Discovery of the role of protein-RNA interactions in protein multifunctionality and cellular complexity

Diogo Ribeiro

Aix-Marseille Université

Faculté des Sciences de Luminy

École Doctorale des sciences de la vie et de la santé

Theories and Approaches of Genomic Complexity (TAGC), INSERM U1090

Thèse de Doctorat d'Aix-Marseille Université

Spécialité: Bioinformatique et Génomique

Découverte du rôle des interactions protéine-ARN dans la multifonctionnalité des protéines et la complexité cellulaire

Présenté par:

Diogo DE ABREU MARQUES RIBEIRO

Soutenue le 5 décembre 2018 devant le jury composé de:

Pr. Jacques VAN HELDEN	Président du jury
Dr. Didier AUBOEUF	Rapporteur
Pr. Ulrich STELZL	Rapporteur
Dr. Anne-Marie FRANÇOIS-BELLAN	Examinatrice
Dr. Gian GAETANO TARTAGLIA	Directeur de thèse
Dr. Christine BRUN	Directrice de thèse

Aix-Marseille University

Faculty of Science of Luminy

Doctoral School of Life and Health Sciences

Theories and Approaches of Genomic Complexity (TAGC), INSERM U1090

PhD Thesis in Bioinformatics and Genomics

Discovery of the role of protein-RNA interactions in protein multifunctionality and cellular complexity

---

Presented by:

Diogo DE ABREU MARQUES RIBEIRO

Defended on the 5th of December 2018 in front of the jury members:

Pr. Jacques VAN HELDEN	President
Dr. Didier AUBOEUF	Reporter
Pr. Ulrich Stelzl	Reporter
Dr. Anne-Marie FRANÇOIS-BELLAN	Examinator
Dr. Gian GAETANO TARTAGLIA	Supervisor
Dr. Christine BRUN	Supervisor

Résumé

Au fil du temps, la vie a évolué pour produire des organismes remarquablement complexes. Pour faire face à cette complexité, les organismes ont développé une pléthore de mécanismes régulateurs. Par exemple, pour chaque ARN messager (ARNm) codant une protéine, des régions non traduites (UTR; *untranslated regions* en anglais) potentiellement régulatrices sont aussi présentes. De plus, les organismes supérieurs transcrivent des milliers d'ARN longs non codants (ARNlnc), accroissant ainsi la capacité régulatrice de leurs cellules. Cependant, la plupart des ARNlnc sont-ils fonctionnels? Le cas échéant, par quels mécanismes peuvent-ils agir? Le rôle d'échafaudage des ARNlnc, formant des ribonucléoprotéines et rapprochant ainsi physiquement les protéines est un concept émergent. Toutefois, la prévalence de ce mécanisme reste encore à déterminer.

De plus, au lieu d'ajouter de nouveaux composants pour augmenter la complexité, les cellules peuvent réutiliser certaines protéines pour exécuter plusieurs fonctions distinctes. C'est le cas des protéines *moonlighting*. Ces protéines exercent souvent des fonctions distinctes dans des environnements différents et peuvent donc être régulées par un changement de localisation cellulaire. Par la formation de complexes protéiques en cours de traduction, les régions 3' non traduites (3'UTRs) peuvent réguler la localisation cellulaire et la fonction de la protéine synthétisée à partir des transcrits auxquels elles appartiennent. Néanmoins, la fréquence ce mécanisme et son rôle dans la régulation des diverses fonctions des protéines *moonlighting* reste à aborder.

Cette thèse a pour objectif de découvrir et comprendre systématiquement deux mécanismes de régulation méconnus impliquant la partie non codante du transcriptome humain. Concrètement, l'assemblage de complexes protéiques promus par les ARNlnc et les 3'UTRs est étudié avec des données d'interactions protéines-protéines et protéines-ARN prédites et expérimentales, à grande échelle. Ceci a permis *(i)* de prédire le rôle de plusieurs centaines d'ARNlnc comme molécules d'échafaudage pour plus de la moitié des complexes protéiques connus, ainsi que *(ii)* d'inférer plus d'un millier de complexes 3'UTR-protéines, dont des cas permettant d'expliquer la localisation cellulaire de protéines *moonlighting*. Ces résultats obtenus à l'échelle du protéome et du transcriptome indiquent qu'une proportion élevée d'ARNlnc et de 3'UTRs pourrait réguler la fonction des protéines en augmentant ainsi la complexité du vivant.

Abstract

Over time, life has evolved to produce remarkably complex organisms. To cope with this complexity, organisms have evolved a plethora of regulatory mechanisms. For instance, for every messenger RNA (mRNA) encoding a protein, regulatory untranslated regions (UTRs) are also present. Additionally, higher organisms transcribe thousands of long non-coding RNAs (lncRNAs), presumably expanding the regulatory capacity of their cells. However, it is questionable whether most lncRNAs are functional, and even though many lncRNAs interact with other cellular components, is yet unclear through which mechanisms they may act. An emerging concept is that lncRNAs can serve as protein scaffolds, forming ribonucleoproteins and bringing proteins in proximity, but the prevalence of this mechanism is yet to be determined.

Besides adding new components to increase complexity, cells can reuse proteins to perform several unrelated functions. Such is the case of the moonlighting proteins. These proteins are often found to perform distinct functions under different environments, and may thus be regulated by a change of cellular localisation. Interestingly, through the formation of protein-complexes during translation, 3'UTRs have been found to regulate the cellular localisation and function of the protein synthesized from their transcript. Yet, if this mechanism is common, and if used to regulate the several functions of moonlighting proteins, remains to be addressed.

This thesis aims to systematically discover and provide insights into two ill-known regulatory mechanisms involving the non-coding portion of the human transcriptome. Concretely, the assembly of protein complexes promoted by lncRNAs and 3'UTRs is investigated using computationally predicted, as well as experimentally determined, large-scale datasets of protein-protein and protein-RNA interactions. This enabled to *(i)* predict hundreds of lncRNAs as possible scaffolding molecules for more than half of the known protein complexes, as well as *(ii)* infer more than a thousand distinct 3'UTR-protein complexes, including cases likely to regulate the cellular localisation of moonlighting proteins. These large-scale results indicate that a high proportion of lncRNAs and 3'UTRs may be employed in regulating protein function, potentially playing a role both as regulators and as components of complexity.

Acknowledgements

Doing a PhD is like taking a long trip, and like any trip, it goes better if you have the right companions, and a good guidance system. This trip has led me where I wanted to go – even if sometimes passing through unexpected but didactic detours – and for this I have to thank all the people that aided me during this endeavour.

First and foremost, I would like to thank my supervisor **Christine Brun** for her continuous support and dedication during my PhD studies. I have learned much about how to do good science with you, and I have always been amazed by your ability to come up with interesting research projects. You gave me a lot and I feel blessed to have had a supervisor who was always there for me and who cared about my professional and personal well-being.

I would like to thank **Gian Gaetano Tartaglia** for all the interesting scientific discussions and ideas, as well as for his right decision to keep believing and digging even when research results deviate from what is expected.

I am very grateful to have had brilliant lab colleagues over the past three years, from whom I have learned a lot. I would like to thank **Lionel Spinelli** for being tough on me when needed, for all his time spent teaching me statistics, programming and how to do science properly. You taught me many valuable lessons which I hope I will always remember. Likewise, I thank my colleague **Andreas Zanzoni** for all the good insights into my projects and for the careful corrections of articles and abstracts. Other colleagues who have come and gone, have helped me grow as a scientist and as a mentor. These include **Elisa Micarelli, Galadriel Briere, Zacharie Menetrier, Adrien Teixeira** and **Paul de Boissier**. Scientific progress requires scientific discussions and with this in mind I would like to thank **Davide Cirillo**, **Philippe Pierre**, as well as the members of my thesis committee **Thien Vu Manh, Nicolas Terrapon** and **Salvatore Spicuglia**.

Over the past three years I was lucky enough to be immersed in a warm work environment, enjoy innumerable beer sessions and make good friends. Therefore, I would like to thank current and past members of the TAGC, especially the Vietnamese crew including **Lan, Minh** and **Khanh** who introduced

me to a whole new culture and literally took me to Vietnam. The more familiar Spanish-speaking crew including **Jaime** (*el bajista del grupo*), **Santiago, José David** and **Claire**. My special thanks to **Lucie**, the glue and defender of all the TAGC students, as well as to **Marie, Florian, Jeanne, Michel** and **Eve-Lyne**, for all the help with understanding France and its bureaucracy, and for making me laugh countless times. **Benoit Ballester**, **Laurence Roder** and **Myriam Ramadour** were also always ready to help desperate PhD students, and I thank everyone at TAGC for making me feel at home.

I would like to acknowledge the funding from A*MIDEX, Aix-Marseille University, and the hosting of TAGC, Inserm U1090 which enabled me to complete my PhD. Moreover, I would like to thank the Bioinformatics pedagogic unit and all my bachelor students who made me laugh and grow as a person.

Besides working on my thesis, perhaps what I will remember the most from this period of my life will be many adventures with friends around and across all the mountains and beaches of Provence. Some of my best memories will surely come from the time spent with the *La République coloc team*, composed of **Alberto** (el tronco), always up for an adventure or a relaxing beer, **Annamaria**, with her unstoppable can-do attitude, and **Guillaume**, continuously improving our French. I will always remember the overcrowded parties in our fancy flat.

Being part of the *Café des Langues* and the international student community of Marseille was an enormous pleasure, and I thank CIELL's **Tom Grainger** for gluing this amazing team together. There I have made unforgettable friends including **Lamia, Alejandro, Elena, Serena, Afroditi**, **Lolita, Andrea** and **Alessandro**. I will always think fondly of each and every one of you and the good times we shared, including travels in France and abroad and - for many of you - my wedding in Poland. Indeed, I would like to express my gratitude to all the friends who provided me with much-needed breaks from my thesis and took me trekking, sailing, kayaking, and put me to play ping pong, tennis, football and volleyball.

Agradeço muito os meus pais e a minha irmã por me criarem da maneira que me criaram e por me permitirem seguir e concretizar os meus sonhos. Vocês foram sempre o meu pilar e porto seguro nesta aventura. Obrigado pela força e motivação que me deram.

Lastly, my deepest thanks to you **Marta**, for all the love and care that you provided me, as well as for keeping me well fed and incessantly correcting my English. *Dziękuję bardzo, kochanie*!

Diogo Ribeiro
Marseille, September 2018

Preface

Having a background in biology, I have always been amazed by how life – cells, organs, organisms, populations – is able to attain a level of complexity and functionality that may well be unmatched by anything else known to humankind. This high complexity, though, renders the study of life – if not impossible – impractical. I believe, however, that even the most complex problems can be simplified, modeled, and eventually, understood.

In this thesis, I report several innovative, largely computational, research projects which attempt to investigate recently found cellular mechanisms whose very prevalence and importance to the complexity of cells has yet to be assessed. The first project (Results, section 2.1) describes an approach to identify hundreds of long non-coding RNAs (lncRNAs) predicted to function as scaffolding molecules for proteins. The second project (Results, section 2.3) builds up on similar ideas, proposing the co-translational formation of thousands of protein complexes promoted by interactions with 3' untranslated regions (3'UTRs). Both works are highly exploratory – to be considered as a first overview on the potential pervasiveness of these mechanisms – as well as extensive, taking advantage of large-scale datasets of protein-protein and protein-RNA interactions. Indeed, my work is very much ingrained in the study of molecular interactions and cellular function on a genome-wide scale, in the context of the regulation of complex systems.

In addition, this thesis describes work on an update of a biological database containing predictions of *moonlighting* proteins (Results, section 2.2) – proteins that perform several unrelated functions – a prime example of cellular complexity. These proteins are used here to study the regulation of multifunctionality by 3'UTRs. Furthermore, I include a manuscript currently in preparation by my colleagues, regarding the regulation of functionally-related messenger RNAs (mRNAs) by proteins (RNA regulon theory), for which I contributed (Results, section 2.4). Lastly, I add a list of publications involving my scientific contributions prior to this thesis (Appendix V).

I will first introduce the reader to the several topics approached by this thesis, with a focus in providing an up-to-date overview of the recent findings and state-of-the-art methods in these fields.

Table of contents

1. Introduction

1.1. Protein-protein and protein-RNA interactions

Biological systems have functions that none of their constituent parts have alone. Cellular functions are performed through interactions between molecules. Thus, the study of macromolecular interactions – such as interactions between nucleic acids, proteins and lipids – is crucial to understand the genetics and evolution of life. Of the possible interactions between macromolecules, protein-protein interactions are one of the most studied, with decades of research producing high-quality interaction maps for several species. Trailing behind, experimental methods to detect protein-RNA interactions may now be reaching their golden-age, but have not yet been sufficiently used to provide a complete network of interactions. Computational approaches to predict protein-RNA interactions have been growing in parallel with experimental ones and are now applicable to genome-wide studies. The data provided by experimental and computational methods can be used to construct interaction networks representing all the interactions of an organism, i.e. an interactome, that can include proteins as well as RNAs. This chapter provides the methodological framework in which my work was carried out, as well as the state of the art of the methods and the data available.

1.1.1. Protein-protein interactions

Proteins have been studied for several decades by structural biologists, molecular biologists, cellular biologists, biochemists and biophysicists, together creating an abundance of knowledge on their many properties and biological functions (De Las Rivas and Fontanillo, 2010). Most proteins perform their biological functions by interacting with other proteins. As a consequence, the ability of proteins to interact and act in concert with one another is perhaps one of the most studied features of proteins. Determining all the protein-protein interactions occurring within an organism would aid our understanding of biology as an integrated system (Cusick *et al.*, 2005).

Types of protein interactions

Protein-protein interactions (PPIs), physical contacts between proteins occurring in a cell, are very diverse and can be distinguished by their type of interaction, such as homo- versus hetero-oligomeric, obligate versus non-obligate and stable versus transient interactions:

- *homo- versus hetero-oligomeric*: interactions can exist between identical or homologous protein units (i.e. homo-oligomers), or between different proteins (i.e. hetero-oligomers). For homo-oligomers, the interaction can involve the same surface for the two monomers (isologous), or different surfaces (heterologous), which can lead to infinite aggregation (Nooren and Thornton, 2003).

- *obligate versus non-obligate*: an interaction is obligate if the components do not have a stable structure on their own *in vivo*. Most hetero-oligomeric complexes involve non-obligate interactions between components that can exist independently (e.g. antibody-antigen, receptor-ligand and enzyme-inhibitor) (Nooren and Thornton, 2003).

- *stable versus transient*: stable interactions are held for a long time, and are often found in permanent complexes (such as obligate complexes), where the proteins are called ‘subunits’. On the other hand, transient interactions are reversible and may occur only briefly, these are generally dependent on the immediate cellular context and regulate the dynamics of biological networks (e.g. interactions of a signalling cascade) (Perkins *et al.*, 2010).

Domain-domain interactions and domain-motif interactions

Across species, more than 600 modular protein domains are known to mediate PPIs, recognizing exposed sites on their binding partners (Xia *et al.*, 2008). Common protein-binding domains are the PDZ domain, the LIM domain and the SH2 and SH3 domains (Amos-Binks *et al.*, 2011). Variations of such domains can be grouped into families. For example, the SH3 domain family includes more than 250 members encoded by the human genome (Pawson and Nash, 2003).

Modular protein-binding domains can interact through: *i)* homo- or heterotypic domain-domain interactions, *ii)* short peptide sequence motifs (Figure 1.1). Moreover, the same domain may perform binding through different mechanisms. For example, PDZ domains can mediate specific PDZ-PDZ domain interactions, but usually recognize C terminal short peptide motifs (Pawson and Nash, 2003).

In addition, other parts of the protein surface are found to mediate interactions (particularly transient interactions), such as intrinsically disordered regions and short linear motifs (SLiMs, also known as linear motifs (LMs)). SLiMs are conserved stretches of around 2 to 8 residues, often occurring within an intrinsically disordered region of a protein, and many are able to interact with globular domains of other proteins, albeit with lower affinity than domain-domain interactions (Perkins *et al.*, 2010; Akiva *et al.*, 2012; Davey *et al.*, 2012). Ongoing efforts to determine interactions mediated by SLiMs in eukaryotes can be found on the Eukaryotic Linear Motif (ELM) resource (Dinkel *et al.*, 2016).

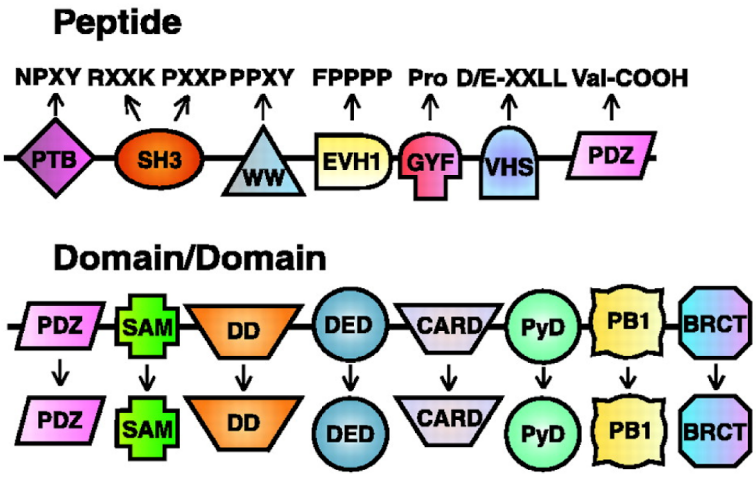


Figure 1.1 | Examples of protein-binding domains interacting with specific peptide motifs (Peptide) or with other homologous domains (Domain/Domain). Figure adapted from (Pawson and Nash, 2003).

Specificity of protein-protein interactions

The cellular environment is known to be crowded with macromolecules (Ellis, 2001), and in this context, a protein could have many potential binding partners. However, to perform a certain function at a given time, the physical contacts between molecules should be specific. In practice, many proteins interact with a specific partner (e.g. enzyme-inhibitors), while others are multispecific, having multiple partners often competing for the same binding interface (Nooren and Thornton, 2003). The specificity can be provided by complementarity in shape and chemistry, but many other factors play a role, such as, *i)* the presence and co-localisation of the proteins in space and time, i.e. their co-expression and co-localisation in the same compartment, *ii)* the current state of the proteins, such as the attachment to a cofactor (e.g. heme group, Mg^{2+} ions) or the post-translational modifications of a protein (e.g. phosphorylation, acetylation)

which can activate, deactivate, change the conformation of the protein or even cover or uncover binding sites (Nooren and Thornton, 2003; De Las Rivas and Fontanillo, 2010; Akiva *et al.*, 2012).

1.1.2. Methods to identify protein-protein interactions

Protein-protein interactions can be determined experimentally at large scale via two types of approaches, those leading to binary interactions and those termed 'co-complex'. Approaches leading to binary interactions measure direct interactions between pairs of proteins, whereas co-complex approaches detect interactions among groups of proteins. Importantly, both types of approaches are complementary in respect to the type of interactors they detect, and both are highly scalable and have been applied to thousands of proteins, identifying tens of thousands of interactions in model organisms (Brückner *et al.*, 2009; De Las Rivas and Fontanillo, 2010; Rao *et al.*, 2014; Rolland *et al.*, 2014). The most used method for each type of approach is presented here. Several other methods to determine protein-protein interactions exist but are used to a lesser extent and are not covered here. These include, not exhaustively, BioID (Roux, Kim and Burke, 2013), FRET (Margineanu *et al.*, 2016), BRET (Dimri, Basu and De, 2016) and BiFC (Miller *et al.*, 2015; Snider *et al.*, 2015).

Yeast two-hybrid (Y2H) method

Methods to identify binary protein-protein interactions measure direct physical interactions between pairs of proteins, the most common method used being the yeast two-hybrid (Y2H) (Fields and Song, 1989). This method, along with BiFC, FRET and others, is a protein complementation assay, in which a molecular complex is formed when inactive fragments of a reporter protein are assembled due to a bait-prey interaction (Morell, Ventura and Avilés, 2009). Particularly, Y2H is carried out by screening the interactions of a protein of interest against potential partners, in a pairwise fashion (Figure 1.2a) (Cusick *et al.*, 2005). Two domains are required for the transcription of a reporter gene in a Y2H assay: *i)* a DNA binding domain (BD) fused to the protein of interest (bait protein), and *ii)* an activation domain (AD), responsible for activating the transcription of DNA, fused to the potential interacting protein (prey protein). Through the fused BD domain, the bait protein binds the upstream activator sequence (UAS) of the promoter of the reporter gene. Only the interaction between the bait and prey proteins can reconstitute a functional transcription factor, which leads to the recruitment of RNA polymerase II and transcription of the reporter gene, whose expression would thus indicate that a physical interaction between prey and bait has occurred (Brückner *et al.*, 2009). Detection of the reporter expression can be detected through the

measurement of an enzymatic activity or a fluorescence in the cell (Rao *et al.*, 2014). To reduce non-specific interactions, studies often assay several reporter genes in parallel (Brückner *et al.*, 2009).

It is generally considered that PPIs identified through Y2H represent biophysically possible interactions, without accounting for spatiotemporal information. Even though Y2H is performed *in vivo*, it is traditionally performed in the nucleus of yeast cells, and thus not taking into account the context in which the interaction exists naturally. Indeed, due to limitations of using yeast cells, some interactions may be systematically missed, such as those involving membrane proteins or involving post-translational modifications not occurring in yeast cells (Koegl and Uetz, 2007; Stynen *et al.*, 2012). However, several Y2H systems were introduced to tackle some of these issues, such as the MAPPIT system used for mammalian cells (Tavernier *et al.*, 2002), and the SCINEX-P system, which screens interactions between extracellular proteins (Urech, Lichtlen and Barberis, 2003). Indeed, the state-of-the-art of Y2H-based methods allows for almost the entire proteome to be amenable to Y2H assays (Brückner *et al.*, 2009).

Y2H methods present several advantages compared to other methods. Even though transient interactions are challenging to identify, due to their timescale, Y2H is sensitive to such interactions (Perkins *et al.*, 2010). More generally, the benefits of Y2H are its low cost and high scalability. Indeed, Y2H can be used to screen a bait against a set of preys in a protein matrix, such as proteome-wide sets of full length open reading frames (ORFs) (Stelzl *et al.*, 2005; Brückner *et al.*, 2009).

Affinity purification coupled to mass spectrometry (AP/MS)

Co-complex approaches to detect PPIs measure physical interactions among groups of proteins by tagging a bait protein with a molecular marker, and “fishing out” the group of proteins (prey proteins) that attach to it, followed by a purification and mass spectrometry (MS) analysis (Figure 1.2b). The most common co-complex method is the affinity purification coupled to mass spectrometry (AP/MS) or variations of this method, but other methods such as co-immunoprecipitation (Co-IP) are also widely used (Rao *et al.*, 2014). In AP/MS, the bait protein is fused with a two-part tag recognising IgG immunoglobulin and calmodulin. The bait is produced in physiological conditions, thus allowing the retrieval of *in vivo* complexes. Subsequently, these complexes are purified and the prey proteins are identified by MS (Dunham, Mullin and Gingras, 2012).

An advantage of AP/MS is that interactions that occur in the native cellular environment are identified. Moreover, if proteins are expressed in their normal cell conditions, identification of post-translational modifications may be detected (Cusick *et al.*, 2005). However, a disadvantage is that purification of

complexes can lead to the loss of real interactions, or even the gain of spurious interactions (Rao *et al.*, 2014). In addition, due to their nature, co-complex methods measure direct but also indirect interactions between proteins, i.e. detect proteins that are in complex with the bait protein, but not necessarily interacting directly with the bait protein. Due to this, several models have been used to represent and further analyse interactions produced by co-complex methods, including the spoke model, which considers only bait-prey interactions, and the matrix model, which takes into account both bait-prey and prey-prey interactions (De Las Rivas and Fontanillo, 2010; Zhang *et al.*, 2015).

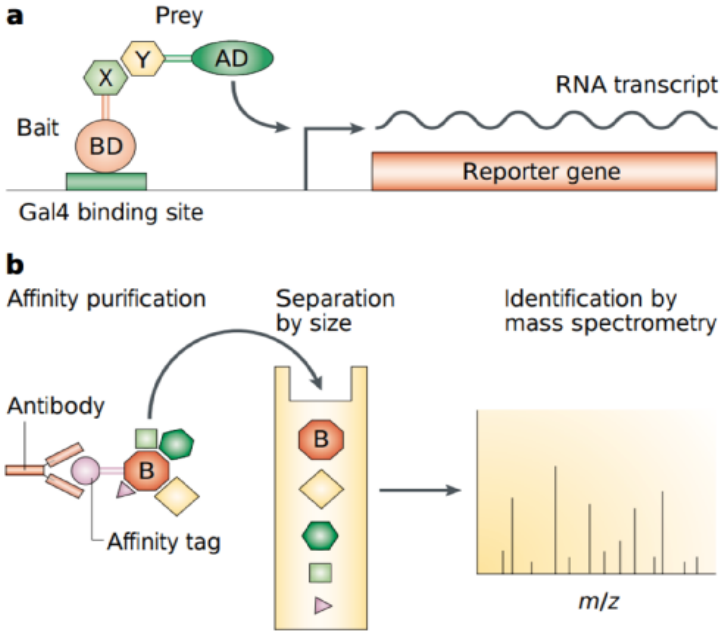


Figure 1.2 | Basic principles of common methods to detect protein-protein interactions. **a**) Yeast two-hybrid assay (Y2H); **b)** Affinity purification followed by mass spectrometry (AP/MS). Figure adapted from (Grünenfelder and Winzeler, 2002).

Resources of protein-protein interactions

Although *in silico* prediction methods exist (reviewed in (Rao *et al.*, 2014)), their use maybe less useful for some species – such as yeast and human – for which experimental PPI knowledge covers most interacting proteins (De Las Rivas and Fontanillo, 2010; Vidal, Cusick and Barabasi, 2011; Rolland *et al.*, 2014). Indeed, in yeast, more than 6,000 ORFs have been cloned and the interactions detected involve more than 70% of the proteome (Uetz *et al.*, 2000; Ito *et al.*, 2001). Likewise, a reference map of more than 14,000 high quality PPIs, including most of the proteome, has been generated for human (Rolland *et al.*, 2014), even though this number is still far from the 130,000 binary interactions estimated for the human interactome (Venkatesan *et al.*, 2009).

Millions of PPIs are stored among several databases, such as BioGRID (Chatr-Aryamontri *et al.*, 2017), IntAct (Orchard *et al.*, 2014) and APID (Alonso-López *et al.*, 2016), and these can be retrieved systematically using the ConsensusPathDB (Kamburov *et al.*, 2013) or the PSICQUIC web service (del-Toro *et al.*, 2013). In an attempt to improve data quality and curation of PPIs, the IMEx (Orchard *et al.*, 2012) and HUPO (Omenn, 2014) international consortiums are involved with defining standards for describing molecular interactions, including specific formats and controlled vocabulary (e.g. PSI-MI) to describe PPIs and the methods used (Hermjakob *et al.*, 2004).

1.1.3. Protein-RNA interactions

Regardless of their type or functionality, RNA molecules interact with proteins even while being transcribed, and continue to do so during all stages of their life (Stefl, Skrisovska and Allain, 2005). Like protein-protein interactions, protein-RNA interactions play an important role in many essential biological systems (Jones *et al.*, 2001). RNA-binding proteins (RBPs) recognize and bind RNA of any class (e.g. non-coding RNA, messenger RNAs (mRNAs)), either transiently or as part of a ribonucleoprotein (RNP) complex, affecting their processing, splicing, localisation, as well as their fate and function (Stefl, Skrisovska and Allain, 2005). For example, pre-mRNAs usually form RNP complexes with proteins called ‘heterogeneous nuclear ribonucleoproteins’ (HNRNPs), which play a role in their splicing, nuclear export, translation and stability (Chaudhury, Chander and Howe, 2010). Likewise, microRNA (miRNA) processing as well as its function is largely dependent on several RBPs such as the Dicer and the Argonaute proteins (Bartel, 2018). Nonetheless, knowledge of the full set of proteins that interact with an individual RNA during its lifetime still remains elusive.

RNA-binding protein interaction modes

The structural details of protein-RNA interactions have been described through X-ray crystallography and nuclear magnetic resonance analysis, such as the structure of the PAZ domain of Argonaute proteins binding several RNA oligonucleotides (Jones *et al.*, 2001; Carlomagno, 2014; Flores, Walshe and Ataide, 2014). For a protein-RNA interaction to occur, both the sequence and the structure of the RNA, i.e. the actual shape of the RNA, are important (Stefl, Skrisovska and Allain, 2005). Indeed, RBPs may use RNA-binding domains (RBDs) that bind sequence and/or structural motifs of the RNA, the most abundant being the zinc-finger motif, the RNA-recognition motif (RRM), the double-stranded RNA-binding motif (dsRBM) and the K homology (KH) motif (Jones *et al.*, 2001; Ray *et al.*, 2013; Hentze *et al.*, 2018). Most RBDs recognize stretches of 3 to 8 nucleotides that often allow a high degree of sequence variation in

them and are found in the majority of mRNAs (Lambert *et al.*, 2014; Mitchell and Parker, 2014). These short motifs can occur multiple times in the same sequence and act synergistically or even cooperatively (Hennig and Sattler, 2015).

Interestingly, recent studies have found that protein-RNA interactions can also be mediated by unconventional RNA binding mechanisms and that this phenomenon may be common (Helder *et al.*, 2016; Hentze *et al.*, 2018). Studies performing the RNA interactome capture method have found hundreds of novel RBPs, expanding the repertoire of RBPs to more than 2000 in human (Gerstberger, Hafner and Tuschl, 2014; Beckmann *et al.*, 2015; Hentze *et al.*, 2018). The RNA interactome capture method involves the crosslinking of RBPs to polyadenylated (poly(A)) RNAs *in vivo*, followed by the RNA capture and subsequent identification of interacting proteins by MS (Castello *et al.*, 2013). A recent variant of the RNA interactome capture technique, named chemistry-assisted RNA interactome capture (CARIC), allows capturing RBPs bound to not only to poly(A) RNAs but also non-poly(A) RNAs (e.g. pre-mRNA, many long non-coding RNAs) (Huang *et al.*, 2018). Surprisingly, a large proportion of the novel RBP identified do not contain any known RBD (Figure 1.3) (Beckmann, Castello and Medenbach, 2016). Up to now, it is not fully understood how these unconventional protein-RNA interactions are formed, but intrinsically disordered regions were found enriched among the RBP binding regions and may be responsible for such interactions (Calabretta and Richard, 2015; Castello *et al.*, 2016). Importantly, the discovery of the extent of unconventional RNA binding has raised important questions about the specificity and biological functions of such interactions. Recently, it was found that as many as 472 of such unconventional RBPs had been previously linked to virus-related processes, even if their RNA-binding potential had not yet been discovered (Garcia-Moreno, Järvelin and Castello, 2018).

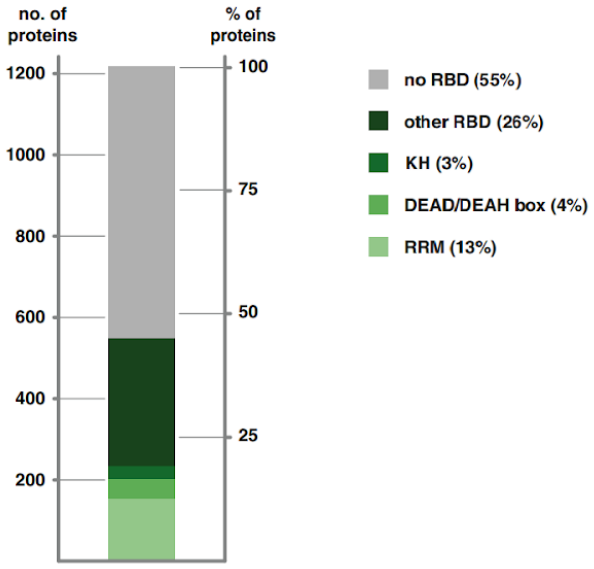


Figure 1.3 | Distribution of RNA-binding domains (RBDs) in RNA-binding proteins (RBPs). RBPs discovered by RNA interactome capture show to contain *i)* canonical RBDs such as the RRM, DEAD/DEAH box and KH domains, *ii)* other canonical and non-canonical RBDs (other RBD), *iii)* no known RBDs. There is a high proportion of RBPs without known RBDs. Figure adapted from (Beckmann, Castello and Medenbach, 2016).

Specificity of protein-RNA interactions

Generally, an RNA can be bound by multiple proteins (Chu *et al.*, 2015) and proteins can bind multiple RNAs, in some cases even thousands of RNAs, such as the subunits of the PRC2 protein complex and proteins involved in RNA processing (Milek, Wyler and Landthaler, 2012; Kretz and Meister, 2014; Van Nostrand *et al.*, 2016). However, *in vivo* studies have shown that specific RBPs may only bind a small fraction (e.g. 15%) of its described potential motifs (Taliaferro *et al.*, 2016; Van Nostrand *et al.*, 2016). About half of all RBPs may associate with RNA sites seemingly devoid of specific sequence or structural motifs. Indeed, some RBPs may bind RNA in a non-selective way to fulfill their functions, e.g. transfer RNA (tRNA) charger, mRNA export factor, RNA degradation (Jankowsky and Harris, 2015).

Assessing the target specificity of RBPs can be very challenging. This can be true even for RBPs with well-described domains, such as the PUF domains that, albeit having a well defined repetitive structure, are able to bind RNAs in many different ways (Koh *et al.*, 2009). Nevertheless, recent efforts have characterised the affinity and specificity of certain unconventional RBPs. These were done performing *in vitro* methods such as the RNAcompete and RNA Bind-n-Seq, in which pertinent RBP sequence motifs

can be determined by analysing the regions binding to sets of $1x10^8$ or more different synthesized oligonucleotides (Jankowsky and Harris, 2015; Ray *et al.*, 2017). Recently, RNA Bind-n-Seq has been systematically applied to 78 human RBPs containing RRM, KH and other RBDs and the results demonstrate the importance of contextual features for RNA recognition, such as the flanking sequences of linear RNA motifs and the secondary structure of the RNA (Dominguez *et al.*, 2018).

Interestingly, beyond the idea that RBPs target RNAs, it is now thought that some RNAs may actually target RBPs themselves (Figure 1.4) (Hentze *et al.*, 2018). This is corroborated by studies that found chromatin-modifying complexes and transcription factors being recruited, organised or inhibited by certain RNAs (Cech and Steitz, 2014).

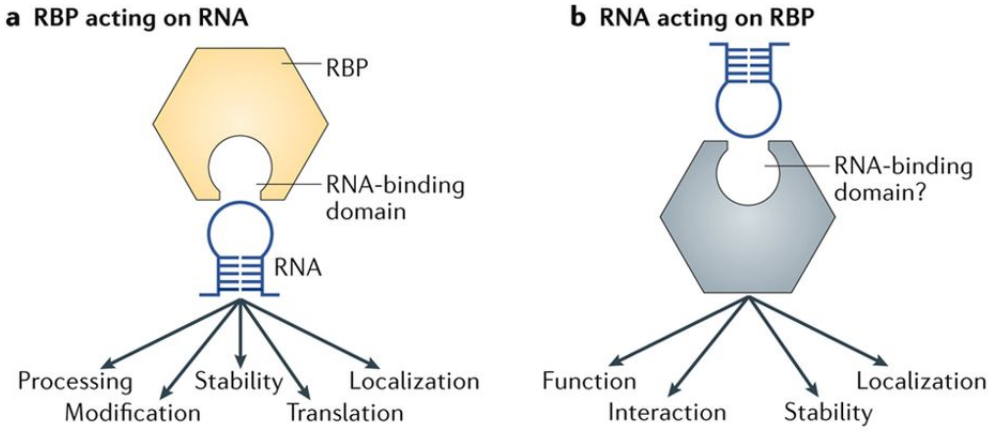


Figure 1.4 | Reciprocity between RNA and RNA-binding protein (RBP) interactions. **a)** RBPs can interact with RNAs in order to regulate RNA metabolism and function. **b)** RNAs can interact with RBPs, affecting their function and fate. Figure adapted from (Hentze *et al.*, 2018).

1.1.4. Methods to identify protein-RNA interactions

Protein-centric methods to detect protein-RNA interactions

Early methods to study protein-RNA interactions include the *in vitro* electrophoretic mobility shift assays (EMSA) (Hellman and Fried, 2007), the 'systematic evolution of ligands by the exponential enrichment' (SELEX) system (Tuerk and Gold, 1990) and the *in vivo* yeast three-hybrid system (Martin, 2012). A more recent *in vivo*, but low-throughput, method is the RNA immunoprecipitation (RIP), which is performed by precipitating the RBP under physiological conditions, thus preserving native complexes (Figure 1.5a). However, by itself, this method is unable to detect the RBP's binding site location (Barra and Leucci, 2017). Yet, along with the widespread of next generation sequencing technologies, RIP has

been coupled with high-throughput sequencing (RIP-seq), enabling the identification of the RNA fragments bound by the RBP and allowing the transcriptome-wide discovery of protein-RNA interactions (Zhao *et al.*, 2010).

Over the past decade, accompanying the recognition of the importance of protein-RNA interactions in many biological aspects, novel high-throughput methods to detect protein-RNA interactions have been developed (McHugh *et al.*, 2014). Most of these methods are based on cross-linking and immunoprecipitation (CLIP) (Ule *et al.*, 2003), followed by sequencing (CLIP-seq) (Wang *et al.*, 2009). Using CLIP, protein-RNA interactions in intact cells are crosslinked by ultraviolet (UV) irradiation, subsequently, protein-RNA complexes comprising the protein of interest are isolated by immunoprecipitation and the attached RNA fragments are sequenced (Figure 1.5a). Importantly, whereas RIP-based methods also find indirect protein-RNA interactions, the crosslinking step in CLIP allows to detect exclusively direct interactions, thus reducing possible artifacts (Barra and Leucci, 2017).

Besides CLIP-seq, many other variants of CLIP-based approaches have been developed, with improvements at the level of the protein capture, background noise, precise binding-site detection and other features. These include the ‘high-throughput sequencing of RNA isolated by UV crosslinking and immunoprecipitation’ (HITS-CLIP) (Darnell, 2010), the ‘photoactivatable-ribonucleoside enhanced CLIP’ (PAR-CLIP) (Danan, Manickavel and Hafner, 2016), the ‘individual-nucleotide resolution CLIP’ (iCLIP) (Konig *et al.*, 2011), the ‘UV-C crosslinking and immunoprecipitation platform infrared-CLIP’ (irCLIP) (Zarnegar *et al.*, 2016) and the ‘enhanced CLIP’ (eCLIP) (Van Nostrand *et al.*, 2016) methods. Several other variants exist and recent reviews comparing the different methodologies can be found on (Lee and Ule, 2018; Wheeler, Van Nostrand and Yeo, 2018). CLIP-based methods have been applied to a repertoire of RBPs. Notably, eCLIP has been applied to more than a hundred RBPs in two different cell lines, detecting tens of thousands of protein-RNA interactions (Van Nostrand *et al.*, 2016).

RNA-centric methods to detect protein-RNA interactions

Apart from the protein-centric methods described above, which allow to determine all the targets of an RBP, another set of protein-RNA interaction detection methods are RNA-centric. These allow the identification of all the proteins bound to an RNA of interest, normally through a pull-down of the RNA via complementary biotinylated oligonucleotides (Cirillo, Livi, *et al.*, 2014; Barra and Leucci, 2017). RNA-centric methods include the ‘chromatin isolation by RNA purification’ (ChIRP), the ‘capture hybridization analysis of RNA targets’ (CHART), as well as the ‘RNA antisense purification’ (RAP) (Figure 1.5b). RAP was initially developed to detect RNA-DNA interactions (Engreitz, Lander and

Guttman, 2015), but it is modifiable to detect RNA-protein interactions, identifying the associated proteins through MS (McHugh *et al.*, 2015). Despite being able to address research questions that the protein-centric methods cannot address, such as discovering the specificity and function of a certain RNAs, RNA-centric methods have yet to be applied for large sets of RNAs.

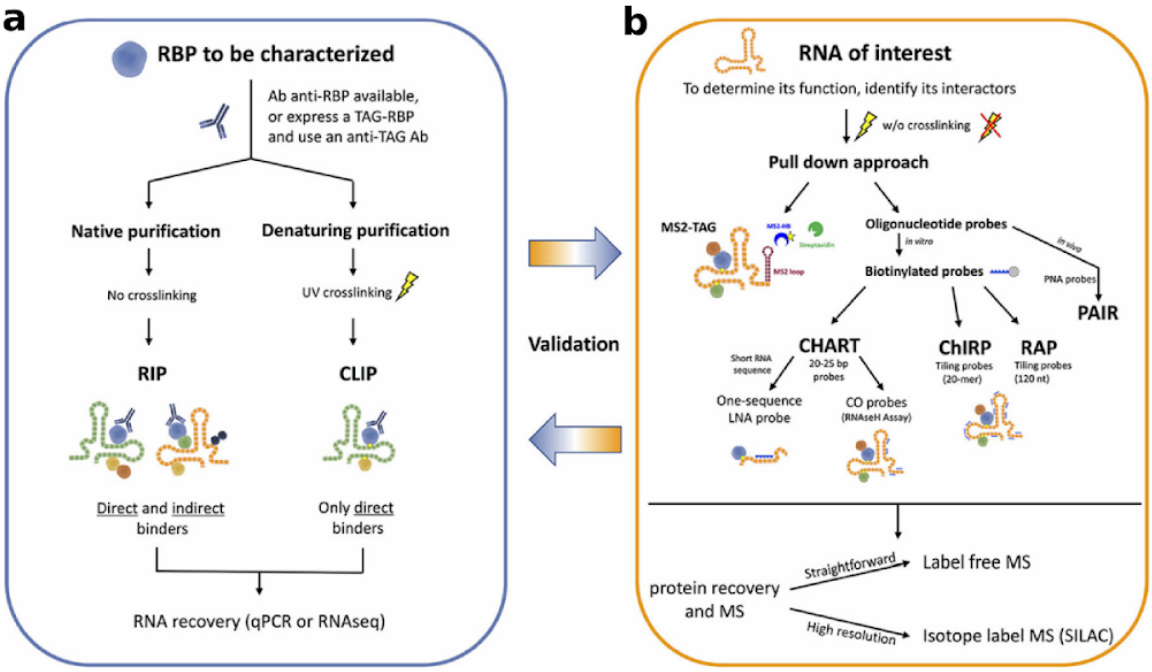


Figure 1.5 | Principles of methods to detect protein-RNA interactions. a) Protein-centric methods, **b)** RNA-centric methods. Figure adapted from (Barra and Leucci, 2017).

Resources of experimental protein-RNA interactions

Several public databases store collections of RNA interactions, including protein-RNA, DNA-RNA or RNA-RNA interactions, from many of the different experimental methods available. Among the dozen databases containing protein-RNA interactions (reviewed in (Yi, Zhao, Huang, *et al.*, 2017)), some of the most up-to-date include RAID (Yi, Zhao, Li, *et al.*, 2017), NPInter (Hao *et al.*, 2016), CLIPdb / POSTAR (Hu *et al.*, 2017) and the continuously updated AURA database, focused on untranslated regions (UTRs) of mRNAs (Dassi *et al.*, 2014). However, despite the development and large-scale use of several experimental methods to detect protein-RNA interactions, the establishment of a comprehensive and high quality protein-RNA network is lagging behind available protein interaction networks. In fact, several biases have been ascribed to several CLIP approaches, namely biases produced by UV crosslinking, such as pyrimidines being more photoactivatable than purines (Wheeler, Van Nostrand and Yeo, 2018), as well

as biases related to RNase over-digestion (Kishore *et al.*, 2011) and biases due to transcript abundance (Krakau, Richard and Marsico, 2017). These have led to the development of bioinformatic tools that re-analyse and attempt to control for some of the biases associated to iCLIP and eCLIP (Krakau, Richard and Marsico, 2017). Moreover, only a fraction of the results from eCLIP replicate experiments overlap (Van Nostrand *et al.*, 2016), although it is unclear if this due to a methodological limitation or a biological effect, as protein-RNA interactions may be highly transient and their interaction space very large. Also for RNA-centric methods, the results from different experiments may not agree, as observed for the XIST RNA, where only one common protein interactor was found among more than 600 proteins detected in five independent studies (Cirillo *et al.*, 2016).

Experimental protein-RNA interaction detection methods are likely to continue to evolve and produce more comprehensive sets of interactions in the next years, but current datasets are still insufficient for a global analysis of all the protein-RNA interactions in a cell.

Computational methods to identify protein-RNA interactions

Computational methods have been used to predict protein-RNA interactions, in an attempt to tackle experimental biases and coverage issues. Available computational methods are generally based on the structure and/or sequence properties of RNAs and proteins (Cirillo, Agostini and Tartaglia, 2013). These include RPI-Pred (Suresh *et al.*, 2015), RPI-Bind (Luo *et al.*, 2017), RNABindR (El-Manzalawy *et al.*, 2016), omiXcore (Armaos, Cirillo and Gaetano Tartaglia, 2017) and catRAPID (Bellucci *et al.*, 2011). The catRAPID method is based on physico-chemical properties of interacting nucleotides and amino acids, derived from X-ray crystallography data, as well as RNA and protein secondary structures, and has shown particularly effective in predicting protein-RNA interactions (Barra and Leucci, 2017). Moreover, the catRAPID *omics* expansion of method stands out due to its ability to predict protein-RNA interactions proteome- and transcriptome-wide (Agostini *et al.*, 2013; Cirillo, Marchese, *et al.*, 2014). However, it is limited to RNAs shorter than 1200 nucleotides and proteins with no more than 750 amino acids. More recently, Global Score was developed in order to predict long non-coding RNA interactions with proteins, without length restrictions, albeit not applicable to large-scale predictions (Cirillo *et al.*, 2016).

1.1.5. The interactome and functional components

The complete repertoire of molecular interactions in a cell or an organism is called an interactome (Sanchez *et al.*, 1999). Since the whole is greater than the sum of its parts, interactome networks are useful to understand complex biological systems, which may underlie most genotype to phenotype

relationships (Hartwell *et al.*, 1999; Vidal, Cusick and Barabasi, 2011). Importantly, interactomes can be represented as graphs, a mathematical object, and are thus amenable to many analysis stemming from graph theory (Pavlopoulos *et al.*, 2011). In that way, many biological insights can be derived from interactomes, such as, *i)* determining a protein or an RNA's function through its partner's functions (*guilty-by-association* principle), *ii)* identifying drug targets and designing effective strategies to treat disease, *iii)* detecting macromolecular complexes and functional modules (Hartwell *et al.*, 1999; Sharan, Ulitsky and Shamir, 2007; Pavlopoulos *et al.*, 2011; Ma and Gao, 2012; Ge, Li and Wang, 2016; Huttlin *et al.*, 2017). Indeed, functional network modules, protein communities and protein complexes have been comprehensively identified through network analysis (Figure 1.6) (Brun *et al.*, 2003; Brun, Herrmann and Guénoche, 2004; Emmanuelle Becker *et al.*, 2012; Didier, Brun and Baudot, 2015; Huttlin *et al.*, 2015, 2017). Together with biological pathway resources such as KEGG (Kanehisa *et al.*, 2017) and Reactome (Fabregat *et al.*, 2018), as well as datasets of protein complexes such as CORUM (Ruepp *et al.*, 2009) and Hu.MAP (Drew *et al.*, 2017), associations between groups of proteins sharing the same function or acting together in the same biological process are now well described.

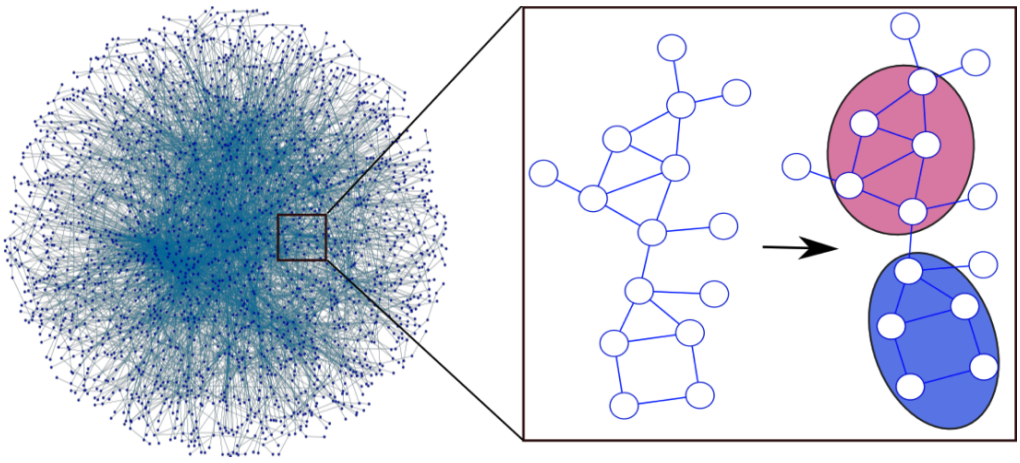


Figure 1.6 | Representation of a protein interactome and functional network modules. Interactomes such as those derived from complete maps of protein-protein interaction networks can be represented as graphs. Using graph theory approaches, such as analysing the community structure of a network, groups of interacting proteins, possibly involved in the same biological process, can be identified. Figure source: courtesy of Dr. Anaïs Baudot.

Even though the term 'interactome' is often synonymous with interactomes created with protein-protein interactions, in 1999, Bernard Jacq and collaborators coined this term to include "the complete repertoire of interactions potentially encoded by [a] genome" (Sanchez *et al.*, 1999). Indeed, interactomes can

include many types of molecular interactions such as protein-protein, protein-RNA interactions, RNA-RNA interactions, metabolic interactions and protein-DNA associations (gene-regulatory networks) (Vidal, Cusick and Barabasi, 2011; Gong *et al.*, 2018). Approaches that combine several types of molecular interactions exist, as a way to integrate different types of data, such as DNA methylation and gene expression, and to improve functional module detection, for example using multiplex biological networks (Didier, Brun and Baudot, 2015; Ma *et al.*, 2017). Moreover, even though protein-RNA interaction networks are far from complete, two recent approaches that combine protein-protein and protein-RNA interactomes were developed, in order to predict RNA function (Junge *et al.*, 2017; Cheng and Leung, 2018). Methods that combine networks or interactomes of different types of data in order to understand complex systems are promising, but yet uncommon.

1.2. Long non-coding RNAs and the protein scaffolding function

Non-coding RNAs (ncRNAs) are increasingly regarded as key players in regulation and their prevalence is correlated with organismal complexity. Research into one of the most common types of ncRNAs, the long non-coding RNAs (lncRNAs), is continuously expanding and novel unexpected functions of lncRNAs are regularly discovered. Indeed, lncRNAs have the potential to perform functions by interacting with DNA, other RNAs and proteins, or even through combinations of these. However, their identification, prevalence and biological importance is often debated. Recent work has put into question whether part of these transcripts can actually code for small peptides, and whether most lncRNAs are just a by-product of bi-directional transcription. On the other hand, the first studies to assess lncRNA functionality at a large-scale are now in place and a growing body of work finds lncRNAs to be associated to disease and various biological mechanisms. This chapter describes lncRNAs and the current views on their biological importance, going through the latest findings in this blooming topic, focusing on their potential as protein scaffolding molecules.

1.2.1. Definition, characteristics and prevalence of lncRNAs

Definition and features of lncRNAs

In the past decade, advances in the depth and quality of the transcriptome sequencing methods have revealed that 60% of the human genome is transcribed, even though only about 2% of the genome encodes proteins (Djebali *et al.*, 2012). A large part of this newly found transcriptional landscape is accounted for lncRNAs.

LncRNAs are defined as RNA molecules with low protein coding potential and larger than 200 nucleotides, a cutoff chosen to distinguish these RNAs from other smaller non-coding RNAs with well established functions (e.g. transfer RNAs, miRNAs, small nuclear RNAs) (Ulitsky, 2016). Molecularly, lncRNAs resemble mRNAs and share many common biogenesis steps (Figure 1.7), such as transcription by RNA polymerase II, 5’-capping, splicing and polyadenylation, although exceptions exist (Quinn and Chang, 2016). Unique features of lncRNAs are their lack of robustly translated open reading frames (ORFs), as well as low expression levels and low sequence conservation when compared to mRNAs (Quinn and Chang, 2016; Ulitsky, 2016).

Importantly, lncRNA genes are often expressed in a tissue-specific manner and their expression has been found to be regulated. Indeed, analysis of the expression of lncRNA genes in different cell lines found that 29% of the lncRNAs were expressed specifically in a single cell type, whereas only 10% of them were expressed in all cell types (Djebali *et al.*, 2012). Moreover, several studies have shown that lncRNAs are differentially expressed during differentiation, development and disease (Sheik Mohamed *et al.*, 2010; Clark and Blackshaw, 2014; Yuan *et al.*, 2017).

Categories of lncRNAs

LncRNAs are often categorised by their position relative to protein-coding genes or other genomic features. These include: *i)* sense-overlapping lncRNAs, in respect to protein-coding genes, and further subdivided into intronic- or exonic-overlapping, *ii)* antisense lncRNAs, if transcribed in the opposite direction of a protein-coding gene, *iii)* intergenic lncRNAs, not overlapping with any protein-coding gene, known as long intergenic non-coding RNAs (lincRNAs) (Figure 1.7) (Derrien *et al.*, 2012; Laurent, Wahlestedt and Kapranov, 2014).

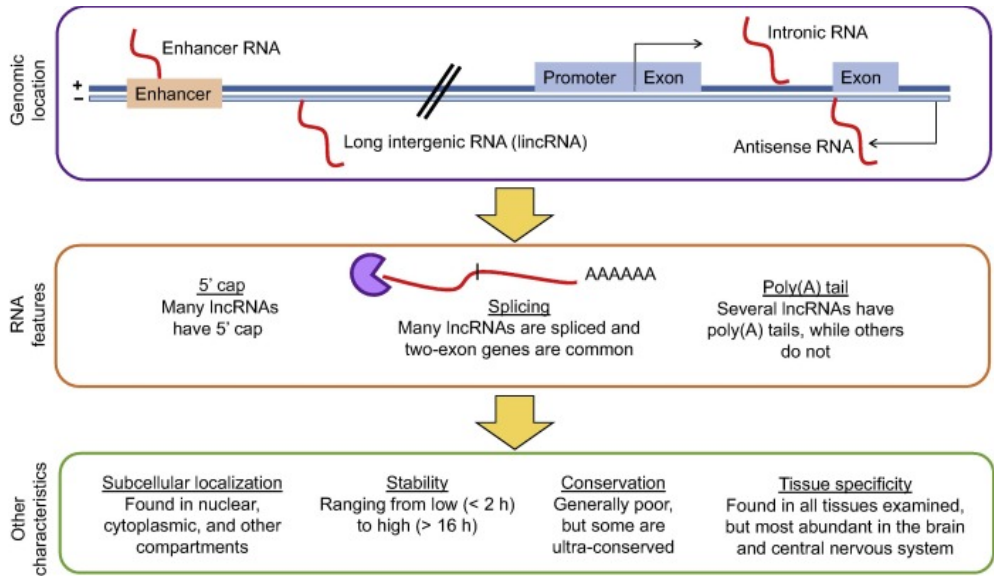


Figure 1.7 | Categories and features of long non-coding RNAs (lncRNAs). Figure adapted from (Fang and Fullwood, 2016).

Depending on their biogenesis origins, lncRNAs can come from the processing of other transcripts. For instance, lncRNAs can be originated from the splicing of introns of protein-coding genes (intronic RNAs). Moreover, lncRNAs can also be originated from the transcription of DNA elements such as promoters (i.e. promoter-associated long RNAs) and enhancers (i.e. enhancer RNAs), as a result of the bi-directional transcription of DNA elements (Laurent, Wahlestedt and Kapranov, 2014). Moreover, due

to their function, size or other specific features, lncRNAs can be further grouped into other categories, such as the competing endogenous RNAs (ceRNAs), which contain miRNA recognition elements, and the very stable circular RNAs (circRNAs), which are produced from pre-mRNA back-splicing circularization (Kartha and Subramanian, 2014; Chen, 2016).

Prevalence of lncRNAs

LncRNAs are found in every branch of life, but their prevalence seems to be correlated with organismal complexity, being more frequent in higher organisms such as primates (Mattick, Taft and Faulkner, 2010). Indeed, a connection between lncRNAs in the nervous system and the evolution of the brain in higher vertebrates has been proposed (Clark and Blackshaw, 2014).

The exact prevalence of lncRNAs in human is debatable but thought to be high, with tens of thousands of lncRNAs consistently detected through next-generation and third generation sequencing (Iyer *et al.*, 2015; Hon *et al.*, 2017; Uszczynska-Ratajczak *et al.*, 2018). Conservative and curated resources, such as GENCODE (Harrow *et al.*, 2012), provide around 15,000 human lncRNA genes producing 28,000 different transcripts. However, datasets that integrate lncRNA genes identified from several sources, such as NONCODE (Fang *et al.*, 2018), include up to 97,000 lncRNA genes producing 172,216 transcripts. Novel lncRNAs are identified in every new study, but a large proportion of them are yet poorly annotated and different lncRNA datasets only mildly overlap (Uszczynska-Ratajczak *et al.*, 2018). Moreover, several ribosome profiling studies proposed that many lncRNAs may in fact have some coding potential (Ruiz-Orera *et al.*, 2014; Mackowiak *et al.*, 2015; Olexiouk, Van Criekinge and Menschaert, 2018), in some cases encoding functional peptides (Nelson *et al.*, 2016). However, the amount of lncRNAs being translated into peptides is an open question, with several studies suggesting these events may be rare and indeed exceptional (Verheggen *et al.*, 2017; Uszczynska-Ratajczak *et al.*, 2018).

1.2.2. Biological functions of LncRNAs

Prevalence of functional lncRNAs

LncRNAs, just as other RNA molecules, have been shown to functionally bind DNA, protein as well as other RNAs. However, the importance and role of most lncRNAs, if any, remains elusive. Indeed, less than 1% of the identified lncRNAs have been experimentally characterised or associated to disease (Uszczynska-Ratajczak *et al.*, 2018). It is thought that many - or even most - lncRNAs are simply the by-products of the bi-directional transcription of promoters and enhancers, and thus deprived of any

function as an independent molecule (Kopp and Mendell, 2018). Large-scale attempts to discover functional human lncRNA loci have found 499 loci affecting cell growth, among 16,401 loci tested (Liu *et al.*, 2017). However, it would have to be tested if these lncRNAs function independently of the DNA sequence from which they are transcribed (Kopp and Mendell, 2018). Regardless, the functions of at least 156 lncRNA molecules are catalogued in lncRNAdb and other compilations, and these are continuously growing (Quek *et al.*, 2015; Marchese, Raimondi and Huarte, 2017). Moreover, several approaches have predicted the function of lncRNA through their genomic proximity and co-expression with protein-coding genes, as well as through their protein, DNA and RNA interactions (Park *et al.*, 2014; Li *et al.*, 2015; Gawronski *et al.*, 2018).

Known lncRNA functions

Most of the lncRNA functions seem to involve some form of regulation, and they do so through many different mechanisms (Figure 1.8). Among other functions, lncRNAs have been shown to: *i)* regulate gene expression in *cis* and in *trans ii)* regulate mRNA processing and stability, *iii)* act as protein or RNA (e.g. miRNA) decoys, also known as ‘sponges’, *iv)* regulate protein activity, as well as *v)* act as scaffolds for high-order complexes (Geisler and Coller, 2013; Marchese, Raimondi and Huarte, 2017; Ransohoff, Wei and Khavari, 2018).

In many cases, only one or a few example lncRNAs have been found to be involved in each function or mechanism, and it is unclear if these are one-time exceptions or if other examples are yet to be found. However, a recurrent function of lncRNAs is their regulation of gene expression (both activation and repression) through the formation of complexes with chromatin-associated proteins, such as the relationship between the polycomb repressive complex 2 (PRC2) protein complex and the HOTAIR lncRNA, as well as several other lncRNAs (Khalil *et al.*, 2009; Spitale, Tsai and Chang, 2011; Hendrickson *et al.*, 2016). Moreover, several lncRNAs were found to regulate gene expression through many other mechanisms such as RNA-DNA triplex formation, binding of RNA polymerase II or its cofactors, as well as through transcription factor binding (Geisler and Coller, 2013; Li, Syed and Sugiyama, 2016).

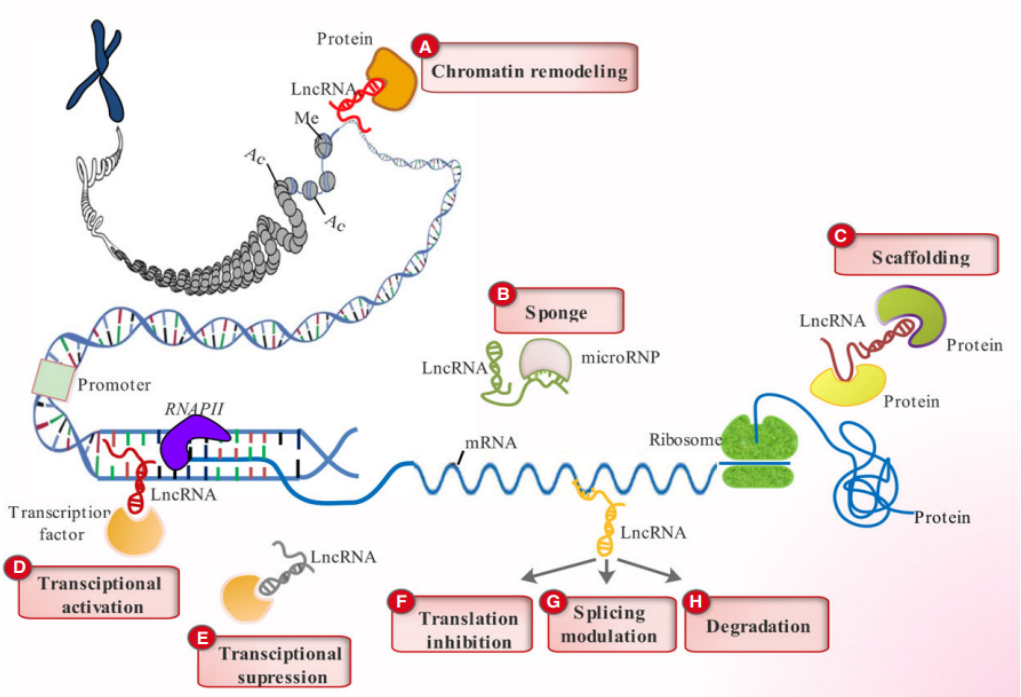


Figure 1.8 | Overview of lncRNA functions. LncRNAs can: (A) recruit chromatin remodelling complexes that affect chromatin organization and thus gene expression; (B) act as 'sponges' or decoys by binding one or several complementary miRNAs; (C) scaffold several proteins that act together in the same biological pathway; (D) activate transcription by guiding transcription factors to their promoters; (E) suppress transcription by sequestering transcription factors. By associating with mRNAs, lncRNAs can (F) inhibit translation; (G) modulate splicing; (H) regulate mRNA degradation. Figure adapted from (Salehi *et al.*, 2017).

Notably, certain lncRNAs are highly expressed and have strong impacts in the cell, such as the NORAD lncRNA, an exceptionally well conserved lncRNA, capable of sequestering virtually all the PUMILIO proteins upon expression stimulation triggered by DNA damage, leading to the maintenance of genomic stability in the cell (Lee *et al.*, 2016). Recently, NORAD was also found to interact with RBMX and be essential in assembling a protein complex dubbed NORAD-activated ribonucleoprotein complex 1 (NARC1), involving RBMX and other DNA replication and repair proteins (Munschauer *et al.*, 2018).

Other well studied lncRNAs include the XIST (X-inactive specific transcript) lncRNA, a main player in the X chromosome inactivation (Briggs and Reijo Pera, 2014), and the MALAT1, responsible for the formation of nuclear speckles and associated to several diseases (Zhang, Hamblin and Yin, 2017). Both interact with several proteins and may be considered scaffolding molecules, a mechanism which will be described further.

1.2.3. LncRNA scaffolding of protein complexes

Molecules acting as protein scaffolds

To perform a cellular function, the components of a complex or a pathway need to be physically close to each other, whether transiently or permanently. In a cell crowded with macromolecules, one way to gather the required components is to employ molecular scaffolds that piece together components in a selective way (Good, Zalatan and Lim, 2011; Spitale, Tsai and Chang, 2011). Proteins are often used as scaffolds for other proteins, especially in signalling pathways, in this way increasing the interaction efficiency between the partner molecules, as well as regulating them allosterically (Shaw and Filbert, 2009; Good, Zalatan and Lim, 2011; Garbett and Bretscher, 2014).

RNA molecules have been engineered to act as protein scaffolds *in vivo* in order to co-localize enzymes. These have shown to be more efficient when compared to DNA or protein scaffolders (Delebecque, Silver and Lindner, 2012; Sachdeva *et al.*, 2014). Moreover, at least theoretically, the use of RNA scaffolds should also present several advantages when employed by living organisms; *i)* due to their size, RNA molecules could potentially bind 5 to 20 proteins per 100 nucleotides, whereas a protein with 100 amino acids would only bind one or two proteins simultaneously; *ii)* unlike proteins, RNAs can act immediately during transcription; *iii)* RNA molecules, particularly lncRNAs, are able to evolve and adapt faster in order to interact with other molecules (Chujo, Yamazaki and Hirose, 2015).

Known lncRNAs acting as protein scaffolds

A dozen lncRNAs may function as protein scaffolds in human, including the LINP1 (Zhang *et al.*, 2016), the GUARDIN (Hu *et al.*, 2018), the SAMMSON (Leucci *et al.*, 2016) and the TERC (Telomerase RNA component) lncRNAs (Zhang, Kim and Feigon, 2011; Cech and Steitz, 2014). The TERC is part of the telomerase ribonucleoprotein, which maintains the terminal portions of eukaryotic chromosomes. Besides providing the template for telomeric DNA synthesis, TERC scaffolds the protein assembly required for this process, using several independent structural domains to recruit and bind several proteins simultaneously (Figure 1.9) (Egan and Collins, 2010).

Importantly, the described cases of lncRNAs acting as protein scaffolds were found through serendipity and no approach to identify them at a large-scale has been undertaken. It is possible that many yet uncharacterized lncRNAs act as protein scaffolds but do so in specific cell-types and stress conditions.

