TESI DOCTORAL UPF/ ANY 2019

COMPUTATIONAL APPROACHES TO CHARACTERIZE RNP GRANULES

Fernando Cid Samper

Directors
Dr. Gian Gaetano Tartaglia
Dr. Natalia Sánchez de Groot

Gene Function and Evolution

Bioinformatics and Genomics Department

Centre for Genomic Regulation (CRG)

Agradecimientos

Siempre se ha presupuesto al científico como una persona marcadamente racional, pero en mi caso ha sido más bien la intuición la que ha guiado hasta ahora el desarrollo de mi vida profesional. No fue una decisión racional sino una decisión de última hora la que me llevó con 17 años a estudiar Biotecnología en vez de Medicina. Tampoco entendía bien qué era aquello de la bioinformática cuando me decidí a hacer un máster en el Reino Unido; y, siendo sinceros, ni siquiera tenía una idea clara de qué tipo de investigación se llevaba a cabo en el CRG cuando lo descubrí buscando por internet un sitio donde hacer el doctorado en Barcelona. Simplemente mi intuición me decía que era el lugar correcto.

A veces guiarte tanto por la intuición puede jugar malas pasadas. Aún recuerdo la cara que puso el prometedor joven jefe de grupo Manu Irimia cuando le pregunté durante el proceso de selección en qué grupo quería él hacer el doctorado. Sin embargo, fue esa misma intuición la que me llevó a pedir una entrevista extra con un tipo italiano llamado Gian Tartaglia, al que tampoco conocía de nada antes de llegar al proceso de selección. Aquella sin duda fue una de las mejores decisiones de mi vida.

Cuatro años después me encuentro presentando una tesis de la que me siento orgulloso y que me ha llevado a experimentar una de las etapas más felices de mi vida. Mi aventura en la ciencia académica seguramente termine aquí, pero me deja la satisfacción de haber aportado mi granito de arena en responder algunas preguntas que no tenían respuesta hace cuatro años. He podido así responder por fin a aquel niño que descubrió el placer de la biología con los VHS de "Érase una vez la vida" y que desde entonces no ha podido parar de preguntarse sobre el mundo que le rodeaba.

Por todo ello, quiero en primer lugar agradecer a Gian el haberme dado esta oportunidad, con billete de ida y vuelta a Japón incluido. Porque él siempre me ha entendido mejor de lo que yo le he entendido a él. Por su infinita paciencia, su profunda humanidad y su continuo sacrificio para crear un laboratorio tan único, estimulante, divertido y cálido que estoy seguro que echaré de menos durante el resto de mi vida.

También debo mucho de mi desarrollo profesional a mis otros mentores en él, a quienes también quiero agradecer profundamente su esfuerzo realizado conmigo. A Davide, por acompañarme en mis primeros pasos en el CRG y ser un modelo de lo que debería ser un doctorando en bioinformática. A Benni y a Natalia por descubrirme el interesantísimo mundo de los gránulos de estrés y por haberme ayudado a salir del laberinto que siempre supone

intentar terminar una nueva publicación. A Aki, por haber sido la mejor anfitriona que un occidental puede esperar en Japón.

Hay tantas cosas que me gustaría agradecerle a Alex, que resultaría complicado enumerarlas todas sin verme obligado a pedir de nuevo un diseño de la portada a CAU La Factoría por exceso en el número de páginas de la tesis. Recuerdo cómo cuando comenzamos me pedías que te explicara los aspectos más básicos de la biología y cómo me has acabado superando en todo. Sigo aprendiendo mucho de ti, sobre todo de tu capacidad para apreciar las cosas importantes de la vida, ésas que nos hacen realmente humanos. Gran parte del genial ambiente del labo es cosa tuya y de tus insultos en griego como música de ambiente en el trabajo. Sería genial que algún día podamos volver a trabajar juntos.

Desde luego voy a echar mucho de menos este genial a ambiente de trabajo. A Nieves, por su comprensión escuchando problemas y su capacidad multitasking. A Riccardo, por sus recomendaciones sobre videojuegos. A Mimma, por su guía para abrir nuevos caminos. A Dasti, por sus inacabables preguntas y sus controvertidas respuestas. A Andrea, por haberme ofrecido la oportunidad de devolverle a alguien el esfuerzo que Davide un día hizo conmigo. A Maria Carla, por recomendarme al mejor fisio del mundo y por contener a Dasti para que nos dejetrabajarporfavor. A Magda, por los divertidos momentos que crea cada vez que aparece por el labo. A Elías, por las inagotables fiestas en el Mata Hari. A Michelle, por su generosidad con el template de latex que nunca llegué a utilizar. Y a todos aquellos que han pasado alguna vez por Tartaglia Lab, por haber creado un lugar único al que acudir cada mañana (eso sí, a partir de las 11).

Además de la gente del laboratorio, también me siento muy afortunado porque durante estos años he iniciado un montón de nuevas amistades que se encuentran entre las mejores que he tenido nunca en la vida. Ha sido genial el poder tener estimulantes charlas y un sinfín de divertidas anécdotas con el grupo de las “Aftework Beers” (Pablo B, Antonio, Cate, Júlia, Paul y Claudia). Siempre recordaré los retreats, las cenas en el Bacoa y los viajes a Lisboa y a Israel, de los mejores que he hecho jamás. Aunque no pertenezca a la estirpe del siaryi, tampoco me quiero olvidar de Pablo H. y de sus siempre interesantes recomendaciones culturales. También recordaré con un cariño especial, el extraño núcleo familiar que constituimos en los pisos de la calle Aragó.

Pero sin duda, la persona más importante que he conocido en este periodo es Sònia. Admiro de ella su enorme empatía y capacidad para comprender a las personas, su cuidado por los pequeños detalles de la vida, su amor hacia el arte, su pasión viajera y aventura, su mente abierta y obstinada al mismo tiempo y su enorme capacidad de sacrificio y profesionalidad. Ella me dio

el consejo más importante para que esta tesis haya sido posible: “Empieza a escribir la tesis en Japón”. T´estimo, amor meu.

Por último, quiero agradecerles profundamente este tesis a toda mi familia, en especial a mi hermana y a mis padres. A mi hermana, porque aunque seamos muy diferentes en casi todo, somos muy parecidos en lo más importante. A mis padres, por todo el esfuerzo que han hecho y siguen haciendo cada día por mí. Porque, aunque no siempre me entiendan, siempre me apoyan en todas las decisiones que tomo en la vida. Ese apoyo incondicional me ha dado la seguridad que necesitaba para emprender cada nueva aventura al saber que, si todo salía mal, ellos estarían a mi lado. Gracias por todo, papás.

Fernando Cid Samper

Barcelona, Septiembre de 2019

Abstract

Ribonucleoprotein granules (RNP granules) are liquid-liquid phase separated complexes composed mainly by proteins and RNA. They are responsible of many processes involved in RNA regulation. Alterations in the dynamics of these protein-RNA complexes are associated with the appearance of several neurodegenerative disorders such as Amyotrophic Lateral Sclerosis ALS or Fragile X Tremor Ataxia Syndrome FXTAS. Yet, many aspects of their organization as well as the specific roles of the RNA on the formation and function of these complexes are still unknown.

In order to study RNP granules structure and formation, we integrated several state of the art high-throughput datasets. This includes protein and RNA composition obtained from RNP pull-downs, protein-RNA interaction data from eCLIP experiments and transcriptome-wide secondary structure information (produced by PARS). We used network analysis and clustering algorithms to understand the fundamental properties of granule RNAs. By integrating these properties, we produced a model to identify scaffolding RNA. Scaffolding RNAs are able to recruit many protein components into RNP granules. We found that the main protein components of stress granules (a kind of RNP granules) are connected through protein-RNA interactions. We also analyzed the contribution of RNA-RNA interactions and RNA post-transcriptional modifications on the granule internal organization.

We applied these findings to understand the biochemical pathophysiology of FXTAS disease, employing as well some novel experimental data. In FXTAS, a mutation on the FMR1 gene produces a 5´microsatellite repetition that enhances its scaffolding ability. This mutated mRNA is able to sequester some important proteins into nuclear RNP granules, such as TRA2A (i.e. a splicing factor), impeding their normal function and therefore producing some symptoms associated with the progress of the disease. The better understanding of the principles governing granules formation and structure will enable to develop novel therapies (e.g. aptamers) to mitigate the development of several neurodegenerative diseases.

Resumen

Los gránulos ribonucleoproteicos (gránulos RNP, por sus siglas en inglés) son complejos producidos mediante separación líquido-líquido y están constituidos principalmente por proteínas y ARN. Son responsables de numerosos procesos involucrados con la regulación del ARN. Alteraciones en la dinámica de estos complejos de proteínas y ARN están asociadas con la aparición de diversas enfermedades neurodegenerativas como el ELA o FXTAS. Sin embargo, todavía se desconocen muchos aspectos relativos a su organización interna así como las contribuciones específicas del RNA en la formación y funcionamiento de estos complejos.

A fin de estudiar la estructura y formación de los gránulos RNP, hemos integrado varias bases de datos de alto rendimiento de reciente aparición. Esto incluye datos sobre la composición proteica y en ARN de los RNP, sobre la interacción de proteínas y ARN extraída de experimentos de eCLIP y sobre la estructura secundaria del transcriptoma (producida mediante PARS). Todos estos datos han sido procesados para comprender las propiedades fundamentales de los ARNs que integran los gránulos, mediante el empleo de métodos computacionales como el análisis de redes o algoritmos de agrupamiento. De esta manera, hemos producido un modelo que integra varias de estas propiedades e identifica candidatos denominados ARNs de andamiaje. Definimos ARNs de andamiaje como moléculas de ARN con una alta propensión a formar gránulos y reclutar un gran número de componentes proteicos a los gránulos RNP. También hemos encontrado que las interacciones proteína-ARN conectan los principales componentes proteicos de consenso de los gránulos de estrés (un tipo específico de gránulos RNP). También hemos estudiado la contribución de las interacciones ARN-ARN y las modificaciones post-transcriptionales del RNA en la organización interna del gránulo.

Hemos aplicado estos resultados para la comprensión de la fisiopatología molecular de FXTAS, empleando también algunos datos experimentales originales. En FXTAS, una mutación en el gen FMR1 produce una repetición de microsatélite en 5´ que incrementa su capacidad como ARN de andamiaje. Este mARN mutado es capaz de secuestrar algunas proteínas importantes como TRA2A (un factor de ayuste alternativo) en gránulos RNP nucleares, impidiendo su normal funcionamiento y por consiguiente produciendo algunos síntomas asociados con el progreso de la enfermedad. Una mejor comprensión de los principios que gobiernan la formación y estructura de los gránulos puede permitir desarrollar nuevas terapias (ej: aptámeros) para mitigar el desarrollo de diversas enfermedades neurodegenerativas.

Preface

The major aim of the present thesis consists on the study of the internal structure and organization of the stress granules (SG), with a special focus on the RNA molecules that undergo granule formation. The results of this research have produced three main publications: a review, a published article and an article ready to be submitted. These publications are presented along the thesis, which is structured into seven chapters. Hereunder I briefly explain the content of each specific chapter of the thesis:

Chapter 1 presents a detailed introduction on the stress granules field, namely a theoretical description covering their definition, composition, formation, structure and function and a brief overview of the main methods employed for their study, both experimental and computational. **Chapter 2** contains a review on the *in vitro* and *in silico* methods developed for detecting the determinants on the RNA that produce the specificity on its binding with proteins. **Chapter 3** enumerates the main objectives addressed on this thesis.

Chapter 4 details the description of the main distinct properties of granule RNAs, proposing the term scaffolding RNA for those molecules with both high granule-forming and protein-interaction propensity. We apply our scaffolding model for the study of the molecular physiopathology of Fragile X Tremor-Ataxia Syndrome, a neurodegenerative disease linked with mutations on the FMR1 gene that alter the granule-forming ability of its mRNA.

Chapter 5 describes the main results obtained after confronting the model obtained on the previous chapter with new sources of experimental data. We include data considering different protein compositions under different cell types and stress conditions, analyze the first stress granule transcriptome, confront data from RNA-RNA interaction databases and study how posttranscriptional modifications may alter granule RNA structure.

Chapter 6 presents the discussion of the results obtained, detailing their main implications as well as suggesting some applications and possible future lines of research. **Chapter 7** summarizes the main conclusions of the results presented through the thesis.

Finally, the **Appendix** provides the list of scientific publications where I contributed during my doctoral studies.

Glossary of common abbreviations

ALS: amyotrophic lateral sclerosis
FXTAS: fragile X-associated tremor/ataxia syndrome
eCLIP: enhanced cross-linking immunoprecipitation
lincRNAs: long intergenic non-coding RNA
lncRNA: long intergenic non-coding RNA
LLPS: liquid-liquid phase separation
MDS: multi-dimensional scaling
miRNA: micro RNA
mRNA: messenger RNA
PPI: protein-protein interactions
PRI: protein-RNA interactions
RBP: RNA-binding protein
RNP: ribonucleoprotein
RRI: RNA-RNA interactions
SG: stress granules
siRNA: silencing RNA
snoRNA: small nucleolar RNA
snRNA: small nuclear RNA

Contents

I. INTRODUCTION

Chapter 1. Overall Introduction

In this chapter I provide an overall introduction to the topics and methods that are relevant for my thesis. The main subject of my work are ribonucleoprotein granules (RNP granules), which are key regulators of cellular metabolism under stress conditions[1,2]. Alterations in the dynamics of protein-RNA complexes are associated with the appearance of neurodegenerative disorders such as ALS or FXTAS[3–5]. I studied the RNP granules by analysing both computational and experimental data.

In the first part of the introduction, I describe extensively RNPs, their definition and classification. More specifically, my thesis is focused on stress granules (SG), which are a kind of cytoplasmic RNP granule[6,7]. Their composition and formation together with their relationship with neurological diseases is also addressed in this part of the introduction.

In the second part, I describe the experimental methods employed to study RNP granules. At this part I explain the concepts underlying protein-protein, protein-RNA and RNA-interactions and I also cover the techniques developed to detect them. Here, I explain the main methods for detecting SG composition as well as some *in situ* microscopy methods employed to observe specific components.

The third part of the introduction explains the *in silico* methods employed to analyze the experimental data described on the previous part. This covers the explanation of the main mathematical concepts employed on network theory and clustering analysis as well as the fundamental concepts behind predictive algorithms employed to obtain complementary data such as catRAPID omics and CROSS alive.

1.1 Stress granules (SG)

SG are defined as cytoplasmic RNP granules induced after stress and translational repression[6,7]. The initial characterization of SG began in the late 1990s when observed that impairment of translation initiation causes the formation of liquid droplets in the cytoplasm[8,9].

They have recently been extensively studied due to their implications in several cellular processes. First, stress granules affect mRNA localization, translation and degradation and therefore represent a crucial step for understanding RNA cell cycle[10]. Second, they are especially important in neuronal tissues where they are master regulators of gene expression [11]. In this sense, mutations affecting their function are causative of

neurodegenerative diseases[5,12]. Finally, RNP granules have been classified as membrane-less organelles, they represent a functional and efficient strategy of cellular organization that are not fully understood[1]. For instance, some metabolic enzymes have been recently shown to be present in this form, which may explain the relative high kinetic rates of metabolic pathways[13,14].

Recently, there has been a vision change in the field from a protein-centric (inherited from the aggregation field) to another in which the RNA can also be a main actor[15,16]. This has led to discover the very first compendium of RNA molecules inside the SG[10]. The work reveals that any mRNA in the cell can be targeted to the SG (at a low proportion in comparison with the rest of the cytoplasm through) as well as many non-coding RNA (ncRNA) species. Remarkably, long noncoding RNAs (lncRNAs) have been described to do important scaffolding functions in other types of RNP granules, such as paraspeckles[31]. Still, there are no clear models for SG formation nor their function while some theories start to develop from the current data and observations obtained[17,18]. A further comprehension of these structures will be fundamental for improving the treatment of many different neurological diseases involved with SG impairment[12].

1.1.1 Definition and classification of RNP granules

RNP granules are liquid-liquid phase separated complexes composed mainly by proteins and RNA[1,2]. They are present across all eukaryotes and conserved from yeast to mammals[7]. Liquid-liquid phase separation (LLPS) is a (bio)-physical demising process that occurs between two immiscible liquids (e.g. oil and water)[19]. This process produces two or several distinct and separate homogeneous mixtures called phases[20]. LLPS is produced when the energy of interaction between macromolecules is greater than the entropic energy reduction that arises from their homogenous mixing[21,22]. Membrane-less compartments formed by LLPS are able to undergo fission, fusion and show rapid components exchange, which can be observed by fluorescence recovery after photobleaching[7,23].

In biology phase separation is a way to organize molecular interactions, since it allows the compartmentalization of biomolecules into organelles without the presence of a specific membrane[24]. Both protein and RNA are able to promote phase separation due to their ability to establish multivalent interactions (i.e. several binding contacts) with other proteins and RNAs respectively[17,25,26]. However, for the specific case of SG, LLPS maintenance seems to be an energy-consuming process [27]. Other roles of the presence of ATP in the SG are being discussed; such it seems to help on the demixing as an hydrotrope[21,22].

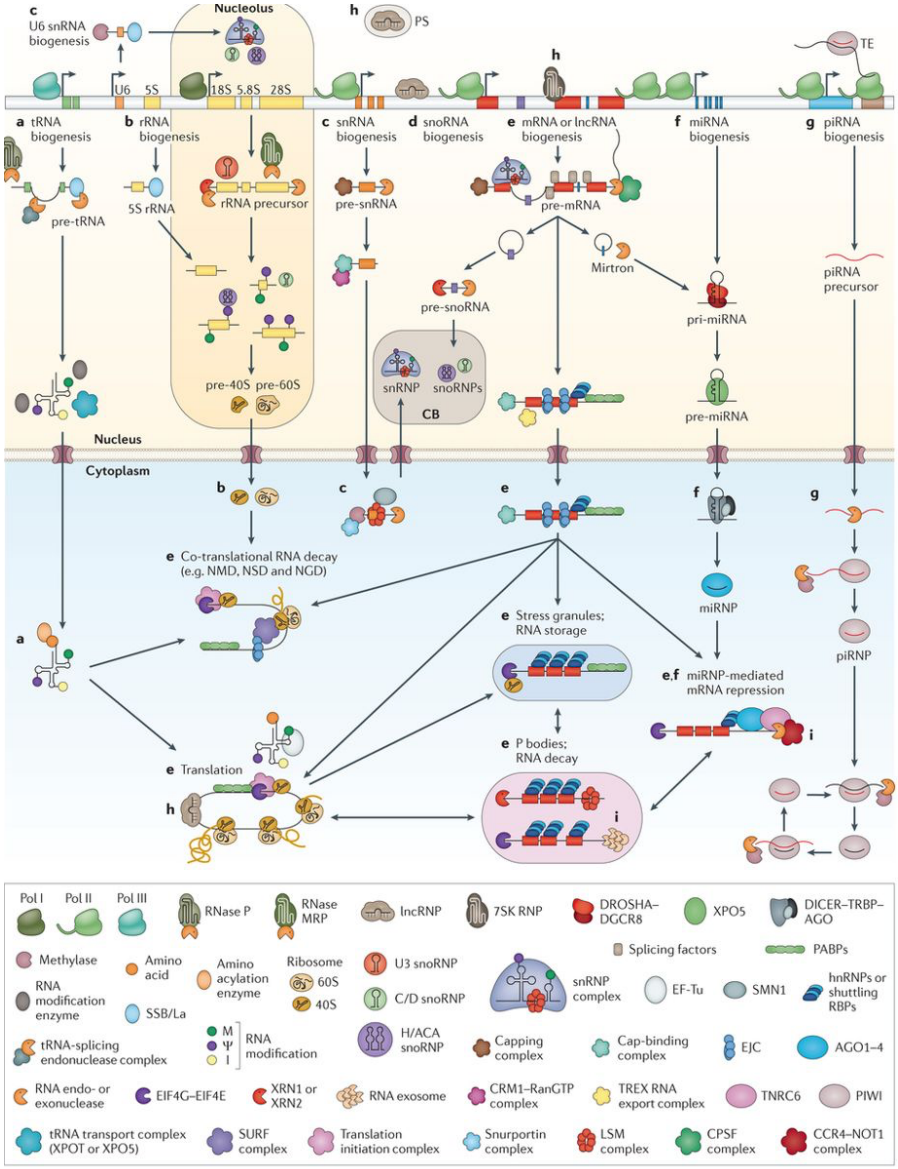


Figure 1. Overview of the main post-transcriptional gene regulation pathways in eukaryotes including the role of the main types of RNP granules. An overview is given for the biogenesis, decay and function of the most abundant RNAs: tRNAs, ribosomal RNAs, small nuclear RNAs (snRNAs), small nucleolar RNAs (snoRNAs), mRNAs, microRNAs (miRNAs), PIWI-interacting RNAs (piRNAs) and long non-coding RNAs (lncRNAs). CB: Caja bodies, PS: paraspeckles. Adapted from S. Gerstberger, 2014 (Nature Reviews).

RNP include highly diverse group of compartments that varies in function and composition depending on the organism, cell type, location and condition (**Figure 1**).

Nucleus: examples include the Cajal bodies, paraspeckles or the nucleolus[28]. Cajal bodies are found in the nucleus of proliferative cells or neurons. They are involved in small nuclear RNA (snRNA) and small nucleolar RNA (snoRNA) regulation although their function is not yet understood[29]. Paraspeckles are present in the interchromatin space and seem to control the translation of certain RNA molecules by retaining them into the nucleus[30,31]. The functions of the nucleus have been extensively described as the center for ribosome biogenesis[32]. Other less-defined nuclear granules are recently proposed as transcriptional enhancers[33].

