Computational analysis of ribonucleoprotein networks

Andrea Vandelli

TESI DOCTORAL UPF / year 2022

DIRECTORS DE LA TESI

Dr. Gian Gaetano Tartaglia
DEPARTMENT OF BIOLOGY AND BIOTECHNOLOGY C. DARWIN
UNIVERSITY OF ROME SAPIENZA

Dr. Marc Torrent Burgas
DEPARTMENT OF BIOCHEMISTRY AND MOLECULAR BIOLOGY
UNIVERSITAT AUTÒNOMA DE BARCELONA

DEPARTMENT OF MEDICINE AND LIFE SCIENCES
POMPEU FABRA UNIVERSITY

”To my family, for being my greatest supporters.”
To Leila, for teaching me that sometimes my fears exist only in my head.”

”Do or do not. There is no try.”

Yoda

Acknowledgments

This is the moment in which I can finally stop and look back to what these four years have meant for me and savor the bittersweet approach of the finish line. I understand that this is not the most interesting part to read but I will try my best to make it worth the effort for the courageous ones who want to go on.

The Ph.D. has been a real adventure for me. I remember that at the beginning there was a little bit of tension and emotion to think that a 4-year-journey was about to start, with a lot of responsibilities associated with it, among which the fact that for the first time I found myself living without my family, in a foreign country and with a real working contract! However, I adjusted to this new life fairly quickly, as well as to the beautiful view from the CRG terrace and the sea at walking distance. I still remember the first time I learned about CRG's existence during the master's degree in Bologna and I just thought 'Wow it would be a dream to be in that place!' and there I was, first for the master's internship and then for the Ph.D.

Working at CRG really was wonderful and there I knew a lot of amazing people, like fellow Ph.D. students and especially my lab members. I honestly think that Tartaglia's lab was one of the funniest and craziest places to be in the entire institute and I had really the best time of my life because talking about science with my colleagues really did not feel like working at all.

Yet, there were surely some challenges too, like when I had to change the lab and leave CRG after Gian started the new group in Italy in January 2020 or, even worse, when the pandemic hit two months later and we were forced to learn to work from home. If I have one regret, it is that the

pandemic prevented me to visit a lot of places I wanted to see and cut a little bit of my social life. However, moving from the CRG to the UAB was less tough than I initially thought; the UAB is huge (I got lost many times in its corridors!) and has a park so big that I felt like being in a forest more than inside a university campus. Globally I can say I am pretty happy about this experience and thanks to this I could grow and mature. Now I feel like I can walk towards the future with more confidence in myself and my skills.

Now it is the time to thank all the people that made this journey exceptional and walked alongside me to reach this milestone. First of all, I would like to thank Gian, who made all of this possible. Since the moment he hired me as his Ph.D. student, he was always present not only for work-related stuff but more importantly, for my well-being. We have kept in touch very often even after he went to Italy and we have managed to continue working well together. He is not only my mentor but my friend as well and I am very proud of him for having created a wonderful working environment, where people can feel free to do research without unjustified pressure or anxieties. I think he is the only boss able to manage more than twenty people spread in different countries, without them feeling isolated or lost. I also thank Marc, my co-supervisor, who came to rescue me when I had to leave CRG. I feel very lucky to know him too, because he has always supported me 100 % and gave me all the freedom I needed, especially during the early pandemic period, when I suffered a bit from anxiety and I was fearful of going physically to the lab. I am happy that we are now able to work on exciting projects together. I also thank my thesis committee members for providing important guidance during the first year.

Now I would like to thank all my lab members: Fernando, the biggest video games and anime fan I know, who made the life in the lab so tremendously funny and who was my mentor in the first year of the PhD, explaining to me all stress granules obscure details; Alex, who still is the programming guru of the lab and even now provides me with incredible advices; he will always be the Jesus of the lab; Riccardo, who I always saw as the experienced PhD guy and we are still keeping in touch often even though he lives in Singapore; Michele, the running man together with Gian, always open for a talk; his switchable knife used to cut cakes in the lab has become legendary, as well as his live breakfasts at every virtual lab meeting; Natalia who, together with Marc, helped me a lot with the details of experimental techniques and made me feel at home both at CRG and when I started working at the UAB, providing constant guidance; we sure shared the struggle when trying to find new ideas to close my first stress granules paper; Elias, the best performer and businessman I know and a person who, I am sure of it, will always listen to me if I ever have problems in life; Alessandro and Maria Carla, for their constant jokes and squabbles that made them a sort of married couples; Nieves, for her experience and more specifically for her chicken mask that made the videos always funny; Teresa, Benedetta, Ben and Iona for their support as postdocs and technicians; Nuria, Dani and Javi, for helping me in the first moments at the UAB when I was drowning in the bureaucracy and for being a very welcoming group. Thanks to you all for making my life as a researcher a moment I'll never forget.

Special thanks also to all the members of Notredame lab and more specifically to Athanasios (Malaka), Alessio, Edgar (Andiamo squadra!) and

Björn for the many laughs and precious moments talking about Ph.D. difficulties.
I would like to thank also all my fellow CRG Ph.D. students for the amazing time together and for giving each other strength whenever we felt down.
Finally, I would like to thank Romina, Imma, Gloria, Montse, Anna and many others, because the Ph.D. would not have been the same without their guidance and support.

Now it is time to thank my Italian friends and family.
Vorrei ringraziare innanzitutto i miei amici del master di Bologna che ho ritrovato a Barcellona e che mi hanno accompagnato in questo percorso: Pat, per mille motivi, per la fissazione coi gatti, il suo sbagliare sempre metro, il suo dimenticarsi gli appuntamenti fissati, il suo abbigliamento stravagante; Cri, D&D e Star Wars nerd, perché con lui potrei parlare per ore e giorni senza stancarmi e trovando sempre nuovi spunti di riflessione. Spero mi sopporti ancora dopo l'ennesimo video di carlini; Damiano, per la sua pazzia e il suo accento veneto troppo divertente; ricorderò per sempre il suo leggendario fil di ferro che è stato utilizzato per generare i più svariati utensili quando vivevamo insieme.

Un grazie particolare anche a tutti gli altri amici del master e della triennale di Bologna e ai miei amici di San Felice; potrei dire tante cose per ognuno ma riassumerò dicendo semplicemente che ne abbiamo passate talmente tante insieme che senza tutti loro io non sarei quello che sono oggi. So che potrò contare su di loro per sempre.

A questo punto vorrei ringraziare la mia famiglia e in particolare i miei

genitori Massimo e Cristina e mio fratello Giacomo, ma anche Paolo, Franco, i miei zii Chiara e Gigi e le mie nonne Sara e Gina. Grazie per avermi sostenuto in questi 4 anni e in generale nel corso di tutta la vita; so che siete fieri di me per la mia determinazione e so altrettanto bene che non è sempre stato facile per voi avermi lontano ma sappiate che siete sempre nei miei pensieri, anche se sapete benissimo che la maggior parte delle volte non lo ammetterei mai.

Per concludere vorrei ringraziare Leila. Lei è il segreto della mia forza e se sono arrivato fin qui con il sorriso sulle labbra e tanti bei ricordi da serbare nel cuore è in gran parte merito suo. Abbiamo affrontato questa avventura insieme, da quando siamo venuti a Barcellona per la prima volta per il tirocinio del master a quando poi abbiamo iniziato il dottorato al CRG. Abbiamo resistito a una pandemia globale, abbiamo combattuto contro vicini di casa, muffe e insetti di ogni tipo e siamo cresciuti e maturati durante questo percorso sostenendoci a vicenda. Aver avuto lei al mio fianco in tutto questo sicuramente mi ha aiutato tanto. Potrei scrivere pagine e pagine su di lei e su di noi ma rischierei di riempire tutto lo spazio della tesi per cui mi limiterò a dirle che sono orgoglioso di lei e che non vedo l'ora di scoprire cosa ci riserva il futuro.

I don't know if someone will manage to read until here, but I will conclude by simply thanking all the people that have been part of this journey and have had a positive impact on my life.

Thank you
Andrea Vandelli

Abstract

Ribonucleoprotein condensates such as stress granules (SGs) and processing bodies (PBs) form in response to specific stimuli in the cell. Even though the key protein and RNA elements of these condensates are starting to be uncovered, we do not yet have a full understanding of the molecular network of interactions that can link them. To answer these questions, we analyzed SGs and PBs components through available high-throughput data, finding that both RNAs and proteins enriched in these condensates are poorly structured and create a dense network of contacts. Based on these results, we developed a database named PRALINE, which stores information about different types of condensates, the relationship between their components and the contribution of disease-related single-nucleotide variants, including both computational and experimental data. In a related work, we predicted that the 5' end of SARS-CoV-2 interacts with elements of the innate immune response that are shared with SGs and PBs and our calculations indicate that strong interactors could be sequestered by SARS-CoV-2 for its viral replication, tampering with the formation of the condensates. Overall our analyses could facilitate the study of the underlying structure of SGs, PBs and other aggregates and how SARS-CoV-2 and other pathogens are able to exploit these mechanisms to help their own survival and infectivity.

Resumen

Los condensados de ribonucleoproteínas, como los gránulos de estrés (GE) y los cuerpos de procesamiento (CP), se forman en respuesta a estímulos específicos en la célula. Dado que los elementos clave de proteínas y ARNs de estos condensados están comenzando a descubrirse, aún no disponemos de una visión completa de la red molecular de interacciones que los caracterizan. Para responder a estas preguntas, hemos analizado los componentes de GE y CP a través de los datos de alto rendimiento disponibles y hemos descubierto que tanto los ARNs como las proteínas enriquecidos en estos condensados están poco estructurados y crean una densa red de contactos. Sobre la base de estos resultados, hemos desarrollado una base de datos llamada PRALINE, que almacena información sobre diferentes tipos de condensados, la relación entre sus componentes y la contribución de las variantes de un solo nucleótido relacionadas con enfermedades, incluyendo tanto datos computacionales como experimentales. En otro trabajo relacionado, hemos predicho que el extremo 5' del SARS-CoV-2 interactúa con elementos de la respuesta inmune innata que se comparten con los GE y los CP y nuestros cálculos indican que el SARS-CoV-2 podría secuestrar interactores fuertes para su replicación viral, alterando la formación de los condensados. En general, nuestros análisis podrían facilitar el estudio de la estructura subyacente de los GE, CP y otros agregados y dilucidar cómo el SARS-CoV-2 y otros patógenos pueden explotar estos mecanismos para ayudar a su propia supervivencia e infectividad.

Preface

My doctoral thesis revolves around the analysis of macromolecular interactions in the context of ribonucleoprotein condensates like stress granules and processing bodies, formed through the biophysical process of phase separation. These organelles exhibit a complex inner structure and organization and are regulated by the network of contacts established by their different protein and RNA components. Alterations in their composition can convert these complexes into irreversible and pathological aggregates, involved in several pathological states. Furthermore, these condensates can be exploited by bacterial and viral pathogens, which can modulate their formation to favor their replication process.

The results of my research have produced five first-author articles (in addition to collaborative papers and other manuscripts currently under submission): a review, three original publications and one pre-print.

These publications are presented along with the thesis, which is structured into four parts and nine chapters.

In **Chapter 1** I will introduce macromolecular interactions patterns and the field of ribonucleoprotein granules and phase-separation, with a focus on the formation, composition and function of liquid-like condensates such as stress granules and processing bodies, their role in neurodegenerative diseases and how they are exploited by the pathogens during their infection process. In **Chapter 2** I will summarize the major points articulated in the introduction and lay out my objectives for the thesis work. Then, I will start presenting the results obtained in my publications, pref-

aced with a short introduction for each chapter. In **Chapter 3** I present a review article about how to create a workflow to study protein-RNA interactions patterns and to catalog potentially new RNA-binding proteins, combining experimental and computational approaches.
Chapter 4 introduces the analysis I did on stress granules and processing bodies, studying their proteomes and transcriptomes and the networks of interactions they form.
This analysis helped the creation of PRALINE that I describe in **Chapter 5**, the new database we developed for the interrogation of proteins and RNAs in liquid-like condensates, such as SGs and PBs, but also in solid-like assemblies including amyloids, containing information about the physicochemical properties of their components and the role of disease-related single-nucleotide variants (SNVs). Then **Chapter 6** describes our study on SARS-CoV-2, focusing on the characterization of its genome and on its interactions with the human host.
This work was expanded with a further study introduced in **Chapter 7** where we compared different human-SARS-CoV-2 protein-RNA interactome experiments available in the literature, to investigate the small set of proteins shared among them.
Chapter 8 of my thesis will be the discussion part in which the main results of the previous chapters are highlighted, together with their importance for the scientific community, while giving my personal opinion on the current limitations and possible future development of the field and **Chapter 9** summarises the main conclusions of the results presented in the thesis.
Finally, in **Appendix A-C** sections I will report the supplementary materials of my papers, in **Appendix D** I will include the complete list of my publications and in **Appendix E** I will show the posters presented at

conferences I attended throughout the Ph.D.
The complete bibliography of the thesis is then reported.

Table of Contents

Part I

INTRODUCTION

CHAPTER 1

INTERNAL ORGANISATION OF RIBONUCLEOPROTEIN CONDENSATES AND IMPLICATIONS FOR NEURODEGENERATIVE DISEASES AND PATHOGENIC INFECTIONS

1.1. Macromolecular interactions

A macromolecule is generally a large molecule involved in one or several biophysical processes. It can be a polymer composed of many single units called monomers or can be a stand-alone large non-polymeric molecule. In biochemistry, nucleic acids and proteins are the most common types of biopolymers, while molecules like lipids are instead non-polymeric (Berg et al., 2002).
In biology and, more specifically, in the crowded cellular environment, macromolecules can come into contact, coordinating with each other to carry out a huge amount of different reactions and biochemical pathways. In this context, the three most common types of macromolecular interactions involving biopolymers are protein-protein, protein-RNA and RNA-RNA interactions, where shape recognition, sequence motives, and secondary and tertiary structure domains or patterns become the basis to regulate these networks of contacts.

1.1.1. Protein-protein interactions

Protein-protein interactions (PPIs) involve two or more proteins that are attracted together thanks to chemical bonds and electrostatic forces.
They can be classified according to the affinity of the binding, the composition of the aggregate and the reversibility of the interaction (Nooren, 2003).
In this context, PPIs can involve different units (hetero-oligomers) or identical molecules (homo-oligomers) and in this case, they can further organize in a bigger structure having structural symmetry (Monod et al., 1965; Goodsell and Olson, 2000).

In addition to a compositional difference, a complex can be distinguished into obligate and non-obligate. In an obligate complex, the single proteins do not exist as stable entities *in vivo* and usually cannot exert their functions on their own. Non-obligate complexes instead involve independently stable proteins that are often not co-localized, including receptor-ligand, antibody-antigen interactions and many signaling complexes.

The reversibility and the duration of the interaction are other classification criteria. Obligate interactions are often permanent, while non-obligate ones can be permanent or transient (Nooren, 2003).
Furthermore, transient interactions can be weak if the bond is broken and formed continuously, or strong if they require a specific trigger to be dissociated (**Figure 1**).

PPIs are also extensively controlled to catalyze the plethora of different reactions of the cell and this can happen in three ways: co-localization in time and space of the interacting proteins leading to their physical encounter; control of the local concentration through mechanisms like gene expression, diffusion and temporary storages; regulation through factors like the concentration of ions, the temperature, pH and phosphorylation to control the local physicochemical environment.
In addition, the control of PPIs depends on the binding affinity of the complex. For example, an interaction between a receptor and its ligand, which are brought together by controlling their localization, is usually an irreversible and high-affinity binding; on the other hand, in a transient PPI usually the affinity among the binders needs to change continuously and in this case a physicochemical control is more appropriate, using ions and other different factors as triggers to regulate the network.

This is linked with the binding specificity of proteins, which derives mainly from chemical and shape complementarity, but also from the local concentration of components, their localization and binding affinity. This gives rise to a plethora of possibilities, with some proteins having multiple potential targets and competing for binders, while others interacting with a single specific ligand (Nooren, 2003).
Finally, the structural characteristics of the binding interfaces can also affect the interaction networks. Techniques like X-ray crystallography, Nuclear magnetic resonance (NMR) spectroscopy and the more recent Cryogenic electron microscopy (cryo-EM) have increased a lot the amount of structural information both on single proteins and on complexes. The X-ray technique is based on the crystallization of a molecule and optimization of its quality, followed by the collection of X-ray diffraction data (Shi, 2014). The NMR spectroscopy instead is based on the nuclei of different isotopes that are subjected to a magnetic field in the spectrometer. As these nuclei magnetize they resonate at different frequencies, which are correlated and combined to extract multidimensional spectra, used by a computer to calculate an ensemble of 3D structures (Kwan et al., 2011). Finally, cryo-EM exploits the transmission electron microscope to visualize a sample at very low temperatures (Milne et al., 2013).

In the context of PPI interfaces, several parameters have been studied such as the polarity and the size of the contact area (Chothia and Janin, 1975; Janin et al., 1988; Jones and Thornton, 1995). In general, the interfaces of obligate complexes are larger and more hydrophobic compared to non-obligate ones (Jones and Thornton, 1996; Lo Conte et al., 1999) which, being able to exist independently as single units, exhibit a more polar interface to increase their solubility and facilitate their folding.

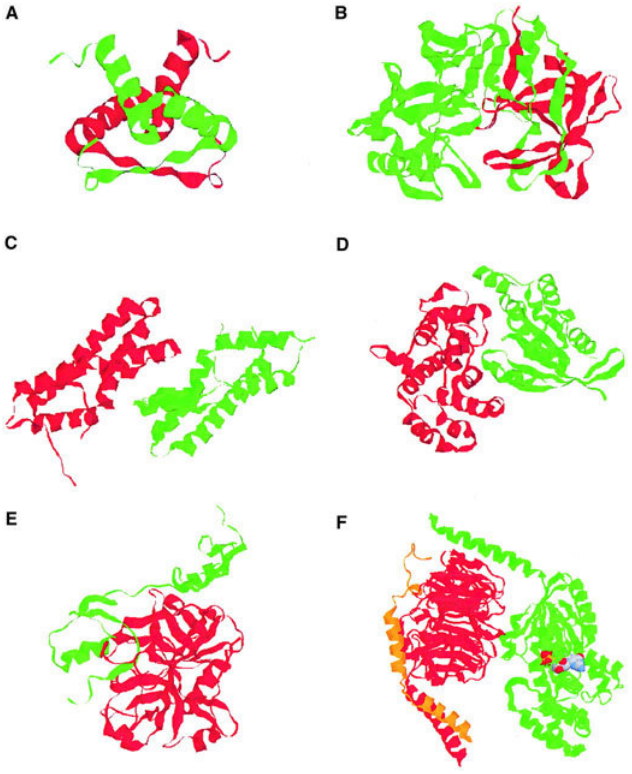


Figure 1: Examples of different types of PPIs, from Nooren and Thornton (2003). In each complex, every monomer is colored differently. (**A**) Obligate homodimer. (**B**) Obligate heterodimer. (**C**) Non-obligate homodimer. (**D**) Non-obligate heterodimer. (**E**) Non-obligate permanent heterodimer. (**F**) Non-obligate transient heterotrimer. In this example the bovine G protein contain a transient interaction between Gα (green) and G$\beta\gamma$ (red, orange).

Furthermore, complexes with interfaces bigger than $\sim$1000 Å^2 seem to undergo conformational changes in the event of an interaction (Lo Conte et al., 1999; Nooren, 2003), such as strong transient PPIs like the heterotrimeric G protein, in which Gα and G$\beta\gamma$ subunits dissociate upon

guanosine triphosphate (GTP) binding, but constitutes a stable trimer interacting with guanosine diphosphate (GDP). Interestingly, the conformational change of the interfaces is more radical in case of strong molecular triggers like phosphorylation or GDP/GTP binding, while factors like pH or temperature generate less powerful modifications and generally influence only smaller interfaces.
Finally, there is no apparent correlation between the size of the interface and its other parameters like the hydrophobicity and the binding energy, with only some exceptions (Brooijmans et al., 2002; Nooren, 2003).

Despite the advances made in recent years in understanding the structural implications and characteristics of the binding between proteins, there is still much to uncover and it is currently difficult to distinguish PPIs on the basis of structural information only. This is the reason why advanced techniques like the cryo-EM will have in the future a major role in understanding single proteins and the complexes they form.

1.1.2. Protein-RNA interactions

Protein-DNA and protein-RNA interactions (PRI) are involved in many fundamental cellular processes like translation, gene regulation, DNA damage repair and many others and these functions are carried out through the binding to different types of RNAs, like messenger RNAs (mRNAs), transfer-RNAs (tRNAs), ribosomal RNAs (rRNAs) and non-coding RNAs (ncRNAs). This is the reason why eukaryotic proteomes show a percentage of RNA-binding proteins (RBPs) of 4-13% (Jones, 2016). These interactions were first characterized through structure determination, at first with X-ray crystallography and afterwards with NMR and the more re-

cent cryo-EM, which have progressively increased the number of available complexes (Grabowski et al., 2016). However, the majority of RBPs are often deposited without RNAs bound to them and this is why there is still the need to develop increasingly advanced computational methods to simulate the docking of the structures.
In general, protein-DNA interactions have been better characterized than PRIs (Luscombe et al., 2000). This is probably due to the higher stability of the regular DNA double helix, while RNAs show a plethora of different secondary and tertiary structural patterns such as bulges, hairpin loops and pseudo-knots (Jones et al., 2001) and proteins seem to interact with RNAs exploiting these structural elements, with the establishment of non-Watson-Crick base pairing that can happen in loop regions (Nagai, 1996; Sanchez de Groot et al., 2019).

In 1999, Draper and colleagues divided protein-RNA structures into a groove binding class and a β-sheet binding class (Draper, 1999). In the first one, the protein binds in the groove of an RNA helix using an α-helix or a loop, while in the second class an unpaired portion of the RNA is bound by the protein's β-sheet. However, both these two classes of protein binding can be observed together with RNAs of different structural content, such as single-stranded RNAs (ssRNAs), single-stranded RNAs with single or multiple loops or double-stranded RNAs, with a different preference depending on the functions the interaction has to exert. In addition, protein-binding tRNAs contain domains performing both classes of binding (Jones et al., 2001).
As the number of studied complexes' structures has grown in time, new trends have emerged in analyzing interaction data. For example, Van der Waals interactions seem to be more relevant for the interactions than pre-

viously anticipated, some amino acids like asparagine, arginine, phenylalanine, tyrosine and threonine are frequent in RNA binding sites and proteins seem to have a preferential binding to guanine (Jones et al., 2001).

Interestingly, despite the many functions RBPs carry out in the cell, they show a relatively small amount of different RNA-binding modules, that are present in multiple copies and in different structural conformations to ensure a variety of binding patterns to different substrates (Lunde et al., 2007). In this way the protein can bind to an RNA combining multiple weak interactions given by the individual domains, resulting in an increased affinity and specificity, an easier regulation in case of disassembly and the possibility of binding a longer RNA stretch or multiple RNAs, arranging the RNA to perform different functions thanks to its flexibility (Sickmier et al., 2006).
In addition, the linker between two different domains is relevant for the binding process and can be either long and disordered, allowing different domains to interact with distinct targets at the same time, or short, to bind a unique long stretch of RNA, undergoing a structural change and forming an α-helix (Deo et al., 1999; Handa et al., 1999; Allain et al., 2000).
Finally, other than enhancing RNA recognition, multiple modules can also allow RBPs to bind RNA through the simultaneous binding with other proteins, for example through a dimerization process. These domains can also work together with enzymatic domains to help the regulation and catalysis of their activity, cooperating in recognizing the RNA through a series of weak protein-protein and protein-RNA interactions (Lunde et al., 2007).
The RNA binding domains (RBDs) can vary in composition and interac-

tion patterns and can be divided into different categories (**Figure 2**).

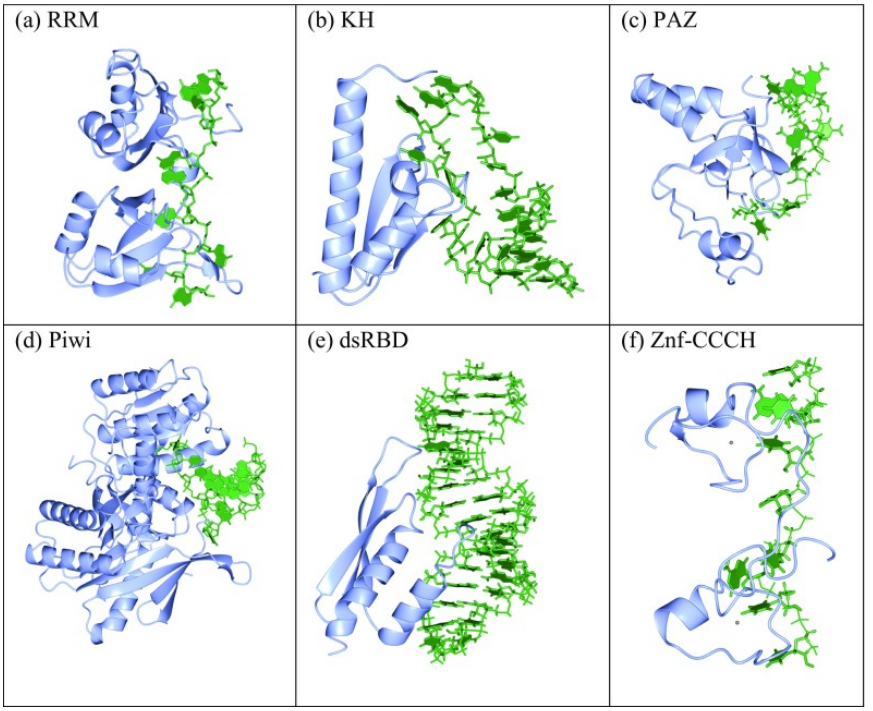


Figure 2: Examples of different types of RNA binding domains, from Jones (2016), obtained with X-ray crystallography and NMR. In each figure, the RNA is shown in green highlighting the different bases, while proteins are represented in blue depicting their secondary structure conformation. (**A**) RNA recognition motif (RRM). (**B**) hnRNP K homology (KH) domain. (**C**) PAZ domain. (**D**) PIWI domain. (**E**) Double-stranded RNA-binding domain (dsRBD). (**F**) Zinc finger CCCH domain.

- The **RNA recognition motif (RRM)** is the most common among RBDs and is involved in many post-transcriptional processes. It is 80-90 amino acids long and it has the $\beta\alpha\beta\beta\alpha\beta$ topology, with the two central β-sheets generally binding RNAs, thanks to two ribonucleoprotein motives containing three Arg/Lys and two aro-

matic residues that mediate the binding (Oubridge et al., 1994). A single RRM in general can only bind 4-8 nucleotides, so the coordination of multiple domains is necessary to ensure binding specificity (Auweter et al., 2006).

- The **hnRNP K homology (KH) domain** can bind both to ssDNA and ssRNA (Backe et al., 2005). It is $\sim$70 amino acids long and can have two alternative topologies, $\beta\alpha\alpha\beta\beta\alpha$ or $\alpha\beta\beta\alpha\alpha\beta$, which can both bind four nucleotides thanks to the GXXG loop (Grishin, 2001). Unlike RRMs, the binding is mediated by electrostatic interactions and hydrogen bonds and not by aromatic residues.

- The **double-stranded RNA-binding domain (dsRBD)** is another domain of 70-90 amino acids with an $\alpha\beta$ topology and it is widespread in eukaryotes and bacteria. It differs from the previous RBDs because the binding is shape-specific and mostly sequence-independent and requires only a sugar-phosphate backbone interaction between the proteins α-helices and the dsRNA helix (Ryter and Schultz, 1998).

- The **zinc fingers** are domains that can bind both DNA and RNA, they are repeated several times within a protein and they are classified according to the residues coordinating the zinc (Carballo et al., 1998). For example, the C2H2 zinc finger can form electrostatic contacts with RNA loops using the protein recognition α-helices located on the fingers (Wolfe et al., 2000), while the CCCH zinc finger establishes direct hydrogen bonds with RNA bases (Lai et al., 2000).

- The **S1 domain** was first identified in the ribosomal protein S1 but it is present in several exonucleases (Subramanian, 1983). It is 70 amino acids long, distributed in five antiparallel β-barrels and one short α-helix (Bycroft et al., 1997). The binding pattern is similar to RRMs, with the two central β-sheets that can interact with RNA thanks to the contribution of several aromatic residues and the surrounding loops and secondary structure elements (Bycroft et al., 1997; Schubert et al., 2004).

- **PAZ domains** are 110 amino acids long and are formed by a β-barrel resembling an S1 domain and a small $\alpha\beta$ motif (Yan et al., 2003), creating a clump-like structure that recognizes 2-nucleotides overhangs at the 3' of small interfering RNAs (siRNAs) through hydrogen bonding (Ma et al., 2004; Macrae et al., 2006).

- **PIWI domains** have a highly conserved binding pocket containing a metal ion that can recognize the 5' phosphate group of siRNAs (Parker et al., 2004, 2005).

- **Pumilio domains** are an example of tandem RBDs. Each motif can recognize one nucleotide, but with multiple repeats, the protein can bind up to eight nucleotides with high specificity and affinity thanks to hydrogen bonds between the RNA and two residues in a domain's α-helix (Wang et al., 2002).

- **TRAP domains** can recognize a GAG triplet through hydrogen bonding and stacking interactions provided by β-sheets, which constitute the eleven subunits of the ring-shaped Tryptophan Regulated Attenuation Protein (TRAP) (Antson et al., 1999).

- **SAM domains** constitute a hydrophobic cavity within three helices, which is able to recognize the RNA stem-loop shape through interactions with a base in the loop and the sugar-phosphate backbone (Oberstrass et al., 2006).

Despite the different ways in which individual domains can interact with RNAs and the enhanced binding ability given by the presence of multiple copies of these motives inside proteins, there are still relatively few available structures of proteins containing multiple RBDs and new structural analyses will be necessary to expand our knowledge on the number of combinations and functions provided by these domains.

1.1.3. RNA-RNA interactions

RNA-RNA interactions (RRIs) are another important type of macromolecular interactions important for different biological functions. In particular, RNAs can exploit their own flexibility to connect different parts of the same molecule through intramolecular interactions, forming complex secondary and tertiary structure conformations that shape the RNA architecture, such as loops, hairpins and pseudo-knots, or they can contact other RNA molecules, either directly through base-pairing or thanks to proteins' mediation (Xue, 2022). For example, intramolecular contacts among loops are necessary to shape the 28S rRNA (Cai et al., 2020), while instead small nuclear RNAs (snRNAs) can bind intronic regions of precursor mRNAs, sometimes through direct base-pairing in essential steps of splicing or for RNA interference (Lee et al., 1993; Valadkhan and Manley, 2001) and sometimes indirectly, like in the case of lncRNA

Malat1, that seems to interact directly with pre-mRNAs through the mediation of different proteic factors (Engreitz et al., 2014). RNA-DNA interactions are also possible and regulate several phenomena such as the X-chromosome inactivation by the hand of lncRNA Xist (Penny et al., 1996; Cerase et al., 2019).
Interestingly, the two types of RRIs can be achieved not only by Watson-Crick base pairing but also by non-canonical ones (e.g. G-U pairing that occurs in RNA secondary strcture and tRNA recognition) and base stacking between single-stranded regions or co-axial stacking of helices (Zanchetta et al., 2008).

Many RRIs inside the cell originate from the combination of intra- and inter-molecular interactions that coordinate to achieve a specific function and different strategies are available nowadays to tackle RRIs identification.
In particular, studying intra-molecular RRIs require the mapping of the RNA secondary structure elements. This can be achieved through either enzymatic probing or chemical probing methods and can be further classified according to their *in vitro* or *in vivo* application.
Among the enzyme-based *in vitro* techniques, the most important are the Parallel Analysis of RNA Structure (PARS) (Kertesz et al., 2010), the Fragmentation Sequencing (FragSeq) (Underwood et al., 2010), the Parallel Analysis of RNA structures with Temperature Elevation (PARTE) (Wan et al., 2012) and the Protein Interaction Profile sequencing (PIP-seq) (Silverman et al., 2014).
In the PARS technique, the RNAs are divided into two pools and are treated with two different enzymes. The first pool is digested with RNase S1 that cleaves single-strand sequences, while the second pool is cut with

RNase V1 in double-stranded regions. The resulting segments are then randomly fragmented and sequenced (Kertesz et al., 2010).
FragSeq is a simpler and faster protocol compared to PARS, where the RNA is cleaved with RNase P1 in single-stranded regions and reverse-transcribed. Then, the cDNA segments are sequenced and the structure of RNA can be mapped by looking at the nuclease digestion sites (Underwood et al., 2010).
PARTE is a technique very similar to PARS, in which the RNA is using RNase V1 at different increasing temperatures in order to assess the RNA folding energies (Wan et al., 2014).
Finally, PIP-seq is a technique to find RBPs binding sites on RNAs in crosslinked or non-crosslinked cells, to understand which regions bind to RBPs and which are insensitive to RNase, in order to obtain information on both protein binding and RNA structures. Initially, cross-linked cells (with UV or formaldehyde) and uncross-linked cells are lysed and divided in an experimental set and an RNase insensitivity control. The first group is treated with dsRNases or ssRNases and treated with proteinase K to remove the RBPs, while the second group is first treated with the proteinase and then with RNases. Then the fragments are reverse-crosslinked and used for strand-specific sequencing (Silverman et al., 2014).
In general, these enzymatic techniques can provide a complete map of single and double-stranded regions but being *in vitro* they can have only limited resolution and they do not take into account potential binding to other proteins that can happen in the cellular environment (Piao et al., 2017; Nguyen et al., 2018).

Chemical methods instead comprehend *in vitro* techniques like Chemical Inference of RNA Structures (CIRS-seq) (Incarnato et al., 2014) and

Selective 2'-hydroxyl Acylation analyzed by Primer Extension and Mutational Profiling (SHAPE-MaP) (Siegfried et al., 2014), while examples of *in vivo* methods are Structure-seq (Ding et al., 2014) Mod-seq (Talkish et al., 2014), Dimethyl Sulfate Sequencing (DMS-seq) (Rouskin et al., 2014) and in vivo Click Selective 2-hydroxyl Acylation And Profiling Experiment (icSHAPE) (Spitale et al., 2015).

CIRS-seq starts with proteinase K treatment to remove proteins bound to RNAs, leaving their secondary structure intact. Then, DMS and N-cyclohexyl- N'-(2-morpholinoethyl) carbodiimide metho-p-toluenesulfonate (CMC) are applied to methylate As and Cs and selectively modifying pseudouridines when the RNA is in single-stranded conformation. Then, the RNA is reverse-transcribed and sequenced. In this way, DMS and CMC can be exploited to identify the locations of secondary structure elements (Incarnato et al., 2014).

SHAPE-MaP has the advantage of retrieving information about RNA secondary structure at a massive scale at single-nucleotide resolution, it is very customizable to investigate different types of RNAs and can target all four nucleotides. It employs the 1-Methyl-7-nitroisatoic anhydride (1M7) to identify secondary RNA structures by binding to the ribose 2'-OH groups. Then, mutational profiling (MaP) is performed to induce non-complementary nucleotide mutations at the moment of reverse transcription. Finally, the sequences are aligned to obtain profiles and study the mutations' position (Siegfried et al., 2014).

Structure-seq and DMS-seq are techniques that can assess RNA secondary structure both *in vitro* and *in vitro* at single-nucleotide resolution and use dimethyl-sulfate to modify the base-pairing faces of Cs and As in RNA loops. The RNA is then reverse-transcribed to obtain a single-stranded cDNA that is PCR amplified and sequenced (Ding et al., 2014; Rouskin

et al., 2014).
Mod-seq is similar to the previous two in the use of DMS but introduces a new feature to allow a high-throughput profiling even of very long RNAs (Talkish et al., 2014).
Finally, icSHAPE is useful to assess PRIs and m^6A modifications *in vivo*. In this protocol, RNA secondary structures are modified by adding a 2-methylnicotinic acid imidazolide (NAI) probe called NAI-N3 and they are marked and subsequently tagged with dibenzocyclooxtyne (DIBO)-biotin after the cell's lysis. Then, the RNA is reverse-transcribed and tagged RNAs are captured by streptavidin beads and subsequently sequenced (Spitale et al., 2015).
In general, these chemical-based techniques have the advantage of being applicable in *in vivo* environments and being able to work at single-nucleotide resolution, but often have the cons of limiting the analysis to only two nucleotides out of four and tackling relatively short RNAs. Combining both chemical and enzymatic approaches could cover some of their specific weaknesses and lead to a major understanding of intramolecular RRIs and RNA secondary structure elements (Nguyen et al., 2018).

In addition to intramolecular interactions, several approaches have been developed to investigate the interactions between different RNA molecules. The first category of these methods are low-throughput approaches, which aim to study and validate specific RRIs, previously predicted through computational methods (Gong et al., 2018). In this case, the RRI can be tested with several biochemical methods.
The first approach is Surface Plasmon Resonance (SPR), a technique in which a studied RNA segment is immobilized on a sensor chip through the interaction between biotin and streptavidin in order to monitor the po-

tential binding with a series of RNA candidates in real-time (Di Primo et al., 2011).
The Electrophoretic Mobility Shift Assay (EMSA) is instead applied to RNA fragments extracted from cells or synthesized to investigate their interaction. Since an interacting RNA pair has greater molecular mass than non-paired transcripts, it will migrate slower in the gel (Bak et al., 2015).
The RNA immobilization is also at the base of single molecule Forster Resonance Energy Transfer (FRET), in which the RNAs molecules are either encapsulated in lipid vesicles or fixed on a quartz surface to be real-time monitored thanks to fluorescent dyes positioned on specific regions of the RNAs. The system is able to produce a signal if an interaction occurs (Yu et al., 2015).
In another technique, the co-sedimentation assay, a mixture of different RNAs are subjected to sucrose or glycerol gradient and interacting RNAs in the same gradient fractions are revealed with Northern blot (Liang and Fournier, 2006).
Finally, the yeast RNA hybrid system is a cellular method developed in Saccharomyces Cerevisiae, in which a reporter gene is activated and expressed only if a bait and a prey RNA interact together (Piganeau et al., 2006).

However, these techniques are very specific for a particular RRI and usually are not able to pinpoint the binding site region.
This led to the need to develop high-throughput techniques for the analysis of RRIs. These approaches usually employ cross-linking and proximity ligation to turn the interacting strands into a chimeric RNA, which is then sequenced to understand the underlying interaction (Nguyen et al., 2018).

The first examples of these methods were targeted techniques, to identify RRIs mediated by a protein or a specific RNA, like Cross-linking Ligation And Sequencing of Hybrid (CLASH) (Kudla et al., 2011) and RNA Hybrid and individual-nucleotide resolution ultraviolet Cross-Linking and ImmunoPrecipitation (HiCLIP) (Sugimoto et al., 2015).
In CLASH, RNA duplexes are UV cross-linked to proteins and affinity-purified and the resulting RNA-RNA hybrids are ligated and reverse-transcribed to cDNAs for the sequencing, revealing RRIs as chimera reads mapped to the two transcripts at high-resolution (Kudla et al., 2011).
HiCLIP is an approach similar to CLASH, in which RNAs are cross-linked with proteins to obtain a duplex and then immunoprecipitated. Then, a specific adapter is ligated to both strands of the duplex and the 3' end of the adapter is bound to the 5' of the other strand, followed by the removal of the proteins with proteinase K and preparation for cloning. The advantage of this technique is that the RNA duplex can be formed by the same RNA or two different transcripts (Sugimoto et al., 2015).

Despite being an upgrade compared to low-throughput techniques, these methods can only identify RRIs for one target RNA or interactions between specific transcripts. This is why a huge improvement was obtained with the introduction of transcriptome-wide methods, that potentially could cover all interactions present in a cell. The most famous of these techniques are shown in **Figure 3** and are Psoralen Analysis of RNA Interactions and Structures (PARIS) (Lu et al., 2016), Sequencing of Psoralen cross-linked, Ligated, And Selected Hybrids (SPLASH) (Aw et al., 2016), LIGation of Interacting RNA followed by high-throughput sequencing (LIGR-seq) (Sharma et al., 2016), MAapping RNA Interactome and structure in vivO (MARIO) (Nguyen et al., 2016) and, more

recently, the RNA In situ Conformation sequencing (RIC-seq) (Cai et al., 2020).

In PARIS, RNAs are cross-linked in living cells using an organic compound, a psoralen derivative called 4'-aminomethyltrioxsalen (AMT) and the duplexes are collected with a 2D electrophoresis, then proximity ligated, reverse-cross-linked and sequenced (Lu et al., 2016).

LIGR-seq is similar to PARIS, but the cross-linked RNA duplexes are digested with nuclease S1, circularized and then treated with RNase R for the following reverse-cross-linking and sequencing (Sharma et al., 2016).

In SPLASH the cross-linking process is obtained with biotinylated psoralen, then streptavidin-coated beads are used to obtain the duplexes, which are again proximity-ligated, reverse-cross-linked and sequenced (Aw et al., 2016).

MARIO differs from the other three techniques because it is an approach designed to identify RNA interactions mediated by proteins. Instead of using psoralen derivatives, it starts with UV cross-linking the RNAs with proteins, followed by protein denaturation with RNase I, biotinylation, the pull-down of the formed complex and the proximity ligation of the RNAs with a biotinylated RNA linker. These chimeric RNAs are then purified and sequenced (Nguyen et al., 2016).

Finally, RIC-seq is another technique to identify protein-mediated RRIs developed to facilitate the generation of 3D RNA interaction maps. It starts again with the cross-linking process on formaldehyde, then the RNAs are randomly cut and RNAs 3' end is labeled with pCp–biotin, followed by proximity ligation *in situ* of RNAs in close proximity without denaturation. In this way, the biotin is located at the junction of two different RNAs. Later the RNAs are extracted and fragmented and the ones containing biotin are enriched and sequenced.

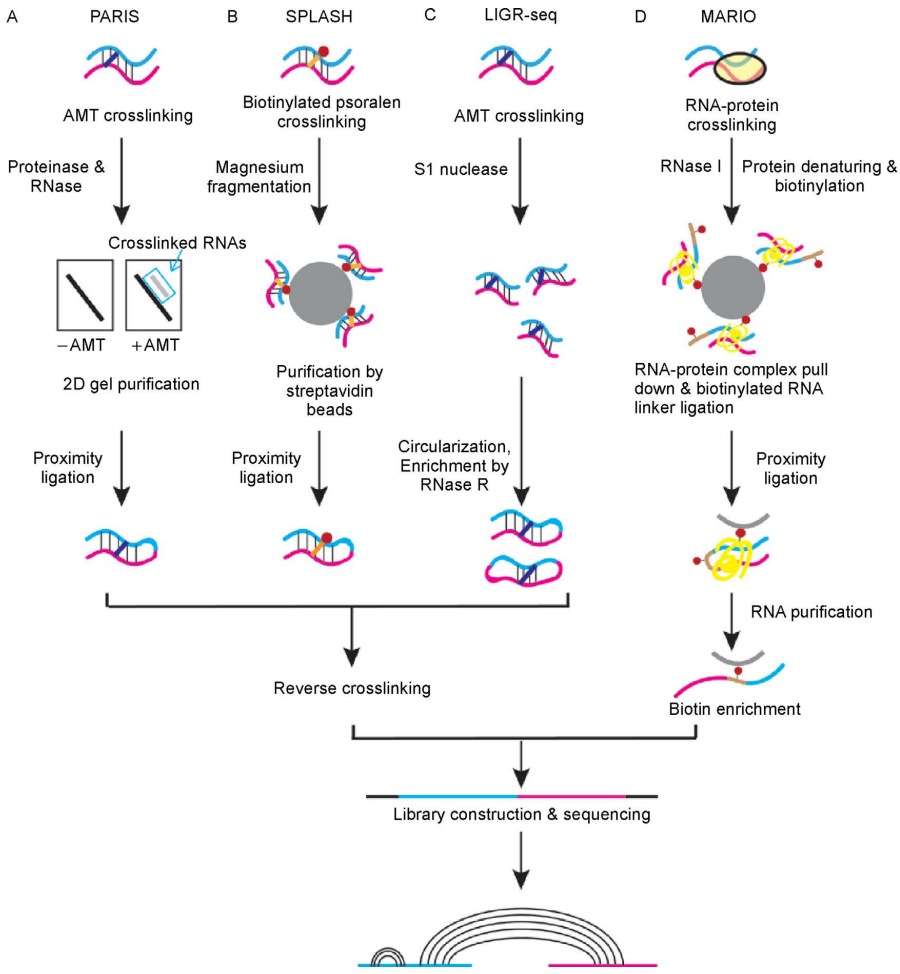


Figure 3: Transcriptome-wide methods for RRIs detection, from Gong et al (2018). (**A**) In the PARIS technique interacting RNAs are cross-linked with AMT and purified using 2D electrophoresis, followed by proximity ligation, reverse cross-linking, reverse transcription and sequencing. Then, chimeric reads mapped to two different transcripts are exploited to identify the RRIs. (**B**) SPLASH is similar to PARIS, but in this case, the RNAs are cross-linked with biotinylated psoralen and purified with streptavidin-coated beads. (**C**) LIGR-seq is another method similar to PARIS, but after cross-linking RNA duplexes are circularized and treated with RNase R for enrichment. (**D**) MARIO is used to identify RNA interactions mediated by proteins. RNAs are initially cross-linked to the proteins, followed by protein denaturation, biotinylation, pull-down and the subsequent proximity-ligation of RNAs with a biotinylated linker and final sequencing.

The use of proximity-ligation *in situ* and in non-denaturing conditions seems to reduce the number of false positives compared to other techniques and has led to the production of a vast collection of high-confidence interactions (Cai et al., 2014; Xue, 2022).

In general, all these methods have helped to increase our knowledge of RRIs at the cellular level and have the advantage of identifying the binding regions at very high resolution. One disadvantage of using psoralen derivatives for cross-linking relies on their preferential activity for pyrimidines (Nawy, 2016), so combining different methods and predictive tools could be a way to reach an even more comprehensive view of direct and mediated RRIs in the cell.

1.2. Ribonucleoprotein condensates

The importance of macromolecular interactions in regulating countless biochemical pathways and cellular activities is particularly relevant in the context of biomolecular condensates, compartments with specific functions in the cell. Even though some of these organelles are physically separated from the external environment with a physical layer such as the nucleus, some are membrane-less and form and organize thanks to a physiological process of phase-separation, in which a single-phase molecular complex separates into a more concentrated phase and a more diluted one. This process is achieved through free energy minimization of the solution thanks to the maximization of weak inter and intra-molecular interactions established among its elements (Alberti et al., 2019; Zbinden et al., 2020). In particular, phase separation is triggered when an increase in the con-

centration of biopolymers reaches a specific saturation concentration limit and depends on environmental variables like temperature, pH and the salt type and it is responsible for moving macromolecules inside liquid, gel or solid-like compartments (Wolf et al., 2014; Munder et al., 2016; Verdile et al., 2019; Garaizar et al., 2022).

Phase separation is a phenomenon not limited to mammals, but can be found in numerous organisms such as yeast and C. Elegans but also in bacteria, fungi and protozoa (Azaldegui et al., 2021; Tweedie and Nissan, 2021).
While liquid-liquid phase separation (LLPS) usually forms reversible structures in response to a critical rising in concentration or temperature, a liquid-solid phase transition is generally an irreversible state that creates aberrant aggregates, such as amyloids, responsible for several neurodegenerative diseases (Hyman et al., 2014; Chiti and Dobson, 2017; Wan et al., 2018; Verdile et al., 2019).
Among the human liquid-like compartments, the term ribonucleoprotein (RNP) granules has been used to address those organelles with a high concentration of proteins and RNAs and can be located both in the nucleus and in the cytoplasm, providing a spatiotemporal control of biological activities (**Figure 4**) (Matera, 1999; Protter and Parker, 2016; Shin and Brangwynne, 2017; Verdile et al., 2019).

Among nuclear granules, the most important are Cajal Bodies, paraspeckles and nucleoli and more recently a mechanism of phase-separation has been proposed also for super-enhancer (Verdile et al., 2019).

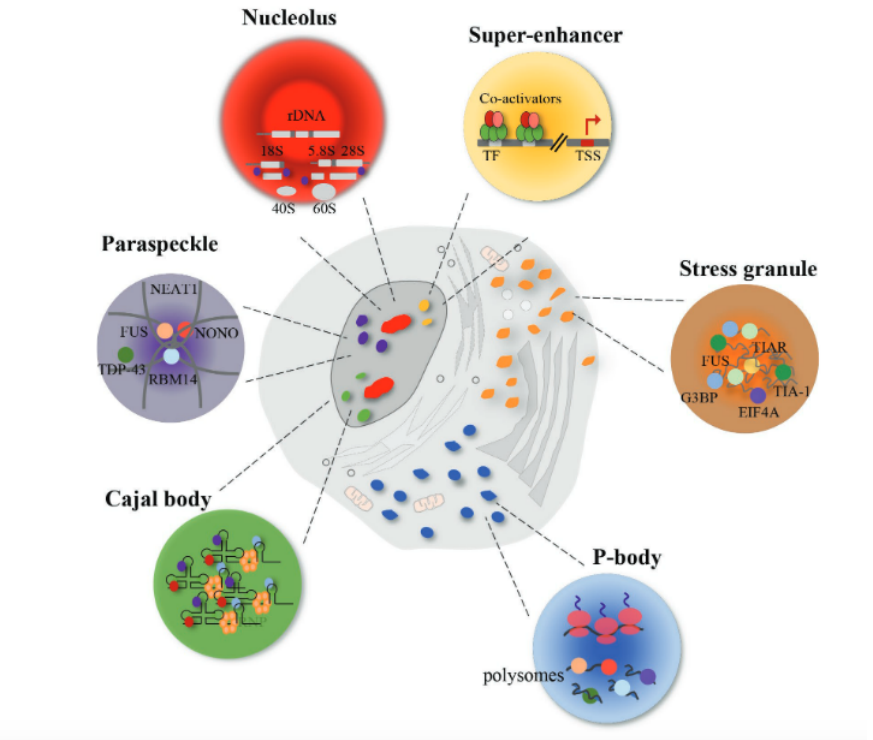


Figure 4: Schematic representation of RNP granules (adapted from Verdile et al. (2019)). Different types of RNP granules can form in the cell's nucleus or cytoplasm.

- **Cajal bodies** are involved in small nuclear SNP (snRNP) and small nucleolar RNP (snoRNP) biogenesis and recycling, as well as other functions such as telomere maturation and spliceosome formation. An essential factor in these assemblies is the coilin protein, capable of aggregating other proteins and RNAs (Gall et al., 1999; Machyna et al., 2013, 2014).

- **Paraspeckles** control DNA repair and gene expression and are generated around the lncRNA NEAT1, which acts as a scaffold for sev-

eral RBPs (Fox et al., 2002; Souquere et al., 2010). These condensates show a subdivision in a denser core and a more diluted outer shell, where the first one forms around the central part of NEAT1, while the second one is generated around the 5' and 3' ends of the RNA. Among the proteins bound by NEAT1, there are FUS (localized in the core), TDP-43 (in the shell) and the family of splicing proteins NONO, RBM14 and PSPC1 (Fox et al., 2002; Hennig et al., 2015).

- **Nucleoli** are responsible for ribosome biogenesis (Andersen et al 2002). They are divided into three subregions: the first one is involved in the transcription of rRNAs and is enriched in RNA polymerase I, the second one is destined for the processing and modification of pre-rRNA transcripts, while the third region, enriched in proteins, is where the ribosomes' assembly occurs (Boisvert et al., 2007).

- **Super enhancers** are clusters of transcriptional enhancers assembled with a high level of binding to master transcription factors and co-activators, together with RNA and RNA polymerase II and this complex drives gene expression (Whyte et al., 2013; Hnisz et al., 2017). Reversible chemical modifications and phase separation can modify the interactome and the activity of these assemblies both in healthy and pathological cellular states.

In the cytoplasm, two of the most studied condensates are stress granules and processing bodies.

1.3. Stress granules and processing bodies

Stress granules (SGs) and processing bodies (PBs) are two types of RNP granules located in the cytoplasm, forming in response to an accumulation of untranslated RNAs in the cell due to a stress condition or a phenomenon that represses translation initiation (i.e drugs, over-expression of repressors, knockdown of translation initiation factors) (Kedersha et al., 2005; Franks and Lykke-Andersen, 2008; Decker and Parker, 2012; Jain et al., 2016).

While SGs' proposed function is mostly related to RNAs' protection from harmful conditions and they contain many translation initiation components, PBs are more involved in post-transcriptional regulation and contain members of the mRNA decay machinery (Kedersha et al., 2005). This has led to the hypothesis that these organelles could have a function related to mRNA degradation (Decker and Parker, 2012; Protter and Parker, 2016; Verdile et al., 2019), even though others disagree, stating that decay factors are not essential for PBs assembly, there is no detectable change in RNAs' decay level upon PBs dissolution and 5' truncated RNAs were not detected, suggesting that these mRNAs could be accumulating inside PBs just for storage upon events of the impaired decay process (Parker and Sheth, 2007; Hubstenberger et al., 2017).

Both organelles can contribute to removing factors and RNAs from the cytosol to increase their local concentration inside the condensates, influencing the equilibrium of the cell's resources and are involved in other functions such as RNA localization in neurons and embryos (Kiebler and Bassell, 2006; Decker and Parker, 2012).

As discovered by Fluorescence Recovery After Photobleaching (FRAP) technique, these assemblies are dynamic entities, capable of exchanging

components with the surrounding cytoplasm and they often physically interact in mammals (Kedersha et al., 2005; Andrei et al., 2005) (**Figure 5**). This plasticity seems to be regulated by several types of ATPases that exploit ATP to promote changes in their inner interactions and composition, implying that SGs are not really uniform structures (Jain et al., 2016). The exchange rate of components is also enhanced thanks to the spherical shape of these condensates (Johansson et al., 1998; Verdile et al., 2019).

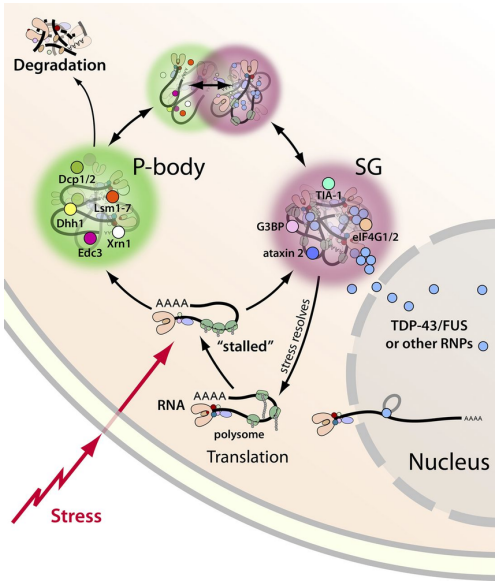


Figure 5: Stress granules and P-bodies can interact with each other, from Li et al. (2013). Following a stress condition, stalled mRNAs resulting from stopped translation can lead to the formation of SGs or PBs, that can co-exist and exchange components with each other, recruiting factors from the surroundings. RNAs can then be degraded or can resume translation if a physiological condition is re-established.

Interestingly, while some proteins and RNA are in common between the

two types of condensates, the key elements that govern their formation are different.
Like other types of liquid-like condensates, SGs and PBs are reversible entities and are usually disassembled upon clearance of the initial stimuli, with a consequent resume of the translation process or autophagy of damaged molecules (Bhattacharyya et al., 2006).

1.3.1. Formation and structure

SGs are not uniform structures, they range from 0.1 to 2 μm and contain denser and stable regions referred to as cores of around 200 nm, where the concentration of proteins and RNAs is higher, and a more sparse shell, potentially more dynamic (Protter and Parker, 2016; Jain et al., 2016). The number of cores directly correlates with the SG's volume, they show higher levels of G3BP1, PABP1 and poly(A+) RNAs and they are more resistant to chemical disruption than the shell, suggesting stronger macromolecular interactions within them.
PBs instead are constitutive assemblies of untranslated RNAs and members of the decay machinery that can exist even in unstressed conditions. They are in the range of 500 nm and their size shifts in presence of stress, they are limited to 4-7 units per cell and show in general a denser interactions network compared to SGs (Hubstenberger et al., 2017).

A first complete overview of the formation of these condensates focused on the proteins' ability to bind other proteins and untranslated mRNAs through their intrinsically disordered regions (IDRs) and/or prion-like domains (PrLDs), coupled with the post-translational modifications of proteins (mainly methylation, phosphorylation and glycosylation) and ex-

ploiting the microtubules as motors for the recruitment of these mRNPs inside the organelles (Protter and Parker, 2016; Jain et al., 2016).

In PBs, after the failure of translation initiation or termination, a complex including proteins Dcp1, Dcp2, Dhh1 and other variable factors are recruited together with the group of Pat1, Xrn1 and Lsm17p proteins onto untranslated mRNA that act as a scaffold, forming an aggregate gradually growing in size thanks to the Pat1 protein multivalency and the mediation of PrLDs, which recruit decapping enzymes (Anderson and Kedersha, 2006).

SGs instead seem to form when a stress condition (DNA damage, heat shock, oxidative stress, chemical shock, etc.) blocks the translation process, either by phosphorylation of eIF2α or inactivating eIF4A, followed by the interaction of untranslated mRNAs to PrLDs of TIA-1, TIA-R and G3BP1 proteins, that seems to trigger SG formation in response to free cytosolic RNA concentrations and it constitutes the central node of granule protein-RNA interactions network (Mokas et al., 2009; Yang et al., 2020). Interestingly, according to the specific condition of the cell, mRNAs can be recruited inside one type of organelle and can later resume translation or become part of other types of condensates through mRNP rearrangements.

Even though the first description of the formation of these condensates was already accurate, there was still confusion about the possible role of RNA-RNA interactions and, in the case of SGs, the order of formation of the core and shell substructure was unknown: a first model suggested that SG formation started with initial nucleation of oligomers gradually

increasing in size to become the cores with a final fusion and the recruitment of the outer shell; the second model instead began with initial phase-separated droplets of untranslated mRNPs, linked together by weak molecular interactions, which further grow in a second stage to form cores thanks to a higher concentration of proteins and RNAs (Protter and Parker, 2016).

In the last years, new details have emerged, expanding our knowledge of SGs and PBs. In the first place, the RNAs were discovered to have a much bigger role in the phase-separation process than previously anticipated, providing a scaffolding platform for other proteins and RNAs that are recruited. Firstly, RNAs can lower the concentration threshold that leads to RBPs phase-separation by binding to them, exemplified by G3BP1 role in partitioning mRNAs inside SGs (Matheny et al., 2021); secondly, a protein-free transcriptome is able to generate droplets, implying a fundamental role of RNA-RNA interactions (Su et al., 2021); thirdly, RNAs *in vitro* usually need a lower concentration to condense compared to disordered proteins (Van Treeck et al., 2018; Campos-Melo et al., 2021).
This generates a possible model in which not only specific proteins are necessary for granule assembly, but their formation depends also on some RNAs with special functions or characteristics. One of these properties is the RNA secondary structure profile, responsible for their interactions patterns. These RNAs could establish a plethora of RNA-RNA interactions and, in parallel, could act as a scaffold for important granule proteins needed for nucleation (Campos-Melo et al., 2021).
This is the reason why condensates nowadays are believed to be the result of a combination of multivalent low-affinity interactions that include protein-protein, protein-RNA and RNA-RNA interactions, exploiting RNA

flexibility, sequence and structural patterns as well as RNA binding domains inside proteins (Matheny et al., 2021; Su et al., 2021).
Interestingly, the high number of molecular interactions present inside these condensates is also seen in unstressed cells, implying an underlying pre-network of interactions that could increase the aggregation speed of these organelles (Markmiller et al., 2018).