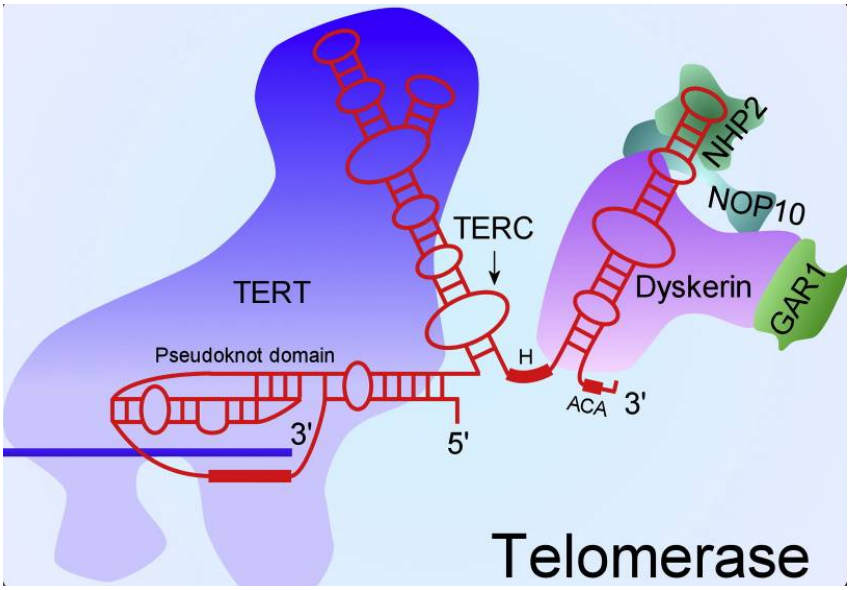


Figure 1.9 | Telomerase RNA component (TERC) as an example of an lncRNA scaffolding proteins. The highly structured TERC lncRNA interacts with several proteins simultaneously, including TERT, Dyskerin and NHP2. Figure adapted from (Kirwan and Dokal, 2009).

Granule-forming lncRNAs

Several RNAs were found to transiently assemble groups of proteins, such as the XIST lncRNA (Chu *et al.*, 2015; Cirillo *et al.*, 2016) and the NEAT1 (Nuclear Paraspeckle Assembly Transcript 1), a lncRNA responsible for the formation of stress-dependent nuclear granules named paraspeckles (Clemson *et al.*, 2009; Fox *et al.*, 2018). Indeed, cellular granules such as Cajal bodies, nuclear speckles, paraspeckles and processing bodies (P-bodies) all involve protein and RNA components and their specific interactions (Helder *et al.*, 2016) and these RNAs may be considered as having a protein scaffolding function.

The paraspeckle formation by the 23-kb NEAT1 lncRNA is one of the best studied examples of RNA granule formation. Out of the more than 40 different proteins found to compose these nuclear granules, several were found to be absolutely required, such as NONO and SFPQ, whereas other proteins were not found to be essential to the granule formation, but may instead regulate the granules (Fox *et al.*, 2018). While their role remains to be fully understood, paraspeckles are thought to sequester proteins as well as mRNAs through binding to inverted Alu-sequences in their 3’ untranslated regions (3’UTRs). Indeed, paraspeckles were found to post-transcriptionally regulate circadian genes as a consequence of their circadian expression and mRNA sequestration (Torres *et al.*, 2016, 2017). Furthermore, paraspeckles have

been associated to female reproduction, nervous system diseases and several types of cancer (Nishimoto *et al.*, 2013; Nakagawa *et al.*, 2014; Fox *et al.*, 2018).

Since the proteins identified in RNA granules often bear intrinsically disordered low-complexity regions which could be responsible for their aggregation into granules (Kato *et al.*, 2012; Protter *et al.*, 2018), it is yet unclear how many RBPs bind the RNAs directly and how many proteins are gathered through other mechanisms. Further work into the formation and dynamics of RNA granules should elucidate the importance of granule-forming RNAs and their potential scaffolding function.

1.3. Roles of 3’-untranslated regions (3’UTRs) in regulation

Found in virtually all the messenger RNAs of bacteria, archaea and eukaryotes, 3’UTRs are well characterised as important sequences that influence mRNA’s fate and thus protein synthesis. New alternative 3’UTR isoforms are continuously discovered for most genes through 3'-end sequencing, and the usage of the 3’UTR isoforms can be specific for some cell types and development stages. Together with findings of novel 3’UTR functions, particularly through their ability to promote the formation of co-translational protein complexes, this suggests that 3’UTRs may be more important for the regulation of biological complexity than previously known. In this chapter, the formation of 3’UTR isoforms and their biological functions are overviewed, with a focus on the recent discovery of a 3’UTR function that involves the formation of protein complexes during protein translation that impacts protein cellular localisation and function.

1.3.1. 3’UTR biogenesis and alternative polyadenylation (APA)

3’UTR biogenesis

Typically, pre-mRNAs undergo transcription, splicing, cleavage, polyadenylation and other RNA processing events leading to the maturation of the mRNAs. Mature mRNAs, ready for translation, are composed of a coding sequence (CDS), 5′ and 3′ untranslated regions (UTRs), as well as a 5’ cap and a 3’ poly(A) tail on each end, respectively. 3’UTRs are thus the region between the stop codon and the start of the poly(A) tail of an mRNA. (Matoulkova *et al.*, 2012).

The pre-mRNA 3’-end processing is a crucial multi-step process consisting in *i)* the cleavage at a polyadenylation site (PA site), often formed of “CA” dinucleotides; *ii)* the attachment of a poly(A) tail, in human usually comprising 50 to 250 adenines, essential for the protection of the transcript against exonucleases and the export of the mRNA to the cytoplasm (Matoulkova *et al.*, 2012). The 3’-end processing requires several protein complexes, together comprising more than 80 proteins, usually recruited by the nascent RNA during transcription through the polyadenylation signal (PAS), although other flanking sequences are also important, and alternative 3’-end processing mechanisms exist (Proudfoot, 2011; Gruber *et al.*, 2014). The PAS, often a “AAUAAA” hexamer, is a highly conserved signal for cleavage, normally localised 10-30 nucleotides upstream of the actual 3’-end cleavage site.

After intron splicing, human 3’UTRs have a median length of around 700 nucleotides. Alternative splicing in 3’UTRs (and 5’UTRs) can lead to mRNAs encoding the same protein but with a different UTR

sequence (Mignone and Pesole, 2018). The 3′ splice site of the last intron and the PAS set the terminal exon boundaries (Herzel *et al.*, 2017). A coupling between polyadenylation and splicing on terminal exons exists, potentially as competing processes (Movassat *et al.*, 2016; Herzel *et al.*, 2017). Indeed, a recent study applying long-read sequencing to nascent RNAs in yeast further suggests that the crosstalk between splicing and 3′ end processing occurs co-transcriptionally (Herzel, Straube and Neugebauer, 2018).

Alternative polyadenylation (APA)

A gene can be cleaved in multiple PA sites and thus produce more than one transcript from a single gene, analogous to alternative splicing. This mechanism is called alternative cleavage and polyadenylation (or simply alternative polyadenylation, APA) (Di Giammartino, Nishida and Manley, 2011). PA sites can be found within 3'-terminal introns and exons, affecting the pre-mRNA splicing and forming mRNAs with alternative 3′-terminal exons. In some cases, APA can lead to the formation of different protein isoforms (Di Giammartino, Nishida and Manley, 2011). For example, intronic polyadenylation can lead to truncated protein production with functional consequences (Lee *et al.*, 2018; Singh *et al.*, 2018). Besides producing alternative protein isoforms, APA frequently leads to the presence of transcripts with variable 3'UTR lengths (Figure 1.10) (Matoulkova *et al.*, 2012). Indeed, it is thought that up to 79% of the human genes may express alternative 3'UTRs through APA (Lianoglou *et al.*, 2013; Mayr, 2017).

Even though APA in 3'UTRs would lead to the same protein isoform, the lengths of 3'UTRs seem to be under selective pressure. In fact, similar to the prevalence of long non-coding RNAs, 3'UTR length has greatly expanded throughout metazoan evolution and is correlated with organism complexity (Chen *et al.*, 2012; Mayr, 2017). 3'UTR length is inversely correlated with mRNA stability and gene expression, i.e. short 3'UTRs are more stable and lead to higher protein synthesis, possibly because short 3'UTRs are less likely to be affected by mechanisms that regulate or degrade long 3'UTRs, such as miRNA targeting (Matoulkova *et al.*, 2012).

Given the diverse functions of 3'UTRs (see below), each mRNA-3'UTR isoform can behave differently. Indeed, many genes undergoing APA have evolved ways to regulate the prevalence of each isoform in different cell types and developmental stages, and APA can be thought as one of the tools that organisms have to achieve tissue-specificity (Ulitsky *et al.*, 2012; Lianoglou *et al.*, 2013). Many factors, such as the 3'-end processing machinery, the splicing machinery and RBPs such as HuR (also known as ELAVL1) can influence APA (Dutertre *et al.*, 2014; Herzel, Straube and Neugebauer, 2018). In fact, APA can be affected by several physiological conditions such as developmental-stage, cell growth, circadian rhythm

and disease, further evidencing its regulatory importance (Di Giammartino, Nishida and Manley, 2011; Torres *et al.*, 2018).

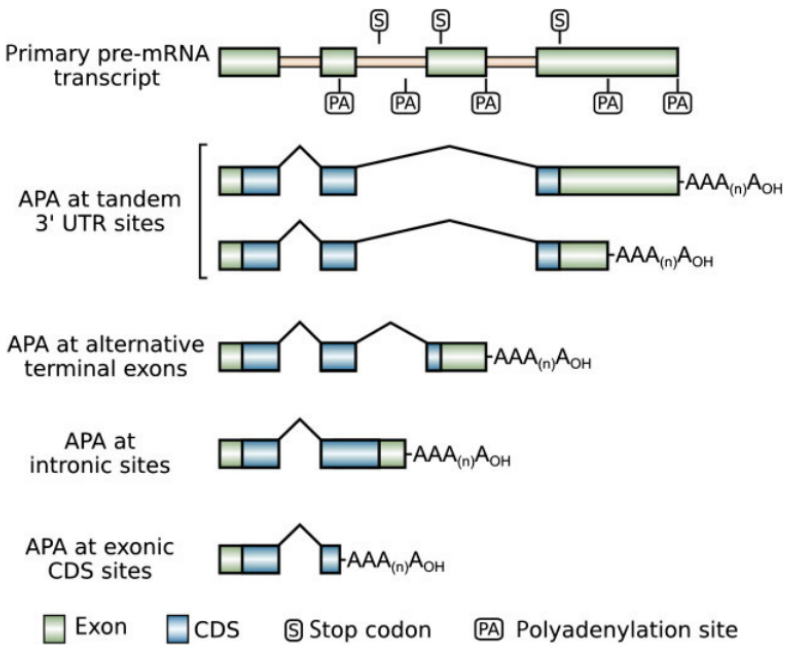


Figure 1.10 | Possible outcomes of alternative polyadenylation (APA). APA at tandem 3'UTR sites leads to transcripts with the same CDS sequence but different 3'UTRs. APA events in other locations can lead truncated transcript isoforms. Figure from (Gruber *et al.*, 2014).

1.3.2. Biological functions of 3'UTRs

Besides being the site for transcript cleavage and polyadenylation, 3'UTRs have been found to be involved in several regulatory processes. 5'UTRs, the untranslated region on the 5' end of mRNAs, contain sequences responsible for translation initiation. 3'UTRs are usually longer than 5'UTRs and the sequence constraints in 3'UTRs are generally more relaxed, thus more available for the evolution of regulatory elements (Barrett, Fletcher and Wilton, 2012).

The importance of 3'UTRs in higher organisms is corroborated by the fact that 3'UTRs contain many stretches of sequence that are as conserved as protein-coding regions, often containing binding sites for miRNAs but also RBPs. Indeed, a phylogenetic study has found more than 3000 hyper conserved elements in vertebrate 3'UTRs and shown that these are often targeted by RBPs, supporting the importance of RBP-3'UTR interactions (Dassi *et al.*, 2013). Moreover, both sequence and structural features have shown to be important for RBP binding of 3'UTRs.

A well known function of 3'UTRs is their ability to regulate their mRNA's translation through several mechanisms. 3'UTRs often contain miRNAs binding sites (i.e. RNA-RNA interactions by nucleotide complementarity), which either induce translational repression or mRNA decay, both occurring through multistep processes involving the recruitment of several protein complexes (Matoulkova *et al.*, 2012; Iwakawa and Tomari, 2015). Indeed, of all the mRNA regions, 3'UTRs are by far the preferential target of miRNAs and it is common for a 3'UTR to contain dozens of miRNA binding sites. Alike the use of certain miRNAs in a tissue-specific or developmental stage-specific manner, alternative 3'UTRs are also usually expressed (e.g. through APA) in a tissue and developmental stage-dependent manner (Barrett, Fletcher and Wilton, 2012). In fact, miRNAs have been shown to affect the evolution of 3'UTRs, for instance, by promoting the expression of tissue-specific 3'UTR isoforms that lack certain miRNA binding sites (Stark *et al.*, 2005).

Besides 3'UTR regulation mediated by other RNAs such as the miRNAs, 3'UTRs have been found to affect mRNAs, and thus cellular protein levels, through several other mechanisms that involve the binding of RBPs and the formation of protein complexes (Szostak and Gebauer, 2013). Indeed, 3'UTRs regulate: *i)* the export of mRNAs to the cytoplasm, through binding of the PABP protein (poly(A) binding protein), *ii)* the targeting of mRNAs to specific subcellular compartments, where they will undergo localised translation, a mechanism that often occurs in response to extracellular cues and is especially relevant in highly polarised cells such as neurons (Andreassi and Riccio, 2009), *iii)* mRNA stabilisation through adenylate-uridylate-rich elements (AU-rich elements, AREs; repetitive stretches of 'AUUUA' nucleotide sequences) which are bound by ARE-binding proteins that promote mRNA decay (Matoulkova *et al.*, 2012; Szostak and Gebauer, 2013).

Moreover, 3'UTRs play a role in translation by interacting with the translation machinery in complex ways, such as forming mRNA pseudo-circularization. Since many RBPs also interfere with the translational process, such as the closed-loop formation and ribosome recruitment, 3'UTRs could determine many other features of translation by associating with RBPs (Szostak and Gebauer, 2013). Interestingly, it was found that under certain conditions, a protein complex that normally behaves as a translation repressor can become a translation activator, by changing its composition or post-translational modifications (Szostak and Gebauer, 2013). This is exemplified by the 4EHP protein, which changes its behaviour upon hypoxia (Uniacke *et al.*, 2012).

1.3.3. Formation of co-translational 3’UTR-protein complexes

In 2015, Berkovits and Mayr (Berkovits and Mayr, 2015) have shown that 3’UTRs can also regulate protein localisation independently of the mRNA localisation, through the formation of protein complexes mediated by the 3’UTR. This has been thoroughly demonstrated for the cell-surface CD47 protein, an ubiquitously expressed protein involved in several processes such as apoptosis, adhesion and phagocytosis (Figure 1.11). The CD47 gene produces two 3’UTR isoforms through APA, a short 3’UTR and a long 3’UTR, both producing proteins with the same amino acid sequence. However, whereas the CD47 protein translated from the short 3’UTR is retained in the endoplasmic reticulum (ER), the protein translated from the long 3’UTR localises to the cell-surface. This occurs because the long 3’UTR, but not the short 3’UTR, is bound by the HuR RBP during translation. Subsequently, the SET protein is recruited to the site of translation and binds the nascent peptide chain of the CD47. Through additional interaction with the RAC1 protein, the CD47 protein – newly translated from the long 3’UTR isoform – is translocated to the cell-surface (Berkovits and Mayr, 2015). This mechanism was termed as 3′UTR-dependent protein localisation (UDPL), and evidences the ability of 3’UTRs to function as a scaffold for several proteins. Importantly, other genes encoding for plasma membrane proteins such as the CD44, Integrin α1 and BAFF-R genes, also had their cell-surface localisation affected by their 3’UTRs, suggesting this mechanism could be more widespread (Berkovits and Mayr, 2015; Mayr, 2017).

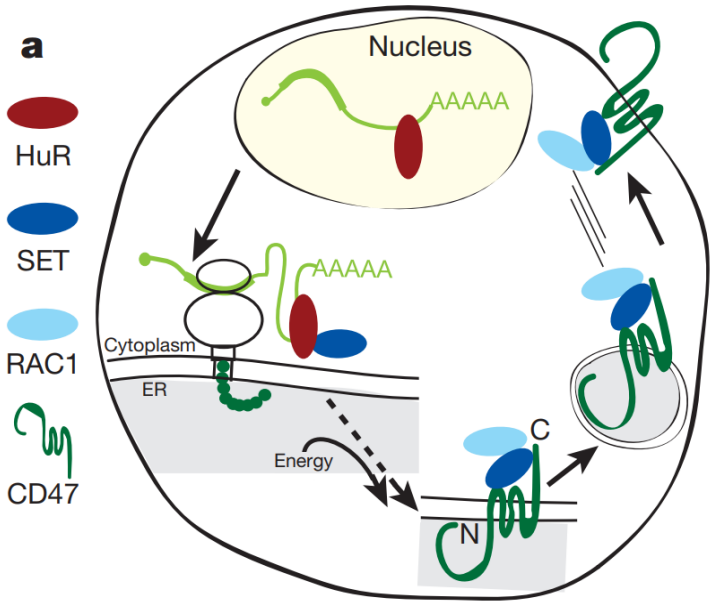


Figure 1.11 | Model of the 3’UTR-dependent protein localization (UDPL) mechanism. Figure adapted from (Berkovits and Mayr, 2015).

Clues of the importance of the UDPL process comes from its regulation by other proteins such as the HNRNPC RBP, whose 3'UTR binding sites were found to be correlated with HuR binding sites, indicating HNRNPC as a potential upstream regulator of UDPL (Gruber *et al.*, 2016). Moreover, other cases of 3'UTRs recruiting proteins co-translationally have also been described, such as the co-translational signal recognition particle (SRP) recruitment by 3'UTRs (Chartron, Hunt and Frydman, 2016). Furthermore, studies in *Schizosaccharomyces pombe* and *Saccharomyces cerevisiae* show that cotranslational formation of protein-protein interactions may be a widespread phenomenon (Duncan and Mata, 2011; Shiber *et al.*, 2018). However, whether the UDPL mechanism is a prevalent mechanism in the cell is yet unknown.

Mechanisms such as the UDPL, in which a 3'UTR facilitates the formation of protein complexes leading to a change in the protein function, could be envisaged as a way for an organism to diversify its proteome function, without recourse to amino acid changes (Mayr, 2016). Interestingly, depending on the cell type and cellular conditions, each alternative 3'UTR isoform could have its own RBP composition, and since an RBP can interact with several other proteins, each 3'UTR could be processed differently and serve a different function. Therefore, it can be thought that, besides affecting the cellular localisation of proteins, the 3'UTR formation of protein complexes could serve to regulate other processes. Indeed, it has been proposed that alternative 3'UTRs could play a role in mediating the multifunctionality of proteins (Mayr, 2017).

1.4. Moonlighting proteins and multifunctionality

Constructing a complex organism does not require a large number of genes. Rather, organism complexity is provided by the ensemble of all available functions and their regulation. Mechanisms such as alternative splicing and RNA editing of transcripts contribute to protein diversity and can increase the number of functions a certain gene can provide. However, using the same exact synthesised protein for several biological functions, allows cells to make more using less resources. Moonlighting proteins are a subset of multifunctional proteins that perform multiple unrelated functions. The multiple functions of moonlighting proteins can add another dimension to cellular complexity and provide a way to coordinate several cellular activities. However, the full prevalence of moonlighting proteins is yet unknown, as these have been found by serendipity in most cases. Large-scale bioinformatic approaches were employed to predict moonlighting proteins and analyse them as a group, discovering features such as a frequent association to disease and higher interaction propensity, thus giving insights into their cellular importance. The several functions of moonlighting proteins can be regulated in space and time in many different ways, such as by changing their cellular location, yet, the exact mechanisms used in this type of regulation are still unknown. This chapter describes the definition and prevalence of moonlighting proteins, as well as the manners in which this interesting set of proteins is found to be regulated.

1.4.1. Moonlighting proteins: definition, function and prevalence

Definitions of moonlighting

Before the structure of DNA was identified, it was thought that each gene would encode a single protein, and that each protein would have a single function (Copley, 2012). However, it is now known that a single gene can encode different proteins (e.g. through alternative splicing or APA). Likewise, with the discovery of moonlighting proteins, it is now known that a single protein can serve multiple unrelated functions, using the same polypeptide sequence (Piatigorsky and Wistow, 1989). The term ‘moonlighting proteins’ was coined by Constance Jeffrey in 1999 (Jeffery, 1999), in analogy to moonlighting in the sense of holding a second job in addition to a regular one. Under a strict definition, moonlighting proteins perform multiple functions “without partitioning these functions into different protein domains” (Huberts and van der Klei, 2010). Moreover, this moonlighting protein definition excludes cases “where the two functions are the result of gene fusions, families of homologous proteins, splice variants, or promiscuous enzyme activities” (Jeffery, 2009). However, less stringent definitions of moonlighting proteins have been

used, sometimes adopting terms such as ‘multitask proteins’ and ‘extreme multifunctional proteins’, focusing on the fact that the proteins are indeed involved in diverse unrelated functions, regardless of their evolutionary history or domain organisation (Chapple and Brun, 2015; Franco-Serrano *et al.*, 2018). In this view, moonlighting functions may be performed by any region of the protein surface, often involving regions other than the one responsible for the canonical function (Copley, 2012).

Importantly, regardless of the exact definition used, protein moonlighting should not be confused with protein multifunctionality. While moonlighting proteins are a subset of multifunctional proteins, multifunctionality can arrive from performing the same action under different contexts, such as acting in distinct pathways or cell lines. For instance, proteins involved in signalling or transcription (e.g. a transcription factor) may have pleiotropic functions because they regulate a large number of processes, but this is not considered moonlighting since their mode of action is the same. In addition, protein moonlighting should be clearly distinguished from gene multifunctionality, which also includes genes that encode distinct protein isoforms that have different functions from each other (van de Peppel and Holstege, 2005).

One of the first proteins discovered to be a moonlighting protein, even before this term was coined, was the mammalian P8 protein (Henderson and Martin, 2014). A first study demonstrated the ability of P8 to bind single-stranded DNA (Tsai and Green, 1973). A few years later, another study discovered that P8 was the glyceraldehyde-3-phosphate dehydrogenase (GAPDH), a key enzyme in glycolysis (Perucho, Salas and Salas, 1977). Notably, this protein continues to be an archetypical example of protein moonlighting, as more recent work on GAPDH continues to find novel moonlighting functions for this protein, including cell signaling, tRNA export and intracellular membrane trafficking in eukaryotes, among other functions (Sirover, 2011; Tristan *et al.*, 2011).

In the 1980s, Piatigorsky & Wistow described other examples of moonlighting proteins, by reporting that the duck ε-crystallin, a structural protein in the eye of vertebrates, was identical and indeed the same as the lactate dehydrogenase B4 protein, a highly conserved glycolytic enzyme (Wistow and Piatigorsky, 1987). Later, the authors shown that in other vertebrate species crystallin proteins turned out to perform other functions, such as the turtle τ-crystallin identified also as the glycolytic enzyme α-enolase, and the *Schistosoma mansoni* α-crystallin identified as the egg antigen p40 (Piatigorsky and Wistow, 1989).

The recruitment of ancient enzymes to function as a crystallins represent interesting examples in which the same moonlighting function is provided by different proteins in distinct lineages of species (Copley,

2012). Conversely, other moonlighting proteins, including GAPDH in bacteria, seem to hold a propensity to gain novel distinct functions in different lineages (Copley, 2012).

Prevalence and functions of moonlighting proteins

Since the original discoveries of moonlighting proteins, a few hundred other cases have been found throughout the evolutionary tree, from prokaryotes to eukaryotes, as well as in viruses (Jeffery, 2018). Overall, moonlighting enzymes perform a large variety of additional functions, for example acting as enzymes, cytokines, protein and RNA chaperones, transcription and translation factors, DNA stabilisers, components of the cytoskeleton, proteasome subunits, receptors and transmembrane channels, among many other functions (Jeffery, 2014; Wang and Jeffery, 2016; Lu and Hunter, 2018). However, about two-thirds of the known moonlighting proteins have an enzymatic function as one of their functions. In fact, many of these moonlighting enzymes are highly conserved ancient enzymes, many involved in the sugar metabolism. For example, 7 out of 10 and 7 out of 8 proteins of the glycolytic pathway and the tricarboxylic acid (TCA) cycle, respectively, are known to moonlight (Huberts and van der Klei, 2010). However, the fact that most moonlighting proteins are found to be enzymes may be due to the way moonlighting proteins were discovered, since protein functions have often been detected through enzyme activity assays, and proteins of the glycolysis and TCA cycle pathways are some of the best characterized proteins.

Many moonlighting proteins play a central role in many diseases, such as cancer, autoimmune disease, heart disease, obesity and diabetes (Jeffery, 2018). Moreover, it has been suggested that disease comorbidities (i.e. several diseases co-occurring in a same individual) with different phenotypes, could be caused by proteins involved in several processes, such as moonlighting proteins (Zanzoni, Chapple and Brun, 2015). Indeed, moonlighting proteins could also cause unforeseen drug side-effects due to interferences with several unexpected biological processes. Comprehensive knowledge of moonlighting proteins would thus aid the development of treatments, in order to avoid targeting a function that is not involved in the disease (Jeffery, 2018). Moreover, moonlighting proteins can also confound genome and protein annotations, which often use sequence homology to attribute function to newly found sequences (Jeffery, 2014).

1.4.2. Resources and detection of moonlighting and multifunctional proteins

Resources and experimental approaches to detect moonlighting proteins

Moonlighting functions can be revealed experimentally when mutation or deletion studies result in unexpected phenotypes. However, unequivocal identification of moonlighting proteins requires several mutational studies, in order to show that some mutations affect both functions of the protein, while others affect only one of the functions (Gancedo, Flores and Gancedo, 2016). Nevertheless, such analysis would only work if the two functions use different parts of the protein. Finally, providing evidence that the several functions are indeed completely unrelated, and not due to pleiotropic effects, can be difficult.

In a recent study, Espinosa-Cantú *et al.* attempted to experimentally address the prevalence of enzymes with moonlighting functions in *Saccharomyces cerevisiae*. They did this by evaluating if the enzyme gene deletion phenotypes are caused solely by the loss of catalytic activity, or if the phenotypes (e.g. cell growth) are due to a yet unknown moonlighting function independent of the catalytic activity (Espinosa-Cantú *et al.*, 2018). For this, 11 enzymes associated to amino acid biosynthesis were chosen, since their known function could be readily confirmed. This study shown that 4 out of 11 tested enzymes may have moonlighting functions, thus suggesting that moonlighting proteins may be highly prevalent, at least in this set of enzymes. Moreover, the number of moonlighting proteins identified here may even be an underestimate, since this experimental approach, performed under constant conditions, is unable to find cases of moonlighting proteins that only display alternative functions when present in different contexts (see below, Introduction 1.4.3).

Over the years, hundreds of moonlighting proteins have been identified experimentally. Two public databases contain multi-species collections of moonlighting proteins described in the literature. These are the MoonProt database (Chen *et al.*, 2018), containing more than 350 proteins pertaining to the strict definition of moonlighting proteins (as described above), and the MultitaskProtDB (Franco-Serrano *et al.*, 2018), a database containing more than 650 proteins, not necessarily sticking to the strict definition of moonlighting proteins.

Computational approaches to detect moonlighting and multifunctional proteins

Even though the number of experimentally determined moonlighting proteins is increasing, their full prevalence is unknown. Indeed, the discovery of moonlighting functions has been largely serendipitous, by discovering that two proteins known to serve distinct functions are in fact the very same protein

(Copley, 2012). Consequently, the number of known multifunctional proteins is probably underestimated. A few large-scale computational methods to predict moonlighting proteins have been developed and predicted hundreds to thousands of such proteins. These used indicators and approaches such as sequence similarity (Khan *et al.*, 2012), text mining (Khan, Bhuiyan and Kihara, 2017), and machine learning classifiers using these and other features, such as gene expression and structural disorder (Khan and Kihara, 2016). However, these methods may have shortcomings. For example, moonlighting may not be readily identified through sequence analysis, since multifunctionality may blur the results of sequence similarity searches (Chapple *et al.*, 2015). Moreover, gene expression analysis are unable to find moonlighting proteins whose function differs upon their cellular export, such as the human RHAMM protein (also known as HMMR) (Maxwell, McCarthy and Turley, 2008).

In 2015, Chapple *et al.* 2015 developed an innovative approach which for the first time combined protein-protein interaction network analysis and Gene Ontology (GO) functional annotations, to predict "extreme multifunctional" (EMF) proteins at a proteome scale (Chapple *et al.*, 2015). Since multifunctional proteins interact with different sets of proteins to perform their different cellular functions, the usage of the PPI network topology for their identification is pertinent (Becker *et al.*, 2012). In addition, Chapple *et al.* 2015 used novel methods to determine functions that are highly dissimilar to each other, a hallmark of protein moonlighting. The 430 human EMF proteins predicted were defined as proteins "whose multiple functions are very dissimilar to one another" (Chapple and Brun, 2015), thus related to moonlighting proteins, but not constricted by the strict definition of moonlighting proteins which is also concerned by the evolutionary history of the protein. Notably, Chapple *et al.* 2015 shown that EMF proteins possess characteristics that set them apart from other proteins. Within a protein interactome, a typical EMF protein is more likely to have a high number of protein partners and to be central to the network. Moreover, EMF proteins were found more likely to be expressed ubiquitously, suggesting that they can perform alternative functions in different tissues (Chapple *et al.*, 2015). Furthermore, EMF proteins were shown to contain more short linear motifs (SLiMs), which may be responsible for transient interactions with other molecules (Perkins *et al.*, 2010). Indeed, EMF proteins were found to be enriched in SLiMs that are regulated by pre- and post-translational switch mechanisms, collected from the *Switch.ELM* resource (Van Roey *et al.*, 2013). This finding suggested such SLiMs could provide the ability for moonlighting proteins to switch between functions.

1.4.3. Regulation of moonlighting protein multifunctionality

It has been proposed that moonlighting proteins can coordinate several cellular activities, serving as switches between pathways and helping to respond to changes in the cellular environment (Jeffery, 1999, 2015). Therefore, regulation of the multiple protein activities, in space and time, is likely to be important for the homeostasis of biological systems. Some moonlighting proteins perform its multiple functions simultaneously, and each of these functions may be independently regulated, while other moonlighting proteins alternate between functions due to certain triggers. In many cases, the switch in function involves the binding of another molecule, such as another protein (Jeffery, 1999, 2018). The switch of the moonlighting protein's functions can be triggered or regulated by several distinct factors, sometimes in combination to each other (Jeffery, 1999, 2018). These include:

- Post-translational modifications (PTMs): these involve dozens of different modifications and molecular additions to the amino acid chains of a protein, some of the most common being phosphorylation, acetylation and glycosylation. PTMs are generally used to regulate protein function, by affecting several properties of the protein, such as their activity state (on/off), structural conformation, cellular localisation and interaction partners (Beltrao *et al.*, 2013).

- *Oligomeric state*: several known moonlighting proteins were found to function differently depending on their oligomeric states. For example, the glyceraldehyde-3-phosphate dehydrogenase acts as a glycolytic enzyme as a tetramer, but functions as a nuclear uracil-DNA glycosylase when in a monomeric form (Meyer-Siegler *et al.*, 1991).

- *Cellular concentration of other molecules*: the function of moonlighting proteins can be switched by the concentration of substrate, ligand or cofactor available. For instance, the human aconitase, an enzyme of the TCA cycle, changes function according to the cellular iron concentration. When cellular iron concentration is low, the aconitase loses its interaction with a Fe-S cluster, changes its conformation and is able to bind iron responsive elements (IREs) in the mRNAs of iron metabolism genes, thus modulating their translation (Volz, 2008).

- *Cell type or tissue expression*: a moonlighting protein can perform distinct functions when expressed in different cell types or tissues, where they encounter a different cellular environment (e.g. tissue-specific proteins, different molecular concentrations). A prime example of this are the proteins moonlighting as crystallins. These act as structural proteins in the eye lens of vertebrates

when highly-expressed in this tissue, but act either as a glycolytic enzyme (e.g. lactate dehydrogenase B4 protein in ducks and α-enolase in turtles) or a egg antigen (in *Schistosoma mansoni)* when present at lower-levels in non-lens tissues (Wistow and Piatigorsky, 1988).

- *Cellular localisation*: the presence of a moonlighting protein in different subcellular (e.g. cytoplasm, nucleus) or extracellular locations (e.g. secreted, extracellular matrix) can also be responsible for a change in function (Yoon, Ryu and Baek, 2018). For example, the *Escherichia coli* PutA protein acts a dehydrogenase when in the plasma membrane, but binds DNA when present in the bacteria's cytoplasm, regulating the *put* operon (Ostrovsky de Spicer and Maloy, 1993). Another example is the human RHAMM protein, which acts as a centrosomal or mitotic-spindle protein in the cytoplasm of normal cells, but at the plasma membrane of tumour cells, binds hyaluronan (hyaluronic acid), which triggers CD44 activation and the regulation of signaling cascades (Maxwell, McCarthy and Turley, 2008).

It has been proposed that "finding a protein in an unexpected location provides a clue that a moonlighting function may exist" (Copley, 2012). Indeed, nuclear or cytoplasmic moonlighting proteins are often used as secreted molecules involved in signalling (Jeffery, 1999; Yoon, Ryu and Baek, 2018). For example, the phosphoglucose isomerase participates in glycolysis in the cytosol, but it can also be secreted as neuroleukin and act as a nerve growth factor, as well as a cytokine involved in B cell maturation (Bonini *et al.*, 2003). Notably, the system or mechanism by which most cytoplasmic moonlighting proteins are secreted is not known (Kainulainen and Korhonen, 2014; Jeffery, 2018). Moreover, several 'housekeeping proteins' such as intracellular chaperones and enzymes in glycolysis and citric acid cycle pathways also function as cell surface receptors or secreted molecules (Kainulainen and Korhonen, 2014; Amblee and Jeffery, 2015; Jeffery, 2018; Yoon, Ryu and Baek, 2018). Interestingly, in 2015, Amblee & Jeffery performed an analysis on 30 distinct prokaryotic and eukaryotic moonlighting proteins, thought to have different functions intracellularly and on the cell surface (Amblee and Jeffery, 2015). They found that none of these proteins contained any signal or motif for cell surface targeting, such as an N-terminal signal peptide or a LPXTG motif, suggesting that their cellular localisation may be regulated by a yet unknown mechanism (Amblee and Jeffery, 2015).

Overall, even though several triggers and regulatory mechanisms for the switch between functions of moonlighting proteins have been found, such as the ones described in this section, for many moonlighting proteins, it remains to be determined what triggers and regulates their switch of functions.

2. Results

2.1. Protein complex scaffolding predicted as a prevalent function of long non-coding RNAs

The biological function of most long non-coding RNAs (lncRNAs), if any, is yet unknown. A few lncRNAs, such as the TERC, NEAT1 and XIST, show the ability to gather several protein components, i.e. to act as protein scaffolds (Introduction, section 1.2.3). Protein scaffolding has since been included as one of the dozen potential functions or mechanisms of action attributed to lncRNAs. However, it remains to be demonstrated whether this mechanism is common in the cell, since prior to this work this mechanism has not been investigated systematically.

Several studies have attempted to systematically assess the functionality of lncRNAs. The first experimental study of large-scale functionality of lncRNAs was the study by Liu *et al.* (Liu *et al.*, 2017). In this study CRISPR interference was applied in human cell lines to repress the expression of more than 16,000 lncRNA loci while measuring a single phenotype, cell growth, which was found affected by 499 lncRNA loci (Liu *et al.*, 2017). Other large-scale works studying function of lncRNAs are largely based on the analysis of lncRNA co-expression with protein-coding genes (Park *et al.*, 2014; Jiang *et al.*, 2015). However, in all of these studies, the mechanism of action of the lncRNAs is not elucidated. Moreover, it is not clear whether the function of the lncRNA loci can be attributed to the a function of the lncRNA molecule itself, independently of their transcription event.

Predicting lncRNA function through their interactions with proteins provides a way to determine their function as a molecule. However, research through such approaches has been hampered by the fact that experimental datasets of protein-RNA interactions are not yet applicable to genome-wide approaches (Introduction, section 1.1.4), and focus mostly on mRNAs, rather than lncRNAs. Several computational methods to predict protein-RNA interactions exist, but most are unsuitable for large-scale analysis, due to large computation times. The catRAPID *omics* software was developed by Gian Gaetano Tartaglia's group, in order to predict genome-wide protein-RNA interactions within a reasonable time frame (Agostini *et al.*, 2013).

In our work, we propose the first computational approach to systematically study the potential function of lncRNAs as protein scaffolding molecules. As a first step, we generated the largest protein-RNA interaction network ever made (more than 6 millions interactions) using catRAPID *omics*. Subsequently, we used an original computational and statistical approach that integrates and analyses protein-lncRNA interaction predictions and protein complexes. This enabled us to *(i)* identify 847 human lncRNAs (~5% of the transcriptome used) as possible scaffolding molecules for known protein complexes, *(ii)* predict that half of the human protein complexes may be scaffolded by lncRNAs and *(iii)* propose that the mechanism of action of several lncRNAs known to be associated to disease involves the scaffolding of specific protein complexes. This work thus predicts that protein scaffolding is a prevalent function of lncRNAs and provides a dataset of thousands of lncRNA-protein-complex combinations involved in this mechanism, publicly available to the community. Moreover, this study suggests that a substantial proportion of lncRNA transcripts play a role in the cell, in a large variety of biological processes.

Ribeiro, DM, Zanzoni, A, Cipriano, A, Delli Ponti, R, Spinelli, L, Ballarino, M, Bozzoni, I, Tartaglia, GG and Brun, C (2018) Protein complex scaffolding predicted as a prevalent function of long non-coding RNAs. *Nucleic acids Research*, 46, 917–928.

Supplementary material is available on the *Appendix I*

Published online 20 November 2017

Nucleic Acids Research, 2018, Vol. 46, No. 2 ***917–928***
doi: 10.1093/nar/gkx1169

Protein complex scaffolding predicted as a prevalent function of long non-coding RNAs

Diogo M. Ribeiro[1], Andreas Zanzoni[1], Andrea Cipriano[2], Riccardo Delli Ponti[3,4], Lionel Spinelli[1], Monica Ballarino[2], Irene Bozzoni[2], Gian Gaetano Tartaglia[3,4,5,*] and Christine Brun[1,6,*]

[1]Aix-Marseille Université, Inserm, TAGC UMR_S1090, Marseille, France, [2]Dept. of Biology and Biotechnology Charles Darwin, Sapienza University, Rome, Italy, [3]Centre for Genomic Regulation (CRG), The Barcelona Institute of Science and Technology, Dr Aiguader 88, 08003 Barcelona, Spain, [4]Universitat Pompeu Fabra (UPF), 08003 Barcelona, Spain, [5]Institucio Catalana de Recerca i Estudis Avançats (ICREA), 23 Passeig Lluıs Companys, 08010 Barcelona, Spain and [6]CNRS, Marseille, France

Received October 06, 2017; Revised November 03, 2017; Editorial Decision November 07, 2017; Accepted November 07, 2017

ABSTRACT

The human transcriptome contains thousands of long non-coding RNAs (lncRNAs). Characterizing their function is a current challenge. An emerging concept is that lncRNAs serve as protein scaffolds, forming ribonucleoproteins and bringing proteins in proximity. However, only few scaffolding lncRNAs have been characterized and the prevalence of this function is unknown. Here, we propose the first computational approach aimed at predicting scaffolding lncRNAs at large scale. We predicted the largest human lncRNA–protein interaction network to date using the *cat*RAPID *omics* algorithm. In combination with tissue expression and statistical approaches, we identified 847 lncRNAs (∼5% of the long non-coding transcriptome) predicted to scaffold half of the known protein complexes and network modules. Lastly, we show that the association of certain lncRNAs to disease may involve their scaffolding ability. Overall, our results suggest for the first time that RNA-mediated scaffolding of protein complexes and modules may be a common mechanism in human cells.

INTRODUCTION

More than 60% of the human genome is transcribed into tens of thousands of RNAs with low coding potential (1). Long non-coding RNAs (lncRNAs) are a subset of those transcripts longer than 200 nt, transcribed by RNA polymerase II, often capped, spliced and polyadenylated (2). The possible function of most of the > 26 000 GENCODE annotated lncRNAs is yet to be addressed (3), and many are thought to be transcription errors or noise. However, thousands of lncRNAs have been found to be differentially expressed in distinct cell types, with dozens shown to be implicated in transcription regulation (4), stress responses (5) and disease (6). Indeed, lncRNAs are versatile molecules able to perform numerous tasks in the cell through binding of proteins, DNA or other RNA molecules (2).

All cellular functions are performed by interactions between molecules, such as interaction between proteins and RNAs. These interactions can be stable, leading to ribonucleoprotein (RNP) complexes such as the ribosome, the spliceosome or the telomerase complex, or transient such as those involved in transport and degradation of nuclear transcripts. Similarly, components of complexes or pathways need to be physically close to each other (either transiently or permanently) in order to perform their function. One way to achieve this, while attaining selectivity in a crowded cell, is to employ platform or scaffold molecules that piece together components of a complex or a pathway (7). Although proteins can and do serve as scaffolds for other proteins (8), the use of RNA scaffolds would present several advantages, since 'one protein comprising 100 amino acids can capture only one or two proteins, whereas one RNA molecule comprising 100 nt can capture around 5–20 proteins', simultaneously (9). Moreover, lncRNAs can act immediately after transcription, while protein scaffolds require at least the step of translation before being functional (2).

Several ncRNAs have been found to function as scaffolds for RNP complexes such as TERC (Telomerase RNA Component), SRP (Signal Recognition Particle RNA) and LINP1 (LncRNA In Nonhomologous End Joining Pathway 1) (2,10,11) or found to transiently assemble groups of proteins as in the case of XIST (X-inactive specific transcript) and both the granule-forming NEAT1 (Nuclear Paraspeckle Assembly Transcript 1) and MALAT1

*To whom correspondence should be addressed. Tel: +33 491828712; Email: christine-g.brun@inserm.fr
Correspondence may also be addressed to Gian Gaetano Tartaglia. Tel: +34 933160116; Email: gian.tartaglia@crg.eu

(Metastasis Associated Lung Adenocarcinoma Transcript 1) (5,12). Although known scaffolding lncRNAs carry out important cellular functions, only a few dozen cases have been uncovered so far (7), many while studying the protein complexes rather than the lncRNAs. We therefore hypothesize that other yet uncharacterized lncRNAs may act as scaffolds.

Recently, with the development of RNA interactome capture methodologies, the repertoire of RNA-binding proteins (RBPs) has greatly expanded (13), leading to the discovery of hundreds of novel RNA-interacting proteins, many of which contain no known RNA-binding domain (RBD). In addition, studies using high-throughput methods to detect RNAs bound by RBPs including iCLIP, PAR-CLIP and recently eCLIP (14), demonstrate that most RBPs bind thousands of different RNA molecules depending on the cell line. However, these investigations have been limited to a set of ∼140 RBPs containing known RBDs (14,15) and do not cover the full extent of the protein–RNA interaction space. Furthermore, only one fraction of the RNAs targeted by the RBPs are found in common by independent replicate experiments, suggesting that the interaction maps of the studied RBPs are far from complete (14). Computational prediction of protein–RNA interactions can therefore help fill the gap in our knowledge of protein–RNA interactions and be applied to large-scale analyses.

In this paper, we study for the first time the prevalence of protein complex scaffolding as a function of lncRNAs. By exploiting a computed protein–RNA interaction network, we developed and applied an original large-scale approach to identify candidate lncRNAs possibly acting as scaffolding molecules for protein complexes and network functional modules. We discovered hundreds of scaffolding lncRNA candidates, suggesting that RNA scaffolding is a prevalent and widespread mechanism in the cell. In addition, we found that more than half of the protein complexes and network modules in the cell may be scaffolded by lncRNAs, reinforcing the widespread nature of their action.

MATERIALS AND METHODS

LncRNA–protein interaction predictions

The *cat*RAPID *omics* protein–RNA interaction predictor (16) was used to predict interactions between the human long non-coding RNA transcriptome (Ensembl v82) and the human canonical proteome, leading to ∼243 million predictions. Predictions with interaction propensity score ≥50 were kept for further analyses (∼30.8 million interactions). See Supplementary Material for details.

Tissue expression filtering

To create a set of high confidence protein–RNA interaction predictions, we restricted the analysis to pairs of lncRNA–proteins that are likely to be found together in at least one tissue. Human tissue expression data from the GTEx v6.0 project (17) was used. We downloaded RPKM (Reads Per Kilobase of transcript per Million mapped reads) information from 8555 samples across 53 tissues, already mapped to human transcripts (GENCODE v19). RPKM values of samples coming from the same tissue were averaged after a step of removing outlier values (below or above 1.5-times the interquartile range). Protein expression was derived from their coding mRNA expression, by selecting the highest RPKM value among the protein's mRNAs for each tissue. Only protein–RNA interactions where both the RNA and the protein have a minimum RPKM value of 1.58 in at least one of the 53 tissues, were retained. This cutoff was determined as the optimal expression cutoff (maximizing the sum of specificity and sensitivity) in a ROC curve experiment between the pre-filtering lncRNA–protein interaction prediction dataset (∼243 million interaction predictions) and a set of 2438 experimentally detected CLIP interactions taken from StarBase v2.0 (18) with at least 100 mapped reads (area under the ROC = 0.71). The expression metric used ('paired expression') was calculated for each protein–RNA pair as the lowest RPKM expression between the protein and RNA for each tissue, to which the maximum RPKM value among tissues for that protein–RNA pair is then withdrawn, i.e.

$$E\ (\text{Protein},\ \text{RNA}) = \max_{t\ \in\ \text{tissues}} \left(\min\left(E_t\,(\text{Protein}),\ E_t\,(\text{RNA})\right)\right)$$

where E(Protein, RNA) denotes the 'paired expression' for each protein–RNA pair and E_t denotes the RPKM expression in tissue t (RPKM values were $\log_{10}$-transformed).

Protein complex and network module datasets

We collected protein complex information from the (i) BioPlex publication (19) Supplementary Table S3, which includes 354 complexes; (ii) list of conserved protein complexes from Wan *et al.* (20), Supplementary Table S4, which includes 981 complexes; (iii) list of non-redundant CORUM (21) complexes from Havugimana *et al.* (22), Supplementary Table S3, which includes 324 complexes, referred to as 'non-redundant CORUM complexes'. Protein network modules were extracted from a human interactome as described in (23). See Supplementary Material for details.

LncRNA–protein complex enrichment analysis

Using the set of predicted interactions between lncRNA and proteins filtered by (i) interaction propensity and (ii) minimum RPKM expression, we performed the following enrichment analysis: for each lncRNA and protein group, we assessed the enrichment of the lncRNA's interacting-proteins among the proteins in the group using a hypergeometric test (one-tailed test; FDR = 5%, multiple test corrected with the Benjamini–Hochberg procedure; Figure 1B), using as background the set of proteins in complexes or modules retaining at least one interaction after interaction filters. We considered only enrichments where: (i) the lncRNA is interacting with at least two proteins of the protein group and (ii) all the proteins in the complex or the module are expressed in a same tissue as the lncRNA with at least 1.58 RPKM. To exclude lncRNAs with high background levels of enrichments, we built a null hypothesis distribution by performing 10 000 hypergeometric tests for each lncRNA, each time randomly shuffling the proteins labels between the protein groups. We excluded lncRNAs

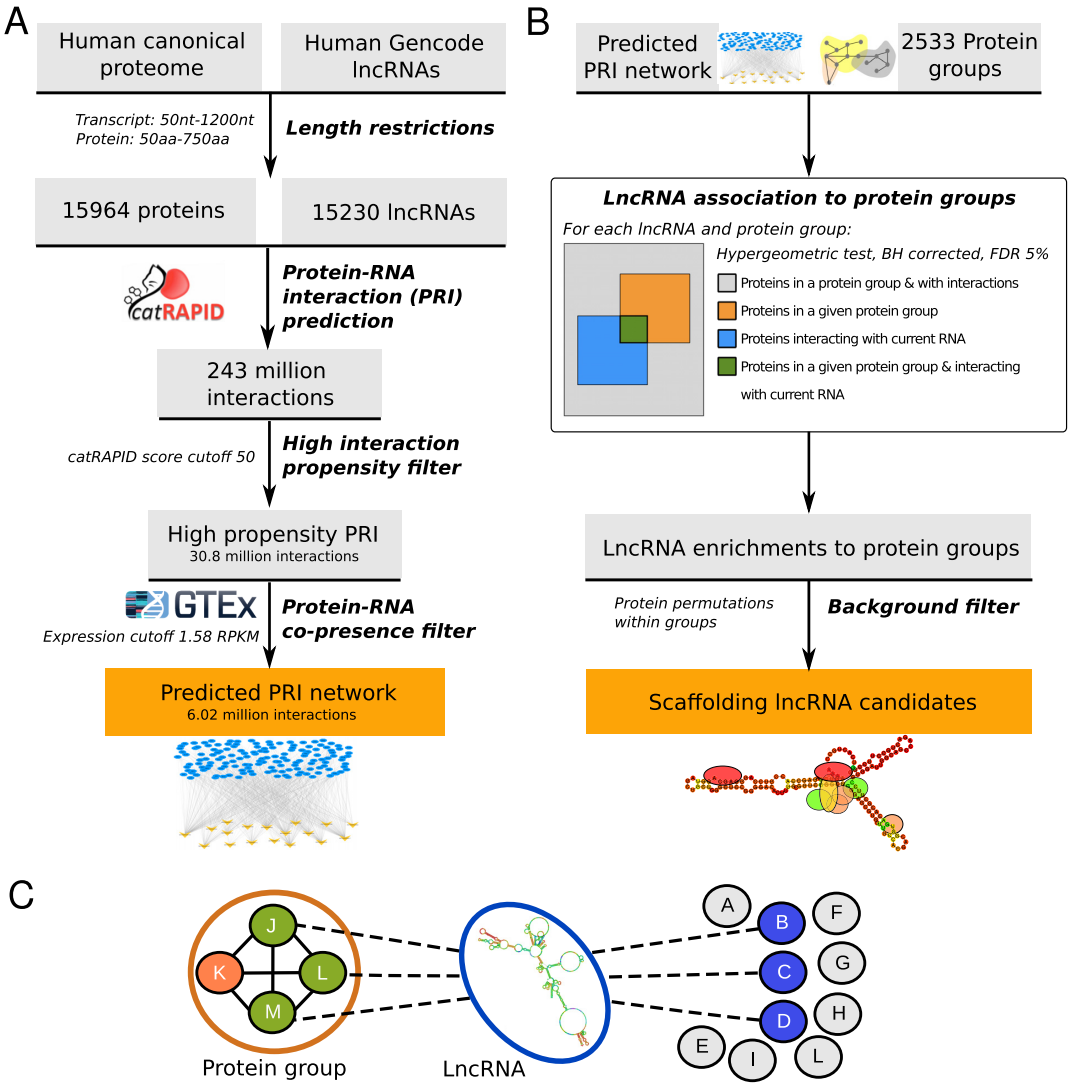


Figure 1. Data production and analysis workflows. (**A**) Predictions of protein-lncRNA interactions (PRI) using *cat*RAPID *omics* for the human proteome and long non-coding transcriptome. Interactions are further filtered by co-presence in the same GTEx tissue. The produced PRI network contains 6.02 million interactions. (**B**) Protein groups and lncRNAs are tested for enrichment in lncRNA protein's targets among groups of proteins. After noise filtering, a final list of scaffolding lncRNA candidates is produced. (**C**) Principle of the enrichment in lncRNA protein's targets among groups of proteins. Colors of nodes correspond to the ones used on the lncRNA association to protein groups box on (B).

with (i) enrichments not significant in respect to the null hypothesis (empirical P-value > 0.01); (ii) an enrichment ratio lower than 2-fold.

RESULTS

A predicted human interaction network between the non-coding transcriptome and the proteome

Aiming to extensively identify lncRNA molecules interacting with protein complexes and potentially acting as protein scaffolds, we first computed the protein–RNA interaction potential between most of the human proteome and the long non-coding transcriptome (79% and 81%, respectively; Supplementary Material) using the *cat*RAPID *omics* algorithm (16) (Figure 1A). The *cat*RAPID algorithm is a protein–RNA interaction predictor based on the physicochemical features of the molecules that has been extensively used and tested on lncRNAs with good performances (16,24,25). With this method we produced 243 million predicted interactions, of which 30.8 million display high interaction propensity scores (*cat*RAPID score $\geq$ 50). Since many lncRNAs have only been found to be expressed at very low levels and often in a tissue-specific manner (26), we only retained 6.02 million protein–lncRNA interactions between molecules co-present in at least one out of the 53 human tissues from the GTEx RNA-seq dataset (17) (see Materials and Methods). Globally, the 6.02 million predicted interactions occur between 12629 proteins and 2799 lncRNAs (Figure 2), i.e. between 80% of the tested proteins and 18% of our initial set of lncRNAs. Individual proteins are predicted to interact with up to 2.5% of the lncRNAs on average (Supplementary Figure S1). When considering only RBPs (Supplementary Material), we predict them to interact with 4.14% of the lncRNAs on average, in the same

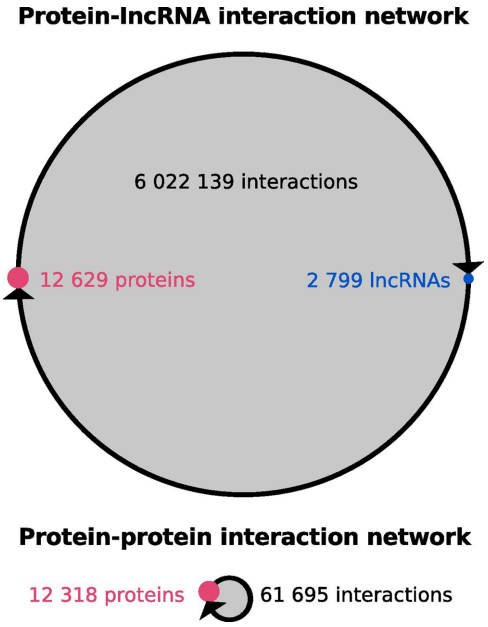


Figure 2. A global lncRNA–protein interaction network. Predicted protein-lncRNA interaction network composed by more than 6 millions interactions (grey circle) between 12629 proteins (pink circle) and 2799 lncRNAs (blue circle). The size of the network is compared to the human binary protein-protein interaction network (see Supplementary Methods). All circles are proportional to their components.

range as eCLIP results on 82 RBPs (14), which interact with 7.98% of the lncRNAs from the same dataset. On the lncRNA side, their median number of protein interactions is 1267 (Supplementary Figure S1), a higher number than suggested by current RNA pull-down studies that report between 126 and 852 interacting proteins per lncRNA (27,28).

As evident in recent high-throughput screenings, the complexity of biological systems challenges interpretation of experimental results due to the specific interactions occurring in different contexts (12,14). Yet our predictions, based on molecular physicochemical properties, represent a set of possible interactions between co-expressed proteins and RNA, independent of the cellular sub-localization and the cellular states. Our predictions therefore cover a larger spectrum of conditions in which protein–RNA interactions may occur, compared to the ones assessed in specific *in vivo* studies. This allows us to detect, for example, lncRNAs acting exclusively upon DNA damage and other stress conditions, or interactions restricted to a few cell types. Despite all this, 9414 of our predicted interactions are found in the relatively small set of eCLIP experiments, a highly significant overlap considering the 82 proteins and 7381 transcripts present in both eCLIP and *cat*RAPID datasets (P-value < 2.2e-271, OR = 1.85, two-tailed Fisher's exact test), therefore increasing our confidence in the predicted network.

Overall, to the best of our knowledge, we have predicted the largest human lncRNA–protein interaction network to date.

Interactions between lncRNAs and protein complexes or network modules

To assess our capacity to computationally predict lncRNA interactions with protein complexes, we studied the possible association between a recently discovered evolutionarily-conserved and muscle-restricted lncRNA, *lnc-405* (29), and the Puro–Purβ–YBX1 protein complex, implicated in gene regulation of muscle cells (30). The *cat*RAPID *omics* algorithm predicts the interaction of human *lnc-405* with Puro, Purβ and YBX1 with moderate to high scores (38.56, 44.05, 67.84, respectively).

To determine if *cat*RAPID correctly predicted the interactions of the lncRNA to the protein complex in a cellular context, we performed endogenous *lnc-405* RNA pull-down from nuclear extracts of C2C12 mouse myotubes followed by a mass spectrometry (MS) analysis. Murine cells were used since *lnc-405* is highly conserved in mouse and very abundant in differentiated C2C12 cells, allowing the easy production of the large amounts of nuclear extracts which are required for the pull-down. Efficient enrichment of *lnc-405* was detected in both odd and even RNA pull-down samples, while no recovery was observed with lacZ control (Supplementary Figure S2A).

Notably, MS analysis applied on the odd, even and lacZ (control) samples allowed the identification of 19 *lnc-405* interactors, including two components of the Puro-Purβ-YBX1 complex (Supplementary Table S1; Supplementary Material). RIP assays performed in mouse and human myotubes, using an antibody against Purβ, allowed to validate the specificity of the interaction with *lnc-405* and to confirm the evolutionary conservation of such interaction (Supplementary Figure S2B and C). Moreover, a GSEA experiment shows that the top interactors of *lnc-405* predicted by *cat*RAPID are enriched in proteins identified in the MS experiment (Figure 3, P-value = 0.017). These results remarkably show that *cat*RAPID is able to correctly predict interactions between lncRNAs and proteins (whether in a complex or not), in line with good *cat*RAPID performances observed for other ncRNAs and reported in previous articles (24,25,31).

We thus proceeded with the exploration of our *cat*RAPID predicted lncRNA–protein interaction network, aiming to test the hypothesis that lncRNAs frequently scaffold known protein complexes through protein–RNA interaction. For this, we investigated three public datasets of human macromolecular complexes. Briefly, we used the (i) non-redundant dataset of 326 CORUM complexes (21) collected by Havugimana *et al.* (22) (hereafter referred to as 'non-redundant CORUM'), (ii) a set of 981 metazoan-conserved complexes produced by Wan *et al.* (20) through biochemical fractionation with quantitative MS (hereafter referred to as 'Wan 2015'), as well as (iii) the BioPlex dataset (19) of 354 complexes detected through affinity purification, MS experiments and interaction network analysis. Moreover, the human cell contains groups of functionally-related proteins that interact more transiently but may nevertheless be assembled or gathered together by lncRNA scaffolds to participate in metabolic or signaling pathways. For these reasons, we also used a dataset of 874 functional modules identified

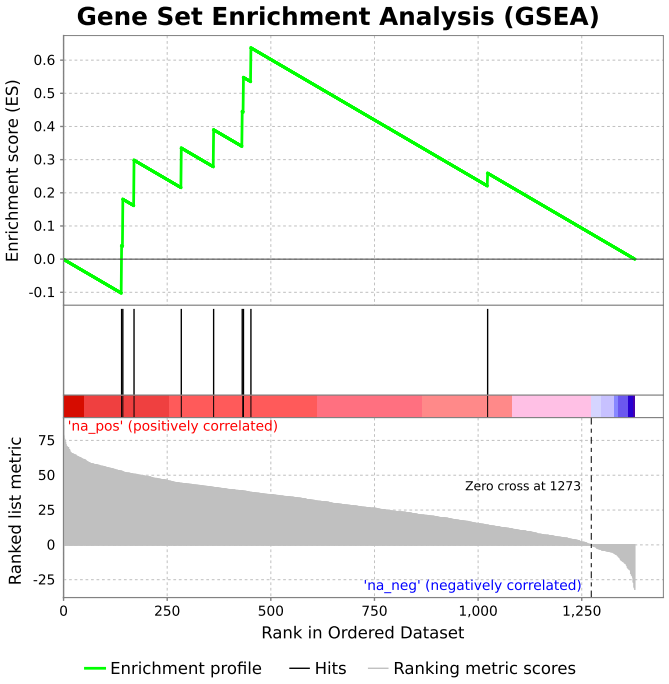


Figure 3. Experimentally-determined *lnc-405*-interacting proteins are enriched as top *cat*RAPID predictions. Gene Set Enrichment Analysis (GSEA) (69) of *cat*RAPID predictions between *lnc-405* lncRNA and 1459 human RBPs (Supplementary Material), using the RBPs identified as interactors in the MS experiment as a gene set. Note that only RBPs with *cat*RAPID predictions (within size restrictions) were considered. *P*-value = 0.017 (10 000 simulations), normalized enrichment score = 1.59).

in the human interactome using OCG, an algorithm that decomposes a network into overlapping modules, based on modularity optimization (32). These modules (hereafter referred to as 'Network modules') are groups of highly interacting proteins, which tend to be involved in the same cellular processes, metabolic or signalling pathways (33) (Supplementary Table S2).

Using these datasets of protein groups and our protein-lncRNA interaction predictions, we identified lncRNAs that may scaffold complexes or modules by assessing first, for each lncRNA, the enrichment of the lncRNA's interacting proteins among those proteins composing each complex or network module (hypergeometric test, Benjamini-Hochberg corrected FDR 5%) (Figure 1B). Second, because some lncRNAs are predicted to bind a large number of proteins, we estimated the number of protein groups we would expect to find enriched by chance for each lncRNA, as a control, by shuffling the protein labels between protein groups (10 000 times). Only lncRNAs predicted to bind significantly more (empirical *P*-value < 0.01), and at least twice as many, protein groups than expected by chance were considered candidates for scaffolding function.

After filtering using the randomised control, we obtained a total of 27 090 statistically significant enrichments between 1517 protein groups and 847 distinct lncRNA transcripts, encoded by 820 lncRNA genes (Supplementary Table S3). These 847 lncRNAs, ∼5% of our 15 230 tested transcripts, are hereafter referred to as 'scaffolding lncRNA candidates' and constitute a set of lncRNAs predicted to be involved in a scaffolding function (Supplementary Table S4). Remarkably, we also predict that ∼56% of the known protein complexes and 66% of the network modules are scaffolded by at least one lncRNA (Supplementary Table S3). These results suggest that lncRNAs scaffolding complexes and modules are highly prevalent. Moreover, as the set of predicted complexes and modules found to be scaffolded by lncRNAs are involved in most cellular biological processes (Supplementary Figure S3), the scaffolding function of lncRNAs appears therefore to be a general feature and not restricted to specific cellular processes.

Although current experimental protein-lncRNA interaction datasets are largely incomplete and limited to 148 RBPs (14,15,18), we find that 832 out of 6186 lncRNA–protein-group interactions including at least one of the 148 RBPs contain one or more known experimental interactions (Supplementary Table S4). Importantly, as a control, when restricting our scaffolding lncRNA candidate detection method to protein–RNA interactions involving only RNA-binding proteins (1459 RBPs; Supplementary Material), instead of the whole proteome, we identify 788 scaffolding lncRNA candidates among which 572 (72.5%) were also found by our proteome-wide approach. This highly significant overlap (*P*-value $< 2.2e{-}16$, OR = 158, Fisher's exact test; Supplementary Figure S4) reinforces the confidence of our predictions.

Overall, our large-scale approach predicted tens of thousands of lncRNA–protein-group interactions between hundreds of lncRNAs and protein groups, many of which containing experimentally determined interactions, suggesting an abundant presence of lncRNA scaffolds.

Global analysis of scaffolded complexes and modules

In order to analyse the patterns of predicted interactions between lncRNAs and protein groups, we represent them as a clustered matrix (Figure 4). Clusters of protein groups with similar enrichment profiles often share proteins, while clusters of lncRNAs with similar enrichment profiles are largely composed of transcript isoforms from the same or paralog genes. While some protein groups and lncRNAs interact specifically, others—protein groups as well as lncRNAs—do so more promiscuously, and this occurs for each of the four protein group datasets used. Indeed, we observe that some lncRNAs are predicted to interact with 1 to 98 protein groups, according to the dataset, i.e. at most 54 (16.7% of total) non-redundant CORUM complexes, 35 (9.9%) in BioPlex, 98 (10.0%) in Wan 2015, 68 (7.8%) in network modules (Figure 4; Supplementary Figure S5A). Likewise, protein complexes are predicted to interact with 1 to 401 lncRNAs i.e. at most 401 lncRNAs (2.6% of total tested) in non-redundant CORUM, 17 (0.1%) in BioPlex, 248 (1.6%) in Wan 2015, 115 (0.7%) in network modules (Figure 4; Supplementary Figure S5B).

Interestingly, some of our predictions corroborate and further extend the current knowledge of protein–RNA complexes. For instance, the polycomb repressive complex 2 (PRC2 complex), previously found associated with lncRNAs (4), is predicted to be scaffolded by 101 different lncRNAs in our analysis. Indeed, the PRC2 complex and some of its constituent proteins have previously been found to bind

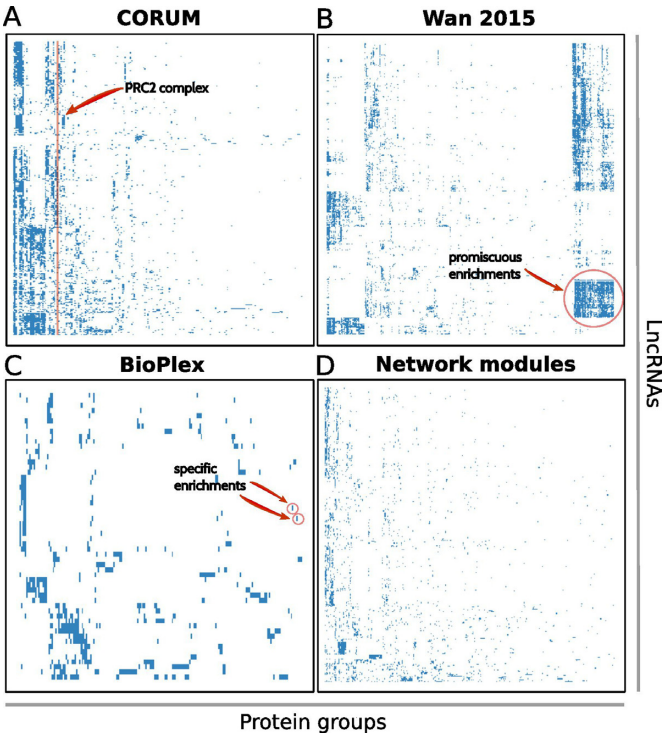


Figure 4. Interactions between lncRNAs and protein groups. Boolean matrix representing enrichment between lncRNAs and protein groups on (**A**) non-redundant CORUM, (**B**) Wan 2015, (**C**) BioPlex and (**D**) Network modules. Blue color represents significant enrichments, white color represents non-significant enrichments. Only lncRNAs/protein groups with at least one significant enrichment are displayed. Matrix was clustered by hierarchical clustering with euclidean distance, dendrograms are not displayed due to the very high number of rows and columns. The PRC2 complex, as well as examples of promiscuous and specific enrichments are highlighted.

hundreds of lncRNAs, presumably as a part of its targeted gene repression mechanism or its regulation by decoy lncRNAs (4,34).

