RNA sequence following cleavage at the poly(A) site. Additionally, disruption of Pol II termination caused by Xrn2 downregulation leads to stabilisation of 3' flanking RNA sequences and has the potential to cause transcriptional read-though into neighbouring genes, creating non-sense transcripts.

In order for torpedo termination to occur, Xrn2 must be able to catch the elongating Pol II complex as it degrades the downstream RNA sequence (Kim *et al* 2004; West *et al* 2004), therefore a termination window exists within the intergenic region immediately downstream of the TES. The termination window in effect could act to slow Pol II elongation through chromatin remodelling and/or pausing at alternative poly(A) sites providing time for Xrn2 to catch the elongating complex (Fong *et al* 2015). While I often observed a greater accumulation of 3' flanking RNA immediately following the proposed cleavage site, read-through RNA was detected as far as ~100 kb downstream of the gene 3' end indicating that the absence of Xrn2 greatly increases the distance required for Pol II termination to occur. Interestingly, despite the near complete depletion of Xrn2 using the AID cell line (Eaton & Davidson 2018; see also **Appendix Figure 1**), termination of Pol II still occurs, either through the action of residual Xrn2 protein undetectable by western blot, or from an alternative termination mechanism, possibly resembling the allosteric model (Osheim *et al* 2002).

Termination by torpedo is not common to all genes and requires more regulation than just 3' endonuclease cleavage. Both histone and snRNA transcripts undergo 3' end formation using alternative pathways independent of cleavage and polyadenylation. Interestingly, several CPA factors are also required for 3' processing of histone RNA maturation, however the activity of CPSF73 in this scenario is not sufficient to direct Xrn2

to the nascent 3' flanking RNA. Likewise, snRNAs are processed at the 3' end via the integrator complex which contains subunits that perform analogous functions to CPA factors, also terminate transcription independently of Xrn2 activity. This analysis demonstrates a strong link between cleavage at the poly(A) site and Xrn2-dependant termination, however it is still unclear why cleavage by CPSF73 in this instance initiates Xrn2-mediated degradation. It is likely that the recruitment and activity of Xrn2 is regulated by the composition of CPA factors bound to *cis* acting poly(A) elements possibly in combination with Pol II CTD modifications.

With the exception of read-though beyond the TES, I did not detect any significant alteration of gene expression after 60 minutes of Xrn2 knockdown. Moreover, many of the false positive differentially expressed genes were in fact due to transcriptional read-though from nearby upstream genes, which would otherwise be efficiently terminated by Xrn2. Given the importance of Xrn2 at enforcing gene punctuation at gene 3' ends, I cannot rule out the possibility that long-term downregulation of Xrn2 would cause significant disruption the composition of the transcriptome, in part, due to reduced recycling of Pol II complexes back to the TSS during subsequent rounds of transcription, but also due to the build-up of non-sense or missense transcripts.

In contrast to Dis3, the activity of Xrn2 during transcription is largely restricted to the 3' end of the gene. truncRNA transcripts released from the template strand during early transcription abortion are not substrates of Xrn2-mediated decay, presumably due to the presence of a 5' cap structure. It is possible that a high proportion of truncRNAs released from the template DNA arise as a consequence of inefficient transcription around the +1 nucleosome sequence which stalls Pol II elongation facilitating termination. A similar mechanism was recently published showing that Pol II pausing at the +1 nucleosome acts as a checkpoint to elongation whereby, Pol II complexes that are unable to enter progressive elongation are terminated through the premature cleavage at nearby poly(A) sites (Chui *et al* 2018).

Finally, stabilisation of eRNA-like transcription cannot explain the increased level of intergenic transcription observed as a consequence of Xrn2 depletion. Although endogenous Dis3 is still active within the XRN2-AID cell line, low level bidirectional transcription was observed over the *de novo* assembled eRNA-like gene intervals which persists even in the absence of Xrn2. Consistent with previous findings, I reasoned that these eRNA-like transcripts are capped at the 5' end making them resistance to Xrn2-mediated degradation, and since eRNAs lack poly-A tails, they remain susceptible to 3'→5' decay by Dis3.

Chapter 6

Discussion

Nuclear RNA surveillance pathways contribute substantially toward the quality of transcriptional output from RNA polymerases. Several nuclear exoribonucleases and auxiliary accessory proteins cooperate, often co-transcriptionally with actively transcribing polymerase complexes to target and degrade a broad range of unwanted RNA molecules thereby protecting the integrity of the transcriptome. Moreover, surveillance of polymerase output in this manner provides a means of "fine tuning" transcriptional output at later stages following initial gene regulation at the chromatin level. Functional studies of the protein complexes involved in human nuclear RNA surveillance has been considerably difficult in the past in part due to the relatively indirect, slow and often incomplete level of gene downregulation achievable by RNAi. In this study, I have incorporated an auxin-inducible degron system into the HCT116 cell line, capable of rapidly and specifically depleting a target protein *in vivo*. In doing so, this thesis provides the most comprehensive atlas, to date, of immediate substrates for three major exoribonucleases present in human nuclei.

The Auxin Degron System is a Viable Alternative to RNAi

Auxin-inducible protein degradation is an efficient system capable of conditionally downregulating gene expression post-translationally within plants (Gray *et al* 2001; Dharmasiri *et al* 2005). In recent years, the AID system has been exploited as a molecular tool for functional genomics in several non-plant metazoans, since it utilizes the same functional components of the SCF complex (minus the F-box protein TIR1) required

for ubiquitin-mediated proteome degradation (Nishimura *et al* 2009; Holland *et al* 2012; Morawska & Ulrich 2013). Thanks to advancements in CRISPR/Cas9 genome engineering technology, it is now possible to incorporate an AID tag to any target gene within the human genome with relative ease (Natsume *et al* 2016), heralding a return to more traditional and direct functional genomic approaches applicable to metazoans.

In this study, I have shown that near complete protein depletion in HCT116 cells expressing the plant TIR1 F-box gene and an engineered AID presenting target protein, is achievable within 60 minutes following the introduction of AUX/IAA into the growth media (**Figure 3.5**; **Appendix Figure 1;** Eaton & Davidson *et al* 2018). Furthermore, protein depletion via the AID system is only possible if both the TIR1 protein and the AID-tagged protein are present. Importantly, the stability of Exosc10 following inclusion of the AID tag to the 3' end was unaltered since the abundance of Exosc10-AID remained comparable to endogenous Exosc10 protein levels, and additionally no observable growth rate was detected despite constitutive expression of plant TIR1 (**Supplementary Figure S3**). Although the AID system is responsible for a small number of off-target effects (**Table 3.1**), namely the upregulation of certain genes involved in the metabolism of uremic toxins caused by IAA in the growth media (Sallee *et al* 2014), they are far fewer compared to the off-target siRNA binding potential within the genome (Qui *et al* 2005; Smith *et al* 2017). Collectively, the rate of protein depletion and direct protein targeting attainable by AID prevents the build-up of indirect effects providing a much clearer understanding of gene function at temporal resolutions unachievable by RNAi.

Exosc10 is vital for Cell Proliferation

In yeast, knockout of the Exosc10 homologue Rrp6, was shown to be not essential for cell viability, and its loss instead contributes to a slow growth phenotype (Briggs *et al* 1998; Januszyk *et al* 2011). Exosc10 in metazoans

however, has been shown to be essential for cell proliferation, reinforced by the discovery that Exosc10 fulfils a secondary function during mitotic spindle assembly in *D. melanogaster* S2 cells (Graham *et al* 2009; Kiss & Andrulis 2010).

Investigation of Exosc10 as part of this study determined that the near complete, prolonged depletion of Exosc10 protein within HCT116 severely reduced cell survivability (**Figure 3.7**). Interestingly, the catalytic activity of Exosc10 appeared to be somewhat dispensable for cell proliferation since co-overexpression of a catalytically inert mutant Exosc10 was able to partially recover cell viability (**Figure 3.10**[**A**]). Since rRNAs are crucial for translation of mRNA to protein, cell death in this instance was initially attributed to the disruption of pre-rRNA processing detected in response to Exosc10 knockdown. However, defects in 5.8S processing appear to be considerably more pronounced in cells expressing the Exosc10 mutant (**Figure 3.10**[**B**]), indicating that the structural presence of Exosc10 within the nucleus is perhaps more important for cell viability. This will be an interesting feature to follow up.

Dis3 is the Major Exoribonuclease Component of the Exosome Complex

Exosc10 and Dis3 both catalyse 3'→5' degradation of substrate RNA, however in spite of this, both exoribonucleases have been proposed to degrade separate groups of RNA substrates. While Dis3 degrades a broad range of largely unstructured RNA (Szczepinska *et al* 2015), Exosc10 has been proposed to be involved in the degradation of small complex structured RNAs such as pre-rRNA and snoRNAs (Januszyk *et al* 2011). Additionally, Exosc10 facilitates substrate targeting to the active site of Dis3 by threading RNA through the central channel of the exosome itself (Mitchell 2014; Kilchert *et al* 2016; Ogami *et al* 2018). Due to their intimate relationship and potential redundancy as part of the exosome, it has therefore been difficult in the past to functionally separate each protein

in order to categorise specific RNA substrates assigned to each exoribonuclease.

By applying the AID degron system to each exosome subunit, I was able to investigate the immediate effects following protein loss before the activation of any potential redundant pathways available in the nucleus. Here, I reveal that loss of Exosc10 has almost no effect on the immediate nuclear RNA composition of the transcriptome. In stark contrast, the loss of Dis3 causes a significant accumulation of numerous types of short RNAs derived from spurious transcription initiation and premature termination events (**Chapter 4**). While the majority of upregulated transcripts are derived from the stabilisation of PROMPT RNAs at known bidirectional protein-coding gene promoters, a considerable proportion of the upregulated transcriptome is composed of intergenic enhancer RNAs. Since Exosc10 downregulation was unable to recapitulate these results, it seems likely due to the preferred nucleolar localisation of Exosc10 (Lykke-Andersen *et al* 2011), that an exosome complex lacking associated Exosc10 may be responsible for the degradation of these transcripts (**Figure 6.1**). However, without reconstitution or visualisation of the exosome complexes in nuclear sub-compartments, this remains a working theory.

How are Exoribonucleases Targeted to Specific RNA Substrates?

The broad range of Dis3-sensitive RNA substrates supports previous finding suggesting that Dis3 is the major exoribonuclease of the core exosome (Dziembowski *et al* 2007; Szczepinska *et al* 2015). However, it is important to consider how the exosome is specifically targeted to RNA substrates. This is especially true when considering transcripts such as PROMPT RNAs which undergo the same co-transcriptional processing steps including, 5' capping and 3' end formation, making them almost indistinguishable in structure from mRNAs (Preker *et al* 2011).

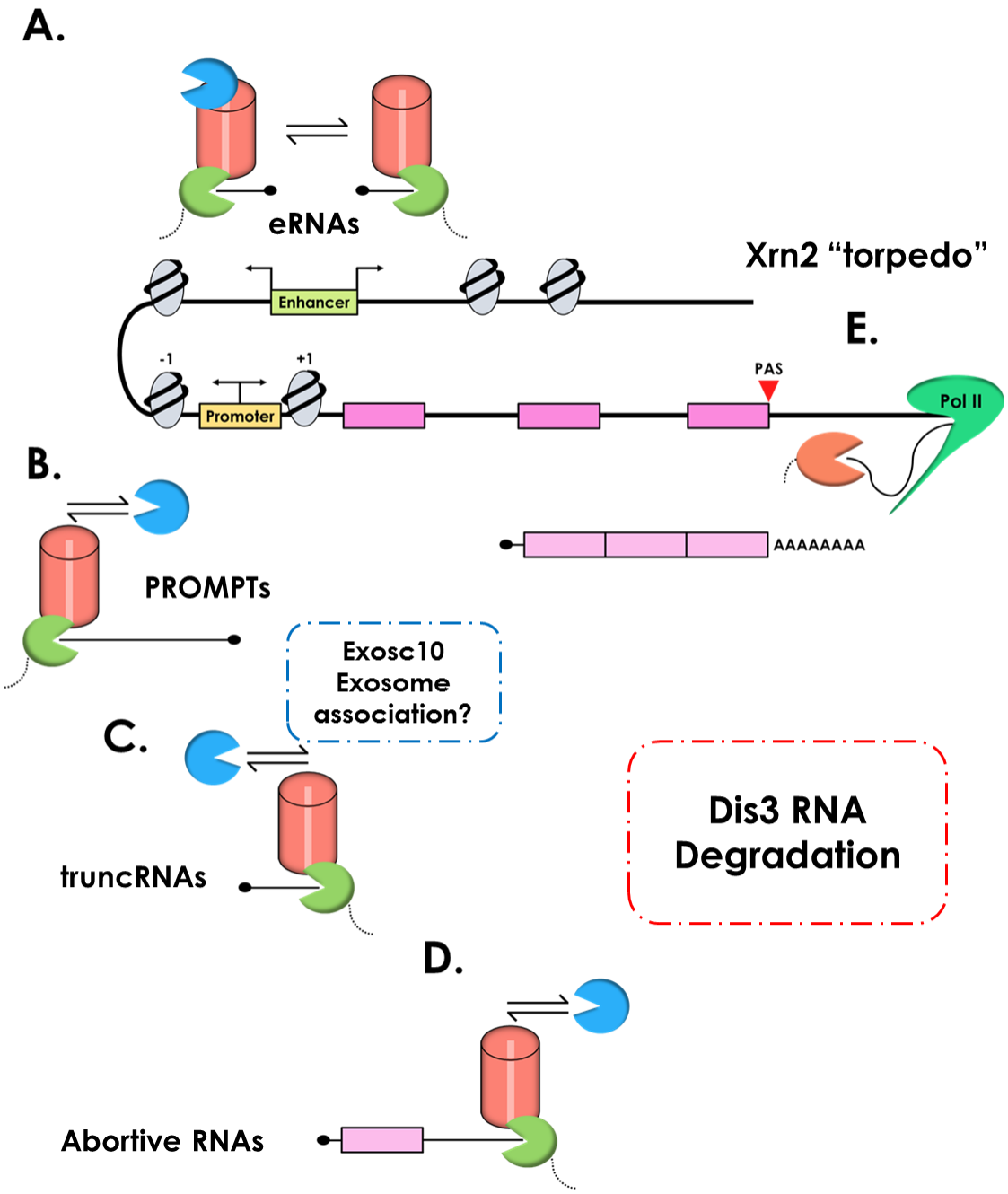


Figure 6.1: Reviewed model of nuclear RNA surveillance pathways in human nuclei during early RNA biogenesis. Dis3 (green pacman) degrades eRNA (**A**), PROMPT (**B**), truncRNA (**C**) and prematurely aborted transcripts (**D**) as part of the exosome complex, either with or without association with Exosc10 (blue pacman). Xrn2 (red pacman) predominantly degrades 3' flanking RNA (**E**) following cleavage at the poly(A) site (PAS) as part of the torpedo termination model. It is still unclear whether Exosc10 is also associated with the exosome as Dis3 during degradation of these pervasive transcripts.

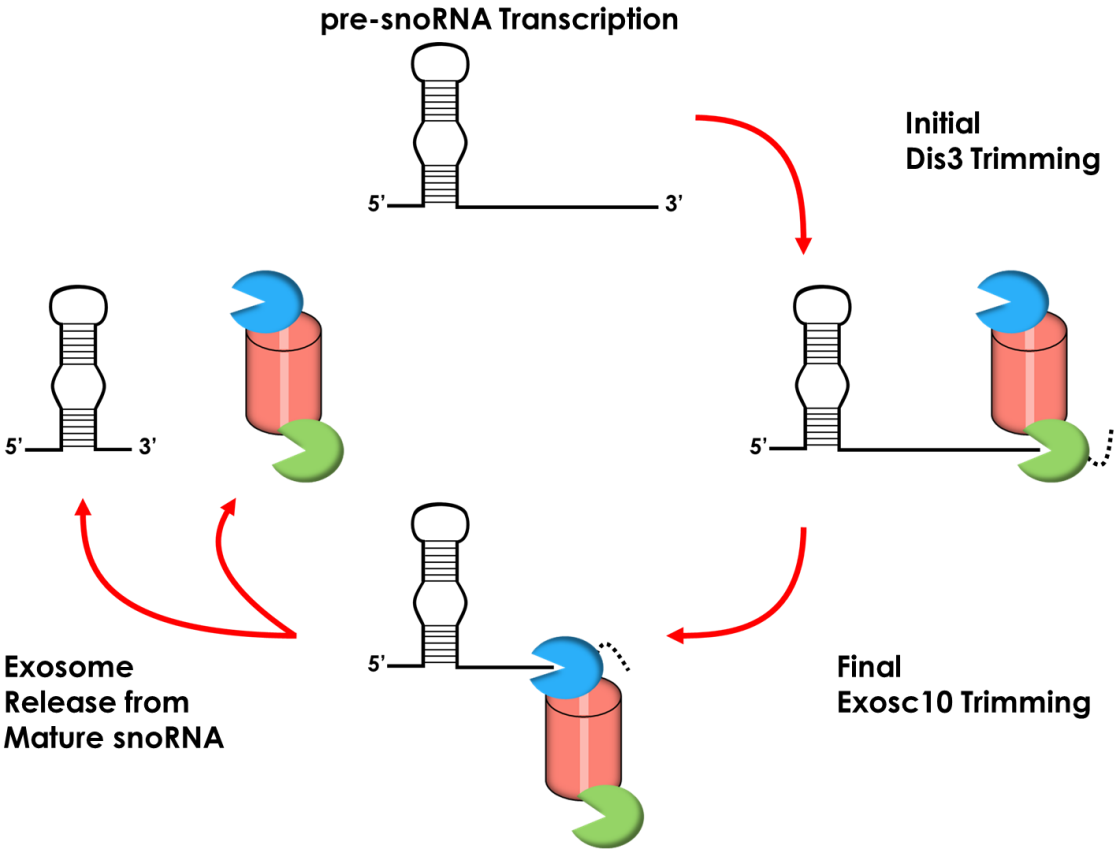


Figure 6.2: Proposed model of 3' extended pre-snoRNA processing by the exosome complex in which, the 3' extension is initially trimmed by Dis3 (green) before subsequent trimming by Exosc10 (blue) releasing the mature snoRNA transcript.

As I mentioned previously in **Chapter 1**, two well-known multi-subunit complexes: TRAMP-like and the NEXT complex are present within human nuclei, and have been shown to interact with the exosome facilitating its loading onto RNA substrates (Ogami *et al* 2018). Additionally, the poly(A) tail exosome targeting (PAXT) complex, has recently been discovered and shown to facilitate exosome-mediated degradation of polyadenylated RNA transcripts (Meola *et al* 2016). Central to the core composition of each of these three complexes is the presence of the RNA helicase Mtr4. In addition to its role as a helicase, Mtr4 is intimately associated with Exosc10 (Lubas *et al* 2011) and acts as an anchor during the assembly of factors required for the formation of each nuclear targeting complex (Meola *et al* 2016). The composition of each complex is therefore a determining factor in directing the exosome to specific transcripts. For example, TRAMP-like is restricted to nucleoli due to its association with the strictly nucleolar zinc finger protein ZCCHC7 (Lubas *et al* 2011), and as such preferentially enhances exosome activity towards nucleolar and snoRNA transcripts. Similarly, NEXT which contains ZCCHC8, is closely associated with the ARS2-associated CBC complex (CBCA) which preferentially targets cryptic RNAs such as PROMPTs and eRNAs, whereas PAXT has been shown to interact with the polyadenylation factor PABPN1 via ZCF3H1 (Ogami *et al* 2018). Therefore, the broad activity of the nuclear exosome complex can be repurposed to target a specific RNA substrate by altering the composition of its associated accessory factors.

Although beyond the scope of this study, an unbiased approach to determine the composition of associated factors recruited to the exosome during the turnover of the RNAs detected in this analysis, could be achieved using proximity protein labelling, followed by mass-spectrometry analysis using the newly developed mini-turbo assay (Branon *et al* 2018).

What is the Termination Mechanism Releasing Dis3-sensitive Abortive and Premature RNAs?

During this analysis, I discovered several RNA substrates that accumulate as a consequence of Dis3 depletion which I suggest are likely to derive from either premature or abortive transcription termination (**Figure 4.8**; **Figure 4.10**). Although I am unable to determine the mechanism involved in the termination of these transcripts, there are numerous publications describing the presence of promoter-proximal (< 5 kb), intronic cryptic poly(A) sites (Kaida *et al* 2010; Berg *et al* 2012) which can act to prematurely terminate Pol II transcription. Additionally, the asymmetrical distribution of poly(A) sites surrounding the promoter region have been shown to enforce promoter directionality in favour of the sense coding gene (Ntini *et al* 2013; Andersson *et al* 2014). Furthermore, recent work performed by Chui *et al* (2018) have described an early transcription QC checkpoint at the +1 nucleosome that can prematurely terminate arrested Pol II through cleavage at nearby cryptic poly(A) sites. Given the striking accumulation of RNA reads aligned to promoter-proximal introns and truncRNAs, it would be interesting to investigate the possibility that Dis3 remains poised alongside +1 nucleosome paused Pol II complexes in order to rapidly remove these aberrant RNAs from the transcriptome.

Pre-snoRNA 3' trimming is a 2-step process involving both Exosc10 and Dis3

Similar to pre-rRNA transcripts, premature snoRNAs are transcribed with an additional 3' extended RNA sequence that requires 3'→5' trimming before release of the mature snoRNA isoform. In a recent study performed by Szczepinska *et al*, Dis3, but not Exosc10 was shown to participate in the 3' processing of C/D box containing snoRNAs. In the same study however, mutant Exosc10 protein also failed to detect 5.8S rRNA processing defects.

In this study, I detected snoRNA processing defects in both Exosc10 and Dis3 depleted cell lines (**Figure 3.16**). Interestingly, in the absence of Exosc10, a 10-30 nt 3' extended snoRNA isoform was detected, consistent with what we observed for 5.8S rRNA. However, following Dis3 downregulation, accumulation of a longer 3' extended RNA sequence (~100 nt) was detected. From these results, I propose that snoRNA processing occurs in 2-steps, whereby Dis3 catalyses the initial trimming of the 3' extension before handing over to Exosc10 which removes the final 30 nt RNA sequence (**Figure 6.2**). This would explain why the 10-30 nt extended snoRNA precursors were able to accumulate in Exosc10 null cells, since Dis3 activity during the initial trimming process remained present. Since the snoRNA extension seen on Dis3 loss does not extend to the 3' end of the intron, there may be an as yet uncharacterised endonuclease cleavage step involved in their processing/release from the intron.

Due to the disparate distribution of Exosc10 and Dis3 within the nucleus, it remains unclear if this 2-step snoRNA processing reaction is performed by the full EXO-9/Exosc10/Dis3 exosome complex, or if the snoRNA intermediate migrates between the nucleoplasm and nucleoli. Moreover, of the two distinct classes of snoRNA (Reichow *et al* 2007; Jorjani *et al* 2016), Exosc10 depletion caused a greater processing defect in HCA/A box containing snoRNAs, compared with C/D box isoforms, however it remains unclear if each class of snoRNA transcripts undergo separate maturation pathways. Finally, despite detecting the 3' extended snoRNA transcripts in this dataset, snoRNAs were not enriched during library preparation, and due to their small transcript length, RNA-Seq of nascent snoRNAs does not provide adequate resolution required to further investigate 3' snoRNA processing. Additional characterisation using techniques such as Individual-nucleotide resolution UV crosslinking and immunoprecipitation (iCLIP), can identify protein-RNA interactions at single nucleotide resolutions may be able to detect the proposed exchange of pre-snoRNA between both exoribonucleases.

Xrn2 Enhances Transcription Termination of a Subset of Genes

Xrn2 is an important factor involved in the termination of elongating Pol II by torpedo downstream of gene 3' ends (Kim *et al* 2004; West *et al* 2004). Following cleavage at the poly(A) site, Xrn2 associates with the free 5'-P where it proceed in the degradation of the 3' flanking RNA sequence. If the rate of RNA degradation is greater that the elongation rate of the transcribing Pol II complex collision ensues, displacing Pol II from the template DNA by an unknown mechanism. Disruption of Xrn2 expression has been previously shown to delay transcription termination, by increasing the distance transcribed by Pol II downstream of the gene 3' end (Fong *et al* 2015).

Corroborating previous data, knockdown of Xrn2 using the AID system caused an observable transcription termination defect predominantly downstream of transcripts that undergo 3' cleavage and polyadenylation (**Chapter 5; Appendix;** Eaton & Davidson *et al* 2018). Interestingly, transcripts that undergo endonuclease cleavage by alternative poly(A) independent mechanisms do not always elicit a termination defect. This is in spite of the recruitment (in the case of histone transcripts) of CPA factors such as CPSF73. This highlights the possibility that recruitment of Xrn2 to the cleaved 3' flanking RNA is tightly regulated, possibly through interactions with the CTD of Pol II. Moreover, termination of Pol II transcription was still observed at distal regions downstream of the gene TES, implying that a redundant termination pathway (or pathways) exists. How Pol II is terminated in the absence of Xrn2 is unclear from this analysis, although a likely possibility is the allosteric model of termination which, triggers a conformational change, slowing the Pol II complex to the point of dissociation (Osheim *et al* 2002). However, reducing the elongation rate of Pol II could also be achieved through rearrangement of the chromatin landscape, which could be determined by ChIP-Seq analysis of the XRN2-AID cell line used in this study.