In addition, RNA post-transcriptional modifications seem to be important in the regulation of phase separation, among which the N6-adenosine methylation (m^6A), a reversible change catalyzed by METTL enzymes with the help of other RNA binding proteins able to modify the RNA structure and the interactions patterns. In recent studies, RNAs undergoing the m^6A modification seem to be able to act as a scaffold to partition YTHDF proteins inside SGs, thanks to the previously described combination of macromolecular interactions, where they can exert multiple functions related to RNA degradation or translation initiation of other m^6A-modified RNAs (Shi et al., 2017; Ries et al., 2019). Furthermore, m^6A can change granule dynamics, for example preventing core proteins like G3BP1 or CAPRIN1 to bind RNAs (Arguello et al., 2017). In particular, this modification was found particularly enriched in tRNAs, rRNAs and at the 3'UTR in mRNAs, with functions related to the regulation of localization, stability, translation and splicing (Wang et al., 2014; Meyer and Jaffrey, 2014; Campos-Melo et al., 2021).
Another example is represented by N1-adenine (m^1A) methylated RNAs, which seem to accumulate in SGs during heat shock or oxidative stress (Su et al., 2021; Alriquet et al., 2021).

1.3.2. Transcriptome composition

Regarding the transcriptome composition of these condensates, on average around 10% of the total human transcriptome can be recruited inside SGs (Jain et al., 2016; Cid-Samper et al., 2018) but, even though nearly every expressed gene can be found inside SGs, there are no RNA species representing more than 1% of the SG transcripts. It is estimated that around 78–95% of SG composition is made of RNAs (Khong et al., 2017). In general, around 80% of SG RNAs are coding transcripts, while the 20% of ncRNAs consist mostly of snRNAs and snoRNAs, found also to be important in Cajal bodies formation, and of a few highly contacted long-non-coding RNAs such as NEAT1 and NORAD that are known scaffolding RNAs, probably involved in macromolecular interactions that aid condensates formation (Khong et al., 2017; Cid-Samper et al., 2018).

In 2017, a comprehensive SG transcriptome was published, obtained through RNA-seq analysis of purified SG cores isolated from U-2 OS cells upon arsenite exposure, combined with a single-molecule fluorescence in situ hybridization (smFISH) validation (Khong et al., 2017). The authors of the study estimated each core to contain around 21 to 106 mRNA molecules, suggesting that the heterogeneity of RNAs recovered is due to the high variability of transcripts among different cores.
Notably, SG transcriptome can also vary under different stress conditions, probably because of stress-specific translation repression processes. In this report, the SG transcriptome was divided into enriched and depleted RNAs, comparing their concentration inside these condensates with the cytoplasm's one. In particular, enriched SG RNAs appear to have much longer coding sequences and 3'UTRs compared to depleted RNAs, with

an average length of 7.5 kb, suggesting that longer RNAs could be more easily recruited inside the condensates, probably thanks to a higher number of possible binding sites, creating a multivalency of interactions with other proteins and RNAs (Jain et al., 2016). Furthermore, RNAs enriched inside SGs seem to have a shorter half-life, are more AU-rich and are more prone to be m^6A modified (Khong et al., 2017; Anders et al., 2018). Interestingly, the isolation of SG mRNAs through Photo-Activatable Ribonucleoside Cross-Linking and Immunoprecipitation (PAR-CLIP) experiments showed that under oxidative stress around 50% of SG transcripts contain the m^6A modification (Anders et al., 2018).

PBs transcriptome composition is very similar to SGs, with enrichment of long, poorly translated and AU-rich RNAs, even though their AU composition is usually higher in PBs than SGs (Hubstenberger et al., 2017). More than one-fifth of the cytoplasmic transcripts can accumulate inside PBs, with a very high proportion of protein-coding transcripts. In particular, there is an enrichment in mRNAs encoding for histone modifiers, regulators of the ubiquitin pathway and factors involved in several processes such as cell division and chromatin remodeling.
On the other hand, PBs seem to be depleted in RNAs coding for mitochondrial elements, translation machinery components and rRNAs, immunity factors metabolic pathways enzymes (Hubstenberger et al., 2017).
Furthermore, PBs transcriptome has a shift in their composition according to the presence or absence of a stress condition. Interestingly, a study comparing unstressed HEK293 cells with stressed U-2 OS cells revealed that stressed PBs seem to increase the percentage of long RNAs and decrease the amount of AU-rich RNAs, constituting a transcriptome that is more similar to the SGs one, compared to an unstressed condition. This

highlights the potential role of the RNA length in defining the composition of these condensates in a stressful situation when the majority of mRNAs are translationally repressed (Matheny et al., 2019).

1.3.3. Proteome composition

SGs and PBs often physically dock together and exchange components between them. This is the reason why, despite having different functions and containing unique factors, several components of their transcriptome and proteome are in common and usually localize at the interface between the two (Buchan and Parker, 2009; Jain et al., 2016).

In general, SGs proteome contains a great percentage of RBPs, several translation initiation factors, DNA/RNA helicases, ATPases and several enzymes for the transfer of methyl and glucosyl groups (Jain et al., 2016; Markmiller et al., 2018; Kuechler et al., 2020). In particular, TIA-1 and G3BP1 are two RBPs that are necessary to start SGs nucleation interacting with free RNAs (Tourrière et al., 2003; Gilks et al., 2004), even though the absence of one or the other could still lead to SG formation according to the type of stress condition (Kedersha et al., 2016).
SGs can be divided into canonical or non-canonical depending on their composition. Canonical SGs contain several pro-apoptotic factors that are sequestered inside them to prevent signal cascades leading to the cell's programmed death, while non-canonical SGs lack these components and this control mechanism (Fujimura et al., 2012; Aulas et al., 2018; Reineke et al., 2018; Reineke and Neilson, 2019). Some of these factors are the receptor of activated protein C kinase 1 (RACK1), ribosomal S6 kinase 2 (RSK2), TNF receptor-associated factor 1 (TRAF2), histone deacety-

lase 6 (HDAC6), mitogen-activated protein kinase 7 (MKK7) and Ras homolog family member A (RhoA) (Kim et al., 2005; Kwon et al., 2007; Eisinger-Mathason et al., 2008; Tsai and Wei, 2010; Wasserman et al., 2010). Interestingly, the formation of canonical or non-canonical SGs seems to be dependent on the type and duration of the stress. In the case of chronic nutrient starvation stress, for example, SGs do not show the presence of RACK1 protein (Reineke et al., 2018), which instead is recruited inside these condensates in the case of acute oxidative stress, promoting cell survival (Arimoto et al., 2008; Reineke et al., 2018; Park et al., 2020). Specific stressors can also determine the formation of non-canonical SGs even if with limited duration, like in the case of selenite-induced stress, inducing the exclusion from SGs of both RACK1 and HDAC6 (Fujimura et al., 2012).

PBs instead contain several proteins involved in mRNA decay and translation repression, such as the decapping enzyme Dcp1, Dcp2 and the components of the deadenylase CCR4-NOT (Sheth and Parker, 2003; Decker and Parker, 2012), while translation initiation factors including eIF4A, B, G and the poly-A binding protein Pabp are usually absent and more commonly found in SGs. However, several components are shared between the two, like XRN1 exoribonuclease, the argonaute ARGO2 that regulates miRNA functions and the translation initiation factor eIF4E (Sheth and Parker, 2003; Kedersha et al., 2005; Decker and Parker, 2012). Compared to SGs, PBs show a much denser network of interactions, due to a higher concentration of RBPs and transiting RNAs inside them (Jain et al., 2016; Hubstenberger et al., 2017; Markmiller et al., 2018). More specifically, PBs show a remarkable 4-fold enrichment in helicases compared to SGs, including DDX6 (Hubstenberger et al., 2017).

An overview of the differences in proteome composition between the two types of condensates is shown in **Figure 6**.
Interestingly, the key factors for the formation of these condensates, as well as the macromolecular interactions they generate, are redundant and context-specific, so diverse stress conditions induce condensation in different ways. For example, G3BP1 and G3BP2 are required for SG formation under oxidative stress but not under an osmotic one (Protter and Parker, 2016).

In recent years, a huge improvement has been made regarding the characterization of these condensates' components.
In 2016 the development of experimental techniques led to the first purification of the cores substructures within SGs. In this work, the cores substructures were obtained from sodium-arsenite stressed U-2 OS cells through an affinity purification of GFP-G3BP followed by mass-spectrometry (Jain et al., 2016). The analysis revealed a diversified proteome with $\sim$ 50% of components being RNA-binding proteins, together with other elements like metabolic or post-translation modification enzymes, probably recruited through protein-protein interactions (Protter and Parker, 2016; Jain et al., 2016).
Several tRNA synthetases and ribosome factors are also present.
In 2018 Markmiller and colleagues investigated the proteome composition of SGs through an *in vivo* proximity labeling approach, employing an engineered ascorbate peroxidase (APEX2) fused to the SG protein G3BP1, upon sodium arsenite exposure (Markmiller et al., 2018). This, combined with an immunofluorescence (IF) screen, led to the identification of stress-dependent and independent interactors associated with G3BP1, as well as sensitive proteins that decrease in concentration during

stress conditions.
Interestingly, the composition of SGs' proteome seems to be also cell-specific, with different interactors found in Hela, HepG2 and Neuronal progenitor cells (NPC), the last one showing the most diverse composition of SGs, being enriched in PQC factors that are responsible for the clearance of misfolded proteins by autophagy (Markmiller et al., 2018).
In parallel, Hubstenberger and colleagues developed a Fluorescence-Activated Particle Sorting (FAPS) method to purify PBs, by expressing a canonical PB marker GFP-LSM14A in HEK293 cells. This was then combined with liquid chromatography-tandem mass spectrometry (LC-MS/MS) and led to the identification of 125 PB's proteins, such as the repression cofactor DDX6, several decapping enzymes like DCP1A, DCP1B and DCP2, mRNA decay factors LSM14A, LSM14B and components of the miRNA pathway like AGO1, AGO2. Ribosomal proteins, several SG proteins and all translation initiation factors except eIF4E were instead depleted (Hubstenberger et al., 2017).

In general, the dynamicity of these condensates is influenced by the presence of ATPases and several DEAD-box helicases, creating a fast exchanging rate of components with the environment by disrupting macromolecular interactions and unwinding nucleic acids when the concentration of stalled mRNP becomes too high. These proteins, coupled with microtubule motors and several chaperones are also the key elements responsible for the assembly and disassembly of SGs during the recovery phase from stress (Protter and Parker, 2016). This is why mutations and imbalances in these components are often the cause of delayed disassembly or even diseases.

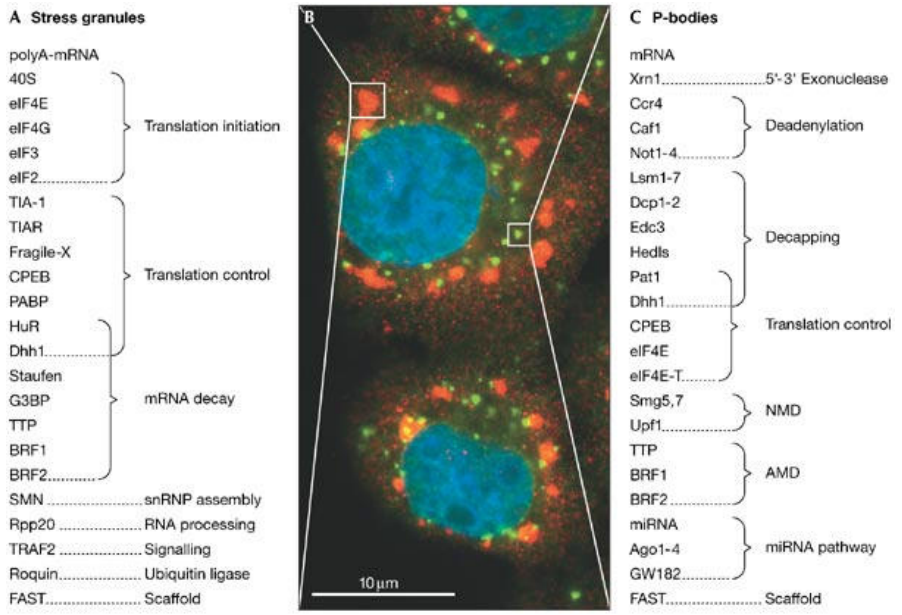


Figure 6: Differences in proteome composition between SGs and PBs, from Newbury et al. (2006). (**A**) SGs contain mainly untranslated poly-A mRNAs, as well as complexes involved in translation initiation and control. (**B**) Immunofluorescence micrograph of human HeLa cells undergoing arsenite-induced oxidative stress. The nuclei of the cells are shown in blue, while SGs are colored in red and PBs in green. (**C**) PBs contain members of decapping and deadenylation machinery, as well as translation initiation factor eIF4E and miRNA regulators.

1.4. Condensates and diseases

Even though SGs and PBs are enriched in RNA content and this helps their dissolution upon stress clearance, mutations and changes in their components' concentration can contribute to cellular toxicity and diseases. Whereas for some proteins an inappropriate liquid-liquid phase separation process is linked to proteins' over-expression (Bolognesi et al.,

2016), in other cases mutations in the proteins' sequence or external factors can promote the formation of solid-like aggregates (Cid-Samper et al., 2018) leading to pathologic states.

The first scenario can be achieved when the over-expression of a dosage-sensitive protein (e.g. concentration-dependent) generates an inappropriate liquid-liquid demixing (Bolognesi et al., 2016). This is the case of Mip6p in yeast, a protein containing two low-complexity regions and four RNA recognition motives that in physiological conditions is lowly expressed and diffused in the cytoplasm, but when over-expressed can relocate in cytoplasmic foci with liquid-like properties establishing interactions with PB components.

The second scenario instead is often linked to a mutated PrLD, enriched in condensates' proteins, that becomes the cause of protein-misfolding diseases (Gotor et al., 2020).
This is especially the case of ATXN1 and ATXN2, where a CAG-repeat in their coding regions causes respectively the type 1 and type 2 spinocerebellar ataxia by inhibiting the shuttling of these proteins between cytoplasm and nucleus (Lorenzetti et al., 1997).
Mutated hnRNPA1 and hnRNPA2B1 in their PrLD have been found in families affected by multisystem proteinopathy (MSP), an anomaly in SG dynamics and autophagy processes that is linked to other diseases such as Amyotrophic Lateral Sclerosis (ALS), frontotemporal lobar degeneration (FTLD), inclusion body myopathy (IBM) and Paget disease of bone (PDB) (Benatar et al., 2013; Le Ber et al., 2014).
Another example is FMR1, where a CGG expansion is responsible for fragile X-associated tremor/ataxia syndrome (FXTAS), sequestering sev-

eral proteins like the splicing regulator TRA2A. (Cid-Samper et al., 2018). A similar case is represented by TDP-43, a protein containing a C-terminal low complexity region and PrLD that by proteolytic cleavage is able to phase separate and aggregate in deposits, leading to the insurgence of ALS and frontotemporal lobar degeneration with ubiquitin-positive inclusions (FTLD-U) (**Figure 7**) (Da Cruz and Cleveland, 2011; Patel et al., 2015; Verdile et al., 2019; Zbinden et al., 2020; Zacco et al., 2022).
FUS is a protein involved similarly to TDP-43 in different processes of RNA metabolism and it has one of the best characterized PrLD in the phase-separation field. A mutation in the RGG/RG domain of this protein can suppress its physiological LLPS behavior by preventing its binding to the nuclear import receptor Transportin-1, causing the ALS disease (Da Cruz and Cleveland, 2011; Kwiatkowski et al., 2009).
Interestingly, TDP-43 and FUS can cross-regulate their concentration and the diseases they induce are mutually exclusive.

The formation of condensates has also been associated with cancer insurgence.
For example, due to chromosomal translocations, the RNA binding domain of FET RBPs (FUS,/TLS, EWS and TAF15) can be replaced with an ETS transcription factor (e.g. FLI1) with the generation of a fusion protein (e.g. EWS-FLI1), which can cause aggressive pediatric tumors called Ewing sarcomas (Araya et al., 2005; Paronetto et al., 2011; Verdile et al., 2019). Another example is the reduction in TIA-1 and TIAR concentration, which has been linked to a rise in cell proliferation and tumor growth (Sánchez-Jiménez et al., 2015). In addition, the core protein G3BP1 seems to enhance cell proliferation, metastasis and chemoresistance (Dou et al., 2016).

Furthermore, cancer cells can induce the formation of condensates to protect their survival chances and to resist chemotherapy, as cancer cells could actively sequester drug molecules inside these condensates to limit their efficacy and develop resistances, changing their pharmacodynamics (Anderson et al., 2015). In this sense, pharmacological inhibition of these organelles could be a way of influencing tumor progression.

Other examples are Tau proteins, cytoplasmic proteins unlike TDP-43 and FUS with no RNA-binding domains, which form cytosolic aggregates called neurofibrillary tangles in the neurons of Alzheimer patients (Wolozin et al., 1986) and are also involved in many cases of frontotemporal dementia (Ling et al., 2013), or α-synuclein, a protein involved in synaptic homeostasis (Lashuel et al., 2013), which in an oligomeric or fibrillar conformation can induce the formation of pathological deposits called Lewy bodies (McKeith et al., 1996; Shahmoradian et al., 2019), responsible for different types of Parkinson disease and dementia and containing high concentrations of this protein together with lipid membranes,
Interestingly, while FUS, TDP-43 and Tau physiologically oligomerize and phase-separate to exert their regulatory functions, α-synuclein oligomerization is usually toxic. In general, while all of these proteins differ in the oligomerization process, in RNA and protein-binding ability and in both intra- and inter-molecular interactions they form, they all contain low-complexity regions that can induce their phase-separation, which will happen differently according to sequence properties and external stimuli, as well as mutations events and post-translational modifications these proteins may undergo. A sudden change in parameters like phosphorylation, methylation and acetylation or even mitochondrial ATP production nec-

essary for maintaining the LLPS can allow a transition to a pathological state (Zbinden et al., 2020).

The link between changes in the composition of these condensates and the occurrence of diseases will be an increasingly relevant area of focus for the scientific community in the next years.

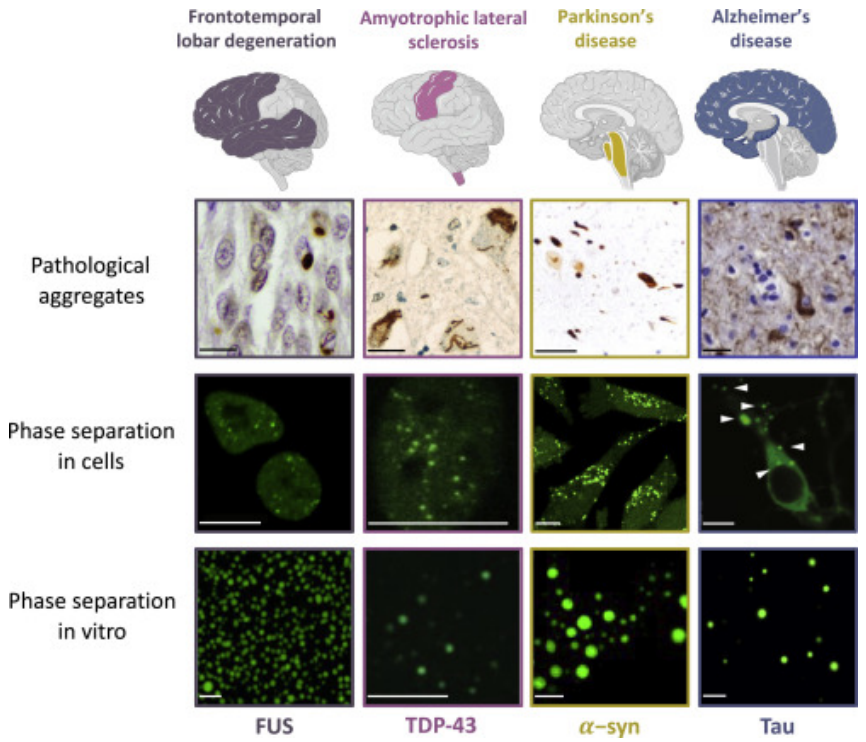


Figure 7: Representation of pathological aggregates forming in different brain areas and associated with FUS, TDP-43, α-synuclein and Tau, from Zbinden et al. (2020). Immunohistochemistry images in postmortem brain samples and phase-separated proteins *in vivo* and *in vitro* are shown in the second, third and fourth rows respectively.

1.5. Pathogens and phase-separation

In addition to temperature and chemicals, another stressor that can induce or alter a phase-separation process is represented by pathogenic infections.
During these events, ribonucleoprotein condensates can recruit several antiviral proteins like RIG-1 and RNase L, promoting the induction of the innate immune response of the cell (Onomoto et al., 2012; Reineke and Lloyd, 2015). This is the reason why several viral and bacterial pathogens have developed mechanisms to block or hijack the formation of these organelles.

In the SARS-CoV-2 infection process, for example, its nucleocapsid (N) protein is a key factor in neutralizing the innate immune response of the cell. The N protein contains two RNA binding domains and three IDRs that are able to induce its phase-separation process, which depends also on several factors such as salt concentration, pH and phosphorylation mechanisms. This protein can initially be recruited inside SGs through a phase-separation process, where it binds to SG's G3BP1, G3BP2 and other factors, sequestering them to inhibit SG formation, so that these condensates decrease both in number and size (Luo et al., 2021). Since SARS-CoV-2 is an enveloped virus, its RNA must be encapsulated to form a mature virion. The N protein is then responsible for the formation of nascent virions, through selective condensation and packaging of the viral genomic RNA (gRNA) and the accumulation of SARS-CoV-2 structural proteins (S, N, E and M) at the ER-Golgi intermediate compartment membrane (ERGIC) (Klein et al., 2020), to which follow the recruitment of the SARS-CoV-2 RNA-dependent RNA polymerase (RdRp) to form

viral replication centers (Savastano et al., 2020). The exact formation of these N-protein condensates containing gRNAs is still not completely clear, as well as whether they form directly on the ERGIC membrane or are recruited there only in a second moment.
In parallel, the N protein deactivates the host immune responses. While in general the double-stranded viral RNA is often detected by host proteins such as RIG-1 or kinase PKR (McNab et al., 2015) which, activating IRF3 that leads to the expression of type -I interferons (IFNs), induce the innate immune response (Hou et al., 2011; Cai et al., 2014), the SARS-CoV-2 N protein phase-separation seems to interfere with the production of IFNs' precursors like MAVS and suppress their expression, also preventing the phosphorylation of the PKR protein and activating the NF-kB signaling pathway through the recruitment of kinases inside the N condensates (Wu et al., 2021), increasing the production of pro-inflammatory cytokines (Cascarina and Ross, 2022), as shown in **Figure 8**.
Other than SARS-CoV-2, several other viruses have become efficient in counteracting the cell's defenses to successfully replicate inside their host (Gaete-Argel et al., 2019).
Certain viruses can suppress SGs formation and this often is achieved with the cleavage of the G3BP1 protein, an essential SG factor, in the case of encephalomyocarditis virus (EMCV) (Ng et al., 2013).
Another example is the Immunodeficiency Virus Type-I (HIV-1), which impairs SGs despite eIF2α phosphorylation thanks to the binding of the viral protein Gag to G3BP1, eIF4E or eEF2 depending on the type of SGs and stressors (Valiente-Echeverría et al., 2014; Poblete-Durán et al., 2016). Similarly, the West Nile Virus (WNV) sequesters TIA-1 and TIAR SG components preventing their formation (Gaete-Argel et al., 2019).

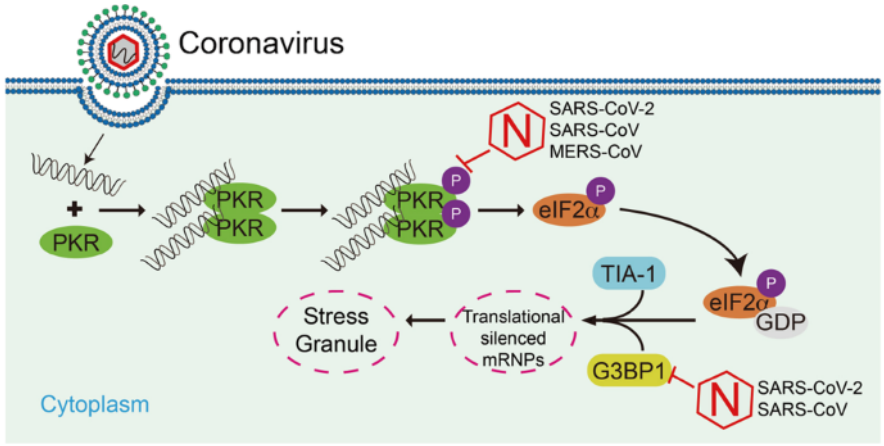


Figure 8: Mechanism of inhibition of SG formation by SARS-CoV-2 N protein, from Zhou. et al. (2020). SARS-CoV-2 N protein can prevent SG formation and tamper with the host innate immune response either by sequestering G3BP1, essential for SG formation, or blocking the kinase PKR that detects environmental stress and activating NF-kB singling pathways to produce inflammatory cytokines, suppressing the production of IFNs.

Other viruses, instead, including the Ebola virus (EBOV), exploit SG formation to sequester key proteins important for the host translational machinery such as, eIF3, eIF4G, PABP and G3BP1 (Nelson et al., 2016), which are relocated into viral replication factories (RFs), sorts of organelles with liquid-like properties assembled in the cytoplasm or nucleus of the host cell (Nevers et al., 2020). These structures usually contain nucleic acids and viral proteins and some key cellular elements that are sequestered to enhance the replication process or to protect the virus from the host's immune defenses. For its replication, the Ebola virus also requires the host nuclear RNA export factor 1 (NXF1), a component of the nuclear mRNA export pathway. This protein interacts with viral mRNAs in RFs and with Ebola nucleoprotein and probably is required for export-

ing viral RNAs to ribosomes for translation (Wendt et al., 2020).
Another example is the Hepatitis C virus (HCV), which induces the formation of RFs in liquid droplets and relocates several SG components like G3BP1, TIA1, DDX3 and ATX2, interfering with the normal functioning of SGs in order to avoid the production of anti-viral molecules (Ariumi et al., 2011; Garaigorta et al., 2012).
Finally, some viral species like Mammalian orthoreovirus (MRV), initially promote SG formation through eIF2α phosphorylation incorporating the viral core proteins inside the condensates, which are later disrupted as the infection proceeds to help the viral proteins' synthesis and replication occurring in RFs in the perinuclear region (Rhim et al., 1962; Qin et al., 2011).

Despite the SGs being a more common target in viral replication events, PBs formation is also blocked or exploited by several pathogens.
For example, adenoviruses accumulate viral mRNAs by reducing the number of PBs in the cell and sequestering numerous important PB components, such as DDX6 and XRN1, which are moved to viral-induced aggresomes to be degraded (Greer et al., 2011).
Other examples are West Nile Virus (WNV) and Dengue Virus (DENV), which recruit PB components inside RFs, while reducing PB assembly (Pijlman et al., 2008; Silva et al., 2010).
Some viruses instead seem to increase the number of PBs per cell, such as for instance SARS-CoV, as its protein Nsp1 is able to increase the mRNA down-regulation and degradation to maximize viral replication (Huang et al., 2011). The same happens with the Cytomegalovirus (HCM), which promotes the PBs accumulation and raises the expression of several PB components, even though the viral RNA does not localize inside them

(Seto et al., 2014).
Finally, similarly to what happens for SGs, SARS-CoV-2 and other coronaviruses seem to induce PB disassembly through the expression of the N protein. This causes a significant rise in the concentration of PB-regulated cytokine transcripts and could be responsible for the abnormal production of proinflammatory molecules observed in severe SARS-CoV-2 infection cases (Kleer et al., 2022).

In summary, RNA viruses seem either to inhibit granule assembly in order to favor their viral cycle, or they accomplish this task by building membrane-less replication organelles, which share many characteristics with SGs and PBs.

Viruses are surely the most studied pathogens in connection with phase-separated organelles but recently other organisms have shown the capability of affecting the innate immune response of the cell, among which several bacteria, fungi and protozoa.
In particular, SGs formation upon infection has been recovered in three species of bacteria (Salmonella Typhimurium, Shigella and Listeria) and in the protozoan parasite Plasmodium (Tweedie and Nissan, 2021).
Salmonella Typhimurium infection induces the phosphorylation of eIF2α, triggering SG formation in a relatively small amount of cells (Abdel-Nour et al., 2019) while Shigella exhibits a more pronounced SG host response, always eIF2α-phosphorylation dependant (Vonaesch et al., 2016). Other organisms instead employ different approaches, with the parasite Plasmodium not showing any SG formation upon infection (Hanson and Mair, 2014) and Listeria showing an oscillating SG induction over time, dependent on eIF2α phosphorylation (Abdel-Nour et al., 2019).

However, the differential composition of these microbial-induced SGs has not been yet addressed, as well as the role of SGs or PBs and the sequestration of their components during these infections, for which a lot instead is known for viral pathogens.
A hypothesis suggests that SGs could regulate the formation of inflammasomes, an important mechanism of the innate immune response to bacterial infection, by sequestering the DEAD-box helicase DDX3X (Samir et al., 2019), or SG assembly could be regulated by the ubiquitin-proteasome system (Lin and Machner, 2017) and these pathways could be exploited and hijacked by microbes to their advantage.

CHAPTER 2

OBJECTIVES

In summary, macromolecular interactions are fundamental mechanisms needed for a plethora of cellular pathways. Protein and RNA components can adopt different conformations and establish a network of contacts that can be the trigger for phase-separation processes leading to the formation of ribonucleoprotein condensates. By recruiting components and altering their concentration, these condensates can control and catalyze biochemical reactions. Among the different types of membrane-less organelles that can form in the nucleus or in the cytoplasm of the cells, SGs and PBs are nowadays two of the best characterized and studied. Their formation is induced by external stimuli, when the RNAs translation rate is slowed down or stopped and they can exert multiple functions related to RNA storage and degradation or protection of important molecules from harmful conditions. They are dynamic entities and they can exchange components among themselves and with the surrounding environment. Furthermore, the proteome and transcriptome composition of these condensates is not fixed but can vary according to the type of stress condition, cell types and incubation period.

Despite being reversible in physiological conditions, mutations or other events changing the biophysical equilibrium of molecules inside the cell can trigger pathological aggregation events that can lead to different neurodegenerative diseases or even cancer. Furthermore, these organelles and the underlying interaction networks can be exploited or repressed by pathogens, which can sequester key components of the cell to antagonize the host's innate immune response and to help the viral replication.

A summary of SGs in different pathological conditions is shown in **Figure 9**.

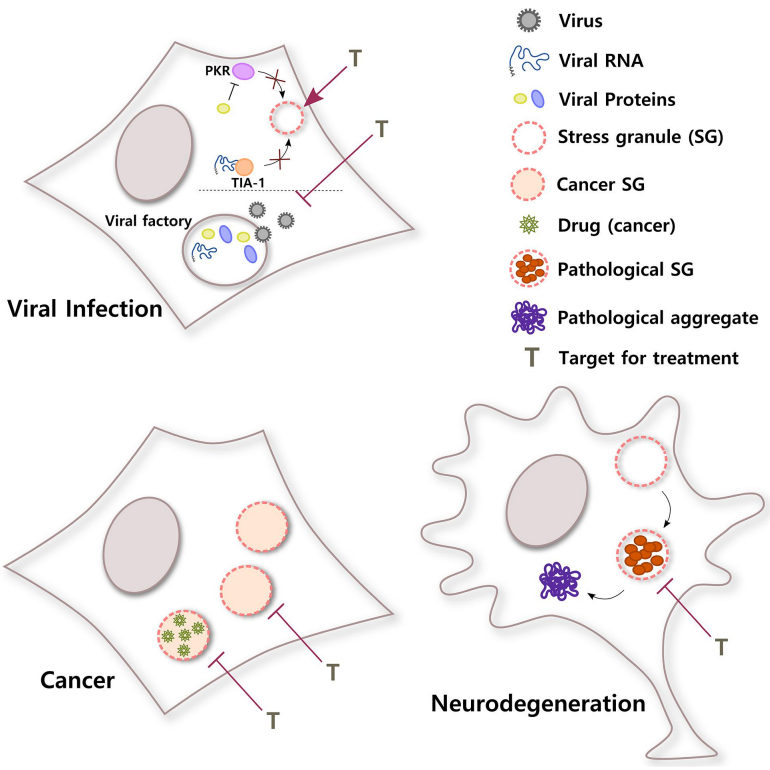


Figure 9: SGs in different pathological conditions, from Campos-Melo et al (2021). SG formation can be inhibited by viruses blocking kinase PKR, which senses environmental adverse conditions and by sequestering key components for granule assembly like TIA-1 or G3BP1, impairing the innate immune response of the cell by suppressing IFNs production. Furthermore, SGs assembly can be vital for cancer cells, which can become resistant to chemotherapies by storing drug molecules inside these condensates, changing their action and concentration. Finally, changes in SGs components can lead to aberrant irreversible aggregates responsible for several neurodegenerative diseases and pathological states.

However, while there is evidence supporting that these condensates are sustained by a combination of protein-protein, protein-RNA and RNA-RNA interactions, little is still known about which are the specific ele-

ments promoting each type of interaction and how these factors can regulate one another. In addition, in the context of infections, pathogens seem to hijack or repress these condensates exploiting some of their key elements, but there is still a lot of work to do to unravel the specific network of interactions that these infectious agents establish with the host and a complete picture of all the cellular factors that are preferentially targeted is still missing, especially for bacteria and viruses of great global interest and relatively new insurgence like SARS-CoV-2.

This thesis aims to answer these questions with the following steps.

- Characterize protein and RNA elements that are enriched in SGs and PBs, unraveling the link between the secondary structure content of both protein and RNAs and how the different types of macromolecular interactions are established and connected.
- The creation of PRALINE, a database that aims to combine the characterization of physicochemical properties of both liquid-like and solid-like condensates' components with information of disease-related SNVs occurring in both proteins and RNAs.
- Predict the protein-RNA interactome between the human host and SARS-CoV-2 genome and identify relevant loci, to highlight potentially relevant factors that could be sequestered or targeted by the virus to increase its infectivity.
- Investigate the reason behind the relatively small overlap of interactors found in different human-SARS-CoV-2 protein-RNA interactomes experiments, speculating their importance in the context of host infection.

Part II

RESULTS

CHAPTER 3

REVIEW: ZOOMING IN ON PROTEIN-RNA INTERACTIONS: A MULTILEVEL WORKFLOW TO IDENTIFY INTERACTION PARTNERS

Review: Zooming in on protein–RNA interactions: a multilevel workflow to identify interaction partners

Ribonucleoprotein complexes have been largely studied in recent years in order to understand the different ways in which RNAs and proteins can exert their function while binding to each other. Every protein that comes in contact with a transcript is commonly defined as an RNA-binding protein (RBP).
However, this interaction is dependent on the cellular environment and its strength can vary according to the context and the affinity between the molecules. Furthermore, while canonical RBPs usually bind RNA thanks to known domains like RNA recognition motives, non-canonical RBPs generally harbor non-canonical domains, often located in poorly structured regions, and are more difficult to study (Castello et al., 2012; Conrad et al., 2016; Monti et al., 2021).
In this context, classifying and defining new RBPs is a necessary task, even though the identification of precise RNA binding domains remains challenging.
Despite the development of sophisticated experimental high-throughput techniques in recent years, there is still the need to combine them with computational tools to filter the vast amount of both relevant and non-specific interactions collected with these techniques in order to select only the most important candidates, aiming at the creation of organized pipelines that can be reproducible and available to the public.

In this review article, we propose a workflow for the classification of

new putative RBPs and the identification of their binding partners, starting from a pool of potential interactors collected by performing high-throughput in-cell technique experiments and progressively narrowing down their numbers with computational predictors, which can help to identify the regions involved in the binding and estimate the binding strength of the interactions.

This workflow could be viewed as a multidisciplinary approach, exploiting powerful high-throughput methods and predictors to create an integrated and productive pipeline useful for the scientific community.

This work was published in the Biochemical Society Transactions journal in 2020.

As a co-first author, I mainly contributed to the review by describing the computational methods for protein-RNA interactions prediction and, more specifically, catRAPID algorithm and its different implementations.

Colantoni, A., Rupert, J., Vandelli, A., Tartaglia, G. G., and Zacco, E. (2020). Zooming in on protein-RNA interactions: a multi-level workflow to identify interaction partners. *Biochemical Society Transactions*, 48(4):1529–1543.
DOI: 10.1042/BST20191059

Biochemical Society Transactions (2020) **48** 1529–1543
https://doi.org/10.1042/BST20191059

PORTLAND PRESS

Review Article

Zooming in on protein–RNA interactions: a multi-level workflow to identify interaction partners

Alessio Colantoni[1,*], Jakob Rupert[2,3,*], Andrea Vandelli[4,5,6,*], Gian Gaetano Tartaglia[1,2,3,4,5,7] and Elsa Zacco[3]

[1]Center for Life Nanoscience, Istituto Italiano di Tecnologia, Viale Regina Elena 291, 00161 Rome, Italy; [2]Department of Biology 'Charles Darwin', Sapienza University of Rome, P.le A. Moro 5, Rome 00185, Italy; [3]Center for Human Technologies, Istituto Italiano di Tecnologia, Via Erico Melen 83, 16152 Genoa, Italy; [4]Centre for Genomic Regulation (CRG), The Barcelona Institute for Science and Technology, Dr. Aiguader 88, 08003 Barcelona, Spain; [5]Universitat Pompeu Fabra (UPF), 08003 Barcelona, Spain; [6]Systems Biology of Infection Lab, Department of Biochemistry and Molecular Biology, Biosciences Faculty, Universitat Autònoma de Barcelona, 08193 Cerdanyola del Vallès, Spain; [7]Institucio Catalana de Recerca i Estudis Avançats (ICREA), 23 Passeig Lluis Companys, 08010 Barcelona, Spain

Correspondence: Elsa Zacco (elsa.zacco@iit.it) or Gian Gaetano Tartaglia (gian.tartaglia@iit.it)

OPEN ACCESS

Interactions between proteins and RNA are at the base of numerous cellular regulatory and functional phenomena. The investigation of the biological relevance of non-coding RNAs has led to the identification of numerous novel RNA-binding proteins (RBPs). However, defining the RNA sequences and structures that are selectively recognised by an RBP remains challenging, since these interactions can be transient and highly dynamic, and may be mediated by unstructured regions in the protein, as in the case of many non-canonical RBPs. Numerous experimental and computational methodologies have been developed to predict, identify and verify the binding between a given RBP and potential RNA partners, but navigating across the vast ocean of data can be frustrating and misleading. In this mini-review, we propose a workflow for the identification of the RNA binding partners of putative, newly identified RBPs. The large pool of potential binders selected by in-cell experiments can be enriched by *in silico* tools such as *cat*RAPID, which is able to predict the RNA sequences more likely to interact with specific RBP regions with high accuracy. The RNA candidates with the highest potential can then be analysed *in vitro* to determine the binding strength and to precisely identify the binding sites. The results thus obtained can furthermore validate the computational predictions, offering an all-round solution to the issue of finding the most likely RNA binding partners for a newly identified potential RBP.

Introduction

Since their discovery and until recently, RNA-binding proteins (RBPs) have been identified by the presence of one or more RNA-binding domains in their sequences [1]. However, concomitantly to a new appreciation for RNA as key biological macromolecule acting at post-transcriptional level [2–4], there has also been a re-evaluation of what constitutes an RBP. Since one of the principal ways by which RNA exerts its function is by the formation of ribonucleoprotein complexes, every protein capable of establishing even weak and extemporary interactions with an RNA molecule may be defined as RBP [5,6]. The interactions of proteins with RNA can be highly dynamic and heavily dependent on the cellular environment [7], which makes the goal of defining the range of affinities and specificities quite challenging [8]. In fact, indiscriminate binding of RNA by RBPs is a quite common phenomenon [9], and the assumption that stronger affinity translates into more relevant biological functions is not necessarily correct [10]. For the scientific community, the revelation of the dynamicity and malleability of the partnership between RBPs and RNA allows for the exploration of new possible interaction mechanisms, networks, genes and protein regulation systems to investigate. It becomes, therefore, increasingly important to complete the catalogue of eukaryotic RBPs, at present

*These authors contributed equally to this work.

Received: 12 June 2020
Revised: 17 July 2020
Accepted: 20 July 2020

Version of Record published: 21 August 2020

PORTLAND PRESS

Biochemical Society Transactions (2020) **48** 1529–1543
https://doi.org/10.1042/BST20191059

containing >315 000 elements (about 6000 orthologues from >150 species) of which 3500 are from *H. sapiens* and can be divided in conventionally-defined RBPs and several classes of non-canonical RBPs [5,11,12].

Large-scale identification of new potential RBPs can reveal unexpected biological and pathological functions and, when confronted with a novel RBP, the identification of its RNA binding partners is a critical step to define the protein's cellular and molecular roles. To achieve this goal, increasingly sophisticated high-throughput methodologies have been developed, spanning from methods that aim to preserve the native cellular RNA–RBP interactions [13–15] to finely-controlled *in vitro* techniques that allow to define kinetics and dynamic parameters for each binding pair [16,17]. However, the field that has probably seen the biggest evolution in the shortest time span is the one of computational prediction algorithms [18–22]. The wide range of computational tools available includes several data-driven methods based on learning models, in which the algorithms are trained using experimental outcomes and databases to identify RNA–RBP binding patterns and define the genome-wide profiling of RNA–protein interactions [22].

In this short review, we propose a work pipeline for the identification of the RNA binding partners for novel putative RBPs. Starting from in-cell data harvesting, we would like to guide the reader through the employment of the different tools offered by *cat*RAPID [23,24], our in-house developed RNA–protein interaction prediction algorithm, and propose some indications on how to validate the outcome experimentally. Our wider goal is to support the scientific community in the identification of novel biologically relevant non-canonical RBPs.

Identification of RNA binding partners in cellular context

The validation and training of computational algorithms for the prediction of protein-RNA interactions is strongly supported by experimental data (Figure 1). Prediction software can be significantly enriched by the output of techniques able to identify a protein's RNA partners within the cellular environment; such procedures are key to defining native interaction pairs and to monitoring responses and variations upon physiological *stimuli* or under pathological stress.

RIP-based approaches

The main tool to obtain information about the RNA binding partners of a target protein in the cellular environment is immunoprecipitation (IP), a widespread technique to pull down the protein of interest together with its physiological RNA binding partners. RNA immunoprecipitation (RIP) requires incubation of cell lysates with an antibody raised against the target protein [25]. RNA molecules bound to the target protein can then be isolated and analysed to reconstruct physiological native complexes formed within the cell. RIP can be coupled to either microarrays (RIP-Chip) or high-throughput sequencing (RIP-Seq). In either case, RIP can only determine the identity of the RNA molecules associated to the target protein, unless digestion-optimized RIP (DO-RIP) is performed [26]. This variation of RIP introduces an RNase digestion step to preserve only the portion of RNA bound to the protein, allowing for binding-site mapping [27]. If, instead, the interest is focused on identifying multi-subunit ribonucleoproteins, the most appropriate RIP variant may be RIP in tandem (RIPiT), which employs two distinct IP steps performed either with antibodies against different proteins of the complex or with antibodies binding different regions of the same target protein [28]. Information about native protein–RNA complexes formed within the cell can also be obtained by employing affinity tags [29], without having to rely on the antibody's specificity and sensitivity.

CLIP-based approaches

To overcome RIP's limitations (enrichment of indirectly bound RNAs, detection of interactions not present in cell but formed after lysis, loss of weaker interactions due to the required stringent washing conditions), cross-linked RNA immunoprecipitation (CLIP) has been developed [30]. CLIP promotes the stabilisation of the bonds between a protein and its interacting RNA, generally by UV radiation [31]. A large amount of CLIP variants is available, and most of them can offer high-resolution results at the single nucleotide level. Pioneers among these are, for example, the high-throughput sequencing CLIP (HITS-CLIP), that enriches the RNA population for sequences corresponding to the RBP binding sites [32]; the individual nucleotide-resolution CLIP (iCLIP) [33,34] and enhanced CLIP (eCLIP) [35], that utilise different oligonucleotide adapter configurations to obtain RNAs of different lengths employed to build the interacting fragments at single-nucleotide resolution; and the photoactivatable ribonucleoside CLIP (PAR-CLIP), that relies on metabolic incorporation of labelled ribonucleoside analogues that yield photo-adducts when cross-linked at selected wavelengths [36]. Radiation-free CLIP variants, such as infrared-CLIP (irCLIP), have also been reported [37]. To overcome the

Biochemical Society Transactions (2020) **48** 1529–1543
https://doi.org/10.1042/BST20191059

PORTLAND PRESS

Figure 1. The workflow of discovering RNA partners for an RBP.

Protein and RNA sequence databases, structural information and results from RIP/CLIP experiments feed computational prediction tools such as *cat*RAPID. The software utilises this information to define RNA sequences with high probability of interacting with a given RBP and rank them accordingly. Several *in vitro* techniques allow for the validation of predicted results, for the calculation of binding strength and the definition of the binding sites.

restrictions imposed by the use of irreversible UV cross-linking, formaldehyde-based techniques such as fCLIP, which are more efficient in capturing interactions with dsRNA [38], can also be employed [39]. One limit of CLIP-Seq approaches is that low crosslinking efficiency makes low abundant transcripts difficult to detect, while transcripts present at high amounts are usually over-represented in IP samples. This issue can be partial solved with a proper bioinformatic analysis.

Bioinformatic analysis

High-throughput sequencing of the RNAs isolated through RIP, CLIP and related protocols yields millions of short 'reads', which represent the sequenced portions of cDNA fragments obtained through RNase titration (a step omitted in standard RIP-Seq), followed by reverse transcription and PCR amplification. The RNase treatment allows to obtain fragments long enough to be uniquely mappable but short enough to identify the binding site with the highest possible resolution. Library preparation ends with the production of cDNA fragments flanked by adapters that allow amplification and sequencing, resulting in the generation of short reads that can undergo bioinformatic analysis aimed at identifying RNA targets for the RBP (Figure 2). Reads pre-processing steps, including demultiplexing and adapter trimming, are often required, especially in those cases in which Unique Molecule Identifiers (UMIs) are used [40]. UMIs are random barcodes which identify unique cDNA fragments, allowing to detect and remove PCR duplicates that are commonly produced during CLIP-Seq library preparation [34,35].

To identify the RNA molecules from which they derive, reads are aligned to a reference genome using splice-aware alignment programs commonly used to analyze standard RNA-Seq data, like TopHat2 [41] or Star [42]. If the RBP under investigation binds mature mRNAs, reads can be mapped directly to the transcriptome [43] using a splice-unaware mapper like Bowtie2 [44]. Reads coming from HITS-CLIP experiments have high mutation rates (usually deletions) at the cross-linking site (CIMS, standing for cross-linking induced mutation sites), which are due to residual amino acids hindering the reverse transcriptase [45–47]. Similarly, the use of 4-thiouridine (4-SU) or 6-thioguanosine (6-SG) in the PAR-CLIP protocol leads to a high number of transition events (T to C or G to A, respectively) at the cross-linking sites [48]. Reads mapping can be improved by taking into account the high rate of such mutations. For instance, the splice-unaware BWA aligner [49] has been modified in order to incorporate an error model that favours PAR-CLIP specific transitions [50,51]. Reads produced by iCLIP, and likely eCLIP, experiments do not require such special treatment, since such protocols enrich for cDNAs truncated at the cross-linking site (CITS, standing for cross-linking induced truncation sites), while only a minor proportion of fragments represent CIMS-containing read-through cDNAs [43].

Post-processing of aligned reads is a mandatory step in RIP-Seq and CLIP-Seq data analysis. Reads aligning to multiple genomic positions are usually removed [52]. However, such multi-mapped reads can be used to identify regulatory RNA sites localized within repetitive regions [53]. To filter out PCR duplicates, reads mapping to the same genomic position are collapsed; UMIs, when present, can be used to avoid removing natural read duplicates, common in case of high sequencing depth [54].

In RIP-Seq experiments, target RNAs can be identified either by transcript enrichment analysis of IP versus control samples [55–57], which can be performed using procedures commonly adopted in differential expression analysis, or with ad-hoc peak-calling tools [58,59]. Binding site identification in HITS-CLIP experiments is usually accomplished by means of peak-calling approaches, that identify regions which are enriched in reads with respect to their genomic context (gene, transcript, metagene region) [59], the background represented by control experiments (input, IgG, mock IP) [60], or baseline expression profiles [59]. Single-nucleotide

PORTLAND PRESS

Biochemical Society Transactions (2020) 48 1529–1543
https://doi.org/10.1042/BST20191059

resolution in HITS-CLIP data analysis can be achieved by looking for statistically significant CIMS [59]. Such resolution in binding site identification can also be attained in PAR-CLIP, by searching for transition events that are not likely to be caused by sequencing noise, single nucleotide polymorphisms or contamination [61]. The first nucleotide of reads from iCLIP and eCLIP experiments most often marks the truncation site and it is therefore located one nucleotide after the cross-linking site. Statistically significant CITS can be identified by using tools like iCount [62] and PureCLIP [63].

In silico prediction of protein–RNA interactions

The outcome of the data analysis obtained by means of CLIP-Seq experiments is of enormous value in approaching the study of a potential RBP with unknown RNA partners. However, there are two major drawbacks in limiting the investigation to CLIP-Seq approaches: the protein regions in direct contact with the target RNAs remain unknown, and there are not sufficient data to speculate on the strength of the interaction for each protein–RNA pair. The use of predictive algorithms such as *cat*RAPID would integrate the results of a CLIP-Seq experiment with this information [24] (Figure 1).

*cat*RAPID

*cat*RAPID is an algorithm able to compute protein–RNA interaction propensities with strong predictive power (area under the Receiver Operating Characteristic curve of 0.78 on >1 000 000 interactions) [64], through the calculation of secondary structures, hydrogen bonding and van der Waals contributions. The algorithm was trained on PDB crystals [24] and was later adapted to predict CLIP-Seq interactions with long non-coding RNAs [65]. For large proteins (>750 amino acids) and RNAs (>1000 nucleotides), the algorithm fragments sequences into overlapping segments and computes the interaction propensity through the analysis of physical–chemical properties and secondary structures of the molecules. According to the chosen implementation (Table 1), the method can either reconstruct the overall interaction propensity score for each protein–RNA pair [66] or rank the fragments according to the predicted interaction strength [67]. The outcome of this analysis allows to map both protein and RNA binding sites and to estimate the overall strength of the interactions [68], overcoming the limitations of CLIP-Seq techniques mentioned above. If used in combination with CLIP-Seq data relative to the protein of interest, *cat*RAPID can be employed to select the best targets based on calculated binding strength, but it could be also useful in predicting putative targets that are not expressed within the cellular system. If no information about the RNA targets is available, *cat*RAPID can represent a promising tool for the investigation of the protein's genome-wide RNA-binding potential against an RNA sequence library. Here we suggest a pipeline that could be followed in such circumstance:

1. If the RNA binding potential of a given protein is unknown, *cat*RAPID *signature* [71] can be used to predict it, along with the putative RNA-binding regions. This approach is particularly recommended if a potential RBP needs to be selected from a panel of candidate proteins;
2. *cat*RAPID *omics* [69] can then be used to predict the interactions between the protein and a precompiled or custom RNA library. The result is a ranked list of protein–RNA pairs;
3. If the protein of interest is human, its co-expression with the putative interactors in different tissues can be evaluated using *cat*RAPID *express* [70];
4. Once the most promising binding partners have been identified, *cat*RAPID *strength* can be employed to evaluate the strength of each interaction [67];
5. Finally, *cat*RAPID *fragments* [66] can be run on the highest scoring protein–RNA pairs to predict the binding sites. Interactions with long RNAs can be analyzed using *Global Score* [65] or *omiXcore* [68].

By narrowing down the number of potential targets and by suggesting the most likely binding sites, such approach can be employed to guide further experimental and computational analyses. Being a predictive tool, there is always the chance that *cat*RAPID may fail in identifying valid RNA targets. Prediction accuracy depends on the set provided to train the algorithm. As more and more data become available, retraining of the algorithm will be necessary to achieve better performances.

*cat*RAPID alternatives

*cat*RAPID is only one of several possible computational methods developed for predicting protein–RNA interactions. We would like to mention here some valid alternatives:

Biochemical Society Transactions (2020) **48** 1529–1543
https://doi.org/10.1042/BST20191059

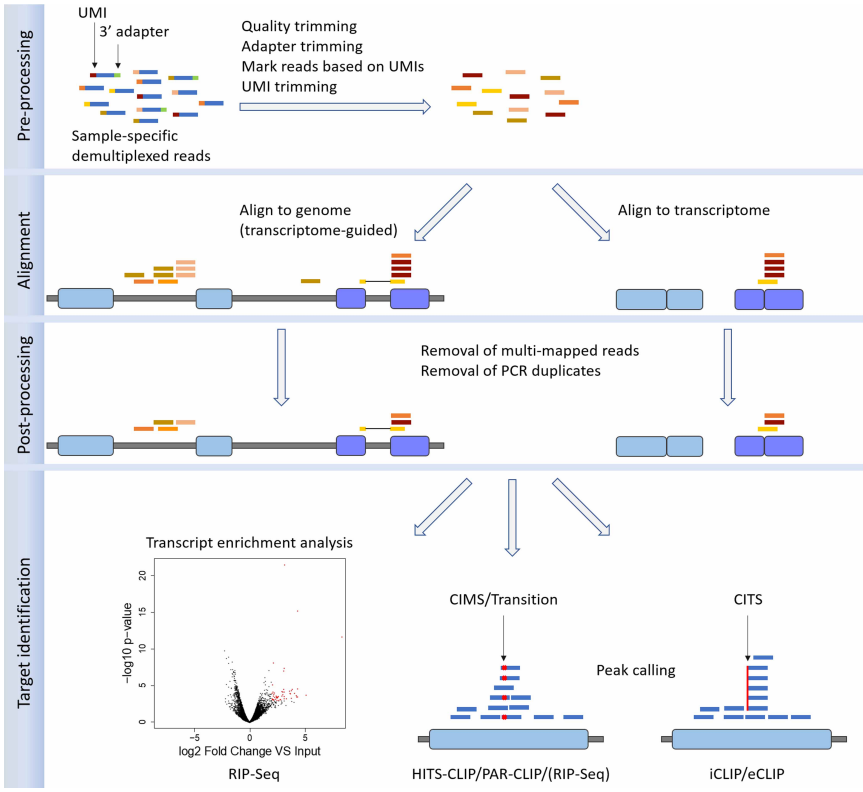


Figure 2. Bioinformatics analysis of CLIP-Seq and RIP-Seq data.

After being de-multiplexed based on sample-specific barcodes, reads undergo a **pre-processing** phase. UMIs are not always used, being more common in iCLIP and eCLIP protocols. When employed, they are sometimes used to remove PCR duplicates directly at this stage, but in most cases reads are simply marked based on UMI sequence, as shown by the colours assigned to trimmed reads. After the **alignment** of reads to the genome or to the transcriptome is performed, **post-processing** is needed to filter out multi-mapped reads and to collapse reads mapping at the same position, that are likely to represent PCR duplicates; if UMI-based read marking occurred, natural duplicates, which map at the same place but have different UMIs, can be retained, as shown here. **RNA target identification** and binding site detection strategy depend on the protocol. Roughly, such approaches can be divided into transcript enrichment analysis (RIP-Seq), which is analogous to differential expression analysis, and peak-calling (all protocols). Single-nucleotide resolution can be achieved using CIMSs in HITS-CLIP, transitions in PAR-CLIP and CITSs in iCLIP/eCLIP. HITS-CLIP experiments do not always produce clear and usable CIMS patterns.

- LncPro [72]: similar to *cat*RAPID in the employment of RNA secondary structure, hydrogen-bonding and van der Waals interaction propensities. It is designed to predict whether a specific long non-coding RNA interacts with one or more protein sequences. Propensities are calculated for both protein and RNA and a probability score ranging from 0 to 100 is generated;

PORTLAND PRESS

Biochemical Society Transactions (2020) **48** 1529–1543
https://doi.org/10.1042/BST20191059

Table 1 A summary of the different *cat*RAPID implementations

Name of the algorithm	Description	Input	Output
*cat*RAPID *fragments* [66]	It divides inputted protein and RNA into fragments and computes the interaction propensity between each fragment.	• A protein sequence in FASTA format. • An RNA sequence in FASTA format.	• Interaction profile plot[a]. • Interaction matrix[b]. • Table of interacting protein–RNA fragments.
Global Score [65] *omiXcore* [68]	A variant of *cat*RAPID fragments calibrated on CLIP-Seq data, it is able to predict interaction with >1000 nt long RNAs and to provide an overall interaction score.	• A protein sequence in FASTA format. • An RNA sequence in FASTA format.	• Interaction profile plot[a] • Interaction matrix[b]. • Table of interacting protein–RNA fragments.
*cat*RAPID *omics* [69]	It computes the interactions between a molecule (protein/RNA) and the reference set (transcriptome/ nucleotide-binding proteome) of a model organism.	• Protein/RNA sequence in FASTA format. • Reference set of protein/RNA sequences.	• Graphical representation of protein sequence/ domains. • Pie chart with ranking distribution[c]. • Table of interacting protein–RNA pairs.
*cat*RAPID *express* [70]	It allows the identification of co-expressed protein–RNA pairs in human tissues.	• A protein sequence in FASTA format • An RNA sequence in FASTA format. • (Only one protein sequence or one RNA sequence is required for the omics option).	• Correlation coefficient representing the coexpression of the protein–RNA pair. • Interaction heatmap[d]. • Table of tissue expression.
*cat*RAPID *signature* [71]	It scans a protein sequence for RNA-binding regions.	• One or more protein sequences in FASTA format.	• Overall binding score. • Binding propensity plot[e].
*cat*RAPID *library* [69]	It allows the creation of a new reference set for *cat*RAPID omics.	• One or more protein or RNA sequences.	• A library ID that can be used in catRAPID *omics*.
*cat*RAPID *strength* [67]	It computes the interaction strength of a protein–RNA pair with respect to a reference set of sequences of similar length.	• A protein sequence in FASTA format • A RNA sequence in FASTA format.	• Table of interaction strength (significance of interaction propensity). • Cumulative distribution function plots of protein-RNA interaction score[f].

[a]The interaction profile plot represents the interaction score (*y*-axis) of the protein along the RNA sequence (*x*-axis), giving information about the transcript regions that are most likely to be bound by the protein;
[b]The interaction matrix is an heatmap showing the interaction propensity between each possible fragment of the protein (*y*-axis) and the RNA (*x*-axis);
[c]The pie chart shows the proportion of targets having High, Moderate and Low star rating score. Star rating score weights the interaction based on the interaction propensity, the presence of RNA/DNA binding domains and the presence of known RNA motifs;
[d]The interaction heatmap shows the interaction score of the individual amino acid-nucleotide pairs;
[e]The binding propensity plot reports, for each amino acid (*x*-axis), the propensity to be part of a binding region;
[f]The Cumulative distribution function plots report the interaction score of the query protein–RNA pair within the distribution of the interaction scores from the reference set.
A more detailed explanation of the different algorithms is available on *cat*RAPID tutorial page (http://s.tartaglialab.com/static_files/shared/tutorial.html) and documentation page (http://s.tartaglialab.com/static_files/shared/documentation.html).

