Prediction of protein and nucleic acid interactions

Davide Cirillo

TESI DOCTORAL UPF / ANY 2016

DIRECTOR DE LA TESI

Dr. Gian Gaetano Tartaglia

DEPARTAMENT OF BIOINFORMATICS AND GENOMICS
AT CENTER FOR GENOMIC REGULATION (CRG)

A Claudio, Gina, Andrea e Giulia.
Grazie per avermi donato le ali
e un nido a cui fare ritorno.

Acknowledgments

"I want to tell you this story again. You already know it, I've told you a thousand times. By now, you are used to hear my endless stories about Barcelona. But you keep asking me to talk to you because for us our words are like bridges crossing the distance. And I never miss a chance to meet you even just on a bridge of words.

I was in the right place at the right time. Paolo and Anna were advertising that doctoral position as if it was the most precious opportunity of a student's life. At the time, it was okay for me to keep playing with my random forests of antibody structures (it sounds weird, isn't it?). I was ready for Berlin but then I chose Barcelona. They were right. That doctoral position turned out to be the most precious opportunity of my life.

New city, new responsibilities. The starting was a real challenge. But day after day I have been discovering that I was able to take that challenge and succeed. There's only one person that I have to thank for teaching me this lesson. Gian has been giving everything to all of us. I thank him for always being so committed and caring for the group like a family. Thanks to him I understood that you build a house one brick after the other; that the secret is exploring multiple threads as a source of learning; that the meaning of what you do is far beyond its difficulties. He taught me the importance of intuition and of being yourself without compromise. And all the virtue that comes from sharing big dreams with people you trust.

Right from the start, I had a good mate with me. I remember the first time that I have talked to Federico. It was on Skype, just before I joined the group (well, at the time he was *the* group!) To be honest, I felt like it was a kind of second interview! I have always seen him as someone to learn from, the guy with a better --or just different-- answer for any sort of question. The doors of his house were always open and those years we had such f#$*!ng good time together with Giovanni and tons of beers. I will never forget it.

In the second act of my doctorate incredible people and great friends were surrounding me. I will never have enough words to thank Mimma for her grace. She is the most generous and real person that I've ever met. Maybe she doesn't know but she really helped me out. Her smile is refreshing. I thank her for the time

spent at work and off, the terrible movies and the amazing dinners, the unspoken confidences and our scandalous love affair!! She makes me feel light and forget about all the bad things. And she is a wonderful co-chef!

Looking back, I can say I am the luckiest man on the planet for all the things that I have learnt from the many, many people who have been with me up to here. Benni, for her wisdom and for inspiring me on a daily basis to get better and do more. Riccardo, for all the *cortados con leche natural* and all the memorable moments that marked our work routine. Sergio, Alessandra, Silvina, (¡*los domingueros!*), for being such beautiful souls and invaluable friends. Nieves, for teaching me how to *atacatá* at work (it still sounds to me like a kind of kung-fu move!). Teresa, for her determination and brightness. Petr, Carmen, Stefanie, for their guidance and help (and all the chocolate, muffins and cheese!). Joana, Natalia, Silvia, Elias, for advice and for being always authentic and sincere. Alex, Fernando, for listening to me and making me feel like a mentor (an awesome one! oh yeah!). Shalu, for teaching me how to punch a bad day with a "Jai Ho!". Francesca, Andrea, Irene, Marta, for all the good times and their contagious energy (apparently not Francesca after waking up early!). Laura, for my birthday tiramisú when I was sick and on edge! Marcos, for his time. Mekayla, JF, Birgit, my unofficial motivational team, for all the funny support. My Thesis Committee members, my official motivational team, for all the serious support. Dario, Daniele, Romano, Adriano, Roberta, Lucrezia, Cristina, my dearest friends in Rome, for being always there every time that I am back. And finally Luciana, Elisa, Valerio, for being so pure (I love you unconditionally!).

And then you. You that are listening to this story once more. Those years would have not made sense without you. I feel so blessed because I met you. I keep on saying it to you every day and I will never stop doing it. Maybe one day we will fill that distance once and for all. And those bridges of words will become again looks and smiles and time to spend together. You have been with me all the days of this chapter of my life. These words are for you.

Davide

Abstract

The purpose of my doctoral studies has been the development of bioinformatics methods to quantitatively evaluate associations between proteins and nucleic acids (NAs). This thesis aims to provide insights into molecular features and relatively unknown mechanisms involving RNA-binding proteins and long noncoding RNAs as well as transcription factors and regulatory DNA elements. In this work, I present two algorithms, *cat*RAPID *omics express* and PA*n*DA, for the prediction of RNA- and DNA-protein interaction respectively. These computational methods offer the possibility to address experimental problems and guide new approaches facilitating experimental design and procedures.

Resumen

Mis estudios de doctorado han tenido como propósito principal el desarrollo de herramientas bioinformáticas para la evaluación de interacciones entre proteínas y ácidos nucleicos (ANs) de forma cuantitativa. Por consiguiente, esta tesis apunta a proporcionar conocimientos sobre características moleculares y mecanismos de asociación proteína-AN relativamente desconocidos; concretamente, la asociación de proteínas a ARNs y ARNs no codificantes, a la vez que factores de transcripción y elementos de regulación del ADN. En este proyecto presento dos algoritmos: *ca*tRAPID *omics express* y PA*n*DA, cuyas finalidades son las de predecir interacciones proteína-ARN y proteína-ADN respectivamente. Dichos métodos computacionales ofrecen la posibilidad de abordar problemas experimentales, así como de guiar el diseño y procedimiento de nuevas estrategias para su resolución.

Preface

The work carried out during my doctoral studies has been mainly focused on computational prediction of protein and nucleic acids (NAs) interactions. Protein-NAs interactions are involved in many cellular processes and can imply either transient or stable nucleoprotein complexes encompassing specific and nonspecific interactions. The study of protein-NA interactions occupies a prominent role in several research areas as well as in a large number of biotechnological and clinical applications. Not surprisingly, a vast number of works have provided deep insights into the functional implication of NA-binding protein complexes features in terms of sequence and structures. Despite of this research effort, difficulties associated with the experimental determination of protein-NA complexes and binding sites led to an urgent need for reliable and accurate computational predictions of NA binding in an automatic fashion.

In this thesis, I report the results obtained while testing and improving performances of the *cat*RAPID suite that is one of the most used computational frameworks for large-scale analysis of protein-RNA associations. Applications of *cat*RAPID approaches are described for several physiological and pathological processes namely neurodegenerative diseases (Chapter I), gene expression regulation (Chapter II and Chapter VII), and cancer (Chapter III). The development of *cat*RAPID *omics express* (Chapters III and VI) as a module of *cat*RAPID suite is presented. Furthermore, performances of PA*n*DA (Chapter IV), a new implementation for the prediction of protein-DNA interactions, are reported.

Both *cat*RAPID *omics express* and PA*n*DA are sequence-based methods that integrate genomic and functional annotations such as expression levels and protein-protein interaction networks. These methodologies pave the way for a better understanding of protein-NAs interaction features and will be valuable in providing information for numerous theoretical and practical applications.

Contents

INTRODUCTION

The association of proteins with nucleic acids is essential to life. This cellular event is a key element of the genetic blueprint and regulates mechanisms for its maintenance and variation. Therefore, the study of this macromolecular interaction is of paramount importance to understand cell growth, development, differentiation, evolution, and disease. Advances in computational biology are mirroring experimental approaches aiming to unveil details on the binding mechanism and regulation. As the experimental determination of binding sites is laborious and not always feasible, computational prediction of protein-NA interactions has been a fast-growing field in computational biology over the past two decades.

1. Chemistry and Biology of Nucleic Acids

1.1. Composition and structure of Nucleic Acid polymers

Nucleic acids (NAs), or polynucleotides, are biopolymers essential for storing and transmitting genetic information in nearly all living systems (Landenmark, Forgan, and Cockell 2015). Nucleic acids include DNA (2'-deoxyribonucleic acid) and RNA (ribonucleic acid), which differ form their structure, functions and stabilities (Alberts 1989; Bryce and Pacini 1998; Brown 2011). DNA and RNA are polymeric molecules composed of monomers known as nucleotides. Each nucleotide consists of a heterocyclic base (nucleobase or nitrogenous base), a pentose (5-carbon) sugar, and a phosphate group derived from phosphoric acid (H_3PO_4).

Organic bases found in nucleic acids are related either to the purine or to the pyrimidine heterocyclic ring systems. There are four heterocyclic bases in DNA: adenine (A), guanine (G), cytosine (C) and thymine (T). The first two are derived from purine, whereas the remaining two are derived from pyrimidine. The fourth base in RNA is not thymine but instead the pyrimidine-derived base, uracil (U).

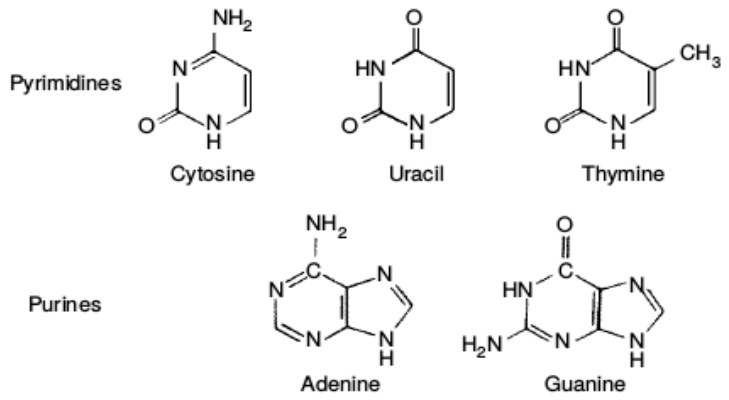


Figure 1. Chemical structures of the heterocyclic bases found in DNA and RNA [adapted from (Bryce and Pacini 1998)]

As for the 5-carbon sugars found in nucleic acids, deoxyribose (β-2'-deoxy-d-ribofuranose) is present solely in DNA while ribose (β-D-ribofuranose) is present solely in RNA. Ribose differs from deoxyribose for a hydroxyl group attached to the 2'-position of the pentose sugar (Figure 2).

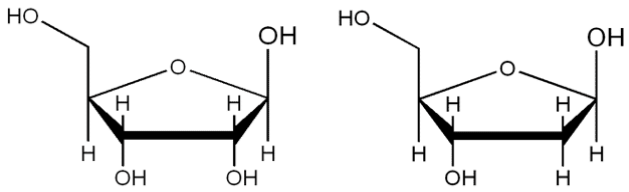


Figure 2. Ribose (right) and deoxyribose (left) 5-carbon sugars found in nucleic acids [adapted from https://en.wikibooks.org/wiki/Structural_Biochemistry]

A glycoside bond joins any one of the bases to either one of the two sugar molecules to form a compound known as a nucleoside. Addition of a phosphate group to the sugar residue of a nucleoside produces a compound known as nucleotide. A dinucleotide (dimer) of DNA or RNA is formed by covalently linking the 5′-phosphate

group of one nucleotide to the 3′-hydroxyl group of another to form a phosphodiester bond. An oligonucleotide (oligomer) is formed when several such bonds are made. Since at physiological pH of 7.4 each phosphodiester group exists as an anion, nucleic acids are highly charged polyanionic molecules (Figure 3).

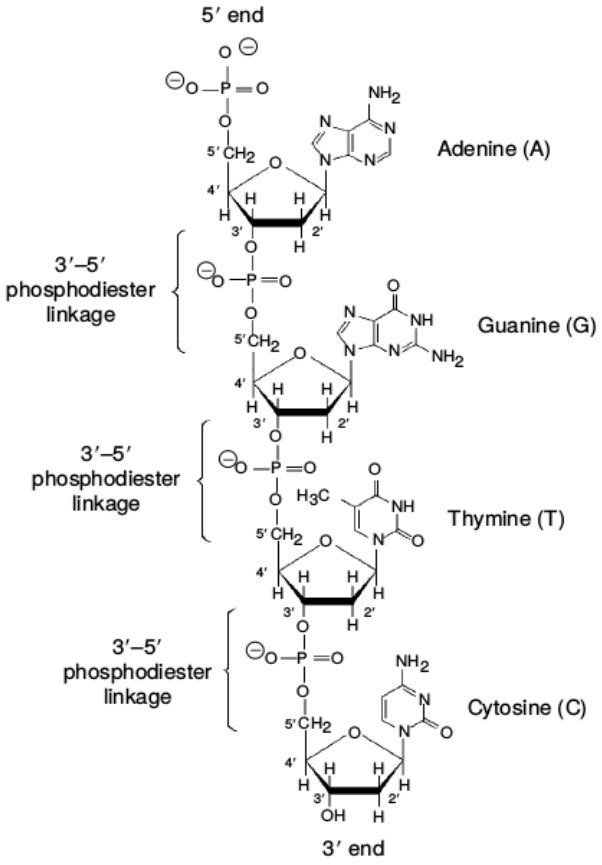


Figure 3. Diagram representation of the structure of an oligonucleotide [adapted from (Bryce and Pacini 1998)]

One end of a nucleic acid strand has a 5′-hydroxyl group (primary hydroxyl) and the other end has a 3′-hydroxylstructure of a polynucleotide group (secondary hydroxyl). The nucleic acid chain therefore has directionality. By convention, nucleic acid sequences are written in the 5' to 3' direction. It is important to stress that

distinct oligonucleotides, i.e. distinct sequences, are distinct molecules with different chemical and biophysical properties.

One of the factors responsible for the folded structure of both DNA and RNA is the intra-molecular base pairing that occurs within double-stranded nucleic acids chains. Dictated by specific hydrogen bonding patterns, the standard or canonical Watson-Crick base pairs [A-U(T) and G-C] (Figure 4) allow the DNA to maintain a regular helical structure that is dependent on its nucleotide sequence. In RNA molecules (e.g., transfer RNA), Watson-Crick base pairs permit the formation of short double-stranded helices, and a wide variety of non-canonical interactions or mismatches (e.g., G-U or A-A) (Figure 4) permit RNAs to fold into a vast range of specific three-dimensional structures. DNA with high GC-content is more stable than DNA with low GC-content, although the stability of the duplex is derived from both hydrogen bonding and base stacking (Yakovchuk, Protozanova, and Frank-Kamenetskii 2006).

It should be mentioned that non-canonical base pairing is possible and modified nucleobases do also occur (a comprehensive catalogue of modified nucleotides can be found at http://mods.rna.albany.edu/mods/).

Because of the canonical base pairing, the sequence of one strand of DNA precisely defines the sequence of the other; the two strands are said to be complementary, and are sometimes called reverse complements of each other. Nucleic acids can adopt different conformations: right-handed helices B-form and A-form, and left-handed helix Z-form. B-form has a wide major groove and a narrow minor groove running around the helix along the entire length of the molecule (Figure 5). B-form is found at low salt concentrations and it is believed to be the native conformation occurring in chromatin, a periodic structure made up of repeating, regularly spaced subunits called nucleosomes. Within the nucleosomes the major part of DNA is wrapped around histones, the remaining DNA known as linker DNA.

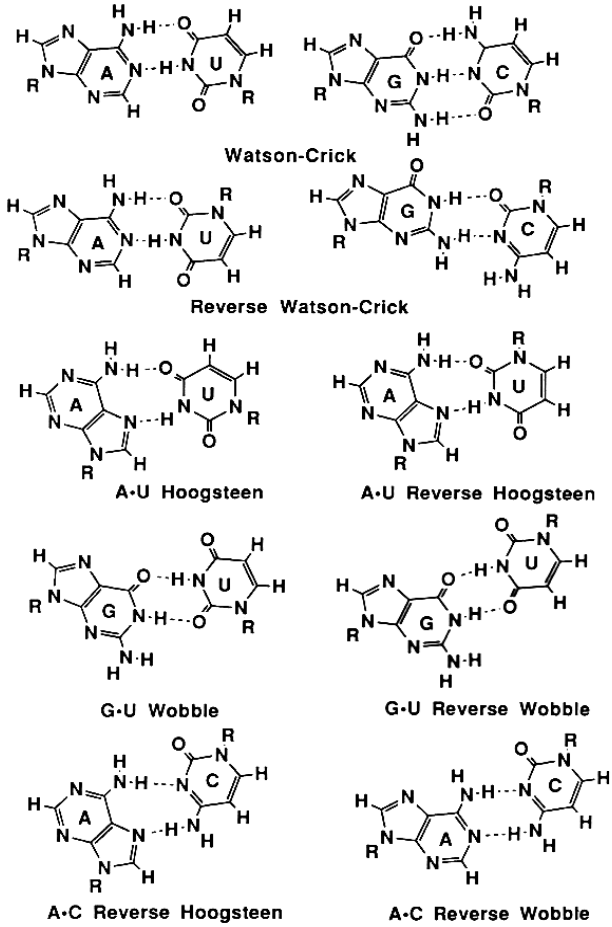


Figure 4. Canonical and non-canonical base pairs [adapted from (Crick 1993)]

A-DNA is found in solutions with higher salt concentrations or with alcohol added. RNA occurs almost exclusively in the A-form (or in a related A'-form). In A-DNA, the major groove is deep and the minor groove very shallow. Z-DNA occurs for alternating poly(dG-dC) sequences in solutions with high salt concentrations or alcohol. In addition, there exist further nucleic acid conformations like C-DNA, H-DNA or others.

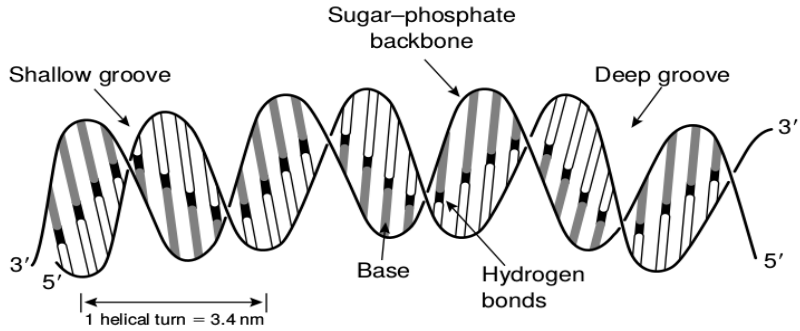


Figure 5. Major and minor grooves in DNA double helix [adapted from (Bryce and Pacini 1998)]

1.2. Properties of Base–Amino acid interface

As chains of amino acid residues, proteins are able to perform a vast range of activities within living systems, including interactions with NAs. DNA- and RNA-binding protein interfaces have diverse nature of binding sites at the atomic contact level (Gromiha 2011). In protein-DNA complexes, the grooves formed by the backbones of base pairs provide an interface for protein binding. In protein-RNA structures, both double- and single-stranded segments are found with some single-stranded regions stabilizing the structure and contributing to protein binding. While protein-DNA associations are governed predominantly by interaction of side-chain amino acids and functional groups in the major groove, RNA recognition is largely mediated by interactions of amide and carbonyl groups in the protein backbone with the edge of the RNA base (Allers and Shamoo 2001).

DNA-binding sites form packed, hydrophilic surfaces capable of direct and water-mediated hydrogen bonds (Susan Jones et al. 1999; Nadassy, Wodak, and Janin 1999). Conversely, RNA-binding sites are less tightly packed (see Table 1) and more frequently involved in van der Waals interactions (Susan Jones et al. 2001; Ellis, Broom, and Jones 2007). The wide range of conformations (e.g. loops, bulges, stems; see Figure 6) exhibited by RNA (Bon et al. 2008) may be responsible for the poor atomic packing. Furthermore, the convex nature of RNA surface that binds to the concave protein surface (Bahadur, Zacharias, and Janin 2008)

determines the typical asymmetry found in protein-RNA interface area (1208 $Å^2$ for protein and 1337 $Å^2$ for RNA). Nonetheless, shape complementarity is a primary feature for the two kind of complex formation (see Table 1). Interestingly, large interfaces comprised into protein-RNA complexes are suspected to be under higher selection pressure (Barik et al. 2015).

	Protein-DNA	**Protein-RNA**
Interface size	3137 $Å^2$ (1600 $Å^2$✶)	2545 $Å^2$
Number of aa	24✶	45
Number of nt	12✶	16
Atomic packing †	6.1 Å	8.9 Å
Shape complementarity ‡	0.65	0.67
Salt bridges	11	8
Stacking interactions	Purines: 35% Pyrimidines: 65%	Purines: 54% Pyrimidines: 46%

Table 1. Average values of relevant structural parameters of DNA- and RNA-protein interfaces. † (S Jones and Thornton 1996); ‡ (Lawrence and Colman 1993); ✶(Nadassy, Wodak, and Janin 1999)[adapted from (Barik et al. 2015)].

The immediate proximity of peptides and nucleotides involves a mutual action at atomic level exhibiting favoured amino acid-base hydrogen bonds and van der Waals contacts (Nicholas M. Luscombe, Laskowski, and Thornton 2001; Treger and Westhof 2001). In particular, for both DNA- and RNA-protein interfaces, positively charged (Arginine and Lysine), polar (Threonine and Asparagine) and aromatic (Phenylalanine) amino acids play a predominant role in mediating specific and nonspecific interactions with certain base types or sequence contexts (Nicholas M.

Luscombe, Laskowski, and Thornton 2001; Treger and Westhof 2001; Susan Jones et al. 2001).

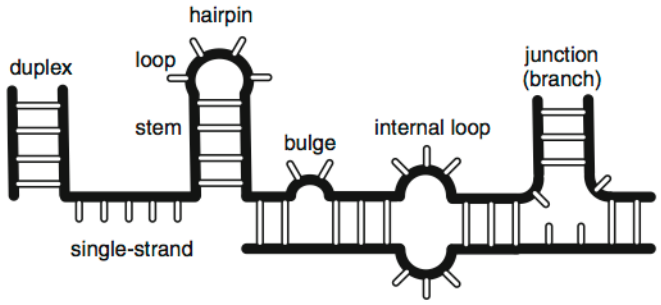


Figure 6. RNA secondary structure feature and naming conventions [adapted from (Zwieb 2014)].

As a consequence of the chemical affinities, many DNA- and RNA-binding proteins can recognize specific base pairing patterns that identify particular regulatory regions of genes and transcripts. A nucleotide sequence pattern with functional or chemical significance (e.g. a DNA site of high affinity for protein molecules) is called a sequence motif (Stormo 2000). Sometimes patterns are defined in terms of probabilistic models, such as Positional Weight Matrices (PWMs) (Stormo 2000), and can be graphically represented using sequence logos (Schneider and Stephens 1990) showing the most conserved bases in a set of aligned sequences. Given a set of sequences, many bioinformatics programs attempt to identify candidate motifs. Renowned *de novo* motif discovery tools include MEME (T. L. Bailey et al. 2015), SeAMotE (Agostini et al. 2014), and many others (Lihu and Holban 2015). Although motif representation of binding site positions is well-established, novel probabilistic paradigms are regularly proposed to yield better prediction performances such as the sparse local inhomogeneous mixture (Slim) model (Keilwagen and Grau 2015) that takes into account inter-position dependence (Mukherjee et al. 2013). Nonetheless, an unbiased and comprehensive evaluation of the differences between binding sites recognized *in vivo* and *in vitro* needs to be attempted.

Notable repositories of DNA-binding proteins recognition motifs are FlyTF (Pfreundt et al. 2010) and LEGO Factors (Stampfel et al.

2015) (*Drosophila melanogaster*), CollecTF (Kiliç et al. 2014) (Bacteria domain), DPInteract (Robison, McGuire, and Church 1998) (*Escherichia coli*), AthaMap (Steffens et al. 2005) (*Arabidopsis thaliana*), ScerTF (Spivak and Stormo 2012) (*Saccharomyces* species), HOCOMOCO (Kulakovskiy et al. 2016) (*Homo sapiens* and *Mus Musculus*), TRANSFAC (Wingender et al. 2000) and Factorbook (J. Wang et al. 2012) (*Homo sapiens*), JASPAR (Mathelier et al. 2016) (Eukaryotes), and UniPROBE (Hume et al. 2015) (several organisms).

Available repositories of RNA-binding proteins recognition motifs are the database of RNA-Binding Protein DataBase (RBPDB) (Cook et al. 2010), RNAcompete compendium (Ray et al. 2013), RNA Bricks (Chojnowski, Walen, and Bujnicki 2014), INTERactions in RNA structures (InterRNA) (Appasamy et al. 2016), RNA Characterization of Secondary Structure Motifs (RNA CoSSMos) (Vanegas et al. 2012), RNA 3D Motif Atlas (Petrov, Zirbel, and Leontis 2013).

A polypeptide sequence pattern with functional or chemical significance (e.g. a protein site of high affinity for DNA molecules) is called a domain (Richardson 1981). A protein domain is a highly conserved structural unit that can function and evolve almost independently of the rest of the protein (Ponting and Russell 2002; Orengo and Thornton 2005). Proteins often include multiple domains, many of which can be traced back to the Last Universal Common Ancestor (LUCA) (Ranea et al. 2006), even though their origin is still poorly understood (Alva, Söding, and Lupas 2016). Protein domain assignments using Pfam (Finn et al. 2014), InterPro (A. Mitchell et al. 2015), and other domain annotation resources are widely used to infer protein evolutionary and functional relationships.

Although some protein domains have clearly understood functions (Forslund and Sonnhammer 2008), many proteins are able to undergo specific processes even in absence of the conforming canonical domain. Remarkable examples are amyloidogenic proteins associated with neurodegenerative disorders acting as transcription factors (Hegde, Vasudevaraju, and Rao 2010; Maloney and Lahiri 2011), and metabolic enzymes acting as RNA-binding proteins (Beckmann et al. 2015; Castello, Hentze, and Preiss 2015).

Those proteins are referred to as *moonlighting* proteins. Especially concerning RNA recognition, recent experimental studies (Castello et al. 2012; Baltz et al. 2012; Kwon et al. 2013; Castello et al. 2016) indicate that a number of RNA-binding proteins contain non-classical RNA-binding domains and are not annotated in RNA-related pathways (Gerstberger, Hafner, and Tuschl 2014). For both classes of NA-binding proteins, structural disorder (Guharoy, Pauwels, and Tompa 2015) emerged as a prevalent and important feature in establishing the interaction with nucleotides (Castello et al. 2012; Klus et al. 2015). Based on such discoveries, a new generation of knowledge-free computational methods for domain detection has been developed (Carmen Maria Livi et al. 2015).

2. Discovery of Protein and Nucleic Acid interactions

Protein–NA interactions play a crucial role in central biological processes, ranging from mechanisms of replication, transcription and recombination to enzymatic events utilizing nucleic acids as substrates. For these reasons, biochemical and structural studies of protein–NA recognition processes are of general relevance. Many multidisciplinary approaches have been posing unique and challenging views on protein-NA interactions. Advances in genomic techniques to identify NA-binding proteins and their targets, as well as methods to elucidate their functions, are calling for the development of novel computational frameworks for the analysis of protein-NA interaction data. Hence, the broad range of methodologies required for a mechanistic understanding of protein–NA interactions and their functions in the cell covers both structural and genomic aspects investigated with both experimental and computational techniques.

2.1 The interaction as a structural event

X-ray crystallography (Shi 2014), along with Nuclear Magnetic Resonance NMR (Marion 2013) and other chemical and physical methods (Hanein and Milligan 2013) are used to discover how proteins and NAs interact with each other. Protein–DNA interactions have been documented for individual structures, and the literature on the subject has been reviewed in detail (N. M. Luscombe et al. 2000) along with protein-RNA interactions (Susan Jones et al. 2001). Currently, the Protein Data Bank (PDB; [http://www.rcsb.org/pdb]) (Berman et al. 2000) includes more than 5000 DNA/RNA-protein complex structures that is about 800 non-redundant DNA/RNA binding protein chains (below 25% sequence identity), corresponding to only 5% of the number of available protein structures (Zhao, Yang, and Zhou 2013; Miao and Westhof 2015).

In addition to PDB, structures of protein–NA complexes are deposited in the Nucleic Acid Database (NDB) (Coimbatore Narayanan et al. 2014), and also specific collections such as Protein-RNA Interface Database (PRIDB) (Lewis et al. 2011), 3D-

footprint (Contreras-Moreira 2010), Nucleic acid–Protein Interaction DataBase (NPIDB) (Kirsanov et al. 2013), Transcription Factor Binding Site Shape (TFBSshape) (L. Yang et al. 2014), Transcription factor-DNA interaction data repository (TFinDit) (Turner, Kim, and Guo 2012), Biological Interaction Database for Protein-Ncucleic Acid (BIPA) (Lee and Blundell 2009), Thermodynamic Database for Protein-Nucleic Acid Interactions (ProNIT) (M. D. S. Kumar et al. 2006) Protein-DNA Structure-Affinity Database (PDSA) (AlQuraishi, Tang, and Xia 2015), and Telomeric Proteins Interaction Network (TeloPIN) (Luo et al. 2015).

Considering practical problems occurring in experimental structural biology such as high costs, poor NMR spectra of larger proteins, and conformational changes due to packing interactions in crystallisation (Acharya and Lloyd 2005), computational modelling of protein–NA complexes represents a powerful alternative to prompt investigation and discovery into the field (Karplus and Lavery 2014; Zhou 2014). Nonetheless, a problem for *in silico* simulations of protein-NA complexes, associated with the training of force fields (MacKerell and Nilsson 2008), makes calculations as limited as the application of experimental methods for determining molecular conformations (Bränd´en and Alwyn Jones 1990). For instance, the Critical Assessment of PRediction of Interactions CAPRI (http://capri.ebi.ac.uk) international challenge (Janin 2010) and the three protein-RNA benchmarks available in the literature up-to-date (Barik et al. 2012; Pérez-Cano, Jiménez-García, and Fernández-Recio 2012; Huang and Zou 2013) show that molecule flexibility still remains a computational issue to overcome. Importantly, known complex structures are still very few compared with known sequence space (only less than 1/1000th proteins of known sequences have experimental structures available (Moult 2008)), and the ease of crystallization confines the set of solved structures to a non-random sampling.

