Predictions of RNA-binding ability and aggregation propensity of proteins

Federico Agostini

TESI DOCTORAL UPF / ANY 2014

DIRECTOR DE LA TESI

Dr. Gian Gaetano Tartaglia
Departament of Bioinformatics and Genomics at the CRG

To my parents, Augusto and Vincenza, with love and gratitute.
You have always believed in me and supported me in all of my endeavors.

Acknowledgments

I would like to thank all the people who helped me during this PhD:

my supervisor,

Gian Gaetano Tartaglia,

current and former members of the lab,

Davide Cirillo, Benedetta Bolognesi, Nieves Lorenzo, Silvia Rodríguez, Domenica Marchese, Petr Klus, Carmen Maria Livi, Maria Teresa Botta-Orfila, Joana Ribeiro Domingues, Marta Baldrighi, Andreas Zanzoni, Matteo Bellucci, Priscilla De Rosa, Marianela Masin,

friends from other labs,

Mekayla Anna Storer, Jean-François Popoff, Birgit Ritschka, Samuel Francis Reid, Alessandra Breschi, Debayan Datta,

friends that are far away but always present,

Giovanni Bussotti, Marco Mariotti, Francesco Guescini, Giuseppe Pio Masone, Valerio Coppola, Alessandro Diele,

and, most importantly, my family.

Abstract

RNA-binding proteins (RBPs) control the fate of a multitude of coding and non-coding transcripts. Formation of ribonucleoprotein (RNP) complexes fine-tunes regulation of post-transcriptional events and influences gene expression. Recently, it has been observed that non-canonical proteins with RNA-binding ability are enriched in structurally disordered and low-complexity regions that are generally involved in functional and dysfunctional associations. Therefore, it is possible that interactions with RNA protect unstructured protein domains from aberrant associations or aggregation. Nevertheless, the mechanisms that prevent protein aggregation and the role of RNA in such processes are not well understood. In this work, I will describe algorithms that I have developed to predict protein solubility and to estimate the ability of proteins and transcripts to interact. I will illustrate applications of computational methods and show how they can be integrated with high throughput approaches. The overarching goal of my work is to provide experimentalists with tools that facilitate the investigation of regulatory mechanisms controlling protein homeostasis.

Resumen

Las proteínas de unión de ARN son responsables de controlar el destino de una multitud de transcriptos codificantes y no codificantes. De hecho, la formación de complejos de ribonucleoproteínas (RNP) afina la regulación de una serie de eventos post-transcripcionales e influye en la expresión génica. Recientemente, se ha observado que las proteínas con capacidad no canónica de unión al ARN se enriquecen en las regiones estructuralmente desordenadas y de baja complejidad, que son las que participan generalmente en asociaciones funcionales y disfuncionales. Por lo tanto, es posible que interactuar con el ARN pudiera ser una manera de proteger las proteínas no estructuradas de asociaciones aberrantes o de agregación. Sin embargo, los mecanismos que impiden la agregación de proteínas y la función del ARN en tales procesos no están bien descritas. En este trabajo, se describen los métodos que he desarrollado para predecir la solubilidad de proteínas y para estimar la capacidad de transcriptos y proteínas de interactuar. De otra parte, voy a ilustrar sus aplicaciones y explicar como los métodos de bajo rendimiento han evolucionado a un mayor rendimiento. El objetivo final es proporcionar instrumentos a los investigadores experimentales que se pueden utilizar para facilitar la investigación de los mecanismos reguladores que controlan la homeostasis molecular.

Preface

In this thesis, I will report two projects in which I have been directly involved. The first project includes the investigation of physico-chemical properties that describe the propensity of proteins to interact with coding and non-coding transcripts (see Chapters I, II and III). The second project focuses on the determination of features that are relevant for the solubility of polypeptide chains (see Chapters VI and VII). In both projects I have been in charge of algorithm development and improvement, as well as application to representative experimental studies. The work presented in this thesis illustrates the flexibility of computational approaches to study molecular processes such as X-chromosome inactivation (see Chapter II) and to perform large-scale simulations (see Chapter III). Finally, I will introduce the recently developed *SeAMotE* algorithm (see Chapter IV), a method to search for protein recognition motifs in the sequence of their nucleic acid targets.

The methods and analyses described here represent small elements of a more ambitious project that aims to the identification of macromolecule features associated with formation of functional and dysfunctional protein-protein and protein-RNA complexes. In this context, the *omics* modules of *cat*RAPID (see Chapter III) and *cc*SOL (see Chapter VII) algorithms are intended to facilitate the investigation of general evolutionary principles. It is my belief that the synergistic use of computational tools will aid the investigation of protein-RNA interactions, which will lead to the discovery of novel associations. By exploring the composition and dynamics of macromolecular structures, such as RNA granules, it will be possible to elucidate their functional and dysfunctional implications in human neuropathology.

Contents

Introduction

Next-generation genomics

In the past two decades, a number of high-throughput approaches have been developed to investigate different aspects of cellular biology. In particular, the application of new technologies, such as next generation sequencing (NGS), has rapidly generated a substantial quantity of genomic, transcriptomic and proteomic data (Hawkins et al., 2010). The integration of data from various sources resulted in an exponential growth of the bioinformatic field and substantial development of computational tools. Very intriguingly, the employment of genome-wide approaches has led to a better understanding of the human genome blueprint. In particular it has provided compelling insights into the complexity and variability of each individual, hereby granting access to previously uncharted areas of cell biology.

Mounting evidence accumulated throughout the past decade has shown that a large portion of DNA, at least 75% (Djebali et al., 2012), is pervasively transcribed. With respect of the human genome, it has been proposed "that the majority of its bases are associated with at least one primary transcript" (Birney et al., 2007). Interestingly, the higher sequencing resolution achieved by means of the new techniques resulted in the discovery of a substantial number of human genes that do not code for proteins, but are nonetheless transcribed to produce small and long non-coding RNAs (ncRNAs) (Djebali et al., 2012; Harrow et al., 2012). Although these molecules

were initially considered a by-product of "junk" DNA regions, increasing evidence indicates that ncRNAs possess an active role in the regulation of several processes inside the cell. Specifically, extensive non-coding portions of the genome show a high level of conservation (Bejerano et al., 2004; Stephen et al., 2008), and the number of non-protein-coding genes seems to correlate with the developmental complexity of eukaryotic organisms (Taft et al., 2007). Collectively, both these results and the observations derived from cytosolic and nuclear RNA mapping led to the proposal of "a model of genome organization where protein-coding genes are at the center of a complex network of overlapping sense and antisense (long) RNA transcription, with interleaved (small) RNAs" (Kapranov et al., 2007).

The versatile and effective regulatory role of ncRNAs is emphasized by their ability to bind to different types of molecules in a sequence- or structure-specific manner, thereby providing specificity to complexes (Hüttenhofer and Schattner, 2006), and directly or indirectly regulating transcription of thousands of genes (Gupta et al., 2010). It is possible that pervasive transcription allows evolutionary pressure to operate its selection on a large dynamic pool of ncRNAs. Indeed, most ncRNAs are subject to less rigid structure-function constraints than protein-coding RNAs (Birney et al., 2007; Heimberg et al., 2008; Meader et al., 2010), and might function via formation of stable secondary and tertiary structures. Intriguingly, these structures can accommodate compensatory nucleotide substitutions without disrupting their functional integrity (Smit et al., 2009).

For the sake of space and consistency with the topics of this composition, the small RNA subject will not be discussed, but more information can be found in recent literature (Castel and Martienssen, 2013; Lui and Lowe, 2013; Sabin et al., 2013).

Long non-coding RNAs

In the last few years, long non-coding RNAs (lncRNAs) have received particular consideration due to their involvement in a variety of cellular processes, of which they play a key role in controlling cellular regulation (Guttman and Rinn, 2012; Guttman et al., 2011; Ørom et al., 2010; Ponting et al., 2009; Rinn and Chang, 2012; Ulitsky and Bartel, 2013; Wang and Chang, 2011). Furthermore, long non-coding molecules have been implicated in carcinogenesis, where they act as oncogenes or tumor suppressors (Gupta et al., 2010; Poliseno et al., 2010; Zhang et al., 2010), and in other complex diseases such as myocardial infarction (Ishii et al., 2006) and Alzheimer's disease (Mus et al., 2007).

LncRNAs are arbitrarily defined as molecules larger than 200 nucleotides (Furuno et al., 2006; Lyle et al., 2000) with low protein coding potential, often represented by the lack of a functional open reading frame (ORF) (Kageyama et al., 2011). It is not possible to absolutely rule out a potential dual function as suggested by the presence of unordinary lncRNAs (Chew et al., 2013; Guttman et al., 2013). The metastasis-associated lung adenocarcinoma transcript 1 (MALAT1) is one such example, which despite the lack of a poly(A) tail can be efficiently translated *in vivo* (Wilusz et al., 2012). A series of studies demonstrated that these molecules have lower expression levels with respect to mRNAs and are more likely to be highly expressed in tissue- or cell-specific patterns (Cabili et al., 2011; Derrien et al., 2012; Guttman et al., 2010; Ravasi et al., 2006). As a matter of fact, the existence of a few lncRNAs has been known since the beginning of the 90s (Brown et al., 1992; Penny et al., 1996). Only with the recent application of comprehensive genome-wide sequencing analyses (Carninci et al., 2005; Guttman et al., 2009; Ravasi et al., 2006) it was possible to establish that these were not isolated cases but belonged to a much larger class of regulatory molecules.

LncRNAs are often defined by their genomic location relative to nearby

protein-coding genes (Rinn and Chang, 2012). LncRNAs are not as highly conserved as most mRNAs and small ncRNAs, showing only a meager level of conservation in their promoters, primary sequences and splice sites (Carninci et al., 2005; Guttman et al., 2010; Marques and Ponting, 2009). A subclass, namely the large intergenic transcripts (lincRNAs), received great consideration over the past few years owing to the peculiarity of their originating sites. These sites are represented by previously un-annotated regions, and a higher degree of sequence conservation throughout evolution with respect to introns and untranscribed intergenic regions (Khalil et al., 2009; Ponjavic et al., 2007). Other regions of genome in which lncRNAs are found include promoters, enhancers, introns, untranslated regions (UTRs), overlapping or non-coding isoforms of coding genes, antisense to other gene products and pseudogenes (Carninci et al., 2005; Engström et al., 2006; Kim et al., 2010).

To date, the roles of most lncRNAs remain elusive and only few of these molecules have been well characterized and shown to be part of almost every level of gene expression (Wapinski and Chang, 2011). The ability of lncRNAs to contribute to post-transcriptional processes, such as protein synthesis, RNA maturation, transport and transcriptional gene silencing (Bernstein and Allis, 2005; Whitehead et al., 2009) is even more striking when considering the promiscuity of interactions and the number of ribonucleoprotein (RNP) complexes they can contribute to (Rinn and Chang, 2012). By participating in the formation of RNP complexes, lncRNAs can be involved in gene expression. The multitude of functions of lncRNAs has been hitherto refined into four archetypes of molecular mechanisms:

I Decoys: lncRNAs can act as decoys that preclude the access of regulatory proteins to DNA;

II Scaffold: lncRNAs can serve as adaptors to bring two or more proteins into discrete complexes (Spitale et al., 2011);

III Guides: lncRNAs can be required for the proper localization of spe-

cific protein complexes, thus combining the binding ability of a protein partner with a mechanism to selectively contact regions of the genome;

IV Enhancers: lncRNAs can interface with the chromatin-modifying machinery, resulting in enhancer-based gene activation.

These categories are endowed with flexible boundaries, as an individual lncRNA may be involved in several functions. Moreover, these classes provide a further example of how apparently complex functions can be constructed from combinatorial usage of archetypal molecular mechanisms (Wang and Chang, 2011).

Currently, a universal function of lncRNAs appears to be directing the activity of chromatin-modifying complexes and transcription factors by specifying their genomic DNA targets and activating or inhibiting their function (Guttman et al., 2011; Huarte et al., 2010; Nagano et al., 2008; Rinn et al., 2007; Zhao et al., 2008). In these contexts, lncRNAs have the ability to act as scaffolds, nucleating the assembly of larger complexes or cellular structures (Clemson et al., 2009; Shevtsov and Dundr, 2011). Nevertheless, although considerable efforts have been made to elucidate the identity of proteins that interact with lncRNAs (Guttman et al., 2009; Meyer et al., 2012), there is still much progress to be made. Perhaps a deeper understanding could be obtained through the analysis of the similarities of action mechanisms, which may eventually facilitate instructive and predictive models of lncRNA function (Wang and Chang, 2011).

Protein and RNA interactions

RBPs have the ability to form dynamic ribonucleoprotein (RNP) complexes, a critical step in the control of mRNA processing (Chen and Manley, 2009; Licatalosi and Darnell, 2010) and ncRNA function (Guttman

and Rinn, 2012; Rinn and Chang, 2012). Therefore, it is not surprising that the proper cellular functions of virtually all ncRNAs depend upon the formation of RNA-protein complexes (Eddy, 2001; Guttman and Rinn, 2012; Rinn and Chang, 2012). The advantage of forming RNP complexes to exert a specific molecular function is not limited to the class of ncRNAs; mRNAs are constantly coated and compacted by RBPs throughout their life cycle. Recent studies corroborated the hypothesis that coding transcripts are bound by multiple and heterogeneous RBPs and that individual RBPs are able to control from a few to thousands of mRNA targets (Keene, 2007; Ascano et al., 2012; Ankö and Neugebauer, 2012). These analyses provided further support to the importance of the formation of RNP complexes, suggesting that the establishment of such structures can often proceed in a highly combinatorial fashion and can affect all aspects of the life of RNA (Glisovic et al., 2008; Janga, 2012; Keene, 2007).

Given the central role of RNPs in gene expression regulatory hubs, alterations in post-transcriptional expression levels or the appearance of mutations in either RBPs or binding sites in targeted transcripts have been linked to a number of human diseases including muscular atrophies, neurological disorders and cancer (Lukong et al., 2008; Musunuru, 2003; Kim et al., 2009; Castello et al., 2013a). As a matter of fact, RNAs exiting the nucleus must be equipped with the necessary RBPs to regulate their localization, translation and decay in the cytoplasm. Hence, RBPs exert a tight control over these RNAs, facilitated by the presence of linear and structural elements onto the RNA sequence, which allows the spatial and temporal confinement of the target transcripts to specific sites within the cytoplasm and the eventual recruitment of additional co-factors. Increased competition between cellular compartments can promote exchanges between RBPs and impairment in intermolecular recognition and variations in the association stoichiometry mechanisms can easily result in the onset of pathological conditions (Kechavarzi and Janga, 2014).

Currently, the number of RBPs in the human genome is estimated to

be aroung 2000 proteins (Dieterich and Stadler, 2013). To date, ~600 RBPs are annotated in mammalian genomes as carriers of canonical RNA-binding domains (RBDs), but functions and *in vivo* binding specificities of most RBPs remains unclear (Müller-McNicoll and Neugebauer, 2013). Recently, the application of a new experimental protocol, the "interactome capture" (Castello et al., 2013b) provided further insights into the identity of proteins that bind to mRNA *in vivo* (Baltz et al., 2012; Castello et al., 2012; Kwon et al., 2013). These groups identified similar numbers of RBPs (797, 865 and 555, respectively), which is higher than ~600 canonical RBPs. Interestingly, 315 proteins from the study conducted by Castello et al. (2012), 245 proteins from Baltz et al. (2012) and 133 protein from Kwon et al. (2013) were not previously annotated as RNA binding, and collectively ~200 proteins had been inferred as RBPs by homology. Nonetheless, it has to be considered that these studies detected only the RBPs that were active in their well-defined experimental settings and, presumably, many more condition- and tissue-specific RBPs await discovery. It is important to stress that the technique utilized in these analyses was aimed to the investigation of mature messenger ribonucleoprotein particles (mRNPs), whereas RBPs bound to pre-mRNA, non-polyadenylated RNA or introns remains elusive. Nonetheless, these studies introduced the concept of novel RBP and provided evidence that mRNPs packaging represents a crucial event in gene expression. Additionally, these studies suggest that many RBPs are employed to influence the structure of mRNAs and to determine localization and fate of RNAs according to their length.

In spite of these recent insights, the extent to which protein composition of RNP changes during the lifetime of an individual RNA, and the number of different proteins that interact, still remain to be investigated. In order to answer these unsolved questions, perhaps the most urgent task in the field is to investigate the proteomic and transcriptomic content of individual RNPs. This would subsequently enable investigation into how proteomic composition is linked to RNP function and to decipher how this changes

temporally and spatially (Müller-McNicoll and Neugebauer, 2013).

RNA granules and neurodegeneration

An important consideration for RNA biology is the subcellular location of RNP complexes. As a matter of fact, localization of mRNAs enables the precise regulation of protein expression both spatially and temporally. In addition to constrains imposed by membrane boundaries within the cell, RNPs often localize by assorting into functionally distinct sub-compartments in a temporally appropriate manner (Wolozin, 2012). Hence, predicting whether remodeling of an RNP will occur after its cellular re-localization is not as simple as comparing protein-RNA binding constants, because the concentrations of both the RNA targets and competing RBPs can contribute to the outcome (Riley and Steitz, 2013).

The functions of RBPs can be broadly divided into nuclear and cytoplasmic regulatory activities. In the nucleus, RBPs orchestrate mRNA maturation, including splicing, RNA helicase activity, RNA polymerase elongation and nuclear export (Liu-Yesucevitz et al., 2011). In the cytoplasm, RBP functions encompass RNA transport, silencing, translation and degradation (Liu-Yesucevitz et al., 2011). Cytoplasmic RBPs regulate transcript activity and distribution by forming RNA aggregates or "RNA granules" that are macromolecular complexes containing RNA binding proteins and mRNA transcripts consolidated to form granules. The initial RNA granules can interact with other particles to grow into dynamic granules, which sometimes reach the size of several microns. Despite their large dimensions, however, such granules are not surrounded by a delimiting membrane, an observation that has puzzled researchers (Alberti, 2013).

Among the multitude of assemblies, RNA granules include Cajal bodies, nuclear speckles and paraspeckles (Caudron-Herger and Rippe, 2012; Mao

et al., 2011) within the nucleus, and neuronal granules, stress granules, and processing bodies (P-bodies) within the cytoplasm. RNA granules vary by molecular composition and function. For example, RNA degradation is mediated by the P-body (Thomas et al., 2011), whereas transport granules play important roles in neurons, where they move transcripts from the soma into the dendritic and axonal arbors (Krichevsky and Kosik, 2001). Most of the transcripts contained in the stress granules are translationally silent and posses distinctive features, such as the lack of the 5'-cap or the presence of internal ribosomal entry (IRE) sites (Anderson and Kedersha, 2008). Upon stress conditions, these structures facilitate the shift of protein production in the cell from more heterogeneous to protective ("housekeeping") functions.

Generation and the dynamics of stress granules are probably the most interesting processes for the pathology of neurodegenerative diseases. As a matter of fact, it has been observed that, in some cases, RNA-protein aggregates are detergent insoluble and show hydrogel-like features (Weber and Brangwynne, 2012), though the degree of insolubility is less than that observed in protein aggregates found in neurodegenerative diseases (Collier et al., 1988; Liu-Yesucevitz et al., 2011). Although the RNA granules components are of extreme interest, their characterization is technically difficult due to the singular nature of these nonmembrane-delimited structures. So far, the most promising results were obtained employing classic methods, commonly used to isolate insoluble proteins aggregates present in brain tissues of subjects with neurodegenerative diseases (Johnson et al., 2009; Liu-Yesucevitz et al., 2010). In a pair of recent publications, the McKnight group identified protein and RNA components of RNA granules that were isolated by precipitation with a small molecule (Han et al., 2012; Kato et al., 2012). Mass spectrometry revealed that an overwhelming majority of the precipitated RBPs bear repetitive motifs of low-complexity sequences (LCSs), which are intrinsically disordered. This finding is consistent with what has been discovered in other related studies, where a large

portion of RBPs contain disordered regions enriched in short repetitive amino acid motifs with unusual RBDs (Castello et al., 2012), indicating that LCSs might play a relevant role in protein-RNA associations and the subsequent formation of aggregates. The most recent reports represent a significant advance in our understanding of the complexity and subcellular organization of RNPs: the existence of RBP aggregates could explain the detection of indirectly associated mRNAs in immunoprecipitates, where the analysis does not include generation of protein-RNA covalent bonds, and some of the experimental variability.RNP complexes mediate the process of activity dependent protein synthesis, which is critical in all aspects of biology. Yet, this function has drawn a particularly strong interest to the synapses, where it controls synaptic plasticity, habituation and memory (Hoeffer and Klann, 2010).

Protein folding, misfolding and disease

Maintenance of protein solubility and hence avoidance of misfolding and aggregation are crucial requirements for proteins to perform their cellular functions (Ellis, 2001; Powers et al., 2009). Indeed, the level of abundance of specific proteins in living systems has been linked to their requirements for chaperones in order to fold successfully (Tartaglia et al., 2010) and to maintain their solubility (Tartaglia et al., 2007). Therefore, to achieve this degree of regulation cells control the behavior of proteins at two complementary levels:

I "molecular" level, in which the properties of the amino acid sequences safeguard protein solubility at the concentrations required by the cell for optimal function;

II "cellular" level, in which quality control mechanisms are in place to maintain homeostasis, with the chaperone response ensuring that any

incipiently misfolded assemblies are prevented from developing further.

Intriguingly, the presence of factors, such as RNA molecules (Schaeffer et al., 2001; Ayala et al., 2011; Zanzoni et al., 2013), capable of interacting with nascent polypeptide chains can be crucial in the initial folding stages or in situations of controlled protein misfolding, especially under conditions that can promote aggregation (Tartaglia et al., 2010).

The mechanisms of stable and beneficial protein misfolding, analogous to the biology of prions, were first examined in yeast. In this organism, the Sup35 protein was shown to misfold in response to environmental stress and to alter the synthesis of proteins in a manner that promotes survival (Serio and Lindquist, 2001). This crucial role is played by the glycine rich domains in Sup35 that mediate the misfolding and give rise to insoluble protein aggregates, much like amyloidogenic proteins that aggregate in neurodegenerative diseases (Goehler et al., 2010). The level of homology between Sup35 glycine rich domains and that of several mammalian RNA binding proteins suggests a possible role of regulated protein aggregation in the biology of RNA binding proteins. However, the aggregation processes characterizing Sup35 and RNA binding proteins differ from the conventional models of protein aggregation in that they perform distinct biological functions and are reversible.

In summary, the relationships between protein solubility, abundance and chaperone usage impose stringent conditions on the amino acid sequences of proteins (Tartaglia et al., 2010). Therefore, to understand how processes such as the formation of RNA granules take place and what role they play in human disease, it is essential to examine the behavior of RNA binding proteins throughout the biological cycle of mRNA processing. To this purpose, the use of experimental techniques and computational methods to identify mutations that can affect RBP homeostasis and, as a consequence or co-factor, the rearrangement of intermolecular interactions, would be of

great assistance in the exploration of this still uncharted area of biology.

Experimental investigation of protein-RNA associations

The experimental approaches applied in the past years to characterize the RNA-protein interaction landscape can be broadly separated into two general categories: 'Early methods' and 'High-throughput methods'.

Early methods

This category comprises a series of biochemical *in vitro* approaches, such as electrophoretic mobility shift assays (EMSA) or ultraviolet (UV) crosslinking of proteins to their target RNAs, which have been used to study RNA-protein interactions and mRNPs assembly over the past three decades. Twenty years ago, the systematic evolution of ligands by the exponential enrichment (SELEX) method was introduced to isolate high-affinity RNA aptamers from highly diverse pools of *in vitro* transcribed RNAs (Klug and Famulok, 1994). Although this method enabled high-affinity RNA recognition elements (RREs) to be identified for numerous RBPs *in vitro*, it cannot be applied *in vivo* (Ankö and Neugebauer, 2012). The yeast three-hybrid system, which was developed in 1996, enables RNA-protein interactions to be monitored *in vivo* for the first time by measuring the expression levels of reporter genes (Martin, 2012).

High-throughput methods

Diverse RNA immunoprecipitation (RIP) protocols established in the past 10 years finally permitted the identification of endogenous RNAs present in complex mRNPs (Niranjanakumari et al., 2002) and have laid the foundation for many structural and functional studies (Khalil and Rinn, 2011). RIP protocols begin with the creation of a lysate of cells or tissue that is then subjected to immunoprecipitation with an antibody directed against an RBP of interest. Formaldehyde or UV crosslinking may or may not be used to link protein-RNA complexes covalently before lysis. RIP, followed by microarray-based identification of protein-bound RNAs (RIP-chip) revealed, for example, that the targets of an RBP could be functionally related transcripts (Ankö et al., 2010; Keene et al., 2006). A major limitation of RIP-chip is the absence of crosslinking, which has been used to recover less stable RNPs, often including non-coding RNAs and other poorly expressed transcripts. Yet, transient interactions are not readily captured by this method. In analyses designed to characterize less stable RNPs, particularly those involving mRNAs, non-crosslinked RNAs and proteins reassociate upon cell lysis, yielding false positive results that do not reflect *in vivo* interactions (Mili and Steitz, 2004; Riley et al., 2012). The demonstrated reproducibility of RIP-chip experiments is ~60-75% (Khalil et al., 2009), complicating analyses and inarguably requiring many replicates, which are not always undertaken. Finally, data from RIP-chip without crosslinking represent the sum of direct and indirect interactions of a protein with RNA (Keene et al., 2006), and binding sites cannot be mapped to the nucleotide resolution.

More recently, techniques based on UV crosslinking and immunoprecipitation followed by deep sequencing (CLIP-seq) were introduced as powerful approaches to determine *in vivo* protein-RNA interactions on a global scale (Ule et al., 2003; König et al., 2010; Licatalosi et al., 2008). UV crosslinking requires direct contact between protein and RNA and does not

promote protein-protein crosslinks. UV light with a wavelength of 254 nm crosslinks the naturally photo-reactive nucleotide bases to specific amino acids, and the photo-reactive thionucleosides 4-thiouridine (4-SU) and 6-thioguanosine (6-SG) that are used in photoactivatable-ribonucleoside-enhanced CLIP (PAR-CLIP) can be crosslinked at 365 nm. PAR-CLIP and another CLIP variant, termed iCLIP, provide resolution of RNA-binding sites that is almost to the nucleotide (Hafner et al., 2010; König et al., 2010). The crosslinked amino acid covalently bound to recovered RNA constitutes a barrier for the reverse transcriptase enzyme and either causes specific nucleotide changes (PAR-CLIP) or truncation (iCLIP) of the cDNA during reverse transcription. After UV light treatment, lysates are subjected to immunoprecipitation, and stringent purification steps are used to isolate RNAs crosslinked to the protein of interest. RNA sequencing then identifies RNA regions directly bound to the RBP, whereby the background noise is very low, and a defined consensus sequence for binding can be derived (König et al., 2012; Milek et al., 2012). As demonstrated by recent analyses, some RBPs are favored by one method but not the other, and neither protocol seems to be superior (Castello et al., 2012; Kishore et al., 2011).

The different CLIP-seq approaches have been successfully used to generate transcriptome-wide RNA maps for numerous RBPs, confirming that one RBP can have few or thousands of mRNA targets. One of the important questions in RNA biology of how many RNAs one particular RBP can bind to can be elucidated with these methods (Ascano et al., 2012). It is worth mentioning that all CLIP procedures are elaborate, multistep processes that require extensive optimization and proper controls. Biases can arise from several sources. The nucleotide composition of the RNA linkers that are ligated to the precipitated RNAs or RNA fragments has been documented to affect ligation efficiency in the creation of small RNA libraries (Hafner et al., 2011). The aforementioned 254 nm and 365 nm UV crosslinking chemistries exhibit differential sequence preferences (Castello

et al., 2012). Furthermore, sequence-specific RNase over-digestion can also bias CLIP results (Kishore et al., 2011). Finally, since any procedure involving immunoprecipitation is subject to noise, replicates of each CLIP experiments are necessary to reduce significantly the background signal (Chi et al., 2009).

The recently developed "interactome capture" and protein occupancy profiling approaches also involve *in vivo* UV crosslinking, which in this case is followed by capture of polyadenylated RNA-proteins complexes via incubation with oligo(dT)-coated beads, subsequent stringent washes to eliminate all non-crosslinked proteins and finally, elution of proteins by nuclease digestion. Released proteins are then analyzed by mass spectrometry (Castello et al., 2013b). At present, the approach cannot provide the protein composition of distinct mRNPs, which are expressed at a given time or within cellular compartments; instead, all expressed mRNAs and their bound RBPs are simultaneously analyzed. Furthermore, some proteins may be indirectly associated with the selected RNA through high-affinity protein-protein interactions that are stable despite high-stringency washing (Müller-McNicoll and Neugebauer, 2013).

Computational tools for predicting protein-RNA associations

Investigation of molecular mechanisms governing protein-RNA interactions continues to improve as new structures of RNA-protein complexes are solved and the spatial architecture of interactions is analyzed. Unfortunately, the experimental determination of RNP complexes, by means of crystallography and nuclear magnetic resonance (NMR) techniques, requires an accurate optimization of the experimental conditions, which is often a slow and difficult process (Ke and Doudna, 2004; Scott and Hen-

nig, 2008). Therefore, given the scarcity of experimentally determined structures of RNP complexes, computational prediction of RNP complex structures can greatly help studying protein-RNA interactions (Puton et al., 2012). Computational methods aim at providing knowledge regarding whether a given protein binds RNA, which residues in the protein sequence are directly involved in making contacts with the RNA, which nucleotides interact with the protein and what is the eventual structure of the protein-RNA complex (Puton et al., 2012). Previous studies have contributed to the widely accepted idea that RNA-binding sites are often positively charged patches exposed to the solvent, able to bind the negatively charged RNA backbone (Stawiski et al., 2003). Hence, the majority of structure-based predictive methods exploit the distribution of charged amino acids and spatial proximity of specific key residues to infer the interaction potential. Conversely, sequence-based approaches employ the same strategy, but replace structural observations with predicted propensities, typically computed using physico-chemical scales.

Structure-based methods

The availability of protein tertiary structures can greatly facilitate the prediction of RNA-binding sites, which are typically identified by surface-exposed residues close to each other spatially, but not necessarily in sequence (Cirillo et al., 2013). Unfortunately, as of February 2014, only 1579 macromolecular complexes involving both protein and RNA components (but excluding RNA/DNA hybrids) were available in the Protein Data Bank (PDB), including 1327 solved by X-ray crystallography, 79 by nuclear magnetic resonance (NMR) spectroscopy, and 173 by other methods. Therefore, the number of proteins that can be investigated by employing the information on tertiary structures is profoundly limited. Moreover, given the difficulty in determining and modeling the tertiary structure of RNA molecules, most of the structure-based methods focus solely on the

inference of the protein RNA-binding site, thus overlooking binary interaction predictions (Liu et al., 2010; Puton et al., 2012; Walia et al., 2012; Zhao et al., 2011b).

The algorithms Struct-NB (Towfic et al., 2010), PRIP (Maetschke and Yuan, 2009), PatchFinderPlus (Shazman and Mandel-Gutfreund, 2008), SPOT (Zhao et al., 2011a) and OPRA (Pérez-Cano and Fernández-Recio, 2010) predict RNA-binding using the properties of protein surfaces. SVM and Naïve Bayes Classifiers (NBCs) trained on structural data are employed to analyze surface features. The RNABindR method combines structural information with sequence-based predictions of hydrophobicity and entropy (Terribilini et al., 2007). In general, the success of structure-based predictive methods can provide great structural detail on substrate-binding clefts, but is greatly limited by the availability of protein-RNA complexes as templates (Zhao et al., 2011a).

Sequence-based methods

There exists a wealth of low-resolution experimental data that can be analyzed to derive the RNA and protein interacting components, associate them to particular functional states, and ultimately exploit this information to predict individual structures and RBP associations. Physico-chemical properties are particularly useful to identify binding regions in protein and RNA molecules, as demonstrated by a number of algorithms, such as RNABindR (Terribilini et al., 2007) and SCRPRED (Fernandez et al., 2011), which have been trained to predict the RNA-binding propensity of proteins using primary structure information. Recent computational methods focus on the simultaneous predictions of contact regions for both protein and RNA, which are essential to capture the specificity of RBP complexes.

The *cat*RAPID algorithm (Bellucci et al., 2011) was the first method to predict protein associations with coding and non-coding transcripts using

the information contained in the primary structure (see Chapter 2) and, since its release, other methods have been developed.

Pancaldi and Bähler (2011) published an approach based on Support Vector Machine (SVM) and Random Forest (RF) classifiers to predict RBP targets in yeast. To rationalize the factors contributing to the formation of ribonucleoprotein complexes, the authors studied several features including untranslated region (UTR) properties, RNA structures, expression levels, gene ontology (GO) associations and physico-chemical features. A subset of 40 RBPs along with their experimental targets and $>$12000 interactions were used to validate the method. The findings of their analysis can be summarized as follows:

1. High nitrogen content and high isoelectric point discriminate RBPs from other proteins;
2. A significant correlation between RNA length and relative amount of Glycine, Isoleucine and Valine has been reported;
3. Proteins with high-isoelectric points tend to bind to long mRNAs containing a large number of stem-loops;
4. RBPs sharing common targets often interact with each other and bind to the mRNAs of their interaction partners, building an autoregulatory system.

To test the predictive power of their method, the authors performed a cross-validation resulting in an accuracy of 0.69, an Area Under the ROC Curve (AUROC) of 0.77 and sensitivity and specificity around 0.7. SVM performed better than RF, but only 14 out of 76 RBP targets could be well discriminated.

Muppirala et al. (2011) developed RPIseq to predict protein-RNA associations using SVM and RF approaches. In this method, RNA sequences are encoded with the normalized frequency of nucleotide tetrads (total of 256

characteristics), while protein sequences are represented using conjoint triads (total of 343 characteristics):

1. Nucleotide tetrads are 4-mer combinations of [A, C, G, U];
2. Protein triad divides the 20 amino acids into 7 classes: [A, G, V], [I, L, F, P], [Y, M, T, S] , [H, N, Q, W], [R, K], [D, E] and [C].

RPIseq training has been performed on two different datasets obtained from the Protein-RNA Interface Database (PRIDB) (Lewis et al., 2011): a larger set containing ribosomal complexes and a smaller set without ribosomal proteins-RNA associations. On both sets, RF outperforms SVM in both accuracy and true positive rate. Furthermore, both approaches demonstrate good performances on the dataset containing ribosomal information (SVM accuracy of 0.87; RF accuracy of 0.89). RPIseq has been additionally applied to predict protein interactions with non-coding RNAs downloaded from NPInter (Wu et al., 2006). When trained on the larger dataset, RF correctly predicted 80% of NPInter interactions, while SVM only 66%.

Wang et al. (2013) developed a sequence-based Naïve Bayes (NB) classifier to predict interactions between RBPs and non-coding RNAs. Three different datasets were used to validate the method: PRIDB (Lewis et al., 2011) with and without ribosomal complexes and NPInter (Wu et al., 2006). The following features were used as input:

1. RNA sequences were analyzed using a 3-mer occurrence of [A, C, G, U];
2. Four classes [D, E], [H, R, K], [C, G, N, Q, S, T, Y] and [A, F, I, L, M, P, V, W] were employed for amino acid frequencies.

In a 10-fold cross validation, NB and extended NB classifiers obtained similar results with accuracies around 0.7, specificities of 0.9 and sensitivities of 0.3-0.4 in all the datasets.

Finally, Lu et al. (2013) implemented an algorithm, lncPro, which uses an approach similar to *cat*RAPID, but with a Fisher's linear discriminant

training of the interaction matrix, and shows comparable performance on RNase P and MRP complexes as well as the HOTAIR network.

Approaches to predict protein solubility

Protein solubility is a thermodynamic property that depends on intrinsic characteristics of the polypeptide chain as well as environmental conditions such as temperature, pH and ionic strength (Tartaglia et al., 2005). As the most insoluble regions of protein sequences are secluded from the solvent through the folding process, the solubility of proteins is strongly dependent on the stability of their native states (Tartaglia and Vendruscolo, 2008).

Wilkinson and Harrison (1991) used a dataset of 81 proteins to rationalize the solubility of proteins in *Escherichia coli* in terms of chemical properties, including charge, propensity for forming turns, hydrophilicity and length of the sequence:

$$solubility = \alpha \left[\frac{N + G + P + S}{n} \right] + \beta \left[\frac{(R + K) - (D + E)}{n} \right] + \gamma \quad (1)$$

In this formula, *n* is the number of amino acids in the protein, α and β are parameters coupled, respectively, to the propensity to form turns and the electrostatic charge (N, G, P, and S are the number of Asn, Gly, Pro, and Ser residues and R, K, D, and E the number of Arg, Lys, Asp, and Glu residues, respectively) and γ is a constant. After this initial work, several other studies have increased our understanding of the relationship between the chemical properties of amino acid sequences and their solubility (Goh et al., 2004; Chiti et al., 1999, 2003).

Tartaglia et al. (2004) introduced an equation to predict the effect of amino acid mutation on aggregation rates v_{mut}/v_{wt} without the use of fitting parameters. The original purpose of the study was to investigate naturally

occurring mutations involved in amyloid disorders such Parkinson's and Alzheimer's diseases (Bolognesi and Tartaglia, 2013).

$$v_{mut}/v_{wt} = \phi_h \phi_\beta \phi_\alpha \phi_c \tag{2}$$

In the formula, the factor ϕ_h captures most of the apolar and polar interactions. An amino acid is called *p* if its side chain carries a charge or dipole; otherwise it is called *a*. For mutations that involve the same type of amino acid $a \rightarrow a$ or $p \rightarrow p$

$$\phi_h = \begin{cases} ASA^a_{mut}/ASA^a_{wt} & a \rightarrow a \\ ASA^p_{wt}/ASA^p_{mut} & p \rightarrow p \end{cases} \tag{3}$$

where ASA^a and ASA^p are the apolar and polar water accessible surface areas of the amino acid chains (Tartaglia et al., 2004). For mutations that involve different types of amino acids ($a \rightarrow p$ or $p \rightarrow a$), we used:

$$\phi_h = \begin{cases} 1/D_{wt} & a \rightarrow a \\ D_{wt} & p \rightarrow p \end{cases} \tag{4}$$

where *D* is the magnitude of the dipole of the amino acid side chains. The factor ϕ_β is related to the ratio of β-propensity:

$$\phi_\beta = \frac{\beta_{mut}}{\beta_{wt}} \tag{5}$$

Functions ϕ_α and ϕ_c take into account the contribution of aromatic residues *A* and total charge *C*:

$$\phi_\alpha \phi_c = e^{\Delta A - \frac{\Delta |c|}{2}} \tag{6}$$

The high accuracy obtained with these simple mathematical formulas (correlation with experimental data >0.85) motivated the development of other sequence-based methods (Fernandez-Escamilla et al., 2004; Conchillo-Solé et al., 2007; Tsolis et al., 2013). Very importantly, this work indicated

that mutations leading to increased aggregation propensity result in severe cell impairment and decrease in organism longevity (Luheshi et al., 2007; Murakami et al., 2012).

The early algorithms for prediction of protein aggregation and solubility were trained on 100 proteins or less. Although accurate (Tartaglia and Caflisch, 2007; Tartaglia and Vendruscolo, 2010), these methods are not built to perform proteome-wide predictions. To have an algorithm for large-scale predictions, one should validate the method against a high number of solubility data. Therefore, to achieve a "omic" descriptor of protein solubility, we took advantage of a study in which the solubility of 70% of *Escherichia coli* proteins was experimentally measured using an *in vitro* translation system (Niwa et al., 2009).

In 2012, I introduced the *cc*SOL method (Chapter VI) to predict protein solubility using only 5 physico-chemical properties: coil/disorder (Deléage and Roux, 1987), hydrophobicity (Engelman et al., 1986), hydrophilicity (Hopp and Woods, 1981), β-turn (Levitt, 1978) and α-helix (Deléage and Roux, 1987). To identify the 5 features, we divided the original database (Niwa et al., 2009) into 2 subsets containing the most soluble (1081 entries, "head set") and least soluble (1078 entries, "tail set") proteins and calculated the discriminative power of 28 physicochemical properties collected through a literature search.

Other methods have been also developed to predict protein solubility using amino acid sequences (Smialowski et al., 2012; Magnan et al., 2009). To build PROSO II (Smialowski et al., 2012), Frishman and colleagues studied occurrence of monopeptides (i.e., frequencies of 20 residues) and dipeptides (400 residues). The training set was build using the pepcDB database (Berman et al., 2009), which stores target and protocol information (i.e., soluble expression) contributed by Protein Structure Initiative centers, while *cc*SOL employs measured soluble fractions of endogenous *E. coli* proteins (Niwa et al., 2009). Hence, two main differences between

*cc*SOL and PROSO II are the variables and training sets employed for validation. Yet, both *cc*SOL and PROSO II perform highly accurate predictions ($>$75%).

Chapter I

The *cat*RAPID algorithm

A major focus of my doctoral thesis has been the improvement of the first computational tool to predict the propensity of protein and RNA to interact. Indeed, I participated and subsequently took the lead of the *cat*RAPID approach development and improvement. The idea behind this method is to exploit physico-chemical properties contained in the primary structure of the input molecules to generate propensity profiles, which are then combined together to produce an interaction score. The algorithm was trained with ribonucleoprotein complexes derived from the PDB and the performances were evaluated using datasets of proteins and RNAs obtained from different sources (e.g. UniProtKB, NPInter, etc.). The results were surprising, showing that our method was able to infer most of the known protein and RNA association and highlighting its ability to precisely infer the binding region on both molecules.

Bellucci, M., Agostini, F., Masin, M., and Tartaglia, G. G. (2011). Predicting protein associations with long noncoding RNAs. *Nature Methods*, 8(6):444–445.
PMID: 21623348
DOI: 10.1038/nmeth.1611

Chapter II

*cat*RAPID and XIST lncRNA

Due to the conformational space of nucleotide chains, prediction of RNA secondary structures is difficult when RNA sequences are larger than few hundreds nucleotides and simulations cannot be completed on standard processors. To overcome this limitation, we implemented in *cat*RAPID *fragments* a procedure that involves division of protein and RNA sequences into fragments followed by prediction of their interaction propensities. With the *uniform* option, protein and RNA sequences are divided into overlapping segments, which is particularly useful to identify the regions involved in the binding (Cirillo et al., 2012). With the *long RNA* fragmentation, the RNALfold algorithm (Hofacker, 2003) is employed to predict the most stable secondary structures in the range 100-200 nt.

Agostini, F., Cirillo, D., Bolognesi, B., and Tartaglia, G. G. (2013). X-inactivation: quantitative predictions of protein interactions in the xist network. *Nucleic Acids Research*, 41(1):e31. PMID: 23093590
DOI: 10.1093/nar/gks968

Published online 22 October 2012

Nucleic Acids Research, 2013, Vol. 41, No. 1 ***e31***
doi:10.1093/nar/gks968

X-inactivation: quantitative predictions of protein interactions in the *Xist* network

Federico Agostini[1,2], Davide Cirillo[1,2], Benedetta Bolognesi[1,2] and Gian Gaetano Tartaglia[1,2,*]

[1]Centre for Genomic Regulation (CRG), Dr. Aiguader 88, 08003 Barcelona and [2]Universidat Pompeu Fabra (UPF), 08003 Barcelona, Spain

Received May 30, 2012; Revised September 21, 2012; Accepted September 25, 2012

ABSTRACT

The transcriptional silencing of one of the female X-chromosomes is a finely regulated process that requires accumulation in *cis* of the long non-coding RNA X-inactive-specific transcript (*Xist*) followed by a series of epigenetic modifications. Little is known about the molecular machinery regulating initiation and maintenance of chromosomal silencing. Here, we introduce a new version of our algorithm catRAPID to investigate *Xist* associations with a number of proteins involved in epigenetic regulation, nuclear scaffolding, transcription and splicing processes. Our method correctly identifies binding regions and affinities of protein interactions, providing a powerful theoretical framework for the study of X-chromosome inactivation and other events mediated by ribonucleoprotein associations.

INTRODUCTION

X-chromosome inactivation (XCI) is a highly regulated process that involves the transcriptional silencing of one of the female X-chromosomes (1). The silencing process is mainly attributable to the long non-coding RNA X-inactive-specific transcript (*Xist*) transcribed from the *Xist* gene located on the XCI inactivation centre (1). *Xist*-mediated X-inactivation involves two distinct phases: initiation and maintenance. First, *Xist* transcript coats *in cis* the entire X-chromosome triggering transcriptional silencing (2). Subsequently, stabilization of the repressed state is facilitated by a number of epigenetic processes, such as DNA methylation and chromatin modifications mediated by the Polycomb group (PcG) proteins (3). Notably, *Xist* is regulated in *cis* by its antisense partner *Tsix* (4), which also interacts with PcG proteins (5).