- RPISeq [73]: protein and RNA sequences are encoded into features that are then used to train Support Vector Machine (SVM) and Random Forests classifiers;
- RPI-Pred [74]: it combines RNA and protein sequences with predicted or actual 3D structures. The features are then used to train an SVM classifier;
- iDeepS [75]: a deep learning-based method that exploits convolutional neural networks (CNNs) trained on RNA sequences and predicted secondary structures. At the end of the pipeline, a classification layer is

Biochemical Society Transactions (2020) **48** 1529–1543
https://doi.org/10.1042/BST20191059

Table 2 Methods for *in vitro* characterisation of protein–RNA binding

Method	Principle of detection	Sample requirements	Detection range	Sample capacity	Direct measurements
EMSA [90]	Detection of RNA–protein complex' electrophoretic mobility properties, typically different compared to free RNA.	• Labelled RNA. • nM concentrations of RNA and protein.	$\geq 10^{-18}$ mol RNA.	0.5–500 µl depending on electrophoresis setup.	K_d, n
Filter binding assay [88]	Quantification of ^{32}P-labelled RNA via imagine screen or scintillation counter.	• About 0.1 µM labelled RNA (usually with ^{32}P). • Purified protein serial dilutions.	$\geq 10^{-15}$ mol RNA.	Multi-well plate dot-blot setup.	k_{on}, k_{off}, K_d, n
Fluorescence anisotropy [91,92]	Changes in fluorescence anisotropy or polarisation of excitation light upon binding.	• Fluorescent labelling of one of the partners. • 1 nM RNA.	nM ranges of fluorophores.	Multi-well plates.	K_d
FRET [93,94]	Energy transfer of between fluorophores detected as a change in fluorescence intensity.	• Two fluorophores, either one on each partner or strategically placed on one for structural studies.	single-molecule experiments.	Single molecule to multi-well plates.	K_d, k_{on}, k_{off}, distance between fluorophores.
SPR [95]	Variations in the refractive index of polarised laser light upon molecular binding.	• About 200 µl 25 nM RNA/sensor. • Variable conc. of protein (ideally 100-times K_d), up to 4 ml of sample. • Immobilisation of one partner required.	1 pM < K_d < 1 mM	Up to 16 channels with microfluidics.	K_d, k_{on}, k_{off}
BLI [96]	Detection of the variation of refracted white light upon the binding of the interaction partner to the immobilised ligand on the optical fibres.	• 1–50 µg/ml of ligand, immobilised on biosensor. • 1 nM–µM of receptor. • 5–250 µl of sample per measurement.	1 nM < K_d < 10 mM	Single channel, 5 min per measurement (BLItz) or multi-well plate, 1–8 simultaneous channels.	K_d, k_{on}, k_{off}
MST [97]	Variations in temperature-induced fluorescence emission of a target as a function of the concentration of a non-fluorescent ligand.	• 1–20 µl, nM–µM concentrations. • Fluorescent labelling.	pM < K_d < mM	Up to 96 samples per run in a multi-capillary system.	K_d
switchSENSE [98]	Voltage-dependent variations of the movement of short fluorescent DNA nanolevers attached to a gold surface upon binding of an analyte.	• Immobilisation of one binding partner. • 20 µl of 1 µM RNA for biochip saturation. • 250 µl of 0.2 µM protein.	nM < K_d < µM	Four flow channels with six microelectrodes for sampling per chip	K_d, k_{on}, k_{off}, R_h
ITC [99]	Measuring the heat consumed/released during titration of sample with the ligand in regard to reference cell.	• 200 µl–2 ml of 1-2 µM receptor. • 40–500 µl 10× concentration ligand.	• K_d > nM (direct measurements). • K_d > pM (competitive binding).	single cell.	K_d, ΔH, n

Kinetic constants are measured directly and are used as basis for equilibrium thermodynamic parameters calculations, apart from ITC where the reaction enthalpy can be obtained without relying on kinetic data. K_d: equilibrium dissociation constant; k_{on}: association rate constant; k_{off}: dissociation rate constant; n: stechiometry of binding; R_h: hydrodynamic radius (radius of a theoretical sphere with the same translational diffusion coefficient); ΔH: reaction enthalpy.

PORTLAND PRESS

Biochemical Society Transactions (2020) 48 1529–1543
https://doi.org/10.1042/BST20191059

responsible for RBP binding sites prediction. Deep learning models are generated individually for each RBP based on available CLIP-Seq data, allowing the formulation of predictions on a limited set of proteins.

Functional characterization of RNA targets and binding sites

Once the RNA targets of a protein have been determined, further analyses are necessary in order to verify the reliability of the results and to gain insights into the biological function of the RBP. Both tasks can be approached by looking at the function of target RNAs. A common way to do that is to perform an Over Representation Analysis (ORA) [76], which consists of identifying over-represented functional categories in a list of genes. Enrichment is evaluated against a background composed of all the expressed genes. A more sophisticated approach, called Gene Set Enrichment Analysis (GSEA) [76], involves ranking genes based on a certain score and evaluating if some categories are enriched at the top or the bottom of the ranked list. An implementation of this method that is specific for CLIP-Seq data is provided by the Seten tool [77]. This program requires as input a set of CLIP-Seq peaks, each with a score assigned by the peak-caller, but it could also work starting from predicted binding sites, as long as an interaction score is provided.

Another common analysis consists in scanning the identified binding sites in order to detect common patterns highlighting RBP binding preferences (motif analysis). Sequence motifs recurring in large sets of binding sites can be discovered using different tools, like MEME-ChIP [78] and SeAMotE [79]. Both tools start from a set of sub-sequences identified in the positive sequence set and evaluate their enrichment with respect to a control sequence set (unbound RNAs). A more recent tool, named mCross, exploits the single-nucleotide resolution offered by CLIP-Seq techniques to enhance the accuracy of *de novo* motif discovery [80].

Sequence alone may not be sufficient to fully explain the binding specificity of an RBP: a sequence motif could be accessible only when put in a proper secondary structure context. Tools like GraphProt [81], ssHMM [82] and BEAM [83] are able to detect motifs encoding both sequence and secondary structure information.

In vitro validation of predicted RBP–RNA interactions

To validate the prediction accuracy of the computational analysis proposed so far, it is ideal to evaluate each interacting pair within a controlled environment. A most accurate validation should start from the screening of potential binders, followed by the precise determination of binding sites and kinetic and thermodynamic parameters, and completed with structural insights into the drivers of the interaction (Figure 1). A comprehensive review of the methodology is beyond the scope of this article and for more details we refer to a number of recent reviews of the field [16,17,84–86].

Kinetics and thermodynamics of RBP–RNA interactions

As for other molecular interactions, the binding between a protein and an RNA molecule is kinetically characterised by the rate at which they associate (k_{on}) and dissociate (k_{off}). Conventionally, the dissociation constant (K_d), which is the ratio between k_{off} and k_{on} at the chemical equilibrium, is used to express the binding affinity: the lower the K_d, the greater the affinity. It is however important to note that binding pairs with the same K_d may have different k_{on} and k_{off} and therefore different binding mechanisms. To add another layer of complexity, many RBPs have multiple binding regions that may vastly differ in their affinity towards the same RNA [87].

There are several established methods for determining the kinetic and thermodynamic parameters of binding (Table 2). Techniques such as electrophoretic mobility shift assay (EMSA) and filter binding assay can be useful to estimate binding affinities with basic molecular biology tools [88,89]. Both these methods represent a viable first step analysis, especially because of short protocols and limited amounts of samples required. The latter criterion can be crucial in protein–RNA interaction studies, since some RBPs can be very difficult to isolate, tagged RNA synthesis can be expensive and advanced methods for more reliable and accurate determination of molecular binding characteristics generally have specific sample requirements and higher operation costs. Despite these advantages, since EMSA and filter binding experiments are performed within conditions very distinct from the *in vivo* ones (polyacrylamide gel and nitrocellulose filter, respectively), more reliable kinetic data may be obtained by other techniques. Examples of such methods are bio-layer interferometry (BLI), multi-channel surface plasmon resonance (SPR), microscale thermophoresis (MST), the employment of

Biochemical Society Transactions (2020) **48** 1529–1543
https://doi.org/10.1042/BST20191059

fluorescence resonance energy transfer (FRET), and the most recent switchSENSE (Table 2, [86–89, 95]). Some of these require the labelling of one of the interactors with fluorescent dyes, as in the case of MST and FRET, or sample immobilisation on biosensors, as needed for SPR and BLI experiments. These requisites can impose certain structural restraints and thus compromise the binding and the results obtained. Comparing how the same RBP binds to RNA molecules that differ only in one nucleotide can help the identification of the RNA portion physically interacting with the protein. However, the determination of the exact nucleotides involved in

Table 3 A short overview of the major structural biology techniques with a comparison of their advantages and disadvantages for the study of protein–RNA interactions

Method	Principle of detection	Resolution	Sample requirements	Pros/Cons
NMR [111,114]	Detection of the electric current, induced by the magnetization of the non-equilibrium spins in a magnetic field. Upon Fourier transform, the results can be used to determine structural constraints and produce a molecular model.	atomic (<2 Å).	• Isotope labelling, side-chain deuteration essential for larger complexes to avoid lengthy relaxation times. • Protein concentration varies according to MW.	• Solution-based, can observe time-resolved experiments and kinetics, most accessible on the list, possibilities of differential isotope labelling, saturation transfer experiments and more. • Poor signal-to-noise ratio, line broadening and complex spectra with higher molecular mass complexes.
X-ray crystallography [115,116]	Detection of diffracted X-ray photons, scattered by the crystal, from which an 3D electron density map is calculated, which is then used to build the molecular structure model.	atomic (<2 Å).	Crystals of the protein–RNA complex, frozen in liquid nitrogen.	• Highest resolution limit with free electron lasers. • Relies on quality crystals, often difficult to obtain.
Cryogenic electron microscopy [117]	Based on electron microscopy, the sample images are grouped into specific projections, with a 3D model calculated based on them.	high (<5 Å).	Monodisperse sample blotted onto grids and frozen under cryogenic conditions.	• Solution based, flexible buffer components, no need for crystals etc. • Maximum resolution limit around 3.5 Å for molecules with a MW ∼50 kDa. • Very difficult to obtain sufficient quality data for determination of structures of elements with MW < 150 kDa under the resolution of 5 Å.
Small angle scattering [110,112,118]	Detection of diffracted X-ray photons (SAXS) or neutrons (SANS) on sample solutions under small angles (typically <10°), from which a scattering curve and a 3D shape can be calculated.	medium (>10 Å).	• Monodisperse sample, dilution series from 1 mg/ml to 20 mg/ml. • Possible deuteration or isotope labelling for SANS studies.	• Investigation of molecule shape as well as other information, selective deuteration can provide valuable contrast (SANS). • Need for monodisperse sample, rather high protein concentrations for the dilution curve, no exact molecular structure.

PORTLAND PRESS

Biochemical Society Transactions (2020) **48** 1529–1543
https://doi.org/10.1042/BST20191059

the binding can be highly challenging without structural studies. Once the RNA targets have been precisely characterised, selective mutations on the protein region thought to be responsible for the binding may be useful to determine which amino acids are directly responsible for the interaction. The same approach could be extended to mutating selected nucleotides or regions of the RNA molecules for further characterisation of the binding spots without obtaining an atomic resolution structure.

Kinetic studies are essential for determining the binding propensity and validating different binding partners [10]. However, experiments conducted *in vitro* with isolated components often do not allow studies under physiologically relevant conditions. Isolated proteins can be sensitive to higher temperatures, especially under prolonged experiments that far exceed their in-cell lifespans. Since chemical equilibriums are temperature-dependent, *in vivo* kinetics may thus significantly differ from those measured *in vitro*. The thermodynamics of binding can also reveal the energetic landscape of protein–RNA interactions [100–102]. It is therefore very important to measure the thermodynamics parameters directly, when experimentally possible, or to calculate them. Isothermal titration calorimetry (ITC) allows direct measurements of the equilibrium constant, stoichiometry and reaction enthalpy (ΔH) at a given temperature [99]. Using the van't Hoff equation, these can then be used to determine the temperature dependency of the equilibrium constant. SPR and MST also allow calculation of thermodynamic data based on the stable temperature of the measurement cell, while BLI is considered less reliable.

Structural approaches of studying RBP–RNA interactions

Kinetic and thermodynamic data obtained from interaction studies provide a good numerical description of the binding and, through the use of RNA-centric methods, a library of RNA sequences with high binding propensities [103]. However, protein and RNA sequences alone may not be sufficient to fully characterise their binding specificity. This is especially true in the case of non-canonical RBPs that do not contain a consensus RNA-recognition sequence [104]. RNA structure has been shown to be the driver behind most non-canonical RBP binding events, with highly structured RNAs having bigger protein interactomes [105,106]. The approaches to determine the structure of separate components have been extensively reviewed [107,108]. Among the most pertinent techniques to determine the macromolecular structure of a complex, there are X-ray crystallography, nuclear magnetic resonance (NMR) and cryogenic electron microscopy (cryo-EM) (Table 3). However, the definition of structural details by these methodologies can be challenging [107]. A large number of RBPs contain intrinsically disordered regions and tend to form macromolecular condensates, making their crystallisation impractical; NMR, that bypasses the need for crystals, is dependent on the molecular weight of the complexes that can make relaxation times slow and signal-to-noise ratio poor; obtaining results at atomic resolution with cryo-EM remains difficult [108]. These limitations make a strong case for the employment of complementary structural biology techniques [107]. A particularly promising development is the advance of small angle scattering and computational methods for data analysis, in particular small angle neutron scattering (SANS) [109]. Selective deuteration enables a higher degree of contrast between binding partners and can therefore provide a rough position for each component of the complex [110]. Data obtained from the above mentioned methods can be used as structural restraints for molecular dynamics simulations and data-driven docking [111,112] and can be integrated together in a hybrid multi-level approach for studying RNA-protein complexes, thus completing the full circle of integrated methodologies [75,113,119,120].

Perspectives

- Our understanding of many physiological and pathological phenomena cannot be exempted from an in-depth knowledge of protein–RNA interactions underlying them.
- Such comprehension, which goes from the identification of targets RNAs to binding modes characterization, requires a multidisciplinary approach involving biochemistry, molecular biology, bioinformatics and physics.

Biochemical Society Transactions (2020) **48** 1529–1543
https://doi.org/10.1042/BST20191059

PORTLAND PRESS

- As new and more powerful high-throughput methods and predictors are being developed, an integrated and productive usage of both approaches becomes more and more feasible. For instance, predictive tools such as *cat*RAPID, which is general enough to be applied to any protein–RNA pair, could also be employed to improve the specificity of omics studies.

Competing Interests
The authors declare that there are no competing interests associated with the manuscript.

Funding
The research leading to this work has been supported by European Research Council (RIBOMYLOME_309545 and ASTRA_855923), by the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 754490 and under projects IASIS_727658 and INFORE_825080, by the Spanish Ministry of Economy and Competitiveness BFU2017-86970-P as well as the collaboration with Peter St. George-Hyslop financed by the Wellcome Trust.

Author Contributions
A.C., J.R. and A.V. drafted the sections of the review pertinent to their expertise. E.Z. and G.G.T. guided, supervised and coordinated the work. All authors contributed to writing and editing the manuscript and to the original production of all figures and tables provided.

Acknowledgements
The authors would like to thank all members of Tartaglia lab.

Abbreviations
BLI, biolayer interferometry; CIMS, crosslinking-induced mutation sites; CITS, crosslinking-induced truncation sites; CLIP, cross-linking followed by immunoprecipitation; CLIP-Seq, CLIP followed by followed by high-throughput sequencing; eCLIP, enhanced CLIP; FRET, Förster resonance energy transfer; HITS-CLIP, high-throughput sequencing CLIP; iCLIP, individual nucleotide-resolution CLIP; IP, immunoprecipitation; MST, microscale thermophoresis; NMR, nuclear magnetic resonance; PAR-CLIP, photoactivatable ribonucleoside CLIP; RBP, RNA-binding protein; RIP, RNA immunoprecipitation; RIP-Seq, RIP followed by high-throughput sequencing; SPR, surface plasmon resonance; SVM, support vector machine; UMI, unique molecule identifier.

References
1 Castello, A., Fischer, B., Frese, C.K., Horos, R., Alleaume, A.M., Foehr, S. et al. (2016) Comprehensive identification of RNA-binding domains in human cells. *Mol. Cell* **63**, 696–710 https://doi.org/10.1016/j.molcel.2016.06.029
2 Blackinton, J.G. and Keene, J.D. (2014) Post-transcriptional RNA regulons affecting cell cycle and proliferation. *Semin. Cell Dev. Biol.* **34**, 44–54 https://doi.org/10.1016/j.semcdb.2014.05.014
3 Bernabò, P., Viero, G. and Lencioni, V. (2020) A long noncoding RNA acts as a posttranscriptional regulator of heat shock protein (HSP70) synthesis in the cold hardy Diamesa tonsa under heat shock. *PLoS One* **15**, e0227172 https://doi.org/10.1371/journal.pone.0227172
4 García-Mauriño, S.M., Rivero-Rodríguez, F., Velázquez-Cruz, A., Hernández-Vellisca, M., Díaz-Quintana, A., De la Rosa, M.A. et al. (2017) RNA binding protein regulation and cross-talk in the control of AU-rich mRNA fate. *Front. Mol. Biosci.* **4**, 71 https://doi.org/10.3389/fmolb.2017.00071
5 Hentze, M.W., Castello, A., Schwarzl, T. and Preiss, T. (2018) A brave new world of RNA-binding proteins. *Nat. Rev. Mol. Cell Biol.* **19**, 327–341 https://doi.org/10.1038/nrm.2017.130
6 Moore, S., Järvelin, A.I., Davis, I., Bond, G.L. and Castello, A. (2018) Expanding horizons: new roles for non-canonical RNA-binding proteins in cancer. *Curr. Opin. Genet. Dev.* **48**, 112–120 https://doi.org/10.1016/j.gde.2017.11.006
7 Backlund, M., Stein, F., Rettel, M., Schwarzl, T., Perez-Perri, J.I., Brosig, A. et al. (2020) Plasticity of nuclear and cytoplasmic stress responses of RNA-binding proteins. *Nucleic Acids Res.* **48**, 4740 https://doi.org/10.1093/nar/gkaa256
8 Gebauer, F., Preiss, T. and Hentze, M.W. (2012) From cis-regulatory elements to complex RNPs and back. *Cold Spring Harb. Perspect. Biol.* **4**, 1–14 https://doi.org/10.1101/cshperspect.a012245
9 Mukherjee, N., Wessels, H.H., Lebedeva, S., Sajek, M., Ghanbari, M., Garzia, A., et al. (2019) Deciphering human ribonucleoprotein regulatory networks. *Nucleic Acids Res.* **47**, 570–581 https://doi.org/10.1093/nar/gky1185
10 Jankowsky, E. and Harris, M.E. (2015) Specificity and nonspecificity in RNA-protein interactions. *Nat. Rev. Mol. Cell Biol.* **16**, 533–544 https://doi.org/10.1038/nrm4032

PORTLAND PRESS

Biochemical Society Transactions (2020) **48** 1529–1543
https://doi.org/10.1042/BST20191059

11 Liao, J.-Y.Y., Yang, B., Zhang, Y.-C.C.Y., Wang, X.-J.J., Ye, Y., Peng, J.-W.W., et al. (2020) EuRBPDB: a comprehensive resource for annotation, functional and oncological investigation of eukaryotic RNA binding proteins (RBPs). *Nucleic Acids Res.* **48**, 307–313 https://doi.org/10.1093/nar/gkz823
12 Gerstberger, S., Hafner, M. and Tuschl, T. (2014) A census of human RNA-binding proteins. *Nat. Rev. Genet.* **15**, 829–845 https://doi.org/10.1038/nrg3813
13 Jensen, K.B. and Darnell, R.B. (2008) CLIP: crosslinking and immunoprecipitation of in vivo RNA targets of RNA-binding proteins. *Methods Mol. Biol.* **488**, 85–98 https://doi.org/10.1007/978-1-60327-475-3_6
14 Ramanathan, M., Majzoub, K., Rao, D.S., Neela, P.H., Zarnegar, B.J., Mondal, S., et al. (2018) RNA-protein interaction detection in living cells. *Nat. Methods* **15**, 207–212 https://doi.org/10.1038/nmeth.4601
15 Ule, J., Jensen, K., Mele, A. and Darnell, R.B. (2005) CLIP: a method for identifying protein-RNA interaction sites in living cells. *Methods* **37**, 376–386 https://doi.org/10.1016/j.ymeth.2005.07.018
16 Dasti, A., Cid-Samper, F., Bechara, E. and Tartaglia, G.G. (2019) RNA-centric approaches to study RNA-protein interactions in vitro and in silico. *Methods* **178**, 11–18 https://doi.org/10.1016/j.ymeth.2019.09.011
17 Ye, X. and Jankowsky, E. (2019) High throughput approaches to study RNA-protein interactions in vitro. *Methods* **178**, 3–10 https://doi.org/10.1016/j.ymeth.2019.09.006
18 Lam, J.H., Li, Y., Zhu, L., Umarov, R., Jiang, H., Héliou, A., et al. (2019) A deep learning framework to predict binding preference of RNA constituents on protein surface. *Nat. Commun.* **10**, 1–13 https://doi.org/10.1038/s41467-018-07882-8
19 Zhou, Y.-K., Shen, Z.-A., Yu, H., Luo, T., Gao, Y. and Du, P.-F. (2020) Predicting lncRNA–protein interactions with miRNAs as mediators in a heterogeneous network model. *Front. Genet.* **10**, 1341 https://doi.org/10.3389/fgene.2019.01341
20 Qiu, J., Bernhofer, M., Heinzinger, M., Kemper, S., Norambuena, T., Melo, F. et al. (2020) ProNA2020 predicts protein–DNA, protein–RNA, and protein–protein binding proteins and residues from sequence. *J. Mol. Biol.* **432**, 2428–2443 https://doi.org/10.1016/j.jmb.2020.02.026
21 Adjeroh, D., Allaga, M., Tan, J., Lin, J., Jiang, Y., Abbasi, A. et al. (2018) Feature-based and string-based models for predicting RNA-protein interaction. *Molecules* **23**, 697 https://doi.org/10.3390/molecules23030697
22 Pan, X., Yang, Y., Xia, C.Q., Mirza, A.H. and Shen, H.B. (2019) Recent methodology progress of deep learning for RNA–protein interaction prediction. *Wiley Interdiscip. Rev. RNA* https://doi.org/10.1002/wrna.1544
23 catRAPID
24 Bellucci, M., Agostini, F., Masin, M. and Tartaglia, G.G. (2011) Predicting protein associations with long noncoding RNAs. *Nat. Methods* **8**, 444–445 https://doi.org/10.1038/nmeth.1611
25 Gagliardi, M. and Matarazzo, M.R. (2016) RIP: RNA immunoprecipitation. *Methods Mol. Biol.* **1480**, 73–86 https://doi.org/10.1007/978-1-4939-6380-5_7
26 Nicholson, C.O., Friedersdorf, M. and Keene, J.D. (2017) Quantifying RNA binding sites transcriptome-wide using DO-RIP-seq. *RNA* **23**, 32–46 https://doi.org/10.1261/rna.058115.116
27 Nicholson, C.O., Friedersdorf, M.B., Bisogno, L.S. and Keene, J.D. (2017) DO-RIP-seq to quantify RNA binding sites transcriptome-wide. *Methods* **118–119**, 16–23 https://doi.org/10.1016/j.ymeth.2016.11.004
28 Singh, G., Ricci, E.P. and Moore, M.J. (2014) RIPiT-Seq: a high-throughput approach for footprinting RNA: protein complexes. *Methods* **65**, 320–332 https://doi.org/10.1016/j.ymeth.2013.09.013
29 Philippe, N., Salson, M., Commes, T. and Rivals, E. (2013) CRAC: an integrated approach to the analysis of RNA-seq reads. *Genome Biol.* **14**, R30 https://doi.org/10.1186/gb-2013-14-3-r30
30 Ule, J., Jensen, K.B., Ruggiu, M., Mele, A., Ule, A. and Darnell, R.B. (2003) CLIP identifies nova-regulated RNA networks in the brain. *Science* **302**, 1212–1215 https://doi.org/10.1126/science.1090095
31 Tiedge, H. (1991) The use of UV light as a cross-Linking agent for cells and tissue sections in in situ hybridization. *DNA Cell Biol.* **10**, 143–147 https://doi.org/10.1089/dna.1991.10.143
32 Licatalosi, D.D., Mele, A., Fak, J.J., Ule, J., Kayikci, M., Chi, S.W., et al. (2008) HITS-CLIP yields genome-wide insights into brain alternative RNA processing. *Nature* **456**, 464–469 https://doi.org/10.1038/nature07488
33 Huppertz, I., Attig, J., D'Ambrogio, A., Easton, L.E., Sibley, C.R., Sugimoto, Y. et al. (2014) iCLIP: protein-RNA interactions at nucleotide resolution. *Methods* **65**, 274–287 https://doi.org/10.1016/j.ymeth.2013.10.011
34 König, J., Zarnack, K., Rot, G., Curk, T., Kayikci, M., Zupan, B. et al. (2010) iCLIP reveals the function of hnRNP particles in splicing at individual nucleotide resolution. *Nat. Struct. Mol. Biol.* **17**, 909–915 https://doi.org/10.1038/nsmb.1838
35 Van Nostrand, E.L., Pratt, G.A., Shishkin, A.A., Gelboin-Burkhart, C., Fang, M.Y., Sundararaman, B., et al. (2016) Robust transcriptome-wide discovery of RNA-binding protein binding sites with enhanced CLIP (eCLIP). *Nat. Methods* **13**, 508–514 https://doi.org/10.1038/nmeth.3810
36 Ascano, M., Hafner, M., Cekan, P., Gerstberger, S. and Tuschl, T. (2012) Identification of RNA-protein interaction networks using PAR-CLIP. *Wiley Interdiscip. Rev. RNA* **3**, 159–177 https://doi.org/10.1002/wrna.1103
37 Zarnegar, B.J., Flynn, R.A., Shen, Y., Do, B.T., Chang, H.Y. and Khavari, P.A. (2016) IrCLIP platform for efficient characterization of protein-RNA interactions. *Nat. Methods* **13**, 489–492 https://doi.org/10.1038/nmeth.3840
38 Kim, B., Jeong, K. and Kim, V.N. (2017) Genome-wide mapping of DROSHA cleavage sites on primary microRNAs and noncanonical substrates. *Mol. Cell* **66**, 258–269.e5 https://doi.org/10.1016/j.molcel.2017.03.013
39 Klockenbusch, C., O'Hara, J.E. and Kast, J. (2012) Advancing formaldehyde cross-linking towards quantitative proteomic applications. *Anal. Bioanal. Chem.* **404**, 1057–1067 https://doi.org/10.1007/s00216-012-6065-9
40 Kivioja, T., Väharautio, A., Karlsson, K., Bonke, M., Enge, M., Linnarsson, S. et al. (2012) Counting absolute numbers of molecules using unique molecular identifiers. *Nat. Methods* **9**, 72–74 https://doi.org/10.1038/nmeth.1778
41 Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R. and Salzberg, S.L. (2013) Tophat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol.* **14**, R36 https://doi.org/10.1186/gb-2013-14-4-r36
42 Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S. et al. (2012) STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 https://doi.org/10.1093/bioinformatics/bts635

Biochemical Society Transactions (2020) **48** 1529–1543
https://doi.org/10.1042/BST20191059

43 Haberman, N., Huppertz, I., Attig, J., König, J., Wang, Z., Hauer, C., et al. (2017) Insights into the design and interpretation of iCLIP experiments. *Genome Biol.* **18**, 7 https://doi.org/10.1186/s13059-016-1130-x
44 Langmead, B. and Salzberg, S.L. (2012) Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 https://doi.org/10.1038/nmeth.1923
45 Granneman, S., Kudla, G., Petfalski, E. and Tollervey, D. (2009) Identification of protein binding sites on U3 snoRNA and pre-rRNA by UV cross-linking and high-throughput analysis of cDNAs. *Proc. Natl Acad. Sci. U.S.A.* **106**, 9613–9618 https://doi.org/10.1073/pnas.0901997106
46 Zhang, C. and Darnell, R.B. (2011) Mapping in vivo protein-RNA interactions at single-nucleotide resolution from HITS-CLIP data. *Nat. Biotechnol.* **29**, 607–614 https://doi.org/10.1038/nbt.1873
47 Kishore, S., Jaskiewicz, L., Burger, L., Hausser, J., Khorshid, M. and Zavolan, M. (2011) A quantitative analysis of CLIP methods for identifying binding sites of RNA-binding proteins. *Nat. Methods* **8**, 559–567 https://doi.org/10.1038/nmeth.1608
48 Hafner, M., Landthaler, M., Burger, L., Khorshid, M., Hausser, J., Berninger, P., et al. (2010) Transcriptome-wide identification of RNA-binding protein and microRNA target sites by PAR-CLIP. *Cell* **141**, 129–141 https://doi.org/10.1016/j.cell.2010.03.009
49 Li, H. and Durbin, R. (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 https://doi.org/10.1093/bioinformatics/btp324
50 Kerpedjiev, P., Frellsen, J., Lindgreen, S. and Krogh, A. (2014) Adaptable probabilistic mapping of short reads using position specific scoring matrices. *BMC Bioinformatics* **15**, 100 https://doi.org/10.1186/1471-2105-15-100
51 Kloetgen, A., Borkhardt, A., Hoell, J.I. and McHardy, A.C. (2016) The PARA-suite: PAR-CLIP specific sequence read simulation and processing. *PeerJ* **4**, e2619 https://doi.org/10.7717/peerj.2619
52 Wang, T., Xiao, G., Chu, Y., Zhang, M.Q., Corey, D.R. and Xie, Y. (2015) Design and bioinformatics analysis of genome-wide CLIP experiments. *Nucleic Acids Res.* **43**, 5263–5274 https://doi.org/10.1093/nar/gkv439
53 Zhang, Z. and Xing, Y. (2017) CLIP-seq analysis of multi-mapped reads discovers novel functional RNA regulatory sites in the human transcriptome. *Nucleic Acids Res.* **45**, 9260–9271 https://doi.org/10.1093/nar/gkx646
54 Fu, Y., Wu, P.H., Beane, T., Zamore, P.D. and Weng, Z. (2018) Elimination of PCR duplicates in RNA-seq and small RNA-seq using unique molecular identifiers. *BMC Genomics* **19**, 531 https://doi.org/10.1186/s12864-018-4933-1
55 Lu, Z., Guan, X., Schmidt, C.A. and Matera, A.G. (2014) RIP-seq analysis of eukaryotic Sm proteins identifies three major categories of Sm-containing ribonucleoproteins. *Genome Biol.* **15**, 1–23 https://doi.org/10.1186/gb-2014-15-1-r1
56 Moore, K.S. and t Hoen, P.A.C. (2019) Computational approaches for the analysis of RNA–protein interactions: a primer for biologists. *J. Biol. Chem.* **294**, 1–9 https://doi.org/10.1074/jbc.REV118.004842
57 Baquero-Perez, B., Antanaviciute, A., Yonchev, I.D., Carr, I.M., Wilson, S.A. and Whitehouse, A. (2019) The tudor SND1 protein is an m^6a RNA reader essential for replication of kaposi's sarcoma-associated herpesvirus. *eLife* **8**, e47261 https://doi.org/10.7554/eLife.47261
58 Li, Y., Zhao, D.Y., Greenblatt, J.F. and Zhang, Z. (2013) RIPSeeker: a statistical package for identifying protein-associated transcripts from RIP-seq experiments. *Nucleic Acids Res.* **41**, e94 https://doi.org/10.1093/nar/gkt142
59 Kucukural, A., Özadam, H., Singh, G., Moore, M.J. and Cenik, C. (2013) ASPeak: an abundance sensitive peak detection algorithm for RIP-Seq. *Bioinformatics* **29**, 2485–2486 https://doi.org/10.1093/bioinformatics/btt428
60 Uren, P.J., Bahrami-Samani, E., Burns, S.C., Qiao, M., Karginov, F.V., Hodges, E. et al. (2012) Site identification in high-throughput RNA-protein interaction data. *Bioinformatics* **28**, 3013–3020 https://doi.org/10.1093/bioinformatics/bts569
61 Golumbeanu, M., Mohammadi, P. and Beerenwinkel, N. (2016) BMix: probabilistic modeling of occurring substitutions in PAR-CLIP data. *Bioinformatics* **32**, 976–983 https://doi.org/10.1093/bioinformatics/btv520
62 Curk, T., Rot, G., Gorup, Č., de los Mozos, I.R., Konig, J., Zmrzlikar, J. et al. (2019) iCount: protein-RNA interaction iCLIP data analysis (in preparation)
63 Krakau, S., Richard, H. and Marsico, A. (2017) PureCLIP: capturing target-specific protein-RNA interaction footprints from single-nucleotide CLIP-seq data. *Genome Biol.* **18**, 240 https://doi.org/10.1186/s13059-017-1364-2
64 Lang, B., Armaos, A. and Tartaglia, G.G. (2019) RNAct: protein-RNA interaction predictions for model organisms with supporting experimental data. *Nucleic Acids Res.* **47**, D601–D606 https://doi.org/10.1093/nar/gky967
65 Cirillo, D., Blanco, M., Armaos, A., Buness, A., Avner, P., Guttman, M. et al. (2016) Quantitative predictions of protein interactions with long noncoding RNAs. *Nat. Methods* **14**, 5–6 https://doi.org/10.1038/nmeth.4100
66 Cirillo, D., Agostini, F., Klus, P., Marchese, D., Rodriguez, S., Bolognesi, B. et al. (2013) Neurodegenerative diseases: quantitative predictions of protein-RNA interactions. *RNA* **19**, 129–140 https://doi.org/10.1261/rna.034777.112
67 Agostini, F., Cirillo, D., Bolognesi, B. and Tartaglia, G.G. (2013) X-inactivation: quantitative predictions of protein interactions in the Xist network. *Nucleic Acids Res.* **41**, e31 https://doi.org/10.1093/nar/gks968
68 Armaos, A., Cirillo, D. and Gaetano Tartaglia, G. (2017) Omixcore: a web server for prediction of protein interactions with large RNA. *Bioinformatics* **33**, 3104–3106 https://doi.org/10.1093/bioinformatics/btx361
69 Agostini, F., Zanzoni, A., Klus, P., Marchese, D., Cirillo, D. and Tartaglia, G.G. (2013) CatRAPID omics: a web server for large-scale prediction of protein-RNA interactions. *Bioinformatics* **29**, 2928–2930 https://doi.org/10.1093/bioinformatics/btt495
70 Cirillo, D., Marchese, D., Agostini, F., Livi, C.M., Botta-Orfila, T. and Tartaglia, G.G. (2014) Constitutive patterns of gene expression regulated by RNA-binding proteins. *Genome Biol.* **15**, R13 https://doi.org/10.1186/gb-2014-15-1-r13
71 Livi, C.M., Klus, P., Delli Ponti, R. and Tartaglia, G.G. (2016) CatRAPID signature: identification of ribonucleoproteins and RNA-binding regions. *Bioinformatics* **32**, 773–775 https://doi.org/10.1093/bioinformatics/btv629
72 Lu, Q., Ren, S., Lu, M., Zhang, Y., Zhu, D., Zhang, X. et al. (2013) Computational prediction of associations between long non-coding RNAs and proteins. *BMC Genomics* **14**, 1 https://doi.org/10.1186/1471-2164-14-1
73 Muppirala, U.K., Honavar, V.G. and Dobbs, D. (2011) Predicting RNA-protein interactions using only sequence information. *BMC Bioinformatics* **12**, 489 https://doi.org/10.1186/1471-2105-12-489
74 Suresh, V., Liu, L., Adjeroh, D. and Zhou, X. (2015) RPI-Pred: predicting ncRNA-protein interaction using sequence and structural information. *Nucleic Acids Res.* **43**, 1370–1379 https://doi.org/10.1093/nar/gkv020
75 Pan, X., Rijnbeek, P., Yan, J. and Shen, H.B. (2018) Prediction of RNA-protein sequence and structure binding preferences using deep convolutional and recurrent neural networks. *BMC Genomics* **19**, 511 https://doi.org/10.1186/s12864-018-4889-1

PORTLAND PRESS

Biochemical Society Transactions (2020) **48** 1529–1543
https://doi.org/10.1042/BST20191059

76 Boyle, E.I., Weng, S., Gollub, J., Jin, H., Botstein, D., Cherry, J.M. et al. (2004) GO::TermFinder-open source software for accessing Gene Ontology information and finding significantly enriched Gene Ontology terms associated with a list of genes INTRODUCTION: MOTIVATION AND DESIGN. *Bioinformatics* **20**, 3710–3715 https://doi.org/10.1093/bioinformatics/bth456
77 Budak, G., Srivastava, R. and Janga, S.C. (2017) Seten: a tool for systematic identification and comparison of processes, phenotypes, and diseases associated with RNA-binding proteins from condition-specific CLIP-seq profiles. *RNA* **23**, 836–846 https://doi.org/10.1261/rna.059089.116
78 Machanick, P. and Bailey, T.L. (2011) MEME-ChIP: motif analysis of large DNA datasets. *Bioinformatics* **27**, 1696–1697 https://doi.org/10.1093/bioinformatics/btr189
79 Agostini, F., Cirillo, D., Delli Ponti, R. and Tartaglia, G.G. (2014) SeAMotE: a method for high-throughput motif discovery in nucleic acid sequences. *BMC Genomics* **15**, 925 https://doi.org/10.1186/1471-2164-15-925
80 Feng, H., Bao, S., Rahman, M.A., Weyn-Vanhentenryck, S.M., Khan, A., Wong, J. et al. (2019) Modeling RNA-binding protein specificity in vivo by precisely registering protein-RNA crosslink sites. *Mol. Cell* **74**, 1189–1204.e6 https://doi.org/10.1016/j.molcel.2019.02.002
81 Maticzka, D., Lange, S.J., Costa, F. and Backofen, R. (2014) Graphprot: modeling binding preferences of RNA-binding proteins. *Genome Biol.* **15**, R17 https://doi.org/10.1186/gb-2014-15-1-r17
82 Heller, D., Krestel, R., Ohler, U., Vingron, M. and Marsico, A. (2017) SSHMM: extracting intuitive sequence-structure motifs from high-Throughput RNA-binding protein data. *Nucleic Acids Res.* **45**, 11004–11018 https://doi.org/10.1093/nar/gkx756
83 Adinolfi, M., Pietrosanto, M., Parca, L., Ausiello, G., Ferrè, F., Ferrè, F. et al. (2019) Discovering sequence and structure landscapes in RNA interaction motifs. *Nucleic Acids Res.* **47**, 4958–4969 https://doi.org/10.1093/nar/gkz250
84 Licatalosi, D.D., Ye, X. and Jankowsky, E. (2020) Approaches for measuring the dynamics of RNA–protein interactions. *Wiley Interdiscip. Rev. RNA* **11**, 1–23 https://doi.org/10.1002/wrna.1565
85 Marchese, D., de Groot, N.S., Lorenzo Gotor, N., Livi, C.M. and Tartaglia, G.G. (2016) Advances in the characterization of RNA-binding proteins. *Wiley Interdiscip. Rev. RNA* **7**, 793–810 https://doi.org/10.1002/wrna.1378
86 Nechay, M. and Kleiner, R.E. (2020) High-throughput approaches to profile RNA-protein interactions. *Curr. Opin. Chem. Biol.* **54**, 37–44 https://doi.org/10.1016/j.cbpa.2019.11.002
87 Zacco, E., Martin, S.R., Thorogate, R. and Pastore, A. (2018) The RNA-recognition motifs of TAR DNA-binding protein 43 May play a role in the aberrant self-assembly of the protein. *Front. Mol. Neurosci.* **11**, 372 https://doi.org/10.3389/fnmol.2018.00372
88 Rio, D.C. (2012) Filter-binding assay for analysis of RNA-protein interactions. *Cold Spring Harb. Protoc.* **7**, 1078–1081 https://doi.org/10.1101/pdb.prot071449
89 Ryder, S.P., Recht, M.I. and Williamson, J.R. (2008) Quantitative analysis of protein-RNA interactions by gel mobility shift. *Methods Mol. Biol.* **488**, 99–115 https://doi.org/10.1007/978-1-60327-475-3_7
90 Hellman, L.M. and Fried, M.G. (2007) Electrophoretic mobility shift assay (EMSA) for detecting protein-nucleic acid interactions. *Nat. Protoc.* **2**, 1849–1861 https://doi.org/10.1038/nprot.2007.249
91 Lakowicz, J.R. (2006) Plasmonics in biology and plasmon-controlled fluorescence. *Plasmonics* **1**, 5–33 https://doi.org/10.1007/s11468-005-9002-3
92 Littler, D.R., Gully, B.S., Colson, R.N. and Rossjohn, J. (2020) Crystal structure of the SARS-CoV-2 non-structural protein 9, Nsp9. *iScience* **23**, 101258 https://doi.org/10.1016/j.isci.2020.101258
93 Beier, D.H., Carrocci, T.J., Van Der Feltz, C., Tretbar, U.S., Paulson, J.C., Grabowski, N. et al. (2019) Dynamics of the DEAD-box ATPase Prp5 RecA-like domains provide a conformational switch during spliceosome assembly. *Nucleic Acids Res.* **47**, 10842–10851 https://doi.org/10.1093/nar/gkz765
94 Meiser, N., Fuks, C. and Hengesbach, M. (2020) Cooperative analysis of structural dynamics in RNA-protein complexes by single-molecule förster resonance energy transfer spectroscopy. *Molecules* **25**, 2057 https://doi.org/10.3390/molecules25092057
95 Vo, T., Paul, A., Kumar, A., Boykin, D.W. and Wilson, W.D. (2019) Biosensor-surface plasmon resonance: a strategy to help establish a new generation RNA-specific small molecules. *Methods* **167**, 15–27 https://doi.org/10.1016/j.ymeth.2019.05.005
96 Sultana, A. and Lee, J.E. (2015) Measuring protein-protein and protein-nucleic acid interactions by biolayer interferometry. *Curr. Protoc. Protein Sci.* **2015**, 19.25.1–19.25.26 https://doi.org/10.1002/0471140864.ps1925s79
97 Moon, M.H., Hilimire, T.A., Sanders, A.M. and Schneekloth, J.S. (2018) Measuring RNA-ligand interactions with microscale thermophoresis. *Biochemistry* **57**, 4638–4643 https://doi.org/10.1021/acs.biochem.7b01141
98 Cléry, A., Sohier, T.J.M., Welte, T., Langer, A. and Allain, F.H.T. (2017) switchSENSE: a new technology to study protein-RNA interactions. *Methods* **118–119**, 137–145 https://doi.org/10.1016/j.ymeth.2017.03.004
99 Feig, A.L. (2009) Studying RNA-RNA and RNA-protein interactions by isothermal titration calorimetry. *Methods Enzymol.* **468**, 409–422 https://doi.org/10.1016/S0076-6879(09)68019-8
100 Samatanga, B., Cléry, A., Barraud, P., Allain, F.H.-T. and Jelesarov, I. (2017) Comparative analyses of the thermodynamic RNA binding signatures of different types of RNA recognition motifs. *Nucleic Acids Res.* **45**, 6037–6050 https://doi.org/10.1093/nar/gkx136
101 Ruminski, D.J., Watson, P.Y., Mahen, E.M. and Fedor, M.J. (2016) A DEAD-box RNA helicase promotes thermodynamic equilibration of kinetically trapped RNA structures in vivo. *RNA* **22**, 416–427 https://doi.org/10.1261/rna.055178.115
102 Gu, S., Jin, L., Zhang, F., Huang, Y., Grimm, D., Rossi, J.J. et al. (2011) Thermodynamic stability of small hairpin RNAs highly influences the loading process of different mammalian Argonautes. *Proc. Natl Acad. Sci. U.S.A.* **108**, 9208–9213 https://doi.org/10.1073/pnas.1018023108
103 Amano, R., Takada, K., Tanaka, Y., Nakamura, Y., Kawai, G., Kozu, T. et al. (2016) Kinetic and thermodynamic analyses of interaction between a high-Affinity RNA aptamer and its target protein. *Biochemistry* **55**, 6221–6229 https://doi.org/10.1021/acs.biochem.6b00748
104 Ray, D., Ha, K.C.H., Nie, K., Zheng, H., Hughes, T.R. and Morris, Q.D. (2017) RNAcompete methodology and application to determine sequence preferences of unconventional RNA-binding proteins. *Methods* **118–119**, 3–15 https://doi.org/10.1016/j.ymeth.2016.12.003
105 Beckmann, B.M., Castello, A. and Medenbach, J. (2016) The expanding universe of ribonucleoproteins: of novel RNA-binding proteins and unconventional interactions. *Pflugers Arch. Eur. J. Physiol.* **468**, 1029–1040 https://doi.org/10.1007/s00424-016-1819-4
106 Sanchez de Groot, N., Armaos, A., Graña-Montes, R., Alriquet, M., Calloni, G., Vabulas, R.M. et al. (2019) RNA structure drives interaction with proteins. *Nat. Commun.* **10**, 1–13 https://doi.org/10.1038/s41467-019-10923-5
107 Schlundt, A., Tants, J.N. and Sattler, M. (2017) Integrated structural biology to unravel molecular mechanisms of protein-RNA recognition. *Methods* **118–119**, 119–136 https://doi.org/10.1016/j.ymeth.2017.03.015

Biochemical Society Transactions (2020) **48** 1529–1543
https://doi.org/10.1042/BST20191059

108 Hennig, J. and Sattler, M. (2014) The dynamic duo: combining NMR and small angle scattering in structural biology. *Protein Sci.* **23**, 669–682 https://doi.org/10.1002/pro.2467
109 Mahieu, E. and Gabel, F. (2018) Biological small-angle neutron scattering: recent results and development. *Acta Crystallogr. Sect. D Struct. Biol.* **74**, 715–726 https://doi.org/10.1107/S2059798318005016
110 Lapinaite, A., Carlomagno, T. and Gabel, F. (2020) Small-angle neutron scattering of RNA–protein complexes. *Methods Mol. Biol.* **2113**, 165–188 https://doi.org/10.1007/978-1-0716-0278-2_13
111 Delhommel, F., Gabel, F. and Sattler, M. (2020) Current approaches for integrating solution NMR spectroscopy and small-angle scattering to study the structure and dynamics of biomolecular complexes. *J. Mol. Biol.* **432**, 2890–2912 https://doi.org/10.1016/j.jmb.2020.03.014
112 Gräwert, T.W. and Svergun, D.I. (2020) Structural modeling using solution small-angle X-ray scattering (SAXS). *J. Mol. Biol.* **432**, 3078–3092 https://doi.org/10.1016/j.jmb.2020.01.030
113 Patel, T.R., Chojnowski, G., Astha, Koul, A., McKenna, S.A. and Bujnicki, J.M. (2017) Structural studies of RNA-protein complexes: a hybrid approach involving hydrodynamics, scattering, and computational methods. *Methods* **118–119**, 146–162 https://doi.org/10.1016/j.ymeth.2016.12.002
114 Yadav, D.K. and Lukavsky, P.J. (2016) NMR solution structure determination of large RNA-protein complexes. *Prog. Nucl. Magn. Reson. Spectrosc.* **97**, 57–81 https://doi.org/10.1016/j.pnmrs.2016.10.001
115 Ke, A. and Doudna, J.A. (2004) Crystallization of RNA and RNA-protein complexes. *Methods* **34**, 408–414 https://doi.org/10.1016/j.ymeth.2004.03.027
116 Luo, X., Wang, X., Gao, Y., Zhu, J., Liu, S., Gao, G. et al. (2020) Molecular mechanism of RNA recognition by zinc-finger antiviral protein. *Cell Rep.* **30**, 46–52 https://doi.org/10.1016/j.celrep.2019.11.116
117 Partin, A.C., Zhang, K., Jeong, B.C., Herrell, E., Li, S., Chiu, W. et al. (2020) Cryo-EM structures of human drosha and DGCR8 in complex with primary microRNA. *Mol. Cell* **78**, 411–422 https://doi.org/10.1016/j.molcel.2020.02.016
118 Haahr Larsen, A., Wang, Y., Bottaro, S., Grudinin, S., Arleth, L. and Lindorff-Larsen, K. (2020) Combining molecular dynamics simulations with small-angle X-ray and neutron scattering data to study multi-domain proteins in solution. *PLoS Comput. Biol.*, **16**, e1007870. https://doi.org/10.1371/journal.pcbi.1007870
119 Shapiro, B.A. and Le Grice, S.F.J. (2016) Advances in RNA structure determination. *Methods* **103**, 1–3 https://doi.org/10.1016/j.ymeth.2016.06.006
120 Bonomi, M., Heller, G.T., Camilloni, C. and Vendruscolo, M. (2017) Principles of protein structural ensemble determination. *Curr. Opin. Struct. Biol.* **42**, 106–116 https://doi.org/10.1016/j.sbi.2016.12.004

CHAPTER 4

THE INTERPLAY BETWEEN DISORDERED REGIONS IN RNAS AND PROTEINS MODULATES INTERACTIONS WITHIN STRESS GRANULES AND PROCESSING BODIES

The interplay between disordered regions in RNAs and proteins modulates interactions within stress granules and processing bodies

Phase separation is a widespread subject of interest in the scientific literature. In particular, liquid-like condensates such as stress granules (SGs) and processing bodies (PBs) are being studied in detail thanks to the development of new experimental techniques, leading to a better understanding of their transcriptomes and proteomes (Decker and Parker, 2012). One of the main questions remaining unsolved revolved around the interaction networks of their proteic and ribonucleic components as to how they shape and sustain these organelles. In particular, a precise physicochemical characterization of such networks was yet to be established.
For this reason, we decided to analyse the proteome and the transcriptome of these condensates and the interactions they establish, focusing both on SGs and PBs in order to unravel the differences and similarities between the two. To do so, we relied on both wet and dry approaches in order to strengthen the results we found.

We show that poorly structured RNA and protein elements seem to interact within themselves and with each other, creating a sort of circular scenario in which disorder of both proteins and transcripts seems to be the driving elements in creating the interaction network that sustains these condensates.

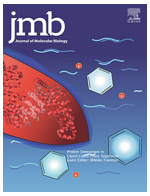
This analysis provides a complete overview of the main players involved in phase-separating condensates and the relevance of their molecular in-

teractions, setting a baseline for the investigation of specific molecules or pathways that could be important for the survival and functioning of these organelles.

This work was published in the Journal of Molecular Biology in 2021.

Vandelli, A, Cid Samper, F., Torrent Burgas, M., Sanchez de Groot, N., and Tartaglia, G. G. (2022). The Interplay Between Disordered Regions in RNAs and Proteins Modulates Interactions Within Stress Granules and Processing Bodies. *Journal of Molecular Biology*, 434(1):167159.
DOI: 10.1016/j.jmb.2021.167159

Research Article



The Interplay Between Disordered Regions in RNAs and Proteins Modulates Interactions Within Stress Granules and Processing Bodies

Andrea Vandelli [1,2,3], Fernando Cid Samper [2,3], Marc Torrent Burgas [1], Natalia Sanchez de Groot [1,3*] and Gian Gaetano Tartaglia [2,3,4,5,6*]

*1 - **Department of Biochemistry and Molecular Biology,** Universitat Autònoma de Barcelona, Bellaterra, 08193 Barcelona, Spain*
*2 - **Universitat Pompeu Fabra (UPF),** 08003 Barcelona, Spain*
*3 - **Centre for Genomic Regulation (CRG),** The Barcelona Institute for Science and Technology, 08003 Barcelona, Spain*
*4 - **Center for Human Technologies,** Istituto Italiano di Tecnologia, 16152 Genova, Italy*
*5 - **Department of Biology** '**Charles Darwin**', Sapienza University of Rome, 00185 Rome, Italy*
*6 - **Institucio Catalana de Recerca i Estudis Avançats (ICREA),** 08010 Barcelona, Spain*

Correspondence to Natalia Sanchez de Groot and Gian Gaetano Tartaglia: Department of Biochemistry and Molecular Biology, Universitat Autònoma de Barcelona, Bellaterra, 08193 Barcelona, Spain (N.S. de Groot). Universitat Pompeu Fabra (UPF), 08003 Barcelona, Spain (G.G. Tartaglia). *natalia.sanchez@uab.cat (N. Sanchez de Groot), gian.tartaglia@iit.it (G.G. Tartaglia)*
https://doi.org/10.1016/j.jmb.2021.167159
Edited by Monika Fuxreiter

Abstract

Condensation, or liquid-like phase separation, is a phenomenon indispensable for the spatiotemporal regulation of molecules within the cell. Recent studies indicate that the composition and molecular organization of phase-separated organelles such as Stress Granules (SGs) and Processing Bodies (PBs) are highly variable and dynamic. A dense contact network involving both RNAs and proteins controls the formation of SGs and PBs and an intricate molecular architecture, at present poorly understood, guarantees that these assemblies sense and adapt to different stresses and environmental changes. Here, we investigated the physico-chemical properties of SGs and PBs components and studied the architecture of their interaction networks. We found that proteins and RNAs establishing the largest amount of contacts in SGs and PBs have distinct properties and intrinsic disorder is enriched in all protein-RNA, protein-protein and RNA-RNA interaction networks. The increase of disorder in proteins is accompanied by an enrichment in single-stranded regions of RNA binding partners. Our results suggest that SGs and PBs quickly assemble and disassemble through dynamic contacts modulated by unfolded domains of their components.

Introduction

Cells exploit the spatiotemporal confinement for efficient organization of biochemical reactions.[1] In the complex and crowded intracellular milieu,[2] condensation in membrane-bound or membrane-less organelles allows to control concentration and interactions of the reactants.[3] These assemblies, located in both cytoplasm and nucleus, participate in multiple cellular functions[4] including stress response, transport channels in the nuclear pore complex and chromatin reorganization.[5]

Molecular condensation is currently the subject of intense investigation and recent advances started to reveal their composition and inner architecture.[3,6,7] Molecular interactions within molecular condensates are not yet understood, but involve proteins and RNAs.[8,9] These assemblies have liquid-like properties and are commonly formed through a process that requires phase separation.[10,11] Valency or

number of interaction sites dictates the contact density of the molecular network regulating the stability, organization and composition of the condensates.[10,12]

Through helical interactions, canonical and non-canonical Watson-Crick base-pairing, RNAs interact with other RNAs and promote phase separation.[13] Yet, technical difficulties in the study of RNA-RNA contacts currently impede our complete understanding of this phenomenon.[14] Protein-protein interactions, and especially prion-like elements, contribute to condensation by promoting protein associations.[10,15] More specifically, perturbation of the native state[16] accompanied by an increase in structural disorder[17] and hydrophobicity[18] enhance the propensity of proteins to aggregate.[15]

Depending on their binding preferences, RNA-binding proteins (RBPs) interact with either single or double-stranded regions of RNAs.[19] Highly structured RNAs attract large amounts of proteins thanks to their intrinsic ability to establish stable interactions.[12,20] RNAs can be often scaffolding elements: whereas a polypeptide of 100 amino acids can interact with one or two proteins, a chain of 100 nucleotides is able to bind to 5–20 proteins.[21] Not only RNA attracts proteins, but also proteins can in turn contribute to change RNA properties: chemical modifications such as N^1-methyladenosine (m1A) and N^6-methyladenosine (m6A) can modify RNA structure[22,23] and influence the formation of ribonucleoprotein condensates.[24,25] Helicases such as the Eukaryotic initiation factor 4A-I can also alter RNA structure by opening up double-stranded regions and altering cellular interactions.[26]

Here, we used a computational approach to investigate the interactions and properties of RNA and protein in the two of the best-known biological condensates: stress granules (SGs)[27] and processing-bodies (PBs).[28] These large assemblies arise upon viral infection or when chemical and physical insults occur to cells. They are thought to form to protect transcripts that would otherwise be aberrantly processed. More specifically, SGs store non-translating mRNAs as indicated by translation initiation factors enriched in the pool of proteins that compose them, whereas PBs facilitate RNA decay because of the abundance in RNA decapping and deadenylation enzymes.[29]

Proteins[7,9] and RNAs[27,30] contained in SGs and PBs are only now starting to be unveiled and their interaction networks are largely unknown. With the present systematic analysis, we aim to characterize how structure influences the interactions sustaining these biological condensates, including both proteins and RNAs and all its possible combinations (RNA-RNA, protein-protein and RNA-protein). Our results show similarities and interconnections between the most contacted players of both molecular types. We report the intriguing result that RNAs enriched in SGs and PBs are disordered and form a large number of contacts with RNAs and proteins. At the same time, proteins enriched in SGs and PBs are disordered and form a large number of contacts with proteins and RNAs. Taken together, our data suggest that structural disorder is a property that distinguishes dynamic fuzzy-like assemblies such as PBs and SGs from solid-like aggregates.[31,32]

Results

RNA structure drives interaction with proteins in SGs and PBs

SGs and PBs are two of the best-known biological condensates. They contain multiple proteins whose concentration changes with stress, cell state and environmental conditions.[9,29] Among them, a small specific set of proteins essential for their formation has been found and they are involved in the recruitment of the other components and in sustaining the condensate. Despite this, there are still uncertainties regarding how the cell regulates their content and assembly.

We recently reported that protein-RNA interactions build up the scaffold of phase-separating organelles[10,33] and their selective recruitment is dictated by RNA physicochemical properties.[12,20] Specifically, we have shown that RNAs engaging in interactions with many protein partners are enriched in double-stranded content (Figure 1(**A**)).[19,20] The origin of this property, observed with a number of different experimental approaches, is that double-stranded regions reduce the flexibility of the polynucleotide chain. Presence of a stable fold favors the formation of stable and well-defined binding sites where the protein can bind. However, our observation does not suggest that protein binding sites and double-stranded regions are the same. If a specific interaction occurs in a small loop at the end of a stem, the overall region is enriched in double-stranded nucleotides, although the exact binding could be in a single-stranded region.

We wondered whether RNA structure drives the interaction with proteins present in SGs and PBs as detected in the whole transcriptome analysis.[19,20] Following up on our previous computational analysis,[19,20] we used protein-RNA interactions available from enhanced CLIP (eCLIP) experiments[34] to rank protein associations with RNAs present in SGs and PBs (**Materials and Methods**). We first selected the transcripts with the largest and lowest amount of protein contacts from the list of RNA reported in SGs[27] and PBs[30] (**Supplementary Table 1**) and then compared their secondary structure content. We used *CROSS*[35] to predict the secondary structure properties of transcripts using the information contained in their sequences and we found that RNAs with more protein contacts in SGs and PBs are significantly more structured (Figure 1(**B**)**; Materials and Methods**).