Overall, we find that some lncRNA candidates may act as general scaffolds for several protein groups, while others are specific to one or a few protein groups. Likewise, some protein groups are predicted to interact with many different lncRNAs, perhaps reflecting their function, exemplified by the PRC2 complex.

Scaffolding lncRNA candidates display functional features

To determine if our scaffolding lncRNA candidates are likely to be functional, we gathered several orthogonal datasets of lncRNAs displaying functional features. Together these include lncRNAs (i) displaying a metabolism profile characteristic of functional transcripts (35), (ii) overlapping eQTLs (36); (iii) that alter cell-growth when subjected to inactivation by CRISPRi (37); (iv) involved in disease (38,39), as well as lncRNAs (v) conserved in tetrapods (40) or (vi) possessing structurally conserved elements (41). Strikingly, even though these functional lncRNAs have been found to act not only through protein-binding but also RNA- and DNA-binding, many were successfully identified by our protein–RNA interaction-based approach (Figure 5A). Indeed, we observe a significant (P-value < 0.05, one-tailed Fisher's exact test) and often strong overlap (OR $>$ 2) between our scaffolding lncRNA candidate dataset and every functional or conserved lncRNA dataset analysed except therian-conserved lncRNAs. This latter result suggests that most human scaffolding lncRNAs may have appeared later in evolution or may be highly species-specific.

Additionally, when considering the different sets of scaffolding lncRNA candidates identified using our four different protein group datasets separately, they are all found significantly enriched in functional or conserved lncRNAs from all tested orthogonal datasets (P-value < 0.05, OR from 1.73 to 1.96, one-tailed Fisher's exact test; Supplementary Figure S6A). Different pertinent lncRNA candidates can therefore be detected from each protein group dataset, consistent with the relatively low overlap observed between lncRNAs candidates found from each dataset (Supplementary Figure S6B).

In agreement with our findings, we observe that mutations in exons of scaffolding long non-coding intergenic RNA (lincRNAs) candidates have a higher predicted consequence on fitness than mutations in other lincRNAs, by measuring their fitCons scores (Figure 5B; Supplementary Material), a metric that takes into account sequence polymorphisms in human and sequence divergence in primates (42).

Altogether, these results suggest that our candidates generally possess the features of functional transcripts, therefore lending further weight to our predictions.

LncRNA-associated disease mechanisms could involve lncRNA scaffolding function

Hundreds of lncRNA genes have been associated with several human diseases and conditions including cancer, diabetes and neurodegenerative diseases. As most of these associations were identified through the analysis of lncRNA differential expression in disease states (38,39), knowledge on the molecular role of these lncRNAs in disease is lacking.

We have found 30 scaffolding lncRNA candidate genes associated with disease in lnc2cancer (38) and LncRNADisease (39) databases (Figure 5A; Supplementary Material). We then assessed whether these lncRNA-disease associations could occur through the predicted protein group scaffolding functions of the lncRNAs. For this, we mapped proteins involved in disease from the OMIM database (43) to protein groups and found that 15 out of 30 scaffolding lncRNA candidate genes associated with disease are possibly interacting with a protein group that includes at least one protein associated with the same or similar disease (Figure 6; Supplementary Table S5).

In several cases (e.g. lncRNA genes SNHG1, SOX2-CT and RP11–356I2.4), lncRNAs and diseases are linked through different protein complexes, and involving different proteins, which provides further evidence of the association.

For instance, the SNHG15 lncRNA gene has been associated to Hereditary Haemorrhagic Telangiectasia (HHT) (44), a disease known to be caused by mutations in genes that modulate the TGF-β superfamily (45). Here, we find that two of its transcripts possibly interact with a com-

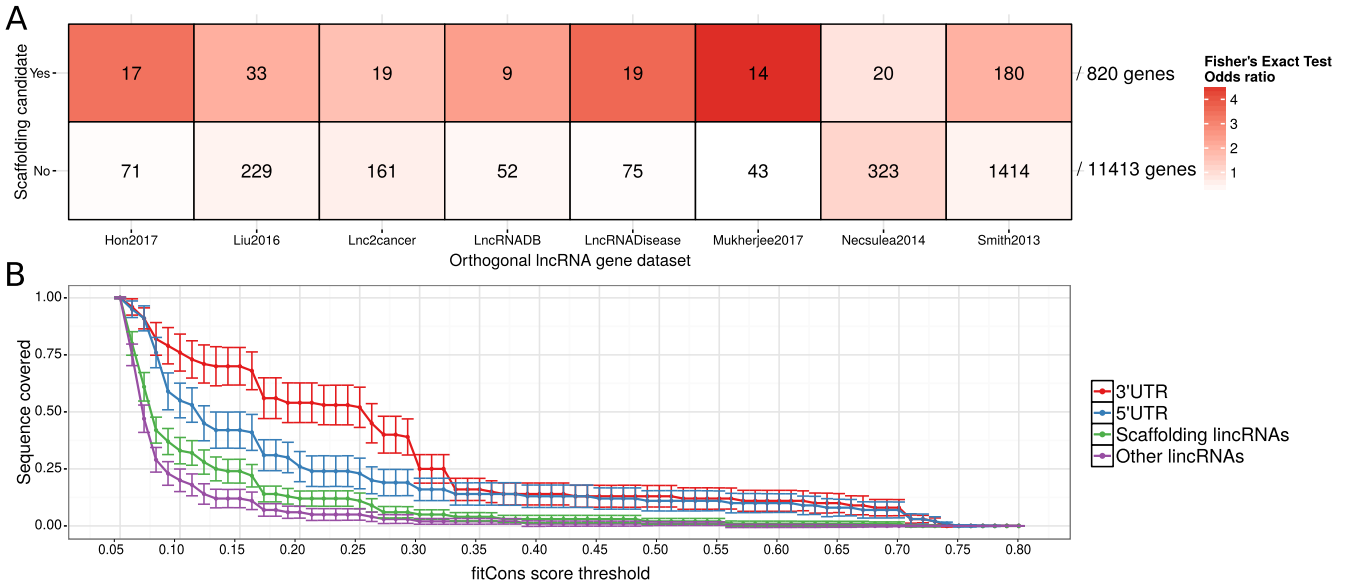


Figure 5. Scaffolding lncRNA candidates display functional features. (**A**) Overlap between scaffolding lncRNA candidate gene (820 genes, 847 transcripts) and the following groups of functional or conserved lncRNA genes characterized in other studies: Hon2017 (36): lncRNAs displaying four features of functionality; Liu2016 (37): lncRNAs affecting cell growth according to CRISPRi experiments; Lnc2cancer (38): lncRNAs involved in cancer; LncRNADB (70): compendium of known functional lncRNAs; LncRNADisease (39): lncRNAs involved in human diseases; Mukherjee2017 (35): lncRNAs with a metabolic profile characteristic of functional transcripts; Necsulea2014 (40): lncRNAs conserved in therians; Smith2013 (41): lncRNAs containing at least one exonic conserved structural element (see Supplementary Material). Enrichment was tested with one-tailed Fisher's exact tests, background included all genes (12233 lncRNA genes, 15230 transcripts) analysed in this study. All *P*-values for the 'Yes' category are significant (*P*-value < 0.05), except for Necsulea2014. (**B**) Proportion of sequence covered with fitCons score above the threshold (x-axis), for different gene features (3′UTR, 5′UTR), lincRNA exons on scaffolding lincRNAs candidates and all other lincRNAs accessed in this study. Error bars: standard deviation of 100 subsampling experiments (with replacement) of 50 genes per category. 'Scaffolding lincRNAs' have a higher proportion of sequence covered above the threshold than 'Other lincRNAs' (one-tailed Kolmogorov–Smirnov test *P*-value = 0.008). As observed in other studies (35,42), lincRNA fitCons scores are lower than UTR regions of protein-coding genes.

plex containing 11 components and regulators of the TGF-β pathway out of 23 proteins (ENST00000585030, non-redundant CORUM complex 81), and with a module composed of signalling proteins and transcription factors (ENST00000578968, network module 686, Supplementary Table S5). Notably, whereas these SNHG15-interacting protein groups are largely composed of different sets of proteins, both contain the SMAD4 protein, a TGF-β pathway component mutated in HHT (46). Overall, further credibility is given to an involvement of SNHG15 in this disease through its predicted scaffolding function.

Moreover, the MEG3 lncRNA gene has been linked to colorectal cancer (47), and has been shown to bind chromatin-remodeling complexes (4). Interestingly, we detected a short MEG3 lncRNA isoform (ENST00000524131, 721 nucleotides) possibly interacting with a complex containing DNA polymerase epsilon subunits as well as chromatin-remodeling proteins (Wan 2015 complex 79), including POLE1, also associated to colorectal cancer.

Finally, the SNHG1 gene is associated to hepatocellular carcinoma (HCC) (48) and non-small cell lung cancer (49). Here we find one of its transcripts (ENST00000539975) interacting with 18 different protein groups associated with one or both of those diseases. Moreover, the interaction of SNHG1 lncRNA with 6 of those protein groups is corroborated by experimental interactions (14,15,18) through five distinct RBPs. Several pathway components of the TNFα/NF-κB signaling pathway have been associated with both lung cancer and HCC, as well as other cancers (50,51). The SNHG1 lncRNA is predicted to interact with the TNFα/NF-κB signaling complex (non-redundant CORUM complex 10) through PAPOLA (poly(A) polymerase α) and CHUK (inhibitor of nuclear factor κ-B kinase subunit α). Notably, the lncRNA interaction with the protein complex is further corroborated by two experimental interactions with two RBPs of the complex, DDX3X and AKAP8L (Supplementary Table S5). Additionally, the SNHG1 lncRNA has been associated with HCC through suppression of miR-195 (52), a microRNA known to target the TNFα/NF-κB pathway by repressing the CHUK protein, and thus suppressing HCC (53) (Figure 7). Given our predictions, we can thus propose that beyond its known effect through miR-195, SNHG1 may regulate elements of the TNFα/NF-κB pathway and therefore directly affect HCC through its possible protein group scaffolding function.

Globally, we propose that the association of 15 lncRNA genes to 22 diseases is due to protein-lncRNA interaction-based mechanisms, notably through the scaffolding of protein complexes and modules by lncRNAs.

DISCUSSION

The current scarcity of experimentally determined lncRNA–protein interaction data hinders the investigation of lncRNA function at large-scale. We thus

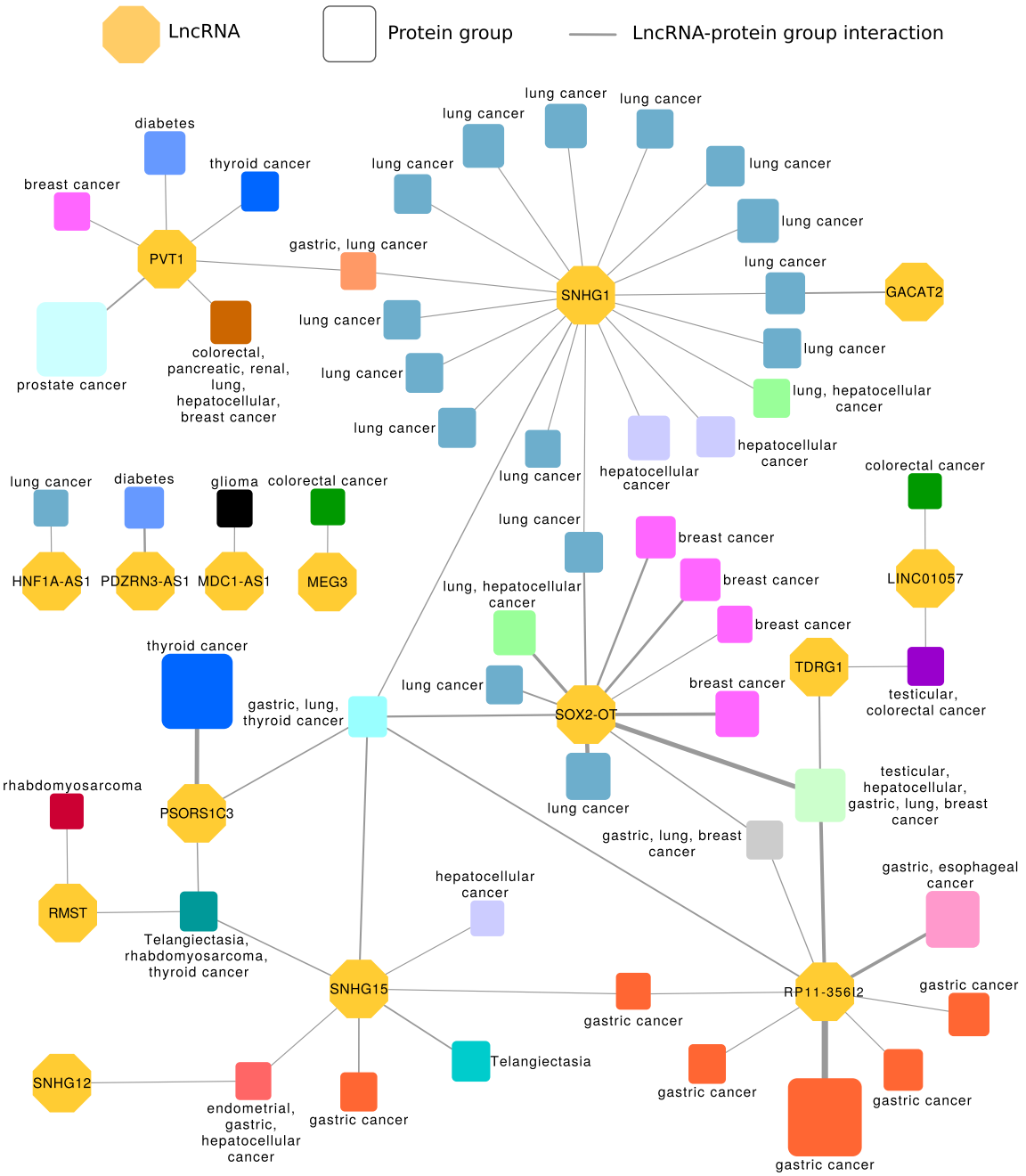


Figure 6. Disease-associated lncRNA network. Network representation of disease-associated lncRNAs (hexagonal nodes in yellow) potentially scaffolding protein groups (square colored nodes) containing at least one protein known to be involved in the same disease. Colors correspond to different diseases. Node size reflects the number of proteins in the group. Edges represent lncRNA–protein-group interactions. Edge width reflects the number of proteins interacting with the lncRNA. LncRNA transcripts were mapped to genes. Some disease names have been abbreviated for simplicity.

computationally predicted a comprehensive lncRNA–protein interaction network in order to better cover the lncRNA–protein interaction space. For this, we used *cat*RAPID, a protein–RNA interaction predictor based on the physicochemical features of the molecules, which can be used large-scale and has been initially validated on a large collection of protein associations with lncRNA (24). Indeed, *cat*RAPID performed well against the NPInter database (area under the receiver operating characteristic (ROC) of 0.88), as well as on the non-nucleic-acid-binding database (area under the ROC curve of 0.92) (31). In addition, we showed herein that *cat*RAPID predictions provide relevant information about lncRNA–protein-complex interactions by experimentally validating that part of the Purα-Purβ-YBX1 complex — predicted here to interact with the *lnc-405* lncRNA — effectively binds the lncRNA *in vivo*.

Noticeably, as the *cat*RAPID predicted interaction network contains the set of biophysically possible interactions between co-expressed molecules, which may differ from in-

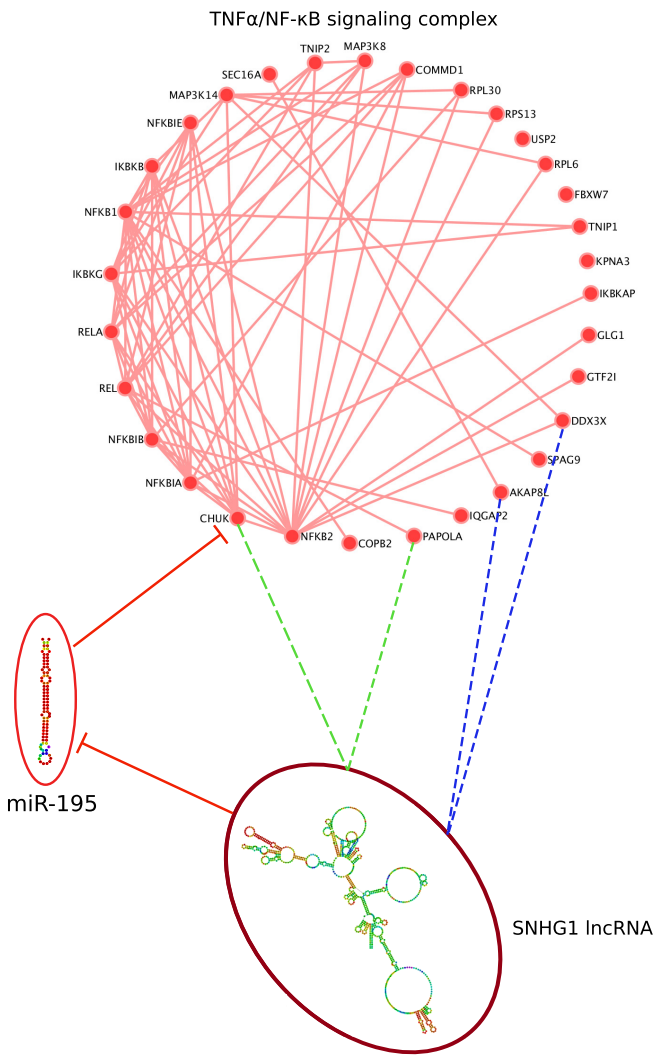


Figure 7. SNHG1 lncRNA gene association to hepatocellular carcinoma through interaction with the TNFα/NF-κB signaling complex. Red nodes represent protein components of the TNFα/NF-κB signaling complex (non-redundant CORUM complex 10). Pink edges correspond to the identified protein-protein interactions between those proteins, downloaded from IntAct (71) on 22 May 2017. Interactions predicted in this study are represented by green dashed edges, experimentally determined ones (see Supplementary Material) by blue dashed edges. Negative regulatory interactions are shown in red and are taken from (53) and (52). ViennaRNA web services were used to predict the secondary structure of SNHG1 and miR-195 (72).

teractions occurring in particular biological contexts, experimentally assessing the quality of the predicted interactions in our network is a desirable goal. However, issues relative to the fraction of interactions to be tested, the sensitivity of the chosen experimental assay, the fraction of interactions identifiable by the chosen assay, and its precision have to be solved beforehand as proposed in the case of the assessment of large-scale binary protein interactomes (54). Moreover, validating the predicted protein complex scaffolding function of lncRNAs is yet another challenge that should involve a wealth of experimental work — e.g., knocking-down of the lncRNA, determination of the localization of the predicted associated complex, its effect on the cell, as well as analysis of binding sites involved in the binding of each protein by the lncRNA (55) — which is beyond the scope of our analysis. Overall, these reasons justify our integration of several orthologous functional datasets to validate our interaction predictions and the possibility of the lncRNA to be indeed functional in the cell.

A growing body of evidence suggests that a significant fraction of lncRNAs has a function (36,37). Large-scale efforts to determine or predict lncRNA function have used their metabolic properties (35,36), sequence or structural conservation (40,41), differential expression in disease (56), lncRNA and protein-coding gene co-expression profiles (57), variant analysis (58), as well as combinations thereof (59). Methods to understand the function of individual lncRNAs through direct interaction with proteins have been exploited to a lesser extent, and are generally restricted to the limited number of known RBPs assessed to date. Hence, there is a clear need for novel large-scale methods to investigate the functions of ncRNAs acting through protein–RNA interactions, such as their ability to scaffold protein groups.

Although protein–RNA interactions are usually perceived as a protein-centric mechanism, they are now also envisioned as a RNA-centric question, where the interactions are driven by the RNA (13). However, even for RNA-centric experiments where the RNA is precipitated and its interacting proteins are identified with MS, each experiment seems to underestimate the number of proteins interacting with lncRNAs. This was observed for the XIST lncRNA, where five independent studies found >600 proteins in total associated with XIST, of which only one is in common between the five studies (12). Hence, we used a method based on proteome-wide and transcriptome-wide interaction predictions combined with tissue-expression information, and predict the presence of millions of protein–RNA interactions in human cells.

As our knowledge of proteins with RNA-binding capabilities is still incomplete (13), we produced proteome-wide protein–RNA interaction predictions to explore the action of lncRNAs at a wider level, going beyond the current knowledge. Indeed, using the *cat*RAPID algorithm, we find that many proteins not yet identified as RBPs have a high propensity to interact with several lncRNAs, as RBPs do. However, with increasingly stringent interaction-propensity cutoffs, we observe a significant increase in the proportion of proteins that are annotated as RBPs (e.g. Spearman's rank correlation coefficient = 0.985, P-value < 2.2e–16, for proteins with at least five interaction partners; Supplementary Figure S7), even though many RBPs display milder binding propensities (e.g. we retain only 79.3% of RBPs with at least 10 interactions above score = 100; 6.6% for score = 200). As RBPs are predicted to interact with lncRNAs with different interaction propensities, we selected an interaction-propensity score cutoff (≥50) that would ensure that we capture biological information, as applied in previous studies (60), while allowing for a large number of possible interactions to be detected.

Due to computational constraints, we have restricted our analysis to lncRNAs of up to 1200 nucleotides, thus excluding well characterized moderately long or very long scaffolding molecules such as MALAT1, NEAT1 and XIST,

that are known to bind dozens to hundreds of proteins (61,62). In addition, lncRNA identification studies have found from tens- to hundreds-of-thousands of novel lncRNAs (63) that are shown to vary according to the methodology used and experimental conditions. This suggests the identification of human lncRNAs is far from complete, but also that lncRNA identification methods are not yet convergent (64). In our study, we therefore restricted our analysis to lncRNAs from the curated dataset of GENCODE, widely considered as the human gene annotation reference standard. However, this also means that several recently found lncRNA scaffolds such as the LUNAR-1 (65), linc-RAM (66) and PARTICLE (67) lncRNAs are not yet present in the GENCODE dataset.

Importantly, we identified for the first time 847 lncRNAs, accounting for ∼5% of the human long non-coding transcriptome, that potentially act as RNA scaffolding molecules for a total of 1517 protein complexes or modules, roughly half of the human protein complexes known to date. As for protein–RNA interactions, knowledge of the human protein complexome is not yet comprehensive. Therefore, we used several datasets of protein complexes to better cover the protein complex space. Indeed, these datasets are largely non-redundant, with 0 to 12.4% of complexes sharing $> = 50\%$ of their constituent proteins with another complex of the same dataset (Supplementary Table S6). In addition, the three datasets are largely complementary, with at the most 20.4% of complexes sharing ≥50% of their proteins between datasets (Supplementary Table S7), and none of the complexes being entirely shared between datasets. A slightly higher inter-dataset overlap (26.2%) is found for network modules, mostly due to the higher module size compared to the protein complexes. As expected, we found that each protein group dataset used allows identifying a different set of scaffolding lncRNA candidates and the majority of the candidates (57%) are detected exclusively with one dataset of protein groups (Supplementary Figure S6B). Overall, this reveals the necessity of considering several datasets for a global analysis of human cellular complexes, as performed in this study.

Notably, our study indicates that RNA scaffolding may be an important regulatory mechanism, not limited to the few well-known cases. We indeed greatly expand the current knowledge on RNA-mediated scaffolding, by proposing that scaffolding occurs with a high prevalence and for most cellular processes. Even though major cellular functions such as telomere repair, signal peptide recognition and translation are known to closely involve RNA components, usual methods to identify cellular macromolecular complexes routinely use an RNA nuclease step before protein purification (2), thereby hindering the possible detection of RNA components in protein complexes. It is therefore likely that many ribonucleoprotein (RNP) complexes have previously been overlooked. These can possibly be retraced with a computational approach, as suggested by our results. Moreover, cellular functions are not only performed via stable macromolecular complexes, but also through stepwise reactions performed by molecules whose temporal and spatial proximity may be mediated by other molecules, as exemplified by the MAYA lncRNA, which links two pathways related to cancer metastasis through protein interaction (68). Such situations are also taken into account by our analyses when investigating interaction enrichment of lncRNAs to functional network modules. Indeed, our data revealed hundreds of modules which may be organized by RNA scaffolding.

Several lncRNAs have been shown to bind protein complexes by interacting with a single protein of the complex. Examples include HOTAIR, MEG3 and Linc-RAM which have been shown to regulate gene expression through their binding to only one component of chromatin-remodeling complexes (PRC2 (4), LSD1 (10), and MyoD–Baf60c–Brg1 complexes (66)). As our enrichment-based approach only allows identification of lncRNAs that bind at least two proteins of the same complex or module, single-protein-binding lncRNAs are beyond the scope of our approach. However, we report a short isoform of the MEG3 gene predicted to interact with several proteins of a chromatin-remodeling-related complex, suggesting that here again, some functional protein–RNA interactions may have been missed by experimental approaches, therefore emphasizing the power of predictive computational analyses.

Overall, our findings suggest the widespread prevalence of scaffolding function for lncRNAs. By proposing that lncRNAs perform such a scaffolding function for a large fraction of protein complexes and functional modules, we further characterize their function and open new questions regarding the importance and essential nature of RNA-mediated scaffolding in the cell.

AVAILABILITY

The filtered protein-lncRNA interaction network produced and analysed in this study is provided at: http://tagc.univ-mrs.fr/MoonDB/protein_lncrna_interaction_network.tsv.gz [38 Mbytes]. Source code used for data processing and analyses can be found at: https://github.com/TAGC-Brun/RAINET-RNA.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We are indebted to Charles E. Chapple for helpful suggestions and careful editing of the manuscript. We thank Anaïs Baudot, Alberto Valdeolivas, Nieves Lorenzo, Davide Cirillo and Alexandros Armaos for fruitful discussions.

FUNDING

Work in IB's lab was partially supported by grants from ERC-2013 [AdG 340172–MUNCODD]; Telethon [GGP16213]; Human Frontiers Science Program Award [RGP0009/2014]; Parent Project Italia, AFM-Telethon [17835]; Epigen-Epigenomics Flagship Project and AriSLA full grant 2014 'ARCI'. Work in GGT's lab was supported by the European Research Council [RIBOMYLOME_309545]; Spanish Ministry of Economy and Competitiveness [BFU2014-55054-P]. The RAINET project leading to this publication has received funding from Excellence Initiative of Aix-Marseille University—A*MIDEX,

a French 'Investissements d'Avenir' programme (to C.B.). Funding for open access charge: Excellence Initiative of Aix-Marseille University—A*MIDEX [RAINET].
Conflict of interest statement. None declared.

REFERENCES

1. Djebali,S., Davis,C.A., Merkel,A., Dobin,A., Lassmann,T., Mortazavi,A., Tanzer,A., Lagarde,J., Lin,W., Schlesinger,F. *et al.* (2012) Landscape of transcription in human cells. *Nature*, **489**, 101–108.
2. Geisler,S. and Coller,J. (2013) RNA in unexpected places: long non-coding RNA functions in diverse cellular contexts. *Nat. Rev. Mol. Cell Biol.*, **14**, 699–712.
3. Harrow,J., Frankish,A., Gonzalez,J.M., Tapanari,E., Diekhans,M., Kokocinski,F., Aken,B.L., Barrell,D., Zadissa,A., Searle,S. *et al.* (2012) GENCODE: the reference human genome annotation for the ENCODE project. *Genome Res.*, **22**, 1760–1774.
4. Mondal,T., Subhash,S., Vaid,R., Enroth,S., Uday,S., Reinius,B., Mitra,S., Mohammed,A., James,A.R., Hoberg,E. *et al.* (2015) MEG3 long noncoding RNA regulates the TGF-β pathway genes through formation of RNA–DNA triplex structures. *Nat. Commun.*, **6**, 7743.
5. Clemson,C.M., Hutchinson,J.N., Sara,S.A., Ensminger,A.W., Fox,A.H., Chess,A. and Lawrence,J.B. (2009) An architectural role for a nuclear noncoding RNA: NEAT1 RNA is essential for the structure of paraspeckles. *Mol. Cell*, **33**, 717–726.
6. Wapinski,O. and Chang,H.Y. (2011) Long noncoding RNAs and human disease. *Trends Cell Biol.*, **21**, 354–361.
7. Spitale,R.C., Tsai,M.C. and Chang,H.Y. (2011) RNA templating the epigenome: Long noncoding RNAs as molecular scaffolds. *Epigenetics*, **6**, 539–543.
8. Good,M.C., Zalatan,J.G. and Lim,W.A. (2011) Scaffold proteins: hubs for controlling the flow of cellular information. *Science*, **332**, 680–686.
9. Chujo,T., Yamazaki,T. and Hirose,T. (2015) Architectural RNAs (arcRNAs): A class of long noncoding RNAs that function as the scaffold of nuclear bodies. *Biochim. Biophys. Acta*, **1859**, 139–146.
10. Tsai,M.-C., Manor,O., Wan,Y., Mosammaparast,N., Wang,J.K., Lan,F., Shi,Y., Segal,E. and Chang,H.Y. (2010) Long noncoding RNA as modular scaffold of histone modification complexes. *Science*, **329**, 689–693.
11. Zhang,Y., He,Q., Hu,Z., Feng,Y., Fan,L., Tang,Z., Yuan,J., Shan,W., Li,C., Hu,X. *et al.* (2016) Long noncoding RNA LINP1 regulates repair of DNA double-strand breaks in triple-negative breast cancer. *Nat. Struct. Mol. Biol.*, **23**, 1–12.
12. Cirillo,D., Blanco,M., Armaos,A., Buness,A., Avner,P., Guttman,M., Cerase,A. and Tartaglia,G.G. (2016) Quantitative predictions of protein interactions with long noncoding RNAs. *Nat. methods*, **14**, 5–6.
13. Beckmann,B.M., Castello,A. and Medenbach,J. (2016) The expanding universe of ribonucleoproteins: of novel RNA-binding proteins and unconventional interactions. *Pflügers Arch. - Eur. J. Physiol.*, **468**, 1029–1040.
14. 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. *et al.* (2016) Robust transcriptome-wide discovery of RNA-binding protein binding sites with enhanced CLIP (eCLIP). *Nat. methods*, **13**, 1–9.
15. Hao,Y., Wu,W., Li,H., Yuan,J., Luo,J., Zhao,Y. and Chen,R. (2016) NPInter v3.0: an upgraded database of noncoding RNA-associated interactions. *Database: J. Biol. databases Curation*, **2016**, baw057.
16. Agostini,F., Zanzoni,A., Klus,P., Marchese,D., Cirillo,D. and Tartaglia,G.G. (2013) CatRAPID omics: a web server for large-scale prediction of protein–RNA interactions. *Bioinformatics*, **29**, 2928–2930.
17. GTEx Consortium (2015) Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science*, **348**, 648–660.
18. Li,J.H., Liu,S., Zhou,H., Qu,L.H. and Yang,J.H. (2014) StarBase v2.0: Decoding miRNA-ceRNA, miRNA-ncRNA and protein–RNA interaction networks from large-scale CLIP-Seq data. *Nucleic Acids Res.*, **42**, 92–97.
19. Huttlin,E.L., Ting,L., Bruckner,R.J., Gebreab,F., Gygi,M.P., Szpyt,J., Tam,S., Zarraga,G., Colby,G., Baltier,K. *et al.* (2015) The BioPlex Network: a systematic exploration of the human interactome. *Cell*, **162**, 425–440.
20. Wan,C., Borgeson,B., Phanse,S., Tu,F., Drew,K., Clark,G., Xiong,X., Kagan,O., Kwan,J., Bezginov,A. *et al.* (2015) Panorama of ancient metazoan macromolecular complexes. *Nature*, **525**, 339–344.
21. Ruepp,A., Waegele,B., Lechner,M., Brauner,B., Dunger-Kaltenbach,I., Fobo,G., Frishman,G., Montrone,C. and Mewes,H.W. (2009) CORUM: the comprehensive resource of mammalian protein complexes-2009. *Nucleic Acids Res.*, **38**, 497–501.
22. Havugimana,P.C., Hart,G.T., Nepusz,T., Yang,H., Turinsky,A.L., Li,Z., Wang,P.I., Boutz,D.R., Fong,V., Phanse,S. *et al.* (2012) A census of human soluble protein complexes. *Cell*, **150**, 1068–1081.
23. Chapple,C.E., Robisson,B., Spinelli,L., Guien,C., Becker,E. and Brun,C. (2015) Extreme multifunctional proteins identified from a human protein interaction network. *Nat. Commun.*, **6**, 7412.
24. Bellucci,M., Agostini,F., Masin,M. and Tartaglia,G.G. (2011) Predicting protein associations with long noncoding RNAs. *Nat. Methods*, **8**, 444–445.
25. Agostini,F., Cirillo,D., Bolognesi,B. and Tartaglia,G.G. (2013) X-inactivation: quantitative predictions of protein interactions in the Xist network. *Nucleic Acids Res.*, **41**, 1–9.
26. Kornienko,A.E., Dotter,C.P., Guenzl,P.M., Gisslinger,H., Gisslinger,B., Cleary,C., Kralovics,R., Pauler,F.M. and Barlow,D.P. (2016) Long non-coding RNAs display higher natural expression variation than protein-coding genes in healthy humans. *Genome Biol.*, **17**, 14.
27. Lee,S., Kopp,F., Chang,T.C., Sataluri,A., Chen,B., Sivakumar,S., Yu,H., Xie,Y. and Mendell,J.T. (2015) Noncoding RNA NORAD regulates genomic stability by sequestering PUMILIO proteins. *Cell*, **164**, 69–80.
28. Minajigi,A., Froberg,J.E., Wei,C., Sunwoo,H., Kesner,B., Colognori,D., Lessing,D., Payer,B., Boukhali,M., Haas,W. *et al.* (2015) Chromosomes. A comprehensive Xist interactome reveals cohesin repulsion and an RNA-directed chromosome conformation. *Science*, **349**, aab2276.
29. Ballarino,M., Cazzella,V., D'Andrea,D., Grassi,L., Bisceglie,L., Cipriano,A., Santini,T., Pinnarò,C., Morlando,M., Tramontano,A. *et al.* (2015) Novel long noncoding RNAs (lncRNAs) in myogenesis: a miR-31 overlapping lncRNA transcript controls myoblast differentiation. *Mol. Cell. Biol.*, **35**, 728–736.
30. Kelm,R.J., Cogan,J.G., Elder,P.K., Strauch,A.R. and Getz,M.J. (1999) Molecular interactions between single-stranded DNA-binding proteins associated with an essential MCAT element in the mouse smooth muscle alpha-actin promoter. *J. Biol. Chem.*, **274**, 14238–14245.
31. Cirillo,D., Agostini,F. and Tartaglia,G.G. (2013) Predictions of protein–RNA interactions. *WIRES Comput. Mol. Sci.*, **3**, 161-175.
32. Becker,E., Robisson,B., Chapple,C.E., Guénoche,A. and Brun,C. (2012) Multifunctional proteins revealed by overlapping clustering in protein interaction network. *Bioinforma.*, **28**, 84–90.
33. Brun,C., Chevenet,F., Martin,D., Wojcik,J., Guénoche,A. and Jacq,B. (2003) Functional classification of proteins for the prediction of cellular function from a protein-protein interaction network. *Genome Biol.*, **5**, R6.
34. Davidovich,C., Zheng,L., Goodrich,K.J. and Cech,T.R. (2013) Promiscuous RNA binding by Polycomb repressive complex 2. *Nat. Struct. Mol. Biol.*, **20**, 1250–1257.
35. Mukherjee,N., Calviello,L., Hirsekorn,A., de Pretis,S., Pelizzola,M. and Ohler,U. (2017) Integrative classification of human coding and noncoding genes through RNA metabolism profiles. *Nat. Struct. Mol. Biol.*, **24**, 86–96.
36. Hon,C., Ramilowski,J., Harshbarger,J., Bertin,N., Rackham,O., Gough,J., Denisenko,E., Schmeier,S., Poulsen,T., Severin,J. *et al.* (2017) An atlas of human long non-coding RNAs with accurate 5′ ends. *Nature*, **543**, 199-204.
37. Liu,S.J., Liu,S.J., Horlbeck,M.A., Cho,S.W., Birk,H.S., Malatesta,M., Attenello,F.J., Villalta,J.E., Cho,M.Y., Chen,Y. *et al.* (2017) CRISPRi-based genome-scale identification of functional long noncoding RNA loci in human cells. *Science*, **355**, aah7111.
38. Ning,S., Zhang,J., Wang,P., Zhi,H., Wang,J., Liu,Y., Gao,Y., Guo,M., Yue,M., Wang,L. *et al.* (2016) Lnc2Cancer: a manually curated

database of experimentally supported lncRNAs associated with various human cancers. *Nucleic Acids Res.*, **44**, D980–D985.
39. Chen,G., Wang,Z., Wang,D., Qiu,C., Liu,M., Chen,X., Zhang,Q., Yan,G. and Cui,Q. (2013) LncRNADisease: a database for long-non-coding RNA-associated diseases. *Nucleic Acids Res.*, **41**, 983–986.
40. Necsulea,A., Soumillon,M., Warnefors,M., Liechti,A., Daish,T., Zeller,U., Baker,J.C., Grützner,F. and Kaessmann,H. (2014) The evolution of lncRNA repertoires and expression patterns in tetrapods. *Nature*, **505**, 635–640.
41. Smith,M.A., Gesell,T., Stadler,P.F. and Mattick,J.S. (2013) Widespread purifying selection on RNA structure in mammals. *Nucleic Acids Res.*, **41**, 8220–8236.
42. Gulko,B., Hubisz,M.J., Gronau,I. and Siepel,A. (2015) A method for calculating probabilities of fitness consequences for point mutations across the human genome. *Nat. Publ. Group*, **47**, 276–283.
43. Hamosh,A., Scott,A.F., Amberger,J.S., Bocchini,C.A. and McKusick,V.A. (2005) Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res.*, **33**, D514–D517.
44. Tørring,P.M., Larsen,M.J., Kjeldsen,A.D., Ousager,L.B., Tan,Q. and Brusgaard,K. (2014) Long non-coding RNA expression profiles in hereditary haemorrhagic telangiectasia. *PLoS One*, **9**, e90272.
45. Dupuis-Girod,S., Bailly,S. and Plauchu,H. (2010) Hereditary hemorrhagic telangiectasia: from molecular biology to patient care. *J. Thromb. Haemostasis: JTH*, **8**, 1447–1456.
46. Gallione,C.J., Repetto,G.M., Legius,E., Rustgi,A.K., Schelley,S.L., Tejpar,S., Mitchell,G., Drouin,E., Westermann,C.J.J. and Marchuk,D.A. (2004) A combined syndrome of juvenile polyposis and hereditary haemorrhagic telangiectasia associated with mutations in MADH4 (SMAD4). *Lancet*, **363**, 852–859.
47. Yin,D.-D., Liu,Z.-J., Zhang,E., Kong,R., Zhang,Z.-H. and Guo,R.-H. (2015) Decreased expression of long noncoding RNA MEG3 affects cell proliferation and predicts a poor prognosis in patients with colorectal cancer. *Tumour Biol. J. Int. Soc. Oncodev. Biol. Med.*, **36**, 4851–4859.
48. Zhang,M., Wang,W., Li,T., Yu,X., Zhu,Y., Ding,F., Li,D. and Yang,T. (2016) Long noncoding RNA SNHG1 predicts a poor prognosis and promotes hepatocellular carcinoma tumorigenesis. *Biomed. Pharmacother.*, **80**, 73–79.
49. You,J., Fang,N., Gu,J., Zhang,Y., Li,X., Zu,L. and Zhou,Q. (2014) Noncoding RNA small nucleolar RNA host gene 1 promote cell proliferation in nonsmall cell lung cancer. *Indian J. Cancer*, **51**, e99–e102.
50. Luedde,T. and Schwabe,R.F. (2011) NF-κB in the liver–linking injury, fibrosis and hepatocellular carcinoma. *Nat. Rev. Gastroenterol. Hepatol.*, **8**, 108–118.
51. Wu,Y. and Zhou,B.P. (2010) TNF-alpha/NF-kappaB/Snail pathway in cancer cell migration and invasion. *Br. J. Cancer*, **102**, 639–644.
52. Zhang,H., Zhou,D., Ying,M., Chen,M., Chen,P., Chen,Z. and Zhang,F. (2016) Expression of long non-coding RNA (lncRNA) small nucleolar RNA host gene 1 (SNHG1) exacerbates hepatocellular carcinoma through suppressing miR-195. *Med. Sci. Monit. Int. Med. J. Exp. Clin. Res.*, **22**, 4820–4829.
53. Ding,J., Huang,S., Wang,Y., Tian,Q., Zha,R., Shi,H., Wang,Q., Ge,C., Chen,T., Zhao,Y. *et al.* (2013) Genome-wide screening reveals that miR-195 targets the TNF-α/NF-κB pathway by down-regulating IκB kinase alpha and TAB3 in hepatocellular carcinoma. *Hepatol.*, **58**, 654–666.
54. Venkatesan,K., Rual,J.-F., Vazquez,A., Stelzl,U., Lemmens,I., Hirozane-Kishikawa,T., Hao,T., Zenkner,M., Xin,X., Goh,K.-I. *et al.* (2009) An empirical framework for binary interactome mapping. *Nat. Methods*, **6**, 83–90.
55. Wang,K.C. and Chang,H.Y. (2012) Molecular mechanisms of long noncoding RNAs. *Mol. Cell*, **43**, 904–914.
56. Li,J., Han,L., Roebuck,P., Diao,L., Liu,L., Yuan,Y., Weinstein,J.N. and Liang,H. (2015) TANRIC: an interactive open platform to explore the function of lncRNAs in cancer. *Cancer Res.*, **75**, 3728–3737.
57. Zhao,Z., Bai,J., Wu,A., Wang,Y., Zhang,J., Wang,Z., Li,Y., Xu,J. and Li,X. (2015) Co-LncRNA: investigating the lncRNA combinatorial effects in GO annotations and KEGG pathways based on human RNA-Seq data. *Database: J. Biol. databases Curation*, **2015**, 1–7.
58. Chen,X., Hao,Y., Cui,Y., Fan,Z., He,S., Luo,J. and Chen,R. (2017) LncVar: a database of genetic variation associated with long non-coding genes. *Bioinformatics*, **33**, 112–118.
59. Park,C., Yu,N., Choi,I., Kim,W. and Lee,S. (2014) LncRNAtor: a comprehensive resource for functional investigation of long non-coding RNAs. *Bioinformatics*, **30**, 2480–2485.
60. Zanzoni,A., Marchese,D., Agostini,F., Bolognesi,B., Cirillo,D., Botta-Orfila,M., Livi,C.M., Rodriguez-Mulero,S. and 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.
61. West,J., Davis,C., Sunwoo,H., Simon,M., Sadreyev,R., Wang,P., Tolstorukov,M. and Kingston,R. (2014) The long noncoding RNAs NEAT1 and MALAT1 bind active chromatin sites. *Mol. Cell*, **55**, 791–802.
62. Cerase,A., Pintacuda,G., Tattermusch,A. and Avner,P. (2015) Xist localization and function: new insights from multiple levels. *Genome Biol.*, **16**, 166.
63. Zhao,Y., Li,H., Fang,S., Kang,Y., Hao,Y., Li,Z., Bu,D., Sun,N., Zhang,M.Q. and Chen,R. (2016) NONCODE 2016: an informative and valuable data source of long non-coding RNAs. **44**, D203–D208.
64. Kashi,K., Henderson,L., Bonetti,A. and Carninci,P. (2015) Discovery and functional analysis of lncRNAs: methodologies to investigate an uncharacterized transcriptome. *Biochim. Biophys. Acta*, **1859**, 3–15.
65. Trimarchi,T., Bilal,E., Ntziachristos,P., Fabbri,G., Dalla-Favera,R., Tsirigos,A. and Aifantis,I. (2014) Genome-wide mapping and characterization of Notch-regulated long noncoding RNAs in acute leukemia. *Cell*, **158**, 593–606.
66. Yu,X., Zhang,Y., Li,T., Ma,Z., Jia,H., Chen,Q., Zhao,Y., Zhai,L., Zhong,R., Li,C. *et al.* (2017) Long non-coding RNA Linc-RAM enhances myogenic differentiation by interacting with MyoD. *Nat. Commun.*, **8**, 14016.
67. O'Leary,V.B., Hain,S., Maugg,D., Smida,J., Azimzadeh,O., Tapio,S., Ovsepian,S.V. and Atkinson,M.J. (2017) Long non-coding RNA PARTICLE bridges histone and DNA methylation. *Sci. Rep.*, **7**, 1790.
68. Li,C., Wang,S., Xing,Z., Lin,A., Liang,K., Song,J., Hu,Q., Yao,J., Chen,Z., Park,P.K. *et al.* (2017) A ROR1-HER3-lncRNA signalling axis modulates the Hippo – YAP pathway to regulate bone metastasis. *Nat. Cell Biol.*, **19**, 106–119.
69. Subramanian,A., Tamayo,P., Mootha,V.K., Mukherjee,S., Ebert,B.L., Gillette,M.A., Paulovich,A., Pomeroy,S.L., Golub,T.R., Lander,E.S. *et al.* (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U.S.A.*, **102**, 15545–15550.
70. Quek,X.C., Thomson,D.W., Maag,J.L.V., Bartonicek,N., Signal,B., Clark,M.B., Gloss,B.S. and Dinger,M.E. (2015) lncRNAdb v2.0: Expanding the reference database for functional long noncoding RNAs. *Nucleic Acids Res.*, **43**, D168–D173.
71. Orchard,S., Ammari,M., Aranda,B., Breuza,L., Briganti,L., Broackes-Carter,F., Campbell,N.H., Chavali,G., Chen,C., del-Toro,N. *et al.* (2014) The MIntAct project–IntAct as a common curation platform for 11 molecular interaction databases. *Nucleic acids Res.*, **42**, D358–D363.
72. Lorenz,R., Bernhart,S.H., Höner Zu Siederdissen,C., Tafer,H., Flamm,C., Stadler,P.F. and Hofacker,I.L. (2011) ViennaRNA Package 2.0. *Algorithms Mol. Biol. AMB*, **6**, 26.

2.2. MoonDB 2.0: an updated database of extreme multifunctional and moonlighting proteins

Moonlighting proteins are a subset of multifunctional proteins that are able to perform several unrelated functions. Even though a few hundred cases of moonlighting proteins have been described, these were found mostly through serendipity, since clear procedures to identify secondary moonlighting functions are inexistent. Thus, the full prevalence of moonlighting proteins in a species is yet unknown. Moreover, the set of known moonlighting proteins may be biased to include the most extensively studied proteins, as well as for proteins with functions that are easily characterised (e.g. enzymatic activity, DNA-binding). Yet, identifying moonlighting proteins is important because these proteins could play relevant regulatory roles in both normal and pathological cells (Jeffery, 2018). In addition, drug-design needs to be aware of the potential moonlighting functions of a protein chosen as drug target, in order to avoid unexpected side-effects due to possible interferences with an undisclosed function. Indeed, it has been suggested that moonlighting proteins are often associated to more than one disease, possibly explaining disease comorbidity patterns (Zanzoni, Chapple and Brun, 2015). Paramountly, methods dedicated to the discovery of moonlighting proteins large-scale are needed.

The MoonGO method to identify extreme multifunctional proteins at a proteome-scale was developed by Christine Brun's group in 2015 (Chapple *et al.*, 2015). This method employed for the first time a combination of protein-protein interaction networks and functional annotations to identify human 'extreme multifunctional' (EMF) proteins, defined as proteins whose multiple functions are very dissimilar to one another. While related to moonlighting proteins, whose definition may be too strict (Introduction, section 1.4.1), the term 'extreme multifunctional' proteins is used to englobe all proteins that have very dissimilar functions, regardless of their domain organisation or evolutionary history (Chapple and Brun, 2015).

Given that moonlighting proteins interact with different sets of proteins to perform their different functions, PPI networks can be used to identify these proteins. Indeed, the MoonGO pipeline (Figure 2.1) takes this into account. First, a PPI network is partitioned into groups of overlapping clusters, which often represent functionally related proteins (Introduction, section 1.1.5), using the OCG algorithm (E. Becker *et al.*, 2012). In this way, proteins that belong to one or more clusters can be retrieved. Second, network clusters are annotated according to the Gene Ontology (GO) annotations (Biological Processes) of their constituent proteins using a majority rule. Finally, EMF protein candidates are identified at the

intersection of clusters involved in unrelated biological processes according to PrOnto GO term association probabilities (Figure 2.1). PrOnto is an original tool developed with the specific purpose of identifying pairs of GO terms that are dissimilar to each other, and is based on the probabilities of finding a pair of GO terms annotated to the same protein or to interacting proteins (Chapple, Herrmann and Brun, 2015). Overall, this multi-level approach ensures that the EMF proteins identified are not only connected to several groups of functionally-related proteins, but also annotated with functions that are highly dissimilar to each other. The initial application of the MoonGO pipeline to identify human EMF proteins retrieved 430 proteins that were deposited into MoonDB, a database of extreme multifunctional and moonlighting proteins, contain also a curated set of known human moonlighting proteins (Chapple *et al.*, 2015). This set of proteins is much larger than other collections of experimental human moonlighting proteins. Moreover, the production of an extensive set of EMF proteins allowed to analyse for the first time these proteins as a group, evidencing a feature signature particular to these proteins, such as a high presence of SLiMs and ubiquitous tissue expression (Introduction, section 1.4.2) (Chapple *et al.*, 2015).

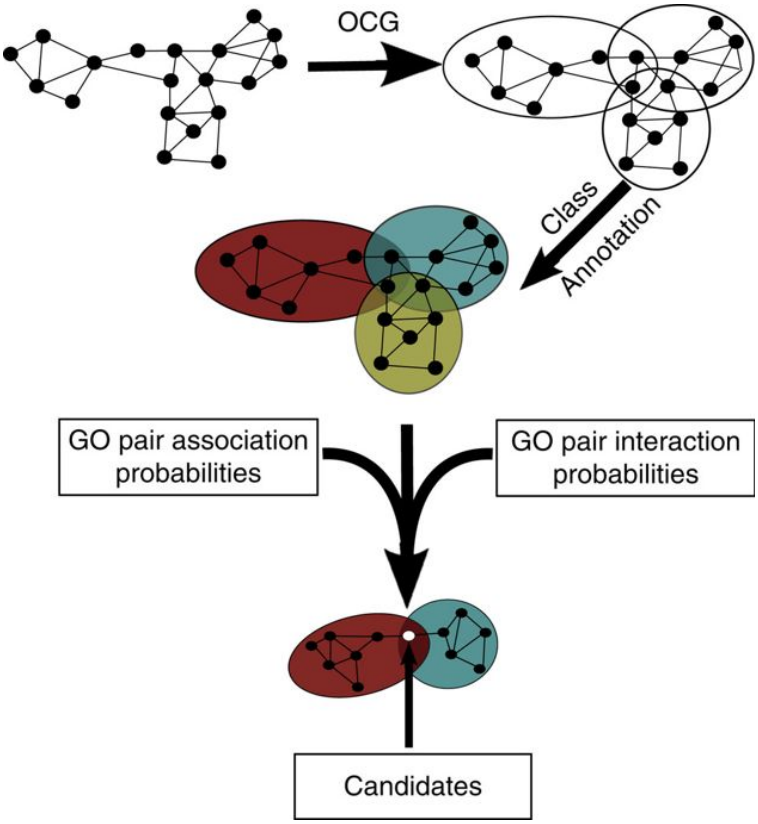


Figure 2.1 | The MoonGO pipeline used for the identification of EMF proteins. Figure from (Chapple *et al.*, 2015).

By exploiting PPI networks and GO term annotations, the ability of the MoonGO pipeline to determine EMF proteins depends on data that is constantly being generated and updated. To be relevant to the community, a database needs to be continuously up-to-date. Therefore, in this thesis I employed the MoonGO pipeline to produce novel EMF protein candidates using the latest protein-protein interaction and GO term annotations, leading to the development of MoonDB 2.0. Besides, MoonDB 2.0 now includes EMF protein predictions for other model organisms such as mouse, worm, fly and yeast, as well as more manually curated moonlighting protein entries. The interface of MoonDB 2.0 was fully modernised and this database is now cross-referenced by the UniProtKB database, thus magnifying the exposure of the database to the general scientific community.

As there are many open questions regarding moonlighting proteins, such as the regulation of their multiple functions, a systematically-detected dataset of extreme multifunctional proteins is of great value for large-scale analysis. In this thesis, this set of proteins is used to analyse the role of 3'UTRs in regulating protein multifunctionality (Results, section 2.3).

Ribeiro, DM, Briere G, Bely, B, Spinelli, L and Brun, C (2018) MoonDB 2.0: an updated database of extreme multifunctional and moonlighting proteins. *Nucleic Acids Research* (under minor revisions).

MoonDB database available on: http://moondb.hb.univ-amu.fr/

Supplementary material is available on the *Appendix II*

MoonDB 2.0: an updated database of extreme multifunctional and moonlighting proteins

Diogo M. Ribeiro[1], Galadriel Briere[1†], Benoit Bely[2], Lionel Spinelli[1], Christine Brun[1,3*]

1. Aix-Marseille Univ, INSERM, TAGC, UMR_S1090, Marseille, France

2. The European Bioinformatics Institute (EMBL-EBI), Wellcome Genome Campus, CB10 1SD, United Kingdom

3. CNRS, Marseille, France

*To whom correspondence should be addressed: Tel: +33 491828712; Email: christine-g.brun@inserm.fr

†Present address: Bordeaux Sciences Agro, 33175 Gradignan, France

ABSTRACT

MoonDB 2.0 (http://moondb.hb.univ-amu.fr/) is a database of predicted and manually curated extreme multifunctional (EMF) and moonlighting proteins, i.e. proteins that perform multiple unrelated functions. We have previously shown that such proteins can be predicted through the analysis of their molecular interaction subnetworks, their functional annotations and their association to distinct groups of proteins that are involved in unrelated functions. In MoonDB 2.0, we updated the set of human EMF proteins (238 proteins), using the latest functional annotations and protein-protein interaction networks. Furthermore, for the first time, we applied our method to four additional model organisms - mouse, fly, worm and yeast - and identified 54 novel EMF proteins in these species. In addition to novel predictions, this update contains 63 human and yeast proteins that were manually curated from literature, including descriptions of moonlighting functions and associated references. Importantly, MoonDB's interface was fully redesigned and improved, and its entries are now cross-referenced in the UniProt Knowledgebase

(UniProtKB). MoonDB will be updated once a year with the novel EMF candidates calculated from the latest available protein interactions and functional annotations.

INTRODUCTION

Moonlighting, multitask and extreme multifunctional proteins are proteins that perform multiple unrelated biological functions, regardless of their domain organisation and their evolutionary history (1-3). A canonical example of a moonlighting protein is the human aconitase, an enzyme of the tricarboxylic acid cycle (TCA cycle) that also functions as a translational regulator, upon a conformational change (4). Extreme multifunctional and moonlighting proteins are present throughout the evolutionary tree, and their unrelated functions may be performed in different tissues or cellular locations, sometimes associated (either as a cause or a consequence) to a change in their interaction partners, conformation or oligomeric states (5). These proteins are often in the intersection - and may coordinate - several pathways or responses to different stimuli (6). Despite their importance, the moonlighting functions of proteins have usually been identified by serendipity, since clear procedures to identify secondary functions have not been proposed. As a consequence, the prevalence of moonlighting proteins in proteomes was unknown. This prompted us to provide in 2015, MoonGO, a computational pipeline to identify extreme multifunctional (EMF) proteins on a large scale (2). EMF proteins were identified by exploiting the topology of protein-protein interaction networks and protein GO term annotations, without any *a priori* knowledge of moonlighting. The first version of MoonDB (2) contained the EMF proteins predicted by MoonGO, complemented with a careful manual curation of literature of moonlighting or EMF proteins. Here, we present MoonDB 2.0, an update that, besides improving predictions and manual curation for human, also includes predicted and curated entries for four other model organisms - mouse, fly, worm and yeast. Our main focus is to provide users with an extensive set of predicted and curated EMF and moonlighting proteins, describing their functions comprehensively.

MATERIALS AND METHODS

Prediction of extreme multifunctional proteins

The method used to predict extreme multifunctional (EMF) proteins was first described in Chapple et al. (2). Briefly, we perform a large-scale search for EMF proteins by *i)* identifying functionally-dissimilar pairs of Biological Process Gene Ontology (GO) terms with PrOnto (7) that uses two metrics of GO *functional dissimilarity* based on the frequency of co-occurrence of GO term pairs in protein annotations; *ii)* clustering the protein interactome into overlapping clusters of proteins using the OCG algorithm (8); *iii)* annotating each cluster with functions (Biological Process GO terms) based on the annotations of its constituent proteins; *iv)* identifying proteins that belong to at least two clusters and are annotated to *dissimilar functions* (after having inherited the annotations of their clusters in addition to their own), hereby labeled as EMF proteins. We used protein-protein interaction data gathered on December 2017 from the PSICQUIC webservice (9), processed as described in Chapple et al. 2015 (2). We only include experimentally identified binary protein-protein interactions, by considering only interactions from certain experimental methods (Supplementary Table S2). GO term annotations and ontologies were collected from the Gene Ontology Consortium (10) on December 2017.

Criteria for manual curation

We provide a list of *bona fide* moonlighting and extreme multifunctional proteins manually curated from literature over the years. Each entry was confirmed independently by at least two members of our team. Specifically, we confirm that the several functions are indeed distinct to each other and not a by-product of the same function under different circumstances (e.g. regulation of two distinct pathways through the same mechanism, such as phosphorylation). In each case, publications describing the different functions of a protein are provided. When available, the conditions that may be related to the change in function are also described (e.g. cellular localisation, oligomerization).

Database architecture and web interface

MoonDB 2.0 has been developed using the SQLAlchemy (v1.2.0) Python (v2.7.6) library for data storage. The web interface is mainly written in PHP (v7.1.14) and JQuery (v3.2.1) and is powered by the Drupal (v8.4) Content Management System (CMS). The database was deployed with Docker (v17.09.0-ce) to ensure stability. We gathered information on protein domains, publications and diseases from UniProtKB (11) in January 2018.

DATABASE CONTENT AND WEB INTERFACE

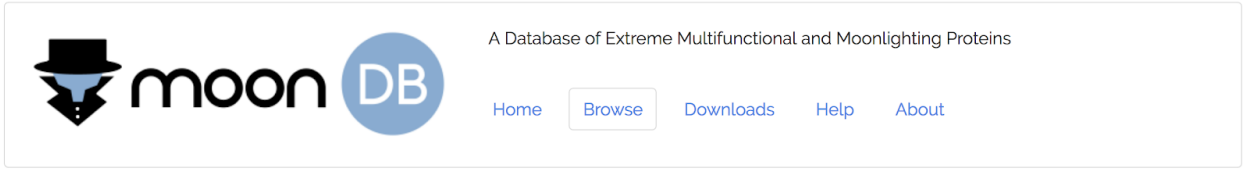
A new dataset of extreme multifunctional proteins

In MoonDB 2.0 we have predicted 292 extreme multifunctional (EMF) proteins in human, and - for the first time - also in mouse, fly, worm and yeast model species (Supplementary Table S1). These have been created *de novo* using the latest protein-protein interaction networks and GO term annotations. The power to predict EMF proteins is dependent on the underlying coverage and quality of the protein interactomes and GO term annotations used (Supplementary Table S1). The new human EMF protein dataset and the one in the previous MoonDB version overlap significantly (Supplementary Figure S1). Interestingly, the analysis of the EMF signature on the new set of EMF proteins (as performed in Chapple *et al (Chapple et al., 2015)*), shows that the new set of EMF proteins produces a similar signature in terms of network properties, tissue expression, as well as domain, structural disorder and Eukaryotic linear motif (ELM) presence, among other features (Supplementary Figure S2). Notably, as observed in other studies (2,12,13), moonlighting and EMF proteins are often associated to disease, a feature also observed for the set of human EMF proteins when considering disease-associations from OMIM ($P = 2.2x10^{-7}$, odds ratio = 2.10; one-tailed Fisher's Exact test).

Besides EMF predictions, in this update we also manually curated 15 yeast moonlighting proteins, describing their unrelated functions, specifying which conditions may influence moonlighting (e.g. cellular localisation) and referencing relevant publications. Similarly, we added functional descriptions for 47 human moonlighting proteins. All these proteins constitute the ‘Reference Set’.

A new user-friendly interface and additional content

To provide our visitors with a clear, fast and easy-to-use database, we completely redesigned MoonDB's web interface and added new functionalities. It is now possible to search a MoonDB entry by gene name, UniProtKB identifier (ID) or UniProtKB accession (AC). Moreover, the 'Browse' page (Figure 1), which displays all protein entries in MoonDB, can now be filtered through any column, thus allowing searches by species, full name of the protein and its presence in the 'Reference Set'. These filters can also be used in combination with each other to make more advanced queries.



Show 15 entries

Copy Excel CSV PDF

MoonDB ID	UniprotKB AC	Gene Name	Protein Full Name	Species	Reference Set
e.g. 14	e.g. P63104	e.g. YWHAE	e.g. Actin	e.g. yeast	e.g. yes
1	Uniprot: Q86VP1	TAX1BP1	Tax1-binding protein 1	human	No
2	Uniprot: Q8WUM4	PDCD6IP	Programmed cell death 6-interacting protein	human	No
3	Uniprot: Q6UWE0	LRSAM1	E3 ubiquitin-protein ligase LRSAM1	human	No
4	Uniprot: O00499	BIN1	Myc box-dependent-interacting protein 1	human	No
5	Uniprot: P06744	GPI	Glucose-6-phosphate isomerase	human	Yes
6	Uniprot: Q8TB24	RIN3	Ras and Rab interactor 3	human	No
7	Uniprot: P62256	UBE2H	Ubiquitin-conjugating enzyme E2 H	human	No
8	Uniprot: Q06124	PTPN11	Tyrosine-protein phosphatase non-receptor type 11	human	No
9	Uniprot: P04792	HSPB1	Heat shock protein beta-1	human	No
10	Uniprot: P32121	ARRB2	Beta-arrestin-2	human	Yes
11	Uniprot: P00519	ABL1	Tyrosine-protein kinase ABL1	human	No
12	Uniprot: O14503	BHLHE40	Class E basic helix-loop-helix protein 40	human	No
13	Uniprot: P02511	CRYAB	Alpha-crystallin B chain	human	Yes
14	Uniprot: P62136	PPP1CA	Serine/threonine-protein phosphatase PP1-alpha catalytic subunit	human	No
15	Uniprot: P63279	UBE2I	SUMO-conjugating enzyme UBC9	human	No

Showing 1 to 15 of 351 entries

First Previous 1 2 3 4 5 ... 24 Next Last

Figure 1. MoonDB 2.0 browse page. The browse page displays the entries of all MoonDB 2.0 proteins and can be searched interactively. For example, the figure displays results with the "Species" and

“Reference Set” filters active. The ‘MoonDB ID’ can be clicked to access each individual MoonDB 2.0 protein entry.

Importantly, MoonDB specifies which pairs of dissimilar (i.e. unrelated) functions led us to propose each predicted EMF and curated protein as moonlighting/extreme-multifunctional (Figure 2, under “MoonDB Dissimilar Functions”). Furthermore, we provide the set of GO terms associated to the protein (Figure 2, under “Network Module GO Annotations”), determined by its participation in network clusters with annotated functions (*guilt-by-association* principle), and the GO terms directly annotating the protein. This information is pertinent in the context of multifunctionality, since EMF proteins associate with several groups of proteins to perform alternative functions. In addition, since the ability to perform unrelated functions may be correlated with the presence of a protein in unrelated subcellular locations, in MoonDB 2.0 we identified pairs of unrelated cellular component GO terms associated to each protein with PrOnto (7) (Figure 2, under “Protein GO Annotations”). Lastly, to fully describe moonlighting and extreme multifunctional proteins, we further cross-link functional data with other orthogonal information such as the protein association to disease, protein domains and publications associated to the protein.