Using an inducible expression system in mouse embryonic stem cells, Wutz *et al.* (6) identified a number of *Xist* domains associated with chromatin localization. Interestingly, these domains do not contain sequence or structural motifs and could be low-affinity protein-binding sites (6). In contrast to the poorly defined sequence properties associated with RNA localization, the 5′-repeat region A (RepA) represents a structured domain involved in X-chromosome silencing (6). Secondary structure predictions indicate that RepA folds in two stem loops of ~200 nt containing a number of repeats (6,7).

To date, the precise mechanisms underlying localization and confinement of *Xist* onto the X-chromosome as well as the molecular details of the silencing process remain poorly understood. Recent experiments suggest that: (i) alternative splicing factor SFRS1 regulates *Xist* processing (8); (ii) transcriptional repressor Ying and Yang (YY1) tethers *Xist* onto the X-chromosome (9); (iii) the RNA-binding domains of scaffold attachment factor SAF-A bind to *Xist*-inducing chromatin reorganization (10) and (iv) the special AT-rich sequence-binding protein SATB1 co-localizes with *Xist* in the nucleus (11). Yet, due to the limited amount of experimental evidence, the challenge of identifying protein–RNA interactions associated with XCI still stands (11).

Here, we use our theoretical framework, catRAPID, to investigate *Xist* interactions with a number of epigenetic modifiers as well as transcription and splicing factors (12). Our approach exploits physicochemical properties of nucleotide and amino acid chains such as secondary structure, hydrogen bonding and van der Waals' propensities to predict protein–RNA associations with a confidence of 78% or higher (12). In the original implementation of the method, we calculated interactions with transcripts <3 kb ('Materials and Methods' section) (12). In order to investigate *Xist*, which is 16–19 kb long and represents the largest non-coding transcript with known function, we developed an extension of the algorithm. In addition to the fine calculation of protein–RNA interactions (interaction propensity), we present here an algorithm to estimate the specificity of associations (interaction strength) and a method to identify binding regions in transcripts (interaction fragments). These new developments are introduced to facilitate the characterization of protein interactions with long non-coding RNA and guide future experimental design. Notably, the new versions of the method do not require introduction of fitting

*To whom correspondence should be addressed. Tel: +34 93 316 01 16; Fax: +34 93 396 99 83; Email: gian.tartaglia@cgr.es

parameters and represent a conceptual and methodological advance to study ribonucleoprotein associations. A new version of our web servers is released at http://tartaglialab.crg.cat/.

MATERIALS AND METHODS

Interaction propensity

We use the catRAPID method to predict protein–RNA interactions (12). In catRAPID, the contributions of secondary structure, hydrogen bonding and van der Waals' are combined together into the 'interaction profile':

$$|\Phi_x\rangle = \alpha_S|S_x\rangle + \alpha_H|H_x\rangle + \alpha_W|W_x\rangle \quad (1)$$

In Equation (1), $|Y\rangle$ indicates the physicochemical profile of a property Y calculated for each amino acid (nucleotide) starting from the N-terminus (5′). For example, the hydrogen bonding profile, denoted by $|H\rangle$, is the hydrogen bonding ability of each amino acid (nucleotide) in the sequence:

$$|H\rangle = H_1, H_2, ..., H_L \quad (2)$$

Similarly, $|S\rangle$ represents the secondary structure occupancy profile and $|W\rangle$ the van der Waals' profile. The variable x indicates RNA ($x = r$) or protein ($x = p$) profiles. Secondary structure, hydrogen bonding and van der Waals contributions are calculated as described in the original articles (12). In particular, the RNA secondary structure is predicted from sequence using the Vienna package including the algorithms RNAfold, RNAsubopt and RNAplot (13). Model structures, ranked by energy, are used as input for *cat*RAPID. For each model structure, the RNAplot algorithm is used to generate secondary structure coordinates. Using the coordinates, we define the 'secondary structure occupancy' by counting the number of contacts within the nucleotide chain. High values of secondary structure occupancy indicate that base pairing occurs in regions with high propensity to form stems, while low values are associated with junctions or multi-loops.

We use discrete Fourier transform to compare interaction profiles of different length:

$$\Psi_{k,x} = \sqrt{\frac{2}{\text{length}}} \sum_{n=0}^{\text{length}} \Phi_{n,x} \cos\left[\frac{\pi}{\text{length}}\left(n+\frac{1}{2}\right)\left(k+\frac{1}{2}\right)\right] \quad k = 0,1,...\ell \quad (3)$$

where the number of coefficients is $\ell = 50$.

The 'interaction propensity' π is defined as the inner product between the protein propensity profile $|\Psi_p\rangle$ and the RNA propensity profile $|\Psi_r\rangle$ weighted by the 'interaction matrix' I:

$$\pi = \langle \Psi_p | I | \Psi_r \rangle \quad (4)$$

To calculate the interaction propensity π, we exploit that the squared norm of π is conserved under Fourier transform:

$$\sum_{i,j}^{\text{length}_p,\text{length}_r} |\langle \ _p|I|\psi_r\rangle|^2 \approx \sum_{i,j}^{\ell_r,\ell_p} |\langle \ _p|I|\psi_r\rangle|^2 \quad (5)$$

The interaction matrix I as well as the parameters α_S, α_H and α_W are derived under the condition that interaction propensities π take maximal values for associations present in the positive training set (and minimal values for those in the negative training set):

$$I: \begin{cases} \max \langle \Psi_p | I | \Psi_r \rangle \ \forall\{r,p\} \in \{\text{positive training set}\} \\ \min \langle \Psi_p | I | \Psi_r \rangle \ \forall\{r,p\} \in \{\text{negative training set}\} \end{cases} \quad (6)$$

In the training and test phases, we used protein and RNA sequences in the range of 50–750 amino acids and 50–3000 nt, respectively (12). We note that prediction of RNA secondary structures results in intense CPU usage when sequences are >1500 nt and simulations cannot be completed on standard processors (2.5 GHz; 4–8 GB memory).

The server to compute the interaction propensity with respect to the negative training set (discriminative power) is available at: http://tartaglialab.crg.cat/catrapid.html.

Interaction strength

Computational models indicate that RNA sequence length and secondary structure free energies are correlated (Supplementary Figure S1a) (14). Hence, one would expect that long RNAs are more stable and prone to bind to proteins than short RNAs (see also section 'interaction fragments'). Indeed, we observe a weak correlation between secondary structure energy and protein–RNA interaction propensity in our algorithm (Pearson's correlation = 20%; $P = 0.07$) (Supplementary Figure S2b). Nevertheless, as no experimental evidence indicates that long transcripts interact more than small RNAs, we eliminated the length dependence introducing a 'reference set' composed by protein and RNA sequences that have exactly the same lengths as the molecules under investigation. In our calculations, we use random associations between polypeptide and nucleotide sequences. Since little interaction propensities are expected from random associations, the reference set represents a 'negative control'.

For each protein–RNA pair under investigation, we use a reference set of 10^2 protein and 10^2 RNA molecules (the number of sequences is chosen to guarantee sufficient statistical sampling). To assess the strength of a particular association, we compute the interaction propensity π and compare it with the interaction propensities $\tilde{\pi}$ of the reference set (total of 10^4 protein–RNA pairs). Using the interaction propensity distribution of the reference set, we generate the 'interaction score':

$$\text{Interaction score} = \frac{\pi - \mu}{\sigma} \quad \begin{cases} \mu = \frac{1}{\Lambda}\sum_{i=1}^{\Lambda} \tilde{\pi}_i \\ \sigma^2 = \frac{1}{\Lambda}\sum_{i=1}^{\Lambda} (\tilde{\pi}_i - \mu)^2 \end{cases} \quad (7)$$

The number of interactions is $\Lambda = 10^4$. From the distribution of interaction propensities, we compute the 'interaction strength':

$$\text{Interaction strength} = P(\tilde{\pi} \leq \pi) = \text{cumulative distribution function (cdf)} \quad (8)$$

Reference sequences have the same lengths as the pair of interest to guarantee that the interaction strength is independent of protein and RNA lengths. The interaction strength ranges from 0 (non-interacting) to 100% (interacting). Interaction strengths >50% indicate propensity to bind. The 'RNA interaction strength' and the 'protein interaction strength' are special cases of the interaction strength in which only a reference set is generated using RNA or protein sequences. The RNA interaction strengths used for the analysis of RepA, 4R and 2R represent the RNA-binding abilities of SUZ12 and EZH2 with respect to the polynucleotide reference set (Figure 1). Similarly, the protein interaction strengths used for SFRS1, SAF-A and SATB1 are the protein-binding abilities of the experimental RNA fragments with respect to the polypeptide reference set (Figures 2 and 4). The interaction strength is also used to compare YY1- and green fluorescent protein (GFP)-binding propensities (proteins are RNA fragments are of different lengths; Figure 3). It should be noted that in the case of *Xist* fragment BC (nt 1898–4940), the RNA sequence is >3 kb. In order to calculate the abilities of fragment BC to interact with YY1 and GFP, we analyzed all *Xist* fragments of size 1500 nt contained in the region 1898–4940 nt, computed the corresponding interaction strengths and averaged the scores.

The server to compute the interaction strength is available at: http://tartaglialab.crg.cat/catrapid.strength.html.

Interaction fragments

The use of RNA fragments is introduced to identify RNA regions involved in protein binding. The RNALfold algorithm from the Vienna package (www.tbi.univie.ac.at/RNA/) is used to select RNA fragments in the range of 100–200 nt with predicted stable secondary structure. Secondary structure stabilities are estimated by calculating the RNA free energy predicted by RNALfold (15). As long RNA segments have lower free energy for the higher number of bases that can be paired (Supplementary Figure S1a) (14), the choice of segments in the range of 100–200 nt is optimal because it allows simultaneously: (i) selection of secondary structures with comparable free energy (Supplementary Figure S1b) and (ii) high sequence coverage (>90%) for long transcripts such as *Xist* (Supplementary Figure S1c). Once the RNA fragments are selected, catRAPID is used to predict their ability to bind to polypeptide chains. Conceptually, the interaction fragments algorithm is a variant of the RNA interaction strength algorithm that allows identification of putative binding areas in long sequences. If the exact protein and/or RNA domains are known, we recommend the use of the interaction strength method to predict the binding specificity (Figure 3).

The server to compute fragment interactions is available at: http://tartaglialab.crg.cat/catrapid.fragments.html.

RESULTS

Xist-mediated X-chromosome silencing implies a complex network of macromolecular associations orchestrated by epigenetic modifiers as well as splicing and transcription factors. *Xist* function at the initiation of X-inactivation has been extensively studied in mouse embryonic stem cells. The mouse system is more accessible to experimental investigation than the human one and is here investigated

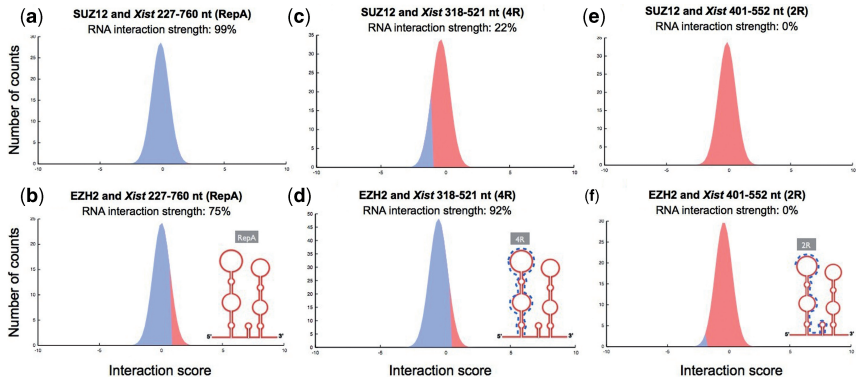


Figure 1. *Xist* RepA, 4R, 2R and PcG proteins. We predict that *Xist* RepA (227–760 nt) binds strongly to (**a**) SUZ12 (RNA interaction strength = 99%), and (**b**) EZH2 (RNA interaction strength = 75%), in agreement with experimental evidence; (**c**) SUZ12 does not bind to repeat 4R (318–521 nt; RNA interaction strength = 22%), while (**d**) EZH2 shows high interaction propensity (RNA interaction strength = 92%). Neither (**e**) SUZ12 nor (**f**) EZH2 are in contact with repeat 2R (401–552 nt; RNA interaction strengths = 0; Supplementary Table S1c) (7). Insets (b, d and f) are secondary structures of RepA (red line), 4R and 2R (blue dots) proposed by Maenner *et al.* (7).

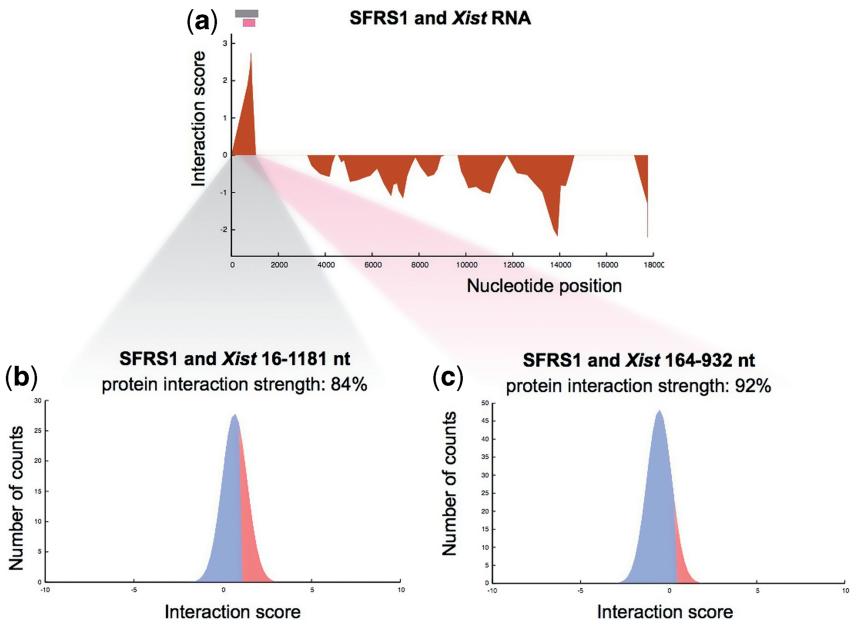


Figure 2. *Xist* and alternative splicing factor SFRS1. The interaction fragments algorithm is used to predict *Xist* ability to interact with SFRS1. (**a**) SFRS1 shows high propensity to contact *Xist* 5′. (**b**) The region studied by Royce-Tolland *et al.* (8) is marked in grey (nt 16–1181). In agreement with experimental evidence, strong interaction propensity is predicted between SFRS1 and nt 16–1181 (protein interaction strength = 84%); (**c**) nucleotides 164–932 nt (marked in red) correspond to an RNA region whose deletion abolishes *Xist* splicing (8). Strong interaction propensity is predicted between SFRS1 and nt 164–930 (protein interaction strength = 92%), as previously reported (8).

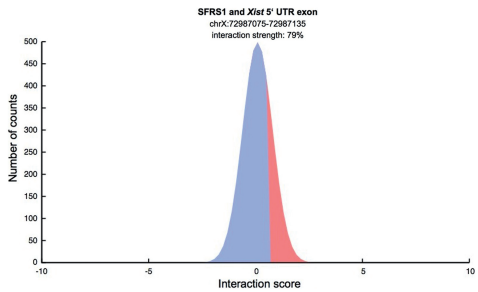


Figure 3. SFRS1 and *Xist* 5′-UTR. We predict that SFRS1 interacts with the 5′-UTR exon region of *Xist*, in agreement with CLIP-seq experiments (18).

using two novel algorithms: interaction strength and interaction fragments.

SUZ12 and EZH2 bind to RepA

The Polycomb repressive complex 2 (PRC2) is one of the two classes of PcG proteins and plays a major role in the epigenetic silencing of X-chromosome (7). More specifically, PRC2 is associated with histone modifications promoting tri-methylation of histone H3 lysine 27 along the X-chromosome, which is thought to generate a repressive compartment for silencing (16). In agreement with experimental evidence, we predict that *Xist* Repeat A region (RepA) interacts with PRC2 (7). More specifically, we find that Suppressor of Zeste 12 (SUZ12) protein homolog and Enhancer of Zeste homolog 2 (EZH2) have

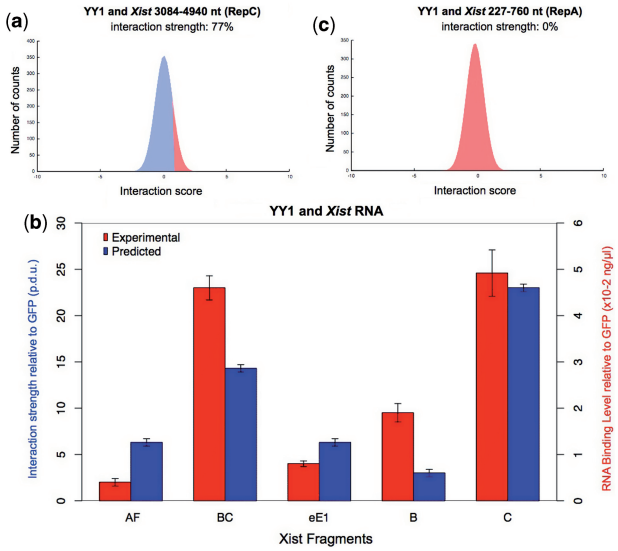


Figure 4. *Xist* and transcriptional repressor Ying and Yang (YY1). The interaction strength algorithm is used to predict YY1 ability to interact with *Xist*. (**a**) High interaction propensity is found between YY1 and *Xist* Repeat C region (RepC; interaction strength = 77%). (**b**) No interaction is predicted between YY1 and RepA (interaction strength = 0%), as previously reported (9). (**c**) Experimental binding levels of AF, B, C, BC and eE1 fragments (red bars) are reproduced by catRAPID (blue bars) with high accuracy (Pearson's correlation = 92%; $P = 0.04$ estimated with analysis of variance, two-tailed *t*-test (9) (Supplementary Table S1b). Interactions strengths and RNA-binding levels are normalized subtracting GFP signals (Supplementary Figure S2b). Errors on catRAPID predictions are evaluated using the second derivative of the cumulative distribution function associated with the interaction strength.

strong propensities to bind to RepA (region 227–760 nt; RNA interaction strengths >75%; Figures 1a and d and 5; 'Material and Methods' section). Hence, our results clearly indicate that *Xist* is able to contact PRC2 without mediation of other molecules (7).

Based on secondary structure predictions, it has been proposed that RepA contains two long stem-loop structures of ~200 nt, each containing four repeats (6,7). Nuclear magnetic resonance studies have given indication that the second loop has higher propensity to pair (17). This pairing propensity can lead to multiple interactions and complex folding. Such folding was indeed observed by structural probing of RepA and a large set of interactions have been observed with no direct evolutionary conservation or consistency with known mutations (7).

By using chemical and enzymatic probes as well as Förster resonance energy transfer experiments, EZH2 was shown to bind to RepA and repeat 4R located at position 318–521 nt within RepA (7) (Figure 5 and Supplementary Table S1a and b). By contrast, SUZ12 was found to interact with RepA and not 4R (7). Our predictions show that both EZH2 and SUZ12 contact RepA (RNA interaction strengths >75%) and that EZH2 binds to 4R (RNA interaction strength = 92%), whereas SUZ12 shows much lower binding propensity (RNA interaction strength = 22%). Moreover, we predict that neither EZH2 nor SUZ12 is able to interact with region 2R; (nt 401–552; RNA interactions strengths = 0%), as previously demonstrated by immunoprecipitation assays and western blot analysis (7) (Supplementary Table S1). In agreement with experimental evidence, we also predict that EZH2 binds to the reverse complement of RepA present in *Tsix* (2073–2239 nt; Supplementary Figure S2a) (5).

SFRS1 associates with RepA

Stochastic differences in *Xist* RNA levels influence the production of spliced RNA in the two X-chromosomes, thus leading to inactivation of one chromosome upon differentiation (8). Using HeLa cell nuclear extracts and ultraviolet cross-linking, Royce-Tolland *et al.* (8) showed that the splicing factor SFRS1 is able to associate with RepA. Here, we use the interaction fragments algorithm to predict the ability of SFRS1 to interact with *Xist*. In our analysis, the interaction propensities are calculated using RNA fragments with predicted stable secondary structure ('Materials and Methods' section). In agreement with *in vitro* and *in vivo* experiments (8), we find that SFRS1 interacts with RepA (nt 682–881, 707–826 and 726–907; Supplementary Table S1a). In particular, we predict that SFRS1 has strong propensity to bind to the domain investigated by Royce-Tolland *et al.* (nt 16–1181; protein interaction strength = 84%; Figure 2b; 'Material and Methods' section) and with a

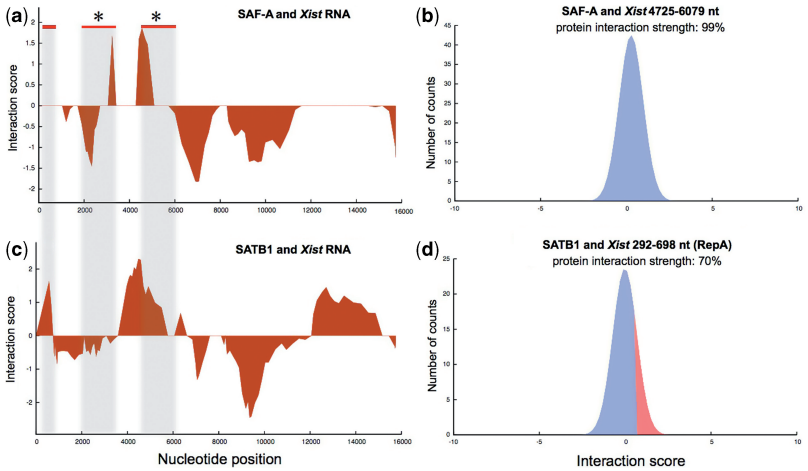


Figure 5. *Xist*, scaffold attachment factor SAF-A and special AT-rich sequence-binding protein SATB1. (**a**) In agreement with experimental evidence, SAF-A is predicted to contact *Xist* in more than one region (10). Red lines and grey boxes indicate experimentally validated regions involved in *Xist* localization (6). Stars mark primers of elements studied by Hasegawa *et al.* (10). (**b**) SAF-A shows strong propensity to bind to *Xist* region 4934–5056 nt (protein interaction strength = 99%). (**c**) Multiple binding sites are predicted between *Xist* and SATB1. (**d**) We predict that SATB1 binds strongly to nt 292–698 (RepA; protein interaction propensity = 70%), as previously suggested (6,11).

fragment whose deletion abrogates *Xist* splicing (nt 164–932; protein interaction strength = 92%; Figure 2c); (8). Thus, our results indicate that SFRS1 is directly recruited for selective inactivation of the X-chromosome (8).

Recently, Sanford *et al.* (18) used cross-linking immunoprecipitation coupled with high-throughput sequencing (CLIP-seq) to characterize SFRS1's interactome. Using HEK293T cells, the authors gathered a large amount of information on the RNA-binding sites targeted by SFRS1. In particular, CLIP-seq experiments indicate that SFRS1 binds to the 5′-UTR exon region of *Xist* (coordinates chrX:72987075–72987135 in the Human Genome Assembly 18) (18). In agreement with this finding, we predict high interaction propensity between SFRS1 and the 5′-UTR exon region (interaction strength: 79%; Figure 3).

We take the opportunity offered by CLIP-seq experiments to assess catRAPID's ability to predict SFSR1's interactions. In our analysis, we use RNA regions containing the highest number of CLIP-seq-binding sites (i.e. CLIP-seq 'clusters'). Using the interaction strength algorithm, we predict that 78 out of 100 large (>50 nt) clusters bind to SFSR1 with average interaction strength of 69% (Supplementary Figure S3a), which indicates strong agreement between observed and predicted interactions. Based on the analysis of SFRS1 CLIP-seq experiments, Wang *et al.* (19) developed the 'RNAMotifModeler' algorithm to predict RNA-binding sites using sequence features and secondary structures. RNAMotifModeler identifies binding motifs in 72 out of 100 large clusters (motifs AGAAGA, AAGAAG and GAAGAA; Supplementary Figure S3a), which is fully compatible with catRAPID's performances. We also analyse the interaction propensity of 100 small (<50 nt) clusters and their corresponding upstream and downstream regions (Supplementary Figure S3b). High interaction propensities are observed for regions containing SFSR1-binding sites (interactions predicted by catRAPID: 76; RNAMotifModeler motifs: 25; Supplementary Figure S3b), while lower interaction strengths and fewer binding motifs are predicted in the flanking regions (interactions predicted by catRAPID: 30; RNAMotifModeler motifs: 10; Supplementary Figure S3b).

YY1 contacts RepC

To study *Xist* RNA localization onto the X-chromosome, Jeon and Lee (9) introduced a doxycycline-inducible *Xist* transgene into female mouse embryonic fibroblasts. Multiple independent clones showed that *Xist* transgenes act on endogenous locus *in trans* and squelch *Xist* RNA clouds on the inactive X (9). The authors reported that RepA elimination does not abolish *Xist* RNA clouds squelching, which indicates that the region is not required for X-chromosome localization (9). By contrast, knocking down of transcriptional repressor YY1 can be correlated with 70% loss of *Xist* clouds. Importantly, pull-down assays showed that *Xist* RNA repeat C, a conserved C-rich element repeated 14 times in tandem (RepC; 3084–4940 nt; Figure 6), has a pronounced ability to bind to YY1 with respect to GFP.

Using the interaction strength approach, we are able to recapitulate all the *in vitro* assays performed by Jeon and Lee to probe YY1 affinity for *Xist* fragments (9).

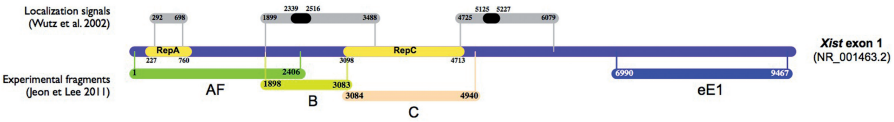


Figure 6. *Xist* first exon. RepA and RepC (yellow lines) encompass nt 227–760 and 3098–4713 (8,15). YY1 interactions investigated by Jeon and Lee (9) correspond to nt 1–2406 (AF), 1898–3083 (B), 3084–4940 (C) and 6990–9467 (eE1). The localization signals identified by Wutz *et al.* (6) are indicated by grey lines at nt 292–698, 1899–3488 and 4725–6079. The primers used by Hasegawa *et al.* correspond to nt 2339–2515 and 5125–5227 (10).

According to our calculations, *Xist* RepC shows very high propensity to interact with YY1 (Figure 4a and b), followed by one region containing an overlap between RepC and Repeat B region (RepB; Figure 4b).

In striking agreement with experimental evidence, we predict that YY1 interacts with *Xist* through RepC and RepB (Figure 4c; Pearson's correlation = 92%; $P = 0.04$) and does not associate directly with RepA (9,20).

SAF-A interacts with *Xist* 5′

Xist chromosomal localization is regulated by *cis*-elements in the 5′-half of the transcript located at nt 292–698 (RepA), 1899–3488 and 4725–6079 (Supplementary Table S1b) (6). Recently, the nuclear scaffold protein SAF-A has been linked with *Xist* localization (21). SAF-A contains three conserved domains: a SAF-box (22) binding to AT-rich DNA regions (23), Spla and Ryanodine receptor (SPRY) domain of unknown function (24) and an arginine–glycine glycine (RGG) RNA-binding domain. Deletion of the RGG-binding domain strongly reduces *Xist* chromosomal localization, suggesting direct interaction with *Xist* (10).

Using co-immunoprecipitation assays, Hasegawa *et al.* (10) reported that SAF-A contacts nt 1899–3488 and 4725–6079 (Figure 6). Employing the interaction fragments method, we find that these regions are highly prone to interact with SAF-A (Figure 5a and Supplementary Table S1a). In particular, we predict that nt 4725–6079 have strong propensity to bind to SAF-A (protein interaction strength = 77%; Figure 4b). In our analysis, we used a protein region spanning residues 50–800, which contain the uncharacterized SPRY region and the RNA-binding domain RGG (Supplementary Table S1c). By sliding a window of 750 amino acids from the N- to the C-terminus of SAF-A, we observe that the interaction fragments profiles correlate significantly (mean Pearson's correlation = 90%; $P = 0.01$; Supplementary Figure S4a). Intriguingly, when the SAF-box is included in the analysis (residues 9–759), we predict an increased ability to bind to RepA (Supplementary Figure S4b). The binding region present in RepA (Supplementary Figure S4b and Table S1b) was not investigated by Hasegawa *et al.* (10), but is consistent with the observations made by Wutz *et al.* (6) and the fact that deletion of SAF-box abolishes *Xist* chromosomal localization (10).

In agreement with experimental data, we expect that direct interaction between *Xist* and SAF-A could have an effect on *Xist* localization in the nuclear matrix, thus facilitating association with chromosomal DNA (6,10).

Does SATB1 binds to multiple *Xist* sites?

In a thymic lymphoma model, the nuclear protein SATB1 was identified as a critical component for gene silencing (25). In fact, it has been shown that viral expression of SATB1 in fibroblasts—in which *Xist* does not induce gene repression—could establish *Xist* silencing (3,25). As SATB1 co-localizes with *Xist* at the initiation of X-inactivation (25), it has been proposed that it could act as an anchor promoting RepA-mediated chromosomal reorganization (26). Nevertheless, it should be noted that SATB1 binds and regulates chromatin domains containing genes, whereas *Xist* overlaps chromosomal regions that are enriched for genomic repeats and deprived of genes. This aspect could lead to the idea that SATB1 makes genes susceptible to *Xist* by positioning gene-rich chromatin, without direct interaction (3).

In our calculations, we use SATB1 residues 23–764 (Supplementary Table S1c), which contain all the functional domains with exclusion of protein localization signals. Employing the interaction fragments method, we predict interactions for two regions identified by Wutz *et al.* (nt 292–698 and 4725–6079; Figure 6) (6). In particular, we find that SATB1 has strong propensity to bind to RepA (region 292–698 nt; interaction strength: 85%; Figure 5d), as suggested by Arthold *et al.* (11). Intriguingly, we observe previously uncharacterized binding sites in correspondence of the 3′-region (Figure 5c), in agreement with the fact that more than one *Xist* region could be involved in low-affinity cooperative binding of protein factors (3,6).

DISCUSSION

XCI is a complex process that requires several regulated events such as the *Xist* localization onto the X-chromosome and its spatial confinement. These steps are controlled by transcriptional factors and nuclear scaffold proteins, which play a role in the selection of chromosome and recruitment of silencing machinery. One of the first processes during XCI is the random selection of the X-chromosome to be silenced. The choice has been suggested to be stochastically determined by levels of spliced *Xist* RNA accumulated on the X-chromosome (8). We find that the splicing factor SFRS1 binds to the 5′-UTR exon (Figure 3) and RepA (Figure 2b and c), which

suggests direct involvement of this protein in the production of mature *Xist* (8). Although RepA is fundamental for PCR2 recruitment and chromosomal silencing (7), we predict that it is unlikely to be involved in the interaction with YY1 (9) (Figure 4b). By contrast, we find that RepC has high interaction propensity for YY1 (Figure 4a and b). Hence, our predictions support the current hypothesis that PRC2 is co-transcriptionally recruited by RepA, while YY1 tethers RepC on the X-inactivation centre (9).

How the *Xist*–PRC2 complex translocates *in cis* along the X-chromosome is an open and tantalizing question. It has been reported that the nuclear scaffold factor SAF-A facilitates the association of *Xist* with nuclear matrix (10). Indeed, the nuclear matrix could provide a highly dynamic structure (27,28) to control *Xist* movements. We observe that the interaction profile of SAF-A correlates (Figure 5a) with that of the nuclear matrix protein SATB1 (Figure 5c) at the 5′, suggesting a possible synergistic mechanism of action to organize *Xist* translocation along the X-chromosome. The involvement of matrix-associated factors in the X-chromosome coating represents an intriguing scenario to be further investigated experimentally.

Our calculations suggest that localization and confinement of *Xist* are finely regulated by multiple factors acting at the interface between chromosome X and the nuclear matrix. Our results are compatible with a model in which following X-chromosome docking mediated by YY1 (9), matrix-associated proteins SAF-A and SATB1 recruit the 5′-half of *Xist* and drive the translocation *in cis* of the *Xist*–PRC2 complex.

In this work, we presented a new version of the catRAPID method to study *Xist* associations with a number of proteins, including SUZ12, EZH2, YY1, SAF-A, SFRS1 and SATB1. In striking agreement with experimental evidence, we demonstrated that our algorithms predict RNA-binding sites and affinities for a number of epigenetic, splicing and transcription factors. In particular, we investigated the association with transcription repressor YY1, which favours *Xist* tethering onto the X-chromosome, and nuclear matrix proteins SAF-A and SATB1, which guide its translocation. We also applied our method to SFRS1's interactome, showing that catRAPID predicts CLIP-seq-binding sites with great accuracy (18). Most importantly, we showed that computational approaches can provide a solid basis for the investigation of protein interactions with long non-coding transcripts (20).

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online: Supplementary Table 1 and Supplementary Figures 1–4.

ACKNOWLEDGEMENTS

The authors thank Domenica Marchese; Prof. R. Guigo'; Dr B. Keyes and Dr L. di Croce for stimulating discussions and Dr J.R. Sanford for having provided CLIP-seq data on SFSR1.

FUNDING

Funding for open access charge: Spanish Ministry of Economy and Competitiveness [SAF2011-26211].

Conflict of interest statement. None declared.

REFERENCES

1. Navarro,P. and Avner,P. (2010) An embryonic story: analysis of the gene regulative network controlling *Xist* expression in mouse embryonic stem cells. *Bioessays*, **32**, 581–588.
2. Tattermusch,A. and Brockdorff,N. (2011) A scaffold for X chromosome inactivation. *Hum. Genet.*, **130**, 247–253.
3. Wutz,A. (2011) Gene silencing in X-chromosome inactivation: advances in understanding facultative heterochromatin formation. *Nat. Rev. Genet.*, **12**, 542–553.
4. Lee,J.T., Davidow,L.S. and Warshawsky,D. (1999) Tsix, a gene antisense to *Xist* at the X-inactivation centre. *Nat. Genet.*, **21**, 400–404.
5. Zhao,J., Sun,B.K., Erwin,J.A., Song,J.-J. and Lee,J.T. (2008) Polycomb proteins targeted by a short repeat RNA to the mouse X-chromosome. *Science*, **322**, 750–756.
6. Wutz,A., Rasmussen,T.P. and Jaenisch,R. (2002) Chromosomal silencing and localization are mediated by different domains of *Xist* RNA. *Nat. Genet.*, **30**, 167–174.
7. Maenner,S., Blaud,M., Fouillen,L., Savoye,A., Marchand,V., Dubois,A., Sanglier-Cianférani,S., Van Dorsselaer,A., Clerc,P., Avner,P. *et al.* (2010) 2-D structure of the A region of *Xist* RNA and its implication for PRC2 association. *PLoS Biol.*, **8**, e1000276.
8. Royce-Tolland,M.E., Andersen,A.A., Koyfman,H.R., Talbot,D.J., Wutz,A., Tonks,I.D., Kay,G.F. and Panning,B. (2010) The A-repeat links ASF/SF2-dependent *Xist* RNA processing with random choice during X inactivation. *Nat. Struct. Mol. Biol.*, **17**, 948–954.
9. Jeon,Y. and Lee,J.T. (2011) YY1 tethers *Xist* RNA to the inactive X nucleation center. *Cell*, **146**, 119–133.
10. Hasegawa,Y., Brockdorff,N., Kawano,S., Tsutui,K., Tsutui,K. and Nakagawa,S. (2010) The matrix protein hnRNP U is required for chromosomal localization of *Xist* RNA. *Dev. Cell*, **19**, 469–476.
11. Arthold,S., Kurowski,A. and Wutz,A. (2011) Mechanistic insights into chromosome-wide silencing in X inactivation. *Hum. Genet.*, **130**, 295–305.
12. Bellucci,M., Agostini,F., Masin,M. and Tartaglia,G.G. (2011) Predicting protein associations with long noncoding RNAs. *Nat. Methods*, **8**, 444–445.
13. Gruber,A.R., Lorenz,R., Bernhart,S.H., Neubock,R. and Hofacker,I.L. (2008) The Vienna RNA Websuite. *Nucleic Acids Res.*, **36**, W70–W74.
14. Pervouchine,D.D., Graber,J.H. and Kasif,S. (2003) On the normalization of RNA equilibrium free energy to the length of the sequence. *Nucleic Acids Res.*, **31**, e49–e49.
15. Hofacker,I.L., Priwitzer,B. and Stadler,P.F. (2004) Prediction of locally stable RNA secondary structures for genome-wide surveys. *Bioinformatics*, **20**, 186–190.
16. Fang,J., Chen,T., Chadwick,B., Li,E. and Zhang,Y. (2004) Ring1b-mediated H2A ubiquitination associates with inactive X chromosomes and is involved in initiation of X inactivation. *J. Biol. Chem.*, **279**, 52812–52815.
17. Duszczyk,M.M., Zanier,K. and Sattler,M. (2008) A NMR strategy to unambiguously distinguish nucleic acid hairpin and duplex conformations applied to a *Xist* RNA A-repeat. *Nucleic Acids Res.*, **36**, 7068–7077.
18. Sanford,J.R., Wang,X., Mort,M., Vanduyn,N., Cooper,D.N., Mooney,S.D., Edenberg,H.J. and Liu,Y. (2009) Splicing factor SFRS1 recognizes a functionally diverse landscape of RNA transcripts. *Genome Res.*, **19**, 381–394.
19. Wang,X., Juan,L., Lv,J., Wang,K., Sanford,J.R. and Liu,Y. (2011) Predicting sequence and structural specificities of RNA binding regions recognized by splicing factor SRSF1. *BMC Genomics*, **12(Suppl. 5)**, S8.
20. Thorvaldsen,J.L., Weaver,J.R. and Bartolomei,M.S. (2011) A YY1 Bridge for X Inactivation. *Cell*, **146**, 11–13.

21. Fackelmayer,F.O. (2005) A stable proteinaceous structure in the territory of inactive X chromosomes. *J. Biol. Chem.*, **280**, 1720–1723.
22. Kipp,M., Göhring,F., Ostendorp,T., van Drunen,C.M., van Driel,R., Przybylski,M. and Fackelmayer,F.O. (2000) SAF-Box, a conserved protein domain that specifically recognizes scaffold attachment region DNA. *Mol. Cell. Biol.*, **20**, 7480–7489.
23. Mirkovitch,J., Gasser,S.M. and Laemmli,U.K. (1987) Relation of chromosome structure and gene expression. *Phil. Trans. R. Soc. Lond. B, Biol. Sci.*, **317**, 563–574.
24. Ponting,C., Schultz,J. and Bork,P. (1997) SPRY domains in ryanodine receptors (Ca(2+)-release channels). *Trends Biochem. Sci.*, **22**, 193–194.
25. Agrelo,R., Souabni,A., Novatchkova,M., Haslinger,C., Leeb,M., Komnenovic,V., Kishimoto,H., Gresh,L., Kohwi-Shigematsu,T., Kenner,L. *et al.* (2009) SATB1 defines the developmental context for gene silencing by *Xist* in lymphoma and embryonic cells. *Dev. Cell*, **16**, 507–516.
26. Chaumeil,J., Le Baccon,P., Wutz,A. and Heard,E. (2006) A novel role for *Xist* RNA in the formation of a repressive nuclear compartment into which genes are recruited when silenced. *Genes Dev.*, **20**, 2223–2237.
27. Albrethsen,J., Knol,J.C. and Jimenez,C.R. (2009) Unravelling the nuclear matrix proteome. *J. Proteomics*, **72**, 71–81.
28. Simon,D.N. and Wilson,K.L. (2011) The nucleoskeleton as a genome-associated dynamic 'network of networks'. *Nat. Rev. Mol. Cell Biol.*, **12**, 695–708.

Chapter III

*cat*RAPID *omics*

To facilitate the investigation of RNA-protein interactions at a genome-wide scale, I developed *cat*RAPID *omics*, which allows fast calculation of ribonucleoprotein associations in a number of model organisms (*Caenorhabditis elegans*, *Danio rerio*, *Drosophila melanogaster*, *Homo sapiens*, *Mus musculus*, *Rattus norvegicus*, *Saccharomyces cerevisiae* and *Xenopus tropicalis*). The algorithm computes the interaction between a molecule (protein/transcript) and the pre-compiled reference library (transcriptome/proteome) for each model organism. In addition to the interaction propensities, discriminative power and interaction strength, the method employs the Pfam (Sonnhammer et al., 1997) and RBPDB (Cook et al., 2011) databases to provide information on the presence of known RNA-binding domains and recognition motifs within the molecules involved in the interaction. The performances have been assessed on low- and high-throughput studies of protein and RNA associations.

Agostini, F., Zanzoni, A., Klus, P., Marchese, D., Cirillo, D., and Tartaglia, G. G. (2013). *cat*RAPID *omics*: a web server for large-scale prediction of protein-RNA interactions. *Bioinformatics (Oxford, England)*, 29(22):2928–2930

DOI: 10.1093/bioinformatics/btt495

BIOINFORMATICS APPLICATIONS NOTE Vol. 29 no. 22 2013, pages 2928–2930
doi:10.1093/bioinformatics/btt495

Sequence analysis

Advance Access publication August 23, 2013

*cat*RAPID *omics*: a web server for large-scale prediction of protein–RNA interactions

Federico Agostini[1,2], Andreas Zanzoni[1,2], Petr Klus[1,2], Domenica Marchese[1,2], Davide Cirillo[1,2] and Gian Gaetano Tartaglia[1,2,*]

[1]Gene Function and Evolution, Bioinformatics and Genomics, Centre for Genomic Regulation (CRG) and [2]Universitat Pompeu Fabra (UPF), 08003 Barcelona, Spain

Associate Editor: Ivo Hofacker

ABSTRACT

Summary: Here we introduce *cat*RAPID *omics, a* server for large-scale calculations of protein–RNA interactions. Our web server allows (i) predictions at proteomic and transcriptomic level; (ii) use of protein and RNA sequences without size restriction; (iii) analysis of nucleic acid binding regions in proteins; and (iv) detection of RNA motifs involved in protein recognition.

Results: We developed a web server to allow fast calculation of ribonucleoprotein associations in *Caenorhabditis elegans, Danio rerio, Drosophila melanogaster, Homo sapiens, Mus musculus, Rattus norvegicus, Saccharomyces cerevisiae* and *Xenopus tropicalis* (custom libraries can be also generated). The *cat*RAPID *omics* was benchmarked on the recently published RNA interactomes of Serine/arginine-rich splicing factor 1 (SRSF1), Histone-lysine N-methyltransferase EZH2 (EZH2), TAR DNA-binding protein 43 (TDP43) and RNA-binding protein FUS (FUS) as well as on the protein interactomes of U1/U2 small nucleolar RNAs, X inactive specific transcript (Xist) repeat A region (RepA) and Crumbs homolog 3 (CRB3) 3′-untranslated region RNAs. Our predictions are highly significant ($P < 0.05$) and will help the experimentalist to identify candidates for further validation.

Availability: *cat*RAPID *omics* can be freely accessed on the Web at http://s.tartaglialab.com/catrapid/omics. Documentation, tutorial and FAQs are available at http://s.tartaglialab.com/page/catrapid_group.

Contact: gian.tartaglia@crg.eu

Received on March 21, 2013; revised on July 30, 2013; accepted on August 16, 2013

1 INTRODUCTION

Increasing evidence indicates that ribonucleoprotein interactions are fundamental for cellular regulation (Khalil and Rinn, 2011). Moreover, several studies highlighted the involvement of RNA molecules in the onset and progression of human diseases including neurological disorders (Johnson *et al.*, 2012). To our knowledge, there are two sequence-based methods for prediction of protein–RNA interactions: *cat*RAPID (Bellucci *et al.*, 2011) and RPISeq (Muppirala *et al.*, 2011). The *cat*RAPID algorithm exploits predictions of secondary structure, hydrogen bonding and van der Waals' contributions to estimate the binding propensity of protein and RNA molecules. RPISeq is based on support vector machine (SVM) and random forest (RF) models predicting protein–RNA interactions from primary structure alone (Muppirala *et al.*, 2011). Both methods show remarkable performances, but *cat*RAPID discriminates positive and negative cases with higher accuracy (Cirillo *et al.*, 2013b) and has been tested on long non-coding RNAs (Agostini *et al.*, 2013).

