TESI DOCTORAL UPF / YEAR 2019

Computational characterization of protein-RNA interactions and implications for phase separation

Alexandros Armaos

Director

Dr. Gian Gaetano Tartaglia

Gene Function and Evolution

Bioinformatics and Genomics Department

Centre for Genomic regulation (CRG)

Acknowledgements

Well, all started almost 5 years ago, when I took the decision to do a PhD. To be honest, at least back then I was motivated to start and accomplish a PhD for personal reasons and not for scientific interest. You can call me selfish dear reader, but I had the ambition and the need to push myself, test my capabilities and see how far can I reach. And doing a PhD could provide me with the opportunity to be creative, innovative and use my personal intuition and imagination. Thinking of a complex "puzzle" and try to solve it, regardless of whether it has a solution or not sound to me really intriguing and fascinating. There is saying that I often use: "When the going gets tough, the tough get going!", and I wanted to "get going"!

So this journey started when I visited Mr. George Paliouras in the Greek National Center for Scientific Research, Demokritos. Mr. Paliouras, first and foremost, is an exceptionally clever person, kind and very polite. He won my admiration and trust by the very first moment I met him. Mr. Paliouras helped me a lot with preparing my application for the PhD call in CRG and gave me many advises, opinions and guidance for my first steps.

The months passed and finally I was selected and invited in CRG for the interviews. First day of the interviews and we were offered a welcome coffee at CRG 5th floor terrace. "WOW!! Is that real? Can I really have this view every day from my working environment?" I thought… Dear reader, this view is the best meditation you can think of. Just ten minutes in the terrace, looking at the horizon and the blue infinity of the Mediterranean sea, will really relax you, clean your mind and motivate you. CRG is such an appealing place to work, clean, modern, human friendly and just 50 meters away from the sea! It is more than obvious that I got supper enthusiastic when I saw all these. "Yes!!", I thought, "this is the right place and the right moment!!"… And it was! After several months of bureaucracy problems, I was informed that I was finally accepted and that by September '15 I would be enrolled as a PhD student. "OH YEAH!!! Barcelona here I come!"

Now that I am writing these lines dear reader, its already summer '19… Four years have flown by the blink of an eye and I am almost finishing this journey. An amazing period of my life reaches its end and offers its turn to the next one to come. I consider myself absolutely lucky that I had this opportunity to experience this adventure. Definitively, this period was one of the greatest and most important of my life. I feel so full of experience, confident for myself, satisfied about my decisions and really, really happy. I truly wish that everyone could experience a PhD adventure like the one I had.

We usually relate places, moments or periods of our lives with the people that were present during those and with people that contributed to these moments. I think that, in the very end, are the people surrounding you, those that make a moment special to you.

I couldn’t thank anyone else more than my dear friend Gian for these amazing years. And I say friend, because to my eyes Gian Gaetano Tartaglia is primarily a real friend and secondary my thesis supervisor. Gian trusted me and really showed me his interest in working with me from the first day of the PhD interviews. He offered me this amazing opportunity to come to Barcelona and to do my great step by being part of the Tartaglia’s lab. He dedicated a lot of his time on me, providing me with his invaluable advises, his clever opinions and showing an absolute trust on my face by involving me in almost all projects of the lab. Both of us knew how the other thinks and works and this perfect combination of characters made my PhD look more like an amusing hobby than anything else. My first day as his PhD student, he told me exactly that: “I want you to enjoy it!”. Back then I didn’t understand exactly what he wanted to say, but now I do, because I enjoyed it at one hundred percent! Even during the really stressful and difficult moments of my PhD, Gian was always there, offering a cold beer while listening to some high quality psychedelic rock music. Yes!!, even our taste in music was identical. His priceless achievement, however, was to create a group, Tartaglia’s lab, that was something beyond a group of students and post-docs. For all of us, this lab is a family and it is an absolute honor of me to have been member of it. I could spend endless pages describing the amazing atmosphere of this lab, I will, however, try to keep this section short.

There were a lot of people that were part of this family during these years, and all of them in their way contributed and helped me to successfully complete the current doctoral thesis.

I would like to give special thanks to Davide Cirillo, a previous PhD student of the lab. He was a real mentor, an inspiring personality and a truly gentle person that I really admire. He was always there, for any kind of help I needed, the prototype PhD student that I was looking forward to be.

Remarkable thanks I would like to give to Natalia, the silent mind of the lab. We had an extraordinary cooperation in the paper that we co-authored. She transmitted to me her outside-of-the-box way of thinking and her broad perception and understanding of how things work. She is an exceptional person and researcher. Someone could give to Natalia three ingredients and she could engineer the way to bake a cake without even dirt the kitchen!

I would like to express my sincere appreciation and thanks to Fernando who is the definitive soul and mascot of the lab. We joined Tartaglia's family the same week and during the first months, while we were working side by side in a different floor from the rest of the lab, we shared the integration and adaptation period to the PhD reality. I can admit that my PhD would have been a lot more boring without him. Daily, we had amazing conversations of any kind, interchanging opinions on science, politics, books, music and of course…girls! Being very clever, with critical mind, explosive temperament and wonderful sense of humor, Fernando is the best "compañero de trabajo". It was an absolute pleasure working with him and I will never forget him shouting: "¿ Me dejáis trabajar por favor ?".

I would also like to express my sincere thanks to Ale, a real friend with whom I daily had the most bizarre and crazy conversations, disclosing our deepest thoughts on any possible subject and without any prejudgment.

I would like to thank Stefanie for her genius intuitions and the inspiring talks that we had regarding artificial intelligence and neural networks. And obviously for the delicious swiss chocolates that she was bringing to the lab every time she was coming back from Switzerland!

I would like to thanks Elias for being such an honest person. During these years he transmitted to the lab and me the necessary tranquility with his endless hugs. I am sure that Elias, if he was not working in research he would have been an actor and/or a singer. With his amazing performances he offered us with unforgettable moments of laugh that could cheer up even the most boring day.

I would also like to express my notable thanks to Mimma and MC (Maria Carla). These two Italian girls that were working just next to my desk provided me with enthusiasm, motivating comments and most predominantly, deep psychological support when I was facing difficult situations.

I would also like to thank Riccardo and Nieves for their sincere support and the funny moments that we daily shared in the lab. But most importantly I would like to appreciate their joyful company during the daily coffee break at the our beloved “Magatzem 03” !

I would also like to thank Petr for all the guidance and knowledge that he gave me to be able to administrate the digital infrastructure of the lab, as well as his priceless advises on informatics and coding.

I would like to thank the newer post-docs of the lab. Ben for being the open encyclopedia of the lab and lastly Michele Monti. Even though that he joined us just some months ago, we have already built a strong friendship and we expect to continue collaborating during the next years.

Finally, I would also like to thank Andrea, Laura and Irene for their clever questions that made me doubt for things that I thought I knew but I didn’t and of course Magda, the hardworking and smiley girl that brought a fresh wind of joy to the lab!

As for the people outside of the lab, I would honestly like to thank all my friends here in Barcelona, Aggelos, Dimitris, Frank, Manos, Panos, Tasos, Anna and Nandia for being by my side daily. They were my second family here in Barcelona, supporting my decisions and being a real source of courage, energy and motivation.

A great thanks to all my friends back in Greece for being a remarkable origin of inspiration for my life. A big part of me is because of them.

They were always providing me with innovative and diverse stimulus and continual motivation for the evolution of my character, my thinking and my personality. Our strong bonds is one of the most important achievements and gifts that I have. I would like to express my sincere apologies for not being with them during these years.

There are two people in my life however, to whom I own everything and to whom this work is dedicated. These are my parents, Korina and Iosif. There are no words that can reflect my feelings for them. There are no words that can describe my thankfulness for everything they have done for me. The present work is mine as much as theirs. They were my eternal source of energy, kindness, motivation… Thank you from the bottom of my heart!

Alexandros Armaos

Barcelona , August 2019

Abstract

Despite what was previously considered, the role of RNA is not only to carry the genetic information from DNA to proteins. Indeed, RNA has proven to be implicated in more complex cellular processes. Recent evidence suggests that transcripts have a regulatory role on gene expression and contribute to the spatial and temporal organization of the intracellular environment. They do so by interacting with RNA-binding proteins (RBPs) to form complex ribonucleoprotein (RNP) networks, however the key determinants that govern the formation of these complexes are still not well understood. In this work, I will describe algorithms that I developed to estimate the ability of RNAs to interact with proteins. Additionally, I will illustrate applications of computational methods to propose an alternative model for the function of *Xist* lncRNA and its protein network.

Finally, I will show how computational predictions can be integrated with high throughput approaches to elucidate the relationship between the structure of the RNA and its ability to interact with proteins. I conclude by discussing open questions and future opportunities for computational analysis of cell's regulatory network.

Overall, the underlying goal of my work is to provide biologists with new insights into the functional association between RNAs and proteins as well as with sophisticated tools that will facilitate their investigation on the formation of RNP complexes.

Resumen

A pesar de lo que se consideraba anteriormente, el papel del ARN no es solo transportar la información genética del ADN a las proteínas. De hecho, el ARN ha demostrado estar implicado en muchos procesos celulares más complejos. La evidencia reciente sugiere que los transcriptos tienen un papel regulador en la expresión génica y contribuyen a la organización espacial y temporal del entorno intracelular. Lo hacen interactuando con proteínas de unión a ARN (RBP) para formar redes complejas de ribonucleoproteína (RNP), sin embargo, los determinantes clave que rigen la formación de estos complejos aún no se conocen bien. En este trabajo, describiré algoritmos que he desarrollado para estimar la capacidad de los ARN de interactuar con las proteínas. Además, ilustraré aplicaciones de métodos computacionales para proponer una maquinaria alternativa para el *Xist* lncRNA y su red de interacciones.

Finalmente, mostraré cómo las predicciones computacionales pueden integrarse con enfoques de alto rendimiento para dilucidar la relación entre la estructura del ARN y su capacidad para interactuar con las proteínas. Concluyo discutiendo preguntas abiertas y oportunidades futuras para el análisis computacional de la red reguladora de la célula.

En general, el objetivo subyacente de mi trabajo es proporcionar a los biólogos nuevas ideas sobre la asociación funcional entre ARN y

proteínas, así como herramientas sofisticadas que facilitarán su investigación sobre la formación de complejos RNP.

Preface

In this thesis I will present the work in which I have been directly involved during my doctorate studies. The first two projects, presented in Chapters I and II, include the development of two algorithms that constitute the evolution of the *cat*RAPID algorithm which was previously published by Tartaglia's group. These two methods use as features the physico-chemical and structural properties of the protein and RNA chains and they make use of neural networks for their prediction procedure. In more detail, Chapter I presents the *Global Score* approach, which is a method for classifying interacting from non-interacting protein-RNA pairs. Chapter II on the other hand, introduces the *omiXcore* approach, a method for predicting the interaction affinity for a protein-RNA pair trained on Enhanced Crosslinking and Immunoprecipitation data. Both methods can be used in complementarity and they are optimized for long transcripts. The Chapter III of this thesis illustrates an application of the *Global Score* method. In more detail, that Chapter focuses on the process of X-chromosome inactivation by *Xist* long non-coding RNA and presents the hypothesis that *Xist* uses phase separation to perform its function.

Finally, Chapter IV demonstrates the results of an extensive investigation and analysis on the involvement of the RNA structure in its ability to bind proteins. Additionally, an evolutionary hypothesis is presented, that connects structured mRNAs with the potential of their protein products to physically interact with other proteins providing new insights into the layers of the cellular regulatory network.

In the Discussion part of the current doctoral thesis, I will highlight the main findings of Chapters I to IV and their importance in the scientific community. I will also propose my new hypothesis that expands the results of Chapter IV and I hope that these perspectives will give stimulating ideas for future investigation on the field.

Table of Contents

Introduction

1 Central dogma of molecular biology

DNA lies at the core of molecular biology since is the key macromolecule for the continuity of life. It stores the hereditary and genetic information that is passed on from parents to children, providing instructions for how and when to make the proteins needed to build and maintain functioning cells.

DNA (deoxyribonucleic acid) and RNA (ribonucleic acid) are the two naturally occurring varieties of nucleic acids, which are macromolecules (polymers) made out of units called nucleotides. In eukaryotes, such as plants and animals, DNA is found in the nucleus and is typically divided into chromosomes. However, in prokaryotes, such as bacteria, the DNA is not enclosed in a membranous envelope, although it is located in a specialized cell region called the nucleoid and in that case, chromosomes are usually smaller and circular. A chromosome may contain tens of thousands of genes, each providing instructions on how to make a particular product needed by the cell. Additionally, many of these genes encode for proteins, meaning that they specify the sequence of amino acids used to build a particular polypeptide. Before this information can be used for protein synthesis, however, an RNA copy (transcript) of the gene must first be made. This type of RNA is called messenger RNA (mRNA) and transmits the genetic information from the DNA to the ribosomes, molecular machines that read mRNA and build proteins. This progression from DNA to RNA to protein is called the *central dogma* of molecular biology.

Not all genes store information for protein production. For instance, some genes encode ribosomal RNAs (rRNAs), which serve as structural components of ribosomes, or transfer RNAs (tRNAs) which are the RNA molecules that bring amino acids to the ribosome for protein synthesis. Still other RNA molecules, such as microRNAs (miRNAs) that act as regulators of other genes or a class of large RNA transcripts not coding for proteins termed long non-coding RNAs (lncRNAs), have provided new perspectives on the important role of the RNA in gene regulation. Processes including protein synthesis, RNA maturation and transport and transcriptional gene silencing by chromatin structure regulation have been shown to be controlled by this class of lncRNAs (Bernstein and Allis, 2005). Intriguingly, there seem to be a linear relationship between the complexity of an organism and the number of non-coding RNAs produced (Taft et al., 2007) suggesting that developmental complexity which is not reflected in the number of protein-coding genes maybe mediated by non-coding RNAs. However, the mechanisms by which non-coding RNAs contribute to this complexity is not completely understood.

2 Chemistry and Biology of Nucleic Acids

A nucleotide is one of the structural components of DNA and RNA and is made up of three parts, a nitrogen-containing ring structure called a nitrogenous base, a five-carbon sugar, and at least one phosphate group derived from phosphoric acid. The sugar molecule has a central position in the nucleotide, with the base attached to one of its carbons and the phosphate group (or groups) attached to another (Figure 1).

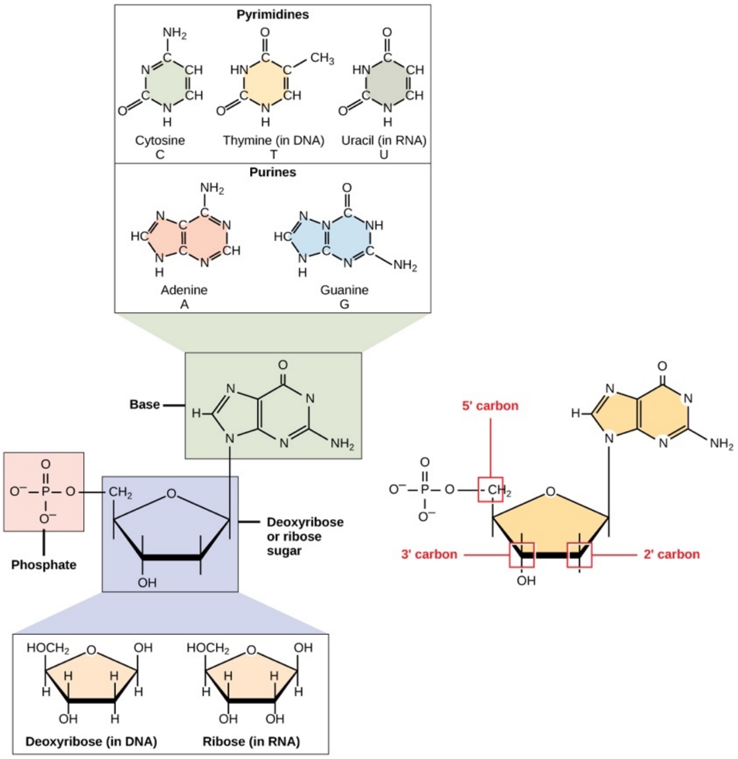


Figure 1) Components of DNA and RNA, including the sugar (deoxyribose or ribose), phosphate group, and nitrogenous base. Bases include the pyrimidine bases (cytosine, thymine in DNA, and uracil in RNA, one ring) and the purine bases (adenine and guanine, two rings). The phosphate group is attached to the 5' carbon. The 2' carbon bears a hydroxyl group in ribose, but no hydroxyl (just hydrogen) in deoxyribose. *[Adapted from Biology. OpenStax CNX]*

2.1 Nitrogenous bases

Each nucleotide in DNA contains one of four possible nitrogenous bases: adenine (A), guanine (G) cytosine (C), and thymine (T). Adenine and guanine are purines, meaning that their structures contain two fused carbon-nitrogen rings. Cytosine and thymine, in contrast, are pyrimidines and have a single carbon-nitrogen ring. RNA nucleotides may also contain adenine, guanine and cytosine bases, but instead of thymine they have another pyrimidine-derived base called uracil (U). As shown in the Figure 1, each base has a unique structure, with its own set of functional groups attached to the ring structure.