Cytoplasmic: two main types in this category are SG and processing bodies (P-bodies). Although they seem to have different distinct function they are both composed by pools of untranslating mRNAs[1,6]. Actually, they share many components (10-25% of their protein components) that they can even interchange by docking together[23,34]. They are generally dynamic (i.e. able to exchange components also with the cytoplasm) and dependent on RNA for their assembly[17,23]. Specifically, SG are formed by mRNAs stalled in translation and they contain several translation initiation factors, a variety of RNA-binding proteins (RBPs) and many non-RNA binding proteins[10]. They are formed during stress promoting survival[24]. P-bodies instead contain mRNAs associated with translation repressors and mRNA decay machinery[35] . Some RNAs contained in them are targeted for autophagy[36].

Tissue specific: such as neuronal granules, germ cell granules and the Balbiani body. Germ granules are involved in specific post-transcriptional regulation events required on neurons for certain mRNAs related to synaptic remodeling[37]. Germ granules seem to be a source of maternal mRNA storage in early development. Balbiani body is a specific structure of female germ cells that contains most of the organelles in dormant oocytes and disappears as the oocyte matures[25].

1.1.2 Protein composition of SG

Proteins and RNAs are the two main components of SG[1]. They enable many different interaction modes mainly named protein-protein (PPI), protein-RNA (PRI) and RNA-RNA interactions (RRI). Usually these interactions are multivalent and create a high dense contact network[38] that would promote SG formation. In this sense, most interactions isolated are not essential, but overall they lead to SG formation by synergistic, emergent processes[18,39].

Specifically, regarding protein composition**,** SG contain: initiation factors (EIF2a/3/4/4b/4G0), 40S ribosomal units, components of miRNA pathway (ZFP36, TNRC6B, AGO2), many translation repressors (Carpin-1, TIA-1/TIAR [Pub1/Ngr1 in yeast]), RBPs related with RNA decay and stabilization (G3BP, DX6 [Ded1 in yeast], TDP-43, PAB1), enzymes with

ATPase activity (RUVBL1/2 [Rvb1/2 in yeast], MCM, CCT), helicases (DDX3 [Ded1 in yeast]) and chaperonins (HSP80, HSP40)[2,27,40]. The function of some of these proteins indicates a possible structure remodelling inside stress granules that could affect formation or disassembly.

Despite presence of chaperonins and helicases, half of protein components of SG are RBPs[27]. SG proteins that do not bind RNA are presumably recruited through protein-protein interactions. Among all different kinds of RBPs, proteins present in SG are specifically enriched in intrinsic disorder regions (IDRs, e.g. hnRNPA1, Ddx4, FUS, Whie3)[20,41]. IDRs are protein sequences that lack a defined 3D structure because of the absence of a core of hydrophobic amino acids. IDRs are very promiscuous interactors, promoting the formation of multivalent assemblies[42,43].

Protein composition varies depending on the stress conditions where the SG are formed. For instance, Gtr1, Rps1b and High1 in yeast promote SG during glucose starvation but suppress it during heat shock[44]. In the case of mammals, GP31, a widely used marker for stress granules, it is crucial for its formation under oxidative stress by interacting with caprin RBP[45,46]. However, GP31 is not necessary for SG formation during osmotic stress or heat shock[47]. Furthermore, protein modifications, such as methylation, phosphorylation and glycosylation may influence SG assembly by altering specific protein-protein interactions[48,49]. For instance, the phosphorylation of G2BP impairs its ability to multimerize, impeding granule assembly[46]. In contrast, phosphorylation of Grb7 and DYRK3 kinase promotes granule disassembly during recovery[49].

This heterogeneity in composition suggests that SG may have different functions for different stresses. In this sense, Markmiller et al, 2018 used ascorbate peroxidase proximity labelling (APEX) paired with mass spectrometry and immunofluorescence to characterized protein composition in SG under different stresses and cell types[50]. They identified 260 SG associated proteins, 20% of them being stress or cell type specific. Analogously, it can be also inferred from their work a set of proteins that was present under all the stresses and cell types studied.

Biotinylation was also applied to study SG composition by a proximity-labelling approach (BioID)[51]. They identified 119 human SG proteins enriched in functions related to mRNA processing, revealing also functional clusters based on proximal protein-protein interactions. Both proximity-labelling studies observe a pre-existing network of interactions between SG components under normal growth conditions. These interactions would coalescence under stress conditions to initiate the SG nucleation, although this process is still unclear[50]. Some models suggest that interactions during normal growth states would be sub-stoichiometric, while interactions become more concentrated in granules during stress[52]. These pre-existing

interactions may drive the preassembly of sub-microscopic granules at early stages of their formation[10,33].

1.1.3 RNA composition of SG

The analysis of the RNA composition has been very elusive until very recently due to the RNA unstable and transient nature. Khong et al, 2017 provided the first compendium of RNA composition of SG in U2OS cells[10]. They purified SG cores using G3P1 protein as bait. These cores are small stable (solid-like) structures located in the inner of the SG that are surrounded by a dynamic LLPS shell. This study described that almost any RNA can be driven into SG, including some lncRNAs and other ncRNA species. However, only 9.4% of the total mRNA in the cell accumulates in SG at a given time. Moreover, there is no highly enriched RNA in SG, being the most common species the actin mRNA, with only 0.5 % of the total SG mRNAs. Despite this, 1841 transcripts (1626 mRNAs and 215 ncRNAs) have been classified as enriched in SG, as they are more concentrated in the granule than in the cytoplasm. Based on the same approach, 2539 transcripts (1780 mRNAs and 759 ncRNAs) have been classified as depleted from SG cores.

Additionally, Namkoong et al, 2018 provided another transcriptome for cytoplasmic droplets composition after endoplasmic reticulum stress[53]. However, this dataset shows higher correlation to P-bodies composition than to SG[54]. Also, many ER markers were almost completely depleted in the droplets formed after endoplasmic reticulum stress[53].

RNAs enriched in SG shown specific properties such a longer 3' UTR and lower translational efficiency rates[10]. Other studies also highlight the importance of RNA structure for its location into SG. For instance, Langdon et al. (2018) showed that recruitment of CLC3 mRNA intro droplets is dependent on its secondary structure[55]. Finally, AU-rich elements (AREs) are also strongly correlated with SG-targeting of mRNAs upon analysis of motifs[53], though consensus sequences are still controversial for mRNA targeting[2].

To conclude, the possibility of the presence of RNA-RNA interactions within the SG has raised attention recently (**Figure 2A**). RNA-RNA interactions may even drive the formation of SG RNA self-aggregates *in vitro* following similar principles of those observed for SG *in vivo*[17,56].

1.1.3 Formation, structure and function of SG

A wide range of stresses can trigger the formation of SG and PB such as heat shock, oxidative stress, UV irradiation, osmotic stress and nutrient starvation[2]. Formation occurs on the scale of minutes after exposure to stress stimuli. First mechanism described for SG formation consists on the

phosphorylation of the translation initiation factor eIF2α[47]. However, alternative mechanisms that also block translation initiation can provoke SG formation. Examples include the knockdown of specific translation initiation factors, the overexpression of RBPs that repress translation or the addition of small molecules with the ability to block translation initiation[39,57,58].

Formation seems to occur either in the proximity of ribosomes or from P-bodies material both in yeast and mammals[34]. Assembling of SG involves several steps in a process not fully understood yet[33]. Initially, small oligomeric assemblies are formed from untranslated mRNPs[10]. In this sense, SG fail to form when mRNA are trapped in polysomes[59]**.** In a similar direction, addition of puromycin (that dissociates ribosomes from mRNAs) triggers SG formation[34,60]. Once the initial nucleation assemblies are created, these start a process of growing by the recruitment of additional mRNPs that form small SG of 200 nm[27] (Figure 2B). In mammals, these smaller granules finally merge and form higher-order assemblies with many core structures surrounded by a so-called shell (Figure 2C). These internal regions or cores have higher concentration of proteins and mRNAs whereas the shell is less dense but more dynamic[27,61]. All the steps from initial nucleation until maturation are helped by the disposition of specific microtubule organization[62,63].

There are two main models that explain the formation of the substructure within the stress granules[1,33]. A first model proposes that the cores observed in mature SG come from the first nucleation complexes that are assembled at the first steps of SG formation. The outer shell would be entirely formed by components attached to the initial core (**Figure 2B**). An alternative model suggests that both processes (initial nucleation and growing of the stress granule and maturation) would be independent. In this model, cores would be constituted after the prior maturation and reorganization of a complete stress granule.

SG are highly dynamic. They flow in the cytosol and can undergo fusion and fission[23]. Also, they are able to disassemble into translating mRNAs and can undergo clearance by autophagy. Fluorescence recovery after photobleaching (FRAP) show that most components are exchanged rapidly between the SG and the cytosol, with half-lies for recovery smaller than 30s[7]. There are components that are less dynamic that presumably belong to the core structure. SG and PBs are able to interact, docking and swapping components but they have unique RNA and protein content[64].

SG are enriched in mRNAs that code for proteins related with translation initiation, translational repression and mRNA degradation[27]. Therefore, they seem to be related with mRNA regulation. This includes functions such as

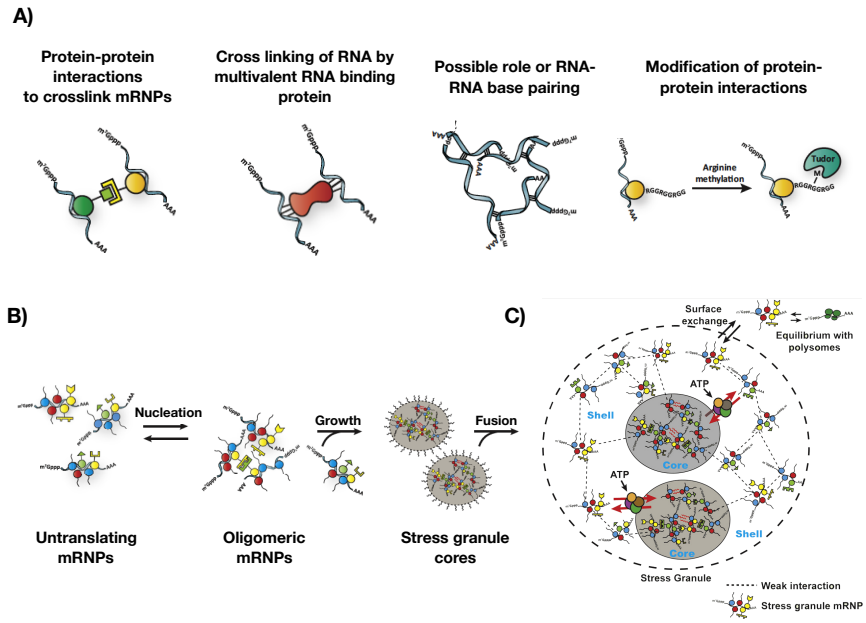


Figure 2. Overview of stress granules (SGs): structure, formation and governing interactions. A) Main kind of interactions involving proteins and RNAs present on the SGs. Proteins and RNAs interact both with themselves and each other. Posttranscriptional modifications also play a role on modulating some interactions. **C) Internal structure of SGs.** Highly dense stable cores are surrounded by a more dynamic and external structure called shell. **B) A model proposed for the formation of mature SG.** This model suggests that SG cores would be the first structures to be constituted and the subsequent addition or molecules to the complex would constitute the shell in the mature SGs. Adapted from DSW. Protter, 2016 (Trends Cell Biol.) and S. Jain, 2016 (Cell).

storage, decay or eventual reintroduction of mRNAs to the translating pool after the stress overcome[2]. Mutants that cannot form mRNP granules are more sensitive to stress[44,65,66]. However, it is difficult to study their function and discriminate the SG contribution from other stress-induced responses in the cell.

It is not clear if proteins maintain their function inside stress granules. The presence of enzymes in SG suggest that they may also help in concentrating and producing components of metabolic reactions[14]. Similarly, the presence of chaperonins may indicate their ability to participate in the folding and maintenance of certain proteins or RNAs[67]. Finally, it has been also observed that the sequestration of certain proteins to SG modulates the activity of signalling pathways as the case of TOR, RACK1 or TRAF2[4,68,69].

1.1.4 Stress Granules and Disease

Mutations that affect SG formation or persistence usually contribute to degenerative diseases such as Amyotrophic Lateral Sclerosis (ALS) and FrontoTemporal Dementia (FTLD)[3,5]. For instance, TDP-43 mislocalization from the nucleus to cytoplasmic inclusions leads to the appearance of ALS[70]. Mutations on the C9ORF72 gene that leads to its accumulation in nuclear foci are also linked to the appearance of ALS[71]. A similar example occurs in the case of fragile X tremor ataxia syndrome (FXTAS) patients, where a CGG expansion present in the FMRI gene also provokes its toxic accumulation in nuclear droplets[4].

Regarding FXTAS disease more specifically, CGG repeats in FMR1 5'UTR are of different lengths (being 30 repeats, the most common allele in Europe) [4]. However, mutations that contain over 200 repeats block the FMRP protein expression by a process of methylation of silencing of the FMR1 gene. It is noteworthy that nuclear foci are the typical hallmark of FXTAS[4]. These droplets are highly dynamic and dissolve upon tautomycin treatment, (characteristic of RNP granules). They contain proteins such as HNRMP, A2/B1, MBNL1, LMNA and INA as well as some splicing regulation factors such as CUGBP1, KHDRBS1 and DGCR8[4]. It is still unknown the molecular physiopathology of FXTAS and therefore there is a lack of molecular targets for a therapeutic intervention[4].

However, RNP granules disorders are not only related to neurodegeneration. For instance, some microsatellite expansion in the UTRs of DMPK and ZNF9 mRNAs are causative of myotonic dystrophy types 1 and 2 respectively[72,73]. In both cases, the mutated RNAs aggregate into nuclear foci, sequestering their function as splicing factors. Similarly, mutations on the TIA1 gene are associated with the Welander distal myopathy[74]. Some studies show even a relation between RNP granules and tumor progression and treatment. For instance, there are chemotherapeutic drugs promote assembly of non-canonical SG[75].

In general, repetitive expansions promote RNP granule assembly and the appearance of associated diseases[76]. Although the mechanism is not completely clear, the most accepted model suggests that pathological mutations would provoke abnormalities on the assembly and clearance of normal SG. These defects would lead to the formation of hyperstable, solid-like amyloid fibrils instead of the normal dynamic and liquid-like SG[1]. This effect will ultimate triggers cell death by altering the regulation, biogenesis and function of several RNAs that would be trapped inside these abnormal droplets.

1.2 Experimental methods for studying RNP granules.

In this section, I describe: (i) the biological basis required to understand how protein and RNA molecules interact; (ii) the main methods employed to obtain protein-protein, protein-RNA and RNA-RNA interaction data, (iii) an approach to detect protein and RNA composition of RNP granules and (iv) techniques (i.e. FISH) to visualize the content of the granules *ex vivo*.

Since obtaining information about the interactions inside the granules is experimentally challenging, it is reasonable to assume that the interactions observe outside the granules would also occur inside if molecules localize together. This assumption is based on the fact that interactions depend solely on the physico-chemical properties of the interacting molecules. These properties should not change in order to maintain the function of the molecules inside the granules. Therefore, the presence of two molecules inside the granule that interact outside should imply also their interaction in the granule. As shown in the results presented on this thesis, this assumption correlates well with further experimental data published during the thesis development.

1.2.1 Fundamental concepts of protein and RNA: sequence and structure

Since proteins and RNAs mainly compose RNP granules, protein-protein, protein-RNA and RNA-RNA interactions govern their internal organization. These interactions are a consequence of the distinct physico-chemical properties of each individual protein or RNA species. Both proteins and RNAs are biological polymers, and their physico-chemical properties depend solely on their primary sequence (e.g. the order and quantity of the monomers that form the complete molecule)[77]. This sequence determines their final secondary (the three dimensional form of local segments of the molecule) and tertiary structure (the three dimensional form of the whole molecule, as a consequence of further folding of the secondary structure segments). The three dimensional structure creates a specific electrostatic and steric interface that enables only certain interactions with complementary interfaces of other specific molecules[78]. In summary, protein and RNA sequences contains the information of their tertiary structure and therefore their set of potential interactors.

The protein sequence is formed by amino acids that consist of an amine (-NH2) and a carboxyl (-COOH) functional groups, along with a side chain (R group) specific to each amino acid[79]. Amino acids are then binding though peptide (amide) bonds between the carboxyl group of a certain amino acid and the amine group of another forming a progression known as peptide chain. Most of the biological proteins are composed by a combination of 20 main amino acids that vary only on their side chains

chemical composition, conferring to each amino acid specific physico-chemical properties. Electrostatic attraction and repulsion together with other weak forces among the side chains of all the amino acids in a peptide chain determine the structure of a protein.

RNA is also a polymeric molecule constituted by nucleotides[80]. Nucleotides are composed of a 5-carbon-ribose sugar, a phosphate group (H3PO4) and a nitrogenous base. They bind each other through phosphodiester linkages between the 5' and 3' carbon atoms of two adjacent riboses. There are four different types of nitrogenous bases that combine in different proportions to form RNA molecules. These are the adenine, the uracile, the guanosine and the cytosine. This nucleotide chain is very flexible and usually single-chained but can be folded onto it to form double-stranded regions that comprise a secondary and sometime tertiary structure. This folding is mainly due to weak hydrogen bonds interactions between complementary nitrogenous bases, namely adenine and uracile or guanine and cytosine. These complementary interactions can either be from different parts of the sequence of a single RNA molecule or come from two different RNA molecules, establishing therefore RNA-RNA interactions.

1.2.2 Interactions within a RNP assembly

Protein-protein

Protein-protein interactions were the first kind of interactions to be studied with high-throughput technologies with the development of yeast-two hybrid system (later optimised for mammalian cells) [108]. Two-hybrid system is based on the cloning of a reporter gene activated by a transcription factor that binds a regulatory promoter sequence (i.e. upstream activating sequence, UAS). The transcription factor is then split into two separated fragments (i.e. the DNA-binding domain, DB and the activating domain, AD). Each of the two proteins that want to be tested for interaction is fused with a different domain of the transcription factor. If proteins do bind together, the transcription factor will be functional and the reporter gene (e.g. LacZ) will be transcribed. The generation of yeast libraries containing different cloned colonies for any possible pair of proteins in a system enables to obtain protein-protein interactions on a high-throughput scale.

Protein-protein interactions datasets are still the largest datasets available on any biological interaction. Specifically, the Biological General Repository for Interaction Datasets (BIOGRID v.3.4, https://thebiogrid.org/) contains a total of 1.559.32 curated interactions in all major model organism species. It stores several kinds of protein-protein interactions, being physical interactions the highest accurate source of direct contacts between proteins.

Protein-RNA

The enhanced Cross-Linking and ImmunoPrecipitation (eCLIP) dataset is the biggest source of protein-RNA interactions available. Recent studies corroborate that eCLIP detected interactions correlate well with known *in vitro* and *in silico* experiments[81]. A complete description of the different methods available for detecting protein-RNA interactions *in vitro* and *in silico* and of the main RNA determinants for protein binding is presented on **Chapter 2** of this thesis.

In general, CLIP methods combine UV cross-linking with immunoprecipitation in order to analyze the target RNAs and binding sites of a certain RBP[82]. UV cross-linking produces the formation of covalent bonds between proteins and nucleic acids in the proximity. Cross-linked cells are then lysed with proteinase K and the protein of interested is isolated via immunoprecipitation. Finally, retrotranscription and amplification with barcoding enables to identify the binding sequence. There are different CLIP techniques. The eCLIP protocol uses specific adaptors to enable decrease the requisite amplification by 1000 fold, as it discards most of the PCR duplicated reads[83].

The eCLIP methodology have been applied to two cell lines: K562 and HepG2. To date, the K562 cell line contains information about the RNA targets of 98 proteins[81].

RNA-RNA

Thanks to recent technical improvements, RNA-RNA interactions are starting to gain importance for the understanding of the RNA biology[17]. RNA molecules can interact with each other through base pairing, such as the case of miRNA and siRNAs[84]. In a similar extent, lncRNAs and mRNAs can interact with each other following the same principles[85]. This highly unexplored interaction world of RNA molecules may regulate important steps of the RNA life cycle, including the formation, maintenance and disassembly of SG.

Recent methods to detect RNA-RNA interactions are based on RNA proximity ligation coupled with high-throughput sequencing. For instance, LIGR-seq enables to detect RNA-RNA interactions in a global-scale[86]. This technique employs 4'-amiomethyltrioxalen and 365 nm UV irradiation to *in vivo* crosslink RNA duplexes. After digestion of linear and structured RNAs by RNase R and purification, high-throughput sequencing is performed. Finally, a computational method uses the sequencing data to discriminate intra or inter-molecular interactions.

A compendium of RNA-RNA interaction experiments is compiled in RISE (database of RNA Interactome from Sequencing Experiments)[85]. RISE contains information from different transcriptome-wide sequencing-based experiments like PARIS, SPLASH, LIGR-seq and MARIO[30,86–88], as well as complementary information from targeted studies like RIA-seq, RAP-RNA and CLASH[89–91]. This represents a total of 328,811 RNA-RNA interactions in human.

1.2.3 Methods microscopy independent for determining SG content

Stress granules protein composition was characterized by stable cores extraction followed by mass spectrometry or sequencing analysis[27]. These cores are extracted using specific differential centrifugation spins in order to subsequently create core-enriched fractions. The cores are then purified by affinity purification using antibodies and dynabeads targeted against GFP attached to G3BP protein (a constitutive SG component). Finally, the protein content can be characterized using mass spectrometry and the RNA components using sequencing[10].

Another method for characterizing SG proteins is the BioID approach[92]. BioID fuses an abortive biotin ligase to a bait protein to mediate biotinylation of proximal polypeptides within ~10 nm[92]. Youn et. al, 2018 employed this method on 119 human proteins involved in mRNA biology as bait to detect SG and P-bodies proteins[51]. They identified 44 proteins as part of the SG proteome.