Here we introduce *cat*RAPID *omics* to perform high-throughput predictions of protein–RNA interactions using the information on protein and RNA domains involved in macromolecular recognition.

2 WORKFLOW AND IMPLEMENTATION

The *cat*RAPID *omics* server provides two main services to explore the interaction potential of (i) a protein of interest with respect to a target transcriptome or (ii) a given RNA with respect to the nucleic acid binding proteome. Several options are available to refine the type of analysis in eight model organisms or custom libraries (see online documentation):

- In the case of a protein query, *cat*RAPID *omics* takes as input the protein sequence (FASTA format): full-length or, alternatively, nucleic acid binding regions.
- For a transcript query (FASTA format), the server uses the full-length sequence if below 1200 nt, or, alternatively, uses fragments with predicted stable secondary structure (Agostini *et al.*, 2013). Full-length proteins and nucleic acid binding regions can be searched.
- The server automatically detects disordered proteins lacking canonical RNA binding domains. Indeed, it has been observed that disordered regions are enriched in RNA binding proteins (Castello *et al.*, 2012).
- As RNA motifs are important for protein recognition (Kazan *et al.*, 2010), a search for these elements is carried out. The motifs were taken from RNA-Binding Protein DataBase (RBPDB) (Cook *et al.*, 2011), SpliceAid-F (Giulietti *et al.*, 2013) and a recent motif compendium (Ray *et al.*, 2013).
- Using the interaction propensities distribution, *cat*RAPID *omics* predicts the RNA binding ability of the input protein (86% accuracy) and ranks RNA interactions (downloadable by the user).

*To whom correspondence should be addressed.

Fig. 1. *cat*RAPID *omics* features and performances. (**A**) Example of the output table showing *Z*-score (interaction propensity normalized with respect to experimental cases), discriminative power (with respect to training sets), interaction strength (enrichment with respect to random interactions) and presence of RNA binding domains as well as RNA motifs. Interaction scores are ranked according to a 'star rating system' ranging from 0 to 3 (http://service.tartaglialab.com/static_files/shared/faqs.html). A click on the text redirects to reference pages. Performances on (**B**) full-length proteins and (**C**) RNA binding protein domains. Gray is used to highlight transcriptomic studies (i.e. RNA sequencing) and red indicates proteomic analyses (i.e. mass spectrometry). The significance of our predictions was assessed using Fisher's exact test (the dashed line corresponds to $P = 0.05$)

In the output page (Fig. 1A), we report all the variables used to estimate protein–RNA associations: interaction propensity (Bellucci *et al*., 2011), discriminative power (Bellucci *et al*., 2011), interaction strength (Agostini *et al*., 2013) and presence of protein RNA binding domains as well as RNA motifs. A 'star rating system' ranks the binding propensities (http://service.tartaglialab.com/static_files/shared/faqs.html). As for the reference sets, ENSEMBL (version 68) is used for retrieval and classification of coding and non-coding RNAs, whereas protein sequences are gathered from the UniProtKB database (release 2012_11). Finally, *cat*RAPID *omics* uses hmmscan, a Hidden Markov Model-based algorithm from the HMMER3 package (Finn *et al*., 2011), to identify known PfamA domains (Finn *et al*., 2009) and recognize protein regions involved in binding nucleic acid molecules. Algorithm hit significance is determined according to the PfamA 'gathering thresholds'.

3 PERFORMANCES

The *cat*RAPID algorithm has been previously validated on a number of protein–RNA associations (Agostini *et al*., 2013; Bellucci *et al*., 2011; Cirillo, *et al*., 2013a; Johnson *et al*., 2012). To evaluate large-scale performances of *cat*RAPID *omics*, we used data from recent large-scale experiments. To compare predicted and experimental interactions, we used Fisher's exact test. As shown in Figure 1B, performances on the human splicing factor serine/arginine-rich splicing factor 1 (SRSF1) (Sanford *et al*., 2009) and murine nucleic acid binding protein Histone-lysine N-methyltransferase EZH2 (EZH2) (Zhao *et al*., 2010) are highly significant (*P*-values: 0.01 and 0.01, respectively). Good performances are found for low-throughput experiments on murine non-coding X inactive specific transcript (Xist) repeat A region (RepA) (Maenner *et al*., 2010; Royce-Tolland *et al*., 2010) and yeast small nuclear RNA U1 (Cvitkovic and Jurica, 2012) (*P*-values: 0.03 and 0.015) (Fig. 1B). To illustrate the ability of *cat*RAPID *omics* to predict interactions with nucleic acid binding domains (Fig. 1C), we used murine FUS (Han *et al*., 2012) and rat TAR DNA-binding protein 43 (TDP43) (Sephton *et al*., 2011) (*P*-values: 3e-05 and 0.002) as well as human Crumbs homolog 3 (CRB3) 3′-untranslated region (Iioka *et al*., 2011) and yeast small nuclear U2 (Cvitkovic and Jurica, 2012) (*P*-values: 0.001 and 2e-0.6). To evaluate *cat*RAPID's performances on high-throughput data, we collected positive interactions (TDP43: 568, FUS: 99, SRSF1: 358, EZH2: 1141) as well as negative controls (same numbers as positives and generated in four random extractions). Comparing the interaction scores of positives and negatives, we found enrichment (calculated as discriminative power) in 72% (TDP43), 88% (FUS), 74% (SRSF1) and 56% (EZH2) of cases. On the same datasets, SVM RPIseq showed enrichment in 58% (TDP43; RF has enrichment in 53%), 83% (FUS; RF has enrichment in 68%), 47% (SRSF1; RF has enrichment in 59%) and 41% (EZH2; RF has enrichment in 48%) of cases.

4 CONCLUSIONS

Despite recent technical developments, detection of protein–RNA associations remains a challenging task. For this reason, we developed an algorithm that can be used to complement experimental efforts (Zanzoni *et al*., 2013). The *cat*RAPID *omics* server offers unique features such as organism-specific proteomic and transcriptomic libraries, possibility to generate custom datasets, analysis of long sequences and calculation of interaction specificities. Moreover, we implemented an algorithm for the detection of RNA motifs as well as protein RNA binding domains, which will help to retrieve recognition motifs embedded in sequences. Our server enables fast calculations of ribonucleoprotein associations and predicts RNA binding activity of proteins with high accuracy, thus resulting in a powerful tool for designing new experiments.

ACKNOWLEDGEMENTS

The authors would like to thank Dr J.R. Sanford CLIP-seq data on SFSR1, Prof R. Guigó and Dr G. Bussotti for stimulating discussions.

Funding: Spanish Ministry of Economy and Competitiveness (SAF2011-26211), the European Research Council (ERC Starting Grant to G.G.T) and the RTTIC project (to A.Z.). 'La Caixa' fellowship (to P.K.). Programa de Ayudas FPI del Ministerio de Economia y Competitividad—BES-2012-052457 (to D.M.).

Conflict of Interest: none declared.

REFERENCES

Agostini,F. *et al.* (2013) X-inactivation: quantitative predictions of protein interactions in the Xist network. *Nucleic Acids Res.*, **41**, e31.

Bellucci,M. *et al.* (2011) Predicting protein associations with long noncoding RNAs. *Nat. Methods*, **8**, 444–445.

Castello,A. *et al.* (2012) Insights into RNA biology from an atlas of mammalian mRNA-binding proteins. *Cell*, **149**, 1393–1406.

Cirillo,D. *et al.* (2013a) Neurodegenerative diseases: quantitative predictions of protein–RNA interactions. *RNA*, **19**, 129–140.

Cirillo,D. *et al.* (2013b) Predictions of protein–RNA interactions. *Wiley Interdiscip. Rev. Comput. Mol. Sci.*, **3**, 161–175.

Cook,K.B. *et al.* (2011) RBPDB: a database of RNA-binding specificities. *Nucleic Acids Res.*, **39**, D301–D308.

Cvitkovic,I. and Jurica,M.S. (2012) Spliceosome database: a tool for tracking components of the spliceosome. *Nucleic Acids Res.*, **41**, D132–D141.

Finn,R.D. *et al.* (2009) The Pfam protein families database. *Nucleic Acids Res.*, **38**, D211–D222.

Finn,R.D. *et al.* (2011) HMMER web server: interactive sequence similarity searching. *Nucleic Acids Res.*, **39**, W29–W37.

Giulietti,M. *et al.* (2013) SpliceAid-F: a database of human splicing factors and their RNA-binding sites. *Nucleic Acids Res.*, **41**, D125–D131.

Han,T.W. *et al.* (2012) Cell-free formation of RNA granules: bound RNAs identify features and components of cellular assemblies. *Cell*, **149**, 768–779.

Iioka,H. *et al.* (2011) Efficient detection of RNA-protein interactions using tethered RNAs. *Nucleic Acids Res.*, **39**, e53.

Johnson,R. *et al.* (2012) Neurodegeneration as an RNA disorder. *Prog. Neurobiol.*, **99**, 293–315.

Kazan,H. *et al.* (2010) RNAcontext: a new method for learning the sequence and structure binding preferences of RNA-binding proteins. *PLoS Comput. Biol.*, **6**, e1000832.

Khalil,A.M. and Rinn,J.L. (2011) RNA-protein interactions in human health and disease. *Semin. Cell Dev. Biol.*, **22**, 359–365.

Maenner,S. *et al.* (2010) 2-D structure of the A Region of Xist RNA and its implication for PRC2 association. *PLoS Biol.*, **8**, e1000276.

Muppirala,U.K. *et al.* (2011) Predicting RNA-protein interactions using only sequence information. *BMC Bioinformatics*, **12**, 489.

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

Royce-Tolland,M.E. *et al.* (2010) The A-repeat links ASF/SF2-dependent Xist RNA processing with random choice during X inactivation. *Nat. Struct. Mol. Biol.*, **17**, 948–954.

Sanford,J.R. *et al.* (2009) Splicing factor SFRS1 recognizes a functionally diverse landscape of RNA transcripts. *Genome Res.*, **19**, 381–394.

Sephton,C.F. *et al.* (2011) Identification of neuronal RNA targets of TDP-43-containing ribonucleoprotein complexes. *J. Biol. Chem.*, **286**, 1204–1215.

Zanzoni,A. *et al.* (2013) Principles of self-organization in biological pathways: a hypothesis on the autogenous association of alpha-synuclein. *Nucleic Acids Res.*, [Epub ahead of print, doi: 10.1093/nar/gkt794, September 3, 2013].

Zhao,J. *et al.* (2010) Genome-wide identification of polycomb-associated RNAs by RIP-seq. *Mol. Cell*, **40**, 939–953.

Chapter IV

The *SeAMotE* algorithm

RNA-binding domains represent evolutionary conserved peptide domains that recognize specific sequence or structural elements embedded in their target RNAs, which are referred to as RNA recognition elements (RREs) (Ascano et al., 2012). Here, I introduce the *SeAMotE* algorithm to perform discriminative motif discovery analyses on large sets of nucleic acid sequences. The approach offers unique features such as the discrimination based on the actual occurrences in the datasets, the choice of multiple reference backgrounds (shuffle, random or custom) and the output of the most significant motifs in the whole span of tested motif widths, thus providing a wide range of solutions. *SeAMotE* ability to find the patterns that best represent the RBPs recognition motifs is compared against the well-established DREME method (Bailey, 2011). To test the performances of both approaches, we use datasets of bound and not bound transcripts derived from recent PAR-CLIP experiments.

This article has been submitted for publication to the "Web Server" issue of "Nucleic Acids Research".

Agostini F, Cirillo D, Ponti RD, Tartaglia GG. SeAMotE: a method for high-throughput motif discovery in nucleic acid sequences. BMC Genomics. 2014 Oct 23; 15: 925. DOI: 10.1186/1471-2164-15-92

SeAMotE: a web-server for high-throughput motif discovery in nucleic acid sequences

Federico Agostini[1,2], Davide Cirillo[1,2] and Gian Gaetano Tartaglia[1,2,*]

[1]Gene Function and Evolution, Centre for Genomic Regulation (CRG), Dr. Aiguader 88, 08003 Barcelona, Spain.

[2]Universitat Pompeu Fabra (UPF), 08003 Barcelona, Spain.

* Corresponding Author. Tel: +34 933160116; Fax: +34 933969983; Email: gian.tartaglia@crg.eu

ABSTRACT

The large amount of data produced by high-throughput sequencing poses a number of computational challenges. In the last decade, several tools have been developed for the identification of gene regulation elements such as transcription and splicing factor binding sites. Here, we introduce the web-server SeAMotE (Sequence Analysis of Motifs Enrichment) for large-scale calculation of discriminative elements in nucleic acid sequences. SeAMotE provides (i) a robust and fast analysis of high-throughput sequence collections, (ii) a motif search based on pattern occurrences within the datasets and (iii) an easy-to-use web-server interface. We applied our approach to recently published data generated with crosslinking immunoprecipitation (CLIP) experiments and compared our results with those of other well-established discriminative motif discovery tools. SeAMotE shows an average accuracy of 80% in finding discriminative motifs and outperforms DREME in 70% of cases. The server can be freely accessed on the Web at http://s.tartaglialab.com/new_submission/seamote.

INTRODUCTION

Transcriptional and post-transcriptional events rely on inter-molecular recognition and interaction mechanisms. These processes involve the interplay between protein effectors and nucleic acid targets, whose physical association is thought to be guided by linear motifs and/or specific structural elements (1–3). In the past decade the advancement of high-throughput technologies contributed to the generation of a large amount of genomic data (4), promoting development of computational methods to detect regulatory elements such as transcription and splicing factor binding sites (5). On the one hand, algorithms for large-scale sequence analysis must be able to identify relevant features (e.g., recognition motifs) in a reasonable time (6, 7). On the other hand, bioinformatics tools should be as comprehensive as possible to provide insights into the nature of regulatory elements in the genomic context, which requires comparison with biologically relevant reference sets (8).

As discussed by Ma *et al.* (9) and Weirauch *et al.* (10), there are several algorithms for *de novo* motif discovery, but only few are capable of performing a discriminative analysis (i.e., comparison between two sets) on high-throughput datasets:

- Dimont (7) is a discriminative method based on 'zero or one occurrence per sequence' (ZOOPS) model (11). It works under the hypothesis that only few binding sites are present within long target sequences. Although Dimont achieves acceptable runtimes (7),it performs the discriminative analysis using only the foreground (i.e., positive or signal) set, thus excluding the possibility of a comparison with an experimentally derived background (i.e., reference) set.

- DREME is a well-established algorithm that restricts the search for motifs to a simplified form of "regular expression" (RE): words over the IUPAC alphabet, which exploits 11 wildcard characters in addition to the standard DNA alphabet, ACGT (12). To save computation time, DREME estimates the significance of candidate RE by a heuristic search without scanning the whole input sequences (12).

- motifRG measures the discriminative power of a motif by a logistic regression model, which shows similarity to DREME (12) and comparable performances for the identification of core motifs (8). Although motifRG provides an efficient iterative process for seeds refinement and extension (8), the algorithm searches for specific patterns, identifying few motifs in both the background and foreground sets.

Despite the variety of motif discrimination approaches, knowledge of programming languages (8, 13) and acquaintance with web-based bioinformatics platforms (7, 14) limit use among experimentalists. In this article, we introduce SeAMotE, a web-server to perform *de novo* discriminative motif discovery in high-throughput nucleic acid datasets. Specifically, we present an approach that enables the exhaustive search of distinctive patterns in large sets of sequences, in a reasonable amount of computational time and with an easy-to-use interface.

MATERIAL AND METHODS

SeAMotE is based on the generation of nucleotide seeds followed by ZOOPS model testing and pattern refinement, techniques used in recently published tools (7, 8). However, SeAMotE includes a number of unique features that dramatically increase the performance of the method. The user can (i) set a coverage threshold, which is employed in the selection of enriched motifs for the positive set (foreground), and (ii) choose among multiple reference set (background) options.

Usage

The SeAMotE server presents an input page that allows the upload of nucleic acid sequences and the selection of the parameters. Default parameters (e.g. reference set, coverage threshold, etc.) are defined according to best settings estimated during testing on known data sets. However, most of the parameters can be modified by the user, which adds flexibility to our web-service. Detailed descriptions of the submission process and variables are provided in the on-line tutorial (see

http://service.tartaglialab.com/static_files/shared/tutorial_seamote.html, sections “Submission form” and “Interpreting the output”).

- At least one input set (FASTA format file) should be provided for the analysis. Currently, the number of sequences is limited to 10^4, with a maximal length of 15,000 nucleotides per sequence.
- A reference set is required to estimate the significance of the discovered motifs. This can be:
    1. Automatically generated as a shuffle set, where the foreground set composition (i.e., single nucleotide alphabet frequencies) and dimensions (i.e., sequence numbers and lengths) are kept constant;
    2. Automatically generated as a random set, where the foreground set dimensions are preserved but the internal composition is based on letter frequencies obtained from the human transcriptome/genome;
    3. Provided by the user (FASTA format file), having the same restrictions as the input set.
- The coverage threshold (i.e. the percentage of sequences matching the searched pattern) represents a threshold that the algorithm uses internally to select the most abundant motifs in the two datasets. The higher the threshold, the faster the process will run and the more stringent the search will be.

Optionally, the user can assign a job name for each submission and request for an email notification upon completion (not required to run the server).

The workflow starts by exploiting a series of pre-generated seed motifs (IUPAC alphabet) of *k*-mers (e.g. ACY = AC[CT]) that are employed to evaluate dataset coverage and discrimination. After the first scan of the datasets, the nucleotide patterns that score above the coverage threshold undergo an expansion process, which works by incorporating another nucleotide in the $k + 1$ position. The extended motifs are then used as seeds for the next round of calculation and the set coverage is re-evaluated. This process is executed iteratively until at least one motif is found above the threshold in the positive set.

The output summary consists of the information about the submission (e.g., identifier, downloadable datasets) and a table (Figure 1A). The latter displays the discovered motifs (IUPAC and RE formats), the logo representations and the statistics used to estimate their significance: motif coverage for positive and reference sets, discrimination factor (Youden's index = Sensitivity + Specificity - 1) and P-value (Fisher’s exact test) associated with each pattern. In addition, it is possible to retrieve the list of motifs tested (txt format), as well as their sequence logos (png format) and positional weighted matrices (txt format) following the links provided in the output page (Figure 1A).

Implementation

The web-server is implemented in Python, HTML and JavaScript, which provides a convenient framework for the pipeline control and the presentation of the output data. User-provided data are validated by Python scripts and passed to the Amazon Web Services (AWS), which manages the queue system, performs the redistribution of the work on our local machines and, once the job is completed successfully, forwards the user to the output page. Local operations are executed by an ANSI C application, whereas significance estimation and sequence logo design are computed using R and WebLogo (15), respectively. Typically, the computations take from between 2-3 and 30-40 minutes.

Documentation

The documentation/tutorial of the SeAMotE web-server is available online, and it can be accessed using the links in the menu at the top of every server page. It contains a brief description of the method, a tutorial and information on the benchmark data. Additionally, the web interface in the output page provides help notes (accessible also through the “mouse-over” function) for table variables and download buttons. Online documentation and “Frequently Asked Questions” (FAQs) sections updates will be provided on a regular basis according to method improvements and users’ inquiries, respectively.

Availability

The SeAMotE server is free and accessible to all users through the main browsers (we tested Safari, Firefox, Explorer and Chrome), and there is no login requirement. After clicking the “submit” button, a web link to the results is provided, which the user can bookmark and access at a later time. This page will refresh automatically every 10 seconds and redirects to the results once the job is successfully completed.

RESULTS

To evaluate SeAMotE performances on large-scale datasets, we collected recent CLIP experiments (16–24) and assessed ability to identify significantly enriched motifs (Fisher’s exact test). In each case analyzed, we compared RNAs bound to a specific protein (foreground set) with the same amount of non-interacting transcripts (background set; Supplementary Table 1). The DREME (12) algorithm was used as a reference to evaluate the performance of our system. Our method achieves both higher discrimination, which is the ability to separate the foreground from the background set, and significance, denoted by lower P-values associated with sequence motifs (Figure 1B). In addition, SeAMoTe also shows very high sensitivity (~90%) and accuracy (80%) (Table 1).

We compared SeAMoTe with DREME because other methods such as motifRG (8) show limited variability of motifs, which results in low abundance of sequence patterns (Supplementary Table 2)

and scarce discrimination between datasets. Indeed, motifRG has been developed for analysis of chromatin immunoprecipitation (ChIP) data with an optimal peak size of ~100nt (8) and is less accurate on larger nucleic acid regions (Supplementary Table 1).

DISCUSSION

Detection of regulatory motifs is a challenging task. For this reason, we developed the SeAMotE web-server, which provides an easy-to-use interface and allows the exhaustive analysis of large-scale datasets. Our approach offers unique features such as the discrimination based on the actual occurrences (i.e. pattern counts are not estimated) in the datasets, the choice of multiple reference backgrounds (shuffle, random or custom) and the output of the most significant motifs in the whole span of tested motif widths, thus providing a wide range of solutions. In conclusion, our web-server is a powerful tool for the identification of enriched sequence patterns that characterize recognition process between proteins and nucleic acids.

ACKNOWLEDGEMENT

The authors would like to thank Guillaume Filion (CRG), Andreas Zanzoni (Inserm, U1090), Giovanni Bussotti (EMBL-EBI) and Samuel Francis Reid (CRG) for stimulating discussions.

FUNDING

Spanish Ministry of Economy and Competitiveness [SAF2011-26211], the European Research Council (ERC Starting Grant RYBOMYLOME to G.G.T). Funding for open access charge: ERC and Spanish Ministry of Economy and Competitiveness [SAF2011-26211].

Conflict of interest statement. None declared.

REFERENCES

1. Coulon,A., Chow,C.C., Singer,R.H. and Larson,D.R. (2013) Eukaryotic transcriptional dynamics: from single molecules to cell populations. *Nat. Rev. Genet.*, **14**, 572–584, doi:10.1038/nrg3484, PMID:23835438.

2. Janga,S.C. (2012) From specific to global analysis of posttranscriptional regulation in eukaryotes: posttranscriptional regulatory networks. *Brief. Funct. Genomics*, **11**, 505–521, doi:10.1093/bfgp/els046, PMID:23124862.

3. Pichon,X., Wilson,L.A., Stoneley,M., Bastide,A., King,H.A., Somers,J. and Willis,A.E.E. (2012) RNA binding protein/RNA element interactions and the control of translation. *Curr. Protein Pept. Sci.*, **13**, 294–304, PMID:22708490.

4. Koboldt,D.C., Steinberg,K.M., Larson,D.E., Wilson,R.K. and Mardis,E.R. (2013) The next-generation sequencing revolution and its impact on genomics. *Cell*, **155**, 27–38, doi:10.1016/j.cell.2013.09.006, PMID:24074859.

5. Dassi,E. and Quattrone,A. (2012) Tuning the engine: an introduction to resources on post-transcriptional regulation of gene expression. *RNA Biol.*, **9**, 1224–1232, doi:10.4161/rna.22035, PMID:22995832.

6. Sinha,S. (2003) Discriminative motifs. *J. Comput. Biol. J. Comput. Mol. Cell Biol.*, **10**, 599–615, doi:10.1089/10665270360688219, PMID:12935347.

7. Grau,J., Posch,S., Grosse,I. and Keilwagen,J. (2013) A general approach for discriminative de novo motif discovery from high-throughput data. *Nucleic Acids Res.*, **41**, e197, doi:10.1093/nar/gkt831, PMID:24057214.

8. Yao,Z., Macquarrie,K.L., Fong,A.P., Tapscott,S.J., Ruzzo,W.L. and Gentleman,R.C. (2013) Discriminative motif analysis of high-throughput dataset. *Bioinforma. Oxf. Engl.*, 10.1093/bioinformatics/btt615.

9. Ma,X., Kulkarni,A., Zhang,Z., Xuan,Z., Serfling,R. and Zhang,M.Q. (2012) A highly efficient and effective motif discovery method for ChIP-seq/ChIP-chip data using positional information. *Nucleic Acids Res.*, **40**, e50, doi:10.1093/nar/gkr1135, PMID:22228832.

10. Weirauch,M.T., Cote,A., Norel,R., Annala,M., Zhao,Y., Riley,T.R., Saez-Rodriguez,J., Cokelaer,T., Vedenko,A., Talukder,S., et al. (2013) Evaluation of methods for modeling transcription factor sequence specificity. *Nat. Biotechnol.*, **31**, 126–134, doi:10.1038/nbt.2486, PMID:23354101.

11. Lawrence,C.E. and Reilly,A.A. (1990) An expectation maximization (EM) algorithm for the identification and characterization of common sites in unaligned biopolymer sequences. *Proteins*, **7**, 41–51, doi:10.1002/prot.340070105, PMID:2184437.

12. Bailey,T.L. (2011) DREME: motif discovery in transcription factor ChIP-seq data. *Bioinforma. Oxf. Engl.*, **27**, 1653–1659, doi:10.1093/bioinformatics/btr261, PMID:21543442.

13. Fauteux,F., Blanchette,M. and Strömvik,M.V. (2008) Seeder: discriminative seeding DNA motif discovery. *Bioinforma. Oxf. Engl.*, **24**, 2303–2307, doi:10.1093/bioinformatics/btn444, PMID:18718942.

14. Giardine,B., Riemer,C., Hardison,R.C., Burhans,R., Elnitski,L., Shah,P., Zhang,Y., Blankenberg,D., Albert,I., Taylor,J., et al. (2005) Galaxy: a platform for interactive large-scale genome analysis. *Genome Res.*, **15**, 1451–1455, doi:10.1101/gr.4086505, PMID:16169926.

15. Crooks,G.E., Hon,G., Chandonia,J.-M. and Brenner,S.E. (2004) WebLogo: a sequence logo generator. *Genome Res.*, **14**, 1188–1190, doi:10.1101/gr.849004, PMID:15173120.

16. Hafner,M., Landthaler,M., Burger,L., Khorshid,M., Hausser,J., Berninger,P., Rothballer,A., Ascano,M., Jungkamp,A.-C., Munschauer,M., et al. (2010) PAR-CliP--a method to identify transcriptome-wide the binding sites of RNA binding proteins. *J. Vis. Exp. JoVE*, 10.3791/2034.

17. Lebedeva,S., Jens,M., Theil,K., Schwanhäusser,B., Selbach,M., Landthaler,M. and Rajewsky,N. (2011) Transcriptome-wide analysis of regulatory interactions of the RNA-binding protein HuR. *Mol. Cell*, **43**, 340–352, doi:10.1016/j.molcel.2011.06.008, PMID:21723171.

18. Kishore,S., Jaskiewicz,L., Burger,L., Hausser,J., Khorshid,M. and Zavolan,M. (2011) A quantitative analysis of CLIP methods for identifying binding sites of RNA-binding proteins. *Nat. Methods*, **8**, 559–564, doi:10.1038/nmeth.1608, PMID:21572407.

19. Mukherjee,N., Corcoran,D.L., Nusbaum,J.D., Reid,D.W., Georgiev,S., Hafner,M., Ascano,M.,Jr, Tuschl,T., Ohler,U. and Keene,J.D. (2011) Integrative regulatory mapping indicates that the RNA-binding protein HuR couples pre-mRNA processing and mRNA stability. *Mol. Cell*, **43**, 327–339, doi:10.1016/j.molcel.2011.06.007, PMID:21723170.

20. Hoell,J.I., Larsson,E., Runge,S., Nusbaum,J.D., Duggimpudi,S., Farazi,T.A., Hafner,M., Borkhardt,A., Sander,C. and Tuschl,T. (2011) RNA targets of wild-type and mutant FET family proteins. *Nat. Struct. Mol. Biol.*, **18**, 1428–1431, doi:10.1038/nsmb.2163, PMID:22081015.

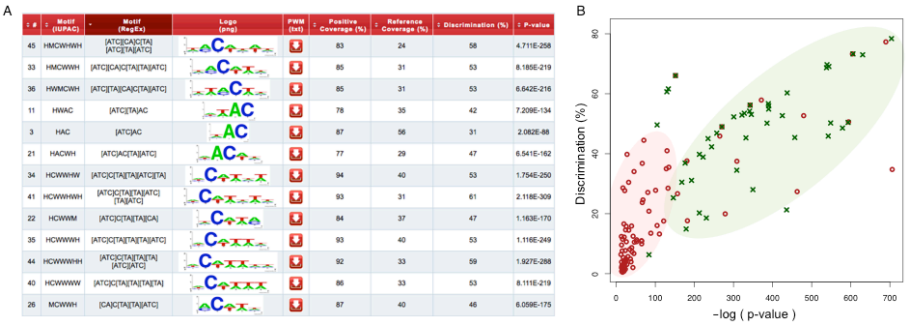
21. Sanford,J.R., Wang,X., Mort,M., Vanduyn,N., Cooper,D.N., Mooney,S.D., Edenberg,H.J. and Liu,Y. (2009) Splicing factor SFRS1 recognizes a functionally diverse landscape of RNA transcripts. *Genome Res.*, **19**, 381–394, doi:10.1101/gr.082503.108, PMID:19116412.

22. Tollervey,J.R., Curk,T., Rogelj,B., Briese,M., Cereda,M., Kayikci,M., König,J., Hortobágyi,T., Nishimura,A.L., Zupunski,V., et al. (2011) Characterizing the RNA targets and position-dependent splicing regulation by TDP-43. *Nat. Neurosci.*, **14**, 452–458, doi:10.1038/nn.2778, PMID:21358640.

23. Wang,Z., Kayikci,M., Briese,M., Zarnack,K., Luscombe,N.M., Rot,G., Zupan,B., Curk,T. and Ule,J. (2010) iCLIP predicts the dual splicing effects of TIA-RNA interactions. *PLoS Biol.*, **8**, e1000530, doi:10.1371/journal.pbio.1000530, PMID:21048981.

24. Ray,D., Kazan,H., Cook,K.B., Weirauch,M.T., Najafabadi,H.S., Li,X., Gueroussov,S., Albu,M., Zheng,H., Yang,A., et al. (2013) A compendium of RNA-binding motifs for decoding gene regulation. *Nature*, **499**, 172–177, doi:10.1038/nature12311.

TABLE AND FIGURES LEGENDS



#	Motif (IUPAC)	Motif (RegEx)	Logo (png)	PWM (txt)	Positive Coverage (%)	Reference Coverage (%)	Discrimination (%)	P-value
45	HMCWHWH	[ATC][CA]C[TA][ATC][TA][ATC]			83	24	58	4.711E-258
33	HMCWWH	[ATC][CA]C[TA][TA][ATC]			85	31	53	8.185E-219
36	HWMCWH	[ATC][TA][CA]C[TA][ATC]			85	31	53	6.642E-216
11	HWAC	[ATC][TA]AC			78	35	42	7.209E-134
3	HAC	[ATC]AC			87	56	31	2.082E-88
21	HACWH	[ATC]AC[TA][ATC]			77	29	47	6.541E-162
34	HCWWHW	[ATC]C[TA][TA][ATC][TA]			94	40	53	1.754E-250
41	HCWWHWH	[ATC]C[TA][TA][ATC][TA][ATC]			93	31	61	2.118E-309
22	HCWWM	[ATC]C[TA][TA][CA]			84	37	47	1.163E-170
35	HCWWWH	[ATC]C[TA][TA][TA][ATC]			93	40	53	1.116E-249
44	HCWWWHH	[ATC]C[TA][TA][TA][ATC][ATC]			92	33	59	1.927E-288
40	HCWWWW	[ATC]C[TA][TA][TA][TA]			86	33	53	8.111E-219
26	MCWWH	[CA]C[TA][TA][ATC]			87	40	46	6.059E-175

Figure 1. SeAMotE output summary and performances. A) Example of output showing the list of best discriminating patterns (IUPAC and RegEx) with their logo representations and positional weighted matrix download button, positive and reference coverage (as percentage of sequences containing at least one occurrence), discrimination (Youden's index) and associated P-value (Fisher's exact test). By clicking on the logo, it is possible to retrieve the image file (png format) of the associated motif. B) Comparison of SeAMotE and DREME performances plotted as the discrimination of top motifs (4- to 7-mers) obtained with DREME (red circles) and SeAMotE (green crosses), against the minus log of the P-value (Fisher's exact test).

CLIP protein	SeAMotE					DREME				
	TPR (%)	SPC (%)	PPV (%)	FDR (%)	ACC (%)	TPR (%)	SPC (%)	PPV (%)	FDR (%)	ACC (%)
ELAVL1 (Hafner)[16]	91.5	69.9	75.2	24.8	80.7	83.0	73.2	75.6	24.4	78.1
ELAVL1 (Lebedeva)[17]	81.0	73.0	75.0	25.0	77.0	75.6	74.5	73.3	26.7	75.0
ELAVL1 (Mnase)	93.4	69.9	75.6	24.4	81.7	86.3	71.6	75.2	24.8	79.0
ELAVL1 (Mukharjee)[19]	90.3	84.3	85.2	14.8	87.3	89.5	81.7	83.1	16.9	85.6
FUS[20]	92.9	66.6	73.5	26.5	79.7	92.2	45.3	62.8	37.2	68.8
IGF2BP1-3[16]	84.5	35.8	56.8	43.2	60.1	92.5	27.5	56.0	44.0	60.0
PUM2[16]	91.8	87.5	88.0	12.0	89.6	84.9	92.4	91.8	8.2	88.7
QKI[16]	91.0	78.1	80.6	19.4	84.6	88.4	84.9	85.4	14.6	86.6
SFSR1[21]	86.5	79.6	80.7	19.3	83.0	86.5	79.6	80.7	19.3	83.0
TAF15[20]	94.9	60.0	70.3	29.7	77.5	91.0	54.9	66.9	33.1	73.0
TARDBP (iCLIP)[22]	91.3	85.7	86.5	13.5	88.5	87.9	93.8	93.5	6.5	90.9
TIA1 (iCLIP)[23]	86.7	62.3	70.4	29.6	74.5	86.7	62.3	70.4	29.6	74.5
TIAL1 (iCLIP)[23]	84.9	65.5	71.3	28.7	75.2	84.4	66.2	71.7	28.3	75.3
TOTAL	89.3	70.6	76.1	23.9	80.0	86.8	69.8	75.9	24.1	78.3

Table 1. Comparison of SeAMotE and DREME methods. Sensitivity (True Positive Rate, TPR), specificity (SPC), precision (Positive Predictive Value, PPV), false discovery rate (FDR) and accuracy (ACC) achieved by the two methods on the experimental datasets.

CLIP protein	Positive Set				Negative Set			
	Number of Sequences	Minimum length	Maximum length	GC content	Number of Sequences	Minimum length	Maximum length	GC content
ELAVL1 (Hafner)[16]	1000	40	160	31.1%	1000	40	160	47.3%
ELAVL1 (Lebedeva)[17]	1445	16	324	29.5%	1445	17	324	41.7%
ELAVL1 (Mnase)	1000	40	166	29.4%	1000	40	160	47.4%
ELAVL1 (Mukharjee)[19]	5625	15	111	19.2%	5625	15	111	40.1%
FUS[20]	1568	16	57	25.0%	1568	17	57	39.3%
IGF2BP1-3[16]	3799	25	2015	41.0%	3799	25	2015	41.6%
PUM2[16]	1000	19	161	24.3%	1000	20	161	44.8%
QKI[16]	1000	17	84	26.5%	1000	18	71	39.7%
SFSR1[21]	310	30	84	52.6%	314	30	85	44.7%
TAF15[20]	1000	15	51	26.3%	1000	16	51	39.6%
TARDBP (iCLIP)[22]	4755	14	1653	43.4%	4745	16	1653	42.2%
TIA1 (iCLIP)[23]	1000	12	215	29.3%	968	12	233	42.2%
TIAL1 (iCLIP)[23]	2117	12	306	30.6%	2093	12	444	41.5%

Supplementary Table 1. Datasets composition. CLIP dataset information: number of sequences, minimum and maximum lengths, and GC content.

CLIP protein	Motif 1	Positive Count	Positive Coverage	Reference Count	Reference Coverage	Motif 2	Positive Count	Positive Coverage	Reference Count	Reference Coverage	Motif 3	Positive Count	Positive Coverage	Reference Count	Reference Coverage
ELAVL1 (Hafner)[16]	AATTTT	129	12%	28	2%	CTTTTT	281	28%	45	4%	ATTTTT	282	28%	51	5%
ELAVL1 (Lebedeva)[17]	ACTTTT	260	17%	73	5%	ATTTTT	579	40%	136	9%	TCTTTT	377	26%	71	4%
ELAVL1 (Mnase)	AATTTT	138	13%	41	4%	TTTTTA	314	31%	41	4%	TTTTTC	243	24%	46	4%
ELAVL1 (Mukharjee)[19]	AATTTT	985	17%	186	3%	TTTTTA	2182	38%	273	4%	TTTTTC	1948	34%	175	3%
FUS[20]	AATAAA	120	7%	30	1%	TTAAAA	150	9%	38	2%	TTTATA	102	6%	26	1%
IGF2BP1-3[16]	ACTTCA	581	15%	283	7%	TCTTCA	701	18%	376	9%	TCCACA	430	11%	189	4%
PUM2[16]	AAATAT	197	19%	26	2%	TGTATA	348	34%	16	1%	TGTAAA	365	36%	19	1%
QKI[16]	ACTAAT	216	21%	7	0%	TTAACA	228	22%	19	1%	CTAACA	196	19%	7	0%
SFSR1[21]	AGGAGA	57	18%	7	2%	GGAAGA	72	23%	7	2%	AGAAGA	76	24%	3	0%
TAF15[20]	ACTTTT	63	6%	11	1%	TATTTA	73	7%	13	1%	CATTTC	61	6%	9	0%
TARDBP (iCLIP)[22]	ATGTGT	1885	39%	245	5%	TGTGTG	3701	77%	260	5%	TGTATG	1663	34%	155	3%
TIA1 (iCLIP)[23]	ACTTTT	123	12%	50	5%	TTTTTA	360	36%	121	12%	TTTTTC	263	26%	78	8%
TIAL1 (iCLIP)[23]	ATTTTT	711	33%	234	11%	CTTTTT	636	30%	170	8%	TGTTTT	503	23%	176	8%

Supplementary Table 2. Results of motifRG on CLIP data. The top 3 motifs discovered using motifRG (R package "motifRG"). The table shows for each motif the number of sequences containing the pattern, count and percentage for both foreground (positive) and background (reference) datasets.

Chapter V

Chaperone networks in *E. coli*

The following work focuses on characterizing the protein interactome of DnaK, the major bacterial chaperone Hsp70, in *E. coli*. We analysed differences in chaperone requirements by investigating the physico-chemical properties encoded in their target sequences. We observed that DnaK substrates bury amino acid residues from solvent less effectively than other DnaK associations, suggesting that enriched interactors populate dynamic intermediate states during folding and expose hydrophobic residues. Importantly, the analyses performed in this work constitute preliminary version of the algorithms presented in Chapter VI and VII, and in Klus et al. (2014).

Calloni G., Chen T., Schermann S.M., Chang H., Genevaux P., Agostini F., Tartaglia G.G., Hayer-Hartl M., Hartl F.U. (2012). DnaK Functions as a Central Hub in the *E. coli* Chaperone Network. *Cell Reports*, 3(1):251–264. PMID: 22832197 DOI: 10.1016/j.celrep.2011.12.007

Cell Reports
Resource

DnaK Functions as a Central Hub in the *E. coli* Chaperone Network

Giulia Calloni,[1,4] Taotao Chen,[1,4] Sonya M. Schermann,[1,5] Hung-chun Chang,[1,6] Pierre Genevaux,[2] Federico Agostini,[3] Gian Gaetano Tartaglia,[3] Manajit Hayer-Hartl,[1,*] and F. Ulrich Hartl[1,*]
[1]Department of Cellular Biochemistry, Max Planck Institute of Biochemistry, 82152 Martinsried, Germany
[2]Laboratoire de Microbiologie et Génétique Moléculaire, Centre National de la Recherche Scientifique and Université Paul Sabatier, F-31000 Toulouse, France
[3]Centre for Genomic Regulation and Universitat Pompeu Fabra, 08003 Barcelona, Spain
[4]These authors contributed equally to this work
[5]Present address: Micromet, Staffelseestr 2, 81477 Munich, Germany
[6]Present address: Department of Biology, Massachusetts Institute of Technology, Cambridge, MA 02139, USA
*Correspondence: mhartl@biochem.mpg.de (M.H.-H.), uhartl@biochem.mpg.de (F.U.H.)
DOI 10.1016/j.celrep.2011.12.007

SUMMARY

Cellular chaperone networks prevent potentially toxic protein aggregation and ensure proteome integrity. Here, we used *Escherichia coli* as a model to understand the organization of these networks, focusing on the cooperation of the DnaK system with the upstream chaperone Trigger factor (TF) and the downstream GroEL. Quantitative proteomics revealed that DnaK interacts with at least ~700 mostly cytosolic proteins, including ~180 relatively aggregation-prone proteins that utilize DnaK extensively during and after initial folding. Upon deletion of TF, DnaK interacts increasingly with ribosomal and other small, basic proteins, while its association with large multidomain proteins is reduced. DnaK also functions prominently in stabilizing proteins for subsequent folding by GroEL. These proteins accumulate on DnaK upon GroEL depletion and are then degraded, thus defining DnaK as a central organizer of the chaperone network. Combined loss of DnaK and TF causes proteostasis collapse with disruption of GroEL function, defective ribosomal biogenesis, and extensive aggregation of large proteins.

INTRODUCTION

In all cell types, molecular chaperones function in preventing protein misfolding and aggregation, typically by shielding hydrophobic surfaces exposed by proteins in their non-native states. Chaperones have essential roles in assisting the folding, assembly and transport of newly synthesized polypeptides and in surveying the conformational status of preexistent proteins (Hartl and Hayer-Hartl, 2009). Although detailed insights into the structure and mechanism of individual chaperone components have been obtained in recent years, how multiple chaperone modules cooperate to maintain conformational proteome integrity (proteostasis) is not yet understood (Balch et al., 2008). What is the degree of functional overlap and specificity among chaperones, and how robust is the network in tolerating disturbances and avoiding collapse? Although there is evidence that the complexity of proteostasis networks has increased during evolution (Gidalevitz et al., 2011), central players, such as the ATP-regulated Hsp70 chaperones, have been highly conserved from bacteria to human. Here we employed quantitative proteomics to analyze the chaperone network of *Escherichia coli* as a tractable model, focusing on the central role of the Hsp70 system.

DnaK, the major bacterial Hsp70, is one of the most abundant constitutively expressed and stress inducible chaperones in the *E. coli* cytosol. Yet it is not essential under nonstress conditions at intermediate temperature (Bukau and Walker, 1989). Indeed, DnaK (together with its co-chaperone DnaJ and regulator GrpE) cooperates in de novo protein folding with the ribosome-bound chaperone Trigger factor (TF). Although DnaK and TF can be deleted individually, their simultaneous deletion results in synthetic lethality at temperatures above 30°C (Deuerling et al., 1999; Genevaux et al., 2004; Teter et al., 1999). Under stress conditions, such as heat shock at 42°C, DnaK becomes indispensable (Bukau and Walker, 1989). TF and DnaK act upstream of the essential GroEL/ES chaperonin, which provides a cage-like compartment for the folding of single protein molecules, unimpaired by aggregation. About 10% of cytosolic proteins (~250 different proteins) have been found to interact with GroEL, of which a subset of ~50–85 proteins (so-called class III substrates) are absolutely GroEL/ES dependent for folding (Fujiwara et al., 2010; Kerner et al., 2005).

