The Selective Partitioning of mRNA in Biomolecular Condensates

Mark Hickling



February 2022

Abstract

Biomolecular condensates (BMCs) are complex assemblies of RNA and protein that form via the process of liquid-liquid phase separation. They have a diverse range of functionalities, often centered around RNA metabolism, and are found throughout the cell. However, how they selectively partition certain molecules whilst excluding others remains a poorly understood, yet pertinent question in biology. Here we addressed this from an mRNA perspective by developing a simplified, *in vitro* approach, utilising cytoplasmic RNA in combination with a panel of model BMCs. We identified that the length, GC %, and G-quadruplex frequency of an mRNA were critical in defining how it partitioned in these models. Remarkably, very similar rules persist in defining mRNA selection in cellular P-bodies (PBs) and stress granules (SGs). Supporting this notion, we demonstrated that the transcriptome captured by our *in vitro* model, DDX4N, was highly comparable to that of PBs. These results strongly suggest that the properties of the mRNA are the fundamental driving force behind how it partitions in BMCs. Moreover, we provide evidence for a novel role of 3' untranslated region alternative cleavage and polyadenylation (3'UTR-APA), in manipulating the length, GC %, and G-quadruplex content of mRNA in order to selectively distribute mRNA

isoforms to BMCs, with profound implications for mRNA metabolism. In addition, we demonstrated that SG and PB proteome components were significantly more likely to exhibit selective isoform partitioning within their respective condensate than the rest of the transcriptome. This raises the intriguing prospect that 3'UTR-APA may also be essential in BMC formation and maintenance.

Declaration

I declare that this thesis is entirely my own work, and except where otherwise stated, describes my own research.

Mark Hickling, Oxford, February 2022.

Acknowledgements

I would like to thank Andre Furger for all his support and guidance during my studies. When I started as a Master student, I had never envisaged that I would join the DPhil program. However, his passion for science was one of the main reasons I changed my mind, and I am very grateful for that looking back. I would also like to thank Andrew Baldwin, not only for helping me establish what has been an incredibly interesting project, but also for his generosity and encouragement, particularly during COVID-19. I would also like to thank all members, past and present, of the Furger, Mellor, and Baldwin groups, with whom I have been able to share this journey. I have always felt comfortable and able to ask for help whenever I have needed it. A special mention goes to Cillian Brophy and Harriet Lester, who have not only been inspirational, but also great friends throughout my time in Oxford. I would like to thank the States of Jersey for providing me with a bursary and enabling me to pursue my ambitions in science. As always, I would like to thank my parents and my family for all their support, without which I would never have been able to enjoy such an amazing time in my four years in Oxford.

Contents

1 Introduction **17**

1.1 Thermodynamics of Liquid-Liquid Phase Separation 19

1.1.1 Gibbs Free Energy of Mixing 19

1.1.2 Flory-Huggins Theory 20

1.2 Formation and Regulation of Biomolecular Condensates 23

1.2.1 Evidence for functional Biomolecular Condensates in Cells 24

1.2.2 Multivalency 25

1.2.3 Characteristics of Phase-Separating Proteins 26

1.2.4 The Influence of RNA Physical Properties on Condensate Formation and RNA Partitioning 32

1.2.5 Alternative Cleaveage and Polyadenylation (APA) . . . 41

1.2.6 RNA in cellular Biomolecular Condensates 43

1.2.7 Physical Properties of Biomolecular Condensates . . . 50

1.2.8 Regulation of Biomolecular Condensate Formation . . . 52

1.3 Cellular Biomolecular Condensates 53

1.3.1 P-Bodies 55

1.3.2 Stress Granules 59
1.3.3 Germ Granules 62
1.3.4 Paraspeckles 63
1.3.5 Nucleoli 64
1.3.6 Other Biomolecular Condensates 65
1.3.7 Biomolecular Condensates and Disease 67
1.4 *In vitro* Models for Studying Biomolecular Condensates 71
1.4.1 *In vitro* models 71
1.4.2 DDX4N 71
1.4.3 EWSR1C 72
1.4.4 FUS-GFP 72
1.5 Thesis Objective 73

2 Materials and Methods **74**

2.1 Materials and Methods 74
2.1.1 Cell Culture 74
2.1.2 Subcellular RNA Fractionation 74
2.1.3 Denaturing agarose RNA Gel Electrophoresis 75
2.1.4 Protein Expression 76
2.1.5 RNA Partitioning Experiments 77
2.1.6 Hanging Drop Assays 78
2.1.7 EWSR1C ThT Amyloid Assay 78
2.1.8 QuantSeq 3' mRNA Sequencing 78
2.1.9 Full RNA-seq Data Acquisition 79
2.1.10 Integrative Genomics Viewer (IGV) 79

2.1.11 Counts Table Generation 79
2.1.12 Analysis Pipelines 80
2.1.13 mRNA Sequences 81
2.1.14 *Drosophila* Orthologs 82
2.1.15 GQFR Analysis 82
2.1.16 Random Sampling Analysis 82
2.1.17 Streme Motif Discovery and Tomtom Motif Comparison 84
2.1.18 Enrichr Gene Ontology Analysis 84
2.1.19 Plots 84
2.1.20 Microscopy 85
2.1.21 Protein Constructs and Amino Acid Sequences 86

3 Investigating Selective mRNA Partitioning in Biomolecular Condensates 88

3.1 Introduction 88
3.1.1 *In vitro* mRNA partitioning approach 92
3.1.2 Establishing a pipeline to analyse gene-specific partitioning of mRNA in BMCs 95
3.2 Results 96
3.2.1 DDX4N 96
3.2.2 EWSR1C 107
3.2.3 FUS-GFP 112
3.2.4 Summary of the results from the *in vitro* model BMCs 119
3.2.5 Selective partitioning of mRNA in *in vivo* BMCs . . . 119
3.3 Discussion 130

3.4 Summary 136

4 Investigating the Role of 3'UTR-APA in Selective Partitioning of mRNA in Biomolecular Condensates 138

4.1 Introduction 138

4.1.1 Establishing pipelines to analyse 3'UTR-APA isoform-specific partitioning of mRNA in BMCs 141

4.2 Results 146

4.2.1 The role of 3'UTR-APA in selective isoform partitioning in *in vitro* model systems 146

4.2.2 Establishing the role of 3'UTR-APA in selective isoform partitioning in *in vivo* BMCs 151

4.3 Discussion 163

4.4 Summary 167

5 Discussion and Future Directions 169

5.1 Thesis Summary 169

5.2 *In vitro* approach 171

5.3 How do the physical properties of mRNA determine its selection? 172

5.3.1 A critical role for 3'UTR-APA in BMC metabolism? . 175

5.4 Conclusion 176

A Appendix 212

A.0.1 DDX4N protein gel 212

A.0.2 Correlation analysis of GQFRs vs transcript length and GC % in the human transcriptome 213

A.0.3 Bar plots displaying percentage of genes with 0 or 1 + GQFRs in each mRNA partitioning cohort, as well as the whole data set, for human and *Drosophila* data sets in DDX4N 214
A.0.4 Analysis of the physical properties of human and *Drosophila* orthologs with and without the same partitioning in DDX4N 215
A.0.5 EWSR1C protein gel 216
A.0.6 Comparison between results obtained from DESeq2 analysis using either the dilute phase or SW480 BAM files as the reference for DDX4N and FUS-GFP 217
A.0.7 Preliminary characterisation of EWSR1C 218
A.0.8 Bar plots displaying percentage of genes with 0 or 1 + GQFRs in each mRNA partitioning cohort, as well as the whole data set, for, FUS-GFP, P-Bodies and Stress Granules 224
A.0.9 Analysis of the length and GC % of the 5'UTR, CDS, and 3'UTR in mRNA partitioning cohorts in all investigated BMCs 225
A.0.10 The length dependency of selective partitioning of human mRNA in DDX4N as a function of its GC % . . . 228
A.0.11 Example IGV and corresponding expression plots depicting two genes with long isoform enriched 3'UTR-APA changes in DDX4N from the human data set . . . 228

List of Tables

1.1 Key PB protein examples and their functions (D'Lima et al. 2017; Kamenska et al. 2014; D. Zhou and S. Chen 2001; Blasco-Moreno et al. 2019; Vilela et al. 2000; Hernández et al. 2018; Rehwinkel et al. 2005; Denis and J. Chen 2003; Scheller et al. 2007; Ingelfinger et al. 2002; Smillie and Sommerville 2002; Ying Li et al. 2012; Fromm et al. 2012) 57

1.2 Key SG protein examples and their functions (Smillie and Sommerville 2002; Shurtleff et al. 2017; Fonseca et al. 2015; Behm-Ansmant et al. 2007; D. Han et al. 2019; A. R. Morris, Mukherjee, and Keene 2008; Singleton et al. 2014; French et al. 2002; Q. Wang et al. 2019; Ayala et al. 2008) 61

1.3 Table displaying phase-separating RNA-binding proteins and their associated diseases (Kwiatkowski et al. 2009; Couthouis et al. 2012; Ian RA Mackenzie and Rademakers 2008; Ian RA Mackenzie and M. Neumann 2012; H. J. Kim et al. 2013; Y. Lee et al. 2018) 68

2.1 Protein constructs and sequences 87

List of Figures

1.1 Phase separation behaviour of a polymer in solution. 22
1.2 Evidence for P-granules being phase separated compartments with liquid-like properties 25
1.3 Phase Separation in the Actin-Regulatory Signalling Pathway is Governed by Multivalent Interactions between Protein Modular Domains 27
1.4 Molecular Interactions that Promote Phase Separation 29
1.5 Regulation of NELF condensation via post-translational modifications 32
1.6 RNA-RNA interactions involved in promoting phase separation 34
1.7 RNA Structural Elements 37
1.8 Possible Mechanisms via which RNA Post-Transcriptional Modifications can Influence Condensate Formation 41
1.9 3'UTR Alternative Cleavage and Polyadenylation 42
1.10 Protein-RNA Interactions in Ribonucleoprotein Granules . . . 44
1.11 The driving forces behind RNA selection in biomolecular condensates 45

1.12 Simple Model of Biomolecular Condensate Composition Regulation 48
1.13 Regulation of Liquid-to-Solid Transition in Biomolecular Condensates 51
1.14 Examples of biomolecular condensates found throughout the cell 55
1.15 P-Body transcriptome is influenced by mRNA physical properties 58
1.16 Stress granule transcriptome is influenced by mRNA physical properties 60
1.17 Liquid-to-solid phase transition of the RNA-binding protein FUS *in vitro* 69

3.1 Model BMCs used for *in vitro* mRNA partitioning assays . . . 92
3.2 Schematic showing the mRNA partitioning approach 94
3.3 99
3.3 Selective partitioning of mRNA in DDX4N 100
3.4 Gene ontology analysis of mRNA partitioning cohorts in DDX4N 104
3.5 Selective partitioning of orthologous human and *Drosophila* mRNA in DDX4N 106
3.6 Selective mRNA partitioning in EWSR1C 111
3.7 115
3.7 Selective partitioning of mRNA in FUS-GFP 116
3.8 Gene ontology analysis of mRNA partitioning cohorts in FUS-GFP 118

3.8 Selective partitioning of mRNA in BMCs 123
3.9 125
3.9 126
3.9 GQFR and GO analysis of mRNA partitioning cohorts in PBs and SGs, and correlation analysis of selective mRNA partitioning in all BMCs 127
3.10 Correlation analysis of the lengths and GC % of the 5'UTR, CDS, and 3'UTR with vs the full mRNA transcript in the human transcriptome 130
3.11 Diagrammatic overview of mRNA selection trends in BMCs . 137

4.1 142
4.1 3'UTR-APA, sequencing methods and subsequent 3'UTR-APA analysis approaches 143
4.2 149
4.2 3'UTR-APA analysis in DDX4N 150
4.3 152
4.3 3'UTR-APA analysis of *in vivo* BMCs 153
4.4 GO Analysis of genes with significant 3'UTR-APA changes in SGs and PBs 156
4.5 Comparison of 3'UTR-APA in PBs and SGs 158
4.6 161
4.6 Analysis of genes with 3'UTR-APA changes in PBs and SGs . 162

A.1 DDX4N final protein gel 212

A.2 Correlation analysis of GQFRs vs transcript length and GC % in the human transcriptome 213
A.3 Bar plots displaying percentage of genes with 0 or 1 + GQFRs in each mRNA partitioning cohort, as well as the whole data set, for human and *Drosophila* data sets in DDX4N 214
A.4 Pearson correlation analysis of the physical properties of human and *Drosophila* orthologs with and without the same partitioning in DDX4N 215
A.5 EWSR1C final protein gel 216
A.6 217
A.6 Comparison between results obtained from DESeq2 analysis using either the dilute phase or SW480 BAM files as the reference for DDX4N and FUS-GFP 218
A.7 219
A.7 DIC microscopy analysis of EWSR1C phase-separation assays 220
A.8 222
A.8 Amyloid detection assay with ThT using EWSR1C and cytoplasmic RNA 223
A.9 224
A.9 Bar plots displaying percentage of genes with 0 or 1 + GQFRs in each mRNA partitioning cohort, as well as the whole data set, for FUS-GFP, P-Bodies and Stress Granules 225
A.10 Boxplots depicting the relationship between the length of the 5'UTR, CDS and 3'UTR in mRNA partitioning cohorts in DDX4N, EWSR1C, FUS-GFP, PBs and SGs. 226

A.11 Boxplots depicting the relationship between the GC % of the 5'UTR, CDS and 3'UTR in mRNA partitioning cohorts in DDX4N, EWSR1C, FUS-GFP, PBs and SGs. 227

A.12 Boxplots depicting the transcript length for each of the partitioning cohorts in the human data set, split into two groups by GC %. 228

A.13 Example IGV and corresponding expression plots depicting two genes with long isoform enriched 3'UTR-APA changes in DDX4N from the human data set 229

Chapter 1

Introduction

Spatial organisation of molecules is a fundamental characteristic of living cells. Often biochemical reactions require mutually exclusive conditions which can be achieved through the use of distinct subcellular compartments. The most prominent compartments are the organelles, such as mitochondria, endoplasmic reticulum or the Golgi apparatus, which are encapsulated by a largely impermeable phospholipid bilayer. In these organelles, the internal environment is physically separated from the cytoplasm, and its composition is regulated through the use of selective membrane transport. In addition to membrane-enclosed organelles, eukaryotic cells also contain a large variety of conserved, membraneless structures known as biomolecular condensates (BMCs). BMCs form via the mechanism of liquid-liquid phase separation, a reversible process that is intimately in tune with changes in the subcellular environment. It provides an alternative means for the cell to compartmentalise and concentrate molecules, and provide solvent conditions that favour specific biological processes. For example, as part of the innate immune

response, activity of cyclic guanosine monophosphate–adenosine monophosphate (cGAMP) synthase (cGAS) is greatly enhanced upon phase-separation, owing to changes in the cGAS catalytic site (Xu Zhang et al. 2014). Accordingly, mutations to cGAS that abrogated the enzymes ability to form condensates dramatically reduced it's activity (Du and Z. J. Chen 2018). Moreover, a key feature of BMCs is their specific composition, implying that they are able to partition some components whilst keeping others out (Banani, H. O. Lee, et al. 2017; Huaiying Zhang et al. 2015). Examples of these include stress granules (SGs), P-bodies (PBs), and germ granules in the cytoplasm, and Cajal bodies, paraspeckles and the nucleoli in the nucleus (Buchan and Parker 2009; Brangwynne, Mitchison, and Hyman 2011; Gall 2003; Santiago Ramón y Cajal 1910; S Ramón Cajal et al. 1903). A large number of BMCs contain, and are selective for, specific RNA molecules, and are directly involved in RNA metabolism (S. Tian, Curnutte, and Trcek 2020). Many are also enriched in RNA-binding proteins (RBPs) and are consequently known as ribonucleoprotein (RNP) granules (Banani, Rice, et al. 2016; Yamazaki et al. 2018). As a result, they are now considered an integral part of gene expression regulation. RNA itself has the intrinsic ability to regulate condensate formation, fine tune their physical properties, and in some cases serve to prevent pathological aggregation associated with diseases such as ALS (Maharana et al. 2018; H. J. Kim et al. 2013). Structural elements encoded within RNA sequences have been shown to enhance condensate formation, and can influence how an RNA molecule is partitioned into them (Poudyal et al. 2021; Huaiying Zhang et al. 2015; Begovich and Wilhelm 2020). In spite of this, there remain many unanswered questions as to why certain

RNAs may be enriched in a specific condensate, whilst others are excluded, and the driving forces that dictate this selection remain poorly understood. Resolving these questions is of critical importance in to the understanding of RNA metabolism and BMC function in general.

1.1 Thermodynamics of Liquid-Liquid Phase Separation

1.1.1 Gibbs Free Energy of Mixing

Liquid-liquid phase separation is a thermodynamically driven process that results in the demixing of a homogeneous solution into two distinct liquid phases. The tendency of this process to occur is governed by the Gibbs free energy of mixing displayed in equation 1.1; where ΔG_{mix} is the free energy change of mixing, ΔH_{mix} is the enthalpy change of mixing, T is the temperature and ΔS_{mix} is the entropy change of mixing:

$$G_{mixed} - G_{unmixed} = \Delta G_{mix} = \Delta H_{mix} - T\Delta S_{mix} \tag{1.1}$$

Since the mixing of two species results in an increase in entropy, in a system where there is no interaction between the molecules from each of the different species (i.e. $\Delta H_{mix} = 0$), a single, homogeneous phase is always favourable. However, in some circumstances, a dual-phase solution may become favourable. This is classically illustrated by the example of oil and vinegar. In simplistic terms, this can be rationalised by energetically-

favourable homotypic interactions between like-molecules outweighing the entropic deficit caused by demixing, resulting in an overall reduction in the free energy relative to the single phase system (Hyman, Weber, and Jülicher 2014). In other words, the entropic term is overcome by the enthalpic term in the equation.

1.1.2 Flory-Huggins Theory

The Flory-Huggins model is an embellishment of the Gibbs free energy of mixing equation (equation 1.1) tailored specifically to the thermodynamics of polymers in solution, where new terms for the enthalpy, ΔH_{mix}, and entropy, ΔS_{mix}, are established (Flory 1941; Huggins 1941). Despite its simplicity it has been widely applied as a model for characterising biological phase separation (C. F. Lee et al. 2013; Nott, Petsalaki, et al. 2015). The model considers a system containing two types of molecule; a homopolymer of length N, and a solvent, with the solution treated as an infinite lattice with coordination number z (Figure 1.1A). Each site in the lattice is occupied exclusively by either a solvent molecule or a monomeric subunit of the polymer. For each of these sites the Flory-Huggins free energy of mixing is defined by the equation:

$$\frac{F}{k_B T} = \frac{\phi}{N}\ln\phi + (1-\phi)\ln(1-\phi) + \chi\phi(1-\phi) \tag{1.2}$$

The volume fraction of the polymer, ϕ_p, plus the volume fraction of the solvent, ϕ_s, is equal to 1, meaning ϕ_s can be defined as $1 - \phi_p$. ϕ_p is then treated as ϕ. χ is the Flory-Huggins interaction constant and accounts

for polymer-solvent interactions, u_{ps}, polymer-polymer interactions, u_{pp} and solvent-solvent interactions, u_{ss}. The constant is defined in equation 1.3 below:

$$\chi = \frac{z}{k_B T}\left[u_{ps} - \frac{1}{2}(u_{pp} + u_{ss})\right] \tag{1.3}$$

As the interactions between polymer and solvent become increasingly unfavourable (i.e. the worse the solvent is), the value of χ increases, as does the free energy of mixing. Eventually the magnitude of χ becomes large enough that the enthalpic term, $\chi\phi(1-\phi)$, outweighs the entropic terms, $\frac{\phi}{N}\ln\phi + (1-\phi)\ln(1-\phi)$ in equation 1.2 resulting in a region of negative curvature, indicated by the coloured lines in figure 1.1B. This corresponds to an increasing instability of the single phase solution which is rectified by separation into two phases, one polymer-rich and the other polymer-poor as shown in figure 1.1C (Brangwynne, P. Tompa, and Pappu 2015).

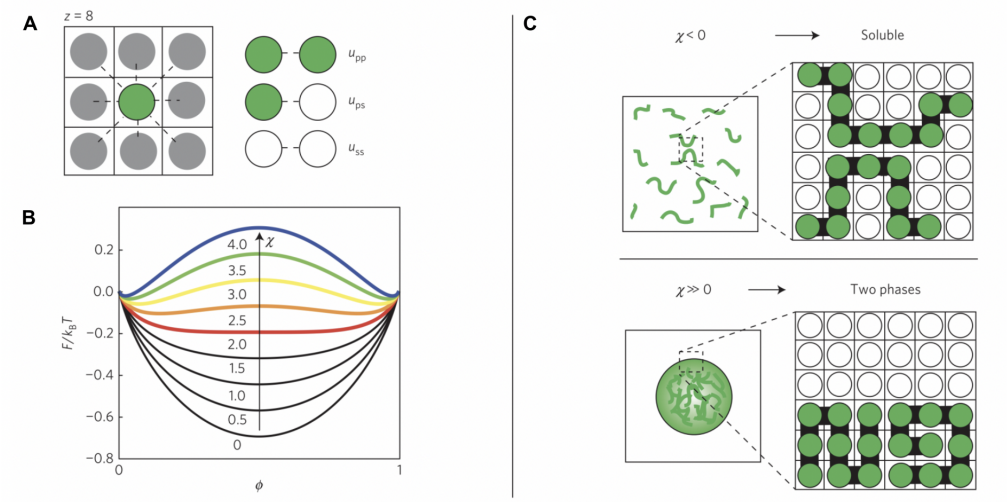


Figure 1.1: Phase separation behaviour of a polymer in solution. A) Illustration of the interaction of a lattice site with surrounding neighbours. Here z is the lattice coordination number, whilst u_{pp}, u_{ps} and u_{ss} represent the interactions between polymer and polymer, polymer and solvent, and solvent and solvent respectively. B) Plot showing the effect increasing χ, which accounts for all the interactions in a given mixture, has on the free energy of a single-phase solution. As χ increases, the single phase solution becomes increasingly unfavourable. C) Diagram showing how χ influences the phase behaviour of a polymer in solution. When χ is small, a single phase solution is favourable, as depicted in the upper panel. As χ becomes large, a single-phase solution becomes unfavourable leading to formation of a dual-phase solution as depicted in the lower panel. Figure adapted from (Brangwynne, P. Tompa, and Pappu 2015)

Whilst further complexities can be added to the model, for example, to account for screened interactions between polyions using the Debye-Hückel theory (Debye and Hückel 1923; Overbeek and Voorn 1957), clearly in biological systems the reality is far more complicated. Despite this, Flory-Huggins theory can serves as a very useful first insight (Zeigler et al. 2021).

1.2 Formation and Regulation of Biomolecular Condensates

The assembly and disassembly of BMCs is under strict spatiotemporal regulation, finely in tune with changes in the cell's environment. Condensates such as SGs only form in response to cellular stress, and it is postulated their assembly serves to promote cell survival under these conditions (Mahboubi and Stochaj 2017; Arimoto et al. 2008; Panas, Ivanov, and P. Anderson 2016). Whilst their precise role in this remains the subject of debate, it is thought that this may in part be due to their sequestration of apoptotic-signalling molecules such as TRAF2 and RACK1 (Arimoto et al. 2008; W. J. Kim et al. 2005; Y.-J. Park et al. 2020). We are just beginning to understand some of the mechanisms underpinning condensate formation, and regulation, as well as the characteristics of the biomolecules that are typically involved. These are critical first steps towards fully elucidating their function.

Molecules, such as proteins with intrinsically disordered regions (IDRs) or RNA, which have the ability to form weak, transient, multivalent interactions, are often key components. Sequence elements such as amino acid motifs in proteins or specific structural elements in RNA can mediate these interactions (Hyman, Weber, and Jülicher 2014). Importantly, cellular mechanisms, such as alternative cleavage and polyadenylation (APA), alternative splicing, base modifications and post-translational modifications provide means to change the physical properties of RNA and proteins and make this a dynamic process. This provides cells with scope to fine-tune the phase separating propensity of these molecules, and align BMC formation

with cellular physiology and states.

1.2.1 Evidence for functional Biomolecular Condensates in Cells

Evidence for non membrane-bound compartments in cells has been around since the 1830's when the nucleolus was first observed inside the nuclei of neuronal cells (Pederson 2011). However, it wasn't until 2009 that Brangwynne, Jülicher, and Hyman provided proof that P-Granules (membraneless, perinuclear compartments comprised of protein and RNA) in the embryos of *Caenohabditis elegans* have liquid-like properties, and form via phase separation (Brangwynne, Eckmann, et al. 2009). They found that these granules displayed four key signatures of liquid behavior over timescales of seconds:

1. Upon contact with other granules, they fuse before returning to a spherical conformation - figure 1.2B
2. They deform under shear stress - figure 1.2D
3. They are spherical.
4. They recover following photobleaching via internal rearrangement - figure 1.2D

This finding has prompted many more phase-separated structures displaying liquid-like properties to be discovered, including SGs, nucleoli, and PBs (Sheth and Parker 2003; Brangwynne, Mitchison, and Hyman 2011; Nover, Scharf, and D. Neumann 1989). Since then a large amount of work

has gone into understanding how BMCs form, a vital first step towards fully understanding their biological roles.

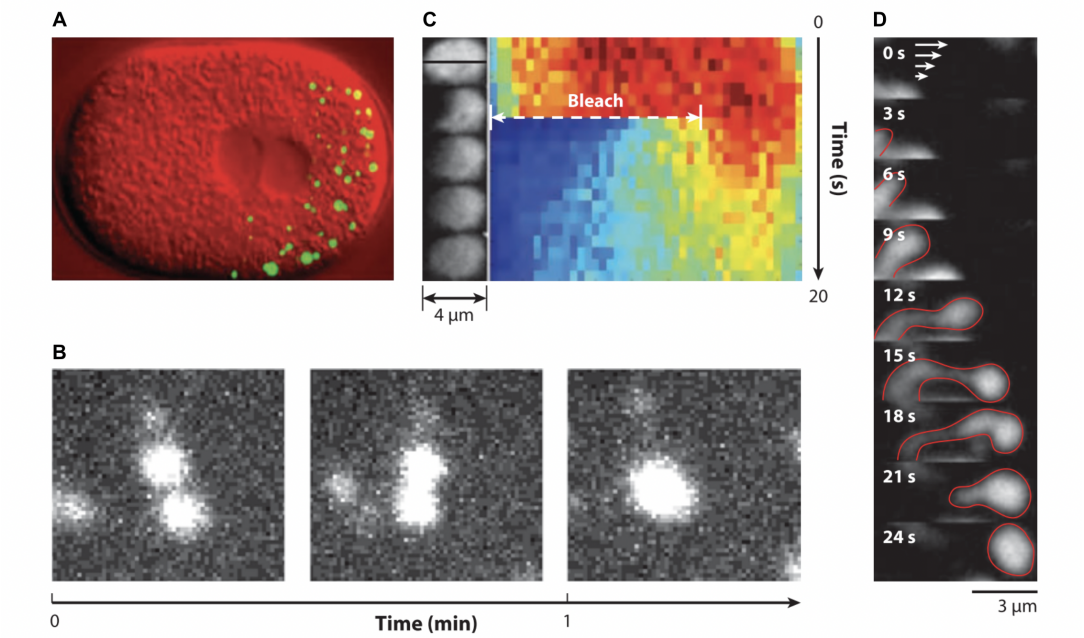


Figure 1.2: Evidence for P-granules being phase separated compartments with liquid-like properties. A) Fluorescently-tagged P Granules localising to the posterior of a *C.elegans embryo*. B) Two P-granules fusing, demonstrating liquid-like properties. C) Recovery of a P-granule from photobleaching, providing evidence for free diffusion of molecules within them, again demonstrating liquid-like properties. D) P-Granule deformed by shear flow, as further evidence that they are liquid-like. Figure adapted from (Hyman, Weber, and Jülicher 2014)

1.2.2 Multivalency

A fundamental characteristic of molecules enriched within BMCs is multivalency (P. Li et al. 2012; O. D. King, Gitler, and Shorter 2012; Nott, Petsalaki, et al. 2015). Multivalent molecules harbour multiple motifs that modulate both inter- and intra-molecular interactions. Multivalent interactions lead to

a dual-phase solution becoming energetically favourable once a critical concentration of components has been reached. Upon mixing, these molecules polymerise, decreasing their overall solubility owing to entropy-driven effects, promoting phase separation (Flory 1953). *In vivo* a diverse range of these molecules exist including proteins with multiple modular interaction domains, proteins that contain low-complexity IDRs and RNA (Banani, H. O. Lee, et al. 2017).

1.2.3 Characteristics of Phase-Separating Proteins

Modular Domains

One of the ways proteins are able to form multivalent contacts is through interactions between modular binding domains and cognate ligands. Here phase separation is promoted by stereospecific interactions between folded domains. The first well characterised example of a system that utilised this mechanism was the actin-regulatory signalling pathway (P. Li et al. 2012; Banjade and Rosen 2014; Zeigler et al. 2021). Here three proteins, neural Wiskott-Aldrich syndrome protein (N-WASP), nephrin and nck, come together to form high-order oligomers, producing phase-separated liquid-droplets, mediated by interactions between various modular domains (Figure 1.3). Liquid-liquid phase separation driven by dimerisation of G3BP1 leading to SG assembly, and actin organisation in T-cells are also governed by this process (Tourrière et al. 2003; Su, Ditlev, Hui, et al. 2016). In the case of G3BP1, it is postulated that it's dimerisation leads to an increase in valence that is essential for condensate formation (Guillén-Boixet et al.

2020). For the "survival of motor neuron protein" (SMN), it was shown that phase separation was induced by the globular tudor domain, as opposed to it's IDR (Courchaine et al. 2021). In this example, this domain binds to dimethylarginine or DMA, promoting condensate formation.

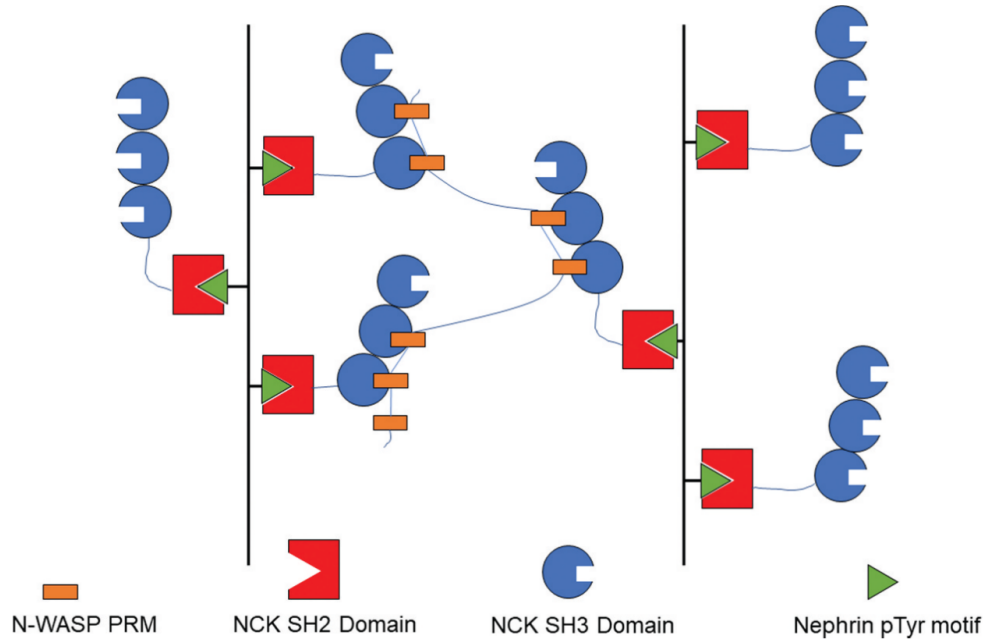


Figure 1.3: Phase Separation in the Actin-Regulatory Signalling Pathway is Governed by Multivalent Interactions between Protein Modular Domains. Here the NCK protein has two different types of modular domain. The first, is a singular SH2 domain, which binds to a proline rich motif (PRM) on the N-wasp protein. Secondly, it contains three SH3 domains which are able to bind phosphorylated tyrosine (pTyr) motifs on nephrin. These interactions promote the assembly of multimeric complexes which then undergo condensation. Figure from (Zeigler et al. 2021)

Intrinsically Disordered Regions

IDRs serve as another way for proteins to achieve multivalency. Indeed, they are enriched in phase-separating proteins, often located at the N- or C-

terminus (Brangwynne, P. Tompa, and Pappu 2015; Michelitsch and Weissman 2000; Castello et al. 2012; Uversky 2013). Many experiments have determined IDRs to be necessary and sufficient for phase separation both *in vitro* and *in vivo* (Castello et al. 2012; Conicella et al. 2016; Nott, Petsalaki, et al. 2015). IDRs are characterised as regions that are unable to fold spontaneously, instead rapidly fluctuating through a range of conformations. The sequence of IDRs tends to be of low complexity, often containing short-linear motifs (SLiMs) that modulate inter- and intra-molecular interactions, and promote phase separation. IDRs primarily consist of hydrophilic amino acids over bulky, hydrophobic ones; the most common amino acid constituents being glycine, glutamine, phenylalanine, tyrosine, asparagine, and serine (P. E. Wright and Dyson 2015). Some IDRs also contain a high proportion of charged residues - lysine, glutamate, aspartate and arginine (O. D. King, Gitler, and Shorter 2012; Nott, Petsalaki, et al. 2015; Reijns et al. 2008). Interactions that contribute to the phase-separating propensity of proteins containing IDRs include charge-charge, cation-π, dipole-dipole, and $\pi - \pi$ stacking (Figure 1.4A).

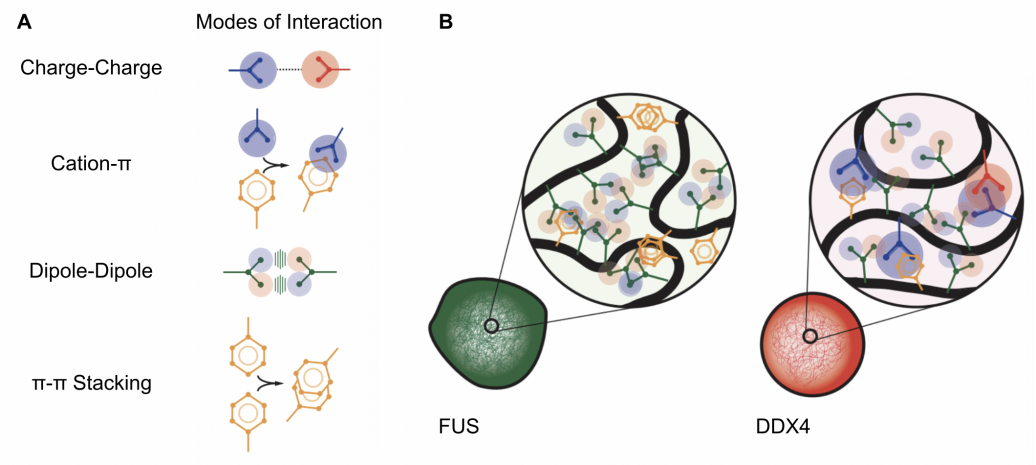


Figure 1.4: Molecular Interactions that Promote Phase Separation A) Modes of interaction that can occur between amino acid residues. B) Interactions predicted to dominate in the dense phase for FUS and DDX4 proteins. Figure adapted from (Brangwynne, P. Tompa, and Pappu 2015)

Much of our understanding in this area has come from *in vitro* screening of individual proteins via mutagenesis, to identify specific residues or sequence motifs essential for phase separation. In the case of the RBP fused in sarcoma (FUS), cation-π interactions between tyrosine residues in the so-called, prion-like domain (PLD) (given this name due to its sequence homology with the yeast prion protein) and arginine residues in the RNA-binding domain (RBD) dominate. Additional favourable interactions involve associations between aromatic residues ($\pi - \pi$ stacking) and oppositely charged residues (charge-charge), while the material properties of the droplets are controlled mainly by glycines, glutamines, and serines, through dipole-dipole interactions (Jie Wang et al. 2018). For the disordered N-terminal region of the DEAD-box helicase protein, DDX4, condensation into droplets is determined mainly by cation-π interactions between phenylalanines and arginines. These results were evidenced by methylation of arginine residues, and mu-

tation/fluorination of phenylalanine residues, which all led to a reduction in droplet stability. In addition to this, it was shown that the alternating blocks of positive and negative charge, found throughout the chain, are also essential for droplet formation (Nott, Petsalaki, et al. 2015). An overview of the interactions described for these two proteins is depicted in Figure 1.4B.

Post-translational Modification of Intrinsically Disordered Regions

Since IDRs often harbour elements that are essential for phase separation, it becomes clear how post-translational modifications, may be used to regulate this process in cells. Indeed, IDRs are enriched in post-translational modifications relative to the rest of the proteome (Iakoucheva et al. 2004; Collins et al. 2008). Post-translational modifications can cause a range of alterations including changing the overall charge, hydrophobicity, and structure of IDRs, through the addition of functional groups such as phosphoryl (Qiyao Li et al. 2017). Classically, they are thought of as on/off switches regulating events like protein signalling, localisation and degradation. In a similar vein, modification of IDRs can serve as a switch for phase-separation, through modulating inter- and intra-molecular interactions. FUS has two putative phosphorylation sites in its N-terminal PLD that regulate its condensation. Phosphorylation of these sites in FUS introduces increased repulsive electrostatic interactions that impairs its ability to form condensates (Monahan et al. 2017). The opposite is true for the protein Tau441, where phosphorylation of an IDR is necessary to stimulate phase-separation (Boyko et al. 2019). Negative elongation factor (NELF) is another protein which harbours multiple post-translational modification sites within it's two IDRs. Upon

conditions of cellular stress, it has been shown that there is a significant decrease in phosphorylation of residues within these IDRs. In addition, these regions then become highly sumoylated by the SUMO E3 ligase. Both of these modifications together are necessary for, and stimulate, the formation of NELF condensates. This process is postulated to sterically prevent interactions between elongation factors, such as P-TEFb, and RNA Pol II, and can thereby serve to downregulate transcription in response to stress (Rawat et al. 2021)(Figure 1.5). This example is a compelling demonstration of how the biological function of a protein can be directly impacted via dynamic regulation of its propensity to phase separate through post-translational modifications.

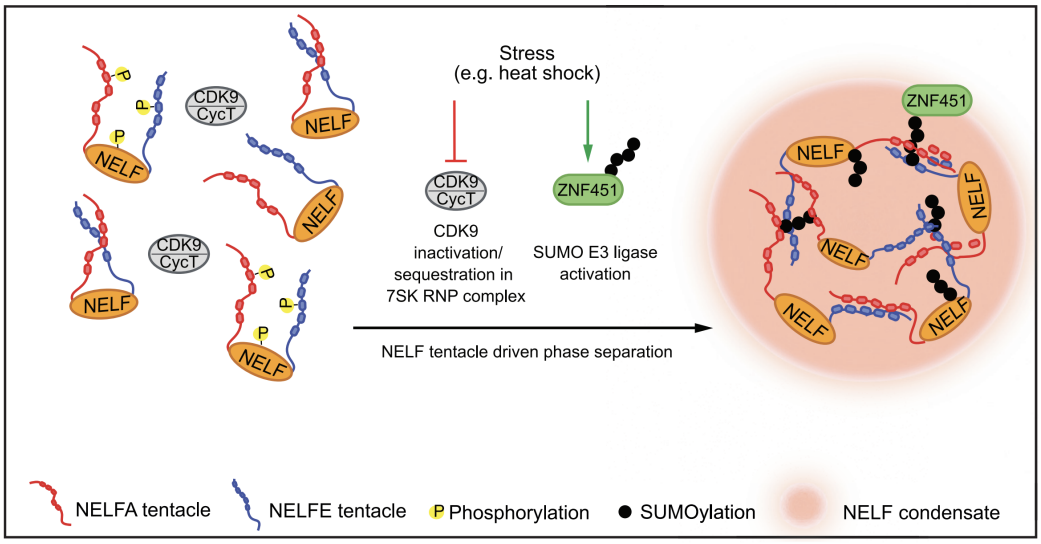


Figure 1.5: Regulation of NELF condensation via post-translational modifications. Under conditions of stress, phosphorylation of the NELFA and NELFB intrinsically disordered regions (tentacles) by CDK9 is inhibited by it's sequestration in the 7SK RNP complex. In addition, SUMO E3 ligase activation leads to sumoylation of these tentacles. Together both modifications serve to induce condensate formation of the NELF protein, which is postulated to induce transcriptional downregulation. Figure from (Rawat et al. 2021)

1.2.4 The Influence of RNA Physical Properties on Condensate Formation and RNA Partitioning

To date the majority of research into BMCs has centred on the role of their protein constituents. This is in spite of the fact that many are enriched in RNA (and indeed RBPs), have roles involving RNA homeostasis, and in some cases require RNA to form (Banani, Rice, et al. 2016; Yamazaki et al. 2018; Decker, J. M. Burke, et al. 2021). It is thus imperative to try to understand the role of RNA has in regulating condensate formation, properties, composition and function.

RNA-RNA Interactions

RNA is an anionic polymer of low sequence complexity that serves as an excellent platform for achieving multivalency. *In vitro*, RNA in the presence of crowding agents can undergo robust self-assembly in the absence of protein (Aumiller Jr et al. 2016; Van Treeck, Protter, et al. 2018). Moreover, *in vitro*, yeast RNA was found to self-assemble into condensates that have an RNA content that resembles that of yeast SGs *in vivo* (Van Treeck, Protter, et al. 2018). These observations suggest that RNA-RNA inter- and intra-molecular interactions may be key driving forces behind RNP granule formation, providing insight into how condensates may achieve selectivity (Van Treeck and Parker 2018).

RNA-RNA interactions can be promiscuous (Figure 1.6A), or specific (Figure 1.6B), and occur through Watson-Crick base-pairing, non-canonical base-pairing and helical stacking. Promiscuous interactions occur at random between RNA molecules. SGs are enriched in long RNA transcripts and it is hypothesised that this may be due to longer transcripts being able to partake in more promiscuous, multivalent interactions than their shorter counterparts. Interestingly, despite this increase in length, SG RNAs actually only showed a modest increase in RNA-binding sites of SG proteins, further highlighting the potential of RNA-RNA interactions as a key driving force behind SG formation (Khong et al. 2017). Additional evidence supporting a key role for RNA-RNA interactions is the high salt and aliphatic alcohol resistance associated with SG condensate cores (S. Jain et al. 2016). If the stability of SG cores was conferred largely by protein-protein interactions

exposure of SGs to such treatments would force their dissolution.

Specific base-pairing between mRNAs is also known to affect mRNA partitioning into condensates and formation of RNP granules. During *Drosophila* oocyte development, *Oskar* and *bicoid* transcripts form homodimers, a process which is essential for their recruitment to RNP granules (Ferrandon et al. 1997; Jambor, Brunel, and Ephrussi 2011). Van Treeck *et al.* suggest another example for how specific interactions can lead to selective RNA partitioning into condensates, by showing that certain antisense non-coding RNAs tended to colocalise to SGs, along with their corresponding mRNA transcripts. The implication of this is that base-pairing between the two RNAs influences their localisation.

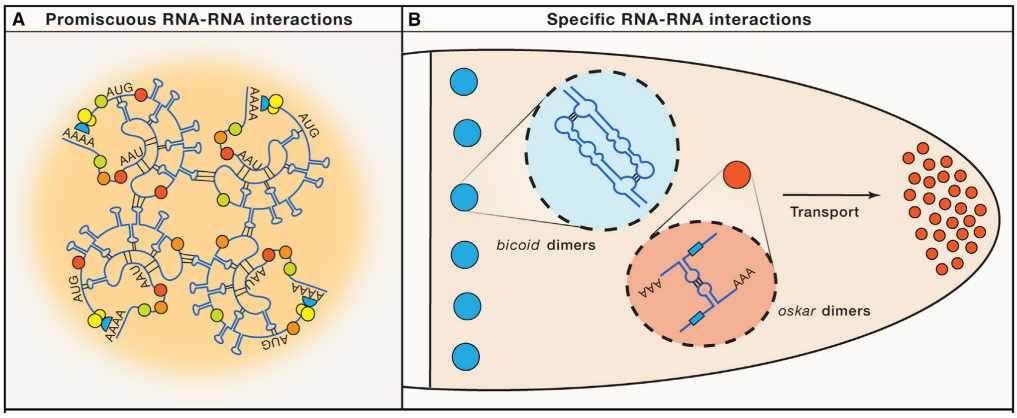


Figure 1.6: RNA-RNA interactions involved in promoting phase separation. A) Promiscuous RNA-RNA interactions between molecules essentially occurring at random. B) Specific RNA-RNA interactions targeted to occur between specific RNA molecules. The blue and red circles represent condensates formed via interactions between *bicoid* dimers and *oskar* dimers respectively. Figure adapted from (Van Treeck and Parker 2018).

RNA primary Sequence and its consequence for phase separation

RNA is composed of four different nucleotides; adenine (A), cytosine (C), gunanine (G) and uracil (U). A and G, the purines, contain two aromatic rings, while C and U, the pyrimidines, contain only one. The stacking free energies of the nucleotides decreases in the order G/A/U/C, owing to the number of aromatic rings (Brown, Andrews, and Elcock 2018). Lu *et al.* investigated how the properties of condensates formed from each of these individual nucleotides would differ, based on the trends in stacking free energies (T. Lu, Nakashima, and Spruijt 2021). They found that the temperature-based stability of the condensates decreased with decreasing stacking free energies of the nucleotides, highlighting how nucleotide composition and accordingly the strength of favourable π-stacking interactions, can influence the properties of BMCs. Coacervates formed from homotypic RNA polymers also have different properties depending on which nucleoside they contain. Poly(A/C/G) RNAs form liquid-like droplets with proline-arginine repeat peptides, whereas polyG RNAs form more gel-like structures, owing to G-quadruplex (GQ) formation. Additionally, poly(A) condensates are more viscous than the equivalent poly(C) or poly(U) condensates (Boeynaems et al. 2019). Moreover, mixtures of poly(A) and poly(U) transcripts will heterotypically coassemble *in vitro* whereas poly(A) and poly(C) transcripts de-mix into immiscible, homotypic droplets thus indicating that RNAs that base-pair with each other are more likely to be co-recruited to liquid droplets (Tauber et al. 2020). RNA polymers *in vivo* are a complex composition of different nucleotides, but its evident from these results that RNA partitioning into BMCs can be

influenced by the composition of the RNA sequence alone.

RNA Secondary Structure

The secondary structure of an RNA is another property that can influence how it interacts with other RNA molecules and proteins. RNA can adopt many structural conformations including hairpin stem-loops, psuedoknots and i-Motifs (Rietveld et al. 1982; Dzatko et al. 2018) (Figure 1.7A). The specific structure an RNA molecule exhibits is dynamic, and can be influenced by changes in the cell's environment or interactions with other protein or RNA molecules. Often a single RNA molecule has many accessible 2D and 3D structures with similar folding free energies (Drino and Schaefer 2018; Sponer et al. 2018; Lorenz et al. 2016). Langdon *et al.* showed that modifying the sequence and therefore, secondary structure of *CLN3* mRNA altered its recruitment to Whi3-*BNI1* condensates both *in vitro*, and *in vivo* (Langdon et al. 2018). Structural analysis of the mRNA revealed that regulation may be achieved through a stem-loop forming region which can selectively display or mask sequences that promote hybridisation with other mRNAs.