Future Work

Pervasive transcription is much more widespread among eukaryotes than previously envisaged. Despite strict regulation of transcription initiation, inappropriate transcription initiation is still responsible for the production of a high volume of spurious non-sense transcripts. Degradation of these transcripts by nuclear RNA surveillance pathways therefore provides an important defence mechanism preventing the build-up of unwanted RNAs, thereby protecting the integrity of the transcriptome.

This study demonstrates that each of the 3 major exoribonucleases present within human nuclei catalyse the degradation and processing of distinct classes of RNA transcript. In the case of Dis3 and Xrn2, as little as 60 minutes of gene downregulation is required to detect significant transcription perturbation, indicating that both nuclear exoribonucleases are likely poised close to the site of transcription, facilitating co-transcriptional RNA decay in agreement with previously published data (Almeida *et al* 2010; Davidson *et al* 2012).

While this investigation primarily focused on the expression of nascent RNA within the transcriptome, further examination of more prolonged periods of protein depletion would provide greater insight into the negative impact of Dis3/Xrn2 downregulation over successive rounds of transcription. Additionally, comparison of increasing intervals of protein downregulation would provide a mechanism to discover potential redundant degradation pathways that may only become active in response to cell stress. For example, the results presented in this study cannot fully rule out the possibility that Exosc10 and Dis3 do not share some overlapping substrate activity, similar to snoRNA processing, since the cells were not given sufficient time to adjust and, given the involvement of TRAMP and NEXT targeting of substrates to the exosome, it is possible that RNA substrates may be redirected to the remaining exoribonuclease.

The results presented in this study failed to definitively identify the cause of cell death following severe downregulation of Exosc10. Despite

the involvement of Exosc10 during pre-rRNA maturation (Januszyk *et al* 2011), I cannot conclude that defects in rRNA processing as a contributing factor, since cells expressing inert Exosc10 mutants continue to survive in perpetuity in spite of this deficiency. Investigating the role of Exosc10 during spindle assembly (Graham *et al* 2009; Kiss & Andrulis 2010) at each stage of the cell cycle would be achievable using the AID degron cell line and may provide a clearer understanding behind the drastic loss of cell viability.

Over the course of this analysis, I determined the function of Exosc10 and Dis3 by studying global changes within the transcriptome as a consequence of their absence. However, both proteins exist as part of the exosome complex, and as previously mentioned, several isoforms of the exosome exist within the nucleus (Lykke-Andersen *et al* 2011). If so then: What is the exosome composition required to degrade each class of RNA substrate? Addressing this using techniques such as mRNP capture assays, designed to identify mRNA-protein complexes, would provide a method of determining how each class of transcript is targeted to the exosome and the mechanism of degradation, in addition to determining an estimated location of RNA decay since both exoribonucleases are enriched within different sub-nuclear compartments.

Although several unanswered questions remain, the results represented in this study provides an initial characterisation of each of the 3 major nuclear exoribonucleases in humans, and in doing so, highlights the effectiveness of the auxin-inducible degron system as a tool for functional genomic studies.

References

Adelman K & Lis JT (2012) Promoter-proximal pausing of RNA polymerase II: emerging roles in metazoans. *Nat Rev Genet* **13:** 720-731

Albrecht TR & Wagner EJ (2012) snRNA 3' end formation requires heterodimeric association of integrator subunits. *Mol Cell Biol* **32**(6): 1112

Almada AE, Wu X, Kriz AJ, Burge CB & Sharp PA (2013) Promoter directionality is controlled by U1 snRNP and polyadenylation signals. *Nature* **449**(7458): 360-363

de Almeida SF, Garcıa-Sacristan A, Custodio N & Carmo-Fonseca M (2010) A link between nuclear RNA surveillance, the human exosome and RNA polymerase II transcriptional termination. *Nucleic Acids Res* **38**(22): 8015-8026

Andersen PR, Domanski M, Kristiansen MS, Storvall H, Ntini E, Verheggen C, Schein A, Bunkenborg J, Poser I, Hallais Y, Sandberg R, Hyman A, LaCava J, Rout MP, Andersen JS, Bertrand E & Jensen TH (2013) The human cap-binding complex is functionally connected to the nuclear RNA exosome. *Nat Struct Mol Biol* **20**(12): 1367-1376

Andersson R, Andersen PR, Valen E, Core LJ, Bornholdt J, Boyd M, Jensen TH & Sandelin A (2014) Nuclear stability and transcriptional directionality separate functionally distinct RNA species. *Nature Comms* DOI: 10.1038/ncomms6336

Andersson R, Gebhard C, Miguel-Escalada I, Hoof I, Bornholdt J, Boyd M, Chen Y, Zhao X, Schmidl C, Suzuki T, Ntini E, Arner E, Valen E, Li K, Schwarzfischer L, Glatz D, Raithel J, Lilje B, Rapin N, Bagger FO, Jørgensen M, Andersen PR, Bertin N, Rackham O, Burroughs AM, Baillie JK, Ishizu Y, Shimizu Y, Furuhata E, Maeda S, Negishi Y, Mungall CJ, Meehan TF, Lassmann T, Itoh M, Kawaji H, Kondo N, Kawai J, Lennartsson A, Daub C O, Heutink P, Hume DA, Jensen TH, Suzuki H, Hayashizaki Y, Muller F, The FANTOM Consortium, Forrest ARR, Carninci P, Rehli M & Sandelin A (2014) An atlas of active enhancers across human cell types and tissues. *Nature* **507:** 455-461

Andrews S (2010) FastQC: a quality control tool for high throughput sequence data. Available online at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc

Annunziato A (2008) DNA packaging: Nucleosomes and chromatin. *Nature Education* **1**(1): 26

Barnett DW, Garrison EK, Quinlan AR, Strömberg MP & Marth GT (2011) BamTools: a C++ API and toolkit for analyzing and managing BAM files. *Bioinformatics* **27**(12): 1691–1692

Benchling [Biology Software]. (2017) Retrieved from https://benchling.com

Bentley DL (2014) Coupling mRNA processing with transcription in time and space. *Nat Rev Genet* **15:** 163-175

Berg MG, Singh LN, Younis I, Liu Q, Pinto AM, Kaida D, Zhang Z, Cho S, Sherrill-Mix S, Wan L & Dreyfuss G (2012) U1 snRNP determines mRNA length and regulates isoform expression. *Cell* **150:** 53-64

Bieberstein NI, Carrillo Oesterreich F, Straube K & Neugebauer KM (2012) First exon length control active chromatin signatures and transcription. *Cell Rep* **2:** 62-68

Boettcher M & McManus MT (2015) Choosing the right tool for the job: RNAi, TALEN or CRISPR. *Mol Cell* **58**(4): 575–585

Brannan K, Kim H, Erickson B, Glover-Cutter K, Kim S, Fong N, Kiemele L, Hansen K, Davis R, Lykke-Andersen J & Bentley DL (2012) mRNA decapping factors and the exonuclease Xrn2 function in widespread premature termination of RNA polymerase II transcription. *Mol Cell* **46**(3): 311-324

Branon TC, Bosch JA, Sanchez AD, Udeshi ND, Svinkina T, Carr SA, Feldman JL, Perrimon N & Ting AY (2017) Directed evolution of TurboID for efficient proximity labeling in living cells and organisms. *bioRxiv* doi: http://dx.doi.org/10.1101/196980

Briggs MW, Burkard KTD & Butler JS (1998) Rrp6p, the yeast homologue of the human PM-Scl 100-kDa autoantigen, is essential for efficient 5.8 S rRNA 3' end formation. *J Biol Chem* **273**(21): 13255–13263

Brouwer R, Pruijn GJM & van Venrooij WJ (2001) The human exosome: an autoantigenic complex of exoribonucleases in myositis and scleroderma. *Arthritis Res* **3:** 102-106

Buratowski S (2009) Progression through the RNA Polymerase II CTD cycle. *Mol Cell* **36**(4): 541–546

Cabili MN, Trapnell C, Goff L, Koziol M, Tazon-Vega B, Regev A & Rinn JL (2011) Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. *Genes Dev* **25:** 1915-1927

Callahan KP & Butler JS (2008) Evidence for core exosome independent function of the nuclear exoribonuclease Rrp6p. *Nucleic Acids Res* **36**(21): 6645-6655

Cong L, Ran FA, Cox D, Lin S, Barretto R, Habib N, Hsu PD, Wu X, Jiang W, Marraffini LA & Zhang F (2013) Multiplex genome engineering using CRISPR/Cas systems. *Science* **339**(6121): 819-823

Carneiro T, Carvalho C, Braga J, Rino J, Milligan L, Tollervey D & Carmo-Fonseca M (2007) Depletion of the yeast nuclear exosome subunit Rrp6 results in accumulation of polyadenylated RNAs in a discrete domain within the nucleolus. *Mol Cell Biol* **27**(11): 4157-4165

Carrozza MJ, Li B, Florens L, Suganuma T, Swanson SK, Lee KK, Shia W, Anderson S, Yates J, Washburn MP & Workman JL (2005) Histone H3 methylation by Set2 directs deacetylation of coding regions by Rpd3S to suppress spurious intragenic transcription. *Cell* **123:** 581-592

Chiu AC, Suzuki HI, Wu X, Mahat DB, Kriz AJ & Sharp PA (2018) Transcriptional pause sites delineate stable nucleosome-associated premature polyadenylation suppressed by U1 snRNP. *Mol Cell* **69**(4): 648-633

Core LJ, Waterfall JJ & Lis JT (2008) Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. *Science* **322**(5909): 1845-1848

Dharmasiri N, Dharmasiri S & Estelle M (2005) The F-box protein TIR1 is an auxin receptor. *Nature* **435:** 441-445

David L, Huber W, Granovskaia M, Toedling J, Palm CJ, Bofkin L, Jones T, Davis RW & Steinmetz LM (2006) A high-resolution map of transcription in the yeast genome. *PNAS* **103**(14): 5320-5325

Davidson L, Kerr A & West S (2012) Co-transcriptional degradation of aberrant pre-mRNA by Xrn2. *EMBO* **31:** 2566-2578

Davidson L, Muniz L & West S (2014) 3' end formation of pre-mRNA and phosphorylation of Ser2 on the RNA polymerase II CTD are reciprocally coupled in human cells. *Genes Dev* **28:** 342-356

Davis CA & Ares M Jr (2006) Accumulation of unstable promoter-associated transcripts upon loss of the nuclear exosome subunit Rrp6p in *Saccharomyces cerevisiae*. *PNAS* **103**(9): 3262-3267

Deaton AM & Bird A (2011) CpG islands and the regulation of transcription. *Genes & Dev* **25:** 1010-1022

Derrien T, Johnson R, Bussotti G, Tanzer A, Djebali S, Tilgner H, Guernec G, Martin D, Merkel A, Knowles DG, Lagarde J, Veeravalli L, Ruan X, Ruan Y, Lassmann T, Carninci P, Brown JB, Lipovich L, Gonzalez JM, Thomas M, Davis CA, Shiekhattar R, Gingeras TR, Hubbard TJ, Notredame C, Harrow J & Guigó R (2012) The GENCODE v7 catalog of human long noncoding RNAs: Analysis of their gene structure, evolution, and expression. *Genome Res* **22:** 1775-1789

van Dijk EL, Chen CL, d'Aubenton-Carafa Y, Gourvennec S, Kwapisz M, Roche V, Bertrand C, Silvain M, Legoix-Né P, Loeillet S, Nicolas A, Thermes C & Morillon A (2011) XUTs are a class of Xrn1-sensitive antisense regulatory non-coding RNA in yeast. *Nature* **475:** 114-117

Djebali S, Davis CA, Merkel A, Dobin A, Lassmann T, Mortazavi A, Tanzer A, Lagarde J, Lin W, Schlesinger F, Xue C, Marinov GK, Khatun J, Williams BA, Zaleski C, Rozowsky J, Röder M, Kokocinski F, Abdelhamid RF, Alioto T, Antoshechkin I, Baer MT, Bar NS, Batut P, Bell K, Bell I, Chakrabortty S, Chen X, Chrast J, Curado J, Derrien T, Drenkow J, Dumais E, Dumais J, Duttagupta R, Falconnet E, Fastuca M, Fejes-Toth K, Ferreira P, Foissac S, Fullwood MJ, Gao H, Gonzalez D, Gordon A, Gunawardena H, Howald C, Jha S, Johnson R, Kapranov P, King B, Kingswood C, Luo OJ, Park E, Persaud K, Preall JB, Ribeca P, Risk B, Robyr D, Sammeth M, Schaffer L, See L, Shahab A, Skancke J, Suzuki AM, 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 SE, Hannon G, Giddings MC, Ruan Y, Wold B, Carninci P, Guigo R & Gingeras TR (2012) Landscape of transcription in human cells. *Nature* **489:** 101-108

Dominski Z, Yang XC & Marzluff WF (2005) The polyadenylation factor CPSF73 is involved in histone–pre-mRNA processing. *Cell* **123:** 37–48

Dominski Z & Marzluff WF (2007) Formation of the 3' end of histone mRNA: Getting closer to the end. *Gene* **396**(2): 373-396

Doudna JA & Charpentier E (2014) The new frontier of genome engineering with CRISPR-Cas9. *Science* **346**(6213): 1258096

Dziembowski A, Lorentzen E, Conti E & Seraphin B (2007) A single subunit, Dis3, is essentially responsible for yeast exosome core activity. *Nat Struct Mol Biol* **14**(1): 15-22

Eaton JD, Davidson L, Bauer DLV, Natsume T, Kanemaki MT & West S (2018) Xrn2 accelerates termination by RNA polymerase II, which is underpinned by CPSF73 activity. *Genes Dev* **32:** 1-13

Eckmann CR, Rammelt C & Wahle E (2011) Control of poly(A) tail length. *WIREs RNA* **2:** 348-361

Elbashir SM, Lendeckel W & Tuschl T (2001) RNA interference is mediated by 21- and 22-nucleotide RNAs. *Genes Dev* **15:** 188-200

The FANTOM Consortium and the RIKEN PMI and CLST (DGT) (2014) A promoter-level mammalian expression atlas. *Nature* **507:**462–470

Flynn RA, Almada AE, Zamudio JR & Sharp PA (2011) Antisense RNA polymerase II divergent transcripts are P-TEFb dependent and substrates for the RNA exosome. *PNAS* **108**(26): 10460-10465

Fong N, Brannan K, Erickson B, Kim H, Cortazar MA, Sheridan RM, Nguyen T, Karp S & Bentley DL (2015) Effects of transcription elongation rate and Xrn2 exonuclease activity on RNA Polymerase II termination suggest widespread kinetic competition. *Mol Cell* **60:** 256-267

Fox, MJ & Mosley AL (2016) Rrp6: Integrated roles in nuclear RNA metabolism and transcription termination. *WIREs RNA* **7:** 91–104 doi: 10.1002/wrna.1317

Fusby B, Kim S, Erickson B, Kim H, Peterson ML & Bentley DL (2016) Coordination of RNA Polymerase II pausing and 3' end processing factor recruitment with alternative polyadenylation. *Mol Cell Biol* **36:** 295-303

Graham AC, Kiss DL, & Andrulis ED (2009) Core exosome-independent roles for Rrp6 in cell cycle progression. *Mol Biol Cell* **20:** 2242-2253

Gray WM, Kepinski S, Rouse D, Leyser O & Estelle M (2001) Auxin regulates SCFTIR1-dependent degradation of AUX/IAA proteins. *Nature* **414:** 271-276

de Groen FLM, Krijgsman O, Tijssen M, Vriend LEM, Ylstra B, Hooijberg E, Meijer GA, Steenbergen RDM & Carvalho B (2014) Gene-dosage dependent overexpression at the 13q amplicon identifies DIS3 as candidate oncogene in colorectal cancer progression. *Genes Chromosomes Cancer* **53:** 339-348

Goff LA & Rinn JL (2015) Linking RNA biology to lncRNAs. *Genome Res* **25:** 1456-1465

Gromak N, West S & Proudfoot NJ (2006) Pause sites promote transcriptional termination of mammalian RNA polymerase II. *Mol Cell Biol* **26**(10): 3986-3996

Gutschner T, Hämmerle M, Eißmann M, Hsu J, Kim Y, Hung G, Revenko A, Arun G, Stentrup M, Groß M, Zörnig M, MacLeod RA, Spector DL & Diederichs S (2013) The non-coding RNA *MALAT1* is a critical regulator of the metastasis phenotype of lung cancer cells. *Cancer Res* **73**(3): 1180-1189

Guttman M, Amit I, Garber M, French C, Lin MF, Feldser D, Huarte M, Zuk O, Carey BW, Cassady JP, Cabili MN, Jaenisch R, Mikkelsen TS, Jacks TJ, Hacohen N, Bernstein BE, Kellis M, Regev A, Rinn JL & Lander ES (2009) Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. *Nature* **458**(7235): 223-227

Hackett PB, Largaespada DA & Cooper LJN (2010) A transposon and transposase system for human application. *Mol Ther* **18**(4): 647-683

Hangauer MJ, Vaughn IW & McManus MT (2013) Pervasive transcription of the human genome produces thousands of previously unidentified long intergenic noncoding RNAs. *PLoS Genet* **9**(6): e1003569

Heidemann M, Hintermair C, Voß K & Eick D (2013) Dynamic phosphorylation patterns of RNA polymerase II CTD during transcription. *Biochim. Biophys. Acta, Gene Regul. Mech* **1829:** 55-62

Heidemann M & Eick D (2012) Tyrosine-1 and threonine-4 phosphorylation marks complete the RNA polymerase II CTD phospho-code. *RNA Biol* **9**(9): 1144-1146

Herzel L, Ottoz DSM, Alpert T & Neugebauer KM (2017) Splicing and transcription touch base: co-transcriptional spliceosome assembly and function. *Nat Rev Mol Cell Biol* **18:** 637-650

Ho CK & Shuman S (1999) Distinct roles for CTD Ser-2 and Ser-5 phosphorylation in the recruitment and allosteric activation of mammalian mRNA capping enzyme. *Mol Cell* **3:** 405-411

Holland AJ, Fachinetti D, Han JS & Cleveland DW (2012). Inducible, reversible system for the rapid and complete degradation of proteins in mammalian cells. *PNAS* E3350–E3357

Hou X, Du Y, Deng Y, Wu J & Cao G (2015) Sleeping Beauty transposon system for genetic etiological research and gene therapy of cancers. *Cancer Biol Ther* **16**(1): 8-16

Hrossova D, Sikorsky T, Potesil D, Bartosovic M, Pasulka J, Zdrahal Z, Stefl R & Vanacova S (2015) RBM7 subunit of the NEXT complex binds U-rich sequences and targets 3'-end extended forms of snRNAs. *Nucleic Acids Res* **43**(8): 4236-4248

Hsin J & Manley JL (2012) The RNA polymerase II CTD coordinates transcription and RNA processing. *Genes Dev* **26:** 2119-2137

Iyer MK, Niknafs YS, Malik R, Singhal U, Sahu A, Hosono Y, Barrette TR, Prensner JR, Evans JR, Zhao S, Poliakov A, Cao X, Dhanasekaran SM, Wu Y, Robinson DR, Beer DG, Feng FY, Iyer HK & Chinnaiyan AM (2015) The landscape of long noncoding RNAs in the human transcriptome. *Nat Genet* **47**(3): 199–208

Jacquier A (2009) The complex eukaryotic transcriptome: unexpected pervasive transcription and novel small RNAs. *Nat Rev Genet* **10**(12): 833-844

Januszyk K, Liu Q & Lima CD (2011). Activities of human RRP6 and structure of the human RRP6 catalytic domain. *RNA* **17:** 1566–1577

Jensen TH, Jacquier A, & Libri D (2013) Dealing with pervasive transcription. *Mol Cell* **52:** 473-484

Jiao X, Chang JH, Kilic T, Tong L & Kiledjian M (2013) A mammalian pre-mRNA 5′-end capping quality control mechanism and an unexpected link of capping to pre-mRNA processing. *Mol Cell* **50**(1): 104-115

Jiao X, Doamekpor SK, Bird JG, Nickels BE, Tong L, Hart RP & Kiledjian M (2017) 5′-end NAD+ cap in human cells promotes RNA decay through DXO-mediated deNADding. *Cell* **168**(6):1015-1027

Jimeno-Gonzalez S, Ceballos-Chavez M & Reyes JC. (2015) A positioned +1 nucleosome enhances promoter-proximal pausing. *Nucleic Acids Res* **43**(6): 3068-3078

Jonkers I & Lis JT (2015) Getting up to speed with transcription elongation by RNA polymerase II. *Nat Rev Mol Cell Biol* **16:** 167-177

Jonkers I, Kwak H & Lis JT (2014) Genome-wide dynamics of Pol II elongation and its interplay with promoter proximal pausing, chromatin, and exons. *Elife* DOI: 10.7554/eLife.02407

Jorjani H, Kehr S, Jedlinski DJ, Gumienny R, Hertel J, Stadler PF, Zavolan M & Gruber AR (2016) An updated human snoRNAome. *Nucleic Acids Res* **44**(11): 5068-5082

Kaida D, Berg MG, Younis I, Kasim M, Singh LN, Wan L & Dreyfuss G (2010) U1 snRNP protects pre-mRNAs from premature cleavage and polyadenylation. *Nature* **468:** 664-668

Kaplan CD, Laprade L & Winston F (2003) Transcription elongation factors repress transcription initiation from cryptic sites. *Science* **301:** 1096-1099

Kapranov P, Cheng J, Dike S, Nix DA, Duttagupta R, Willingham AT, Stadler PF, Hertel J, Hackermüller J, Hofacker IL, Bell I, Cheung E, Drenkow J, Dumais E, Patel S, Helt G, Ganesh M, Ghosh S, Piccolboni A, Sementchenko V, Tammana H & Gingeras TR (2007) RNA maps reveal new RNA classes and a possible function for pervasive transcription. *Science* **316:** 1484-1488

Khalil AM, Guttman M, Huarte M, Garber M, Raj A, Morales DR, Thomas K, Presser A, Bernstein BE, van Oudenaarden A, Regev A, Lander ES & Rinn JL (2009) Many human large intergenic noncoding RNAs associate with chromatin-modifying complexes and affect gene expression. *PNAS* **106**(28): 11667-11672

Kilchert C, Wittmann S & Vasiljeva L (2016) The regulation and functions of the nuclear RNA exosome complex. *Nat Rev Mol Cell* **17:** 227-239

Kim D, Langmead B & Salzberg SL (2015) HISAT: a fast spliced aligner with low memory requirements. *Nat Methods* **12**(4): 357–360.

