

The evolution, modifications and interactions of proteins and RNAs



Ananth Prakash Surappa-Narayanappa

Hughes Hall

University of Cambridge

European Bioinformatics Institute

A dissertation submitted for the degree of

Doctor of Philosophy

August 2017

To,
Amma and Appa

“Extraordinary claims require extraordinary evidence.”

– Carl Sagan

Declaration

This dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration except as declared in the Preface and specified in the text.

It is not substantially the same as any that I have submitted, or, is being concurrently submitted for a degree or diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text. I further state that no substantial part of my dissertation has already been submitted, or, is being concurrently submitted for any such degree, diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text.

This dissertation does not exceed the prescribed word limit of the Degree Committee for the Faculty of Biology.

Summary

Proteins and RNAs are two of the most versatile macromolecules that carry out almost all functions within living organisms. In this thesis I have explored evolutionary and regulatory aspects of proteins and RNAs by studying their structures, modifications and interactions.

In the first chapter of my thesis I investigate *domain atrophy*, a term I coined to describe large-scale deletions of core structural elements within protein domains. By looking into truncated domain boundaries across several domain families using Pfam, I was able to identify rare cases of domains that showed atrophy. Given that even point mutations can be deleterious, it is surprising that proteins can tolerate such large-scale deletions. Some of the structures of atrophied domains show novel protein-protein interaction interfaces that appear to compensate and stabilise their folds.

Protein-protein interactions are largely influenced by the surface and charge complementarity, while RNA-RNA interactions are governed by base-pair complementarity; both interaction types are inherently different and these differences might be observed in their interaction networks. Based on this hypothesis I have explored the protein-protein, RNA-protein and the RNA-RNA interaction networks of yeast in the second chapter. By analysing the three networks I found no major differences in their network properties, which indicates an underlying uniformity in their interactomes despite their individual differences.

In the third chapter I focus on RNA-protein interactions by investigating post-translational modifications (PTMs) in RNA-binding proteins (RBPs). By comparing occurrences of PTMs, I observe that RBPs significantly undergo more PTMs than non-RBPs. I also found that within RBPs, PTMs are more frequently targeted at regions that directly interact with RNA compared to regions that do not. Moreover disorder and amino acid composition were not observed to significantly influence the differential PTMs observed between RBPs and non-RBPs. The results point to a direct regulatory role of PTMs in RNA-protein interactions of RBPs.

In the last chapter, I explore regulatory RNA-RNA interactions. Using differential expression data of mRNAs and lncRNAs from mouse models of hereditary hemochromatosis, I investigated competing regulatory interactions between mRNA, lncRNA and miRNA. A mutual interaction network was created from the predicted miRNA interaction sites on mRNAs and lncRNAs to identify regulatory RNAs in the disease. I also observed interesting relations between the sense-antisense mRNA-lncRNA pairs that indicate mutual regulation of expression levels through a yet unknown mechanism.

Acknowledgements

The more I think about what to write here, the more difficult it is proving to be - how do I thank all the people on a sheet of paper, who have been with me on this scientific and emotional journey over the last four years. As I reflect on the past, all I can think of are sweet memories that come rushing back. It is impossible for me to pick one memory over another or thank a few people among the many, whom I have met over these past few years; for them I am ever so grateful to have made this happen.

One person who I am very grateful to is my supervisor, Alex Bateman. If it weren't for him, things would have been very different for me. By giving me an opportunity to do Ph.D. at the EBI, he has definitely changed my life for good. I have not only learned from him how to do good science, but also how to be a better scientist. I thank him for being patient and kind with me, guiding me and for having confidence in me at times when I had doubts about myself. He is my best supervisor, ever.

I also thank my Thesis Advisory Committee members – Dr. Laura Itzhaki, Dr. Toby Gibson, Dr. Sarah Teichmann, Dr. Anton Enright and Dr. Marco Marcia, who over the years have given valuable advice on my projects. I thank Dr. Martina Muckenthaler and Dr. Kamesh Babu for providing me with experimental data on lncRNAs.

I am lucky to have had Penny, Ruth and Gera as my colleagues earlier during my Ph.D. I'd like to thank Neil, for educating me on birds, wildlife, classical antiquity and ancient Egyptian history on our daily discussions at tea in the mornings and also for proofreading my thesis. I also thank Matt, Aleix and my fellow predocs for scientific discussions.

I thank my friends for bringing fun and colour in my life during these years. Hugs to my two bestest friends Jag & Swaathi – for taking me in and making me a part of their family. I shall always have a big smile on my face when I think of Uma, Navis, Venkat, Thawfeek, Netra, Phani, Kedar, Nitin and Rizwan.

I've dedicated this thesis to my Mum and Dad, who have always shown love and support. I've had their emotional support and comfort all through these years. Mum has been my greatest source of strength in life. She has encouraged me, laughed and shared my happiness over those innumerable phone calls everyday over the last four years. I would also like to thank my brother and sister for their moral support.

Contents

Preface.....	xiii
Chapter 1	1
Protein domain atrophy – identification and characterisation of functional partial protein domains.....	1
1.1 Introduction	1
1.2 Methods.....	7
1.2.2 Nomenclature.....	7
1.2.3 Atrophy Score.....	9
1.2.4 Filtering.....	12
1.2.5 Manual inspection	12
1.2.6 Structure visualisation	15
1.2.7 Phylogenetic analysis.....	15
1.3 Results.....	16
1.3.1 N-terminal end-bounded atrophy	19
1.3.2 C-terminal end-bounded atrophy.....	26
1.3.3 Upstream domain-bounded atrophy.....	28
1.3.4 Downstream domain-bounded atrophy.....	28
1.3.5 Within-domain atrophy	33
1.4 Conclusion.....	36
1.5 References.....	42
Chapter 2.....	49
Comparative analysis of the yeast non-coding RNA interaction network.....	49

2.1 Introduction	49
2.2 Methods.....	56
2.2.1 Data collection and curation.....	56
2.2.2 Network analysis	56
2.2.3 Degree distribution.....	57
2.2.4 Clustering coefficient (Transitivity)	58
2.2.5 Betweenness centrality.....	58
2.2.6 Closeness centrality.....	58
2.2.7 Neighbourhood connectivity	58
2.3 Results.....	59
2.3.1 Degree distribution.....	63
2.3.2 Clustering coefficient (Transitivity)	67
2.3.3 Betweenness centrality.....	71
2.3.4 Closeness centrality	74
2.3.5 Neighbourhood connectivity	76
2.4 Conclusion.....	80
2.5 References.....	83
Chapter 3.....	89
Post-translational modifications of RNA-binding proteins.....	89
3.1 Introduction	89
3.2 Methods.....	96
3.2.1 RNA-binding peptides.....	96
3.2.2 Non RNA-binding proteins	98
3.2.3 RNA-binding proteins.....	98
3.2.4 DNA-binding proteins.....	98

3.2.5 Post-translational modifications.....	99
3.2.6 Globular and disordered regions.....	99
3.2.7 Structural validation.....	99
3.2.8 Protein abundance.....	100
3.3 Results.....	100
3.3.1 Overview of RBDpep and candidate RBDpep datasets.....	100
3.3.2 Post-translational modifications in RNA-binding proteins and non RNA-binding proteins.....	102
3.3.3 Post-translational modifications in RNA-binding peptides and non RNA-binding peptides.....	111
3.3.4 Disorderedness in RNA-binding proteins.....	116
3.3.5 Functional classification of RNA-binding proteins.....	118
3.3.6 Amino acid abundance.....	120
3.3.7 Protein abundance.....	122
3.3.8 Structural validation.....	124
3.3.9 Regulation of RNA-protein interactions mediated by post-translational modifications.....	126
3.4 Conclusion.....	130
3.5 References.....	133
Chapter 4.....	141
Long non-coding RNA mediated regulation of gene expression in hereditary hemochromatosis.....	141
4.1 Introduction.....	141
4.2 Methods.....	150
4.2.1 Mouse models of hereditary hemochromatosis.....	150

4.2.2 RNA sequencing and differential expression analysis.....	150
4.2.3 Homology of lncRNAs	151
4.2.5 miRNA target site prediction.....	152
4.2.6 Competing endogenous RNA network	152
4.2.7 Sense-antisense mRNA-lncRNA pairs.....	153
4.2.8 LincRNA-adjacent mRNA pairs.....	154
4.2.9 Gene Ontology	154
4.3 Results.....	155
4.3.1 Overview of Fpn-C326S and Fpn-Trp datasets.....	155
4.3.2 Homologues of lncRNAs.....	157
4.3.3 miRNA target site prediction and conservation.....	159
4.3.4 Competing endogenous RNA network	164
4.3.5 Co-expression of sense-antisense mRNA-lncRNA pairs	171
4.3.6 Correlation of expression of lincRNAs and adjacent mRNAs	177
4.3.7 Gene ontology enrichment	181
4.4 Conclusion.....	181
4.5 References.....	186
Appendix	195
Table A1.....	197
Table A2.....	200
Figure A1.....	202
Figure A2.....	203
Table A3.....	204
Table A4.....	207

Preface

The work on this thesis began four years ago, in August 2013, during the Ph.D. selection interviews. It was during the interview my supervisor Dr. Alex Bateman and I discussed how we had both previously noticed structures of protein domains that had undergone degradation. We realised that there could be many other such structurally degraded protein domains out there, which may have not been discovered. Literature search confirmed our doubts that partial protein domains do not appear to have been systematically studied, except for a few sporadic reports in the literature. We agreed that I would begin work on my Ph.D. by systematically looking for partial protein domains. During the course of this work, in 2014, Alex met Dr. William Pearson and Dr. Deborah Triant, from the University of Virginia, at the Intelligent Systems for Molecular Biology (ISMB) conference in Boston, USA, where both groups discovered that they have been working on a related topic. While Pearson and Triant focussed on the bioinformatics causes of partial domain artefacts, we focussed on identifying cases of true partial domains. Due to the similar nature of work, we borrowed their nomenclature for describing different domain atrophy types. In the end both groups, using different perspectives, came to a similar conclusion on partial domains – that true cases of partial domains are extremely rare and most cases are sequence artefacts. This work is presented in Chapter 1. Results from both groups were published as back-to-back research articles in *Genome Biology*. *Genome Biology* also carried out a research highlight article commenting on partial protein domains by Lawrence Kelley and Michael Sternberg from Imperial College London. The studies were also featured in news outlets such as BioMed Central's blog network and EMBL etc.

Following the publication of this study and discussions with the Thesis Advisory Committee, Alex and I planned the next course of projects, which focussed on studying interactions involving RNAs. Firstly, I would compare and analyse protein-protein, protein-RNA and RNA-RNA interaction networks, then I would

focus on post-translational modifications and their influence on protein-RNA interactions and finally investigate RNA-RNA interactions by studying regulation of messenger RNA (mRNA) expression by non-coding RNAs (ncRNAs). During this period, in November 2014, Dr. Martina Muckenthaler and Dr. Kamesh Rajendra Babu, from the University Hospital Heidelberg, Germany, contacted us regarding a collaborative project on non-coding RNAs. I accepted to work on this topic as it fitted well into my research plans. Their experimental group at Heidelberg had characterised coding and non-coding transcripts that were differentially expressed in mouse models of hereditary hemochromatosis – a genetic condition that causes abnormality in iron homeostasis. I explored various computational methods to understand if non-coding transcripts such as long non-coding RNAs and microRNAs would form mutual regulatory interactions with mRNAs to control their expression. This topic was new and challenging to me and offered a good understanding and appreciation of the complexities of the regulatory RNA-world. I passed on the results to the experimental group, which were then taken up by them for further experimental analyses. Chapter 4 documents the outcome of this work.

By March 2016 I had started to analyse the protein and RNA interaction networks in yeast and humans. This work was carried out using the macromolecular interaction data curated from literature by Dr. Sandra Orchard from the IntAct team at the EBI, UK and Dr. Simona Panni from the University of Calabria, Italy. In this study I have compared the physical network properties of three macromolecular interaction networks – protein-protein, protein-ncRNA and ncRNA-ncRNA. The yeast non-coding RNA-RNA interaction network from this study is the first such reported non-coding RNA interaction network. The result from this study is presented in Chapter 2. The results were accepted for publication by the RNA journal and the manuscript is currently in the pre-publication process.

Lastly, I investigated post-translational modifications (PTMs) of RNA-binding proteins. The aim of this study was to understand how PTMs could influence interactions with the RNA. By using data from a recently published study in

August 2016, on RNA-binding peptides, I was able to map PTM sites onto RNA-binding proteins, which then allowed me to distinguish PTMs in RNA-binding regions from non RNA-binding regions. The outcome of this work is presented in Chapter 3. I plan to draft the results from this study and submit the manuscript for publication.

A small project that I undertook at the beginning of my PhD, that is not part of the thesis, was in collaboration with Dr. David Thomas from the Department of Medicine, University of Cambridge, UK. In this project I analysed homology of an uncharacterised mouse protein C17ORF62 later named Eros. This result as part of the larger study was published in *The Journal of Experimental Medicine*.

I have also presented the results of my work in regional and international meetings. The domain atrophy study was presented at the 2nd student symposium organised by International Society for Computational Biology's Regional Student Group (ISCB-RSG) – UK chapter, 2015 in Norwich. Results from the lncRNA-mediated regulation of hereditary hemochromatosis were presented at RNA 2016 The 21st Annual Meeting of the RNA Society, in Kyoto, Japan.

Chapter 1

Protein domain atrophy – identification and characterisation of functional partial protein domains

1.1 Introduction

Protein domains are key to the diversity of structure and functions observed in proteins. Domains are composed of a defined set of secondary structural elements, which are spatially arranged to form distinct folded stable 3-dimensional structures. In their billions of years of evolution, domains have evolved from simple folds with basic functions to large multifunctional complex subunits. In the traditional sense, protein domains are viewed as indivisible structural and functional building blocks; however, a few recent studies have identified proteins that are composed of structurally partial or incomplete domains. Existence of such partial protein domains is interesting as they shed light on the evolution, function and stability of these domains. In this chapter I have carried out large-scale systematic analysis of protein domain families in Pfam in order to identify cases of partial structural domains. I quantify the magnitude of structural loss in protein domains and discuss the nature of deletions, their functions and the mechanisms that stabilise these domains. Finally, I discuss some of the bioinformatics artefacts that plague the identification of true partial domains.

Domains are spatially distinct structural units within a protein that are characterised by conserved sequence, geometrical compactness and the ability to fold and function independently (Ponting and Russell, 2002). One of the characteristic features of protein domains is their recurrence in different contexts, i.e., the domain is observed in one or more different multidomain proteins (Ponting and Russell, 2002; Vogel et al., 2004). This modularity allows protein domains to be combined in many ways giving rise to proteins with diverse structures and functions.

Protein domains have distinct 3-dimensional folds that have evolved along with their functions. The term protein fold commonly denotes the topology of secondary structural elements and their unique spatial arrangement within a domain. Evolutionarily related functionally similar domains often exhibit the same fold, however it is also commonly observed that the same protein fold is shared between protein domains that are functionally diverse (Martin et al., 1998). One of the best-known examples of one-fold-many-functions is the TIM barrel fold, which is commonly observed among protein domains that catalyse a wide range of chemical reactions (Nagano et al., 2002). Given the intrinsic physical constraints of protein folding, it is assumed that there is a limited number of ways a domain can fold, which has led to an observation that the multitude of protein functions are carried out by only a few thousand unique protein folds (Chothia, 1992; Finkelstein et al., 1993).

Identification of a large number of protein domains have led to the systematic classification of domains into families and superfamilies either based on their 3-dimensional structures or folds (Andreeva et al., 2014; Sillitoe et al., 2015) or based on their amino acid sequences (Finn et al., 2016). The SCOP and CATH databases are two of the well-known structure-based classifiers of protein domains, which group domains based on their similarity in structure or folds with or without detectable sequence similarity (Andreeva et al., 2014; Sillitoe et al., 2015). The Pfam database classifies protein domains based on sequence similarity; domains with highly similar sequences are grouped into families and clans and also have similar functions (Finn et al., 2016). Some of the other

databases that classify protein domains based on sequence and/or structure include FSSP (Holm and Sander, 1998), CDD (Marchler-Bauer et al., 2015), SMART (Letunic et al., 2012), ProDom (Servant et al., 2002) and PROSITE (Sigrist et al., 2013) among others. At present there are 1,393 different folds defined by SCOP (v1.75, 2017) and 1,375 unique folds (topologies) defined by CATH (v4.0.0, 2017) and 16,712 protein domain families defined by Pfam (31.0, 2017).

A large fraction of proteins among prokaryotes and eukaryotes are composed of two or more domains; about two-thirds of prokaryotic proteins and about 80% of eukaryotic proteins are multidomain (Chothia et al., 2003; Teichmann et al., 1998). It is often observed that multidomain proteins evolve through domain duplications followed by functional modification either through sequence divergence or by recombination with other domains (Lynch and Conery, 2000; Vogel et al., 2004). Proteins have also been observed to lose single or multiple domains during their course of evolution through mechanisms such as insertion of new start and stop codons, gene fusion and gene fission (Buljan and Bateman, 2009; Weiner et al., 2006).

Apart from the domain gain or domain loss events, protein domains, at a smaller scale of modification, can gain or lose secondary structural elements through insertions or deletions (indels) of amino acid residues. Analyses of a large number of domain superfamily sequences and structures have shown variability in domain lengths, which are attributed to indels in loops, coils or a few secondary structural elements that leave the domain core largely unaltered (Pascarella and Argos, 1992; Sandhya et al., 2008; Sandhya et al., 2009; Taylor et al., 2004). Variations in domain lengths caused by gain of accessory secondary structural elements or 'embellishments' are well studied (Reeves et al., 2006). One of the examples of embellishments in protein domains is the HUP superfamily (CATH: 3.40.50.620), wherein domains exhibit large structural variations around its domain core (Dessailly et al., 2010). Some of the other superfamily members that exhibit large-scale domain embellishments include galactose binding domain-like superfamily (CATH: 2.60.120.260), cupredoxin superfamily (CATH: 2.60.40.420), dihydrodipicolinate reductase domain 2

superfamily (CATH: 3.30.360.10), ATP-dependent amine/thiol ligase superfamily (CATH: 3.30.0470.20) and the $\alpha\beta$ -hydrolase superfamily (CATH: 3.40.50.1820) (Reeves et al., 2006). These embellishments have been observed to influence interactions, affect substrate specificity, binding and stability and degradation (Dessailly et al., 2010; Reeves et al., 2006).

Similar to domain embellishments, protein domains could also undergo large-scale loss or degradation of secondary structural elements. I propose the term 'domain atrophy' for events that lead to a large-scale loss of core secondary structural elements in protein domains. Figure 1.1 schematically illustrates the theory of domain atrophy using an example of a multi-domain protein. In the course of protein evolution one or more domains in a protein undergo truncation or a significant loss of its structural elements due to a mutation or other cellular events. The protein with a truncated or atrophied domain that still retains its active sites or original function, such as enzymatic or structural, may be positively selected while the protein with a non-functional atrophied domain is lost. Unlike cases wherein only a few peripheral secondary structural elements are lost, domain atrophy refers to large-scale deletions of 'core' structural elements.

Mutational events that lead to such large-scale loss of domain structure are often detrimental to protein stability and function. Proteins are only marginally stable such that a mutation of a single amino acid residue can drastically influence its folding. For example, single missense mutations within the sucrose domain of sucrose-isomaltase leads to defects in protein folding and transport (Alfalah et al., 2009). It is also observed that the core mutation of a single amino acid residue determines the folding stability of the N-terminal domains of P-type copper ATPases CopAa and CopAb (Banci et al., 2003). Large-scale deletions are also expected to significantly alter stability, however cases exist wherein such deletions only marginally affect protein stability. For example, removal of a stretch of amino acids such as the whole β -strand of the Ig-domain of the human muscle protein titin only marginally decreases stability by 2.8 kcal/mol (Fowler et al., 2002). Single or large-scale deletions or mutations of amino acid residues

affect protein stability and function differentially based on their location. Mutations of amino acids that are part of peripheral secondary structures or loops, termini, or those present on the protein surface are more likely to be tolerated, whereas mutations within the hydrophobic core that disrupt packing are not (Bowie et al., 1990).

Unlike domain elaborations, which is commonly seen in protein structures, structural data and literature on domain atrophy is very scarce. Only three cases of structural partial domains have been observed in the past, which include 'truncated globin family' (Nardini et al., 2007) and bacterial luciferases (Grishin, 2001) and recently in a domain of unknown function DUF2172 (Das et al., 2014). One of the main reasons that the atrophied domains are less studied could be due to their rarity because of the reasons described above. However, natural occurrences of a few stable protein domains with large-scale structural deletions are intriguing and suggest that compensatory mechanisms that help stabilise these atrophied domains must exist.

In this chapter, using sequence-based profile hidden Markov models (HMMs) of protein domain families, I have devised an algorithm to identify potential new cases of domain atrophy. I introduce a new metric called the 'atrophy score' to quantify the magnitude of structural loss. Using the algorithm I have identified several new cases of domain atrophy. For sequences where experimental structures were not available I have instead mapped sequences of atrophied domains on to complete homologous structures, as reference, to infer the extent of atrophy. I have also identified cases that confound the discovery of true atrophied domains. Using a series of filters and through manual curation I avoid cases of computational artefacts and discuss their possible origins. Finally, I discuss compensatory mechanisms that stabilise the folds of atrophied domains.

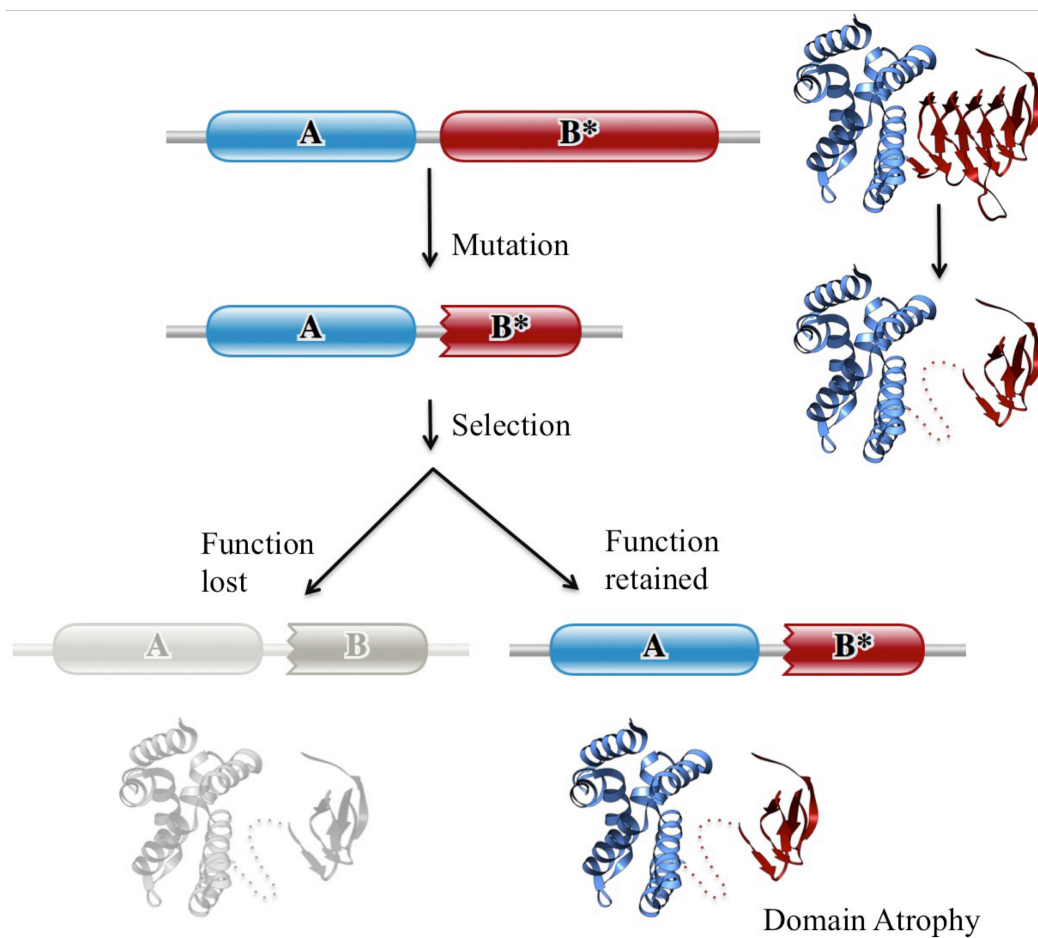


Figure 1.1 Schematic representation of domain atrophy event. A protein with a particular architecture comprising domains A and B, wherein domain B has the active site residues, undergoes mutation resulting in truncation of domain B. The protein is positively selected if the truncated domain retains its functional state (enzymatic or structural), while the protein with a non-viable truncated domain is lost. Complete Pfam domain boundaries are denoted by smooth edges and incomplete domain boundaries are denoted by toothed edges. Dotted line in the toy example shows the region of atrophy. Figure reused from (Prakash and Bateman, 2015), doi: 10.1186/s13059-015-0655-8.

1.2 Methods

The following sections (1.2.1 to 1.2.7) are taken verbatim from (Prakash and Bateman, 2015).

1.2.1 Data

To identify potential cases of domain atrophy I use matches of the UniProt sequence database (release 2012_06) against the profile HMM models from the Pfam database release 27.0 (Finn et al., 2014). This set of matches contains 28,738,352 Pfam-A protein domain instances across 14,831 families in 18,523,877 protein sequences.

1.2.2 Nomenclature

Domain atrophy events were classified into five types, based on the domain location (architecture) in the protein and the region of atrophy in the domain. Figure 1.2 shows the schematic representations of the five types atrophied domains, which are described below.

- (1) N-terminal end-bounded atrophy: structural loss at the N-terminal region of the N-terminal domain.
- (2) C-terminal end-bounded atrophy: structural loss at the C-terminal region of the C-terminal domain.
- (3) Upstream domain-bounded atrophy: structural loss at the N-terminal region of an inner domain, also including the N-terminal region of the C-terminal domain.
- (4) Downstream domain-bounded atrophy: structural loss at the C-terminal region of an inner domain, also including the C-terminal region of the N-terminal domain.
- (5) Within-domain atrophy: structural loss within the domain.

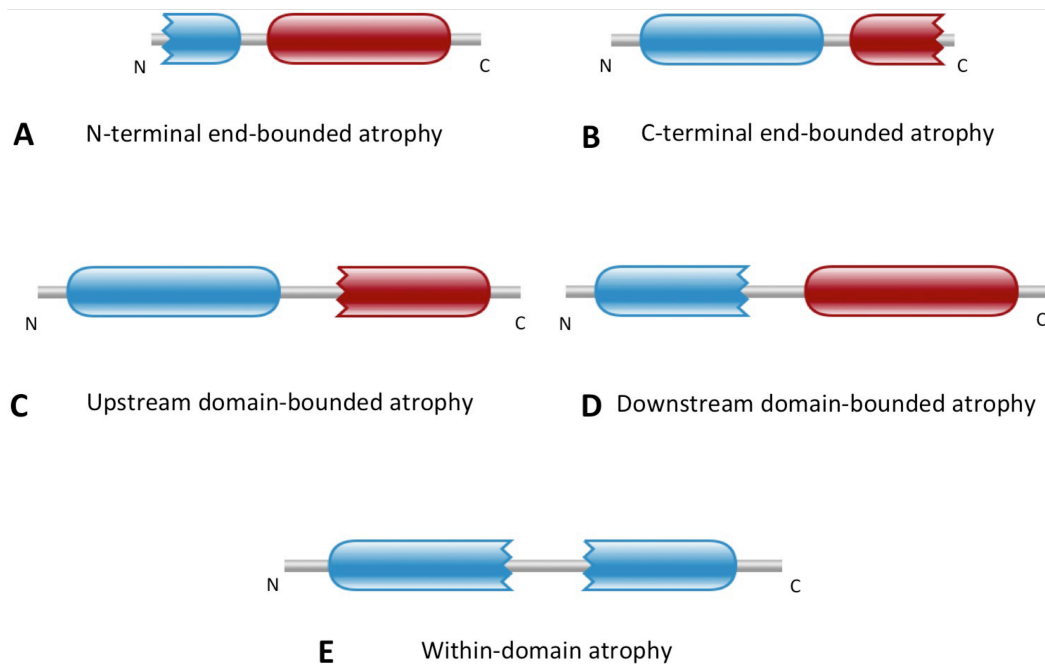


Figure 1.2 The five classes of domain atrophy events. Complete Pfam domain boundaries are represented as smooth edges. Domain boundaries with incomplete or partial matches to Pfam HMM models are represented as toothed-edges. Figure reused from (Prakash and Bateman, 2015), doi: 10.1186/s13059-015-0655-8.

1.2.3 Atrophy Score

To determine which domains may be cases of atrophy I calculated a measure called the Atrophy Score (AS) at both the N-terminal (AS_N) and C-terminal (AS_C) boundaries of each domain instance using their protein sequence and HMM coordinates.

An HMM is a statistical model that describes observable events that depend on internal factors, through a visible process of observable symbols and an invisible process of hidden states. These models have been applied in a wide range of applications from speech recognition to passive sonar detection. In the case of biology, the HMM architecture that is used today was introduced by Sjölander and Haussler (Krogh et al., 1994) and became known as profile HMMs. Protein profile HMMs can be used to infer homology and predict secondary and tertiary structures. A Pfam profile HMM representing a particular protein family is built from an aligned set of good quality homologous protein sequences. This profile HMM comprises a number of hidden states, which correspond to columns of a multiple sequence alignment. As the HMM progresses from one state to another according to the state-transition probabilities, the state emits an amino acid residue (or symbol) according to its symbol-emission probabilities (Eddy, 2004). Once the end state is reached the observable sequence of amino acid residues (or symbols) is generated. As well as acting as a generative model, profile HMMs can be used to score sequences to see how well they fit the model. The general scoring scheme is to calculate the probability of the sequence given the model, normalised by the probability of the sequence being generated by a null or random model.

Figure 1.3 shows a schematic representation of the parameters used in calculating domain atrophy score. The equations used to calculate the atrophy score are shown below:

$$AS_N = (D_N - d_N) / L \dots\dots\dots (Eqn. 1)$$

$$AS_C = (D_C - d_C) / L \dots\dots\dots (Eqn. 2)$$

Where, AS_N is the atrophy score at the N-terminus of the domain, D_N is the number of unmatched HMM match states at the N-terminus of the domain, d_N is the inter-domain distance or domain interval, i.e., the number of amino acid residues between the domain and its adjacent upstream domain or the sequence start site in the case of an N-terminal domain, and L is the HMM model length of the domain family. Similarly AS_C , D_C and d_C correspond to atrophy score at the C-terminus of the domain, the number of unmatched HMM match states at the C-terminus of the domain and the inter-domain distance to the start site of the downstream domain or, in the case of a C-terminal domain, its sequence end site respectively. Instances of within-domain atrophy can be identified in cases where the profile HMM matches to a single domain have been split into two profile HMM matches, with the first corresponding to the N-terminal part of the domain and the second corresponding to the C-terminal part of the domain. Instances of within-domain atrophy were distinguished from tandem repeats by considering the HMM match states of each domain. The start HMM-match state of the downstream domain is greater than the end HMM-match state of the upstream domain in cases of split domains. The computation of within-domain atrophy score (AS_w) is similar to AS_N .

$$AS_w = (D_w - d_w)/L \dots\dots\dots (Eqn. 3)$$

Where D_w is the number of unmatched HMM match states within the domain, d_w is the domain interval within the domain and L is the HMM model length of the domain family. Alignment co-ordinates of each domain are considered to calculate the inter-domain interval. An intuitive description of the atrophy score would be that a score of 0.33 means that one-third of the length of the domain has been lost to domain atrophy.

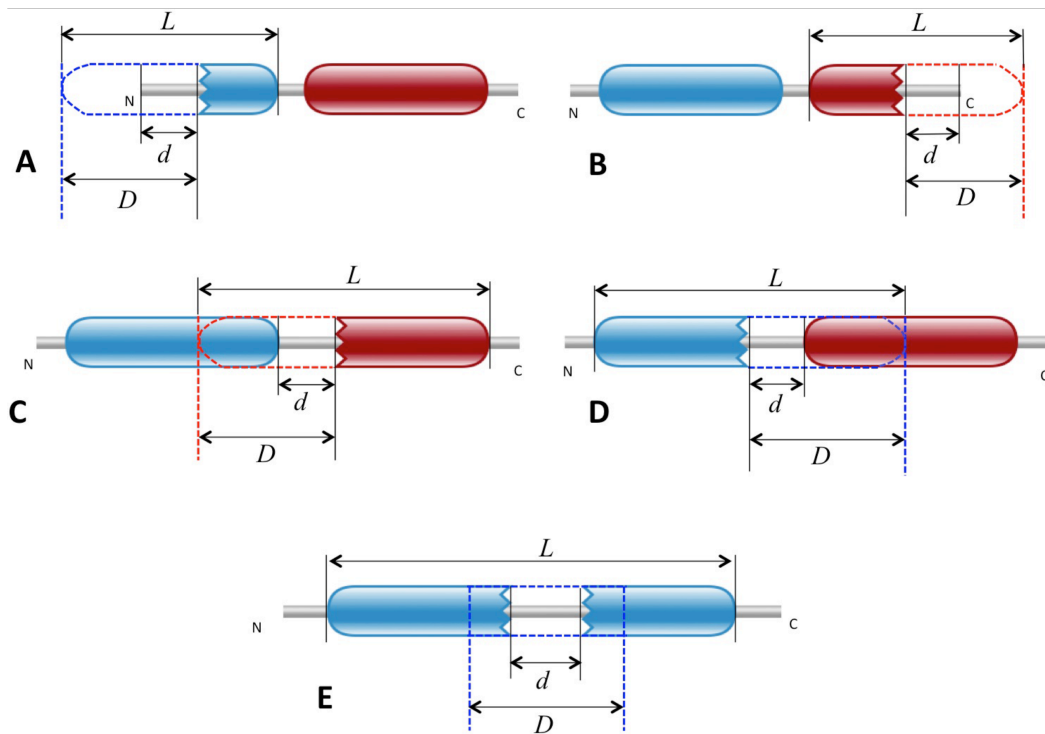


Figure 1.3 Calculation of domain atrophy score. Atrophy score is the ratio of the missing region ($D-d$) of the domain to the domain HMM-model length (L). (A) N-terminal end-bounded atrophy, (B) C-terminal end-bounded atrophy, (C) Upstream domain-bounded atrophy, (D) Downstream domain-bounded atrophy and (E) Within-domain atrophy. Figure reused from (Prakash and Bateman, 2015), doi: 10.1186/s13059-015-0655-8.

1.2.4 Filtering

Initial results from applying the atrophy score to all UniProt proteins showed that there were numerous common failure modes (Figure 1.4) that would mask the ability to find genuine domain atrophy events. Therefore, I applied a set of filters to reduce the number of false positive matches with high atrophy scores.

Of a total of 23,193,494 sequences in the database, 18,523,877 sequences had at least one Pfam-A domain instance and these were used in the analysis. Initial filtering was applied to exclude domain models from sequences with protein existence (PE) levels of 2 to 5. These are enriched in gene prediction errors and fragment proteins. This reduced the number of sequence considered from 18,523,877 to 77,305. Proteins with a protein existence level of 1 have clear experimental evidence for the existence of the protein from Edman sequencing, mass spectrometry, X-ray, NMR or other experimental evidence. Although not strictly a measure of protein sequence quality these proteins usually have highly accurate protein sequences. We also removed sequences annotated as fragments in UniProt, which further reduced the set of sequences considered from 77,305 to 75,435.

Adjacent domains that are of the same clan, similar to figure 1.4B, could lead to ambiguous domain boundary assignments at the interval and hence such cases were filtered out to avoid detecting false atrophy events. The resulting final-set comprising 114,303 domain instances from 75,435 sequences were included in the analysis. The algorithm pipeline was implemented in Perl to calculate atrophy scores across the set of domains. Domain instances with an atrophy score of 0.15 or more were further investigated.