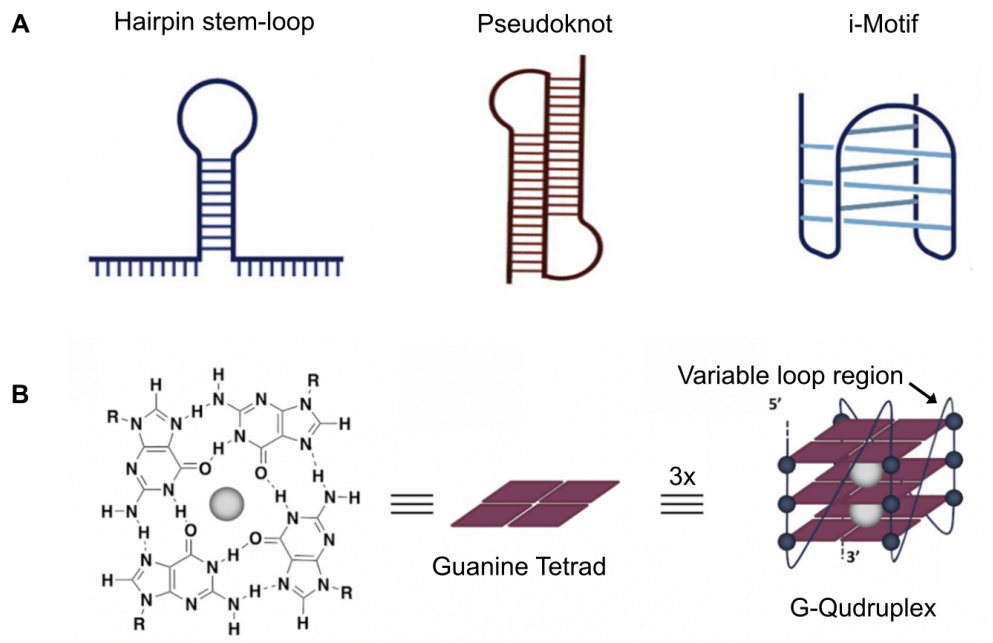


Figure 1.7: RNA Structural Elements. A) Depictions of three common RNA secondary structures; hairpin stem-loops, pseudoknots and i-motifs. B) G-quadruplex formation. Four guanine bases associate via Hoogsteen hydrogen bonding to form a guanine tetrad, stabilised through coordination of a metal ion (grey sphere). Tetrads then stack to form the G-quadruplex structure. Figure adapted from (Tassinari, S. N. Richter, and Gandellini 2021)

Moreover, RNA secondary structure provides a mechanism by which SARS CoV-2 is able to selectively partition its own gRNA from a very crowded cellular RNA environment. The 30 kb SARS CoV-2 gRNA contains specific structural elements that the nucleocapsid protein is able to recognise, and requires in order to undergo phase separation. This process allows for its selective packaging into a virion, with exclusion of host RNA molecules (Iserman et al. 2020). For PGL-3, a main constituent of *C. elegans* P granules, phase separation was shown to be induced in the presence of mRNA but not in the presence of ribosomal RNA (rRNA). rRNA is highly structured and it's hypothesised that the intramolecular base-pairing that

causes this interferes with it's ability to form the intermolecular multivalent contacts required for condensate formation. When the rRNA was denatured by heating however, droplets were then able to form (Saha et al. 2016). More recently, a study using small bacterial riboswitch RNAs showed condensate formation was dependent upon intermolecular base-pairing and subsequent multimerisation. In particular, they found specific palindromic loop regions that were responsible for this, that would form intermolecular loops and drive the formation of higher order RNA oligomers, promoting condensation. In this way, they provide a mechanism by which specific structural elements in RNA, formed intermolecularly, contribute to both condensate formation, and which RNA molecules are selectively partitioned (Poudyal et al. 2021). Moreover, specific stem-loop RNAs have the capacity to trigger formation of SG-like condensates *in vitro*, substantiating the role of RNA-RNA interactions in BMC formation (Begovich and Wilhelm 2020). However, this area still remains relatively unexplored, and there is a lot more to understand in order to fully elucidate the role RNA secondary structure has on condensate formation and RNA selection. Indeed, most of the work done to date has focussed on a few RNA molecules in isolation, often using short sequences. This leaves many open questions regarding its influence in a crowded cellular environment consisting of large numbers of heterotypically structured RNA molecules.

RNA G-Quadruplexes

GQs are structures found in DNA and RNA, that consist of multiple guanine tetrads stacked on top of each other. Each guanine tetrad is formed

through Hoogsteen hydrogen bonding, with a metal ion coordinated at the centre (usually sodium or potassium) (Millevoi, Moine, and Vagner 2012). They are able to form inter- or intra-molecularly, and have a diverse range of biological roles, including regulation of transcription and translation (1.7B) (Huppert and Balasubramanian 2007; Rhodes and Lipps 2015). Now, increasing amounts of evidence suggests that they may be integral to the formation and regulation of certain BMCs. mRNAs containing GQs have been shown to phase-separate *in vitro* in the absence of protein, as is the case with *Arabidopsis SHR* mRNA (Yueying Zhang et al. 2019). Phase separation of the *SHR* mRNA was dependent on containing a GQ, and the stability of the droplet phase could be altered by changing the GQ structure. GQs with longer loop regions (arrow in 1.7B) and more guanine tetrads produced more stable droplets, perhaps due to increased multivalency (Yueying Zhang et al. 2019). *In vivo*, GQs can be stabilised or destabilised in many ways, for example through changes in the cellular environment, or by interactions with helicases such WRN (Kamath-Loeb, Loeb, and Fry 2012). Regulation of GQ formation may provide a mechanism by which the cell can both influence condensate formation and alter selective partitioning of transcripts into them, and poses an interesting avenue for further research.

Post-Transcriptional Modifications

Currently, over 150 post-transcriptional RNA modifications are known (Boccaletto et al. 2018). In spite of this, their biological role remains relatively unclear. Post-transcriptional modifications are a mechanism by which the sequence identity, charge, structure, and protein-binding sites of an RNA can

be dynamically regulated in response to changes in the cells environment. All of these properties, as discussed, can be integral to an RNAs role in BMC formation and regulation. Methylation of adenosine residues at position 6 (m^6A) is the most prevalent post-transcriptional modification on eukaryote RNA (Wei et al. 2017). m^6A modifications can destabilise RNA duplexes by inducing base rotation. In the case of unstructured RNA, they can induce stronger base-stacking interactions as a result of its hydrophobic interaction with the phosphate backbone (Roost et al. 2015). Proteins with a YTH domain bind to m^6A-containing RNAs and YTH-protein-mRNA complexes can undergo phase separation. mRNAs with only a single m^6A modification did not induce phase separation, but polymethylated mRNAs did (Ries et al. 2019). This highlights how dynamic RNA modifications can alter how RNA participates in BMC formation and regulation, as well as the importance of multivalency in this process. *In vivo*, mammalian SGs enriched for m^6A-modified mRNAs, with modification levels inside the granule being 50 % higher than those outside. This is perhaps unsurprising since SGs also enrich for YTHDF-family proteins (Fu and Zhuang 2020; M. Anders et al. 2018). There remain many unanswered questions concerning how RNA modifications can influence selective partitioning, which modifications can induce condensate formation, and whether condensates can facilitate removal or addition of RNA modifications. Figure 1.8 displays some examples of how this process may influence condensate metabolism, through alteration of RNA-protein interactions and intra- or inter-molecular RNA-RNA interactions:

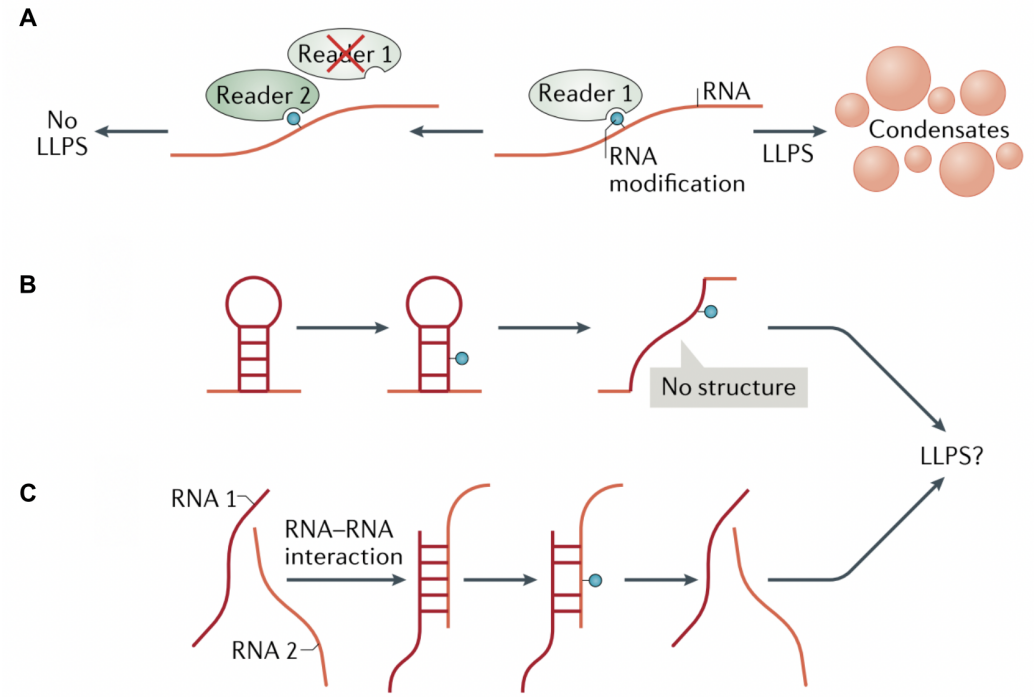


Figure 1.8: Possible Mechanisms via which RNA Post-Transcriptional Modifications can Influence Condensate Formation. A) "Reader" proteins that recognise specific post-transcriptional modifications may either inhibit or induce phase separation B) Post-transcriptional modifications may alter intramolecular RNA structures and promote phase separation C) Post-transcriptional modifications may affect intermolecular RNA-RNA interactions leading to liquid-liquid phase separation. Figure adapted from (Roden and Gladfelter 2021)

1.2.5 Alternative Cleaveage and Polyadenylation (APA)

Alternative cleavage and polyadenylation (APA) is a biological process by which the cell can produce multiple mRNA isoforms from the same gene. Over 70 % of mammalian genes contain multiple polyadenylation sites (PASs) which are subject to dynamic selection (Yongsheng Shi 2012; Mayr 2016; Neve and Furger 2014; Neve, Burger, et al. 2016; Neve, R. Patel, et al. 2017). 3' untranslated region APA (3'UTR-APA) is where mRNA isoforms

are produced that encode the same protein, but differ in the length of their 3'UTRs (Figure 1.9). In some cases, the length difference between isoforms can be considerable, and it is postulated that this can lead to significant differences in their metabolism, affecting processes such as localisation, stability and translational efficiency (Mayr 2016; B. Tian and Manley 2017). The 3'UTR-APA profile of a cell can vary drastically between cell types, and is known to change in response to stress (Yuting Liu et al. 2013; Hollerer et al. 2016; E. T. Wang et al. 2008).

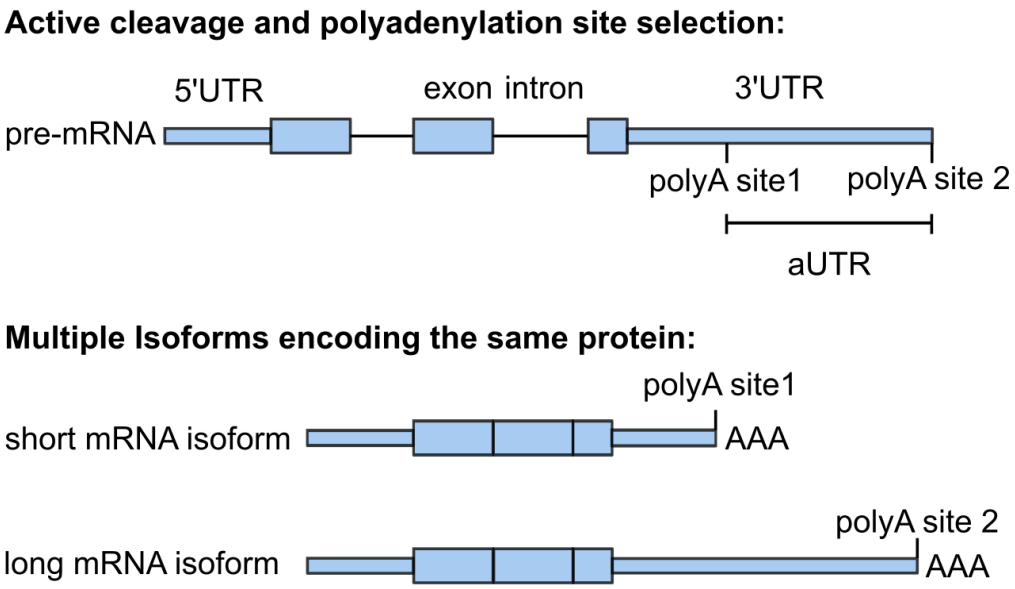


Figure 1.9: 3'UTR Alternative Cleavage and Polyadenylation. pre-mRNA can harbour multiple cleavage and polyadenylation sites. In the diagram, cleavage at polyA site 1 results in production of the "Short" mRNA isoform, while cleavage at polyA site 2 results in production of the "long" isoform. These transcripts differ by a stretch of sequence known as the alternative untranslated region (aUTR) but encode the same protein isoform.

As already discussed, the length, sequence and structure of an RNA can affect both condensate formation and how it partitions into them. Given this,

it becomes apparent that APA, a cellular process which can alter all of these elements, has the potential to be a key mechanism in condensate regulation, and determination of mRNA fate, by selective targeting transcripts to condensates involved in RNA metabolism. One study suggests that transcripts with longer 3'UTRs are more likely to be bound by TIA1 (a key SG component) and consequently more likely to associate with SGs. It was shown that U-rich sequences (known TIA1 binding sites) within these UTRs seems to drive this (D. Zheng et al. 2018). However, to date, there has been very little work done examining this process through the lens of phase separation, despite its clear potential to play a critical role in shaping this process.

1.2.6 RNA in cellular Biomolecular Condensates

Many BMCs are enriched in RNA and RBPs and are known as ribonucleoprotein (RNP) granules, which are highly conserved and ubiquitous. As well as containing an RBD, it has been shown that RBPs are significantly enriched in IDRs compared to other protein groups, and evidence suggests that both domains can play a role in condensate formation (Kroschwald, Munder, et al. 2018; Cid-Samper et al. 2018; Chen Wang, Uversky, and Kurgan 2016; Varadi et al. 2015). As with protein-protein and RNA-RNA interactions; RNA-protein interactions can be both specific and non-specific. RBDs (such as zinc fingers and K-homology domains) mediate specific interactions with certain RNA structural or sequence elements (Castello et al. 2012). For example, the FUS C-terminal RGG motif that specifically recognises RNA GQ structures which promotes phase separation (Imperatore et al. 2020). Non-

specific interactions occur between protein IDRs and RNA, including charge-charge interactions with the negatively-charged RNA phosphate backbone, as well as π-stacking, or cation-π interactions between RNA base-pairs and amino acid residues (Dutagaci et al. 2021; Vernon et al. 2018)(Figure 1.10A). Figure 1.10B displays how both specific and non-specific multivalent interactions between an RNA molecule, and multiple RBPs contribute to the formation of a complex called a ribonucleoprotein (RNP). These RNP complexes can interact with one another to assemble into macromolecular RNP granules (Figure 1.10C).

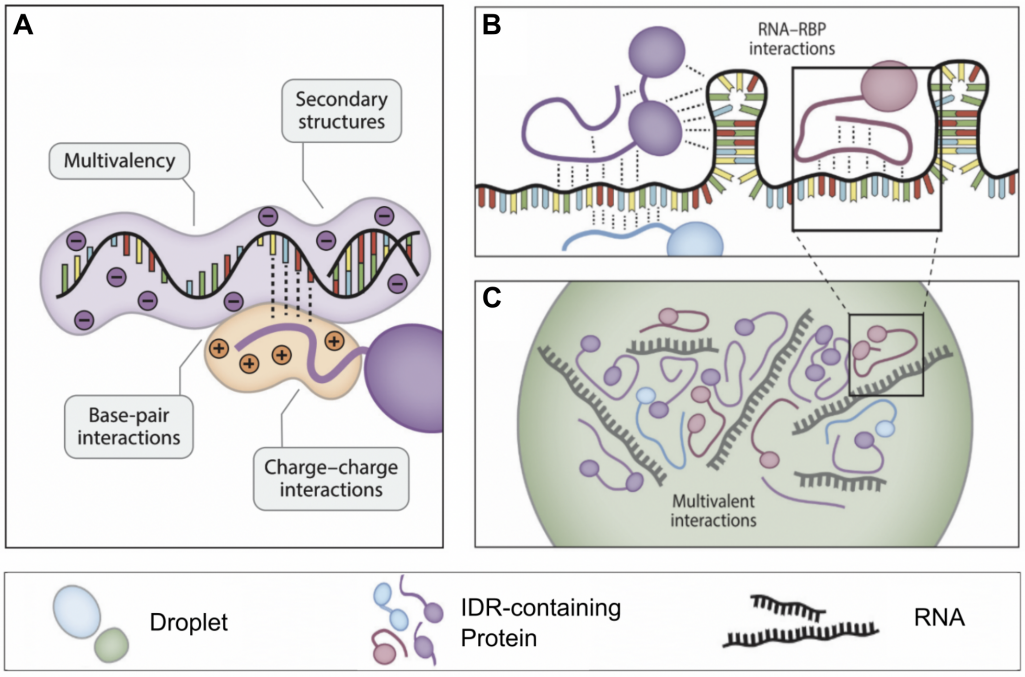


Figure 1.10: Protein-RNA Interactions in Ribonucleoprotein Granules. A) Modes of interaction that can occur between RNA and proteins, including charge-charge interactions and base-pair interactions. B) Interactions between an RNA molecule and multiple RBPs results in the formation of (RNPs) C) RNP molecules are then able to coalesce to form RNP granules in a process mediated by multivalent interactions. Figure adapted from (Rhine, Vidaurre, and Myong 2020)

Diagrammatic summary of the driving forces behind RNA selection in BMCs

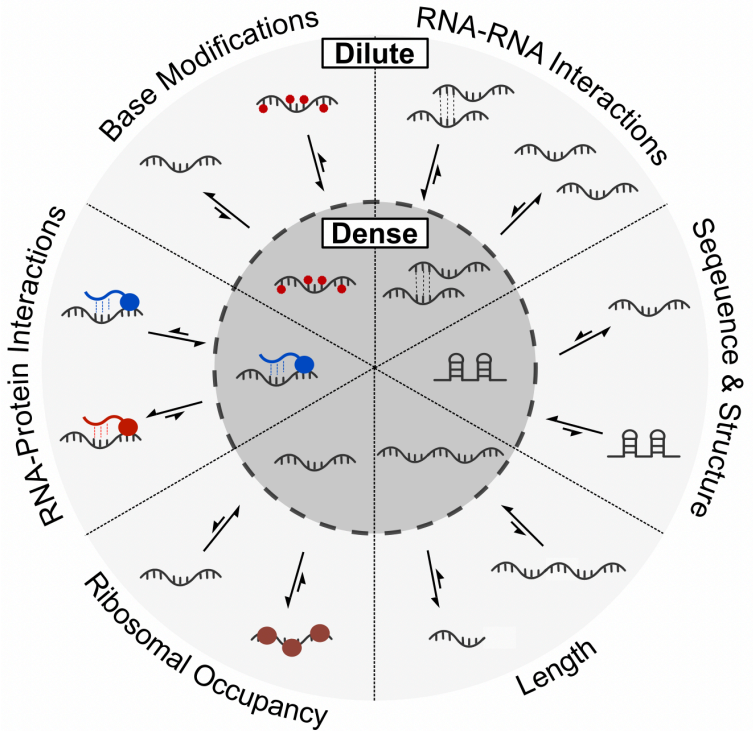


Figure 1.11: The driving forces behind RNA selection in biomolecular condensates. The diagram displays six key driving forces that may determine how an RNA is partitioned in BMCs that have been discussed in Sections 1.2.2 to 1.2.6. Here, "base modifications", "sequence structure" and "length" refer to properties of the RNA molecule.

RNA's Role in Nucleating RNP Granules

In order to undergo phase separation, many proteins are required to be at a high concentration threshold, otherwise known as C_{sat}. However, RNA is able to decrease the C_{sat}, through providing additional multivalent contacts. Many RBPs can bind to a single RNA, which can concentrate them in close proximity and promote oligomeric interactions (Figure 1.10B). For FUS, ad-

dition of RNA significantly reduces the C_{sat} allowing phase separation at concentrations of 5 μM (Maharana et al. 2018). Multiple FUS molecules can bind to a single RNA and this fits with the idea that RNA can serve to seed phase-separation through oligmerisation of RBPs and promotion of intermolecular interactions. Interestingly it was observed that FUS-crosslinking with RNA followed by treatment with a nuclease did not destroy the complexes. This suggests that protein-protein contacts promote stability as well as RNA-protein or RNA-RNA interactions, but that the RNA was necessary to allow enough of these interactions to occur, in order to reach the critical phase-separation threshold (Schwartz et al. 2013).

There are many clear examples where RNA has been shown to be essential for BMC formation both *in vitro*, and *in vivo*. NEAT1 is a long non-coding RNA found in the nucleus that is responsible for formation of paraspeckles (Bond and Fox 2009; Clemson et al. 2009). NEAT1 has two isoforms, one that is 4 kb and polyadenylated (short isoform) and the other that is 23 kb and not polyadenylated (long isoform). The long isoform is required for paraspeckle formation, and accordingly paraspeckle formation is induced following a switch in transcription from the short to the long isoform (Yamazaki et al. 2018). Many RBPs associate with a single long NEAT1 RNA and these RNPs then coalesce to form paraspeckles. Using RNA interference to ablate NEAT1 expression resulted in complete eradication of paraspeckle formation in cells (Clemson et al. 2009). Indeed, more recently, it was demonstrated that RNA is essential to the formation and maintenance of most cellular BMCs (Decker, J. M. Burke, et al. 2021). This was evidenced by degradation of cellular RNA by RNase L, which resulted in a reduction of the number

of P-bodies, alteration of the nucleolar structure, as well as affecting the size and shape of Cajal bodies and paraspeckles.

Regulation of Biomolecular Condensate Composition

Individual BMCs can contain large numbers of distinct molecular components. For example, SGs can contain over 1800 RNA transcripts as well as more than 300 proteins (Markmiller et al. 2018; Khong et al. 2017). Some components are specific to certain condensates, whilst others can be shared between different types. Moreover, composition can differ drastically depending on cellular conditions and alter rapidly in response to cellular signals (Youn et al. 2018; Weidtkamp-Peters et al. 2008; Matheny, Rao, and Parker 2019). Additionally, most components are also able to rapidly exchange between the droplet phase and the surrounding environment (Woodruff et al. 2017). In most instances, formation is dependent on only a subset of key condensate components, referred to as "scaffolds"." The protein PML is an example of a scaffold; when it is knocked out, PML bodies are unable to form, and when it is overexpressed, there is an increase in the number of PML bodies (Zhong et al. 2000; De Stanchina et al. 2004). Components that are not critical for condensate assembly and that may interact transiently are referred to as "clients". Generally, it can be considered that scaffold molecules should be present in higher concentrations than client molecules in condensates. Often, *in vivo*, it can be difficult to distinguish between the two classes and this can vary dependent on cellular conditions, but it can serve as useful terminology when trying to understand the principles underlying component selection.

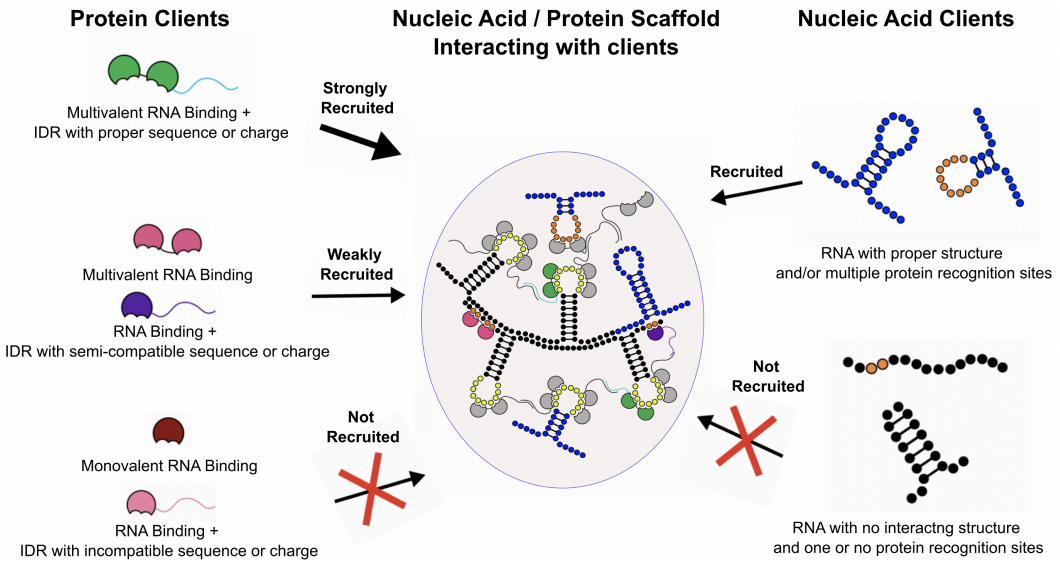


Figure 1.12: Simple Model of Biomolecular Condensate Composition Regulation. A more detailed explanation of the interactions depicted in this figure is provided in the text. Figure adapted from (Ditlev, Case, and Rosen 2018)

In sections 1.2.2 to 1.2.6, protein-protein, RNA-RNA and protein-RNA interactions that promote phase separation were discussed. Figure 1.12 is a model that displays how an assimilation of these interactions can lead to condensate formation, and how each type of interaction can play a role in influencing the molecules it partitions. The scaffold molecules are an RBP with multiple RBDs (grey), and an IDR, as well as RNA (black) with a specific structural element (yellow). A combination of specific protein-RNA interactions, between the RBDs and the RNA, non-specific protein-RNA interactions, between the protein IDR and the RNA, protein-protein interactions, between IDRs, and RNA-RNA interactions, between base-pairs, lead to stabilisation of the complex. The strength of recruitment to the condensate depends on the number of favourable interactions a client molecule can have with the scaffolds. Protein clients with multiple RBDs and IDRs

with favourable charge or sequence motifs are strongly recruited, whereas proteins with monovalent RNA binding or IDRs with unfavourable charge or sequence are excluded. Similarly RNAs with favourable structure and sequence are more strongly recruited than RNAs with no interacting structure, and few protein recognition sites. Essentially the more favourable multivalent interactions a client molecule can have with scaffold molecules, the more strongly recruited it is to the droplet phase. Clearly this is a simplified model but it serves as a good conceptual framework for understanding regulation of condensate composition.

Post-transcriptional and post-translational modifications can serve to dynamically regulate recruitment. By changing the chemical properties or multivalency of a molecule, the way it interacts with other condensate constituents can be altered. Banani *et al.* (2016) displayed this using a simple system with two scaffold proteins, one consisting of SUMO modules and the other SUMO-interacting modules (SIMs). Changing which scaffold was in excess altered which client molecules were recruited to the complex. When the SUMO-containing scaffold was in excess, clients with more SIMs were more strongly recruited. However, when the scaffold containing SIMs was in excess, the opposite was true with clients containing more SUMO residues being enriched (Banani, Rice, et al. 2016). This serves as compelling evidence for how altering the relative valency of scaffolds can affect client partitioning and additionally how post-translational modifications such as sumoylation enable cells to dynamically change the composition of BMCs in response to cues.

1.2.7 Physical Properties of Biomolecular Condensates

The constituents of a condensate determine its physical properties, but over time, these properties are susceptible to change. Droplet hardening is the process by which a condensate transitions from a fluid-like state towards a solid-like state, corresponding to decreased exchange of molecules with the surrounding environment (A. Patel et al. 2015). This process has been observed in multiple condensates formed from proteins with IDRs, including PUB1 and FUS (A. Patel et al. 2015; S. Xiang et al. 2015). In the dense phase, disordered proteins have a propensity to adopt certain conformations for extended periods of time, including β-strands which have been shown to interact and promote formation of amyloid-like structures (Lin et al. 2015; T. W. Han et al. 2012). Additionally, over time, the molecules within a condensate can become increasingly entangled which can also reduce its fluidity (Watanabe 1999). Dis-regulation of this progression from liquid-to-solid has now been shown associated with numerous diseases including amyotrophic lateral sclerosis (ALS) and frontomporal lobar dementia (FTD), where RNP granules convert to solid-like fibrils in motor-neurons (discussed further in Section 1.3) (A. Patel et al. 2015). Given this observation, the importance of regulating this process becomes apparent. RNP granules have been shown to be enriched in ATP-dependent machineries, including chaperons, disaggregases and molecular motors (S. Jain et al. 2016; Kroschwald, Maharana, et al. 2015). Accordingly, depletion of ATP was shown to increase the viscosity of certain BMCs (Brangwynne, Mitchison, and Hyman 2011; S. Jain et al. 2016). These ATP-driven proteins can regulate the degree of crosslinking

and fiber formation within condensates, and as such help to tune the physical properties of a particular droplet. Figure 1.13 shows the dynamic regulation of liquid-to-solid transitions in BMCs. A liquid-like condensate can rapidly exchange molecules with the surrounding environment and undergo internal rearrangement. As it transitions to a more solid-like state, this ability becomes reduced, but can be rectified by ATP-dependent machineries.

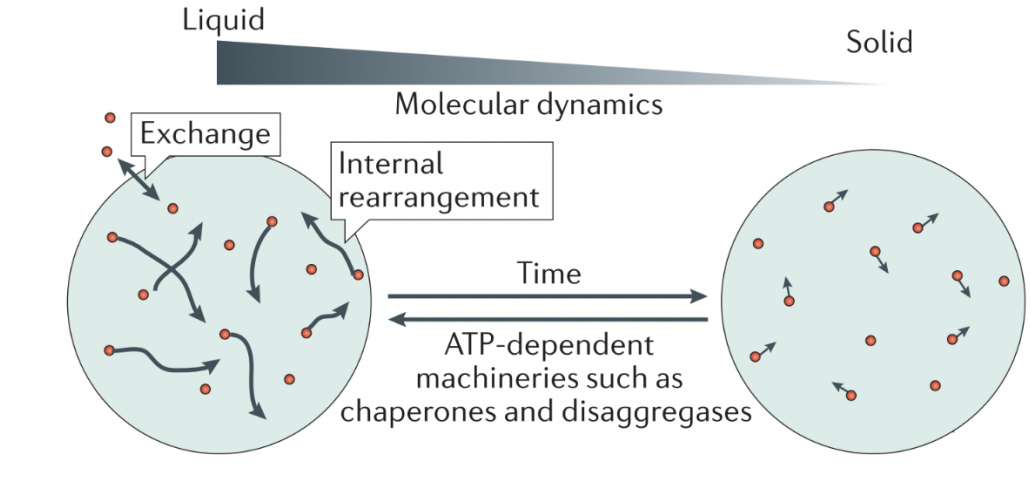


Figure 1.13: Regulation of Liquid-to-Solid Transition in Biomolecular Condensates. Over time droplets age and transition from a fluid-like-state to a solid-like state. This process is regulated by machineries such as chaperones and disaggregases. An example of the former is Hsp70/Hsp40 which is postulated to limit extensive formation of amyloid-like fibers in SGs (S. Jain et al. 2016; Ramaswami, J. P. Taylor, and Parker 2013). Whilst disaggregases such as Hsp104, serve to dissolve protein aggregates within SGs (Kroschwald, Maharana, et al. 2015). Figure from (Banani, H. O. Lee, et al. 2017)

Interestingly, it has also been shown that RNA can also help control the fluidity of RNP granules (Elbaum-Garfinkle et al. 2015; Maharana et al. 2018; N. Taylor et al. 2016). Fluorescence recovery after photobleaching (FRAP) experiments showed that LAF-1 droplets containing RNA recovered

more quickly than those that without RNA, indicative of increased fluidity (Elbaum-Garfinkle et al. 2015). Additionally, RNA appears able to regulate the material exchange rates of as well as the shape and rigidity (T. E. Audas et al. 2016; Huaiying Zhang et al. 2015). However, these findings are predominantly based on observation, and how the specific physical properties of an RNA may contribute to modulating the characteristics of condensates remains largely unknown.

1.2.8 Regulation of Biomolecular Condensate Formation

BMC formation *in vivo* is strictly spatiotemporally regulated. One of the main ways a cell can achieve this is through control of the concentration of specific condensate components. Once the concentration of these components surpasses a critical threshold, phase separation occurs. RNA expression levels are controlled through transcription, stabilisation, and degradation, and the concentration of RNA can directly influence condensate formation and maintenance. For example, transcription of ribosomal RNA results in nucleoli formation, and similarly transcription of NEAT1 leads to paraspeckle formation (Berry et al. 2015; Bond and Fox 2009; Clemson et al. 2009). RNA concentration can act to buffer phase separation with low levels of RNA promoting phase separation, and high concentrations inhibiting it (Maharana et al. 2018). In a similar way, levels of protein components can also be regulated through translation and degradation. Overexpression of PML leads to an increase in the number of PML bodies that formed (Zhong et al. 2000;

De Stanchina et al. 2004). In addition to this, the cell can regulate condensate formation by altering specific local concentrations of condensate components, or inhibitory elements. P-granule localisation to the posterior of *C. elegans* embryos following post-symmetry breaking is mediated by creating a concentration gradient of the inhibitory MEX-5 protein, with the lowest concentration being at the posterior of the embryo (Brangwynne, Eckmann, et al. 2009). This results in condensate formation being promoted at the posterior of the cell, and as a result, concentration of P-granule components in this region. Changes in the cell's environment have also been shown to induce phase separation processes. Poly(A)-binding protein (PAB1) forms condensates in response to changes in the pH or temperature of the cell, that have been shown to promote cell survival (Riback et al. 2017). This was demonstrated as yeast with mutant PAB1 that were exposed to heat shock, displayed reduced fitness, despite displaying normal growth behaviour in the absence of stress.

1.3 Cellular Biomolecular Condensates

BMCs have now been implicated in a vast array of biological processes throughout the cell. Many are RNP granules that constitute specific subsets of RNA and RBPs (Bond and Fox 2009; Sheth and Parker 2003; White and Lloyd 2012; Boisvert et al. 2007). For example, the nucleolus consists primarily of nascent ribosomal RNA, whilst paraspeckles form from NEAT1 RNA and specific RBPs such as FUS (Clemson et al. 2009; Berry et al. 2015). It is thought many have a role in RNA metabolism, including in transcrip-

tional regulation, RNA splicing, and pre-mRNA processing (Lyon, Peeples, and Rosen 2021; Ying et al. 2017; Tatomer et al. 2016; Hirose et al. 2014). Additionally, liquid-liquid phase separation has also been attributed to processes such as immune signalling, where condensates form on the plasma membrane to promote signal transduction (Su, Ditlev, Hui, et al. 2016). Some BMCs, such as PBs, are widely spread and found in a diverse array of cell types, whilst others which are more specialised, such as germ granules, are found only in a distinct subset (Hyman, Weber, and Jülicher 2014). Condensates provide a means of compartmentalising particular components of specific biological reactions, and enhancing the reactivity of their components by crowding them. It has been shown *in vitro* that the activity of the hammerhead ribozyme can be increased 70-fold when it is sequestered with RNA substrates in phase-separated condensates (Strulson et al. 2012). It is therefore easy to envisage how this kind of augmentation also occurs *in vivo*. One issue with concentrating large numbers of molecules in close proximity is that if they are improperly regulated, strong interactions may result in a transition from a liquid-like state and result in formation of pathological aggregates (Alberti and Hyman 2016). This process has now been attributed to many diseases, most prominently, neurodegenerative diseases including ALS and FTD (A. Patel et al. 2015).

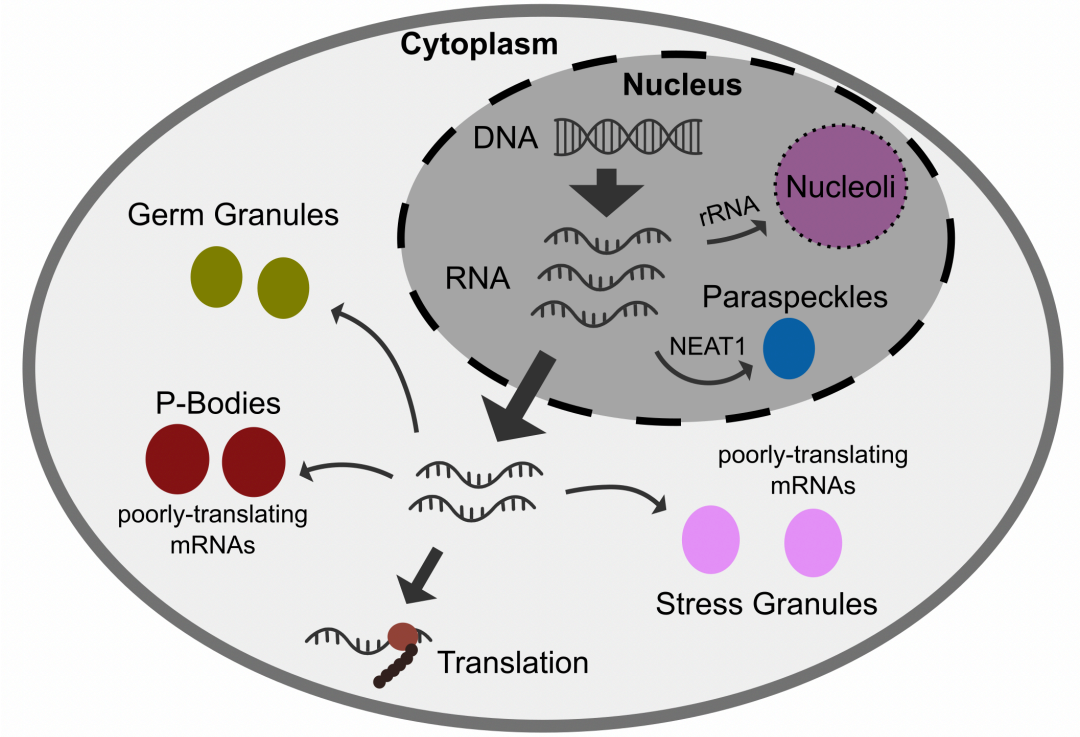


Figure 1.14: Examples of biomolecular condensates found throughout the cell along with their subcellular locations. Many BMCs are postulated to be involved in RNA metabolism and this figure depicts specific types of RNA that are known to associate with specific condensates.

1.3.1 P-Bodies

PBs were first observed in yeast cells in 2003, and have since been subject to intense investigation (Sheth and Parker 2003). They are highly conserved in eukaryotes and found in the cytoplasm, comprising predominantly of translationally repressed mRNAs, as well as proteins related to 5'-3' mRNA decay. These observations initially led to the idea that PBs were sites of mRNA degradation. However, it has since been demonstrated that macroscopically observable PBs were not necessary for mRNA decay to occur, and that this

process still occurs throughout the cytoplasm. Additionally, mRNAs that are released from PBs are able to re-associate with translating polysomes (Eulalio, Behm-Ansmant, and Izaurralde 2007; Decker, Teixeira, and Parker 2007; Brengues, Teixeira, and Parker 2005; Hubstenberger et al. 2017). Thus, an alternative model has emerged that suggests that PBs act as cellular stores of non-translating mRNAs (Parker and Sheth 2007; Kroschwald, Maharana, et al. 2015). Interestingly, PB assembly is dependent on translational repressors, like DDX6, rather than decay machineries, which further emphasises a role for PBs outside the remit of mRNA decay (Ayache et al. 2015). Furthermore, mRNAs have been identified in PBs that are stabilised and sheltered from degradation (Congwei Wang et al. 2018). One hypothesis is that PBs simply act as a storage hub to regulate translational flow through sequestering and releasing mRNAs in response to cues from the cellular environment. Whilst PBs are undoubtedly highly important structures, their precise role in regulation of gene expression remains to be determined. Table 1.1 displays some of the key PB protein components:

P-body protein	**Function**
NBDY	mRNA decay
4E-T	Translational repression
PNRC2	Nonsense-mediated mRNA decay
XRN1	5'-3' exoribonuclease mRNA decay
DCP1	mRNA decapping
EDC4	Enhancer of mRNA decapping
DCP2	mRNA decapping
CCR4-NOT	mRNA deadenylation
PATL1	mRNA decapping activator
LSm1-7	mRNA deadenylation
DDX6	mRNA decapping activator
LSm14A	Translational repression
EDC3	mRNA decapping activator

Table 1.1: Key PB protein examples and their functions (D'Lima et al. 2017; Kamenska et al. 2014; D. Zhou and S. Chen 2001; Blasco-Moreno et al. 2019; Vilela et al. 2000; Hernández et al. 2018; Rehwinkel et al. 2005; Denis and J. Chen 2003; Scheller et al. 2007; Ingelfinger et al. 2002; Smillie and Sommerville 2002; Ying Li et al. 2012; Fromm et al. 2012)

Several studies have identified resident PB proteins, including 4-ET and EDC4, that are essential for PB maintenance, with depletion resulting in a loss of PB formation (Eulalio, Behm-Ansmant, and Izaurralde 2007; Andrei et al. 2005; Serman et al. 2007) (Table 1.1). In addition, PB formation has been shown to be dependent on non-translating mRNAs. Treatment of cells with cyclohexamide, which prevents translation but also sequesters

mRNA in polysomes, results in a loss of PB formation (Teixeira et al. 2005). Importantly, some interactions are redundant in PB assembly, and knocking out several key PB constituents isn't sufficient to prevent PB formation (Rao and Parker 2017). This finding indicates that a complex network of protein-protein, protein-RNA and RNA-RNA interactions (with the exception of a small number of key factors), is more important to PB assembly than the individual components.

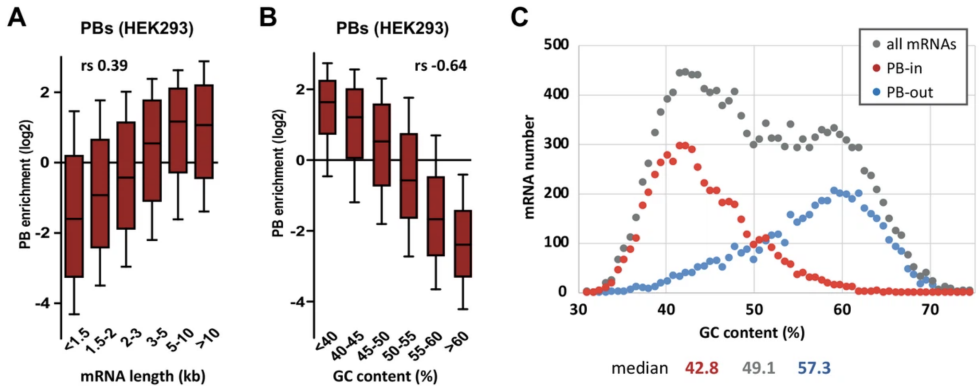


Figure 1.15: P-Body transcriptome is influenced by mRNA physical properties. A) Correlation of mRNA length with PB enrichment. B) Correlation of mRNA GC content and PB enrichment. C) GC content distribution in all human mRNAs, PB-enriched mRNAs and PB-depleted mRNAs. Figure from (Courel et al. 2019)

Recently it has become possible to isolate PBs from cells using fluorescence-activated particle sorting. Hubstenberger *et al.* showed that mRNA that was enriched in these bodies tended to be more poorly translated, with the proteins they encode being less abundant (Hubstenberger et al. 2017). Later analysis revealed that the mRNAs found in PBs tended to be longer, and more AU-rich, than those which were depleted, as shown in Figure 1.15. mRNAs that were preferentially sequestered in PBs also tended to be enriched

with low-usage codons. It is thought that this may manifest in low ribosomal occupancy of these transcripts, which may result in their enrichment in PBs. Using cell extracts, synthetic *in vitro* transcribed RNA, and recombinant DDX6, it was further demonstrated that AU-rich RNA stimulated droplet formation whilst GC-rich RNA prevented it (Courel et al. 2019). This raises the possibility that the physical properties of the mRNA itself, rather than its association with ribosomes, may be critically important in PB formation. Interestingly, the PB transcriptome appears to be highly dynamic and varies in response to stress (Matheny, Rao, and Parker 2019). In the absence of stress, PBs were found to consist primarily of poorly translating mRNAs, whilst under stress, the composition shifts to be primarily composed of longer mRNAs with a weaker correlation with mRNA translation.

1.3.2 Stress Granules

SGs form in the cytoplasm in response to various cellular stresses including heat shock, viral infection and hypoxia (White and Lloyd 2012; Khong et al. 2017; Adivarahan et al. 2018). They are highly dynamic and disappear within minutes of removal of stress (Wheeler et al. 2016). Under stress conditions, eIF2α becomes phosphorylated in a process known as the integrated stress response. This process involves transcriptional upregulation of genes involved in cell survival and homeostasis, and limitation of translation initiation (Haigis and Yankner 2010; Taniuchi et al. 2016). This leads to a growing pool of poorly-translating RNAs following ribosomal run-off, that subsequently condense along with their associated RBPs to form SGs (Prot-

ter and Parker 2016; Buchan and Parker 2009). SGs consist of a dense "core", where it is believed that majority of SG RNA is located, surrounded by a dynamic outer "shell" (S. Jain et al. 2016). Recent advances have allowed for purification of SG cores via immunoprecipitation with antibodies to either accessible SG components, or a tagged SG component (Khong et al. 2017; S. Jain et al. 2016). The SG transcriptome has been shown to consist of long, poorly-translating mRNAs (Figure 1.16) (Khong et al. 2017; Namkoong et al. 2018; Matheny, Rao, and Parker 2019).

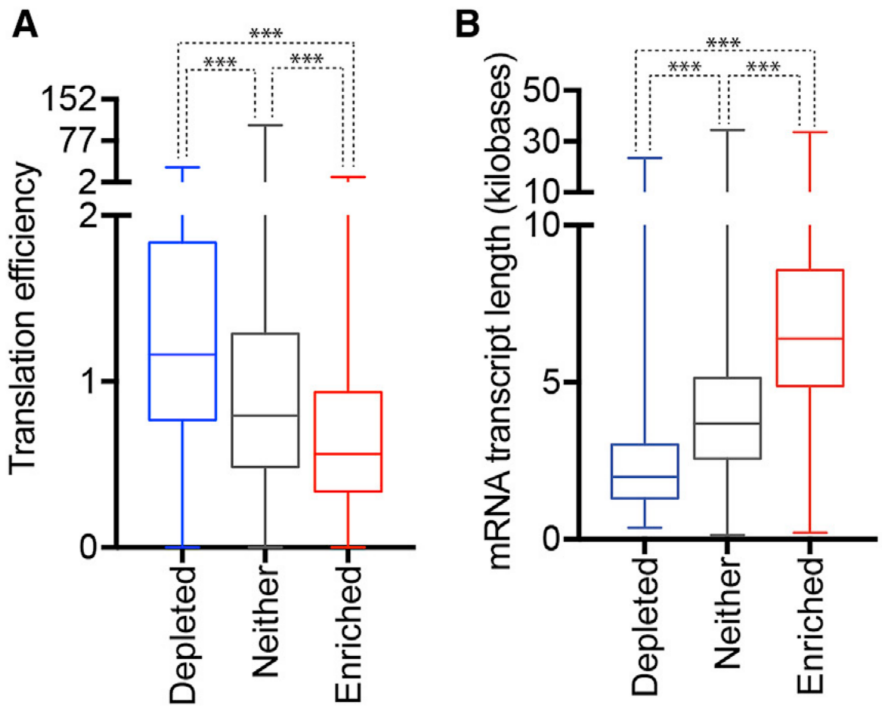


Figure 1.16: Stress granule transcriptome is influenced by mRNA physical properties. A) Correlation of translation efficiency with SG enrichment B) Correlation of mRNA length with SG enrichment. Figure adapted from (Khong et al. 2017)

Interestingly, observations suggest that only 10 % of the total, cellular, mRNA molecules actually accumulate within SGs, which indicates that they

will have a relatively minor effect on the bulk mRNA (Khong et al. 2017). Furthermore, cells which have G3BP1 & 2, essential SG proteins, knocked out, fail to form SGs, but still effectively repress global translation (Kedersha et al. 2016). Thus, as with PBs, there still remains much debate as to the precise function of these BMCs. In addition to requiring non-translating mRNAs for formation, SGs are dependent on a number of essential protein components, a few of which are listed in table 1.2:

Stress Granule protein	**Function**
DDX6	Translational suppression and mRNA decay
YBX1	Translational suppression
LARP1	Translational suppression
PABPC1	PolyA lengthening and translation termination
YTHDF1	Degradation of m^6A-containing transcripts
PUM1	Translational suppression
OGFOD1	Translational termination
G3BP1,2	DNA- and RNA-unwinding
TIAL1	Translational regulation
TDP-43	Transcriptional and translational regulation

Table 1.2: Key SG protein examples and their functions (Smillie and Sommerville 2002; Shurtleff et al. 2017; Fonseca et al. 2015; Behm-Ansmant et al. 2007; D. Han et al. 2019; A. R. Morris, Mukherjee, and Keene 2008; Singleton et al. 2014; French et al. 2002; Q. Wang et al. 2019; Ayala et al. 2008)

SG formation is suggested to be dependent on a summation of protein-protein, RNA-protein and RNA-RNA interactions (Van Treeck and Parker 2018; Van Treeck, Protter, et al. 2018; Matheny, Van Treeck, et al. 2021).

As discussed above, knockouts of essential SG proteins, such as G3BP1 and G3BP2, impair the cell's ability to form SGs, providing evidence for the role of protein-mediated interactions (Table 1.2). Similarly, since the SG transcriptome can be largely recapitulated *in vitro*, via RNA-self assembly, and proteins which inhibit RNA self-assembly *in vitro* also inhibit SG formation *in vivo*, it is clear that RNA-RNA interactions are also central to SG formation.

