THE UNIVERSITY OF WESTERN AUSTRALIA

Identifying the RNA targets of DBHS proteins to give insights into their role in gene regulation and paraspeckle formation.

Ellen Fortini
Bachelor of Science (Honours)

The School of Medicine and Pharmacology
The University of Western Australia
February 2014

This thesis is submitted in fulfillment of the requirements for the degree of Doctor of Philosophy.

Abstract

RNA binding proteins are important in many aspects of cellular function. In particular, RNA binding proteins regulate gene expression at multiple levels, a process that is de-regulated in many diseases. One family of RNA binding proteins, called DBHS (**D**rosophila **b**ehaviour and **h**uman **s**plicing) proteins are of interest as they regulate gene expression at the transcriptional and post-transcriptional levels. In addition, the three members of the DBHS protein family in mammals, Non-POU domain containing octamer binding protein (NONO), splicing factor proline/glutamine-rich protein (SFPQ) and paraspeckle protein component 1 (PSPC1), along with the long non-coding RNA (lncRNA) NEAT1 (Nuclear Paraspeckle Assembly Transcript 1) are involved in the formation of sub-nuclear structures called paraspeckles.

Paraspeckles are RNA:Protein complexes located in mammalian cell nuclei, and are unusual in that their formation and maintenance relies on the specific interaction between the paraspeckle proteins and NEAT1. Paraspeckles are thought to play a role in the regulation of gene expression via the sub-nuclear sequestration of specific proteins, including the DBHS proteins, to attenuate their function. In addition, paraspeckles are involved in binding and retaining specific RNAs in the nucleus as a means of post-transcriptional regulation.

At present there is a lack of knowledge about the molecular interactions occurring within paraspeckles between the RNA binding proteins and the RNAs they interact with. These interactions are the key to both the formation and function of paraspeckles, as well as the component proteins and RNAs.

In this project I optimized and applied PAR-CLIP (**P**hoto**a**ctivatable **r**ibonucleoside enhanced **c**ross**l**inking and **i**mmuno**p**recipitation) to isolate RNAs bound by the DBHS proteins NONO and SFPQ. The isolated RNA was sequenced and a bioinformatics analysis pipeline implemented with the program PARalyzer to identify, to single nucleotide resolution, the NONO and SFPQ binding sites in RNA.

This work has identified a periodic pattern of NONO and SFPQ binding sites along NEAT1_v2, possibly a reflection of the proteins oligomerization along this architectural RNA. This has given important insights into the internal organization of the paraspeckle complex that can also be applied to understanding other lncRNA:Protein complexes. In addition, this work has revealed NONO and SFPQ bind to several biologically relevant lncRNAs, presenting novel new functions for NONO and SFPQ in the regulation of gene expression. Possibly these proteins are regulating the cellular functions of their target lncRNAs. Alternatively the proteins may themselves be regulated by the lncRNAs they bind. Lastly, this study has identified a number of protein-coding RNAs bound by NONO and SFPQ. This work identified a preference for these proteins in binding in long first introns, possibly to stabilize the nascent RNAs as they are made. In addition, a number of the NONO and SFPQ bound RNAs encode important molecules in cancer progression and a number encode paraspeckle proteins, implicating NONO and SFPQ in important diseases and stress response pathways.

In summary, this work is the first detailed, transcriptome wide analysis of NONO and SFPQ RNA binding. Understanding not only the interactions occurring within paraspeckles, but also their interactions with other lncRNAs and mRNAs is the first step towards therapeutically targeting these complexes. This large scale identification of binding targets of two disease relevant RNA binding proteins will direct future work into the functions of this binding and therapeutic management of it.

Acknowledgements

Over the last 3 years 10 months and 18 days there has been a constant stream of people who have left their mark on my life. You helped me, you taught me, you made life fun. I am a firm believer that life is not about the things you do or the things you have, it is about the lives you touch. I can only hope that I have been as important in all of your lives as you have been in mine.

To my supervisors, Archa and Charlie, thank you for everything you have taught me. Archa thank you for being so available for all manner of things, from offering motivation and inspiration, to helping me with protocols, making me focus and giving me direction. Most of all though thank you for the genuine care and kindness you have shown me. Charlie thank you for your patience while teaching me to talk to the computer and the enthusiasm and excitement you have for the world of science. I could not have found two better teachers or nicer people to work for and I will be forever grateful.

Aside from the guidance from my supervisors I was blessed to have three amazing mentors in the lab for each of my three years. You have been my life mentors, science teachers and most of all my cherished friends. Sven, Luciano and Simon, I would have not made it this far without you. Sven thank you for everything you taught me the first year of my PhD, for taking me under your wing and making WAIMR such a fun place to come to work, can't wait to have the pleasure of visiting Sven's lab. Luciano thank you for seeing me through the infamous second-year-of-PhD tough times. Your amazing work ethic and clear-headed approaches to research never ceased to amaze me and is something I continue to aspire to. Simon thank you for listening to all my bla bla bla and stories of college shenanigans this past year. Sorry for the times I accidently gave you early (wrong) versions of protocols and thank you for every bit of support, skill and friendship you have shown me in this last challenging and fun year. You three are true scientists, and the world is a better place for it.

To some of the best friends I could ever hope for, the people I chose and the ones who chose me. Though you may not exactly understand what I do, you have

nonetheless commiserated the failed experiments and celebrated my successes. Jade, you have been with me since kindy and are still supporting me through my never-ending education journey. You are one of the strongest people I know and an inspiration to me. Stacey thanks for being an ear for all my worries and complaints through the ups and downs. Thank you for making me a part of the beautiful family you have built, it has been my absolute privilege. Hannah, your support has extended across the country, proving how strong it is. Your positive outlook and sense of fun has helped me through hard times ever since the Brisbane floods of 2010. Shimmy my voice of reason and partner in crime, never underestimate the impact having you in my life has had. From Ikea dates to pie making, talking science and life, thanks hubby. Pauline, I have told you many times you have created a monster, and I don't really have the words to express how grateful I am to you for that. Loving the life you have and living the life you love is not something many people get, but the opportunities and inspiration you have given me have made that my reality.

And lastly, to my family. You are the most important and cherished thing in my life, yesterday, today and forever. Nanna and Grandad, Nonna and Nonno, I have been so lucky to have had you around while I was growing up and beyond. Thank you for being proud of me, which has, in turn, inspired me to live a life I am proud of. My brothers, Andy and Daniel, two pretty awesome humans. Andy, your free spirit (hippy!), wisdom and courage are an inspiration to me. When ever in doubt I ask myself, "WWAFD?" You are the second most intelligent person I know… it must have been all the espresso we had as babies. Daniel, I have never known a more caring person with a bigger heart. You are always there to help us and look out for us and I love you for that. I am very proud to be your sister. And finally, to my Mum and Dad. I know you say we didn't get our brains from you but I beg to differ….Believe me…I'm a scientist. You gave us all the love and care we would ever need. You gave us opportunities but also helped us find and make our own. You gave us freedom to take our lives in the direction we desired, never intervening but always ready to let us discuss it and run ideas past you. I am the person I am because of you.

Table of Contents

I. Abbreviations

(v/v)	Volume/volume
(w/w)	Weight/weight
4-SU	4-Thiouradine
A	Adenosine
A-to-I	Adenosine to Inosine
ADAR	Adenosine deaminase that act on RNA
ALS	Amyotrophic lateral sclerosis
BASP1	Brain abundant, membrane attached signal protein 1
BSA	Bovine serum albumin
C	Cytosine
CCAT1	Colon cancer associated transcript 1
cDNA	Complementary DNA
CLIP	Crosslinking and immunoprecipitation
CPSF6	Cleavage and polyadenylation specificity factor subunit 6
CTN RNA	CAT2 transcribed nuclear RNA
DAZ	Deleted in azoospermia
DAZAP1	DAZ associated protein 1
DBHS	Drosophila behaviour and human splicing
ddH2O	Double distilled H2O
DNA	Deoxyribonucleic acid
dNTPs	Dinucleotide triphosphates
DRB	5,5-dichloro-1,3-D-ribofuranosylbenzimidazole
dsRNA	Double stranded RNA
DTT	Dithiothreitol
EDTA	Ethylene diamine tetraacetic acid
EGFP	Enhanced green fluorescent protein
EGTA	Ethylene glycol tetaacetic acid
eIF2α	Eukaryotc initiation factor 2 alpha
eRNAs	Enhancer RNAs
FMRP	Fragile X mental retardation 1

FPK	Fragments per kilobase
FTLD	Frontotemporal lobar degeneration
FUS	Fused in sarcoma
Fwd	Forward
G	Guanine
GNB1	Guanine nucleotide binding protein
GPI	Glycosylphosphatidyliositol
HFM1	Helicase family member 1
HNRNPK	Heterogeneous nuclear ribonucleoprotein K
HOTAIR	HOX transcript antisense RNA
HOTTIP	HOXA transcript at the distal tip
HOXC	Homeobox C
iCLIP	Individual nucleotide resolution crosslinking and immunoprecipitation
IGF1R	Insulin growth factor 1 receptor
IP	Immunoprecipitation
IR Alus	Inverted repeat Alu
kD	Kilo daltons
LARP4B	La ribonucleoprotein domain family, member 4B
Lin28	Lin-gene 28
LincRNA	Long intergenic non-coding RNA
lncRNAs	Long non-coding RNAs
LSINCT5	Long stress induced non-coding transcript 5
MAFG	V-maf avian musculoaponeurotic fibrosarcoma oncogene homolog G
MALAT1	Metastasis associated lung adenocarcinoma transcript 1
MATR3	Matrin 3
mCAT2	Mouse cationic amino acid transporter 2
min	Minutes
MOPS	3-N-morpholino-propanesulfonic acid
mRNA	Messenger RNA
Mw	Molecular weight
ncRNA	non-coding RNA
NEAT1	Nuclear Paraspeckle Assembly Transcript 1Nuclear enriched abundant transcript 1

NONO	Non-POU domain containing octamer binding protein
NOPS	NONA/paraspeckle domain
NUDT21	Nucleoside diphosphate linked moiety X type motif 21
NUFIP2	nuclear fragile X mental retardation protein interacting protein 2
p54NRB	Nuclear RNA binding protein 54kDa
PAGE	Polyacrylamide gel electrophoresis
PANDA	P21 associated ncRNA DNA damage activated
PAPD7	PAP associated domain containing 7
PAR-CLIP	Photoactivatable ribonucleoside enhanced crosslinking and immunoprecipitation
PBS	Phosphate buffered saline
PCR	Polymerase chain reaction
PDE3A	Phosphodiesterase 3A
Poly I:C	Polyinosinic:polycytidylic acid
PPEF2	Protein phosphatase, EF- hand calcium binding domain 2
PPIA	Peptidylprolyl isomerase A
PSF	Polypyrimidine tract-bindigng protein-associated-splicing factor
PSP1	Paraspeckle protein 1
PSPC1	Paraspeckle protein component 1
PTBP1	Polyprimidine tract binding protein 1
RBM14	RNA binding motif protein 14
Rev	Reverse
RNA	Ribonucleic acid
RNAi	RNA interference
RNPS1	RNA binding protein S1
RRM	RNA recognition motif
RYR2	Ryanodine receptor 2
scaRNAs	Small cajal body associated RNAs
SDS	Sodium dodecyl-sulfate
SEC14L	Ubiquitous tocopherol associated protein 1
SELEX	Selection of evolution by exponential enrichment
SEM	Standard error of the mean
SFPQ	Splicing factor proline/glutamine-rich protein

SINES	Short interspersed nuclear elements
SNP	Single nucleotide polymorphism
snRNPs	Small nuclear ribonucleic proteins
SR proteins	Ser/Arg containing proteins
SRm160	Ser/Arg-related nuclear matrix protein of 160 kD
STAU1	Staufen 1
T	Thymine
TDP-43	TAR DNA binding protein 43
TINCR	Tissue differentiation-inducing non-protein coding RNA
TP53INP1	Tumor protein 53 induced nuclear protein 1
Tris	Tris-hydroxymethyl-aminomethane
UCSC	University of California, Santa Cruz
UTR	Untranslated region
WDR5	WD repeat domain 5
XIST	X-inactive specific transcript

II. Figures

III. Tables

1. Introduction

1.1 Regulation of gene expression

The expression of the DNA in our cells is a highly controlled, complex multistep process that is responsible for making us who we are. However, humans are 99.9% identical at the genome level [1]. Our genome is 98% identical to that of gorillas [2] and our protein coding DNA is 85% identical to that of mice [3]. Our DNA instructions are therefore quite similar, but the way the instructions are carried out, that is, the regulation of the genome, holds the key to our complexity.

This regulation of the genome, the turning on or off of different genes at different times, and controlling how much of the gene product is made, is a phenomenon called the regulation of gene expression. The expression of the information encoded in our genomes is a highly controlled and complex process. Regulation of gene expression occurs at many levels, from the accessibility of the chromatin for transcription factors to bind, recruitment of the transcriptional machinery, co-transcriptional processing of the transcript, through to RNA localization, degradation and translation. Each of these steps is a massively complex network of interactions and processes.

In this study I focus on the post-transcriptional control of gene expression mediated by RNA binding proteins.

1.1.1 RNA binding proteins are key molecules in the regulation of gene expression

RNA binding proteins are essential molecules in the cell for the regulation of gene expression. They form complexes with each other, with other proteins as well as a variety of coding and non-coding RNAs to modify their expression or downstream action.

7.5% of all protein coding genes in the human genome contain one or more recognized RNA binding domains, suggesting the ability to bind RNA is widespread and important [4]. There are many different RNA binding domains in human proteins. These include, but are not limited to, RNA recognition motifs (RRMs), K-

homology domains (HNRNPK homology domain), PAZ (Piwi Argonaut and Zwille) domains and S1 domains (Ribosomal protein S1 domain) [4,5]. A number of studies on RNA binding domains indicates that individually they often only recognize short stretches of nucleotides (~2-7nt) and thus may not achieve precise RNA binding specificity [6,7]. For this reason, proteins often contain multiple RNA binding domains or form complexes with other proteins to bring together multiple binding domains to achieve target specificity [5]. In addition, the binding of RNA by proteins containing well-recognised RNA-binding domains may represent only a fraction of the proteins that actually do bind RNA in the cell. A recent unbiased study of all human Protein:RNA interactions revealed many proteins lacking well-recognized RNA binding domains nevertheless bind RNA in the cell [8]. Thus the full-gamut of regulatory roles played by RNA binding proteins is yet to be revealed.

1.1.2 Non-coding RNAs in the human genome

One of the exciting roles for RNA-binding proteins that has emerged over the past decade is their pivotal importance in binding non-coding RNAs (ncRNAs). The importance of non-coding RNA was one of the major surprises of the human genome project: when it was revealed that only ~2% of our DNA had protein coding potential [9,10]. This was game changing for the field of molecular biology, as up to this point it had been widely accepted that there were upwards of 50,000 genes [11,12] responsible for human complexity and that proteins were the principal functional molecules in the cell [13] . Some scientists argued the accumulation of 'junk' DNA sequences was evolutionary debris [14], however, technological advances showed the genome was pervasively transcribed, with upwards of 80% of the DNA being transcribed to RNA [15-18]. Whilst there is still controversy about the proportion of this ncDNA/ncRNA that is functional [19] many scientists have turned to characterizing the function of this pervasive transcription and the role these ncRNAs are playing in the cell.

It is one school of thought that these non-coding sequences may be responsible for organism complexity, as, unlike genome size and protein-coding gene number, the amount of ncDNA scales with increasing complexity [20]. Further, less than 1% of the single nucleotide polymorphisms (SNPs) identified to date result in a change in the

amino acid sequence of a protein, suggesting regulatory changes or ncRNA changes may be the basis of SNP functionality [10]. Additionally, variation in ncDNA is highly correlated with complex diseases, for example, 80% of cancer-associated SNPs identified in genome wide association studies are located in introns and intergenic regions and therefore possibly contribute to the regulation of these transcripts [21].

MiRNAs (micro RNAs) were among the first functional ncRNAs to be described. Originally found in *C. elegans* [22], miRNAs have since been attributed to a variety of human diseases through their ability to regulate the expression of genes post transcriptionally. MiRNAs bind via complementary base pairing to target transcripts and facilitate their degradation by the RISC (RNA induced silencing) complex [23]. De-regulation of miRNA expression lies at the heart of many types of cancers and other diseases, with these ncRNAs also regulated by a repertoire of RNA binding proteins [24]. However, the work presented here focuses on mRNAs, and another type of ncRNA, the lncRNAs, and, as such, these molecules are the focus of this introduction.

1.1.3 Long non-coding RNAs are an important class of molecules in the cell

Long non-coding RNAs (lncRNAs) are newly identified players in the network of gene expression regulation and often form regulatory complexes with one or more RNA binding protein [25,26]. These molecules, just starting to be catalogued and characterized, can influence many levels of the gene expression regulation network, with new functions being found for them every day [27].

LncRNAs are arbitrarily classified as ncRNA transcripts longer than 200nt that can arise between protein-coding genes (intergenic), within the introns of coding genes, overlapping coding genes in the sense orientation, or anti-sense to coding genes [25,28]. Conservation of lncRNA promoters through evolution [29], secondary structure constraints [30] and high stability [31] all suggest that lncRNAs have the potential to perform important functions in the cell.

LncRNAs, like protein coding transcripts, are tightly regulated, some have very low expression levels and are differentially expressed in different cell types and at

different stages of development [29]. This differential expression of lncRNAs is also seen in many diseases, including cancer and in some cases have become the hallmarks of these diseases [32-36].

1.1.4 lncRNAs contribute to regulation of gene expression at multiple levels

The field of lncRNA biology has grown rapidly in the last decade, and we can now appreciated that some common themes in lncRNA action are emerging. It has been proposed that there are 4 key mechanisms lncRNAs employ in the cell to regulate gene expression, acting as signals, decoys, guides and scaffolds [37]. Each of these are discussed below.

1.1.4.1 LncRNAs act as signals

It has long been known that chromatin extracts contain abundant RNA, sometimes matching DNA content [38-40]. LncRNAs may represent a large fraction of chromatin-associated RNA as they interact with chromatin at specific sites, and in doing so, act as signals for other molecules, such as histone modification complexes or transcription factors. The well-known process of X-chromosome inactivation is achieved in part by the lncRNA XIST binding to the X-chromosome resulting in transcriptional and epigenetic silencing [41,42].

1.1.4.2 LncRNAs act as decoys

Some lncRNAs can act as decoys to draw genome regulation machinery away from their target chromatin. For example, the lncRNA PANDA (P21 associated ncRNA DNA damage activated) is induced with DNA damage and forms a complex with the chromatin modifying protein NY-YA. This complex formation prevents NY-YA from activating the genes that would normally allow the cell to begin the process of apoptosis [43].

1.1.4.3 LncRNAs act as guides

lncRNAs can guide proteins to otherwise non-targeted transcripts, resulting in stabilization or destabilization of the target RNA. This is best illustrated by the

effect two lncRNAs have on the protein Staufen 1 (STAU1) which facilitates STAU1 mediated mRNA decay [44]. STAU1 binds double stranded RNA (dsRNA) and can be directed to degrade certain mRNAs that are tethered to a novel lncRNA by complementary base pairing [45]. In contrast, STAU1 binds another lncRNA, TINCR and together they stabilize mRNAs that would otherwise be targets of STAU1 [46]. This stabilization results in the accumulation of mRNAs that control differentiation in human epithelial cells [46] and illustrates how lncRNAs can not only direct proteins to target genes but can also influence the function that protein performs at the target.

1.1.4.4 LncRNAs act as scaffolds.

In some cases lncRNAs can serve as the structural scaffold for large multi-protein complexes, with some of these complexes large enough to form a subnuclear structure. Acting in this way, lncRNA bind and bring together multiple molecules that can act on target genes or transcripts. LncRNAs as scaffolds is an emerging field, with the novel lncRNA NEAT1, the structural scaffold of paraspeckles the first discovered to recruit proteins into a sub-nuclear complex [47].

Other lncRNAs can form scaffolds to bring together the protein subunits of regulatory complexes and localize them to their sites of action. The lncRNA HOTAIR (Homeobox transcript antisense intergenic RNA) acts in this way. HOTAIR contains a number of binding sites for different chromatin modifying complexes including PRC2 (polycomb repressive complex 2), LSD1 (Histone lysine demethylase), CoREST (co-repressor for elements-1-silencing transcription factor) and REST (repressor for elements-1-silencing transcription factor) [48]. HOTAIR not only forms a scaffold along which these proteins organize but also acts as a bridge to bring these chromatin modifying complexes to target sites in the chromatin. Though direct interactions with the chromatin [49], HOTAIR actively recruits the bound chromatin modifying complexes to their target genes where they trigger histone methylation or demethylation to regulate their expression [50,51].

Many studies utilizing new and high throughput technologies are detecting lncRNAs that can fit into one of the above categories, however it is likely that many more functions will also emerge with time.

1.1.5 Long non-coding RNAs in cancer

LncRNAs are emerging as important regulators of gene expression in a variety of diseases, including cancer. Microarray analysis and RNA deep sequencing of multiple cancerous and healthy tissue samples has been used to identify lncRNAs that may be responsible for the development, progression or prognostic outcome of cancers [52-55]. The ability to form complexes that can alter the expression of target genes is a property that has been observed for many of the lncRNAs identified in these profiling experiments. A hallmark of cancer is the mis-regulation of expression of key genes, and this may be partially caused by aberrant expression and function of the gene regulatory lncRNAs. There are several examples of specific lncRNAs that have been identified and characterized for their role in cancer.

HOTAIR, though originally identified for its role in regulating the expression of genes at the HOX locus [56] now also has a well documented role in the development and progression of a number of cancers. As detailed above, HOTAIR forms a scaffold along which a number of chromatin modifying complexes bind and recruits them to specific chromatin sites where they facilitate histone modification[48]. Over-expression of HOTAIR in human breast tumours and breast cancer cell lines correlates with an increase in metastasis and invasiveness of the cancer cells [57]. HOTAIR over-expression has also been attributed to the increased metastasis and tumor cell invasion that is observed in colorectal and pancreatic cancer [51,58]. The genes differentially regulated by HOTAIR vary in different cancers[58] however the general mechanisms by which HOTAIR induces oncogenesis are thought to be similar. Under this model, HOTAIR over expression increases the recruitment of the chromatin modifying complexes to tumor suppressor genes and silences them [57].

MALAT1 (Metastasis associated lung adenocarcinoma transcript 1) is a well characterized lncRNA that has been implicated many cancers including colorectal [59], gall bladder [60], cervical [61] and liver cancer [62]. MALAT1 was first discovered through its correlation with metastasis in a study comparing the transcriptomes of metastatic potential in primary lung tumors [63]. MALAT1 knockdown results in dis-

regulation of several genes involved in metastasis, making it a potential diagnostic and prognostic marker and therapeutic target [64]. MALAT1 binds to the splicing factor SRSF1 and regulates its phosphorylation within nuclear speckles [65]. MALAT1 is also a key regulator of cell cycle progression with its over expression in cancer cells resulting in hyper proliferation [66,67].

LncRNAs can be found in both the nucleus and cytoplasm of the cell, however, to date, many of the most well-characterized lncRNAs with roles in gene regulation are nuclear. The fields of nuclear cell biology and lncRNA have become intertwined in several respects, as outlined below.

1.2 Nuclear organisation

The eukaryotic cell nucleus is highly organized with chromosomes occupying discrete territories and many sub-nuclear bodies or 'compartments' [68]. It is thought that sub-nuclear compartments allow the various nuclear functions to be performed in an organized and controlled way by forcing local high concentrations of key effector molecules. For example Polycomb bodies bring together polycomb group proteins and localize them to polycomb regulated genes [69]. Many of the sub-nuclear compartments contain lncRNAs, which either form the structural scaffold of these compartments or localize to them [70]. These lncRNA sub-nuclear compartments are summarized in figure 1.1 and include nuclear speckles enriched in MALAT1, nuclear stress bodies forming around repetitive satellite III ncRNA, nucleoli forming around ribosomal RNA genes, gomafu speckles, omega speckles found in *Drosophila*, and the structure relevant to this thesis, paraspeckles.

1.3 Paraspeckles

Paraspeckles were first discovered when an analysis of novel nucleolar proteins unexpectedly found one that was not nucleolar and instead localized to 10-20 distinct foci in the nucleus [71]. This protein, named paraspeckle protein component 1 (PSPC1) was observed in foci adjacent to nuclear speckles, hence this new sub-nuclear compartment was named 'paraspeckles'. Paraspeckles are mammalian

specific and are observed in both primary and transformed cell lines [71] and in many tissues throughout mammals [72] but not in embryonic stem cells or pluripotent cells [73]. Paraspeckles were identified based on the localization of the PSPC1 protein, but are now known to contain up to 40 nuclear RNA binding proteins. Paraspeckle integrity and formation is dependent on active transcription, as incubation of cells with transcriptional inhibitors such as Actinomycin D or D-ribofuranosylbenzimidazole (DRB) result in paraspeckle disintegration and subsequent re-localization of paraspeckle proteins to the nucleolus [71]. Paraspeckles persist in the cell through interphase and for most of mitosis [74], and are largely non-mobile throughout the nucleus unlike many other nuclear bodies [75]. There are typically 2-20 paraspeckles in the average HeLa cell [71] and they have varying lengths, ranging from 0.3 to 2 μm long but a uniform width of 300 nm in human cells [76].

1.3.1 Paraspeckle structural components

1.3.1.1 NEAT1 architectural RNA

The observations that paraspeckles are dependent on active transcription and that RNA depletion with RNase A dissolves paraspeckles [71,77] led to the search for the RNA or RNAs that were essential for paraspeckle formation and maintenance in the cell. In 2009, three groups reported that the lncRNA NEAT1 (first called Nuclear enriched abundant transcript 1, or Men ε / β, now known as Nuclear Enriched Paraspeckle Assembly Transcript 1) is the essential RNA structural backbone of paraspeckles [47,78,79]. Ablation of NEAT1 with siRNA in cell lines results in loss of paraspeckles [47,80] and NEAT1 deletion in the mouse genome results in a loss of paraspeckles [81].

NEAT1 is expressed as two isoforms that overlap at their 5' ends, both generated from the same promoter on human chromosome 11 [82]. The short isoform, NEAT1_v1 (also called Men ε) is 3,700nt long and the long isoform, NEAT1_v2 (also called Men β) is 23,000nt [78]. The NEAT1 gene is located on chromosome 19 in mouse and has a conserved gene structure also giving rise to two isoforms from the same 5' promoter, with mNeat1_v1 being 3,100nt in length and mNeat1_v2

20,000nt long (Figure 1.2). Strikingly, despite it's role as the structural scaffold for paraspeckles in both mice and humans [47,72], the NEAT1 primary sequence is highly divergent. Publicly available RNA-sequencing data shows that the majority of the NEAT1_v1 and v2 RNA exists as single-exon unspliced transcripts, although there is some evidence of a short intron within NEAT1_v1 being spliced out in some tissues.

NEAT1_v1 is generated when a natural polyadenylation (polyA) site is utilized, resulting in transcriptional termination, cleavage and polyadenylation of the transcript [81]. Due to a complex interplay of different paraspeckle proteins, including HNRNPK and the cleavage factors NUDT21 and CPSF6, in some instances this polyA site is missed, resulting in extension of the transcript and generation of NEAT1_v2. NEAT1_v2 contains a cleavable triple helical structure at its 3' end [79,83] that confers transcript stability and nuclear retention [84].

Two key experiments demonstrated that NEAT1_v2 is the essential isoform for paraspeckle formation: firstly, specific reduction of NEAT1_v1 levels by siRNA inhibition of the cleavage factors did not affect paraspeckles [81], and secondly, rescue of cells lacking paraspeckles due to NEAT1 deletion could only be achieved with overexpression of NEAT1_v2, not NEAT1_v1 [72].

1.3.1.2 Paraspeckle Proteins

NEAT1_v2 now has an established role as the architectural backbone of paraspeckles. However, many paraspeckle proteins also play essential roles in paraspeckle formation, indicating it is the combination of both an architectural RNA and it's binding proteins that are necessary to build a paraspeckle. PSPC1 was the first paraspeckle protein. However, soon after the discovery of paraspeckles, it was established that two other proteins with high sequence similarity to PSPC1 were also enriched in paraspeckles: these are NONO (non-POU domain containing octamer binding protein) and SFPQ (splicing factor proline/glutamine-rich protein) [71,77]. Together, these three proteins make up the mammalian members of the DBHS protein family (Drosophila Behaviour Human Splicing). Interestingly, PSPC1 is not required for paraspeckle formation, as siRNA knockdown of the protein in HeLa cells does not result in loss of paraspeckles [78]. However, the two

related protein, NONO and SFPQ, are essential for paraspeckle formation, as knocking them down with siRNA in HeLa cells results in the loss of paraspeckles, and a reduction in NEAT1_v2 levels [78].

Along with the DBHS proteins, there are now 40 proteins (listed in appendix 1) that have been observed to localize to paraspeckles [81]. These proteins were identified in a high-throughput screen of localization of fluorescent protein fusions of 18,467 human protein coding cDNAs, in which co-localization with SFPQ was used as the marker for paraspeckles [81]. The majority of proteins localizing to paraspeckles contain one or more RNA binding domains, either RRMs (RNA recognition motifs) or KH (K homology) domain. The fact that the majority of the paraspeckle proteins have the capacity to bind RNA provides strong evidence that this subnuclear compartment is intimately involved in the regulation of gene expression at the RNA level.

1.3.2 Paraspeckle formation and structure

In 2010 it was shown that active transcription of NEAT1 is needed to recruit paraspeckle proteins for the formation and persistence of paraspeckles. Mao and colleagues used an elegant gene array system, stably incorporating a tagged NEAT1 construct, to visualize nascent NEAT1_v2 transcripts and subsequent recruitment of proteins to form paraspeckles[85]. They showed that the accumulation of paraspeckle proteins co-localizing with the NEAT1 transcript occurs very soon after the initiation of NEAT1 transcription [85].

Once the proteins bind the RNA, the molecules then form an ordered structure. This spatial organization of the molecules has been particularly well-characterized for NEAT1. Electron microscopy coupled with in situ hybridisation targeting different regions of NEAT1 demonstrated that NEAT1_v1, and the 5' and 3' ends of NEAT1_v2, were located around the periphery of the paraspeckle, whilst the middle part of NEAT1_v2 was seen in the center of the paraspeckle [76]. From these observations, a model of NEAT1_v1 and NEAT1_v2 organization within the paraspeckle was proposed (Figure 1.3). This model accounts for the dependence on NEAT1_v2 for paraspeckle formation, showing NEAT1_v2 forms the core of the paraspeckle, whilst NEAT1_v1 packs around the periphery. Importantly, it is

thought that this organization is laid down by the binding of various paraspeckle proteins to both NEAT1_v1 and NEAT1_v2. As mentioned above, NONO and SFPQ are both essential for this process of paraspeckle formation. In addition, four more paraspeckle proteins are essential for paraspeckle formation in HeLa cells. Naganuma and colleagues showed that knockdown of NONO, SFPQ, HNRNPK, RBM14, DAZAP1 or FUS resulted in the loss of paraspeckles [81].

1.4 Paraspeckles function and mechanisms

To date, there have been two mechanisms characterized to explain the function of paraspeckles in gene regulation. Firstly, paraspeckles trap and retain certain types of RNAs in the nucleus, preventing their translation or downstream function [73,77,86]. Second, paraspeckles sequester proteins away from the rest of the nucleoplasm, thereby preventing their normal nuclear function [87,88].

1.4.1 Mechanism 1: Nuclear retention of A-to-I edited RNA.

Early studies on the fate of double-stranded viral RNAs showed that mammalian cells combat this RNA by subjecting it to adenosine-to-inosine (A-to-I) RNA editing and retaining it in the nucleus [89,90]. The ADAR (Adenosine deaminase that acts on RNA) family of enzymes are responsible for carrying out the A-to-I editing [91-93]. But what happens to the RNA once edited? Searching for the proteins that bind this edited RNA, Zhang and Carmichael purified a complex containing NONO, SFPQ and Matrin3 from HeLa extracts [90]. It subsequently emerged that in uninfected cells, there is nevertheless a significant amount of dsRNA produced, mostly derived from transcribed repetitive elements that form intra-transcript dsRNA regions. In human, the most abundant repeat is the Alu element (named for the Alu1 restriction site found within it), and many Alus are found in an inverted orientation (IR-Alu) within transcribed regions. Prasanth et al showed that NONO and SFPQ bind one such IR-Alu containing transcript, called CTN-RNA, and that CTN-RNA is localized to paraspeckles [77].

CTN-RNA is an alternative isoform of the cationic amino acid transporter 2 (mCAT2) mRNA transcript, with a longer 3'untranslated region (UTR) that contains the IR-Alu. The mechanism of nuclear retention involved NONO/SFPQ

binding the A-to-I edited IR-Alu, sequestering CTN-RNA within paraspeckles, thereby preventing its nuclear export [77]. When the cell was subsequently stressed, the long 3'UTR was cleaved off, resulting in export, translation, and increased mCAT2 protein levels.

How wide-spread might this mechanism be? There are 333 human genes with IR-Alus in their 3' UTRs [86]. IR-Alus from two of these genes (NICN1 and LIN28) can silence a fused reporter gene [86]. The concept that paraspeckles are key to nuclear retention of A-to-I edited IR-Alus was strengthened by the finding that siRNA against NEAT1, and subsequent loss of paraspeckles resulted in less efficient nuclear retention of IR-Alu containing RNAs [86].

1.4.1.1 Nuclear retention in paraspeckles is not the main mechanism of post transcriptional gene regulation for A-to-I edited IR-Alu containing transcripts.

Recent work has suggested that nuclear retention is less important than first thought in terms of regulating the expression of these IR-Alu transcripts. Rather, transcripts with A-to-I edited IR-Alus in their 3' UTR are often present in the cytoplasm and bound to ribosomes [94]. In addition, ADAR knockdown does not affect the translational output of cells, indicating that both A-to-I edited and non-edited IR-Alu containing genes are translated to the same extent [95]. New evidence suggests that the majority of the A-to-I edited IR-Alu mediated silencing is actually occurring in the cytoplasm on polysomes, via a translational inhibition mechanism [96]. STAU1 competes with NONO to bind A-to-I edited IR-Alus, and once STAU1 is bound it triggers nuclear export of these RNAs [96]. Once in the cytoplasm, STAU1 can be displaced by PKR (Protein kinase R) in turn mediating eIF2 α phosphorylation, leading to translational inhibition of the IR-Alu containing transcript [96]. Thus nuclear retention in paraspeckles is just one part of a long regulatory pathway for A-to-I edited IR-Alu transcripts, that may only be used in restricted cell types with low STAU1 levels.

1.4.2 Mechanism 2: Sequestration of proteins

Paraspeckles can also regulate gene expression by the sub-nuclear sequestration of specific proteins. This may prove to be a more generalizable mechanism than

nuclear retention as it is likely applicable to all paraspeckle-containing cell types. It was first observed that increasing NEAT1 levels led to larger paraspeckles [88]. The building of larger paraspeckles, consisting of both NEAT1 and paraspeckle proteins, led to a depletion of paraspeckle proteins from the nuclear pool [88]. The effect of this depletion was investigated for the essential paraspeckle protein SFPQ that has a well-characterised transcription factor activity [97-99]. Increased sequestration within paraspeckles led to altered levels of several SFPQ target genes, including IL8 (Interleukin-8), which is repressed by SFPQ, and ADARB2, which is up-regulated by SFPQ. The altered transcription was associated with a decrease in SFPQ promoter binding, as demonstrated by Chromatin-IP [87,88].

Looking at the list of 40 paraspeckle proteins (Appendix 1), it is easy to see how lowering the available nuclear pool of many proteins by paraspeckle sequestration could result in dramatically altered gene regulation. Several of these proteins are very well characterized and are involved in numerous biological pathways. However, beyond the role of SFPQ as a transcription factor [88], it is not known which other functions of paraspeckle proteins may be affected by this mechanism. In order to fully understand the effects of sequestration on paraspeckle proteins, it is important to understand the full function of these proteins, both inside and outside paraspeckles. For the purposes of this thesis, I have focused on the two essential paraspeckle proteins NONO and SFPQ, therefore their functions are discussed in greater detail in chapter 1.6.

1.5 The biological function of NEAT1/paraspeckles

1.5.1 NEAT1 knockout mouse reveals a role for paraspeckles in female reproduction

Nakagawa and colleagues created a Neat1 knockout mouse by insertion of the lacZ reporter gene immediately downstream of the *Neat1* transcription start site [72]. The homozygous mouse is viable, has dramatically reduced Neat1_v1 levels and undetectable Neat1_v2 levels and does not contain any cells with paraspeckles [72]. The Neat1 knockout mouse was first reported to have no gross phenotype.