The ATP-dependent reaction cycle of DnaK is regulated by the Hsp40 co-chaperone DnaJ and the nucleotide exchange factor GrpE (reviewed in Hartl et al., 2011; Mayer, 2010). DnaJ functions in presenting non-native substrate proteins to DnaK (Figure 1A). Substrate binding and release by Hsp70 is achieved through the allosteric coupling of a N-terminal ATPase domain with a C-terminal peptide-binding domain, the latter consisting of a β sandwich subdomain and an α-helical lid segment. The β sandwich recognizes extended, ~7 residue segments enriched with hydrophobic amino acids, preferentially when they are framed by positively charged residues (Rüdiger et al., 1997; Zhu et al.,

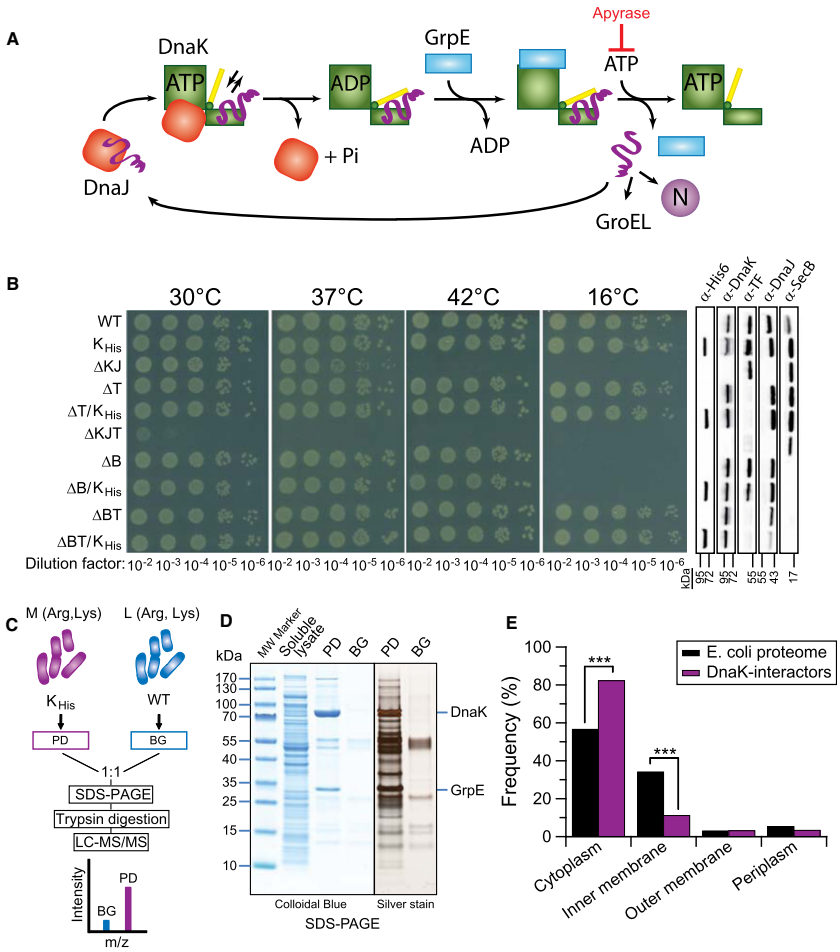


Figure 1. Isolation of DnaK-Interactor Complexes

(A) Schematic representation of the DnaK reaction cycle. Upon DnaJ-mediated delivery of non-native protein substrate to ATP-bound DnaK, hydrolysis of ATP to ADP results in closing of the α-helical lid (yellow) and tight binding of substrate by DnaK. Stable DnaK-substrate complexes are accumulated by depleting ATP with apyrase upon cell lysis.

(B) In vivo functionality of the chromosomally encoded DnaK-His6. The *dnaK* gene (WT) was replaced with *dnaK-His6* (K_{His}) in MC4100 and where indicated in the isogenic chaperone mutant strains Δ*dnaKdnaJ* (ΔKJ), Δ*tig* (ΔT), Δ*tig/dnaK-His6* (ΔT/K_{His}), Δ*dnaKdnaJΔtig* (ΔKJT), Δ*secB* (ΔB), Δ*secB/dnaK-His6* (ΔB/K_{His}), Δ*secBΔtig* (ΔBT), and Δ*secBΔtig/dnaK-His6* (ΔBT/K_{His}) as described in Extended Experimental Procedures. Cells in mid-log phase were serially diluted, spotted on LB agar plates and incubated for 1 day at 30°C, 37°C, or 42°C and for 5 days at 16°C.

(C) Schematic of the SILAC approach used to identify DnaK interactors. L (light) and M (medium) Arg, Lys isotope media. DnaK-substrate complexes isolated from M-labeled cells containing the DnaK-His6 (pulldown, PD) were mixed 1:1 with L-labeled proteins isolated with the same procedure from an equal amount of cell lysate containing non-tagged DnaK (background, BG). The mixture was subsequently separated by SDS-PAGE, followed by in-gel trypsin digestion and LC-MS/MS analysis.

(D) Isolation of DnaK-substrate complexes. Soluble lysate, PD and BG fractions were analyzed by 4%–12% gradient SDS-PAGE, followed by Colloidal Blue or silver staining, as indicated.

(E) Cellular localization of DnaK interactors compared to the genome-based *E. coli* proteome (Yu et al., 2011). ***$p \leq 0.001$ based on χ^2 test.

1996). The α-helical lid and a conformational change in the β sandwich domain regulate the affinity state for peptide through an ATP-dependent, allosteric mechanism (Zhuravleva and Gierasch, 2011). In the ATP-bound state, the lid adopts an open conformation, resulting in high on- and off-rates for peptide (Figure 1A). Hydrolysis of ATP to ADP is accelerated by DnaJ, leading to lid closure and stable peptide binding. Following ATP-hydrolysis, DnaJ dissociates and GrpE binds to the DnaK ATPase domain, catalyzing ADP release. Binding of ATP then results in lid-opening and substrate release for folding or transfer to other chaperones (Figure 1A).

Attempts to identify DnaK substrates have been limited to the analysis of proteins that aggregated in cells lacking TF upon depletion of DnaK (Deuerling et al., 2003). Here we developed an approach for the direct isolation of DnaK-substrate complexes and their identification by quantitative proteomics from wild-type, TF-deleted, or GroEL-depleted cells. In parallel, we analyzed global proteome changes under conditions of single or combined chaperone deletion. Our measurements show that DnaK normally interacts with at least ~700 newly synthesized and preexistent proteins, which we characterized based on their relative enrichment on DnaK. Individual deletion of TF or depletion of GroEL/ES leads to specific changes in the DnaK interactome and in global proteome composition. These effects are highly informative as to the functional cooperativity of chaperone modules. We conclude that DnaK is the central hub in the cytosolic *E. coli* chaperone network, interfacing extensively with the upstream TF and the downstream chaperonin. The functional interconnection of these major chaperone systems is critical for robust proteostasis control.

RESULTS

Isolation and Identification of the DnaK Interactome

To isolate DnaK-substrate complexes, we generated an *E. coli* MC4100 strain in which the wild-type *dnaK* gene was replaced by *dnaK-His6*, encoding DnaK with a C-terminal His6-tag (henceforth called K_{His}). K_{His} cells grew like wild-type (WT) on agar plates or in liquid culture at 30°C–37°C, or under heat shock conditions at 42°C (Figure 1B and Figure S1A available online). Quantitative proteomic analysis using SILAC (stable isotope labeling with amino acids in cell culture) (Ong et al., 2002) did not detect significant differences in protein abundance between WT and K_{His} cells (Figure S1B). DnaK-His6 also supported normal growth of TF-deleted (ΔT) cells above 30°C and of cells lacking the protein export chaperone SecB (ΔB) (Figure 1B and Figure S1A) (Deuerling et al., 1999; Genevaux et al., 2004; Smock et al., 2010; Teter et al., 1999). ΔB cells are cold-sensitive and this defect is compensated by deletion of TF (Ullers et al., 2007).

We isolated DnaK interactors by immobilized metal affinity chromatography (IMAC) from K_{His} cells growing exponentially at 37°C. DnaK-substrate complexes were stabilized during cell lysis by rapidly (within <10 s) depleting ATP with apyrase to inhibit substrate cycling (Figure 1A) (Teter et al., 1999). The K_{His} cells were SILAC-labeled with medium (M) Arg/Lys isotopes, lysate prepared and subjected to IMAC pulldown (PD). DnaK-His6 complexes were mixed 1:1 with a background (BG) sample obtained by the same IMAC procedure from lysate of unlabeled WT cells (light isotopes, L) (Figure 1C and Extended Experimental Procedures). The composition of PD and BG samples prior to mixing is shown in Figure 1D. Note that GrpE was co-isolated with DnaK as a stoichiometric complex (Figures 1A and 1D), whereas DnaJ was present in substoichiometric amounts (Table S1). The molar ratio of DnaK/GrpE/DnaJ was ~30:20:1, as estimated from the number of peptides identified by MS using exponentially modified Protein Abundance Index (emPAI) scores (Ishihama et al., 2005).

A total of 674 DnaK interactors (Table S1) were identified by LC-MS/MS with >95% confidence (Figure S1C and Extended Experimental Procedures), including proteins either not identified in BG samples or having a >2-fold enrichment (M/L ratio) in PD over BG in at least two of three independent experiments (biological repeats). Most of these (503 proteins) were >4-fold enriched over BG (Table S1). The identification of DnaK interactors approached saturation in consecutive experiments (Extended Experimental Procedures). For the vast majority of interactors (>95%), the amount of protein co-isolated with DnaK was strongly diminished upon incubation of cell lysate with ATP (Figure S1D), indicating that these proteins bind DnaK in an ATP-regulated manner. While GrpE was released, DnaJ was enriched on DnaK in the presence of ATP (data not shown) (Figure 1A). The DnaJ homolog cbpA and the two small heat shock proteins, IbpA and IbpB (Hsp20), were also identified in DnaK pulldowns in the presence of ATP, suggesting that these chaperones functionally cooperate with DnaK (data not shown).

Approximately 80% of the DnaK interactors are predicted to be cytosolic, ~11% are inner membrane proteins, ~3% outer membrane proteins, and ~3% are located in the periplasm (Figure 1E and Table S1). Thus, the identified DnaK interactors comprise ~25% of the cytosolic proteome. As a collective, they are similar to a set of 1,938 proteins identified in soluble cell lysates (list available at the Proteome Commons Tranche repository) in terms of molecular weight (Figure S1E) and other physico-chemical properties, such as isoelectric point, average hydrophobicity, and aggregation propensity (data not shown), indicating that DnaK has a broad substrate specificity.

Classification of DnaK Substrates by Enrichment on DnaK

We next analyzed each of the different DnaK interactors to determine what fraction of the total protein is DnaK-bound, assuming that this parameter correlates with chaperone dependence, as was observed for the substrates of GroEL (Kerner et al., 2005). Unlabeled soluble cell lysate (L) was mixed at a defined proportion with DnaK complexes isolated from cells labeled with heavy Arg/Lys isotopes (H) and H/L ratios were determined by LC-MS/MS. H/L ratios were obtained for 666 DnaK interactors, reflecting their relative enrichment on DnaK. The relative enrichment factors (REF) displayed a broad, bimodal distribution (Figure 2A). By setting thresholds at the 20th and 70th percentiles of the distribution, we grouped 142 proteins as less-enriched, 183 proteins as enriched, and 341 proteins as medium enriched on DnaK (Figure 2A and Table S1). Interestingly, the enriched proteins are below average in cellular abundance but together make up ~40% of all identified DnaK interactors by mass

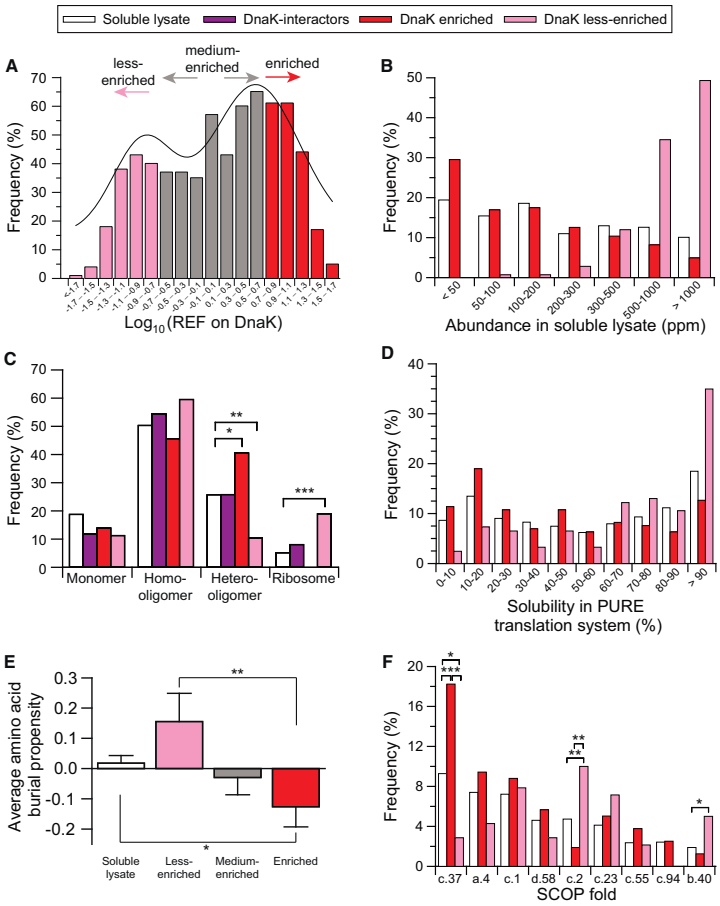


Figure 2. Classification and Characterization of DnaK Interactors

(A) Relative enrichment of interactor proteins on DnaK. The histogram shows the distribution of relative enrichment factors (REF) for 666 DnaK interactors identified in 3 independent experiments. REF indicates the fraction of total cellular protein bound to DnaK. Enriched and less-enriched sets of interactors were selected at the extremes of the distribution for further analysis as described in Extended Experimental Procedures.

(B–F) Properties of enriched and less-enriched DnaK interactors compared to soluble lysate proteins. (B) Abundance in soluble lysate determined based on cumulative abundance values (emPAI) (Ishihama et al., 2005). (C) Oligomeric state of all, the enriched and less-enriched DnaK interactors compared to lysate proteins. Ribosomal proteins are analyzed separately. (D) Solubility upon in vitro translation in the absence of chaperones (Niwa et al., 2009). (E) Average propensity of soluble lysate proteins and DnaK interactors to bury amino acid residues from solvent, calculated using the burial propensity scale of specific amino acids by Janin (1979) (see Extended Experimental Procedures). Shown is the mean burial propensity for each class; error bars correspond to the SEM. P values based on Mann-Whitney test: *$p \leq 0.05$; **$p \leq 0.01$. (F) SCOP fold distribution. c.37, P loop containing nucleotide triphosphate hydrolases; a.4, DNA/RNA binding 3-helical bundle; c.1, TIM β/α barrel; d.58, Ferredoxin-like; c.2, NAD(P)-binding Rossmann-fold domains; c.23, Flavodoxin-like; c.55, Ribonuclease H-like motif; c.94, Periplasmic binding protein-like II; b.40, OB-fold. Statistical significance for categorical variables is based on a χ^2 test: *$p \leq 0.05$; **$p \leq 0.01$; ***$p \leq 0.001$.

(Figure 2B and Table S1), as based on emPAI scores. For these proteins ~5% of cellular content is bound to DnaK. In contrast, the less-enriched DnaK interactors are highly abundant proteins but amount to only ~13% of all DnaK interactors by mass (Figure 2B and Table S1). On average, only 0.1% or less of their cellular content is DnaK-bound. The medium enriched

interactors amount to ~47% of all DnaK interactors by mass, with ~1% of cellular content being bound.

The DnaK-enriched proteins cover a wide range of cellular functions, prominently including DNA replication, recombination and repair (COG class L), and cell division and chromosome partitioning (COG class D) (Figure S2A and Table S1). The less-enriched DnaK interactors have a significant preference for proteins involved in translation, ribosomal structure, and biogenesis (COG class J) and include 24 ribosomal proteins. Essential proteins (Gerdes et al., 2003) are more frequent among the less-enriched substrates (Figure S2B). The enriched interactors include some very large proteins >100 kDa (Figure S2C) and tend to have more predicted DnaK binding sites than the less enriched substrates (both absolute and when corrected for size) (Figure S2D) (Van Durme et al., 2009). Furthermore, they are more frequently part of heterooligomeric complexes (when ribosomal proteins are considered separately) (Figure 2C).

To determine whether the proteins enriched on DnaK have a high propensity to aggregate, we took advantage of the study by Niwa et al. (2009) who analyzed the solubility of *E. coli* proteins upon in vitro translation in the absence of chaperones. Indeed, the DnaK-enriched proteins are more aggregation prone upon translation than average proteins of soluble cell lysate, whereas the less-enriched DnaK interactors are more soluble (Figure 2D). This is consistent with the finding that the enriched proteins frequently display pI values close to neutral pH (Figure S2E). Moreover, these proteins are predicted to bury amino acid residues less effectively from solvent than the less enriched DnaK interactors and average soluble proteins (Figure 2E) (Tartaglia et al., 2010). This suggests that the enriched interactors populate dynamic intermediate states during folding and expose hydrophobic residues, although their average sequence hydrophobicity is not increased (data not shown). Similar properties were previously found for the obligate (class III) GroEL substrates (Figure 2E) (Raineri et al., 2010; Tartaglia et al., 2010). Interestingly, ~18% (29 proteins) of the DnaK-enriched substrates with assigned fold (159 proteins) contain at least one domain with SCOP fold c.37 (P loop containing nucleoside triphosphate hydrolases), compared to only ~8% of soluble lysate proteins and ~3% of the less-enriched DnaK interactors (Figure 2F). The c.37 fold is characterized by a complex α/β topology and is highly represented in heterooligomeric proteins (Figure S2F). The less-enriched substrates have a preference for the SCOP folds c.2 (NAD(P)-binding Rossmann-fold) and b.40 (OB-fold) (Figure 2F), which are found in abundant metabolic enzymes and in ribosomal proteins, respectively.

In summary, the relative enrichment of proteins on DnaK correlates with their propensity to aggregate during folding. The ~180 most enriched DnaK interactors amount to ~40% of total mass of DnaK substrates. They are of relatively low cellular abundance, tend to contain more predicted DnaK binding sites than the less-enriched interactors, and frequently assemble with other proteins to heterooligomeric complexes.

Effects of DnaK Deletion at the Proteome Level

To analyze the global effects of deleting the DnaK chaperone system, we performed quantitative proteomics of Δ*dnaKdnaJ* (ΔKJ) cells (H-labeled) in comparison to K_{His} cells (M-labeled). The cells were grown at 30°C where deletion of DnaK/DnaJ is well tolerated in liquid culture (Figure 1B and Figure S1A). Out of ~1,400 proteins quantified, 105 proteins were increased in abundance in ΔKJ cells (Table S2A). These proteins include 42 identified DnaK interactors, presumably reflecting a compensatory response. In addition, the major cytosolic chaperones and proteases (GroEL/ES, HtpG, IbpA, IbpB, ClpB, Hsp33, HslU, HslV, Lon) were upregulated 5- to >10-fold, consistent with a 4.5-fold increase in abundance of the central heat shock regulator rpoH (σ32), which is negatively controlled by DnaK and DnaJ (Table S1) (Gamer et al., 1992; Straus et al., 1990). Indeed, the genes of 56 of the 105 upregulated proteins contain known or putative σ32 binding sites in their upstream regions (Zhao et al., 2005) (Table S2A). Interestingly, 87 proteins were reproducibly reduced in abundance in ΔKJ cells by ~40%–95% (median 60%; $p < 0.05$) (Table S2B). SILAC pulse-labeling demonstrated similar rates of synthesis in K_{His} and ΔKJ cells (Figure S2G), indicating that the observed decrease in protein abundance was largely due to degradation. Notably, among the degraded proteins were 40 identified DnaK interactors (four essential proteins), including the periplasmic chaperones of acid-denatured proteins hdeA and hdeB, and several amino acid metabolic enzymes. The DnaK interactors that are degraded in ΔKJ cells are above average enriched on DnaK in DnaK-His6 cells (Table S2B) and exhibit relatively low solubility upon in vitro translation (Figure S2H) (Niwa et al., 2009).

To determine whether proteins aggregate in cells lacking DnaK/DnaJ, we analyzed the insoluble and soluble fractions of ΔKJ cells compared to K_{His} cells. In total, 474 proteins were significantly increased in the insoluble fraction of ΔKJ cells, including 201 identified DnaK interactors (Table S3). However, only 65 proteins were substantially depleted from the soluble fraction by 5%–90% (median ~9%; $p < 0.05$) due to aggregation. Among this group of proteins were 30 identified DnaK interactors (eight essential proteins), such as excision nuclease subunit A (uvrA) (Table S3). The extent of aggregation correlated strongly with enrichment on DnaK (Table S3).

These findings demonstrate that a subset of DnaK interactors are specifically dependent on the DnaK system in vivo. These proteins tend to be degraded or aggregate in ΔKJ cells, even at a growth temperature of 30°C, where the loss of DnaK is otherwise well compensated.

DnaK Interacts with Newly Synthesized and Preexistent Proteins

To determine whether proteins interact with DnaK only during initial folding or return to DnaK later for conformational maintenance (newly synthesized and preexistent interactors, respectively), we performed pulse-SILAC experiments. K_{His} cells grown at 37°C with unlabeled Arg/Lys (L) were shifted to M-labeled Arg/Lys for 2.5 min to label newly synthesized polypeptides. WT cells grown in heavy (H) Arg/Lys served as background control. Lysates from the K_{His} and WT cells were mixed 1:1 and DnaK interactors identified. The M/L ratio of a DnaK interactor relative to its M/L ratio in the lysate (the latter correcting for rates of synthesis and turnover) was used to indicate whether it bound to DnaK preferentially as a newly synthesized or preexistent protein (Figure 3A). Isotope ratios were obtained

Cell PRESS
Open ACCESS

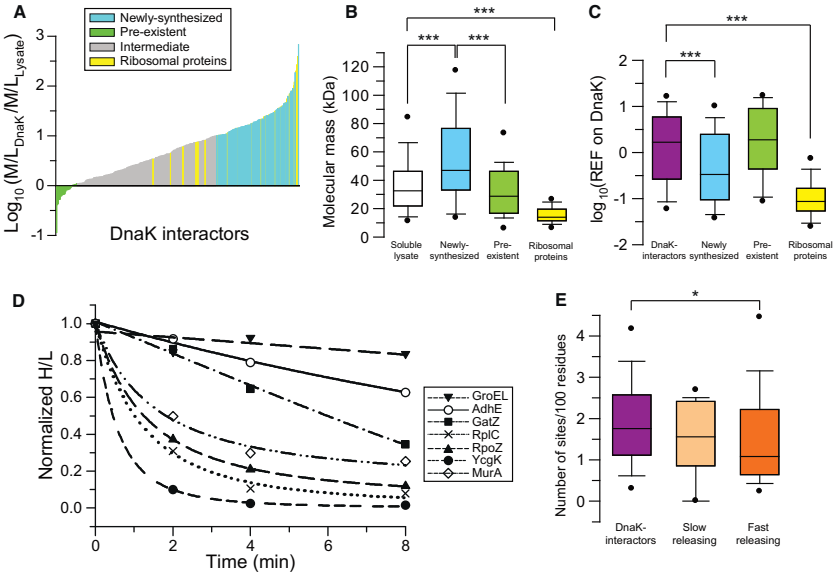


Figure 3. Characterization of Proteins that Interact with DnaK upon Synthesis or for Conformational Maintenance

(A–C) Analysis of DnaK interactors classified by pulse-SILAC as preexistent or newly synthesized proteins. (A) Ratios of medium to light isotopes (M/L) of DnaK interactors relative to the M/L ratios for the same proteins in soluble cell lysate. Positive values of the log transformed ratio of ratios indicate a preferential interaction with DnaK as newly synthesized proteins. Groups of proteins are color coded: blue and green, strong tendency to interact as newly synthesized or preexistent proteins; gray, intermediate tendency to interact as newly synthesized proteins; yellow, ribosomal proteins. Molecular weight (B) and relative enrichment factors (REF) on DnaK (C) of the substrates preferentially interacting as preexistent or newly synthesized polypeptides as compared to *E. coli* soluble lysate proteins and all DnaK interactors, respectively. The ribosomal proteins among DnaK interactors are analyzed separately. Horizontal line indicates the median, whisker caps and circles indicate 10th/90th and 5th/95th percentiles, respectively. P values based on Mann-Whitney test: ***$p \leq 0.001$.

(D and E) Time-dependent dissociation of proteins from DnaK as determined by pulse-chase SILAC. (D) Kinetics of dissociation from DnaK shown for selected proteins. Data were fitted to exponential decay. (E) Distribution of the number of predicted DnaK binding sites (Van Durme et al., 2009) for the DnaK-interactor sets with fast and slow release kinetics as compared to all DnaK interactors. Horizontal line indicates the median, whisker caps and circles indicate 10th/90th and 5th/95th percentiles, respectively. P value based on Mann-Whitney test: *$p \leq 0.05$.

for ~300 DnaK interactors (Table S4). Log_{10} M/L ratios lower than 0, indicating a strong preference for interaction as preexistent protein, were observed for only 20 interactors (Figure 3A, green). Most other proteins bound to DnaK upon synthesis (Figure 3A, gray and blue), including ~100 proteins with a strong preference for interaction as newly synthesized polypeptides (log_{10} M/L ratio $\geq$ 1) (Figure 3A, blue). Ribosomal proteins are included in this group (Figure 3A, yellow), in support of the proposed role of DnaK in ribosome assembly (Maki et al., 2002; René and Alix, 2011).

We next compared the physico-chemical properties of the newly synthesized and preexistent DnaK interactors (Figure 3A). Ribosomal proteins were analyzed separately, as their unusual size and charge properties would introduce a strong bias. Interestingly, the 71 proteins with a strong preference to interact with DnaK upon synthesis are significantly shifted to large sizes and thus are likely to have complex folding pathways (Figure 3B). They are of average to above average cellular abundance and comparable to lysate proteins in terms of hydrophobicity and aggregation propensities, as calculated with the Z_{agg} algorithm based on their amino acid sequence properties (Tartaglia et al., 2008) (Figures S3A and S3B). Interestingly, only a few of these proteins aggregated in ΔKJ cells (Table S3), suggesting that they can utilize multiple chaperones for folding or, perhaps less likely, have only low chaperone dependence. As expected, the ribosomal proteins have very low aggregation scores, consistent with their charge character (Figure S3B). Furthermore, the tendency to interact with DnaK upon synthesis correlates with a relatively lower enrichment on DnaK, suggesting that these proteins interact only transiently (Figure 3C). On the other hand, the preexistent interactors are similar in size to the average of lysate proteins (Figure 3B) and are more enriched on DnaK than the newly synthesized substrates (Figure 3C), arguing for longer residence times on DnaK. They are also characterized by higher intrinsic aggregation propensities than the newly synthesized interactors ($p < 0.05$) (Figure S3B) and several of

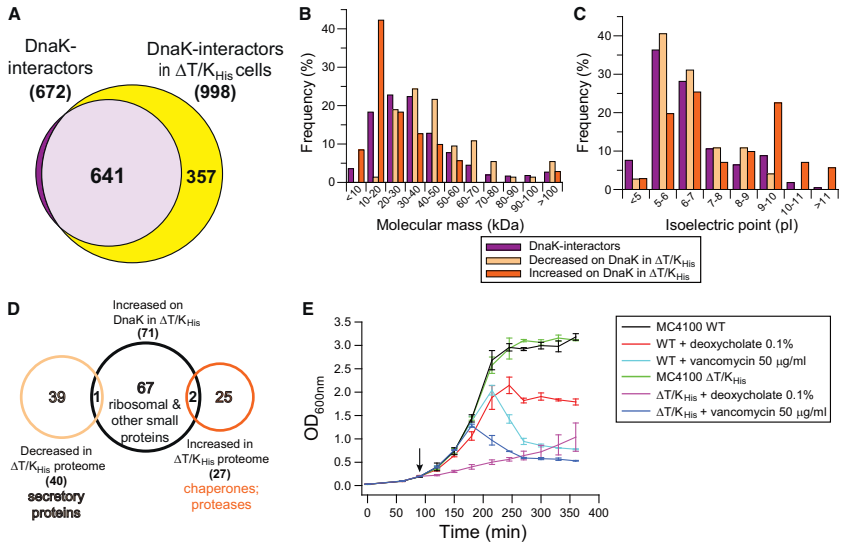


Figure 4. Effect of TF Deletion on the DnaK Interactome and Cellular Proteome

(A) Overlap of DnaK interactors in K_{His} (672 proteins) and $\Delta T/K_{His}$ cells (998 proteins).

(B and C) Change in the distribution of molecular weight (B) and isoelectric point (C) of the DnaK interactome in $\Delta T/K_{His}$ cells compared to all DnaK interactors in K_{His} cells.

(D) Minimal overlap of the sets of proteins significantly decreased or increased in abundance in the proteome of $\Delta T/K_{His}$ cells with the set of proteins that increased on DnaK in $\Delta T/K_{His}$ cells.

(E) Outer membrane destabilization in $\Delta T/K_{His}$ cells. Growth curves at 37°C of K_{His} and $\Delta T/K_{His}$ cells in the presence and absence of 0.1% deoxycholate or 50 μg/ml vancomycin. Arrow indicates the time of addition of deoxycholate or vancomycin after dilution of an overnight culture in fresh M63 medium at an OD_{600nm} of 0.025. Error bars represent SD of three independent measurements.

these proteins were found to be degraded in ΔKJ cells (Table S2B). Thus, a prominent feature of the preexistent interactors is their intrinsic tendency to populate aggregation-prone states, consistent with a requirement for conformational maintenance by DnaK.

Protein Flux through DnaK

To measure the residence time of interactors on DnaK directly, K_{His} cells were pulse-labeled with Arg/Lys isotopes (H) for 2.5 min, followed by a chase with excess unlabeled amino acids (L). Time course data based on the time-dependent decrease of the H/L ratio were collected for 91 proteins, with apparent dissociation rates from DnaK corresponding to half-times of ~30 s to ~25 min (Figure 3D and Table S5). Proteins with slow release rates (≤30th percentile of the rate distribution; 27 proteins) include GroEL, IbpA, and SecB, presumably reflecting the functional cooperation of these chaperones with DnaK. Slow releasing proteins (excluding chaperones) are characterized by above average enrichment on DnaK and an average number of predicted binding sites compared to all DnaK interactors (Figure 3E). Proteins with fast release rates (≥70th percentile of the rate distribution; 27 proteins) display average enrichment on DnaK and tend to have a lower number of binding sites (Figure 3E). Most of these substrates are predicted to bury hydrophobic regions effectively (data not shown). Thus, the residence time (and consequently the enrichment) on DnaK appears to be regulated, at least in part, by the frequency of potential DnaK recognition motifs in the polypeptide chain and by the efficiency of their burial during folding. Fast release rates correlate generally with the tendency of proteins to interact with DnaK only upon synthesis, whereas proteins that utilize DnaK also for maintenance have longer residence times.

Partial Functional Redundancy of DnaK and TF

To understand how DnaK interfaces with other modules of the chaperone network, we first investigated how the DnaK interactome changes upon deletion of the upstream chaperone TF. We identified the DnaK interactors in the TF deletion strain at 37°C by pulldown from H-labeled Δ*tig*/*dnaK-His6* ($\Delta T/K_{His}$) cells using L-labeled ΔT cells as the background. The number of identified DnaK interactors increased to 998 (see Proteome Commons Tranche repository), including ~95% of the 672 DnaK interactors identified in K_{His} cells (Figure 4A). (DnaK levels increased ~1.4-fold in $\Delta T/K_{His}$ cells; see Table S6D below.) The additional

proteins have physico-chemical properties similar to the DnaK interactors in WT cells (data not shown). To detect quantitative changes in the DnaK interactome upon TF deletion, we performed a comparative analysis of H-labeled ΔT/K_{His} and M-labeled K_{His} cells. Seventy-one proteins interacted to a significantly greater extent (1.4- to 3.3-fold) and 74 proteins to a lesser extent with DnaK in ΔT/K_{His} cells (Tables S6A and S6B). The proteins accumulating on DnaK are typically small in size (<20 kDa) (Figure 4B) and include 17 highly basic ribosomal proteins as well as ten basic nonribosomal proteins (pI ≥ 9) (Figure 4C), such as several secretory proteins with functions in cell envelope and outer membrane biogenesis (Skp, AmpH, YcgK, SlyB) (Table S6A). The proteins that interacted less extensively with DnaK in ΔT/K_{His} cells are larger in size (Figure 4B) and overlap with the large newly synthesized DnaK interactors described above (Figure 3B). Apparently, they bind to DnaK less efficiently in the absence of TF (Figure 4B).

We also performed a proteomic analysis of total cell lysate of ΔT/K_{His} cells as compared to K_{His} cells. Only 40 of 1,490 quantified proteins were reproducibly reduced in abundance by 25%–85% ($p < 0.05$) when TF was deleted (Table S6C), although their rates of synthesis were unchanged (Figure S4); 27 proteins were increased in abundance (Figure 4D and Table S6D). Both these groups of proteins show minimal overlap with the proteins that accumulated on DnaK upon TF deletion (Figure 4D). Strikingly, most of the proteins that decreased in ΔT/K_{His} cells carry predicted signal sequences (Table S6C). These proteins include numerous outer membrane β-barrel proteins, suggesting that TF has a specific role in outer membrane protein biogenesis. Indeed, ΔT/K_{His} cells proved to be sensitive to treatment with detergents like deoxycholate and the antibiotic vancomycin, which is indicative of a weakening of the outer membrane (Figure 4E) (Nichols et al., 2011). The proteins that increased in ΔT/K_{His} cells include cytosolic chaperones and proteases, as well as the ATPase SecA required for membrane translocation of outer membrane proteins (Table S6D). However, these proteins are only moderately upregulated (~1.5-fold), suggesting that loss of TF function at 37°C causes only limited proteome stress.

These results indicate a functional redundancy of TF and DnaK in the folding/assembly of ribosomal and other small, basic proteins. In addition, TF has a specific role in the biogenesis of outer membrane β-barrel proteins. This function cannot be performed by DnaK and is only partially replaced by other chaperones, resulting in outer membrane destabilization.

Interplay of DnaK and GroEL/ES

GroEL and its cofactor GroES are upregulated in both the ΔKJ and ΔT/K_{His} cells (Tables S2 and S6D), suggesting that these chaperone systems form a functional network. Strikingly, 119 of the identified DnaK interactors are known GroEL substrates (Fujiwara et al., 2010; Kerner et al., 2005), and together these proteins amount to ~30% of all DnaK interactors by mass. The number of GroEL substrates on DnaK increases to 152 in ΔT/K_{His} cells (Figure 5A, Table S1, and Proteome Commons Tranche repository). Forty-two of these DnaK interactors are obligate GroEL-dependent (class III) and thus must be delivered to GroEL for folding, whereas 80 belong to class II and 30 to class I (Figure 5B). Class II substrates are highly chaperone dependent but can utilize either GroEL/ES or DnaK/DnaJ for folding in vitro, whereas class I proteins have a lower chaperone dependence (Kerner et al., 2005). About 90 of the previously identified GroEL substrates were not detected on DnaK. Many of these proteins are of low abundance and thus may have very low steady-state levels on DnaK.

To investigate how the depletion of GroEL/ES affects the spectrum of DnaK interactors, K_{His} cells carrying the *groELS* operon under arabinose control were shifted from arabinose (LS+/K_{His}) to glucose for 3.5 hr (LS−/K_{His}) at 37°C, which resulted in ~97% depletion of GroEL/ES (Kerner et al., 2005; McLennan and Masters, 1998). Note that the cells grow normally during the first 5 hr of GroEL/ES depletion (data not shown). Ninety-two proteins increased on DnaK (2- to 60-fold) upon GroEL depletion and 54 proteins decreased (Tables S7A and S7B). The former include 38 GroEL substrates (19 class III proteins) (Figure 5B). They are enriched in domains with SCOP fold c.1 (TIM-barrel) (Figure S5A), which is prominently represented among obligate GroEL substrates (Fujiwara et al., 2010; Kerner et al., 2005) (Table S7A). The proteins depleted from DnaK include 11 GroEL substrates, mostly of class II (Table S7B), suggesting that they are partially displaced from DnaK by class III substrates that are unable to fold.

We also analyzed the consequences of GroEL depletion at the proteome level. Depletion of GroEL resulted in a ~35%–95% decrease in abundance of 114 proteins and a ≥2-fold increase of 95 proteins (Tables S7C and S7D). The former include 44 GroEL substrates (24 class III proteins) (Figure 5B and Table S7C) that are apparently degraded (Figure S5B). Strikingly, 18 of these GroEL substrates nevertheless accumulated on DnaK (Figure 5C), suggesting that they are stabilized by DnaK for transfer to the degradation machinery. The proteins that are upregulated upon GroEL depletion include chaperones and proteases (upregulated ~2-fold) as well as 19 GroEL substrates (Table S7C). In addition to degradation, loss of GroEL function at 37°C also resulted in substantial aggregation of many obligate GroEL substrates (Figure 5D) (Chapman et al., 2006; Kerner et al., 2005).

In summary, GroEL substrates interact extensively with DnaK. Upon depletion of GroEL, obligate GroEL substrates accumulate further on DnaK and are either transferred to the degradation machinery or eventually aggregate.

Proteostasis Collapse upon Combined Deletion of DnaK and TF

Prevention of protein aggregation is considered fundamental to proteostasis maintenance (Hartl et al., 2011). To define the relative contribution of the different chaperone systems to aggregation prevention, we performed a comparative analysis of the aggregated proteomes upon individual and combined chaperone deletion at 30°C, where cells lacking DnaK/DnaJ and TF (ΔKJT) still grow, albeit slowly (Figure 1B and Figure S1A). As described above, 474 proteins increased significantly in the insoluble fraction of ΔKJ cells, compared to only 15 proteins in ΔT/K_{His} cells and 33 proteins in GroEL/ES depleted (LS−/K_{His}) cells (Figure 6A and Table S8), indicating a major role of the DnaK system in aggregation prevention. Upon combined deletion of DnaK/DnaJ and TF, 1,087 proteins aggregated, including 403 DnaK interactors identified in this study (Figure 6A and

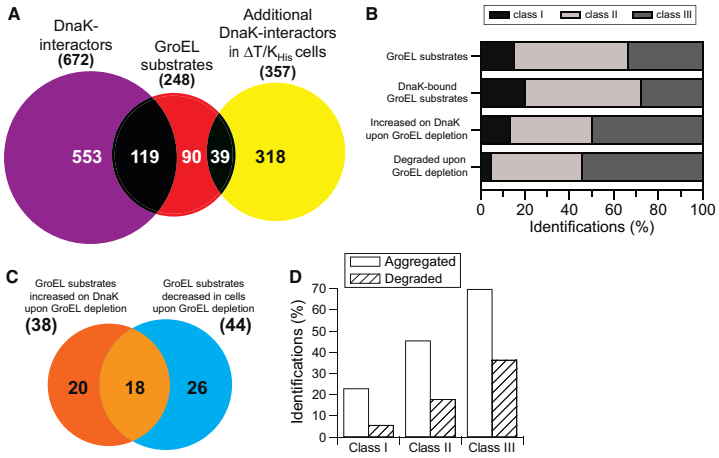


Figure 5. GroEL Substrates Accumulate on DnaK in GroEL-Depleted Cells

(A) Overlap of previously identified GroEL substrates (248 proteins) (Kerner et al., 2005) with DnaK interactors in K_{His} (672 proteins) and $\Delta T/K_{His}$ (357 additional DnaK interactors).

(B) Class distribution of the GroEL substrates (Kerner et al., 2005) compared to the class distribution of GroEL substrates bound to DnaK in K_{His} cells, increased on DnaK upon GroEL-depletion and class distribution of GroEL substrates that are partially degraded in total cell lysate upon GroEL-depletion. Class I, chaperone independent; class II, chaperone dependent; class III, obligate GroEL substrates.

(C) Overlap of the GroEL substrates that are significantly decreased in the proteome of LS−/K_{His} cells due to degradation and that increase on DnaK upon GroEL-depletion.

(D) Class distribution of GroEL substrates identified at 37°C in the aggregate fraction of LS−/K_{His} cells and of GroEL substrates decreased in the total proteome of LS−/K_{His} cells due to degradation.

Table S8). The aggregated proteins also included 149 of a total of 196 proteins shown previously to aggregate in cells lacking DnaK/DnaJ and TF (Deuerling et al., 2003; Martinez-Hackert and Hendrickson, 2009). Moreover, whereas in ΔKJ cells only 65 proteins aggregated substantially (>5% depletion from the soluble fraction), the number of substantially aggregated proteins increased 4-fold upon additional loss of TF (Figure 6A and Table S8). Strikingly, the size distribution of these proteins showed a strong shift to large proteins >50 KDa, frequently containing multiple domains (Figure 6B and Table S8). Many of these proteins (~50%) are identified DnaK substrates of average or above average abundance (Table S1) and represent a subset of the large proteins that interact with DnaK on synthesis (Figure 3B). Apparently, these proteins are chaperone-dependent but can utilize either the DnaK system or TF for efficient folding.

Remarkably, ~70% of the previously identified GroEL substrates (167 proteins) (Kerner et al., 2005) were also recovered in the aggregate fraction of ΔKJT cells (Figure 6A), many of which aggregated substantially (Table S8). This effect was observed despite a ~10-fold upregulation of GroEL/ES (data not shown). Aggregation of GroEL substrates was essentially undetectable in ΔKJ or $\Delta T/K_{His}$ cells at 30°C (Table S8). Thus, substrate delivery to GroEL critically depends on the upstream chaperones but can be performed by either DnaK or TF.

A large number of ribosomal proteins also accumulated in the insoluble fraction of ΔKJT cells, almost all of which were identified as DnaK interactors (Table S8). Deletion of DnaK/DnaJ alone resulted in the aggregation of only three ribosomal proteins, whereas none of these proteins aggregated in the $\Delta T/K_{His}$ cells. Although aggregation did not cause a significant depletion of ribosomal proteins from the soluble fraction, these results suggested that ΔKJT cells have a defect in ribosomal biogenesis. Indeed, numerous ribosomal proteins were reduced in abundance by 10%−30% in total lysates of ΔKJT cells compared to K_{His} cells and the single chaperone deletions (Figure 6C). Interestingly, two small ribosomal proteins (RpsQ S17 and RpsT S20) were increased in abundance, an effect that was also detected in ΔKJ cells (Figure 6C).

In summary, the combined loss of DnaK/DnaJ and TF results in a pronounced proteostasis collapse characterized by the aggregation of large, multidomain proteins, disruption of proper protein flux through GroEL/ES and defective ribosomal biogenesis. These findings explain the strong growth defect of ΔKJT cells at 30°C and their inability to grow at higher temperatures.

DISCUSSION

The DnaK Interactome

The DnaK interactome characterized here comprises at least ~700 proteins in WT cells and ~1,000 proteins in TF-deleted cells, demonstrating the pervasive role of the Hsp70 chaperone system in protein folding and proteostasis. While the vast

Cell PRESS
Open ACCESS

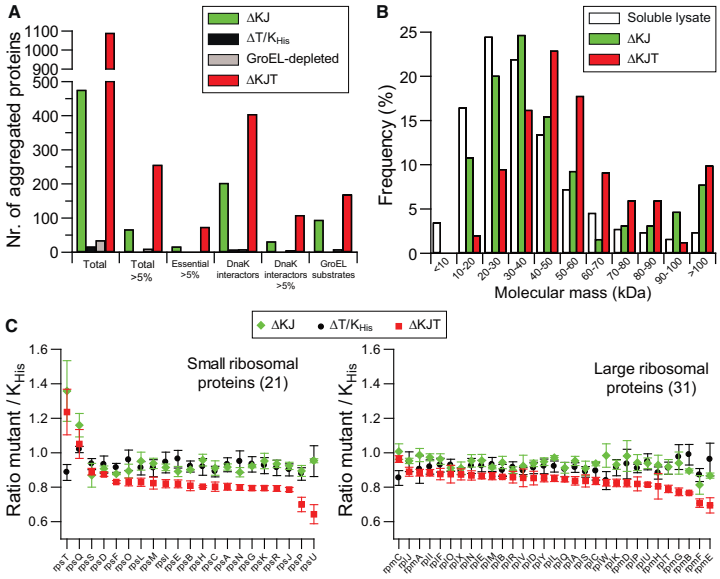


Figure 6. Proteostasis Collapse in Δ*dnaKdnaJ*/Δ*tig* Cells

(A) Protein aggregation in ΔKJ, ΔT/K$_{His}$, LS−/K$_{His}$, and ΔKJT cells. The number of aggregated proteins were analyzed according to the following categories: total, aggregated to >5% (Total > 5%), essential proteins aggregated to >5% (Essential > 5%), DnaK interactors, DnaK interactors aggregated to >5% and GroEL substrates.

(B) Size distribution of >5% aggregated proteins in ΔKJ and ΔKJT cells in comparison to soluble lysate proteins of WT cells.

(C) Changes in abundance of small and large ribosomal proteins in ΔKJ, ΔT/K$_{His}$, and ΔKJT relative to K$_{His}$ cells. SILAC ratios (mutant/K$_{His}$) are shown with standard deviations from three independent experiments for 21 small and 31 large ribosomal proteins.

majority of DnaK substrates (~80%) are cytosolic, DnaK also interacts with a small subset of proteins of the inner membrane, periplasm and outer membrane. Under conditions of exponential cell growth at 37°C, proteins bind to DnaK preferentially upon synthesis and may return to DnaK during their life time for conformational maintenance.