However, since SG may have different protein compositions under different conditions, it is important to characterize SG on different cell types and after different stress inductions. In this sense Markmiller et al., 2018 employed ascorbate peroxidase (APEX)-mediated *in vivo* proximity labelling paired with mass spectrometry and immunofluorescence to identify SG proteins under different stress conditions (NaAsO2 and heat shock) and cell types (HepG2, HeLa, NPCs)[50]. APEX purification tag was also fused with the G3BP1 protein as in similar studies. The ~20% of SG proteins identified by this approach were specific on certain cell types or stresses.

1.2.4 Methods microscopy dependent for determining SG content

Fluorescence *in situ* hybridization (FISH) represents one of the main techniques to observe the location in the SG of individual RNAs *ex vivo*[93]. FISH uses fluorescent probes that bind to specific nucleic acid sequences and therefore allows to track targeted RNA molecules[94]. Single-molecule FISH is a variant of the method that allows to detect and quantify both mRNA and lncRNAs[95]. This enables to study how different species of RNAs or mutations of the same gene can affect to their granule (mis)-

location and concentration. In this sense, effects of gene mutations for granule formation and its implications on certain diseases can be explored in detail. Also, this technique is useful to understand which specific properties make certain RNAs to become more granule-prone.

Other imaging techniques employed to study SG are able to track individual molecules *in vivo*. This allows to track in detail the formation and remodeling of SG as a dynamic event[96]. These techniques include the use of oligodeoxynucleotides, molecular beacons (hairpin shaped structures with a fluorophore and quencher attached) or aptamer-fluorogen systems such as the Spinach 3,5-difluoro-4-hydroxybenzylidene imidazolinone system. These labelling methods can be used paired with multifocus microscopy (MFM), light-sheet microscopy or fluorescence correlation spectroscopy to observe dynamically *in vivo* specific RNA molecules[97].

1.3 Computational methods for studying RNP granules.

SG are complex structures comprising many elements interacting in several modes[2]. In this sense, some of their properties arise by the specific interaction among the elements[33]. As explained above, it is difficult to define a set of crucial proteins or RNAs required to granule formation because this process involves redundant interactions[1]. Also, it has been proposed by several studies the idea of a pre-exiting network of interactions before the formation of SG structures[50]. This pre-existing network would produce nucleation structures when specific conditions arise. These conditions are met under SG formation by an increase of the concentration of the interacting elements due to a general translational repression[33].

Therefore, network analysis represents an interesting approach to study SG. It provides a methodology to study complex systems with a global perspective and to detect the importance of long-range interactions for the organization of the droplet. Here, I describe the basic concepts of network science and the tools employed in this thesis. In a more general extent, I detail clustering methods such as hierarchical clustering and multidimensional scaling that allow to detect or visualize the internal organization of complex structures such as SG. Also, I briefly cover some predictive methods such as catRAPID omics and CROSS alive that are useful for exploring systems with a lower data availability.

1.3.1 Network analysis

Many complex systems such as chemical systems, neural networks or Internet can be modelled as graphs or networks composed of nodes representing the interactions between them[98]. This approach describes how

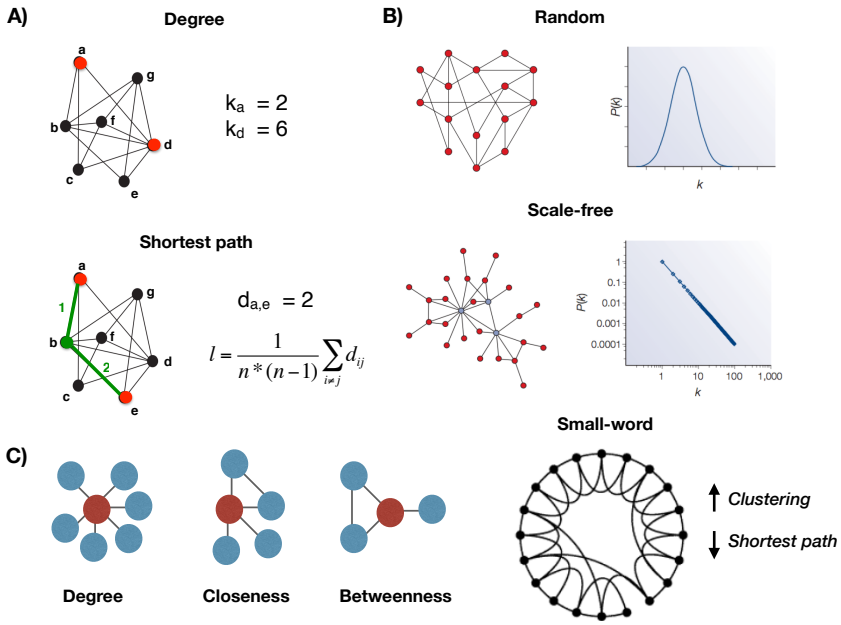


Figure 3. Summary of main concepts regarding network analysis and network properties. A) Degree and shortest path definitions. Networks are defined as a series of nodes connected though a series of edges. Degree (k) is the number of edges that a certain node has. Shortest path is the minimum number of nodes required to traverse from a certain node A to another node. **B) Main kinds of networks according to their degree distribution**. In random networks, all the nodes have degrees similar to the average degree. In contrast, scale-free networks present certain nodes called hubs that have a much higher number of connections than the rest of the network. The properties of scale-free networks enable to reach any node of the network in a few steps **C) Network centrality measures**. Degree is the number of connections a node has. Closeness measures the ability to reach in fewer steps on average all the other nodes in the network. Betweenness is defined as the number of shortest paths in the network that pass though a specific node.

these systems behave as a whole and therefore it allows detecting emergent properties that arise from the interplay between topology and dynamics, which were not deductible by the individual analysis of its components[99].

Emergent properties are especially important in biology, where the function of a given cell component cannot be usually understood neglecting its relationships with the rest of the system[100]. Moreover, the raise of high-throughput technologies and their ability of detecting how and when most of the cellular components interact, have allowed building very detailed and reliable biological networks. Protein-protein, protein-RNA or RNA-RNA interactions are examples of networks whose study can help us to understand the emergent properties of complex biological systems.

The description of biological systems as interacting networks of different components has interesting applications. For example, Guimera et al. (2005) analyzed the metabolic networks of twelve organisms discovering a module functional structure[101]. Moreover, a work by Li et al. (2004) based on a dynamical model of the yeast cell-cycle regulatory network demonstrated that the network structure had stable and robust properties that were important for the cell-cycle dynamics[102].

Despite the remarkable diversity of networks in nature, their architecture is governed by simple principles that are common to most scientific and technological networks [99,103]. In this sense, there are specific network measures that allow to compare and characterize their properties. First, the degree (k), represents the number of links that connect a node to others (**Figure 3A**). Second, the degree distribution P(k), gives the probability that a selected node forms exactly k links and it is obtained by counting the number of nodes N(k) with k =1, 2, …, n links and dividing it by the total number of nodes N (**Figure 3B**).

P(k) allows to distinguish between different classes of networks with different basic properties. For instance, random networks tend to have a P(k) following a normal distribution, where most of the nodes have a degree similar to the mean degree of the network (**Figure 3B**). However, random networks cannot explain most of the properties of real world networks, that are better modelled as scale-free networks[100]. P(k) on scale-free networks typically follows a power law ($P(k) \sim k^{\gamma}$, $2 < \gamma < 3$)[104]. This distribution reflects a non-uniform behaviour where most of the nodes have only a few links in contrast to a few nodes (often called hubs) that hold a very large number of connections (**Figure 3B**). Hubs have usually important functional roles within the networks, such as regulatory points in biological networks.

Degree is considered a centrality measure, where hub nodes are considered more central (Figure 3C). Centrality is associated with the importance of a node within a network[100]. Since there are different definitions on the importance of a node, there are different centrality measures. In general, they weight the role of a certain node on connecting different parts of the networks and therefore the grade of disturbance on the network if the node would be removed. Nodes with higher centrality values usually represent elements that would isolate or disrupt parts of the network if removed. In this sense, nodes with high degree (high number of interactions) are considered central since they can regulate a many other nodes. Other important centrality measures are based on the concept of shortest path between two given nodes in the network, these include betweenness and closeness centrality[100]. **Shortest path** is the minimum number of nodes required to connect two given nodes (**Figure 3A**). **Betweenness** is defined as the number of shortest paths in the network that go through a certain

node (**Figure 3C**). **Closeness** is defined as the inverse of the average of the shortest paths between a certain node and all the other nodes in the network (**Figure 3C**). In general, in any complex network two nodes are connected through a path of a few links only. This effect is known as small-world and indicates that information transmission within the network is highly efficient (**Figure 3B**)[99].

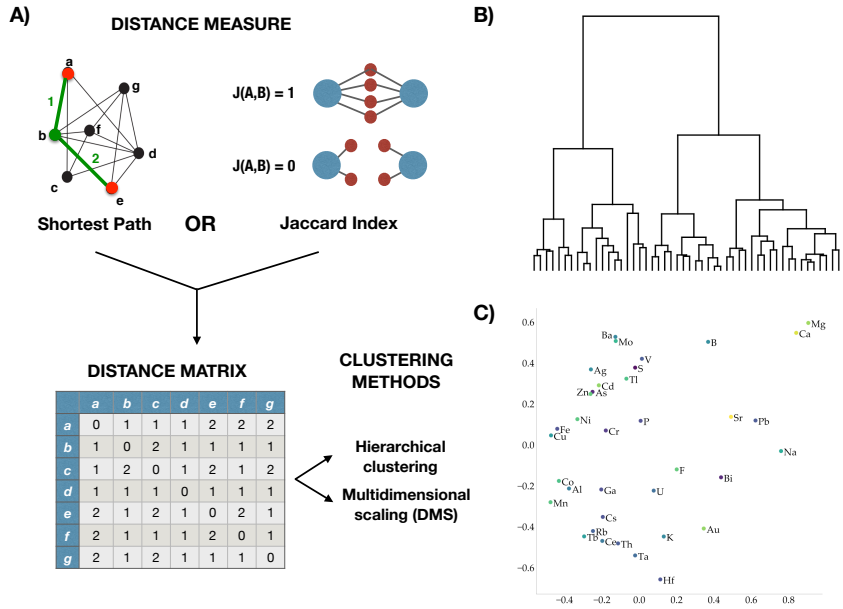
Networks can also be used to identify different communities commonly referred as clusters. Clusters are defined as group of nodes that are more connected to each other than with the rest of the network. When talking about biological networks, clustering has been postulated to have strong biological implications. Commonly, the emergence of modules within biological networks results in sets of interacting agents sharing functional ontologies[104,105].

1.3.2 Classification algorithms: clustering methods

Clustering analysis is an exploratory data analysis technique that allows discovering associations and structures within the data that are non-obvious. In this manner, hierarchical groups could be identified to describe different functional groups and give a global view of the system[106]. Clustering analysis works with distance matrices that indicate the proximity between two elements in the system (**Figure 4A**). These distances are not necessary physical distances but measures that reflect similarity or functional relationship. For instance, the shortest path can be used as a distance measure for systems described as networks (**Figure 4A**). The shortest path between two nodes will indicate their proximity in the network and therefore their functional relationship, since two nodes that work on a similar function tend to be interacting together or with the same partners. Shortest path allows detecting and comparing long-range relations between any two nodes in the network[100]. For stressing local relations instead, Jaccard index is more appropriate[107]. **Jaccard index** gives a measure of the amount of common partners of two given nodes (e.g. overlap of RNA targets between a two RBPs). The Jaccard index (**Figure 4A**) is computed as the intersection of targets of two given nodes (their common partners) divided by the union of their total targets (in order to normalize for nodes with a high number of interactions).

Using any of this distance measures, a distance matrix can be computed (**Figure 4A**). Then, several clustering algorithms to detect and interpret the clustering relations in the system can be applied on these distance matrices. These methods include for instance hierarchical clustering and multidimensional scaling algorithms.

Hierarchical clustering produces a dendrogram or tree that reflects the hierarchy between all the elements in the system (**Figure 4B**)[106]. Based on



	a	b	c	d	e	f	g
a	0	1	1	1	2	2	2
b	1	0	2	1	1	1	1
c	1	2	0	1	2	1	2
d	1	1	1	0	1	1	1
e	2	1	2	1	0	2	1
f	2	1	1	1	2	0	1
g	2	1	2	1	1	1	0

Figure 4. Brief explanation of the clustering methods. A) Schematic representation of clustering algorithms pipeline. Clustering methods employ a distance measure to generate a distance matrix that represents similarity betweene the elements in the system of interest. **B) Example dendrogram**. Dendrograms are the main output of hierarchical clustering algorithms. Each leaf on the tree corresponds to a different element on the dataset. Closer elements are represented as leafs merged together a lower height. **C) Example of a multidimensial scalling (MDS) output.** MDS projects the information of a distance matrix into a 2D representation that mimics the distances between elements (points) in the original matrix.

the distance matrix, hierarchical clustering algorithms work searching the lower distance value between two elements of the matrix. Then, these two elements are considered as a single point and all the distances between this new cluster and the rest of the elements of the system are recalculated. The new distances can be recalculated using different methods. For example, in single linkage (also known as nearest neighbour clustering) the distance between two groups is defined as the distance between their two closest members. Finally, this process is iterated and the result can be represented as a dendrogram, which is a graphic tree representation that summarizes the clustering process. In the dendrogram, each branch in the tree represents a certain element. Branches that represent similar elements split later in the tree and therefore tend to appear as close leaves.

Multidimensional scaling (MDS) is an approach that seeks a configuration in low-dimensional space such that the distances between points in the space match the (dis)similarities contained in a distance matrix as closely as

possible (**Figure 4C**). The degree of correspondence between the distances among points represented in the space and the input matrix is measured by a *stress* function:

$$\sqrt{\frac{\sum\sum(f(x_{ij})-d_{ij})^2}{scale}}$$

Equation 1

where d_{ij} refers to the euclidean distance, across all dimensions, between points i and j on the map, $f(x_{ij})$ is a function of the input data, and *scale* refers to a constant scaling factor, used to keep stress values between 0 and 1. If the MDS space perfectly fits the input data stress is zero. Practically, the smaller the stress, the better the representation. The function of the input values $f(x_{ij})$ used varies between MDS algorithms (metric or non-metric scaling). In metric scaling $f(x_{ij}) = x_{ij}$, so the input data is compared directly to the distances in the space. In non-metric scaling $f(x_{ij})$ is a weakly monotonic transformation of the input data that minimizes the stress function, that is computed via a monotonic regression.

The stress value also depends on the number of dimensions: in general, increasing the number of dimensions leads to a decrease in stress. This happens because for any given dataset, it may be impossible to perfectly represent the input data in two or other small number of dimensions. On the other hand, any dataset can be perfectly represented using $n-1$ dimensions, where n is the number of items scaled. Even if the stress is not 0, a certain amount of distortion is always tolerated according to benchmark thresholds.

1.3.3 Prediction algorithms (*cat*RAPID and CROSSalive)

Predictive algorithms apply machine-learning approaches on validated datasets of experimental data to identify novel outcomes with high accuracy. This provides predicted estimations of biological systems that are difficult to study by experimental methods. For instance, protein-RNA interactions are difficult to study at large-scale since the state of the art technology requires the purification of a cross-linked protein with high quality, which is not straightforward under current methods. In this sense, I briefly describe here the two main algorithms employed in this thesis: *cat*RAPID omics and CROSS alive.

*cat*RAPID omics (http://s.tartaglialab.com/page/catrapid_omics_group) enables to predict protein-RNA interactions on a large-scale by considering both physico-chemical features of the primary sequences interrogated as well the presence of motifs and RNA-binding domains.

CROSS alive (http://service.tartaglialab.com/new_submission/crossalive) is optimize to predict changes on the RNA secondary structure due to post-transcriptional modifications, mimicking the *in vivo* structure of RNA molecules. The algorithm is trained on icSHAPE data on presence and absence of N6 methyladenosine modification (m6a+ and m6a- respectively).

Chapter 2. *In vitro* and *in-silico* methods for determining the RNA specificity on RBP binding

Protein-RNA interactions are crucial for granule structure. In this sense, several properties of the RNA molecules may influence its binding with proteins and therefore, alter granule formation. However, the specific determinants that enable the binding to proteins are poorly understood. In this sense, I present here a review detailing the general *in vitro* and *in silico* methods to detect the contributions of different RNA properties influencing its protein binding. My collaborators and me covered this topic not only examining granule components but also any protein-RNA interaction in the cell, since their contacts are governed by common physico-chemical principles.

We focused on the RNA since it has been less studied than its protein counterpart. While RBPs have precisely identified RNA binding motifs (like the KH or RRM domains), the RNA motifs that they bind very noisy and overrepresented on random sequences. We described the main methods trying to identify all the sources of specificity on the RNA binding to proteins, including the influence of its structure both locally and globally. We also explained some differences between *in vitro* and *in vivo* methods together with the contributions produced by post-transcriptional modifications or cell compartmentalization.

Cid-Samper F., Dasti A., Bechara E., Tartaglia, G. G. (2019). RNA-centric Approaches to Study RNA-Protein Interactions *in vitro* and *in silico*. *Methods. Accepted for publication.*

Dasti A, Cid-Samper F, Bechara E, Tartaglia GG. RNA-centric approaches to study RNA-protein interactions in vitro and in silico. Methods. 2020 Jun 1;178:11-18. DOI: 10.1016/j.ymeth.2019.09.011

Chapter 3. Objectives

The central aims of the present thesis are summarized as follows:

- **To understand the specific properties of granule RNA in relation to non-granule RNA.** Specifically, to investigate how different is the protein-RNA network for the case of granule components.

- **To assess the implications of the properties of granule RNAs at biological level.** We plan to analyze if any distinct properties of granule RNAs may indicate some hints of their function or how they are recruited into the granule.

- **To explore the role of the RNA in the appearance of certain neurodegenerative diseases.** Since some of these diseases are linked with mutations in certain mRNAs, we want to associate some granule RNA properties to their pathogenesis.

- **To employ those properties to build a model able to estimate the propensity of a certain RNA to be granule-forming.** We hypothesize that any RNA molecule whose characteristics are similar to those observed for the majority of the granule-RNAs, is highly likely to be as well granule-prone. We are interested on detecting RNA molecules with highest granule-propensity in the transcriptome, employing distinct granule features.

- **To integrate protein-protein, protein-RNA and RNA-RNA interactions.** SG are composed mainly of proteins and RNA. These molecules are able to form and maintain the granule structure through different modes of interaction. We want to decipher how these different kinds of interactions are related to each other.

- **To study the link between formation, structure and function in SG.** In biology, structure correlates with function. In this sense, we believe that understanding the internal structure of SG (e.g. the topology of the protein-RNA network and the similarities between the protein-protein network), we will understand better their functionality. In a similar extent, the structure of a mature SG may point towards a model of their formation compatible with the final composition.

- **To apply network and clustering methods to address the points described above.** Since SG are sustained thought multiple interactions among different molecule types, we think that network properties might reflect fundamental characteristics of the SG. Similarly, we will apply clustering methods to find groups of densely connected elements in the SG network.

II. RESULTS

Chapter 4. A model for determining scaffolding RNAs and their relation with FXTAS disease

In this Chapter, I present a work published on Cell Reports. In this project, we employed computational methods such as network analysis and clustering algorithms on publicly available experimental data. This analysis produced a model that integrates several distinct properties of RNAs molecules (such as expression, structure and length) to identify RNAs highly prone to be granule-forming. All these properties are related with a higher propensity to interact with proteins, and therefore with the ability to act as scaffolding molecules.

We applied this model on the study of the Fragile X Tremor/Ataxia Syndrome (FXTAS). In FXTAS, a 5´-microsatellite expansion of a CGG repetition on the FMR1 produces an increase in the scaffolding ability of the mRNA of the gene. This abnormal increase sequesters some important proteins into nuclear RNP granules, such as TRA2A, impeding their normal function and producing some symptoms associated with the progress of the disease. This novel experimental data also corroborates other postulates of our model, such as the inefficiency of protein-protein interactions to recruit elements to the granule.

Cid-Samper, F., Gelabert-Baldrich, M., Lang, B., Lorenzo-Gotor, N., Ponti, R. D., Severijnen, L. A. W., ... & Tartaglia, G. G. (2018). An Integrative Study of Protein-RNA Condensates Identifies Scaffolding RNAs and Reveals Players in Fragile X-Associated Tremor/Ataxia Syndrome. ***Cell reports*****,** ***25*****(12), 3422-3434. doi: https://doi.org/10.1016/j.celrep.2018.11.076 Epub 2018 Dec 18. PMID: 30566867**

Cid-Samper F, Gelabert-Baldrich M, Lang B, Lorenzo-Gotor N, Ponti RD, Severijnen LAWFM, et al. An Integrative Study of Protein-RNA Condensates Identifies Scaffolding RNAs and Reveals Players in Fragile X-Associated Tremor/Ataxia Syndrome. Cell Rep. 2018 Dec 18;25(12):3422-3434.e7. DOI: 10.1016/j.celrep.2018.11.076

Chapter 5. RNA implication on granule structure: a hypothesis for SG formation

During the development of the study presented above, several groups published new lists of proteins and RNAs contained in SG. We decided to expand our first model by integrating these newly available datasets. In this sense, we dealt with differences on SG composition under different cell types and stresses by defining a consensus SG proteome, whose components are present under most conditions. We observed that this consensus SG proteome is densely connected through protein-RNA interactions whereas did not have any special connectivity in terms of protein-protein interactions. This observation corroborated our previous model that confers importance on the RNA components of a RNP and particularly to protein-RNA interactions.

However, the first published SG transcriptome revealed a significant amount of RNAs without any known protein interaction, suggesting unrevealed strategies for the recruitment of those elements into the granule. In this sense, we studied the importance of RNA-RNA on the SG structure. SG RNA-RNA interaction network showed similar properties that those observed on the granule protein-RNA network. Finally, we explored the role of post-transcriptional modifications on modulating the structure of certain RNAs and their implications for SG formation and organization.