1.2.5 Manual inspection

Domain instances that were obtained after applying the above filters were then selected for manual inspection. Only those domains that had an atrophy score of 0.15 or above were checked manually for identification of false positives. Each

potential domain atrophy case was checked for evidence of any of the following failure modes (see Figure 1.4):

- (1) Gene prediction errors: I checked whether the missing part of a domain was to be found in an adjacent gene or due to an incomplete gene prediction.
- (2) Nested domains: I checked whether a high atrophy score was due to a domain nesting within another. These were considered as false positives.
- (3) Multi-domain families: Due to incorrect Pfam domain definitions some Pfam domains actually correspond to multiple structural domains that can be found independently. I checked the structure of each Pfam family to confirm whether this was the cause of a high atrophy score.
- (4) Small domains: Domains of length less than 30 amino acid residues were not considered since atrophy score of 0.15 and above of small domains correspond to loss of a single secondary structural element or a part thereof, which is not considered true atrophy.
- (5) Circular permutations: While circular permuted domains are complete domains, the rearrangement of domain HMM start-site and HMM end-site with respect to their domain HMM-model would result in misidentification of such cases as domain atrophies.
- (6) Short repeats: Domains composed of tandem structural motifs, such as β -propeller, β - or α -helix, are made of short repeating sequence motifs were considered as false positives. Addition or removal of repeats is often tolerated in terms of protein mutation.
- (7) Disordered domains: Inferring domain atrophy among intrinsically disordered protein domains is not straightforward mainly owing to their

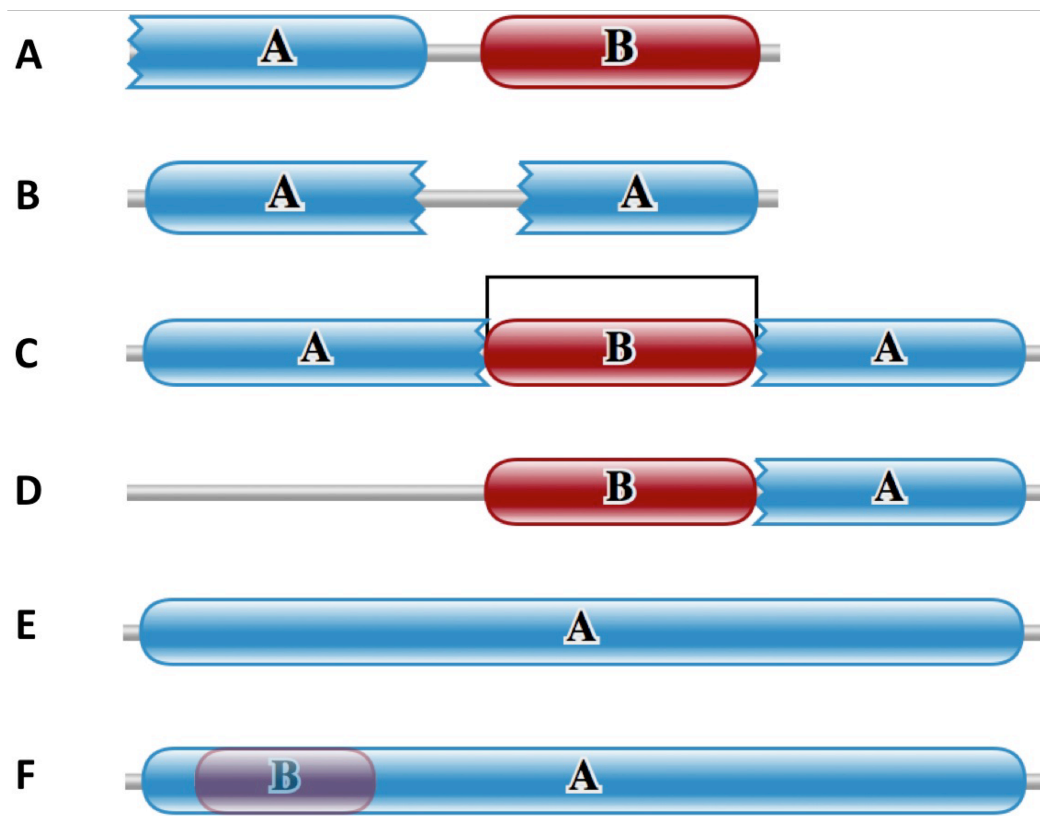


Figure 1.4 Failure modes of the pipeline. Some of the commonly observed domain architectures that were classified as false-positives or failure modes by the pipeline. (A) Incorrect gene prediction or partial sequence: sequence fragment or incorrect gene prediction could lead to events that look like domain atrophy, example: Aldo/keto reductase family (UniProt: P43546, Pfam: PF00248) (B) Tandem repeat: a tandem repeat is distinguished from single domain instances that are split/predicted in two parts, by considering its HMM match coordinates; for tandem domain instances the downstream domain start-HMM-match state is less than the upstream domain end-HMM-match state, example: Peroxidase (UniProt: A0QXX7, Pfam: PF00141). (C) Nested domain: this architecture results in an atrophy score greater than 1 for domain hosting nested domain(s), example: Peptidase_M20 (UniProt: A0Z6B3, Pfam: PF01546) (D) Unmatched domain region: missing region of the domain containing a nested domain, example: Lon_C (UniProt: A4ILZ1, Pfam: PF05362). (E) Multi-domain family: a single-domain architecture comprising more than one domain. (F) Domain overlap, example: 4Fe-4S single cluster domain (UniProt: A6L094, Pfam: PF13353). Figure reused from (Prakash and Bateman, 2015), doi: 10.1186/s13059-015-0655-8.

lack of native ordered tertiary structure and such cases were considered false positives. Apart from the above failure modes domain atrophy cases whose structures were theoretically modelled or had no other reference structures in the family to compare with were also treated as failure modes. Other cases of complete structural domains but identified as domain atrophy were treated as false positives.

From a total of 1,362 domain instances, with atrophy scores between 0.15 and 1, which were manually checked, 1,287 domain instances were classified as failure modes or false positives. The positive predictive value (PPV) of my method to identify domain atrophy is $(75/(75 + 1287)) = 0.055$.

1.2.6 Structure visualisation

Structures were visualised with Chimera (Pettersen et al., 2004). Where experimental structures of atrophied domains were not available, the shortest full-length domain structure within the domain family was chosen as the reference. The extent of domain atrophy was then inferred by a pairwise sequence alignment guided mapping of unaligned sequence regions onto the reference structure. Instances of putative domain atrophy where no full-length reference structure was available for the family were not considered further.

1.2.7 Phylogenetic analysis

Evolutionary information was inferred from phylogenetic trees constructed from multiple sequence alignment of domain family seed sequences and homologous sequences from a JackHMMER search (Finn et al., 2011). A non-redundant set of sequences of 90% identity or less was aligned with MAFFT (Kato and Standley, 2013). Alignments were visualised with Belvu (Sonnhammer and Hollich, 2005) and phylogenetic trees constructed using the neighbour-joining method present in Belvu using default parameters.

1.3 Results

Amino acid sequences from UniProt were scanned against the Pfam profile HMM models to identify potential cases of atrophied domains. I investigated the domains that showed partial matches to the profile HMM models and calculated the atrophy scores. Atrophy score quantifies the magnitude of structural loss and is equivalent to the fraction of the Pfam profile HMM model that is missing from the domain. A negative atrophy score indicates that domain is complete or there is no structural loss, while a positive atrophy score indicates an incomplete match to the profile HMM and structural loss. I note that a partial match of domain sequences to profile HMMs does not always denote domain atrophy; matches to the termini of profile HMM models could be missed due to low sensitivity of the model to the terminal sequences, hence for cases wherein a domain sequence does not completely match the profile HMM, similarity can often be found to the full-length domain by extending the sequence through simple sequence similarity comparison. In order to avoid detecting such partial profile HMM matches of full-length domains as domain atrophy, I have used domain boundaries of adjacent domains or the sequence terminus to constrain partial profile HMM matches. Full-length sequences of domains that are partially matched to profile HMMs cannot be extended over to the adjacent non-homologous domains or sequence terminus, which suggest that they represent potential cases of domain atrophy. Therefore I have only focussed on partial domains that are end-bounded in this study.

I investigated instances of domains with an atrophy score ≥ 0.15 , wherein at least 15% of the domain is lost. Instances of domains with atrophy scores below 0.15 were neglected since they might represent cases of peripheral structural loss, which are not true cases of atrophy. Since the atrophy score is calculated considering the domain boundaries of adjacent domains, I observed that, due to the negative inter-domain distance, the atrophy scores of nested domains have the chance of reaching values of 1 or higher, which is an artefact of the scoring system. Therefore, I have excluded nested domains from further analyses.

I manually examined domains with an atrophy score between 0.15 and 1. Among these I identified 1,287 instances of false positives (or failure modes) that were incorrectly assigned as atrophied domains due to various reasons such as gene prediction errors, circular-permuted domains, profile HMM models comprising more than one domains and others (refer to section 1.2.5). Table 1.1 lists all the failure modes or false positives.

I classify partial domains into 5 different types, based on the site of atrophy and its end-boundary, as following: (1) N-terminal end-bounded atrophy; (2) C-terminal end-bounded atrophy; (3) Upstream domain-bounded atrophy; (4) Downstream domain-bounded atrophy; and (5) Within-domain atrophy. A detailed description of the five types of atrophy is discussed in section 1.2.2.

8 true domain atrophy events with evidence from known 3-dimensional structures and a further 67 putative domain atrophy events using homologous structures were confirmed. Among the 8 instances of true domain atrophy, 6 cases are representatives of N-terminal end-bounded atrophy (2 examples from the bacterial luciferase domain and 4 examples from the AMP-binding domain) and 2 cases are representatives of downstream domain-bounded atrophy. Some of the examples of true and putative domain atrophy are discussed below. For the complete list of atrophied domains, including 67 putative cases, refer to the appendix (Table A1).

Types of failure modes or false positives	Number of instances
Gene prediction error	9
Containing nested domain	268
Multi-domain family	316
Small domain	173
Circular permutation	54
Short repeat	112
Disordered domain	85
Theoretical model	9
No other reference structure available	27
Complete structural domain	234
Total	1,287

Table 1.1 False positives and failure modes of the domain atrophy pipeline.

1.3.1 N-terminal end-bounded atrophy

Type example: Bacterial luciferase domain (Pfam: PF00296)

The bacterial luciferase domain of the non-fluorescent flavoprotein (NFP) from *Photobacterium leiognathi* luxF (UniProt: P09142) shows an atrophy score of 0.31, indicating a loss of nearly one-third of the domain's structure. The second example within the same domain family includes NFP from *Photobacterium phosphoreum* luxF (UniProt: P12745) with a similar atrophy score of 0.31.

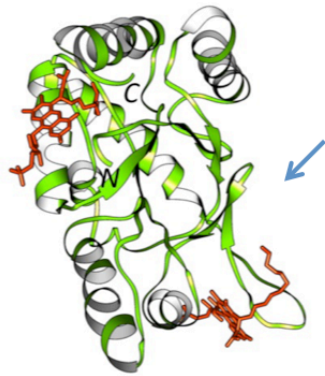
Figure 1.5A shows a schematic representation of the partial match of *P. leiognathi* luxF to the profile HMM model of the bacterial luciferase family (Pfam: PF00296). The 288 amino acid residue long luxF sequence is partially aligned with the profile HMM model beginning from HMM alignment start site coordinate 170. The HMM match state co-ordinate 170 corresponds to the amino acid residue 72 in the sequence, which indicates that there is no alignment between the N-terminal 71 amino acid residues and the first 169 unmatched states of the profile HMM model. Due to low sequence similarity the N-terminal 71 amino acid residues are not matched by the profile HMM model, but these residues are still part of the domain structure. It is clearly evident that these unmatched N-terminal 71 amino acid residues are far fewer compared to the number of unmatched HMM states; there are no amino acid residues that can be extended beyond the N-terminal sequence start site, which can completely cover or align with the remaining unmatched HMM states, therefore indicating a true loss of sequences at the N-terminal at the *P. leiognathi* luxF domain compared to a full-length canonical bacterial luciferase domain. Therefore using these values within equation 1 (refer to section 1.2.3) gives an N-terminal atrophy score $AS_N = (169 - 71)/307 = 0.31$. The C-terminal end sequence of the domain can be extended by sequence similarity to completely cover the HMM-profile and exhibits no atrophy.

The bacterial luciferase domain, homolog of the bacterial luciferase subunits (Moore and James, 1995), is present mostly among members of gammaproteobacteria. The NFP acts as a 'molecular sponge' to sequester myristylated flavine mononucleotide, the side-product of the bio-luminescence pathway (Moore and James, 1995). The structure of NFPs from *Photobacterium leiognathi* (PDB: 1NFP) and *Photobacterium phosphoreum* (PDB: 1FVP) resemble a partial TIM-barrel-like fold missing a β -strand and three α -helices (Kita et al., 1996; Moore and James, 1995) (Figure 1.5B). To compare the extent of structural loss I compared the atrophied domain with the full-length reference structure of bacterial luciferase domain from *Bacillus cereus* (PDB: 2B81), which has a complete $(\beta/\alpha)_8$ TIM-barrel fold with characteristic β -barrel structure consisting of eight alternating β -strands and α -helices (Figure 1.5C). Although from the HMM-model the atrophy was initially identified at the N-terminus of the domain, structural superpositions show that structural elements, β_1 , α_1 and β_2 , at the N-termini of 1NFP and 1FVP are intact, however the atrophied domains have no secondary structural elements that are equivalent to α_2 , β_3 , α_3 and α_4 , of the reference domain 2B81 (residues 61-125, 132-192), indicating that the atrophy is within the domain rather than at the N-terminus. Sequence alignment with luxB, a homologue of luxF, also shows atrophy within the domain (Figure 1.6A).

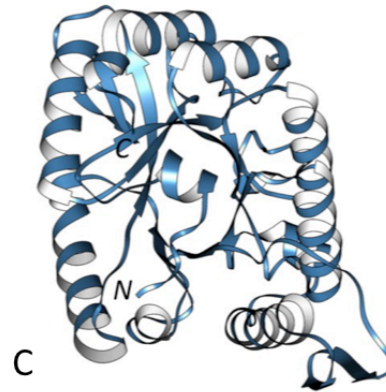
It is well known that domain cores are highly hydrophobic and their exposure to solvent leads to unfolding or instability (Miller et al., 1987; Rose et al., 1985). The hydrophobic β -barrel core of the TIM-barrel fold is shielded from the solvent by the peripheral α -helices. It is therefore interesting to know how the atrophied bacterial luciferase domains with a solvent-exposed hydrophobic cleft are stable by tolerating such large deletion. The crystal structures of atrophied domains (PDB: 1FVP) shows that the atrophied domain buries its solvent exposed hydrophobic core by forming homo-dimeric interactions (Figure 1.5D) (Kita et al., 1996; Moore et al., 1993). The two solvent-exposed clefts face each other forming a new dimer interface that shields the exposed core from the solvent and thereby forming stabilising interactions. Interestingly a similar homo-dimeric interaction is observed in the complete full-length domain 2B81 at the same side of the molecule that exhibit atrophy (Figure 1.5E).

HMM co-ordinates	170	246	307	
Sequence-alignment co-ordinates	1	72	153	228

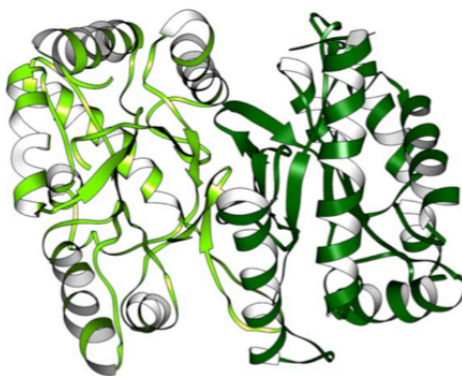
A



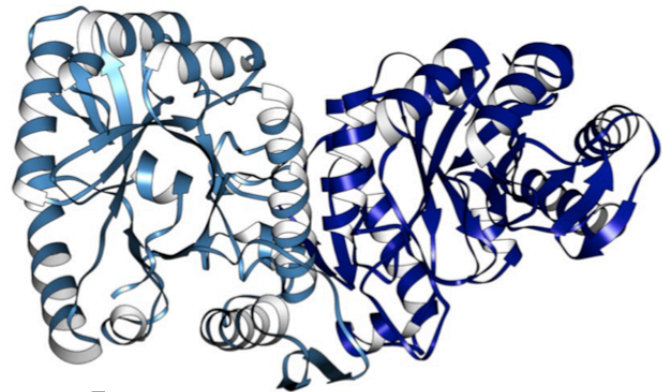
B



C



D



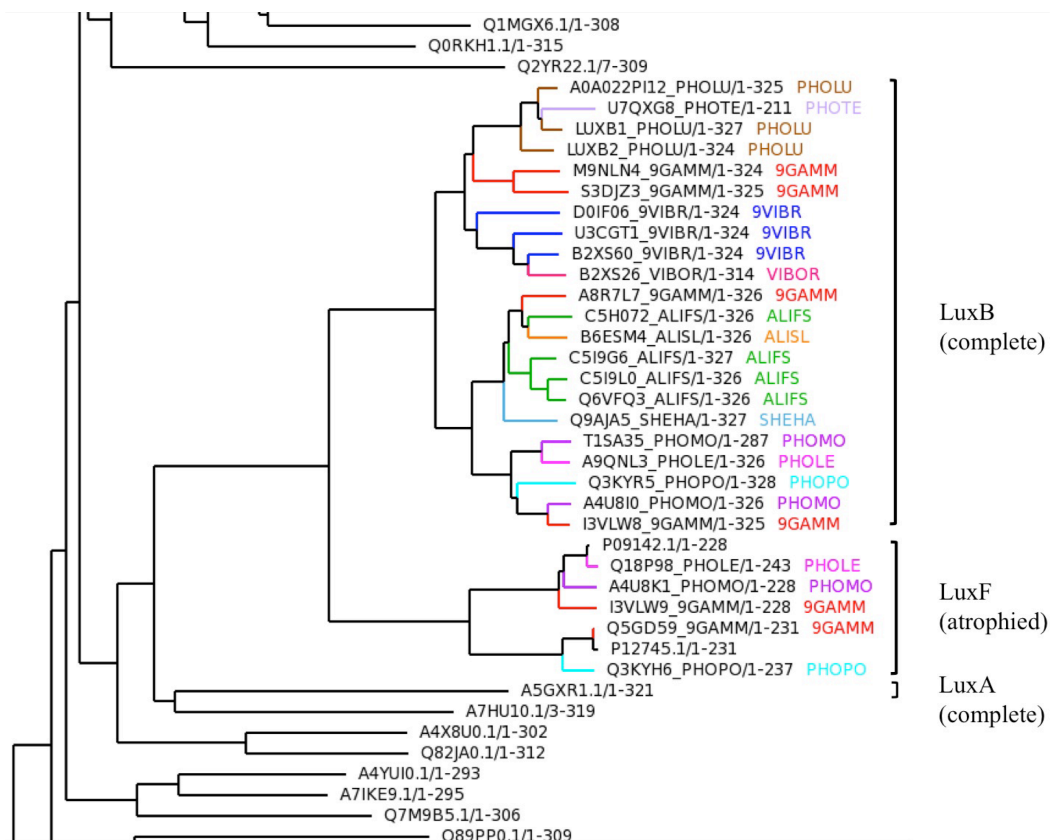
E

Figure 1.5 N-terminal end-bounded atrophy of bacterial luciferase domain.

(A) Schematic representation of Pfam bacterial luciferase domain (Pfam: PF00296) from *P. leiognathi* luxF sequence (UniProt: P09142). Numbers at the top indicate co-ordinates of profile HMM and numbers at the bottom denote co-ordinates of amino acid sequence aligned with the profile HMM. Dotted green lines indicate the missing unmatched region of the profile HMM (B) Monomer of the atrophied bacterial luciferase domain of *P. phosphoreum* non-fluorescent flavoprotein (PDB: 1FVP, light green) bound to ligand FMA (6-(3-tetradecanoic acid) flavine mononucleotide), orange. Arrow shows the solvent exposed atrophied region. (C) Monomer of the *B. cereus* reference structure (PDB: 2B81, light blue). (D) Homo-dimer complex of 1FVP. The exposed hydrophobic core due to domain atrophy is stabilised by the new dimer interface. (E) Homo-dimer complex of 2B81 showing dimerisation on the same side of the molecule.

A9QNL3_LuxB	1	M---	NFGLFFLN	QPEGMTS-	EMVLDNMVDT	VALVDKDDYHY	NRVLVSEHFSK	NGIIG	E																	
P09142_LuxF	1	MTKWN	YGVFFLN	FYHVG	QQEPLTMS	NALET	LRIIDEDT	SIYDVF	SEHHIDKS-----																	
A9QNL3_LuxB	57	PLTAVS	FLLGLTK	RKLGSL	NQVITTH	HPVRI	GEQTGL	LDQMSY	GRFILGLSDCVNDFEM																	
P09142_LuxF	56	-----	-----	-----	-----	-----	YNDETK	LAPFVSLG	KQI-----																	
A9QNL3_LuxB	117	DDFKR	QRSSQK	QFEACY	EILNEAL	TTHY	CHADDDFF	NFPRI	SVNPHCINEIKQYILASS																	
P09142_LuxF	73	-----	-----	-----	-----	-----	-----	-----	-----HILATS																	
A9QNL3_LuxB	177	MEV	EWAAK	GLPLTY	RWSDK	LAEKE	YYQRY	LAVAK	ENNVDVSNVDHQFP	LLVNIENR																
P09142_LuxF	79	PET	VVKA	AKYG	MPLLFK	WDD	SQKRI	ELLNH	YQAAA	AKFNVDIAGVPHRLMLFVNVNDNP																
A9QNL3_LuxB	237	RVAR	DEV	RKYI	ESY	VAEAY	PTDPNI	ELR	IEELLE	QHAVGKMD	EYDPTM	HAVK-----	VT													
P09142_LuxF	139	TQAK	AEL	SIY	LEDY	LSYT	-----	QAET	SI	DEI	INS	NAA	GN-----	FDTCLHHVAEMAQGLN												
A9QNL3_LuxB	292	GSKN	VLLS	FES	MKNK	DDVTK	L	IN	M	NQKI	----	KD	NLIK	326												
P09142_LuxF	190	NKVD	FL	CF	FES	MKD	QEN	KK	S	L	M	I	N	F	DKR	V	I	N	R	K	E	H	N	L	N	228

A



B

Figure 1.6 Pairwise sequence alignments and phylogenetic analysis of the bacterial luciferase domains. (A) The luciferase B subunit protein luxB (UniProt: A9QNL3) is a homologue of luxF. Compared to the homologue, luxF shows deletions of amino acid residues within the domain. (B) The bacterial luciferase, non-fluorescent flavoprotein (LuxF) and the alkanal monooxygenase beta (LuxB) share a common ancestor. The ancestral fold of luciferase is a complete TIM-barrel fold observed in LuxB and LuxA proteins. Figure B reused from (Prakash and Bateman, 2015), doi: 10.1186/s13059-015-0655-8.

The core fold of a domain, during the course of evolution, can embellish secondary structural elements – termed domain elaborations, which is the opposite mechanism of domain atrophy. Therefore, to distinguish domain atrophy from domain elaboration, it is important to determine the phylogenetic relationships between the two variations of domain folds. To determine whether the ancestral fold of bacterial luciferase domains is a complete TIM-barrel fold, I analysed the phylogenetic relation between three clades of the bacterial luciferase family- luxA, luxB and luxF. The luciferase protein is a hetero-dimeric complex of luxA and luxB polypeptide chains (Close et al., 2012) and they both exhibit complete $(\beta/\alpha)_8$ TIM-barrel fold (Fisher et al., 1996). I observe that the luxF protein clade is completely enclosed by luxB (Figure 1.6B) and that the ancestral fold must be the complete TIM-barrel fold observed in luxA and luxB proteins.

Type example: AMP-binding domain (Pfam: PF00501)

This is the second example of a true domain atrophy event observed at the N-terminal end of the domain. The phenylacetate-coenzymeA ligase, Paak1, from *Burkholderia cenocepacia* (UniProt: B4E7B5) is a 432 amino acid residue long protein composed of two domains – the N-terminal AMP-binding domain (Pfam: PF00501) and the AMP-binding C-terminal domain (AMP-binding_C2, Pfam: PF14535). Sequence scan against the Pfam HMM model for this domain family shows that the N-terminal adenosine monophosphate (AMP) binding domain has atrophy at the N-terminus. The AMP-binding C-terminal domain does not show any atrophy.

Figure 1.7A shows a schematic representation of the partial match of the N-terminal AMP-binding domain. The HMM model of the N-terminal AMP-binding domain has a length of 417, of which co-ordinates 152 through to 417 align with the domain sequence covering amino acid residues 72 to 334. It should be noted here that the N-terminal region from amino acid 5 to 71 is probabilistically matched to the profile HMM (region coloured in light green, Figure 1.7A), which indicates that the N-terminal domain boundary within this region is not accurate.

Therefore to strictly demarcate the domain boundaries while computing atrophy scores I have used the alignment co-ordinates of protein sequences, rather than their envelope co-ordinates, which results in an atrophy score of 0.19.

Phenylacetate-coenzymeA ligases are adenylate-forming enzymes involved in the metabolism of phenylacetate (Martinez-Blanco et al., 1990). The enzyme links the phosphoryl moiety of AMP to the carboxyl group of the substrate, activating it before transferring to the acceptor CoA (Martinez-Blanco et al., 1990). The full length N-terminal AMP-binding domain of 4-chlorobenzoyl CoA ligase from *Alcaligenes sp.* (PDB: 3CW9) was used as the reference structure to identify structural loss within the N-terminal AMP-binding domain of *B. cenocepacia*. The reference N-terminal AMP-binding domain is an α/β structure comprising three distinct β -sheet with a cleft containing the binding pocket (Reger et al., 2008) (Figure 1.7C). Structural comparison of *B. cenocepacia* N-terminal AMP-binding domain (PDB: 2Y27) with the reference structure 3CW9 indicates that the atrophied N-terminal AMP-binding domain lacks region comprising residues 1 to 150 of the reference structure that forms the first sub-domain (Figure 1.7B, D). It can be seen that the first sub-domain of the reference structure has minimal interactions with the substrates during adenylation and thioester formation process (Reger et al., 2008), therefore the absence of this region in the atrophied N-terminal AMP-binding domain may not have affected its catalytic function. The crystal structure of the *B. cenocepacia* Paak1, 2Y27, shows the atrophied N-terminal AMP-binding domain in homo-dimeric interactions (Figure 1.7E) (Law and Boulanger, 2011). Interestingly this interaction involving the second sub-domain region (residues 151 to 322) results in the formation of an intramolecular β -sheet at the interface that appears to mimic the structural arrangement of the deleted first sub-domain observed in full-length reference structure (Law and Boulanger, 2011), shown by region circled in red in Figure 1.7E.

Again, similar to the atrophied bacterial luciferase domain discussed above, homo-dimeric interactions are observed to bury the solvent exposed core residues and compensating for the deleted structural region. In addition to the

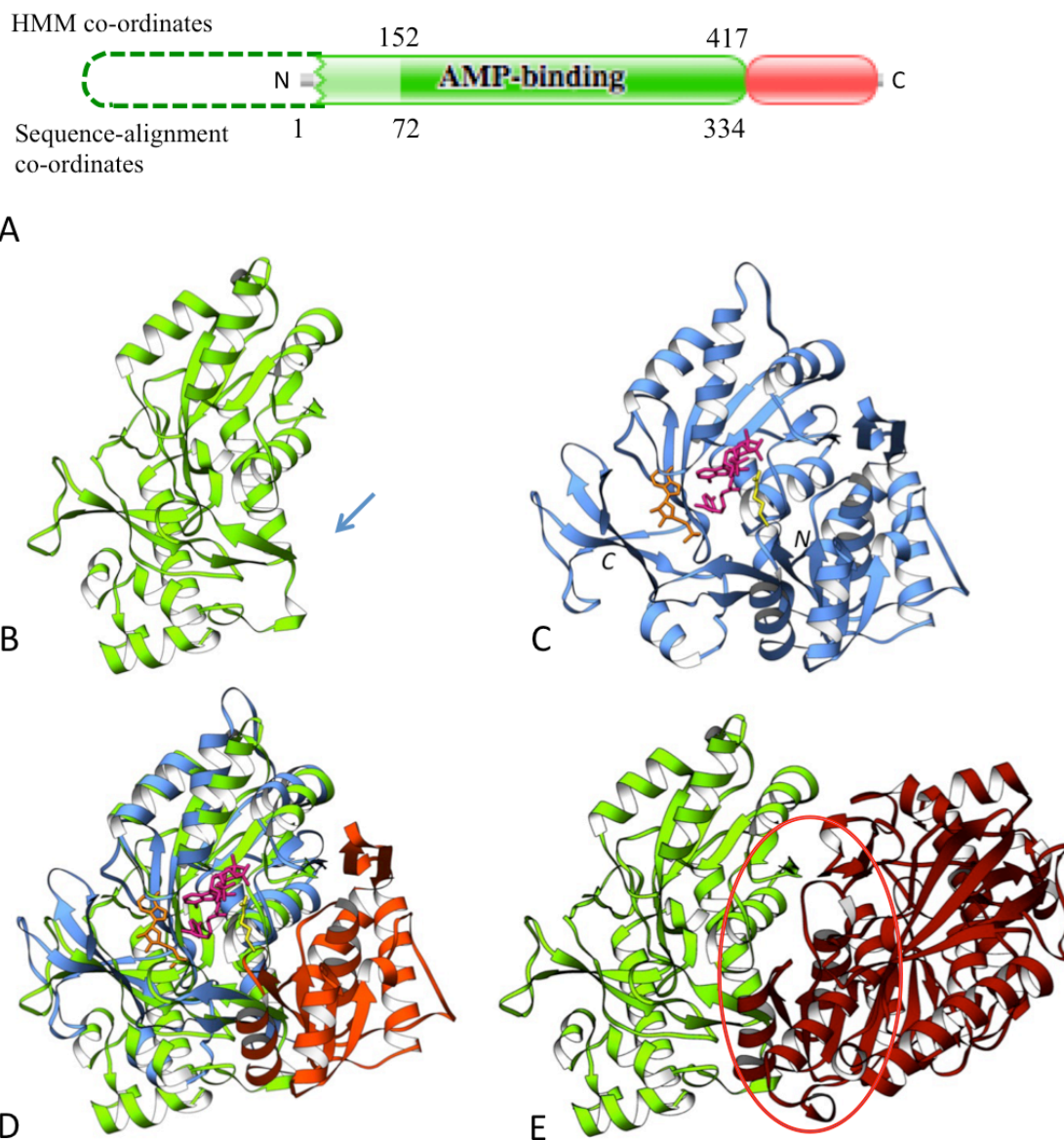


Figure 1.7 N-terminal end-bounded atrophy of AMP-binding domain. (A) Schematic representation of N-terminal AMP-binding domain (Pfam: PF00501, green) and the AMP-binding C-terminal domain (AMP-binding_C2, Pfam: PF14535, red) of *B. cenocepacia* phenylacetate-coenzymeA ligase (UniProt: B4E7B5). Sequence region probabilistically matched to the profile HMM is shown in light green. Dashed lines indicate the missing unmatched region of the profile HMM (B) Monomer of the atrophied N-terminal AMP-binding domain (PDB: 2Y27); arrow indicates region of structural loss at the N-terminus. (C) Reference structure of full length N-terminal AMP-binding domain of 4-chlorobenzoyl CoA ligase from *Alcaligenes sp.* (PDB: 3CW9). (D) Structural superposition of 2Y27 (green) and 3CW9 (blue), the atrophied first sub-domain is shown in orange. (E) Homo-dimer of atrophied domain (PDB: 2Y27) showing structural arrangement (red circle) that mimics the deleted first sub-domain of the reference structure.

atrophy observed in *B. cenocepacia* phenylacetate-coenzymeA ligase, similar cases of atrophied N-terminal AMP-binding domains were observed in *Bacteroides thetaiotaomicron* phenylacetate-coenzymeA ligase (UniProt: Q8AAN6, PDB: 3QOV) and *Enterobacter agglomerans* phenazine antibiotic biosynthesis protein (UniProt: Q8GPH0, PDB: 3HGU), each with an atrophy score of 0.16.

1.3.2 C-terminal end-bounded atrophy

Type example: Ral-GTPase-activating protein domain (Pfam: PF02145)

The Ral-GTPase-activating protein (RapGAP) domain of rat Ral-GTPase-activating subunit α -1 isoform-1 (UniProt: O55007) was predicted to exhibit atrophy at the C-terminal end of the domain at the sequence end.

The isoform-1 is 747 amino acid residues long with a single RapGAP domain at its C-terminus and no detectable Pfam domains at its N-terminal region. The sequence only aligns to the first 75 match states of the domain profile HMM model, which corresponds to sequence region 650 to 722. The remaining sequence of 25 amino acid residues at the C-terminal, that are not aligned with the profile HMM, is not long enough to be extended further to completely match the missing 113 profile HMM match states (Figure 1.8A). Calculation of the C-terminal atrophy using equation 2 (refer to section 1.2.3) results in an atrophy score of $AS_c = (113 - 25)/188 = 0.46$, indicating structural loss of nearly half the domain.

The GTPase-activating proteins (GAPs) terminate G-protein signalling by inducing hydrolysis of bound GTP to GDP (Bos et al., 2007). The full-length Rap1-GAP catalytic domain of the human Rap1-GAP protein (PDB: 3BRW) was used as the reference structure. There are no experimental structures solved of the rat RapGAP domain, therefore, instead of structural superposition I used pairwise sequence alignment and mapped the sequence deletion on to the domain's homologous structure to infer atrophy of the rat RapGAP domain. Pairwise

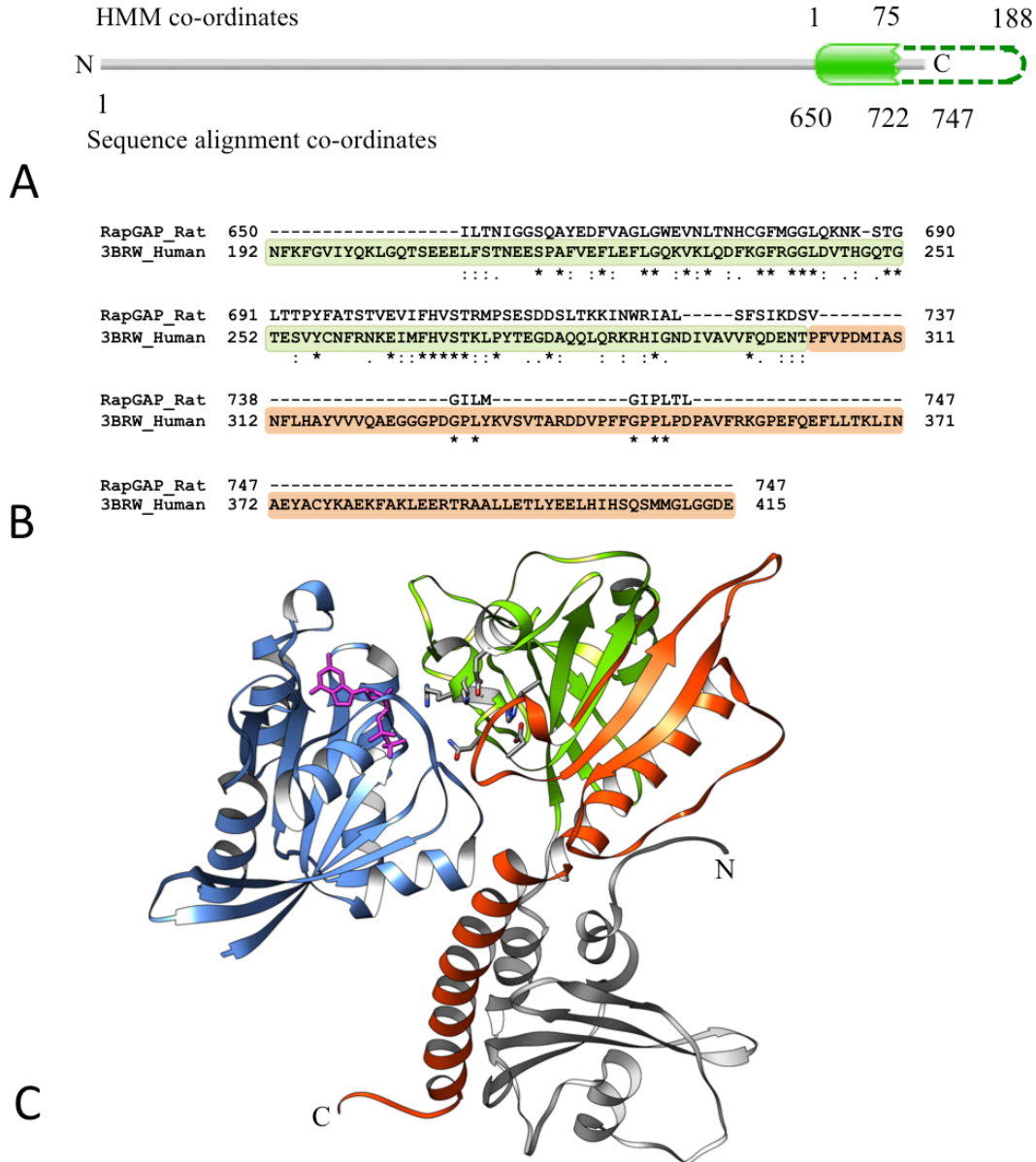


Figure 1.8 C-terminal end-bounded atrophy of rat RapGAP domain. (A) Schematic representation of RapGAP domain or rat Ral-GTPase-activating subunit α -1 isoform-1 (UniProt: O55007). Dashed lines indicate Pfam profile HMM model that is not aligned to the sequence. (B) Pairwise sequence alignment of the rat RapGAP and human RapGAP domains. The human RapGAP domain sequence highlighted in green corresponds to that region in the rat RapGAP domain that is matched by the profile HMM and the region highlighted in orange corresponds to that region in the rat RapGAP domain that is absent or atrophied. (C) Sequence mapping from pairwise alignment on to the homologous reference structure in human (PDB: 3BRW) shows the region of rat RapGAP domain that remains (green) and the region of RapGAP domain that is atrophied (orange). The dimerisation domain at the N-terminal is shown in gray and the interacting protein Rap1B is shown in blue. Amino acid residues involved in interactions with Rap1B are shown as gray sticks.

alignment of the rat RapGAP and the human Rap1GAP domain sequences clearly indicates a large deletion at the C-terminal end of the rat RapGAP domain, while showing 30% sequence identity in the aligned region (Figure 1.8B). Sequence mapping of the rat RapGAP domain onto the human Rap1-GAP protein (PDB: 3BRW) indicates structural loss in the catalytic domain (residues 301-414) (Figure 1.8C, orange). The catalytic domain is an α/β structure with mixed parallel/antiparallel arrangement of β -strands and a conserved C-terminal alpha helix and interacts with the Rap1B protein (Scrima et al., 2008). The catalytic centre comprising Asn290 is close to the nucleotide-binding region and the protein interface (Daumke et al., 2004; Scrima et al., 2008). The observed atrophy does not affect the catalytic centre or residues involved in Rap1B interaction, therefore suggesting that the atrophied domain may be functional.