2.2. The interaction as a genomic event

In addition to efforts to improve knowledge on protein-NA interactions at the structural level, other approaches, both experimental and computational, have been developed based on

chemical properties and cellular context of the recognition or binding event. Such techniques are based on next-generation sequencing (van Dijk et al. 2014) and proteomics (Larance and Lamond 2015) and exploit chemical specificities of interacting molecules in a large-scale context. Indeed, *in vivo* experimental techniques to check protein-DNA and protein-RNA interactions share the same principles, which often imply immunoprecipitation (IP) (see Figure 7), i.e. precipitation of the protein of interest using a specific antibody. Such techniques form a precious toolbox for understanding protein-NA interactions at the finest resolution and broadest scale. Notable online resources for experimental protein-NA interactions are ENCODE (https://www.encodeproject.org/) and NCBI GEO (http://www.ncbi.nlm.nih.gov/geo/). Repositories of protein-RNA interactions are CLIPZ (Khorshid, Rodak, and Zavolan 2011), iCounts (Anders et al. 2012), Atlas of UTR Regulatory Activity (AURA) (Dassi et al. 2014), CLIPdb (Y.-C. T. Yang et al. 2015), Database of RNA interactions in post-transcriptional regulation (DoRiNA) (Blin et al. 2015). Computational methods build upon more highly heterogeneous approaches [reviewed in (Cirillo, Agostini, and Tartaglia 2013; K 2013; Cirillo, Livi, et al. 2014; Si et al. 2015)]. The development of algorithms for a comprehensive analysis of high-throughput data represents one of the main challenges for the bioinformatics community and a rational aid to experimental scientists. Computational methods offer the possibility to address experimental problems, prompt functional hypotheses, and guide new approaches. Indeed, the prediction of structural and functional properties of macromolecules could largely facilitate the process of designing experimental procedures and protocols.

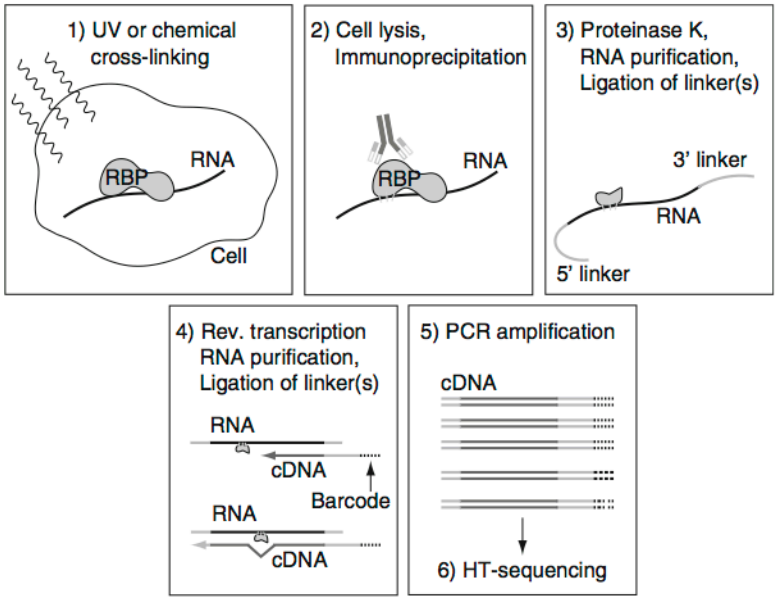


Figure 7. Overview of CLIP-based methods [adapted from (Re et al. 2014)]

3. Experimental methods for protein and Nucleic Acids interactions detection

Experimental characterization of protein-NA interactions can be broken down into *in vitro* approaches, which determine the specificity of NA-binding proteins free from other cellular factors, and *in vivo* approaches, which measure a snapshot of proteins binding to DNAs or expressed RNAs.

3.1 *In vitro* protein-NAs interactions

In vitro methods for the determination of DNA-binding protein targets are DNA Electrophoretic Mobility Shift Assay (EMSA) (Hellman and Fried 2007), which is based on the observation that the rate of DNA migration is shifted or retarded upon protein binding when subjected to non-denaturing polyacrylamide or agarose gel electrophoresis; the DNA Pull-down Assay, which purifies the components of a protein-DNA complex either by Western blot/mass spectrometry when using a biotinylated DNA, or by Southern blotting/PCR when using a protein labelled with an affinity tag; and Microplate Capture and Detection Assay (Gibellini et al. 1993), which uses immobilized DNA probes to capture specific protein entities.

In vitro methods for the determination of RNA-binding protein targets are Systematic evolution of ligands by exponential enrichment (SELEX) (Ellington and Szostak 1990), which consists of multiple rounds of binding and amplification of RNA molecules; SEQRS (Campbell et al. 2012), which modifies traditional SELEX by sequencing the bound RNA pool at each round; RNAcompete (Ray et al. 2009), which assays the bound pool of designed RNA using microarray; and RNA Bind-n-Seq (Lambert et al. 2014), which sequence RNAs bound to various amounts of proteins after incubation.

3.2 *In vivo* protein-DNA interactions

Chromatin immunoprecipitation (ChIP) experiments (Pillai, Dasgupta, and Chellappan 2015) allow the capture of a snapshot of specific protein-DNA interactions and to quantitate them by means

of quantitative polymerase chain reaction (qPCR). *In vivo* crosslinking, traditionally achieved with formaldehyde, covalently stabilizes protein-DNA complexes, allowing even transient interactions to be trapped (Jackson 1978). Importantly, formaldehyde fixation step is also able to stabilize protein-protein interactions (Hoffman et al. 2015; Gavrilov, Razin, and Cavalli 2014; Cirillo, Botta-Orfila, and Tartaglia 2015).

Crosslinked protein-DNA complexes are extracted with a lysis step that dissolves the cell membrane with detergent based solutions. DNA is sheared by sonication or digestion with micrococcal nuclease MNase (C. Wu and Allis 2004). ChIP validated antibodies are then used to isolate the complex of interest. Alternatively, affinity tags such as HA, myc or GST fused to target proteins can be used to immunoprecipitate target proteins lacking qualified antibodies (Lichty et al. 2005). Beaded antibody binding resin like protein A, G or A/G, or immobilized streptavidin in the case of biotinylated antibodies, are used to affinity purify the complex using blocking buffers such as salmon sperm DNA and a generic protein source. Importantly, increase in bead volume increases non-specific binding (Lin, Tirichine, and Bowler 2012). Crosslinks are reversed typically through extensive heat incubations or through digestion of the protein component with proteinase K, which cleaves at the carboxy-side of aliphatic, aromatic or hydrophobic residues and also eliminates nucleases from the purified DNA preventing degradation. After DNA purification using phenol-chloroform, DNA levels can be determined by agarose gel electrophoresis or more commonly by quantitative polymerase chain reaction (qPCR). The direct correlation between the amounts of immunoprecipitated complex and bound DNA (Blecher-Gonen et al. 2013) makes qPCR procedures sufficiently accurate to enable measurement of target protein-DNA levels in different experimental conditions.

ChIP technology has fostered advanced specializations and offshoot techniques. ChIP coupled with microarray analysis (ChIP-chip) (Ren et al. 2000) allows genome-wide analysis of protein or protein modification distribution. Purified DNA sample and a control (the input sample or an IP with a non-specific antibody) are each fluorescently labelled and co-hybridized to a microarray (Aparicio, Geisberg, and Struhl 2004). Despite the relatively inexpensive

costs, the main disadvantages of ChIP-chip are the inherent restrictions of microarray technology, and the limited resolution and higher signal to noise ratio compared to sequencing technologies (Ho et al. 2011; Massie and Mills 2012). ChIP coupled with quantitative next-generation sequencing technology (ChIP-seq) (Lieb et al. 2001; Johnson et al. 2007) identifies binding sites of DNA-associated proteins detecting enrichment of chromatogram peaks. ChIP-seq main disadvantages are its costs and its limitations in the case of rare sample types (Gilfillan et al. 2012). Remarkably, ChIP-seq is the primary technology used in the ENCODE (Encyclopedia of DNA Elements) project (Landt et al. 2012; T. Bailey et al. 2013).

To better address biological questions or to modify the resolution and scale of the experiments, researchers have created a specialized version of ChIP. ChIP-exo (Rhee and Pugh 2012) is used to specifically map binding sites in the genome via the addition of a DNA digestion step to ChIP-seq. ChIA-PET (Chromatin Interaction Analysis by Paired-End Tag Sequencing) (J. Zhang et al. 2012) couples ChIP with chromatin conformation capture (3C) technology (Sajan and Hawkins 2012) to detect the interaction of distant DNA regions via a protein of interest.

3.3 *In vivo* protein RNA-interactions

The two major approaches for analyzing protein-RNA interactions *in vivo* are RNA Immunoprecipitation (RIP) (Jain et al. 2011) and Cross-Linking Immunoprecipitation (CLIP) (Milek, Wyler, and Landthaler 2012; Riley and Steitz 2013). Both RIP and CLIP are similar to DNA-based ChIP in that they use antibodies to isolate specific nucleic acid-protein interactions.

RIP involves IP of an RNA-binding protein (RBP) of interest using an antibody. RIP can be coupled to microarray (RIP-ChIP) (Keene, Komisarow, and Friedersdorf 2006) or sequencing (RIP-seq) (Cloonan et al. 2008). The use of a recombinant protein to probe an isolated total RNA sample is known as recombinant RIP (rRIP) (Townley-Tilson et al. 2006). Disadvantages of RIP protocols include lack of RBP binding site detection, non-specific RNA

interaction identification, and high signal-to-noise ratio (Mili and Steitz 2004).

CLIP technologies (Milek, Wyler, and Landthaler 2012; Riley and Steitz 2013) differ from RIP in their use of UV crosslinking (Brimacombe et al. 1988). As weak and non-specific protein interactions are not crosslinked, CLIP protocols allows stringent isolation conditions, hence a reduced background noise and an increased resolution leading to actual RBP binding sites identification to within a few nucleotides. CLIP is generally coupled to sequencing as in the case of CLIP-seq also known as HITS-CLIP (high-throughput sequencing of RNA isolated by crosslinking immunoprecipitation) (Darnell 2010). PAR CLIP (photoactivatable-ribonucleoside-enhanced crosslinking and immunoprecipitation) (Ascano et al. 2012) attempts to enhance efficiency of crosslinking and resolution of RBP binding site by incorporating photoreactive ribonucleoside analogs into nascent RNA of living cells, which undergoes characteristic mutations in the sequence upon crosslinking (Spitzer et al. 2014). The major disadvantage of PAR-CLIP is its limitation to cell cultures which are the only experimental system for animal studies to be conducive to incorporation of the ribonucleoside analog (Ascano et al. 2012). iCLIP (Konig et al. 2011; Huppertz et al. 2014) allows single-nucleotide resolution of RBP binding sites. After partial digestion of the protein by proteinase K, cDNA synthetized from an adapter sequence ligated at the 3'end of RNA is circularized resulting in its location next to the last nucleotide before the RBP binding site. The very detailed protocol and specialized data analysis tools represent the main disadvantages of iCLIP technique (Huppertz et al. 2014; Chen et al. 2014).

ChIRP (chromatin isolation by RNA purification), CHART (capture hybridization analysis of RNA targets) and RAP (RNA antisense purification) exploit biotinylated oligonucleotides complementary to the RNA of interest as a way to pull down associated proteins (C. Chu et al. 2011; Simon et al. 2011). Mass spectrometry and next-generation sequencing are employed to identify proteins associated with RNA and genomic locations at which these interactions occur.

4. Computational methods for protein and Nucleic Acid interactions prediction

In the context of theoretical methods for computational biology, selection of relevant information is essential to build predictive models (T. M. Mitchell 1982). As predictions are based upon previous experience, predictive approaches are intrinsically limited by the initial hypotheses and the choice of features to describe the system.

Computational prediction of protein-NA binding requires experimental knowledge on whether a given protein binds NAs. This information can be retrieved from structural (e.g. X-ray crystallography) or genomic (e.g. ChIP and CLIP technologies) approaches. The definition of a NA binding residue changes if it is based on distance cut-offs (generally from 3.5 to 7Å), on non-covalent contacts (e.g. hydrogen bonding and Van de Waals interactions), or even on changes in accessible surface area ($\Delta ASA > 0 Å^2$ or >10% area change) (Miao and Westhof 2015). Moreover, single point mutations are able to maintain the structure of a binding site but disable the binding ability (Arnaud et al. 2011), and a binding site can be associated with multiple activities like in the case of *moonlighting* proteins (Huberts and van der Klei 2010).

Based on the kind of features exploited, protein-NA interaction prediction methods can be roughly divided in two major categories: structure-based methods and sequence-based methods (Cirillo, Agostini, and Tartaglia 2013). Sequence based methods take advantage of the information collected within primary sequences of protein and NAs. In general, statistical analysis of a large collection of sequences known to be involved in an interaction leads to the creation of a model that is further used to identifying novel binding regions. In contrast, structure-based methods use the geometric shape of protein and NAs to describe interactions at atomic level and derive binding affinities and rules of binding recognition. Both methodologies can return binary predictions (binding or not-binding) or score-based predictions generally including arbitrary cut-offs.

Common features used in sequence-based methods are:

- Nucleic acids and amino acid composition (e.g. sequence binary encoding);
- Sequence similarity (e.g. multiple sequence alignment coupled to conservation scoring such as Shannon entropy, Scorecons (Valdar 2002), etc.);
- Evolutionary information [e.g. position-specific scoring matrix (Stormo 2000)].

Common features used in structure-based methods are:

- Secondary structure [e.g. assessed from the structure using DSSPcont (Carter, Andersen, and Rost 2003)];
- Accessible surface area [e.g. assessed from the structure using NACCESS (Hubbard and Thornton 1993)];
- Physicochemical features (e.g. hydrophobicity, electrostatic patches, cleft size, charge, dipole, quadrupole moments, etc.).

Secondary structure and physicochemical features can be predicted using primary structure (Bellucci et al. 2011).

An up-to-date selection of protein-NA interaction prediction methods is reported in Appendix.

CHAPTER I

Protein-RNA interactions in Neurodegenerative diseases

Established in 2010, research in Tartaglia's lab at Center for Genomic Regulation (CRG) of Barcelona, Spain, focuses on neurodegenerative diseases. Although neurodegenerative diseases are traditionally described as protein disorders leading to amyloidosis, recent evidence indicates that protein-RNA associations are involved in their onset. In this work, we used *cat*RAPID to investigate a number of protein–RNA associations involved in neuronal function and misfunction such as protein–RNA interactions associated with fragile X syndrome; protein sequestration in CGG aggregates; the TDP-43 noncoding interactome; FMRP and TDP-43 autogenous regulation; iron-mediated translation of APP and α-synuclein transcripts; and prion proteins and RNA aptamers. The strong agreement of our calculations with experimental evidence encouraged us to propose putative candidates in the disease mechanism to be further investigated by experimental studies. This work also introduces two new modules of catRAPID suite: catRAPID *strength* and catRAPID *fragments*. catRAPID *strength* is a tool to estimate the specificity of an interaction under study; catRAPID *fragments* allows the analysis of associations between molecules with long sequences. The work also set the basis to new lines of research at Tartaglia's lab such as ribonucleoprotein associations of triplet repeat expansions; design of RNA aptamers for neurodegenerative diseases; and the autogenous regulation of gene expression (Zanzoni et al. 2013).

Cirillo D, Agostini F, Klus P, Marchese D, Rodriguez S, Bolognesi B, Tartaglia GG. Neurodegenerative diseases: quantitative predictions of protein-RNA interactions. RNA. 2013 Feb;19(2):129-40. doi: 10.1261/rna.034777.112. Epub 2012 Dec 21. PMID: 23264567

Cirillo D, Agostini F, Klus P, Marchese D, Rodriguez S, Bolognesi B, Tartaglia GG. Neurodegenerative diseases: quantitative predictions of protein-RNA interactions. RNA. 2013 Feb;19(2):129-40. doi: 10.1261/rna.034777.112

CHAPTER II

Regulatory functions of ncRNAs

My PhD involved data analysis projects complementary to the methodological part. In this chapter I present a work carried out in collaboration with the Bellvitge Institute for Biomedical Research (IDIBELL). The aim of the work was to produce and analyse the allelic expression screen of an imprinted domain on mouse chromosome 10 comprising the paternally expressed *Plagl1* gene. One result of the study was the identification of two unspliced ncRNAs, *Hymai* and *Plagl1it*. My contribution to the project had been to help defining the interaction propensity between *Hymai* and *Plagl1it*, and Trithorax chromatin regulators. This analysis allowed the identification of a potential regulatory function at the imprinted domain. This work was published in 2012 in PLoS One, and is a typical example of how the wealth of high-throughput data sets new challenges for protein-RNA interaction prediction.

Iglesias-Platas I, Martin-Trujillo A, Cirillo D, Court F, Guillaumet-Adkins A, Camprubi C, Bourc'his D, Hata K, Feil R, Tartaglia G, Arnaud P, Monk D. Characterization of novel paternal ncRNAs at the Plagl1 locus, including Hymai, predicted to interact with regulators of active chromatin. PLoS One. 2012;7(6):e38907. doi: 10.1371/journal.pone.0038907. Epub 2012 Jun 19. PMID: 22723905

Iglesias-Platas I, Martin-Trujillo A, Cirillo D, Court F, Guillaumet-Adkins A, Camprubi C, Bourc'his D, Hata K, Feil R, Tartaglia G, Arnaud P, Monk D. Characterization of novel paternal ncRNAs at the Plagl1 locus, including Hymai, predicted to interact with regulators of active chromatin. PLoS One. 2012;7(6):e38907. doi: 10.1371/journal.pone.0038907.

CHAPTER III

Interaction determines expression

In this work, published in *Genome Biology*, we use *cat*RAPID algorithm to integrate computational predictions of protein-RNA interactions with experimental expression profiles. Remarkably, our analysis uncovered novel regulatory paradigms concerning proliferation and differentiation processes. The work linked experimentally determined tissue-specific expression patterns of known human mRNA-binding proteins (RBPs) and thousands of mRNAs. As such associations are experimentally known for just a small subset of molecules, our computational strategy allowed to generalize to a proteomic scale and reach an unprecedented scope. We found that mRNA-RBP pairs for which the *cat*RAPID algorithm predicts a high interaction propensity tend to have strongly correlated or strongly anti-correlated expression patterns in human tissues. By analysing functional categories, we detected a strong enrichment of functions related to cell-cycle control among the positively correlated patterns and those for survival, growth and differentiation among negatively correlated patterns. Furthermore, over 90% of genes in both categories are listed as cancer-related genes. Due to its large-scale implications and the soundness of predictions, the study has high potential to guide and inspire future experimental work. As commented by colleagues, "the overall picture painted embodies important principles that are here to stay, robust to false discoveries in the prediction set" (Zagrovic 2014).

Cirillo D, Marchese D, Agostini F, Livi CM, Botta-Orfila T, Tartaglia GG. Constitutive patterns of gene expression regulated by RNA-binding proteins. Genome Biol. 2014 Jan 2;15(1):R13. doi: 10.1186/gb-2014-15-1-r13. PMID: 24401680

Cirillo D, Marchese D, Agostini F, Livi CM, Botta-Orfila T, Tartaglia GG. Constitutive patterns of gene expression regulated by RNA-binding proteins. Genome Biol. 2014 Jan 2;15(1):R13. doi: 10.1186/gb-2014-15-1-r13.

CHAPTER IV

PA*n*DA, Protein And DNA Associations

Transcription factors are proteins that bind to specific patterns of DNA sequences to control how genes are turned on or off. The way this function is achieved is still unknown. To gain insights into this mechanism, we analysed a large collection of ENCODE ChIP-seq data to study how transcription factors interact together with specific DNA regions. We found that the association of multiple transcription factors is a fundamental feature to explain their localization onto DNA. We developed a computational method that uses this feature to predict where a transcription factor will localize in the genome. This tool is called PA*n*DA (Protein And DNA Associations). The very high accuracy of PA*n*DA shows that the network itself contains enough information to localize transcription factors on DNA even in absence of known recognition motifs. The most innovative aspect of our work is that it introduces a cell-specific view of transcription factors networks, which opens up the way for efficient and effective manipulation of cellular processes. PA*n*DA tool will raise new fundamental questions in the field and will inspire future research on topics like the evolution of regulatory networks and the formation of macromolecular complexes.

Cirillo D, Botta-Orfila T, Tartaglia GG. By the company they keep: interaction networks define the binding ability of transcription factors. Nucleic Acids Res. 2015 Oct 30;43(19):e125. doi: 10.1093/nar/gkv607. Epub 2015 Jun 18. PMID: 26089389

Cirillo D, Botta-Orfila T, Tartaglia GG. By the company they keep: interaction networks define the binding ability of transcription factors. Nucleic Acids Res. 2015 Oct 30;43(19):e125. doi: 10.1093/nar/gkv607.

CHAPTER V

Refining Xist interactome

This chapter presents a recent submission for publication in *Nature Structural and Molecular Biology*. Mammalian female-specific process of X Chromosome Inactivation (XCI) is critically dependent on a long non-coding RNA called *Xist*. At the onset of X inactivation, *Xist* spreads in *cis* on the future inactive X and triggers gene silencing by recruitment of repressive DNA and chromatin modifiers. In this study I explored the protein interactome of *Xist* through a multifaceted approach aiming at identify direct Xist binders. Five proteomic and genetic studies recently revealed a large and heterogeneous list of binding proteins containing *bona fide* interactors as well as transient and spurious interactions. The *Global Score* method based on the *cat*RAPID *fragment* algorithm (Cirillo et al. 2013) (Chapter I) was applied to identify specific and direct associations. Using enhanced individual-nucleotide resolution Cross-Linking and ImmunoPrecipitation (eCLIP), we validated our predictions for Spen, Hnrnpk, Lbr, Ptbp1, and Hnrnpu/Saf-A proteins, reporting a global prediction accuracy of ~80%. An innovative aspect of this approach is the investigation of protein networks involved in *Xist* regulation. The computational method and pipeline presented in this work can be easily applied to the study of other lncRNAs.

Cirillo D, Blanco M, Buness A, Avner P, Guttman M, Tartaglia GG, Cerase A. A Computational Approach Reveals Direct Protein Interactions to the Long Non-Coding RNA *Xist*. Nature Structural and Molecular Biology (submitted).

A computational approach for identification of protein-RNA interactions uncovers direct binders of Xist lncRNA

Davide Cirillo[1,5,*], Mario Blanco[2,5,*], Andreas Buness[3], Philip Avner[3], Mitchell Guttman[2], Gian GaetanoTartaglia[1, 4, 5,**] and Andrea Cerase[3,**]

1) Centre for Genomic Regulation (CRG), The Barcelona Institute of Science and Technology, Dr Aiguader 88, 08003 Barcelona, Spain
2) Division of Biology and Biological Engineering, California Institute of Technology, 1200 E. California Blvd, MC 156-29, Pasadena, CA 91125
3) EMBL-Monterotondo, Via Ramarini 32, 00015 Monterotondo (RM), Italy
4) Universitat Pompeu Fabra (UPF), 08003 Barcelona, Spain
5) Institucio Catalana de Recerca i Estudis Avanc ats (ICREA), 23 Passeig Lluıs Companys, 08010 Barcelona, Spain

*) Equal contribution
**) To whom correspondence should be addressed andrea.cerase@embl.it, gian@tartaglialab.com

Abstract

Computational frameworks predicting protein-RNA networks provide an important source of information for understanding the regulation of long non-coding RNAs (lncRNAs) and complement experimental approaches. We here introduce the *cat*RAPID *Global Score* to calculate direct binders of *Xist* lncRNA, the master regulator of X Chromosome Inactivation (XCI). Using enhanced individual nucleotide CLIP method (eCLIP), we validated our predictions for five candidates. We proved that *Global Score* can efficiently predict which protein domain and RNA fragments mediates the interaction. We showed that *Global Score* can be used as a tool to prioritize *bona fide* direct interactors from high-throughput data or Gene Ontology functional categories. Our approach paves the way for a novel approach to study ribonucleoprotein interactions involved in non-coding RNA regulation.

Glossary:

CHART-seq: Capture Hybridisation Analysis of RNA Target-Sequencing
ChIRP-MS: Chromatin Isolation by RNA Purification-Mass Spectrometry
ESC: Embryonic Stem Cells
GO: Gene Ontology
HnrnpK: Heterogeneous nuclear ribonucleoprotein K
H3K27me3: Histone 3 lysine 27 trimethylation
KD: Knock Down
KO: Knock Out
lncRNA: long non-coding RNA
Fbxw7: F-box and WD-40 domain protein 7
Hnrnpab: heterogeneous nuclear ribonucleoprotein A/B
Hnrnpc: heterogeneous nuclear ribonucleoprotein C
Hnrnpf: heterogeneous nuclear ribonucleoprotein F
Hnrnpk: heterogeneous nuclear ribonucleoprotein K
Hnrnpl : heterogeneous nuclear ribonucleoprotein L
Matr3: matrin 3
Pcbp2: poly(rC) binding protein 2
Ptbp1: polypyrimidine tract binding protein 1
Raly: hnRNP-associated with lethal yellow
Rbm3: RNA binding motif protein 3
Srsf3: serine/arginine-rich splicing factor 3
Srsf9: serine/arginine-rich splicing factor 9
Tardbp: TAR DNA binding protein
Hnrnpq: synaptotagmin binding, cytoplasmic RNA interacting protein
Myef2: myelin basic protein expression factor 2, repressor
Fubp3: far upstream element (FUSE) binding protein 3
Rbm15: RNA binding motif protein 15
Hnrnpa2b: heterogeneous nuclear ribonucleoprotein A2/B1
Hnrnpm: heterogeneous nuclear ribonucleoprotein M
Hnrnpu: heterogeneous nuclear ribonucleoprotein U
Lbr: lamin B receptor
Thoc2: THO complex 2
Celf1: CUGBP, Elav-like family member 1
Sf3b3: splicing factor 3b, subunit 3
Tcf7l1: transcription factor 7 like 1 (T cell specific, HMG box)
Spen: SPEN homolog, transcriptional regulator (Drosophila)
Wtap: Wilms tumour 1-associating protein
ncRNA: non-coding RNA
RAP-Seq: RNA Antisense Purification-Sequencing
RAP-MS: RNA Antisense-Mass Spectrometry
RepA: A repeats of Xist RNA
Rmb15:RNA binding motive protein 15
RNA-IP: RNA-immunoprecipitation
Rnf12/RLIM: Ring finger protein LIM domain interacting
i-CLIP: Individual nucleotide resolution Crosslinking and Immunoprecipitation
Pol II: RNA Polymerase II
PRC1/2: Polycomb Repressive Complex 1/2
SAF-A/hnrnpU: Scaffold attachment factor A/heterogeneous ribonucleoprotein U
SHARP/Spen: SMRT-and HDAC-associated Repressor Complex/Msx2-interacting protein
S/MAR: scaffold/matrix attachment region
SMRT/NCoR: Silencing Mediator for Retinoic Acid Receptor and Thyroid Hormone Receptor/Nuclear Receptor Co-Repressor
Xa: active X chromosome
XCI: X chromosome inactivation
Xi: inactive X chromosome
XIC: X Inactivation Center
Xist: Inactive X specific transcript

Introduction

Many non-coding transcripts carry out their functional roles by physically interacting with RNA-binding proteins (RBPs). A number of reports show that non-coding RNAs are tightly associated with various ribonucleoprotein complexes and chromatin regulators in order to target enzymatic activities to appropriate locations in the genome (1, 2). Accordingly, understanding how non-coding RNAs regulate gene expression requires investigation of protein-RNA complexes *in vivo*.

To date, most approaches to determine protein-RNA interactions exploit immunoprecipitation (3), which requires prior knowledge about which proteins might interact in order to test an interaction. For this reason, the direct protein interactions of most lncRNAs remains unknown.

Recently, work by several groups have developed unbiased mass spectrometry methods to comprehensively define the proteins that directly interact with a given lncRNA (4-6). These approaches were applied to the well-studied *Xist* lncRNA and uncovered many previously unknown proteins that have now been shown to be required for Xist-mediated transcriptional silencing (4-7). Although these approaches are powerful for defining direct interactions, they require significant resources – including significant time and cell numbers – in order to study each individual lncRNA.

Here we develop a novel computational method for defining direct RNA-protein interactions that exploits some important property of biochemistry of interactions, we call this approach *Global Score*. We show that this approach performs really well by gauging it again known Xist interactions (4-8). Our results show that there are 38 direct interactions with Xist and that we can validate many of these by eCLIP. Together, our approach provides a robust computational framework that enable identification of bona fide RNA-protein interactions that can be used for prioritized IP based follow-up.

Results

*cat*RAPID *Global Score*

The *cat*RAPID algorithm is extensively used in experimental works (9-11). Nonetheless, the method has the main limitation to be restricted to transcripts shorter than 1000 nt due to the complexity of their conformational space (12).

To identify proteins interacting with longer transcripts and especially lncRNAs, we implemented a new module called *Global Score*. The algorithm is based on the observation that binding sites are identifiable by fragmenting protein and RNA sequences (13, 14). As in the original *cat*RAPID method (15), the interaction propensity is calculated considering secondary structures, van der Waals' and hydrogen-bonding potentials. Using fragmentation, we reported accurate predictions for interactions involving FMRP, TDP-43, p53 and other proteins (12, 16), indicating that the procedure is particularly suitable to discriminate between binding and non-binding sites. The ability to predict both protein and RNA contacting regions makes *cat*RAPID a valid tool to complement CLIP experiments and identify protein regions involved in RNA recognition. Indeed, *cat*RAPID shows high performances when compared with CLIP and other high-throughput approaches (14).