Interestingly, the lack of an obvious phenotype for a ncRNA knockout is not restricted to NEAT1. The ncRNA BC1 knockout initially showed no gross

phenotypic differences from the wild types, however, when the mice were moved to an outdoor pen, they exhibited behavioral defects [100]. In addition, the Malat1 knockout mouse was shown to be viable [101-103]. These ncRNA knockout studies illustrate the need for more inventive and comprehensive phenotypic screening procedures for ncRNA knockouts, as, unlike mRNA knockouts, these molecules are often involved in fine-tuning gene expression in response to the environment, rather than development [104].

Two studies were published in 2014 exploring the Neat1 knockout phenotype more thoroughly. The first study found lack of Neat1 resulted in a defect in mammary gland formation [105]. The finding was triggered by the observation that pups were not surviving when being fed from Neat1-/- mothers. The lack of Neat1 was associated with defects in proliferation of epithelial cells that give rise to milk producing cells in the mammary glands, defective branching and duct elongation [105]. Whilst it is not possible to discern whether this phenotype is due to loss of Neat1 or loss of paraspeckles, the luminal epithelial cells of the mammary glands in wild type mice, where these defects stem from, do contain abundant paraspeckles [105].

The second study reported that only a small number of Neat1-/- females were able to sustain pregnancy, giving birth to only half as many pups as wildtype animals [106]. This defect was attributed to the post ovulatory ovary, which is responsible for building the corpus luteum and secreting progesterone, the spike in which maintains the pregnancy [106]. In wild type animals, Neat1_v2 is highly expressed in the cells that precede the corpus luteum (and paraspeckles are large and abundant), with the highest levels seen at the early stages of pregnancy [106]. The inability to form a functioning corpus luteum, and the consequent absence of progesterone secreted causes failure of pregnancy in Neat1-/- mothers [106]. However, in a subset of Neat1-/- mothers, these corpus luteum defects were not observed and pregnancy was normal. In light of this, the authors conclude that Neat1 is required in a subset of pregnancies, possibly due to the exposure to some kind of environmental condition [106].

1.5.2 Cancer

Two recent studies have implicated NEAT1 and paraspeckles in the determination of cancer severity, with up-regulation of NEAT1 resulting in a more aggressive disease phenotype. In prostate cancer, NEAT1 is highly up-regulated by Estrogen receptor alpha (ERα) [107]. Chromatin isolation by RNA purification (ChIRP) revealed NEAT1 at oncogene promoters [107]. Xenograft experiments showed that NEAT1 over-expression was associated with increased tumorigenesis, with reduced NEAT1 being associated with smaller tumor size [107]. Appearance of paraspeckles was confirmed in tumor biopsies but the mechanism of gene regulation for NEAT1 is not addressed in the paraspeckle context.

A study of breast cancer found that expression of NEAT1 was induced with hypoxia, resulting in more aggressive disease and a poor prognosis [108]. While the overall levels of NONO and PSPC1 (SFPQ was not evaluated in this study) do not change with hypoxia, their localization does, with these proteins moving from the nucleoplasm into paraspeckles. The increase in NEAT1 and the number of paraspeckles was observed to promote proliferation and inhibit apoptosis in the breast cancer cell lines studied [108].

1.5.3 Viral infection

NEAT1 is up-regulated in response to many viruses. One of the first reports of NEAT1 was as a novel transcript that was induced in mouse brain tissues infected with Japanese encephalitis virus and rabies [109].

In a more recent study, it was observed that influenza infection, or mock infection with PolyI:C dsRNA stimulates increased transcription of NEAT1, leading to increased expression of genes involved in the innate immune response [87]. As described in chapter 1.4.2, the mechanism used here to regulate gene expression is that more NEAT1 results in larger paraspeckles, with an increased sequestration of SFPQ, resulting in relief of SFPQ repression of the IL8 promoter.

There are also indicators that paraspeckles modulate viral infection at the post-transcriptional level. Zang et al. demonstrated that NEAT1 regulates production of virus particles by forming paraspeckles, thereby preventing nuclear to cytoplasmic export of HIV-1 INS containing transcripts [110]. When NEAT1 was knocked down, there were fewer paraspeckles and an increase in HIV-1 INS containing RNA in the cytoplasm, resulting in increased translation of the viral proteins and increased virus production [110]. In each of these viral studies there was no observed change in the overall levels of SFPQ, therefore, its ability to affect gene expression during viral infection is dependent upon its sub-cellular localization, which, in turn, regulates its nucleic acid binding [110].

1.6 The biological functions of the DBHS proteins

It is important to realize that formation of paraspeckles is but one of many functions for DBHS proteins. Indeed, whilst NEAT1 is only found in mammalian genomes, DBHS orthologues can be traced back to *C. elegans*. This suggests many paraspeckle-independent roles for these proteins.

1.6.1 Multifunctional proteins

DBHS proteins share a conserved 2D structure, consisting of two adjacent RNA recognition motifs (RRMs), a NONO-paraspeckle (NOPS) domain and an extended coil-coil region (Figure 1.4). In mammals, NONO, SFPQ and PSPC1 share over 70% sequence identity in the DBHS region [71,111]. Of the three mammalian DBHS proteins, SFPQ is the most distinct, with a long N-terminal region containing a putative DNA-binding domain. DBHS proteins are predominantly nuclear, however there is evidence for cytoplasmic enrichment in some cell types [112].

DBHS proteins have been implicated in the regulation of transcription, in post-transcriptional RNA processing, in RNA stability, RNA localization and in RNA degradation. The many ways the DBHS proteins can influence gene expression through their ability to interact with proteins and nucleic acids are summarized in Table 1.1.

As this work focuses on the RNA binding properties of NONO and SFPQ, Table 1.1 is a summary of the way these proteins can act on a RNA at each stage of its journey from transcription to degradation. Firstly the DBHS proteins are able to influence transcription of target genes, for example, the androgen receptor and ADARB2 [88,113]. The proteins influence transcription of target genes by forming a complex with RNA polymerase II, and, in some cases, they bridge transcription and processing of the nascent RNA [114]. Processing of newly formed RNAs facilitated by the DBHS proteins includes alternate splicing, polyadenylation and cleavage [115,116]. In addition, the DBHS proteins can facilitate RNA degradation [117], promote RNA stability [118] and facilitate RNA transport [119].

1.6.2 NONO and SFPQ bind lncRNAs to promote tumorigenesis in cancer

Besides binding NEAT1 in paraspeckles, there have recently been reports of DBHS proteins binding to other lncRNA molecules, thus suggesting lncRNA binding may be a general mechanism for DBHS proteins, particularly in controlling gene expression in cancer. NONO was reported to bind a novel lncRNA, lncUSMycN, found proximal to the MYC-N gene, an oncogene driving proliferation in several cancers. Liu and colleagues showed that NONO bound lncUSMycN, as well as MYC-N RNA, resulting in up-regulation of MYC-N expression in neuroblastoma, increased cellular proliferation and poor prognosis [36]. In a separate study it was shown that SFPQ associates with the lncRNA MALAT1, resulting in SFPQ disassociating from the oncogenic transcription factor PTBP2, releasing its repression of the oncogene and promoting its metastatic effects in the cell [59].

1.6.3 DBHS proteins in neurological disease

There have been several studies that link the DBHS proteins, in particular SFPQ, with brain development and neuronal disease. Experiments in zebrafish revealed that mutants of the SFPQ orthologue have gross defects in brain development caused by widespread cell death and defects in neuronal cell differentiation [120]. The necessity for SFPQ for proper brain development has also been suggested in mouse, with differential SFPQ expression in different neuronal cell types thought to contribute to splicing events occurring at specific stages of neuron development [121]. A number of studies have identified SFPQ as a possible mediator of neuronal

disease. One study identified SFPQ in a complex that regulates the transcription of a key dyslexia susceptibility gene [122]. In addition SFPQ was shown to be up-regulated in the brain of bipolar patients, suggesting it could also be influencing gene expression in this disease [123]. More recently, a study looking at gene expression in neurons and astrocytes from patients with frontotemporal lobar degeneration (FTLD) saw that SFPQ was almost completely depleted from the nucleus, instead, forming cytoplasmic aggregates [124]. This is intriguing as cytoplasmic accumulation of two other paraspeckle proteins, TDP-43 and FUS (both RNA binding proteins) is also observed in neuronal disease, hinting at the possibility that paraspeckles may sequester proteins in the nucleus to modulate their functions in the cell [88,125,126].

1.6.4 DBHS protein structure and interactions

How do the DBHS proteins carry out their myriad functions? One way the proteins can increase their functional diversity is to dimerize. Indeed, DBHS proteins exist as obligate dimers, as demonstrated by Yeast-2-Hybrid and Co-immuno-precipitation (IP) experiments [127], as well as *in vitro* studies. In HeLa cells, NONO and SFPQ are much more abundant than PSPC1 and the most common heterodimer observed is between SFPQ and NONO [74]. PSPC1 preferentially dimerizes with NONO, rather than SFPQ [74]. However, all dimer combinations form when the proteins are over-expressed [128] or as endogenous proteins in specific cell types [129].

Several studies have delineated which domains within DBHS proteins are responsible for different interactions. Dimerization of the proteins is mediated by RRM2, NOPS and a short stretch of the coiled coil [130]. The crystal structure of the NONO/PSPC1 heterodimer shows that the NOPS domain, a newly described Protein:Protein interaction domain, binds to the second RRM in the partner protein in the NONO/PSPC1 heterodimer, forming a tight interface that makes up 65% of the heterodimer interaction surface [130] (Figure 1.5 A). Key amino acids at the junction of the NOPS and coiled-coil domains were demonstrated to be essential for dimerization [130]. It is thought that subtle differences in sequence

between NONO, PSPC1 and SFPQ at this region result in the varying affinities for each protein to form heterodimers.

In contrast to dimerization, we have less insight into the molecular determinants of nucleic acid binding by DBHS proteins, largely as the crystal structure was solved in the absence of nucleic acid. RNA binding is likely facilitated, at least in part, by the two RRMs. While RRM1 is a canonical RNA recognition motif and likely binds 4-6 nucleotides, RRM2 is non-canonical and has potentially lost the ability to bind RNA, instead contributing to Protein:Protein interactions. One possible model for RNA binding involves the heterodimerization of DBHS proteins to bring the two canonical RRMs together to confer binding site recognition and specificity. This interaction would be further stabilized by the oligomerization of heterodimers to bring additional RRMs into contact with the RNA as the proteins bind along the transcript.

DNA binding is thought to occur for dimers containing SFPQ, utilizing the DNA binding domain unique to SFPQ. SFPQ binding to DNA was first demonstrated through EMSA (Electro mobility shift assay) [131]. Since then, the SFPQ DNA binding sequence has been identified through homology analysis of the SFPQ target gene IL8 in mammals [87] and it has been shown that SFPQ acts as a transcription factor through binding to several gene promoters and is also rapidly localized to sites of DNA damage [88,97,113,132].

1.6.4.1 RNA binding insights gleaned from paraspeckle localization

The regions within DBHS proteins required for their targeting to paraspeckles have been delineated, and this has given clues to how DBHS proteins bind each other, as well as RNA, in the cell. DBHS proteins require intact RRMs, as well as extended coiled-coils to localize to paraspeckles. This was demonstrated primarily by examining the localization of PSPC1 truncations and mutants. In these studies a PSPC1 construct with two mutated RRM motifs (in which the canonical RNA binding residues were substituted with alanine) failed to localize to paraspeckles when overexpressed [74]. In addition, the dimerization and extended coiled-coil region of PSPC1 was also shown to be necessary for paraspeckle localisation [74,130].

The minimal PSPC1 construct that can localize to paraspeckles contains RRM1, RRM2, NOPS, and a coiled-coil extending to residue 337, analogous to the extended coiled-coil of SFPQ shown in figure 1.5 B [130]. This requirement for an extended coiled-coil has led to a model of extended dimer oligomerisation being necessary for paraspeckle localization: ie. many dimers must oligomerize to be able to bind RNA in paraspeckles (Figure 1.5C). This model is strengthened by the observation that NONO and SFPQ containing extended coil-coil regions have poor solubility in solution [133,134], an effect that could be explained by the coiled-coil driven oligomerization. The Bond lab has recently solved the crystal structure of a construct containing this extended coiled-coil and have demonstrated oligomerization (Figure 1.5B). If the coiled-coil is in fact driving the oligomerization of the DBHS proteins to facilitate their inclusion into paraspeckle complexes, this mechanism of interaction is reminiscent of other nuclear bodies which form through protein self organization dictated by Protein:Protein interaction regions [135,136].

DBHS proteins are thought to be some of the first proteins recruited to NEAT1 once it is transcribed, and are speculated to directly bind NEAT1. IP of DBHS proteins and RT-qPCR of the associated RNA showed that NEAT1 was co-purified [47,78]. However, this co-purification may still arise via additional proteins in the complex, therefore direct binding *in vivo* between NEAT1 and DBHS proteins has yet to be definitively demonstrated. It has been shown that NONO and SFPQ stabilize NEAT1_v2, as siRNA against either of these proteins results in lower NEAT1_v2 levels [78].

1.6.4.2 What sequence of RNA do DBHS proteins bind?

The RNA binding sites recognized and occupied by the DBHS proteins has been under investigation for some time. SELEX (Systematic evolution of ligands by exponential enrichment) experiments with SFPQ showed that it has a preference for purine binding (50-67%) and the consensus sequence 5'-UGGAGAGGAAC-3' was reported [127]. An earlier SELEX experiment with NONO showed it has a preference for binding to G-rich regions containing one or more copies of the 5'-AGGGA/U-3' sequence [137]. SELEX experiments performed with purified PSPC1

homodimers identified the sequence 5'-UUUGUAA-3' (Passon 2012 Thesis). It is also highly likely that the DBHS proteins recognize and bind RNA based on its secondary structure, perhaps in addition to sequence recognition. A model for RNA binding may be sequence-specific binding by RRM1, followed by a structured RNA bound by RRM2 (facing inward, towards the center of the structure), however this model is yet to be tested.

It is important to note that, to date, studies to identify an RNA binding motif for NONO and SFPQ have been performed *in vitro*. Thus it is unlikely that these experiments faithfully mimic the complex interactions mediated *in vivo*.

1.7 Aims

The over-arching aim of this project was to investigate the *in vivo* RNA binding by DBHS proteins, in order to understand paraspeckle organization, as well as gaining insight into the greater role of DBHS proteins in the cell. At present, very little is known about the molecular interactions that occur to build and maintain lncRNA:Protein complexes such as paraspeckles. In addition, we do not understand the dynamic interactions that contribute to the function of these complexes in the cell.

This study had three aims:

1. Optimize a technique known as PAR-CLIP **P**hoto**a**ctivatable **r**ibonucleoside enhanced **c**ross**l**inking and **i**mmuno**p**recipitation) coupled with next generation RNA sequencing and bioinformatics analysis to purify and identify RNAs bound by the DBHS proteins NONO and SFPQ.

2. Identify the DBHS protein binding motifs in NEAT1 to give insights into how the core paraspeckle proteins are organized along the structural lncRNA.

3. Identify other coding and non-coding RNAs bound by NONO and SFPQ and characterize the role of NONO and SFPQ on the regulation of these RNAs.

2. Materials and Methods

2.1 General methods

The following methods apply to all of the non-PAR-CLIP experiments described in this thesis. For PAR-CLIP specific methods, please see chapter 2.2

2.1.1 Cell culture

All cell lines were cultured in Dulbecco's modified eagle medium (DMEM) with high glucose (4500 mg/L), L-glutamine (584 mg/L) and sodium pyruvate (110 mg/L) (Life Technology) supplemented with 10% Fetal Calf Serum (Gibco). Cells were grown at 37°C in 5% CO2. HeLa cells were late passage lab stocks. NIH3T3 cells were a gift from the lab of John Mattick.

For passaging, cells were washed with Phosphate Buffered Saline (PBS) (Gibco) and trypsinised with TrypLE Express Enzyme (Gibco). DMEM was used to inactivate the trypsin, cells were centrifuged for 5 min at 193 g, and the cell pellet re-suspended in standard culture medium before seeding into fresh dishes.

2.1.2 Buffers and Reagents

All buffer and reagent details can be found in appendix 2.

2.1.3 Transfection

NONO was knocked down using NONO Silencer Select siRNA (Ambion) and SFPQ knockdown was achieved with SFPQ_human_14 Silencer siRNA (Invitrogen). Negative Silencer Select siRNA (Ambion) was used in control samples. Both NONO and SFPQ targeting siRNAs, as well as the scramble control, were used at 40 pmol siRNA per well of a 12 well plate or 80 pmol per well of a 6 well plate. SiRNA sequences are outlined in appendix 2.

Cells were transfected on day 0 with the reverse transfection method. This involved trypsinising and counting cells on the day of transfection and seeding 0.1

X 10^6 cells per well of a 12 well plate (in 1 ml volume) or 0.3 X 10^6 cells per well of a 6 well plate (in 2 ml volume). In microcentrifuge tubes, 100 μl Opti-MEM was incubated with 2 μl Lipofectamine RNAi Max (Intitrogen) (12 well plate transfection) or 300 μl Opti-MEM incubated with 4 μl Lipofectamin RNAiMax (6 well plate transfection) for 5 min at room temperature. siRNAs were diluted in 100 μlOpti-MEM (12 well transfection) or 300 μlOpti-MEM (6 well transfection).

The Lipofectamine mixture was combined with the siRNA mixture and incubated for 20 min at RT before being applied to the cell suspension in the wells and incubated for 24 h. On Day 1 the transfection was repeated in a forward manner (ie the existing cells in the wells were not trypsinised prior to transfection) and cells incubated for a further 24 h. On day 2 the media was changed and cells were harvested on day 3.

RNA isolation was carried out as in chapter 2.1.5. Alternatively, cell lysates for protein analysis were made by the removal of the medium, followed by a PBS wash. 2 X SDS loading dye was then added to the cells directly, incubated for 2 min at RT, and the sample scraped into a Qiashredder (Qiagen) to shear the DNA. Protein samples were stored at -20°C.

2.1.4 SDS-PAGE and Western Blotting

Gel electrophoresis was carried out with protein samples in 2 X SDS loading dye. Precision Plus dual colour standard (Biorad) was loaded for reference bands. Samples were heated for 10 min at 90°C prior to being loaded onto NuPAGE Novex 4-12% Bis-Tris Protein Gels (Novex) in 1 X NuPAGE MOPS SDS running buffer (Novex) and run at 150 V until the dye front reached the bottom of the gel.

For total protein visualization, gels were stained with Coomassie stain followed by destaining with coomassie destain solution and visualized with the Odyssey CLx (LI-COR).

For western blotting, the wet transfer method was used. PVDF Immobilon-P membrane (Millipore) was activated in methanol by brief rinsing and the transfer

set up as per XCell II Blot module (Novex) manufacturer instructions with 1 X Transfer buffer. Transfer was carried out at 25 V for 2 h.

The membrane was blocked for 1 h at room temperature in 5% skim milk/PBS-0.05% Tween20. For antibodies used, dilutions and blotting conditions see table 2.1. In all cases, each antibody incubation step was followed by three 5 min wash steps with copious PBS-0.05% Tween20 to remove non-specifically bound antibody. For detection of the primary antibodies with the Odyssey imager, IRDye 800CW Donkey anti-Mouse IgG, IRDye 800CW Donkey anti-Rabbit IgG and IRDye 680RD Donkey anti-Mouse IgG (all LI-COR) were diluted 1:10,000 in 5% skim milk/PBS-0.05% Tween20 and incubated on membrane for 1 h at room temperature. Western blots were imaged with the Odyssey CLx (LI-COR).

2.1.5 RNA extraction from total cell lysate

Total RNA was extracted using the Paris Kit (Ambion). Briefly, cells were trypsinised from the plate or flask and pelleted at 193 X g for 5 min at 4°C. Cell pellets were washed in 1 ml DEPC (Diethylpyrocarbonate) treated PBS and re-suspended in 300 μl Cell Disruption buffer. 300 μl Lysis/Binding solution and 300 μl 100% ethanol was added and mixed by inverting. Samples were pipetted into a filter cartridge and bound to the membrane by centrifuging at 10,000 X g for 30 sec at room temperature. Filters were washed once with 700 μl Wash Buffer 1 and twice with 500 μl Wash Buffer 2. Residual wash buffer was removed by centrifuging at 10,000 X g for 3 min at room temperature. RNA was eluted from the filter cartridge with 50 μl of Elution solution heated to 95°C and centrifuged at 10,000 X g for 30 sec. RNA was stored at -80°C.

2.1.6 Nuclear/Cytoplasmic fractionation and RNA extraction

Nuclear and cytoplasmic fractionation was performed with the Paris kit (Ambion). Briefly, cells were trypsinised from the plate or flask and pelleted at 193 X g for 5 min at 4°C. Cell pellets were washed in 1 ml DEPC PBS and re-suspended in 50 μl DEPC PBS. 500 μl cell fractionation buffer was added gently to the cell suspension

and incubated on ice for 3 min. Cells were centrifuged at 500 X g for 3 min at 4°C to pellet nuclei. 300 μl of the cytoplasm supernatant was removed to a fresh tube and kept on ice. The remaining 200 μl supernatant was discarded to prevent disturbing the nuclei and contaminating the cytoplasmic fraction. 300 μl Cell Disruption buffer was added to the nuclei and vortexed to homogenise. 300 μl Lysis/Binding solution was added to each of the cytoplasmic and nuclear fractions and mixed by inverting. 300 μl 100% ethanol was added to the fractions and they were bound to a filter cartridge by centrifuging at 10,000 X g for 30 sec at room temperature. Filters were washed once with 700 μl Wash Buffer 1 and twice with 500 μl Wash Buffer 2. Residual wash buffer was removed by centrifuging at 10,000 X g for 3 min at room temperature. RNA was eluted from the filter cartridge with 50 μl of Elution solution heated to 95° and centrifuged at 10,000 X g for 30 sec. RNA was stored at -80°C.

2.1.7 Reverse transcription and quantitative PCR

Reverse transcription was performed with the QuantiTect Reverse Transcription kit (Qiagen) according to manufacturer instructions. 500 ng RNA was used in each reverse transcription reaction. In the cases where there was insufficient RNA yield from the extraction, an equal amount from each sample in the experiment was used. Genomic DNA was eliminated by incubating RNA with 1 X gDNA wipeout buffer in a final volume of 14 µl at 42°C for 2 min. 1µl of Quantiscript Reverse Transcriptase, 4 µl of 5X Quantiscript RT buffer and 1 µl RT primer mix (all Qiagen) was added to the entire genomic DNA elimination reaction and incubated for 15 min at 42°C followed by 3 min at 95°C. cDNA was diluted 1:2 with 20 µl MilliQ water and stored at -20°C.

Quantitative PCR reactions were set up using mastermixes containing 1 X SensiMix SYBR Hi-ROX (Bioline) and 500 nM of forward and reverse primers and H_2O to achieve a final volume of 15 µl with 2 µl cDNA. qPCR primers are detailed in table 2.2. Reactions were set up in duplicate or triplicate and run in a Rotor Gene Q real time PRC machine (Qiagen) with the following cycling profile:

Cycles	Temperature	Time
1	95°C	5 min
50	95°C	15 sec
	59°C	15 sec
	72°C	15 sec

2.1.8 Q-PCR data analysis and statistical significance test.

For each experimental and control condition, RNA was isolated from three replicate siRNA knockdown experiments. Two cDNA synthesis reactions were done for each RNA sample and cDNAs were measured in triplicate with qPCR. Where appropriate, outliers were removed to bring the standard deviation for replicate Ct values below 0.5.

2.1.8.1 Q-PCR on RNA from total cell lysates

The comparative CT method was used to analyze the effect of DBHS protein knockdown on the total RNA levels of candidate bound transcripts [138]. In each qPCR run, β-actin was the housekeeping gene analyzed in parallel to normalize the CT values and account for run variability and differences in cDNA synthesis efficiency. The scramble control CT values served as the reference sample, and calibrator values calculated by normalizing the level of target gene in the reference sample to β-actin. The delta CT for the candidate gene and the reference sample was calculated for each knockdown experiment. Finally the fold change in expression of the candidate RNAs in each knockdown relative to the scramble control was calculated.

2.1.8.2 QPCR on RNA from nuclear and cytoplasmic fractions

The comparative CT method was also used to analyze the effect DBHS protein knockdown had on the level of candidate bound transcripts in the nuclear and cytoplasmic cellular fractions. Again, β-actin was used as the housekeeping gene

run in parallel. Candidate gene CTs in the nuclear isolated RNA were normalized to β-actin in the nucleus, and candidate gene CTs in the cytoplasmic isolated RNA normalized to β-actin in the cytoplasm. Scramble control CT values in each RNA fraction served as the nuclear and cytoplasmic reference samples, and the calibrator values were calculated by normalizing the level of target gene in the reference sample to β-actin in the nucleus and cytoplasm respectively. The delta CTs for the candidate gene and the reference sample in each fraction were calculated for each knockdown experiment. From this, two comparisons were made; firstly, the amount of candidate gene in each cellular fraction was compared to its level in that fraction without DBHS knockdown (ie, the scramble control for that cellular fraction was set to 1 and the levels of candidates following each knockdown was relative to that). Secondly, a comparison was made between the level of the candidate gene in the nucleus compared to its level in the cytoplasm following DBHS protein knockdown (ie, for each knockdown, the transcript level in the cytoplasm was set to 1, and its level in the nucleus calculated relative to that).

2.1.8.3 Statistical significance testing

T-tests were performed and the results graphed using GraphPad Prism software (Version 6, San Diego California USA, www.graphpad.com). Unpaired, two tailed Student's T-tests were used to test for significant differences ($P<0.05$) between the scramble control and DBHS protein knockdown samples.

2.2 PAR-CLIP experimental methods

PAR-CLIP experiments were performed according to the published protocol [139] with optimization as outlined below. See figure 2.1 for a schematic representation of the PAR-CLIP protocol.

2.2.1 Antibody conjugation to Dynabeads

300 μl protein G Dynabeads (Invitrogen) were used to prepare 600 μl of NONO antibody conjugated beads. Dynabeads were washed 3 times in low salt NP40 lysis

buffer to remove the preservative they were stored in. In all cases, a magnetic rack was employed for separation of liquids from Dynabeads. Mouse monocolonal Anti-NONO antibody (Monoclonal Antibody Facility, Harry Perkins Institute for Medical Research) was used for both NONO IP and western blot and recognizes the epitope CSQGNFEGPNKRRRY. 75 μg of NONO antibody was incubated with the Dynabeads and made to a final volume of 1 ml in a 1.5 ml tube with low salt NP40 lysis buffer. The antibody was bound to the beads by incubation at 4°C overnight on a rotating wheel. Beads were collected with a magnet and supernatant removed. To conjugate the antibody to the Dynabeads, beads were washed twice in 1 ml 0.1 M sodium borate, pH 9 and then incubated with 1 ml of sodium borate with 20 mM dimethyl Pimelimidate (DMP) (Note, the DMP was stored as a solid, and the required amount weighed out and added just prior to use) and incubated for 30 min at RT on a rotating wheel. Beads were collected and re-suspended for a second time in 1 ml of fresh sodium borate/ 20 mM DMP for a further 30 min. Beads were then collected and washed twice in 1 ml 50 mM glycine, pH 2.5. Beads were then washed 4 times with 1 ml low salt NP40 lysis buffer to remove glycine and made to a final volume of 600 μl with low salt NP40 lysis buffer and stored at 4°C.

2.2.2 Cell growth and 4-thiouridine incorporation

HeLa and NIH3T3 cells were seeded at a density of 2.5 X 10^6 cells per 15 cm dish. Generally, 20 dishes were processed at a time. When cells were ~80% confluent, they were supplemented with 0.5 M 4-thiouridine (4-SU, Sigma) to a final concentration of 100 μM. Cells were grown in 4-SU supplemented media overnight (14 hours).

2.2.3 Crosslinking, cell lysis and RNase T1 treatment

Following overnight incubation, 4-SU supplemented growth media was removed and cells washed once with PBS. The PBS was removed and the cells (largely devoid of a liquid covering) were irradiated with 365 nm UV at 0.15 J/cm^2 in a customized Stratalinker 2400 fitted with the appropriate wavelength UV lamps (Stratagene). Cells were collected with a cell scraper in 1 ml ice cold PBS per plate

and pelleted by centrifugation at 193 X g for 5 min. The cell pellet was re-suspended in 3 cell pellet volumes of high or low salt NP40 lysis buffer (A pellet from 20 dishes was generally re-suspended in ~6 ml NP40 lysis buffer). The re-suspended pellet was incubated on ice for 10 min and the lysate was passed through a 32 gauge needle 10 times and centrifuged at 14,000 X g for 15 min at 4°C to clear lysate. The supernatant was moved to a fresh tube and the lysate incubated with RNase T1 (Fermentas) at RT according to the extent of RNase digestion required (Table 2.3).

2.2.4 Immunoprecipitation

Lysate was pre-cleared with Protein G Dynabeads (20 μl beads/1 ml lysate) for 30 min at 4°C on a rotating wheel. Following pre-clearing, the lysate was moved to fresh tubes and incubated with 100 μl of NONO antibody conjugated Protein G Dynabeads per 1ml lysate for 2 h at 4°C on rotating wheel. Typically 6 ml of lysate was processed at a time, resulting in 600 μl pooled beads. Following immunoprecipitation (IP), beads were collected into one tube and were washed 4 times with 1 ml high or low salt NP40 lysis buffer. The salt concentration of the NP40 lysis buffer was optimized to minimize non-specific protein interactions in the IP, low salt buffer being less stringent that high salt buffer. For PAR-CLIP_Harsh, a second RNase T1 digestion was performed on beads in NP40 lysis buffer (Table 2.3). Beads were re-suspended in one bead volume dephosphorylation buffer and Calf intestinal alkaline phosphatase (CIP, New England Biolabs) at a final concentration of 0.5 U/μl for 10 min at 37°C. Following incubation, beads were washed twice with 1 ml Phosphatase wash buffer and then twice with 1 ml Polynucleotide kinase buffer (PNK), leaving beads in a final volume of 50 μl PNK buffer.

2.2.5 Radiolabelling

Beads were incubated with γ-^{32}P-ATP (Perkin Elmer) at a final concentration of 0.5 µCi/µl and T4 PNK (Polynucleotide kinase) at 0.8 U/µl for 30 min at 37°C. Non-radioactive ATP (NEB) was added to a final concentration of 100 µM and incubated

for a further 5 min at 37°C. Beads were washed 5 times with 1 ml PNK buffer and were finally re-suspended in 40 µl 2 X SDS loading dye.

2.2.6 SDS-PAGE and electro-elution

The radiolabelled bead suspension was incubated at 95°C for 5 min. Following this incubation, beads were collected with a magnet and supernatant loaded in to one lane of a 10 well NuPAGE Novex 4-12% Bis-Tris Protein Gels (Novex) and run at 150 V for 1 h 10 min. Following electrophoresis, the gel was wrapped in plastic cling wrap and exposed to film (Amersham hyperfilm ECL, GE Healthcare) and visualized. Film was aligned to the gel and used to direct excision of the bands corresponding to RNA:Protein complexes. D-Tube Dialyzer midi tubes (Merk Millipore) were used for electroelution of RNA:Protein complexes from gel slices as per manufacturers instructions. Briefly, D-tubes were equilibrated with 800 μl DEPC H_2O for 5 min at room temperature. Water was removed and 400 μl 1 X NuPAGE MOPS SDS running buffer (made with DEPC H_2O) added to each D-tube, along with a gel slice. Electroelution was performed in a Mini-Sub Cell GT tank (BioRad) filled with 1 X NuPAGE MOPS SDS running buffer (made with DEPC H_2O) and run at 100 V for 2 h.

2.2.7 Proteinase K treatment

Electroeluent was moved to a fresh 1.5 ml tube and supplemented with CaCl to a final concentration of 3 mM and SDS to a final concentration of 1% in a final volume of 500 μl. Proteinase K (Fermentas) was added to a final concentration of 1.2 mg/ml and the sample incubated at 55°C for 30 min. TRIsure (Bioline) was added at a ratio of 3:1 (vol/vol), trizol:sample and stored at -80°C.

2.2.8 RNA purification

TriSure samples from multiple IPs were pooled for RNA purification. For Harsh RNAse treatment sample, a total of 100 15 cm dishes of cells were used. For PAR-

CLIP_Mild and PAR-CLIP_Medium, 60 15 cm dishes of cells were used. Chloroform (Sigma) was added to trizol samples at a ratio of 1:5, chloroform:sample volume, vortexed and incubated at room temperature for 3 min. Samples were centrifuged at 12,000 x RPM in a microcentrifuge for 15 min at 4°C and the aqueous phase moved to fresh tube and mixed with one aqueous phase volume of 70% ethanol by vortexing. RNA extraction was then performed with a miRNAeasy Micro Kit (Qiagen) as per manufacturers instructions. Briefly, the ethanol/RNA solution was passed through an RNeasy MinElute spin column by centrifuging at 8,000 X rpm in a microcentrifuge for 15 sec at room temperature. The flow through (containing the small RNA fraction) was moved to a fresh tube and mixed with 1 volume of 100% ethanol. The sample was then passed through an RNeasy MinElute spin column. Both the columns (the first containing the large RNAs and the second containing the small RNAs) were then washed with 700 μl buffer RTW then 500 μl buffer RPE. The spin columns were dried with an extra 2 min spin and RNA was eluted in 20 μl warm RNase free water yielding a total of 40 μl combined eluted RNA.

2.2.9 RNA quantification and library preparation

RNA was quantified using the Qubit Fluorometer (Life technology) and sent on dry ice for library preparation. Libraries for PAR-CLIP_Harsh and PAR-CLIP_Mild experiments were prepared by Gene Works (Adelaide, Australia) and PAR-CLIP_Medium and NIH3T3 PAR-CLIP_Medium libraries prepared by AGRF (Melbourne, Australia). In both cases the Illumina TruSeq small RNA kit v2 (Illumina) was used. Briefly, RNA adapters were ligated to the 3' and 5' ends of the RNA molecules. The molecules were then reverse transcribed and the resulting cDNA underwent 15 cycles of PCR amplification. The resulting libraries were gel purified, with gel cuts made from just above the adapter dimer to include as much of the library band as possible.

2.3 PAR-CLIP bioinformatics methods

2.3.1 Sequencing platforms

cDNA libraries from the NONO and SFPQ bands from PAR-CLIP_Mild were sequenced on the Illumina MiSeq in separate lanes of a flow cell with 75 nt single end read lengths. cDNA libraries from the NONO and SFPQ bands from PAR-CLIP_Medium experiments (HeLa and NIH3T3 PAR-CLIP) were sequenced on the Illumina HiSeq, multiplexed in one lane with 75 nt single end reads. For sequencing on the MiSeq, approximately 20 million reads/library was achieved and for the HiSeq approximately 30million reads/library.

2.3.2 Galaxy for FastX toolkit, grooming and clipping

Raw reads from each library were uploaded to the Galaxy Server (http://galaxyproject.org) for read processing. Illumina sequencing adapters were removed from the dataset using the FastX Clipper tool from the FastX Toolkit (http://hannonlab.cshl.edu/fastx_toolkit/index.html). Reads that contained only adapter sequence, reads that were too short (<15 nt) after clipping, or reads that contained 'N' bases were removed from the dataset. The dataset containing NONO and SFPQ clipped reads was downloaded for mapping with bowtie.

2.3.3 Bowtie for mapping to the human genome

Bowtie (http://bowtie.cbcb.umed.edu) version 1.0.0 was used to map reads from each dataset against the human genome. The *H. sapiens*, UCSC hg19 pre built index was used as the reference genome. Reads containing up to three mismatches and that aligned to less than 10 locations in the best alignment stratum (bowtie parameters –v3 –m10 –-best –-strata) were reported to output for further processing. Aligned reads for each dataset were collapsed using the FastX collapse tool from the FastX Toolkit and the resulting mapped reads processed with PARalyzer.

2.3.4 PARalyzer for transition analysis

For each protein dataset, PARalyzer [140] was used to identify protein binding sites in the mapped reads. PARalyzer utilizes the T-to-C conversion event introduced by 4-SU crosslinking to identify potential protein binding sites in the transcripts. PARalyzer takes into account the read depth over a transition site and the number of transition locations within groups of overlapping reads to report potential binding sites. A schematic representation of the PARalyzer methodology is outlined in figure 2.2. Briefly, the files containing collapsed aligned reads for the protein were input into PARalyzer (Figure 2.2 A). PARalyzer grouped overlapping reads, and reported groups that contained a minimum number of reads as specified (Figure 2.2 B). The output of this is a "groups" file. PARalyzer then defined clusters within each group according to input parameters that specified the minimum conversion location, minimum conversion count and the minimum number of reads that contained a conversion at that location (Figure 2.2 C). The output of this is a "clusters" file that reports among other things the cluster ID, genome co-ordinates, sequence, read count, conversion event count and conversion location count. The optimized PARalyzer parameters used to analyze each protein dataset are outlined in table 2.4. Strict parameters were used to analyze PAR-CLIP_Mild datasets and PAR-CLIP_Medium in HeLa and NIH3T3 cells. In addition, HeLa PAR-CLIP_Medium datasets were also analyzed with relaxed PARalyzer parameters. See PARalyzer scripts (as .ini files) in appendix 3 for exact parameters.