The dependence of proteins on DnaK for folding or maintenance is partially buffered by TF and other chaperones, but is reflected by the relative enrichment of substrates on DnaK. By measuring for each protein the fraction of total that is chaperone-bound, we defined a set of ~180 interactors that are highly enriched on DnaK and amount to about 40% of the total mass of DnaK substrates. DnaK-enriched proteins are generally of average or below average cellular abundance and of low solubility. Essential proteins are underrepresented among this group. DnaK dependence tends to correlate with the number of predicted DnaK binding sites in polypeptide sequences, and the propensity of proteins to populate structurally dynamic intermediates. Moreover, proteins that interact extensively with DnaK are often part of heterooligomeric complexes (Figure 7A).

Our findings suggest that proteins of lower abundance are frequently prone to misfolding or aggregation and thus have high chaperone requirements (with dependence on a specific chaperone system), whereas the folding properties of abundant (and often essential) proteins have been optimized in evolution, resulting in a reduced (or less specific) chaperone requirement. This is consistent with the existence of a negative correlation between the calculated aggregation-propensity of proteins and their cellular abundance (Tartaglia et al., 2010; Tartaglia et al., 2007). Indeed, the less abundant, DnaK-enriched substrates are aggregation-prone upon translation in vitro (Niwa et al., 2009) and are either degraded or aggregate in the absence of DnaK in vivo, reflecting their specific requirement for DnaK for folding and conformational maintenance. Interestingly, several DnaK-enriched substrates contain at least one domain with SCOP fold c.37 (Figure 7A). Proteins with this and other complex α/β topologies have to form many long-range interactions during folding and are thus likely to populate dynamic folding intermediates exposing hydrophobic residues (Gromiha and Selvaraj, 2004). Moreover, proteins with c.37 domains often assemble into heterooligomeric complexes (Figure S2G), a process that may be facilitated by DnaK's ability to bind partially structured protein regions in addition to extended peptide segments (Schlecht et al., 2011). On the other hand, many large proteins of higher cellular abundance but lower enrichment on DnaK are adapted to utilize either DnaK or TF for de novo folding. These

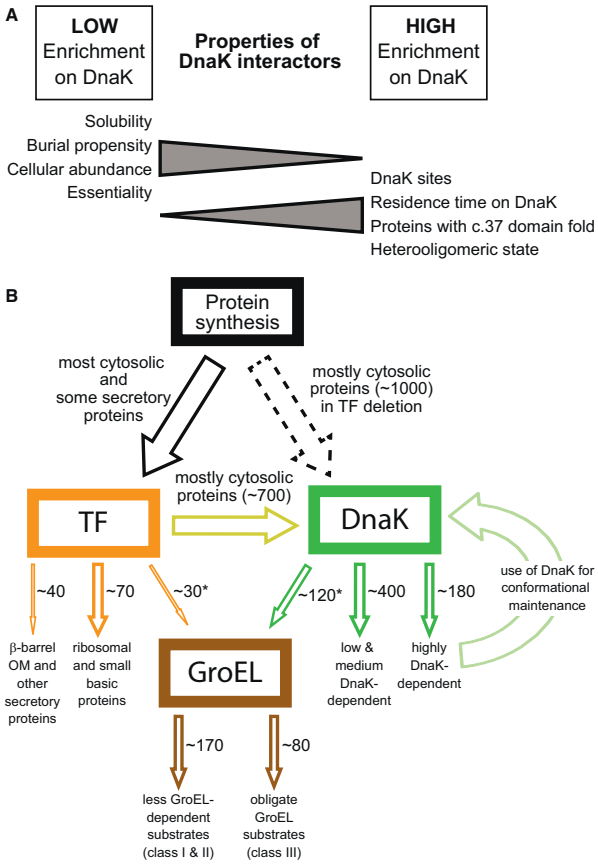


Figure 7. Central Role of DnaK in the Cytosolic Chaperone Network

(A) Properties of DnaK substrates correlating with their relative enrichment on DnaK (fraction of total cellular protein bound to DnaK).

(B) Functional redundancies and specificities in the chaperone network formed by TF, the DnaK system, and GroEL. Numbers denote numbers of proteins determined in this work and in the study by Kerner et al. (2005). *Note that ~100 known GroEL substrates were either not or not reproducibly identified as DnaK interactors. These proteins are of low cellular abundance and thus may have very low steady-state levels on DnaK.

proteins aggregate substantially only in the absence of both chaperones, defining sequence length, and hence multidomain topology, as a property strongly correlated with chaperone dependence.

Proteins with essential functions are underrepresented among the DnaK-enriched substrates. However, our identification of the essential tubulin homolog, FtsZ, and the cooperating MinCDE proteins as strong DnaK binders (Table S1) would explain why DnaK mutant cells have defects in cell division (Bukau and Walker, 1989). Furthermore, the sensitivity of Δ*dnaK* cells to antibiotics causing DNA damage (Nichols et al., 2011) is consistent with the finding that proteins of COG class L (DNA replication, recombination and repair), such as the nucleotide excision repair protein UvrA, are overrepresented among the DnaK-enriched substrates (Figure S2). UvrA is a large, heterooligomeric protein with two c.37 domains; which aggregates to 30% of total in the absence of DnaK/DnaJ already at 30°C. The sensitivity of Δ*dnaK* cells to antibiotics inhibiting protein synthesis (Nichols et al., 2011) would correlate with the extensive interaction of DnaK with ribosomal proteins and the degradation in ΔKJ cells of several DnaK interactors of COG class E (amino acid transport and metabolism). Finally, the sensitivity of Δ*dnaK* cells to acidic conditions (Nichols et al., 2011) is consistent with the 80%–97% degradation in ΔKJ cells of the periplasmic chaperones of acid-denatured proteins, hdeA, and hdeB.

Interplay between Chaperone Modules

Our analysis of the DnaK interactome in cells lacking TF or depleted of GroEL/ES underscores the significance of DnaK as a central hub in the chaperone network. The observed accumulation of a subset of proteins on DnaK in the absence of TF defines in a quantitative manner the functional redundancy between these two chaperone systems described earlier (Deuerling et al., 1999; Teter et al., 1999). Interestingly, these substrates comprise mostly ribosomal and other small (<20 kDa), positively charged proteins (Figure 7B), which may normally interact predominantly with TF, but shift to DnaK when TF is absent. Indeed, TF has a negative net charge (Ferbitz et al., 2004; Martinez-Hackert and Hendrickson, 2009), which may facilitate its interaction with positively charged nascent polypeptides, and a role of TF in the folding/assembly of ribosomal proteins has been suggested (Martinez-Hackert and Hendrickson, 2009).

In contrast, the DnaK system is unable to replace the role of TF in the biogenesis of a set of secretory proteins, prominently including β-barrel proteins of the outer membrane (Figure 7B). These proteins undergo partial degradation in cells lacking TF, suggesting a specific role of TF in translocation of outer membrane preproteins across the inner membrane. Such a function of TF would be consistent with the initial identification of TF as a chaperone of proOmpA translocation in vitro (Crooke et al., 1988) and with the finding that TF modulates the kinetics of

protein export (Lee and Bernstein, 2002; Ullers et al., 2007). It is also of interest in this context that TF structurally resembles the periplasmic chaperone for outer membrane proteins, SurA (Bitto and McKay, 2002; Ferbitz et al., 2004).

An effective functional cooperation apparently exists between TF and the DnaK system in the folding of a group of large multidomain proteins that aggregate substantially only in the absence of both chaperones. These proteins may normally interact sequentially with TF and DnaK during translation or with multiple TF molecules in the absence of DnaK, as shown with large model proteins in vitro (Agashe et al., 2004; Kaiser et al., 2006). Notably, most of these proteins would be unable to interact productively with GroEL, as the capacity of the GroEL/ES folding compartment is limited to proteins up to ~60 kDa (Kerner et al., 2005).

The DnaK interactome overlaps extensively with the set of previously identified GroEL substrates (Figure 7B), most of which are below 50 kDa in size (Kerner et al., 2005). GroEL substrates amount to nearly 30% of the total mass of DnaK interactors, indicating that protein transfer between DnaK and GroEL is a central function of the chaperone network. We estimate that only a minor fraction of GroEL substrates (~20%) are transferred directly from TF to GroEL, circumventing DnaK (Figure 7B). Notably, upon GroEL-depletion, obligate GroEL substrates accumulate further on DnaK, reflecting an important role of DnaK as a buffer in stabilizing these proteins until GroEL is available or in transferring them to the proteolytic system.

Defining Proteostasis Collapse

Upon growth at 30°C, the loss of individual chaperone modules—DnaK/DnaJ, TF or GroEL/ES—is remarkably well tolerated by *E. coli* cells in terms of preventing major protein aggregation. Instead, degradation is the strongly preferred fate of misfolded proteins under these conditions (Tables S2B, S6C, and S7C). However, proteostasis collapse characterized by extensive aggregation of relatively abundant proteins occurs when TF is deleted in addition to DnaK/DnaJ. Apparently, the chaperone capacity available for the folding and stabilization of large proteins in particular becomes severely limiting and, as a result, aggregation is favored relative to degradation. Furthermore, our data show that the loss of the upstream chaperones, TF and DnaK/DnaJ, disrupts the normal protein flux to GroEL, resulting in wide-spread aggregation of GroEL-substrates, despite a ~10-fold upregulation of GroEL under these conditions. The failure of newly synthesized GroEL substrates to reach the abundantly expressed chaperonin signifies the systematic collapse of the chaperone network.

EXPERIMENTAL PROCEDURES

Bacterial Strains

The *E. coli* strains used were based on MC4100 (WT) and are described in Extended Experimental Procedures.

Isolation of DnaK-Interactor Complexes

SILAC labeling of cells was performed at 37°C in M63 medium supplemented with light (L), medium (M), or heavy (H) arginine and lysine isotopes (see Extended Experimental Procedures). DnaK interactors were isolated from cells growing exponentially (OD_{600nm} ~1). In pulse-SILAC experiments, L-labeled cells were shifted to M-medium for 2.5 min; in pulse chase-SILAC experiments, L-labeled cells were shifted to H-medium for 2 min and then chased by addition of a 100-fold excess of L arginine and lysine for 2, 4, and 8 min. Spheroplasts were prepared as described (Ewalt et al., 1997) and lysed in hypo-osmotic buffer containing apyrase. Soluble cell lysate was prepared by centrifugation (20,000 × g, 30 min). Talon beads (Invitrogen) were used to isolate the DnaK-His6 and its interactors. The eluates containing bound proteins obtained from equal amounts of L-, M-, or H-labeled cells were mixed (see Extended Experimental Procedures). Samples were prepared for LC-MS/MS as described (Figure 1C) (Ong and Mann, 2006). The spectra were interpreted using MaxQuant version 1.0.13.13 (Cox and Mann, 2008) combined with Mascot version 2.2 (Matrix Science, www.matrixscience.com).

The MaxQuant tables along with a full list of identified proteins and quantitations are available at the Proteome Commons Tranche repository (https://proteomecommons.org/) by inserting the following tranche code: RKuFJcuu8iBnZHlBN4Nth0pz+HQgPNvV0zJvdnMN8wyAN8l7ifufEnmW6j2Cwn0msEakFC6eulEqlYv+7B+dFALonrgAAAAAAAAKDA== and passphrase: vqXGUF93rGLCraqm1yll.

The raw data is accessible using tranche codes and passphrase given in Extended Experimental Procedures.

Fractionation of Total Cell Lysate

E. coli MC4100 *dnaK-His6* (K_{His}) and chaperone mutant strains were grown to OD_{600nm} ~1 at 30°C or 37°C, as indicated, in the respective SILAC medium. Cells were collected, flash frozen, and lysed by sonication. Whole proteome analyses on total lysates, soluble, and detergent insoluble fractions (Deuerling et al., 2003) were performed by LC-MS/MS as described in Extended Experimental Procedures.

Bioinformatic Analysis

Bioinformatics and statistical analyses of the physico-chemical properties of protein sequences were performed as detailed in Extended Experimental Procedures. Note that both DnaJ and GrpE, the cofactors of DnaK, are excluded in all bioinformatic analyses.

SUPPLEMENTAL INFORMATION

Supplemental Information includes Extended Experimental Procedures, five figures, and eight tables and can be found with this article online at doi:10.1016/j.celrep.2011.12.007.

LICENSING INFORMATION

ACKNOWLEDGMENTS

We thank R.B. Körner and A. Ries for support with mass spectrometry and S. Pinkert for his help with data management. Financial support from EU Framework 7 Integrated Project PROSPECTS, the Deutsche Forschungsgemeinschaft (SFB 594), and the Körber Foundation is acknowledged. G.C. was supported by an EMBO long-term fellowship.

Received: October 31, 2011
Revised: December 4, 2011
Accepted: December 23, 2011
Published online: March 8, 2012

REFERENCES

Agashe, V.R., Guha, S., Chang, H.C., Genevaux, P., Hayer-Hartl, M., Stemp, M., Georgopoulos, C., Hartl, F.U., and Barral, J.M. (2004). Function of trigger factor and DnaK in multidomain protein folding: increase in yield at the expense of folding speed. Cell *117*, 199–209.

Balch, W.E., Morimoto, R.I., Dillin, A., and Kelly, J.W. (2008). Adapting proteostasis for disease intervention. Science *319*, 916–919.

Bitto, E., and McKay, D.B. (2002). Crystallographic structure of SurA, a molecular chaperone that facilitates folding of outer membrane porins. Structure *10*, 1489–1498.

Bukau, B., and Walker, G.C. (1989). Cellular defects caused by deletion of the *Escherichia coli* dnaK gene indicate roles for heat shock protein in normal metabolism. J. Bacteriol. *171*, 2337–2346.

Chapman, E., Farr, G.W., Usaite, R., Furtak, K., Fenton, W.A., Chaudhuri, T.K., Hondorp, E.R., Matthews, R.G., Wolf, S.G., Yates, J.R., et al. (2006). Global aggregation of newly translated proteins in an Escherichia coli strain deficient of the chaperonin GroEL. Proc. Natl. Acad. Sci. USA *103*, 15800–15805.

Cox, J., and Mann, M. (2008). MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. Nat. Biotechnol. *26*, 1367–1372.

Crooke, E., Guthrie, B., Lecker, S., Lill, R., and Wickner, W. (1988). ProOmpA is stabilized for membrane translocation by either purified E. coli trigger factor or canine signal recognition particle. Cell *54*, 1003–1011.

Deuerling, E., Patzelt, H., Vorderwülbecke, S., Rauch, T., Kramer, G., Schaffitzel, E., Mogk, A., Schulze-Specking, A., Langen, H., and Bukau, B. (2003). Trigger Factor and DnaK possess overlapping substrate pools and binding specificities. Mol. Microbiol. *47*, 1317–1328.

Deuerling, E., Schulze-Specking, A., Tomoyasu, T., Mogk, A., and Bukau, B. (1999). Trigger factor and DnaK cooperate in folding of newly synthesized proteins. Nature *400*, 693–696.

Ewalt, K.L., Hendrick, J.P., Houry, W.A., and Hartl, F.U. (1997). In vivo observation of polypeptide flux through the bacterial chaperonin system. Cell *90*, 491–500.

Ferbitz, L., Maier, T., Patzelt, H., Bukau, B., Deuerling, E., and Ban, N. (2004). Trigger factor in complex with the ribosome forms a molecular cradle for nascent proteins. Nature *431*, 590–596.

Fujiwara, K., Ishihama, Y., Nakahigashi, K., Soga, T., and Taguchi, H. (2010). A systematic survey of in vivo obligate chaperonin-dependent substrates. EMBO J. *29*, 1552–1564.

Gamer, J., Bujard, H., and Bukau, B. (1992). Physical interaction between heat shock proteins DnaK, DnaJ, and GrpE and the bacterial heat shock transcription factor sigma 32. Cell *69*, 833–842.

Genevaux, P., Keppel, F., Schwager, F., Langendijk-Genevaux, P.S., Hartl, F.U., and Georgopoulos, C. (2004). In vivo analysis of the overlapping functions of DnaK and trigger factor. EMBO Rep. *5*, 195–200.

Gerdes, S.Y., Scholle, M.D., Campbell, J.W., Balázsi, G., Ravasz, E., Daugherty, M.D., Somera, A.L., Kyrpides, N.C., Anderson, I., Gelfand, M.S., et al. (2003). Experimental determination and system level analysis of essential genes in Escherichia coli MG1655. J. Bacteriol. *185*, 5673–5684.

Gidalevitz, T., Prahlad, V., and Morimoto, R.I. (2011). The stress of protein misfolding: from single cells to multicellular organisms. Cold Spring Harb. Perspect. Biol. *3*, 3.

Gromiha, M.M., and Selvaraj, S. (2004). Inter-residue interactions in protein folding and stability. Prog. Biophys. Mol. Biol. *86*, 235–277.

Hartl, F.U., Bracher, A., and Hayer-Hartl, M. (2011). Molecular chaperones in protein folding and proteostasis. Nature *475*, 324–332.

Hartl, F.U., and Hayer-Hartl, M. (2009). Converging concepts of protein folding in vitro and in vivo. Nat. Struct. Mol. Biol. *16*, 574–581.

Ishihama, Y., Oda, Y., Tabata, T., Sato, T., Nagasu, T., Rappsilber, J., and Mann, M. (2005). Exponentially modified protein abundance index (emPAI) for estimation of absolute protein amount in proteomics by the number of sequenced peptides per protein. Mol. Cell. Proteomics *4*, 1265–1272.

Janin, J. (1979). Surface and inside volumes in globular proteins. Nature *277*, 491–492.

Kaiser, C.M., Chang, H.C., Agashe, V.R., Lakshmipathy, S.K., Etchells, S.A., Hayer-Hartl, M., Hartl, F.U., and Barral, J.M. (2006). Real-time observation of trigger factor function on translating ribosomes. Nature *444*, 455–460.

Kerner, M.J., Naylor, D.J., Ishihama, Y., Maier, T., Chang, H.C., Stines, A.P., Georgopoulos, C., Frishman, D., Hayer-Hartl, M., Mann, M., and Hartl, F.U. (2005). Proteome-wide analysis of chaperonin-dependent protein folding in Escherichia coli. Cell *122*, 209–220.

Lee, H.C., and Bernstein, H.D. (2002). Trigger factor retards protein export in Escherichia coli. J. Biol. Chem. *277*, 43527–43535.

Maki, J.A., Schnobrich, D.J., and Culver, G.M. (2002). The DnaK chaperone system facilitates 30S ribosomal subunit assembly. Mol. Cell *10*, 129–138.

Martinez-Hackert, E., and Hendrickson, W.A. (2009). Promiscuous substrate recognition in folding and assembly activities of the trigger factor chaperone. Cell *138*, 923–934.

Mayer, M.P. (2010). Gymnastics of molecular chaperones. Mol. Cell *39*, 321–331.

McLennan, N., and Masters, M. (1998). GroE is vital for cell-wall synthesis. Nature *392*, 139.

Nichols, R.J., Sen, S., Choo, Y.J., Beltrao, P., Zietek, M., Chaba, R., Lee, S., Kazmierczak, K.M., Lee, K.J., Wong, A., et al. (2011). Phenotypic landscape of a bacterial cell. Cell *144*, 143–156.

Niwa, T., Ying, B.W., Saito, K., Jin, W., Takada, S., Ueda, T., and Taguchi, H. (2009). Bimodal protein solubility distribution revealed by an aggregation analysis of the entire ensemble of Escherichia coli proteins. Proc. Natl. Acad. Sci. USA *106*, 4201–4206.

Ong, S.E., Blagoev, B., Kratchmarova, I., Kristensen, D.B., Steen, H., Pandey, A., and Mann, M. (2002). Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics. Mol. Cell. Proteomics *1*, 376–386.

Ong, S.E., and Mann, M. (2006). A practical recipe for stable isotope labeling by amino acids in cell culture (SILAC). Nat. Protoc. *1*, 2650–2660.

Raineri, E., Ribeca, P., Serrano, L., and Maier, T. (2010). A more precise characterization of chaperonin substrates. Bioinformatics *26*, 1685–1689.

René, O., and Alix, J.-H. (2011). Late steps of ribosome assembly in E. coli are sensitive to a severe heat stress but are assisted by the HSP70 chaperone machine. Nucleic Acids Res. *39*, 1855–1867.

Rüdiger, S., Buchberger, A., and Bukau, B. (1997). Interaction of Hsp70 chaperones with substrates. Nat. Struct. Biol. *4*, 342–349.

Schlecht, R., Erbse, A.H., Bukau, B., and Mayer, M.P. (2011). Mechanics of Hsp70 chaperones enables differential interaction with client proteins. Nat. Struct. Mol. Biol. *18*, 345–351.

Smock, R.G., Rivoire, O., Russ, W.P., Swain, J.F., Leibler, S., Ranganathan, R., and Gierasch, L.M. (2010). An interdomain sector mediating allostery in Hsp70 molecular chaperones. Mol. Syst. Biol. *6*, 414.

Straus, D., Walter, W., and Gross, C.A. (1990). DnaK, DnaJ, and GrpE heat shock proteins negatively regulate heat shock gene expression by controlling the synthesis and stability of sigma 32. Genes Dev. *4*(12A), 2202–2209.

Tartaglia, G.G., Dobson, C.M., Hartl, F.U., and Vendruscolo, M. (2010). Physicochemical determinants of chaperone requirements. J. Mol. Biol. *400*, 579–588.

Tartaglia, G.G., Pawar, A.P., Campioni, S., Dobson, C.M., Chiti, F., and Vendruscolo, M. (2008). Prediction of aggregation-prone regions in structured proteins. J. Mol. Biol. *380*, 425–436.

Tartaglia, G.G., Pechmann, S., Dobson, C.M., and Vendruscolo, M. (2007). Life on the edge: a link between gene expression levels and aggregation rates of human proteins. Trends Biochem. Sci. *32*, 204–206.

Teter, S.A., Houry, W.A., Ang, D., Tradler, T., Rockabrand, D., Fischer, G., Blum, P., Georgopoulos, C., and Hartl, F.U. (1999). Polypeptide flux through bacterial Hsp70: DnaK cooperates with trigger factor in chaperoning nascent chains. Cell *97*, 755–765.

Ullers, R.S., Ang, D., Schwager, F., Georgopoulos, C., and Genevaux, P. (2007). Trigger Factor can antagonize both SecB and DnaK/DnaJ chaperone functions in Escherichia coli. Proc. Natl. Acad. Sci. USA *104*, 3101–3106.

Van Durme, J., Maurer-Stroh, S., Gallardo, R., Wilkinson, H., Rousseau, F., and Schymkowitz, J. (2009). Accurate prediction of DnaK-peptide binding via homology modelling and experimental data. PLoS Comput. Biol. *5*, e1000475.

Yu, N.Y., Laird, M.R., Spencer, C., and Brinkman, F.S. (2011). PSORTdb—an expanded, auto-updated, user-friendly protein subcellular localization database for Bacteria and Archaea. Nucleic Acids Res. *39*(Database issue), D241–D244.

Zhao, K., Liu, M., and Burgess, R.R. (2005). The global transcriptional response of Escherichia coli to induced σ^{32} protein involves σ^{32} regulon activation followed by inactivation and degradation of σ^{32} in vivo. J. Biol. Chem. *280*, 17758–17768.

Zhu, X.T., Zhao, X., Burkholder, W.F., Gragerov, A., Ogata, C.M., Gottesman, M.E., and Hendrickson, W.A. (1996). Structural analysis of substrate binding by the molecular chaperone DnaK. Science *272*, 1606–1614.

Zhuravleva, A., and Gierasch, L.M. (2011). Allosteric signal transmission in the nucleotide-binding domain of 70-kDa heat shock protein (Hsp70) molecular chaperones. Proc. Natl. Acad. Sci. USA *108*, 6987–6992.

Chapter VI

The *cc*SOL algorithm

Escherichia coli is one of the most widely used hosts for the production of recombinant proteins. However, very often the target protein accumulates into insoluble aggregates in a misfolded and biologically inactive form (Ventura, 2005). For this reason, I took advantage of experimental data on ~70% of the *E. coli* proteins (Niwa et al., 2009) to design a sequence-based algorithm to predict protein solubility. In this method, a number of physico-chemical properties are used to describe the polypeptide: coil/disorder, hydrophobicity, hydrophilicity, β-turn, α-helix propensities. These features are then provided as input to a support vector machine classifier, which allows the calculation of protein solubility with great accuracy.

Agostini, F., Vendruscolo, M., and Tartaglia, G. G. (2012). Sequence-based prediction of protein solubility. *Journal of Molecular Biology*, 421(2-3):237–241. PMID: 22172487 DOI: 10.1016/j.jmb.2011.12.005

Chapter VII

*cc*SOL *omics*

The aforementioned *cc*SOL method was introduced in 2012 to predict the *in vitro* solubility of *E. coli* proteins using physico-chemical information contained in primary structure. However, in its original implementation *cc*SOL allowed predictions of one protein at a time, making it difficult to perform large-scale analyses. Therefore, we decided to develop *cc*SOL *omics*, the first web-server for proteome-wide predictions of solubility. The training was performed using a neural network on data retrieved from the Target Track database and filtered using criteria similar to Smialowski et al. (2012). In this new release we introduce three significant improvements: i) validation of *omics* performances using large sets comprising $>$65000 non-redundant proteins (similarity $\leq$30%); ii) solubility profiles to reveal the most/least soluble regions within each protein sequence and iii) detection of susceptible regions that change solubility upon single point mutation. The algorithm shows an overall accuracy of 78% in predicting protein solubility.

This article has been submitted for publication as "Application Note" to the "Bioinformatics" journal.

Agostini F, Cirillo D, Livi CM, Delli Ponti R, Tartaglia GG. ccSOL omics: a webserver for solubility prediction of endogenous and heterologous expression in Escherichia coli. Bioinformatics. 2014 Oct 15; 30(20): 2975-7. DOI: 10.1093/bioinformatics/btu420

Application Note

*cc*SOL *omics*: a web server for large-scale prediction of protein solubility

Federico Agostini[1,2], Davide Cirillo[1,2] , Carmen Maria Livi[1,2] and Gian Gaetano Tartaglia[1,2,*]

[1]Gene Function and Evolution, Centre for Genomic Regulation (CRG), Dr. Aiguader 88, 08003 Barcelona, Spain.
[2] Universitat Pompeu Fabra (UPF), 08003 Barcelona, Spain.
* Corresponding author: Gian Gaetano Tartaglia. Telephone +34 933160116. Email: gian.tartaglia@crg.es.

ABSTRACT

Summary: Here we introduce *cc*SOL *omics*, a webserver for large-scale calculations of protein solubility. Our method allows (i) proteome-wide predictions; (ii) identification of soluble fragments within each sequences; iii) exhaustive single-point mutation analysis.

Results: Using coil/disorder, hydrophobicity, hydrophilicity, β-sheet and α-helix propensities, we perform large-scale predictions of protein solubility. Our approach shows an accuracy of 79% on the training set (36990 Target Track entries). Furthermore, cross-validation on three independent sets indicates that *cc*SOL *omics* discriminates soluble and insoluble proteins with an accuracy of 74% on 31760 proteins sharing less than 30% sequence similarity.

Availability: *cc*SOL *omics* can be freely accessed on the web at http://s.tartaglialab.com/page/ccsol_group.

Supplementary information: Documentation and tutorial are available at http://s.tartaglialab.com/static_files/shared/tutorial_ccsol_omics.html.

1 INTRODUCTION

The early methods for prediction of protein aggregation and solubility were trained on 100 proteins or less (Tartaglia *et al.*, 2005). Although able to identify patterns in proteome-wide analyses (Fernandez-Escamilla *et al.*, 2004; Conchillo-Solé *et al.*, 2007), these algorithms were not built to make large-scale predictions of protein solubility. As a matter of fact, to perform proteome-wide predictions of protein solubility, one should train models using a high number of solubility data. In 2012, we introduced the *cc*SOL method (Agostini *et al.*, 2012) to predict protein solubility using 5 physico-chemical properties: coil/disorder, hydrophobicity, hydrophilicity, β-sheet and α-helix. To identify these features, we divided the original database, consisting of experimental solubility data (Niwa *et al.*, 2009), into two subsets containing the most soluble (1081 entries) and least soluble (1078 entries) proteins and calculated the discriminative power of a number of physicochemical properties collected through a literature search. Other methods have been developed to predict protein solubility using amino acid sequences alone. For instance, PROSO II (Smialowski *et al.*, 2012) exploits occurrence of monopeptides and dipeptides. Two main differences between *cc*SOL and PROSO II are i) the input variables and ii) the training sets employed for validation. PROSO II was trained on the pepcDB database [now Target Track (Berman *et al.*, 2009)] that stores target and protocol information provided by Protein Structure Initiative centers, while *cc*SOL employs soluble fractions of *E. coli* proteins (Niwa *et al.*, 2009). Both *cc*SOL and PROSO II perform accurate predictions when used to predict endogenous or heterologous soluble expressions, respectively [*cc*SOL: 76% accuracy; PROSO II: 75% accuracy (Smialowski *et al.*, 2012)]. However, we note that the experimental status of a number of entries has been updated in Target Track (http://sbkb.org/tt/), which can be used to derive solubility of proteins (see Supplementary information). Here, we introduce a new implementation of the *cc*SOL method to perform large-scale solubility predictions of both endogenous and heterologous expression in *E. coli*. Specifically, our algorithm exploits Target Track information and allows identification of soluble and insoluble regions within protein sequences.

2 WORKFLOW AND IMPLEMENTATION

The *cc*SOL *omics* server allows the investigation of large protein datasets (see online documentation). Once the user provides sequences in FASTA format, the algorithm calculates:

- *Solubility profiles.* To identify soluble fragments within each polypeptide chain, protein sequences are divided into elements and calculate individual solubility propensities. Starting from the N-terminus of a protein, we use a sliding window of 21 amino acids that is moved one residue at a time until the C-terminus is reached. The solubility propensity profile of each window is calculated with *cc*SOL as defined in our original publication (Agostini *et al.*, 2012).
- *Sequence susceptibility.* For each sequence analyzed, the algorithm computes the effect of single amino acid mutations at different positions. This approach is particularly useful to identify regions susceptible to change solubility upon mutation. All variants are reported along with their scores, which provides a basis to engineer protein sequences and test hypotheses such as the occurrence of specific mutations in pathology.
- *Solubility score.* The solubility profile represents a unique *signature* containing information on all fragments arranged in sequential order. The profile can be used to estimate the solubility upon expression in the *E. coli* system. As sequences have different lengths, we exploit a method based on Fourier's transform (Bellucci *et al.*, 2011; Tartaglia *et al.*, 2007) that allows comparison of polypeptide chains with different sizes. Using 100 Fourier's coefficients, we trained an algorithm that has the same architecture developed for the analysis of expression levels in *E. coli* [neural network with 10 inner neurons and one output (Tartaglia *et al.*, 2009)].

These three types of analyses are performed for each protein in the submitted dataset if its size is below 500 entries (only the solubility score will be computed otherwise).

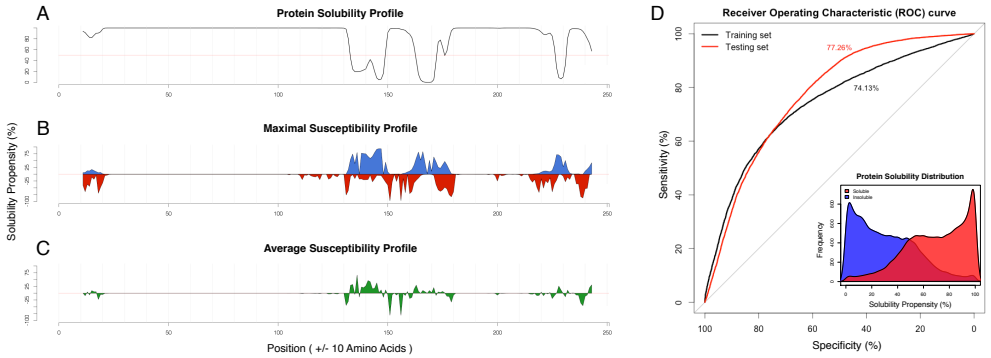


Figure 1. Human Prion Solubility and *cc*SOL Performances. A) Starting from the N-terminus, *cc*SOL computes the solubility profile using a sliding window moved towards the C-terminus. *cc*SOL identifies the fragment 130-170 as the most insoluble within the C-terminus of human PrP. B-C) Maximal and average susceptibility upon single point mutation. D) We trained on the Target Track set (AUROC = 74.13%) and tested the approach on *E. coli* proteins [AUROC = 92.2%; (Niwa *et al*., 2009)], SOLpro [AUROC = 79.0%; (Magnan *et al*., 2009)] and PROSO II [AUROC = 75.6%; (Smialowski *et al*., 2012)]. Inset: overall score distribution for soluble (red) and insoluble (blue) proteins.

3 PERFORMANCES

As an illustrative example, we report human prion protein. Prion diseases are a group of neurodegenerative disorders associated with a conformational transformation of the prion protein (PrP^{C}) into a self-replicating conformer PrP^{Sc}. Our algorithm correctly identifies the fragment 130-170 as the most insoluble (Figure 1A-C) together with region 231-253 (not present in the mature form). This finding is very well in agreement with what has been discussed in previous reports (Tartaglia *et al.*, 2005, 2008). The analysis of susceptible fragments identifies a number of experimentally validated mutations (e.g. G131V, S132I, R148H, V176I, D178N) associated with lower solubility and located in the region promoting PrP^{Sc} conversion (Corsaro *et al.*, 2012). We validated *cc*SOL *omics* with a 10-fold cross validation on Target Track [total of 36990 entries with 30% redundancy (Fu *et al.*, 2012)] and observed 79% accuracy in discriminating between soluble and insoluble proteins. Furthermore, we tested the algorithm on three independent datasets containing protein expression data [total of 31760 entries taken from: *E. coli* (Niwa *et al.*, 2009), SOLpro (Magnan *et al.*, 2009) and PROSO II (Smialowski *et al.*, 2012) and found 74% accuracy (Figure 1D), which indicates that our tool achieves good performances (see Supplementary information).

4 CONCLUSIONS

The *cc*SOL *omics* algorithm shows excellent performances in predicting solubility of endogenous and heterologous genes in *E. coli*. We hope that the webserver will be useful for biotechnological purpose, as it could be for instance employed to design fusion tags for soluble expression (Wilkinson and Harrison, 1991). Moreover, we plan to develop a new implementation of the algorithm to assess protein solubility in other expression systems, such as *S. cerevisiae*, which will lead to a better understanding of i) sequence evolution, ii) post-translational modifications and iii) environmental conditions. In the future, it will be also important to develop new methods to understand the role of chaperones in preventing protein aggregation (Tartaglia *et al.*, 2010) and evaluate if other molecules, such as RNA, can contribute to protein solubility (Choi *et al.*, 2009).

ACKNOWLEDGMENTS

The authors would like to thank Prof R. Guigó and Dr G. Bussotti for stimulating discussions.

Funding: The research leading to these results has received funding from the European Union Seventh Framework Programme (FP7/2007-2013), through the European Research Council, under grant agreement RIBOMYLOME_309545, and from the Spanish Ministry of Economy and Competitiveness (SAF2011-26211). We also acknowledge support from the Spanish Ministry of Economy and Competitiveness, 'Centro de Excelencia Severo Ochoa 2013-2017' (SEV-2012-0208)..

REFERENCES

Agostini,F. *et al.* (2012) Sequence-based prediction of protein solubility. *J. Mol. Biol.*, **421**, 237–241.
Bellucci,M. *et al.* (2011) Predicting protein associations with long noncoding RNAs. *Nat. Methods*, **8**, 444–445.
Berman,H.M. *et al.* (2009) The protein structure initiative structural genomics knowledgebase. *Nucleic Acids Res.*, **37**, D365–368.
Choi,S.I. *et al.* (2009) RNA-mediated chaperone type for de novo protein folding. *RNA Biol*, **6**, 21–24.
Conchillo-Solé,O. *et al.* (2007) AGGRESCAN: a server for the prediction and evaluation of 'hot spots' of aggregation in polypeptides. *BMC Bioinformatics*, **8**, 65.
Corsaro,A. *et al.* (2012) Role of prion protein aggregation in neurotoxicity. *Int J Mol Sci*, **13**, 8648–8669.
Fernandez-Escamilla,A.-M. *et al.* (2004) Prediction of sequence-dependent and mutational effects on the aggregation of peptides and proteins. *Nat Biotechnol*, **22**, 1302–6.
Fu,L. *et al.* (2012) CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics*, **28**, 3150–3152.
Magnan,C.N. *et al.* (2009) SOLpro: accurate sequence-based prediction of protein solubility. *Bioinformatics*, **25**, 2200–2207.
Niwa,T. *et al.* (2009) Bimodal protein solubility distribution revealed by an aggregation analysis of the entire ensemble of Escherichia coli proteins. *Proc. Natl. Acad. Sci. U.S.A*, **106**, 4201–4206.
Smialowski,P. *et al.* (2012) PROSO II – a new method for protein solubility prediction. *FEBS Journal*, **279**, 2192–2200.
Tartaglia,G.G. *et al.* (2009) A relationship between mRNA expression levels and protein solubility in E. coli. *J. Mol. Biol*, **388**, 381–389.
Tartaglia,G.G. *et al.* (2010) Physicochemical determinants of chaperone requirements. *J. Mol. Biol*, **400**, 579–588.
Tartaglia,G.G. *et al.* (2005) Prediction of aggregation rate and aggregation-prone segments in polypeptide sequences. *Protein Sci*, **14**, 2723–34.
Tartaglia,G.G. *et al.* (2008) Prediction of aggregation-prone regions in structured proteins. *J Mol Biol*, **380**, 425–36.

Discussion

In this thesis, I presented two distinct but intimately connected topics: the ability of proteins to interact with RNA molecules and the relation between protein solubility and aggregation propensities. Specifically, I focused on the development of methods to characterize ribonucleoprotein associations (Chapters I, II, III and IV) and physico-chemical features defining protein solubility (Chapters V, VI and VII). I applied these computational approaches to investigate proteins binding to the long non-coding Xist (Chapter II) as well as to unravel determinants for DnaK chaperone recognition (Chapter V). The link between RNA-binding ability and protein aggregation is particularly relevant if one considers that a number of amyloidogenic proteins have an RNA-binding ability. For example, as discussed in two recent publications that I co-authored (Cirillo et al., 2012; Zanzoni et al., 2013), proteins involved in neurodegenerative diseases such as TDP-43 and FUS (amyotrophic lateral sclerosis) and FMRP (fragile X mental retardation and tremor/ataxia syndrome), regulate a large part of the transcriptome and possess an intrinsic propensity to aggregate (Liu-Yesucevitz et al., 2011). Intriguingly, TDP-43 and FMRP are able to bind to their own mRNAs (Ayala et al., 2011; Schaeffer et al., 2001; Zanzoni et al., 2013), which can be regarded as a way to control their expression levels and, possibly, avoid high concentration and aggregation.

The role of protein-RNA interactions has been intensively studied for its centrality in transcriptional and post-transcriptional events (Bernhardt, 2012; Keren et al., 2010). Alternative splicing represents one example of

the importance of protein-RNA interactions (Keren et al., 2010). In fact, splicing gradually creates protein functionalities without the need of additional genes and without affecting existing products (Gal-Mark et al., 2009; Keren et al., 2010; Zarnack et al., 2013). Protein interaction networks have been studied for decades yielding remarkable results concerning the understanding of basic biological mechanisms (Ryan et al., 2013; Yu et al., 2013). Nevertheless, the recent discovery of a plethora of RNAs, including lncRNAs and other previously uncharacterized transcripts, demanded a reexamination of biological networks to include these new effectors in the established protein-centric landscape (Prasanth and Spector, 2007). This led to the formulation of a more heterogeneous perspective, where RNA ceased to be considered a mere carrier of information but rather an active player in nearly every cellular process.

RNA as the key player

It has now become clear that every RNA inside the cell makes contact with a wide variety of molecules throughout its lifetime and that the ability to participate in the interactions is independent of the coding or non-coding potential of the transcript. Compelling evidence indicates that the complexity of higher organisms correlates with relative amount of non-coding RNA rather than the number of protein-coding genes (Barrett et al., 2012). One possible explanation of this phenomenon derives from the observation that the number and complexity of regulatory pathways increases in higher organisms (Levine and Tjian, 2003), and that within these pathways, there is less conservation within non-coding sequence, producing phenotypic variation between both individuals and species (Mattick, 2001). The largest and best-characterized group of proteins engaging coding and non-coding transcripts in the formation of RNPs is represented by the RBPs. These are generally defined as proteins with the ability to

bind to RNA, which is usually associated with one or multiple RBDs. The human genome harbors 1783 genes encoding known RBPs or proteins annotated to contain at least one RBD (Ascano et al., 2013). Nonetheless, the number of proteins with identified RNA-binding ability, either possessing canonical or non-canonical RBDs (Lunde et al., 2007), is increasing (Baltz et al., 2012; Castello et al., 2012; Kwon et al., 2013). The fact that some proteins are able to bind to RNA with domains or regions that are not specifically evolved to this precise purpose (Castello et al., 2012; Kwon et al., 2013) is quite intriguing. This suggests a scenario where unexpected players can exert crucial functions in domains that were previously thought of as exclusively regulated by selected RBD-containing proteins. With this in mind, computational models represent an important source of information that can be exploited to identify hidden trends and understand the basics of molecular recognition. As a matter of fact, bioinformatic tools can perform exhaustive analyses and extract distinctive features, hence facilitating the design of new experiments. For example, it has been shown in several studies that the composition of primary protein structure, and the physico-chemical properties associated with it, can be used to describe the amino acid regions that are more likely to be involved in binding to RNA molecules (Terribilini et al., 2007; Fernandez et al., 2011). This practice, however, cannot be applied to transcripts as very little is known of the features and specificities of RNA-binding sites. Furthermore, due to the limitations of current experimental approaches, it remains difficult to simultaneously investigate the plethora of RBPs bound to a single transcript and the RNA regions that are likely to be involved in the binding. This has resulted in experimentalists having to rely on protein analysis to investigate specific signatures. Nonetheless, experimental studies and computational analyses, such as those presented in this thesis, are providing compelling insights into the rules that govern RNP formation.

RNA recognition elements

A typical procedure for experimental determination of the RNP composition is the identification of exact protein binding sites and the derivation of the underlying RNA recognition elements (RREs) (Gerstberger et al., 2013). As a matter of fact, computational definition of patterns is extremely useful to detect false positives and negatives, and to highlight presence of degenerate motifs and putative co-binding factors (Li et al., 2014). On the one hand, RBPs can carry one or multiple RBDs that recognize the RNA with different affinities and *in vivo* they can compete for the same or proximal binding sites (Ascano et al., 2013). On the other hand the binding sites within transcripts generally consists of sequence patterns, from three to eight nucleotides in length, and each of these motifs can form or possess different secondary structure conformations. Therefore, although recognition of RNA binding site and assembly of multimeric complexes is mainly performed by proteins (Wan et al., 2011), it is also important to consider the specific features of the target molecules. Hence, the identification of sequence motifs, structural patterns and interplay between them would greatly advance our understanding of the mechanisms by which transcripts participate in post-transcriptional regulation (Li et al., 2014).

RNA motifs

Motif discovery methods can be roughly classified as profile-based, such as MEME (Bailey and Elkan, 1994), or pattern-based like CONSENSUS (Hertz and Stormo, 1999) (see Tompa et al. (2005) for a review and performance study of popular motif discovery tools). Most of these tools, however, compare the signal against a non-informative null distribution and do not easily scale to very large datasets. To address these problems, a number of discriminative motif discovery methods that perform large-scale analyses have been developed (Bailey, 2011; Thomas-Chollier et al., 2011). The

strength of these approaches mainly consists in the calculation of motif enrichment in a foreground dataset against an explicitly stated background dataset, which is carefully selected to eliminate systematic biases present in the foreground (Yao et al., 2013). The *SeAMotE* algorithm has been introduced as an approach to perform discriminative large-scale motif discovery analyses in datasets of nucleic acid sequences. The performances of the method were evaluated on a series of CLIP data sets and the identified patterns compared to the ones obtained with DREME (Bailey, 2011)and other methods. Both algorithms are able to impartially discriminate bound and unbound sequences by identifying representative k-mers motifs. Performances seem to correlate in most of the cases, regardless of the discrimination level achieved in the individual analysis. As a matter of fact, both tools evaluate nucleotide composition of the primary sequence, but do not take into account other sources of information. Therefore, cases showing poor performances can be explained by involvement of other factors, such as secondary structure or other positional features, in the determination of binding propensities. As an example, the RNA motifs method was recently introduced to identify enriched groups of non-degenerate or degenerate nucleotide tetramers, also including the positional information along the transcripts (Cereda et al., 2014). The approach was applied to evaluate clusters of tetramers in three regions around the splice sites of alternative exons. By integrating the positional information with motif search, the authors were able to successfully infer the binding sites for a number of validated RBPs (Cereda et al., 2014). This result indicates that, although the analyses performed using primary sequences provide quick and informative results, predictions of RBP motifs can be further improved (Hiller et al., 2006). As a matter of fact, the implementation of additional aspects, such as sequence conservation, RNA structure and analysis of non-clustered contiguous motifs, into motif discovery methods would definitely facilitate the comprehension of poorly understood and non-canonical RNA-protein recognition mechanisms.