Here we propose a method to integrate the signal coming from the binding propensities of fragments into a variable called *Global Score* that predicts the overall interaction ability of a protein-RNA pair. The introduction of the *Global Score* module allows us to compute interactions with large RNAs (>1000 nt), thus extending the general applicability of our approach (Methods*, Global Score*). Briefly, we calculate a total of 10^4 interactions are calculated, we weighted according to their interaction propensities and sum up into an overall score.

We trained the *Global Score* on 1500 ribonucleoprotein interactions detected by CLIP involving all the RNA-binding proteins reported in "The Atlas of UTR Regulatory Activity" (AURA, version 2014) (17). In a 5-fold cross-validation, we discriminated interacting and non-interacting protein-RNA pairs with an area

under the ROC curve (AUC) of 0.84 (Fig. 2). We performed an independent cross-validation on 800 interactions between transcripts longer than 1000 nt (Methods*, Global Score*) and protein partners identified by protein microarrays (18). The performances on the test set were considerably high (AUC=0.80; Fig. 2), indicating that *Global Score* can predict protein interactions with large RNAs with good accuracy. In our test set, we also used 50 proteins reported by recent studies to have RNA-binding ability but lacking canonical domains (19, 20). As *Global Score* correctly predicts that 85% of the non-canonical RBPs (i.e. 43 out of 50) bind to their RNA targets, we can conclude that the algorithm does not show particular preference for specific RBP classes.

Analysis of Xist interacting proteins

Recent publications created an unprecedented wealth of information on *Xist* interactions and functional players in XCI (4-8). While proteomic approaches (4-6) reveal proteins associating with *Xist*, they cannot directly differentiate between functional Xist-interactors and other cellular processes (such as RNA-processing and polyadenylation). In addition, proteomic analyses are deprived of proteins targeting RNA species that are used as reference controls in the experiments (21). By contrast, genetic screens select important regulators of XCI, but fail to provide information of direct protein interactions (7, 8). Often, due to their experimental set-up genetic screens are devoid of proteins interfering with other cellular functions (22).

About 350 candidates (including unpublished data; Material and Methods) have been reported in proteomic studies (Fig. 1). By contrast, about 50 proteins were identified through genetic screens. Yet, it should be mentioned that potentially-important low-ranking hits of genetic screen might be consequence of i) inefficient knock-down (i.e. Sh/SiRNA screening), ii) spatially limited integration sites (i.e., small genes in insertional mutagenesis screens), iii) depletion affecting cell viability or cell cycle control. Another important aspect to consider in genetic analyses is how complete are the screenings. Hits from the Monfort *et al.* screen are biased against short genes indicating that the screen is not complete (25). On the other hand, most of shRNAs were recovered in the work of Moindrot *et al.,* showing a higher degree of saturation.

In addition, published hits from Moindrot *et al.* (7) show better overlap with proteomic data over Monfort *et al.* (8). In this work, we decided to re-rank functional data by Moindrot *et al.* (7) and use them for our analysis [Fig. 1; Material and Methods; (7, 21)].

Having developed the *Global Score* method, we sought to determine which of these interactions are likely to be direct Xist interactions. We used the *Global Score* to measure the interaction strength of each protein predicted by the three proteomic studies (Fig. 2). Our predictions show that the two datasets by McHugh *et al.* (5) [published (I) and plus unpublished results (II) (Material and Methods)] are associated with the highest predictive power (Area under the ROC curve AUC > 0.9) followed by Chu *et al.* (4) (AUC=0.82) and Minajigi *et al.* (6) (AUC=0.75). In the calculations, we considered as negative controls those proteins reported by Minagiji *et al*. (6) that were depleted in the male vs female spectral counting [log(FC) < -0.75]. Accordingly, *Global Score* predicts them as non-interacting in ~80% of cases (Methods: *Global Score*). Thus, *cat*RAPID calculations suggest that the database by McHugh *et al.* (5) is the most enriched in direct targets.

Predictions of direct protein-RNA interactions

For the >600 Xist interacting proteins considered in this study, we observed that the *Global Score* values correlate with the lines of evidence supporting these interaction (Fig. 3A). Using values above the *Global Score* of Spen (also called SHARP), which is the only protein reported in all the experiments (*Global Score*=0.59; Fig. 3A dashed-line), we identified 58 candidates. Considering hits appearing in at least 2 datasets, we selected 38 proteins (Fig. 3B and Table 1) showing medium- (**) to high- (***) interaction propensities (Table 1). Notably, 29 out of 38 proteins have high-propensity (***) of interaction and 20 are associated with *Global Score* ≥0.95, which is highly significant with respect to the negative set (28 out 200 proteins have *Global Score* ≥0.95; p-value = 7.979e-07; Fisher's exact test) as well as proteomic (73 out of the 343 proteins; p-value = 8.123e-0.5; Fisher's exact test) and genomic (82 out 298 proteins; p-value = 0.0024; Fisher's exact test) datasets (Fig. 3B).

We note that Polycomb Repressive complex proteins PRC2 did not rank high in our analysis due to the fact that PRC2 elements were not over-represented in proteomic (4-6) or genetic screens (7, 8). Yet, PRC1 catalytic subunit Ring1B (also known as Rnf2) showed a high *cat*RAPID score (*Global Score* = 0.98, ***). Given the fact that this protein was found in only one of the three proteomic datasets and Chu *et al.* identified a number of non-direct binders by using formaldehyde fixation conditions (4), we excluded this candidate for further analysis.

GO analysis and network analysis of selected candidates

We then screened our candidates for cellular localization (i.e. direct interactors have to be nuclear), functional categories (i.e. RNA metabolism, gene-silencing), protein association network (i.e. STRING-network) and expression-level (i.e. expressed in early embryogenesis, when available; for details Methods and Tables S1-3).

GO analysis reveals that 21 out of the 38 candidates are part of the Hnrnp protein network (Fig. 4A). Importantly, Hnrnpu and Hnrnpk are crucial regulators of XCI: they are respectively necessary, for *Xist*-localization to chromatin (and, in turn, gene-silencing) (5, 23) as well as Polycomb recruitment (4). Our analysis indicates another sub-network between Rbm15/Spen and Rbm3, which is important for Ncor-complex recruitment to the inactive X (4, 5).

Almost all of the 38 candidates are in the RNA-related functional categories (35 out of 38 genes). We found functional associations with RNA-related processes, especially post-transcriptional regulation, splicing and nuclear export. The last category is particularly interesting as *Xist*, a poly-adenylated, spliced RNA never leaves the nucleus (24). A considerable fraction of our selected genes (20 out of 38; Fig. 4B and Table S4) cluster in the transcriptional regulation category. Other candidates are part of the silencing machinery [Ncor2 (Spen) and Hdac1 complex (Rbm14)] or are important for RNA processing and stabilization (Hnrnp-proteins) (25, 26). Three out of 38 genes are also part of the nuclear matrix (Lbr, Matr3, Hnrnpm), a sub-compartment that is involved in silencing and contacts *Xist* (27, 28).

To infer functional relationships among the selected candidates, we clustered the initial pool of 58 genes based on enriched GO terms of interactions (Supl. Table 3 and Methods *Gene ontology clustering*). Our analysis identified two major groups: one related to RNA splicing and transport, and another related to transcription regulation and protein degradation (Suppl. Fig. 4). The two classes contain genes that are important for *Xist* spreading and localization to the chromatin (Hnrnpu/Saf-A) (24) and are relevant for Xist localization to the nuclear lamina (Lbr) and may be relevant for Xist localization to the nucleolus (29).

Prediction of protein interactions at nucleotide resolution

We selected 5 representative genes for further investigation: Hnrnpu/Saf-A and Spen that have a role in *Xist* mediated silencing (Hnrnpu/Saf-A with *Global Score* = 0.66 **, and Spen with *Global Score* = 0.59 **,), Lbr and Hnrnpk that have been described to have a role in gene-silencing or Polycomb recruitment (Lbr with *Global Score* = 0.79, **, and Hnrnpk with *Global Score* = 0.99, ***), and Ptbp1 (*Global Score* = 0.99, ***) that associates with *Xist* but its role seems to be redundant in XCI establishement (4, 5).

Another powerful feature of the *Global Score* algorithm is the ability to define the protein domain and RNA binding sites that interact. We made use of this property to predict the binding sites of the 5 representative candidates (Fig. 5; Table S2 and S3). To determine what *Xist* regions are specifically contacted by each protein, we calculated interactions using fragments containing RNA-binding domains annotated in Gerstberger *et al*. 2014 (11) and NPIDB ('RNA' and 'hybrid' families) (30): Hnrnpk (P61979) 363-414 aa (KH domain); Hnrnpu/Saf-A (Q8VEK3) 1-52 aa (SAP domain); Ptbp1 (P17225) 76-127 aa (RRM domain); Spen (Q62504) 332-477 aa (RRM domain) (Methods: *Binding sites assessment*). In the case of Lbr, we used amino acids 51-102 that are predicted by our method to be most interactive (Methods: *Binding sites assessment;* Suppl. Fig. 2).

In agreement with previous work, Spen is predicted to interact with the *Xist* A-repeats, a region of *Xist* that is necessary for gene-silencing (31). Instead

Hnrnpk is predicted to bind to the B-repeats of *Xist* that have been also associated to Polycomb recruitment by Heard's lab (4, 32). Lbr and Ptbp1 are predicted to bind to the 3'-part of *Xist* corresponding to the E-repeats, while Hnrnpu/Saf-A is predicted to bind in multiple locations (Suppl. Fig. 3) (31).

Validation of *cat*RAPID predictions

We went on to map the binding sites of Spen, Hnrnpk, Lbr, Ptbp1 and Hnrnpu/Saf-A using an enhanced individual nucleotide CLIP method (eCLIP). As shown in Fig. 6, the five representative proteins bind to different regions: Spen and Hnrpk have propensity for the 5' of *Xist* (i.e., 1-5000 nt), Lbr and Ptbp1 show binding sites in the central region (i.e., 9000-13000 nt) and Hnrnrpu/Saf-A has a dispersed signal (Suppl. Fig. 3). Notably, *cat*RAPID does not predict binding sites in the 3' of *Xist*, which indicates marginal role in macromolecular recognition, as suggested by the poor sequence conservation of the region (33). Nevertheless, a recent paper suggested a role for Xist exon 7 in its localization on the chromatin (34).

Overall, eCLIP data are in very good agreement with the *cat*RAPID predictions: the portion of *Xist* predicted to interact with our selected proteins (highest *cat*RAPID score) was verified in all cases. Interestingly, eCLIP data confirm the specific interaction of Spen to *Xist* A-repeats and Hnrnpk to *Xist* B-repeat. Ptbp1 instead interacts, as predicted, to *Xist* E-repeats. Lbr interacts with a region downstream of *Xist* A-repeats, mostly around the E-repeat and a *Xist* 3'-end (Fig. 6 and Table 1). In all the cases, at least 1 out of 3 top-interacting regions are matched by experimental validation. *cat*RAPID, however, reports that Hnrnpu/Saf-A binds to a region in the central part of *Xist*, while eCLIP experiments reveal that the protein binds broadly across the whole transcript. The binding regions identified by Hasegawa *et al*. (nt 1899–3488 and nt 4725–6079) are included in our predictions and experimental validation (23). As Hnrnpu/Saf-A is mostly implicated in *Xist* localization onto the chromatin, it is possible that *Xist* interactions are non-specific to support *Xist* attachment to the nuclear matrix (23). Hnrnpu/Saf-A dispersed eCLIP profile may also be an artefact of cell-population sequencing experiments. i.e. Saf-A binds to *Xist* in

each cell but its binding site profile may differ between individual cells. As a consequence, the Saf-A eCLIP profile result broad/dispersed in cell population assays (24).

Discussion

In this study we introduced the *Global Score* method based on the *cat*RAPID *fragment* algorithm (12) to predict the interaction propensity of proteins that directly interact with a lncRNA. Our approach is based on the hypothesis that the information to detect protein interactions is contained in RNA domains identifiable by fragmentation of the molecule into sub-elements (13, 14). Previous estimates indicate that *cat*RAPID has an accuracy of 80% in predicting protein-RNA interactions (3), which is perfectly in-line with the results reported in this study. Using this approach, we explored the protein interactome of *Xist* through a multifaceted approach aiming to identify direct Xist binders and showed that we can recapitulate previous proteomic and genetic screens and can even further separate *bona fide* direct interactions across the five previous studies. Importantly, this approach correctly identifies the binding sites on RNA in 4/5 cases tests further highlighting the power of this approach. The sole exception (SAF-A) reflects the fact that the protein interacts with several regions of the *Xist* transcript - although, our predictions match the most highly interacting regions (Fig. 6).

Our approach will provide a powerful method for the lncRNA community because currently there are no straightforward methods for predicting likely protein interactions with a lncRNA – all current methods are time consuming and expensive and only provide partial information. Therefore, this computational method, which can be used on any lncRNA and protein set can provide a rapid platform for evaluating likely interactions for biochemical and functional followup.

Our eCLIP analysis refines our knowledge of the binding sites of *Xist*-interacting proteins (5, 23, 35). Indeed, we identified binding sites, at nucleotide resolution, showing which regions are important for Xist interaction

and function. Our results are in agreement with previous studies mapping Spen-Xist interaction to the Xist A-repeats, a key region for the establishment of Xist-mediated silencing (4, 31). The Ptbp1 eCLIP profile reveals that this protein may have a role in Xist spreading and localization to the chromatin, although its function may be redundant (5). HnrnpK eCLIP mapping is particularly in the light of HnrnpK role in Polycomb recruitment. The Xist-HnrnpK interacting region was previously mapped to between Xist repeats F and B (4). Jarid2, an important cofactor of Xist-mediated PRC2 recruitment, also interacts with the same region of Xist RNA (32). It is tempting to speculate that HnrnpK and Jarid2 may interact to recruit PRC2 on the inactive X chromosome and repB binding is essential for Polycomb recruitment. It is known that Xi localizes to the nuclear lamina (36). It is possible that Lbr mediates this interaction. In this case, we predicted the RNA-binding region to aa 51-102. This region largely overlaps with the RS domain, which has been implicated in nucleic acid recognition (37). As we predicted the region prone to interact with RNA without previous knowledge of the domains, we can conclude that our method can be used to predict novel RNA-binding domains for proteins with non-canonical RNA-binding regions.

HnrnpU/SafA's broad interaction is instead unexpected. Further studies are needed to understand whether only few regions of Xist are needed to sustain this interaction.

Intriguingly, our analysis identifies Fbxw7 and Tcf7l1 as novel potential Xist direct interactors selected from the genetic screens (7, 8). We do not currently know which proteins tether Xist to the nucleolus (29). It is tempting to speculate that proteins from the Ubiquitin-Proteasome pathways (UPS) may be involved in this process. A potential candidate is Fbxw7, which was independently found from the Brockdorff and Wutz's laboratories (7, 8). Fbwx7 is a component of the SCF (SKP1-CUL1-F-box protein) E3-ligase complex that is important for poly-ubiquitination of target substrates for subsequent proteasome degradation (38). However, as candidates coming from a Ubiquitin-Proteasome-System (UPS) in Moindrot *et al*. ranked low (7), this protein may have an indirect role in Xist mediated silencing.

We believe that the computational method presented here can be applied to the study of other lncRNAs. Our work paves new avenues for protein-RNA interaction studies as we can predict, with high accuracy, which regions of RNA are bound by proteins as well as the protein domains mediating the binding. As *cat*RAPID can be used as a tool to design mutations to any *lncRNA* and interacting proteins, a tempting possibility is to generate *Xist* deletion to uncouple Xist spreading from Polycomb recruitment, gene silencing or nuclear lamina/nucleolus tethering.

Materials and methods

Dataset ranking

In the genetic screening by Moindrot *et al*, (7) the effect of shRNAs targeting specific genes was calculated by dividing final counts ("sorted") over initial counts ("input"). The ratio of each individual shRNA was standardized by subtracting the median ratio of the dataset followed by division with median absolute deviation. The third highest standardized ratio of shRNAs targeting the same gene was used as score for the ranking. Thus, at least three individual shRNAs show higher or equal enrichment in counts were employed to assure consistent results and avoid off-targets shRNAs. The overlap between Moindroit *et al*. top-300 with joint proteomic datasets [Chu *et al*. (4). McHugh *et al.* (5) and Minajigi *et al.* (6)] is of 50 proteins. As the overlap between the genes list by Moindroit *et al.* (7) and proteomic datasets was only marginally increasing by considering lists of 500 or 1000 cases, we used the top 300 candidates in our analysis. By contrast, the overlap between the genes list by Monfort *et al.* (8) and proteomic datasets was of 5 candidates.

Database generation

In our study we integrated the results of two genetic [Moindrot *et al.* (7). Monfort *et al.* (8)] and three proteomic [Chu *et al*. (4). McHugh *et al.* (5) and Minajigi *et al.* (6)] screening.

Genetic screens comprise ~300 genes of which 8 ncRNAs (8 genes: Senp2, Prmt1, Dgkh, Fance, Zfp326, Ube2d2b, Snapc4, Ufd1l) from Moindrot *et al.* 2015; and 22 genes from Monfort *et al.* 2015. Proteomic screens comprise 81 genes from Chu et al. 2015; 1768 proteins (1765 genes: Actb, Parp1, Hmga1, Ppid/Ppif have duplicated entries) from Minajigi *et al*. 2015; and 20 genes from McHugh *et al.* 2015.

As for the datasets reported by Chu *et al*. (4). McHugh *et al.* (5) and Minajigi *et al.* (6), we used gene symbols to retrieve non-redundant sets of protein sequences through Uniprot (http://www.ebi.ac.uk/reference_proteomes). In the case of Moindrot *et al.* (7) and Minajigi *et al.,* we used the provided protein

identifiers (Moindrot RefSeq IDs were converted to UniProt IDs with 100% sequence similarity).

catRAPID

We used the *cat*RAPID *fragment* approach (12, 15) to identify putative binding sites between *Xist* and proteins and the *Global Score* algorithm to assess the overall interaction propensity.

In the *catRAPID* method, contributions of secondary structure, hydrogen bonding and van der Waals' are combined together into the *interaction profile*:

$$\vec{\Phi}_x = \alpha_H \vec{H}_x + \alpha_W \vec{W}_x + \alpha_S \vec{S}_x \quad (1)$$

where the variable x indicates RNA ($x = r$) or protein ($x = p$). The hydrogen bonding profile, denoted by $\vec{H}$, is the hydrogen bonding ability of each amino acid (or nucleotide) in a protein (or RNA) sequence:

$$\vec{H} = H_1, H_2, \dots, H_{length} \quad (2)$$

Similarly, $\vec{S}$ represents the secondary structure occupancy profile and $\vec{W}$ the van der Waals' profile. The *interaction propensity* π is defined as the product between the protein propensity profile $\vec{\Psi}_p$and the RNA propensity profile $\vec{\Psi}_r$ weighted by the *interaction matrix* I:

$$\pi = \vec{\Psi}_p \mathrm{I}\, \vec{\Psi}_r \quad (3)$$

The algorithm predicts the interaction propensity of a protein-RNA pair reporting the discriminative power, which is a measure of interaction strength with respect to the training sets.

Due to computational requirements, the *cat*RAPID graphic algorithm accepts only protein sequences with a length ranging between 50aa and 750aa and RNA sequences between 50nt and 1200nt (3). When the input sequences exceed the size compatible with our computational requirements, *catRAPID*

cannot be used to calculate the interaction propensity. To overcome this limitation, we developed a procedure called *fragmentation*, which involves division of polypeptide and nucleotide sequences into overlapping fragments followed by prediction of the interaction propensities (3). Following the procedure described in (12), the RNA fragment size employed in this study is 700 nt,

Global Score

To estimate the overall interaction potential of protein and RNA molecules using *uniform fragmentation*, we built the *Global Score* approach. To train the algorithm, we used the interactomes of RNA-binding proteins reported in AURA (AGO1, AGO2, AGO4, ELAVL1, QKI, PUM1, PUM2, TNRC6A, TNRC6B, TNRC6C, NCL, IGF2BP1, IGF2BP2, IGF2BP3 and ELAVL1) (17). We filtered out similar UTR sequences using CD-HIT [sequence identity > 80%] (39). To avoid biased training towards proteins with many RNA partners, we selected 50 UTRs for each protein to generate the positive binding dataset. The negative non-binding set was built shuffling the UTRs partners of the positive pool.

We computed protein-RNA interactions using *uniform fragmentation* (total of 750 positives and 750 negatives)(3). For each protein-RNA association, we clustered the interaction propensity scores π (Eq. 3) as follows

$$f_i = \vartheta(\pi - i)[1 - \vartheta(\pi - i - 1)] \quad (4)$$

where $\vartheta(x)$ is the Heaviside function that is 1 if $x > 0$ and zero otherwise. The values f_i are weighted to norm 1:

$$F_i = f_i / \sum_{i=min}^{max} f_i \quad (5)$$

where $min = -50$ and $max = 50$. To determine the relative contribution F_i of fragments, we compute h_k:

$$h_k = tanh\left(\omega_k^i F_i\right) \quad (6)$$

where $tanh(\mathrm{x})$ is the hyperbolic tangent of x. The global score Π is evaluated using h_k:

$$\Pi = tanh\left(\Omega^{k} h_{k}\right) \quad (7)$$

The weights ω_k^i and Ω^k have been determined by optimizing the match between experimental and predicted interactions (same number of positive and negative cases). To avoid over-fitting, we varied the number of internal weights proportionally to the size of the training set and performed a 5-fold cross-validation at each optimization. For $i = 100$ and $k = 10$, we obtained an AUC of 0.84 (Fig. 2A) in discriminating interacting and non-interacting protein-RNA pairs.

We performed an independent cross-validation using 8 transcripts (*Myc, Bcl2, Igf2rnc, Pwrn1, Sox2oy, lincRBM26, Occ1* and *Tp53*) > 1200 nucleotides whose binding partners have been determined through protein microarrays technology (8). For each RNA molecule, we selected 50 top-ranked (i.e., high-affinity) and 50 bottom-ranked (i.e., low-affinity) proteins and used *cat*RAPID *fragment* and the *Global Score* algorithm to classify. Also on this test set, the performances were high (AUC=0.80; Figure 2A).

The algorithms to compute protein-RNA interactions are available at our group webpage http://service.tartaglialab.com/page/catrapid_group and the new algorithm *Global Score* can be accessed at http://service.tartaglialab.com/new_submission/catrapid_fragments_ultra [upon acceptance of the paper, the link will replace previous web-address http://service.tartaglialab.com/new_submission/catrapid_fragments].

Binding sites predictions

To visualize *Xist* binding sites, *cat*RAPID scores were Z-normalized using interaction propensities calculated on proteins associated with poor spectral counts in the study by Minajigi *et al.* [200 proteins with log2(FC) < –0.75] (6). Notably, *cat*RAPID predicts 80% of these proteins as non-interacting.

To determine what *Xist* regions are specifically contacted by protein candidates, we selected fragments containing RNA-binding domains retrieved from Gerstberger *et al*. 2014 (11). NPIDB ['RNA' and 'hybrid' families; update September 2015] (30). *Xist*-protein interactions were ranked (Fig. 3 and Table S3) to identify high-confidence regions. As for Lbr, our approach identifies amino acids 51-102 as the most prone to interact with RNA (Suppl. Fig. 2)

To localise high-confidence binding sites, we calculated the coordinates of the highest-scoring regions (top 2%) and filtered out fragments falling outside the resolution of our approach (average distance > 5 times the size of the RNA fragment). We observe that Hnrnpu/Saf-A has the largest signal dispersion (Suppl. Fig. 3), which suggests that binding is non-specific, as revealed by eCLIP experiments (Fig. 6). For this case, the filter on *cat*RAPID resolution has been removed (Fig. 6)

The *Global Score* is used to divide proteins in three groups of 1000 entries: low-affinity interactions (*Global Score* < 0.02; one-star *), medium-affinity interactions (0.02 < *Global Score* < 0.80; two-stars **), and high-affinity (*Global Score* > 0.80; three-stars ***).

Binding sites assessment

Predicted binding regions of Hnrnpk, Hnrnpu/Saf-A, Lbr, Ptbp1, and Spen have been assessed with eCLIP data. Highest scoring associations falling within experimentally validated binding sites (>50% coverage) are listed: Hnrnpk (P61979) 363-414 aa (KH domain) interacting with Xist 2507-3224 nt (0.98 percentile, Fig. 5); Hnrnpu/Saf-A (Q8VEK3) 1-52 aa (SAP domain) with Xist 3956-4673 nt (0.99 percentile, Fig. 5); Lbr (Q3U9G9) 51-102 aa (most interacting Lbr fragment, Suppl. Fig. 2) with 10025-10742 nt (0.98 percentile, Fig. 5); Ptbp1 (P17225) 76-127 aa (RRM domain) with Xist 10741-11458 nt (0.99 percentile, Fig. 5); Spen (Q62504) 332-477 aa (RRM domain) with Xist 18-735 (0.98 percentile, Fig. 5).

Interaction network

The network of protein-protein interactions among 40 candidate genes has been constructed using STRING database (40) with several confidence scores (high confidence score of 0.70 to highest confidence score of 0.90). Most of those interactions are reported with highest confidence score of 0.90. Interactions among Spen, Rbm15 and Rbm3 have been manually curated (4, 5).

Gene ontology clustering

We clustered candidate genes using functional macro-categories of interest ("Chromatin remodeling", "Nuclear matrix and envelop", "RNA processing and splicing", "Transcription regulation"). Gene Ontologies (GO) terms (PMID: 10802651) are assigned to a macro-category querying their definitions using keywords (i.e. the words in the macro-category).

In order to infer functional relationship among 58 candidate genes, we downloaded their interactors from STRING (highest confidence score of 0.9) and compute GO term enrichment (Dunn–Šidák correction for multiple testing). Based on enriched GO terms, we compute Jaccard index to built a similarity matrix to be used to cluster candidate genes (hierarchical clustering, Ward's method). Optimal cluster number was estimated using the Calinski-Harabasz criterion. We associated each cluster to the top 3 unique most enriched GO terms. All computations were performed using R statistical environment.

eCLIP experiments

We crosslinked 6 hours doxycycline-induced pSM33 mouse male ES cells with 0.4J of UV254.Cells were lysed in 1 ml lysis buffer (50 mM Tris pH 7.5, 100mM NaCl, 1% NP-40, 0.5% Sodium Deoxycholate, 1x Protease inhibitor cocktail). RNA was digested with Ambion RNase I (1:4000 dilution) to achieve a size range of 100-500 nucleotides in length. Lysate preparations were precleared by mixing with Protein G beads for 1hr at 4C. Target proteins were immunoprecipitated from 5 million cells with 10 ug of antibody and 75 ul of Protein G beads in 100uL lysis buffer. The antibodies were pre-coupled to the

beads for 1 hr at room temperature with mixing before incubating the precleared lysate to the beads-antibody overnight at 4C. After the immunoprecipitation, the beads were washed four times with High salt wash buffer (50 mM Tris-HCl pH 7.4, 1 M NaCl, 1 mM EDTA, 1% NP-40, 0.1% SDS, 0.5% sodium deoxycholate) and four times with Wash buffer (20 mM Tris-HCl pH 7.4, 10 mM MgCl2, 0.2% Tween-20). RNAs were then eluted by incubating at 50C in NLS elution buffer (20 mM Tris-HCl pH 7.5, 10 mM EDTA, 2% N-lauroylsacrosine, 2.5 mM TCEP) supplemented with 100 mM DTT for 20 minutes. Samples were then run through a standard SDS-PAGE gel and transferred to a nitrocellulose membrane, and a region 75 kDa above the molecular size of the protein of interest was isolated and treated with Proteinase K (NEB) followed by buffer exchange and concentration with RNA Clean & Concentrator™-5 (Zymo). We then made sequencing libraries from these samples as previously described in Engreitz et al. 2014 (*Cell*, doi: 10.1016/j.cell.2014.08.018) and Shishkin et al. 2015 (*Nature Methods*, doi: 10.1038/nmeth.3313).

Additional annotations

Cellular localization information (Tables 1, S2) was retrieved from UniProt (41) and LOCATE (42) (*experimental* evidence) databases. Expression levels in ES-E14 cell line and E18 mouse (Central Nervous System) were retrieved from ENCODE (43) RNA-seq data averaging RPKMs of replicates with IDR<0.1.

Authors contributions

AC, MG and GGT conceived this study. AC, GGT, DC and MG wrote the paper. DC performed the *in silico* work, MB performed the experimental validation of selected candidates. AB performed the ranking and the statistical analysis on selected datasets.

Acknowledgements

We want to thank Deanne Whitworth and Kristina Havas for critical reading of the manuscript. 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 (Gian Gaetano Tartaglia), and from the Spanish Ministry of Economy and Competitiveness (BFU2014-55054-P). We also acknowledge support from AGAUR (2014 SGR 00685), the Spanish Ministry of Economy and Competitiveness, 'Centro de Excelencia Severo Ochoa 2013–2017' (SEV-2012-0208). We thank the EMBL grant to PA.