1.3.3 Germ Granules

Germ granules are unique to germ cells and appear to be inextricably related to germ cell biology. They are involved in regulatory post-transcriptional processes that dictate germ cell fate across metazoans. Like with other RNP granules, some components are essential and conserved across species, including the DEAD-box RNA helicase Vasa (DDX4 in mammals or GLH-1 in *C.elegans*) and nanos (nos) mRNA. Germ granules can exhibit diverse morphologies and functions dependent on the stage of development of an organism (Pitt, Schisa, and Priess 2000; Ephrussi, Dickinson, and Lehmann 1991). During early development, germ granules surround the nuclei and are known in this instance as nuage. It is thought that nuage has a role in the biogenesis of small piwi protein-interacting RNAs (piRNAs) that mediate the self-/non-self recognition of germline mRNAs in *C.elegans* and transposon silencing in *Drosophila* (Updike et al. 2011; Xiawei Huang, Tóth, and Aravin 2017). During early embryogenesis, germ granules that form in the germ plasm, are known as polar granules in *Drosophila* and P granules in *C. ele-*

gans. The germ plasm is a specialised cytoplasm that locates to the posterior of zygotes in *D. Melanogaster* and *C. elegans*. Here polar and P granules, respectively, accumulate mRNA that encodes for proteins essential for the specification of germ cell fate. In this way, germ cell lineage is spatio-temporally regulated (Voronina and Seydoux 2010; Trcek and Lehmann 2019). Interestingly, knocking out two key germ granule proteins, MEG-3 and MEG-4, in *C. elegans* and preventing germ granule formation lead to defects in RNA interference. This resulted in aberrant silencing of germline-expressed genes and epigenetic defects that were propagated between generations (Dodson and Kennedy 2019). These findings are indicative of a central role of germ granules in ensuring the correct transfer of genetic information between generations and promoting offspring viability.

1.3.4 Paraspeckles

Paraspeckles are nuclear bodies that are localised to the interchromatin space in mammalian cells and are formed by RBPs, such as FUS and NONO, associating with the long non-coding RNA, NEAT1 (Fox, Lam, et al. 2002; Clemson et al. 2009; Bond and Fox 2009). One major function of paraspeckles is nuclear retention of mRNA. This process has been linked with the stress response specifically in the case of the gene CTN (Fox and Lamond 2010). The CTN gene encodes multiple mRNA isoforms, and under normal cellular conditions, a "long" mRNA is produced through cleavage and polyadenylation (CPA) at a distal CPA site, which is sequestered in paraspeckles. Stress signals can mediate a change in CPA site usage to a more proximal CPA

site in the pre-mRNA, and this "short" isoform no longer associates with paraspeckles, and is released from the nucleus, allowing for rapid cellular response (Prasanth et al. 2005). Modified nucleotides within the aUTR sequence of CTN are responsible for retaining the "long" mRNA isoform in paraspeckles (Fox and Lamond 2010). Many of the paraspeckle-associated proteins also interact with a large subset of transcription factors, RNA and DNA. As a result, paraspeckles may also be intimately involved with processes such as transcription initiation, elongation and termination, as well as facilitating co-transcriptional processing and regulation of mRNA localisation and export (Knott, Bond, and Fox 2016).

1.3.5 Nucleoli

The nucleolus is the largest membraneless organelle and it is the site of ribosome biogenesis, modulating processes including RNA polymerase(I)-driven transcription, modification and processing of ribosomal RNA (rRNA), and the assembly of rRNA complexes (Berry et al. 2015; Brangwynne, Mitchison, and Hyman 2011; Boisvert et al. 2007). Lafontaine *et al.* demonstrated that the nucleolus is actually a multiphase liquid condensate that has three nested layers, each associated with a different stage of ribosome biogenesis (Lafontaine et al. 2021). The process begins in the inner core where rRNA transcription takes place in the fibrillar centre (FC). rRNA processing and rRNA modification then occurs in the next layer which is known as the dense fibrillar component (DFC) and finally, rRNP assembly, rRNP transport and rRNP surveillance then occur in the granular component (GC) (Lafontaine

et al. 2021). The nucleolus provides an excellent example of how phase separation can be used to mediate and compartmentalise specific biological processes. Once rRNA has completed its first stage of processing, one can postulate that the post-transcriptional modifications or structural changes it undergoes lead to changes in its physiochemical properties that result in it being released from the first nucleolar compartment and into the second for the next stage of processing. In this way a very efficient reaction pathway is created, whereby processed substrates subsequently become excluded, and can move on to the next stage of processing, or go and effect their biological roles within the cell.

1.3.6 Other Biomolecular Condensates

There are many different types of BMCs found throughout cell, with a wide array of functionalities, and more and more are being discovered. Cajal bodies (CB) are nuclear membraneless organelles whose assembly is initiated by small nuclear RNAs, as well as small Cajal body-specific RNAs. These interact with RBPs which subsequently recruit other proteins in the presence of an essential aggregation factor, coilin (Machyna, Heyn, and Neugebauer 2013). They have been implicated in roles such as small nuclear ribonucleoprotein biogensis, telomere maturation and genesis and spliceosome formation (Machyna, Heyn, and Neugebauer 2013). Phase separation has also been suggested as a mechanism for transcriptional control. It can lead to concentration of groups of transcription factors and transcriptional enhancers together referred to as super-enhancers, at transcription sites enhancing gene

expression in this region (Whyte et al. 2013). Reversible, post-translational modifications to super-enhancers can serve to regulate their phase separation in a manner finely in tune with changes in the cell's environment (Hnisz et al. 2017; Boija et al. 2018).

The role of phase separation in immune signalling is also becoming more apparent. Condensates can form in contact with the cytoplasmic membrane in response to external signals, as is the case with T cell receptor signalling (Su, Ditlev, Hui, et al. 2016). Here the T cell receptor (TCR complex) becomes phosphorylated by LCK upon antigen engagement. This triggers a signalling cascade which results in phosphorylation of "linker for activation of T-cells" (LAT), and induces phase separation of LAT. LAT condensates are enriched with proteins involved in activating downstream responses, including RAS signalling and calcium influx. Ubiquitination by CBL, an E3 ubiquitin ligase, serves to disrupt these assemblies and stop signal transduction (Grakoui et al. 1999; Su, Ditlev, Rosen, et al. 2017; Q. Xiao, McAtee, and Su 2021). Interestingly, the overall composition of these clusters is such that they are negatively charged, and this serves to exclude negatively charged phosphatases, which would otherwise interfere with the kinase-mediated signalling (Su, Ditlev, Hui, et al. 2016). In addition to signal transduction from the plasma membrane, phase separation can also affect intracellular immune signalling. Cyclic GMP-AMP synthase (cGAS) acts as a sensor for abnormal cytosolic dsDNA which can be derived from either nuclear or mitochondrial damage, or pathogens (L. Sun et al. 2013). Binding to the dsDNA induces cGAS condensate formation and results in the synthesis of 2',3'-cyclic GMP-AMP (cGAMP) which then activates genes involved in the interferon

response (Jiaxi Wu et al. 2013; W. Xie et al. 2019). Additionally, some viral proteins produce phase-separated condensates that are able to interfere with the host immune response (Heinrich et al. 2018; Monette et al. 2020). Respiratory syncytial virus (RSV) induces sequestration of the NF-{kappa}B subunit p65 in perinuclear intracytoplasmic puncta which prevent it from activating transcription of genes involved in antiviral response (Jobe et al. 2020). Additional BMCs have been associated with DNA damage response, redox imbalance sensing, and metabolic switching.

1.3.7 Biomolecular Condensates and Disease

The role of aberrant phase separation in disease is becoming increasingly apparent, with implications in cancer and neurodegeneration (Alberti and Hyman 2016; Forman-Kay, Kriwacki, and Seydoux 2018; Shin and Brangwynne 2017). There are a few hypotheses for how these disease phenotypes may arise. Firstly, there may be an alteration to the assembly mechanism of a condensate, which can occur through mutation leading to changes in the solubility of components. Secondly, a regulator of phase-separation, such as a protein kinase, may become mutated, resulting in disregulation and aberrant phase transitions. Thirdly, changes in the cell's physiochemical environment may affect regulation of condensates. For example, many condensates require the action of ATP-driven enzymes, such as disaggregases, to maintain fluidity, and disruption of the cell's energy metabolism can interfere with this (Alberti and Dormann 2019). One of the best studied ways that condensate disregulation can result in disease is when key condensate components mu-

tate. A few examples of proteins where this has been shown to occur are listed in table 1.3 below, along with their associated diseases:

Phase-separating protein	**Disease Association**
FUS	ALS, FTD, cancer
EWSR1	ALS, FTD, cancer
TDP-43	ALS
TAF15	ALS, FTD, cancer
HNRNPA1	MSP, ALS
TIA-1	ALS, WDM

Table 1.3: Table displaying phase-separating RNA-binding proteins and their associated diseases (Kwiatkowski et al. 2009; Couthouis et al. 2012; Ian RA Mackenzie and Rademakers 2008; Ian RA Mackenzie and M. Neumann 2012; H. J. Kim et al. 2013; Y. Lee et al. 2018)

Mutations to FUS, and the aggregates it forms as a consequence, have been widely linked to patients suffering from neurodegenerative diseases, including amylotrophic lateral sclerosis (ALS), and frontotemporal dementia (FTD) (Kwiatkowski et al. 2009; Vance et al. 2009; M. Neumann et al. 2009). It has been demonstrated that patient-linked mutations in FUS' prion-like domain accelerate the rate at which FUS droplets transition from liquid to fibrous aggregates *in vitro* (A. Patel et al. 2015) (Figure 1.17). Abnormal FUS phase transitions, as a consequence of mutation, have been shown to result in aberrant paraspeckle and SG formation, both of which have been associated with neurodegenerative diseases (De Santis et al. 2019; Shelkovnikova et al. 2014).

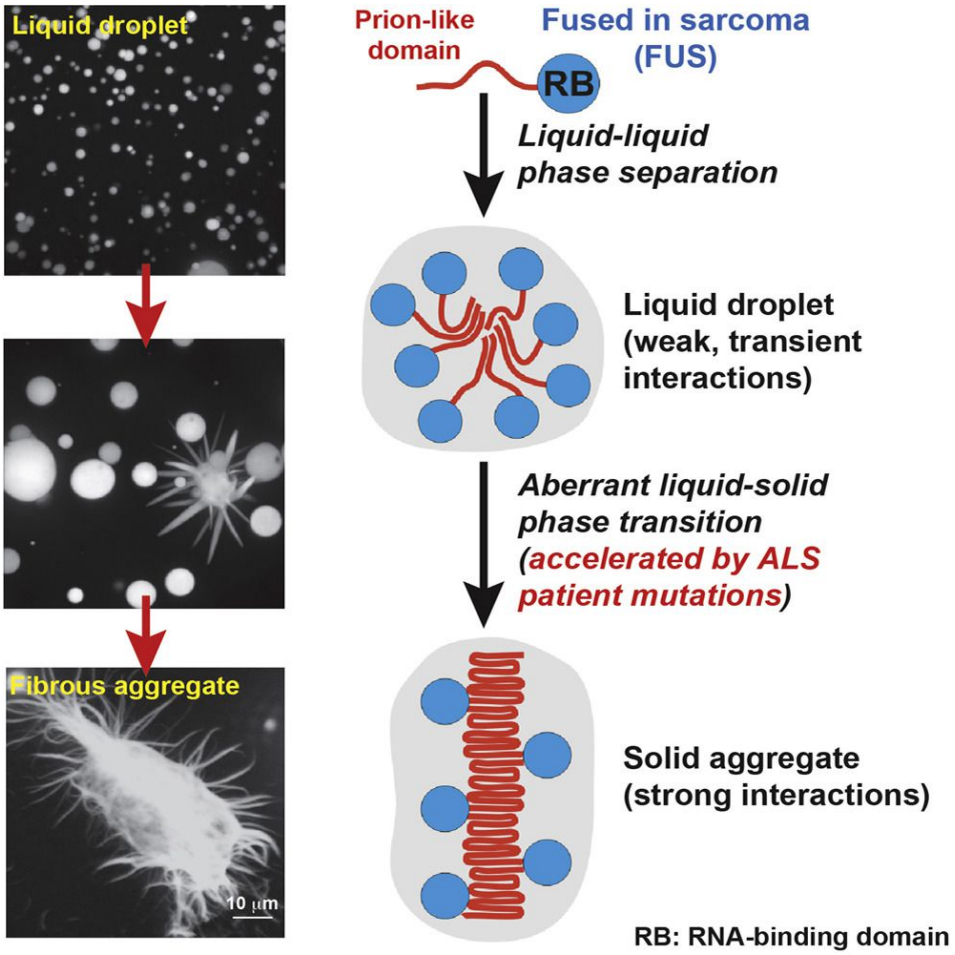


Figure 1.17: Liquid-to-solid phase transition of the RNA-binding protein FUS *in vitro*. RBP FUS undergoes liquid-liquid phase separation in a process governed by weak, transient, multivalent interactions. Over time, this liquid state transitions to a solid-like state due to stronger interactions leading to fibril formation. The transition time is accelerated in FUS mutants found in patients with ALS. Figure from (A. Patel et al. 2015)

Homotypic RNA phase separation is another cause of pathological aggregate formation. Repeat expansion disorders of the hexanucleotide repeat $(GGGGCC)_n$ in the *C9ORF72*(C9) gene is the known to be the most common cause of ALS and FTD (Renton et al. 2011; DeJesus-Hernandez et al. 2011). RNA consisting of this repeat forms GQs, and phase separates to

form gels *in vitro*, and *in vivo* (A. Jain and Vale 2017). These RNA foci have been shown to form both in the nucleus, and the cytoplasm, and can recruit and sequester RBPs, most notably FUS and HNRNPH, leading to the disruption of many cellular processes (Conlon et al. 2016; Fay, P. J. Anderson, and Ivanov 2017).

The advantages of concentrating large numbers of molecules together in a phase separated condensate to produce favourable reaction conditions comes with the cost, that over time, it can lead to the formation of strong interactions and transition from a dynamic, liquid phase, to a solid, pathological phase. Slight perturbations in the balance of regulatory mechanisms, as a result of mutation or changes in the cells physiochemical environment, can lead to disregulation, and ultimately disease (Alberti and Dormann 2019). The interplay between RNA and RBPs is critical to this and needs to be finely tuned in order to maintain proper cellular homeostasis and to avoid the genesis of pathological aggregates. Indeed, there remains much to be understood about the precise way the cell is able to mediate these processes and much to be gained from a therapeutic perspective as we began to untangle the mechanisms responsible for these types of disease.

1.4 *In vitro* Models for Studying Biomolecular Condensates

1.4.1 *In vitro* models

BMCs often contain vast numbers of different components. However, it has been demonstrated that their composition is often dominated by a small subset of key constituents, which likely impart the condensate's overall properties (Lafontaine et al. 2021; P. Yang et al. 2020; Feric et al. 2016). Thus, isolation of major protein components for use in *in vitro* studies has been widely employed, and much of our current knowledge of BMCs has been obtained by this method. We utilised the following three protein constructs for use as *in vitro* models.

1.4.2 DDX4N

DDX4 is a DEAD-box helicase expressed exclusively in human germ cells. It is a major component of germ granules (a cytosolic BMC) and has been shown to be essential to spermatogenesis. It is also a component in the piRNA pathway and functions in repressing transposable elements during meiosis (Castrillon et al. 2000). It contains a DEAD-box helicase domain flanked by two IDRs. The N-terminal IDR (DDX4N, amino acids 1-236) has been extensively studied, and the condensates it forms *in vitro* have been demonstrated to be indistinguishable from those it forms *in vivo* (Nott, Petsalaki, et al. 2015; Nott, Craggs, and A. J. Baldwin 2016).

1.4.3 EWSR1C

EWSR1 is an RBP that has been associated with many different cellular processes including transcription, RNA processing and RNA transport, and is predominantly found in the nucleus. It belongs to the FET family proteins, and is a paralog of the protein FUS. it has been shown to be associated with BMCs involved in DNA damage repair, mediated by the DNA damage-induced polymer poly(ADP-ribose) (Spegg and Altmeyer 2021). EWSR1 contains an RBD flanked by two IDRs similar to DDX4. We employ the C terminal IDR (EWSR1C, amino acids 450-655) for use in this study.

1.4.4 FUS-GFP

FUS is another member of the FET family proteins, like EWSR1. It is predominantly found in the nucleus, and is associated with a wide range of cellular processes, most commonly to do with RNA metabolism, including transcription, pre-mRNA splicing and mRNA transport (Dormann and Haass 2013). It is a component of a variety of different BMCs including SGs, paraspeckles and condensates involved in DNA damage repair (Levone et al. 2021; Shelkovnikova et al. 2014; Sama et al. 2013). The FUS protein consists of a C-terminal RBD and an N-terminal IDR, known as the prion-like domain due to its association with the formation of prions in yeast (O. D. King, Gitler, and Shorter 2012). FUS has been strongly linked to a variety of neurodegenerative diseases, including ALS, and as a consequence, has been subject to intense investigation. It's phase separation behaviour, and the driving forces that are responsible, have been well characterised. In this

study we were gifted FUS-GFP which constitutes the full length wild-type protein with a linker to a C-terminal EGFP.

1.5 Thesis Objective

The objective of this study was to investigate how the intrinsic physical properties of mRNA influence it's partitioning in BMCs. We aimed to utilise deep sequencing, in combination with bioinformatics to establish how these properties defined mRNA selection in a panel of simplified *in vitro* BMCs, and to determine whether the same principles would also apply in *in vivo* systems. Additionally, we sought to investigate the role of 3'UTR-APA in shaping this process. Studies have often investigated this phenomenon from a protein perspective, but here we wanted to focus on the mRNA itself. In doing so, we endeavoured to gain a unique insight into the driving forces behind selective partitioning of mRNA in BMCs.

Chapter 2

Materials and Methods

2.1 Materials and Methods

2.1.1 Cell Culture

Human primary adenocarcinoma cells (SW480) were grown in DMEM (10 % FCS, 1 % Penicillin / Streptomycin). At 60-80 % confluency, these cells were detached using TrypLE Express trypsin (Gibco) and diluted into fresh cell culture media. *Drosophila* Schneider 2 (S2) cells were grown in Schneider's Drosophila Medium (10 % heat-inactivated FCS, 1 % Penicillin / Streptomycin). At 80-90 % cells were detached by pipetting up and down.

2.1.2 Subcellular RNA Fractionation

Cells were detached from the culture flask as described above. They were washed twice with ice-cold PBS and then pelleted by centrifugation at 1000 g for 5 minutes. The pellet was re-suspended in 1 mL lysis buffer (10 mM

Tris, 1.5mM $MgCl_2$, 140mM NaCl, 0.5% NP-40). The lysate was centrifuged at 1000 g for 3 minutes at 4°C. 500 μL of the supernatant was kept, representative of cytoplasmic isolate. This was centrifuged at 12,000 g for 1 minute at 4°C. 500 μL Trizol® was added and the sample was left at room temperature for 5 minutes. 100 μL chloroform was added and the mixture was vortexed vigorously prior to centrifugation at 12,000 g for 15 minutes at 4°C. The upper aqueous phase containing the RNA was extracted and the RNA precipitated in 750 μL isopropanol at -20 °C for a minimum of 1 hour. The resulting RNA pellets were DNase 1-treated and RNase-out treated (Roche) as per manufacturers instructions. RNA was purified by phenol-chloroform extraction and precipitated in 2.5x volume ethanol, 10 % sodium acetate and 1 μL glycogen for a minimum of 1 hour at -20 °C. The sample was subsequently centrifuged at 13,000 RPM for 30 minutes at 4 °C. The pellet was resuspended in RNase-free water and the concentration was determined using a NanoDrop ND1000.

2.1.3 Denaturing agarose RNA Gel Electrophoresis

A 1 % denaturing formaldehyde/agarose gel (1 x formaldehyde gel running buffer, 3.3 % formaldehyde, 1 % agarose) were used to determine the quality of the cytoplasmic RNA obtained by subcellular fractionation. 2 μg of RNA was mixed with formaldehyde gel running buffer, formaldehyde and formamide. The samples were incubated at 65 °C for 15 minutes. Gels were run for 2 hours at 65 V using 1 x formaldehyde running buffer. Gels were visualised with UV and imaged using a Kodak Gel Camera system.

2.1.4 Protein Expression

DDX4N

The DDX4N expression plasmid (an IPTG-inducible pET SUMO vector) was a gift received from Timothy Nott (Department of Biochemistry, University of Oxford). The protein was expressed overnight at 20 °C, using *E. coli* BL21-CodonPlus (DE3)-RIL Competent Cells from Agilent. Cell pellets were resuspended in buffer (20 mM disodium phosphate, 6 mM GdmCl, 10 mM imidazole, 3 mM DTT, pH 8.0) and lysed by microfluidiser. The his-tagged protein was purified by metal affinity chromatography (HisTrap^{TM}; Merck or HisPur^{TM} Ni-NTA Resin). The tag was cleaved with ULP1 protease and the construct further purified by metal affinity chromatography and buffer-exchanged by size-exclusion chromatography into an experimental buffer (20 mM Tris, 300 mM NaCl, 5 mM TCEP, pH 8.0). Purified protein was concentrated to 325 μM, flash-frozen in liquid nitrogen and stored at -80 °C. The sequence for the purified protein construct is displayed in table 2.1.

EWSR1C

The EWSR1C expression plasmid (an IPTG-inducible pETM-30 vector) was a gift received from Timothy Nott (Department of Biochemistry, University of Oxford). The his-GST tagged protein was expressed overnight at 20 °C, using *E. coli* BL21-CodonPlus (DE3)-RIL Competent Cells from Agilent. Cell pellets were resuspended in buffer (50 mM Tris, 500 mM NaCl, 20 mM imidazole, 2 mM TCEP, pH 8.0) and lysed by microfluidiser. The protein was purified using a GST spin column (Pierce ^{TM}: from Sigma-Aldrich). The tag

was cleaved with TEV protease and the construct further purified by metal affinity chromatography and buffer-exchanged by size-exclusion chromatography into an experimental buffer (20 mM Tris, 300 mM NaCl, 5 mM TCEP, pH 8.0). Purified protein was concentrated to 225 μM, flash-frozen in liquid nitrogen and stored at -80 °C. The sequence for the purified protein construct is displayed in table 2.1.

FUS-GFP

FUS-GFP was provided as a gift by the Hyman group (Max Planck Institute of Molecular Cell Biology and Genetics). The protein was at a concentration of 170 mM in a high-salt buffer (50 mM HEPES, 500 mM KCl and 5% glycerol, at pH 7.25). The sequence for the purified protein construct is displayed in table 2.1.

2.1.5 RNA Partitioning Experiments

At 4 °C, 10 μg of cytoplasmic RNA in appropriate non-salty buffer was added to each model condensate to induce phase separation. The mixture was centrifuged at 1000 g at 4 °C to separate the two phases. The dilute and dense phase were extracted by pipette. RNA from each phase was purified by phenol-chloroform extraction and precipitated in 2.5x volume ethanol, 10 % sodium acetate and 1 μL glycogen for a minimum of 1 hour at -20 °C. The RNA pellets were resuspended in 6 μL RNase-free water. 1 μL was used for determining the RNA concentration with the NanoDrop ND1000. The remaining 5 μL was used for library preparation.

2.1.6 Hanging Drop Assays

Hanging drop assays were performed by applying 2 μL of sample to a cover slip and sealing it above a well in a 24-well plate with vacuum grease. Each well contained 1 mL of appropriate buffer. The drops were imaged at various time points by DIC microscopy as described in section 2.1.20.

2.1.7 EWSR1C ThT Amyloid Assay

The ThT assay was performed using a FLUOstar Omega spectrophotometer (BMG LabTech). Excitation was measured at 440 nm and emission at 485 nm, and data was recorded every 60 s. Orbital shaking at 300 rpm was performed for 20 seconds after each cycle.

2.1.8 QuantSeq 3' mRNA Sequencing

QuantSeq 3'mRNA-Seq Library Preparation

Library preparation was performed utilising the Lexogen QuantSeq 3' mRNA-Seq kit as per the manufacturer's instructions using the 5 μL samples obtained from the RNA partitioning experiments. 14-16 PCR cycles were used for amplification.

hsDNA Bioanalyzer Assay

Quality control for the libraries generated from QuantSeq 3'mRNA-Seq library preparation was performed with a high sensitivity DNA (hsDNA) chip on the Agilent 2100 Bioanalyzer as per manufacturer's instructions. Here library concentrations were also assessed.

RNA-seq on the Ion Proton Platform

3'QuantSeq libraries were diluted to 350 pM for use with the Ion PI™ Hi-Q™ Chef Kit and ion Chef™ system as per the manufacturer's instructions. For human samples, reads were aligned to the hg19 genome using the Ion Torrent Server TMAP aligner with default settings. For *Drosophila* samples reads were aligned to the dm6 genome using the Ion Torrent Server TMAP aligner with default settings.

2.1.9 Full RNA-seq Data Acquisition

FASTQ files for PB (Hubstenberger *et al.* 2017) and SG (Khong *et al.* 2017) full RNA-seq data sets were obtained using the SRA accession numbers via the "Faster Download and Extract Reads in FASTQ" function on Galaxy (Afgan et al. 2018). The single-end PB FASTQs were were aligned using HISAT2 on galaxy with default settings (Yun Zhang et al. 2021). For the paired-end SG data, the reverse FASTQ files were aligned using HISAT2 on galaxy with default settings (Yun Zhang et al. 2021).

2.1.10 Integrative Genomics Viewer (IGV)

IGV was used to visualise normalised reads obtained from 3'RNA-seq or full RNA-seq (Robinson et al. 2011).

2.1.11 Counts Table Generation

For the 3'RNA-seq data, human cleavage and polyadenylation (CPA) site annotations were obtained from PolyA_DB 3 and *Drosophila* CPA sites were

obtained from the Tian laboratory (R. Wang, Nambiar, et al. 2018; X. Liu et al. 2017). CPA windows were made by extending 20 nucleotides (nt) 3' and 200 nt 5' from the cleavage site. Where windows overlapped on the same strand they were combined into a single CPA annotation. Mapped reads were refined to their 3' most nt and these were overlapped with the CPA windows to obtain counts for each. For the full RNA-seq data, a GTF file containing exon positions for the hg19 genome was obtained from UCSC (Lander et al. 2001). Reads were refined to their 3' most nt and overlapped with the exon positions to obtain counts for each.

2.1.12 Analysis Pipelines

Gene-specific Partitioning Pipeline

For the 3'RNA-seq data, counts for each gene were obtained by combining the counts from CPA windows that belonged to each gene. For the full RNA-seq data, counts for each gene were obtained by combining the counts for each exon belonging to each gene. Differential partitioning was then analysed using the DESeq algorithm within the DESeq2 R package (Love, Huber, and S. Anders 2014). This produces a table that includes normalised read counts, $\log_2$ fold change in RNA levels between the two samples being analysed, p-value evaluated by a Walt test, and adjusted pvalue (padj) acquired using the Benjamini-Hochberg method, for each gene.

3'UTR-APA Partitioning Pipelines

For the 3'RNA-seq data, the CPA sites counts table was used for differential 3'UTR-isoform partitioning analysed using the DEXseq algorithm (Reyes et al. 2013). This generates a table with normalised read counts, $\log_2$ fold change, pvalue determined by a Wald test, and padj value acquired using the Benjamini-Hochberg method, for each CPA site. The relative isoform abundance (RA) value for each gene was calculated using equation 4.1. The two most significant CPA sites were retained for each gene for downstream analysis. For the full RNA-seq data, the Apalyzer program was utilised for 3'UTR-APA analysis (R. Wang and B. Tian 2020). This generates a table containing the relative expression difference (RED), calculated using equation 4.2, which indicates the change in isoform abundance between the two samples being analysed for each gene, p-value determined by a Fisher's exact test, and a false discovery rate (FDR) adjusted p-value for each gene.

2.1.13 mRNA Sequences

The sequence for the longest mRNA transcript for each gene from UCSC using the TxDB and BSgenome packages in R, for human and *Drosophila* (Pagès 2018; Carlson and Maintainer 2015b; Carlson and Maintainer 2015a). Isoform sequences were approximated by combining the longest 5'UTR sequence, longest coding sequence, and the sequence from either the end of the coding sequence (CDS) to CPA site 1 (short isoform), or the sequence from the end of the CDS to CPA site 2 (long isoform), using the aforementioned packages in R. Length, GC % and GQFRs were calculated using these

sequences.

2.1.14 *Drosophila* Orthologs

Drosophila orthologs for human genes were obtained from DIOPT (Y. Hu et al. 2011). Orthologs with a DIOPT score of less than 3 were excluded.

2.1.15 GQFR Analysis

GQFR analysis was performed in python utilising a pipeline created by Andrew Baldwin incorporating the QGRS Mapper (Kikin, D'Antonio, and Bagga 2006). A G-score (an output of the QGRS mapper which indicates how well matched a sequence is to the GGGGGGTGGGGGGTGGGGGGTGGGGGG consensus sequence) of greater than 60 was required in order to be counted as a GQFR.

2.1.16 Random Sampling Analysis

In the general method, we used random sampling to generate 5000 %s that were equivalent to our observed %. We obtained the mean, standard deviation from this data set and calculated the Z score and p-value corresponding to the likelihood our observed value was due to random chance. Normal distributions were plotted using these results. This was performed in python.

Human and *Drosophila* Orthologs Same Partitioning

Tables were generated that were the same length as the number of orthologs that were present in the human and *Drosophila* refined gene-partitioning

data sets. The tables had two columns, one with partitioning preferences randomly sampled from the *Drosophila* ortholog partitioning preferences, and the other from the human equivalent. From these, the percentage of rows with the same partitioning preferences were calculated and used as described in Section 2.1.16.

PB and SG Proteome Components that underwent 3'UTR-APA

For each, tables were generated that were the same length as the number of proteome components by randomly sampling the corresponding 3'UTR-APA table. The percentages of genes in these tables with significant 3'UTR-APA changes were calculated and used as described in Section 2.1.16.

Genes with the same significant 3'UTR-APA changes in both PBs and SGs

Tables were generated that were the same length as the number of genes with significant 3'UTR-APA changes in both PBs and SGs. The tables contained two columns one with a significant 3'UTR-APA change randomly sampled from the list of genes with significant 3'UTR-APA changes in PBs, and the other with the SG equivalent. The percentage of rows with the same 3'UTR-APA change was calculated and used as described in Section 2.1.16.

GQFR frequency in PB and SG aUTRs

For each, tables were generated that were the same length as either the SIE or LIE cohort by randomly sampling each 3'UTR-APA data set. The

percentage of rows where transcripts contained 1+ GQFR was calculated and used as described in Section 2.1.16.

2.1.17 Streme Motif Discovery and Tomtom Motif Comparison

Streme motif discovery was performed using the Meme Suite online tool (T. L. Bailey 2021). Test sequences were compared with a set of control sequences, obtained from the same data set, of the same length. A motif width of 7 was set and default parameters were used. Motifs with a p-value of less than 0.05 were analysed with the Tomtom motif comparison software (Gupta et al. 2007; Ray et al. 2013). Identified RNA-binding protein motifs were considered significant if the p-value was less than 0.05.

2.1.18 Enrichr Gene Ontology Analysis

Gene ontology analysis was performed using Enrichr's online tool (E. Y. Chen et al. 2013; Kuleshov et al. 2016). Results with a padj of less than 0.01 were considered significant.

2.1.19 Plots

Boxplots

Boxplots were made using either the ggplot2 and ggpubr packages in R or Seaborn in python (Waskom 2021; Wickham 2016; Kassambara 2020). T-tests were used to determine significant differences between group means.

Scatterplots and Pearson Correlation Analysis

Scatterplots and Pearson correlation analysis were obtained using the ggplot2 and ggpubr packages in R (Wickham 2016; Kassambara 2020)

Bar Graphs

Bar graphs were made using the ggplot2 package in R (Wickham 2016)

2D Histograms

2D histograms were plotted in python using Seaborn (Waskom 2021). GQFR scatter plots that were overlayed were plotted using gnuplot.

Venn Diagrams

Venn diagrams were generated using the Bio Venn package in R (Hulsen and Tim 2021; T. Hulsen 2021).

2.1.20 Microscopy

Light Microscopy

Light microscopy images of DDX4N and EWSR1C condensates were obtained using a 5x/0.12 objective on an Axiovert 25 light microscope.

Differential Interference Contrast (DIC) Microscopy

DIC images of the EWSR1C, DDX4N and RNA hanging drop assays were obtained by using a 10x differential interference contrast (DIC) objective on an Olympus BX43 microscope.

2.1.21 Protein Constructs and Amino Acid Sequences

Sequences for the protein constructs employed as *in vitro* model BMCs are listed in the following table (Table 2.1):

Construct	Sequence
DDX4N	GAMGSMGDEDWEAEINPHMSSYVPIFEKDRYSGENGDNFN RTPASSSEMDDGPSRRDHFMKSGFASGRNFGNRDAGECNKR DNTSTMGGFGVGKSFGNRGFSNSRFEDGDSSGFWRESSNDC EDNPTRNRGFSKRGGYRDGNNSEASGPYRRGGRGSFRGCR GGFGLGSPNNDLDPDECMQRTGGLFGSRRPVLSGTGNGDTS QSRSGSGSERGGYKGLNEEVITGSGKNSWKSEAEGGES
EWSR1C	GAMGSMGDEDWEAEINPHMSSYVPIFEKDRYSGENGDNFN RTPASSSEMDDGPSRRDHFMKSGFASGRNFGNRDAGECNKR DNTSTMGGFGVGKSFGNRGFSNSRFEDGDSSGFWRESSNDC EDNPTRNRGFSKRGGYRDGNNSEASGPYRRGGRGSFRGCR GGFGLGSPNNDLDPDECMQRTGGLFGSRRPVLSGTGNGDTS QSRSGSGSERGGYKGLNEEVITGSGKNSWKSEAEGGES
FUS-GFP	MASNDYTQQATQSYGAYPTQPGQGYSQQSSQPYGQQSYSG YSQSTDTSGYGQSSYSSYGQSQNTGYGTQSTPQGYGSTGGY GSSQSSQSSYGQQSSYPGYGQQPAPSSTSGSYGSSSQSSSYGQ PQSGSYSQQPSYGGQQQSYGQQQSYNPPQGYGQQNQYNSSS GGGGGGGGGGNYGQDQSSMSSGGGSGGGYGNQDQSGGGG SGGYGQQDRGGRGRGGSGGGGGGGGGGYNRSSGGYEPRG RGGGRGGRGGMGGSDRGGFNKFGGPRDQGSRHDSEQDNS DNNTIFVQGLGENVTIESVADYFKQIGIIKTNKKTGQPMINLY TDRETGKLKGEATVSFDDPPSAKAAIDWFDGKEFSGNPIKV SFATRRADFNRGGGNGRGGRGRGGPMGRGGYGGGGSGGG GRGGFPSGGGGGGGGQQRAGDWKCPNPTCENMNFSWRNEC NQCKAPKPDGPGGGPGGSHMGGNYGDDRRGGRGGYDRG GYRGRGGDRGGFRGGRGGGDRGGFGPGKMDSRGEHRQDR RERPYGAPGSAGSAAGSGMVSKGEELFTGVVPILVELDGDV NGHKFSVSGEGEGDATYGKLTLKFICTTGKLPVPWPTLVTT LTYGVQCFSRYPDHMKQHDFFKSAMPEGYVQERTIFFKDD GNYKTRAEVKFEGDTLVNRIELKGIDFKEDGNILGHKLEYN YNSHNVYIMADKQKNGIKVNFKIRHNIEDGSVQLADHYQQN TPIGDGPVLLPDNHYLSTQSALSKDPNEKRDHMVLLEFVTA AGITLGMDELYK

Table 2.1: Protein constructs and sequences

Chapter 3

Investigating Selective mRNA Partitioning in Biomolecular Condensates

3.1 Introduction

A critical property of biomolecular condensates (BMCs) is their ability to enrich for certain molecules whilst excluding others. Their liquid-like properties facilitate exchange of molecules with the surrounding environment, as well as allowing for rapid formation and disassembly in response to external stimuli; SGs for example, only form under stress, and rapidly dissipate upon its removal (Wheeler et al. 2016; Khong et al. 2017). Cells need to orchestrate a vast number of chemical reactions at any given time in order to maintain homeostasis and BMCs provide an excellent means to be able to do this, in a way that can be dynamically regulated. Each condensate is unique

in composition, despite lacking a membrane, and therefore, selective partitioning of molecules must occur via means that are alternative to processes such as active transport.

These systems can be considered analogous to organic solvents; molecules which are more soluble in the condensate phase than outside are preferentially sequestered there, and vice versa. As outlined in the Introduction, many BMCs including PBs and SGs are enriched in specific RNA and proteins that are functionally associated with post-transcriptional regulation of mRNA. To date, a number of fundamental questions about these so-called RNA granules are being intensely investigated: What subset of mRNAs do they partition? What is the precise way they compartmentalise this subset of mRNAs? and what are the biological consequences of partitioning them? This is particularly important in the context of RNA granule disregulation which, as previously discussed, is the hallmark of many neurodegenerative diseases (Savas et al. 2008; A. Patel et al. 2015). Consequently, elucidating their specific functions and how they enact them can not only provide essential information about RNA metabolism, but also can provide critical insight into human diseases and how we may treat, or prevent them.

In vivo, RNA granules contain highly complex mixtures of many different RNAs. Moreover, as discussed in the Introduction, a multitude of factors can influence RNA selection, including the length, sequence, and structure of an RNA molecule (See Figure 1.11). In an attempt to reduce this complexity, studies have often employed short, non-native, homotypic, RNA polymers. Such molecules tend to be single-stranded and often lack any secondary structure, and thus, inferences drawn from such experiments must be treated with

caution with respect to the mechanisms that might underpin *in vivo* BMC assembly and dynamics (Boeynaems et al. 2019; Roden and Gladfelter 2021). On the other hand, results obtained from the isolation of RNA granules from cells can often be very difficult to interpret owing to the fact that a multitude of possible factors might contribute to any specific observation. In order to bridge the gap between overly simplistic and overly complex, we have developed an *in vitro* approach to specifically address how the physical properties of mRNA may account for its selection into BMCs. We employ protein-free cytoplasmic RNA and combine it at physiological pH and salt conditions with three recombinant proteins to induce formation of dual-phase systems. RNA from the protein dense and dilute phases is then extracted and subjected to 3'RNA-seq to identify mRNA that are selectively partitioned in either phase. This approach maintains cellular complexity at an mRNA level, whilst at the same time, reducing the complexity associated with the multimeric protein nature of *in vivo* condensates.

The proteins used in this study as model BMCs were an IDR from the N-terminal domain of the dead-box helicase DDX4 (DDX4N), an IDR from the C-terminal domain of the RBP EWSR1 (EWSR1C), and the full length, RBP FUS, fused with Green Fluorescent Protein (GFP) (FUS-GFP) (Figure 3.1). All are derived from wild-type proteins that are known to be key components of various biomolecular condensates *in vivo*, and in the case of EWSR1 and FUS, implicated in human disease. The phase diagram of DDX4N has been extensively studied, and importantly, the physical characteristics of the *in vitro* condensates are indistinguishable from those formed *in vivo* in cells (Nott, Petsalaki, et al. 2015). Critically, the recombinant DDX4N fragment

lacks any recognised RNA-binding domain (RBD) and thus any potential selectivity for specific mRNAs is independent of corresponding sequence motifs. This is of considerable importance, as mRNA selection into BMCs is largely attributed to to the presence of specific motifs that are recognised by RBPs, that in turn carry bound mRNAs as passengers into condensate. EWSR1C was chosen as an orthogonal condensate model system to DDX4N. At pH 8, DDX4N is negatively charged whilst EWSR1C carries a positive charge. Similar to DDX4N, recombinant EWSR1C features an IDR that lacks a recognised RBD. Although phase separation of EWSR1C has not been previously characterised it offers the opportunity to assess the importance of charge in mRNA selectivity. In contrast to DDX4N and EWSR1C protein fragments, recombinant FUS-GFP corresponds to wild-type full length FUS protein with a C-terminally fused GFP. It retains its native RNA-recognition motif, which binds a variety of mRNA molecules (X. Wang, Schwartz, and Cech 2015).

We employed the three recombinant proteins in combination with human and *Drosophila* cytoplasmic RNA to form *in vitro* model RNA-protein condensates, to establish whether they create selective solvents, and to identify the intrinsic features of mRNA that confer selectivity. We compared our findings with those from *in vivo* systems to determine whether the trends we had identified persisted.

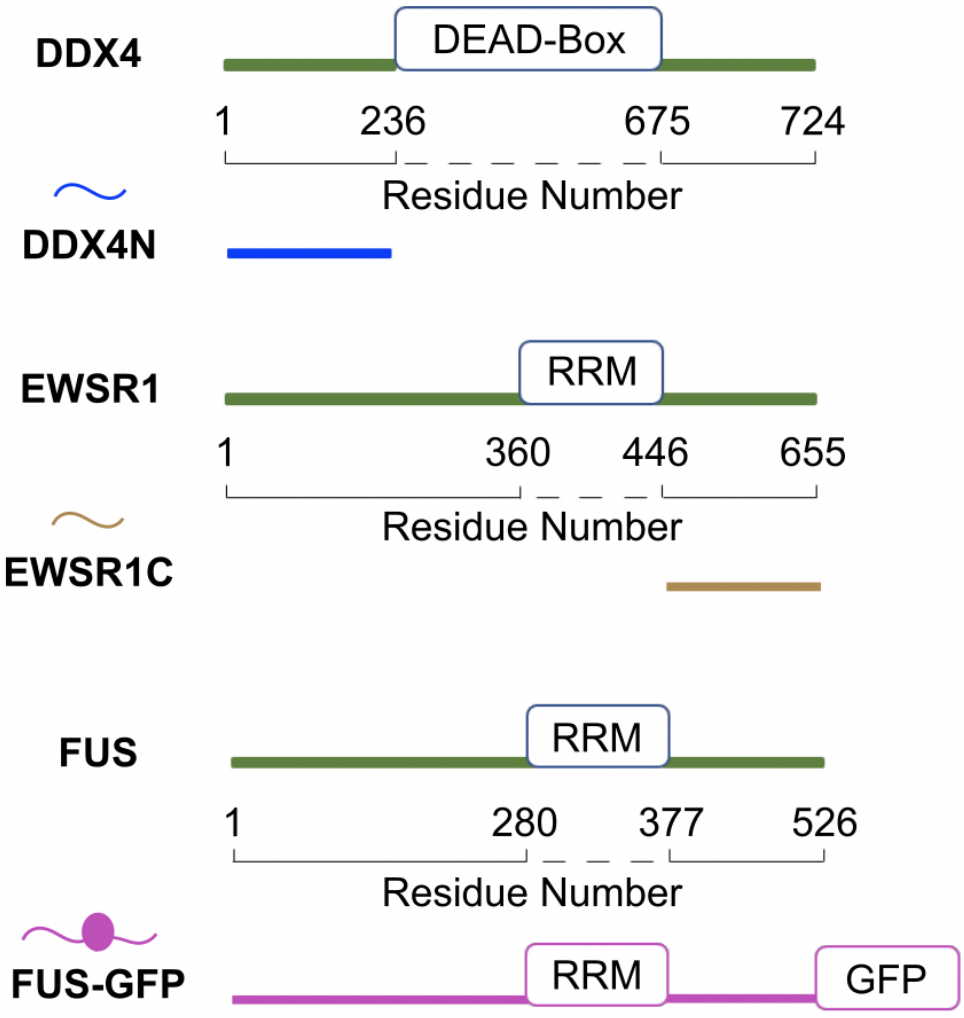


Figure 3.1: Model BMCs used for *in vitro* mRNA partitioning assays. Schematics depicting the three constructs, DDX4N, EWSR1C and FUS-GFP, along with their wild-type equivalents are shown.

3.1.1 *In vitro* mRNA partitioning approach

EWSR1C and DDX4N were expressed using bacterial vectors and FUS-GFP was a gift from the Hyman group (Max Planck Institute of Molecular Cell Biology and Genetics). Cytoplasmic RNA was extracted from SW480 cells (a

human colorectal cancer cell line), or from *Drosophila melanogaster* Schneider cells (S2). Cytoplasmic RNA was used in order to ensure that the mRNA had been fully processed as this would be important for downstream analysis.

Our workflow begins with a solution containing 10 µg of cytoplasmic RNA that is added to a single-phase protein solution at 4°C, resulting in the formation of a dual-phase protein-RNA solution (Figure 3.2A,B). The two phases are separated by centrifugation (Figure 3.2C), and RNA is isolated from the dilute and dense phase respectively (Figure 3.2D). Extracted RNA was subjected to 3'RNA-seq on the ion proton sequencing platform and mapped to the hg19 human genome (or the dm6 Drosophila melanogaster genome), using the inbuilt Torrent Mapping Alignment Software (TMAP) to generate BAM files containing aligned reads for each of the RNA cohorts, to be used for downstream analysis (Figure 3.2E). Two replicates were performed for each individual partitioning experiment.

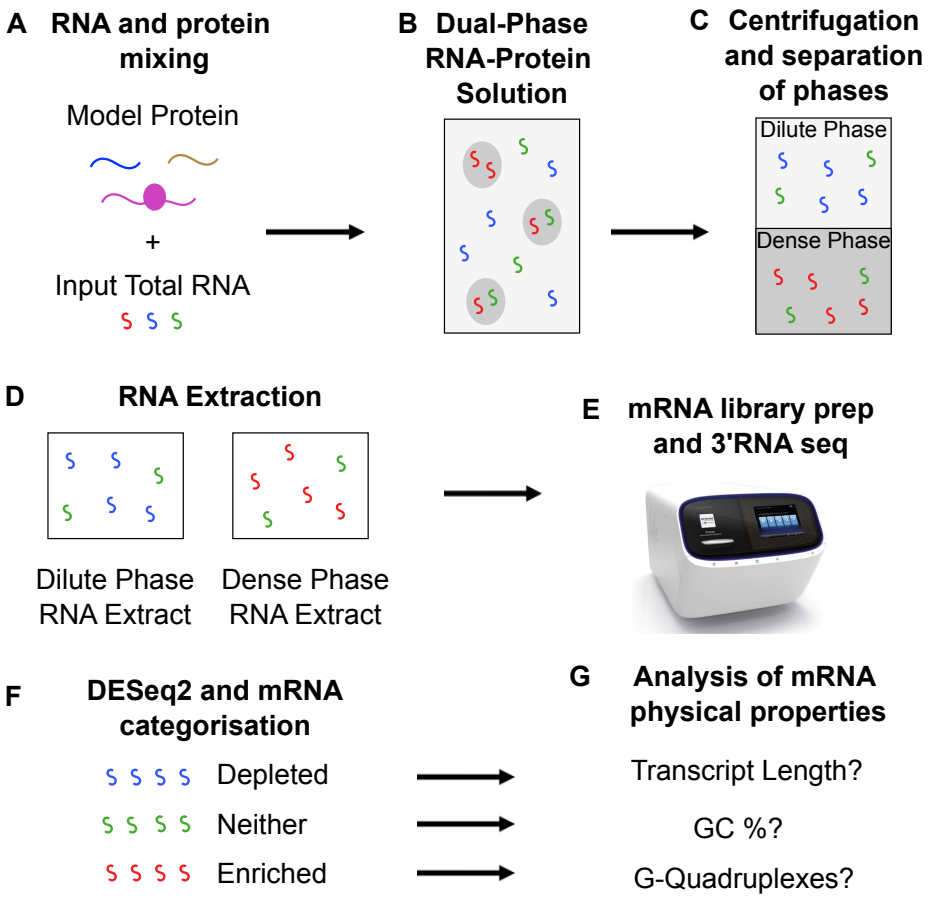


Figure 3.2: Schematic showing mRNA partitioning approach. A) A single-phase model protein solution is mixed with cytoplasmic RNA B) A dual-phase solution is yielded from the mixing step. C) The dilute phase and the dense phase in the solution are separated by centrifugation. D) The dilute phase and the dense phase are extracted and the RNA from each of the two phases is then subsequently extracted and purified. E) The purified RNA from each of the samples is then used for mRNA library prep and subsequent 3'RNA seq using the Ion Torrent platform. F) DESeq2 is performed using the dilute phase BAM files as the reference and the dense phase BAM files as the test, in order to determine whether transcripts are depleted, neither or enriched in the dense phase. G) These cohorts of transcripts are then analysed on the basis of their physical properties, looking specifically at transcript length, GC % and G-quadruplex forming regions.