Cid-Samper F. Vandelli A., Sanchez de Groot N. & Tartaglia, G. G. (2019). Stress granules network analysis reveals a RNA scaffolding population. ***Submitted to Molecular Systems Biology.***

REPORT for Molecular Systems Biology

Network analysis of stress granules reveals a RNA scaffolding population

Fernando Cid-Samper[1,2,‡], Andrea Vandelli[1,2,‡], Natalia Sanchez de Groot[1,*] and Gian Gaetano Tartaglia[1,2,3,*]

1 Centre for Genomic Regulation (CRG), The Barcelona Institute for Science and Technology, Dr. Aiguader 88, 08003 Barcelona, Spain
2 Universitat Pompeu Fabra (UPF), 08003 Barcelona, Spain
3 Institució Catalana de Recerca i Estudis Avançats (ICREA), 23 Passeig Lluís Companys, 08010 Barcelona, Spain
‡ Both authors contributed equally.
* To whom correspondence should be addressed:
NSG: natalia.sanchez@crg.eu; GGT: gian.tartaglia@crg.eu or gian@tartaglialab.com Tel: +34 93 316 01 16; Fax: +34 93 396 99 83;

ABSTRACT

Membrane-less organelles organize the activities of both proteins and RNAs in the cell. Often these assemblies are formed through a process known as phase separation that allows the spatio-temporal isolation of biochemical reactions. One of the best-characterized membrane-less compartments is the stress granule that regulates transcripts metabolism upon physical and chemical insults. At present it is still unclear how the granule is organized at the molecular level. We addressed this question by analysing proteins and RNAs interaction networks. We found that specific RNAs can act as scaffolds of the granule and there are specific transcripts able to promote both RNA-RNA and RNA-protein interactions within the assembly. Our work supports the vision change that RNA is much more than a mere intermediate in the transmission of genetic information from DNA to proteins. By shading light on the inner complexity of granules we hope to open new avenues to understand their formation, organization and function.

INTRODUCTION

Inside the cell, biochemical reactions should be optimized and regulated to happen at the required moment and without disturbing the other cellular processes. To achieve this goal, the cell is organized in different organelles and compartments that could be isolated with membranes or by physical segregation[1]. Interestingly, the study of these membrane-less organelles is one of the most prolific and promising research lines in cell biology. At present, more than 22 types of membrane-less organelles have been described[2], all of them with different composition and biophysical properties. They are assemblies formed thanks to the phase separation capacity of their different components, which usually interact with each other in a multivalent manner (i.e. one molecule establishes several interactions). One of the most valuable properties associated to this multivalence is their dynamicity, which allows to exchange components with the surrounding environment, change their composition or disassembly when required[3].

Stress granules (SG) are one of the most studied biological condensates. Their formation and composition depends on the cellular state (i.e. stress conditions) and they are involved in regulating the availability and half-life of a large number of transcripts[4]. As a result, they are mainly composed of proteins and RNAs. Many stress granules components are associated to diseases. For instance, amyotrophic lateral sclerosis (ALS) patients present a mutated form of the TDP-43 protein that leads to its accumulation into stress granules[5]. The analysis of SG shows the existence of some constitutive components that are critical for its formation coupled with the presence of other elements that are specific of the environmental conditions. In addition, SG contain semi-stable solid-like cores surrounded by a more dynamic liquid-like shell that coexist in a proportional equilibrium[3] (**Figure 1a**).

Even though SG proteins have been largely studied, only recently the first systematic transcriptome analysis was published [6]. This study found that SG contain mRNAs from essentially every expressed gene, however, no single RNA represents more than 1% of the SG RNA molecules[6]. Intriguing, there is just a small bias in the binding of SG proteins to mRNAs enriched in SG, moderating the dominant role attributed to the SGs' RNA binding proteins (RBPs) and opening the question about how the mRNAs locate to SGs[7]. At this context, we aim to analyse, thought network analysis tools, the interactions sustaining the SGs with especial interest on those involving RNA molecules.

Network analysis is a powerful tool used to interpret the existent experimental data. In this sense, it allows to detect properties that are invisible if one studies the elements of the network individually[8]. We have recently applied network analysis at the biological condensates field to screen for transcripts with the ability to scaffold ribonucleoprotein (RNP) assemblies and, consequently, with potential to be involved in human diseases and to be therapeutic targets[9]. Specifically, this analysis showed that highly contacted RNAs are structured, have long UTRs, and contain nucleotide repeat expansions. In this way, we discovered that the expansion of CGG repeats on the *FMR1* transcript increases its scaffolding abilities and is implicated in fragile X-associated tremor/ataxia syndrome (FXTAS). We validated that *FMR1* CGG expansion affects its interaction with TRA2A and that both (FMR1-TRA2A) co-aggregate in mouse model and in post-mortem human samples, demonstrating their phase separation and disease triggering capacities[9].

Here we show, through network analysis, how the main components of the SG interact with each other and how these connections define the SG arrangement. Overall, these analyses show that transcripts have a crucial role in arranging the SG proteins but, more interestingly, we revealed the pre-existence of a dense, virtually independent and RNA-RNA network containing a core of RNAs optimized to enhance both RNA-RNA and RNA-protein interactions. Due the high overlapping observed, the RNA-RNA contacts may serve as a platform where build protein-RNA interactions. Our results reveal a complex SG inner structure and opens new venues to understand their formation, organization and function.

RESULTS AND DISCUSSION

The study of how proteins and RNA interact between each other, and later how these contacts lead to the formation of biological condensates has been closely related with the study of SGs. These RNP assemblies have been systematically analysed through numerous strategies and under numerous conditions. In this way, it is currently known its composition at different cell types, organisms and environmental conditions[10] (**Figure 1b, Supplementary table 1**). It is worth to note that, under stress conditions, once the SG is build, it is composed by two phases: (i) a dynamic liquid-like shell that surrounds (ii) several small solid-like cores[3].

Despite it is know which RNAs and proteins are located in the SG, it is not understood how these elements interact between each other and how they are organised (**Figure 1c**). This information can be crucial to understand their function and their relationship with human disorders. To unravel the contact network that sustains the SG we analysed all the different types of interactions that can occur in these assemblies, regarding to the molecules involved: protein-protein (PPI), protein-RNA (PRI) and RNA-RNA (RRI) (**Figure 1c**).

Protein-RNA interactions as a central cohesive force

We first analysed how proteins are organised in the SG. Due to the high amount of data reported and the discrepancies between different techniques and cell types, we focused our analyses on three different sets containing proteins specifically identified at the SG: 144 proteins defined by proximity-based proteomics[11], 300 protein detected by APEX proximity levelling[10] and 411 proteins detected in the SG core[6] (**Figure 1b, Supplementary table 1**). Accordingly, we defined a set of consensus SG proteins containing those identified by at least two different techniques (**Figure 1d, Supplementary table 1**). Then we look how the PPI and PRI attract and organise these consensus SG proteins.

For the PPI interactions we employed those deposited on BIOGRID v.3.4, (https://thebiogrid.org/), as physical interactions. For the PRI interactions we employed all the interactions reported by eCLIP (including a total of **93** proteins) and the 13838 RNAs detected by eCLIP[12] (**Figure 1e**, **see Methods**). For each clustering analysis we measured: (i) proteins grouped by the number of PPI shared (those that bind similar proteins) (**Figure 1f**) and (ii) proteins grouped by the number of PRI shared (those that bind similar RNAs) (**Figure 1g**, see **Methods**).

The clustering analysis shows that the PPIs are able to cluster the consensus SG proteins (6 proteins, **Figure 1f, p-value = 0.007 , Fisher's test**). However, the PRIs clustered much better a higher number of proteins (10 proteins, **Figure 1g**, p-value = 4.8e-6, Fisher's test). This result suggests that the proteins by themselves (i.e. just a network of PPIs), are not able to gather each other to build the SG consensus core and supports the hypothesis that the presence of RNA is required to assemble a SG[7].

Our data also confirms that the SG consensus proteins have a central role on recruiting and retaining transcripts: "transcription process regulation" (**Figure 1d**). In agreement, just 17 SG consensus proteins clustered by the PRIs interact with 90% of the eCLIP transcriptome, an amount significantly larger than any other group of proteins of the same size (**Figure 1e**, p-value < 1e-5). In addition, this data, based on impartial high-throughput analyses, agrees with the recently published observation about that nearly any RNA can eventually be part of the SG[6]. These data also suggest that the cell evolutionarily selected just a small set of RBPs (the SG consensus) to recruit, when required, nearly any RNA at the SGs.

SG enriched RNAs act as organizers

Despite that SGs have been typically defined by their protein composition[13] , the network analysis presented above indicates that the presence of RNA is also important to put together these proteins. However, the rationale behind this it is not clear since each consensus SG protein can be forming contacts within a set of around 5000 putative RNA partners[12]. So, how a promiscuous interaction network can help to put together 17 specific proteins to build a stress granule? Unfortunately, due to the late publication of a SG transcriptome, there is still no information about how the interactions between these RNAs influence the SG arrangement.

Thus, with aim to understand better the transcripts located in the SGs[6] (**Figure 2a**), we analysed this network of interactions. The SG transcriptome is composed of transcripts that have been found by purifying the solid cores (**Figure 1a**). This set contains more than 70% of the human transcriptome and has been divided in three groups (depleted, neither, enriched) regarding to the RNA abundance with respect to the rest of the cell (**Figure 2a**). Accordingly, here we consider that the most consistent SG RNA set is the enriched one and, thus, we use its interaction network as the one that may occur in the SGs.

Consistently with the previous section, we analysed how the PRI organise the RNAs present in the SG. The 2D representation (obtained by applying multidimensional scaling algorithms, **see Methods**) of the number of PRI shared (those RNAs that bind similar proteins), shows a network in which the proteins do not specially organize the RNAs neither the SG enriched ones (**Figure 2b-c**). This suggests that the protein interaction with the RNAs present in the SGs is promiscuous and low specific. In addition, these data also support that, from a network point of view, the main characteristic differentiating the consensus SG proteins from the other proteins (**Figure 1g**) is their ability to bind loads of different RNAs.

We analysed how PRI contacts allow to group consensus SG proteins together and we especially studied if among different RNAs there is a set that better clusters these proteins. To achieve this, we focused on three RNA sets with different SGs concentration. When we represent the 2D distribution of the PRI contacts we obtain better clustering power for the enriched SG transcripts (**Figure 2d-f**, p-value [enriched] = 8e-6, p-value [neither] = 0.0006, p-value [depleted] = 0.007, Fisher´s test), which is linked with a higher number of protein contacts (**Figure 2g**). Although, enriched transcripts have on average more proteins contacts than neither or depleted ones **Figure 2g,** p-value = 6.7e-41, p-value = 2.2.e-05 respectively, Wilcoxon rank-sum test).), it is important to note that there is a large amount of enriched RNAs with no reported protein contacts (**Figure 2h**), which agrees with the small bias in the RBP binding previously measured by Khong and co-workers[6]. Overall, these results support that the role of the RNA in the organization of the SG is more important than previously thought and points that many transcripts should be attracted to the SG not only by proteins but also though RNA-RNA interactions[7,14]

SGs are sustained by highly contacted transcripts

To measure how the RNA-RNA contact network influences the SG arrangement, we looked at the number and density of contacts between the enriched and depleted transcripts. We first observed that the enriched transcripts have a higher number of RNA-RNA interactions (**Figure 3a**, p-value < 2.2e-16 , Wilcoxon rank-sum test). Moreover, these transcripts are more closely connected and showed lower average shortest path (i.e. lower average distance among elements on the network, **see Methods**) and jaccard distance (i.e. more RNA contact sharing, **see Methods**) in comparison to the depleted ones (**Figure 3b-c**, p-value < 2.2e-16, Wilcoxon rank-sum test and p-value < 2.2e-16, Kolmogorov-Smirnov test respectively, Wilcoxon rank-sum test). This observation agrees with what was recently observed *in vivo* and *in vitro* studies: RNAs from the SG

cores are prone to interact each other[6,15]. In a similar extent, enriched transcripts have high betweenness (**Figure 3e** p-value < 2.2e-16, Wilcoxon rank-sum test) in the global RNA-RNA network, indicating than any transcript in the cell can reach them by few interaction steps (**see Methods**). Despite that the enriched transcripts are highly connected with all the RNAs, they also present a lower ratio of self-contacts when compared to the depleted RNAs, which means that they usually contact with many non-enriched RNAs (**Figure 3f**, p-value < 2.2e-16, Kolmogorov-Smirnov test). These results point toward a model where the enriched transcripts build a highly contacted and dense RNA-RNA network able to attract other RNA and protein components.

The characteristics shared by the enriched transcripts (high protein and RNA contacts, act as a cohesive force) resemble those associated to the scaffolding RNAs[9,16]. These RNAs are characterised by a high structured content, which has been previously associated with the number of protein interactions[9,17]. However, despite of being highly contacted by proteins (**Figure 2g**), the SG enriched RNAs are less structured than depleted ones (**Figure 3g**, p-value = 0.0042, Wilcoxon rank-sum test). Yet, it must be noted that the N6 methyladenosilation (m6A+) modification is involved in the regulation of phase separation, specifically it enhances the granule formation by reducing the amount of secondary structure[18]. To analyse how the presence or absence of m6A can affect to the structure of the SG transcripts we employed CROSSalive. This algorithm predicts the changes on the RNA secondary structure due to post-transcriptional modifications. In this way. the analysis on the 200 most enriched versus the 200 most depleted transcripts shows that the m6a+ modification inverts the structural differences between the two RNA sets, becoming the enriched one the more structured (**Figure 3h,** p-value = 1.26e-5 , Wilcoxon rank-sum test). Thus, the effect of the RNA modifications may be in charge of the enriched SG RNAs higher interactivity (**Figure 2g**).

Overall, our data points that the enriched SG RNAs may be developing a scaffolding role. Based on this observation we decided to analyse the scaffolding propensity of the different RNAs found in the SGs. Between the top scaffolding RNAs we found important known lncRNAs such as NEAT1 and MALAT1, that are involved in nuclear paraspeckles (**Supplementary Figure 1)**. Looking for RNAs enriched in SG, we found NORAD among the top scaffolding candidates on eCLIP (**Figure 3i-l)**. Interestingly, looking at the sequence of these powerful scaffolding RNAs we observed a high correlation between their protein-RNA and RNA-RNA binding sites (NORAD r = 0.67, Pearson´s; NEAT1 r = 0.41, MALA1 r = 0.46, Pearsons´s, **Figure 3m, Supplementary Figure 1**). RNA structure is supposed to provide the special characteristics that ensure specificity to the binding site. However, proteins could not differentiate if this secondary structure is built through a single RNA chain (intra-molecular base-pairing) or different RNA chains (inter-molecular base-pairing). Hence, the overlapping between binding sites may be indicating that the RNA-RNA interactions build by the scaffolding RNAs regulate and promote the interaction with proteins.

SG pre-existing interaction network

Our RNA-RNA interaction data is based on RISE database[19], which collects the multiple unrelated experiments performed mainly under no stress conditions. As other studies previously suggested, this fact indicates the existence of a pre-existing network of interactions that would eventually form the granule. These observations point that the cell is equipped with a set of proteins and RNAs designed for, when required, find each other in a rapidly. The protein set is small but highly promiscuous (**Figure 1e**) favouring its attachment to places enriched in RNA. The enriched SG RNAs are highly expressed[9], **low translated**[6] and very prone to form contacts between each other (**Figure 3f**). Importantly, all these properties exist before any stress condition appear. When the stress comes the set of RNAs are rapidly released from the ribosome and, due to their dense contact network,

they can easily find each other. This can favour the assembly of local high RNA concentrations that can attract the promiscuous SG proteins, which in turn can drag a diverse number of transcripts.

By the integration of different previous studies, we have described a consensus SG proteome that would be recruited thought protein-RNA interactions. This points towards a common function of the stress granules regardless of the specific conditions or cell types where they are formed. Of course, different compositions under different stresses will led to slightly different specialized functions[10], but the presence of the consensus proteome points towards a central common function. Overall, our model supports the recent change of paradigm in the field that highlights the role of the RNA in the SGs[7,15]. We provide a further insight into the internal organization of RNA networks into the SGs. We expect that our model will help on the experimental identification of the core components of the SG.

REFERENCES

1. Boeynaems, S. *et al.* Protein Phase Separation: A New Phase in Cell Biology. *Trends in Cell Biology* **28**, 420–435 (2018).

2. Banani, S. F., Lee, H. O., Hyman, A. A. & Rosen, M. K. Biomolecular condensates: organizers of cellular biochemistry. *Nat Rev Mol Cell Biol* **18**, 285–298 (2017).

3. Jain, S. *et al.* ATPase-Modulated Stress Granules Contain a Diverse Proteome and Substructure. *Cell* **164**, 487–498 (2016).

4. Protter, D. S. W. & Parker, R. Principles and Properties of Stress Granules. *Trends in Cell Biology* **26**, 668–679 (2016).

5. Fang, M. Y. *et al.* Small-Molecule Modulation of TDP-43 Recruitment to Stress Granules Prevents Persistent TDP-43 Accumulation in ALS/FTD. *Neuron* **103**, 802-819.e11 (2019).

6. Khong, A. *et al.* The Stress Granule Transcriptome Reveals Principles of mRNA Accumulation in Stress Granules. *Molecular Cell* **68**, 808-820.e5 (2017).

7. Van Treeck, B. & Parker, R. Emerging Roles for Intermolecular RNA-RNA Interactions in RNP Assemblies. *Cell* **174**, 791–802 (2018).

8. Barabási, A.-L. & Oltvai, Z. N. Network biology: understanding the cell's functional organization. *Nat Rev Genet* **5**, 101–113 (2004).

9. Cid-Samper, F. *et al.* An Integrative Study of Protein-RNA Condensates Identifies Scaffolding RNAs and Reveals Players in Fragile X-Associated Tremor/Ataxia Syndrome. *Cell Reports* **25**, 3422-3434.e7 (2018).

10. Markmiller, S. *et al.* Context-Dependent and Disease-Specific Diversity in Protein Interactions within Stress Granules. *Cell* **172**, 590-604.e13 (2018).

11. Youn, J.-Y. *et al.* High-Density Proximity Mapping Reveals the Subcellular Organization of mRNA-Associated Granules and Bodies. *Molecular Cell* **69**, 517-532.e11 (2018).

12. Van Nostrand, E. L. *et al.* Robust transcriptome-wide discovery of RNA-binding protein binding sites with enhanced CLIP (eCLIP). *Nat Methods* **13**, 508–514 (2016).

13. Lin, Y., Protter, D. S. W., Rosen, M. K. & Parker, R. Formation and Maturation of Phase-Separated Liquid Droplets by RNA-Binding Proteins. *Molecular Cell* **60**, 208–219 (2015).

14. Boeynaems, S. *et al.* Spontaneous driving forces give rise to protein−RNA condensates with coexisting phases and complex material properties. *PNAS* **116**, 7889–7898 (2019).

15. Treeck, B. V. *et al.* RNA self-assembly contributes to stress granule formation and defining the stress granule transcriptome. *PNAS* **115**, 2734–2739 (2018).

16. Chujo, T., Yamazaki, T. & Hirose, T. Architectural RNAs (arcRNAs): A class of long noncoding RNAs that function as the scaffold of nuclear bodies. *Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms* **1859**, 139–146 (2016).

17. Sanchez de Groot, N. *et al.* RNA structure drives interaction with proteins. *Nat Commun* **10**, 3246 (2019).

18. Ries, R. J. *et al.* m6A enhances the phase separation potential of mRNA. *Nature* **571**, 424–428 (2019).

19. Gong, J. *et al.* RISE: a database of RNA interactome from sequencing experiments. *Nucleic Acids Res* **46**, D194–D201 (2018).

FIGURE CAPTIONS

Figure 1. Consensus stress granule proteome. a) Scheme of SG internal organization. Stress granules are mainly composed by proteins and RNA. They contain several more stable solid-like cores surrounded by a more liquid-like and dynamic shell. **b) Set of consensus granule proteins present on eCLIP data.** We considered three different studies reporting "core" granule proteins for defining a set of consensus granules proteins. Consensus proteins were those reported by at least 2 different studies (17 proteins). **c) Interaction modes.** Protein and RNA can either interact with each other or self-interact, generating protein-protein, protein-RNA and RNA-RNA networks. These networks contain specific topologies that define the SG organization. **d) GO enrichment on transcripts targeted by consensus SG proteins**. Regulatory RNAs involved in transcription process regulation are the most common targets of most of the consensus SG proteins. **e) Total number of RNA targets of consensus SG proteins.** Most of the total number of RNAs detected on eCLIP (90%) interact with at least one consensus granule protein. **f-g) Clustering of consensus granule proteins according to their jaccard distance on the PPI (f) and PRI (g).** Multidimensional-scaling projection and hierarchical clustering analysis of the consensus granule proteins out of the total number of proteins on eCLIP. Consensus granule proteins cluster better by protein-RNA interaction than by protein-protein.

Figure 2. Protein-RNA interactions on SG enriched transcripts. a) Different kinds of transcripts according to their concentration on the SG. Khong et. Al, 2017 describe a set of transcripts that are enriched on the SG (more concentrated than in the cytosol), depleted on the SG (less concentration that in the cytosol) or neither (no significant concentration on the SG nor the cytosol). We considered enriched transcripts as the highest granule-prone RNA molecules. Both eCLIP and SG transcriptome cover a high fraction of the total transcriptome of K562 cells (~70%). **b-c) Clustering of SG transcripts (b) and only enriched SG transcripts (c) according to their jaccard distance on the PRI.** Neither all the RNAs nor the SG enriched show any distinct organization by their sharing of protein contacts. While RNA has a role on recruiting granule proteins by protein-RNA contacts, proteins do not seem to have the same function on organizing RNA into the granule. **d-f) Clustering of consensus granule proteins according to their jaccard distance on the PRI considering only enriched (d), depleted (e) or neither (f) transcripts.** Enriched transcripts show the highest ability for recruiting the consensus SG proteins. **g) Protein contacts of SG transcripts by kind**. Enriched transcripts have on average a higher number of protein contacts than depleted or neither transcripts. **h) Protein contacts distribution of enriched transcripts.** 17% of enriched transcripts transcripts do not have any known protein

interactions. This suggest the existence of alternative pathways of enriched transcripts into SG rather than protein-RNA interactions.