Kim JH, Lee S, Li L, Park H, Park J, Lee KY, Kim M, Shin BA & Choi S (2011) High cleavage efficiency of a 2A peptide derived from porcine teschovirus-1 in human cell lines, zebrafish and mice. *PLoS One* **6**(4): e18556

Kim M, Krogan NJ, Vasiljeva L, Rando OJ, Nedea E, Greenblatt GF & Buratowski S (2004) The yeast Rat1 exonuclease promotes transcription termination by RNA polymerase II. *Nature* **432**(25): 517-522

Kim NV (2003) RNA interference in functional genomics and medicine. *J Korean Med Sci* **18:** 309-318

Kim T, Hemberg M, Gray JM, Costa AM, Bear DM, Wu J, Harmin DA, Laptewicz M, Barbara-Haley K, Kuersten S, Markenscoff-Papadimitriou E, Kuhl D, Bito H, Worley PF, Kreiman G & Greenberg ME (2010) Widespread transcription at neuronal activity-regulated enhancers. *Nature* **465:** 182-187

Kim T, Hemberg M & Gray JM (2015) Enhancer RNAs: A class of long noncoding RNAs synthesized at enhancers. *Cold Spring Harb Perspect Biol* **7:** a018622

Kiss DL & Andrulis ED (2010) Genome-wide analysis reveals distinct substrate specificities of Rrp6, Dis3, and core exosome subunits. *RNA* **16:** 781-791

Kodzius R, Kojima M, Nishiyori H, Nakamura M, Fukuda S, Tagami M, Sasaki D, Imamura K, Kai C, Harbers M, Hayashizaki M & Carninci P (2006) CAGE: cap analysis of gene expression. *Nat Methods* **3**(3): 211-222

Komor AC, Badran AH & Liu DR (2017) CRISPR-based technologies for the manipulation of eukaryotic genomes. *Cell* **168**(3): 20-36

Kopylova E, Noé L & Touzet H (2012) SortMeRNA: fast and accurate filtering of ribosomal RNAs in metatranscriptomic data. *Bioinformatics* **28**(24): 3211-3217

Kowarz E, Löscher D & Marschalek R (2015) Optimized Sleeping Beauty transposons rapidly generate stable transgenic cell lines. *Biotechnol J* **10:** 647-653

Kreidenweiss A, Hopkins AV & Mordmuller B (2013) 2A and the auxin-based degron system facilitate control of protein levels in *Plasmodium falciparum*. *PLoS One* **8**(11): e78661

Krueger F (2012) Trim Galore! A wrapper tool around Cutadapt and FastQC to consistently apply quality and adapter trimming to FastQ files, with some extra functionality for MspI-digested RRBS-type (Reduced Representation Bisufite-Seq) libraries. Available online at: https://www.bioinformatics.babraham.ac.uk/projects/trim_galore

Kuhn U, Gundel M, Knoth A, Kerwitz Y, Rudel S & Wahle E (2009) Poly(A) tail length is controlled by the nuclear poly(A)-binding protein regulating the interaction between poly(A) polymerase and the cleavage and polyadenylation specificity factor. *J Biol Chem* **284**(34): 22803-22814

Kwak H & Lis JT (2013) Control of transcriptional elongation. *Annu Rev Genet* **47:** 483-508

Labno A, Warkocki Z, Kulinski T, Krawczyk PS, Bijata K, Tomecki R & Dziembowski A (2016) Perlman syndrome nuclease DIS3L2 controls cytoplasmic non-coding RNAs and provides surveillance pathway for maturing snRNAs. *Nucleic Acids Res* **44**(21): 10437-10453

LaCava J, Houseley J, Saveanu C, Petfalski E, Thompson E, Jacquier A & Tollervey D (2005) RNA degradation by the exosome is promoted by a nuclear polyadenylation complex. *Cell* **121:** 713-724

Lawrence M, Gentleman R & Carey V (2009) rtracklayer: an R package for interfacing with genome browsers. *Bioinformatics* **25:** 1841-1842 doi: 10.1093/bioinformatics/btp328

Lawrence M, Huber W, Pagès H, Aboyoun P, Carlson M, Gentleman R, Morgan M & Carey V (2013) Software for computing and annotating genomic ranges. *PLoS Comput Biol* **9:** doi: 10.1371/journal.pcbi.1003118

Lewis JD & Izaurralde E (1997) The role of the cap structure in RNA processing and nuclear export. *Eur J Biochem* **247:** 461-469

Lizio M, Harshbarger J, Shimoji H, Severin J, Kasukawa T, Sahin S, Abugessaisa I, Fukuda S, Hori F, Ishikawa-Kato S, Mungall CJ, Arner E, Baillie KJ, Bertin N, Bono H, de Hoon M, Diehl AD, Dimont E, Freeman TC, Fujieda K, Hide W, Kaliyaperumal R, Katayama T, Lassmann T, Meehan TF, Nishikata K, Ono H, Rehli M, Sandelin A, Schultes EA, 't Hoen PAC, Tatum Z, Thompson M, Toyoda T, Wright DW, Daub CO, Itoh M, Carninci P, Hayashizaki Y, Forrest ARR & Kawaji H (2015) Gateways to the FANTOM5 promoter level mammalian expression atlas. *Genome Biol* **16**(22): DOI 10.1186/s13059-014-0560-6

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R & 1000 Genome Project Data Processing Subgroup (2009) The Sequence alignment/map (SAM) format and SAMtools. *Bioinformatics* **25:** 2078-2079

Li H (2011) A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* **27**(21): 2987-93

Liao Y, Smyth GK & Shi W (2013) The Subread aligner: fast, accurate and scalable read mapping by seed-and-vote. *Nucleic Acids Res* **41**(10): e108

Liao Y, Smyth GK & Shi W (2014) featureCounts: an efficient general-purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30**(7): 923-930

Love MI, Huber W & Anders S (2014) Moderated estimation of fold change and dispersion for RNA-Seq data with DESeq2. *Genome Biol* **15**(550): DOI 10.1186/s13059-014-0550-8

Lubas M, Christensen MS, Kristiansen MS, Domanski M, Falkenby LG, Lykke-Andersen S, Andersen JS, Dziembowski A & Jensen1 TH (2011) Interaction profiling identifies the human nuclear exosome targeting complex. *Mol Cell* **43:** 624-637

Lykke-Andersen S, Tomecki R, Jensen TH & Dziembowski A (2011) The eukaryotic RNA exosome: Same scaffold but variable catalytic subunits. *RNA Biol* **8**(1): 61-66

Maes L, Blockmans D, Verschueren P, Westhovens R, Op De Beéck K, Vermeersch P, Van den Bergh K, Burlingame RW, Mahler M & Bossuyt X (2010) Anti-PM/Scl-100 and anti-RNA-polymerase III antibodies in scleroderma. *Clinica Chimica Acta* **411:** 965-971

Mahler M & Raijmakers R (2007) Novel aspects of autoantibodies to the PM/Scl complex: Clinical, genetic and diagnostic insights. *Autoimmun Rev* **6:** 432-437

Mali P, Esvelt KM & Church GM (2013) Cas9 as a versatile tool for engineering biology. *Nat Methods* **10**(10): 957-963

Martin M (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet J* **17**(1): 10-12

Mates L, Chuah MK, Belay E, Jerchow B, Manoj N, Acosta-Sanchez A, Grzela DP, Schmitt A, Becker K, Matrai J, Ma L, Samara-Kuko E, Gysemans C, Pryputniewicz D, Miskey C, Fletcher B, VandenDriessche T, Ivics Z & Izsvak Z (2009) Molecular evolution of a novel hyperactive Sleeping Beauty transposase enables robust stable gene transfer in vertebrates. *Nat Genet* **41**(6): 753-761

Meola N, Domanski M, Karadoulama E, Chen Y, Gentil C, Pultz D, Vitting-Seerup K, Lykke-Andersen S, Andersen JS, Sandelin A & Jensen TH (2016) Identification of a nuclear exosome decay pathway for processed transcripts *Mol Cell* **64:** 520-533

Mercer TR & Mattick, JS (2013) Structure and function of long noncoding RNAs in epigenetic regulation. *Nat Struct Mol Biol* **20**(3): 300-307

Mitchell P, Petfalski E, Shevchenko A, Mann M & Tollervey D (1997) The exosome: A conserved eukaryotic RNA processing complex containing multiple 3'→5' exoribonucleases. *Cell* **91:** 457-466

Mitchell P (2014) Exosome substrate targeting: the long and short of it. *Biochem Soc Trans* **42:** 1129-1134

Morawska & Ulrich HD (2013) An expanded tool kit for the auxin-inducible degron system in budding yeast. *Yeast* **30:** 341-351

Morris MR, Astuti D & Maher ER (2013) Perlman syndrome: Overgrowth, Wilms tumor predisposition and DIS3L2. *Am J Med Genet Part C Semin Med Genet* **163C:** 106-113

Nagalakshmi U, Wang Z, Waern K, Shou C, Raha D, Gerstein M & Snyder M (2008) The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science* **320**(5881): 1344-1349

Natsume T, Kiyomitsu T, Saga Y & Kanemaki MT (2016). Rapid protein depletion in human cells by auxin-inducible degron tagging with short homology donors. *Cell Rep* **15:** 210–218.

Neil H, Malabat C, d'Aubenton-Carafa Y, Xu Z, Steinmetz LM & Jacquier A (2009) Widespread bidirectional promoters are the major source of cryptic transcripts in yeast. *Nature* **457:** 1038-1042

Neph S, Kuehn SM, Reynolds AP, Haugen E, Thurman RE, Johnson AK, Rynes E, Maurano MT, Vierstra J, Thomas S, Sandstrom R, Humbert R & Stamatoyannopoulos JA (2012) BEDOPS: high-performance genomic feature operations. *Bioinformatics* **28**(14): 1919-1920 doi:10.1093/bioinformatics/bts277

Ng D, Toure O, Wei M, Arthur DC, Abbasi F, Fontaine L, Marti GE, Fraumeni JF Jr., Goldin LR, Caporaso N & Toro JR (2007) Identification of a novel chromosome region, 13q21.33-q22.2, for susceptibility genes in familial chronic lymphocytic leukemia. *Blood* **109**(3): 916-925

Nishimura K, Fukagawa T, Takisawa H, Kakimoto T & Kanemaki M (2009) An auxin-based degron system for the rapid depletion of proteins in non-plant cells. *Nat Methods* **6**(12): 917-922

Norrander J, Kempe T & Messing J (1983) Construction of improved M13 vectors using oligodeoxynucleotide-directed mutagenesis. *Gene* **26**(1): 101-106

Nojima T, Gomes T, Grosso ARF, Kimura H, Dye MJ, Dhir S, Carmo-Fonseca M & Proudfoot NJ (2015) Mammalian NET-Seq Reveals Genome-wide Nascent Transcription Coupled to RNA Processing. *Cell* **161:** 526-540

Ntini E, Järvelin AI, Bornholdt J, Chen Y, Boyd M, Jørgensen M, Andersson R, Hoof I, Schein A, Andersen PR, Andersen PK, Preker P, Valen E, Zhao X, Pelechano V, Steinmetz LM, Sandelin A & Jensen TH (2013) Polyadenylation site–induced decay of upstream transcripts enforces promoter directionality. *Nat Struct Mol Biol* **20**(8): 923-928

Ogami K, Chen Y & Manley JL (2018) RNA surveillance by the nuclear RNA exosome: mechanisms and significance. *Non-coding RNA* **4**(8): doi:10.3390/ncrna4010008

O'Reilly D, Kuznetsova OV, Laitem C, Zaborowska J, Dienstbier M & Murphy S (2014) Human snRNA genes use polyadenylation factors to promote efficient transcription termination. *Nucleic Acids Res* **42**(1): 264-275

Osheim YN, Sikes ML & Beyer AL (2002) EM visualization of Pol II genes in Drosophila: most genes terminate without prior 3′ end cleavage of nascent transcripts. *Chromosoma* **111:** 1-12

Pertea M, Kim D, Pertea G, Leek JT & Salzberg SL (2016) Transcript-level expression analysis of RNA-Seq experiments with HISAT, StringTie, and Ballgown. *Nat Protoc* **11**(9): 1650–1667

Pontier DB & Gribnau J (2011) *XIST* regulation and function eXplored. *Hum Genet* **130:** 223-236

Preker P, Almvig K, Christensen MS, Valen E, Mapendano CK, Sandelin A & Jensen TH (2011) PROMoter uPstream Transcripts share characteristics with mRNAs and are produced upstream of all three major types of mammalian promoters. *Nucleic Acids Res* **39**(16): 7179-7193

Preker P, Nielsen J, Kammler S, Lykke-Andersen S, Christensen MS, Mapendano CK, Schierup MH & Jensen TH (2008) RNA exosome depletion reveals transcription upstream of active human promoters. *Science* **322:** 1851-1854

Proudfoot NJ (2011) Ending the message: poly(A) signals then and now. *Genes Dev* **25:** 1770-1782

Quaresma AJC, Bugai A & Barboric M (2016) Cracking the control of RNA polymerase II elongation by *7SK* snRNP and P-TEFb. *Nucleic Acids Res* **44**(16): 7527-7539

Qiu S, Adema CM & Lane T (2005) A computational study of off-target effects of RNA interference. *Nucleic Acids Res* **33**(6): 1834-1847

Quinlan AR & Hall IM (2010) BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26:** 841-842

Ramirez F, Dundar F, Diehl S, Gruning BA & Manke T (2014) deepTools: a flexible platform for exploring deep-sequencing data. *Nucleic Acids Res* **42:** W187-W191

Ramírez F, Ryan DP, Grüning B, Bhardwaj V, Kilpert F, Richter AS, Heyne S, Dündar F & Manke T (2016) deepTools2: a next generation web server for deep-sequencing data analysis. *Nucleic Acids Res* **44:** W160–W165

Ran FA, Hsu PD, Wright J, Agarwala V, Scott DA & Zhang F (2013) Genome engineering using the CRISPR-Cas9 system. *Nat Protoc* **8**(11): 2281-2308

Reichow SL, Hamma T, Ferre-D'Amare AR & Varani G (2007) The structure and function of small nucleolar ribonucleoproteins. *Nucleic Acids Res* **35**(5): 1452-1464

Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G & Mesirov JP (2011) Integrative Genomics Viewer. *Nat Biotechnol* **29:** 24–26

Robinson SR, Oliver AW, Chevassut TJ & Newbury SF (2015) The 3' to 5' Exoribonuclease DIS3: From Structure and Mechanisms to Biological Functions and Role in Human Disease. *Biomolecules* **5:** 1515-1539

Robinson SR, Viegas SC, Matos RG, Domingues S, Bedir M, Stewart HJS, Chevassut TJ, Oliver AW, Arraiano CM & Newbury SF (2018) DIS3 isoforms vary in their endoribonuclease activity and are differentially expressed within haematological cancers. *Biochem J* 475**:** 2091-2105

Sainsbury S, Bernecky C & Cramer P (2015) Structural basis of transcription initiation by RNA polymerase II. *Nat Rev Mol Cell Biol* **16:** 129-143

Saldi T, Cortazar MA, Sheridan RM & Bentley DL (2016) Coupling of RNA polymerase II transcription elongation with pre- mRNA splicing. *J Mol Biol* **428**(12): 2623-2635

Sallee M, Dou L, Cerni C, Poitevin S, Brunet P & Burtey S (2014) The aryl hydrocarbon receptor-activating effect of uremic toxins from tryptophan metabolism: A new concept to understand cardiovascular complications of chronic kidney disease. *Toxins* **6:** 934-949

Schindelin J, Arganda-Carreras I, Frise E, Kaynig V, Longair M, Pietzsch T, Preibisch S, Rueden C, Saalfeld S, Schmid B, Tinevez JY, White DJ, Hartenstein V, Eliceiri K, Tomancak P & Cardona A (2012) Fiji: an open-source platform for biological-image analysis. *Nat methods* **9**(7): 676-682

Seila AC, Calabrese JM, Levine SS, Yeo GW, Rahl PB, Flynn RA, Young RA & Sharp PA (2008) Divergent transcription from active promoters. *Science* **322**(5909): 1849-1851

Sen GL & Blau HM (2006) A brief history of RNAi: the silence of the genes. *The FASEB Journal* **20:** 1293-1299

Shi Y, Campigli Di Giammartino D, Taylor D, Sarkeshik A, Rice WJ, Yates III JR, Frank J & Manley JL (2009) Molecular architecture of the human pre-mRNA 3' processing complex. *Mol Cell* **33**(3): 365-376

Skipper KA, Andersen PR, Sharma N & Mikkelsen JG (2013) DNA transposon-based gene vehicles – scenes from an evolutionary drive. *J Biomed Sci* **20**(92): http://www.jbiomedsci.com/content/20/1/92

Smith I, Greenside PG, Natoli T, Lahr DL, Wadden D, Tirosh I, Narayan R, Root DE, Golub TR, Subramanian A & Doench JG (2017) Evaluation of RNAi and CRISPR technologies by large-scale gene expression profiling in the Connectivity Map. *PLoS Biol* **15**(11): e2003213

St Laurent G, Wahlestedt C & Kapranov P (2015) The Landscape of long non-coding RNA classification. *Trends Genet* **31**(5): 239-251

Steinmetz EJ, Warren CL, Kuehner JN, Panbehi B, Ansari AZ & Brow DA (2006) Genome-wide distribution of yeast RNA polymerase II and its control by Sen1 helicase. *Mol Cell* **24:** 735-746

Szczepinska T, Kalisiak K, Tomecki R, Labno A, Borowski LS, Kulinski TM, Adamska D, Kosinska J & Dziembowski A (2015) DIS3 shapes the RNA polymerase II transcriptome in humans by degrading a variety of unwanted transcripts. *Genome Res* **25:** 1622-1633

Tanny JC (2014) Chromatin modification by the RNA Polymerase II elongation complex. *Transcription* **5**(5): e988093

Thiebaut M, Kisseleva-Romanova E, Rougemaille M, Boulay J & Libri D (2006) Transcription termination and nuclear degradation of cryptic unstable transcripts: A role for the Nrd1-Nab3 pathway in genome surveillance. *Mol Cell* **23:** 853-864

Thorvaldsdóttir H, Robinson JT & Mesirov JP (2013) Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief Bioinform* **14:** 178-192

Towler BP, Jones CI, Harper KL, Waldron JA & Newbury SF (2016) A novel role for the 30-50 exoribonuclease Dis3L2 in controlling cell proliferation and tissue growth. *RNA Biol* **13**(12): 1286-1299

Ulitsky I & Bartel DP (2013) lincRNAs: genomics, evolution, and mechanisms. *Cell* **154**(1): 26-46

Vanacova S, Wolf J, Martin G, Blank, D, Dettwiler S, Friedlein A, Langen H, Keith G & Keller W (2005) A new yeast poly(A) polymerase complex involved in RNA quality control. *PLoS Biol* **3**(6): 0986-0997

Vilborg A, Passarelli MC, Yario TA, Tycowski KT & Steitz JA (2015) Widespread inducible transcription downstream of human genes. *Mol Cell* **59**(3): 449-461

Wasmuth EV & Lima C (2017) The Rrp6 C-terminal domain binds RNA and activates the nuclear RNA exosome. *Nucleic Acids Res* **45**(2): 846-860

West S, Gromak N & Proudfoot NJ (2004) Human 5′ → 3′ exonuclease Xm2 promotes transcription termination at co-transcriptional cleavage sites. *Nature* **432:** 522-525

Wickham H (2009) ggplot2: Elegant graphics for data analysis. *J Stat Soft* **35**(1)

Wyers F, Rougemaille M, Badis G, Rousselle JC, Dufour ME, Boulay J, Régnault B, Devaux F, Namane A, Séraphin B, Libri D & Jacquier A (2005) Cryptic pol II transcripts are degraded by a nuclear quality control pathway involving a new poly(A) polymerase. *Cell* **121:** 725-737

Xiang K, Tong L & Manley JL (2014) Delineating the structural blueprint of the pre-mRNA 3'-end processing machinery. *Mol Cell Biol* **34**(11) 1894-1910

Xie C, Yuan J, Li H, Li M, Zhao G, Bu D, Zhu W, Wu W, Chen R & Zhao Y (2014) NONCODEv4: exploring the world of long non-coding RNA genes. *Nucleic Acids Res* **42:** D98-D103

Xu Z, Wei W, Gagneur J, Perocchi F, Clauder-Münster S, Camblong J, Guffanti E, Stutz F, Huber W & Steinmetz LM (2009) Bidirectional promoters generate pervasive transcription in yeast. *Nature* **457:** 1033-1037

Young RS, Kumar Y, Bickmore WA & Taylor MS (2017) Bidirectional transcription initiation marks accessible chromatin and is not specific to enhancers. *Genome Biol* **18**(242)

Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, Nusbaum C, Myers RM, Brown M, Li W & Liu XS (2008) Model-based analysis of ChIP-seq (MACS). *Genome Biol* **9:** R137

Supplementary Figures

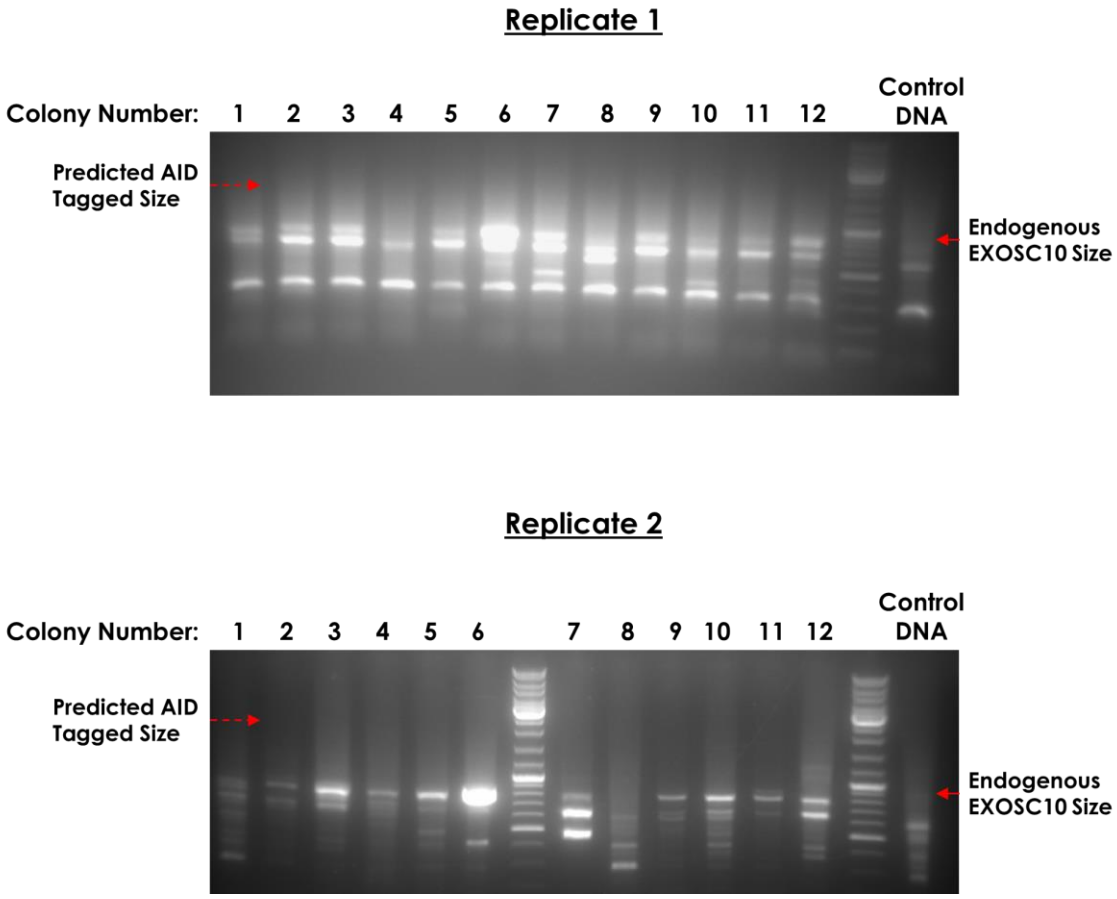


Figure S1: Nested PCR of genomic DNA isolated from 12 CRISPR/Cas9 edited HCT116 EXOSC10-AID colonies following antibiotic resistance selection (10-14 days). Both replicate PCRs were achieved by first amplifying a larger product from the genomic DNA surrounding the poly(A) site of the EXOSC10. This product was then diluted and used as a template to generate the PCR amplicons above. In both instances Q5 polymerase was used, the 2nd replicate was performed using a higher annealing temperature in an attempt to eliminate the numerous small non-specific PCR products amplified. Negative colony PCR size was expected to be ~900 nt in length whereas AID modified EXOSC10 products should be 2000 nt for neomycin and 2300 nt for hygromycin respectively.

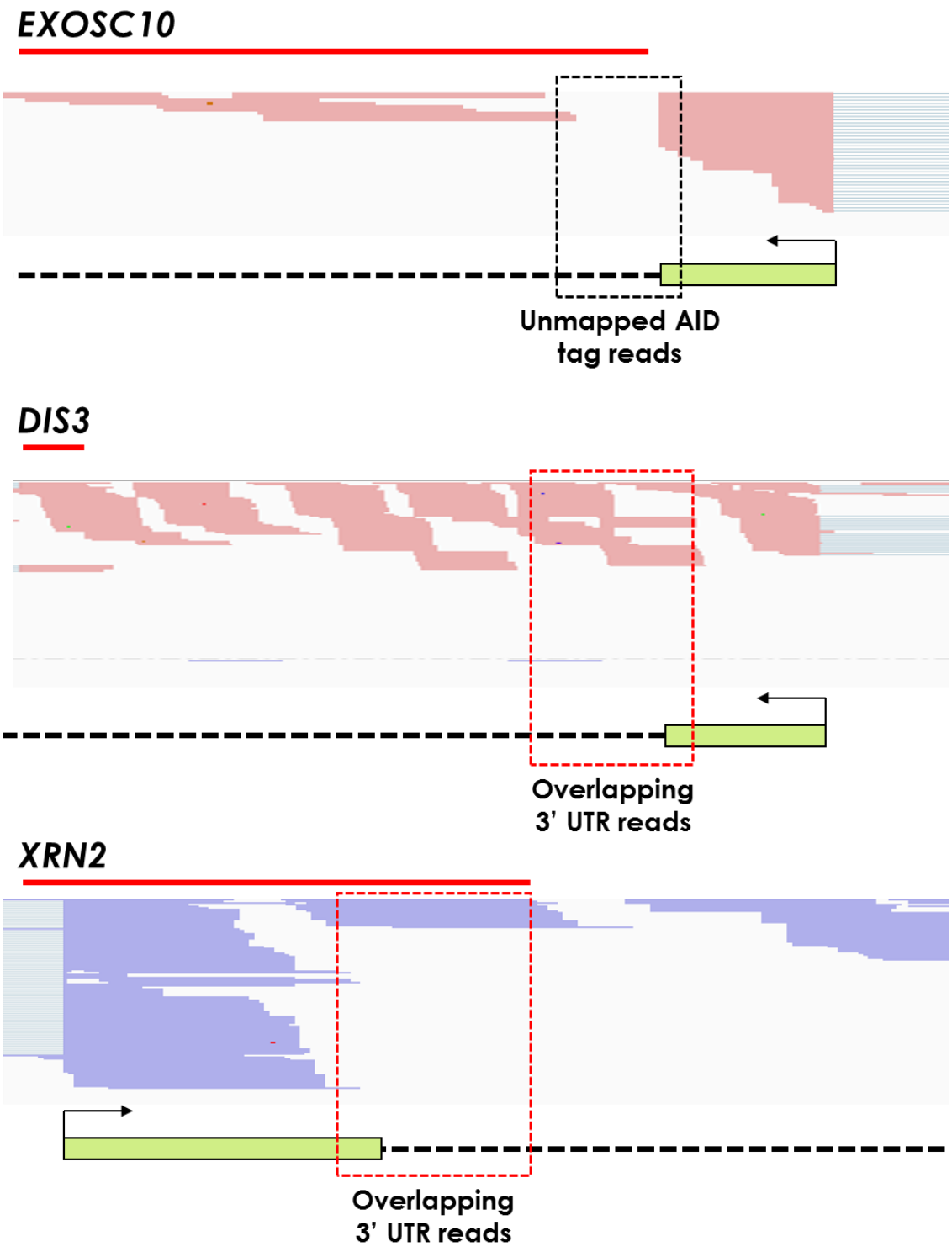
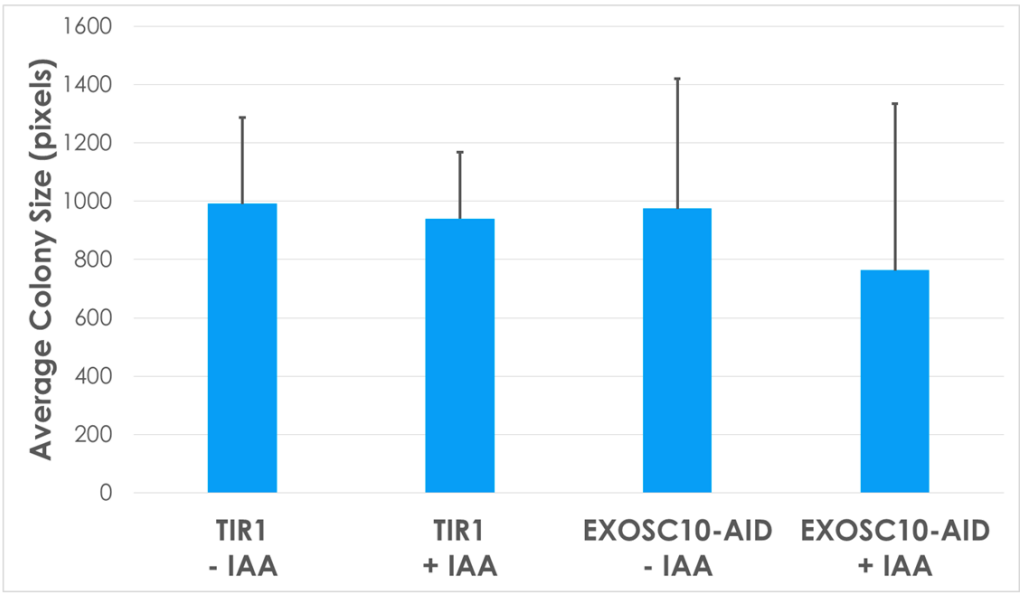


Figure S2: RNA-Seq aligned reads from EXOSC10-AID modified HCT116 cell lines. Reads are shown to overlap the final exon and 3'UTR region (dashed line) of 3 genes; EXOSC10, DIS3 and XRN2. The EXOSC10 gene has been modified with the 3' AID tag which cannot map to the reference genome. Scale bars relative to 100 nt are shown in red.