3.1.2 Establishing a pipeline to analyse gene-specific partitioning of mRNA in BMCs

The aim of the partitioning experiment was to analyse how the physical properties, including length, GC %, and G-quadruplex forming regions (GQFRs) of an mRNA might influence its selection into our model BMCs. To this end we designed a bioinformatics "partitioning" pipeline to determine a) mRNAs were preferentially partitioned into the model condensates (the "enriched" cohort), b) which mRNAs exhibited no partitioning preference at all (the "neither" cohort) and c) which mRNAs were preferentially excluded (the "depleted" cohort). Briefly, the analysis pipeline begins with the intersection of the mapped BAM files with a table that contained the coordinates for regions surrounding cleavage and polyadenylation sites (CPA sites) in the genome (human or *Drosophila*), according to the PolyA DB (R. Wang, Nambiar, et al. 2018). The output provides the number of reads that mapped to all mRNA isoforms that were present in the isolates. Reads for all mRNA isoforms belonging to the same gene were combined into a counts table that contained the total raw reads for each gene across all our samples. To identify mRNAs that are differentially distributed in the two phases, the tables were analysed with DESeq2 (Love, Huber, and S. Anders 2014). Genes which had less than five normalised reads per sample were excluded from the resulting output table. Genes which displayed a positive $\log_2$ fold change (LFC) and had a p-adjusted value (padj) of less than 0.01 were classified as "enriched". Genes which displayed a negative LFC with a padj value of less than 0.01 were classified as "depleted". Genes which had an LFC

of between -0.1 and 0.1, and additionally, a standard deviation divided by mean normalised read numbers across each gene (σ/μ) that was less than 0.5, were classified as "neither" (Figure 3.2F). The physical properties including GC content, length and G-quadruplex content of the mRNA partitioning cohorts were subsequently interrogated using the longest mRNA transcript annotation (UCSC) for each gene. GQFRs were predicted using a pipeline incorporating QGRS mapper (Figure 3.2G) (See methods). As discussed in the introduction, these three properties were interrogated because of their potential to influence mRNA selection into biomolecular condensates.

3.2 Results

3.2.1 DDX4N

Design of the DDX4N mRNA partitioning experiment

DDX4N was concentrated to 325 μM during purification and stored in a high salt buffer (300 mM NaCl). A gel for the final purified protein product is shown in the Appendix (Figure A.1). Phase separation can be induced by the addition of buffer with no salt, reducing the salt concentration to 100 mM (Figure 3.1A) (Nott, Petsalaki, et al. 2015). In the partitioning experiment this buffer contains cytoplasmic RNA, however, it was unknown how this may affect the stability of DDX4N condensates. Therefore, in order to investigate this, we challenged DDX4N with three different input masses of cytoplasmic RNA; 0 μg, 10 μg and 20 μg. As a proxy for stability, the mass of protein in the dense phase was measured using the Nanodrop. Our results showed

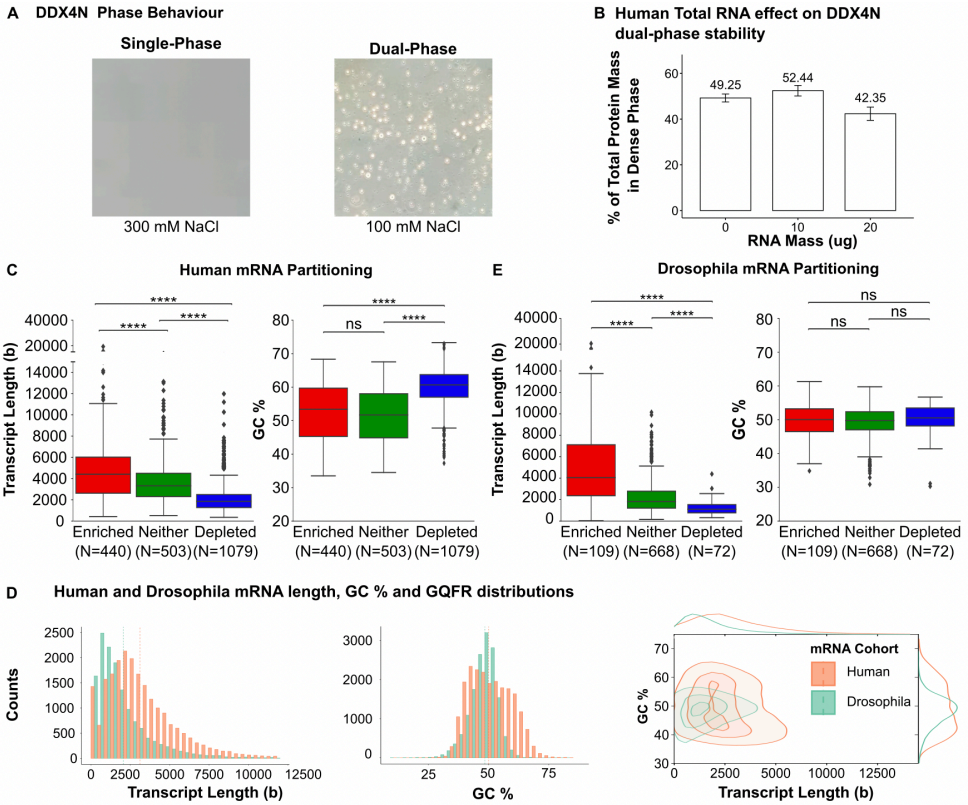
no reduction in stability with the addition of 10 μg of cytoplasmic RNA. However, addition of 20 μg was shown to lower the stability (Figure 3.1B). 10 μg cytoplasmic RNA was therefore decided as the optimal input mass for our partitioning experiment.

DDX4N partitions human mRNA on the basis of its length and GC %

We first challenged DDX4N with human mRNA and utilised the approach described in section 3.1.2 to establish three mRNA partitioning cohorts. Strikingly, we found significant differences in the physical properties of human mRNA, which dictated which partitioning cohort they were found in. Transcripts that were enriched were significantly longer (mean of 4769.8 b) than those that were neither (mean of 3669.8 b) and those that were depleted (mean of 2065.8 b). Similarly, transcripts that were depleted were also significantly shorter than those that were neither. Analysis of the GC % revealed that there was no significant difference between the enriched (mean of 52.4 %) and the neither (mean of 51.5 %) but that both were significantly lower in GC % than the depleted cohort (mean of 60.1 %) (Figure 3.3C). GC % can be considered as a proxy for secondary structure, since mRNAs with a higher GC % form more stable secondary structures (Chan et al. 2009). Together, these results demonstrate that DDX4N enriches for long mRNAs whilst excluding short, highly structured mRNAs.

DDX4N partitions *Drosophila* mRNA based on its length

To interrogate whether the trends that we saw in length and GC % would be applicable to any generic pool of RNA, and not specific to that particular experiment, we challenged DDX4N with cytoplasmic RNA extracted from *Drosophila* cells. We hypothesised that mRNA with similar physical properties would partition in the same way, independent of the species. Furthermore, the human and *Drosophila* transcriptomes differ significantly, particularly in their GC % distributions, and as such utilising *Drosophila* RNA provided an excellent comparison with which to further examine this concept (Figure 3.3D). The human transcriptome has a much wider GC % distribution, split between two peaks at approximately 41 and 57 % GC, whilst the *Drosophila* transcriptome has a much smaller spread, with a main peak at approximately 49 %. Our findings revealed, that like human mRNA, *Drosophila* mRNA also partitioned with a clear dependency on the length of the transcript, with longer mRNAs being significantly enriched (Figure 3.3E). Interestingly, our analysis of the GC % of transcripts showed that there was no significant difference between any of the partitioning cohorts for the *Drosophila* RNA. However, this result is likely explained by the GC % distribution in the *Drosophila* transcriptome being much narrower than in humans.


A DDX4N Phase Behaviour
Single-Phase
Dual-Phase
300 mM NaCl
100 mM NaCl
B Human Total RNA effect on DDX4N dual-phase stability
% of Total Protein Mass in Dense Phase
49.25
52.44
42.35
RNA Mass (ug)
C Human mRNA Partitioning
Transcript Length (b)
GC %
Enriched (N=440)
Neither (N=503)
Depleted (N=1079)
E Drosophila mRNA Partitioning
Enriched (N=109)
Neither (N=668)
Depleted (N=72)
D Human and Drosophila mRNA length, GC % and GQFR distributions
Counts
Transcript Length (b)
GC %
mRNA Cohort
Human
Drosophila

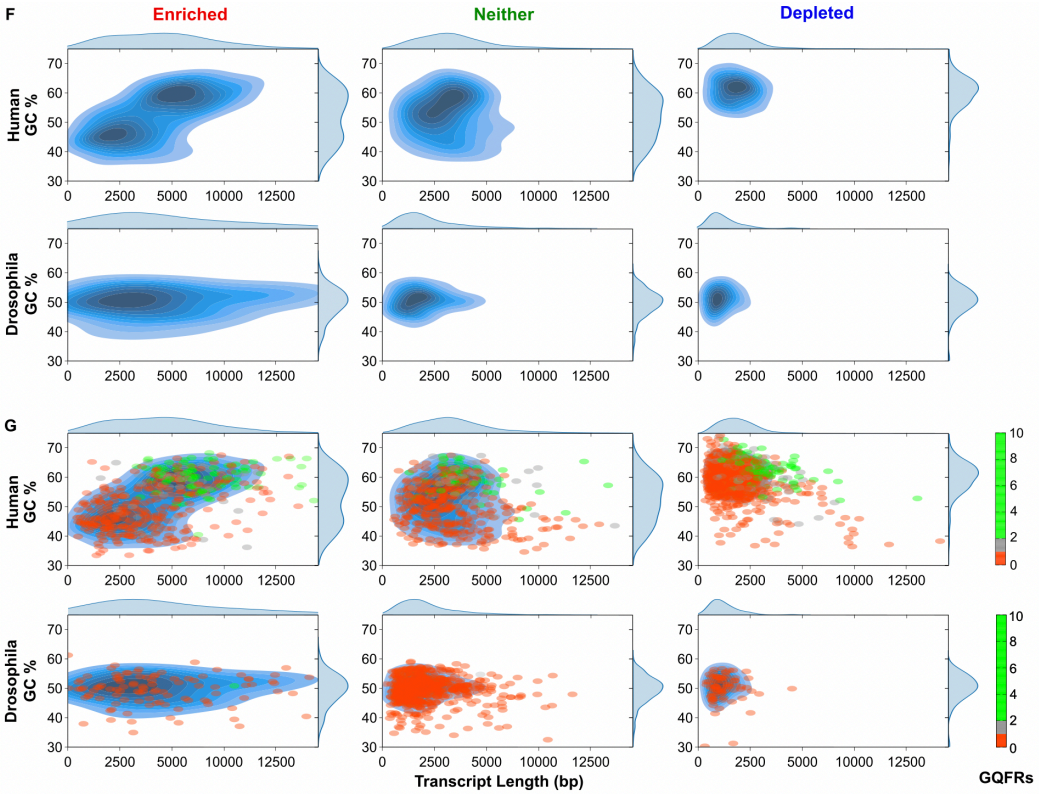


Figure 3.3 *(previous page)*: Selective partitioning of mRNA in DDX4N. A) Light microscopy images displaying DDX4N in a single-phase solution at high salt concentrations (300 mM NaCl) and DDX4N in a dual-phase solution at low salt concentrations B) The stability of the DDX4N dense phase, evaluated by measuring the mass of the protein in the dense phase when different masses of cytoplasmic RNA were added with the mean displayed above each group C) Boxplots depicting transcript length and GC % for each of the three mRNA partitioning cohorts in the Human mRNA partitioning experiment D) Comparison between Human and *Drosophila* transcriptomes displaying the transcript length (b) and the GC %. The 2D histogram displays the length and GC % of the two transcriptomes overlayed. E) Boxplots depicting transcript length and GC % for each of the three mRNA partitioning cohorts in the *Drosophila* RNA experiment. F) 2D histograms displaying GC % vs transcript length for each of the three mRNA partitioning cohorts for the Human and *Drosophila* experiments. G) 2D histograms from F) displayed with overlayed points corresponding to individual genes, coloured according to the number of GQFRs they contain; 0 GQFRs = red, 1 GQFR = grey, 2+ GQFRs = green. Significance values displayed above C and E were acquired using a T-test; ns = p > 0.05, * = p <= 0.05, ** = p <= 0.01, *** = p <= 0.001, **** = p <= 0.0001.

G-quadruplex forming regions promote enrichment of mRNA in DDX4N

We demonstrated for both the human and *Drosophila* RNA cohorts that the length of the transcript was a key determinant in how it partitioned in DDX4N; longer transcripts were enriched, whilst shorter transcripts were depleted. However, we found that *Drosophila* mRNA selection appeared independent of its GC %. To interrogate this further, we combined the length and GC % metrics to generate density plots for each of the partitioning cohorts in both data sets. This revealed some key insights that weren't apparent from examination of each individually. In the human data set, the enriched cohort contained two sub-populations of transcripts; one which was AU-rich and short (average length of approximately 2500 b and GC % of approximately 45 %) and one which was GC-rich and long (average length of approximately 5000 b and an average GC % of approximately 60 %)(Figure 3.3F). Only one population of transcripts was observed in the enriched cohort in the *Drosophila* data set.

We next investigated how GQFRs might influence DDX4N mRNA selection. As described previously, these structures can affect how mRNA partitions in BMCs. In addition, RGG-rich domains have been shown to associate with RNA G-Quadruplexes and DDX4N contains four RGG motifs (Imperatore et al. 2020). Since G-quadruplexes are formed from guanine tetrads, we thought that the GC % and the length of an mRNA might be influential in the number of GQFRs it contained. To investigate this, we performed Pearson correlation analysis of the GQFR-frequency, against the length and GC

% obtained from the longest transcript for every human gene. We found that GC % correlated with an R value of 0.36 and length only correlated with an R value of 0.24 (Figure A.2). Given that both the correlations were fairly weak, we determined that GQFRs could be considered as an independent analysis metric. Initially, we examined the GQFR-frequencies of the overall human and *Drosophila* transcriptomes and found that 47.0 % of human mRNAs contained at least one GQFR, whilst for *Drosophila*, this number was only 4.2 %. Interestingly, inspection of the partitioning cohorts in both revealed that enriched transcripts were enriched in GQFRs relative to their respective data sets. In the human data set, we found that in the enriched cohort 59.5 % of transcripts contained at least one GQFR, whereas only 47 % of genes in the total data set contained at least one. For the *Drosophila* data set, we observed a similar result; 13 % of the enriched mRNAs contained at least one GQFR, whilst this value was only 4 % for the total data set (Figure A.3). We overlayed these results with the 2D histograms from Figure 3.3F and made the remarkable observation that the two distinct populations within the enriched cohort in the human data set could be distinguished by the number of GQFRs they contained (Figure 3.3G). The AU-rich, short population were GQFR-poor, whilst the GC-rich, long population were GQFR-rich. These findings strongly suggest that GQFRs play a role in determining selective partitioning in DDX4N.

DDX4N sequesters functionally related mRNAs

We had demonstrated that DDX4N selected for transcripts on the basis of their physical properties. Next, we wanted to explore if DDX4N enriched for

human and *Drosophila* mRNAs with similar functionalities. Gene ontology (GO) analysis, using Enrichr, revealed remarkable similarities in the function of genes, in the enriched and depleted mRNA cohorts, between the human and *Drosophila* data sets (Figure 3.4) (E. Y. Chen et al. 2013; Kuleshov et al. 2016; Z. Xie et al. 2021). In both depleted cohorts, the most significantly enriched term for GO biological processes was "translation", whilst in the enriched cohorts, this term was "gene expression". "Regulation of translation" was also in the top five most significant terms in the enriched cohort for both data sets. GO cellular components analysis again demonstrated clear similarities, where the most significant result for the enriched cohort in both instances was "nucleus", whilst in the depleted cohorts we saw terms involved with the mitochondria and ribosome. Most strikingly, we saw the term "P granule" for the enriched cohort in *Drosophila*. This is the *Drosophila* equivalent of human germ granules, of which a DDX4 homolog, VASA, is a key component (Raz 2000). These results suggest that DDX4N not only acts as a solvent to sequester mRNA on the basis of its physical properties, but also in a manner that is related to its function.

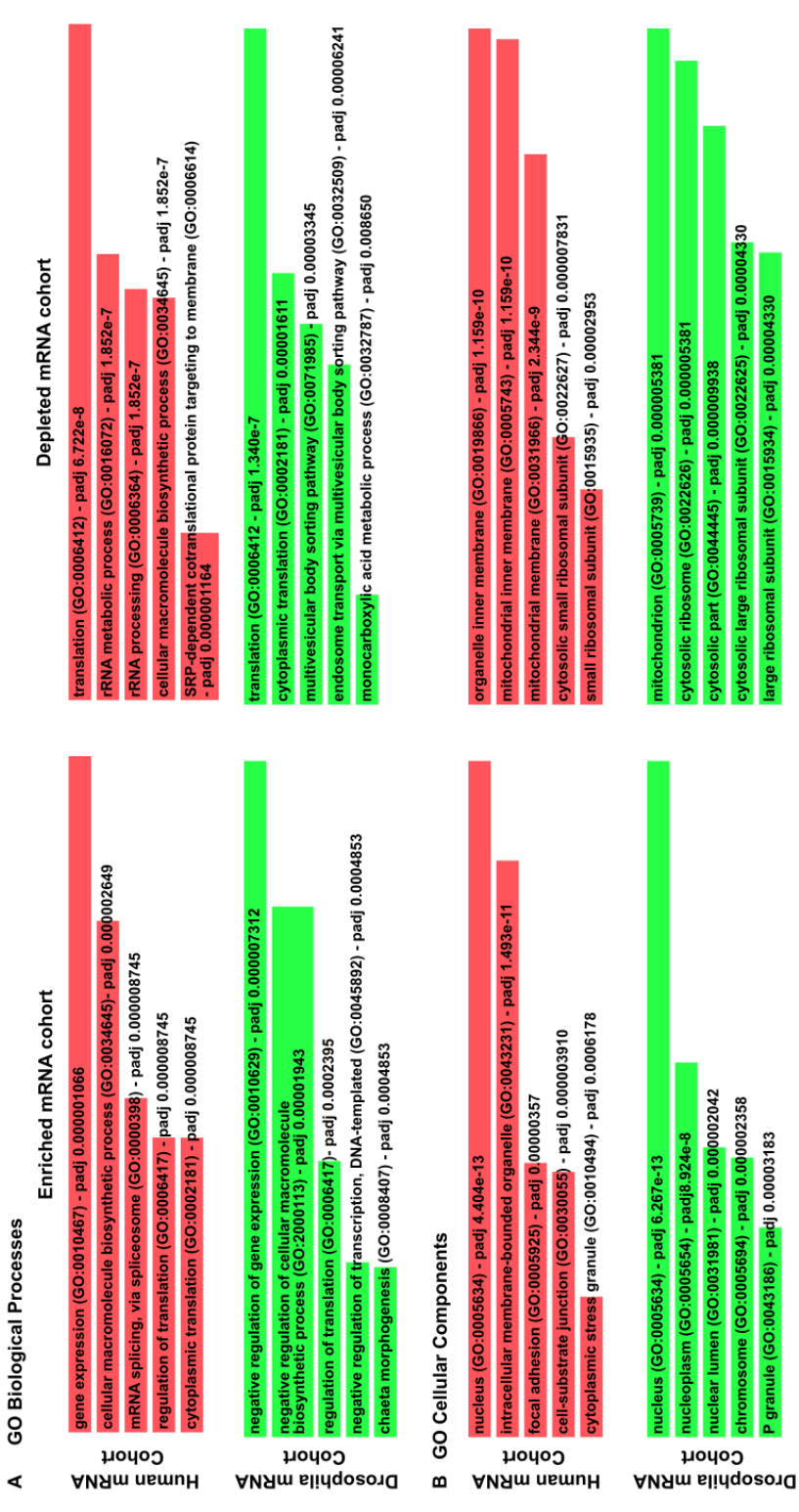


Figure 3.4: Gene ontology analysis of mRNA partitioning cohorts in DDX4N. The top five results for each gene ontology analysis are displayed along with their respective p-adjusted (padj) values with transcripts grouped by species and mRNA partitioning cohort. The size of the bar for each GO term is a reflection of the significance of the result, with the most significant result having the largest bar in each group. A) Gene ontology analysis of biological processes. B) Gene ontology analysis of cellar components.

To investigate this concept further, we explored the partitioning behaviour of orthologous mRNA in each data set. We obtained a list of human genes and their *Drosophila* orthologs using the DIOPT ortholog finder, removing any orthologs with a DIOPT score of less than 3 (Y. Hu et al. 2011). We used this to create a table that displayed how each ortholog was partitioned in their respective data set. We then compared each of the partitioning cohorts, and found 36.6 % of orthologous mRNA partitioned in the same way in both data sets (Figure 3.5A). We tested the significance of this value by generating similar tables through random sampling and found that this was significantly higher than the mean, with a p-value of 0.0015 (Figure 3.5B) (See methods for more details). This demonstrates that this observation is highly unlikely to be due to random chance, suggesting orthologous mRNAs are more likely to partition in the same way, relative to non-orthologous mRNAs. This result provides further evidence that DDX4N likely sequesters functionally-related transcripts. Next, we examined the physical properties of orthologs with and without the same partitioning behaviour. We found that orthologs that partitioned in the same way correlated more strongly on transcript length (R = 0.66) than those that didn't (R = 0.60). The same was true of the GC % although the correlations were both very weak (R=0.17 and R=0.1 respectively) (Figure A.4). Overall, this analysis raises some interesting possibilities. Since orthologs were more likely to partition in the same way, this suggests that mRNA partitioning into BMCs may be evolutionarily conserved, in manner that is determined by the intrinsic properties of the mRNA itself.

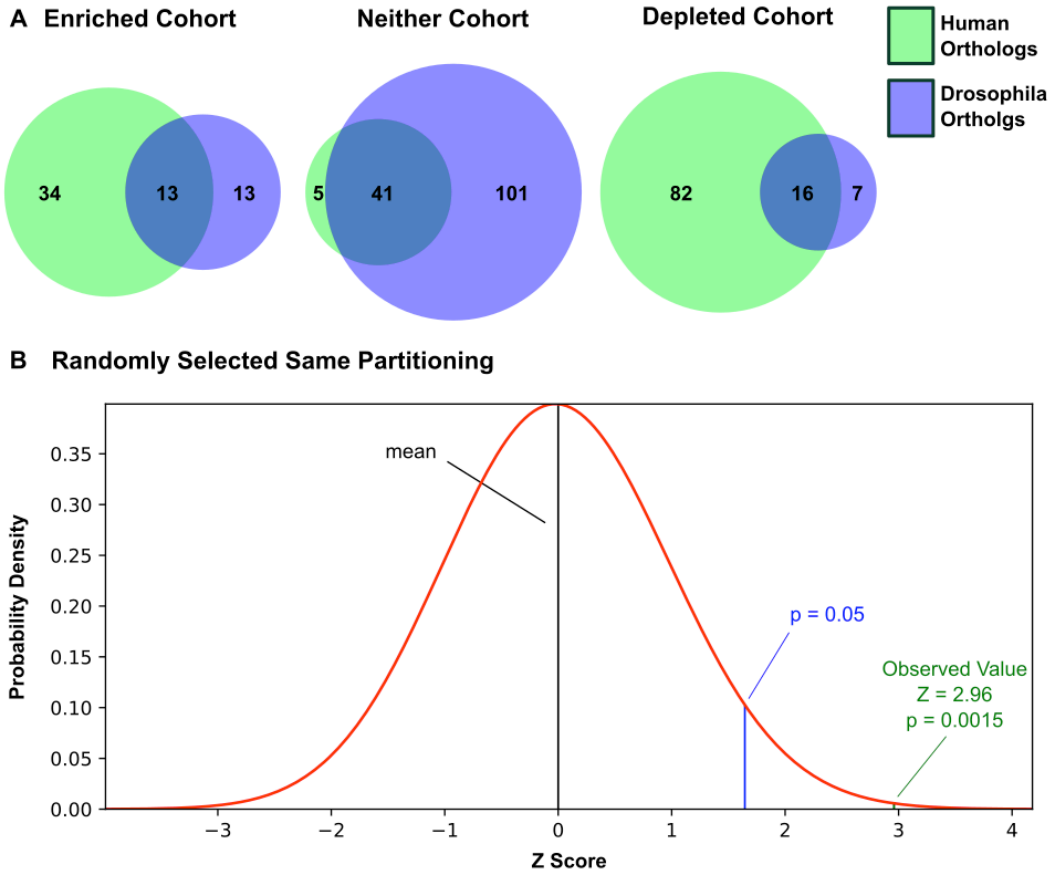


Figure 3.5: Selective partitioning of orthologous human and *Drosophila* mRNA in DDX4N. A) Venn diagrams displaying how orthologous mRNA transcripts from each partitioning cohort for each data set overlapped. B) Normal distribution plot for the % of orthologs with the same partitioning behaviour generated via random sampling of our data set. The mean of the normal distribution, the Z score where the p-value was equal to 0.05, and the Z score for the observed value from our data set, are displayed.

3.2.2 EWSR1C

Initial characterisation of EWSR1C and design of the mRNA partitioning experiment

In order to investigate how mRNA partitioning might be affected by the solvent properties of condensates, we utilised an orthogonally charged model to DDX4N, EWSR1C. EWSR1C is positively charged, but like DDX4N, is also intrinsically disordered, and lacking in an RBD, making it an ideal model to examine. Unlike DDX4N, the phase diagram of EWSR1C was previously uncharacterised and a protocol for its purification hadn't been established. EWSR1C purification proved challenging as it was prone to aggregation during the concentration stages, and as a result, on multiple occasions we lost a significant portion of our yield. Secondly, with no prior knowledge of the phase diagram, we concentrated the protein to be in the same regime as DDX4N, at 225 μM. A gel for the final purified protein product is shown in the Appendix (Figure A.5). However, we subsequently determined that this was far higher than was necessary for inducing phase separation and as a result we were unable to perform as many experiments as we had hoped. This was compounded by the yield of EWSR1C being approximately 6-fold lower than that for DDX4N. Unfortunately owing to Covid-19 we were unable to further improve the purification procedure and were restricted to the yield of two combined purifications with which to perform the partitioning experiment and a limited number of characterisation experiments. Initial experimentation revealed the phase diagram to be significantly different to that of DDX4N; reducing the salt concentration from 300 mM to 100 mM

wasn't sufficient to induce phase separation. However, we demonstrated that addition of cytoplasmic mRNA was sufficient, likely as a result of complex coacervation (Figure 3.6A). Therefore, for the RNA partitioning experiment, we used 10 μg cytoplasmic RNA, to replicate the DDX4N experiment.

As with its purification, the EWSR1C RNA partitioning experiment also proved difficult. Firstly, leaving the dual-phase solution to equilibrate for the same period of time as for the DDX4N experiment, lead to the formation of insoluble aggregates. Reduction of the waiting time, following addition of the cytoplasmic RNA, mitigated this issue and allowed us to observe, extract and resuspend the dense phase prior to phenol / chloroform purification. Secondly, the mRNA library preparation for the dilute phase produced abnormal looking libraries with very short cDNA lengths. Figure 3.6B displays the output from the Bioanalyser, which is used to characterise a cDNA library based on its length distribution. A typical library has a broad length distribution from between 150 to 500 bp with an average of around 250 bp, as shown by the example dense phase library. In contrast, a typical dilute phase library had an average length of around 100 bp with a very narrow distribution. This is indicative of issues with either the input RNA for the library preparation, or issues with the library preparation itself. Bioanalyser profiles like the one shown for a typical dilute phase library, are usually produced by self-annealing primers during the PCR stage of library preparation, owing to low concentrations or poor quality of the input mRNA. We performed the partitioning experiment followed by library preparation multiple times, but in each instance the result for the cDNA library was the same, indicating that it was most likely an issue with the input RNA. Readings obtained from

the Nanodrop indicated that on average 1.7 μg of RNA was in the dilute phase, whilst an average of 2.1 μg was in the dense phase. This left 6.2 μg of the 10 μg input RNA unaccounted for a significantly higher amount than observed for DDX4N. Although RNA was present in the dilute phase, the Nanodrop doesn't distinguish between mRNA and other RNAs such as ribosomal RNA. This leaves the possibility that the dilute phase contained no mRNA and that it was entirely sequestered in EWSR1C condensates. However, we cannot be certain from our results and cannot rule out other causes such as experimental error.

We sequenced the dilute phase libraries, but analysis of the nucleotide composition of the reads confirmed that they were caused by self-annealing of the oligoDT primer, as they were shown to be almost exclusively either adenine or thymine stretches (Figure 3.6C). As a result, we were unable to use the dilute phase BAM files as the reference for downstream analysis. Previously mRNA extracted from SW480 cells had been sequenced by Jonathan Neve (Department of Biochemistry, Oxford University). We therefore, decided to perform the DESeq2 analysis using the SW480 BAM files as the reference for EWSR1C, in lieu of dilute phase BAM files, as they would be reflective of the RNA we had used for the input. We tested the validity of this approach first, by performing this analysis with DDX4N and FUS-GFP to see if the results yielded were comparable to those obtained from analysis with the dilute phase BAM files. We found that the trends we observed were very similar in both cases (Figure A.6A). We next performed Pearson correlation analysis on the enrichment of genes from both methods. This revealed that whilst there was a positive correlation in both cases (R = 0.62 and

R= 0.58 for DDX4N and FUS-GFP respectively) it wasn't as strong as we might have anticipated (Figure A.6B). Thus, this analysis method would be suitable for inferring how the physical properties of mRNA governed mRNA selection in EWSR1C and these results will be discussed subsequently.

EWSR1C partitions human mRNA solely based on its GC %

Given the orthogonal nature of EWSR1C to DDX4N, we anticipated that different mRNA selection principles may apply. In line with this, our results showed that there was no significant difference in the length of transcripts in each of the three partitioning cohorts. The mean lengths were 3416.2 b for the enriched cohort, 3309.8 b for the neither cohort and 3381.8 b for the depleted cohort. However, partitioning analysis revealed a clear GC % dependency. mRNA in the enriched cohort was significantly lower in GC (mean of 46.2 %) than the neither cohort (mean of 52.0 %) and the depleted cohort (mean of 57.1 %). The neither cohort is significantly lower in GC than the depleted cohort. Unlike with DDX4N, 2D histograms revealed a single transcript population governed by GC %, independent of length (Figure 3.6C). These results suggest that DDX4N and EWSR1C, two models with unique characteristics, do indeed have different mRNA selection principles. This suggests partitioning of mRNA on the basis of its length and GC % is solvent dependent, which may serve as critical insight into how BMCs are able to maintain their unique compositions *in vivo*.

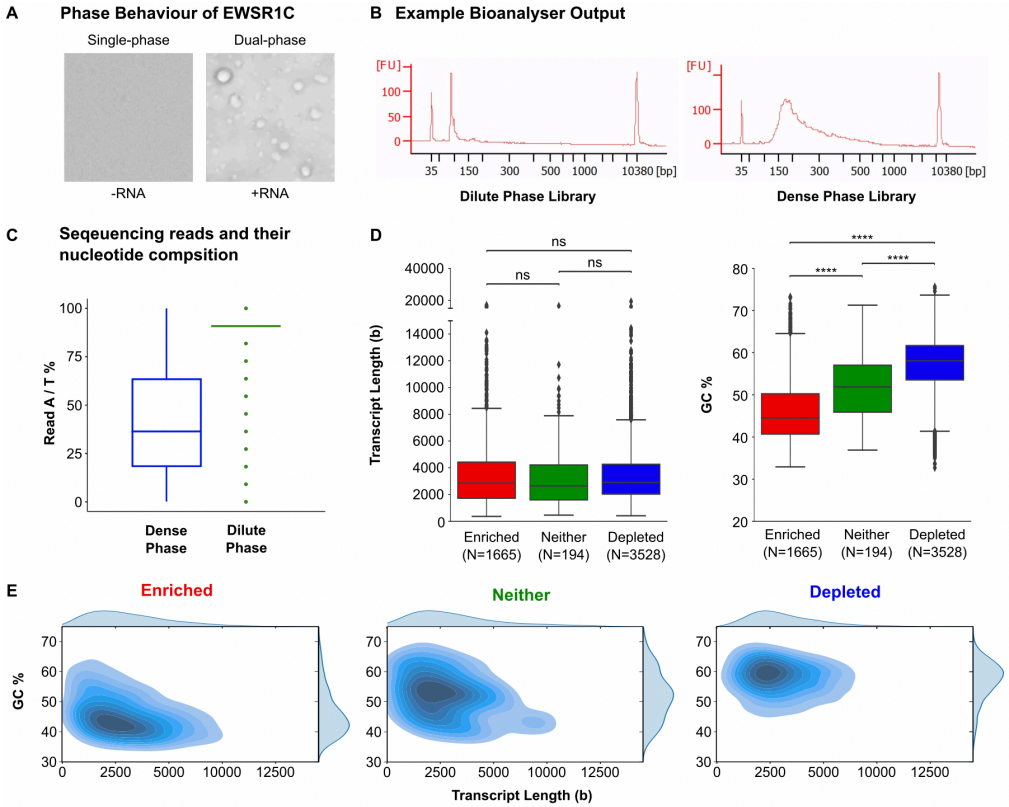


Figure 3.6: Selective mRNA partitioning in EWSR1C. A) Light microscopy images displaying EWSR1C in a single-phase solution without addition of cytoplasmic RNA and EWSR1C in a dual-phase solution when cytoplasmic RNA was added. B) Example output from bioanalyser library quality control showing a typical EWSR1C dilute phase library vs a typical EWSR1C dense phase library. The x axis displays the length in base pairs of the cDNA and the y axis is a proxy for the frequency that a fragment of a particular length is present C) Boxplot displaying the nucleotide composition of sequencing reads found in each phase. A % was obtained from plus strand reads and a T % was obtained from minus strand reads. D) Boxplots depicting transcript length and GC % for each of the three mRNA partitioning cohorts for EWSR1C. E) 2D histograms displaying GC % vs transcript length for each of the three mRNA partitioning cohorts. Significance values displayed above D were acquired using a T-test; ns = p > 0.05, * = p <= 0.05, ** = p <= 0.01, *** = p <= 0.001, **** = p <= 0.0001.

Further characterisation of EWSR1C

Since this was the first study to use EWSR1C, very little was previously known about its phase separating behaviour. We therefore performed a microscopy assay in order to further interrogate this, utilising EWSR1C at varying concentrations, in combination with either cytoplasmic RNA, DDX4N, or both. We found that over time, EWSR1C and RNA condensates were prone to aggregation, and also observed that EWSR1C phase separates in the presence of DDX4N, likely forming multiphase condensates. Our observation that EWSR1C forms aggregates lead us to speculate whether these were amyloids. FUS is known to form amyloids and this has been linked to neurodegenerative diseases, such as ALS (A. Patel et al. 2015). Therefore, we speculated that EWSR1C might do the same. To investigate this, we performed a ThT assay, but our results demonstrated that EWSR1C likely doesn't form amyloids in the presence of RNA. The results for these experiments are discussed in section A.0.7.

3.2.3 FUS-GFP

Design of FUS-GFP mRNA partitioning experiment

FUS-GFP was the final condensate we examined, and it provided another unique dimension to investigate. It was the only model we utilised that contained a recognised RBD, and we were interested to see how this might influence mRNA partitioning. FUS-GFP was obtained as a gift, and as a consequence, we were able to perform the mRNA partitioning experiment, but no further characterisation experiments. However, numerous studies have

investigated the behaviour of FUS-GFP in the presence of total RNA, and we used this information to design an RNA partitioning protocol. The aliquots we received were concentrated at 160 μM, in 500 mM KCl, and the protocol from the Hyman group (Max Planck Institute of Molecular Cell Biology and Genetics) demonstrated that phase separation could be induced by reducing the salt concentration to 125 mM, in the absence of RNA. In addition, 50 ng / μL of total RNA has been shown to stabilise FUS-GFP condensates (Figure 3.7A) (Maharana et al. 2018). Combining these observations, we deduced that addition of 10 μg of RNA to the FUS-GFP aliquots would yield the most stable condensates, at a final concentration of FUS-GFP of 32 μM. Therefore, we were able to utilise the same mass of input RNA as we had used in our previous partitioning experiments. The dilute and dense phases were extracted and sequenced as per sections 3.1.1 and 3.1.2.

FUS-GFP partitions human mRNA based on its GC and GQFR-frequency %

FUS-GFP was the only model protein we examined with a recognised RBD. However, our results showed remarkable similarities to the selection principles in EWSR1C. We found that FUS-GFP selection was also independent of the length of the mRNA transcript and instead solely dependent on its GC % (Figure 3.7B). Similarly, we observed a single population in the 2D histogram, reflective of a selection dependency on GC % alone (Figure 3.7C). FUS contains an RGG domain that interacts with G-quadruplexes and we hypothesised that this may influence mRNA selection (Imperatore et al. 2020). Counterintuitively, we found that within the enriched cohort, only 38.3 % of

transcripts contained at least 1 GQFR, whilst in the neither and depleted cohort, we saw that 47.2 % and 51.5 % did respectively (Figure 3.7D). Again, we observed that the physcial properties of mRNA defined it selection, as we had seen for the other two model condensates. Interestingly, like EWSR1C, FUS-GFP appears to select for mRNA independent of length, and primarily its GC %, suggesting that the solvent properties of the condensates formed by both models, may be highly similar. This is intuitive, since both are positively charged and from the same family of proteins (FET family), with high sequence homology.

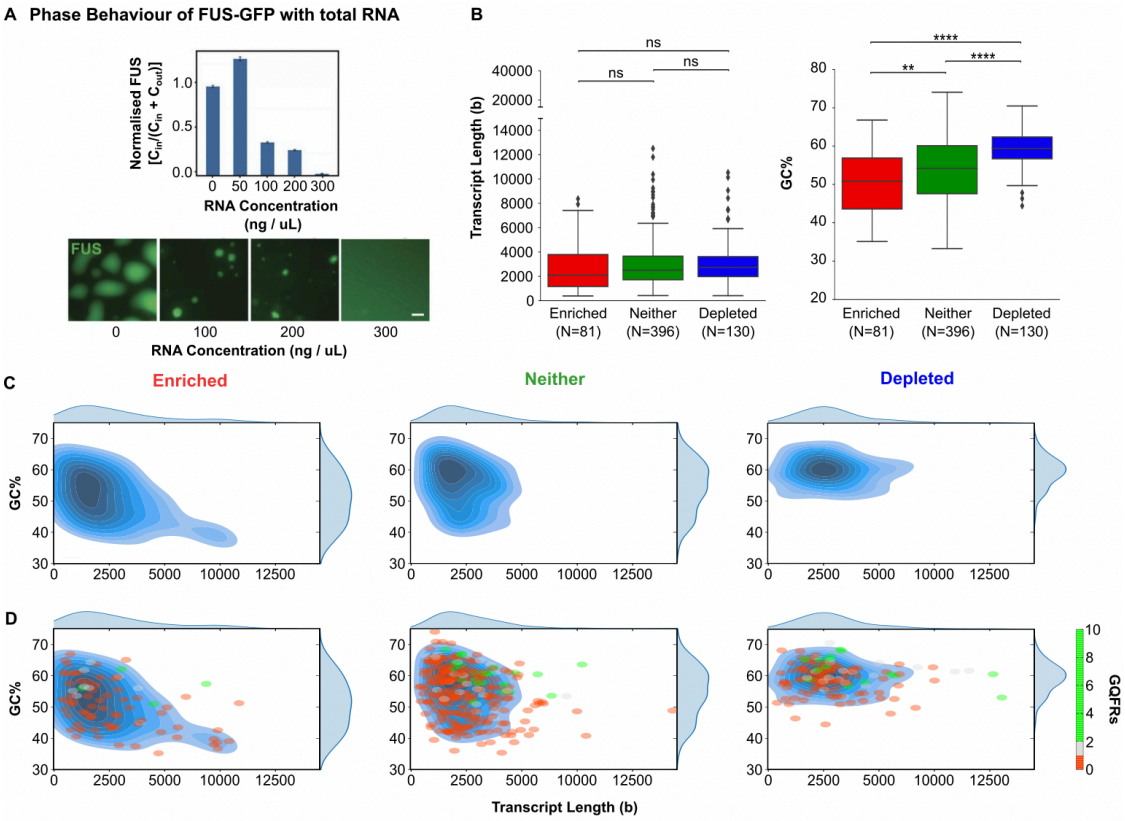

A Phase Behaviour of FUS-GFP with total RNA
Normalised FUS [C_in/(C_in + C_out)]
RNA Concentration (ng / uL)
FUS
B
Transcript Length (b)
GC%
Enriched (N=81)
Neither (N=396)
Depleted (N=130)
C
Enriched
Neither
Depleted
D
GQFRs
Transcript Length (b)
E Enriched Motif Analysis
Streme Motif
Top Tomtom RNA Binding Proteins
1) SNRPA
2) MSI1
3) MSI1
1) KHDRBS1
2) A1CF
3) KHDRBS3
1) RBM46
2) TARDBP
3) RBM5

Figure 3.7 *(previous page)*: Selective partitioning of mRNA in FUS-GFP. A) Phase behaviour of FUS-GFP in the presence of total RNA. The bar graph displays how the stability of FUS-GFP condensates varies with the concentration of total RNA. Stability is measured as the ratio of protein found in the dense phase compared to the total protein input (protein found in the dilute phase (C_{out}) + protein found in the dense phase (C_{in})). Microscopy images showing FUS-GFP with varying total RNA concentrations are also displayed. Figures are adapted from (Maharana et al. 2018). B) Boxplots depicting transcript length and GC % for each of the three mRNA partitioning cohorts C) 2D histograms displaying GC % vs transcript length for each of the three mRNA partitioning cohorts. D) 2D histograms from C) displayed with overlayed points corresponding to individual transcripts. These are coloured according to the number of GQFRs they contain; 0 GQFRs = red, 1 GQFR = grey, 2+ GQFRs = green. E) Top three most significant motifs identified in the enriched mRNA cohort by streme. The top three RNA-binding proteins identified Tomtom that bind each motif are shown. Significance values displayed above B were acquired using a T-test; ns = p > 0.05, * = p <= 0.05, ** = p <= 0.01, *** = p <= 0.001, **** = p <= 0.0001.

The RNA-binding domain is likely not a strong determinant in overall mRNA partitioning in FUS-GFP

To investigate the role of the RBD in FUS-GFP, we performed motif enrichment analysis using streme (a software for discovering enriched motifs) on the mRNAs in the enriched cohort, using a control group of the same length, obtained from mRNA in the neither cohort (T. L. Bailey 2021). Using the three most significant motifs from this analysis, we identified motifs associated with RBPs using the Tomtom motif comparison software (Gupta et al. 2007; Ray et al. 2013). However, surprisingly, this didn't reveal any motifs associated with FUS binding (Figure 3.7E). Strikingly, further inspection of the genes within each mRNA cohort revealed that FUS itself was one of the

most significantly enriched genes in the FUS-GFP condensates. It was found to be enriched with a padj of 0.00026 and a $\log_2$ fold change of 2.55, and in addition, it was the only model condensate where FUS was significantly enriched. FUS is known to bind its own mRNA, and this finding may provide evidence of an autoregulatory mechanism through which FUS sequesters its own mRNA into condensates, as a means of regulating its own expression (Y. Zhou et al. 2013). Together our results suggest that the RNA binding domain of FUS-GFP was likely not a strong determinant in influencing mRNA partitioning overall, but that it might preferentially sequester a small subset of mRNAs, including FUS itself.

FUS-GFP enriches for mRNAs with distinct functionalities from DDX4N

We next investigated whether mRNAs with similar functionalities to DDX4N may also be enriched in FUS-GFP. However, GO analysis revealed that there were no similarities, and indeed, where terms were present in both data sets, they were attributed to opposite partitioning cohorts (Figure 3.8) (E. Y. Chen et al. 2013; Kuleshov et al. 2016; Z. Xie et al. 2021). One of the most significant terms for GO biological processes, in the enriched cohort for FUS-GFP, was "cytoplasmic translation", a term which was present in the depleted cohort in DDX4N (Figure 3.4). This further emphasises the notion that the properties of the condensate define which mRNAs they enrich, and again serves as evidence for how condensates may maintain their unique compositions.

Figure 3.8: Gene ontology analysis of mRNA partitioning cohorts in FUS-GFP. The top five results for each gene ontology analysis are displayed along with their respective p-adjusted (padj) values with transcripts grouped by mRNA partitioning cohort. The size of the bar for each GO term is a reflection of the significance of the result, with the most significant result having the largest bar in each group. A) GO analysis of biological processes. B) GO analysis of cellar components.

3.2.4 Summary of the results from the *in vitro* model BMCs

In each model BMC we investigated, we saw clear trends in the physical properties of mRNA that determined how it was partitioned. Importantly however, we found that the rules that defined selection were unique to each. We found transcript length was only a determinant in DDX4N, whilst enrichment in EWSR1C and FUS-GFP was primarily dependent on GC %. In addition, DDX4N enriched for mRNAs with a high frequency of GQFRs, whilst FUS-GFP depleted them. Interestingly, the trends we observed were most alike in the two positively-charged condensates, EWSR1C and FUS-GFP indicating that BMCs enrich for mRNAs on the basis of their physical properties in a manner conferred by the solvent properties of the condensate itself.

3.2.5 Selective partitioning of mRNA in *in vivo* BMCs

Comparative analysis of selective partitioning of mRNA in *in vivo* PBs and SGs

We had established clear trends in the physical properties of mRNA that defined how it partitioned in a panel of model BMCs. Next, we wanted to investigate whether these trends would persist in highly complex *in vivo* systems. Recently, techniques have been developed that have allowed for purification of certain BMCs from cells, and subsequent analysis of their transcriptomes. SGs have been shown to enrich long, poorly-translating mRNAs, whilst PBs enrich for long, AU-rich, poorly-translating mRNAs (S. Jain et

al. 2016; Khong et al. 2017; Hubstenberger et al. 2017; Courel et al. 2019). To directly compare these results with the results from our *in vitro* model systems, we obtained the sequencing data from the Courel *et al.* (PBs), and Khong *et al.* (SGs) studies, and analysed them using the same partitioning pipeline we had developed for our *in vitro* analysis. The data was acquired via Galaxy and aligned to the hg19 genome using HISAT2 (Yun Zhang et al. 2021; Afgan et al. 2018). The sequencing method used to obtain both data sets was full RNA-seq and this required use of a different intersection file for generating the read tables (hg19 GTF file from UCSC). All subsequent downstream analysis was otherwise identical to the method described in section 3.1.2, and we obtained genes in each of the three partitioning cohorts for each *in vivo* BMC.

mRNA length, GC % and GQFR-frequency also define mRNA selection in *in vivo* BMCs

In agreement with the results from Courel *et al.* we found that selective partitioning of mRNA in PBs was dependent on both the length and GC % of transcripts (Courel et al. 2019; Hubstenberger et al. 2017). Long, AU-rich transcripts were significantly enriched, whilst short, GC-rich transcripts were depleted (Figure 3.8A). Prior to this study, the role of GQFRs hadn't been established in PBs. However, we found that the frequency of GQFRs in each of the partitioning cohorts was highly similar, with no clear trend, suggesting that GQFRs likely aren't a discriminator in gene-specific mRNA selection in PBs (Figure A.9).

Our analysis of the SG transcriptome agreed with the results from Khong

et al (Khong et al. 2017). Longer mRNAs were significantly enriched, whilst short mRNAs were significantly depleted (Figure 3.8A). Surprisingly, we found there was also dependency on GC % in SGs. Enriched mRNAs were significantly lower in GC % (mean of 48.5 %) than depleted mRNAs (mean of 51.1 %) (Figure 3.8A). As can be seen from the mean values however, this effect is less pronounced than the trend observed in PBs. Interestingly, our 2D histogram analysis identified two potential transcript populations (Figure 3.8B). GQFR analysis further demonstrated that these two peaks could be defined by their GQFR frequency, as we had observed for DDX4N (Figure 3.9A). We found that enriched mRNAs were also enriched in GQFRs with 54.4 % containing at least one, compared to only 26.6 % in the depleted cohort (Figure A.9). Our findings support a role for GQFRs in selectively targeting mRNAs to SGs which agrees with the observation that GQFR-containing mRNAs can promote robust SG assembly *in vivo* (Fay, P. J. Anderson, and Ivanov 2017).