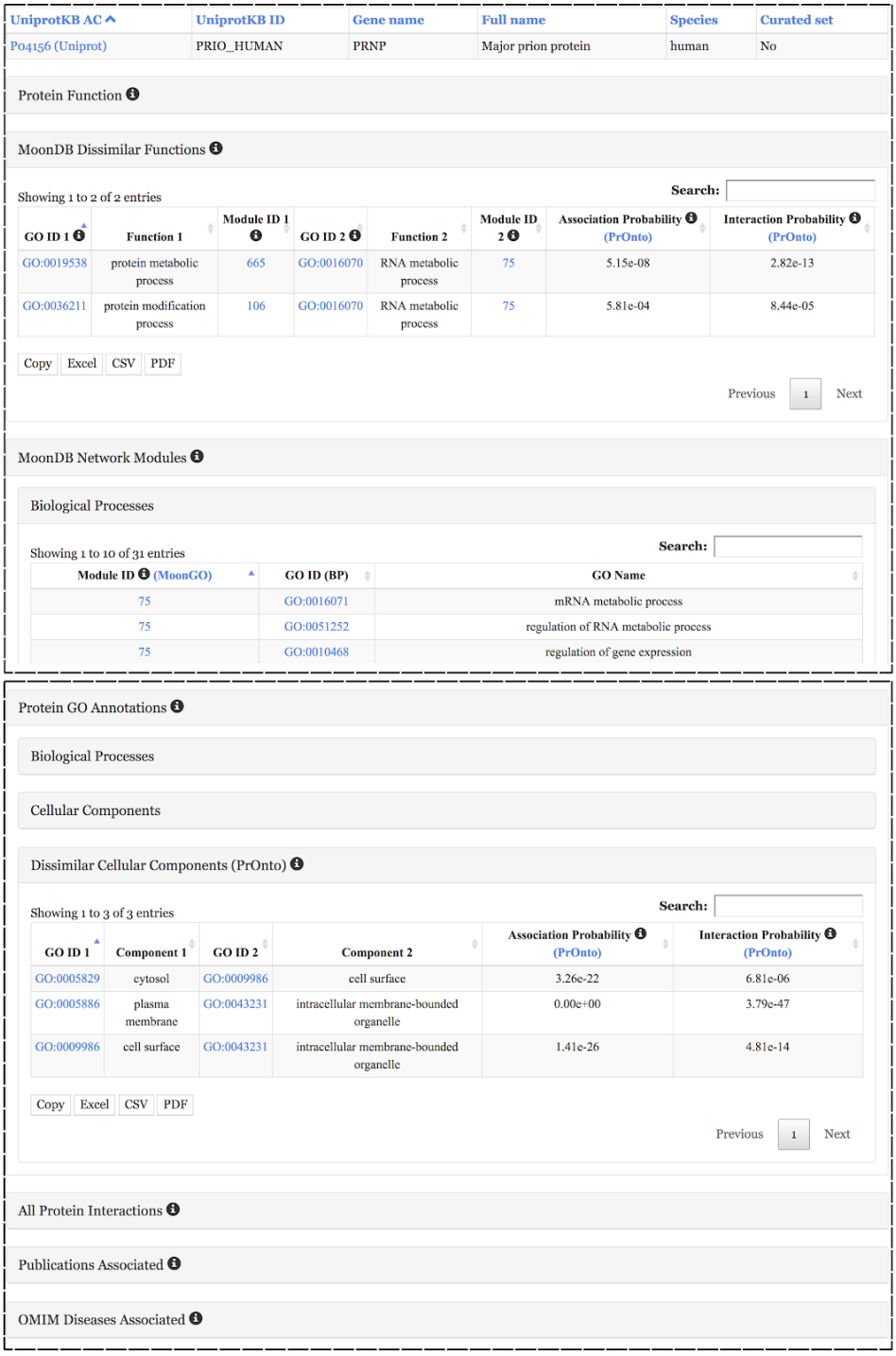


Figure 2. Example of a MoonDB's protein entry. Protein entries provide extensive functional information such as the dissimilar function annotations and GO term annotations from network modules, as well as publications, diseases and domains associated with the protein.

DISCUSSION AND CONCLUSION

The MoonDB 2.0 database is accessible at http://moondb.hb.univ-amu.fr/ and now contains data for human, mouse, fly, worm and yeast. MoonDB 2.0 stands out compared to the two other current databases of moonlighting proteins MoonProt (14) and Multitask-II (12) because MoonDB 2.0 combines curated and predicted proteins. We consider our dataset to be more comprehensive as well as highly complementary to other available databases. Whereas other databases are dependent on the available literature, and thus limited to providing information which is already known, our dataset of predictions goes beyond current propositions of moonlighting and provides novel candidates. EMF prediction is large-scale and detection does not require *a priori* knowledge besides protein interactions and GO term annotations.

Importantly, as protein interactomes and GO term annotations of model organisms will continue to grow towards completion in the following years, MoonDB will be updated every year with EMF predictions made from the latest interactomes and GO term annotations. Both power and reliability will progressively increase with future releases. This will be particularly important in the cases of mouse, which possesses high-quality GO term annotations (average of ~19 GO terms per protein), but an incomplete protein interactome (<15% of the proteome covered), as well as in fly, whose interactome is better covered (>40% proteome), but GO term annotations are available for less than half of the interactome. Consequently, only few EMF proteins in mouse and fly were detected with our method. However, 5 out of 14 mouse EMF proteins are orthologs of human EMF proteins, suggesting that even when data is limited, the EMF proteins predicted are reliable. Indeed, the ability for genes to be multifunctional is conserved across orthologs of different organisms (15) and some orthologous proteins are known to have moonlighting functions in different organisms (16). Notably, orthologs were also found between human and worm (UBE2I/ubc-9 and SUMO1; 2 out of 6 MoonDB 2.0 worm entries) and even between the distant human and yeast species (SKP1 gene), although our method does not use ortholog relationships for EMF prediction. Together, these findings further underline the quality of our predictions and designates

MoonDB 2.0 as a valuable data repository for one interested in studying the extreme multifunctionality and moonlighting of proteins, possibly across species.

We believe that MoonDB is of interest not only to bioinformaticians working on multifunctionality, but also to any biologist who may profit from knowing whether their protein of study is likely to perform unexpected functions aside from the ones generally known. Due to the high frequency of EMF proteins involved in multiple diseases, often in comorbidity (13), the extensive functional information provided in MoonDB 2.0 is of interest to help designing therapies that are aware of the several functions of the protein. Importantly, MoonDB 2.0 is now cross-referenced in the UniProt Knowledgebase (UniProtKB) (11). We consider that this greatly magnifies the exposure of our database to the general scientific community, as UniProtKB is the reference database for protein-related data and widely used by biologists, biochemists, bioinformaticians and others.

AVAILABILITY

MoonDB 2.0 is freely available at http://moondb.hb.univ-amu.fr/. Files containing EMF protein lists for each species, as well as the protein-protein interaction networks used in this study are freely available for download in MoonDB 2.0, and can be used in accordance with the GNU Public License and the license of primary data sources.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

FUNDING

The project leading to this publication has received funding from Excellence Initiative of Aix-Marseille University - A*MIDEX, a French “Investissements d’Avenir” programme (to CB).

ACKNOWLEDGEMENTS

We would like to thank Zacharie Menetrier for creating the MoonDB 2.0 logo, Benoit Ballester for tips regarding biological databases and Andreas Zanzoni for critically reading the manuscript and testing the database.

REFERENCES

1. Jeffery,C.J. (1999) Moonlighting proteins. Trends Biochem. Sci., 24, 8–11.

2. Chapple,C.E., Robisson,B., Spinelli,L., Guien,C., Becker,E. and Brun,C. (2015) Extreme multifunctional proteins identified from a human protein interaction network. Nat. Commun., 6, 7412.

3. Chapple,C.E. and Brun,C. (2015) Redefining protein moonlighting. Oncotarget, 6, 16812–16813.

4. Volz,K. (2008) The functional duality of iron regulatory protein 1. Curr. Opin. Struct. Biol., 18, 106–111.

5. Jeffery,C.J. (2014) An introduction to protein moonlighting. Biochem. Soc. Trans., 42, 1679–1683.

6. Jeffery,C.J. (2015) Why study moonlighting proteins? Front. Genet., 6, 211.

7. Chapple,C.E., Herrmann,C. and Brun,C. (2015) PrOnto database: GO term functional dissimilarity inferred from biological data. Front. Genet., 6, 200.

8. Becker,E., Robisson,B., Chapple,C.E., Guenoche,A. and Brun,C. (2012) Multifunctional proteins revealed by overlapping clustering in protein interaction network. Bioinformatics, 28, 84-90.

9. del-Toro,N., Dumousseau,M., Orchard,S., Jimenez,R.C., Galeota,E., Launay,G., Goll,J., Breuer,K., Ono,K., Salwinski,L., et al. (2013) A new reference implementation of the PSICQUIC web service. Nucleic acids Res., 41, W601–W606.

10. The Gene Ontology Consortium (2017) Expansion of the Gene Ontology knowledgebase and resources. Nucleic acids Res., 45, D331–D338.

11. UniProt Consortium,T. (2018) UniProt: the universal protein knowledgebase. Nucleic acids Res., 46, 2699.

12. Franco-Serrano,L., Hernández,S., Calvo,A., Severi,M.A., Ferragut,G., Pérez-Pons,J., Piñol,J., Pich,Ò., Mozo-Villarias,Á., Amela,I., et al. (2018) MultitaskProtDB-II: an update of a database of multitasking/moonlighting proteins. Nucleic acids Res., 46, D645–D648.

13. Zanzoni,A., Chapple,C.E. and Brun,C. (2015) Relationships between predicted moonlighting proteins, human diseases, and comorbidities from a network perspective. Front. Physiol., 6, 171.

14. Chen,C., Zabad,S., Liu,H., Wang,W. and Jeffery,C. (2018) MoonProt 2.0: an expansion and update of the moonlighting proteins database. Nucleic acids Res., 46, D640–D644.

15. Pritykin,Y., Ghersi,D. and Singh,M. (2015) Genome-Wide Detection and Analysis of Multifunctional Genes. PLoS Comput. Biol., 11, e1004467.

16. Copley,S.D. (2012) Moonlighting is mainstream: paradigm adjustment required. BioEssays: news Rev. Mol. Cell. Dev. Biol., 34, 578–588.

2.3. Prediction of human 3′UTR-protein complex assembly reveals a role in the regulation of protein multifunctionality

3’UTRs are known to influence protein synthesis and the fate of mRNAs. Berkovits & Mayr, in 2015, found that 3’UTRs can also post-translationally affect the function of the protein encoded by the mRNA molecule they belong to (Berkovits and Mayr, 2015). Particularly, 3’UTRs were shown to promote the formation of co-translational protein complexes that interact with the newly synthesised protein, eventually altering its subcellular localisation, in a mechanism dubbed 3′UTR–dependent protein localisation (UDPL). This mechanism has been demonstrated to translocate the CD47 protein to the plasma membrane, but its full prevalence is unknown.

Moonlighting proteins are able to perform several unrelated functions, often promoted by a change of environment such as a change of cellular localisation (Jeffery, 2018). Interestingly, many moonlighting proteins are known to have alternative functions when localised in the plasma membrane, but most of them are translocated to the plasma membrane through yet unknown mechanisms (Amblee and Jeffery, 2015). Christine Brun’s group has previously made efforts to categorise human EMF proteins, describing several distinguishing features of these proteins, including a higher number of protein interactions, expression in a wide number of tissues and a significant involvement in disease (Chapple *et al.*, 2015).

In this thesis, I used the up-to-date set of human EMF proteins from MoonDB 2.0, described in the previous section (Results, section 2.2), and found that these proteins often have more isoforms and longer 3’UTRs than other proteins, thus expanding the feature signature of moonlighting proteins. Moreover, I confirmed that these proteins are more often associated to the plasma membrane, even though most lack N-terminal signals normally associated to plasma membrane translocation. Based on these findings, I study whether a mechanism such as the UDPL could explain the multifunctionality of EMF proteins, particularly through a change in cellular localisation. For this, we predicted the formation of 3’UTR-protein complexes by the 3’UTRs of EMF proteins, based on the largest available experimental datasets of protein-protein and protein-3’UTR interactions. This approach rendered us with more than a thousand possible complexes involving EMF proteins, including complexes potentially associated with mechanisms alike the UDPL. Indeed, this study proposes new ways in which EMF protein localisation, and thus its cellular function, may be affected. Furthermore, our study predicts that 3’UTR-protein complex formation occurs frequently, indicating that 3’UTR-protein complex formation may be a

common phenomenon in human cells and may play other roles beyond cellular localisation (Discussion, section 3.2.2).

Further analysis of the results presented here will confirm whether the 3'UTR-protein complexes predicted could represent cases of regulation through differential RBP binding of alternative 3'UTRs, as in the case of CD47 and UDPL. For this, the RBP binding to the various 3'UTR isoforms (when present) of the mRNAs of EMF proteins will be evaluated, using up-to-date datasets of RBP-3'UTR interactions.

Furthermore, in this study I have predicted that, in general, 3'UTR-protein complexes occur more often than expected by chance. Future work in this project will involve the experimental validation of several predicted 3'UTR-protein complexes potentially involved in the translocation of EMF proteins to the plasma membrane. Indeed, to fully understand the likelihood of 3'UTR-protein complex formation and the accuracy of our predictions, several complexes should be experimentally validated. Experimental methods that label proteins during translation and follow the newly synthesised protein cellular localisation may be used to validate such complexes (Iwasaki and Ingolia, 2017). These methods include fluorescence-activated cell sorting-based assays using amino acid puromycylation (SUnSET method) or fluorescent tags (e.g. FUNCAT method) (Schmidt *et al.*, 2009; Dieterich *et al.*, 2010). By combining these techniques with *in situ* proximity ligation (PLA) assays, it is possible to monitor the translation of a specific target protein (tom Dieck *et al.*, 2015). Such methods have subcellular resolution and can be used to assess the cell surface localisation of proteins. Moreover, these methods could validate the 3'UTR regulation of cellular localisation by comparing the localisation of proteins synthesised from mRNAs differing only in their 3'UTRs. Other proteins involved in the process could further be identified through mass spectrometry. In addition, an alternative way for mRNAs to regulate the cellular localisation of the protein they encode is to employ local translation of the protein (Glock, Heumüller and Schuman, 2017). Interestingly, methods that measure protein synthesis localisation have found translation to occur in the nucleus (David *et al.*, 2012), and 3'UTR isoform alternative usage has been found to affect mRNA translation in axons and dendrites of neuronal cells (Glock, Heumüller and Schuman, 2017).

Finally, we plan to expand the computational search of 3'UTR-protein complexes to the whole human proteome, with the aim of ascertaining the general prevalence of these complexes in the cell, outside the context of multifunctionality.

Ribeiro, DM, Teixeira, A, Zanzoni, A, Spinelli, L, Brun, C. Prediction of human 3′UTR-protein complex assembly reveals a role in the regulation of protein multifunctionality. (in preparation)

Supplementary material is available on the *Appendix III*

Prediction of human 3′UTR-protein complex assembly reveals a role in the regulation of protein multifunctionality

Authors: Diogo M. Ribeiro[1], Adrien Teixeira[1], Andreas Zanzoni[1], Lionel Spinelli[1], Christine Brun[1,2]

1. Aix-Marseille Université, Inserm, TAGC U1090, Marseille, France
2. CNRS, Marseille, France

Abstract

Moonlighting proteins are a subset of multifunctional proteins that perform multiple unrelated functions. The several functions of moonlighting proteins can be regulated in space and time in diverse ways, such as by a change in cellular localisation. Indeed, many moonlighting proteins perform different functions when localised to the plasma membrane, but most are translocated by unknown mechanisms. Recently, a mechanism termed 3′ UTR–dependent protein localisation (UDPL) demonstrated the ability of alternative 3’UTRs in regulating the cellular localisation of newly synthesised proteins through the co-translational formation of 3’UTR-protein complexes. This mechanism has been demonstrated to translocate the CD47 protein to the plasma membrane, but its full prevalence is unknown. Here, we set out to decipher the extent of 3'UTR-protein complex formation in human proteins and evaluate their role in regulating cellular localisation and multifunctionality. For this, we used 238 computationally identified ‘extreme multifunctional’ (EMF) proteins, moonlighting protein candidates, and revealed that mRNAs encoding these proteins have a high number of alternative 3’UTR isoforms. Using large-scale protein-protein and RBP-3’UTR interaction networks, we comprehensively predicted all the 3'UTR-protein complexes involving EMF proteins plausible to be formed. We identified 1557 possible 3'UTR-protein complexes formed by several hundred distinct protein components, involving 140 EMF proteins, indicating that 3’UTR-protein complex formation is a common phenomenon in human cells. Notably, these complexes include 42 EMF proteins (out of 140), including the Alpha-enolase, which have been found in the plasma membrane but lack N-terminal translocation signals, hinting that the UDPL mechanism may be widely employed in moonlighting protein translocation to the plasma membrane.

Introduction

Constructing a complex organism does not require a large number of genes. Rather, organism complexity is provided by the ensemble of all available functions and their regulation. Protein multifunctionality, like alternative splicing, allows cells to make more with less. Moonlighting proteins are a subset of

multifunctional proteins that perform multiple unrelated functions (Piatigorsky and Wistow, 1989; Jeffery, 1999). A well-studied example of a moonlighting protein is the human aconitase, an enzyme of the tricarboxylic acid cycle (TCA cycle) that also functions as a translation regulator, upon a iron-dependent conformational change (Volz, 2008).

Regardless of their importance, the moonlighting functions of proteins have usually been identified by serendipity, since clear procedures to identify secondary functions are inexistent (Copley, 2012). Recently, we have made available in MoonDB 2.0 (http://moondb.hb.univ-amu.fr/; Ribeiro, Briere, Bely, Spinelli & Brun, in revision) a set of 238 human moonlighting protein candidates – termed “extreme multifunctional” (EMF) proteins – that were computationally identified through a large-scale approach that analyses protein-interaction networks and protein annotations that we previously proposed (Chapple et al., 2015). Previously, we have demonstrated that this group of proteins is characterized by specific features, constituting a signature of extreme multifunctionality (Chapple et al., 2015). For example, within a protein interactome, a typical EMF protein is likely to have a high number of protein partners and to be central to the network. Moreover, EMF proteins contain more short linear motifs (SLiMs) than other proteins (Chapple et al., 2015). These short conserved sequences are mostly located in structurally disordered regions and can mediate transient interactions and be used as molecular switches between functions (Perkins et al., 2010). In addition, EMF proteins are more likely to be involved in multiple diseases (Zanzoni, Chapple and Brun, 2015) and to be expressed ubiquitously, suggesting that they can perform alternative functions in different tissues (Chapple et al., 2015).

The manner in which the distinct functions of moonlighting proteins can be performed and are regulated are largely unknown. However, in some cases the multiple functions of moonlighting proteins have been found to be performed in different tissues or cellular locations, sometimes associated to a change in their interaction partners, conformation or oligomeric states (Jeffery, 2014). Indeed, the presence of a moonlighting protein in different cellular compartments (e.g. nucleus, cytoplasm, plasma membrane) has been found in many cases to be responsible for a change in function (Ostrovsky de Spicer and Maloy, 1993; Jeffery, 2018). Several intracellular chaperones, enzymes in glycolysis and citric acid (TCA) cycle pathways, as well as other ‘housekeeping’ proteins, have been found to function as cell surface receptors (Amblee and Jeffery, 2015; Jeffery, 2018). For instance, the Hyaluronan-mediated motility receptor (HMMR/RHAMM) protein acts intracellularly as a mitotic-spindle or centrosomal protein (Maxwell et al., 2003; Joukov et al., 2006) in normal cells, whereas extracellularly, RHAMM is a hyaluronan-binding protein and partners with the CD44 cell-surface protein, controlling signalling through RAS proteins (Maxwell, McCarthy and Turley, 2008) in tumour cells. Interestingly, the RHAMM protein lacks a membrane spanning domain or other export signals such as a N-terminal signal peptide, in contrast to many other cell-surface receptors (Simpson, Mateos and Pepperkok, 2007). Notably, out of 30 different

multi-species moonlighting proteins having different functions intracellularly and on the cell surface, none contained any signal or motif for cell surface targeting, such as an N-terminal signal peptide or a LPXTG motif. This suggests their cellular localisation may be regulated by a yet unknown mechanism (Amblee and Jeffery, 2015).

In a breakthrough work, Berkovitz & Mayr described in 2015 (Berkovits and Mayr, 2015) a novel plasma membrane translocation mechanism, termed UTR-dependent protein localisation (UDPL). This mechanism involves the interaction between 3' untranslated regions (UTR) and RNA-binding proteins (RBPs) during translation, facilitating the formation of protein complexes that interact with the nascent peptide chain (Berkovits and Mayr, 2015; Mayr, 2016, 2017). In this manner, 3'UTRs were shown to affect the function of its cognate proteins, without recourse to amino acid changes. The relationship between alternative 3'UTRs, subcellular localisation and protein complex formation has been demonstrated in detail for CD47, a cell-surface protein involved in a range of cellular processes, including apoptosis, adhesion, migration and phagocytosis (Soto-Pantoja, Kaur and Roberts, 2015). Due to the UDPL mechanism, whereas the CD47 protein translated from a short 3'UTR-mRNA is retained in the endoplasmic reticulum, the protein translated from the long 3'UTR-mRNA localises to the plasma membrane. This is achieved through the recruitment by the long 3'UTR-mRNA of specific protein partners (SET protein and RAC1), necessary for addressing the CD47 protein to the plasma membrane. The formation of this complex is mediated by the HuR RBP (also known as ELAVL1), by recognising a binding site on the long 3'-UTR that is absent from the short one (Berkovits and Mayr, 2015).

It is thought that the UDPL mechanism has the potential to be a widespread trafficking mechanism for membrane proteins (Berkovits and Mayr, 2015). Moreover, it has been proposed that through this mechanism alternative 3'UTRs could play a role in mediating the multifunctionality of proteins (Mayr, 2017). However, so far, this mechanism has only been proposed for a few other cell-surface proteins (CD44, ITGA1 and TNFRSF13C) and there is a need to interrogate its full prevalence and determine whether the formation of 3'UTR-protein complexes is a major contributor to the diversification of protein function.

In this study, we set out to reveal the extent of 3'UTR-protein complex formation in human proteins and determine its role in cellular localisation and multifunctionality by investigating EMF proteins (moonlighting protein candidates). We first determined that mRNAs encoding EMF proteins have longer and more 3'UTR isoforms, and that EMF proteins are present in more cellular locations (including the plasma membrane) than other groups of proteins, therefore confirming that EMF proteins represent a pertinent model system to study UDL. We then predict all 3'UTR-protein complexes plausible to be formed with the 238 EMF proteins, using large-scale protein-protein and RBP-3'UTR interaction networks. With this approach, we identified more than a thousand possible 3'UTR-protein complexes on more than 140 EMF

proteins, indicating that formation of 3'UTR-protein complexes may be a common phenomenon. Moreover, among the 140 EMF proteins, we identify 42 that have been found in the plasma membrane but lack N-terminal translocation signals. Several of these proteins are predicted to participate to 3'UTR-protein complexes with proteins involved in protein transport, suggesting the UDPL mechanism or similar mechanisms may be widely employed in moonlighting protein translocation to the plasma membrane.

Experimental Procedures

Protein-protein interaction network, EMF proteins and protein groups

Predicted human extreme multifunctional (EMF) proteins (238 proteins) and a human protein-protein interaction (PPI) network (14046 proteins, 92348 interactions) were downloaded from MoonDB 2.0 (http://moondb.hb.univ-amu.fr/; Ribeiro, Briere, Bely, Spinelli & Brun, in revision) on January 2018 (Chapple et al., 2015). The human PPI network was constructed by interactions retrieved from the PSICQUIC web service on January 2018, as described in (Chapple et al., 2015). This PPI network does not contain interactions between the same protein ('self interactions'). Network modules from the PPI network were extracted using OCG (Becker et al., 2012), a clustering algorithm that allows proteins to belong to more than one cluster. EMF, 'multi-clustered' and 'mono-clustered' protein groups were then identified as described in (Chapple et al., 2015)EMF proteins (238 proteins) are proteins that belong to two or more network modules whose GO term annotations (Biological Processes) contain at least two terms that are dissimilar to each other according to PrOnto (Chapple, Herrmann and Brun, 2015). Control groups are formed by 'mono-clustered' proteins that belong to only one network module (10468 proteins) and 'multi-clustered' proteins that belong to more than one network module annotated to similar functions, therefore, not EMF proteins (3340 proteins). Analysis involving the 'proteome' protein group used a human proteome (20349 proteins) retrieved from UniProt ('reviewed' proteins only) on June 2018 (UniProt Consortium, 2018).

Datasets of 3'UTRs and polyadenylation sites

Ensembl v90 spliced 3'UTR sequences for all human transcripts were downloaded from the Ensembl BioMart service (Kinsella et al., 2011). The maximum 3'UTR length was calculated for each protein in the human proteome (UniprotKB AC) by selecting the longest 3'UTR among all transcripts encoding for a certain protein. Genome-wide polyadenylation sites for human were downloaded from APADB v2 (Müller et al., 2014) as well as PolyASite version r1.0 (Gruber et al., 2016), on December 2017. APADB polyadenylation sites per kb were calculated for proteins produced from transcripts with 3'UTRs longer than 1000 nt, taking into account the length of the longest 3'UTR. For PolyASite, polyadenylation sites on

terminal-exon “TE” category were considered. Gene names and Ensembl transcript IDs were converted to UniprotKB AC using the Uniprot ID mapping tool (UniProt Consortium, 2018).

RBP-3’UTR interaction network

Interactions between RBPs and 3’UTRs were retrieved from the Atlas of UTR Regulatory Activity (AURA) v2.4.3 database (AURAlight dataset) on January 2018 (Dassi et al., 2014). The AURA database contains interactions between 3’UTRs and RBPs collected and mapped from various experiments, including several types of cross-linking and immunoprecipitation (CLIP) methods. Gene and coding-transcript identifiers were mapped to reviewed UniprotKB ACs using UniProt ID cross-referencing files (HUMAN_9606_idmapping.dat) (UniProt Consortium, 2018). Only interactions involving proteins present in the PPI network were used. In total, 469266 interactions between 201 RBPs and the 3’UTRs of 11494 proteins were used.

Prediction of 3’UTR-protein complexes

3’UTR-protein complexes were predicted with in-house Python v2.7 scripts using protein-protein and RBP-3’UTR interaction networks described above. RBP-3’UTR interactions were converted to RBP-’nascent’ protein interactions through the correspondence between the 3’UTR’s mRNA and the protein encoded by the mRNA, using UniProt ID cross-referencing files (HUMAN_9606_idmapping.dat) (UniProt Consortium, 2018). Each 3’UTR-protein complex includes: (i) an interaction between the RBP and the nascent protein (based on the RBP-3’UTR interaction dataset), (ii) an interaction between the intermediate protein (i.e. the protein which interact with both the RBP and the nascent protein) and the nascent protein, (iii) an interaction between the intermediate protein and the RBP. We only considered the presence of one intermediate protein, and complexes formed without any intermediate protein were not examined (i.e., RBP interacting directly with the nascent protein). Since the PPI network used does not contain self interactions, the intermediate protein must be different to the nascent and the RBP. 3’UTR-protein complexes were detected for EMF, multi-clustered and mono-clustered groups of nascent proteins. Only proteins with (i) protein-protein interactions, (ii) 3’UTR-RBP interactions and (iii) presence in at least one HPA tissue (see below) were liable to be assessed for 3’UTR-protein complexes as ‘nascent’ proteins. Overall, these included 7373 nascent proteins, as well as 173, 2078 and 5122 proteins on the EMF, multi-clustered and mono-clustered protein groups, respectively. Overall, 8260 and 157 proteins formed the intermediate and RBP background sets, respectively. 3’UTR-protein complexes were further filtered according to protein tissue presence, as described below.

Protein tissue presence filter

Tissue protein presence from Human Protein Atlas (HPA) version 18 (Jan-2018) (Uhlén et al., 2015) was used to filter 3'UTR-protein complexes. This dataset contains data on 58 normal tissues. Information on cell type associated to tissue names was not used in this study. Gene names and Ensembl Gene IDs were converted to 13044 reviewed UniprotKB AC using the Uniprot ID mapping tool (UniProt Consortium, 2018). Proteins with reliability score (level of reliability of the protein expression pattern) indicated as 'uncertain' and proteins with presence level 'not detected' were excluded. We only considered 3'UTR-protein complexes where all proteins of the complex are present in at least one of the 58 tissues.

Proteins localised in plasma membrane

Plasma membrane proteins were retrieved from two datasets: (i) UniProt (UniProt Consortium, 2018), querying reviewed Homo sapiens proteins with the GO term 'plasma membrane' (GO:0005886) (4602 proteins) and (ii) plasma membrane proteins experimentally detected by HPA version 18 (Uhlén et al., 2015), querying for the subcellular locations "plasma membrane" and "cell junctions" (1734 genes mapped to 1776 UniProtKB ACs using the UniProt ID mapping tool). Note that both datasets include proteins that are integral to the plasma membrane (e.g. membrane receptors) as well as peripheral membrane proteins that may attach to integral membrane proteins or penetrate the peripheral regions of the membrane (e.g. receptor-interacting proteins). Information on the presence or absence of signal peptide and transmembrane domains was obtained from UniProt on June 2018 (UniProt Consortium, 2018). The set of 7373 nascent proteins liable to form 3'UTR-protein complexes (i.e. having protein-protein and protein-RNA interactions, as well as present in HPA), even though not enriched in plasma membrane proteins, contain a higher proportion of proteins localised in plasma membrane without a signal peptide or transmembrane domains than the proteome (51.0% using UniProt data; 835 out of 1636 plasma membrane proteins). Thus, to avoid potential biases, statistical comparisons were done against this set of proteins instead of the proteome in plasma membrane-related analysis.

Results

3'UTRs of mRNAs encoding extreme multifunctional proteins are longer and more diverse

The usage of alternative 3'UTRs has been found to regulate tissue-specific expression and subcellular localisation of proteins (Lianoglou et al., 2013; Berkovits and Mayr, 2015). Correspondingly, the several functions of moonlighting proteins have been found to be regulated by their tissue-specific expression and subcellular localisation, hinting that 3'UTRs could play a role in moonlighting protein regulation (Jeffery, 2014). Thus, we first set out to determine if moonlighting proteins often display alternative 3'UTRs. For this, we used the MoonDB 2.0 database that contains extreme multifunctional (EMF) proteins, moonlighting protein candidates, identified from protein-protein interaction (PPI) networks and Gene

Ontology (GO) annotations (see Experimental Procedures) (Chapple et al., 2015). We recently updated this database, which now contains 238 human EMF proteins (Ribeiro, Briere, Bely, Spinelli & Brun, in revision).

Here, using 3'UTR models from the Ensembl database (see Experimental Procedures), we have found that mRNAs encoding EMF proteins have significantly longer 3' untranslated regions (3'UTRs) compared to the mRNAs of (i) all human proteins ('proteome'), (ii) proteins in the interactome that belong to several OCG protein clusters but are not considered EMF proteins due to lack of dissimilar GO terms ('multi-clustered'; 3340 proteins), (iii) proteins in the interactome that belong to only one OCG protein cluster ('mono-clustered'; 10468) (Figure 1a). Moreover, using polyadenylation sites from APADB and PolyASite databases (Müller et al., 2014), we found that mRNAs encoding EMF proteins bear a higher number of alternative polyadenylation (APA) sites in 3'UTRs than mRNAs encoding the other sets of proteins (Supplementary Material Figure S1). This trend is still observed when accounting for the previously observed differences in 3'UTR length between the protein groups, by calculating the number of APA sites per kb of 3'UTR (Figure 1b).

Consistent with the previous findings, when counting transcript models with distinct 3'UTRs using Ensembl annotations (Kinsella et al., 2011), EMF proteins have significantly more 3'UTR isoforms than the proteome, as well as multi-clustered or mono-clustered proteins (Mann Whitney U, two-sided FDR < 5%, after applying the Benjamini-Hochberg procedure; Supplementary Material Figure S2).

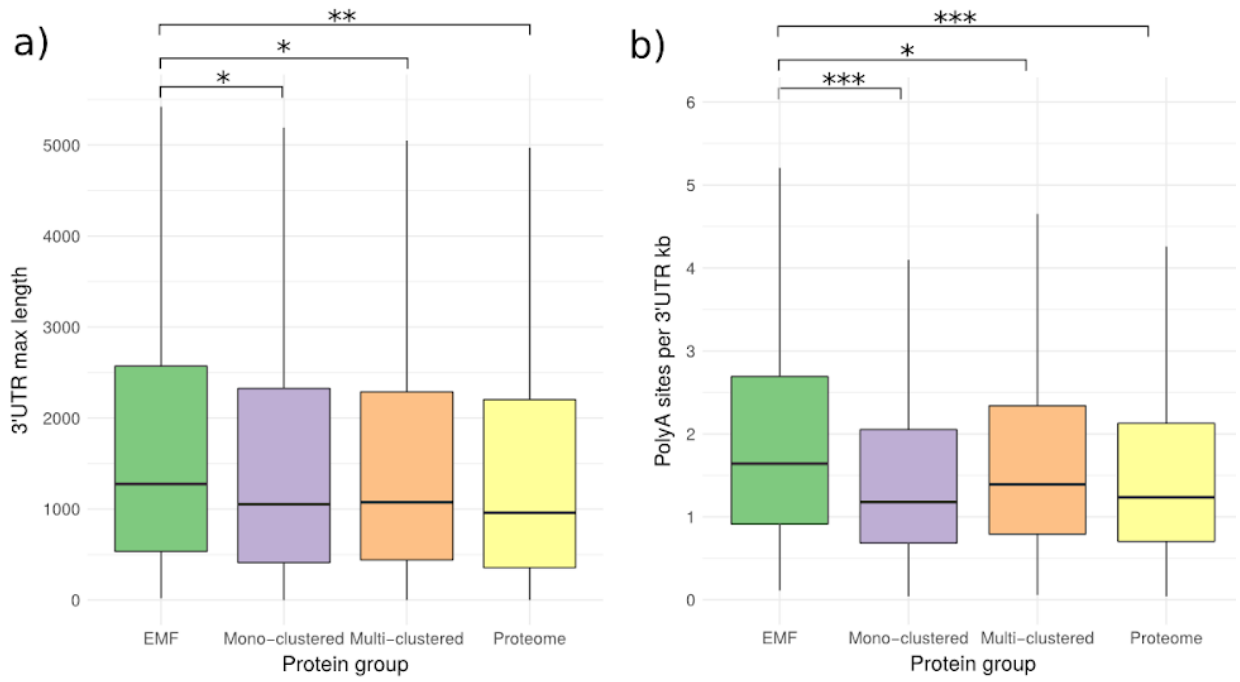


Figure 1. 3'UTR-related features of EMF proteins. (a) Comparison of maximum 3'UTR lengths. **(b)** Number of

polyadenylation (polyA) sites per kb of 3'UTR. Only transcripts with 3'UTRs longer than 1000 nt were considered. 'EMF' represents the MoonDB 2.0 moonlighting protein candidates. 'Mono-clustered' represents the proteins in the interactome that belong to only one protein cluster. 'Multi-clustered' proteins in the interactome that belong to several protein clusters but are not considered EMF proteins. Mann-Whitney U tests were performed to test for statistical significance. The Benjamini-Hochberg procedure was applied for multiple test correction. Significance: '*' indicates a FDR < 0.05; '**' indicates a FDR < 0.01; '***' indicates a FDR < 0.001.

The prevalence of transcript isoforms with distinct 3'UTR have been found to differ between cell types and developmental stages, suggesting that the production of transcripts differing in 3'UTRs helps to achieve tissue- or developmental- specificity (Ulitsky et al., 2012; Lianoglou et al., 2013). EMF proteins have both longer 3'UTRs and more alternative 3'UTR isoforms, potentially produced by APA events. Together, these results suggest that mRNAs encoding EMF proteins are more likely to be regulated by their 3'UTRs than other protein groups.

EMF proteins are present in more cellular locations, including the plasma membrane

Moonlighting proteins have been found to perform different functions when localised in different cellular compartments, such as the bacterial PutA and the Alpha-enolase (ENO1 gene) proteins (Ostrovsky de Spicer and Maloy, 1993; Díaz-Ramos et al., 2012). Using 'Cellular Component' (CC) GO term annotations of EMF proteins, we found that EMF proteins are associated to significantly more CC GO terms than other groups of proteins (Figure 2a). Indeed, on average a EMF protein is associated to 7.8 CC GO terms, whereas the proteome average is 4.0 CC GO terms.

Next, given the fact that many moonlighting proteins have been found to perform moonlighting functions when associated to the plasma membrane, as in the case of the RHAMM receptor, we researched if EMF proteins are often associated to the plasma membrane (Maxwell, McCarthy and Turley, 2008; Amblee and Jeffery, 2015). Indeed, using proteome-wide GO term annotations to plasma membrane (GO:0005886), we found that 65 out of 238 (27.3%) EMF proteins have been found in the plasma membrane, significantly more than expected when compared to the human protein interactome (14046 proteins, see Experimental Procedures) (two-sided Fisher's Exact Test, $P = 1.22 \times 10\text{-}2$; Figure 2b). This enrichment was not found for 'Multi-clustered' or 'Mono-clustered' proteins (two-sided Fisher's Exact Test, P = 0.59 and P = 0.2, respectively). These results suggest that the functions of EMF proteins are prone to be affected by the cellular localisation of the protein, particularly an association to the plasma membrane.

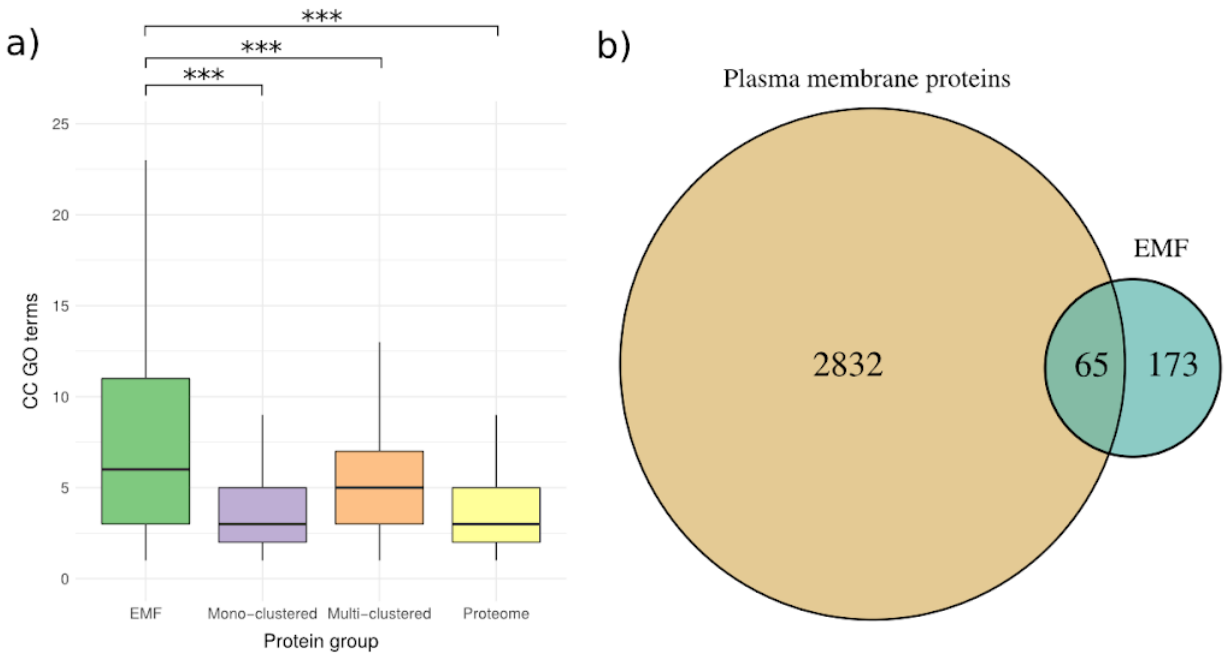


Figure 2. Cellular localisation of EMF proteins. (a) Comparison of number of Cellular component (CC) GO term annotations between protein groups. **(b)** Venn diagram of EMF proteins annotated as being localised in plasma membrane, using UniProt GO term annotations, compared to all the proteins in the protein interactome with the same annotation. Mann-Whitney U tests were performed to test for statistical significance. The Benjamini-Hochberg procedure was applied for multiple test correction. Significance: '*' indicates a FDR < 0.05; '**' indicates a FDR < 0.01; '***' indicates a FDR < 0.001.

Together with longer and more variable 3'UTRs, the EMF protein association with plasma membrane and other cellular locations lead us to propose the EMF proteins as a pertinent model to study the UDPL mechanism, in which 3'UTRs affect the cell-surface localisation of their cognate proteins.

3'UTR-protein complex prediction on EMF proteins

The UDPL mechanism involves the recruitment of RNA-binding proteins (RBPs) to the site of translation by 3'UTRs, which may in turn promote the co-translational formation of protein complexes that interact with the nascent peptide chain (Berkovits and Mayr, 2015; Mayr, 2016, 2017).

Conceptually, the co-translational 3'UTR-protein complex formation may involve the following components: (i) an mRNA with a 3'UTR; (ii) the cognate protein being translated (hereby known as 'nascent' protein); (iii) an RBP able to bind the 3'UTR; (iv) one or more other proteins (hereby called 'intermediate' proteins), that interact with the RBP and the nascent protein. Such protein complexes may thus alter the cellular location, function or in some way regulate the nascent protein.

To investigate the potential occurrence of the UDPL mechanism at a large-scale, we predicted the formation 3'UTR-protein complexes on the EMF proteins (as nascent proteins). For this, we used

large-scale experimental datasets of RBP-3'UTR interactions (AURA database (Dassi et al., 2014)) and protein-protein interactions (MoonDB 2.0; Ribeiro, Briere, Bely, Spinelli & Brun, in revision; see Experimental Methods) and identified sets of co-interacting 3'UTRs, RBPs, nascent and intermediate proteins (Figure 3). To simplify our approach, we only considered the presence of one intermediate protein. Moreover, we only analysed 3'UTR-protein complexes where the RBP, nascent, intermediate proteins are co-present in at least one of the 58 Human Protein Atlas (HPA) normal tissues (Uhlén et al., 2015). Note that the RBP and intermediate proteins are not limited to proteins known to be involved in the UDPL mechanism, but try to cover the whole realm of possibilities.

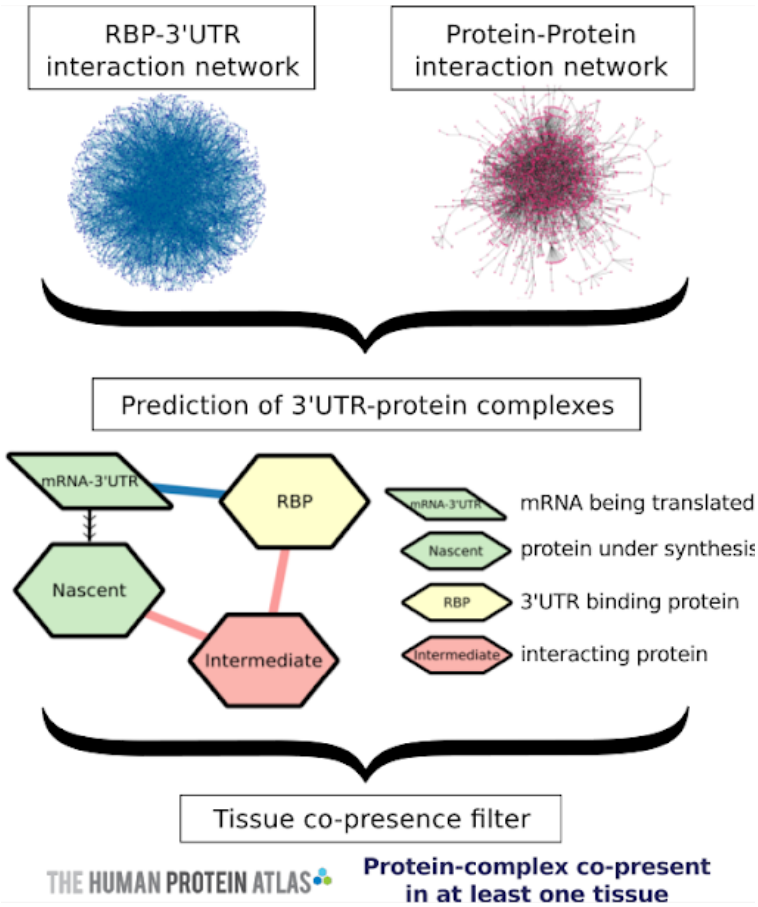


Figure 3. Workflow for the prediction 3'UTR-protein complexes. First, experimental RNA-binding protein (RBP) interactions with 3'-untranslated regions (UTR) of mRNA genes are retrieved from AURA v2 database (Dassi et al., 2014), as well as a large-scale protein-protein interaction network from MoonDB 2.0 (Ribeiro, Briere, Bely, Spinelli & Brun, in revision; see Experimental Methods). Second, 3'UTR complexes are predicted by finding cases in which the 3'UTR of a 'nascent' protein (protein under synthesis) interacts with an RBP, which in turn interacts with another protein ('intermediate') that closes the loop by interacting with the nascent protein. Finally, only 3'UTR complexes where the nascent, RBP and intermediate proteins are present in at least one same tissue were kept (Human Protein Atlas (HPA) (Uhlén et al., 2015), 58 normal tissues).

Using our original approach, we predicted thousands of 3'UTR-protein complexes for 238 EMF proteins.

Notably, we found that 140 EMF proteins (58.8% of the total) may form at least one 3'UTR-protein complex (Table 1, Figure 4). A total of 1557 distinct complexes comprising EMF proteins are predicted to be formed, using a combination of 92 RBPs and 460 interacting intermediate proteins. As a comparison, 1133 (33.9%) and 1372 (13.1%) multi-clustered and mono-clustered nascent proteins form complexes, respectively, a much lower percentage than for EMF proteins.

Table 1. 3'UTR-protein complex prediction on EMF, multi-clustered and mono-clustered protein groups. Percentages under the nascent column are relative to the initial number of proteins in the dataset.

Protein group	Nascents	RBPs	Intermediates	Complex combinations
EMF	140 (58.8%)	92	460	1557
Multi-clustered	1133 (33.9%)	106	739	5990
Mono-clustered	1372 (13.1%)	101	439	3572

The observed cases of the UDPL mechanism involved the HuR RBP binding to 3'UTRs (Berkovits and Mayr, 2015). Here, we predicted that 26 EMF nascent proteins form 31 distinct 3'UTR-protein complexes mediated by the HuR RBP in combination with 8 intermediates (Figure 4). One of these complexes involves the Sorting nexin-1 (SNX1 gene) EMF nascent protein and the Syntenin-1 (SDCBP gene) intermediate protein. Sorting nexin-1 is a membrane-interacting protein involved in intracellular trafficking, including endosome-to-plasma membrane transport for cargo protein recycling (Zhong et al., 2002), and based on MoonDB 2.0, associated also to nucleobase metabolism. Interestingly, the intermediate protein of this complex, Syntenin-1, is also a EMF protein and is found in various cellular locations such as nucleus, cytoplasm, plasma membrane and adherens junction. Syntenin-1 contains two PDZ domains (protein-interacting domains) and one of the several functions of Syntenin-1 is the trafficking of proteins, such as transmembrane proteins, to the plasma membrane (Fernández-Larrea et al., 1999; Philley, Kannan and Dasgupta, 2016). In addition, the intermediate Syntenin-1 and the RBP HuR form complexes with 8 other nascent proteins, including the membrane-related Calmodulin-1 (CALM1), Programmed cell death 6-interacting protein (PDCD6IP) and Abl interactor 2 (ABI2) proteins (Figure 4).

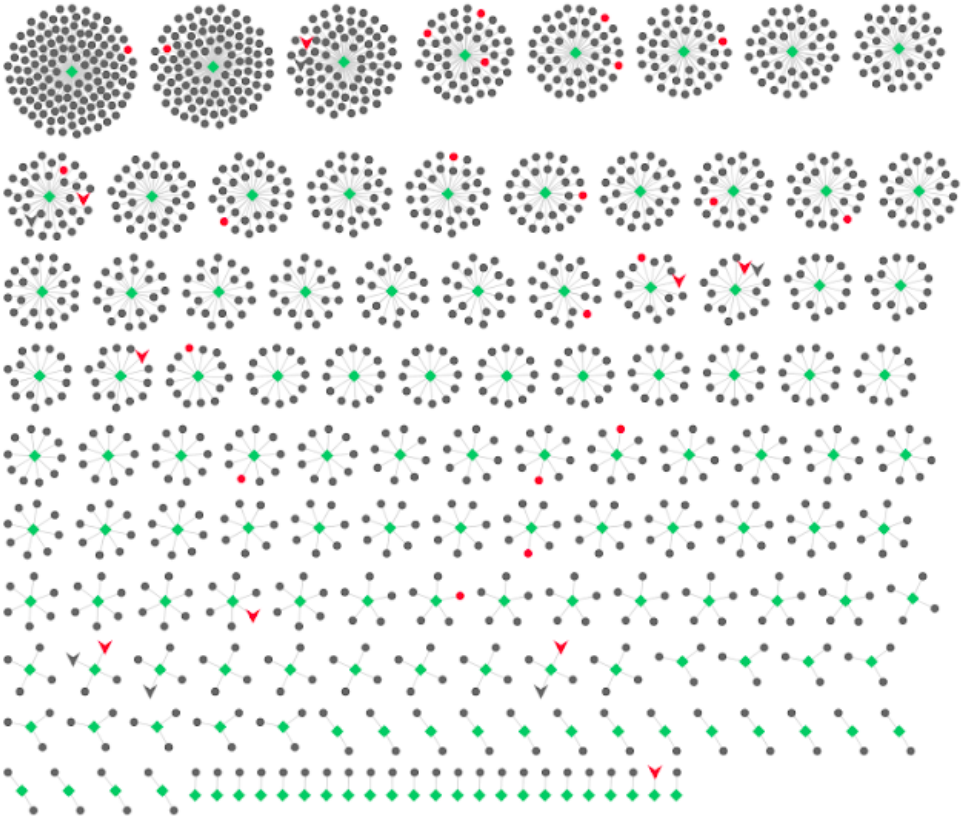


Figure 4. Network representation of EMF protein 3'UTR-protein complexes. The 1557 EMF protein 3'UTR-protein predicted complexes are represented as follows. Each green node represents a EMF nascent protein associated to at least one complex (i.e. the 140 EMF proteins in a complex). Each edge represents the 1557 distinct 3'UTR-protein complex predictions (a combination of RBP and intermediate interacting with the nascent). Pink nodes refer to complexes that include the HuR RBP (31 colored nodes). 'V-shaped' nodes denote complexes which contain Syntenin-1 as an intermediate protein (15 nodes, 9 of them in combination with HuR). The nodes of other complexes are dark grey.

The PRKCA-binding protein (PICK1), like Syntenin-1, is another protein containing PDZ domains and responsible for organising the subcellular localisation of several membrane proteins (Torres et al., 2001; Hirbec et al., 2002). Here we find that the PICK1 participates (as an intermediate protein) in 3'UTR-protein complexes with 7 nascent EMF proteins in 13 complex combinations. Interestingly, one of the nascent EMF proteins in complex with PICK1 is the receptor tyrosine-protein kinase erbB-2 protein (ERBB2 gene), and these two proteins are known to interact through PICK1's PDZ domain (Jaulin-Bastard et al., 2001). Indeed, PDZ domains are associated to protein complex assembly and it has been found that PDZ domains affect the localisation and function of their associated proteins (Hung and Sheng, 2001). Proteins containing PDZ domains may thus be associated to the UDPL mechanism or similar mechanisms. In total, we find 6 PDZ domain-containing proteins predicted to act as intermediate proteins of 3'UTR-protein complexes, including the Syntenin-1, PICK1, Tyrosine-protein phosphatase non-receptor type 3 (PTPN3) and the 'PDZ and LIM domain protein 7' (PDLIM7). Both PICK1 and

Syntenin-1 proteins are part of the top 20 intermediates (out of 460) participating in a higher number of predicted 3'UTR-protein complexes (Supplementary Material Table S1). This list of intermediates includes several other non-PDZ transport-related proteins such as the 'LIM and SH3 domain protein 1' (LASP1) and the Zinc finger protein RFP (TRIM27).

Similarly, when analysing the 20 most commonly found RBPs participating in 3'UTR-protein complexes (Supplementary Material Table S2), we identified the Heterogeneous nuclear ribonucleoproteins C1/C2 (HNRNPC), an RBP involved in 3'-UTR binding, participating in 52 complexes. Interestingly, this protein has been previously associated to the UDPL mechanism, found to be correlated with HuR RBP targets and affecting the CD47 cell-surface localisation, possibly by competing with HuR binding sites (Gruber et al., 2016). The list of RBPs includes several other heterogeneous nuclear ribonucleoproteins known to bind 3'UTRs, as well as HuR and several RNA splicing factors (Supplementary Material Table S2).

Given the large size of the protein-protein and 3'UTR-RBP interaction networks (see Experimental Procedures) used in this study, the co-interaction between three components can occur by chance, without indicating any inherent functionality. To estimate the rate of 3'UTR-protein complex formation by chance, we predicted complexes while shuffling all proteins in the protein-protein interaction network 1000 times. We found that the number of 3'UTR-protein complexes attributed to chance is lower than the one observed with the real interaction network (Supplementary Material Table S3). The fact that we predict more 3'UTR-protein complexes than expected suggests these may indeed be used and selected for in human cells.

Overall, these results suggest that the formation of 3'UTR-protein complexes could be a common mechanism employed by EMF and other proteins, not exclusive to the HuR RBP and the SET protein, but potentially using a wide range of different RBPs and intermediate proteins.

EMF protein 3'UTR-protein complex formation could explain plasma membrane localisation

Given the high prevalence of EMF proteins with predicted 3'UTR-protein complexes, and the EMF protein enrichment in plasma membrane proteins, we decided to explore the relationship between 3'UTR-protein complex formation with unconventional plasma membrane translocation, as observed with the CD47 protein use of the UDPL mechanism (Berkovits and Mayr, 2015).

We first confirmed that the 140 EMF proteins predicted to constitute at least one 3'UTR-protein complex are still highly annotated with a plasma membrane GO term. Indeed, 47 out of the 140 (33.6%) EMF proteins in complexes have been associated to the plasma membrane, a significant enrichment when compared to the human protein interactome (two-sided Fisher's Exact Test, $P = 3.19 \times 10\text{-}4$). Put differently, 47 out of 65 (72.3%) plasma membrane EMF proteins form 3'UTR-protein complexes, whereas

only 140 of a total of 238 (58.8%) EMF proteins form complexes.

N-terminal signal sequences such as the signal peptide or transmembrane domains are common features known to be critical in translocating proteins to the membrane (von Heijne, 2006; Zimmermann et al., 2011), although alternative mechanisms exist, such as the case of the interleukin 1β, the 'High mobility group protein B1' (HMCG1) protein and others (Nickel and Seedorf, 2008). The lack of plasma membrane translocation signals has been observed for dozens of moonlighting proteins, including the human Alpha-enolase (Amblee and Jeffery, 2015). Notably, we found that 42 out of 47 (89.3%) EMF proteins in complex and associated to the plasma membrane are not annotated as having a signal peptide or transmembrane domains. Finding 42 proteins satisfying this criteria is significantly more than expected by chance (two-sided Fisher's Exact Test, P = 1.28 × 10-9; Table 2). Moreover, this is a much higher proportion than the generality of the plasma-membrane associated proteome, where in 4602 proteins associated to the plasma membrane, 1443 of them (31.4%) have no signal peptide or transmembrane domain (see Experimental Procedures). The proportion is also higher than the one found for multi-clustered or mono-clustered proteins (79.4% and 61.7%). In addition, we confirmed these results using an independent set of plasma membrane proteins from the HPA database instead of UniProt GO term annotations (Supplementary Material Table S4) (Uhlén et al., 2015).

Table 2. Numbers of nascent proteins in 3'UTR-protein complexes localised in the plasma membrane and without conventional translocation signals. Percentages denote the proteins retained compared to the previous column. Where indicated, Fisher's exact tests were performed to test for statistical significance using as background the set of 7373 nascent proteins liable to be assessed for 3'UTR-protein complexes (see Experimental Procedures). The Benjamini-Hochberg procedure was applied for multiple test correction. Significance: '*' indicates a FDR < 0.05; '**' indicates a FDR < 0.01; '***' indicates a FDR < 0.001.

Protein group	Nascents in complex	of which, localised in plasma membrane	of which, contain no signal peptide or transmembrane domain
EMF	140 (58.8%)	47 (33.6%)	42 (89.3%) ***
Multi-clustered	1133 (33.9%)	218 (19.2%)	173 (79.4%) ***
Mono-clustered	1372 (13.1%)	238 (17.4%)	147 (61.7%) N.S.

Importantly, the set of 42 EMF proteins includes the Alpha-enolase (ENO1 gene), a well-known moonlighting protein, which functions as a receptor and activator of plasminogen on the cell surface, but is translocated by a yet unknown mechanism (Díaz-Ramos et al., 2012). We found that Alpha-enolase forms 3'UTR-protein complexes with the Actin and Desmin proteins, two components of the intermediate filaments in the cellular cytoskeleton. Interestingly, an association between vesicular trafficking and intermediate filaments including Desmin has been proposed (Jones et al., 2017). Moreover, intermediate

filaments have also shown an association with plasma membrane proteins, such as receptors and adhesion molecules (Jones et al., 2017).

Similarly, we predicted 3'UTR-protein complexes formed with nascent protein HMMR/RHAMM, a moonlighting protein annotated as curated in MoonDB 2.0 (Ribeiro, Briere, Bely, Spinelli & Brun, in revision), also known to function in the plasma membrane but translocated by unknown mechanisms. We predict that HMMR/RHAMM forms a single complex, interacting with the Dynactin subunit 1 (DCTN1 gene) protein intermediate, a protein known to be involved in endoplasmic reticulum (ER) to Golgi transport, providing a link between specific cargos, microtubules and dynein (Ayloo et al., 2014).

Overall, we have found that the vast majority of EMF proteins in 3'UTR-protein complexes and associated to the plasma membrane lack conventional translocation signals. This leads us to believe that 3'UTR-protein complexes could play a role in the translocation of these moonlighting candidates, perhaps through the UDPL mechanism or similar mechanisms involving PDZ domain-containing proteins or other proteins involved in macromolecule transport.

Discussion

The UDPL mechanism to translocate membrane proteins has been described for CD47 and proposed for several other proteins, but its prevalence is otherwise unknown. However, other cases of 3'UTRs co-translationally recruiting other sets of proteins have been described, such as the signal recognition particle (SRP) recruitment by 3'UTRs (Chartron, Hunt and Frydman, 2016). Moreover, a study in yeast showed that more than 12 out of the 31 sampled proteins copurified with mRNAs that encode for their protein interactors, suggesting that the co-translational recruitment of proteins that interact with the protein encoded by the mRNA may be a widespread phenomenon (Duncan and Mata, 2011).

Our study allows us to estimate and decipher the prevalence of an ill-known regulation mechanism and evaluate its role on protein multifunctionality. Indeed, with our large-scale approach, we predicted the formation of thousands of 3'UTR-protein complexes, suggesting this may be a common phenomenon in human cells. The lack of conventional signals for plasma membrane translocation in moonlighting proteins, and the high prevalence for these proteins to form 3'UTR-protein complexes, hint at a role of such complexes in protein translocation, along the lines of the UDPL mechanism. The localisation of proteins in the plasma membrane often requires the influence of additional interacting proteins, including trafficking proteins, chaperones and transcription factors (Sharma et al., 2018). In this study, we have found several 3'UTR-protein complexes in which RBPs interacted with intermediate proteins that are motor proteins or associated to macromolecule transport, as well as transcription factors and proteins involved in other biological processes. Moreover, cases in which 3'UTRs are bound by RBPs that in turn

associate with motor proteins have been previously described for the budding yeast ASH1 mRNA, leading to the transport of the mRNA-protein complex to the bud tip through actin microfilaments (Niedner, Edelmann and Niessing, 2014), as well as for the Drosophila oskar mRNA which is transported along microtubules (Marchand, Gaspar and Ephrussi, 2012; Jambor et al., 2014; Mayr, 2017).

Moreover, the predicted complexes included hundreds of distinct RBPs and effector proteins with diverse functions. Theoretically, depending on the cell type and cellular conditions, each alternative 3'UTR isoform could have its own RBP composition, and since an RBP may interact with several other proteins, each 3'UTR could be processed differently and serve a different function. Indeed, it is predictable that besides affecting the cellular localisation of proteins, the formation of 3'UTR-protein complexes could function to regulate other cellular processes. For example, a recent study showed that co-translational protein complex assembly occurs frequently in eukaryotic cells (Shiber et al., 2018). By performing selective ribosome profiling on 12 hetero-oligomeric Saccharomyces cerevisiae protein complexes, Shiber et al. showed that, for 9 out of these 12 protein complexes, pairs of subunits of the same protein complex interact and co-assemble during their translation. It has been proposed that this protein assembly mechanism may involve interactions with their mRNA molecules, and 3'UTR-protein complex formation, in a model involving RBP binding of 3'UTRs and the RBP recruitment of the partner protein subunit (intermediate protein) to the nascent protein translation site (Mayr, 2018a). Indeed, an alternative 3'UTR of the human E3 ubiquitin ligase BIRC3 has been recently implicated in the assembly of a protein complex involving its cognate nascent protein, IQGAP1 and the Ras-GTPase RALA, thus affecting the function of the nascent protein (Mayr, 2018b).

Notably, as many as 4 out of the 9 S. cerevisiae protein complexes formed co-translationally may involve known moonlighting proteins or proteins found to be multifunctional in yeast (Shiber et al., 2018). These include the 6-phosphofructokinase subunit alpha (PFK1) of the phosphofructokinase complex (manually curated in MoonDB 2.0) (Yuan et al., 1997; Gancedo and Flores, 2008), the EGD2 subunit of the nascent chain associated chaperone (NAC) complex (Kogan and Gvozdev, 2014; Franco-Serrano et al., 2018) and the GluRS and MetRS subunits of the aminoacyl-tRNA synthetase complex (Guo and Schimmel, 2013; Frechin et al., 2014; Shiber et al., 2018). Interestingly, as part of their alternative function as ATP-synthase expression regulators, MetRS and GluRS possess nuclear and mitochondrial localisation signals, respectively, which are only revealed when these proteins are not in complex with each other (and ARC1) (Frechin et al., 2014). It is thus possible that their co-translation complex assembly (Shiber et al., 2018) may regulate their moonlighting function.

Our approach to predict 3'UTR-protein complexes involved the use of large-scale interaction networks. While experimental human protein-protein interaction networks are thought to cover most interacting proteins (Rolland et al., 2014), public interaction databases may not include all the interactions known in

the literature or interactions only discovered recently. Cellular 3'UTR-protein complexes that are formed by proteins and RNA interactions not yet present in experimental interaction datasets are thus missed. Indeed, the interaction between protein SET and CD47 is not present in the protein-protein interaction dataset used here, thus the CD47-HuR-SET protein complex identified in Berkovits & Mayr could not be retrieved (Berkovits and Mayr, 2015). Moreover, current RBP-3'UTR interaction datasets contain data for only a subset of an increasingly growing number of proteins thought to interact with RNAs (Dassi et al., 2014; Hentze et al., 2018). Furthermore, cellular 3'UTR-protein complexes may involve an undefined amount of proteins or other components and indeed be more complex than the complexes predicted here. Particularly, certain 3'UTR-protein complexes may involve more than one intermediate protein in order for the action on the nascent protein to be effective (Berkovits and Mayr, 2015). Lastly, 3'UTR-protein complex mechanisms such as the UDPL use alternative 3'UTRs to regulate the nascent protein. Even though EMF proteins have shown to contain more 3'UTR isoforms than other groups of proteins, in our analysis we considered 3'UTR-protein complex formation on all EMF nascent proteins, regardless of the presence of alternative 3'UTRs.

Finally, in this work we have expanded the feature signature of EMF proteins, by finding that the mRNAs of this group of moonlighting candidate proteins have longer 3'UTRs and more isoforms than those of other protein groups. EMF proteins were also found annotated to more GO term cellular locations and enriched in plasma membrane proteins. Moreover, the set of EMF proteins form more complexes than other protein groups, such as multi-clustered proteins. The number of EMF protein complexes may be related to some of these features, as well as previously identified features, such as higher number of protein interactions. However, it is hard to distinguish technical bias from biological reality, as these proteins may indeed be involved in more interactions in order to be more tightly regulated by diverse mechanisms, such as 3'UTR-protein complex formation. Increased knowledge on moonlighting proteins and their regulation is important because these proteins (i) are implicated in the regulation of the biological processes through their coordination and switch, (ii) contribute to the complexity of the genotype-phenotype relationship by diversifying phenotypes, and (iii) cause unexpected drug side-effects due to their multiple functions. In this study we have deepened our knowledge into how moonlighting proteins may be regulated.

Acknowledgements

We would like to thank Philippe Pierre for fruitful scientific discussions. The project leading to this publication has received funding from Excellence Initiative of Aix-Marseille University - A*MIDEX, a French "Investissements d'Avenir" programme (to CB).

References

Amblee, V. and Jeffery, C. J. (2015) “Physical Features of Intracellular Proteins that Moonlight on the Cell Surface.,” PloS one, 10(6), p. e0130575.

Ayloo, S. et al. (2014) “Dynactin functions as both a dynamic tether and brake during dynein-driven motility.,” Nature communications, 5, p. 4807.

Becker, E. et al. (2012) “Multifunctional proteins revealed by overlapping clustering in protein interaction network.,” Bioinformatics (Oxford, England), 28(1), pp. 84–90.

Berkovits, B. D. and Mayr, C. (2015) “Alternative 3’ UTRs act as scaffolds to regulate membrane protein localization.,” Nature, 522, pp. 363–367.

Chapple, C. E. et al. (2015) “Extreme multifunctional proteins identified from a human protein interaction network.,” Nature communications, 6, p. 7412.

Chapple, C. E., Herrmann, C. and Brun, C. (2015) “PrOnto database: GO term functional dissimilarity inferred from biological data.,” Frontiers in genetics, 6, p. 200.

Chartron, J. W., Hunt, K. C. L. and Frydman, J. (2016) “Cotranslational signal-independent SRP preloading during membrane targeting.,” Nature, 536(7615), pp. 224–228.

Copley, S. D. (2012) “Moonlighting is mainstream: paradigm adjustment required.,” BioEssays: news and reviews in molecular, cellular and developmental biology, 34(7), pp. 578–588.

Dassi, E. et al. (2014) “AURA 2: Empowering discovery of post-transcriptional networks,” Translation. Taylor & Francis, 2(1).

Díaz-Ramos, A. et al. (2012) “α-Enolase, a multifunctional protein: its role on pathophysiological situations.,” Journal of biomedicine & biotechnology, 2012, p. 156795.

Duncan, C. D. S. and Mata, J. (2011) “Widespread cotranslational formation of protein complexes.,” PLoS genetics, 7(12), p. e1002398.

Fernández-Larrea, J. et al. (1999) “A role for a PDZ protein in the early secretory pathway for the targeting of proTGF-alpha to the cell surface.,” Molecular cell, 3(4), pp. 423–433.