A.



B.

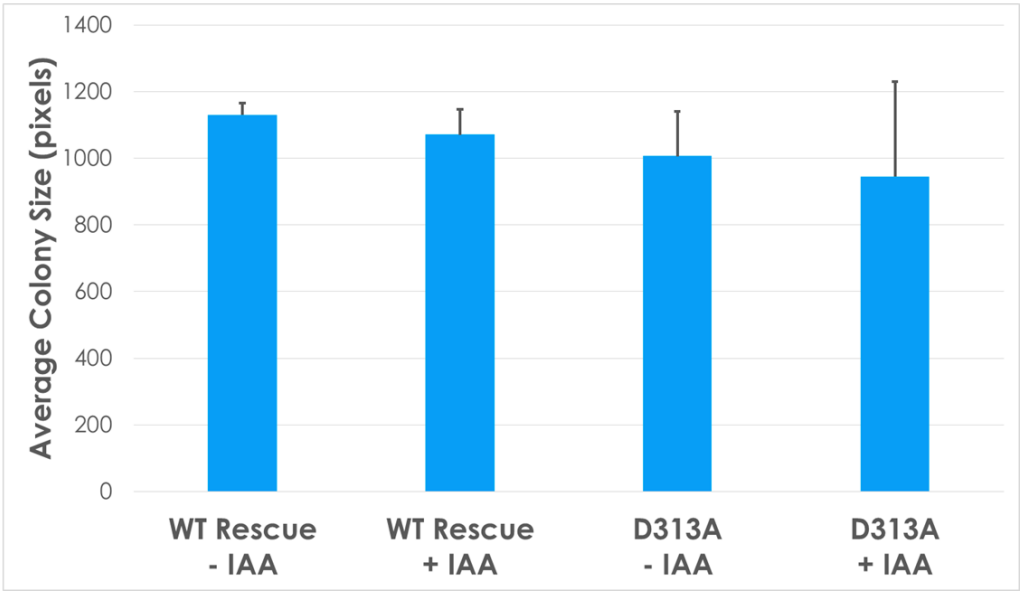


Figure S3: Average colony size of 3 biological replicate colony formation assays represented in **Figure 3.7** and **Figure 3.10**(**A**). Colony pixel density of (**A**) parent TIR1 and EXOSC10-AID cells and (**B**) WT and D313A EXOSC10 rescue cell lines was calculated using ImageJ. Error bars = standard deviation.

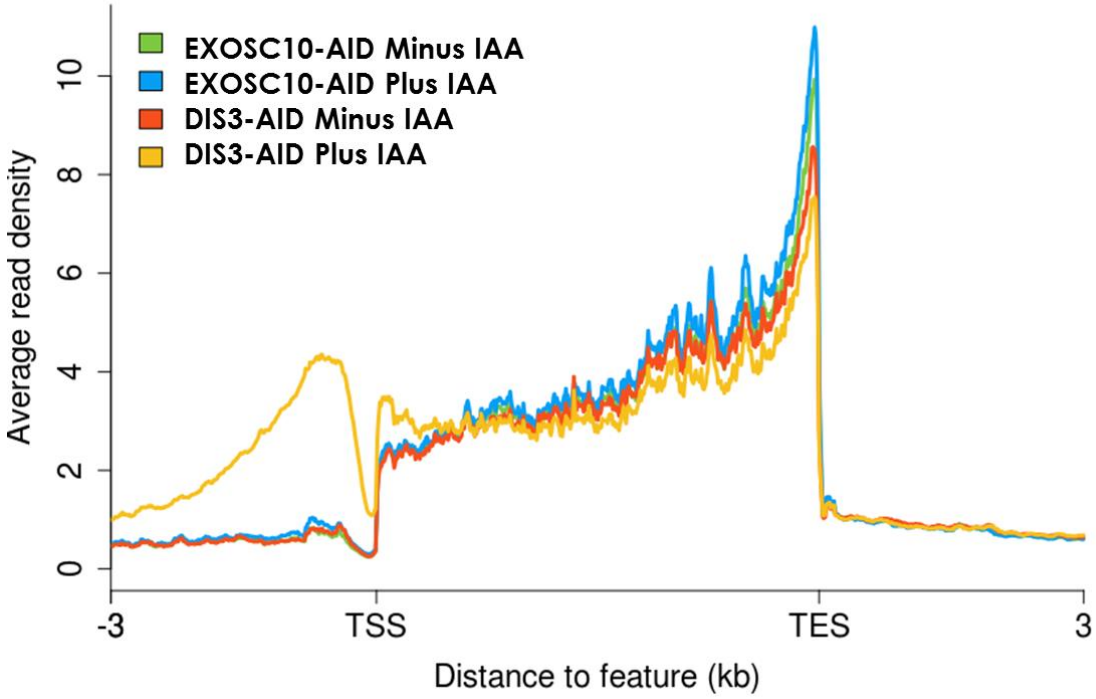


Figure S4: Additional replicate metagene read coverage profile comparison of non-overlapping expressed genes with a 3 kb inclusion window flanking the TSS and TES (n = 4701). The gene body was scaled to 5 kb.

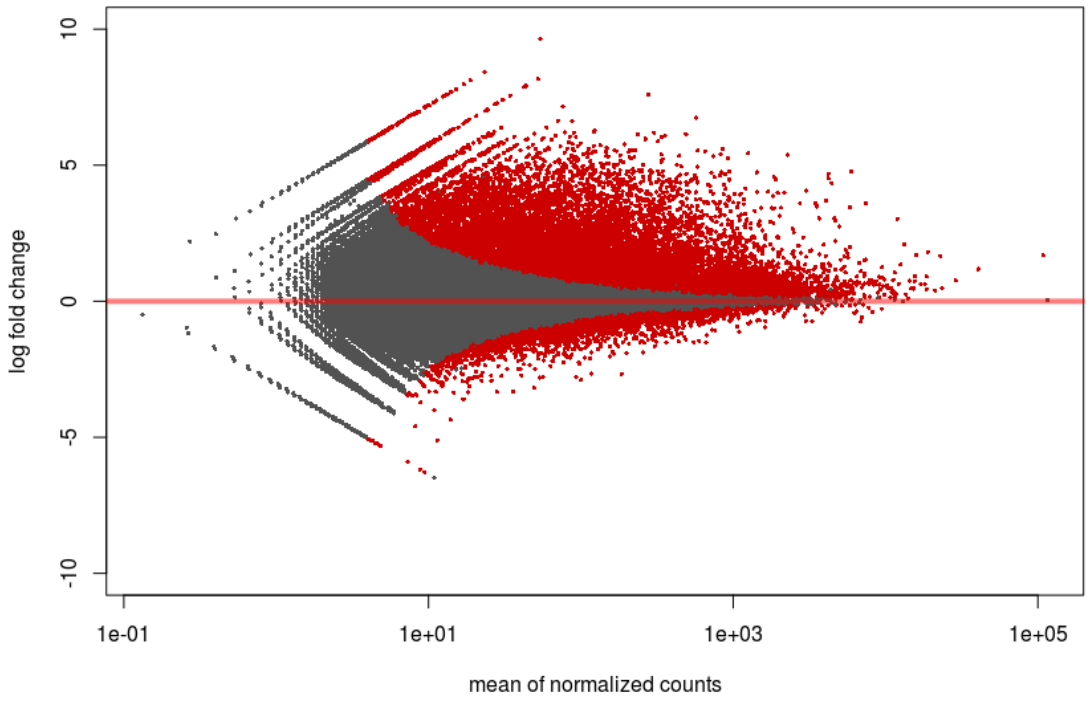


Figure S5: MA plot representation of differential expression of all synthetic introns (n = 280,045); significantly altered expression of 9252 introns were identified (fold ≥ 2-fold, padj < 0.05) from the 16,093 hits (red points).

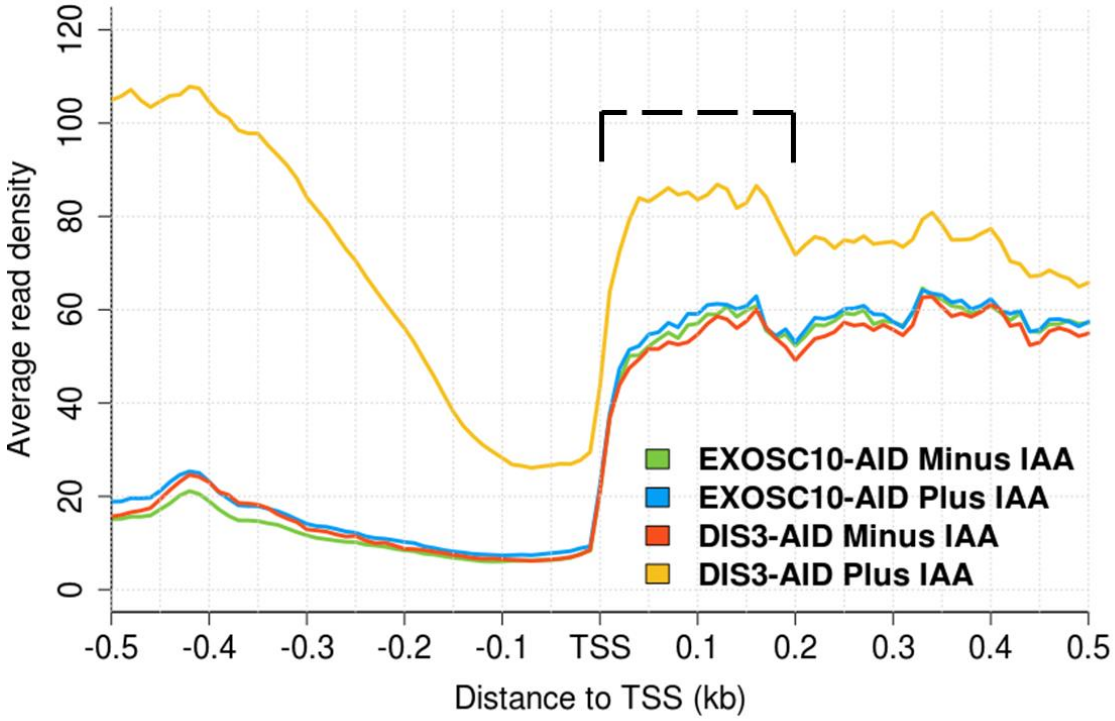


Figure S6: Additional biological replicate metagene read coverage profile comparison centred on the TSS of 4701 non-overlapping genes.

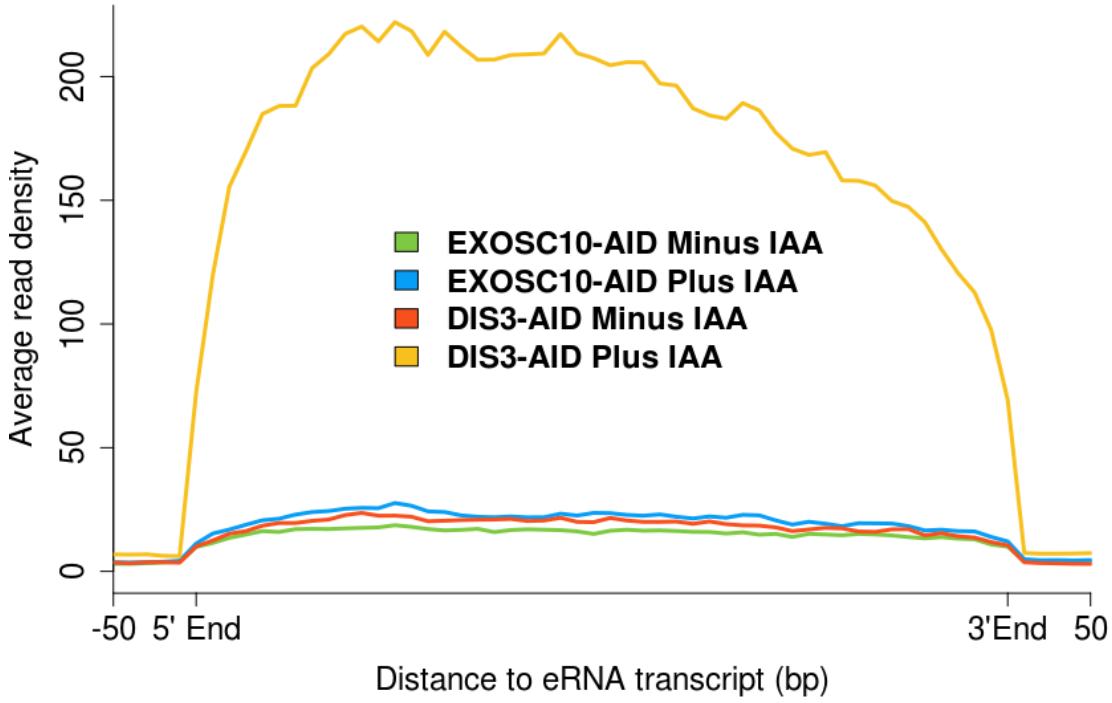


Figure S7: Second biological replicate metagene profile of eRNA transcription in DIS3-AID and EXOSC10-AID cell lines (n = 960).

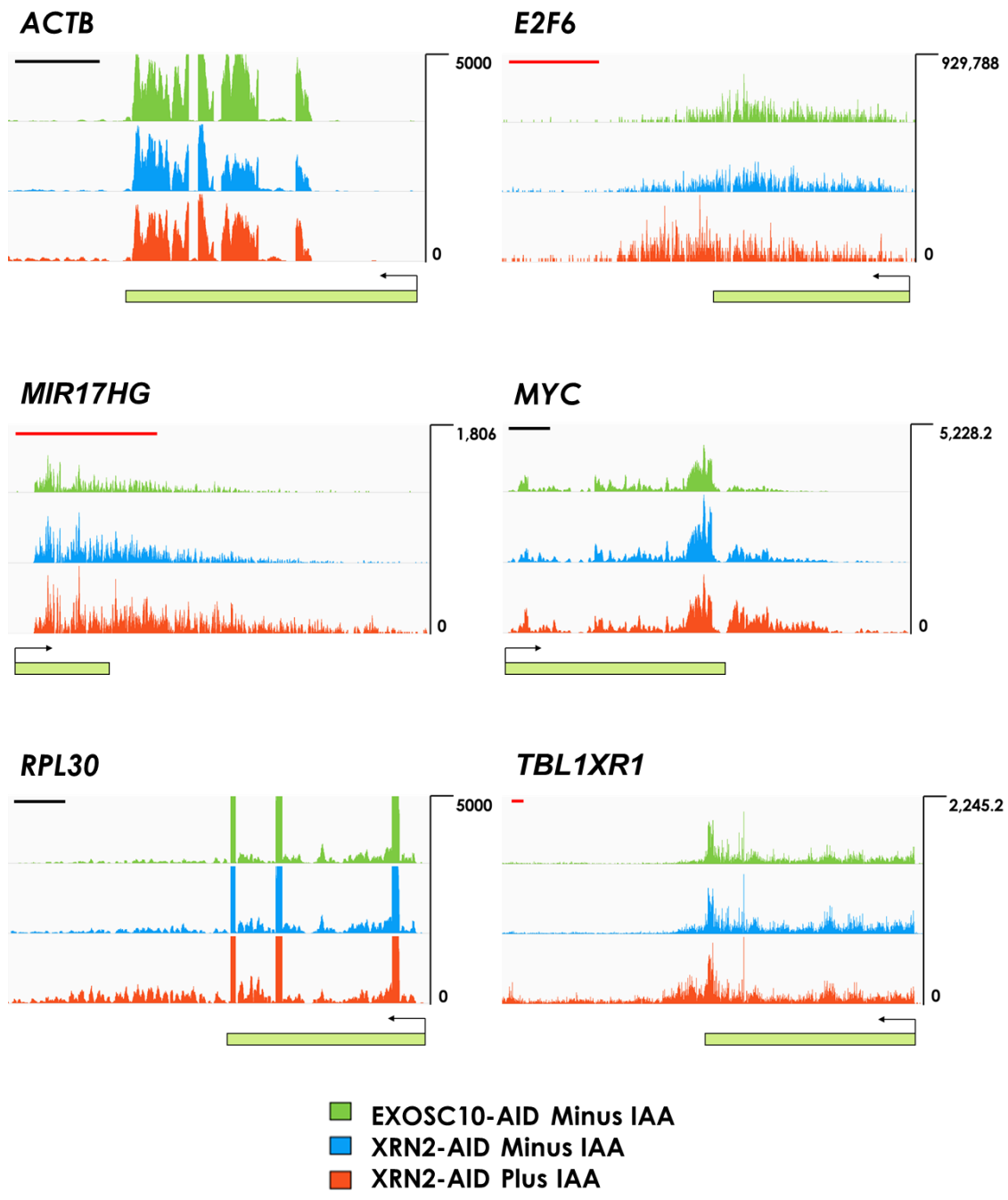


Figure S8: Additional biological replicate of RPKM normalised read coverage in Xrn2 depleted cells. Black scale bars = 1 kb, apply to every coverage track within *ACTB*, *MYC* and *RPL30* gene images respectively. Likewise, red scale bars = 10 kb were applied to each track in *E2F6*, *MIR17HG* and *TBLXR1* genes

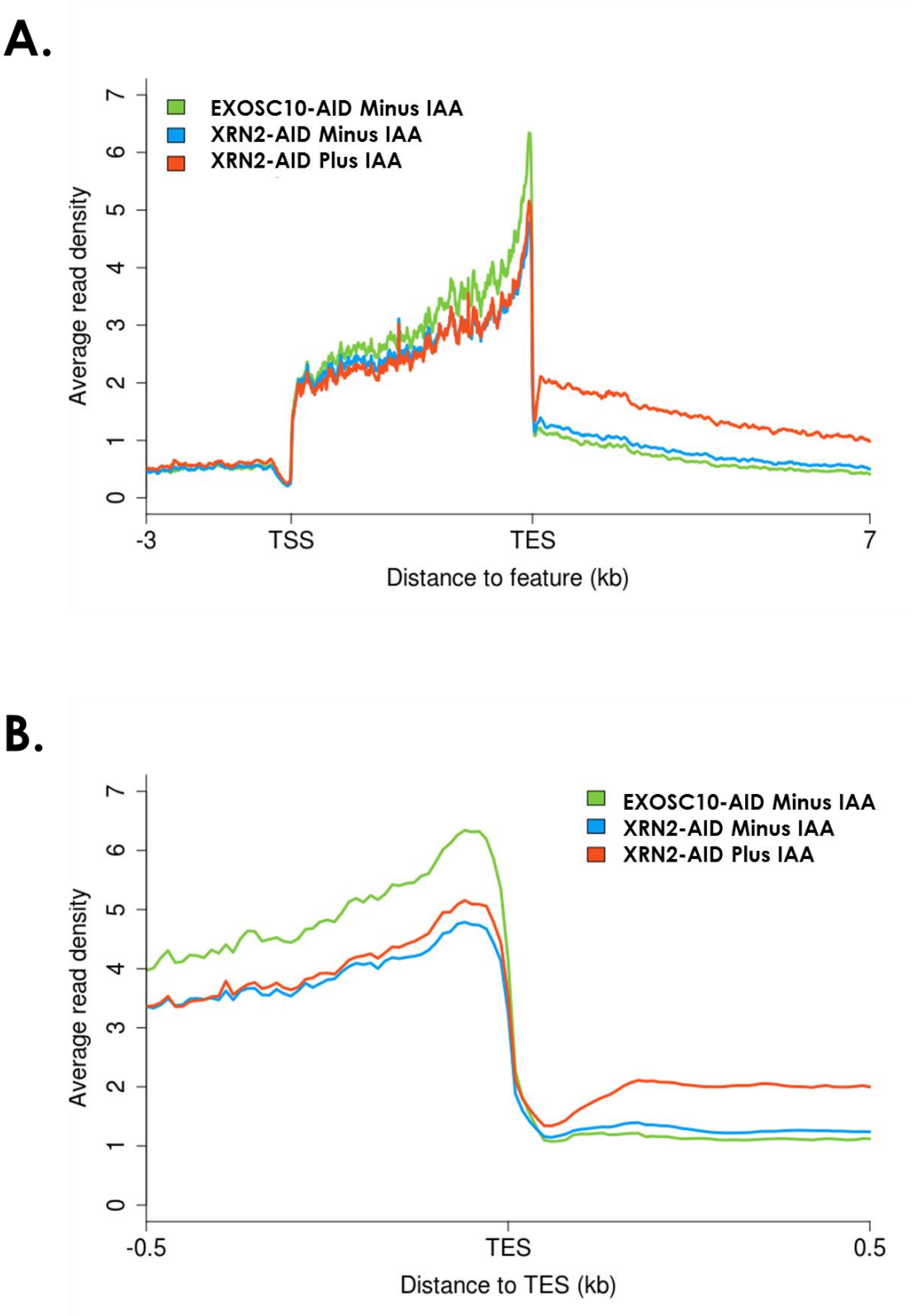


Figure S9: (**A**) Replicate 2 scaled metagene coverage plot of 4701 non-overlapping genes including an inclusion window of 3 kb upstream of the TSS and 7 kb downstream of the TES. Gene body scaled to 5 kb. (**B**) Enhanced image of the read coverage in (**A**) over the TES.