Interestingly this domain atrophy is not observed in the other isoforms of rat RapGAP protein. Isoform-2 (906 amino acids) and isoform-3 (2,035 amino acids) have complete full-length RapGAP domains suggesting that exon loss mediated by alternative splicing could be a probable mechanism in mediating atrophy at the C-terminal end of RapGAP domain in isoform-1.

1.3.3 Upstream domain-bounded atrophy

125 cases were initially identified where atrophy was predicted in domain regions that were bounded by an upstream non-homologous domain. After manual inspection none of the identified examples were determined as true cases of domain atrophy, but were failure modes of the pipeline.

1.3.4 Downstream domain-bounded atrophy

Type example: 2-hydroxyacid dehydrogenase, NAD binding domain (2-Hacid_dh_C) (Pfam: PF02826)

The 2-hydroxyacid dehydrogenase NAD-binding (2-Hacid_dh_C) domain in *Staphylococcus aureus* PurK (UniProt: A6QFS4) was identified with atrophy at

the C-terminus of the domain. The domain is found to have nearly a quarter of the canonical structure missing indicated by the atrophy score of 0.23 (Figure 1.9A). The C-terminal end of the 2-Hacid_dh_C domain is bounded by the downstream ATP-grasp domain. Similar atrophy is also observed in the 2-Hacid_dh_C domain of *Bacillus anthracis* (UniProt: C3PBM5, PDB: 3Q2O). These two cases represent true downstream domain-bounded atrophy events, which are validated by experimental structures.

2-Hacid_dh_C domains are found in dehydrogenases and oxidoreductases in prokaryotes and eukaryotes. The bacterial PurK and PurE proteins are involved in a two-step conversion of 5-aminoimidazole ribonucleotide to 4-carboxy-5-aminoimidazole ribonucleotide (Brugarolas et al., 2011). The crystal structure of *S. aureus* 2-Hacid_dh_C domain (PDB: 3ORQ) adopts a partial Rossmann fold and exhibits secondary structural embellishments including an additional alpha helix (α_2) and β -strand (β_3) (Figure 1.9B) (Brugarolas et al., 2011). The canonical full-length 2-Hacid_dh_C reference structure for comparison was chosen from *Lactobacillus jensenii* D-lactate dehydrogenase (PDB: 4PRL), which adopts a Rossmann fold with six parallel β -strands bound to NAD (Figure 1.9C) (Kim et al., 2014).

Superposition of structures of the atrophied domain (3ORQ) and the reference domain (4PRL) indicates a partial loss of secondary structural elements at the C-terminus of 3ORQ. The β -strands β_2 - β_1 - β_4 of the atrophied domain align with the strands β_2 - β_1 - β_3 of the reference domain, respectively, but there are no equivalent residues in the atrophied domain to match strands β_4 - β_5 - β_6 of the reference domain (residues 225-298). Interestingly the atrophy at the C-terminus of the domain does not affect the ligand-binding site at the N-terminus. The atrophied domain retains the 'reverse' Rossmann fold motif (GXXGXG) in the loop connecting β_1 and α_1 of the atrophied domain. I observe that the binding-motif and interaction sites within the atrophied and the reference domain are conserved and present at structurally equivalent locations (3ORQ; residues: 16-21, 38-41 and 4PRL; residues: 153-158, 175-178). The presence of the conserved binding motif suggests that the atrophied domain might bind to the adenine

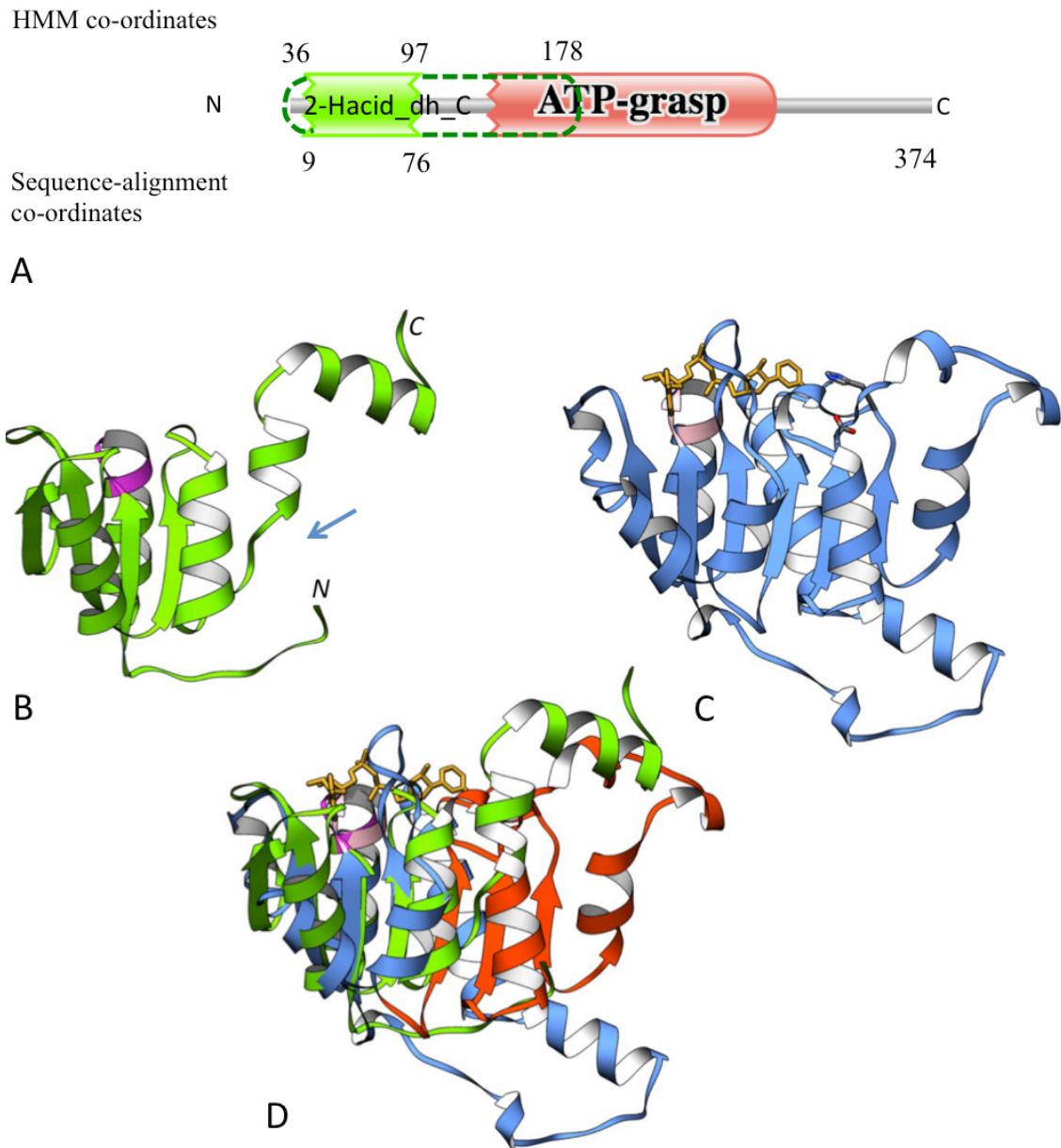


Figure 1.9 Downstream domain-bounded atrophy of 2-Hacid_dh_C domain.

(A) Schematic representation of domain architecture in *Staphylococcus aureus* PurK (UniProt: A6QFS4). Atrophy at the C-terminal region of 2-Hacid_dh_C domain (green) is indicated by dashed lines. (B) Crystal structure of the atrophied *S. aureus* 2-Hacid_dh_C domain (PDB: 3ORQ). Arrow indicates region of structural loss at the C-terminus. (C) The complete reference structure of *L. jensenii* 2-Hacid_dh_C domain (PDB: 4PRL). The conserved NAD (yellow stick) binding 'reverse' Rossmann motifs in atrophied and reference domains are highlighted in dark pink and light pink respectively. (D) Superposition shows the structural elements of 4PRL, highlighted in orange, that are atrophied in 3ORQ.

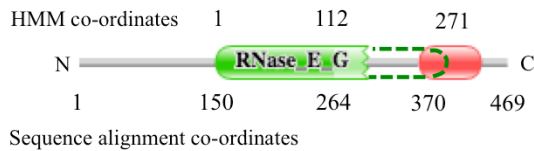
moiety, however, other important residues at the C-terminus of the reference domain such as Asp259 and the catalytically important His295 located near the nicotinamide moiety are absent (Kim et al., 2014; Tishkov et al., 1996).

Type example: RNase_E_G domain (Pfam: PF10150)

The RNase_E_G domain is an example of a putative downstream domain-bounded atrophy, for which structural loss was inferred through sequence mapping onto a homologous structure. The RNase E domain of *Pyrococcus furiosus* RNA-binding protein AU-1 was identified with atrophy at the C-terminus indicated by an atrophy score of 0.19.

The *P. furiosus* RNA-binding protein AU-1 (UniProt: Q8U4Q7) is 469 amino acid residues long and consists of the RNase_E_G domain (Pfam: PF10150; Figure 1.10A, green) and the domain of unknown function DUF402 (Pfam: PF04167; Figure 1.10A, red) at the C-terminus. The RNA-binding protein is a large oligomeric complex that binds specifically to AU-rich RNA sequences and involved in RNA metabolic processes (Kanai et al., 2003).

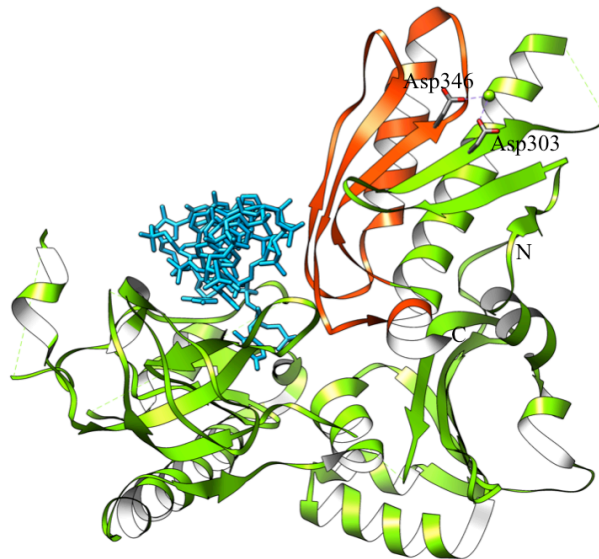
The N-terminal region of the profile HMM matches the *P. furiosus* RNase_E_G domain from amino acid residues 150 to 264. The sequence region between the domains (residues 265 to 369) is poorly matched by the RNase_E_G profile HMM; but although poorly matched this region can still be part of the RNase_E_G domain suggesting that the maximum size the RNase_E_G domain within this sequence could have is from residues 150 to 369. Therefore I have extended the RNase_E_G domain boundary till the amino acid residue 369 to include sequences that were missed by the profile HMM. Pairwise sequence alignment of the extended *P. furiosus* RNase_E_G domain (amino acid residues 150 to 369) with the *E. coli* RNase_E_G reference domain (UniProt: P21513; PDB: 2C0B, amino acid residues 121 to 391) shows sequence loss at the C-terminal end of the *P. furiosus* RNase_E_G domain (Figure 1.10B) indicating that despite extending the domain boundary the RNase_E_G domain is atrophied at the C-terminal end.



A

Q8U4Q7_Pfur	150	TIPGDYAVLIPKPIGVQRHVKISRKIKDP	---	RERLRILGL	-----	SVDLGE	195
2C0B_Ecoli	121	SLAGSYLVLMPNNPRA	---	GGISRRIEGDDRTELKEALASLEL	PEGMGLIVRTAGVGKSA		177
		:: * . * * * : :		*** : * . . .	: * * * *	. * . .	
Q8U4Q7_Pfur	196	WGVLRWRTAAAYKDWNTRLDELVRLSKIADKLKEAEKFSAPAEIIEGREIYEIEFGGGVKK					255
2C0B_Ecoli	178	EALQWDLRFRLKHWEAIKKAES	-----	RPAPFLIHQESNVIVRAFRDYL	RQ		224
		. : * : * . * : : : .		** * : : :	* . : :		
Q8U4Q7_Pfur	256	KLDEIRN	--	EVVPTIEGHHQFKSYDPEFTLAVDVAEGILAKLPSQ	--	RQKISKGFLEAII	311
2C0B_Ecoli	225	DIGEILIDNPKVLELARQHIAALGRPDFSSKIKLYTGEIPLFSHYQIESAFQR	---				281
		. . . * * * : * : * * : : : * : : : . . . * . . . *					
Q8U4Q7_Pfur	312	TSKGPKVGWIFTLNHVKPDGQIIKIGPGEVIE	-	VSTDPLKVTIKRYLRPGKFYDGL	EV		369
2C0B_Ecoli	282	-----	EVRLPSGG	SIVIDSTEALTAIDINSARATRGDIEETA	FNLEAAD		328
			: * . * * * . * . : : : : * : . . * . . . *				
Q8U4Q7_Pfur	369	-----					369
2C0B_Ecoli	329	EIARQLRLRDLGGLIVIDFIDMTPVRHQRAVENRLREAVRQDRARIQISHISRFGLEMS					388
Q8U4Q7_Pfur	369	---					369
2C0B_Ecoli	389	RQR					391

B



C

Figure 1.10 Downstream domain-bounded atrophy of RNase_E_G domain.

(A) Schematic representation of RNase_E_G domain from *P. furiosus* RNA-binding protein AU-1 (UniProt: Q8U4Q7) with RNase_E_G domain (green) and DUF402 (red). Dotted lines denote the missing region from the C-terminal of the domain. (B) Pairwise sequence alignment showing sequence within the *E. coli* reference domain that is deleted (orange) in the *P. furiosus* RNase_E_G domain. (C) Sequence mapping on to the structure of *E. coli* RNase_E_G (PDB: 2C0B) shows structural elements that are lost in *P. furiosus* RNase_E_G domain, highlighted in orange. Single strand RNA is shown in blue, and active site residues are shown as gray sticks.

The *E. coli* RNase_E_G domain (PDB: 2C0B) is a large multi-domain structure consisting of S1, 5' sensing region, RNase H and DNase I subdomains (Callaghan et al., 2005). The DNase I subdomain is the catalytic centre of the complex and is made of two α -helices and six antiparallel β -strands in the order $\beta_{1-2-3-4-6-5}$. The active site residues Asp303 and Asp346, present on β_3 and β_4 respectively, coordinate a Magnesium ion, which cleaves the scissile phosphate on the RNA backbone through nucleophilic attack (Callaghan et al., 2005). Sequence mapping shows the region of atrophy in the C-terminus of the DNase I subdomain (residues 339 to 393) (Figure 1.10C, orange), with which the atrophied domain shares 21% sequence identity. The atrophied domain region as seen from sequence mapping shows loss of one of the active site residue Asp346. The *P. furiosus* RNA-binding protein AU-1 is a homo-oligomer trimeric complex (Kanai et al., 2003). Since the experimental structure of the atrophied domain is not available, I hypothesize that the interactions with the C-terminal domain, DUF402 might stabilise the atrophied domain within the trimeric complex.

1.3.5 Within-domain atrophy

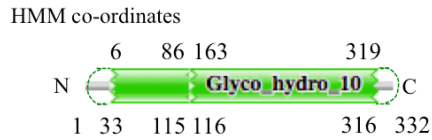
Type example: Glycosyl hydrolase family 10 (Pfam: PF00331)

The glycosyl hydrolase family 10 (Glyco_hydro_10) domain from the endo-1,4-beta-xylanase (UniProt: P07529) of *Cryptococcus albidus* was identified with atrophy within the domain interior.

The endo-1,4-beta-xylanase protein is 332 amino acid residues long made of a single glyco_hydro_10 domain. Xylanase are found in bacteria, fungi and other microbes (Beg et al., 2001; Polizeli et al., 2005), which degrade hemicellulose by breaking down beta-1,4-xylan into xylose. The *C. albidus* endo-1,4-beta-xylanase is an inducible extracellular enzyme with xylobiose as its natural inducer (Biely et al., 1980). Sequence scan of *C. albidus* endo-1,4-beta-xylanase against the profile HMM of the glyco_hydro_10 domain family matches the full-length sequence except at the amino acid residues 115 and 116, wherein the match state co-ordinates 85 to 162 of the profile HMM are not found, indicating

deletion of sequences from within the domain (Figure 1.11A). Using equation 3, the domain atrophy score of *C. albidus* glyco_hydro_10 was calculated to be $AS_w = (77 - 0)/320 = 0.24$.

For comparison I used the complete glyco_hydro_10 domain from endo-1,4-beta-xylanase of *Thermotoga petrophila* (PDB: 3NJ3) as the reference domain, which has a TIM-barrel fold (Santos et al., 2010). Pairwise sequence alignment clearly indicates deletion of a large stretch of residues within the *C. albidus* glyco_hydro_10 domain (Figure 1.11B). Using sequence mapping I observe that the deleted sequence corresponds to β -strand β_4 and two core alpha helices, α_3 and α_4 (residues 110-176) of the reference domain (Figure 1.11C, orange). It can be seen that one of the active site residues Glu150, present on β_4 , which interacts with xylobiose, is lost due to the atrophy in *C. albidus* glyco_hydro_10 domain. Interestingly the *C. albidus* endo-1,4-beta-xylanase was reported to be inefficient in degrading xylobiose and hydrolyses xylotriose at very slow rates compared to the longer substrates such as xylotetraose (Biely et al., 1981; Biely et al., 1980). I suggest that the inefficiency of the enzyme is due to the loss of one of its active site residues Glu150. Since the atrophy within the domain exposes its hydrophobic β -barrel to the solvent, I predict that similar to the atrophied bacterial luciferase domain, the *C. albidus* glyco_hydro_10 domain also highly likely undergoes oligomerisation, presumably by forming homo-dimeric interactions to stabilise its atrophied structure.

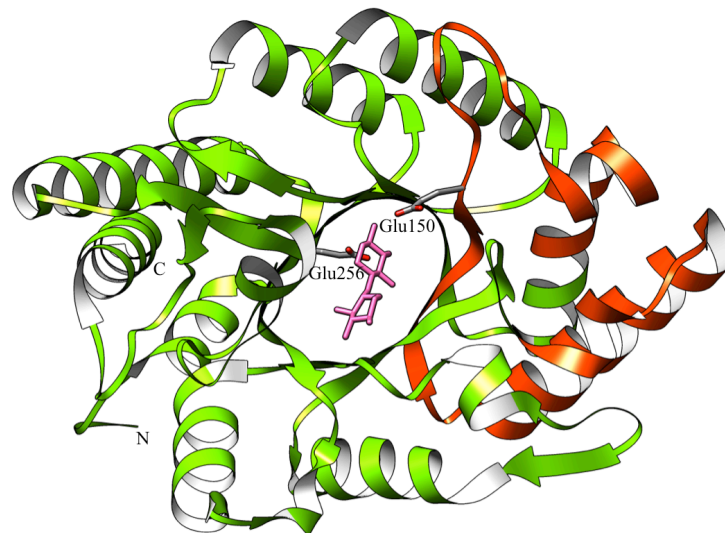


Sequence alignment co-ordinates

A

P07529_Calb	1	--MLSSTLLAILLSALALTSVQAAPADKNSLDYLANKAGKRYLGTAVQ--SPQLVPGSQY	56
3NJ3_Tpet	1	MKILPSV-LILLGCVPV-----FSSQNVSLRELAEKL-NIYIGFAAINNFWLSDEEKY	53
		:* *. * :*: *	
P07529_Calb	57	VQILESQFDAITPENEMKWEVVEPTEGNFDFGTGDKIVAEAKKTGSLLRGHNICWDSQLR	116
3NJ3_Tpet	54	MEVARREFNILTPENQMKWDTIHPERDRYNFTPAEKHVEFAEENNMIVHGHTLVVHNQLP	113
		: : . : * : : * * * : * * * : : * : : : : : * : : * * : : * * : : * * : : *	
P07529_Calb	116	-----	116
3NJ3_Tpet	114	GWITGREWTKHEELNVLEDHIKTVVSHFKGRVKIWDVVNEAVSDSGTYRESVWYKTIGFE	173
P07529_Calb	117	-----YAHEVAPKMKLCINDYNIETVNAKSQAMAKVAAGLLAKGAPLHCIGMFKNAKR	169
3NJ3_Tpet	174	YIEKAFRWTKADPDAILIYNDYSIEEINAKSNFVYVMIKELKEKGVDPVDGIGFQMHIDY	233
		: : * . * . * * * . * * : * * * : : : : * * * . * * : : .	
P07529_Calb	170	RSSGLLIRTASSGLESHFIGGSTPKDIPAAMNLFSDQGLEVPMTELDVRIPVNGNDMPAN	229
3NJ3_Tpet	234	R-----GLNYDSFRNLERFAKLGQLIYITEMDVRIPLSGSEDYY-	273
		* : : * . * * : : * * * * : * * : * * * * : * . :	
P07529_Calb	230	ATVAKEQVDDYYSVSACLGNLDCPGVSIWQFADPTSWIPGVFKGLIAVSCVTFSGCLLQ	289
3NJ3_Tpet	274	---LKKQAEICAKIFDIDLNPVAKAIQFWGFTDKYSWVPGFFK-----	314
		* : * : . . . * * * . . : : * * * * * * * * * * * * * * * * * * *	
P07529_Calb	290	YCVGYGAALLYDAQYQPKSTYYVYVQALKDGNKSGSKFHGIKL	332
3NJ3_Tpet	315	---GYGKALLFDENYNPKPCYYAIKEVLEKKIEERK-----	347
		* *	

B



C

Figure 1.11 Within-domain atrophy of glyco_hydro_10 domain. (A) Schematic representation of glyco_hydro_10 domain from *C. albidus* endo-1,4-beta-xylanase (UniProt: P07529). (B) Pairwise sequence alignment shows amino acids in *T. petrophila* (highlighted in orange) that are deleted in *C. albidus*. (C) Sequence mapping onto *T. petrophila* glyco_hydro_10 reference domain (PDB: 3NJ3) shows region of atrophy (orange) within the *C. albidus* glyco_hydro_10 domain. Active site residues are shown in gray and xylobiose in pink.

1.4 Conclusion

Protein evolution is marked by gradual changes that affect both their sequence and structural composition. On a minor scale mutations of single amino acid residues are frequently observed and on larger scales evolutionary events such as domain duplication, deletion, recombination and exon shuffling are less frequently observed but they significantly influence the structure and/or modify the functions of proteins. Many studies have explored these well-known mechanisms in trying to understand the evolution of protein domains. However, one such domain evolutionary mechanism that has so far not been subject of systematic analysis is domain atrophy. In this study I have investigated this less-known evolutionary event, in which domains undergo significant loss of core structural elements leading to a partial structure or incomplete fold, yet may still be functional.

By analysing the variations of domain boundaries across 14,831 Pfam domains, I find that the occurrence of domain atrophy is extremely rare. Only 0.005% of the total domain instances analysed showed significant loss of core structures. Deletion of structural elements, as observed in atrophied domains, would contribute to significant changes in the energetics of the fold leading to fold instability and it is one of the main reasons for such rare occurrences of atrophied domains. It is known that amino acid residues within the hydrophobic core are largely conserved so much so that mutation of even a single residue could lead to unfolding (Lee et al., 2010). But on the other hand domain elaborations do not influence the domain core in a similar way, since most elaborations occur on the surface where they are easily tolerated (Sandhya et al., 2009) and therefore elaborations occur more frequently than domain atrophy. If deletion or mutation of a single amino acid residue could lead to unfolding then it is highly likely that atrophied domains employ mechanisms that help stabilise their fold and functions.

One of the mechanism that is commonly observed among atrophied domains is the formation of homo-dimeric complexes. The atrophied regions within the

homo-dimeric complex interact with each other, which prevent the exposure of their hydrophobic cores to the solvent. For example, the atrophied bacterial luciferase domain (Figure 1.5D) forms homo-dimeric complexes, wherein their mutual interaction shields their exposed β -barrels from the solvent. The juxtaposition of the atrophied regions also prevents these exposed surfaces, with 'sticky' residues, in forming run-away homopolymeric interactions that could lead to protein aggregations. For example the atrophied AMP-binding domain (Figure 1.7E) forms a homo-dimeric complex wherein the two domains are oriented in head-to-tail fashion (PDB: 2Y27) (Law and Boulanger, 2011) (Figure 1.12). This symmetric orientation causes the head of one monomer to interact with the atrophied tail of the other, thus mutually burying their exposed surfaces. As shown in Figure 1.12E,F any non-symmetrical homo-dimeric interactions for burying the exposed atrophied surface would lead to a homopolymeric interactions causing aggregation.

An important issue is to understand the events that cause domain atrophy. Mutations causing a premature stop codon could lead to atrophy of the domain at the C-terminus and similarly mutational events leading to creation of a new downstream start codon could lead to domain atrophy at the N-terminus. Exon loss or deletions are other mechanisms that could cause domain atrophy. Alternative splicing is one of the mechanisms that allows exploration of domain sequence while still retaining functions (Birzele et al., 2008). Analysis of splice variants from the ENCODE dataset have shown a large number of splice sites within functional domains (Tress et al., 2007). Alternative splicing in human proteins was previously shown to result in fewer partial domains than expected and these partial domains were reported to affect a significant number of functional sites (Kriventseva et al., 2003). Table 1.2 summaries the number of domain atrophy cases observed within various categories in this study. It is unclear whether the distribution of domain atrophy events in these categories is the reflection of true proportions or due to the methodology of the study.

The protein degradation machinery targets unfolded or misfolded proteins by recognising their solvent exposed hydrophobic residues. It is unclear how partial

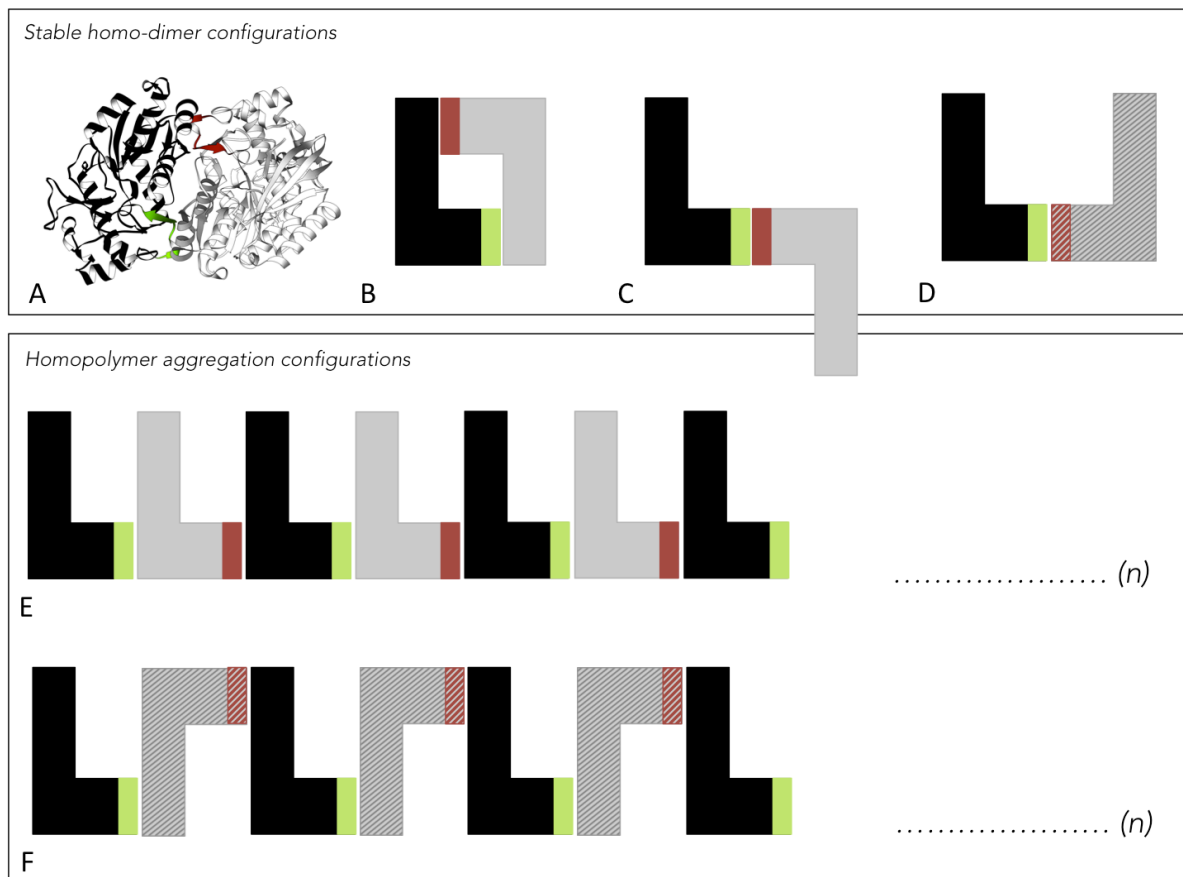


Figure 1.12 Homo-dimeric interactions of atrophied domains. Illustrations show the various modes of possible homo-dimeric interactions involving atrophied interfaces. (A) The homo-dimeric interactions of atrophied N-terminal AMP-binding domains of *B. cenocepacia* (PDB: 2Y27). The monomers are shown in black and gray and their atrophied regions at the interface are coloured green and brown respectively. (B) A schematic representation of the homo-dimeric interactions seen in (A). (B, C, D) show symmetric homo-dimeric interactions wherein the atrophied regions undergo ‘closure’, such that the solvent exposed hydrophobic regions with ‘sticky’ ends are stabilised by their mutual interactions. The surfaces with solid and hatched lines show the anterior and posterior faces or vice versa. (E, F) show asymmetric interactions, wherein the interactions at the atrophied regions do not result in closure thereby forming homopolymeric interactions that could lead to aggregations.

		N-terminal end-bounded atrophy	C-terminal end-bounded atrophy	Upstream domain-bounded atrophy	Downstream domain-bounded atrophy	Within-domain atrophy
	Type	499 (100%)	468 (100%)	125 (100%)	331 (100%)	213 (100%)
Structure available	True domain atrophy	6 (1.20%)	0 (0.00%)	0 (0.00%)	2 (0.60%)	0 (0.00%)
	False positive	161 (32.26%)	119 (25.42%)	27 (21.60%)	104 (31.41%)	45 (21.13%)
Homologous structure available	Putative domain atrophy	34 (6.81%)	26 (5.55%)	0 (0.00%)	3 (0.90%)	4 (1.88%)
	False positive	216 (43.28%)	218 (46.58%)	91 (72.80%)	207 (62.53%)	99 (46.48%)
No structure available	Unknown	82 (16.43%)	105 (22.43%)	7 (5.60%)	15 (4.53%)	65 (30.52%)

Table 1.2 Summary of various types of domain atrophy. Classification of domains identified by the pipeline into various classes after manual inspection. Only domain atrophy instances with atrophy scores between 0.15 and 1 were manually inspected.

domains evade degradation. Chaperones such as BiP maintain unfolded or misfolded proteins in a folding-competent state and evade degradation until then (Schroder and Kaufman, 2005). I speculate that atrophied domains evade recognition by the degradation machinery by burying their hydrophobic core through interactions with other subunits or proteins or are protected by chaperones until stably folded upon complex formation.

During the course of this work Dr. William Pearson and Dr. Deborah Triant at the University of Virginia carried out an independent study, which investigated the causes of partial domains (Triant and Pearson, 2015). While our study focuses on identifying true cases of domain atrophy, (Triant and Pearson, 2015) have focussed on addressing the computational origins of partial domains. The authors had found that 5% to 10% of protein domains in Pfam have only a fraction of the sequences present in them and nearly 4% of domains have more than half of their domain sequences missing. The authors investigated 290,148 Pfam domains from 270,776 protein sequences in which they identified 30,961 partial domains. These partial domains were grouped into 3 types based on their sequence context, such as i) split domains – non-contiguous matches of the HMM model to a sequence resulting in a domain being ‘split’ into several parts, ii) bounded partials – domains that are bound/delimited by the end of protein sequence or non-homologous domains and iii) unbounded partials – partial domains that are not bound/delimited by sequence termini or non-homologous domains and for which a complete domain instance can be found by extending the alignment (Triant and Pearson, 2015). Due to the similar nature of the problem that was being investigated, both research groups agreed upon using the same nomenclature to classify atrophied domains. I therefore have used their nomenclature such as end-bounded and terminal-bounded atrophy in this study. Further investigation into the true nature of these partial domains, by examining sequence alignments and Pfam HMM models, indicated that these partial domains were annotation or computational artefacts caused by either alignment errors, incorrect genome assemblies or incorrect Pfam domain boundaries (Triant and Pearson, 2015), similar to the failure modes listed in table 1.1. After computational filtering and manual verification the authors identified 18

putative partial domains, which were not artefacts. These domains were built by Pfam using two smaller domains, which could sometimes be found in different sequence contexts.

New cases of domain atrophy have been brought to light after the publication of (Prakash and Bateman, 2015) further corroborating the theory of domain atrophy. A recent study by Dr. Jennifer Potts, at the University of York, and group has determined the structure of a membrane protein SasG, from *Staphylococcus aureus* (Gruszka et al., 2015). The SasG protein promotes host adherence and biofilm formation by forming extended fibrils. SasG consists of tandem repeats of two structurally related domains – E and G5, which form single layer triple-stranded β -sheets. It was observed that the smaller E domain exhibits atrophy at the N-terminus (Gruszka et al., 2015) (Figure 1.13). The N-terminal β -sheet of E domain is much shorter than that of the G5 domain due to the truncation of three β -strands. Interestingly the E domain is disordered in isolation but folds to form elongated G5-E-G5 structure.