Franco-Serrano, L. et al. (2018) “MultitaskProtDB-II: an update of a database of multitasking/moonlighting proteins.,” Nucleic acids research, 46(D1), pp. D645–D648.

Frechin, M. et al. (2014) “Expression of nuclear and mitochondrial genes encoding ATP synthase is synchronized by disassembly of a multisynthetase complex.,” Molecular cell, 56(6), pp. 763–776.

Gancedo, C. and Flores, C.-L. (2008) “Moonlighting proteins in yeasts.,” Microbiology and molecular biology reviews: MMBR, 72(1), p. 197–210, table of contents.

Gruber, A. J. et al. (2016) “A comprehensive analysis of 3’ end sequencing data sets reveals novel polyadenylation signals and the repressive role of heterogeneous ribonucleoprotein C on cleavage and polyadenylation.,” Genome research, 26(8), pp. 1145–1159.

Guo, M. and Schimmel, P. (2013) “Essential nontranslational functions of tRNA synthetases.,” Nature chemical biology, 9(3), pp. 145–153.

von Heijne, G. (2006) “Membrane-protein topology.,” Nature Reviews Molecular Cell Biology. Center for Biomembrane Research and Stockholm Bioinformatics Center, Department of Biochemistry and Biophysics, Stockholm University, SE-10691 Stockholm. gunnar@dbb.su.se: Nature Publishing Group, 7, pp. 909–918.

Hentze, M. W. et al. (2018) “A brave new world of RNA-binding proteins.,” Nature reviews. Molecular cell biology, 19(5), pp. 327–341.

Hirbec, H. et al. (2002) “The PDZ proteins PICK1, GRIP, and syntenin bind multiple glutamate receptor

subtypes. Analysis of PDZ binding motifs.," The Journal of biological chemistry, 277(18), pp. 15221–15224.

Hung, A. Y. and Sheng, M. (2001) "PDZ Domains: Structural Modules for Protein Complex Assembly," Journal of Biological Chemistry, 277(8).

Jambor, H. et al. (2014) "A stem-loop structure directs oskar mRNA to microtubule minus ends.," RNA (New York, N.Y.), 20(4), pp. 429–439.

Jaulin-Bastard, F. et al. (2001) "The ERBB2/HER2 receptor differentially interacts with ERBIN and PICK1 PSD-95/DLG/ZO-1 domain proteins.," The Journal of biological chemistry, 276(18), pp. 15256–15263.

Jeffery, C. J. (1999) "Moonlighting proteins.," Trends in biochemical sciences, 24(1), pp. 8–11.

Jeffery, C. J. (2014) "An introduction to protein moonlighting.," Biochemical Society transactions, 42(6), pp. 1679–1683.

Jeffery, C. J. (2018) "Protein moonlighting: what is it, and why is it important?," Philosophical transactions of the Royal Society of London. Series B, Biological sciences, 373(1738).

Jones, J. C. R. et al. (2017) "Intermediate Filaments and the Plasma Membrane.," Cold Spring Harbor perspectives in biology, 9(1).

Joukov, V. et al. (2006) "The BRCA1/BARD1 heterodimer modulates ran-dependent mitotic spindle assembly," Cell, 127, pp. 539–52.

Kinsella, R. J. et al. (2011) "Ensembl BioMarts: a hub for data retrieval across taxonomic space.," Database: the journal of biological databases and curation, 2011, p. bar030.

Kogan, G. L. and Gvozdev, V. A. (2014) "[Multifunctional protein complex NAC (nascent polypeptide associated complex].," Molekuliarnaia biologiia, 48(2), pp. 223–231.

Lianoglou, S. et al. (2013) "Ubiquitously transcribed genes use alternative polyadenylation to achieve tissue-specific expression.," Genes & development, 27(21), pp. 2380–2396.

Marchand, V., Gaspar, I. and Ephrussi, A. (2012) "An intracellular transmission control protocol: assembly and transport of ribonucleoprotein complexes.," Current opinion in cell biology, 24(2), pp. 202–210.

Maxwell, C. A. et al. (2003) "RHAMM is a centrosomal protein that interacts with dynein and maintains spindle pole stability.," Molecular biology of the cell, 14(6), pp. 2262–2276.

Maxwell, C. A., McCarthy, J. and Turley, E. (2008) "Cell-surface and mitotic-spindle RHAMM: moonlighting or dual oncogenic functions?," Journal of cell science, 121(Pt 7), pp. 925–932.

Mayr, C. (2016) "Evolution and Biological Roles of Alternative 3'UTRs," Trends in Cell Biology. Elsevier Ltd, 26, pp. 227–237.

Mayr, C. (2017) "Regulation by 3'-Untranslated Regions.," Annual review of genetics, 51, pp. 171–194.

Mayr, C. (2018a) "Protein complexes assemble as they are being made.," Nature, 561(7722), pp. 186–187.

Mayr, C. (2018b) "What Are 3' UTRs Doing?," Cold Spring Harbor perspectives in biology.

Müller, S. et al. (2014) "APADB: a database for alternative polyadenylation and microRNA regulation events.," Database: the journal of biological databases and curation, 2014.

Nickel, W. and Seedorf, M. (2008) "Unconventional mechanisms of protein transport to the cell surface of eukaryotic cells.," Annual review of cell and developmental biology, 24, pp. 287–308.

Niedner, A., Edelmann, F. T. and Niessing, D. (2014) "Of social molecules: The interactive assembly of ASH1 mRNA-transport complexes in yeast," RNA Biology. Taylor & Francis, 11(8), pp. 998–1009.

Ostrovsky de Spicer, P. and Maloy, S. (1993) "PutA protein, a membrane-associated flavin dehydrogenase, acts as a redox-dependent transcriptional regulator.," Proceedings of the National

Academy of Sciences of the United States of America, 90(9), pp. 4295–4298.

Perkins, J. R. et al. (2010) “Transient protein-protein interactions: structural, functional, and network properties.,” Structure (London, England: 1993), 18(10), pp. 1233–1243.

Philley, J. V., Kannan, A. and Dasgupta, S. (2016) “MDA-9/Syntenin Control.,” Journal of cellular physiology, 231(3), pp. 545–550.

Piatigorsky, J. and Wistow, G. J. (1989) “Enzyme/crystallins: gene sharing as an evolutionary strategy.,” Cell, 57(2), pp. 197–199.

Ribeiro, Diogo M, Galadriel Briere, Benoit Bely, Lionel Spinelli, Christine Brun. 2018. “MoonDB 2.0: an updated database of extreme multifunctional and moonlighting proteins.” (in revision)

Rolland, T. et al. (2014) “A proteome-scale map of the human interactome network.,” Cell, 159(5), pp. 1212–1226.

Sharma, S. et al. (2018) “Genome-scale identification of cellular pathways required for cell surface recognition.,” Genome research, 28(9), pp. 1372–1382.

Shiber, A. et al. (2018) “Cotranslational assembly of protein complexes in eukaryotes revealed by ribosome profiling.,” Nature.

Simpson, J. C., Mateos, A. and Pepperkok, R. (2007) “Maturation of the mammalian secretome.,” Genome biology, 8(4), p. 211.

Soto-Pantoja, D. R., Kaur, S. and Roberts, D. D. (2015) “CD47 signaling pathways controlling cellular differentiation and responses to stress.,” Critical reviews in biochemistry and molecular biology, 50(3), pp. 212–230.

Torres, G. E. et al. (2001) “Functional interaction between monoamine plasma membrane transporters and the synaptic PDZ domain-containing protein PICK1.,” Neuron, 30(1), pp. 121–134.

Uhlén, M. et al. (2015) “Tissue-based map of the human proteome,” Science. American Association for the Advancement of Science, 347(6220), p. 1260419.

Ulitsky, I. et al. (2012) “Extensive alternative polyadenylation during zebrafish development.,” Genome research, 22(10), pp. 2054–2066.

UniProt Consortium, T. (2018) “UniProt: the universal protein knowledgebase.,” Nucleic acids research, 46(5), p. 2699.

Volz, K. (2008) “The functional duality of iron regulatory protein 1.,” Current opinion in structural biology, 18(1), pp. 106–111.

Yuan, W. et al. (1997) “Glucose-induced microautophagy in Pichia pastoris requires the alpha-subunit of phosphofructokinase.,” Journal of cell science, 110 (Pt 16), pp. 1935–1945.

Zanzoni, A., Chapple, C. E. and Brun, C. (2015) “Relationships between predicted moonlighting proteins, human diseases, and comorbidities from a network perspective.,” Frontiers in physiology, 6, p. 171.

Zhong, Q. et al. (2002) “Endosomal localization and function of sorting nexin 1.,” Proceedings of the National Academy of Sciences of the United States of America, 99(10), pp. 6767–6772.

Zimmermann, R. et al. (2011) “Protein translocation across the ER membrane.,” Biochimica et biophysica acta, 1808(3), pp. 912–924.

2.4. Predicted protein-RNA interactions reveal distinct post-transcriptional regulatory patterns

To respond efficiently to environmental signals, cells have evolved mechanisms that coordinate the expression of functionally-related proteins in space and time (Keene, 2007). Prokaryotes such as *E. coli* often organize genes into DNA operons, in which genes that operate in the same pathway are physically linked (Jacob and Monod, 1961). Since most eukaryotic genes are monocistronic, each having a promoter and transcription terminator, the coordination of expression is achieved and influenced by other processes, including the chromatin structure, RNA processing and protein synthesis (Imig, Kanitz and Gerber, 2012). In fact, in eukaryotes, post-transcriptional coordination can be mediated by RBPs that interact with certain groups of mRNAs coding for functionally-related proteins, affecting their splicing, localisation, stability and translation (Keene, 2007). These RBPs interact with regulatory elements within mRNAs, most lying in the non-coding regions, but some residing in the coding part. Such RNP assemblies are called 'RNA regulons' and constitute a conserved feature of the post-transcriptional regulation in eukaryotes (Scherrer *et al.*, 2011). For example, in yeast, RNA regulons are found among five RBPs of the PUF family, such as the PUF3 binding of mRNAs coding for mitochondrial proteins (Gerber, Herschlag and Brown, 2004). In mammals, groups of mRNAs coding for proteins involved in inflammation, cell cycle and other processes have been found to be regulated by RBPs such as ELAVL1, HNRNPL and PUM1 (Anderson, 2010; Blackinton and Keene, 2014). Nevertheless, a deeper understanding of the extent of post-transcriptional regulation by RNA regulons is yet lacking, and their study has been mostly centered on specific RBPs or biological processes.

The work presented in this section aims at assessing the RNA regulon theory transcriptome-wide, determining which biological functions may be regulated in such a way, and through which factors. For this, Zanzoni *et al.* applies novel methods to infer the post-transcriptional regulation of human biological pathways and protein complexes mediated by RBPs. Concretely, this work uses large-scale datasets of predicted (by catRAPID) and available experimental (eCLIP results) protein-mRNA interactions, to determine which sets of mRNAs coding for proteins of the same pathway/complex may be regulated by RBPs. This study shows that approximately 10% of the 2977 groups of mRNAs tested may take part in RNA regulons. These comprise mRNAs encoding proteins involved in various biological functions such as chromatin organisation, signaling pathways and DNA transcription. Likewise, we predict that several hundred RBPs take part in a complex regulatory system, where the interaction to specific groups of functionally-related RNAs is either promoted or avoided. Overall, this work charts the functional

regulatory landscape of human RBPs, revealing particular patterns of post-transcriptional regulation of cellular functions.

Zanzoni, A, Spinelli, L, **Ribeiro DM**, Tartaglia GG, Brun, C. Predicted protein-RNA interactions reveal distinct post-transcriptional regulatory patterns. (in preparation)

Supplementary material is available on the *Appendix IV*

Predicted protein-RNA interactions reveal distinct post-transcriptional regulatory patterns

Andreas Zanzoni[1,#], Lionel Spinelli[1], Diogo M. Ribeiro[1], Gian Gaetano Tartaglia[2,3,4], Christine Brun[1,5,#]

[1]Aix-Marseille Univ, INSERM, TAGC, UMR_S1090, Marseille, France;

[2]Centre for Genomic Regulation (CRG), The Barcelona Institute of Science and Technology, Dr Aiguader 88, 08003 Barcelona, Spain;

[3]Universitat Pompeu Fabra (UPF), 08003 Barcelona, Spain;

[4]Institucio Catalana de Recerca i Estudis Avançats (ICREA), 23 Passeig Lluıs Companys, 08010 Barcelona, Spain;

[5]CNRS, Marseille, France.

[#]Address correspondence to: Andreas Zanzoni (andreas.zanzoni@univ-amu.fr) and Christine Brun (christine-g.brun@inserm.fr).

Abstract

Previous studies revealed that the coordinated synthesis of functionally-related proteins can be achieved at the post-transcriptional level by the action of common regulatory molecules, such as RNA-binding proteins (RBPs). Despite the advances in the genome-wide identification of RBPs and their binding transcripts, the protein-RNA interaction space is still largely unexplored, thus hindering a broader understanding of the extent of the post-transcriptional regulation of related coding RNAs. Here, we propose a computational approach that combines protein-mRNA interaction predictions and statistical analyses to generate an inferred regulatory landscape for more than 800 human RBPs and identify the cellular processes that can be regulated at the post-transcriptional level. We show that only 10% of the tested sets of functionally-related mRNAs can be post-transcriptionally regulated. Moreover, we reveal that established and novel RBPs have distinct behaviors in the inferred functional landscape, which could be explained can by their different propensity to be regulated at the post-translational level. We also show that the inferred functional landscape is a useful resource to make new hypotheses on the cellular role of both well-characterized and novel RBPs in the context of human disease.

Introduction

Genome-wide analysis tools have stimulated the study of eukaryotic gene expression programs. These revealed the presence of highly coordinated control mechanisms that guarantee proper spatial and temporal activity of cell's molecular components in response to external cues[1]. While transcription is a significant contributor to this coordinated gene expression, several studies highlighted the discordance between mRNAs levels and protein production[2]. This indicates that the regulation of mRNA transcripts is key to achieve coordinated protein synthesis. Indeed, it has been shown that sets of transcripts coding for functionally related proteins are bound by common regulatory molecules, such as RNA-binding proteins (RBPs) and/or non-coding RNAs, thus forming the so-called RNA regulons[3,4].

Early protein-RNA interaction mapping studies in yeast demonstrated that many RBPs bind specific mRNAs coding for proteins involved in the same biological process (e.g., ribosome biogenesis, chromatin architecture, oxidative phosphorylation) or that are cytotopically related (e.g., cell wall, endoplasmic reticulum, mitochondrion)[5,6]. In mammalian cells, several sets of related mRNAs may be part of RNA regulons as well, *e.g.*, histone mRNAs bound by the stem-loop binding protein (SLBP)[7], transcripts involved in inflammation regulated by the RBPs ELAVL1, HNRNPL and TTP[8], those implicated in DNA damage response and regulated by the RBPs BCLAF1, ELAVL1 and THRAP3[9,10], and mRNAs coding for cell cycle and proliferation factors bound by Dead end protein homolog 1 (DND1) and Pumilio 1 (PUM1) proteins[9].

As this regulatory phenomenon has been observed in different species, RNA regulons represent a conserved feature of the post-transcriptional regulation in eukaryotes[3,4,11]. However, even though RNA regulon perturbations can lead to the onset of neurological diseases and cancers in human[12–14], the combinatorial control of these regulatory circuits exerted by RBPs is rather sketchy[15,16], therefore calling for further scrutiny.

A deeper understanding of the extent of the post-transcriptional regulation of related coding transcripts is subordinate to the availability of experimentally verified protein-mRNA interaction data. Over the last years, studies based on high-throughput methods to detect RNA molecules bound by RBPs, such as RNA immunoprecipitation and CLIP-based techniques[17,18] allowed to identify thousands of protein-RNA interactions. However, these studies have focused on the binding ability of a reduced number of established RBPs in a few cell lines[18], indicating that the protein-RNA interactions space is largely unexplored. Moreover, thanks to the recent development of RNA interactome capture technologies, the catalogue of RBPs has dramatically increased (*e.g.*, [19–24]). Importantly, many of these RBPs lack a known RNA-binding domain and their role in RNA biology has not been characterized yet[22]. In this context,

large-scale computational prediction of protein-RNA interactions can provide a better coverage of the protein-RNA interaction space and improve our understanding of post-transcriptional regulation.

What is the extent of the regulon theory at the coding transcriptome scale? What are the cellular functions regulated at the post-transcriptional level? Can RBPs be classified based on the regulation they exert? In order to answer these questions, we inferred the functional landscape of the post-transcriptional regulation mediated by the human RBPs, by assessing the RNA regulon theory at different levels of organization of the cellular processes, such as biological pathways and protein complexes. For this, we developed and applied an original large-scale approach to identify human cellular processes post-transcriptionally regulated by RBPs, using both predicted and experimentally identified protein-RNA interaction combined with protein-protein interaction network data and statistical analyses. We showed that the post-transcriptional regulation of functionally-related mRNAs by common RBPs only concern 10% of the groups that we tested in the inferred regulatory landscape. Furthermore, we identified 3 groups of RBPs possibly regulating these groups by using different molecular strategies.

Results

A predicted large-scale human RBP-mRNA interaction network

In order to find the cellular processes potentially regulated through the binding of RBPs, we first computed the interaction propensities of 877 experimentally identified human RBPs with a representative set of 13,984 mRNA sequences, covering ~63% of the human protein-coding genes (see Methods), using the *cat*RAPID *omics* algorithm[25] (Figure 1A). This algorithm predicts protein-RNA interactions by exploiting the physicochemical properties of both molecules[26], and has extensively been used and tested on different RNA and protein datasets with good performances [27–30]. We generated more than 12 million protein-mRNA interaction predictions, of which 3.2 million show high interaction propensity score (*cat*RAPID score ≥ 50) (see Methods) between the tested RBPs and ~87% of the initial coding transcripts (12,215 mRNAs). RBPs are predicted to interact with 3176 mRNAs on average (26% of the tested mRNAs) (Figure S1A), *i.e.*, twice as much as the average number of transcripts found to bind 112 RBPs using the eCLIP technology[18] (see Methods) on the same set of coding transcripts (Figure S1C). Similarly, *cat*RAPID predicts that mRNAs interact with a higher average number of RBPs (256 RBPs/mRNA, ~30% of the whole set) (Figure S1B) compared to eCLIP detected interactions (8 RBPs/mRNA, 7.5% of the whole set) (Figure S2D). Such differences are expected as *cat*RAPID predictions represent a set of biophysically possible interactions that are independent of the cellular localization of the interacting molecules and the experimental conditions in which *in vitro* and *in vivo* studies are carried out. Nevertheless, for 49 out of 74 eCLIP RBPs with *cat*RAPID predictions, we found an enrichment of

experimentally identified binding transcripts among predicted interactors at high interaction propensity score (two-sided Fisher's Exact test, BH-corrected P-value < 0.05) (Table S1), thus strengthening the confidence of our predictions.

Overall, to the best of our knowledge, we have predicted the largest human mRNA–RBP interaction network to date.

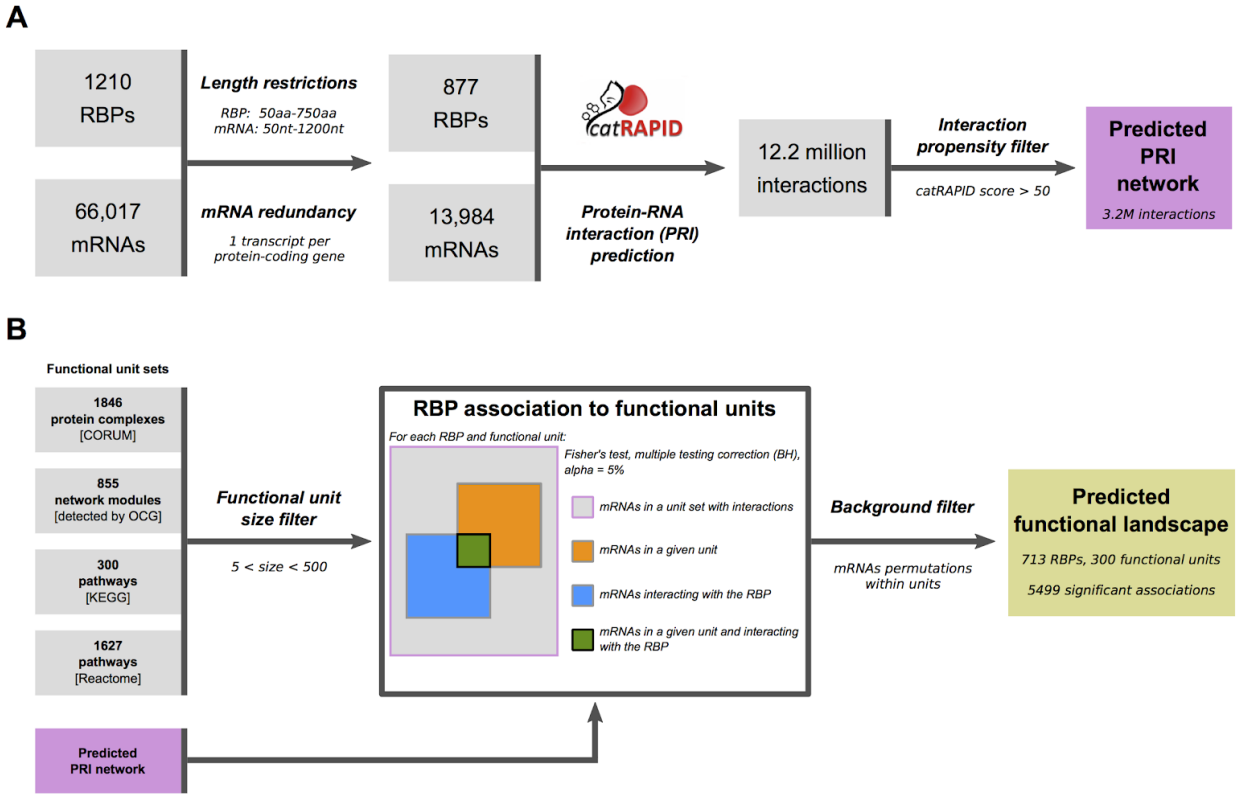


Figure 1. Workflows of our computational pipeline. (A) Prediction of protein-mRNA interactions (PRI) using the *cat*RAPID *omics* algorithm between experimentally identified human RBPs and a representative set of the human coding transcriptome. The resulting PRI network contains 3.2 million interactions. (B) Different functional units are tested for enrichment and depletions among RBP predicted targets in the PRI. This approach generated 5499 functional associations.

An inferred post-transcriptional regulatory landscape

We aimed at identifying the cellular functions that can be potentially regulated at the post-transcriptional level by RBPs. According to the regulon theory[3], a RBP can regulate a given biological process by binding a substantial fraction of mRNAs encoding its constituent proteins. We then expect to detect a statistically significant over-representation of mRNAs bound by the RBP among the functionally-related coding transcripts. We therefore investigated the predicted mRNA-RBP interaction network to characterize the

functional landscape of 877 RBPs (Figure 1B). We first gathered the transcripts encoding proteins involved in the same biological process or pathway, taken from four datasets representing different levels of organization of the cellular functions, and collectively named hereafter “functional units”: *(i)* 1846 manually curated protein macromolecular complexes from the CORUM database[31]; *(ii)* 873 functional modules detected in a human protein-protein interaction network using the OCG algorithm, which decomposes a network into overlapping modules based on modularity optimization[32]; *(iii)* 300 pathways described in the KEGG database[33]; and *(iv)* 1627 pathways from the Reactome knowledgebase[34] (see Methods).

Next, for each functional unit, we have computed the ratio of interacting *vs.* non-interacting transcripts with every RBP and assessed its significance to be higher or lower than expected by chance by performing a two-sided Fisher’s Exact test (see Methods). We chose this strategy to gain a broader view on the relationships between RBPs and their functional targets. Indeed, a statistically significant over-representation of predicted interactors hints a putative post-transcriptional regulatory event by a given RBP, whereas a statistically significant under-representation suggests that certain functional units may avoid the binding of a RBP and its possible regulatory action.

Seven hundred thirteen RBPs out of 877 (81% of the tested RBPs) showed at least one statistically significant result (5499 in total, BH-corrected P-value <0.05), namely 3185 significant enrichments (58%) and 2314 significant depletions (42%) involving 300 functional units out of the 2977 tested (see Methods). Because some RBPs are predicted to bind a large number of transcripts, we estimated the number of functional units expected to be found over- or under-represented by chance for each RBPs as a control, by randomly shuffling the protein names within the functional units 1000 times (see Methods). All the 713 RBPs passed this test, as their targets were enriched or depleted in a significantly higher number of functional units compared to random. Thus, they were kept for further study (Table S2).

Overall, this two-step statistical analysis allowed us to define the potential post-transcriptional regulatory landscape of numerous cellular processes by identifying *(i)* those functional units that can be regulated at the post-transcriptional level and *(ii)* the RBPs responsible for such regulation.

Statistical enrichments and depletions of RBP binding as an indication of post-transcriptional regulation

Our analysis reveals an interesting pattern of functional enrichments/depletions. Indeed, it allows grouping RBPs and functional units in three broad categories each (Figure 2A, Table S3 and S4).

On the one hand, a relatively small number of RBPs show only enrichments in predicted targets among functional units (75 RBPs, ~10% of the RBPs with significant results, named hereafter RBP-1 set), indicating that these RBPs display an exclusive binding preference for a number of FUs. A second category accounting for 427 RBPs shows both significant enrichments and depletions of their predicted targets among functional units (~60%, RBP-2 set) suggesting that they bind the mRNAs of certain functional units and avoid those of others. Finally, the third category contains 211 RBPs that display only significant depletions (~30%, RBP-3 set) within functional units, illustrating that some functional units avoid RBP binding (Figure 2B).

On the other hand, from the perspective of the FUs, we observe a mirrored situation. Most functional units (223 functional units, 74% of the units with significant results, named hereafter FU-1 set) are exclusively enriched in targets of at least one RBP, thus possibly regulated at the post-transcriptional level through the binding of those RBPs. Few functional units, namely 27 (9%, FU-2 set), are both enriched and depleted in RBP predicted targets, indicating that they may be regulated by the binding of certain RBPs and the avoidance of others. Finally, 50 functional units (~17%, FU-3 set) show only significant depletions thereby indicating that their post-transcriptional regulation consists uniquely in the avoidance of RBPs binding (Figure 2B).

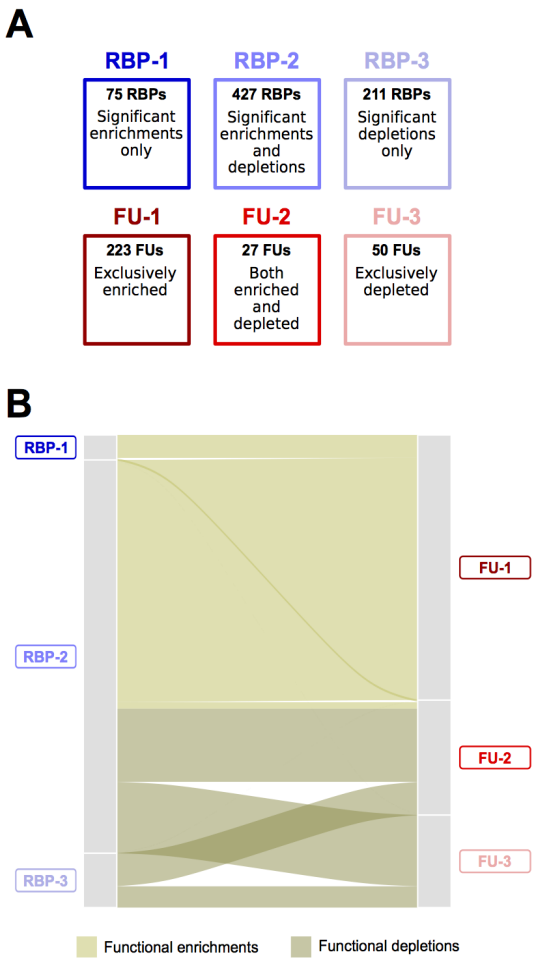


Figure 2. The predicted functional regulatory landscape. (A) Summary of the composition of the three RBP (shades of blue color) and functional unit (FU, shade of red color) groups. (B) Alluvial plot depicting the functional relationships among RBP and FU groups in the predicted functional regulatory landscape. The thickness of each stream is proportional to the number of enrichment or depletions between two given groups. The size of the grey blocks is proportional to the number of enrichments/depletions in which a given RBP or FU group is involved.

To assess whether the observed enrichment/depletion patterns of the predicted landscape do not depend on the *cat*RAPID interaction propensity threshold chosen, we carried out a threshold-free statistical analysis based on the GSEA method[35] (see Methods). Importantly, we also found the three distinct categories for both RBPs and functional units, with the RBP-2 set being involved in a similar fraction of the significant functional enrichments and depletions (Figure S2A), therefore supporting the observed pattern in the threshold-based predicted functional landscape. However, the fraction of RBPs in the RBP-3 set is lower (9%) (Supplementary note, Table S5) when using the threshold-free method compared to the fraction detected by the threshold-based approach (30%), indicating the influence of the chosen *cat*RAPID interaction propensity threshold on this category.

To further corroborate our results on predicted RBP-mRNA interactions, we carried out the functional analysis based on Fisher's Exact test on experimentally identified mRNA interactors of 112 RBPs using eCLIP (Supplementary note, Table S6). As before, we detected the three groups of functional units (*i.e.*, exclusively enriched, both enriched and depleted and depleted only) as well as the RBP-1 and RBP-2 sets, but none of the analyzed RBP had exclusively depleted functional units among its interactors (*i.e.*, RBP-3 set). This discrepancy between our predicted regulatory landscape and the results obtained on eCLIP data corroborates the influence of the chosen *cat*RAPID interaction propensity threshold (i.e., score >=50) already observed with the GSEA method on this latter category. Altogether, these assessments show that by using our strict parameters, we may have limited the occurrence of potential false positives cases by favoring sensitivity rather than specificity.

Overall, the results obtained on both predicted and experimental RBP-mRNA interactions suggest that RBPs can adopt several possible regulation strategies and can be classified accordingly.

The predicted regulatory landscape from the RBP perspective

The classification of RBPs in distinct groups based on the functional analysis of their interactors motivate us to assess whether the RBPs have distinct functional and sequence features as well as system-level properties (Table S3).

First, we observed that RBPs in the RBP-2 set have a statistically significant higher number of enrichments (average=6.7, median=4, P-value=7.6×10^{-6}, Mann-Whitney *U* test, one-sided) and depletions (average=3.9, median=4, P-value=7.4×10^{-13}, Mann-Whitney *U* test, one-sided) compared to those of the RBP-1 (average=3.8, median=2) and the RBP-3 (average=3, median=3) sets respectively. This suggests that the more numerous RBP group in our classification (RBP-2 set) can potentially regulate the larger number of FUs (Figure 2).

Second, we checked whether RBPs of the three groups were characterized by an over-representation of different types of RBPs according to a previously proposed functional classification[22] (see Methods). Indeed, Beckmann and colleagues annotated RBPs into four classes: (*i*) established RBPs (*i.e.*, proteins with a known role in RNA biology); (*ii*) RBPs carrying a characterized RNA-binding domain (RBD); (*iii*) enigmRBPs, which are proteins found to bind RNA but lacking a canonical RBD and with no previous evidence of involvement in RNA fate; (*iv*) RNA-binding enzymes, which have a RNA-independent metabolic activity.

We found that RBPs with a defined role in RNA biology are depleted in the RBP-1 set (odds ratio = 0.53, P-value = 0.009, Fisher's Exact test, one-sided), which is otherwise enriched in enigmRBPs (odds ratio =

2, P-value = 0.004, Fisher's Exact test, one-sided) (Figure 3A). In the RBP-2 set, we detected a significant over-representation of RBPs with recognized RNA-binding domains (RBDs) (odds ratio = 1.27, P-value = 0.04, Fisher's Exact test, one-sided) and a significant depletion of enigmRBP (odds ratio = 0.75, P-value = 0.04, Fisher's Exact test, one-sided). We did not observe any statistically significant over- or under-representation among the RBP-3 set. We also checked whether the RBPs in the three groups showed difference in the binding preference of other RNA biotypes based on previous knowledge {Gerstberger et al., 2014}. Interestingly, we observed that the RBPs binding predominantly mRNAs more frequently in the RBP-1 (82%) compared to the RBP-2 (66%) and RBP-3 (64%) sets, in which we observed an higher fraction of ribosomal proteins and RBPs binding small RNAs (Table S8). Recent reports showed that many RNA-binding sites are found in intrinsically disordered regions[24] and that RBPs are enriched in low complexity sequence stretches[19]. Hence, we compared the disorder propensity and low complexity content of the RBP sequences belonging to the three different groups using state-of-the-art tools (see Methods). The RBP-1 set has a slightly higher disorder (Figure S3A and Figure S3B) and low complexity content (Figure S3C) compared to the other two groups. However, these differences are not statistically significant, meaning that these features cannot entirely explain the different enrichment/depletion patterns.

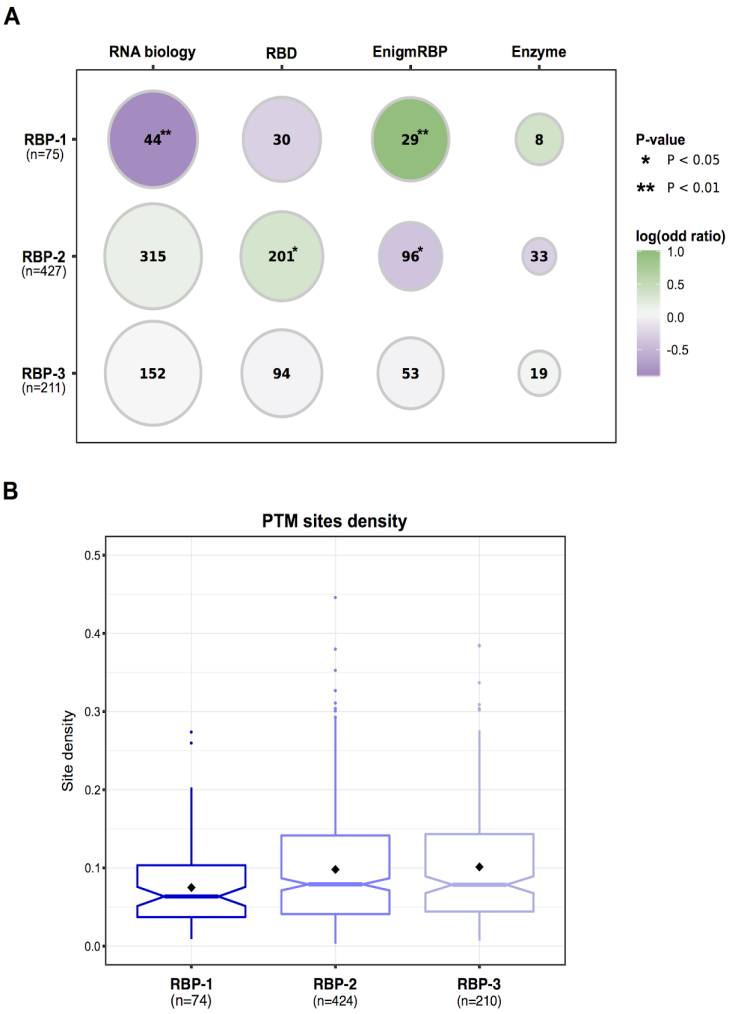


Figure 3. RBPs belonging to the three sets have distinct features. (A) Enrichments (circles filled in green) and depletions (circles filled in violet) of different types of RNA-binding proteins among the three groups of RBPs were assessed using the Fisher's Exact test. Size of the circles is proportional to the fraction of RBPs of a given type that are present in each of the RBP groups, and their frequency is reported as a number within the circle. Significant enrichment and depletions are denoted by one (P-value <0.05) or two (P-value<0.01) asterisks. (B) Distribution of the overall post-translational modification (PTM) density in the sequences of the three RBP groups. Densities for every RBP are computed as the number of experimentally identified PTM sites divided by the RBP sequence length. Black diamonds represent density mean values. Boxplot colors correspond to the RBP group colors in Figure 2.

RBPs are generally ubiquitously expressed given their central role in gene regulation[37]. In a compendium of 58 human tissues (see Methods), we did not observe any statistically significant difference among the three groups (Figure S3D), suggesting that the functional enrichment/depletion patterns are independent of the expression breath of the RBPs.

The function of regulatory proteins – such as protein kinases[38], transcription[39] and chromatin remodeling factors[40,41] – is fine-tuned through post-translational modifications (PTMs). Increasing evidence indicates that the activity of RBPs can also be regulated by PTMs[24,42]. We collected the modification site data for seven PTM types from the PhosphoSitePlus database[43] (see Methods) and mapped them onto the RBP sequences of the three groups. We found that RBPs of the RBP-1 set have a significantly lower PTM density (Figure 3B) compared to RBP-2 (Kruskal-Wallis test followed by post-hoc Dunn's test, corrected P-value = 0.016) and RBP-3 (Kruskal-Wallis test followed by post-hoc Dunn's test, corrected P-value = 0.029) (Table S9). When considering individual PTM types alone, a lower density is still observed for the RBP-1 set (Figure S4), which is statistically significant for acetylation and phosphorylation (Table S9). These results indicate that the function of RBPs belonging to the RBP-2 and RBP-3 sets can be more finely regulated at the post-translational level than the RBPs of the RBP-1 set.

In conclusion, our analyses identified several features discriminating the RBPs belonging to the different groups that could explain the regulatory behaviour they may have on functional units.

The predicted regulatory landscape from the functional unit perspective

What are the cellular processes embodied by the 300 functional units present in the predicted regulatory landscape (Table S4). The FU-1 units are exclusively enriched among the predicted targets of 480 RBPs (average number of RBPs per unit: 13.8), whereas FU-3 units show significant depletions only among the interactors of 499 RBPs (average number of RBPs per unit: 21.5). The few functional units in the FU-2 groups are enriched among the targets of 74 RBPs (average number of RBPs per unit: 3.7) and depleted among the interactors of 600 RBPs (average number of RBPs per unit: 45.8).

FU-1 units are involved in processes related to gene expression, such as chromatin organization and regulation, transcription initiation, protein degradation, which are known to be coupled[1,44]. Among the FUs related to chromatin organization and transcription activation, we found SWI/SNF-containing complexes and distinct forms of the Mediator complex from CORUM, as well as several network modules and Reactome pathways involved in DNA methylation and RNA Polymerase I transcription initiation. Notably, both SWI/SNF and Mediator complexes have been implicated in RNA processing[45,46] and their subunit transcripts are regulated post-transcriptionally by miRNAs[47,48]. Moreover, many of these FUs contain histones, whose expression can be controlled at the post-transcriptional level[49]. Altogether, our results underline the role of protein-RNA interactions in coordinating the different steps of gene expression programs, as it has been shown for the regulation of chromatin structure and DNA transcription[50,51].

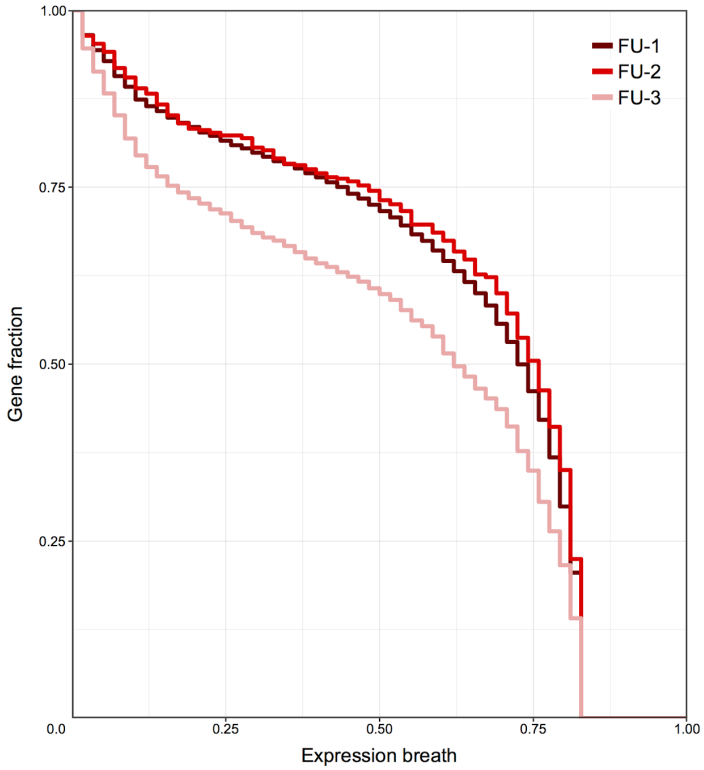


Figure 4. Tissue expression distributions of the proteins annotated in the three FU groups. The color of each distribution correspond to the FU colors in Figure 2.

Additional enriched FUs are related to cellular processes localized in the mitochondria. Indeed, we identified as frequently-enriched the large subunit of the mitochondrial ribosome from CORUM, four Reactome pathways related to mitochondrial translation as well as complexes (*e.g.*, the respiratory chain complex I) and pathways (*e.g.*, TCA cycle, oxidative phosphorylation) involved in energy production. Interestingly, these results corroborate the known post-transcriptional regulation of the mitochondrial components[52–54].

FU-2 units are involved in several signaling pathways. Indeed, we found that two pathways related to olfactory signaling (one from KEGG and the other from Reactome) are depleted in interactors of around two-thirds of the tested RBPs. However, they are exclusively enriched in those coded by the ERAL1, G3BP1, G3BP2, MKRN2 and TUFM genes, all expressed in brain tissues, according to Human Protein Atlas[55], and their coding transcripts have been detected in olfactory sensory neurons (G3BP1, G3BP2, MKRN2, TUFM) or epithelium (ERAL1, TUFM)[56]. Our results indicate that these RBPs could potentially regulate the fate of olfactory signaling mRNAs.

Finally, the most frequently depleted units in FU-3 are related to glutamate receptor signalling, defensins and glycosylation of mucins, as well as some units related to cytoskeleton organization. Interestingly, proteins in FU-3 are expressed in a lower number of tissues compared to those in FU-1 (Kolgomorov-Smirnov test, P-value<$2.2x10^{-16}$) and FU-2 (Kolgomorov-Smirnov test, P-value=$1.7x10^{-10}$), respectively. This suggests that RBP-binding avoidance may participate to the proper tissue-specific expression of some functional unit components.

Disease pathways are targeted by common RBPs

Among the 223 exclusively enriched functional units (FU-1) we found 20 disease-related pathways from the KEGG database. The majority of them (*i.e.*, 13) are related to viral and bacterial infections, whereas the other disease functional units are linked to immune-related, neurological and metabolic disorders (Figure 5). Notably, 17 disease FUs can be regulated by common RBPs, which can also target other non-disease related FUs. For instance, 4 viral infection FUs and one immunological disorder unit are all enriched among the predicted targets of the BTB/POZ domain-containing protein KCTD12, an enigmRBP {Beckmann et al., 2015}. KCTD12 predicted interactors are also enriched among coding transcripts annotated in three FUs related to immune system pathways (Figure 5), suggesting that this novel and uncharacterized RBP may be involved in immunity and in infection-related processes.

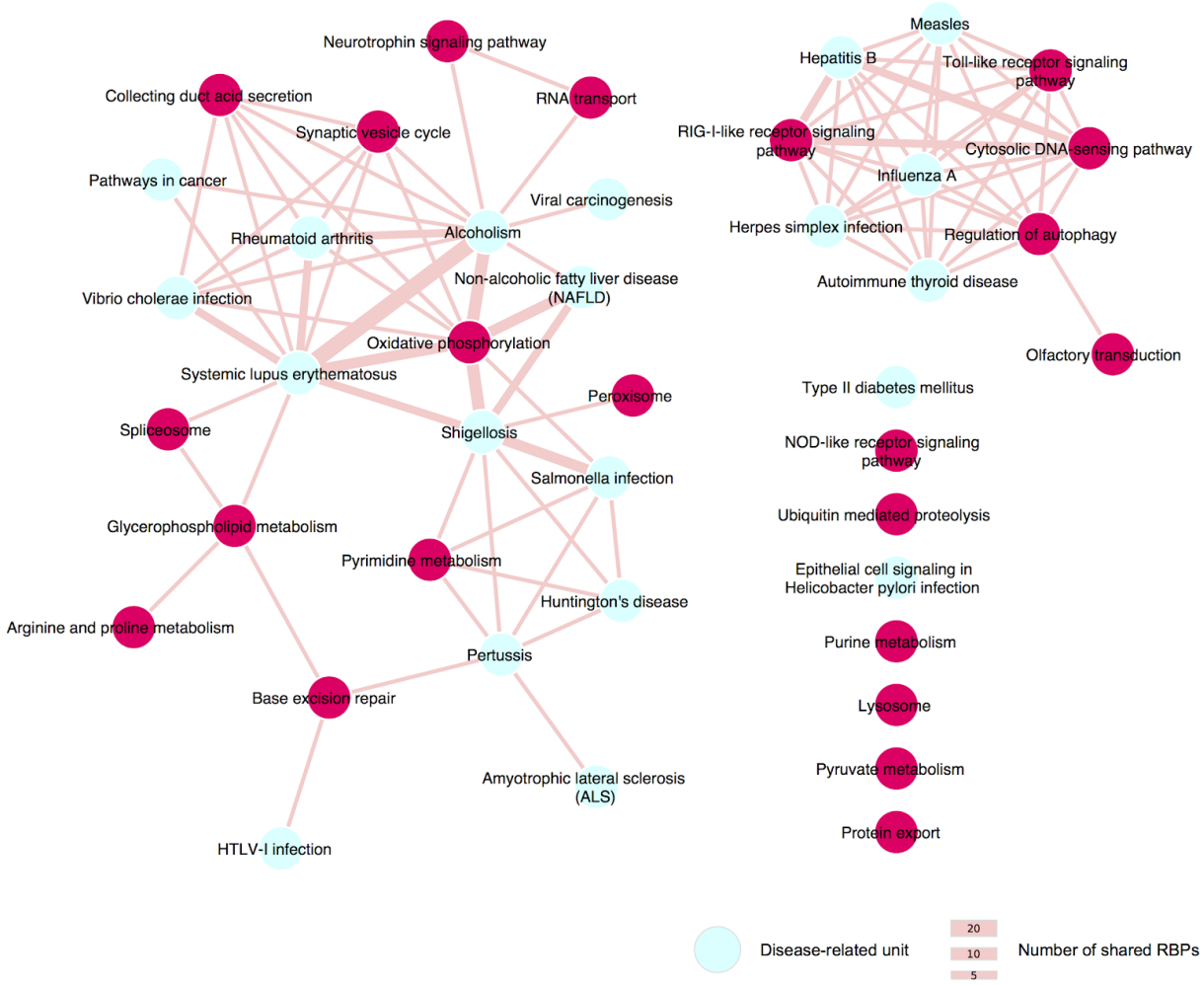


Figure 5. Network representation of disease-related units sharing common RBPs. The size of the edges is proportional to the number of shared RBPs by the two units. Disease units, depicted in cyan, share also RBPs with non-disease related units, depicted in magenta. For sake of clarity, we included only non-disease FUs from the KEGG database. The network was generated using the Cytoscape app (www.cytoscape.org).

We also observed this commonality among FUs-related to bacterial infection as well (*i.e.*, Shigellosis, Pertussis and Salmonella infection pathways). These units are enriched in interactors of the PRKC apoptosis WT1 regulator protein encoded by the PAWR gene (also known as PAR4), which it has been implicated in mRNA splicing in cancer cells {Lu et al., 2013}. Furthermore, the pyrimidine metabolism pathway is also enriched among PAWR predicted interactions. Interestingly, it has been shown that intracellular pathogenic bacteria -- such as *Salmonella*, *Shigella* and *Bordetella* (the etiological agent of pertussis) -- can modulate several host cell metabolism pathways for their own benefit, including

nucleotide biosynthesis {Eisenreich et al., 2013}, indicating a potential role of PAWR in the post-transcriptional regulation of genes involved in bacterial infection response.

Overall, these results show that our predicted functional landscape is an useful resource to formulate new hypotheses on the cellular role of both established and novel RBPs.

Discussion

In this work, we explored the post-transcriptional regulation of functionally-related mRNAs by RBPs to first, detect different behaviours, if present, among the hundreds of RBPs analysed, and second, estimate the prevalence of the regulon theory at the coding transcriptome scale. As experimentally determined mRNA–protein interactions are too scarce to allow a large-scale investigation of the post-transcriptional regulation, we computationally predicted an interaction network between representative sets of RBPs and mRNAs in order to better cover the interaction space. For this, we used *cat*RAPID *omics*, a large-scale protein–RNA interaction predictor that exploits the physicochemical features of the interacting molecules {Bellucci et al., 2011; Agostini et al. 2013}, which has been initially validated on a large collection of experimentally identified protein-RNA associations[26,30]. Noticeably, our computational analyses on the *in silico* predicted network are also performed on available experimentally identified mRNA-protein interactions in order to support and validate all observations.

By studying both types of data, we detected statistically significant over- and under-representations of the mRNAs bound by the RBPs among the functionally related coding transcripts. First of all, these results allow an estimation of the prevalence of the regulon theory. Among the 2977 functional units that we tested, comprising protein complexes, network modules and pathways, only ~10% (300) have been found possibly regulated in the predicted functional landscape. These results are affected by two factors: *(i)* some FUs may be partially overlapping (*e.g.*, some protein complexes may play a role in some pathways) or redundant, therefore leading to an overestimation; *(ii)* the choice of a strict *cat*RAPID threshold may have led to an under-estimation of the number of potentially regulated FUs. Moreover, as by construction, our statistical approach detects regulation events by considering a pairwise combination of FU and RBP, ignoring possible combinatorial and/or dynamic regulation modulations that could involve several RBPs {Dassi et al., 2017}, the regulon prevalence could have been underestimated. Indeed, the analysis carried out on the eCLIP data provides an higher proportion of enriched/depleted FUs (40%, see Supplementary note), thus suggesting that the underestimation is the most plausible scenario.

Second, the different patterns of enrichments and depletions for the RBP binding to functional unit transcripts revealed by our analysis, lead to a post-transcriptional landscape shaped the RBP-mRNA interactions. It reveals that 57,2% of the 877 tested RBPs regulate FUs by possibly binding to their mRNAs whereas 72% do so by being avoided, therefore indicating the prevalence of this latter RBP

regulatory mode. On the other hand, the groups of functionally related mRNAs (the 2977 FUs) appear to rather be regulated through the binding than through the avoidance of the RBPs (8% enriched, 2,6% depleted). Notwithstanding this, 90% of the FUs do not appear as being particularly regulated by RBPs.

Indeed, promiscuous RBPs interacting with a vast majority of mRNA targets in the coding transcriptome and FUs interacting with those are not expected to be detected as significant by our approach since the spread of the RBP targets precludes the detection of a statistically significant signal. This could be the case for 18% of the RBPs (164 RBPs) and 90% of the FUs (2677 FUs) for which no statistical signal has been detected.

We observed 3 different patterns of enrichments and depletions for the RBP binding of functional unit transcripts. These patterns may reflect different possible FU molecular regulation strategies by the RBPs, involving *(i)* the presence of RBP binding in the case of RBP targets enrichment, *(ii)* its avoidance in the case of depletions, or *(iii)* presence or avoidance of binding, when both enrichments and depletions are observed for a given RBP. Indeed, whereas some RBPs (the RBP-1 set) appear to act exclusively through their binding to the mRNAs of the FUs (*i.e.*, presence of binding), some others (the RBP-3 set) are excluded from binding by having less targets than expected by chance among the mRNAs of the FUs (*i.e.*, avoidance of binding). Finally, for other RBPs (the RBP-2 set), both strategies, presence and avoidance of binding are observed.

What do represent the ‘presence’ and the ‘avoidance’ of RBP binding? As *cat*RAPID identifies RNA-protein interactions, the ‘presence’ is the physical ability for a RBP to regulate the FUs through its binding, independently of the binding status itself, bound or unbound, which may change with conditions. Conversely, the ‘avoidance’ is the physical inability for the RBP to bind, *e.g.*, because of the lack of binding sites. As well as the ability, the inability to bind can lead to a regulation event.

Interestingly, the observed depletion or avoidance of binding could reflect a molecular mechanism limiting inappropriate binding that could interfere with correct gene expression. Indeed, it has been proposed recently by Savisaar and Hurst[36] that coding sequences are evolutionarily constrained to avoid certain RBP binding motifs, in order to prevent inappropriate interactions with RBPs that could impair, for instance, their correct mRNA processing. Such avoidance of regulatory elements has also been observed for target sites of microRNAs within 3’UTRs {Stark et al., 2005} and to limit spurious transcription binding sites. Our striking observation that some functional units could contain the information to not interact with certain RBPs could therefore represent a cellular regulatory mechanism *per se*. However, further work is needed to investigate this hypothesis.

We further studied the properties of the RBPs belonging to the three sets and found that are several features that can distinguish them. For instance, the RBP-1 set is characterized by an enrichment in

enigmRBPs that lack canonical RBDs and for which a role in RNA biology has not been established so far. Among the 29 enigmRBPs in the RBP-1 set, there are 8 metabolic enzymes, including the moonlighting protein Leukotriene A-4 hydrolase (LTA4H) {MoonProt, Chen et al., 2018}, and several signaling and structural proteins. In addition, RBP in this group have a significant low density in PTM sites, which can regulate, for instance, RNA binding or dictate the subcellular localization of a given RBP[42]. Altogether, this suggests that these RBPs are putative multifunctional proteins whose RNA binding activity, which represents one of their possible molecular tasks, can be potentially modulated by a not yet identified molecular signal.

Conversely, the RBP-2 set is enriched in RBPs with canonical RBDs showing a significantly higher PTM density compared to the RBP in set 1, consistent with the current knowledge that the function of established RBPs is modulated by post-translational modifications, as in case of SR splicing factors {Colwill et al., 1996}, ELAVL1 {Abdelmohsen et al., 2007; Yu et al., 2011} and FMR1 proteins {Dolzhanskay et al., 2006}. Moreover, RBPs in set 2, as well those in set 3, show a wider range of binding preferences among RNA biotypes compared to the RBP-1 set, which comprises an high fraction of RBPs binding preferentially/exclusively mRNAs. Overall, our analysis indicates that RBPs in set 1 have distinct features that discriminate them from the two other groups. Consequently, further experimental studies are needed to identify the *in vivo* RNA interactors of RBPs in set 1 (only 4 have been tested with the eCLIP technology) and, in the case of the enigmRBPs, decipher their role in mRNA fate.

Altogether, our analyses defined a post-transcriptional regulatory landscape occupied by functionally related mRNA differently regulated by RBPs, thereby allowing us to provide a novel classification of the RBPs. This classification may help understanding the regulatory of action of the continuously increasing number of newly discovered RBPs.

Methods

RNA-binding proteins and coding transcripts. We collected a list of 1217 human RBP protein-coding genes identified by mRNA interactome capture by Beckmann et al.[22] and their corresponding amino acid sequences from the UniprotKB human reference proteome[60] (May 2016). We downloaded the human coding transcriptome cDNA sequences (66,017 mRNAs) from Ensembl v82[61] (September 2015).

RNA-binding protein annotations. For each RBP in our dataset, we gathered from the original article the following annotations: whether a role in RNA biology is known (based on Gene Ontology annotations), presence or absence of a recognized RNA-binding domain according to the classification proposed in Castello et al.[19], whether it has been categorized as 'classic' metabolic enzyme (i.e.,

non-RNA-related enzymes). Those RBPs lacking a recognized RNA-binding and with no established role in RNA biology are labelled as enigmRBP[22].

Protein-RNA interaction predictions. We used the *cat*RAPID *omics* algorithm[25], which allows large-scale predictions between transcript and protein sequences, to compute the interaction propensities between human RBPs and coding transcripts. Due to *cat*RAPID computational constraints, we selected mRNA sequences between 50 and 1200 nucleotides of length, as well as protein sequences between 50 and 750 amino acids. Around 72% of the RBPs (877 proteins) and 57% of the human coding transcriptome (37,788 mRNAs) respected the length criterion. To avoid functional biases in subsequent analyses, we further reduced sequence redundancy among mRNAs (i.e., transcript isoforms) by selecting, for each protein-coding gene, the longest transcript as the representative sequence. Doing so, we retained 13,984 transcripts coded by ~63% of the annotated protein-coding genes in Ensembl v82 (22,029 genes). We then predicted more than 12 million protein-RNA interactions between 877 RBPs and 13,984 mRNAs.

Dataset of experimentally identified protein-RNA interactions. We retrieved interaction information from the ENCODE enhanced CLIP (eCLIP) dataset[18] gathering 159 experiments for 112 RBPs. We mapped BED peak coordinates referencing the GRCh38 human assembly to Ensembl v82 coding transcripts models using BEDTools intersect v2.17[62] with flags *–w* and *-a*. Interactions from replicates and different cell lines were pooled. To have an interaction set comparable to *cat*RAPID predictions, interactions involving transcript isoforms were mapped to the corresponding coding gene and counted as one. Doing so, we obtained a final list of 131,366 experimental interactions between 112 RBPs and at least one transcript encoded by 11,647 genes.

Compendium of functional units. We built a wide compendium of 4646 functional units and processes by gathering annotations from different sources: 1846 manually annotated human protein complexes from the CORUM database[31]; 873 functional network modules, defined as groups of proteins densely connected through their interactions and involved in the same biological process, detected by the OCG algorithm (Becker et al. 2012) on a human protein binary interactome built and annotated as previously described[63,64, zanzoni et al., 2017]; 300 maps and 1627 biological pathways from KEGG and Reactome databases, respectively[33,34] (Kanehisa et al. 2012; Croft et al. 2014). The gene lists annotated in CORUM complexes and biological pathways from KEGG/Reactome were downloaded from the gProfiler webserver[65] (rev1477, October 2015, based on Ensembl v82), which provides Ensembl identifiers for annotated genes. The genes/proteins annotated in the OCG network modules were mapped to the corresponding Ensembl v82 gene identifiers through the Ensembl BioMart service. We restricted

subsequent analyses to complexes, modules and pathways having at least 5 and no more than 500 genes/proteins (*i.e.*, 2977 functional units).

Functional unit enrichment analysis. To assess the over- and under-representation of the functional units among RBP predicted interactions, as done previously[30,66], we considered as interacting all RBP-mRNA pairs with a *cat*RAPID interaction propensity score of at least 50 and non-interacting all those with a score below 50. We next computed, for each functional unit in a given annotation dataset, the log2-transformed ratio of annotated mRNAs among RBP interacting and non-interacting transcripts and assessed its significance by performing a two-sided Fisher's Exact test. P-values were corrected for multiple testing using the Benjamini-Hochberg procedure and we considered as significant only those enrichments/depletions with a corrected P-value below 0.05. As RBPs are predicted to bind to many mRNAs, we further evaluated the number of enrichments/depletions expected by chance in each dataset by shuffling 1000 proteins labels among functional units. Only RBPs having a significantly higher number of enrichments/depletions (empirical P-value<0.05) were kept.

In a second approach, we carried out a Gene Set Enrichment Analysis[35] (GSEA) using annotated mRNAs in a given functional unit as gene set. We selected as significant only those enrichments (normalized enrichment score > 0) or depletions (normalized enrichment score < 0) with a false discovery rate (FDR) < 0.05 based on 1000 gene set permutations. In both tests, we used annotated mRNAs in the *cat*RAPID interaction space as statistical background.

The functional enrichment analysis of eCLIP interaction data was carried out using a Fisher's Exact test followed by a multiple testing correction (Benjamini-Hochberg procedure).

Intrinsic disorder and sequence complexity. We computed protein residue disorder propensity using the stand-alone version of two state-of-the art disorder prediction algorithms: IUPred[67] (both long and short predictions) and DISOPRED3[68]. An amino acid was considered disordered if its probability score was greater than 0.4. We calculated the RBP sequence low complexity using the NCBI segmasker application, which is based on the SEG algorithm[69], using default parameters. For each RBP, we computed the fraction of the number of predicted disordered and low complexity amino-acid residues divided by the sequence length.

Post-translational modification sites. We collected post-translational modification (PTM) information for 18,030 proteins from PhosphositePlus[43], which stores data for seven different PTMs: acetylation (20'854 sites in 6874 proteins), methylation (15'195 sites in 5347 proteins), O-GalnAc (2115 sites in 476 proteins), O-GlcnAc (420 sites in 166 proteins), phosphorylation (227'514 sites in 17'464 proteins), sumoylation

(7932 sites in 2500 proteins) and ubiquitination (62’256 sites in 10’325 proteins). We extracted PTM data for the RBPs and computed their PTM densities as the number of PTM sites over the sequence length.

Protein expression profiles. We downloaded protein expression data in human tissue based on immunohistochemistry from the Human Protein Atlas (version 18)[55]. We considered as expressed 10,579 protein-coding genes with a qualitative expression level of at least 'low' a reliability score equal to 'approved' or higher. For each protein-coding gene, we computed the expression breath as the fraction of tissues in which the given gene is considered as expressed over the total number of tissues present in the Human Protein Atlas (*i.e.*, 58).

Statistical analyses. Distributions of disorder propensity and low complexity content fractions, PTM densities and tissue expression breath ratios were compared by using a two-sided Kruskal-Wallis test (significance level=0.05), a non-parametric analysis of variance method. In case of a null-hypothesis rejection, we applied a *post hoc* Dunn Test, which performs multiple pairwise comparisons between the individual distributions (BH-corrected P-value significance level=0.05).

Data availability. All data generated or analyzed during this study are included in this published article and its supplementary information files. The predicted protein-RNA interactions are available from the corresponding authors on reasonable request.

Acknowledgements

The authors would like to thank Davide Cirillo (CRG, Barcelona), Guillaume Charbonnier (TAGC), Denis Puthier (TAGC) and Thien-Phong Vu Manh (CIML, Marseille) for fruitful discussion and advice; and Elisa Micarelli (TAGC) for providing the list of the RNA biotype targets for the RBPs. The RAINET project leading to this publication has received funding from Excellence Initiative of Aix-Marseille University - A*MIDEX, a French “Investissements d’Avenir” programme (to CB). GGT acknowledges the support of the European Research Council (RIBOMYLOME_309545) and the Spanish Ministry of Economy and Competitiveness (BFU2014-55054-P).

Author's contributions

AZ, GGT and CB conceived the study; AZ, LS and CB designed the experiments; AZ and DMR performed the experiments; AZ, LS and CB analyzed the data; AZ and CB wrote the manuscript with inputs from all the other authors.

Competing interests

The author(s) declare no competing interests.

References

1. Komili, S. & Silver, P. A. Coupling and coordination in gene expression processes: a systems biology view. *Nat. Rev. Genet.* **9,** 38–48 (2008).
2. Schwanhäusser, B. *et al.* Global quantification of mammalian gene expression control. *Nature* **473,** 337–342 (2011).
3. Keene, J. D. RNA regulons: coordination of post-transcriptional events. *Nat. Rev. Genet.* **8,** 533–543 (2007).
4. Imig, J., Kanitz, A. & Gerber, A. P. RNA regulons and the RNA-protein interaction network. *Biomol Concepts* **3,** 403–414 (2012).
5. Gerber, A. P., Herschlag, D. & Brown, P. O. Extensive association of functionally and cytotopically related mRNAs with Puf family RNA-binding proteins in yeast. *PLoS Biol.* **2,** E79 (2004).
6. Hogan, D. J., Riordan, D. P., Gerber, A. P., Herschlag, D. & Brown, P. O. Diverse RNA-binding proteins interact with functionally related sets of RNAs, suggesting an extensive regulatory system. *PLoS Biol.* **6,** e255 (2008).
7. Townley-Tilson, W. H. D., Pendergrass, S. A., Marzluff, W. F. & Whitfield, M. L. Genome-wide analysis of mRNAs bound to the histone stem-loop binding protein. *RNA* **12,** 1853–1867 (2006).
8. Anderson, P. Post-transcriptional regulons coordinate the initiation and resolution of inflammation. *Nat. Rev. Immunol.* **10,** 24–35 (2010).
9. Blackinton, J. G. & Keene, J. D. Post-transcriptional RNA regulons affecting cell cycle and proliferation. *Semin. Cell Dev. Biol.* **34,** 44–54 (2014).
10. Vohhodina, J. *et al.* The RNA processing factors THRAP3 and BCLAF1 promote the DNA damage response through selective mRNA splicing and nuclear export. *Nucleic Acids Res.* **45,** 12816–12833 (2017).
11. Scherrer, T., Femmer, C., Schiess, R., Aebersold, R. & Gerber, A. P. Defining potentially conserved RNA regulons of homologous zinc-finger RNA-binding proteins. *Genome Biol.* **12,** R3 (2011).
12. Fernández, E., Rajan, N. & Bagni, C. The FMRP regulon: from targets to disease convergence. *Front Neurosci* **7,** 191 (2013).
13. Galloway, A. & Turner, M. Cell cycle RNA regulons coordinating early lymphocyte development. *Wiley Interdiscip Rev RNA* **8,** (2017).
14. Bisogno, L. S. & Keene, J. D. RNA regulons in cancer and inflammation. *Curr. Opin. Genet. Dev.* **48,** 97–103 (2018).
15. Iadevaia, V. & Gerber, A. P. Combinatorial Control of mRNA Fates by RNA-Binding Proteins and Non-Coding RNAs. *Biomolecules* **5,** 2207–2222 (2015).
16. Dassi, E. Handshakes and Fights: The Regulatory Interplay of RNA-Binding Proteins. *Front Mol Biosci* **4,** 67 (2017).
17. McHugh, C. A., Russell, P. & Guttman, M. Methods for comprehensive experimental identification of RNA-protein interactions. *Genome Biol.* **15,** 203 (2014).
18. Van Nostrand, E. L. *et al.* Robust transcriptome-wide discovery of RNA-binding protein binding sites with enhanced CLIP (eCLIP). *Nat. Methods* **13,** 508–514 (2016).
19. Castello, A. *et al.* Insights into RNA biology from an atlas of mammalian mRNA-binding proteins. *Cell* **149,** 1393–1406 (2012).
20. Baltz, A. G. *et al.* The mRNA-bound proteome and its global occupancy profile on protein-coding transcripts. *Mol. Cell* **46,** 674–690 (2012).
21. Matia-González, A. M., Laing, E. E. & Gerber, A. P. Conserved mRNA-binding proteomes in eukaryotic organisms. *Nat. Struct. Mol. Biol.* **22,** 1027–1033 (2015).