2.2 Sugars

DNA and RNA nucleotides, apart from having slightly different sets of bases, they also have different sugars. The five-carbon sugar in DNA is called deoxyribose, while in RNA, the sugar is ribose and differs from deoxyribose for a hydroxyl group attached to the 2'-position of the pentose sugar (Figure 2).

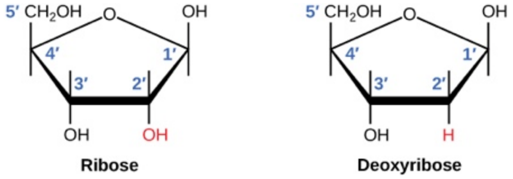


Figure 2) The difference between the ribose found in RNA and the deoxyribose found in DNA is that ribose has a hydroxyl group at the 2′ carbon. [*Adapted from Concepts of Biology, 1st Canadian Edition, Charles Molnar and Jane Gair*]

2.3 Phosphate

In a cell, a nucleotide before to be added to the end of a polynucleotide chain will contain three phosphate groups. When the nucleotide is added to a DNA or RNA chain, it loses two phosphate groups and, consequently, in a polynucleotide chain each nucleotide has just one phosphate group.

2.4 Polynucleotide chains

At the 5’ end, or the beginning of the chain, the first nucleotide has a 5’ phosphate group. At the other end, called the 3’ end, the last nucleotide has a 3’ hydroxyl structure. As a consequence of the above structure, the polynucleotide chain has directionality which means that it has two ends that are different from each other.

DNA sequences are as a rule written in the 5' to 3' direction, meaning that the nucleotide at the 5' end comes first and the nucleotide at the 3' end comes last. New nucleotides are added to a strand of DNA or RNA, at the 3’ end, with the 5′ phosphate of an incoming nucleotide attaching to the hydroxyl group at the 3’ end of the chain. This makes a chain with each sugar joined to its neighbours by phosphodiester linkages.

3 Properties of DNA and RNA

Deoxyribonucleic acid, or DNA, chains are typically found in a double helix, a structure in which two matching (complementary) chains are stuck together, as shown in Figure 3. The sugars and phosphates lie on the outside of the helix, forming the backbone of the DNA. The nitrogenous bases that extend into the interior are bound to each other by hydrogen bonds.

The two strands of the helix run in opposite directions, meaning that the 5′ end of one strand is paired up with the 3′ end of its matching strand. This is referred to as antiparallel orientation and is important for the copy of DNA. The base pairing is highly specific meaning that A can only pair with T or U by two hydrogen bonds, and G can only pair with C by three hydrogen bonds (Figure 3). Because of the canonical base pairing, the sequence of one strand can precisely define the sequence of the other in a complementary way.

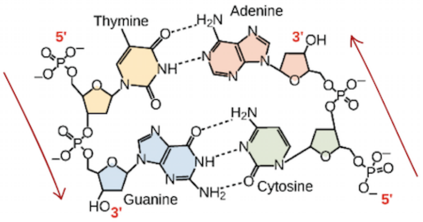


Figure 3) Hydrogen bonding between complementary bases holds DNA strands together in a double helix of antiparallel strands. Thymine forms two hydrogen bonds with adenine and guanine forms three hydrogen bonds with cytosine. *[Adapted from Biology. OpenStax CNX]*

Contrary to the DNA, the RNA in its native state exists in a single-stranded conformation. This feature allows the RNA to be more dynamic and with a higher degree of freedom than the double-stranded DNA. Because of its higher flexibility, the RNA is able to form interactions along its primary structure. The pattern of the internal interactions for an entire RNA molecule is defined as its secondary structure (Doty et al., 1959). Between the possible base pairing interactions of the four RNA bases (Adenine, Cytosine, Guanine and Uracil), only six are stable (AU, GU, GC, UA, UG, CG). Accordingly, these are the most common interactions within RNA molecules. The GC/CG pairs are the strongest and thus the most stable ones since they are formed by 3 hydrogen bonds, while the rest four are formed only by 2 hydrogen bonds. The consequence of these base pairing interactions, is that the RNA is able to form different structural elements of various lengths. These structural elements are the following:

- Stems: double-stranded regions. The most stable RNA structural motif and usually the longest one.
- Hairpin-loop: a very common structure which is the combination of strong complementary bases separated by unpaired nucleotides.
- Internal-loop: a loop that is internal to consecutive stems and has the same number of nucleotides on the left and on the right side.
- Bulge: a specific sub-class of internal-loop where only one side of the loop has unpaired nucleotides while the other is connected to the stem.
- Multibranch-loop: a complex structure composed of different sub-loop structured.

- Pseudoknots: the most complex structure. Is formed when a loop region and bases outside of the loop interact.

4 Physical and Chemical properties of Proteins

Proteins are one of the most abundant organic molecules in living systems and have the most diverse range of functions of all macromolecules since they are implicated in the majority of cell's processes. Each cell in a living system may contain thousands of proteins with varying structures and functions. They are all, however, polymers of amino acids, arranged in a linear sequence. Each amino acid has the same fundamental structure, which consists of a central carbon atom, also known as the alpha (α) carbon, bonded to an amino group (NH_2), a carboxyl group (COOH), and to a hydrogen atom. Every amino acid also has another atom or group of atoms bonded to the central atom known as the R group (Figure 4 left).

In the simplest amino acid, glycine, the R group is hydrogen (–H), but in other naturally occurring amino acids, the R group may be an alkyl group or a substituted alkyl group, a carboxylic group, or an aryl group. The nature of the R group determines the particular chemical properties of each amino (that is, whether it is acidic, basic, polar, or nonpolar). In total, there are 20 standard amino acids commonly present in proteins. Ten of these are considered essential amino acids in humans because the human body cannot produce them and they are obtained from the diet (Figure 4 centre).

Each amino acid is attached to another amino acid covalently by a peptide bond, and the resulting chain is known as a polypeptide. The sequence and number of amino acids ultimately determine the protein's shape, size, and function (Figure 4 right). Each polypeptide has a free amino group at one end, called the N terminal or the amino terminal. At the other end has a free carboxyl group, also known as the C or carboxyl terminal. After protein synthesis (translation), most proteins are modified. These are known as post-translational modifications, such as: cleavage, phosphorylation, or the addition of other chemical groups. Only after these modifications the protein is fully functional.

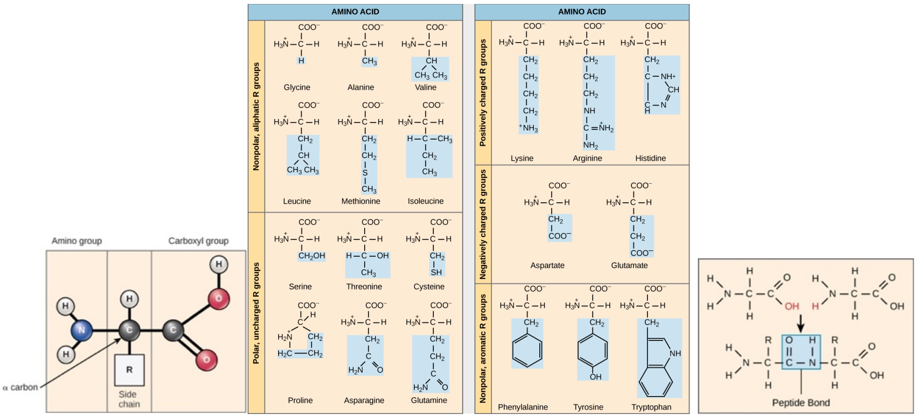


Figure 4: **Left)** Amino acids have a central asymmetric carbon to which an amino group, a carboxyl group, a hydrogen atom, and a side chain (R group) are attached. **Center)** There are 20 amino acids commonly found in proteins, each with a different R group (variant group) that determines its chemical nature. **Right)** Peptide bond formation is a dehydration synthesis reaction. The carboxyl group of one amino acid is linked to the amino group of the incoming amino acid. In the process, a molecule of water is released.

5 The protein-RNA interaction event

RNA and proteins are two interconnected molecules, which means that they are involved in the regulation of many aspects of each other's life. From their transcription to the end of their life RNAs are coated with proteins. The RNA-binding proteins (RBPs) orchestrate all phases of post-transcriptional RNA regulation, including splicing, polyadenylation, localization, transport, translation, stability and degradation (Jankowsky and Harris, 2015). Nevertheless, RNA also has a regulatory role on the fate of the proteins. There has been substantial work showing that there is a functional crosstalk between RNAs and their protein partners (Delaunay and Frye, 2019; Yao et al., 2019). For instance, it has been demonstrated that lncRNAs can influence proteins by acting as guides, scaffolds, signals or decoys mediating their functions including processing, modification, localization stability and translation (Wang and Chang, 2011). This relationship results in a very complex regulatory interacting network, since different partners are necessary for each process and they can either bind simultaneously, subsequently, or in a mutually exclusive manner.

Perturbations or mis-regulation of these networks can lead to cellular dysfunctions that have been linked with diseases such for instance amyotrophic lateral sclerosis (ALS), Creutzfeuld-Jakob, Alzheimer's, and Parkinson's diseases (Cid-Samper et al., 2018; Marchese et al., 2017). In this context, it is crucial to study protein-RNA assemblies to better understand the etiopathogenesis of specific diseases and to design new therapeutic strategies. Hopefully, due to exciting advances in experimental technologies we have obtained a better insight of the binding preferences and specificities of ribonucleoprotein complexes. These advances include many high-throughput methods, that can identify RNAs bound by specific proteins *in vivo*, methods that can predict the RNA binding potential of a protein, and methods that identify the RNA binding sites on a genome-wide scale.

5.1 RNA-binding proteins and modes of binding

RBPs are a class of proteins that have the ability to bind RNA. It has been estimated that the human genome contains around 2000 RBPs, however the estimation of their exact number, nature and function has been proven to be a difficult task (Hentze et al., 2018). Many RBPs are able to interact with their RNA targets through a set of structurally well-defined RNA binding domains (RBD). Some examples of these domains are RNA recognition motif (RRM), hnRNP K homology domain (KH), DEAD motif, double-stranded RNA-binding motif (DSRM) or zinc-finger domain. These domains can work independently or synergistically during the interaction event. They often occur multiple times in the sequence and can exist as a combination of different RBDs in order to engage RNA, which may happen in sequence and/or structure specific manner.

Although initially RBDs were considered the key factors to recognize the RNA targets, recent advances in determining the elements and structures of protein-RNA complexes have revealed the existence of a high number of interactions that do not require the presence of canonical RBDs (Castello et al., 2016; Cirillo et al., 2013). These findings indicate that inferring RNA binding from the protein sequence alone is not a trivial task and that there are still many open questions about the mechanism by which proteins bind RNAs.

5.2 Experimental methods to study protein-RNA interactions

There has been a large advance in the methods for studying the physical interactions between RNA and protein. Regarding to the target molecule of interest, these methods can be classified to (1) RNA-centric methods, when they study the proteins that bind a target RNA; and (2) protein-centric methods, when they analyze the RNAs that bind a target protein. Each method has particular advantages and drawbacks, and thus their selection must be tailored to the relevant biological question.

5.2.1 RNA-centric methods

In RNA-centric approaches, the goal is to identify the binding positions of a broad spectrum of RBPs on a specific transcript or set of transcripts and can be subdivided in to two categories (Marchese et al., 2016):

- *in vitro* approaches, where a tagged RNA construct is generated and bound to a solid support. Cell lysate is prepared and proteins from lysate are captured using the tagged RNA *in vitro*.

- *in vivo* approaches, where the target RNA is crosslinked to specific interacting RNA-binding proteins in living cells using UV, formaldehyde or other cross-linkers. Cells are lysed and the RNA-protein complexes captured from solution.

In both cases, the complex is washed to remove non-specific interactions, and finally mass spectrometry is most commonly used to identify the bound RBPs.

5.2.2 Protein-centric methods

Protein-centric methods start with a protein of interest to characterize its interaction with RNA. These approaches can be then classified as *in vitro* or *in vivo* assays.

Several new *in vitro* methods allow the screening of interactions between proteins and libraries of randomly generated RNA sequences. For the interaction analysis they combine the use of microarray and microfluidic platforms with molecule fluorescent labelling and RNA sequencing technologies. RNAcompete for instance, is a high-throughput *in vitro* binding assay that captures a more complete specificity profile by quantifying the relative affinity of an RBP to a pre-defined set of 250,000 RNA fragments (Ray et al., 2017).

On the contrary, it is common of *in vivo* assays to either directly purify the protein to find associated RNAs or use selective chemical modification of RNA in a way that relies on its association with the protein of interest (McHugh et al., 2014).

The overwhelming majority of studies that identify RNAs bound to a given protein employ CLIP-seq, term that embraces a set of methods based on UV-crosslinking followed by protein immunoprecipitation and sequencing. The UV-crosslinking consist in radiate an *in vivo* sample with UV light at approximately 254nm to create covalent bonds between RNA and protein (Smith and Aplin, 1966). Then the RNA is fragmented by a RNase treatment and the protein of interest is immunoprecipitated together with its associated cross-linked RNAs. These RNAs are then reverse transcribed, PCR amplified and finally high throughput sequenced to retrieve reads that uniquely map to the genome. Bioinformatics analysis is then used to map reads back to their transcripts of origin and identify protein binding sites (Kishore et al., 2011; Zhang and Darnell, 2011).

In order to overcome limitations of the first employed CLIP-seq protocol such as the DNA mutations caused by UV light (König et al., 2012), a vast number of alternative methods have been proposed and from which PAR-CLIP, iCLIP and eCLIP are the most common.

- In PAR-CLIP (photoactivatable ribonucleoside CLIP) protocol, cells are preincubated with photo-reactive ribonucleosides, which enables the use of UVA light (365 nm) for crosslinking (Spitzer et al., 2014). Interactions are isolated and the protein linkages are removed allowing RNA purification and reverse transcription, while mutations are left by the linkages at the interaction points. After sequencing, these characteristic mutations are easily identified against the reference sequence, allowing single-nucleotide resolution of binding events.

Unfortunately, this protocol is restricted to conditions that allow RNA alteration such as cell culture and single-celled organisms.

- iCLIP (individual-nucleotide resolution CLIP) can be used in most experimental systems (Konig et al., 2011). It uses a 3’ exonuclease to degrade protein-bound RNA. This enzyme digests the isolated RNA but stops at the cross-linked protein. An adapter is then ligated to this position. In order to recover truncated cDNA, which may constitute a large fraction of the total cDNA fragments, a single adaptor is ligated to the 3′-end of RNA fragments before reverse transcription. After circularization, re-linearization, reverse transcription and sequencing, the presence of this adapter in the sequence immediately follows the exact binding site in the RNA.