In conclusion, I have identified a few cases of domains that exhibit partial structures of canonical folds. Traditionally domains have been viewed as indivisible, basic, building blocks of proteins, but domain atrophy sheds new light into the evolution of partial protein domains. The cases of atrophied domains identified in this study represent a significant increase in the known number of cases so far.

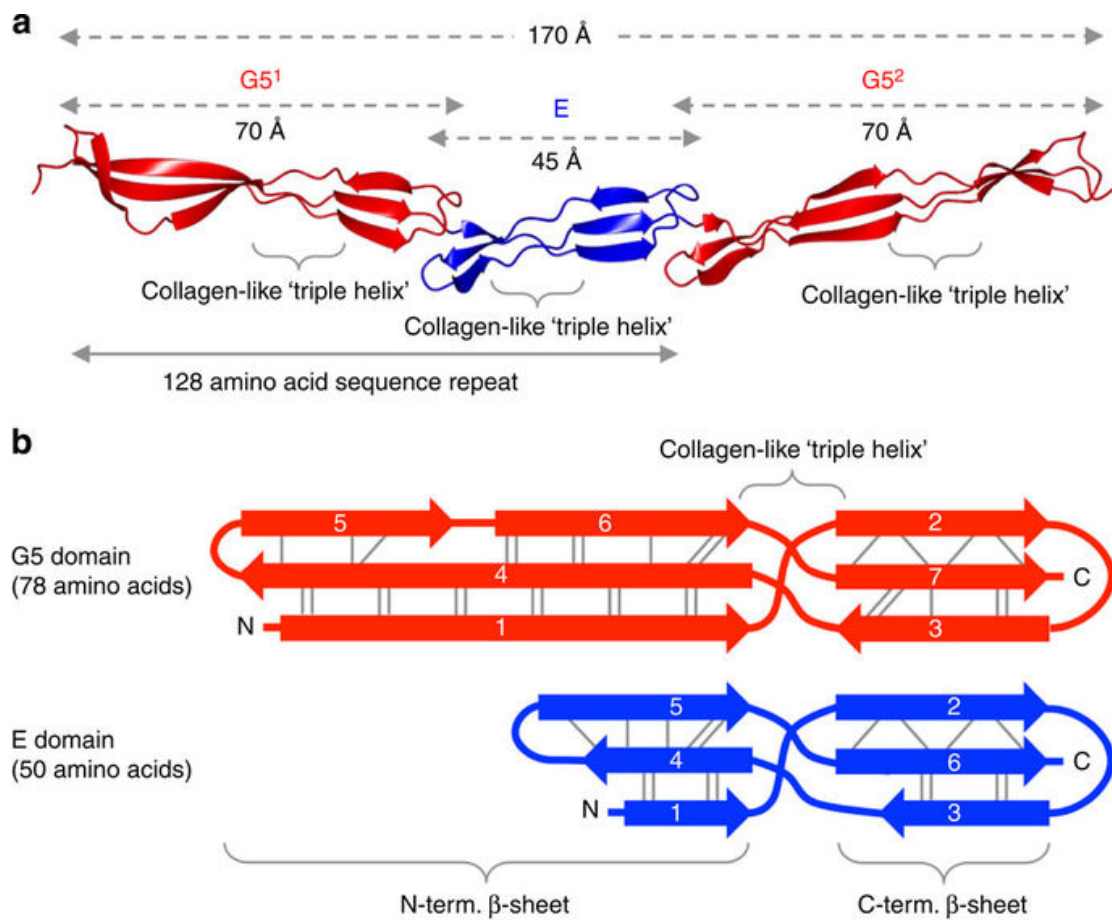


Figure 1.13 SasG system in *Staphylococcus aureus*. The E domain is shorter than the G5 domain and the N-terminal β -sheet of E domain appears truncated indicating atrophy. Figure adapted from (Gruszka et al., 2015). DOI:10.1038/ncomms8271

1.5 References

- Alfalah, M., Keiser, M., Leeb, T., Zimmer, K.P., and Naim, H.Y. (2009). Compound heterozygous mutations affect protein folding and function in patients with congenital sucrase-isomaltase deficiency. *Gastroenterology* 136, 883-892.
- Andreeva, A., Howorth, D., Chothia, C., Kulesha, E., and Murzin, A.G. (2014). SCOP2 prototype: a new approach to protein structure mining. *Nucleic acids research* 42, D310-314.
- Banci, L., Bertini, I., Ciofi-Baffoni, S., Gonnelli, L., and Su, X.C. (2003). A core mutation affecting the folding properties of a soluble domain of the ATPase protein CopA from *Bacillus subtilis*. *Journal of molecular biology* 331, 473-484.
- Beg, Q.K., Kapoor, M., Mahajan, L., and Hoondal, G.S. (2001). Microbial xylanases and their industrial applications: a review. *Applied microbiology and biotechnology* 56, 326-338.
- Biely, P., Kratky, Z., and Vrsanska, M. (1981). Substrate-binding site of endo-1,4-beta-xylanase of the yeast *Cryptococcus albidus*. *European journal of biochemistry* 119, 559-564.
- Biely, P., Kratky, Z., Vrsanska, M., and Urmanicova, D. (1980). Induction and inducers of endo-1,4-beta-xylanase in the yeast *Cryptococcus albidus*. *European journal of biochemistry* 108, 323-329.
- Birzele, F., Csaba, G., and Zimmer, R. (2008). Alternative splicing and protein structure evolution. *Nucleic acids research* 36, 550-558.
- Bos, J.L., Rehmann, H., and Wittinghofer, A. (2007). GEFs and GAPs: critical elements in the control of small G proteins. *Cell* 129, 865-877.
- Bowie, J.U., Reidhaar-Olson, J.F., Lim, W.A., and Sauer, R.T. (1990). Deciphering the message in protein sequences: tolerance to amino acid substitutions. *Science* 247, 1306-1310.
- Brugarolas, P., Duguid, E.M., Zhang, W., Poor, C.B., and He, C. (2011). Structural and biochemical characterization of N5-carboxyaminoimidazole ribonucleotide synthetase and N5-carboxyaminoimidazole ribonucleotide mutase from *Staphylococcus aureus*. *Acta crystallographica. Section D, Biological crystallography* 67, 707-715.
- Buljan, M., and Bateman, A. (2009). The evolution of protein domain families. *Biochemical Society transactions* 37, 751-755.
- Callaghan, A.J., Marcaida, M.J., Stead, J.A., McDowall, K.J., Scott, W.G., and Luisi, B.F. (2005). Structure of *Escherichia coli* RNase E catalytic domain and implications for RNA turnover. *Nature* 437, 1187-1191.

Chothia, C. (1992). Proteins. One thousand families for the molecular biologist. *Nature* 357, 543-544.

Chothia, C., Gough, J., Vogel, C., and Teichmann, S.A. (2003). Evolution of the protein repertoire. *Science* 300, 1701-1703.

Close, D., Xu, T., Smartt, A., Rogers, A., Crossley, R., Price, S., Ripp, S., and Sayler, G. (2012). The evolution of the bacterial luciferase gene cassette (lux) as a real-time bioreporter. *Sensors* 12, 732-752.

Das, D., Murzin, A.G., Rawlings, N.D., Finn, R.D., Coggill, P., Bateman, A., Godzik, A., and Aravind, L. (2014). Structure and computational analysis of a novel protein with metallopeptidase-like and circularly permuted winged-helix-turn-helix domains reveals a possible role in modified polysaccharide biosynthesis. *BMC bioinformatics* 15, 75.

Daumke, O., Weyand, M., Chakrabarti, P.P., Vetter, I.R., and Wittinghofer, A. (2004). The GTPase-activating protein Rap1GAP uses a catalytic asparagine. *Nature* 429, 197-201.

Dessailly, B.H., Redfern, O.C., Cuff, A.L., and Orengo, C.A. (2010). Detailed analysis of function divergence in a large and diverse domain superfamily: toward a refined protocol of function classification. *Structure* 18, 1522-1535.

Eddy, S.R. (2004). What is a hidden Markov model? *Nature biotechnology* 22, 1315-1316.

Finkelstein, A.V., Gutun, A.M., and Badretdinov, A. (1993). Why are the same protein folds used to perform different functions? *FEBS letters* 325, 23-28.

Finn, R.D., Bateman, A., Clements, J., Coggill, P., Eberhardt, R.Y., Eddy, S.R., Heger, A., Hetherington, K., Holm, L., Mistry, J., *et al.* (2014). Pfam: the protein families database. *Nucleic acids research* 42, D222-230.

Finn, R.D., Clements, J., and Eddy, S.R. (2011). HMMER web server: interactive sequence similarity searching. *Nucleic acids research* 39, W29-37.

Finn, R.D., Coggill, P., Eberhardt, R.Y., Eddy, S.R., Mistry, J., Mitchell, A.L., Potter, S.C., Punta, M., Qureshi, M., Sangrador-Vegas, A., *et al.* (2016). The Pfam protein families database: towards a more sustainable future. *Nucleic acids research* 44, D279-285.

Fisher, A.J., Thompson, T.B., Thoden, J.B., Baldwin, T.O., and Rayment, I. (1996). The 1.5-Å resolution crystal structure of bacterial luciferase in low salt conditions. *The Journal of biological chemistry* 271, 21956-21968.

Fowler, S.B., Best, R.B., Toca Herrera, J.L., Rutherford, T.J., Steward, A., Paci, E., Karplus, M., and Clarke, J. (2002). Mechanical unfolding of a titin Ig domain: structure of unfolding intermediate revealed by combining AFM, molecular dynamics simulations, NMR and protein engineering. *Journal of molecular biology* 322, 841-849.

Grishin, N.V. (2001). Fold change in evolution of protein structures. *Journal of structural biology* 134, 167-185.

Gruszka, D.T., Whelan, F., Farrance, O.E., Fung, H.K., Paci, E., Jeffries, C.M., Svergun, D.I., Baldock, C., Baumann, C.G., Brockwell, D.J., *et al.* (2015). Cooperative folding of intrinsically disordered domains drives assembly of a strong elongated protein. *Nature communications* 6, 7271.

Holm, L., and Sander, C. (1998). Touring protein fold space with Dali/FSSP. *Nucleic acids research* 26, 316-319.

Kanai, A., Oida, H., Matsuura, N., and Doi, H. (2003). Expression cloning and characterization of a novel gene that encodes the RNA-binding protein FAU-1 from *Pyrococcus furiosus*. *The Biochemical journal* 372, 253-261.

Katoh, K., and Standley, D.M. (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular biology and evolution* 30, 772-780.

Kim, S., Gu, S.A., Kim, Y.H., and Kim, K.J. (2014). Crystal structure and thermodynamic properties of d-lactate dehydrogenase from *Lactobacillus jensenii*. *International journal of biological macromolecules* 68, 151-157.

Kita, A., Kasai, S., Miyata, M., and Miki, K. (1996). Structure of flavoprotein FP390 from a luminescent bacterium *Photobacterium phosphoreum* refined at 2.7 Å resolution. *Acta crystallographica. Section D, Biological crystallography* 52, 77-86.

Kriventseva, E.V., Koch, I., Apweiler, R., Vingron, M., Bork, P., Gelfand, M.S., and Sunyaev, S. (2003). Increase of functional diversity by alternative splicing. *Trends in genetics : TIG* 19, 124-128.

Krogh, A., Brown, M., Mian, I.S., Sjolander, K., and Haussler, D. (1994). Hidden Markov models in computational biology. Applications to protein modeling. *Journal of molecular biology* 235, 1501-1531.

Law, A., and Boulanger, M.J. (2011). Defining a structural and kinetic rationale for paralogous copies of phenylacetate-CoA ligases from the cystic fibrosis pathogen *Burkholderia cenocepacia* J2315. *The Journal of biological chemistry* 286, 15577-15585.

Lee, S., Mahler, B., Toward, J., Jones, B., Wyatt, K., Dong, L., Wistow, G., and Wu, Z. (2010). A single destabilizing mutation (F9S) promotes concerted unfolding of an entire globular domain in gammaS-crystallin. *Journal of molecular biology* 399, 320-330.

Letunic, I., Doerks, T., and Bork, P. (2012). SMART 7: recent updates to the protein domain annotation resource. *Nucleic acids research* 40, D302-305.

Lynch, M., and Conery, J.S. (2000). The evolutionary fate and consequences of duplicate genes. *Science* 290, 1151-1155.

Marchler-Bauer, A., Derbyshire, M.K., Gonzales, N.R., Lu, S., Chitsaz, F., Geer, L.Y., Geer, R.C., He, J., Gwadz, M., Hurwitz, D.I., *et al.* (2015). CDD: NCBI's conserved domain database. *Nucleic acids research* 43, D222-226.

Martin, A.C., Orengo, C.A., Hutchinson, E.G., Jones, S., Karmirantzou, M., Laskowski, R.A., Mitchell, J.B., Taroni, C., and Thornton, J.M. (1998). Protein folds and functions. *Structure* 6, 875-884.

Martinez-Blanco, H., Reglero, A., Rodriguez-Aparicio, L.B., and Luengo, J.M. (1990). Purification and biochemical characterization of phenylacetyl-CoA ligase from *Pseudomonas putida*. A specific enzyme for the catabolism of phenylacetic acid. *The Journal of biological chemistry* 265, 7084-7090.

Miller, S., Janin, J., Lesk, A.M., and Chothia, C. (1987). Interior and surface of monomeric proteins. *Journal of molecular biology* 196, 641-656.

Moore, S.A., and James, M.N. (1995). Structural refinement of the non-fluorescent flavoprotein from *Photobacterium leiognathi* at 1.60 Å resolution. *Journal of molecular biology* 249, 195-214.

Moore, S.A., James, M.N., O'Kane, D.J., and Lee, J. (1993). Crystal structure of a flavoprotein related to the subunits of bacterial luciferase. *The EMBO journal* 12, 1767-1774.

Nagano, N., Orengo, C.A., and Thornton, J.M. (2002). One fold with many functions: the evolutionary relationships between TIM barrel families based on their sequences, structures and functions. *Journal of molecular biology* 321, 741-765.

Nardini, M., Pesce, A., Milani, M., and Bolognesi, M. (2007). Protein fold and structure in the truncated (2/2) globin family. *Gene* 398, 2-11.

Pascarella, S., and Argos, P. (1992). Analysis of insertions/deletions in protein structures. *Journal of molecular biology* 224, 461-471.

Pettersen, E.F., Goddard, T.D., Huang, C.C., Couch, G.S., Greenblatt, D.M., Meng, E.C., and Ferrin, T.E. (2004). UCSF Chimera--a visualization system for exploratory research and analysis. *Journal of computational chemistry* 25, 1605-1612.

Polizeli, M.L., Rizzatti, A.C., Monti, R., Terenzi, H.F., Jorge, J.A., and Amorim, D.S. (2005). Xylanases from fungi: properties and industrial applications. *Applied microbiology and biotechnology* 67, 577-591.

Ponting, C.P., and Russell, R.R. (2002). The natural history of protein domains. *Annual review of biophysics and biomolecular structure* 31, 45-71.

Prakash, A., and Bateman, A. (2015). Domain atrophy creates rare cases of functional partial protein domains. *Genome biology* 16, 88.

Reeves, G.A., Dallman, T.J., Redfern, O.C., Akpor, A., and Orengo, C.A. (2006). Structural diversity of domain superfamilies in the CATH database. *Journal of molecular biology* 360, 725-741.

Reger, A.S., Wu, R., Dunaway-Mariano, D., and Gulick, A.M. (2008). Structural Characterization of a 140° Domain Movement in the Two-Step Reaction Catalyzed by 4-Chlorobenzoate:CoA Ligase†‡. *Biochemistry* 47, 8016-8025.

Rose, G.D., Geselowitz, A.R., Lesser, G.J., Lee, R.H., and Zehfus, M.H. (1985). Hydrophobicity of amino acid residues in globular proteins. *Science* 229, 834-838.

Sandhya, S., Pankaj, B., Govind, M.K., Offmann, B., Srinivasan, N., and Sowdhamini, R. (2008). CUSP: an algorithm to distinguish structurally conserved and unconserved regions in protein domain alignments and its application in the study of large length variations. *BMC structural biology* 8, 28.

Sandhya, S., Rani, S.S., Pankaj, B., Govind, M.K., Offmann, B., Srinivasan, N., and Sowdhamini, R. (2009). Length variations amongst protein domain superfamilies and consequences on structure and function. *PloS one* 4, e4981.

Santos, C.R., Meza, A.N., Hoffmam, Z.B., Silva, J.C., Alvarez, T.M., Ruller, R., Giesel, G.M., Verli, H., Squina, F.M., Prade, R.A., *et al.* (2010). Thermal-induced conformational changes in the product release area drive the enzymatic activity of xylanases 10B: Crystal structure, conformational stability and functional characterization of the xylanase 10B from *Thermotoga petrophila* RKU-1. *Biochemical and biophysical research communications* 403, 214-219.

Schroder, M., and Kaufman, R.J. (2005). The mammalian unfolded protein response. *Annual review of biochemistry* 74, 739-789.

Scrima, A., Thomas, C., Deaconescu, D., and Wittinghofer, A. (2008). The Rap–RapGAP complex: GTP hydrolysis without catalytic glutamine and arginine residues. *The EMBO journal* 27, 1145-1153.

Servant, F., Bru, C., Carrere, S., Courcelle, E., Gouzy, J., Peyruc, D., and Kahn, D. (2002). ProDom: automated clustering of homologous domains. *Briefings in bioinformatics* 3, 246-251.

Sigrist, C.J., de Castro, E., Cerutti, L., Cucho, B.A., Hulo, N., Bridge, A., Bougueleret, L., and Xenarios, I. (2013). New and continuing developments at PROSITE. *Nucleic acids research* 41, D344-347.

Sillitoe, I., Lewis, T.E., Cuff, A., Das, S., Ashford, P., Dawson, N.L., Furnham, N., Laskowski, R.A., Lee, D., Lees, J.G., *et al.* (2015). CATH: comprehensive structural and functional annotations for genome sequences. *Nucleic acids research* 43, D376-381.

Sonnhammer, E.L., and Hollich, V. (2005). Scoredist: a simple and robust protein sequence distance estimator. *BMC bioinformatics* 6, 108.

Taylor, M.S., Ponting, C.P., and Copley, R.R. (2004). Occurrence and consequences of coding sequence insertions and deletions in Mammalian genomes. *Genome research* *14*, 555-566.

Teichmann, S.A., Park, J., and Chothia, C. (1998). Structural assignments to the *Mycoplasma genitalium* proteins show extensive gene duplications and domain rearrangements. *Proceedings of the National Academy of Sciences of the United States of America* *95*, 14658-14663.

Tishkov, V.I., Matorin, A.D., Rojkova, A.M., Fedorchuk, V.V., Savitsky, P.A., Dementieva, L.A., Lamzin, V.S., Mezentzev, A.V., and Popov, V.O. (1996). Site-directed mutagenesis of the formate dehydrogenase active centre: role of the His332-Gln313 pair in enzyme catalysis. *FEBS letters* *390*, 104-108.

Tress, M.L., Martelli, P.L., Frankish, A., Reeves, G.A., Wesselink, J.J., Yeats, C., Olason, P.I., Albrecht, M., Hegyi, H., Giorgetti, A., *et al.* (2007). The implications of alternative splicing in the ENCODE protein complement. *Proceedings of the National Academy of Sciences of the United States of America* *104*, 5495-5500.

Triant, D.A., and Pearson, W.R. (2015). Most partial domains in proteins are alignment and annotation artifacts. *Genome biology* *16*, 99.

Vogel, C., Bashton, M., Kerrison, N.D., Chothia, C., and Teichmann, S.A. (2004). Structure, function and evolution of multidomain proteins. *Current opinion in structural biology* *14*, 208-216.

Weiner, J., 3rd, Beaussart, F., and Bornberg-Bauer, E. (2006). Domain deletions and substitutions in the modular protein evolution. *The FEBS journal* *273*, 2037-2047.

Chapter 2

Comparative analysis of the yeast non-coding RNA interaction network

2.1 Introduction

Macromolecular interactions are key to various cellular functions. Biological macromolecules, such as proteins and nucleic acids, have complex roles achieved in large part through association with their interacting partners. Macromolecular interactions have been well studied, from investigating interactions between a pair of molecules in exquisite detail to large-scale high-throughput interactomes. Representing interactomes as physical interaction networks makes them amenable to compute various network properties, which provide insights into their nature of interactions and regulation. In this chapter I have analysed the physical properties of three large-scale biological networks – the protein-protein interaction network, the RNA-protein interaction network and the RNA-RNA interaction network, using manually curated high quality interactions from yeast. I have compared the three networks in order to investigate if their physical network properties reflect the differences observed in the nature of physical interaction between these macromolecules.

Most biological macromolecules do not function in isolation within a cell; instead, they interact with their environment, which could include small chemical moieties, lipids, proteins or nucleotides. Their interactions enable molecular synthesis, signalling, transport, assembly and degradation of by-products. Based on the nature and complexity of function, a cellular process

could involve interactions between binary pairs or a large number of partners. The interactions could either involve molecules of the same kind - such as protein-protein or RNA-RNA, or molecules of different kinds - such as RNA-protein, protein-cofactor or RNA-protein-cofactor. Given that the cell is packed with molecules of all kinds and sizes, it is challenging for the macromolecules to establish specific interactions from the non-specific background molecules to form biologically meaningful outcomes.

Macromolecules have different strategies for establishing interactions *in vivo*. Protein interactions are mainly guided by properties such as size, shape, charge, flexibility (Jones and Thornton, 1996), and also include cellular location and concentration among others. Proteins can either self-assemble to form homomeric interactions with copies of themselves or form heteromeric assemblies with distinct protein subunits (Marsh and Teichmann, 2015). Compared to heteromeric complexes, the homomeric interactions significantly bury large solvent accessible surface areas, show less planarity of interaction interfaces and have high preference for hydrophobic residues at the interface (Jones and Thornton, 1996). In some proteins, interaction interfaces are formed through cooperative folding driven disorder-to-order transition upon binding (Shammas et al., 2016). Interactions among RNA are mainly guided by sequence and the hydrogen bond donor-acceptor composition of the bases. Internucleotide interactions include base-pairing, base-stacking and base-phosphate backbone interactions and among larger nucleotide structures, such as rRNA, long-range interactions assist in helix packing (Sweeney et al., 2015). The loops, bulges and non Watson-Crick base pairs in RNA also play a role in determining identity and discriminating between specific and non-specific interactions (Giege et al., 1998; Sumner-Smith et al., 1991).

While proteins interact with each other through shape, hydrophobic and electrostatic complementarity (Keskin et al., 2008), the nucleic acids DNA and RNA largely interact through base pair complementarity. One can imagine interactions between proteins as analog and interactions between nucleotides as digital; recognition and interaction between proteins is largely independent of

their primary sequence but instead is tertiary structure and surface electrostatic charge dependent, whereas DNA or RNA interactions are largely sequence dependent. In addition to base pairing RNA also takes part in tertiary contacts, which are facilitated by long-range interactions. The long-range RNA-RNA interactions facilitate contacts between distant regions of the nucleotide through hydrogen bonding between the ribose sugars, nucleotide bases and the phosphodiester backbone (Ulyanov and James, 2010). These long-range interactions are abundantly observed in ribosomes (Nissen et al., 2001), ribozymes (Zheng et al., 2017) and riboswitches (Schroeder et al., 2011) and play an important role in stabilizing contacts between RNA-helices, promote compact helical packing and stability of tertiary and quaternary structures (Xin et al., 2008).

Unlike a strict requirement of structural complementarity among proteins to interact with each other, base pairing between RNA largely tolerates mismatches. The strength of intermolecular RNA-RNA interactions can also be easily manipulated by shortening or extending complementarity. The ability to tolerate mismatches, the diversity of nucleotide sequences and the variations in interaction lengths may allow a large number of intermolecular RNA-RNA interactions compared to that allowed in proteins (Figure 2.1). The interactions between RNA and proteins, on the other hand, is achieved through a combination of features that are observed among protein-protein and RNA-RNA interactions, which involves recognising combination of bases, nucleotide backbone features, complementary surface geometry and surface electrostatic potential (Auweter et al., 2006; Ellis et al., 2007; Jones et al., 2001). Most RNA-binding proteins have RNA-recognition and binding modules or motifs, which establish interactions with the RNA (Lunde et al., 2007). Some of the other nucleic acid-binding proteins such as the PUF and TALE proteins are composed of tandem repeated units, which bind nucleotide bases in a highly sequence specific modular way (Filipovska and Rackham, 2012).

It is easy to make or break interactions between proteins through manipulating interaction affinity by mutating key amino acid residues or “hotspots” at the

interface (Bogan and Thorn, 1998; Jubb et al., 2016). For example, in the interaction between immunity protein (Im)-DNase, each mutation of conserved hotspot residues Y54 and Y55 in the Im protein to alanine reduces binding affinity with DNase by ~ 5 kcal mol⁻¹ (Meenan et al., 2010; Wallis et al., 1998) and in the interaction between *Streptococcal* protein G and human Fc fragment of IgG, the deletion of hotspot residue E27 on the B1 domain of protein G completely abolishes interaction with the IgG (Sloan and Hellinga, 1999). However it is observed that interactions between RNA are robust to such point mutations due to their plasticity in incorporating perturbations (Kladwang et al., 2011; Rodrigo and Fares, 2012). It is likely that single nucleotide base changes or mismatches are tolerated within RNA-RNA complexes base-paired over longer sequence lengths. Therefore unlike proteins the interactions between RNA-RNA are relatively difficult to break.

Associations between macromolecules are studied using various techniques, based on the level of interaction detail sought and the scope of their interaction within the cell. For example, on a finer scale, the associations between macromolecules can be studied using X-ray crystallography, NMR or electron microscopy techniques. These techniques provide a very detailed view of the interacting partners and the interactions between them, however they do not indicate where these macromolecular associations are placed within the pathway. On the other hand associations between macromolecules can be studied on a larger scale using data from high-throughput experimental approaches such as yeast two-hybrid screens and mapping the interactions onto graphical interaction networks, which provides a broad systems-wide view of these associations and their relation with other macromolecules within the context of pathways (Alm and Arkin, 2003), but at the expense of detailed inter-atomic interactions.

Several high-throughput techniques identify interactions between proteins and/or RNA in vivo. Techniques such as co-immunoprecipitation, phage display and tandem affinity purification methods identify interactions between proteins on a large scale (Goodfellow and Bailey, 2014; Goodyear and Silverman, 2008;

van der Geer, 2014). UV crosslinking based techniques such as CLASH, CRAC, iCLIP and PAR-CLIP identify interactions between RNA-protein and RNA-RNA (Bohnsack et al., 2012; Danan et al., 2016; Helwak and Tollervey, 2014; Huppertz et al., 2014). The interactions identified through such experiments are deposited or are manually curated from published literature into public databases such as STRING, IntAct, RAID, RAIN, etc. (Junge et al., 2017; Orchard et al., 2014; Szklarczyk et al., 2015; Zhang et al., 2014). These high quality curated interactions are reliable sources to build macromolecular interaction networks.

Interaction networks are created by depicting macromolecules as nodes and the interactions between them as edges connecting these nodes. Various kinds of interaction networks have been studied such as protein-protein interaction networks (Schwikowski et al., 2000), transcriptional regulatory networks (DNA-protein) (Lee et al., 2002), signal transduction networks (Papin et al., 2005) and metabolic networks (Forster et al., 2003). Protein-protein interaction (PPI) networks have been studied in detail in various organisms, either as complete proteomes or as specific modules (proteins that belong to a certain pathway or complex). RNA-protein interaction (RPI) networks have been described in detail for few organisms (Stoiber et al., 2015) and very little information is available regarding network analysis of RNA-RNA interactions. One of the reasons RNA-RNA interaction (RRI) networks are less studied could be due to the limited availability of experimentally validated large-scale RNA-RNA interaction data.

Unlike proteins, interactions among ncRNA lack systematic manual curation, which makes it difficult to build high quality RNA interaction networks. The lack of a unique, unambiguous identifier for each ncRNA molecule presented a major challenge to consistently identify and annotate interactions among various ncRNAs. RNACentral (<http://rnacentral.org/>) now provides universal identifiers that can uniquely identify each ncRNA and integrate annotation from numerous ncRNA databases (The RNACentral Consortium, 2017). With the availability of a single identifier, the IntAct database has recently began to annotate ncRNA interactions by manually curating interactions from published literature.

Due to the inherently distinct ways in which proteins and nucleotides interact, I postulate that the physical network properties of PPI, RPI and RRI networks to be different from each other. To test the hypothesis I have built protein-protein, ncRNA-protein and ncRNA-ncRNA interaction networks of yeast from manually curated interactions by Dr. Simona Panni and Dr. Sandra Orchard, and compared their network topological properties. To the best of my knowledge, this study also presents the first described analysis of ncRNA interaction network in yeast.

The work described here is in collaboration with Dr. Simona Panni and Dr. Sandra Orchard from the IntAct consortium. Dr. Panni and Dr. Orchard carried out the literature curation and populating the IntAct database. I have carried out the analysis of the interaction networks. The main focus of the study presented in this chapter is to investigate similarities or dissimilarities between PPI and RRI networks using yeast as the model dataset, since the data curated for the RRI in yeast is of good quality than compared to other organisms. The human interaction networks are only used for comparative purposes.

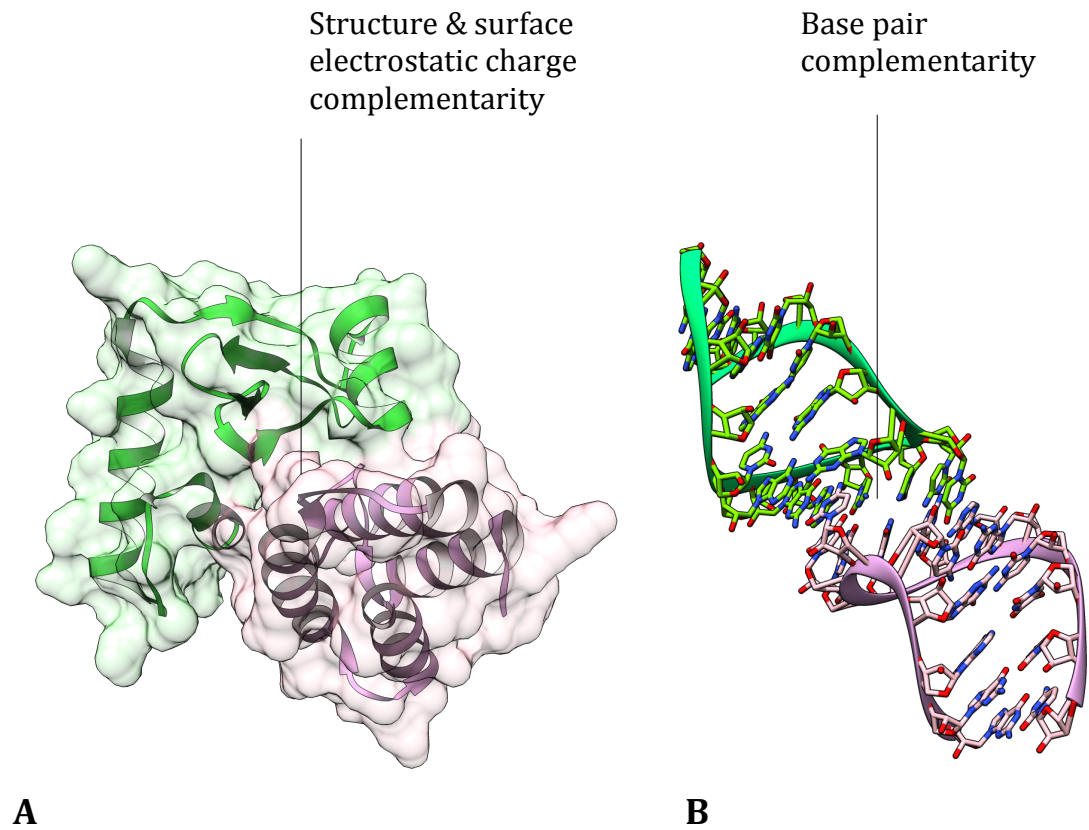


Figure 2.1 Schematic representations of protein-protein interactions and RNA-RNA interactions. (A) Proteins interact through shape, hydrophobic and electrostatic complementarity, while (B) RNA mainly interacts through base pair complementarity. Proteins and RNA are coloured by their subunits. (A) PDB: 1TFK, *E. coli* C-terminal domain of Colicin-D toxin (green) complexed with its inhibitor Colicin-D immunity protein (pink). (B) PDB: 2JLT, RNA kissing complex of HIV-1 trans-activating responsive (TAR) RNA stem (green) with a high-affinity RNA aptamer (pink).

2.2 Methods

2.2.1 Data collection and curation

The data collection and manual curation work described here was carried out by Dr. Simona Panni and Dr. Sandra Orchard of the IntAct consortium at EMBL-EBI. *Saccharomyces cerevisiae* was selected for curating ncRNA interactions, since it contains a limited number of ncRNAs, most of which are well characterised in literature and annotated in databases. RNAcentral identifiers were assigned to each ncRNA by searching the database with Saccharomyces Genome Database (SGD) identifiers. For tRNAs the database was searched with GTRNAdb identifier. RNA sequence search was used for a few tRNA precursors. The precursors and mature RNAs were considered different interactors.

Non-coding RNAs were manually curated from literature following curation standards established by the IMEx Consortium. ncRNA databases such as Rfam (Nawrocki et al., 2015), SGD (Cherry et al., 2012), LncRNAdb (Quek et al., 2015) and yeast snoRNA databases (Piekna-Przybylska et al., 2007) were queried in addition to published literature to draw up a complete list of *S. cerevisiae* ncRNAs. Relevant articles were queried from PubMed abstracts containing at least one ncRNA name and "yeast" or "*S. cerevisiae*" terms. Other keywords that were queried in PubMed include "CLIP", "CLIP-seq", "CLASH", "rna rna interaction" "rna protein interaction". The resulting several hundred articles from PubMed was manually filtered down to a total of 120 articles, which were then manually curated. RNA-RNA and RNA-protein interactions were manually curated using the IntAct editor, according to the IMEx standards.

2.2.2 Network analysis

The yeast protein-protein interactions were downloaded from IntAct (<http://www.ebi.ac.uk/intact/search>) (November 2016) using the query 'ptypeA:protein AND ptypeB:protein'. Similarly interactions between ncRNA and protein were queried with using term '((ptypeA:RNA AND ptypeB:protein) OR

(ptypeA:protein AND ptypeB:RNA)'. Interactions between ncRNAs was searched using the term 'ptypeA:RNA AND ptypeB:RNA'. The searches were limited to yeast (taxonomy ID: 559292) by using the term "*Saccharomyces cerevisiae*" in the advanced search option 'Organism'.

The physical network properties of interaction networks were computed using igraph package (<http://igraph.org>) in R. Only a single edge was kept for instances of duplicate edges and self-interaction edges (loops) were removed before computing network properties. To compare network properties between the RRI network and the PPI network, the PPI network was down-sampled such that the down-sampled network consisted of the same number of edges as in RRI network and mean values from one hundred down-sampled networks were considered. A random network was generated, to compare network properties with biological networks, using the Erdős-Renyi model with the same number of nodes and edges as in PPI network (nodes: 6,091, edges: 77,620).

The human interaction data was used for comparison. Protein-protein interactions in human were downloaded from IntAct using the query 'ptypeA:protein AND ptypeB:protein'. Similarly interactions between RNA and protein were queried with using term '((ptypeA:RNA AND ptypeB:protein) OR (ptypeA:protein AND ptypeB:RNA))'. The searches were limited to human (taxonomy ID: 9606) by using the term "*Homo sapiens*" in the advanced search option 'Organism'. Interactions between RNAs were downloaded from RAID v2.0 database (<http://www.rna-society.org/raid/>) (Yi et al., 2017). The random interaction network comprises same number of nodes and edges as in PPI network (nodes: 17,522, edges: 110,917). Networks were visualised in Cytoscape (Shannon et al., 2003).