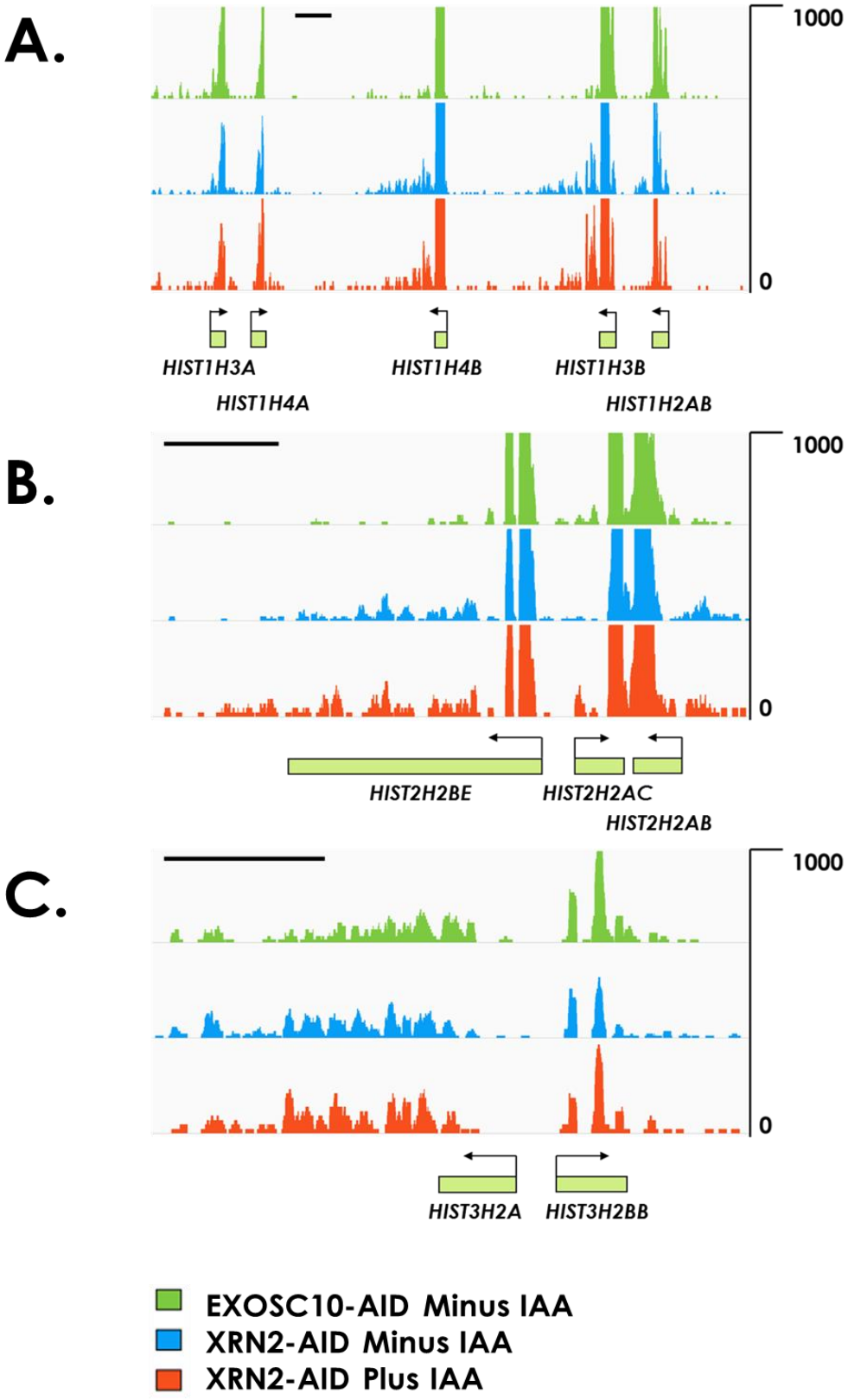


Figure S10: Additional biological replicate of coverage tracks over histone genes in cluster 1 (**A**), 2 (**B**) and 3 (**C**) respectively. Scale bars = 1 kb apply to all 3 coverage tracks within each gene image respectively

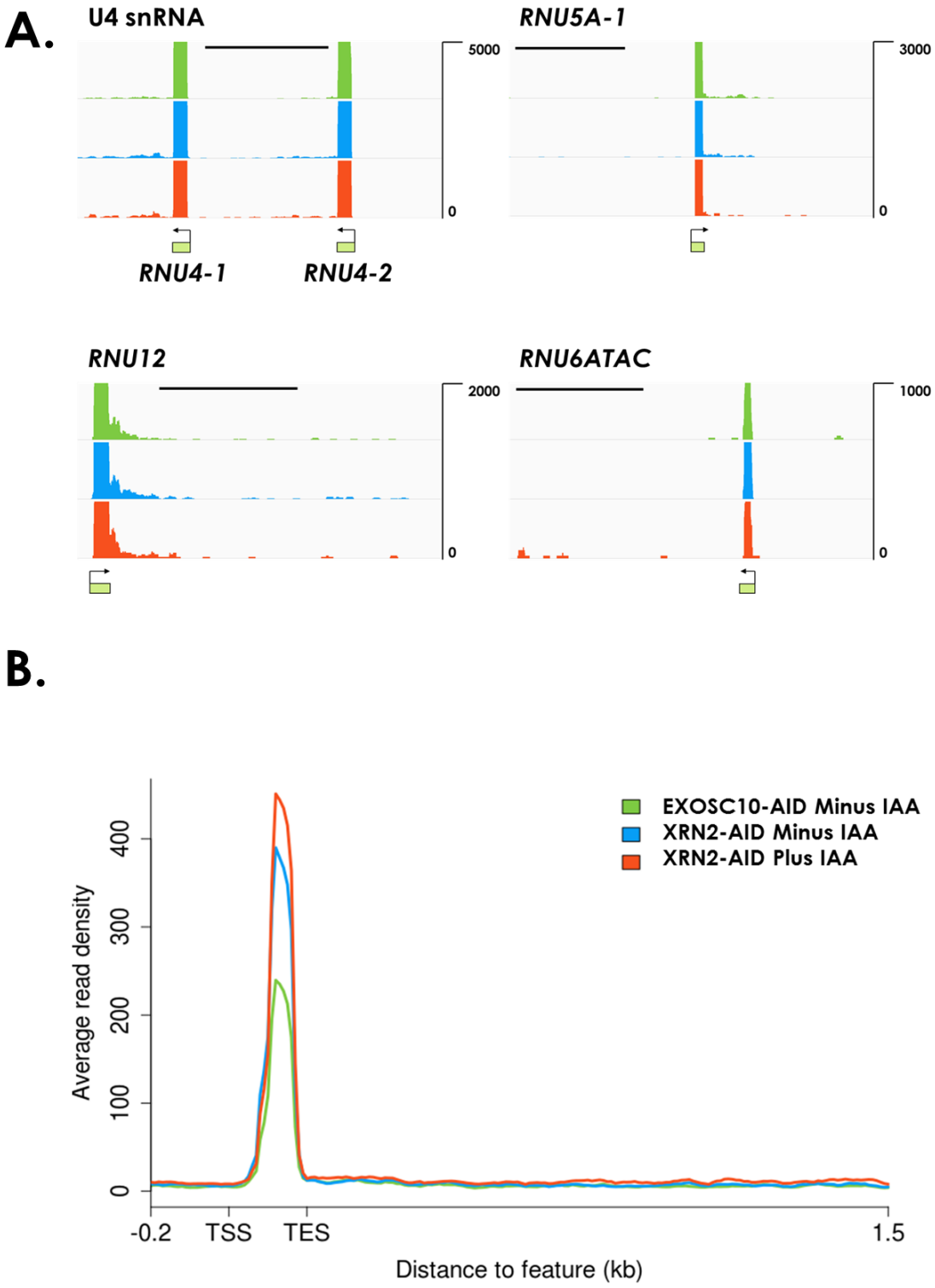


Figure S11: A second biological replicate analysis of (**A**) RPKM normalised coverage tracks of read density in snRNA genes. Scale bars = 1 kb apply to a 3 coverage tracks within each gene image respectively. (**B**) the transcription profile of non-overlapping snRNA genes with a 1.5 kb extended region flanking the TES (n = 707).

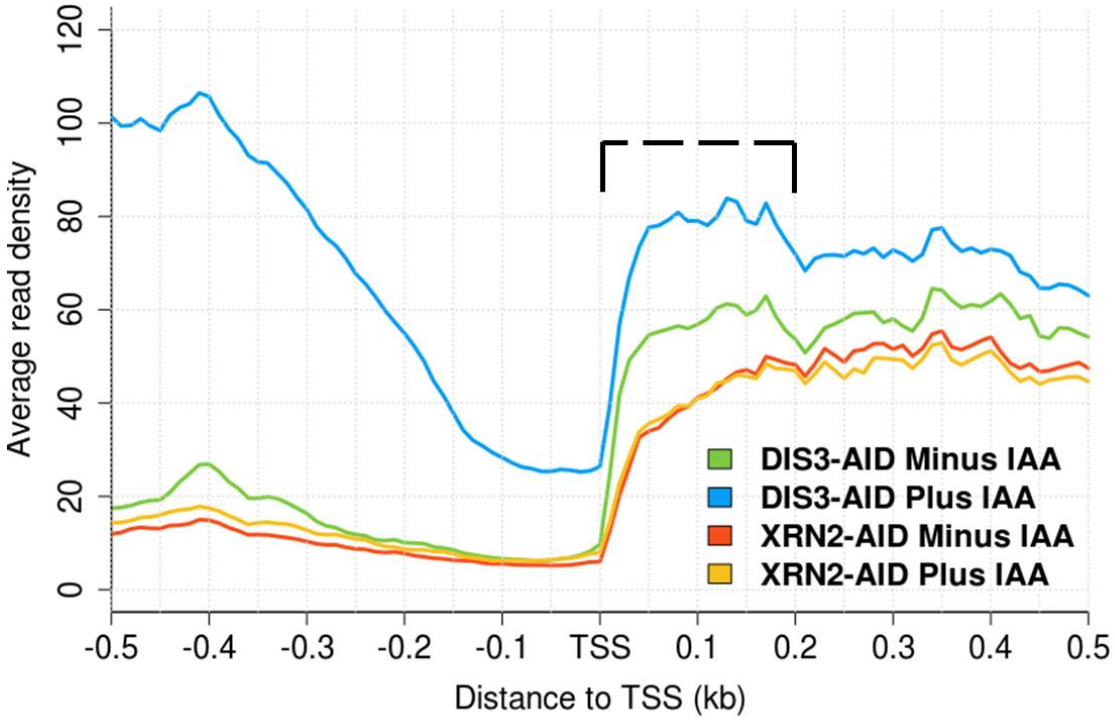


Figure S12: Additional biological replicate metagene read coverage profile comparison centred on the TSS of 4701 non-overlapping genes.

Appendix

The Publication

Xrn2 accelerates termination by RNA polymerase II, which is underpinned by CPSF73 activity

Joshua D. Eaton,[1,6] Lee Davidson,[1,2,6] David L.V. Bauer,[3] Toyoaki Natsume,[4,5] Masato T. Kanemaki,[4,5] and Steven West[1]

[1]The Living Systems Institute, University of Exeter, Exeter EX4 4QD, United Kingdom; [2]Department of Molecular Biology and Biotechnology, University of Sheffield, Sheffield S10 2TN, United Kingdom; [3]Sir William Dunn School of Pathology, University of Oxford, Oxford OX1 3RE, United Kingdom; [4]Division of Molecular Cell Engineering, National Institute of Genetics, Research Organization of Information and Systems (ROIS), Mishima, Shizuoka 411-8540, Japan; [5]Department of Genetics, Graduate University for Advanced Studies (SOKENDAI), Mishima, Shizuoka 411-8540, Japan

Abstract

Termination is a ubiquitous phase in every transcription cycle but is incompletely understood and a subject of debate. We used gene editing as a new approach to address its mechanism through engineered conditional depletion of the 5′ → 3′ exonuclease Xrn2 or the polyadenylation signal (PAS) endonuclease CPSF73 (cleavage and polyadenylation specificity factor 73). The ability to rapidly control Xrn2 reveals a clear and general role for it in cotranscriptional degradation of 3′ flanking region RNA and transcriptional termination. This defect is characterized genome-wide at high resolution using mammalian native elongating transcript sequencing (mNET-seq). An Xrn2 effect on termination requires prior RNA cleavage, and we provide evidence for this by showing that catalytically inactive CPSF73 cannot restore termination to cells lacking functional CPSF73. Notably, Xrn2 plays no significant role in either Histone or small nuclear RNA (snRNA) gene termination even though both RNA classes undergo 3′ end cleavage. In sum, efficient termination on most protein-coding genes involves CPSF73-mediated RNA cleavage and cotranscriptional degradation of polymerase-associated RNA by Xrn2. However, as CPSF73 loss caused more extensive readthrough transcription than Xrn2 elimination, it likely plays a more underpinning role in termination.

[Keywords: Xrn2; transcriptional termination; CPSF73; torpedo; allosteric; RNA polymerase II] Supplemental material is available for this article.

Supplemental material is available for this article.

Received October 23, 2017; revised version accepted January 5, 2018.

[6]These authors contributed equally to this work.

Corresponding author: s.west@exeter.ac.uk

Article published online ahead of print. Article and publication date are online at http://www.genesdev.org/cgi/doi/10.1101/gad.308528.117.

Introduction

Transcriptional termination can be defined as the cessation of RNA polymerization and dissolution of the ternary complex of RNA polymerase II (Pol II), DNA, and RNA. Termination is a biologically important process, as it prevents transcriptional interference of genes and ensures that polymerases are available for new rounds of gene expression. Despite the fact that all transcription ends this way, it is perhaps the least understood phase in the cycle. A polyadenylation signal (PAS) is a prerequisite for termination, and mutations within it were shown decades ago to cause extended transcriptional readthrough (Whitelaw and Proudfoot 1986; Connelly and Manley 1988). Two models, the allosteric and torpedo, have since framed efforts to understand PAS-dependent termination (Porrua and Libri 2015; Proudfoot 2016). In the allosteric mechanism, transcription of a PAS causes a change in Pol II structure or alters the composition of the elongation complex to promote termination. In the torpedo model, RNA cleavage generates a Pol II-associated RNA substrate for 5′ → 3′ degradation that triggers termination by pursuing and catching the polymerase (Connelly and Manley 1988; Proudfoot 1989). Multiple studies provide support for both models, with the actual mechanism likely to incorporate aspects of each. However, their relative contributions are debated due to different results obtained in a variety of experimental systems (Libri 2015).

Early support for the torpedo model came from observations that depletion of the nuclear 5′ → 3′ exonuclease Xrn2 caused termination defects on transfected plasmids (West *et al*. 2004). Its homolog, Rat1, was simultaneously found to promote termination more widely in budding yeast (Kim *et al*. 2004), with recent transcriptome-wide analysis supporting this finding (Baejen *et al*. 2017). The broader role of Xrn2 in human cells has been less clear. RNAi of Xrn2 showed no general function in termination at the 3′ ends of protein-coding genes (Nojima *et al*. 2015), but a significant effect was later observed upon concurrent expression of catalytically dead Xrn2 (Fong *et al*. 2015). It is likely that the inactive protein binds Xrn2

substrates and blocks their degradation by the diminished levels of endogenous Xrn2. As such, RNAi may not always reveal the complete set of functions for some proteins.

Rat1 was shown to promote the recruitment of some polyadenylation factors to budding yeast genes and so may sometimes affect termination indirectly through impacting PAS function (Luo *et al*. 2006). In this instance, cotranscriptional degradation of PAS-cleaved RNA was insufficient to cause termination on some genes, highlighting the possibility that RNA degradation may not always release polymerase (Luo *et al*. 2006). Even so, catalytically inactive Rat1 does not support termination on other yeast genes, and Rat1, Xrn1, and Xrn2 can all dissociate Pol II from DNA in purified systems (Kim *et al*. 2004; Park *et al*. 2015). In *Caenorhabditis elegans*, Xrn2 depletion does not affect termination on the majority of protein-coding genes, suggesting that the torpedo mechanism is less widely used in that organism (Miki *et al*. 2017).

To understand the extent to which the allosteric and torpedo models explain the termination mechanism, it is important to distinguish the role of PAS recognition from PAS cleavage, which is difficult to do in vivo. A human PAS is recognized by several multisubunit complexes that bind to its AAUAAA hexamer and downstream G/ U-rich motif (Proudfoot 2012). AAUAAA is recognized by the CPSF30 and WDR33 subunits of cleavage and polyadenylation specificity factor (CPSF), with endonuclease activity provided by CPSF73 (Mandel *et al*. 2006; Shi *et al*. 2009; Chan *et al*. 2014; Schonemann *et al*. 2014). Although CPSF73 was identified as the nuclease over a decade ago (Mandel *et al*. 2006), its function in termination is not fully characterized. This issue has been tackled using in vitro systems competent for transcription and RNA processing, which revealed that a PAS can promote termination in the absence of cleavage (Zhang *et al*. 2015). While highlighting the capacity of PAS recognition to affect Pol II activity, it is unknown whether this mechanism promotes termination in cells.

Therefore, several aspects of termination in human cells are incompletely understood, especially in terms of their generality, and understanding of the process has lagged behind that of other model organisms. It is not known whether Xrn2 degrades PAS-cleaved RNA generally or whether this process is cotranscriptional, as was envisaged in the torpedo model. Possible effects of Xrn2 on PAS cleavage are also not established in a global manner. It is also unclear whether PAS cleavage is required for termination or whether polymerase release can be promoted by cleavage-independent factors, which is an issue that has an impact on the applicability of current models.

As RNAi approaches take days and since protein depletion is often incomplete, we adopted gene editing to engineer conditional depletion of Xrn2 or CPSF73 on faster time scales. This was used to show that Xrn2 degrades the 3′ product of PAS cleavage cotranscriptionally and promotes efficient termination genome-wide, which we mapped transcriptome-wide at high resolution. Importantly, we show that CPSF73 activity is required for efficient termination, confirming a primary mechanism in which PAS cleavage precedes degradation of polymerase-associated RNA. However, CPSF73 elimination causes stronger termination defects than the loss of Xrn2, suggesting that it might promote termination by additional mechanisms when the primary process fails.

Results

An auxin-inducible degron (AID) system for rapid Xrn2 depletion

To set up a system for rapid elimination of Xrn2, CRISPR/ Cas9 was used to tag XRN2 with an AID (Fig. 1A,B). AID-tagged proteins are degraded upon addition of indole-3-acetic acid (referred to here as auxin [IAA]) in a manner dependent on plant Tir1 protein (Nishimura *et al*. 2009; Natsume *et al*. 2016). HCT116 cells were chosen for this experiment due to their diploid nature. Cells expressing Tir1 were subjected to CRISPR/Cas9 genome editing using repair templates that incorporated three tandem

mini-AID degrons and hygromycin or neomycin selection markers (Kubota *et al*. 2013; Natsume *et al*. 2016). Selection markers were separated from the tag by a P2A sequence that was cleaved during translation (Kim *et al*. 2011). Transfection of these two constructs together with an XRN2-specific guide RNA expressing Cas9 plasmid yielded multiple resistant colonies, and homozygous modification was demonstrated by PCR (Fig. 1C).

Western blotting confirmed homozygous targeting in two selected positive clones, shown by the higher-molecular-weight Xrn2 and the absence of any signal at the size expected for native Xrn2 (Fig. 1D). It is notable that Xrn2-AID is present at lower levels than endogenous Xrn2, suggesting a destabilizing effect of the tag. Even so, XRN2-AID cells showed no growth defects (Supplemental Fig. 1A). Further RNA analyses performed throughout this study also showed that RNA degradation functions are virtually unimpaired in XRN2-AID cells.

To test Xrn2-AID depletion, Western blotting was performed over a time course of auxin addition (Fig. 1E). Xrn2-AID was detected through the Flag epitope present within the AID tag, with specificity shown by a lack of signal in unmodified HCT116 cells. Importantly, Xrn2-AID levels are reduced within 30 min of auxin treatment and were virtually undetectable after 1 h. As such, this system allows rapid and conditional depletion of Xrn2. The addition of auxin to the culture medium of XRN2-AID cells completely prevented cell colony formation, showing that Xrn2 is an essential protein (Supplemental Fig. 1B).

Xrn2 plays a general role in the degradation of 3′ flanking region RNA

Next, we tested the effect of Xrn2 loss on PAS cleavage and the stability of 3′ flanking region RNA from MYC and ACTB using quantitative RT–PCR (qRT–PCR). RNA was isolated over the same time course as for the Western blot in Figure 1E, and primers were used to detect non-PAS-cleaved (UCPA) RNA or 3′ flanking transcripts (Fig. 2A). An accumulation of 3′ flanking region RNA was seen for both genes by 30 min of auxin treatment.

An even greater effect was seen after 60 min that was maintained (but not enhanced) after 120 min. In contrast, Xrn2-AID loss had no obvious effect on PAS cleavage, as no accumulation of UCPA species was observed for either gene at any time point. This experiment shows that in these two cases, Xrn2 degrades RNA beyond the PAS without affecting PAS cleavage. The latter conclusion is further supported by observations that Xrn2-AID loss has no impact on the recruitment of the polyadenylation factor Pcf11 to ACTB (Supplemental Fig. 2A). Importantly, 3′ flanking region RNA was stabilized only in the combined presence of the AID tag, Tir1, and auxin, showing that no individual factor indirectly causes the effect (Supplemental Fig. 2B). These findings are unlikely to result from secondary effects due to the speed of Xrn2-AID depletion, especially by comparison with RNAi, with the near-complete elimination of Xrn2-AID revealing function without overexpression of the inactive protein.

We then sought to test the generality of the effects seen on Xrn2-AID loss using nuclear RNA sequencing (RNA-Seq) carried out on XRN2-AID cells treated with auxin or untreated. We also performed this analysis on a HCT116 cell line that was unmodified at XRN2 and grown in the absence of auxin. Analysis of individual gene tracks confirmed the effect on MYC and ACTB, where an enhanced signal beyond their PASs was observed upon Xrn2-AID elimination (Fig. 2B). Further examples of Xrn2 effects are shown for E2F6 and RPL30 (Fig. 2C). XRN2-AID cells grown in the absence of auxin gave slightly elevated levels of 3′ flanking RNA as compared with cells unmodified at XRN2, suggesting that Xrn2-AID can carry out almost all 3′ flanking RNA degradation. Interestingly, strong effects of Xrn2 depletion were seen downstream from where Drosha cleaves microRNA (miRNA) precursors (Supplemental Fig. S3A,B), showing other ways of Xrn2 substrate generation.

Metagene plots were then generated for protein-coding genes that were separated from any reads within 3 kb of their transcription start site (TSS) and 7 kb of the PAS (denoted as TES [transcript end site]). This left 4701 genes for analysis and revealed a clear enhancement of 3′ flanking

region RNA upon auxin treatment of XRN2-AID cells (Fig. 2D). Xrn2-AID samples obtained in the absence of auxin showed slightly raised levels of 3′ flanking region RNA compared with the cell line unmodified at XRN2, arguing that reduced levels of Xrn2-AID do not cause significant readthrough defects. Metagene plots generated from an independent biological replicate showed a similar result (Supplemental Fig. 3C). We note that Xrn2-AID loss is associated with a slight reduction in reads upstream of the PAS, potentially reflecting mildly reduced gene expression that might be caused by Pol II recycling defects. Finally, closer analysis of the TES (PAS) region showed that read counts at this position are similar in all samples (Fig. 2E; Supplemental Fig. 3D). This again suggests that major PAS cleavage defects are not widespread following Xrn2 loss, which is consistent with the analysis of MYC and ACTB shown above.

Xrn2 degrades 3′ flanking RNA cotranscriptionally and promotes termination

The validity of the torpedo model of termination depends on cotranscriptional degradation of 3′ flanking region RNA taking place (Connelly and Manley 1988; Proudfoot 1989), but this has not been shown for Xrn2. To address this, we immunoprecipitated Pol II-associated RNA following cross-linking of XRN2-AID cells treated with auxin or untreated and analyzed it by qRT–PCR (Fig. 2F). Levels of UCPA RNA and 3′ flanking region RNA produced from MYC were assayed, and Xrn2 loss caused a substantial increase in the latter but not the former. This is consistent with Xrn2 involvement in the cotranscriptional degradation of 3′ flanking region RNA.

As a second measure of cotranscriptional degradation, we isolated nuclei from control or auxin-treated XRN2-AID cells and subjected them to nuclear run-on (NRO) analysis in the presence of 4-thio UTP (4sUTP). In this experiment, transcriptionally engaged Pol II was allowed to run on and label the 3′ ends of nascent transcripts in vitro. These were purified via

linkage of biotin onto 4sUTP followed by streptavidin capture (see the Materials and Methods) and subjected to qRT–PCR to analyze UCPA and 3′ flanking region transcripts from MYC (Fig. 2G). This experiment yielded a result similar to that shown in Figure 2F in that Xrn2 loss increased 3′ flanking region RNA but not UCPA transcripts. The analysis of additional genes confirmed the role of Xrn2 in cotranscriptional degradation of 3′ flanking region RNA (Supplemental Figure 3E,F). Finally, stable integration of wild-type or catalytically inactive (D235A) XRN2 into XRN2-AID cells demonstrated that both RNA degradation and termination defects caused by Xrn2-AID elimination are completely rescued by wild-type Xrn2 but not by D235A (Supplemental Fig. 4). The Xrn2 effects on transcriptional termination therefore require its exoribonuclease function.

Mammalian native elongating transcript sequencing (mNET-seq) reveals a global termination defect on Xrn2 loss

Next, we precisely interrogated the global function of Xrn2 in transcriptional termination using mNET-seq (Nojima *et al*. 2015). In this method, the position of Pol II is revealed genome-wide at single-nucleotide resolution through its immunoprecipitation and the deep sequencing of RNA extracted from its active site. An antibody was used to capture all forms of Pol II.

MYC and RPL30 mNET-seq profiles are shown in Figure 3, A and B (ACTB in Supplemental Fig. 5A). In cells not treated with auxin, termination occurs downstream from the PAS, where the mNET-seq signal reaches background. When Xrn2 is eliminated, a clear termination defect is observed, and, due to the high resolution of mNET-seq, it is possible to visualize two manifestations of this. First, where flanking region signal is detected in control cells, it is frequently elevated over the same positions in cells lacking Xrn2. This can be seen in the MYC and RPL30 examples in Figure 3, A and B (blue arrows), and is consistent with polymerase stalling over termination regions facilitating termination by Xrn2. While this

provides evidence that Xrn2 might not always have to pursue a still-transcribing Pol II, an additional effect of Xrn2 loss is an enhanced mNET-seq signal beyond where termination takes place in control cells. An example of this is marked by the red bracket on the RPL30 gene plot in Figure 3B and suggests that normal termination sites can be ignored, with polymerases potentially having escaped pursuit by Xrn2.