Structural patterns

Although primary sequence can be used for binding site predictions, recent studies indicate that knowledge of target-site structural conformation increases the *in vivo* inference of the binding site (Li et al., 2010; Ray et al., 2013). As a matter of fact, comparison of transcripts, such as lncRNAs, have been difficult owing to the poor degree of sequence conservation of most of these genes (Derrien et al., 2012). As the strongest signal contained in these class of molecules is usually represented by evolutionarily conserved secondary structures (Johnsson et al., 2014), lncRNAs are challenging our ability of comparison, classification and search using conventional alignment tools (Bussotti et al., 2013). Nevertheless, as shown by Siebert and Backofen (2005) and Wilm et al. (2008), the combination of multiple sequence alignment with the information on RNA secondary structures is currently representing one the most promising approaches. A similar strategy has been adopted by RNA context (Kazan et al., 2010), which takes advantage of established RNA secondary structure predictive methods (Bernhart et al., 2006; Bompfünewerer et al., 2008) to calculate accessibility of the putative binding motifs. A number of algorithms exist that are able to infer RNA secondary structures using minimum free energy (MFE) or stochastic free context grammar (SCFG) approaches (Hofacker, 2014; Giegerich, 2014; Lai et al., 2013; Seetin and Mathews, 2012). The use of such techniques, however, is often limited. Indeed, these methods do not take into account the contribution of the environment, and predictions may not accurately represent the typical base pairing that occurs in the structure (Bussotti et al., 2013). To address these concerns, some methods consider the ensemble of all possible structures (Bernhart et al., 2006; Bompfünewerer et al., 2008; Zuker and Stiegler, 1981). *cat*RAPID relies on the ViennaRNA package (Hofacker, 2003), which has an accuracy of ~76% (Lorenz et al., 2011), to generate predictions of secondary structure ensembles. These structures are then dissected to extract information on the pairing profile of each nucleotide. By means of this procedure, the

probability of *cat*RAPID predicting a protein-RNA interaction has a 72% correlation with secondary structure information. However, a higher correlation factor is consistently expected with the enhancement of secondary structure prediction accuracies. Furthermore, as the predictive power of global RNA structure becomes less accurate as the length of the RNA increases (Doshi et al., 2004), we developed the *cat*RAPID *fragments* module that exploits the RNALfold algorithm (Lorenz et al., 2011) to determine interactions for the most stable local structure.

A number of lncRNAs (Chapter II), inherently possess evolutionarily conserved secondary structures (Smith et al., 2013; Johnsson et al., 2014). Disruption of such special organization can lead to a complete loss of RNA-protein interaction and, ultimately, of RNP function (Mercer and Mattick, 2013). For example, the RepA of the *Xist* transcript (Brown et al., 1992) contains eight repeated sequence elements arranged in a specific three-stem-loop architecture, whose progressive impairment has been demonstrated to prevent the recruitment of some, to all of the PRC2 complex subunits (Maenner et al., 2010). We demonstrated that the *cat*RAPID method is able to identify the interaction of the RepA region and the PRC2 complex. Furthermore, the gradual deletion of the repeated elements, involved in the stem-loop formation, was reflected by a decrease of the predicted interacting score. Although no evidence has been provided for the existence of a structural configuration in the RepC of *Xist*, the presence of repeated elements and the site-specific binding of the YY1 protein (Jeon and Lee, 2011) suggest the possibility that a similar recognition mechanism could be involved.

The implementation of *in vitro*/*in vivo* derived information in *cat*RAPID calculations represent an enticing direction for the prediction of protein-RNA associations. As a matter of fact, the use of high quality experimentally derived secondary structure information (Kertesz et al., 2010) will definitely improve the performance of computational models in predicting RNA regions involved in the binding. Classical experimental methods for

RNA structure determination include X-ray crystallography, NMR, cryo-electron microscopy and chemical and enzymatic probing. However, so far these methods have only been applied in the analysis of single RNA and the length of probed transcripts often limits their use. To address these issues, new and promising large-scale techniques are rapidly emerging in the field. Parallel Analysis of RNA Structure (PARS) is based on the generation of precise RNA fragments, by digestion using a single strand specific enzyme (S1) and a double-strand specific enzyme (V1), followed by deep sequencing (Kertesz et al., 2010).

Similarly, high-throughput sequencing of fragments generated by single-strand specific nuclease (P1) has been applied to study RNA structures in different cells. In this case, the selective 2'-hydroxyl acylation analyzed by primer extension (SHAPE) chemistry, combined with multiplexed bar coding and next generation sequencing, was able to measure the structures of a complex pool of RNAs (Lucks et al., 2011). Finally, the Parallel Analysis of RNA structures with Temperature Elevation (PARTE) was conceived in an attempt to combine the two approached described above (Wan et al., 2012). In this approach, RNA footprinting using RNase V1 is coupled to high-throughput sequencing to probe for double-stranded regions across five temperatures, from 23 to 75ºC. Methods based on technologies such as PARS, PARTE and SHAPE will be very useful for the determination of RNA structure *in vivo* and will provide large-scale data to be exploited by new and powerful predictive algorithms. Indeed, despite significant global correspondences, there are substantial differences between experimental results and computational predictions. A fact that might be due, in part to noise in the employed approaches but also due to known inaccuracies of folding algorithms (Kertesz et al., 2010). Therefore, it is advisable to couple data obtained by means of these techniques with the use computational algorithms to have a better estimate of RNA secondary structures and folding energies.

Homeostasis and RNA-protein interactions

Protein homeostasis is crucial for the maintenance of proteins in their soluble state (Chapter V). Even relatively small impairments in the quality control mechanisms that regulate the protein concentration can eventually lead to aberrant conditions (Vendruscolo et al., 2011). It has been shown that proteins are present in the cytosol at concentrations at which they are only marginally soluble (Tartaglia et al., 2007) and that different types of stress conditions can lead to widespread aggregation in living organisms (Geiler-Samerotte et al., 2011; Narayanaswamy et al., 2009; Olzscha et al., 2011). At present, determining how even small changes in amino acid composition can alter local physico-chemical context and inadvertently lead to aggregation presents an exciting challenge (Weber and Brangwynne, 2012). As a matter of fact, not all regions of a polypeptide chain are equally important for determining the aggregation propensities. Very short specific amino acid stretches can act as facilitators or inhibitors to the incorporation of globular proteins into pathological aggregates (Chiti et al., 2003; Ventura et al., 2004; Tartaglia et al., 2005; Calloni et al., 2005). Intriguingly, recent experiments demonstrated that several disease-related mutations in RBPs, such as TDP-43 and FUS, promote granule formation (Murakami et al., 2012; Ramaswami et al., 2013). By associating with their target transcripts, RBPs can influence protein production at different stages of the mRNA lifetime (transcription, translation, and mRNA degradation). This suggests that mRNP complexes, by rapidly undergoing extensive remodeling, can be responsible for both the spatial and temporal expression of a number of potentially dangerous targets, such as proteins with a high propensity to aggregate. In this context, theoretical approaches such as *cc*SOL (Chapters VI-VII) and *cat*RAPID (Chapters I-III) can be considered powerful instruments to rapidly assess the presence of insoluble/hydrophobic amino acid patches and their ability to interact with transcripts, respectively. In this thesis, I have shown (Chapters V-VII)

that solubility and aggregation propensities are tightly related to physico-chemical features such as hydrophobicity, secondary structure propensity and solvent accessibility (Tartaglia et al., 2005; Tartaglia and Vendruscolo, 2008; Vendruscolo and Tartaglia, 2008). Moreover, I have illustrated that similar physico-chemical features can be employed to predict protein and RNA associations and to define the regions involved in the interaction. Together with the evidence that non-canonical RBPs are enriched in unstructured and low-complexity sequence regions (Castello et al., 2012; Kwon et al., 2013), these findings indicate that solubility/aggregation and RNA-binding propensity can be analyzed simultaneously. Hence, the synergistic use of *cc*SOL and *cat*RAPID would enhance previously collected data on the general promiscuity of natively unfolded proteins in protein-protein interaction networks, at the protein-RNA level (Babu et al., 2011; Olzscha et al., 2011). Ultimately, it will provide a better understanding on the role of RBPs in maintaining homeostasis throughout post-transcriptional regulation of mRNAs, and of coding and non-coding transcripts in promoting the formation of RNP aggregates.

RNPs and control of gene expression

Previously, it has been observed that protein and RNA physico-chemical properties impose stringent conditions on their intracellular localization and expression levels (Tartaglia et al., 2009; Tartaglia and Vendruscolo, 2009). However, while evolutionary constraints on tissue-specific gene expression patterns have been extensively investigated (Brawand et al., 2011; Chan et al., 2009; Merkin et al., 2012; Ravasi et al., 2010), the regulation of RBP-mediated interactions is still poorly understood (Hogan et al., 2008; Masuda et al., 2012). In a recent publication, our group showed for the first time that *cat*RAPID *omics* predictions (Chapter III) can be integrated with expression profile data (Harrow et al., 2012; Uhlen et al., 2010) to

guide the discovery of distinct features of RBP biological functions (Cirillo et al., 2014). Specifically, we observed that an enrichment of unique and functionally related GO terms for RBP-mRNA pairs associates with high interaction propensities and specific expression patterns. Co-expression of interacting partners are linked to the constitutive regulation of processes such as proliferation and cell cycle control. In contrast, anti-expression patterns are likely to be a distinctive characteristic of specific spatiotemporal events involved in survival, growth and differentiation processes. Nevertheless, as spatial and temporal separation, and limited chemical reactivity could be ways to avoid aberrant associations (Quenneville et al., 2012), it is not possible to rule out the possibility that associations with lower interaction affinities might have relevant implications. As a matter of fact, proteins harboring disordered regions are highly reactive (Gsponer and Babu, 2012) and the degree of multimerization or low-complexity content can endow them with a wider spectrum of intermolecular affinities (Lunde et al., 2007). Moreover, increasing evidence indicates that unexpected protein features, such as multivalency and low complexity, are capable of driving the conversion between small complexes and large, dynamic, macromolecular assemblies (Han et al., 2012; Kato et al., 2012; Li et al., 2012). The latter are enriched in multivalent proteins and nucleic acids (Lunde et al., 2007; Parker and Sheth, 2007) and include a great variety of cellular structures such as Cajal bodies, P bodies and RNP granules (Buchan and Parker, 2009; Matera et al., 2009). Recruitment of RNA inside these granules, however, is not surprising, as non-canonical RBPs often bear multivalent domains in association with structurally disordered or low complexity regions (Castello et al., 2012; Kwon et al., 2013). By using *cat*RAPID on a number of disease-associated pathways, we found that structurally disordered regions are able to interact with transcripts (Cirillo et al., 2014; Zanzoni et al., 2013), a fact that corroborate the possible involvement of RNA in aggregation (Olzscha et al., 2011) and toxicity mechanisms (Vavouri et al., 2009). Furthermore, we observed that contribution of structural disorder to the interaction seems to be greater in proteins lacking classical

RBDs, thus indicating a putative mimicking activity of the RNA-binding function (Cirillo et al., 2014). In addition, it is possible that stable RNA secondary structures, especially those enriched in GC content, contribute to the spatial rearrangement of disordered protein regions (Zanzoni et al., 2013). Collectively, these results suggest that protein-RNA interactions, followed by transition into dynamic aggregates, need to be tightly regulated in order to control homeostasis and avoid potential damage. Indeed, it has been proposed that the packaging of cytoplasmic mRNA into discrete RNP granules regulates gene expression by delaying or preventing the translation of specific transcripts (Kedersha and Anderson, 2007). Currently, the boundaries that separate normal RNP granules assemblies from pathological transitions to amyloid structures are not clear. Therefore, the application of bioinformatics tools will be critical to understand the principles of assembly, disassembly, and clearance of RNP aggregates in both normal and pathological conditions. Ultimately, this will help to clarify the role of RNA in the pathology of neurodegenerative diseases and, possibly, it will suggests strategies for diagnostic and therapeutic interventions (Ramaswami et al., 2013).

In summary, the methods and ideas discussed here have been developed in an exciting moment of the post-genomic era (Harrow et al., 2012). Experimental and computational approaches have started to unveil the complexity of our genomes and RNA-protein interactions emerged as key events in a large number of regulatory processes (Altman, 2013). As shown in this thesis, albeit some work has been accomplished, there are still areas that need to be explored and discovered. New and ever more sophisticated algorithms will definitely help facing these challenges. I believe that my methods will provide valid assistance in carrying out the simultaneous investigation of RNA-binding ability and aggregation propensity. Understanding of these processes will be the key to elucidate the pathogenesis of several disorders, including neurodegeneration and cancer (Wolozin, 2012).

Conclusions

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

I The development of *cat*RAPID, the first sequence-based method to perform RNA-protein interaction predictions. The algorithm exploits the physico-chemical information derived from the primary structure of transcripts and proteins to compute their probability of interaction. The method was trained using ribonucleoprotein complexes acquired from the PDB and its performance assessed on a number of proteins and RNAs obtained from the NNBP and NPInter databases. The *cat*RAPID approach is suitable for the investigation of protein association with coding and non-coding transcripts;

II The application of the *cat*RAPID approach to investigate a number of regulatory mechanisms mediated by ribonucleoproteins. In particular, I hereby presented the interactions of the long non-coding Xist with several proteins involved in the X-inactivation process: EZH2, SUZ12, YY1, SAF-A/hnRNP-U and SATB1. To analyze the 17kb long Xist RNA, I implemented two additional modules to the main algorithm: i) *cat*RAPID fragments, to generate smaller fragments from

very large transcripts and ii) *cat*RAPID strength, to better appreciate the degree of interaction;

III The transition to large-scale analysis with *cat*RAPID *omics*, which offers unique features such as i) organism-specific proteomic and transcriptomic pre-generated libraries, ii) use of custom datasets, iii) analysis of long sequences and iv) identification of interaction specificities. The main advantages of this method are the fast calculation of ribonucleoprotein associations and the prediction of the RNA binding activity of proteins with high accuracy, thus resulting in a powerful tool for designing new experiments;

IV The development of *SeAMotE* to perform discriminative motif discovery in large sets of nucleic acid sequences. The approach offers features such as i) discrimination, based on the actual occurrences in the datasets, ii) multiple reference backgrounds (shuffle, random or custom) and iii) output of the most significant motifs in the whole span of tested motif widths;

V The application of physico-chemical properties to characterize the interaction network of DnaK, the major bacterial chaperone Hsp70, in *E. coli*. By using a number of features we were able to identify burial propensity as the attribute that best discriminates the set of proteins enriched on DnaK from the depleted and from the lysate. The result obtained in this study represents the starting point from which we derived the rationale to design *cc*SOL and other algorithms;

VI The development of *cc*SOL, a sequence-based method to predict the solubility of proteins in the *E. coli* expression system. This method exploits a SVM that was trained using approximately 70% of the *Escherichia coli* proteins, for which the solubility was experimentally measured *in vitro*. I demonstrated that 5 physico-chemical properties (coil/disorder, hydrophobicity, hydrophilicity, β-turn, α-helix) are sufficient to accurately infer protein solubility *in silico*;

VII The implementation of the *cc*SOL *omics* module, which performs large-scale analyses of proteins' solubility also providing information on: i) soluble and insoluble regions along the protein sequence, and ii) areas more susceptible to mutations. The algorithm employs a neural network trained on more than 30000 proteins obtained from the TargetTrack database and tested on an equal size dataset.

Appendix

Papers published during my doctoral studies

- **Agostini F.**, Cirillo D., Tartaglia GG. (2014) SeAMotE: a web-server for high-throughput motif discovery in nucleic acid sequences. *Manuscript Submitted*
- **Agostini F.**, Cirillo D., Livi CM., Tartaglia GG. (2014) ccSOL omics: a web server for large-scale prediction of protein solubility. *Manuscript Submitted*
- Cirillo D., Livi CM., **Agostini F.**, Tartaglia GG. (2014) Discovery of Protein-RNA Networks. *Molecular BioSystems.* Ahead of publication. DOI: 10.1039/C4MB00099D
- Klus P., **Agostini F.**, Bolognesi B., Zanzoni A. and Tartaglia GG. (2014) The cleverSuite Approach for Protein Characterization: Predictions of RNA-Binding Ability, Dosage Sensitivity, Solubility and Chaperone Requirements. *Bioinformatics (Oxford, England).* Ahead of publication. DOI: 10.1093/bioinformatics/btu074
- Cirillo D. and Marchese D., **Agostini F.**, Livi CM., Botta-Orfila T., Tartaglia GG. (2013) Constitutive patterns of gene expression regulated by RNA-binding proteins. *Genome Biol.* 15(1):R13.
- Zanzoni A., Marchese D., **Agostini F.**, Bolognesi B., Cirillo D., Botta-Orfila T., Livi CM., Rodriguez-Mulero S., Tartaglia, GG. (2013) Principles

of Self-Organization in Biological Pathways: A Hypothesis on the Autogenous Association of Alpha-Synuclein. *Nucleic Acids Res.* gkt794

- **Agostini F.**, Zanzoni A., Klus P., Marchese P., Cirillo D. Tartaglia GG. (2013) *cat*RAPID omics: a web server for large-scale prediction of protein-RNA interactions. *Bioinformatics (Oxford, England).* 29(22):2928–2930
- Cirillo D., **Agostini F.** and Tartaglia GG. (2013) Predictions of protein-RNA interactions. *Wiley Interdisciplinary Reviews: Computational Molecular Science.* 3, 161-175
- **Agostini F.**, Cirillo D., Bolognesi B. and Tartaglia GG. (2013) X-inactivation: quantitative predictions of protein interactions in the Xist network. *Nucleic Acids Res.* 41, e31
- Cirillo D. and **Agostini F.**, Klus P., Marchese D., Rodriguez-Mulero S., Bolognesi B. and Tartaglia GG. (2012) Neurodegenerative diseases: Quantitative predictions of protein-RNA interactions. *RNA.* 19, 129-140
- Calloni G., Chen T., Schermann SM., Chang HC., Genevaux P., **Agostini F.**, Tartaglia GG., Hayer-Hartl M. and Hartl FU. (2012) DnaK functions as a central hub in the *E. coli* chaperone network. *Cell Rep.* 1, 251-264
- **Agostini F.**, Vendruscolo M. and Tartaglia GG. (2012) Sequence-based prediction of protein solubility. *J. Mol. Biol.* 421, 237-241
- Mossuto MF., Bolognesi B., Guixer B., Dhulesia A., **Agostini F.**, Kumita JR., Tartaglia GG., Dumoulin M., Dobson CM. and Salvatella X. (2011) Disulfide bonds reduce the toxicity of the amyloid fibrils formed by an extracellular protein. *Angew. Chem. Int. Ed. Engl.* 50, 7048-7051
- Bellucci, M., **Agostini, F.**, Masin, M., and Tartaglia, G. G. (2011). Predicting protein associations with long noncoding RNAs. *Nature Methods*, 8(6):444–445

Supplementary material for Chapter I

Supplementary Methods

Training Set

Structural data were collected in March 2010 and consisted of 858 RNA-protein complexes (8367 protein-RNA pairs) available from the RCSB databank (http://www.pdb.org/). A cutoff of 7 Å for physical contacts was employed to discriminate between interacting and non-interacting protein-RNA pairs. The cutoff was decided according to the average resolution of structural complexes and led to define a positive dataset containing 7409 interacting protein-RNA pairs and a negative set containing 958 non-interacting protein-RNA pairs. The CD-HIT tool (http://weizhong-lab.ucsd.edu/cdhit_suite/cgi-bin/index.cgi) was used to filter out RNA and protein sequences with identities higher than 80% and 60%, respectively. After redundancy removal, the database contained 410 interacting ("Positive set") and 182 non-interacting ("Negative set") protein-RNA pairs. With regards to the composition of the Positive and Negative sets, protein-RNA associations were grouped into five functional classes: "Ribosome and protein synthesis", "Splicing", "Transcription", "tRNA synthesis and Viral RNA assemblies", which account for 70%, 10%, 8%, 12% and 10% of the entire training set. Performances were estimated using a ten-fold cross-validation approach, in which a representative set of each functional class was sampled. In the analysis, the data set of interactions was randomly partitioned into ten subsamples requiring the condition that all the partitions carry the same distribution of functional classes. One subsample was retained for testing, and the remaining nine were used for training the algorithm. The cross-validation process was repeated ten times with each of the ten subsamples used exactly once as the validation data. The significance of our predictions was evaluated by calculating p-values (two-tail t-test). See also section **Discriminative Power**.

We tested catRAPID's performance on the identification of binding regions. For each protein-RNA complex in the redundant set, we calculated interaction propensities of all possible associations between amino acid and nucleotide chains and ranked their scores from lowest to highest. Protein binding sites were top-ranked in 87% of cases while RNA binding sites were ranked in 75% of cases. Simultaneous identification of

Nature Methods: doi.10.1038/nmeth.1611

both protein and RNA binding regions was top-ranked in 62% of cases. Indeed, these results underline the extreme accuracy in identifying interaction sites (**Fig. 1a**).

Physico-chemical Properties

Secondary Structure Propensities. The secondary structure of the RNA molecule is predicted from its nucleotide sequence using the Vienna package[1] (including the algorithms RNAfold, RNAsubopt and RNAplot). Although the average predictive power of the RNAfold algorithm is 70%, lower performances are expected for long non-coding RNAs because these transcripts are poorly characterized. To increase the amount of information that can be extracted from secondary structure predictions, we adopted a strategy that exploits the generation of ensembles produced with the RNAsubopt algorithm. The sampling of structures was performed with probabilities estimated through Boltzmann weighting and stochastic backtracking in the partition function. Six model structures, ranked by energy, are used as input for catRAPID. For each model structure, the RNAplot algorithm was employed to generate secondary structure coordinates. Using the coordinates we defined the "secondary structure occupancy" by counting the number of contacts made by each nucleotide within the different regions of the chain. High values of secondary structure occupancy indicate that base pairing occurs in regions with high propensity to form hairpin-loops, while low values are associated with junctions or multi-loops. The secondary structure of proteins was taken into account in our model by calculating the Chou-Fasman[2] and Deleage-Roux[3] propensities for turn, β-strand and α-helical elements. As the average predictive power of these models is around 60%, we preferred to combine together the individual propensities to have better performances. The correlation between interaction propensities and secondary structure contributions is 73% (**Interaction Propensities**).

Hydrogen-Bonding Propensities. The structural information on purine and pyrimidine contacts was extracted from a set of 41 non-redundant protein-RNA complexes[4]. Both the number and the frequency of hydrogen-bond contacts are used in our method. With respect to proteins, we used Grantham's and Zimmerman's scales[5,6] to estimate the propensity of amino acids to form hydrogen bonds. Other propensity scales were disregarded because they showed lower predictive power.

Nature Methods: doi.10.1038/nmeth.1611

The correlation between interaction propensities and hydrogen bonding contributions is 58% (**Interaction Propensities**).

Van der Waals' Propensities. The information on purine and pyrimidine contacts was taken from a set of 41 protein-RNA complexes[4]. Both the number and the frequency of van der Waals' contacts were used in catRAPID. With respect to proteins, we employed Kyte-Dolittle and Bull-Breese scales[7,8] to estimate the propensity to form van der Waals' contacts. Other propensity scales were disregarded because they showed lower predictive power. The correlation between interaction propensities and Van der Waals' contributions is 26% (estimated with a ten-fold cross-validation).

Fitting coefficients for Secondary Structure, Hydrogen-Bonding and Van der Waals' Contributions are reported in **Supplementary Table 2**.

Interaction Propensity

Secondary structure, hydrogen bonding and van der Waals propensities were combined together into the interaction profile:

$$|\Phi_x\rangle = \alpha_S|S_x\rangle + \alpha_H|H_x\rangle + \alpha_W|W_x\rangle \quad (1)$$

We used the symbol $|\ \rangle$ to indicate the profile associated with a specific physico-chemical property. For example, the van der Waal's profile of a protein is denoted by $|W_p\rangle$ and contains the van der Waal's contributions of each amino acid:

$$|W_p\rangle = W_{p1}, W_{p2}, ..., W_{pL} \quad (2)$$

Where L is the protein's sequence length. Similarly, $|H\rangle$ represents the hydrogen bonding profile and $|S\rangle$ the secondary structure profile. The variable x is used to distinguish between RNA ($x = r$) and protein ($x = p$) profiles.

Nature Methods: doi.10.1038/nmeth.1611

In order to deal with molecules of different length, we approximated each propensity profile using plane-waves:

$$\tilde{\Phi}_x^k = \sqrt{\frac{2}{length}} \sum_{n=0}^{length} \Phi_x^n \cos\left[\frac{\pi}{length}\left(n+\frac{1}{2}\right)\left(k+\frac{1}{2}\right)\right] \qquad k = 0,1,\ldots L-1 \qquad (3)$$

The number of plane waves employed to approximate each profile is L = 50 as the discriminative power does not improve by increasing L.

The following condition was employed to derive the interaction matrix I:

$$\mathrm{I}:\ \max \left\langle \tilde{\Phi}_r \right| \mathrm{I} \left| \tilde{\Phi}_p \right\rangle \ \text{for}\ (r,p) \in \{\text{positive set}\} \qquad (4)$$

The interaction propensity score $\pi = \left\langle \tilde{\Phi}_r \right| \mathrm{I} \left| \tilde{\Phi}_p \right\rangle$ is defined as the inner product between the protein profile $\left| \tilde{\Phi}_r \right\rangle$ and the RNA profile $\left| \tilde{\Phi}_p \right\rangle$, weighted by the interaction matrix I:

$$\pi = \left\langle \tilde{\Phi}_r \right| \mathrm{I} \left| \tilde{\Phi}_p \right\rangle = \sum_{l,m} \tilde{\Phi}_r^l \, \mathrm{I}_{l,m} \tilde{\Phi}_p^m = \sum_{l,m} \tilde{\Lambda}_{l,m} \qquad (5)$$

The interaction propensity matrix $\Lambda_{l,m}$ is obtained by applying Eq. (3) to $\tilde{\Lambda}_{l,m}$.

The interaction matrix I is given by applying Eq. (3) to the parameters $\tilde{\mathrm{I}}_{n,k}$ reported in **Supplementary Table 3**.

Discriminative Power

In order to evaluate the ability of catRAPID to distinguish between interacting and non-interacting RNA-protein associations, we introduced the concept of discriminative power (dp):

Nature Methods: doi.10.1038/nmeth.1611

$$\mathrm{dp} = \frac{\sum_i \sum_n \vartheta(\pi_i - \pi_n)}{\sum_i \sum_n \vartheta(\pi_i - \pi_n) + \vartheta(\pi_n - \pi_i)} = 1 - (\mathrm{I} \cap \mathrm{N}) \quad (6)$$

Where π_i indicates the interaction propensity of an interacting RNA-protein pairs, π_n represents the interaction propensity of non-interacting molecules, I is the score distribution associated with the positive set and N is the score distribution associated with the negative set. The definition of π is given in the section **Interaction Propensity**. The function $\vartheta(\pi_i - \pi_n)$ is 1 if $\pi_i - \pi_n > 0$ and 0 otherwise. According to the definition given in Eq. (6), the discriminative power ranges from 0% to 100%. The significance of predictions was evaluated by calculating p-values (two-tail t-test).

With regards to catRAPID's performances, the discriminative power associated with the non-redundant training dataset is 78%. The discriminative power associated with the redundant training dataset is 90%. If a consistent number of protein or RNA sequences are moved from the negative to the positive set (or vice-versa), the distribution of interaction propensities associated with the positive and negative sets tend to overlap. When the number of sequences transferred from the negative to the positive set equals half the size of the positive set, dp is 42%. If Fourier's coefficients associated with RNA or protein sequences are scrambled (i.e., their order is modified in a random way), dp is < 50%. If we use the unitary matrix in Eq. 3, the algorithm shows a dp of 65% on the training set, which increases up to 71% when the NPInter dataset is also considered.

Interaction Propensity

Using the score distribution f_n associated with the negative training set, we calculated the probability $p(v) = p(\pi \leq v)$ that the score π takes values less than or equal to v (interaction probability):

$$p(v) = \int_{-\infty}^{v} f_n(\pi) d\pi \quad (7)$$

Nature Methods: doi.10.1038/nmeth.1611

Similarly, using the score distribution f_p of the positive training set, we estimated the probability that the score π takes values more than or equal to v (non-interaction probability):

$$n(v) = \int_{v}^{\infty} f_p(\pi) d\pi \quad (8)$$

The two probabilities $p(v)$ and $n(v)$ were then combined together to define the interaction propensity $P(v)$:

$$P(v,x) = \frac{x[1-n(v)]p(v)}{[1-n(v)]p(v)[1-x] + x[1-p(v)]n(n)} \quad (9)$$

where $x = 0.5$

Test Sets

The NPInter database[9] (http://www.bioinfo.org.cn/NPInter/) was used to evaluate the ability of the algorithm to predict interactions between proteins and long non-coding RNAs. RNA sequences were obtained from the fRNAdb database (http://www.ncrna.org/frnadb/). We excluded micro-RNAs from our analysis because their size significantly differs from that of molecules used for training. The long non-coding database contains 405 interactions from 6 model organisms. Only for a subset of the NPInter database direct physical evidence for protein-RNA interactions is reported (Fig. 1b; class "The ncRNA binds the protein" accounting for 59% of the NPInter dataset and class "The protein as a factor affects the ncRNA's function" accounting for 22% of the NPInter dataset). We also estimated the significance of our predictions on the entire database by calculating p-values (two-tail t-test): 0.04 for class "The ncRNA is regulated by the protein", 0.21 for class "Special linkage between the ncRNA and the Protein" 0.11 for class "Genetic interaction between the ncRNA gene and the protein", 0.03 for class "The ncRNA regulates the mRNA"), 0.20 for class "The ncRNA indirectly regulates a gene" and 0.6 for class "The ncRNA as a

Nature Methods: doi.10.1038/nmeth.1611

factor affects the protein's function". The average discriminative power is 85% and was evaluated by comparing the interaction propensities of the different NPInter classes with the interaction propensities of the non-redundant negative set (and increases up to 90% by comparing with the redundant negative set).

The Non-Nucleid-acid-Binding database NNBP[10] was employed to evaluate the ability of catRAPID to identify proteins that have little propensity to interact with RNA molecules. The original set comprises 246 proteins, among which 62 were selected after a search on the Uniprot database (http://www.uniprot.org/) for molecules that are exclusively involved in protein-protein interactions. A total of 12000 random associations were generated with RNA sequences of the positive set. The discriminative power of the algorithm was evaluated by comparing the interaction propensities of the negative set (**Training Set**) with those of the random list. The significance of predictions was evaluated by calculating p-values (two-tail t-test) (Supplementary Table 4).

DNA-binding (DNA BP) and RNA-binding (RNA BP) proteins were obtained from the Uniprot database. DNA BP were collected by searching for molecules that bind "with DNA and not with RNA" (7535 hits), while RNA BP were obtained by selecting molecules that bind "with RNA and not with DNA" (84 hits). The CD-HIT tool was used to filter out sequences with identities higher than 60%. After filtering we counted a total of 5410 entries for DNA BP and 65 entries for RNA BP). Random associations were generated with RNA sequences present in the positive training set (130000 associations for DNA-binding and 12000 for RNA-binding, respectively). The discriminative power of the algorithm was evaluated by comparing interaction propensities of the negative set (**Training Set**) with those of the random lists. The significance of predictions was evaluated by calculating p-values (two-tail t-test) (Supplementary Table 5).

The Human MRP and RNase P Complexes

The human MRP complex is comprised of ten protein subunits (hPop1, hPop5, Rpp14, Rpp20, Rpp21, Rpp25, Rpp29, Rpp30, Rpp38 and Rpp40) and one RNA unit (266 nucleotides). The RNA shows a catalytic core domain with evolutionary

Nature Methods: doi.10.1038/nmeth.1611

conserved structural features in domain I (P1-P3 helices), and a variable portion named domain II (P8, P9, P12, eP19 helices) with unknown function. The human RNase P complex shares protein components with the MRP system. It includes one RNA unit (344 nucleotides) that possesses analogous structural features compared to the *MRP RNA*, with a more extended P12 stem and additional P7, P10, P11 elements. The two complexes display different catalytic activities: MRP mediates the processing of rRNA precursors while RNAse P is required for processing pre-tRNAs in functional tRNAs molecules.

Several studies were carried out to identify protein-RNA interactions in human, yeast and bacterial MRP complexes, using a wide variety of techniques[11]. The most detailed picture of the human system was given by Welting and coworkers[12] who demonstrated, using GST pull-down data, that hPop1, Rpp20, Rpp21, Rpp25, Rpp29 and Rpp38 directly interact with RNA, whereas hPop5 and Rpp14 are part of the assembly but do not contact the transcript. Interaction data for Rpp30 and Rpp40 are missing because of the poor solubility of the proteins. It has been observed that Rpp20 and Rpp25 bind strictly to the P3 helix, whereas Rpp29 mediate additional contacts in the P12 stem by associating with more than one RNA region. The interaction between RNA, Rpp20 and Rpp25 was confirmed by the very recent release of the crystal structure of the *MRP RNA* P3 stem in complex with yeast homologues of Rpp20 and Rpp25[13].

Comparisons between our predictions and experimental evidences can be summarized as follows (**Supplementary Table 6**, **Supplementary Fig. 1**): i) Rpp20 and Rpp21 binds the P3 stem that can be considered a nucleation center. The predicted binding region for Rpp20 - *MRP RNA* corresponds to the one observed in the crystal structure of yeast *MRP RNA* P3 portion in complex with the yeast homolog POP7[13]. ii) Rpp29 and Rpp38 mediate multiple interactions between P3 helix and P12 stem. These results are in complete agreement with the known interaction map of Rpp29 which connects domain I and II[12]. iii) Rpp25 is predicted to have lower propensity to interact with RNA. This finding can be explained by considering that Rpp25 is able to recognize the P3 element of *MRP RNA* only after association with Rpp20[14]. iv) Rpp14, Rpp30 and Rpp40 are predicted to be non-interacting with MRP RNA, in agreement with what was reported in literature[12]. v) hPop5 is predicted to mediate weak interactions with the MRP RNA in the P3 area.

Nature Methods: doi.10.1038/nmeth.1611

This finding is in accordance with activity assays conducted on the archeal homolog *Pho*Pop5[15].

With regards to the RNase P system, similar interaction propensities were found for Rpp20, Rpp21, Rpp25, Rpp29 and Rpp38 (**Supplementary Fig. 2**). In general, an increase in the intensity of signals is observed together with an enhanced binding preference for the P3 stem region. This finding could be explained by considering the different substrate specificity and catalytic activity of the two RNA-protein assemblies.

Association of the PRC-2 with *Xist* and *HOTAIR*

The Polycomb Repressive Complex is comprised of four protein units: Ezh2, Eed, Suz12 and Rbap48. Ezh2 and Eed are predicted by catRAPID to contact approximately the same RNA regions (330-680 and 330-530 for *Xist* A Region; 1-240 and 1-220 for the 5' domain of *HOTAIR*; **Supplementary Fig. 3**), which is well in agreement with the ability of these proteins to heterodimerize[16]. Eed shows similar binding propensities with both 2R (431-531; **Supplementary Fig. 3**) and 4R (371-531; **Supplementary Fig. 3**) segments, as shown by immuno-precipitation assays[17]. According to previous experimental evidences[18] and in agreement with our predictions on repeat regions, Ezh2 can be regarded as the main RNA-binding subunit, representing the catalytic core of the PCR2 complex. Higher propensity to bind 2R is found for Rbap48, which might arise from its involvement in mediating protein-protein interactions in addition to RNA binding[19].

Databases used for MRP, Xist and HOTAIR

RNA sequences (human *MRP RNA*, FR355912; human *RNase P RNA*, FR174566) were downloaded from the fRNAdb database (http://www.ncrna.org/frnadb/). Protein sequences were retrieved from Uniprot database (hPop5, Q969H6; Rpp14, O95059; Rpp20, O75817; Rpp21, Q9H633; Rpp25, Q9BUL9; Rpp29, O95707; Rpp30, P78346; Rpp38, P78345; Rpp40, O75818). The catRAPID algorithm was employed to predict the interaction propensity of all protein subunits except for hPop1 whose large size does not fit with our computational requirements. The three-dimensional structure of the *MRP* P3 domain in complex with POP6-POP7 was displayed using

Nature Methods: doi.10.1038/nmeth.1611

the UCSF Chimera visualization tool (http://www.cgl.ucsf.edu/chimera/). The crystal structure of the yeast *MRP* P3 domain in complex with the POP6-POP7 protein heterodimer (PDB code: 3iab) was released in July 2010.

The RNA sequences of human *Xist* (M97168.1) and *HOTAIR* (DQ926657.1) were downloaded from the NCBI database. Regions of interest were selected on the basis of available experimental data (sequence numbering is reported): *Xist* A Region, 330-796; *Xist* 4R, 371-531; 5' *HOTAIR*, 1-300; 3' *HOTAIR*, 1500-2146. The catRAPID algorithm was used to predict the interaction propensity of the four PRC2 protein subunits, whose Uniprot IDs are: Ezh2, Q15910; Eed, O75530; Suz12, Q15022; Rbap48, Q09028.

Nature Methods: doi.10.1038/nmeth.1611

References

1. Gruber, A.R., Lorenz, R., Bernhart, S.H., Neubock, R. & Hofacker, I.L. The Vienna RNA Websuite. *Nucleic Acids Research* **36**, W70-W74 (2008).
2. Chou, P.Y. & Fasman, G.D. Prediction of the secondary structure of proteins from their amino acid sequence. *Adv. Enzymol. Relat. Areas Mol. Biol* **47**, 45-148 (1978).
3. Deléage, G. & Roux, B. An algorithm for protein secondary structure prediction based on class prediction. *Protein Eng* **1**, 289-294 (1987).
4. Morozova, N., Allers, J., Myers, J. & Shamoo, Y. Protein-RNA interactions: exploring binding patterns with a three-dimensional superposition analysis of high resolution structures. *Bioinformatics* **22**, 2746-2752 (2006).
5. Grantham, R. Amino acid difference formula to help explain protein evolution. *Science* **185**, 862-864 (1974).
6. Zimmerman, J.M., Eliezer, N. & Simha, R. The characterization of amino acid sequences in proteins by statistical methods. *J. Theor. Biol* **21**, 170-201 (1968).
7. Kyte, J. & Doolittle, R.F. A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol* **157**, 105-132 (1982).
8. Bull, H.B. & Breese, K. Surface tension of amino acid solutions: a hydrophobicity scale of the amino acid residues. *Arch. Biochem. Biophys* **161**, 665-670 (1974).
9. Wu, T. et al. NPInter: the noncoding RNAs and protein related biomacromolecules interaction database. *Nucleic Acids Res* **34**, D150-152 (2006).
10. Stawiski, E.W., Gregoret, L.M. & Mandel-Gutfreund, Y. Annotating nucleic acid-binding function based on protein structure. *J. Mol. Biol* **326**, 1065-1079 (2003).
11. Esakova, O. & Krasilnikov, A.S. Of proteins and RNA: the RNase P/MRP family. *RNA* **16**, 1725-1747 (2010).
12. Welting, T.J.M., van Venrooij, W.J. & Pruijn, G.J.M. Mutual interactions between subunits of the human RNase MRP ribonucleoprotein complex. *Nucleic Acids Res* **32**, 2138-2146 (2004).
13. Perederina, A., Esakova, O., Quan, C., Khanova, E. & Krasilnikov, A.S. Eukaryotic ribonucleases P/MRP: the crystal structure of the P3 domain. *EMBO J* **29**, 761-769 (2010).
14. Hands-Taylor, K.L.D. et al. Heterodimerization of the human RNase P/MRP subunits Rpp20 and Rpp25 is a prerequisite for interaction with the P3 arm of RNase MRP RNA. *Nucleic Acids Res* **38**, 4052-4066 (2010).
15. Tsai, H., Pulukkunat, D.K., Woznick, W.K. & Gopalan, V. Functional reconstitution and characterization of Pyrococcus furiosus RNase P. *Proceedings of the National Academy of Sciences* **103**, 16147 -16152 (2006).
16. Han, Z. et al. Structural basis of EZH2 recognition by EED. *Structure* **15**, 1306-1315 (2007).
17. Maenner, S. et al. 2-D structure of the A region of Xist RNA and its implication for PRC2 association. *PLoS Biol* **8**, e1000276 (2010).

Nature Methods: doi.10.1038/nmeth.1611

18. Zhao, J., Sun, B.K., Erwin, J.A., Song, J. & Lee, J.T. Polycomb proteins targeted by a short repeat RNA to the mouse X chromosome. *Science* **322**, 750-756 (2008).
19. Qian, Y.W. et al. A retinoblastoma-binding protein related to a negative regulator of Ras in yeast. *Nature* **364**, 648-652 (1993).
20. Welting, T.J., Kikkert, B.J., van Venrooij, W.J. & Pruijn, G.J. Differential association of protein subunits with the human RNase MRP and RNase P complexes. *RNA* **12**, 1373-1382 (2006).
21. Honda, T., Hara, T., Nan, J., Zhang, X. & Kimura, M. Archaeal homologs of human RNase P protein pairs Pop5 with Rpp30 and Rpp21 with Rpp29 work on distinct functional domains of the RNA subunit. *Biosci. Biotechnol. Biochem* **74**, 266-273 (2010).

Nature Methods: doi.10.1038/nmeth.1611

Supplementary material for Chapter II

a

RNA	sequence	reference	protein	method	top peak or strength%
Xist (NR_001463.2)	1–15890	6	SAF-A	Fragments	198-306; 2790-2970; 4934–5056
			SATB1		198-306; 4223-4336; 13883-14013
	1-17919	7	SFRS1		682-881; 707-826; 726-907
RepA	227-760	7	SUZ12	Protein and RNA strengths	99%
			EZH2		75%
		9	YY1		0%
			GFP		4%
	292-698	11	SATB1		70%
	164-932	8	SFRS1		92%
RepC	3084-4940	9	YY1		77%
Tsix R1 (NR_002844.2)	2073-2239	5	EZH2		99%

b

	name	sequence	reference
experimental fragments	AF	1–2406	9
	B	1898–3083	
	BC	1898–4940	
	C	3084–4940	
	eE1	6990–9467	
	2R	318-521	7
	4R	401-552	
localization signals	*A-D*	292-698	6
	B-Ev	1899-3488	
	Nc-H	4725-6079	
RT-PCR primers	X2	2339-2516	10
	X3	5125-5227	

c

name	length	UniProtKB AC
EZH2	746	Q61188
SUZ12	741	Q80U70
SFRS1	248	Q6PDM2
YY1	414	Q00899
SAF-A$_{50-800}$ SAF-A$_{9-757}$	750	Q8VEK3
SATB1$_{23-764}$	742	Q60611
GFP	238	P42212

Table S1: *Sequence information*. a) Xist and Tsix regions, protein names, methods and predictions; b) Xist fragments, localization signals (restriction enzymes) and primers; c) Proteins, sequence lengths and Uniprot codes.