The descriptions of physical network properties described below (sections 2.2.3 to 2.2.7) are taken verbatim from the manuscript (Panni et al., 2017).

2.2.3 Degree distribution

The degree of a node n is the number of edges linked to it. The number of nodes ordered by their increasing degree gives the degree distribution of a network.

2.2.4 Clustering coefficient (Transitivity)

The clustering coefficient of a node is n defined as $C(n) = 2e/(k(k-1))$, where e is the number of edges between neighbours of node n and k is the number of neighbours of n . The value of clustering coefficient lies between values 0 and 1 and is highest if all the neighbours of the node directly interact with each other (i.e., edges form triangles) and 0 when none of the neighbours are connected with each other.

2.2.5 Betweenness centrality

Betweenness centrality of a node n is defined as $B(n) = \sum_{a \neq n \neq b} (\sigma_{ab}(n)/\sigma_{ab})$, where σ_{ab} is the shortest number of paths between nodes a , b and $\sigma_{ab}(n)$ is the shortest number of paths between nodes a , b through node n . Betweenness centrality is divided by the normalising factor $(N-1)(N-2)/2$, where N is the total number of nodes in the connected network.

2.2.6 Closeness centrality

Closeness centrality of a node n is defined as the reciprocal of average shortest path length, $K(n) = 1/\text{average}(L(n,a))$, where $L(n,a)$ is the shortest path between two nodes n and a . The closeness centrality measure was computed on sub-graphs with the highest number of interconnected nodes.

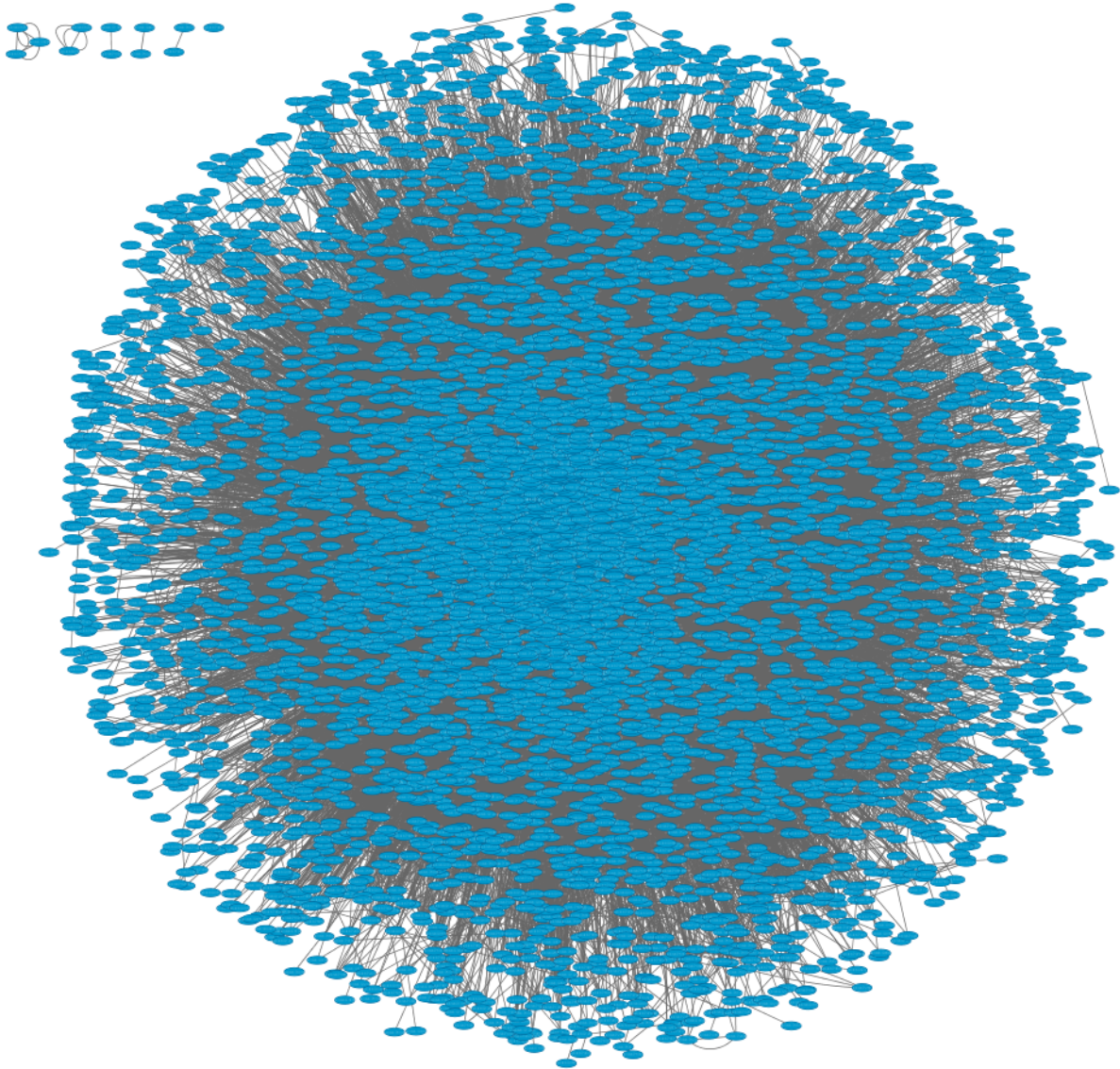
2.2.7 Neighbourhood connectivity

Neighbourhood connectivity of a node n is defined as the average connectivity (degree) of all its neighbour nodes.

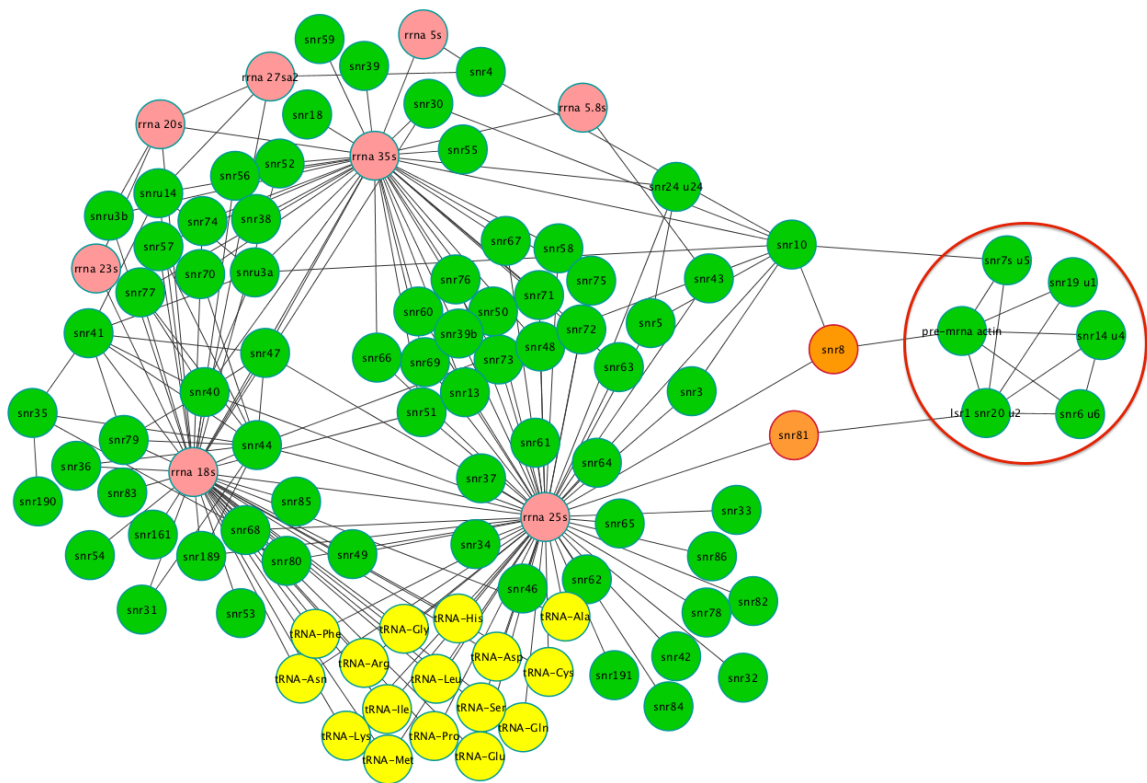
2.3 Results

The undirected and unweighted interaction networks of proteins and RNAs from yeast and human were built after removing duplicate edges and loops. The complete yeast proteome PPI network comprises 77,620 interactions (edges) among 6,091 proteins (nodes), the RPI network was inferred from 596 interactions between 105 ncRNAs and 153 proteins, and the RRI network from 195 interactions among 102 ncRNAs. The human PPI network consists of 110,917 interactions between 17,522 proteins, the RPI network has 111 interactions between 34 ncRNAs and 89 proteins and the RRI network comprises 2,412 interactions between 2,011 ncRNAs. Figure 2.2 shows the yeast PPI, RPI and RRI networks.

Representation of macromolecular interaction systems as networks enables analyses of their topological properties, such as connectivity between nodes, shortest paths between nodes, centrality of nodes, properties of edges, global network properties among others (Ma'ayan, 2011). I have compared the topological properties within biological networks as well as with the properties of random generated networks to investigate the principles of macromolecular interactions on a global scale.



A



C

Figure 2.2 Network representations of undirected *S. cerevisiae* macromolecular interactions. (A) Protein-protein interaction network (B) RNA-protein interaction network; nodes representing proteins are coloured blue and RNA in green and (C) RNA-RNA interaction (RRI) network; nodes representing ribosomal RNAs are coloured pink, tRNAs are coloured yellow, snRNAs are coloured green and the nodes snr8 and snr81 with high betweenness but low centrality (HBLC) scores are coloured orange. The spliceosomal module comprising spliceosomal RNAs is circled in red.

2.3.1 Degree distribution

Degree distribution of a network describes the connectivity of nodes within a network. Degree distribution measures the probability of nodes within a network to interact with k other nodes. The degree of a node refers to the number of interactions with other nodes, which in biological sense refers to the number of interaction partners of a molecule. As seen from figure 2.3A the degree distribution of PPI network follows the power law, wherein a large number of nodes interact with few partners (low degree) and a small number of nodes, called hubs, interact with a large number of partners (large degree). The power-law distribution denotes that the probability of an event P is an inverse power of its value k , i.e., $P(k) \sim k^{-\gamma}$, where γ is a constant (Arita, 2005). In most real-world networks, including biological networks, the power law exponent (γ) ranges between $2 < \gamma < 3$ (Barabasi and Oltvai, 2004; Chung and Lu, 2002). Networks with a power-law degree distribution exhibit scale-free character i.e., the topology of the network structure does not vary with the scale of the network, independent of whether the network is viewed locally or globally (Arita, 2005). Biological networks such as PPI network, metabolic networks, have been shown to exhibit scale-freeness (Barabasi and Oltvai, 2004; Nacher et al., 2009; Rajarathinam and Lin, 2006). The power-law degree distribution and the scale-freeness of biological networks are alluded to the property of 'preferential attachment' of nodes, wherein during network growth a new node preferentially associates with a well connected node (hub) rather than associating with a node that has fewer links (Barabasi and Oltvai, 2004). For example in the *E. coli* metabolic network, novel enzymes, which are evolved through gene duplication, maintain some compounds involved in the original reaction catalysed by the ancestral enzyme suggesting that the newly formed node links with the already connected metabolite (Light et al., 2005).

Hub proteins are critical for the functioning of a network and their removal can result in failure of the system (Jeong et al., 2001). The yeast PPI network has an average 12.74 interactions per node (median = 11). The PPI network consists of 651 hubs that interact with more than 50 proteins. The major hubs include heat

shock proteins SSB1 (UniProt: P11484), SSA1 (UniProt: P10591) and SSA2 (UniProt: P10592) with 3493, 2751 and 2444 interactions respectively. In comparison both the yeast RPI and RRI networks are sparsely connected with an average 2.13 (median = 2) and 1.90 (median = 2) interactions per node respectively. The GAR1 protein, a subunit of H/ACA ribonucleoprotein complex, (UniProt: P28007) and the small nucleolar RNA U3a (snru3a) (IntAct: EBI-10821792, RNACentral: URS0000444F9B) form the major protein and RNA hubs in the RPI network with 32 and 51 edges respectively (Figure 2.2B). On the other hand, the 18S and 25S ribosomal RNAs (rRNAs) dominate the yeast RRI network (Figure 2.2C).

Since the scale of PPI and RRI networks is not similar, comparing network properties between them could potentially introduce sampling bias. To eliminate this bias and to compare network properties between equally sized networks I randomly down-sampled the PPI network to build a down-sampled network consisting 195 edges (equivalent to number of edges in RRI network). 100 rounds of this random down-sampling were performed and the network properties for each of these 100 down-sampled networks were analysed. Figure 2.4 shows the distributions of power law exponent (γ) for some of the network properties. The mean value of γ for degree distribution of down-sampled PPI networks is 3.50, which is relatively higher than the degree distribution power law exponent γ of RRI (2.64), suggesting that the degree distribution could be under estimated in the RRI networks. Table 2.1 lists the mean values of power-law exponents for distributions of various network properties.

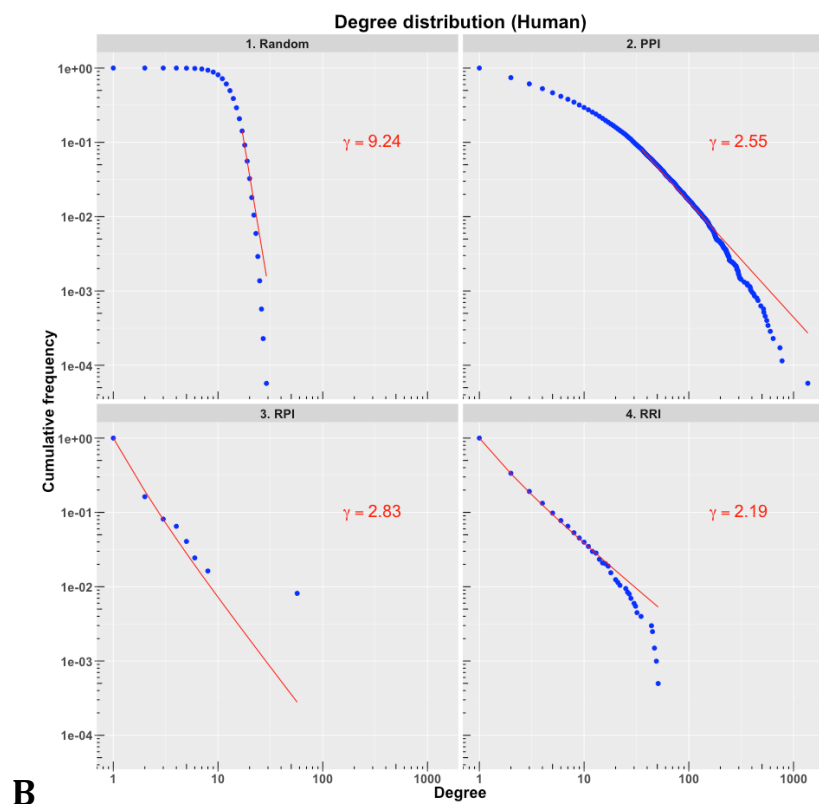
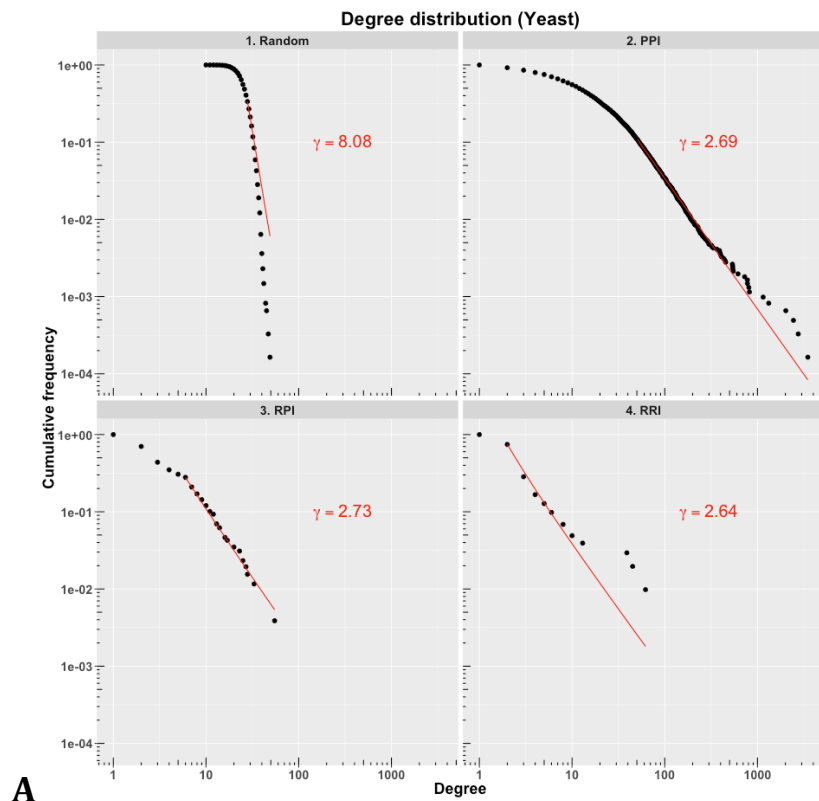


Figure 2.3 Degree distributions of (A) yeast and (B) human interaction networks. PPI: Protein-protein interaction network, RPI: RNA-protein interaction network, RRI: RNA-RNA interaction network.

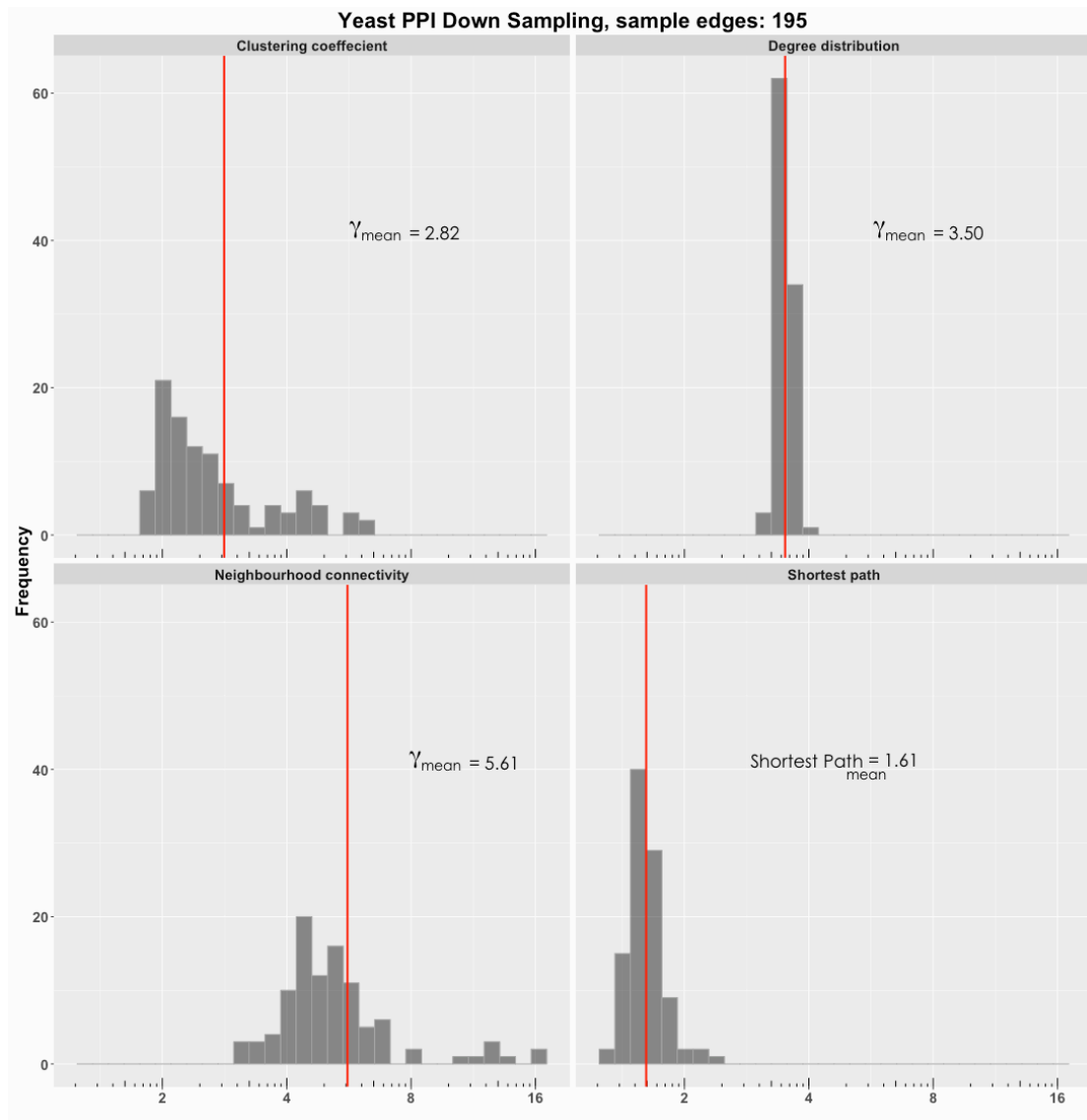


Figure 2.4 Distributions of power law exponent (γ) for clustering coefficient, degree distribution and neighbourhood connectivity of 100 random down-sampled yeast PPI networks. The distribution of average shortest paths is also shown. The red line intercept on the x-axis denotes the mean value of the distributions.

Comparison of yeast networks with human networks show similar trends of degree distributions (Figure 2.3B). The human PPI network consists of 837 hubs that interact with more than 50 proteins. Some of the largest hub nodes are transcription factor AP-1 or JUN (UniProt: P05412), Myc proto-oncogene (UniProt: P01106) and growth factor receptor-bound protein 2 GRB2 (UniProt: P62993), which interact with 1369, 777 and 742 proteins respectively. The human PPI network is also very dense with an average 6.33 interacting partners per protein. In the RPI network the let-7a miRNA precursor (IntAct: EBI-2462028) is the largest hub with 57 protein interactions, while the telomerase reverse transcriptase TERT (UniProt: O14746) is the largest protein interacting with 8 ncRNAs. The main hub nodes of human RRI network include miR-155, miR-21 and miR-145 with 51, 49 and 47 interactions respectively.

2.3.2 Clustering coefficient (Transitivity)

Clustering coefficient, or transitivity, measures the likelihood of nodes in a network to form clusters or sub-networks (modules). Nodes tend to have a high likelihood of clustering if their neighbours are directly connected to each other (Watts and Strogatz, 1998). In comparison to random networks most real world networks, including PPI networks, display a high level of clustering, which signifies grouping of functionally related nodes into modules (Barabasi and Oltvai, 2004). Modules are composed of a small fraction of cell components, each with discrete functions that form interactions to carry out a biological process (Hartwell et al., 1999). This modular organisation observed in cell biology could be due to separation through spatial localisation or chemical specificity (Hartwell et al., 1999). For example the DNA replication or the ribosome module are spatially localised and comprises components with distinct but related functions that are involved in synthesizing a biopolymer. Such kind of modular architecture is absent from random networks. Apart from modularity, biological networks also exhibit hierarchical organisation (Ravasz, 2009). The modules within the network are not isolated but instead connect to form larger but less cohesive groups, which in turn connect to other modules, in a hierarchical fashion, to form even larger and less connected clusters (Ravasz, 2009). Modules

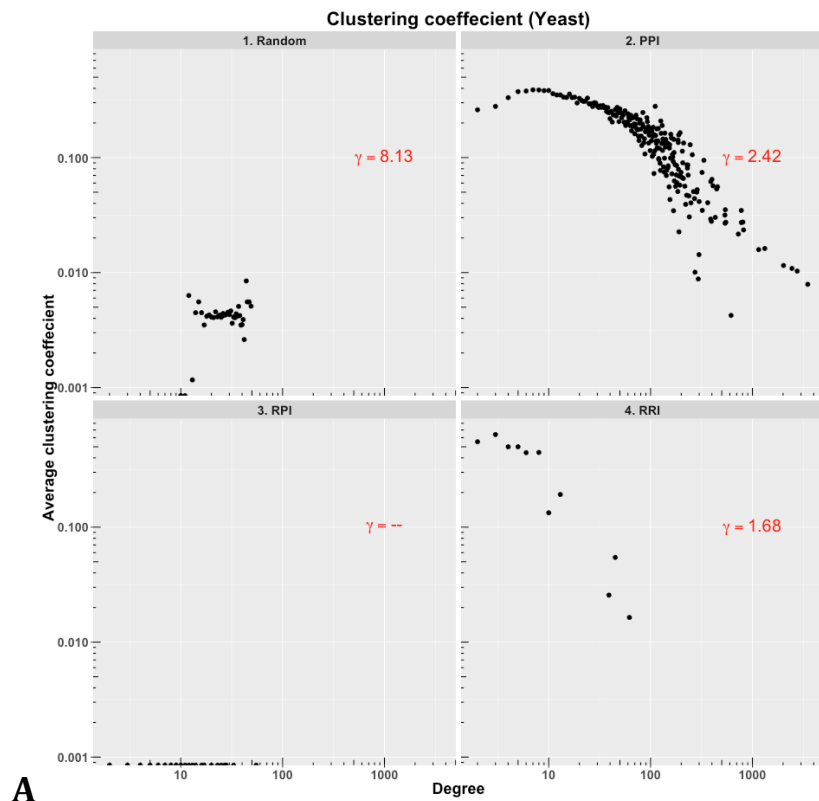
that share metabolites within metabolic networks show nested hierarchical topology (Ravasz, 2009). Self-organisation or nesting of modules into each other is observed among networks with high clustering coefficient (Galeota et al., 2015). The clustering coefficient of a node lies between values 0 and 1 and is highest if all the neighbours of the node directly interact with each other (i.e., edges form triangles) and 0 when none of the neighbours are connected with each other. For many real-world networks the clustering coefficient value typically ranges from 0.1 to 0.5 (Girvan and Newman, 2002).

In the yeast PPI network the average clustering coefficient of nodes decrease with the increase in node degree and follows a power-law scaling behaviour (Figure 2.5A). Higher clustering coefficient of nodes that have fewer interacting partners (low degree) indicates that interactions within smaller modules are dense with all interacting partners communicating with each other. For example the oligo(A)/oligo(T)-binding protein DAT1 (UniProt: P13483) has a clustering coefficient of 0.93. It interacts with six different chaperones including prefoldin subunit 1 (PFD1) (UniProt: P46988) and heat shock proteins SSA1, SSA2, SSB1, SSB2 and SSE1. The interacting partners of DAT1 represent a functional unit, which are related by similar molecular functions and biological process. As a result this highly connected small sub-network has a high clustering coefficient. On the other hand a large hub with its many interacting partners has a very low clustering coefficient, since all interacting partners do not show mutual interactions between each other. For example, a large hub such as the ATP-dependent molecular chaperone HSP82 (HSP90) (UniProt: P02829) interacts with 1,152 proteins, which is 18.9% of the yeast proteome, but only has clustering coefficient of 0.01. The chaperone HSP82 is abundantly synthesized in eukaryotic cells and is essential for protein homeostasis and promotes structural maintenance of target proteins (Borkovich et al., 1989; Zhao and Houry, 2007). However proteins that interact with HSP82 perform diverse functions such as transcription regulation, lipid metabolism, signal transduction among others (Rizzolo et al., 2014) and proteins from these different functional units do not necessarily interact with each other.

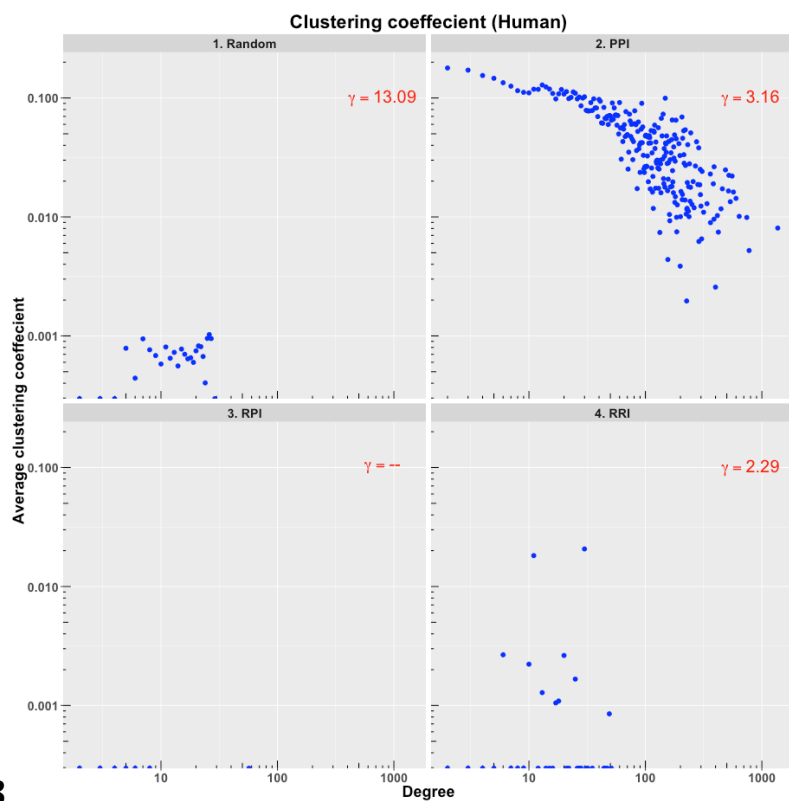
Similar to the PPI network the yeast RRI network shows a decrease in clustering coefficient values with increasing node degree and the distribution follows power-law (Figure 2.5A). The yeast small nucleolar RNA snR47 (IntAct: EBI-10921939) interacts with 6 nodes and has a clustering coefficient of 0.66. It interacts with the ribosomal RNA subunits 18S, 25S, 35S rRNA and snRNAs snR40, snR41 and snR44. SnRNA47, snR40, snR41 and snR44 guide 2'-O-methylation of large and small rRNA subunits (Cherry et al., 2012). The common targets of these snoRNAs are ribosomal subunits and they also interact with each other forming a small sub-network with mutual interactions, which indicates that discrete functional small sub-units have high scores of clustering coefficient. The large hub nodes the rRNA subunits 18S, 25S and 35S rRNA each have 45, 62 and 39 edges (degree) and clustering coefficient of 0.05, 0.01 and 0.02 respectively. As observed in PPI networks, these large rRNA hubs interact with other RNAs that have diverse functions, for example snoRNAs and tRNAs (Figure 2.2C), which do not share mutual interactions with each other and therefore have low potential for clustering. Comparison of clustering coefficient distributions between RRI and the down-sampled PPI network show similar values of (γ) (Table 2.1)

The clustering coefficient property of the RPI network shows a completely different behaviour. The RPI network forms a bipartite network; the nodes belong to two disjoint sets - proteins and RNAs - with no edges between two nodes of the same set (i.e., no triangles) and therefore have zero clustering coefficient. Random networks, as expected, show poor clustering coefficient indicating the absence of modular sub-networks.

The human PPI, RPI and RRI networks have similar distributions of clustering coefficient when compared to the yeast PPI, RPI and RRI networks respectively (Figure 2.5B).



A



B

Figure 2.5 Distributions of average clustering coefficients of nodes in (A) yeast and (B) human interaction networks.

2.3.3 Betweenness centrality

Betweenness centrality is one of the node centrality measures, the other being closeness centrality, that evaluates the crucial role of a node as a mediator of interactions within the network. A node is considered to have high betweenness in the network if it lies on the shortest paths between all the other nodes (Pavlopoulos et al., 2011). The betweenness centrality value of a node ranges from 0 to 1. Nodes with high betweenness centrality value have a large influence on the directed networks such as signal transduction networks or metabolic pathways; since the shortest paths between all the other nodes pass through them, they act as bottlenecks through which transfer of signals between nodes can be regulated (Yu et al., 2007). Betweenness centrality is also an indicator of how crucial the nodes are for the functioning of a network. Studies in the eukaryotic protein interaction networks in yeast, worm and fly have shown that proteins with high betweenness centrality values are essential for organism survival and their rate of evolution is much slower compared to other proteins (Hahn and Kern, 2005).

In the yeast PPI network, nodes with high degree also exhibit high betweenness (Figure 2.6A). The heat shock protein SSA1 (UniProt: P10591) is a hub in the PPI network with 2751 edges and a betweenness centrality value of 0.15. The enzyme chorismate mutase ARO7 (UniProt: P32178) has a degree of 6, but with a very small betweenness centrality value of $4.72e-08$. Similarly hubs in the RRI network exhibit high betweenness. The ribosomal RNAs rRNA 25S, 18S and 35S have betweenness centrality values of 0.55, 0.30 and 0.22 respectively. Hubs cover a large number of paths that connect nodes within a network and therefore tend to show high betweenness, but on the other hand nodes that interact with fewer number of proteins do not have many edges or shortest paths passing through them, they show low betweenness centrality values and therefore are not central within the network.

Interestingly the yeast PPI and RRI networks contain nodes that have low connectivity (degree) but relatively high betweenness centralities. These include

proteins such as meiotic nuclear division protein 1 (UniProt: P53102) (degree: 2, betweenness centrality: 0.32e-03), mitochondrial inner membrane protease subunit 1 (UniProt: P28627) (degree: 2, betweenness centrality: 0.32e-03) and mitochondrial rhomboid protein 1 (UniProt: P53259) (degree: 2, betweenness centrality: 0.32e-03) in the PPI network and snoRNAs SNR8 (IntAct: EBI-10921271) (degree: 3, betweenness centrality: 0.04) and SNR81 (IntAct: EBI-10918031) (degree: 2, betweenness centrality: 0.03) in the RRI network. Such nodes with 'high betweenness but low connectivity' (HBLC) were identified in a previous study of the yeast proteome (Joy et al., 2005). It was shown that nodes in PPI network with HBLC tend to be important connectors that link various modules (or clusters) within the network and are essential proteins of recent evolutionary origin (Joy et al., 2005). For example in the RRI network snoRNAs SNR8 and SNR81 connect the main ribosomal module with the spliceosomal module forming bridging interactions (Figure 2.2C).

It has been proposed that HBLC nodes in PPI networks evolve by the addition of nodes with edges and random rewiring of these edges, as a result of gene duplication and point mutations (Joy et al., 2005). Although node addition (gene duplications) and random rewiring (mutations) may answer the presence of HBLC nodes in PPI networks, the same evolutionary model cannot be extended to RPI and RRI networks. Protein-coding genes and non-coding genes evolve by different mechanisms and under different evolutionary constraints; while duplication and divergence is suggested as the major mechanism for expanding protein-coding gene repertoire (Dujon, 2010; Guan et al., 2007), mechanisms such as retroposition (Schmitz et al., 2008; Weber, 2006), intragenic duplications (Shao et al., 2009) and de novo emergence (Meunier et al., 2013) are suggested to drive the expansion of short ncRNA genes. Moreover mutation rates of protein-coding genes and non-coding RNAs, which can potentially rewire interactions in the network, are different thereby affecting edge dynamics of the networks.

The human interaction networks show similar distributions of betweenness centrality values, wherein nodes with high degree have high betweenness centrality scores (Figure 2.6B).