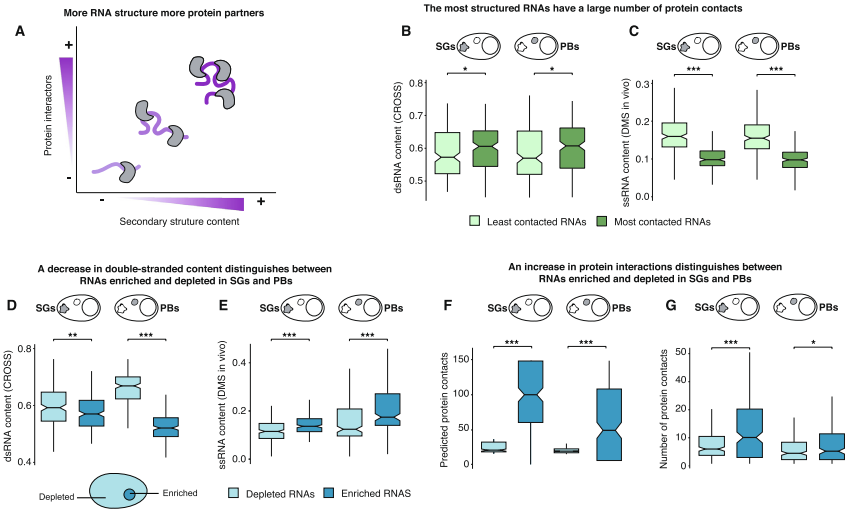


Figure 1. RNAs enriched in SGs and PBs are less structured and contact a larger amount of proteins. A. Graphical representation of the relationship between number of protein interactions and double-stranded content of RNAs. The trend was identified by using different computational and experimental techniques. **B.** Double-stranded content dsRNA (*CROSS* predictions) of RNAs present in SGs and PBs. The RNAs are categorized in two classes: least- and most-contacted depending on the amount of protein interactions detected by eCLIP. An equal amount of 200 transcripts is used in each class (SGs and PBs, least and most contacted RNAs). Significant differentiation is found (SG p-value < 0.013 and PB p-value < 0.029, Wilcoxon test). **C.** Single stranded content ssRNA (dimethyl sulfate modification, DMS, measured *in vivo*) for RNAs most and least contacted by proteins in SGs and PBs. RNA classes follow the definition given in panel **B.** Significant differentiation is found (SG p-value < 1.15e-34, PB p-value < 3.19e-35, Wilcoxon test). **D.** Double stranded content dsRNA (CROSS predictions) for RNAs enriched or depleted in SGs and PBs. An equal amount of 200 transcripts is used for each category (SGs and PBs, depleted and enriched RNAs). Significant differentiation is found (SG p-value < 0.006, PB: p-value < 1.88e-54, Wilcoxon test). **E.** Single stranded content ssRNA (DMS, measured *in vivo*) for RNAs enriched or depleted in SGs and PBs. RNA classes follow the definition given in panel **D.** Significant differentiation is found (SG p-value < 4.51e-67, PB p-value < 2.43e-67, Wilcoxon test). **F.** *cat*RAPID predictions of protein interactions with RNAs enriched or depleted in SGs and PBs. RNA classes follow the definition given in panel **D.** Significant differentiation is found (SG p-value < 3.69e-33, PB p-value < 4.62e-21, Wilcoxon test); **G.** eCLIP detection of protein interactions with RNAs enriched or depleted in SGs and PBs. RNA classes follow the definition of panel **D**. Significant differentiation is found (SG p-value < 1.44e-06, PB p-value < 0.075, Wilcoxon test). Significance indicated in the plots: * p-value < 0.1, ** p-value < 0.01 and *** p-value < 0.001.

CROSS reproduces transcriptomic experiments such as *in vivo* click selective 2-hydroxyl acylation and profiling experiment (icSHAPE)[22] and Parallel Analysis of RNA Structure (PARS)[36] with accuracies higher than 0.80.[37] To assess whether the calculations are in agreement with experimental data, we used data coming from dimethyl sulfate (DMS) foot-printing experiments carried out *in vitro* and *in vivo*[38] (**Materials and Methods**). DMS modification of the unpaired adenosine and cytidine nucleotides is commonly used for revealing structural properties of RNA molecules.[39] The results are in complete accordance with *CROSS* predictions, with the most contacted RNAs being more structured than the least contacted ones (Figure 1(**C**) and **Supplementary Figure 1**). Although the conditions in which DMS experiments were performed did not take into account formation of SGs and PBs, our results show that for both SGs and PBs the amount of double-stranded regions is statistically associated with the number of protein contacts SG's and PB's RNAs can form (Figure 1(**C**)).

Our results indicate that RNAs establishing interactions with a large number of proteins[12,40] act as scaffolds for the formation of ribonucleoprotein complexes,[33,41] which suggests that specific

transcripts could be the 'hubs' in the transcriptional and post-transcriptional layers of regulation.[19,20] This observation indicates that RNAs could be regarded as network connectors or 'kinetic condensers' sustaining and capturing the different components of the biological condensates. Actually, recent evidence indicates that RNA interactions with other RNAs occur spontaneously, thus an additional level exists in the inner regulation of SGs and PBs architecture.[13]

RNA enriched in SGs and PBs are less structured

Since SGs can contain mRNAs from essentially every expressed gene,[27] we decided to study a subset of RNAs that are specifically enriched in SGs or PBs. Indeed, it is possible to distinguish two subsets of transcripts, *enriched* and *depleted*, depending on their abundance in SGs and PBs relative to the rest of the transcriptome (**Materials and Methods** and **Supplementary Table 1**). We stress that the distribution in these groups is independent of the total transcript abundance or the AU content.[42] We also note that the overlap between SGs and PBs is just 25%, and this percentage varies when comparing different sets. Despite these differences, enriched RNAs share similar properties in both cases: they are composed by transcripts with less translation efficiency and longer sequences.[27,28]

Since longer sequences have higher probability to have a larger number of interaction partners,[7,9] we expected to find an enrichment of double-stranded regions in PBs and SGs.[20] However, our predictions carried out with *CROSS* indicate that these RNAs contain more single-stranded regions than depleted transcripts (Figure 1(**D**)).

To assess whether the calculations are in agreement with experimental data, we compared our predictions with DMS experiments.[38] The results are in complete accordance, with enriched RNAs being more unstructured than depleted RNAs (Figure 1(**E**)). Interestingly, the 5′ UTRs, CDS and 3′ UTRs consistently show a lower amount of structure, which indicates that the trend identified is particularly robust (**Supplementary Figure 2**). We obtained similar results using another experimental approach to reveal RNA secondary structure, PARS (**Supplementary Figure 3**).[36] PARS is an approach based on deep sequencing fragments of RNAs treated with structure-specific enzymes[43] (**Materials and Methods**). Again, RNAs enriched in SGs and PBs have a significantly increased number of single-stranded regions.

RNAs enriched in SGs and PBs bind a large amount of proteins

We next investigated protein interactions with RNAs enriched in SGs and PBs. In this context, we previously showed that the interactions between proteins and RNAs could scaffold the formation of phase-separating organelles[12,21,33] and the incorporation of RNAs depends on their physico-chemical properties.[19,20] We used the *cat*RAPID approach to predict RNA interactions with proteins (**Materials and Methods**).[44,45] *cat*RAPID exploits secondary structure predictions coupled with hydrogen bonding and van der Waals calculations to estimate the binding affinity of protein-RNA pairs with an average accuracy of 0.78.[46,47] For both SGs and PBs, our predictions indicate that enriched RNAs have a significantly larger number of interactions with proteins than depleted RNAs (Figure 1(**F**)).

To experimentally validate our predictions, we retrieved protein-RNA interactions available from eCLIP experiments (**Materials and Methods**).[34] On the same set of proteins investigated with *cat*RAPID, we observed that RNAs enriched in PBs and SGs have an increased number of protein partners (Figure 1(**G**)).

Although predictions and experiments used in our analysis do not consider the cellular context in which SGs and PBs are formed, our models are based on physico-chemical properties of the molecules involved and they are therefore expected to have general validity.[19,20] Intriguingly, RNAs enriched in SGs and PBs establish a dense network of contacts with proteins despite their increase in single-stranded content. This contradicts the trend previously identified and suggests that these RNAs might have an interaction network that deviate from those characterizing the average transcriptome.[19,20]

Single stranded regions are involved in RNA-RNA interactions

Even though protein interactions correlate with the amount of double stranded regions found in them (Figure 1(**B**)), it is possible that other RNA properties are involved in different interactions. We hypothesized that RNAs enriched in SGs and PBs may interact among themselves through a mechanism of base-pairing recognition in single-stranded regions.[48] To investigate if an increase in single-stranded regions is a property favouring contacts among RNAs, we compared the structures of RNAs that build a larger number of contacts with RNAs and those more prone to interact with proteins (Figure 2(**A**)). The analysis of the DMS structure shows that the RNAs interacting with a larger number of RNAs are more single-stranded.

Using IntaRNA to predict RNA-RNA interactions (**Materials and Methods**),[48] we then compared the binding ability of the most enriched and depleted RNAs in SGs and PBs. Our results clearly show that enriched RNAs are more prone to interact with RNAs (Figure 2(**B**)).

We then searched available experimental data to validate our predictions. To this aim, we used the RISE database containing RNA-RNA interactions assessed through high-throughput approaches.[49]

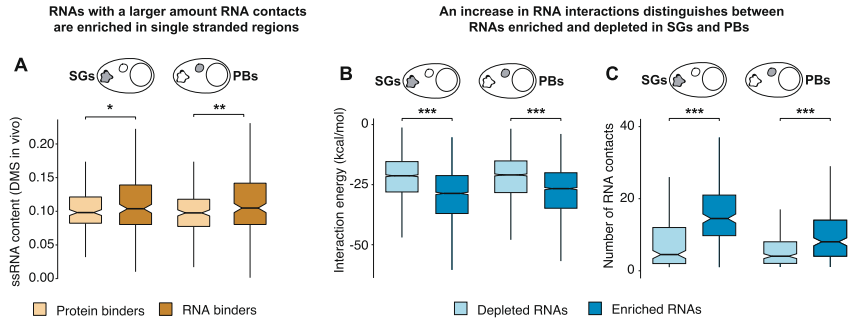


Figure 2. RNAs enriched in SGs and PBs are less structured and contact a larger amount of RNAs. A. Single-stranded content ssRNA (DMS measured *in vivo*) of RNAs enriched in protein interactions (eCLIP experiments) and RNAs enriched in RNA interactions (RISE database). An equal amount of 200 transcripts is used in each category (SGs and PBs, protein and RNA binders). Significant differentiation is found (SG p-value < 0.091, PB p-value < 0.007, Wilcoxon test). **B.** IntaRNA predictions of the energies associated with RNA-RNA interactions. An equal amount of 2500 predicted interactions are used in each category (50 transcripts from SGs and PBs, enriched and depleted interacting with 50 random human RNAs). Significant differentiation is found (SG p-value < 2.54e-108, PB p-value < 1.57e-69, Wilcoxon test). For each plot the interaction bonding energy is reported (kcal/mol). **C.** Number of RNA-RNA interactions (RISE database). An equal amount of 200 transcripts is used in each category (SGs and PBs, enriched and depleted). Significant differentiation is found (SG p-value < 1.54e-19, PB p-value < 5.89e-09, Wilcoxon test). Significance indicated in the plots: * p-value < 0.1, ** p-value < 0.01 and *** p-value < 0.001.

By counting the number of binding partners that each transcript has with other transcripts (**Materials and Methods**), we found that the enriched RNAs are associated with a large number of binding partners (Figure 2(**C**)). Altogether, our results indicate that enriched RNAs are more single-stranded and base-pair with multiple RNAs to establish a larger number of contacts.

Thus, RNAs enriched in SGs and PBs are able to establish a dense network of contacts not just with proteins but also with RNAs. This result suggests that RNAs in SGs and PBs could act as central players sustaining their inner architecture.

Enriched RNAs are populated by master regulators of protein- and RNA-binding

To understand how enriched RNAs are able to create a dense network of contacts, we studied their molecular composition. Starting from the pool of enriched transcripts, we calculated the intersection between the RNAs showing the largest and smallest amounts of protein contacts (eCLIP experiments)[34] against the RNAs showing the largest and smallest amounts of RNA contacts (RISE database).[49] This approach is useful to reveal the existence of a particular subset specialised in binding specific molecular types. The results are shown in **Supplementary Figure 4**, where we report the intersection of the strongest and poorest protein and RNA binders, comparing them with a control. Despite the vast majority of RNAs does not show significant preference for a certain molecular type, we detected an enrichment in the set of RNAs that binds extensively both proteins and RNAs (**Supplementary Table 3**). Thus, these data confirm that the interactivity of the RNAs enriched in SGs and PBs with boths proteins and RNAs is significantly higher than the other transcripts, supporting their importance in sustaining the network of these biological condensates.

SGs and PBs protein pairs are enriched in structural disorder

We next investigated the properties of proteins accumulating in SGs and PBs to better understand how they contribute to their interaction network. First, we analyzed how structure affects the formation of protein pairs in these biological condensates in comparison with the rest of the proteins.

We retrieved from BioGRID[50] all binary protein-protein interactions (PPIs) involving proteins located in SGs and PBs (**Materials and Methods** and **Supplementary Table 2**) and, as a control, an equal amount of PPI with interactors that were not found therein (extracted multiple times, **Supplementary Figure 5**). In this analysis, we measured the amount of disorder available using MobiDB (mean disHL disorder score for each pair)[51] (Figure 3(**A**) and **Supplementary Figures 5 and**

6). In addition, we also measured the amount of disorder of single condensates and non-condensates proteins (**Supplementary Figure 7**). Both analyses indicated that proteins from PBs and SGs are more disordered than the rest of the proteome.

Thus, in addition to RNAs with increased content of single-stranded regions, our results indicate that SGs and PBs contact networks are enriched in proteins with a lower amount of structure. Since it is known that structural disorder promotes allosteric interactions and favours binding with many protein partners,[52] we speculated that SG and PB proteins could have a large number of contacts.

To this aim, we took proteins from SGs and PBs and as a control an equal number of proteins that were not found therein (extracted multiple times, **Supplementary Figure 8**). We then counted how many interactions were reported in BioGRID (Figure 3(**B**) and **Supplementary Figure 8**).[50] Our results indicate that SG and PB proteins have a significantly larger number of partners, suggesting that they have a denser contact network than the rest of the proteome. Therefore, we found an equiv-

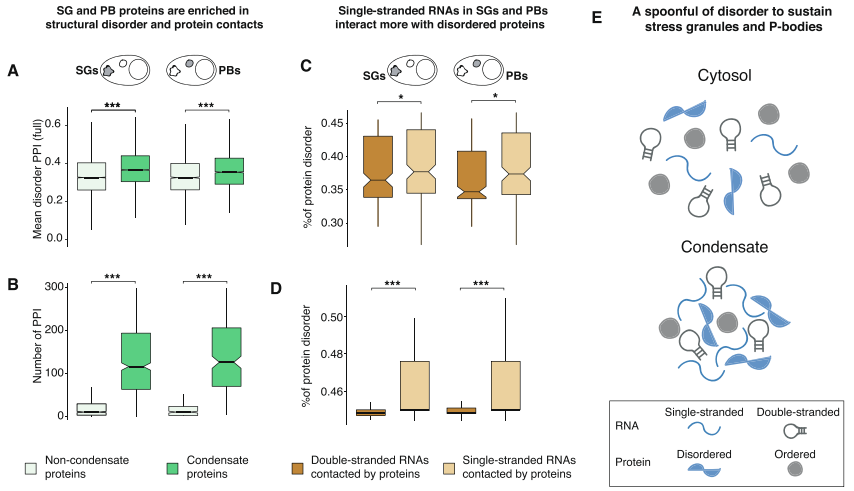


Figure 3. Protein interaction in SGs and PBs is lead by disorder. A. Disorder content of protein–protein interactions associated with SGs and PBs (BioGRID database). For each organelle (SG and PB), an equal number of protein pairs (9336 for SG, 3920 for PB) with the non-condensate control is used. The mean disorder content of each pair was retrieved from the MobiDB database (disHL score). Significant differentiation is found (SG p-value < 1.05e-157, PB p-value < 2.21e-33, Wilcoxon test). **B.** Number of protein–protein interactions associated with SGs and PBs proteins (BioGRID database). For each organelle (SG and PB), an equal number of proteins (586 for SG, 231 for PB) with the non-condensate control is used. Significant differentiation is found (SG p-value < 3.36e-126, PB p-value < 1.51e-58, Wilcoxon test). **C.** *cat*RAPID predictions of protein interactions with RNAs most single stranded and double stranded in SGs and PBs and calculation of mean proteins disorder content. An equal amount of 200 transcripts is used in each category (SGs and PBs, most single-stranded and double-stranded RNAs). The mean disorder content of the interacting proteins for each RNA is retrieved from MobiDB (disHL score). Significant differentiation is found (SG p-value < 0.056, PB p-value < 0.067, Wilcoxon test). **D.** Disorder content of eCLIP proteins interacting with SG and PB most single stranded and double stranded RNAs. An equal amount of 200 transcripts is used in each category (SGs and PBs, most single-stranded and double-stranded RNAs). The mean disorder content of the interacting proteins for each RNA is retrieved from MobiDB (disHL score). Significant differentiation is found (SG p-value < 2.89e-17, PB p-value < 4.97e-19, Wilcoxon test). **E.** Graphical representation of interaction patterns found in our analysis. Condensate enriched RNAs and proteins are responsible for the creation of a contact network that involves both other RNAs and proteins. This leads to the hypothesis that granule proteins and enriched RNAs are crossed "hubs" recruiting and sustaining the different components of SGs and PBs. Significance indicated in the plots: * p-value < 0.1, ** p-value < 0.01 and *** p-value < 0.001.

alence between RNAs and proteins enriched in SGs and PBs, in which both are characterized by a larger number of contacts.

Disorder proteins in SGs and PBs interact more with non-structured RNA

Since RNA binding proteins (RBPs) contain disordered regions,[53] and SG and PB contact networks are enriched in disordered proteins, we investigated which type of structural properties regulate the interactions between RNAs and proteins. Based on the increased amount of single-stranded regions in enriched RNAs and their capacity to form a larger number of interactions with proteins, we expected an increased amount of disorder in RBPs. To test this hypothesis, we analyzed the least and most structured RNAs (data from DMS measured *in vivo*)[38] in SGs and PBs and measured the disorder content (disHL score from MobiDB) of the interacting proteins[51] (**Materials and Methods**). In this analysis we focused on proteins that bind to RNA as predicted by *cat*RAPID and for which the eCLIP interactome is available.[34] The analysis shows that single-stranded RNAs in SGs and PBs are preferentially contacted by disorder proteins (Figure 3 (**C**)). The same result was obtained considering interactions from eCLIP experiments, which confirms the validity of our predictions (Figure 3(**D**)). In the same way, this analysis, carried out only on the RNAs enriched in SGs and PBs, also reproduces this trend (**Supplementary Figure 9**).

From two independent points of view we arrived at the same conclusion about the organization of molecules contained in SGs and PBs (Figure 3 (**E**)). Enriched RNAs, which are more single-stranded, form a larger number of interactions with RNAs but also have a strong potential to interact with proteins. Disordered proteins are enriched in SGs and PBs, have a larger number of PPIs, but also can form more contacts with single-stranded RNAs. So, the two molecular sets that we detected as the most interacting, RNAs and proteins, are both depleted in structure, and form strong interactions between them. This finding indicates that proteins and RNAs in SGs and PBs act together as "hubs" that recruit and sustain the different components of the assemblies (Figure 3(**E**)).

Discussion

We previously observed that RNAs enriched in double-stranded regions attract a large number of proteins.[19,20] The origin of this trend, also identified in SG and PB analyses, is that double-stranded regions favor stable interactions with proteins by reducing the intrinsic flexibility of polynucleotide chains.[19,20] While for each amino acid residue there are two torsional degrees of freedom, RNA conformational space is greater - for each nucleotide residue there are seven independent torsion angles.

Here, we report the novel result that RNAs enriched in SGs and PBs contain single-stranded regions that increase the structural disorder. Since recent reports indicate that single-stranded RNAs have strong ability to act as scaffolds of SGs and PBs,[13,30] we focused our analyses on their interactions with proteins and RNAs.

We first found that RNAs enriched in single-stranded regions are prone to engage in RNA-RNA contacts. This result is not unexpected since single-stranded transcripts are able to base-pair[48,54] and, by doing so, can establish a network of stable interactions. We note that the analysis of RNA-RNA interactions does not take into account the cellular context in which SGs and PBs are formed, thus our results are compatible with a scenario in which specific transcripts are highly prone to interact to quickly promote molecular condensation.[13]

In parallel, the analysis of the SGs and PBs protein interaction networks revealed that proteins enriched in disordered elements form a larger number of contacts with other proteins. This result is very well in line with recent reports indicating that unstructured regions modulate the formation of phase separated assemblies.[31] Indeed, phase separation is a widespread phenomenon in the cell[55] and disordered interactions greatly contribute to the assembly formation.[56] By reporting that single-stranded RNAs preferably contact disorder proteins, we extended the concept of "fuzziness" to RNA molecules. Our work leads to the intriguing result that the two molecular sets identified as the most interacting in the proteome and transcriptome are both depleted in structure and bind one to the other. Thus, specific elements in proteins and RNAs have the ability to recruit and sustain all the components of SGs and PBs (Figure 3(**E**)).

In conclusion, our work suggests that there is not only great diversity in the interaction partners (RNA-RNA, protein-protein, and RNA-protein) but also in their binding modes.[57,58] In this complex scheme, RNA ability to induce phase separation can have an impact on both ordered and disordered proteins: while structural elements can irreversibly sequester globular proteins,[12] disordered regions dynamically engage in interactions that lead to phase separation.[59] Our study shows that the inner architectures of SGs and PBs are intrinsically governed by RNAs and proteins with an increased amount of structurally disordered domains. Thanks to the dynamicity of these regions, protein-RNA complexes are able to assemble and disassemble without the need of strong efforts by the cell. RNA-RNA interactions, at present poorly investigated, are expected to greatly contribute to establishing molecular associations within SGs and PBs.[13] Indeed, RNA molecules are versatile platforms[1,40] capable of interacting with all other molecules,[19] thus

promoting the efficient coordination of transcriptional and post-transcriptional layers of regulation.[20]

Materials and methods

SG and PB transcriptomes

SG transcriptome was collected from Khong *et al.*[27]. The data was generated through RNA-sequencing (RNA-seq) analysis of purified SG cores and single-molecule fluorescence *in situ* hybridization (smFISH) validation. The PB transcriptome was retrieved from Hubstenberger *et al.*[30]; in which a fluorescence-activated particle sorting (FAPS) method was used to purify cytosolic PBs from human epithelial cells. In our statistical analysis, we applied filtering and retained only RNAs with an experimental p-value < 0.01. Within the transcriptome, we distinguished two subsets of transcripts depending on their abundance with respect to the cell transcriptome: enriched (fold-change >=2) and depleted (fold-change<=0.5).

SG and PB proteomes

SG proteome data was retrieved from experiments in various stress conditions and different cell types[7,9,60] for a total of 632 proteins. The first dataset was obtained purifying SG cores from Sodium Arsenite ($NaAsO_2$) stressed U-2 OS cells using a series of differential centrifugations and then affinity purification of GFP-G3BP. The second dataset was obtained using a combination of ascorbate peroxidase (APEX)-mediated *in vivo* proximity labeling with quantitative mass spectrometry (MS) and an RBP-focused immunofluorescence (IF) to identify SG proteins in neuronal and non-neuronal cells and under different types of stress conditions (heat shock, ER stress and oxidative stress). The third dataset employs systematic in vivo proximity-dependent biotinylation (BioID) analysis to identify core components of SGs and PBs. PB proteome data was retrieved combining two studies[30,53] for a total of 259 proteins. In the first study, a fluorescence-activated particle sorting (FAPS) method was developed to purify cytosolic PBs from human epithelial cells, while the second dataset is the one mentioned before, which identified core proteins for both PB and SG using BioID analysis.

RNA secondary structure prediction

We predicted the secondary structure of transcripts using *CROSS* (Computational Recognition of Secondary Structure).[35] The algorithm predicts the structural profile (single- and double-stranded state) at single-nucleotide resolution using sequence information only and without sequence length restrictions (scores > 0 indicate double stranded regions). The obtained scores are then averaged to obtain a secondary structure propensity score for each transcript.

RNA secondary structure measured by DMS

Data on RNA structural content measured by dimethyl sulfate modification (DMS) *in vitro* and *in vivo* conditions were retrieved from Rouskin *et al.*[38]. The number of reads of each transcript was normalized to the highest value (as in the original publication) and averaged.

RNA secondary structure measured by PARS

To profile the secondary structure of human transcripts, we used Parallel Analysis of RNA Structure (PARS) data.[36] To measure PARS structural content for each transcript, we computed the fraction of double-stranded regions over the entire sequence. Given the stepwise function $\vartheta(x) = 1$ for $\times > 0$ and $\vartheta(x) = 0$ otherwise, we computed the fraction of structured domains as:

$$\text{PARSstructuralcontent} = \frac{1}{l}\sum_{i}^{l}\vartheta\left(\frac{V(i)}{S(i)}\right)$$

where $V(i)$ and $S(i)$ are the number of double- and single-stranded reads.

To measure the secondary structure content of the human transcripts 5′- and 3′- UTR and CDS, we retrieved the corresponding locations of the 5′- and 3′-UTR from Ensembl database and repeated the same procedure described above simply considering only the corresponding part of the sequence.

Protein-RNA interaction prediction

Predicted interactions with human proteins were retrieved from RNAct,[47] a database of protein-RNA interactions calculated using *cat*RAPID *omics*,[61] an algorithm to estimate the binding propensity of protein-RNA pairs by combining secondary structure, hydrogen bonding and van der Waals contributions.[44] As reported in the analysis of about half a million of experimentally validated human interactions,[47] the algorithm is able to separate interacting vs non-interacting pairs with an area under the ROC curve of 0.78.[62] The output is filtered according to the Z-score column, which is the interaction propensity normalised by the mean and standard deviation calculated over the reference RBP set. For our analysis, we considered only predicted interactions with a Z-score > 1.

Experimental data on Protein-RNA interactions

RNA interactions for 151 RBPs were retrieved from eCLIP experiments[63] performed in K562 and HepG2 cell lines. In order to measure the fraction of protein binders for each transcript, we applied stringent cut-offs [$-\log_{10}$(*p*-value) > 5 and $-\log_2$(fold_enrichment) > 3] as in previous work.[63] Furthermore, in case of interactions established in

one cell line, only interactions seen in 2 replicates were retained, while in case of two cell lines, only interactions seen in at least 3 out of 4 replicates were retained.

RNA-RNA interactions predictions

RNA-RNA interactions were predicted using the stand-alone IntaRNA software,[48] a program for the fast and accurate prediction of interactions between two RNA molecules. It has been designed to predict mRNA target sites for given non-coding RNAs, such as eukaryotic microRNAs (miRNAs) or bacterial small RNAs (sRNAs), but it can be used to predict other types of RNA-RNA interactions. For each predicted RNA-RNA interaction we retrieved the most optimal one and considered the associated interaction energy.

Experimental data on RNA-RNA interactions

Information about human RNA-RNA interactions were retrieved from RNA Interactome from Sequencing Experiments (RISE) database.[49] RISE is a comprehensive repository of RNA-RNA interactions that mainly come from transcriptome-wide sequencing-based experiments such as PARIS; SPLASH, LIGRseq, and MARIO, and targeted studies like RIAseq, RAP-RNA, and CLASH. Currently it hosts 328,811 RNA-RNA interactions mainly coming from three species (human, mouse, yeast). Human RNA-RNA interactions were filtered, and we retrieved only those in which both partners had an available Ensembl ID.

Experimental data on protein-protein interactions

We used BioGRID (version 4.2.193) for experimental data on protein-protein interactions data.[50] BioGRID is a biomedical interaction repository with data compiled through comprehensive curation efforts, and it contains protein and genetic interactions, chemical interactions and post translational modifications from major model organism species. We used BioGRID to retrieve protein-protein interactions involving condensates proteins against a control. To further strengthen our results, our analyses were done considering both the entire available human BioGRID interactome and physical interactions.

Protein disorder information

Information about human protein disorder predictions were retrieved from MobiDB database (version 4.0)[51]; that contains several data resources and features for protein disorder. Structural and functional properties of disordered regions are based on third party databases and a set of prediction methods, which are assembled to provide a comprehensive view of properties of disordered regions at the residue level. From the whole set of predictive methods, we selected scores obtained with DisEMBL tool with hot loops threshold (DisHL), developed for the prediction of loops with a high degree of mobility, considered important for the definition of protein disorder.[64]

Statistical analysis

To assess the significance of the different trends throughout the analysis, we used the Wilcoxon rank sum test (two-sided). It is a non-parametric test that can be used to compare two independent groups of samples. In order to have analysis with balanced groups, for each comparison performed in our study we used the same number of RNAs/proteins for each category, except when stated otherwise.

CRediT authorship contribution statement

Andrea Vandelli: Data curation, Formal analysis, Writing - original draft, Writing - review & editing. **Fernando Cid Samper:** Conceptualization, Data curation. **Marc Torrent Burgas:** Writing - original draft. **Natalia Sanchez de Groot:** Conceptualization, Funding acquisition, Supervision, Writing - original draft, Writing - review & editing. **Gian Gaetano Tartaglia:** Conceptualization, Funding acquisition, Supervision, Writing - original draft, Writing - review & editing.

Acknowledgements

The authors would like to thank all the members of Tartaglia's lab at the CRG and IIT and especially Alexandros Armaos.

Funding

The research leading to these results has been supported by European Research Council [RIBOMYLOME_309545 and ASTRA_855923], the H2020 projects [IASIS_727658 and INFORE_825080] and the Spanish Ministry of Science and Innovation (RYC2019-026752-I and PID2020-117454RA-I00).

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jmb.2021.167159.

Received 1 May 2021;
Accepted 9 July 2021;
Available online 16 July 2021

Keywords:
structural disorder;
interaction networks;
liquid-liquid phase separation;
fuzzy interactions;
RNA and proteins

References

1. Marchese, D., de Groot, N.S., Lorenzo Gotor, N., Livi, C. M., Tartaglia, G.G., (2016). Advances in the characterization of RNA-binding proteins, Wiley Interdiscip Rev. *RNA*, **7** (6), 793–810. https://doi.org/10.1002/wrna.1378.
2. Tartaglia, G.G., Vendruscolo, M., (2009). Correlation between mRNA expression levels and protein aggregation propensities in subcellular localisations. *Mol. BioSyst.*, **5**, 1873–1876. https://doi.org/10.1039/B913099N.
3. Hyman, A.A., Weber, C.A., Jülicher, F., (2014). Liquid-liquid phase separation in biology. *Annu. Rev. Cell Dev. Biol.*, **30** (1), 39–58. https://doi.org/10.1146/annurev-cellbio-100913-013325.
4. Banani, S.F., Lee, H.O., Hyman, A.A., Rosen, M.K., (2017). Biomolecular condensates: organizers of cellular biochemistry. *Nat. Rev. Mol. Cell Biol.*, **18** (5), 285–298. https://doi.org/10.1038/nrm.2017.7.
5. Gomes, E., Shorter, J., (2019). The molecular language of membraneless organelles. *J. Biol. Chem.*, **294** (18), 7115–7127. https://doi.org/10.1074/jbc.TM118.001192.
6. Brangwynne, C.P., Eckmann, C.R., Courson, D.S., Rybarska, A., Hoege, C., Gharakhani, J., Julicher, F., Hyman, A.A., (2009). Germline P granules are liquid droplets that localize by controlled dissolution/condensation. *Science*, **324** (5935), 1729–1732. https://doi.org/10.1126/science:1172046.
7. Jain, S., Wheeler, J., Walters, R., Agrawal, A., Barsic, A., Parker, R., (2016). ATPase-modulated stress granules contain a diverse proteome and substructure. *Cell*, **164** (3), 487–498. https://doi.org/10.1016/j.cell.2015.12.038.
8. Van Treeck, B., Parker, R., (2018). Emerging roles for intermolecular RNA-RNA interactions in RNP assemblies. *Cell*, **174** (4), 791–802. https://doi.org/10.1016/j.cell.2018.07.023.
9. Markmiller, S., Soltanieh, S., Server, K.L., Mak, R., Jin, W., Fang, M.Y., Luo, E.-C., Krach, F., Yang, D., Sen, A., Fulzele, A., Wozniak, J.M., Gonzalez, D.J., Kankel, M.W., Gao, F.-B., Bennett, E.J., Lécuyer, E., Yeo, G.W., (2018). Context-dependent and disease-specific diversity in protein interactions within stress granules. *Cell*, **172** (3), 590–604. e13. https://doi.org/10.1016/j.cell.2017.12.032.
10. Lorenzo Gotor, N., Armaos, A., Calloni, G., Torrent Burgas, M., Vabulas, R.M., De Groot, N.S., Tartaglia, G.G., (2020). RNA-binding and prion domains: the Yin and Yang of phase separation. *Nucleic Acids Res.*, **48**, 9491–9504. https://doi.org/10.1093/nar/gkaa681.
11. Bolognesi, B., Lorenzo Gotor, N., Dhar, R., Cirillo, D., Baldrighi, M., Tartaglia, G.G., Lehner, B., (2016). A concentration-dependent liquid phase separation can cause toxicity upon increased protein expression. *Cell Rep.*, **16** (1), 222–231. https://doi.org/10.1016/j.celrep.2016.05.076.
12. Cid-Samper, F., Gelabert-Baldrich, M., Lang, B., Lorenzo-Gotor, N., Ponti, R.D., Severijnen, L.-A., Bolognesi, B., Gelpi, E., Hukema, R.K., Botta-Orfila, T., 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 Rep.*, **25** (12), 3422–3434.e7. https://doi.org/10.1016/j.celrep.2018.11.076.
13. Van Treeck, B., Protter, D.S.W., Matheny, T., Khong, A., Link, C.D., Parker, R., (2018). RNA self-assembly contributes to stress granule formation and defining the stress granule transcriptome. *PNAS*, **115** (11), 2734–2739. https://doi.org/10.1073/pnas.1800038115.
14. Tian, S., Curnutte, H.A., Trcek, T., Granules, R.N.A., (2020). A view from the RNA perspective. *Molecules*, **25** https://doi.org/10.3390/molecules25143130.
15. Tartaglia, G.G., Pawar, A.P., Campioni, S., Dobson, C.M., Chiti, F., Vendruscolo, M., (2008). Prediction of aggregation-prone regions in structured proteins. *J. Mol. Biol.*, **380**, 425–436. https://doi.org/S0022-2836(08)00567-6.
16. Chiti, F., Taddei, N., White, P.M., Bucciantini, M., Magherini, F., Stefani, M., Dobson, C.M., (1999). Mutational analysis of acylphosphatase suggests the importance of topology and contact order in protein folding. *Nat. Struct. Mol. Biol.*, **6**, 1005–1009. https://doi.org/10.1038/14890.
17. Linding, R., Schymkowitz, J., Rousseau, F., Diella, F., Serrano, L., (2004). A comparative study of the relationship between protein structure and beta-aggregation in globular and intrinsically disordered proteins. *J. Mol. Biol.*, **342**, 345–353. https://doi.org/15313629.
18. Conchillo-Solé, O., de Groot, N.S., Avilés, F.X., Vendrell, J., Daura, X., Ventura, S., (2007). AGGRESCAN: a server for the prediction and evaluation of "hot spots" of aggregation in polypeptides. *BMC Bioinf.*, **8**, 65. https://doi.org/10.1186/1471-2105-8-65.
19. Armaos, A., Zacco, E., Sanchez de Groot, N., Tartaglia, G. G., (2021). RNA-protein interactions: Central players in coordination of regulatory networks. *BioEssays*, **43** (2), 2000118. https://doi.org/10.1002/bies.v43.210.1002/bies.202000118.
20. Sanchez de Groot, N., Armaos, A., Graña-Montes, R., Alriquet, M., Calloni, G., Vabulas, R.M., Tartaglia, G.G., (2019). RNA structure drives interaction with proteins. *Nat. Commun.*, **10**, 3246. https://doi.org/10.1038/s41467-019-10923-5.
21. Ribeiro, D.M., Zanzoni, A., Cipriano, A., Delli Ponti, R., Spinelli, L., Ballarino, M., Bozzoni, I., Tartaglia, G.G., Brun, C., (2018). Protein complex scaffolding predicted as a prevalent function of long non-coding RNAs. *Nucleic Acids Res.*, **46**, 917–928. https://doi.org/10.1093/nar/gkx1169.
22. Spitale, R.C., Flynn, R.A., Zhang, Q.C., Crisalli, P., Lee, B., Jung, J.-W., Kuchelmeister, H.Y., Batista, P.J., Torre, E.A., Kool, E.T., Chang, H.Y., (2015). Structural imprints in vivo decode RNA regulatory mechanisms. *Nature*, **519** (7544), 486–490. https://doi.org/10.1038/nature14263.

23. Ponti, R.D., Armaos, A., Vandelli, A., Tartaglia, G.G., (2020). CROSSalive: a web server for predicting the in vivo structure of RNA molecules. *Bioinformatics*, **36**, 940–941. https://doi.org/10.1093/bioinformatics/btz666.
24. Alriquet, M., Calloni, G., Martínez-Limón, A., Delli Ponti, R., Hanspach, G., Hengesbach, M., Tartaglia, G.G., Vabulas, R.M., (2020). The protective role of m1A during stress-induced granulation. *J. Mol. Cell Biol.*, **12**, 870–880. https://doi.org/10.1093/jmcb/mjaa023.
25. Ries, R.J., Zaccara, S., Klein, P., Olarerin-George, A., Namkoong, S., Pickering, B.F., Patil, D.P., Kwak, H., Lee, J.H., Jaffrey, S.R., (2019). m6A enhances the phase separation potential of mRNA. *Nature*, **571** (7765), 424–428. https://doi.org/10.1038/s41586-019-1374-1.
26. Tauber, D., Tauber, G., Parker, R., (2020). Mechanisms and regulation of RNA condensation in RNP granule formation. *Trends Biochem. Sci.*, **45** (9), 764–778. https://doi.org/10.1016/j.tibs.2020.05.002.
27. Khong, A., Matheny, T., Jain, S., Mitchell, S.F., Wheeler, J. R., Parker, R., (2017). The stress granule transcriptome reveals principles of mRNA accumulation in stress granules. *Mol. Cell.*, **68** (4), 808–820.e5. https://doi.org/10.1016/j.molcel.2017.10.015.
28. Courel, M., Clément, Y., Bossevain, C., Foretek, D., Vidal Cruchez, O., Yi, Z., Bénard, M., Benassy, M.-N., Kress, M., Vindry, C., Ernoult-Lange, M., Antoniewski, C., Morillon, A., Brest, P., Hubstenberger, A., Roest Crollius, H., Standart, N., Weil, D., (2019). GC content shapes mRNA storage and decay in human cells. *Elife*, **8** https://doi.org/10.7554/eLife.49708.
29. Decker, C.J., Parker, R., (2012). P-bodies and stress granules: possible roles in the control of translation and mRNA degradation a012286 *Cold Spring Harb. Perspect. Biol.*, **4** (9) https://doi.org/10.1101/cshperspect.a012286.
30. Hubstenberger, A., Courel, M., Bénard, M., Souquere, S., Ernoult-Lange, M., Chouaib, R., Yi, Z., Morlot, J.-B., Munier, A., Fradet, M., Daunesse, M., Bertrand, E., Pierron, G., Mozziconacci, J., Kress, M., Weil, D., (2017). P-body purification reveals the condensation of repressed mRNA regulons. *Mol. Cell*, **68** (1), 144–157.e5. https://doi.org/10.1016/j.molcel.2017.09.003.
31. Fuxreiter, M., Vendruscolo, M., (2021). Generic nature of the condensed states of proteins. *Nat Cell Biol.*, **23** (6), 587–594. https://doi.org/10.1038/s41556-021-00697-8.
32. M. Monti, A. Armaos, M. Fantini, A. Pastore, G.G. Tartaglia, Aggregation is a context-dependent constraint on protein Evolution, Frontiers in Molecular Biosciences. 8 (2021) in press. https://doi.org/10.3389/fmolb.2021.678115
33. Cerase, A., Armaos, A., Neumayer, C., Avner, P., Guttman, M., Tartaglia, G.G., (2019). Phase separation drives X-chromosome inactivation: a hypothesis. *Nat. Struct. Mol. Biol.*, **26** (5), 331–334. https://doi.org/10.1038/s41594-019-0223-0.
34. Van Nostrand, E.L., Freese, P., Pratt, G.A., Wang, X., Wei, X., Xiao, R., Blue, S.M., Chen, J.-Y., Cody, N.A.L., Dominguez, D., Olson, S., Sundararaman, B., Zhan, L., Bazile, C., Bouvrette, L.P.B., Bergalet, J., Duff, M.O., Garcia, K.E., Gelboin-Burkhart, C., Hochman, M., Lambert, N.J., Li, H., McGurk, M.P., Nguyen, T.B., Palden, T., Rabano, I., Sathe, S., Stanton, R., Su, A., Wang, R., Yee, B.A., Zhou, B., Louie, A.L., Aigner, S., Fu, X.-D., Lécuyer, E., Burge, C.B., Graveley, B.R., Yeo, G.W., (2020). A large-scale binding and functional map of human RNA-binding proteins. *Nature*, **583** (7818), 711–719. https://doi.org/10.1038/s41586-020-2077-3.
35. Delli Ponti, R., Marti, S., Armaos, A., Tartaglia, G.G., (2017). A high-throughput approach to profile RNA structure e35–e35 *Nucleic Acids Res.*, **45** https://doi.org/10.1093/nar/gkw1094.
36. Wan, Y., Qu, K., Zhang, Q.C., Flynn, R.A., Manor, O., Ouyang, Z., Zhang, J., Spitale, R.C., Snyder, M.P., Segal, E., Chang, H.Y., (2014). Landscape and variation of RNA secondary structure across the human transcriptome. *Nature*, **505** (7485), 706–709. https://doi.org/10.1038/nature12946.
37. R. Delli Ponti, A. Armaos, S. Marti, Gian Gaetano Tartaglia, A method for RNA structure prediction shows evidence for structure in lncRNAs, J. Mol. Cell Biol. 5 (2018) 111. https://doi.org/10.3389/fmolb.2018.00111
38. Rouskin, S., Zubradt, M., Washietl, S., Kellis, M., Weissman, J.S., (2014). Genome-wide probing of RNA structure reveals active unfolding of mRNA structures in vivo. *Nature*, **505** (7485), 701–705. https://doi.org/10.1038/nature12894.
39. Tijerina, P., Mohr, S., Russell, R., (2007). DMS footprinting of structured RNAs and RNA-protein complexes. *Nat. Protoc.*, **2** (10), 2608–2623. https://doi.org/10.1038/nprot.2007.380.
40. D. Marchese, T. Botta-Orfila, D. Cirillo, J.A. Rodriguez, C. M. Livi, R. Fernández-Santiago, M. Ezquerra, M.J. Martí, E. Bechara, G.G. Tartaglia, A. Ávila, À. Bayés, T. Botta-Orfila, N. Caballol, M. Calopa, J. Campdelacreu, Y. Compta, M. Ezquerra, O. de Fàbregues, R. Fernández-Santiago, D. Girado, J. Hernández-Vara, S. Jaumà, D. Marchese, M.J. Martí, J. Pagonabarraga, P. Pastor, L. Planellas, C. Pont-Sunyer, V. Puente, M. Pujol, J. Saura, G.G. Tartaglia, E. Tolosa, F. Valldeoriola, Discovering the 3′ UTR-mediated regulation of alpha-synuclein, Nucleic Acids Res. 45 (2017) 12888–12903. https://doi.org/10.1093/nar/gkx1048.
41. Cerase, A., Tartaglia, G.G., (2020). Long non-coding RNA-polycomb intimate rendezvous. *Open Biol.*, **10** (9), 200126. https://doi.org/10.1098/rsob.200126.
42. Matheny, T., Van Treeck, B., Huynh, T.N., Parker, R., (2021). RNA partitioning into stress granules is based on the summation of multiple interactions. *RNA*, **27** (2), 174–189. https://doi.org/10.1261/rna.078204.120.
43. Kertesz, M., Wan, Y., Mazor, E., Rinn, J.L., Nutter, R.C., Chang, H.Y., Segal, E., (2010). Genome-wide measurement of RNA secondary structure in yeast. *Nature*, **467** (7311), 103–107. https://doi.org/10.1038/nature09322.
44. Bellucci, M., Agostini, F., Masin, M., Tartaglia, G.G., (2011). Predicting protein associations with long noncoding RNAs. *Nat. Meth.*, **8** (6), 444–445. https://doi.org/10.1038/nmeth.1611.
45. Cirillo, D., Blanco, M., Armaos, A., Buness, A., Avner, P., Guttman, M., Cerase, A., Tartaglia, G.G., (2017). Quantitative predictions of protein interactions with long noncoding RNAs. *Nat. Meth.*, **14** (1), 5–6. https://doi.org/10.1038/nmeth.4100.
46. Colantoni, A., Rupert, J., Vandelli, A., Tartaglia, G.G., Zacco, E., (2020). Zooming in on protein-RNA interactions: a multi-level workflow to identify interaction partners. *Biochem. Soc. Trans.*, **48**, 1529–1543. https://doi.org/10.1042/BST20191059.

47. Lang, B., Armaos, A., Tartaglia, G.G., (2019). RNAct: Protein-RNA interaction predictions for model organisms with supporting experimental data. *Nucleic Acids Res.*, **47**, D601–D606. https://doi.org/10.1093/nar/gky967.
48. Mann, M., Wright, P.R., Backofen, R., (2017). IntaRNA 2.0: enhanced and customizable prediction of RNA-RNA interactions. *Nucleic Acids Res.*, **45**, W435–W439. https://doi.org/10.1093/nar/gkx279.
49. Gong, J., Shao, D., Xu, K., Lu, Z., Lu, Z.J., Yang, Y.T., Zhang, Q.C., (2018). RISE: a database of RNA interactome from sequencing experiments. *Nucleic Acids Res.*, **46**, D194–D201. https://doi.org/10.1093/nar/gkx864.
50. Oughtred, R., Stark, C., Breitkreutz, B.-J., Rust, J., Boucher, L., Chang, C., Kolas, N., O'Donnell, L., Leung, G., McAdam, R., Zhang, F., Dolma, S., Willems, A., Coulombe-Huntington, J., Chatr-Aryamontri, A., Dolinski, K., Tyers, M., (2019). The BioGRID interaction database: 2019 update. *Nucleic Acids Res.*, **47** https://doi.org/10.1093/nar/gky1079. D529–D541.
51. Piovesan, D., Necci, M., Escobedo, N., Monzon, A.M., Hatos, A., Mičetić, I., Quaglia, F., Paladin, L., Ramasamy, P., Dosztányi, Z., Vranken, W.F., Davey, N.E., Parisi, G., Fuxreiter, M., Tosatto, S.C.E., (2021). MobiDB: intrinsically disordered proteins in 2021. *Nucleic Acids Res.*, **49**, D361–D367. https://doi.org/10.1093/nar/gkaa1058.
52. Tompa, Peter, (2014). Multisteric regulation by structural disorder in modular signaling proteins: an extension of the concept of allostery. *Chem. Rev.*, **114** (13), 6715–6732. https://doi.org/10.1021/cr4005082.
53. A. Balcerak, A. Trebinska-Stryjewska, R. Konopinski, M. Wakula, E.A. Grzybowska, RNA–protein interactions: disorder, moonlighting and junk contribute to eukaryotic complexity, Open Biology. 9 (n.d.) 190096. https://doi.org/10.1098/rsob.190096
54. Ma, W., Zheng, G., Xie, W., Mayr, C., (2021). In vivo reconstitution finds multivalent RNA–RNA interactions as drivers of mesh-like condensates. *ELife*, **10**, https://doi.org/10.7554/eLife.64252 e64252.
55. Hardenberg, Maarten, Horvath, Attila, Ambrus, Viktor, Fuxreiter, Monika, Vendruscolo, Michele, (2020). Widespread occurrence of the droplet state of proteins in the human proteome. *Proc. Natl. Acad. Sci. U. S. A.*, **117** (52), 33254–33262. https://doi.org/10.1073/pnas.2007670117.
56. Miskei, Marton, Horvath, Attila, Vendruscolo, Michele, Fuxreiter, Monika, (2020). Sequence-based prediction of fuzzy protein interactions. *J. Mol. Biol.*, **432** (7), 2289–2303. https://doi.org/10.1016/j.jmb.2020.02.017.
57. Alberti, Simon, Gladfelter, Amy, Mittag, Tanja, (2019). Considerations and challenges in studying liquid-liquid phase separation and biomolecular condensates. *Cell*, **176** (3), 419–434. https://doi.org/10.1016/j.cell.2018.12.035.
58. Garcia-Jove Navarro, M., Kashida, S., Chouaib, R., Souquere, S., Pierron, G., Weil, D., Gueroui, Z., (2019). RNA is a critical element for the sizing and the composition of phase-separated RNA-protein condensates. *Nat. Commun.*, **10**, 3230. https://doi.org/10.1038/s41467-019-11241-6.
59. Corley, Meredith, Burns, Margaret C., Yeo, Gene W., (2020). How RNA-binding proteins interact with RNA: Molecules and mechanisms. *Mol. Cell.*, **78** (1), 9–29. https://doi.org/10.1016/j.molcel.2020.03.011.
60. Youn, J.-Y., Dunham, W.H., Hong, S.J., Knight, J.D.R., Bashkurov, M., Chen, G.I., Bagci, H., Rathod, B., MacLeod, G., Eng, S.W.M., Angers, S., Morris, Q., Fabian, M., Côté, J.-F., Gingras, A.-C., (2018). High-density proximity mapping reveals the subcellular organization of mRNA-associated granules and bodies 517-532.e11 *Mol. Cell.*, **69** https://doi.org/10.1016/j.molcel.2017.12.020.
61. Agostini, F., Zanzoni, A., Klus, P., Marchese, D., Cirillo, D., Tartaglia, G.G., (2013). catRAPID omics: a web server for large-scale prediction of protein-RNA interactions. *Bioinformatics*, **29** (22), 2928–2930. https://doi.org/10.1093/bioinformatics/btt495.
62. A. Armaos, A. Colantoni, G. Proietti, J. Rupert, G.G. Tartaglia, catRAPID omics v2.0: going deeper and wider in the prediction of protein-RNA interactions, Nucleic Acids Res. (2021) gkab393. https://doi.org/10.1093/nar/gkab393.
63. Van Nostrand, E.L., Pratt, G.A., Shishkin, A.A., Gelboin-Burkhart, C., Fang, M.Y., Sundararaman, B., Blue, S.M., Nguyen, T.B., Surka, C., Elkins, K., Stanton, R., Rigo, F., Guttman, M., Yeo, G.W., (2016). Robust transcriptome-wide discovery of RNA-binding protein binding sites with enhanced CLIP (eCLIP). *Nat. Meth.*, **13** (6), 508–514. https://doi.org/10.1038/nmeth.3810.
64. Linding, R., Jensen, L.J., Diella, F., Bork, P., Gibson, T.J., Russell, R.B., (2003). Protein disorder prediction. *Structure*, **11** (11), 1453–1459. https://doi.org/10.1016/j.str.2003.10.002.

CHAPTER 5

THE PRALINE DATABASE: PROTEIN AND RNA HUMAN SINGLE NUCLEOTIDE VARIANTS IN CONDENSATES

The PRALINE database: Protein and Rna humAn singLe nucleotIde variaNts in condEnsates

Phase separation is a phenomenon that in physiological conditions is reversible and can help the control of biological reactions, with the resulting liquid-like condensates exerting important functions that include the protection of important molecules from harmful conditions to the cell (Decker and Parker, 2012; Protter and Parker, 2016). On the other hand, liquid-to-solid phase transition usually generates irreversible aggregates that are responsible for pathological states. However, despite liquid-liquid phase separation being usually a reversible state, changes in the composition or concentration of condensates' components can also lead to protein misfolding or aberrant accumulation of molecules, resulting in toxic aggregates and subsequent neurodegenerative diseases (Bolognesi et al., 2016; Cid-Samper et al., 2018; Campos-Melo et al., 2021).

In this context, we created the PRALINE database (Protein and Rna human singLe nucleotIde variatioNs in condEnsates), which contains information about proteins and RNAs components enriched in liquid-like (e.g. SGs and PBs) or solid-like (e.g. amyloids) condensates and their interactions, providing data collected from high-throughput experimental and computational approaches. Compared to similar databases, PRALINE is the first to combine experimental and predicted physicochemical properties of the condensates and their inner macromolecular interactions with disease-related single-nucleotide variants (SNVs), describing how these variations can affect these condensates' equilibrium and change the structure of their components.

Considering the existing link between mutations in condensates' components and the insurgence of pathological states, this database will help the design of experiments to study condensates' formation and implication in human diseases.

This work is currently under submission and available as a BioRxiv preprint.

Vandelli, A, Arnal Segura, M., Monti, M., Fiorentino, J., Broglia, L., Colantoni, A., Sanchez de Groot, N., Torrent Burgas, M., Armaos, A., and Tartaglia, G. G. (2022). The PRALINE database: Protein and Rna humAn singLe nucleotIde variaNts in condEnsates. *BioRxiv Preprint*,
DOI: 10.1101/2022.12.03.518982

The PRALINE database: Protein and Rna humAn singLe nucleotIde variaNts in condEnsates

Andrea Vandelli[1,2], Magdalena Arnal Segura[3,4], Michele Monti[3], Jonathan Fiorentino[3], Laura Broglia[3], Alessio Colantoni[4], Natalia Sanchez de Groot[1], Marc Torrent Burgas[1], Alexandros Armaos[3,*] and Gian Gaetano Tartaglia[3,4,*]

[1] Department of Biochemistry and Molecular Biology, Universitat Autònoma de Barcelona, Barcelona, 08193, Spain
[2] Universitat Pompeu Fabra (UPF), 08003 Barcelona, Spain
[3] Center for Human Technologies (CHT) and Center for Life Nano Science (CNLS), Istituto Italiano di Tecnologia (IIT), Via Enrico Melen, 83, 16152 Genova GE.
[4] Department of Biology and Biotechnologies, University Sapienza Rome, Via Aldo Moro 5, 00185, Roma, Italy
* Corresponding authors: AA and GGT. Email: alexandros.armaos@iit.it (AA); gian.tartaglia@iit.it (GGT); Tel. +34 933160116.

ABSTRACT

Summary: Biological condensates are membraneless organelles with different material properties. Proteins and RNAs are the main components, but most of their interactions are still unknown. Here we introduce PRALINE, a database for the interrogation of proteins and RNAs contained in stress-granules, processing bodies, and other assemblies including droplets and amyloids. PRALINE provides information about the predicted and experimentally validated protein-protein, protein-RNA and RNA-RNA interactions. For proteins, it reports the liquid-liquid phase separation and liquid-solid phase separation propensities. For RNAs, it provides information on predicted secondary structure content. PRALINE shows detailed information on human single-nucleotide variants, their clinical significance and presence in protein and RNA binding sites, and how they can affect condensates' physical properties.

Availability: PRALINE is freely accessible on the web at http://alvinlee.bio.uniroma1.it/praline.

Supplementary information: General information is at http://alvinlee.bio.uniroma1.it/praline/about, where we provide a detailed description of the datasets and the tools employed in the database. Data provided in PRALINE are available at http://alvinlee.bio.uniroma1.it/praline/downloads. The tutorial is at http://alvinlee.bio.uniroma1.it/praline/tutorial.

INTRODUCTION

Although the exact composition and functions of the different condensates are unknown, they are enriched in protein and RNA molecules that interact through protein-protein, protein-RNA and RNA-RNA networks. Solid-like condensates, and in particular amyloids, are generally considered to be inherently irreversible aberrant clumps (Dobson, 2017), while liquid-like condensates are dynamic entities that exchange components with the surrounding environment and grow, collapse and fuse in the nucleus and cytoplasm (Marchese et al., 2016). Liquid-like condensates perform different functions on RNA molecules such as storage in the germline, localization in neurons and protection from harmful conditions. The most known liquid-like condensates are processing bodies (PBs) and stress granules (SGs), both enriched in RNA that allows them to form and dissolve rapidly (Lorenzo Gotor et al., 2020). Yet, subtle changes in the composition or concentration of condensates' constituents can induce the formation of solid-like assemblies (Cid-Samper et al., 2018). This is the case of Amyotrophic Lateral Sclerosis (ALS), where single-nucleotide variants (SNVs) in FUS trigger a liquid-to-solid phase transition (Patel et al., 2015). Structural properties of the RNA and changes upon mutations are important, since they play a role in the process of condensation. Highly structured RNAs attract large amounts of proteins thanks to their intrinsic ability to establish stable interactions (Sanchez de Groot et al., 2019). Moreover, RNAs can act as scaffolding elements (Armaos et al., 2021): whereas a polypeptide of 100 amino acids can interact with one or two proteins, a chain of 100 nucleotides is able to bind to 5–20 proteins (Vandelli et al., 2022). Poorly structured transcripts also induce condensation, as they base-pair with other RNAs establishing a dense network of contacts (Treeck et al., 2018). All this information is gathered in PRALINE, a database that provides information on different condensates' components, their interaction networks, and disease-related variants.

METHODS

RNA set. All RNAs sequences were retrieved from Ensembl version 105 (Cunningham et al., 2022). 1841 Stress Granule (SG) enriched RNAs were taken from Khong et al. (Khong et al., 2017) selecting the ones with log2(fold-change) >=1 and p-value <=0.01. 4852 Processing Body (PB) enriched RNAs are taken from Hubstenberger et al. (Hubstenberger et al., 2017), retrieving the ones with log2(fold-change) >=1 and q-value <= 0.01) for a total of 5614 unique genes upon removal of obsolete gene ids in the ensembl version 105. For each gene, the longest isoform was taken. The CROSS algorithm for secondary structure prediction of RNAs was launched on the 5614 RNA sequences with Globalscore parameters (Delli Ponti et al., 2017).

Protein set. 997 UniprotKB IDs were retrieved from different sources. Stress Granule (SG) proteins were retrieved from Markmiller et al. (Markmiller et al., 2018), Youn et al. (Youn et al., 2018) and Jain et al. (Jain et al., 2016) studies. From Jain work we retrieved 411 proteins that were either already known SG proteins or newly discovered through mass spec and IF. From Markmiller and colleagues, we retrieved 397 proteins either already known from previous experiments or found with APEX technique in hek293, NPC and IPSC cells, as well as proteins found to be stress / cell specific or independent. 60 SG proteins were retrieved from Youn's work. We retrieved proteins with Non-negative matrix factorization (NMF) values = 9 or in case of a different NMF value, we collected proteins found to co-localise with G3BP1.

Processing Body (PB) proteins were retrieved from Hustemberger et al. Hubstenberger and Youn et al. (Youn et al., 2018). From the Hustemberger's study, we collected 125 proteins found to be significantly enriched in PBs with p-value < 0.025 as reported in the paper. From Gingras and colleagues' work, we retrieved 42 proteins either with NMF value =8 or that co-localize with DCP1A.

From Vendruscolo and Fuxreiter work we retrieved 280 droplet forming proteins and 68 amyloid forming ones (Vendruscolo and Fuxreiter, 2022).

SNVs. DisGeNet (release 7.0) curated variant-disease associations (Piñero et al., 2020) and ClinVar variants (Landrum et al., 2014; Mj et al., 2018) with a review status higher than one ("practice guideline", "reviewed by expert panel" and "criteria provided, multiple submitters, no conflicts") were downloaded in May 2022. From these datasets, we retrieved disease-related single nucleotide changes (SNVs). 13857 SNVs from DisGeNet and 48671 SNVs from ClinVar fell in the coding region of RNAs enriched in SG and PB and their relative position in the transcripts was retrieved from the Ensembl Variation 105 (Cunningham et al., 2022) in Human Short Variants dataset, excluding insertions and deletions. The CROSS algorithm with Globalscore parameters was launched on RNA fragments of 51 nt with the SNV in the center to calculate the difference in secondary structure content between the reference and alternative RNA sequences (Delli Ponti et al., 2017). To avoid smaller-size fragments, SNVs falling at the beginning and at the end of the RNAs were removed. The reference and alternative CROSS secondary structure propensity profiles of the 51 nt RNA fragments were represented against each other in plots. The mean difference in secondary structure between reference and alternative sequence was computed on a window size of 11 nt upstream and downstream the SNV. Information on the expression quantitative trait loci (eQTL) and splicing quantitative trait loci (sQTL) of the SNVs was obtained from GTEx V8 (GTEx Consortium, 2020).