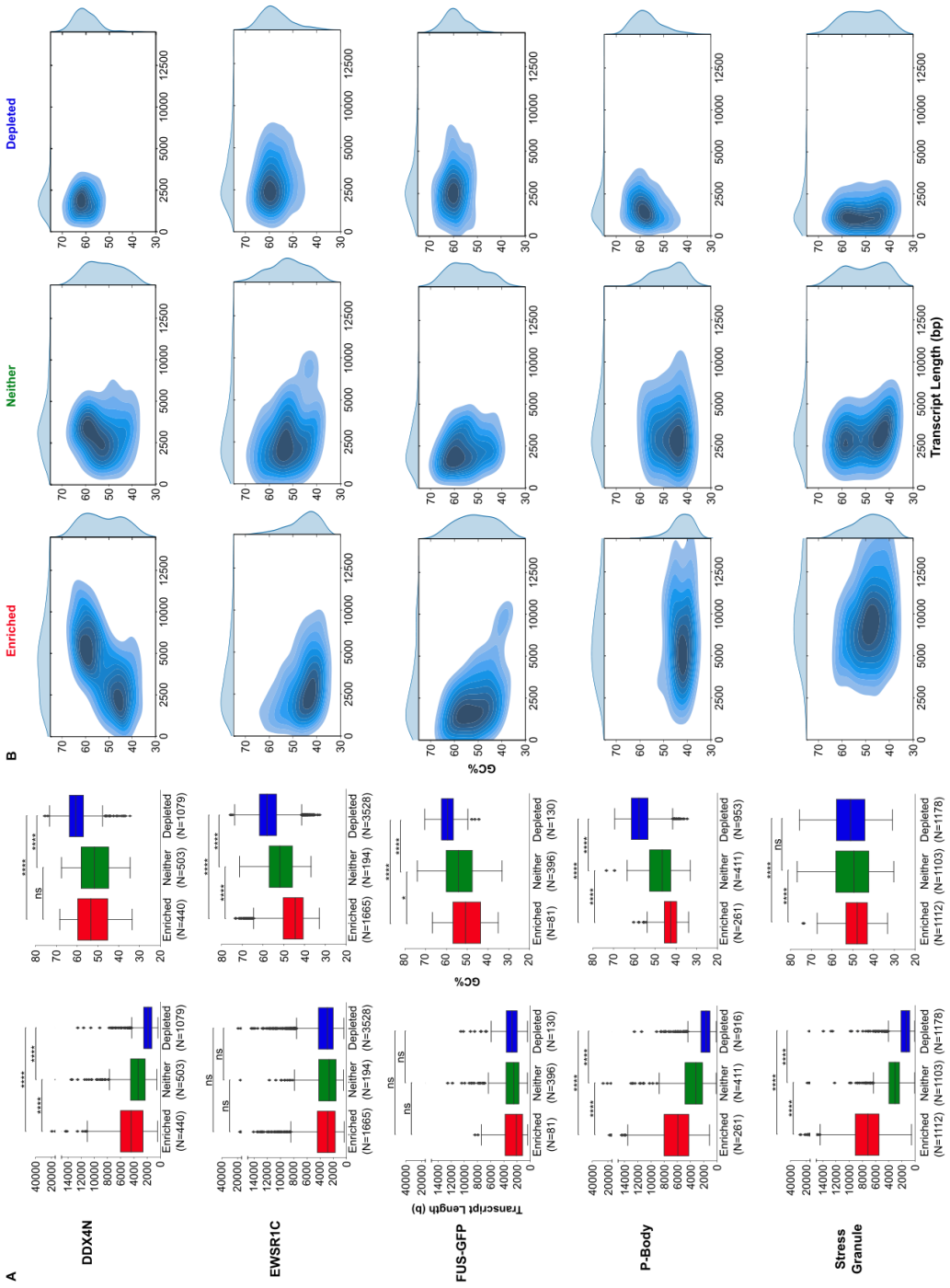

A
B
DDX4N
EWSR1C
FUS-GFP
P-Body
Stress Granule
Enriched
Neither
Depleted
Transcript Length (b)
GC%
Transcript Length (bp)
Enriched (N=440)
Neither (N=503)
Depleted (N=1079)
Enriched (N=1665)
Neither (N=194)
Depleted (N=3528)
Enriched (N=81)
Neither (N=396)
Depleted (N=130)
Enriched (N=261)
Neither (N=411)
Depleted (N=953)
Enriched (N=1112)
Neither (N=1103)
Depleted (N=1178)
ns

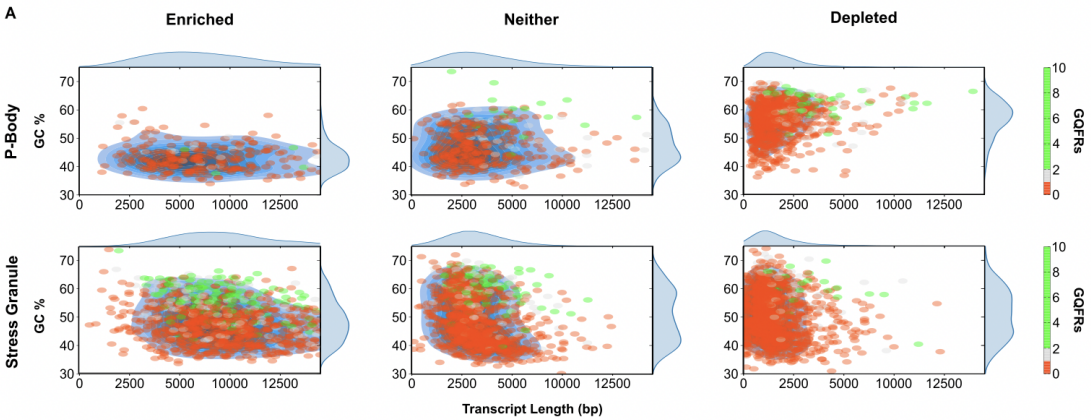
Figure 3.8 *(previous page)*: Selective partitioning of mRNA in BMCs. A) Boxplots depicting transcript length and GC % for each of the three mRNA partitioning cohorts for each BMC. B) 2D histograms displaying GC % vs transcript length for each of the three mRNA partitioning cohorts in each BMC.

Our findings demonstrate that mRNA selection in simple *in vitro* systems, and highly complex *in vivo* systems, can be well characterised on the basis of a transcripts physical properties alone (Figure 3.8A and B). This strongly suggests that the fundamental determinant in selective partitioning of mRNA in BMCs is dictated by the physical properties of the mRNA, rather than through specific interactions with RBPs or ribosomes. In addition, the trends we observed followed the same direction for all condensates; where length was a determinant, longer mRNAs were enriched, where GC % was a determinant, AU-rich mRNAs were enriched. It is difficult to imagine how such trends would persist unless the same driving forces determine mRNA selection in both types of system.

The *in vitro* DDX4N transcriptome shares remarkable similarities with the *in vivo* PB transcriptome

We had determined that similar trends in mRNA physical properties determined its selection in both *in vitro* and *in vivo* systems. We next addressed whether the function of mRNAs partitioned by PBs and SGs was similar to those in our *in vitro* models. To investigate, we performed GO analysis examining the mRNA in the enriched and depleted cohorts of PBs and SGs (E. Y. Chen et al. 2013; Kuleshov et al. 2016; Z. Xie et al. 2021). Interest-

ingly, we found that the cohorts of mRNA enriched and depleted in SGs and PBs were almost identical in function (Figure 3.9B). Comparison with the terms for DDX4N revealed some surprising similarities, particularly in the function of depleted mRNAs. Terms for GO biological processes "cytoplasmic translation" and "translation" were present in all three, whilst the same was true for GO cellular component terms "large ribosomal subunit" and "mitochondrial membrane" (Figure 3.4). However, the functions of genes in the FUS-GFP cohorts were very different. We found that where the same terms appeared, they were in the opposite partitioning cohort; FUS-GFP enriched for transcripts with functions that were depleted for the other BMCs (Figure 3.8).



B GO Biological Processes

	Enriched mRNA cohort	Depleted mRNA cohort
P-Bodies	regulation of transcription by RNA polymerase II (GO:0006357) - padj 1.631e-30	SRP-dependent cotranslational protein targeting to membrane (GO:0006614)) - padj 5.701e-66
	regulation of transcription, DNA-templated (GO:0006355) - padj 1.941e-29	cotranslational protein targeting to membrane (GO:0006613)) - padj 4.142e-64
	positive regulation of transcription, DNA-templated (GO:0045893) - padj 9.896e-13	cytoplasmic translation (GO:0002181)) - padj 4.415e-63
	positive regulation of transcription by RNA polymerase II (GO:0045944) - padj 9.896e-13	translation (GO:0006412)) - padj 1.468e-62
	negative regulation of transcription, DNA-templated (GO:0045892) - padj 6.931e-11	protein targeting to ER (GO:0045047)) - padj 7.282e-62
Stresss Granules	regulation of transcription by RNA polymerase II (GO:0006357)) - padj 3.061e-33	SRP-dependent cotranslational protein targeting to membrane (GO:0006614)) - padj 2.375e-72
	regulation of transcription, DNA-templated (GO:0006355)) - padj 2.941e-29	cytoplasmic translation (GO:0002181)) - padj 2.375e-72
	positive regulation of transcription, DNA-templated (GO:0045893)) - padj 1.230e-21	cotranslational protein targeting to membrane (GO:0006613)) - padj 2.840e-68
	negative regulation of transcription, DNA-templated (GO:0045892)) - padj 1.510e-17	protein targeting to ER (GO:0045047)) - padj 1.218e-66
	positive regulation of transcription by RNA polymerase II (GO:0045944)) - padj 3.944e-13	peptide biosynthetic process (GO:0043043)) - padj 1.708e-60

GO Cellular Components

P-Bodies	nucleus (GO:0005634)) - padj 7.275e-28	mitochondrial inner membrane (GO:0005743)) - padj 2.187e-38
	intracellular membrane-bounded organelle (GO:0043231)) - padj 5.507e-23	organelle inner membrane (GO:0019866)) - padj 1.426e-37
	cytoskeleton (GO:0005856)) - padj 0.0006041	mitochondrial membrane (GO:0031966)) - padj 4.127e-35
	nucleolus (GO:0005730)) - padj 0.001160	cytosolic large ribosomal subunit (GO:0022625)) - padj 3.356e-34
	nuclear lumen (GO:0031981)) - padj 0.001199	large ribosomal subunit (GO:0015934)) - padj 3.874e-34
Stresss Granules	intracellular membrane-bounded organelle (GO:0043231)) - padj 1.454e-28	large ribosomal subunit (GO:0015934)) - padj 1.879e-46
	nucleus (GO:0005634) - padj 6.883e-27	cytosolic large ribosomal subunit (GO:0022625)) - padj 5.678e-44
	cytoskeleton (GO:0005856)) - padj 6.123e-8	mitochondrial inner membrane (GO:0005743)) - padj 7.056e-41
	intracellular non-membrane-bounded organelle (GO:0043232)) - padj 0.00001020	mitochondrial membrane (GO:0031966)) - padj 1.103e-39
	microtubule (GO:0005874)) - padj 0.00005969	organelle inner membrane (GO:0019866)) - padj 7.458e-39

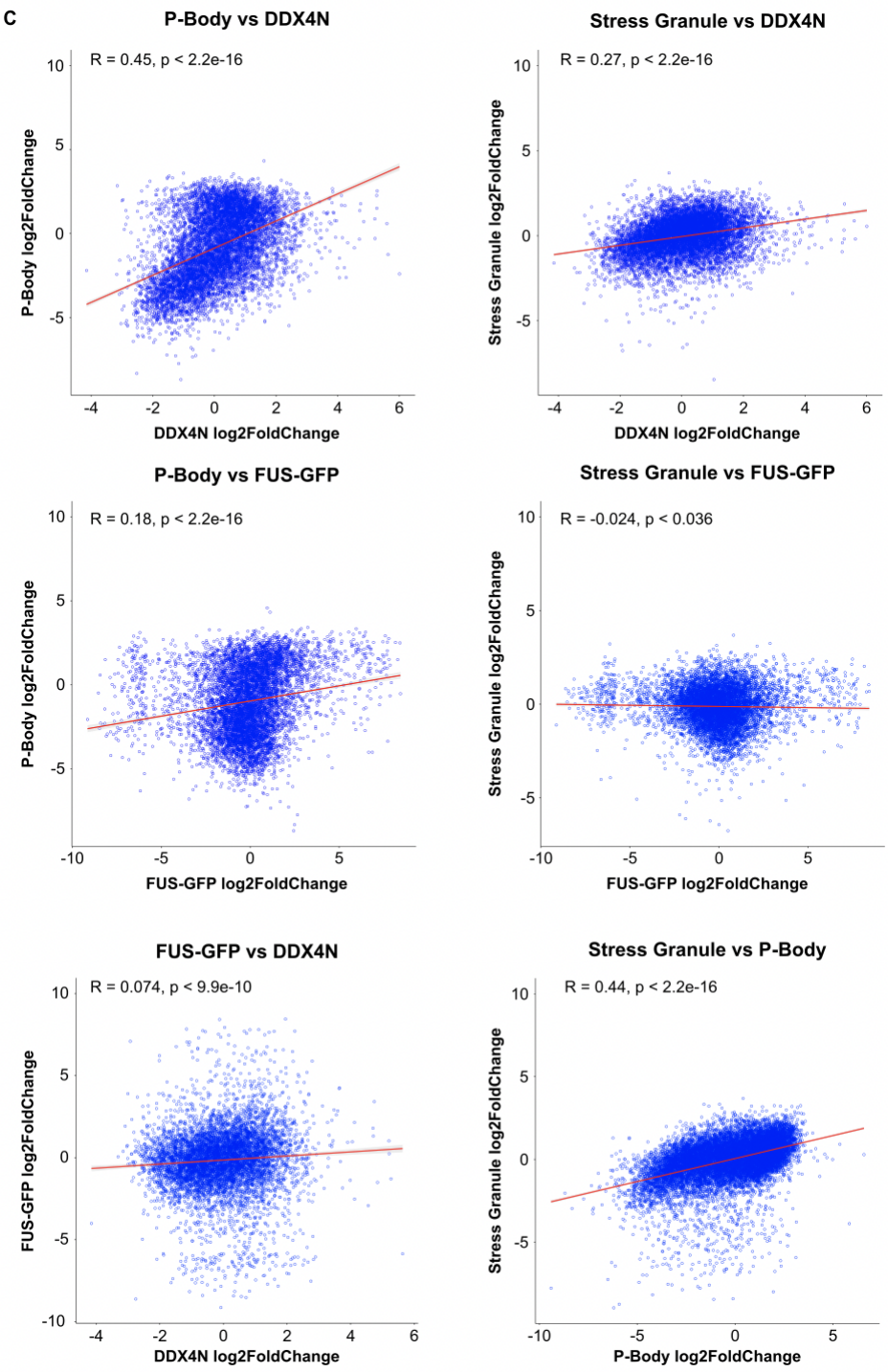

C
P-Body vs DDX4N
R = 0.45, p < 2.2e-16
P-Body log2FoldChange
DDX4N log2FoldChange
Stress Granule vs DDX4N
R = 0.27, p < 2.2e-16
Stress Granule log2FoldChange
DDX4N log2FoldChange
P-Body vs FUS-GFP
R = 0.18, p < 2.2e-16
P-Body log2FoldChange
FUS-GFP log2FoldChange
Stress Granule vs FUS-GFP
R = -0.024, p < 0.036
Stress Granule log2FoldChange
FUS-GFP log2FoldChange
FUS-GFP vs DDX4N
R = 0.074, p < 9.9e-10
FUS-GFP log2FoldChange
DDX4N log2FoldChange
Stress Granule vs P-Body
R = 0.44, p < 2.2e-16
Stress Granule log2FoldChange
P-Body log2FoldChange

Figure 3.9 *(previous page)*: GQFR, and GO analysis of mRNA partitioning cohorts in PBs and SGs, and Pearson correlation analysis of selective mRNA partitioning in all BMCs. A) 2D histograms from Figure 3.8B displayed with overlayed points corresponding to individual transcripts. These are coloured according to the number of GQFRs they contain; 0 GQFRs = red, 1 GQFR = grey, 2+ GQFRs = green. B) Gene ontology analysis of mRNA partitioning cohorts in *in vivo* BMCs. The top five results for each gene ontology analysis are displayed along with their respective p-adjusted (padj) values with transcripts grouped by mRNA partitioning cohort. The size of the bar for each GO term is a reflection of the significance of the result, with the most significant result having the largest bar in each group. GO analysis of biological processes and GO analysis of cellar components are displayed. C) Correlation analysis comparing the $\log_2$ fold change of genes between different BMC data sets. The Pearson correlation coefficients are shown with each plot.

To further examine these relationships, we performed Pearson correlation analysis exploring how the same mRNAs were enriched in different pairs of condensates. We utilised the $\log_2$ fold change values that we had obtained from DESeq2 for the genes in the entire data sets (i.e. not the refined cohorts). Astonishingly, we found that the strongest correlation was between DDX4N and PBs with an R value of 0.45. This is surprisingly high for a number of reasons (Figure 3.9C). Most importantly, the DDX4N partitioning experiment was performed *in vitro* whilst the PBs were extracted from *in vivo*. This is further evidence that the same rules must dictate mRNA selection in both types of system. Additionally, since the only similarity between the two types of system is the complexity of the pool of available mRNA, this strongly suggests that the mRNA itself defines how it partitions. This is further evidenced by both condensates exhibiting very similar selection principles based on transcript length and GC %. Moreover, since

DDX4N lacks in any recognised RBD, this again strongly suggests that the properties of the mRNA are the critical determinant in its selection in BMCs, rather than specific interactions with RBPs. We found that the correlation between DDX4N and SGs was weaker with an R value of 0.27. This is likely a reflection of less comparable selection principles. In addition, FUS-GFP correlated poorly with each of the other BMCs, emphasising our findings from GO analysis.

Individual regions in mRNA are unlikely to independently influence partitioning

Multiple studies have alluded to different regions of an mRNA having different potentials in influencing how it partitions (Khong et al. 2017; Courel et al. 2019). It has been shown that the length and the GC % of the coding sequence (CDS), and the 3' untranslated region (3'UTR), correlate with enrichment in SGs and PBs respectively. However, this correlation was found to be much weaker with the 5' untranslated regions (5'UTRs) of mRNAs. To investigate this, we performed a similar analysis, examining each of these regions individually, assessing how their length and GC % influenced mRNA partitioning in each of our model condensates, as well as repeating the analysis for the *in vivo* BMCs. We found that in DDX4N, similar trends in the lengths were observed with a less pronounced dependency on the length of the 5'UTR (Figure A.10). We also demonstrated that EWSR1C and FUS-GFP had no dependency on the length of either of the three regions, consistent with neither showing a dependency on the length of the full transcript. In addition to this, the trends we observed in the GC % for each condensate,

followed the same pattern, being less prominent in the 5'UTR region (Figure A.11). In our model systems it becomes less apparent how these regions may differentially influence selective partitioning since the mRNA was not associated with ribosomes or RBPs. This prompted us to examine how the length and GC % of the 5'UTR, CDS and 3'UTR correlated with the length and GC % of the full mRNA transcript. Surprisingly, we found that the length and GC % of the of the 5'UTR had significantly weaker correlations with the length and GC % of the full transcript, compared with the CDS and 3'UTR regions (Figure 3.10). This finding, together with the fact that we observe similar trends in our *in vitro* systems, implies that the poorer correlation of the 5'UTR with the enrichment of mRNA may rather be a consequence of its poorer correlation with the physical properties of the full mRNA transcript. This suggests that different regions of the mRNA are likely to have a less influential role in defining mRNA selection than the overall properties of the full mRNA molecule. In addition, this finding further emphasises the role of the physical properties of mRNA in determining partitioning, rather than its association with RBPs.

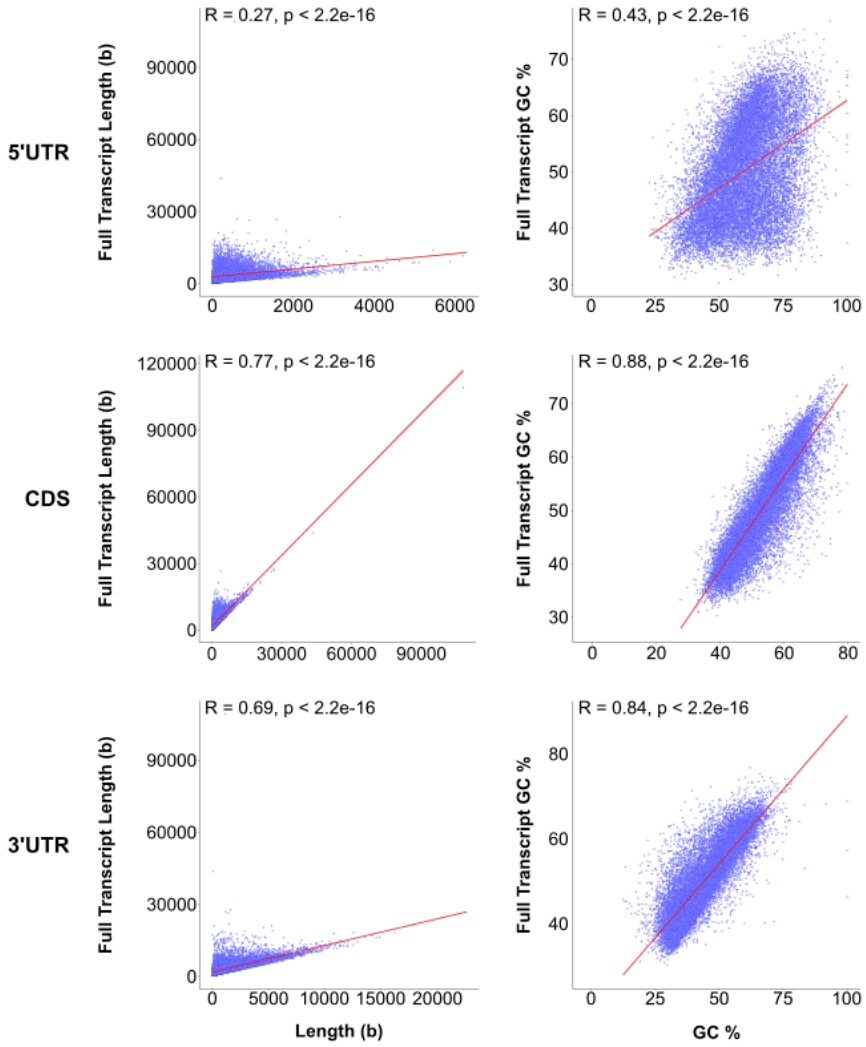


Figure 3.10: Correlation analyses of the lengths and GC % of the 5'UTR, CDS, and 3'UTR with vs the full mRNA transcript in the human transcriptome. The Pearson correlation coefficients are shown with each plot.

3.3 Discussion

The aim of this chapter was to investigate how the physical properties of mRNA influence its partitioning in a variety of BMCs. We developed a novel, simplified, *in vitro* mRNA partitioning experiment coupled with 3'RNA se-

quencing that allowed us to analyse the transcriptomes of a panel of model BMCs, each with unique characteristics. Our data revealed clear trends in the length, GC %, and GQFRs of mRNA, that defined how it was selectively partitioned, in a manner that was specific to the properties of each model. Very similar trends persist in highly complex *in vivo* systems, which strongly suggests that mRNA selection may largely be defined by its physical properties, rather than through its association with RBPs.

We challenged DDX4N with cytoplasmic mRNA extracted from human and *Drosophila* cells and found that in both cases, longer mRNAs were enriched (Figure 3.3). This is consistent with the notion that longer mRNAs are able to engage in more, stabilising, multivalent interactions than their shorter counterparts (Khong et al. 2017; Matheny, Rao, and Parker 2019). GO analysis highlighted remarkable similarities in the function of mRNA found in the enriched and depleted cohorts between the two species indicating that mRNA selection into BMCs may be evolutionarily conserved. Substantiating this claim, we demonstrated that mRNA orthologs were also significantly more likely to have the same partitioning behaviour. We further demonstrated that selection of human mRNA was dependent on its GC % and the number of GQFRs it contained. This wasn't the case for the *Drosophila* data set but this is likely explained by the *Drosophila* transcriptome having a much narrower GC % distribution and being sparse in GQFRs. The influence of RNA GQs on phase-separation is becoming increasingly apparent; poly-guanine nucleic acids assemble into condensates in the absence of protein, owing to the formation of GQ structures and RNAs produced in repeat expansion disorders containing these structures can seed

SG formation *in vivo* (Fay, P. J. Anderson, and Ivanov 2017; Boeynaems et al. 2019). We found that human mRNAs could be split into two distinct partitioning cohorts dependent on whether they contained a GQFR or not. Homologs of DDX4 are known to bind RNA GQs and DDX4N itself also contains four RGG motifs which may promote interaction with GQs (Imperatore et al. 2020; G. Wu et al. 2019; McRae et al. 2017). This suggests that RNA-protein interactions between GQFR-containing transcripts and DDX4N, may influence mRNA selection. In addition, GQ-containing RNA is also known to self-associate and our results may also be demonstrative of GQFR regions engaging in RNA-RNA interactions and shaping condensate composition (Yueying Zhang et al. 2019).

Next, we investigated mRNA partitioning in two orthogonal models to DDX4N, EWSR1C and FUS-GFP. Prior to this study, the phase-separating properties of EWSR1C were uncharacterised, and we demonstrated that EWSR1C condensate formation was dependent on the addition of cytoplasmic RNA. Although there remain some uncertainties surrounding the mRNA content of the dilute phase, we were able to infer selection principles by comparing the condensate transcriptome to that of the input SW480 RNA. mRNA selection in EWSR1C was independent of the length of the mRNA and instead exclusively dependent on it's GC %. FUS-GFP had near identical selection trends, again enriching for AU-rich mRNA independent of the length of the transcript. The intermolecular forces that stabilise BMCs also define the molecules they sequester, and our results concur with this observation; the condensates formed by two positively charged proteins, from the same protein family, shared near identical selection principles that were dif-

ferent to those of the negatively charged DDX4N condensates (Ditlev, Case, and Rosen 2018; Nott, Craggs, and A. J. Baldwin 2016; Banani, H. O. Lee, et al. 2017; Maharana et al. 2018). Common to all models however, was an enrichment of AU-rich mRNAs and a depletion of GC-rich ones. Consistent with this finding, AU-rich mRNAs promote assembly of PB-like granules with cell-free extracts, whilst GC-rich mRNAs don't (Courel et al. 2019). GC-rich mRNAs form more stable secondary structures than AU-rich mRNAs, and multiple studies have demonstrated that this property can be highly influential in condensate formation and RNA selection. Highly structured ribosomal RNA is unable to induce phase separation of PGL-3 unless it has been denatured, whilst NEAT1, which is also highly structured, doesn't induce G3BP1 condensate formation (Faure et al. 2016; Saha et al. 2016; Guillén-Boixet et al. 2020; Chan et al. 2009). Together with our results, these findings imply that inter- or intramolecular base-pairing of mRNA can occlude its bases from engaging in the essential, stabilising, multivalent interactions that are responsible for condensate enrichment.

FUS-GFP provided another dimension to the selection process since it also contained an RBD. However, surprisingly, motif enrichment analysis didn't identify FUS-binding motifs in the enriched mRNA cohort. Fascinatingly, however, we observed that one of the most significantly enriched genes in FUS-GFP was FUS itself. Supporting this observation, FUS has been shown to bind its own mRNA in a conserved region of exon 7. Abrogation of this binding, through mutation, leads to aberrant FUS expression *in vivo* (Y. Zhou et al. 2013). Our result hints at a mechanism for this process, whereby FUS protein expression is autoregulated through sequestration of it's own

mRNA in BMCs. FUS is also known to interact with GQs; GQ-containing mRNA has been shown to seed FUS condensate formation (Imperatore et al. 2020). However, we discerned that mRNA in the enriched cohort of FUS-GFP had a lower frequency of GQFRs than in the depleted. This result may possibly be explained by the inclusion of the GFP-tag in our construct interfering with its GQ-binding capacity. Additionally, however, depletion of GQFR-containing mRNAs may also be a consequence of the solvent properties created by FUS-GFP condensates, rather than by specific protein-RNA interactions.

Several studies have shown that SGs and PBs enrich for long, poorly-translating mRNAs, with PBs also preferentially enriching AU-rich mRNAs (Courel et al. 2019; Khong et al. 2017; S. Jain et al. 2016; Hubstenberger et al. 2017; Matheny, Rao, and Parker 2019). The explanation for these results has often centered around how these mRNA properties are a proxy for low ribosomal occupancy or increased RBP association. However, in this chapter we demonstrate that remarkably similar trends in the physical properties of mRNA govern its selection in highly simplified *in vitro* systems, in two instances with models entirely lacking in RBDs. Van treeck et al. made the critical observation that the transcriptome of condensates formed *in vitro*, solely from yeast RNA in the presence of a crowding agent, largely recapitulated that of *in vivo* SGs. This finding highly implies that the fundamental mechanisms governing mRNA partitioning into BMCs may largely be imparted by RNA-RNA interactions and the physical properties of the mRNA alone. We demonstrate a similarly striking finding, that the *in vitro* DDX4N transcriptome correlates remarkably well with that of *in vivo* PBs

(Figure 3.9B). The only commonality between the two systems is the diversity of the RNA, which again, strongly suggests the physical properties of the mRNA, as well as RNA-RNA interactions, are critical in shaping its selection. Accordingly, very similar partitioning trends in transcript length and GC % were observed for both condensates, which likely manifested in the highly comparable transcriptomes. Additionally, like DDX4N, we demonstrated that mRNA enriched in SGs was shown to be augmented in GQFRs. GQ formation is dynamically regulated *in vivo* and these structures provide an axis for interaction with proteins and other RNA molecules (Varshney et al. 2020). GQFR-containing mRNAs are known to instigate robust assembly of SGs *in vivo*, and in addition, SGs are enriched in many proteins known for GQ-binding, such as the SRSF-family proteins (Markmiller et al. 2018; S. Jain et al. 2016; Fay, P. J. Anderson, and Ivanov 2017). Together, these observations suggest that GQFR structures may serve as a signal for recruitment of RBPs to promote SG assembly.

Multiple studies have alluded that distinct regions of an mRNA transcript may affect its partitioning differentially (Khong et al. 2017; Courel et al. 2019). It has been suggested that the 5'UTR is less influential than both the CDS and 3'UTR in this process, with the emphasis being that these regions associate differentially with ribosomes or RBPs. We demonstrated similar patterns in our *in vitro* model systems where the role of these regions becomes much less apparent. Furthermore, we showed that of the three mRNA regions, the physical properties of the 5'UTR was much more poorly correlated with the physical properties of the full mRNA transcript than either of the CDS or the 3'UTR (Figure 3.10). This implies that the poor correlation

between the 5'UTR and mRNA enrichment isn't due to differential regulation of these regions but simply due to its poor relationship with the overall properties of the full transcript. This seems to be a significant oversight, and lends further emphasis to the idea that the most critical aspect to mRNA selection in BMCs is the transcripts overall physical properties rather than its association with RBPs or ribosomes.

3.4 Summary

In summary, we first demonstrated that the length, GC % and number of GQFRs of mRNA were critical determinants in governing its selection in a panel of *in vitro* model BMCs. Moreover, we showed that the enrichment of transcripts was achieved in a model-specific manner, with the rules determining selection being defined by the characteristics of each. We then demonstrated a remarkable likeness in the trends dictating mRNA selection in two *in vivo* BMCs. Together these findings strongly suggest that partitioning of mRNA in BMCs is fundamentally a consequence of its physical properties rather than a product of its association with RBPs.

Diagrammatic overview of mRNA selection trends in BMCs studied in this Chapter

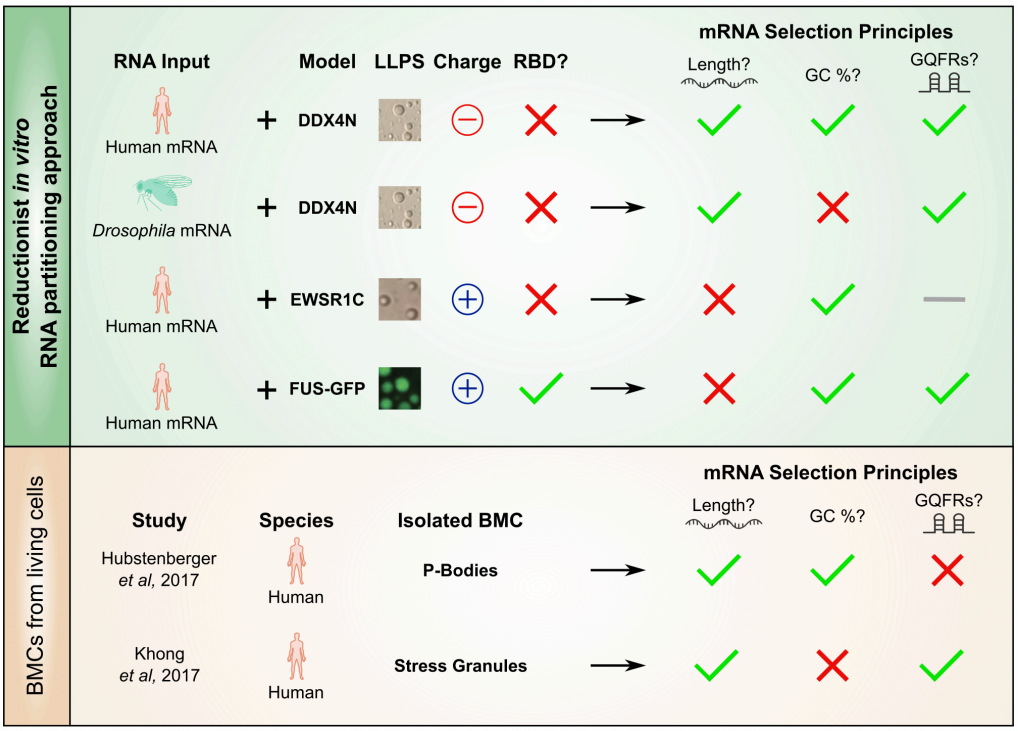


Figure 3.11: Diagrammatic overview of mRNA selection trends in BMCs. The table displays the species of RNA that was partitioned, the BMC that was studied, and the selection trends based on the physical properties of mRNA. LLPS indicates liquid-liquid phase separation and depicts images taken of phase-separated model condensates. RBD stands for RNA-binding domain. Ticks show there is a dependency on that property, crosses show there isn't and the dash indicates this area wasn't investigated.

Chapter 4

Investigating the Role of 3'UTR-APA in Selective Partitioning of mRNA in Biomolecular Condensates

4.1 Introduction

In the previous chapter, we demonstrated that the physical properties of an mRNA shape how it is selectively partitioned in both *in vitro*, and *in vivo* BMCs. There are a number of ways that the cell can actively modify these properties. It can do so post-transcriptionally, via post-transcriptional modifications like m^6A-methylation or alteration of mRNA structural elements by the action of helicases. It can also do so co-transcriptionally, via the processes of alternative splicing, and alternative cleavage and polyadenylation

(APA). Cleavage and polyadenylation (CPA) is a 3'end processing reaction where pre-mRNA is cleaved at a specific site (CPA site), and a stretch of adenosines is added by polyadenylate polymerase (Balbo and Bohm 2007). 70 % of human genes contain multiple CPA sites, and the process by which the cell actively regulates cleavage site selection is known as APA (Neve and Furger 2014; Derti et al. 2012; Hoque, W. Li, and B. Tian 2014). There are two main types of APA; coding region-APA (CR-APA) and 3' untranslated region-APA (3'UTR-APA). In CR-APA, CPA sites are found upstream of the 3'UTR which results in multiple mRNA isoforms being produced that encode multiple protein isoforms. In 3'UTR-APA CPA sites are found within the 3'UTR and mRNA isoforms are produced that encode the same protein, but differ by a stretch of sequence known as the alternative untranslated region (aUTR), which in some cases can be very large (Figure 4.1A) (Hoque, Ji, et al. 2013). Whilst the consequences of CR-APA are apparent; the cell is able produce protein isoforms with different localisation or function; the role of 3'UTR-APA is somewhat less intuitive. 3'UTR-APA profiles are highly dynamic and have been shown to be shaped by cell type and environmental conditions, as well as being linked in a number of cases to disease, including cancer and neurodegenerative diseases such as ALS and Alzheimer's (R. Patel et al. 2019; Fischl et al. 2019; Haibo Zhang, J. Y. Lee, and B. Tian 2005; E. T. Wang et al. 2008; Hollerer et al. 2016; D. Zheng et al. 2018). It has been suggested that 3'UTR-APA can influence mRNA fate in three main ways; firstly by altering the stability of isoforms, secondly by affecting their translation efficiencies and thirdly, by influencing their subcellular localisation (Neve, R. Patel, et al. 2017; Neve and Furger 2014; Mayr 2019;

Mayr and Bartel 2009; Akman, Can, and Erson-Bensan 2012; Andreassi and Riccio 2009; Neve, Burger, et al. 2016). However, in spite of these findings, there is still no clear consensus as to its precise function. In light of this, and given the ability of 3'UTR-APA to manipulate the physical properties of an mRNA, we investigated how this process might influence selective partitioning of mRNA in BMCs.

To date, there has been remarkably little investigation into the role that APA might have in targeting mRNA to BMCs. One study has shown that enrichment of uridine-motifs within aUTRs of mRNA correlated with SG enrichment and TIA1 (a major SG component) binding (D. Zheng et al. 2018). However, as has been commonplace for studies investigating *in vivo* BMCs, the narrative is centred around how the aUTR might promote interactions with RNA-binding proteins and result in selective partitioning, rather than how this could be affected by altering the physical properties of the transcript itself. An example where RNA transcribed from the same gene have different roles in BMC formation is the case of NEAT1 (Yamazaki et al. 2018; Taiana et al. 2020). Here the shorter polyadenylated isoform does not promote paraspeckle formation, whilst the longer, non-polyadenylated isoform is essential to it. This serves as evidence that mRNA isoforms have the potential to exhibit vastly different behaviours with respect to their roles in both the formation of, and the selective partitioning into, BMCs. With the context of the results from the previous chapter, we explore how 3'UTR-APA influenced the partitioning of mRNA, through alteration of their physical properties, in both our *in vitro* model systems and *in vivo* BMCs. In doing so, we sought to establish a novel role for this process with profound con-

sequences for our understanding of mRNA metabolism and gene expression regulation.

4.1.1 Establishing pipelines to analyse 3'UTR-APA isoform-specific partitioning of mRNA in BMCs

In the previous chapter, we were able to utilise the same partitioning pipeline to examine both the *in vitro*, and *in vivo* data sets, since we were combining all the reads that belonged to the same gene, regardless of which isoform they corresponded to. However, owing to the difference in sequencing methods this was not possible for isoform-specific analysis. Figure 4.1B highlights the key differences between the two sequencing methods; reads from 3'RNA-seq are clustered to specific CPA sites and can be assigned to specific isoforms, whilst reads from full RNA-seq are distributed across the whole gene. We therefore, employed two separate 3'UTR-APA partitioning pipelines for analysis of the two types of system.

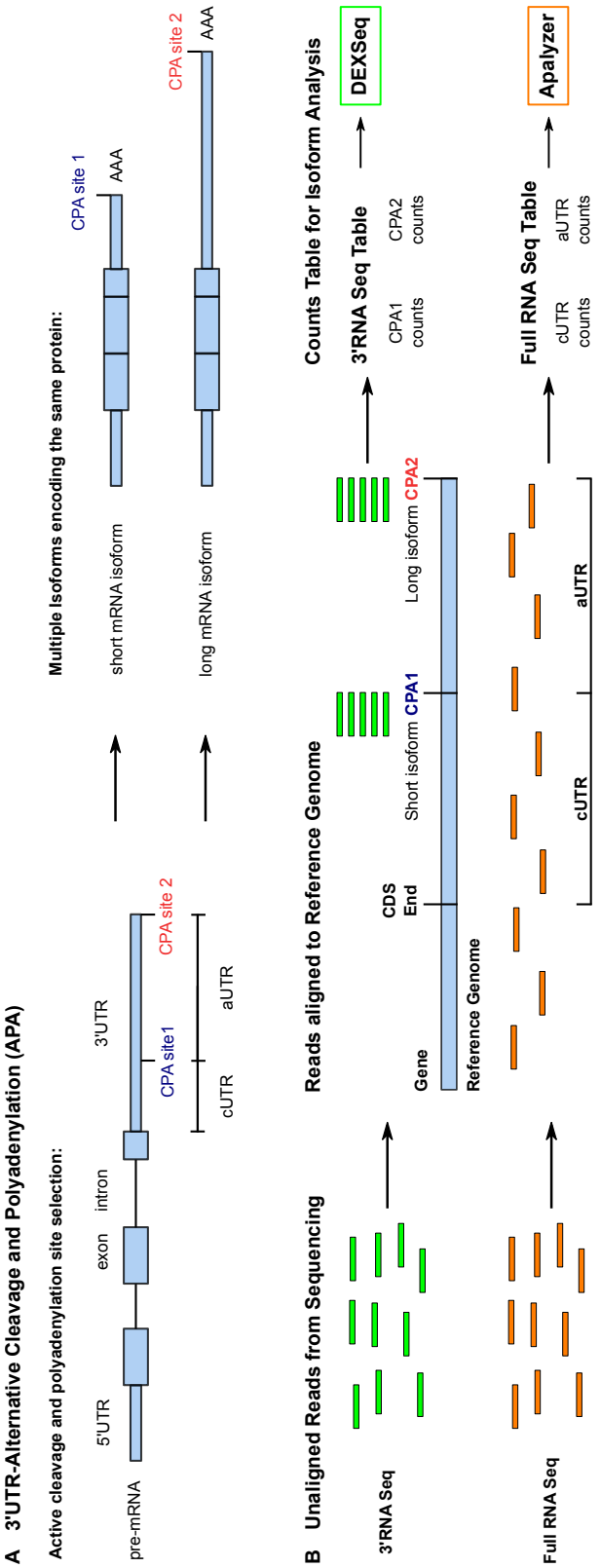

A 3'UTR-Alternative Cleavage and Polyadenylation (APA)
Active cleavage and polyadenylation site selection:
pre-mRNA
5'UTR
exon
intron
3'UTR
CPA site1
CPA site 2
cUTR
aUTR
Multiple Isoforms encoding the same protein:
CPA site 1
AAA
short mRNA isoform
CPA site 2
AAA
long mRNA isoform
B Unaligned Reads from Sequencing
3'RNA Seq
Full RNA Seq
Reads aligned to Reference Genome
Gene
CDS
End
Short isoform CPA1
Long isoform CPA2
Reference Genome
cUTR
aUTR
Counts Table for Isoform Analysis
3'RNA Seq Table
CPA1 counts
CPA2 counts
DEXSeq
Full RNA Seq Table
cUTR counts
aUTR counts
Apalyzer

Figure 4.1 *(previous page)*: 3'UTR-APA, sequencing methods and subsequent 3'UTR-APA analysis approaches. A) pre-mRNA is transcribed from a gene which contains two cleavage and polyadenylation sites (CPA sites) within its 3'untranslated region (3'UTR). The sequence between these two CPA sites is known as the alternative untranslated region. B) The pre-mRNA is processed and cleaved and polyadenylated at either of the CPA sites to produce multiple mRNA isoforms that differ only by the length of their 3'UTRs. Both isoforms encode the same protein. B) Unaligned reads are obtained from 3' RNA sequencing (3' RNA seq) and full RNA sequencing (full RNA seq). Reads are aligned to the reference genome. Reads from 3' RNA seq cluster around cleavage and polyadenylation sites (e.g. polyadenylation site 1 (CPA1)), whilst reads from full RNA seq disperse across the gene. Counts tables are generated from the aligned reads. For 3'RNA seq, the reads can be identified according to the isoform they are from. I.e. reads that are mapped to CPA1 belong to the first isoform of the gene, whilst reads mapped to CPA2 belong to the second isoform of the gene. The counts table from 3' RNA seq can then be used to perform DEXSeq. For full RNA seq, two regions are defined; the cUTR region, which spans from the end of the coding sequences (CDS end) to CPA1 and the aUTR region which spans from CPA1 to the final polyadenylation site in the gene (in this example, CPA2). The counts for each region are then used for analysis with the APAlyzer.

For the *in vitro* data sets, we designed a pipeline to make use of the resolution provided by 3'RNA-sequencing. Briefly, we utilised the same counts table we obtained as discussed in section 3.1.2, however, instead of combining the reads to each gene, we performed our analysis using the reads for each CPA site individually. We incorporated DEXSeq (a pipeline originally designed for differential exon usage analysis) to identify how isoforms were differentially distributed between the two phases (Reyes et al. 2013). We then filtered the results to investigate the two isoforms with the most significant changes in each gene, and dubbed these the "short" and "long" isoforms, defined by their proximity to the end of the CDS (Figure 4.2A). Genes with

less than 1 raw read for each isoform in each sample were removed. To assess the change in isoform distribution we used the following equation (Equation 4.1):

$$\Delta\text{RA} = \frac{\text{Short(dense)}}{\text{Short(dense)} + \text{Long(dense)}} - \frac{\text{Short(dilute)}}{\text{Short(dilute)} + \text{Long(dilute)}} \quad (4.1)$$

Here, "RA" stands for relative abundance, "short" for short isoform reads, "long" for long isoform reads, and "dense"/"dilute" represents which phase the reads were from. Where ΔRA is greater than 0, abundance of the short isoform relative to the long isoform is greater in the dense phase, and vice versa. We established two isoform partitioning cohorts, "short isoform enriched" (SIE) and "long isoform enriched" (LIE), with which to classify genes. SIE genes were defined as having a padj $<= 0.1$ and $\Delta\text{RA} > 0$, whilst LIE genes were defined as having a padj $<= 0.1$ and $\Delta\text{RA} < 0$. The sequences for each the short and long isoforms were obtained by combining the sequences for longest 5'UTR and CDS (UCSC) with the sequence from the end of the CDS to either of the two respective CPA sites. The aUTR sequence is defined as the sequence between the two CPA sites (Figure 4.1).

For analysis of the *in vivo* data sets, we utilised a program called the Apalyzer, specifically designed for investigation of APA in data sets obtained from full RNA sequencing (R. Wang and B. Tian 2020). Here two regions are defined; the "cUTR" which spans the CDS end to the first CPA site and the "aUTR" which spans the first CPA site to the last (Figure 4.1B). The read density is calculated (normalised reads per kilobase) for each region and

the change in relative expression of the two regions between the two phases is calculated with the following equation (Equation 4.2):

$$\text{RED} = log_2 \left[\frac{\text{aUTR(dense)}}{\text{aUTR(dense)} + \text{cUTR(dense)}} \right] - log_2 \left[\frac{\text{aUTR(dilute)}}{\text{aUTR(dilute)} + \text{cUTR(dilute)}} \right] \quad (4.2)$$

Here, "RED" stands for relative expression difference, "aUTR(dilute/dense)" stands for the read density in the aUTR region and "cUTR(dilute/dense)" stands for the read density in the cUTR region. When the RED value is greater than 0, the read density of the aUTR region relative to the cUTR region is greater in the dense phase than in the dilute phase, and vice versa. As with the previous method, we then defined two cohorts of genes, using the same terms. The SIE cohort was defined as having a padj <= 0.1 and a RED value of < 0, while the LIE cohort was defined as having a padj <= 0.1 and a RED value of > 0.

4.2 Results

4.2.1 The role of 3'UTR-APA in selective isoform partitioning in *in vitro* model systems

3'UTR-APA defines isoform-specific selection in DDX4N in a predictable manner

We had previously established that mRNA selection in DDX4N is dependent on the length of the mRNA and we therefore, anticipated 3'UTR-APA, a process that specifically alters this property, may play a crucial role in this process. Confirming this, our analysis revealed 178 (85.6 %) LIE changes and only 30 (14.4 %) SIE changes in the human data set (Figure 4.2B). The significant bias towards genes with LIE changes corroborates our finding that DDX4N enriches for long mRNAs. Two examples of human LIE genes are depicted in Figure A.13. Surprisingly, for the *Drosophila* data set, we only observed 9 significant 3'UTR-APA changes (4 SIE changes and 5 LIE changes). However, analysis of the length of the aUTRs provided provided an explanation for this result. We found the average length of the aUTRs in the human data set was 1814.0 b whilst the average length for the *Drosophila* transcriptome was only 736.0 b. Closer inspection revealed genes with significant 3'UTR-APA changes in the human data set had significantly longer aUTRs (average of 2465.3 b) than those without (average of 1710.2 b) (Figure 4.2C). Together these findings suggest that in order to induce a significant change in how the isoforms partitioned, the aUTRs needed to be sufficiently long. Since *Drosophila* aUTRs are much shorter, this explains why we observed so

few significant changes. Interestingly, in the human data set, we found that genes with significant 3'UTR-APA changes were also largely distinct from genes with significant gene-specific partitioning changes (Figure 4.2D). This suggests that two different gene cohorts were present; one which is subject to isoform-specific partitioning and the other to gene-specific partitioning.