22. Beckmann, B. M. *et al.* The RNA-binding proteomes from yeast to man harbour conserved enigmRBPs. *Nat Commun* **6,** 10127 (2015).
23. Conrad, T. *et al.* Serial interactome capture of the human cell nucleus. *Nat Commun* **7,** 11212 (2016).
24. Castello, A. *et al.* Comprehensive Identification of RNA-Binding Domains in Human Cells. *Mol. Cell* **63,** 696–710 (2016).
25. Agostini, F. *et al.* catRAPID omics: a web server for large-scale prediction of protein-RNA interactions. *Bioinformatics (Oxford, England)* **29,** 2928–2930 (2013).
26. Bellucci, M., Agostini, F., Masin, M. & Tartaglia, G. G. Predicting protein associations with long noncoding RNAs. *Nat. Methods* **8,** 444–445 (2011).
27. Agostini, F., Cirillo, D., Bolognesi, B. & Tartaglia, G. G. X-inactivation: quantitative predictions of protein interactions in the Xist network. *Nucleic Acids Res.* **41,** e31 (2013).
28. Cirillo, D. *et al.* Neurodegenerative diseases: Quantitative predictions of protein-RNA interactions. *RNA* **19,** 129–140 (2013).
29. Cirillo, D. *et al.* Constitutive patterns of gene expression regulated by RNA-binding proteins. *Genome Biol.* **15,** R13 (2014).
30. Ribeiro, D. M. *et al.* Protein complex scaffolding predicted as a prevalent function of long non-coding RNAs. *Nucleic Acids Res.* **46,** 917–928 (2018).
31. Ruepp, A. *et al.* CORUM: the comprehensive resource of mammalian protein complexes--2009. *Nucleic Acids Res.* **38,** D497-501 (2010).
32. Becker, E., Robisson, B., Chapple, C. E., Guénoche, A. & Brun, C. Multifunctional proteins revealed by overlapping clustering in protein interaction network. *Bioinformatics* **28,** 84–90 (2012).
33. Kanehisa, M., Goto, S., Sato, Y., Furumichi, M. & Tanabe, M. KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic acids research* **40,** D109–114 (2012).
34. Croft, D. *et al.* The Reactome pathway knowledgebase. *Nucleic Acids Res.* **42,** D472-477 (2014).
35. Subramanian, A. *et al.* Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *PNAS* **102,** 15545–15550 (2005).
36. Savisaar, R. & Hurst, L. D. Both Maintenance and Avoidance of RNA-Binding Protein Interactions Constrain Coding Sequence Evolution. *Mol. Biol. Evol.* **34,** 1110–1126 (2017).
37. Gerstberger, S., Hafner, M. & Tuschl, T. A census of human RNA-binding proteins. *Nat. Rev. Genet.* **15,** 829–845 (2014).
38. Nolen, B., Taylor, S. & Ghosh, G. Regulation of protein kinases; controlling activity through activation segment conformation. *Mol. Cell* **15,** 661–675 (2004).
39. Filtz, T. M., Vogel, W. K. & Leid, M. Regulation of transcription factor activity by interconnected post-translational modifications. *Trends Pharmacol. Sci.* **35,** 76–85 (2014).
40. Bannister, A. J. & Kouzarides, T. Regulation of chromatin by histone modifications. *Cell Res.* **21,** 381–395 (2011).
41. Wotton, D., Pemberton, L. F. & Merrill-Schools, J. SUMO and Chromatin Remodeling. *Adv. Exp. Med. Biol.* **963,** 35–50 (2017).
42. Lovci, M. T., Bengtson, M. H. & Massirer, K. B. Post-Translational Modifications and RNA-Binding Proteins. *Adv. Exp. Med. Biol.* **907,** 297–317 (2016).
43. Hornbeck, P. V. *et al.* PhosphoSitePlus, 2014: mutations, PTMs and recalibrations. *Nucleic Acids Res.* **43,** D512-520 (2015).
44. Braunschweig, U., Gueroussov, S., Plocik, A. M., Graveley, B. R. & Blencowe, B. J. Dynamic integration of splicing within gene regulatory pathways. *Cell* **152,** 1252–1269 (2013).
45. Tyagi, A., Ryme, J., Brodin, D., Ostlund Farrants, A. K. & Visa, N. SWI/SNF associates with nascent pre-mRNPs and regulates alternative pre-mRNA processing. *PLoS Genet.* **5,** e1000470 (2009).
46. Huang, Y. *et al.* Mediator complex regulates alternative mRNA processing via the MED23 subunit. *Mol. Cell* **45,** 459–469 (2012).

47. Grueter, C. E. *et al.* A cardiac microRNA governs systemic energy homeostasis by regulation of MED13. *Cell* **149,** 671–683 (2012).
48. Wade, S. L., Langer, L. F., Ward, J. M. & Archer, T. K. MiRNA-Mediated Regulation of the SWI/SNF Chromatin Remodeling Complex Controls Pluripotency and Endodermal Differentiation in Human ESCs. *Stem Cells* **33,** 2925–2935 (2015).
49. Rattray, A. M. J. & Müller, B. The control of histone gene expression. *Biochem. Soc. Trans.* **40,** 880–885 (2012).
50. G Hendrickson, D., Kelley, D. R., Tenen, D., Bernstein, B. & Rinn, J. L. Widespread RNA binding by chromatin-associated proteins. *Genome Biol.* **17,** 28 (2016).
51. He, C. *et al.* High-Resolution Mapping of RNA-Binding Regions in the Nuclear Proteome of Embryonic Stem Cells. *Mol. Cell* **64,** 416–430 (2016).
52. Antonicka, H. & Shoubridge, E. A. Mitochondrial RNA Granules Are Centers for Posttranscriptional RNA Processing and Ribosome Biogenesis. *Cell Rep* (2015). doi:10.1016/j.celrep.2015.01.030
53. Sirey, T. M. & Ponting, C. P. Insights into the post-transcriptional regulation of the mitochondrial electron transport chain. *Biochem. Soc. Trans.* **44,** 1491–1498 (2016).
54. Pearce, S. F. *et al.* Regulation of Mammalian Mitochondrial Gene Expression: Recent Advances. *Trends Biochem. Sci.* **42,** 625–639 (2017).
55. Uhlén, M. *et al.* Tissue-based map of the human proteome. *Science* **347,** 1260419 (2015).
56. Olender, T. *et al.* The human olfactory transcriptome. *BMC Genomics* **17,** 619 (2016).
57. Pancaldi, V. & Bähler, J. In silico characterization and prediction of global protein-mRNA interactions in yeast. *Nucleic Acids Res.* **39,** 5826–5836 (2011).
58. Brun, C. *et al.* Functional classification of proteins for the prediction of cellular function from a protein-protein interaction network. *Genome biology* **5,** R6 (2003).
59. Nabieva, E., Jim, K., Agarwal, A., Chazelle, B. & Singh, M. Whole-proteome prediction of protein function via graph-theoretic analysis of interaction maps. *Bioinformatics* **21 Suppl 1,** i302-310 (2005).
60. Breuza, L. *et al.* The UniProtKB guide to the human proteome. *Database (Oxford)* **2016,** (2016).
61. Cunningham, F. *et al.* Ensembl 2015. *Nucleic Acids Res.* **43,** D662-669 (2015).
62. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26,** 841–842 (2010).
63. Chapple, C. E. *et al.* Extreme multifunctional proteins identified from a human protein interaction network. *Nat Commun* **6,** 7412 (2015).
64. Zanzoni, A. & Brun, C. Integration of quantitative proteomics data and interaction networks: Identification of dysregulated cellular functions during cancer progression. *Methods* **93,** 103–109 (2016).
65. Reimand, J., Arak, T. & Vilo, J. g:Profiler--a web server for functional interpretation of gene lists (2011 update). *Nucleic Acids Res.* **39,** W307-315 (2011).
66. Zanzoni, A. *et al.* Principles of self-organization in biological pathways: a hypothesis on the autogenous association of alpha-synuclein. *Nucleic Acids Res.* **41,** 9987–9998 (2013).
67. Dosztányi, Z., Csizmok, V., Tompa, P. & Simon, I. IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content. *Bioinformatics* **21,** 3433–3434 (2005).
68. Jones, D. T. & Cozzetto, D. DISOPRED3: precise disordered region predictions with annotated protein-binding activity. *Bioinformatics* **31,** 857–863 (2015).
69. Wootton, J. C. & Federhen, S. Analysis of compositionally biased regions in sequence databases. *Meth. Enzymol.* **266,** 554–571 (1996).

3. General discussion & perspectives

3.1. Integration of protein-protein and protein-RNA interactions

The complexity of higher organisms requires the existence of a concerted series of interactions between several types of macromolecules, such as those between proteins, DNA and RNA. Having a complete map of all the interactions in a cell would allow us to comprehend deeply the regulatory mechanisms employed by the cell, and predict the outcome of a change in the system (Vidal, 2001; Costanzo *et al.*, 2016). However, this is an enormous challenge, akin to mapping all the connections between the human brain's neurons in order to understand brain function and behavior (Van Essen *et al.*, 2013).

3.1.1. Choice of the interaction datasets used

Current macromolecule interaction maps are relatively scarce for several types of molecules, as in the case of protein-RNA interactions. Thus, large-scale analysis of protein-RNA interaction networks have been hindered by the lack of experimental interactions. Indeed, these have only been applied to a subset of molecules (e.g. a few hundred RBPs for protein-centric approaches, a few dozen RNAs for RNA-centric approaches), and have often excluded some types of molecules such as non-canonical RBPs and RNAs lacking poly(A) tails (Van Nostrand *et al.*, 2016; Huang *et al.*, 2018). In addition, considering the high number of distinct transcripts in mammalian cells and that many RBPs have hundreds or even thousands of RNA targets (Van Nostrand *et al.*, 2016), the protein-RNA interaction space appears to be much larger than the protein-protein interaction space (Ribeiro *et al.*, 2018; See Figure 2). Therefore, mapping all protein-RNA interactions would require a vast effort. Making this task more difficult, current high-throughput experimental methods to detect protein-RNA interactions (e.g. PAR-CLIP and eCLIP) have been shown to suffer from several biases (Introduction, section 1.1.4) (Krakau, Richard and Marsico, 2017) and experimental results differ between different method variants or even between technical replicates of the same method (Van Nostrand *et al.*, 2016). When high-throughput methods to detect protein-protein interactions became available, one of the initial concerns was the apparent lack of robustness in the results (mostly due to false-positives), which varied in every replicate experiment.

However, efforts were made to produce high quality protein interaction maps, with the addition of several replicates and probability scores for interactions (Braun *et al.*, 2009; Venkatesan *et al.*, 2009). Due to these efforts, nowadays, protein-protein interaction methods such Y2H can be deemed reliable. For high-throughput protein-RNA interaction methods, robust protocols and frameworks are yet to be developed in order for their results to be consistent and reliable, but the development, improvement and standardisation of high-throughput protein-RNA interaction methods is a current challenge and a highly researched topic (Lee and Ule, 2018). Indeed, a recent review article discusses data science issues in analysing CLIP data, describing the impact of using different methods on the results and advocating the need of applying computational quality controls and standardising steps such as read alignment, peak calling and computational modeling of protein–RNA binding sites (Chakrabarti *et al.*, 2018). It is thus predictable that the quality of results from such methods will improve over the next years.

Currently, an alternative to using experimental protein-RNA interaction datasets is to computationally predict interactions. This was the prefered option when analysing protein-lncRNA interactions to research the scaffolding function of lncRNAs and also when predicting RNA regulons (Results, sections 2.1 and 2.4, respectively), since the scope of these studies was proteome- and transcriptome-wide. In fact, most lncRNAs, as well as certain mRNAs, are expressed in a tissue- and condition-specific manner (e.g. cellular stress, DNA damage) (Djebali *et al.*, 2012; Su *et al.*, 2018). Therefore, by being able to predict all possible interactions independently of the context, methods that computationally predict protein-RNA interactions have the advantage of finding interactions that may only occur in certain conditions and thus be difficult to find experimentally. However, computational methods such as catRAPID *omics* also have limitations, namely the maximum length limit for RNAs (1200 nucleotides) (Agostini *et al.*, 2013). Even though certain very long scaffolding lncRNAs such as XIST and NEAT1 are missed, the majority of lncRNAs (>80%) can be assessed with this method. Nevertheless, the size restriction often excludes long 3’UTRs, which were found to be the drivers of mechanisms such as the UDPL, unlike short 3’UTRs (Berkovits and Mayr, 2015). This is one of the reasons why, for the project exploring the formation of protein complexes promoted by 3’UTRs (Results, section 2.3), I decided to primarily use experimentally determined protein-RNA interactions. In addition, publicly available interactions for mRNAs (including 3’UTRs) are more numerous than those for lncRNAs, and certain databases, like the AURA database, provide an extensive compendium of interactions focused on UTRs (Dassi *et al.*, 2014).

3.1.2. Extensive protein-RNA complex prediction by integrating different types of interactions

In this work, I integrate protein-protein interactions with protein-RNA interactions to predict a role of RNAs in assembling protein complexes. Several methods to detect protein-protein interactions and protein-RNA interactions exist (Introduction, sections 1.1.2 and 1.1.4, respectively), as well as comprehensive datasets of protein complexes (Drew *et al.*, 2017), however, the detection of protein-RNA complexes has been largely overlooked. Protein-RNA complexes have been mostly identified through X-ray crystallography and cryo-electron microscopy, on a case by case basis (Patel *et al.*, 2017). Large-scale approaches to study protein-RNA complexes, like the analysis of thousands of MS experiments to produce a compendium of protein complexes, are nonexistent (Huttlin *et al.*, 2017; Patel *et al.*, 2017). The results exposed in this thesis suggest that the formation of protein-RNA complexes is highly widespread, with lncRNAs predicted to scaffold hundreds of known protein complexes, and 3'UTRs predicted to promote the formation of more than a thousand different protein complexes. While an unknown amount of these predictions may be false positives, both the computational and experimental interaction datasets used suggest that RNAs can often be associated to several proteins that are known to act together. One consideration is that the protein-RNA complexes predicted here did not take into account that the binding of two interactors may affect the binding of a third component of the complex. Indeed, catRAPID *omics* does not detect the interaction sites between RNAs and RBPs, and even though several experimental protein-RNA interaction methods may be able identify RBP binding sites and protein-RNA complexes, the datasets used here (AURA database) provide interactions as binary interactions (i.e. one molecule interacting with another single molecule). However, *in vivo*, it is possible that *(i)* two proteins compete for the same RNA binding site, *(ii)* the protein-protein interactions between two or more proteins would prevent an interaction with the RNA. The first issue does not affect the 3'UTR-protein complex predictions, since only the interaction between one RBP and a 3'UTR is required, but it could affect the predictions of lncRNAs scaffolding protein complexes. On the other hand, the second issue could occur for both types of predictions made here.

In this work, expression filters were applied to improve the likelihood of the protein-RNA complex predictions, requiring that all components of a complex can be found in a same tissue. This thus excludes computationally predicted protein-RNA interactions between molecules that had good propensity to interact but were unlikely to occur biologically. Moreover, expression filters retain only complexes with components that are found in the same tissue, excluding cases in which one pair of components interact in

a tissue, but the other components of a complex interact only in a distinct tissue. This last point is relevant for both protein-protein and protein-RNA experimentally identified interactions. However, the human datasets of RNA or protein tissue expression used, although large-scale as in the case of GTEx and HPA (Consortium, 2015; Uhlén *et al.*, 2015), may be incomplete and lead us to miss certain interactions and possible complexes. One such example is the lack of detected expression for the telomerase RNA component (TERC) in the GTEx v6 dataset, a well known scaffolding lncRNA, potentially due to the fact that this RNA (and the Telomerase) is repressed in most normal adult tissues (Cong, Wright and Shay, 2002). Indeed, TERC may exemplify a group of functional transcripts that only act in certain conditions and respond to certain stimuli, such as the RNAs responsible for forming stress-dependent nuclear granules like the NEAT1-promoted paraspeckles (Introduction, section 1.2.3) (Clemson *et al.*, 2009; Fox *et al.*, 2018).

3.2. Prevalence of novel non-coding RNA functions

The complexity of higher organisms is correlated with the amount of non-coding RNAs, such as UTRs and lncRNAs, rather than the number of protein-coding genes of an organism (Mattick, Taft and Faulkner, 2010; Barrett, Fletcher and Wilton, 2012). Non-coding RNAs are found to play an active role in most cellular processes, through interaction with other RNAs, DNA molecules or proteins (Wilusz, Sunwoo and Spector, 2009; Geisler and Coller, 2013). Given the demonstrated impact of non-coding RNAs in biological networks, RNAs are no longer perceived merely as carriers of information, but important parts of biological pathways and their control. Indeed, in light of this, Sidney Altman has proposed to label today's world as the "RNA–Protein World", instead of the most commonly used "Protein World" (Altman, 2013). Thus, the identification and functional characterisation of all new non-coding RNAs is vital to the understanding of cellular mechanisms.

Protein-RNA interactions are key events in a large number of regulatory processes, such as gene imprinting, differentiation and development (Geisler and Coller, 2013; Marchese, Raimondi and Huarte, 2017). Studying protein-RNA interactions can thus give us a broad perspective on lncRNA function. Moreover, I think that with the discovery that the RNA-binding proteome is larger than previously known (Introduction, section 1.1.3), RNAs may impact more biological processes than previously expected, perhaps through yet undisclosed mechanisms. In this thesis, I predict the prevalence of two distinct but mechanistically related RNA functions which are still largely undescribed. I have predicted that the formation of both lncRNA-protein complexes (lncRNA scaffolding; Results, section 2.1) and

3’UTR-protein complexes (Results, section 2.3) in human tissues occurs often, and more than expected by chance. Indeed, my results suggest that as much as half of the known human protein complexes have at least two proteins that may be scaffolded by a certain lncRNA. Similarly, my results indicate that the 3’UTRs of mRNAs encoding EMF proteins can gather hundreds of different combinations of co-interacting proteins.

3.2.1. Scaffolding function of lncRNAs

Even though the definition of lncRNA serving as protein scaffolds (in this thesis termed simply ‘lncRNA scaffolding’) is progressively being established by the community, I consider that a minimum requirement should be the simultaneous, and functional, physical interaction of the lncRNA with two or more proteins. In this thesis, I *in silico* predicted lncRNAs acting as protein scaffolds, based on the interaction between a lncRNA and at least two proteins of a same protein complex or functional protein network module, without information on simultaneous binding (as discussed above) or the functionality of the interactions. This exploratory approach rendered us with more than 800 lncRNAs that may act as scaffolding molecules to thousands of protein complexes and functional network modules, seen as scaffolding candidates whose exact function need to be further characterised. These results indicate that lncRNA scaffolding could be a rather common mechanism in human cells, and that scaffolding lncRNAs possess features that distinguish them from other lncRNAs, such as displaying a metabolic profile characteristic of functional transcripts and containing structurally conserved elements (Ribeiro *et al.*, 2018).

In vivo, RNA scaffolding has been found to be performed by several lncRNA molecules in a number of different contexts. In my perspective, these include: *(i)* lncRNAs that perform their functions as ribonucleoprotein particles (RNPs), such as the TERC RNA (Cech and Steitz, 2014), *(ii)* lncRNAs that interact with specific proteins to form cellular granules, including the NEAT1 lncRNA formation of nuclear paraspeckles (Fox *et al.*, 2018); *(iii)* lncRNAs that gather several components of a pathway (i.e. acting in a similar way to protein scaffolders), such as the LINP1 lncRNA (Zhang *et al.*, 2016); *(iv)* lncRNAs that functionally assemble groups of proteins, like the XIST and HOTAIR lncRNAs (Tsai *et al.*, 2010; Creamer and Lawrence, 2017). The extensive search for scaffolding lncRNAs made in this thesis should englobe all of these cases, and potentially others yet undiscovered.

While the debate of whether most lncRNA molecules are functional or transcriptional noise is ongoing, novel examples of scaffolding lncRNAs are gradually being unveiled (Introduction, section 1.2.2) (Kopp and Mendell, 2018; Uszczynska-Ratajczak *et al.*, 2018). For example, even though RNAs have been

found in the kinetochores in the 1970s, only recently it was revealed that several RNAs transcribed from the centromeric region (albeit at low levels) are bound by kinetochore proteins and incorporated into centromeric chromatin, playing a role in protein colocalization and stabilisation (Talbert and Henikoff, 2018). These include a 1.3kb centromeric lncRNA that associates with HJURP and CENP-A proteins, thus serving essential functions, such as the maintenance of centromere integrity (Quénet and Dalal, 2014), by employing mechanisms that can be considered to be lncRNA scaffolding.

A common known function of lncRNAs is to regulate gene expression, with several dozen nuclear lncRNAs having been implicated in transcription activation and repression, as well as post-transcriptional regulation events (Sun, Hao and Prasanth, 2018). Although for some lncRNAs this regulation seems to be associated to their transcriptional event, many lncRNA molecules have been found to act through the interaction with proteins or protein complexes, such as RNA polymerase II, DNA-binding proteins and chromatin-remodeling complexes, including the polycomb repressive complex 1 and 2 (PRC1 and PRC2) (Long *et al.*, 2017). In my work, I have predicted that the PRC2 complex, as well as other complexes, is scaffolded by more than a hundred lncRNAs. Indeed, some proteins or protein-complexes have also been found to interact with thousands of RNAs in cells, including the PRC2, heterogeneous nuclear RNP (hnRNP) proteins and FUS proteins, although it is unclear if this represents functional binding or is simply the outcome of promiscuity of the interactors (Hendrickson *et al.*, 2016; Long *et al.*, 2017). In such cases, further work needs to be performed to assess the functionality of the lncRNA association with the proteins. In addition, my work predicts that hundreds of lncRNAs may have a scaffolding function. While this function can only be thoroughly verified using experimental methods, I have provided orthogonal evidence that this set of lncRNAs comprises transcripts known to be functional, associated to disease and possess other features of functionality (Ribeiro *et al.*, 2018). Moreover, we validated some of the interactions of the lnc-405 lncRNA with a known protein complex. The mouse ortholog version of this lncRNA (renamed as “Charme” lncRNA) has recently been found to cause myogenic defects and heart remodeling (Ballarino *et al.*, 2018).

Besides acting on gene expression regulation, an increasing amount of lncRNAs and other RNAs are being implicated in RNP granule formation, such as the germ granules, paraspeckles, Cajal bodies, stress granules and P-bodies (Van Treeck and Parker, 2018). Recent findings suggest that the formation of these granules initially stems from specific protein-protein and protein-RNA interactions, followed by promiscuous binding mostly through intrinsically disordered regions in RBPs (Protter *et al.*, 2018). However, it was recently found that intermolecular and as well as self RNA-RNA interactions also contribute to RNP granule formation, particularly contributing to the recruitment of specific RNAs to

stress granules (Van Treeck *et al.*, 2018). This means that protein-protein, protein-RNA and RNA-RNA interactions can work together in order to form RNP assemblies (Figure 3.1). RNA-centric experimental methods to detect protein-RNA interactions focus on single RNAs, whereas protein-centric methods may identify many protein-interacting RNAs (Barra and Leucci, 2017), but in neither case it is evident whether these RNAs may interact together. Novel methods to predict lncRNA scaffolding could integrate RNA-RNA interaction data in order to produce more detailed models and find cases in which groups of proteins may be scaffolded by a group of lncRNAs, as observed in certain RNP granules.

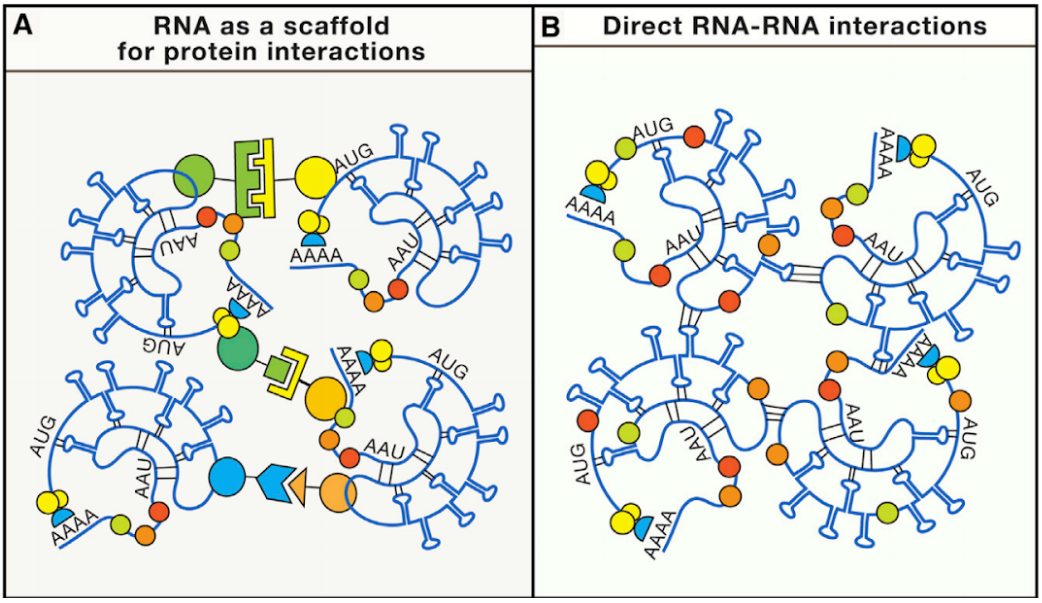


Figure 3.1 | Model of granule formation mediated by RNAs. (A) RNAs serve as scaffolds by interacting with RBPs, which may in turn interact with other proteins. (B) RNAs can also play a role in granule formation through interaction with RBPs but also with other RNAs. Figure adapted from (Van Treeck *et al.*, 2018).

It is now known that mRNAs can have functions unrelated to their coding-potential (Auboeuf, 2018). Indeed, mRNAs and other transcripts stemming from protein-coding genes have been found to participate in RNP granule formation, such as P-bodies and stress granules, where functionally-related mRNAs can be assembled and stored in the cytoplasm, allowing coordinated co-translation upon granule dissolution (Mitchell *et al.*, 2013; Nielsen, Hansen and Christiansen, 2016; Hubstenberger *et al.*, 2017). Interestingly, the expression coordination through RNA regulons (Results, section 2.4) may involve RNA granule formation, as in the case of the *Drosophila* SFPQ RBP regulating multiple mRNAs encoding proteins that promote axon survival (Cosker *et al.*, 2016). Besides known mRNA transcripts, other sense-strand RNAs that overlap protein-coding regions have been found to interact with DNA methyltransferases such as

DNMT1 and DNMT3a (Savell *et al.*, 2016). Furthermore, both pre-mRNA and intronic RNA sequences were found to functionally interact with the PRC2 complex, thus playing a role in regulating transcription and chromatin function through protein interactions (Guil *et al.*, 2012; Skalska *et al.*, 2017). Overall, it is possible that, similarly to some non-coding genes, certain protein-coding genes could produce transcripts that act through mechanisms involving protein binding, such as RNA scaffolding. The novel methods developed during this thesis to predict lncRNA scaffolding could further be applied to discover the scaffolding potential of transcripts derived from protein-coding genes, something that has not been attempted before.

3.2.2. 3'UTR-protein complex formation

Alike non-coding RNAs, 3'UTR regions of mRNAs have been found to perform functions through protein interactions. 3'UTR protein binding occurs throughout the life cycle of the mRNA, being important for the mRNA processing, transport, stability and translation (Introduction, section 1.3.2) (Szostak and Gebauer, 2013; Tian and Manley, 2017). Moreover, 3'UTRs have been found to mediate protein interactions in protein complexes formed during translation (Duncan and Mata, 2011; Berkovits and Mayr, 2015; Chartron, Hunt and Frydman, 2016). In particular, the UTR-dependent protein localisation (UDPL) mechanism was found to translocate CD47 to the plasma membrane, evidencing a case in which the interaction between a 3'UTR and an RBP promotes the recruitment to the translation site of other proteins that interact with the 3'UTR/mRNA's cognate protein (Berkovits and Mayr, 2015). So far, before the work in this thesis, the UDPL mechanism has not been searched systematically and has only been found in a few cases. On the other hand, formation of mRNA-promoted protein complexes involving the mRNA's cognate protein seem to be widespread in yeast. For example, the SET1 mRNA and nascent protein co-purify with several proteins of the SET1 histone methyltransferase complex during translation (Halbach *et al.*, 2009). Moreover, in 2011, Duncan & Mata showed that cotranslational complexes involving the nascent protein and their cognate mRNA occured in 38% of the tested cases (Duncan and Mata, 2011).

The formation of 3'UTR-protein complexes can be particularly interesting because 3'UTRs are found to interact with hundreds of different RBPs and these binding events can be dynamic and dependent on the local environment (Freeberg *et al.*, 2013; Dassi *et al.*, 2014). For example, post-transcriptional modifications in RBPs can alter their interaction with 3'UTRs (Mayr, 2017) and alternatively processed pre-mRNAs can have alternative cellular functions, such as the p53 mRNA nuclear trafficking control of the MDM2 protein through its interaction (Gajjar *et al.*, 2012; Auboeuf, 2018). Indeed, the 3'UTR

alternative polyadenylation of the CD47 mRNA was found to be responsible for the differential cellular localisation of its nascent protein (Berkovits and Mayr, 2015). Given the presence, absence and availability of binding motifs in 3'UTRs, RBPs may bind only one of the 3'UTR isoforms produced and not others (Mayr, 2017). Besides CD47, this was also shown for differential HuR RBP binding of the α-synuclein (SNCA) mRNA depending on the 3'UTR isoform expressed (Marchese *et al.*, 2017). Moreover, the competition and binding of different RBPs to a same 3'UTR could change the ability to recruit different RBP-interacting proteins, thus potentially forming 3'UTR-protein complexes with diverse compositions, possibly with different functions. It can also be speculated that the composition of 3'UTR-protein complexes involving a certain mRNA can vary depending on the cell type and cellular conditions in which they are present, alike moonlighting proteins performing different functions in distinct conditions (Introduction, section 1.4.3) (Jeffery, 2018).

It is predictable that, besides affecting the cellular localisation of proteins, the formation of 3'UTR-protein complexes could function to regulate other cellular processes. For example, the two heteromeric ion channel subunits (hERG 1a and 1b) assemble co-translationally in association with their cognate mRNAs (Liu *et al.*, 2016). In fact, co-translational protein complex assembly was recently suggested to occur for most cytoplasmic protein complexes of *Saccharomyces cerevisiae* (Shiber *et al.*, 2018). This was found by performing selective ribosome profiling, which isolates ribosomes synthesizing nascent proteins already interacting with another protein, such as another subunit of the same protein complex (Shiber *et al.*, 2018). Considering this, two models in which the mRNAs encoding for the several protein complex subunits could play a role in the protein complex assembly have been proposed (Mayr, 2018a). In one model, the two mRNAs may be brought to proximity through RBP binding, and possibly involving RNA granule formation (Figure 3.2a,b). Alternatively, an RBP may bind the 3'UTR of an mRNA and recruit a protein that assembles with its cognate nascent protein (Figure 3.2c). Occurences of the latter model are prone to be identified by the approaches undertaken in this thesis (Results, section 2.3). Moreover, these models are also consistent with the RNA regulon theory, in which an RBP regulates functionally-related mRNAs (Results, section 2.4).

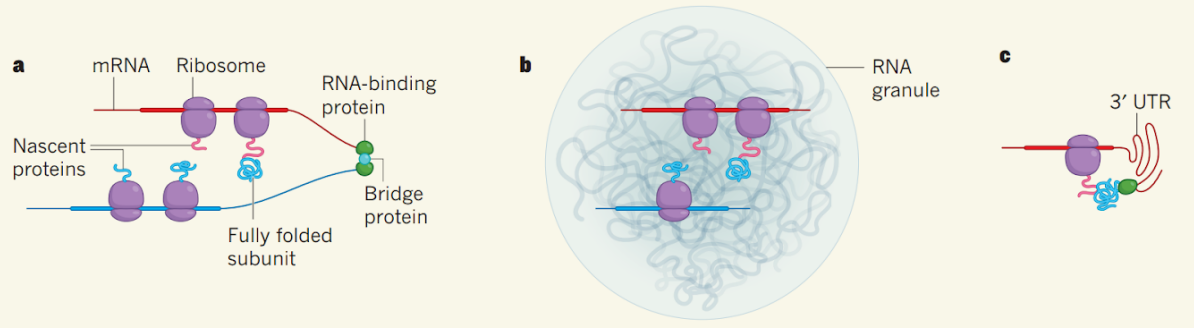


Figure 3.2 | mRNA binding in promoting protein complex assembly. (a) the mRNAs encoding two subunits of the same protein complex are bound by a same RBP, bringing the nascent proteins into proximity; (b) the mRNAs of the protein complex subunits are found in the same RNA granule; (c) the 3'UTR of the mRNA interacts with an RBP that in its turn interacts with another subunit of the protein complex (i.e. the 3'UTR-protein complex formation described in this thesis). Figure adapted from (Mayr, 2018a).

The results described in this thesis (Results, section 2.3) indicate that hundreds of different 3'UTR-protein complexes may be formed via the mRNAs of moonlighting protein candidates (EMF proteins), comprising more than a thousand different protein compositions. The effector proteins found in these predicted complexes include motor proteins and other proteins involved in macromolecule transport, which could be involved in plasma membrane translocation of the nascent protein, but most complexes involve other proteins participating in many diverse biological processes. It is possible that many of the 3'UTR-protein complexes predicted play a role in protein complex assembly, as described above. In fact, Christine Mayr has recently reported that an alternative 3'UTR of the human E3 ubiquitin ligase BIRC3 is implicated in the assembly of a protein complex involving its cognate nascent protein, evidencing the possibility of co-translational protein complex assembly to be mediated by 3'UTRs (Mayr, 2018b). However, a limitation of my work is that 3'UTR-protein complexes were predicted using protein-protein and protein-RNA interactions that have been determined in diverse cellular compartments, not specifically associated to ribosomes and co-translational interactions. Indeed, large-scale datasets of co-translational protein-protein or protein-RNA interactions, such as from selective ribosome profiling methods, are yet inexistant. Regardless, the full extent of 3'UTR-protein complex formation is yet unknown and the work presented in this thesis provides a preliminary estimation of the amount of 3'UTR-protein complexes that could possibly be formed in human cells, based on interactions found experimentally. Even though this work focused on moonlighting protein candidates, whose characteristics make them more prone to form 3'UTR-protein complexes, thousands of complexes involving other proteins were also found. Further efforts will involve the analysis of 3'UTR-protein complexes liable to be formed throughout the human proteome, outside the context of protein multifunctionality, since the

formation of these complexes may be a general and prevalent cellular mechanism. Indeed, although there is evidence that certain biological processes, such as metabolic pathways (Shiber *et al.*, 2018), can be coordinated cotranslationally, the function of many co-translational protein interactions is still unknown.

3.2.3. Moonlighting proteins and their regulation by 3'UTRs

How cells cope with proteins that can perform very different biological functions is an intriguing question. The mechanisms employed by the cell to regulate these activities, even though described for several proteins (Introduction, section 1.4.3), remain to be elucidated for many moonlighting proteins. In this thesis I suggest that the formation of 3'UTR-protein complexes may be one way in which cognate mRNAs and their interactors regulate the multifunctionality of proteins. More particularly, these complexes could play a role in the switch between the several functions of moonlighting proteins by changing the subcellular localisation of the nascent protein, as with the UDPL mechanism (Berkovits and Mayr, 2015). Moreover, since 3'UTR-protein complexes can be formed by many different protein components, and thus associating the nascent protein with different effector (intermediate) proteins, I speculate that these complexes may regulate moonlighting protein function through other mechanisms. For example, 3'UTR-proteins complexes could affect the nascent protein folding (e.g. through a chaperon effector), the participation in different protein complexes, the PTMs of the nascent protein, or indeed any other property or state that involves interactions with proteins. Interestingly, several of the 9 protein complexes found to be assembled co-translation in Shiber *et al.* include known yeast moonlighting or multifunctional proteins (PFK1, EGD2, GluRS, MetRS), suggesting that cotranslational protein complex formation may indeed be important in the regulation of moonlighting protein function (Shiber *et al.*, 2018).

Human extreme multifunctional (EMF) proteins were chosen as a model in which to study 3'UTR-protein complex formation due to their features, such as having long 3'UTRs (Results, section 2.3). The use of a large and up-to-date dataset of EMF proteins, such as the one produced here and present in MoonDB 2.0 (Results, section 2.2), has been instrumental to perform large-scale analysis on the potential regulatory role of 3'UTR-protein complex formation in protein multifunctionality. Indeed, the prevalence of moonlighting proteins in model organisms is not fully established and their discovery, either through experimental or computational methods, is a continuous effort (Chapple *et al.*, 2015; Khan, Bhuiyan and Kihara, 2017; Espinosa-Cantú *et al.*, 2018). The computational prediction of EMF proteins employed here was found to be highly dependent on the amount of protein-protein interaction and GO term annotation data available for a species. Both types of data are expected to continuously grow over the years for

model species, and thus the development of a semi-automated pipeline such as the one created for MoonDB 2.0 facilitates the keeping of an up-to-date dataset of EMF proteins (del-Toro *et al.*, 2013; The Gene Ontology Consortium, 2017). Indeed, MoonDB will be updated every year with novel candidates and functional annotations, thus providing a reliable resource to the community.

4. Conclusion

The work carried out during my Ph.D. studies at the TAGC, Inserm U1090 (Marseille, France), partly in collaboration with the CRG (Barcelona, Spain), has been composed into a Ph.D. thesis entitled ‘Discovery of the role of protein-RNA interactions in protein multifunctionality and cellular complexity’.

This thesis presents my contributions to several lines of research which can be presently considered as ‘hot topics’, receiving increasing attention from the community. Indeed, this thesis tackles several open uncharted questions, namely regarding the potential extent and importance of 3’UTR-protein complex formation in human cells and the prevalence and regulation of moonlighting proteins. Furthermore, the results presented here feed the debate concerning the importance of non-coding regions of the genome, by suggesting that a substantial fraction of human lncRNAs may function as scaffolding molecules.

To the best of my knowledge, neither the prevalence of scaffolding lncRNAs, nor the extent of possible 3’UTR-protein complex formation had been previously revealed transcriptome-wide for any species. In fact, prior to my work, no methods to tackle these questions large-scale were available. Thus, this thesis introduces very innovative and extensive methods to approach these subjects, giving a first general overview of the potential impact of these ill-known RNA functions. Partially due to the lack of knowledge in the topics approached here, some of the conclusions described in this thesis can be viewed as speculative. Indeed, this work is largely based upon predictive methods that yet require experimental validation. Effectively, this computational work leaves many questions open, namely the actual essentiality of lncRNA with scaffolding functions in cells, and the functionality of the 3’UTR-protein complexes predicted, which can only be verified experimentally.

Due to its originality and large-scale scope, I believe that my work can have broad implications for future research. Indeed, here I provide large predictive datasets of protein-RNA interaction networks, scaffolding lncRNAs, 3’UTR-protein complexes as well as moonlighting proteins, which can be further exploited by the community. Overall, the innovative approaches developed and applied in this thesis can help pave the way to a better understanding of the role of key molecular components, such as the non-coding parts of our genomes, in the making of complex systems.

Bibliography

Agostini, F. et al. (2013) “catRAPID omics: a web server for large-scale prediction of protein-RNA interactions.,” Bioinformatics (Oxford, England), 29(22), pp. 2928–2930.

Akiva, E. et al. (2012) “A dynamic view of domain-motif interactions.,” PLoS computational biology, 8(1), p. e1002341.

Alonso-López, D. et al. (2016) “APID interactomes: providing proteome-based interactomes with controlled quality for multiple species and derived networks.,” Nucleic acids research, 44(W1), pp. W529–W535.

Altman, S. (2013) “The RNA-Protein World.,” RNA (New York, N.Y.), 19(5), pp. 589–590.

Amblee, V. and Jeffery, C. J. (2015) “Physical Features of Intracellular Proteins that Moonlight on the Cell Surface.,” PloS one, 10(6), p. e0130575.

Amos-Binks, A. et al. (2011) “Binding site prediction for protein-protein interactions and novel motif discovery using re-occurring polypeptide sequences.,” BMC bioinformatics, 12, p. 225.

Anderson, P. (2010) “Post-transcriptional regulons coordinate the initiation and resolution of inflammation.,” Nature reviews. Immunology, 10(1), pp. 24–35.

Andreassi, C. and Riccio, A. (2009) “To localize or not to localize: mRNA fate is in 3’UTR ends.,” Trends in cell biology, 19(9), pp. 465–474.

Armaos, A., Cirillo, D. and Gaetano Tartaglia, G. (2017) “omiXcore: a web server for prediction of protein interactions with large RNA.,” Bioinformatics (Oxford, England), 33(19), pp. 3104–3106.

Auboeuf, D. (2018) “Alternative mRNA processing sites decrease genetic variability while increasing functional diversity,” Transcription. Taylor & Francis, 9(2), pp. 75–87.

Ballarino, M. et al. (2018) “Deficiency in the nuclear long noncoding RNA causes myogenic defects and heart remodeling in mice.,” The EMBO journal, 37(18).

Barra, J. and Leucci, E. (2017) “Probing Long Non-coding RNA-Protein Interactions.,” Frontiers in molecular biosciences, 4, p. 45.

Barrett, L. W., Fletcher, S. and Wilton, S. D. (2012) “Regulation of eukaryotic gene expression by the untranslated gene regions and other non-coding elements.,” Cellular and molecular life sciences : CMLS, 69(21), pp. 3613–3634.

Bartel, D. P. (2018) “Metazoan MicroRNAs.,” Cell, 173(1), pp. 20–51.

Becker, E. et al. (2012) “Multifunctional proteins revealed by overlapping clustering in protein interaction network.,” Bioinformatics (Oxford, England), 28(1), pp. 84–90.

Beckmann, B. M. et al. (2015) “The RNA-binding proteomes from yeast to man harbour conserved enigmRBPs.,” Nature communications, 6, p. 10127.

Beckmann, B. M., Castello, A. and Medenbach, J. (2016) “The expanding universe of ribonucleoproteins: of novel RNA-binding proteins and unconventional interactions.,” Pflugers Archiv : European journal of physiology, 468(6), pp. 1029–1040.

Bellucci, M. et al. (2011) “Predicting protein associations with long noncoding RNAs.,” Nature methods, 8(6), pp. 444–445.

Beltrao, P. et al. (2013) “Evolution and functional cross-talk of protein post-translational modifications,” Molecular Systems Biology. Wiley Online Library, 9.

Berkovits, B. D. and Mayr, C. (2015) “Alternative 3’ UTRs act as scaffolds to regulate membrane protein localization.,” Nature, 522(7556), pp. 363–367.

Blackinton, J. G. and Keene, J. D. (2014) “Post-transcriptional RNA regulons affecting cell cycle and proliferation.,” Seminars in cell & developmental biology, 34, pp. 44–54.

Bonini, S. et al. (2003) “Nerve growth factor: neurotrophin or cytokine?,” International archives of allergy and immunology, 131(2), pp. 80–84.

Braun, P. et al. (2009) “An experimentally derived confidence score for binary protein-protein interactions.,” Nature methods, 6(1), pp. 91–97.

Briggs, S. F. and Reijo Pera, R. A. (2014) “X chromosome inactivation: recent advances and a look forward.,” Current opinion in genetics & development, 28, pp. 78–82.

Brückner, A. et al. (2009) “Yeast two-hybrid, a powerful tool for systems biology.,” International journal of molecular sciences, 10(6), pp. 2763–2788.

Brun, C. et al. (2003) “Functional classification of proteins for the prediction of cellular function from a protein-protein interaction network.,” Genome biology, 5(1), p. R6.

Brun, C., Herrmann, C. and Guénoche, A. (2004) “Clustering proteins from interaction networks for the prediction of cellular functions.,” BMC bioinformatics, 5, p. 95.

Calabretta, S. and Richard, S. (2015) “Emerging Roles of Disordered Sequences in RNA-Binding Proteins.,” Trends in biochemical sciences, 40(11), pp. 662–672.

Carlomagno, T. (2014) “Present and future of NMR for RNA-protein complexes: a perspective of integrated structural biology.,” Journal of magnetic resonance (San Diego, Calif. : 1997), 241, pp.

126–136.

Castello, A. et al. (2013) “System-wide identification of RNA-binding proteins by interactome capture.,” Nature protocols, 8(3), pp. 491–500.

Castello, A. et al. (2016) “Comprehensive Identification of RNA-Binding Domains in Human Cells.,” Molecular cell, 63(4), pp. 696–710.

Cech, T. R. and Steitz, J. A. (2014) “The noncoding RNA revolution-trashing old rules to forge new ones.,” Cell, 157(1), pp. 77–94.

Chakrabarti, A. M. et al. (2018) “Data Science Issues in Studying Protein–RNA Interactions with CLIP Technologies,” Annual Review of Biomedical Data Science, 1(1).

Chapple, C. E. et al. (2015) “Extreme multifunctional proteins identified from a human protein interaction network.,” Nature communications, 6, p. 7412.

Chapple, C. E. and Brun, C. (2015) “Redefining protein moonlighting.,” Oncotarget, 6(19), pp. 16812–16813.

Chapple, C. E., Herrmann, C. and Brun, C. (2015) “PrOnto database : GO term functional dissimilarity inferred from biological data.,” Frontiers in genetics, 6, p. 200.

Chartron, J. W., Hunt, K. C. L. and Frydman, J. (2016) “Cotranslational signal-independent SRP preloading during membrane targeting.,” Nature, 536(7615), pp. 224–228.

Chatr-Aryamontri, A. et al. (2017) “The BioGRID interaction database: 2017 update.,” Nucleic acids research, 45(D1), pp. D369–D379.

Chaudhury, A., Chander, P. and Howe, P. H. (2010) “Heterogeneous nuclear ribonucleoproteins (hnRNPs) in cellular processes: Focus on hnRNP E1’s multifunctional regulatory roles.,” RNA (New York, N.Y.), 16(8), pp. 1449–1462.

Chen, C. et al. (2018) “MoonProt 2.0: an expansion and update of the moonlighting proteins database.,” Nucleic acids research, 46(D1), pp. D640–D644.

Chen, C.-Y. et al. (2012) “Lengthening of 3’UTR increases with morphological complexity in animal evolution.,” Bioinformatics (Oxford, England), 28(24), pp. 3178–3181.

Chen, L.-L. (2016) “The biogenesis and emerging roles of circular RNAs.,” Nature reviews. Molecular cell biology, 17(4), pp. 205–211.

Cheng, L. and Leung, K.-S. (2018) “Identification and characterization of moonlighting long non-coding RNAs based on RNA and protein interactome.,” Bioinformatics (Oxford, England).

Chu, C. et al. (2015) “Systematic discovery of Xist RNA binding proteins.,” Cell, 161(2), pp. 404–416.

Chujo, T., Yamazaki, T. and Hirose, T. (2015) “Architectural RNAs (arcRNAs): A class of long

noncoding RNAs that function as the scaffold of nuclear bodies," Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms. Elsevier B.V., pp. 1–8.

Cirillo, D., Marchese, D., et al. (2014) "Constitutive patterns of gene expression regulated by RNA-binding proteins.," Genome biology, 15(1), p. R13.

Cirillo, D., Livi, C. M., et al. (2014) "Discovery of protein-RNA networks.," Molecular bioSystems, 10(7), pp. 1632–1642.

Cirillo, D. et al. (2016) "Quantitative predictions of protein interactions with long noncoding RNAs.," Nature methods, 14(1), pp. 5–6.

Cirillo, D., Agostini, F. and Tartaglia, G. G. (2013) "Predictions of protein-RNA interactions: Protein-RNA interactions," Wiley Interdisciplinary Reviews: Computational Molecular Science, 3(2).

Clark, B. S. and Blackshaw, S. (2014) "Long non-coding RNA-dependent transcriptional regulation in neuronal development and disease.," Frontiers in genetics, 5, p. 164.

Clemson, C. M. et al. (2009) "An architectural role for a nuclear noncoding RNA: NEAT1 RNA is essential for the structure of paraspeckles.," Molecular cell, 33(6), pp. 717–726.

Cong, Y.-S., Wright, W. E. and Shay, J. W. (2002) "Human telomerase and its regulation.," Microbiology and molecular biology reviews : MMBR, 66(3), p. 407–25, table of contents.

Consortium, T. Gt. (2015) "The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans," Science (80-.)., 348, pp. 648–660.

Copley, S. D. (2012) "Moonlighting is mainstream: paradigm adjustment required.," BioEssays : news and reviews in molecular, cellular and developmental biology, 34(7), pp. 578–588.

Cosker, K. E. et al. (2016) "The RNA-binding protein SFPQ orchestrates an RNA regulon to promote axon viability.," Nature neuroscience, 19(5), pp. 690–696.

Costanzo, M. et al. (2016) "A global genetic interaction network maps a wiring diagram of cellular function.," Science (New York, N.Y.), 353(6306).

Creamer, K. M. and Lawrence, J. B. (2017) "RNA: a window into the broader role of RNA in nuclear chromosome architecture.," Philosophical transactions of the Royal Society of London. Series B, Biological sciences, 372(1733).

Cusick, M. E. et al. (2005) "Interactome: gateway into systems biology.," Hum Mol Genet, 14 Spec No. 2, p. 171.

Danan, C., Manickavel, S. and Hafner, M. (2016) "PAR-CLIP: A Method for Transcriptome-Wide Identification of RNA Binding Protein Interaction Sites.," Methods in molecular biology (Clifton, N.J.), 1358, pp. 153–173.

Darnell, R. B. (2010) “HITS-CLIP: panoramic views of protein-RNA regulation in living cells.,” Wiley interdisciplinary reviews. RNA, 1(2), pp. 266–286.

Dassi, E. et al. (2013) “Hyper conserved elements in vertebrate mRNA 3’-UTRs reveal a translational network of RNA-binding proteins controlled by HuR.,” Nucleic acids research, 41(5), pp. 3201–3216.

Dassi, E. et al. (2014) “AURA 2: Empowering discovery of post-transcriptional networks,” Translation. Taylor & Francis, 2(1).

Davey, N. E. et al. (2012) “Attributes of short linear motifs.,” Molecular bioSystems, 8(1), pp. 268–281.

David, A. et al. (2012) “Nuclear translation visualized by ribosome-bound nascent chain puromycylation.,” The Journal of cell biology, 197(1), pp. 45–57.

De Las Rivas, J. and Fontanillo, C. (2010) “Protein-protein interactions essentials: key concepts to building and analyzing interactome networks.,” PLoS computational biology, 6(6), p. e1000807.

Delebecque, C. J., Silver, P. A. and Lindner, A. B. (2012) “Designing and using RNA scaffolds to assemble proteins in vivo.,” Nature protocols, 7(10), pp. 1797–1807.

Derrien, T. et al. (2012) “The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression.,” Genome research, 22(9), pp. 1775–1789.

Di Giammartino, D. C., Nishida, K. and Manley, J. L. (2011) “Mechanisms and consequences of alternative polyadenylation,” Molecular cell. United States: Elsevier Inc, 43(6), pp. 853–866.

Didier, G., Brun, C. and Baudot, A. (2015) “Identifying communities from multiplex biological networks.,” PeerJ, 3, p. e1525.

tom Dieck, S. et al. (2015) “Direct visualization of newly synthesized target proteins in situ.,” Nature methods, 12(5), pp. 411–414.

Dieterich, D. C. et al. (2010) “In situ visualization and dynamics of newly synthesized proteins in rat hippocampal neurons.,” Nature neuroscience, 13(7), pp. 897–905.

Dimri, S., Basu, S. and De, A. (2016) “Use of BRET to Study Protein-Protein Interactions In Vitro and In Vivo.,” Methods in molecular biology (Clifton, N.J.), 1443, pp. 57–78.

Dinkel, H. et al. (2016) “ELM 2016--data update and new functionality of the eukaryotic linear motif resource.,” Nucleic acids research, 44(D1), pp. D294–D300.

Djebali, S. et al. (2012) “Landscape of transcription in human cells.,” Nature, 489(7414), pp. 101–108.

Dominguez, D. et al. (2018) “Sequence, Structure, and Context Preferences of Human RNA Binding Proteins.,” Molecular cell, 70(5), p. 854–867.e9.

Drew, K. et al. (2017) “Integration of over 9,000 mass spectrometry experiments builds a global map of

human protein complexes.," Molecular systems biology, 13(6), p. 932.

Duncan, C. D. S. and Mata, J. (2011) "Widespread cotranslational formation of protein complexes.," PLoS genetics, 7(12), p. e1002398.

Dunham, W. H., Mullin, M. and Gingras, A.-C. (2012) "Affinity-purification coupled to mass spectrometry: basic principles and strategies.," Proteomics, 12(10), pp. 1576–1590.

Dutertre, M. et al. (2014) "A recently evolved class of alternative 3'-terminal exons involved in cell cycle regulation by topoisomerase inhibitors.," Nature communications, 5, p. 3395.

Egan, E. D. and Collins, K. (2010) "Specificity and stoichiometry of subunit interactions in the human telomerase holoenzyme assembled in vivo.," Molecular and cellular biology, 30(11), pp. 2775–2786.

Ellis, R. J. (2001) "Macromolecular crowding: obvious but underappreciated," Trends in Biochemical Sciences. Elsevier, 26, pp. 597–604.

El-Manzalawy, Y. et al. (2016) "FastRNABindR: Fast and Accurate Prediction of Protein-RNA Interface Residues.," PloS one, 11(7), p. e0158445.

Engreitz, J., Lander, E. S. and Guttman, M. (2015) "RNA antisense purification (RAP) for mapping RNA interactions with chromatin.," Methods in molecular biology (Clifton, N.J.), 1262, pp. 183–197.

Espinosa-Cantú, A. et al. (2018) "Protein Moonlighting Revealed by Noncatalytic Phenotypes of Yeast Enzymes.," Genetics, 208(1), pp. 419–431.

Fabregat, A. et al. (2018) "The Reactome Pathway Knowledgebase.," Nucleic acids research, 46(D1), pp. D649–D655.

Fang, S. et al. (2018) "NONCODEV5: a comprehensive annotation database for long non-coding RNAs.," Nucleic acids research, 46(D1), pp. D308–D314.

Fang, Y. and Fullwood, M. J. (2016) "Roles, Functions, and Mechanisms of Long Non-coding RNAs in Cancer.," Genomics, proteomics & bioinformatics, 14(1), pp. 42–54.

Fields, S. and Song, O. (1989) "A novel genetic system to detect protein-protein interactions.," Nature, 340(6230), pp. 245–246.

Flores, J. K., Walshe, J. L. and Ataide, S. F. (2014) "RNA and RNA–Protein Complex Crystallography and its Challenges," Australian Journal of Chemistry, 67(12).

Fox, A. H. et al. (2018) "Paraspeckles: Where Long Noncoding RNA Meets Phase Separation.," Trends in biochemical sciences, 43(2), pp. 124–135.

Franco-Serrano, L. et al. (2018) "MultitaskProtDB-II: an update of a database of multitasking/moonlighting proteins.," Nucleic acids research, 46(D1), pp. D645–D648.

Freeberg, M. A. et al. (2013) "Pervasive and dynamic protein binding sites of the mRNA transcriptome in

Saccharomyces cerevisiae.,” Genome biology, 14(2), p. R13.

Gajjar, M. et al. (2012) “The p53 mRNA-Mdm2 interaction controls Mdm2 nuclear trafficking and is required for p53 activation following DNA damage.,” Cancer cell, 21(1), pp. 25–35.

Gancedo, C., Flores, C.-L. and Gancedo, J. M. (2016) “The Expanding Landscape of Moonlighting Proteins in Yeasts.,” Microbiology and molecular biology reviews : MMBR, 80(3), pp. 765–777.

Garbett, D. and Bretscher, A. (2014) “The surprising dynamics of scaffolding proteins.,” Molecular biology of the cell, 25(16), pp. 2315–2319.

Garcia-Moreno, M., Järvelin, A. I. and Castello, A. (2018) “Unconventional RNA-binding proteins step into the virus-host battlefront.,” Wiley interdisciplinary reviews. RNA, p. e1498.

Gawronski, A. R. et al. (2018) “MechRNA: prediction of lncRNA mechanisms from RNA-RNA and RNA-protein interactions.,” Bioinformatics (Oxford, England), 34(18), pp. 3101–3110.

Ge, M., Li, A. and Wang, M. (2016) “A Bipartite Network-based Method for Prediction of Long Non-coding RNA-protein Interactions.,” Genomics, proteomics & bioinformatics, 14(1), pp. 62–71.

Geisler, S. and Coller, J. (2013) “RNA in unexpected places: long non-coding RNA functions in diverse cellular contexts.,” Nature reviews. Molecular cell biology, 14(11), pp. 699–712.

Gerber, A. P., Herschlag, D. and Brown, P. O. (2004) “Extensive Association of Functionally and Cytotopically Related mRNAs with Puf Family RNA-Binding Proteins in Yeast,” PLoS Biol. Public Library of Science, 2(3), p. e79.

Gerstberger, S., Hafner, M. and Tuschl, T. (2014) “A census of human RNA-binding proteins.,” Nature reviews. Genetics, 15(12), pp. 829–845.

Glock, C., Heumüller, M. and Schuman, E. M. (2017) “mRNA transport & local translation in neurons.,” Current opinion in neurobiology, 45, pp. 169–177.

Gong, J. et al. (2018) “RISE: a database of RNA interactome from sequencing experiments.,” Nucleic acids research, 46(D1), pp. D194–D201.

Good, M. C., Zalatan, J. G. and Lim, W. A. (2011) “Scaffold proteins: hubs for controlling the flow of cellular information.,” Science (New York, N.Y.), 332(6030), pp. 680–686.

Gruber, A. J. et al. (2016) “A comprehensive analysis of 3’ end sequencing data sets reveals novel polyadenylation signals and the repressive role of heterogeneous ribonucleoprotein C on cleavage and polyadenylation.,” Genome research, 26(8), pp. 1145–1159.

Gruber, A. R. et al. (2014) “Means to an end: mechanisms of alternative polyadenylation of messenger RNA precursors.,” Wiley interdisciplinary reviews. RNA, 5(2), pp. 183–196.

Grünenfelder, B. and Winzeler, E. A. (2002) “Treasures and traps in genome-wide data sets: case

examples from yeast," Nature reviews. Genetics, 3(9), pp. 653–61.

Guil, S. et al. (2012) "Intronic RNAs mediate EZH2 regulation of epigenetic targets.," Nature structural & molecular biology, 19(7), pp. 664–670.

Halbach, A. et al. (2009) "Cotranslational assembly of the yeast SET1C histone methyltransferase complex.," The EMBO journal, 28(19), pp. 2959–2970.

Hao, Y. et al. (2016) "NPInter v3.0: an upgraded database of noncoding RNA-associated interactions.," Database : the journal of biological databases and curation, 2016.

Harrow, J. et al. (2012) "GENCODE: The reference human genome annotation for The ENCODE Project," pp. 1760–1774.

Hartwell, L. H. et al. (1999) "From molecular to modular cell biology," Nature, 402, p. 47.

Helder, S. et al. (2016) "Determinants of affinity and specificity in RNA-binding proteins.," Current opinion in structural biology, 38, pp. 83–91.

Hellman, L. M. and Fried, M. G. (2007) "Electrophoretic mobility shift assay (EMSA) for detecting protein–nucleic acid interactions," Nature protocols. Nature Publishing Group, 2(8), pp. 1849–1861.

Henderson, B. and Martin, A. C. R. (2014) "Protein moonlighting: a new factor in biology and medicine.," Biochemical Society transactions, 42(6), pp. 1671–1678.

Hendrickson, DG. et al. (2016) "Widespread RNA binding by chromatin-associated proteins.," Genome biology, 17, p. 28.

Hennig, J. and Sattler, M. (2015) "Deciphering the protein-RNA recognition code: combining large-scale quantitative methods with structural biology.," BioEssays : news and reviews in molecular, cellular and developmental biology, 37(8), pp. 899–908.

Hentze, M. W. et al. (2018) "A brave new world of RNA-binding proteins.," Nature reviews. Molecular cell biology, 19(5), pp. 327–341.

Hermjakob, H. et al. (2004) "IntAct: an open source molecular interaction database," Nucleic Acids Res, (32 Database), pp. D452-455.

Herzel, L. et al. (2017) "Splicing and transcription touch base: co-transcriptional spliceosome assembly and function.," Nature reviews. Molecular cell biology, 18(10), pp. 637–650.

Herzel, L., Straube, K. and Neugebauer, K. M. (2018) "Long-read sequencing of nascent RNA reveals coupling among RNA processing events.," Genome research, 28(7), pp. 1008–1019.

Hon, C.-C. et al. (2017) "An atlas of human long non-coding RNAs with accurate 5' ends.," Nature, 543(7644), pp. 199–204.

Hu, B. et al. (2017) "POSTAR: a platform for exploring post-transcriptional regulation coordinated by

RNA-binding proteins.,” Nucleic acids research, 45(D1), pp. D104–D114.

Hu, W.L. et al. (2018) “GUARDIN is a p53-responsive long non-coding RNA that is essential for genomic stability.,” Nat Cell Biol, 20(4):492-502.

Huang, R. et al. (2018) “Transcriptome-wide discovery of coding and noncoding RNA-binding proteins.,” Proceedings of the National Academy of Sciences of the United States of America, 115(17), pp. E3879–E3887.

Huberts, D. H. E. W. and van der Klei, I. J. (2010) “Moonlighting proteins: an intriguing mode of multitasking.,” Biochimica et biophysica acta, 1803(4), pp. 520–525.

Hubstenberger, A. et al. (2017) “P-Body Purification Reveals the Condensation of Repressed mRNA Regulons.,” Molecular cell, 68(1), p. 144–157.e5.

Huttlin, E. L. et al. (2015) “The BioPlex Network: A Systematic Exploration of the Human Interactome.,” Cell, 162(2), pp. 425–440.

Huttlin, E. L. et al. (2017) “Architecture of the human interactome defines protein communities and disease networks.,” Nature, 545(7655), pp. 505–509.

Imig, J., Kanitz, A. and Gerber, A. P. (2012) “RNA regulons and the RNA-protein interaction network.,” Biomolecular concepts, 3(5), pp. 403–414.

Ito, T. et al. (2001) “A comprehensive two-hybrid analysis to explore the yeast protein interactome,” Proceedings of the National Academy of Sciences of the United States of America, 98(8), pp. 4569–4574.

Iwakawa, H.-O. and Tomari, Y. (2015) “The Functions of MicroRNAs: mRNA Decay and Translational Repression.,” Trends in cell biology, 25(11), pp. 651–665.

Iwasaki, S. and Ingolia, N. T. (2017) “The Growing Toolbox for Protein Synthesis Studies.,” Trends in biochemical sciences, 42(8), pp. 612–624.

Iyer, M. K. et al. (2015) “The landscape of long noncoding RNAs in the human transcriptome.,” Nature genetics, 47(3), pp. 199–208.

Jacob, F. and Monod, J. (1961) “Genetic regulatory mechanisms in the synthesis of proteins,” Journal of molecular biology. Elsevier, 3(3), pp. 318–356.

Jankowsky, E. and Harris, M. E. (2015) “Specificity and nonspecificity in RNA-protein interactions.,” Nature reviews. Molecular cell biology, 16(9), pp. 533–544.

Jeffery, C. J. (1999) “Moonlighting proteins.,” Trends in biochemical sciences, 24(1), pp. 8–11.

Jeffery, C. J. (2009) “Moonlighting proteins--an update.,” Molecular bioSystems, 5(4), pp. 345–350.

Jeffery, C. J. (2014) “An introduction to protein moonlighting.,” Biochemical Society transactions, 42(6),

pp. 1679–1683.

Jeffery, C. J. (2015) “Why study moonlighting proteins?,” Frontiers in genetics, 6, p. 211.

Jeffery, C. J. (2018) “Protein moonlighting: what is it, and why is it important?,” Philosophical transactions of the Royal Society of London. Series B, Biological sciences, 373(1738).

Jiang, Q. et al. (2015) “LncRNA2Function: a comprehensive resource for functional investigation of human lncRNAs based on RNA-seq data.,” BMC genomics, 16 Suppl 3, p. S2.

Jones, S. et al. (2001) “Protein-RNA interactions: a structural analysis,” Nucl. Acids Res., 29(4), pp. 943–954.

Junge, A. et al. (2017) “RAIN: RNA-protein Association and Interaction Networks.,” Database : the journal of biological databases and curation, 2017.

Kainulainen, V. and Korhonen, T. K. (2014) “Dancing to another tune-adhesive moonlighting proteins in bacteria.,” Biology, 3(1), pp. 178–204.

Kamburov, A. et al. (2013) “The ConsensusPathDB interaction database: 2013 update,” Nucleic Acids Research, 41(D1).

Kanehisa, M. et al. (2017) “KEGG: new perspectives on genomes, pathways, diseases and drugs.,” Nucleic acids research, 45(D1), pp. D353–D361.

Kartha, R. V. and Subramanian, S. (2014) “Competing endogenous RNAs (ceRNAs): new entrants to the intricacies of gene regulation.,” Frontiers in genetics, 5, p. 8.

Kato, M. et al. (2012) “Cell-free formation of RNA granules: low complexity sequence domains form dynamic fibers within hydrogels,” Cell. Elsevier, 149, pp. 753–767.

Keene, J. D. (2007) “RNA regulons: coordination of post-transcriptional events,” Nat Rev Genet, 8, pp. 533–43.

Khalil, A. M. et al. (2009) “Many human large intergenic noncoding RNAs associate with chromatin-modifying complexes and affect gene expression,” Proceedings of the National Academy of Sciences. National Acad Sciences, 106(28), pp. 11667–11672.

Khan, I. et al. (2012) “Evaluation of function predictions by PFP, ESG,and PSI-BLAST for moonlighting proteins.,” BMC proceedings, 6 Suppl 7, p. S5.

Khan, I. K., Bhuiyan, M. and Kihara, D. (2017) “DextMP: deep dive into text for predicting moonlighting proteins,” Bioinformatics, 33(14).

Khan, I. K. and Kihara, D. (2016) “Genome-scale prediction of moonlighting proteins using diverse protein association information,” Bioinformatics, 32(15).