RNA-RNA interactions. In May 2020, we retrieved human RNA-RNA interactions from the RISE database (version 1.0) focusing on the experimental interactions obtained with PARIS technique (Gong et al., 2018). For each RNA, binding site location were mapped to the longest transcript in Ensembl 105, using blastn algorithm (Cunningham et al., 2022). We retrieved a total of 25.232 RNA-RNA interactions with at least one interactor being an enriched RNA in SG or PB. In 934 of those interactions we found at least a SNV located inside a binding site.

Protein-RNA interactions. We provide experimentally determined protein-RNA interactions validated through eCLIP experiments available in May 2022 from https://www.encodeproject.org/eclip/ (Van Nostrand et al., 2020) as well as catRAPID predictions available in RNAct (Lang et al., 2019). For the experimentally validated interactions, binding sites are displayed.

Protein-Protein interactions. As binding sites related to protein-protein interactions are not physically available, we provide links to an external database. Human curated protein-protein interactions are linked to BioGRID database version 4.4 (Oughtred et al., 2021).

LLPS and LSPT Propensities. We computed the propensity to undergo liquid-like and solid-like condensation for the set of 997 proteins detailed previously and for their 6152 natural variants (involving 632 of them), retrieved from UniProtKB (release 2022_01) (The UniProt Consortium, 2021). We considered 5949 single point mutations and 203 deletions. To quantify the extent to which each sequence is prone to undergo LSPT and LLPS we used the Zyggregator (Tartaglia et al., 2008) and catGRANULE algorithms (Bolognesi et al., 2016), which compute the liquid-solid and liquid-liquid propensities, respectively.

We note that LLPS and LSPT are promoted by both intrinsic and extrinsic contributions. By intrinsic contributions we mean physico-chemical properties of the polypeptide chain such as the hydrophobicity for Zyggregator and the structural disorder for catGRANULE. The extrinsic contributions relate to environmental factors, such as concentration, pH, ionic strength, but not only, for instance crowding agents are also to be included. Our methods predict intrinsic properties of the polypeptide chain and do not take into account the different extrinsic contributions at present.

For the proteins that do not have natural variants present in UniProtKB we only report the WT scores of Zyggregator and catGRANULE.

INTERPRETATION AND USE OF THE DATABASE

PRALINE can be accessed using protein or RNA names provided as: Gene Name, Ensembl Gene / Transcript ID (https://www.ensembl.org/) and UniprotKB ID (https://www.uniprot.org/; **Figure 1A**).

- Searching for a specific protein, the user can retrieve information on the condensate state (Droplet/liquid-like or Amyloid/solid-like) and the organelle in which it has been found (SG/PB). The predicted Liquid-liquid phase separation (LLPS) and liquid-solid phase transition (LSPT) propensities and profiles of the wild-type sequence are provided, calculated with *cat*GRANULE (Bolognesi et al., 2016) and Zyggregator (Tartaglia et al., 2008) methods, respectively (>0.80 accuracy in predicting regions of the proteins involved in protein condensation; **Figure 1B**). Experimentally validated protein-protein interactions are available through links to BioGRID (https://thebiogrid.org/; **Methods**), while experimental and predicted protein-RNA interactions can be retrieved from RNAct (https://rnact.crg.eu/; **Methods**). Protein-RNA interactions are calculated using *cat*RAPID, an algorithm trained on NMR and X-ray structures (AUC of 0.77 on eCLIP interactions) (Lang et al., 2019). The number of SNVs is shown for the protein of interest and, for each SNV, it is possible to interrogate the amino acid position, the difference in LSPT and LLPS propensities compared to the reference (i.e., wild-type protein) and retrieve information related to disease (Landrum et al., 2014; Piñero et al., 2020). LSPT and LLPS scores and profiles are provided (**Figure 1C**; **Methods**).

- Searching for a specific RNA, the user can retrieve information on the condensate state (SG/PB), the RNA secondary structure content (table and profile predicted using CROSS, http://s.tartaglialab.com/page/cross_group), the experimentally validated RNA interactions (RISE database, http://rise.life.tsinghua.edu.cn/) and the predicted or experimentally validated protein interactions reported in RNAact (https://rnact.crg.eu/) for both the reference sequence and SNVs (Mj et al., 2018; Piñero et al., 2020)(**Figure 1D,E**; **Methods**). The RNA-RNA interactions table reports information on different binding partners, if the interactors belong to a condensate, binding sites location in the transcripts and related SNVs (**Figure 1F**; **Methods**). The SNV section reports the position in the transcript, the difference in secondary structure compared to the reference (a numerical value and a profile image are provided) (Delli Ponti et al., 2017), associated diseases and interactions with RNAs (RISE database) as well as proteins (eCLIP https://www.encodeproject.org/eclip/; **Methods**) that involve the SNV containing region (**Figure 1F**; **Methods**).

For most genes, information is available at both the protein and RNA levels, so it is possible to navigate from one molecule to the other, revealing the links between them.

APPLICATIONS

PRALINE is a database that provides a comprehensive view of protein and RNA interactions and SNVs in human liquid-like and solid-like condensates. Information about experimentally validated and predicted molecular interactions, including protein-protein, protein-RNA and RNA-RNA, is provided, as well as the predicted RNA secondary structure content and both LLPS and LSPT propensities of proteins.

For each SNV, we provide a description of the associated diseases, the binding sites and the change in RNA secondary structure, LLPS and LSPT propensites. Combining physico-chemical properties of molecules and disease-related annotations, PRALINE helps to unravel macromolecular connections that sustain different types of condensates and how variants can affect their equilibrium. PRALINE is the first database providing LLPS and LSPT predictions for SNPs, and we envisage that it would greatly facilitate the design of experiments to study condensates' formation and implication in human diseases. We note that although tested extensively and validated experimentally, *cat*GRANULE predictions could not be benchmarked against a database of individual SNVs causing LLPS, due to lack of adequate published resources. The availability of such databases will lead to a more precise understanding of the relationship between SNVs, structural conformations, protein-RNA assembly and diseases.

ACKNOWLEDGEMENTS

The authors would like to thank Adriano Setti for the RNA-RNA interactions section and Leila Mansouri for the database name.

Funding: Our research was supported by the ERC ASTRA_855923 and H2020 projects IASIS_727658 and INFORE_825080.

REFERENCES

Armaos, A., Zacco, E., Sanchez de Groot, N., and Tartaglia, G. G. (2021). RNA-protein interactions: Central players in coordination of regulatory networks. *BioEssays* 43, 2000118. doi: 10.1002/bies.202000118.

Bolognesi, B., Lorenzo Gotor, N., Dhar, R., Cirillo, D., Baldrighi, M., Tartaglia, G. G., et al. (2016). A Concentration-Dependent Liquid Phase Separation Can Cause Toxicity upon Increased Protein Expression. *Cell Rep* 16, 222–231. doi: 10.1016/j.celrep.2016.05.076.

Cid-Samper, F., Gelabert-Baldrich, M., Lang, B., Lorenzo-Gotor, N., Ponti, R. D., Severijnen, L.-A. W. F. M., et al. (2018). An Integrative Study of Protein-RNA Condensates Identifies Scaffolding RNAs and Reveals Players in Fragile X-Associated Tremor/Ataxia Syndrome. *Cell Rep* 25, 3422-3434.e7. doi: 10.1016/j.celrep.2018.11.076.

Cunningham, F., Allen, J. E., Allen, J., Alvarez-Jarreta, J., Amode, M. R., Armean, I. M., et al. (2022). Ensembl 2022. *Nucleic Acids Res* 50, D988–D995. doi: 10.1093/nar/gkab1049.

Delli Ponti, R., Marti, S., Armaos, A., and Tartaglia, G. G. (2017). A high-throughput approach to profile RNA structure. *Nucleic Acids Res* 45, e35–e35. doi: 10.1093/nar/gkw1094.

Dobson, C. M. (2017). The Amyloid Phenomenon and Its Links with Human Disease. *Cold Spring Harb Perspect Biol* 9, a023648. doi: 10.1101/cshperspect.a023648.

Gong, J., Shao, D., Xu, K., Lu, Z., Lu, Z. J., Yang, Y. T., et al. (2018). RISE: a database of RNA interactome from sequencing experiments. *Nucleic Acids Res* 46, D194–D201. doi: 10.1093/nar/gkx864.

GTEx Consortium (2020). The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science* 369, 1318–1330. doi: 10.1126/science.aaz1776.

Hubstenberger, A., Courel, M., Bénard, M., Souquere, S., Ernoult-Lange, M., Chouaib, R., et al. (2017). P-Body Purification Reveals the Condensation of Repressed mRNA Regulons. *Molecular Cell* 68, 144-157.e5. doi: 10.1016/j.molcel.2017.09.003.

Jain, S., Wheeler, J. R., Walters, R. W., Agrawal, A., Barsic, A., and Parker, R. (2016). ATPase-Modulated Stress Granules Contain a Diverse Proteome and Substructure. *Cell* 164, 487–498. doi: 10.1016/j.cell.2015.12.038.

Khong, A., Matheny, T., Jain, S., Mitchell, S. F., Wheeler, J. R., and Parker, R. (2017). The Stress Granule Transcriptome Reveals Principles of mRNA Accumulation in Stress Granules. *Mol. Cell* 68, 808-820.e5. doi: 10.1016/j.molcel.2017.10.015.

Landrum, M. J., Lee, J. M., Riley, G. R., Jang, W., Rubinstein, W. S., Church, D. M., et al. (2014). ClinVar: public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Res* 42, D980-985. doi: 10.1093/nar/gkt1113.

Lang, B., Armaos, A., and Tartaglia, G. G. (2019). RNAct: Protein-RNA interaction predictions for model organisms with supporting experimental data. *Nucleic Acids Res* 47, D601–D606. doi: 10.1093/nar/gky967.

Lorenzo Gotor, N., Armaos, A., Calloni, G., Torrent Burgas, M., Vabulas, R. M., De Groot, N. S., et al. (2020). RNA-binding and prion domains: the Yin and Yang of phase separation. *Nucleic Acids Research* 48, 9491–9504. doi: 10.1093/nar/gkaa681.

Marchese, D., de Groot, N. S., Lorenzo Gotor, N., Livi, C. M., and Tartaglia, G. G. (2016). Advances in the characterization of RNA-binding proteins. *Wiley Interdiscip Rev RNA* 7, 793–810. doi: 10.1002/wrna.1378.

Markmiller, S., Soltanieh, S., Server, K. L., Mak, R., Jin, W., Fang, M. Y., et al. (2018). Context-Dependent and Disease-Specific Diversity in Protein Interactions within Stress Granules. *Cell* 172, 590-604.e13. doi: 10.1016/j.cell.2017.12.032.

Mj, L., Jm, L., M, B., Gr, B., C, C., S, C., et al. (2018). ClinVar: improving access to variant interpretations and supporting evidence. *Nucleic acids research* 46. doi: 10.1093/nar/gkx1153.

Oughtred, R., Rust, J., Chang, C., Breitkreutz, B.-J., Stark, C., Willems, A., et al. (2021). The

BioGRID database: A comprehensive biomedical resource of curated protein, genetic, and chemical interactions. *Protein Science* 30, 187–200. doi: 10.1002/pro.3978.

Patel, A., Lee, H. O., Jawerth, L., Maharana, S., Jahnel, M., Hein, M. Y., et al. (2015). A Liquid-to-Solid Phase Transition of the ALS Protein FUS Accelerated by Disease Mutation. *Cell* 162, 1066–1077. doi: 10.1016/j.cell.2015.07.047.

Piñero, J., Ramírez-Anguita, J. M., Saüch-Pitarch, J., Ronzano, F., Centeno, E., Sanz, F., et al. (2020). The DisGeNET knowledge platform for disease genomics: 2019 update. *Nucleic Acids Res* 48, D845–D855. doi: 10.1093/nar/gkz1021.

Sanchez de Groot, N., Armaos, A., Graña-Montes, R., Alriquet, M., Calloni, G., Vabulas, R. M., et al. (2019). RNA structure drives interaction with proteins. *Nat Commun* 10, 3246. doi: 10.1038/s41467-019-10923-5.

Tartaglia, G. G., Pawar, A. P., Campioni, S., Dobson, C. M., Chiti, F., and Vendruscolo, M. (2008). Prediction of aggregation-prone regions in structured proteins. *J Mol Biol* 380, 425–36. doi: S0022-2836(08)00567-6.

The UniProt Consortium (2021). UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Research* 49, D480–D489. doi: 10.1093/nar/gkaa1100.

Treeck, B. V., Protter, D. S. W., Matheny, T., Khong, A., Link, C. D., and Parker, R. (2018). RNA self-assembly contributes to stress granule formation and defining the stress granule transcriptome. *PNAS*, 201800038. doi: 10.1073/pnas.1800038115.

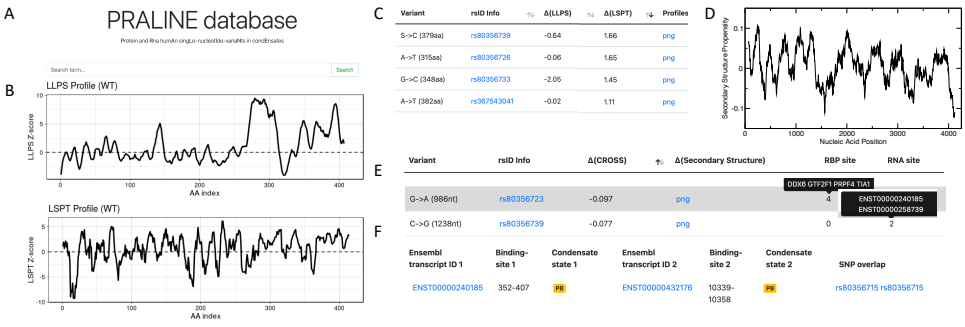
Van Nostrand, E. L., Freese, P., Pratt, G. A., Wang, X., Wei, X., Xiao, R., et al. (2020). A large-scale binding and functional map of human RNA-binding proteins. *Nature* 583, 711–719. doi: 10.1038/s41586-020-2077-3.

Vandelli, A., Cid Samper, F., Torrent Burgas, M., Sanchez de Groot, N., and Tartaglia, G. G. (2022). The Interplay Between Disordered Regions in RNAs and Proteins Modulates Interactions Within Stress Granules and Processing Bodies. *J Mol Biol* 434, 167159. doi: 10.1016/j.jmb.2021.167159.

Vendruscolo, M., and Fuxreiter, M. (2022). Sequence Determinants of the Aggregation of Proteins Within Condensates Generated by Liquid-liquid Phase Separation. *Journal of Molecular Biology* 434, 167201. doi: 10.1016/j.jmb.2021.167201.

Youn, J.-Y., Dunham, W. H., Hong, S. J., Knight, J. D. R., Bashkurov, M., Chen, G. I., et al. (2018). High-Density Proximity Mapping Reveals the Subcellular Organization of mRNA-Associated Granules and Bodies. *Molecular Cell* 69, 517-532.e11. doi: 10.1016/j.molcel.2017.12.020.

Figure 1. PRALINE database (**A**) Search bar. The input can be a protein or an RNA in different ID formats (**B**) Liquid-liquid phase-separation (LLPS) and liquid-solid phase-transition (LSPT) propensity profiles of a protein are predicted using *cat*GRANULE and Zyggregator algorithms. (**C**) Protein SNVs description table: the difference in LLPS and LSPT compared to the WT is provided. (**D**) CROSS secondary structure propensity profile image of a RNA sequence. (**E**) RNA SNVs description table: the difference in CROSS secondary structure propensity, compared to the WT, corresponding to a 11-nt window around the mutation is provided, as well as proteins and RNAs interacting with the query transcript containing the SNV. (**F**) Example of an RNA-RNA interaction table. The information about RNAs' binding sites, condensates localization and SNVs falling inside at least one of the binding sites are reported. The examples **B-F** relate to *TARDBP*.


A
PRALINE database
Protein and RNA human single-nucleotide variants in condensates
Search term...
Search
B
LLPS Profile (WT)
LLPS Z-score
AA index
LSPT Profile (WT)
LSPT Z-score
AA index
C
Variant
rsID Info
Δ(LLPS)
Δ(LSPT)
Profiles
S->C (379aa)
rs80356739
-0.64
1.66
png
A->T (315aa)
rs80356726
-0.06
1.65
png
G->C (348aa)
rs80356733
-2.05
1.45
png
A->T (382aa)
rs367543041
-0.02
1.11
png
D
Secondary Structure α-Propensity
Nucleic Acid Position
E
Variant
rsID Info
Δ(CROSS)
Δ(Secondary Structure)
RBP site
RNA site
DDX6 GTF2F1 PRPF4 TIA1
G->A (986nt)
rs80356723
-0.097
png
4
ENST00000240185
ENST00000258739
C->G (1238nt)
rs80356739
-0.077
png
0
2
F
Ensembl transcript ID 1
Binding-site 1
Condensate state 1
Ensembl transcript ID 2
Binding-site 2
Condensate state 2
SNP overlap
ENST00000240185
352-407
PS
ENST00000432176
10339-10358
PS
rs80356715 rs80356715

CHAPTER 6

STRUCTURAL ANALYSIS OF SARS-COV-2 GENOME AND PREDICTIONS OF THE HUMAN INTERACTOME

Structural analysis of SARS-CoV-2 genome and predictions of the human interactome

In early 2020, the insurgence of the Covid-19 pandemic raised the need to quickly investigate the molecular pathways of SARS-CoV-2 infection and to identify the key players involved in the host-virus interactions. At the time, many groups decided to tackle the difficult task of studying this new 30kb virus by comparing several known coronavirus strains using multiple sequence alignments tools, in an attempt to highlight possible similarities (Li et al., 2020; Alqahtani et al., 2020).
In this context, our group focused on the computational analysis of structurally relevant viral genomic regions and on predicting the interactions between SARS-CoV-2 and the human proteome, with the aim of providing a list of potential candidates that could be relevant in the infection process.

Comparing different coronavirus species and strains, we show that the viral regions corresponding to the nucleotides 22000-23000 are highly conserved at the structural level, while the region one thousand nucleotides upstream is very variable. These two regions code for a domain in spike S protein that binds to the human ACE2 receptor, and is responsible in MERS-CoV respiratory syndrome for the interaction with sialic acids. The 5' and 3' of the viral genome are instead predicted to be highly structured attractors of proteins, among which some seem to be known condensate components and are implicated in other known viral infection processes. Furthermore, some of these candidates are involved in the innate immune response of the cell, so could be potentially targeted by the

virus to escape the organism's defenses, preventing the storage of these elements inside phase-separating condensates.
In this regard, in more recent studies several DEAD-box helicases (e.g. DDX1, DDX6) have been proven experimentally to be exploited by SARS-CoV-2 to increase its infectivity and to tamper with SGs and PBs formation and a similar workflow could be used to unravel infection mechanisms of other pathogens.

An experimental follow-up of this project is currently under submission and available in BioRxiv at https://www.biorxiv.org/content/10.1101/2022.07.18.499583v1.full.pdf+html

This work was published in the Nucleic Acids Research journal in 2020.

Vandelli, A, Monti, M., Milanetti, E., Armaos, A., Rupert, J., Zacco, E., Bechara, E., Delli Ponti, R., and Tartaglia, G. G. (2020). Structural analysis of SARS-CoV-2 genome and predictions of the human interactome. *Nucleic Acids Research*, 48(20): 11270–11283. DOI: 10.1093/nar/gkaa864

11270–11283 *Nucleic Acids Research, 2020, Vol. 48, No. 20* *Published online 17 October 2020*
doi: 10.1093/nar/gkaa864

Structural analysis of SARS-CoV-2 genome and predictions of the human interactome

Andrea Vandelli[1,2], Michele Monti[1,3], Edoardo Milanetti[4,5], Alexandros Armaos[1,3], Jakob Rupert[3,6], Elsa Zacco[3], Elias Bechara[1,3], Riccardo Delli Ponti[7,*] and Gian Gaetano Tartaglia[1,3,6,8,*]

[1]Centre for Genomic Regulation (CRG), The Barcelona Institute for Science and Technology, Dr. Aiguader 88, 08003 Barcelona, Spain and Universitat Pompeu Fabra (UPF), 08003 Barcelona, Spain, [2]Systems Biology of Infection Lab, Department of Biochemistry and Molecular Biology, Biosciences Faculty, Universitat Autònoma de Barcelona, 08193 Cerdanyola del Vallès, Spain, [3]Center for Human Technologies, Istituto Italiano di Tecnologia, Via Enrico Melen 83, 16152 Genoa, Italy, [4]Department of Physics, Sapienza University, Piazzale Aldo Moro 5, 00185 Rome, Italy, [5]Center for Life Nanoscience, Istituto Italiano di Tecnologia, Viale Regina Elena 291, 00161 Rome, Italy, [6]Department of Biology 'Charles Darwin', Sapienza University of Rome, P.le A. Moro 5, Rome 00185, Italy, [7]School of Biological Sciences, Nanyang Technological University, 60 Nanyang Drive, Singapore 637551, Singapore and [8]Institucio Catalana de Recerca i Estudis Avançats (ICREA), 23 Passeig Lluis Companys, 08010 Barcelona, Spain

Received August 19, 2020; Revised September 15, 2020; Editorial Decision September 17, 2020; Accepted September 25, 2020

ABSTRACT

Specific elements of viral genomes regulate interactions within host cells. Here, we calculated the secondary structure content of >2000 coronaviruses and computed >100 000 human protein interactions with severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). The genomic regions display different degrees of conservation. SARS-CoV-2 domain encompassing nucleotides 22 500–23 000 is conserved both at the sequence and structural level. The regions upstream and downstream, however, vary significantly. This part of the viral sequence codes for the Spike S protein that interacts with the human receptor angiotensin-converting enzyme 2 (ACE2). Thus, variability of Spike S is connected to different levels of viral entry in human cells within the population. Our predictions indicate that the 5′ end of SARS-CoV-2 is highly structured and interacts with several human proteins. The binding proteins are involved in viral RNA processing, include double-stranded RNA specific editases and ATP-dependent RNA-helicases and have strong propensity to form stress granules and phase-separated assemblies. We propose that these proteins, also implicated in viral infections such as HIV, are selectively recruited by SARS-CoV-2 genome to alter transcriptional and post-transcriptional regulation of host cells and to promote viral replication.

INTRODUCTION

A disease named Covid-19 by the World Health Organization and caused by the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) has been recognized as responsible for the pneumonia outbreak that started in December 2019 in Wuhan City, Hubei, China (1) and spread in February to Milan, Lombardy, Italy (2) becoming pandemic.

SARS-CoV-2 is a positive-sense single-stranded RNA virus that shares similarities with other beta-coronavirus such as severe acute respiratory syndrome coronavirus (SARS-CoV) and Middle East respiratory syndrome coronavirus (MERS-CoV) (3). Bats have been identified as the primary host for SARS-CoV and SARS-CoV-2 (4,5) but the intermediate host linking SARS-CoV-2 to humans is still unknown, although a recent report indicates that pangolins could be involved (6).

Coronaviruses use species-specific proteins to mediate the entry in the host cell and the spike S protein activates the infection in human respiratory epithelial cells in SARS-CoV, MERS-CoV and SARS-CoV-2 (7). Spike S is assembled as a trimer and contains around 1300 amino acids within each unit (8,9). The receptor binding domain (RBD) of Spike S, which contains around 300 amino acids, mediates the binding with angiotensin-converting enzyme (ACE2), attacking respiratory cells. A region upstream of the RBD, present in

*To whom correspondence should be addressed. Tel: +39 010 28976204; Fax: +39 010 2897 621; Email: gian.tartaglia@iit.it
Correspondence may also be addressed to Riccardo Delli Ponti. Email: riccardo.ponti@ntu.edu.sg

MERS-CoV but not in SARS-CoV, is involved in the adhesion to sialic acid-containing oligosaccharides and plays a key role in regulating viral infection (7,10).

At present, few molecular details are available on SARS-CoV-2 and its interactions with the human host, which are mediated by specific RNA elements (11). To study the RNA structural content, we used *CROSS* (12) that was previously developed to investigate large transcripts such as the human immunodeficiency virus HIV-1 (13). *CROSS* predicts the structural profile of RNA molecules (single- and double-stranded state) at single-nucleotide resolution using sequence information only. Here, we performed sequence and structural alignments among SARS-CoV-2 strains available and identified the conservation of specific elements in the spike S region, which provides clues on the evolution of domains involved in the binding to ACE2 and sialic acid.

As highly structured RNAs have strong propensity to form stable contacts with different proteins (14) and promote specific assembly of complexes (15,16), SARS-CoV-2 domains enriched in double-stranded content are expected to establish interactions within host cells that are important to replicate the virus (17). To investigate the interactions of SARS-CoV-2 RNA with human proteins, we employed *cat*RAPID (18,19). *cat*RAPID (20) estimates the binding potential of a specific protein for an RNA molecule through van der Waals, hydrogen bonding and secondary structure propensities allowing identification of interaction partners with high confidence (21). The computational analysis of more than 100 000 interactions with SARS-CoV-2 RNA reveals that the 5′ end of SARS-CoV-2 has strong propensity to bind to human proteins involved in viral infection and reported to be associated with HIV infection. A comparison between SARS-CoV and HIV reveals similarities (22) that are still unexplored. Interestingly, HIV and SARS-CoV-2, but not SARS-CoV nor MERS-CoV, have a furin-cleavage site occurring in the spike S protein, which could explain the high velocity spread of SARS-CoV-2 compared to SARS-CoV and MERS-CoV (23,24).

We hope that our large-scale calculations of structural properties and binding partners of SARS-CoV-2 will be useful to identify the mechanisms of virus replication within the human host.

MATERIALS AND METHODS

Structure prediction

We computed the secondary structure of transcripts using *CROSS* (Computational Recognition of Secondary Structure) (12,13). The algorithm predicts the structural profile (single- and double-stranded state) at single-nucleotide resolution using sequence information only and without sequence length restrictions (scores > 0 indicate double stranded regions). We used the *Vienna* RNA Package (25) to further investigate the RNA secondary structure of minima and maxima identified with *CROSS* (13).

CROSS alive was employed to predict SARS-CoV-2 secondary structure *in vivo* (26). *CROSS alive* (m6A+ fast option) predicts long range interactions and can identify pseudoknots of 50–100 nucleotides. The *RF-Fold* algorithm of the *RNAFramework* suite (26) was used to identify pseudoknots in SARS-CoV-2. In this analysis, the partition function was calculated using CROSS calculations as soft-constraints. RNA was then folded employing *Vienna* RNA Package (25) and pseudo-knotted bases were hard-constrained to be single-stranded.

MN908947 predictions are available at http://crg-webservice.s3.amazonaws.com/submissions/2020-05/270257/output/index.html?unlock=fd65439e7b (*CROSS*) and also http://crg-webservice.s3.amazonaws.com/submissions/2020-05/271372/output/index.html?unlock=1de1d3a54a (*CROSS alive*).

Structural conservation

We used *CROSSalign* (12,13), an algorithm based on Dynamic Time Warping (DTW), to check and evaluate the structural conservation between different viral genomes (13). *CROSSalign* was previously employed to study the structural conservation of ∼5000 HIV genomes. SARS-CoV-2 fragments (1000 nt, not overlapping) were searched inside other complete genomes using the OBE (open begin and end) module, in order to search a small profile inside a larger one. The lower the structural distance, the higher the structural similarities (with a minimum of 0 for almost identical secondary structure profiles). The significance is assessed as in the original publication (13).

The Infernal package (version 1.1.3) was employed to build covariance models (CMs) for fragments 22, 23 and 24 (27). The package was then used to search for sequence and structural similarities among RNAs in our database (267 representative sequences), which allows to identify a series of matches below a specific E-value threshold (0.1, 1 and 10). The analysis shows agreement with *CROSSalign* (12,13) results. The minimum and maximum number of identified motifs were 224 and 4878 (E-value of 10), 136 and 3093 (*E*-value of 1) and 94 and 1060 (E-value of 0.1). The motifs in Spike S region were counted for annotated coronaviruses (239 genomes out of 246, of which 161 within *E*-value of 0.1).

Sequence collection

The FASTA sequences of the complete genomes of SARS-CoV-2 were downloaded in March 2020 from Virus Pathogen Resource (VIPR; www.viprbrc.org), for a total of 62 strains. An additional non-redundant set was downloaded in August 2020 for further analyses (462 sequences). Regarding the other coronaviruses, the sequences were downloaded in March 2020 from NCBI selecting only complete genomes, for a total of 2040 genomes. The reference Wuhan sequence with available annotation (EPI_ISL_402119) was downloaded from Global Initiative on Sharing All Influenza Data in March 2020 (GISAID https://www.gisaid.org/).

Protein-RNA interaction prediction

Interactions between each fragment of target sequence and the human proteome were predicted using *cat*RAPID *omics* (18,19), an algorithm that estimates the binding propensity of protein–RNA pairs by combining secondary structure, hydrogen bonding and van der Waals

11272 *Nucleic Acids Research, 2020, Vol. 48, No. 20*

contributions. As reported in a recent analysis of about half a million of experimentally validated interactions (21), the algorithm is able to separate interacting vs non-interacting pairs with an area under the ROC curve of 0.78. The complete list of interactions between the 30 fragments and the human proteome is available at http://crg-webservice.s3.amazonaws.com/submissions/2020-03/252523/output/index.html?unlock=f6ca306af0. The output then is filtered according to the *Z*-score column, which is the interaction propensity normalised by the mean and standard deviation calculated over the reference RBP set (http://s.tartaglialab.com/static_files/shared/faqs.html#4). We used three different thresholds in ascending order of stringency: *Z* greater or equal than 1.50, 1.75 and 2 respectively and for each threshold we then selected the proteins that were unique for each fragment for each threshold. *omiXscore* calculations of ADAR and ADARB1 are interactions are respectively at http://crg-webservice.s3.amazonaws.com/submissions/2020-04/263420/output/index.html?unlock=f9375fdbf9 and http://crg-webservice.s3.amazonaws.com/submissions/2020-04/263140/output/index.html?unlock=bb28d715ea.

GO terms analysis

*clever*GO (28), an algorithm for the analysis of Gene Ontology annotations, was used to determine which fragments present enrichment in GO terms related to viral processes. Analysis of functional annotations was performed in parallel with *GeneMania* (29). The link to *clever*GO analyses for fragment 1 is at http://www.tartaglialab.com/GO_analyser/render_GO_universal/3073/0d66e887c3/ ($Z \geq 2$).

RNA and protein alignments

We used *Clustal W* (30) for 62 SARS-CoV-2 strains alignments and *T-Coffee* (31) for spike S proteins alignments. The variability in the spike S region was measured by computing Shannon entropy on translated RNA sequences. The Shannon entropy is computed as follows:

$$S(a) = -\sum_{i} P(a,i)\log P(a,i)$$

where *a* correspond to the amino acid at the position *i* and $P(a,i)$ is the frequency of a certain amino-acid *a* at position *i* of the sequence. Low entropy indicates poorly variability: if $P(a,x) = 1$ for one *a* and 0 for the rest, then $S(x) = 0$. By contrast, if the frequencies of all amino acids are equally distributed, the entropy reaches its maximum possible value.

Predictions of phase separation

*cat*GRANULE (32) was employed to identify proteins assembling into biological condensates. Scores >0 indicate that a protein is prone to phase separate. Structural disorder, nucleic acid binding propensity and amino acid patterns such as arginine–glycine and phenylalanine–glycine are key features combined in this computational approach (32).

RESULTS

SARS-CoV-2 contains highly structured elements

Structured elements within RNA molecules attract proteins (14) and reveal regions important for interactions with the host (33). Indeed, each gene expressed from SARS-CoV-2 is preceded by conserved transcription-regulating sequences that act as signal for the transcription complex during the synthesis of the RNA minus strand to promote a strand transfer to the leader region to resume the synthesis. This process is named discontinuous extension of the minus strand and is a variant of similarity-assisted template switching that operates during viral RNA recombination (17).

To analyze SARS-CoV-2 structure (reference Wuhan strain MN908947.3), we employed *CROSS* (12) that was previously developed to predict the double- and single-stranded content of RNA genomes such as HIV-1 (13). We found the highest density of double-stranded regions in the 5′ end (nucleotides 1–253), membrane M protein (nucleotides 26 523–27 191), spike S protein (nucleotides 21 563–25 384), and nucleocapsid N protein (nucleotides 28 274–29 533; Figure 1A) (34). The lowest density of double-stranded regions were observed at nucleotides 6 000–6 250 and 20 000–21 500 and correspond to the regions between the non-structural proteins nsp14 and nsp15 and the upstream region of the spike surface protein S (Figure 1) (34). In addition to the maximum corresponding to nucleotides 22 500–23 000, the structural content of Spike S protein shows minima at around nucleotides 21 500–22 000 and 23 500–24 000 (Figure 1). We used the *Vienna* method (25) to further investigate the RNA secondary structure of specific regions identified with *CROSS* (13). Employing a 100-nucleotide window centered around *CROSS* maxima and minima, we found good match between *CROSS* scores and Vienna free energies (Figure 1).

RNA structure *in vitro* and *in vivo* could be significantly different due to interactions with proteins and other molecules (26). Using *CROSS alive* to predict the double- and single-stranded content of SARS-CoV-2 in the cellular context, we found that both the 5′ and 3′ ends are the most structured regions followed by nucleotides 22 500–23 000 in the Spike S region, while nucleotides 6 000–6 250 and 20 000–21 500 have the lowest density of double-stranded regions (Figure 1B). The region corresponding to nucleotides 13 400–13 600 shows high density of contacts. This part of SARS-CoV-2 sequence has been proposed to form a pseudoknot (35) that is also visible in *CROSS* profile (Figure 1A), but *CROSS alive* is able to identify long range interactions and better identifies the region. Additionally, we used the *RF-Fold* algorithm of the *RNAFramework* suite (36) (Material and Methods) to search for pseudoknots. Employing *CROSS* as a soft-constraint for RF-Fold, we predicted 6 pseudoknots (nucleotides 3 394–3 404, 13 723–13 732, 14 677–14 711, 16 867–16 905, 24 844–24 884, 27 969–27 990). The pseudoknot at nucleotides 13 723–13 732 is in close proximity to the one proposed for SARS-CoV-2 (35) and the one at nucleotides 27 969–27 990 is at the 3′ end, where pseudonoknots have been shown to occur in coronaviruses (37).

Nucleic Acids Research, 2020, Vol. 48, No. 20 11273

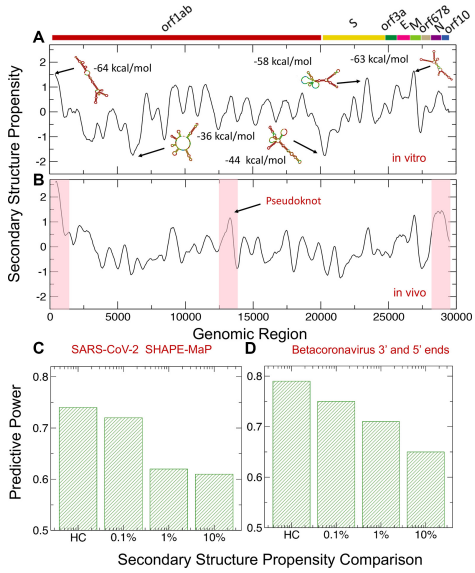


Figure 1. Predictions of SARS-CoV-2 structure. (**A**) Using the *CROSS* approach (12,13), (**A**) we predicted the structural content of SARS-CoV-2 *in vitro*. We found the highest density of double-stranded regions in the 5′ end (nucleotides 1–250) and within membrane M protein (nucleotides 26 500–27 000), and spike S protein (nucleotides 22 500–23 000) regions. Regions with the highest structural content are predicted by *Vienna* to have the lowest free energies. (**B**) Using *CROSS alive* (26), we studied the structural content of SARS-CoV-2 *in vivo*. The 5′ and 3′ ends (indicated by red boxes) are predicted to be highly structured. In addition, nucleotides 22 500–23 000 in Spike S region and nucleotides 13 400–13 600 (indicated by a red box) forming a pseudoknot (35) show high density of contacts. (**C**) Comparison of *CROSS* predictions with the secondary structure landscape of SARS-CoV-2 revealed by SHAPEMaP (38). From low (10%) to high (0.1%) confidence scores, the predictive power, measured as the Area Under the Curve (AUC) of Receiver Operating Characteristics (ROC), increases monotonically (HC corresponds to 10 nucleotides with highest/lowest scores). (**D**) CROSS performances on betacoronavirus 5′ and 3′ ends (39–42). Using different confidence scores, we show that CROSS is able to identify double and single stranded regions with great predictive power.

To validate our results, we compared *CROSS* predictions of double- and single-stranded content (as released in March 2020) with the secondary structure landscape of SARS-CoV-2 revealed by SHAPE mutational profiling (SHAPEMaP) (38). In their experimental work, Manfredonia *et al*. carried out *in vitro* refolding of RNA followed by probing with 2-methylnicotinic acid imidazolide. In our comparison, balanced lists of single and double stranded regions were used for the calculations: A confidence score of 10% indicates that we compared the SHAPE reactivity values of 3000 nucleotides associated with the highest *CROSS* scores (i.e. double stranded) and 3000 nucleotides associated with the lowest *CROSS* scores (i.e. single stranded). From low (10%) to high (0.1%) confidence scores, we observed that the predictive power, measured as the Area Under the Curve (AUC) of Receiver Operating Characteristics (ROC), increases monotonically reaching the value of 0.73 (the AUC is 0.74 for the 10 highest/lowest scores; Figure 1C), which indicates that *CROSS* reproduces SHAPEMaP in great detail.

We also assessed *CROSS* performances on structures of betacoronavirus 5′ and 3′ ends (39–42) (Figure 1D). In this analysis, we used RFAM multiple sequence alignments of betacoronavirus 5′ and 3′ ends and relative consensus structures (RF03117 and RF03122) (39–42). We generated the 2D representation of nucleotide chains of consensus structures. We extracted the 'secondary structure occupancy', as defined in a previous work (20), and counted the contacts present around each nucleotide. Following the procedure used for the comparison with SHAPEMaP, different progressive cut-offs were used for ranking all the structures using balanced lists of single and double stranded regions: 10% indicates that we compared 600 nucleotides associated with the highest amount of contacts and 600 nucleotides associated with the lowest amount of contacts. From low (10%) to high (0.1%) confidence scores we

11274 *Nucleic Acids Research, 2020, Vol. 48, No. 20*

observed that the AUC of ROC increases monotonically reaching the value of 0.75 (10 highest/lowest scores have an AUC of 0.78; Figure 1D), which indicates that *CROSS* is able to identify known double and single stranded regions reported in great detail. We also tested the ability of *CROSS* to recognize specific secondary structures in representative cases for which we studied both the 3′ and 5′ ends: NC_006213 or Human coronavirus OC43 strain ATCC VR-759, NC_019843 or Middle East respiratory syndrome coronavirus, NC_026011 or Betacoronavirus HKU24 strain HKU24-R05005I, NC_001846 or Mouse hepatitis virus strain MHV-A59 C12 and NC_012936 or Rat coronavirus Parker (Supplementary Figure S1).

In summary, our analysis identifies several structural elements in SARS-CoV-2 genome (11). Different lines of experimental and computational evidence indicate that transcripts containing a large amount of double-stranded regions have a strong propensity to recruit proteins (14,43) and can act as scaffolds for protein assembly (15,16). We therefore expected that the 5′ end attracts several host proteins because of the enrichment in secondary structure elements. The binding would not just involve proteins interacting with double-stranded regions. If a specific protein contact occurs in a loop at the end of a long RNA stem, the overall region is enriched in double-stranded nucleotides but the specific interaction takes place in a single-stranded element.

Structural comparisons reveal that a spike S region of SARS-CoV-2 is conserved among coronaviruses

We employed *CROSS*align (13) to study the structural conservation of SARS-CoV-2 in different strains (Materials and Methods).

In our analysis, we compared the Wuhan strain MN908947.3 with 2040 coronaviruses (reduced to 267 sequences upon redundancy removal at 95% sequence similarity (44); Figure 2; full data shown in Supplementary Figure S2).

We note that the regulatory regions located at the 3′ end are slightly longer (about 250–500 nts containing a bulged stem loop, a pseudoknot plus a poly-A tail) than the ones at the 5′ end (the 1–4 stem loops are within the first 200 nucleotides) and their structural elements are therefore better recognized within the 1000 nucleotides window that we use for our analysis (45). Although the 5′ end is variable, it is more structured in SARS-CoV-2 than other coronaviruses (average structural content of 0.56, indicating that 56% of the *CROSS* signal is >0). The 3′ end is less variable and slightly less structured (average structural content of 0.49). By contrast, the other coronaviruses have lower average structural content of 0.49 in the 5′ end and 0.42 in the 3′ end.

One conserved region falls inside the Spike S genomic locus between nucleotides 22 000 and 23 000 and exhibits an intricate and stable secondary structure (RNA*fold* minimum free energy = −285 kcal/mol) (25). High conservation of a structured region suggests a functional activity that is relevant for host infection.

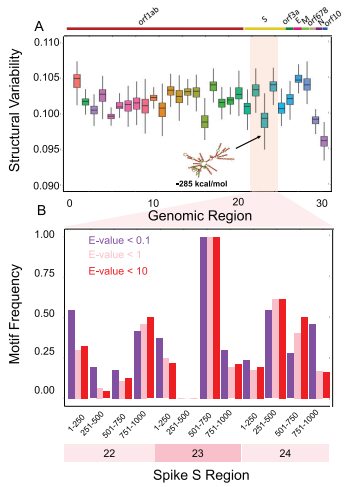


Figure 2. Structural comparisons of coronaviruses. (**A**) We employed the *CROSS*align approach (12,13) to compare Wuhan strain MN908947.3 with other coronaviruses. One of the regions with the lowest structural variability encompasses nucleotides 22 000–23 000. The centroid structure and free energy computed with the Vienna method (25) are displayed. (**B**) We studied the conservation of nucleotides 22 000–23 000 (fragment 23) and the adjacent regions using structural motives identified with RF-Fold algorithm of the RNAFramework suite (36) with *CROSS* as soft-constraint. We found that nucleotides 501–750 within fragment 23 are the ones with the highest number of matches at confidence thresholds (E-values).

To demonstrate the conservation of nucleotides 22 000–23 000 (fragment 23), we divided this region and the adjacent ones (nucleotides 21 000–22 000 and 23 000–24 000) into sub-fragments. We then used the *RF-Fold* algorithm of the *RNAFramework* suite (36) to fold the different subregions using *CROSS* predictions as soft-constraints. The structural motives identified with this procedure were employed to build covariance models (CMs) that were then searched in our set of coronaviruses using the 'Infernal' package (27). We found that nucleotides 501–750 within fragment 23 have the highest number of matches for different confidence thresholds, implying a higher chance of sequence and structure conservation across coronaviruses (*E*-values of 10,1, 0.1; Figure 2B). We specifically counted the matches falling in the Spike S region (±1000 nucleotides to take into account the division of the genome into fragments; Supplementary Table S1). For the large majority of annotated sequences, we found a match falling in the Spike S region (239 genomes out of 246, of which 161 with *E*-value

below 0.1) This further emphasizes the conservation of the region in exam.

Sequence and structural comparisons among SARS-CoV-2 strains

To better investigate the sequence conservation of SARS-CoV-2, we compared 62 strains isolated from different countries during the pandemic (including China, USA, Japan, Taiwan, India, Brazil, Sweden and Australia; data from NCBI and in VIPR www.viprbrc.org; Materials and Methods). Our analysis aims to determine the relationship between structural content and sequence conservation.

Using *ClustalW* for multiple sequence alignments (30), we observed general conservation of the coding regions (Figure 3A). The 5′ and 3′ ends show high variability due to practical aspects of RNA sequencing and are discarded in this analysis (46). Indeed, their sequences are less well characterized (47), and their variation results higher than other parta of the viral sequence. One highly conserved region is between nucleotides 22 000 and 23 000 in the Spike S genomic locus, while sequences up- and downstream are variable (purple bars in Figure 3A). We then used *CROSSalign* (13) to compare the structural content (Material and Methods). High variability of structure is observed for both the 5′ and 3′ ends and for nucleotides 21 000–22 000 as well as 24 000–25 000, associated with the Spike S region (purple bars in Figure 3A). The rest of the regions are significantly conserved at a structural level (P-value < 0.0001; Fisher's test).

We note that sequence conservation (Figure 3A) and secondary structure profiles (Figure 1A) are statistically related. Following the analysis to compare *CROSS* and SHAPE scores, we selected balanced groups of nucleotides with the highest and lowest sequence conservation and measured their single and double stranded content: a conservation score of 1% indicates that we compared 300 nucleotides with the highest sequence similarity and 300 nucleotides with the lowest sequence similarity. At conservation score of 1% (or less stringent threshold of 10%), the match between similarity and structure, measured as the AUC of ROC is 0.76 (or 0.60, respectively). The association is statistically significant: shuffling the sequence conservation profiles, the empirical P-values are <0.02 (at both 10% and 1% conservation scores).

We also compared protein sequences coded by the Spike S genomic locus (NCBI reference QHD43416) and found that both sequence (Figure 3A) and structure (Figure 2) of nucleotides 22 000–23 000 are highly conserved. The region corresponds to amino acids 460–520 that contact the host receptor angiotensin-converting enzyme 2 (ACE2) (48) promoting infection and provoking lung injury (24,49). By contrast, the region upstream of the binding site receptor ACE2 and located in correspondence to the minimum of the structural profile at around nucleotides 22 500–23 000 (Figure 1) is highly variable (31), as indicated by *T-coffee* multiple sequence alignments (31) (Figure 3A). This part of the Spike S region corresponds to amino acids 243–302 that in MERS-CoV binds to sialic acids regulating infection through cell–cell membrane fusion (Figure 3B; see related manuscript by E. Milanetti *et al.*) (10,50,51).

Our analysis suggests that the structural region between nucleotides 22 000 and 23 000 of Spike S region is conserved among coronaviruses (Figure 2) and that the binding site for ACE2 has poor variation in human SARS-CoV-2 strains (Figure 3B). By contrast, the region upstream of the ACE2 binding site, which has also propensity to bind sialic acids (10,50,51), showed poor structural content and high variability (Figure 3B). The region downstream of the ACE2 binding site and located at the beginning of S2 domain shows high variability (Figure 3B). The results are confirmed by analysing a pool of 462 genomes having a ± 5 nucleotides length difference with respect to MN908947.3 (August 2020; Supplementary Figure S3).

Analysis of human interactions with SARS-CoV-2 identifies proteins involved in viral replication

In order to obtain insights on how the virus replicates in human cells, we predicted SARS-CoV-2 interactions with the whole RNA-binding human proteome. Following a protocol to study structural conservation in viruses (13), we first divided the Wuhan sequence in 30 fragments of 1000 nucleotides each moving from the 5′ to 3′ end and then calculated the protein–RNA interactions of each fragment with *cat*RAPID *omics* (3 340 canonical and putative RNA-binding proteins, or RBPs, for a total 102 000 interactions) (18). Proteins such as Polypyrimidine tract-binding protein 1 PTBP1 (Uniprot P26599) showed the highest interaction propensity (or Z-score; Materials and Methods) at the 5′ end while others such as heterogeneous nuclear ribonucleoprotein Q HNRNPQ (O60506) showed the highest interaction propensity at the 3′end, in agreement with previous studies on coronaviruses (Figure 4A) (52).

For each fragment, we predicted the most significant interactions by filtering according to the Z score. We used three different thresholds in ascending order of stringency: $Z \geq 1.50$, 1.75 and 2 respectively and we removed from the list the proteins that were predicted to interact promiscuously with more than one fragment. Fragment 1 corresponds to the 5′ end and is the most contacted by RBPs ($\sim$120 with $Z \geq 2$ high-confidence interactions; Figure 4B), which is in agreement with the observation that highly structured regions attract a large number of proteins (14). Indeed, the 5′ end contains multiple stem loop structures that control RNA replication and transcription (53,54). By contrast, the 3′ end and fragment 23 (Spike S), which are still structured but to a lesser extent, attract fewer proteins (10 and 5, respectively) and fragment 20 (between Orf1ab and Spike S) that is predicted to be unstructured, does not have predicted binding partners. Fragments 1 and 29 together with the adjacent regions are also predicted to be the most structured *in vivo* and show the highest amount of contacts for different Z scores (Figure 1B).

The interactome of each fragment was analysed using *clever*GO, a tool for Gene Ontology (GO) enrichment analysis (28). Proteins interacting with fragments 1, 2 and 29 were associated with annotations related to viral processes (Figure 4C; Supplementary Table S2). Considering the three thresholds applied (Material and Methods), we found 23 viral proteins (including 2 pseudogenes), for fragment 1, 2 proteins for fragment 2 and 11 proteins for

11276 *Nucleic Acids Research, 2020, Vol. 48, No. 20*

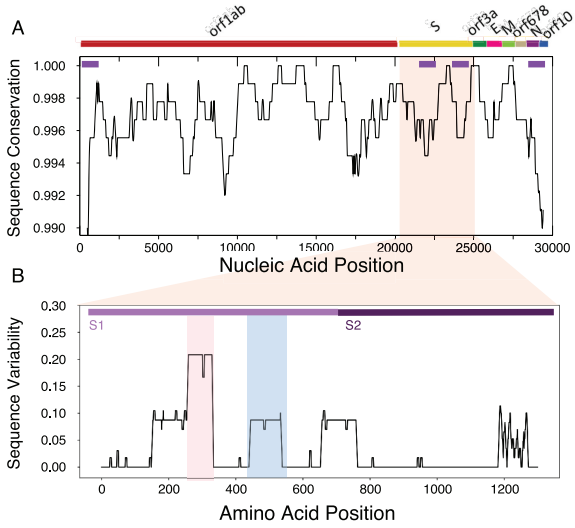
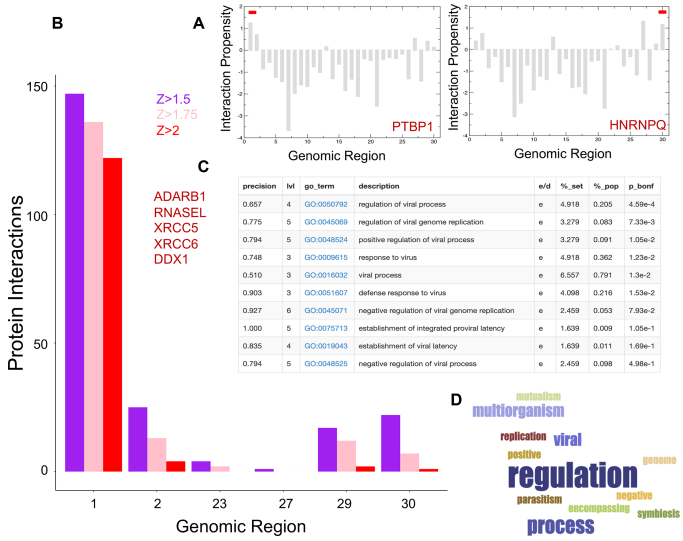


Figure 3. Sequence and structural comparison of human SARS-CoV-2 strains. (**A**) Strong sequence conservation (*ClustalW* multiple sequence alignments (28)) is observed in coding regions, including the region between nucleotides 22 000 and 23 000 of spike S protein. High structural variability (purple bars on top) is observed for both the UTRs and for nucleotides between 21 000 and 22 000 as well as 24 000 and 25 000, associated with the S region. The rest of the regions are significantly conserved at a structural level. (**B**) The sequence variability (Shannon entropy computed on T-Coffee multiple sequence alignments (31)) in the spike S protein indicate conservation between amino-acids 460 and 520 (blue box) binding to the host receptor angiotensin-converting enzyme 2 ACE2. The region encompassing amino-acids 243 and 302 is highly variable and is implicated in sialic acids in MERS-CoV (red box). The S1 and S2 domains of Spike S protein are displayed.

fragment 29 (Figure 4D). Among the high-confidence interactors of fragment 1, we discovered RBPs involved in positive regulation of viral processes and viral genome replication, such as double-stranded RNA-specific editase 1 ADARB1 (Uniprot P78563), 2–5A-dependent ribonuclease RNASEL (Q05823) and 2–5-oligoadenylate synthase 2 OAS2 (P29728; Figure 5A). Interestingly, 2–5-oligoadenylate synthase 2 OAS2 has been reported to be upregulated in human alveolar adenocarcinoma (A549) cells infected with SARS-CoV-2 (log fold change of 4.2; *P*-value of 10^{-9} and *q*-value of 10^{-6}) (55). While double-stranded RNA-specific adenosine deaminase ADAR (P55265) is absent in our library due to its length that does not meet *cat*RAPID *omics* requirements (18), the *omiXcore* extension of the algorithm specifically developed for large molecules (56) attributes the same binding propensity to both ADARB1 and ADAR, thus indicating that the interactions with SARS-CoV-2 are likely to occur (Materials and Methods). Moreover, experimental works indicate that the family of ADAR deaminases is active in bronchoalveolar lavage fluids derived from SARS-CoV-2 patients (57) and is upregulated in A549 cells infected with SARS-CoV-2 (log fold change of 0.58; *P*-value of 10^{-8} and *q*-value of 10^{-5}) (55).

We also identified proteins related to the establishment of integrated proviral latency, including X-ray repair cross-complementing protein 5 XRCC5 (P13010) and X-ray repair cross-complementing protein 6 XRCC6 (P12956; Figure 5A). In accordance with our calculations, comparison of A549 cells responses to SARS-CoV-2 and respiratory syncytial virus, indicates upregulation of XRCC6 in SARS-CoV-2 (log fold-change of 0.92; *P*-value of 0.006 and *q*-value of 0.23) (55). Moreover, previous evidence suggests that the binding of XRCC6 takes places at the 5′ end of SARS-CoV-2, thus giving further support to our predictions (58). Nucleolin NCL (P19338), a protein known to be involved in coronavirus processing, was also predicted to bind tightly to the 5′ end (Supplementary Table S2) (59).

Importantly, we found proteins related to defence response to viruses, such as ATP-dependent RNA helicase DDX1 (Q92499), that are involved in negative regulation of viral genome replication. Some DNA-binding proteins



precision	lvl	go_term	description	e/d	%_set	%_pop	p_bonf
0.657	4	GO:0050792	regulation of viral process	e	4.918	0.205	4.59e-4
0.775	5	GO:0045069	regulation of viral genome replication	e	3.279	0.083	7.33e-3
0.794	5	GO:0048524	positive regulation of viral process	e	3.279	0.091	1.05e-2
0.748	3	GO:0009615	response to virus	e	4.918	0.362	1.23e-2
0.510	3	GO:0016032	viral process	e	6.557	0.791	1.3e-2
0.903	3	GO:0051607	defense response to virus	e	4.098	0.216	1.53e-2
0.927	6	GO:0045071	negative regulation of viral genome replication	e	2.459	0.053	7.93e-2
1.000	5	GO:0075713	establishment of integrated proviral latency	e	1.639	0.009	1.05e-1
0.835	4	GO:0019043	establishment of viral latency	e	1.639	0.011	1.69e-1
0.794	5	GO:0048525	negative regulation of viral process	e	2.459	0.098	4.98e-1

Figure 4. Predictions of protein interactions with SARS-CoV-2 RNA. (**A**) In agreement with studies on coronaviruses (52), PTBP1 shows the highest interaction propensity at the 5′ and HNRNPQ at 3′ (indicated by red bars). (**B**) Number of RBP interactions for different SARS-CoV-2 regions (colours indicate catRAPID (18,19) confidence levels: $Z = 1.5$ or low $Z = 1.75$ or medium and $Z = 2.0$ or high; regions with scores lower than $Z = 1.5$ are omitted); (**C**) enrichment of viral processes in the 5′ of SARS-CoV-2 (precision = term precision calculated from the GO graph structure lvl = depth of the term; go_term = GO term identifier, with link to term description at AmiGO website; description = label for the term; e/d = enrichment / depletion compared to the population; %_set = coverage on the provided set; %_pop = coverage of the same term on the population; p_bonf = *P*-value of the enrichment. To correct for multiple testing bias, use Bonferroni correction) (28); (**D**) viral processes are the third largest cluster identified in our analysis;

such as Cyclin-T1 CCNT1 (O60563), Zinc finger protein 175 ZNF175 (Q9Y473) and Prospero homeobox protein 1 PROX1 (Q92786) were included because they could have potential RNA-binding ability (Figure 5A) (60). As for fragment 2, we found two canonical RBPs: E3 ubiquitin-protein ligase TRIM32 (Q13049) and E3 ubiquitin−protein ligase TRIM21 (P19474), which are listed as negative regulators of viral release from host cell, negative regulators of viral transcription and positive regulators of viral entry into host cells. Among these genes, DDX1 (log fold change of 0.36; *P*-value of 0.007 and *q*-value of 0.23) and TRIM21 (log fold change of 0.44; *P*-value of 0.003 and *q*-value of 0.18) are also upregulated in A549 cells infected with SARS-CoV-2 (55). Ten of the 11 viral proteins detected for fragment 29 are members of the Gag polyprotein family, that perform different tasks during HIV assembly, budding, and maturation. More than just scaffold elements, these proteins are elements that accompany viral and host proteins as they traffic to the cell membrane (Supplementary Table S2) (61). Finally, among the RBPs with the highest interaction propensity for fragment 23, we found nucleosome assembly protein 1-like 1 NAP1L1 and E3 ubiquitin-protein ligase makorin-1 MKRN1, which could have an effect on the regulation of cell proliferation.

Analysis of functional annotations carried out with *GeneMania* (29) revealed that proteins interacting with the 5′ of SARS-CoV-2 RNA are associated with regulatory pathways involving NOTCH2, MYC and MAX that have been previously connected to viral infection processes (Figure 5A) (62,63). Interestingly, some proteins, including DDX1, CCNT1 and ZNF175 for fragment 1 and TRIM32 for fragment 2, have been shown to be necessary for HIV functions and replication inside the cell, as well as SARS-CoV-1. DDX1 has been shown to enable the switch from discontinuous to continuous transcription in SARS-CoV-1 infection and its knockdown reduced the number of longer sub-genomic mRNA (sgmRNA) through interaction with the SARS-CoV-1 nucleocapsid protein N (64) and Nsp14 (65). It functions as a bidirectional helicase, which distinguishes it from the coronaviruses helicases, which can only unwind RNA in the 5′ to 3′ direction (66), a very important function in highly structured RNA such SARS-CoV-2. DDX1 is also required for HIV-1 Rev as well as for avian coronavirus IBV replication and it

11278 *Nucleic Acids Research, 2020, Vol. 48, No. 20*

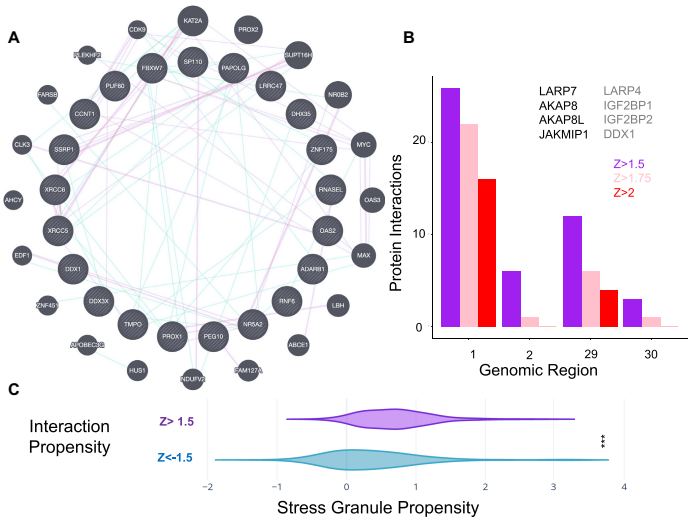


Figure 5. Characterization of protein interactions with SARS-CoV-2 RNA. (**A**) Protein interaction network of SARS-CoV-2 5′ end (inner circle) and associations with other human genes retrieved from literature (blue: genetic associations; purple: physical associations); (**B**) number of RBP interactions identified by Gordon *et al.* (71) and Schmidt *et al.* (76) for different SARS-CoV-2 regions. Representative cases are shown in black (Gordon *et al.* (71)) and gray (Schmidt *et al.* (76)). (**C**) Proteins binding to the 5′ with *Z* score ≥1.5 show high propensity to accumulate in stress-granules (same number of proteins with *Z* score <−1.5 are used in the comparison; *** *P*-value <0.0001; Kolmogorov–Smirnoff).

binds to the RRE sequence of HIV-1 RNAs (67,68), while CCNT1 binds to 7SK snRNA and regulates transactivation domain of the viral nuclear transcriptional activator, Tat (69,70).