- In the case of eCLIP (enhanced CLIP) notably, adaptors are ligated first at the 3′-end of RNA and next at the 3′-end of the cDNA, hence bypassing a relatively low-yield circularization step (Van Nostrand et al., 2016). In addition, eCLIP includes a parallel analysis of the size-matched input (SM-input) control to identify the most abundant non-specific RNA fragments contributing to background signal.

A common difficulty in CLIP-seq methods is the amount of immunopurified cross-linked RNA, which can become a problem due to poor crosslinking efficiency or low RNA–ribonucleoprotein complex abundance. If sufficient UV cross-linked complexes can’t be purified then the standard method is RIP-seq, which conceptually can be thought as a CLIP-seq method without the removal of non-crosslinked RNAs, but with the expense of lower signal-to-noise ratios.

More recently two protein-centric methods appeared that do not require protein purification or UV cross-linking and rely on RNA chemical modifications. In TRIBE method, the RBP of interest is coupled to the catalytic domain of the Drosophila RNA-editing enzyme ADAR and the fusion protein is expressed *in vivo* (McMahon et al., 2016). RBP targets are marked with novel RNA editing events and identified by RNA sequencing. In RNA-tagging protocol the RBP is fused to the enzyme poly(U) polymerase, which adds poly(U) tails to bound RNAs. These tagged RNAs are then identified from a pool of total RNA using both targeted and high-throughput assays.

5.3 Computational methods to study protein-RNA interactions

Despite the technical advances mentioned above, the experimental time, effort, and expenses have created a demand for computational methods that can predict the binding partners or sites in RNA-protein complexes. Computational tools are particularly useful to predict potential ribonucleoprotein associations and to narrow down a list of interaction partners for experimental validation, inexpensively and quickly.

The majority of these tools make use of features that can be derived either from the sequence of the protein and RNA or from the structure. Respectively, depending on the kind of features exploited, the computational methods can be classified as sequence-based and/or structured-based methods. Sequence based methods take advantage of the information collected within primary sequences of protein and RNA. In general, statistical analysis of a large collection of sequences (training data) known to be involved in an interaction leads to the creation of a model that is further used to identify novel binding partners or binding sites. In contrast, structure-based methods use the geometric shapes of protein and/or RNAs to derive this information. Both methodologies however can return binary predictions (binding or not binding) or scored based predictions (e.g. affinity of interaction)

5.3.1 Sequence-based features of proteins and RNA

The methodologies that fall in this category extract features and properties for each molecule (protein/RNA) looking at their primary sequence. The most common features that they use are:

- Amino acid composition

The simplest way to encode a sequence of amino acids, each of which can have 20 different values, is by standard binary encoding also known as one-hot encoding, which encodes each amino acid into a 20-dimensional binary vector $\nu(\alpha)$. This vector, representing an amino acid of type α, is defined by a binary encoding, with $v(i) = 1$ if $i = a$, and zero otherwise. A sequence S of length N is thus represented with an array of numerical values of size 20N (Baldassi et al., 2014; Jones et al., 2012).

- Chemical and Physical Features

There are more than hundred different physicochemical features for each amino acid. Especially, hydrophobicity, structural disorder and polarity are relevant to characterize the RNA-binding ability of proteins and thus the use of the corresponding scales is quite common among computational tools. Methods that consider this kind of information usually translate the sequence of amino acids to a sequence of values derived from the corresponding physicochemical scale, or make use of encoding methodologies such as for instance encoding by Composition, Transition and Distribution (CTD) (Govindan and Nair, 2011).

- Sequence Similarity

Sequence similarity (also referred to as sequence conservation) is frequently used for RNA-binding site prediction. The BLAST and PSI-BLAST programs are used to compare the similarities among various protein sequences. Generally, multiple sequence alignment (MSA) were obtained by comparing query sequences against the NCBI non-redundant database and if the homologous sequences are known to be RNA-binding, then the query protein can also be regarded as an RNA-binding protein (Kumar et al., 2008).

- Evolutionary Information

Evolutionary information has often been introduced in functional site predictors in recent studies, including RNA-binding site prediction. Previous studies showed that Position-Specific Scoring Matrix (PSSM) (an important form of evolutionary information) greatly improved the performance of RBPs prediction. PSSMs were used widely in pervious prediction studies because they provide the likelihood of a particular residue substitution based on evolutionary information (Fernandez et al., 2011).

5.3.2 Structure-based features

- The Secondary Structure (SS)

The secondary structure (SS) provides local and geometric patterns, which can be obtained in two ways: i) if the structure is available, the real SS could be obtained by publicly accessible databases, like the Protein Data Bank (PDB); ii) if not, protein or RNA SS can be predicted by a vast number of methods. For instance, in the case of RNA, there two main approaches that can infer the SS, thermodynamics-based approaches, such as Vienna package and phenomenological potentials such as CROSS (Delli Ponti et al., 2017; Gruber et al., 2015).

- Accessible Surface Area (ASA)

The accessible surface area (ASA) or solvent-accessible surface area (SASA) is the surface area of a biomolecule that is accessible to a solvent. Since binding residues tend to be exposed in order to interact, calculation of solvent accessibility focus on predicting the binding-site of the interaction. The relative ASA could be calculated if the molecular structure is available or predicted otherwise (Faraggi et al., 2014).

The Table 2 summarizes some of the most popular computational methods for the identification of RNA-binding proteins and protein-RNA interactions.

5.3.3 Methodological Approaches

5.3.3.1 Machine – learning

The majority of the computational methods to study protein-RNA interactions use machine-learning. The term “machine learning” refers to a broad list of computational algorithms able to derive from a data set relevant prediction patterns that are not known in advance. In this sense, the algorithm “learns” the underlying rules from the data itself.

Machine-learning algorithms can be divided into unsupervised and supervised super-classes. An unsupervised algorithm tries to make sense out of un-labelled data by extracting features and patterns on its own, common examples include the Principle Component Analysis and the k-means clustering. Interestingly, unsupervised clustering is important for the analysis of single-cell RNA sequencing (scRNA-seq) data as it provides large catalogues detailing the transcriptomes of individual cells and to define cell types.

In contrast, supervised algorithms predict an output based on labelled data. Then they can use this output to evaluate their performances by comparing it with the training data. The most widely used algorithmic methodology are the support vector machines (SVMs), that have as their fundamental basis the estimation of a threshold that can maximally separate the labelled classes of the training set into the feature space. An example is RNApred which combines amino acid composition and PSSM profiles and uses the SVM method to discriminate between RBPs and non-RBPs (Kumar et al., 2011).

5.3.3.2 Artificial Neural networks and deep learning approaches

In recent years, advanced algorithms based upon Artificial Neural Networks (ANNs) are becoming increasingly popular in studying RNA-protein interactions at a transcriptome wide level (Zhang et al., 2016). As the name suggests, ANNs were inspired by the biological function of neurons as they operate in image processing (Cao et al., 2018). In the context of ANNs, a “neuron” takes multiple data inputs (analogous to neurotransmitters within a synapse) and applies a weight to each signal to provide information for the so-called activation function. Depending on the application, the output may be either binary or continuous. A neural network can be constructed by grouping many such computational neurons in layers, so that the output from one neuron may be used as the input in the next layer. The layers between the input and output units are referred to as “hidden”, as the values within are not observed at the input or output data.

Deep Neural Networks (DNNs) can be understood as an ANN consisting of several non-linear layers, that is, the activation function is nonlinear, and the neural network may contain loops or cycles between layers. The deep learning architecture seeks to cyclically optimize the weight parameters in each layer. One of the first attempts on integrating deep-learning concepts in bioinformatics was the DeepBind approach that allows prediction of sequence specificities of DNA and RNA-binding proteins (Alipanahi et al., 2015).

The success of machine learning and deep learning approaches on predicting interactions between transcripts and proteins is largely dependent on the availability of large training datasets which usually come from multiple experiments (i.e. the RNA sequence specificities for an RBP of interest). Thus, a major limitation of these approaches comes when the training data has limited size or is biased. In the first scenario, when the size of the training data is small, the model will under-fit, meaning that the model hasn't seen and trained on enough data, thus being unable to make accurate predictions. On the other hand, when the training data is biased towards some specific data points, will lead the model to overfit, which essentially means that the function of the algorithm has closely fitted to a limited set of data points and cannot successfully generalize. Consequently, there is a huge discussion and research in order to regularize the predicting methods and avoid overfitting and underfitting.

5.3.4 The catRAPID approach

*cat*RAPID is an algorithm that was developed in our group to evaluate the interaction propensities of polypeptides and nucleotide chains (Bellucci et al., 2011). The features that it uses are the physicochemical properties of the two chains as well as structural information, that in the case of the RNA, is predicted by Vienna package (see chapter 5.3.2), while in the case of the protein is derived directly from the sequence. *cat*RAPID was trained on a large set of protein–RNA pairs available in the Protein Data Bank (Berman et al., 2000) to discriminate interacting and non-interacting molecules and can be applied to predict the protein associations with coding and non-coding RNAs. When the input sequences exceed the length compatible with the computational requirements (i.e.: protein length > 750aa or RNA length > 1200 nt), *cat*RAPID cannot be directly used to calculate the interaction propensity. To overcome this limitation, a procedure called fragmentation was developed, which cuts polypeptide and nucleotide sequences into fragments followed by the prediction of the individual interaction propensities. Chapters I and II present two complementary methods for integrating the individual scores coming from the fragments into one representative score.

Prediction	Examples	Advantages	Disadvantage	References
Binding motif (RNA)	MEME	*de novo* binding site discovery	High-throughput data are required as input	(Bailey et al., 2009)
	SeAMotE		Sequence complexity is a limitation	(Agostini et al., 2014)
Binding residue	Pprint	Evolutionary information	RNA-binding domains cannot be identified	(Kumar et al., 2008)
	BindN+			(Wang et al., 2010)
	RNAbindR+			(Walia et al., 2014)
Domain (protein)	HMMER	Domain recognition	Annotation of RNA-binding domains are required	(Finn et al., 2011)
	*cat*RAPID *signature*	Annotation of RNA-binding domains are not required	Single amino acid resolution has not been implemented	(Livi et al., 2016)
RNA–protein interaction	*cat*RAPID	Runs on high-throughput data	RNA < 1200 nt	(Agostini et al., 2013b; Bellucci et al., 2011)
			Protein < 750 aa	
	RPISeq	High sensitivity	Low specificity	(Muppirala et al., 2011)
			Max 100 sequences per run	

Table 2. List of Computational Methods for the Identification of Protein–RNA Interactions

6 Beyond the Protein-RNA complex: Membrane-less organelles

Organization of the densely packed intracellular environment requires compartmentalization. This is particularly important for gene expression as coordinated processes must occur in an ordered fashion. In eukaryotic cells, double stranded DNA (dsDNA) is sequestered in the nucleus and packaged by histones. Within the nucleus, DNA is organized into heterochromatin and euchromatin to control the relative access to the transcriptional machinery. Transcribed mRNA undergoes splicing, polyadenylation, and capping prior to export to the cytoplasm. Each of these processes is under spatiotemporal control that ensures correct processing and localization.

Just as membrane-enclosed organelles (e.g., nuclei, mitochondria, endoplasmic reticulum, golgi apparatus) serve to organize biological processes into discrete cellular domains, non-membrane enclosed domains organize biological activities throughout the cell. These assemblies have been conceptualized as Ribonucleoprotein (RNP)-containing hubs. There, complex biochemical reactions take place and are referred to as membrane-less organelles (MLO) due to their ability to concentrate factors associated with a biological process. MLOs have been found either to the nucleus or in the cytoplasm. Several of those that occur in the nucleus include the nucleoli, nuclear Speckles and Cajal bodies while common examples in the cytoplasm are the processing-bodies (P-bodies) and the stress granules (Brangwynne et al., 2011; Nott et al., 2015).

Regardless of their sub-cellular location, there are three basic principles underlying the formation of membrane-less organelles. They arise from a phase separation of proteins or proteins and nucleic acids from the surrounding milieu. They remain in a liquid state but with properties distinct from those of the surrounding matter and importantly, proteins exchange with these bodies in a matter of seconds opposite to what is observed in stable complexes, supporting the notion of constant access within these highly concentrated molecular assemblies (Brangwynne, 2013; Brangwynne et al., 2011, 2009; Weber and Brangwynne, 2012). In addition to the aforementioned principles, various MLOs appear inherently structured with cores surrounded by shells, suggesting not just RNP aggregation but the existence and maintenance of higher-order structures as a result of liquid-liquid phase separation (LLPS) of different components (Feric et al., 2016; Jain et al., 2016; Wheeler et al., 2016).

From the observations described above, a picture emerges suggesting that the material and the interactions among the molecules in a given MLOs determines its biological functionality. Importantly, MLOs can lose their liquid-like characteristics by transitioning into a more rigid and gelatinous state (Qamar et al., 2018). Indeed, evidence is accumulating indicating that both aberrant formation of MLOs and imbalances between liquid-like and solid-like states of particular MLO components could be crucial for the cause of many diseases. For instance, many neurodegenerative diseases (i.e., Parkinson’s Disease, PD; Amyotrophic

Lateral Sclerosis, ALS; Frontotemporal Lobar Degeneration, FTLD), amyloidoses, prion diseases as well as a number of inherited myopathies are presently characterized as stress-induced protein conformational disorders or proteinopathies (Wolozin, 2012).

6.1 Phase separation – Protein Characteristics

Different groups have attempted the identification of the protein components of membrane-less organelles (Andersen et al., 2005; Boke et al., 2016; Fong et al., 2013; Jain et al., 2016) and define the molecular determinants of phase separation. These studies suggest that multivalency, which refers to the effective numbers of adhesive domains/motifs that provide specificity in intra- as well intermolecular interactions, is a defining feature of proteins (and perhaps RNA molecules) that drive phase transitions. Multivalency can come about in at least one of three ways:

(i) folded proteins, with well-defined interaction surfaces, can form oligomers that engender multivalency of other associative patches, which participate in stereospecific interactions,

(ii) folded domains can be interspersed by flexible spacers to generate linear multivalent proteins and

(iii) intrinsically disordered regions (IDRs) can serve as scaffolds for multiple, distinctive short linear motifs.

The common conclusion however, of the majority of these studies is that the list of proteins that phase separate into droplets is enriched in low complexity amino acid composition domains (LCDs) including tandem repeats (TRs) of individual amino acids or amino acid motifs, such as polyglutamine (polyQ) and polyasparagine (polyN) domains (Altmeyer et al., 2015). These LCD-containing proteins belong to the general class of intrinsically disordered proteins (IDPs) and can undergo liquid-liquid phase separation more readily, either self-aggregating or upon binding to nucleic acids or other proteins as in many liquid-like RNP granules where IDRs can aid in their assembly (Decker et al., 2007; Gilks et al., 2004; Reijns et al., 2008). It is important to note that IDPs are about a third of the eukaryotic proteome (Dunker et al., 2015; Toretsky and Wright, 2014; van der Lee et al., 2014). Hence, we can speculate that many unique and uncharacterized liquid droplets could exist and efforts in order to characterize them will help in our understanding of the organization of matter in the cell.