References

1. Rinn JL & Chang HY (2012) Genome regulation by long noncoding RNAs. *Annual review of biochemistry* 81:145-166.
2. Wang KC & Chang HY (2011) Molecular mechanisms of long noncoding RNAs. *Molecular cell* 43(6):904-914.
3. Cirillo D, Livi CM, Agostini F, & Tartaglia GG (2014) Discovery of protein-RNA networks. *Molecular bioSystems* 10(7):1632-1642.
4. Chu C, *et al.* (2015) Systematic discovery of xist RNA binding proteins. *Cell* 161(2):404-416.
5. McHugh CA, *et al.* (2015) The Xist lncRNA interacts directly with SHARP to silence transcription through HDAC3. *Nature* 521(7551):232-236.
6. Minajigi A, *et al.* (2015) Chromosomes. A comprehensive Xist interactome reveals cohesin repulsion and an RNA-directed chromosome conformation. *Science* 349(6245).
7. Moindrot B, *et al.* (2015) A Pooled shRNA Screen Identifies Rbm15, Spen, and Wtap as Factors Required for Xist RNA-Mediated Silencing. *Cell reports* 12(4):562-572.
8. Monfort A, *et al.* (2015) Identification of Spen as a Crucial Factor for Xist Function through Forward Genetic Screening in Haploid Embryonic Stem Cells. *Cell reports* 12(4):554-561.
9. Ng SY, Johnson R, & Stanton LW (2012) Human long non-coding RNAs promote pluripotency and neuronal differentiation by association with chromatin modifiers and transcription factors. *The EMBO journal* 31(3):522-533.
10. Yan B, *et al.* (2014) Aberrant expression of long noncoding RNAs in early diabetic retinopathy. *Investigative ophthalmology & visual science* 55(2):941-951.
11. Gerstberger S, Hafner M, & Tuschl T (2014) A census of human RNA-binding proteins. *Nature reviews. Genetics* 15(12):829-845.
12. Cirillo D, *et al.* (2013) Neurodegenerative diseases: quantitative predictions of protein-RNA interactions. *RNA* 19(2):129-140.
13. Somarowthu S, *et al.* (2015) HOTAIR forms an intricate and modular secondary structure. *Molecular cell* 58(2):353-361.
14. Agostini F, *et al.* (2013) catRAPID omics: a web server for large-scale prediction of protein-RNA interactions. *Bioinformatics* 29(22):2928-2930.
15. Bellucci M, Agostini F, Masin M, & Tartaglia GG (2011) Predicting protein associations with long noncoding RNAs. *Nature methods* 8(6):444-445.
16. Zanzoni A, *et al.* (2013) Principles of self-organization in biological pathways: a hypothesis on the autogenous association of alpha-synuclein. *Nucleic acids research* 41(22):9987-9998.
17. Dassi E, *et al.* (2012) AURA: Atlas of UTR Regulatory Activity. *Bioinformatics* 28(1):142-144.
18. Siprashvili Z, *et al.* (2012) Identification of proteins binding coding and non-coding human RNAs using protein microarrays. *BMC genomics* 13:633.

19. Baltz AG, *et al.* (2012) The mRNA-bound proteome and its global occupancy profile on protein-coding transcripts. *Molecular cell* 46(5):674-690.
20. Castello A, *et al.* (2012) Insights into RNA biology from an atlas of mammalian mRNA-binding proteins. *Cell* 149(6):1393-1406.
21. Moindrot B & Brockdorff N (2016) RNA binding proteins implicated in Xist-mediated chromosome silencing. *Seminars in cell & developmental biology*.
22. Grimm S (2004) The art and design of genetic screens: mammalian culture cells. *Nature reviews. Genetics* 5(3):179-189.
23. Hasegawa Y, *et al.* (2010) The matrix protein hnRNP U is required for chromosomal localization of Xist RNA. *Developmental cell* 19(3):469-476.
24. Cerase A, Pintacuda G, Tattermusch A, & Avner P (2015) Xist localization and function: new insights from multiple levels. *Genome biology* 16:166.
25. Coelho MB, *et al.* (2015) Nuclear matrix protein Matrin3 regulates alternative splicing and forms overlapping regulatory networks with PTB. *The EMBO journal* 34(5):653-668.
26. Salton M, *et al.* (2011) Matrin 3 binds and stabilizes mRNA. *PloS one* 6(8):e23882.
27. Linnemann AK & Krawetz SA (2009) Silencing by nuclear matrix attachment distinguishes cell-type specificity: association with increased proliferation capacity. *Nucleic acids research* 37(9):2779-2788.
28. Stauffer DR, Howard TL, Nyun T, & Hollenberg SM (2001) CHMP1 is a novel nuclear matrix protein affecting chromatin structure and cell-cycle progression. *Journal of cell science* 114(Pt 13):2383-2393.
29. Zhang LF, Huynh KD, & Lee JT (2007) Perinucleolar targeting of the inactive X during S phase: evidence for a role in the maintenance of silencing. *Cell* 129(4):693-706.
30. Kirsanov DD, *et al.* (2013) NPIDB: Nucleic acid-Protein Interaction DataBase. *Nucleic acids research* 41(Database issue):D517-523.
31. Wutz A, Rasmussen TP, & Jaenisch R (2002) Chromosomal silencing and localization are mediated by different domains of Xist RNA. *Nature genetics* 30(2):167-174.
32. da Rocha ST, *et al.* (2014) Jarid2 Is Implicated in the Initial Xist-Induced Targeting of PRC2 to the Inactive X Chromosome. *Molecular cell* 53(2):301-316.
33. Nesterova TB, *et al.* (2001) Characterization of the genomic Xist locus in rodents reveals conservation of overall gene structure and tandem repeats but rapid evolution of unique sequence. *Genome research* 11(5):833-849.
34. Yamada N, *et al.* (2015) Xist Exon 7 Contributes to the Stable Localization of Xist RNA on the Inactive X-Chromosome. *PLoS genetics* 11(8):e1005430.
35. Chow JC, *et al.* (2010) LINE-1 activity in facultative heterochromatin formation during X chromosome inactivation. *Cell* 141(6):956-969.
36. Yang F, *et al.* (2015) The lncRNA Firre anchors the inactive X chromosome to the nucleolus by binding CTCF and maintains H3K27me3 methylation. *Genome biology* 16:52.
37. Takano M, *et al.* (2002) The binding of lamin B receptor to chromatin is regulated by phosphorylation in the RS region. *European journal of biochemistry / FEBS* 269(3):943-953.

38. Welcker M & Clurman BE (2005) The SV40 large T antigen contains a decoy phosphodegron that mediates its interactions with Fbw7/hCdc4. *The Journal of biological chemistry* 280(9):7654-7658.
39. Huang Y, Niu B, Gao Y, Fu L, & Li W (2010) CD-HIT Suite: a web server for clustering and comparing biological sequences. *Bioinformatics* 26(5):680-682.
40. Szklarczyk D*, et al.* (2015) STRING v10: protein-protein interaction networks, integrated over the tree of life. *Nucleic acids research* 43(Database issue):D447-452.
41. UniProt C (2015) UniProt: a hub for protein information. *Nucleic acids research* 43(Database issue):D204-212.
42. Sprenger J*, et al.* (2008) LOCATE: a mammalian protein subcellular localization database. *Nucleic acids research* 36(Database issue):D230-233.
43. Consortium EP (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature* 489(7414):57-74.

Name	Ensembl ID	Full name	Datasets	GlobalScore	Cellula Localization	Xist interacting region
Rbm3	ENSMUSG00000031167	RNA binding motif protein 3	Minajigi;McHugh;Chu	1***	nucleus;cytoplasm;cell projection:dendrite	7178-7895;11815-12532;7161-7878
Hnrnpk	ENSMUSG00000021546	heterogeneous nuclear ribonucleoprotein K	Minajigi;McHugh;Chu	0,999984***	cytoplasm;nucleus:nucleoplasm;cell projection:podosome	4655-5372;3956-4673;2507-3224
Rbm14	ENSMUSG00000006456	RNA binding motif protein 14	Minajigi;Chu	0,999939***	nucleus;nucleus:nucleolus	10758-11475;10741-11458;16469-17186
Srsf7	ENSMUSG00000024097	serine/arginine-rich splicing factor 7	Minajigi;Chu	0,999696***	nucleus;cytoplasm	11116-11833;10741-11458;10042-10759
Srsf10	ENSMUSG00000028676	serine/arginine-rich splicing factor 10	Minajigi;Chu	0,999097***	nucleus speckle;nucleus	5080-5747;5013-5730;10741-11458
Hnrnpab	ENSMUSG00000020358	heterogeneous nuclear ribonucleoprotein A/B	Minajigi;McHugh;Chu	0,998922***	nucleus;cytoplasm	10042-10759;12548-13265;7178-7895
Ptbp1	ENSMUSG00000006498	polypyrimidine tract binding protein 1	Moindrot;Minajigi;McHugh;Chu	0,998363***	nucleus¶	10400-11117;7178-7895;12173-12890
Hnrnpc	ENSMUSG00000060373	heterogeneous nuclear ribonucleoprotein C	Moindrot;Minajigi;McHugh;Chu	0,998289***	nucleus	11116-11833;12173-12890;10383-11100
Raly	ENSMUSG00000027593	hnRNP-associated with lethal yellow	Minajigi;McHugh;Chu	0,99797***	nucleus	12173-12890;7178-7895;11099-11816
Fbxw7	ENSMUSG00000028086	F-box and WD-40 domain protein 7	Moindrot;Monfort	0,994495***	nucleus:nucleoplasm;nucleus;cytoplasm;endoplasmic reticulum;golgi apparatus	12173-12890;10741-11458;9326-10043
Myef2	ENSMUSG00000027201	myelin basic protein expression factor 2 repressor	Minajigi;McHugh;Chu	0,993986***	nucleus;nuclear	11099-11816;8951-9668;7178-7895
Pcbp2	ENSMUSG00000056851	poly(rC) binding protein 2	Moindrot;Minajigi	0,991756***	nucleus;cytoplasm	10383-11100;10741-11458;9309-10026
Matr3	ENSMUSG00000037236	matrin 3	Moindrot;Minajigi;Chu	0,987189***	nucleus	10741-11458;12548-13265;10758-11475
Hnrnpd	ENSMUSG00000000568	heterogeneous nuclear ribonucleoprotein D	Minajigi;Chu	0,984306***	nucleus;cytoplasm	10025-10742;7178-7895;11116-11833
Hnrnpa0	ENSMUSG00000007836	heterogeneous nuclear ribonucleoprotein A0	Moindrot;Minajigi;Chu	0,980204***	nucleus	11116-11833;10025-10742;2865-3582
Fubp3	ENSMUSG00000026843	far upstream element (FUSE) binding protein 3	Minajigi;Chu	0,978294***	nucleus¶;cytoplasm¶	10383-11100;12173-12890;7178-7895
Srsf3	ENSMUSG00000071172	serine/arginine-rich splicing factor 3	Minajigi;McHugh;Chu	0,968346***	nucleus;cytoplasm	12190-12907;11099-11816;11116-11833
Hnrnpf	ENSMUSG00000042779	heterogeneous nuclear ribonucleoprotein F	Minajigi;McHugh	0,964498***	nucleus:nucleoplasm;nucleus	15412-16129;12548-13265;13980-14697
Khsrp	ENSMUSG00000007670	KH-type splicing regulatory protein	Minajigi;Chu	0,960451***	nucleus;cytoplasm	5729-6446;5746-6463;7178-7895
Khdrbs1	ENSMUSG00000028790	KH domain containing RNA binding signal transduction associated 1	Minajigi;Chu	0,953148***	nucleus;membrane;nuclear	5388-6105;5371-6088;4297-5014
Hnrnpq	ENSMUSG00000032423	synaptotagmin binding cytoplasmic RNA interacting protein	Minajigi;McHugh;Chu	0,939472***	nucleus;cytoplasm	9309-10026;10758-11475;12548-13265
Srsf9	ENSMUSG00000029538	serine/arginine-rich splicing factor 9	Minajigi;McHugh;Chu	0,936926***	nucleus	8593-9310;376-1093;8610-9327
Erh	ENSMUSG00000021131	enhancer of rudimentary homolog (Drosophila)	Minajigi;Chu	0,933699***	membrane¶;midbody¶	7178-7895;11099-11816;12173-12890
Rbfox2	ENSMUSG00000033565	RNA binding protein fox-1 homolog (C. elegans) 2	Minajigi;Chu	0,931653***	nucleus;cytoplasm	7178-7895;4297-5014;6820-7537
Rbm15	ENSMUSG00000048109	RNA binding motif protein 15	Moindrot;Minajigi;McHugh;Chu	0,892984***	nuclear speck¶;nucleoplasm¶;nucleus¶	6820-7537;10400-11117;1808-2525
Nxf1	ENSMUSG00000010097	nuclear RNA export factor 1	Moindrot;Minajigi	0,892436***	nucleus:nucleoplasm;nucleus speckle;cytoplasm;nucleus	7178-7895;10741-11458;5030-5747
Preb	ENSMUSG00000045302	prolactin regulatory element binding	Moindrot;Minajigi	0,888005***	nucleus;membrane;single-pass membrane protein;endoplasmic reticulum;nuclear;cytoplasmic	4297-5014;4655-5372;9309-10026
Hnrnpl	ENSMUSG00000015165	heterogeneous nuclear ribonucleoprotein L	Moindrot;Minajigi;Chu	0,879665***	nucleus:nucleoplasm;cytoplasm;nucleus	376-1093;10741-11458;10042-10759
Tardbp	ENSMUSG00000041459	TAR DNA binding protein	Minajigi;McHugh;Chu	0,875843***	nucleus	12548-13265;12173-12890;7178-7895
Lbr	ENSMUSG00000044880	lamin B receptor	Minajigi;McHugh	0,790350**	nucleus inner membrane;multi-pass membrane protein;membrane associated;nuclear;nucleus	12190-12907;10025-10742
Hnrnpa2b1	ENSMUSG00000004980	heterogeneous nuclear ribonucleoprotein A2/B1	Minajigi;McHugh;Chu	0,769678**	nucleus:nucleoplasm;cytoplasm;nucleus	9309-10026;10758-11475;16469-17186
Celf1	ENSMUSG00000005506	CUGBP Elav-like family member 1	Minajigi;McHugh;Chu	0,751637**	nucleus;cytoplasm	11116-11833;10383-11100;7519-8236
Hnrnpm	ENSMUSG00000059208	heterogeneous nuclear ribonucleoprotein M	Minajigi;McHugh;Chu	0,750406**	nucleus	5729-6446;9309-10026;4655-5372
Tcf7l1	ENSMUSG00000055799	transcription factor 7 like 1 (T cell specific HMG box)	Moindrot;Monfort	0,690085**	nucleus¶	12548-13265;11116-11833;5746-6463
Hnrnpr	ENSMUSG00000066037	heterogeneous nuclear ribonucleoprotein R	Minajigi;Chu	0,694462**	nucleus	10383-11100;8951-9668;12548-13265
Hnrnpu	ENSMUSG00000039630	heterogeneous nuclear ribonucleoprotein U	Minajigi;McHugh;Chu	0,663432**	nucleus	10400-11117;10741-11458;12548-13265
Safb	ENSMUSG00000071054	scaffold attachment factor B	Minajigi;Chu	0,605143**	nucleus¶	12548-13265;10741-11458;10025-10742
Spen	ENSMUSG00000040761	SPEN homolog transcriptional regulator (Drosophila)	Moindrot;Monfort;Minajigi;McHugh;Chu	0,592912**	mitochondrial inner membrane;nucleus	1808-2525;18-735;1-718
Below CatRapis threshold						
Thoc2	ENSMUSG00000037475	THO complex 2	Moindrot;Minajigi	0,552522**	nucleus	12531-13248;4655-5372;13980-14697
Med25	ENSMUSG00000002968	mediator complex subunit 25	Moindrot;Monfort	0,542981**	nucleus;nuclear	8610-9327;5746-6463;13980-14697
Ptbp2	ENSMUSG00000028134	polypyrimidine tract binding protein 2	Minajigi;Chu	0,311822**	nucleus	14338-15055;10741-11458;15412-16129
Hnrnpa3	ENSMUSG00000059005	heterogeneous nuclear ribonucleoprotein A3	Minajigi;Chu	0,289574**	nucleus	4297-5014;6820-7537;10741-11458
Ddx5	ENSMUSG00000020719	DEAD (Asp-Glu-Ala-Asp) box polypeptide 5	Minajigi;Chu	0,288154**	nucleus:nucleolus;nucleus	5729-6446;11116-11833;10025-10742
Eif4a3	ENSMUSG00000025580	eukaryotic translation initiation factor 4A3	Minajigi;Chu	0,160177**	nucleus;nucleus speckle;cytoplasm	4297-5014;10383-11100;10741-11458
Zfr	ENSMUSG00000022201	zinc finger RNA binding protein	Minajigi;Chu	0,125391**	nucleus	12190-12907;9309-10026;5746-6463
Elavl1	ENSMUSG00000040028	ELAV (embryonic lethal abnormal vision)-like 1 (Hu antigen R)	Minajigi;Chu	0,114546**	cytoplasm;nucleus	10383-11100;7178-7895;4297-5014
Hnrnpul2	ENSMUSG00000071659	heterogeneous nuclear ribonucleoprotein U-like 2	Minajigi;Chu	0,085841**	nucleus	5729-6446;5371-6088;5388-6105
Xrn2	ENSMUSG00000027433	5'-3' exoribonuclease 2	Minajigi;Chu	0,041415**	nucleus	4297-5014;10741-11458;5030-5747
Sap18	ENSMUSG00000021963	Sin3-associated polypeptide 18	Moindrot;Chu	0,024243**	nucleus¶;cytoplasm¶;nuclus speckle¶	5371-6088;5388-6105;7178-7895
Wtap	ENSMUSG00000060475	Wilms tumour 1-associating protein	Moindrot;Chu	0,004528*5	nucleus:nucleoplasm;nucleus speckle;nuclear;cytoplasmic;nucleus	12531-13248;16469-17186;10741-11458
Dhx9	ENSMUSG00000042699	DEAH (Asp-Glu-Ala-His) box polypeptide 9	Minajigi;Chu	0,002593*	nucleus	5729-6446;7519-8236;717-1434
Mrps23	ENSMUSG00000023723	mitochondrial ribosomal protein S23	Moindrot;Minajigi	0,001270*5	mitochondrion;mitochondrial inner membrane	6820-7537;10383-11100;10758-11475
Ssb	ENSMUSG00000068882	Sjogren syndrome antigen B	Minajigi;Chu	0,000366*5	nucleus	11815-12532;11116-11833;6820-7537
Klf4	ENSMUSG00000034311	kinesin family member 4	Moindrot;Minajigi	0,000139*	nucleus	6820-7537;5729-6446
Msh2	ENSMUSG00000024151	mutS homolog 2 (E. coli)	Moindrot;Minajigi	9,05E-05*	nucleus	10383-11100;11815-12532;11116-11833
Mybbp1a	ENSMUSG00000040463	MYB binding protein (P160) 1a	Minajigi;Chu	6,05E-05*	nucleus	11815-12532;5729-6446;8610-9327
Mcm3	ENSMUSG00000041859	minichromosome maintenance deficient 3 (S. cerevisiae)	Moindrot;Minajigi	5,60E-05*	nuclear;nucleus	13980-14697;9309-10026;5729-6446
Smarcad1	ENSMUSG00000029920	SWI/SNF-related matrix-associated actin-dependent regulator of chromatin subfamily a containing DEAD/H box 1	Moindrot;Monfort	1,85E-05*	nucleus	10042-10759;11116-11833;7178-7895

¶ manually cured annotation

Table 1. *Annotations of 58 putative candidates of XCI.*

	color code	experimental approach	murine cell line	reference
proteomic assays		ChIRP-MS	ESCs	Chu et al. *Cell* 2015
		iDRiP	MEFs	Minajigi et al. *Science* 2015
		RAP-MS	ESCs	McHugh et al. *Nature* 2015
genetic assays		genetic screen	ESCs	Monfort et al. *Cell Reports* 2015
		shRNA screen	ESCs	Moindrot et al. *Cell Reports* 2015

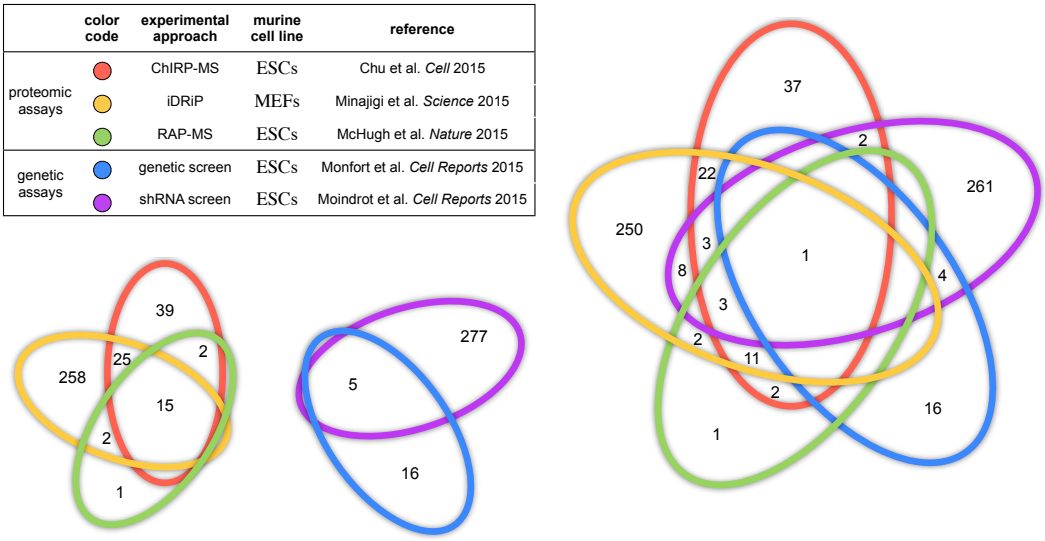


Figure 1. *Intersections of genes reported in 5 studies of Xist interactomes*. Five laboratories investigated *Xist* interactions: three groups used biochemical approaches to identify *Xist* interactors (4-6), and two others used a genetic strategy to reveal *Xist* functional partners that mediating gene-silencing (7, 8). Only one protein, Spen, has been found in all the assays. 58 genes are present in at least two assays and 17 candidates are in common in at least one proteomic or one genetic assay.

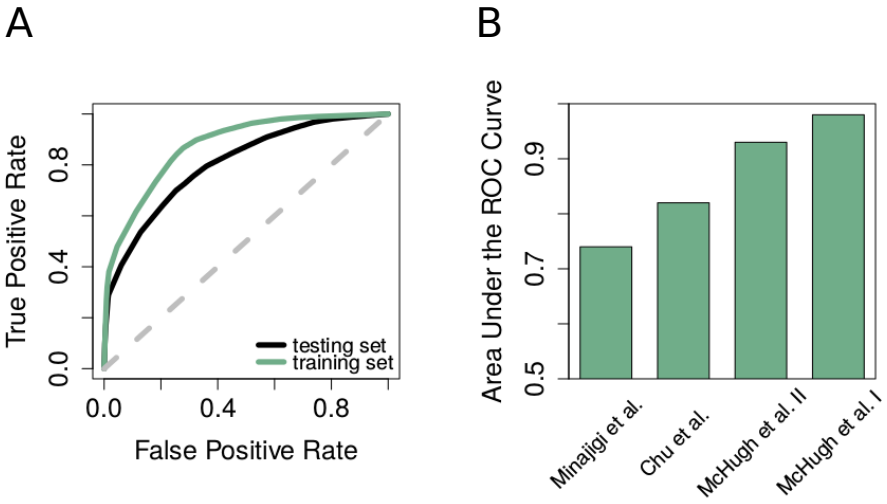


Figure 2. *Performances of Global Score algorithm* A) Receiver Operating Characteristic (ROC) curves of training and testing sets: In the 5-fold cross-validation, we discriminated interacting and non-interacting protein-RNA pairs with an area under the ROC curve (AUC) of 0.84. On the test set, performances were comparable to those of the training set (AUC=0.80). B) Area Under the ROC curve of proteomic assays [Minajigi *et al.* (6), Chu *et al.* (4), McHugh *et al.* II [ranked 11-20 in the publication (5)], McHugh et al. (ranked 1-10 in the publication (5))].

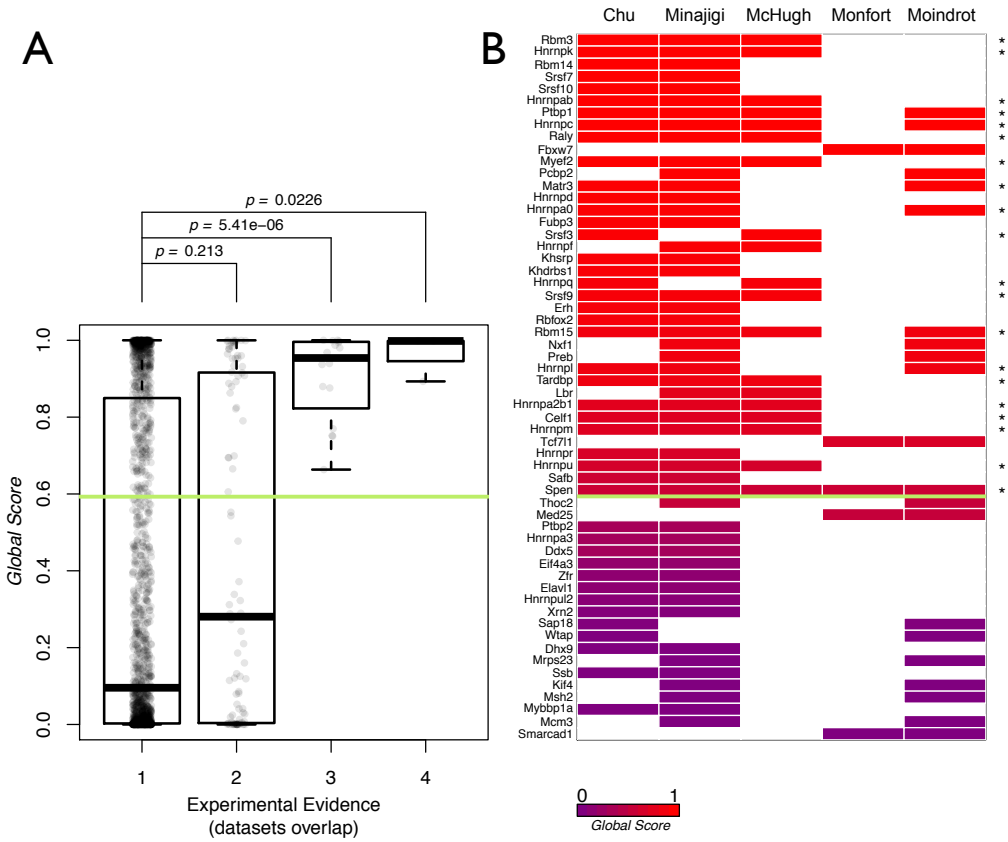


Figure 3. *Selection of Xist-interacting proteins* A) *Global Score* distribution of protein groups classified by experimental evidence. Predicted interaction propensities correlate with lines of experimental evidence (Wilcoxon signed-rank test). The green line indicates the only experimental case reported in all the screenings (Spen; *Global Score* = 0.59). B) List of candidate proteins analysed in this study. We identified 38 proteins associated with at least two lines of evidence and *Global Score* > 0.59. Above *Global Score* > 0.59, 20 proteins experiments (highlighted with a star on the right) appear in three or four experiments.

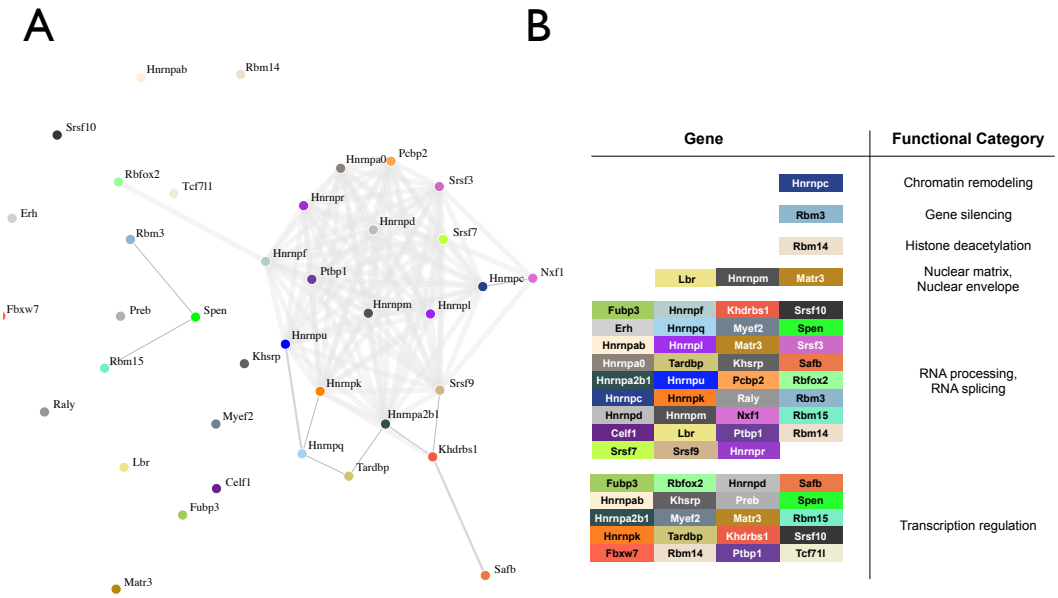


Figure 4. *Network and functional analysis of candidate mediators of XCI Protein-protein*. A) Interaction networks of 38 candidate factors (Methods: Interaction network). B) Functional categories associated with candidate mediators of XCI (Methods: Gene ontology clustering): Chromatin remodeling, Nuclear matrix/envelop, RNA processing/splicing, and Transcription regulation. Twenty out of 38 genes cluster in the transcriptional regulation category. Ncor2 (Spen), Hdac1 complex (Rmb14) are part of the splicing machinery and Matr3 is important for *Xist* processing and stabilization (25, 26). Three out of 38 genes are also part of the nuclear matrix (Lbr, Matr3, Hnrnpm), a nuclear sub-compartment that is important for gene silencing and is known to contact *Xist.* In the plot, grey lines connect interacting proteins (grey line with is proportional to STRING confidence score: thick lines indicate a confidence score of 0.9; thin lines indicate a confidence score of 0.7).