We next addressed the generality of Xrn2 function in termination by generating metagene plots from control and auxin-treated cells. We analyzed expressed genes separated from upstream and downstream reads by at least 1 and 15 kb, respectively, which revealed a general transcriptional termination defect upon loss of Xrn2 (Fig. 3C). Interestingly, mNET-seq signal declined even in the absence of Xrn2, suggesting the existence of termination mechanisms that do not depend on it. The metagene plot of a separate biological replicate of this experiment showed the same general termination defect upon Xrn2-AID loss (Supplemental Fig. 5B). Interestingly, some genes were especially sensitive to Xrn2 elimination and showed more extensive readthrough than the genome-wide trend—as exemplified by TBL1XR1 in Figure 3D. Nuclear RNA-Seq analysis confirmed the extended readthrough over TBL1XR1 (Supplemental Fig. 5C).

PAS cleavage is not the only mechanism to generate RNA 3′ ends. For instance, Drosha processes miRNAs, and a small number of noncoding RNA genes use this mechanism of 3′ end formation (Dhir *et al*. 2015). We tested whether Xrn2 promoted termination of two examples of these long noncoding primary miRNAs (lncpri-miRNAs): MIR17HG and MIR31HG (Fig. 3E,F). Cotranscriptional miRNA cleavage is visible (Fig. 3E,F, red asterisks) in both cases due to the known capacity of mNET-seq to detect Drosha cleavage products that remain associated with transcribing Pol II (Nojima *et al*. 2015). For MIR17HG, nascent transcription is detected in Xrn2 depleted samples beyond where termination occurs in the control experiment. There is also a higher read density beyond the MIR31HG miRNA sequence upon Xrn2 loss, with a noticeable defect emphasized by

the reduced read count upstream of the Drosha cleavage site. This supports the notion that Xrn2 promotes efficient transcriptional termination from multiple cleavage processes, as suggested previously (Fong *et al*. 2015).

Transcriptional termination on Histone and small nuclear RNA (snRNA) genes is unaffected by Xrn2 loss

Although not polyadenylated, Histone RNAs also use CPSF for 3′ end formation, which could provide an entry site for Xrn2 (Dominski *et al*. 2005; Kolev *et al*. 2008), and we were interested in whether this was the case. Figure 4A shows mNET-seq traces of the HIST1 cluster in XRN2-AID cells treated with auxin or untreated. Interestingly, there is no impact of Xrn2 loss on transcriptional termination of any of the genes in this cluster, strongly suggesting that Xrn2 does not play a prominent role in Histone gene termination. This result was confirmed for other examples of Histone genes, and, similarly, RNA-Seq showed little to no effect of Xrn2 elimination on 3′ flanking region RNA deriving from these genes (Supplemental Fig. 6). snRNAs also undergo 3′ end cleavage by the integrator complex, and this may also precede Xrn2 activity (Baillat *et al*. 2005). However, as for Histone genes, our mNET-seq and RNA-Seq analyses showed no major role for Xrn2 in their transcriptional termination or in the degradation of their 3′ flanking region transcripts (Fig. 4B; Supplemental Fig. 7). As such, 3′ end cleavage is not always sufficient to promote an Xrn2-dependent termination process.

Conditional depletion of CPSF73 causes a strong PAS cleavage and termination defect

For Xrn2 to function in termination, RNA cleavage is required, and this presumably occurs most often at the PAS. CPSF73 is the PAS endonuclease in humans, and its depletion by RNAi causes strong termination defects genome-wide, confirming its general function in the process (Nojima *et al*.

2015). However, depletion of CPSF73 cannot establish whether its catalytic center or physical presence underlies its function in termination. To begin testing this, we generated cells in which the PAS endonuclease CPSF73 could be manipulated in a manner similar to Xrn2-AID. As we were unable to make an AID-tagged version of CPSF73, we tagged its C terminus with an *Escherichia coli* DHFR-based degron using the system used for Xrn2-AID (Iwamoto *et al*. 2010; Sheridan and Bentley 2016). In this system, cells are grown in the presence of trimethoprim (TMP), the withdrawal of which triggers degradation of the tagged protein. Western blotting confirmed homozygous tagging of CPSF73 with DHFR, as CPSF73-DHFR was seen to migrate at a higher molecular weight than the native protein for which there was no signal in the CRISPR-modified cell line (Fig. 5A). Withdrawal of TMP from the medium promoted near elimination of CPSF73-DHFR after 10 h. This rate of depletion is slower than for Xrn2-AID but more than sevenfold faster than what we used previously for functional depletion of CPSF73 by RNAi (Davidson *et al*. 2014).

We tested the impact of CPSF73-DHFR elimination on 3′ end processing of MYC and ACTB transcripts by qRT– PCR of total RNA from CPSF73-DHFR cells grown in the presence or absence of TMP (Fig. 5B). For both genes, there was a significant reduction of PAS cleavage, demonstrated by an accumulation of UCPA RNA. Notably, the magnitude of effect (sevenfold to 12-fold) was threefold to fourfold greater than we observed previously by RNAi of CPSF73 (Davidson *et al*. 2014), highlighting the enhanced effects gained from this system.

To analyze the effect of CPSF73 depletion on termination, Pol II chromatin immunoprecipitation (ChIP) was performed in CPSF73-DHFR cells grown in the presence or absence of TMP. Pol II occupancy was monitored downstream from MYC and ACTB (Fig. 5C,D). In both cases, CPSF73 loss caused a general reduction in transcription, as evidenced by the lower Pol II signal upstream of the PAS (denoted as US). This is consistent with observations that PAS mutations or polyadenylation factor depletion negatively impacts transcription (Mapendano *et al*. 2010). Despite this, a

large termination defect was evident on both genes through the accumulation of Pol II beyond the normal site of termination.

CPSF73 elimination causes more extensive readthrough than loss of Xrn2

We next tested whether CPSF73 and Xrn2 produced differential effects on readthrough transcription. For this, Pol II ChIP was compared for CPSF73-DHFR cells ±TMP, on XRN2-AID cells, and on D235A XRN2-AID cells +auxin(Fig. 6A,B). D235A XRN2-AID cells stably express catalytically inactive Xrn2 that is not sensitive to auxin. When these cells are treated with auxin, 5′ → 3′ degradation of readthrough RNA and termination are more strongly impaired than in auxin-treated XRN2-AID cells (Supplemental Fig. 4). Pol II occupancy over extended readthrough regions of MYC and ACTB was plotted relative to the signal from upstream of the PAS. For both genes, CPSF73 depletion resulted in greater signals over extended positions than elimination of Xrn2 function, suggesting that termination is more adversely effected by loss of CPSF73. qRT–PCR analysis of readthrough RNA over the same positions confirmed this result (Supplemental Fig. 8A). Inhibition of CPSF30 function by influenza NS1A protein (Nemeroff *et al*. 1998) also caused more extensive transcriptional readthrough than Xrn2, further arguing for a more crucial function of CPSF in promoting termination (Supplemental Fig. 8B,C).

Although auxin-treated D235A cells represent the scenario most lacking in 5′ → 3′ degradation of RNA, the smaller effect on termination relative to CPSF73 loss may be due to incomplete Xrn2 depletion or other 5′ → 3′ nucleases acting in its absence. To address this, we analyzed the turnover rate of 3′ flanking region transcripts from MYC and ACTB in more detail. A time course was used in XRN2-AID cells treated with auxin or untreated and in D235A cells treated with auxin following transcriptional inhibition by actinomycin D (Act D) (Fig. 6C). In XRN2-AID cells not treated with auxin, Act D induced a strong reduction in the level of 3′ flanking region RNA, consistent with rapid degradation. The addition of auxin

resulted in greater recovery of 3′ flanking region RNA following Act D treatment that was more pronounced in D235A cells treated with auxin. This confirms the role of Xrn2 in their degradation. However, degradation was incompletely blocked by Xrn2 elimination, as ~40%–60% of these transcripts were still degraded after transcriptional inhibition even in auxin-treated D235A cells.

The degradation of 3′ flanking region RNA in auxin-treated D235A cells could be by alternative 5′ → 3 ′ exonucleases or from the 3′ end by the exosome. To address this, we treated D235A cells with control or human Rrp40 (hRrp40)-specific siRNAs before auxin addition (Fig. 6D; Supplemental Fig. 9A,B). The same experiment was performed on XRN2-AID cells not treated with auxin to determine any exclusive effects of hRrp40 depletion. We first tested the effects of these conditions on the levels of MYC and ACTB 3′ flanking region RNA. hRrp40 depletion alone gave no substantial effect, whereas auxin treatment of D235A cells gave the expected strong accumulation. When hRrp40 was depleted from D235A cells treated with auxin, there was an accumulation of 3′ flanking region RNA above what was seen upon manipulation of Xrn2 that was most marked for MYC transcripts. Therefore, the exosome contributes to readthrough RNA degradation in the absence of Xrn2 function. The level of UCPA transcripts was similar under each of these conditions, arguing that PAS cleavage is unaffected (Supplemental Fig. 9C).

Next, the impact of the exosome on degradation of 3′ flanking RNAs was assessed after 20 min of Act D treatment (Fig. 6E). In the XRN2-AID sample treated with control siRNA, Act D treatment caused depletion of ACTB and MYC flanking transcripts as expected, and hRrp40 depletion gave a similar result. In auxin-treated D235A cells, ~40%–60% of 3′ flanking region RNA was again degraded in the absence of Xrn2 function. Importantly, hRrp40 depletion from D235A cells grown in auxin essentially blocked degradation, as the level of RNA recovered after transcription inhibition was similar to before Act D addition. This shows that the exosome rather than other 5′ → 3′ exonucleases is responsible for the degradation

of RNA that occurs in the absence of functional Xrn2. As such, auxin treatment of D235A cells effectively blocks degradation of 3′ flanking region transcripts from their 5′ ends. A similar result was obtained when transcription was inhibited using flavopiridol (Supplemental Fig. 9D). Act D time course analysis also revealed that CPSF73 elimination prevented turnover of 3′ flanking region RNA, arguing that PAS cleavage is necessary to promote their degradation (Supplemental Fig. 9E). These data argue that the differential effect of Xrn2 and CPSF73 on transcriptional termination is unlikely to be due to an incomplete block of 5′ → 3′ degradation when Xrn2 is manipulated. As such, they support the existence of additional termination mechanisms that occur in the absence of 5′ → 3′ degradation.

A CPSF73 active site mutant cannot support efficient transcriptional termination

A primary termination pathway involving Xrn2 predicts a requirement for PAS cleavage. To test whether active CPSF73 is required for termination, we generated plasmids containing either wild-type CPSF73 or a point-mutated derivative (H73A) shown previously to have diminished nuclease activity (Kolev *et al*. 2008). The plasmid system was used because repeated attempts to stably integrate H73A into CPSF73-DHFR cells failed, potentially because of its deleterious effect. Plasmids also incorporated puromycin selection markers to enrich for transfected cells. Western blotting confirmed similar expression of wild-type and H73A proteins in CPSF73-DHFR cells and the expected absence of endogenous-sized CPSF73 in empty vector transfected samples (Fig. 7A).

To test the ability of H73A to function in termination, CPSF73-DHFR cells were transfected with empty vector, wild type, or H73A. Transfected cells were then enriched for by puromycin selection before removal (or not) of CPSF-DHFR via 10 h of TMP withdrawal. Chromatin-associated RNA was then isolated to study termination via the extent of nascent RNA

transcription, which was assayed by qRT–PCR for MYC and ACTB genes (Fig. 7B–D). In empty vector transfected cells, TMP withdrawal induced the expected accumulation of UCPA RNA and a strong enhancement of readthrough transcripts extending beyond the PAS. These readthrough defects were substantially suppressed in the absence of TMP by wild-type CPSF73.

Discussion

Our study reveals a clear role for CPSF73 activity and 5′ → 3′ degradation in efficient termination on protein-coding genes as envisioned by the torpedo model. They are most consistent with a primary mechanism in which PAS site cleavage precedes cotranscriptional degradation of Pol II-associated RNA by Xrn2. However, we also observed some termination in situations where 5′ → 3′ degradation of RNA was blocked, arguing for alternative secondary mechanisms. In particular, ablation of CPSF73 activity caused more readthrough than seen on loss of Xrn2, suggesting additional roles for CPSF73 in termination. The observation that miRNA cleavage is capable of promoting Xrn2-dependent termination argues that RNA cleavage may more widely underpin the process beyond protein-coding genes.

Previous reports have reached different conclusions on the role of Xrn2 in termination. Originally, RNAi of Xrn2 caused a termination defect on transfected β-globin plasmids, while a subsequent global analysis found no genome-wide function for Xrn2 in termination at gene 3′ ends using mNET-seq (West *et al*. 2004; Nojima *et al*. 2015). An explanation for this came through observations that RNAi of Xrn2 caused termination defects when catalytically inactive Xrn2 was also expressed (Fong *et al*. 2015). Our results support the view that trace levels of active Xrn2 can provide false negative results in RNAi experiments because Xrn2-AID is virtually eliminated in our system, with its levels likely falling below a critical threshold. Moreover, although Xrn2-AID protein is at substantially reduced

levels compared with native Xrn2, this is still sufficient to promote termination, suggesting that a fraction of normal levels supports this function. Finally, expression of inactive Xrn2 in XRN2-AID cells has a dominant-negative effect on termination in our system (Supplemental Fig. 4E). These observations may be of importance beyond Xrn2, as they suggest that a degron-based approach can yield a fuller repertoire of functions for some proteins than RNAi alone.

Another finding in our study is that termination is not readily observed in the absence of CPSF73 activity, suggesting that PAS cleavage is required. In vitro experiments suggest that PAS cleavage is not absolutely required for termination (Zhang *et al*. 2015). However, additional cellular factors may be absent from in vitro systems. Moreover, RNA degradation improved termination in that system, consistent with our finding on the importance of Xrn2 in cells. We do note that H73A CPSF73 has been shown to immunoprecipitate other CPSF components slightly less efficiently than wild-type CPSF73 (Kolev *et al*. 2008). This means that the presence of incomplete or unstable CPSF complexes might account for the inability of the H73A mutant to promote termination. If this is true, then it would identify CPSF assembly or activation as providing a crucial function in the process rather than PAS cleavage itself. This would still be an important observation, but we favor PAS cleavage as important for several reasons. First, H73A proved an effective dominant-negative inhibitor of PAS cleavage. Second, partial defects in complex formation might be expected to result in partial termination defects instead of the very strong effect caused by exclusive H73A expression. Moreover, recent results show that polyadenylation factors, exemplified by CstF64, assemble on inactive intronic PASs, but this is insufficient to cause termination unless cleavage is activated by U1 snRNA inhibition (Oh *et al*. 2017). Finally, the widespread requirement for Xrn2 in efficient termination is most readily explained by PAS cleavage preceding its action.

We also suggest that CPSF73 is required for termination even in the absence of Xrn2. The evidence for this conclusion is that the termination

defect is larger upon loss of CPSF73 than when Xrn2 is absent. This could be due to allosteric effects induced by CPSF assembly or activity. Alternatively, such termination could be via the RNA:DNA helicase activity of Senataxin (Skourti-Stathaki *et al*. 2011), given that its budding yeast homolog, Sen1, can terminate polymerase in purified systems (Porrua and Libri 2013). The exosome may also terminate Pol II by degrading RNA that protrudes from the front of backtracked polymerase (Lemay *et al*. 2014). Our data argue that these possibilities, including an allosteric mechanism, would require PAS cleavage, given the inability of inactive CPSF73 to support termination. A termination mechanism underpinned by cleavage may also apply following miRNA cleavage. We show an Xrn2 effect on this process; however, the readthrough caused is less than previously observed when miRNA cleavage was prevented by Drosha depletion (Dhir *et al*. 2015). Drosha depletion caused MIR17HG transcription to enter the downstream GPC5 gene, whereas transcription terminates before this point following Xrn2 loss (Fig. 3E).

While it is difficult to interrogate some molecular details of termination in cells, important principles are consolidated here. In particular, we provide strong evidence that PAS cleavage and cotranscriptional degradation of Pol II-associated RNA are key components of the most efficient termination mechanism. Our results align with predictions of the torpedo model made using highly purified in vitro systems, where it was shown that Xrn2-, Rat1-, and Xrn1-mediated RNA degradation terminates Pol II (Park *et al*. 2015). In those cases, termination improved when Pol II-associated RNA was longer or when Pol II progression was prevented by nucleotide misincorporation, suggesting that nuclease momentum or polymerase stalling may facilitate the process in cells. Polymerase backtracking over termination regions was inferred from transient transcriptome sequencing (TT-seq) (Schwalb *et al*. 2016). Moreover, our mNET-seq shows signal accumulation over termination regions that may result from pausing or backtracking. As this signal is often enhanced by loss of Xrn2 (denoted by the blue arrows in Fig. 3),

polymerases prone at these sites may be more vulnerable to termination by Xrn2. As we also observed a signal beyond termination sites upon loss of Xrn2, it will be interesting to establish whether this represents polymerases that resume transcription following pausing or those normally terminated by a pause-independent process. In sum, our results provide important details on the termination mechanism in human cells, especially regarding CPSF73 and Xrn2 activities. Our AID system provides a rationale for why RNAi of Xrn2 led to controversy over its role in the process, and our DHFR approach gives strong evidence that PAS cleavage precedes termination.

Materials and Methods

Plasmids, primers, and DNA sequences

Primer sequences used for ChIP and qRT–PCR, sequences of repair templates, homology arms, and guide RNA target sites are provided in the Supplemental Material.

Antibodies

The antibodies used were Pol II (CMA601; MBL Technologies), CPSF73 (Abcam, ab72295), CPSF73 for Figure 5A (Bethyl Laboratories, A301-090A), Flag (Sigma, F3165), HA (Roche, 3F10), Xrn2 (Bethyl Laboratories, A301-101), SF3b155 (Abcam, ab39578), Myc (Sigma, 9E10), Pcf11 (Bethyl Laboratories, A303-705 and A303-706), and NS1A (gift from Aldolfo Garcia-Sastre).

Cell culture

HCT116 cells were maintained in DMEM with 10% fetal calf serum. Transfections were with JetPrime (polyplus). For CRISPR, 1 µg of guide RNA plasmid and 1 µg of each repair plasmid were transfected into six-well

dishes. Twenty-four hours later, culture medium was changed, and, a further 24 h later, cells were split into a 100-mm dish containing 800 µg/mL neomycin and 150 µg/mL hygromycin. After ~10 d of selection, single colonies were transferred to a 24-well plate and screened by PCR or Western blotting. The presence of repair cassettes at XRN2 or CPSF73 was confirmed by Sanger sequencing. An optimized sleeping beauty transposon system (Kowarz *et al*. 2015) was used to generate Tir1-expressing parental cell lines and cells in which Xrn2 derivatives were stably transfected. A 24-well dish was transfected with 300 ng of sleeping beauty plasmid (derived from pSBbi-puro/pSBbi-blast)) and 100 ng of pCMV(CAT) T7-SB100. Twenty-four hours later, cells were put under selection with 1 µg/mL puromycin or 20 µg/mL blasticidin. For Tir1-expressing cells, single colonies were isolated; for Xrn2 rescue experiments, the entire population was studied. Auxin (Sigma) was added to 500 nM for 60 min unless stated otherwise. TMP (Sigma) was maintained at 20 µM, and, for depletion, cells were grown in medium lacking TMP for 10 h unless stated otherwise. Act D and flavopiridol were used at 5 µg/mL and 1 µM, respectively.

qRT–PCR

TRI Reagent (Sigma) was used to isolate total RNA following the manufacturers' guidelines, and RNA was treated with Turbo DNase (Life Technologies) for 1 h. In all cases, reverse transcription of 1 µg of RNA was primed with random hexamers using Protoscript II (New England Biolabs). qPCR was performed using Brilliant III (Agilent Technologies) in a Qiagen Rotorgene instrument. Comparative quantitation was used to establish fold effects.

ChIP and RNA immunoprecipitation

For ChIP, one 100-mm dish of cells was cross-linked for 10 min in 0.5% formaldehyde, and cross-links were quenched in 125 mM glycine for 5 min.

Cells were collected (500g for 5 min) and resuspended in 400 µL of RIPA buffer (150 mM NaCl, 1% NP40, 0.5% sodium deoxycholate, 0.1% SDS, 50 mM Tris-HCl at pH 8, 5 mM EDTA at pH 8). Samples were sonicated in a Bioruptor sonicator (30 sec on and 30 sec off) 10 times on high setting. Tubes were spun at 13,000 rpm for 10 min. Supernatant was then split into two and added to 30 µL of Dynabeads (Life Technologies) that had been incubated for 2 h with 3 µg of antibody or, as a control, mock-treated. Ten percent of the supernatant was kept for input. Immunoprecipitation was for 2–14 h at 4°C, and beads were then washed twice in RIPA, three times in high-salt wash buffer (500 mM NaCl, 1% NP40, 1% sodium deoxycholate, 100 mM Tris-HCl at pH 8.5), and once in RIPA. Samples were eluted (0.1 M NaHCO3 + 1% SDS), and cross-links were reversed overnight at 65°C. DNA was purified by phenol chloroform extraction and ethanol precipitation. Samples were generally resuspended in 100 µL of water, with 1 µL used per PCR reaction. For RNA immunoprecipitation, cross-links were reversed for 45 min at 65°C. RNA was purified by phenol chloroform extraction and ethanol precipitation followed by DNase treatment and reverse transcription.

Chromatin RNA isolation

Nuclei were isolated from cells by resuspending cell pellets from a 100-mm dish in hypotonic lysis buffer (HLB; 10 mM Tris at pH 7.5, 10 mM NaCl, 2.5 mM MgCl2, 0.5% NP40). This was underlayered with HLB + 10% sucrose and spun at 500g for 5 min. Nuclei were resuspended in 100 µL of NUN1 (20 mM Tris-HCl at pH 7.9, 75 mM NaCl, 0.5 mM EDTA, 50% glycerol, 0.85 mM DTT). One milliliter of NUN2 (20 mM HEPES at pH 7.6, 1 mM DTT, 7.5 mM MgCl2, 0.2 mM EDTA. 0.3 M NaCl, 1 M urea, 1% NP40) was added before incubation for 15 min on ice with regular vortexing. Chromatin pellets were isolated by centrifugation at 13,000 rpm in a benchtop centrifuge, and RNA was isolated using Trizol.

4sUTP NRO

Nuclei were isolated as for chromatin-associated RNA. These were resuspended in an equal volume of 2× transcription buffer (40 mM Tris-HCl at pH 7.9, 300 mM KCl, 10 mM MgCl2, 40% glycerol). This was supplemented with rA, C, and G together with 4sUTP (final concentration ~0.1 mM). Following incubation for 15 min at 30°C, RNA was isolated, and biotin linkage and capture were performed as described in Duffy *et al*. (2015) with some modification. RNA (15–20 µg) was biotinylated in a volume of 250 µL containing 10 mM HEPES (pH 7.5) and 5 µg of MTSEA Biotin-XX (Iris Biotech) dissolved in dimethyl formamide. After incubation in the dark for 90 min, biotinylated RNA was phenol chloroform-extracted and ethanol-precipitated. This was resuspended in RPB (300 mM NaCl, 10 mM Tris at pH 7.5, 5 mM EDTA) and incubated with 150 µL of streptavidin-coated paramagnetic particles (Promega) for 15 min. Beads were washed five times in 100 mM Tris-HCl (pH 7.4), 10 mM EDTA, 1 M NaCl, and 0.1% Tween-20 preheated to 60°C. RNA was eluted in 100 µL of 0.1 M DTT for 15 min at 37°C before final phenol chloroform extraction and ethanol precipitation.

Nuclear RNA-Seq

Following 1 h of auxin or mock treatment, nuclei were isolated as for chromatin-associated RNA. Nuclear RNA was extracted using Trizol reagent. rRNA was removed using Ribo-Zero Gold rRNA removal kit (Illumina) according to the user manual. Libraries were prepared using TruSeq stranded total RNA library preparation kit (Illumina) according to the manual and purified using Ampure XP beads (Beckman Coulter). Libraries were screened for fragment size and concentration by Tapestation D1000 (Agilent) and sequenced using HiSeq 2500 (Illumina).

Raw single-end 50-base-pair (bp) reads were screened for sequencing quality using FASTQC, adapter sequences were removed using Trim Galore (wrapper for Cutadapt), and trimmed reads <20 bp

were discarded. Reads were aligned to the GRCh38 human genome using Hisat2 (Kim *et al*. 2015) with splice site annotation from Ensembl. Unmapped and low MAPQ reads were discarded. For metagene analyses, expression levels were calculated for each gene, and genes with low or no expression were removed. A transcriptional window was then applied (TSS − 3 kb and TES + 7 kb). Genes with overlapping reads in this window were discarded (Quinlan and Hall 2010). Metagene profiles were generated using the deeptools suite (Ramirez *et al*. 2016), with further graphical processing performed in the R environment (http://www.R-project.org). Normalized gene coverage plots were visualized using the Integrated Genome Viewer suite (Robinson *et al*. 2011).

mNET-seq

A detailed description of the mNET-seq protocol can be found in the study by Nojima *et al*. (2016). XRN2-AID cells were treated for 2 h with auxin or left untreated. NEBNext small RNA libraries were sequenced using HiSeq 2500 (Illumina). Raw 50-bp paired-end sequences had adapter sequences removed using Trim Galore, and resultant reads with a quality of <20 and fragment size of <19 bp were discarded. Reads were aligned using HiSat2 against GRCh38 (Ensemble) with known splice site annotation (Gencode), and concordantly mapped read pairs were selected (Kim *et al*. 2015).