Figure 3. Properties of the SG RNA-RNA interaction network. a-f) RNA-RNA network properties comparison between enriched and depleted transcripts. SG enriched transcripts have **(a)** a higher number of RNA-RNA contacts, **(b)** lower average shortest path, **(c)** higher jaccard index on their RNA-RNA targets and **(d)** higher betweenness. All these properties together indicate that enriched transcripts interact though a highly dense network of RNA-RNA interaction. However, they are also able to reach transcripts that are not enriched in SG as indicated by their **(f)** lower number of self RNA-RNA contacts (i.e. contacts with transcripts of the same kind, enriched or depleted, normalized by the total number of transcripts in the group. **g-h) Changes on structural content of SG transcripts due to methylation**. While depleted transcripts seem more structured *in vitro* than the enriched ones **(g)**, this trend is inverted when considering m6A+ modification on the RNA **(h)**. **i-l) NORAD scaffolding properties**. In red, position of NORAD on the eCLIP distribution of the main properties related with scaffolding propensity: **i)** protein contacts, **j)** RNA contacts, **k)** structural content (PARS), **l)** transcript length. **m) NORAD binding profiles**. Protein-protein and protein-RNA binding profile according to eCLIP and RISE databases respectively. Nucleotide positions correspond to the average window profiling method (see **Methods**).

Supplementary Table 1. List of consensus granule proteins. We defined as consensus granule proteins (highlighted in yellow) as those detected as "core" SG proteins by at least two different studies.

Supplementary Figure 1. NEAT1 and MALAT1 binding profiles. Protein-protein and protein-RNA binding profile according to eCLIP and RISE databases respectively. Nucleotide positions correspond to the average window profiling method (see **Methods**).

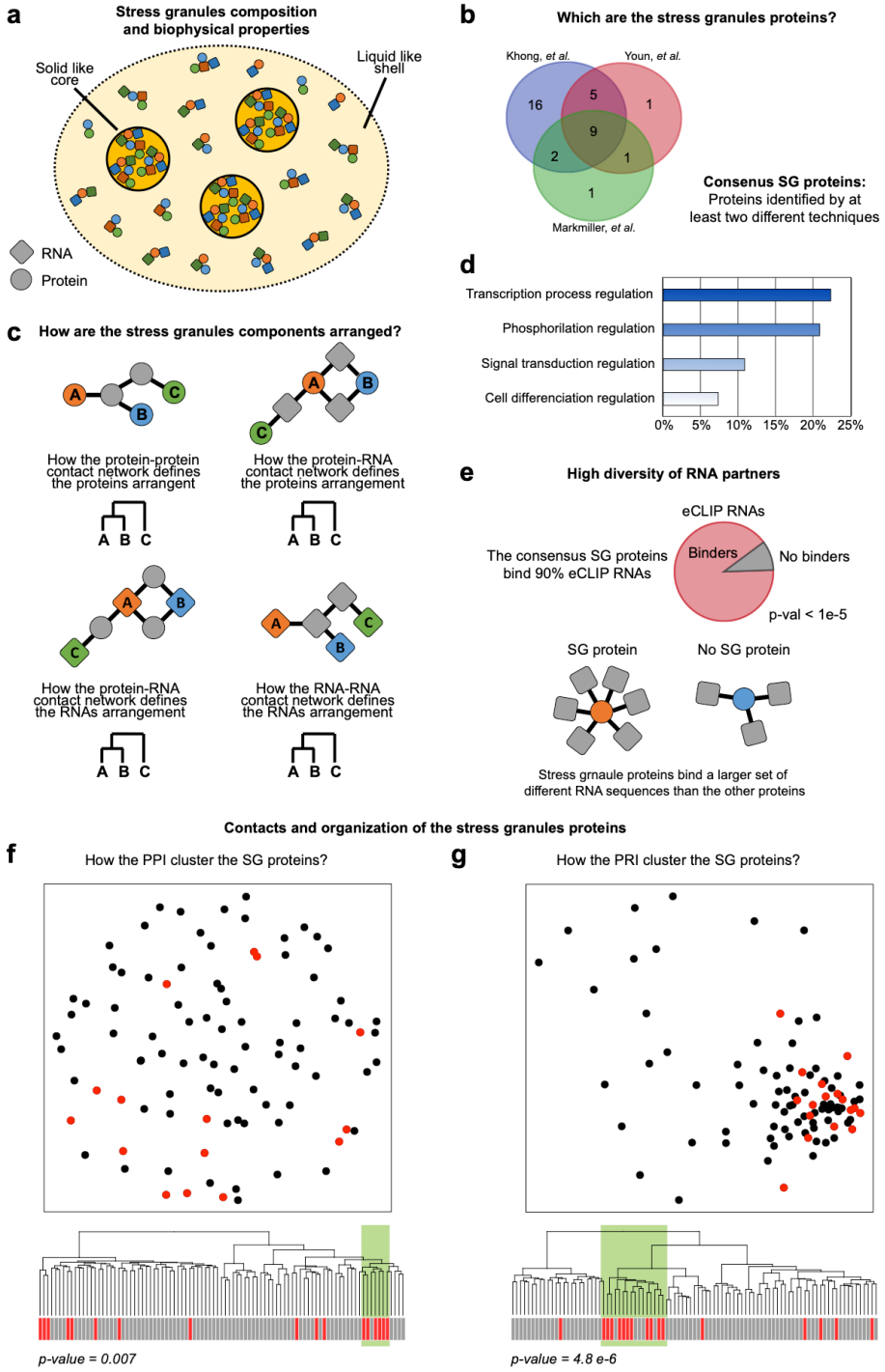

a
Stress granules composition and biophysical properties
Solid like core
Liquid like shell
RNA
Protein
b
Which are the stress granules proteins?
Khong, et al.
Youn, et al.
Markmiller, et al.
16
5
1
9
2
1
1
Consenus SG proteins:
Proteins identified by at least two different techniques
c
How are the stress granules components arranged?
How the protein-protein contact network defines the proteins arrangent
How the protein-RNA contact network defines the proteins arrangement
How the protein-RNA contact network defines the RNAs arrangement
How the RNA-RNA contact network defines the RNAs arrangement
A B C
d
Transcription process regulation
Phosphorilation regulation
Signal transduction regulation
Cell differenciation regulation
0% 5% 10% 15% 20% 25%
e
High diversity of RNA partners
eCLIP RNAs
The consensus SG proteins bind 90% eCLIP RNAs
Binders
No binders
p-val < 1e-5
SG protein
No SG protein
Stress grnaule proteins bind a larger set of different RNA sequences than the other proteins
Contacts and organization of the stress granules proteins
f
How the PPI cluster the SG proteins?
p-value = 0.007
g
How the PRI cluster the SG proteins?
p-value = 4.8 e-6

Figure 1

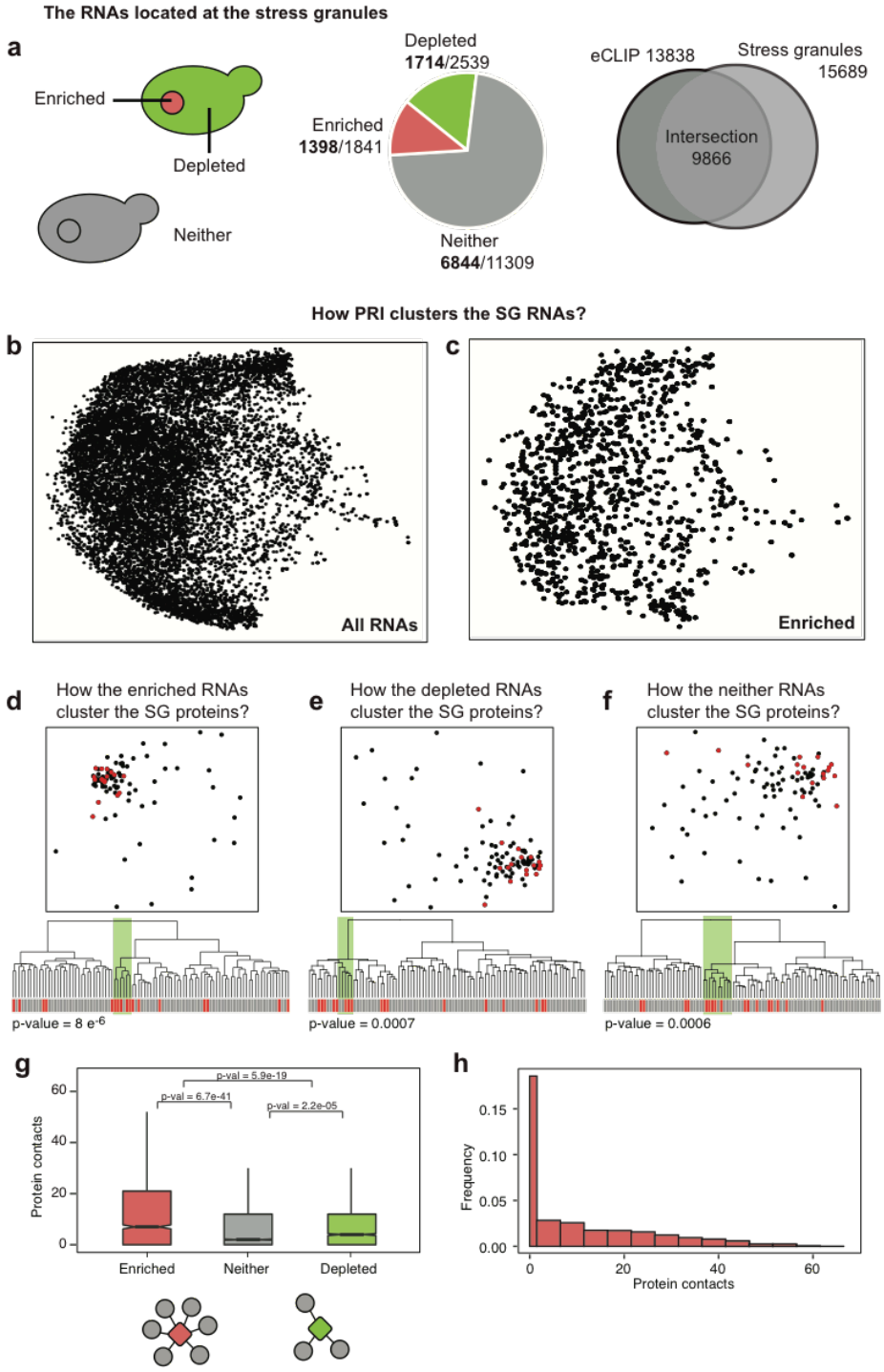

The RNAs located at the stress granules
a
Enriched
Depleted
Neither
Depleted
1714/2539
Enriched
1398/1841
Neither
6844/11309
eCLIP 13838
Stress granules
15689
Intersection
9866
How PRI clusters the SG RNAs?
b
All RNAs
c
Enriched
d
How the enriched RNAs cluster the SG proteins?
e
How the depleted RNAs cluster the SG proteins?
f
How the neither RNAs cluster the SG proteins?
p-value = 8 e-6
p-value = 0.0007
p-value = 0.0006
g
p-val = 5.9e-19
p-val = 6.7e-41
p-val = 2.2e-05
Protein contacts
Enriched
Neither
Depleted
h
Frequency
Protein contacts

Figure 2

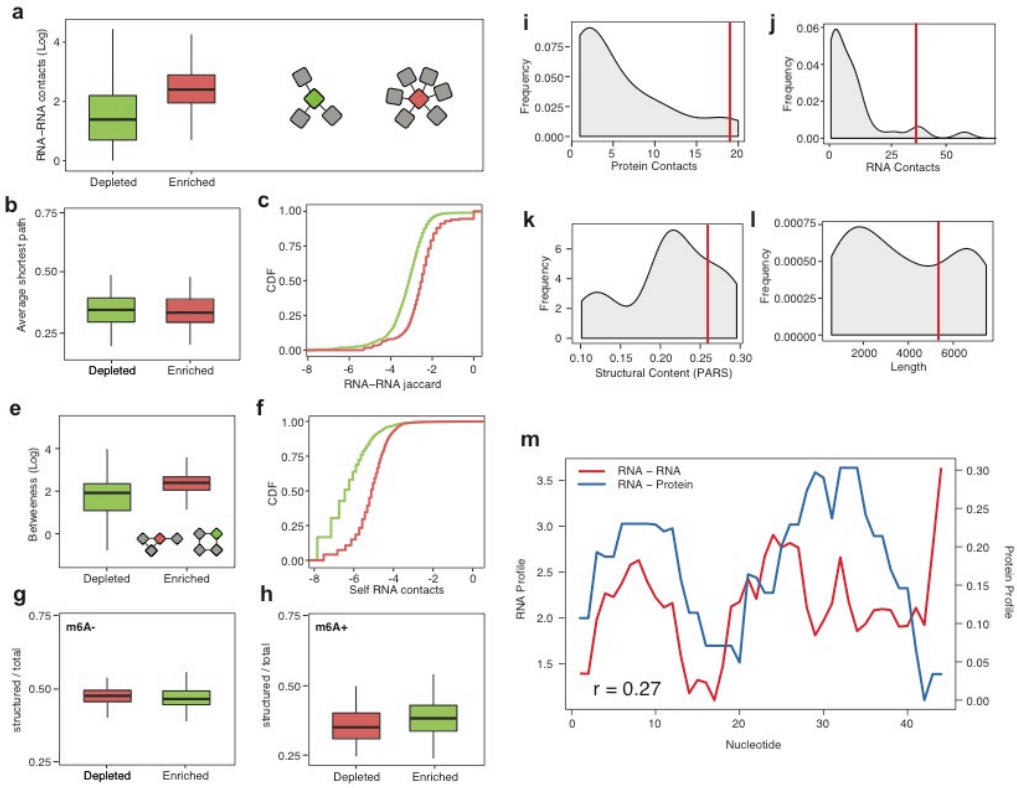
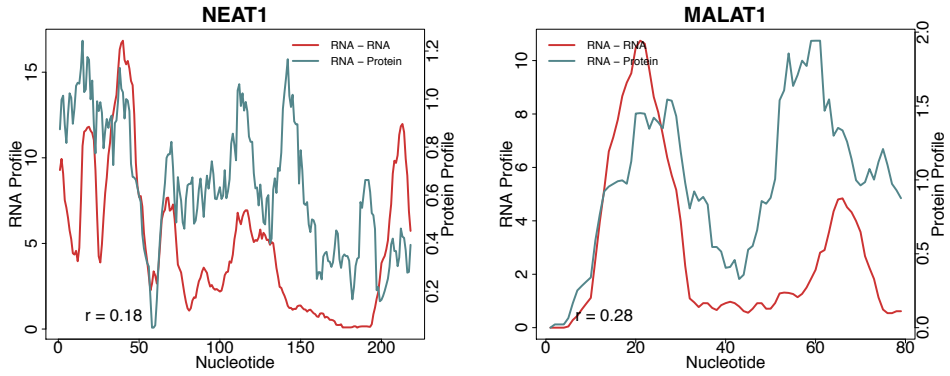

a
RNA–RNA contacts (Log)
Depleted
Enriched
b
Average shortest path
Depleted
Enriched
c
CDF
RNA–RNA jaccard
e
Betweeness (Log)
Depleted
Enriched
f
CDF
Self RNA contacts
g
m6A-
structured / total
Depleted
Enriched
h
m6A+
structured / total
Depleted
Enriched
i
Frequency
Protein Contacts
j
Frequency
RNA Contacts
k
Frequency
Structural Content (PARS)
l
Frequency
Length
m
RNA – RNA
RNA – Protein
RNA Profile
Protein Profile
r = 0.27
Nucleotide

Figure 3

PROTEIN	Index	Khong, et al.	Marmiller, et al.	Youn, et al., Cell type	Youn, et al. - Stress type	Counts
CPSF6	1	Y	N	N	N	1
DDX3X	2	Y	Y	Y	Y	4
DDX6	3	Y	Y	N	N	2
DHX30	4	Y	N	N	N	1
DROSHA	5	N	N	Y	Y	2
EIF4G2	6	Y	Y	N	N	2
EWSR1	7	Y	N	N	N	1
FAM120A	8	Y	Y	Y	Y	4
FMR1	9	Y	Y	Y	Y	4
FXR1	10	Y	Y	Y	Y	4
FXR2	11	Y	Y	Y	Y	4
HNRNPA1	12	Y	N	N	N	1
HNRNPK	13	Y	N	N	N	1
HNRNPUL1	14	Y	N	N	N	1
IGF2BP1	15	Y	Y	Y	Y	4
IGF2BP2	16	Y	Y	N	N	2
KHDRBS1	17	Y	N	N	N	1
KHSRP	18	Y	N	N	N	1
LARP4	19	Y	Y	N	N	2
METAP2	20	N	N	Y	Y	2
NONO	21	Y	N	N	N	1
NSUN2	22	Y	N	N	N	1
PUM2	23	Y	Y	Y	Y	4
SAFB2	24	Y	N	N	Y	2
SERBP1	25	Y	N	N	N	1
SND1	26	Y	Y	Y	Y	4
SRSF1	27	Y	N	N	N	1
TAF15	28	Y	N	N	N	1
TARDBP	29	Y	N	N	N	1
TIA1	30	Y	N	N	N	1
TNRC6A	31	N	Y	N	N	1
U2AF1	32	Y	N	N	N	1
UPF1	33	Y	Y	N	N	2
YBX3	34	Y	Y	Y	N	3

Supplementary Table 1



Supplementary Figure 1

METHODS

Data acquisition and composition

Regarding stress granule RNA composition, we employed the transcriptome described by Khong et al., 2019[1]. In order to build to stress granule core proteome, we combined three different studies on the stress granule protein composition: G3P1 pull-down[2], BioID[3] and APEX-proximity labelling[4]. Note, that APEX-proximity labelling was applied to obtain two different sets of proteins (i.e. cell-type constitutive and stress-type constitutive). We consider a protein as a consensus granule component if it was detected as stress-granule forming on at least two out of the four studies analysed.

Protein-protein were taken from the BIOGRID database version 3.4.[5], comprising a total of 1.559.32 interactions. We retrieved protein-RNA interactions from the data deposited on ENCODE corresponding to the eCLIP experiments[6]. For avoiding cell-type biases, we only consider the data available for the K562 cell type. We processed the eCLIP normalizing the number of reads by gene expression[7]. We extracted the data on March, 2018 which consists of a set of 93 proteins and 157263 interactions. RNA-RNA interactions were extracted from RISE database (database of RNA Interactome from Sequencing Experiments)[8]. RISE is a compendium containing both data from targeted studies like RAP-RNA or CLASH as well as transcriptome-wide sequencing-based experiments like PARIS, SPALSH, LIGR-seq and MARIO[8]. In total, it contains 328,811 RNA-RNA interactions among all the different types of RNA species.

Network analysis

Protein-protein, protein-RNA and RNA-RNA networks were defined as a set of nodes (either proteins or RNAs) connected though edges (i.e. biological interactions)[9]. Network measurements were computed employing the igraph package (http://igraph.org) in the R environment (http://www.r-project.org). Shortest path and betweenness were computed by the build-in functions.

The shortest path is the main distance measure in network science for computing the distance between a pair of nodes. It is defined as the minimum number of edges needed to connect a pair of nodes. Betweenness represents a centrality measure and is based on computing the shortest path for every possible pair of nodes in the network. In this sense, betweenness of a node is defined as the fraction of the total shortest paths on the network including this node.

Whereas shortest path performs well at representing long-range distances on the network, it can be inappropriate for small distances due to a lack of

resolution coming from the fact that it is a discrete value. In this sense, all the nodes at a given value of shortest path distance will be equally close according to this variable. However, some of these nodes can be forming a cluster (i.e. they would be "closer" among them) and the shortest path would be able to detect these differences. To avoid this issue, we employed the jaccard index for exploring short-range distances. We defined the jaccard index (J) as the set of common interactors of two given nodes (a and b) of the network (i.e. nodes with connected with a shortest path value of two) normalized by the total number of interactors of the two nodes analysed. Being A and B set of nodes directly connected to the nodes a and b respectively:

$$J_{a,b} = |A \cap B| / |A \cup B|$$

The jaccard index have a possible range of values between 0 and 1, being Ja,b = 1 when there is total overlap of targets between two given nodes and Ja,b = 0 when there is not a single target in common between the nodes a and b. Since we consider "closer" those nodes sharing a higher amount of targets we transformed the jaccard index in order to define the jaccard distance (JD) as JDa,b = 1 - Ja,b.

Clustering methods

Clustering analysis allows to identify groups (i.e. "clusters") of elements with similar properties. Clustering methods require to define a distance measure in order to identify which elements are "close" among them. This distance does not have to be a physical distance but a measure of similarity or functional relationship (e.g. jaccard distance). We employed two kind of algorithms for clustering analysis: hierarchical clustering and multidimensional scaling (MDS).