We next addressed the differences between SIE and LIE genes in the human data set. GQFR-analysis revealed that the aUTRs of LIE genes were twice as likely to contain a GQFR as SIE genes, which corroborates our previous finding that DDX4N enriches for mRNAs containing GQFRs (Figure 4.2E). Next, we examined the properties of the long and short isoforms for both 3'UTR-APA cohorts. We had previously established that a long aUTR was necessary to induce a significant partitioning change, and this was evident from the isoforms in both groups of genes (Figure 4.2F). Inspection of the GC % revealed a key difference between the two groups of genes; SIE genes were AU-rich, whilst LIE genes were GC-rich. These trends appeared to be identical to our findings from the previous chapter, where we showed DDX4N enriched for AU-rich, short and GQFR-poor mRNAs or long, GC-rich, GQFR-rich mRNAs. To confirm this result, we combined the enriched isoform from each group of genes (short isoform from SIE genes and long isoform from LIE genes) to create an enriched isoform cohort. We plotted the GC % vs the length of these isoforms and coloured the points by the numbers of GQFRs they contained, and compared this with the enriched, gene-specific, partitioning cohort from the previous chapter. The two plots were highly comparable, as we again saw a splitting in two peaks, defined by GQFRs (Figure 4.2G). To further investigate why SIE genes were AU-

rich whilst LIE genes were GC-rich, we split our data set from the previous chapter into two groups of above and below 50 % GC. We then investigated the length dependency of these two groups of genes separately (Figure A.12). AU-rich genes showed little dependency on length, and in fact, the enriched cohort were significantly shorter than the neither cohort. Conversely, GC-rich genes were strongly dependent on the length of the transcript, with longer transcripts being significantly more enriched. These results clearly explain the differences between the two 3'UTR-APA cohorts and why we observed AU-rich genes with SIE changes and GC-rich genes with LIE changes. Together these findings substantiate the notion that the physical properties of mRNA is critical in shaping its selection, and demonstrate that 3'UTR-APA can play a key role in manipulating these properties to influence this process. Additionally, this is further emphasised by the fact that two largely independent analyses, evidenced by the lack of overlap between genes with significant 3'UTR-APA changes and genes with significant gene-specific partitioning changes, converged on identical results for the rules that define mRNA selection in DDX4N.

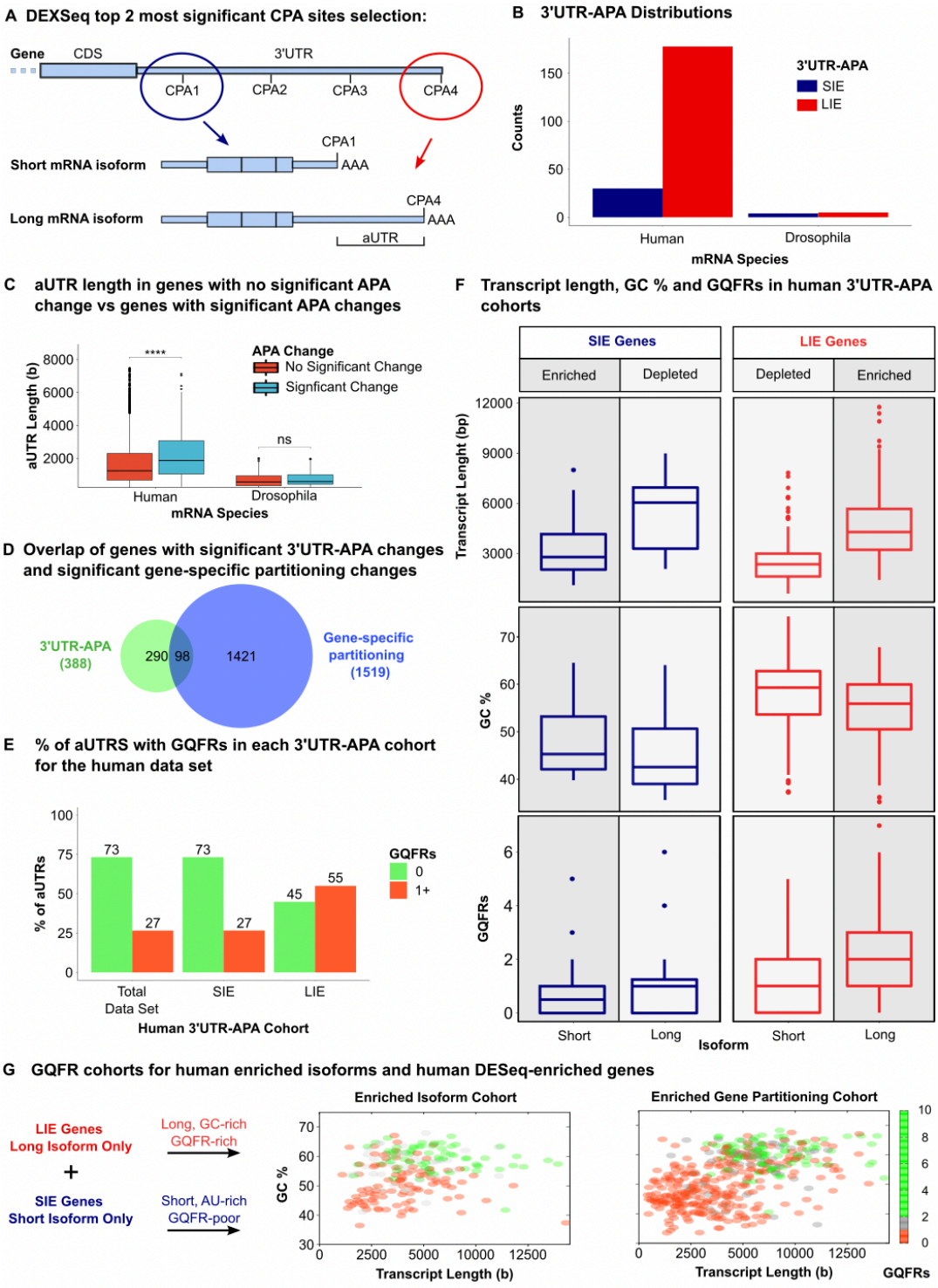

A DEXSeq top 2 most significant CPA sites selection:
Gene
CDS
3'UTR
CPA1
CPA2
CPA3
CPA4
Short mRNA isoform
Long mRNA isoform
AAA
aUTR
B 3'UTR-APA Distributions
Counts
3'UTR-APA
SIE
LIE
Human
Drosophila
mRNA Species
C aUTR length in genes with no significant APA change vs genes with significant APA changes
aUTR Length (b)
APA Change
No Significant Change
Signficant Change
****
ns
D Overlap of genes with significant 3'UTR-APA changes and significant gene-specific partitioning changes
3'UTR-APA (388)
290
98
1421
Gene-specific partitioning (1519)
E % of aUTRS with GQFRs in each 3'UTR-APA cohort for the human data set
% of aUTRs
GQFRs
0
1+
Total Data Set
SIE
LIE
Human 3'UTR-APA Cohort
F Transcript length, GC % and GQFRs in human 3'UTR-APA cohorts
SIE Genes
LIE Genes
Enriched
Depleted
Transcript Lenght (bp)
GC %
GQFRs
Short
Long
Isoform
G GQFR cohorts for human enriched isoforms and human DESeq-enriched genes
LIE Genes
Long Isoform Only
+
SIE Genes
Short Isoform Only
Long, GC-rich
GQFR-rich
Short, AU-rich
GQFR-poor
Enriched Isoform Cohort
Enriched Gene Partitioning Cohort
GC %
Transcript Length (b)
GQFRs

Figure 4.2 *(previous page)*: 3'UTR-APA analysis in DDX4N. A) Diagram depicting selection of the most significant CPA sites according to DEXSeq and their corresponding mRNA isoforms. B) Bar graph depicting the distribution of significant 3'UTR-APA changes in the human and *Drosophila* RNA cohorts. C) Boxplot depicting the lengths of the aUTRs for genes with and without significant 3'UTR-APA changes in the human and *Drosophila* RNA cohorts. D) Overlap between genes with significant 3'UTR-APA changes and significant gene partitioning changes in the human mRNA cohort. E) Bar plots depicting the % of aUTRs which contained 0 or 1+ GQFRs for each 3'UTR-APA cohort for the human RNA cohort. F) Boxplots depicting the length, GC % and number of GQFRs for each isoform for SIE and LIE genes. The boxplots are shaded to depict which isoform was preferentially enriched. G) The short isoforms from the SIE cohort and the long isoforms from the LIE cohort were combined to give an "enriched isoform" cohort for the human data set. Scatter plot for the enriched isoform cohort of GC % vs transcript length was plotted with the points coloured dependent on the number of GQFRs they contain; 0 GQFRs = red, 1 GQFR = grey, 2+ GQFRs = green. The same scatter plot was also shown for the enriched gene partitioning cohort from the previous chapter. Significance values displayed above C) were acquired using a T-test; ns = p > 0.05, * = p <= 0.05, ** = p <= 0.01, *** = p <= 0.001, **** = p <= 0.0001.

3'UTR-APA has no influence on isoform selection in FUS-GFP

We next performed 3'UTR-APA analysis on the FUS-GFP data set, but this revealed no significant results. However, this finding is in agreement with our results from the previous chapter where we demonstrated that FUS-GFP exhibited no selection dependency on the length of the mRNA. Since 3'UTR-APA primarily alters the length, rather than the overall GC % of an mRNA, it's intuitive that this process would play little role in influencing the partitioning of different mRNA isoforms.

4.2.2 Establishing the role of 3'UTR-APA in selective isoform partitioning in *in vivo* BMCs

3'UTR-APA appears highly influential in isoform-specific partitioning in *in vivo* PBs and SGs

Since the length of an mRNA defined its selection in both PBs and SGs, we hypothesised that 3'UTR-APA may play a crucial role in dictating isoform-specific partitioning in these two BMCs. Utilising the approach described in section 4.1.1, we performed the first 3'UTR-APA analysis of the PB and SG transcriptomes. In both instances, we observed large numbers of genes with significant 3'UTR-APA changes. Surprisingly however, given that PBs enriched for longer mRNAs, our results showed that for the PB data set, there were more SIE 3'UTR-APA changes (356, 69.4 %) than LIE changes (157, 30.6 %) (Figure 4.3A). Conversely, for the SG data set our findings were as expected, and we observed a bias towards LIE changes (1421, 64.2 %), in agreement with our previous results. Similarly to DDX4N, the cohorts of genes with significant 3'UTR-APA changes appeared to be largely distinct from the genes with significant gene-specific partitioning changes for both PBs and SGs (Figure 4.3B).

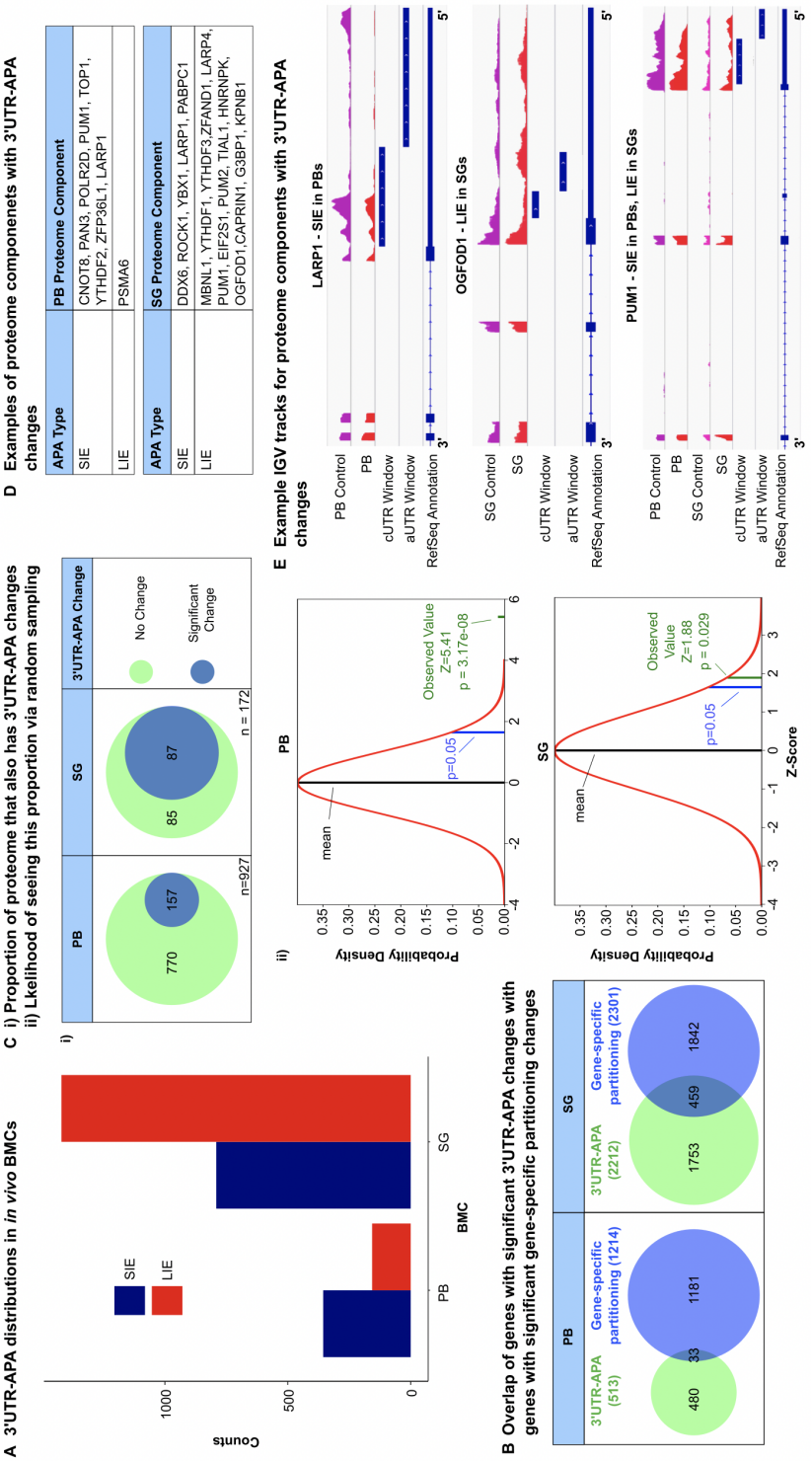

A 3'UTR-APA distributions in in vivo BMCs
SIE
LIE
Counts
1000
500
0
PB
SG
BMC
B Overlap of genes with significant 3'UTR-APA changes with genes with significant gene-specific partitioning changes
PB
SG
3'UTR-APA (513)
Gene-specific partitioning (1214)
3'UTR-APA (2212)
Gene-specific partitioning (2301)
480
33
1181
1753
459
1842
C i) Proportion of proteome that also has 3'UTR-APA changes
ii) Lkelihood of seeing this proportion via random sampling
i)
PB
SG
3'UTR-APA Change
770
157
85
87
No Change
Significant Change
n=927
n = 172
ii)
PB
SG
mean
Observed Value
Z=5.41
p = 3.17e-08
Observed Value
Z=1.88
p = 0.029
p=0.05
Probability Density
Z-Score
D Examples of proteome componenets with 3'UTR-APA changes
APA Type
PB Proteome Component
SIE
CNOT8, PAN3, POLR2D, PUM1, TOP1, YTHDF2, ZFP36L1, LARP1
LIE
PSMA6
APA Type
SG Proteome Component
SIE
DDX6, ROCK1, YBX1, LARP1, PABPC1
LIE
MBNL1, YTHDF1, YTHDF3,ZFAND1, LARP4, PUM1, EIF2S1, PUM2, TIAL1, HNRNPK, OGFOD1,CAPRIN1, G3BP1, KPNB1
E Example IGV tracks for proteome components with 3'UTR-APA changes
LARP1 - SIE in PBs
PB Control
PB
cUTR Window
aUTR Window
RefSeq Annotation
3'
5'
OGFOD1 - LIE in SGs
SG Control
SG
cUTR Window
aUTR Window
RefSeq Annotation
PUM1 - SIE in PBs, LIE in SGs
PB Control
PB
SG Control
SG
cUTR Window
aUTR Window
RefSeq Annotation

Figure 4.3 *(previous page)*: 3'UTR-APA analysis of *in vivo* BMCs. A) Bar graph depicting the distribution of significant 3'UTR-APA changes in the SG and PB BMC data sets. B) Overlap between genes with significant 3'UTR-APA changes and genes with significant gene partitioning changes in both BMC data sets. C) i) Venn diagrams displaying the proportion of the PB and SG proteomes that displayed 3'UTR-APA changes. ii) Normal distribution plots for the % of genes that had 3'UTR-APA changes generated by random sampling of our 3'UTR-APA data sets. The mean of the normal distribution, the Z score where the p value was equal to 0.05, and the Z score for the observed value in our actual data set, are displayed in both cases. D) Examples of key proteome components for PBs and SGs that also show 3'UTR-APA changes. E) Screenshots from the integrated genome browser (IGV) displaying examples of proteome components that had 3'UTR-APA changes in PBs and SGs. The cUTR and aUTR windows are displayed beneath the normalised read values.

PB and SG proteome components are significantly more likely to exhibit 3'UTR-APA changes within their respective condensates than the rest of the transcriptome

Next, we performed GO analysis on genes with significant 3'UTR-APA changes in PBs and SGs to determine whether this process may target genes with specific functionalities. Interestingly, as we had observed for the cohorts of genes with gene-specific partitioning changes, there were many similarities between the two data sets (Figure 4.4). Most notably, the term "intracellular non-membrane-bounded organelle" appeared enriched in both data sets, which hinted at a potential role for 3'UTR-APA in BMC metabolism. To interrogate this, we obtained lists of proteins that were identified as enriched in either PBs or SGs (S. Jain et al. 2016; Hubstenberger et al. 2017). Of the 927 enriched PB proteins, 157 (16.9 %) showed significant 3'UTR-APA changes, while 87 (50.6 %) of the 172 SG protein did (Figure 4.3C i). To test

the significance of these findings we performed random sampling to generate normal distributions based on these percentages (See methods). Remarkably, we found that both percentages were significant with p-values of 3.17e-08, and 0.029 for PBs and SGs respectively (Figure 4.3C ii). This demonstrated that mRNA isoforms of PB and SG proteome components were significantly more likely to exhibit differential selection within their respective condensates than the rest of the transcriptome. This raises the intriguing possibility that 3'UTR-APA may be part of an autoregulatory mechanism for SG and PB condensate formation and maintenance. It is plausible that protein expression of key components is regulated by selective targeting of their mRNA isoforms to their respective condensates.

Figure 4.3D displays a selection of these that are subject to such 3'UTR-APA changes, and figure 4.3E displays the IGV tracks for three. LARP1 is an RNA-binding protein that binds to TOP motifs and has been shown to both positively and negatively regulate transcript stability. The protein localises to both SGs and PBs and in both data sets we observed an SIE 3'UTR-APA change (Hopkins et al. 2016). There is currently much debate as to the function of different LARP1 mRNA isoforms with only one main protein isoform being identified (Schwenzer et al. 2021). Whilst we can't provide resolution on precisely which isoforms are responsible for this change, this finding may provide insight into how LARP1 mRNA isoforms are differentially regulated via isoform-specific enrichment in PBs and SGs. OGFOD1 is intimately linked with the stress response through regulation of eIF2α and it has been shown to localise to SGs (Wehner, Schütz, and Sarnow 2010). PUM1 isoforms were differentially selected in PBs and SGs, with an SIE

change and an LIE change respectively. PUM1 is a post-transcriptional repressor that regulates mRNA fate and is a key component of both SGs and PBs, and it is an intriguing observation that the two condensates appear to enrich for different PUM1 mRNA isoforms (Filipovska et al. 2011; Hubstenberger et al. 2017; Vessey et al. 2006). Overall, our findings demonstrate that 3'UTR-APA likely plays a critical role in isoform-specific partitioning in *in vivo* BMCs. In addition, through specifically targeting the isoforms of key proteome components, it may also regulate PB and SG metabolism.

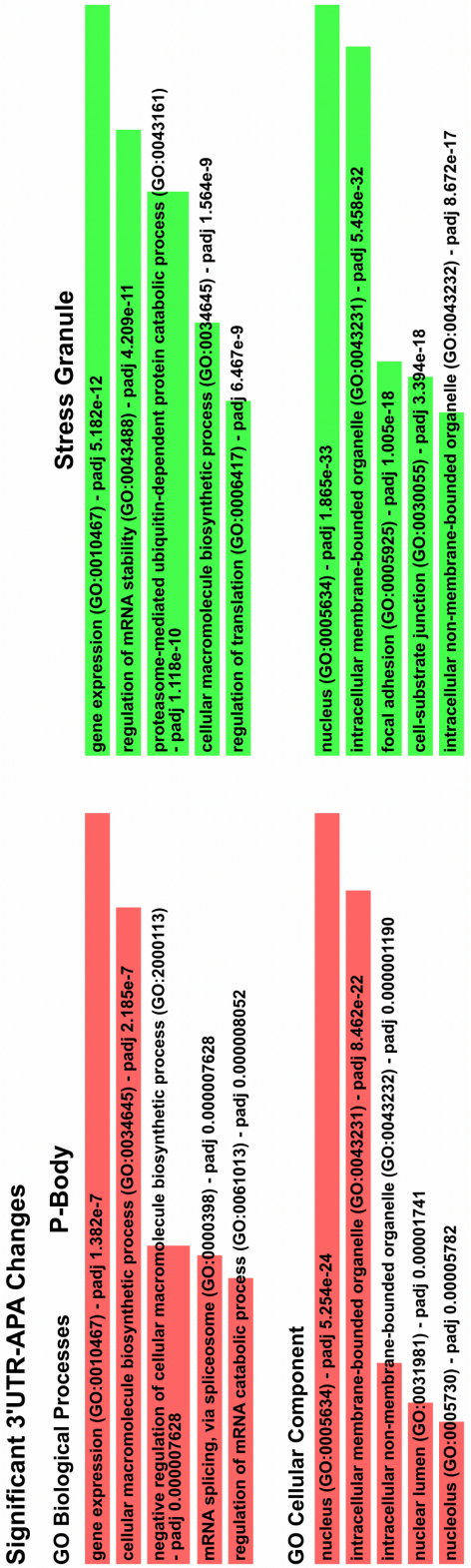


Figure 4.4: GO Analysis of genes with significant 3'UTR-APA changes in SGs and PBs. The top 5 most significant results for GO biological processes and GO cellular components are displayed for each BMC.

PBs and SGs have very similar 3'UTR-APA profiles

Our earlier results corroborated analysis from Matheny *et al.* which indicated that the gene-specific PB and SG transcriptomes were surprisingly comparable (Figure 3.9C) (Matheny, Rao, and Parker 2019). We developed this concept further by investigating the correlation between the 3'UTR-APA profiles. To do so, we utilised the full 3'UTR-APA data sets and performed Pearson correlation analysis using the RED values for each gene in both data sets. Our results showed that the 3'UTR-APA profiles also correlated remarkably well with an R value of 0.42 (Figure 4.5A). Examination of genes with significant 3'UTR-APA changes in both data sets further substantiated this, as we found 82.9 % of these had the same change in distribution in both condensates (i.e. both SIE changes) (Figure 4.5B i). This was particularly surprising since we had demonstrated that 3'UTR-APA changes in PBs were biased towards SIE changes whilst SGs biased towards LIE changes. Random sampling to generate normal distributions based of this percentage confirmed that this observation was highly unlikely to be due to random chance, with the p-value being 2.42e-30 for our observed value (Figure 4.5B ii). Together these findings demonstrate that the rules determining isoform-specific selection in PBs and SGs are likely very similar.

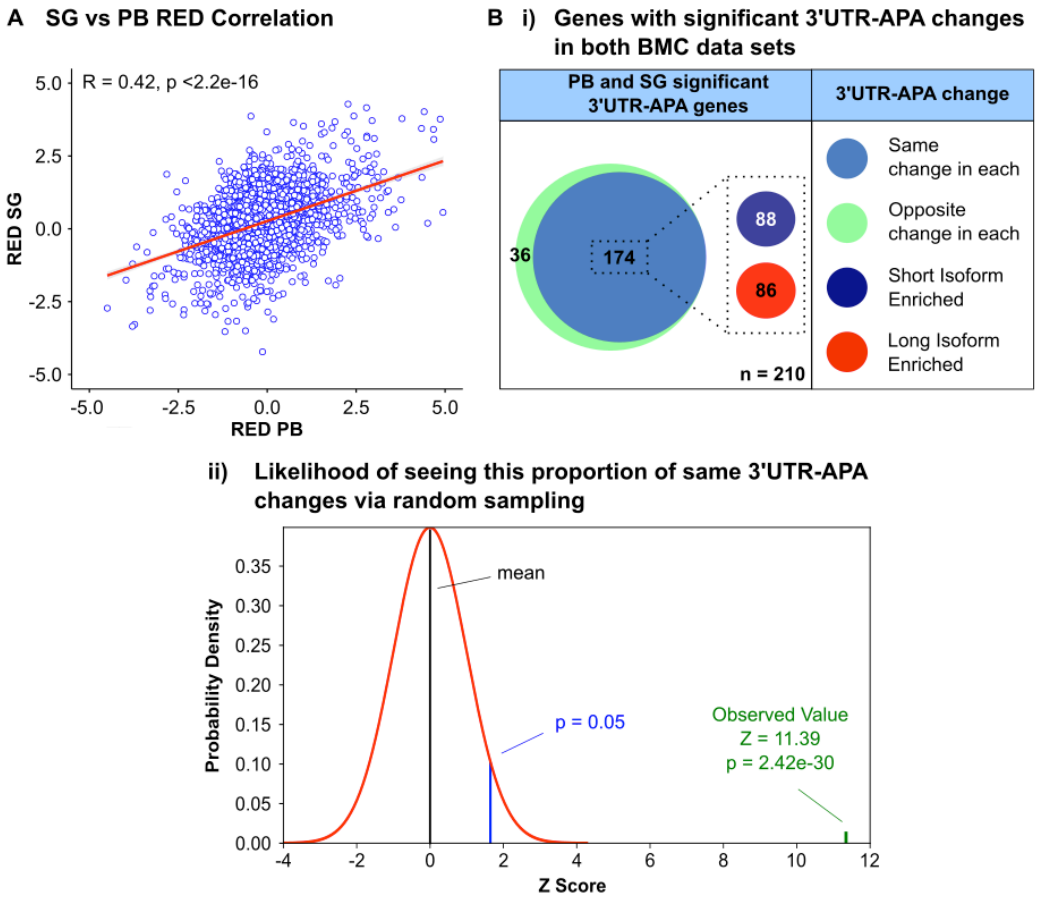


Figure 4.5: Comparison of 3'UTR-APA in PBs and SGs. A) Pearson correlation analysis displaying the RED values for SGs vs PBs, with R value indicated. B) i) Overlap of genes with significant 3'UTR-changes in both the PB and SG data sets and whether they had the same change in both (i.e. both SIE) or whether they were the opposite (i.e SIE in PBs and LIE in SGs). ii) Normal distribution plots for the % of genes that had the same 3'UTR-APA changes generated by random sampling of our 3'UTR-APA data sets. The mean of the normal distribution, the Z score where the p value was equal to 0.05, and the Z score for the observed value for our data set, are displayed.

SIE and LIE gene cohorts have distinct physical properties in PBs and SGs

We demonstrated for DDX4N that the genes in the two 3'UTR-APA cohorts could be distinguished by their physical properties, therefore, we predicted that we may observe similar patterns in PBs and SGs. However, as evidenced in Figure 4.1B, full RNA-seq doesn't provide the same isoform-specific resolution afforded by 3'RNA-seq, therefore, we had to use some approximations in order to perform this analysis. As a proxy for inferring the properties of each gene, we obtained the longest transcript for each (UCSC), as we had done in the previous chapter. In addition, we defined the aUTR for each gene as being between the two most used CPA sites according to the polyA_DB (R. Wang, Nambiar, et al. 2018). Interestingly, for both the longest transcript and aUTR analyses, we observed significant differences between SIE and LIE genes in both data sets (Figure 4.6A). We found in both instances that SIE genes were significantly longer than LIE genes, but that their aUTRs were significantly shorter. Together this suggests that 3'UTR-APA converges on isoforms of similar lengths being enriched in both gene cohorts. For PBs, both the longest transcript and aUTRs were significantly lower in GC % in the SIE cohort, similar to what we had observed for DDX4N. GQFR analysis of the aUTRs, followed by significance testing via random sampling revealed some surprising results (See methods). For PBs we found that the aUTRs of SIE genes were significantly depleted in GQFRs, whilst the aUTRs of LIE genes were significantly enriched in them (Figure 4.6B i,ii). Conversely neither cohort in SGs showed any significant enrichment or depletion. In

the previous chapter, we had demonstrated gene-specific selection in PBs appeared to be independent of GQFRs whilst SGs exhibited a strong dependency. Together, these observations suggest that GQFRs within the aUTRs can serve to induce specific-isoform enrichment in PBs, whilst GQFRs induce gene-specific enrichment in SGs.

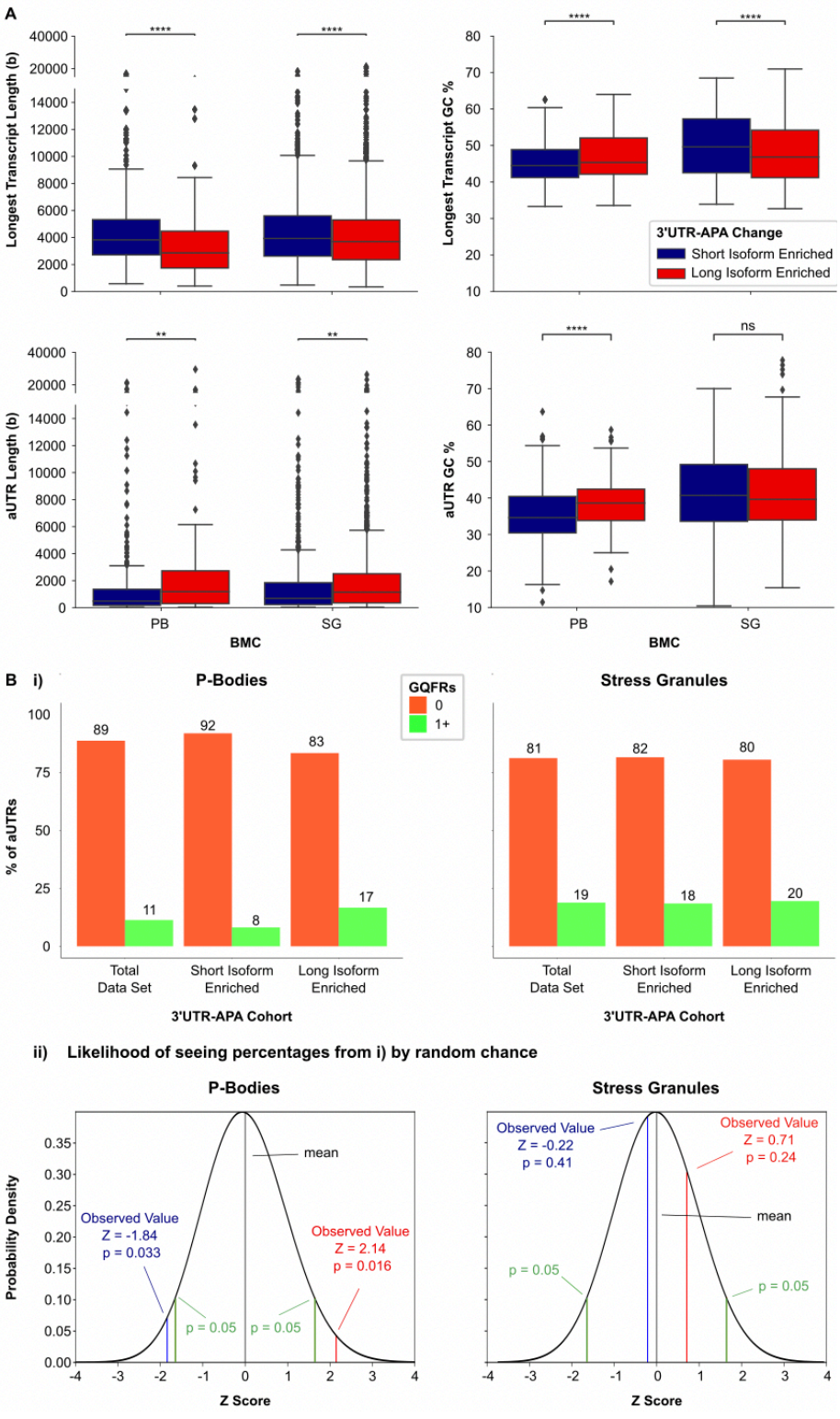

A
Longest Transcript Length (b)
Longest Transcript GC %
aUTR Length (b)
aUTR GC %
****
**
ns
3'UTR-APA Change
Short Isoform Enriched
Long Isoform Enriched
PB
SG
BMC
B i)
P-Bodies
Stress Granules
GQFRs
0
1+
% of aUTRs
89
11
92
8
83
17
81
19
82
18
80
20
Total Data Set
Short Isoform Enriched
Long Isoform Enriched
3'UTR-APA Cohort
ii) Likelihood of seeing percentages from i) by random chance
Probability Density
Z Score
mean
Observed Value
Z = -1.84
p = 0.033
Observed Value
Z = 2.14
p = 0.016
p = 0.05
Observed Value
Z = -0.22
p = 0.41
Observed Value
Z = 0.71
p = 0.24

Figure 4.6 *(previous page)*: Analysis of genes with 3'UTR-APA changes in PBs and SGs. A) Boxplots depicting the length and GC % of the longest transcript and aUTRs of genes with 3'UTR-APA changes in PBs and SGs. The genes are split by whether they were in the SIE cohort or LIE cohort in each data set. B) i) Bar plots displaying percentage of aUTRs with 1 or more GQFR in each 3'UTR-APA cohort, as well as the whole data set, for P-Bodies and Stress Granules. Percentages for each bar are displayed above. ii) Normal distribution plots displaying the likelihood of seeing the percentages from B i) by random chance. The observed value in blue represents the value for the SIE cohort, whilst the value in red represents the value for the LIE cohort. Significance values displayed above A were acquired using a T-test; ns = p > 0.05, * = p <= 0.05, ** = p <= 0.01, *** = p <= 0.001, **** = p <= 0.0001.

RNA-binding sites in the aUTRs of LIE genes likely don't influence isoform-specific selection in SGs and PBs

It has been suggested previously that TIA1, an RBP and SG component, can target U-rich sequences in the aUTRs of mRNA, promoting their inclusion into SGs (D. Zheng et al. 2018). We addressed this by performing streme analysis on the aUTRs of genes with LIE changes in both data sets (T. L. Bailey 2021). Surprisingly however, we found no motifs that were significantly enriched (p-value <= 0.05) in these sequences in either of the two data sets. This finding suggests that enrichment of the long isoform in LIE genes is likely independent of specific RBPs associating with sequences within their aUTRs.

4.3 Discussion

The focus of this chapter was to investigate how 3'UTR-APA can affect mRNA selection in BMCs. 70 % of human genes encode multiple mRNA isoforms with the same protein coding capacity, that differ only in the length of their 3'UTRs (Neve and Furger 2014; Derti et al. 2012; Hoque, W. Li, and B. Tian 2014; Neve, R. Patel, et al. 2017). Global 3'UTR-APA profiles are influenced by cell type, stress, and disease but there remains little consensus as to the biological consequences of this phenomenon (R. Patel et al. 2019; Mohanan et al. 2021; Fischl et al. 2019; Neve and Furger 2014). We demonstrated in the previous chapter that the physical properties of mRNA are a critical determinant in how it partitions. Here we propose a novel function for 3'UTR-APA in modifying these physcial properties, allowing the cell to selectively target mRNA isoforms to BMCs, which may have consequences for mRNA metabolism and gene expression regulation.

Our 3'RNA-sequencing approach provided us with isoform-specific resolution with which to interrogate 3'UTR-APA in DDX4N and FUS-GFP. This analysis entirely corroborated our previous findings, with 3'UTR-APA altering the partitioning of mRNA in a predictable manner, governed by the selection principles we had previously established for each model condensate. Consistent with DDX4N enriching for longer mRNA, we found that 86 % of human genes had LIE 3'UTR-APA changes. We observed very few significant changes for the *Drosophila* data set. However, this was readily be explained by *Drosophila* aUTRs being very short, and thus not able to significantly alter the properties of, and consequently the partitioning of, dif-

ferent mRNA isoforms. This is corroborated by the observation that genes with significant 3'UTR-APA changes in the human data set had significantly longer aUTRs than those without (Figure 4.2D). Furthermore, in agreement with our finding that mRNA selection in FUS-GFP condensates is independent of mRNA length, we showed that 3'UTR-APA had no effect on selective isoform partitioning.

We demonstrated that the 3'UTR-APA cohorts from the DDX4N human data set could be distinguished by GC %, with LIE genes having a much higher GC %. This finding converged on an identical result to the one we had previously made that DDX4N enriches for short, AU-rich mRNA or long GC-rich mRNA. Thus, two fundamentally different analysis approaches came to the same conclusion, further emphasising the crucial role mRNA physical properties have on condensate selection. In addition to this, we again showed GQFRs are highly important in DDX4N selection; the aUTRs of LIE genes were twice as likely to contain a GQFR compared to the SIE genes (Figure 4.2G). Most genes that encode mRNAs with GQFRs within their UTRs also encode mRNAs without a GQFR within their UTR (D. S. Lee, Ghanem, and Barash 2020). Our result may demonstrate a function of this, as it suggests that 3'UTR-APA can equip mRNAs with additional GQ-structures that can help to selectively target isoforms to BMCs.

In this chapter, we performed the first in depth 3'UTR-APA analysis on the PB and SG transcriptomes. We show that 3'UTR-APA is highly likely to play a critical role in shaping the composition of both transcriptomes, with our results revealing large numbers of genes with significant changes (Figure 4.3A). Interestingly, genes that underwent 3'UTR-APA changes were in

both instances, largely distinct from those with significant gene partitioning changes (Figure 4.3B). This observation was also true for our DDX4N system and suggests that selection can be either "gene-specific" or "isoform-specific". In agreement with previous findings, and in line with SGs enriching for longer mRNAs, we saw a bias towards LIE changes in SGs (Khong et al. 2017; D. Zheng et al. 2018). Surprisingly, the opposite was true for PBs, which is contrary to the idea that PBs enrich for longer mRNAs (Courel et al. 2019; Matheny, Rao, and Parker 2019). To investigate this further, we examined the properties of the aUTRs and longest transcripts for each gene in each 3'UTR-APA cohort for each BMC. Significant differences were found in both, with SIE genes having significantly longer longest transcript lengths but significantly shorter aUTRs than LIE genes (Figure 4.6A/B). This suggests that the length of the enriched isoform in both cohorts may converge to be very similar. Surprisingly, GQFRs in the aUTRs of genes appeared influential in PBs but not in SGs. This is contrary to our finding in the previous chapter that GQFRs were a key determinant in SG enrichment and not PBs (Figure 4.6C). Collectively, these observations strongly agree with the notion that the physical properties of an mRNA readily influence its selection in both condensates. However, whilst this is clearly the case, we do observe some discrepancies with the results from the previous chapter which can't be readily explained. Owing to the limitations of isoform-specific analysis with data obtained from full RNA-seq, further investigation will be necessary, perhaps through 3'RNA-seq or nanopore sequencing, to uncover the precise rules responsible for these results. Nevertheless, the fact that we do observe differences in the results between the two approaches empha-

sises the need to reexamine this process in an isoform-specific manner for these condensates. We were able to do this for DDX4N and consequently our findings were highly consistent.

Most remarkably we found that the genes encoding the SG and PB proteomes were significantly more likely to exhibit 3'UTR-APA changes within their respective BMCs (Figure 4.3C). This result may provide evidence for a complex autoregulatory mechanism for controlling condensate formation. PBs and SGs enrich in poorly-translating mRNAs and it is possible that through isoform-specific selection, the cell has a mechanism with which to manipulate the protein expression of these components and coordinate it with condensate formation and maintenance (Hubstenberger et al. 2017; Khong et al. 2017; S. Jain et al. 2016). 3'UTR-APA is known to affect the translational efficiency of mRNA isoforms and whether this is a consequence of isoform-specific sequestration in BMCs or via other means, it's clear how this may be influential in condensate formation (Yoon et al. 2013; Neve and Furger 2014). In addition to this, it is highly possible that the different mRNA isoforms of these components, may also have different potential with regard to seeding formation of these condensates. This concept has been demonstrated in the instance of NEAT1 RNA where the long non-coding transcript is able to seed paraspeckle formation, whilst the shorter polyadenylated transcript is unable to (Yamazaki et al. 2018; Taiana et al. 2020). Since 3'UTR-APA is also highly sensitive to changes in the cells environment, it is intuitive how this pathway could integrate signals, for example from stress, and provide a way to rapidly induce changes in condensate abundance (D. Zheng et al. 2018; Hollerer et al. 2016; Neve, Burger, et al. 2016).

Gene-specific enrichment correlation analysis performed in chapter 3, corroborated previous findings that the PB and SG transcriptomes were remarkably similar (Matheny, Rao, and Parker 2019). Here we revealed that the same applied to their 3'UTR-APA profiles, which correlated with an R value of 0.42 (Figure 4.5A). Supporting this idea, we found that genes with significant 3'UTR-APA changes in both data sets, were significantly more likely to exhibit the same type of 3'UTR-APA change. This finding emphasises the notion that similar selection principles may determine mRNA enrichment in PBs and SGs (Matheny, Rao, and Parker 2019).

4.4 Summary

In summary, we performed the first comprehensive study examining the role of 3'UTR-APA in selective recruitment of mRNA isoforms to *in vitro* and *in vivo* BMCs. Our *in vitro* findings entirely corroborated and substantiated our results from the previous chapter, as we determined 3'UTR-APA manipulated isoform partitioning in tune with the selection principles we had established. We showed large numbers of genes were subject to 3'UTR-APA regulation in SGs and PBs, indicating this process is highly likely to also play a critical role *in vivo*. Critically, our analysis revealed that genes encoding proteome components for both, were significantly more likely to exhibit 3'UTR-APA changes. This finding raises the intriguing possibility that 3'UTR-APA may be involved in autoregulation of these condensates, both at the protein and the mRNA level. Overall, we propose a novel role for 3'UTR-APA which the cell can utilise to selectively target specific mRNA

isoforms to BMCs.

Chapter 5

Discussion and Future Directions

5.1 Thesis Summary

The driving forces behind selective partitioning of mRNA in biomolecular condensates are poorly understood. Many studies have previously focused exclusively on the importance of ribosomes, or RNA-binding proteins in this process. Here we addressed this issue from an mRNA perspective. We utilised deep sequencing in combination with bioinformatic analysis to interrogate how the physical properties of mRNA defined its selection in *in vitro* and *in vivo* BMCs. We performed the first comprehensive analysis examining a novel role for 3'UTR-APA in shaping the distribution of mRNA isoforms in these systems. Overall, our findings strongly suggest that the prevailing determinant of mRNA selection in BMCs is imparted by the physical properties of the mRNA itself.

The importance of RNA in assembly and maintenance of BMCs is only just beginning to emerge (Decker, J. M. Burke, et al. 2021). In Chapter 3 we developed a simplified *in vitro* system which allowed us to address this concept. By challenging a single protein component with cytoplasmic RNA, we were able to interrogate specifically the role of RNA in governing its own partitioning. We established a bioinformatic partitioning pipeline to characterise the transcriptomes of three model BMCs, and evaluate the physical properties of mRNAs that were enriched, neither, and depleted in each condensate. We utilised the same pipeline to compare our findings with data obtained from highly complex *in vivo* systems. Our findings highlighted the importance of mRNA physical properties in construing how it partitioned and this lead us to investigate a cellular process capable of manipulating these. To this end, in Chapter 4, we employed two separate bioinformatics pipelines to establish a crucial role for 3'UTR-APA in altering the partitioning of mRNA isoforms, with significant consequences for both mRNA and condensate metabolism.

The key findings in this study are as follows:

1. mRNA length, GC %, and GQFRs are critical determinants in governing its selection in both *in vitro* and *in vivo* BMCs
2. Commonalities in the function of the human and *Drosophila* mRNA that were enriched and depleted in DDX4N suggests an evolutionary conservation in selection, designated by an mRNA's intrinsic properties
3. The *in vitro* DDX4N transcriptome shares remarkable similarities with

the *in vivo* PB transcriptome, indicative highly similar rules dictating mRNA selection in both systems.

4. 3'UTR-APA has the capacity to alter mRNA isoform distribution in BMCs based on evidence from both *in vitro* and *in vivo* systems.

5. 3'UTR-APA may be involved in autoregulation of condensate assembly *in vivo*.

5.2 *In vitro* approach

One of the major concerns with *in vitro* studies of BMCs and BMC components is that they fail to capture the complexity of *in vivo* systems. Consequently, it is often hard to determine whether the results are reflective of what happens *in vivo*, or whether they are simply artifacts. There are two major findings in this study which we strongly consider to be evidence that our results are highly relevant. The first is that the trends we observed in our *in vitro* systems were very similar to those observed in *in vivo* systems; where the length was influential in mRNA selection, longer transcripts were enriched; where the GC % was influential, AU-rich transcripts were enriched. It is hard to envisage that such consistencies would emerge if the same driving forces weren't responsible in both systems. Secondly, we observed a significant correlation between the transcriptome captured by our model DDX4N system and that of *in vivo* PBs. The two condensates shared the most similar selection principles based on the mRNA physical properties and this further substantiates the idea that similar rules must be responsible for defining selection in both. In support of this observation, yeast RNA, in the presence

of crowding agent, has been demonstrated to self-assemble *in vitro* and recapitulate the transcriptome observed in yeast stress granules (Van Treeck, Protter, et al. 2018). Both results converge on a model where RNA-RNA interactions likely dominate in defining mRNA condensate selection. Indeed, this is an idea that is now central to many of the most recent studies (Begovich and Wilhelm 2020; Matheny, Van Treeck, et al. 2021). This narrative was only made apparent through the use of simplified *in vitro* systems, which demonstrates their continued importance in developing our understanding of these complex biological phenomena.

5.3 How do the physical properties of mRNA determine its selection?

Our approach enabled us to identify the physical properties of mRNA that defined how it partitioned, but the precise molecular mechanisms that underpin our observations remain to be fully established. It is hypothesised that enrichment for longer mRNAs is a consequence of longer mRNAs being able to engage in more stabilising, multivalent interactions than shorter mRNAs (Khong et al. 2017; Namkoong et al. 2018; Hubstenberger et al. 2017). Whilst we found that length served as a discriminator in three of the systems we investigated, two also displayed length-independent selection, demonstrating that this isn't the sole determinant of mRNA selection in BMCs. The role of mRNA GC % becomes harder to rectify outside of a cellular context. Explanations for why low GC % corresponds to enrichment in BMCs have

focused on how this corresponds to low ribosomal occupancy or an increased frequency of AU-rich elements, which are the target sites of RNA-binding proteins such as TIA-1 (Courel et al. 2019; Namkoong et al. 2018; D. Zheng et al. 2018). However, we observed the same trend in systems that lack both ribosomes and proteins with RNA-binding domains, which strongly suggests an alternative mechanism may be responsible. One possibility is that mRNAs with higher GC % form more stable secondary structures that may prevent the bases involved from engaging in additional multivalent interactions (Chan et al. 2009; Faure et al. 2016). Another explanation may be due to different post-transcriptional modifications of the nucleic acid bases. Indeed, m^6A modification is the most abundant modification in eukaryotes, and it's feasible that this may contribute to preferential enrichment of AU-rich mRNAs (Wei et al. 2017). Unfortunately, both full RNA-sequencing and 3'RNA-sequencing do not provide resolution on post-transcriptional modifications and as such more work will be required to investigate this prospect. We demonstrated that GQFRs were another property with the potential to influence mRNA partitioning. It has been shown that GQFR-containing mRNAs have a propensity to phase separate in the absence of protein, with the suggestion that these structures provide an axis for RNA-RNA interactions (Yueying Zhang et al. 2019; Conlon et al. 2016; Fay, P. J. Anderson, and Ivanov 2017). This may provide an explanation for why we saw a bias towards enrichment of GQFR-containing mRNAs, as these transcripts may recruit one another into BMCs. Additionally, repeat expansion mRNAs containing GQs have been demonstrated to promote robust assembly of SGs *in vivo* (Fay, P. J. Anderson, and Ivanov 2017). Therefore, it is feasible that SG

formation may be a consequence of GQFR-containing mRNAs coalescing.

A key observation is that whilst there are many similarities in the trends determining selection in each of the studied BMCs, there are no set of rules that apply universally (Figure 3.8). This suggests that each creates a unique solvent environment that enriches for mRNA with specific physical properties. Interestingly, we demonstrated that the selection principles defining our positively-charged models differed from our negatively-charged model. This may serve as insight into how BMCs are able to maintain their unique compositions and enrich for mRNAs specifically related to their function. Moreover, we demonstrated that DDX4N enriched for functionally-related human and *Drosophila* mRNAs, suggesting that mRNA selection into BMCs may be conserved, encoded by the transcripts physical properties (Figure 3.4).