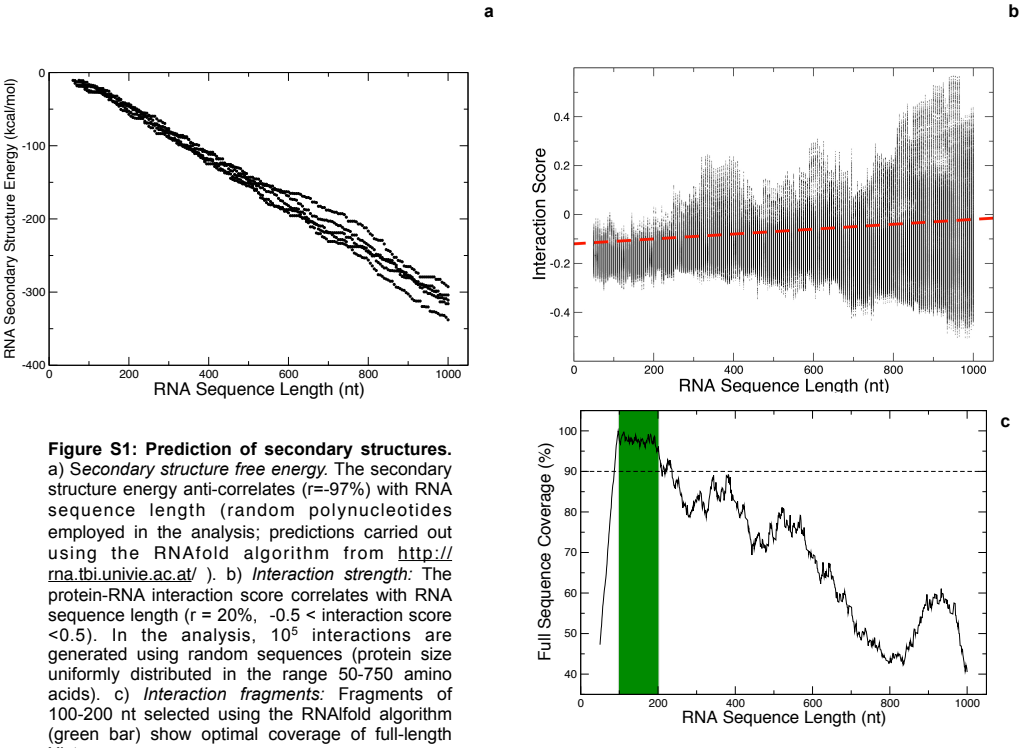


Figure S1: Prediction of secondary structures. a) S*econdary structure free energy.* The secondary structure energy anti-correlates (r=-97%) with RNA sequence length (random polynucleotides employed in the analysis; predictions carried out using the RNAfold algorithm from http://rna.tbi.univie.ac.at/). b) *Interaction strength:* The protein-RNA interaction score correlates with RNA sequence length (r = 20%, -0.5 < interaction score <0.5). In the analysis, 10^5 interactions are generated using random sequences (protein size uniformly distributed in the range 50-750 amino acids). c) *Interaction fragments:* Fragments of 100-200 nt selected using the RNAlfold algorithm (green bar) show optimal coverage of full-length Xist sequence.

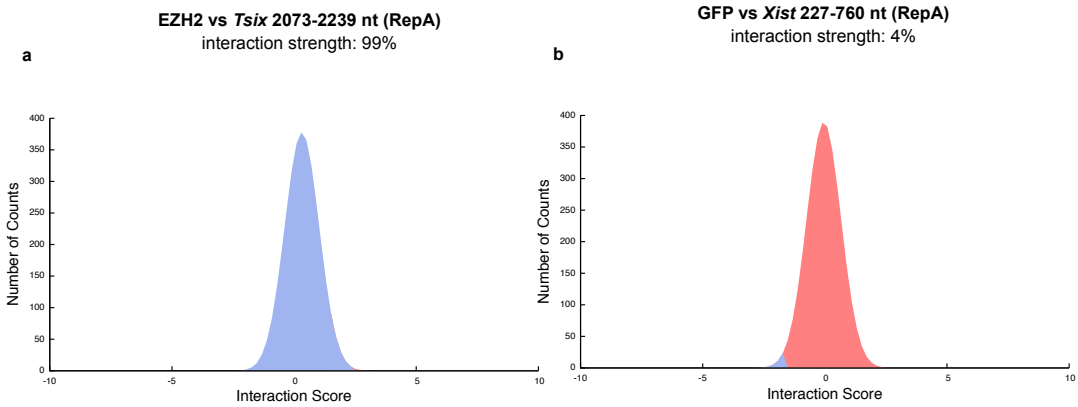
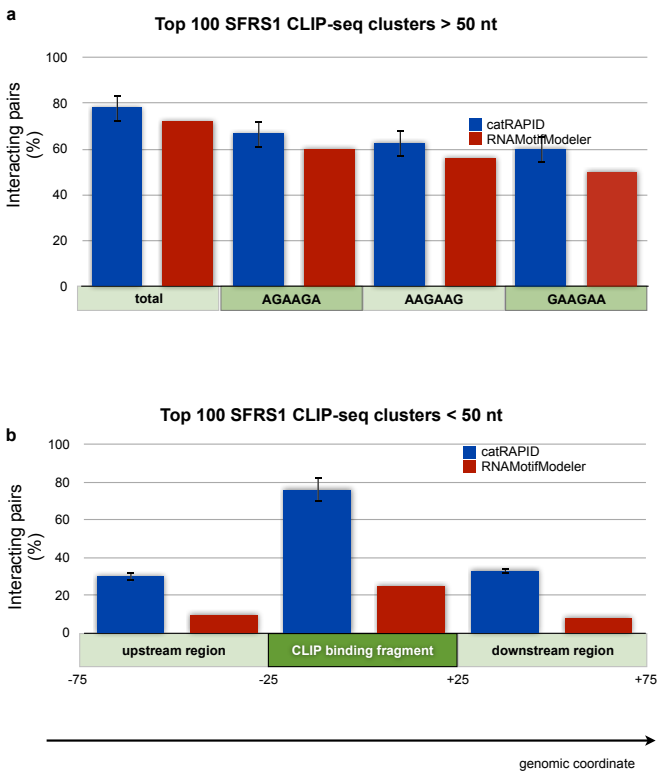


Figure S2: Positives and negative interactions. a) *Tsix* and EZH2. In agreement with experimental evidence, we predict that EZH2 binds to Tsix nucleotides 2073-2239, which correspond to an antisense region of RepA (5). b) Green Fluorescent Protein (GFP) and Xist nucleotides 227-760 (RepA). GFP and RepA are predicted to have poor propensity to interact, as previously reported (9).

Figure S3: *SFRS1 and its RNA interactome.* a) *SFRS1 and CLIP-seq clusters > 50 nt.* The top 100 CLIP-seq large clusters (i.e., RNA sequences containing the highest number of SFRS1 binding sites > 50 nt) are predicted to have strong propensity to bind SFRS1. catRAPID predictions (blue bars; interacting pairs = 78; average interaction strength =69%) are in good agreement with *RNAMotifModeler* performances (red bars; binding motifs AGAAGA, AAGAAG and GAAGAA present in 72 sequences) (19); b) *SFRS1 and CLIP-seq clusters < 50 nt.* The top 100 CLIP-seq regions < 50 nt show higher interaction strengths than the upstream and downstream flanking regions (binding sites: 76 protein-RNA pairs predicted by catRAPID; 25 motifs found by *RNAMotifModeler*; flanking regions: 30 protein-RNA pairs predicted by catRAPID; 10 motifs found by *RNAMotifModeler*).



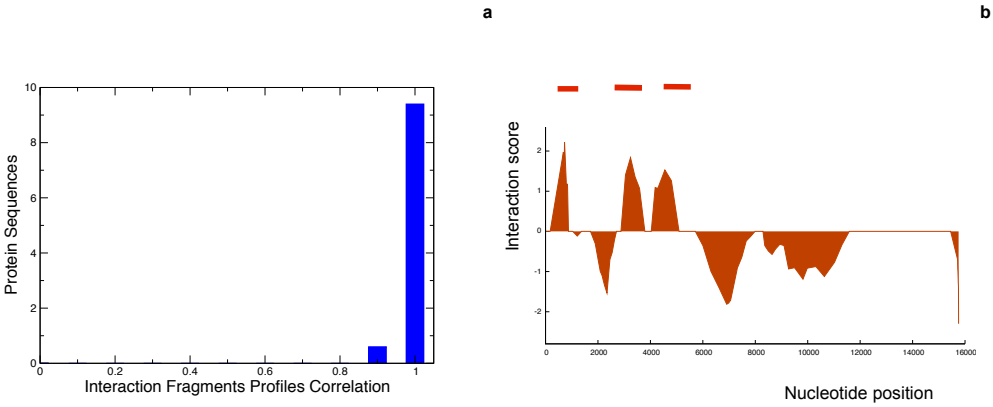


Figure S4: *Xist* interactions with SAF-A. a) *Correlation between interaction fragment profiles.* High correlation is found between SAF-$A_{50\text{-}800}$ and SAF-$A_{x\text{-}750+x}$ (x is the first amino acid in the window sliding from N- to C-terminus) b) *SAF-$A_{9\text{-}759}$ Interaction fragments profile.* SAF-$A_{9\text{-}759}$ is predicted to have three binding regions in correspondence to *Xist* nucleotides 298-678,1899-3488 and 4725-6079, which correspond to the localization signals identified by Wutz *et al.* (6).

Glossary

Intrinsically disordered region (IDR) Part of protein sequence that cannot accept stable secondary or tertiary structures. Disordered regions are very flexible and often serve as binding sites for other proteins.

Long non-coding RNA (lncRNA) Transcripts longer than 200 nucleotides that have little or no protein-coding capacity.

Low-complexity sequence (LCS) Sequence region of biased nucleotide or amino acid composition. Low-complexity regions in amino acid sequences typically assume non-globular structure in proteins.

Physico-chemical property Physical molecular property of a compound. Typical properties are solubility, acidity, lipophilicity, polar surface area, shape, flexibility, etc..

Protein aggregate An abnormal protein assembly that results from the cohesion of two or more misfolded monomeric proteins.

Ribonucleoprotein complex (RNP complex) A multimolecular complex that is composed of RNAs and associated proteins.

RNA granule Macromolecular structure enriched with RNA and RNA-binding proteins, thought to be involved in the preservation and transport of mRNA.

RNA recognition element (RRE) Sequence or structural elements embedded in the target RNAs bound by specific RBPs.

RNA-binding protein (RBP) Protein that bind to RNA through an RNA-binding motif. The binding may regulate the translation of RNA or induce post-transcriptional changes, such as RNA splicing and editing.

Stress granules Dense cytosolic protein and RNA aggregations that appear under conditions of cellular stress.

Bibliography

Alberti, S. (2013). Aggregating the message to control the cell cycle. *Developmental Cell*, 25(6):551–552.

Altman, S. (2013). The RNA–Protein world. *RNA*, 19(5):589–590. PMID: 23592800.

Anderson, P. and Kedersha, N. (2008). Stress granules: the tao of RNA triage. *Trends in biochemical sciences*, 33(3):141–150. PMID: 18291657.

Ankö, M.-L., Morales, L., Henry, I., Beyer, A., and Neugebauer, K. M. (2010). Global analysis reveals SRp20- and SRp75-specific mRNPs in cycling and neural cells. *Nature structural & molecular biology*, 17(8):962–970. PMID: 20639886.

Ankö, M.-L. and Neugebauer, K. M. (2012). RNA-protein interactions in vivo: global gets specific. *Trends in biochemical sciences*, 37(7):255–262. PMID: 22425269.

Ascano, M., Gerstberger, S., and Tuschl, T. (2013). Multi-disciplinary methods to define RNA-Protein interactions and their regulatory networks. *Current opinion in genetics & development*, 23(1):20–28. PMID: 23453689 PMCID: PMC3624891.

Ascano, M., Hafner, M., Cekan, P., Gerstberger, S., and Tuschl, T. (2012). Identification of RNA-protein interaction networks using PAR-CLIP. *Wiley interdisciplinary reviews. RNA*, 3(2):159–177. PMID: 22213601 PMCID: PMC3711140.

Ayala, Y. M., De Conti, L., Avendaño-Vázquez, S. E., Dhir, A., Romano, M., D'Ambrogio, A., Tollervey, J., Ule, J., Baralle, M., Buratti, E., and Baralle, F. E. (2011). TDP-43 regulates its mRNA levels through a negative feedback loop. *The EMBO journal*, 30(2):277–288. PMID: 21131904 PMCID: PMC3025456.

Babu, M. M., van der Lee, R., de Groot, N. S., and Gsponer, J. (2011). Intrinsically disordered proteins: regulation and disease. *Current opinion in structural biology*, 21(3):432–440. PMID: 21514144.

Bailey, T. L. (2011). DREME: motif discovery in transcription factor ChIP-seq data. *Bioinformatics (Oxford, England)*, 27(12):1653–1659. PMID: 21543442 PMCID: PMC3106199.

Bailey, T. L. and Elkan, C. (1994). Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proceedings / ... International Conference on Intel-*

ligent Systems for Molecular Biology ; ISMB. International Conference on Intelligent Systems for Molecular Biology, 2:28–36. PMID: 7584402.

Baltz, A. G., Munschauer, M., Schwanhäusser, B., Vasile, A., Murakawa, Y., Schueler, M., Youngs, N., Penfold-Brown, D., Drew, K., Milek, M., Wyler, E., Bonneau, R., Selbach, M., Dieterich, C., and Landthaler, M. (2012). The mRNA-bound proteome and its global occupancy profile on protein-coding transcripts. *Molecular cell*, 46(5):674–690. PMID: 22681889.

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*, 69(21):3613–3634. PMID: 22538991 PMCID: PMC3474909.

Bejerano, G., Pheasant, M., Makunin, I., Stephen, S., Kent, W. J., Mattick, J. S., and Haussler, D. (2004). Ultraconserved elements in the human genome. *Science (New York, N.Y.)*, 304(5675):1321–1325. PMID: 15131266.

Bellucci, M., Agostini, F., Masin, M., and Tartaglia, G. G. (2011). Predicting protein associations with long noncoding RNAs. *Nature methods*, 8(6):444–445. PMID: 21623348.

Berman, H. M., Westbrook, J. D., Gabanyi, M. J., Tao, W., Shah, R., Kouranov, A., Schwede, T., Arnold, K., Kiefer, F., Bordoli, L., Kopp, J., Podvinec, M., Adams, P. D., Carter, L. G., Minor, W., Nair, R., and La Baer, J. (2009). The protein structure initiative structural genomics knowledgebase. *Nucleic acids research*, 37(Database issue):D365–368. PMID: 19010965 PMCID: PMC2686438.

Bernhardt, H. S. (2012). The RNA world hypothesis: the worst theory of the early evolution of life (except for all the others)(a). *Biology direct*, 7:23. PMID: 22793875 PMCID: PMC3495036.

Bernhart, S. H., Hofacker, I. L., and Stadler, P. F. (2006). Local RNA base pairing probabilities in large sequences. *Bioinformatics (Oxford, England)*, 22(5):614–615. PMID: 16368769.

Bernstein, E. and Allis, C. D. (2005). RNA meets chromatin. *Genes & development*, 19(14):1635–1655. PMID: 16024654.

Birney, E., Stamatoyannopoulos, J. A., Dutta, A., Guigó, R., Gingeras, T. R., Margulies, E. H., Weng, Z., Snyder, M., Dermitzakis, E. T., Stamatoyannopoulos, J. A., Thurman, R. E., Kuehn, M. S., Taylor, C. M., Neph, S., Koch, C. M., Asthana, S., Malhotra, A., Adzhubei, I., Greenbaum, J. A., Andrews, R. M., Flicek, P., Boyle, P. J., Cao, H., Carter, N. P., Clelland, G. K., Davis, S., Day, N., Dhami, P., Dillon, S. C., Dorschner, M. O., Fiegler, H., Giresi, P. G., Goldy, J., Hawrylycz, M., Haydock, A., Humbert, R., James, K. D., Johnson, B. E., Johnson, E. M., Frum, T. T., Rosenzweig, E. R., Karnani, N., Lee, K., Lefebvre, G. C., Navas, P. A., Neri, F., Parker, S. C. J., Sabo,

P. J., Sandstrom, R., Shafer, A., Vetrie, D., Weaver, M., Wilcox, S., Yu1, M., Collins, F. S., Dekker, J., Lieb, J. D., Tullius, T. D., Crawford, G. E., Sunyaev, S., Noble, W. S., Dunham, I., Dutta, A., Guigó, R., Denoeud, F., Reymond, A., Kapranov, P., Rozowsky, J., Zheng, D., Castelo, R., Frankish, A., Harrow, J., Ghosh, S., Sandelin, A., Hofacker, I. L., Baertsch, R., Keefe, D., Flicek, P., Dike, S., Cheng, J., Hirsch, H. A., Sekinger, E. A., Lagarde, J., Abril, J. F., Shahab, A., Flamm, C., Fried, C., Hackermüller, J., Hertel, J., Lindemeyer, M., Missal, K., Tanzer, A., Washietl, S., Korbel, J., Emanuelsson, O., Pedersen, J. S., Holroyd, N., Taylor, R., Swarbreck, D., Matthews, N., Dickson, M. C., Thomas, D. J., Weirauch, M. T., Gilbert, J., Drenkow, J., Bell, I., Zhao, X., Srinivasan, K. G., Sung, W.-K., Ooi, H. S., Chiu, K. P., Foissac, S., Alioto, T., Brent, M., Pachter, L., Tress, M. L., Valencia, A., Choo, S. W., Choo, C. Y., Ucla, C., Manzano, C., Wyss, C., Cheung, E., Clark, T. G., Brown, J. B., Ganesh, M., Patel, S., Tammana, H., Chrast, J., Henrichsen, C. N., Kai, C., Kawai, J., Nagalakshmi, U., Wu, J., Lian, Z., Lian, J., Newburger, P., Zhang, X., Bickel, P., Mattick, J. S., Carninci, P., Hayashizaki, Y., Weissman, S., Dermitzakis, E. T., Margulies, E. H., Hubbard, T., Myers, R. M., Rogers, J., Stadler, P. F., Lowe, T. M., Wei, C.-L., Ruan, Y., Snyder, M., Birney, E., Struhl, K., Gerstein, M., Antonarakis, S. E., Gingeras, T. R., Brown, J. B., Flicek, P., Fu, Y., Keefe, D., Birney, E., Denoeud, F., Gerstein, M., Green, E. D., Kapranov, P., Karaöz, U., Myers, R. M., Noble, W. S., Reymond, A., Rozowsky, J., Struhl, K., Siepel, A., Stamatoyannopoulos, J. A., Taylor, C. M., Taylor, J., Thurman, R. E., Tullius, T. D., Washietl, S., Zheng, D., Liefer, L. A., Wetterstrand, K. A., Good, P. J., Feingold, E. A., Guyer, M. S., Collins, F. S., Margulies, E. H., Cooper, G. M., Asimenos, G., Thomas, D. J., Dewey, C. N., Siepel, A., Birney, E., Keefe, D., Hou, M., Taylor, J., Nikolaev, S., Montoya-Burgos, J. I., Löytynoja, A., Whelan, S., Pardi, F., Massingham, T., Brown, J. B., Huang, H., Zhang, N. R., Bickel, P., Holmes, I., Mullikin, J. C., Ureta-Vidal, A., Paten, B., Seringhaus, M., Church, D., Rosenbloom, K., Kent, W. J., Stone, E. A., Gerstein, M., Antonarakis, S. E., Batzoglou, S., Goldman, N., Hardison, R. C., Haussler, D., Miller, W., Pachter, L., Green, E. D., Sidow, A., Weng, Z., Trinklein, N. D., Fu, Y., Zhang, Z. D., Karaöz, U., Barrera, L., Stuart, R., Zheng, D., Ghosh, S., Flicek, P., King, D. C., Taylor, J., Ameur, A., Enroth, S., Bieda, M. C., Koch, C. M., Hirsch, H. A., Wei, C.-L., Cheng, J., Kim, J., Bhinge, A. A., Giresi, P. G., Jiang, N., Liu, J., Yao, F., Sung, W.-K., Chiu, K. P., Vega, V. B., Lee, C. W. H., Ng, P., Shahab, A., Sekinger, E. A., Yang, A., Moqtaderi, Z., Zhu, Z., Xu, X., Squazzo, S., Oberley, M. J., Inman, D., Singer, M. A., Richmond, T. A., Munn, K. J., Rada-Iglesias, A., Wallerman, O., Komorowski, J., Clelland, G. K., Wilcox, S., Dillon, S. C., Andrews, R. M., Fowler, J. C., Couttet, P., James, K. D., Lefebvre, G. C., Bruce, A. W., Dovey, O. M., Ellis, P. D., Dhami, P., Langford, C. F., Carter, N. P., Vetrie, D., Kapranov, P., Nix, D. A., Bell, I., Patel, S., Rozowsky, J., Euskirchen, G.,

Hartman, S., Lian, J., Wu, J., Urban, A. E., Kraus, P., Calcar, S. V., Heintzman, N., Kim, T. H., Wang, K., Qu, C., Hon, G., Luna, R., Glass, C. K., Rosenfeld, M. G., Aldred, S. F., Cooper, S. J., Halees, A., Lin, J. M., Shulha, H. P., Zhang, X., Xu, M., Haidar, J. N. S., Yu, Y., Birney*, E., Weissman, S., Ruan, Y., Lieb, J. D., Iyer, V. R., Green, R. D., Gingeras, T. R., Wadelius, C., Dunham, I., Struhl, K., Hardison, R. C., Gerstein, M., Farnham, P. J., Myers, R. M., Ren, B., Snyder, M., Thomas, D. J., Rosenbloom, K., Harte, R. A., Hinrichs, A. S., Trumbower, H., Clawson, H., Hillman-Jackson, J., Zweig, A. S., Smith, K., Thakkapallayil, A., Barber, G., Kuhn, R. M., Karolchik, D., Haussler, D., Kent, W. J., Dermitzakis, E. T., Armengol, L., Bird, C. P., Clark, T. G., Cooper, G. M., Bakker, P. I. W. d., Kern, A. D., Lopez-Bigas, N., Martin, J. D., Stranger, B. E., Thomas, D. J., Woodroffe, A., Batzoglou, S., Davydov, E., Dimas, A., Eyras, E., Hallgrímsdóttir, I. B., Hardison, R. C., Huppert, J., Sidow, A., Taylor, J., Trumbower, H., Zody, M. C., Guigó, R., Mullikin, J. C., Abecasis, G. R., Estivill, X., Birney, E., Bouffard, G. G., Guan, X., Hansen, N. F., Idol, J. R., Maduro, V. V. B., Maskeri, B., McDowell, J. C., Park, M., Thomas, P. J., Young, A. C., Blakesley, R. W., Muzny, D. M., Sodergren, E., Wheeler, D. A., Worley, K. C., Jiang, H., Weinstock, G. M., Gibbs, R. A., Graves, T., Fulton, R., Mardis, E. R., Wilson, R. K., Clamp, M., Cuff, J., Gnerre, S., Jaffe, D. B., Chang, J. L., Lindblad-Toh, K., Lander, E. S., Koriabine, M., Nefedov, M., Osoegawa, K., Yoshinaga, Y., Zhu, B., and Jong, P. J. d. (2007). Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature*, 447(7146):799–816.

Bolognesi, B. and Tartaglia, G. G. (2013). Physicochemical principles of protein aggregation. *Progress in molecular biology and translational science*, 117:53–72. PMID: 23663965.

Bompfünewerer, A. F., Backofen, R., Bernhart, S. H., Hertel, J., Hofacker, I. L., Stadler, P. F., and Will, S. (2008). Variations on RNA folding and alignment: lessons from benasque. *Journal of mathematical biology*, 56(1-2):129–144. PMID: 17611759.

Brawand, D., Soumillon, M., Necsulea, A., Julien, P., Csárdi, G., Harrigan, P., Weier, M., Liechti, A., Aximu-Petri, A., Kircher, M., Albert, F. W., Zeller, U., Khaitovich, P., Grützner, F., Bergmann, S., Nielsen, R., Pääbo, S., and Kaessmann, H. (2011). The evolution of gene expression levels in mammalian organs. *Nature*, 478(7369):343–348. PMID: 22012392.

Brown, C. J., Hendrich, B. D., Rupert, J. L., Lafrenière, R. G., Xing, Y., Lawrence, J., and Willard, H. F. (1992). The human XIST gene: analysis of a 17 kb inactive x-specific RNA that contains conserved repeats and is highly localized within the nucleus. *Cell*, 71(3):527–542. PMID: 1423611.

Buchan, J. R. and Parker, R. (2009). Eukaryotic stress granules: the ins and outs of translation. *Molecular cell*, 36(6):932–941. PMID: 20064460 PMCID: PMC2813218.

Bussotti, G., Notredame, C., and Enright, A. J. (2013). Detecting and comparing non-coding RNAs in the high-throughput era. *International Journal of Molecular Sciences*, 14(8):15423–15458. PMID: 23887659 PMCID: PMC3759867.

Cabili, M. N., Trapnell, C., Goff, L., Koziol, M., Tazon-Vega, B., Regev, A., and Rinn, J. L. (2011). Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. *Genes & development*, 25(18):1915–1927. PMID: 21890647 PMCID: PMC3185964.

Calloni, G., Zoffoli, S., Stefani, M., Dobson, C. M., and Chiti, F. (2005). Investigating the effects of mutations on protein aggregation in the cell. *The Journal of biological chemistry*, 280(11):10607–10613. PMID: 15611128.

Carninci, P., Kasukawa, T., Katayama, S., Gough, J., Frith, M. C., Maeda, N., Oyama, R., Ravasi, T., Lenhard, B., Wells, C., Kodzius, R., Shimokawa, K., Bajic, V. B., Brenner, S. E., Batalov, S., Forrest, A. R. R., Zavolan, M., Davis, M. J., Wilming, L. G., Aidinis, V., Allen, J. E., Ambesi-Impiombato, A., Apweiler, R., Aturaliya, R. N., Bailey, T. L., Bansal, M., Baxter, L., Beisel, K. W., Bersano, T., Bono, H., Chalk, A. M., Chiu, K. P., Choudhary, V., Christoffels, A., Clutterbuck, D. R., Crowe, M. L., Dalla, E., Dalrymple, B. P., de Bono, B., Della Gatta, G., di Bernardo, D., Down, T., Engstrom, P., Fagiolini, M., Faulkner, G., Fletcher, C. F., Fukushima, T., Furuno, M., Futaki, S., Gariboldi, M., Georgii-Hemming, P., Gingeras, T. R., Gojobori, T., Green, R. E., Gustincich, S., Harbers, M., Hayashi, Y., Hensch, T. K., Hirokawa, N., Hill, D., Huminiecki, L., Iacono, M., Ikeo, K., Iwama, A., Ishikawa, T., Jakt, M., Kanapin, A., Katoh, M., Kawasawa, Y., Kelso, J., Kitamura, H., Kitano, H., Kollias, G., Krishnan, S. P. T., Kruger, A., Kummerfeld, S. K., Kurochkin, I. V., Lareau, L. F., Lazarevic, D., Lipovich, L., Liu, J., Liuni, S., McWilliam, S., Madan Babu, M., Madera, M., Marchionni, L., Matsuda, H., Matsuzawa, S., Miki, H., Mignone, F., Miyake, S., Morris, K., Mottagui-Tabar, S., Mulder, N., Nakano, N., Nakauchi, H., Ng, P., Nilsson, R., Nishiguchi, S., Nishikawa, S., Nori, F., Ohara, O., Okazaki, Y., Orlando, V., Pang, K. C., Pavan, W. J., Pavesi, G., Pesole, G., Petrovsky, N., Piazza, S., Reed, J., Reid, J. F., Ring, B. Z., Ringwald, M., Rost, B., Ruan, Y., Salzberg, S. L., Sandelin, A., Schneider, C., Schönbach, C., Sekiguchi, K., Semple, C. A. M., Seno, S., Sessa, L., Sheng, Y., Shibata, Y., Shimada, H., Shimada, K., Silva, D., Sinclair, B., Sperling, S., Stupka, E., Sugiura, K., Sultana, R., Takenaka, Y., Taki, K., Tammoja, K., Tan, S. L., Tang, S., Taylor, M. S., Tegner, J., Teichmann, S. A., Ueda, H. R., van Nimwegen, E., Verardo, R., Wei, C. L., Yagi, K., Yamanishi, H., Zabarovsky, E., Zhu, S., Zimmer, A., Hide, W., Bult, C., Grimmond, S. M., Teasdale, R. D., Liu, E. T., Brusic, V., Quackenbush, J., Wahlestedt, C., Mattick, J. S., Hume, D. A., Kai, C., Sasaki, D., Tomaru, Y., Fukuda, S., Kanamori-Katayama, M., Suzuki, M., Aoki, J., Arakawa, T., Iida, J., Imamura, K., Itoh, M., Kato, T., Kawaji, H., Kawagashira, N., Kawashima, T., Kojima,

M., Kondo, S., Konno, H., Nakano, K., Ninomiya, N., Nishio, T., Okada, M., Plessy, C., Shibata, K., Shiraki, T., Suzuki, S., Tagami, M., Waki, K., Watahiki, A., Okamura-Oho, Y., Suzuki, H., Kawai, J., Hayashizaki, Y., FANTOM Consortium, and RIKEN Genome Exploration Research Group and Genome Science Group (Genome Network Project Core Group) (2005). The transcriptional landscape of the mammalian genome. *Science (New York, N.Y.)*, 309(5740):1559–1563. PMID: 16141072.

Castel, S. E. and Martienssen, R. A. (2013). RNA interference in the nucleus: roles for small RNAs in transcription, epigenetics and beyond. *Nature reviews. Genetics*, 14(2):100–112. PMID: 23329111.

Castello, A., Fischer, B., Eichelbaum, K., Horos, R., Beckmann, B. M., Strein, C., Davey, N. E., Humphreys, D. T., Preiss, T., Steinmetz, L. M., Krijgsveld, J., and Hentze, M. W. (2012). Insights into RNA biology from an atlas of mammalian mRNA-binding proteins. *Cell*, 149(6):1393–1406. PMID: 22658674.

Castello, A., Fischer, B., Hentze, M. W., and Preiss, T. (2013a). RNA-binding proteins in mendelian disease. *Trends in genetics: TIG*, 29(5):318–327. PMID: 23415593.

Castello, A., Horos, R., Strein, C., Fischer, B., Eichelbaum, K., Steinmetz, L. M., Krijgsveld, J., and Hentze, M. W. (2013b). System-wide identification of RNA-binding proteins by interactome capture. *Nature protocols*, 8(3):491–500. PMID: 23411631.

Caudron-Herger, M. and Rippe, K. (2012). Nuclear architecture by RNA. *Current opinion in genetics & development*, 22(2):179–187. PMID: 22281031.

Cereda, M., Pozzoli, U., Rot, G., Juvan, P., Schweitzer, A., Clark, T., and Ule, J. (2014). RNAmotifs: prediction of multivalent RNA motifs that control alternative splicing. *Genome Biology*, 15(1):R20. PMID: 24485098.

Chan, E. T., Quon, G. T., Chua, G., Babak, T., Trochesset, M., Zirngibl, R. A., Aubin, J., Ratcliffe, M. J. H., Wilde, A., Brudno, M., Morris, Q. D., and Hughes, T. R. (2009). Conservation of core gene expression in vertebrate tissues. *Journal of biology*, 8(3):33. PMID: 19371447 PMCID: PMC2689434.

Chen, M. and Manley, J. L. (2009). Mechanisms of alternative splicing regulation: insights from molecular and genomics approaches. *Nature reviews. Molecular cell biology*, 10(11):741–754. PMID: 19773805 PMCID: PMC2958924.

Chew, G.-L., Pauli, A., Rinn, J. L., Regev, A., Schier, A. F., and Valen, E. (2013). Ribosome profiling reveals resemblance between long non-coding RNAs and 5' leaders of coding RNAs. *Development (Cambridge, England)*, 140(13):2828–2834. PMID: 23698349 PMCID: PMC3678345.

Chi, S. W., Zang, J. B., Mele, A., and Darnell, R. B. (2009). Argonaute HITS-CLIP decodes microRNA-mRNA interaction maps. *Nature*, 460(7254):479–486. PMID: 19536157 PMCID: PMC2733940.

Chiti, F., Stefani, M., Taddei, N., Ramponi, G., and Dobson, C. M. (2003). Rationalization of the effects of mutations on peptide and protein aggregation rates. *Nature*, 424(6950):805–808. PMID: 12917692.

Chiti, F., Taddei, N., White, P. M., Bucciantini, M., Magherini, F., Stefani, M., and Dobson, C. M. (1999). Mutational analysis of acylphosphatase suggests the importance of topology and contact order in protein folding. *Nature structural biology*, 6(11):1005–1009. PMID: 10542090.

Cirillo, D., Agostini, F., Klus, P., Marchese, D., Rodriguez, S., Bolognesi, B., and Tartaglia, G. G. (2012). Neurodegenerative diseases: Quantitative predictions of protein-RNA interactions. *RNA (New York, N.Y.)*. PMID: 23264567.

Cirillo, D., Agostini, F., and Tartaglia, G. G. (2013). Predictions of protein–RNA interactions. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, 3(2):161–175.

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 biology*, 15(1):R13. PMID: 24401680.

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. *Molecular cell*, 33(6):717–726. PMID: 19217333 PMCID: PMC2696186.

Collier, N. C., Heuser, J., Levy, M. A., and Schlesinger, M. J. (1988). Ultrastructural and biochemical analysis of the stress granule in chicken embryo fibroblasts. *The Journal of cell biology*, 106(4):1131–1139. PMID: 3283146 PMCID: PMC2114993.

Conchillo-Solé, O., de Groot, N. S., Avilés, F. X., Vendrell, J., Daura, X., and Ventura, S. (2007). AGGRESCAN: a server for the prediction and evaluation of "hot spots" of aggregation in polypeptides. *BMC bioinformatics*, 8:65. PMID: 17324296 PMCID: PMC1828741.

Cook, K. B., Kazan, H., Zuberi, K., Morris, Q., and Hughes, T. R. (2011). RBPDB: a database of RNA-binding specificities. *Nucleic acids research*, 39(Database issue):D301–308. PMID: 21036867.

Deléage, G. and Roux, B. (1987). An algorithm for protein secondary structure prediction based on class prediction. *Protein engineering*, 1(4):289–294. PMID: 3508279.

Derrien, T., Johnson, R., Bussotti, G., Tanzer, A., Djebali, S., Tilgner, H., Guernec, G., Martin, D., Merkel, A., Knowles, D. G., Lagarde, J., Veeravalli, L., Ruan, X., Ruan, Y., Lassmann, T., Carninci, P., Brown, J. B., Lipovich, L., Gonzalez, J. M., Thomas, M., Davis, C. A., Shiekhattar, R., Gingeras, T. R., Hubbard, T. J., Notredame, C., Harrow, J., and Guigó, R. (2012). The GENCODE v7 catalog of human long noncoding

RNAs: analysis of their gene structure, evolution, and expression. *Genome research*, 22(9):1775–1789. PMID: 22955988 PMCID: PMC3431493.

Dieterich, C. and Stadler, P. F. (2013). Computational biology of RNA interactions. *Wiley interdisciplinary reviews. RNA*, 4(1):107–120. PMID: 23139167.

Djebali, S., Davis, C. A., Merkel, A., Dobin, A., Lassmann, T., Mortazavi, A., Tanzer, A., Lagarde, J., Lin, W., Schlesinger, F., Xue, C., Marinov, G. K., Khatun, J., Williams, B. A., Zaleski, C., Rozowsky, J., Röder, M., Kokocinski, F., Abdelhamid, R. F., Alioto, T., Antoshechkin, I., Baer, M. T., Bar, N. S., Batut, P., Bell, K., Bell, I., Chakrabortty, S., Chen, X., Chrast, J., Curado, J., Derrien, T., Drenkow, J., Dumais, E., Dumais, J., Duttagupta, R., Falconnet, E., Fastuca, M., Fejes-Toth, K., Ferreira, P., Foissac, S., Fullwood, M. J., Gao, H., Gonzalez, D., Gordon, A., Gunawardena, H., Howald, C., Jha, S., Johnson, R., Kapranov, P., King, B., Kingswood, C., Luo, O. J., Park, E., Persaud, K., Preall, J. B., Ribeca, P., Risk, B., Robyr, D., Sammeth, M., Schaffer, L., See, L.-H., Shahab, A., Skancke, J., Suzuki, A. M., Takahashi, H., Tilgner, H., Trout, D., Walters, N., Wang, H., Wrobel, J., Yu, Y., Ruan, X., Hayashizaki, Y., Harrow, J., Gerstein, M., Hubbard, T., Reymond, A., Antonarakis, S. E., Hannon, G., Giddings, M. C., Ruan, Y., Wold, B., Carninci, P., Guigó, R., and Gingeras, T. R. (2012). Landscape of transcription in human cells. *Nature*, 489(7414):101–108. PMID: 22955620 PMCID: PMC3684276.

Doshi, K. J., Cannone, J. J., Cobaugh, C. W., and Gutell, R. R. (2004). Evaluation of the suitability of free-energy minimization using nearest-neighbor energy parameters for RNA secondary structure prediction. *BMC bioinformatics*, 5:105. PMID: 15296519 PMCID: PMC514602.

Eddy, S. R. (2001). Non-coding RNA genes and the modern RNA world. *Nature reviews. Genetics*, 2(12):919–929. PMID: 11733745.

Ellis, R. J. (2001). Macromolecular crowding: obvious but underappreciated. *Trends in biochemical sciences*, 26(10):597–604. PMID: 11590012.

Engelman, D. M., Steitz, T. A., and Goldman, A. (1986). Identifying nonpolar transbilayer helices in amino acid sequences of membrane proteins. *Annual review of biophysics and biophysical chemistry*, 15:321–353. PMID: 3521657.

Engström, P. G., Suzuki, H., Ninomiya, N., Akalin, A., Sessa, L., Lavorgna, G., Brozzi, A., Luzi, L., Tan, S. L., Yang, L., Kunarso, G., Ng, E. L.-C., Batalov, S., Wahlestedt, C., Kai, C., Kawai, J., Carninci, P., Hayashizaki, Y., Wells, C., Bajic, V. B., Orlando, V., Reid, J. F., Lenhard, B., and Lipovich, L. (2006). Complex loci in human and mouse genomes. *PLoS genetics*, 2(4):e47. PMID: 16683030 PMCID: PMC1449890.

Fernandez, M., Kumagai, Y., Standley, D. M., Sarai, A., Mizuguchi, K., and Ahmad, S. (2011). Prediction of dinucleotide-specific RNA-binding sites in proteins. *BMC bioinformatics*, 12 Suppl 13:S5. PMID: 22373260 PMCID: PMC3278845.

Fernandez-Escamilla, A.-M., Rousseau, F., Schymkowitz, J., and Serrano, L. (2004). Prediction of sequence-dependent and mutational effects on the aggregation of peptides and proteins. *Nature biotechnology*, 22(10):1302–1306. PMID: 15361882.

Furuno, M., Pang, K. C., Ninomiya, N., Fukuda, S., Frith, M. C., Bult, C., Kai, C., Kawai, J., Carninci, P., Hayashizaki, Y., Mattick, J. S., and Suzuki, H. (2006). Clusters of internally primed transcripts reveal novel long noncoding RNAs. *PLoS genetics*, 2(4):e37. PMID: 16683026 PMCID: PMC1449886.

Gal-Mark, N., Schwartz, S., Ram, O., Eyras, E., and Ast, G. (2009). The pivotal roles of TIA proteins in 5' splice-site selection of alu exons and across evolution. *PLoS genetics*, 5(11):e1000717. PMID: 19911040 PMCID: PMC2766253.

Geiler-Samerotte, K. A., Dion, M. F., Budnik, B. A., Wang, S. M., Hartl, D. L., and Drummond, D. A. (2011). Misfolded proteins impose a dosage-dependent fitness cost and trigger a cytosolic unfolded protein response in yeast. *Proceedings of the National Academy of Sciences of the United States of America*, 108(2):680–685. PMID: 21187411 PMCID: PMC3021021.

Gerstberger, S., Hafner, M., and Tuschl, T. (2013). Learning the language of post-transcriptional gene regulation. *Genome biology*, 14(8):130. PMID: 23998708.

Giegerich, R. (2014). Introduction to stochastic context free grammars. *Methods in molecular biology (Clifton, N.J.)*, 1097:85–106. PMID: 24639156.

Glisovic, T., Bachorik, J. L., Yong, J., and Dreyfuss, G. (2008). RNA-binding proteins and post-transcriptional gene regulation. *FEBS letters*, 582(14):1977–1986. PMID: 18342629 PMCID: PMC2858862.

Goehler, H., Dröge, A., Lurz, R., Schnoegl, S., Chernoff, Y. O., and Wanker, E. E. (2010). Pathogenic polyglutamine tracts are potent inducers of spontaneous sup35 and rnq1 amyloidogenesis. *PloS one*, 5(3):e9642. PMID: 20224794 PMCID: PMC2835767.

Goh, C.-S., Lan, N., Douglas, S. M., Wu, B., Echols, N., Smith, A., Milburn, D., Montelione, G. T., Zhao, H., and Gerstein, M. (2004). Mining the structural genomics pipeline: identification of protein properties that affect high-throughput experimental analysis. *Journal of molecular biology*, 336(1):115–130. PMID: 14741208.

Gsponer, J. and Babu, M. M. (2012). Cellular strategies for regulating functional and nonfunctional protein aggregation. *Cell reports*, 2(5):1425–1437. PMID: 23168257 PMCID: PMC3607227.

Gupta, R. A., Shah, N., Wang, K. C., Kim, J., Horlings, H. M., Wong, D. J., Tsai, M.-C., Hung, T., Argani, P., Rinn, J. L., Wang, Y., Brzoska, P., Kong, B., Li, R., West, R. B., van de Vijver, M. J., Sukumar, S., and Chang, H. Y. (2010). Long non-coding RNA HOTAIR reprograms chromatin state to promote cancer metastasis. *Nature*, 464(7291):1071–1076. PMID: 20393566 PMCID: PMC3049919.

Guttman, M., Amit, I., Garber, M., French, C., Lin, M. F., Feldser, D., Huarte, M., Zuk, O., Carey, B. W., Cassady, J. P., Cabili, M. N., Jaenisch, R., Mikkelsen, T. S., Jacks, T., Hacohen, N., Bernstein, B. E., Kellis, M., Regev, A., Rinn, J. L., and Lander, E. S. (2009). Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. *Nature*, 458(7235):223–227. PMID: 19182780 PMCID: PMC2754849.

Guttman, M., Donaghey, J., Carey, B. W., Garber, M., Grenier, J. K., Munson, G., Young, G., Lucas, A. B., Ach, R., Bruhn, L., Yang, X., Amit, I., Meissner, A., Regev, A., Rinn, J. L., Root, D. E., and Lander, E. S. (2011). lincRNAs act in the circuitry controlling pluripotency and differentiation. *Nature*, 477(7364):295–300. PMID: 21874018 PMCID: PMC3175327.

Guttman, M., Garber, M., Levin, J. Z., Donaghey, J., Robinson, J., Adiconis, X., Fan, L., Koziol, M. J., Gnirke, A., Nusbaum, C., Rinn, J. L., Lander, E. S., and Regev, A. (2010). Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs. *Nature biotechnology*, 28(5):503–510. PMID: 20436462 PMCID: PMC2868100.

Guttman, M. and Rinn, J. L. (2012). Modular regulatory principles of large non-coding RNAs. *Nature*, 482(7385):339–346. PMID: 22337053.

Guttman, M., Russell, P., Ingolia, N. T., Weissman, J. S., and Lander, E. S. (2013). Ribosome profiling provides evidence that large noncoding RNAs do not encode proteins. *Cell*, 154(1):240–251.

Hafner, M., Landthaler, M., Burger, L., Khorshid, M., Hausser, J., Berninger, P., Rothballer, A., Ascano, M., Jungkamp, A.-C., Munschauer, M., Ulrich, A., Wardle, G. S., Dewell, S., Zavolan, M., and Tuschl, T. (2010). PAR-CliP–a method to identify transcriptome-wide the binding sites of RNA binding proteins. *Journal of visualized experiments: JoVE*, (41). PMID: 20644507 PMCID: PMC3156069.

Hafner, M., Renwick, N., Brown, M., Mihailović, A., Holoch, D., Lin, C., Pena, J. T. G., Nusbaum, J. D., Morozov, P., Ludwig, J., Ojo, T., Luo, S., Schroth, G., and Tuschl, T. (2011). RNA-ligase-dependent biases in miRNA representation in deep-sequenced small RNA cDNA libraries. *RNA (New York, N.Y.)*, 17(9):1697–1712. PMID: 21775473 PMCID: PMC3162335.

Han, T. W., Kato, M., Xie, S., Wu, L. C., Mirzaei, H., Pei, J., Chen, M., Xie, Y., Allen, J., Xiao, G., and McKnight, S. L. (2012). Cell-free formation of RNA granules: bound RNAs identify features and components of cellular assemblies. *Cell*, 149(4):768–779. PMID: 22579282.

Harrow, J., Frankish, A., Gonzalez, J. M., Tapanari, E., Diekhans, M., Kokocinski, F., Aken, B. L., Barrell, D., Zadissa, A., Searle, S., Barnes, I., Bignell, A., Boychenko, V., Hunt, T., Kay, M., Mukherjee, G., Rajan, J., Despacio-Reyes, G., Saunders, G.,

Steward, C., Harte, R., Lin, M., Howald, C., Tanzer, A., Derrien, T., Chrast, J., Walters, N., Balasubramanian, S., Pei, B., Tress, M., Rodriguez, J. M., Ezkurdia, I., van Baren, J., Brent, M., Haussler, D., Kellis, M., Valencia, A., Reymond, A., Gerstein, M., Guigó, R., and Hubbard, T. J. (2012). GENCODE: the reference human genome annotation for the ENCODE project. *Genome research*, 22(9):1760–1774. PMID: 22955987 PMCID: PMC3431492.