2.3.5 Analysis of PARalyzer clusters for binding site features

A method to match clusters from PAR-CLIP_Mild to annotated transcripts was implemented in python. The analysis pipeline is outlined in figure 2.3.
The table of UCSC Known Genes based on the hg19 database (UCSC_knownGene, downloaded December 2013) was used to cross-reference the clusters to obtain the names of the genes the clusters mapped to. The translation status (i.e. mRNA or non-coding RNA) of each transcript a cluster mapped to could be extracted using information from the UCSC_knownGene table.

Clusters from PAR-CLIP_Medium were matched to transcripts manually as there were few enough clusters in PAR-CLIP_Medium for this method to be feasible.

Information about the transcript feature type (i.e. first intron, non-first intron, 3'UTR etc.) each cluster mapped to was manually curated, and proportions of clusters in each feature type calculated.

2.3.6 Motif finding for NEAT1 clusters

MEME (Multiple Em for Motif Elicitation)[141] was run through the MEME Suite (http://meme.nbcr.net/meme/). Data files containing the sequences of each of the clusters that aligned to NEAT1 in each protein dataset were uploaded to MEME. MEME was set to find motifs that occurred at a minimum of 5 and a maximum of 100 sites with any number of repetitions, with a minimum width of 4 nucleotides and a maximum of 10 nucleotides in the given strand only.

2.3.7 Ingenuity Pathway analysis for Gene Ontology assignment

Transcripts were analysed with QIAGEN's Ingenuity Pathway Analysis (IPA, QIAGEN Redwood City, www.qiagen.com/ingenuity).

3 Optimization of PAR-CLIP for the isolation of transcripts bound by the DBHS proteins

PAR-CLIP (**P**hoto**A**ctivatable **R**ibonucleoside enhanced **C**ross**L**inking and **I**mmuno**P**recipitation) is a relatively new technique used to discover the target transcripts of RNA binding proteins. It utilizes *in vivo* crosslinking and subsequent T-to-C transition of the uridine analogue incorporated into the RNA to identify protein-bound transcripts and map the binding sites to single nucleotide resolution.

PAR-CLIP is one of several techniques available for the identification and characterization of protein binding sites in RNA. CLIP (**C**ross**l**inking and **i**mmuno**p**recipitation) experiments have traditionally been employed to identify the DNA fragments bound by a variety of proteins. These techniques utilize ultraviolet irradiation to stably crosslink proteins to the DNA they bind, and then isolate and identify the DNA bound to the protein of interest by immunoprecipitation (IP), DNA isolation and sequencing [142]. UV irradiation can be applied to live cells, meaning crosslinks are formed *in vivo*, providing a snapshot of nucleic acid binding at the precise moment of crosslinking. The formation of stable crosslinks also means the protein IP and washing conditions can be very stringent, as the nucleic acids are attached and will not disassociate if the protein denatures.

The interest in RNA binding proteins has increased in recent years, given the growing appreciation of the importance of post-transcriptional control of gene expression. As such, CLIP techniques have been modified and optimized for the identification of the RNA targets of proteins, and the characterization of RNA binding motifs for different proteins [143]. Figure 3.1 outlines and compares the various CLIP methods. Some of the first RNA immunoprecipitation (RIP) experiments utilized the reversible crosslinking properties of formaldehyde to facilitate *in vivo* preservation of RNA binding proteins in complex with their target RNAs [144]. The ability to reverse formaldehyde crosslinks meant the RNA could be isolated without the need to digest the bound proteins with a protease such as proteinase K. However, formaldehyde not only mediates RNA:Protein crosslinks, but also induces DNA:Protein and Protein:Protein crosslinks, making it

problematic for the study of large RNA:Protein complexes that contain many proteins. In these cases, it was difficult to discern which of the proteins in the complex was binding each isolated transcript.

254 nm UV irradiation of live cells overcomes this problem because it does not induce Protein:Protein crosslinks [145], meaning the isolation of the protein of interest and its bound RNA is dependent upon the specificity of the antibody and the stringency of the IP. 254 nm UV crosslinking has been applied to the identification of the RNA targets of some key proteins by crosslinking *in vivo*, digesting the RNA to obtain fragments of an optimal size for purification and Sanger sequencing [146].

Newer CLIP protocols now utilize next generation sequencing to give a transcriptome wide picture of protein binding sites. The higher read count and depth achieved with next generation sequencing aids in the identification of binding motifs. CLIP-Seq was utilized to identify targets of FOX2 in human embryonic stem cells to give a detailed picture of the targets of this protein [147]. HITS-CLIP (**Hi**gh **t**hroughput **s**equencing of **c**ross**l**inking **i**mmuno**p**recipitation) was developed to investigate the binding targets of the neuron specific splicing factor NOVA on a genome wide scale [148]. Previous CLIP experiments on NOVA using Sanger sequencing identified 2,481 NOVA bound RNAs from 5 experiments, while HITS-CLIP identified 412,686 CLIP tags from 3 experiments, illustrating the power of coupling CLIP with deep sequencing [149].

Whilst high throughput sequencing allows the identification of a larger number of bound RNAs, the increased number of sequences can include non-specifically bound RNAs, thus complicating motif identification. With such a large number of sequences (a typical RNA-Sequencing run yields ~15-30 million reads) it is useful to have a method that allows the identification of true binding sites. New variations on CLIP address this by incorporating a step that allows the identification of protein binding sites to single nucleotide resolution, thus enabling the identification and removal of fragments that were not directly bound by the protein of interest.

iCLIP (**I**ndividual nucleotide **c**ross**l**inking and **i**mmuno**p**recipitation) utilizes the fact that during proteinase K digest a peptide is left at the site where the protein was cross-linked to the RNA following 254 nm UV irradiation [150]. Reverse transcriptase cannot read through this peptide-bound site resulting in a truncated cDNA terminating at that position [151].

PAR-CLIP is another technique that allows the identification of protein binding sites in RNA to single nucleotide resolution[139]. PAR-CLIP relies on the pre-incorporation of 4-thiouridine (4-SU) into the RNA transcripts within the cell. The 4-SU is then specifically crosslinked to bound proteins when irradiated with long wave UV light (365 nm UV) [152,153]. The long wave UV irradiation of 4-SU is much more efficient than the 254 nm UV irradiation used in other CLIP techniques. Incubation with RNase results in the degradation of unprotected RNA, leaving only the 'footprinted' RNA cross-linked to the protein. The RNase incubation also releases any other RNP (Ribonucleoprotein) complexes from the protein of interest. Following IP with a specific antibody to the protein of interest, the small RNA fragments cross-linked to the protein are reverse transcribed, during which guanosine is mis-incorporated opposite the cross-linked 4-SU nucleotide analogue. Analysis of the genomic sequence of the mapped reads will display a T-to-C transition at the site of 4-SU incorporation and crosslinking, indicating that a protein was in direct contact with that nucleotide [139,154,155].

Unlike iCLIP, in which the sequence read ends at the protein bound nucleotide, PAR-CLIP allows multiple protein bound nucleotides (T-C transitions) to be seen in the same read. PAR-CLIP not only allows for transcriptome wide identification of binding sites, but also allows better characterization of the mechanisms of protein binding, as multiple binding sites can be resolved from the same sequenced fragment, and used to inform secondary structure prediction. For these reasons, PAR-CLIP was adopted to identify the RNA targets of NONO and SFPQ and identify to single nucleotide resolution the protein binding sites in the target transcripts. One drawback of PAR-CLIP is that it relies on cells in culture for the starting material , due to the requirement for 4-SU incorporation into the RNA. Other limitations of PAR-CLIP, and considerations for the data analysis are described in chapter 6.

This chapter details the various experiments and analyses that were performed to optimize PAR-CLIP for the isolation of RNA bound by the DBHS proteins. These optimizations resulted in isolation of RNAs that were bound specifically by NONO and/or SFPQ. A sufficient quantity of RNA was isolated to be used for deep sequencing library preparation. The RNA fragments were optimized for length to achieve a balance between being long enough to map uniquely to the genome and being short enough to give higher resolution maps of NONO and SFPQ binding sites.

3.1 HeLa lysis and NONO immuno-precipitation in NP40 lysis buffer

The first priority was to confirm that the PAR-CLIP cell lysis protocol (Chapter 2.2) gave complete lysis of HeLa cells to release NONO from the nucleus in a soluble form. To examine this, HeLa cells were grown as outlined (Chapter 2.2.2) with 4-SU incorporation for 14-16 h prior to harvest. Lysates were prepared as described in (2.2.3) and were pre-cleared with beads to measure non-specific association of proteins with beads, followed by the lysate being used for IP of NONO with antibody conjugated beads (Chapter 2.2.4.). The NONO antibody was made by the Fox lab, and was raised against an epitope at the extreme C terminus of NONO [76]. Various samples were collected throughout this procedure and analysed by SDS-PAGE and western blotting with NONO and β-actin antibodies (Chapter 2.1.4). The western blot for NONO and β-actin on lysates and IP samples (Figure 3.2) shows NONO was clearly present in the lysate (lane 1), indicating the NP40 lysis buffer effectively lysed the HeLa cells and resulted in solubilisation of NONO. Interestingly, both NONO and β-actin were present on the beads used for pre-clearing (Fig 3.2, lane 2), suggesting there is some non-specific binding of the target and control proteins to the beads. This is not surprising given the known property of NONO as a sticky protein commonly identified as a contaminant in mass spectrometric experiments of resin-based IPs [156]. However, despite this non-specific association of NONO with the resin, there was still a considerable amount of NONO remaining in the input, which was then used for NONO IP (Figure 3.2, lane 3). Reassuringly, there was no NONO or β-actin observed on the control IP

beads (incubated in parallel to the NONO IP, Figure 3.2, lane 6). This suggests that all the protein that interacts non-specifically with the beads had been removed either in the pre-clearing step or with the high salt NP40 lysis buffer washes. Post IP lysates for control IP and NONO IP are samples of lysate taken following IP with 'empty' Dynabeads (ie, Dynabeads with no antibody conjugated to them) or NONO-antibody coupled Dynabeads respectively. Figure 3.2, lane 5 shows that there was some NONO present in the lysate following NONO IP. Therefore, the amount of antibody used was insufficient to fully deplete NONO from the lysate. The amount of antibody coupled to the beads was increased for all future experiments. NONO, but not β-actin, was eluted off the NONO beads (Figure 3. 2, lane 7). Also eluted off the NONO beads was a protein running just above NONO at approximately 56 kD. Previous experiments carried out in the Fox laboratory have demonstrated that this protein is PSPC1, which heterodimerizes with NONO. It was previously shown that NONO co-immunoprecipitates with PSPC1 [74], hence it is likely that PSPC1 is co-immunoprecipitating with NONO here and the antibody is cross-reacting with PSPC1 when present at high levels. The peptide the antibody is raised against is in a region of sequence similarity with PSPC1. Three faint larger proteins are visible around 160 kD that may contain NONO in a larger protein complex. A faint band just below NONO, at approximately 52 kD is also visible, however the identity of this protein is unknown.

Given NONO is predicted to oligomerize into multi-protein complexes (see Figure 1.5 C) it was necessary to check whether the introduction of crosslinks (part of the PAR-CLIP procedure) to stabilize these complexes would effect NONO IP. *In vivo* crosslinks were introduced by irradiating cells with 254 nm UV light and cell lysates were incubated with either NONO antibody-coupled beads or control beads. The IP was performed at a large enough scale that proteins could be detected without western blotting. SDS-PAGE and coomassie blue staining of the samples eluted from the beads show clearly that crosslinking has not deleteriously affected the IP (Figure 3.3, compare lane 2, no crosslinking, with lane 4 that had crosslinking). NONO IP resulted in purification of NONO and PSPC1 as previously seen (Figure 3.2). In addition, NONO IP also pulled down a higher molecular weight protein, known to be the third member of the mammalian DBHS protein family, SFPQ. SFPQ is isolated with NONO in a 1:1 ratio in both crosslinked and non-

crosslinked lysates. This indicates it forms a stable heterodimer with NONO that is not disrupted by the high salt NP40 lysis buffer washes, and the NONO/SFPQ complex only dissociates during gel electrophoresis. This experiment also confirms that the stringent high salt NP40 lysis buffer IP and washes largely prevents co-IP of other non-specifically bound proteins. However, there are two additional bands visible in both crosslinked samples in the presence and absence of NONO IP (lane 3 and 4 respectively). These bands could represent proteins that precipitate upon crosslinking and stick non-specifically to the Dynabeads. There is a fuzzy band running at the bottom of the gel that is present in both NONO IP samples (lane 2 and 4). This may be a degradation product of either NONO, PSPC1 or SFPQ or an as yet unknown protein that forms a stable complex with one of the DBHS proteins.

3.2 4-Thiouridine incorporation and crosslinking is essential for co-immunoprecipitation of RNA with NONO.

Using the lysis conditions as optimized above, I next carried out a full PAR-CLIP experiment, including controls, and obtained an autoradiograph of the SDS-PAGE of the radiolabelled RNA:Protein complexes (Figure 3.4). This autoradiograph showed a number of important features. Firstly, the SDS-PAGE separated two major species of RNA:Protein complexes, corresponding to NONO (Figure 3.4, lane 4, lower band), and SFPQ (Figure 3.4, lane 4, upper band). This interpretation was possible, due to the information obtained in figure 3.3, showing that SFPQ was pulled down with NONO with a 1:1 ratio. Figure 3.4 also reveals that the 365 nm UV crosslinking was essential for the preservation of NONO and SFPQ bound RNA within the PAR-CLIP experiment. In the absence of 4-SU incorporation and crosslinking, no RNA was found in complex with NONO or SFPQ (Figure 3.4, compare lanes 2 and 4), despite the fact that these IP conditions pull out the proteins (shown in Figure 3.2). A radioactive smear is visible at the bottom of the lane containing non-crosslinked NONO IP sample (lane 2). This smear could represent RNA that was bound by NONO and/or SFPQ, but that disassociated when the proteins were denatured. This same low molecular weight RNA smear is also present in the NONO IP sample that had 4-SU incorporation and crosslinking (lane 4) but is fainter, indicating there is less RNA dissociating from the protein under these conditions. This suggests that crosslinking secures the majority of bound

RNA to the proteins. There is no radioactively labeled RNA detected in the control IP samples (Figure 3.4, lanes 1 and 3) indicating that the RNA isolated in the PAR-CLIP experiment is present as a result of the NONO antibody being present, enriching for NONO-RNA and SFPQ-RNA complexes.

3.3 Optimization of salt concentration in the NP40 lysis buffer to maximize IP stringency

Pilot PAR-CLIP experiments suggested that the harsh RNase treatment and high salt buffer washes gave very clean IPs, but very low RNA yields. Optimizations were carried out to assess whether the salt concentration in the NP40 lysis buffer could be altered to reduce the stringency of the IP to give higher RNA yields while still preventing non-specific RNA binding. Two different concentrations of salt in the NP40 lysis buffer were trialed with and without RNase T1 digest to assess the effect the stringency of the IP washes had on RNA yield.

Figure 3.5 is an autoradiograph of the SDS-PAGE from a PAR-CLIP experiment comparing these salt concentrations and conditions. As expected, the low salt NP40 lysis buffer (150 mM KCl) resulted in a much higher background RNA signal (Figure 3.5, compare lanes 3 and 4, with lanes 1 and 2). This result likely reflects extensive non-specific binding of proteins and RNA onto the beads. Where RNase T1 digestion was performed with the low salt samples (Figure 3.5, lane 4) the RNA background smear was darkest at the bottom of the gel, suggesting this sample contains partially digested RNAs that were not completely washed off the beads with the low salt washes. In contrast, where the high salt NP40 buffer (1.5 M KCl) was used with RNase treated samples (Figure 3.5, lane 2) there was no radioactive smear at the bottom of the gel, confirming the high salt lysis buffer was able to wash away all the weakly bound RNAs and RNA:Protein complexes.

An interesting observation was made with the undigested samples (Figure 3.5, lanes 1 and 3); in the absence of RNase T1 digest, the low salt buffer washes gave a very high background (Figure 3.5, lane 3), suggesting there was extensive RNA:Protein complexes non-specifically co-IP'd with NONO in this sample. In contrast, when the high salt buffer was used for IP without RNase T1 digestion, no RNA was visible, surprisingly, not even in complex with NONO or SFPQ (Figure 3.5,

lane 1). Previous results had confirmed that the proteins can IP in the absence of RNase T1 digest (see Figure 3.3). One interpretation of this is that in the absence of RNase, the NONO-RNA and SFPQ-RNA complexes are so large that they do not effectively IP onto the beads. Instead, it is likely that the protein pulled down under these conditions is a pool of 'free' protein that is not part of a larger RNP complex. Given these results, a high salt buffer was used for subsequent IP and wash steps, with the number of dishes of HeLa cells scaled up to isolate sufficient quantities of RNA for deep sequencing.

3.4 Titrating the RNase T1 to optimize RNA fragment length

RNase digestion of extraneous bound RNA is a key element of PAR-CLIP experiments. RNA digestion facilitates the preferential isolation of those RNAs that are proximally bound to proteins, as these fragments are protected by the protein and not digested. However, the extent to which protein-bound RNAs are digested depends on the binding properties of the protein of interest. If the proteins bind weakly, or like beads on a string, the RNA will be prone to digestion by RNases to a greater extent than if the protein binds strongly and wraps the RNA in a tight complex. Thus, it is essential to carefully assess and optimize RNase digest conditions for the RNA binding mode of the protein of interest. It is also important to note that whilst RNase T1, which cleaves specifically 3' of G nucleotides, was used in these experiments, there are many other RNases that could have been used, each with different sequence recognition sites and activity [157]. Time and costs nessitated choosing just one RNase to pursue and given RNase T1 wwas used in the original PAR-CLIP protocol this was the one that was adopted. The RNase T1 digest in the experiments reported here did not result in observable depletion of G nucleotides in the isolated RNA fragments.

Having established that an RNase T1 digestion step was essential to purify RNA in complex with NONO and SFPQ (see Chapter 3.3), it was of interest to test how minimal this RNase digest could be. The aim was to establish PAR-CLIP conditions that could give RNA fragments of varying lengths with the premise that the longer RNA fragments would give indicators of genome-wide protein binding sites, while the smaller fragments would allow precise binding sites to be resolved (more on this in chapter 4).

3.4.1 PAR-CLIP_Harsh experiment gives RNA fragments that are too short for analysis.

PAR-CLIP_Harsh experiments incorporated the RNase T1 digest conditions that were used in the first published PAR-CLIP study[139]. Briefly, 1 U of RNase T1 per 1 μl of lysate was added to the lysate prior to IP for 15 min at RT, and following IP, the beads were then incubated with 100 U RNase T1 per 1 μl of IP bead suspension for 15 min at RT. A representative PAR-CLIP_Harsh experiment (Figure 3.6 A) shows that under these conditions, no RNA was isolated in the control IP (Figure 3.6 A, lanes 1 and 3) and well-separated radioactive bands corresponding to NONO and SFPQ bound RNA (Figure 3.6 A, lane 4) could be excised from the gel.

RNA samples were purified from these bands and used for library preparation and sequencing. Despite a clean autoradiograph and nicely separated bands, PAR-CLIP_Harsh resulted in RNA fragments that were too short and not sufficient in quantity for binding site analysis. HeLa PAR-CLIP_Harsh was sequenced on the MiSeq, and ~2 million raw sequence reads returned for NONO and SFPQ (Table 3.1, HeLa PAR-CLIP_Harsh). FastQC analysis of the raw sequence reads showed PCR adapters accounted for a significant proportion of over-represented sequences. These adapters were removed with the FastX clipper tool, leaving 39,801 and 51,614 adapter extracted reads for NONO and SFPQ respectively. The high proportion of adapter contamination is a result of the very small RNA fragments that were used in the library preparation. Mapping was performed with the Bowtie aligner [158] allowing up to 3 mismatches. Three mismatches were allowed as the 4-SU incorporation and crosslinking induce a T-to-C transition that will result in reads that do not match perfectly to the genome. In addition, reads aligning to more than 10 sites in the genome were discarded. These reads are considered multi-mappers as it is not known from which transcript they originated. Using these parameters, PAR-CLIP_Harsh yielded 9,552 and 14,047 mapped reads for NONO and SFPQ, respectively. Thus, PAR-CLIP_Harsh resulted in very poor genome coverage, with reads not reaching a sufficient depth to be carried through the analysis pipeline. The high adapter contamination and large

number of multi-mapped reads was a result of very short input RNAs. As such, the RNase T1 digest was reduced for the next experiment, termed PAR-CLIP_Mild.

3.4.2 PAR-CLIP_Mild isolated longer RNAs.

PAR-CLIP_Mild (Figure 3.6 C) was established where RNA was digested minimally with 1 U/ μ l RNase T1 in lysate prior to IP for 5 min at RT. There was no second RNase T1 step, as it was judged that the RNase digestion after the IP was too harsh. An autoradiograph of the mild RNase T1 treated samples (Figure 3.6 C) shows bands corresponding to NONO and SFPQ in the crosslinked NONO IP sample, with no obvious signal in the controls (Figure 3.6 C, compare lane 4 with lanes 1-3). Interestingly, in the absence of the second RNase digest, the result is a radioactive smear extending from the top of the SFPQ band to the top of the gel (Figure 3.6 C, lane 4). This suggests this sample contained long RNA fragments crosslinked to SFPQ that migrated slowly during gel electrophoresis.

The PAR-CLIP_Mild experiment was performed and RNA isolated for deep sequencing. Figure 3.6 C shows a representative radiograph of the PAR-CLIP_Mild sample. The bands corresponding to NONO and SFPQ were not so obvious above the background, aside from a slight outward bulging in the gel lane, therefore the ladder was used to direct where gel slices were taken from. There is very little separation between the NONO and SFPQ radioactive bands, therefore, it is possible that there is some overlap in the RNA isolated. Sequencing was carried out on pooled RNA from 5 such experiments (totaling 100 dishes of HeLa, approximately 20 X 10^8 cells) and sequenced. The libraries made from the PAR-CLIP_Mild experiment (Table 3.1) contained RNA in a high yield and suitable length for library preparation, and the deep sequencing results gave good genome coverage, with distinguishable binding sites in several transcripts (the results are described in Chapter 4 and 5).

As with the PAR-CLIP_Harsh, the PAR-CLIP_Mild was sequenced on the MiSeq, but in contrast to PAR-CLIP_Harsh, over 20 million raw sequence reads were reported for each library (Table 3.1, HeLa PAR-CLIP_Mild). This indicated that increasing the number of cells to use in PAR-CLIP, and reducing the extent of the RNase digest

resulted in a better library for sequencing. Again, over half the sequence reads were removed with adapter clipping (as they contained solely adapter sequence, or were <15nt following adapter clipping), however, Bowtie was able to map ~5 million reads to the genome, and these mapped reads could be carried through the analysis pipeline.

A replicate PAR-CLIP_Mild experiment (PAR-CLIP_Mild_2) was sequenced on the IonTorrent platform, however, this data was found to be unsuitable for further analysis for two reasons. First, the Ion Torrent sequencing generated half as many raw reads for PAR-CLIP_Mild_2 as were reported from sequencing the original PAR-CLIP_Mild on the MiSeq platform. Secondly, alignment of the IonTorrent reads to the genome was performed using Bowtie and the same parameters as all other PAR-CLIP datasets, however, when aligned to the genome, the depth of sequencing along NEAT1 (which serves as a positive control) was insufficient for further analysis.

3.4.3 A PAR-CLIP_Medium experiment was carried out to refine NEAT1 binding sites.

One consequence of reducing the RNase T1 digest for PAR-CLIP_Mild was an over-abundance of reads across NEAT1, resulting in a loss of resolution for mapping binding sites (the binding of the proteins to NEAT1 is described fully in Chapter 4). Therefore, a PAR-CLIP_Medium digest condition was established as a compromise between mild and harsh. In this 'medium' protocol, 1 U/μl RNase T1 was incubated in the lysate for 15 min prior to IP (as opposed to the 5 min incubation for the 'mild'). An autoradiograph of the medium RNase T1 treated samples (Figure 3.6 B) shows bands corresponding to NONO and SFPQ in the crosslinked NONO IP sample, with no obvious signal in the controls (Figure 3.6 B, compare lane 4 with lanes 1-3). PAR-CLIP_Medium conditions were used to purify NONO- and SFPQ-bound RNA that was made into libraries and sequenced on the Illumina platform (Table 3.1).

The sequencing platform for PAR-CLIP_Medium was different to PAR-CLIP_Mild and Harsh. In this case, PAR-CLIP_Medium libraries were multiplexed with a total

of 6 samples in one lane of the Illumina HiSeq. The HiSeq produced many more raw reads than the MiSeq, and unlike sequences from the Harsh and Mild experiments, adapter contaminants accounted for fewer than half the reads in the Medium experiments. For these datasets from PAR-CLIP_Medium, Bowtie alignment reported ~15 million reads mapped to the human genome (Table 3.1).

3.4.4 A PAR-CLIP_Medium experiment was carried out in murine NIH3T3 cells

PAR-CLIP_Medium was also performed on mouse NIH3T3 fibroblasts to give insights into binding sites along NEAT1 in mouse as well as transcriptome wide binding sites. In this PAR-CLIP experiment only NONO bound RNA was able to be isolated and made into a library (Table 3.1). The NIH3T3 library was sequenced on the Illumina Hiseq, and approximately 15 million reads were mapped to the mouse genome as a result of this experiment.

3.5 Discussion

3.5.1 SFPQ is co-immunoprecipitated with NONO in PAR-CLIP experiments

One of the main findings of this chapter is that SFPQ and its associated RNA can co-IP in PAR-CLIP experiments with NONO in a 1:1 ratio (Figure 3.3 and 3.4). Despite its high sequence similarity to NONO, the isolation of SFPQ is not due to cross reactivity of the NONO antibody, as this antibody does not detect SFPQ when used for western blot (Figure 3.2, lane 7). Rather, the co-IP of SFPQ with NONO occurs from the formation of a stable heterodimer between the two proteins, an interaction that has been observed on multiple occasions and one that is believed to be important for the actions of these proteins in the cell [90,98,114]. This interaction is strong enough to withstand the 1.5 M salt conditions used here for the IP (eg. Figure 3.5, lane 2). It is known that these proteins form obligate dimers and the NONO/SFPQ heterodimer (as opposed to the PSPC1/NONO heterodimer) seems to be the preferred dimer for these proteins, at least *in vivo*, where PSPC1 levels are limiting (Gavin Knott, personal communication, and this study).

The other mammalian DBHS protein, PSPC1, is also present in the co-IP with NONO. Unlike SFPQ co-IP, it is not possible to determine if PSPC1 is being isolated due to its heterodimerization with NONO [130] or because of cross-reactivity with the NONO antibody, which recognizes an epitope in a region of homology between PSPC1 and NONO (Figure 3.2, lane 7). Given that the two proteins migrate closely during SDS-PAGE (migrating at 54 kD for NONO and 56 kD for PSPC1), and that the autoradiograph shows diffuse radioactivity around 55 kD, it is not possible to distinguish between NONO-RNA complex and PSPC1-RNA complex. However, given that the western blot shows only a small amount of PSPC1 is present in the co-IP with NONO (Figure 3.2, lane 7) it is reasonable to assume the majority of the isolated RNA is bound by NONO. The DBHS proteins share 70% sequence identity throughout their RNA binding domains and key RNA-binding residues are invariant. Furthermore the NONO/PSPC1 structure shows they are virtually indistinguishable from a structural perspective [130], therefore the presence of partner proteins in the IPs was not deemed to be a problem for further experiments, but is a factor that must nevertheless be taken into account.

3.5.2 4-SU incorporation and crosslinking is essential for isolation of RNA bound by NONO and SFPQ in PAR-CLIP

Crosslinking the RNA to the proteins is essential for the preservation of the RNA in complex with NONO and SFPQ within a PAR-CLIP experiment. In the absence of crosslinking, there was no RNA co-migrating with either NONO or SFPQ in SDS-PAGE (Figure 3.4), despite the fact that western blotting confirmed that both proteins were isolated in the absence of crosslinking (Figure 3.3, lane 2). It is possible that in the absence of crosslink formation, either the associated RNA is completely digested by RNases, or the RNA dissociates from the proteins when they denature either at the point of lysis or during SDS-PAGE.

3.5.2.1 Assessment of the effect 4-SU incubation and incorporation into RNA transcripts has on HeLa cells

Even though 4-SU incorporation and crosslinking is essential for the isolation of NONO and SFPQ bound RNA with PAR-CLIP, it is necessary to take into account the effect 4-SU may be having on the cells and interpret the results of PAR-CLIP in light of this. In the original PAR-CLIP experiments, Hafner et al. used 4-SU at 100 μM

concentrations for 14 h and did not observe any toxicity in HEK293 cells [139]. In contrast, Huppertz et al. recommend only a 60 min incubation of 4-SU in cells, in order to limit cell death [159]. In addition, Burger et al. reported that 4-SU at 100 μM concentrations impaired 47S rRNA synthesis in the cultured cells studied, including HeLa [160], and that this effect increases the longer the cells are incubated in the presence of 4-SU. It is likely that short incubation times with 4-SU will result in labeling of only the most highly expressed or rapidly turned-over transcripts, biasing the results. In the experiments reported here, incubation of HeLa cells with 100 μM 4-SU for 14 h did not result in any observable cell death. Thus, 14 h incorporation was employed as standard.

3.5.3 RNase T1 digest conditions were optimized for isolation of a RNA that is specifically bound by NONO and SFPQ in a sufficient quantity for deep sequencing analysis.

In the PAR-CLIP experiments, the RNA associated with the protein in the IP is visualized through radiolabelling and the SDS-PAGE is visualized as an autoradiograph. Thus, even if there are non-specifically bound proteins present in the IP eluent, they will not be visible unless they are binding RNA in a quantity sufficient for radiolabelling and visualization. To prevent non-specific binding of RNA and co-IP of other RNP complexes, the RNase digest conditions were optimized. The RNase conditions employed in PAR-CLIP_Harsh resulted in very clear, compact bands corresponding to SFPQ and NONO in complex with their bound RNAs (Figure 3.4, lane 4). The absence of any diffuse radioactivity extending above the bands allowed us to be confident the gel slices excised, the RNA extracted and the datasets generated from sequencing were highly enriched for the specific SFPQ or NONO bound RNA fragments. However, even though this harsh digest was able to give informative results when applied to other RNA binding proteins [154,155] it was not appropriate for isolation of DBHS protein-bound RNA. The RNA sequencing results from PAR-CLIP_Harsh treatment showed that this extensive RNase T1 digest resulted in a quantity of RNA that was too low for effective deep sequencing library preparation, and that the fragments were too short to map uniquely to the genome (Table 3.1). I trialed different times for which the RNase T1 digest should be carried out for, with the 5 minute digest in PAR-CLIP_Mild (Figure 3.6 C) and 15 minute digest in PAR-CLIP_Medium (Figure 3.6 B)

both giving higher background radioactivity than the PAR-CLIP_Harsh experiment (figure 3.6 A), suggesting that under less extensive digest conditions there is either more RNA bound to the IP beads or that the RNA bound by the proteins is longer.
The increased amount of RNA bound was reflected in the raw sequencing reads, with PAR-CLIP_Mild and Medium resulting in more raw reads than the Harsh digest experiment (Table 3.1).
It is important to note that even though the RNase digest step was shorter and therefore less extensive, the high salt conditions in the lysis and wash steps ensure a clean IP and lack of non-specific co-IP'd proteins. Thus the increased smears above the NONO and SFPQ bands likely reflect longer SFPQ- and NONO-bound RNA, rather than other RNA:Protein complexes. The only exception to this is the possibility of a small amount of PSPC1-RNA complexes contaminating the NONO fraction, as mentioned above.

Overall, the RNase digest conditions were optimized to achieve a balance between isolating enough RNA for sequencing library preparation, and isolating fragments that are large enough to map uniquely to the genome. In addition, the RNA fragments were small enough to obtain binding site specificity information. The optimum protocol for this purpose was the PAR-CLIP_Mild, with one 5 minute RNAse T1 digest in lysate, which allowed identification of transcriptome wide binding sites. In addition PAR-CLIP_Medium, with one 15 minute RNase T1 digest in lysate was employed in an attempt to refine protein binding sites in NEAT1.

4. Non-coding RNA targets of NONO and SFPQ identified in PAR-CLIP

In Chapter 3 I presented how PAR-CLIP was optimized for the isolation of RNA crosslinked to NONO and SFPQ. Following sequencing, the challenge was to extract meaningful biological data out of these reads. In this chapter I describe the experiments and analysis focused on the identification and characterization of NONO and SFPQ binding sites in lncRNAs. The initial focus is on NONO and SFPQ binding in NEAT1, with results indicating patterns of clusters along the transcript, potential binding motifs and the potential for NONO to bind structured RNA. The focus then shifts to NONO and SFPQ binding to MALAT1, an interaction that has not previously been reported. Other lncRNAs are then examined and experiments determining the effect NONO and SFPQ are having on the overall levels of these transcripts are described. In Chapter 5, I discuss NONO and SFPQ binding to coding RNAs.

To identify transcriptome-wide binding sites for NONO and SFPQ, PARalyzer software was used to analyze PAR-CLIP sequencing data [140,161]. PARalyzer is a software pipeline that generates groups and clusters from alignment data of PAR-CLIP reads. Figure 2.2 shows a schematic representation of PARalyzer analysis, outlining how groups and clusters are generated. Groups generally span a greater distance along transcripts, while clusters are typically shorter regions within groups. Clusters are reported when the nucleotides within the cluster meet a user-defined threshold of T-to-C conversion locations and events. In this study, the PARalyzer clusters were taken to represent high confidence, refined, protein binding sites in transcripts.

There are two key elements of PAR-CLIP that PARalyzer utilizes to identify NONO and SFPQ binding sites in their target RNAs to single nucleotide resolution. Firstly, due to extensive RNase T1 digestion, read depth is higher over the bound and protected nucleotides, as opposed to nucleotides that were not protected by the protein and thus digested away. PARalyzer parameters specify a minimum read count for groups and clusters to be reported, meaning that a defined number of reads must cover the region. In addition, PARalyzer analysis parameters also

specify that in order for a nucleotide position to be included in the cluster, it must be spanned by a defined number of reads. The number of reads spanning a nucleotide is dependent upon the read depth achieved in the sequencing run and this should be considered when setting read count for cluster inclusion. In the PAR-CLIP experiments outlined here, PAR-CLIP_Mild was sequenced on the MiSeq (average of 4-5 million reads per library), and thus had much lower read depth than PAR-CLIP_Medium, which was sequenced on the HiSeq (average 20-25 million reads per library). As such, a higher minimum read count for cluster inclusion was used for PAR-CLIP_Medium analysis to minimize false positives, as the read depth was higher across the genome so more sites would meet the thresholds for cluster reporting.

Secondly, 4-SU incorporation and irradiation with UV light induces crosslinking where the 4-SU is in contact with protein. These crosslink sites are seen as T-to-C nucleotide transitions in the mapped reads and can be used to determine protein-binding sites. PARalyzer parameters can be adjusted to specify the minimum number of separate locations required to have a conversion (T-to-C transition) for a cluster to be reported. In addition, PARalyzer only reports a cluster if a minimum conversion count is achieved across all the reads making up that cluster.

Unless otherwise specified, I carried out all bioinformatics optimizations and analysis detailed in this chapter.

4.1 PARalyzer was used to generate clusters representing binding sites in NONO and SFPQ bound transcripts

A process of optimization was undertaken to determine the PARalyzer parameters that would allow the identification of NONO and SFPQ binding sites in NEAT1 as well as other target RNAs across the whole transcriptome. A minimum read count for groups and clusters was set to 50 and the minimum conversion location and event count for a cluster was set to 5. NONO and SFPQ are abundant nuclear RNA binding proteins, and as such likely bind many RNAs transiently or non-specifically. This threshold of 5 conversions at 5 locations was found to minimize clusters reported in non-specifically bound transcripts (as defined by their mapping to repetitive elements, low transition event count, low read count or

absence of additional clusters to indicate genuine protein binding). The effectiveness of these parameters in minimizing reporting of non-specifically bound transcripts was assessed by the ability to report clusters in the positive control NEAT1, while reporting only a few, if any clusters in transcripts reported to be common PAR-CLIP contaminants (discussed in chapter 4.8.2).