To further probe these findings it would be possible to employ an *in vitro* transcription approach to synthesise mRNAs with distinct physical characteristics. For example, mRNAs with and without GQFRs could be transcribed and fluorescently-labelled, and their partitioning in our model systems could be characterised by confocal microscopy. Another option would be to use nanopore sequencing which would provide precise mRNA sequences for the condensate transcriptomes, as well as information about post-transcriptional modifications. Another aspect of this project that requires further investigation is the function of RNA-binding domains in conferring mRNA selectivity. We propose a model whereby the mRNA physical properties are the key determinant in this process, largely independent of specific RNA binding. One way to address this would be to utilise the full length DDX4 protein, and

compare the mRNA it sequesters with that of our DDX4N model to establish how it may differ, if at all.

5.3.1 A critical role for 3'UTR-APA in BMC metabolism?

This is the first study to deeply investigate 3'UTR-APA in BMCs and we propose a novel role for this process in selectively disseminating mRNA isoforms to these structures. In our *in vitro* systems, we were able to demonstrate that 3'UTR-APA altered partitioning in a highly predictable manner, in tune with the selection principles of the model condensate. 3'UTR-APA profiles are highly dynamic and can rapidly respond to changes in the cells environment (Neve and Furger 2014; D. Zheng et al. 2018). Moreover, it has been demonstrated that mRNA is not only often a critical component in BMCs, but additionally, that it can serve as a molecular seed in inducing their formation (Bond and Fox 2009; Clemson et al. 2009; Yamazaki et al. 2018; Decker, J. M. Burke, et al. 2021; Teixeira et al. 2005). Together, this raises the intriguing possibility that 3'UTR-APA could integrate cellular signals to produce isoform cohorts with the specific physical properties requisite for the assembly of BMCs *in vivo*. 3'UTR-APA is highly conserved in eukaryotes, however there remains much debate as to why (Neve and Furger 2014). It is possible that mRNA isoforms have a functionality outside of a protein-coding capacity, which is to serve primarily as a key components in RNA granules. Perhaps indicative of this, we found that PB and SG proteome components were significantly more likely to exhibit 3'UTR-APA changes in their respective condensates than the rest of the transcriptome

(Figure 4.3C). This alludes to a highly complex autoregulatory mechanism whereby 3'UTR-APA provides the cell with a means to alter the concentrations of constituent mRNA or protein BMC components in order to induce their assembly or disassembly.

Further investigation is an absolute necessity to fully explore these ideas. It may be possible to use an approach such as RNA interference to specifically target certain mRNA isoforms that are enriched in BMCs and investigate how condensate assembly is affected as a consequence. However, it has been demonstrated by numerous studies investigating the effect of knocking down specific protein BMC components that there are often redundancies (Matsuki et al. 2013; Hofmann et al. 2021). In addition, owing to the limitations of full-RNA-seq we were forced to make several approximations when investigating 3'UTR-APA in SGs and PBs. This led to some discrepancies between our results from Chapter 3 and Chapter 4 for these two condensates, whilst our results for DDX4N were highly consistent. As a consequence, there is a critical importance to reinvestigate these results utilising a different sequencing approach in order to fully elucidate isoform-specific selection in these *in vivo* BMCs.

5.4 Conclusion

By using simplified *in vitro* systems, this study was in a unique position to interrogate the driving forces behind mRNA selection in BMCs. We demonstrated that highly similar trends persisted between our model systems and *in vivo* systems. By removing many of the complexities of *in vivo* systems

(such as RBPs and ribosomes), that have often been attributed to defining this process, we were able to establish that the physical properties of mRNA are the critical factor in dictating how it partitions. We propose a novel role for 3'UTR-APA in producing mRNA isoforms with distinct partitioning properties. This has profound implications not only for mRNA metabolism and gene expression, but also BMC regulation.

However, there still remain many unanswered questions. What are the molecular interactions that make the physical characteristics of mRNA so important in this process? How intimately linked is 3'UTR-APA to condensate metabolism? Do the 3'UTR-APA profiles seen in disease also play a role in aberrant BMC assembly or dynamics?

Bibliography

Adivarahan, Srivathsan et al. (2018). "Spatial organization of single mRNPs at different stages of the gene expression pathway". In: *Molecular cell* 72.4, pp. 727–738.

Afgan, Enis et al. (2018). "The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update". In: *Nucleic acids research* 46.W1, W537–W544.

Akman, Begum H, Tolga Can, and A Elif Erson-Bensan (2012). "Estrogen-induced upregulation and 3-UTR shortening of CDC6". In: *Nucleic acids research* 40.21, pp. 10679–10688.

Alberti, Simon and Dorothee Dormann (2019). "Liquid–liquid phase separation in disease". In: *Annual review of genetics* 53, pp. 171–194.

Alberti, Simon and Anthony A Hyman (2016). "Are aberrant phase transitions a driver of cellular aging?" In: *BioEssays* 38.10, pp. 959–968.

Anders, Maximilian et al. (2018). "Dynamic m6A methylation facilitates mRNA triaging to stress granules". In: *Life science alliance* 1.4.

Andreassi, Catia and Antonella Riccio (2009). "To localize or not to localize: mRNA fate is in 3 UTR ends". In: *Trends in cell biology* 19.9, pp. 465–474.

Andrei, Maria Alexandra et al. (2005). "A role for eIF4E and eIF4E-transporter in targeting mRNPs to mammalian processing bodies". In: *Rna* 11.5, pp. 717–727.

Arimoto, Kyoko et al. (2008). "Formation of stress granules inhibits apoptosis by suppressing stress-responsive MAPK pathways". In: *Nature cell biology* 10.11, pp. 1324–1332.

Audas, Timothy E et al. (2016). "Adaptation to stressors by systemic protein amyloidogenesis". In: *Developmental cell* 39.2, pp. 155–168.

Aumiller Jr, William M et al. (2016). "RNA-based coacervates as a model for membraneless organelles: formation, properties, and interfacial liposome assembly". In: *Langmuir* 32.39, pp. 10042–10053.

Ayache, Jessica et al. (2015). "P-body assembly requires DDX6 repression complexes rather than decay or Ataxin2/2L complexes". In: *Molecular biology of the cell* 26.14, pp. 2579–2595.

Ayala, Youhna M et al. (2008). "Structural determinants of the cellular localization and shuttling of TDP-43". In: *Journal of cell science* 121.22, pp. 3778–3785.

Bailey, Timothy L (2021). "STREME: accurate and versatile sequence motif discovery". In: *Bioinformatics* 37.18, pp. 2834–2840.

Balbo, Paul B and Andrew Bohm (2007). "Mechanism of poly (A) polymerase: structure of the enzyme-MgATP-RNA ternary complex and kinetic analysis". In: *Structure* 15.9, pp. 1117–1131.

Banani, Salman F, Hyun O Lee, et al. (2017). "Biomolecular condensates: organizers of cellular biochemistry". In: *Nature reviews Molecular cell biology* 18.5, pp. 285–298.

Banani, Salman F, Allyson M Rice, et al. (2016). "Compositional control of phase-separated cellular bodies". In: *Cell* 166.3, pp. 651–663.

Banjade, Sudeep and Michael K Rosen (2014). "Phase transitions of multivalent proteins can promote clustering of membrane receptors". In: *Elife* 3, e04123.

Begovich, Kyle and James E Wilhelm (2020). "An in vitro assembly system identifies roles for RNA nucleation and ATP in yeast stress granule formation". In: *Molecular Cell* 79.6, pp. 991–1007.

Behm-Ansmant, Isabelle et al. (2007). "A conserved role for cytoplasmic poly (A)-binding protein 1 (PABPC1) in nonsense-mediated mRNA decay". In: *The EMBO journal* 26.6, pp. 1591–1601.

Berry, Joel et al. (2015). "RNA transcription modulates phase transition-driven nuclear body assembly". In: *Proceedings of the National Academy of Sciences* 112.38, E5237–E5245.

Blasco-Moreno, Bernat et al. (2019). "The exonuclease Xrn1 activates transcription and translation of mRNAs encoding membrane proteins". In: *Nature communications* 10.1, pp. 1–15.

Boccaletto, Pietro et al. (2018). "MODOMICS: a database of RNA modification pathways. 2017 update". In: *Nucleic acids research* 46.D1, pp. D303–D307.

Boeynaems, Steven et al. (2019). "Spontaneous driving forces give rise to protein- RNA condensates with coexisting phases and complex material properties". In: *Proceedings of the National Academy of Sciences* 116.16, pp. 7889–7898.

Boija, Ann et al. (2018). "Transcription factors activate genes through the phase-separation capacity of their activation domains". In: *Cell* 175.7, pp. 1842–1855.

Boisvert, François-Michel et al. (2007). "The multifunctional nucleolus". In: *Nature reviews Molecular cell biology* 8.7, pp. 574–585.

Bond, Charles S and Archa H Fox (2009). "Paraspeckles: nuclear bodies built on long noncoding RNA". In: *Journal of Cell Biology* 186.5, pp. 637–644.

Boyko, Solomiia et al. (2019). "Liquid–liquid phase separation of tau protein: the crucial role of electrostatic interactions". In: *Journal of Biological Chemistry* 294.29, pp. 11054–11059.

Brangwynne, Clifford P, Christian R Eckmann, et al. (2009). "Germline P granules are liquid droplets that localize by controlled dissolution/condensation". In: *Science* 324.5935, pp. 1729–1732.

Brangwynne, Clifford P, Timothy J Mitchison, and Anthony A Hyman (2011). "Active liquid-like behavior of nucleoli determines their size and shape in Xenopus laevis oocytes". In: *Proceedings of the National Academy of Sciences* 108.11, pp. 4334–4339.

Brangwynne, Clifford P, Peter Tompa, and Rohit V Pappu (2015). "Polymer physics of intracellular phase transitions". In: *Nature Physics* 11.11, pp. 899–904.

Brengues, Muriel, Daniela Teixeira, and Roy Parker (2005). "Movement of eukaryotic mRNAs between polysomes and cytoplasmic processing bodies". In: *Science* 310.5747, pp. 486–489.

Brown, Reid F, Casey T Andrews, and Adrian H Elcock (2018). "Correction to Stacking Free Energies of All DNA and RNA Nucleoside Pairs and

Dinucleoside-Monophosphates Computed Using Recently Revised AMBER Parameters and Compared with Experiment". In: *Journal of chemical theory and computation* 14.12, pp. 6742–6743.

Buchan, J. Ross and Roy Parker (Dec. 2009). "Eukaryotic Stress Granules: The Ins and Out of Translation". In: *Molecular cell* 36.6, p. 932. DOI: 10.1016/J.MOLCEL.2009.11.020. URL: /pmc/articles/PMC2813218/%20/pmc/articles/PMC2813218/?report=abstract%20https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2813218/.

Cajal, S Ramón et al. (1903). "Un sencillo metodo de coloracion seletiva del reticulo protoplasmatico y sus efectos en los diversos organos nerviosos de vertebrados e invertebrados". In: *Trab. Lab. Invest. Biol.(Madrid)* 2, pp. 129–221.

Cajal, Santiago Ramón y (1910). *El núcleo de las células piramidales del cerebro humano y de algunos mamıferos*.

Carlson, Marc and Bioconductor Package Maintainer (2015a). *TxDb.Dmelanogaster.UCSC.dm Annotation package for TxDb object(s)*. R package version 3.2.2.

— (2015b). *TxDb.Hsapiens.UCSC.hg19.knownGene: Annotation package for TxDb object(s)*. R package version 3.2.2.

Castello, Alfredo et al. (2012). "Insights into RNA biology from an atlas of mammalian mRNA-binding proteins". In: *Cell* 149.6, pp. 1393–1406.

Castrillon, Diego H et al. (2000). "The human VASA gene is specifically expressed in the germ cell lineage". In: *Proceedings of the National Academy of Sciences* 97.17, pp. 9585–9590.

Chan, Chi Yu et al. (2009). "A structural interpretation of the effect of GC-content on efficiency of RNA interference". In: *BMC bioinformatics* 10.1, pp. 1–7.

Chen, Edward Y et al. (2013). "Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool". In: *BMC bioinformatics* 14.1, pp. 1–14.

Cid-Samper, Fernando et al. (2018). "An integrative study of protein-RNA condensates identifies scaffolding RNAs and reveals players in fragile X-associated tremor/ataxia syndrome". In: *Cell reports* 25.12, pp. 3422–3434.

Clemson, Christine M et al. (2009). "An architectural role for a nuclear non-coding RNA: NEAT1 RNA is essential for the structure of paraspeckles". In: *Molecular cell* 33.6, pp. 717–726.

Collins, Mark O et al. (2008). "Phosphoproteomic analysis of the mouse brain cytosol reveals a predominance of protein phosphorylation in regions of intrinsic sequence disorder". In: *Molecular & Cellular Proteomics* 7.7, pp. 1331–1348.

Conicella, Alexander E et al. (2016). "ALS mutations disrupt phase separation mediated by α-helical structure in the TDP-43 low-complexity C-terminal domain". In: *Structure* 24.9, pp. 1537–1549.

Conlon, Erin G et al. (2016). "The C9ORF72 GGGGCC expansion forms RNA G-quadruplex inclusions and sequesters hnRNP H to disrupt splicing in ALS brains". In: *Elife* 5, e17820.

Courchaine, Edward M et al. (2021). "DMA-tudor interaction modules control the specificity of in vivo condensates". In: *Cell*.

Courel, Maıté et al. (2019). "GC content shapes mRNA storage and decay in human cells". In: *Elife* 8, e49708.

Couthouis, Julien et al. (2012). "Evaluating the role of the FUS/TLS-related gene EWSR1 in amyotrophic lateral sclerosis". In: *Human molecular genetics* 21.13, pp. 2899–2911.

D'Lima, Nadia G et al. (2017). "A human microprotein that interacts with the mRNA decapping complex". In: *Nature chemical biology* 13.2, pp. 174–180.

De Santis, Riccardo et al. (2019). "Mutant FUS and ELAVL4 (HuD) aberrant crosstalk in amyotrophic lateral sclerosis". In: *Cell reports* 27.13, pp. 3818–3831.

De Stanchina, Elisa et al. (2004). "PML is a direct p53 target that modulates p53 effector functions". In: *Molecular cell* 13.4, pp. 523–535.

Debye, Peter and Erich Hückel (1923). "Zur Theorie der Elektrolyte. I. Gefrierpunktserniedrigung und verwandte Erscheinungen". In: *Physikalische Zeitschrift* 24.185, p. 305.

Decker, Carolyn J, James M Burke, et al. (2021). "RNA is required for the maintenance of multiple cytoplasmic and nuclear membrane-less organelles." In: *bioRxiv*.

Decker, Carolyn J, Daniela Teixeira, and Roy Parker (2007). "Edc3p and a glutamine/asparagine-rich domain of Lsm4p function in processing body assembly in Saccharomyces cerevisiae". In: *The Journal of cell biology* 179.3, pp. 437–449.

DeJesus-Hernandez, Mariely et al. (2011). "Expanded GGGGCC hexanucleotide repeat in noncoding region of C9ORF72 causes chromosome 9p-linked FTD and ALS". In: *Neuron* 72.2, pp. 245–256.

Denis, Clyde L and Junji Chen (2003). "The CCR4-NOT complex plays diverse roles in mRNA metabolism." In: *Progress in nucleic acid research and molecular biology* 73, pp. 221–250.

Derti, Adnan et al. (2012). "A quantitative atlas of polyadenylation in five mammals". In: *Genome research* 22.6, pp. 1173–1183.

Ditlev, Jonathon A, Lindsay B Case, and Michael K Rosen (2018). "Who's in and who's out—compositional control of biomolecular condensates". In: *Journal of molecular biology* 430.23, pp. 4666–4684.

Dodson, Anne E and Scott Kennedy (2019). "Germ granules coordinate RNA-based epigenetic inheritance pathways". In: *Developmental cell* 50.6, pp. 704–715.

Dormann, Dorothee and Christian Haass (2013). "Fused in sarcoma (FUS): an oncogene goes awry in neurodegeneration". In: *Molecular and Cellular Neuroscience* 56, pp. 475–486.

Drino, Aleksej and Matthias R Schaefer (2018). "RNAs, phase separation, and membrane-less organelles: Are post-transcriptional modifications modulating organelle dynamics?" In: *BioEssays* 40.12, p. 1800085.

Du, Mingjian and Zhijian J Chen (2018). "DNA-induced liquid phase condensation of cGAS activates innate immune signaling". In: *Science* 361.6403, pp. 704–709.

Dutagaci, Bercem et al. (2021). "Charge-driven condensation of RNA and proteins suggests broad role of phase separation in cytoplasmic environments". In: *Elife* 10, e64004.

Dzatko, Simon et al. (2018). "Evaluation of the stability of DNA i-Motifs in the nuclei of living mammalian cells". In: *Angewandte Chemie International Edition* 57.8, pp. 2165–2169.

Elbaum-Garfinkle, Shana et al. (2015). "The disordered P granule protein LAF-1 drives phase separation into droplets with tunable viscosity and dynamics". In: *Proceedings of the National Academy of Sciences* 112.23, pp. 7189–7194.

Ephrussi, Anne, Laura K Dickinson, and Ruth Lehmann (1991). "Oskar organizes the germ plasm and directs localization of the posterior determinant nanos". In: *Cell* 66.1, pp. 37–50.

Eulalio, Ana, Isabelle Behm-Ansmant, and Elisa Izaurralde (2007). "P bodies: at the crossroads of post-transcriptional pathways". In: *Nature reviews Molecular cell biology* 8.1, pp. 9–22.

Faure, Guilhem et al. (2016). "Role of mRNA structure in the control of protein folding". In: *Nucleic acids research* 44.22, pp. 10898–10911.

Fay, Marta M, Paul J Anderson, and Pavel Ivanov (2017). "ALS/FTD-associated C9ORF72 repeat RNA promotes phase transitions in vitro and in cells". In: *Cell reports* 21.12, pp. 3573–3584.

Feric, Marina et al. (2016). "Coexisting liquid phases underlie nucleolar subcompartments". In: *Cell* 165.7, pp. 1686–1697.

Ferrandon, Dominique et al. (1997). "RNA–RNA interaction is required for the formation of specific bicoid mRNA 3 UTR–STAUFEN ribonucleoprotein particles". In: *The EMBO journal* 16.7, pp. 1751–1758.

Filipovska, Aleksandra et al. (2011). "A universal code for RNA recognition by PUF proteins". In: *Nature chemical biology* 7.7, pp. 425–427.

Fischl, Harry et al. (2019). "hnRNPC regulates cancer-specific alternative cleavage and polyadenylation profiles". In: *Nucleic acids research* 47.14, pp. 7580–7591.

Flory, Paul J (1941). "Thermodynamics of high polymer solutions". In: *The Journal of chemical physics* 9.8, pp. 660–660.

— (1953). *Principles of polymer chemistry*. Cornell University Press.

Fonseca, Bruno D et al. (2015). "La-related protein 1 (LARP1) represses terminal oligopyrimidine (TOP) mRNA translation downstream of mTOR complex 1 (mTORC1)". In: *Journal of Biological Chemistry* 290.26, pp. 15996–16020.

Forman-Kay, Julie D, Richard W Kriwacki, and Geraldine Seydoux (2018). "Phase separation in biology and disease". In: *Journal of molecular biology* 430.23, p. 4603.

Fox, Archa H, Yun Wah Lam, et al. (2002). "Paraspeckles: a novel nuclear domain". In: *Current Biology* 12.1, pp. 13–25.

Fox, Archa H and Angus I Lamond (2010). "Paraspeckles". In: *Cold Spring Harbor perspectives in biology* 2.7, a000687.

French, Juliet et al. (2002). "The expression of Ras–GTPase activating protein SH3 domain-binding proteins, G3BPs, in human breast cancers". In: *The Histochemical journal* 34.5, pp. 223–231.

Fromm, Simon A et al. (2012). "The structural basis of Edc3-and Scd6-mediated activation of the Dcp1: Dcp2 mRNA decapping complex". In: *The EMBO journal* 31.2, pp. 279–290.

Fu, Ye and Xiaowei Zhuang (2020). "m 6 A-binding YTHDF proteins promote stress granule formation". In: *Nature chemical biology* 16.9, pp. 955–963.

Gall, Joseph G (2003). "The centennial of the Cajal body". In: *Nature reviews Molecular cell biology* 4.12, pp. 975–980.

Grakoui, Arash et al. (1999). "The immunological synapse: a molecular machine controlling T cell activation". In: *Science* 285.5425, pp. 221–227.

Guillén-Boixet, Jordina et al. (2020). "RNA-induced conformational switching and clustering of G3BP drive stress granule assembly by condensation". In: *Cell* 181.2, pp. 346–361.

Gupta, Shobhit et al. (2007). "Quantifying similarity between motifs". In: *Genome biology* 8.2, pp. 1–9.

Haigis, Marcia C and Bruce A Yankner (2010). "The aging stress response". In: *Molecular cell* 40.2, pp. 333–344.

Han, Dali et al. (2019). "Anti-tumour immunity controlled through mRNA m 6 A methylation and YTHDF1 in dendritic cells". In: *Nature* 566.7743, pp. 270–274.

Han, Tina W et al. (2012). "Cell-free formation of RNA granules: bound RNAs identify features and components of cellular assemblies". In: *Cell* 149.4, pp. 768–779.

Heinrich, Bianca S et al. (2018). "Phase transitions drive the formation of vesicular stomatitis virus replication compartments". In: *MBio* 9.5, e02290–17.

Hernández, Gonzalo et al. (2018). "Decapping protein EDC4 regulates DNA repair and phenocopies BRCA1". In: *Nature communications* 9.1, pp. 1–11.

Hirose, Tetsuro et al. (2014). "NEAT1 long noncoding RNA regulates transcription via protein sequestration within subnuclear bodies". In: *Molecular biology of the cell* 25.1, pp. 169–183.

Hnisz, Denes et al. (2017). "A phase separation model for transcriptional control". In: *Cell* 169.1, pp. 13–23.

Hofmann, Sarah et al. (2021). "Molecular mechanisms of stress granule assembly and disassembly". In: *Biochimica et Biophysica Acta (BBA)-Molecular Cell Research* 1868.1, p. 118876.

Hollerer, Ina et al. (2016). "The differential expression of alternatively polyadenylated transcripts is a common stress-induced response mechanism that modulates mammalian mRNA expression in a quantitative and qualitative fashion". In: *RNA* 22.9, pp. 1441–1453.

Hopkins, Thomas G et al. (2016). "The RNA-binding protein LARP1 is a post-transcriptional regulator of survival and tumorigenesis in ovarian cancer". In: *Nucleic acids research* 44.3, pp. 1227–1246.

Hoque, Mainul, Zhe Ji, et al. (2013). "Analysis of alternative cleavage and polyadenylation by 3 region extraction and deep sequencing". In: *Nature methods* 10.2, pp. 133–139.

Hoque, Mainul, Wencheng Li, and Bin Tian (2014). "Accurate mapping of cleavage and polyadenylation sites by 3 region extraction and deep sequencing". In: *Polyadenylation*. Springer, pp. 119–129.

Hu, Yanhui et al. (2011). "An integrative approach to ortholog prediction for disease-focused and other functional studies". In: *BMC bioinformatics* 12.1, pp. 1–16.

Huang, Xiawei, Katalin Fejes Tóth, and Alexei A Aravin (2017). "piRNA biogenesis in Drosophila melanogaster". In: *Trends in Genetics* 33.11, pp. 882–894.

Hubstenberger, Arnaud et al. (2017). "P-body purification reveals the condensation of repressed mRNA regulons". In: *Molecular cell* 68.1, pp. 144–157.

Huggins, Maurice L (1941). "Solutions of long chain compounds". In: *The Journal of chemical physics* 9.5, pp. 440–440.

Hulsen, Tim (2021). *BioVenn: Create Area-Proportional Venn Diagrams from Biological Lists*. R Package Version 1.1.3. URL: `https://cran.r-project.org/package=BioVenn`.

— (2021). "BioVenn–an R and Python package for the comparison and visualization of biological lists using area-proportional Venn diagrams". In: *Data Science* 4.1, pp. 51–61. URL: `https://dx.doi.org/10.3233/DS-210032`.

Huppert, Julian L and Shankar Balasubramanian (2007). "G-quadruplexes in promoters throughout the human genome". In: *Nucleic acids research* 35.2, pp. 406–413.

Hyman, Anthony A, Christoph A Weber, and Frank Jülicher (2014). "Liquid-liquid phase separation in biology". In: *Annual review of cell and developmental biology* 30, pp. 39–58.

Iakoucheva, Lilia M et al. (2004). "The importance of intrinsic disorder for protein phosphorylation". In: *Nucleic acids research* 32.3, pp. 1037–1049.

Imperatore, Joshua A et al. (2020). "FUS recognizes G quadruplex structures within neuronal mRNAs". In: *Frontiers in molecular biosciences* 7, p. 6.

Ingelfinger, Dierk et al. (2002). "The human LSm1-7 proteins colocalize with the mRNA-degrading enzymes Dcp1/2 and Xrn1 in distinct cytoplasmic foci". In: *Rna* 8.12, pp. 1489–1501.

Iserman, Christiane et al. (2020). "Specific viral RNA drives the SARS CoV-2 nucleocapsid to phase separate". In: *BioRxiv*.

Jain, Ankur and Ronald D Vale (2017). "RNA phase transitions in repeat expansion disorders". In: *Nature* 546.7657, pp. 243–247.

Jain, Saumya et al. (2016). "ATPase-modulated stress granules contain a diverse proteome and substructure". In: *Cell* 164.3, pp. 487–498.

Jambor, Helena, Christine Brunel, and Anne Ephrussi (2011). "Dimerization of oskar 3 UTRs promotes hitchhiking for RNA localization in the Drosophila oocyte". In: *Rna* 17.12, pp. 2049–2057.

Jobe, Fatoumatta et al. (2020). "Respiratory syncytial virus sequesters NF-κB subunit p65 to cytoplasmic inclusion bodies to inhibit innate immune signaling". In: *Journal of virology* 94.22, e01380–20.

Kamath-Loeb, Ashwini, Lawrence A Loeb, and Michael Fry (2012). "The Werner syndrome protein is distinguished from the Bloom syndrome pro-

tein by its capacity to tightly bind diverse DNA structures". In: *PloS one* 7.1, e30189.

Kamenska, Anastasiia et al. (2014). "Human 4E-T represses translation of bound mRNAs and enhances microRNA-mediated silencing". In: *Nucleic acids research* 42.5, pp. 3298–3313.

Kassambara, Alboukadel (2020). *ggpubr: 'ggplot2' Based Publication Ready Plots*. R package version 0.4.0. URL: `https://CRAN.R-project.org/package=ggpubr`.

Kedersha, Nancy et al. (2016). "G3BP–Caprin1–USP10 complexes mediate stress granule condensation and associate with 40S subunits". In: *Journal of Cell Biology* 212.7.

Khong, Anthony et al. (2017). "The stress granule transcriptome reveals principles of mRNA accumulation in stress granules". In: *Molecular cell* 68.4, pp. 808–820.

Kikin, Oleg, Lawrence D'Antonio, and Paramjeet S Bagga (2006). "QGRS Mapper: a web-based server for predicting G-quadruplexes in nucleotide sequences". In: *Nucleic acids research* 34.suppl_2, W676–W682.

Kim, Hong Joo et al. (2013). "Mutations in prion-like domains in hnRNPA2B1 and hnRNPA1 cause multisystem proteinopathy and ALS". In: *Nature* 495.7442, pp. 467–473.

Kim, Woo Jae et al. (2005). "Sequestration of TRAF2 into stress granules interrupts tumor necrosis factor signaling under stress conditions". In: *Molecular and cellular biology* 25.6, pp. 2450–2462.

King, Oliver D, Aaron D Gitler, and James Shorter (2012). "The tip of the iceberg: RNA-binding proteins with prion-like domains in neurodegenerative disease". In: *Brain research* 1462, pp. 61–80.

Knott, Gavin J, Charles S Bond, and Archa H Fox (2016). "The DBHS proteins SFPQ, NONO and PSPC1: a multipurpose molecular scaffold". In: *Nucleic acids research* 44.9, pp. 3989–4004.

Kroschwald, Sonja, Shovamayee Maharana, et al. (2015). "Promiscuous interactions and protein disaggregases determine the material state of stress-inducible RNP granules". In: *elife* 4, e06807.

Kroschwald, Sonja, Matthias C Munder, et al. (2018). "Different material states of Pub1 condensates define distinct modes of stress adaptation and recovery". In: *Cell reports* 23.11, pp. 3327–3339.

Kuleshov, Maxim V et al. (2016). "Enrichr: a comprehensive gene set enrichment analysis web server 2016 update". In: *Nucleic acids research* 44.W1, W90–W97.

Kwiatkowski, T Jr et al. (2009). "Mutations in the FUS/TLS gene on chromosome 16 cause familial amyotrophic lateral sclerosis". In: *Science* 323.5918, pp. 1205–1208.

Lafontaine, Denis LJ et al. (2021). "The nucleolus as a multiphase liquid condensate". In: *Nature Reviews Molecular Cell Biology* 22.3, pp. 165–182.

Lander, Eric S et al. (2001). "Initial sequencing and analysis of the human genome". In:

Langdon, Erin M et al. (2018). "mRNA structure determines specificity of a polyQ-driven phase separation". In: *Science* 360.6391, pp. 922–927.

Lee, Chiu Fan et al. (2013). "Spatial organization of the cell cytoplasm by position-dependent phase separation". In: *Physical review letters* 111.8, p. 088101.

Lee, David SM, Louis R Ghanem, and Yoseph Barash (2020). "Integrative analysis reveals RNA G-quadruplexes in UTRs are selectively constrained and enriched for functional associations". In: *Nature communications* 11.1, pp. 1–12.

Lee, YouJin et al. (2018). "TIA1 variant drives myodegeneration in multisystem proteinopathy with SQSTM1 mutations". In: *The Journal of clinical investigation* 128.3, pp. 1164–1177.

Levone, Brunno R et al. (2021). "FUS-dependent liquid–liquid phase separation is important for DNA repair initiation". In: *Journal of Cell Biology* 220.5.

Li, Pilong et al. (2012). "Phase transitions in the assembly of multivalent signalling proteins". In: *Nature* 483.7389, pp. 336–340.

Li, Qiyao et al. (2017). "Global post-translational modification discovery". In: *Journal of proteome research* 16.4, pp. 1383–1390.

Li, Ying et al. (2012). "LSm14A is a processing body-associated sensor of viral nucleic acids that initiates cellular antiviral response in the early phase of viral infection". In: *Proceedings of the National Academy of Sciences* 109.29, pp. 11770–11775.

Lin, Yuan et al. (2015). "Formation and maturation of phase-separated liquid droplets by RNA-binding proteins". In: *Molecular cell* 60.2, pp. 208–219.

Liu, Xiaochuan et al. (2017). "Transcription elongation rate has a tissue-specific impact on alternative cleavage and polyadenylation in Drosophila melanogaster". In: *Rna* 23.12, pp. 1807–1816.

Liu, Yuting et al. (2013). "Cold-induced RNA-binding proteins regulate circadian gene expression by controlling alternative polyadenylation". In: *Scientific reports* 3.1, pp. 1–11.

Lorenz, Ronny et al. (2016). "Predicting RNA secondary structures from sequence and probing data". In: *Methods* 103, pp. 86–98.

Love, Michael I., Wolfgang Huber, and Simon Anders (2014). "Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2". In: *Genome Biology* 15 (12), p. 550. DOI: 10.1186/s13059-014-0550-8.

Lu, Tiemei, Karina K Nakashima, and Evan Spruijt (2021). "Temperature-Responsive Peptide–Nucleotide Coacervates". In: *The Journal of Physical Chemistry B* 125.12, pp. 3080–3091.

Lyon, Andrew S, William B Peeples, and Michael K Rosen (2021). "A framework for understanding the functions of biomolecular condensates across scales". In: *Nature Reviews Molecular Cell Biology* 22.3, pp. 215–235.

Machyna, Martin, Patricia Heyn, and Karla M Neugebauer (2013). "Cajal bodies: where form meets function". In: *Wiley Interdisciplinary Reviews: RNA* 4.1, pp. 17–34.

Mackenzie, Ian RA and Manuela Neumann (2012). "FET proteins in frontotemporal dementia and amyotrophic lateral sclerosis". In: *Brain research* 1462, pp. 40–43.

Mackenzie, Ian RA and Rosa Rademakers (2008). "The role of TDP-43 in amyotrophic lateral sclerosis and frontotemporal dementia". In: *Current opinion in neurology* 21.6, p. 693.

Maharana, Shovamayee et al. (2018). "RNA buffers the phase separation behavior of prion-like RNA binding proteins". In: *Science* 360.6391, pp. 918–921.

Mahboubi, Hicham and Ursula Stochaj (2017). "Cytoplasmic stress granules: Dynamic modulators of cell signaling and disease". In: *Biochimica et Biophysica Acta (BBA)-Molecular Basis of Disease* 1863.4, pp. 884–895.

Markmiller, Sebastian et al. (2018). "Context-dependent and disease-specific diversity in protein interactions within stress granules". In: *Cell* 172.3, pp. 590–604.

Matheny, Tyler, Bhalchandra S Rao, and Roy Parker (2019). "Transcriptome-wide comparison of stress granules and P-bodies reveals that translation plays a major role in RNA partitioning". In: *Molecular and Cellular Biology* 39.24, e00313–19.

Matheny, Tyler, Briana Van Treeck, et al. (2021). "RNA partitioning into stress granules is based on the summation of multiple interactions". In: *RNA* 27.2, pp. 174–189.

Matsuki, Hideaki et al. (2013). "Both G 3 BP 1 and G 3 BP 2 contribute to stress granule formation". In: *Genes to Cells* 18.2, pp. 135–146.

Mayr, Christine (2016). "Evolution and biological roles of alternative 3 UTRs". In: *Trends in cell biology* 26.3, pp. 227–237.

— (2019). "What are 3 UTRs doing?" In: *Cold Spring Harbor Perspectives in Biology* 11.10, a034728.

Mayr, Christine and David P Bartel (2009). "Widespread shortening of 3 UTRs by alternative cleavage and polyadenylation activates oncogenes in cancer cells". In: *Cell* 138.4, pp. 673–684.

McRae, Ewan KS et al. (2017). "Human DDX21 binds and unwinds RNA guanine quadruplexes". In: *Nucleic acids research* 45.11, pp. 6656–6668.

Michelitsch, Melissa D and Jonathan S Weissman (2000). "A census of glutamine/asparagine-rich regions: implications for their conserved function and the prediction of novel prions". In: *Proceedings of the National Academy of Sciences* 97.22, pp. 11910–11915.

Millevoi, Stefania, Hervé Moine, and Stéphan Vagner (2012). "G-quadruplexes in RNA biology". In: *Wiley Interdisciplinary Reviews: RNA* 3.4, pp. 495–507.

Mohanan, Neeraja K et al. (2021). "Alternative polyadenylation: An enigma of transcript length variation in health and disease". In: *Wiley Interdisciplinary Reviews: RNA*, e1692.

Monahan, Zachary et al. (2017). "Phosphorylation of the FUS low-complexity domain disrupts phase separation, aggregation, and toxicity". In: *The EMBO journal* 36.20, pp. 2951–2967.

Monette, Anne et al. (2020). "Pan-retroviral nucleocapsid-mediated phase separation regulates genomic RNA positioning and trafficking". In: *Cell reports* 31.3, p. 107520.

Morris, Adam R, Neelanjan Mukherjee, and Jack D Keene (2008). "Ribonomic analysis of human Pum1 reveals cis-trans conservation across

species despite evolution of diverse mRNA target sets". In: *Molecular and cellular biology* 28.12, pp. 4093–4103.

Namkoong, Sim et al. (2018). "Systematic characterization of stress-induced RNA granulation". In: *Molecular cell* 70.1, pp. 175–187.

Neumann, Manuela et al. (2009). "A new subtype of frontotemporal lobar degeneration with FUS pathology". In: *Brain* 132.11, pp. 2922–2931.

Neve, Jonathan, Kaspar Burger, et al. (2016). "Subcellular RNA profiling links splicing and nuclear DICER1 to alternative cleavage and polyadenylation". In: *Genome research* 26.1, pp. 24–35.

Neve, Jonathan and André Furger (2014). "Alternative polyadenylation: less than meets the eye?" In: *Biochemical Society Transactions* 42.4, pp. 1190–1195.

Neve, Jonathan, Radhika Patel, et al. (2017). "Cleavage and polyadenylation: Ending the message expands gene regulation". In: *RNA biology* 14.7, pp. 865–890.

Nott, Timothy J, Timothy D Craggs, and Andrew J Baldwin (2016). "Membraneless organelles can melt nucleic acid duplexes and act as biomolecular filters". In: *Nature chemistry* 8.6, pp. 569–575.

Nott, Timothy J, Evangelia Petsalaki, et al. (2015). "Phase transition of a disordered nuage protein generates environmentally responsive membraneless organelles". In: *Molecular cell* 57.5, pp. 936–947.

Nover, LUTZ, KLAUS-DIETER Scharf, and DIETER Neumann (1989). "Cytoplasmic heat shock granules are formed from precursor particles and are associated with a specific set of mRNAs". In: *Molecular and cellular biology* 9.3, pp. 1298–1308.

Overbeek, J Th G and MJ Voorn (1957). "Phase separation in polyelectrolyte solutions. Theory of complex coacervation". In: *Journal of Cellular and Comparative Physiology* 49.S1, pp. 7–26.

Pagès, Hervé (2018). *BSgenome: Software infrastructure for efficient representation of full genomes and their SNPs*. R package version 1.50.0.

Panas, Marc D, Pavel Ivanov, and Paul Anderson (2016). "Mechanistic insights into mammalian stress granule dynamics". In: *Journal of Cell Biology* 215.3, pp. 313–323.

Park, Ye-Jin et al. (2020). "Stress granule formation attenuates RACK1-mediated apoptotic cell death induced by morusin". In: *International journal of molecular sciences* 21.15, p. 5360.

Parker, Roy and Ujwal Sheth (2007). "P bodies and the control of mRNA translation and degradation". In: *Molecular cell* 25.5, pp. 635–646.

Patel, Avinash et al. (2015). "A liquid-to-solid phase transition of the ALS protein FUS accelerated by disease mutation". In: *Cell* 162.5, pp. 1066–1077.

Patel, Radhika et al. (2019). "Alternative cleavage and polyadenylation of genes associated with protein turnover and mitochondrial function are deregulated in Parkinson's, Alzheimer's and ALS disease". In: *BMC medical genomics* 12.1, p. 60.

Pederson, Thoru (2011). "The nucleolus". In: *Cold Spring Harbor perspectives in biology* 3.3, a000638.

Pitt, Jason N, Jennifer A Schisa, and James R Priess (2000). "P granules in the germ cells of Caenorhabditis elegans adults are associated with

clusters of nuclear pores and contain RNA". In: *Developmental biology* 219.2, pp. 315–333.

Poudyal, Raghav R et al. (2021). "RNA sequence and structure control assembly and function of RNA condensates". In: *RNA* 27.12, pp. 1589–1601.

Prasanth, Kannanganattu V et al. (2005). "Regulating gene expression through RNA nuclear retention". In: *Cell* 123.2, pp. 249–263.

Protter, David SW and Roy Parker (2016). "Principles and properties of stress granules". In: *Trends in cell biology* 26.9, pp. 668–679.

Ramaswami, Mani, J Paul Taylor, and Roy Parker (2013). "Altered ribostasis: RNA-protein granules in degenerative disorders". In: *Cell* 154.4, pp. 727–736.

Rao, Bhalchandra S and Roy Parker (2017). "Numerous interactions act redundantly to assemble a tunable size of P bodies in Saccharomyces cerevisiae". In: *Proceedings of the National Academy of Sciences* 114.45, E9569–E9578.

Rawat, Prashant et al. (2021). "Stress-induced nuclear condensation of NELF drives transcriptional downregulation". In: *Molecular cell* 81.5, pp. 1013–1026.

Ray, Debashish et al. (2013). "A compendium of RNA-binding motifs for decoding gene regulation". In: *Nature* 499.7457, pp. 172–177.

Raz, Erez (2000). "The function and regulation of vasa-like genes in germ-cell development". In: *Genome biology* 1.3, pp. 1–6.

Rehwinkel, JAN et al. (2005). "A crucial role for GW182 and the DCP1: DCP2 decapping complex in miRNA-mediated gene silencing". In: *Rna* 11.11, pp. 1640–1647.

Reijns, Martin AM et al. (2008). "A role for Q/N-rich aggregation-prone regions in P-body localization". In: *Journal of cell science* 121.15, pp. 2463–2472.

Renton, Alan E et al. (2011). "A hexanucleotide repeat expansion in C9ORF72 is the cause of chromosome 9p21-linked ALS-FTD". In: *Neuron* 72.2, pp. 257–268.

Reyes, Alejandro et al. (2013). "Drift and conservation of differential exon usage across tissues in primate species." In: *PNAS* 110, pp. -5. DOI: 10.1073/pnas.1307202110.

Rhine, Kevin, Velinda Vidaurre, and Sua Myong (2020). "RNA droplets". In: *Annual review of biophysics* 49, pp. 247–265.

Rhodes, Daniela and Hans J Lipps (2015). "G-quadruplexes and their regulatory roles in biology". In: *Nucleic acids research* 43.18, pp. 8627–8637.

Riback, Joshua A et al. (2017). "Stress-triggered phase separation is an adaptive, evolutionarily tuned response". In: *Cell* 168.6, pp. 1028–1040.

Ries, Ryan J et al. (2019). "m 6 A enhances the phase separation potential of mRNA". In: *Nature* 571.7765, pp. 424–428.

Rietveld, Krijn et al. (1982). "The tRNA-Uke structure at the 3 terminus of turnip yellow mosaic virus RNA. Differences and similarities with canonical tRNA". In: *Nucleic acids research* 10.6, pp. 1929–1946.

Robinson, James T et al. (2011). "Integrative genomics viewer". In: *Nature biotechnology* 29.1, pp. 24–26.

Roden, Christine and Amy S Gladfelter (2021). "RNA contributions to the form and function of biomolecular condensates". In: *Nature Reviews Molecular Cell Biology* 22.3, pp. 183–195.

Roost, Caroline et al. (2015). "Structure and thermodynamics of N6-methyladenosine in RNA: a spring-loaded base modification". In: *Journal of the American Chemical Society* 137.5, pp. 2107–2115.

Saha, Shambaditya et al. (2016). "Polar positioning of phase-separated liquid compartments in cells regulated by an mRNA competition mechanism". In: *Cell* 166.6, pp. 1572–1584.

Sama, Reddy Ranjith K et al. (2013). "FUS/TLS assembles into stress granules and is a prosurvival factor during hyperosmolar stress". In: *Journal of cellular physiology* 228.11, pp. 2222–2231.

Savas, Jeffrey N et al. (2008). "Huntington's disease protein contributes to RNA-mediated gene silencing through association with Argonaute and P bodies". In: *Proceedings of the National Academy of Sciences* 105.31, pp. 10820–10825.

Scheller, Nicoletta et al. (2007). "Identification of PatL1, a human homolog to yeast P body component Pat1". In: *Biochimica Et Biophysica Acta (BBA)-Molecular Cell Research* 1773.12, pp. 1786–1792.

Schwartz, Jacob C et al. (2013). "RNA seeds higher-order assembly of FUS protein". In: *Cell reports* 5.4, pp. 918–925.

Schwenzer, Hagen et al. (2021). "LARP1 isoform expression in human cancer cell lines". In: *RNA biology* 18.2, pp. 237–247.

Serman, Aurélia et al. (2007). "GW body disassembly triggered by siRNAs independently of their silencing activity". In: *Nucleic acids research* 35.14, pp. 4715–4727.

Shelkovnikova, Tatyana A et al. (2014). "Compromised paraspeckle formation as a pathogenic factor in FUSopathies". In: *Human molecular genetics* 23.9, pp. 2298–2312.

Sheth, Ujwal and Roy Parker (2003). "Decapping and decay of messenger RNA occur in cytoplasmic processing bodies". In: *Science* 300.5620, pp. 805–808.

Shi, Yongsheng (2012). "Alternative polyadenylation: new insights from global analyses". In: *Rna* 18.12, pp. 2105–2117.

Shin, Yongdae and Clifford P Brangwynne (2017). "Liquid phase condensation in cell physiology and disease". In: *Science* 357.6357.

Shurtleff, Matthew J et al. (2017). "Broad role for YBX1 in defining the small noncoding RNA composition of exosomes". In: *Proceedings of the National Academy of Sciences* 114.43, E8987–E8995.

Singleton, Rachelle S et al. (2014). "OGFOD1 catalyzes prolyl hydroxylation of RPS23 and is involved in translation control and stress granule formation". In: *Proceedings of the National Academy of Sciences* 111.11, pp. 4031–4036.

Smillie, David A and John Sommerville (2002). "RNA helicase p54 (DDX6) is a shuttling protein involved in nuclear assembly of stored mRNP particles". In: *Journal of cell science* 115.2, pp. 395–407.

Spegg, Vincent and Matthias Altmeyer (2021). "Biomolecular condensates at sites of DNA damage: more than just a phase". In: *DNA repair* 106, p. 103179.

Sponer, Jiri et al. (2018). "RNA structural dynamics as captured by molecular simulations: a comprehensive overview". In: *Chemical reviews* 118.8, pp. 4177–4338.

Strulson, Christopher A et al. (2012). "RNA catalysis through compartmentalization". In: *Nature chemistry* 4.11, pp. 941–946.

Su, Xiaolei, Jonathon A Ditlev, Enfu Hui, et al. (2016). "Phase separation of signaling molecules promotes T cell receptor signal transduction". In: *Science* 352.6285, pp. 595–599.

Su, Xiaolei, Jonathon A Ditlev, Michael K Rosen, et al. (2017). "Reconstitution of TCR signaling using supported lipid bilayers". In: *The Immune Synapse*. Springer, pp. 65–76.

Sun, Lijun et al. (2013). "Cyclic GMP-AMP synthase is a cytosolic DNA sensor that activates the type I interferon pathway". In: *Science* 339.6121, pp. 786–791.

Taiana, Elisa et al. (2020). "LncRNA NEAT1 in paraspeckles: a structural scaffold for cellular DNA damage response systems?" In: *Non-coding RNA* 6.3, p. 26.

Taniuchi, Shusuke et al. (2016). "Integrated stress response of vertebrates is regulated by four eIF2α kinases". In: *Scientific reports* 6.1, pp. 1–11.

Tassinari, Martina, Sara N Richter, and Paolo Gandellini (2021). "Biological relevance and therapeutic potential of G-quadruplex structures in

the human noncoding transcriptome". In: *Nucleic Acids Research* 49.7, pp. 3617–3633.

Tatomer, Deirdre C et al. (2016). "Concentrating pre-mRNA processing factors in the histone locus body facilitates efficient histone mRNA biogenesis". In: *Journal of Cell Biology* 213.5, pp. 557–570.

Tauber, Devin et al. (2020). "Modulation of RNA condensation by the DEAD-box protein eIF4A". In: *Cell* 180.3, pp. 411–426.

Taylor, Nicole et al. (2016). "Biophysical characterization of organelle-based RNA/protein liquid phases using microfluidics". In: *Soft matter* 12.45, pp. 9142–9150.

Teixeira, Daniela et al. (2005). "Processing bodies require RNA for assembly and contain nontranslating mRNAs". In: *Rna* 11.4, pp. 371–382.

Tian, Bin and James L Manley (2017). "Alternative polyadenylation of mRNA precursors". In: *Nature reviews Molecular cell biology* 18.1, pp. 18–30.

Tian, Siran, Harrison A Curnutte, and Tatjana Trcek (2020). "RNA granules: A view from the RNA perspective". In: *Molecules* 25.14, p. 3130.

Tourrière, Hélène et al. (2003). "The RasGAP-associated endoribonuclease G3BP assembles stress granules". In: *The Journal of cell biology* 160.6, pp. 823–831.

Trcek, Tatjana and Ruth Lehmann (2019). "Germ granules in Drosophila". In: *Traffic* 20.9, pp. 650–660.

Updike, Dustin L et al. (2011). "P granules extend the nuclear pore complex environment in the C. elegans germ line". In: *Journal of Cell Biology* 192.6, pp. 939–948.

Uversky, Vladimir N (2013). "The most important thing is the tail: multitudinous functionalities of intrinsically disordered protein termini". In: *FEBS letters* 587.13, pp. 1891–1901.