The mNET-seq traces used single-nucleotide resolution BAM files corresponding to the 3′ end of the RNA fragment (Nojima *et al*. 2015). For metagene profiles, gene expression was determined by converting raw read counts into transcripts per million (TPM) for each annotated gene (Li and Dewey 2011; Wagner *et al*. 2012; Liao *et al*. 2014). For protein-coding metaplots, genes were selected where no other expressed annotated gene overlapped the exclusion range (TES − 1250 bp to TES + 15,250 bp). For each nucleotide across the region, fragments were counted in a 5-bp sliding window and converted to TPM. The normalized metagene profiles

represent the average nascent RNA fragment density against relative position from the TES. mNET-seq and RNA-Seq data have been deposited with Gene Expression Omnibus (accession no. GSE109003).

Acknowledgments

We thank Karen Moore and Exeter University Sequencing Service for help with RNA-Seq and mNET-seq. We thank members of our laboratory for reading the manuscript. Ervin Fodor is thanked for supporting D.L.V.B. through a program grant from the Medical Research Council. Oksana Gonchar is thanked for assistance with Western blotting. This work was supported by a Wellcome Trust Investigator Award (WT107791/Z/15/Z) and a Lister Institute Research Fellowship held by S.W.

References

Baejen C, Andreani J, Torkler P, Battaglia S, Schwalb B, Lidschreiber M, Maier KC, Boltendahl A, Rus P, Esslinger S, *et al*. 2017. Genome-wide analysis of RNA polymerase II termination at protein-coding genes. *Mol Cell* 66: 38–49 e36.

Baillat D, Hakimi MA, Naar AM, Shilatifard A, Cooch N, Shiekhattar R. 2005. Integrator, a multiprotein mediator of small nuclear RNA processing, associates with the C-terminal repeat of RNA polymerase II. *Cell* 123: 265–276.

Chan SL, Huppertz I, Yao C, Weng L, Moresco JJ, Yates JR III, Ule J, Manley JL, Shi Y. 2014. CPSF30 and Wdr33 directly bind to AAUAAA in mammalian mRNA 3′ processing. *Genes Dev* 28: 2370–2380.

Connelly S, Manley JL. 1988. A functional mRNA polyadenylation signal is required for transcription termination by RNA polymerase II. *Genes Dev* 2: 440–452.

Davidson L, Muniz L, West S. 2014. 3′ end formation of pre-mRNA and phosphorylation of Ser2 on the RNA polymerase II CTD are reciprocally coupled in human cells. *Genes Dev* 28: 342–356.

Dhir A, Dhir S, Proudfoot NJ, Jopling CL. 2015. Microprocessor mediates transcriptional termination of long noncoding RNA transcripts hosting microRNAs. *Nat Struct Mol Biol* 22: 319–327.

Dominski Z, Yang XC, Marzluff WF. 2005. The polyadenylation factor CPSF73 is involved in histone–pre-mRNA processing. *Cell* 123: 37–48.

Duffy EE, Rutenberg-Schoenberg M, Stark CD, Kitchen RR, Gerstein MB, Simon MD. 2015. Tracking distinct RNA populations using efficient and reversible covalent chemistry. *Mol Cell* 59: 858–866.

Fong N, Brannan K, Erickson B, Kim H, Cortazar MA, Sheridan RM, Nguyen T, Karp S, Bentley DL. 2015. Effects of transcription elongation rate and Xrn2 exonuclease activity on RNA polymerase II termination suggest widespread kinetic competition. *Mol Cell* 60: 256–267.

Iwamoto M, Bjorklund T, Lundberg C, Kirik D, Wandless TJ. 2010. A general chemical method to regulate protein stability in the mammalian central nervous system. *Chem Biol* 17: 981–988.

Kim M, Krogan NJ, Vasiljeva L, Rando OJ, Nedea E, Greenblatt JF, Buratowski S. 2004. The yeast Rat1 exonuclease promotes transcription termination by RNA polymerase II. *Nature* 432: 517–522.

Kim JH, Lee SR, Li LH, Park HJ, Park JH, Lee KY, Kim MK, Shin BA, Choi SY. 2011. High cleavage efficiency of a 2A peptide derived from porcine teschovirus-1 in human cell lines, zebra-fish and mice. *PLoS One* 6: e18556.

Kim D, Langmead B, Salzberg SL. 2015. HISAT: a fast spliced aligner with low memory requirements. *Nat Methods* 12: 357–360.

Kolev NG, Yario TA, Benson E, Steitz JA. 2008. Conserved motifs in both CPSF73 and CPSF100 are required to assemble the active endonuclease for histone mRNA 3′ -end maturation. *EMBO Rep* 9: 1013–1018.

Kowarz E, Loscher D, Marschalek R. 2015. Optimized sleeping beauty transposons rapidly generate stable transgenic cell lines. *Biotechnol J* 10: 647–653.

Kubota T, Nishimura K, Kanemaki MT, Donaldson AD. 2013. The Elg1 replication factor C-like complex functions in PCNA unloading during DNA replication. *Mol Cell* 50: 273–280.

Lemay JF, Larochelle M, Marguerat S, Atkinson S, Bahler J, Bach and F. 2014. The RNA exosome promotes transcription termination of backtracked RNA polymerase II. *Nat Struct Mol Biol* 21: 919–926.

Li B, Dewey CN. 2011. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* 12: 323.

Liao Y, Smyth GK, Shi W. 2014. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* 30: 923–930.

Libri D. 2015. Endless quarrels at the end of genes. *Mol Cell* 60: 192–194.

Luo W, Johnson AW, Bentley DL. 2006. The role of Rat1 in coupling mRNA 3′ -end processing to transcription termination: implications for a unified allosteric-torpedo model. *Genes Dev* 20: 954–965.

Mandel CR, Kaneko S, Zhang H, Gebauer D, Vethantham V, Manley JL, Tong L. 2006. Polyadenylation factor CPSF73 is the pre-mRNA 3′ -end-processing endonuclease. *Nature* 444: 953–956.

Mapendano CK, Lykke-Andersen S, Kjems J, Bertrand E, Jensen TH. 2010. Crosstalk between mRNA 3′ end processing and transcription initiation. *Mol Cell* 40: 410–422.

Miki TS, Carl SH, Grosshans H. 2017. Two distinct transcription termination modes dictated by promoters. *Genes Dev* 31: 1870–1879.

Natsume T, Kiyomitsu T, Saga Y, Kanemaki MT. 2016. Rapid protein depletion in human cells by auxin-inducible degron tagging with short homology donors. *Cell Rep* 15: 210–218.

Nemeroff ME, Barabino SM, Li Y, Keller W, Krug RM. 1998. Influenza virus NS1 protein interacts with the cellular 30 kDa sub-unit of CPSF and inhibits 3′ end formation of cellular pre-mRNAs. *Mol Cell* 1: 991–1000.

Nishimura K, Fukagawa T, Takisawa H, Kakimoto T, Kanemaki M. 2009. An auxin-based degron system for the rapid depletion of proteins in nonplant cells. *Nat Methods* 6: 917– 922.

Nojima T, Gomes T, Grosso ARF, Kimura H, Dye MJ, Dhir S, Carmo-Fonseca M, Proudfoot NJ. 2015. Mammalian NET-Seq reveals genome-wide nascent transcription coupled to RNA processing. *Cell* 161: 526–540.

Nojima T, Gomes T, Carmo-Fonseca M, Proudfoot NJ. 2016. Mammalian NET-seq analysis defines nascent RNA profiles and associated RNA processing genome-wide. *Nat Protoc* 11: 413–428.

Oh JM, Di C, Venters CC, Guo J, Arai C, So BR, Pinto AM, Zhang Z, Wan L, Younis I, *et al*. 2017. U1 snRNP telescripting regulates a size-function-stratified human genome. *Nat Struct Mol Biol* 24: 993–999.

Park J, Kang M, Kim M. 2015. Unraveling the mechanistic features of RNA polymerase II termination by the 5′ -3′ exoribonuclease Rat1. *Nucleic Acids Res* 43: 2625–2637.

Porrua O, Libri D. 2013. A bacterial-like mechanism for transcription termination by the Sen1p helicase in budding yeast. *Nat Struct Mol Biol* 20: 884–891.

Porrua O, Libri D. 2015. Transcription termination and the control of the transcriptome: why, where and how to stop. *Nat Rev Mol Cell Biol* 16: 190–202.

Proudfoot NJ. 1989. How RNA polymerase II terminates transcription in higher eukaryotes. *Trends Biochem Sci* 14: 105–110.

Proudfoot NJ. 2012. Ending the message: poly(A) signals then and now. *Genes Dev* 25: 1770–1782.

Proudfoot NJ. 2016. Transcriptional termination in mammals: stopping the RNA polymerase II juggernaut. *Science* 352: aad9926.

Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* 26: 841–842.

Ramirez F, Ryan DP, Gruning B, Bhardwaj V, Kilpert F, Richter AS, Heyne S, Dundar F, Manke T. 2016. deepTools2: a next generation web server for deep-sequencing data analysis. *Nucleic Acids Res* 44: W160–W165.

Robinson JT, Thorvaldsdottir H, Winckler W, Guttman M, Lander ES, Getz G, Mesirov JP. 2011. Integrative genomics viewer. *Nat Biotechnol* 29: 24–26.

Schonemann L, Kuhn U, Martin G, Schafer P, Gruber AR, Keller W, Zavolan M, Wahle E. 2014. Reconstitution of CPSF active in polyadenylation: recognition of the polyadenylation signal by WDR33. *Genes Dev* 28: 2381–2393.

Schwalb B, Michel M, Zacher B, Fruhauf K, Demel C, Tresch A, Gagneur J, Cramer P. 2016. TT-seq maps the human transient transcriptome. *Science* 352: 1225–1228.

Sheridan RM, Bentley DL. 2016. Selectable one-step PCR-mediated integration of a degron for rapid depletion of endogenous human proteins. *Biotechniques* 60: 69–74.

Shi Y, Di Giammartino DC, Taylor D, Sarkeshik A, Rice WJ, Yates JR III, Frank J, Manley JL. 2009. Molecular architecture of the human pre-mRNA 3' processing complex. *Mol Cell* 33: 365–376.

Skourti-Stathaki K, Proudfoot NJ, Gromak N. 2011. Human senataxin resolves RNA/DNA hybrids formed at transcriptional pause sites to promote Xrn2-dependent termination. *Mol Cell* 42: 794–805.

Wagner GP, Kin K, Lynch VJ. 2012. Measurement of mRNA abundance using RNA-Seq data: RPKM measure is inconsistent among samples. *Theory Biosci* 131: 281–285.

West S, Gromak N, Proudfoot NJ. 2004. Human 5′ → 3′ exonuclease Xrn2 promotes transcription termination at co-transcriptional cleavage sites. *Nature* 432: 522–525.

Whitelaw E, Proudfoot N. 1986. a-Thalassaemia caused by a poly(A) site mutation reveals that transcriptional termination is linked to 3′ end processing in the human β globin gene. *EMBO J* 5: 2915–2922.

Zhang H, Rigo F, Martinson HG. 2015. Poly(A) signal-dependent transcription termination occurs through a conformational change mechanism that does not require cleavage at the poly(A) site. *Mol Cell* 59: 437–448.

Figure Legends

Figure 1. (A) Diagram showing the basis of auxin-dependent depletion of AID-tagged proteins. In the presence of auxin (star), Tir1 facilitates ubiquitination (blue circles) of the AID tag and rapid protein degradation. (B) Strategy for AID tagging of Xrn2. Homology arms (HAs) flanked repair cassettes containing 3× miniAID sequences, preceded by a Flag tag and separated from an antibiotic resistance gene (denoted as Abr and either Neo or Hyg) by a P2A cleavage site, with 3′ end processing driven by an SV40 PAS. (C) Diagnostic PCR of genomic DNA from antibiotic-resistant cell colonies following CRISPR gene editing. The presence of a tag increases the size of the PCR product compared with the smaller product derived from the unmodified gene. Homozygous modification is shown by the lack of unmodified product in the four drug-resistant colonies (#1–#4). (M) DNA marker. (D) Western blot confirmation of Xrn2 tagging. The top panel shows Xrn2 in two unmodified cell samples (C) and two gene-edited colonies (#1 and #2). Successful biallelic tagging is shown by the higher-molecular-weight species and the lack of native-sized Xrn2 in CRISPR-modified cells. SF3b155 was probed for as a loading control. (E) Time course of auxin addition on XRN2-AID cells. Xrn2-AID was detected by anti-Flag, and specificity is shown by the lack of product in Tir1 HCT116 cells,

which are not modified at XRN2. Tir1 was probed for as a loading control via its myc tag.

Figure 2. (A) qRT–PCR analysis of UCPA and 3′ flanking RNA from MYC and ACTB genes from total RNA during a time course of auxin addition. Values are plotted relative to those obtained at t0 after normalization to unspliced RNA levels from the respective genes. The diagram depicts the positions of UCPA amplicons and 3′ flank amplicons for both genes (+1.7 kb for ACTB and +1.8 kb for MYC). Asterisks denote $P < 0.05$ for changes relative to t0 in the absence of auxin. (B) Nuclear RNA sequencing (RNA-Seq) traces of MYC and ACTB genes in samples obtained from XRN2 unmodified cells and XRN2-AID cells treated with auxin for 1 h or untreated. The Y-axis shows RPKM (reads per kilobase transcript per million mapped reads). Bars, 1 kb. (C) As in B but showing E2F6 and RPL30genes. (D) Metagene plots from nuclear RNA-Seq on XRN2 unmodified cells and XRN2-AID cells treated with auxin or untreated. The graph shows the region from 3 kb upstream of the transcription start site (TSS) up to 7 kb beyond the PAS (denoted as transcript end site [TES]). (E) A zoomed in view of ±0.5 kb of the TES from the same metagene presented in D. (F) Pol II RNA immuno-precipitation analysis of UCPA and 3′ flanking (+1.8 kb) RNA from MYC in cells depleted of Xrn2-AID (1 h of auxin treatment) or not. Quantitation is shown for +auxin samples relative to −auxin after normalizing to the level of unspliced MYC RNA. The asterisk denotes the difference between +auxin and −auxin, where $P < 0.05$. (G) 4-thio UTP (4sUTP) nuclear run-on (NRO) analysis of UCPA and 3′ flanking (+1.8 kb) RNA from MYC in cells depleted of Xrn2-AID (1 h of auxin treatment) or not. Quantitation is shown for +auxin samples expressed relative to −auxin after normalizing to the level of unspliced MYC RNA. The asterisk denotes the difference between +auxin and −auxin, where $P < 0.05$. All error bars show standard deviation from at least three independent experiments.

Figure 3. (A) MYC mNET-seq trace from XRN2-AID cells treated (orange) with auxin or untreated (blue) for 2 h. The X-axis shows a position relative to the gene TSS in kilobases. Reads are plotted as abundance per 108 reads. Blue arrows denote a signal enhanced in the absence of Xrn2. (B) As in A but for RPL30. Additionally, the red bracket marks readthrough upon Xrn2 loss. (C) Metagene plot to analyze transcriptional termination on protein-coding genes in mNET-seq data from XRN2-AID cells grown with or without auxin. The average read density is shown over positions extending from 1 kb upstream of the TES to 15 kb downstream. The signal less than zero is transcription from the opposite strand, which is at or close to background. (D) As in A but for TBL1XR1. The red bracket denotes the region of extended read-through. (E) As in A but for MIR17HG. In this case, a red asterisk marks the miRNA cleavage events, and a red bracket marks readthrough. (F) As in E but for MIR31HG. In each diagram, the expressed gene is shown in orange, with non-expressed genes in gray.

Figure 4. (A) mNET-seq profiles over the HIST1 cluster from XRN2-AID cells treated with auxin or untreated. The Y-axes show signals per 108 mapped reads. It should be noted that reads <0 represent examples of Histone genes expressed on the opposite strand. (B) mNET-seq metagene analyses of snRNA genes from XRN2-AID cells treated with auxin or untreated. The Y-axes show average read density and are scaled to zoom into the termination region where signals are much lower than the snRNA gene body.

Figure 5. (A) Western blot showing successful tagging of CPSF73 with DHFR and a time course of CPSF73-DHFR depletion in the absence of TMP. The top panel shows native CPSF73 in unmodified HCT116 cells and the higher-molecular-weight CPSF73-DHFR in CRISPR-modified cells. CPSF73-DHFR levels are depleted in the absence of TMP. SF3b155 was detected as a loading control. (B) qRT–PCR analysis of UCPA RNA from MYC or ACTB

genes in CPSF73-DHFR cells grown in the presence or absence of TMP. Values are expressed relative to those obtained in cells grown in TMP after normalizing to unspliced RNA levels from each gene to account for any effects of transcription. Asterisks denote $P < 0.05$ for differences between +TMP and −TMP. (C) Pol II chromatin immunoprecipitation (ChIP) on MYC in CPSF73-DHFR cells grown in the presence or absence of TMP. Values are expressed as the percentage of input, and asterisks denote differences between +TMP and −TMP samples with $P < 0.05$. (D) As in C but on ACTB. All error bars show standard deviation from at least three independent experiments.

Figure 6. (A) Analysis of Pol II occupancy at +5 and +15 kb beyond the MYC PAS expressed relative to that upstream of the PAS (US) in CPSF73-DHFR cells ±TMP, XRN2-AID cells, and XRN2-AID + D235A cells +auxin (1 h). Asterisks denote $P < 0.05$ between CPSF73-DHFR – TMP and XRN2-AID + D235A + auxin. (B) As in A but for 6.3 and 12 kb beyond the ACTB PAS. (C) qRT–PCR analysis of ACTB and MYC 3′ flanking region RNA degradation in XRN2-AID cells treated with auxin or D235A cells treated with auxin (all auxin for 1 h) followed by 10 or 20 min of actinomycin D (Act D) treatment. For each sample set, RNA levels are expressed relative to that recovered at t0. (D) qRT–PCR analysis of ACTB and MYC 3′ flanking region RNA in control or human Rrp40 (hRrp40) siRNA-treated XRN2-AID cells or D235A cells treated with auxin (all auxin for 1 h). RNA levels are expressed as a fold change relative to those recovered in XRN2-AID cells treated with control siRNA following normalization to the level of unspliced MYC or ACTB transcripts. Asterisks denote $P < 0.05$ versus XRN2-AID cells treated with control siRNA. (E) qRT–PCR analysis of ACTB and MYC 3′ flanking region RNA under the conditions used in D but after 20 min of Act D treatment. Values are expressed as a percentage of RNA remaining under each condition relative to the amounts recovered in each sample at t0. Asterisks denote $P < 0.05$ versus the 0 time point. All error bars show standard deviation from at least three independent experiments.

Figure 7. (A) Western blotting of CPSF73-DHFR cells transfected with H73A CPSF73, wild-type (WT) CPSF73, or empty vector (EV) and probed with anti-HA (to detect CPSF73-DHFR) or anti-CPSF73 (to additionally detect protein derived from transfected constructs). (B) qRT–PCR analysis of chromatin-associated RNA isolated from CPSF-DHFR cells transfected with empty vector, wild-type, or H73A DHFR ± TMP. Primers were used to detect UCPA Myc RNA or RNA from +1.8 kb beyond the PAS. Values are expressed relative to those in empty vector samples in the presence of TMP after normalizing to unspliced RNA levels. Asterisks display $P < 0.05$ for comparison of the ability or inability of wild-type or H73A CPSF73 to restore termination in relation to the situation lacking CPSF73-DHFR. (C) As in B but showing signals for +5 and +15 kb beyond the MYC PAS. (D) As in B but detecting RNA from positions +6.3 or +12 kb beyond the ACTB PAS. All error bars show standard deviation from at least three independent experiments.

Figures

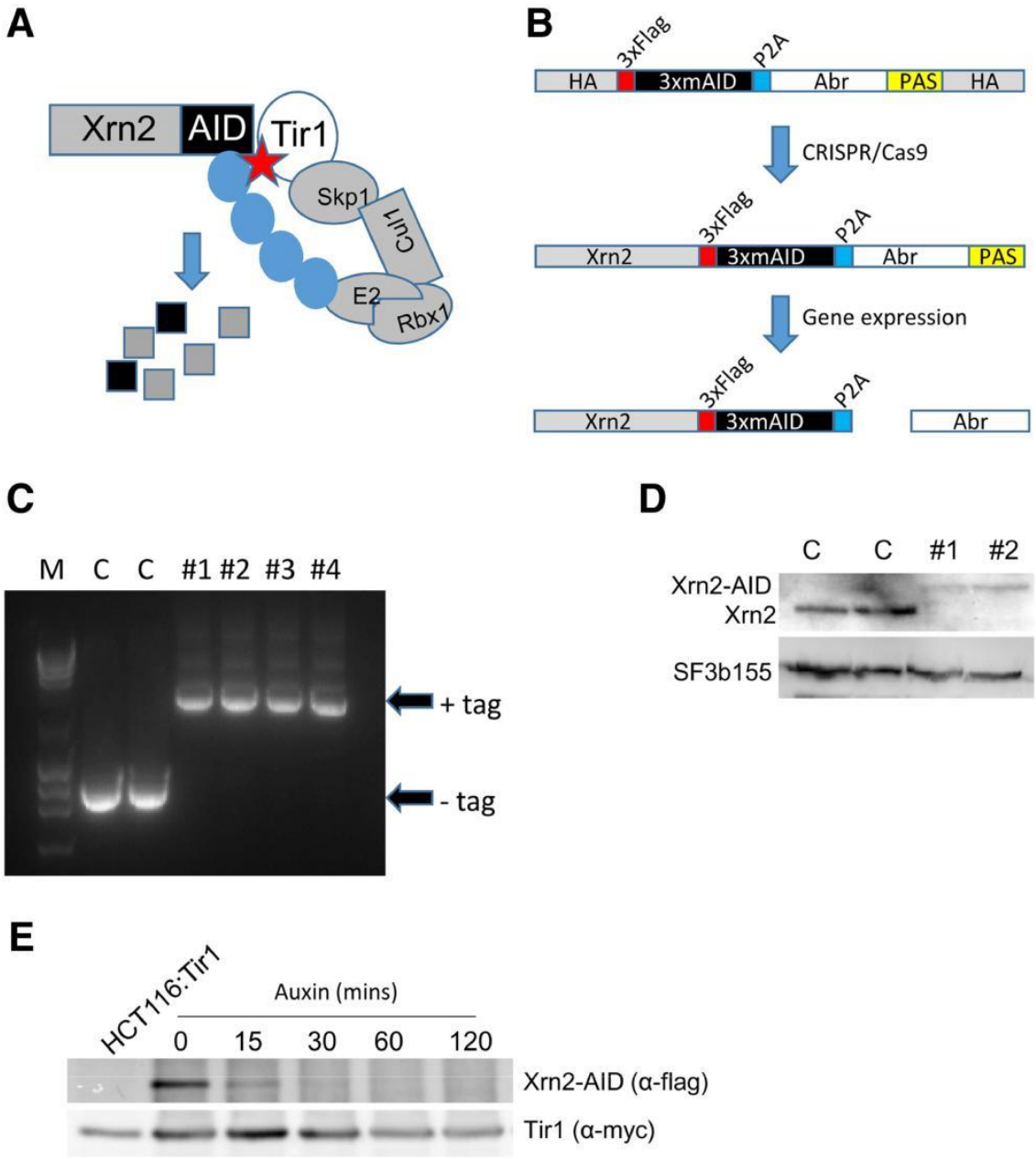

A
Xrn2
AID
Tir1
Skp1
Cul1
E2
Rbx1
B
3xFlag
P2A
HA
3xmAID
Abr
PAS
HA
CRISPR/Cas9
Xrn2
Gene expression
C
M C C #1 #2 #3 #4
+ tag
- tag
D
C C #1 #2
Xrn2-AID
Xrn2
SF3b155
E
HCT116:Tir1
Auxin (mins)
0 15 30 60 120
Xrn2-AID (α-flag)
Tir1 (α-myc)