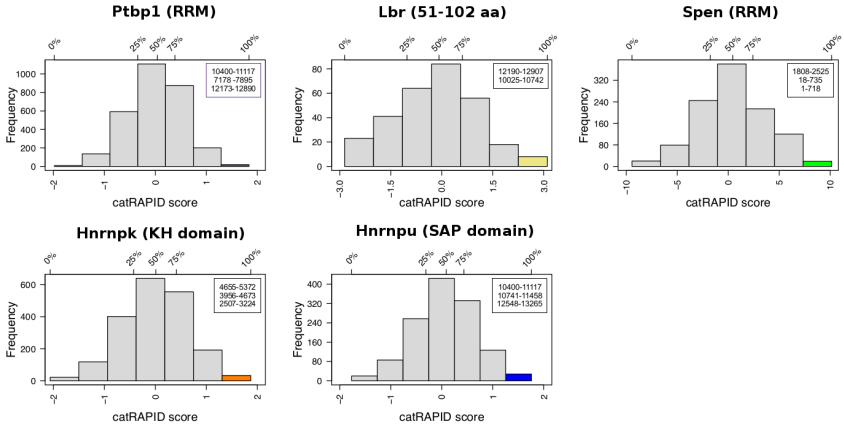


Figure 5. *Interaction propensities of Xist regions with Spen, Hnrnpk, Hrnnpu/Saf-A, Lbr and Ptbp1* For each RNA-binding domain, we used catRAPID to predict the interacting *Xist* regions: Hnrnpk (P61979) 363-414 aa (KH domain) interacting with Xist 2507-3224 nt; Hnrnpu/Saf-A (Q8VEK3) 1-52 aa (SAP domain) with *Xist* 3956-4673 nt; Lbr (Q3U9G9) 51-102 aa (most interacting Lbr fragment, Suppl. Fig. 2) with 10025-10742 nt; Ptbp1 (P17225) 76-127 aa (RRM domain) with *Xist* 10741-11458 nt; Spen (Q62504) 332-477 aa (RRM domain) with *Xist* 18-735.

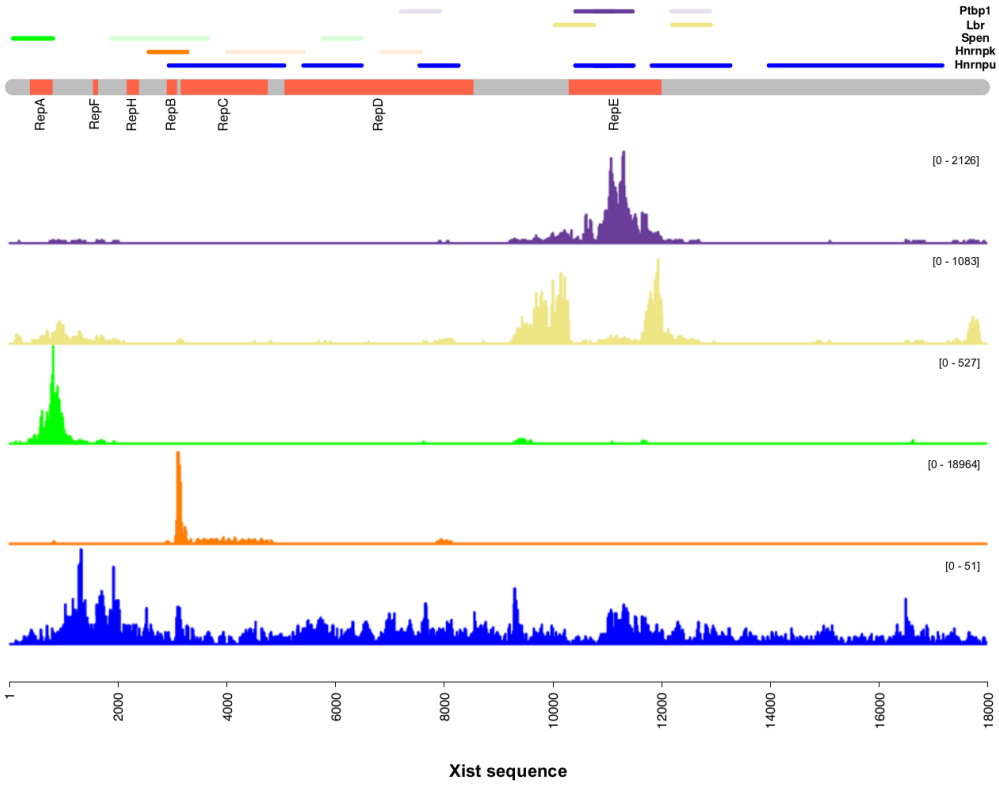
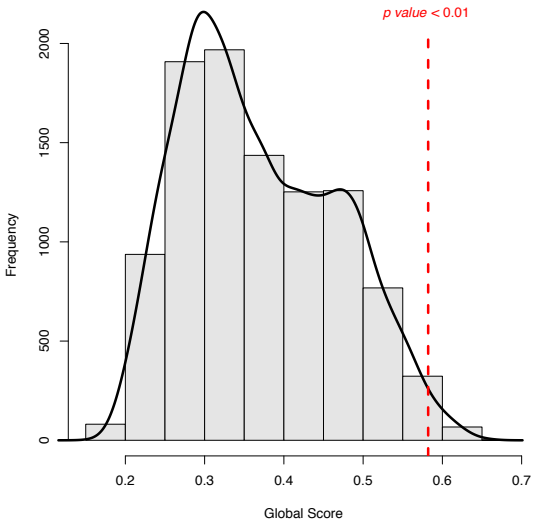
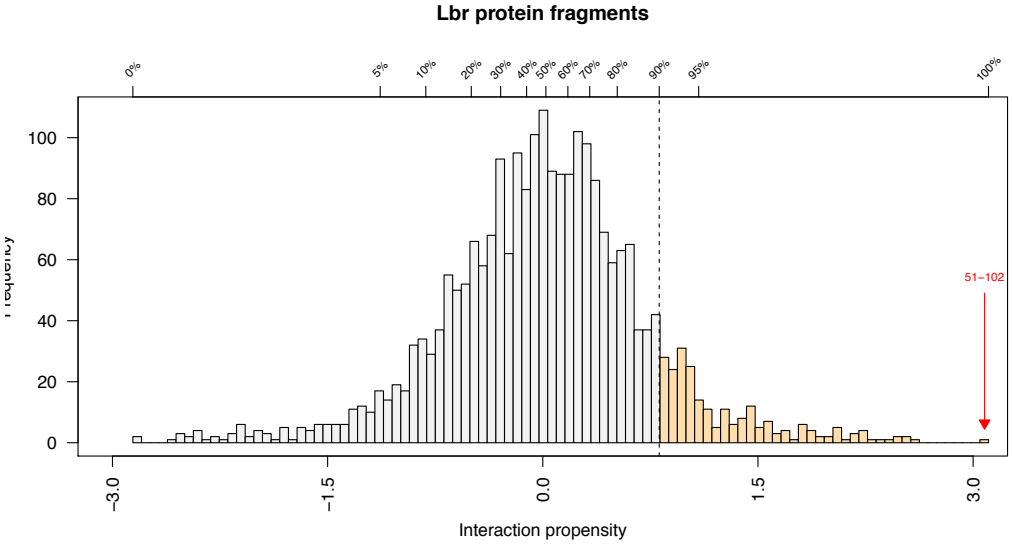


Figure 6. *eCLIP validation of Spen, Hnrnpk, Hrnnpu/Saf-A, Lbr and Ptbp1 binding sites* eCLIP and *cat*RAPID predictions show high agreement: Spen and Hnrpk bind in the 5' of *Xist* (< 5000 nt; respectively A-repeats and B-repeats), Lbr and Ptbp1 show binding sites in the central region (i.e., 9000-13000 nt; E-repeats) while Hnrnrpu/Saf-A has a dispersed signal (throughout all sequence). Predicted binding regions are reported along *Xist* sequence (Methods: Binding sites predictions). Matches are highlighted with colour shades.

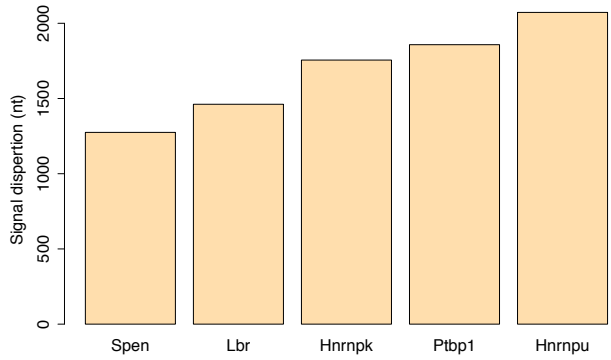
Supplementary Material



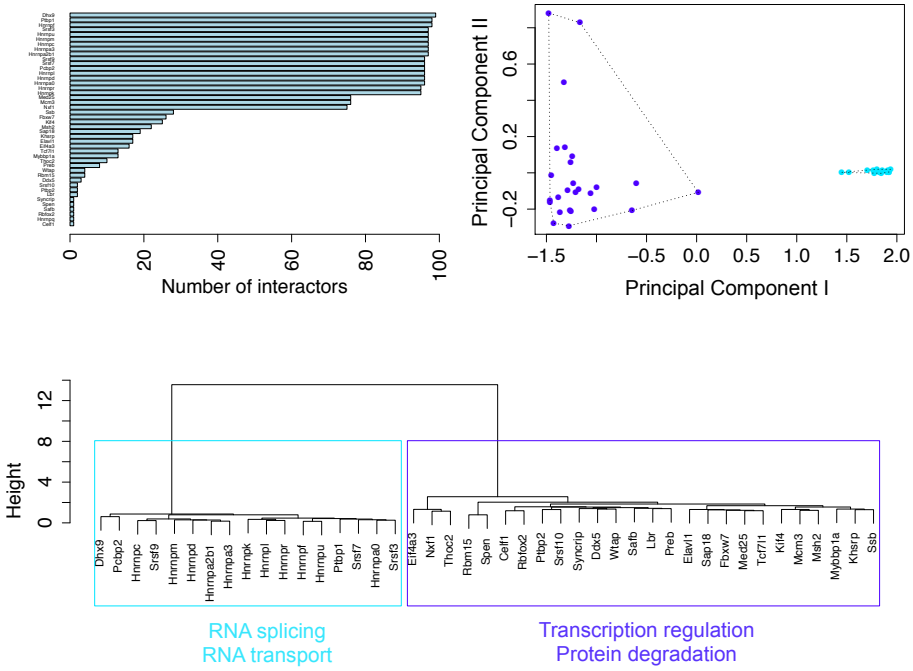
Supplementary Figure 1. *Significance of candidates selection.* By bootstrapping (10^4 randomizations of associations between *Global Score* and experimental evidence values; see also Fig. 3), we observed that predictions above Spen score (*Global Score*=0.59) are significantly associated with values of experimental evidence ≥ 2 (p value<0.01).



Supplementary Figure 2. *Histograms of interaction propensities* catRAPID scores for interactions between Xist and Lbr fragments. Histogram bins containing associations corresponding to the top 10% are highlighted in yellow. The highest interaction propensity is achieved by Lbr fragment 51-102.



Supplementary Figure 3. *Signal dispersion* Variance associated with binding sites predicted by *cat*RAPID.



Supplementary Figure 4. *Analysis of putative XCI candidates* (top left) Number of interactors retrieved from STRING (highest confidence score). (top right) Multidimentional scaling (MDS) of similarities of functions associated to high confidence interactors of putative candidates. (bottom) Dendogram of hierarchical clustering of putative candidates. Most enriched functional categories of putative candidates interactors are reported in cyan and blue.

CHAPTER VI

Reviews on computational methods for protein-RNA interaction prediction

During my Ph.D. studies, I had the opportunity to participate in the writing of two reviews about the state-of-the-art of computational methods for protein-RNA interaction prediction. The two reviews are complementary and cover a broad overview of the main approaches used in the field. The first review uncovers details about *cat*RAPID method (Bellucci et al. 2011) offering illustrative predictions on long noncoding RNA biology and prokaryotic RNA regulation. A collection of notable sequence-based and structure-based (Appendix II) methods is reported. The second review deals with popular experimental and computational methods to detect protein-RNA interactions. A comparison of *cat*RAPID and RIPseq (Muppirala, Honavar, and Dobbs 2011) on autogenous interactions (Zanzoni et al. 2013) is presented. This review presents a practical synopsis of *cat*RAPID modules and implementations. It also introduces *cat*RAPID *omics express* module that is built upon *cat*RAPID *express* (Cirillo, Marchese, et al. 2014) (Chapter III). Moreover, it prefigures future research line in Tartaglia's lab such as the implementation of RNA secondary structure models based on experimental data (e.g. PARS, SHAPE), and the study of ribonucleoprotein aggregates (e.g. nucleoli, stress granules and Cajal bodies) in neurodegenerative diseases.

Cirillo D, Agostini F, Tartaglia GG. Predictions of protein–RNA interactions. Wiley Interdisciplinary Reviews: Computational Molecular Science. Volume 3, Issue 2, Volume 3, pages 161–175, March/April 2013. doi: 10.1002/wcms.1119. Epub 2012 Sep 25. Review.

Cirillo D, Livi CM, Agostini F, Tartaglia GG. Discovery of protein-RNA networks. Mol Biosyst. 2014 Jul;10(7):1632-42. doi: 10.1039/c4mb00099d. Epub 2014 Apr 23. Review. PMID: 24756571

Cirillo D, Agostini F, Tartaglia GG. Predictions of protein-RNA interactions. Wiley Interdisciplinary Reviews: ComputationalMolecular Science. 2013; 3 (2): 161-175. DOI: 10.1002/wcms.1119

Cirillo D, Livi CM, Agostini F, Tartaglia GG. Discovery of protein-RNA networks. Mol Biosyst. 2014 Jul;10(7):1632-42. doi: 10.1039/c4mb00099d. Epub 2014 Apr 23. Review. PubMed PMID: 24756571.

DISCUSSION

Information contained in biological sequences

Protein interaction with nucleic acids (NAs) is at the heart of gene regulation. Recognition sites of NA-binding proteins have been found to be highly sequence-specific in prokaryotes and much less so in eukaryotes (Villar, Flicek, and Odom 2014). As a matter of fact, the complexity of higher eukaryotes requires a concerted series of actions involving transcription factors (TFs) interacting with other proteins and DNA (Stampfel et al. 2015) (*transcriptional* gene regulation) as well as RNA-binding proteins (RBPs) interacting together and with transcripts (Campbell and Wickens 2015) (*post-transcriptional* gene regulation). Furthermore, the chromatin state and its spatial organization (Grubert et al. 2015), and NAs-NAs interactions (e.g. noncoding RNA-DNA interactions) (Holoch and Moazed 2015) add further layers of complexity to the entire process. A comprehensive computational modelling of eukaryotic gene regulation is an ambitious endeavour (Ahsendorf et al. 2014). Nonetheless, simple approaches, although sometimes reductionist (Regenmortel 2004), are useful to improve our understanding of the key parts of the whole process. Examples of such approaches are sequence-based methods integrating functional knowledge like *cat*RAPID *omics express* (Chapter 4) and PA*n*DA (Chapter 5) (see Table 2). These methods are of very general applicability thanks to the main role of primary structure in determining cellular events. Features such as secondary structure and macromolecular assembly are encoded in the sequence from which both tri-dimensional constraints and binding specificities can be derived. Moreover, primary structure is the most complete and reliable source of information due to the wide availability of sequencing data (NCBI Resource Coordinators 2016; Kersey et al. 2016).

From motif-based methods to integrative approaches

Prediction of protein-NA interactions relies on information extracted from RNA and DNA sequences that are recognized by proteins (Introduction section 1.2). A comprehensive knowledge of experimentally-determined recognition sites (Introduction, section 3) is critical to understanding the mechanisms underlying the

binding. The difficulty of acquiring the necessary data makes sequence-based predictions of quantitative estimate of NA-binding not easy to accomplish. Motif-finding algorithm RNAcontext (http://www.rnamotif.org/) (Kazan et al. 2010) is an example of a sequence-oriented approach based on affinity data provided by the *in vitro* assay RNAcompete (Ray et al. 2009). The intuition behind this method is the observation that RBP target recognition is determined by both base content and its tridimensional conformation (i.e. paired, in a hairpin loop, unstructured, and miscellaneous) or structural accessibility (X. Li et al. 2010). Moreover, the predictive value of the structural context highlighted by this work is also taken into account in *cat*RAPID approach (Bellucci et al. 2011). Although *cat*RAPID was not designed to find motifs in RNA sequences, the importance of RNA structure information for protein-RNA interaction prediction is reflected in the comparable performances of the two methods (reported to be 70% to 80%). The role of RNA secondary structure in RBPs recognition is supported by the chemoaffinity structure probing methodology called *in vivo* Click Selective 2'-hydroxyl acylation analysed by primer extension icSHAPE (Spitale et al. 2015). Authors implemented a Support Vector Machine (SVM) algorithm combining icSHAPE signals (*in vivo* and *in vitro*), genomic locations, and sequence conservation to predict RNA binding sites for a number of RBPs with high accuracy.

Regardless of the use of motif models, sequence degeneration represents the main reason that makes motif-based computational approaches inevitably suffer from high error rates (Hannenhalli 2008). In addition, motif-based classification of NA-binding proteins does not necessarily correspond to structural and functional properties of protein-NA complexes. Indeed, it has been demonstrated that structurally-related proteins can recognize the same motif, and proteins recognizing distinct motifs can be part of the same structural group (Prabakaran et al. 2006). Recent detection of secondary motifs shared by multiple TFs in addition to their primary ones (J. Wang et al. 2012; Gerstein et al. 2012) corroborates the sheer complexity of properties and rules that govern protein-NA recognition (Cirillo, Livi, et al. 2014; Cirillo, Botta-Orfila, and Tartaglia 2015).

To reduce the ambiguity of motif-based binding site prediction, novel approaches for recognition sites detection have been developed using (i) improved or alternative binding motif representations and (ii) additional biological information (*integrative* approaches). The first type of methods exploits mainly motif subtypes (Kel et al. 2004; Hannenhalli and Wang 2005; Georgi and Schliep 2006; Bais, Kaminski, and Benos 2011; Chan et al. 2012) and inter-position dependence models (Osada, Zaslavsky, and Singh 2004; Quader and Huang 2012; Keilwagen and Grau 2015). The second type of methods relies on relevant attributes of cellular context such as 'omics' profiles (i.e. transcriptomic, genomic, proteomic, and epigenomic data). Those features can capture the spatio-temporal background of the binding event and have a bearing on a more accurate selection of binding sites. A critical advantage of integrative approaches is the possibility of increasing the coverage of available data (dataset integration) and reducing 'noise' by assessing the reliability of the retrieved information (confidence estimation).

In Chapters 4 and 5, I introduced *cat*RAPID *omics express* and PA*n*DA, two novel large-scale methods for protein-RNA and protein-DNA-interaction prediction, respectively. The two algorithms apply integrative approaches for predicting multiple interacting partners:

- *cat*RAPID *omics express* integrates the Interaction Propensity score of *cat*RAPID method (Bellucci et al. 2011) with expression data (Cirillo, Marchese, et al. 2014) and sequence annotations (protein domains and RNA motifs). The integration is a linear operation resulting in a ranking score that allows transcriptome- or proteome-wide selection of candidate partners among human coding or noncoding RNAs and proteins (full-length or RNA/DNA binding domains, and possibly disordered regions).
- PA*n*DA integrates protein-protein interaction networks, expression levels, and sequence annotations (DNA motifs) to identify putative binding modes of transcription factors onto sets of DNA sequences. Using several machine-learning algorithms, the integration results in a ranking score that allow the selection of mediators of TFs.

DNA- and RNA-binding proteins: akin by nature?

Recently, a large-scale benchmark of NA binding site prediction algorithms (Miao and Westhof 2015) revealed that most of the assessed programs exhibit prediction abilities on both DNA- and RNA-binding proteins, some with AUC values >0.7 on all the datasets, demonstrating that similar interaction rules during NA binding are operating. In line with this finding, a new generation of advanced algorithms are being developed for predicting NA-binding regardless of DNA- and RNA-binding differences.

DeepBind (Alipanahi et al. 2015) is method for predicting NA-binding sites based on convolutional neural networks (CNNs) modelling binding scores directly from raw data of high-throughput experiments (PBM, SELEX, ChIP- and CLIP-seq). Currently, models for 538 distinct TFs and 194 distinct RBPs have been generated that can be used to score new sequences. The two interesting aspects of DeepBind are the following: (i) DeepBind is a sequence-based method that applies the same theoretical framework (i.e. CNNs) to both DNA and RNA binding predictions, highlighting the importance of sequence patterns in NAs recognition process; (ii) While most existing methods are trained on the strongest interacting regions (e.g. the top few hundred peaks of a ChIP-seq experiment), DeepBind models are trained using all sequencing data and reach better accuracies, showing how informative 'extra' sequences could be.

Proteins able to bind both types of nucleic acids are called DNA- and RNA-binding proteins (DRBPs). A list of 149 experimentally validated human DRBPs have been manually curated, containing several regulatory enzymes (Hudson and Ortlund 2014). Indeed, DRBPs undergo many cellular functions ranging from DNA/RNA-related activities to unexpected processes (e.g. apoptosis and response to extreme temperatures). Binding of DNA and RNA can be competitive, simultaneous, or combinatorial, allowing a powerful multi-level regulation of gene expression, often mediated by lncRNAs. It would be extremely compelling to apply the methods and ideas discussed here to carry out the simultaneous investigation of DNA and RNA binding abilities of such engaging set of proteins.

	cat*RAPID *omic express	**PA*n*DA**
Input	Protein or RNA sequences	Protein and DNA sequences
Length restrictions	>50 amino acids or nucleotides	none
Output	• Interaction Propensity score • Tissues expression correlation • Domains and motifs presence • Statistics (Discriminative Power (Bellucci et al. 2011), Interaction Strength (Agostini, Cirillo, et al. 2013), Ranking distribution)	• Binding Propensity score • Binding modes • Statistics (Candidate targets distribution, Mappability (Cirillo, Botta-Orfila, and Tartaglia 2015))
Organism	Homo sapiens	Homo sapiens
Scope	Large-scale	Large-scale
Parameters	• Entire proteins (<750 aa) or NA-binding domains • Disordered regions (optional) • Coding or noncoding transcripts	*Default* mode: • Cell line *Expert* mode: • Cell line • Motif database • Expression threshold • Machine Learning method • Protein-protein Interaction database
Expression data	RNA-seq (Harrow et al. 2012)	RNA-seq (Djebali et al. 2012)
Method	Weighted inner product of linear functions of structural (nucleotide contact frequencies), biochemical (amino acid physico-chemical scales), and predicted (nucleotide secondary structure occupancy) features of protein and RNA sequences.	Four supervised models of several learning methods (K-nearest neighbors, Adaptive Boosting, Support Vector Machine, Random Forest) based on motif occurrences of interacting transcription factors with optimal cell-specific expression levels.

Table 2. Similarities and differences between catRAPID omics express (Cirillo, Livi, et al. 2014) and PAnDA (Cirillo, Botta-Orfila, and Tartaglia 2015) algorithms.

Present and future challenges of integrative approaches

As for many integrative methods, the major limitation of both catRAPID *omics express* and PA*n*DA resides in the availability of high-quality training data: many novel binding regions of protein (Carmen Maria Livi et al. 2015) and RNA (Agostini et al. 2014) sequences are still to be discovered; high-quality high-throughput interaction experiments are to be designed with better efficiency (Rao et al. 2014); several cells and tissues of many organisms will have to be sequenced (Hornett and Wheat 2012).

Such restrictions limit the extent to which data integration can be employed effectively. For instance, for both *cat*RAPID *omics express* and PA*n*DA, predictions are limited to *Homo sapiens* due to the inaccessibility of comprehensive resources of similar data for other organisms. Despite this incompleteness, new experimental approaches exhibit potential to partly fill this methodological gap and lead to better quality predictions.

Even if limited by the availability of efficient antibodies and the multiplicity of cell-lines, tissues and developmental stages, *in vivo* transcriptome-wide discovery of RNA binding sites has improved dramatically over the last decade (Rinn and Ule 2014). Very recently, a catalog of validated IP-quality antibodies against 365 unique RBPs has been released (Sundararaman et al. 2015) based on an 'RBP compilation' of 1072 proteins comprising domain-based (Lunde, Moore, and Varani 2007) and interactome-captured RBPs (Castello et al. 2012). By means of this catalog, eCLIP (enhanced CLIP) (Van Nostrand EL et al, manuscript under preparation) experiments in K562 and HepG2 cells are being performed in the context of ENCODE project (data available at https://www.encodeproject.org). eCLIP is a radioactive-free CLIP protocol which reduces execution time to almost 4 days and requires much many fewer cycles of PCR amplification to get enough material to sequence.

Such a new wealth of genomic features can improve current performances and has even been used to re-train predictive methods. In the case of *cat*RAPID, an interaction potential based on sequence-derived physico-chemical features could be generated

using high-quality data on RBPs for which both CLIP-related and mass spectrometry (MS) data are available. To date, RNA binding sites (PAR-CLIP, iCLIP, CLIP-SEQ/HITS-CLIP, eCLIP, and RIP-seq experiments) and MS-validated enzymatic/nonenzymatic RBDs (Gerstberger et al. 2014) are available for 70 human RBPs. Considering the high-throughput nature of these experiments, the database of 'interacting regions' is expected to substantially exceed that of *cat*RAPID original training set (Bellucci et al. 2011), although the variability of protein sequences is to be increased.

In the case of PA*n*DA, evaluation of chromatin accessibility (Tsompana and Buck 2014) and chromosome conformation (Cao et al. 2015) could be used to better select candidate target regions. The main drawback of PA*n*DA algorithm is that parameters such as optimal expression thresholds and mappability (i.e. a measure of co-factors' motif coverage) must be derived anew whenever expanded databases of expression levels, interaction networks and binding motifs make part of an updated version. In addition, the use of RNA-seq data as a proxy for TF protein concentration implies a direct proportionality between protein and mRNA expression levels which is still a debated issue (Vogel and Marcotte 2012; Cirillo, Marchese, et al. 2014). As an alternative, data on protein abundances could be used (M. Wang et al. 2015).

A perspective on NA-binding protein assemblies

Integrative approaches are key for protein-NA interaction predictions (Levo and Segal 2014). A peculiar feature of such methods is that they provide meaningful information that extends pretty much over the practical value of a single prediction score. Indeed, multiple features combined into a model designed to reveal interaction propensities will help to unveil the broader context in which a physical event is occurring, for instance the cooperation of multiple NA-binding proteins in the same activity.

As for RNA-binding, the method iONMF (Stražar et al. 2016) represents an example of high-throughput data integration that yields remarkable improvements on prediction accuracy and downstream applications, such as the interpretation of RNA recognition determinants. By means of multiple matrix factorization

technique, iONMF is able to generate models for multiple RNA-binding proteins using several data sources: RNA secondary structure prediction (Denman 1993), functional annotations (Ashburner et al. 2000), as well as RBP co-binding, *k*-mer composition, and region type (exon, intron, 5'UTR, 3'UTR, CDS) derived from a large collection of iCLIP, PAR-CLIP, CLIP-SEQ/HITS-CLIP experiments [(Anders et al. 2012) and http://icount.biolab.si]. Interestingly, the two most informative data sources revealed by iONMF for RBP binding are RNA structure and RBPs co-binding within the same gene region.

As highlighted in Chapters 5 and 6, cooperation between different proteins is essential to recapitulate how proteins bind to RNA and DNA, respectively (Mascareñas 2008). Cooperative binding is critical for biological function of NA-binding proteins like transcription factors (Levine and Tjian 2003; Spitz and Furlong 2012). Cooperative TFs are clustered within protein interaction networks (Manke, Bringas, and Vingron 2003), are found in shortened distance along DNA sequences (Aguilar and Oliva 2008), and are evolutionary conserved (He et al. 2011). This combinatorial interplay is suspected to be responsible for driving distinct functions and regulatory control mechanisms (Farnham 2009; MacQuarrie et al. 2011; Stampfel et al. 2015). Also RBPs engage in homo- and hetero-oligomeric interactions (Danner 2002). An illustrative example is Hnrnp complex (Krecic and Swanson 1999), which I recently found to be an essential part of RBPs interactome of long noncoding RNA *Xist* (Chapter 5).

Over the last decade, more than 20 different methods have been proposed for complex prediction (Srihari et al. 2015) based (i) solely on protein-protein interaction (PPI) network topology or (ii) combined with auxiliary biological insights. The study of protein complexes allows the identification of *modules* or groups of interacting molecules regulating specific biological processes (Hartwell et al. 1999). Integrative methods for NA-protein interaction prediction such as PA*n*DA and *cat*RAPID *omics express* have the inherent ability to identify functional modules. In the case of PA*n*DA, predicted TF binding modes based on cell-specific PPI network bring out key mediators of TF activity. In the case of *cat*RAPID *omics express*, co-expressed RBPs with high interaction propensities might bind cooperatively to the same RNA targets.

Hence, both methods permit meaningful further analysis towards organisation, function and dynamics of NA-related modules.

Interestingly, the theoretical methodologies developed here pave the way to design a 'multibody' simulation for protein-NA interactions using components of PPI networks. In physics, a multibody (or *n*-body) simulation is a representation of a dynamic system of objects under the influence of physical forces. In the context of molecular dynamics, all-atom simulation of large macromolecular assemblies remains a computational challenge (Pankavich and Ortoleva 2015). Nonetheless, by eliminating some of the interaction details, a 'coarse-grained' description of the system can help to overcome computational limitations. I speculate that the use of phenomenological constraints can further speed up calculations, especially in the case of interaction (or *docking*) simulations (Krippahl and Barahona 2015). In theory, a protein-RNA interaction simulation could be constrained to *cat*RAPID *omics express* predicted binding sites onto RNA, protein and co-expressed RBPs belonging to the same PPI network. This approach could dramatically accelerate the simulation process.