Analyses of SARS-CoV-2 proteins interactomes reveal common protein targets

Recently, Gordon *et al.* reported a list of human proteins binding to Open Reading Frames (ORFs) translated from SARS-CoV-2 (71). Identified through affinity purification followed by mass spectrometry quantification, 332 proteins from HEK-293T cells interact with viral ORF peptides. By selecting 274 proteins binding at the 5′ with Z score ≥1.5 (Supplementary Table S2), of which 140 are exclusively interacting with fragment 1 (Figure 4B), we found that 8 are also reported in the list by Gordon *et al.* (71), which indicates significant enrichment (representation factor of 2.5; *P*-value of 0.02; hypergeometric test with human proteome in background). The fact that our list of protein-RNA binding partners contains elements identified also in the protein-protein network analysis is not surprising, as ribonucleoprotein complexes evolve together (14) and their components sustain each other through different types of interactions (16).

We note that out of 332 interactions, 60 are RBPs (as reported in Uniprot), which represents a considerable fraction (i.e. 20%), considering that there are around 1500 RBPs in the human proteome (i.e. 6%). Comparing the RBPs present in Gordon *et al.* (71) and those present in our list (79 RBP annotated in Uniprot), we found an overlap of six proteins (representation factor = 26.5; *P*-value $< 10^{-8}$; hypergeometric test), including: Janus kinase and microtubule-interacting protein 1 JAKMIP1 (Q96N16), A-kinase anchor protein 8 AKAP8 (O43823) and A-kinase anchor protein 8-like AKAP8L (Q9ULX6), which in case of HIV-1 infection is involved as a DEAD/H-box RNA helicase binding (72), signal recognition particle subunit SRP72 (O76094), binding to the 7S RNA in presence of SRP68, La-related protein 7, LARP7 (Q4G0J3) and La-related protein 4B LARP4B (Q92615), which are part of a system for transcriptional regulation acting by means of the 7SK RNP system (73) (Figure 5B; Supplementary Table S3). We speculate that sequestration of these elements is orchestrated by a viral program aiming to recruit host genes (74). LARP7 is also upregulated in A549 cells infected with SARS-CoV-2 (log fold change of 0.48; *P*-value of 0.006 and *q*-value of 0.23) (55).

Moreover, by directly analysing the RNA interaction potential of all the 332 proteins by Gordon *et al.* (71),

*cat*RAPID identified 38 putative binders at the 5′ end (*Z* score ≥ 1.5; 27 occurring exclusively in the 5′ end and not in other regions of the RNA) (18), including Serine/threonine-protein kinase TBK1 (Q9UHD2), among which 10 RBPs (as reported in Uniprot) such as: Splicing elements U3 small nucleolar ribonucleoprotein protein MPP10 (O00566) and Pre-mRNA-splicing factor SLU7 (O95391), snRNA methylphosphate capping enzyme MEPCE involved in negative regulation of transcription by RNA polymerase II 7SK (Q7L2J0) (75), Nucleolar protein 10 NOL10 (Q9BSC4) and protein kinase A Radixin RDX (P35241; in addition to those mentioned above; Supplementary Table S3).

Using the liver cell line HuH7 a recent experimental study by Schmidt *et al*. (76). identified SARS-CoV-2 RNA associations within the human host (76). Through the RAP-MS approach, 571 interactions were detected, of which 250 are RBPs (as reported in Uniprot) (76).

In common with our library we found an overlap of 148 proteins. We compared predicted (as released in March 2020) and experimentally-validated interactions employing balanced lists of high-affinity (high fold-change with respect to RNA Mitochondrial RNA Processing Endoribonuclease RMRP) and low-affinity (low fold-change with respect to RNA Mitochondrial RNA Processing Endoribonuclease RMRP) associations: a confidence score of 25% indicates that we compared the interaction scores of 35 proteins with the highest fold-change values and 35 interactions associated with the lowest fold-change values. From low (25%) to high (5%) confidence scores, we observed that the predictive power, measured as the AUC of ROC, increases monotonically reaching the remarkable value of 0.99 (the AUC is 0.72 for 25% confidence score; Supplementary Figure S4), which indicates strong agreement between predictions and experiments. In addition to DDX1 and DDX3X (O00571), other interactions corresponding to *cat*RAPID scores >1.5 and fold-change >1 include Insulin-like growth factor 2 mRNA-binding protein 1 IGF2BP1 (Q9Y6M1), Insulin-like growth factor 2 mRNA-binding protein 2 IGF2BP2 2 (Q9Y6M1) and La-related protein 4 LARP4 (Q71RC2; also in Gordon *et al*. (71)).

By directly analysing RNA interactions of all the 571 proteins by Schmidt *et al*. (76), *cat*RAPID identified 18 strong RBP binders at the 5′ end (Z score ≥ 1.5; fold-change >1; *P*-value of 0.008 computed with respect to all the interactions; Fisher exact test; Supplementary Table S4), including Helicase MOV-10 (Q9HCE1), Cold shock domain-containing protein E1 CSDE1 (O75534), Staphylococcal nuclease domain-containing protein 1 SND1 (Q7KZF4), Pumilio homolog 1 PUM1 (Q14671), and La-related protein 1 LARP1 (Q6PKG0), among other interactions (Supplementary Table S4).

The 5′ end is enriched in host interactions implicated in other viral infections

In the list of 274 proteins binding to the 5′ end (fragment 1) with *Z* score ≥1.5, we found 10 hits associated with HIV (Supplementary Table S5), which represents a significant enrichment (*P*-value = 0.0004; Fisher's exact test), considering that the total number of HIV-related proteins is 35 in the whole *cat*RAPID library (3340 elements). The complete list of proteins includes ATP-dependent RNA helicase DDX1 (Q92499), ATP-dependent RNA helicase DDX3X (O00571 also involved in Dengue and Zika Viruses), Tyrosine-protein kinase HCK (P08631, nucleotide binding), Arf-GAP domain and FG repeat-containing protein 1 (P52594), Double-stranded RNA-specific editase 1 ADARB1 (P78563), Insulin-like growth factor 2 mRNA-binding protein 1 IGF2BP1 (Q9NZI8), A-kinase anchor protein 8-like AKAP8L (Q9ULX6; its partner AKAP8 is also found in Gordon *et al.* (71)) Cyclin-T1 CCNT1 (O60563; DNA-binding) and Forkhead box protein K2 FOXK2 (Q01167; DNA-binding; Figures 4B and 5A; Supplementary Table S5).

Smaller enrichments were found for proteins related to Hepatitis B virus (HBV; *P*-value = 0.01; three hits out of seven in the whole *cat*RAPID library; Fisher's exact test), including Nuclear receptor subfamily 5 group A member 2 NR5A2 (DNA-binding; O00482), Interferon-induced, double-stranded RNA-activated protein kinase EIF2AK2 (P19525), and SRSF protein kinase 1 SRPK1 (Q96SB4) as well as Influenza A (*P*-value = 0.03; two hits out of four; Fisher's exact test), including Synaptic functional regulator FMR1 (Q06787) and RNA polymerase-associated protein RTF1 homologue (Q92541; Supplementary Table S5). By contrast, no significant enrichments were found for other viruses such as for instance Ebola.

Very importantly, specific chemical compounds have been developed to interact with HIV- and HVB-related proteins. The list of HIV-related targets reported in ChEMBL (77) includes ATP-dependent RNA helicase DDX1 (CHEMBL2011807), ATP-dependent RNA helicase DDX3X (CHEMBL2011808), Cyclin-T1 CCNT1 (CHEMBL2348842) and Tyrosine-protein kinase HCK (CHEMBL2408778), among other targets. In addition, HVB-related targets are Nuclear receptor subfamily 5 group A member 2 NR5A2 (CHEMBL3544), Interferon-induced, double-stranded RNA-activated protein kinase EIF2AK2 (CHEMBL5785) and SRSF protein kinase 1 SRPK1 (CHEMBL4375). We hope that this list can be the starting point for further pharmaceutical studies.

Phase-separating proteins are enriched in the 5′ end interactions

As SARS-CoV-2 represses host gene expression through a number of unknown mechanisms, sequestration of cell transcription machinery elements could be exploited to alter biological pathways in the host cell. A number of proteins identified in our *cat*RAPID calculations have been previously reported to coalesce in large ribonucleoprotein assemblies similar to stress granules. Among these proteins, we found double-stranded RNA-activated protein kinase EIF2AK2 (P19525), Nucleolin NCL (P19338), ATP-dependent RNA helicase DDX1 (Q92499), Cyclin-T1 CCNT1 (O60563), signal recognition particle subunit SRP72 (O76094), LARP7 (Q4G0J3) and La-related protein 4B LARP4B (Q92615) as well as Polypyrimidine tract-binding protein 1 PTBP1 (P26599) and Heterogeneous nuclear ribonucleoprotein Q HNRNPQ (O60506) (78). To further investigate the propensity of these proteins to phase

11280 *Nucleic Acids Research, 2020, Vol. 48, No. 20*

separate, we used the *cat*GRANULE algorithm (Material and Methods) (32). Differently from other methods to predict solid-like aggregation (79,80), *cat*GRANULE estimates the propensity of proteins to form liquid-like assemblies such as stress granules (81). We found that the 274 proteins binding to the 5′ end (fragment 1) with Z score $\geq$1.5 are highly prone to accumulate in assemblies similar to stress-granules (274 proteins with the lowest Z score are used in the comparison; P-value <0.0001; Kolmogorov−Smirnoff; Figure 5C; Supplementary Table S6). We note that there is not a direct correlation between RNA-binding scores (*cat*RAPID) and phase-separation propensities (*cat*GRANULE; Supplementary Figure S5).

Supporting this hypothesis, DDX1 and CCNT1 have been shown to condense in membrane-less organelles such as stress granules (82–84) that are the direct target of RNA viruses (85). DDX1 is also the primary component of distinct nuclear *foci* (86), together with factors associated with pre-mRNA processing and polyadenylation. Similarly, SRP72, LARP7 and LARP4B proteins have been found to assemble in stress granules (78,87,88). A recent work also suggests that the binding of LARP4 and XRCC6 takes places at the 5′ end of SARS-CoV-2 and contributes to SARS-CoV-2 phase separation (58). Moreover, emerging evidence indicates that the SARS-CoV-2 nucleocapsid protein N has a strong phase separation propensity that is modulated by the viral genome (58,89,90) and can enter into host cell protein condensates (89), suggesting a possible mechanism of cell protein sequestration. Notably, *cat*GRANULE does predict that nucleocapsid protein N is the viral protein with highest propensity to phase separate (91).

As is the case with molecular chaperones (92), RNAs can influence the liquid-like or solid-like state of proteins (93). This observation is particularly relevant because RNA viruses are known to antagonize stress granules formation (85). Stress granules and other phase-separated assemblies such as processing bodies regulate translation suppression and RNA decay, which could have a strong impact on virus replication (94).

DISCUSSION

Our study is motivated by the need to identify molecular mechanisms involved in Covid-19 spreading. Using advanced computational approaches, we investigated the structural content of SARS-CoV-2 genome and predicted human proteins that bind to it.

We employed *CROSS* (12,13) to compare the structural properties of ∼2000 coronaviruses and identified elements conserved in SARS-CoV-2 strains. The regions containing the highest amount of structure are the 5′ end as well as glycoproteins spike S and membrane M.

We found that the Spike S protein domain encompassing amino acids 460–520 is conserved across SARS-CoV-2 strains. This result suggests that Spike S must have evolved to specifically interact with its host partner ACE2 (48) and mutations increasing the binding affinity should be highly infrequent. As nucleic acids encoding for this region are enriched in double-stranded content, we speculate that the structure might attract host regulatory elements, such as nucleosome assembly protein 1-like 1 NAP1L1 and E3 ubiquitin-protein ligase makorin-1 MKRN1, further constraining its variability. The fact that this region of the Spike S region is highly conserved among all the analysed SARS-CoV-2 strains suggests that a specific drug could be designed to prevent interactions within the host.

The highly variable region at amino acids 243–302 in spike S protein corresponds to the binding site of sialic acids in MERS-CoV (7,10,51) and could play a role in infection (50). The fact that the binding region is highly variable suggests different affinities for sialic acid-containing oligosaccharides and polysaccharides such as heparan sulfate, which provides clues on the specific responses in the human population. At present, a glycan microarray technology indicated that SARS-CoV-2 Spike S binds more tightly to heparan sulfate than sialic acids (95).

Using *cat*RAPID (18,19) we computed >100 000 protein interactions with SARS-CoV-2 and found previously reported interactions such as Heterogeneous nuclear ribonucleoprotein Q HNRNPQ and Nucleolin NCL (59), among others. We discovered that the highly structured region at the 5′ end has the largest number of protein partners including ATP-dependent RNA helicase DDX1, which was previously reported to be essential for HIV-1 and coronavirus IBV replication (96,97), and the double-stranded RNA-specific editases ADAR and ADARB1, which catalyse the hydrolytic deamination of adenosine to inosine. Other predicted interactions are XRCC5 and XRCC6 members of the HDP-RNP complex associating with ATP-dependent RNA helicase DHX9 (98) as well as and 2–5A-dependent ribonuclease RNASEL and 2–5-oligoadenylate synthase 2 OAS2 that control viral RNA degradation (99,100). Interestingly, DDX1, XRCC6 and OAS2 were found upregulated in human alveolar adenocarcinoma cells infected with SARS-CoV-2 (55) and DDX1 knockdown has been shown to reduce the number of sgmRNA in SARS-CoV-1 infected cells (64). In agreement with our predictions, recent experimental work indicates that the family of ADAR deaminases is active in bronchoalveolar lavage fluids derived from SARS-CoV-2 patients (57).

Comparison with protein-RNA interactions detected in the liver cell line HuH7 (76) shows agreement with our predictions. We note that the experiments have been carried out 24 h after infection (76) and the protein interaction landscape might have changed with respect to the early events of replication. Yet, the accordance with our calculations indicates participation of elements involved in controlling RNA processing and editing (DDX1, DDX3X) and translation (IGF2BP1 and IGF2BP2), although proteins such as ADAR and XRCC5 were reported to have poorer binding capacity (76).

A significant overlap exists with the list of protein interactions reported by Gordon *et al.* (71) and among the candidate partners we identified AKAP8L, involved as a DEAD/H-box RNA helicase binding protein involved in HIV infection (72). In general, proteins associated with retroviral replication are expected to play different roles in SARS-CoV-2. As SARS-CoV-2 massively represses host gene expression (74), we hypothesize that the virus hijacks host pathways by recruiting transcriptional and post-transcriptional elements interacting with polymerase II

genes and splicing factors such as for instance A-kinase anchor protein 8-like AKAP8L and La-related protein 7 LARP7. In concordance with our predictions LARP7 has been reported to be upregulated in human alveolar adenocarcinoma cells infected with SARS-CoV-2 (55). The link to proteins previously studied in the context of HIV and other viruses, if further confirmed, is particularly relevant for the repurposing of existing drugs (77).

The idea that SARS-CoV-2 sequesters different elements of the transcriptional machinery is particularly intriguing and is supported by the fact that a large number of proteins identified in our screening are found in stress granules (78). Indeed, stress granules protect the host innate immunity and are hijacked by viruses to favour their own replication (94). As coronaviruses transcription uses discontinuous RNA synthesis that involves high-frequency recombination (59), it is possible that pieces of the viruses resulting from a mechanism called defective interfering RNAs (101) could act as scaffold to attract host proteins (14,15). In agreement with our hypothesis, it has been very recently shown that the coronavirus nucleocapsid protein N can form protein condensates based on viral RNA scaffold and can merge with the human cell protein condensates (89), which provides a potential mechanism of host protein sequestration.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

The authors would like to thank Dr Mattia Miotto, Dr Lorenzo Di Rienzo, Dr Alexandros Armaos, Dr Alessandro Dasti, Dr Claudia Giambartolomei for discussions. We are particularly grateful to Prof. Annalisa Pastore for critical reading, Dr Gilles Mirambeau for the RT versus RdRP analysis, Dr Andrea Cerase for the discussing on stress granules and Dr Roberto Giambruno for pointing to PTBP1 and HNRNPQ experiments. We are very much in debt with Dr Tommaso Muto, Giampaolo Fiore for their advice and friendship.

Authors contributions: G.G.T. and R.D.P. conceived the study. A.V. and A.A. carried out *cat*RAPID analysis of protein interactions, R.D.P. calculated *CROSS* structures of coronaviruses, G.G.T., M.M. and E.M. performed and analysed sequence alignments, J.R., E.Z. and E.B. analysed the prediction results. A.V., R.D.P. and G.G.T. wrote the paper.

FUNDING

European Research Council [RIBOMYLOME_309545, ASTRA_855923]; H2020 projects [IASIS_727658 and INFORE_825080]; Spanish Ministry of Economy and Competitiveness [BFU2017-86970-P]; collaboration with Peter St. George-Hyslop financed by the Wellcome Trust. Funding for open access charge: ERC [ASTRA 855923].

Conflict of interest statement. None declared.

REFERENCES

1. Zhu,N., Zhang,D., Wang,W., Li,X., Yang,B., Song,J., Zhao,X., Huang,B., Shi,W., Lu,R. *et al.* (2020) A novel coronavirus from patients with pneumonia in China, 2019. *N. Engl. J. Med.*, **382**, 727–733.
2. D'Antiga,L. (2020) Coronaviruses and immunosuppressed patients. The facts during the third epidemic. *Liver Transpl.*, **26**, 832–834.
3. Cascella,M., Rajnik,M., Cuomo,A., Dulebohn,S.C. and Di Napoli,R. (2020) Features, evaluation and treatment coronavirus (COVID-19). In: *StatPearls*. StatPearls Publishing, Treasure Island.
4. Ge,X.-Y., Li,J.-L., Yang,X.-L., Chmura,A.A., Zhu,G., Epstein,J.H., Mazet,J.K., Hu,B., Zhang,W., Peng,C. *et al.* (2013) Isolation and characterization of a bat SARS-like coronavirus that uses the ACE2 receptor. *Nature*, **503**, 535–538.
5. Follis,K.E., York,J. and Nunberg,J.H. (2006) Furin cleavage of the SARS coronavirus spike glycoprotein enhances cell-cell fusion but does not affect virion entry. *Virology*, **350**, 358–369.
6. Xiao,K., Zhai,J., Feng,Y., Zhou,N., Zhang,X., Zou,J.-J., Li,N., Guo,Y., Li,X., Shen,X. *et al.* (2020) Isolation of SARS-CoV-2-related coronavirus from Malayan pangolins. *Nature*, **583**, 286–289.
7. Park,Y.-J., Walls,A.C., Wang,Z., Sauer,M.M., Li,W., Tortorici,M.A., Bosch,B.-J., DiMaio,F. and Veesler,D. (2019) Structures of MERS-CoV spike glycoprotein in complex with sialoside attachment receptors. *Nat. Struct. Mol. Biol.*, **26**, 1151–1157.
8. Walls,A.C., Tortorici,M.A., Bosch,B.-J., Frenz,B., Rottier,P.J.M., DiMaio,F., Rey,F.A. and Veesler,D. (2016) Cryo-electron microscopy structure of a coronavirus spike glycoprotein trimer. *Nature*, **531**, 114–117.
9. Ou,X., Liu,Y., Lei,X., Li,P., Mi,D., Ren,L., Guo,L., Guo,R., Chen,T., Hu,J. *et al.* (2020) Characterization of spike glycoprotein of SARS-CoV-2 on virus entry and its immune cross-reactivity with SARS-CoV. *Nat. Commun.*, **11**, 1620.
10. Li,W., Hulswit,R.J.G., Widjaja,I., Raj,V.S., McBride,R., Peng,W., Widagdo,W., Tortorici,M.A., van Dieren,B., Lang,Y. *et al.* (2017) Identification of sialic acid-binding function for the Middle East respiratory syndrome coronavirus spike glycoprotein. *Proc. Natl. Acad. Sci. U.S.A.*, **114**, E8508–E8517.
11. Yang,D. and Leibowitz,J.L. (2015) The structure and functions of coronavirus genomic 3′ and 5′ ends. *Virus Res*, **206**, 120–133.
12. Delli Ponti,R., Marti,S., Armaos,A. and Tartaglia,G.G. (2017) A high-throughput approach to profile RNA structure. *Nucleic Acids Res.*, **45**, e35.
13. Delli Ponti,R., Armaos,A., Marti,S. and Tartaglia,G.G. (2018) A method for RNA structure prediction shows evidence for structure in lncRNAs. *Front. Mol. Biosci.*, **5**, 111.
14. Sanchez de Groot,N., Armaos,A., Graña-Montes,R., Alriquet,M., Calloni,G., Vabulas,R.M. and Tartaglia,G.G. (2019) RNA structure drives interaction with proteins. *Nat. Commun.*, **10**, 3246.
15. Cid-Samper,F., Gelabert-Baldrich,M., Lang,B., Lorenzo-Gotor,N., Ponti,R.D., Severijnen,L.-A.W.F.M., Bolognesi,B., Gelpi,E., Hukema,R.K., Botta-Orfila,T. *et al.* (2018) An integrative study of protein-RNA condensates identifies scaffolding RNAs and reveals players in fragile X-Associated Tremor/Ataxia syndrome. *Cell Rep.*, **25**, 3422–3434.
16. Cerase,A., Armaos,A., Neumayer,C., Avner,P., Guttman,M. and Tartaglia,G.G. (2019) Phase separation drives X-chromosome inactivation: a hypothesis. *Nat. Struct. Mol. Biol.*, **26**, 331.
17. Moreno,J.L., Zúñiga,S., Enjuanes,L. and Sola,I. (2008) Identification of a coronavirus transcription enhancer. *J. Virol.*, **82**, 3882–3893.
18. Agostini,F., Zanzoni,A., Klus,P., Marchese,D., Cirillo,D. and Tartaglia,G.G. (2013) catRAPID omics: a web server for large-scale prediction of protein-RNA interactions. *Bioinformatics*, **29**, 2928–2930.
19. Cirillo,D., Blanco,M., Armaos,A., Buness,A., Avner,P., Guttman,M., Cerase,A. and Tartaglia,G.G. (2017) Quantitative predictions of protein interactions with long noncoding RNAs. *Nat Meth*, **14**, 5–6.
20. Bellucci,M., Agostini,F., Masin,M. and Tartaglia,G.G. (2011) Predicting protein associations with long noncoding RNAs. *Nat. Methods*, **8**, 444–445.
21. Lang,B., Armaos,A. and Tartaglia,G.G. (2019) RNAct: Protein–RNA interaction predictions for model organisms with supporting experimental data. *Nucleic Acids Res.*, **47**, D601–D606.

22. Kliger,Y. and Levanon,E.Y. (2003) Cloaked similarity between HIV-1 and SARS-CoV suggests an anti-SARS strategy. *BMC Microbiol*, **3**, 20.
23. Hallenberger,S., Bosch,V., Angliker,H., Shaw,E., Klenk,H.D. and Garten,W. (1992) Inhibition of furin-mediated cleavage activation of HIV-1 glycoprotein gp160. *Nature*, **360**, 358–361.
24. Glowacka,I., Bertram,S., Herzog,P., Pfefferle,S., Steffen,I., Muench,M.O., Simmons,G., Hofmann,H., Kuri,T., Weber,F. *et al.* (2010) Differential downregulation of ACE2 by the spike proteins of severe acute respiratory syndrome coronavirus and human coronavirus NL63. *J. Virol.*, **84**, 1198–1205.
25. Lorenz,R., Bernhart,S.H., Höner zu Siederdissen,C., Tafer,H., Flamm,C., Stadler,P.F. and Hofacker,I.L. (2011) ViennaRNA Package 2.0. *Algorith. Mol. Biol.*, **6**, 26.
26. Ponti,R.D., Armaos,A., Vandelli,A. and Tartaglia,G.G. (2020) CROSSalive: a web server for predicting the in vivo structure of RNA molecules. *Bioinformatics*, **36**, 940–941.
27. Nawrocki,E.P. and Eddy,S.R. (2013) Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics*, **29**, 2933–2935.
28. Klus,P., Ponti,R.D., Livi,C.M. and Tartaglia,G.G. (2015) Protein aggregation, structural disorder and RNA-binding ability: a new approach for physico-chemical and gene ontology classification of multiple datasets. *BMC Genomics*, **16**, 1071.
29. Warde-Farley,D., Donaldson,S.L., Comes,O., Zuberi,K., Badrawi,R., Chao,P., Franz,M., Grouios,C., Kazi,F., Lopes,C.T. *et al.* (2010) The GeneMANIA prediction server: biological network integration for gene prioritization and predicting gene function. *Nucleic Acids Res.*, **38**, W214–W220.
30. Madeira,F., Park,Y. mi, Lee,J., Buso,N., Gur,T., Madhusoodanan,N., Basutkar,P., Tivey,A.R.N., Potter,S.C., Finn,R.D. *et al.* (2019) The EMBL-EBI search and sequence analysis tools APIs in 2019. *Nucleic Acids Res.*, **47**, W636–W641.
31. Di Tommaso,P., Moretti,S., Xenarios,I., Orobitg,M., Montanyola,A., Chang,J.-M., Taly,J.-F. and Notredame,C. (2011) T-Coffee: a web server for the multiple sequence alignment of protein and RNA sequences using structural information and homology extension. *Nucleic Acids Res.*, **39**, W13–W17.
32. Bolognesi,B., Lorenzo Gotor,N., Dhar,R., Cirillo,D., Baldrighi,M., Tartaglia,G.G. and Lehner,B. (2016) A Concentration-Dependent liquid phase separation can cause toxicity upon increased protein expression. *Cell Rep.*, **16**, 222–231.
33. Gultyaev,A.P., Richard,M., Spronken,M.I., Olsthoorn,R.C.L. and Fouchier,R.A.M. (2019) Conserved structural RNA domains in regions coding for cleavage site motifs in hemagglutinin genes of influenza viruses. *Virus Evol.*, **5**, vez034.
34. Wu,A., Peng,Y., Huang,B., Ding,X., Wang,X., Niu,P., Meng,J., Zhu,Z., Zhang,Z., Wang,J. *et al.* (2020) Genome composition and divergence of the novel coronavirus (2019-nCoV) Originating in China. *Cell Host Microbe*, **27**, 325–328.
35. Kelly,J.A., Olson,A.N., Neupane,K., Munshi,S., San Emeterio,J., Pollack,L., Woodside,M.T. and Dinman,J.D. (2020) Structural and functional conservation of the programmed -1 ribosomal frameshift signal of SARS coronavirus 2 (SARS-CoV-2). *J. Biol. Chem.*, **295**, 10741–10748.
36. Incarnato,D., Morandi,E., Simon,L.M. and Oliviero,S. (2018) RNA Framework: an all-in-one toolkit for the analysis of RNA structures and post-transcriptional modifications. *Nucleic Acids Res.*, **46**, e97.
37. Williams,G.D., Chang,R.-Y. and Brian,D.A. (1999) A phylogenetically conserved hairpin-type 3′ untranslated region pseudoknot functions in coronavirus RNA replication. *J Virol*, **73**, 8349–8355.
38. Manfredonia,I., Nithin,C., Ponce-Salvatierra,A., Ghosh,P., Wirecki,T.K., Marinus,T., Ogando,N.S., Snider,E.J., Hemert,M.J. van, Bujnicki,J.M. and Incarnato,D. (2020) Genome-wide mapping of therapeutically-relevant SARS-CoV-2 RNA structures. bioRxiv doi: https://doi.org/10.1101/2020.06.15.151647, 15 June 2020, preprint: not peer reviewed.
39. Goebel,S.J., Taylor,J. and Masters,P.S. (2004) The 3′ cis-acting genomic replication element of the severe acute respiratory syndrome coronavirus can function in the murine coronavirus genome. *J. Virol.*, **78**, 7846–7851.
40. Yang,D. and Leibowitz,J.L. (2015) The structure and functions of coronavirus genomic 3′ and 5′ ends. *Virus Res.*, **206**, 120–133.
41. Sola,I., Mateos-Gomez,P.A., Almazan,F., Zuñiga,S. and Enjuanes,L. (2011) RNA-RNA and RNA-protein interactions in coronavirus replication and transcription. *RNA Biol*, **8**, 237–248.
42. Madhugiri,R., Fricke,M., Marz,M. and Ziebuhr,J. (2016) Coronavirus cis-Acting RNA elements. *Adv. Virus Res.*, **96**, 127–163.
43. Agostini,F., Cirillo,D., Bolognesi,B. and Tartaglia,G.G. (2013) X-inactivation: quantitative predictions of protein interactions in the Xist networkitle. *Nucleic Acids Res.*, **41**, e31.
44. Li,W. and Godzik,A. (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, **22**, 1658–1659.
45. Yang,D. and Leibowitz,J.L. (2015) The structure and functions of coronavirus genomic 3′ and 5′ ends. *Virus Res*, **206**, 120–133.
46. Hrdlickova,R., Toloue,M. and Tian,B. (2017) RNA-Seq methods for transcriptome analysis. *Wiley Interdiscip. Rev. RNA*, **8**, e1364.
47. Ozsolak,F. and Milos,P.M. (2011) RNA sequencing: advances, challenges and opportunities. *Nat Rev Genet*, **12**, 87–98.
48. Andersen,K.G., Rambaut,A., Lipkin,W.I., Holmes,E.C. and Garry,R.F. (2020) The proximal origin of SARS-CoV-2. *Nat. Med.*, **26**, 450–452.
49. Zhou,P., Yang,X.-L., Wang,X.-G., Hu,B., Zhang,L., Zhang,W., Si,H.-R., Zhu,Y., Li,B., Huang,C.-L. *et al.* (2020) A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature*, **579**, 270–273.
50. Qing,E., Hantak,M., Perlman,S. and Gallagher,T. (2020) Distinct roles for sialoside and protein receptors in coronavirus infection. *mBio*, **11**, e02764–19.
51. Milanetti,E., Miotto,M., Di Rienzo,L., Monti,M., Gosti,G. and Ruocco,G. (2020) In-Silico evidence for two receptors based strategy of SARS-CoV-2. bioRxiv doi: https://doi.org/10.1101/2020.03.24.006197, 27 March 2020, preprint: not peer reviewed.
52. Galán,C., Sola,I., Nogales,A., Thomas,B., Akoulitchev,A., Enjuanes,L. and Almazán,F. (2009) Host cell proteins interacting with the 3′ end of TGEV coronavirus genome influence virus replication. *Virology*, **391**, 304–314.
53. Lu,K., Heng,X. and Summers,M.F. (2011) Structural determinants and mechanism of HIV-1 genome packaging. *J. Mol. Biol.*, **410**, 609–633.
54. Fehr,A.R. and Perlman,S. (2015) Coronaviruses: an overview of their replication and pathogenesis. *Methods Mol Biol*, **1282**, 1–23.
55. Blanco-Melo,D., Nilsson-Payant,B.E., Liu,W.-C., Uhl,S., Hoagland,D., Møller,R., Jordan,T.X., Oishi,K., Panis,M., Sachs,D. *et al.* (2020) Imbalanced host response to SARS-CoV-2 drives development of COVID-19. *Cell*, **181**, 1036–1045.
56. Armaos,A., Cirillo,D., Tartaglia and,G.G. (2017) omiXcore: a web server for prediction of protein interactions with large RNA. *Bioinformatics*, **33**, 3104–3106.
57. Giorgio,S.D., Martignano,F., Torcia,M.G., Mattiuz,G. and Conticello,S.G. (2020) Evidence for host-dependent RNA editing in the transcriptome of SARS-CoV-2. *Science Advances*, **6**, eabb5813.
58. Iserman,C., Roden,C., Boerneke,M., Sealfon,R., McLaughlin,G., Jungreis,I., Park,C., Boppana,A., Fritch,E., Hou,Y.J. *et al.* (2020) Specific viral RNA drives the SARS CoV-2 nucleocapsid to phase separate. bioRxiv doi: https://doi.org/10.1101/2020.06.11.147199, 12 June 2020, preprint: not peer reviewed.
59. Sola,I., Mateos-Gomez,P.A., Almazan,F., Zuñiga,S. and Enjuanes,L. (2011) RNA-RNA and RNA-protein interactions in coronavirus replication and transcription. *RNA Biol*, **8**, 237–248.
60. Castello,A., Fischer,B., Eichelbaum,K., Horos,R., Beckmann,B.M., Strein,C., Davey,N.E., Humphreys,D.T., Preiss,T., Steinmetz,L.M. *et al.* (2012) Insights into RNA biology from an atlas of mammalian mRNA-binding proteins. *Cell*, **149**, 1393–1406.
61. Bell,N.M. and Lever,A.M.L. (2013) HIV Gag polyprotein: processing and early viral particle assembly. *Trends Microbiol.*, **21**, 136–144.
62. Hayward,S.D. (2004) Viral interactions with the Notch pathway. *Semin. Cancer Biol.*, **14**, 387–396.
63. Dudley,J.P., Mertz,J.A., Rajan,L., Lozano,M. and Broussard,D.R. (2002) What retroviruses teach us about the involvement of c- Myc in leukemias and lymphomas. *Leukemia*, **16**, 1086–1098.
64. Wu,C.-H., Chen,P.-J. and Yeh,S.-H. (2014) Nucleocapsid phosphorylation and RNA helicase DDX1 recruitment enables

coronavirus transition from discontinuous to continuous transcription. *Cell Host Microbe*, **16**, 462–472.
65. Xu,L., Khadijah,S., Fang,S., Wang,L., Tay,F.P.L. and Liu,D.X. (2010) The cellular RNA helicase DDX1 interacts with coronavirus nonstructural protein 14 and enhances viral replication. *J. Virol.*, **84**, 8571–8583.
66. Shu,T., Huang,M., Wu,D., Ren,Y., Zhang,X., Han,Y., Mu,J., Wang,R., Qiu,Y., Zhang,D.-Y. *et al.* (2020) SARS-Coronavirus-2 Nsp13 possesses NTPase and RNA helicase activities that can be inhibited by bismuth salts. *Virol. Sin.*, **35**, 321–329.
67. Edgcomb,S.P., Carmel,A.B., Naji,S., Ambrus-Aikelin,G., Reyes,J.R., Saphire,A.C.S., Gerace,L. and Williamson,J.R. (2012) DDX1 is an RNA-dependent ATPase involved in HIV-1 rev function and virus replication. *J Mol Biol*, **415**, 61–74.
68. Xu,L., Khadijah,S., Fang,S., Wang,L., Tay,F.P.L. and Liu,D.X. (2010) The cellular RNA helicase DDX1 interacts with coronavirus nonstructural protein 14 and enhances viral replication. *J. Virol.*, **84**, 8571–8583.
69. Ivanov,D., Kwak,Y.T., Nee,E., Guo,J., García-Martínez,L.F. and Gaynor,R.B. (1999) Cyclin T1 domains involved in complex formation with Tat and TAR RNA are critical for tat-activation. *J. Mol. Biol.*, **288**, 41–56.
70. Kwak,Y.T., Ivanov,D., Guo,J., Nee,E. and Gaynor,R.B. (1999) Role of the human and murine cyclin T proteins in regulating HIV-1 tat-activation. *J. Mol. Biol.*, **288**, 57–69.
71. Gordon,D.E., Jang,G.M., Bouhaddou,M., Xu,J., Obernier,K., White,K.M., O'Meara,M.J., Rezelj,V.V., Guo,J.Z., Swaney,D.L. *et al.* (2020) A SARS-CoV-2 protein interaction map reveals targets for drug repurposing. *Nature*, **583**, 459–468.
72. Xing,L., Zhao,X., Guo,F. and Kleiman,L. (2014) The role of A-kinase anchoring protein 95-like protein in annealing of tRNALys3 to HIV-1 RNA. *Retrovirology*, **11**, 58.
73. Markert,A., Grimm,M., Martinez,J., Wiesner,J., Meyerhans,A., Meyuhas,O., Sickmann,A. and Fischer,U. (2008) The La-related protein LARP7 is a component of the 7SK ribonucleoprotein and affects transcription of cellular and viral polymerase II genes. *EMBO Rep.*, **9**, 569–575.
74. Kim,D., Lee,J.-Y., Yang,J.-S., Kim,J.W., Kim,V.N. and Chang,H. (2020) The architecture of SARS-CoV-2 transcriptome. *Cell*, **181**, 914–921.
75. Jeronimo,C., Forget,D., Bouchard,A., Li,Q., Chua,G., Poitras,C., Thérien,C., Bergeron,D., Bourassa,S., Greenblatt,J. *et al.* (2007) Systematic analysis of the protein interaction network for the human transcription machinery reveals the identity of the 7SK capping enzyme. *Mol. Cell*, **27**, 262–274.
76. Schmidt,N., Lareau,C.A., Keshishian,H., Melanson,R., Zimmer,M., Kirschner,L., Ade,J., Werner,S., Caliskan,N., Lander,E.S. *et al.* (2020) A direct RNA-protein interaction atlas of the SARS-CoV-2 RNA in infected human cells. bioRxiv doi: https://doi.org/10.1101/2020.07.15.204404, 15 July 2020, preprint: not peer reviewed.
77. Mendez,D., Gaulton,A., Bento,A.P., Chambers,J., De Veij,M., Félix,E., Magariños,M.P., Mosquera,J.F., Mutowo,P., Nowotka,M. *et al.* (2019) ChEMBL: towards direct deposition of bioassay data. *Nucleic Acids Res.*, **47**, D930–D940.
78. Markmiller,S., Soltanieh,S., Server,K.L., Mak,R., Jin,W., Fang,M.Y., Luo,E.-C., Krach,F., Yang,D., Sen,A. *et al.* (2018) Context-dependent and disease-specific diversity in protein interactions within stress granules. *Cell*, **172**, 590–604.
79. Lee,Y., Zhou,T., Tartaglia,G.G., Vendruscolo,M. and Wilke,C.O. (2010) Translationally optimal codons associate with aggregation-prone sites in proteins. *Proteomics*, **10**, 4163–4171.
80. Agostini,F., Cirillo,D., Livi,C.M., Ponti,R.D. and Tartaglia,G.G. (2014) ccSOL omics: a webserver for large-scale prediction of endogenous and heterologous solubility in E. coli. *Bioinformatics*, **30**, 2975–2977.
81. Gotor,N.L., Armaos,A., Calloni,G., Torrent Burgas,M., Vabulas,R.M., De Groot,N.S. and Tartaglia,G.G. (2020) RNA-binding and prion domains: the Yin and Yang of phase separation. *Nucleic Acids Res.*, **48**, 9491–9504.
82. Ning,W., Guo,Y., Lin,S., Mei,B., Wu,Y., Jiang,P., Tan,X., Zhang,W., Chen,G., Peng,D. *et al.* (2020) DrLLPS: a data resource of liquid–liquid phase separation in eukaryotes. *Nucleic Acids Res.*, **48**, D288–D295.
83. Shin,Y., Chang,Y.-C., Lee,D.S.W., Berry,J., Sanders,D.W., Ronceray,P., Wingreen,N.S., Haataja,M. and Brangwynne,C.P. (2018) Liquid nuclear condensates mechanically sense and restructure the genome. *Cell*, **175**, 1481–1491.
84. Su,X., Ditlev,J.A., Hui,E., Xing,W., Banjade,S., Okrut,J., King,D.S., Taunton,J., Rosen,M.K. and Vale,R.D. (2016) Phase separation of signaling molecules promotes T cell receptor signal transduction. *Science*, **352**, 595–599.
85. White,J.P. and Lloyd,R.E. (2012) Regulation of stress granules in virus systems. *Trends Microbiol*, **20**, 175–183.
86. Uversky,V.N. (2017) Intrinsically disordered proteins in overcrowded milieu: membrane-less organelles, phase separation, and intrinsic disorder. *Curr. Opin. Struct. Biol.*, **44**, 18–30.
87. Schäffler,K., Schulz,K., Hirmer,A., Wiesner,J., Grimm,M., Sickmann,A. and Fischer,U. (2010) A stimulatory role for the La-related protein 4B in translation. *RNA*, **16**, 1488–1499.
88. Küspert,M., Murakawa,Y., Schäffler,K., Vanselow,J.T., Wolf,E., Juranek,S., Schlosser,A., Landthaler,M. and Fischer,U. (2015) LARP4B is an AU-rich sequence associated factor that promotes mRNA accumulation and translation. *RNA*, **21**, 1294–1305.
89. Perdikari,T.M., Murthy,A.C., Ryan,V.H., Watters,S., Naik,M.T. and Fawzi,N.L. (2020) SARS-CoV-2 nucleocapsid protein undergoes liquid-liquid phase separation stimulated by RNA and partitions into phases of human ribonucleoproteins. bioRxiv doi: https://doi.org/10.1101/2020.06.09.141101, 10 June 2020, preprint: not peer reviewed.
90. Cubuk,J., Alston,J.J., Incicco,J.J., Singh,S., Stuchell-Brereton,M.D., Ward,M.D., Zimmerman,M.I., Vithani,N., Griffith,D., Wagoner,J.A. *et al.* (2020) The SARS-CoV-2 nucleocapsid protein is dynamic, disordered, and phase separates with RNA. bioRxiv doi: https://doi.org/10.1101/2020.06.17.158121, 18 June 2020, preprint: not peer reviewed.
91. Cascarina,S.M. and Ross,E.D. (2020) A proposed role for the SARS-CoV-2 nucleocapsid protein in the formation and regulation of biomolecular condensates. *FASEB J.*, doi:10.1096/fj.202001351.
92. Tartaglia,G.G., Dobson,C.M., Hartl,F.U. and Vendruscolo,M. (2010) Physicochemical determinants of chaperone requirements. *J. Mol. Biol.*, **400**, 579–588.
93. Zacco,E., Graña-Montes,R., Martin,S.R., de Groot,N.S., Alfano,C., Tartaglia,G.G. and Pastore,A. (2019) RNA as a key factor in driving or preventing self-assembly of the TAR DNA-binding protein 43. *J. Mol. Biol.*, **431**, 1671–1688.
94. Zhang,Q., Sharma,N.R., Zheng,Z.-M. and Chen,M. (2019) Viral regulation of RNA granules in infected cells. *Virol. Sin.*, **34**, 175–191.
95. Hao,W., Ma,B., Li,Z., Wang,X., Gao,X., Li,Y., Qin,B., Shang,S., Cui,S. and Tan,Z. (2020) Binding of the SARS-CoV-2 Spike Protein to Glycans. bioRxiv doi: https://doi.org/10.1101/2020.05.17.100537, 17 May 2020, preprint: not peer reviewed.
96. Fang,J., Kubota,S., Yang,B., Zhou,N., Zhang,H., Godbout,R. and Pomerantz,R.J. (2004) A DEAD box protein facilitates HIV-1 replication as a cellular co-factor of Rev. *Virology*, **330**, 471–480.
97. Xu,L., Khadijah,S., Fang,S., Wang,L., Tay,F.P.L. and Liu,D.X. (2010) The cellular RNA helicase DDX1 interacts with coronavirus nonstructural protein 14 and enhances viral replication. *J. Virol.*, **84**, 8571–8583.
98. Zhang,S., Schlott,B., Görlach,M. and Grosse,F. (2004) DNA-dependent protein kinase (DNA-PK) phosphorylates nuclear DNA helicase II/RNA helicase A and hnRNP proteins in an RNA-dependent manner. *Nucleic Acids Res.*, **32**, 1–10.
99. Sarkar,S.N., Ghosh,A., Wang,H.W., Sung,S.S. and Sen,G.C. (1999) The nature of the catalytic domain of 2′-5′-oligoadenylate synthetases. *J. Biol. Chem.*, **274**, 25535–25542.
100. Sarkar,S.N., Bandyopadhyay,S., Ghosh,A. and Sen,G.C. (1999) Enzymatic characteristics of recombinant medium isozyme of 2′-5′ oligoadenylate synthetase. *J. Biol. Chem.*, **274**, 1848–1855.
101. Pathak,K.B. and Nagy,P.D. (2009) Defective interfering RNAs: foes of viruses and friends of virologists. *Viruses*, **1**, 895–919.

CHAPTER 7

PHASE SEPARATION DRIVES SARS-COV-2 REPLICATION: A HYPOTHESIS

Phase separation drives SARS-CoV-2 replication: a hypothesis

In recent years, many experimental works have been published with the aim of capturing as much information as possible on the SARS-CoV-2 interactivity with the human host and were carried out in different cell types following various purification techniques (Gordon et al., 2020; Flynn et al., 2021; Schmidt et al., 2021; Kamel et al., 2021; Lee et al., 2021).
However, while this collective effort led to the identification and confirmation of several important players, we noticed that only a small overlap of interactors was actually found across all experiments, while the majority of them were either unique to one study or shared by a few of them.

Following this thread, in this work we compare four interactome experiments performed in different conditions and identify twenty-one proteins shared by all of them. These elements are predicted to bind preferentially to the 5' of the viral genome and contain proteins involved in stress granule formation, pre-mRNA regulators and factors involved in the replication process of SARS-CoV-2 and other viral species. Furthermore, this group of interactors also shows the highest binding propensity to the viral genome. This indicates that strong-affinity binders are highly reproducible and therefore more likely to be found in a greater number of experiments. In addition, the enrichment of proteins belonging to phase-separating condensates constitutes further evidence of SARS-CoV-2 sequestration of these elements to antagonize the cell's defenses.

This work was published in Frontiers in Molecular Biosciences journal in

2022.

As a co-first author, I contributed to the work by conceiving the study and computing the predictions of binding strength and phase-separating propensity of the different protein datasets.

Vandelli, A, Vocino, G., and Tartaglia, G. G. (2022). Phase Separation Drives SARS-CoV-2 Replication: A Hypothesis. *Frontiers in Molecular Biosciences*, 9:893067.
DOI: 10.3389/fmolb.2022.893067

frontiers | Frontiers in Molecular Biosciences

PERSPECTIVE
published: 11 May 2022
doi: 10.3389/fmolb.2022.893067

Phase Separation Drives SARS-CoV-2 Replication: A Hypothesis

Andrea Vandelli[1,2†], *Giovanni Vocino*[3†] *and Gian Gaetano Tartaglia*[4,5,6*]

[1]Department of Biochemistry and Molecular Biology, Universitat Autònoma de Barcelona, Barcelona, Spain, [2]Universitat Pompeu Fabra (UPF), Barcelona, Spain, [3]Department of Pharmacy and Biotechnology, University of Bologna, Bologna, Italy, [4]Center for Human Technologies, Istituto Italiano di Tecnologia, Genova, Italy, [5]Department of Biology 'Charles Darwin', Sapienza University of Rome, Rome, Italy, [6]Institucio Catalana de Recerca i Estudis Avançats (ICREA), Barcelona, Spain

Identifying human proteins that interact with SARS-CoV-2 genome is important to understand its replication and to identify therapeutic strategies. Recent studies have unveiled protein interactions of SARS-COV-2 in different cell lines and through a number of high-throughput approaches. Here, we carried out a comparative analysis of four experimental and one computational studies to characterize the interactions of SARS-CoV-2 genomic RNA. Although hundreds of interactors have been identified, only twenty-one appear in all the experiments and show a strong propensity to bind. This set of interactors includes stress granule forming proteins, pre-mRNA regulators and elements involved in the replication process. Our calculations indicate that DDX3X and several editases bind the 5′ end of SARS-CoV-2, a regulatory region previously reported to attract a large number of proteins. The small overlap among experimental datasets suggests that SARS-CoV-2 genome establishes stable interactions only with few interactors, while many proteins bind less tightly. In analogy to what has been previously reported for *Xist* non-coding RNA, we propose a mechanism of phase separation through which SARS-CoV-2 progressively sequesters human proteins hijacking the host immune response.

Keywords: viral RNA, phase separation, stress granules, protein-RNA interactions, RNA-binding proteins

OPEN ACCESS

Edited by:
Barbara Bardoni,
UMR7275 Institut de Pharmacologie Moléculaire et Cellulaire (IPMC), France

Reviewed by:
Venu Raman,
Johns Hopkins Medicine, United States
Lorena Zubovic,
University of Trento, Italy

***Correspondence:**
Gian Gaetano Tartaglia
gian.tartaglia@iit.it

†These authors share first authorship

Specialty section:
This article was submitted to RNA Networks and Biology, a section of the journal Frontiers in Molecular Biosciences

Received: *09 March 2022*
Accepted: *25 April 2022*
Published: *11 May 2022*

Citation:
Vandelli A, Vocino G and Tartaglia GG (2022) Phase Separation Drives SARS-CoV-2 Replication: A Hypothesis. Front. Mol. Biosci. 9:893067. doi: 10.3389/fmolb.2022.893067

INTRODUCTION

Identification of viral interactions within the host cell can lead to the design of novel strategies against infection. Recently, different high-throughput strategies have been implemented to characterize host interactions with SARS-CoV-2 proteins and genomic RNA.

Non-structural proteins of SARS-CoV-2 have been used for affinity purification to retrieve host binding partners using mass spectrometry in HEK-293T/17 cells (Gordon et al., 2020). A total of 332 interactions between human and SARS-CoV-2 proteins have been identified. Around 40% of SARS-CoV-2 interacting proteins are associated with vesicle trafficking pathways and endomembrane compartments.

Here, we focus on four experimental studies aiming to characterize interactions with SARS-CoV-2 genomic RNA.

In one experiment, a multi-omic approach was employed to identify which viral and human RNA-binding proteins (RBPs) are involved in SARS-CoV-2 infection (Kamel et al., 2021). The "comparative RNA interactome capture" (cRIC) method was developed to find in which way the RNA-bound proteome responds to the infection. The results show that SARS-CoV-2 genome is the epicenter of critical interactions with host proteins: many cellular RBP networks are remodeled upon SARS-CoV-2 infection and around 300 proteins are affected, mostly related to RNA metabolic

processes and antiviral defenses. A second approach called "viral RNA interactome capture" (vRIC) was employed to identify cellular and viral proteins interacting with SARS-CoV-2 genomic RNA (Kamel et al., 2021). Inhibition of specific proteins interacting with viral RNA was shown to impair SARS-CoV-2 infection.

In another study (Lee et al., 2021), the repertoire of host proteins associated with SARS-CoV-2 and HCoV-OC43 genomes was identified. The work relies on a robust nucleoprotein (RNP) capture protocol. More than 100 host factors directly binding to SARS-CoV-2 RNA were detected. By applying RNP capture on HCoV-OC43, evolutionary conserved interactions between the viral RNAs and the host proteins could be identified. Upon knockdown experiments and transcriptome analysis, Lee *et al.* identified 17 antiviral and 8 pro-viral RBPs that have a role in several steps of the mRNA life cycle. The authors identified La-related protein 1 (LARP1), a downstream target of the mTOR signaling pathway, as an important antiviral host factor that interacts with SARS-CoV-2 RNA.

Another group exploited an approach in which a comprehensive identification of RBPs followed by mass spectrometry (ChIRP-MS) led to the identification of host proteins that bind SARS-CoV-2 genomic RNA during active infection (Flynn et al., 2021). The results were corroborated with analyses from three RNA viruses and contributed to characterize the specificity of virus-host interactions. Flynn *et al.* also carried out a series of targeted CRISPR screens that highlighted the fact that a big portion of functional RNA-binding proteins act as host's protectors from virus-induced cell death. Comparative CRISPR screens, performed across seven RNA viruses, reveal both shared and SARS-specific antiviral factors. By combining the RNA-centric approach and the functional CRISPR screens, the authors found a functional connection between SARS-CoV-2 and mitochondria, showing that this organelle is a platform for antiviral activity.

A slightly different experiment led to the identification of more than 100 human proteins that directly and specifically bind to SARS-CoV-2 RNAs in infected cells, performing RNA antisense purification and mass spectrometry. Schmidt et al. linked SARS-CoV-2 interactome with changes in proteome abundance induced by viral infection, identifying cellular pathways relevant to SARS-CoV-2 infections. The authors demonstrated by genetic perturbation that both Cellular Nucleic-acid Binding Protein (CNBP) and LARP1, which are two of the most enriched viral RNA binders, have the ability to restrict SARS-CoV-2 replication in infected cells and provide a general map of their direct RNA contact sites. The authors demonstrated a reduced viral replication rate in two human cell lines after a pharmacological inhibition of three other binding partners (PPIA, ATP1A1, ARP2/3 complex).

As experimental studies require time and resources and are affected by intrinsic limitations (for instance mass-spec cannot identify every protein with the same efficiency), computational methods can be exploited to prioritize candidate targets. We previously used the CROSS method (Delli Ponti et al., 2017) to predict secondary structure content of and the *cat*RAPID approach (Bellucci et al., 2011; Agostini et al., 2013b; Cirillo et al., 2017) to compute >100000 human protein interactions with SARS-CoV-2 genomic RNA (Vandelli et al., 2020). The 5′ and 3′ end of SARS-CoV-2 were found to be highly structured, in agreement with subsequent experimental reports (Manfredonia et al., 2020) and show strong propensity to interact with human proteins. Among the identified interactors we identified there are several RNA editases and ATP-dependent RNA helicases that play a role in viral RNA processing and have a high propensity to participate in large macromolecular complexes. A number of proteins are predicted to be sequestered by SARS-CoV-2 genome and their recruitment contributes is thought to modify both the transcriptional and post-transcriptional regulations of host cells.

Here, we analyzed four experimental and one computational studies on human RBPs interactions with SARS-CoV-2 genomic RNA. We exploited the *cat*RAPID algorithm to estimate the ability of proteins to bind SARS-CoV-2 and identified a tight correlation between the number of experiments in which a specific protein is detected experimentally and its predicted binding affinity. Finally, we propose a model in which SARS-CoV-2 RNA promotes the formation of a phase-separated assembly by sequestering specific human proteins.

RESULTS

Interactomes Comparison

To retrieve interactions relevant for SARS-CoV-2 infection, we analysed four protein-RNA interactome experiments (**Supplementary Material S1**).

Twenty-one proteins were found in common to the four datasets (Flynn et al., 2021; Kamel et al., 2021; Lee et al., 2021; Schmidt et al., 2021) (**Figure 1A**). The list includes PABPC1 (Polyadenylate-binding protein 1), SND1 (Staphylococcal nuclease domain-containing protein 1), PPIA (Peptidyl-prolyl cis-trans isomerase A), DDX3X (ATP-dependent RNA helicase DDX3X), HNRNPA2B1 (Heterogeneous nuclear ribonucleoproteins A2/B1), HNRNPA0 (Heterogeneous nuclear ribonucleoprotein A), G3BP1 (Ras GTPase-activating protein-binding protein 1), G3BP2 (Ras GTPase-activating protein-binding protein 2), EIF4B (Eukaryotic translation initiation factor 4B), RPS2 (40S ribosomal protein S2), RPS3 (40S ribosomal protein S3), EIF3G (Eukaryotic translation initiation factor 3 subunit G) and YBX1 (Y-box-binding protein 1), **Supplementary Tables S1, S2**).

These proteins form a dense protein-protein network (**Figure 1B**) containing several stress granule components (G3BP1, G3BP2, EIF4B, DDX3X, YBX1, PABPC1), ribosomal units (RPS2 and RPS3) and pre-mRNA processing units (HNRNPA1/B2, HNRNPA0, YBX1) (Warde-Farley et al., 2010). The biological relevance of these interactions is confirmed by the fact that SARS-CoV-2 N protein impairs stress granule by sequestering G3BP1 (Lu et al., 2021; Zheng et al., 2021). RPS2 and RPS3 are important because the NSP1 protein of SARS-CoV-2 is responsible for the impairment of mRNA translation by blocking the entry access to the ribosome. The docking within the ribosomal entry channel occurs through binding with RPS2 and RPS3 together with 18S RNA (Mendez et al., 2021).

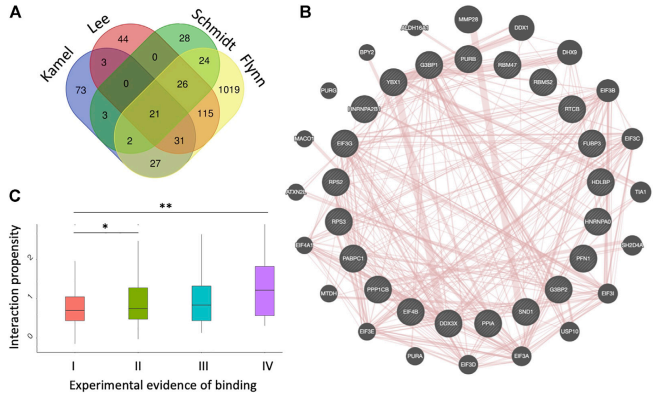


FIGURE 1 | Datasets of protein interactions with SARS-CoV-2 genome. **(A)** experimental datasets (Flynn et al., 2021; Kamel et al., 2021; Lee et al., 2021; Schmidt et al., 2021). The name of each dataset is shown above the diagrams. **(B)** diagram showing the protein-protein interactions among the 21 proteins identified in the four experiments, as annotated by GeneMANIA (Warde-Farley et al., 2010). **(C)** *cat*RAPID interaction scores (Agostini et al., 2013b; Armaos et al., 2021) correlate with the number of experiments reporting a protein to interact with SARS-CoV-2, indicating that strong binding proteins are more likely to be identified; *p-value < 0.05; **p-value < 0.01 (Wilcoxon rank sum test); I II, II, III and IV indicate proteins detected in 1,2,3 or 4 experiments, respectively.

Some of these proteins have been shown to be also relevant for other viruses' infection. SND1 is involved in Epstein-Barr infection (Tong et al., 1995); PABPC1 positively regulates Dengue virus infection (Suzuki et al., 2016); PPIA acts as a mediator for SARS-CoV nucleoprotein during the cell invasion process and stimulates RNA-binding ability of HCV NS5A (Chen et al., 2005; Foster et al., 2011); EIF3G is involved in FCV infection process (Pöyry et al., 2007) and DDX3X has been shown to facilitate the viral replication of other several viruses, such as HIV-1, Dengue, Zyka, Venezuelan equine encephalitis and hepatitis C virus (Yedavalli et al., 2004; Amaya et al., 2016; Doñate-Macián et al., 2018). DDX3X has been identified as a suitable target to fight against SARS-CoV-2 infection by Ciccosanti et al. (2021). More precisely, DDX3X has the capability of unfolding viral RNA secondary structures (Kukhanova et al., 2020) as reported for HIV-1 (Brai et al., 2020) in which it enhances both translation and nucleus-to-cytoplasm transport (Stunnenberg et al., 2018), and West Nile (Brai et al., 2019). DDX3X belongs to the DEAD-box family of ATP-dependent RNA helicases and assumes a crucial role in an important variety of processes concerning RNA metabolism, including transcription, splicing, and initial phase of translation (Ariumi, 2014). Importantly, DDX3X interacts with the N protein of SARS-CoV-2 and is required to infect both Vero E6 and Calu-3 cells (Ciccosanti et al., 2021). Additionally, SARS-CoV-2 protein N interacts with DDX3X to inhibit its activity in the antiviral response (Winnard et al., 2021). For these reasons, treating cells with DDX3X inhibitors represents a promising approach to block SARS-CoV-2 replication and viral production (Maga et al., 2011; Brai et al., 2020).