6.2 Phase separation – The RNA view

Despite the flurry of all these studies demonstrating the importance of IDPs in the process of phase separation, the role of RNAs is yet less understood. MLOs frequently contain nucleic acids, especially RNA and the proteins associated with them often possess RNA-binding domains or motifs (Boeynaems et al., 2017; Jain et al., 2016), suggesting that the RNA has an important role in the formation of MLOs. On the one hand, many nuclear MLOs form in coordination with transcription of specific RNAs. For instance, the assembly of *nucleoli* is coordinated by prerRNA transcription (Falahati et al., 2016) and nuclear paraspeckles are formed exclusively from NEAT1 transcription sites (Clemson et al., 2009). The function of these RNAs has been summarized with the term "architectural RNAs" (arcRNAs) (Yamazaki et al., 2018) and by dynamically regulating their transcription and turnover, the formation and dissolution of nuclear MLOs can be achieved. On the other hand, the formation of cytoplasmic MLOs is probably primarily governed by the availability of existing RNPs, proteins, or RNAs serving structural functions. However, regardless of whether RNA itself scaffolds the formation of MLOs or becomes recruited into an existing network, its content affects the properties of assembled MLOs, such as for instance the material exchange rates as well as their rigidity and their shape (Audas et al., 2016; Elbaum-Garfinkle et al., 2015; Jain and Vale, 2017; Langdon et al., 2018; Van Treeck et al., 2018; Zhang et al., 2015). More accurately, RNAs facilitate LLPS of particular IDR-containing proteins by reducing the protein concentration needed for their phase-separation (Burke et al., 2015; Lin et al., 2015; Molliex et al., 2015). For example in the case of fragile X-associated tremor/ataxia syndrome (FXTAS), our group reported that FMR1 is responsible for the sequestration and finally co-aggregation with TRA2A (Cid-Samper et al., 2018). This sequestration is depended to the presence of CGG repeats present in the 3' UTR of FMR1. TRA2A shows diffuse pattern in cells that do not overexpress those repeats while opposite and granular patterns are identified in a FXTAS permutation carrier with overexpressed CGG repeats.

Overall, a main driver that influences this formation is concentration. Interestingly, a recent study showed that several RBPs implicated in neurodegeneration, phase-separate *in vitro* in concentrations similar to those physiologically found in the nucleus (Maharana et al., 2018). On the contrary, inside cellular nucleus the same proteins remained diffused even under cellular stress conditions, that would drive their condensation into cytoplasmic stress granules. The authors showed that in vitro high amounts of short, nonspecific RNAs keep prion-like RBPs soluble. Conversely, longer RNAs that can form higher order secondary structures and which specifically bind RBPs promote phase-separation. For example, highly structured RNAs such as Neat1 act as scaffolds that promote the nucleation of condensates in the high–RNA concentration environment of the nucleus. A similar scenario may apply for stress granules in the cytoplasm, which contain large amounts of structured polyadenylated mRNAs (Cerase et al., 2019).

6.3 The role of RNA structure

RNAs form various structural elements, often in well-defined contexts, such as short and long stems, hairpins, helical regions, tetra-loops, or G-quadruplexes, which contribute to the overall complexity of the three-dimensional space a given RNA can obtain (Miao and Westhof, 2017). In addition, these structural motifs also attract non-specific protein binders, potentially through interactions with IDRs or Prion like Domains (PrLDs) (Jankowsky and Harris, 2015). Importantly, since the number of putative RNA sequence motifs that could be recognized by proteins vastly exceeds the number of known RNA-binding proteins it is likely that RNA structure plays a more decisive role than the sequence context in discriminating protein-RNA interactions (Singh and

Valcárcel, 2005). Indeed, particular protein domains that bind nucleic acids recognized structure over sequence identity (Ding et al., 2014; Taliaferro et al., 2016). For example, evolutionary conserved as well as repetitive sequences in the lncRNA NEAT1 were shown to be essential for protein binding resulting in nuclear paraspeckle formation (Yamazaki et al., 2018). Moreover, structural changes in specific mRNAs can influence MLO identity. More specifically, protein droplet identity was not only established through intermolecular mRNA-mRNA interactions, but particular RNA structures selectively exposed or masked RNA sequences capable of interacting with other RNAs, thereby directing mRNAs into specialized MLOs. Formed MLOs further became stabilized through additional interactions with RRM- and IDR-containing proteins (Langdon et al., 2018).

A particular RNA structure is only partially determined by primary sequence context as it represents an equilibrium of possible structures (Lorenz et al., 2016), which can be affected by various parameters. Charge, temperature, ion concentrations, nucleotide modifications (Ries et al., 2019) and interactions with proteins allow major changes of RNA structure in response to, for instance, environmental signals or stress conditions (Boccaletto et al., 2018). It follows that the propensity of RNAs to self-assemble and form higher-order structures is likely one of the defining properties for its influence on LLPS and MLO dynamics (Sanchez de Groot et al., 2019)

Chapter I – *Global Score:* Quantitative predictions of protein interactions with long noncoding RNAs.

The long-noncoding RNA *Xist,* the master regulator of the Mammalian female-specific process of X Chromosome Inactivation (XCI) was identified almost 25 years ago. At the onset of X inactivation, *Xist* spreads in *cis* on the future inactive X and triggers gene silencing by recruitment of repressive DNA and chromatin modifiers. We are still just beginning to understand *Xist* network of interactions. In fact, five genomic and proteomic studies recently revealed a quite heterogeneous list of *Xist* binding proteins indicating that there remain much to learn about how and with which partners it interacts.

The *cat*RAPID *Global Score* method based on the *cat*RAPID *fragment* algorithm (Cirillo et al., 2013) was applied to identify specific and direct associations. It was an ongoing project when I first joined Tartaglia's lab where Davide Cirillo had started the research on exploring the protein interactome of *Xist.* I was involved in a series of data analysis of that project mainly to validate our predictions for Spen, Hnrnpk, Lbr, Ptbp1 and Hnrnpu/Saf-A proteins using the eCLIP data. The computational method and pipeline in this work was applied to the study of other lncRNAs in other works.

Cirillo, Davide, Mario Blanco, Alexandros Armaos, Andreas Buness, Philip Avner, Mitchell Guttman, Andrea Cerase, and Gian Gaetano Tartaglia. 2017. "Quantitative Predictions of Protein Interactions with Long Noncoding RNAs." *Nature Methods* 14 (1): 5–6. https://doi.org/10.1038/nmeth.4100.

Cirillo D, Blanco M, Armaos A, Buness A, Avner P, Guttman M, et al. Quantitative predictions of protein interactions with long noncoding RNAs. Nat Methods. 2017 Jan 29;14(1):5–6. DOI: 10.1038/nmeth.4100

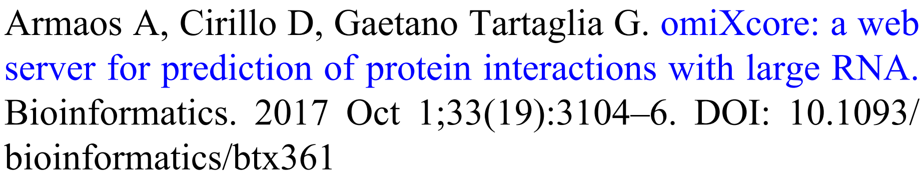
Chapter II – *omiXcore:* Predicting Protein-RNA interaction affinity based on eCLIP data.

omiXcore was developed during the second year of my PhD and was published in Bioinformatics in 2017. This approach is the latest addition in the *cat*RAPID suite and aims to work specifically with large transcripts including coding and non-coding RNAs. As in the case of *Global Score* (Cirillo et al., 2016) approach, *omiXcore* was based on the *cat*RAPID *fragment* algorithm (Cirillo et al., 2013) which includes the division of the polypeptide and nucleotide sequences in overlapping fragments when their lengths exceed the system limitations.

The non-linear implementation captures the contribution of the individual scores coming from the fragments while providing an integrated and unique score that resembles the interaction affinity of a protein – RNA pair. *omiXcore* was trained on the enhanced UV CrossLinking and ImmunoPrecipitation (eCLIP) (Van Nostrand et al., 2016) data from the ENCODE project. The use of eCLIP read counts that were normalized to transcript's expression levels, allowed the training of the algorithm on a score in a continuous range that approximates the interaction affinity.

Similarly to *Global Score, omiXcore* has wide applicability on transcripts without length restrictions and the pre-compiled library that is available through the webserver, allows for further exploration of long intergenic RNAs and candidate prioritization for further experimental validation.

Armaos, Alexandros, Davide Cirillo, and Gian Gaetano Tartaglia. 2017. "OmiXcore: A Web Server for Prediction of Protein Interactions with Large RNA." *Bioinformatics (Oxford, England)*, June. https://doi.org/10.1093/bioinformatics/btx361.



Armaos A, Cirillo D, Gaetano Tartaglia G. omiXcore: a web server for prediction of protein interactions with large RNA. Bioinformatics. 2017 Oct 1;33(19):3104–6. DOI: 10.1093/bioinformatics/btx361

Chapter III – A hypothesis: X Chromosome silencing through phase-separation

Phase separation allows for functional compartmentalization in the cell, resulting in droplets where key factors are concentrated, thereby facilitating biochemical processes (see Introduction). Proteins with intrinsically disordered or low complexity domains interact to form 'hubs,' which are ensembles of phase-separated molecules with hydrogel-like properties (Shin and Brangwynne, 2017). The assembly of phase separated compartments is facilitated by the presence of RNA (Molliex et al., 2015). Many proteins with disordered regions interact with RNA (Castello et al., 2013a; Livi et al., 2016), suggesting that the diversity of lncRNA sequences, expression patterns, and protein-binding properties might contribute to specifying compositionally and functionally distinct phase-separated compartments. Two examples include MALAT1 and NEAT1 lncRNAs that are important for the assembly of two phase-separated bodies in the nucleus, speckles, and paraspeckles, respectively.

This study was motivated by recent advances in the characterization of phase-separation in the cell. Also here I used predictions carried out with *Global Score* approach (Chapter I) to propose that the X chromosome heterochromatisation in female mammals is facilitated by the process of liquid-liquid phase-separation. This work occupied the 3rd and 4th year of my PhD and was published in Nature Structural and Molecular Biology during 2019.

Cerase, Andrea*, Alexandros Armaos*, Christoph Neumayer, Philip Avner, Mitchell Guttman, and Gian Gaetano Tartaglia. 2019. “Phase Separation Drives X-Chromosome Inactivation: A Hypothesis.” *Nature Structural & Molecular Biology* 26 (5): 331–34. https://doi.org/10.1038/s41594-019-0223-0.

Cerase A, Armaos A, Neumayer C, Avner P, Guttman M, Tartaglia GG. Phase separation drives X-chromosome inactivation: a hypothesis. Nat Struct Mol Biol. 2019 May 6;26(5):331–4. DOI: 10.1038/s41594-019-0223-0

Chapter IV – RNA structure drives protein interaction

The structure of RNA molecules is known to be involved in gene regulation through RNA stabilization and localization (Goodarzi et al., 2012) while it is also known to be critical to the biogenesis and function of many noncoding RNAs. During the second half of my PhD I investigated the relationship between RNA secondary structure and its ability to bind proteins in multiple low- and high-throughput data. Interestingly and independently of the experimental data I used (PARS, DMS, microarray, X-ray, NMR, eCLIP, PAR-CLIP, HITS-CLIP and iCLIP), the algorithms that I employed (*cat*RAPID and RPISeq as well as CROSS to mimic SHAPE data) or the organism I studied (PDB database), I found that the number of protein contacts correlates with the RNA structural content. This interconnection suggests that the stable and less variable conformations of structured RNAs create well-defined binding sites that promote specific interactions, with functional roles in gene regulation.

To validate the results of my analysis, the senior post-doc of Tartaglia's lab, Natalia Sanchez de Groot, introduced an experiment showing that a structured mRNA, Hsp70, has the ability to reorganize the composition of a protein aggregate.

This work that was co-authored by Natalia and me was published in Nature Communication in 2019.

Sanchez de Groot, Natalia, Alexandros Armaos, Ricardo Graña-Montes, Marion Alriquet, Giulia Calloni, R. Martin Vabulas, and Gian Gaetano Tartaglia. 2019. “RNA Structure Drives Interaction with Proteins.” *Nature Communications* 10 (1): 3246. https://doi.org/10.1038/s41467-019-10923-5.

Sanchez de Groot N, Armaos A, Graña-Montes R, Alriquet M, Calloni G, Vabulas RM, et al. RNA structure drives interaction with proteins. Nat Commun. 2019 Dec 19;10(1):3246. DOI: 10.1038/s41467-019-10923-5

Chapter V – Discussion

In this thesis, I presented two distinct but intimately connected topics: the ability of RNA molecules to interact with proteins and their potential to phase-separate in processes such as the X chromosome inactivation.

The first topic occupied two separate periods of my PhD. During the first period which expanded during the first two years, I was involved in the development of methods to characterize ribonucleoprotein associations (Chapters I, II). In the second period, I combined my predictions with findings that I obtained through the analysis of multiple and heterogeneous experimental data with the purpose to investigate the association of RNA secondary structure with its ability to bind proteins (Chapter IV).

The second subject of my thesis presents my hypothesis on the process of the X chromosome inactivation. This study was motivated by recent advances in the characterization of phase-separation in the cell. Here I used predictions carried out with *Global Score* approach (Chapter I) to propose that the X chromosome heterochromatisation in female mammals is facilitated by the process of liquid-liquid phase-separation (Chapter III).

The link between the two topics is particularly relevant if one considers that *Xist* lncRNA is an RNA with very well conserved and structured regions that promote stable and specific interactions with proteins (Delli Ponti et al., 2018). A high percentage of *Xist* direct interactors (40%), as identified in the first Chapter, are predicted to contain IDRs or have partners that have high intrinsic disorder potential. This local increase in local concentration of disordered proteins, directly or indirectly associated with *Xist*, is suggested to promote phase separation of the *Xist* RNA and its protein interacting network from the surrounding milieu (Cid-Samper et al., 2018).

Computational methods: An essential source of information

The role of protein-RNA interactions has been intensively studied for its centrality in transcriptional and post-transcriptional events (Bernhardt, 2012; Keren et al., 2010). RNA-binding proteins (RBPs) are present in every aspect of RNA biology, from transcription, pre-mRNA splicing and polyadenylation to RNA modification, transport, localization, translation and turnover. The RBPs not only influence each of these processes, but also provide a link between them (Hilleren et al., 2001; Kyburz et al., 2006; Millevoi et al., 2006; Rigo and Martinson, 2008). Proper functioning of these intricate networks is essential for the coordination of complex post-transcriptional events, and their perturbation can lead to diseases. The human genome harbors around 2000 genes encoding known RBPs or proteins annotated to contain at least one RNA-binding domain RBD (Castello et al., 2013b; Hentze et al., 2018). Nonetheless, the number of proteins with identified RNA-binding ability, either possessing canonical or non-canonical RBDs, is increasing and the fact that some proteins are able to bind to RNA with domains or regions that are not specifically evolved to this precise purpose is quite intriguing (Castello et al., 2016; Livi et al., 2016). Additionally, the recent discovery of a plethora of RNAs, including long non-coding RNAs (lncRNAs) and other previously uncharacterized transcripts (Iyer et al., 2015) like NEAT1 (Yamazaki et al., 2018) or SAMMSON (Vendramin et al., 2018), demanded a re-examination of biological processes and biological networks to include these new effectors in the established protein-centric landscape

Nonetheless, relying exclusively on experimental approaches in order to characterize the elements of the above-mentioned biological networks could result in misleading assumptions. For instance, despite the recent advances of CLIP-seq approaches (see Introduction), it remains difficult to simultaneously detect the many RBPs bound to a single transcript and the RNA regions that are likely to be involved in the binding. The reason behind that is that CLIP-seq protocol and procedure, similarly to any other RNA-seq experimental approach, suffer from sensitivity or specificity biases due, in part to noise in the employed experimental approaches but also due to known inaccuracies in the experimental protocols (Chakrabarti et al., 2018). In more detail, HITS-CLIP and PAR-CLIP suffer from limited sensitivity due to the loss of cDNAs truncated at crosslink sites (Chakrabarti et al., 2018). On the other hand eCLIP achieves low specificity since the protein-RNA complex is not validated and because the blind cutting from the membrane, when normalizing the enrichment scores by the size-matched input (SMI), generates artefacts (Chakrabarti et al., 2018).

With all the above in mind, computational models represent an important source of information that can be exploited to identify hidden trends, interpret experimental results and understand the basics of molecular recognition. Computational tools have the advantage to perform exhaustive analyses and extract distinctive features fast and inexpensive, facilitating the design of new experiments. Experimental studies and computational analyses, such as those presented in this thesis, aim to provide compelling insights into the rules that govern RNP formation.