Table 4.1 contains a summary of PARalyzer analysis and identified transcripts for PAR-CLIP experiments. PARalyzer analysis of PAR-CLIP_Mild reported 3,126 and 3,875 groups for NONO and SFPQ respectively, consisting of 438 and 477 clusters. To take into account the increased sequencing depth for PAR-CLIP_Medium, compared to PAR-CLIP_Mild, two different PARalyzer analyses were run to optimize detection of binding sites and minimize background. The parameters for 'strict' and 'relaxed' PARalyzer analysis of PAR-CLIP_Medium datasets were identical, with the exception of the minimum number of reads required for cluster inclusion: 'strict' with a minimum of 180 NONO reads and 205 SFPQ reads, and 'relaxed' with 90 NONO reads and 105 SFPQ reads (PARalyzer .ini files with full parameters are presented in appendix 3).

PARalyzer strict analysis reported 7,735 groups for NONO with 123 clusters, and 9,451 groups for SFPQ with 161 clusters. PARalyzer relaxed analysis reported the same number of groups, but these groups contained 379 clusters for NONO and 424 clusters for SFPQ.

Because mice and human have highly similar NONO and SFPQ, but divergent NEAT1 sequences (see chapter 1.3.1.1), PAR-CLIP_Medium was carried out in the NIH3T3 mouse cell line to give insights into how the DBHS proteins bind along NEAT1 in these cells. Only NONO associated RNA was isolated in NIH3T3 PAR-CLIP_Medium, as the SFPQ band in the PAR-CLIP autoradiograph was indistinguishable from background (Appendix 4). PARalyzer analysis reported 13,497 groups with 305 clusters for NONO PAR-CLIP_Medium in NIH3T3 cells.

4.2 Transcriptome wide identification of NONO- and SFPQ-bound transcripts from PAR-CLIP in HeLa cells.

To identify the transcripts in which clusters were found, clusters from each PAR-CLIP experiment for each protein were cross-referenced to the relevant genome (Chapter 2.3.5, hg19 for HeLa and mm9 for NIH3T3). Briefly, genome co-ordinates for clusters were cross-matched to the table of UCSC Known Genes, and gene names were extracted. Some clusters mapped to multiple locations in the same transcript, and in these cases for whole transcript analysis, the gene name was only counted once. In some cases clusters mapped to a region common to multiple transcript isoforms, making it difficult to discern whether one or all isoforms were bound. In these cases alternate isoforms were collapsed to one count. In cases where the isoforms were a mixture of coding and non-coding transcripts, for example UBE2V1, which was identified bound by NONO in PAR-CLIP_Mild, the transcript was attributed to both the 'coding' and 'noncoding' categories, as it was not possible to tell whether the protein bound the mRNA or ncRNA isoforms. This was also the case for read-through transcripts, for example RBM14-RBM4, bound by both NONO and SFPQ in PAR-CLIP_Mild, in which the clusters could correspond to RBM14, RBM4 or the read though transcript. Transcripts the proteins bind to in each PAR-CLIP experiment are detailed in table 4.2 (NONO bound transcripts) and table 4.3 (SFPQ bound transcripts).

Several transcripts identified with relaxed PARalyzer analysis of PAR-CLIP_Medium were not identified in PAR-CLIP_Mild (Table 4.4, asterisk) and are likely non-specifically bound transcripts that were detected when the threshold was lowered. From now on, unless otherwise stated, PARalyzer analysis of PAR-CLIP_Medium refers to the output from strict PARalyzer parameters, as these parameters prevent detection of non-specifically bound or 'background' transcripts.

4.2.1 Gene ontology and functional analysis give insights into the biological functions of NONO and SFPQ bound transcripts.

Once the NONO and SFPQ clusters had been cross-referenced to transcripts, it was of interest to investigate the types of biological functions these predicted NONO and SFPQ bound transcripts had in the cell. Gene ontology analysis was performed with Ingenuity Pathway Analysis (Chapter 2.3.7).

Ingenuity reported top diseases and biological functions for the 85-90% of PAR-CLIP_Mild transcripts for which information was available (Table 4.5). Overwhelmingly, NONO and SFPQ bound transcripts have a role in Cell death and survival networks. Cancer is the major disease caused by alterations in these pathways, however, cancer was not one of the top diseases reported for NONO bound transcripts in this analysis. In addition, both NONO and SFPQ are reported to play a role in infectious disease response networks, in agreement with recent literature reports.

Gene ontology analysis of the high confidence transcripts in PAR-CLIP_Medium reported roles in cell death and survival, cell cycle and cell signaling (Table 4.5). In addition, the majority of these transcripts in PAR-CLIP_Medium are implicated in cancer (64% and 60% respectively). Notably, these transcripts from PAR-CLIP_Medium are also reported to play a role in reproductive system disease. This is an interesting observation given the reproductive system defects associated with the Neat1 knockout mouse, suggesting a link between NONO and SFPQ RNA binding and paraspeckles in reproductive system function.

4.2.2 NONO and SFPQ bind both mRNA and ncRNA transcripts

NONO and SFPQ target transcripts were annotated as ncRNAs or mRNAs, with the proportions summarized in figure 4.1. The individual transcripts identified in each dataset are also detailed (Table 4.2 outlines the NONO bound transcripts and table 4.3 the SFPQ bound transcripts). 22% of the NONO-bound transcripts reported in PAR-CLIP_Mild are ncRNAs, while 78% are mRNA (Figure 4.1 A). Approximately the same proportions are seen with SFPQ bound transcripts detected in PAR-

CLIP_Mild (Figure 4.1 C). This suggests that the proteins could be binding together along the same transcripts. PAR-CLIP_Medium experiments detect a greater proportion of ncRNA transcripts for both proteins, indicating that ncRNAs may be the stronger and/or more abundant RNA targets of DBHS proteins (Figure 4.1 B and D).

4.2.3 NONO and SFPQ appear to coordinately bind several transcripts

NONO and SFPQ are known to interact and exist as heterodimers. This binding is strong enough to withstand the high salt washes used in PAR-CLIP, with the NONO/SFPQ heterodimer not separating until the IP sample is run on denaturing SDS-PAGE. Studies that investigate either NONO or SFPQ often detect the partner protein also, making it difficult to discern which protein is responsible for the function under investigation. The RNase digest steps employed in PAR-CLIP_Mild and Medium can help in determining those transcripts genuinely bound by the heterodimer compared to each homodimer. The PAR-CLIP_Mild experiment yielded a longer list of transcripts that were bound by both proteins: 93 transcripts were present in both the NONO and SFPQ RNA libraries in PAR-CLIP_Mild (Table 4.4 A). As expected, there is a much smaller list of common transcripts detected in PAR-CLIP_Medium, with only 9 transcripts bound by both NONO and SFPQ (Table 4.4 B).

For PAR-CLIP_Medium, relaxing the parameters resulted in the detection of more protein-associated transcripts that were common to both SFPQ and NONO (Table 4.4 C) although 7 of the mRNAs detected in this relaxed analysis (Table 4.4 C, asterisk) do not appear in the PAR-CLIP_Mild datasets, thus raising questions about their validity.

4.3 PAR-CLIP in mouse NIH3T3 cells predominantly reports clusters in Neat1 and Malat1

Clusters from PAR-CLIP in NIH3T3 cells for NONO were manually cross-referenced to the transcripts they mapped to (Appendix 5). Of the 11 transcripts that were

detected for NONO, 8 were mRNAs and 3 ncRNAs. Importantly, the vast majority of clusters reported for NIH3T3 PAR-CLIP mapped to Neat1 (205/305; 67%) and Malat1 (56/305; 18%).

Whilst it was expected that PAR-CLIP_Medium would digest most other RNA targets, leaving an abundance of Neat1 fragments, this is still a far greater proportion of Neat1:other RNA reads than was reported for the HeLa experiments. The same pattern was seen for binding in Malat1. This suggests that in human cells, NONO and SFPQ may be performing a wider variety of functions in binding different RNA transcripts, while in mouse, they predominantly bind Neat1 and Malat1. Alternatively, this may be due to the fact that HeLa are a transformed cell line, while NIH3T3 are immortalized fibroblasts. It could be that in transformed cells, NONO and SFPQ take on a wider variety of functions, binding a larger range of transcripts to regulate a wider variety of pathways.

4.4 PAR-CLIP identifies potential NONO and SFPQ binding sites in human NEAT1

To address the question of NONO and SFPQ binding sites in NEAT1, the clusters aligning to NEAT1 were separated for analysis. Sequences of the clusters located in NEAT1 for NONO and SFPQ are presented in appendix 6.

Figure 4.2 shows the clusters along NEAT1 for NONO, in both the PAR-CLIP_Mild and PAR-CLIP_Medium experiments. In PAR-CLIP_Mild, there were 1.56 times more clusters for NONO in NEAT1 (128 clusters) than in PAR-CLIP_Medium (82 clusters), indicating the clusters were greatly refined with the harsher RNase digest. Reassuringly, whilst the harsher RNase experiment reports fewer clusters overall, these clusters have much higher transition event counts than PAR-CLIP_Mild. Also of note is that PAR-CLIP_Medium analysis fails to report some of the high transition count clusters observed in PAR-CLIP_Mild. For example, the cluster with the highest transition count in PAR-CLIP_Mild (522 transitions) is not detected in the PAR-CLIP_Medium (Figure 4.2, arrow). This may be explained by random variations in 4-SU incorporation meaning that this segment of the transcript was not as tightly crosslinked in PAR-CLIP_Medium. Alternatively, the

absence of the high transition cluster in PAR-CLIP_Medium may indicate the fragment is located on the outside of the protein complex and thus not protected from the harsher RNase digest.

Figure 4.3 shows the clusters along NEAT1 for SFPQ, and a similar pattern to NONO is observed with respect to refinement in the harsher RNase experiment. However, in contrast to the NONO clusters, the maximum transition event count for a SFPQ cluster is similar between the PAR-CLIP_Mild and PAR-CLIP_Medium (transition event count shown only to 1000 in figure 4.3, arrows denote highest transition event clusters in each PAR-CLIP experiment). The maintenance of a high transition event count across the SFPQ clusters in both PAR-CLIP experiments is indicative of strong RNA:Protein interactions that are maintained despite the increased RNase digest. This suggests that SFPQ may have a more stable and reproducible binding to NEAT1 than NONO.

4.4.1 PAR-CLIP_Medium clusters along NEAT1 appear in a periodic pattern

The clusters along NEAT1 with the highest transition event counts were examined to see if they gave any clues about the macromolecular interactions between NEAT1, NONO and SFPQ within paraspeckles. SFPQ clusters with 700 or more transition events, along with the corresponding NONO clusters were highlighted and distances between these clusters calculated (Figure 4.4). A first striking observation was that high transition event count clusters common to NONO and SFPQ appear periodically along NEAT1 (Figure 4.4, blue boxes). This periodic pattern is most obvious in the PAR-CLIP_Medium experiment. A likely reason for this is that the increased RNase T1 digest left only the most tightly bound fragments that were protected by the proteins. The periodic pattern of clusters occurs within the longer isoform, NEAT1_v2, beginning at approximately nucleotide 4000 and continue to approximately 6000nt before the 3' end of the transcript.

Interestingly, DBHS protein dimers are predicted to form extended oligomers via interactions between residues in the coiled coil domain (Bond Lab, personal communication, Figure 1.5C). Extended oligomers of proteins are hypothesised to

bind along NEAT1, to allow proteins to organize into the paraspeckle complex along the scaffold RNA. These high transition event count clusters localize to the center of the transcript, suggesting this portion of NEAT1 could be located in the center of the complex where it is most in contact with the proteins.

4.4.2 Motif finding in the clusters along NEAT1 reveals protein binding is unlikely to solely depend on sequence

The clusters along NEAT1 represent protein-binding sites in the transcript. To investigate potential modes of binding to NEAT1, I examined the clusters for sequence motifs that occur frequently. Motifs occurring in multiple clusters would be indicative of protein recognition sites.

PAR-CLIP_Mild and PAR-CLIP_Medium clusters for each protein that aligned to NEAT1 were analyzed with the motif finding software MEME (Chapter 2.3.6). The three motifs with the lowest E-value (where E-value is the likelihood that that motif would occur by chance) are reported. The three top motifs in NEAT1 for NONO binding are shown in Figure 4.5 A and B. Unfortunately, these motifs cannot account for all binding as they do not occur in all clusters (the top site occurs 13 times in the 128 clusters for PAR_CLIP_Mild, and 9 times in 82 clusters for PAR-CLIP_Medium). Interestingly, however, a common theme of G-enrichment is apparent (the first 7 nucleotides in the top motif from PAR-CLIP_Mild are G[AG]GG[AC][TAG]G (Figure 4.5 A, #1). Looking beyond the clusters at the whole NEAT1 sequence, GGAGG appears 16 times in NEAT1_v1 and 54 times in NEAT1_v2, concentrating in the 5' and 3' ends of the transcript (Figure 4.5 E). This GGAGG motif is interesting, but it cannot explain NONO and SFPQ binding, given that this motif is not found in every cluster.

As with NONO, the top sequence motif in NEAT1 for SFPQ does not occur in many of the clusters (6 out of 154 clusters from Mild PAR-CLIP and 5 out of 104 from Medium). However, similar to the top NONO_Mild motif, this SFPQ motif is A and G rich. Motifs #2 and #3 for SFPQ in both experiments lack a clear consensus sequence, but do have a clear TG pattern. This TG also occurs in some of the reported NONO motifs. Overall the reported motifs for both NONO and SFPQ are

generally G rich. However, there are no motifs that occur in the majority of clusters, and it is difficult to build a consensus sequence in many cases. It is possible that the low number of RNA targets input into MEME have limited the ability of the software to identify potential binding motifs. From this analysis, while sequence-specific recognition cannot be discounted, it is unlikely to be the primary molecular phenomenon that directs the proteins to bind at specific sites in NEAT1.

4.4.3 Evolutionary conserved secondary structure prediction reports a number of structured regions among the NONO clusters in NEAT1

The absence of a clear sequence motif for NONO and SFPQ binding in NEAT1 suggests that the proteins may be recognizing and binding structured RNA regions instead of or in addition to a simple sequence. Based on the crystal structure of the NONO/PSPC1 heterodimer and work since (Bond lab, personal communication) we hypothesised that the RNA binding interface for a single DBHS protein monomer is likely to be approximately 4-6 nucleotides, on the assumption that RRM1 is binding RNA in a canonical sequence-specific manner. Given NEAT1_v2 is 23,000 nt long, it is highly likely it forms secondary structures, and that these structures are necessary for recognition and binding by the DBHS proteins. This is also a reasonable assumption given the DBHS proteins also bind mouse Neat1, which despite a divergent primary sequence, is likely to have conserved structural elements. To identify potential secondary structures the proteins may bind, the cluster sequences were examined for evolutionary conserved secondary structures. It is important to note that the field of evolutionary-conserved RNA structure prediction is in its infancy, therefore results need to be approached with caution.

NONO clusters from PAR-CLIP_Mild were analyzed by Martin Smith (Garvan Institute, Sydney). The method aligns the transcripts of interest from multiple species and predicts regions that have been structurally conserved through evolution. Detailed outline of the analysis pipeline is provided in chapter 2.3.5. The pipeline depends on a minimum sequence length of 100nt. Hence, cluster co-ordinates were merged if they were within 50nt of each other and extended by

50nt at both ends where necessary. Pairwise comparison of the clusters was used to calculate a tree of relatedness (Figure 4.6 A, high resolution image in appendix 7), with the NEAT1 regions in the red box denoting the regions that are the most significantly related at the structure level after 10K bootstrapping. Bootstrapping is a statistical method where resampling from the sample dataset multiple times (ie 10K, 20K) can improve the accuracy of the prediction of the population the dataset represents. Upon receiving these results, I extracted the sequences from the most statistically significant regions (Figure 4.6 A, chromosome coordinates in red box). These regions were viewed as a track on the UCSC genome browser (Figure 4.6 B, red box), alongside NONO clusters from PAR-CLIP_Mild (the clusters from which the conserved structure prediction was done) and PAR-CLIP_Medium (with refined clusters in NEAT1, giving more likely interaction regions). The 21 statistically significant related structures align along NEAT1, with 6 of these evolutionarily structurally conserved regions in the first 3.7 kb (common to NEAT1_v1 and NEAT1_v2). Interestingly, only a few of these regions correspond to clusters with the highest transition event count in PAR-CLIP_Mild. However, the PAR-CLIP_Medium NONO cluster with the highest transition event count (1402 transitions) is spanned by one of the statistically significant related structures, therefore this region was investigated further (Figure 4.6 B, green arrow).

4.4.3.1 One of the NONO clusters that is structurally conserved is predicted to form a G-Quadruplex

G-quadruplexes are a structure formed by DNA or RNA, that consist of stacks of planar tetramers of G nucleotides. G-quadruplex-containing RNAs are known to be involved in cancer [162,163]. G-quadruplexes were relevant in this project due to a recent report that NONO recognized a G-quadruplex structure [164]. Thus, the statistically significant related clusters (Figure 4.6 A, red box) were analyzed with QGRS Mapper [165], a program which predicts the likelihood of a sequence forming a G-Quadruplex. QGRS was used to generate G-scores (a scale of 1-100 for likelihood of G-Quadruplex formation) for each of the statistically significant clusters. The highest G-score (41) was returned for a sequence within the region shown (Figure 4.6 B, green arrow). Close inspection of this region (Figure 4.6 C) shows the predicted G-quadruplex lies just 6 bp upstream of the PAR-CLIP_Medium cluster

with the highest transition event count of 1402 transitions. This predicted G-Quadruplex also corresponds to the 3' end of a cluster in PAR-CLIP_Mild containing 71 transitions. Strikingly, the cluster with 1402 transitions in PAR-CLIP_Medium is absent in PAR-CLIP_Mild, suggesting this region was not bound by NONO to an extent that was detected in PAR-CLIP_Mild.

To test if NONO co-localized with G-quadruplex regions in RNA, NONO and G-quadruplex immunofluorescence was attempted. The BG4 antibody, which recognizes DNA and RNA G-quadruplexes was applied to HeLa cells, with co-staining with NONO antibody. Unfortunately these results were inconclusive as no G-quadruplex signal was observed within the nucleus. To follow up on this with a different approach, the G-quadruplex structure occurring in NEAT1 has been synthesized and is being investigated using *in vitro* protein binding assays within the lab of Charlie Bond.

4.5 PAR-CLIP identifies potential NONO binding sites in mouse Neat1

Mouse Neat1 was next investigated to determine if common themes for DBHS-NEAT1 binding would emerge. NONO_Medium clusters along Neat1 are shown in figure 4.7. As stated above, Neat1 has the highest density of clusters with the highest transition event count of all the transcripts crosslinked to NONO in NIH3T3 cells, indicating that, in this experiment at least, NONO binds Neat1 to a greater extent than any other transcript.

The highest transition event cluster, with 2136 transition events, is located in Neat1_v2, approximately 1000nt past the end of Neat1_v1 (Figure 4.7, arrow). In addition another high transition event count cluster is located at the 3' end of Neat1_v2, indicating that these two regions of the transcript are likely most often, and most extensively bound to NONO. Unlike the HeLa PAR-CLIP_Medium experiments, these data do not reveal any obvious periodic patterns of crosslinking. However, the data indicating the importance of Neat1_v2 binding sites (as opposed to Neat1_v1) is in accordance with the HeLa observations.

4.6 PAR-CLIP identifies potential NONO and SFPQ binding sites in MALAT1

MALAT1 is a highly expressed lncRNA implicated in cancer and cell cycle progression [60,67]. MALAT1 localizes to nuclear speckles in a variety of cell types, but unlike NEAT1 it is not essential for the structural integrity of these domains [47,82]. MALAT1 is a common contaminant in PAR-CLIP experiments due to its high expression, however, the numerous clusters along MALAT1 reported in the NONO and SFPQ PAR-CLIP experiments are comparable to those observed along NEAT1 (Figure 4.8 A). In light of this, it is possible that MALAT1 is a true NONO and SFPQ target transcript rather than a contaminant.

Surprisingly, the transition event count for NONO_Mild clusters is higher in MALAT1 than for the clusters in NEAT1. For all other datasets the transition event counts for the clusters is lower in MALAT1 than NEAT1. Figure 4.8 B shows the NONO and SFPQ clusters for PAR-CLIP_Mild and Medium along MALAT1. Accounting for the differences in transcript length, NONO clusters occur in NEAT1:MALAT1 with a 2:1 ratio, whereas SFPQ clusters occur with a 4:1 ratio. This suggests that perhaps SFPQ may have a higher affinity for NEAT1 than NONO. NONO may then, in turn, bind more widely to other RNA transcripts.

4.6.1 NONO and SFPQ binding motifs in human MALAT1 are likely not solely sequence dependent.

The NONO and SFPQ clusters mapping to MALAT1 were analyzed with MEME to determine if there was any sequence motif that NONO and SFPQ bind (Chapter 2.3.6). The top motifs in MALAT1 clusters for both NONO and SFPQ are very different to those found within NEAT1 (Figure 4.9). The PAR-CLIP_Medium motifs (Figure 4.9 B and D) do not match the motifs found in the Mild PAR-CLIP, instead, quite different motifs are reported in these clusters. This suggests that either the binding sites varied from experiment to experiment, or that MEME is not the most appropriate tool for this analysis (see chapter 4.8.4.1 for alternate methods for binding motif identification).

Overall, motif finding for NONO and SFPQ in MALAT1 did not result in a core consensus motif and could not be used to inform potential binding sites in other transcripts. As with NONO and SFPQ binding in NEAT1, although the requirement for a sequence recognition site cannot be ruled out, it is likely not the main protein-binding mode.

4.6.2 NONO binding in mouse Malat1 occurs to a greater extent than NONO binding in human MALAT1

The Neat1 and Malat1 genes are localized on chromosome 19 in the mouse genome, 23kb apart. PARalyzer analysis of PAR-CLIP samples in mouse NIH3T3 cells generated clusters indicative of NONO binding sites, the majority of which localized to Neat1 and Malat1 (Figure 4.10 A). The highest cluster in Neat1 has 2136 transition events, while in Malat1 the maximum is 1952 transitions in a cluster. Clusters with higher transition event counts are located in the 5' half of the transcript (Figure 4.10 B).

A comparison of human and mouse NONO binding to Malat1 indicates that NONO binding in Malat1 may be a more frequent event in mice than humans. Despite the fact that mouse Malat1 is nearly 2000nt shorter than human Malat1 (mouse Malat1 is 6,982 kb compared to 8,707kb for human), there are many more NONO clusters in Malat1 from NIH3T3 PAR-CLIP than HeLa PAR-CLIP. Even with relaxed PARalyzer parameters, which report 43 clusters in MALAT1 for HeLa PAR-CLIP, the NONO binding to mouse Malat1 is still occurring at a higher rate compared to the rest of the transcriptome. It is important to consider that transition events can only occur at T nucleotides, thus transcripts devoid of U's may have fewer clusters reported due to the transition event count threshold not being reached. To investigate whether the difference in the number of clusters in Malat1 for mouse and human was due to different transcript constitutions, the percentage of T nucleotides was calculated. For both human and mouse Malat1, T's account for 28% of the nucleotide content of the genes, suggesting the difference in the number of clusters reported is not due to an under representation and therefore absence of crosslinking in human Malat1.

4.7 PAR-CLIP reveals NONO and SFPQ bind a number of other ncRNAs

NcRNAs, and in particular lncRNAs are amongst the most dynamically regulated molecules in disease progression, with altered expression levels of many lncRNAs now being used as diagnostic and prognostic markers. More recently, lncRNA:Protein complexes have emerged as promising new therapeutic targets in cancer. In the context of NONO and SPFQ, two proteins that have established roles in the cells response to DNA damage, tumorigenesis and viral infection, their binding to a number of lncRNAs is beginning to reveal new mechanisms of action. SFPQ was recently shown to bind the lncRNA CTBP1-AS (an RNA transcribed anti-sense to the CTBP1 gene), resulting in the recruitment of SFPQ to the CTBP1 gene locus, where it triggers histone deacetylation to silence CTBP1, promoting tumor growth in prostate cancer [166]. NONO has also been shown to contribute to the increased expression of oncogenic genes via an interaction with a novel lncRNA, lncUSmycN [36]. Both these interactions represent novel therapeutic drug targets for the prevention of tumor proliferation. Important to note, while neither of these targets were detected in the PAR-CLIP experiment performed here in HeLa cells, it was of interest to interrogate the PAR-CLIP data to identify other lncRNAs bound by NONO and/or SFPQ, as these may give insight into other biologically relevant interactions and potential therapeutic targets. From table 4.4, we can see that there are 16 lncRNAs (besides NEAT1 and MALAT1) bound by both NONO and SFPQ in PAR-CLIP_Mild and 2 lncRNAs (besides NEAT1 and MALAT1) bound by both proteins in PAR-CLIP_Medium.

4.7.1 LINC00473

LINC00473 is a long intergenic non-coding RNA (LincRNA) transcript identified in the PAR-CLIP_Mild dataset. LINC00473, also known as C6orf176, was identified in a genome wide screen to find genes up-regulated with activation of the cAMP signaling pathway [167]. There appears to be co-operative binding by SFPQ and NONO to this transcript as the clusters are found at the same transcript locations within the SFPQ and NONO datasets (Figure 4.11). Interestingly, NONO and SFPQ appear to be preferentially bound to the unspliced transcript, in that the clusters

appear within the intron for LINC00473. This phenomenon of intron-binding is described in further detail in Chapter 5. The appearance of multiple clusters along LINC00473, even though with low transition event counts, suggests this transcript may be a genuine target. In addition, two adjacent SFPQ clusters at the 5' end of LINC00473 detected in PAR-CLIP_Medium suggests this transcript may have been bound tightly by SFPQ. For these reasons LINC00473 was chosen for further analysis (see below).

4.7.2 CCAT1

CCAT1 (Colon Cancer Associated transcript 1) is another lncRNA with highly relevant biological functions. CCAT1 induces the expression of the C-MYC oncogene in *trans* by facilitating chromosomal looping to bring the MYC enhancers and its transcription start site together [168]. CCAT1 has several NONO and SFPQ clusters along it, again primarily appearing within the intron (Figure 4.12). The clusters surrounding the annotated CCAT1 transcript are potentially informative. Unlike clusters reported for NEAT1 and MALAT1, which were strictly localized within the annotated transcript, the clusters are found beyond the 5' and 3' regions of the UCSC annotated CCAT1 transcript (Figure 4.12). Interestingly, there have been reports of a long isoform of CCAT1 [169], CCAT1-L, however CCAT1-L has a length of ~2,600nt, which still does not account for the extra length RNA observed here. When the UCSC track for lincRNA and TUCP transcripts is displayed [170] it is apparent that these clusters lie in the potential lincRNA TCONS_00015169, which spans beyond the UCSC annotated CCAT1 gene, and could potentially represent a transcript NONO and SFPQ are binding. Whether this TCONS_00015169 transcript is a longer isoform of CCAT1 or a different, novel transcript is unknown.

4.7.3 LINC00473 and CCAT1 RNA levels were unaffected by NONO and SFPQ knockdown by siRNA

To follow up on the observation of NONO and SFPQ binding to LINC00473 and CCAT1, their RNA levels were measured by RT-qPCR following knockdown of SFPQ, NONO, or both proteins. I hypothesized that the binding of NONO and SFPQ may result in some changes to the overall RNA levels and that any changes in RNA

level following DBHS knockdown could assist in identifying the role the proteins have in binding these novel lncRNAs. For detailed methods on knockdown, RNA extraction, RT-qPCR and qPCR data analysis see chapter 2.1.5, 2.1.6 and 2.1.7. Briefly, siRNA knockdown was performed in HeLa cells in triplicate, with cells harvested 72 hours post transfection for RNA extraction and western blotting. qPCR was performed on cDNA from the knockdowns to measure levels of candidate bound transcripts. CT values were normalized to β-actin and fold change relative to the scramble control graphed. T-tests were performed to determine whether the target RNA levels were statistically different between the scramble control and DBHS protein knockdown samples. The results shown are combined data from the triplicate experiments.

A representative western blot showing β-actin, NONO, PSPC1 (through cross-reactivity with the NONO antibody) and SFPQ levels following siRNA knockdown is shown (Figure 4.13 A). The optimized knockdown conditions I established resulted in a very efficient knockdown of SFPQ (Figure 4.13 B, compare SFPQ in lane 1 with lanes 3 and 4). NONO siRNA treatment, though significantly reducing the level of NONO in the cells, did not result in a complete reduction (Figure 4.13 A, compare NONO in lane 1 with lanes 2 and 4). Recently it was reported that PSPC1 was up-regulated upon NONO knockdown [171], this effect was also observed here, with PSPC1 up-regulation thought to functionally compensate for NONO loss.

NEAT1 levels were assessed following DBHS protein knockdown to serve as a positive control that the knockdown had affected bound lncRNAs as it was previously reported that NONO and SFPQ knockdown resulted in decreased levels of NEAT1 [81]. NONO knockdown results in an almost 10-fold reduction in NEAT1 levels compared with the scramble control, whilst SFPQ and double knockdown (NONO +SFPQ siRNA) results in a 4 fold change (Figure 4.13 B). This illustrates that the DBHS protein knockdown was effective and able to elicit a change in RNA levels.

LINC00473 RNA levels were measured following NONO, SFPQ and double knockdown in HeLa cells, with the fold change in RNA graphed relative to the scramble control (Figure 4.13 C). Unlike NEAT1, NONO and SFPQ knockdown did

not significantly alter the levels of LINC00473. Although a 5-10 fold increase can be seen, these values are not statistically significant, likely due to the large standard errors of the mean (SEM) showing variation from experiment to experiment.

CCAT1 levels are not significantly changed by either NONO or SFPQ protein knockdown or double knockdown (Figure 4.13 D). Like the LINC00473 qPCR results, CCAT1 qPCR also showed high SEM. However, CCAT1 levels in the knock down samples are only 1-1.5 fold higher than the scramble control, suggesting that, even though the SEM is considerable, CCAT1 is likely not affected by NONO or SFPQ knockdown to a large extent.

Thus, despite an effective knockdown, NONO and SFPQ depletion does not affect the levels of two target lncRNAs, LINC00473 and CCAT1. LncRNAs are known to be highly dynamic, with levels fluctuating through the cell cycle and in response to stress. Likely these natural fluctuations may be affecting the measured levels of the transcript and giving high SEM, which could be masking the effects of the knockdown. However, whilst there is no obvious effect on RNA levels, it is still possible NONO and SFPQ binding is influencing the actions of these, as well as other target lncRNAs in the cells, perhaps in combination with other proteins, or through other mechanisms.

4.8 Discussion

This chapter has detailed the optimizations that were made for the bioinformatic analysis of PAR-CLIP data and described investigations into the association with ncRNAs.

Binding sites in NEAT1 were assessed to give insights into possible mechanisms of NONO and SFPQ binding that may trigger the formation and facilitate the persistence of paraspeckles in the cell. A better understanding of the macromolecular interactions that occur within paraspeckles is important for future therapeutic targeting and as a model system for insights into the formation of lncRNA:protein complexes in general. In addition, the identification of other

lncRNAs bound by NONO and SFPQ will likely lead to future studies investigating their involvement in potentially novel pathways and gene expression mechanisms.

4.8.1 Careful optimization of the PAR-CLIP conditions and PARalyzer parameters is important to identify RNA targets of the proteins of interest.

PAR-CLIP conditions for the isolation of protein bound RNA, as well as the bioinformatics parameters to get meaningful data out of the sequencing need to be carefully tailored for the specific proteins of interest. The original PAR-CLIP study performed by Hafner et al. utilized a harsh RNase T1 digest [139]. It was suggested that such extensive digest with site specific RNase T1 (that cleaves at G nucleotides) would lead to depletion of binding sites containing Gs [157]. I established that harsh RNase T1 digest was not appropriate for DBHS proteins due to short, un-mappable RNA being isolated. This is potentially a reflection of the way NONO and SFPQ make contact and bind their target RNAs.

4.8.1.1 NONO and SFPQ form heterodimers and extended oligomers, with these interactions reflected in the binding site locations in target RNAs

The structure of the NONO/PSPC1 heterodimer suggests the proteins can make contact with 4-6nt of RNA, if binding with a classical RRM sequence-specific recognition mode [130]. Given the structural similarity of the DBHS proteins, the NONO/SFPQ heterodimer that is isolated in the PAR-CLIP experiments reported here is likely to bind a similar number of nucleotides. As the proteins are potentially interacting with such a short stretch of nucleotides, extensive digest will chew away all but the most protected fragment, some 4-6nt, which will not map to the genome. The short RNA isolated using the published PAR-CLIP RNase conditions supports the notion of a small RNA footprint for DBHS proteins.

Fortunately, the DBHS proteins oligomerize via their coiled-coil domain ([130] and unpublished work). These extended oligomers are speculated to make multiple contacts with RNA, thus, even if the binding interface is only 4nt for each protein, the combined effect increases the binding specificity and the oligomer of proteins could protect the RNA from extensive RNase T1 degradation.

In fact, NONO and SFPQ binding as a heterodimer is likely responsible for the large number of clusters reported in the target transcripts. In several other PAR-CLIP studies, it was the PARalyzer groups, not the clusters, that were taken as binding sites [161]. However, the nature of the NONO/SFPQ heterodimer, and the fact that it likely forms larger protein complexes means potentially many transcripts are so tightly bound and wound up in the complex that multiple RNA fragments were isolated. These longer overlapping reads mean the groups are very large, making it difficult to discern the interaction region. For this reason, clusters were used to infer binding sites, as, on the whole, they spanned shorter distances along the transcript than the groups. Nevertheless, analysis of the reported groups would be a simple way to get a transcriptome wide picture of protein bound RNAs. As the reporting of groups does not take into account transition events, they lie in transcripts that are bound loosely or in RNA fragments that may not be crosslinked. As a result, many thousands of target RNAs may result from this analysis. This is important information as it gives insights into possible transient interaction partners or those bound loosely, however, to validate these interaction it would be necessary to perform microarray analysis or RNA-Seq following protein knockdown to see which are directly affected by NONO and/or SFPQ.

4.8.2 PARalyzer parameters prevented the reporting of clusters in non-specifically bound transcripts

The PAR-CLIP experimental and bioinformatics protocols were carefully optimized to achieve conditions that minimized the reporting of non-specifically bound transcripts while maximizing the ability to detect genuine RNA targets across the transcriptome. The DBHS protein case is unusual compared to other RNA binding proteins subjected to PAR-CLIP. In other PAR-CLIP experiments, the abundant lncRNA NEAT1 is a common reported contaminant, however, in the case of the DBHS proteins, it is a genuine and biologically relevant target RNA.

The PARalyzer parameters were set to require a high read count for cluster reporting, resulting in appropriate inclusion of genuine transcripts, and exclusion of non-specific RNAs. The minimum read count for cluster inclusion was increased

for the analysis of the PAR-CLIP_Medium experiment so high confidence binding sites could emerge above the high read count background. However, simply taking the high read count sites as binding sites would result in a large number of false positives [172]. To overcome this, in addition to the requirement of a minimum read count, clusters were required to contain a minimum number of T-C transition locations. RNAs that are bound loosely by the protein may only make contact at a few locations, thus, when the cells are irradiated, minimal crosslinking will occur, conversely, more tightly-bound RNAs have more frequent crosslinks. It is important to remember here, however, that crosslinks in this PAR-CLIP are only made at the sites of 4-SU incorporation, which occurs at (genome-encoded) T nucleotides. Therefore, RNA from a gene devoid of Ts will not crosslink even if it is bound by the protein [172]. To overcome this, another ribonucleoside analogue could be used in future experiments, such as 6-SG (6-thioguanosine, incorporated in place of G nucleotides) and crosslinks determined by G-to-A transitions [173].

Part of the PARalyzer optimization process involved analyzing the dataset with several different parameters for transition location and transition event for cluster inclusion. This analysis focused on human chromosome 11 (the location of NEAT1) and the X chromosome (the location of XIST and FTX, two common PAR-CLIP contaminants). Parameters that were strict enough to prevent reporting clusters in contaminants while still detecting them in other transcripts on chromosome 11 ensured PARalyzer was able to detect transcriptome wide binding sites. I also aimed to find PARalyzer parameters that were strict enough to refine the clusters reported along NEAT1 to determine binding site motifs or patterns. However, the large number of transition locations and events along NEAT1 meant that any parameters that brought a few clusters above the background minimized the detection of clusters on other transcripts in the genome.

To overcome this, and refine binding sites along NEAT1, the experimental conditions for PAR-CLIP were altered. The harsher RNase T1 digest refined binding sites in NEAT1 resulting in the emergence of a periodic pattern, while also enhancing the ability to detect only the most tightly bound transcripts. The transcripts that had reported clusters in both PAR-CLIP_Mild and Medium (Table 4.4) are considered to be reproducibly bound and will form the basis of future follow up experiments.