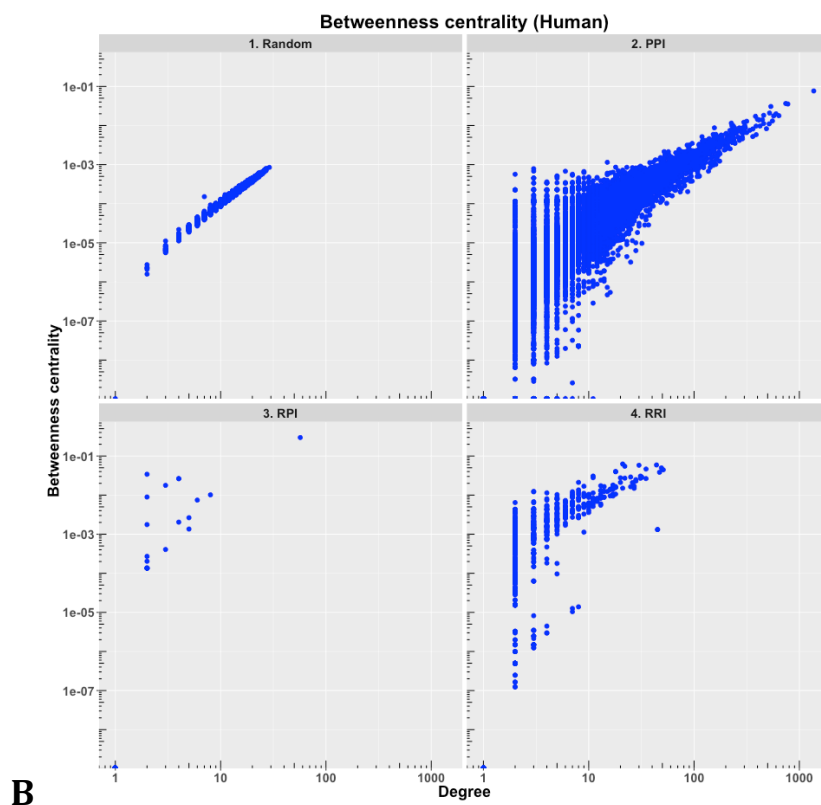
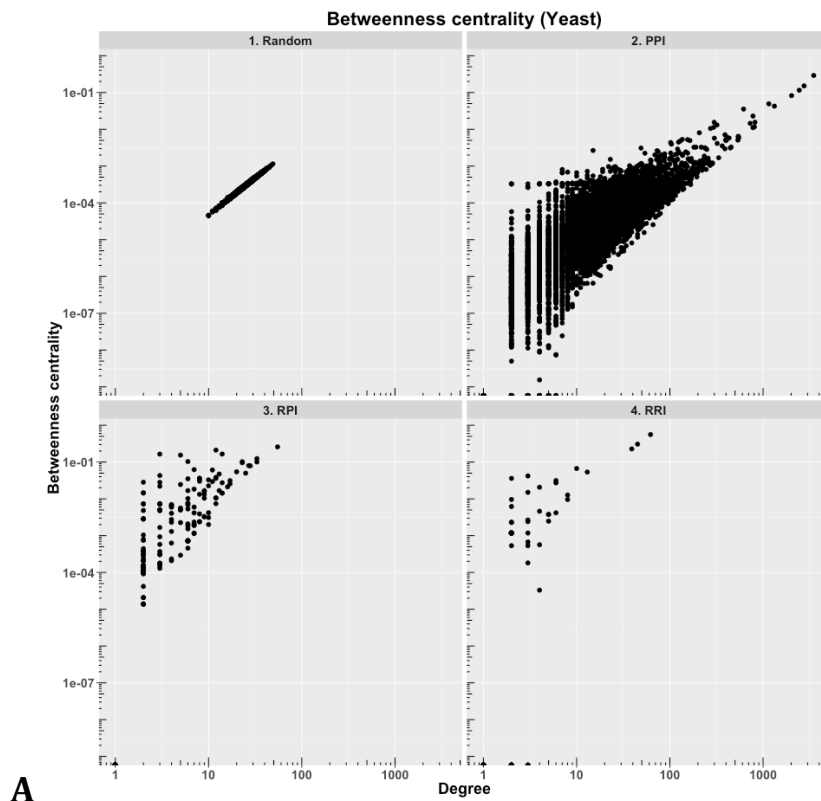


Figure 2.6 Distributions of betweenness centrality values of nodes in (A) yeast and (B) human interaction networks.

2.3.4 Closeness centrality

Closeness centrality is another measure to infer node centrality in a network. Closeness measures the average number of nodes connecting a node with all other nodes (Hahn and Kern, 2005). It is computed by calculating the inverse of the sum of the shortest distances between a node and every other node in the network (Koschutzki and Schreiber, 2008). A node with high value of closeness centrality indicates that it is near to all other nodes in a network, which suggest that these important nodes can communicate quickly with other nodes within the network (Pavlopoulos et al., 2011). For example, in the host-pathogen PPI networks, nodes that exhibit high betweenness and closeness centrality measures are considered potential drug targets since they represent key nodes that are crucial for network navigability; targeted attack of these nodes makes the system vulnerable (Mulder et al., 2014).

Since closeness centrality measures the shortest distance between nodes, disjointed networks cannot be considered for computation, as there are no links between them. In PPI and RPI networks, the closeness centrality measure was only computed on sub-graphs with the largest number of interconnected nodes. The normalised values of closeness centrality ranges between 0 and 1, wherein nodes have a score 0 if the node is isolated and a score of 1 if the node is directly connected to all other nodes. Compared to random interaction networks, in the yeast interaction networks I observe that the closeness centrality measure is significantly higher among nodes that have many interactions (Figure 2.7A). In the PPI network, the hubs have the highest closeness centrality values; for example the ribosome associated molecular chaperone SSB1 (UniProt: P11484) has the closeness centrality score of 0.68, while the mitochondrial protein SOM1 (UniProt: Q05676) has only one interactions and the lowest closeness centrality score of 0.20 in the network.

The yeast RPI and RRI networks too follow similar distributions, however the closeness centrality values for hub nodes in the RRI network is much higher than that of RPI network. In the RPI network, snoRNA snru3a (IntAct: EBI-10821792)

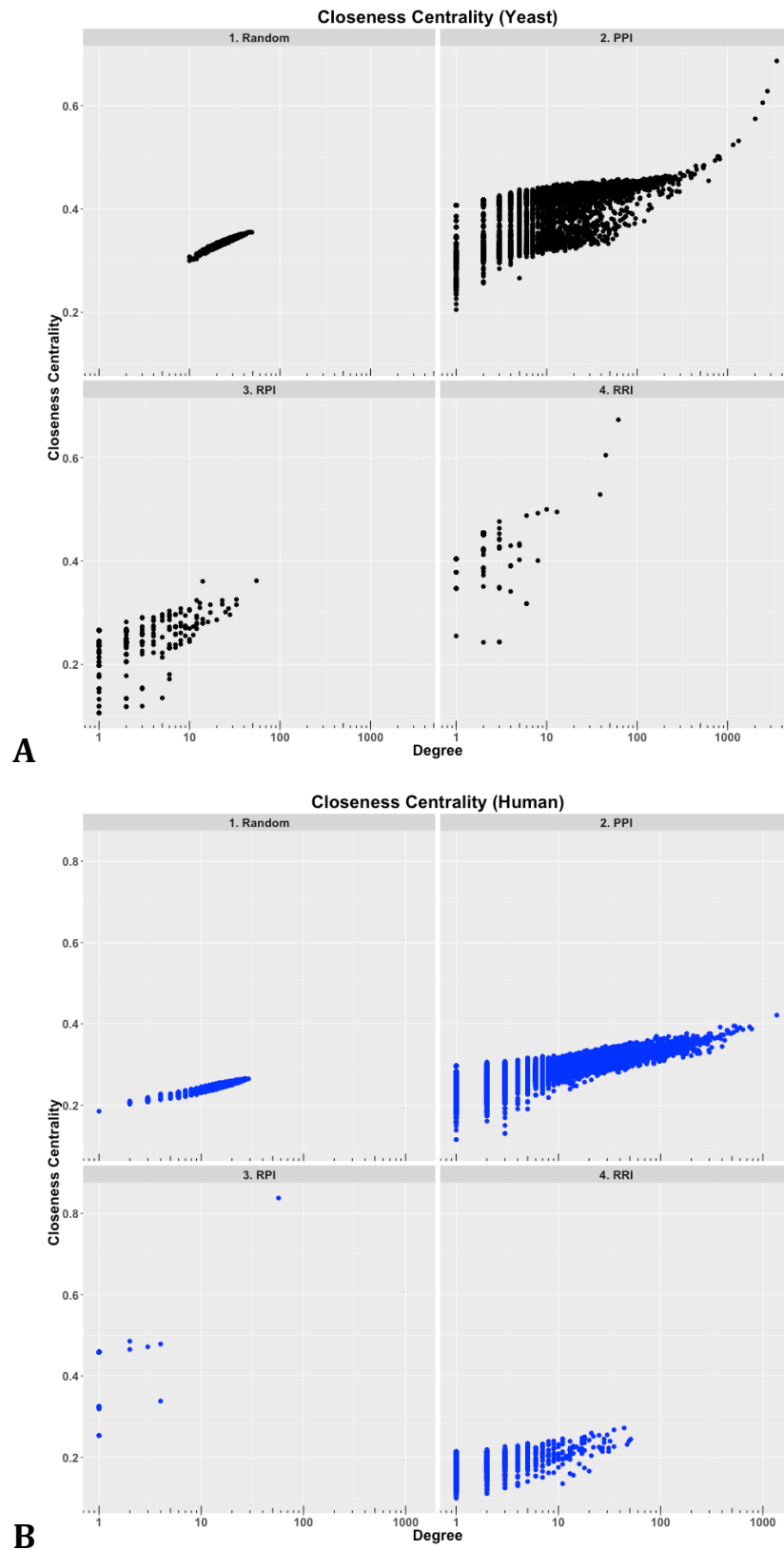


Figure 2.7 Distributions of closeness centrality values of nodes in (A) yeast and (B) human interaction networks.

has the closeness centrality score of 0.36 in the network followed by the pre-mRNA splicing factor RNA helicase PRP43 (UniProt: P53131) with a score of 0.36. These nodes represent the ncRNA and protein hubs respectively and are directly connected to most of the nodes with many edges passing through them and therefore have high closeness centrality scores. In the RRI network the highest closeness centrality scores belong to ribosomal RNA hubs, rRNA 25S, 18S and 35S with a score of 0.67, 0.60 and 0.52 respectively. On the other hand the less connected nodes, snoRNAs of the spliceosomal module, snr6_u6 (IntAct: EBI-10824938), snr14_u4 (IntAct: EBI-10054797) and snr19_u1 (IntAct: EBI-10054789) have a closeness centrality score of 0.24 each.

By comparison the distributions of closeness centrality values with respect to the node degrees among human interaction networks are similar to the yeast interaction networks (Figure 2.7B).

2.3.5 Neighbourhood connectivity

The number of neighbours (or degree) of a node is its connectivity and the neighbourhood connectivity is a measure of the average connectivity of all neighbours of a node (Maslov and Sneppen, 2002). Neighbourhood connectivity describes the likelihood of nodes with different degrees to connect to each other (Maslov and Sneppen, 2002). Figure 2.8A shows the distributions of neighbourhood connectivity scores in yeast interaction networks. I observe that the neighbourhood connectivity shows a decreasing trend with an increase in node connectivity in PPI, RPI and RRI networks and the distribution of the scores follow power-law. Hub nodes have low scores of neighbourhood connectivity compared to the nodes with fewer connections. This asymmetric nature of connectivity has been observed in protein-interaction networks (Maslov and Sneppen, 2002). To further understand the nature of neighbourhood connectivity I computed the assortativity coefficient of these networks.

Assortativity coefficient denotes the preference for a network's nodes to other similar nodes. The value of coefficient ranges between -1 and +1; the coefficient

value closer to +1 suggests that the network is assortative, i.e., the nodes within the network tend to connect to other nodes with similar degree values; while the coefficient value closer to -1 suggests that the network is disassortative and the high degree nodes connect to low degree nodes (Sharma et al., 2013). The assortativity coefficients of PPI, RPI and RRI networks are -0.14, -0.31, -0.60 respectively, which indicates that the highly connected nodes of PPI, RPI and RRI networks, on average, tend to associate with sparsely connected nodes (low degree nodes) and therefore the networks are disassortative. By linking highly connected nodes with sparsely connected nodes in the disassortative network, the likelihood of cross talk between different functional modules within the cell is decreased (Maslov and Sneppen, 2002).

The assortativity coefficient of the random network is 0.002 which, compared to the biological networks, suggests that there is no preferential attachment of nodes. The human interaction networks display similar distributions of clustering coefficient values (Figure 2.8B).

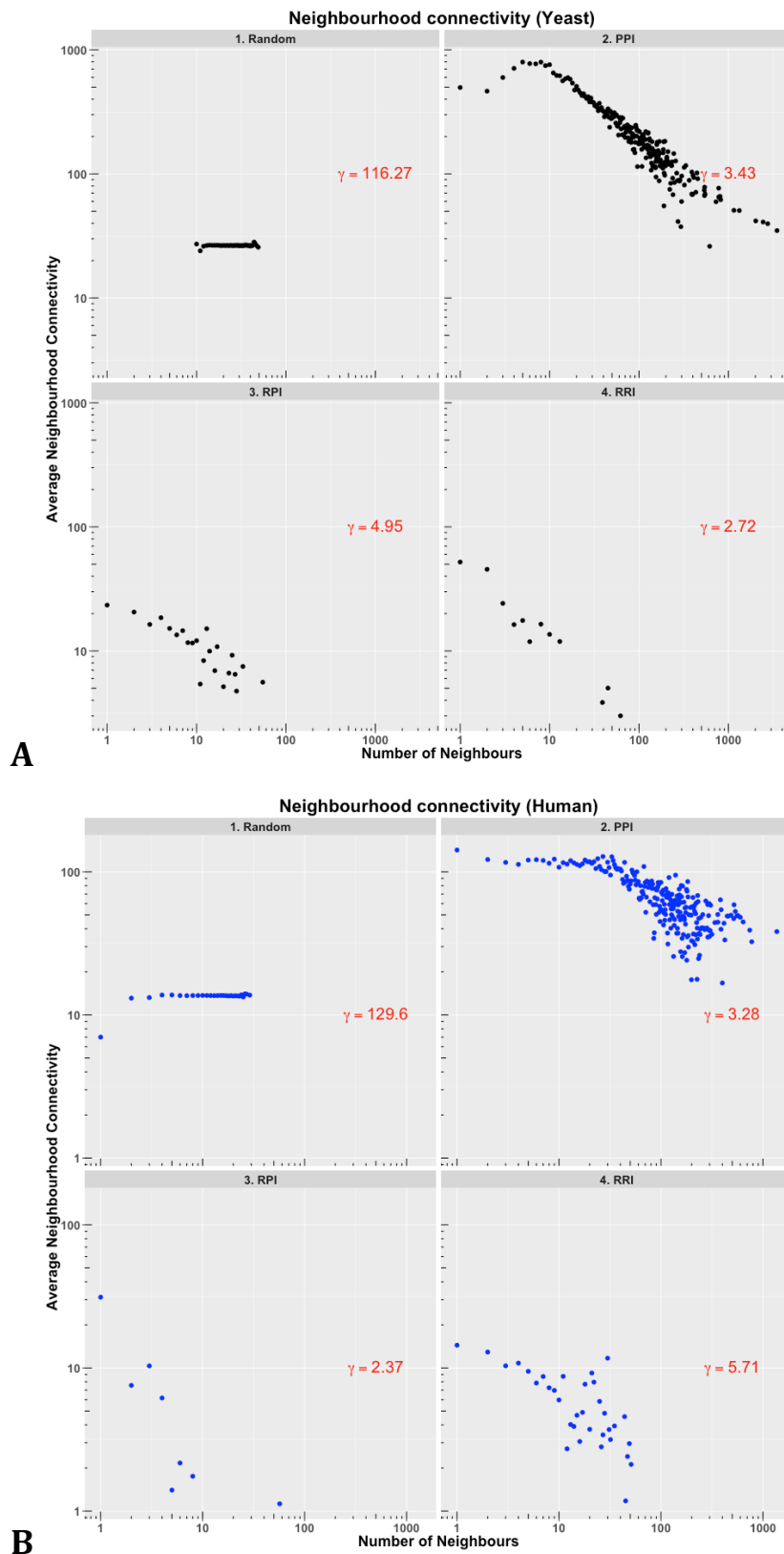


Figure 2.8 Distributions of neighbourhood connectivity values of nodes in (A) yeast and (B) human interaction networks.

Network property	PPI network (nodes: 6,091, edges: 77620)	Random network (nodes: 6,091, edges: 77,620)	RPI network (nodes: 258, edges: 596)	RRI network (nodes: 102, edges: 195)	Down-sampled PPI network* (edges: 195)
Degree distribution exponent (γ)	2.69	9.45	2.73	2.64	3.50
Clustering coefficient exponent (γ)	2.42	20.9	NA	1.68	2.82
Neighbourhood connectivity exponent (γ)	3.43	180.60	5.95	2.72	5.61
Assortativity coefficient (r)	-0.14	0.002	-0.31	-0.60	-0.14
Average shortest path	2.58	2.96	4.27	2.44	1.61

Table 2.1 Mean values of distributions of various network properties in yeast interaction networks. * denotes mean value of distribution from 100 down-sampled networks.

2.4 Conclusion

Representation of biological processes or systems in terms of networks offers a broad perspective of the cellular mechanism involved and it also allows investigation and identification of key links or interactions between molecules that are crucial for the system to function. In this chapter, using expert curated data of protein and RNA interactions from yeast, I have computed interaction networks of proteins and RNA and analysed their physical network properties. The study was conceived with the hypothesis that the fundamental differences in the way proteins and RNA interact with one another and with themselves would be reflected in their interaction networks and network properties, and the underlying similarities or differences in macromolecular interactions at large-scale could be inferred by comparing their network properties with each other.

By comparing PPI, RPI and RRI networks, I observe that their network properties exhibit similarity, despite the differences in how these macromolecules interact with each other. As in all real-world networks, a very few number of nodes called hubs interact with a large number of nodes. In PPI networks, these hub nodes represent molecular chaperones that are essential for mediating proper folding and functioning of a large number of proteins. In RRI network the ribosomal RNA subunits function as hubs that interact with snoRNAs and tRNAs. Hubs are central to the functioning of the network; biological networks are robust against the deletion of a peripheral node, however since hubs interact with a large number of nodes, there is a high probability of them engaging in interactions that are crucial for the organism's survival and therefore deletion of a hub in the network could lead to lethality (He and Zhang, 2006; Jeong et al., 2001).

Similarly other aspects of nodes within the network such as the tendency to form clusters, to have high centralities that help in communicating quickly with other nodes and high connectivity with neighbourhood nodes are similar between PPI and RRI networks. The only major difference observed among network properties is the absence of clustering in the RPI network, which is due to the

nature of RPI network. There are no links between RNA and protein, thus this leads to a bipartite graph for which the clustering coefficient is zero. Comparison of network properties of yeast with the human interaction data looks similar. There is poor correlation of biological networks with the random generated networks, indicating the absence of an organisational hierarchy among them.

Results from large-scale high-throughput studies represent a highly valuable resource to study interaction networks; however there are certain caveats that have to be taken into consideration when analysing and interpreting the resultant networks. I highlight a few of these limitations here. Firstly, a large number of experimental methods identify interacting partners between proteins and/or RNA, but they suffer from technical limitations wherein only certain interaction types or interactions between certain molecules are identified (Droit et al., 2005; Wheeler et al., 2017). For example, the yeast two-hybrid system, a widely used powerful method for identifying protein-protein interactions, cannot detect interactions between three or more proteins or those interactions that depend on post-translational modifications (Ito et al., 2002) and the high-throughput RNA immunoprecipitation (RIP) methods, such as RIP-chip and RIP-seq, may suffer from not detecting low affinity bound proteins to RNA (Wheeler et al., 2017). Second, it has been observed that cellular abundances and the number of interacting partners of a protein are correlated (Ivanic et al., 2009). In PPI networks, determined using interactions identified by affinity purification methods, proteins that represent hubs have high cellular abundances, but this correlation is absent in networks derived from interactions identified using the yeast two-hybrid system (Ivanic et al., 2009). Due to their high abundances and importance in diseases, some proteins and their interactions are more often identified by some techniques or frequently studied than others and could be represented as hubs in interaction networks. Another limitation to be noted is the inability of certain methods to distinguish specific from non-specific interactions resulting in high false-positive rates (Droit et al., 2005). These limitations suggest that networks derived from interactions identified by different experimental methods can have different properties. The choice of method therefore can potentially bias the observations.

By integrating high-confidence interactions that are identified using various techniques it is possible to overcome these experimental biases. In addition, pooling interactions from all techniques increases the number of interactions and our statistical power to make accurate inferences. I have used the interaction data from *Saccharomyces cerevisiae* to study the interaction network properties, since it represents one of the well-curated datasets. One of the important aspects of studying interaction networks is the availability of good quality data. The IntAct database (Orchard et al., 2014) contains high-quality manually curated interaction data obtained from both small-scale and high-throughput experimental studies. The IntAct database uses identifiers of ncRNA from RNACentral (The RNACentral Consortium, 2017) to unambiguously identify ncRNAs and link interaction data to a specific ncRNA.

In this study I have shown that the analysis of various network properties in yeast and human PPI, RPI and RRI networks indicate that, despite the differences in how proteins and RNA interact with each other or with themselves, on a large scale they exhibit similarity in their network characteristics. Comparison of interaction network properties derived from individual experimental methods is beyond the scope of this chapter. However in the future it would be interesting to analyse protein and RNA interaction networks derived from such experimental methods.

2.5 References

- Alm, E., and Arkin, A.P. (2003). Biological networks. *Current opinion in structural biology* 13, 193-202.
- Arita, M. (2005). Scale-freeness and biological networks. *Journal of biochemistry* 138, 1-4.
- Auweter, S.D., Oberstrass, F.C., and Allain, F.H. (2006). Sequence-specific binding of single-stranded RNA: is there a code for recognition? *Nucleic acids research* 34, 4943-4959.
- Barabasi, A.L., and Oltvai, Z.N. (2004). Network biology: understanding the cell's functional organization. *Nature reviews. Genetics* 5, 101-113.
- Bogan, A.A., and Thorn, K.S. (1998). Anatomy of hot spots in protein interfaces. *Journal of molecular biology* 280, 1-9.
- Bohnsack, M.T., Tollervey, D., and Granneman, S. (2012). Identification of RNA helicase target sites by UV cross-linking and analysis of cDNA. *Methods in enzymology* 511, 275-288.
- Borkovich, K.A., Farrelly, F.W., Finkelstein, D.B., Taulien, J., and Lindquist, S. (1989). hsp82 is an essential protein that is required in higher concentrations for growth of cells at higher temperatures. *Molecular and cellular biology* 9, 3919-3930.
- Cherry, J.M., Hong, E.L., Amundsen, C., Balakrishnan, R., Binkley, G., Chan, E.T., Christie, K.R., Costanzo, M.C., Dwight, S.S., Engel, S.R., *et al.* (2012). *Saccharomyces Genome Database: the genomics resource of budding yeast*. *Nucleic acids research* 40, D700-705.
- Chung, F., and Lu, L. (2002). The average distances in random graphs with given expected degrees. *Proceedings of the National Academy of Sciences of the United States of America* 99, 15879-15882.
- Danan, C., Manickavel, S., and Hafner, M. (2016). PAR-CLIP: A Method for Transcriptome-Wide Identification of RNA Binding Protein Interaction Sites. *Methods in molecular biology* 1358, 153-173.
- Droit, A., Poirier, G.G., and Hunter, J.M. (2005). Experimental and bioinformatic approaches for interrogating protein-protein interactions to determine protein function. *Journal of molecular endocrinology* 34, 263-280.
- Dujon, B. (2010). Yeast evolutionary genomics. *Nature reviews. Genetics* 11, 512-524.
- Ellis, J.J., Broom, M., and Jones, S. (2007). Protein-RNA interactions: structural analysis and functional classes. *Proteins* 66, 903-911.
- Filipovska, A., and Rackham, O. (2012). Modular recognition of nucleic acids by PUF, TALE and PPR proteins. *Molecular bioSystems* 8, 699-708.
- Forster, J., Famili, I., Fu, P., Palsson, B.O., and Nielsen, J. (2003). Genome-scale reconstruction of the *Saccharomyces cerevisiae* metabolic network. *Genome research* 13, 244-253.

- Galeota, E., Gravila, C., Castiglione, F., Bernaschi, M., and Cesareni, G. (2015). The hierarchical organization of natural protein interaction networks confers self-organization properties on pseudocells. *BMC systems biology* 9 Suppl 3, S3.
- Giege, R., Sissler, M., and Florentz, C. (1998). Universal rules and idiosyncratic features in tRNA identity. *Nucleic acids research* 26, 5017-5035.
- Girvan, M., and Newman, M.E. (2002). Community structure in social and biological networks. *Proceedings of the National Academy of Sciences of the United States of America* 99, 7821-7826.
- Goodfellow, I., and Bailey, D. (2014). Detection of protein-protein interactions using tandem affinity purification. *Methods in molecular biology* 1177, 121-133.
- Goodyear, C.S., and Silverman, G.J. (2008). Phage-display methodology for the study of protein-protein interactions: overview. *CSH protocols* 2008, pdb top48.
- Guan, Y., Dunham, M.J., and Troyanskaya, O.G. (2007). Functional analysis of gene duplications in *Saccharomyces cerevisiae*. *Genetics* 175, 933-943.
- Hahn, M.W., and Kern, A.D. (2005). Comparative genomics of centrality and essentiality in three eukaryotic protein-interaction networks. *Molecular biology and evolution* 22, 803-806.
- Hartwell, L.H., Hopfield, J.J., Leibler, S., and Murray, A.W. (1999). From molecular to modular cell biology. *Nature* 402, C47-52.
- He, X., and Zhang, J. (2006). Why do hubs tend to be essential in protein networks? *PLoS genetics* 2, e88.
- Helwak, A., and Tollervey, D. (2014). Mapping the miRNA interactome by cross-linking ligation and sequencing of hybrids (CLASH). *Nature protocols* 9, 711-728.
- Huppertz, I., Attig, J., D'Ambrogio, A., Easton, L.E., Sibley, C.R., Sugimoto, Y., Tajnik, M., Konig, J., and Ule, J. (2014). iCLIP: protein-RNA interactions at nucleotide resolution. *Methods* 65, 274-287.
- Ito, T., Ota, K., Kubota, H., Yamaguchi, Y., Chiba, T., Sakuraba, K., and Yoshida, M. (2002). Roles for the two-hybrid system in exploration of the yeast protein interactome. *Molecular & cellular proteomics : MCP* 1, 561-566.
- Ivanic, J., Yu, X., Wallqvist, A., and Reifman, J. (2009). Influence of protein abundance on high-throughput protein-protein interaction detection. *PLoS one* 4, e5815.
- Jeong, H., Mason, S.P., Barabasi, A.L., and Oltvai, Z.N. (2001). Lethality and centrality in protein networks. *Nature* 411, 41-42.
- Jones, S., Daley, D.T., Luscombe, N.M., Berman, H.M., and Thornton, J.M. (2001). Protein-RNA interactions: a structural analysis. *Nucleic acids research* 29, 943-954.
- Jones, S., and Thornton, J.M. (1996). Principles of protein-protein interactions. *Proceedings of the National Academy of Sciences of the United States of America* 93, 13-20.

- Joy, M.P., Brock, A., Ingber, D.E., and Huang, S. (2005). High-betweenness proteins in the yeast protein interaction network. *Journal of biomedicine & biotechnology* 2005, 96-103.
- Jubb, H.C., Pandurangan, A.P., Turner, M.A., Ochoa-Montano, B., Blundell, T.L., and Ascher, D.B. (2016). Mutations at protein-protein interfaces: Small changes over big surfaces have large impacts on human health. *Progress in biophysics and molecular biology*.
- Junge, A., Refsgaard, J.C., Garde, C., Pan, X., Santos, A., Alkan, F., Anthon, C., von Mering, C., Workman, C.T., Jensen, L.J., *et al.* (2017). RAIN: RNA-protein Association and Interaction Networks. *Database : the journal of biological databases and curation* 2017.
- Keskin, O., Gursoy, A., Ma, B., and Nussinov, R. (2008). Principles of protein-protein interactions: what are the preferred ways for proteins to interact? *Chemical reviews* 108, 1225-1244.
- Kladwang, W., Cordero, P., and Das, R. (2011). A mutate-and-map strategy accurately infers the base pairs of a 35-nucleotide model RNA. *Rna* 17, 522-534.
- Koschutzki, D., and Schreiber, F. (2008). Centrality analysis methods for biological networks and their application to gene regulatory networks. *Gene regulation and systems biology* 2, 193-201.
- Lee, T.I., Rinaldi, N.J., Robert, F., Odom, D.T., Bar-Joseph, Z., Gerber, G.K., Hannett, N.M., Harbison, C.T., Thompson, C.M., Simon, I., *et al.* (2002). Transcriptional regulatory networks in *Saccharomyces cerevisiae*. *Science* 298, 799-804.
- Light, S., Kraulis, P., and Elofsson, A. (2005). Preferential attachment in the evolution of metabolic networks. *BMC genomics* 6, 159.
- Lunde, B.M., Moore, C., and Varani, G. (2007). RNA-binding proteins: modular design for efficient function. *Nature reviews. Molecular cell biology* 8, 479-490.
- Ma'ayan, A. (2011). Introduction to network analysis in systems biology. *Science signaling* 4, tr5.
- Marsh, J.A., and Teichmann, S.A. (2015). Structure, dynamics, assembly, and evolution of protein complexes. *Annual review of biochemistry* 84, 551-575.
- Maslov, S., and Sneppen, K. (2002). Specificity and stability in topology of protein networks. *Science* 296, 910-913.
- Meenan, N.A., Sharma, A., Fleishman, S.J., Macdonald, C.J., Morel, B., Boetzel, R., Moore, G.R., Baker, D., and Kleanthous, C. (2010). The structural and energetic basis for high selectivity in a high-affinity protein-protein interaction. *Proceedings of the National Academy of Sciences of the United States of America* 107, 10080-10085.
- Meunier, J., Lemoine, F., Soumillon, M., Liechti, A., Weier, M., Guschanski, K., Hu, H., Khaitovich, P., and Kaessmann, H. (2013). Birth and expression evolution of mammalian microRNA genes. *Genome research* 23, 34-45.
- Mulder, N.J., Akinola, R.O., Mazandu, G.K., and Rapanoel, H. (2014). Using biological networks to improve our understanding of infectious diseases. *Computational and structural biotechnology journal* 11, 1-10.

- Nacher, J.C., Hayashida, M., and Akutsu, T. (2009). Emergence of scale-free distribution in protein-protein interaction networks based on random selection of interacting domain pairs. *Bio Systems* 95, 155-159.
- Nawrocki, E.P., Burge, S.W., Bateman, A., Daub, J., Eberhardt, R.Y., Eddy, S.R., Floden, E.W., Gardner, P.P., Jones, T.A., Tate, J., *et al.* (2015). Rfam 12.0: updates to the RNA families database. *Nucleic acids research* 43, D130-137.
- Nissen, P., Ippolito, J.A., Ban, N., Moore, P.B., and Steitz, T.A. (2001). RNA tertiary interactions in the large ribosomal subunit: the A-minor motif. *Proceedings of the National Academy of Sciences of the United States of America* 98, 4899-4903.
- Orchard, S., Ammari, M., Aranda, B., Breuza, L., Briganti, L., Broackes-Carter, F., Campbell, N.H., Chavali, G., Chen, C., del-Toro, N., *et al.* (2014). The MIntAct project--IntAct as a common curation platform for 11 molecular interaction databases. *Nucleic acids research* 42, D358-363.
- Panni, S., Prakash, A., Bateman, A., and Orchard, S. (2017). Yeast non-coding RNA interaction network. *Rna*.
- Papin, J.A., Hunter, T., Palsson, B.O., and Subramaniam, S. (2005). Reconstruction of cellular signalling networks and analysis of their properties. *Nature reviews. Molecular cell biology* 6, 99-111.
- Pavlopoulos, G.A., Secrier, M., Moschopoulos, C.N., Soldatos, T.G., Kossida, S., Aerts, J., Schneider, R., and Bagos, P.G. (2011). Using graph theory to analyze biological networks. *BioData mining* 4, 10.
- Piekna-Przybylska, D., Decatur, W.A., and Fournier, M.J. (2007). New bioinformatic tools for analysis of nucleotide modifications in eukaryotic rRNA. *Rna* 13, 305-312.
- Quek, X.C., Thomson, D.W., Maag, J.L., Bartonicek, N., Signal, B., Clark, M.B., Gloss, B.S., and Dinger, M.E. (2015). lncRNADB v2.0: expanding the reference database for functional long noncoding RNAs. *Nucleic acids research* 43, D168-173.
- Rajaratnam, T., and Lin, Y.H. (2006). Topological properties of protein-protein and metabolic interaction networks of *Drosophila melanogaster*. *Genomics, proteomics & bioinformatics* 4, 80-89.
- Ravasz, E. (2009). Detecting hierarchical modularity in biological networks. *Methods in molecular biology* 541, 145-160.
- Rizzolo, K., Wong, P., Tillier, E.R.M., and Houry, W.A. (2014). The Interaction Network of the Hsp90 Molecular Chaperone. 111-131.
- Rodrigo, G., and Fares, M.A. (2012). Describing the structural robustness landscape of bacterial small RNAs. *BMC evolutionary biology* 12, 52.
- Schmitz, J., Zemmann, A., Churakov, G., Kuhl, H., Grutzner, F., Reinhardt, R., and Brosius, J. (2008). Retroposed SNOfall--a mammalian-wide comparison of platypus snoRNAs. *Genome research* 18, 1005-1010.
- Schroeder, K.T., Daldrop, P., and Lilley, D.M. (2011). RNA tertiary interactions in a riboswitch stabilize the structure of a kink turn. *Structure* 19, 1233-1240.

Schwikowski, B., Uetz, P., and Fields, S. (2000). A network of protein-protein interactions in yeast. *Nature biotechnology* *18*, 1257-1261.

Shammas, S.L., Crabtree, M.D., Dahal, L., Wicky, B.I., and Clarke, J. (2016). Insights into Coupled Folding and Binding Mechanisms from Kinetic Studies. *The Journal of biological chemistry* *291*, 6689-6695.

Shannon, P., Markiel, A., Ozier, O., Baliga, N.S., Wang, J.T., Ramage, D., Amin, N., Schwikowski, B., and Ideker, T. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome research* *13*, 2498-2504.

Shao, P., Yang, J.H., Zhou, H., Guan, D.G., and Qu, L.H. (2009). Genome-wide analysis of chicken snoRNAs provides unique implications for the evolution of vertebrate snoRNAs. *BMC genomics* *10*, 86.

Sharma, A., Costantini, S., and Colonna, G. (2013). The protein-protein interaction network of the human Sirtuin family. *Biochimica et biophysica acta* *1834*, 1998-2009.

Sloan, D.J., and Hellinga, H.W. (1999). Dissection of the protein G B1 domain binding site for human IgG Fc fragment. *Protein science : a publication of the Protein Society* *8*, 1643-1648.

Stoiber, M.H., Olson, S., May, G.E., Duff, M.O., Manent, J., Obar, R., Guruharsha, K.G., Bickel, P.J., Artavanis-Tsakonas, S., Brown, J.B., *et al.* (2015). Extensive cross-regulation of post-transcriptional regulatory networks in *Drosophila*. *Genome research* *25*, 1692-1702.

Sumner-Smith, M., Roy, S., Barnett, R., Reid, L.S., Kuperman, R., Delling, U., and Sonenberg, N. (1991). Critical chemical features in trans-acting-responsive RNA are required for interaction with human immunodeficiency virus type 1 Tat protein. *Journal of virology* *65*, 5196-5202.

Sweeney, B.A., Roy, P., and Leontis, N.B. (2015). An introduction to recurrent nucleotide interactions in RNA. *Wiley interdisciplinary reviews. RNA* *6*, 17-45.

Szklarczyk, D., Franceschini, A., Wyder, S., Forslund, K., Heller, D., Huerta-Cepas, J., Simonovic, M., Roth, A., Santos, A., Tsafou, K.P., *et al.* (2015). STRING v10: protein-protein interaction networks, integrated over the tree of life. *Nucleic acids research* *43*, D447-452.

The RNAcentral Consortium (2017). RNAcentral: a comprehensive database of non-coding RNA sequences. *Nucleic acids research* *45*, D128-D134.

Ulyanov, N.B., and James, T.L. (2010). RNA structural motifs that entail hydrogen bonds involving sugar-phosphate backbone atoms of RNA. *New journal of chemistry = Nouveau journal de chimie* *34*, 910-917.

van der Geer, P. (2014). Analysis of protein-protein interactions by coimmunoprecipitation. *Methods in enzymology* *541*, 35-47.