Relationship Between Experimental Interactomes and Computational Predictions

We used the *cat*RAPID method to understand the relationship between experimental evidence of binding and predicted interaction propensity that estimates interaction affinity (Agostini et al., 2013a; Cid-Samper et al., 2018). For this analysis we followed a procedure previously introduced to study the interactome of the long non-coding RNA *Xist* (Cirillo et al., 2017). We computed all SARS-CoV-2 interactions with proteins reported in the four experimental datasets and counted how many times they were identified (**Supplementary Material S1**). We observed a distinct correlation between occurrence and strength of interactions, indicating that high-affinity interactions are more likely to be detected (**Figure 1C**). We note that in the case of *Xist*, strong interaction proteins were predicted to initiate the formation of a phase-separated assembly (Cerase et al., 2019, 2022), as recently confirmed experimentally (Markaki et al., 2021; Jachowicz et al., 2022).

Evaluation of the Predictions of SARS-CoV-2 Protein Interactions

The vRIC dataset by Kamel *et al.* contains both enriched and depleted interactions (Kamel et al., 2021) and thus can be used to assess the ability of *cat*RAPID to distinguish between binding and non-binding proteins. To analyze the vRIC interactome, we

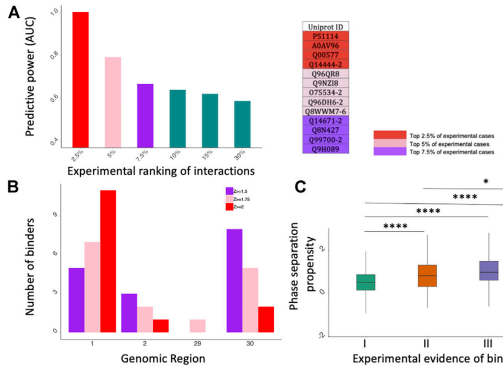


FIGURE 2 | catRAPID and catGRANULE predictions of protein interactions. **(A)** catRAPID performance evaluation. On the *X* axis we report different portions of the experimental dataset ranked by fold change and on the *Y* axis there is the corresponding predictive power (Area Under the ROC Curve, AUC). On the right, we report a summary table showing the Uniprot IDs of top 2.5%, 5% and 7.5% experimental cases. **(B)** Distribution of specific binders for Kamel et al. dataset (Kamel et al., 2021). The most contacted SARS-CoV-2 genomic regions correspond to the 5′ (first fragment) e 3′ (30th fragment). **(C)** catGRANULE phase separation propensity scores correlate with the number of experiments reporting a protein to interact with SARS-CoV-2 (Bolognesi et al., 2016; Cid-Samper et al., 2018); **p*-value < 0.05; *****p*-value < 0.0001 (Wilcoxon rank sum test); I II, II, III and IV indicate proteins detected in 1,2,3 or 4 experiments, respectively.

computed *cat*RAPID predictions of interactions with an experimental FDR <0.10 for SARS-CoV-2 RNA following a procedure detailed in a previous work (Vandelli et al., 2020) (**Supplementary Material S1**).

As shown in **Figure 2A**, *cat*RAPID performs extremely well when the proteins are ranked according to their experimental scores (fold change; **Supplementary Table S3**): the predictive power is proportional to the significance of protein interactions: the Area Under the ROC Curve (AUC) increases from 0.60 to 0.99 while the experimental scores move from 30% (i.e., the 30% strongest positives vs. the 30% strongest negatives) to 2.5% (i.e., the 2.5% strongest positives vs. the 2.5% strongest negatives). Thus, in agreement with the results presented in **Figure 1C**, computational approaches such as *cat*RAPID can be exploited to address the problem of which proteins bind more tightly to SARS-CoV-2 genome.

Specific Binders to SARS-CoV-2 Genomic Fragments

*cat*RAPID was employed for the localization of protein binding sites on SARS-CoV-2 genomic RNA. To identify which regions of SARS-CoV-2 bind to specific proteins, we computed interactions for the four experimental protein datasets (30 fragments; **Supplementary Material S4**), a procedure already proven to be efficient in a previous work (Vandelli et al., 2020).

For each dataset the proteins bound to one fragment at a certain interaction threshold were retained as interactors. We applied three Z-score thresholds ($Z \geq 1.5$, $Z \geq 1.75$ and $Z \geq 2$) in order to evaluate the binding at the different levels of stringency. Higher Z-scores correspond to higher interaction strength (**Supplementary Material S5**).

Regions encompassing nucleotides 1–1000, 1001–2000, 22001–23000, 26001–27000, 28001–29000, 29001–29903 (Fragments 1, 2, 23, 27, 29 and 30 respectively) are the most contacted SARS-CoV-2 regions, with a high number of interactors in fragments 1, 2 and 30. (**Figure 2B**; **Supplementary Figures S1–S3**). In particular, fragment 1, corresponding to the 5′ end of SARS-CoV-2 genome, is the region showing the highest number of specific interactors in all four datasets, as previously discovered (Vandelli et al., 2020). DDX3X is the only common interactor reported in the experimental and computational studies. At a $Z \geq 1.75$ we DDX3X is found to bind specifically to fragment 1 of SARS-CoV-2.

Experimental Interactors Have a High Propensity to Phase-Separate

Stress granules facilitate the establishment of an antiviral state by limiting viral protein accumulation and regulating signaling cascades that affect replication (McCormick and Khaperskyy, 2017). The sequestration of G3BP1, G3BP2, EIF4B, DDX3X, YBX1, PABPC1, among other proteins, is part of a mechanism through which SARS-CoV-2 eludes the host immune response by weakening the formation of stress granules (Lu et al., 2021; Zheng et al., 2021). Biochemically, stress granule proteins form labile protein-protein and protein-RNA interactions (Balcerak et al., 2019; Vandelli et al., 2022), which induces the condensation in liquid-liquid phase

separated assemblies (Gotor et al., 2020). We reasoned that the relatively small overlap among experimental datasets (**Figure 1A**) could be caused by the establishment of weak molecular interactions with SARS-CoV-2 RNA. In agreement with this observation, previous studies have suggested that phase separation could be a mechanism through which SARS-CoV-2 attracts host proteins (Iserman et al., 2020; Vandelli et al., 2020).

Using the *cat*GRANULE algorithm to predict phase separation propensities (Bolognesi et al., 2016; Cid-Samper et al., 2018) we analyzed the interactomes of the four experimental datasets. We discovered that the phase separation propensity correlates with how many times proteins are identified experimentally (**Figure 2C**). Considering that strong binding propensities are associated with proteins reported in the four experiments (**Figure 1C**) and the reliability of our approach (**Figure 2A**), we speculate that a possible mechanism of action for SARS-CoV-2 is to target proteins that attract other partners through phase separation.

DISCUSSION

This work is a comparative analysis on protein-RNA interactomes reported in experimental and computational studies. We found several proteins shared by the four experiments, including PABPC1, SND1, PPIA, EIF3G and DDX3X, which previous studies have shown to regulate replication of viruses.

DDX3X is found in all the experimental studies and it has been proven fundamental in SARS-CoV-2 biological processes and in the replication process of other viruses (Maga et al., 2011; Ariumi, 2014; Stunnenberg et al., 2018; Brai et al., 2019, 2020; Kukhanova et al., 2020; Ciccosanti et al., 2021; Winnard et al., 2021). *cat*RAPID predictions of human protein interactions with SARS-CoV-2 showed a prevalence of specific binders to the 5′ end of the virus, with DDX3X being one of them. Since *cat*RAPID reproduces experimental data to a remarkable extent, as assessed by directly comparing performances at different cut-offs, we believe that this information on the localization of protein interactions is to be taken into account for future analyses.

Predictive studies always have a margin of error, so further work will be necessary for a complete understanding of the specific binding sites and the role(s) of proteins in the context of infection.

In a recent study (Cirillo et al., 2017), we reported that the long non-coding RNA *Xist* physically interacts with few specific proteins that attract several other proteins (Cerase et al., 2019) forming a phase-separated assembly that silences the X chromosome (Cerase et al., 2022; Jachowicz et al., 2022). The relatively poor overlap of interactors among SARS-CoV-2 studies (only 21 proteins in common out of hundreds identified in total) suggests a mechanism similar to the one identified for *Xist*. The fact that SARS-CoV-2 binding proteins are either stress granules components or have high phase separation propensity supports our hypothesis. Indeed, phase separation is caused by weak protein-protein or protein-RNA interactions (Balcerak et al., 2019; Vandelli et al., 2022), which renders the identification of binding partners particularly difficult at the experimental level (Tartaglia, 2016; Cerase and Tartaglia, 2020) and could hamper their reproducibility. Moreover, the fact that proteins with the highest interaction and phase separation propensities were identified in all experimental studies suggests that they could act as the primary attractors to ignite the formation of an assembly that is capable of using host elements for replication. Further work is needed to study this fundamental aspect of SARS-CoV-2 biology and how it could be exploited to prevent viral infection. For example, molecular chaperones (Tartaglia et al., 2010; Alagar Boopathy et al., 2022) could be important players (Guihur et al., 2020) to be investigated in more detail.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**, further inquiries can be directed to the corresponding author.

AUTHOR CONTRIBUTIONS

AV and GT conceived the study. AV and GV carried out the analysis. AV, GV, and GT wrote the paper.

FUNDING

The research leading to these results has been supported by European Research Council (RIBOMYLOME_309545 and ASTRA_855923), the H2020 projects (IASIS_727658 and INFORE_825080) and the Spanish Ministry of Science and Innovation (RYC 2019-026752-I and PID 2020-117454RA-I00).

ACKNOWLEDGMENTS

The authors would like to thank Prof. Alberto Danielli and Prof. Marc Torrent Burgas for illuminating discussions.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmolb.2022.893067/full#supplementary-material

Supplementary Table S1 | Twenty-one proteins shared by the four experimental studies. Uniprot IDs and gene names of the proteins are provided.

Supplementary Table S2 | Proteins found in each different experimental study.

Supplementary Table S3 | Dataset used for the evaluation of *cat*RAPID performances. For each protein, its Uniprot ID, experimental fold-change and normalized *cat*RAPID score are reported. The top experimental cases are highlighted in different colors according to the relative portion of the dataset.

Supplementary Table S4 | Sequence and genomic location of each of the 30 fragments of SARS-CoV-2.

Supplementary Table S5 | List of interactors for each experimental dataset, predicted to bind univocally each of the 30 SARS-CoV-2 fragments.

REFERENCES

Agostini, F., Cirillo, D., Bolognesi, B., and Tartaglia, G. G. (2013a). X-inactivation: Quantitative Predictions of Protein Interactions in the Xist Network. *Nucleic Acids Res.* 41, e31. doi:10.1093/nar/gks968

Agostini, F., Zanzoni, A., Klus, P., Marchese, D., Cirillo, D., and Tartaglia, G. G. (2013b). catRAPID Omics: a Web Server for Large-Scale Prediction of Protein-RNA Interactions. *Bioinformatics* 29, 2928–2930. doi:10.1093/bioinformatics/btt495

Alagar Boopathy, L. R., Jacob-Tomas, S., Alecki, C., and Vera, M. (2022). Mechanisms Tailoring the Expression of Heat Shock Proteins to Proteostasis Challenges. *J. Biol. Chem.*, 101796. doi:10.1016/j.jbc.2022.101796

Amaya, M., Brooks-Faulconer, T., Lark, T., Keck, F., Bailey, C., Raman, V., et al. (2016). Venezuelan Equine Encephalitis Virus Non-structural Protein 3 (nsP3) Interacts with RNA Helicases DDX1 and DDX3 in Infected Cells. *Antivir. Res.* 131, 49–60. doi:10.1016/j.antiviral.2016.04.008

Ariumi, Y. (2014). Multiple Functions of DDX3 RNA Helicase in Gene Regulation, Tumorigenesis, and Viral Infection. *Front. Genet.* 5. 423. doi:10.3389/fgene.2014.00423

Armaos, A., Colantoni, A., Proietti, G., Rupert, J., and Tartaglia, G. G. (2021). catRAPID Omics v2.0: Going Deeper and Wider in the Prediction of Protein-RNA Interactions. *Nucleic Acids Res.* 49, W72–W79. gkab393. doi:10.1093/nar/gkab393

Balcerak, A., Trebinska-Stryjewska, A., Konopinski, R., Wakula, M., and Grzybowska, E. A. (2019). RNA-protein Interactions: Disorder, Moonlighting and Junk Contribute to Eukaryotic Complexity. *Open Biol.* 9, 190096. doi:10.1098/rsob.190096

Bellucci, M., Agostini, F., Masin, M., and Tartaglia, G. G. (2011). Predicting Protein Associations with Long Noncoding RNAs. *Nat. Methods* 8, 444–445. doi:10.1038/nmeth.1611

Bolognesi, B., Lorenzo Gotor, N., Dhar, R., Cirillo, D., Baldrighi, M., Tartaglia, G. G., et al. (2016). A Concentration-dependent Liquid Phase Separation Can Cause Toxicity upon Increased Protein Expression. *Cell Rep.* 16, 222–231. doi:10.1016/j.celrep.2016.05.076

Brai, A., Martelli, F., Riva, V., Garbelli, A., Fazi, R., Zamperini, C., et al. (2019). DDX3X Helicase Inhibitors as a New Strategy to Fight the West Nile Virus Infection. *J. Med. Chem.* 62, 2333–2347. doi:10.1021/acs.jmedchem.8b01403

Brai, A., Riva, V., Saladini, F., Zamperini, C., Trivisani, C. I., Garbelli, A., et al. (2020). DDX3X Inhibitors, an Effective Way to Overcome HIV-1 Resistance Targeting Host Proteins. *Eur. J. Med. Chem.* 200, 112319. doi:10.1016/j.ejmech.2020.112319

Cerase, A., Armaos, A., Neumayer, C., Avner, P., Guttman, M., and Tartaglia, G. G. (2019). Phase Separation Drives X-Chromosome Inactivation: a Hypothesis. *Nat. Struct. Mol. Biol.* 26, 331–334. doi:10.1038/s41594-019-0223-0

Cerase, A., Calabrese, J. M., and Tartaglia, G. G. (2022). Phase Separation Drives X-Chromosome Inactivation. *Nat. Struct. Mol. Biol.* 29, 183–185. doi:10.1038/s41594-021-00697-0

Cerase, A., and Tartaglia, G. G. (2020). Long Non-coding RNA-Polycomb Intimate Rendezvous. *Open Biol.* 10, 200126. doi:10.1098/rsob.200126

Chen, Z., Mi, L., Xu, J., Yu, J., Wang, X., Jiang, J., et al. (2005). Function of HAb18G/CD147 in Invasion of Host Cells by Severe Acute Respiratory Syndrome Coronavirus. *J. Infect. Dis.* 191, 755–760. doi:10.1086/427811

Ciccosanti, F., Di Rienzo, M., Romagnoli, A., Colavita, F., Refolo, G., Castilletti, C., et al. (2021). Proteomic Analysis Identifies the RNA Helicase DDX3X as a Host Target against SARS-CoV-2 Infection. *Antivir. Res.* 190, 105064. doi:10.1016/j.antiviral.2021.105064

Cid-Samper, F., Gelabert-Baldrich, M., Lang, B., Lorenzo-Gotor, N., Ponti, R. D., Severijnen, L. W. F. M., et al. (2018). An Integrative Study of Protein-RNA Condensates Identifies Scaffolding RNAs and Reveals Players in Fragile X-Associated Tremor/Ataxia Syndrome. *Cell Rep.* 25, 3422–e7. e7. doi:10.1016/j.celrep.2018.11.076

Cirillo, D., Blanco, M., Armaos, A., Buness, A., Avner, P., Guttman, M., et al. (2017). Quantitative Predictions of Protein Interactions with Long Noncoding RNAs. *Nat. Methods* 14, 5–6. doi:10.1038/nmeth.4100

Delli Ponti, R., Marti, S., Armaos, A., and Tartaglia, G. G. (2017). A High-Throughput Approach to Profile RNA Structure. *Nucleic Acids Res.* 45, e35. doi:10.1093/nar/gkw1094

Doñate-Macián, P., Jungfleisch, J., Pérez-Vilaró, G., Rubio-Moscardo, F., Perálvarez-Marín, A., Diez, J., et al. (2018). The TRPV4 Channel Links Calcium Influx to DDX3X Activity and Viral Infectivity. *Nat. Commun.* 9, 2307. doi:10.1038/s41467-018-04776-7

Flynn, R. A., Belk, J. A., Qi, Y., Yasumoto, Y., Wei, J., Alfajaro, M. M., et al. (2021). Discovery and Functional Interrogation of SARS-CoV-2 RNA-Host Protein Interactions. *Cell* 184, 2394–2411. doi:10.1016/j.cell.2021.03.012

Foster, T. L., Gallay, P., Stonehouse, N. J., and Harris, M. (2011). Cyclophilin A Interacts with Domain II of Hepatitis C Virus NS5A and Stimulates RNA Binding in an Isomerase-dependent Manner. *J. Virol.* 85, 7460–7464. doi:10.1128/JVI.00393-11

Gordon, D. E., Jang, G. M., Bouhaddou, M., Xu, J., Obernier, K., White, K. M., et al. (2020). A SARS-CoV-2 Protein Interaction Map Reveals Targets for Drug Repurposing. *Nature* 583, 459–468. doi:10.1038/s41586-020-2286-9

Gotor, N. L., Armaos, A., Calloni, G., Torrent Burgas, M., Vabulas, R. M., De Groot, N. S., et al. (2020). RNA-binding and Prion Domains: the Yin and Yang of Phase Separation. *Nucleic Acids Res.* 48, 9491–9504. doi:10.1093/nar/gkaa681

Iserman, C., Roden, C. A., Boerneke, M. A., Sealfon, R. S. G., McLaughlin, G. A., Jungreis, I., et al. (2020). Genomic RNA Elements Drive Phase Separation of the SARS-CoV-2 Nucleocapsid. *Mol. Cell* 80, 1078–1091. e6. doi:10.1016/j.molcel.2020.11.041

Jachowicz, J. W., Strehle, M., Banerjee, A. K., Blanco, M. R., Thai, J., and Guttman, M. (2022). Xist Spatially Amplifies SHARP/SPEN Recruitment to Balance Chromosome-wide Silencing and Specificity to the X Chromosome. *Nat. Struct. Mol. Biol.* 29, 239–249. doi:10.1038/s41594-022-00739-1

Kamel, W., Noerenberg, M., Cerikan, B., Chen, H., Järvelin, A. I., Kammoun, M., et al. (2021). Global Analysis of Protein-RNA Interactions in SARS-CoV-2-Infected Cells Reveals Key Regulators of Infection. *Mol. Cell* 81, 2851–2867. doi:10.1016/j.molcel.2021.05.023

Kukhanova, M. K., Karpenko, I. L., and Ivanov, A. V. (2020). DEAD-box RNA Helicase DDX3: Functional Properties and Development of DDX3 Inhibitors as Antiviral and Anticancer Drugs. *Molecules* 25, 1015. doi:10.3390/molecules25041015

Lee, S., Lee, Y.-s., Choi, Y., Son, A., Park, Y., Lee, K.-M., et al. (2021). Young-sukThe SARS-CoV-2 RNA Interactome. *Mol. Cell* 81, 2838–2850. doi:10.1016/j.molcel.2021.04.022

Lu, S., Ye, Q., Singh, D., Cao, Y., Diedrich, J. K., Yates, J. R., et al. (2021). The SARS-CoV-2 Nucleocapsid Phosphoprotein Forms Mutually Exclusive Condensates with RNA and the Membrane-Associated M Protein. *Nat. Commun.* 12, 502. doi:10.1038/s41467-020-20768-y

Maga, G., Falchi, F., Radi, M., Botta, L., Casaluce, G., Bernardini, M., et al. (2011). Toward the Discovery of Novel Anti-HIV Drugs. Second-Generation Inhibitors of the Cellular ATPase DDX3 with Improved Anti-HIV Activity: Synthesis, Structure-Activity Relationship Analysis, Cytotoxicity Studies, and Target Validation. *ChemMedChem* 6, 1371–1389. doi:10.1002/cmdc.201100166

Manfredonia, I., Nithin, C., Ponce-Salvatierra, A., Ghosh, P., Wirecki, T. K., Marinus, T., et al. (2020). Genome-wide Mapping of SARS-CoV-2 RNA Structures Identifies Therapeutically-Relevant Elements. *Nucleic Acids Res.* 48, 12436–12452. doi:10.1093/nar/gkaa1053

Markaki, Y., Gan Chong, J., Wang, Y., Jacobson, E. C., Luong, C., Tan, S. Y. X., et al. (2021). Xist Nucleates Local Protein Gradients to Propagate Silencing across the X Chromosome. *Cell* 184, 6174–6192. doi:10.1016/j.cell.2021.10.022

McCormick, C., and Khaperskyy, D. A. (2017). Translation Inhibition and Stress Granules in the Antiviral Immune Response. *Nat. Rev. Immunol.* 17, 647–660. doi:10.1038/nri.2017.63

Mendez, A. S., Ly, M., González-Sánchez, A. M., Hartenian, E., Ingolia, N. T., Cate, J. H., et al. (2021). The N-Terminal Domain of SARS-CoV-2 Nsp1 Plays Key Roles in Suppression of Cellular Gene Expression and Preservation of Viral Gene Expression. *Cell Rep.* 37, 109841. doi:10.1016/j.celrep.2021.109841

Pöyry, T. A. A., Kaminski, A., Connell, E. J., Fraser, C. S., and Jackson, R. J. (2007). The Mechanism of an Exceptional Case of Reinitiation after Translation of a Long ORF Reveals Why Such Events Do Not Generally Occur in Mammalian mRNA Translation. *Genes Dev.* 21, 3149–3162. doi:10.1101/gad.439507

Schmidt, N., Lareau, C. A., Keshishian, H., Ganskih, S., Schneider, C., Hennig, T., et al. (2021). The SARS-CoV-2 RNA-Protein Interactome in Infected Human Cells. *Nat. Microbiol.* 6, 339–353. doi:10.1038/s41564-020-00846-z

Stunnenberg, M., Geijtenbeek, T. B. H., and Gringhuis, S. I. (2018). DDX3 in HIV-1 Infection and Sensing: A Paradox. *Cytokine. Growth Factor Rev.* 40, 32–39. doi:10.1016/j.cytogfr.2018.03.001

Suzuki, Y., Chin, W.-X., Han, Q. E., Ichiyama, K., Lee, C. H., Eyo, Z. W., et al. (2016). Characterization of RyDEN (C19orf66) as an Interferon-Stimulated Cellular Inhibitor against Dengue Virus Replication. *PLOS Pathog.* 12, e1005357. doi:10.1371/journal.ppat.1005357

Tartaglia, G. G., Dobson, C. M., Hartl, F. U., and Vendruscolo, M. (2010). Physicochemical Determinants of Chaperone Requirements. *J. Mol. Biol.* 400, 579–588. doi:10.1016/j.jmb.2010.03.066

Tartaglia, G. G. (2016). The Grand Challenge of Characterizing Ribonucleoprotein Networks. *Front. Mol. Biosci.* 3, 24. doi:10.3389/fmolb.2016.00024

Tong, X., Drapkin, R., Yalamanchili, R., Mosialos, G., and Kieff, E. (1995). The Epstein-Barr Virus Nuclear Protein 2 Acidic Domain Forms a Complex with a Novel Cellular Coactivator that Can Interact with TFIIE. *Mol. Cell. Biol.* 15, 4735–4744. doi:10.1128/MCB.15.9.4735

Vandelli, A., Cid Samper, F., Torrent Burgas, M., Sanchez de Groot, N., and Tartaglia, G. G. (2022). The Interplay between Disordered Regions in RNAs and Proteins Modulates Interactions within Stress Granules and Processing Bodies. *J. Mol. Biol.* 434, 167159. doi:10.1016/j.jmb.2021.167159

Vandelli, A., Monti, M., Milanetti, E., Armaos, A., Rupert, J., Zacco, E., et al. (2020). Structural Analysis of SARS-CoV-2 Genome and Predictions of the Human Interactome. *Nucleic Acids Res.* 48, 11270–11283. doi:10.1093/nar/gkaa864

Warde-Farley, D., Donaldson, S. L., Comes, O., Zuberi, K., Badrawi, R., Chao, P., et al. (2010). The GeneMANIA Prediction Server: Biological Network Integration for Gene Prioritization and Predicting Gene Function. *Nucleic Acids Res.* 38, W214–W220. doi:10.1093/nar/gkq537

Winnard, P. T., Vesuna, F., and Raman, V. (2021). Targeting Host DEAD-Box RNA Helicase DDX3X for Treating Viral Infections. *Antivir. Res.* 185, 104994. doi:10.1016/j.antiviral.2020.104994

Yedavalli, V. S. R. K., Neuveut, C., Chi, Y.-h., Kleiman, L., and Jeang, K.-T. (2004). Requirement of DDX3 DEAD Box RNA Helicase for HIV-1 Rev-RRE Export Function. *Cell* 119, 381–392. doi:10.1016/j.cell.2004.09.029

Zheng, Z.-Q., Wang, S.-Y., Xu, Z.-S., Fu, Y.-Z., and Wang, Y.-Y. (2021). SARS-CoV-2 Nucleocapsid Protein Impairs Stress Granule Formation to Promote Viral Replication. *Cell Discov.* 7, 38. doi:10.1038/s41421-021-00275-0

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Part III

CLOSING REMARKS

CHAPTER 8

GENERAL DISCUSSION

In my thesis, I presented two distinct but interconnected topics: the characterization of ribonucleoprotein condensates and the analysis of human-SARS-CoV-2 interactions networks.

In the first part of my Ph.D., I focused my attention on protein and RNA interactions (Chapter 3) occurring in SGs and PBs, two condensates formed through a process of liquid-liquid phase separation (Chapter 4). My goal was to study how different types of macromolecular interactions (protein-protein, protein-RNA and RNA-RNA interactions) and structural information could be linked together to define the network sustaining these condensates. My main finding is that poorly structured RNA and protein elements are key ingredients to reversibly aggregate components of SGs and PBs. My observations were based on a previous study of different experimental and computational methods to assess physicochemical determinants of protein-protein, RNA-RNA and especially protein-RNA interactions (Chapter 3).
My analysis of RNP condensates was integrated with information on amyloid aggregates, disease-related SNVs and their effect on these organelles' components in a new database called PRALINE (Chapter 5), a very complete resource containing both experimental and computational high-throughput data.

The second part of my Ph.D. is related to studies on SARS-CoV-2. Since very little information was available in March 2020, I employed computational methods to assess the structural properties of the virus, to compare them with other coronaviruses and to generate a complete interactome with human proteins (Chapter 6). Comparing my predictions and available experimental interactomes, I proposed that few protein interac-

tors found in multiple experiments and different conditions promote progressive accumulation of protein components involved in viral replication (Chapter 7).

The link between the two topics of my thesis is evident if we consider that biological condensates store essential protein and RNA elements of the cell to promote survival in stress conditions, induced not only by physicochemical shocks but also by infection. Several viral and bacterial pathogens are known to modulate condensate formation to protect themselves from host defenses. This can be achieved by cleavage of essential condensates proteins (e.g. West Nile virus), or by sequestration of components into viral factories (e.g. Hepatitis C virus) (Ariumi et al., 2011; Gaete-Argel et al., 2019).
This is one of the reasons why at the beginning of the pandemic, I hypothesized that the SARS-CoV-2 genome could establish interactions with the human host targeting specific molecules essential for host immune response, which otherwise would be recruited inside condensates to be protected or would impair the viral replication.

8.1. Macromolecular interactions in RNP condensates

Phase separation is nowadays an extensively studied topic that seems to influence and coordinate a lot of cellular mechanisms, controlling the concentration of molecules and biochemical reactions and can lead to the formation of condensates to store important components in stressful moments of the cell or can organize chromatin like in the case of

X-chromosome inactivation (Cerase et al., 2019).
SGs and PBs are two of the best characterized macromolecular assemblies formed through liquid-liquid phase-separation and despite having different essential elements and different functions, they share many protein and RNA factors and communicate with each other (Li et al., 2013).
Despite previous studies on essential components of these condensates and various attempts made to characterize each type of macromolecular interactions separately, there were still limitations in what we knew about these condensates, and in particular about how structural information of both protein and RNAs could influence the way in which protein-protein, protein-RNA and RNA-RNA interactions can regulate each other.

This is what I addressed in the first part of my Ph.D. and in particular in Chapter 4, in which I compared SGs and PBs employing high-throughput experimental data from techniques for RNA secondary structure assessment and inter-molecular interactions characterization, combined with computational approaches to strengthen the signal obtained experimentally.
Given that RNAs enriched in SGs and PBs are very long and I found them to be globally poorly structured, I hypothesized they could be prone to bind multiple molecules and to act as scaffolds to lead the formation and stability of these granules, just like the lncRNA NEAT1 for paraspeckles (Souquere et al., 2010). Indeed, longer RNAs can guarantee larger single-stranded regions to bind to other RNAs and, at the same time, they could still contain structured regions prone to protein binding (Sanchez de Groot et al., 2019).
Looking at the signal derived from both experimental and computational data, I found that a population of RNAs enriched in these granules have an

extremely high degree of interactivity with both proteins and other RNAs, confirming my hypothesis. In parallel, I focused on protein-protein interactions. In agreement with a previous report indicating that proteins with a higher structural disorder are generally more prone to aggregate (Tartaglia et al., 2008), I found that highly disordered proteins enriched in condensates have higher interactivity among themselves compared to non-granule proteins. This led me to the realization that poorly structured RNAs and proteins preferentially tend to bind each other and that disordered regions in both proteins and RNAs of these condensates are the key factors to link all types of macromolecular interactions together, creating a circular scenario in which precise protein and RNAs elements could act as scaffolds and generate a dense network of contacts.

There are of course limitations in my analysis. First of all, the composition of the condensates and in particular of SGs is strongly dependent on stress and cell types, with a dramatic change even in the essential protein and RNA components (Markmiller et al., 2018). In addition, what we know about the SGs is strongly limited to their cores, easier to purify, while the composition of the shell is more difficult to study due to its diluted phase (Jain et al., 2016). Since each SG has multiple cores with only a limited amount of proteins and RNAs (Khong et al., 2017), the population of proteins and RNAs with high multivalency that we discussed in Chapter 4 could be actually a series of different small networks that belong to different cores rather than a unique set of interactions. Secondly, experimental techniques to assess macromolecular interactions are often carried out in physiological conditions of the cell, so in absence of stress, even though real-time monitoring of SGs dynamics is partially possible nowadays thanks to particular probes binding to core components (Shao

et al., 2021). In addition, the techniques for secondary structure determination and interactions identification are not bias-free. For example, cross-linking procedures for RNA-RNA interactions detection which employing AMT and derivatives have limited efficiency (Garrett-Wheeler et al., 1984; Harris and Christian, 2009), mainly due to the preferential binding of AMT to pyrimidines and the limitations in the ligation approach, causing drops in performances, especially in highly structured small RNAs (Sharma et al., 2016; Lu et al., 2016). This is then reflected in multiple computational tools using these data as training sets.

However, the establishment of stress-free interactions could also be interpreted as a mechanism to form a series of pre-networks of contacts that facilitates the recruitment of essential components under actual harmful conditions, to form firstly the different cores of SGs and then the entire condensates, helping also their dissolution upon stress clearance. Further analyses will be needed to confirm or reject this interesting hypothesis.
All these factors, coupled with the dynamicity of these aggregates, make the study of condensates' structure and composition fairly complex. New techniques for real-time imaging tracking different components simultaneously, together with the development of more precise methods for the assessment of the interactions will certainly help the characterization of these assemblies.

SGs, as well as other types of ribonucleoprotein condensates, are transient and reversible entities that form in the cell during harmful conditions and can dissolve upon stress clearance. However, changes in the composition or concentration of components inside these assemblies can lead to the formation of irreversible, solid-like assemblies causing oftentimes neu-

rodegenerative diseases (Murakami et al., 2015; Bolognesi et al., 2016).
The analysis carried out in Chapter 4 highlighted the role of structural content and macromolecular interactions in liquid-like condensates, involving both proteins and RNAs. This is the reason why I decided to build PRALINE (Chapter 5) that includes both experimental and computational high-throughput data of SGs and PBs.
Furthermore, I included information on proteins coalescing in solid-like aggregates to have a more comprehensive view. Since SNVs seem to determine the transition between liquid-like to solid-like behavior, I provided information on disease-related SNVs falling inside coding regions of RNAs and proteins enriched in SGs, PBs and amyloids. Being my focus on structure and interactions, I calculated changes in the secondary structure propensity of mutated RNA sequences compared to the wild-type and I collected variants falling inside binding regions in the context of condensates' macromolecular interactions.
In addition, PRALINE is the first database to show the predicted difference in liquid-liquid phase-separation (LLPS) and liquid-solid-phase transition (LSPT) potential of mutated proteins compared to the wild-type sequences (Tartaglia et al., 2008; Bolognesi et al., 2016).
Mutations and SNVs can influence the conformations and interfaces of macromolecules and consequently the already mentioned network of contacts inside the condensates and has to be considered another layer that adds up to the already complex scenario created by these assemblies.

As mentioned in the analysis done in Chapter 4, the accuracy of our data is constrained by current limitations affecting the experimental techniques for the interactions assessment and the predictive methods based on them. Furthermore, predictions of LLPS obtained with catGRANULE

algorithm (Bolognesi et al., 2016) for individual SNVs could not be tested on experimental data, due to the lack of available databases of SNVs' influence on phase-separation, even though mutational assays were carried out in previous works starting from this algorithm's predictions (Bolognesi et al., 2016; Gotor et al., 2020). The introduction of new experimental data could help to increase the resolution of the method, providing new insights into the SNVs' influence on condensate dynamics.

8.2. SARS-CoV-2 infection and interactions with human cells

The second part of my Ph.D. started in 2020 when the Covid-19 pandemic hit the world. While I was investigating the composition and functions of the condensates, I realized that many viral pathogens were already known either to suppress or regulate the formation of these assemblies, or to sequester components of these organelles as one of their defense mechanisms to tamper with the innate immune response of the cell (Gaete-Argel et al., 2019).

In this context, I hypothesized that SARS-CoV-2 could exploit similar mechanisms once come in contact with the human host. This fueled the work done in Chapter 6. Firstly, I compared its sequence with others of known coronavirus to identify regions with high structural content that could act as attractors of many proteins. Secondly, I predicted the interactions of the viral genome with the human proteome to search for human factors that could be targeted by the virus to enhance its replication. The results confirmed my previous observations, as especially the 5' of the SARS-CoV-2 genome was predicted to be a highly structured region and

an attractor of many human proteins known to be condensates components and belonging to the innate immune response of the cell, like the helicases DDX1 and DDX3X.

At the time of my first analysis, there were still few published human-SARS-CoV-2 interactome experiments to confirm my predictions (Flynn et al., 2021; Schmidt et al., 2021; Kamel et al., 2021; Lee et al., 2021), but I already noticed that the number of human interactors found simultaneously in multiple experiments was very little, while the majority of them was collected only in one or two studies. Since the experiments were carried out in different cell types and following different protocols, I speculated that the most shared binders were either the ones with the strongest affinity to bind the viral genome, or were particularly important targets to protect the virus survival, or both. This set the baseline for the work carried out in Chapter 7, where I found that these factors had both the highest propensity to bind the SARS-CoV-2 genome and the highest propensity to phase-separate, with some elements already known components of SG and PBs and already linked to the infection processes of several other viruses like Dengue and HIV-1, confirming that probably SARS-CoV-2 has an interest in sequestering these components to help its survival and replication while avoiding their entry into phase-separating assemblies.

Certainly, our methods are limited by the use of only computational tools and predictors and by the massive length of the SARS-CoV-2 genome. To use catRAPID algorithm (Bellucci et al., 2011) for the calculation of protein-RNA interactions predictions, we divided the viral genomic sequence into multiple fragments in order to retain as much structural and

physicochemical information as possible while avoiding calculation problems. This fragmentation procedure though had already been employed effectively in previous viral studies (Delli Ponti et al., 2018) and several works have demonstrated that catRAPID predictions can be validated experimentally, while the algorithm itself has been updated regularly to allow the analysis of entire proteomes and transcriptomes for multiple organisms (Armaos et al., 2021).

Since the publication of my work on SARS-CoV-2, other studies have detailed aspects of the infection process, including critical steps in the translation and transcription of the genome once entered in the cell thanks to the binding to ACE2, the packaging of new virions at the ER-Golgi interface (Klein et al., 2020), mechanisms of defense antagonizing the production of interferons (Wu et al., 2021) and exploiting host components to regulate its infectivity, such as DEAD-box helicases DDX1, DDX6 and several others (Ariumi, 2022), while ribonucleoprotein condensates are regulated by the phase-separation of the viral N protein (Luo et al., 2021). Many other questions have yet to be answered, including the differential modulation of host defenses by distinct SARS-CoV-2 variants and the difference in the infection's response at the individual level.

8.3. Future perspectives

These new developments are in agreement with my predictions and have opened new lines of investigation.
Currently, I am planning to improve the accuracy of the prediction of phase-separating propensity, either by integrating new proteomic data now

available in the literature or by developing an algorithm for predicting the RNA ability to be recruited inside condensates, that is currently unavailable in the field.
I am also contributing to the experimental validation of other host interactors that could be binding to SARS-CoV-2 in order to study other potential host mechanisms of modulation of viral infectivity.
At the same time, I am involved in the analysis of other pathogens' infection pathways, following a similar protocol adopted for SARS-CoV-2. On the one hand, I am investigating the family of Flaviviridae, composed of a group of single-strand RNA viruses causing hemorrhagic fevers like Dengue and Zika. During the infection process, their genomic RNA is targeted by the host cell's 5'-3' exoribonuclease 1 (XRN1) that stalls at 3'UTR due to the presence of stem-loops-like secondary structures. This results in an undigested fragment called subgenomic flavivirus RNA (sfRNA) that accumulates in the cell and antagonizes the cell's innate immune response (Funk et al., 2010; Chapman et al., 2014). In this context, my work could help in identifying which host proteins are actually targeted and sequestered by these viruses to help their infection process.
On the other hand, I am studying Salmonella Typhimurium, a bacterium causing severe gastrointestinal symptoms. The infection is caused by effector proteins, which are secreted by the bacterium and enter the host cell through endocytosis, where they generate a cascade of reactions and are known to induce the formation of SGs (Abdel-Nour et al., 2019). However, the potential presence of bacterial effector proteins inside these granules has yet to be investigated and, even though some of these proteins have RNA-binding potential and present RNA-binding domains (Gerovac et al., 2020), there is currently no evidence of them binding to host RNAs during the infection process.

I speculate that some of these effectors could phase-separate and behave like the N protein in the context of SARS-CoV-2 infection to modulate the formation and dissolution of SGs. Furthermore, I think that some of these bacterial proteins could bind host RNAs during the infection process and prevent their translation or their scaffolding ability to form condensates. This is why I propose to use an approach similar to the one adopted for SARS-CoV-2 but focusing on the prediction of the human transcriptome bound to Salmonella effectors. This analysis and the assessment of the RNA-binding ability of these effectors proteins are already pointing to some promising candidates showing high phase-separating potential which will be tested experimentally.

CHAPTER 9

CONCLUSIONS

The work carried out during my Ph.D. can be divided into two different stages. The first focused on the characterization of the molecular network of interactions inside liquid-like phase-separating condensates such as SGs and PBs, followed by the creation of a database on the physicochemical properties of these organelles and amyloid formations, including information collected from experimental and computational high-throughput data and focusing on the impact that disease-related SNVs have on their components.
In the second stage, I investigated the molecular interactions established between SARS-CoV-2 and the human host and identified the viral regions acting as attractors of the majority of the binders. I predicted a complete human-SARS-CoV-2 protein-RNA interactome while analyzing structural conserved elements shared among different coronavirus species and identified several human proteins that could be targeted by the virus to enhance its replication process and to avoid storage inside phase-separating organelles. Finally, I compared different experimental protein-RNA interactomes already published, investigating the scarcity of conserved interactors across different studies and the link between binding strength and phase-separation propensity of such binders.

Collectively, the main findings of my thesis can be summarized in the following points:

- Stress granules and processing bodies are enriched in poorly structured elements that show a high degree of interactivity. Single-stranded RNAs enriched in condensates show a high degree of RNA-RNA interactions, while granule proteins enriched in disordered regions form a large number of contacts with other protein partners.

These two molecular sets identified as the most interacting at the proteome and transcriptome level are both depleted in structure and bind one to the other, creating a circular pattern of macromolecular interactions that regulate the condensates.

- PRALINE is the first database to combine physicochemical properties of phase-separating condensates with disease-related single-nucleotide variants information, including high-throughput experimental and computational data to describe their components.

- The genomic locus corresponding to nucleotides 22000-23000 is highly conserved at the sequence and structural level and shared among several coronavirus species, while the region upstream is highly variable. These regions code for a viral spike S protein domain which binds to the human ACE2 receptor and favors the interaction with sialic acids in MERS-CoV respiratory syndrome. This variability could be one of the different possible causes for the differences in symptoms across the human population.

- The 5' end of SARS-CoV-2 is a structured RNA region predicted to be an attractor for many different human proteins, among which there are many granules components and members of the innate immune response machinery that the virus could sequester to help its own replication process.

- Only strong-affinity binders to the SARS-CoV-2 genome can be found in multiple human-virus protein-RNA interactome experiments. These proteins have also a high propensity to phase separate and were found in infection processes involving several other

viruses. This indicates that these strong interactors could act as primary attractors and be exploited by SARS-CoV-2 to modulate its viral replication.

Part IV

Appendix

APPENDIX A

Supplementary Materials of Chapter 4

Supplementary Materials of Chapter 4 are available at https://www.sciencedirect.com/science/article/pii/S0022283621003880?via%3Dihub#s0145

APPENDIX B

Supplementary Materials of Chapter 6

Supplementary Materials of Chapter 6 are available at https://academic.oup.com/nar/article/48/20/11270/5929227#supplementary-data

APPENDIX C

Supplementary Materials of Chapter 7

Supplementary Materials of Chapter 7 are available at https://www.frontiersin.org/articles/10.3389/fmolb.2022.893067/full#SM1

APPENDIX D

List of Publications

1. Vecchi, G., Sormanni, P., Mannini, B., **Vandelli, A**, Tartaglia, G. G., Dobson, C. M., Hartl, F. U., and Vendruscolo, M. (2020). Proteome-wide observation of the phenomenon of life on the edge of solubility. *Proceedings of the National Academy of Sciences of the United States of America*, 117(2):1015–1020. DOI: 10.1073/pnas.1910444117

2. Ponti, R. D., Armaos, A., **Vandelli, A**, and Tartaglia, G. G. (2020). CROSSalive: a web server for predicting the *in vivo* structure of RNA molecules. *Bioinformatics (Oxford, England)*, 36(3):940–941. DOI: 10.1093/bioinformatics/btz666

3. Colantoni, A., Rupert, J., **Vandelli, A**, Tartaglia, G. G., and Zacco, E. (2020). Zooming in on protein-RNA interactions: a multilevel workflow to identify interaction partners. *Biochemical Society Transactions*, 48(4):1529–1543. DOI: 10.1042/BST20191059

4. **Vandelli, A**, Monti, M., Milanetti, E., Armaos, A., Rupert, J., Zacco, E., Bechara, E., Delli Ponti, R., and Tartaglia, G. G. (2020). Structural analysis of SARS-CoV-2 genome and predictions of the human interactome. *Nucleic Acids Research*, 48(20):11270–11283. DOI: 10.1093/nar/gkaa864

5. Battistelli, C., Garbo, S., Riccioni, V., Montaldo, C., Santangelo, L., **Vandelli, A**, Strippoli, R., Tartaglia, G. G., Tripodi, M., and Cicchini, C. (2021). Design and Functional Validation of a Mutant Variant of the LncRNA HOTAIR to Counteract Snail Function in Epithelial-to-Mesenchymal Transition. *Cancer Research*, 81(1):103–113. DOI: 10.1158/0008-5472.CAN-20-1764

6. **Vandelli, A**, Vocino, G., and Tartaglia, G. G. (2022). Phase Separation Drives SARS-CoV-2 Replication: A Hypothesis. *Frontiers in Molecular Biosciences*, 9:893067.
DOI: 10.3389/fmolb.2022.893067

7. **Vandelli, A**, Cid Samper, F., Torrent Burgas, M., Sanchez de Groot, N., and Tartaglia, G. G. (2022). The Interplay Between Disordered Regions in RNAs and Proteins Modulates Interactions Within Stress Granules and Processing Bodies. *Journal of Molecular Biology*, 434(1):167159. DOI: 10.1016/j.jmb.2021.167159

8. Ponti, R. D., Broglia, L., **Vandelli, A**, Armaos, A., Burgas, M. T., Sanchez de Groot, N., and Tartaglia, G. G. (2022) A high-throughput approach to predict A-to-I effects on RNA structure indicates a change of double-stranded content in non-coding RNAs. *IUBMB Life*, iub.2673. DOI: 10.1002/iub.2673

9. Giambruno, R., Zacco, E., Ugolini, C., **Vandelli, A.**, Mulroney, L., D'Onghia, M., Criscuolo, E., Castelli, M., Clementi, N., Clementi, M., Mancini, N., Bonaldi, T., Gustincich, S., Leonardi, T., Tartaglia, G. G., and Nicassio, F. (2022). Discovering host protein interactions specific for SARS-CoV-2 RNA genome. *preprint*, Molecular Biology. DOI: 10.1101/2022.07.18.499583

10. **Vandelli, A.**, Arnal Segura, M., Monti, M., Fiorentino, J., Broglia, L., Colantoni, A., Sanchez de Groot, N., Torrent Burgas, M., Armaos, A., and Tartaglia, G. G. (2022). The PRALINE database: Protein and Rna humAn singLe nucleotIde variaNts in condEnsates. *preprint*, Bioinformatics. DOI: 10.1101/2022.12.03.518982

APPENDIX E

Posters

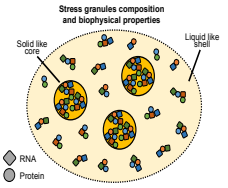
The hidden side of stress granules

Andrea Vandelli, Fernando Cid Samper,
Natalia Sanchez de Groot, Gian Gaetano Tartaglia

Background

Context

Stress granules (SGs) are membrane-less compartments formed by a process of phase separation in stress condition of the cell.[1,2] The translation of mRNA is slow or stopped in SGs. Their main components are proteins and RNAs, which assembly results in the formation of several dense solid-like cores surrounded by liquid-like shell.



The problem

- Only one SG transcriptome and few SG proteomes has been published so far.
- Our knowledge about the interactions between the molecules inside SG is poor.

The aim

- Analyze the SGs interaction network to understand the molecular contacts involved in its formation and arrangement.

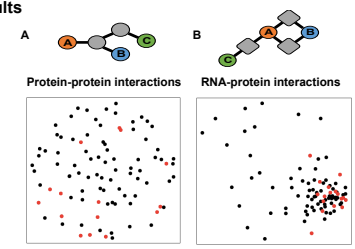
Results

1) RNAs are involved in SG proteins arrangement

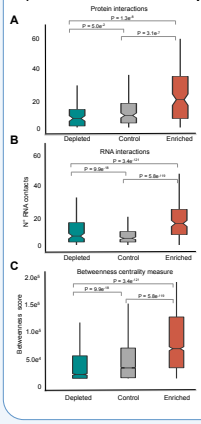
RNA-protein contacts are able to cluster together SG proteins better than protein-protein contacts.

(**A**) Proteins clustered by the protein-protein network, interactions obtained from BioGRID. In red are represented the proteins that are constitutively present in SG, in black the non granule proteins.

(**B**) Proteins clustered by the protein-RNA network, interactions obtained from the eCLIP dataset[4]. In both plots, the interaction between each pair of proteins is represented by the Jaccard distance, calculated as: intersection of protein targets/union of protein targets, subtracted to 1. A similar result is obtained by the use of shortest path length as a measure of distance (data not shown). In the schematic representations squares are RNAs and circles are proteins.



2) Enriched RNAs form a pre-network of interactions

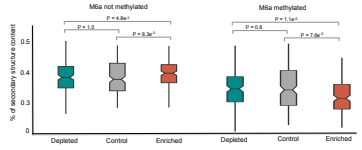


Cell diagram: SGs RNAs have been classified in 3 group accordingly to their abundance in respect to the total RNAs in the cell: enriched (more in the SGs), depleted (less in the SGs) and control (no significant).[2]

Enriched RNAs are the best binders for both proteins and RNAs in vivo. (**A**) Number of protein interactions from eCLIP. (**B**) Number of RNA interactions from RISE[5]. Both in non-stress conditions.

Enriched RNAs are hubs of the network of RNA-RNA interactions in the cell.
(**C**) Betweenness centrality score for each of the three classes of SG RNAs. It calculates the shortest path between every pair of nodes in a connected graph (RNA-RNA network). Each node (RNA) receives a score based on the number of shortest paths passing through the node. Nodes that lies more frequently on these path have higher scores.

3) Enriched RNAs m6a modification enhances linearity



The boxplots show the percentage of secondary structure content, for the three classes of RNAs in SG, predicted by CROSSalive[6] algorithm in presence or absence of m6a methylation. This algorithm, developed in our lab, predicts the secondary structure in vivo using contribution from RNA sequence and protein partners interactions. It is trained on icSHAPE data.

It has been reported that **M6a methylation enhances the phase separation potential of mRNAs**[3]. The comparison between the above left and right boxplots show that **RNAs undergoing this modification become more linear and prone to bind other RNAs**, two properties associated with the SGs material state.

CROSS Alive

http://s.tartaglialab.com/update_submission/227435/d043ccf62a

Discussion

- RNAs are important for proteins arrangement in SG formation.
- Enriched SG RNAs pre-network of interactions can favor SG formation (e.g. acceleration).
- Enriched RNAs linearity is enhanced after m6a modification which may be crucial to keep the SGs liquid state.

References

[1] D. S W Protter and Roy Parker. Principles and Properties of Stress Granules. Trends in Cell Biology. 2016.
[2] A. Khong, et al. The Stress Granule Transcriptome Reveals Principles of mRNA Accumulation in Stress Granules. Molecular Cell. 2017.
[3] R.J Ries, et al. M6a enhances the phase separation potential of mRNA. Nature. 2019
[4] E. L Van Nostrand, et al. Robust transcriptome-wide discovery of RNA-binding protein binding sites with enhanced CLIP (eCLIP). Nature Methods. 2016.
[5] J. Gong, et al. RISE: a database of RNA interactome from sequencing experiments. NAR. 2018
[6] R. Delli Ponti et. Al. CROSSalive: a web server for predicting the in vivo structure of RNA molecules. Bioinformatics. 2019

Structural Analysis of SARS-CoV-2 genome and predictions of the human interactome[1]

Andrea Vandelli
Systems Biology of infection lab, Universitat Autonoma de Barcelona (UAB)

Introduction

SARS-CoV-2 is a positive-sense single-stranded RNA virus similar to other beta-coronavirus. It has a 30k bp genome, codifying for 4 structural proteins: S, E, M and N.

Spike S protein is fundamental for the viral infection. It interacts with the human ACE2 receptor[2]. It is the target of the available vaccines.

Some studies suggest that the virus can then interact with other human proteins, and can exploit mechanisms like phase-separation to alter the regulation of host cells[3].

Our study aims to:

- Analyze the level of sequence and secondary structure conservation of Spike S genomic locus across different coronavirus and SARS-CoV-2 strains.
- Predict interactions between SARS-CoV-2 genome and human proteome and analyze potentially valid candidates.

Methods

Different algorithms were used in the different sections of the analysis.

1) Secondary structure conservation among different coronavirus

CROSSalign[4] ⟶ Structural alignment of 2040 coronavirus genomes (reduced to 267 sequences upon redundancy removal at 95% sequence similarity)

2) Sequence and secondary structure conservation among SARS-CoV-2 strains

Clustal W[5] ⟶ Multiple sequence alignment of ~60 genomes of SARS-CoV-2 strains from different countries

CROSSalign ⟶ Structural alignment of the same ~60 genomes

T-Coffee[6] ⟶ Multiple sequence alignment of Spike S protein sequences the ~60 strains

3) Prediction and selection of RNA-protein interactions

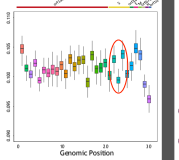
CatRAPID omics[7] ⟶ Prediction of viral genome (divided in 30 fragments) vs human RNA binding proteome

CatGRANULE[8] ⟶ Prediction of phase separating propensity of best (Z>1.5) vs worst (Z<1.5) protein interactors

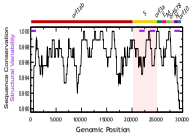
Results

1) Secondary structure conservation among different coronavirus

The viral genomic region (nt 22000-23000) containing ACE2 binding domain has relatively high structural conservation



2) Sequence and secondary structure conservation among SARS-CoV-2 strains

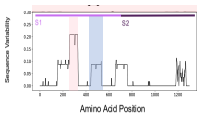


ACE2 binding domain genomic region (22000-23000) shows high sequence and structural conservation

The flanking genomic regions (21000-22000, 23000-24000) show high sequence and structural variability

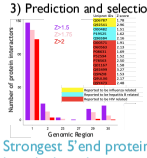
ACE2 binding domain (in blue) amino acid composition is relatively conserved

Sialic acid binding region (in pink) amino acid composition shows high variability



⟶ *Possible cause of differential infection behaviour across individuals*

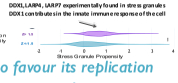
3) Prediction and selection of RNA-protein interactions



>120 proteins predicted to bind the 5' end of SARS-CoV-2 genome

The 5' end is enriched in host interactions also implicated in other viral infections

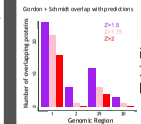
Strongest 5'end protein interactors have higher phase-separating propensity



⟶ *SARS-CoV-2 hijacks stress granules to favour its replication and protect itself from the cellular immune response*

Experimental Validation

2 datasets on host-virus interactions:



1) 332 host-virus high-confidence protein-protein interactions[9].

2) 570 host-virus protein-RNA interactions in HuH7 cells through RAP-MS approach[10].

- *High overlap with predicted protein interactions*
- *Strong agreement between experimental fold-change and predicted Z score (AUC = 0.72 at 25% confidence score)*

Conclusions

- *SARS-CoV-2 infection is a very complex reality, that can vary across individuals maybe due to the high variability of its genomic region containing sialic acid binding domain.*
- *SARS-CoV-2 can exploit host cellular mechanisms, such as phase separation, to promote its replication and protect itself from the cellular immune response.*
- *Some of our findings are in agreement with the latest literature and have been proven experimentally.*

References

1) Vandelli A. et al. (2020) *Structural analysis of SARS-CoV-2 genome and predictions of the human interactome.* Nucleic acids research, 48, 11270-11283
2) Park, Y. et al. (2019) *Structure of MERS-CoV spike glycoprotein in complex with sialoside attachment receptors.* Nat. Struct. Mol. Biol., 26, 1151–1157.
3) Yang, J. et al. (2021) *SARS-CoV-2 nucleocapsid protein undergoes liquid-liquid phase separation into stress granules through its N-terminal intrinsically disordered region.* Cell Discov. 7.5
4) Chelli Piovesi, R. et al. (2017) *A high-throughput approach to profile RNA structure.* Nucleic Acids Res., 45, e35.
5) Madeira, F. et al. (2019) *The EMBL-EBI search and sequence analysis tools APIs in 2019.* Nucleic Acids Res., 47, W636–W641
6) Di Tommaso, P. et al. (2011) *T-Coffee: a web server for the multiple sequence alignment of protein and RNA sequences using structural information and homology extension.* Nucleic Acids Res., 39, W1 3–W1 7
7) Agostini, F. et al. (2013) *catRAPID omics: a web server for large-scale prediction of protein-RNA interactions.* Bioinformatics, 29, 2928–2930
8) Bolognesi, B. et al. (2016) *A Concentration-Dependent Liquid Phase Separation Can Cause Toxicity upon Increased Protein Expression.* Cell Rep., 16, 222–231
9 Gordon, D.E. et al. (2020) *A SARS-CoV-2 protein interaction map reveals targets for drug repurposing.* Nature, 583, 459–468
10) Schmidt, N. et al. (2021) *The SARS-CoV-2 RNA–protein interactome in infected human cells.* Nat Microbiol 6, 339–353

GOBIERNO DE ESPAÑA · MINISTERIO DE ECONOMÍA Y COMPETITIVIDAD

UAB Universitat Autònoma de Barcelona

Andrea.Vandelli@uab.cat

Bibliography

Abdel-Nour, M., Carneiro, L. A. M., Downey, J., Tsalikis, J., Outlioua, A., Prescott, D., Da Costa, L. S., Hovingh, E. S., Farahvash, A., Gaudet, R. G., Molinaro, R., van Dalen, R., Lau, C. C. Y., Azimi, F. C., Escalante, N. K., Trotman-Grant, A., Lee, J. E., Gray-Owen, S. D., Divangahi, M., Chen, J.-J., Philpott, D. J., Arnoult, D., and Girardin, S. E. (2019). The heme-regulated inhibitor is a cytosolic sensor of protein misfolding that controls innate immune signaling. *Science*, 365(6448):eaaw4144.

Alberti, S., Gladfelter, A., and Mittag, T. (2019). Considerations and Challenges in Studying Liquid-Liquid Phase Separation and Biomolecular Condensates. *Cell*, 176(3):419–434.

Allain, F. H., Bouvet, P., Dieckmann, T., and Feigon, J. (2000). Molec-

ular basis of sequence-specific recognition of pre-ribosomal RNA by nucleolin. *The EMBO journal*, 19(24):6870–6881.

Alqahtani, W., Alneghery, I., Alqahtani, A., ALKahtani, M., Alkahtani, S., and Princess Nourah bint Abdulrahman University (2020). A review of comparison study between Corona Viruses (SARS-CoV, MERS-CoV) and Novel Corona Virus (Covid-19). *Revista Mexicana de Ingeniería Química*, 19(1):201–212.

Alriquet, M., Calloni, G., Martínez-Limón, A., Delli Ponti, R., Hanspach, G., Hengesbach, M., Tartaglia, G. G., and Vabulas, R. M. (2021). The protective role of m1A during stress-induced granulation. *Journal of Molecular Cell Biology*, 12(11):870–880.

Anders, M., Chelysheva, I., Goebel, I., Trenkner, T., Zhou, J., Mao, Y., Verzini, S., Qian, S.-B., and Ignatova, Z. (2018). Dynamic m ^{6} A methylation facilitates mRNA triaging to stress granules. *Life Science Alliance*, 1(4):e201800113.

Anderson, P. and Kedersha, N. (2006). RNA granules. *Journal of Cell Biology*, 172(6):803–808.