Kirwan, M. and Dokal, I. (2009) “Dyskeratosis congenita, stem cells and telomeres.,” Biochimica et

biophysica acta, 1792(4), pp. 371–379.

Kishore, S. et al. (2011) “A quantitative analysis of CLIP methods for identifying binding sites of RNA-binding proteins,” Nat Methods, 8, pp. 559–61.

Koegl, M. and Uetz, P. (2007) “Improving yeast two-hybrid screening systems.,” Briefings in functional genomics & proteomics, 6(4), pp. 302–312.

Koh, Y. Y. et al. (2009) “A single C. elegans PUF protein binds RNA in multiple modes.,” RNA (New York, N.Y.), 15(6), pp. 1090–1099.

Konig, J. et al. (2011) “iCLIP--transcriptome-wide mapping of protein-RNA interactions with individual nucleotide resolution.,” Journal of visualized experiments : JoVE, (50).

Kopp, F. and Mendell, J. T. (2018) “Functional Classification and Experimental Dissection of Long Noncoding RNAs.,” Cell, 172(3), pp. 393–407.

Krakau, S., Richard, H. and Marsico, A. (2017) “PureCLIP: capturing target-specific protein-RNA interaction footprints from single-nucleotide CLIP-seq data.,” Genome biology, 18(1), p. 240.

Kretz, M. and Meister, G. (2014) “RNA binding of PRC2: promiscuous or well ordered?,” Molecular cell, 55(2), pp. 157–158.

Lambert, N. et al. (2014) “RNA Bind-n-Seq: quantitative assessment of the sequence and structural binding specificity of RNA binding proteins.,” Molecular cell, 54(5), pp. 887–900.

Laurent, G. S., Wahlestedt, C. and Kapranov, P. (2014) “The Landscape of long noncoding RNA classification,” Trends in Genetics, 31(5):239-51.

Lee, F. C. Y. and Ule, J. (2018) “Advances in CLIP Technologies for Studies of Protein-RNA Interactions.,” Molecular cell, 69(3), pp. 354–369.

Lee, S. et al. (2016) “Noncoding RNA NORAD Regulates Genomic Stability by Sequestering PUMILIO Proteins.,” Cell, 164(1–2), pp. 69–80.

Lee, S.-H. et al. (2018) “Widespread intronic polyadenylation inactivates tumour suppressor genes in leukaemia.,” Nature, 561(7721), pp. 127–131.

Leucci, E. et al. (2016) “Melanoma addiction to the long non-coding RNA SAMMSON.,” Nature, 531(7595), pp. 518–522.

Li, Y. et al. (2015) “LncRNA ontology: inferring lncRNA functions based on chromatin states and expression patterns.,” Oncotarget, 6(37), pp. 39793–39805.

Li, Y., Syed, J. and Sugiyama, H. (2016) “RNA-DNA Triplex Formation by Long Noncoding RNAs.,” Cell chemical biology, 23(11), pp. 1325–1333.

Lianoglou, S. et al. (2013) “Ubiquitously transcribed genes use alternative polyadenylation to achieve

tissue-specific expression.,” Genes & development, 27(21), pp. 2380–2396.

Liu, F. et al. (2016) “Cotranslational association of mRNA encoding subunits of heteromeric ion channels.,” Proceedings of the National Academy of Sciences of the United States of America, 113(17), pp. 4859–4864.

Liu, S. J. et al. (2017) “CRISPRi-based genome-scale identification of functional long noncoding RNA loci in human cells.,” Science (New York, N.Y.), 355(6320).

Long, Y. et al. (2017) “How do lncRNAs regulate transcription?,” Science advances, 3(9), p. eaao2110.

Lu, Z. and Hunter, T. (2018) “Metabolic Kinases Moonlighting as Protein Kinases.,” Trends in biochemical sciences, 43(4), pp. 301–310.

Luo, J. et al. (2017) “RPI-Bind: a structure-based method for accurate identification of RNA-protein binding sites.,” Scientific reports, 7(1), p. 614.

Ma, X. et al. (2017) “Multiple network algorithm for epigenetic modules via the integration of genome-wide DNA methylation and gene expression data.,” BMC bioinformatics, 18(1), p. 72.

Ma, X. and Gao, L. (2012) “Biological network analysis: insights into structure and functions.,” Briefings in functional genomics, 11(6), pp. 434–442.

Mackowiak, S. D. et al. (2015) “Extensive identification and analysis of conserved small ORFs in animals.,” Genome biology, 16, p. 179.

Marchese, D. et al. (2017) “Discovering the 3’ UTR-mediated regulation of alpha-synuclein.,” Nucleic acids research, 45(22), pp. 12888–12903.

Marchese, F. P., Raimondi, I. and Huarte, M. (2017) “The multidimensional mechanisms of long noncoding RNA function.,” Genome biology, 18(1), p. 206.

Margineanu, A. et al. (2016) “Screening for protein-protein interactions using Förster resonance energy transfer (FRET) and fluorescence lifetime imaging microscopy (FLIM).,” Scientific reports, 6, p. 28186.

Martin, F. (2012) “Fifteen years of the yeast three-hybrid system: RNA-protein interactions under investigation.,” Methods (San Diego, Calif.), 58(4), pp. 367–375.

Matoulkova, E. et al. (2012) “The role of the 3’ untranslated region in post-transcriptional regulation of protein expression in mammalian cells.,” RNA Biology. Taylor & Francis, 9(5), pp. 563–576.

Mattick, J. S., Taft, R. J. and Faulkner, G. J. (2010) “A global view of genomic information--moving beyond the gene and the master regulator.,” Trends in genetics : TIG, 26(1), pp. 21–28.

Maxwell, C. A., McCarthy, J. and Turley, E. (2008) “Cell-surface and mitotic-spindle RHAMM: moonlighting or dual oncogenic functions?,” Journal of cell science, 121(Pt 7), pp. 925–932.

Mayr, C. (2016) “Evolution and Biological Roles of Alternative 3’UTRs.,” Trends in cell biology, 26(3),

pp. 227–237.

Mayr, C. (2017) “Regulation by 3’-Untranslated Regions.,” Annual review of genetics, 51, pp. 171–194.

Mayr, C. (2018a) “Protein complexes assemble as they are being made.,” Nature, 561(7722), pp. 186–187.

Mayr, C. (2018b) “What Are 3’ UTRs Doing?,” Cold Spring Harbor perspectives in biology.

McHugh, C. A. et al. (2014) “Methods for comprehensive experimental identification of RNA-protein interactions,” Genome Biology, 15, p. 203.

McHugh, C. A. et al. (2015) “The Xist lncRNA interacts directly with SHARP to silence transcription through HDAC3.,” Nature, 521(7551), pp. 232–236.

Meyer-Siegler, K. et al. (1991) “A human nuclear uracil DNA glycosylase is the 37-kDa subunit of glyceraldehyde-3-phosphate dehydrogenase.,” Proceedings of the National Academy of Sciences of the United States of America, 88(19), pp. 8460–8464.

Mignone, F. and Pesole, G. (2018) “mRNA Untranslated Regions (UTRs),” in eLS. John Wiley & Sons, Ltd, pp. 1–6.

Milek, M., Wyler, E. and Landthaler, M. (2012) “Transcriptome-wide analysis of protein-RNA interactions using high-throughput sequencing.,” Seminars in cell & developmental biology, 23(2), pp. 206–212.

Miller, K. E. et al. (2015) “Bimolecular Fluorescence Complementation (BiFC) Analysis: Advances and Recent Applications for Genome-Wide Interaction Studies.,” Journal of molecular biology, 427(11), pp. 2039–2055.

Mitchell, S. F. et al. (2013) “Global analysis of yeast mRNPs.,” Nature structural & molecular biology, 20(1), pp. 127–133.

Mitchell, S. F. and Parker, R. (2014) “Principles and properties of eukaryotic mRNPs.,” Molecular cell, 54(4), pp. 547–558.

Morell, M., Ventura, S. and Avilés, F. X. (2009) “Protein complementation assays: approaches for the in vivo analysis of protein interactions.,” FEBS letters, 583(11), pp. 1684–1691.

Movassat, M. et al. (2016) “Coupling between alternative polyadenylation and alternative splicing is limited to terminal introns,” RNA Biology. Taylor & Francis, 13(7), pp. 646–655.

Munschauer, M. et al. (2018) “The NORAD lncRNA assembles a topoisomerase complex critical for genome stability.,” Nature, 561(7721), pp. 132–136.

Nakagawa, S. et al. (2014) “The lncRNA Neat1 is required for corpus luteum formation and the establishment of pregnancy in a subpopulation of mice.,” Development (Cambridge, England), 141(23),

pp. 4618–4627.

Nelson, B. R. et al. (2016) “A peptide encoded by a transcript annotated as long noncoding RNA enhances SERCA activity in muscle.,” Science (New York, N.Y.), 351(6270), pp. 271–275.

Nielsen, F. C., Hansen, H. T. and Christiansen, J. (2016) “RNA assemblages orchestrate complex cellular processes.,” BioEssays : news and reviews in molecular, cellular and developmental biology, 38(7), pp. 674–681.

Nishimoto, Y. et al. (2013) “The long non-coding RNA nuclear-enriched abundant transcript 1_2 induces paraspeckle formation in the motor neuron during the early phase of amyotrophic lateral sclerosis.,” Molecular brain, 6, p. 31.

Nooren, I. M. A. and Thornton, J. M. (2003) “Diversity of protein-protein interactions.,” The EMBO journal, 22(14), pp. 3486–3492.

Olexiouk, V., Van Criekinge, W. and Menschaert, G. (2018) “An update on sORFs.org: a repository of small ORFs identified by ribosome profiling.,” Nucleic acids research, 46(D1), pp. D497–D502.

Omenn, G. S. (2014) “The strategy, organization, and progress of the HUPO Human Proteome Project.,” Journal of proteomics, 100, pp. 3–7.

Orchard, S. et al. (2012) “Protein interaction data curation: the International Molecular Exchange (IMEx) consortium.,” Nature methods, 9(4), pp. 345–350.

Orchard, S. et al. (2014) “The MIntAct project--IntAct as a common curation platform for 11 molecular interaction databases.,” Nucleic acids research, 42(Database issue), pp. D358–D363.

Ostrovsky de Spicer, P. and Maloy, S. (1993) “PutA protein, a membrane-associated flavin dehydrogenase, acts as a redox-dependent transcriptional regulator.,” Proceedings of the National Academy of Sciences of the United States of America, 90(9), pp. 4295–4298.

Park, C. et al. (2014) “lncRNAtor: a comprehensive resource for functional investigation of long non-coding RNAs.,” Bioinformatics (Oxford, England), 30(17), pp. 2480–2485.

Patel, T. R. et al. (2017) “Structural studies of RNA-protein complexes: A hybrid approach involving hydrodynamics, scattering, and computational methods.,” Methods (San Diego, Calif.), 118–119, pp. 146–162.

Pavlopoulos, G. A. et al. (2011) “Using graph theory to analyze biological networks.,” BioData mining, 4, p. 10.

Pawson, T. and Nash, P. (2003) “Assembly of cell regulatory systems through protein interaction domains.,” Science (New York, N.Y.), 300(5618), pp. 445–452.

van de Peppel, J. and Holstege, F. C. P. (2005) “Multifunctional genes.,” Molecular systems biology, 1, p.

2005.0003.

Perkins, J. R. et al. (2010) “Transient protein-protein interactions: structural, functional, and network properties.,” Structure (London, England : 1993), 18(10), pp. 1233–1243.

Perucho, M., Salas, J. and Salas, M. L. (1977) “Identification of the mammalian DNA-binding protein P8 as glyceraldehyde-3-phosphate dehydrogenase.,” European journal of biochemistry, 81(3), pp. 557–562.

Piatigorsky, J. and Wistow, G. J. (1989) “Enzyme/crystallins: gene sharing as an evolutionary strategy.,” Cell, 57(2), pp. 197–199.

Protter, D. S. W. et al. (2018) “Intrinsically Disordered Regions Can Contribute Promiscuous Interactions to RNP Granule Assembly.,” Cell reports, 22(6), pp. 1401–1412.

Proudfoot, N. J. (2011) “Ending the message: poly (A) signals then and now,” Genes & development, 25(17), pp. 1770–1782.

Quek, X. C. et al. (2015) “lncRNAdb v2.0: expanding the reference database for functional long noncoding RNAs.,” Nucleic acids research, 43(Database issue), pp. D168–D173.

Quénet, D. and Dalal, Y. (2014) “A long non-coding RNA is required for targeting centromeric protein A to the human centromere.,” eLife, 3, p. e03254.

Quinn, J. J. and Chang, H. Y. (2016) “Unique features of long non-coding RNA biogenesis and function.,” Nature reviews. Genetics, 17(1), pp. 47–62.

Ransohoff, J. D., Wei, Y. and Khavari, P. A. (2018) “The functions and unique features of long intergenic non-coding RNA.,” Nature reviews. Molecular cell biology, 19(3), pp. 143–157.

Rao, V. S. et al. (2014) “Protein-protein interaction detection: methods and analysis.,” International journal of proteomics, 2014, p. 147648.

Ray, D. et al. (2013) “A compendium of RNA-binding motifs for decoding gene regulation.,” Nature, 499(7457), pp. 172–177.

Ray, D. et al. (2017) “RNAcompete methodology and application to determine sequence preferences of unconventional RNA-binding proteins.,” Methods (San Diego, Calif.), 118–119, pp. 3–15.

Ribeiro, D. M. et al. (2018) “Protein complex scaffolding predicted as a prevalent function of long non-coding RNAs.,” Nucleic acids research, 46(2), pp. 917–928.

Rolland, T. et al. (2014) “A proteome-scale map of the human interactome network.,” Cell, 159(5), pp. 1212–1226.

Roux, K. J., Kim, D. I. and Burke, B. (2013) “BioID: a screen for protein-protein interactions.,” Current protocols in protein science, 74, p. Unit 19.23.

Ruepp, A. et al. (2009) “CORUM: The comprehensive resource of mammalian protein complexes-2009,”

Nucleic Acids Research, 38, pp. 497–501.

Ruiz-Orera, J. et al. (2014) “Long non-coding RNAs as a source of new peptides.,” eLife, 3, p. e03523.

Sachdeva, G. et al. (2014) “In vivo co-localization of enzymes on RNA scaffolds increases metabolic production in a geometrically dependent manner.,” Nucleic acids research, 42(14), pp. 9493–9503.

Salehi, S. et al. (2017) “State of the art technologies to explore long non-coding RNAs in cancer.,” Journal of cellular and molecular medicine, 21(12), pp. 3120–3140.

Sanchez, C. et al. (1999) “Grasping at molecular interactions and genetic networks in Drosophila melanogaster using FlyNets, an Internet database.,” Nucleic acids research, 27(1), pp. 89–94.

Savell, K. E. et al. (2016) “Extra-coding RNAs regulate neuronal DNA methylation dynamics.,” Nature communications, 7, p. 12091.

Scherrer, T. et al. (2011) “Defining potentially conserved RNA regulons of homologous zinc-finger RNA-binding proteins.,” Genome biology, 12(1), p. R3.

Schmidt, E. K. et al. (2009) “SUnSET, a nonradioactive method to monitor protein synthesis.,” Nature methods, 6(4), pp. 275–277.

Sharan, R., Ulitsky, I. and Shamir, R. (2007) “Network-based prediction of protein function.,” Mol Syst Biol, 3, p. 88.

Shaw, A. S. and Filbert, E. L. (2009) “Scaffold proteins and immune-cell signalling.,” Nature reviews. Immunology, 9(1), pp. 47–56.

Sheik Mohamed, J. et al. (2010) “Conserved long noncoding RNAs transcriptionally regulated by Oct4 and Nanog modulate pluripotency in mouse embryonic stem cells.,” RNA (New York, N.Y.), 16(2), pp. 324–337.

Shiber, A. et al. (2018) “Cotranslational assembly of protein complexes in eukaryotes revealed by ribosome profiling.,” Nature, 561(7722), pp. 268–272.

Singh, I. et al. (2018) “Widespread intronic polyadenylation diversifies immune cell transcriptomes.,” Nature communications, 9(1), p. 1716.

Sirover, M. A. (2011) “On the functional diversity of glyceraldehyde-3-phosphate dehydrogenase: biochemical mechanisms and regulatory control.,” Biochimica et biophysica acta, 1810(8), pp. 741–751.

Skalska, L. et al. (2017) “Regulatory feedback from nascent RNA to chromatin and transcription.,” Nature reviews. Molecular cell biology, 18(5), pp. 331–337.

Snider, J. et al. (2015) “Fundamentals of protein interaction network mapping.,” Molecular systems biology, 11(12), p. 848.

Spitale, R. C., Tsai, M.-C. and Chang, H. Y. (2011) “RNA templating the epigenome: Long noncoding

RNAs as molecular scaffolds," Epigenetics. Taylor & Francis, 6(5), pp. 539–543.

Stark, A. et al. (2005) "Animal MicroRNAs confer robustness to gene expression and have a significant impact on 3'UTR evolution," Cell. Cell Press, 123(6), pp. 1133–1146.

Stefl, R., Skrisovska, L. and Allain, F. H.-T. (2005) "RNA sequence- and shape-dependent recognition by proteins in the ribonucleoprotein particle.," EMBO reports, 6(1), pp. 33–38.

Stelzl, U. et al. (2005) "A human protein-protein interaction network: a resource for annotating the proteome.," Cell, 122, pp. 957–968.

Stynen, B. et al. (2012) "Diversity in genetic in vivo methods for protein-protein interaction studies: from the yeast two-hybrid system to the mammalian split-luciferase system.," Microbiology and molecular biology reviews : MMBR, 76(2), pp. 331–382.

Su, M. et al. (2018) "LncRNAs in DNA damage response and repair in cancer cells.," Acta biochimica et biophysica Sinica, 50(5), pp. 433–439.

Sun, Q., Hao, Q. and Prasanth, K. V. (2018) "Nuclear Long Noncoding RNAs: Key Regulators of Gene Expression.," Trends in genetics : TIG, 34(2), pp. 142–157.

Suresh, V. et al. (2015) "RPI-Pred: predicting ncRNA-protein interaction using sequence and structural information.," Nucleic acids research, 43(3), pp. 1370–1379.

Szostak, E. and Gebauer, F. (2013) "Translational control by 3'-UTR-binding proteins.," Briefings in functional genomics, 12(1), pp. 58–65.

Talbert, P. B. and Henikoff, S. (2018) "Transcribing Centromeres: Noncoding RNAs and Kinetochore Assembly.," Trends in genetics : TIG, 34(8), pp. 587–599.

Taliaferro, J. M. et al. (2016) "RNA Sequence Context Effects Measured In Vitro Predict In Vivo Protein Binding and Regulation.," Molecular cell, 64(2), pp. 294–306.

Tavernier, J. et al. (2002) "MAPPIT: a cytokine receptor-based two-hybrid method in mammalian cells.," Clinical and experimental allergy : journal of the British Society for Allergy and Clinical Immunology, 32(10), pp. 1397–1404.

The Gene Ontology Consortium (2017) "Expansion of the Gene Ontology knowledgebase and resources.," Nucleic acids research, 45(D1), pp. D331–D338.

Tian, B. and Manley, J. L. (2017) "Alternative polyadenylation of mRNA precursors.," Nature reviews. Molecular cell biology, 18(1), pp. 18–30.

del-Toro, N. et al. (2013) "A new reference implementation of the PSICQUIC web service.," Nucleic acids research, 41(Web Server issue), pp. W601–W606.

Torres, M. et al. (2016) "Circadian RNA expression elicited by 3'-UTR IRAlu-paraspeckle associated

elements.,” eLife, 5.

Torres, M. et al. (2017) “Paraspeckles as rhythmic nuclear mRNA anchorages responsible for circadian gene expression,” Nucleus. Taylor & Francis, 8(3), pp. 249–254.

Torres, M. et al. (2018) “Circadian processes in the RNA life cycle.,” Wiley interdisciplinary reviews. RNA, 9(3), p. e1467.

Tristan, C. et al. (2011) “The diverse functions of GAPDH: views from different subcellular compartments.,” Cellular signalling, 23(2), pp. 317–323.

Tsai, M.-C. et al. (2010) “Long noncoding RNA as modular scaffold of histone modification complexes.,” Science (New York, N.Y.), 329(5992), pp. 689–693.

Tsai, R. L. and Green, H. (1973) “Studies on a mammalian cell protein (P8) with affinity for DNA in vitro.,” Journal of molecular biology, 73(3), pp. 307–316.

Tuerk, C. and Gold, L. (1990) “Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase.,” Science (New York, N.Y.), 249(4968), pp. 505–510.

Uetz, P. et al. (2000) “A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae,” Nature, 403(6770), pp. 623–627.

Uhlén, M. et al. (2015) “Tissue-based map of the human proteome,” Science. American Association for the Advancement of Science, 347(6220), p. 1260419.

Ule, J. et al. (2003) “CLIP identifies Nova-regulated RNA networks in the brain,” Science (New York, N.Y.), 302, pp. 1212–1215.

Ulitsky, I. et al. (2012) “Extensive alternative polyadenylation during zebrafish development.,” Genome research, 22(10), pp. 2054–2066.

Ulitsky, I. (2016) “Evolution to the rescue: using comparative genomics to understand long non-coding RNAs.,” Nature reviews. Genetics, 17(10), pp. 601–614.

Uniacke, J. et al. (2012) “An oxygen-regulated switch in the protein synthesis machinery.,” Nature, 486(7401), pp. 126–129.

Urech, D. M., Lichtlen, P. and Barberis, A. (2003) “Cell growth selection system to detect extracellular and transmembrane protein interactions.,” Biochimica et biophysica acta, 1622(2), pp. 117–127.

Uszczynska-Ratajczak, B. et al. (2018) “Towards a complete map of the human long non-coding RNA transcriptome.,” Nature reviews. Genetics, 19(9), pp. 535–548.

Van Essen, D. C. et al. (2013) “The WU-Minn Human Connectome Project: an overview.,” NeuroImage, 80, pp. 62–79.

Van Nostrand, E. L. et al. (2016) “Robust transcriptome-wide discovery of RNA-binding protein binding

sites with enhanced CLIP (eCLIP).,” Nature methods, 13(6), pp. 508–514.

Van Roey, K. et al. (2013) “The switches.ELM resource: a compendium of conditional regulatory interaction interfaces.,” Science signaling, 6(269), p. rs7.

Van Treeck, B. et al. (2018) “RNA self-assembly contributes to stress granule formation and defining the stress granule transcriptome.,” Proceedings of the National Academy of Sciences of the United States of America, 115(11), pp. 2734–2739.

Van Treeck, B. and Parker, R. (2018) “Emerging Roles for Intermolecular RNA-RNA Interactions in RNP Assemblies.,” Cell, 174(4), pp. 791–802.

Venkatesan, K. et al. (2009) “An empirical framework for binary interactome mapping,” 6(1), pp. 83–90.

Verheggen, K. et al. (2017) “Noncoding after All: Biases in Proteomics Data Do Not Explain Observed Absence of lncRNA Translation Products.,” Journal of proteome research, 16(7), pp. 2508–2515.

Vidal, M. (2001) “A biological atlas of functional maps.,” Cell, 104(3), pp. 333–339.

Vidal, M., Cusick, M. E. and Barabasi, A.-L. (2011) “Interactome networks and human disease,” Cell. Elsevier, 144(6), pp. 986–998.

Volz, K. (2008) “The functional duality of iron regulatory protein 1.,” Current opinion in structural biology, 18(1), pp. 106–111.

Wang, W. and Jeffery, C. J. (2016) “An analysis of surface proteomics results reveals novel candidates for intracellular/surface moonlighting proteins in bacteria.,” Molecular bioSystems, 12(5), pp. 1420–1431.

Wang, Z. et al. (2009) “CLIP: construction of cDNA libraries for high-throughput sequencing from RNAs cross-linked to proteins in vivo.,” Methods (San Diego, Calif.), 48(3), pp. 287–293.

Wheeler, E. C., Van Nostrand, E. L. and Yeo, G. W. (2018) “Advances and challenges in the detection of transcriptome-wide protein-RNA interactions.,” Wiley interdisciplinary reviews. RNA, 9(1).

Wilusz, J. E., Sunwoo, H. and Spector, D. L. (2009) “Long noncoding RNAs: functional surprises from the RNA world.,” Genes & development, 23(13), pp. 1494–1504.

Wistow, G. J. and Piatigorsky, J. (1988) “Lens crystallins: the evolution and expression of proteins for a highly specialized tissue.,” Annual review of biochemistry, 57, pp. 479–504.

Wistow, G. and Piatigorsky, J. (1987) “Recruitment of enzymes as lens structural proteins.,” Science (New York, N.Y.), 236(4808), pp. 1554–1556.

Xia, K. et al. (2008) “Impacts of protein-protein interaction domains on organism and network complexity.,” Genome research, 18(9), pp. 1500–1508.

Yi, Y., Zhao, Y., Huang, Y., et al. (2017) “A Brief Review of RNA-Protein Interaction Database

Resources.,” Non-coding RNA, 3(1).

Yi, Y., Zhao, Y., Li, C., et al. (2017) “RAID v2.0: an updated resource of RNA-associated interactions across organisms,” Nucleic Acids Research, 45(D1).

Yoon, J. H., Ryu, J. and Baek, S. J. (2018) “Moonlighting Activity of Secreted Inflammation-Regulatory Proteins.,” Yonsei medical journal, 59(4), pp. 463–469.

Yuan, M. et al. (2017) “Long noncoding RNA profiling revealed differentially expressed lncRNAs associated with disease activity in PBMCs from patients with rheumatoid arthritis.,” PloS one, 12(11), p. e0186795.

Zanzoni, A., Chapple, C. E. and Brun, C. (2015) “Relationships between predicted moonlighting proteins, human diseases, and comorbidities from a network perspective.,” Frontiers in physiology, 6, p. 171.

Zarnegar, B. J. et al. (2016) “irCLIP platform for efficient characterization of protein-RNA interactions.,” Nature methods, 13(6), pp. 489–492.

Zhang, Q., Kim, N.-K. and Feigon, J. (2011) “Architecture of human telomerase RNA.,” Proceedings of the National Academy of Sciences of the United States of America, 108(51), pp. 20325–20332.

Zhang, X., Hamblin, M. H. and Yin, K.-J. (2017) “The long noncoding RNA Malat1: Its physiological and pathophysiological functions,” RNA Biology. Taylor & Francis, 14(12), pp. 1705–1714.

Zhang, X.-F. et al. (2015) “Identifying binary protein-protein interactions from affinity purification mass spectrometry data.,” BMC genomics, 16, p. 745.

Zhang, Y. et al. (2016) “Long noncoding RNA LINP1 regulates repair of DNA double-strand breaks in triple-negative breast cancer.,” Nature structural & molecular biology, 23(6), pp. 522–530.

Zhao, J. et al. (2010) “Genome-wide identification of polycomb-associated RNAs by RIP-seq.,” Molecular cell, 40(6), pp. 939–953.

Abbreviations

- 3’UTR: three prime untranslated region
- 5’UTR: five prime untranslated region
- AP/MS: affinity purification coupled to mass spectrometry
- APA: alternative polyadenylation
- CLIP: cross-linking and immunoprecipitation
- DNA: deoxyribonucleic acid
- EMF: extreme multifunctional
- lncRNA: long non-coding RNA
- miRNA: microRNA
- mRNA: messenger ribonucleic acid
- MS: mass spectrometry
- ncRNA: non-coding RNA
- ORF: open reading frame
- P-bodies: processing bodies
- PA site: polyadenylation site
- PAS: polyadenylation signal
- poly(A): polyadenylated / polyadenylation
- PPI: protein-protein interaction
- PTM: post-translational modification
- RBD: RNA-binding domain
- RBP: RNA-binding protein
- RIP: RNA immunoprecipitation
- RNA: ribonucleic acid
- RNP: ribonucleoprotein
- SLiM: short linear motif
- UDPL: 3′UTR-dependent protein localisation
- UTR : untranslated region
- Y2H: yeast two-hybrid

Appendices

- **Appendix I**: Supplementary material of the article entitled ‘Protein complex scaffolding predicted as a prevalent function of long non-coding RNAs’
- **Appendix II**: Supplementary material of the article entitled ‘MoonDB 2.0: an updated database of extreme multifunctional and moonlighting proteins‘
- **Appendix III**: Supplementary material of the article entitled ‘Prediction of 3′UTR-protein complex assembly reveals a role in the regulation of protein multifunctionality’
- **Appendix IV**: Supplementary material of the article entitled ‘Predicted protein-RNA interactions reveal distinct post-transcriptional regulatory patterns’
- **Appendix V:** Scientific contributions outside the scope of this thesis

Appendix I: Article supplementary material

SUPPLEMENTARY MATERIAL

Protein complex scaffolding predicted as a prevalent function of long non-coding RNAs

Diogo M. Ribeiro, Andreas Zanzoni, Andrea Cipriano, Riccardo Delli Ponti, Lionel Spinelli, Monica Ballarino, Irene Bozzoni, Gian Gaetano Tartaglia, Christine Brun

Contents:

SUPPLEMENTARY FIGURES S1-S6

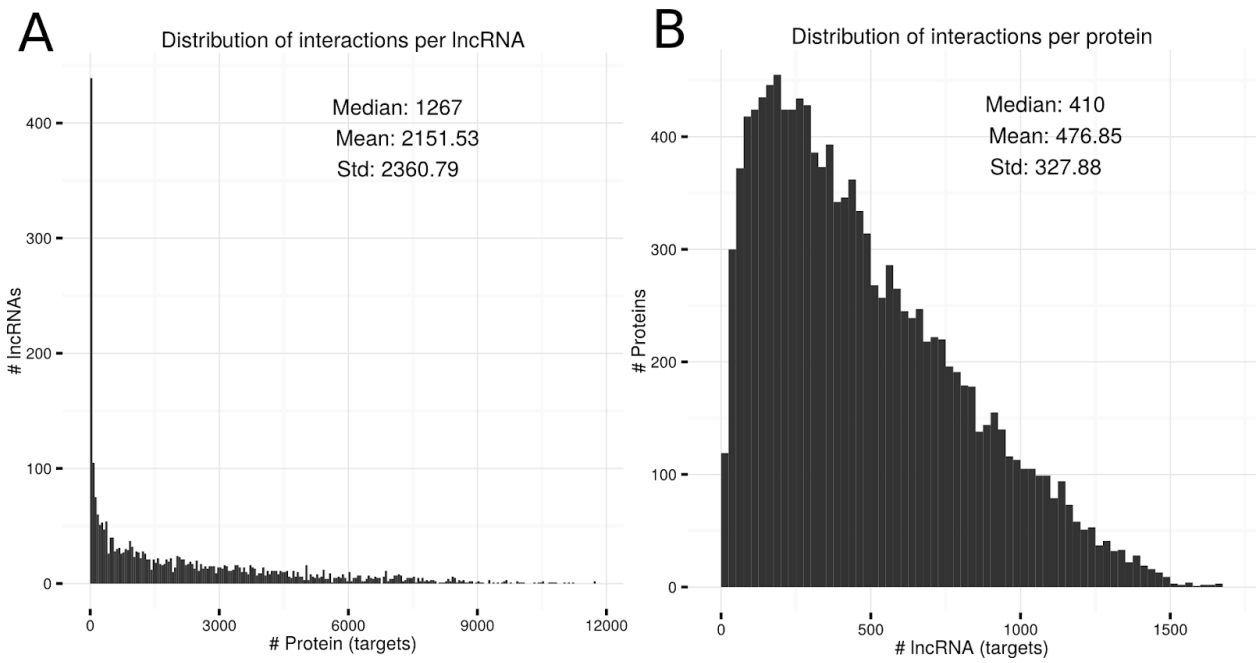


Figure S1. Distribution of 6 million protein-lncRNA interactions between 2799 lncRNAs and 12629 proteins. (A) distribution of the numbers of interacting proteins per lncRNA. (B) distribution of the numbers of interacting lncRNAs per protein.

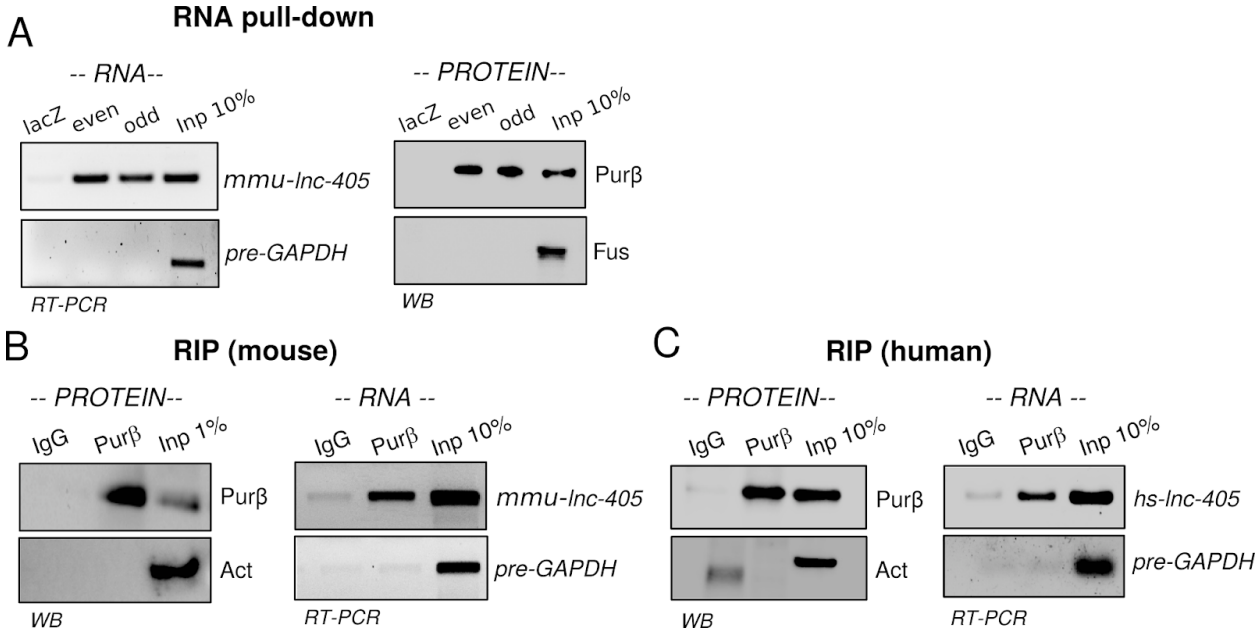


Figure S2. Validation of *lnc-405* interaction with Purβ in myotubes via RIP assays. (A) *lnc-405* RNA pull-down from nuclear extracts of differentiated myotubes RT-PCR quantification of *lnc-405* (left) and western blot analysis of Purβ interactors (right) in lacZ, odd, even and input (10%) samples are shown (B) Mouse RNA immunoprecipitation (RIP) of Purβ from nuclear extracts of differentiated myotubes. Western blot (WB) analysis of Purβ (left) and RT-PCR quantification of lnc-405 recovery (right) are shown. (C) Human RNA immunoprecipitation (RIP) of Purβ from total extracts of 5-days differentiated primary cells. Western blot (left) and qRT-PCR (right) show the recovery of Purβ and *lnc-405* from the indicated samples. Pre-GAPDH RNA, Actin protein (Act) and Fus protein serve as negative controls.

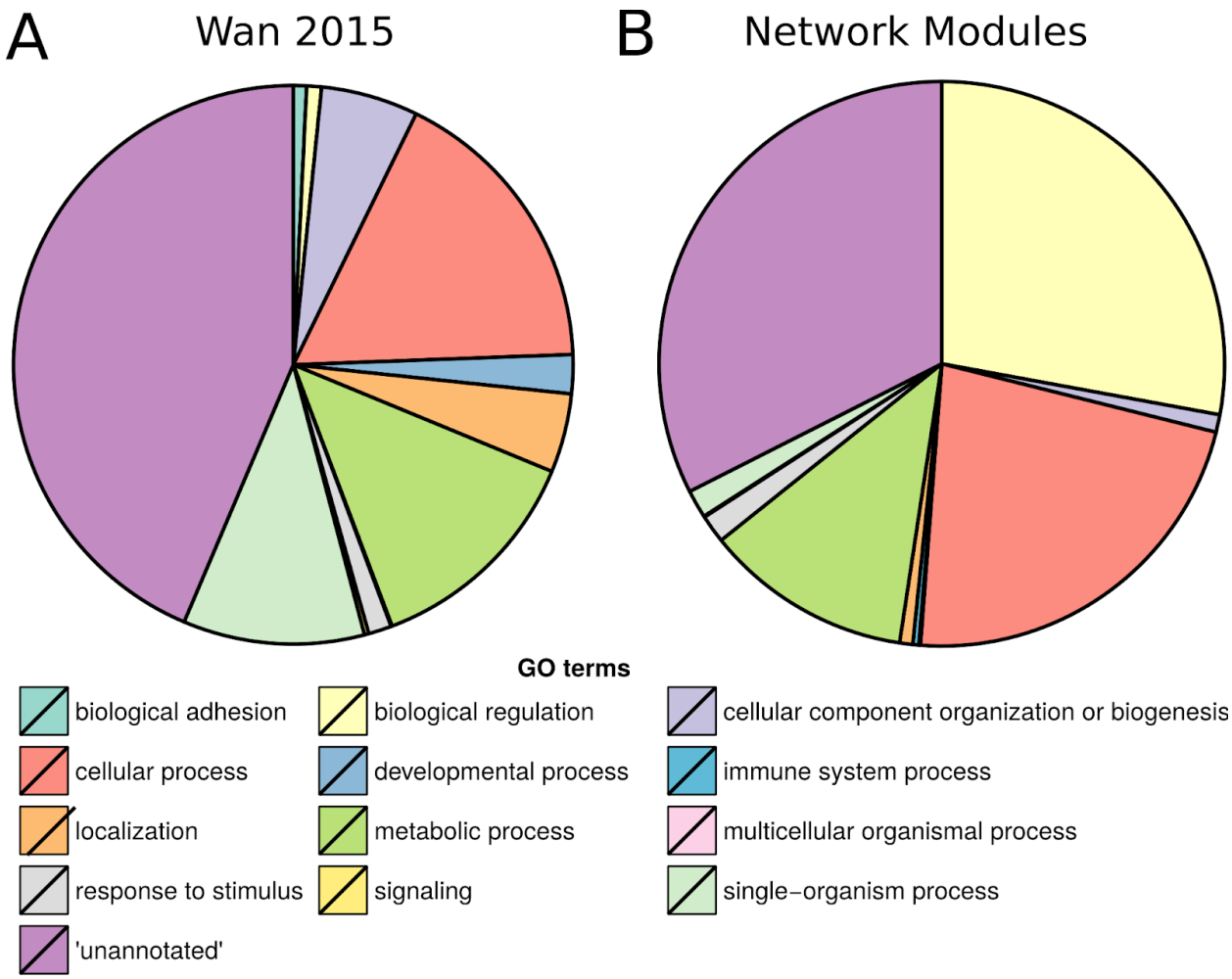


Figure S3. Proportion of gene ontology biological processes for Wan 2015 scaffolded complexes and network modules. For each gene ontology (GO) term, all parents terms with depth level 1 (broad GO terms) were obtained and displayed. The proportion measure is weighted by the number of annotations in each protein group, so that the proportion represents the total number of protein groups in the dataset. (A) Proportion of GO biological processes for Wan 2015. GO term annotations for 525 complexes interacting with at least one lncRNA were obtained from [1]. (B) Proportion of GO biological processes for network modules. GO term annotations for 579 modules interacting with at least one lncRNA were obtained from [2].

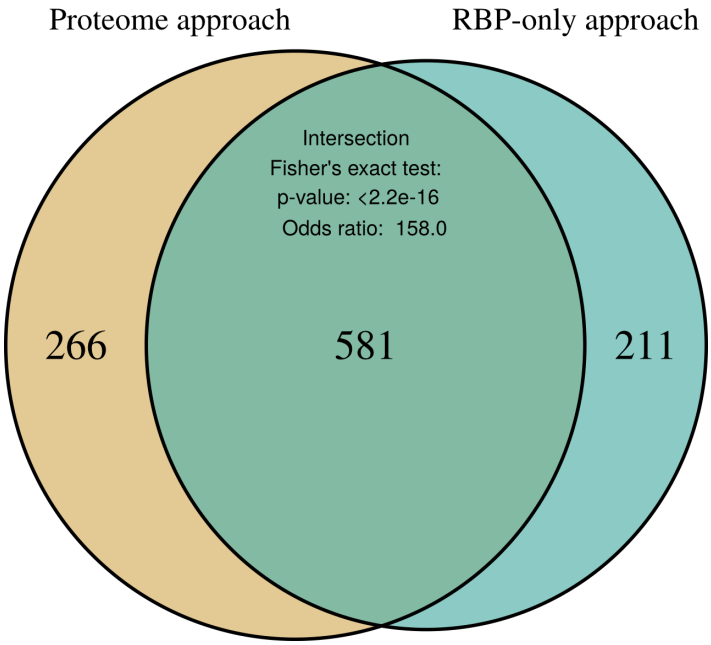


Figure S4. Comparison of proteome-wide and RBP-only approaches. Overlap between the lncRNA scaffolding candidates identified by the proteome-wide interaction dataset and the RBP-only interaction dataset. Fisher exact test background included all lncRNAs (15230 transcripts) analysed in this study.

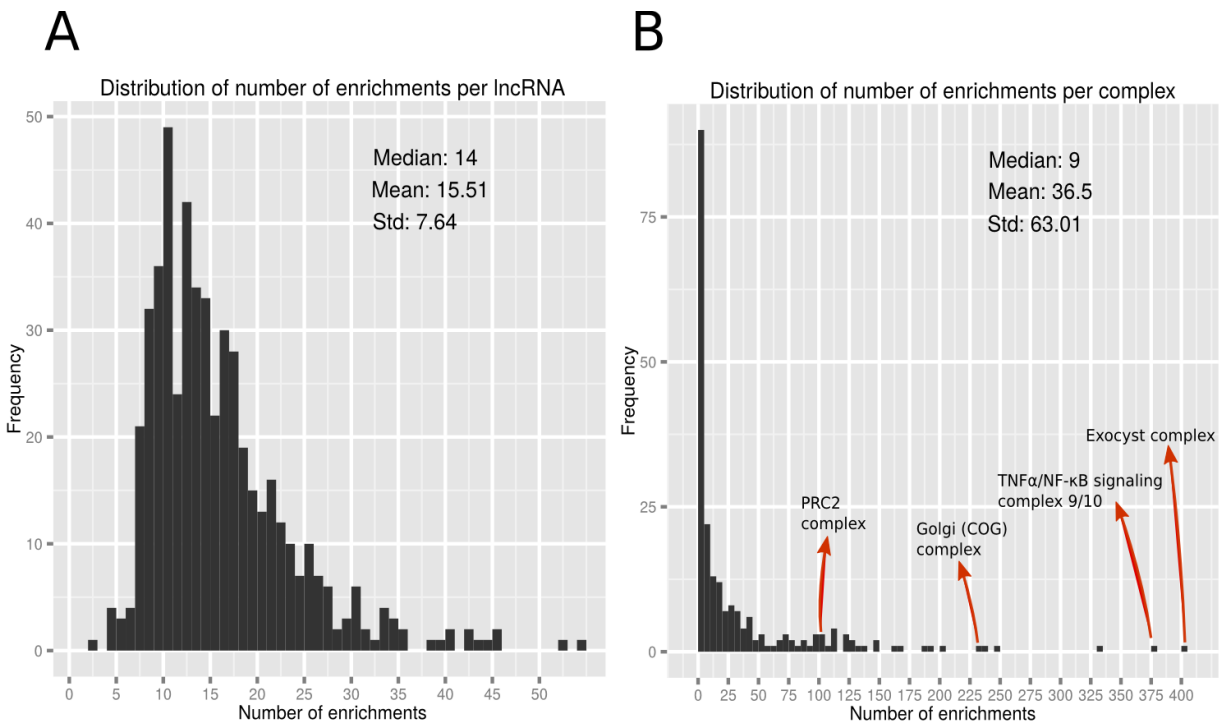


Figure S5. Distribution of the numbers of enrichments for the non-redundant CORUM dataset analysis. (A) distribution of the numbers of enrichments per lncRNA. (B) distribution of the numbers of enrichments per protein complex.

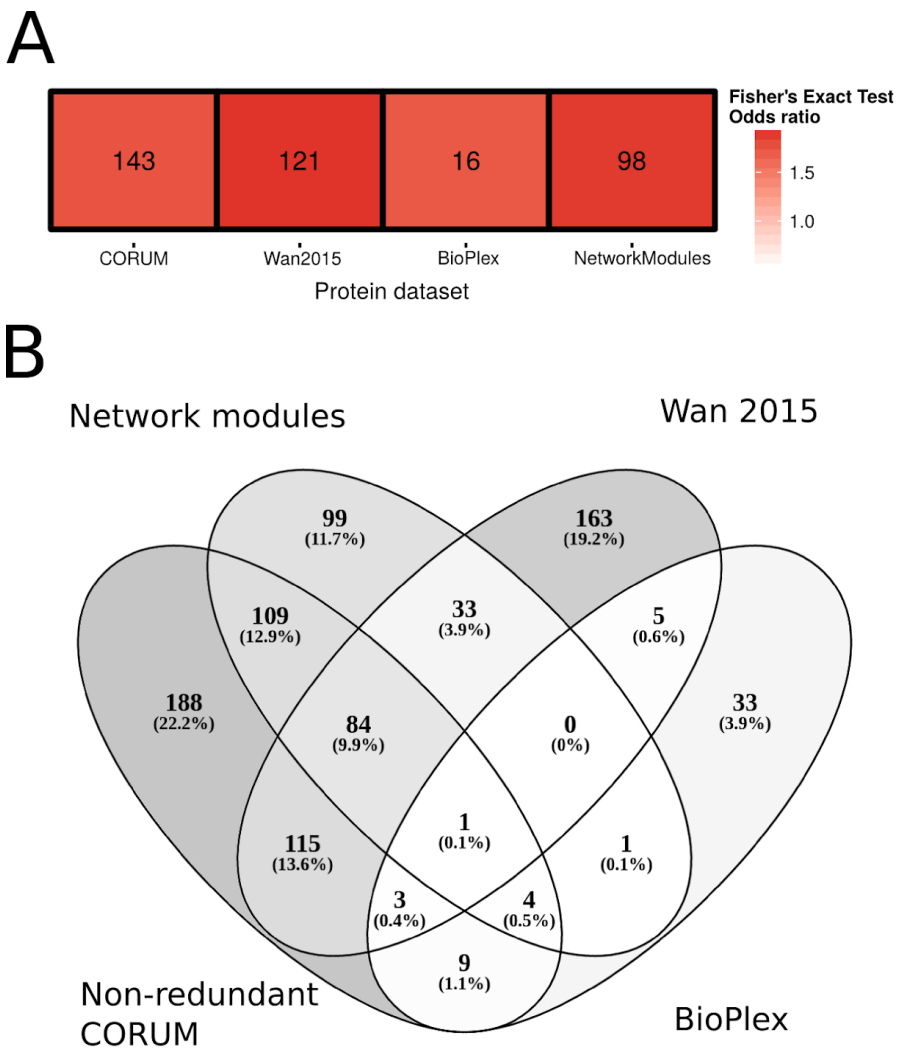


Figure S6. Comparison of lncRNA scaffolding candidates independently identified with each dataset of protein groups. (A) Summary of one-way Fisher's exact test between a collection of functional or conserved lncRNA genes (2013 genes) and scaffolding gene candidates detected with different protein complex or module datasets. Fisher's exact test background included all genes (12233 lncRNA genes, 15230 transcripts) analysed in this study. All p-values are significant (p-value < 0.05, except for BioPlex where p-value = 0.07). (B) Overlap of lncRNAs identified with each dataset of protein groups, produced with Venny 2.1 [3].

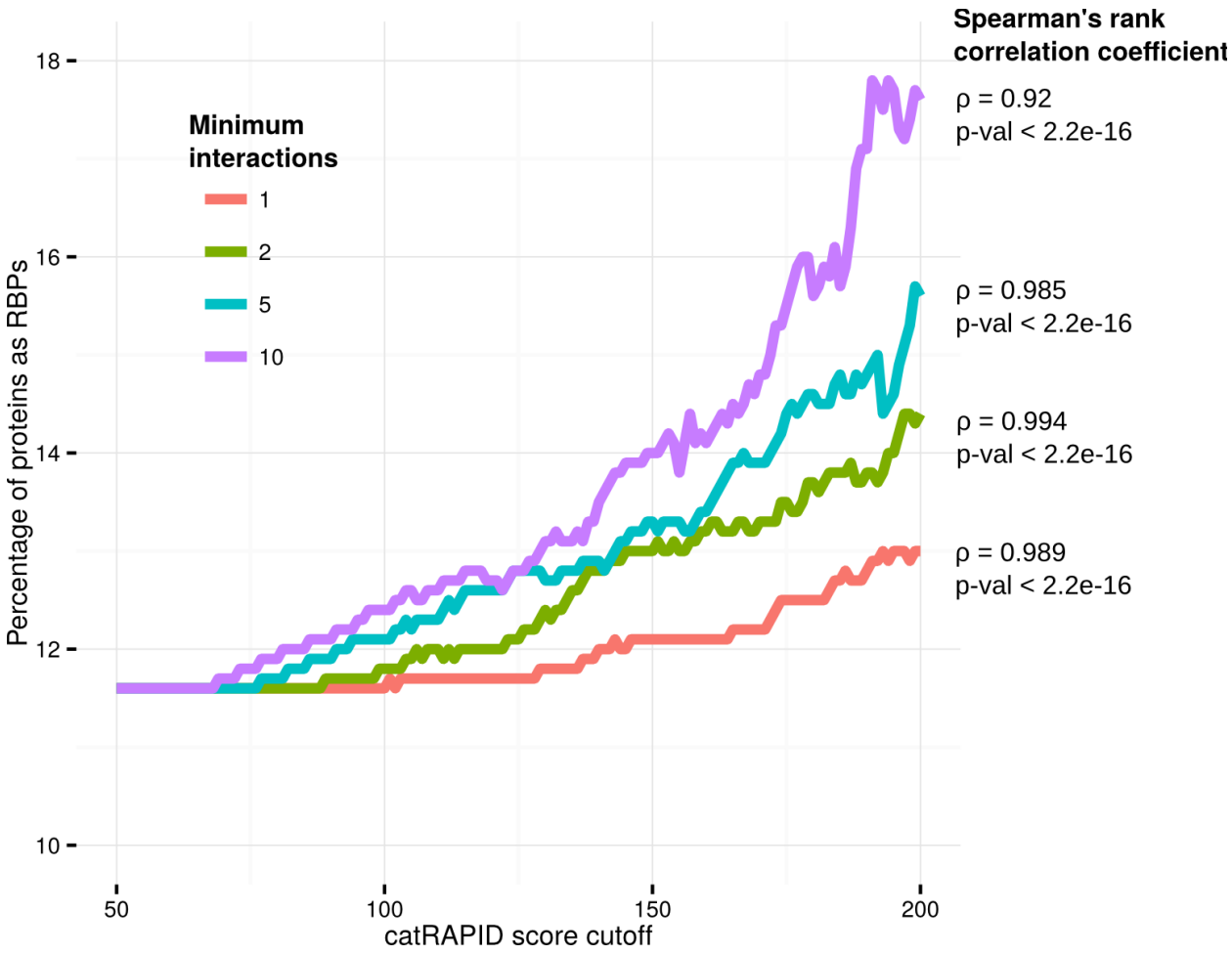


Figure S7. Percentage of proteins annotated as RBPs increases with the stringency of catRAPID score cutoffs. The curves represent the percentage of RBPs among all proteins with at least 1, 2, 5 or 10 interactions to lncRNAs above a certain catRAPID score cutoff (x axis). The increase of the percentage of RBPs with higher catRAPID score cutoffs is statistically significant for all 'minimum interactions' parameters (Spearman rank correlation test; p-values < 2.2e-16).

SUPPLEMENTARY TABLES S6-S7

Table S6. Protein group dataset summary and protein overlap between protein groups of the same dataset. For each protein group in a dataset and for each overlap cutoff (50%, 80%, 100%), we calculate the percentage of protein groups overlapping at least 1 other protein group (from the same dataset) equally or above the overlap cutoff.

Dataset	# protein groups	# proteins in groups	% protein groups >= 50% overlap	% protein groups >= 80% overlap	% protein groups = 100% overlap
Wan 2015	981	2151	12.44	1.12	0
Non-redundant CORUM	324	2133	0	0	0
BioPlex	354	2176	1.41	0.56	0
Network Modules	874	12093	2.97	0.23	0

Table S7. Protein overlap between protein groups of different datasets. For each protein group in dataset and for each overlap cutoff (50%, 80%), we calculate the % of protein groups overlapping at least 1 other protein group (from the other dataset) equally or above the overlap cutoff. The comparison is made both ways (Dataset1 vs Dataset2, Dataset2 vs Dataset1).

Dataset1	Dataset2	% Dataset1 protein groups >= 50% overlap	% Dataset1 protein groups >= 80% overlap	% Dataset2 protein groups >= 50% overlap	% Dataset2 protein groups >= 80% overlap
Wan 2015	Non-redundant CORUM	20.39	15.8	9.88	7.53
Wan 2015	BioPlex	12.23	7.44	11.3	1.98
Wan 2015	Network Modules	12.03	6.63	0	0.11
Non-redundant CORUM	BioPlex	7.1	4.7	14.69	2.82
Non-redundant CORUM	Network Modules	26.23	3.09	0.23	0.11
BioPlex	Network Modules	14.12	1.69	0	0

SUPPLEMENTARY MATERIALS AND METHODS

LncRNA-protein interaction predictions

catRAPID is a protein-RNA interaction predictor trained with public Protein Data Bank protein-RNA structures and using physicochemical features of both molecule types, i.e., secondary structures, hydrogen bonding and van der Waals contributions (1). The catRAPID omics version of the algorithm allows predictions at a large-scale between the transcriptome and the proteome (2). The catRAPID algorithm has been largely benchmarked on different datasets of coding and non-coding RNAs (2–4)

Here we used it to compute interaction propensities between (i) the Ensembl v82 transcripts annotated as GENCODE BASIC and having 'lncRNA' biotypes as defined by GENCODE v24 (5) and (ii) the human canonical proteome as defined by UniProt (6) on May 10, 2016.

Due to computational limitations, catRAPID predictions are restricted to RNA sequences between 50 and 1200 nucleotides of length, as well as to protein sequences between 50 and 750 amino acids. Therefore, we assess ~79% of the human canonical proteome (15974 proteins) and ~81% of the human long non-coding transcriptome (15230 transcripts), producing more than 243 million protein-RNA interactions (Fig. 1A). As applied in previous works (4), only predictions with interaction propensity score of at least 50 were determined as positive and kept for further analyses.

Collection of network modules

Protein network modules were extracted from a human interactome assembled as described in Chapple et al 2015 (9). Briefly, protein interaction data were gathered from several databases through the PSICQUIC query interface (10). Binary (i.e., likely direct) interactions (according to the experimental detection method) were kept. Sequence redundancy at 95% identity among the proteins of the interactome was reduced using the CD-HIT algorithm (11). A human binary interactome containing 61695 interactions between 12318 proteins (February 2016) has been obtained. By applying the Overlapping Cluster Generator algorithm that identifies overlapping clusters based on modularity (12) to this network, 874 network modules have been detected.

Cell culture conditions and transfection

C2C12 murine myoblasts were cultured as previously described (13). and differentiated in 0.5% Fetal Bovine Serum (FBS). Human control healthy myoblasts from the Telethon Neuromuscular Biobank

were cultured as previously described (14) and differentiated with human skeletal muscle differentiation medium (PromoCell).

Endogenous lnc-405 RNA pull-down

lnc-405 pull-down was performed on nuclear extract obtained with some modification of the Rinn's protocol (15). Nuclear pellet was resuspended in 3ml NT2 Buffer (50mM TrisHCl, 150mM NaCl, 10mM EDTA, 0,05%NP40, 1mM MgCl, 1mM DTT, PMSF, RNasin ribonuclease and protease inhibitors). Resuspended nuclei were dounced with 15-25 strokes. Nuclear membranes and debris were pelleted by centrifugation at 13000 rpm, 10 min at 4°C. 6mg of nuclear extract (2mg/sample) were pre-cleared with 300µl of magnetic beads (Promega, ref#Z5481) for 30 min at room temperature (RT) and subsequently incubated 2 hours on rotator with 100nM of the biotinilated probes. For RNA precipitation, 200µl of the beads were added to the extract and incubated 30 min at RT. Co-precipitated proteins were isolated by resuspending 4/5 of the beads in 30µl of elution buffer (10 glycerol, 2% SDS, 60mM Tris-HCl, 1mM DTT and protease inhibitor cocktail (Roche)) and analysed by Mass-spec. 1/5 of the beads were resuspended in 1ml of Trizol for RNA analyses.

Mass spectrometry analysis

lnc-405 co-purified proteins were run on a stacking SDS-PAGE gel and digested in trypsin solution 12.5ng/ml (Trypsin porcine, Promega). Samples were analysed using an Orbitrap ELITE mass spectrometer (Thermo Scientific, San Jose California). Peptide mixtures were separated on C18 Accucore nanoColumn (75µm ID x 50 cm, Thermo Fisher Scientific) with a 2 hours gradient long. For data analysis, proteins were identified by database searching using SequestHT (Thermo Fisher Scientific) with Proteome Discoverer 1.4 software (Thermo Fisher Scientific) against the spMouse_2015_02 database (24721 entries, canonical & isoforms). Peptides were filtered with a false discovery rate (FDR) at 1% and 2 unique peptides minimum/proteins. Proteins annotated as 'Keratin' were excluded as known contaminants. Proteins where the score (sum of the ion scores of identified peptides) of either the ODD or EVEN samples is at least 1.5-times higher than the LacZ sample (control) are considered as lnc-405 interactors (19 final interactors).

RNA immunoprecipitation

For each condition (IP, IgG and BO) 1.5mg of pre-cleared extract was immunoprecipitated (4°C, O.N.) with 100µl of protein G agarose beads (Millipore cat#16-201) and 10µg of immobilized antibody or IgG. Antibody treatments included anti-Purβ (Bethyl, cat#A303-650A) or rabbit IgG (Santacruz, cat#sc2027). Beads were washed 5 times in 500µl of NT2 buffer (50mM Tris pH 8, 150mM NaCl, 1mM MgCl2, 0.5% NP40, 20mM EDTA, 1mM DTT, PMSF, RNasin ribonuclease and protease inhibitors) and

resuspended in 200µl. Co-precipitated proteins were isolated by resuspending 4/5 of the beads in the elution buffer (10 glycerol, 2% SDS, 60mM Tris-HCl, 1mM DTT and protease inhibitors (Roche)) and analysed by Western Blot. Co-precipitated RNAs were isolated by resuspending 1/5 of the beads in 1ml of Trizol prior RT-PCR analyses.

List and sequences of the oligonucleotides and probes used

qRT-PCR

mmu-lnc-405 FW	5'-gcaggaagcaaaagatcagc-3'
mmu-lnc-405 RV	5'-aagtcagccgaggtctttca-3'
mmu-GAPDH FW	5'-ggctcatggtatgtaggcagt-3'
mmu-GAPDH RV	5'-gaaaacacgggggcaatgagt-3'
hs-GADPH FW	5'ggaaggtgaaggtcggagtc3'
hs-GADPH RV	5'ttaccagagttaaaagcagcc3'
hs-lnc-405 RV	5'-gcagattcaggagcccact-3'
hs-lnc-405 FW	5'-aggattccacgcactcagaa-3'

Biotinilated probes (mouse)

lnc-405_ODD_1	5'-ttcattgctgatgcctgaag-3'
lnc-405_ODD_2	5'-tagcagctgcaggttttcag-3
lnc-405_ODD_3	5'-ccaaacacacagatggcaga-3'
lnc-405_ODD_4	5'-ctagtatagctgggttgcag-3'
lnc-405_EVEN_1	5'-ctctccgagctgatcttttg-3'
lnc-405_EVEN_2	5'-gcaggtgtacagatgttttc-3'
lnc-405_EVEN_3	5'-ggcaggtgagtcctaagaag-3'
lnc-405_EVEN_4	5'-tggtgacagagtcccattag-3'
LACZ_1	5'-ccagtgaatccgtaatcatg-3'

LACZ_2	5'-tcacgacgttgtaaaacgac-3'
LACZ_3	5'-agatgaaacgccgagttaac-3'
LACZ_4	5'-tttctccggcgcgtaaaaa-3'

Dataset of experimentally determined protein-RNA interactions

For the creation of a compendium of known protein-lncRNA interactions, we collected information from (i) the NPInter v3.0 (16) human ncRNA-protein binding interaction dataset (on 01 April 2016), NONCODE IDs were converted into Ensembl transcript IDs using NONCODE 2016 (17); (ii) StarBase v2.0 (8) low-stringency RBP-LncRNA interaction dataset, with at least 100 reads mapped to hg19 (on 20 April 2016); (iii) ENCODE enhanced CLIP (eCLIP) dataset (18), 159 experiments (from 112 RBPs). BED peak coordinates referencing the GRCh38 human assembly were mapped to Gencode v24 transcripts models using BEDTools intersect v2.17 (19) with flags -w and -a. Interactions from replicates and different cell lines were combined.

Protein-lncRNA interactions from these three datasets were combined and intersected with the dataset of protein-lncRNA catRAPID predictions to form our “known protein-lncRNA interactions” dataset including 125384 interactions between 148 RBPs and 10965 lncRNAs. When required, transcript gene names were converted to Ensembl v82 transcript IDs using the Ensembl BioMart service (expanding interaction to all transcripts of the gene).

Functional lncRNA datasets

Datasets of lncRNA transcripts or genes identified to be functional or to possess functional features were retrieved from several databases and publications: (i) Hon et al. 2017 (20): list of 124 Ensembl genes predicted to be functional with four functional evidences (Supplementary table 17); (ii) Liu et al. 2016 (21): list of 689 Ensembl genes that affect cell growth in CRISPRi experiments (Supplementary Table 1); (iii) Lnc2cancer (22): list of 381 Ensembl genes retrieved from Lnc2cancer database on January 3, 2017; (iv) LncRNADisease (23): list of 215 Ensembl genes retrieved from LncRNADisease database on January 3, 2017; (v) LncRNADB (24): list of 118 Ensembl genes retrieved from LncRNADB v2.0 on December 30, 2016; (vi) Mukherjee et al. 2017 (25): list of 8580 Ensembl genes (coding and non-coding) which possess features of functionality (Clusters c1, c2 and c3, Supplementary table 3); (vii) Necsulea et al. 2014 (26): list of 457 Ensembl human genes predicted to be conserved in Therians or Eutherians (Supplementary Data 1); (viii) Smith et al. 2013 (27): list of 17704 Ensembl coding and non-coding genes with a conserved structural element in one of its exons. Briefly, we obtained human genome regions with conserved structural features (ECS congruous), converted GRCh37 coordinates to GRCh38 using UCSC

liftOver tool (28) and mapped to Gencode V24 exon annotations using BEDTools intersect with parameters -s -f 1.0 (ensure same strandedness and ensure the complete conserved structural region is inside the exon).

For certain analysis, lncRNA genes from the above datasets were combined and intersected with our dataset of protein-lncRNA interaction predictions, resulting in 2013 functional genes.

When required, transcript gene names were converted to Ensembl v82 transcript IDs using the Ensembl BioMart service (expanding interaction to all transcripts of the gene).

Fitness consequence scores

Mutation fitness consequence scores (fitCons) on human genome were obtained from Gulko et al. 2015 (29) (v1.01, integrated across cell types). Coordinates were converted to GRCh38 using UCSC liftOver tool (28), fitCons scores were mapped to Gencode v24 UTR and lincRNA exons and their distribution across different sets of transcripts and features was assessed. Protein-coding genes were randomly subsampled to the same number of lincRNA genes assessed in this study (7448 genes, 330 of them being scaffolding candidates), and their 3'UTRs and 5'UTRs were used as a control. This analysis was performed on lincRNA instead of all lncRNAs to avoid a potential mutational bias of lncRNAs overlapping protein-coding genes.

Disease-associated lncRNAs and proteins

Lists of lncRNAs associated to disease where collected from Lnc2cancer database (22) and LncRNADisease database (23) (see 'Functional lncRNA datasets'). Lists of proteins associated to disease where collected from Online Mendelian Inheritance in Man (OMIM) (30), downloaded through Ensembl v87 BioMart (on 04 Jan 2017). HGNC protein identifiers were used and converted to UniprotKB accession numbers (ACs). A total of 5397 protein-disease associations between 3499 proteins and 4387 diseases were collected.

Compendium of RNA-binding proteins

A total of 2129 human RNA-binding proteins were gathered from several sources: (i) Neelamraju et al. 2015 Supplementary Table 1 (31) (ii); Gerstberger et al. 2014 (32) Supplementary information S1; (iii) Beckmann et al. 2015 (33) Supplementary Data Set 2; (iv) Castello et al. 2016 (34) Supplementary Table 1, "RBDpep" sheet; (v) Conrad et al. 2016 (35) Supplementary Table 1 sheet "identified RBPs" found in either "chromatin" or "nuclei". When IDs were provided as gene or protein names, these were converted to UniprotKB ACs. In this study we produce and assess interaction predictions for 1459 of these 2129 RBPs (after filterings).

SUPPLEMENTARY MATERIAL REFERENCES

1. Bellucci,M., Agostini,F., Masin,M. and Tartaglia,G.G. (2011) Predicting protein associations with long noncoding RNAs. Nat. methods, 8, 444–445.

2. Agostini,F., Zanzoni,A., Klus,P., Marchese,D., Cirillo,D. and Tartaglia,G.G. (2013) CatRAPID omics: A web server for large-scale prediction of protein-RNA interactions. Bioinformatics, 29, 2928–2930.

3. Cirillo,D., Marchese,D., Agostini,F., Livi,C.M., Botta-Orfila,T. and Tartaglia,G.G. (2014) Constitutive patterns of gene expression regulated by RNA-binding proteins. Genome Biol., 15, R13.

4. Zanzoni,A., Marchese,D., Agostini,F., Bolognesi,B., Cirillo,D., Botta-Orfila,M., Livi,C.M., Rodriguez-Mulero,S. and 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.