Wallis, R., Leung, K.Y., Osborne, M.J., James, R., Moore, G.R., and Kleanthous, C. (1998). Specificity in protein-protein recognition: conserved Im9 residues are the major determinants of stability in the colicin E9 DNase-Im9 complex. *Biochemistry* *37*, 476-485.

Watts, D.J., and Strogatz, S.H. (1998). Collective dynamics of 'small-world' networks. *Nature* 393, 440-442.

Weber, M.J. (2006). Mammalian small nucleolar RNAs are mobile genetic elements. *PLoS genetics* 2, e205.

Wheeler, E.C., Van Nostrand, E.L., and Yeo, G.W. (2017). Advances and challenges in the detection of transcriptome-wide protein-RNA interactions. *Wiley interdisciplinary reviews. RNA*.

Xin, Y., Laing, C., Leontis, N.B., and Schlick, T. (2008). Annotation of tertiary interactions in RNA structures reveals variations and correlations. *Rna* 14, 2465-2477.

Yi, Y., Zhao, Y., Li, C., Zhang, L., Huang, H., Li, Y., Liu, L., Hou, P., Cui, T., Tan, P., *et al.* (2017). RAID v2.0: an updated resource of RNA-associated interactions across organisms. *Nucleic acids research* 45, D115-D118.

Yu, H., Kim, P.M., Sprecher, E., Trifonov, V., and Gerstein, M. (2007). The importance of bottlenecks in protein networks: correlation with gene essentiality and expression dynamics. *PLoS computational biology* 3, e59.

Zhang, X., Wu, D., Chen, L., Li, X., Yang, J., Fan, D., Dong, T., Liu, M., Tan, P., Xu, J., *et al.* (2014). RAID: a comprehensive resource for human RNA-associated (RNA-RNA/RNA-protein) interaction. *Rna* 20, 989-993.

Zhao, R., and Houry, W.A. (2007). Molecular Interaction Network of the Hsp90 Chaperone System. *594*, 27-36.

Zheng, L., Mairhofer, E., Teplova, M., Zhang, Y., Ma, J., Patel, D.J., Micura, R., and Ren, A. (2017). Structure-based insights into self-cleavage by a four-way junctional twister-sister ribozyme. *Nature communications* 8, 1180.

Chapter 3

Post-translational modifications of RNA-binding proteins

3.1 Introduction

RNA is a versatile macromolecule of diverse functions, which can be found in various functional contexts such as genetic, structural, regulatory and catalytic. To perform such diverse functions RNA interacts with a wide array of macromolecules, which include small molecules, nucleic acids and most importantly proteins. At the heart of RNA-protein interactions are RNA-binding proteins (RBPs), which regulate function through dynamic associations or disassociations with RNA based on various environmental cues. RNA-protein interactions are often specific and involve amino acid residues or nucleic acid bases that are essential for recognition and/or catalysis. Changes in macromolecular interactions can occur when one or more of these crucial residues undergo change through mutation or covalent modifications such as post-transcriptional or post-translational modifications. Although there is an abundance of experimentally validated post-translational modification data in public databases compared to post-transcriptional modification data, there are no systematic large-scale studies that focus on the influence of post-translational modifications on RNA-protein interactions. In this chapter, I present a comparative analysis of the post-translational modifications in RNA-binding proteins. First, I compare occurrences of various post-translational modifications between RNA-binding, non RNA-binding and DNA-binding proteins. Next, I compare the occurrences of various post-translational modifications in RNA-binding regions compared to non RNA-binding sites within RNA-binding proteins. I also investigate the relation between post-translational modification,

amino acid abundance, protein abundance and structural disorderedness. Finally, I investigate interactions between RNA and protein by comparing RNA-binding peptides from experimental data with experimental structures of RNA-protein complexes.

RNA serves as an important molecule at the core of many cellular functions; it serves as genetic material in single and double-stranded RNA viruses, as a template to transcribe the genetic code in the form of messenger RNA (mRNA), as an adaptor or structural component during protein synthesis in the form of transfer RNAs (tRNA) and ribosomal RNAs (rRNA) and as a gene regulatory elements in the form of small and long non-coding RNAs, among many other functions. Each of these events is associated with RNA interacting transiently or stably with RNA-binding proteins (RBPs) to form ribonucleoprotein (RNP) complexes. RBPs contain various structural motifs such as RNA recognition motifs (RRM), K-homology (KH) domain, double stranded RNA-binding domain and RNA-binding zinc-finger (ZnF) domains, through which they recognise and bind RNA (Lunde et al., 2007). RBPs form a diverse range that can bind different class, type and sequences of RNA. Certain RBPs, on the one hand, are monomeric, form small complexes and are highly specific in binding specific class of RNAs; for example, the argonaute proteins bind to small non-coding RNAs such as microRNAs (miRNAs), short interfering RNAs (siRNAs) and PIWI-associated RNAs (piRNAs) and form RNA Induced Silencing Complex (RISC) regulating gene expression (Meister, 2013). The pre-mRNP and mRNP complexes, on the other hand, are large multimegadalton complexes that bind pre-mRNA and mRNA, and comprise five small nuclear RNAs (snRNAs) - U1, U2, U4, U5 and U6, and numerous proteins, which are involved in mRNA splicing, polyadenylation, stabilisation, localisation and translation (Muller-McNicoll and Neugebauer, 2013; Will and Luhrmann, 2011). Some RBPs such as the CUG triplet RNA binding protein 1 (CUGBP1) and muscleblind-like protein 1 (MBNL1) bind to sequence specific tri and tetra-nucleotide mRNA repeats respectively and are sequestered, which leads to myotonic dystrophy (Ranum and Day, 2004; Timchenko et al., 1996), while others are generic RNA-binding proteins with no sequence specificity that form RNPs. Other class of RBPs specifically bind single

and double-stranded RNAs using single or double stranded-RNA-binding motifs/domains (Antson, 2000; Tian et al., 2004). Interestingly, recent studies have discovered numerous moonlighting non-canonical RNA-binding proteins, which involve metabolic enzymes such as aconitase 1 (ACO1) and glyceraldehyde-3-phosphate dehydrogenase (GAPDH) (Castello et al., 2015).

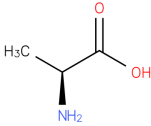
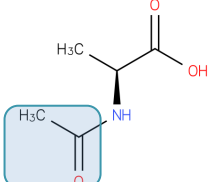
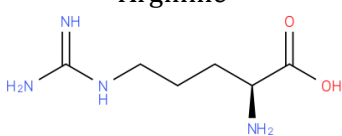
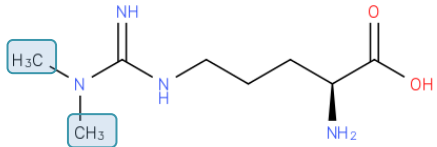
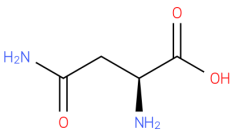
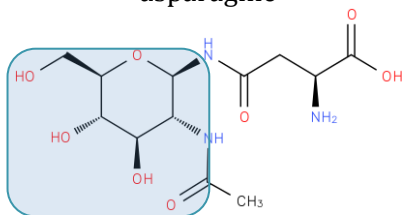
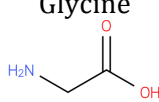
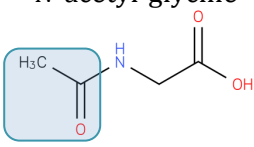
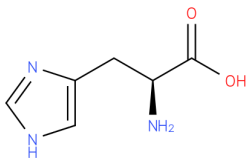
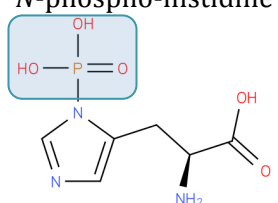
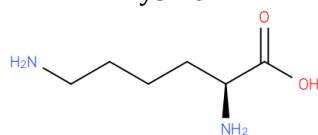
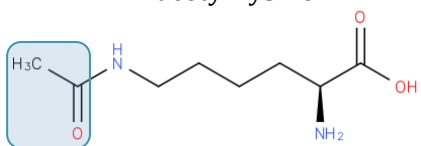
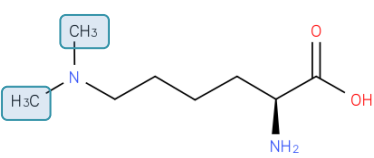
The development of large-scale quantitative methods using immunoprecipitation, mass spectrometry and deep-sequencing has facilitated genome-wide identification of RBPs and their RNA targets (Konig et al., 2012). A large number of RNA-binding proteins have been identified using methods such as RNA immunoprecipitation (RIP), cross-linking immunoprecipitation (CLIP) followed by sequencing (CLIP-seq) (Darnell, 2010; Ule et al., 2005; Yang et al., 2015), photoactivable ribonucleoside-enhanced CLIP (PAR-CLIP) (Hafner et al., 2010) and individual nucleotide-resolution CLIP (iCLIP) (Yao et al., 2014). Recent studies with advancements of the above-mentioned methods have provided a comprehensive atlas of RNA-binding proteins in humans (Baltz et al., 2012) (Castello et al., 2012). Various studies have estimated ~500 RBPs in mice and humans (Cook et al., 2011; McKee et al., 2005), ~700 RBPs in humans including RNA-binding domains (RBDs) involved in other aspects of RNA metabolism (Anantharaman et al., 2002), ~1,900 human RBPs obtained through automated functional annotations (Ashburner et al., 2000) and 1,542 human RBPs using Pfam RBDs in RNA-related proteins (Gerstberger et al., 2014). Availability of good quality data of RNA-binding proteins is better for studying various aspects of RNA-protein interaction. The RNA-binding domains map (RBDmap) method (Castello et al., 2016), an improvement over the RNA interactome capture technique (Castello et al., 2013), comprehensively identifies proteins cross-linked to RNA at the peptide level with high-resolution (for a detailed description of experimental method see method section 3.2.1). The availability of RNA-bound peptide data was one of the main factors that influenced using this dataset for the study.

Although numerous RBPs have been identified it is still unclear how specificity, or non-specificity, is achieved by RBPs (Jankowsky and Harris, 2015) or how

interactions between RNA and proteins are regulated. It is suggested that post-translational modifications of RBPs influence their associations, enzymatic activities and localisations with their RNA targets (Turner et al., 2014). Post-translational modifications (PTMs) alter the surface electrostatic potential of binding sites and dynamically regulate interaction with other proteins, nucleic acids or small ligands. PTMs are usually enzyme mediated covalent modifications of proteins through addition or removal of either small functional groups such as a phosphate, acetate, methyl or carbohydrate moieties, or small proteins such as ubiquitin or small ubiquitin-like modifier (SUMO). Phosphorylation is the most abundant and well-studied PTM, which is involved in regulating a large number of cellular processes. Phosphorylation usually occurs at serine, threonine, tyrosine and histidine residues, whereby addition or removal of phosphate (PO_4)³⁻ by protein kinases or phosphatases respectively. For example, a recent study of ELAV/Hu RNA-binding proteins indicates direct phosphorylation of residues at RNA-binding surface results in abolishing or decreasing affinity to RNA (Brauer et al., 2014). Acetylation is the second most commonly observed PTM and is mostly observed in metabolic enzymes and proteins involved in gene expression regulation (Verdin and Ott, 2015). Acetyltransferases and deacetylases catalyse the addition or removal of acetyl group (CH_3CO) respectively either onto the N-terminus of proteins or lysine residues (Drazic et al., 2016). Acetylation of TDP-43, an RNA-binding protein, was shown to impair RNA-binding and promote protein accumulation (Cohen et al., 2015). Protein methylation is well studied in histones and is targeted at lysine or arginine residues by amino acid residue specific methyltransferases, which can add one or two methyl (CH_3) groups onto arginine and up to three methyl groups onto lysine residues. Both acetylation and methylation are important PTMs that are implicated in gene expression regulation by regulating chromatin structure and transcription (Drazic et al., 2016; Lee et al., 2005). Acetylation and methylation of various RNA-binding proteins and on histones is shown to alter specificity and binding with RNA or DNA (Blackwell and Ceman, 2012; Cohen et al., 2015; Lee et al., 2005; Rothbart and Strahl, 2014). In a recent study ubiquitination was shown to influence RNA-binding and catalysis, wherein ubiquitination of *Leishmania donovani* cycling sequence binding protein (LdCSBP) resulted in the loss of

endoribonuclease activity (Bhandari et al., 2011). SUMOylation of RNA-binding protein La, was shown to facilitate small RNA oligonucleotide binding and de-SUMOylation impaired RNA-binding activity in the cells (Kota et al., 2016). Although certain functionalities are linked to a specific PTM, most proteins undergo more than one type of PTM under different cellular conditions (Pejaver et al., 2014), and implicated in multiple signalling pathways (Zeidan and Hart, 2010). It was recently shown that cross talk exists between certain pairs of PTMs – such as phosphorylation and O-linked glycosylation (Butt et al., 2012; Wang et al., 2014; Zeidan and Hart, 2010) or protein degradation through promotion of ubiquitination by phosphorylation (Vodermaier, 2004), which could result in diverse and fine-tuned functional outcomes. Table 3.1 lists some of the common PTMs.

Experimentally observed PTMs from a large number of proteins have been reported in numerous papers. Databases such as PhosphoSitePlus (Hornbeck et al., 2015) and UniProt store modification sites manually curated from literature and identified by large-scale proteomics methods. Databases such as dbPTM (Huang et al., 2016) and PHOSIDA (Gnad et al., 2011), in addition also provide tools to predict modification sites in protein sequences. For a large number of proteins, their PTM sites have been integrated with protein functions; PTMcode, a database of known and predicted functional associations between PTMs in proteins, annotates nearly 17 million functional associations with 1.6 million PTM sites in 100,000 proteins (Minguez et al., 2013). It is clear that PTMs and their functional associations with respect to a protein or a pathway have been well characterised, and a few studies have investigated the role of PTMs in protein-protein interactions within a system (Chavez et al., 2013; Duan and Walther, 2015; Woodsmith et al., 2013). Compared to protein-protein interactions, studies on the influence of PTMs on RNA-protein interactions within a network are scarce. A recent study has shed some light on regulating RNA-protein interaction through PTM by identifying regulatory motifs in the 3' untranslated region (UTR) of RNA that are sensitive to PTM of a specific RBP and interact with the RBP in a PTM-dependent manner (Brown et al., 2015).

Amino acid	Modified amino acid	Biological significance	Reference
<p>Alanine</p> 	<p><i>N</i>-acetyl-alanine</p> 	Modulate protein-protein interactions, protein stability & translocation	(Arnesen, 2011)
<p>Arginine</p> 	<p><i>N</i>ω-<i>N</i>ω-dimethyl-arginine</p> 	Epigenetic regulation of gene expression.	(Bedford and Clarke, 2009)
<p>Asparagine</p> 	<p><i>N</i>⁴-(β-<i>N</i>-acetyl-D-glucosaminyl)-asparagine</p> 	Facilitate and stabilise protein folding.	(O'Connor and Imperiali, 1996)
<p>Glycine</p> 	<p><i>N</i>-acetyl-glycine</p> 	Modulate protein-protein interactions, protein stability & translocation	(Arnesen, 2011)
<p>Histidine</p> 	<p><i>N</i>-phospho-histidine</p> 	Signal transduction.	(Stegg et al., 2003)
<p>Lysine</p> 	<p><i>N</i>⁶-acetyl-lysine</p> 	Epigenetic regulation of gene expression.	(Fukuda et al., 2006)
	<p><i>N</i>⁶-<i>N</i>⁶-dimethyl-lysine</p> 	Epigenetic regulation of gene expression.	(Martin and Zhang, 2005)

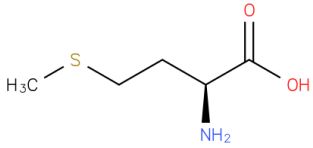
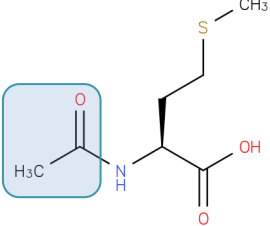
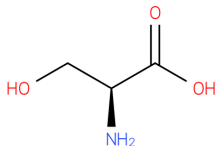
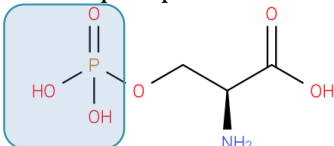
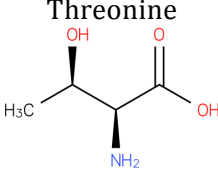
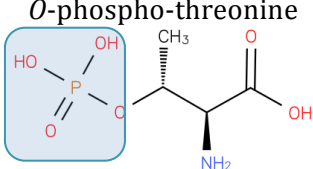
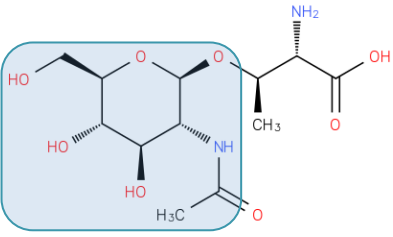
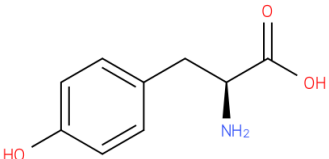
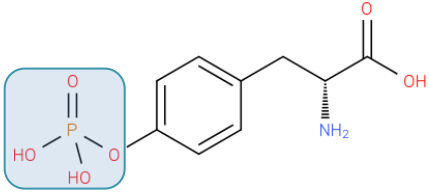
<p>Methionine</p> 	<p>N-acetyl-methionine</p> 	<p>Modulate protein-protein interactions, protein stability & translocation</p>	<p>(Arnesen, 2011)</p>
<p>Serine</p> 	<p>O-phospho-serine</p> 	<p>Basic cellular processes such as metabolism, growth, division, immunity, differentiation, membrane transport, etc.</p>	<p>(Manning et al., 2002)</p>
<p>Threonine</p> 	<p>O-phospho-threonine</p> 	<p>Basic cellular processes such as metabolism, growth, division, immunity, differentiation, membrane transport, etc.</p>	<p>(Manning et al., 2002)</p>
	<p>O-(N-acetyl-β-D-glucosaminyl)-L-threonine</p> 	<p>Immunity, protein-protein interactions</p>	<p>(Szymanski and Wren, 2005)</p>
<p>Tyrosine</p> 	<p>O⁴-phospho-tyrosine</p> 	<p>Basic cellular processes such as metabolism, growth, division, immunity, differentiation, membrane transport, etc.</p>	<p>(Manning et al., 2002)</p>

Table 3.1 Common post-translational modifications. Post-translational modifications are highlighted in blue boxes. Figures are adapted from ChEBI.

3.2 Methods

3.2.1 RNA-binding peptides

RNA-binding peptides identified by (Castello et al., 2016) in human HeLa cells were used in this study. The RNA-bound peptides are termed 'RBDpep' (enriched at 1% FDR) and 'candidate RBDpep' (enriched at 10% FDR) (Castello et al., 2016). The RNA-binding peptides of RBDpep and candidate RBDpep dataset comprise peptides that are cross-linked to RNA after U.V. treatment (X-link) and ~17 amino acids long native peptides adjacent to crosslinking site (N-link). Together the X-link and N-link peptides constitute RNA-binding peptides (Figure 3.1). The RBDpep dataset consists of 1,380 overlapping regions of RNA-bound peptides from 529 RNA-binding proteins, and the candidate RBDpep dataset consists of 2,079 overlapping regions of RNA-bound peptides from 865 candidate RNA-binding proteins. The number of proteins that are in common between RBDpep data and the candidate RBDpep datasets is 392 (74.1% of RBDpep) – the reason for partial overlap was suggested in the experimental design, where the two datasets were generated with two different proteases (LysC or ArgC) and analysed independently (Alfredo Castello, personal communication, September 9, 2016). Therefore, instead of combining the two datasets, I have independently used the two datasets for downstream analyses. Data in sheets named "RBDpep" and "CandidateRBDpep" in supplementary file "Table S1" of (Castello et al., 2016) correspond to RBDpep and candidate RBDpep dataset respectively.

The overlapping RNA-bound peptide segments were manually filtered to remove duplicate entries and to identify peptide contigs. A total of 784 contigs of RBDpep and 1,433 contigs of candidate RBDpep were identified. The overall lengths of RNA-bound peptides (contigs) for each protein were then calculated. Accession numbers of 9 RNA-binding proteins in RBDpep and candidate RBDpep datasets were replaced or obsolete in UniProt (release 2016_10), hence these proteins were not included in downstream analyses.

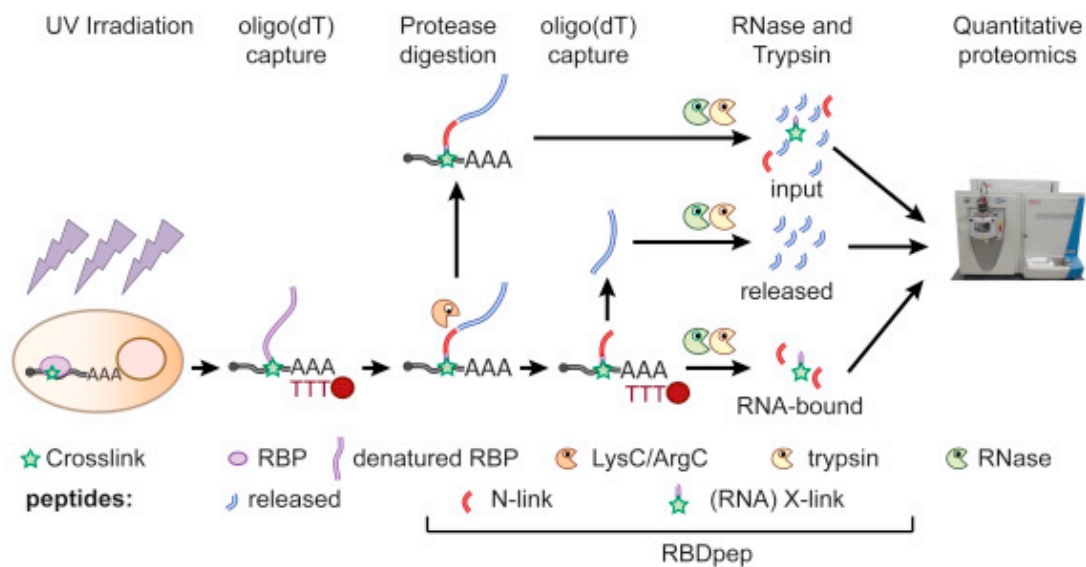


Figure 3.1 Schematic representation of RBDmap workflow. RNA is cross-linked to proteins using UV light. After cell lysis, proteins cross-linked to poly(A) RNA are isolated using oligo(dT) magnetic beads. RBPs are proteolytically digested using either ArgC or LysC, which releases unbound peptides. The RNA bound peptides are eluted and treated with trypsin and RNase, which results in the release of peptides that are directly cross-linked to RNA (X-link) and its neighbouring peptides (N-link), which together are named “RBDpep”. The eluted peptides are fed to mass spectrometer. The N-link peptides are identified using peptide search algorithms and the X-linked peptides are re-derived *in-silico* by extending the mass spectrometer identified peptides to the nearest LysC or ArgC cleavage sites. Figure adapted from (Castello et al., 2016) DOI: 10.1016/j.molcel.2016.06.029.

3.2.2 Non RNA-binding proteins

Although databases annotate proteins as 'RNA-binding' based on experimental evidences, a complete exhaustive list of true 'non RNA-binding' proteins is not available. In order to create a reference set of non RNA-binding proteins (non-RBPs) as a negative control, RNA-binding proteins from RBDpep dataset were removed from the human proteome downloaded from UniProt. A total of 17,305 proteins without isoforms were used as non-RBPs. Similarly a control set of non-RBPs was created using RNA-binding proteins from the candidate RBDpep data. Although this approach would give a set of non-RBPs, it cannot be guaranteed that all proteins within this set would not have RNA-binding functions, a small fraction of proteins in the non-RBP datasets may be involved in some aspects of RNA-binding. However, if an upper limit of 1,900 human RBPs, as suggested by (Ashburner et al., 2000), were considered, it would suggest that there would be 1,371 RBPs (1,900 RBPs – 529 RBPs) present within the non-RBP dataset, which is only 7.92% of the entire non-RBP set and may not significantly influence the observations. Therefore given the current limitations of experimental protocols in completely identifying all RNA-binding proteins this approach of generating a list of non-RBPs was assumed reasonable.

3.2.3 RNA-binding proteins

Apart from lists of RBPs in RBDpep and candidate RBDpep datasets, a third independent dataset comprising human RBPs from Swiss-Prot was used as a comparative set. 659 manually curated human RBPs were downloaded from UniProt (release 2016_10) using the terms 'RNA-binding [KW-0694]' and 'Homo sapiens [9606]' in the advanced 'Keyword [KW]' and 'Organism [OS]' search options respectively. A negative control list of 17,189 non-RBPs was created similarly as described above.

3.2.4 DNA-binding proteins

A fourth dataset independent of RBPs was used to assess the nature of various post-translational modifications in DNA-binding proteins. This dataset consisting of 1,998 Swiss-Prot manually curated DNA-binding proteins in human were downloaded from UniProt (release 2016_10) using the terms 'DNA-binding [KW-0238]' and 'Homo sapiens [9606]' in the advanced 'Keyword [KW]' and 'Organism [OS]' search options respectively. A negative control list of 15,900 non DNA-binding proteins was created similarly as described above.

3.2.5 Post-translational modifications

Post-translational modifications of human proteins were downloaded from PhosphoSitePlus (July 3, 2016) (Hornbeck et al., 2015). The manually curated dataset consists of 299,538 protein modification sites (amino acid sequence coordinates) of experimentally observed post-translational modifications such as acetylation, methylation, phosphorylation, SUMOylation, ubiquitination and glycosylation (O-GalNAc and O-GlcNAc) from 20,488 human proteins. Perl scripts were developed to map post-translational modification sites onto RBDpep, candidate RBDpep, RNA and DNA-binding protein datasets. The methylation dataset was filtered to remove duplicate residue entries with di- and tri-methylation marks. Statistical analyses and graphs were plotted using the R-package.

3.2.6 Globular and disordered regions

Globular and disordered regions in RBPs were identified using IUPred (Dosztanyi et al., 2005). Amino acid sequences were analysed with the IUPred standalone package with a disorder cut-off value of 0.5. Amino acids with a score above 0.5 were considered to contribute to disorder.

3.2.7 Structural validation

The RNA-bound peptides identified by (Castello et al., 2016) were validated using structural data of RNA-protein complexes from the Protein Data Bank

(August 16, 2016). The Protein Data Bank (PDB) was queried for human RBPs using the advanced search option. A total of 312 PDB entries of RNA-protein complexes were downloaded, which include 462 unique UniProt accessions. Other protein structures that are in complex with DNA-RNA hybrid, or comprised only of C-alpha trace and structures solved using electron microscopy were neglected. Out of the 462 proteins, 94 proteins from 215 PDB entries were common with RBDpep dataset (Castello et al., 2016). This set was further filtered to remove those proteins that were not bound to RNA. Finally a set of 66 proteins from 49 PDB entries was obtained, that were common with the RBDpep dataset (Castello et al., 2016). The interatomic RNA-protein contacts were computed using the WHAT IF server (Vriend, 1990), which calculates the distance between Van der Waals surfaces of any two atoms of protein and RNA using the standard Van der Waals radii. A total of 2,317 RNA-protein contacts were inferred. Residue-level mapping between UniProt and PDB entries was carried out using SIFTS (Velankar et al., 2013).

3.2.8 Protein abundance

The relative protein abundance data of the human proteome (whole organism, integrated) was downloaded from the PaxDb protein abundance database, version 4.0. (Wang et al., 2015).

3.3 Results

3.3.1 Overview of RBDpep and candidate RBDpep datasets

Castello et al., 2016, provides a comprehensive atlas of RNA-binding domains present within RNA-binding proteins in humans. In order to verify this set of RBPs, I compared this data with a list of known RBPs from Swiss-Prot. It is observed that the RBDpep dataset and Swiss-Prot share 163 proteins (24.73% of RNA-binding proteins in Swiss-Prot) between them (Figure 3.2A), while 242 proteins from the candidate RBDpep data are common with RNA-binding

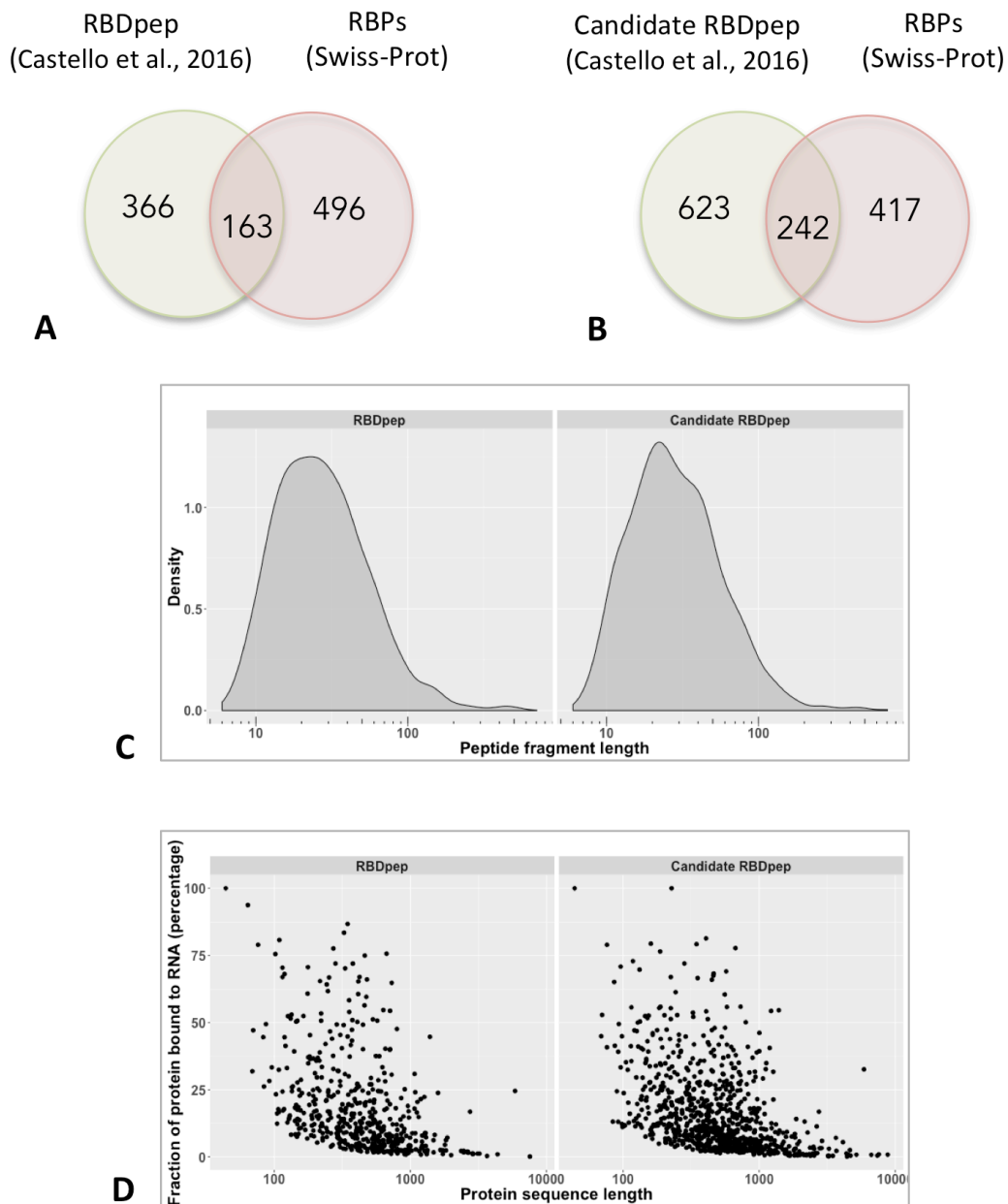


Figure 3.2 Features of RBDpep and candidate RBDpep datasets. Number of RNA-binding proteins that are in common between annotated RNA-binding proteins in Swiss-Prot and (A) RBPs in RBDpep dataset and (B) RBPs in candidate RBDpep dataset. (C) Distribution of peptide lengths for RNA-binding peptides in RBDpep dataset and candidate RBDpep dataset. The average RNA-binding peptide lengths in RBDpep and candidate RBDpep datasets are 37 and 38 amino acids respectively. (D) Overall length of RNA-bound peptides represented as a fraction of total protein length.

proteins from Swiss-Prot (36.72% of RNA-binding proteins in Swiss-Prot) (Figure 3.2B). Some of the reasons that are suggested for not detecting other known RBPs by RBDmap are (i) non-binding to poly(A) RNAs, (ii) low cross-linking efficiency (iii) interactions with the phospho-sugar backbone, but not the nucleotide bases or (iv) lack of cleavage sites for trypsin within LysC or ArgC proteolytic fragments (Castello et al., 2016). RNA-binding peptides (RBDpep and candidate RBDpep) identified by (Castello et al., 2016) have an average length of 37 amino acids (RBDpep) and 38 amino acids (candidate RBDpep) (Figure 3.2C). The shortest and longest peptides in the RBDpep dataset are 6 and 539 amino acids respectively; peptides from candidate RBDpep dataset similarly range from 6 to 706 amino acids.

The RNA-binding peptide contigs were used to infer the coverage of proteins in terms of their RNA-binding. The overall RNA-binding peptide lengths of RBDpep and candidate RBDpep datasets indicate that for nearly 77% of the proteins in the two datasets, the RNA-binding fraction of protein is less than 25% (Figure 3.2D), however, the percentage of amino acid residues directly interacting with RNA within RNA-binding peptides is presumed to be lower. The percentage of RNA-binding residues was observed to be between 30% in ribosomal proteins and 9% in non-ribosomal proteins (Chen et al., 2014).

3.3.2 Post-translational modifications in RNA-binding proteins and non RNA-binding proteins

Comparisons of RBPs with non-RBPs have indicated higher protein abundance, higher half-life (higher stability), lower levels of biological noise (tightly regulated gene expression) (Mittal et al., 2009) and significantly higher tissue expression levels (Kechavarzi and Janga, 2014) in RBPs. To investigate the prevalence of PTMs, I compared occurrences of PTMs in RBPs and non-RBPs. It is seen that RBPs have a higher frequency of sites for PTM than non-RBPs; 7.08% of the total amino acid residues in RBPs are post-translationally modified, which is significantly higher compared to 2.47% of residues that are post-translationally modified in non-RBPs (Table 3.2) (P-value < 2.2×10^{-16} , Chi-

square test. N. B. This is the lowest possible value for this statistical test in R.). It is also found that the PTM occupancy of an amino acid residue (i.e., number of modifications per modified amino acid) in RBPs is significantly more compared to PTM occupancy of amino acid residues in non-RBPs (P-value = 1.862×10^{-09} , Chi-square test); a total of 23,370 PTMs were observed in 21,468 modified amino acids residues in RBPs (occupancy of 1.08 PTM per modified amino acid residue), in comparison to 265,048 PTMs observed in 258,227 modified amino acid residues in non-RBPs (occupancy of 1.02 PTM per modified amino acid residue). The number of RBPs, as discussed earlier, can vary between different experimental methods and may not completely agree with each other. Therefore, I have used an independent set of manually curated RBPs from Swiss-Prot as a gold-standard positive control dataset for comparison. Analysing PTM sites using RBPs curated in Swiss-Prot further supports these observations; 4.78% amino acid residues are modified in RBPs and are frequently targeted compared to 2.52% post-translationally modified amino acid residues in non-RBPs.

A large majority of RBPs are involved in at least one of these cellular processes: synthesis, folding, transport, assembly and clearance of RNA. RNA is seldom found naked *in vivo*, but is normally bound with one or more proteins (Mitchell and Parker, 2014). Unlike a majority of non-RBPs, whose targets do not involve being acted upon by a large number of proteins at the same time, most RBPs act co-operatively i.e., form multimeric complexes with their substrates, for example the transcription complex, spliceosome, editosome complex, mRNA localization complex, translation complex and exosome complex, which involve more than 20 proteins interacting with RNA either all together at the same time or at different stages in RNA's life (Jurica and Moore, 2003; Panigrahi et al., 2006). The components of these multimeric complexes (both proteins and their cognate RNA targets) have to be available in the correct stoichiometry and in their active states to function. It has been shown that PTMs such as acetylation preferably target large macromolecular complexes involved in processes such as chromatin remodelling, cell cycle, nuclear transport and splicing (Choudhary et al., 2009). Therefore, it is reasoned here that PTMs play a larger role in maintaining cellular availability and influencing functions of RBPs than non-RBPs.