From local to global predictions of protein-RNA interactions

We previously developed *cat*RAPID to predict the interaction propensity of protein and RNA sequences using their physico-chemical properties (Bellucci et al., 2011). The method, which was designed to complement experimental studies, has an average accuracy of 78% in predicting binding partners and works for transcripts shorter than 1000 nt due to the difficulty of modelling the structure of larger sequences. Indeed, the size of the configuration space makes structural predictions difficult for thermodynamic approaches (HafezQorani et al., 2016; Lange et al., 2012).

Previous pilot projects indicate that division of sequences into sub-elements is useful to identify contacting regions (Cirillo et al., 2013; Zanzoni et al., 2013). By fragmenting protein and RNA sequences, it is possible to detect the binding sites of Fragile X mental retardation protein FMRP, TAR-DNA binding protein 43 TDP-43 and

Serine/Arginine splicing factor 2 SRSF2 (Agostini et al., 2013a, 2013b). Yet, when proteins bind with low affinity to multiple regions of RNA sequences, identification of binding regions cannot be directly exploited to predict the binding strength between two molecules. For instance, Histone-lysine N-methyltransferase Ezh2 is predicted to associate with *Xist* in several sites within the repetitive region A, but the interactions have low interaction propensities (Agostini et al., 2013a).

As described above, while the division of the sequences into sub-elements can be very useful to successfully detect potential binding sites, there is a major limitation when trying to estimate an overall interaction score for long RNA sequences. This is partially due to the lack of sophisticated and accurate thermodynamic approaches able to predict the folding of large RNA molecules. For instance, RNAstructure (Reuter and Mathews, 2010) and Vienna (Gruber et al., 2015), which are the golden standards for RNA secondary structure prediction, are limited by the length of the sequence, with their accuracy dropping for sequences larger than 700-1000 nucleotides (Hajiaghayi et al., 2012). Even if they are forced to process larger sequences, the computational times are so huge that make these approaches unsuitable for high throughput analysis. This drawback, makes the fragmentation procedure essential for maintaining trustable folding information and thus, the estimation of the overall interaction score should be based on the individual fragments and on the knowledge that is encoded in each of them. The key problem though, for predicting global features form local properties, is the integration of the individual signals. While knowledge of features encoded by fragments is informative, the overall context should be taken into account to accurately predict folding propensities and interaction abilities.

Obtaining an overall score is of high importance for those that wish to prioritize targets when designing experiments for further validation. A simple solution for integrating the signal from all individual fragments could be the estimation of their mean, median or maximum score. However, we demonstrated that these simple functions are suboptimal and fail to integrate all individual interaction propensities in one representative score (Figure 1). With that in mind, we proposed that the function to be applied on the individual scores should be non-linear with the aim to capture the individual contributions of each fragment on the overall binding estimation. At that point, there were two distinct but related questions that needed to be answered. The first one was whether a given protein – RNA pair has the potential to interact or not and the second one was with what affinity would they interact. Chapters I and II present my approaches to answer the above two questions. These two non-linear implementations integrate the information contained in the interaction propensities of the individual protein and RNA fragments.

The first approach, called *Global Score,* aimed to answer the first of the two aforementioned questions. Since the objective of this project was to classify protein-RNA pairs into interacting or non-interacting, *Global Score* was trained and validated using different sets of binding (positives) and non-binding (negatives) protein-RNA pairs. The classification into positives and negatives allowed to make predictions independently of the statistical distributions of experimental affinities, which are intrinsically linked to each individual technique, thus ensuring wide applicability of the approach. Additionally, *Global Score* output is a score in a contiguous range [0,1] which ensures flexibility in the training phase, as the use of a binary score would increase the number of unclassifiable cases (in between the two states).

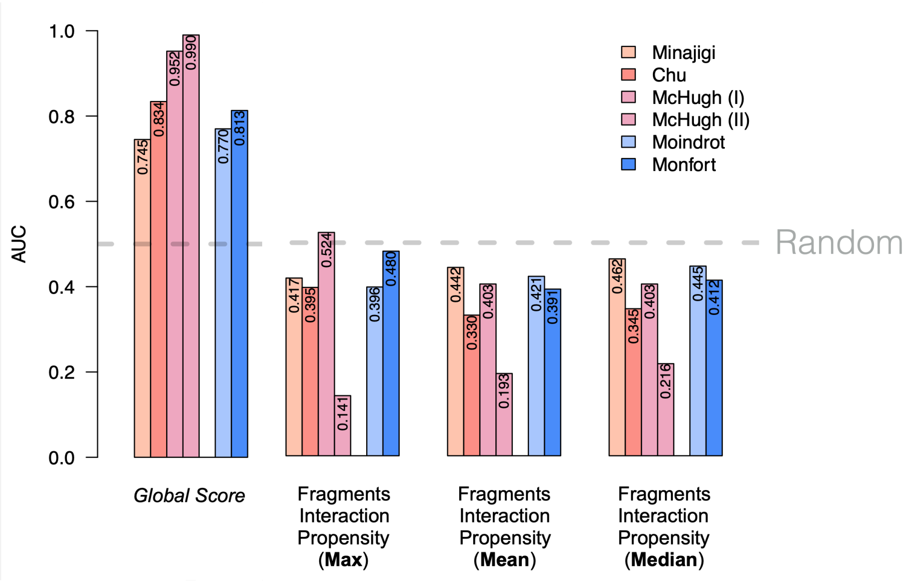


Figure 1) *Xist* interactions with RBPs reported by (Chu et al., 2015; McHugh et al., 2014; Minajigi et al., 2015) (proteomic studies) as well as (Moindrot et al., 2015; Monfort et al., 2015) (genomic studies). For each set of protein and RNA fragments, we measured mean, median and maximum of the interaction propensities calculated with *cat*RAPID (Bellucci et al., 2011). *Global Score* outperforms *cat*RAPID-based analyses for large lncRNAs

The second approach I developed is called *omiXcore* (Armaos et al., 2017). This method aimed to provide an overall estimation of the interaction affinity for a given protein – RNA pair and can be used in complementarity with *Global Score.* It was trained and validated on, by that time available and recently published, eCLIP data.

It is important to note that the use of the raw number of reads as a proxy of affinity has been proven to be inaccurate since eCLIP experiments suffer from biases related to the abundances of transcripts (Chakrabarti et al., 2018). The reason behind that is that the number of RNA-seq reads generated from a transcript is directly proportional to its relative abundance in the sample (Trapnell et al., 2012). Thus, we introduced a normalization step on the eCLIP number of reads to the expression levels before generating the training and validation sets.

In conclusion, these two methods presented in this thesis have been developed in an interesting moment of the post-genomic era. New exciting technological advances on the characterization of protein – RNA complexes have been developed such as for instance the eCLIP protocol (Van Nostrand et al., 2016) or Proximity-CLIP protocol (Benhalevy et al., 2018). Experimental and computational approaches have started to unveil the complexity of our genomes and RNA-protein interactions emerged as key events in a large number of regulatory processes (Hentze et al., 2018). I believe that my methods will provide valid assistance for the interpretation of experimental results and propose potential candidates to those that investigate the key players on the formation of RNP-complexes.

New insights into the compartmentalization inside the cell

Phase separation is now appreciated as a pervasive form of organization in the cell (Boeynaems et al., 2018). Emerging work suggests that this behavior may facilitate complex organization of assemblies with different hierarchies of shells and cores (Jain et al., 2016). The resulting biological consequences can be quite diverse and in fact there are many insights that continue to arise. For instance even the RNA, independent of protein, can assemble into liquid-like droplets, opening new dimensions of phase separation biology that have yet to be explored (Jain and Vale, 2017). In that context and motivated by the similarities with other, well studied, RNA mediated processes, I proposed in this thesis my hypothesis in the context of X chromosome inactivation. I suggest that *Xist*, together with its direct and indirect partners, promotes phase separation. These similarities can be summarized accordingly:

- *Xist foci* are similar in size and morphology to paraspeckles and stress granules;
- *Xist* contains nucleotide repeats that are present in scaffold RNAs and promote protein sequestration;
- *Xist* interactome contains components of paraspeckles and stress granules and is significantly enriched for structurally disordered proteins with a strong propensity for phase separation;
- Most importantly, binding partners of *Xist* and Neat1 diffuse in a liquid-like manner.

However adequate experiments should be conducted in order to better understand if and how *Xist* undergoes phase separation with its protein partners. Doubtless though, the number of cellular phase-separated body types appears to be growing and we are beginning to understand some fundamental facts about them. In fact, many questions remain to be answered:

- How many are the phase-separated body types and how do they interact with each other?
- How structured and dynamic are the assemblies and how heterogeneous are they with respect to composition and function?
- Do these phase-separated bodies have unique compositions or do they share components? What makes one distinct from another? As for example, two mRNA-processing bodies, P-bodies and stress granules, share a number of common components, but also have distinct protein species (Parker and Sheth, 2007; Ramaswami et al., 2013; Teixeira and Parker, 2007).
- Concerning evolution, how did phase-separated bodies evolve and if they are evolvable, what is it about them that evolves? Is the evolution and selection of some, for example LCD-containing, proteins associated with the evolution of droplets?

Towards our understanding of cell's regulatory network

RNA structure promotes protein binding.

The last Chapter of my thesis presents findings on the relationship between RNA secondary structure and the ability to bind proteins. In this work, I suggested the existence of regulation layer between structured mRNAs and their protein products. All the analyses I performed were coming from multiple and heterogenous sources (more than 10 experimental approaches), a fact that make my observations very robust.

My results indicate that RNA structure is crucial for many processes, which is not completely unexpected. In fact, many RNA related processes, such as for instance transcript stabilization are relatively independent from the primary sequence and instead they may be linked to RNA structure (Goodarzi et al., 2012). Another example are the class of non-coding RNAs where it has been shown that the RNA structure has a critical role in their function such as for instance in the case of *Xist* (Pintacuda et al., 2017) and NEAT1 (Yamazaki et al., 2018, p. 1).

As for its implication in the protein interactivity, RNA structure is known to impact RBP binding and regulation (Hiller et al., 2007; Li et al., 2010; Warf et al., 2009). RBPs can be classified according to their RNA mode of binding in two distinct categories; RBPs that bind double-stranded and RBPs that bind single-stranded regions of RNAs. In the first scenario RBPs bind paired nucleotides, a fact that makes the presence of RNA structure mandatory. Examples include DGCR8 and DICER1, important in siRNA and microRNA biogenesis (Macias et al., 2012; Rybak-Wolf et al., 2014). According to the second and most common scenario however, RBPs favor reduced base-pairing of the motif itself, but interestingly they show preference for structure at the flanking positions (Dominguez et al., 2018). In other words, single stranded RBPs bind unpaired regions that are exposed in loops or other secondary structure elements (Lunde et al., 2007). For example, large hairpin loops allow binding of multiple KH domains to the RNA as has been observed in a crystal structure of NOVA1 (Teplova et al., 2011) and in SELEX analysis of PCBP2 (Thisted et al., 2001).

Clearly from both scenarios discussed above, the presence of structured regions in the RNA molecule facilitates the binding with proteins and positively contributes to the formation of stable binding sites, even when these regions don’t overlap with the actual binding sites. On the contrary, complete lack of structure would perturbate or disfavor the protein binding, since it is linked to more flexible and variable conformations, thus a shorter residence of proteins. Moreover, it should be considered that presence of a native fold favors the formation of stable and well-defined binding sites that promote functional roles and, in turn, evolutionary selection (Seemann et al., 2017).

I hope that the results discussed in Chapter IV will provide motivation for further computational and experimental research in order to give answers to multiple open questions that have been raised:

- How far from a structured region does the binding site take place? Is there a minimum distance between them?
- Moreover, do different RNA motifs require the same amount of structural content to promote protein binding?
- From the evolutionary point of view, did the RNA structure evolve in order to make accessible specific binding sites and oppositely to mask others that would be evolutionary unnecessary or even dangerous?

New insights into the layers of cellular regulatory network.

Highly contacted proteins participate in many cellular processes and thus require to be tightly controlled (Jeong et al., 2001; Mitchell and Parker, 2014). Proteins, however are just the last element in the chain (see Introduction) which means that there might exist a regulation layer at the transcriptional level, providing control over the encoded proteins. Motivated by this hypothesis, I concluded to the second important finding of Chapter IV. More precisely, highly structured RNAs and thus highly contacted, tend to code for proteins that have as well many protein contacts. This observation is very significant as it reveals new aspects on cellular regulatory network and shows the tight relationship between the structure of mRNAs and the implication of their protein products in cellular processes.

My speculation at this point is that the aforementioned regulatory network might start even during transcription. Following up an intuition on the regulation layers described above, I suspected that there could be some control at the transcriptional level. Indeed, highly structured RNAs are highly contacted by proteins, thus they might require control even before they are transcribed. In other words, the genes from which these RNAs are transcribed should be tightly controlled. The obvious hypothesis is that these genes might be controlled by a large number of transcription factors.

The big image behind this hypothesis, is that products (proteins) that are implicated in many functions should be also regulated from their begging of their life. In detail, I speculate that genes, whose transcription is controlled by a large number of transcription factors, produce RNAs that are highly structured in order to be highly contacted by proteins and thus their life to be tightly controlled. These RNAs in turn encode for proteins that are associated with multiple functions. RNA binding proteins, chaperone proteins and stress granule proteins are some common examples of protein groups that are linked with complex processes (Boeynaems et al., 2018; Ganassi et al., 2016; Hentze et al., 2018) and that makes them candidate groups to investigate.

There has been an extended part in the introduction of the current thesis dedicated to the importance of RBPs. Similarly, chaperone proteins as discussed in the paper I authored and presented in Chapter IV, have an important molecular role. Between other functions, they promote folding of proteins into the native state (Wang and Chang, 2011) and organize the assembly of phase separate RNP assemblies (Mateju et al., 2017) . In the same way, stress granule proteins are a natural choice for the analysis since they as well, are implicated in multiple functions such as for instance the modulation of the formation of granules as a response to stress. This procedure should be carefully controlled, since perturbation of this formation might lead to neurodegenerative diseases and some cancers (Jain et al., 2016).

An optimal choice as a reference set is represented by genes from genomic regions that show an increased copy number variation (CNV) that is linked to ‘benign’ clinical interpretation. CNVs are regions of the genome that are duplicated or deleted in some individuals in a population. When the variation of a genomic region is not linked to pathogenic phenotypes we can assume that genes falling in these areas do not require tight control, because perturbations of their relative abundances are tolerable by the cellular mechanism. The aforementioned fact makes this set of genes (termed ‘Benign CNV’) an ideal background set to test my hypothesis. It is important to note however, that on the contrary to the ‘Benign CNV’ set, the prevailing hypothesis of genes in CNV regions that show pathogenic phenotype, is that it is due to their dosage sensitivity and thus I suspect that their expression to be subject of higher regulation.

In Figure 2, I present my preliminary results that seem to be quite encouraging and in the line with my hypothesis. Obviously, extensive analysis is required to prove whether the proposed structure of cellular regulatory network exists. This network has proven to be more complicated and sophisticated than initially thought to be. Thus, every step we do towards characterizing is crucial to understand the underlying processes that govern it and which perturbations are source to the pathogenesis of several diseases and disorders.