Anderson, P., Kedersha, N., and Ivanov, P. (2015). Stress granules, P-bodies and cancer. *Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms*, 1849(7):861–870.

Andrei, M. A., Ingelfinger, D., Heintzmann, R., Achsel, T., Rivera-Pomar, R., and Lührmann, R. (2005). A role for eIF4E and eIF4E-transporter in targeting mRNPs to mammalian processing bodies. *RNA (New York, N.Y.)*, 11(5):717–727.

Antson, A. A., Dodson, E. J., Dodson, G., Greaves, R. B., Chen, X., and Gollnick, P. (1999). Structure of the trp RNA-binding attenuation protein, TRAP, bound to RNA. *Nature*, 401(6750):235–242.

Araya, N., Hiraga, H., Kako, K., Arao, Y., Kato, S., and Fukamizu, A. (2005). Transcriptional down-regulation through nuclear exclusion of EWS methylated by PRMT1. *Biochemical and Biophysical Research Communications*, 329(2):653–660.

Arguello, A. E., DeLiberto, A. N., and Kleiner, R. E. (2017). RNA Chemical Proteomics Reveals the N ^{6} -Methyladenosine (m ^{6} A)-Regulated Protein–RNA Interactome. *Journal of the American Chemical Society*, 139(48):17249–17252.

Arimoto, K., Fukuda, H., Imajoh-Ohmi, S., Saito, H., and Takekawa, M. (2008). Formation of stress granules inhibits apoptosis by suppressing stress-responsive MAPK pathways. *Nature Cell Biology*, 10(11):1324–1332.

Ariumi, Y. (2022). Host Cellular RNA Helicases Regulate SARS-CoV-2 Infection. *Journal of Virology*, 96(6):e0000222.

Ariumi, Y., Kuroki, M., Kushima, Y., Osugi, K., Hijikata, M., Maki, M., Ikeda, M., and Kato, N. (2011). Hepatitis C Virus Hijacks P-Body and Stress Granule Components around Lipid Droplets. *Journal of Virology*, 85(14):6882–6892.

Armaos, A., Colantoni, A., Proietti, G., Rupert, J., and Tartaglia, G. (2021). *cat* RAPID *omics v2.0* : going deeper and wider in the prediction of protein–RNA interactions. *Nucleic Acids Research*, 49(W1):W72–W79.

Aulas, A., Lyons, S. M., Fay, M. M., Anderson, P., and Ivanov, P. (2018). Nitric oxide triggers the assembly of "type II" stress granules linked to decreased cell viability. *Cell Death & Disease*, 9(11):1129.

Auweter, S. D., Oberstrass, F. C., and Allain, F. H.-T. (2006). Sequence-specific binding of single-stranded RNA: is there a code for recognition? *Nucleic Acids Research*, 34(17):4943–4959.

Aw, J. G. A., Shen, Y., Wilm, A., Sun, M., Lim, X. N., Boon, K.-L., Tapsin, S., Chan, Y.-S., Tan, C.-P., Sim, A. Y. L., Zhang, T., Susanto, T. T., Fu, Z., Nagarajan, N., and Wan, Y. (2016). In Vivo Mapping of Eukaryotic RNA Interactomes Reveals Principles of Higher-Order Organization and Regulation. *Molecular Cell*, 62(4):603–617.

Azaldegui, C. A., Vecchiarelli, A. G., and Biteen, J. S. (2021). The emergence of phase separation as an organizing principle in bacteria. *Biophysical Journal*, 120(7):1123–1138.

Backe, P. H., Messias, A. C., Ravelli, R. B. G., Sattler, M., and Cusack, S. (2005). X-ray crystallographic and NMR studies of the third KH domain of hnRNP K in complex with single-stranded nucleic acids. *Structure (London, England: 1993)*, 13(7):1055–1067.

Bak, G., Han, K., Kim, K.-s., and Lee, Y. (2015). Electrophoretic mobility shift assay of RNA-RNA complexes. *Methods in Molecular Biology (Clifton, N.J.)*, 1240:153–163.

Bellucci, M., Agostini, F., Masin, M., and Tartaglia, G. G. (2011). Predicting protein associations with long noncoding RNAs. *Nature Methods*, 8(6):444–445.

Benatar, M., Wuu, J., Fernandez, C., Weihl, C. C., Katzen, H., Steele, J., Oskarsson, B., and Taylor, J. P. (2013). Motor neuron involvement in multisystem proteinopathy: Implications for ALS. *Neurology*, 80(20):1874–1880.

Berg, J. M., Tymoczko, J. L., Stryer, L., and Stryer, L. (2002). *Biochemistry*. W.H. Freeman, New York, 5th ed edition.

Bhattacharyya, S. N., Habermacher, R., Martine, U., Closs, E. I., and Filipowicz, W. (2006). Relief of microRNA-mediated translational repression in human cells subjected to stress. *Cell*, 125(6):1111–1124.

Boisvert, F.-M., van Koningsbruggen, S., Navascués, J., and Lamond, A. I. (2007). The multifunctional nucleolus. *Nature Reviews Molecular Cell Biology*, 8(7):574–585.

Bolognesi, B., Lorenzo Gotor, N., Dhar, R., Cirillo, D., Baldrighi, M., Tartaglia, G. G., and Lehner, B. (2016). A Concentration-Dependent Liquid Phase Separation Can Cause Toxicity upon Increased Protein Expression. *Cell Reports*, 16(1):222–231.

Brooijmans, N., Sharp, K. A., and Kuntz, I. D. (2002). Stability of macromolecular complexes. *Proteins*, 48(4):645–653.

Buchan, J. R. and Parker, R. (2009). Eukaryotic stress granules: the ins and outs of translation. *Molecular Cell*, 36(6):932–941.

Bycroft, M., Hubbard, T. J., Proctor, M., Freund, S. M., and Murzin, A. G. (1997). The solution structure of the S1 RNA binding domain: a member of an ancient nucleic acid-binding fold. *Cell*, 88(2):235–242.

Cai, X., Chen, J., Xu, H., Liu, S., Jiang, Q.-X., Halfmann, R., and Chen, Z. J. (2014). Prion-like polymerization underlies signal transduction in antiviral immune defense and inflammasome activation. *Cell*, 156(6):1207–1222.

Cai, Z., Cao, C., Ji, L., Ye, R., Wang, D., Xia, C., Wang, S., Du, Z., Hu, N., Yu, X., Chen, J., Wang, L., Yang, X., He, S., and Xue, Y. (2020). RIC-seq for global in situ profiling of RNA–RNA spatial interactions. *Nature*, 582(7812):432–437.

Campos-Melo, D., Hawley, Z. C. E., Droppelmann, C. A., and Strong, M. J. (2021). The Integral Role of RNA in Stress Granule Formation and Function. *Frontiers in Cell and Developmental Biology*, 9:621779.

Carballo, E., Lai, W. S., and Blackshear, P. J. (1998). Feedback inhibition of macrophage tumor necrosis factor-alpha production by tristetraprolin. *Science (New York, N.Y.)*, 281(5379):1001–1005.

Cascarina, S. M. and Ross, E. D. (2022). Phase separation by the SARS-CoV-2 nucleocapsid protein: Consensus and open questions. *The Journal of Biological Chemistry*, 298(3):101677.

Castello, A., Fischer, B., Eichelbaum, K., Horos, R., Beckmann, B. M., Strein, C., Davey, N. E., Humphreys, D. T., Preiss, T., Steinmetz, L. M., Krijgsveld, J., and Hentze, M. W. (2012). Insights into RNA biology from an atlas of mammalian mRNA-binding proteins. *Cell*, 149(6):1393–1406.

Cerase, A., Armaos, A., Neumayer, C., Avner, P., Guttman, M., and Tartaglia, G. G. (2019). Phase separation drives X-chromosome in-

activation: a hypothesis. *Nature Structural & Molecular Biology*, 26(5):331–334.

Chapman, E. G., Costantino, D. A., Rabe, J. L., Moon, S. L., Wilusz, J., Nix, J. C., and Kieft, J. S. (2014). The structural basis of pathogenic subgenomic flavivirus RNA (sfRNA) production. *Science (New York, N.Y.)*, 344(6181):307–310.

Chiti, F. and Dobson, C. M. (2017). Protein Misfolding, Amyloid Formation, and Human Disease: A Summary of Progress Over the Last Decade. *Annual Review of Biochemistry*, 86(1):27–68.

Chothia, C. and Janin, J. (1975). Principles of protein-protein recognition. *Nature*, 256(5520):705–708.

Cid-Samper, F., Gelabert-Baldrich, M., Lang, B., Lorenzo-Gotor, N., Ponti, R. D., Severijnen, L.-A. W., Bolognesi, B., Gelpi, E., Hukema, R. K., Botta-Orfila, T., and 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.e7.

Conrad, T., Albrecht, A.-S., de Melo Costa, V. R., Sauer, S., Meierhofer, D., and Ørom, U. A. (2016). Serial interactome capture of the human cell nucleus. *Nature Communications*, 7(1):11212.

Da Cruz, S. and Cleveland, D. W. (2011). Understanding the role of TDP-43 and FUS/TLS in ALS and beyond. *Current Opinion in Neurobiology*, 21(6):904–919.

Decker, C. J. and Parker, R. (2012). P-bodies and stress granules: possible roles in the control of translation and mRNA degradation. *Cold Spring Harbor Perspectives in Biology*, 4(9):a012286.

Delli Ponti, R., Armaos, A., Marti, S., and Tartaglia, G. G. (2018). A Method for RNA Structure Prediction Shows Evidence for Structure in lncRNAs. *Frontiers in Molecular Biosciences*, 5:111.

Deo, R. C., Bonanno, J. B., Sonenberg, N., and Burley, S. K. (1999). Recognition of polyadenylate RNA by the poly(A)-binding protein. *Cell*, 98(6):835–845.

Di Primo, C., Dausse, E., and Toulmé, J.-J. (2011). Surface plasmon resonance investigation of RNA aptamer-RNA ligand interactions. *Methods in Molecular Biology (Clifton, N.J.)*, 764:279–300.

Ding, Y., Tang, Y., Kwok, C. K., Zhang, Y., Bevilacqua, P. C., and Assmann, S. M. (2014). In vivo genome-wide profiling of RNA secondary structure reveals novel regulatory features. *Nature*, 505(7485):696–700.

Dou, N., Chen, J., Yu, S., Gao, Y., and Li, Y. (2016). G3BP1 contributes to tumor metastasis via upregulation of Slug expression in hepatocellular carcinoma. *American Journal of Cancer Research*, 6(11):2641–2650.

Draper, D. E. (1999). Themes in RNA-protein recognition. *Journal of Molecular Biology*, 293(2):255–270.

Eisinger-Mathason, T. K., Andrade, J., Groehler, A. L., Clark, D. E., Muratore-Schroeder, T. L., Pasic, L., Smith, J. A., Shabanowitz, J.,

Hunt, D. F., Macara, I. G., and Lannigan, D. A. (2008). Codependent Functions of RSK2 and the Apoptosis-Promoting Factor TIA-1 in Stress Granule Assembly and Cell Survival. *Molecular Cell*, 31(5):722–736.

Engreitz, J., Sirokman, K., McDonel, P., Shishkin, A. A., Surka, C., Russell, P., Grossman, S., Chow, A., Guttman, M., and Lander, E. (2014). RNA-RNA Interactions Enable Specific Targeting of Noncoding RNAs to Nascent Pre-mRNAs and Chromatin Sites. *Cell*, 159(1):188–199.

Flynn, R. A., Belk, J. A., Qi, Y., Yasumoto, Y., Wei, J., Alfajaro, M. M., Shi, Q., Mumbach, M. R., Limaye, A., DeWeirdt, P. C., Schmitz, C. O., Parker, K. R., Woo, E., Chang, H. Y., Horvath, T. L., Carette, J. E., Bertozzi, C. R., Wilen, C. B., and Satpathy, A. T. (2021). Discovery and functional interrogation of SARS-CoV-2 RNA-host protein interactions. *Cell*, 184(9):2394–2411.e16.

Fox, A. H., Lam, Y. W., Leung, A. K., Lyon, C. E., Andersen, J., Mann, M., and Lamond, A. I. (2002). Paraspeckles. *Current Biology*, 12(1):13–25.

Franks, T. M. and Lykke-Andersen, J. (2008). The Control of mRNA Decapping and P-Body Formation. *Molecular Cell*, 32(5):605–615.

Fujimura, K., Sasaki, A. T., and Anderson, P. (2012). Selenite targets eIF4E-binding protein-1 to inhibit translation initiation and induce the assembly of non-canonical stress granules. *Nucleic Acids Research*, 40(16):8099–8110.

Funk, A., Truong, K., Nagasaki, T., Torres, S., Floden, N., Balmori Melian, E., Edmonds, J., Dong, H., Shi, P.-Y., and Khromykh,

A. A. (2010). RNA structures required for production of subgenomic flavivirus RNA. *Journal of Virology*, 84(21):11407–11417.

Gaete-Argel, A., Márquez, C. L., Barriga, G. P., Soto-Rifo, R., and Valiente-Echeverría, F. (2019). Strategies for Success. Viral Infections and Membraneless Organelles. *Frontiers in Cellular and Infection Microbiology*, 9:336.

Gall, J. G., Bellini, M., Wu, Z., and Murphy, C. (1999). Assembly of the Nuclear Transcription and Processing Machinery: Cajal Bodies (Coiled Bodies) and Transcriptosomes. *Molecular Biology of the Cell*, 10(12):4385–4402.

Garaigorta, U., Heim, M. H., Boyd, B., Wieland, S., and Chisari, F. V. (2012). Hepatitis C Virus (HCV) Induces Formation of Stress Granules Whose Proteins Regulate HCV RNA Replication and Virus Assembly and Egress. *Journal of Virology*, 86(20):11043–11056.

Garaizar, A., Espinosa, J. R., Joseph, J. A., and Collepardo-Guevara, R. (2022). Kinetic interplay between droplet maturation and coalescence modulates shape of aged protein condensates. *Scientific Reports*, 12(1):4390.

Garrett-Wheeler, E., Lockard, R. E., and Kumar, A. (1984). Mapping of psoralen cross-linked nucleotides in RNA. *Nucleic Acids Research*, 12(7):3405–3424.

Gerovac, M., El Mouali, Y., Kuper, J., Kisker, C., Barquist, L., and Vogel, J. (2020). Global discovery of bacterial RNA-binding proteins by RNase-sensitive gradient profiles reports a new FinO domain protein. *RNA (New York, N.Y.)*, 26(10):1448–1463.

Gilks, N., Kedersha, N., Ayodele, M., Shen, L., Stoecklin, G., Dember, L. M., and Anderson, P. (2004). Stress Granule Assembly Is Mediated by Prion-like Aggregation of TIA-1. *Molecular Biology of the Cell*, 15(12):5383–5398.

Gong, J., Ju, Y., Shao, D., and Zhang, Q. C. (2018). Advances and challenges towards the study of RNA-RNA interactions in a transcriptome-wide scale. *Quantitative Biology*, 6(3):239–252.

Goodsell, D. S. and Olson, A. J. (2000). Structural Symmetry and Protein Function. *Annual Review of Biophysics and Biomolecular Structure*, 29(1):105–153.

Gordon, D. E., Jang, G. M., Bouhaddou, M., Xu, J., Obernier, K., White, K. M., O'Meara, M. J., Rezelj, V. V., Guo, J. Z., Swaney, D. L., Tummino, T. A., Hüttenhain, R., Kaake, R. M., Richards, A. L., Tutuncuoglu, B., Foussard, H., Batra, J., Haas, K., Modak, M., Kim, M., Haas, P., Polacco, B. J., Braberg, H., Fabius, J. M., Eckhardt, M., Soucheray, M., Bennett, M. J., Cakir, M., McGregor, M. J., Li, Q., Meyer, B., Roesch, F., Vallet, T., Mac Kain, A., Miorin, L., Moreno, E., Naing, Z. Z. C., Zhou, Y., Peng, S., Shi, Y., Zhang, Z., Shen, W., Kirby, I. T., Melnyk, J. E., Chorba, J. S., Lou, K., Dai, S. A., Barrio-Hernandez, I., Memon, D., Hernandez-Armenta, C., Lyu, J., Mathy, C. J. P., Perica, T., Pilla, K. B., Ganesan, S. J., Saltzberg, D. J., Rakesh, R., Liu, X., Rosenthal, S. B., Calviello, L., Venkataramanan, S., Liboy-Lugo, J., Lin, Y., Huang, X.-P., Liu, Y., Wankowicz, S. A., Bohn, M., Safari, M., Ugur, F. S., Koh, C., Savar, N. S., Tran, Q. D., Shengjuler, D., Fletcher, S. J., O'Neal, M. C., Cai, Y., Chang, J. C. J., Broadhurst, D. J., Klippsten, S., Sharp, P. P., Wenzell, N. A., Kuzuoglu-Ozturk, D.,

Wang, H.-Y., Trenker, R., Young, J. M., Cavero, D. A., Hiatt, J., Roth, T. L., Rathore, U., Subramanian, A., Noack, J., Hubert, M., Stroud, R. M., Frankel, A. D., Rosenberg, O. S., Verba, K. A., Agard, D. A., Ott, M., Emerman, M., Jura, N., von Zastrow, M., Verdin, E., Ashworth, A., Schwartz, O., d'Enfert, C., Mukherjee, S., Jacobson, M., Malik, H. S., Fujimori, D. G., Ideker, T., Craik, C. S., Floor, S. N., Fraser, J. S., Gross, J. D., Sali, A., Roth, B. L., Ruggero, D., Taunton, J., Kortemme, T., Beltrao, P., Vignuzzi, M., García-Sastre, A., Shokat, K. M., Shoichet, B. K., and Krogan, N. J. (2020). A SARS-CoV-2 protein interaction map reveals targets for drug repurposing. *Nature*, 583(7816):459–468.

Gotor, N. L., Armaos, A., Calloni, G., Torrent Burgas, M., Vabulas, R., De Groot, N. S., and Tartaglia, G. G. (2020). RNA-binding and prion domains: the Yin and Yang of phase separation. *Nucleic Acids Research*, 48(17):9491–9504.

Grabowski, M., Niedzialkowska, E., Zimmerman, M. D., and Minor, W. (2016). The impact of structural genomics: the first quindecennial. *Journal of Structural and Functional Genomics*, 17(1):1–16.

Greer, A. E., Hearing, P., and Ketner, G. (2011). The adenovirus E4 11k protein binds and relocalizes the cytoplasmic P-body component Ddx6 to aggresomes. *Virology*, 417(1):161–168.

Grishin, N. V. (2001). KH domain: one motif, two folds. *Nucleic Acids Research*, 29(3):638–643.

Handa, N., Nureki, O., Kurimoto, K., Kim, I., Sakamoto, H., Shimura, Y., Muto, Y., and Yokoyama, S. (1999). Structural basis for recog-

nition of the tra mRNA precursor by the Sex-lethal protein. *Nature*, 398(6728):579–585.

Hanson, K. K. and Mair, G. R. (2014). Stress granules and *Plasmodium* liver stage infection. *Biology Open*, 3(1):103–107.

Harris, M. E. and Christian, E. L. (2009). RNA crosslinking methods. *Methods in Enzymology*, 468:127–146.

Hennig, S., Kong, G., Mannen, T., Sadowska, A., Kobelke, S., Blythe, A., Knott, G. J., Iyer, K. S., Ho, D., Newcombe, E. A., Hosoki, K., Goshima, N., Kawaguchi, T., Hatters, D., Trinkle-Mulcahy, L., Hirose, T., Bond, C. S., and Fox, A. H. (2015). Prion-like domains in RNA binding proteins are essential for building subnuclear paraspeckles. *Journal of Cell Biology*, 210(4):529–539.

Hnisz, D., Shrinivas, K., Young, R. A., Chakraborty, A. K., and Sharp, P. A. (2017). A Phase Separation Model for Transcriptional Control. *Cell*, 169(1):13–23.

Hou, F., Sun, L., Zheng, H., Skaug, B., Jiang, Q.-X., and Chen, Z. J. (2011). MAVS forms functional prion-like aggregates to activate and propagate antiviral innate immune response. *Cell*, 146(3):448–461.

Huang, C., Lokugamage, K. G., Rozovics, J. M., Narayanan, K., Semler, B. L., and Makino, S. (2011). SARS Coronavirus nsp1 Protein Induces Template-Dependent Endonucleolytic Cleavage of mRNAs: Viral mRNAs Are Resistant to nsp1-Induced RNA Cleavage. *PLoS Pathogens*, 7(12):e1002433.

Hubstenberger, A., Courel, M., Bénard, M., Souquere, S., Ernoult-Lange, M., Chouaib, R., Yi, Z., Morlot, J.-B., Munier, A., Fradet, M., Daunesse, M., Bertrand, E., Pierron, G., Mozziconacci, J., Kress, M., and Weil, D. (2017). P-Body Purification Reveals the Condensation of Repressed mRNA Regulons. *Molecular Cell*, 68(1):144–157.e5.

Hyman, A. A., Weber, C. A., and Jülicher, F. (2014). Liquid-Liquid Phase Separation in Biology. *Annual Review of Cell and Developmental Biology*, 30(1):39–58.

Incarnato, D., Neri, F., Anselmi, F., and Oliviero, S. (2014). Genome-wide profiling of mouse RNA secondary structures reveals key features of the mammalian transcriptome. *Genome Biology*, 15(10):491.

Jain, S., Wheeler, J. R., Walters, R. W., Agrawal, A., Barsic, A., and Parker, R. (2016). ATPase-Modulated Stress Granules Contain a Diverse Proteome and Substructure. *Cell*, 164(3):487–498.

Janin, J., Miller, S., and Chothia, C. (1988). Surface, subunit interfaces and interior of oligomeric proteins. *Journal of Molecular Biology*, 204(1):155–164.

Johansson, H.-O., Karlström, G., Tjerneld, F., and Haynes, C. A. (1998). Driving forces for phase separation and partitioning in aqueous two-phase systems. *Journal of Chromatography B: Biomedical Sciences and Applications*, 711(1-2):3–17.

Jones, S. (2016). Protein–RNA interactions: structural biology and computational modeling techniques. *Biophysical Reviews*, 8(4):359–367.

Jones, S., Daley, D. T., Luscombe, N. M., Berman, H. M., and Thornton, J. M. (2001). Protein-RNA interactions: a structural analysis. *Nucleic Acids Research*, 29(4):943–954.

Jones, S. and Thornton, J. M. (1995). Protein-protein interactions: a review of protein dimer structures. *Progress in Biophysics and Molecular Biology*, 63(1):31–65.

Jones, S. and Thornton, J. M. (1996). Principles of protein-protein interactions. *Proceedings of the National Academy of Sciences of the United States of America*, 93(1):13–20.

Kamel, W., Noerenberg, M., Cerikan, B., Chen, H., Järvelin, A. I., Kammoun, M., Lee, J. Y., Shuai, N., Garcia-Moreno, M., Andrejeva, A., Deery, M. J., Johnson, N., Neufeldt, C. J., Cortese, M., Knight, M. L., Lilley, K. S., Martinez, J., Davis, I., Bartenschlager, R., Mohammed, S., and Castello, A. (2021). Global analysis of protein-RNA interactions in SARS-CoV-2-infected cells reveals key regulators of infection. *Molecular Cell*, 81(13):2851–2867.e7.

Kedersha, N., Panas, M. D., Achorn, C. A., Lyons, S., Tisdale, S., Hickman, T., Thomas, M., Lieberman, J., McInerney, G. M., Ivanov, P., and Anderson, P. (2016). G3BP–Caprin1–USP10 complexes mediate stress granule condensation and associate with 40S subunits. *Journal of Cell Biology*, 212(7):e201508028.

Kedersha, N., Stoecklin, G., Ayodele, M., Yacono, P., Lykke-Andersen, J., Fritzler, M. J., Scheuner, D., Kaufman, R. J., Golan, D. E., and Anderson, P. (2005). Stress granules and processing bodies are dy-

namically linked sites of mRNP remodeling. *Journal of Cell Biology*, 169(6):871–884.

Kertesz, M., Wan, Y., Mazor, E., Rinn, J. L., Nutter, R. C., Chang, H. Y., and Segal, E. (2010). Genome-wide measurement of RNA secondary structure in yeast. *Nature*, 467(7311):103–107.

Khong, A., Matheny, T., Jain, S., Mitchell, S. F., Wheeler, J. R., and Parker, R. (2017). The Stress Granule Transcriptome Reveals Principles of mRNA Accumulation in Stress Granules. *Molecular Cell*, 68(4):808–820.e5.

Kiebler, M. A. and Bassell, G. J. (2006). Neuronal RNA Granules: Movers and Makers. *Neuron*, 51(6):685–690.

Kim, W. J., Back, S. H., Kim, V., Ryu, I., and Jang, S. K. (2005). Sequestration of TRAF2 into Stress Granules Interrupts Tumor Necrosis Factor Signaling under Stress Conditions. *Molecular and Cellular Biology*, 25(6):2450–2462.

Kleer, M., Mulloy, R. P., Robinson, C.-A., Evseev, D., Bui-Marinos, M. P., Castle, E. L., Banerjee, A., Mubareka, S., Mossman, K., and Corcoran, J. A. (2022). Human coronaviruses disassemble processing bodies. *PLOS Pathogens*, 18(8):e1010724.

Klein, S., Cortese, M., Winter, S. L., Wachsmuth-Melm, M., Neufeldt, C. J., Cerikan, B., Stanifer, M. L., Boulant, S., Bartenschlager, R., and Chlanda, P. (2020). SARS-CoV-2 structure and replication characterized by in situ cryo-electron tomography. *Nature Communications*, 11(1):5885.

Kudla, G., Granneman, S., Hahn, D., Beggs, J. D., and Tollervey, D. (2011). Cross-linking, ligation, and sequencing of hybrids reveals RNA–RNA interactions in yeast. *Proceedings of the National Academy of Sciences*, 108(24):10010–10015.

Kuechler, E. R., Budzyńska, P. M., Bernardini, J. P., Gsponer, J., and Mayor, T. (2020). Distinct Features of Stress Granule Proteins Predict Localization in Membraneless Organelles. *Journal of Molecular Biology*, 432(7):2349–2368.

Kwan, A. H., Mobli, M., Gooley, P. R., King, G. F., and Mackay, J. P. (2011). Macromolecular NMR spectroscopy for the non-spectroscopist. *The FEBS journal*, 278(5):687–703.

Kwiatkowski, T. J., Bosco, D. A., LeClerc, A. L., Tamrazian, E., Vanderburg, C. R., Russ, C., Davis, A., Gilchrist, J., Kasarskis, E. J., Munsat, T., Valdmanis, P., Rouleau, G. A., Hosler, B. A., Cortelli, P., de Jong, P. J., Yoshinaga, Y., Haines, J. L., Pericak-Vance, M. A., Yan, J., Ticozzi, N., Siddique, T., McKenna-Yasek, D., Sapp, P. C., Horvitz, H. R., Landers, J. E., and Brown, R. H. (2009). Mutations in the *FUS/TLS* Gene on Chromosome 16 Cause Familial Amyotrophic Lateral Sclerosis. *Science*, 323(5918):1205–1208.

Kwon, S., Zhang, Y., and Matthias, P. (2007). The deacetylase HDAC6 is a novel critical component of stress granules involved in the stress response. *Genes & Development*, 21(24):3381–3394.

Lai, W. S., Carballo, E., Thorn, J. M., Kennington, E. A., and Blackshear, P. J. (2000). Interactions of CCCH zinc finger proteins with

mRNA. Binding of tristetraprolin-related zinc finger proteins to Au-rich elements and destabilization of mRNA. *The Journal of Biological Chemistry*, 275(23):17827–17837.

Lashuel, H. A., Overk, C. R., Oueslati, A., and Masliah, E. (2013). The many faces of α-synuclein: from structure and toxicity to therapeutic target. *Nature Reviews. Neuroscience*, 14(1):38–48.

Le Ber, I., Van Bortel, I., Nicolas, G., Bouya-Ahmed, K., Camuzat, A., Wallon, D., De Septenville, A., Latouche, M., Lattante, S., Kabashi, E., Jornea, L., Hannequin, D., and Brice, A. (2014). hnRNPA2B1 and hnRNPA1 mutations are rare in patients with "multisystem proteinopathy" and frontotemporal lobar degeneration phenotypes. *Neurobiology of Aging*, 35(4):934.e5–934.e6.

Lee, R. C., Feinbaum, R. L., and Ambros, V. (1993). The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. *Cell*, 75(5):843–854.

Lee, S., Lee, Y.-s., Choi, Y., Son, A., Park, Y., Lee, K.-M., Kim, J., Kim, J.-S., and Kim, V. N. (2021). The SARS-CoV-2 RNA interactome. *Molecular Cell*, 81(13):2838–2850.e6.

Li, F.-F., Zhang, Q., Wang, G.-Y., and Liu, S.-L. (2020). Comparative analysis of SARS-CoV-2 and its receptor ACE2 with evolutionarily related coronaviruses. *Aging*, 12(21):20938–20945.

Li, Y. R., King, O. D., Shorter, J., and Gitler, A. D. (2013). Stress granules as crucibles of ALS pathogenesis. *Journal of Cell Biology*, 201(3):361–372.

Liang, X.-H. and Fournier, M. J. (2006). The helicase Has1p is required for snoRNA release from pre-rRNA. *Molecular and Cellular Biology*, 26(20):7437–7450.

Lin, Y.-H. and Machner, M. P. (2017). Exploitation of the host cell ubiquitin machinery by microbial effector proteins. *Journal of Cell Science*, page jcs.188482.

Ling, S.-C., Polymenidou, M., and Cleveland, D. (2013). Converging Mechanisms in ALS and FTD: Disrupted RNA and Protein Homeostasis. *Neuron*, 79(3):416–438.

Lo Conte, L., Chothia, C., and Janin, J. (1999). The atomic structure of protein-protein recognition sites. *Journal of Molecular Biology*, 285(5):2177–2198.

Lorenzetti, D., Bohlega, S., and Zoghbi, H. Y. (1997). The expansion of the CAG repeat in ataxin-2 is a frequent cause of autosomal dominant spinocerebellar ataxia. *Neurology*, 49(4):1009–1013.

Lu, Z., Zhang, Q. C., Lee, B., Flynn, R. A., Smith, M. A., Robinson, J. T., Davidovich, C., Gooding, A. R., Goodrich, K. J., Mattick, J. S., Mesirov, J. P., Cech, T. R., and Chang, H. Y. (2016). RNA Duplex Map in Living Cells Reveals Higher-Order Transcriptome Structure. *Cell*, 165(5):1267–1279.

Lunde, B. M., Moore, C., and Varani, G. (2007). RNA-binding proteins: modular design for efficient function. *Nature Reviews. Molecular Cell Biology*, 8(6):479–490.

Luo, L., Li, Z., Zhao, T., Ju, X., Ma, P., Jin, B., Zhou, Y., He, S., Huang, J., Xu, X., Zou, Y., Li, P., Liang, A., Liu, J., Chi, T., Huang, X., Ding, Q., Jin, Z., Huang, C., and Zhang, Y. (2021). SARS-CoV-2 nucleocapsid protein phase separates with G3BPs to disassemble stress granules and facilitate viral production. *Science Bulletin*, 66(12):1194–1204.

Luscombe, N. M., Austin, S. E., Berman, H. M., and Thornton, J. M. (2000). An overview of the structures of protein-DNA complexes. *Genome Biology*, 1(1):reviews001.1.

Ma, J.-B., Ye, K., and Patel, D. J. (2004). Structural basis for overhang-specific small interfering RNA recognition by the PAZ domain. *Nature*, 429(6989):318–322.

Machyna, M., Heyn, P., and Neugebauer, K. M. (2013). Cajal bodies: where form meets function: Cajal bodies. *Wiley Interdisciplinary Reviews: RNA*, 4(1):17–34.

Machyna, M., Kehr, S., Straube, K., Kappei, D., Buchholz, F., Butter, F., Ule, J., Hertel, J., Stadler, P., and Neugebauer, K. (2014). The Coilin Interactome Identifies Hundreds of Small Noncoding RNAs that Traffic through Cajal Bodies. *Molecular Cell*, 56(3):389–399.

Macrae, I. J., Zhou, K., Li, F., Repic, A., Brooks, A. N., Cande, W. Z., Adams, P. D., and Doudna, J. A. (2006). Structural basis for double-stranded RNA processing by Dicer. *Science (New York, N.Y.)*, 311(5758):195–198.

Markmiller, S., Soltanieh, S., Server, K. L., Mak, R., Jin, W., Fang, M. Y., Luo, E.-C., Krach, F., Yang, D., Sen, A., Fulzele, A., Wozniak, J. M., Gonzalez, D. J., Kankel, M. W., Gao, F.-B., Bennett, E. J.,

Lécuyer, E., and Yeo, G. W. (2018). Context-Dependent and Disease-Specific Diversity in Protein Interactions within Stress Granules. *Cell*, 172(3):590–604.e13.

Matera, A. (1999). Nuclear bodies: multifaceted subdomains of the interchromatin space. *Trends in Cell Biology*, 9(8):302–309.

Matheny, T., Rao, B. S., and Parker, R. (2019). Transcriptome-Wide Comparison of Stress Granules and P-Bodies Reveals that Translation Plays a Major Role in RNA Partitioning. *Molecular and Cellular Biology*, 39(24):e00313–19.

Matheny, T., Van Treeck, B., Huynh, T. N., and Parker, R. (2021). RNA partitioning into stress granules is based on the summation of multiple interactions. *RNA (New York, N.Y.)*, 27(2):174–189.

McKeith, I., Galasko, D., Kosaka, K., Perry, E., Dickson, D., Hansen, L., Salmon, D., Lowe, J., Mirra, S., Byrne, E., Lennox, G., Quinn, N., Edwardson, J., Ince, P., Bergeron, C., Burns, A., Miller, B., Lovestone, S., Collerton, D., Jansen, E., Ballard, C., de Vos, R., Wilcock, G., Jellinger, K., and Perry, R. (1996). Consensus guidelines for the clinical and pathologic diagnosis of dementia with Lewy bodies (DLB): Report of the consortium on DLB international workshop. *Neurology*, 47(5):1113–1124.

McNab, F., Mayer-Barber, K., Sher, A., Wack, A., and O'Garra, A. (2015). Type I interferons in infectious disease. *Nature Reviews. Immunology*, 15(2):87–103.

Meyer, K. D. and Jaffrey, S. R. (2014). The dynamic epitranscriptome:

N6-methyladenosine and gene expression control. *Nature Reviews Molecular Cell Biology*, 15(5):313–326.

Milne, J. L. S., Borgnia, M. J., Bartesaghi, A., Tran, E. E. H., Earl, L. A., Schauder, D. M., Lengyel, J., Pierson, J., Patwardhan, A., and Subramaniam, S. (2013). Cryo-electron microscopy–a primer for the non-microscopist. *The FEBS journal*, 280(1):28–45.

Mokas, S., Mills, J. R., Garreau, C., Fournier, M.-J., Robert, F., Arya, P., Kaufman, R. J., Pelletier, J., and Mazroui, R. (2009). Uncoupling Stress Granule Assembly and Translation Initiation Inhibition. *Molecular Biology of the Cell*, 20(11):2673–2683.

Monod, J., Wyman, J., and Changeux, J.-P. (1965). On the nature of allosteric transitions: A plausible model. *Journal of Molecular Biology*, 12(1):88–118.

Monti, M., Guiducci, G., Paone, A., Rinaldo, S., Giardina, G., Liberati, F. R., Cutruzzolá, F., and Tartaglia, G. G. (2021). Modelling of SHMT1 riboregulation predicts dynamic changes of serine and glycine levels across cellular compartments. *Computational and Structural Biotechnology Journal*, 19:3034–3041.

Munder, M. C., Midtvedt, D., Franzmann, T., Nüske, E., Otto, O., Herbig, M., Ulbricht, E., Müller, P., Taubenberger, A., Maharana, S., Malinovska, L., Richter, D., Guck, J., Zaburdaev, V., and Alberti, S. (2016). A pH-driven transition of the cytoplasm from a fluid- to a solid-like state promotes entry into dormancy. *eLife*, 5:e09347.

Murakami, T., Qamar, S., Lin, J. Q., Schierle, G. S. K., Rees, E., Miyashita, A., Costa, A. R., Dodd, R. B., Chan, F. T. S., Michel,

C. H., Kronenberg-Versteeg, D., Li, Y., Yang, S.-P., Wakutani, Y., Meadows, W., Ferry, R. R., Dong, L., Tartaglia, G. G., Favrin, G., Lin, W.-L., Dickson, D. W., Zhen, M., Ron, D., Schmitt-Ulms, G., Fraser, P. E., Shneider, N. A., Holt, C., Vendruscolo, M., Kaminski, C. F., and St George-Hyslop, P. (2015). ALS/FTD Mutation-Induced Phase Transition of FUS Liquid Droplets and Reversible Hydrogels into Irreversible Hydrogels Impairs RNP Granule Function. *Neuron*, 88(4):678–690.

Nagai, K. (1996). RNA-protein complexes. *Current Opinion in Structural Biology*, 6(1):53–61.

Nawy, T. (2016). Topographical transcriptomes. *Nature Methods*, 13(7):544–545.

Nelson, E. V., Schmidt, K. M., Deflubé, L. R., Doğanay, S., Banadyga, L., Olejnik, J., Hume, A. J., Ryabchikova, E., Ebihara, H., Kedersha, N., Ha, T., and Mühlberger, E. (2016). Ebola Virus Does Not Induce Stress Granule Formation during Infection and Sequesters Stress Granule Proteins within Viral Inclusions. *Journal of Virology*, 90(16):7268–7284.

Nevers, Q., Albertini, A. A., Lagaudrière-Gesbert, C., and Gaudin, Y. (2020). Negri bodies and other virus membrane-less replication compartments. *Biochimica et Biophysica Acta (BBA) - Molecular Cell Research*, 1867(12):118831.

Ng, C. S., Jogi, M., Yoo, J.-S., Onomoto, K., Koike, S., Iwasaki, T., Yoneyama, M., Kato, H., and Fujita, T. (2013). Encephalomyocarditis Virus Disrupts Stress Granules, the Critical Platform for Triggering

Antiviral Innate Immune Responses. *Journal of Virology*, 87(17):9511–9522.

Nguyen, T. C., Cao, X., Yu, P., Xiao, S., Lu, J., Biase, F. H., Sridhar, B., Huang, N., Zhang, K., and Zhong, S. (2016). Mapping RNA-RNA interactome and RNA structure in vivo by MARIO. *Nature Communications*, 7:12023.

Nguyen, T. C., Zaleta-Rivera, K., Huang, X., Dai, X., and Zhong, S. (2018). RNA, Action through Interactions. *Trends in genetics: TIG*, 34(11):867–882.

Nooren, I. M. (2003). NEW EMBO MEMBER'S REVIEW: Diversity of protein-protein interactions. *The EMBO Journal*, 22(14):3486–3492.

Oberstrass, F. C., Lee, A., Stefl, R., Janis, M., Chanfreau, G., and Allain, F. H.-T. (2006). Shape-specific recognition in the structure of the Vts1p SAM domain with RNA. *Nature Structural & Molecular Biology*, 13(2):160–167.

Onomoto, K., Jogi, M., Yoo, J.-S., Narita, R., Morimoto, S., Takemura, A., Sambhara, S., Kawaguchi, A., Osari, S., Nagata, K., Matsumiya, T., Namiki, H., Yoneyama, M., and Fujita, T. (2012). Critical role of an antiviral stress granule containing RIG-I and PKR in viral detection and innate immunity. *PloS One*, 7(8):e43031.

Oubridge, C., Ito, N., Evans, P. R., Teo, C. H., and Nagai, K. (1994). Crystal structure at 1.92 A resolution of the RNA-binding domain of the U1A spliceosomal protein complexed with an RNA hairpin. *Nature*, 372(6505):432–438.

Park, Y.-J., Choi, D. W., Cho, S. W., Han, J., Yang, S., and Choi, C. Y. (2020). Stress Granule Formation Attenuates RACK1-Mediated Apoptotic Cell Death Induced by Morusin. *International Journal of Molecular Sciences*, 21(15):5360.

Parker, J. S., Roe, S. M., and Barford, D. (2004). Crystal structure of a PIWI protein suggests mechanisms for siRNA recognition and slicer activity. *The EMBO journal*, 23(24):4727–4737.

Parker, J. S., Roe, S. M., and Barford, D. (2005). Structural insights into mRNA recognition from a PIWI domain-siRNA guide complex. *Nature*, 434(7033):663–666.

Parker, R. and Sheth, U. (2007). P Bodies and the Control of mRNA Translation and Degradation. *Molecular Cell*, 25(5):635–646.

Paronetto, M., Miñana, B., and Valcárcel, J. (2011). The Ewing Sarcoma Protein Regulates DNA Damage-Induced Alternative Splicing. *Molecular Cell*, 43(3):353–368.

Patel, A., Lee, H., Jawerth, L., Maharana, S., Jahnel, M., Hein, M., Stoynov, S., Mahamid, J., Saha, S., Franzmann, T., Pozniakovski, A., Poser, I., Maghelli, N., Royer, L., Weigert, M., Myers, E., Grill, S., Drechsel, D., Hyman, A., and Alberti, S. (2015). A Liquid-to-Solid Phase Transition of the ALS Protein FUS Accelerated by Disease Mutation. *Cell*, 162(5):1066–1077.

Penny, G. D., Kay, G. F., Sheardown, S. A., Rastan, S., and Brockdorff, N. (1996). Requirement for Xist in X chromosome inactivation. *Nature*, 379(6561):131–137.

Piao, M., Sun, L., and Zhang, Q. C. (2017). RNA Regulations and Functions Decoded by Transcriptome-wide RNA Structure Probing. *Genomics, Proteomics & Bioinformatics*, 15(5):267–278.

Piganeau, N., Schauer, U. E., and Schroeder, R. (2006). A yeast RNA-hybrid system for the detection of RNA–RNA interactions in vivo. *RNA*, 12(1):177–184.

Pijlman, G. P., Funk, A., Kondratieva, N., Leung, J., Torres, S., van der Aa, L., Liu, W. J., Palmenberg, A. C., Shi, P.-Y., Hall, R. A., and Khromykh, A. A. (2008). A Highly Structured, Nuclease-Resistant, Noncoding RNA Produced by Flaviviruses Is Required for Pathogenicity. *Cell Host & Microbe*, 4(6):579–591.

Poblete-Durán, N., Prades-Pérez, Y., Vera-Otarola, J., Soto-Rifo, R., and Valiente-Echeverría, F. (2016). Who Regulates Whom? An Overview of RNA Granules and Viral Infections. *Viruses*, 8(7):180.

Protter, D. S. W. and Parker, R. (2016). Principles and Properties of Stress Granules. *Trends in Cell Biology*, 26(9):668–679.

Qin, Q., Carroll, K., Hastings, C., and Miller, C. L. (2011). Mammalian Orthoreovirus Escape from Host Translational Shutoff Correlates with Stress Granule Disruption and Is Independent of eIF2α Phosphorylation and PKR. *Journal of Virology*, 85(17):8798–8810.

Reineke, L. C., Cheema, S. A., Dubrulle, J., and Neilson, J. R. (2018). Chronic starvation induces non-canonical pro-death stress granules. *Journal of Cell Science*, page jcs.220244.

Reineke, L. C. and Lloyd, R. E. (2015). The stress granule protein G3BP1 recruits protein kinase R to promote multiple innate immune antiviral responses. *Journal of Virology*, 89(5):2575–2589.

Reineke, L. C. and Neilson, J. R. (2019). Differences between acute and chronic stress granules, and how these differences may impact function in human disease. *Biochemical Pharmacology*, 162:123–131.

Rhim, J. S., Jordan, L. E., and Mayor, H. D. (1962). Cytochemical, fluorescent-antibody and electron microscopic studies on the growth of reovirus (ECHO 10) in tissue culture. *Virology*, 17(2):342–355.

Ries, R. J., Zaccara, S., Klein, P., Olarerin-George, A., Namkoong, S., Pickering, B. F., Patil, D. P., Kwak, H., Lee, J. H., and Jaffrey, S. R. (2019). m6A enhances the phase separation potential of mRNA. *Nature*, 571(7765):424–428.

Rouskin, S., Zubradt, M., Washietl, S., Kellis, M., and Weissman, J. S. (2014). Genome-wide probing of RNA structure reveals active unfolding of mRNA structures in vivo. *Nature*, 505(7485):701–705.

Ryter, J. M. and Schultz, S. C. (1998). Molecular basis of double-stranded RNA-protein interactions: structure of a dsRNA-binding domain complexed with dsRNA. *The EMBO journal*, 17(24):7505–7513.

Samir, P., Kesavardhana, S., Patmore, D. M., Gingras, S., Malireddi, R. K. S., Karki, R., Guy, C. S., Briard, B., Place, D. E., Bhattacharya, A., Sharma, B. R., Nourse, A., King, S. V., Pitre, A., Burton, A. R., Pelletier, S., Gilbertson, R. J., and Kanneganti, T.-D. (2019). DDX3X acts as a live-or-die checkpoint in stressed cells by regulating NLRP3 inflammasome. *Nature*, 573(7775):590–594.

Sanchez de Groot, N., Armaos, A., Graña-Montes, R., Alriquet, M., Calloni, G., Vabulas, R. M., and Tartaglia, G. G. (2019). RNA structure drives interaction with proteins. *Nature Communications*, 10(1):3246.

Savastano, A., Ibáñez de Opakua, A., Rankovic, M., and Zweckstetter, M. (2020). Nucleocapsid protein of SARS-CoV-2 phase separates into RNA-rich polymerase-containing condensates. *Nature Communications*, 11(1):6041.

Schmidt, N., Lareau, C. A., Keshishian, H., Ganskih, S., Schneider, C., Hennig, T., Melanson, R., Werner, S., Wei, Y., Zimmer, M., Ade, J., Kirschner, L., Zielinski, S., Dölken, L., Lander, E. S., Caliskan, N., Fischer, U., Vogel, J., Carr, S. A., Bodem, J., and Munschauer, M. (2021). The SARS-CoV-2 RNA–protein interactome in infected human cells. *Nature Microbiology*, 6(3):339–353.

Schubert, M., Edge, R. E., Lario, P., Cook, M. A., Strynadka, N. C. J., Mackie, G. A., and McIntosh, L. P. (2004). Structural characterization of the RNase E S1 domain and identification of its oligonucleotide-binding and dimerization interfaces. *Journal of Molecular Biology*, 341(1):37–54.

Seto, E., Inoue, T., Nakatani, Y., Yamada, M., and Isomura, H. (2014). Processing bodies accumulate in human cytomegalovirus-infected cells and do not affect viral replication at high multiplicity of infection. *Virology*, 458-459:151–161.

Shahmoradian, S. H., Lewis, A. J., Genoud, C., Hench, J., Moors, T. E., Navarro, P. P., Castaño-Díez, D., Schweighauser, G., Graff-Meyer, A., Goldie, K. N., Sütterlin, R., Huisman, E., Ingrassia, A., Gier, Y. d.,

Rozemuller, A. J. M., Wang, J., Paepe, A. D., Erny, J., Staempfli, A., Hoernschemeyer, J., Großerüschkamp, F., Niedieker, D., El-Mashtoly, S. F., Quadri, M., Van IJcken, W. F. J., Bonifati, V., Gerwert, K., Bohrmann, B., Frank, S., Britschgi, M., Stahlberg, H., Van de Berg, W. D. J., and Lauer, M. E. (2019). Lewy pathology in Parkinson's disease consists of crowded organelles and lipid membranes. *Nature Neuroscience*, 22(7):1099–1109.

Shao, W., Zeng, S.-T., Yu, Z.-Y., Tang, G.-X., Chen, S.-B., Huang, Z.-S., Chen, X.-C., and Tan, J.-H. (2021). Tracking Stress Granule Dynamics in Live Cells and *In Vivo* with a Small Molecule. *Analytical Chemistry*, 93(49):16297–16301.

Sharma, E., Sterne-Weiler, T., O'Hanlon, D., and Blencowe, B. J. (2016). Global Mapping of Human RNA-RNA Interactions. *Molecular Cell*, 62(4):618–626.

Sheth, U. and Parker, R. (2003). Decapping and decay of messenger RNA occur in cytoplasmic processing bodies. *Science (New York, N.Y.)*, 300(5620):805–808.

Shi, H., Wang, X., Lu, Z., Zhao, B. S., Ma, H., Hsu, P. J., Liu, C., and He, C. (2017). YTHDF3 facilitates translation and decay of N6-methyladenosine-modified RNA. *Cell Research*, 27(3):315–328.

Shi, Y. (2014). A Glimpse of Structural Biology through X-Ray Crystallography. *Cell*, 159(5):995–1014.

Shin, Y. and Brangwynne, C. P. (2017). Liquid phase condensation in cell physiology and disease. *Science*, 357(6357):eaaf4382.

Sickmier, E. A., Frato, K. E., Shen, H., Paranawithana, S. R., Green, M. R., and Kielkopf, C. L. (2006). Structural basis for polypyrimidine tract recognition by the essential pre-mRNA splicing factor U2AF65. *Molecular Cell*, 23(1):49–59.

Siegfried, N. A., Busan, S., Rice, G. M., Nelson, J. A. E., and Weeks, K. M. (2014). RNA motif discovery by SHAPE and mutational profiling (SHAPE-MaP). *Nature Methods*, 11(9):959–965.

Silva, P. A. G. C., Pereira, C. F., Dalebout, T. J., Spaan, W. J. M., and Bredenbeek, P. J. (2010). An RNA Pseudoknot Is Required for Production of Yellow Fever Virus Subgenomic RNA by the Host Nuclease XRN1. *Journal of Virology*, 84(21):11395–11406.

Silverman, I. M., Li, F., Alexander, A., Goff, L., Trapnell, C., Rinn, J. L., and Gregory, B. D. (2014). RNase-mediated protein footprint sequencing reveals protein-binding sites throughout the human transcriptome. *Genome Biology*, 15(1):R3.

Souquere, S., Beauclair, G., Harper, F., Fox, A., and Pierron, G. (2010). Highly Ordered Spatial Organization of the Structural Long Noncoding NEAT1 RNAs within Paraspeckle Nuclear Bodies. *Molecular Biology of the Cell*, 21(22):4020–4027.

Spitale, R. C., Flynn, R. A., Zhang, Q. C., Crisalli, P., Lee, B., Jung, J.-W., Kuchelmeister, H. Y., Batista, P. J., Torre, E. A., Kool, E. T., and Chang, H. Y. (2015). Structural imprints in vivo decode RNA regulatory mechanisms. *Nature*, 519(7544):486–490.

Su, Y., Maimaitiyiming, Y., Wang, L., Cheng, X., and Hsu, C.-H. (2021). Modulation of Phase Separation by RNA: A Glimpse on N6-

Methyladenosine Modification. *Frontiers in Cell and Developmental Biology*, 9:786454.

Subramanian, A. R. (1983). Structure and functions of ribosomal protein S1. *Progress in Nucleic Acid Research and Molecular Biology*, 28:101–142.

Sugimoto, Y., Vigilante, A., Darbo, E., Zirra, A., Militti, C., D'Ambrogio, A., Luscombe, N. M., and Ule, J. (2015). hiCLIP reveals the in vivo atlas of mRNA secondary structures recognized by Staufen 1. *Nature*, 519(7544):491–494.

Sánchez-Jiménez, C., Ludeña, M. D., and Izquierdo, J. M. (2015). T-cell intracellular antigens function as tumor suppressor genes. *Cell Death & Disease*, 6(3):e1669–e1669.

Talkish, J., May, G., Lin, Y., Woolford, J. L., and McManus, C. J. (2014). Mod-seq: high-throughput sequencing for chemical probing of RNA structure. *RNA (New York, N.Y.)*, 20(5):713–720.

Tartaglia, G. G., Pawar, A. P., Campioni, S., Dobson, C. M., Chiti, F., and Vendruscolo, M. (2008). Prediction of Aggregation-Prone Regions in Structured Proteins. *Journal of Molecular Biology*, 380(2):425–436.

Tourrière, H., Chebli, K., Zekri, L., Courselaud, B., Blanchard, J. M., Bertrand, E., and Tazi, J. (2003). The RasGAP-associated endoribonuclease G3BP assembles stress granules. *Journal of Cell Biology*, 160(6):823–831.

Tsai, N.-P. and Wei, L.-N. (2010). RhoA/ROCK1 signaling regulates

stress granule formation and apoptosis. *Cellular Signalling*, 22(4):668–675.

Tweedie, A. and Nissan, T. (2021). Hiding in Plain Sight: Formation and Function of Stress Granules During Microbial Infection of Mammalian Cells. *Frontiers in Molecular Biosciences*, 8:647884.

Underwood, J. G., Uzilov, A. V., Katzman, S., Onodera, C. S., Mainzer, J. E., Mathews, D. H., Lowe, T. M., Salama, S. R., and Haussler, D. (2010). FragSeq: transcriptome-wide RNA structure probing using high-throughput sequencing. *Nature Methods*, 7(12):995–1001.

Valadkhan, S. and Manley, J. L. (2001). Splicing-related catalysis by protein-free snRNAs. *Nature*, 413(6857):701–707.

Valiente-Echeverría, F., Melnychuk, L., Vyboh, K., Ajamian, L., Gallouzi, I.-E., Bernard, N., and Mouland, A. J. (2014). eEF2 and Ras-GAP SH3 domain-binding protein (G3BP1) modulate stress granule assembly during HIV-1 infection. *Nature Communications*, 5(1):4819.

Van Treeck, B., Protter, D. S. W., Matheny, T., Khong, A., Link, C. D., and Parker, R. (2018). RNA self-assembly contributes to stress granule formation and defining the stress granule transcriptome. *Proceedings of the National Academy of Sciences*, 115(11):2734–2739.

Verdile, V., De Paola, E., and Paronetto, M. P. (2019). Aberrant Phase Transitions: Side Effects and Novel Therapeutic Strategies in Human Disease. *Frontiers in Genetics*, 10:173.

Vonaesch, P., Campbell-Valois, F.-X., Dufour, A., Sansonetti, P. J., and Schnupf, P. (2016). *Shigella flexneri* modulates stress granule compo-

sition and inhibits stress granule aggregation: Shigella flexneri *affects stress granule formation*. *Cellular Microbiology*, 18(7):982–997.

Wan, G., Fields, B. D., Spracklin, G., Shukla, A., Phillips, C. M., and Kennedy, S. (2018). Spatiotemporal regulation of liquid-like condensates in epigenetic inheritance. *Nature*, 557(7707):679–683.

Wan, Y., Qu, K., Ouyang, Z., Kertesz, M., Li, J., Tibshirani, R., Makino, D. L., Nutter, R. C., Segal, E., and Chang, H. Y. (2012). Genome-wide measurement of RNA folding energies. *Molecular Cell*, 48(2):169–181.

Wan, Y., Qu, K., Zhang, Q. C., Flynn, R. A., Manor, O., Ouyang, Z., Zhang, J., Spitale, R. C., Snyder, M. P., Segal, E., and Chang, H. Y. (2014). Landscape and variation of RNA secondary structure across the human transcriptome. *Nature*, 505(7485):706–709.

Wang, X., Lu, Z., Gomez, A., Hon, G. C., Yue, Y., Han, D., Fu, Y., Parisien, M., Dai, Q., Jia, G., Ren, B., Pan, T., and He, C. (2014). N6-methyladenosine-dependent regulation of messenger RNA stability. *Nature*, 505(7481):117–120.

Wang, X., McLachlan, J., Zamore, P. D., and Hall, T. M. T. (2002). Modular recognition of RNA by a human pumilio-homology domain. *Cell*, 110(4):501–512.

Wasserman, T., Katsenelson, K., Daniliuc, S., Hasin, T., Choder, M., and Aronheim, A. (2010). A Novel c-Jun N-terminal Kinase (JNK)-binding Protein WDR62 Is Recruited to Stress Granules and Mediates a Nonclassical JNK Activation. *Molecular Biology of the Cell*, 21(1):117–130.

Wendt, L., Brandt, J., Bodmer, B. S., Reiche, S., Schmidt, M. L., Traeger, S., and Hoenen, T. (2020). The Ebola Virus Nucleoprotein Recruits the Nuclear RNA Export Factor NXF1 into Inclusion Bodies to Facilitate Viral Protein Expression. *Cells*, 9(1):187.

Whyte, W., Orlando, D., Hnisz, D., Abraham, B., Lin, C., Kagey, M., Rahl, P., Lee, T., and Young, R. (2013). Master Transcription Factors and Mediator Establish Super-Enhancers at Key Cell Identity Genes. *Cell*, 153(2):307–319.

Wolf, M., Roosen-Runge, F., Zhang, F., Roth, R., Skoda, M. W., Jacobs, R. M., Sztucki, M., and Schreiber, F. (2014). Effective interactions in protein–salt solutions approaching liquid–liquid phase separation. *Journal of Molecular Liquids*, 200:20–27.

Wolfe, S. A., Nekludova, L., and Pabo, C. O. (2000). DNA recognition by Cys2His2 zinc finger proteins. *Annual Review of Biophysics and Biomolecular Structure*, 29:183–212.

Wolozin, B. L., Pruchnicki, A., Dickson, D. W., and Davies, P. (1986). A Neuronal Antigen in the Brains of Alzheimer Patients. *Science*, 232(4750):648–650.

Wu, Y., Ma, L., Cai, S., Zhuang, Z., Zhao, Z., Jin, S., Xie, W., Zhou, L., Zhang, L., Zhao, J., and Cui, J. (2021). RNA-induced liquid phase separation of SARS-CoV-2 nucleocapsid protein facilitates NF-kB hyperactivation and inflammation. *Signal Transduction and Targeted Therapy*, 6(1):167.

Xue, Y. (2022). Architecture of RNA-RNA interactions. *Current Opinion in Genetics & Development*, 72:138–144.

Yan, K. S., Yan, S., Farooq, A., Han, A., Zeng, L., and Zhou, M.-M. (2003). Structure and conserved RNA binding of the PAZ domain. *Nature*, 426(6965):468–474.

Yang, P., Mathieu, C., Kolaitis, R.-M., Zhang, P., Messing, J., Yurtsever, U., Yang, Z., Wu, J., Li, Y., Pan, Q., Yu, J., Martin, E. W., Mittag, T., Kim, H. J., and Taylor, J. P. (2020). G3BP1 Is a Tunable Switch that Triggers Phase Separation to Assemble Stress Granules. *Cell*, 181(2):325–345.e28.

Yu, D., Qin, P., and Cornish, P. V. (2015). Single molecule studies of RNA-RNA interactions. *Methods in Molecular Biology (Clifton, N.J.)*, 1240:97–112.

Zacco, E., Kantelberg, O., Milanetti, E., Armaos, A., Panei, F. P., Gregory, J., Jeacock, K., Clarke, D. J., Chandran, S., Ruocco, G., Gustincich, S., Horrocks, M. H., Pastore, A., and Tartaglia, G. G. (2022). Probing TDP-43 condensation using an in silico designed aptamer. *Nature Communications*, 13(1):3306.

Zanchetta, G., Nakata, M., Buscaglia, M., Clark, N. A., and Bellini, T. (2008). Liquid crystal ordering of DNA and RNA oligomers with partially overlapping sequences. *Journal of Physics: Condensed Matter*, 20(49):494214.

Zbinden, A., Pérez-Berlanga, M., De Rossi, P., and Polymenidou, M. (2020). Phase Separation and Neurodegenerative Diseases: A Disturbance in the Force. *Developmental Cell*, 55(1):45–68.