4.8.2.1 lncRNAs are common PAR-CLIP contaminants, but the PARalyzer parameters were optimized to minimize their detection

Nuclear RNA binding proteins are notorious for binding non-specifically to RNA. In my data, several transcripts with mapped clusters are potentially non-specifically bound RNAs. For example, ANKD30BL is a lncRNA with several NONO and SFPQ clusters (data not shown). Notably, these clusters align to a region of ANKRD30BL with an rRNA repeat element, suggesting these clusters may be an artifact of a high number of reads misaligning to the repeat.

There are several ncRNAs that have been reported as common PAR-CLIP contaminants, including NEAT1. NEAT1_v1 is highly abundant in the nucleus, however the long isoform, NEAT1_v2, is found at a much lower level [78]. The presence of high transition event count clusters along NEAT1_v2, which is far less abundant than NEAT1_v1, supports specific binding, as does the literature on an essential role for SFPQ and NONO in NEAT1 binding in paraspeckles [47,78].

The ncRNA FTX, which is another reported common contaminant in PAR-CLIP experiments [161], is detected in NONO and SFPQ PAR-CLIP_Mild, with one cluster found in this transcript (Data not shown). Encouragingly, PAR-CLIP_Medium reported no clusters in FTX, indicating the harsher RNase treatment and the strict PARalyzer parameters were sufficient to remove this non-specifically bound RNA.

XIST and MALAT1 are two other reported common contaminants of PAR-CLIP datasets [172]. XIST is not present in any of the PAR-CLIP experiments, and, although its abundance was not measured in these HeLa cells, its expression is high in HeLa-S3 ([18], Long RNA from ENCODE/Cold Spring Harbor Lab, GEO accession number GSM897079). MALAT1, which has been reported to be the most abundant RNA in cancer cells [174], is detected here with both NONO and SFPQ clusters localized throughout. The fact that MALAT1 clusters are also detected with the harsher RNase treatment, the high number of transitions across multiple clusters, and the detection in both mouse and human MALAT1 all argue that this is a real target. Nevertheless, given the high abundance of MALAT1 this result should still be

treated with caution. The possible significance of the MALAT1-DBHS interaction is discussed further below in chapter 4.8.5.

4.8.3 The biological functions of the NONO and SFPQ bound RNAs agree with the reported roles for these proteins

The Ingenuity Pathway Analysis for NONO and SFPQ target transcripts supports a role for these proteins in cancer and infectious disease (Table 4.5). Interestingly, despite well documented roles in cancer [175-177], the transcripts bound by NONO did not fall into this category. This may, in part, be due to the fact that transcripts bound to NONO in PAR-CLIP_Mild have roles in other diseases besides cancer. However, the molecular and cellular functions assigned to the NONO bound transcripts include cell death and survival and cellular growth and proliferation, both functions that implicate these molecules in cancer progression. In addition, the majority of transcripts bound by NONO from PAR-CLIP_Medium are involved in cancer. Given PAR-CLIP_Medium isolates only the most tightly bound transcripts, it is strong evidence that NONO plays a role in cancer through its RNA binding function.

There are many lines of evidence that NEAT1 and paraspeckles pay a role in the cellular response to viral infection. NEAT1 is up-regulated in mouse cells infected with Japanese encephalitis virus [109]. NEAT1 is also up-regulated in human cells transfected with poly I:C dsRNA to mimic viral infection [87]. In this latter study the molecular mechanism was described as SFPQ sequestration into paraspeckles thereby influencing gene expression. It was therefore intriguing to observe a number of transcripts crosslinked to NONO and SFPQ that are involved in infectious disease pathways, suggesting that these transcripts may be bound and regulated by these proteins, perhaps in paraspeckles.

The interactions NONO and SFPQ form, both as part of the larger paraspeckle complex or as simple heterodimers in the nucleoplasm raises interesting possibilities about therapeutically targeting these complexes. In the case of certain types of cancer, targeting the interactions that form paraspeckles may help prevent the anti-proliferation role it seems to play [108]. Conversely, in an immune response, enhancing the interaction may increase the antiviral effect observed [87].

4.8.4 Identification of NONO and SFPQ binding sites in NEAT1 gives insight into the internal organization of paraspeckles

The binding of NONO and SFPQ along NEAT1 is considered to be the basis for paraspeckle formation. However, the molecular details of this binding are mysterious.

The NONO/SFPQ heterodimer is predicted to bind a short RNA sequence. Looking for such a small sequence motif is difficult with motif finding software such as MEME as it does not take into account gaps [178]. Given the propensity of NONO and SFPQ to form extended oligomers, any short motifs that are bound and separated by gaps will not be found. MEME analysis can, at best, predict that the proteins are binding G-rich regions in NEAT1. This may be a biologically relevant finding as another paraspeckle protein, TDP-43 has been reported to bind simple UG repeats in NEAT1 [179]. Given 4-SU incorporates at the sites of U nucleotides, these regions in NEAT1 would be expected to crosslink well to the proteins and display high transition event counts. SELEX experiments to determine recognition motifs for NONO and SFPQ both yielded G-rich motifs [127,137]. Given PAR-CLIP determines *in vivo* binding sites, this strengthens the idea that NONO and SFPQ are binding G-rich regions. The G-quadruplex finding is encouraging in this regard, however the absence of clear predicted G-quadruplexes in each of the binding sites suggests that either the predictions are not accurate, or the structure is more complex than a G-quadruplex.

RRMs are unlikely to function solely via the recognition of sequence motifs, as 4-6 nucleotides is too short for specificity and would appear many times in the transcriptome [180]. To overcome this, RRM containing proteins often have more than one binding module to give target specificity. This combined binding effect is best illustrated by PUF family proteins, where each binding module in the protein only recognizes a single RNA nucleotide, but a tandem arrangement of binding modules combine to give target specificity [181].

Even more target specificity is achieved when RNA binding proteins form dimers, for example bringing modules in the protein together to modulate their interaction with target RNA [182]. This is likely to be the way the DBHS protein achieve target specificity. By forming homodimers and heterodimers, these proteins can bring their RRMs together in different combinations to target different RNAs. This presents an added layer of complexity to the analysis of PAR-CLIP data. Many of the high transition event count clusters observed in NEAT1 are common to both SFPQ and NONO (Figure 4.4, blue boxes). The linker between RRMs in a protein have been postulated to confer RNA binding affinity [183,184] and the periodic pattern of binding sites in NEAT1 could be a reflection of the arrangement of RRMs in the heterodimer and the distance between them. In addition, oligomerisation via the coiled-coil would also increase the binding specificity and influence the way the proteins interact with the RNA. To investigate this possibility, the structure of NONO/SFPQ heterodimers could be examined to predict the likely distance between two RRM binding sites, based on the length of the NOPS domain and the lengths between heterodimers in an extended oligomer. However, such length predictions will need to be approached with caution, as RNA has a propensity to fold to form secondary structures, and thus predicting binding sites based on the length of RNA is not as simple as counting nucleotides.

RNA secondary structure is essential for RNA recognition and binding, especially by tandemly arranged RRMs. The protein HuD recognizes a simple sequence motif, but that motif must be embedded within a specific RNA structural conformation to confer binding specificity [185]. The protein U1A recognizes a sequence when it is located within a stem loop with much higher affinity than when it is in single stranded RNA region [186]. In contrast, some RNA binding proteins recognize specifically linear RNA. Structural studies on the sxl (sex lethal) protein in drosophila reveal it binds to a U-rich sequence in its target RNA [187]. This stretch of uridines does not form base pairs, so the tandemly arranged RRMs bind ssRNA [187]. The recognition and binding of NEAT1 by the DBHS proteins in both mice and humans, despite divergent RNA sequences, suggests that recognition and binding specificity is due to RNA secondary structure. The presence of structured RNA within the binding sites predicted with PAR-CLIP was examined, however, RNA structure prediction is an evolving field, and has its limitations.

4.8.4.1 The challenges of RNA structure prediction

Originally, RNA structure prediction was performed *in vitro*, with various kinds of RNase protection assays. These involved *in vitro* transcription of the RNA, RNase digestion and interpreting the pattern of resulting RNA fragments to infer structure [188]. However, RNase protection assays have limitations including variable kinetics of the digest, the interpretation of the foot-printing gel, effects confounded by a mixture of sequences [189] and non-detection of secondary structures triggered by protein binding [190,191].

Computational methods for structure prediction are now used widely. The software can predict structures such as hairpins and loops [192,193] and has evolved to predict complex tertiary structure predictions involving interactions between multiple RNA species [194,195]. More recently there has been a shift towards taking RNA sequences from multiple organisms and looking for evolutionarily conserved secondary structures. This specifically addresses structure prediction in ncRNAs. The sequences for ncRNAs are under much less evolutionary constraint that the sequences of coding RNAs, as such, the sequences of ncRNAs can be highly divergent between organisms. The NEAT1 sequence differences in human and mouse exemplify this. While mutations can be tolerated in ncRNAs to a greater extent than in mRNAs, if the ncRNA function is dependent upon its folding into a structured molecule, mutations will only be tolerated as long as complementary base pairing between nucleotides involved in the structure is maintained [196]. Several programs now use RNA sequences from several organisms to identify evolutionarily conserved secondary structures [30,197].

Perhaps the best RNA structure prediction tool is one that combines computational structure prediction with results from *in vivo* assays. One such structural analysis utilizes dimethyl sulfate (DMS). DMS is added to growing cells, where it triggers methylation of unprotected Adenosine and Cytosine residues. In structured RNA, such unprotected nucleotides are found in bulges, loops and linker regions, thus identification of unprotected nucleotides is used to infer structure. DMS methylation sites can be identified with high throughput sequencing in a technique

known as Structure-Seq[198]. Other *in vivo* chemical modifications of RNA are possible[199] with these also able to be taken into consideration by structure prediction software [200].

However, evolutionarily conserved structure prediction using the PAR-CLIP clusters revealed a limitation of combining *in vivo* and computation experiments. While PAR-CLIP was optimized to identify NONO binding sites to single nucleotide resolution, in many cases the clusters were too short for multiple species alignments and structure prediction and needed to be extended to 100 nt in these cases. Thus the precise binding site information obtained through PARalyzer analysis was lost in these extended clusters. New software packages that overcome these limitations will be invaluable for future analysis.

4.8.5 LncRNAs fulfill a variety of biologically relevant roles in the cell, with a few identified bound by NONO and/or SFPQ particularly interesting

4.8.5.1 MALAT1

This study is the first to show DBHS proteins binding MALAT1 with PAR-CLIP. Supporting the validity of this finding, a recent mass spectrometric analysis of proteins binding to MALAT1 identified both NONO and SFPQ [201]. The distribution of clusters along the MALAT1 transcript is more similar to that for NEAT1 than any other transcript in the genome (eg other lncRNAs or mRNAs). For the majority of the other reported transcripts, the clusters are fewer and farther between. This suggests the mechanism of DBHS binding to MALAT1 may be similar to NEAT1. Alternatively, since these are two of the most abundant single-exon lncRNAs in the genome, it could be due to abundance and gene structure. While MALAT1 does not localize to paraspeckles, NONO and SFPQ are found diffusely throughout the nucleus, as well as residing in paraspeckles [78], therefore it is likely they are binding MALAT1 outside of paraspeckles. Recently, both MALAT1 and NEAT1 were shown to crosslink to chromatin of active genes [201]. Thus, a possible role for DBHS binding to both MALAT1 and NEAT1 may be to target these lncRNAs to pre-mRNAs of active genes, which is likely to be a paraspeckle-independent role. In addition, there is further evidence that MALAT1 can bind SFPQ: Ji et al. found that the

chromatin modifier PTB2, SFPQ and MALAT1 compete for binding, with a flow-on effect on gene regulation, enhancing proliferation and metastasis in colorectal cancer [59].

4.8.5.2 LINC00473

The identification of multiple NONO and SFPQ clusters in LINC00473 (Figure 4.11) suggest this transcript is specifically bound by these DBHS proteins. However, knockdown of these proteins had no effect on the overall levels of LINC00473, suggesting the proteins were not affecting the stability of this lncRNA, or its transcription. This presents the possibility that NONO and SFPQ are binding this lncRNA to fulfill some other role in the cell. Little is known about LINC00473, besides its induction with cAMP pathways. NONO itself is involved in the activation of the cAMP signaling pathway, yet the relationship with LINC00473 is unclear in this context [202]. One possibility is that LINC00473 acts to remove NONO from its target sites, acting as a 'sponge', to prevent activation of cAMP signaling. Alternatively, NONO may bind and trap LINC00473 to prevent it from triggering expression of some target genes as part of the cellular network mediating the response to cAMP signaling [167]. Future experiments will be required to test these hypotheses.

4.8.5.3 MYC-associated LncRNAs

Two additional lncRNAs found crosslinked to NONO and SPFQ originate from loci near the C-MYC gene: PVT1 and CCAT1. The C-MYC gene encodes a transcription factor that is essential for development, but importantly is one of the strongest factors promoting transformation [203] and is also a proto-oncogene [204]. Constitutive MYC activation, either through mutation or activation of the pathways that trigger its expression promotes cancer development [204,205]. C-MYC expression is affected by several lncRNAs, including CCAT1, which up-regulates C-MYC by facilitating chromatin looping to bring distal enhancers closer to the C-MYC locus [169]. Interestingly, Xiang et al used NONO as a negative control for studying proteins binding to CCAT1, showing an absence of NONO binding in *in vitro* assays. This discrepancy with the PAR-CLIP observation reported here can be explained, given that the CCAT1 transcript used in their RNA pull down study was derived from the

spliced cDNA, therefore it only contained exons, wheras binding to the CCAT1 intron was reported in my PAR-CLIP data (Figure 4.12).

Excitingly, it may be that there is a general mechanism for NONO and SFPQ binding lncRNAs proximal to MYC genes in order to enhance proliferation. A recent report showed NONO bound the lncRNA LncUSMycN, a novel lncRNA found near to the N-MYC gene [36]. This is relevant for aggressive neuroblastoma patients with amplifications of the regions surrounding the N-MYC gene, including the gene for lncUSMycN. All three molecules, NONO, LncUSMycN and N-MYC act to drive proliferation in neuroblastoma. Further, NONO over expression correlates with disease progression of neuroblastoma and poor patient prognosis [36]. Unfortunately, the PAR-CLIP experiments described in this thesis could not examine this system, as N-MYC is not expressed in HeLa cells. Nevertheless, given the association of DBHS proteins with lncRNAs expressed near to C-MYC in Hela, this is a mechanism that should be investigated in future experiments.

4.8.5.4 The lncRNAs LINC00473 and CCAT1 may be acting as guides to bring NONO and SFPQ to specific locations in the cell

The fact that DBHS knockdown did not alter the levels of some of the candidate lncRNAs tested here does not preclude a genuine biological interaction. Indeed, DBHS binding contributing to the stability of these molecules cannot be eliminated, as it may be possible PSPC1 up-regulation with NONO knockdown is compensating for its loss. Beyond stabilization however, other mechanisms may also be at play. One possibility is that these lncRNAs are acting as competitors to target NONO and SFPQ away from other lncRNAs. This would be similar to the way the lncRNA AIR targets the chromatin modifying complex G9a to certain loci [206] and lincRNA-p21 targets HNRNP-K to its target sites in chromatin [207]. Future experiments involving ablation of these novel lncRNAs would be interesting in this context.

In summary, I have presented the NONO and SFPQ binding sites identified in ncRNAs. In chapter 5, I will address the NONO and SFPQ binding sites in mRNAs, to address how these proteins act on coding transcripts and the possible biological outcomes of this binding.

5. mRNA targets of NONO and SFPQ Identified in PAR-CLIP

It is thought that a significant proportion of DBHS protein function is mediated by RNA binding. Whilst NEAT1 is the most well known NONO- and SFPQ-bound transcript, there are many recent reports linking them to many cellular pathways through their ability to bind a variety of other RNAs [36,114]. This work is the first transcriptome-wide identification of NONO and SFPQ binding sites. It has not only identified a number of DBHS target ncRNAs, but also many target mRNAs. Identifying and characterizing mRNA binding by NONO and SPFQ will give insights into the cellular pathways these proteins are implicated in.

In this chapter I focus on the RNAs with coding potential, and assess the location of binding sites and features of these RNAs. I also examine the effect of DBHS protein knockdown on some candidate RNAs. Unlike the lncRNAs evaluated, NONO and SFPQ knockdown altered the expression of their target mRNAs, indicating they may be acting to post-transcriptionally regulate these targets.

5.1 NONO and SFPQ bind predominantly in the first intron of mRNAs

NONO and SFPQ fulfill a number of roles in the cell, and some of these roles require they bind specific transcripts (Chapter 1.5.1 has a summary of previously known NONO and SFPQ RNA targets). With PAR-CLIP it is possible, for the first time, to get an overall picture of the types of transcript features NONO and SFPQ bind.

NONO and SFPQ clusters from PAR-CLIP_Mild were analyzed with a custom-written python script to cross reference cluster co-ordinates with transcript names and IDs from the UCSC known genes file (see materials and methods chapter 2.3.5 for more details of analysis pipeline and script information). The clusters that fell in mRNA transcripts were annotated for the location within the transcript: 3'-UTR, 5'-UTR, exons, first introns, non-first introns (ie. Introns, but not the first intron) and splice junctions. Cross-referencing the cluster coordinates with these features for each transcript yielded two striking observations (Figure 5.1). Firstly none of the SPFQ clusters and only 9% of NONO clusters were found within exons. The low proportion of clusters in exons (Figure 5.1, green bar in NONO, absent in SFPQ) suggests the proteins either have a strong preference for binding pre-mRNAs or do

not have any binding sites in exons that were detected with this PAR-CLIP_Mild experiment. It is important to note that whilst only a small portion of the transcribed RNA in human cells consists of exonic sequences [9], these are nevertheless the most abundant parts of a transcript, as can be seen with any RNA-seq experiment. The 'abundance' is a reflection of the stability of the mRNA compared to the pre-mRNA. Further, PAR-CLIP experiments and data analysis were optimized to minimize the reporting of clusters from random reads aligning to the more abundant regions of the genome. Thus, thus absence of clusters in exons is remarkable. Secondly, over 89% of the clusters for both NONO and SFPQ map to introns in these transcripts (Figure 5.1, combine blue and red bars). Of the remaining clusters, 2% of those in the NONO dataset and 3% in the SFPQ dataset map to 3'-UTRs (figure 5.1, purple bars) and none of the reported clusters mapped to 5' UTRs.

5.1.1 NONO and SFPQ bind in long first introns

Examination of individual transcripts bound by the proteins showed many contained clusters in the first intron, suggesting that there may be a predilection for binding long first introns (for example, see figure 5.9 for clusters in the long first intron of ADARB2 mRNA). I therefore calculated the length of the first introns that contained clusters, and determined the proportion of first introns for different lengths (Figure 5.2). There were no first introns with NONO or SFPQ binding sites shorter than 1 kb. Strikingly, NONO and SFPQ predominantly bind in first introns that are between 10 and 100kb. Further, 18% of the first introns NONO binds and 23% of those with SFPQ binding sites are very long, at over 100 kb. The PAR-CLIP data shows that NONO and SFPQ bind first introns that are on average 65-78 kb long with a median length of 51-60 kb. The median first intron length of UCSC annotated transcripts is approximately 3kb, suggesting NONO and SFPQ are binding in exceptionally long first introns. This observation is particularly intriguing given recent reports of the two other paraspeckle proteins, TDP-43 and FUS, binding in long first introns to stabilize these transcripts [208,209]. These paraspeckle proteins, and their connection with DBHS proteins will be discussed further in chapter 5.5.1.2.

5.2 PAR-CLIP Mild revealed a number of DBHS bound transcripts, however, PAR-CLIP_Medium clusters revealed high confidence binding sites.

The moderate RNase T1 digest in PAR-CLIP_Mild resulted in many PARalyzer clusters reported and a long list of transcripts bound by NONO and SFPQ (Table 4.1, HeLa PAR-CLIP_Mild). However, as only a mild RNase digest was employed, it is possible that some non-specifically or weakly-bound transcripts were isolated. Nevertheless, PAR-CLIP_Mild was able to give a picture of the types of transcripts the proteins were binding, the feature types these binding sites were located in (Figure 5.1) and the possible cellular functions this binding was achieving (Figure 4.1 GO analysis).

The harsher RNase T1 digest allowed the refinement of the list of target transcripts. In PAR-CLIP_Medium a small number of transcripts were crosslinked to NONO and SFPQ (Figure 4.1). This result suggests that, while many transcripts bind weakly to the proteins and will be isolated following a mild RNase T1 digest, these transcripts fall below a stringency threshold following a harsher digest, or, more specifically, may not be as well protected by the protein, leaving fragments that are too small to be isolated, sequenced or mapped, or with too few T-C transitions to meet PARalyzer criteria.

PAR-CLIP_Medium not only refined the list of transcripts, but also served as a replicate PAR-CLIP experiment of sorts. Whilst the RNase T1 digest conditions employed in the two PAR-CLIP experiments were different, the other conditions were the same and thus transcripts that were detected in both experiments can be viewed with high confidence. There were 5 transcripts with coding potential crosslinked to both NONO and SFPQ in both the Mild and Medium PAR-CLIP experiments: HFM1, DAZAP1, GPI and PDE3A (Table 4.4). RYR2 was also identified in both PAR-CLIP experiments with both proteins, however these clusters mapped to repeat elements so this transcript was not pursued for further investigation.

5.3 DBHS proteins bind coding RNAs involved in cancer and infectious diseases

5.3.1 The mRNAs for a number of cancer associated proteins seem to be preferentially bound by SFPQ

The harsher RNAse treatment employed in PAR-CLIP_Medium was optimized for the isolation of the most tightly bound RNA fragments. Interestingly, four transcripts with SFPQ clusters in the Medium experiment did not have corresponding NONO clusters, although many were detected crosslinked to both NONO and SFPQ in the PAR-CLIP_Mild experiment. The fact that these RNAs were only detected with SFPQ in the harsher RNase experiment was puzzling given SFPQ is isolated in PAR-CLIP through its interaction with NONO, therefore, any transcripts with SFPQ binding sites were likely part of a complex with NONO also. Nevertheless, these transcripts have interesting biological attributes and may be worthy of future investigation. The first of these transcripts encodes the IGF1R (insulin growth factor 1 receptor), an extracellular receptor that binds insulin growth factors triggering a signaling cascade that activates a number of protein kinases to trigger cell growth and differentiation, cell division, proliferation and transformation [210]. IGFR1 is up-regulated in a number of cancers [211,212], where it contributes to tumorigenesis. NONO and SFPQ clusters were identified in the long second intron of IRF1R mRNA (data not shown). Other SFPQ-specific target transcripts were BASP1 (Brain abundant, membrane attached signal protein1) and PAPD7, both with clusters in their long first intron. MATR3 (Matrin 3) was originally identified as a nuclear matrix protein [213], and has since been implicated in binding and stabilizing viral RNA to facilitate its nuclear export and expression [214]. MATR3 mRNA was isolated with SFPQ in the Medium but not the Mild PAR-CLIP experiment (Figure 5.3). The SFPQ cluster maps to the first intron of MATR3, contains 164 transition events and does not correspond to any repetitive DNA elements suggesting it is likely to be a real binding site.

5.3.2 High confidence RNA targets bound by both NONO and SFPQ

Given NONO and SFPQ predominantly exists in the cell as a heterodimer, it was of interest to look closely at the transcripts that were bound by both proteins. The

reporting of clusters in both protein PAR-CLIP datasets gives more confidence that the transcripts they map to are genuine DBHS targets. In addition, binding by both proteins suggests they are acting as a heterodimer to facilitate their action on the transcript.

There are 75 mRNA transcripts that contain both NONO and SFPQ clusters in PAR-CLIP_Mild (Table 4.4). These transcripts are predominantly involved in cancer (75% of the transcripts) and infectious disease (24% of the transcripts) when analyzed with Ingenuity Pathway Analysis. The 5 transcripts bound by both proteins in PAR-CLIP_Medium (Table 4.4) likely represent transcripts that are most stably and reproducibly bound by NONO and SFPQ in HeLa cells. Again, these transcripts are all associated with cancer.

5.3.2.1 HFM1 has both NONO and SFPQ clusters, suggesting it is bound by both proteins.

HFM1 mRNA codes for a DNA helicase thought to help maintain the integrity of the genome during cell division, in particular in germ line cells [215]. At least one cluster is reported for each protein in each PAR-CLIP experiment, with all clusters mapping to a middle intron of HRM1 (Figure 5.4 A). All clusters are located in close proximity, suggesting the proteins bind tightest within this intron. While all clusters have a high transition event count, the highest occurs in an SFPQ_Mild cluster, with 4363 transitions. A closer view of these clusters (Figure 5.4 B) shows that there are no SNPs within the clusters that would be counted as transition events (if the SNP was T-to-C). The large number of transitions within the SPFQ cluster is an effect of extensive crosslinking and a large number of reads making up this cluster, both due to a tight association between SFPQ and this region of HFM1.

5.3.2.2 PDE3A is bound by both NONO and SFPQ

PDE3A is another mRNA that has several clusters for each protein in PAR-CLIP_Mild and a smaller number of clusters in PAR-CLIP_Medium (Figure 5.5 A). PDE3A mRNA encodes a phosphodiesterase that has higher expression in a number of cancer types, including testis, carcinoid, ovarian and endometrial (www.proteinatlas.org and [216]). The clusters are located close together towards the 3' end of the first PDE3A intron (Figure 5.5A). The section of PDE3A these

clusters map to contains a rRNA repeat element (Figure 5.5 B), that could suggest the reads that make up these clusters were misaligned here. However, given there are clusters from the Medium experiment that map at this location, and the transition event count is high for these clusters (as opposed to other mis-mapped rRNA repeat clusters), it is likely these clusters are genuine DBHS crosslinked RNAs. In addition, the multiple clusters upstream of these shared clusters, in the 5' end of the first intron, suggest that this transcript was indeed bound by NONO and SFPQ (Figure 5.5 C). Individually these clusters have a low transition event count, however, there are multiple clusters identified, in particular for NONO and SFPQ in PAR-CLIP_Mild, suggesting that the proteins bind weakly but at multiple locations along the transcript.

5.3.2.3 GPI is bound by both NONO and SFPQ

GPI mRNA encodes a glucose phosphate isomerase protein that has a variety of mechanistically distinct functions. The clusters for NONO and SFPQ lie in a middle intron of GPI (Figure 5.6 A). None of these clusters correspond to repetitive elements. The transition event count in the clusters is much higher for the Medium PAR-CLIP experiments, due to a higher number of reads making up these clusters. Figure 5.6 B shows a close up of the clusters along the transcript. There is one region that is spanned by a cluster for both proteins in both experiments (Figure 5.6 B, arrow). This is indicative of a high confidence NONO and SFPQ binding site in GPI.

5.3.2.4 DAZAP1 appears to be bound by both NONO and SFPQ.

Interestingly, DAZAP1 is an RNA binding protein that has been reported to localize to paraspeckles, indeed, DAZAP1 is essential for paraspeckle formation as siRNA knockdown of DAZAP1 results in loss of paraspeckles [81]. It was therefore intriguing that DAZAP1 mRNA was found crosslinked to NONO and SFPQ. Figure 5.7 shows the NONO and SFPQ clusters from each PAR-CLIP experiment that map to DAZAP1. Importantly, none of these clusters align within repetitive elements, and most are found in multiple experiments. The NONO and SFPQ binding sites are at the same location, with the SFPQ cluster slightly longer than the NONO cluster, 1 nucleotide longer at the 5' end and two longer at the 3' end (not shown). It is

interesting to see that the mRNA of a paraspeckle protein may be regulated by other paraspeckle proteins, and thus may represent a regulatory loop related to paraspeckle formation.

5.3.2.4.1 NONO knockdown lowers the levels of DAZAP1 protein

To investigate the effect of NONO and SFPQ on DAZAP1, the DBHS proteins were knocked down with siRNA and the levels of DAZAP1 assessed by western blot. As previously seen, the optimized knockdown conditions resulted in efficient SFPQ knockdown (Figure 5.8, compare lane 1 with lanes 3 and 4) and, in the case of the DAZAP1 blot, nearly total NONO Knockdown was achieved (Figure 5.8, compare lane 1 with lane 2 and 3). In other experiments the DAZAP1 antibody was demonstrated to be clean and specific, only detecting one band of the correct molecular weight in HeLa western blots (data not shown). Interestingly, NONO knockdown alone and in combination with SFPQ knockdown appeared to result in a subtle decrease in DAZAP1 protein levels (Figure 5.8, compare DAZAP1 in lane 1 with the levels in lane 2 and 4). From this blot the effect SFPQ knockdown alone has on DAZAP1 is not clear, however it appears the levels do not change relative to the scramble control (Figure 5.8, compare lane 1 with lane 3). The decrease in DAZAP1 expression following NONO knockdown suggests NONO may be having an on DAZAP1 expression, although mechanisms for this activity were not investigated further in this project.

5.4 ADARB2 and TP53INP1 are two mRNAs with low expression that were identified as NONO and/or SFPQ bound transcripts.

Abundant transcripts can contaminate CLIP experiments. With PAR-CLIP, T-to-C transitions help to eliminate these non-specific transcripts. However, if certain RNAs are very abundant, nuclear RNA binding proteins may come into contact with them and crosslinks may form due to spatial proximity. With this in mind, it is beneficial to determine the expression levels of RNA targets identified in the cell type being studied, so that this may be taken into consideration when determining whether reported binding sites are indeed true. The PAR-CLIP experiments described here were carried out in HeLa cells, for which there are publicly available RNA-seq tracks, however HeLa cells vary markedly from lab to lab,

therefore it was deemed desirable to carry out RNA-seq on these specific HeLa cells. Unfortunately, technical issues prevented total RNA sequencing of the HeLa cells used in PAR-CLIP experiments reported here from being included in this thesis.

Nevertheless, to illustrate the specificity of binding detected in PAR-CLIP, the presence of clusters in highly expressed common PAR-CLIP contaminants has been assessed in Chapter 4. Here, I examine the ability of this PAR-CLIP to detect RNA binding sites in transcripts that are expressed at very low levels.

ADARB2 (Adenosine Deaminase RNA specific B2) and TP53INP1 (Tumour protein 53 induced nuclear protein 1) are two transcripts detected in the DBHS PAR-CLIP libraries that are expressed at very low levels. ADARB2 was first identified in the DBHS context as a transcriptional target of SFPQ, which binds to the ADARB2 promoter and activates transcription of the gene [88]. Previous experiments in the lab have shown that both RNA and protein levels of ADARB2 are very low in HeLa cells. TP53INP1 mRNA codes for a protein that triggers cell death, and is therefore expressed at very low levels in live cells. The DBHS PAR-CLIP experiments showed binding to these two transcripts, illustrating the ability of this experiment to not only prevent detection of non-specifically bound abundant transcripts, but also to report binding in transcripts with very low expression.

5.4.1 ADARB2 is possibly regulated by SFPQ at both the transcriptional and post transcriptional level

ADARB2 mRNA encodes a double stranded RNA binding protein with structural similarity to the other ADAR proteins that catalyze the deamination of adenosine to inosine in RNA. Interestingly, ADARB2, lacks key residues enabling catalytic activity. Instead, ADARB2 is thought to act as a regulator of ADARs, by competing for binding to substrates [217]. There is an unusual confluence of function in this target, given that NONO and SFPQ have been previously identified as binding to A-I edited dsRNA transcripts, the end-product of the ADAR-catalyzed reaction.

Despite the low ADARB2 expression, clusters were reported in ADARB2 for both NONO and SFPQ in the PAR-CLIP_Mild experiment. These clusters are located in the first intron of ADARB2 and, despite low transition event counts, the presence of multiple adjacent clusters indicates specific binding sites (Figure 5.9 A). Given that SFPQ has previously been shown to bind the ADARB2 promoter by Chromatin-IP [88] it is highly relevant that this study suggests SFPQ is also binding nascent ADARB2 RNA as well as the ADARB2 promoter.

To assess the effect of NONO and SFPQ binding, ADARB2 mRNA levels were analyzed by qPCR following NONO and/or SFPQ siRNA knockdown. As expected, SFPQ knockdown results in a dramatic decrease in ADARB2 mRNA levels compared to the scramble control (Figure 5.9 B). This result mirrors the published findings that ADARB2 levels depend on the transcriptional activation properties of SFPQ [88]. Unfortunately, the function SFPQ plays as a DNA-binding transcription factor cannot be distinguished from any other activity as an RNA binding protein in this assay. NONO knockdown alone does not result in a significant difference in ADARB2 expression compared with the scramble control (Figure 5.9 B). However, given there is only one low transition even count cluster reported for NONO in ADARB2, it is likely NONO binding was weaker than SFPQ binding, and may not have had a dramatic effect.

5.4.2 TP53INP1 transcript levels are repressed by NONO and SFPQ

TP53INP1 is induced by p53 upon cellular stress. It acts in the nucleus to induce apoptosis [218] and decrease metastasis [219]. TP53INP1 is reported to be down-regulated in several cancers [220], suggesting its anti-proliferative and pro-apoptotic function is lost in these cells resulting in increased tumorigenesis.

PARAlyzer analysis of NONO and SFPQ PAR-CLIP libraries did not output any clusters in TP53INP1. However, careful assessment of the reads mapped to the genome revealed a large number of the sequenced reads from NONO and SFPQ in PAR-CLIP_Mild that were mapped to the TP53INP1 transcript (Figure 5.10 A, red box). A close look at the aligned reads showed a number contained a T-to-C transition at a single position, indicating these fragments were cross-linked to NONO and SFPQ (Figure 5.10 A, note that a single read on this figure represents

multiple collapsed reads). It is also interesting that these reads show an A-to-G conversion, suggestive of either Adenosine to Inosine editing or a SNP.

This region was reported as a group for both NONO and SFPQ datasets by PARalyzer. However a cluster was not reported because T-C transitions only occurred at one location, which is below the PARalyzer threshold of 5 conversion locations. As discussed earlier, this requirement of at least 5 T-C conversions was used because of the extensive NEAT1 crosslinking. The read count for each nucleotide in the group, as well as the percentage of nucleotides that have to T-to-C transition can be graphed, with the group for NONO and SFPQ (Figure 5.10 B and C) spanning the same region. For both NONO and SFPQ, the read count across the nucleotides in the group remains consistent, indicating the proteins were likely binding and protecting this whole region from degradation by RNase. However, despite multiple T nucleotides in this group, only one has a transition, suggesting that either only this nucleotide was in direct contact with the proteins, or only this nucleotide was substituted for 4-SU.

The nucleotide position with close to 100% T-C conversion does not correspond to any known SNPs (Figure 5.10 D), indicating the transition is likely to be genuinely due to crosslinking to the proteins. Another possibility is that it is a spontaneous mutation in this HeLa cell line. Without total RNA sequencing results from the HeLa cells used in this PAR-CLIP it is not possible to rule this out, however, the consistently high read count for this group suggests a large number of fragments were isolated bound by NONO and SFPQ. The groups do not correspond to a repetitive element (Figure 5.10 D) indicating the mapped reads originated from TP53INP1 mRNA. Taken together, the genuine transition event and the high read count for the group mapping to TP53INP1 suggests this transcript was bound by the NONO/SFPQ heterodimer.

5.4.2.1 TP53INP1 total cellular RNA levels increase with NONO and SFPQ knockdown

To further confirm NONO and SFPQ binding to TP53INP1 mRNA as well as characterizing the effect of this binding, RT-qPCR was performed to measure TP53INP1 mRNA levels following NONO and SFPQ knockdown.

Interestingly, both NONO and SFPQ knockdown resulted in significant increases in TP53INP1 mRNA levels (Figure 5.10 E). NONO knockdown alone results in a two fold increase in TP53INP1 levels, while SFPQ knockdown and the double knockdown causes a 4-6 fold increase in TP53INP1 levels. While the SEM is large for NONO and SPFQ single knockdown and the results should be viewed with caution, it does present a model whereby NONO and SFPQ bind TP53INP1 mRNA preventing its anti-cancer role in the cell.