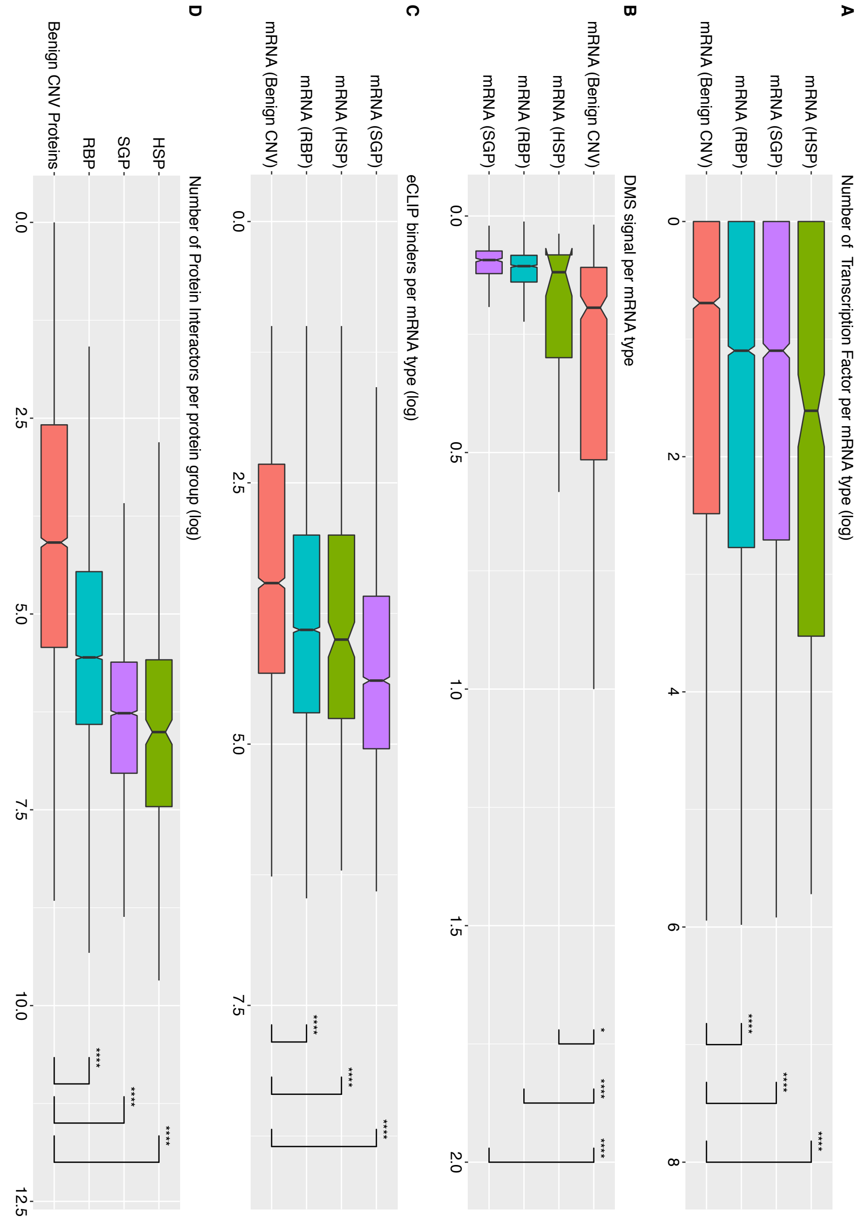


Figure 2) Comparison between four sets of genes at gene, mRNA and protein level. The four sets under comparison are i) genes coding for Heat Shock Proteins (HSP), ii) genes coding for Stress Granule proteins (SGP) (Jain et al., 2016), iii) genes coding for RBPs (Hentze et al., 2018) and iv) Benign CNV genes. **A**. Number of transcription factors binding at the promoter of each gene (log scale), **B.** Structural content measured by dimethyl sulfate modification (DMS) (Rouskin et al., 2014), **C.** Number of protein contacts for the transcribed mRNAs measured by enhanced CrossLinking and ImmunoPrecipitation (eCLIP) (Van Nostrand et al., 2016), **D.** Physical interactions of the corresponding coded proteins, collected from BioGRID; p values estimated with t-test.

Conclusions

The work carried out during my PhD can be divided in two separate stages. The first involving the development of core algorithms and the fine-tuning of these on a number of well-studied cases. The second consisting of an expansion of the current approaches to perform large-scale analysis and the ability to derive general information on post-transcriptional regulatory mechanisms. Collectively, the thesis can be summarized in the following Chapters:

- I. The development of *cat*RAPID *Global Score*, which is a sequence-based predictor that expands the applicability of *cat*RAPID method to transcripts larger than 1000nt. The method exploits the physico-chemical information derived from the primary structure of transcripts and proteins to distinguish interacting from non-interacting pairs protein-RNA. This method was trained on data coming from multiple and heterogenous sources in order to ensure its wide applicability on the identification of specific and direct associations.
- II. The development of *cat*RAPID *omiXcore*, which is as well an expansion of *cat*RAPID method. It inherits the fragmentation procedure from *cat*RAPID *fragments* approach and uses a non-linear implementation to capture the individual contributions coming from the fragments. Is a method that can be used in complementarity with *cat*RAPID *Global Score* since its aim is to provide a unique score that resembles the interaction affinity of a protein – RNA pair. For that reason it was trained on the

enhanced UV CrossLinking and ImmunoPrecipitation (eCLIP) (Van Nostrand et al., 2016) data from the ENCODE project.

- III. The application of the *cat*RAPID *Global Score* approach to investigate the direct interactors of *Xist* long non-coding RNA in a study that presents the hypothesis that *Xist* uses phase separation to perform its function. In order to formulate this hypothesis, we gathered evidence from the literature and from computational analysis and we showed that *Xist* assemblies are similar in size, shape and composition to phase-separated condensates.
- IV. The investigation of the relationship between the structure of an RNA and its ability to interact with proteins. Analyzing in silico, in vitro and in vivo experiments, we find that the amount of double-stranded regions in an RNA correlates with the number of protein contacts. This relationship -which we call structure-driven protein interactivity- allows classification of RNA types, plays a role in gene regulation and could have implications for the formation of phase-separated ribonucleoprotein assemblies. We validate our hypothesis by showing that a highly structured RNA can rearrange the composition of a protein aggregate. We report that the tendency of proteins to phase-separate is reduced by interactions with specific RNAs.

Appendix

1 Supplementary Figures for Chapter I

Cirillo D, Blanco M, Armaos A, Buness A, Avner P, Guttman M, et al. Quantitative predictions of protein interactions with long noncoding RNAs. Nat Methods. 2017 Jan 29;14(1):5–6. DOI: 10.1038/nmeth.4100

2 Supplementary Figures for Chapter II

Armaos A, Cirillo D, Gaetano Tartaglia G. omiXcore: a web server for prediction of protein interactions with large RNA. Bioinformatics. 2017 Oct 1;33(19):3104–6. DOI: 10.1093/bioinformatics/btx361

3 Supplementary Figures for Chapter IV

Sanchez de Groot N, Armaos A, Graña-Montes R, Alriquet M, Calloni G, Vabulas RM, et al. RNA structure drives interaction with proteins. Nat Commun. 2019 Dec 19;10(1):3246. DOI: 10.1038/s41467-019-10923-5

4 List of publications

Armaos, Alexandros, Davide Cirillo, and Gian Gaetano Tartaglia. 2017. “OmiXcore: A Web Server for Prediction of Protein Interactions with Large RNA.” *Bioinformatics (Oxford, England)*, June. https://doi.org/10.1093/bioinformatics/btx361.

Cerase, Andrea, Alexandros Armaos, Christoph Neumayer, Philip Avner, Mitchell Guttman, and Gian Gaetano Tartaglia. 2019. “Phase Separation Drives X-Chromosome Inactivation: A Hypothesis.” *Nature Structural & Molecular Biology* 26 (5): 331–34. https://doi.org/10.1038/s41594-019-0223-0.

Cirillo, Davide, Mario Blanco, Alexandros Armaos, Andreas Buness, Philip Avner, Mitchell Guttman, Andrea Cerase, and Gian Gaetano Tartaglia. 2016. “Quantitative Predictions of Protein Interactions with Long Noncoding RNAs.” *Nature Methods* 14 (1): 5–6. https://doi.org/10.1038/nmeth.4100.

Delli Ponti, Riccardo, Alexandros Armaos, Stefanie Marti, and Gian Gaetano Tartaglia. 2018. “A Method for RNA Structure Prediction Shows Evidence for Structure in LncRNAs,” April. https://doi.org/10.1101/284869.

Delli Ponti, Riccardo, Stefanie Marti, Alexandros Armaos, and Gian Gaetano Tartaglia. 2017. “A High-Throughput Approach to Profile RNA Structure.” *Nucleic Acids Research* 45 (5): e35–e35. https://doi.org/10.1093/nar/gkw1094.

Guiducci, Giulia, Alessio Paone, Angela Tramonti, Giorgio Giardina, Serena Rinaldo, Amani Bouzidi, Maria C Magnifico, et al. 2019. “The Moonlighting RNA-Binding Activity of Cytosolic Serine Hydroxymethyltransferase Contributes to Control Compartmentalization of Serine Metabolism.” *Nucleic Acids Research* 47 (8): 4240–54. https://doi.org/10.1093/nar/gkz129.

Lang, Benjamin, Alexandros Armaos, and Gian G. Tartaglia. 2019. “RNAct: Protein-RNA Interaction Predictions for Model Organisms with Supporting

Experimental Data." *Nucleic Acids Research* 47 (D1): D601–6. https://doi.org/10.1093/nar/gky967.

Ponti, Riccardo Delli, Alexandros Armaos, and Gian Gaetano Tartaglia. 2019. "*CROSSalive*: A Web Server for Predicting the *in Vivo* Structure of RNA Molecules." Preprint. Bioinformatics. https://doi.org/10.1101/626085.

Sanchez de Groot, Natalia, Alexandros Armaos, Ricardo Graña-Montes, Marion Alriquet, Giulia Calloni, R. Martin Vabulas, and Gian Gaetano Tartaglia. 2019. "RNA Structure Drives Interaction with Proteins." *Nature Communications* 10 (1): 3246. https://doi.org/10.1038/s41467-019-10923-5.

Vendramin, Roberto, Yvessa Verheyden, Hideaki Ishikawa, Lucas Goedert, Emilien Nicolas, Kritika Saraf, Alexandros Armaos, et al. 2018. "SAMMSON Fosters Cancer Cell Fitness by Concertedly Enhancing Mitochondrial and Cytosolic Translation." *Nature Structural & Molecular Biology* 25 (11): 1035–46. https://doi.org/10.1038/s41594-018-0143-4.

Bibliography

Agostini, F., Cirillo, D., Bolognesi, B., Tartaglia, G.G., 2013a. X-inactivation: quantitative predictions of protein interactions in the Xist network. Nucleic Acids Res. 41, e31. https://doi.org/10.1093/nar/gks968

Agostini, F., Cirillo, D., Ponti, R., Tartaglia, 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

Agostini, F., Zanzoni, A., Klus, P., Marchese, D., Cirillo, D., Tartaglia, G.G., 2013b. catRAPID omics: a web server for large-scale prediction of protein-RNA interactions. Bioinformatics 29, 2928–2930. https://doi.org/10.1093/bioinformatics/btt495

Alipanahi, B., Delong, A., Weirauch, M.T., Frey, B.J., 2015. Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. Nat. Biotechnol. 33, 831–838. https://doi.org/10.1038/nbt.3300

Altmeyer, M., Neelsen, K.J., Teloni, F., Pozdnyakova, I., Pellegrino, S., Grøfte, M., Rask, M.-B.D., Streicher, W., Jungmichel, S., Nielsen, M.L., Lukas, J., 2015. Liquid demixing of intrinsically disordered proteins is seeded by poly(ADP-ribose). Nat Commun 6, 8088. https://doi.org/10.1038/ncomms9088

Andersen, J.S., Lam, Y.W., Leung, A.K.L., Ong, S.-E., Lyon, C.E., Lamond, A.I., Mann, M., 2005. Nucleolar proteome dynamics. Nature 433, 77–83. https://doi.org/10.1038/nature03207

Armaos, A., Cirillo, D., Tartaglia, G.G., 2017. omiXcore: a web server for prediction of protein interactions with large RNA. Bioinformatics. https://doi.org/10.1093/bioinformatics/btx361

Audas, T.E., Audas, D.E., Jacob, M.D., Ho, J.J.D., Khacho, M., Wang, M., Perera, J.K., Gardiner, C., Bennett, C.A., Head, T., Kryvenko, O.N., Jorda, M., Daunert, S., Malhotra, A., Trinkle-Mulcahy, L., Gonzalgo, M.L., Lee, S., 2016. Adaptation to Stressors by Systemic Protein Amyloidogenesis. Dev. Cell 39, 155–168. https://doi.org/10.1016/j.devcel.2016.09.002

Bailey, T.L., Boden, M., Buske, F.A., Frith, M., Grant, C.E., Clementi, L., Ren, J., Li, W.W., Noble, W.S., 2009. MEME SUITE: tools for motif discovery and searching. Nucleic Acids Res. 37, W202-208. https://doi.org/10.1093/nar/gkp335

Baldassi, C., Zamparo, M., Feinauer, C., Procaccini, A., Zecchina, R., Weigt, M., Pagnani, A., 2014. Fast and accurate multivariate Gaussian modeling of protein families: predicting residue contacts and protein-interaction partners. PLoS ONE 9, e92721. https://doi.org/10.1371/journal.pone.0092721

Bellucci, M., Agostini, F., Masin, M., Tartaglia, G.G., 2011. Predicting protein associations with long noncoding RNAs. Nature Methods 8, 444–445. https://doi.org/10.1038/nmeth.1611

Benhalevy, D., Anastasakis, D.G., Hafner, M., 2018. Proximity-CLIP provides a snapshot of protein-occupied RNA elements in subcellular compartments. Nat. Methods 15, 1074–1082. https://doi.org/10.1038/s41592-018-0220-y

Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., Bourne, P.E., 2000. The Protein Data Bank. Nucleic Acids Res. 28, 235–242. https://doi.org/10.1093/nar/28.1.235

Bernhardt, H.S., 2012. The RNA world hypothesis: the worst theory of the early evolution of life (except for all the others)(a). Biol. Direct 7, 23. https://doi.org/10.1186/1745-6150-7-23

Bernstein, E., Allis, C.D., 2005. RNA meets chromatin. Genes Dev. 19, 1635–1655. https://doi.org/10.1101/gad.1324305

Boccaletto, P., Machnicka, M.A., Purta, E., Piatkowski, P., Baginski, B., Wirecki, T.K., de Crécy-Lagard, V., Ross, R., Limbach, P.A., Kotter, A., Helm, M., Bujnicki, J.M., 2018. MODOMICS: a database of RNA modification pathways. 2017 update. Nucleic Acids Res. 46, D303–D307. https://doi.org/10.1093/nar/gkx1030

Boeynaems, S., Alberti, S., Fawzi, N.L., Mittag, T., Polymenidou, M., Rousseau, F., Schymkowitz, J., Shorter, J., Wolozin, B., Van Den Bosch, L., Tompa, P., Fuxreiter, M., 2018. Protein Phase Separation: A New Phase in Cell Biology. Trends in Cell Biology 28, 420–435. https://doi.org/10.1016/j.tcb.2018.02.004

Boeynaems, S., Bogaert, E., Kovacs, D., Konijnenberg, A., Timmerman, E., Volkov, A., Guharoy, M., De Decker, M., Jaspers, T., Ryan, V.H., Janke, A.M., Baatsen, P., Vercruysse, T., Kolaitis, R.-M., Daelemans, D., Taylor, J.P., Kedersha, N., Anderson, P., Impens, F., Sobott, F., Schymkowitz, J., Rousseau, F., Fawzi, N.L., Robberecht, W., Van Damme, P., Tompa, P., Van Den Bosch, L., 2017. Phase Separation of C9orf72 Dipeptide Repeats Perturbs Stress Granule Dynamics. Mol. Cell 65, 1044-1055.e5. https://doi.org/10.1016/j.molcel.2017.02.013

Boke, E., Ruer, M., Wühr, M., Coughlin, M., Lemaitre, R., Gygi, S.P., Alberti, S., Drechsel, D., Hyman, A.A., Mitchison, T.J., 2016. Amyloid-like Self-Assembly of a Cellular Compartment. Cell 166, 637–650. https://doi.org/10.1016/j.cell.2016.06.051