Hawkins, R. D., Hon, G. C., and Ren, B. (2010). Next-generation genomics: an integrative approach. *Nature reviews. Genetics*, 11(7):476–486. PMID: 20531367 PMCID: PMC3321268.

Heimberg, A. M., Sempere, L. F., Moy, V. N., Donoghue, P. C. J., and Peterson, K. J. (2008). MicroRNAs and the advent of vertebrate morphological complexity. *Proceedings of the National Academy of Sciences of the United States of America*, 105(8):2946–2950. PMID: 18287013 PMCID: PMC2268565.

Hertz, G. Z. and Stormo, G. D. (1999). Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. *Bioinformatics (Oxford, England)*, 15(7-8):563–577. PMID: 10487864.

Hiller, M., Pudimat, R., Busch, A., and Backofen, R. (2006). Using RNA secondary structures to guide sequence motif finding towards single-stranded regions. *Nucleic acids research*, 34(17):e117. PMID: 16987907 PMCID: PMC1903381.

Hoeffer, C. A. and Klann, E. (2010). mTOR signaling: at the crossroads of plasticity, memory and disease. *Trends in neurosciences*, 33(2):67–75. PMID: 19963289 PMCID: PMC2821969.

Hofacker, I. L. (2003). Vienna RNA secondary structure server. *Nucleic Acids Research*, 31(13):3429–3431.

Hofacker, I. L. (2014). Energy-directed RNA structure prediction. *Methods in molecular biology (Clifton, N.J.)*, 1097:71–84. PMID: 24639155.

Hogan, D. J., Riordan, D. P., Gerber, A. P., Herschlag, D., and Brown, P. O. (2008). Diverse RNA-Binding proteins interact with functionally related sets of RNAs, suggesting an extensive regulatory system. *PLoS Biol*, 6(10):e255.

Hopp, T. P. and Woods, K. R. (1981). Prediction of protein antigenic determinants from amino acid sequences. *Proceedings of the National Academy of Sciences of the United States of America*, 78(6):3824–3828. PMID: 6167991 PMCID: PMC319665.

Huarte, M., Guttman, M., Feldser, D., Garber, M., Koziol, M. J., Kenzelmann-Broz, D., Khalil, A. M., Zuk, O., Amit, I., Rabani, M., Attardi, L. D., Regev, A., Lander, E. S., Jacks, T., and Rinn, J. L. (2010). A large intergenic noncoding RNA induced by p53 mediates global gene repression in the p53 response. *Cell*, 142(3):409–419. PMID: 20673990 PMCID: PMC2956184.

Hüttenhofer, A. and Schattner, P. (2006). The principles of guiding by RNA: chimeric RNA-protein enzymes. *Nature reviews. Genetics*, 7(6):475–482. PMID: 16622413.

Ishii, N., Ozaki, K., Sato, H., Mizuno, H., Saito, S., Takahashi, A., Miyamoto, Y., Ikegawa, S., Kamatani, N., Hori, M., Saito, S., Nakamura, Y., and Tanaka, T. (2006). Identification of a novel non-coding RNA, MIAT, that confers risk of myocardial infarction. *Journal of human genetics*, 51(12):1087–1099. PMID: 17066261.

Janga, S. C. (2012). From specific to global analysis of posttranscriptional regulation in eukaryotes: posttranscriptional regulatory networks. *Briefings in functional genomics*, 11(6):505–521. PMID: 23124862.

Jeon, Y. and Lee, J. T. (2011). YY1 tethers xist RNA to the inactive x nucleation center. *Cell*, 146(1):119–133. PMID: 21729784 PMCID: PMC3150513.

Johnson, B. S., Snead, D., Lee, J. J., McCaffery, J. M., Shorter, J., and Gitler, A. D. (2009). TDP-43 is intrinsically aggregation-prone, and amyotrophic lateral sclerosis-linked mutations accelerate aggregation and increase toxicity. *The Journal of biological chemistry*, 284(30):20329–20339. PMID: 19465477 PMCID: PMC2740458.

Johnsson, P., Lipovich, L., Grandér, D., and Morris, K. V. (2014). Evolutionary conservation of long non-coding RNAs; sequence, structure, function. *Biochimica et biophysica acta*, 1840(3):1063–1071. PMID: 24184936 PMCID: PMC3909678.

Kageyama, Y., Kondo, T., and Hashimoto, Y. (2011). Coding vs non-coding: Translatability of short ORFs found in putative non-coding transcripts. *Biochimie*, 93(11):1981–1986. PMID: 21729735.

Kapranov, P., Cheng, J., Dike, S., Nix, D. A., Duttagupta, R., Willingham, A. T., Stadler, P. F., Hertel, J., Hackermüller, J., Hofacker, I. L., Bell, I., Cheung, E., Drenkow, J., Dumais, E., Patel, S., Helt, G., Ganesh, M., Ghosh, S., Piccolboni, A., Sementchenko, V., Tammana, H., and Gingeras, T. R. (2007). RNA maps reveal new RNA classes and a possible function for pervasive transcription. *Science (New York, N.Y.)*, 316(5830):1484–1488. PMID: 17510325.

Kato, M., Han, T. W., Xie, S., Shi, K., Du, X., Wu, L. C., Mirzaei, H., Goldsmith, E. J., Longgood, J., Pei, J., Grishin, N. V., Frantz, D. E., Schneider, J. W., Chen, S., Li, L., Sawaya, M. R., Eisenberg, D., Tycko, R., and McKnight, S. L. (2012). Cell-free formation of RNA granules: low complexity sequence domains form dynamic fibers within hydrogels. *Cell*, 149(4):753–767. PMID: 22579281.

Kazan, H., Ray, D., Chan, E. T., Hughes, T. R., and Morris, Q. (2010). RNAcontext: a new method for learning the sequence and structure binding preferences of RNA-Binding proteins. *PLoS Comput Biol*, 6(7):e1000832.

Ke, A. and Doudna, J. A. (2004). Crystallization of RNA and RNA-protein complexes. *Methods (San Diego, Calif.)*, 34(3):408–414. PMID: 15325657.

Kechavarzi, B. and Janga, S. C. (2014). Dissecting the expression landscape of RNA-binding proteins in human cancers. *Genome Biology*, 15(1):R14. PMID: 24410894.

Kedersha, N. and Anderson, P. (2007). Mammalian stress granules and processing bodies. *Methods in enzymology*, 431:61–81. PMID: 17923231.

Keene, J. D. (2007). RNA regulons: coordination of post-transcriptional events. *Nature reviews. Genetics*, 8(7):533–543. PMID: 17572691.

Keene, J. D., Komisarow, J. M., and Friedersdorf, M. B. (2006). RIP-Chip: the isolation and identification of mRNAs, microRNAs and protein components of ribonucleoprotein complexes from cell extracts. *Nature protocols*, 1(1):302–307. PMID: 17406249.

Keren, H., Lev-Maor, G., and Ast, G. (2010). Alternative splicing and evolution: diversification, exon definition and function. *Nature reviews. Genetics*, 11(5):345–355. PMID: 20376054.

Kertesz, M., Wan, Y., Mazor, E., Rinn, J. L., Nutter, R. C., Chang, H. Y., and Segal, E. (2010). Genome-wide measurement of RNA secondary structure in yeast. *Nature*, 467(7311):103–107. PMID: 20811459 PMCID: PMC3847670.

Khalil, A. M., Guttman, M., Huarte, M., Garber, M., Raj, A., Rivea Morales, D., Thomas, K., Presser, A., Bernstein, B. E., van Oudenaarden, A., Regev, A., Lander, E. S., and Rinn, J. L. (2009). Many human large intergenic noncoding RNAs associate with chromatin-modifying complexes and affect gene expression. *Proceedings of the National Academy of Sciences of the United States of America*, 106(28):11667–11672. PMID: 19571010 PMCID: PMC2704857.

Khalil, A. M. and Rinn, J. L. (2011). RNA-protein interactions in human health and disease. *Seminars in cell & developmental biology*, 22(4):359–365. PMID: 21333748.

Kim, M. Y., Hur, J., and Jeong, S. (2009). Emerging roles of RNA and RNA-binding protein network in cancer cells. *BMB reports*, 42(3):125–130. PMID: 19335997.

Kim, T.-K., Hemberg, M., Gray, J. M., Costa, A. M., Bear, D. M., Wu, J., Harmin, D. A., Laptewicz, M., Barbara-Haley, K., Kuersten, S., Markenscoff-Papadimitriou, E., Kuhl, D., Bito, H., Worley, P. F., Kreiman, G., and Greenberg, M. E. (2010). Widespread transcription at neuronal activity-regulated enhancers. *Nature*, 465(7295):182–187. PMID: 20393465 PMCID: PMC3020079.

Kishore, S., Jaskiewicz, L., Burger, L., Hausser, J., Khorshid, M., and Zavolan, M. (2011). A quantitative analysis of CLIP methods for identifying binding sites of RNA-binding proteins. *Nature methods*, 8(7):559–564. PMID: 21572407.

Klug, S. J. and Famulok, M. (1994). All you wanted to know about SELEX. *Molecular biology reports*, 20(2):97–107. PMID: 7536299.

Klus, P., Bolognesi, B., Agostini, F., Marchese, D., Zanzoni, A., and Tartaglia, G. G. (2014). The cleverSuite approach for protein characterization: predictions of struc-

tural properties, solubility, chaperone requirements and RNA-binding abilities. *Bioinformatics (Oxford, England)*. PMID: 24493033.

Krichevsky, A. M. and Kosik, K. S. (2001). Neuronal RNA granules: a link between RNA localization and stimulation-dependent translation. *Neuron*, 32(4):683–696. PMID: 11719208.

Kwon, S. C., Yi, H., Eichelbaum, K., Föhr, S., Fischer, B., You, K. T., Castello, A., Krijgsveld, J., Hentze, M. W., and Kim, V. N. (2013). The RNA-binding protein repertoire of embryonic stem cells. *Nature structural & molecular biology*, 20(9):1122–1130. PMID: 23912277.

König, J., Zarnack, K., Luscombe, N. M., and Ule, J. (2012). Protein–RNA interactions: new genomic technologies and perspectives. *Nature Reviews Genetics*, 13(2):77–83.

König, J., Zarnack, K., Rot, G., Curk, T., Kayikci, M., Zupan, B., Turner, D. J., Luscombe, N. M., and Ule, J. (2010). iCLIP reveals the function of hnRNP particles in splicing at individual nucleotide resolution. *Nature structural & molecular biology*, 17(7):909–915. PMID: 20601959 PMCID: PMC3000544.

Lai, D., Proctor, J. R., and Meyer, I. M. (2013). On the importance of cotranscriptional RNA structure formation. *RNA*, 19(11):1461–1473. PMID: 24131802 PMCID: PMC3851714.

Levine, M. and Tjian, R. (2003). Transcription regulation and animal diversity. *Nature*, 424(6945):147–151. PMID: 12853946.

Levitt, M. (1978). Conformational preferences of amino acids in globular proteins. *Biochemistry*, 17(20):4277–4285. PMID: 708713.

Lewis, B. A., Walia, R. R., Terribilini, M., Ferguson, J., Zheng, C., Honavar, V., and Dobbs, D. (2011). PRIDB: a protein-RNA interface database. *Nucleic acids research*, 39(Database issue):D277–282. PMID: 21071426 PMCID: PMC3013700.

Li, P., Banjade, S., Cheng, H.-C., Kim, S., Chen, B., Guo, L., Llaguno, M., Hollingsworth, J. V., King, D. S., Banani, S. F., Russo, P. S., Jiang, Q.-X., Nixon, B. T., and Rosen, M. K. (2012). Phase transitions in the assembly of multivalent signalling proteins. *Nature*, 483(7389):336–340. PMID: 22398450 PMCID: PMC3343696.

Li, X., Kazan, H., Lipshitz, H. D., and Morris, Q. D. (2014). Finding the target sites of RNA-binding proteins. *Wiley interdisciplinary reviews. RNA*, 5(1):111–130. PMID: 24217996.

Li, X., Quon, G., Lipshitz, H. D., and Morris, Q. (2010). Predicting in vivo binding sites of RNA-binding proteins using mRNA secondary structure. *RNA (New York, N.Y.)*, 16(6):1096–1107. PMID: 20418358 PMCID: PMC2874161.

Licatalosi, D. D. and Darnell, R. B. (2010). RNA processing and its regulation: global insights into biological networks. *Nature reviews. Genetics*, 11(1):75–87. PMID: 20019688 PMCID: PMC3229837.

Licatalosi, D. D., Mele, A., Fak, J. J., Ule, J., Kayikci, M., Chi, S. W., Clark, T. A., Schweitzer, A. C., Blume, J. E., Wang, X., Darnell, J. C., and Darnell, R. B. (2008). HITS-CLIP yields genome-wide insights into brain alternative RNA processing. *Nature*, 456(7221):464–469. PMID: 18978773 PMCID: PMC2597294.

Liu, Z.-P., Wu, L.-Y., Wang, Y., Zhang, X.-S., and Chen, L. (2010). Prediction of protein-RNA binding sites by a random forest method with combined features. *Bioinformatics (Oxford, England)*, 26(13):1616–1622. PMID: 20483814.

Liu-Yesucevitz, L., Bassell, G. J., Gitler, A. D., Hart, A. C., Klann, E., Richter, J. D., Warren, S. T., and Wolozin, B. (2011). Local RNA translation at the synapse and in disease. *The Journal of neuroscience: the official journal of the Society for Neuroscience*, 31(45):16086–16093. PMID: 22072660 PMCID: PMC3241995.

Liu-Yesucevitz, L., Bilgutay, A., Zhang, Y.-J., Vanderweyde, T., Vanderwyde, T., Citro, A., Mehta, T., Zaarur, N., McKee, A., Bowser, R., Sherman, M., Petrucelli, L., and Wolozin, B. (2010). Tar DNA binding protein-43 (TDP-43) associates with stress granules: analysis of cultured cells and pathological brain tissue. *PloS one*, 5(10):e13250. PMID: 20948999 PMCID: PMC2952586.

Lorenz, R., Bernhart, S. H., Siederdissen, C. H. z., Tafer, H., Flamm, C., Stadler, P. F., and Hofacker, I. L. (2011). ViennaRNA package 2.0. *Algorithms for Molecular Biology*, 6(1):26. PMID: 22115189.

Lu, Q., Ren, S., Lu, M., Zhang, Y., Zhu, D., Zhang, X., and Li, T. (2013). Computational prediction of associations between long non-coding RNAs and proteins. *BMC Genomics*, 14(1):651. PMID: 24063787.

Lucks, J. B., Mortimer, S. A., Trapnell, C., Luo, S., Aviran, S., Schroth, G. P., Pachter, L., Doudna, J. A., and Arkin, A. P. (2011). Multiplexed RNA structure characterization with selective 2'-hydroxyl acylation analyzed by primer extension sequencing (SHAPE-Seq). *Proceedings of the National Academy of Sciences of the United States of America*, 108(27):11063–11068. PMID: 21642531 PMCID: PMC3131332.

Luheshi, L. M., Tartaglia, G. G., Brorsson, A.-C., Pawar, A. P., Watson, I. E., Chiti, F., Vendruscolo, M., Lomas, D. A., Dobson, C. M., and Crowther, D. C. (2007). Systematic in vivo analysis of the intrinsic determinants of amyloid beta pathogenicity. *PLoS biology*, 5(11):e290. PMID: 17973577 PMCID: PMC2043051.

Lui, L. and Lowe, T. (2013). Small nucleolar RNAs and RNA-guided post-transcriptional modification. *Essays in biochemistry*, 54:53–77. PMID: 23829527.

Lukong, K. E., Chang, K.-w., Khandjian, E. W., and Richard, S. (2008). RNA-binding proteins in human genetic disease. *Trends in genetics: TIG*, 24(8):416–425. PMID: 18597886.

Lunde, B. M., Moore, C., and Varani, G. (2007). RNA-binding proteins: modular design for efficient function. *Nature Reviews Molecular Cell Biology*, 8(6):479–490.

Lyle, R., Watanabe, D., te Vruchte, D., Lerchner, W., Smrzka, O. W., Wutz, A., Schageman, J., Hahner, L., Davies, C., and Barlow, D. P. (2000). The imprinted antisense RNA at the igf2r locus overlaps but does not imprint mas1. *Nature genetics*, 25(1):19–21. PMID: 10802648.

Maenner, S., Blaud, M., Fouillen, L., Savoye, A., Marchand, V., Dubois, A., Sanglier-Cianférani, S., Van Dorsselaer, A., Clerc, P., Avner, P., Visvikis, A., and Branlant, C. (2010). 2-d structure of the a region of xist RNA and its implication for PRC2 association. *PLoS Biol*, 8(1):e1000276.

Maetschke, S. R. and Yuan, Z. (2009). Exploiting structural and topological information to improve prediction of RNA-protein binding sites. *BMC bioinformatics*, 10:341. PMID: 19835626 PMCID: PMC2774325.

Magnan, C. N., Randall, A., and Baldi, P. (2009). SOLpro: accurate sequence-based prediction of protein solubility. *Bioinformatics (Oxford, England)*, 25(17):2200–2207. PMID: 19549632.

Mao, Y. S., Zhang, B., and Spector, D. L. (2011). Biogenesis and function of nuclear bodies. *Trends in genetics: TIG*, 27(8):295–306. PMID: 21680045 PMCID: PMC3144265.

Marques, A. C. and Ponting, C. P. (2009). Catalogues of mammalian long noncoding RNAs: modest conservation and incompleteness. *Genome biology*, 10(11):R124. PMID: 19895688 PMCID: PMC3091318.

Martin, F. (2012). Fifteen years of the yeast three-hybrid system: RNA-protein interactions under investigation. *Methods (San Diego, Calif.)*, 58(4):367–375. PMID: 22841566.

Masuda, K., Kuwano, Y., Nishida, K., and Rokutan, K. (2012). General RBP expression in human tissues as a function of age. *Ageing research reviews*, 11(4):423–431. PMID: 22326651.

Matera, A. G., Izaguire-Sierra, M., Praveen, K., and Rajendra, T. K. (2009). Nuclear bodies: random aggregates of sticky proteins or crucibles of macromolecular assembly? *Developmental cell*, 17(5):639–647. PMID: 19922869 PMCID: PMC3101021.

Mattick, J. S. (2001). Non-coding RNAs: the architects of eukaryotic complexity. *EMBO reports*, 2(11):986–991. PMID: 11713189 PMCID: PMC1084129.

Meader, S., Ponting, C. P., and Lunter, G. (2010). Massive turnover of functional sequence in human and other mammalian genomes. *Genome research*, 20(10):1335–1343. PMID: 20693480 PMCID: PMC2945182.

Mercer, T. R. and Mattick, J. S. (2013). Structure and function of long noncoding RNAs in epigenetic regulation. *Nature structural & molecular biology*, 20(3):300–307. PMID: 23463315.

Merkin, J., Russell, C., Chen, P., and Burge, C. B. (2012). Evolutionary dynamics of gene and isoform regulation in mammalian tissues. *Science (New York, N.Y.)*, 338(6114):1593–1599. PMID: 23258891 PMCID: PMC3568499.

Meyer, K. D., Saletore, Y., Zumbo, P., Elemento, O., Mason, C. E., and Jaffrey, S. R. (2012). Comprehensive analysis of mRNA methylation reveals enrichment in 3' UTRs and near stop codons. *Cell*, 149(7):1635–1646. PMID: 22608085 PMCID: PMC3383396.

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):206–212. PMID: 22212136.

Mili, S. and Steitz, J. A. (2004). Evidence for reassociation of RNA-binding proteins after cell lysis: implications for the interpretation of immunoprecipitation analyses. *RNA (New York, N.Y.)*, 10(11):1692–1694. PMID: 15388877 PMCID: PMC1370654.

Muppirala, U. K., Honavar, V. G., and Dobbs, D. (2011). Predicting RNA-protein interactions using only sequence information. *BMC bioinformatics*, 12:489. PMID: 22192482.

Murakami, T., Yang, S.-P., Xie, L., Kawano, T., Fu, D., Mukai, A., Bohm, C., Chen, F., Robertson, J., Suzuki, H., Tartaglia, G. G., Vendruscolo, M., Kaminski Schierle, G. S., Chan, F. T. S., Moloney, A., Crowther, D., Kaminski, C. F., Zhen, M., and St George-Hyslop, P. (2012). ALS mutations in FUS cause neuronal dysfunction and death in caenorhabditis elegans by a dominant gain-of-function mechanism. *Human molecular genetics*, 21(1):1–9. PMID: 21949354 PMCID: PMC3235006.

Mus, E., Hof, P. R., and Tiedge, H. (2007). Dendritic BC200 RNA in aging and in alzheimer's disease. *Proceedings of the National Academy of Sciences of the United States of America*, 104(25):10679–10684. PMID: 17553964 PMCID: PMC1965572.

Musunuru, K. (2003). Cell-specific RNA-binding proteins in human disease. *Trends in cardiovascular medicine*, 13(5):188–195. PMID: 12837581.

Müller-McNicoll, M. and Neugebauer, K. M. (2013). How cells get the message: dynamic assembly and function of mRNA-protein complexes. *Nature reviews. Genetics*, 14(4):275–287. PMID: 23478349.

Nagano, T., Mitchell, J. A., Sanz, L. A., Pauler, F. M., Ferguson-Smith, A. C., Feil, R., and Fraser, P. (2008). The air noncoding RNA epigenetically silences transcription by targeting g9a to chromatin. *Science (New York, N.Y.)*, 322(5908):1717–1720. PMID: 18988810.

Narayanaswamy, R., Levy, M., Tsechansky, M., Stovall, G. M., O'Connell, J. D., Mirrielees, J., Ellington, A. D., and Marcotte, E. M. (2009). Widespread reorganization of metabolic enzymes into reversible assemblies upon nutrient starvation. *Proceedings*

of the National Academy of Sciences of the United States of America, 106(25):10147–10152. PMID: 19502427 PMCID: PMC2691686.

Niwa, T., Ying, B.-W., Saito, K., Jin, W., Takada, S., Ueda, T., and Taguchi, H. (2009). Bimodal protein solubility distribution revealed by an aggregation analysis of the entire ensemble of escherichia coli proteins. *Proceedings of the National Academy of Sciences of the United States of America*, 106(11):4201–4206. PMID: 19251648 PMCID: PMC2657415.

Olzscha, H., Schermann, S. M., Woerner, A. C., Pinkert, S., Hecht, M. H., Tartaglia, G. G., Vendruscolo, M., Hayer-Hartl, M., Hartl, F. U., and Vabulas, R. M. (2011). Amyloid-like aggregates sequester numerous metastable proteins with essential cellular functions. *Cell*, 144(1):67–78. PMID: 21215370.

Pancaldi, V. and Bähler, J. (2011). In silico characterization and prediction of global protein-mRNA interactions in yeast. *Nucleic acids research*, 39(14):5826–5836. PMID: 21459850 PMCID: PMC3152324.

Parker, R. and Sheth, U. (2007). P bodies and the control of mRNA translation and degradation. *Molecular cell*, 25(5):635–646. PMID: 17349952.

Penny, G. D., Kay, G. F., Sheardown, S. A., Rastan, S., and Brockdorff, N. (1996). Requirement for xist in x chromosome inactivation. *Nature*, 379(6561):131–137. PMID: 8538762.

Poliseno, L., Salmena, L., Zhang, J., Carver, B., Haveman, W. J., and Pandolfi, P. P. (2010). A coding-independent function of gene and pseudogene mRNAs regulates tumour biology. *Nature*, 465(7301):1033–1038. PMID: 20577206 PMCID: PMC3206313.

Ponjavic, J., Ponting, C. P., and Lunter, G. (2007). Functionality or transcriptional noise? evidence for selection within long noncoding RNAs. *Genome research*, 17(5):556–565. PMID: 17387145 PMCID: PMC1855172.

Ponting, C. P., Oliver, P. L., and Reik, W. (2009). Evolution and functions of long noncoding RNAs. *Cell*, 136(4):629–641. PMID: 19239885.

Powers, E. T., Morimoto, R. I., Dillin, A., Kelly, J. W., and Balch, W. E. (2009). Biological and chemical approaches to diseases of proteostasis deficiency. *Annual review of biochemistry*, 78:959–991. PMID: 19298183.

Prasanth, K. V. and Spector, D. L. (2007). Eukaryotic regulatory RNAs: an answer to the 'genome complexity' conundrum. *Genes & Development*, 21(1):11–42. PMID: 17210785.

Puton, T., Kozlowski, L., Tuszynska, I., Rother, K., and Bujnicki, J. M. (2012). Computational methods for prediction of protein-RNA interactions. *Journal of structural biology*, 179(3):261–268. PMID: 22019768.

Pérez-Cano, L. and Fernández-Recio, J. (2010). Optimal protein-RNA area, OPRA: a propensity-based method to identify RNA-binding sites on proteins. *Proteins*, 78(1):25–35. PMID: 19714772.

Quenneville, S., Turelli, P., Bojkowska, K., Raclot, C., Offner, S., Kapopoulou, A., and Trono, D. (2012). The KRAB-ZFP/KAP1 system contributes to the early embryonic establishment of site-specific DNA methylation patterns maintained during development. *Cell reports*, 2(4):766–773. PMID: 23041315 PMCID: PMC3677399.

Ramaswami, M., Taylor, J. P., and Parker, R. (2013). Altered ribostasis: RNA-Protein granules in degenerative disorders. *Cell*, 154(4):727–736.

Ravasi, T., Suzuki, H., Cannistraci, C. V., Katayama, S., Bajic, V. B., Tan, K., Akalin, A., Schmeier, S., Kanamori-Katayama, M., Bertin, N., Carninci, P., Daub, C. O., Forrest, A. R. R., Gough, J., Grimmond, S., Han, J.-H., Hashimoto, T., Hide, W., Hofmann, O., Kamburov, A., Kaur, M., Kawaji, H., Kubosaki, A., Lassmann, T., van Nimwegen, E., MacPherson, C. R., Ogawa, C., Radovanovic, A., Schwartz, A., Teasdale, R. D., Tegnér, J., Lenhard, B., Teichmann, S. A., Arakawa, T., Ninomiya, N., Murakami, K., Tagami, M., Fukuda, S., Imamura, K., Kai, C., Ishihara, R., Kitazume, Y., Kawai, J., Hume, D. A., Ideker, T., and Hayashizaki, Y. (2010). An atlas of combinatorial transcriptional regulation in mouse and man. *Cell*, 140(5):744–752. PMID: 20211142 PMCID: PMC2836267.

Ravasi, T., Suzuki, H., Pang, K. C., Katayama, S., Furuno, M., Okunishi, R., Fukuda, S., Ru, K., Frith, M. C., Gongora, M. M., Grimmond, S. M., Hume, D. A., Hayashizaki, Y., and Mattick, J. S. (2006). Experimental validation of the regulated expression of large numbers of non-coding RNAs from the mouse genome. *Genome research*, 16(1):11–19. PMID: 16344565 PMCID: PMC1356124.

Ray, D., Kazan, H., Cook, K. B., Weirauch, M. T., Najafabadi, H. S., Li, X., Gueroussov, S., Albu, M., Zheng, H., Yang, A., Na, H., Irimia, M., Matzat, L. H., Dale, R. K., Smith, S. A., Yarosh, C. A., Kelly, S. M., Nabet, B., Mecenas, D., Li, W., Laishram, R. S., Qiao, M., Lipshitz, H. D., Piano, F., Corbett, A. H., Carstens, R. P., Frey, B. J., Anderson, R. A., Lynch, K. W., Penalva, L. O. F., Lei, E. P., Fraser, A. G., Blencowe, B. J., Morris, Q. D., and Hughes, T. R. (2013). A compendium of RNA-binding motifs for decoding gene regulation. *Nature*, 499(7457):172–177.

Riley, K. J. and Steitz, J. A. (2013). The "Observer effect" in genome-wide surveys of protein-RNA interactions. *Molecular cell*, 49(4):601–604. PMID: 23438856 PMCID: PMC3719848.

Riley, K. J., Yario, T. A., and Steitz, J. A. (2012). Association of argonaute proteins and microRNAs can occur after cell lysis. *RNA (New York, N.Y.)*, 18(9):1581–1585. PMID: 22836356 PMCID: PMC3425773.

Rinn, J. L. and Chang, H. Y. (2012). Genome regulation by long noncoding RNAs. *Annual review of biochemistry*, 81:145–166. PMID: 22663078 PMCID: PMC3858397.

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., and Chang, H. Y. (2007). Functional demarcation of active and silent chromatin domains in human HOX loci by noncoding RNAs. *Cell*, 129(7):1311–1323. PMID: 17604720 PMCID: PMC2084369.

Ryan, C. J., Cimermančič, P., Szpiech, Z. A., Sali, A., Hernandez, R. D., and Krogan, N. J. (2013). High-resolution network biology: connecting sequence with function. *Nature reviews. Genetics*, 14(12):865–879. PMID: 24197012.

Sabin, L. R., Delás, M. J., and Hannon, G. J. (2013). Dogma derailed: the many influences of RNA on the genome. *Molecular cell*, 49(5):783–794. PMID: 23473599 PMCID: PMC3825098.

Schaeffer, C., Bardoni, B., Mandel, J. L., Ehresmann, B., Ehresmann, C., and Moine, H. (2001). The fragile x mental retardation protein binds specifically to its mRNA via a purine quartet motif. *The EMBO journal*, 20(17):4803–4813. PMID: 11532944 PMCID: PMC125594.

Scott, L. G. and Hennig, M. (2008). RNA structure determination by NMR. *Methods in molecular biology (Clifton, N.J.)*, 452:29–61. PMID: 18563368.

Seetin, M. G. and Mathews, D. H. (2012). RNA structure prediction: an overview of methods. *Methods in molecular biology (Clifton, N.J.)*, 905:99–122. PMID: 22736001.

Serio, T. R. and Lindquist, S. L. (2001). [PSI+], SUP35, and chaperones. *Advances in protein chemistry*, 57:335–366. PMID: 11447696.

Shazman, S. and Mandel-Gutfreund, Y. (2008). Classifying RNA-binding proteins based on electrostatic properties. *PLoS computational biology*, 4(8):e1000146. PMID: 18716674 PMCID: PMC2518515.

Shevtsov, S. P. and Dundr, M. (2011). Nucleation of nuclear bodies by RNA. *Nature cell biology*, 13(2):167–173. PMID: 21240286.

Siebert, S. and Backofen, R. (2005). MARNA: multiple alignment and consensus structure prediction of RNAs based on sequence structure comparisons. *Bioinformatics (Oxford, England)*, 21(16):3352–3359. PMID: 15972285.

Smialowski, P., Doose, G., Torkler, P., Kaufmann, S., and Frishman, D. (2012). PROSO II–a new method for protein solubility prediction. *The FEBS journal*, 279(12):2192–2200. PMID: 22536855.

Smit, S., Knight, R., and Heringa, J. (2009). RNA structure prediction from evolutionary patterns of nucleotide composition. *Nucleic acids research*, 37(5):1378–1386. PMID: 19129237 PMCID: PMC2655677.

Smith, M. A., Gesell, T., Stadler, P. F., and Mattick, J. S. (2013). Widespread purifying

selection on RNA structure in mammals. *Nucleic Acids Research*, 41(17):8220–8236. PMID: 23847102.

Sonnhammer, E. L., Eddy, S. R., and Durbin, R. (1997). Pfam: a comprehensive database of protein domain families based on seed alignments. *Proteins*, 28(3):405–420. PMID: 9223186.

Spitale, R. C., Tsai, M.-C., and Chang, H. Y. (2011). RNA templating the epigenome: long noncoding RNAs as molecular scaffolds. *Epigenetics: official journal of the DNA Methylation Society*, 6(5):539–543. PMID: 21393997 PMCID: PMC3230545.

Stawiski, E. W., Gregoret, L. M., and Mandel-Gutfreund, Y. (2003). Annotating nucleic acid-binding function based on protein structure. *Journal of molecular biology*, 326(4):1065–1079. PMID: 12589754.

Stephen, S., Pheasant, M., Makunin, I. V., and Mattick, J. S. (2008). Large-scale appearance of ultraconserved elements in tetrapod genomes and slowdown of the molecular clock. *Molecular biology and evolution*, 25(2):402–408. PMID: 18056681.

Taft, R. J., Pheasant, M., and Mattick, J. S. (2007). The relationship between non-protein-coding DNA and eukaryotic complexity. *BioEssays: news and reviews in molecular, cellular and developmental biology*, 29(3):288–299. PMID: 17295292.

Tartaglia, G. G. and Caflisch, A. (2007). Computational analysis of the s. cerevisiae proteome reveals the function and cellular localization of the least and most amyloidogenic proteins. *Proteins*, 68(1):273–278. PMID: 17407164.

Tartaglia, G. G., Cavalli, A., Pellarin, R., and Caflisch, A. (2004). The role of aromaticity, exposed surface, and dipole moment in determining protein aggregation rates. *Protein science: a publication of the Protein Society*, 13(7):1939–1941. PMID: 15169952 PMCID: PMC2279921.

Tartaglia, G. G., Cavalli, A., Pellarin, R., and Caflisch, A. (2005). Prediction of aggregation rate and aggregation-prone segments in polypeptide sequences. *Protein science: a publication of the Protein Society*, 14(10):2723–2734. PMID: 16195556 PMCID: PMC2253302.

Tartaglia, G. G., Dobson, C. M., Hartl, F. U., and Vendruscolo, M. (2010). Physicochemical determinants of chaperone requirements. *Journal of molecular biology*, 400(3):579–588. PMID: 20416322.

Tartaglia, G. G., Pechmann, S., Dobson, C. M., and Vendruscolo, M. (2007). Life on the edge: a link between gene expression levels and aggregation rates of human proteins. *Trends in biochemical sciences*, 32(5):204–206. PMID: 17419062.

Tartaglia, G. G., Pechmann, S., Dobson, C. M., and Vendruscolo, M. (2009). A relationship between mRNA expression levels and protein solubility in e. coli. *Journal of molecular biology*, 388(2):381–389. PMID: 19281824.

Tartaglia, G. G. and Vendruscolo, M. (2008). The zyggregator method for predicting protein aggregation propensities. *Chemical Society reviews*, 37(7):1395–1401. PMID: 18568165.

Tartaglia, G. G. and Vendruscolo, M. (2009). Correlation between mRNA expression levels and protein aggregation propensities in subcellular localisations. *Molecular bioSystems*, 5(12):1873–1876. PMID: 19763336.

Tartaglia, G. G. and Vendruscolo, M. (2010). Proteome-level interplay between folding and aggregation propensities of proteins. *Journal of molecular biology*, 402(5):919–928. PMID: 20709078.

Terribilini, M., Sander, J. D., Lee, J.-H., Zaback, P., Jernigan, R. L., Honavar, V., and Dobbs, D. (2007). RNABindR: a server for analyzing and predicting RNA-binding sites in proteins. *Nucleic acids research*, 35(Web Server issue):W578–584. PMID: 17483510 PMCID: PMC1933119.

Thomas, M. G., Loschi, M., Desbats, M. A., and Boccaccio, G. L. (2011). RNA granules: the good, the bad and the ugly. *Cellular signalling*, 23(2):324–334. PMID: 20813183 PMCID: PMC3001194.

Thomas-Chollier, M., Herrmann, C., Defrance, M., Sand, O., Thieffry, D., and Helden, J. v. (2011). RSAT peak-motifs: motif analysis in full-size ChIP-seq datasets. *Nucleic Acids Research*, page gkr1104. PMID: 22156162.

Tompa, M., Li, N., Bailey, T. L., Church, G. M., De Moor, B., Eskin, E., Favorov, A. V., Frith, M. C., Fu, Y., Kent, W. J., Makeev, V. J., Mironov, A. A., Noble, W. S., Pavesi, G., Pesole, G., Régnier, M., Simonis, N., Sinha, S., Thijs, G., van Helden, J., Vandenbogaert, M., Weng, Z., Workman, C., Ye, C., and Zhu, Z. (2005). Assessing computational tools for the discovery of transcription factor binding sites. *Nature biotechnology*, 23(1):137–144. PMID: 15637633.

Towfic, F., Caragea, C., Gemperline, D. C., Dobbs, D., and Honavar, V. (2010). StructNB: predicting protein-RNA binding sites using structural features. *International journal of data mining and bioinformatics*, 4(1):21–43. PMID: 20300450 PMCID: PMC2840657.

Tsolis, A. C., Papandreou, N. C., Iconomidou, V. A., and Hamodrakas, S. J. (2013). A consensus method for the prediction of 'aggregation-prone' peptides in globular proteins. *PloS one*, 8(1):e54175. PMID: 23326595 PMCID: PMC3542318.

Uhlen, M., Oksvold, P., Fagerberg, L., Lundberg, E., Jonasson, K., Forsberg, M., Zwahlen, M., Kampf, C., Wester, K., Hober, S., Wernerus, H., Björling, L., and Ponten, F. (2010). Towards a knowledge-based human protein atlas. *Nature biotechnology*, 28(12):1248–1250. PMID: 21139605.

Ule, J., Jensen, K. B., Ruggiu, M., Mele, A., Ule, A., and Darnell, R. B. (2003).

CLIP identifies nova-regulated RNA networks in the brain. *Science (New York, N.Y.)*, 302(5648):1212–1215. PMID: 14615540.

Ulitsky, I. and Bartel, D. P. (2013). lincRNAs: genomics, evolution, and mechanisms. *Cell*, 154(1):26–46. PMID: 23827673.

Vavouri, T., Semple, J. I., Garcia-Verdugo, R., and Lehner, B. (2009). Intrinsic protein disorder and interaction promiscuity are widely associated with dosage sensitivity. *Cell*, 138(1):198–208. PMID: 19596244.

Vendruscolo, M., Knowles, T. P. J., and Dobson, C. M. (2011). Protein solubility and protein homeostasis: a generic view of protein misfolding disorders. *Cold Spring Harbor perspectives in biology*, 3(12). PMID: 21825020 PMCID: PMC3225949.

Vendruscolo, M. and Tartaglia, G. G. (2008). Towards quantitative predictions in cell biology using chemical properties of proteins. *Molecular bioSystems*, 4(12):1170–1175. PMID: 19396379.

Ventura, S. (2005). Sequence determinants of protein aggregation: tools to increase protein solubility. *Microbial cell factories*, 4(1):11. PMID: 15847694 PMCID: PMC1087874.

Ventura, S., Zurdo, J., Narayanan, S., Parreño, M., Mangues, R., Reif, B., Chiti, F., Giannoni, E., Dobson, C. M., Aviles, F. X., and Serrano, L. (2004). Short amino acid stretches can mediate amyloid formation in globular proteins: the src homology 3 (SH3) case. *Proceedings of the National Academy of Sciences of the United States of America*, 101(19):7258–7263. PMID: 15123800 PMCID: PMC409906.

Walia, R. R., Caragea, C., Lewis, B. A., Towfic, F., Terribilini, M., El-Manzalawy, Y., Dobbs, D., and Honavar, V. (2012). Protein-RNA interface residue prediction using machine learning: an assessment of the state of the art. *BMC bioinformatics*, 13:89. PMID: 22574904 PMCID: PMC3490755.

Wan, Y., Kertesz, M., Spitale, R. C., Segal, E., and Chang, H. Y. (2011). Understanding the transcriptome through RNA structure. *Nature Reviews Genetics*, 12(9):641–655.

Wan, Y., Qu, K., Ouyang, Z., Kertesz, M., Li, J., Tibshirani, R., Makino, D. L., Nutter, R. C., Segal, E., and Chang, H. Y. (2012). Genome-wide measurement of RNA folding energies. *Molecular cell*, 48(2):169–181. PMID: 22981864 PMCID: PMC3483374.

Wang, K. C. and Chang, H. Y. (2011). Molecular mechanisms of long noncoding RNAs. *Molecular cell*, 43(6):904–914. PMID: 21925379 PMCID: PMC3199020.

Wang, Y., Chen, X., Liu, Z.-P., Huang, Q., Wang, Y., Xu, D., Zhang, X.-S., Chen, R., and Chen, L. (2013). De novo prediction of RNA-protein interactions from sequence information. *Molecular bioSystems*, 9(1):133–142. PMID: 23138266.

Wapinski, O. and Chang, H. Y. (2011). Long noncoding RNAs and human disease. *Trends in cell biology*, 21(6):354–361. PMID: 21550244.

Weber, S. C. and Brangwynne, C. P. (2012). Getting RNA and protein in phase. *Cell*, 149(6):1188–1191. PMID: 22682242.

Whitehead, J., Pandey, G. K., and Kanduri, C. (2009). Regulation of the mammalian epigenome by long noncoding RNAs. *Biochimica et biophysica acta*, 1790(9):936–947. PMID: 19015002.

Wilkinson, D. L. and Harrison, R. G. (1991). Predicting the solubility of recombinant proteins in escherichia coli. *Bio/technology (Nature Publishing Company)*, 9(5):443–448. PMID: 1367308.

Wilm, A., Higgins, D. G., and Notredame, C. (2008). R-coffee: a method for multiple alignment of non-coding RNA. *Nucleic Acids Research*, 36(9):e52. PMID: 18420654 PMCID: PMC2396437.

Wilusz, J. E., JnBaptiste, C. K., Lu, L. Y., Kuhn, C.-D., Joshua-Tor, L., and Sharp, P. A. (2012). A triple helix stabilizes the 3' ends of long noncoding RNAs that lack poly(a) tails. *Genes & development*, 26(21):2392–2407. PMID: 23073843 PMCID: PMC3489998.

Wolozin, B. (2012). Regulated protein aggregation: stress granules and neurodegeneration. *Molecular neurodegeneration*, 7:56. PMID: 23164372 PMCID: PMC3519755.

Wu, T., Wang, J., Liu, C., Zhang, Y., Shi, B., Zhu, X., Zhang, Z., Skogerbø, G., Chen, L., Lu, H., Zhao, Y., and Chen, R. (2006). NPInter: the noncoding RNAs and protein related biomacromolecules interaction database. *Nucleic acids research*, 34(Database issue):D150–152. PMID: 16381834 PMCID: PMC1347388.

Yao, Z., Macquarrie, K. L., Fong, A. P., Tapscott, S. J., Ruzzo, W. L., and Gentleman, R. C. (2013). Discriminative motif analysis of high-throughput dataset. *Bioinformatics (Oxford, England)*. PMID: 24162561.

Yu, D., Kim, M., Xiao, G., and Hwang, T. H. (2013). Review of biological network data and its applications. *Genomics & informatics*, 11(4):200–210. PMID: 24465231 PMCID: PMC3897847.

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 Research*, page gkt794.

Zarnack, K., König, J., Tajnik, M., Martincorena, I., Eustermann, S., Stévant, I., Reyes, A., Anders, S., Luscombe, N. M., and Ule, J. (2013). Direct competition between hnRNP c and U2AF65 protects the transcriptome from the exonization of alu elements. *Cell*, 152(3):453–466. PMID: 23374342 PMCID: PMC3629564.

Zhang, X., Gejman, R., Mahta, A., Zhong, Y., Rice, K. A., Zhou, Y., Cheunsuchon, P., Louis, D. N., and Klibanski, A. (2010). Maternally expressed gene 3, an imprinted

noncoding RNA gene, is associated with meningioma pathogenesis and progression. *Cancer research*, 70(6):2350–2358. PMID: 20179190 PMCID: PMC2987571.

Zhao, H., Yang, Y., and Zhou, Y. (2011a). Highly accurate and high-resolution function prediction of RNA binding proteins by fold recognition and binding affinity prediction. *RNA biology*, 8(6):988–996. PMID: 21955494 PMCID: PMC3360076.

Zhao, H., Yang, Y., and Zhou, Y. (2011b). Structure-based prediction of RNA-binding domains and RNA-binding sites and application to structural genomics targets. *Nucleic acids research*, 39(8):3017–3025. PMID: 21183467 PMCID: PMC3082898.

Zhao, J., Sun, B. K., Erwin, J. A., Song, J.-J., and Lee, J. T. (2008). Polycomb proteins targeted by a short repeat RNA to the mouse x chromosome. *Science (New York, N.Y.)*, 322(5902):750–756. PMID: 18974356 PMCID: PMC2748911.

Zuker, M. and Stiegler, P. (1981). Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. *Nucleic acids research*, 9(1):133–148. PMID: 6163133 PMCID: PMC326673.

Ørom, U. A., Derrien, T., Beringer, M., Gumireddy, K., Gardini, A., Bussotti, G., Lai, F., Zytnicki, M., Notredame, C., Huang, Q., Guigo, R., and Shiekhattar, R. (2010). Long noncoding RNAs with enhancer-like function in human cells. *Cell*, 143(1):46–58. PMID: 20887892.