Conclusions

The work carried out during my Ph.D. studies at Centre for Genomic Regulation (CRG) of Barcelona, Spain, has been compiled in the form of a thesis entitled "Protein and Nucleic Acid Interactions". The thesis presents my personal contribution to the field of computational prediction of macromolecular interactions.

In the first half of my career as a Ph.D. student I have been extensively working on testing and improving the performances of several modules of *cat*RAPID suite for protein-RNA interaction prediction. By employing *cat*RAPID algorithm, I investigated a number of protein-RNA associations involved in many physiological and pathological processes such as neurodegenerative diseases (Chapter I), chromatin regulation (Chapter II), and cancer (Chapter III). Subsequently, I implemented and applied two novel algorithms for protein-NAs interaction prediction: *cat*RAPID *omics express* (Chapters III and VI) and PA*n*DA (Chapter IV).

*cat*RAPID *omics express* is a module of *cat*RAPID suite that computes the interaction propensity of human proteome and transcriptome taking into account expression levels. The implementation of *cat*RAPID *omics express* was prompted by insights on the relation between interaction propensity and correlation in expression of protein and RNAs in human tissues. PA*n*DA predicts the interaction between DNA and assemblies of TFs. The algorithm is built upon the finding that PPI networks and cell-specific expression levels improve performances in predicting binding events.

Overall, the two algorithms are sequence-based methods integrating genomic and functional annotations such as expression levels and PPI interaction networks. This new way of approaching protein-NA interaction prediction has been recently applied to disentangle *Xist* interactome (Chapter VII) paving the way to the study of other long noncoding RNAs using similar computational approaches.

Appendix I

Selection of protein-DNA interaction prediction methods
[adapted from (Nagarajan, Ahmad, and Michael Gromiha 2013)]

Sequence-based methods (sorted by time of publication):

PAnDA
http://service.tartaglialab.com/new_submission/panda (Cirillo, Botta-Orfila, and Tartaglia 2015)

SNBRFinder
http://ibi.hzau.edu.cn/SNBRFinder/ (X. Yang et al. 2015)

INTERACT-O-FINDER
http://interacto.eurekanow.org/index.html (Samant, Jethva, and Hasija 2014)

newDNA-Prot
http://sourceforge.net/projects/newdnaprot/ (Y. Zhang et al. 2014)

iDNA-Prot|dis http://bioinformatics.hitsz.edu.cn/iDNA-Prot_dis/ (B. Liu et al. 2014)

MuMoD
Program available upon request from the authors (Narlikar 2013)

DNABR
http://www.cbi.seu.edu.cn/DNABR/ (Ma et al. 2012)

MetaDBSite
http://projects.biotec.tu-dresden.de/metadbsite/ (Si et al. 2011)

NAPS
http://omictools.com/naps-tool (Carson, Langlois, and Lu 2010)

BindN+
http://bioinfo.ggc.org/bindn+/ (L. Wang et al. 2010)

hPDI
http://bioinfo.wilmer.jhu.edu/PDI/ (Xie et al. 2010)

BindN-RF
http://bioinfo.ggc.org/bindn-rf/ (L. Wang, Yang, and Yang 2009)

DBindR
http://www.cbi.seu.edu.cn/DBindR/DBindR.htm (J. Wu et al. 2009)

ProteDNA
http://serv.csbb.ntu.edu.tw/ProteDNA/ (W.-Y. Chu et al. 2009)

DISIS
http://cubic.bioc.columbia.edu/services/disis (Ofran, Mysore, and Rost 2007)

DP-Bind
http://lcg.rit.albany.edu/dp-bind/ (Hwang, Gou, and Kuznetsov 2007)

TFmodeller
http://maya.ccg.unam.mx/~tfmodell/ (Contreras-Moreira, Branger, and Collado-Vides 2007)

BindN
http://bioinfo.ggc.org/bindn/ (L. Wang and Brown 2006)

DBS-PSSM
http://dbspssm.netasa.org/ (Ahmad and Sarai 2005)

DBS-PRED
http://www.abren.net/dbs-pred/ (Ahmad, Gromiha, and Sarai 2004)

Structure-based methods:

NuProPlot

http://www.nuproplot.com/ (Pradhan and Nam 2015)

SPOT-Struct-DNA
http://sparks-lab.org/yueyang/server/SPOT-Struct-DNA/ (Zhao et al. 2014)

CONSRANK
https://www.molnac.unisa.it/BioTools/consrank/ (Chermak et al. 2014)

DBSI
https://mitchell-lab.biochem.wisc.edu/DBSI_Server/index.php (Zhu, Ericksen, and Mitchell 2013)

DNABind
http://mleg.cse.sc.edu/DNABind/ (R. Liu and Hu 2013)

PreDNA
http://202.207.14.178/predna/ (T. Li et al. 2013)

Nucleos
nucleos.bio.uniroma2.it/nucleos/ (Parca et al. 2013)

DBD2BS
http://dbd2bs.csbb.ntu.edu.tw/ (Chien et al. 2012)

3DTF
http://www.gene-regulation.com/pub/programs/3dtf/ (Gabdoulline et al. 2012)

CONS-COCOMAPS
https://www.molnac.unisa.it/BioTools/conscocomaps/ (Vangone, Oliva, and Cavallo 2012)

COCOMAPS
https://www.molnac.unisa.it/BioTools/cocomaps/ (Vangone et al. 2011)

iDBPs
http://idbps.tau.ac.il/ (Nimrod et al. 2010)

PDA
http://bioinfozen.uncc.edu/webpda/ (R. Kim and Guo 2009)

DBD-Threader
http://cssb.biology.gatech.edu/skolnick/webservice/DBD-Threader/index.html (Gao and Skolnick 2009b)

DBD-Hunter
http://cssb.biology.gatech.edu/skolnick/webservice/DBD-Hunter/index.html (Gao and Skolnick 2008)

DISPLAR
http://pipe.scs.fsu.edu/displar.html (Tjong and Zhou 2007)

DNABINDPROT
http://www.prc.boun.edu.tr/appserv/prc/dnabindprot/ (Ozbek et al. 2010)

DP-dock
http://cssb.biology.gatech.edu/skolnick/webservice/DP-dock/index.html (Gao and Skolnick 2009a)

PFplus
http://pfp.technion.ac.il/ (Shazman et al. 2007)

Appendix II

Selection of protein-RNA interaction prediction methods
[adapted from (Cirillo, Agostini, and Tartaglia 2013; Si et al. 2015)]

Sequence-based methods (sorted by time of publication):

PRIPU
http://admis.fudan.edu.cn/projects/pripu.htm (Cheng, Zhou, and Guan 2015)

Oli
Program available upon request from the authors (Carmen M Livi and Blanzieri 2014)

RNABindRPlus
http://einstein.cs.iastate.edu/RNABindRPlus/ (Walia et al. 2014)

catRAPID
http://s.tartaglialab.com/catrapid (Bellucci et al. 2011; Agostini, Zanzoni, et al. 2013)

SRCPred
http://tardis.nibio.go.jp/netasa/srcpred (Fernandez et al. 2011)

SPOT
http://sparks.informatics.iupui.edu (Zhao, Yang, and Zhou 2011)

PRBR
http://www.cbi.seu.edu.cn/PRBR/ (Ma et al. 2011)

RNAPred
http://www.imtech.res.in/raghava/rnapred/ (M. Kumar, Gromiha, and Raghava 2011)

RPISeq
http://pridb.gdcb.iastate.edu/RPISeq/ (Muppirala, Honavar, and Dobbs 2011)

BindN+

http://bioinfo.ggc.org/bindn+/ (L. Wang et al. 2010)

NAPS
http://prediction.bioengr.uic.edu/ (Carson, Langlois, and Lu 2010)

PiRaNhA
http://bioinformatics.sussex.ac.uk/PIRANHA (Murakami et al. 2010)

PRNA
http://www.sysbio.ac.cn/datatools.asp (Z.-P. Liu et al. 2010)

RNA
http://mcgill.3322.org/RNA/ (Q. Li, Cao, and Liu 2010)

RISP
http://grc.seu.edu.cn/RISP (Tong, Jiang, and Lu 2008)

PRINTR
http://210.42.106.80/printr/ (Y. Wang et al. 2008)

PPRInt
http://www.imtech.res.in/raghava/pprint/ (M. Kumar, Gromiha, and Raghava 2008)8)

RNABindR
http://bindr2.gdcb.iastate.edu/RNABindR/ (Terribilini et al. 2007)

BindN
http://bioinfo.ggc.org/bindn/ (L. Wang and Brown 2006)

SVMProt
http://jing.cz3.nus.edu.sg/cgi-bin/svmprot.cgi (Han et al. 2004)

Structure-based methods:

RBPDetector
http://ibi.hzau.edu.cn/rbrdetector (X.-X. Yang, Deng, and Liu 2014)

SPOT-Seq-RNA
http://sparks-lab.org/server/SPOT-Seq-RNA/ (Y. Yang et al. 2014)

DRNA
http://sparks.informatics.iupui.edu/yueyang/DFIRE/dRdR-DB-service (Zhao, Yang, and Zhou 2011)

OPRA
Program available upon request from the authors (Pérez-Cano et al. 2010)

Struct-NB
http://www.public.iastate.edu/~ftowfic (Towfic et al. 2010)

PRIP
http://www.qfab.org/PRIP (Maetschke and Yuan 2009)

PatchFinderPlus
http://pfp.technion.ac.il/ (Shazman and Mandel-Gutfreund 2008)

KYG
http://cib.cf.ocha.ac.jp/KYG/ (O. T. P. Kim, Yura, and Go 2006)

Bibliography

Acharya, K. Ravi, and Matthew D. Lloyd. 2005. "The Advantages and Limitations of Protein Crystal Structures." *Trends in Pharmacological Sciences* 26 (1): 10–14. doi:10.1016/j.tips.2004.10.011.

Agostini, Federico, Davide Cirillo, Benedetta Bolognesi, and Gian Gaetano Tartaglia. 2013. "X-Inactivation: Quantitative Predictions of Protein Interactions in the Xist Network." *Nucleic Acids Research* 41 (1): e31. doi:10.1093/nar/gks968.

Agostini, Federico, Davide Cirillo, Riccardo Delli Ponti, and Gian Gaetano Tartaglia. 2014. "SeAMotE: A Method for High-Throughput Motif Discovery in Nucleic Acid Sequences." *BMC Genomics* 15 (1): 925. doi:10.1186/1471-2164-15-925.

Agostini, Federico, Andreas Zanzoni, Petr Klus, Domenica Marchese, Davide Cirillo, and Gian Gaetano Tartaglia. 2013. "catRAPID Omics: A Web Server for Large-Scale Prediction of Protein-RNA Interactions." *Bioinformatics (Oxford, England)* 29 (22): 2928–30. doi:10.1093/bioinformatics/btt495.

Aguilar, Daniel, and Baldo Oliva. 2008. "Topological Comparison of Methods for Predicting Transcriptional Cooperativity in Yeast." *BMC Genomics* 9: 137. doi:10.1186/1471-2164-9-137.

Ahmad, Shandar, M. Michael Gromiha, and Akinori Sarai. 2004. "Analysis and Prediction of DNA-Binding Proteins and Their Binding Residues Based on Composition, Sequence and Structural Information." *Bioinformatics (Oxford, England)* 20 (4): 477–86. doi:10.1093/bioinformatics/btg432.

Ahmad, Shandar, and Akinori Sarai. 2005. "PSSM-Based Prediction of DNA Binding Sites in Proteins." *BMC Bioinformatics* 6: 33. doi:10.1186/1471-2105-6-33.

Ahsendorf, Tobias, Felix Wong, Roland Eils, and Jeremy Gunawardena. 2014. "A Framework for Modelling Gene Regulation Which Accommodates Non-Equilibrium Mechanisms." *BMC Biology* 12: 102. doi:10.1186/s12915-014-0102-4.

Alberts, Bruce. 1989. *Molecular Biology of the Cell*.

Alipanahi, Babak, Andrew Delong, Matthew T. Weirauch, and Brendan J. Frey. 2015. "Predicting the Sequence Specificities of DNA- and RNA-Binding Proteins by Deep Learning." *Nature Biotechnology* 33 (8): 831–38. doi:10.1038/nbt.3300.

Allers, J, and Y Shamoo. 2001. "Structure-Based Analysis of Protein-RNA Interactions Using the Program ENTANGLE." *Journal of Molecular Biology* 311 (1): 75–86. doi:10.1006/jmbi.2001.4857.

AlQuraishi, Mohammed, Shengdong Tang, and Xide Xia. 2015. "An Affinity-Structure Database of Helix-Turn-Helix: DNA Complexes with a Universal Coordinate System." *BMC Bioinformatics* 16 (1): 390. doi:10.1186/s12859-015-0819-2.

Alva, Vikram, Johannes Söding, and Andrei N. Lupas. 2016. "A Vocabulary of Ancient Peptides at the Origin of Folded Proteins." *eLife* 4 (January): e09410. doi:10.7554/eLife.09410.

Anders, Gerd, Sebastian D. Mackowiak, Marvin Jens, Jonas Maaskola, Andreas

Kuntzagk, Nikolaus Rajewsky, Markus Landthaler, and Christoph Dieterich. 2012. “doRiNA: A Database of RNA Interactions in Post-Transcriptional Regulation.” *Nucleic Acids Research* 40 (D1): D180–86. doi:10.1093/nar/gkr1007.

Aparicio, Oscar, Joseph V. Geisberg, and Kevin Struhl. 2004. “Chromatin Immunoprecipitation for Determining the Association of Proteins with Specific Genomic Sequences in Vivo.” *Current Protocols in Cell Biology / Editorial Board, Juan S. Bonifacino ... [et Al.]* Chapter 17 (September): Unit 17.7. doi:10.1002/0471143030.cb1707s23.

Appasamy, Sri Devan, Hazrina Yusof Hamdani, Effirul Ikhwan Ramlan, and Mohd Firdaus-Raih. 2016. “InterRNA: A Database of Base Interactions in RNA Structures.” *Nucleic Acids Research* 44 (D1): D266–71. doi:10.1093/nar/gkv1186.

Arnaud, Nicolas, Tom Lawrenson, Lars Østergaard, and Robert Sablowski. 2011. “The Same Regulatory Point Mutation Changed Seed-Dispersal Structures in Evolution and Domestication.” *Current Biology: CB* 21 (14): 1215–19. doi:10.1016/j.cub.2011.06.008.

Ascano, Manuel, Markus Hafner, Pavol Cekan, Stefanie Gerstberger, and Thomas Tuschl. 2012. “Identification of RNA–protein Interaction Networks Using PAR-CLIP.” *Wiley Interdisciplinary Reviews. RNA* 3 (2): 159–77. doi:10.1002/wrna.1103.

Ashburner, M, C A Ball, J A Blake, D Botstein, H Butler, J M Cherry, A P Davis, et al. 2000. “Gene Ontology: Tool for the Unification of Biology. The Gene Ontology Consortium.” *Nature Genetics* 25 (1): 25–29. doi:10.1038/75556.

Bahadur, Ranjit Prasad, Martin Zacharias, and Joël Janin. 2008. “Dissecting protein–RNA Recognition Sites.” *Nucleic Acids Research* 36 (8): 2705–16. doi:10.1093/nar/gkn102.

Bailey, Timothy, Pawel Krajewski, Istvan Ladunga, Celine Lefebvre, Qunhua Li, Tao Liu, Pedro Madrigal, Cenny Taslim, and Jie Zhang. 2013. “Practical Guidelines for the Comprehensive Analysis of ChIP-Seq Data.” *PLoS Computational Biology* 9 (11): e1003326. doi:10.1371/journal.pcbi.1003326.

Bailey, Timothy L., James Johnson, Charles E. Grant, and William S. Noble. 2015. “The MEME Suite.” *Nucleic Acids Research* 43 (W1): W39–49. doi:10.1093/nar/gkv416.

Bais, Abha Singh, Naftali Kaminski, and Panayiotis V. Benos. 2011. “Finding Subtypes of Transcription Factor Motif Pairs with Distinct Regulatory Roles.” *Nucleic Acids Research* 39 (11): e76. doi:10.1093/nar/gkr205.

Baltz, Alexander G, Mathias Munschauer, Björn Schwanhäusser, Alexandra Vasile, Yasuhiro Murakawa, Markus Schueler, Noah Youngs, et al. 2012. “The mRNA-Bound Proteome and Its Global Occupancy Profile on Protein-Coding Transcripts.” *Molecular Cell* 46 (5): 674–90. doi:10.1016/j.molcel.2012.05.021.

Barik, Amita, Nithin C, Smita P. Pilla, and Ranjit Prasad Bahadur. 2015. “Molecular Architecture of Protein-RNA Recognition Sites.” *Journal of Biomolecular Structure and Dynamics* 33 (12): 2738–51. doi:10.1080/07391102.2015.1004652.

Barik, Amita, Nithin C, Manasa P, and Ranjit Prasad Bahadur. 2012. “A Protein-

RNA Docking Benchmark (I): Nonredundant Cases." *Proteins* 80 (7): 1866–71. doi:10.1002/prot.24083.

Beckmann, Benedikt M., Rastislav Horos, Bernd Fischer, Alfredo Castello, Katrin Eichelbaum, Anne-Marie Alleaume, Thomas Schwarzl, et al. 2015. "The RNA-Binding Proteomes from Yeast to Man Harbour Conserved enigmRBPs." *Nature Communications* 6: 10127. doi:10.1038/ncomms10127.

Bellucci, Matteo, Federico Agostini, Marianela Masin, and Gian Gaetano Tartaglia. 2011. "Predicting Protein Associations with Long Noncoding RNAs." *Nature Methods* 8 (6): 444–45. doi:10.1038/nmeth.1611.

Berman, Helen M., John Westbrook, Zukang Feng, Gary Gilliland, T. N. Bhat, Helge Weissig, Ilya N. Shindyalov, and Philip E. Bourne. 2000. "The Protein Data Bank." *Nucleic Acids Research* 28 (1): 235–42. doi:10.1093/nar/28.1.235.

Blecher-Gonen, Ronnie, Zohar Barnett-Itzhaki, Diego Jaitin, Daniela Amann-Zalcenstein, David Lara-Astiaso, and Ido Amit. 2013. "High-Throughput Chromatin Immunoprecipitation for Genome-Wide Mapping of in Vivo Protein-DNA Interactions and Epigenomic States." *Nature Protocols* 8 (3): 539–54. doi:10.1038/nprot.2013.023.

Blin, Kai, Christoph Dieterich, Ricardo Wurmus, Nikolaus Rajewsky, Markus Landthaler, and Altuna Akalin. 2015. "DoRiNA 2.0--Upgrading the doRiNA Database of RNA Interactions in Post-Transcriptional Regulation." *Nucleic Acids Research* 43 (Database issue): D160–67. doi:10.1093/nar/gku1180.

Bon, Michael, Graziano Vernizzi, Henri Orland, and A. Zee. 2008. "Topological Classification of RNA Structures." *Journal of Molecular Biology* 379 (4): 900–911. doi:10.1016/j.jmb.2008.04.033.

Bränd´en, Carl-Ivar, and T. Alwyn Jones. 1990. "Between Objectivity and Subjectivity." *Nature* 343 (6260): 687–89. doi:10.1038/343687a0.

Brimacombe, R., W. Stiege, A. Kyriatsoulis, and P. Maly. 1988. "Intra-RNA and RNA-Protein Cross-Linking Techniques in Escherichia Coli Ribosomes." *Methods in Enzymology* 164: 287–309.

Brown, Tom. 2011. *Nucleic Acids Book*. ATDBio. http://www.atdbio.com/nucleic-acids-book.

Bryce, C. F. A., and D. Pacini. 1998. *The Structure and Function of Nucleic Acids*. Biochemical Society.

Campbell, Zachary T., Devesh Bhimsaria, Cary T. Valley, Jose A. Rodriguez-Martinez, Elena Menichelli, James R. Williamson, Aseem Z. Ansari, and Marvin Wickens. 2012. "Cooperativity in RNA-Protein Interactions: Global Analysis of RNA Binding Specificity." *Cell Reports* 1 (5): 570–81. doi:10.1016/j.celrep.2012.04.003.

Campbell, Zachary T., and Marvin Wickens. 2015. "Probing RNA-Protein Networks: Biochemistry Meets Genomics." *Trends in Biochemical Sciences* 40 (3): 157–64. doi:10.1016/j.tibs.2015.01.003.

Cao, Jun, Zhengyu Luo, Qingyu Cheng, Qianlan Xu, Yan Zhang, Fei Wang, Yan Wu, and Xiaoyuan Song. 2015. "Three-Dimensional Regulation of Transcription." *Protein & Cell* 6 (4): 241–53. doi:10.1007/s13238-015-0135-7.

Carson, Matthew B., Robert Langlois, and Hui Lu. 2010. "NAPS: A Residue-

Level Nucleic Acid-Binding Prediction Server." *Nucleic Acids Research* 38 (Web Server issue): W431–35. doi:10.1093/nar/gkq361.

Carter, Phil, Claus A. F. Andersen, and Burkhard Rost. 2003. "DSSPcont: Continuous Secondary Structure Assignments for Proteins." *Nucleic Acids Research* 31 (13): 3293–95.

Castello, Alfredo, Bernd Fischer, Katrin Eichelbaum, Rastislav Horos, Benedikt M Beckmann, Claudia Strein, Norman E Davey, et al. 2012. "Insights into RNA Biology from an Atlas of Mammalian mRNA-Binding Proteins." *Cell* 149 (6): 1393–1406. doi:10.1016/j.cell.2012.04.031.

Castello, Alfredo, Matthias W. Hentze, and Thomas Preiss. 2015. "Metabolic Enzymes Enjoying New Partnerships as RNA-Binding Proteins." *Trends in Endocrinology and Metabolism: TEM* 26 (12): 746–57. doi:10.1016/j.tem.2015.09.012.

Castello, Alfredo, Rastislav Horos, Claudia Strein, Bernd Fischer, Katrin Eichelbaum, Lars M. Steinmetz, Jeroen Krijgsveld, and Matthias W. Hentze. 2016. "Comprehensive Identification of RNA-Binding Proteins by RNA Interactome Capture." *Methods in Molecular Biology (Clifton, N.J.)* 1358: 131–39. doi:10.1007/978-1-4939-3067-8_8.

Chan, Tak-Ming, Kwong-Sak Leung, Kin-Hong Lee, Man-Hon Wong, Terrence Chi-Kong Lau, and Stephen Kwok-Wing Tsui. 2012. "Subtypes of Associated protein–DNA (Transcription Factor-Transcription Factor Binding Site) Patterns." *Nucleic Acids Research* 40 (19): 9392–9403. doi:10.1093/nar/gks749.

Chen, Beibei, Jonghyun Yun, Min Soo Kim, Joshua T. Mendell, and Yang Xie. 2014. "PIPE-CLIP: A Comprehensive Online Tool for CLIP-Seq Data Analysis." *Genome Biology* 15 (1): R18. doi:10.1186/gb-2014-15-1-r18.

Cheng, Zhanzhan, Shuigeng Zhou, and Jihong Guan. 2015. "Computationally Predicting Protein-RNA Interactions Using Only Positive and Unlabeled Examples." *Journal of Bioinformatics and Computational Biology* 13 (3): 1541005. doi:10.1142/S021972001541005X.

Chermak, Edrisse, Andrea Petta, Luigi Serra, Anna Vangone, Vittorio Scarano, Luigi Cavallo, and Romina Oliva. 2014. "CONSRANK: A Server for the Analysis, Comparison and Ranking of Docking Models Based on Inter-Residue Contacts." *Bioinformatics*, December, btu837. doi:10.1093/bioinformatics/btu837.

Chien, Ting-Ying, Chih-Kang Lin, Chih-Wei Lin, Yi-Zhong Weng, Chien-Yu Chen, and Darby Tien-Hao Chang. 2012. "DBD2BS: Connecting a DNA-Binding Protein with Its Binding Sites." *Nucleic Acids Research* 40 (Web Server issue): W173–79. doi:10.1093/nar/gks564.

Chojnowski, Grzegorz, Tomasz Walen, and Janusz M. Bujnicki. 2014. "RNA Bricks--a Database of RNA 3D Motifs and Their Interactions." *Nucleic Acids Research* 42 (Database issue): D123–31. doi:10.1093/nar/gkt1084.

Chu, Ci, Kun Qu, Franklin L Zhong, Steven E Artandi, and Howard Y Chang. 2011. "Genomic Maps of Long Noncoding RNA Occupancy Reveal Principles of RNA-Chromatin Interactions." *Molecular Cell* 44 (4): 667–78. doi:10.1016/j.molcel.2011.08.027.

Chu, Wen-Yi, Yu-Feng Huang, Chun-Chin Huang, Yi-Sheng Cheng, Chien-Kang Huang, and Yen-Jen Oyang. 2009. "ProteDNA: A Sequence-Based Predictor of Sequence-Specific DNA-Binding Residues in Transcription

Factors." *Nucleic Acids Research* 37 (Web Server issue): W396–401. doi:10.1093/nar/gkp449.

Cirillo, Davide, Federico Agostini, Petr Klus, Domenica Marchese, Silvia Rodriguez, Benedetta Bolognesi, and Gian Gaetano Tartaglia. 2013. "Neurodegenerative Diseases: Quantitative Predictions of Protein-RNA Interactions." *RNA (New York, N.Y.)* 19 (2): 129–40. doi:10.1261/rna.034777.112.

Cirillo, Davide, Federico Agostini, and Gian Gaetano Tartaglia. 2013. "Predictions of protein–RNA Interactions." *Wiley Interdisciplinary Reviews: Computational Molecular Science* 3 (2): 161–75. doi:10.1002/wcms.1119.

Cirillo, Davide, Teresa Botta-Orfila, and Gian Gaetano Tartaglia. 2015. "By the Company They Keep: Interaction Networks Define the Binding Ability of Transcription Factors." *Nucleic Acids Research* 43 (19): e125. doi:10.1093/nar/gkv607.

Cirillo, Davide, Carmen Maria Livi, Federico Agostini, and Gian Gaetano Tartaglia. 2014. "Discovery of Protein-RNA Networks." *Molecular bioSystems* 10 (7): 1632–42. doi:10.1039/c4mb00099d.

Cirillo, Davide, Domenica Marchese, Federico Agostini, Carmen Maria Livi, Teresa Botta-Orfila, and Gian Gaetano Tartaglia. 2014. "Constitutive Patterns of Gene Expression Regulated by RNA-Binding Proteins." *Genome Biology* 15 (1): R13. doi:10.1186/gb-2014-15-1-r13.

Cloonan, Nicole, Alistair R. R. Forrest, Gabriel Kolle, Brooke B. A. Gardiner, Geoffrey J. Faulkner, Mellissa K. Brown, Darrin F. Taylor, et al. 2008. "Stem Cell Transcriptome Profiling via Massive-Scale mRNA Sequencing." *Nature Methods* 5 (7): 613–19. doi:10.1038/nmeth.1223.

Coimbatore Narayanan, Buvaneswari, John Westbrook, Saheli Ghosh, Anton I. Petrov, Blake Sweeney, Craig L. Zirbel, Neocles B. Leontis, and Helen M. Berman. 2014. "The Nucleic Acid Database: New Features and Capabilities." *Nucleic Acids Research* 42 (Database issue): D114–22. doi:10.1093/nar/gkt980.

Contreras-Moreira, Bruno. 2010. "3D-Footprint: A Database for the Structural Analysis of Protein-DNA Complexes." *Nucleic Acids Research* 38 (Database issue): D91–97. doi:10.1093/nar/gkp781.

Contreras-Moreira, Bruno, Pierre-Alain Branger, and Julio Collado-Vides. 2007. "TFmodeller: Comparative Modelling of Protein-DNA Complexes." *Bioinformatics (Oxford, England)* 23 (13): 1694–96. doi:10.1093/bioinformatics/btm148.

Cook, Kate B., Hilal Kazan, Khalid Zuberi, Quaid Morris, and Timothy R. Hughes. 2010. "RBPDB: A Database of RNA-Binding Specificities." *Nucleic Acids Research*, October, gkq1069. doi:10.1093/nar/gkq1069.

Crick, Francis. 1993. *The RNA World: Monograph 24*. Edited by Raymond F. Gesteland and John F. Atkins. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press,U.S. https://cshmonographs.org/index.php/monographs/issue/view/087969380.24.

Danner, Dean J. 2002. "RNA Binding Proteins: New Concepts in Gene Regulation." *American Journal of Human Genetics* 71 (5): 1255.

Darnell, Robert B. 2010. "HITS-CLIP: Panoramic Views of Protein-RNA

Regulation in Living Cells." *Wiley Interdisciplinary Reviews. RNA* 1 (2): 266–86. doi:10.1002/wrna.31.

Dassi, Erik, Angela Re, Sara Leo, Toma Tebaldi, Luigi Pasini, Daniele Peroni, and Alessandro Quattrone. 2014. "AURA 2: Empowering Discovery of Post-Transcriptional Networks." *Translation (Austin, Tex.)* 2 (1): e27738. doi:10.4161/trla.27738.

Denman, R. B. 1993. "Using RNAFOLD to Predict the Activity of Small Catalytic RNAs." *BioTechniques* 15 (6): 1090–95.

Djebali, Sarah, Carrie A Davis, Angelika Merkel, Alex Dobin, Timo Lassmann, Ali Mortazavi, Andrea Tanzer, et al. 2012. "Landscape of Transcription in Human Cells." *Nature* 489 (7414): 101–8. doi:10.1038/nature11233.

Ellington, Andrew D., and Jack W. Szostak. 1990. "In Vitro Selection of RNA Molecules That Bind Specific Ligands." *Nature* 346 (6287): 818–22. doi:10.1038/346818a0.

Ellis, Jonathan J., Mark Broom, and Susan Jones. 2007. "Protein-RNA Interactions: Structural Analysis and Functional Classes." *Proteins* 66 (4): 903–11. doi:10.1002/prot.21211.