We performed hierarchical clustering on the set of eCLIP proteins to group them according to their distance defined as the jaccard index in the protein-protein and the protein-rna network. The results of the hierarchical clustering algorithm is a dendrogram (i.e. clustering tree) that visually represents all the distances between every element of the studied set. Dendrogram is build detecting first the pair with the lowest distance. This pair is grouped together and considered as a cluster. Then, the distances between all the elements in the set and this new cluster are recomputed and the process is iterated until the tree is finished. There are several hierarchical clustering methods depending on how to recompute the distances on each iteration on the algorithm. We employed the Ward method for a being a common and robust algorithm. (cite)

MDS is an approach that seeks a configuration in low-dimensional space such that the distances between points in the space match the

(dis)similarities contained in a distance matrix as closely as possible. We used approach called SMACOF (**S**tress Majorization of a **Co**mplicated **F**unction algorithm), an iterative MDS algorithm in which disparities are fixed, then points in the MDS space are moved ($\mathbf{X}t_t \rightarrow \mathbf{X}_{t+1}$) , so that the distances of **X**t+1 minimize the stress function. This is done by an operation that is called iterative majorization and it works replacing iteratively the original, optionally complicated, function f **(**x**)**, by an auxiliary function g**(**x; z**)**, where z is some fixed value.

g(x; **z)** should be simpler to minimize than **f (x)**
f (x) < **g(x**; **z)** (original function must always be smaller than or at most equal to
the auxiliary function)
f (z) = g(z; **z)** (auxiliary function should touch the surface at the so-called supporting point **z**)

The algorithm is available in R as package "smacof" (version 2.0-0). We used the following parameters: dimension = 2, type = ordinal, ties = secondary, verbose = T, other parameters were kept as default.

CROSS alive

CROSS alive is an algorithm that enables to detect changes on secondary structure due to post-transcriptional modifications on the RNA and therefore study changes between *in vivo* and *in vitro* on the RNA structure[10] **(**http://service.tartaglialab.com/new_submission/crossalive**)**. The algorithm is trained on icSHAPE data on presence and absence of N6 methyladenosine modification (m6a+ and m6a- respectively). We employed the algorithm to study the structure of the top 200 most enriched transcripts in stress granules in comparison to the 200 most depleted transcripts[1]. We normalized the profiles with a Z-score transformation (i.e. subtracting the mean signal value and dividing by the standard deviation to each nucleotide value). Then, the structural content of each transcript was defined as the percentage of nucleotides on its sequence with a Z-score higher than zero (i.e. double stranded nucleotides) over the length of the sequence.

Statistical analysis

For computing the significance on how consensus proteins are clustered on the hierarchical clustering analysis, we employed the Fisher's test. Fisher test is an exact test applied on contingency tables for small sample sizes. When comparing non-exponential distributions (data showed in boxplots), we used the Wilcoxon test (also called Mann-Whitney U test). This is a non-parametric test that assesses if there is a statistically significant difference on the means of two compared distributions. For comparing exponential distribution (data showed in cumulative distribution functions),

we used the Kolmogorov-Smirnov´s test, which is as well a non-parametric test that checks for statistical differences on two distributions that cannot be well-estimated by their means.

For computing the correlation of the protein-RNA and RNA profiles in the case of NORAD, NEAT1 and MALAT1, we performed the Pearson correlation test on the average profile. We employed the R function 'rollapply' on the package 'zoo' with the following parameters: wsize = 1000, ssize = 500.

References

1. Khong, A. *et al.* The Stress Granule Transcriptome Reveals Principles of mRNA Accumulation in Stress Granules. *Molecular Cell* **68**, 808-820.e5 (2017).
2. Jain, S. *et al.* ATPase-Modulated Stress Granules Contain a Diverse Proteome and Substructure. *Cell* **164**, 487–498 (2016).
3. Youn, J.-Y. *et al.* High-Density Proximity Mapping Reveals the Subcellular Organization of mRNA-Associated Granules and Bodies. *Molecular Cell* **69**, 517-532.e11 (2018).
4. Markmiller, S. *et al.* Context-Dependent and Disease-Specific Diversity in Protein Interactions within Stress Granules. *Cell* **172**, 590-604.e13 (2018).
5. Oughtred, R. *et al.* The BioGRID interaction database: 2019 update. *Nucleic Acids Research* **47**, D529–D541 (2019).
6. Van Nostrand, E. L. *et al.* Robust transcriptome-wide discovery of RNA-binding protein binding sites with enhanced CLIP (eCLIP). *Nat Methods* **13**, 508–514 (2016).
7. Cirillo, D. *et al.* Quantitative predictions of protein interactions with long noncoding RNAs. *Nat Methods* **14**, 5–6 (2017).
8. Gong, J. *et al.* RISE: a database of RNA interactome from sequencing experiments. *Nucleic Acids Res* **46**, D194–D201 (2018).
9. Barabási, A.-L. & Oltvai, Z. N. Network biology: understanding the cell's functional organization. *Nat Rev Genet* **5**, 101–113 (2004).
10. Delli Ponti, R., Armaos, A., Vandelli, A. & Tartaglia, G. G. CROSSalive: a web server for predicting the in vivo structure of RNA molecules. *Bioinformatics* btz666 (2019). doi:10.1093/bioinformatics/btz666

III. DISCUSSION

Chapter 6. Summarizing discussion

The main objective of my thesis was to investigate the distinct properties of granule RNA and their implications for granule functional organization. To do so, we started by studying differences of interactions present in granule RNAs. We found that granule RNAs show a higher dense and central protein-RNA network in comparison to non-granule RNAs. We did not observe a similar trend when analysing protein-protein networks. Indeed, granule RNAs are shared between different granule proteins, have a higher number of protein contacts, and show increased betweenness, closeness and degree of the protein-RNA network. These properties suggest a scaffolding function for granule RNAs (i.e. an ability for attracting protein contacts).

We further explored which biophysical properties are characteristic of granule RNAs to make them 'scaffolds'. Actually, we found that granule RNAs are enriched in properties that favour protein-binding propensity such as secondary structure content (i.e. percentage of the RNA sequence with a double-stranded structure), expression level and length (specially of 5' UTR). We validated these findings with experimental results produced in our lab related with FXTAS disease. In FXTAS, there is a CGG repetition on the FMR1 gene that increases the mRNA scaffolding ability. This causes the sequestration of many splicing factors including TRA2A, impeding their function and altering cell physiology in general.

Since new experimental data related to granule composition was published during the progress of the thesis, we decided to further validate our findings. We found indeed that consensus granule proteins (i.e. proteins that are found within the SG under different cell types and conditions) were mainly recruited by protein-RNA interactions rather than by protein-protein contacts. However, by exploring the first published transcriptome of granule RNAs, we discovered that many granule RNAs don't have any known protein interaction. This requires additional strategies for recruiting RNAs into the granule. We identified that RNA-RNA interactions are also enriched in granules. This RNA-RNA network could also have important roles on granule formation and organization. Finally, we explored how some posttranscriptional modifications increase specifically the structural content of certain RNAs, favouring their enrichment into SG.

Paradigm shift on RNP granules research: from a protein-centric to an RNA-centric perspective

The first RNP granules were detected on 1903 by Ramon y Cajal. However, it was not until the 90s with the SG characterisation[8,9] and the discovery of the nucleolus liquid like properties[20,55] that the field took off. Subsequent studies began studying the conditions of formation of RNPs by either knocking down or overexpressing several translation initiation factors and other related RNA-binding proteins[6]. In this sense, these initial studies were focused on the protein contribution for the RNP granule formation. In addition, these membrane less, liquid-liquid phase separated droplets have also attracted the researchers attention due to their relation with prion like proteins and neurodegenerative dieases[20,55].

This protein-centric view was also favoured by the fact that proteins are easier to extract and analyze. Due to its instability and ease for degradation, RNA is more difficult to isolate and characterize. RNA world is poorly understood in comparison to the function of most of the proteins. Function of most non-coding RNAs remain elusive. For instance, one of the biggest groups of non-coding RNAs, lincRNAs, is not classified according to functional criteria but by a negative definition (non-short, non-coding transcripts). By contrast, RNA is believed to have both DNA and protein functions (i.e. information storage and catalysis) during early stages of life. This primitive function could be maintained though evolution for some non-coding transcripts and this would be the reason why some RNAs show catalytic or structural functions in a similar extent as proteins do.

The lack of knowledge on the RNA side of RNP granules until recently also extends to its interactions. Interactions are crucial for understanding granule organization and function. On molecular biology, function raises from interaction, since molecules affect and influence each other though biophysical contacts. However, protein-RNA and RNA-RNA interactions are worse characterized than protein-protein interactions generally. The two-hybrid system, able to detect high-throughput protein-protein interactions, was first developed in 1989 (initially on yeast and later on mammalian cells)[108].

Protein-RNA interactions were not detectable on a high-throughput scale until the development of the eCLIP methodology on 2016, which is still noisy and limited but provides the RNA interactions of about 200 proteins in two specific cell lines[83]. Finally, RNA-RNA interactions are still very elusive. Whereas interactions between DNA strands are fundamental for the DNA structure, RNA-RNA interactions have not been extensively studied yet. Best characterized cases involve the interaction of miRNAs and snRNAs but same biochemical composition of other classes of RNAs such as mRNAs and lincRNAs suggest that they could also undergo RNA-RNA

interactions. In fact, some high-throughput technologies have been developed in the recent years to detect these kinds of interactions. Examples include the PARIS, SPLASH, MARIO and LIGR-seq methods[85].

Recent experimental developments on RNA biochemistry have allowed a paradigm shift on RNP research. Since the appearance of the first SG transcriptome on 2017, other studies tried to better understand the role of the RNA within granules[17,109,110]. This includes an emphasis on deciphering the role of RNA-RNA interactions. Evidence showed that RNA can self-assembly to form SG-like assemblies but other studies state that the correlation between RNA concentration and granule formation may be different depending the biological context (nucleus or cytoplasm for instance)[111].

In any case, since granules are composed of proteins and RNA it is very likely that RNA would exert some kind of function simply because of its necessary interaction with other elements of the granule in order to be recruited.

In **Chapter 4**, we showed how granule RNAs form a distinct protein-RNA network at different levels. They tend (i) to interact with a higher number of proteins, (ii) to be more shared by protein pairs and (iii) to globally increase the centrality of the protein-RNA network in comparison to the RNAs in the rest of cell. In contrast, protein-protein networks showed similar proteins for granule and non-granule components. Also, consensus granule proteins are not especially densely connected nor close regarding protein-protein contacts. In this sense, protein-RNA interactions seem more important organizing the granule than protein-protein interactions.

Functional role of RNA in RNP granules: RNA scaffolds

In **Chapter 4,** we described how granule RNAs have specific properties that increase their ability to bind proteins. This points that granule RNAs may act as scaffolding molecules, highly interacting prone molecules that have the ability to attract many other molecules. In the context of SG, RNA scaffolds are able to recruit proteins into the granule, affecting its protein composition.

To find the basis of this scaffolding ability we analyzed the biophysical properties enriched in the granules RNA. In this way, we obtained that structural content, length of the 5´UTR and expression level favour the interaction with proteins[55]. RNA structural content is important to generate specific, stable binding sites[112]. The length is proportional to the protein interactions as the longer a sequence is, the more likely it will generate protein binding sites. 5´-UTR is one of the main RNA regulatory regions,

and consequently one of the most contacted by proteins. Finally, the expression level increases the likelihood of two potential interactors to bind together.

Other examples of scaffolding RNAs have been reported such as the case of architectural RNAs: lncRNAs that have the function of scaffold nuclear bodies[113]. For instance, NEAT1 lincRNA is crucial for paraspeckles (nuclear RNP granules) formation and organization[30,31]. These observations indicate that non-coding RNAs are good candidates for performing scaffolding functions on SG. In fact, our analysis detected NORAD as a highly likely scaffolding candidate. NORAD is a highly abundant lincRNA induced after cell stress that regulates genomic stability by sequestering PUMILIO proteins[114]. Moreover, NORAD was highlighted as one of the first lincRNAs detected on SG[10].

Implications of RNA scaffolding on neurological diseases

Regardless of the protein function they code for, RNA molecules on their own have important functions such as determining granule composition. Consequently, changes at the RNA level may produce granule alterations that ultimately affect the cell physiology. For instance, as described **Chapter 4**, FXTAS patients present mutations on FMR1 RNA that increase its scaffolding ability to sequester proteins into *foci* (e.g. TRA2A) and consequently impair the cell function[4]. At health conditions scaffolding RNAs have a physiological role but when overpassed they can became toxic. This may be the cause of several neurological diseases related with RNP granules and mutations that produce RNAs with repetitive sequences[76].

The discovery of abnormal scaffolding RNAs as drug targets has potential clinical applications. Nucleic acids aptamers are molecules designed to bind specifically certain RNAs (or proteins) in order to modify and regulate their binding activity[115]. They have several advantages such as their quick chemical production, high stability or lack of immunogenicity[116]. Nowadays, there are already three aptamers either approved by US Food and Drug Administration or in late stage or development. There are also six other RNA aptamers undergoing clinical trials[117].

Understanding RNP granules on different conditions and locations

SG are cytoplasmic RNP granules, whereas RNP granules implied in FXTAS are nuclear. However, the main physicochemical principles governing the RNA interactivity are similar for SG and nuclear *foci*.

Despite this, we have to consider that differences exist between nuclear and cytoplasmic granules. Different concentration in different cell compartments like the nucleus and the cytoplasm would influence how RNA influences foci formation[111]. Indeed, a recent study states that specific RNA species suppress phase separation at high concentrations in the nucleus and stimulate it at lower concentration.

Also, just SG formed in different conditions (e.g. different stresses or cell types) have different compositions. In **Chapter 5**, we defined a consensus SG proteome by integrating different experiments on different conditions and selecting those proteins that were always present in SG[50,92]. In this way, we detected that these consensus proteins share more RNA interaction targets than SG proteins from less restrictive sets. Surprisingly, these consensus proteins do not present differences at the protein-protein network level when compared with the rest of the proteome (even with completely non-granule ones).

In summary, different conditions produce different RNP granules and this would presumably lead to slightly different functions. We consider that a deeper consideration on these questions will be an interesting and important source of research in the near future.

Interplay between protein-RNA and RNA-RNA interactions on RNP granules

Although protein-RNA interactions are essential for SG stability, many granule RNAs do not have any known protein interactions. This fact suggests the presence of alternative ways to recruit RNA molecules into the granule. In **Chapter 5**, we described how SG are also enriched in RNA-RNA interactions, having similar distinct features as the protein-RNA network does (i.e. increased number of contacts and centrality values to respect to the non-granule network). However, what is the possible role of these RNA-RNA interactions within the granule?

We have to note that PRI and RRI are not mutually exclusive. We recall again that granule RNAs are enriched in protein contacts. One of the factors that contribute to promote protein interactions lays on the fact that they are on average more structured (more double-stranded) than cytoplasmic RNAs. In this sense, RNA-RNA interactions could also promote protein-RNA interactions in a similar extent. RNA molecules are much bigger than proteins and therefore, proteins only recognize local regions on the RNA for binding. In this context, proteins would not be able to differentiate a double stranded RNA coming from two different interacting molecules (RNA-RNA intermolecular interactions) or from the folding of a single

RNA molecule (RNA-RNA intramolecular interactions). Either intra or inter molecular RNA-RNA interactions would promote protein interactions in a similar extent.

Finally, we have to note that it is still not clear if RNA changes its composition when it enters into the granule. Appearance of helicases and chaperones suggest. Someone says that is one of the granule function the structure remodelling.

Pre-existing interaction networks and its relation with granule formation

Recent models suggest the hypothesis of a network of interactions that exists before granule formation[50]. Events triggering SG formation would increase the concentration of these interacting molecules and once a certain threshold is reached, the phase-separation would occur and the granule would be formed. This idea is behind the assumptions we made for our analysis since we extrapolated the information of protein-RNA interactions detected outside the granule. For the case of the results presented on **Chapter 4**, we considered as granule RNAs those detected to interact with known granule proteins. This list, retrieved considering this assumption, showed a significant overlap with the published atlas of transcripts enriched in SG (about 90% of coincidence)[10]. We consider that physico-chemical forces that are in place regardless of the environmental conditions govern both PPI and PRI. The presence of this pre-existing network of interactions may have a direct link with the mechanism of the granule formation that would be interesting to further address in the future.

A second interesting aspect regarding the interaction network within the granule is the presence of many redundant interactions[50]. Some granule proteins interact with the same RNAs and vice versa. As a result, only a subset of the interactions would be required to recruit all the elements. This may explain how SG can be formed under different conditions and with slightly different compositions. Thus, this redundancy of interactions is helping on maintaining the robustness of the network.

RNP granules as catalytic complexes

The study of the RNP granules is still quite recent and therefore highly unexplored. RNP granules have been extensively studied for their link with aggregation and neurodegeneration but their functions are still far from being completely covered. In this sense, we might expect some novel and interesting functions to be discovered in the future. For instance, some authors suggest that different steps of metabolic pathways could be physically grouped together in the cell employing these phase-separated

droplets[118–120]. This would explain the unexpectedly high catalytic rates in the cells that surpass the reaction rates that would be expected from random molecular encounters. In a similar extent, foci formation have been recently described to be involved in transcription factor complexes, enabling the regulation of distant regions of the DNA at the same time[121–123]. Definitely, granule field is an exciting and fertile area of research that could change our vision of cell biology in the near future.

Chapter 7. Conclusions

- **The protein-protein interaction network does not discriminate granule and non-granule networks.** Despite that SGs are highly dense and concentrated structures, their main proteins (in this thesis defined as consensus granule proteins) form connections in the same way as the non-granule proteins. Specifically, (i) they have similar connectedness and centrality and (ii) they are not more densely connected.

- **RNA has a key structural role in the granule network.** This role is performed at various interaction levels and is a distinctive characteristic of the granule network. Granule (i) RNAs interact with more proteins, (ii) are more shared and (iii) make the protein-RNA network more central in comparison with non-granule RNAs and (iv) are able to cluster together consensus granule proteins.

- **Granule RNAs properties support a scaffolding role.** Granule RNAs are longer and more structured than non-granule RNAs. These properties favor a higher number of interactions. These properties matched with those previously described for the scaffolding RNAs at nuclear foci (such as NEAT1).

- **A CGG repeat expansion increases FMR1 mRNA scaffolding ability.** This effect provokes the sequestration of proteins such as TRA2A into nuclear foci. Both FMR1 and TRA2A co-aggregate in mouse model and in post-mortem human samples. Function impairment of sequestered proteins such TRA2A may explain certain pathophysiological aspects of FXTAS disease.

- **RNA-RNA interactions are enriched in SG and could organize the granule structure.** RNA-RNA may promote protein-RNA interactions by establishing inter-molecular double stranded regions recognized by granule RBPs that bind preferentially to double-stranded regions in the RNA.

- **There is a pre-existing network of interactions before granule-formation.** RNA molecules targeted by granule proteins under non-stress conditions resemble with high accuracy the SG transcriptome. In addition, the RNA-RNA interactions measured under non-stress conditions reveal a distinct network organization for granule RNAs.

Appendix

List of publications

1. Ciryam, P., Lambert-Smith, I. A., Bean, D. M., Freer, R., Cid, F., Tartaglia, G. G., ... & Dobson, C. M. (2017). Spinal motor neuron protein supersaturation patterns are associated with inclusion body formation in ALS. *Proceedings of the National Academy of Sciences*, *114*(20), E3935-E3943.

2. Cid-Samper, F., Gelabert-Baldrich, M., Lang, B., Lorenzo-Gotor, N., Ponti, R. D., Severijnen, L. A. W., ... & Tartaglia, G. G. (2018). An Integrative Study of Protein-RNA Condensates Identifies Scaffolding RNAs and Reveals Players in Fragile X-Associated Tremor/Ataxia Syndrome. *Cell reports*, *25*(12), 3422-3434.

3. Chatterji, P., Williams, P. A., Whelan, K. A., Samper, F. C., Andres, S. F., Simon, L. A., ... & Liang, S. (2019). Posttranscriptional regulation of colonic epithelial repair by RNA binding protein IMP1/IGF2BP1. *EMBO reports*, *20*(6).

4. Cid-Samper F., Dasti A., Bechara E., Tartaglia, G. G. (2019). RNA-centric Approaches to Study RNA-Protein Interactions *in vitro* and *in silico*. *Methods. Accepted for publication.*

5. Cid-Samper F. Vandelli A., Sanchez de Groot N. & Tartaglia, G. G. (2019). Stress granules network analysis reveals a RNA scaffolding population. *To be submitted to Molecular Systems Biology.*

6. Ciryam, P., Antalek, M., Cid, F., Tartaglia, G. G., Dobson, C. M., Guttsches, A. K., ... & Morimoto, R. I. (2019). Escalating protein supersaturation underlies inclusion formation in muscle proteinopathies. *bioRxiv*, 762245

Supplementary material of Chapter 4

Supplementary Figure 1 [related to Figure 1]. Datasets A) Granule RBPs Red circle: granule- forming proteins, Blue circle: RBPs, as defined in Gerstberger et al, 2014 (Gerstberger et al., 2014). Intersection represents granule RBPs. **B)** Number of interactions. Red circle: granule-forming proteins. Blue circle: RBPs with known targets. Intersection represents granule RBPs with known targets. **Distribution of centrality values of granule and non-granule RBPs in different interaction networks. C)** Centrality distributions for the human dataset. Up: Protein-protein network. (p-value (left) = 0.39, p-value (centre) = 0.41, p-value (right) = 0.36. Down: Protein-RNA network (p-value (left) = 0.003, p-value (centre) = 0.007, p-value (right) = 0.01. **D)** Centrality distributions for the yeast dataset. Up: Protein-protein network. (p-value (left) = 0.26, p-value (centre) = 0.30, p-value (right) = 0.18. Down: Protein-RNA network (p-value (left) = 0.02, p-value (centre) = 0.05, p-value (right) = 0.01.

Supplementary Figure 2 [related to Figure 1]. Number of RNA targets of granule and non- granule RBPs: A) First quartile of the reads/expression distribution (Q1). **B)** Second quartile (Q2).