Brangwynne, C.P., 2013. Phase transitions and size scaling of membrane-less organelles. J. Cell Biol. 203, 875–881. https://doi.org/10.1083/jcb.201308087

Brangwynne, C.P., Eckmann, C.R., Courson, D.S., Rybarska, A., Hoege, C., Gharakhani, J., Jülicher, F., Hyman, A.A., 2009. Germline P granules are liquid droplets that localize by controlled dissolution/condensation. Science 324, 1729–1732. https://doi.org/10.1126/science.1172046

Brangwynne, C.P., Mitchison, T.J., Hyman, A.A., 2011. Active liquid-like behavior of nucleoli determines their size and shape in Xenopus laevis oocytes. Proc. Natl. Acad. Sci. U.S.A. 108, 4334–4339. https://doi.org/10.1073/pnas.1017150108

Burke, K.A., Janke, A.M., Rhine, C.L., Fawzi, N.L., 2015. Residue-by-Residue View of In Vitro FUS Granules that Bind the C-Terminal Domain of RNA Polymerase II. Mol. Cell 60, 231–241. https://doi.org/10.1016/j.molcel.2015.09.006

Cao, C., Liu, F., Tan, H., Song, D., Shu, W., Li, W., Zhou, Y., Bo, X., Xie, Z., 2018. Deep Learning and Its Applications in Biomedicine. Genomics Proteomics Bioinformatics 16, 17–32. https://doi.org/10.1016/j.gpb.2017.07.003

Castello, A., Fischer, B., Frese, C.K., Horos, R., Alleaume, A.-M., Foehr, S., Curk, T., Krijgsveld, J., Hentze, M.W., 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

Castello, A., Fischer, B., Hentze, M.W., Preiss, T., 2013a. RNA-binding proteins in Mendelian disease. Trends Genet. 29, 318–327. https://doi.org/10.1016/j.tig.2013.01.004

Castello, A., Horos, R., Strein, C., Fischer, B., Eichelbaum, K., Steinmetz, L.M., Krijgsveld, J., Hentze, M.W., 2013b. System-wide identification of RNA-binding proteins by interactome capture. Nat Protoc 8, 491–500. https://doi.org/10.1038/nprot.2013.020

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, 331–334. https://doi.org/10.1038/s41594-019-0223-0

Chakrabarti, A.M., Haberman, N., Praznik, A., Luscombe, N.M., Ule, J., 2018. Data Science Issues in Studying Protein–RNA Interactions with CLIP Technologies. Annu. Rev. Biomed. Data Sci. 1, 235–261. https://doi.org/10.1146/annurev-biodatasci-080917-013525

Chu, C., Zhang, Q.C., da Rocha, S.T., Flynn, R.A., Bharadwaj, M., Calabrese, J.M., Magnuson, T., Heard, E., Chang, H.Y., 2015. Systematic discovery of Xist RNA binding proteins. Cell 161, 404–416. https://doi.org/10.1016/j.cell.2015.03.025

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., 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, 3422-3434.e7. https://doi.org/10.1016/j.celrep.2018.11.076

Cirillo, D., Agostini, F., Klus, P., Marchese, D., Rodriguez, S., Bolognesi, B., Tartaglia, G.G., 2013. Neurodegenerative diseases: quantitative predictions of protein-RNA interactions. RNA 19, 129–140. https://doi.org/10.1261/rna.034777.112

Cirillo, D., Blanco, M., Armaos, A., Buness, A., Avner, P., Guttman, M., Cerase, A., Tartaglia, G.G., 2016. Quantitative predictions of protein interactions with long noncoding RNAs. Nature Methods 14, 5–6. https://doi.org/10.1038/nmeth.4100

Clemson, C.M., Hutchinson, J.N., Sara, S.A., Ensminger, A.W., Fox, A.H., Chess, A., Lawrence, J.B., 2009. An architectural role for a nuclear noncoding RNA: NEAT1 RNA is essential for the structure of paraspeckles. Mol. Cell 33, 717–726. https://doi.org/10.1016/j.molcel.2009.01.026

Decker, C.J., Teixeira, D., Parker, R., 2007. Edc3p and a glutamine/asparagine-rich domain of Lsm4p function in processing body assembly in Saccharomyces cerevisiae. J. Cell Biol. 179, 437–449. https://doi.org/10.1083/jcb.200704147

Delaunay, S., Frye, M., 2019. RNA modifications regulating cell fate in cancer. Nat. Cell Biol. 21, 552–559. https://doi.org/10.1038/s41556-019-0319-0

Delli Ponti, R., Armaos, A., Marti, S., Tartaglia, G.G., 2018. A method for RNA structure prediction shows evidence for structure in lncRNAs. https://doi.org/10.1101/284869

Delli Ponti, R., Marti, S., Armaos, A., Tartaglia, G.G., 2017. A high-throughput approach to profile RNA structure. Nucleic Acids Research 45, e35–e35. https://doi.org/10.1093/nar/gkw1094

Ding, Y., Tang, Y., Kwok, C.K., Zhang, Y., Bevilacqua, P.C., Assmann, S.M., 2014. In vivo genome-wide profiling of RNA secondary structure reveals novel regulatory features. Nature 505, 696–700. https://doi.org/10.1038/nature12756

Dominguez, D., Freese, P., Alexis, M.S., Su, A., Hochman, M., Palden, T., Bazile, C., Lambert, N.J., Van Nostrand, E.L., Pratt, G.A., Yeo, G.W., Graveley, B.R., Burge, C.B., 2018. Sequence, Structure, and Context Preferences of Human RNA Binding Proteins. Mol. Cell 70, 854-867.e9. https://doi.org/10.1016/j.molcel.2018.05.001

Doty, P., Boedtker, H., Fresco, J.R., Haselkorn, R., Litt, M., 1959. SECONDARY STRUCTURE IN RIBONUCLEIC ACIDS. Proc. Natl. Acad. Sci. U.S.A. 45, 482–499. https://doi.org/10.1073/pnas.45.4.482

Dunker, A.K., Bondos, S.E., Huang, F., Oldfield, C.J., 2015. Intrinsically disordered proteins and multicellular organisms. Semin. Cell Dev. Biol. 37, 44–55. https://doi.org/10.1016/j.semcdb.2014.09.025

Elbaum-Garfinkle, S., Kim, Y., Szczepaniak, K., Chen, C.C.-H., Eckmann, C.R., Myong, S., Brangwynne, C.P., 2015. The disordered P granule protein LAF-1 drives phase separation into droplets with tunable viscosity and dynamics. Proc. Natl. Acad. Sci. U.S.A. 112, 7189–7194. https://doi.org/10.1073/pnas.1504822112

Falahati, H., Pelham-Webb, B., Blythe, S., Wieschaus, E., 2016. Nucleation by rRNA Dictates the Precision of Nucleolus Assembly. Curr. Biol. 26, 277–285. https://doi.org/10.1016/j.cub.2015.11.065

Faraggi, E., Zhou, Y., Kloczkowski, A., 2014. Accurate single-sequence prediction of solvent accessible surface area using local and global features. Proteins 82, 3170–3176. https://doi.org/10.1002/prot.24682

Feric, M., Vaidya, N., Harmon, T.S., Mitrea, D.M., Zhu, L., Richardson, T.M., Kriwacki, R.W., Pappu, R.V., Brangwynne, C.P., 2016. Coexisting Liquid Phases Underlie Nucleolar Subcompartments. Cell 165, 1686–1697. https://doi.org/10.1016/j.cell.2016.04.047

Fernandez, M., Kumagai, Y., Standley, D.M., Sarai, A., Mizuguchi, K., Ahmad, S., 2011. Prediction of dinucleotide-specific RNA-binding sites in proteins. BMC Bioinformatics 12 Suppl 13, S5. https://doi.org/10.1186/1471-2105-12-S13-S5

Finn, R.D., Clements, J., Eddy, S.R., 2011. HMMER web server: interactive sequence similarity searching. Nucleic Acids Research 39, W29–W37. https://doi.org/10.1093/nar/gkr367

Fong, K.-W., Li, Y., Wang, W., Ma, W., Li, K., Qi, R.Z., Liu, D., Songyang, Z., Chen, J., 2013. Whole-genome screening identifies proteins localized to distinct nuclear bodies. J. Cell Biol. 203, 149–164. https://doi.org/10.1083/jcb.201303145

Ganassi, M., Mateju, D., Bigi, I., Mediani, L., Poser, I., Lee, H.O., Seguin, S.J., Morelli, F.F., Vinet, J., Leo, G., Pansarasa, O., Cereda, C., Poletti, A., Alberti, S., Carra, S., 2016. A Surveillance Function of the HSPB8-BAG3-HSP70 Chaperone Complex Ensures Stress Granule Integrity and Dynamism. Mol. Cell 63, 796–810. https://doi.org/10.1016/j.molcel.2016.07.021

Gilks, N., Kedersha, N., Ayodele, M., Shen, L., Stoecklin, G., Dember, L.M., Anderson, P., 2004. Stress granule assembly is mediated by prion-like aggregation of TIA-1. Mol. Biol. Cell 15, 5383–5398. https://doi.org/10.1091/mbc.e04-08-0715

Goodarzi, H., Najafabadi, H.S., Oikonomou, P., Greco, T.M., Fish, L., Salavati, R., Cristea, I.M., Tavazoie, S., 2012. Systematic discovery of structural elements governing stability of mammalian messenger RNAs. Nature 485, 264–268. https://doi.org/10.1038/nature11013

Govindan, G., Nair, A.S., 2011. Composition, Transition and Distribution (CTD) &#x2014; A dynamic feature for predictions based on hierarchical structure of cellular sorting, in: 2011 Annual IEEE India Conference. Presented at the 2011 Annual IEEE India Conference (INDICON), IEEE, Hyderabad, India, pp. 1–6. https://doi.org/10.1109/INDCON.2011.6139332

Gruber, A.R., Bernhart, S.H., Lorenz, R., 2015. The ViennaRNA web services. Methods Mol. Biol. 1269, 307–326. https://doi.org/10.1007/978-1-4939-2291-8_19

HafezQorani, S., Lafzi, A., de Bruin, R.G., van Zonneveld, A.J., van der Veer, E.P., Son, Y.A., Kazan, H., 2016. Modeling the combined effect of RNA-binding proteins and microRNAs in post-transcriptional regulation. Nucleic Acids Res 44, e83–e83. https://doi.org/10.1093/nar/gkw048

Hajiaghayi, M., Condon, A., Hoos, H.H., 2012. Analysis of energy-based algorithms for RNA secondary structure prediction. BMC Bioinformatics 13, 22. https://doi.org/10.1186/1471-2105-13-22

Hentze, M.W., Castello, A., Schwarzl, T., 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

Hiller, M., Zhang, Z., Backofen, R., Stamm, S., 2007. Pre-mRNA secondary structures influence exon recognition. PLoS Genet. 3, e204. https://doi.org/10.1371/journal.pgen.0030204

Hilleren, P., McCarthy, T., Rosbash, M., Parker, R., Jensen, T.H., 2001. Quality control of mRNA 3'-end processing is linked to the nuclear exosome. Nature 413, 538–542. https://doi.org/10.1038/35097110

Iyer, M.K., Niknafs, Y.S., Malik, R., Singhal, U., Sahu, A., Hosono, Y., Barrette, T.R., Prensner, J.R., Evans, J.R., Zhao, S., Poliakov, A., Cao, X., Dhanasekaran, S.M., Wu, Y.-M., Robinson, D.R., Beer, D.G., Feng, F.Y., Iyer, H.K., Chinnaiyan, A.M., 2015. The landscape of long noncoding RNAs in the human transcriptome. Nat. Genet. 47, 199–208. https://doi.org/10.1038/ng.3192

Jain, A., Vale, R.D., 2017. RNA phase transitions in repeat expansion disorders. Nature 546, 243–247. https://doi.org/10.1038/nature22386

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

Jankowsky, E., 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

Jeong, H., Mason, S.P., Barabási, A.L., Oltvai, Z.N., 2001. Lethality and centrality in protein networks. Nature 411, 41–42. https://doi.org/10.1038/35075138

Jones, D.T., Buchan, D.W.A., Cozzetto, D., Pontil, M., 2012. PSICOV: precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments. Bioinformatics 28, 184–190. https://doi.org/10.1093/bioinformatics/btr638

Keren, H., Lev-Maor, G., Ast, G., 2010. Alternative splicing and evolution: diversification, exon definition and function. Nat. Rev. Genet. 11, 345–355. https://doi.org/10.1038/nrg2776

Kishore, S., Jaskiewicz, L., Burger, L., Hausser, J., Khorshid, M., Zavolan, M., 2011. A quantitative analysis of CLIP methods for identifying binding sites of RNA-binding proteins. Nat. Methods 8, 559–564. https://doi.org/10.1038/nmeth.1608

König, J., Zarnack, K., Luscombe, N.M., Ule, J., 2012. Protein-RNA interactions: new genomic technologies and perspectives. Nat. Rev. Genet. 13, 77–83. https://doi.org/10.1038/nrg3141

Konig, J., Zarnack, K., Rot, G., Curk, T., Kayikci, M., Zupan, B., Turner, D.J., Luscombe, N.M., Ule, J., 2011. iCLIP--transcriptome-wide mapping of protein-RNA interactions with individual nucleotide resolution. J Vis Exp. https://doi.org/10.3791/2638

Kumar, M., Gromiha, M.M., Raghava, G.P.S., 2011. SVM based prediction of RNA-binding proteins using binding residues and evolutionary information. J. Mol. Recognit. 24, 303–313. https://doi.org/10.1002/jmr.1061

Kumar, M., Gromiha, M.M., Raghava, G.P.S., 2008. Prediction of RNA binding sites in a protein using SVM and PSSM profile. Proteins 71, 189–194. https://doi.org/10.1002/prot.21677

Kyburz, A., Friedlein, A., Langen, H., Keller, W., 2006. Direct interactions between subunits of CPSF and the U2 snRNP contribute to the coupling of pre-mRNA 3' end processing and splicing. Mol. Cell 23, 195–205. https://doi.org/10.1016/j.molcel.2006.05.037

Langdon, E.M., Qiu, Y., Ghanbari Niaki, A., McLaughlin, G.A., Weidmann, C.A., Gerbich, T.M., Smith, J.A., Crutchley, J.M., Termini, C.M., Weeks, K.M., Myong, S., Gladfelter, A.S., 2018. mRNA structure determines specificity of a polyQ-driven phase separation. Science 360, 922–927. https://doi.org/10.1126/science.aar7432

Lange, S.J., Maticzka, D., Möhl, M., Gagnon, J.N., Brown, C.M., Backofen, R., 2012. Global or local? Predicting secondary structure and accessibility in mRNAs. Nucleic Acids Research 40, 5215–5226. https://doi.org/10.1093/nar/gks181

Li, X., Quon, G., Lipshitz, H.D., Morris, Q., 2010. Predicting in vivo binding sites of RNA-binding proteins using mRNA secondary structure. RNA 16, 1096–1107. https://doi.org/10.1261/rna.2017210

Lin, Y., Protter, D.S.W., Rosen, M.K., Parker, R., 2015. Formation and Maturation of Phase-Separated Liquid Droplets by RNA-Binding Proteins. Mol. Cell 60, 208–219. https://doi.org/10.1016/j.molcel.2015.08.018

Livi, C.M., Klus, P., Delli Ponti, R., Tartaglia, G.G., 2016. catRAPID signature: identification of ribonucleoproteins and RNA-binding regions. Bioinformatics 32, 773–775. https://doi.org/10.1093/bioinformatics/btv629

Lorenz, R., Wolfinger, M.T., Tanzer, A., Hofacker, I.L., 2016. Predicting RNA secondary structures from sequence and probing data. Methods 103, 86–98. https://doi.org/10.1016/j.ymeth.2016.04.004

Lunde, B.M., Moore, C., Varani, G., 2007. RNA-binding proteins: modular design for efficient function. Nat. Rev. Mol. Cell Biol. 8, 479–490. https://doi.org/10.1038/nrm2178