Farnham, Peggy J. 2009. "Insights from Genomic Profiling of Transcription Factors." *Nature Reviews Genetics* 10 (9): 605–16. doi:10.1038/nrg2636.

Fernandez, Michael, Yutaro Kumagai, Daron M Standley, Akinori Sarai, Kenji Mizuguchi, and Shandar Ahmad. 2011. "Prediction of Dinucleotide-Specific RNA-Binding Sites in Proteins." *BMC Bioinformatics* 12 Suppl 13 (November): S5. doi:10.1186/1471-2105-12-S13-S5.

Finn, Robert D., Alex Bateman, Jody Clements, Penelope Coggill, Ruth Y. Eberhardt, Sean R. Eddy, Andreas Heger, et al. 2014. "Pfam: The Protein Families Database." *Nucleic Acids Research* 42 (D1): D222–30. doi:10.1093/nar/gkt1223.

Forslund, Kristoffer, and Erik L. L. Sonnhammer. 2008. "Predicting Protein Function from Domain Content." *Bioinformatics (Oxford, England)* 24 (15): 1681–87. doi:10.1093/bioinformatics/btn312.

Gabdoulline, R., D. Eckweiler, A. Kel, and P. Stegmaier. 2012. "3DTF: A Web Server for Predicting Transcription Factor PWMs Using 3D Structure-Based Energy Calculations." *Nucleic Acids Research* 40 (Web Server issue): W180–85. doi:10.1093/nar/gks551.

Gao, Mu, and Jeffrey Skolnick. 2008. "DBD-Hunter: A Knowledge-Based Method for the Prediction of DNA-Protein Interactions." *Nucleic Acids Research* 36 (12): 3978–92. doi:10.1093/nar/gkn332.

Gao, Mu, and Jeffrey Skolnick. 2009a. "From Nonspecific DNA-Protein Encounter Complexes to the Prediction of DNA-Protein Interactions." *PLoS Computational Biology* 5 (3): e1000341. doi:10.1371/journal.pcbi.1000341.

Gao, Mu, and Jeffrey Skolnick. 2009b. "A Threading-Based Method for the Prediction of DNA-Binding Proteins with Application to the Human Genome." *PLoS Computational Biology* 5 (11): e1000567. doi:10.1371/journal.pcbi.1000567.

Gavrilov, Alexey, Sergey V. Razin, and Giacomo Cavalli. 2014. "In Vivo Formaldehyde Cross-Linking: It Is Time for Black Box Analysis." *Briefings in Functional Genomics*, September, elu037.

doi:10.1093/bfgp/elu037.
Georgi, Benjamin, and Alexander Schliep. 2006. “Context-Specific Independence Mixture Modeling for Positional Weight Matrices.” *Bioinformatics (Oxford, England)* 22 (14): e166–73. doi:10.1093/bioinformatics/btl249.
Gerstberger, Stefanie, Markus Hafner, Manuel Ascano, and Thomas Tuschl. 2014. “Evolutionary Conservation and Expression of Human RNA-Binding Proteins and Their Role in Human Genetic Disease.” *Advances in Experimental Medicine and Biology* 825: 1–55. doi:10.1007/978-1-4939-1221-6_1.
Gerstberger, Stefanie, Markus Hafner, and Thomas Tuschl. 2014. “A Census of Human RNA-Binding Proteins.” *Nature Reviews. Genetics* 15 (12): 829–45. doi:10.1038/nrg3813.
Gerstein, Mark B, Anshul Kundaje, Manoj Hariharan, Stephen G Landt, Koon-Kiu Yan, Chao Cheng, Xinmeng Jasmine Mu, et al. 2012. “Architecture of the Human Regulatory Network Derived from ENCODE Data.” *Nature* 489 (7414): 91–100. doi:10.1038/nature11245.
Gibellini, D., M. Zerbini, M. Musiani, S. Venturoli, G. Gentilomi, and M. La Placa. 1993. “Microplate Capture Hybridization of Amplified Parvovirus B19 DNA Fragment Labelled with Digoxigenin.” *Molecular and Cellular Probes* 7 (6): 453–58. doi:10.1006/mcpr.1993.1067.
Gilfillan, Gregor D., Timothy Hughes, Ying Sheng, Hanne S. Hjorthaug, Tobias Straub, Kristina Gervin, Jennifer R. Harris, Dag E. Undlien, and Robert Lyle. 2012. “Limitations and Possibilities of Low Cell Number ChIP-Seq.” *BMC Genomics* 13: 645. doi:10.1186/1471-2164-13-645.
Gromiha, M. Michael. 2011. *Protein Bioinformatics: From Sequence to Function*. Academic Press.
Grubert, Fabian, Judith B. Zaugg, Maya Kasowski, Oana Ursu, Damek V. Spacek, Alicia R. Martin, Peyton Greenside, et al. 2015. “Genetic Control of Chromatin States in Humans Involves Local and Distal Chromosomal Interactions.” *Cell* 162 (5): 1051–65. doi:10.1016/j.cell.2015.07.048.
Guharoy, Mainak, Kris Pauwels, and Peter Tompa. 2015. “SnapShot: Intrinsic Structural Disorder.” *Cell* 161 (5): 1230–1230.e1. doi:10.1016/j.cell.2015.05.024.
Hanein, Dorit, and Ron Milligan. 2013. “Structural Analysis of Supramolecular Assemblies by Hybrid Methods.” *Journal of Structural Biology* 184 (1): 1. doi:10.1016/j.jsb.2013.09.014.
Han, Lian Yi, Cong Zhong Cai, Siew Lin Lo, Maxey C. M. Chung, and Yu Zong Chen. 2004. “Prediction of RNA-Binding Proteins from Primary Sequence by a Support Vector Machine Approach.” *RNA (New York, N.Y.)* 10 (3): 355–68.
Hannenhalli, Sridhar. 2008. “Eukaryotic Transcription Factor Binding Sites--Modeling and Integrative Search Methods.” *Bioinformatics (Oxford, England)* 24 (11): 1325–31. doi:10.1093/bioinformatics/btn198.
Hannenhalli, Sridhar, and Li-San Wang. 2005. “Enhanced Position Weight Matrices Using Mixture Models.” *Bioinformatics (Oxford, England)* 21 Suppl 1 (June): i204–12. doi:10.1093/bioinformatics/bti1001.
Harrow, Jennifer, Adam Frankish, Jose M Gonzalez, Electra Tapanari, Mark Diekhans, Felix Kokocinski, Bronwen L Aken, et al. 2012.

"GENCODE: The Reference Human Genome Annotation for The ENCODE Project." *Genome Research* 22 (9): 1760–74. doi:10.1101/gr.135350.111.

Hartwell, Leland H., John J. Hopfield, Stanislas Leibler, and Andrew W. Murray. 1999. "From Molecular to Modular Cell Biology." *Nature* 402 (December): C47–52. doi:10.1038/35011540.

Hegde, Muralidhar L., Padmaraju Vasudevaraju, and Kosagisharaf Jagannatha Rao. 2010. "DNA Induced Folding/fibrillation of Alpha-Synuclein: New Insights in Parkinson's Disease." *Frontiers in Bioscience (Landmark Edition)* 15: 418–36.

Hellman, Lance M., and Michael G. Fried. 2007. "Electrophoretic Mobility Shift Assay (EMSA) for Detecting Protein-Nucleic Acid Interactions." *Nature Protocols* 2 (8): 1849–61. doi:10.1038/nprot.2007.249.

He, Qiye, Anaïs F. Bardet, Brianne Patton, Jennifer Purvis, Jeff Johnston, Ariel Paulson, Madelaine Gogol, Alexander Stark, and Julia Zeitlinger. 2011. "High Conservation of Transcription Factor Binding and Evidence for Combinatorial Regulation across Six Drosophila Species." *Nature Genetics* 43 (5): 414–20. doi:10.1038/ng.808.

Hoffman, Elizabeth A., Brian L. Frey, Lloyd M. Smith, and David T. Auble. 2015. "Formaldehyde Crosslinking: A Tool for the Study of Chromatin Complexes." *Journal of Biological Chemistry* 290 (44): 26404–11. doi:10.1074/jbc.R115.651679.

Ho, Joshua W. K., Eric Bishop, Peter V. Karchenko, Nicolas Nègre, Kevin P. White, and Peter J. Park. 2011. "ChIP-Chip versus ChIP-Seq: Lessons for Experimental Design and Data Analysis." *BMC Genomics* 12: 134. doi:10.1186/1471-2164-12-134.

Holoch, Daniel, and Danesh Moazed. 2015. "RNA-Mediated Epigenetic Regulation of Gene Expression." *Nature Reviews Genetics* 16 (2): 71–84. doi:10.1038/nrg3863.

Hornett, Emily A, and Christopher W Wheat. 2012. "Quantitative RNA-Seq Analysis in Non-Model Species: Assessing Transcriptome Assemblies as a Scaffold and the Utility of Evolutionary Divergent Genomic Reference Species." *BMC Genomics* 13 (August): 361. doi:10.1186/1471-2164-13-361.

Huang, Sheng-You, and Xiaoqin Zou. 2013. "A Nonredundant Structure Dataset for Benchmarking Protein-RNA Computational Docking." *Journal of Computational Chemistry* 34 (4): 311–18. doi:10.1002/jcc.23149.

Hubbard, SJ, and JM Thornton. 1993. "'NACCESS', Computer Program."

Huberts, Daphne H. E. W., and Ida J. van der Klei. 2010. "Moonlighting Proteins: An Intriguing Mode of Multitasking." *Biochimica et Biophysica Acta (BBA) - Molecular Cell Research* 1803 (4): 520–25. doi:10.1016/j.bbamcr.2010.01.022.

Hudson, William H., and Eric A. Ortlund. 2014. "The Structure, Function and Evolution of Proteins That Bind DNA and RNA." *Nature Reviews Molecular Cell Biology* 15 (11): 749–60. doi:10.1038/nrm3884.

Hume, Maxwell A., Luis A. Barrera, Stephen S. Gisselbrecht, and Martha L. Bulyk. 2015. "UniPROBE, Update 2015: New Tools and Content for the Online Database of Protein-Binding Microarray Data on Protein-DNA Interactions." *Nucleic Acids Research* 43 (Database issue): D117–22.

doi:10.1093/nar/gku1045.

Huppertz, Ina, Jan Attig, Andrea D'Ambrogio, Laura E Easton, Christopher R Sibley, Yoichiro Sugimoto, Mojca Tajnik, Julian König, and Jernej Ule. 2014. "iCLIP: Protein-RNA Interactions at Nucleotide Resolution." *Methods (San Diego, Calif.)* 65 (3): 274–87. doi:10.1016/j.ymeth.2013.10.011.

Hwang, Seungwoo, Zhenkun Gou, and Igor B. Kuznetsov. 2007. "DP-Bind: A Web Server for Sequence-Based Prediction of DNA-Binding Residues in DNA-Binding Proteins." *Bioinformatics (Oxford, England)* 23 (5): 634–36. doi:10.1093/bioinformatics/btl672.

Jackson, V. 1978. "Studies on Histone Organization in the Nucleosome Using Formaldehyde as a Reversible Cross-Linking Agent." *Cell* 15 (3): 945–54.

Jain, Ritu, Tiffany Devine, Ajish D. George, Sridar V. Chittur, Timothy E. Baroni, Luiz O. Penalva, and Scott A. Tenenbaum. 2011. "RIP-Chip Analysis: RNA-Binding Protein Immunoprecipitation-Microarray (Chip) Profiling." *Methods in Molecular Biology (Clifton, N.J.)* 703: 247–63. doi:10.1007/978-1-59745-248-9_17.

Janin, Joël. 2010. "Protein-Protein Docking Tested in Blind Predictions: The CAPRI Experiment." *Molecular bioSystems* 6 (12): 2351–62. doi:10.1039/c005060c.

Johnson, David S., Ali Mortazavi, Richard M. Myers, and Barbara Wold. 2007. "Genome-Wide Mapping of in Vivo Protein-DNA Interactions." *Science (New York, N.Y.)* 316 (5830): 1497–1502. doi:10.1126/science.1141319.

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

Jones, Susan, David T. A. Daley, Nicholas M. Luscombe, Helen M. Berman, and Janet M. Thornton. 2001. "Protein–RNA Interactions: A Structural Analysis." *Nucleic Acids Research* 29 (4): 943–54.

Jones, Susan, Paul van Heyningen, Helen M. Berman, and Janet M. Thornton. 1999. "Protein-DNA Interactions: A Structural analysis1." *Journal of Molecular Biology* 287 (5): 877–96. doi:10.1006/jmbi.1999.2659.

Karplus, Martin, and Richard Lavery. 2014. "Significance of Molecular Dynamics Simulations for Life Sciences." *Israel Journal of Chemistry* 54 (8-9): 1042–51. doi:10.1002/ijch.201400074.

Kazan, Hilal, Debashish Ray, Esther T. Chan, Timothy R. Hughes, and Quaid Morris. 2010. "RNAcontext: A New Method for Learning the Sequence and Structure Binding Preferences of RNA-Binding Proteins." *PLoS Comput Biol* 6 (7): e1000832. doi:10.1371/journal.pcbi.1000832.

Keene, Jack D, Jordan M Komisarow, and Matthew B Friedersdorf. 2006. "RIP-Chip: The Isolation and Identification of mRNAs, microRNAs and Protein Components of Ribonucleoprotein Complexes from Cell Extracts." *Nature Protocols* 1 (1): 302–7. doi:10.1038/nprot.2006.47.

Keilwagen, Jens, and Jan Grau. 2015. "Varying Levels of Complexity in Transcription Factor Binding Motifs." *Nucleic Acids Research*, June, gkv577. doi:10.1093/nar/gkv577.

Kel, Alexander, Yury Tikunov, Nico Voss, Jürgen Borlak, and Edgar Wingender. 2004. "Application of Kernel Method to Reveal Subtypes of TF Binding

Motifs." In *Regulatory Genomics*, edited by Eleazar Eskin and Christopher Workman, 42–51. Lecture Notes in Computer Science 3318. Springer Berlin Heidelberg. http://link.springer.com/chapter/10.1007/978-3-540-32280-1_5.

Kersey, Paul Julian, James E. Allen, Irina Armean, Sanjay Boddu, Bruce J. Bolt, Denise Carvalho-Silva, Mikkel Christensen, et al. 2016. "Ensembl Genomes 2016: More Genomes, More Complexity." *Nucleic Acids Research* 44 (D1): D574–80. doi:10.1093/nar/gkv1209.

Khorshid, Mohsen, Christoph Rodak, and Mihaela Zavolan. 2011. "CLIPZ: A Database and Analysis Environment for Experimentally Determined Binding Sites of RNA-Binding Proteins." *Nucleic Acids Research* 39 (Database issue): D245–52. doi:10.1093/nar/gkq940.

Kiliç, Sefa, Elliot R. White, Dinara M. Sagitova, Joseph P. Cornish, and Ivan Erill. 2014. "CollecTF: A Database of Experimentally Validated Transcription Factor-Binding Sites in Bacteria." *Nucleic Acids Research* 42 (Database issue): D156–60. doi:10.1093/nar/gkt1123.

Kim, Oanh T P, Kei Yura, and Nobuhiro Go. 2006. "Amino Acid Residue Doublet Propensity in the Protein-RNA Interface and Its Application to RNA Interface Prediction." *Nucleic Acids Research* 34 (22): 6450–60. doi:10.1093/nar/gkl819.

Kim, RyangGuk, and Jun-tao Guo. 2009. "PDA: An Automatic and Comprehensive Analysis Program for Protein-DNA Complex Structures." *BMC Genomics* 10 Suppl 1: S13. doi:10.1186/1471-2164-10-S1-S13.

Kirsanov, Dmitry D., Olga N. Zanegina, Evgeniy A. Aksianov, Sergei A. Spirin, Anna S. Karyagina, and Andrei V. Alexeevski. 2013. "NPIDB: Nucleic Acid-Protein Interaction DataBase." *Nucleic Acids Research* 41 (Database issue): D517–23. doi:10.1093/nar/gks1199.

Klus, Petr, Riccardo Delli Ponti, Carmen Maria Livi, and Gian Gaetano Tartaglia. 2015. "Protein Aggregation, Structural Disorder and RNA-Binding Ability: A New Approach for Physico-Chemical and Gene Ontology Classification of Multiple Datasets." *BMC Genomics* 16 (1): 1071. doi:10.1186/s12864-015-2280-z.

Konig, Julian, Kathi Zarnack, Gregor Rot, Tomaz Curk, Melis Kayikci, Blaz Zupan, Daniel J. Turner, Nicholas M. Luscombe, and Jernej Ule. 2011. "iCLIP--Transcriptome-Wide Mapping of Protein-RNA Interactions with Individual Nucleotide Resolution." *Journal of Visualized Experiments: JoVE*, no. 50. doi:10.3791/2638.

Krecic, A. M., and M. S. Swanson. 1999. "hnRNP Complexes: Composition, Structure, and Function." *Current Opinion in Cell Biology* 11 (3): 363–71. doi:10.1016/S0955-0674(99)80051-9.

Krippahl, Ludwig, and Pedro Barahona. 2015. "Protein Docking with Predicted Constraints." *Algorithms for Molecular Biology : AMB* 10 (February). doi:10.1186/s13015-015-0036-6.

Kulakovskiy, Ivan V., Ilya E. Vorontsov, Ivan S. Yevshin, Anastasiia V. Soboleva, Artem S. Kasianov, Haitham Ashoor, Wail Ba-alawi, et al. 2016. "HOCOMOCO: Expansion and Enhancement of the Collection of Transcription Factor Binding Sites Models." *Nucleic Acids Research* 44 (D1): D116–25. doi:10.1093/nar/gkv1249.

Kumar, Manish, M Michael Gromiha, and Gajendra P S Raghava. 2011. "SVM Based Prediction of RNA-Binding Proteins Using Binding Residues and Evolutionary Information." *Journal of Molecular Recognition: JMR* 24 (2): 303–13. doi:10.1002/jmr.1061.

Kumar, Manish, M. Michael Gromiha, and G. P. S. Raghava. 2008. "Prediction of RNA Binding Sites in a Protein Using SVM and PSSM Profile." *Proteins* 71 (1): 189–94. doi:10.1002/prot.21677.

Kumar, M. D. Shaji, K. Abdulla Bava, M. Michael Gromiha, Ponraj Prabakaran, Koji Kitajima, Hatsuho Uedaira, and Akinori Sarai. 2006. "ProTherm and ProNIT: Thermodynamic Databases for Proteins and Protein-Nucleic Acid Interactions." *Nucleic Acids Research* 34 (Database issue): D204–6. doi:10.1093/nar/gkj103.

K, Usha. 2013. "Computational Tools for Investigating RNA-Protein Interaction Partners." *Journal of Computer Science & Systems Biology* 06 (04). doi:10.4172/jcsb.1000115.

Kwon, S. Chul, Hyerim Yi, Katrin Eichelbaum, Sophia Föhr, Bernd Fischer, Kwon Tae You, Alfredo Castello, Jeroen Krijgsveld, Matthias W. Hentze, and V. Narry Kim. 2013. "The RNA-Binding Protein Repertoire of Embryonic Stem Cells." *Nature Structural & Molecular Biology* 20 (9): 1122–30. doi:10.1038/nsmb.2638.

Lambert, Nicole, Alex Robertson, Mohini Jangi, Sean McGeary, Phillip A. Sharp, and Christopher B. Burge. 2014. "RNA Bind-N-Seq: Quantitative Assessment of the Sequence and Structural Binding Specificity of RNA Binding Proteins." *Molecular Cell* 54 (5): 887–900. doi:10.1016/j.molcel.2014.04.016.

Landenmark, Hanna K. E., Duncan H. Forgan, and Charles S. Cockell. 2015. "An Estimate of the Total DNA in the Biosphere." *PLoS Biol* 13 (6): e1002168. doi:10.1371/journal.pbio.1002168.

Landt, Stephen G., Georgi K. Marinov, Anshul Kundaje, Pouya Kheradpour, Florencia Pauli, Serafim Batzoglou, Bradley E. Bernstein, et al. 2012. "ChIP-Seq Guidelines and Practices of the ENCODE and modENCODE Consortia." *Genome Research* 22 (9): 1813–31. doi:10.1101/gr.136184.111.

Larance, Mark, and Angus I. Lamond. 2015. "Multidimensional Proteomics for Cell Biology." *Nature Reviews Molecular Cell Biology* 16 (5): 269–80. doi:10.1038/nrm3970.

Lawrence, Michael C., and Peter M. Colman. 1993. "Shape Complementarity at Protein/Protein Interfaces." *Journal of Molecular Biology* 234 (4): 946–50. doi:10.1006/jmbi.1993.1648.

Lee, Semin, and Tom L. Blundell. 2009. "BIPA: A Database for Protein-Nucleic Acid Interaction in 3D Structures." *Bioinformatics (Oxford, England)* 25 (12): 1559–60. doi:10.1093/bioinformatics/btp243.

Levine, Michael, and Robert Tjian. 2003. "Transcription Regulation and Animal Diversity." *Nature* 424 (6945): 147–51.

Levo, Michal, and Eran Segal. 2014. "In Pursuit of Design Principles of Regulatory Sequences." *Nature Reviews Genetics* 15 (7): 453–68. doi:10.1038/nrg3684.

Lewis, Benjamin A., Rasna R. Walia, Michael Terribilini, Jeff Ferguson, Charles Zheng, Vasant Honavar, and Drena Dobbs. 2011. "PRIDB: A protein–

RNA Interface Database." *Nucleic Acids Research* 39 (Database issue): D277–82. doi:10.1093/nar/gkq1108.

Lichty, Jordan J., Joshua L. Malecki, Heather D. Agnew, Daniel J. Michelson-Horowitz, and Song Tan. 2005. "Comparison of Affinity Tags for Protein Purification." *Protein Expression and Purification* 41 (1): 98–105. doi:10.1016/j.pep.2005.01.019.

Lieb, J. D., X. Liu, D. Botstein, and P. O. Brown. 2001. "Promoter-Specific Binding of Rap1 Revealed by Genome-Wide Maps of Protein-DNA Association." *Nature Genetics* 28 (4): 327–34. doi:10.1038/ng569.

Lihu, Andrei, and Ştefan Holban. 2015. "A Review of Ensemble Methods for de Novo Motif Discovery in ChIP-Seq Data." *Briefings in Bioinformatics* 16 (6): 964–73. doi:10.1093/bib/bbv022.

Lin, Xin, Leïla Tirichine, and Chris Bowler. 2012. "Protocol: Chromatin Immunoprecipitation (ChIP) Methodology to Investigate Histone Modifications in Two Model Diatom Species." *Plant Methods* 8 (December): 48. doi:10.1186/1746-4811-8-48.

Li, Quan, Zanxia Cao, and Haiyan Liu. 2010. "Improve the Prediction of RNA-Binding Residues Using Structural Neighbours." *Protein and Peptide Letters* 17 (3): 287–96.

Li, Tao, Qian-Zhong Li, Shuai Liu, Guo-Liang Fan, Yong-Chun Zuo, and Yong Peng. 2013. "PreDNA: Accurate Prediction of DNA-Binding Sites in Proteins by Integrating Sequence and Geometric Structure Information." *Bioinformatics (Oxford, England)* 29 (6): 678–85. doi:10.1093/bioinformatics/btt029.

Liu, Bin, Jinghao Xu, Xun Lan, Ruifeng Xu, Jiyun Zhou, Xiaolong Wang, and Kuo-Chen Chou. 2014. "iDNA-Prot|dis: Identifying DNA-Binding Proteins by Incorporating Amino Acid Distance-Pairs and Reduced Alphabet Profile into the General Pseudo Amino Acid Composition." *PloS One* 9 (9): e106691. doi:10.1371/journal.pone.0106691.

Liu, Rong, and Jianjun Hu. 2013. "DNABind: A Hybrid Algorithm for Structure-Based Prediction of DNA-Binding Residues by Combining Machine Learning- and Template-Based Approaches." *Proteins* 81 (11): 1885–99. doi:10.1002/prot.24330.

Liu, Zhi-Ping, Ling-Yun Wu, Yong Wang, Xiang-Sun Zhang, and Luonan Chen. 2010. "Prediction of Protein-RNA Binding Sites by a Random Forest Method with Combined Features." *Bioinformatics (Oxford, England)* 26 (13): 1616–22. doi:10.1093/bioinformatics/btq253.

Livi, Carmen Maria, Petr Klus, Riccardo Delli Ponti, and Gian Gaetano Tartaglia. 2015. "catRAPID Signature: Identification of Ribonucleoproteins and RNA-Binding Regions." *Bioinformatics (Oxford, England)*, October. doi:10.1093/bioinformatics/btv629.

Livi, Carmen M, and Enrico Blanzieri. 2014. "Protein-Specific Prediction of mRNA Binding Using RNA Sequences, Binding Motifs and Predicted Secondary Structures." *BMC Bioinformatics* 15 (April): 123. doi:10.1186/1471-2105-15-123.

Li, Xiao, Gerald Quon, Howard D. Lipshitz, and Quaid Morris. 2010. "Predicting in Vivo Binding Sites of RNA-Binding Proteins Using mRNA Secondary Structure." *RNA* 16 (6): 1096–1107. doi:10.1261/rna.2017210.

Lunde, Bradley M., Claire Moore, and Gabriele Varani. 2007. "RNA-Binding Proteins: Modular Design for Efficient Function." *Nature Reviews. Molecular Cell Biology* 8 (6): 479–90. doi:10.1038/nrm2178.

Luo, Zhenhua, Zhiming Dai, Xiaowei Xie, Xuyang Feng, Dan Liu, Zhou Songyang, and Yuanyan Xiong. 2015. "TeloPIN: A Database of Telomeric Proteins Interaction Network in Mammalian Cells." *Database: The Journal of Biological Databases and Curation* 2015. doi:10.1093/database/bav018.

Luscombe, Nicholas M., Roman A. Laskowski, and Janet M. Thornton. 2001. "Amino Acid–base Interactions: A Three-Dimensional Analysis of protein–DNA Interactions at an Atomic Level." *Nucleic Acids Research* 29 (13): 2860–74.

Luscombe, N. M., S. E. Austin, H. M. Berman, and J. M. Thornton. 2000. "An Overview of the Structures of Protein-DNA Complexes." *Genome Biology* 1 (1): REVIEWS001. doi:10.1186/gb-2000-1-1-reviews001.

MacKerell, Alexander D., and Lennart Nilsson. 2008. "Molecular Dynamics Simulations of Nucleic Acid-Protein Complexes." *Current Opinion in Structural Biology* 18 (2): 194–99. doi:10.1016/j.sbi.2007.12.012.

MacQuarrie, Kyle L, Abraham P Fong, Randall H Morse, and Stephen J Tapscott. 2011. "Genome-Wide Transcription Factor Binding: Beyond Direct Target Regulation." *Trends in Genetics: TIG* 27 (4): 141–48. doi:10.1016/j.tig.2011.01.001.

Maetschke, Stefan R, and Zheng Yuan. 2009. "Exploiting Structural and Topological Information to Improve Prediction of RNA-Protein Binding Sites." *BMC Bioinformatics* 10: 341. doi:10.1186/1471-2105-10-341.

Maloney, Bryan, and Debomoy K. Lahiri. 2011. "The Alzheimer's Amyloid β-Peptide (Aβ) Binds a Specific DNA Aβ-Interacting Domain (AβID) in the APP, BACE1, and APOE Promoters in a Sequence-Specific Manner: Characterizing a New Regulatory Motif." *Gene* 488 (1-2): 1–12. doi:10.1016/j.gene.2011.06.004.

Manke, T., R. Bringas, and M. Vingron. 2003. "Correlating Protein-DNA and Protein-Protein Interaction Networks." *Journal of Molecular Biology* 333 (1): 75–85. doi:10.1016/j.jmb.2003.08.004.

Marion, Dominique. 2013. "An Introduction to Biological NMR Spectroscopy." *Molecular & Cellular Proteomics*, July, mcp.O113.030239. doi:10.1074/mcp.O113.030239.

Mascareñas, José L. 2008. "Protein–Nucleic Acid Interactions: Structural Biology. Edited by Phoebe A. Rice, and Carl C. Correll." *ChemBioChem* 9 (13): 2162–63. doi:10.1002/cbic.200800519.

Massie, Charles E., and Ian G. Mills. 2012. "Mapping Protein-DNA Interactions Using ChIP-Sequencing." *Methods in Molecular Biology (Clifton, N.J.)* 809: 157–73. doi:10.1007/978-1-61779-376-9_11.

Mathelier, Anthony, Oriol Fornes, David J. Arenillas, Chih-Yu Chen, Grégoire Denay, Jessica Lee, Wenqiang Shi, et al. 2016. "JASPAR 2016: A Major Expansion and Update of the Open-Access Database of Transcription Factor Binding Profiles." *Nucleic Acids Research* 44 (D1): D110–15. doi:10.1093/nar/gkv1176.

Ma, Xin, Jing Guo, Hong-De Liu, Jian-Ming Xie, and Xiao Sun. 2012. "Sequence-Based Prediction of DNA-Binding Residues in Proteins with

Conservation and Correlation Information." *IEEE/ACM Transactions on Computational Biology and Bioinformatics / IEEE, ACM* 9 (6): 1766–75. doi:10.1109/TCBB.2012.106.

Ma, Xin, Jing Guo, Jiansheng Wu, Hongde Liu, Jiafeng Yu, Jianming Xie, and Xiao Sun. 2011. "Prediction of RNA-Binding Residues in Proteins from Primary Sequence Using an Enriched Random Forest Model with a Novel Hybrid Feature." *Proteins* 79 (4): 1230–39. doi:10.1002/prot.22958.

Miao, Zhichao, and Eric Westhof. 2015. "A Large-Scale Assessment of Nucleic Acids Binding Site Prediction Programs." *PLoS Computational Biology* 11 (12). doi:10.1371/journal.pcbi.1004639.