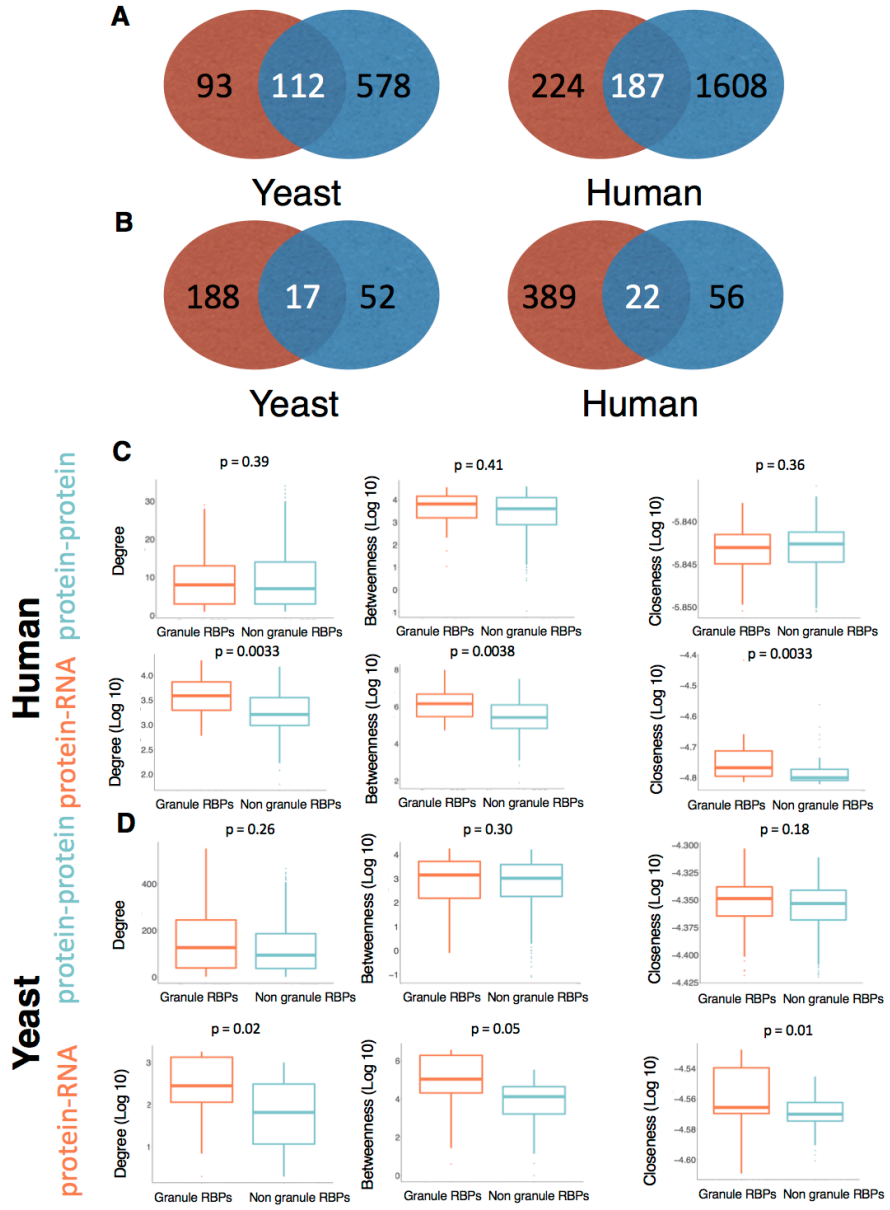
Supplementary Figure 3 [related to Figure 1,2]. Properties of granule RNAs. A) RNAs interacting exclusively with granule forming RBPs have higher number of protein contacts (p-value = 0.04, Wilcoxon test). Human transcripts: **B)** Granule RNAs have more structured UTRs (p-value = 0.007; KS test). PARS analysis on 3'UTR of granule and non-granule RNAs**.** Yeast granule RNA are **C)** structured (p-value = 0.001; KS test; PARS data), and **D)** more abundant (p-value = 2.2e-16; KS test) than non-granule RNAs. The UTR analysis was not performed due to the lack of annotation.

Supplementary Figure 4 [related to Figure 2,3]. Computational predictions of granule-forming components. A) Granule transcripts are predicted to be more structured (structural content according is measured using CROSS; p-value < 2.2e-16, KS test). **B** and **C)** *cat*GRANULE performances on human and yeast experimentally described granule-forming proteins. AUC (Area under the ROC curve) is used to measure the discriminative power of the method. **D)** Distribution of *cat*GRANULE scores for the whole human proteome. TRA2A (*cat*GRANULE score = 2.14) ranks 188^{th} out of 20190 human proteins (i.e. 1% of the distribution).

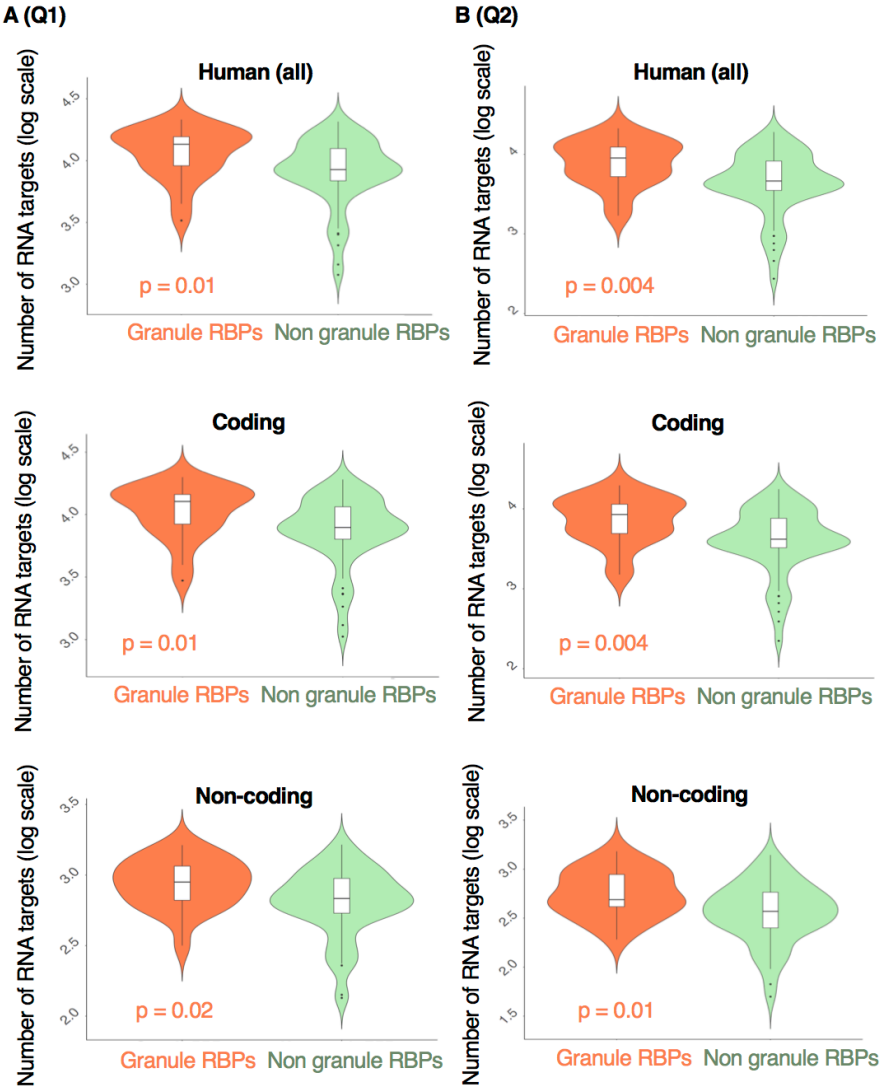
Supplementary Figure 5 [related to Figure 4]. TRA2A levels in human lymphocytes and COS-7 cell model. A) Human lymphocytes from control (A) or pre mutation-carrier (B) were lysated and both RNA and protein were isolated (*** p-value < 0.01). Relative TRA2A RNA expression (left panel) and TRA2A protein (right panel) are represented. **B)** COS-7 cells were transfected with CGG(60X) and compared to controls. After 24h, 48h or 72h of transfection cells were pelleted and RNA and protein extraction was performed. Relative TRA2A RNA expression (left panel) and TRA2A protein (right panel) are represented.

Supplementary Figure 6 [related to Figure 6]. TRA2B over-expression and TRA2A knock-down. A) Control COS-7 cells (without CGG(60X) transfection) were transfected with GFP-TRA2B and siTRA2A. **B)** COS-7 cells were transfected with CGG(60X), GFP-TRA2B and siTRA2A. In both A and B, after 48 hours, cells were hybridized with Cy3-GGC(8X) probe and immunostained with an antibody against TRA2B. The graph represents TRA2B/CGG levels**. TRA2A over-expression and TRA2B knock-down. C)** Control COS-7 cells were transfected with siTRA2B and GFP-TRA2A (in absence of CGG(60X) transfection). **D)** COS-7 cells were transfected with CGG(60X), siTRA2B and GFP-TRA2A. In both A and B, after 48 hours of transfection cells were hybridized with Cy3- GGC(8X) probe and immunostained with antiGFP. The graphs represent TRA2A/CGG levels. **E)** TRA2B protein levels in COS-7 cells treated as in B. **TRA2A and TRA2B over-expression** COS-7 cells were transfected with GFP-TRA2A **F)** or GFP-TRA2B **G)** and CGG(60X). After 48 hours, cells were hybridized with Cy3-GGC(8X) probe and immunostained with an antibody against either TRA2A or TRA2B. Graphs represent TRA2A/TRA2B/CGG levels.

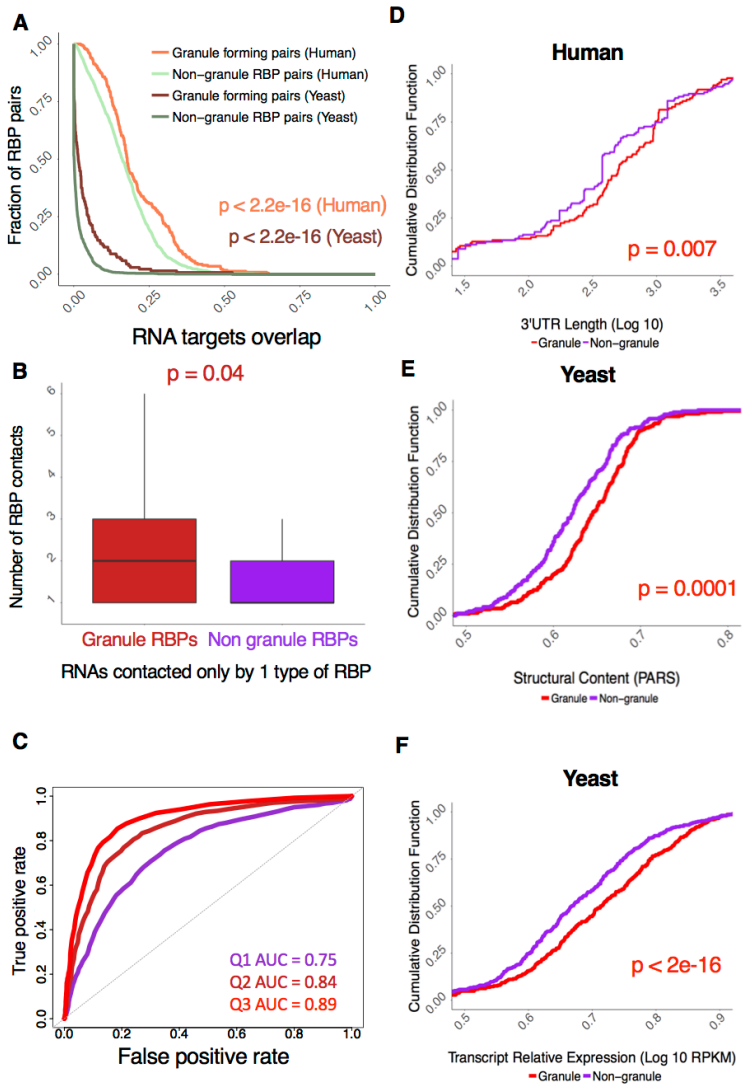
Supplementary Figure 7 [related to Figure 10]. A-F) TRA2A immunohistochemistry in human hippocampus from FXTAS. **G-H)** TRA2A immunohistochemistry in premutated mouse model (counterstaining is done with haematoxylin; the arrows points to the inclusions).



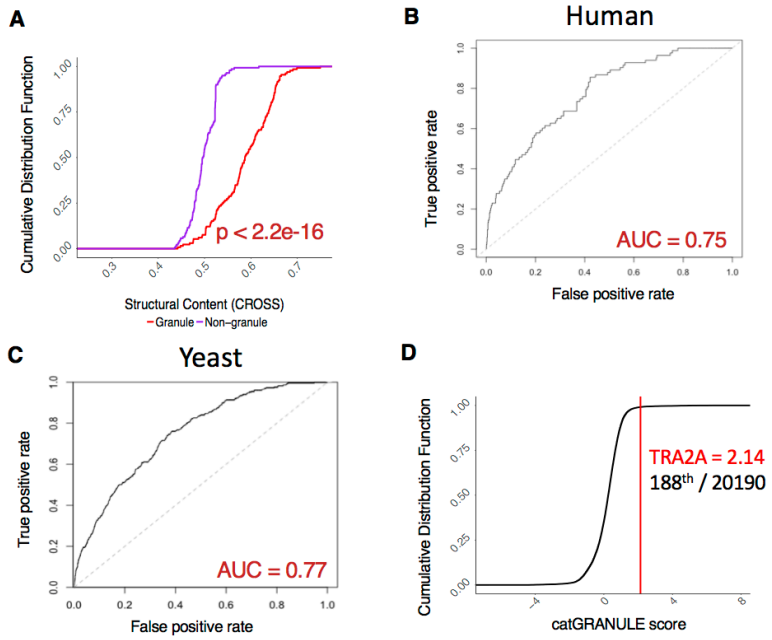
Supplementary Figure 1



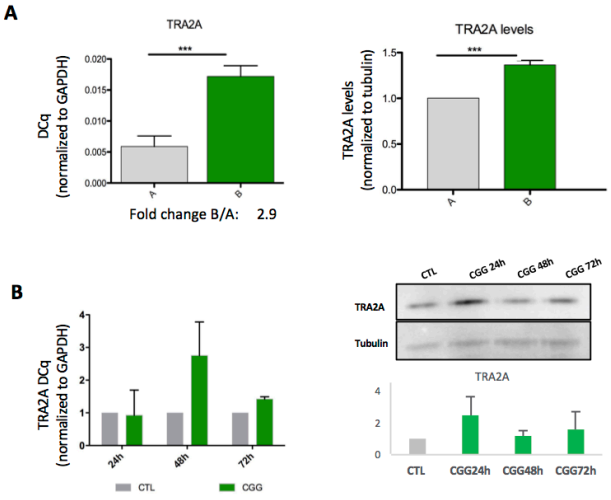
Supplementary Figure 2



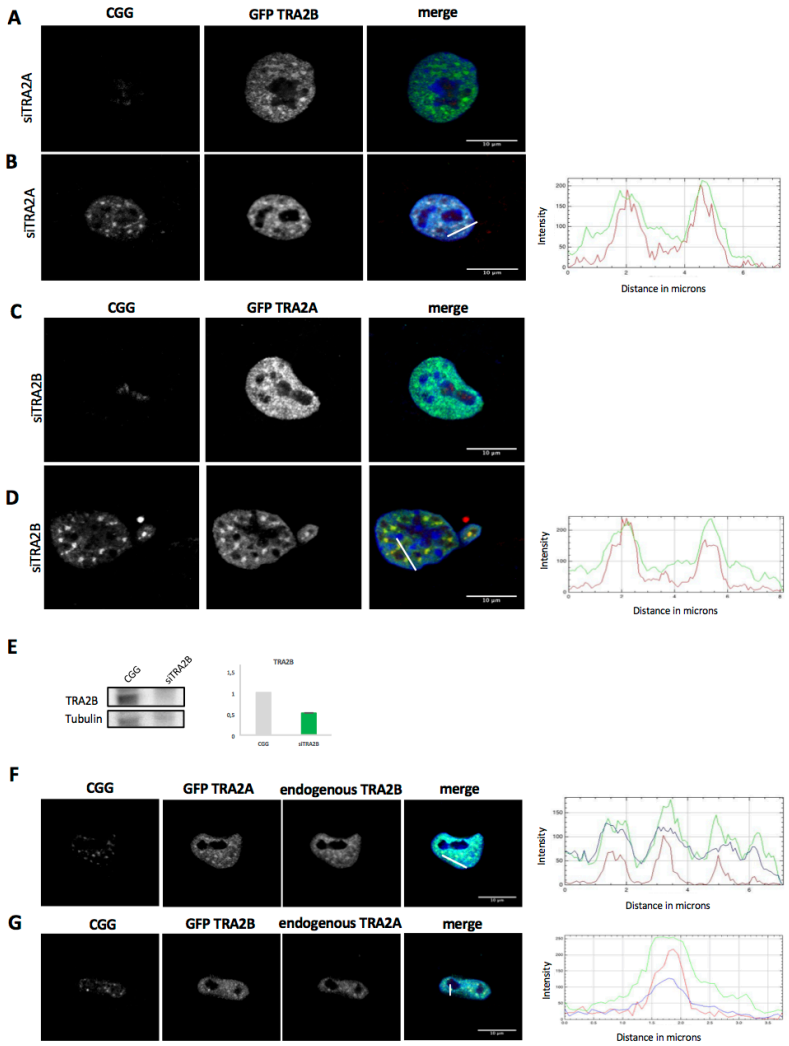
Supplementary Figure 3



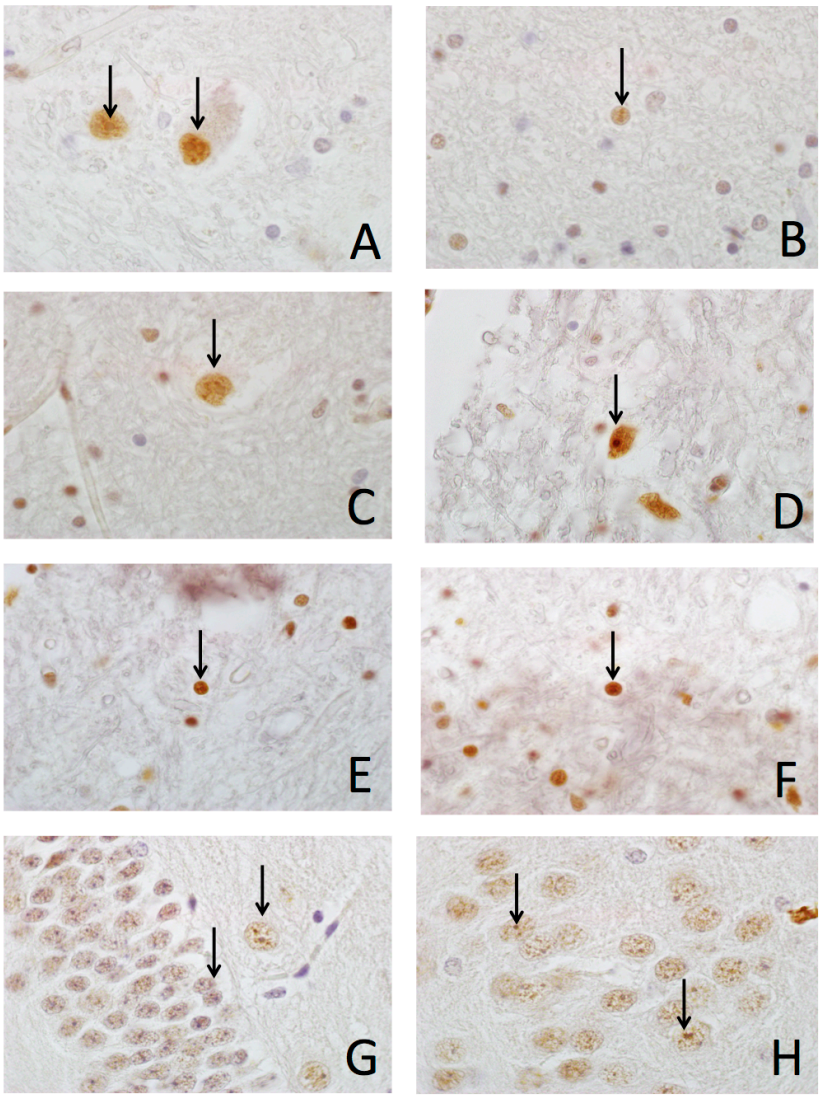
Supplementary Figure 4



Supplementary Figure 5



Supplementary Figure 6



Supplementary Figure 7

Bibliography

1. Protter, D. S. W. & Parker, R. Principles and Properties of Stress Granules. *Trends Cell Biol.* **26**, 668–679 (2016).
2. Guzikowski, A. R., Chen, Y. S. & Zid, B. M. Stress-induced mRNP granules: Form and function of processing bodies and stress granules. *Wiley Interdiscip. Rev. RNA* **10**, e1524 (2019).
3. Li, Y. R., King, O. D., Shorter, J. & Gitler, A. D. Stress granules as crucibles of ALS pathogenesis. *J. Cell Biol.* **201**, 361–372 (2013).
4. Cid-Samper, F. *et al.* An integrative study on ribonucleoprotein condensates identifies scaffolding RNAs and reveals a new player in Fragile X-associated Tremor/Ataxia Syndrome. *bioRxiv* (2018). doi:10.1101/298943
5. Ramaswami, M., Taylor, J. P. & Parker, R. Altered Ribostasis: RNA-Protein Granules in Degenerative Disorders. *Cell* **154**, 727–736 (2013).
6. Anderson, P. & Kedersha, N. RNA granules: post-transcriptional and epigenetic modulators of gene expression. *Nat. Rev. Mol. Cell Biol.* **10**, 430–436 (2009).
7. Buchan, J. R. & Parker, R. Eukaryotic stress granules: the ins and outs of translation. *Mol. Cell* **36**, 932–941 (2009).
8. Kedersha, N. L., Gupta, M., Li, W., Miller, I. & Anderson, P. RNA-Binding Proteins Tia-1 and Tiar Link the Phosphorylation of Eif-2α to the Assembly of Mammalian Stress Granules. *J. Cell Biol.* **147**, 1431–1442 (1999).
9. Bashkirov, V. I., Scherthan, H., Solinger, J. A., Buerstedde, J.-M. & Heyer, W.-D. A Mouse Cytoplasmic Exoribonuclease (mXRN1p) with Preference for G4 Tetraplex Substrates. *J. Cell Biol.* **136**, 761–773 (1997).
10. Khong, A. *et al.* The Stress Granule Transcriptome Reveals Principles of mRNA Accumulation in Stress Granules. *Mol. Cell* **68**, 808–820.e5 (2017).
11. Barbee, S. A. *et al.* Staufen- and FMRP-Containing Neuronal RNPs Are Structurally and Functionally Related to Somatic P Bodies. *Neuron* **52**, 997–1009 (2006).
12. Patel, A. *et al.* A Liquid-to-Solid Phase Transition of the ALS Protein FUS Accelerated by Disease Mutation. *Cell* **162**, 1066–1077 (2015).