5.4.2.2 Nuclear and cytoplasmic ratios of TP53INP1 RNA do not significantly change with NONO or SFPQ knockdown

I next wanted to assess the possibility that NONO/SFPQ was binding and retaining TP53INP1 RNA in the nucleus to inhibit its translation. I therefore carried out cellular fractionation and qPCR of nuclear and cytoplasmic RNA. If NONO and SFPQ were involved in the retention of TP53INP1, its levels would be expected to be higher in the cytoplasm following NONO and/or SFPQ siRNA knockdown compared with the control. The fractionation protocol was first carefully optimized to determine lysis buffer composition and centrifugation speeds that would give clean cytoplasmic fractions devoid of contaminating nuclear RNAs. NEAT1 RNA was used as a nuclear marker to ensure nuclei remained intact during the separation procedure. After considerable optimization, clean nuclear and cytoplasmic fractions were obtained from DBHS siRNA knockdowns in HeLa using the Paris Kit and optimized spin protocol (See Materials and Methods chapter 2.1.6 for fractionation protocol and chapter 2.1.8.2 for qPCR data analysis). NEAT1 levels were 50-100 fold higher in the nucleus compared with the cytoplasm for the scramble control and all of the knockdowns (Figure 5.11 A). Figure 5.11 A also shows that the cytoplasmic:nuclear ratio of NEAT1 is not statistically different between the different samples. The amount of TP53INP1 RNA in the nuclear and cytoplasmic fractions of the scramble and experimental samples was then assessed, to determine whether the knockdown of NONO, SFPQ or a double knockdown would increase the level of this RNA in the cytoplasm. While TP53INP1 levels are higher in the knockdowns compared to the scramble control (as was seen in the assessment of total RNA levels, figure 5.10 E) there is no significant difference in the cytoplasmic:nuclear ratio of TP53INP1 in any of the knockdown samples compared to the scramble control (Figure 5.11 B). While this data is preliminary, it does suggest that depletion of NONO or SFPQ does not result in

dramatic export of TP53INP1 to the cytoplasm. This presents the possibility that NONO and SPFQ are affecting the regulation of TP53INP1 in some other way, perhaps at the translation level. Alternatively, there may be other proteins involved in the nuclear retention of TP53INP1, for example, PSPC1 has been shown to increase following NONO depletion [171], and may be acting to retain this RNA. The large SEM, especially for the SFPQ knockdown sample, indicates the difficulty of getting accurate readings from a transcript with such low expression levels.

5.5 Discussion

This chapter has outlined the identification and characterization of NONO and SFPQ binding in mRNAs. The results of the PAR-CLIP experiments were analyzed to identify NONO and SFPQ binding sites in mRNA transcripts, with these findings supporting the notion that these proteins do much more than facilitate nuclear retention of RNA, instead forming part of a larger network of gene regulation, that likely involves regulating mRNA transcripts with long introns as they are made. Several biologically relevant mRNAs were isolated crosslinked to NONO and SFPQ, with some of these looked at in closer detail. The effect of NONO and SFPQ binding on some of these transcripts was validated and potential models proposed for the role NONO and SFPQ are playing by binding these RNAs.

5.5.1 Analysis of the NONO and SFPQ binding sites in mRNAs shows they bind in introns.

The absence of DBHS binding sites in exons strongly suggests that NONO and SFPQ are binding to pre-mRNAs. This binding within unspliced RNA was also observed for some lncRNAs that contain introns. Although NONO and SFPQ have been implicated in the transport of RNA to the cytoplasm [221] there is little evidence of this observed in this PAR-CLIP experiment. Instead, it appears, in HeLa cells at least, NONO and SFPQ are binding newly transcribed RNAs, prior to their processing.

Whilst it appears NONO and SFPQ bind pre-mRNAs, one cannot exclude the possibility that the RNA fragments derive from introns already spliced out of the transcript. Additional experiments are needed to definitively say whether NONO

and SFPQ are binding to pre-mRNAs. For example, FISH against specific exons and intronic sequences of target RNAs, coupled with staining with DBHS antibodies could be performed to reveal which parts of the RNA colocalise with the proteins in the nucleus. In addition, qPCR on the fragments isolated following protein IP with primers specific to introns and exons, and even intron/exon boundaries could shed more light on the processing the bound transcripts have undergone.

Both NONO and SFPQ have reported roles in alternate splicing [117,222], however, in this experiment, none of the clusters mapped to splice sites (Figure 5.1). One possible reason for the absence of clusters in intron/exons junctions was that the read mapping did not account for this. The Bowtie aligner does not consider gaps when mapping reads, as such, an RNA fragment that spanned a splice site would not have mapped with Bowtie and would have required another alignment program, such as TopHat [223]. Alternatively, DBHS proteins may be binding in other regions aside from intron/exon boundaries, such as in introns to facilitate acceptor or donor site selection. SFPQ was initially found as an accessory factor to PTBP1 (polyprimidine tract binding protein 1), a sequence feature located in introns 5' to the end of the region to be spliced [224,225].

5.5.1.1 None of the NONO or SFPQ binding sites in mRNAs detected with PAR-CLIP lie in IR-Alus in the 3'UTRs

Interestingly, paraspeckle targeting of RNAs via protein binding in their 3' UTRs is a characterized mechanism implicating paraspeckles in gene regulation. However, 3' UTRs are a feature notably devoid of clusters in the PAR-CLIP experiments performed here (Figure 5.1).

A seminal study by Zhang and Carmichael found that NONO and SFPQ bind A-to-I edited regions in dsRNA [90]. This led to a model in which NONO/SFPQ bound A-I edited dsRNA in mRNA 3'UTRs, targeting the RNA to paraspeckles [77]. The finding that only 2% of NONO binding sites and 3% of SFPQ binding sites in mRNAs map to 3' UTRs suggests that binding in these regions was not occurring to an appreciable extent in the experiment reported here (Figure 5.1). To investigate the accepted model of NONO binding in IR-Alus in 3' UTRs, the 3' UTR clusters in each

of the bound transcripts were assessed for the presence of IR-Alus. Surprisingly, none of the binding sites in 3' UTRs corresponded to IR-Alus for either NONO or SPFQ. In fact, none of these transcripts contained IR-Alus in their 3' UTRs.

One possibility is that this mechanism of nuclear retention by NONO binding to IR-Alus in 3' UTRs does not occur to an appreciable extent in HeLa cells. In several of the studies examining nuclear retention, HEK293T and NIH3T3 cells lines were used. Alternatively, the absence of NONO and SFPQ binding in IR-Alus may indicate that any role for these proteins in nuclear retention represents just a tiny fraction of the full RNA-binding capability of DBHS proteins.

It is essential here to keep in mind the features and limitations of the PAR-CLIP and bioinformatics analysis when evaluating the apparent absence of binding sites in 3' UTR IR-Alus. For example, while the PAR-CLIP protocol isolates only short RNA fragments, it is likely that the bowtie cut off for mismatches is not allowing reads to map to the genome if they contain a high number of A-to-I editing sites. The dsRNA editing enzyme ADAR is capable of editing ~50% of the As in a transcript, with such extensively edited transcripts unlikely to map to the genome. However, given PAR-CLIP isolates only short RNA fragments, the effect editing has on the mapping of these is not known.

5.5.1.2 NONO and SFPQ bind predominantly in long first introns and this is similar to other paraspeckle-associated proteins with a role in neurobiology

In this chapter I showed that NONO and SFPQ bind transcripts with long first introns. Binding in long first introns is an attribute shared by other paraspeckle proteins, particularly FUS and TDP-43, two proteins with important roles in neurodegeneration [208]. TDP-43 and FUS are involved in maintaining the levels of mRNAs that have long introns, with many of these mRNAs essential for proper neuron function [208,209]. A CLIP experiment to identify RNA binding sites for FUS in human and mouse brain revealed this protein predominantly binds in long (>100nt) first introns [208] at multiple binding sites. This was similar to what was observed with NONO and SFPQ clusters along mRNAs, with clusters preferentially located in first introns, and these introns being much longer than the transcript

median intron length. It is possible that NONO and SFPQ binding in long first introns is maintaining the levels of these transcripts in a similar way to TDP-43 and FUS. Possibly NONO and SFPQ bind the nascent transcript and oligomerize along the first intron, an action that would 'protect' the transcript as it is being made, preventing premature splicing and degradation. Alternatively they may be binding and directing the splicing machinery to act as soon as the second exon is transcribed, in this way facilitating co-transcriptional splicing. Another possibility is that NONO and SFPQ may be recruited to these long first introns so they can form additional interactions with RNA polymerase to promote transcript elongation. In this latter case it is interesting that there are several reports of NONO/SFPQ interacting with the CTD of RNA Pol II (See table 1.1).

Interestingly, TDP-43 or FUS knockdown results in the reduced abundance of a small number of mRNAs, with the majority of the transcripts affected by this knockdown contained long first introns being involved in neuron function [208,209]. In my PAR-CLIP experiments, Ingenuity pathway analysis reported that 13% of the NONO crosslinked transcripts were implicated in neurological disorders (Table 4.5). This indicates that DBHS proteins may have a role similar to that of FUS in regulating the levels of these transcripts, perhaps by binding in the intron to regulate it before splicing can commence.

FUS and TDP-43 also localize to paraspeckles [81]. iCLIP, performed in human brain from patients with FTLD showed increased binding of FUS and TDP-43 to NEAT1 and MALAT1 in disease [179]. Despite its binding to NEAT1, TDP-43 depletion does not disrupt paraspeckles to an appreciable extent, nor does it lower NEAT1 levels in the cell [81]. In contrast, FUS depletion by siRNA knockdown reduces the number of paraspeckles in the cell, but not the levels of NEAT1 [81]. This new finding that NONO and SFPQ act in a similar manner to FUS and TDP-43, and all are paraspeckle proteins, presents an interesting connection between paraspeckle protein localisation and the regulation of mRNAs necessary for neuronal function.

There are also potential new roles for SFPQ in the development of neurodegenerative disease. In animal models of neurodegenerative disease and brain samples from Alzheimer's and frontotemporal lobar degeneration (FTLD)

patients, SFPQ was observed re-localizing from the nucleus to the cytoplasm, where it forms cytoplasmic aggregates [124]. This cytoplasmic re-localization and aggregation is similar to the behavior of TDP-43 and FUS in amyotrophic lateral sclerosis (ALS) [125,126]. In addition, FUS inclusion bodies in ALS were recently shown to contain NONO and SFPQ, suggesting these proteins interact to form protein aggregates in FUS linked neurodegenerative disorders [126].

5.5.2 The identification of novel binding sites for NONO and SFPQ reveals new cellular pathways these proteins may be influencing

Several of the transcripts that were identified as bound by NONO and/or SFPQ have interesting biology in line with reported roles of NONO and SFPQ in the cell, namely in the development and progression of cancer and the cellular response to viral infection.

5.5.2.1 The effect of NONO and SFPQ binding on IGF1R

IGF1R is a novel DBHS protein target with proliferative properties. Given this binding was observed in a cancer cell line, it is likely IGF1R is up-regulated and contributing to the anti-apoptotic effect in these cells. Further experiments are required to determine what mechanism DBHS proteins are using binding the long second intron in the IGF1R transcript, and if there are any effects on IGF1R with DBHS knockdown. If it could be established whether DBHS protein binding to IGF1R mRNA was occurring in an attempt to prevent its expression, this interaction could be enhanced therapeutically. Conversely, if the interaction is stabilizing IGF1R mRNA and resulting in its up-regulation, disrupting this interaction may be beneficial. There are several drugs already available to either block the interaction of the IGF1R with its ligand [226] or to down regulate the receptor by RNAi [227]. Targeting the interaction between the DBHS proteins and IGF1R mRNA could present a novel treatment strategy.

5.5.2.2 MATR3 protein has been identified in a complex with NONO and SFPQ and now there is evidence the mRNA is also bound

MATR3 RNA is another novel DBHS target. The NONO/SFPQ and MATR3 proteins have been co-purified in a complex associated with several nuclear functions: double-stranded A-I edited RNA binding, viral RNA export and DNA repair [90,221,228]. It is fascinating that this is an example of NONO/SFPQ binding both a protein and its cognate RNA (DAZAP1 is another example, to be discussed below).

Despite being an abundant nuclear RNA-binding protein that interacts with NONO/SFPQ, there is no evidence that the MATR3 protein localizes to paraspeckles[81]. Nevertheless, MATR3 may have a role in paraspeckle accumulation and persistence in the cell, primarily as it is found in a complex with many essential paraspeckle proteins, not just NONO/SFPQ, but also HNRNPK [229].

5.5.2.3 DAZAP1 is a paraspeckle protein, now there is evidence the RNA is bound also

DAZAP1 is another essential paraspeckle protein [81]. The presence of two DBHS binding clusters in two different regions of the DAZAP1 transcript, one in the first intron and the other further down (Figure 5.7) suggests it is a genuine target, as these binding sites are distinct, and the clusters were found in multiple experiments. DAZAP1 was first identified in a screen for DAZ (deleted in azoosperima, a protein expressed almost exclusively in testis) binding partners [230]. Characterization of DAZAP1 in mouse cells, which is 98% similar to human DAZAP1, show it is expressed most abundantly in the testis [231]. Comparison of RNA-seq libraries from DAZAP1 knock down versus control cells showed DAZAP1 was contributing to alternate splicing events [232].

DAZAP1 stimulates translation by binding to specific mRNAs in their 3'UTR and enhancing their association with polysomes [233], in this way, DAZAP1 regulates the expression of certain genes at the post transcriptional level. It most likely achieves this by shuttling mRNAs from the nucleus to the cytoplasm, as its localization changes as the germ cells mature [234]. DAZAP1 is essential for spermatogenesis, and

while female knockout mice have normal ovaries and can fall pregnant, the pregnancies are not sustained due to failure in early embryonic development [233,235].

NONO knockdown and combined NONO/SFPQ knockdown triggers a subtle decrease in DAZAP1 protein levels (Figure 5.8). This suggests that NONO and SFPQ are regulating the expression of DAZAP1, potentially via binding to DAZAP1 mRNA as the PAR-CLIP has reported binding in this transcript. Further experiments should be conducted to establish if NONO and SFPQ affect the DAZAP1 mRNA levels in addition to the protein level. Should future experiments reveal that RNA levels are unaffected by knockdown, it may be possible that cytoplasmic:nuclear ratios are altered and this can also be assessed.

5.5.2.4 TP53INP1 mRNA encodes a pro-apoptotic protein that is repressed by NONO and SFPQ

TP53INP1 is a pro-apoptotic protein, down-regulated in transformed cells. P53 is expressed as part of the cell's response to stress and DNA damage, and triggers apoptosis. A p53 binding site was identified in the TP53INP1 gene promoter, and its induction via p53 with a number of stressors was demonstrated [236-238]. TP53INP1 binds to the kinases HIPK2 [239]and PKC δ [240] to trigger phosphorylation of p53, inducing apoptosis. A summary of the action of TP53INP1 in the cell is outlined in figure 5.12.

NONO and SFPQ binding in TP53INP1 mRNA, and the increase in mRNA levels following NONO and SFPQ knockdown suggests these proteins are regulating TP53INP1 at the post-transcriptional level. This activity aligns with NONO/SFPQ acting as oncogenes, ie suppressing a pro-apoptotic protein, in this context. To get clues as to the mechanism of regulation, a nuclear/cytoplasmic separation experiment was carried out from control and DBHS knockdown cells. These experiments showed that the nuclear/cytoplasmic ratios of TP53INP1 mRNA do not change significantly when the DBHS proteins levels are reduced, suggesting that nuclear retention of the RNA is not a primary mechanism. This is interesting given the presence of a potential A-I editing site in the pre-mRNA for TP53INP1.

However, one confounding factor in these experiments was the extremely low total abundance of the TP53INP1 transcript, meaning that there was considerable variation in the replicates. Ideally, these experiments should be repeated with another cell type that has higher TP53INP1 levels. Another difficulty was that the TP53INP1 protein is associated with apoptosis and was therefore difficult to detect by western blotting (data not shown).

If the NONO and SFPQ interaction with TP53INP1 could be targeted and disrupted, that may result in up-regulation of TP53INP1, as suggested by the increased mRNA levels (and predicted increase in protein following NONO and SFPQ knockdown). PAR-CLIP results indicate that TP53INP1 mRNA is bound at one particular position by the proteins, as the T-to-C transition diagnostic of crosslinking occurs here (Figure 4.11 B and C). Future experiments could involve using genome engineering to mutate this region, thus validating that this is the interaction site. It is anticipated that eventual *in vitro* NONO/SFPQ-RNA structural analysis may be used for generating inhibitors of these binding events as potential anti-cancer agents.

The role of NONO and SFPQ in binding mRNAs involved in a number of diseases, in particular cancer and viral infection, presents many novel therapeutic targets. If the molecular mechanisms the proteins employ to bind target mRNAs can be established with *in vitro* experiments, a mechanism of modulating these interactions can be formulated. In addition, the overall outcome of this binding, including the pathway it is a part of and the outcome it has for the cell must be assessed. Once known, these interactions can be targeted and stabilized or destabilized, resulting in better outcomes for patients.

6. General Discussion

NONO and SFPQ are key RNA binding proteins involved in a variety of cellular processes, including regulation of gene expression and the formation of paraspeckles. Whilst much is known about the protein interaction partners of the DBHS proteins, very little is known about the RNAs they bind and potentially regulate.

This thesis is the first large scale, transcriptome wide, identification and analysis of the RNAs bound by NONO and SFPQ, conducted in two different cell types. Identifying the features of the types of RNAs bound by these proteins has given new insights into the mechanisms NONO and SFPQ may be using to regulate gene expression. In addition, new RNA targets of important disease relevance, particularly for cancer and infectious diseases have been identified. The details of the molecular interaction between the proteins and their target RNAs will potentially allow these complexes to be utilized as therapeutic targets to attenuate their action in the cell.

6.1 PAR-CLIP and PARalyzer were optimized for NONO and SFPQ analysis

While a number of CLIP protocols have been published and widely used, PAR-CLIP was the method used here for the identification of RNA molecules bound by NONO and SFPQ. The sample generation protocol, as well as the bioinformatic analysis, was optimized to identify the transcriptome-wide targets of NONO and SFPQ and reveal binding sites in the RNA to single nucleotide resolution.

6.1.1 PAR-CLIP preserves *in vivo* NONO and SFPQ interactions with their target RNAs.

CLIP protocols, as opposed to REMSA or SELEX , capture RNA:Protein interactions in living cells. We wanted to capture NONO and SFPQ interactions *in vivo* as NEAT1 is long and complex and the paraspeckle cannot be reconstituted *in vitro*, and

secondly, as DBHS oligomerization (difficult to establish in a controlled way in vitro) appeared to be playing an important role in RNA binding.

6.1.2 PAR-CLIP was performed on endogenous NONO, allowing the identification of native RNA:Protein interactions

Many other PAR-CLIP experiments have relied on over expression of a tagged protein for the IP [154,155,173], however the availability of large quantities of a very specific NONO antibody meant the PAR-CLIP reported here isolated endogenous NONO and SFPQ in complex with their crosslinked RNAs. This was beneficial as the homo- and heterodimers the DBHS protein form are present in their endogenous ratios, with these dimers and the arrangement of the RRMs within likely conferring target specificity. In an overexpression system, excess NONO would likely form additional homodimers, indeed it is well established in the Fox lab that DBHS protein over-expression results in large nuclear aggregates, which would likely distort the results. A disadvantage of using endogenous protein is that a large quantity of cells was required to isolate enough RNA for library preparation and sequencing. This experimental scale means knocking down DBHS proteins or NEAT1 followed by PAR-CLIP would be an expensive experiment. Rather than performing siRNA knockdown on ~20 x 10^8 cells (as were used here), it would be beneficial to first generate an inducible stable cell line to over-express biologically meaningful amounts of DBHS proteins, determine how this changes the reported PAR-CLIP RNAs and take this into consideration in future experiments.

6.1.3 Using 6-SG in addition to 4-SU will overcome the reliance of binding to uridine to facilitate crosslinking

One weakness of PAR-CLIP is the fact that crosslinking is dependent on the incorporation of 4-SU. Thus, crosslinking will not occur where 4-SU is not incorporated, such as in regions of the transcript devoid of U's, or in low-abundance transcripts with limited 4-SU incorporation. This sequence bias is one possible explanation as to why many of the evolutionary-conserved predicted secondary structures in NEAT1 do not correspond to high transition event count

clusters (Chapter 4.4.3). Alternatively, it is possible that the DBHS proteins are not directly binding to these structured regions, and thus, there would be no crosslinks at these sites. These structures may instead form to bring other RNA binding sites into closer contact and it is these sites that contain the high transition event counts. To address this, a combination of 4-SU and 6-SG could be incorporated into transcripts, with crosslinks made at both U and G nucleotides.

6.1.4 PARalyzer clusters are a better indicator of NONO and SFPQ binding sites than the groups

The bioinformatics analysis centered around the program PARalyzer, which incorporated read depth and T-to-C transitions in mapped reads to determine protein binding sites in transcripts. In addition, the PARalyzer parameters were optimized to report only high confidence crosslinked transcripts and minimize detection of common PAR-CLIP contaminants.

In contrast to other PAR-CLIP studies reported [161], I have used the PARalyzer clusters as opposed to the groups as an indicator of protein interaction sites. This was essential for NEAT1, where groups spanned many hundreds of bases and therefore were not useful for identifying binding sites. However, it may perhaps be preferable to assess the groups when looking transcriptome wide for binding sites. This issue of groups versus clusters was illustrated by the case of TP53INP1 (Chapter 5.4.2). TP53INP1 had no reported clusters, but did contain groups called by PARalyzer. As this interaction was shown to be biologically relevant, it may be useful to re-analyse the data based on group, rather than cluster calling.

6.1.5 Bowtie alignment parameters could be altered to tolerate a greater number of mismatches while lowering the cut off for reporting of multi-mapped reads

A common problem with CLIP analysis is the reporting of peaks that lie in repetitive elements (such as rRNA repeats). This occurs as RNA fragments mis-align at these elements, increasing the read count, leading to the threshold for cluster reporting being reached [161]. Although PARalyzer attempts to minimize this problem by requiring a specified transition count is met, often these highly

abundant transcripts will interact non-specifically with the protein and will be crosslinked. To address this, one could lower the Bowtie parameter for number of alignments allowed. In this study up to 10 alignment co-ordinates could be reported for a read. This meant that multi-mapping sequences, such as those arising from repetitive elements, were reported. If this cut-off were lowered, it may mean that reads originating from repetitive elements would not be reported and therefore not used as input into PARalyzer. Determining the number of alignments allowed is a balance between ensuring short fragments, that may align to multiple locations, are reported, illustrating again the importance of optimizing the RNase digest conditions to give RNA fragments that are long enough to align uniquely to the genome, while still giving binding site specificity.

The Bowtie parameter specifying the number of mismatches tolerated in a sequence for an alignment to be reported is also a relevant consideration. PARalyzer recommends a total of 3 mismatches are allowed for an alignment to be reported [140], thus sequences do not align if they contain more than three crosslinks. In addition, and particularly relevant for the DBHS proteins, RNA editing will also introduce mismatches. Thus, if the proteins bind edited RNA and get crosslinked at several sites, leading to more than three mismatches then that alignment will not be reported. This may explain the lack of reported NONO and SFPQ binding sites in IR-Alus, sequences that are known to be hyper-edited (up to 50% A-to-I editing).

Although the small number of mapped reads in PAR-CLIP_Harsh was clearly due to an over-representation of adapter sequence due to the very short fragment length, it is possible that editing sites and T-to-C transitions inhibited the alignment of reads from the Mild and Medium PAR-CLIP experiments. Approximately 22% of NONO and SFPQ raw reads from PAR-CLIP_Mild aligned to the genome. In addition, the pool of longer RNA fragments isolated in complex with NONO and SFPQ would be expected to contain multiple crosslink sites. However, if more than three crosslinks occurred in any one fragment then it could not align. In light of this, it would be of interest to assess how many of the longer RNA sequence reads from PAR-CLIP did not align to the genome.

The mapping for PAR-CLIP_Medium was better, with approximately 50% of reads mapped to the genome. A possible reason for the difference in the proportion of mapped reads may be that in the mild experiment, longer fragments were isolated, with these possibly containing binding sites for other proteins that would have resulted in crosslinks and more transitions. In contrast, the harsher digest in PAR-CLIP_Medium digested away all but the region bound and protected by NONO and/or SFPQ, and, as such, crosslink sites induced by other proteins would not have been isolated with NONO IP.

A starting point for future Bowtie optimization would be to increase the number of mis-matches allowed but simultaneously lower the number of alignments allowed for a sequence to map to the genome. The effectiveness of this could be assessed by examining if IR-Alu containing elements in known NONO target RNAs, such as PAICS, NUP43 and PCCB were reported [73].

6.1.6 Background transcript expression levels in the cell line assayed should be determined by RNA-Seq, and reported bound transcripts assessed in light of this

Another limitation of PAR-CLIP is that 4-SU induced crosslinks will be under-represented in transcripts with low expression levels, or with long half-lives. This highlights the need for RNA-Seq data for the cell line being assayed. Measurements of individual transcript abundance could then be used to determine which clusters are reported due to high expression and 4-SU incorporation, versus those that have lower expression. To compensate for the absence of total RNA-Seq data, I required a high transition event count for cluster reporting, with the aim that this would minimize false positives from transcripts that are highly abundant (so have a high read count, but fewer transitions). This was effective, as many common PAR-CLIP contaminants were not reported, whilst some RNAs with low expression in HeLa cells (ADARB2 and TP53INP1) were reported (Chapter 5.4).

This study has made great progress in creating a robust PAR-CLIP protocol and bioinformatics analysis pipeline specifically optimized for the isolation and identification of NONO and SFPQ bound transcripts. In future, this PAR-CLIP protocol will be used to study the RNAs these proteins are binding in other cell

types, including different cancer cell models, to uncover additional cancer specific interactions and to reveal the overlap with the HeLa transcripts identified here. Future investigations could also include determining the impact of paraspeckle ablation on the RNAs bound by DBHS proteins. DBHS PAR-CLIP could be performed comparing immortalized MEF cells from wildtype and NEAT1 knockout mice, or genome-engineered NEAT1-/- human cell lines. These experiments will be useful to address NEAT1 dependent and independent RNA-binding roles for these proteins in a variety of cell types and tissues.

6.2 Identification of NONO and SFPQ interaction sites in NEAT1

The insights gained from this work have contributed to a model for paraspeckle formation (Figure 6.1). When NEAT1_v1 is initially transcribed, NONO and SFPQ bind along this transcript, however it is not until transcription extends into NEAT1_v2 that many strong and periodically spaced binding events occur. This repetitive binding is potentially strengthened by oligomerization of the proteins, thus the protein oligomers would bring distal segments of the transcript into closer proximity. The formation of DBHS protein-RNA structures may then be involved in higher-order folding of the long RNA. Future experiments could test the model of paraspeckle formation by preventing the periodic DBHS-NEAT1 interactions identified here. The CRISPR (Clustered regularly interspaced short palindromic repeats) system could be utilized to make mutations in, or delete the putative binding sites in NEAT1_v2, followed by assessment of the effect this has on paraspeckle formation.

It is also important to consider the role of additional paraspeckle proteins in this model, as 23 of the 40 known paraspeckle proteins (excluding NONO and SFPQ) contain one or more RRM, with another two proteins containing KH domains [81]. Thus, any or all of these proteins may also be binding NEAT1. Indeed, FUS and TDP-43 binding to NEAT1 has already been reported, albeit without the same nucleotide resolution that was achieved in the PAR-CLIP experiments described in this thesis [179].

Whilst we know that NEAT1 localises to paraspeckles, it is indeed, the paraspeckle backbone, we do not know the localization of the other RNA molecules identified as DBHS targets in this study. Whilst most are presumed to be nuclear, especially the pre-mRNA, there is also evidence of some of the RNAs in the cytoplasm. Within the nucleus it would be of interest to know if any of the DBHS target RNAs are also found within paraspeckles, as this could give clues to novel regulatory mechanisms. FISH against these RNA targets, including with intron-specific probes, should be carried out in future experiments to address this.

6.3 NONO and SFPQ interact with a variety of lncRNAs

Besides NEAT1, a number of other lncRNAs were found to contain NONO and SFPQ binding sites in PAR-CLIP. In particular, this study presented new evidence that NONO and SFPQ are binding MALAT1, a lncRNA highly expressed and significant in cancer (Chapter 4.6). Although MALAT1 does not localize to paraspeckles, it is possible that it functions as part of a complex with NONO and SFPQ outside of paraspeckles. Given new evidence for overlapping roles for NEAT1_v1 and MALAT1 in targeting transcriptionally active genes [201] and MALAT1 competition for SFPQ binding [59], further experiments investigating the mechanism of binding of DBHS proteins to MALAT1 would be interesting.

The two MYC-proximal lncRNAs, CCAT1 and PVT1 are also targeted by NONO and SFPQ. Whilst DBHS knockdown did not affect CCAT1 levels, it would be useful in the future to assess the levels of C-MYC, to see if there is an indirect effect. In this context, it is important to note that NONO knockdown affected levels of N-MYC in neuroblastoma, where NONO binds the N-MYC proximal lncRNA lncUSMycN [36]. The possibility that NONO and SFPQ act by binding RNAs derived from nearby genes and thereby bridging the gap between distal gene elements is illustrated in figure 6.2 B. Further investigation of this mechanism could be done through the CHART (capture hybridization analysis of RNA targets) or ChIRP (chromatin isolation by RNA purification) techniques to see whether any of the reported NONO and SFPQ bound lncRNAs interact with chromatin, followed by assessment of those downstream transcript levels following NONO and SFPQ knockdown.

NONO and SFPQ siRNA knockdown did not alter the total cellular levels of CCAT1 and LINC00473, suggesting that the proteins are not affecting RNA stability. Instead, the lncRNAs may act as decoys to compete for NONO and SFPQ binding to other lncRNAs such as NEAT1. Alternatively, these lncRNAs may be targeting NONO and SFPQ to specific regions in the cell, such as specific chromatin sites (Figure 6.2 A). All these possible mechanisms of action are the focus of future experiments.

6.4 NONO and SFPQ bind a number of disease associated RNAs

Many of the RNAs bound by NONO and/or SFPQ are involved in cancer and infectious disease pathways (Chapter 4.2.1). This implicates these proteins in the cell's response to these diseases, thereby the interactions between the proteins and the bound RNAs represent novel therapeutic targets. However, in order to therapeutically modulate these interaction in a meaningful way, more needs to be know about the role the proteins are playing in binding to these RNAs.

6.4.1 NONO and SFPQ bind in long first introns, possibly to stabilize or direct processing of these pre-mRNA

NONO and SFPQ bind a number of coding RNAs in the HeLa cell PAR-CLIP experiments reported here. The majority of NONO and SFPQ binding sites were located within introns, indicating these proteins likely bind and modulate pre-mRNAs, or introns once spliced out (Chapter 5.1). NONO and SFPQ appear to show a preference for binding in long first introns, a property that is also demonstrated by two other paraspeckle and neurodegenerative disorder proteins, TDP-43 and FUS, that bind and stabilize these transcripts [208]. The effect of NONO and SFPQ binding in long first introns is yet to be determined; they may stabilize these transcripts co-transcriptionally (Figure 6.2 D) or recruit and direct splicing factors (Figure 6.2 E). Future experiments investigating the role of NONO and SFPQ in regulating mRNAs with long first introns would be interesting, however it may be more beneficial to perform these experiments in a neuronal cell line, due to the connection with neurobiology for other proteins that bind long first introns.

6.4.2 The DBHS proteins appear to alter the expression of two key disease related molecules, DAZAP1 and TP53INP1

DAZAP1 is a key paraspeckle protein, as its depletion reduces the number of paraspeckles in the cell. In addition, DAZAP1 acts in the cell to bind RNAs and deliver them to polysomes, a process that regulates the expression of their transcripts. The PAR-CLIP data suggested that NONO and SFPQ targeted the DAZAP1 RNA, and knockdown of DBHS proteins resulted in a subtle decrease in the levels of DAZAP1 protein in HeLa cells. Although these results are preliminary, they do suggest that NONO and SFPQ may regulate DAZAP1 expression, which would have knock on effects for the DAZAP1 target transcripts. Future work should involve assessing DAZAP1 mRNA levels with DBHS knockdown, in order to determine where NONO and SFPQ act in the life of this transcript. Possibly they prevent its transcription; alternatively they could trigger its degradation or retain it in the nucleus to prevent its translation.

A cancer-related target of NONO and SFPQ identified for the first time in this study is TP53INP1. As outlined in chapter 5.5.2.4, TP53INP1 is induced by P53 upon DNA damage and triggers apoptosis, preventing metastasis. The identification of a NONO and SFPQ binding site in TP53INP1 mRNA suggest they are potential regulators of this transcript. The increased expression of TP53INP1 mRNA following NONO and SFPQ depletion indicates they are performing an anti-apoptotic role by repressing the expression of this gene. Although NONO and SFPQ knockdown increased the overall cellular levels of TP53INP1, nuclear/cytoplasmic fractionation showed no change in the localization of this mRNA (Figure 5.11).

It is possible that TP53INP1 transcription is repressed by NONO and/or SFPQ, reminiscent of the repression mediated by SFPQ for the IL-8 gene [87]. When the proteins are depleted, the repression is released. This could be addressed with CHIP experiments to determine if SFPQ/NONO were acting at the TP53INP1 promoter. Alternately, DBHS proteins may facilitate TP53INP1 degradation or translation efficiency. Experiments addressing the effect of NONO and SFPQ on TP53INP1 mRNA levels should be performed in a broad range of cancer cell types to determine if the repression of TP53INP1 by NONO and SFPQ is a universal

mechanism in the inhibition of apoptosis and the promotion of metastasis. If this is the case, this would add weight to future development of DBHS proteins as therapeutic targets in cancer.

6.5 Summary

This study has made substantial steps in identifying the Protein:RNA interactions made by two important RNA binding proteins in the cell. Through careful and extensive optimizations, PAR-CLIP was tailored for the identification of NONO and SFPQ RNA targets in a human cancer and a mouse fibroblast cell line. This optimized protocol can, in future, be used to identify the interaction partners of these proteins in other cells and tissues, in the progression of diseases such as cancer and viral infection and in the absence of paraspeckles, a model lncRNA:Protein complex implicated in the cellular response to stress. This work has made extensive progress in identifying and characterizing the interactions between NONO and SFPQ and the lncRNA NEAT1 that facilitate the formation and persistence of paraspeckles. In addition, characterization of other NONO and SFPQ lncRNA and mRNA targets has implicated these RNA binding proteins in new cellular networks where they bind RNAs in different ways to achieve different outcomes. Future work will delve deeper into how the proteins affect different transcripts, either through directing their retention, stabilizing them as they are made or directing pre-mRNA processing. Conversely, it is highly likely that a number of the associated lncRNAs may regulate the actions of NONO and SFPQ, either through recruiting them to sites in the chromatin, pulling them off promoters or competing for binding to other proteins or lncRNA.