Milek, Miha, Emanuel Wyler, and Markus Landthaler. 2012. "Transcriptome-Wide Analysis of Protein-RNA Interactions Using High-Throughput Sequencing." *Seminars in Cell & Developmental Biology* 23 (2): 206–12. doi:10.1016/j.semcdb.2011.12.001.

Mili, Stavroula, and Joan A. Steitz. 2004. "Evidence for Reassociation of RNA-Binding Proteins after Cell Lysis: Implications for the Interpretation of Immunoprecipitation Analyses." *RNA* 10 (11): 1692–94. doi:10.1261/rna.7151404.

Mitchell, Alex, Hsin-Yu Chang, Louise Daugherty, Matthew Fraser, Sarah Hunter, Rodrigo Lopez, Craig McAnulla, et al. 2015. "The InterPro Protein Families Database: The Classification Resource after 15 Years." *Nucleic Acids Research* 43 (Database issue): D213–21. doi:10.1093/nar/gku1243.

Mitchell, Tom M. 1982. "Generalization as Search." *Artificial Intelligence* 18 (2): 203–26. doi:10.1016/0004-3702(82)90040-6.

Moult, John. 2008. "Comparative Modeling in Structural Genomics." *Structure* 16 (1): 14–16. doi:10.1016/j.str.2007.12.001.

Mukherjee, Rithun, Perry Evans, Larry N. Singh, and Sridhar Hannenhalli. 2013. "Correlated Evolution of Positions within Mammalian Cis Elements." *PLoS ONE* 8 (2): e55521. doi:10.1371/journal.pone.0055521.

Muppirala, Usha K, Vasant G Honavar, and Drena Dobbs. 2011. "Predicting RNA-Protein Interactions Using Only Sequence Information." *BMC Bioinformatics* 12: 489. doi:10.1186/1471-2105-12-489.

Murakami, Yoichi, Ruth V. Spriggs, Haruki Nakamura, and Susan Jones. 2010. "PiRaNhA: A Server for the Computational Prediction of RNA-Binding Residues in Protein Sequences." *Nucleic Acids Research* 38 (Web Server issue): W412–16. doi:10.1093/nar/gkq474.

Nadassy, Katalin, Shoshana J. Wodak, and Joël Janin. 1999. "Structural Features of Protein−Nucleic Acid Recognition Sites." *Biochemistry* 38 (7): 1999–2017. doi:10.1021/bi982362d.

Nagarajan, R., Shandar Ahmad, and M. Michael Gromiha. 2013. "Novel Approach for Selecting the Best Predictor for Identifying the Binding Sites in DNA Binding Proteins." *Nucleic Acids Research* 41 (16): 7606–14. doi:10.1093/nar/gkt544.

Narlikar, Leelavati. 2013. "MuMoD: A Bayesian Approach to Detect Multiple Modes of Protein-DNA Binding from Genome-Wide ChIP Data." *Nucleic Acids Research* 41 (1): 21–32. doi:10.1093/nar/gks950.

NCBI Resource Coordinators. 2016. "Database Resources of the National Center

for Biotechnology Information." *Nucleic Acids Research* 44 (D1): D7–19. doi:10.1093/nar/gkv1290.

Nimrod, Guy, Maya Schushan, András Szilágyi, Christina Leslie, and Nir Ben-Tal. 2010. "iDBPs: A Web Server for the Identification of DNA Binding Proteins." *Bioinformatics (Oxford, England)* 26 (5): 692–93. doi:10.1093/bioinformatics/btq019.

Ofran, Yanay, Venkatesh Mysore, and Burkhard Rost. 2007. "Prediction of DNA-Binding Residues from Sequence." *Bioinformatics (Oxford, England)* 23 (13): i347–53. doi:10.1093/bioinformatics/btm174.

Orengo, Christine A., and Janet M. Thornton. 2005. "Protein Families and Their Evolution-a Structural Perspective." *Annual Review of Biochemistry* 74: 867–900. doi:10.1146/annurev.biochem.74.082803.133029.

Osada, Robert, Elena Zaslavsky, and Mona Singh. 2004. "Comparative Analysis of Methods for Representing and Searching for Transcription Factor Binding Sites." *Bioinformatics (Oxford, England)* 20 (18): 3516–25. doi:10.1093/bioinformatics/bth438.

Ozbek, Pemra, Seren Soner, Burak Erman, and Turkan Haliloglu. 2010. "DNABINDPROT: Fluctuation-Based Predictor of DNA-Binding Residues within a Network of Interacting Residues." *Nucleic Acids Research* 38 (Web Server issue): W417–23. doi:10.1093/nar/gkq396.

Pankavich, Stephen, and Peter Ortoleva. 2015. "A Review of Two Multiscale Methods for the Simulation of Macromolecular Assemblies: Multiscale Perturbation and Multiscale Factorization." *Computation* 3 (1): 29–57. doi:10.3390/computation3010029.

Parca, Luca, Fabrizio Ferré, Gabriele Ausiello, and Manuela Helmer-Citterich. 2013. "Nucleos: A Web Server for the Identification of Nucleotide-Binding Sites in Protein Structures." *Nucleic Acids Research* 41 (Web Server issue): W281–85. doi:10.1093/nar/gkt390.

Pérez-Cano, Laura, Brian Jiménez-García, and Juan Fernández-Recio. 2012. "A Protein-RNA Docking Benchmark (II): Extended Set from Experimental and Homology Modeling Data." *Proteins* 80 (7): 1872–82. doi:10.1002/prot.24075.

Pérez-Cano, Laura, Albert Solernou, Carles Pons, and Juan Fernández-Recio. 2010. "Structural Prediction of Protein-RNA Interaction by Computational Docking with Propensity-Based Statistical Potentials." *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, 293–301.

Petrov, Anton I., Craig L. Zirbel, and Neocles B. Leontis. 2013. "Automated Classification of RNA 3D Motifs and the RNA 3D Motif Atlas." *RNA* 19 (10): 1327–40. doi:10.1261/rna.039438.113.

Pfreundt, Ulrike, Daniel P. James, Susan Tweedie, Derek Wilson, Sarah A. Teichmann, and Boris Adryan. 2010. "FlyTF: Improved Annotation and Enhanced Functionality of the Drosophila Transcription Factor Database." *Nucleic Acids Research* 38 (Database issue): D443–47. doi:10.1093/nar/gkp910.

Pillai, Smitha, Piyali Dasgupta, and SrikumarP. Chellappan. 2015. "Chromatin Immunoprecipitation Assays: Analyzing Transcription Factor Binding and Histone Modifications In Vivo." In *Chromatin Protocols*, edited by Srikumar P. Chellappan, 429–46. Methods in Molecular Biology 1288.

Springer New York. http://dx.doi.org/10.1007/978-1-4939-2474-5_25.
Ponting, Chris P., and Robert R. Russell. 2002. "The Natural History of Protein Domains." *Annual Review of Biophysics and Biomolecular Structure* 31 (1): 45–71. doi:10.1146/annurev.biophys.31.082901.134314.
Prabakaran, Ponraj, Jörg G. Siebers, Shandar Ahmad, M. Michael Gromiha, Maria G. Singarayan, and Akinori Sarai. 2006. "Classification of Protein-DNA Complexes Based on Structural Descriptors." *Structure* 14 (9): 1355–67. doi:10.1016/j.str.2006.06.018.
Pradhan, Lagnajeet, and Hyun Joo Nam. 2015. "NuProPlot: Nucleic Acid and Protein Interaction Analysis and Plotting Program." *Acta Crystallographica. Section D, Biological Crystallography* 71 (Pt 3): 667–74. doi:10.1107/S1399004715000139.
Quader, Saad, and Chun-Hsi Huang. 2012. "Effect of Positional Dependence and Alignment Strategy on Modeling Transcription Factor Binding Sites." *BMC Research Notes* 5: 340. doi:10.1186/1756-0500-5-340.
Ranea, Juan A. G., Antonio Sillero, Janet M. Thornton, and Christine A. Orengo. 2006. "Protein Superfamily Evolution and the Last Universal Common Ancestor (LUCA)." *Journal of Molecular Evolution* 63 (4): 513–25. doi:10.1007/s00239-005-0289-7.
Rao, V. Srinivasa, K. Srinivas, G. N. Sujini, G. N. Sunand Kumar, V. Srinivasa Rao, K. Srinivas, G. N. Sujini, and G. N. Sunand Kumar. 2014. "Protein-Protein Interaction Detection: Methods and Analysis, Protein-Protein Interaction Detection: Methods and Analysis." *International Journal of Proteomics, International Journal of Proteomics* 2014, 2014 (February): e147648. doi:10.1155/2014/147648, 10.1155/2014/147648.
Ray, Debashish, Hilal Kazan, Esther T. Chan, Lourdes Peña Castillo, Sidharth Chaudhry, Shaheynoor Talukder, Benjamin J. Blencowe, Quaid Morris, and Timothy R. Hughes. 2009. "Rapid and Systematic Analysis of the RNA Recognition Specificities of RNA-Binding Proteins." *Nature Biotechnology* 27 (7): 667–70. doi:10.1038/nbt.1550.
Ray, Debashish, Hilal Kazan, Kate B Cook, Matthew T Weirauch, Hamed S Najafabadi, Xiao Li, Serge Gueroussov, et al. 2013. "A Compendium of RNA-Binding Motifs for Decoding Gene Regulation." *Nature* 499 (7457): 172–77. doi:10.1038/nature12311.
Re, Angela, Tejal Joshi, Eleonora Kulberkyte, Quaid Morris, and Christopher T. Workman. 2014. "RNA-Protein Interactions: An Overview." *Methods in Molecular Biology (Clifton, N.J.)* 1097: 491–521. doi:10.1007/978-1-62703-709-9_23.
Regenmortel, Marc H.V. Van. 2004. "Reductionism and Complexity in Molecular Biology." *EMBO Reports* 5 (11): 1016–20. doi:10.1038/sj.embor.7400284.
Ren, B., F. Robert, J. J. Wyrick, O. Aparicio, E. G. Jennings, I. Simon, J. Zeitlinger, et al. 2000. "Genome-Wide Location and Function of DNA Binding Proteins." *Science (New York, N.Y.)* 290 (5500): 2306–9. doi:10.1126/science.290.5500.2306.
Rhee, Ho Sung, and B. Franklin Pugh. 2012. "ChIP-Exo Method for Identifying Genomic Location of DNA-Binding Proteins with near-Single-Nucleotide Accuracy." *Current Protocols in Molecular Biology / Edited by Frederick M. Ausubel ... [et Al.]* Chapter 21 (October): Unit 21.24.

doi:10.1002/0471142727.mb2124s100.
Richardson, J. S. 1981. “The Anatomy and Taxonomy of Protein Structure.” *Advances in Protein Chemistry* 34: 167–339.
Riley, Kasandra J., and Joan A. Steitz. 2013. “The ‘Observer Effect’ in Genome-Wide Surveys of Protein-RNA Interactions.” *Molecular Cell* 49 (4): 601–4. doi:10.1016/j.molcel.2013.01.030.
Rinn, John L., and Jernej Ule. 2014. “ ’Oming in on RNA–protein Interactions.” *Genome Biology* 15: 401. doi:10.1186/gb4158.
Robison, K., A. M. McGuire, and G. M. Church. 1998. “A Comprehensive Library of DNA-Binding Site Matrices for 55 Proteins Applied to the Complete Escherichia Coli K-12 Genome.” *Journal of Molecular Biology* 284 (2): 241–54. doi:10.1006/jmbi.1998.2160.
Sajan, Samin A., and R. David Hawkins. 2012. “Methods for Identifying Higher-Order Chromatin Structure.” *Annual Review of Genomics and Human Genetics* 13: 59–82. doi:10.1146/annurev-genom-090711-163818.
Samant, Monika, Minesh Jethva, and Yasha Hasija. 2014. “INTERACT-O-FINDER: A Tool for Prediction of DNA-Binding Proteins Using Sequence Features.” *International Journal of Peptide Research and Therapeutics* 21 (2): 189–93. doi:10.1007/s10989-014-9446-4.
Schneider, T. D., and R. M. Stephens. 1990. “Sequence Logos: A New Way to Display Consensus Sequences.” *Nucleic Acids Research* 18 (20): 6097–6100.
Shazman, Shula, Gershon Celniker, Omer Haber, Fabian Glaser, and Yael Mandel-Gutfreund. 2007. “Patch Finder Plus (PFplus): A Web Server for Extracting and Displaying Positive Electrostatic Patches on Protein Surfaces.” *Nucleic Acids Research* 35 (Web Server issue): W526–30. doi:10.1093/nar/gkm401.
Shazman, Shula, and Yael Mandel-Gutfreund. 2008. “Classifying RNA-Binding Proteins Based on Electrostatic Properties.” *PLoS Computational Biology* 4 (8): e1000146. doi:10.1371/journal.pcbi.1000146.
Shi, Yigong. 2014. “A Glimpse of Structural Biology through X-Ray Crystallography.” *Cell* 159 (5): 995–1014. doi:10.1016/j.cell.2014.10.051.
Si, Jingna, Jing Cui, Jin Cheng, and Rongling Wu. 2015. “Computational Prediction of RNA-Binding Proteins and Binding Sites.” *International Journal of Molecular Sciences* 16 (11): 26303–17. doi:10.3390/ijms161125952.
Si, Jingna, Zengming Zhang, Biaoyang Lin, Michael Schroeder, and Bingding Huang. 2011. “MetaDBSite: A Meta Approach to Improve Protein DNA-Binding Sites Prediction.” *BMC Systems Biology* 5 Suppl 1: S7. doi:10.1186/1752-0509-5-S1-S7.
Simon, Matthew D, Charlotte I Wang, Peter V Kharchenko, Jason A West, Brad A Chapman, Artyom A Alekseyenko, Mark L Borowsky, Mitzi I Kuroda, and Robert E Kingston. 2011. “The Genomic Binding Sites of a Noncoding RNA.” *Proceedings of the National Academy of Sciences of the United States of America* 108 (51): 20497–502. doi:10.1073/pnas.1113536108.
Spitale, Robert C., Ryan A. Flynn, Qiangfeng Cliff Zhang, Pete Crisalli, Byron Lee, Jong-Wha Jung, Hannes Y. Kuchelmeister, et al. 2015. “Structural

Imprints in Vivo Decode RNA Regulatory Mechanisms." *Nature* 519 (7544): 486–90. doi:10.1038/nature14263.

Spitzer, Jessica, Markus Hafner, Markus Landthaler, Manuel Ascano, Thalia Farazi, Greg Wardle, Jeff Nusbaum, et al. 2014. "PAR-CLIP (Photoactivatable Ribonucleoside-Enhanced Crosslinking and Immunoprecipitation): A Step-by-Step Protocol to the Transcriptome-Wide Identification of Binding Sites of RNA-Binding Proteins." *Methods in Enzymology* 539: 113–61. doi:10.1016/B978-0-12-420120-0.00008-6.

Spitz, François, and Eileen E. M. Furlong. 2012. "Transcription Factors: From Enhancer Binding to Developmental Control." *Nature Reviews. Genetics* 13 (9): 613–26. doi:10.1038/nrg3207.

Spivak, Aaron T., and Gary D. Stormo. 2012. "ScerTF: A Comprehensive Database of Benchmarked Position Weight Matrices for Saccharomyces Species." *Nucleic Acids Research* 40 (Database issue): D162–68. doi:10.1093/nar/gkr1180.

Srihari, Sriganesh, Chern Han Yong, Ashwini Patil, and Limsoon Wong. 2015. "Methods for Protein Complex Prediction and Their Contributions towards Understanding the Organisation, Function and Dynamics of Complexes." *FEBS Letters* 589 (19 Pt A): 2590–2602. doi:10.1016/j.febslet.2015.04.026.

Stampfel, Gerald, Tomáš Kazmar, Olga Frank, Sebastian Wienerroither, Franziska Reiter, and Alexander Stark. 2015. "Transcriptional Regulators Form Diverse Groups with Context-Dependent Regulatory Functions." *Nature* 528 (7580): 147–51. doi:10.1038/nature15545.

Steffens, Nils Ole, Claudia Galuschka, Martin Schindler, Lorenz Bülow, and Reinhard Hehl. 2005. "AthaMap Web Tools for Database-Assisted Identification of Combinatorial Cis-Regulatory Elements and the Display of Highly Conserved Transcription Factor Binding Sites in Arabidopsis Thaliana." *Nucleic Acids Research* 33 (Web Server issue): W397–402. doi:10.1093/nar/gki395.

Stormo, G D. 2000. "DNA Binding Sites: Representation and Discovery." *Bioinformatics (Oxford, England)* 16 (1): 16–23.

Stražar, Martin, Marinka Žitnik, Blaž Zupan, Jernej Ule, and Tomaž Curk. 2016. "Orthogonal Matrix Factorization Enables Integrative Analysis of Multiple RNA Binding Proteins." *Bioinformatics (Oxford, England)*, January. doi:10.1093/bioinformatics/btw003.

Sundararaman, Balaji, Lijun Zhan, Steven Blue, Rebecca Stanton, Keri Elkins, Sara Olson, Xintao Wei, et al. 2015. "Resources for the Comprehensive Discovery of Functional RNA Elements." *bioRxiv*, November, 030486. doi:10.1101/030486.

Terribilini, Michael, Jeffry D Sander, Jae-Hyung Lee, Peter Zaback, Robert L Jernigan, Vasant Honavar, and Drena Dobbs. 2007. "RNABindR: A Server for Analyzing and Predicting RNA-Binding Sites in Proteins." *Nucleic Acids Research* 35 (Web Server issue): W578–84. doi:10.1093/nar/gkm294.

Tjong, Harianto, and Huan-Xiang Zhou. 2007. "DISPLAR: An Accurate Method for Predicting DNA-Binding Sites on Protein Surfaces." *Nucleic Acids Research* 35 (5): 1465–77. doi:10.1093/nar/gkm008.

Tong, Jing, Peng Jiang, and Zu-Hong Lu. 2008. “RISP: A Web-Based Server for Prediction of RNA-Binding Sites in Proteins.” *Computer Methods and Programs in Biomedicine* 90 (2): 148–53. doi:10.1016/j.cmpb.2007.12.003.

Towfic, Fadi, Cornelia Caragea, David C Gemperline, Drena Dobbs, and Vasant Honavar. 2010. “Struct-NB: Predicting Protein-RNA Binding Sites Using Structural Features.” *International Journal of Data Mining and Bioinformatics* 4 (1): 21–43.

Townley-Tilson, W. H. Davin, Sarah A. Pendergrass, William F. Marzluff, and Michael L. Whitfield. 2006. “Genome-Wide Analysis of mRNAs Bound to the Histone Stem-Loop Binding Protein.” *RNA (New York, N.Y.)* 12 (10): 1853–67. doi:10.1261/rna.76006.

Treger, M., and E. Westhof. 2001. “Statistical Analysis of Atomic Contacts at RNA-Protein Interfaces.” *Journal of Molecular Recognition: JMR* 14 (4): 199–214. doi:10.1002/jmr.534.

Tsompana, Maria, and Michael J. Buck. 2014. “Chromatin Accessibility: A Window into the Genome.” *Epigenetics & Chromatin* 7: 33. doi:10.1186/1756-8935-7-33.

Turner, Daniel, RyangGuk Kim, and Jun-tao Guo. 2012. “TFinDit: Transcription Factor-DNA Interaction Data Depository.” *BMC Bioinformatics* 13: 220. doi:10.1186/1471-2105-13-220.

Valdar, William S. J. 2002. “Scoring Residue Conservation.” *Proteins* 48 (2): 227–41. doi:10.1002/prot.10146.

van Dijk, Erwin L., Hélène Auger, Yan Jaszczyszyn, and Claude Thermes. 2014. “Ten Years of next-Generation Sequencing Technology.” *Trends in Genetics* 30 (9): 418–26. doi:10.1016/j.tig.2014.07.001.

Vanegas, Pamela L., Graham A. Hudson, Amber R. Davis, Shannon C. Kelly, Charles C. Kirkpatrick, and Brent M. Znosko. 2012. “RNA CoSSMos: Characterization of Secondary Structure Motifs--a Searchable Database of Secondary Structure Motifs in RNA Three-Dimensional Structures.” *Nucleic Acids Research* 40 (Database issue): D439–44. doi:10.1093/nar/gkr943.

Vangone, Anna, Romina Oliva, and Luigi Cavallo. 2012. “CONS-COCOMAPS: A Novel Tool to Measure and Visualize the Conservation of Inter-Residue Contacts in Multiple Docking Solutions.” *BMC Bioinformatics* 13 (Suppl 4): S19. doi:10.1186/1471-2105-13-S4-S19.

Vangone, Anna, Raffaele Spinelli, Vittorio Scarano, Luigi Cavallo, and Romina Oliva. 2011. “COCOMAPS: A Web Application to Analyze and Visualize Contacts at the Interface of Biomolecular Complexes.” *Bioinformatics* 27 (20): 2915–16. doi:10.1093/bioinformatics/btr484.

Villar, Diego, Paul Flicek, and Duncan T. Odom. 2014. “Evolution of Transcription Factor Binding in Metazoans — Mechanisms and Functional Implications.” *Nature Reviews Genetics* 15 (4): 221–33. doi:10.1038/nrg3481.

Vogel, Christine, and Edward M Marcotte. 2012. “Insights into the Regulation of Protein Abundance from Proteomic and Transcriptomic Analyses.” *Nature Reviews. Genetics* 13 (4): 227–32. doi:10.1038/nrg3185.

Walia, Rasna R., Li C. Xue, Katherine Wilkins, Yasser El-Manzalawy, Drena Dobbs, and Vasant Honavar. 2014. “RNABindRPlus: A Predictor That

Combines Machine Learning and Sequence Homology-Based Methods to Improve the Reliability of Predicted RNA-Binding Residues in Proteins." *PLoS ONE* 9 (5): e97725. doi:10.1371/journal.pone.0097725.

Wang, Jie, Jiali Zhuang, Sowmya Iyer, XinYing Lin, Troy W Whitfield, Melissa C Greven, Brian G Pierce, et al. 2012. "Sequence Features and Chromatin Structure around the Genomic Regions Bound by 119 Human Transcription Factors." *Genome Research* 22 (9): 1798–1812. doi:10.1101/gr.139105.112.

Wang, Liangjiang, and Susan J. Brown. 2006. "BindN: A Web-Based Tool for Efficient Prediction of DNA and RNA Binding Sites in Amino Acid Sequences." *Nucleic Acids Research* 34 (suppl 2): W243–48. doi:10.1093/nar/gkl298.

Wang, Liangjiang, Caiyan Huang, Mary Q. Yang, and Jack Y. Yang. 2010. "BindN+ for Accurate Prediction of DNA and RNA-Binding Residues from Protein Sequence Features." *BMC Systems Biology* 4 (Suppl 1): S3. doi:10.1186/1752-0509-4-S1-S3.

Wang, Liangjiang, Mary Qu Yang, and Jack Y Yang. 2009. "Prediction of DNA-Binding Residues from Protein Sequence Information Using Random Forests." *BMC Genomics* 10 (Suppl 1): S1. doi:10.1186/1471-2164-10-S1-S1.

Wang, Mingcong, Christina J. Herrmann, Milan Simonovic, Damian Szklarczyk, and Christian von Mering. 2015. "Version 4.0 of PaxDb: Protein Abundance Data, Integrated across Model Organisms, Tissues, and Cell-Lines." *Proteomics* 15 (18): 3163–68. doi:10.1002/pmic.201400441.

Wang, Y, Z Xue, G Shen, and J Xu. 2008. "PRINTR: Prediction of RNA Binding Sites in Proteins Using SVM and Profiles." *Amino Acids* 35 (2): 295–302. doi:10.1007/s00726-007-0634-9.

Wingender, E., X. Chen, R. Hehl, H. Karas, I. Liebich, V. Matys, T. Meinhardt, M. Prüss, I. Reuter, and F. Schacherer. 2000. "TRANSFAC: An Integrated System for Gene Expression Regulation." *Nucleic Acids Research* 28 (1): 316–19.

Wu, Carl, and C. David Allis. 2004. *Chromatin and Chromatin Remodeling Enzymes, Part A: Methods in Enzymoglogy*. Academic Press.

Wu, Jiansheng, Hongde Liu, Xueye Duan, Yan Ding, Hongtao Wu, Yunfei Bai, and Xiao Sun. 2009. "Prediction of DNA-Binding Residues in Proteins from Amino Acid Sequences Using a Random Forest Model with a Hybrid Feature." *Bioinformatics (Oxford, England)* 25 (1): 30–35. doi:10.1093/bioinformatics/btn583.

Xie, Zhi, Shaohui Hu, Seth Blackshaw, Heng Zhu, and Jiang Qian. 2010. "hPDI: A Database of Experimental Human Protein-DNA Interactions." *Bioinformatics (Oxford, England)* 26 (2): 287–89. doi:10.1093/bioinformatics/btp631.

Yakovchuk, Peter, Ekaterina Protozanova, and Maxim D. Frank-Kamenetskii. 2006. "Base-Stacking and Base-Pairing Contributions into Thermal Stability of the DNA Double Helix." *Nucleic Acids Research* 34 (2): 564–74. doi:10.1093/nar/gkj454.

Yang, Lin, Tianyin Zhou, Iris Dror, Anthony Mathelier, Wyeth W. Wasserman, Raluca Gordân, and Remo Rohs. 2014. "TFBSshape: A Motif Database for DNA Shape Features of Transcription Factor Binding Sites." *Nucleic*

Acids Research 42 (Database issue): D148–55. doi:10.1093/nar/gkt1087.

Yang, Xiao-Xia, Zhi-Luo Deng, and Rong Liu. 2014. "RBRDetector: Improved Prediction of Binding Residues on RNA-Binding Protein Structures Using Complementary Feature- and Template-Based Strategies." *Proteins* 82 (10): 2455–71. doi:10.1002/prot.24610.

Yang, Xiaoxia, Jia Wang, Jun Sun, and Rong Liu. 2015. "SNBRFinder: A Sequence-Based Hybrid Algorithm for Enhanced Prediction of Nucleic Acid-Binding Residues." *PloS One* 10 (7): e0133260. doi:10.1371/journal.pone.0133260.

Yang, Yu-Cheng T., Chao Di, Boqin Hu, Meifeng Zhou, Yifang Liu, Nanxi Song, Yang Li, Jumpei Umetsu, and Zhi John Lu. 2015. "CLIPdb: A CLIP-Seq Database for Protein-RNA Interactions." *BMC Genomics* 16: 51. doi:10.1186/s12864-015-1273-2.

Yang, Yuedong, Huiying Zhao, Jihua Wang, and Yaoqi Zhou. 2014. "SPOT-Seq-RNA: Predicting Protein-RNA Complex Structure and RNA-Binding Function by Fold Recognition and Binding Affinity Prediction." *Methods in Molecular Biology (Clifton, N.J.)* 1137: 119–30. doi:10.1007/978-1-4939-0366-5_9.

Zagrovic, Bojan. 2014. "Of RNA-Binding Proteins and Their Targets: Interaction Determines Expression." *Genome Biology* 15 (1): 102. doi:10.1186/gb4155.

Zanzoni, Andreas, Domenica Marchese, Federico Agostini, Benedetta Bolognesi, Davide Cirillo, Maria Botta-Orfila, Carmen Maria Livi, Silvia Rodriguez-Mulero, and Gian Gaetano Tartaglia. 2013. "Principles of Self-Organization in Biological Pathways: A Hypothesis on the Autogenous Association of Alpha-Synuclein." *Nucleic Acids Research* 41 (22): 9987–98. doi:10.1093/nar/gkt794.

Zhang, Jingyao, Huay Mei Poh, Su Qin Peh, Yee Yen Sia, Guoliang Li, Fabianus Hendriyan Mulawadi, Yufen Goh, et al. 2012. "ChIA-PET Analysis of Transcriptional Chromatin Interactions." *Methods (San Diego, Calif.)* 58 (3): 289–99. doi:10.1016/j.ymeth.2012.08.009.

Zhang, Yanping, Jun Xu, Wei Zheng, Chen Zhang, Xingye Qiu, Ke Chen, and Jishou Ruan. 2014. "newDNA-Prot: Prediction of DNA-Binding Proteins by Employing Support Vector Machine and a Comprehensive Sequence Representation." *Computational Biology and Chemistry* 52 (October): 51–59. doi:10.1016/j.compbiolchem.2014.09.002.

Zhao, Huiying, Jihua Wang, Yaoqi Zhou, and Yuedong Yang. 2014. "Predicting DNA-Binding Proteins and Binding Residues by Complex Structure Prediction and Application to Human Proteome." *PloS One* 9 (5): e96694. doi:10.1371/journal.pone.0096694.

Zhao, Huiying, Yuedong Yang, and Yaoqi Zhou. 2011. "Structure-Based Prediction of RNA-Binding Domains and RNA-Binding Sites and Application to Structural Genomics Targets." *Nucleic Acids Research* 39 (8): 3017–25. doi:10.1093/nar/gkq1266.

Zhao, Huiying, Yuedong Yang, and Yaoqi Zhou. 2013. "Prediction of RNA Binding Proteins Comes of Age from Low Resolution to High Resolution." *Molecular bioSystems* 9 (10): 2417–25. doi:10.1039/c3mb70167k.

Zhou, Ruhong. 2014. *Molecular Modeling at the Atomic Scale: Methods and*

Applications in Quantitative Biology. CRC Press.
Zhu, Xiaolei, Spencer S. Ericksen, and Julie C. Mitchell. 2013. "DBSI: DNA-Binding Site Identifier." *Nucleic Acids Research* 41 (16): e160. doi:10.1093/nar/gkt617.
Zwieb, Christian. 2014. "The Principles of RNA Structure Architecture." *Methods in Molecular Biology (Clifton, N.J.)* 1097: 33–43. doi:10.1007/978-1-62703-709-9_2.