13. Beckmann, B. M. *et al.* The RNA-binding proteomes from yeast to man harbour conserved enigmRBPs. *Nat. Commun.* **6**, (2015).
14. Jang, S. *et al.* The Glycolytic Protein Phosphofructokinase Dynamically Relocalizes into Subcellular Compartments with Liquid-like Properties. *bioRxiv* (2019). doi:10.1101/636449
15. Calabretta, S. & Richard, S. Emerging Roles of Disordered Sequences in RNA-Binding Proteins. *Trends Biochem. Sci.* **40**, 662–672 (2015).
16. Polymenidou, M. The RNA face of phase separation. *Science* **360**, 859–860 (2018).
17. Van Treeck, B. *et al.* RNA self-assembly contributes to stress granule formation and defining the stress granule transcriptome. *Proc. Natl. Acad. Sci.* **115**, 2734–2739 (2018).
18. Wheeler, J. R., Matheny, T., Jain, S., Abrisch, R. & Parker, R. Distinct stages in stress granule assembly and disassembly. *eLife* **5**, (2016).
19. Alberti, S. Phase separation in biology. *Curr. Biol.* **27**, R1097–R1102 (2017).
20. Zhang, H. *et al.* RNA Controls PolyQ Protein Phase Transitions. *Mol. Cell* **60**, 220–230 (2015).
21. Patel, A. *et al.* ATP as a biological hydrotrope. *Science* **356**, 753–756 (2017).
22. Rice, A. M. & Rosen, M. K. ATP controls the crowd. *Science* **356**, 701–702 (2017).
23. Kedersha, N. *et al.* Stress granules and processing bodies are dynamically linked sites of mRNP remodeling. *J. Cell Biol.* **169**, 871–884 (2005).
24. Franzmann, T. M. & Alberti, S. Protein Phase Separation as a Stress Survival Strategy. *Cold Spring Harb. Perspect. Biol.* **11**, a034058 (2019).
25. Boke, E. *et al.* Amyloid-like Self-Assembly of a Cellular Compartment. *Cell* **166**, 637–650 (2016).
26. Rayman, J. B., Karl, K. A. & Kandel, E. R. TIA-1 Self-Multimerization, Phase Separation, and Recruitment into Stress Granules Are Dynamically Regulated by Zn 2+. *Cell Rep.* **22**, 59–71 (2018).
27. Jain, S. *et al.* ATPase-Modulated Stress Granules Contain a Diverse Proteome and Substructure. *Cell* **164**, 487–498 (2016).
28. Brasch, K. & Ochs, R. L. Nuclear bodies (NBs): A newly 'rediscovered' organelle. *Exp. Cell Res.* **202**, 211–223 (1992).

29. Gall, J. G. Cajal Bodies: The First 100 Years. *Annu. Rev. Cell Dev. Biol.* **16**, 273–300 (2000).
30. Clemson, C. M. *et al.* An Architectural Role for a Nuclear Noncoding RNA: NEAT1 RNA Is Essential for the Structure of Paraspeckles. *Mol. Cell* **33**, 717–726 (2009).
31. Chujo, T., Yamazaki, T. & Hirose, T. Architectural RNAs (arcRNAs): A class of long noncoding RNAs that function as the scaffold of nuclear bodies. *Biochim. Biophys. Acta BBA - Gene Regul. Mech.* **1859**, 139–146 (2016).
32. Boisvert, F.-M., van Koningsbruggen, S., Navascués, J. & Lamond, A. I. The multifunctional nucleolus. *Nat. Rev. Mol. Cell Biol.* **8**, 574–585 (2007).
33. Hnisz, D., Shrinivas, K., Young, R. A., Chakraborty, A. K. & Sharp, P. A. A Phase Separation Model for Transcriptional Control. *Cell* **169**, 13–23 (2017).
34. Buchan, J. R., Muhlrad, D. & Parker, R. P bodies promote stress granule assembly in Saccharomyces cerevisiae. *J. Cell Biol.* **183**, 441–455 (2008).
35. Brangwynne, C. P. *et al.* Germline P Granules Are Liquid Droplets That Localize by Controlled Dissolution/Condensation. *Science* **324**, 1729–1732 (2009).
36. Buchan, J. R., Kolaitis, R.-M., Taylor, J. P. & Parker, R. Eukaryotic Stress Granules Are Cleared by Autophagy and Cdc48/VCP Function. *Cell* **153**, 1461–1474 (2013).
37. Voronina, E., Seydoux, G., Sassone-Corsi, P. & Nagamori, I. RNA Granules in Germ Cells. *Cold Spring Harb. Perspect. Biol.* **3**, a002774–a002774 (2011).
38. Nakamura, H. *et al.* Intracellular production of hydrogels and synthetic RNA granules by multivalent molecular interactions. *Nat. Mater.* **17**, 79–89 (2018).
39. Mokas, S. *et al.* Uncoupling Stress Granule Assembly and Translation Initiation Inhibition. *Mol. Biol. Cell* **20**, 2673–2683 (2009).
40. Sheth, U. Decapping and Decay of Messenger RNA Occur in Cytoplasmic Processing Bodies. *Science* **300**, 805–808 (2003).
41. Molliex, A. *et al.* Phase separation by low complexity domains promotes stress granule assembly and drives pathological fibrillization. *Cell* **163**, 123–133 (2015).
42. Babu, M. M., van der Lee, R., de Groot, N. S. & Gsponer, J. Intrinsically disordered proteins: regulation and disease. *Curr. Opin. Struct. Biol.* **21**, 432–440 (2011).

43. Uversky, V. N. Intrinsically disordered proteins in overcrowded milieu: Membrane-less organelles, phase separation, and intrinsic disorder. *Curr. Opin. Struct. Biol.* **44**, 18–30 (2017).
44. Yang, X. *et al.* Stress Granule-Defective Mutants Deregulate Stress Responsive Transcripts. *PLoS Genet.* **10**, e1004763 (2014).
45. Solomon, S. *et al.* Distinct Structural Features ofCaprin-1 Mediate Its Interaction with G3BP-1 and Its Induction of Phosphorylation of Eukaryotic Translation Initiation Factor 2α, Entry to Cytoplasmic Stress Granules, and Selective Interaction with a Subset of mRNAs. *Mol. Cell. Biol.* **27**, 2324–2342 (2007).
46. Tourrière, H. *et al.* The RasGAP-associated endoribonuclease G3BP assembles stress granules. *J. Cell Biol.* **160**, 823–831 (2003).
47. Kedersha, N. *et al.* G3BP–Caprin1–USP10 complexes mediate stress granule condensation and associate with 40S subunits. *J. Cell Biol.* **212**, 845–860 (2016).
48. Tsai, N.-P., Ho, P.-C. & Wei, L.-N. Regulation of stress granule dynamics by Grb7 and FAK signalling pathway. *EMBO J.* **27**, 715–726 (2008).
49. Wippich, F. *et al.* Dual Specificity Kinase DYRK3 Couples Stress Granule Condensation/Dissolution to mTORC1 Signaling. *Cell* **152**, 791–805 (2013).
50. Markmiller, S. *et al.* Context-Dependent and Disease-Specific Diversity in Protein Interactions within Stress Granules. *Cell* **172**, 590–604.e13 (2018).
51. Youn, J.-Y. *et al.* High-Density Proximity Mapping Reveals the Subcellular Organization of mRNA-Associated Granules and Bodies. *Mol. Cell* **69**, 517–532.e11 (2018).
52. Ma, W.-X., Yong, X. & Zhang, H.-Q. Diversity of interaction solutions to the (2+1)-dimensional Ito equation. *Comput. Math. Appl.* **75**, 289–295 (2018).
53. Namkoong, S., Ho, A., Woo, Y. M., Kwak, H. & Lee, J. H. Systematic Characterization of Stress-Induced RNA Granulation. *Mol. Cell* **70**, 175–187.e8 (2018).
54. Hubstenberger, A. *et al.* P-Body Purification Reveals the Condensation of Repressed mRNA Regulons. *Mol. Cell* **68**, 144–157.e5 (2017).
55. Langdon, E. M. *et al.* mRNA structure determines specificity of a polyQ-driven phase separation. *Science* **360**, 922–927 (2018).

56. Aumiller, W. M., Pir Cakmak, F., Davis, B. W. & Keating, C. D. RNA-Based Coacervates as a Model for Membraneless Organelles: Formation, Properties, and Interfacial Liposome Assembly. *Langmuir* **32**, 10042–10053 (2016).
57. Dang, Y. *et al.* Eukaryotic Initiation Factor 2α-independent Pathway of Stress Granule Induction by the Natural Product Pateamine A. *J. Biol. Chem.* **281**, 32870–32878 (2006).
58. Mazroui, R. *et al.* Inhibition of Ribosome Recruitment Induces Stress Granule Formation Independently of Eukaryotic Initiation Factor 2α Phosphorylation. *Mol. Biol. Cell* **17**, 4212–4219 (2006).
59. Teixeira, D. Processing bodies require RNA for assembly and contain nontranslating mRNAs. *RNA* **11**, 371–382 (2005).
60. Kedersha, N. *et al.* Dynamic Shuttling of Tia-1 Accompanies the Recruitment of mRNA to Mammalian Stress Granules. *J. Cell Biol.* **151**, 1257–1268 (2000).
61. Souquere, S. *et al.* Unravelling the ultrastructure of stress granules and associated P-bodies in human cells. *J. Cell Sci.* **122**, 3619–3626 (2009).
62. Chernov, K. G. *et al.* Role of Microtubules in Stress Granule Assembly: MICROTUBULE DYNAMICAL INSTABILITY FAVORS THE FORMATION OF MICROMETRIC STRESS GRANULES IN CELLS. *J. Biol. Chem.* **284**, 36569–36580 (2009).
63. Ivanov, P. A., Chudinova, E. M. & Nadezhdina, E. S. Disruption of microtubules inhibits cytoplasmic ribonucleoprotein stress granule formation. *Exp. Cell Res.* **290**, 227–233 (2003).
64. Wilbertz, J. H. *et al.* Single-Molecule Imaging of mRNA Localization and Regulation during the Integrated Stress Response. *Mol. Cell* **73**, 946–958.e7 (2019).
65. Riback, J. A. *et al.* Stress-Triggered Phase Separation Is an Adaptive, Evolutionarily Tuned Response. *Cell* **168**, 1028–1040.e19 (2017).
66. Bolognesi, B. *et al.* Dosage sensitivity caused by increased protein concentration triggering a liquid phase separation. *Cell Reports* (2016).
67. Helder, S., Blythe, A. J., Bond, C. S. & Mackay, J. P. Determinants of affinity and specificity in RNA-binding proteins. *Curr. Opin. Struct. Biol.* **38**, 83–91 (2016).
68. Arimoto, K., Fukuda, H., Imajoh-Ohmi, S., Saito, H. & Takekawa, M. Formation of stress granules inhibits apoptosis by

suppressing stress-responsive MAPK pathways. *Nat. Cell Biol.* **10**, 1324–1332 (2008).
69. Kim, W. J., Back, S. H., Kim, V., Ryu, I. & Jang, S. K. Sequestration of TRAF2 into Stress Granules Interrupts Tumor Necrosis Factor Signaling under Stress Conditions. *Mol. Cell. Biol.* **25**, 2450–2462 (2005).
70. Neumann, M. *et al.* Ubiquitinated TDP-43 in Frontotemporal Lobar Degeneration and Amyotrophic Lateral Sclerosis. *Science* **314**, 130–133 (2006).
71. Jovičić, A. *et al.* Modifiers of C9orf72 dipeptide repeat toxicity connect nucleocytoplasmic transport defects to FTD/ALS. *Nat. Neurosci.* **18**, 1226–1229 (2015).
72. Orengo, J. P. *et al.* Expanded CTG repeats within the DMPK 3' UTR causes severe skeletal muscle wasting in an inducible mouse model for myotonic dystrophy. *Proc. Natl. Acad. Sci.* **105**, 2646–2651 (2008).
73. Botta, A. *et al.* Effect of the [CCTG]n repeat expansion on ZNF9 expression in myotonic dystrophy type II (DM2). *Biochim. Biophys. Acta BBA - Mol. Basis Dis.* **1762**, 329–334 (2006).
74. Mateju, D. *et al.* An aberrant phase transition of stress granules triggered by misfolded protein and prevented by chaperone function. *EMBO J.* **36**, 1669–1687 (2017).
75. Anderson, P., Kedersha, N. & Ivanov, P. Stress granules, P-bodies and cancer. *Biochim. Biophys. Acta BBA - Gene Regul. Mech.* **1849**, 861–870 (2015).
76. Jain, A. & Vale, R. D. RNA phase transitions in repeat expansion disorders. *Nature* **546**, 243–247 (2017).
77. Cordes, M. H., Davidson, A. R. & Sauer, R. T. Sequence space, folding and protein design. *Curr. Opin. Struct. Biol.* **6**, 3–10 (1996).
78. Castello, A. *et al.* Insights into RNA biology from an atlas of mammalian mRNA-binding proteins. *Cell* **149**, 1393–1406 (2012).
79. Allers, J. & Shamoo, Y. Structure-based analysis of protein-RNA interactions using the program ENTANGLE. *J. Mol. Biol.* **311**, 75–86 (2001).
80. Landenmark, H. K. E., Forgan, D. H. & Cockell, C. S. An Estimate of the Total DNA in the Biosphere. *PLOS Biol.* **13**, e1002168 (2015).

81. Van Nostrand, E. L. *et al.* A Large-Scale Binding and Functional Map of Human RNA Binding Proteins. *bioRxiv* (2018). doi:10.1101/179648
82. Kishore, S. *et al.* A quantitative analysis of CLIP methods for identifying binding sites of RNA-binding proteins. *Nat. Methods* **8**, 559–564 (2011).
83. Van Nostrand, E. L. *et al.* Robust transcriptome-wide discovery of RNA-binding protein binding sites with enhanced CLIP (eCLIP). *Nat. Methods* **13**, 508–514 (2016).
84. He, L. & Hannon, G. J. MicroRNAs: small RNAs with a big role in gene regulation. *Nat. Rev. Genet.* **5**, 522–531 (2004).
85. Qiangfeng Cliff Zhang, J. G. RISE: a database of RNA interactome from sequencing experiments. *Nucleic Acids Res.* **Vol. 46**, D194–D201 (2018).
86. Sharma, E., Sterne-Weiler, T., O'Hanlon, D. & Blencowe, B. J. Global Mapping of Human RNA-RNA Interactions. *Mol. Cell* **62**, 618–626 (2016).
87. Nguyen, T. C. *et al.* Mapping RNA–RNA interactome and RNA structure in vivo by MARIO. *Nat. Commun.* **7**, (2016).
88. Aw, J. G. A. *et al.* In Vivo Mapping of Eukaryotic RNA Interactomes Reveals Principles of Higher-Order Organization and Regulation. *Mol. Cell* **62**, 603–617 (2016).
89. Terai, G., Iwakiri, J., Kameda, T., Hamada, M. & Asai, K. Comprehensive prediction of lncRNA–RNA interactions in human transcriptome. *BMC Genomics* **17**, (2016).
90. Engreitz, J. M. *et al.* RNA-RNA Interactions Enable Specific Targeting of Noncoding RNAs to Nascent Pre-mRNAs and Chromatin Sites. *Cell* **159**, 188–199 (2014).
91. Kudla, G., Granneman, S., Hahn, D., Beggs, J. D. & Tollervey, D. Cross-linking, ligation, and sequencing of hybrids reveals RNA-RNA interactions in yeast. *Proc. Natl. Acad. Sci.* **108**, 10010–10015 (2011).
92. Kim, D. I. *et al.* An improved smaller biotin ligase for BioID proximity labeling. *Mol. Biol. Cell* **27**, 1188–1196 (2016).
93. Levsky, J. M. Fluorescence in situ hybridization: past, present and future. *J. Cell Sci.* **116**, 2833–2838 (2003).
94. Langer-Safer, P. R., Levine, M. & Ward, D. C. Immunological method for mapping genes on Drosophila polytene chromosomes. *Proc. Natl. Acad. Sci.* **79**, 4381–4385 (1982).

95. Raj, A., van den Bogaard, P., Rifkin, S. A., van Oudenaarden, A. & Tyagi, S. Imaging individual mRNA molecules using multiple singly labeled probes. *Nat. Methods* **5**, 877–879 (2008).
96. Moon, K. *et al.* Subsurface Nanoimaging by Broadband Terahertz Pulse Near-Field Microscopy. *Nano Lett.* **15**, 549–552 (2015).
97. Moon, S. L. *et al.* Multicolour single-molecule tracking of mRNA interactions with RNP granules. *Nat. Cell Biol.* **21**, 162–168 (2019).
98. Boccaletti, S., Latora, V., Moreno, Y., Chavez, M. & Hwang, D. Complex networks: Structure and dynamics. *Phys. Rep.* **424**, 175–308 (2006).
99. Albert, R. & Barabási, A.-L. Statistical mechanics of complex networks. *Rev. Mod. Phys.* **74**, 47–97 (2002).
100. Barabási, A.-L. & Oltvai, Z. N. Network biology: understanding the cell's functional organization. *Nat. Rev. Genet.* **5**, 101–113 (2004).
101. Guimerà, R. & Nunes Amaral, L. A. Functional cartography of complex metabolic networks. *Nature* **433**, 895–900 (2005).
102. Li, F., Long, T., Lu, Y., Ouyang, Q. & Tang, C. The yeast cell-cycle network is robustly designed. *Proc. Natl. Acad. Sci.* **101**, 4781–4786 (2004).
103. Dorogovtsev, S. N. & Mendes, J. F. F. Evolution of networks with aging of sites. *Phys. Rev. E* **62**, 1842–1845 (2000).
104. Barabási, A. Emergence of Scaling in Random Networks. *Science* **286**, 509–512 (1999).
105. Girvan, M. & Newman, M. E. J. Community structure in social and biological networks. *Proc. Natl. Acad. Sci.* **99**, 7821–7826 (2002).
106. Rand, W. M. Objective Criteria for the Evaluation of Clustering Methods. *J. Am. Stat. Assoc.* **66**, 846–850 (1971).
107. Levandowsky, M. & Winter, D. Distance between Sets. *Nature* **234**, 34–35 (1971).
108. Fields, S. & Song, O. A novel genetic system to detect protein–protein interactions. *Nature* **340**, 245–246 (1989).
109. Lee, C.-Y. & Seydoux, G. Dynamics of mRNA entry into stress granules. *Nat. Cell Biol.* **21**, 116–117 (2019).
110. Berry, J., Weber, S. C., Vaidya, N., Haataja, M. & Brangwynne, C. P. RNA transcription modulates phase transition-driven nuclear body assembly. *Proc. Natl. Acad. Sci.* **112**, E5237–E5245 (2015).

111. Maharana, S. *et al.* RNA buffers the phase separation behavior of prion-like RNA binding proteins. *Science* **360**, 918–921 (2018).
112. Sanchez de Groot, N. *et al.* RNA structure drives interaction with proteins. *Nat. Commun.* **10**, (2019).
113. Ribeiro, D. M. *et al.* Protein complex scaffolding predicted as a prevalent function of long non-coding RNAs. *Nucleic Acids Res.* **46**, 917–928 (2018).
114. Lee, S. *et al.* Noncoding RNA NORAD Regulates Genomic Stability by Sequestering PUMILIO Proteins. *Cell* **164**, 69–80 (2016).
115. Bjerregaard, N., Andreasen, P. A. & Dupont, D. M. Expected and unexpected features of protein-binding RNA aptamers: Protein-binding RNA aptamers. *Wiley Interdiscip. Rev. RNA* **7**, 744–757 (2016).
116. Zhou, J. & Rossi, J. Aptamers as targeted therapeutics: current potential and challenges. *Nat. Rev. Drug Discov.* **16**, 181–202 (2017).
117. Rusconi, C. P. *et al.* RNA aptamers as reversible antagonists of coagulation factor IXa. *Nature* **419**, 90–94 (2002).
118. Castello, A., Hentze, M. W. & Preiss, T. Metabolic Enzymes Enjoying New Partnerships as RNA-Binding Proteins. *Trends Endocrinol. Metab. TEM* **26**, 746–757 (2015).
119. Prouteau, M. & Loewith, R. Regulation of Cellular Metabolism through Phase Separation of Enzymes. *Biomolecules* **8**, 160 (2018).
120. Hentze, M. W. Enzymes as RNA-binding proteins: a role for (di)nucleotide-binding domains? *Trends Biochem. Sci.* **19**, 101–103 (1994).
121. Boija, A. *et al.* Transcription Factors Activate Genes through the Phase-Separation Capacity of Their Activation Domains. *Cell* **175**, 1842–1855.e16 (2018).
122. Gurumurthy, A., Shen, Y., Gunn, E. M. & Bungert, J. Phase Separation and Transcription Regulation: Are Super-Enhancers and Locus Control Regions Primary Sites of Transcription Complex Assembly? *BioEssays* **41**, 1800164 (2019).
123. Erdel, F. & Rippe, K. Formation of Chromatin Subcompartments by Phase Separation. *Biophys. J.* **114**, 2262–2270 (2018).