5. Harrow,J., Frankish,A., Gonzalez,J.M., Tapanari,E., Diekhans,M., Kokocinski,F., Aken,B.L., Barrell,D., Zadissa,A., Searle,S., et al. (2012) GENCODE: The reference human genome annotation for the ENCODE project. Genome Res., 22, 1760–1774.

6. Wasmuth,E.V. and Lima,C.D. (2016) UniProt: the universal protein knowledgebase. Nucleic Acids Res., 45, 1–12.

7. GTEx Consortium (2015) Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. Sci., 348, 648–660.

8. Li,J.H., Liu,S., Zhou,H., Qu,L.H. and Yang,J.H. (2014) StarBase v2.0: Decoding miRNA-ceRNA, miRNA-ncRNA and protein-RNA interaction networks from large-scale CLIP-Seq data. Nucleic Acids Res., 42, 92–97.

9. Chapple,C.E., Robisson,B., Spinelli,L., Guien,C., Becker,E. and Brun,C. (2015) Extreme multifunctional proteins identified from a human protein interaction network. Nat. Commun., 6, 7412.

10. Aranda,B., Blankenburg,H., Kerrien,S., Brinkman,F.S.L., Ceol,A., Chautard,E., Dana,J.M., De Las Rivas,J., Dumousseau,M., Galeota,E., et al. (2011) PSICQUIC and PSISCORE: accessing and scoring molecular interactions. Nat. methods, 8, 528–529.

11. Fu,L., Niu,B., Zhu,Z., Wu,S. and Li,W. (2012) CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics, 28.

12. Becker,E., Robisson,B., Chapple,C.E., Guénoche,A. and Brun,C. (2012) Multifunctional proteins revealed by overlapping clustering in protein interaction network. Bioinforma., 28, 84–90.

13. Ballarino,M., Cazzella,V., D'Andrea,D., Grassi,L., Bisceglie,L., Cipriano,A., Santini,T., Pinnarò,C., Morlando,M., Tramontano,A., et al. (2015) Novel long noncoding RNAs (lncRNAs) in myogenesis: a miR-31 overlapping lncRNA transcript controls myoblast differentiation. Mol. Cell. Biol., 35, 728–736.

14. Cazzella,V., Martone,J., Pinnarò,C., Santini,T., Twayana,S.S., Sthandier,O., D'Amico,A., Ricotti,V., Bertini,E., Muntoni,F., et al. (2012) Exon 45 skipping through U1-snRNA antisense molecules recovers the Dys-nNOS pathway and muscle differentiation in human DMD myoblasts. Mol. Ther. J. Am. Soc. Gene Ther., 20, 2134–2142.

15. Rinn,J.L., Kertesz,M., Wang,J.K., Squazzo,S.L., Xu,X., Brugmann,S.A., Goodnough,L.H., Helms,J.A., Farnham,P.J., Segal,E., et al. (2007) Functional demarcation of active and silent chromatin domains in human HOX loci by noncoding RNAs. Cell, 129, 1311–1323.

16. Hao,Y., Wu,W., Li,H., Yuan,J., Luo,J., Zhao,Y. and Chen,R. (2016) NPInter v3.0: an upgraded database of noncoding RNA-associated interactions. Database: J. Biol. databases Curation, 2016, baw057.

17. Zhao,Y., Li,H., Fang,S., Kang,Y., Hao,Y., Li,Z., Bu,D., Sun,N., Zhang,M.Q. and Chen,R. (2016) NONCODE 2016: an informative and valuable data source of long non-coding RNAs. 44, D203–D208.

18. 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., et al. (2016) Robust transcriptome-wide discovery of RNA-binding protein binding sites with enhanced CLIP (eCLIP). Nat. methods, 13, 1–9.

19. Quinlan,A.R. and Hall,I.M. (2010) BEDTools: A flexible suite of utilities for comparing genomic features. Bioinformatics, 26, 841–842.

20. Hon,C., Ramilowski,J., Harshbarger,J., Bertin,N., Rackham,O., Gough,J., Denisenko,E., Schmeier,S., Poulsen,T., Severin,J., et al. (2017) An atlas of human long non-coding RNAs with accurate 5‘ ends. Nature, 9, 543.

21. Liu,S.J., Liu,S.J., Horlbeck,M.A., Cho,S.W., Birk,H.S., Malatesta,M., Attenello,F.J., Villalta,J.E., Cho,M.Y., Chen,Y., et al. (2016) CRISPRi-based genome-scale identification of functional long noncoding RNA loci in human cells. Sci., 355.

22. Ning,S., Zhang,J., Wang,P., Zhi,H., Wang,J., Liu,Y., Gao,Y., Guo,M., Yue,M., Wang,L., et al. (2016) Lnc2Cancer: A manually curated database of experimentally supported lncRNAs associated with various human cancers. Nucleic Acids Res., 44, D980–D985.

23. Chen,G., Wang,Z., Wang,D., Qiu,C., Liu,M., Chen,X., Zhang,Q., Yan,G. and Cui,Q. (2013) LncRNADisease: A database for long-non-coding RNA-associated diseases. Nucleic Acids Res., 41, 983–986.

24. Quek,X.C., Thomson,D.W., Maag,J.L.V., Bartonicek,N., Signal,B., Clark,M.B., Gloss,B.S. and Dinger,M.E. (2015) lncRNAdb v2.0: Expanding the reference database for functional long noncoding RNAs. Nucleic Acids Res., 43, D168–D173.

25. Mukherjee,N., Calviello,L., Hirsekorn,A., de Pretis,S., Pelizzola,M. and Ohler,U. (2017) Integrative classification of human coding and noncoding genes through RNA metabolism profiles. Nat. Struct. & Mol. Biol., 24, 86–96.

26. Necsulea,A., Soumillon,M., Warnefors,M., Liechti,A., Daish,T., Zeller,U., Baker,J.C., Grützner,F. and Kaessmann,H. (2014) The evolution of lncRNA repertoires and expression patterns in tetrapods. Nature, 505, 635–640.

27. Smith,M.A., Gesell,T., Stadler,P.F. and Mattick,J.S. (2013) Widespread purifying selection on RNA structure in mammals. Nucleic Acids Res., 41, 8220–8236.

28. Tyner,C., Barber,G.P., Casper,J., Clawson,H., Diekhans,M., Eisenhart,C., Fischer,C.M., Gibson,D., Gonzalez,J.N. and Guruvadoo,L. (2016) The UCSC Genome Browser database: 2017 update. Nucleic Acids Res., 45, D626–D634.

Appendix II: Article supplementary material

SUPPLEMENTARY MATERIAL

MoonDB 2.0: an updated database of extreme multifunctional and moonlighting proteins

Content:

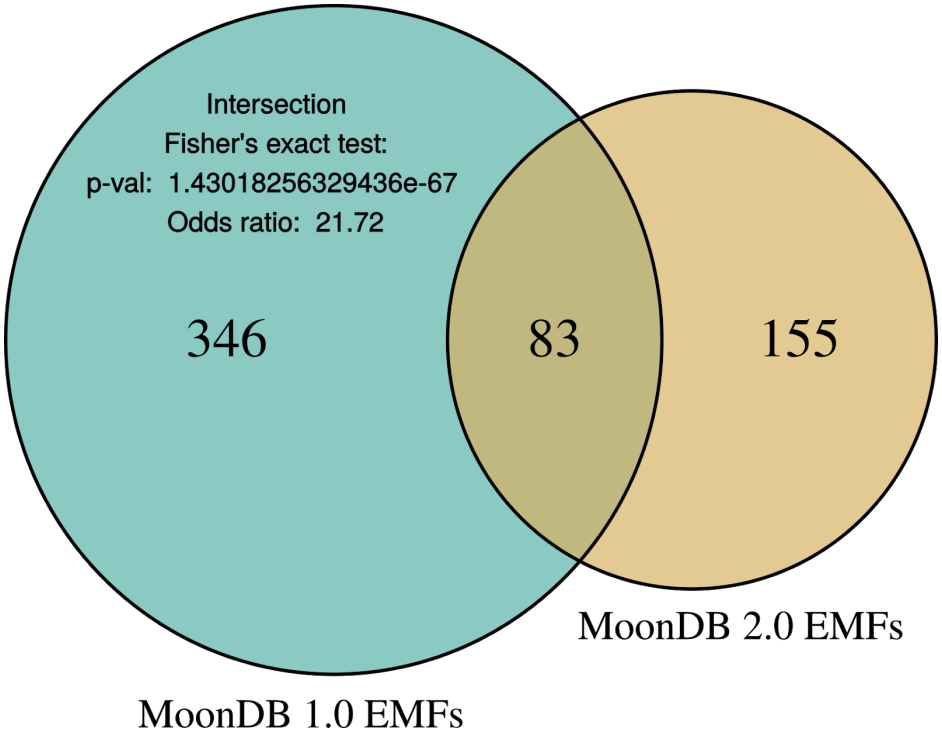


Figure S1. **Overlap between the human EMFs predicted in MoonDB 1.0 and MoonDB 2.0.** Fisher's Exact test (two-sided). Background included all proteins in the protein-protein interaction network (14074 proteins) analysed in this study.

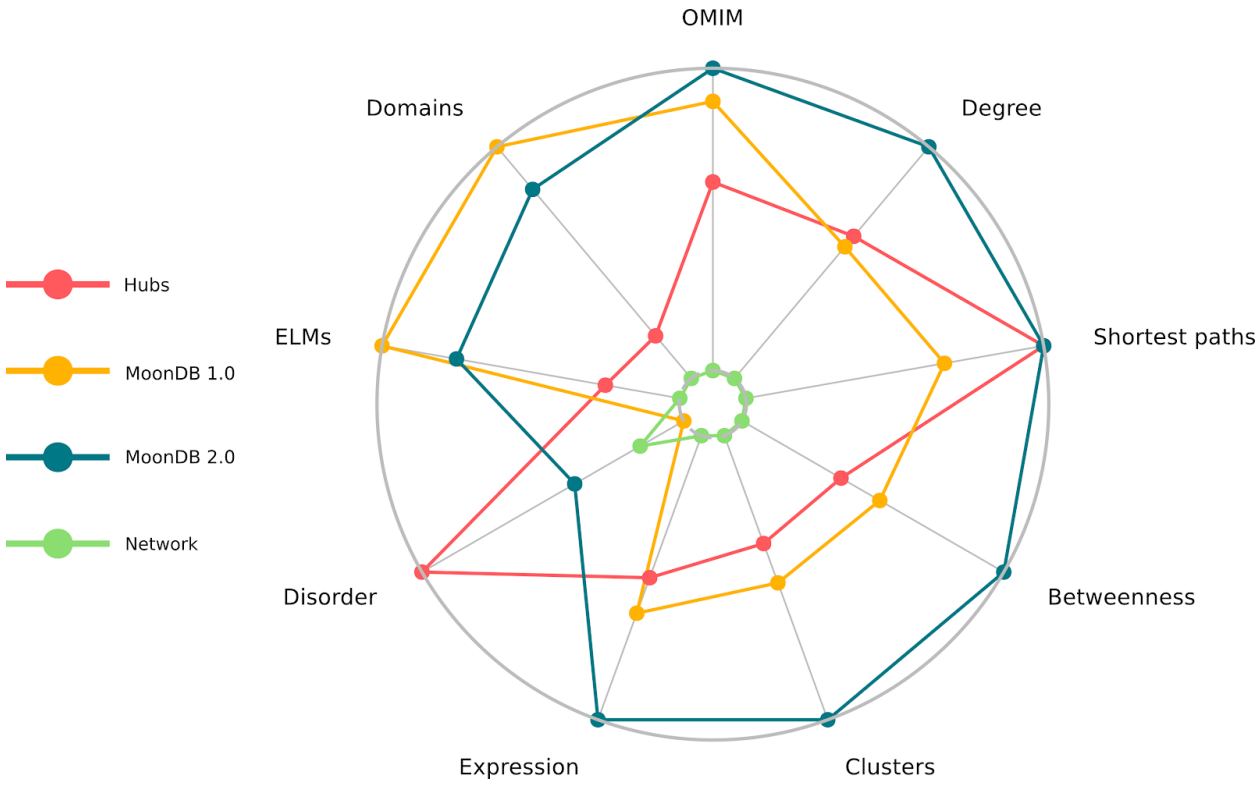


Figure S2. **Radar plot comparing features of EMFs predicted in MoonDB 1.0 and MoonDB 2.0**. The features '**Degree**', '**Shortest paths'** and **'Betweenness'** were calculated using the igraph (v0.7.1) Python (v2.7.6) library for each protein group with the updated human interactome used in this study. **'Clusters'** represents the mean number of network modules where the proteins are present. **'Expression'** refers to the number of different tissues where a protein can be found, obtained from the Human Protein Atlas v18 (1), excluding entries with presence as 'not detected' or with 'unsupported' reliability. **'Disorder'** represents the proportion of amino acids with a MobiDB v3.0 (2) predicted consensus disorder score of 0.5 or more. **'ELMs'** (Eukaryotic Linear Motifs) were obtained from ELM DB (3) on February 2018 and represent the number of ELMs per amino acid, considering only disordered amino acids (see above). **'Domains'** refer to number of domains present in the proteins and were obtained from UniProt on January 2018. **'OMIM'** represents the Fisher's Exact test odds ratio for a protein group to have more proteins with at least one OMIM (data retrieved on June 2016) (4) disease than expected by chance, using 14074 human proteins with at least one interaction as the background (all cases are significantly different, FDR < 5% after applying Benjamini-Hochberg procedure). For **'Shortest paths'**, the outermost

data point is the most connected, that is, has the smallest shortest path value. For all others, the outermost data point is the one with the highest value. The '**Network**' protein group is composed of all the protein interactome network nodes and the '**hubs**' protein group is composed of those nodes which have a degree that is at least twice the network average.

Table S1. **Number of proteins in interactome, protein-protein interactions, GO term annotations and extreme multifunctional proteins predicted, per species**. Only Biological Process (BP) GO terms annotating proteins in the interactome were considered.

Data	Human	Mouse	Fly	Worm	Yeast
Proteins in interactome	14047	2247	9042	4921	4347
Protein-protein interactions	92345	3329	33998	13158	10364
BP GO terms annotated	105487	42663	35727	34850	30483
EMF proteins predicted	238	14	2	6	32

Table S2. **List of protein-protein interaction experimental methods selected to build the interactome used in MoonDB.** PSI-MI IDs refer to the Molecular Interactions Controlled Vocabulary ontology.

PSI-MI ID	PSI-MI name
MI:0009	bacterial display
MI:0010	beta galactosidase complementation
MI:0011	beta lactamase complementation
MI:0012	bioluminescence resonance energy transfer
MI:0014	adenylate cyclase complementation
MI:0016	circular dichroism
MI:0017	classical fluorescence spectroscopy
MI:0018	two hybrid
MI:0031	protein cross-linking with a bifunctional reagent
MI:0034	display technology
MI:0041	electron nuclear double resonance
MI:0042	electron paramagnetic resonance
MI:0043	electron resonance
MI:0047	far western blotting
MI:0048	filamentous phage display
MI:0049	filter binding
MI:0051	fluorescence technology
MI:0052	fluorescence correlation spectroscopy
MI:0053	fluorescence polarization spectroscopy
MI:0055	fluorescent resonance energy transfer
MI:0065	isothermal titration calorimetry

MI:0066	lambda phage display
MI:0073	mrna display
MI:0081	peptide array
MI:0084	phage display
MI:0089	protein array
MI:0090	protein complementation assay
MI:0092	protein in situ array
MI:0095	proteinchip(r) on a surface-enhanced laser desorption/ionization
MI:0097	reverse ras recruitment system
MI:0098	ribosome display
MI:0099	scintillation proximity assay
MI:0107	surface plasmon resonance
MI:0108	t7 phage display
MI:0111	dihydrofolate reductase reconstruction
MI:0112	ubiquitin reconstruction
MI:0114	x-ray crystallography
MI:0115	yeast display
MI:0231	mammalian protein protein interaction trap
MI:0232	transcriptional complementation assay
MI:0369	lex-a dimerization assay
MI:0370	tox-r dimerization assay
MI:0397	two hybrid array
MI:0398	two hybrid pooling approach
MI:0399	two hybrid fragment pooling approach
MI:0405	competition binding
MI:0406	deacetylase assay
MI:0411	enzyme linked immunosorbent assay
MI:0420	kinase homogeneous time resolved fluorescence
MI:0423	in-gel kinase assay
MI:0424	protein kinase assay
MI:0425	kinase scintillation proximity assay
MI:0434	phosphatase assay
MI:0435	protease assay
MI:0437	protein three hybrid
MI:0440	saturation binding
MI:0508	deacetylase radiometric assay
MI:0509	phosphatase homogeneous time resolved fluorescence
MI:0510	homogeneous time resolved fluorescence
MI:0511	protease homogeneous time resolved fluorescence
MI:0512	zymography
MI:0513	collagen film assay
MI:0514	in gel phosphatase assay
MI:0515	methyltransferase assay
MI:0516	methyltransferase radiometric assay
MI:0655	lambda repressor two hybrid
MI:0678	antibody array
MI:0695	sandwich immunoassay
MI:0696	polymerase assay
MI:0726	reverse two hybrid
MI:0727	lexa b52 complementation
MI:0728	gal4 vp16 complementation
MI:0809	bimolecular fluorescence complementation
MI:0813	proximity ligation assay

MI:0824	x-ray powder diffraction
MI:0825	x-ray fiber diffraction
MI:0827	x-ray tomography
MI:0841	phosphotransferase assay
MI:0870	demethylase assay
MI:0872	atomic force microscopy
MI:0887	histone acetylase assay
MI:0889	acetylase assay
MI:0892	solid phase assay
MI:0894	electron diffraction
MI:0895	protein kinase A complementation
MI:0899	p3 filamentous phage display
MI:0900	p8 filamentous phage display
MI:0905	amplified luminescent proximity homogeneous assay
MI:0916	lexa vp16 complementation
MI:0921	surface plasmon resonance array
MI:0946	ping
MI:0947	bead aggregation assay
MI:0949	gdp/gtp exchange assay
MI:0953	polymerization
MI:0968	biosensor
MI:0969	bio-layer interferometry
MI:0972	phosphopantetheinylase assay
MI:0979	oxidoreductase assay
MI:0984	deaminase assay
MI:0989	amidase assay
MI:0991	lipoproteine cleavage assay
MI:0992	defarnesylase assay
MI:0993	degeranylase assay
MI:0994	demyristoylase assay
MI:0995	depalmitoylase assay
MI:0996	deformylase assay
MI:0997	ubiquitinase assay
MI:0998	deubiquitinase assay
MI:0999	formylase assay
MI:1000	hydroxylase assay
MI:1001	lipidase assay
MI:1002	myristoylase assay
MI:1003	geranylgeranylase assay
MI:1004	palmitoylase assay
MI:1005	adp ribosylase assay
MI:1006	deglycosylase assay
MI:1007	glycosylase assay
MI:1008	sumoylase assay
MI:1009	desumoylase assay
MI:1010	neddylase assay
MI:1011	deneddylase assay
MI:1016	fluorescence recovery after photobleaching
MI:1019	protein phosphatase assay
MI:1026	diphtamidase assay
MI:1030	excimer fluorescence
MI:1031	protein folding/unfolding
MI:1036	nucleotide exchange assay

MI:1037	Split renilla luciferase complementation
MI:1038	silicon nanowire field-effect transistor
MI:1087	monoclonal antibody blockade
MI:1088	phenotype-based detection assay
MI:1089	nuclear translocation assay
MI:1111	two hybrid bait or prey pooling approach
MI:1112	two hybrid prey pooling approach
MI:1113	two hybrid bait and prey pooling approach
MI:1137	carboxylation assay
MI:1138	decarboxylation assay
MI:1142	aminoacylation assay
MI:1145	phospholipase assay
MI:1147	ampylation assay
MI:0729	luminescence based mammalian interactome mapping
MI:1356	validated two hybrid
MI:0729	luminescence based mammalian interactome mapping
MI:0946	miniaturized immunoprecipitation
MI:2168	conditional site labelling
MI:1314	proximity-dependent biotin identification
MI:2189	avexis
MI:2167	kinetic exclusion assay
MI:0071	molecular sieving
MI:1236	proteine isomerase assay
MI:1325	sulfurtranferase assay
MI:1229	uridylation assay
MI:1342	qcmd
MI:1311	differential scanning calorimetry
MI:1219	enzyme-mediated activation of radical sources
MI:1086	equilibrium dialysis
MI:2171	complemented donor-acceptor resonante energy transfer
MI:0859	intermolecular force
MI:2170	bimolecular fluoresence complementation
MI:2169	luminescence technology
MI:1247	microscale thermophoresis
MI:0077	nuclear magnetic resonance
MI:1104	solid state nmr
MI:1103	solution state nmr
MI:0938	rheology measurement
MI:1235	thermal shift binding
MI:1203	split luciferase complementation
MI:1204	split firefly luciferase complementation
MI:1320	membrane yeast two hybrid
MI:1321	ire1 reconstruction

References:

1. Thul,P.J. and Lindskog,C. (2018) The human protein atlas: A spatial map of the human proteome. Protein science : a publication of the Protein Society, 27, 233–244.

2. Piovesan,D., Tabaro,F., Paladin,L., Necci,M., Micetic,I., Camilloni,C., Davey,N., Dosztányi,Z., Mészáros,B., Monzon,A.M., et al. (2018) MobiDB 3.0: more annotations for intrinsic disorder, conformational diversity and interactions in proteins. Nucleic acids Res., 46, D471–D476.

3. Gouw,M., Michael,S., Sámano-Sánchez,H., Kumar,M., Zeke,A., Lang,B., Bely,B., Chemes,L.B., Davey,N.E., Deng,Z., et al. (2018) The eukaryotic linear motif resource - 2018 update. Nucleic acids Res., 46, D428–D434.

4. Hamosh,A., Scott,A.F., Amberger,J.S., Bocchini,C.A. and McKusick,V.A. (2005) Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. Nucleic Acids Res., 33.

Appendix III: Article supplementary material

SUPPLEMENTARY MATERIAL

Prediction of 3′UTR-protein complex assembly reveals a role in the regulation of protein multifunctionality

Authors: Diogo M. Ribeiro, Adrien Teixeira, Andreas Zanzoni, Lionel Spinelli, Christine Brun

Contents:

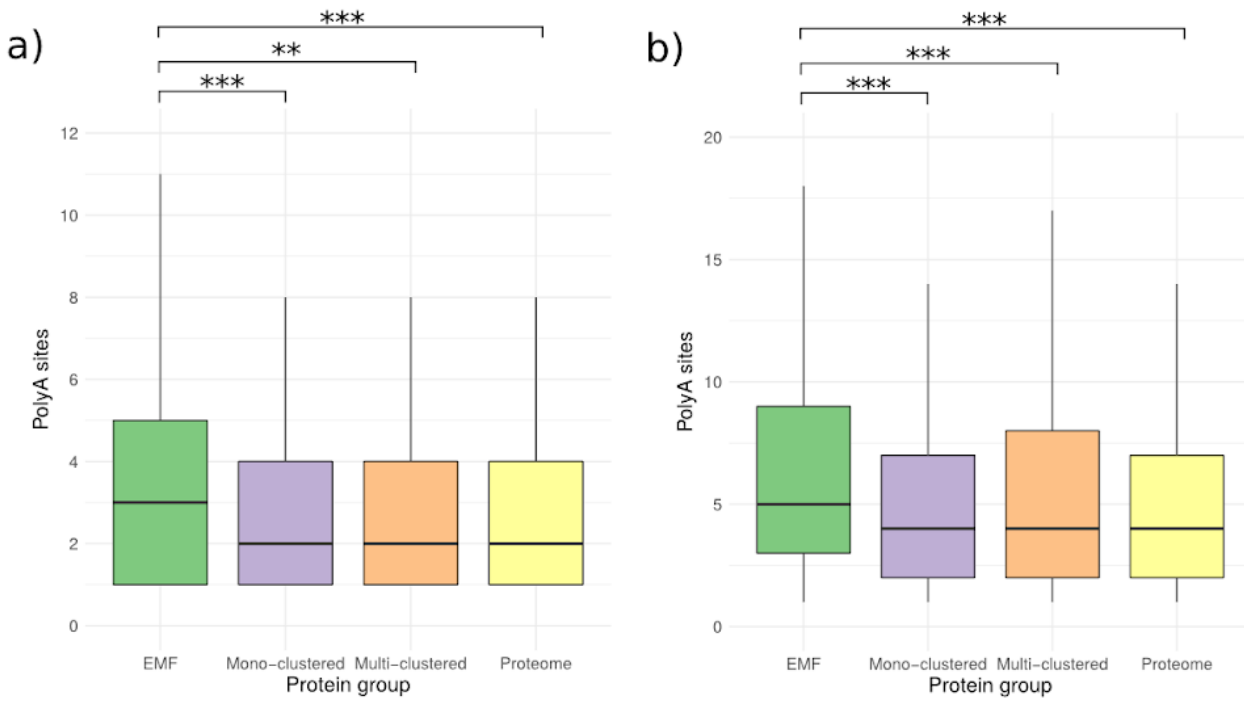


Figure S1. Number of polyadenylation sites in EMF proteins is higher than in other protein groups. (a) Number of APADB database polyadenylation sites in 3'UTRs. **(b)** Number of PolyASite database polyadenylation sites in terminal exons. 'Mono-clustered' represents the proteins in the interactome that belong to only one protein cluster. 'Multi-clustered' proteins in the interactome that belong to several protein clusters but are not considered EMF proteins. Mann-Whitney U tests were performed to test for statistical significance. The Benjamini-Hochberg procedure was applied for multiple test correction. Significance: '*' indicates a FDR < 0.05; '**' indicates a FDR < 0.01; '***' indicates a FDR < 0.001.

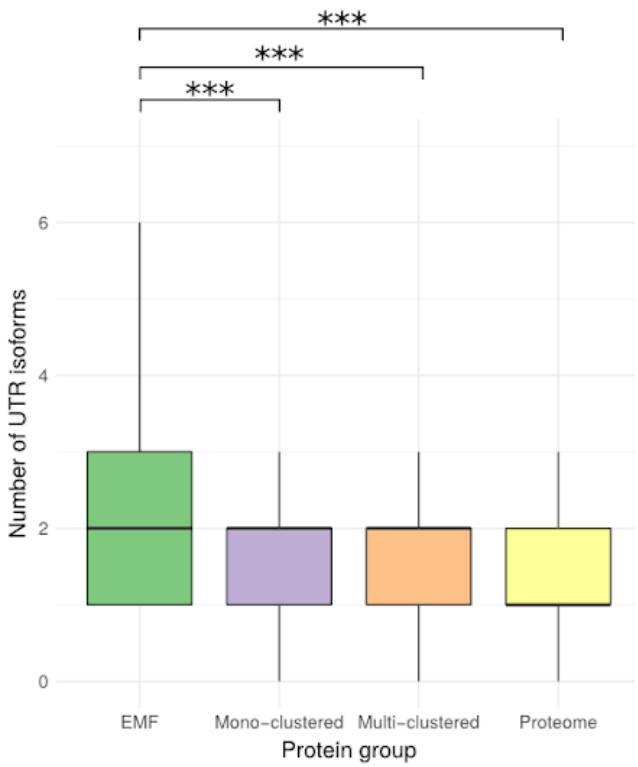


Figure S2. Number of 3'UTR isoforms in EMF proteins is higher than in other protein groups. Number of distinct 3'UTR isoforms among transcripts coding for the same protein. EMF proteins have on average 2.3 isoforms of 3'UTRs, 'Mono-clustered' proteins have 1.9, 'Multi-clustered' proteins have 2.0 and the 'Proteome' average is 1.8. Mann-Whitney U tests were performed to test for statistical significance. The Benjamini-Hochberg procedure was applied for multiple test correction. Significance: '*' indicates a FDR < 0.05; '**' indicates a FDR < 0.01; '***' indicates a FDR < 0.001.

Table S1. List of top 20 intermediate proteins participating in 3'UTR-protein complexes.

Protein ID	Protein names	Gene names	# complexes
P61978	Heterogeneous nuclear ribonucleoprotein K	HNRNPK	43
P78362	SRSF protein kinase 2	SRPK2	33
P68400	Casein kinase II subunit alpha	CSNK2A1	31
P50222	Homeobox protein MOX-2	MEOX2	28
P54253	Ataxin-1	ATXN1	28
P04792	Heat shock protein beta-1	HSPB1	27
P63279	SUMO-conjugating enzyme UBC9	UBE2I	21
Q8TBB1	E3 ubiquitin-protein ligase LNX	LNX1	21
Q15637	Splicing factor 1	SF1	20
P26368	Splicing factor U2AF 65 kDa subunit	U2AF2	19
P62993	Growth factor receptor-bound protein 2	GRB2	19

Q14103	Heterogeneous nuclear ribonucleoprotein D0	HNRNPD	18
P08238	Heat shock protein HSP 90-beta	HSP90AB1	17
P14373	Zinc finger protein RFP	TRIM27	17
Q01081	Splicing factor U2AF 35 kDa subunit	U2AF1	17
Q15427	Splicing factor 3B subunit 4	SF3B4	17
P38398	Breast cancer type 1 susceptibility protein	BRCA1	16
Q14847	LIM and SH3 domain protein 1	LASP1	16
O00560	Syntenin-1	SDCBP	15
Q9NRD5	PRKCA-binding protein	PICK1	13

Table S2. List of top 20 RBPs participating in 3'UTR-protein complexes.

Protein ID	Protein names	Gene names	# complexes
Q93062	RNA-binding protein with multiple splicing	RBPMS	214
P51116	Fragile X mental retardation syndrome-related protein 2	FXR2	129
Q14103	Heterogeneous nuclear ribonucleoprotein D0	HNRNPD	88
P26196	Probable ATP-dependent RNA helicase DDX6	DDX6	64
P61978	Heterogeneous nuclear ribonucleoprotein K	HNRNPK	53
P07910	Heterogeneous nuclear ribonucleoproteins C1/C2	HNRNPC	52
Q01844	RNA-binding protein EWS	EWSR1	47
P38919	Eukaryotic initiation factor 4A-III	EIF4A3	46
P98175	RNA-binding protein 10	RBM10	44
P52597	Heterogeneous nuclear ribonucleoprotein F	HNRNPF	40
Q07955	Serine/arginine-rich splicing factor 1	SRSF1	38
P26368	Splicing factor U2AF 65 kDa subunit	U2AF2	35
Q01130	Serine/arginine-rich splicing factor 2	SRSF2	34
Q9UKV8	Protein argonaute-2	AGO2	33
P84103	Serine/arginine-rich splicing factor 3	SRSF3	32
Q15717	ELAV-like protein 1 (also know as HuR)	ELAVL1	31
Q92900	Regulator of nonsense transcripts 1	UPF1	29
Q15910	Histone-lysine N-methyltransferase EZH2	EZH2	28
Q99700	Ataxin-2	ATXN2	26
Q9UPQ9	Trinucleotide repeat-containing gene 6B protein	TNRC6B	25

Table S3. Numbers of 3'UTR-protein complexes expected by chance. The total number of 3'UTR-protein complexes expected by chance was measured when using protein-protein interaction networks that had their protein

labels randomly shuffled without replacement (e.g. ‘protein 1’ becomes ‘protein 2’, ‘protein 2’ becomes ‘protein 55’ etc). Empirical p-values for the number of 3’UTR-protein complexes formed using the real protein-protein interaction network to be higher than the null hypothesis distribution of 1000 randomisations was calculated. To discard a potential bias of the degree of proteins (i.e. the number of interactions in the protein interactome), we accounted for the protein degree by building another null hypothesis distribution (1000 randomisations) where all protein labels were substituted among proteins with the same degree (e.g. ‘protein 1 with degree 10’ substituted by ‘protein 123 with degree 10’, etc). This analysis was performed for the whole set of possible nascent proteins (7373 proteins).

Experiment	Total complexes	Empirical p-value
Real network	11119	NA
Randomised networks	3926.5 (1087.5)	0.001
Degree-controlled randomised networks	9877.8 (727.6)	0.047

Table S4. Numbers of nascent proteins localised in the plasma membrane and without conventional translocation signals, using Human Protein Atlas annotations. Percentages denote the proteins retained compared to the previous column. Where indicated, Fisher’s exact tests were performed to test for statistical significance using as background the set of 7373 nascent proteins liable to be assessed for 3’UTR-protein complexes (see Experimental Procedures). The Benjamini-Hochberg procedure was applied for multiple test correction. Significance: ‘*’ indicates a FDR < 0.05; ‘**’ indicates a FDR < 0.01; ‘***’ indicates a FDR < 0.001.

Protein group	Nascents in complex	of which, localised in plasma membrane	of which, contain no signal peptide or transmembrane domain
EMF	140 (58.8%)	28 (20.0%)	26 (92.9%) ***
Multi-clustered	1133 (33.9%)	136 (12.0%)	115 (84.6%) *
Mono-clustered	1372 (13.1%)	144 (10.5%)	120 (83.3%) N.S.

Appendix IV: Article supplementary material

SUPPLEMENTARY MATERIAL

Predicted protein-RNA interactions reveal distinct post-transcriptional regulatory patterns

Andreas Zanzoni[1,#], Lionel Spinelli[1], Diogo M. Ribeiro[1], Gian Gaetano Tartaglia[2,3,4], Christine Brun[1,5,#,§]

[1] Aix-Marseille Univ, INSERM, TAGC, UMR_S1090, Marseille, France;
[2]Centre for Genomic Regulation (CRG), The Barcelona Institute of Science and Technology, Dr Aiguader 88, 08003 Barcelona, Spain;
[3] Universitat Pompeu Fabra (UPF), 08003 Barcelona, Spain;
[4]Institucio Catalana de Recerca i Estudis Avançats (ICREA), 23 Passeig Lluıs Companys, 08010 Barcelona, Spain;
[5] CNRS, Marseille, France.
[#]Address correspondence to: Andreas Zanzoni (andreas.zanzoni@univ-amu.fr) and Christine Brun (christine-g.brun@inserm.fr).

Contents

Supplementary note

An inferred post-transcriptional regulation landscape generated by GSEA. The choice of interaction propensity threshold (*i.e.*, catRAPID score >=50) used to define the positive and negative predicted interactions sets was based on our previous work (*e.g.*, [1,2]). However, to exclude that the chosen score threshold could affect our results based on the Fisher's Exact test, we also performed a threshold-free statistical assessment. For each RBP, we ranked predicted interactors according to their propensity score (*i.e.*, from high to low) and we tested the functional units for enrichment, or depletion, using the gene set enrichment analysis (GSEA) algorithm[3]. We obtained 33'603 significant results (27'270 enrichments and 6333 depletions) for 876 RBPs, being the hepatoma-derived growth factor protein, coded by the gene HDGF, the only with no significant enrichments nor depletions. The number of detected enriched and depleted functional units is twice as much as in the threshold-based test (*i.e.*, 604 functional units compared to 300).

As for the threshold-based predicted functional landscape, we observed a similar enrichment/depletion pattern for RBPs and functional units (Figure S2A). Twenty-four RBPs had exclusively enriched functional units among their predicted targets (RBP-1 set. The clear majority of RBPs (*i.e.*, 775, RBP-2 set) had both significant enrichments and depletions. Only 77 showed only significant depletions (RBP-3 set).

We identified three groups of functional units: a predominant subset of exclusively enriched units (*i.e.*, 358, FU-1) and two smaller groups of both enriched/depleted functional units (*i.e.*, 130, FU-2) and exclusively depleted (*i.e.*, 116, FU-3). Finally, around 90% of both functional enrichments (88.5%) and depletions (92.4%) in the threshold-based predicted functional landscape were detected as such by the threshold-free analysis.

This comparison, on one hand, confirms the results obtained by the threshold-based approach and, on the other, complements it by expanding the RBP predicted functional landscapes. The GSEA results are provided in Table S5.

Functional enrichments and depletions in the eCLIP dataset. We collected interaction for 112 RBPs from the ENCODE eCLIP dataset (see Methods) and applied the function unit enrichment analysis based on the Fisher's Exact test as presented in the main text. We obtained 14'660 significant results (13'119 enrichments and 1541 depletions) for ninety-nine RBPs (88% of the total). The number of functional units with significant results is considerably higher (i.e., 1339 units, ~45% of the functional units tested) compared to the one obtained based on predicted interactions.

We also observed different patterns of enrichments and depletions for both RBPs and functional units (Figure S2B). Indeed, 20 RBPs of them had exclusively enriched functional units among their targets (RBP-1 set), whereas 79 showed both functional enrichments and depletions (RBP-2 set). We did not detect any RBP with significant depletion only. We detected three groups of functional units with distinct enrichment/depletion patterns: 1074 exclusively enriched units, 207 depleted only and 58 functional units that were both enriched and depleted. The functional annotation of eCLIP interactions are provided in Table S6.

References

1. Zanzoni, A. *et al.* Principles of self-organization in biological pathways: a hypothesis on the autogenous association of alpha-synuclein. *Nucleic Acids Res.* **41,** 9987–9998 (2013).

2. Ribeiro, D. M. *et al.* Protein complex scaffolding predicted as a prevalent function of long non-coding RNAs. *Nucleic Acids Res.* **46,** 917–928 (2018).

3. Subramanian, A. *et al.* Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *PNAS* **102,** 15545–15550 (2005).

4. Gerstberger, S., Hafner, M. & Tuschl, T. A census of human RNA-binding proteins. *Nat. Rev. Genet.* **15,** 829–845 (2014).

Supplementary Figures

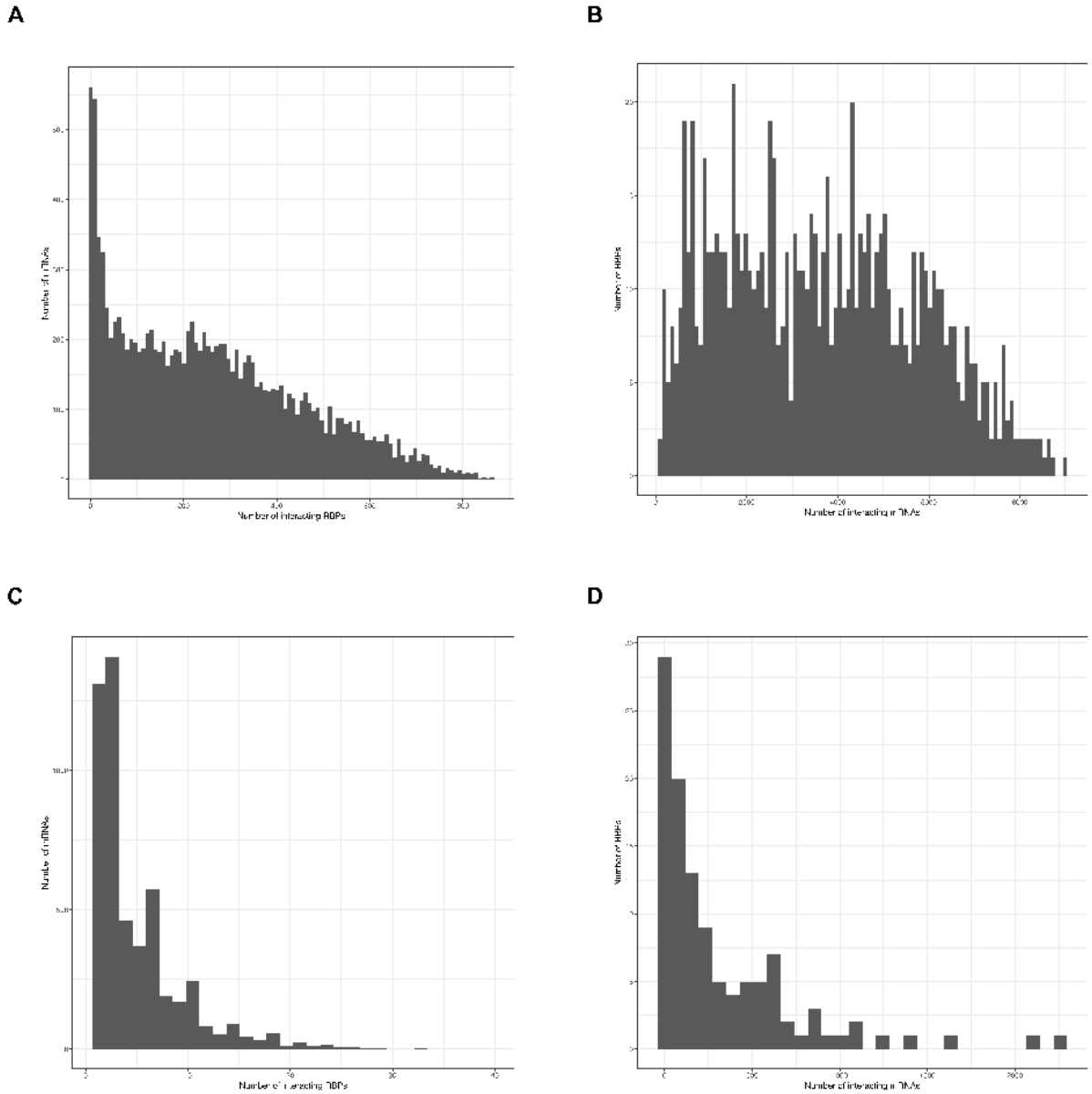


Figure S1. Distributions of protein-mRNA interactions in the PRI network and eCLIP interaction dataset. (A) Distribution of the numbers of interacting RBP per coding transcript in the PRI network. (B)

Distribution of the numbers of interacting mRNA per RBP in the PRI network. (C) Distribution of the numbers of interacting RBPs per coding transcript in the eCLIP interaction dataset. (D) Distribution of the numbers of interacting mRNA per RBP in the eCLIP interaction dataset.

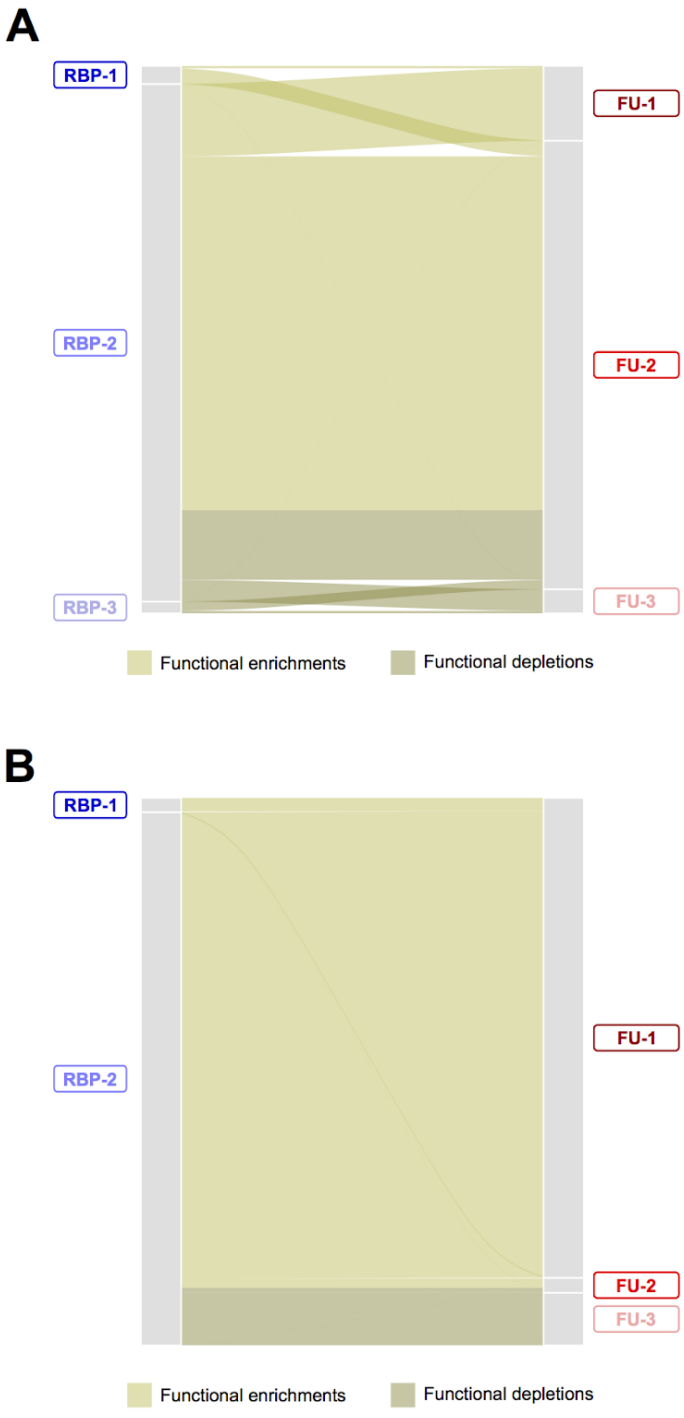


Figure S2. Alluvial plot depicting the functional relationships among RBP (shades of blue color) and FU (FU, shade of red color) groups in (A) the GSEA-predicted functional regulatory landscape and (B) in the eCLIP interaction dataset. The thickness of each stream is proportional to the number of enrichment or

depletions between two given groups. The size of the grey blocks is proportional to the number of enrichments/depletions in which a given RBP or FU group is involved.

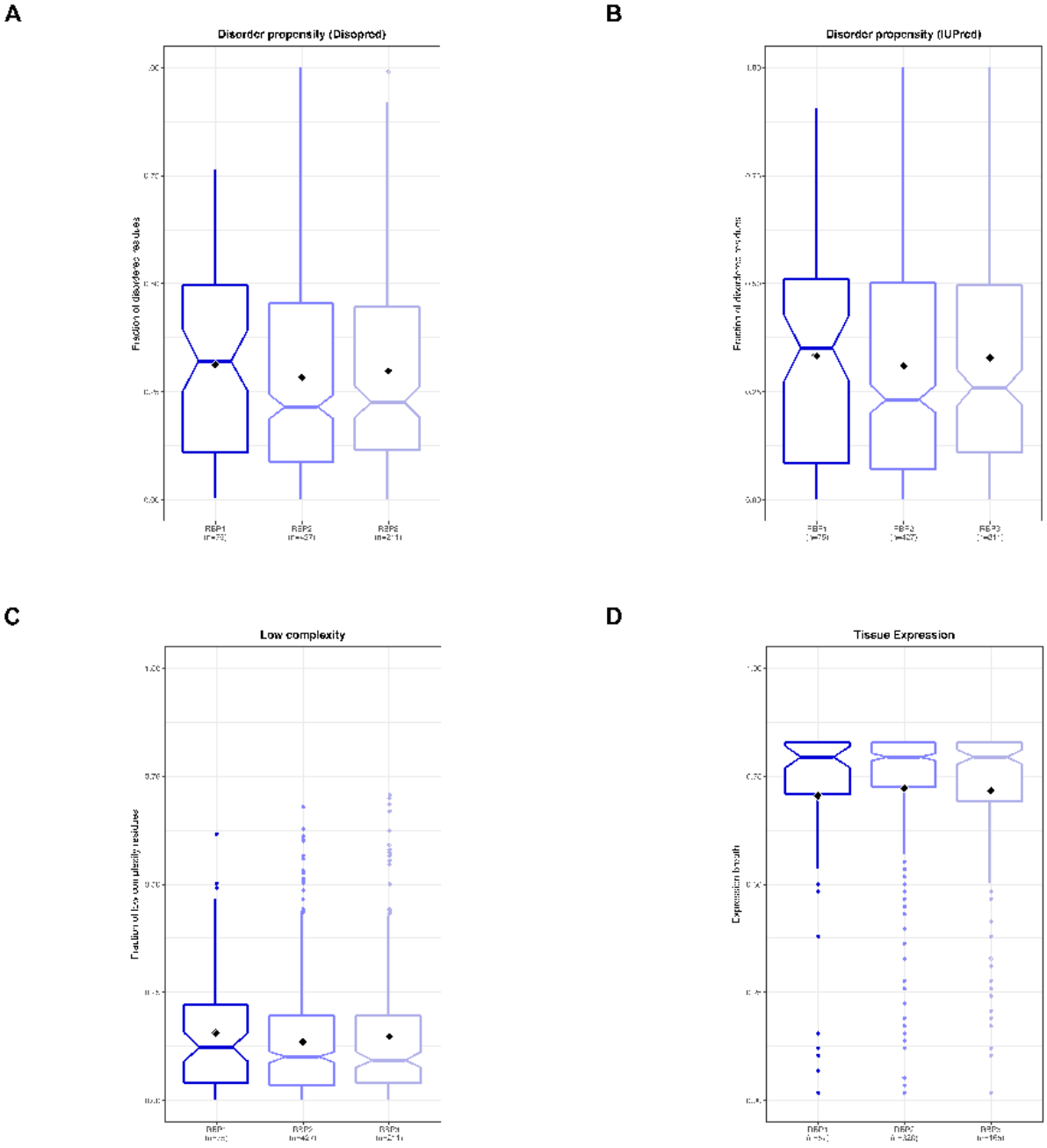


Figure S3 – Sequence and functional properties of RBPs. (A) Disorder distributions of the sequences of the three RBP groups based on DISOPRED3 predictions. Disorder content is estimated as the number predicted disordered residues divided by the RBP sequence length. (B) Disorder distributions of the sequences of the three RBP groups based on IUPred 'long' predictions. (C) Low complexity distributions of the sequences of the three RBP groups predicted by the SEG algorithm. (D) Tissue expression distributions of the sequences of the three RBP groups based on Human Protein Atlas (HPA) data. Tissue

expression is estimated as expression breath, that is the number of tissues in which a given RBP is detected divided by the total number of tissues present in HPA (i.e., 58).

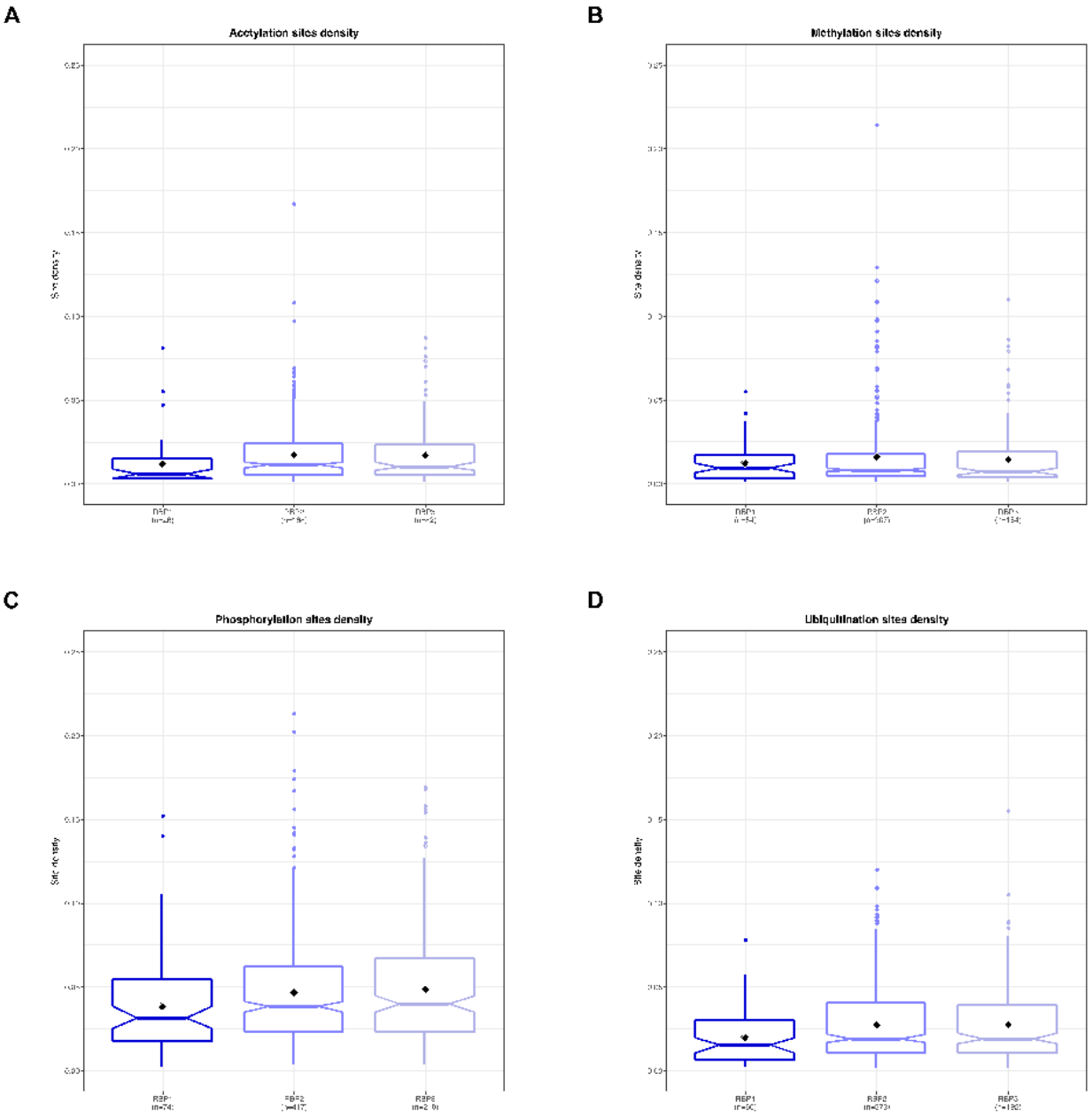


Figure S4 – Distribution of the post-translational modification (PTM) density in the sequences of the three RBP groups. Densities for every RBP are computed as the number of experimentally identified PTM sites divided by the RBP sequence length. Black diamonds represent density mean values. Boxplot colors

correspond to the RBP group colors in Figure 2. (A) Acetylation. (B) Methylation, (C) Phosphorylation and (D) Ubiquitination. See Table S8 for the statistical results.

Supplementary Tables

Table S8. Classification of the RBPs in the three sets based on their RNA biotype binding preference. Predominant RNA target biotypes were defined based on the "consensus target" annotation from Gerstberger *et al.*[4]. For those RBPs were assigned an "unknown" target or were not annotated, as they were identified in mRNA-interactome capture experiments, we assign mRNA as their consensus target. For each RBP set the number and the fraction of RBP with a given predominant RNA biotype target is reported.

RNA biotype	**RBP-1**	**RBP-2**	**RBP-3**
mRNA	62 (82.6%)	283 (66.2%)	135 (64%)
ncRNA	8 (10.6%)	15 (3.5%)	4 (1.8%)
ribosome	2 (2.6%)	59 (14%)	27 (18.8%)
rRNA	5 (6.6%)	25 (5.8%)	23 (11%)
snoRNA	2 (2.6%)	11 (2.5%)	3 (1.4%)
snRNA	0 (0.0%)	15 (3.5%)	6 (2.8%)
diverse	1 (1.3%)	1 (<1%)	1 (2%)

Table S9. Results of the comparison of the distributions of the post-translational modification (PTM) densities in the sequences of the three RBP groups. The Dunn's test was performed only for those PTM for which the Kruskal-Wallis P-value was below 0.05.

PTM types	**P-value (Krustal-Wallis)**		**BH-corrected P-values (Dunn's Test)**		
			RBP-1	*RBP-2*	*RBP-3*
Acetylation	**0.0252**	*RBP-1*		**0.0126**	**0.0089**
		RBP-2			0.4689
Methylation	0.7871				
Phosphorylation	**0.04718**	*RBP-1*		**0.0176**	**0.0248**
		RBP-2			0.3262
Ubiquitination	0.0769				

Appendix V: Scientific contributions outside the scope of this thesis

Before the start of my thesis I worked as a bioinformatician for a few years at the Instituto Gulbenkian de Ciência, Portugal (Dr. Alekos Athanasiadis' lab) and the Wellcome Sanger Institute, UK (Dr. Matthew Berriman's lab). My contributions were mostly centered on *(i)* studying the effect of A-to-I RNA editing on protein evolution, *(ii)* genome assembly and quality control of large helminth genomes, *(iii)* gene discovery and functional annotation, *(iv)* analysis of single nucleotide polymorphisms (SNPs) across helminth strains, *(v)* miRNA target site predictions, *(vi)* comparative genomics analysis of more than 80 helminth species (>1.4 million genes). My work contributed towards several peer-reviewed publications and a protocol, briefly described in this section.

International Helminth Genomes Consortium (2018) "Comparative genomics of the major parasitic worms" ***Nature Genetics*** **(in press).**

BioRxiv: https://www.biorxiv.org/content/early/2017/12/20/236539

Abstract: Parasitic nematodes (roundworms) and platyhelminths (flatworms) cause debilitating chronic infections of humans and animals, decimate crop production and are a major impediment to socioeconomic development. Here we report the broadest comparative study to date of the genomes of parasitic and non-parasitic worms, involving 81. We have identified gene family births and hundreds of expanded gene families at key nodes in the phylogeny that are relevant to parasitism. Examples include gene families that modulate host immune responses, enable parasite migration though host tissues or allow the parasite to feed. We reveal extensive lineage-specific differences in core metabolism and protein families historically targeted for drug development. From an in silico screen, we have identified and prioritised new potential drug targets and compounds for testing. This comparative genomics resource provides a much needed boost for the research community to understand and combat parasitic worms.

Ribeiro D, Coghlan A, Harsha B & Berriman M (2018) "Identification of lineage-specific gene family expansions in a database of gene families" ***Protocol Exchange,*** **doi:10.1038/protex.2018.057.**

Abstract: Gene families specific to, or with significantly changed membership in, particular lineages compared to outgroups may reflect important lineage-specific changes in biology. Here we describe a computational protocol to identify gene families that vary greatly in gene count across a species tree. This protocol uses three different metrics to capture aspects of this variability, and calculates them for each

family in an in-house database of gene families (e.g. built using the Ensembl Compara pipeline). One metric (Cv) identifies families that vary a lot in gene count across the species tree, and the other two (Emax, Zmax) identify families that have an elevated gene count in a certain clade of the species tree. Our protocol controls for differences in gene counts due to fragmented assemblies.

Protasio AV, Dongen S, Collins J, Quintais L, Ribeiro DM, (...), Berriman M (2017) "MiR-277/4989 regulate transcriptional landscape during juvenile to adult transition in the parasitic helminth *Schistosoma mansoni*" *PLoS Neglected Tropical Diseases*, 11(5), p. e0005559.

Abstract: Schistosomes are parasitic helminths that cause schistosomiasis, a disease affecting circa 200 million people, primarily in underprivileged regions of the world. Schistosoma mansoni is the most experimentally tractable schistosome species due to its ease of propagation in the laboratory and the high quality of its genome assembly and annotation. Although there is growing interest in microRNAs (miRNAs) in trematodes, little is known about the role these molecules play in the context of developmental processes. We use the completely unaware "miRNA-blind" bioinformatics tool Sylamer to analyse the 3'-UTRs of transcripts differentially expressed between the juvenile and adult stages. We show that the miR-277/4989 family target sequence is the only one significantly enriched in the transition from juvenile to adult worms. Further, we describe a novel miRNA, sma-miR-4989 showing that its proximal genomic location to sma-miR-277 suggests that they form a miRNA cluster, and we propose hairpin folds for both miRNAs compatible with the miRNA pathway. In addition, we found that expression of sma-miR-277/4989 miRNAs are up-regulated in adults while their predicted targets are characterised by significant down-regulation in paired adult worms but remain largely undisturbed in immature "virgin" females. Finally, we show that sma-miR-4989 is expressed in tegumental cells located proximal to the oesophagus gland and also distributed throughout the male worms' body. Our results indicate that sma-miR-277/4989 might play a dominant role in post-transcriptional regulation during development of juvenile worms and suggest an important role in the sexual development of female schistosomes.

Hunt VL, Tsai IJ, Coghlan A, (…), Ribeiro DM, (…), Berriman M (2016) "The genomic basis of parasitism in the Strongyloides clade of nematodes" *Nature Genetics*, 48(3), pp. 299–307.

Abstract: Soil-transmitted nematodes, including the Strongyloides genus, cause one of the most prevalent neglected tropical diseases. Here we compare the genomes of four Strongyloides species, including the human pathogen Strongyloides stercoralis, and their close relatives that are facultatively parasitic (Parastrongyloides trichosuri) and free-living (Rhabditophanes sp. KR3021). A significant paralogous

expansion of key gene families--families encoding astacin-like and SCP/TAPS proteins--is associated with the evolution of parasitism in this clade. Exploiting the unique Strongyloides life cycle, we compare the transcriptomes of the parasitic and free-living stages and find that these same gene families are upregulated in the parasitic stages, underscoring their role in nematode parasitism.

Bennett HM, Mok HP, Klotsas E, (...), Ribeiro DM, (...), Berriman M (2014) "The genome of the sparganosis tapeworm Spirometra erinaceieuropaei isolated from the biopsy of a migrating brain lesion" *Genome Biology*, 15(11).

Abstract: BACKGROUND: Sparganosis is an infection with a larval Diphyllobothriidea tapeworm. From a rare cerebral case presented at a clinic in the UK, DNA was recovered from a biopsy sample and used to determine the causative species as Spirometra erinaceieuropaei through sequencing of the cox1 gene. From the same DNA, we have produced a draft genome, the first of its kind for this species, and used it to perform a comparative genomics analysis and to investigate known and potential tapeworm drug targets in this tapeworm. RESULTS: The 1.26 Gb draft genome of S. erinaceieuropaei is currently the largest reported for any flatworm. Through investigation of β-tubulin genes, we predict that S. erinaceieuropaei larvae are insensitive to the tapeworm drug albendazole. We find that many putative tapeworm drug targets are also present in S. erinaceieuropaei, allowing possible cross application of new drugs. In comparison to other sequenced tapeworm species we observe expansion of protease classes, and of Kuntiz-type protease inhibitors. Expanded gene families in this tapeworm also include those that are involved in processes that add post-translational diversity to the protein landscape, intracellular transport, transcriptional regulation and detoxification. CONCLUSIONS: The S. erinaceieuropaei genome begins to give us insight into an order of tapeworms previously uncharacterized at the genome-wide level. From a single clinical case we have begun to sketch a picture of the characteristics of these organisms. Finally, our work represents a significant technological achievement as we present a draft genome sequence of a rare tapeworm, and from a small amount of starting material.

Résumé

Au fil du temps, la vie a évolué pour produire des organismes remarquablement complexes. Pour faire face à cette complexité, les organismes ont développé une pléthore de mécanismes régulateurs. Par exemple, pour chaque ARN messager (ARNm) codant une protéine, des régions non traduites (UTR; *untranslated regions* en anglais) potentiellement régulatrices sont aussi présentes. De plus, les organismes supérieurs transcrivent des milliers d'ARN longs non codants (ARNlnc), accroissant ainsi la capacité régulatrice de leurs cellules. Cependant, la plupart des ARNlnc sont-ils fonctionnels? Le cas échéant, par quels mécanismes peuvent-ils agir? Le rôle d'échafaudage des ARNlnc, formant des ribonucléoprotéines et rapprochant ainsi physiquement les protéines est un concept émergent. Toutefois, la prévalence de ce mécanisme reste encore à déterminer.

De plus, au lieu d'ajouter de nouveaux composants pour augmenter la complexité, les cellules peuvent réutiliser certaines protéines pour exécuter plusieurs fonctions distinctes. C'est le cas des protéines *moonlighting*. Ces protéines exercent souvent des fonctions distinctes dans des environnements différents et peuvent donc être régulées par un changement de localisation cellulaire. Par la formation de complexes protéiques en cours de traduction, les régions 3' non traduites (3'UTRs) peuvent réguler la localisation cellulaire et la fonction de la protéine synthétisée à partir des transcrits auxquels elles appartiennent. Néanmoins, la fréquence ce mécanisme et son rôle dans la régulation des diverses fonctions des protéines *moonlighting* reste à aborder.

Cette thèse a pour objectif de découvrir et comprendre systématiquement deux mécanismes de régulation méconnus impliquant la partie non codante du transcriptome humain. Concrètement, l'assemblage de complexes protéiques promus par les ARNlnc et les 3'UTRs est étudié avec des données d'interactions protéines-protéines et protéines-ARN prédites et expérimentales, à grande échelle. Ceci a permis *(i)* de prédire le rôle de plusieurs centaines d'ARNlnc comme molécules d'échafaudage pour plus de la moitié des complexes protéiques connus, ainsi que *(ii)* d'inférer plus d'un millier de complexes 3'UTR-protéines, dont des cas permettant d'expliquer la localisation cellulaire de protéines *moonlighting*. Ces résultats obtenus à l'échelle du protéome et du transcriptome indiquent qu'une proportion élevée d'ARNlnc et de 3'UTRs pourrait réguler la fonction des protéines en augmentant ainsi la complexité du vivant.

Abstract

Over time, life has evolved to produce remarkably complex organisms. To cope with this complexity, organisms have evolved a plethora of regulatory mechanisms. For instance, for every messenger RNA (mRNA) encoding a protein, regulatory untranslated regions (UTRs) are also present. Additionally, higher organisms transcribe thousands of long non-coding RNAs (lncRNAs), presumably expanding the regulatory capacity of their cells. However, it is questionable whether most lncRNAs are functional, and even though many lncRNAs interact with other cellular components, is yet unclear through which mechanisms they may act. An emerging concept is that lncRNAs can serve as protein scaffolds, forming ribonucleoproteins and bringing proteins in proximity, but the prevalence of this mechanism is yet to be determined.

Besides adding new components to increase complexity, cells can reuse proteins to perform several unrelated functions. Such is the case of the moonlighting proteins. These proteins are often found to perform distinct functions under different environments, and may thus be regulated by a change of cellular localisation. Interestingly, through the formation of protein-complexes during translation, 3'UTRs have been found to regulate the cellular localisation and function of the protein synthesized from their transcript. Yet, if this mechanism is common, and if used to regulate the several functions of moonlighting proteins, remains to be addressed.

This thesis aims to systematically discover and provide insights into two ill-known regulatory mechanisms involving the non-coding portion of the human transcriptome. Concretely, the assembly of protein complexes promoted by lncRNAs and 3'UTRs is investigated using computationally predicted, as well as experimentally determined, large-scale datasets of protein-protein and protein-RNA interactions. This enabled to *(i)* predict hundreds of lncRNAs as possible scaffolding molecules for more than half of the known protein complexes, as well as *(ii)* infer more than a thousand distinct 3'UTR-protein complexes, including cases likely to regulate the cellular localisation of moonlighting proteins. These large-scale results indicate that a high proportion of lncRNAs and 3'UTRs may be employed in regulating protein function, potentially playing a role both as regulators and as components of complexity.