Van Treeck, Briana and Roy Parker (2018). "Emerging roles for intermolecular RNA-RNA interactions in RNP assemblies". In: *Cell* 174.4, pp. 791–802.

Van Treeck, Briana, David SW Protter, et al. (2018). "RNA self-assembly contributes to stress granule formation and defining the stress granule transcriptome". In: *Proceedings of the National Academy of Sciences* 115.11, pp. 2734–2739.

Vance, Caroline et al. (2009). "Mutations in FUS, an RNA processing protein, cause familial amyotrophic lateral sclerosis type 6". In: *Science* 323.5918, pp. 1208–1211.

Varadi, Mihaly et al. (2015). "Functional advantages of conserved intrinsic disorder in RNA-binding proteins". In: *PloS one* 10.10, e0139731.

Varshney, Dhaval et al. (2020). "The regulation and functions of DNA and RNA G-quadruplexes". In: *Nature Reviews Molecular Cell Biology* 21.8, pp. 459–474.

Vernon, Robert McCoy et al. (2018). "Pi-Pi contacts are an overlooked protein feature relevant to phase separation". In: *elife* 7, e31486.

Vessey, John P et al. (2006). "Dendritic localization of the translational repressor Pumilio 2 and its contribution to dendritic stress granules". In: *Journal of Neuroscience* 26.24, pp. 6496–6508.

Vilela, Cristina et al. (2000). "The eukaryotic mRNA decapping protein Dcp1 interacts physically and functionally with the eIF4F translation initiation complex". In: *The EMBO journal* 19.16, pp. 4372–4382.

Voronina, Ekaterina and Geraldine Seydoux (2010). "The C. elegans homolog of nucleoporin Nup98 is required for the integrity and function of germline P granules". In: *Development* 137.9, pp. 1441–1450.

Wang, Chen, Vladimir N Uversky, and Lukasz Kurgan (2016). "Disordered nucleiome: Abundance of intrinsic disorder in the DNA-and RNA-binding proteins in 1121 species from Eukaryota, Bacteria and Archaea". In: *Proteomics* 16.10, pp. 1486–1498.

Wang, Congwei et al. (2018). "Context-dependent deposition and regulation of mRNAs in P-bodies". In: *Elife* 7, e29815.

Wang, Eric T et al. (2008). "Alternative isoform regulation in human tissue transcriptomes". In: *Nature* 456.7221, pp. 470–476.

Wang, Jie et al. (2018). "A molecular grammar governing the driving forces for phase separation of prion-like RNA binding proteins". In: *Cell* 174.3, pp. 688–699.

Wang, Qi et al. (2019). "miR-223-3p/TIAL1 interaction is involved in the mechanisms associated with the neuroprotective effects of dexmedetomidine on hippocampal neuronal cells in vitro". In: *Molecular medicine reports* 19.2, pp. 805–812.

Wang, Ruijia, Ram Nambiar, et al. (2018). "PolyA_DB 3 catalogs cleavage and polyadenylation sites identified by deep sequencing in multiple genomes". In: *Nucleic acids research* 46.D1, pp. D315–D319.

Wang, Ruijia and Bin Tian (2020). "APAlyzer: a bioinformatics package for analysis of alternative polyadenylation isoforms". In: *Bioinformatics* 36 (12), pp. -2. DOI: 10.1093/bioinformatics/btaa266.

Wang, Xueyin, Jacob C Schwartz, and Thomas R Cech (2015). "Nucleic acid-binding specificity of human FUS protein". In: *Nucleic acids research* 43.15, pp. 7535–7543.

Waskom, Michael L. (2021). "seaborn: statistical data visualization". In: *Journal of Open Source Software* 6.60, p. 3021. DOI: 10.21105/joss.03021. URL: https://doi.org/10.21105/joss.03021.

Watanabe, H (1999). "Viscoelasticity and dynamics of entangled polymers". In: *Progress in Polymer Science* 24.9, pp. 1253–1403.

Wehner, Karen A, Sylvia Schütz, and Peter Sarnow (2010). "OGFOD1, a novel modulator of eukaryotic translation initiation factor 2α phosphorylation and the cellular response to stress". In: *Molecular and cellular biology* 30.8, pp. 2006–2016.

Wei, Wenqiang et al. (2017). "Regulatory Role of N6-methyladenosine (m6A) methylation in RNA processing and human diseases". In: *Journal of cellular biochemistry* 118.9, pp. 2534–2543.

Weidtkamp-Peters, Stefanie et al. (2008). "Dynamics of component exchange at PML nuclear bodies". In: *Journal of cell science* 121.16, pp. 2731–2743.

Wheeler, Joshua R et al. (2016). "Distinct stages in stress granule assembly and disassembly". In: *Elife* 5, e18413.

White, James P and Richard E Lloyd (2012). "Regulation of stress granules in virus systems". In: *Trends in microbiology* 20.4, pp. 175–183.

Whyte, Warren A et al. (2013). "Master transcription factors and mediator establish super-enhancers at key cell identity genes". In: *Cell* 153.2, pp. 307–319.

Wickham, Hadley (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. ISBN: 978-3-319-24277-4. URL: `https://ggplot2.tidyverse.org`.

Woodruff, Jeffrey B et al. (2017). "The centrosome is a selective condensate that nucleates microtubules by concentrating tubulin". In: *Cell* 169.6, pp. 1066–1077.

Wright, Peter E and H Jane Dyson (2015). "Intrinsically disordered proteins in cellular signalling and regulation". In: *Nature reviews Molecular cell biology* 16.1, pp. 18–29.

Wu, Guanhui et al. (2019). "DDX5 helicase resolves G-quadruplex and is involved in MYC gene transcriptional activation". In: *Proceedings of the National Academy of Sciences* 116.41, pp. 20453–20461.

Wu, Jiaxi et al. (2013). "Cyclic GMP-AMP is an endogenous second messenger in innate immune signaling by cytosolic DNA". In: *Science* 339.6121, pp. 826–830.

Xiang, Siheng et al. (2015). "The LC domain of hnRNPA2 adopts similar conformations in hydrogel polymers, liquid-like droplets, and nuclei". In: *Cell* 163.4, pp. 829–839.

Xiao, Qian, Ceara K McAtee, and Xiaolei Su (2021). "Phase separation in immune signalling". In: *Nature Reviews Immunology*, pp. 1–12.

Xie, Wei et al. (2019). "Human cGAS catalytic domain has an additional DNA-binding interface that enhances enzymatic activity and liquid-phase

condensation". In: *Proceedings of the National Academy of Sciences* 116.24, pp. 11946–11955.

Xie, Zhuorui et al. (2021). "Gene set knowledge discovery with Enrichr". In: *Current protocols* 1.3, e90.

Xu, Shujuan et al. (2016). "Thioflavin T as an efficient fluorescence sensor for selective recognition of RNA G-quadruplexes". In: *Scientific Reports* 6.1, pp. 1–9.

Yamazaki, Tomohiro et al. (2018). "Functional domains of NEAT1 architectural lncRNA induce paraspeckle assembly through phase separation". In: *Molecular cell* 70.6, pp. 1038–1053.

Yang, Peiguo et al. (2020). "G3BP1 is a tunable switch that triggers phase separation to assemble stress granules". In: *Cell* 181.2, pp. 325–345.

Ying, Yi et al. (2017). "Splicing activation by Rbfox requires self-aggregation through its tyrosine-rich domain". In: *Cell* 170.2, pp. 312–323.

Yoon, YoneJung et al. (2013). "Anxiety-associated alternative polyadenylation of the serotonin transporter mRNA confers translational regulation by hnRNPK". In: *Proceedings of the National Academy of Sciences* 110.28, pp. 11624–11629.

Youn, Ji-Young et al. (2018). "High-density proximity mapping reveals the subcellular organization of mRNA-associated granules and bodies". In: *Molecular cell* 69.3, pp. 517–532.

Zeigler, Travis M et al. (2021). "Protein phase separation: physical models and phase-separation-mediated cancer signaling". In: *Advances in Physics: X* 6.1, p. 1936638.

Zhang, Haibo, Ju Youn Lee, and Bin Tian (2005). "Biased alternative polyadenylation in human tissues". In: *Genome biology* 6.12, pp. 1–13.

Zhang, Huaiying et al. (2015). "RNA controls PolyQ protein phase transitions". In: *Molecular cell* 60.2, pp. 220–230.

Zhang, Xu et al. (2014). "The cytosolic DNA sensor cGAS forms an oligomeric complex with DNA and undergoes switch-like conformational changes in the activation loop". In: *Cell reports* 6.3, pp. 421–430.

Zhang, Yueying et al. (2019). "G-quadruplex structures trigger RNA phase separation". In: *Nucleic acids research* 47.22, pp. 11746–11754.

Zhang, Yun et al. (2021). "Rapid and accurate alignment of nucleotide conversion sequencing reads with HISAT-3N". In: *Genome Research*, gr–275193.

Zheng, Dinghai et al. (2018). "Cellular stress alters 3 UTR landscape through alternative polyadenylation and isoform-specific degradation". In: *Nature communications* 9.1, pp. 1–14.

Zhong, Sue et al. (2000). "Role of SUMO-1–modified PML in nuclear body formation". In: *Blood, The Journal of the American Society of Hematology* 95.9, pp. 2748–2752.

Zhou, Dujin and Shiuan Chen (2001). "PNRC2 is a 16 kDa coactivator that interacts with nuclear receptors through an SH3-binding motif". In: *Nucleic acids research* 29.19, pp. 3939–3948.

Zhou, Yueqin et al. (2013). "ALS-associated FUS mutations result in compromised FUS alternative splicing and autoregulation". In: *PLoS genetics* 9.10, e1003895.

Appendix A

Appendix

A.0.1 DDX4N protein gel

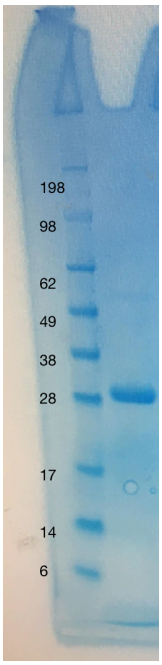


Figure A.1: DDX4N final protein gel. A molecular weight ladder is displayed on the left hand side.

A.0.2 Correlation analysis of GQFRs vs transcript length and GC % in the human transcriptome

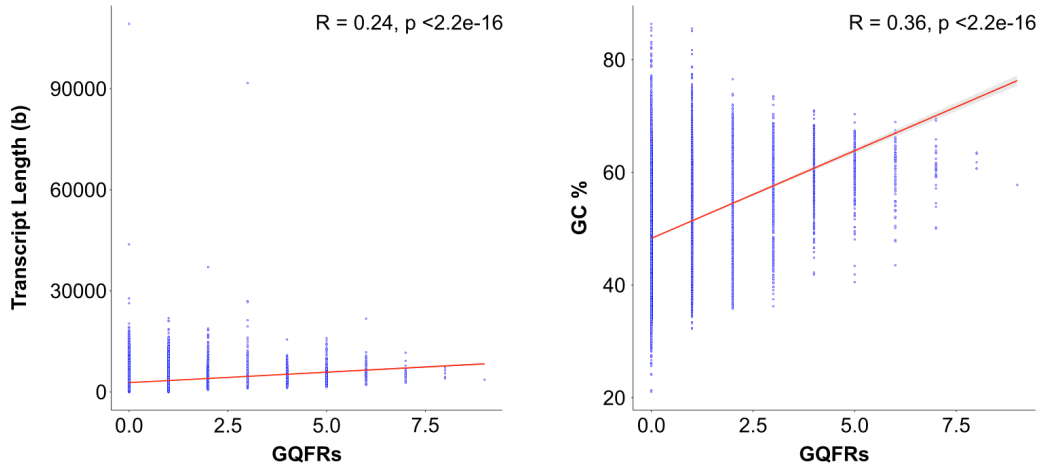
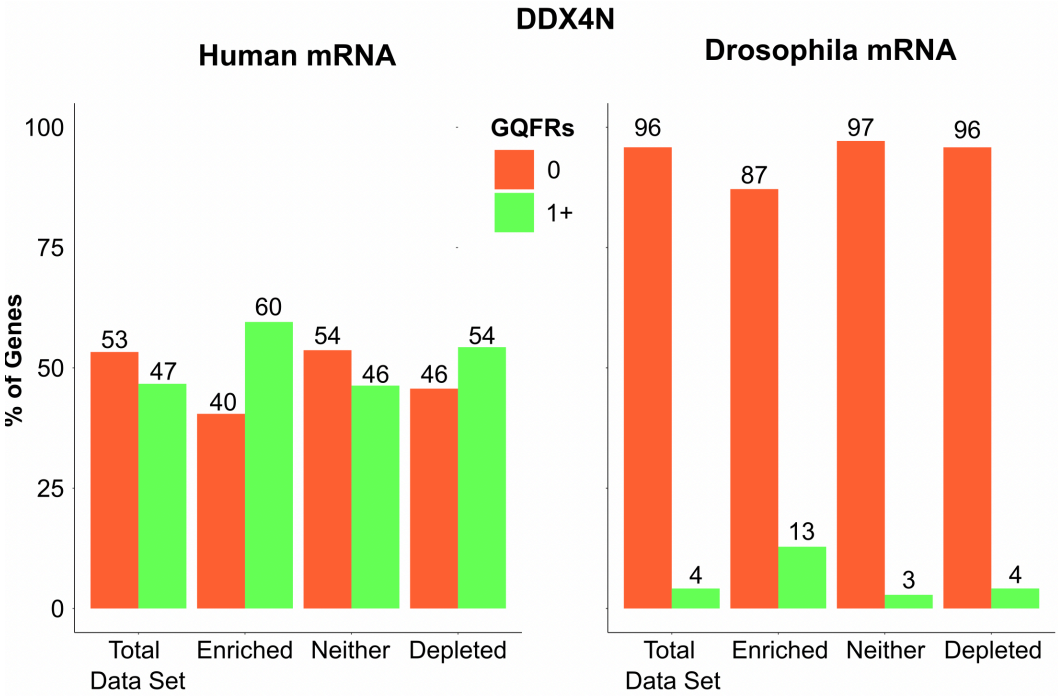


Figure A.2: Correlation analysis of GQFRs vs transcript length and GC % in the human transcriptome. Pearson correlation R values are displayed with each plot.

A.0.3 Bar plots displaying percentage of genes with 0 or 1 + GQFRs in each mRNA partitioning cohort, as well as the whole data set, for human and *Drosophila* data sets in DDX4N



Bar plots displaying percentage of genes with 0 or 1 + GQFRs in each mRNA partitioning cohort, as well as the whole data set, for human and *Drosophila* data sets in DDX4N. Percentages for each bar are displayed above.

A.0.4 Analysis of the physical properties of human and *Drosophila* orthologs with and without the same partitioning in DDX4N

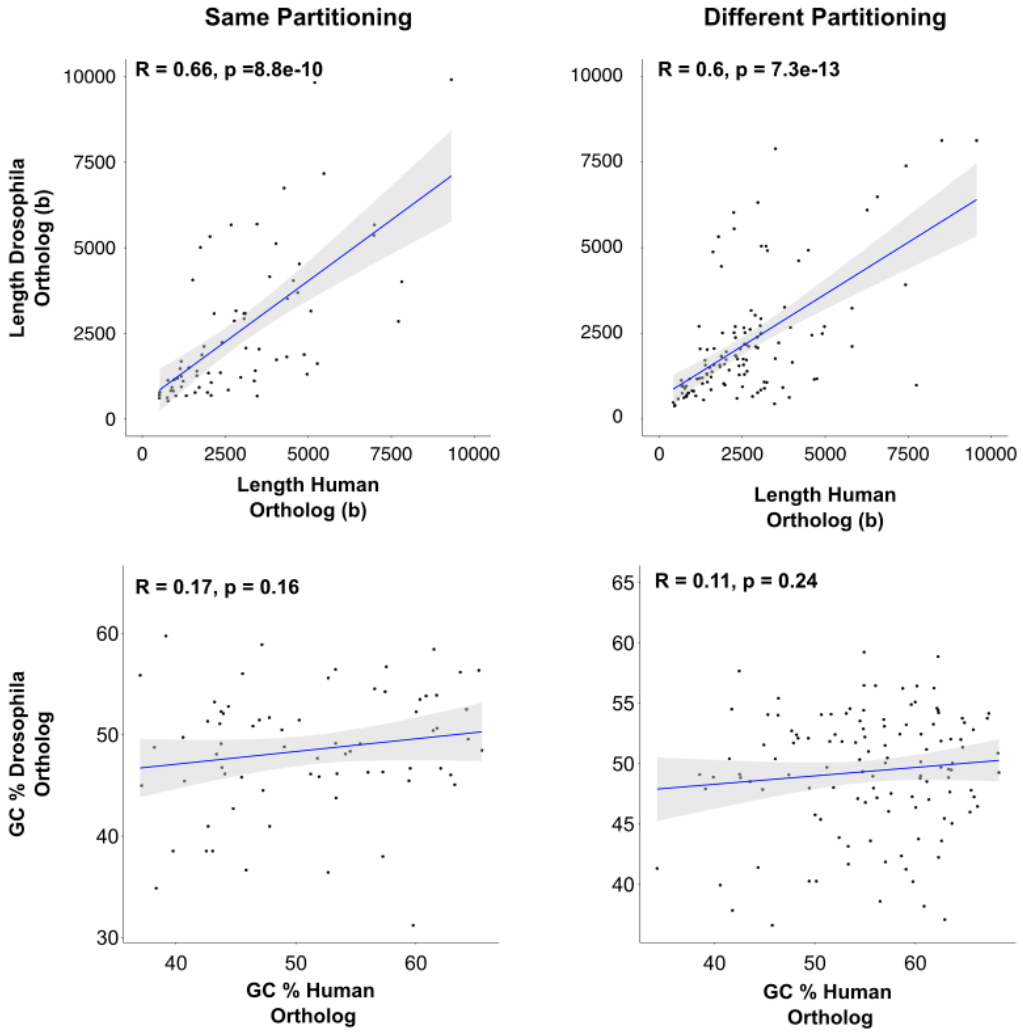


Figure A.4: Pearson correlation analysis of the physical properties of human and *Drosophila* orthologs with and without the same partitioning in DDX4N. The first row displays scatter plots of correlating the length of human and *Drosophila* orthologs with and without the same partitioning. The second row displays scatter plots of correlating the GC % of human and *Drosophila* orthologs with and without the same partitioning

A.0.5 EWSR1C protein gel

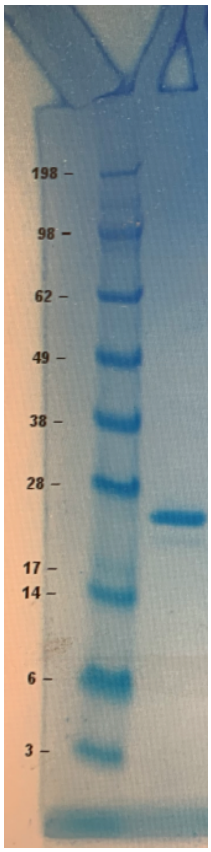


Figure A.5: EWSR1C final protein gel. A molecular weight ladder is displayed on the left hand side.

A.0.6 Comparison between results obtained from DE-Seq2 analysis using either the dilute phase or SW480 BAM files as the reference for DDX4N and FUS-GFP

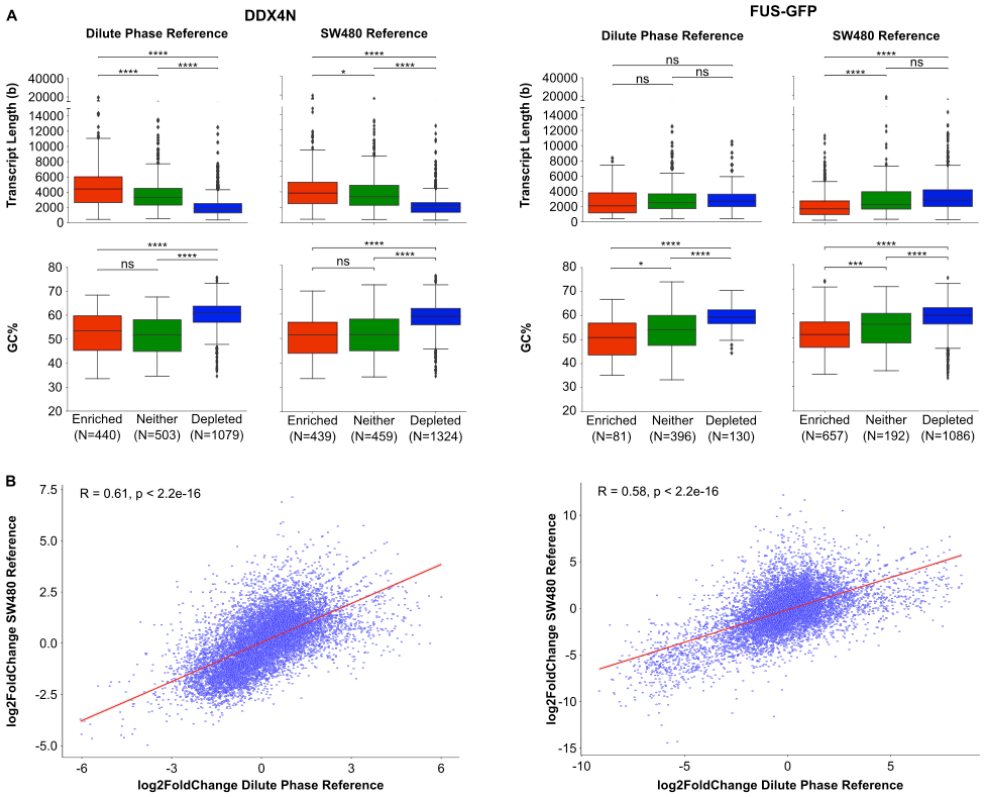


Figure A.6 *(previous page)*: Comparison between results obtained from DESeq2 analysis using either the dilute phase or SW480 BAM files as the reference for DDX4N and FUS-GFP. A) Boxplots depicting the transcript length and GC % for the mRNA the partitioning cohorts in DDX4N and FUS-GFP for each reference BAM are shown. B) Correlation analysis of results obtained from from DESeq2 analysis using either the dilute phase or SW480 BAM files as the reference for DDX4N and FUS-GFP. The R value for each Pearson correlation is displayed on each plot. Significance values displayed above A) were acquired using a T-test; ns = p > 0.05, * = p <= 0.05, ** = p <= 0.01, *** = p <= 0.001, **** = p <= 0.0001.

A.0.7 Preliminary characterisation of EWSR1C

DIC microscopy analysis of EWSR1C, with cytoplasmic RNA, DDX4N and both together

In order to further characterise the phase-separating properties of EWSR1C we performed a hanging drop assay in combination with DIC microscopy. We used different concentrations of EWSR1C and added various concentrations of cytoplasmic RNA, DDX4N or both together to see how whether the system would phase separate and what the differences between each experiment would be. We anticipated that EWSR1C would also phase separate in the presence of DDX4N since DDX4N is negatively charged and the two proteins would form complex coacervates.

A
EWSR1C uM
45
22.5
0 h
0.213 ug / uL RNA
7 h
10 μm
B
DDX4N:EWSR1C
Concentration uM
45:45
22.5:22.5
6 h
C
45:22.5
5 h
100 μm

Figure A.7 *(previous page)*: DIC microscopy analysis of EWSR1C phase-separation assays. Two time points were used for each assay to see how the systems developed. These are displayed above each image. The concentrations for EWSR1C, cytoplasmic RNA and DDX4N are indicated where applicable. Each row contains a results from a different assay. A) EWSR1C and cytoplasmic RNA assay. B) EWSR1C and DDX4N assay. C) EWSR1C, DDX4N and cytoplasmic RNA assay.

The assay we performed, using just EWSR1C and cytoplasmic RNA, showed that at time 0 h, there was a dual-phase system. However, we found that over time, this transitioned to aggregates. The morphology of the aggregates seemed to be dependent on the EWSR1C:cytoplasmic RNA ratio, with a lower ratio leading to aggregates with increased numbers of branches (Figure A.7A). The assay performed with EWSR1C and DDX4N showed that DDX4N also promoted phase separation of EWSR1C. At time point 0 h there was what appeared to be a dual-phase system. Over time, unlike in the presence of cytoplasmic RNA we observed no aggregation. However, what we observed what appeared to be three separate phases (Figure A.7B). More work will need to be done to confirm whether these phases constitute the solvent, and the two proteins individually, and indeed which constituents are responsible for which. This is a phenomenon which has been observed in the nucleolus which is a multiphase liquid condensate and therefore, it is a rational observation that this might be what we are observing here (Lafontaine et al. 2021). Finally our assay where we combined EWSR1C, DDX4N and cytoplasmic RNA, again showed a dual-phase solution at time point 0 h. Over time, we observed what appeared to be a state that exists between those of the previous two systems (Figure A.7C). There was clearly some

aggregation but the patterning largely resembled that of the EWSR1C and DDX4N system at later time points. This suggests that it is the RNA that is driving the transitions from liquid to solid. Again more work will need to be done to verify and elucidate these results.

Thioflavin T (ThT) amyloid assay

Since we observed transition from liquid to solid when cytoplasmic RNA was added to EWSR1C, we examined whether this was due to amyloid formation. ThT is a chemical which omits fluorescence when it binds amyloids and we used this in order to investigate whether or not EWSR1C was forming amyloids with cytoplasmic RNA.

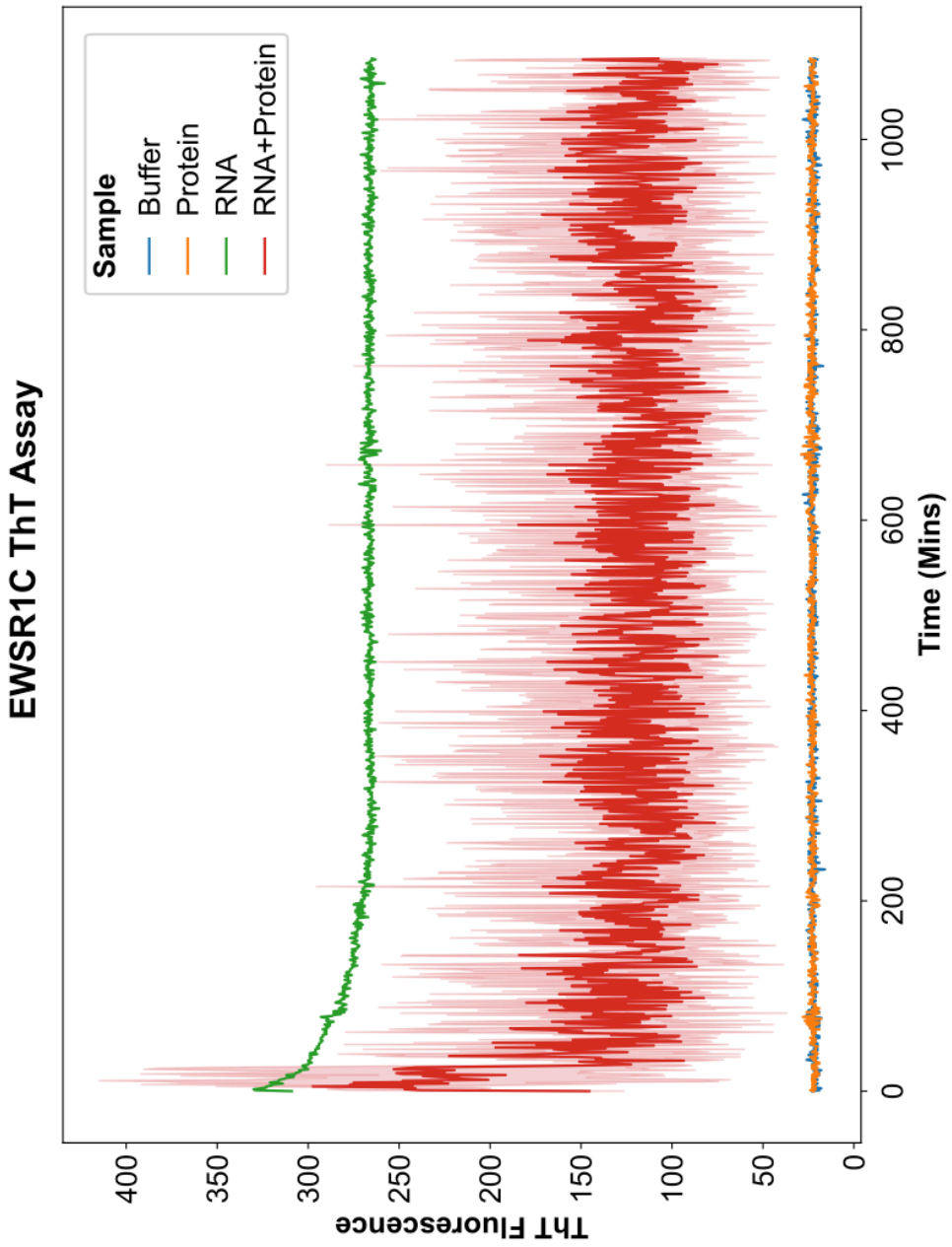

EWSR1C ThT Assay
Sample
Buffer
Protein
RNA
RNA+Protein
ThT Fluorescence
0
50
100
150
200
250
300
350
400
Time (Mins)
0
200
400
600
800
1000

Figure A.8 *(previous page)*: Amyloid detection assay with ThT using EWSR1C and cytoplasmic RNA. The y axis displays the fluorescence while the x axis displays the time. Four samples were run in duplicate with three controls (buffer only, protein only and RNA only) and one experiment (RNA + protein). The error is shown in a lighter colour for each line.

The assay revealed some surprising results. We saw that in the absence of RNA our controls of buffer plus ThT and protein plus ThT omitted only low levels of background fluorescence indicative of no ThT binding. However, surprisingly, we found that in the RNA only sample (green), that there appeared to be ThT binding as there was a higher level of fluorescence, and this occurred at time point 0. We then found that EWSR1C plus cytoplasmic RNA assay omitted fluorescence that fluctuated between the RNA control and the protein only and buffer controls. We found a study that suggests that ThT has the capacity to bind to RNA G-quadruplexes and reasoned that this might be the reason that we saw increased fluorescence in both the samples that contained cytoplasmic RNA (Xu et al. 2016). The reason that we might see a difference between the two samples could be that the protein is interfering with ThTs ability to bind these and hence why we saw fluctuations in the protein plus cytoplasmic RNA sample. Overall,the results suggest that it is unlikely that EWSR1C plus cytoplasmic RNA lead to the formation of amyloids.

A.0.8 Bar plots displaying percentage of genes with 0 or 1 + GQFRs in each mRNA partitioning cohort, as well as the whole data set, for, FUS-GFP, P-Bodies and Stress Granules

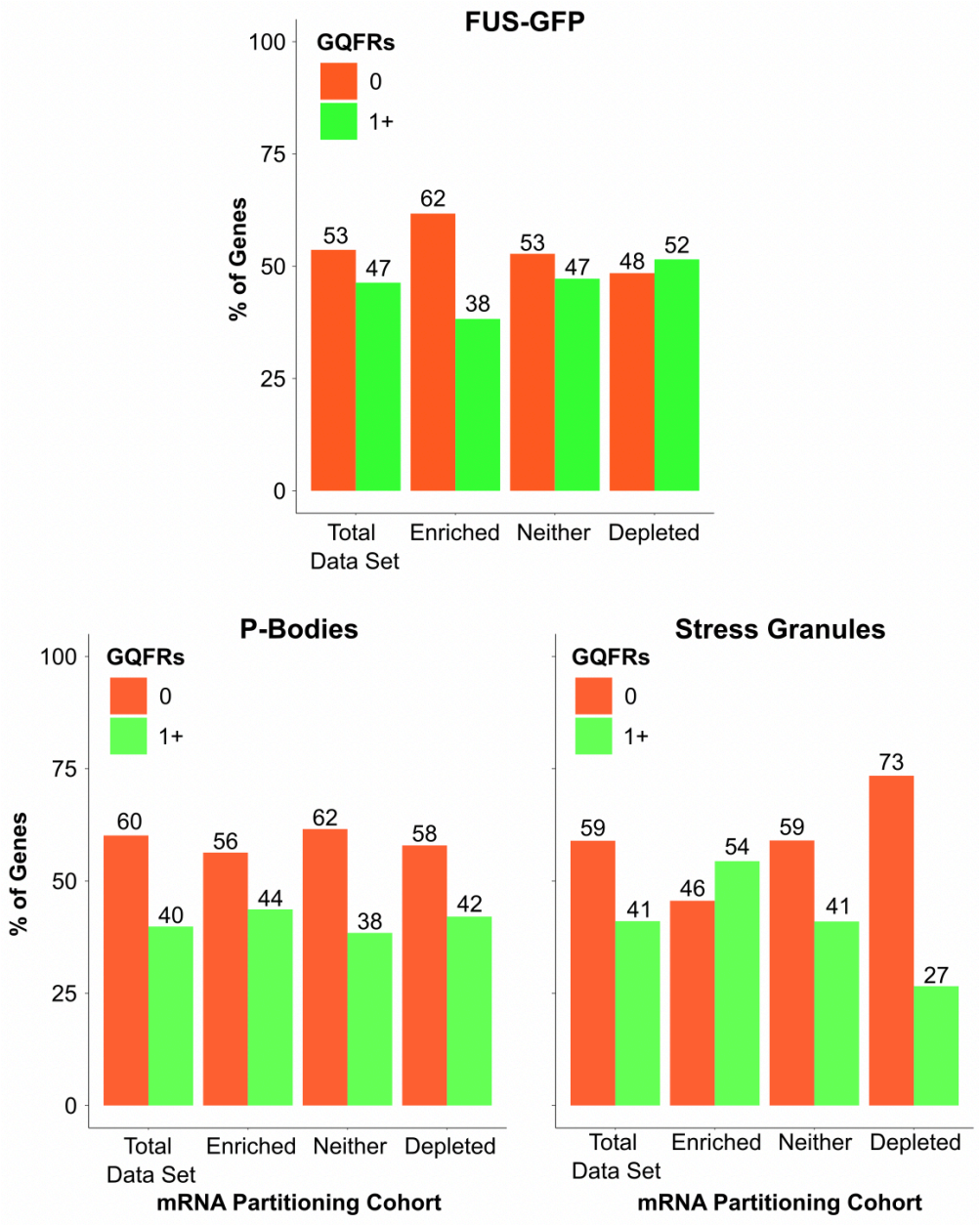
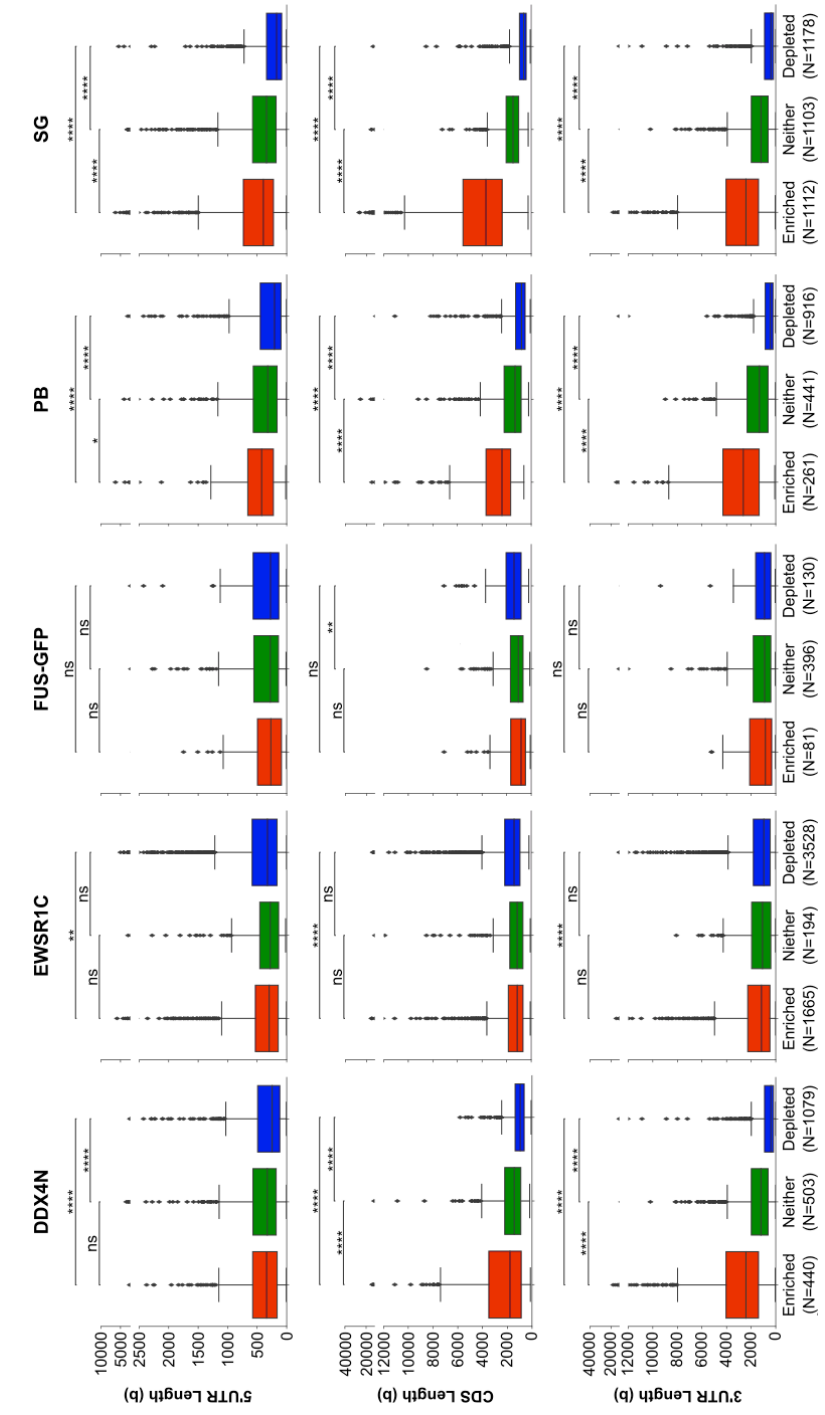
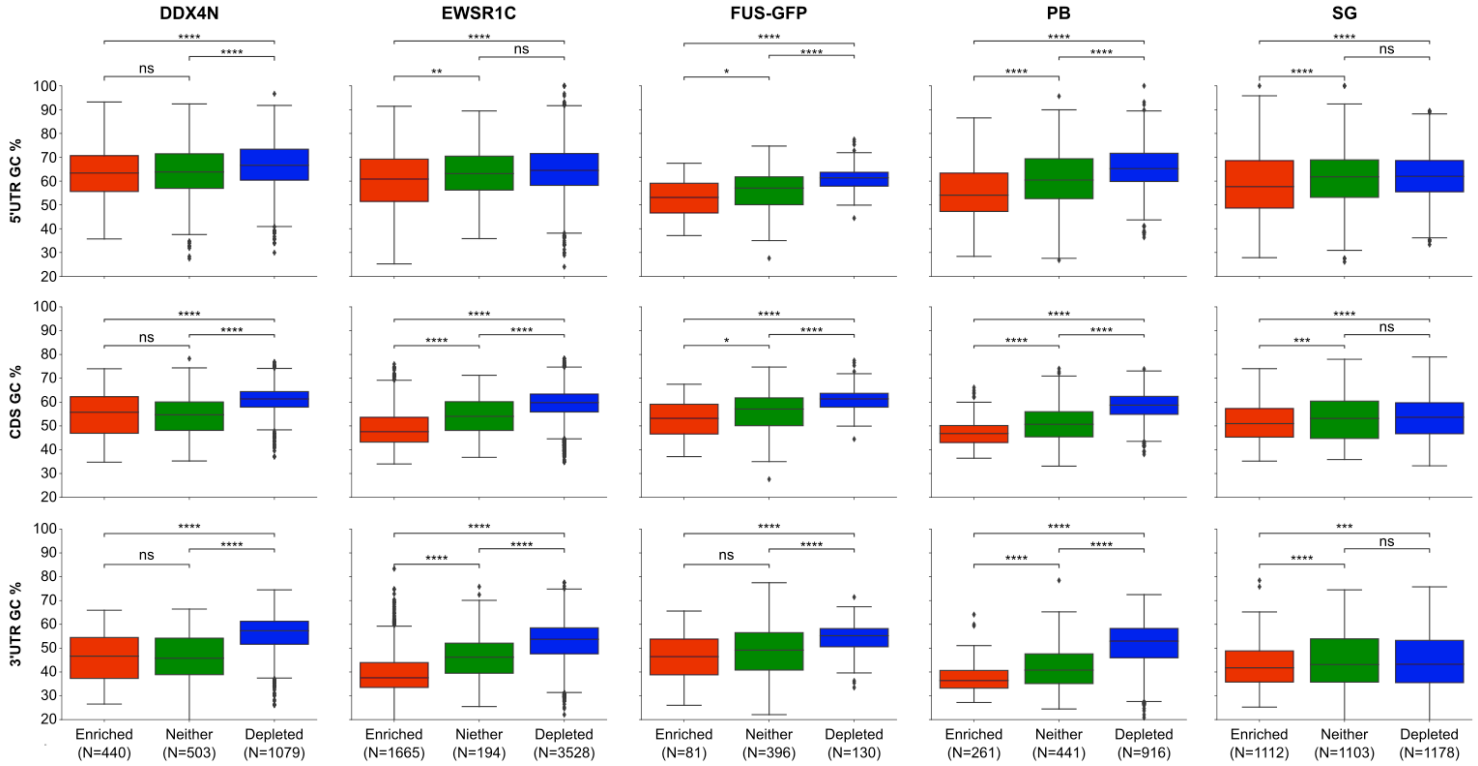


Figure A.9 *(previous page)*: Bar plots displaying percentage of genes with 1 or more GQFR in each mRNA partitioning cohort, as well as the whole data set, for FUS-GFP, P-Bodies and Stress Granules. Percentages for each bar are displayed above.

A.0.9 Analysis of the length and GC % of the 5'UTR, CDS, and 3'UTR in mRNA partitioning cohorts in all investigated BMCs



Boxplots depicting the relationship between the length of the 5'UTR, CDS and 3'UTR in mRNA partitioning cohorts in DDX4N, EWSR1C, FUS-GFP, PBs and SGs. Significance values displayed above were acquired using a T-test; ns = p > 0.05, * = p <= 0.05, ** = p <= 0.01, *** = p <= 0.001, **** = p <= 0.0001.



Boxplots depicting the relationship between the GC % of the 5'UTR, CDS and 3'UTR in mRNA partitioning cohorts in DDX4N, EWSR1C, FUS-GFP, PBs and SGs. Significance values displayed above were acquired using a T-test; ns = p > 0.05, * = p <= 0.05, ** = p <= 0.01, *** = p <= 0.001, **** = p <= 0.0001.

A.0.10 The length dependency of selective partitioning of human mRNA in DDX4N as a function of its GC %

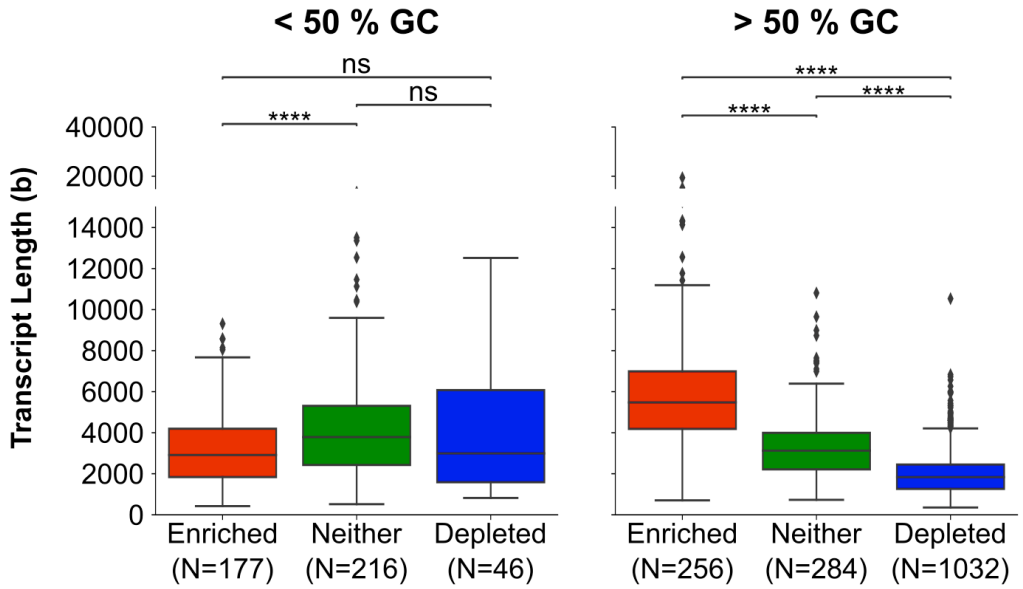


Figure A.12: Boxplots depicting the transcript length for each of the partitioning cohorts in the human data set, split by GC %. Significance values displayed above were acquired using a T-test; ns = p > 0.05, * = p <= 0.05, ** = p <= 0.01, *** = p <= 0.001, **** = p <= 0.0001.

A.0.11 Example IGV and corresponding expression plots depicting two genes with long isoform enriched 3'UTR-APA changes in DDX4N from the human data set

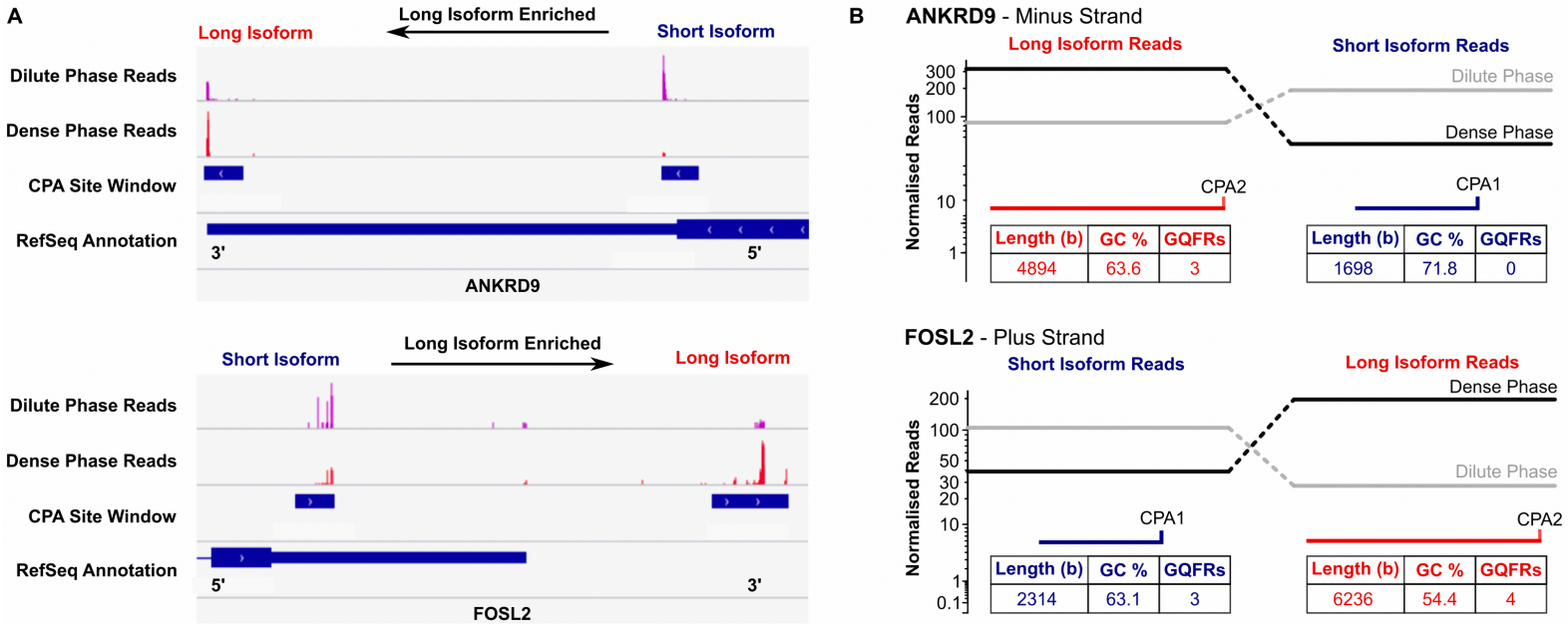


Figure A.13: IGV and corresponding expression plots depicting two genes with long isoform enriched 3'UTR-APA changes in DDX4N from the human data set. A) IGV screenshots showing lineplots for the normalised reads for dilute (purple) and dense (red) phases with each repeat overlayed. The CPA site windows within which the reads are counted are displayed as well as the RefSeq annotation for each gene. B) Normalised read expression graphs depicting the reads for each isoform in both the dense and dilute phases. The tables displays the length, GC % and number of GQFRs for each corresponding isoform.