Figure 1

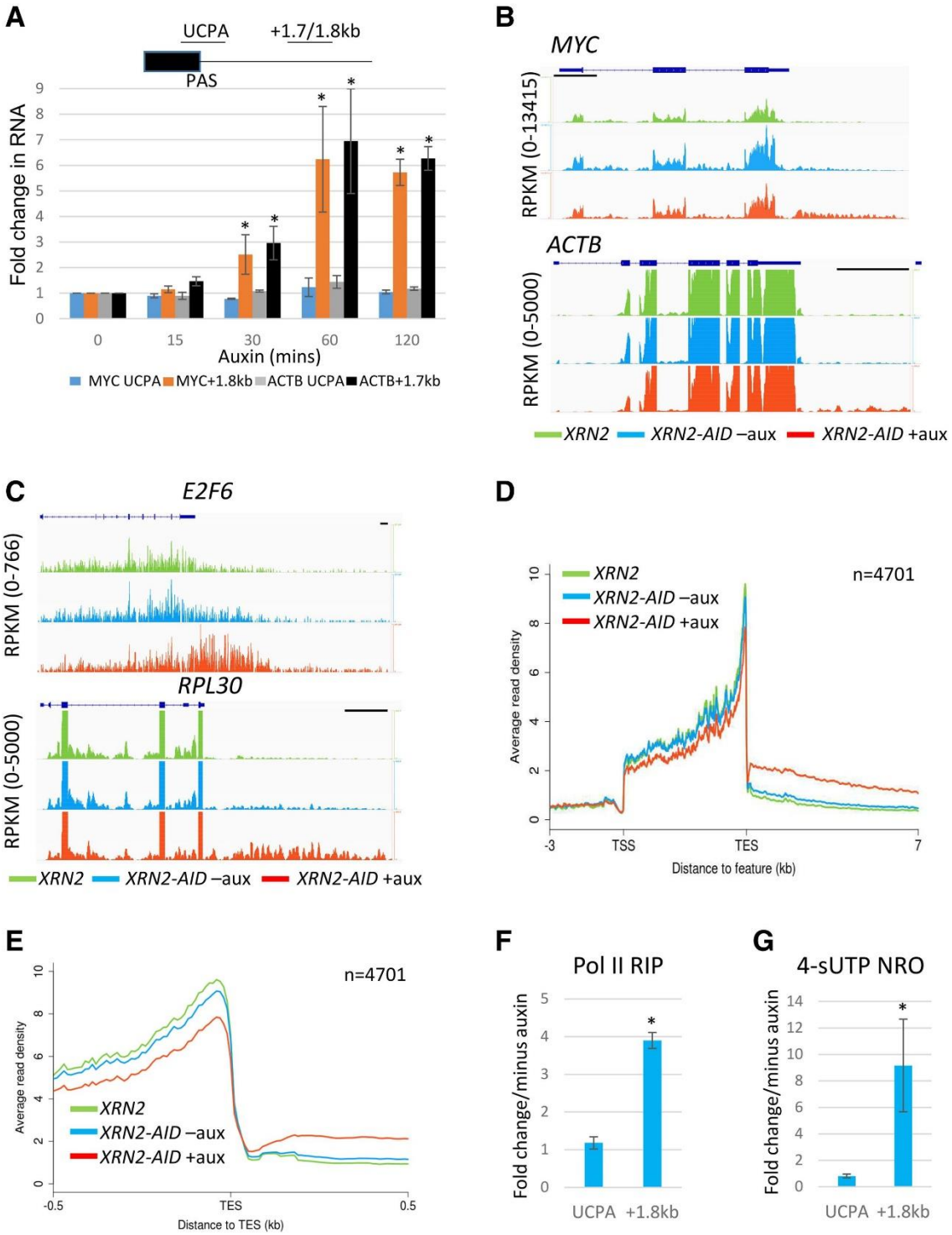

A
UCPA
+1.7/1.8kb
PAS
Fold change in RNA
Auxin (mins)
MYC UCPA
MYC+1.8kb
ACTB UCPA
ACTB+1.7kb
B
MYC
RPKM (0-13415)
ACTB
RPKM (0-5000)
XRN2
XRN2-AID –aux
XRN2-AID +aux
C
E2F6
RPKM (0-766)
RPL30
RPKM (0-5000)
XRN2
XRN2-AID –aux
XRN2-AID +aux
D
n=4701
XRN2
XRN2-AID –aux
XRN2-AID +aux
Average read density
TSS
TES
Distance to feature (kb)
E
n=4701
Average read density
XRN2
XRN2-AID –aux
XRN2-AID +aux
TES
Distance to TES (kb)
F
Pol II RIP
Fold change/minus auxin
UCPA
+1.8kb
G
4-sUTP NRO
Fold change/minus auxin
UCPA
+1.8kb

Figure 2

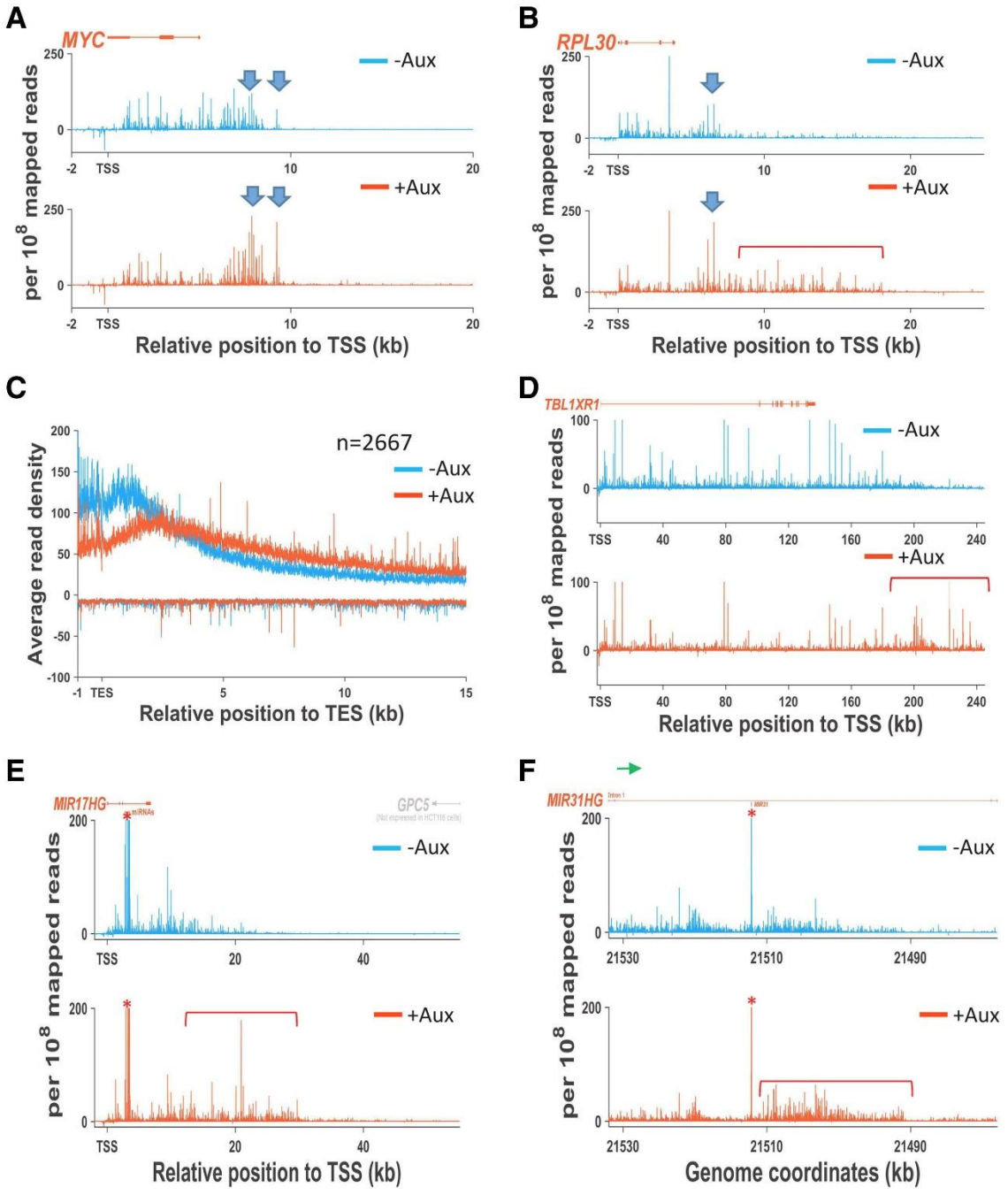

A
MYC
-Aux
+Aux
per 10^8 mapped reads
Relative position to TSS (kb)
B
RPL30
-Aux
+Aux
per 10^8 mapped reads
Relative position to TSS (kb)
C
n=2667
-Aux
+Aux
Average read density
Relative position to TES (kb)
D
TBL1XR1
-Aux
+Aux
per 10^8 mapped reads
Relative position to TSS (kb)
E
MIR17HG
GPC5
-Aux
+Aux
per 10^8 mapped reads
Relative position to TSS (kb)
F
MIR31HG
-Aux
+Aux
per 10^8 mapped reads
Genome coordinates (kb)

Figure 3

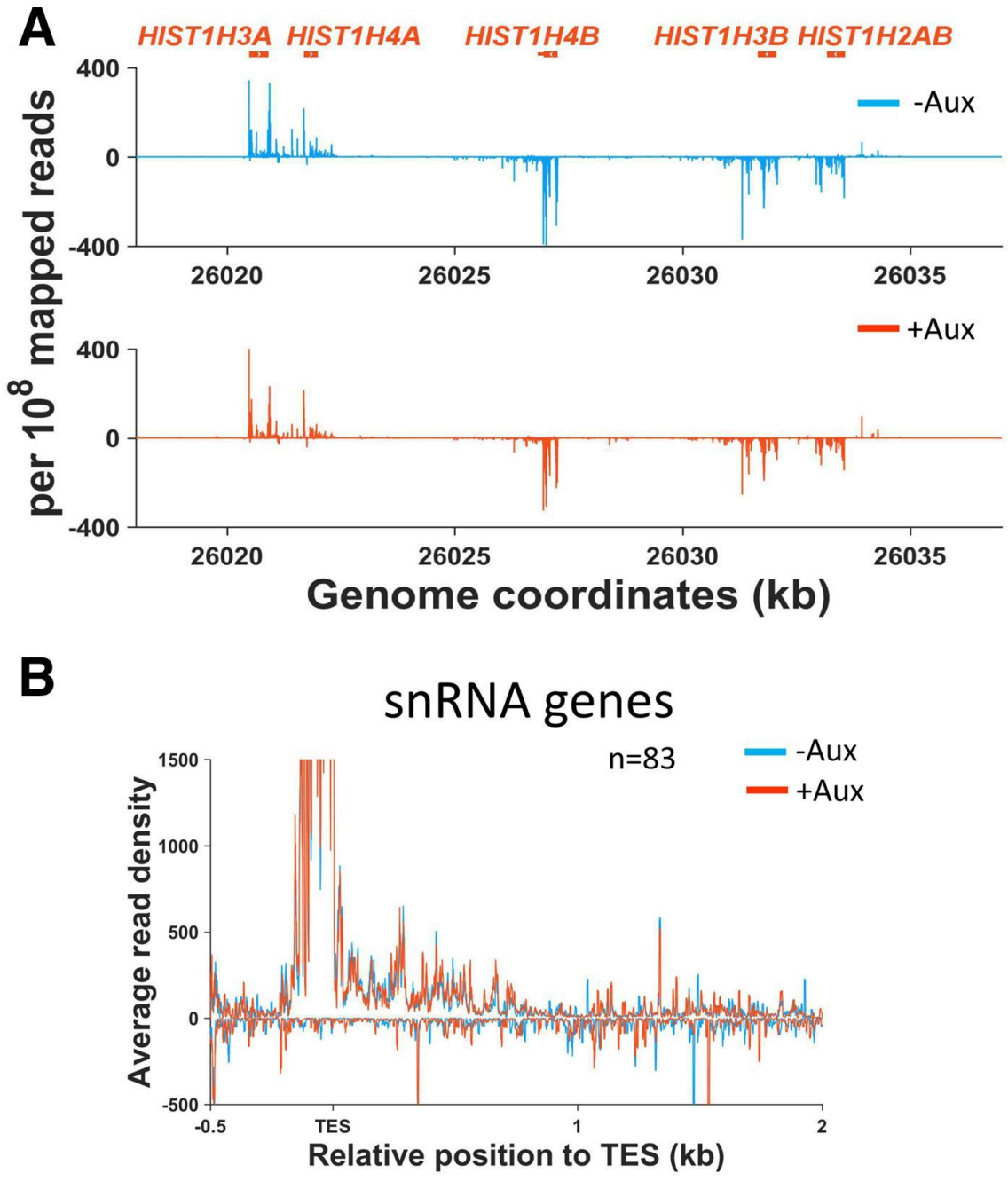

A
HIST1H3A
HIST1H4A
HIST1H4B
HIST1H3B
HIST1H2AB
400
0
-400
-Aux
+Aux
per 10^8 mapped reads
26020
26025
26030
26035
Genome coordinates (kb)
B
snRNA genes
n=83
-Aux
+Aux
1500
1000
500
0
-500
Average read density
-0.5
TES
1
2
Relative position to TES (kb)

Figure 4

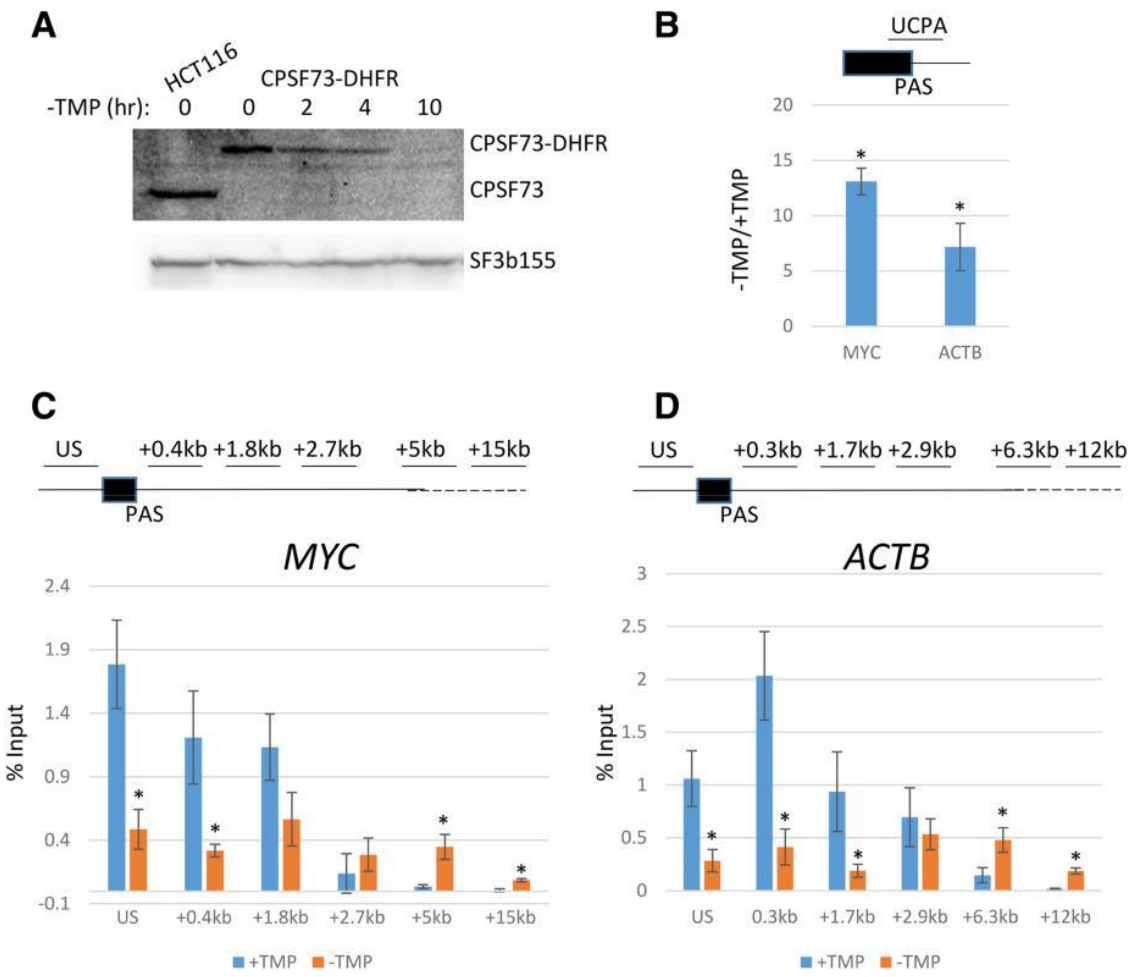

A
HCT116
CPSF73-DHFR
-TMP (hr): 0 0 2 4 10
CPSF73-DHFR
CPSF73
SF3b155
B
UCPA
PAS
-TMP/+TMP
20
15
10
5
0
MYC
ACTB
C
US
+0.4kb +1.8kb +2.7kb +5kb +15kb
PAS
MYC
% Input
2.4
1.9
1.4
0.9
0.4
-0.1
US
+0.4kb
+1.8kb
+2.7kb
+5kb
+15kb
+TMP
-TMP
D
US
+0.3kb +1.7kb +2.9kb +6.3kb +12kb
PAS
ACTB
% Input
3
2.5
2
1.5
1
0.5
0
US
0.3kb
+1.7kb
+2.9kb
+6.3kb
+12kb
+TMP
-TMP

Figure 5

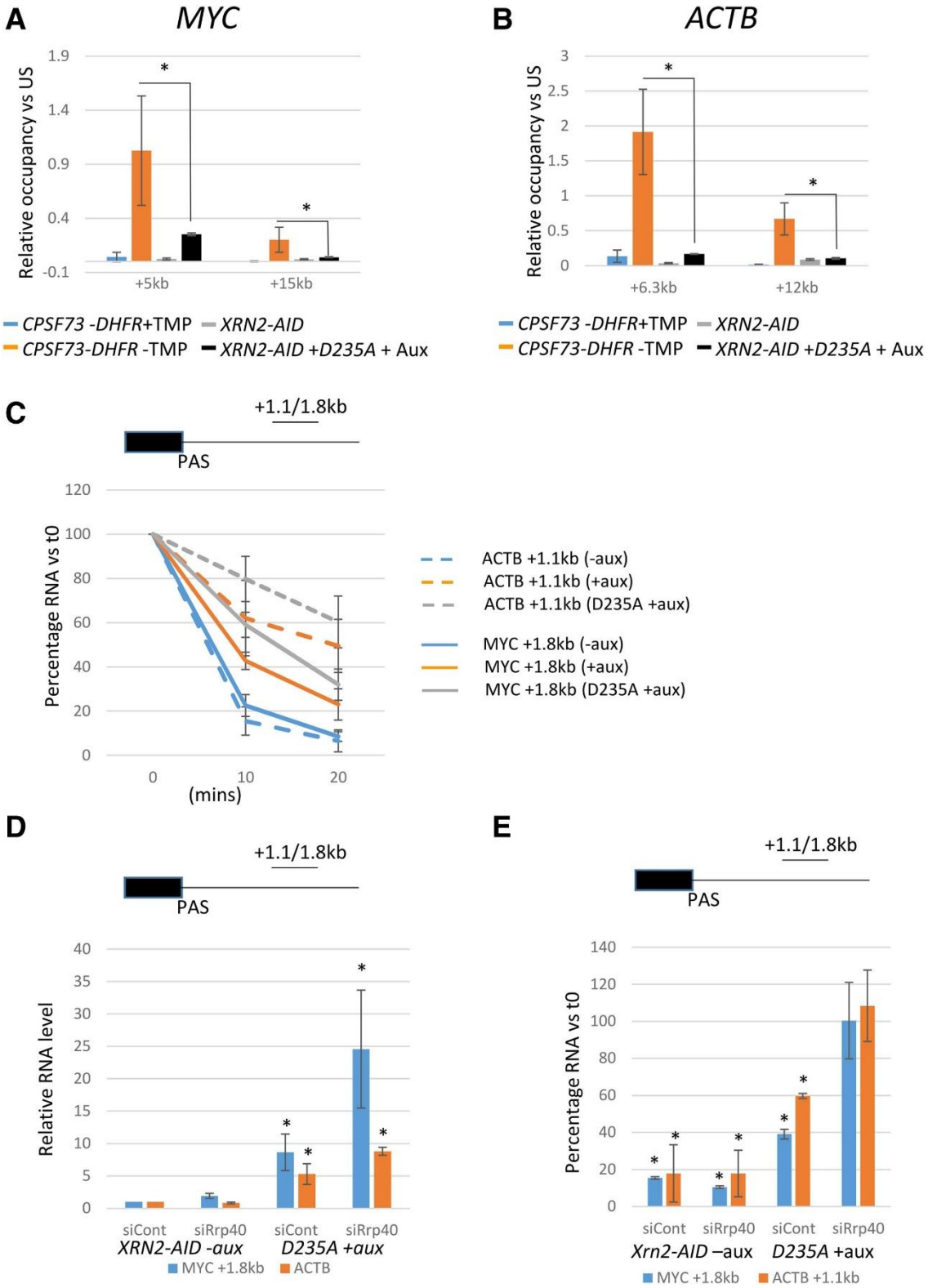

A
MYC
Relative occupancy vs US
1.9
1.4
0.9
0.4
-0.1
+5kb
+15kb
CPSF73 -DHFR+TMP
XRN2-AID
CPSF73-DHFR -TMP
XRN2-AID +D235A + Aux
B
ACTB
Relative occupancy vs US
3
2.5
2
1.5
1
0.5
0
+6.3kb
+12kb
CPSF73 -DHFR+TMP
XRN2-AID
CPSF73-DHFR -TMP
XRN2-AID +D235A + Aux
C
+1.1/1.8kb
PAS
Percentage RNA vs t0
120
100
80
60
40
20
0
0
10
20
(mins)
ACTB +1.1kb (-aux)
ACTB +1.1kb (+aux)
ACTB +1.1kb (D235A +aux)
MYC +1.8kb (-aux)
MYC +1.8kb (+aux)
MYC +1.8kb (D235A +aux)
D
+1.1/1.8kb
PAS
Relative RNA level
40
35
30
25
20
15
10
5
0
siCont
siRrp40
siCont
siRrp40
XRN2-AID -aux
D235A +aux
MYC +1.8kb
ACTB
E
+1.1/1.8kb
PAS
Percentage RNA vs t0
140
120
100
80
60
40
20
0
siCont
siRrp40
siCont
siRrp40
Xrn2-AID –aux
D235A +aux
MYC +1.8kb
ACTB +1.1kb

Figure 6

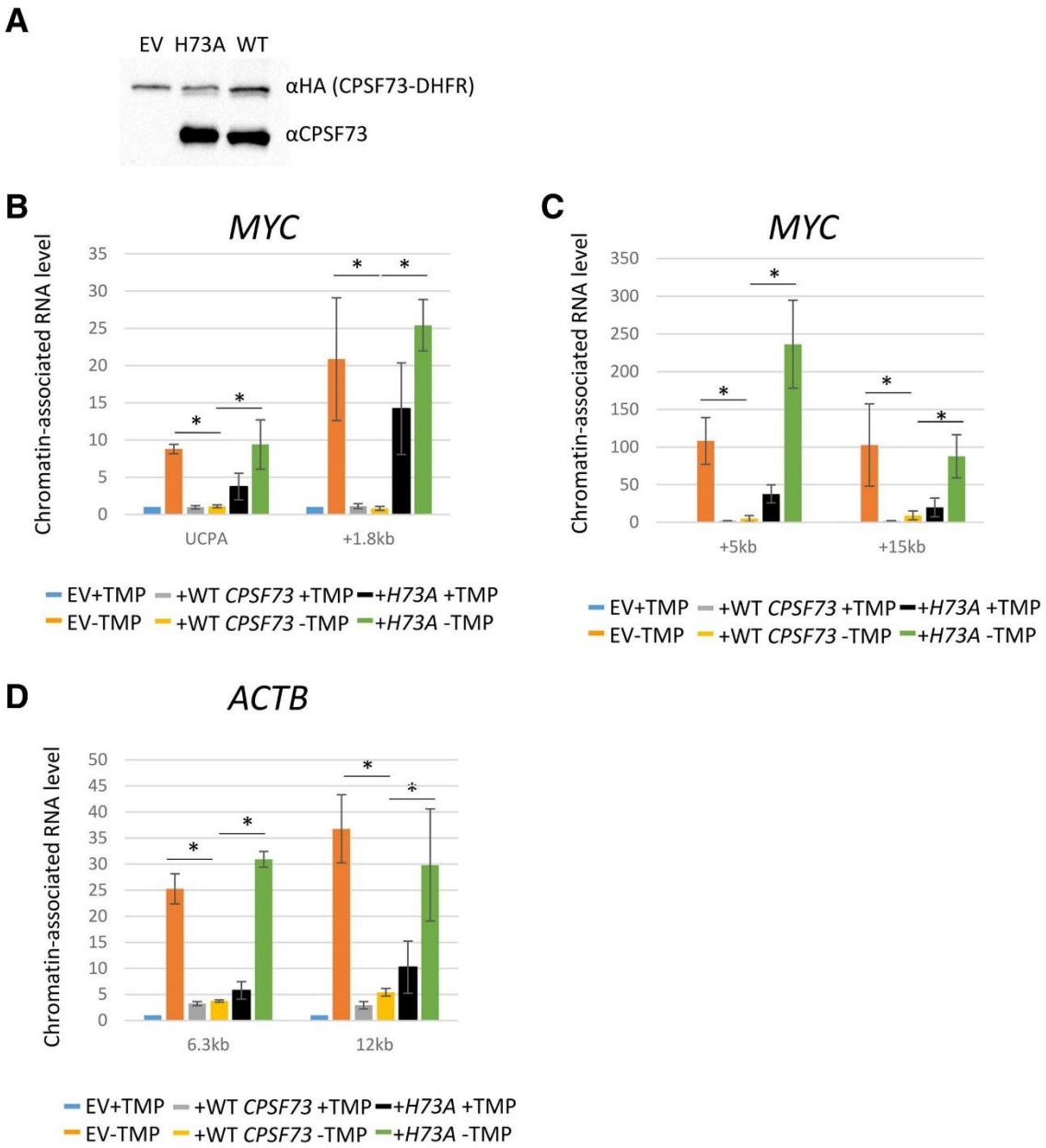

A
EV H73A WT
αHA (CPSF73-DHFR)
αCPSF73
B
MYC
Chromatin-associated RNA level
UCPA
+1.8kb
EV+TMP +WT CPSF73 +TMP +H73A +TMP
EV-TMP +WT CPSF73 -TMP +H73A -TMP
C
MYC
Chromatin-associated RNA level
+5kb
+15kb
EV+TMP +WT CPSF73 +TMP +H73A +TMP
EV-TMP +WT CPSF73 -TMP +H73A -TMP
D
ACTB
Chromatin-associated RNA level
6.3kb
12kb
EV+TMP +WT CPSF73 +TMP +H73A +TMP
EV-TMP +WT CPSF73 -TMP +H73A -TMP

Figure 7