Macias, S., Plass, M., Stajuda, A., Michlewski, G., Eyras, E., Cáceres, J.F., 2012. DGCR8 HITS-CLIP reveals novel functions for the Microprocessor. Nat. Struct. Mol. Biol. 19, 760–766. https://doi.org/10.1038/nsmb.2344

Maharana, S., Wang, J., Papadopoulos, D.K., Richter, D., Pozniakovsky, A., Poser, I., Bickle, M., Rizk, S., Guillén-Boixet, J., Franzmann, T., Jahnel, M., Marrone, L., Chang, Y.-T., Sterneckert, J., Tomancak, P., Hyman, A.A., Alberti, S., 2018. RNA buffers the phase separation behavior of prion-like RNA binding proteins. Science. https://doi.org/10.1126/science.aar7366

Marchese, D., Botta-Orfila, T., Cirillo, D., Rodriguez, J.A., Livi, C.M., Fernández-Santiago, R., Ezquerra, M., Martí, M.J., Bechara, E., Tartaglia, G.G., Catalan MSA Registry (CMSAR), 2017. Discovering the 3’ UTR-mediated regulation of alpha-synuclein. Nucleic Acids Res. 45, 12888–12903. https://doi.org/10.1093/nar/gkx1048

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, 793–810. https://doi.org/10.1002/wrna.1378

Mateju, D., Franzmann, T.M., Patel, A., Kopach, A., Boczek, E.E., Maharana, S., Lee, H.O., Carra, S., Hyman, A.A., Alberti, S., 2017. An aberrant phase transition of stress granules triggered by misfolded protein and prevented by chaperone function. EMBO J. 36, 1669–1687. https://doi.org/10.15252/embj.201695957

McHugh, C.A., Russell, P., Guttman, M., 2014. Methods for comprehensive experimental identification of RNA-protein interactions. Genome Biol. 15, 203. https://doi.org/10.1186/gb4152

McMahon, A.C., Rahman, R., Jin, H., Shen, J.L., Fieldsend, A., Luo, W., Rosbash, M., 2016. TRIBE: Hijacking an RNA-Editing Enzyme to Identify Cell-Specific Targets of RNA-Binding Proteins. Cell 165, 742–753. https://doi.org/10.1016/j.cell.2016.03.007

Miao, Z., Westhof, E., 2017. RNA Structure: Advances and Assessment of 3D Structure Prediction. Annu Rev Biophys 46, 483–503. https://doi.org/10.1146/annurev-biophys-070816-034125

Millevoi, S., Loulergue, C., Dettwiler, S., Karaa, S.Z., Keller, W., Antoniou, M., Vagner, S., 2006. An interaction between U2AF 65 and CF I(m) links the splicing and 3' end processing machineries. EMBO J. 25, 4854–4864. https://doi.org/10.1038/sj.emboj.7601331

Minajigi, A., Froberg, J.E., Wei, C., Sunwoo, H., Kesner, B., Colognori, D., Lessing, D., Payer, B., Boukhali, M., Haas, W., Lee, J.T., 2015. A comprehensive Xist interactome reveals cohesin repulsion and an RNA-directed chromosome conformation. Science 349, aab2276–aab2276. https://doi.org/10.1126/science.aab2276

Mitchell, S.F., Parker, R., 2014. Principles and properties of eukaryotic mRNPs. Mol. Cell 54, 547–558. https://doi.org/10.1016/j.molcel.2014.04.033

Moindrot, B., Cerase, A., Coker, H., Masui, O., Grijzenhout, A., Pintacuda, G., Schermelleh, L., Nesterova, T.B., Brockdorff, N., 2015. A Pooled shRNA Screen Identifies Rbm15, Spen, and Wtap as Factors Required for Xist RNA-Mediated Silencing. Cell Reports 12, 562–572. https://doi.org/10.1016/j.celrep.2015.06.053

Molliex, A., Temirov, J., Lee, J., Coughlin, M., Kanagaraj, A.P., Kim, H.J., Mittag, T., Taylor, J.P., 2015. Phase separation by low complexity domains promotes stress granule assembly and drives pathological fibrillization. Cell 163, 123–133. https://doi.org/10.1016/j.cell.2015.09.015

Monfort, A., Di Minin, G., Postlmayr, A., Freimann, R., Arieti, F., Thore, S., Wutz, A., 2015. Identification of Spen as a Crucial Factor for Xist Function through Forward Genetic Screening in Haploid Embryonic Stem Cells. Cell Reports 12, 554–561. https://doi.org/10.1016/j.celrep.2015.06.067

Muppirala, U.K., Honavar, V.G., Dobbs, D., 2011. Predicting RNA-protein interactions using only sequence information. BMC Bioinformatics 12, 489. https://doi.org/10.1186/1471-2105-12-489

Nott, T.J., Petsalaki, E., Farber, P., Jervis, D., Fussner, E., Plochowietz, A., Craggs, T.D., Bazett-Jones, D.P., Pawson, T., Forman-Kay, J.D., Baldwin, A.J., 2015. Phase transition of a disordered nuage protein generates environmentally responsive membraneless organelles. Mol. Cell 57, 936–947. https://doi.org/10.1016/j.molcel.2015.01.013

Parker, R., Sheth, U., 2007. P bodies and the control of mRNA translation and degradation. Mol. Cell 25, 635–646. https://doi.org/10.1016/j.molcel.2007.02.011

Pintacuda, G., Young, A.N., Cerase, A., 2017. Function by Structure: Spotlights on Xist Long Non-coding RNA. Front Mol Biosci 4, 90. https://doi.org/10.3389/fmolb.2017.00090

Qamar, S., Wang, G., Randle, S.J., Ruggeri, F.S., Varela, J.A., Lin, J.Q., Phillips, E.C., Miyashita, A., Williams, D., Ströhl, F., Meadows, W., Ferry, R., Dardov, V.J., Tartaglia, G.G., Farrer, L.A., Kaminski Schierle, G.S., Kaminski, C.F., Holt, C.E., Fraser, P.E., Schmitt-Ulms, G., Klenerman, D., Knowles, T., Vendruscolo, M., St George-Hyslop, P., 2018. FUS Phase Separation Is Modulated by a Molecular Chaperone and Methylation of Arginine Cation-π Interactions. Cell 173, 720-734.e15. https://doi.org/10.1016/j.cell.2018.03.056

Ramaswami, M., Taylor, J.P., Parker, R., 2013. Altered ribostasis: RNA-protein granules in degenerative disorders. Cell 154, 727–736. https://doi.org/10.1016/j.cell.2013.07.038

Ray, D., Ha, K.C.H., Nie, K., Zheng, H., Hughes, T.R., 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

Reijns, M.A.M., Alexander, R.D., Spiller, M.P., Beggs, J.D., 2008. A role for Q/N-rich aggregation-prone regions in P-body localization. J. Cell. Sci. 121, 2463–2472. https://doi.org/10.1242/jcs.024976

Reuter, J.S., Mathews, D.H., 2010. RNAstructure: software for RNA secondary structure prediction and analysis. BMC Bioinformatics 11, 129. https://doi.org/10.1186/1471-2105-11-129

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, 424–428. https://doi.org/10.1038/s41586-019-1374-1

Rigo, F., Martinson, H.G., 2008. Functional coupling of last-intron splicing and 3'-end processing to transcription in vitro: the poly(A) signal couples to splicing before committing to cleavage. Mol. Cell. Biol. 28, 849–862. https://doi.org/10.1128/MCB.01410-07

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, 701–705. https://doi.org/10.1038/nature12894

Rybak-Wolf, A., Jens, M., Murakawa, Y., Herzog, M., Landthaler, M., Rajewsky, N., 2014. A variety of dicer substrates in human and C. elegans. Cell 159, 1153–1167. https://doi.org/10.1016/j.cell.2014.10.040

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

Seemann, S.E., Mirza, A.H., Hansen, C., Bang-Berthelsen, C.H., Garde, C., Christensen-Dalsgaard, M., Torarinsson, E., Yao, Z., Workman, C.T., Pociot, F., Nielsen, H., Tommerup, N., Ruzzo, W.L., Gorodkin, J., 2017. The identification and functional annotation of RNA structures conserved in vertebrates. Genome Res. 27, 1371–1383. https://doi.org/10.1101/gr.208652.116

Shin, Y., Brangwynne, C.P., 2017. Liquid phase condensation in cell physiology and disease. Science 357. https://doi.org/10.1126/science.aaf4382

Singh, R., Valcárcel, J., 2005. Building specificity with nonspecific RNA-binding proteins. Nat. Struct. Mol. Biol. 12, 645–653. https://doi.org/10.1038/nsmb961

Smith, K.C., Aplin, R.T., 1966. A mixed photoproduct of uracil and cysteine (5-S-cysteine-6-hydrouracil). A possible model for the in vivo cross-linking of deoxyribonucleic acid and protein by ultraviolet light. Biochemistry 5, 2125–2130.

Spitzer, J., Hafner, M., Landthaler, M., Ascano, M., Farazi, T., Wardle, G., Nusbaum, J., Khorshid, M., Burger, L., Zavolan, M., Tuschl, T., 2014. PAR-CLIP (Photoactivatable Ribonucleoside-Enhanced Crosslinking and Immunoprecipitation): a step-by-step protocol to the transcriptome-wide identification of binding sites of RNA-binding proteins. Meth. Enzymol. 539, 113–161. https://doi.org/10.1016/B978-0-12-420120-0.00008-6

Taft, R.J., Pheasant, M., Mattick, J.S., 2007. The relationship between non-protein-coding DNA and eukaryotic complexity. Bioessays 29, 288–299. https://doi.org/10.1002/bies.20544

Taliaferro, J.M., Lambert, N.J., Sudmant, P.H., Dominguez, D., Merkin, J.J., Alexis, M.S., Bazile, C., Burge, C.B., 2016. RNA Sequence Context Effects Measured In Vitro Predict In Vivo Protein Binding and Regulation. Mol. Cell 64, 294–306. https://doi.org/10.1016/j.molcel.2016.08.035

Teixeira, D., Parker, R., 2007. Analysis of P-body assembly in Saccharomyces cerevisiae. Mol. Biol. Cell 18, 2274–2287. https://doi.org/10.1091/mbc.e07-03-0199

Teplova, M., Malinina, L., Darnell, J.C., Song, J., Lu, M., Abagyan, R., Musunuru, K., Teplov, A., Burley, S.K., Darnell, R.B., Patel, D.J., 2011. Protein-RNA and protein-protein recognition by dual KH1/2 domains of the neuronal splicing factor Nova-1. Structure 19, 930–944. https://doi.org/10.1016/j.str.2011.05.002

Thisted, T., Lyakhov, D.L., Liebhaber, S.A., 2001. Optimized RNA targets of two closely related triple KH domain proteins, heterogeneous nuclear ribonucleoprotein K and alphaCP-2KL, suggest Distinct modes of RNA recognition. J. Biol. Chem. 276, 17484–17496. https://doi.org/10.1074/jbc.M010594200

Toretsky, J.A., Wright, P.E., 2014. Assemblages: functional units formed by cellular phase separation. J. Cell Biol. 206, 579–588. https://doi.org/10.1083/jcb.201404124

Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D.R., Pimentel, H., Salzberg, S.L., Rinn, J.L., Pachter, L., 2012. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nature Protocols 7, 562–578. https://doi.org/10.1038/nprot.2012.016

van der Lee, R., Lang, B., Kruse, K., Gsponer, J., Sánchez de Groot, N., Huynen, M.A., Matouschek, A., Fuxreiter, M., Babu, M.M., 2014. Intrinsically disordered segments affect protein half-life in the cell and during evolution. Cell Rep 8, 1832–1844. https://doi.org/10.1016/j.celrep.2014.07.055

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). Nature Methods 13, 508–514. https://doi.org/10.1038/nmeth.3810

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. Proc. Natl. Acad. Sci. U.S.A. 115, 2734–2739. https://doi.org/10.1073/pnas.1800038115

Vendramin, R., Verheyden, Y., Ishikawa, H., Goedert, L., Nicolas, E., Saraf, K., Armaos, A., Delli Ponti, R., Izumikawa, K., Mestdagh, P., Lafontaine, D.L.J., Tartaglia, G.G., Takahashi, N., Marine, J.-C., Leucci, E., 2018. SAMMSON fosters cancer cell fitness by concertedly enhancing mitochondrial and cytosolic translation. Nat. Struct. Mol. Biol. 25, 1035–1046. https://doi.org/10.1038/s41594-018-0143-4

Walia, R.R., Xue, L.C., Wilkins, K., El-Manzalawy, Y., Dobbs, D., Honavar, V., 2014. RNABindRPlus: a predictor that combines machine learning and sequence homology-based methods to improve the reliability of predicted RNA-binding residues in proteins. PLoS ONE 9, e97725. https://doi.org/10.1371/journal.pone.0097725

Wang, K.C., Chang, H.Y., 2011. Molecular mechanisms of long noncoding RNAs. Mol. Cell 43, 904–914. https://doi.org/10.1016/j.molcel.2011.08.018

Wang, L., Huang, C., Yang, M.Q., Yang, J.Y., 2010. BindN+ for accurate prediction of DNA and RNA-binding residues from protein sequence features. BMC Syst Biol 4, S3. https://doi.org/10.1186/1752-0509-4-S1-S3

Warf, M.B., Diegel, J.V., von Hippel, P.H., Berglund, J.A., 2009. The protein factors MBNL1 and U2AF65 bind alternative RNA structures to regulate splicing. Proc. Natl. Acad. Sci. U.S.A. 106, 9203–9208. https://doi.org/10.1073/pnas.0900342106

Weber, S.C., Brangwynne, C.P., 2012. Getting RNA and protein in phase. Cell 149, 1188–1191. https://doi.org/10.1016/j.cell.2012.05.022

Wheeler, J.R., Matheny, T., Jain, S., Abrisch, R., Parker, R., 2016. Distinct stages in stress granule assembly and disassembly. Elife 5. https://doi.org/10.7554/eLife.18413

Wolozin, B., 2012. Regulated protein aggregation: stress granules and neurodegeneration. Mol Neurodegener 7, 56. https://doi.org/10.1186/1750-1326-7-56

Yamazaki, T., Souquere, S., Chujo, T., Kobelke, S., Chong, Y.S., Fox, A.H., Bond, C.S., Nakagawa, S., Pierron, G., Hirose, T., 2018. Functional Domains of NEAT1 Architectural lncRNA Induce Paraspeckle Assembly through Phase Separation. Mol. Cell 70, 1038-1053.e7. https://doi.org/10.1016/j.molcel.2018.05.019

Yao, R.-W., Wang, Y., Chen, L.-L., 2019. Cellular functions of long noncoding RNAs. Nat. Cell Biol. 21, 542–551. https://doi.org/10.1038/s41556-019-0311-8

Zanzoni, A., Marchese, D., Agostini, F., Bolognesi, B., Cirillo, D., Botta-Orfila, M., Livi, C.M., Rodriguez-Mulero, S., Tartaglia, G.G., 2013. Principles of self-organization in biological pathways: a hypothesis on the autogenous association of alpha-synuclein. Nucleic Acids Res. 41, 9987–9998. https://doi.org/10.1093/nar/gkt794

Zhang, C., 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

Zhang, H., Elbaum-Garfinkle, S., Langdon, E.M., Taylor, N., Occhipinti, P., Bridges, A.A., Brangwynne, C.P., Gladfelter, A.S., 2015. RNA Controls PolyQ Protein Phase Transitions. Mol. Cell 60, 220–230. https://doi.org/10.1016/j.molcel.2015.09.017

Zhang, S., Zhou, J., Hu, H., Gong, H., Chen, L., Cheng, C., Zeng, J., 2016. A deep learning framework for modeling structural features of RNA-binding protein targets. Nucleic Acids Res. 44, e32. https://doi.org/10.1093/nar/gkv1025