References

(1) Rosenberg, N. A., Pritchard, J. K., Weber, J. L., Cann, H. M., Kidd, K. K., Zhivotovsky, L. A., and Feldman, M. W. (2002) Genetic structure of human populations. *Science 298*, 2381–2385.
(2) Chimpanzee Sequencing and Analysis Consortium. (2005) Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature 437*, 69–87.
(3) Mouse Genome Sequencing Consortium, Lindblad-Toh, K., Birney, E., Rogers, J., Abril, J. F., Agarwal, P., Agarwala, R., Ainscough, R., Alexandersson, M., An, P., Antonarakis, S. E., Attwood, J., Baertsch, R., Bailey, J., Barlow, K., Beck, S., Berry, E., Birren, B., Bloom, T., Bork, P., Botcherby, M., Bray, N., Brent, M. R., Brown, D. G., Brown, S. D., Bult, C., Burton, J., Butler, J., Campbell, R. D., Carninci, P., Cawley, S., Chiaromonte, F., Chinwalla, A. T., Church, D. M., Clamp, M., Clee, C., Collins, F. S., Cook, L. L., Copley, R. R., Coulson, A., Couronne, O., Cuff, J., Curwen, V., Cutts, T., Daly, M., David, R., Davies, J., Delehaunty, K. D., Deri, J., Dermitzakis, E. T., Dewey, C., Dickens, N. J., Diekhans, M., Dodge, S., Dubchak, I., Dunn, D. M., Eddy, S. R., Elnitski, L., Emes, R. D., Eswara, P., Eyras, E., Felsenfeld, A., Fewell, G. A., Flicek, P., Foley, K., Frankel, W. N., Fulton, L. A., Fulton, R. S., Furey, T. S., Gage, D., Gibbs, R. A., Glusman, G., Gnerre, S., Goldman, N., Goodstadt, L., Grafham, D., Graves, T. A., Green, E. D., Gregory, S., Guigo, R., Guyer, M., Hardison, R. C., Haussler, D., Hayashizaki, Y., Hillier, L. W., Hinrichs, A., Hlavina, W., Holzer, T., Hsu, F., Hua, A., Hubbard, T., Hunt, A., Jackson, I., Jaffe, D. B., Johnson, L. S., Jones, M., Jones, T. A., Joy, A., Kamal, M., Karlsson, E. K., Karolchik, D., Kasprzyk, A., Kawai, J., Keibler, E., Kells, C., Kent, W. J., Kirby, A., Kolbe, D. L., Korf, I., Kucherlapati, R. S., Kulbokas, E. J., Kulp, D., Landers, T., Leger, J. P., Leonard, S., Letunic, I., Levine, R., Li, J., Li, M., Lloyd, C., Lucas, S., Ma, B., Maglott, D. R., Mardis, E. R., Matthews, L., Mauceli, E., Mayer, J. H., McCarthy, M., McCombie, W. R., McLaren, S., McLay, K., McPherson, J. D., Meldrim, J., Meredith, B., Mesirov, J. P., Miller, W., Miner, T. L., Mongin, E., Montgomery, K. T., Morgan, M., Mott, R., Mullikin, J. C., Muzny, D. M., Nash, W. E., Nelson, J. O., Nhan, M. N., Nicol, R., Ning, Z., Nusbaum, C., O'Connor, M. J., Okazaki, Y., Oliver, K., Overton-Larty, E., Pachter, L., Parra, G., Pepin, K. H., Peterson, J., Pevzner, P., Plumb, R., Pohl, C. S., Poliakov, A., Ponce, T. C., Ponting, C. P., Potter, S., Quail, M., Reymond, A., Roe, B. A., Roskin, K. M., Rubin, E. M., Rust, A. G., Santos, R., Sapojnikov, V., Schultz, B., Schultz, J., Schwartz, M. S., Schwartz, S., Scott, C., Seaman, S., Searle, S., Sharpe, T., Sheridan, A., Shownkeen, R., Sims, S., Singer, J. B., Slater, G., Smit, A., Smith, D. R., Spencer, B., Stabenau, A., Stange-Thomann, N., Sugnet, C., Suyama, M., Tesler, G., Thompson, J., Torrents, D., Trevaskis, E., Tromp, J., Ucla, C., Ureta-Vidal, A., Vinson, J. P., Niederhausern, Von, A. C., Wade, C. M., Wall, M., Weber, R. J., Weiss, R. B., Wendl, M. C., West, A. P., Wetterstrand, K., Wheeler, R., Whelan, S., Wierzbowski, J., Willey, D., Williams, S., Wilson, R. K., Winter, E., Worley, K. C., Wyman, D., Yang, S., Yang, S.-P., Zody, M. C., and Lander, E. S. (2002) Initial sequencing and comparative analysis of the mouse genome. *Nature 420*, 520–562.
(4) Gerstberger, S., Hafner, M., and Tuschl, T. (2014) A census of human RNA-binding proteins. *Nat. rev. Genet. 15*, 829–845.
(5) Lunde, B. M., Moore, C., and Varani, G. (2007) RNA-binding proteins: modular design for efficient function. *Nat. Rev. Mol. Cell Biol. 8*, 479–490.
(6) Änkö, M.-L., and Neugebauer, K. M. (2012) RNA–protein interactions in vivo: global gets specific. *Trends Biochem. Sci. 37*, 255–262.
(7) Ray, D., Kazan, H., Cook, K. B., Weirauch, M. T., Najafabadi, H. S., Li, X.,

Gueroussov, S., Albu, M., Zheng, H., Yang, A., Na, H., Irimia, M., Matzat, L. H., Dale, R. K., Smith, S. A., Yarosh, C. A., Kelly, S. M., Nabet, B., Mecenas, D., Li, W., Laishram, R. S., Qiao, M., Lipshitz, H. D., Piano, F., Corbett, A. H., Carstens, R. P., Frey, B. J., Anderson, R. A., Lynch, K. W., Penalva, L. O. F., Lei, E. P., Fraser, A. G., Blencowe, B. J., Morris, Q. D., and Hughes, T. R. (2013) A compendium of RNA-binding motifs for decoding gene regulation. *Nature 499*, 172–177.
(8) Castello, A., Fischer, B., Eichelbaum, K., Horos, R., Beckmann, B. M., Strein, C., Davey, N. E., Humphreys, D. T., Preiss, T., Steinmetz, L. M., Krijgsveld, J., and Hentze, M. W. (2012) Insights into RNA Biology from an Atlas of Mammalian mRNA-Binding Proteins. *Cell 149*, 1393–1406.
(9) Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., and FitzHugh, W. (2001) Initial sequencing and analysis of the human genome. *Nature 409*, 860–921.
(10) Venter, J. C. (2001) The Sequence of the Human Genome. *Science 291*, 1304–1351.
(11) Zhuo, D., Zhao, W. D., Wright, F. A., Yang, H.-Y., Wang, J.-P., Sears, R., Baer, T., Kwon, D.-H., Gordon, D., Gibbs, S., Dai, D., Yang, Q., Spitzner, J., Krahe, R., Stredney, D., Stutz, Al, and Yuan, B. (2001) Assembly, Annotation, and Integration of UNIGENE Clusters into the Human Genome Draft. *Genome Research 11*, 904–918.
(12) Wright, F. A., Lemon, W. J., Zhao, W. D., Sears, R., Zhuo, D., Wang, J. P., Yang, H. Y., Baer, T., Stredney, D., Spitzner, J., Stutz, A., Krahe, R., and Yuan, B. (2001) A draft annotation and overview of the human genome. *Genome Biol. 2*, 1–18.
(13) Crick, F. (1970) Central dogma of molecular biology. *Nature 227*.
(14) Orgel, L. E., and Crick, F. H. (1980) Selfish DNA: the ultimate parasite. *Nature 284*, 604–607.
(15) Bertone, P., Stolc, V., Royce, T. E., Rozowsky, J. S., Urban, A. E., Zhu, X., Rinn, J. L., Tongprasit, W., Samanta, M., Weissman, S., Gerstein, M., and Snyder, M. (2004) Global Identification of Human Transcribed Sequences with Genome Tiling Arrays. *Science 306*, 2242–2246.
(16) Cheng, J., Kapranov, P., Drenkow, J., Dike, S., Brubaker, S., Patel, S., Long, J., Stern, D., Tammana, H., Helt, G., Sementchenko, V., Piccolboni, A., Bekiranov, S., Bailey, D. K., Ganesh, M., Ghosh, S., Bell, I., Gerhard, D. S., and Gingeras, T. R. (2005) Transcriptional maps of 10 human chromosomes at 5-nucleotide resolution. *Science 308*, 1149–1154.
(17) Kapranov, P., Crawley, S. E., Drenkow, J., Bekiranov, S., Strausberg, R. L., Fodor, S. P. A., and Gingeras, T. R. (2002) Large-Scale Transcriptional Activity in Chromosomes 21 and 22. *Science 296*, 916–919.
(18) ENCODE Project Consortium. (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature 489*, 57–74.
(19) Graur, D., Zheng, Y., Price, N., Azevedo, R. B. R., Zufall, R. A., and Elhaik, E. (2013) On the immortality of television sets: “function” in the human genome according to the evolution-free gospel of ENCODE. *Genome Biol. Evol. 5*, 570–590.
(20) Mattick, J. S. (2004) RNA regulation: a new genetics? *Nat. rev. Genet. 5*, 316–323.
(21) Cheetham, S. W., Gruhl, F., and Mattick, J. S. (2013) Long noncoding RNAs and the genetics of cancer. *Brit. J. Cancer* 2419–2425.
(22) Lee, R. C., Feinbaum, R. L., and Ambros, V. (1993) The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. *Cell 75*, 843–854.
(23) Rana, T. M. (2007) Illuminating the silence: understanding the structure and function of small RNAs. *Nat. Rev. Mol. Cell Biol. 8*, 23–36.

(24) van Kouwenhove, M., Kedde, M., and Agami, R. (2011) MicroRNA regulation by RNA-binding proteins and its implications for cancer. *Nat. Rev. Cancer 11*, 644–656.
(25) Rinn, J. L., and Chang, H. Y. (2012) Genome Regulation by Long Noncoding RNAs. *Annu. Rev. Biochem. 81*, 145–166.
(26) Wapinski, O., and Chang, H. Y. (2011) Long noncoding RNAs and human disease. *Trends Cell Biol. 21*, 354–361.
(27) Yoon, J.-H., Abdelmohsen, K., and Gorospe, M. (2013) Posttranscriptional Gene Regulation by Long Noncoding RNA. *J. Mol. Biol. 425*, 3723–3730.
(28) Mercer, T. R., Dinger, M. E., and Mattick, J. S. (2009) Long non-coding RNAs: insights into functions. *Nat. rev. Genet. 10*, 155–159.
(29) The FANTOM Consortium. (2005) The Transcriptional Landscape of the Mammalian Genome. *Science 309*, 1559–1563.
(30) Smith, M. A., Gesell, T., Stadler, P. F., and Mattick, J. S. (2013) Widespread purifying selection on RNA structure in mammals. *Nucleic Acids Res. 41*, 8220–8236.
(31) Clark, M. B., Johnston, R. L., Inostroza-Ponta, M., Fox, A. H., Fortini, E., Moscato, P., Dinger, M. E., and Mattick, J. S. (2012) Genome-wide analysis of long noncoding RNA stability. *Genome Research 22*, 885–898.
(32) Gutschner, T., and Diederichs, S. (2012) The hallmarks of cancer. A long non-coding RNA point of view. *RNA Biol. 9*, 703–719.
(33) Gutschner, T., Hämmerle, M., and Diederichs, S. (2013) MALAT1 - a paradigm for long noncoding RNA function in cancer. *J. Mol. Med. 91*, 791–801.
(34) Yang, F., Xue, X., Bi, J., Zheng, L., Zhi, K., Gu, Y., and Fang, G. (2012) Long noncoding RNA CCAT1, which could be activated by c-Myc, promotes the progression of gastric carcinoma. *J. Cancer Res. Clin. Oncol. 139*, 437–445.
(35) Zhang, A., Xu, M., and Mo, Y.-Y. (2014) Role of the lncRNA-p53 regulatory network in cancer. *J. Mol. Cell Biol. 6*, 181–191.
(36) Liu, P. Y., Erriquez, D., Marshall, G. M., Tee, A. E., Polly, P., Wong, M., Liu, B., Bell, J. L., Zhang, X. D., Milazzo, G., Cheung, B. B., Fox, A., Swarbrick, A., Hüttelmaier, S., Kavallaris, M., Perini, G., Mattick, J. S., Dinger, M. E., and Liu, T. (2014) Effects of a novel long noncoding RNA, lncUSMycN, on N-Myc expression and neuroblastoma progression. *J. Natl. Cancer Inst. 106*.
(37) Wang, K. C., and Chang, H. Y. (2011) Molecular mechanisms of long noncoding RNAs. *Mol. Cell 43*, 904–914.
(38) Paul, I. J., and Duerksen, J. D. (1975) Chromatin-associated RNA content of heterochromatin and euchromatin. *Mol. Cell Biochem. 9*, 9–16.
(39) Rodríguez-Campos, A., and Azorín, F. (2007) RNA Is an Integral Component of Chromatin that Contributes to Its Structural Organization. *PLoS ONE 2*, e1182.
(40) Mondal, T., Rasmussen, M., Pandey, G. K., Isaksson, A., and Kanduri, C. (2010) Characterization of the RNA content of chromatin. *Genome Research 20*, 899–907.
(41) Brown, C. J., Ballabio, A., Rupert, J. L., Lafreniere, R. G., Grompe, M., Tonlorenzi, R., and Willard, H. F. (1991) A gene from the region of the human X inactivation centre is expressed exclusively from the inactive X chromosome. *Nature 349*, 38–44.
(42) Penny, G. D., Kay, G. F., Sheardown, S. A., Rastan, S., and Brockdorff, N. (1996) Requirement for Xist in X chromosome inactivation. *Nature 379*, 131–137.
(43) Hung, T., Wang, Y., Lin, M. F., Koegel, A. K., Kotake, Y., Grant, G. D., Horlings, H. M., Shah, N., Umbricht, C., Wang, P., Wang, Y., Kong, B., Langerød, A., Børresen-Dale, A.-L., Kim, S. K., van de Vijver, M., Sukumar, S., Whitfield, M. L., Kellis, M., Xiong, Y., Wong, D. J., and Chang, H. Y. (2011) Extensive and coordinated transcription of

noncoding RNAs within cell-cycle promoters. *Nat. Genet. 43*, 621–629.
(44) Park, E., and Maquat, L. E. (2013) Staufen-mediated mRNA decay. *WIREs RNA 4*, 423–435.
(45) Gong, C., and Maquat, L. E. (2011) lncRNAs transactivate STAU1-mediated mRNA decay by duplexing with 3' UTRs via Alu elements. *Nature 470*, 284–288.
(46) Kretz, M., Siprashvili, Z., Chu, C., Webster, D. E., Zehnder, A., Qu, K., Lee, C. S., Flockhart, R. J., Groff, A. F., Chow, J., Johnston, D., Kim, G. E., Spitale, R. C., Flynn, R. A., Zheng, G. X. Y., Aiyer, S., Raj, A., Rinn, J. L., Chang, H. Y., and Khavari, P. A. (2014) Control of somatic tissue differentiation by the long non-coding RNA TINCR. *Nature 493*, 231–235.
(47) Clemson, C. M., Hutchinson, J. N., Sara, S. A., Ensminger, A. W., Fox, A. H., Chess, A., and Lawrence, J. B. (2009) An architectural role for a nuclear noncoding RNA: NEAT1 RNA is essential for the structure of paraspeckles. *Mol. Cell 33*, 717–726.
(48) Tsai, M. C., Manor, O., Wan, Y., Mosammaparast, N., Wang, J. K., Lan, F., Shi, Y., Segal, E., and Chang, H. Y. (2010) Long Noncoding RNA as Modular Scaffold of Histone Modification Complexes. *Science 329*, 689–693.
(49) Chu, C., Qu, K., Zhong, F. L., Artandi, S. E., and Chang, H. Y. (2011) Genomic Maps of Long Noncoding RNA Occupancy Reveal Principles of RNA-Chromatin Interactions. *Mol. Cell 44*, 667–678.
(50) Rinn, J. L., Kertesz, M., Wang, J. K., Squazzo, S. L., Xu, X., Brugmann, S. A., Goodnough, L. H., Helms, J. A., Farnham, P. J., Segal, E., and Chang, H. Y. (2007) Functional demarcation of active and silent chromatin domains in human HOX loci by noncoding RNAs. *Cell 129*, 1311–1323.
(51) Kogo, R., Shimamura, T., Mimori, K., Kawahara, K., Imoto, S., Sudo, T., Tanaka, F., Shibata, K., Suzuki, A., Komune, S., Miyano, S., and Mori, M. (2011) Long noncoding RNA HOTAIR regulates polycomb-dependent chromatin modification and is associated with poor prognosis in colorectal cancers. *Cancer Res. 71*, 6320–6326.
(52) Ding, X., Zhu, L., Ji, T., Zhang, X., Wang, F., Gan, S., Zhao, M., and Yang, H. (2014) Long intergenic non-coding RNAs (LincRNAs) identified by RNA-seq in breast cancer. *PLoS ONE 9*, e103270.
(53) Silva, J. M., Perez, D. S., Pritchett, J. R., Halling, M. L., and Bosserhoff, A. K. (2010) Identification of long stress-induced non-coding transcripts that have altered expression in cancer. *Genomics 95*, 355–362.
(54) Li, J., Chen, Z., Tian, L., Zhou, C., He, M. Y., Gao, Y., Wang, S., Zhou, F., Shi, S., Feng, X., Sun, N., Liu, Z., Skogerboe, G., Dong, J., Yao, R., Zhao, Y., Sun, J., Zhang, B., Yu, Y., Shi, X., Luo, M., Shao, K., Li, N., Qiu, B., Tan, F., Chen, R., and He, J. (2014) LncRNA profile study reveals a three-lncRNA signature associated with the survival of patients with oesophageal squamous cell carcinoma. *Gut 63*, 1700–1710.
(55) Liu, W.-T., Lu, X., Tang, G.-H., Ren, J.-J., Liao, W.-J., Ge, P.-L., and Huang, J.-F. (2014) LncRNAs expression signatures of hepatocellular carcinoma revealed by microarray. *World J. Gastroenterol. 20*, 6314–6321.
(56) Kmita, M. (2003) Organizing Axes in Time and Space; 25 Years of Colinear Tinkering. *Science 301*, 331–333.
(57) Gupta, R. A., Shah, N., Wang, K. C., Kim, J., Horlings, H. M., Wong, D. J., Tsai, M.-C., Hung, T., Argani, P., Rinn, J. L., Wang, Y., Brzoska, P., Kong, B., Li, R., West, R. B., van de Vijver, M. J., Sukumar, S., and Chang, H. Y. (2011) Long non-coding RNA HOTAIR reprograms chromatin state to promote cancer metastasis. *Nature 464*, 1071–1076.
(58) Kim, K., Jutooru, I., Chadalapaka, G., Johnson, G., Frank, J., Burghardt, R., Kim, S., and Safe, S. (2012) HOTAIR is a negative prognostic factor and exhibits pro-

oncogenic activity in pancreatic cancer. *Oncogene 32*, 1616–1625.
(59) Ji, Q., Zhang, L., Liu, X., Zhou, L., Wang, W., Han, Z., Sui, H., Tang, Y., Wang, Y., Liu, N., Ren, J., Hou, F., and Li, Q. (2014) Long non-coding RNA MALAT1 promotes tumour growth and metastasis in colorectal cancer through binding to SFPQ and releasing oncogene PTBP2 from SFPQ/PTBP2 complex. *Brit. J. Cancer 111*, 736–748.
(60) Wu, X.-S., Wang, X.-A., Wu, W.-G., Hu, Y.-P., Li, M.-L., Ding, Q., Weng, H., Shu, Y.-J., Liu, T.-Y., Jiang, L., Cao, Y., Bao, R.-F., Mu, J.-S., Tan, Z.-J., Tao, F., and Liu, Y.-B. (2014) MALAT1 promotes the proliferation and metastasis of gallbladder cancer cells by activating the ERK/MAPK pathway. *Cancer Biol. Ther. 15*, 806–814.
(61) Jiang, Y., Li, Y., Fang, S., Jiang, B., Qin, C., Xie, P., Zhou, G., and Li, G. (2014) The role of MALAT1 correlates with HPV in cervical cancer. *Oncol. Lett. 7*, 2135–2141.
(62) Wang, J., Wang, H., Zhang, Y., Zhen, N., Zhang, L., Qiao, Y., Weng, W., Liu, X., Ma, L., Xiao, W., Yu, W., Chu, Q., Pan, Q., and Sun, F. (2014) Cellular Signalling. *Cell. Signal. 26*, 1048–1059.
(63) Ji, P., Diederichs, S., Wang, W., Böing, S., Metzger, R., Schneider, P. M., Tidow, N., Brandt, B., Buerger, H., and Bulk, E. (2003) MALAT-1, a novel noncoding RNA, and thymosin β4 predict metastasis and survival in early-stage non-small cell lung cancer. *Oncogene 22*, 8031–8041.
(64) Gutschner, T., Hammerle, M., Eissmann, M., Hsu, J., Kim, Y., Hung, G., Revenko, A., Arun, G., Stentrup, M., Gross, M., Zornig, M., MacLeod, A. R., Spector, D. L., and Diederichs, S. (2013) The Noncoding RNA MALAT1 Is a Critical Regulator of the Metastasis Phenotype of Lung Cancer Cells. *Cancer Res. 73*, 1180–1189.
(65) Tripathi, V., Ellis, J. D., Shen, Z., Song, D. Y., Pan, Q., Watt, A. T., Freier, S. M., Bennett, C. F., Sharma, A., Bubulya, P. A., Blencowe, B. J., Prasanth, S. G., and Prasanth, K. V. (2010) The Nuclear-Retained Noncoding RNA MALAT1 Regulates Alternative Splicing by Modulating SR Splicing Factor Phosphorylation. *Mol. Cell 39*, 925–938.
(66) Yang, F., Yi, F., Han, X., Du, Q., and Liang, Z. (2013) MALAT-1 interacts with hnRNP C in cell cycle regulation. *FEBS Lett. 587*, 3175–3181.
(67) Tripathi, V., Shen, Z., Chakraborty, A., Giri, S., Freier, S. M., Wu, X., Zhang, Y., Gorospe, M., Prasanth, S. G., Lal, A., and Prasanth, K. V. (2013) Long noncoding RNA MALAT1 controls cell cycle progression by regulating the expression of oncogenic transcription factor B-MYB. *PLoS Genet. 9*, e1003368.
(68) Spector, D. L. (2001) Nuclear domains. *J. Cell Sci. 114*, 2891–2893.
(69) Pirrotta, V., and Li, H.-B. (2012) A view of nuclear Polycomb bodies. *Curr. Opin. Genet. Dev. 22*, 101–109.
(70) Fortini, E., Li, R., and Fox, A. H. (2013) Long Non-coding RNAs and Nuclear Body Formation and Function, in *Molecular Biology of Long Non-coding RNAs*, pp 197–215. Springer New York, New York, NY.
(71) Fox, A. H., Lam, Y. W., Leung, A. K. L., Lyon, C. E., Andersen, J., Mann, M., and Lamond, A. I. (2002) Paraspeckles: a novel nuclear domain. *Curr. Biol. 12*, 13–25.
(72) Nakagawa, S., Naganuma, T., Shioi, G., and Hirose, T. (2011) Paraspeckles are subpopulation-specific nuclear bodies that are not essential in mice. *J. Cell Biol. 193*, 31–39.
(73) Chen, L.-L., and Carmichael, G. G. (2009) Altered nuclear retention of mRNAs containing inverted repeats in human embryonic stem cells: functional role of a nuclear noncoding RNA. *Mol. Cell 35*, 467–478.
(74) Fox, A. H., Bond, C. S., and Lamond, A. I. (2005) P54nrb forms a heterodimer with PSP1 that localizes to paraspeckles in an RNA-dependent manner. *Mol. Biol. Cell 16*, 5304–5315.

(75) Platani, M., Goldberg, I., Swedlow, J. R., and Lamond, A. I. (2000) In vivo analysis of Cajal body movement, separation, and joining in live human cells. *J. Cell Biol. 151*, 1561–1574.
(76) Souquere, S., Beauclair, G., Harper, F., and Pierron, G. (2010) Highly ordered spatial organization of the structural long noncoding NEAT1 RNAs within paraspeckle nuclear bodies. *Mol. Biol. Cell 21*, 4020–4027.
(77) Prasanth, K. V., Prasanth, S. G., Xuan, Z., Hearn, S., Freier, S. M., Bennett, C. F., Zhang, M. Q., and Spector, D. L. (2005) Regulating gene expression through RNA nuclear retention. *Cell 123*, 249–263.
(78) Sasaki, Y., Ideue, T., Sano, M., Mituyama, T., and Hirose, T. (2009) MENε/β noncoding RNAs are essential for structural integrity of nuclear paraspeckles. *Proc. Natl. Acad. Sci. USA 106*, 2525–2530.
(79) Sunwoo, H., Dinger, M. E., Wilusz, J. E., Amaral, P. P., Mattick, J. S., and Spector, D. L. (2009) MEN epsilon/beta nuclear-retained non-coding RNAs are up-regulated upon muscle differentiation and are essential components of paraspeckles. *Genome Research 19*, 347–359.
(80) Nakagawa, S., and Hirose, T. (2012) Paraspeckle nuclear bodies--useful uselessness? *Cell. Mol. Life Sci. 69*, 3027–3036.
(81) Naganuma, T., Nakagawa, S., Tanigawa, A., Sasaki, Y. F., Goshima, N., and Hirose, T. (2012) Alternative 3'-end processing of long noncoding RNA initiates construction of nuclear paraspeckles. *EMBO J 31*, 4020–4034.
(82) Hutchinson, J. N., Ensminger, A. W., Clemson, C. M., Lynch, C. R., Lawrence, J. B., and Chess, A. (2007) A screen for nuclear transcripts identifies two linked noncoding RNAs associated with SC35 splicing domains. *BMC Genomics 8*, 39.
(83) Wilusz, J. E., Freier, S. M., and Spector, D. L. (2008) End Processing of a Long Nuclear-Retained Noncoding RNA Yields a tRNA-like Cytoplasmic RNA. *Cell 135*, 919–932.
(84) Brown, J. A., Valenstein, M. L., Yario, T. A., Tycowski, K. T., and Steitz, J. A. (2012) Formation of triple-helical structures by the 3'-end sequences of MALAT1 and MENβ noncoding RNAs. *Proc. Natl. Acad. Sci. USA 109*, 19202–19207.
(85) Mao, Y. S., Sunwoo, H., Zhang, B., and Spector, D. L. (2011) Direct visualization of the co-transcriptional assembly of a nuclear body by noncoding RNAs. *Nat. Cell Biol. 13*, 95–101.
(86) Chen, L.-L., DeCerbo, J. N., and Carmichael, G. G. (2008) Alu element-mediated gene silencing. *EMBO J 27*, 1694–1705.
(87) Imamura, K., Imamachi, N., Akizuki, G., Kumakura, M., Kawaguchi, A., Nagata, K., Kato, A., Kawaguchi, Y., Sato, H., Yoneda, M., Kai, C., Yada, T., Suzuki, Y., Yamada, T., Ozawa, T., Kaneki, K., Inoue, T., Kobayashi, M., Kodama, T., Wada, Y., Sekimizu, K., and Akimitsu, N. (2014) Long Noncoding RNA NEAT1-Dependent SFPQ Relocation from Promoter Region to Paraspeckle Mediates IL8 Expression upon Immune Stimuli. *Mol. Cell 53*, 393–406.
(88) Hirose, T., Virnicchi, G., Tanigawa, A., Naganuma, T., Li, R., Kimura, H., Yokoi, T., Nakagawa, S., Benard, M., Fox, A. H., and Pierron, G. (2014) NEAT1 long noncoding RNA regulates transcription via protein sequestration within subnuclear bodies. *Mol. Biol. Cell 25*, 169–183.
(89) Kumar, M., and Carmichael, G. G. (1997) Nuclear antisense RNA induces extensive adenosine modifications and nuclear retention of target transcripts. *Proc. Natl. Acad. Sci. USA 94*, 3542–3547.
(90) Zhang, Z., and Carmichael, G. G. (2001) The fate of dsRNA in the nucleus: a p54nrb-containing complex mediates the nuclear retention of promiscuously A-to-I edited RNAs. *Cell 106*, 465–476.

(91) Bass, B. L., and Weintraub, H. (1988) An unwinding activity that covalently modifies its double-stranded RNA substrate. *Cell 55*, 1089–1098.
(92) Bass, B. L. (2002) RNA editing by adenosine deaminases that act on RNA. *Annu. Rev. Biochem. 71*, 817–846.
(93) Athanasiadis, A., Rich, A., and Maas, S. (2004) Widespread A-to-I RNA editing of Alu-containing mRNAs in the human transcriptome. *Plos Biol. 2*, e391.
(94) Hundley, H. A., Krauchuk, A. A., and Bass, B. L. (2008) C. elegans and H. sapiens mRNAs with edited 3' UTRs are present on polysomes. *RNA 14*, 2050–2060.
(95) Capshew, C. R., Dusenbury, K. L., and Hundley, H. A. (2012) Inverted Alu dsRNA structures do not affect localization but can alter translation efficiency of human mRNAs independent of RNA editing. *Nucleic Acids Res. 40*, 8637–8645.
(96) Elbarbary, R. A., Li, W., Tian, B., and Maquat, L. E. (2013) STAU1 binding 3' UTR IRAlus complements nuclear retention to protect cells from PKR-mediated translational shutdown. *Gene. Dev. 27*, 1495–1510.
(97) Urban, R. J., Bodenburg, Y., Kurosky, A., Wood, T. G., and Gasic, S. (2000) Polypyrimidine tract-binding protein-associated splicing factor is a negative regulator of transcriptional activity of the porcine p450scc insulin-like growth factor response element. *Mol. Endocrinol. 14*, 774–782.
(98) Mathur, M., Tucker, P. W., and Samuels, H. H. (2001) PSF is a novel corepressor that mediates its effect through Sin3A and the DNA binding domain of nuclear hormone receptors. *Mol. Cell Biol. 21*, 2298–2311.
(99) Wang, G., Cui, Y., Zhang, G., Garen, A., and Song, X. (2009) Regulation of proto-oncogene transcription, cell proliferation, and tumorigenesis in mice by PSF protein and a VL30 noncoding RNA. *Proc. Natl. Acad. Sci. USA 106*, 16794–16798.
(100) Lewejohann, L., Skryabin, B. V., Sachser, N., Prehn, C., Heiduschka, P., Thanos, S., Jordan, U., Dell'Omo, G., Vyssotski, A. L., Pleskacheva, M. G., Lipp, H. P., Tiedge, H., Brosius, J., and Prior, H. (2004) Role of a neuronal small non-messenger RNA: behavioural alterations in BC1 RNA-deleted mice. *Behavioural Brain Research 154*, 273–289.
(101) Nakagawa, S., Ip, J. Y., Shioi, G., Tripathi, V., Zong, X., Hirose, T., and Prasanth, K. V. (2012) Malat1 is not an essential component of nuclear speckles in mice. *RNA*.
(102) Zhang, B., Arun, G., Mao, Y. S., Lazar, Z., Hung, G., Bhattacharjee, G., Xiao, X., Booth, C. J., Wu, J., Zhang, C., and Spector, D. L. (2012) The lncRNA Malat1 is dispensable for mouse development but its transcription plays a cis-regulatory role in the adult. *Cell Rep. 2*, 111–123.
(103) Eißmann, M., Gutschner, T., Hämmerle, M., Günther, S., Caudron-Herger, M., Gross, M., Schirmacher, P., Rippe, K., Braun, T., Zörnig, M., and Diederichs, S. (2012) Loss of the abundant nuclear non-coding RNA MALAT1 is compatible with life and development. *RNA Biol. 9*, 1076–1087.
(104) Sauvageau, M., Goff, L. A., Lodato, S., Bonev, B., Groff, A. F., Gerhardinger, C., Sanchez-Gomez, D. B., Hacisuleyman, E., Li, E., Spence, M., Liapis, S. C., Mallard, W., Morse, M., Swerdel, M. R., D'Ecclessis, M. F., Moore, J. C., Lai, V., Gong, G., Yancopoulos, G. D., Frendewey, D., Kellis, M., Hart, R. P., Valenzuela, D. M., Arlotta, P., and Rinn, J. L. (2013) Multiple knockout mouse models reveal lincRNAs are required for life and brain development. *Elife 2*, e01749.
(105) Standaert, L., Adriaens, C., Radaelli, E., Van Keymeulen, A., Blanpain, C., Hirose, T., Nakagawa, S., and Marine, J.-C. (2014) The long noncoding RNA Neat1 is required for mammary gland development and lactation. *RNA 20*, 1844–1849.
(106) Nakagawa, S., Shimada, M., Yanaka, K., Mito, M., Arai, T., Takahashi, E., Fujita, Y., Fujimori, T., Standaert, L., Marine, J.-C., and Hirose, T. (2014) The lncRNA Neat1 is required for corpus luteum formation and the establishment of pregnancy in a

subpopulation of mice. *Development 141*, 4618–4627.
(107) Chakravarty, D., Sboner, A., Nair, S. S., Giannopoulou, E., Li, R., Hennig, S., Mosquera, J. M., Pauwels, J., Park, K., Kossai, M., MacDonald, T. Y., Fontugne, J., Erho, N., Vergara, I. A., Ghadessi, M., Davicioni, E., Jenkins, R. B., Palanisamy, N., Chen, Z., Nakagawa, S., Hirose, T., Bander, N. H., Beltran, H., Fox, A. H., Elemento, O., and Rubin, M. A. (2014) The oestrogen receptor alpha-regulated lncRNA NEAT1 is a critical modulator of prostate cancer. *Nat. Comm. 5*, 5383.
(108) Choudhry, H., Albukhari, A., Morotti, M., Hider, S., Moralli, D., Smythies, J., del, J. S. O., Green, C. M., Camps, C., Buffa, F., Ratcliffe, P., Ragoussis, J., Harris, A. L., and Mole, D. R. (2014) Tumor hypoxia induces nuclear paraspeckle formation through HIF-2. *Oncogene* 1–9.
(109) Saha, S., Murthy, S., and Rangarajan, P. N. (2006) Identification and characterization of a virus-inducible non-coding RNA in mouse brain. *J. Gen. Virol. 87*, 1991–1995.
(110) Zhang, Q., Chen, C.-Y., Yedavalli, V. S. R. K., and Jeang, K.-T. (2013) NEAT1 long noncoding RNA and paraspeckle bodies modulate HIV-1 posttranscriptional expression. *MBio 4*, e00596–12.
(111) Dong, B., Horowitz, D. S., Kobayashi, R., and Krainer, A. R. (1993) Purification and cDNA cloning of HeLa cell p54nrb, a nuclear protein with two RNA recognition motifs and extensive homology to human splicing factor PSF and Drosophila NONA/BJ6. *Nucleic Acids Res. 21*, 4085–4092.
(112) Kanai, Y., Dohmae, N., and Hirokawa, N. (2004) Kinesin transports RNA: isolation and characterization of an RNA-transporting granule. *Neuron 43*, 513–525.
(113) Dong, X., Sweet, J., Challis, J. R. G., Brown, T., and Lye, S. J. (2007) Transcriptional activity of androgen receptor is modulated by two RNA splicing factors, PSF and p54nrb. *Mol. Cell Biol. 27*, 4863–4875.
(114) Emili, A., Shales, M., McCracken, S., Xie, W., Tucker, P. W., Kobayashi, R., Blencowe, B. J., and Ingles, C. J. (2002) Splicing and transcription-associated proteins PSF and p54nrb/nonO bind to the RNA polymerase II CTD. *RNA 8*, 1102–1111.
(115) Marko, M., Leichter, M., Patrinou-Georgoula, M., and Guialis, A. (2010) hnRNP M interacts with PSF and p54(nrb) and co-localizes within defined nuclear structures. *Exp. Cell Res. 316*, 390–400.
(116) Gozani, O., Patton, J. G., and Reed, R. (1994) A novel set of spliceosome-associated proteins and the essential splicing factor PSF bind stably to pre-mRNA prior to catalytic step II of the splicing reaction. *EMBO J 13*, 3356–3367.
(117) Kaneko, S., Rozenblatt-Rosen, O., Meyerson, M., and Manley, J. L. (2007) The multifunctional protein p54nrb/PSF recruits the exonuclease XRN2 to facilitate pre-mRNA 3' processing and transcription termination. *Gene. Dev. 21*, 1779–1789.
(118) Hall-Pogar, T., Liang, S., Hague, L. K., and Lutz, C. S. (2007) Specific trans-acting proteins interact with auxiliary RNA polyadenylation elements in the COX-2 3'-UTR. *RNA 13*, 1103–1115.
(119) Kozlova, N., Braga, J., Lundgren, J., Rino, J., Young, P., Carmo-Fonseca, M., and Visa, N. (2006) Studies on the role of NonA in mRNA biogenesis. *Exp. Cell Res. 312*, 2619–2630.
(120) Lowery, L. A., Rubin, J., and Sive, H. (2007) whitesnake/sfpqis required for cell survival and neuronal development in the zebrafish. *Dev. Dyn. 236*, 1347–1357.
(121) Chanas-Sacré, G., Mazy-Servais, C., Wattiez, R., Pirard, S., Rogister, B., Patton, J. G., Belachew, S., Malgrange, B., Moonen, G., and Leprince, P. (1999) Identification of PSF, the polypyrimidine tract-binding protein-associated splicing factor, as a

developmentally regulated neuronal protein. *J. Neurosci. Res. 57*, 62–73.
(122) Tapia-Páez, I., Tammimies, K., Massinen, S., Roy, A. L., and Kere, J. (2008) The complex of TFII-I, PARP1, and SFPQ proteins regulates the DYX1C1 gene implicated in neuronal migration and dyslexia. *FASEB J. 22*, 3001–3009.
(123) Nakatani, N., Hattori, E., Ohnishi, T., Dean, B., Iwayama, Y., Matsumoto, I., Kato, T., Osumi, N., Higuchi, T., Niwa, S.-I., and Yoshikawa, T. (2006) Genome-wide expression analysis detects eight genes with robust alterations specific to bipolar I disorder: relevance to neuronal network perturbation. *Hum. Mol. Genet. 15*, 1949–1962.
(124) Ke, Y., Dramiga, J., Schütz, U., Kril, J. J., Ittner, L. M., Schröder, H., and Götz, J. (2012) Tau-Mediated Nuclear Depletion and Cytoplasmic Accumulation of SFPQ in Alzheimer"s and Pick"s Disease. *PLoS ONE 7*, e35678.
(125) Ling, S.-C., Albuquerque, C. P., Han, J. S., Lagier-Tourenne, C., Tokunaga, S., Zhou, H., and Cleveland, D. W. (2010) ALS-associated mutations in TDP-43 increase its stability and promote TDP-43 complexes with FUS/TLS. *Proc. Natl. Acad. Sci. USA 107*, 13318–13323.
(126) Shelkovnikova, T. A., Robinson, H. K., Troakes, C., Ninkina, N., and Buchman, V. L. (2013) Compromised paraspeckle formation as a pathogenic factor in FUSopathies. *hum. Mol. Genet.*
(127) Peng, R., Dye, B. T., Pérez, I., Barnard, D. C., Thompson, A. B., and Patton, J. G. (2002) PSF and p54nrb bind a conserved stem in U5 snRNA. *RNA 8*, 1334–1347.
(128) Myojin, R., Kuwahara, S., Yasaki, T., Matsunaga, T., Sakurai, T., Kimura, M., Uesugi, S., and Kurihara, Y. (2004) Expression and functional significance of mouse paraspeckle protein 1 on spermatogenesis. *Biology of Reproduction 71*, 926–932.
(129) Kuwahara, S., Ikei, A., Taguchi, Y., Fujimoto, N., Obinata, M., Uesugi, S., and Kurihara, Y. (2006) PSPC1, NONO, and SFPQ are expressed in mouse Sertoli cells and may function as coregulators of androgen receptor-mediated transcription. *Biology of Reproduction 75*, 352–359.
(130) Passon, D. M., Lee, M., Rackham, O., Stanley, W. A., Sadowska, A., Filipovska, A., Fox, A. H., and Bond, C. S. (2012) Structure of the heterodimer of human NONO and paraspeckle protein component 1 and analysis of its role in subnuclear body formation. *Proc. Natl. Acad. Sci. USA 109*, 4846–4850.
(131) Akhmedov, A. T., and Lopez, B. S. (2000) Human 100-kDa homologous DNA-pairing protein is the splicing factor PSF and promotes DNA strand invasion. *Nucleic Acids Res. 28*, 3022–3030.
(132) Bladen, C. L., Udayakumar, D., Takeda, Y., and S, D. W. (2005) Identification of the Polypyrimidine Tract Binding Protein-associated Splicing Factor·p54(nrb) Complex as a Candidate DNA Double-strand Break Rejoining Factor. *J. Biol. Chem.*
(133) Passon, D. M., Lee, M., Fox, A. H., and Bond, C. S. (2011) Crystallization of a paraspeckle protein PSPC1-NONO heterodimer. *Acta Crystal F67*, 1231–1234.
(134) Lee, M., Passon, D. M., Hennig, S., Fox, A. H., and Bond, C. S. (2011) Construct optimization for studying protein complexes: obtaining diffraction-quality crystals of the pseudosymmetric PSPC1-NONO heterodimer. *Acta Crystal D67*, 981–987.
(135) Hebert, M. D., and Matera, A. G. (2000) Self-association of coilin reveals a common theme in nuclear body localization. *Mol. Biol. Cell 11*, 4159–4171.
(136) Chen, T., Boisvert, F. M., Bazett-Jones, D. P., and Richard, S. (1999) A role for the GSG domain in localizing Sam68 to novel nuclear structures in cancer cell lines. *Mol. Biol. Cell 10*, 3015–3033.
(137) Basu, A., Dong, B., Krainer, A. R., and Howe, C. C. (1997) The intracisternal A-particle proximal enhancer-binding protein activates transcription and is identical to the RNA- and DNA-binding protein p54nrb/NonO. *Mol. Cell Biol. 17*, 677–686.

(138) Schmittgen, T. D., and Livak, K. J. (2008) Analyzing real-time PCR data by the comparative CT method. *Nat. Protoc. 3*, 1101–1108.
(139) Hafner, M., Landthaler, M., Burger, L., Khorshid, M., Hausser, J., Berninger, P., Rothballer, A., Ascano, M., Jr., Jungkamp, A.-C., Munschauer, M., Ulrich, A., Wardle, G. S., Dewell, S., Zavolan, M., and Tuschl, T. (2010) Transcriptome-wide identification of RNA-binding protein and microRNA target sites by PAR-CLIP. *Cell 141*, 129–141.
(140) Corcoran, D. L., Georgiev, S., Mukherjee, N., Gottwein, E., Skalsky, R. L., Keene, J. D., and Ohler, U. (2011) PARalyzer: definition of RNA binding sites from PAR-CLIP short-read sequence data. *Genome Biol. 12*, R79.
(141) Bailey, T. L., and Elkan, C. (1994) Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc Int Conf Intell Syst Mol Biol 2*, 28–36.
(142) Alexander, P., and Moroson, H. (1962) Cross-linking of deoxyribonucleic acid to protein following ultra-violet irradiation of different cells. *Nature 194*, 882–883.
(143) König, J., Zarnack, K., Luscombe, N. M., and Ule, J. (2011) Protein–RNA interactions: new genomic technologies and perspectives : Abstract : Nature Reviews Genetics. *Nat. rev. Genet. 13*, 77–83.
(144) Niranjanakumari, S., Lasda, E., Brazas, R., and Garcia-Blanco, M. A. (2002) Reversible cross-linking combined with immunoprecipitation to study RNA-protein interactions in vivo. *Methods 26*, 182–190.
(145) Greenberg, J. R. (1979) Ultraviolet light-induced crosslinking of mRNA to proteins. *Nucleic Acids Res. 6*, 715–732.
(146) Ule, J., Jensen, K., Mele, A., and Darnell, R. B. (2005) CLIP: A method for identifying protein–RNA interaction sites in living cells. *Methods 37*, 376–386.
(147) Yeo, G. W., Coufal, N. G., Liang, T. Y., Peng, G. E., Fu, X.-D., and Gage, F. H. (2009) An RNA code for the FOX2 splicing regulator revealed by mapping RNA-protein interactions in stem cells. *Nat. Struct. Biol. 16*, 130–137.
(148) Ule, J., Jensen, K. B., Ruggiu, M., Mele, A., Ule, A., and Darnell, R. B. (2003) CLIP identifies Nova-regulated RNA networks in the brain. *Sci. Sig. 302*, 1212.
(149) Licatalosi, D. D., Mele, A., Fak, J. J., Ule, J., Kayikci, M., Chi, S. W., Clark, T. A., Schweitzer, A. C., Blume, J. E., Wang, X., Darnell, J. C., and Darnell, R. B. (2008) HITS-CLIP yields genome-wide insights into brain alternative RNA processing. *Nature 456*, 464–469.
(150) König, J., Zarnack, K., Rot, G., Curk, T., Kayikci, M., Zupan, B., Turner, D. J., Luscombe, N. M., and Ule, J. (2010) iCLIP reveals the function of hnRNP particles in splicing at individual nucleotide resolution. *Nat. Struct. Biol. 17*, 909–915.
(151) König, J., Zarnack, K., Rot, G., Curk, T., Kayikci, M., Zupan, B., Turner, D. J., Luscombe, N. M., and Ule, J. (2011) iCLIP - Transcriptome-wide Mapping of Protein-RNA Interactions with Individual Nucleotide Resolution. *JoVE*.
(152) Meisenheimer, K. M., and Koch, T. H. (1997) Photocross-linking of nucleic acids to associated proteins. *Crit. Rev. Biochem. Mol. 32*, 101–140.
(153) Favre, A., Moreno, G., Blondel, M. O., Kliber, J., Vinzens, F., and Salet, C. (1986) 4-Thiouridine photosensitized RNA-protein crosslinking in mammalian cells. *Biochem. Bioph. Res. Co. 141*, 847–854.
(154) Hafner, M., Max, K. E. A., Bandaru, P., Morozov, P., Gerstberger, S., Brown, M., Molina, H., and Tuschl, T. (2013) Identification of mRNAs bound and regulated by human LIN28 proteins and molecular requirements for RNA recognition. *RNA 19*, 613–626.
(155) Lebedeva, S., Jens, M., Theil, K., Schwanhausser, B., Selbach, M., Landthaler, M., and Rajewsky, N. (2011) Transcriptome-wide Analysis of Regulatory Interactions of the RNA-Binding Protein HuR. *Mol. Cell 43*, 340–352.

(156) Trinkle-Mulcahy, L., Boulon, S., Lam, Y. W., Urcia, R., Boisvert, F.-M., Vandermoere, F., Morrice, N. A., Swift, S., Rothbauer, U., Leonhardt, H., and Lamond, A. (2008) Identifying specific protein interaction partners using quantitative mass spectrometry and bead proteomes. *J. Cell Biol. 183*, 223–239.
(157) Kishore, S., Jaskiewicz, L., Burger, L., Hausser, J., Khorshid, M., and Zavolan, M. (2011) A quantitative analysis of CLIP methods for identifying binding sites of RNA-binding proteins. *Nat. Methods 8*, 559.
(158) Langmead, B., Trapnell, C., Pop, M., and Salzberg, S. L. (2009) Bowtie paper Ultrafast and memory efficient alignment or short DNA sequences to the human genome. *Genome Biol. R25*.
(159) Huppertz, I., Attig, J., D'Ambrogio, A., Easton, L. E., Sibley, C. R., Sugimoto, Y., Tajnik, M., König, J., and Ule, J. (2014) iCLIP: Protein-RNA interactions at nucleotide resolution. *Methods 65*, 274–287.
(160) Burger, K., Mühl, B., Kellner, M., Rohrmoser, M., Gruber-Eber, A., Windhager, L., Friedel, C. C., Dölken, L., and Eick, D. (2013) 4-thiouridine inhibits rRNA synthesis and causes a nucleolar stress response. *RNA Biol. 10*, 1623–1630.
(161) De, S., Srikantan, S., Abdelmohsen, K., Grammatikakis, I., Kim, J., Kim, K. M., Noh, J. H., White, E. J. F., Martindale, J. L., Yang, X., Kang, M.-J., Wood, W. H., Hooten, N. N., Evans, M. K., Becker, K. G., Tripathi, V., Prasanth, K. V., Wilson, G. M., Tuschl, T., Ingolia, N. T., Hafner, M., Yoon, J.-H., and Gorospe, M. (2014) PAR-CLIP analysis uncovers AUF1 impact on target RNA fate and genome integrity. *Nat. Comm. 5*, 1–15.
(162) Faudale, M., Cogoi, S., and Xodo, L. E. (2011) Photoactivated cationic alkyl-substituted porphyrin binding to g4-RNA in the 5′-UTR of KRAS oncogene represses translation. *Chem. Commun. 48*, 874.
(163) Wolfe, A. L., Singh, K., Zhong, Y., Drewe, P., Rajasekhar, V. K., Sanghvi, V. R., Mavrakis, K. J., Jiang, M., Roderick, J. E., Van der Meulen, J., Schatz, J. H., Rodrigo, C. M., Zhao, C., Rondou, P., de Stanchina, E., Teruya-Feldstein, J., Kelliher, M. A., Speleman, F., Porco, J. A., Pelletier, J., Rätsch, G., and Wendel, H.-G. (2015) RNA G-quadruplexes cause eIF4A- dependent oncogene translation in cancer. *Nature 513*, 65–70.
(164) Haeusler, A. R., Donnelly, C. J., Periz, G., Simko, E. A. J., Shaw, P. G., Kim, M.-S., Maragakis, N. J., Troncoso, J. C., Pandey, A., Sattler, R., Rothstein, J. D., and Wang, J. (2014) C9orf72 nucleotide repeat structures initiate molecular cascades of disease. *Nature 507*, 195–200.
(165) Kikin, O., D'Antonio, L., and Bagga, P. S. (2006) QGRS Mapper: a web-based server for predicting G-quadruplexes in nucleotide sequences. *Nucleic Acids Res. 34*, W676–82.
(166) Takayama, K.-I., Horie-Inoue, K., Katayama, S., Suzuki, T., Tsutsumi, S., Ikeda, K., Urano, T., Fujimura, T., Takagi, K., Takahashi, S., Homma, Y., Ouchi, Y., Aburatani, H., Hayashizaki, Y., and Inoue, S. (2013) Androgen-responsive long noncoding RNA CTBP1-AS promotes prostate cancer. *EMBO J 32*, 1665–1680.
(167) Reitmair, A., Sachs, G., Im, W. B., and Wheeler, L. (2012) C6orf176: a novel possible regulator of cAMP-mediated gene expression. *Physiol. Genomics 44*, 152–161.
(168) Xiang, J.-F., Yin, Q.-F., Chen, T., Zhang, Y., Zhang, X.-O., Wu, Z., Zhang, S., Wang, H.-B., Ge, J., Lu, X., Yang, L., and Chen, L.-L. (2014) Human colorectal cancer-specific CCAT1-L lncRNA regulates long-range chromatin interactions at the MYC locus. *Cell Res. 24*, 1150.
(169) Xiang, J.-F., Yin, Q.-F., Chen, T., Zhang, Y., Zhang, X.-O., Wu, Z., Zhang, S., Wang, H.-B., Ge, J., Lu, X., Yang, L., and Chen, L.-L. (2014) Human colorectal cancer-specific

CCAT1-L lncRNA regulates long-range chromatin interactions at the MYC locus. *Cell Res. 24*, 513–531.
(170) Cabili, M. N., Trapnell, C., Goff, L., Koziol, M., Tazon-Vega, B., Regev, A., and Rinn, J. L. (2011) Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. *Gene. Dev. 25*, 1915–1927.
(171) Li, S., Li, Z., Shu, F.-J., Xiong, H., Phillips, A. C., and S, D. W. (2014) Double-strand break repair deficiency in NONO knockout murine embryonic fibroblasts and compensation by spontaneous upregulation of the PSPC1 paralog. *Nucleic Acids Res. 42*, 9771–9780.
(172) Friedersdorf, M. B., and Keene, J. D. (2014) Advancing the functional utility of PAR-CLIP by quantifying background binding to mRNAs and lncRNAs. *Genome Biol. 15*, R2.
(173) Hafner, M., Landthaler, M., Burger, L., Khorshid, M., Hausser, J., Berninger, P., Rothballer, A., Ascano, M., Jungkamp, A.-C., Munschauer, M., Ulrich, A., Wardle, G. S., Dewell, S., Zavolan, M., and Tuschl, T. (2010) PAR-CliP - A Method to Identify Transcriptome-wide the Binding Sites of RNA Binding Proteins. *JoVE*.
(174) Forrest, A. R. R., Kawaji, H., Rehli, M., Kenneth Baillie, J., de Hoon, M. J. L., Haberle, V., Lassmann, T., Kulakovskiy, I. V., Lizio, M., Itoh, M., Andersson, R., Mungall, C. J., Meehan, T. F., Schmeier, S., Bertin, N., Jørgensen, M., Dimont, E., Arner, E., Schmidl, C., Schaefer, U., Medvedeva, Y. A., Plessy, C., Vitezic, M., Severin, J., Semple, C. A., Ishizu, Y., Young, R. S., Francescatto, M., Alam, I., Albanese, D., Altschuler, G. M., Arakawa, T., Archer, J. A. C., Arner, P., Babina, M., Rennie, S., Balwierz, P. J., Beckhouse, A. G., Pradhan-Bhatt, S., Blake, J. A., Blumenthal, A., Bodega, B., Bonetti, A., Briggs, J., Brombacher, F., Maxwell Burroughs, A., Califano, A., Cannistraci, C. V., Carbajo, D., Chen, Y., Chierici, M., Ciani, Y., Clevers, H. C., Dalla, E., Davis, C. A., Detmar, M., Diehl, A. D., Dohi, T., Drabløs, F., Edge, A. S. B., Edinger, M., Ekwall, K., Endoh, M., Enomoto, H., Fagiolini, M., Fairbairn, L., Fang, H., Farach-Carson, M. C., Faulkner, G. J., Favorov, A. V., Fisher, M. E., Frith, M. C., Fujita, R., Fukuda, S., Furlanello, C., Furuno, M., Furusawa, J.-I., Geijtenbeek, T. B., Gibson, A. P., Gingeras, T., Goldowitz, D., Gough, J., Guhl, S., Guler, R., Gustincich, S., Ha, T. J., Hamaguchi, M., Hara, M., Harbers, M., Harshbarger, J., Hasegawa, A., Hasegawa, Y., Hashimoto, T., Herlyn, M., Hitchens, K. J., Ho Sui, S. J., Hofmann, O. M., Hoof, I., Hori, F., Huminiecki, L., Iida, K., Ikawa, T., Jankovic, B. R., Jia, H., Joshi, A., Jurman, G., Kaczkowski, B., Kai, C., Kaida, K., Kaiho, A., Kajiyama, K., Kanamori-Katayama, M., Kasianov, A. S., Kasukawa, T., Katayama, S., Kato, S., Kawaguchi, S., Kawamoto, H., Kawamura, Y. I., Kawashima, T., Kempfle, J. S., Kenna, T. J., Kere, J., Khachigian, L. M., Kitamura, T., Peter Klinken, S., Knox, A. J., Kojima, M., Kojima, S., Kondo, N., Koseki, H., Koyasu, S., Krampitz, S., Kubosaki, A., Kwon, A. T., Laros, J. F. J., Lee, W., Lennartsson, A., Li, K., Lilje, B., Lipovich, L., Mackay-sim, A., Manabe, R.-I., Mar, J. C., Marchand, B., Mathelier, A., Mejhert, N., Meynert, A., Mizuno, Y., de Lima Morais, D. A., Morikawa, H., Morimoto, M., Moro, K., Motakis, E., Motohashi, H., Mummery, C. L., Murata, M., Nagao-Sato, S., Nakachi, Y., Nakahara, F., Nakamura, T., Nakamura, Y., Nakazato, K., van Nimwegen, E., Ninomiya, N., Nishiyori, H., Noma, S., Nozaki, T., Ogishima, S., Ohkura, N., Ohmiya, H., Ohno, H., Ohshima, M., Okada-Hatakeyama, M., Okazaki, Y., Orlando, V., Ovchinnikov, D. A., Pain, A., Passier, R., Patrikakis, M., Persson, H., Piazza, S., Prendergast, J. G. D., Rackham, O. J. L., Ramilowski, J. A., Rashid, M., Ravasi, T., Rizzu, P., Roncador, M., Roy, S., Rye, M. B., Saijyo, E., Sajantila, A., Saka, A., Sakaguchi, S., Sakai, M., Sato, H., Satoh, H., Savvi, S., Saxena, A., Schneider, C., Schultes, E. A., Schulze-Tanzil, G. G., Schwegmann, A., Sengstag, T., Sheng, G., Shimoji, H., Shimoni, Y., Shin, J. W., Simon, C., Sugiyama, D., Sugiyama, T., Suzuki, M., Suzuki, N., Swoboda, R. K., t Hoen, P. A. C., Tagami, M., Takahashi, N.,

Takai, J., Tanaka, H., Tatsukawa, H., Tatum, Z., Thompson, M., Toyoda, H., Toyoda, T., Valen, E., van de Wetering, M., van den Berg, L. M., Verardo, R., Vijayan, D., Vorontsov, I. E., Wasserman, W. W., Watanabe, S., Wells, C. A., Winteringham, L. N., Wolvetang, E., Wood, E. J., Yamaguchi, Y., Yamamoto, M., Yoneda, M., Yonekura, Y., Yoshida, S., Zabierowski, S. E., Zhang, P. G., Zhao, X., Zucchelli, S., Summers, K. M., Suzuki, H., Daub, C. O., Kawai, J., Heutink, P., Hide, W., Freeman, T. C., Lenhard, B., Bajic, V. B., Taylor, M. S., Makeev, V. J., Sandelin, A., Hume, D. A., Carninci, P., and Hayashizaki, Y. (2014) A promoter-level mammalian expression atlas. *Nature 507*, 462–470.
(175) Tsofack, S. P., Garand, C., Sereduk, C., Chow, D., Aziz, M., Guay, D., Yin, H. H., and Lebel, M. (2011) NONO and RALY proteins are required for YB-1oxaliplatin induced resistance in colonadenocarcinoma cell lines. *Mol. Cancer 10*, 145.
(176) Schiffner, S., Zimara, N., Schmid, R., and Bosserhoff, A. K. (2011) p54nrb is a new regulator of progression of malignant melanoma. *Carcinogenesis 32*, 1176–1182.
(177) Pavao, M., Huang, Y., Hafer, L. J., Moreland, R. B., and Traish, A. M. (2001) Immunodetection of mnt55/p54nrb isoforms in human breast cancer. *BMC cancer 1*, 1–10.
(178) Bailey, T. L., Bodén, M., Buske, F. A., Frith, M., Grant, C. E., Clementi, L., Ren, J., Li, W. W., and Noble, W. S. (2009) MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Res. 37*, W202–8.
(179) Tollervey, J. R., Curk, T., Rogelj, B., Briese, M., Cereda, M., Kayikci, M., König, J., Hortobágyi, T., Nishimura, A. L., Župunski, V., Patani, R., Chandran, S., Rot, G., Zupan, B., Shaw, C. E., and Ule, J. (2011) Characterizing the RNA targets and position-dependent splicing regulation by TDP-43. *Nat. Neurosci. 14*, 452–458.
(180) Maris, C., Dominguez, C., and allain, F. H. T. (2005) The RNA recognition motif, a plastic RNA-binding platform to regulate post-transcriptional gene expression. *FEBS J. 272*, 2118–2131.
(181) Wang, X., McLachlan, J., Zamore, P. D., and Tanaka Hall, T. M. (2002) Modular Recognition of RNAby a Human Pumilio-Homology Domain. *Cell 110*, 501–512.
(182) Ramos, A., Hollingworth, D., Major, S. A., Adinolfi, S., Kelly, G., Muskett, F. W., and Pastore, A. (2002) Role of Dimerization in KH/RNA Complexes: The Example of Nova KH3. *Biochemistry 41*, 4193–4201.
(183) Shamoo, Y., Abdul-Manan, N., Patten, A. M., Crawford, J. K., Pellegrini, M. C., and Williams, K. R. (1994) Both RNA-binding domains in heterogenous nuclear ribonucleoprotein A1 contribute toward single-stranded-RNA binding. *Biochemistry 33*, 8272–8281.
(184) Shamoo, J., Abdul-Manan, N., and Williams, K. R. (1995) MultipleRNA binding domains (RBDs) just don't add up. *Nucleic Acids Res. 23*, 725–728.
(185) Wang, X., and Tanaka Hall, T. M. (2001) Structural basis for recognition of AU-rich element RNA by the HuD protein. *Nat. Struct. Biol. 8*, 141–145.
(186) Oubridge, C., Ito, N., Evans, P. R., Teo, C. H., and Nagai, K. (1994) Crystal structure at 1.92 A resolution of the RNA-binding domain of the U1A spliceosomal protein complexed with an RNA hairpin. *Nature 372*, 432–438.
(187) Handa, N., Nureki, O., Kurimoto, K., Kim, I., Sakamoto, H., Shimura, Y., Muto, Y., and Yokoyama, S. (1999) Structural basis for recognition of the tra mRNA precursor by the Sex-lethal protein. *Nature 398*, 579–585.
(188) Winter, E., Yamamoto, F., Almoguera, C., and Perucho, M. (1985) A method to detect and characterize point mutations in transcribed genes: amplification and overexpression of the mutant c-Ki-ras allele in human tumor cells. *Proc. Natl. Acad. Sci. USA 82*, 7575–7579.

(189) Aranda, M. A., Fraile, A., Garcia-Arenal, F., and Malpica, J. M. (1995) Experimental evaluation of the ribonuclease protection assay method for the assessment of genetic heterogeneity in populations of RNA viruses. *Arch. Virol. 140*, 1373–1383.
(190) Herschlag, D. (1995) RNA chaperones and the RNA folding problem. *J. Biol. Chem. 270*, 20871–20874.
(191) Bokinsky, G., Nivón, L. G., Liu, S., Chai, G., Hong, M., Weeks, K. M., and Zhuang, X. (2006) Two Distinct Binding Modes of a Protein Cofactor with its Target RNA. *J. Mol. Biol. 361*, 771–784.
(192) Zuker, M. (2003) Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res. 31*, 3406–3415.
(193) Sato, K., Hamada, M., Asai, K., and Mituyama, T. (2009) CENTROIDFOLD: a web server for RNA secondary structure prediction. *Nucleic Acids Res. 37*, W277–80.
(194) Sharma, S., Ding, F., and Dokholyan, N. V. (2008) iFoldRNA: three-dimensional RNA structure prediction and folding. *Bioinformatics 24*, 1951–1952.
(195) Das, R., and Baker, D. (2007) Automated de novo prediction of native-like RNA tertiary structures. *Proc. Natl. Acad. Sci. USA 104*, 14664–14669.
(196) Lindgreen, S., Gardner, P. P., and Krogh, A. (2006) Measuring covariation in RNA alignments: physical realism improves information measures. *Bioinformatics 22*, 2988–2995.
(197) Bernhart, S. H., Hofacker, I. L., Will, S., Gruber, A. R., and Stadler, P. F. (2008) RNAalifold: improved consensus structure prediction for RNA alignments. *BMC Bioinformatics 9*, 474.
(198) Ding, Y., Tang, Y., Kwok, C. K., Zhang, Y., Bevilacqua, P. C., and Assmann, S. M. (2014) In vivo genome-wide profiling of RNA secondary structure reveals novel regulatory features. *Nature 505*, 696–700.
(199) Weeks, K. M. (2010) Advances in RNA structure analysis by chemical probing. *Curr. Opin. Struc. Biol. 20*, 295–304.
(200) Mathews, D. H., Disney, M. D., Childs, J. L., Schroeder, S. J., Zuker, M., and Turner, D. H. (2004) Incorporating chemical modification constraints into a dynamic programming algorithm for prediction of RNA secondary structure. *Proc. Natl. Acad. Sci. USA 101*, 7287–7292.
(201) West, J. A., Davis, C. P., Sunwoo, H., Simon, M. D., Sadreyev, R. I., Wang, P. I., Tolstorukov, M. Y., and Kingston, R. E. (2014) The Long Noncoding RNAs NEAT1 and MALAT1 Bind Active Chromatin Sites. *Mol. Cell* 1–51.
(202) Ong, S. A., Tan, J. J., Tew, W. L., and Chen, K.-S. (2011) Rasd1 modulates the coactivator function of NonO in the cyclic AMP pathway. *PLoS ONE 6*, e24401.
(203) Takahashi, K., Tanabe, K., Ohnuki, M., Narita, M., Ichisaka, T., Tomoda, K., and Yamanaka, S. (2007) Induction of Pluripotent Stem Cells from Adult Human Fibroblasts by Defined Factors. *Cell 131*, 861–872.
(204) Adams, J. M., Harris, A. W., Pinkert, C. A., Corcoran, L. M., Alexander, W. S., Cory, S., Palmiter, R. D., and Brinster, R. L. (1985) The c-myc oncogene driven by immunoglobulin enhancers induces lymphoid malignancy in transgenic mice. *Nature 318*, 533–538.
(205) Nilsson, J. A., and Cleveland, J. L. (2003) Myc pathways provoking cell suicide and cancer. *Oncogene 22*, 9007–9021.
(206) Nagano, T., Mitchell, J. A., Sanz, L. A., Pauler, F. M., Ferguson-Smith, A. C., Feil, R., and Fraser, P. (2008) The Air noncoding RNA epigenetically silences transcription by targeting G9a to chromatin. *Science 322*, 1717–1720.
(207) Huarte, M., Guttman, M., Feldser, D., Garber, M., and Bosserhoff, A. K. (2010)

A large intergenic noncoding RNA induced by p53 mediates global gene repression in the p53 response. *Cell 142*, 409–419.
(208) Lagier-Tourenne, C., Polymenidou, M., Hutt, K. R., Vu, A. Q., Baughn, M., Huelga, S. C., Clutario, K. M., Ling, S.-C., Liang, T. Y., Mazur, C., Wancewicz, E., Kim, A. S., Watt, A., Freier, S., Hicks, G. G., Donohue, J. P., Shiue, L., Bennett, C. F., Ravits, J., Cleveland, D. W., and Yeo, G. W. (2012) Divergent roles of ALS-linked proteins FUS/TLS and TDP-43 intersect in processing long pre-mRNAs. *Nat. Neurosci. 15*, 1488–1497.
(209) Polymenidou, M., Lagier-Tourenne, C., Hutt, K. R., Huelga, S. C., Moran, J., Liang, T. Y., Ling, S.-C., Sun, E., Wancewicz, E., Mazur, C., Kordasiewicz, H., Sedaghat, Y., Donohue, J. P., Shiue, L., Bennett, C. F., Yeo, G. W., and Cleveland, D. W. (2011) Long pre-mRNA depletion and RNA missplicing contribute to neuronal vulnerability from loss of TDP-43. *Nat. Neurosci. 14*, 459–468.
(210) Werner, H., and LeRoith, D. (1996) The role of the Insulin-like growth factor system in human cancers. *Advances in Cancer Research 68*, 184–223.
(211) Hellawell, G. O., Turner, G. D. H., Davies, D. R., Poulsom, R., Brewster, S. F., and Macaulay, V. M. (2002) Expression of the type 1 insulin-like growth factor receptor is up-regulated in primary prostate cancer and commonly persists in metastatic disease. *Cancer Res. 62*, 2942–2950.
(212) Krueckl, S. L., Sikes, R. A., Edlund, N. M., Bell, R. H., Hurtado-Coll, A., Fazli, L., Gleave, M. E., and Cox, M. E. (2004) Increased insulin-like growth factor I receptor expression and signaling are components of androgen-independent progression in a lineage-derived prostate cancer progression model. *Cancer Res. 64*, 8620–8629.
(213) Belgrader, P., Dey, R., and Berezney, R. (1991) Molecular cloning of matrin 3. A 125-kilodalton protein of the nuclear matrix contains an extensive acidic domain. *J. Biol. Chem. 266*, 9893–9899.
(214) Yedavalli, V. S. R. K., and Jeang, K.-T. (2011) Matrin 3 is a co-factor for HIV-1 Rev in regulating post-transcriptional viral gene expression. *Retrovirology 8*, 61.
(215) Tanaka, K., Miyamoto, N., Shouguchi-Miyata, J., and Ikeda, J.-E. (2006) HFM1, the human homologue of yeast Mer3, encodes a putative DNA helicase expressed specifically in germ-line cells. *DNA Sequence 17*, 242–246.
(216) Pontén, F., Jirström, K., and Uhlen, M. (2008) The Human Protein Atlas-a tool for pathology. *J. Pathol. 216*, 387–393.
(217) Paz, N., Levanon, E. Y., Amariglio, N., Heimberger, A. B., Ram, Z., Constantini, S., Barbash, Z. S., Adamsky, K., Safran, M., Hirschberg, A., Krupsky, M., Ben-Dov, I., Cazacu, S., Mikkelsen, T., Brodie, C., Eisenberg, E., and Rechavi, G. (2007) Altered adenosine-to-inosine RNA editing in human cancer. *Genome Research 17*, 1586–1595.
(218) Shahbazi, J., Lock, R., and Liu, T. (2013) Tumor Protein 53-Induced Nuclear Protein 1 Enhances p53 Function and Represses Tumorigenesis. *Front. Genet. 4*, 80.
(219) Seux, M., Peuget, S., Montero, M. P., Siret, C., Rigot, V., Clerc, P., Gigoux, V., Pellegrino, E., Pouyet, L., Guessan, P. N. A., Garcia, S., Dufresne, M., Iovanna, J. L., Carrier, A., eacute, F. A., and Dusetti, N. J. (2011) TP53INP1 decreases pancreatic cancer cell migration by regulating SPARC expression. *Oncogene 30*, 3049–3061.
(220) Gironella, M., Seux, M., Xie, M.-J., Cano, C., Tomasini, R., Gommeaux, J., Garcia, S., Nowak, J., Yeung, M. L., Jeang, K.-T., Chaix, A., Fazli, L., Motoo, Y., Wang, Q., Rocchi, P., Russo, A., Gleave, M., Dagorn, J.-C., Iovanna, J. L., Carrier, A., Pebusque, M.-J., and Dusetti, N. J. (2007) Tumor protein 53-induced nuclear protein 1 expression is repressed by miR-155, and its restoration inhibits pancreatic tumor development. *Proc. Natl. Acad. Sci. USA 104*, 16170–16175.
(221) Kula, A., Gharu, L., and Marcello, A. (2013) HIV-1 pre-mRNA commitment to

Rev mediated export through PSF and Matrin 3. *Virology 435*, 329–340.
(222) Kameoka, S., Duque, P., and Konarska, M. M. (2004) p54(nrb) associates with the 5' splice site within large transcription/splicing complexes. *EMBO J 23*, 1782–1791.
(223) Trapnell, C., Pachter, L., and Salzberg, S. L. (2009) TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics 25*, 1105–1111.
(224) Patton, J. G., Mayer, S. A., Tempst, P., and Nadal-Ginard, B. (1991) Characterization and molecular cloning of polypyrimidine tract-bnding protein: a component of a complex necessary for pre-mRNA splicing. *Gene. Dev. 5*, 1237–1251.
(225) Patton, J. G., Porro, E. B., Galceran, J., Tempst, P., and Nadal-Ginard, B. (1993) Cloning and characterization of PSF, a novel pre-mRNA splicing factor. *Gene. Dev. 7*, 393–406.
(226) Li, S. L., Liang, S. J., Guo, N., Wu, A. M., and Fujita-Yamaguchi, Y. (2000) Single-chain antibodies against human insulin-like growth factor I receptor: expression, purification, and effect on tumor growth. *Cancer Immunol. Immun. 49*, 243–252.
(227) Nakamura, K., Hongo, A., Kodama, J., Miyagi, Y., Yoshinouchi, M., and Kudo, T. (2000) Down-regulation of the insulin-like growth factor I receptor by antisense RNA can reverse the transformed phenotype of human cervical cancer cell lines. *Cancer Res. 60*, 760–765.
(228) Salton, M., Lerenthal, Y., Wang, S. Y., and Chen, D. J. (2010) Involvement of Matrin 3 and SFPQ/NONO in the DNA damage response. *Cell Cycle 9*, 1568–1576.
(229) Salton, M., Elkon, R., Borodina, T., Davydov, A., Yaspo, M.-L., Halperin, E., and Shiloh, Y. (2011) Matrin 3 binds and stabilizes mRNA. *PLoS ONE 6*, e23882.
(230) Tsui, S., Dai, T., Roettger, S., Schempp, W., Salido, E. C., and Yen, P. H. (2000) Identification of Two Novel Proteins That Interact with Germ-Cell-Specific RNA-Binding Proteins DAZ and DAZL1. *Genomics 65*, 266–273.
(231) Dai, T., Vera, Y., Salido, E. C., and Yen, P. H. (2001) Characterization of the mouse Dazap1 gene encoding an RNA-binding protein that interacts with infertility factors DAZ and DAZL. *BMC Genomics 2*, 6.
(232) Choudhury, R., Roy, S. G., Tsai, Y. S., Tripathy, A., Graves, L. M., and Wang, Z. (2014) The splicing activator DAZAP1 integrates splicing control into MEK/Erk-regulated cell proliferation and migration. *Nat. Comm. 5*.
(233) Smith, R. W. P., Anderson, R. C., Smith, J. W. S., Brook, M., Richardson, W. A., and Gray, N. K. (2011) DAZAP1, an RNA-binding protein required for development and spermatogenesis, can regulate mRNA translation. *RNA 17*, 1282–1295.
(234) Vera, Y., Dai, T., Hikim, A. P. S., Lue, Y., Salido, E. C., Swerdloff, R. S., and Yen, P. H. (2002) Deleted in azoospermia associated protein 1 shuttles between nucleus and cytoplasm during normal germ cell maturation. *J. Androl. 23*, 622–628.
(235) Hsu, L. C.-L., Chen, H.-Y., Lin, Y.-W., Chu, W.-C., Lin, M.-J., Yan, Y.-T., and Yen, P. H. (2008) DAZAP1, an hnRNP protein, is required for normal growth and spermatogenesis in mice. *RNA 14*, 1814–1822.
(236) Okamura, S., Arakawa, H., Tanaka, T., Nakanishi, H., Ng, C. C., Taya, Y., Monden, M., and Nakamura, Y. (2001) p53DINP1, a p53-inducible gene, regulates p53-dependent apoptosis. *Mol. Cell 8*, 85–94.
(237) Tomasini, R., Samir, A. A., Pebusque, M.-J., Calvo, E. L., Totaro, S., Dagorn, J.-C., Dusetti, N. J., and Iovanna, J. L. (2002) P53-dependent expression of the stress-induced protein (SIP). *Eur. J. Cell Biol. 81*, 294–301.
(238) Tomasin, R., Samir, A. A., Vaccaro, M. I., Pebusque, M. J., Dagorn, J. C., Iovanna, J. L., and Dusetti, N. J. (2001) Molecular and functional characterization of the stress-induced protein (SIP) gene and its two transcripts generated by alternative

splicing - SIP induced by stress and promotes cell death. *J. Biol. Chem. 276*, 44185–44192.
(239) Tomasini, R., Samir, A. A., Carrier, A., Isnardon, D., Cecchinelli, B., Soddu, S., Malissen, B., Dagorn, J. C., Iovanna, J. L., and Dusetti, N. J. (2003) TP53INP1s and Homeodomain-interacting Protein Kinase-2 (HIPK2) Are Partners in Regulating p53 Activity. *J. Biol. Chem. 278*, 37722–37729.
(240) Yoshida, K., Liu, H. S., and Miki, Y. (2006) Protein kinase C delta regulates Ser(46) phosphorylation of p53 tumor suppressor in the apoptotic response to DNA damage. *J. Biol. Chem. 281*, 5734–5740.

Appendices