	Number of PTMs	Number of amino acids with PTM	Total number of amino acids	PTM occupancy (PTM/modified amino acid)	Percentage of amino acid residues modified
RNA-binding proteins (RBPs)	23,370	21,468	302,895	1.08	7.08%
Non RNA-binding proteins (non-RBPs)	265,048	258,227	10,439,949	1.02	2.47%
Swiss-Prot RBPs	19,459	18,382	383,963	1.05	4.78%
Swiss-Prot non-RBPs	269,340	261,758	10,363,333	1.02	2.52%
Swiss-Prot DNA-binding proteins (DBPs)	41,841	39,515	1,178,225	1.05	3.35%
Swiss-Prot non DNA-binding proteins (non-DBPs)	247,150	240,515	9,569,743	1.02	2.51%

Table 3.2 Comparison of post-translational modifications across nucleic acid binding and non-binding proteins in different datasets.

Recent studies have shown an increasing number of proteins that bind both RNA and DNA, which include p53, the heterogeneous ribonucleoproteins (hnRNPs) and transcription factors TFIIIA and FUS (Cassiday and Maher, 2002; Hudson and Ortlund, 2014). Due to the dual and sometimes overlapping functionalities of DNA-binding proteins (DBPs) and RBPs to bind targets competitively or cooperatively, they are subject to tight regulation. In order to understand if PTMs play a differential role in regulating DBPs and RBPs, I compared the occurrences of PTM sites in DBPs and non-DBPs with RBPs and non-RBPs respectively (Table 3.2). DBPs have significantly more PTM sites than non-DBPs (P-value < 2.2×10^{-16} , Chi-square test); about 3.35% of amino acid residues in DBPs are modified compared to 2.51% in non-DBPs. Modified amino acid residues are significantly more frequently targeted in DBPs than non-DBPs (P-value = 6.737×10^{-5} , Chi-square test). These trends are similar but weaker compared to those observed in RBPs.

PTM sites were further analysed by the type of PTM. By comparing occurrences of various PTMs between nucleic acid binding proteins and non nucleic acid binding proteins, it is observed that almost all types of PTMs are overrepresented in RBPs and DBPs compared to their non nucleic acid binding counterparts (Figure 3.3). Among RBPs 4.90% amino acid residues are phosphorylated which is significantly higher than 1.93% amino acid residues phosphorylated in non-RBPs (P-value < 2.2×10^{-16} , Chi-square test). Table 3.3 shows the counts and percentages of various PTMs in RBPs and non-RBPs. However an opposite trend is observed in both RBPs and DBPs undergoing O-linked N-acetylglucosamine (O-GlcNAc) glycosylation, where it is frequent in non-RBPs and non-DBPs. O-linked N-acetylgalactosamine (O-GalNAc) and O-GlcNAc glycosylation are present in a small number of RBPs and are fewer in numbers relative to other modifications. Phosphorylation is by far the predominant PTM in eukaryotic proteins among other PTM types and dominates the PTM landscape because of its versatility and ready reversibility among various other properties (Hunter, 2012), which is the reason for a prominent phosphorylation signal in Figure 3.3. A similar PTM trend is also observed in candidate RBPs (see appendix Figure A1).

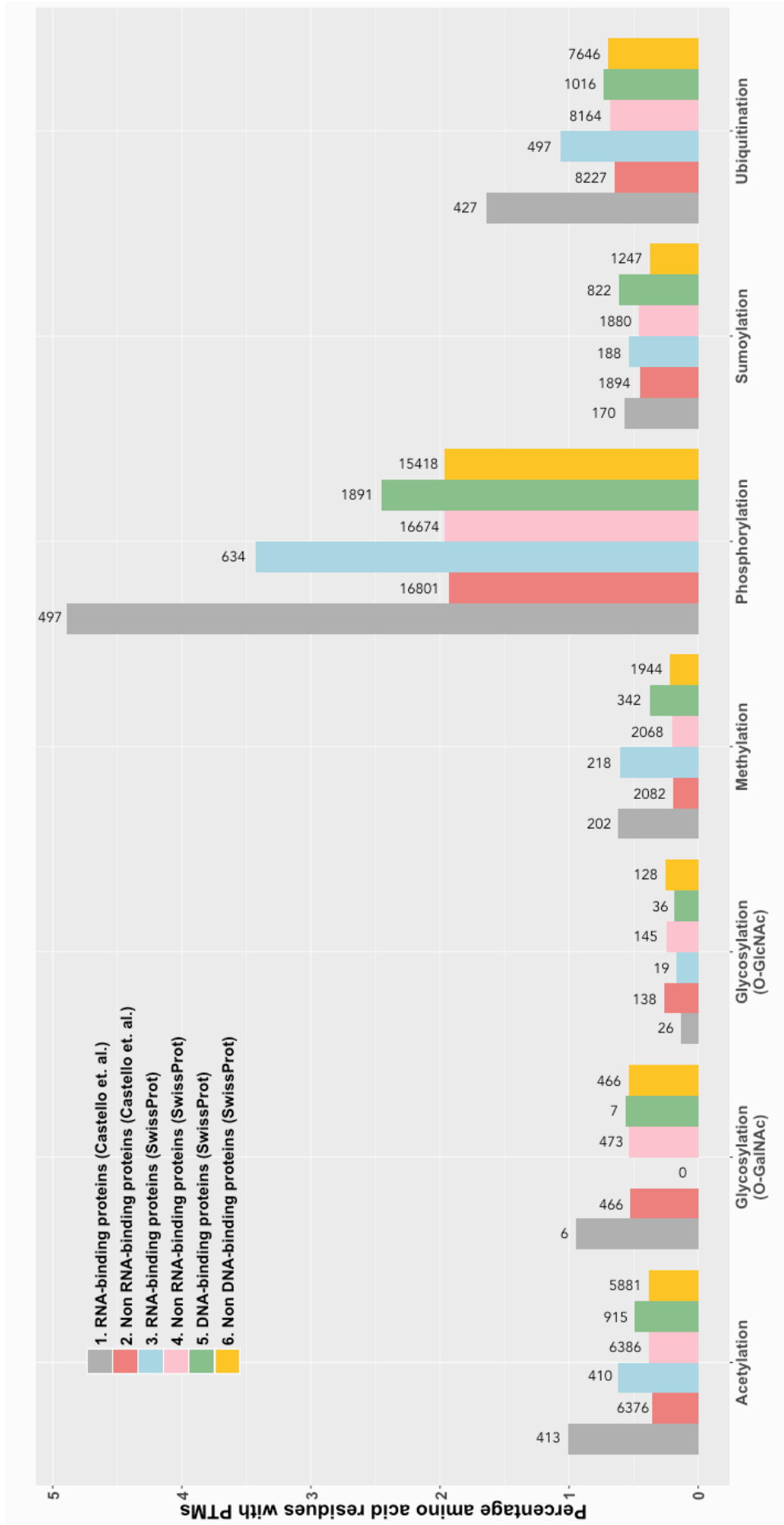


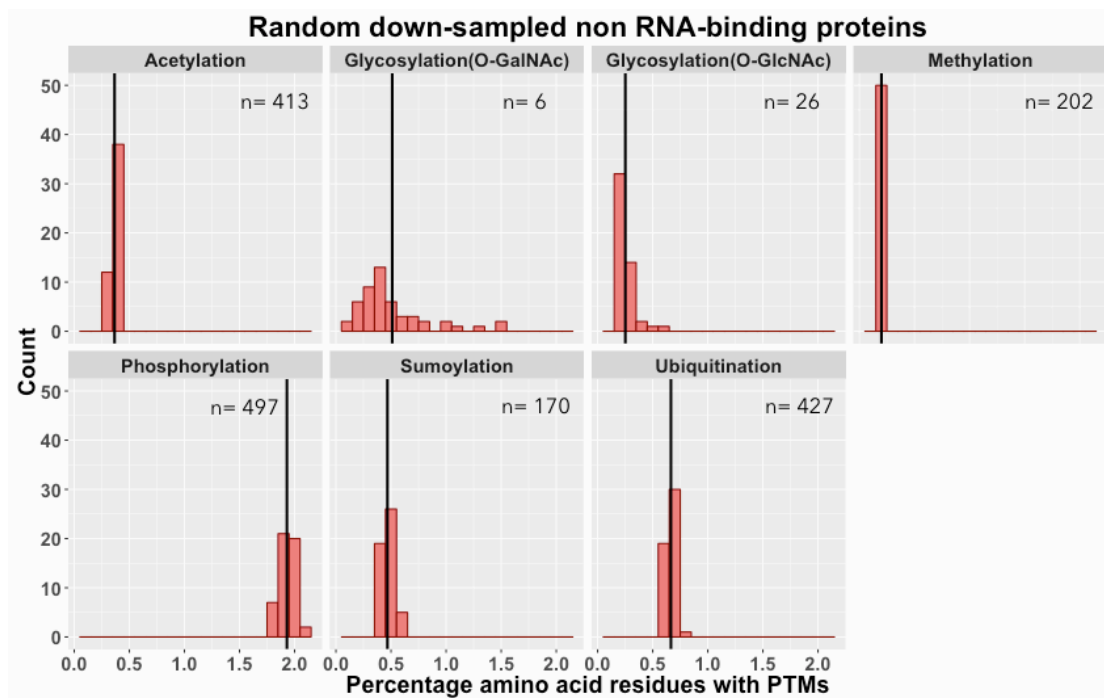
Figure 3.3 Comparison of PTMs between proteins. Percentage of amino acid residues with post-translational modifications in various protein datasets grouped according to their modification types. Each bar represents percentage of the total protein sequence length. Numbers on top of each bar denotes the number of proteins with the particular PTM.

	RNA-binding proteins (RBPs)			Non RNA-binding proteins (non-RBPs)			P-value
	Number of PTMs	Total amino acids	Percentage of amino acids	Number of PTMs	Total amino acids	Percentage of amino acids	
Acetylation	2,617	259,074	1.01%	17,748	4,872,560	0.36%	$< 2.2 \times 10^{-16}$
O-GalNAc	46	4,798	0.95%	2,053	385,642	0.53%	2.2×10^{-02}
O-GlcNAc	53	38,425	0.13%	358	134,159	0.26%	5.8×10^{-04}
Methylation	881	140,022	0.62%	3,811	1,908,462	0.19%	$< 2.2 \times 10^{-16}$
Phosphorylation	14,834	302,895	4.89%	199,253	10298431	1.93%	$< 2.2 \times 10^{-16}$
SUMOylation	626	109,339	0.57%	6,107	1,337,840	0.45%	3.0×10^{-04}
Ubiquitination	4,313	261,589	1.64%	35,718	5,444,070	0.65%	$< 2.2 \times 10^{-16}$

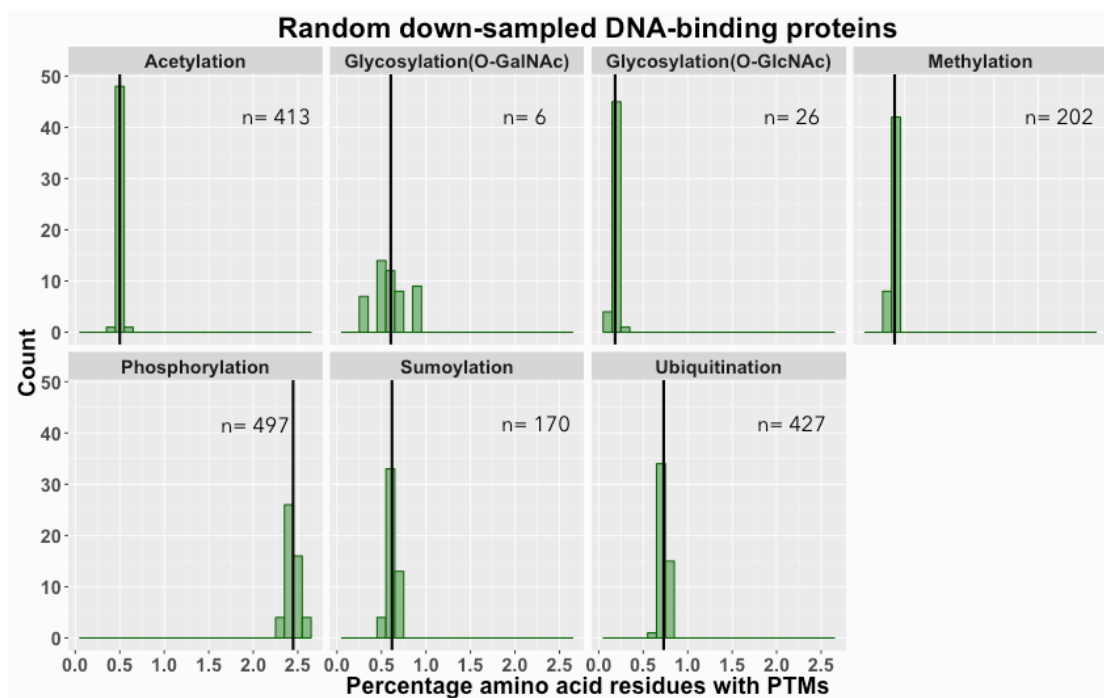
Table 3.3 Breakdown of various PTMs between RBPs and non-RBPs. Details of PTMs in Swiss-Prot RBP, non-RBP, DBP and non-DBP datasets are not shown here.

Some control experiments were performed to confirm any sampling or sequence bias in the different datasets. The number of RBPs and non-RBPs for each PTM type analysed is different; RBPs are fewer in number compared to non-RBPs (Figure 3.3) and this may have introduced a bias in the observed percentage of amino acid residues that are modified in RBPs and non-RBPs. To investigate any such bias introduced by sampling, I randomly down-sampled non-RBPs to match the number of RBPs in their respective PTM groups. This random-down sampling was also performed on the DBP dataset. For example, there are 413 RBPs which are acetylated in comparison to acetylation in 6,376 non-RBPs and 915 DBPs, I randomly down-sampled 413 non-RBPs and DBPs from 6,376 non-RBPs and 915 DBPs and calculated the percentage of amino acid residues that were acetylated in these two datasets. This random down sampling was performed fifty times and the mean value of the percentage of amino acid residues modified was calculated from the distribution. Figure 3.4 shows the distribution of percentage of amino acid residues modified in various PTMs after down sampling. The mean values of percentages calculated were similar to the percentages observed in the full data, which indicates that the observed difference is not due to the differences in the number of proteins in RBP, non-RBP or DBP datasets.

Next, I investigated the relation between protein sequence length and PTM; longer protein sequences have more PTM sites than shorter protein sequences. The number of modified amino acid residues in RBPs and non-RBPs show a positive correlation with the protein sequence length (Figure 3.5A). Comparison of protein sequence lengths show that on average, RBPs, non-RBPs, DBPs and non-DBPs have similar protein sequence lengths of 608.22, 603.79, 615.90 and 601.87 amino acids respectively (Figures 3.5B), which suggests that the numerous modification sites present in RBPs compared to other proteins is not due to differences in protein sequence lengths.

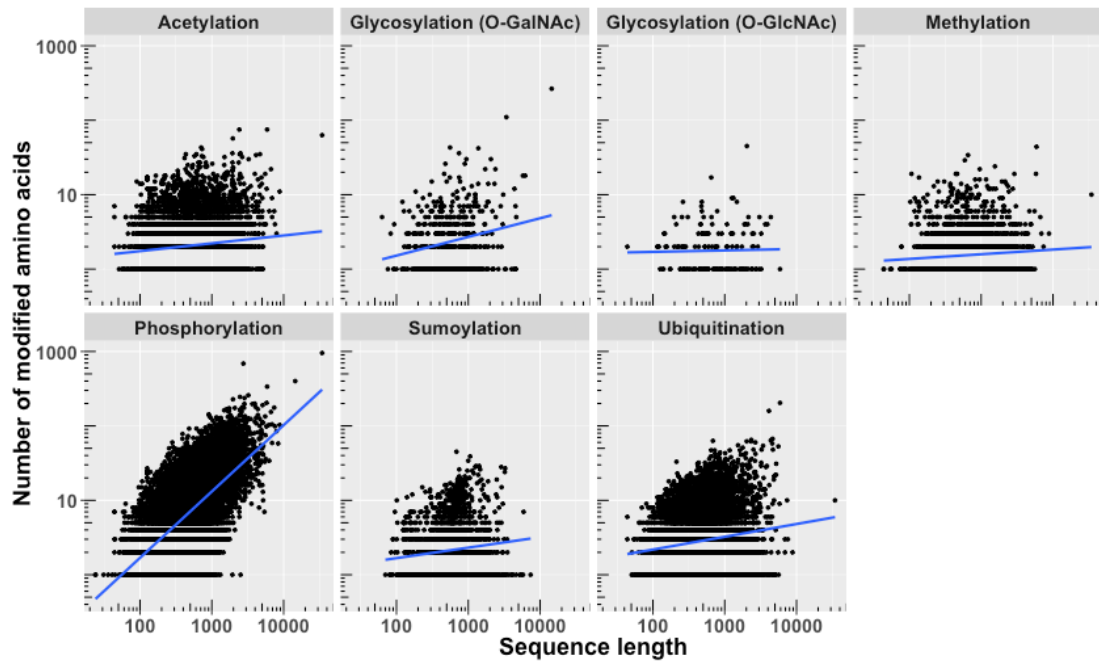


A

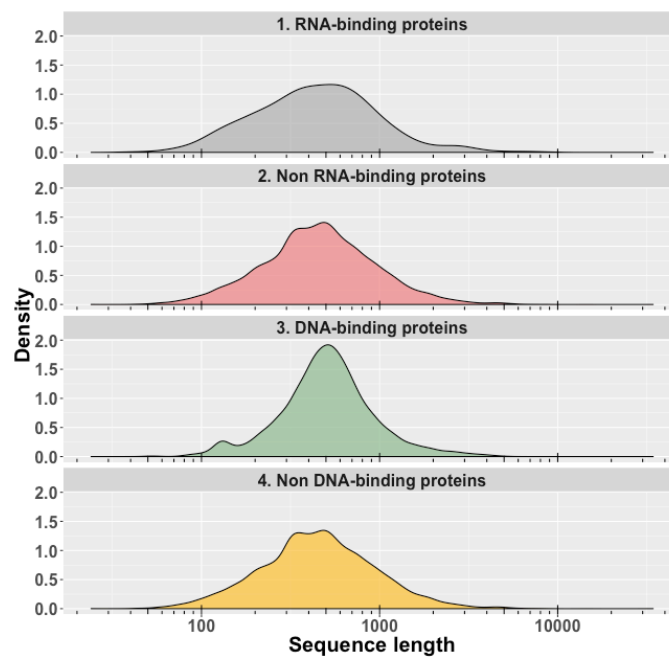


B

Figure 3.4 Distribution of percentage of amino acid residues with post-translational modifications in random down-sampled data. Mean values of distributions for various PTMs in (A) non-RBPs and (B) DBPs are similar to those observed in their respective PTM-types for complete dataset in Figure 3.3. Black lines show mean values, n denotes the number of downsampled proteins.



A



B

Figure 3.5 Relation between PTMs and protein sequence length. (A) Sequence lengths and the number of modified amino acids are positively correlated in RBPs and non-RBPs. (B) The distribution of sequence lengths of RBPs, non-RBPs, DBPs and non-DBPs are similar.

3.3.3 Post-translational modifications in RNA-binding peptides and non RNA-binding peptides

In the previous section comparisons of PTMs between RBPs and non-RBPs have shown that RBPs are modified more than non-RBPs. Next, I investigated the occurrences of PTM sites within RBPs, where I compare residues that are modified in regions of protein that interact with RNA, termed “RNA-binding peptides (RBDpeps)” with residues that are modified elsewhere in the protein that do not interact with RNA, termed “non-RNA binding peptides (non-RBDpeps)”. I observed that more residues are modified in RBDpeps compared to non-RBDpeps; 9.81% of amino acid residues within RBDpeps are modified, which is significantly higher compared to 6.70% of amino acid residues that are modified within non-RBDpeps (P-value < 2.2×10^{-16} , Chi-square test) (Table 3.4). However, the PTM occupancy of amino acid residues in RBDpeps is not significantly different from PTM occupancy of amino acid residues in non-RBDpeps (P-value = 0.13, Chi-square test); a total of 4,097 PTMs were observed in 3,648 modified amino acids residues in RBDpeps (occupancy of 1.12 PTM per modified amino acid residue), in comparison to 19,273 PTMs observed in 17,820 modified amino acid residues in non-RBDpeps (occupancy of 1.08 PTM per modified amino acid residue).

Similar analysis on modified amino acid residues was investigated within the candidate RBDpep dataset (candidate RBPs) by comparing modification sites in candidate RNA-binding peptides (candidate RBDpeps) and candidate non-RNA binding peptides (candidate non-RBDpeps). I observe a similar trend, wherein amino acids in candidate RBDpeps are targeted more frequently than amino acids in candidate non-RBDpeps (P-value < 2.2×10^{-16} , Chi-square test) (Table 3.4).

	Number of PTMs	Number of amino acids with PTM	Total number of amino acids	PTM occupancy (PTM/modified amino acid)	Percentage of amino acid residues modified
RNA-binding peptides (RBDpeps)	4,097	3,648	37,150	1.12	9.81%
Non RNA-binding peptides (non-RBDpeps)	19,273	17,820	265,745	1.08	6.70%
Candidate RNA-binding peptides (candidate RBDpeps)	5,967	5,414	58,536	1.10	9.24%
Candidate non RNA-binding peptides (candidate non-RBDpeps)	31,118	29,251	540,579	1.06	5.41%

Table 3.4 Comparison of post-translational modifications within proteins. Occurrences of PTMs within RNA-binding proteins (RBDpep and non-RBDpep) and candidate RNA-binding proteins (candidate RBDpep and candidate non-RBDpep) are shown. Occurrences of PTMs in non RNA-binding proteins and candidate non RNA-binding proteins were not calculated because of the non-availability of RNA-binding peptide data in these proteins.

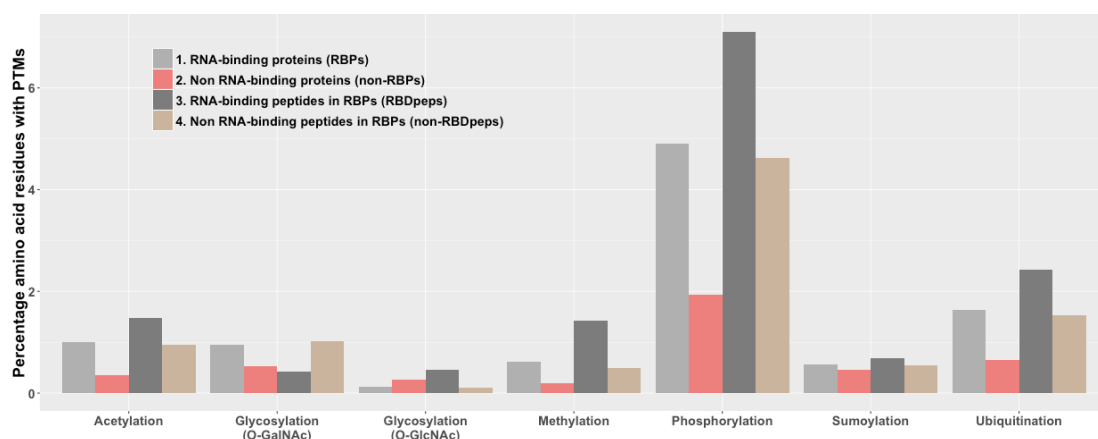
Proteins interact with RNA through the major or minor grooves where surface electrostatic interactions play a major role in recognition, specificity and binding (De Guzman et al., 1998). Modification of residues at the recognition or binding interface is more effective in modulating protein-RNA interaction, since they directly influence the local surface electrostatic potential and interfere with hydrogen-bonding contacts, compared to modifications of residues elsewhere in the protein that may indirectly influence interaction through changes in protein conformation.

Classifying modified amino acid residues in RNA-binding and non-binding peptides by their PTM type show a significant presence of acetylation, methylation, phosphorylation and ubiquitination in RBDpeps (Table 3.5, Figure 3.6A), but no significant difference in the SUMOylation of RBDpeps and non-RBDpeps. It is difficult to comment on the significance of glycosylation due to low statistical power. Acetylation, methylation, phosphorylation and ubiquitination were also seen to be enriched in candidate RBDpeps compared to candidate non-RBDpeps (Figure 3.6B), however SUMOylation is significant in candidate RBDpep in this dataset (P-value = 4.81×10^{-15} , Chi-square test. Data not shown.). Comparison of observed and expected frequencies between RBDpep and non-RBDpep shows that PTMs are observed in RBDpep more than expected by chance.

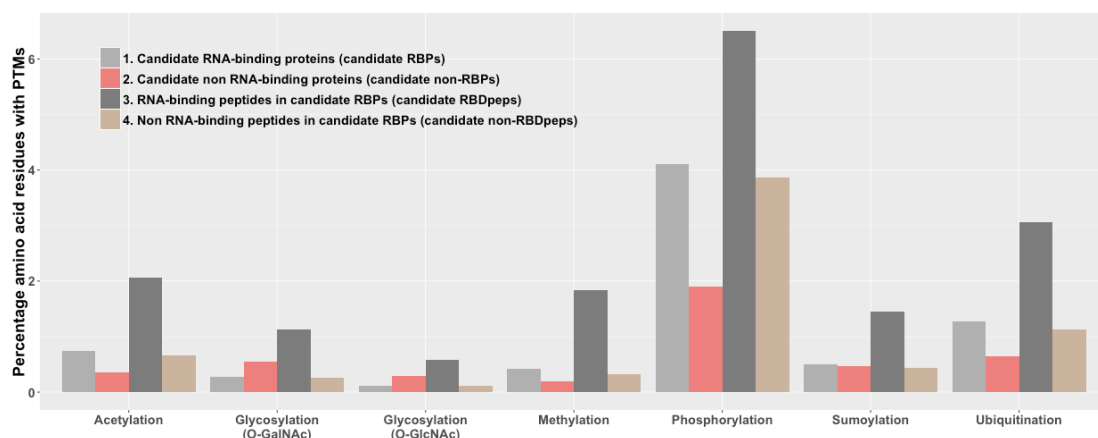
To summarise findings from sections 3.3.2 and 3.3.3, evidences from the above analyses indicate that RNA-binding proteins are relatively more targeted for post-translational modifications than non RNA-binding proteins. And within RNA-binding proteins, regions that bind to RNA are enriched in post-translational modifications compared to those regions that do not.

	RNA-binding peptides (RBDpeps)			Non RNA-binding peptides (non-RBDpeps)			P-value
	Number of PTMs	Total amino acids	Percentage of amino acids	Number of PTMs	Total amino acids	Percentage of amino acids	
Acetylation	475	32,236	1.47%	2,142	226,838	0.94%	1.2×10^{-08}
O-GalNAc	2	478	0.41%	44	4,320	1.01%	0.5*
O-GlcNAc	12	2,599	0.46%	41	35,826	0.11%	0.4×10^{-01}
Methylation	304	21,411	1.41%	577	118,611	0.48%	$< 2.2 \times 10^{-16}$
Phosphorylation	2,370	33,353	7.10%	12,464	269,542	4.63%	$< 2.2 \times 10^{-16}$
SUMOylation	113	16,609	0.68%	513	92,730	0.55%	0.2
Ubiquitination	821	33,811	2.42%	3,492	227,778	1.53%	1.2×10^{-14}

Table 3.5 Breakdown of various PTMs between RBDpeps and non-RBDpeps within RNA-binding proteins. Details of PTMs in candidate RBDpep and candidate non-RBDpep datasets are not shown here. * indicates low statistical power for computing p-value.



A



B

Figure 3.6 Comparisons of PTMs within proteins. PTMs are compared between (A) RNA-binding peptides (dark gray) and non RNA-binding peptides (brown) within RBP and (B) candidate RNA-binding peptides (dark gray) and candidate non RNA-binding peptides (brown) within candidate RBP. Comparison of PTMs between RNA-binding proteins (gray) and non RNA-binding proteins (red) is included for reference. Gray and red bars denote the percentage of their total sequence lengths. Dark gray and brown bars denote the percentage of their total RBDpep length and non-RBDpep lengths respectively.

3.3.4 Disorderedness in RNA-binding proteins

Most eukaryotic proteins are composed of both globular domains and disordered regions; nearly 22% of the human proteome, for instance, is predicted to contain disordered regions with a length of 50 amino acid residues or more (Ward et al., 2004). Disordered regions play an important role in molecular recognition, molecular assembly, modulating specificity or affinity for ligand binding, activation by cleavage, assisted protein folding and protein modification (Dunker et al., 2002). Interestingly PTMs have been shown to occur more in the disordered regions (Kurotani et al., 2014); and disordered regions have an intrinsic RNA-binding activity (Calabretta and Richard, 2015). Therefore, in order to understand if the observed enrichment of PTMs in RBPs is due to the presence of more disordered regions, I investigated the level of disorder in RBPs and non-RBPs and among RBDpeps and non-RBDpeps. Firstly, I observe that RBPs are significantly more disordered (37.44% of total amino acid residues in RBPs) when compared to non-RBPs (25.97% of total amino acid residues in non-RBPs) (P-value < 2.2×10^{-16} , Chi-square test), as also observed by (Varadi et al., 2015). Next, I grouped modified and unmodified sites based on their presence within disordered or globular regions of RBPs and non-RBPs (Figure 3.7 A, B). Comparison of the disordered regions between the two sets of proteins shows that 9.70% of amino acid residues within the disordered regions of RBPs are modified, which is significant compared to just 4.07% of amino acid residues that are modified within the disordered regions of non-RBPs (P-value < 2.2×10^{-16} , Chi-square test) (Figure 3.7 A, B). Similarly, a significant fraction of amino acid residues within the globular regions of RBPs are modified (5.52%) than amino acids within the globular regions of non-RBPs (1.91%). Thus indicating that the observed effect of enriched PTMs in RBPs is not due to the presence of more disordered regions in them.

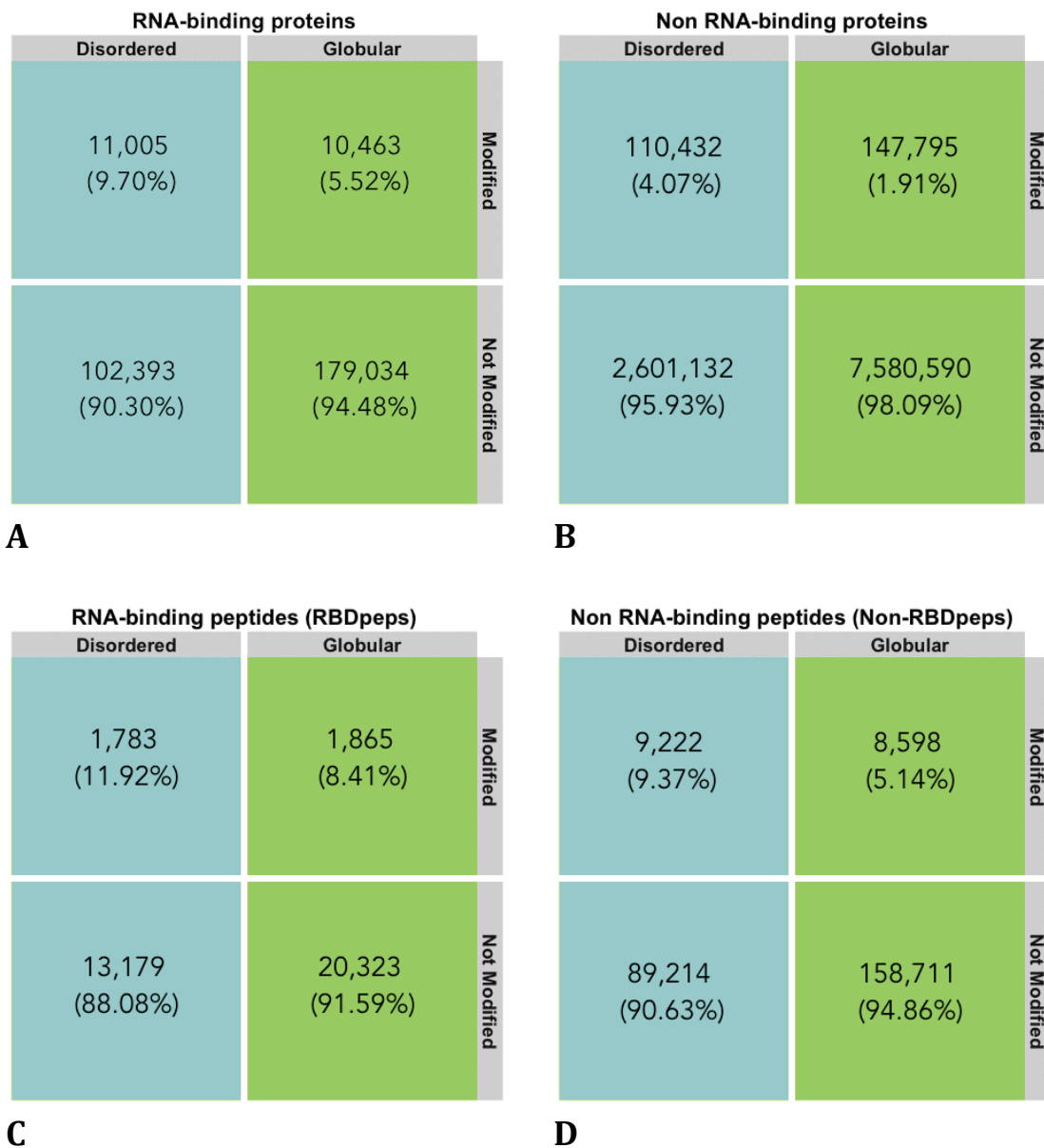


Figure 3.7 Comparison of globular and disorderedness in RNA-binding proteins. (A) Disordered regions in RBPs are significantly enriched in post-translational modification sites than non-RBPs (B). (C) Disordered regions within RBDpeps are significantly enriched with modified amino acid residues than non-RBDpeps (D). Percentages shown are calculated from the total amino acid residues in disordered and globular regions.

Further inspection of disorder within RBDpeps and non-RBDpeps shows that RBDpeps are more disordered (40.27% of amino acid residues within RBDpeps), which is significantly higher than non-RBDpeps (37.04% of amino acid residues within non-RBDpeps) (P-value = 3.30×10^{-11}). And finally, I grouped the modified and non-modified amino acid residues based on their presence within the disordered or globular regions of RBDpeps and non-RBDpeps (Figure 3.7 C, D). Among disordered regions I observe that 11.92% of amino acid residues are modified in RBDpeps, compared to 9.37% of amino acid residues that are modified in non-RBDpeps (P-value = 3.35×10^{-10}) (Figure 3.7 C, D).

Many disordered regions are known to bind RNA through short linear motifs (SLiMs) or low complexity sequences (Calabretta and Richard, 2015). PTMs may also induce disordered-to-ordered transition enabling RNA-binding. The presence of a higher number of PTM sites in structurally disordered regions of RBPs, in contrast to non-RBPs, could suggest that disordered regions might play an important role in regulating RNA-binding as also shown by other studies (Calabretta and Richard, 2015).

3.3.5 Functional classification of RNA-binding proteins

To understand the distribution of PTMs across diverse RBP functions, RBPs and non-RBPs were grouped into three broad functional categories – i) information storage and processing, ii) cellular processes and signalling and iii) metabolism, with an additional sub-classification as described in the EggNOG database (Huerta-Cepas et al., 2016). Functional classification of RBPs show that apart from their traditional functions such as transcription, translation, RNA processing, recombination and repair, they perform diverse functions including signal transduction, energy production, carbohydrate transport and metabolism among others (Figure 3.8). Interestingly, RBPs are also predominantly involved in post-translational modifications of other proteins. PTMs such as acetylation, phosphorylation and ubiquitination are uniformly present among different functional classes, but methylation is observed mainly in RBPs whose main functions involve transcription, RNA processing, transcription and maintaining

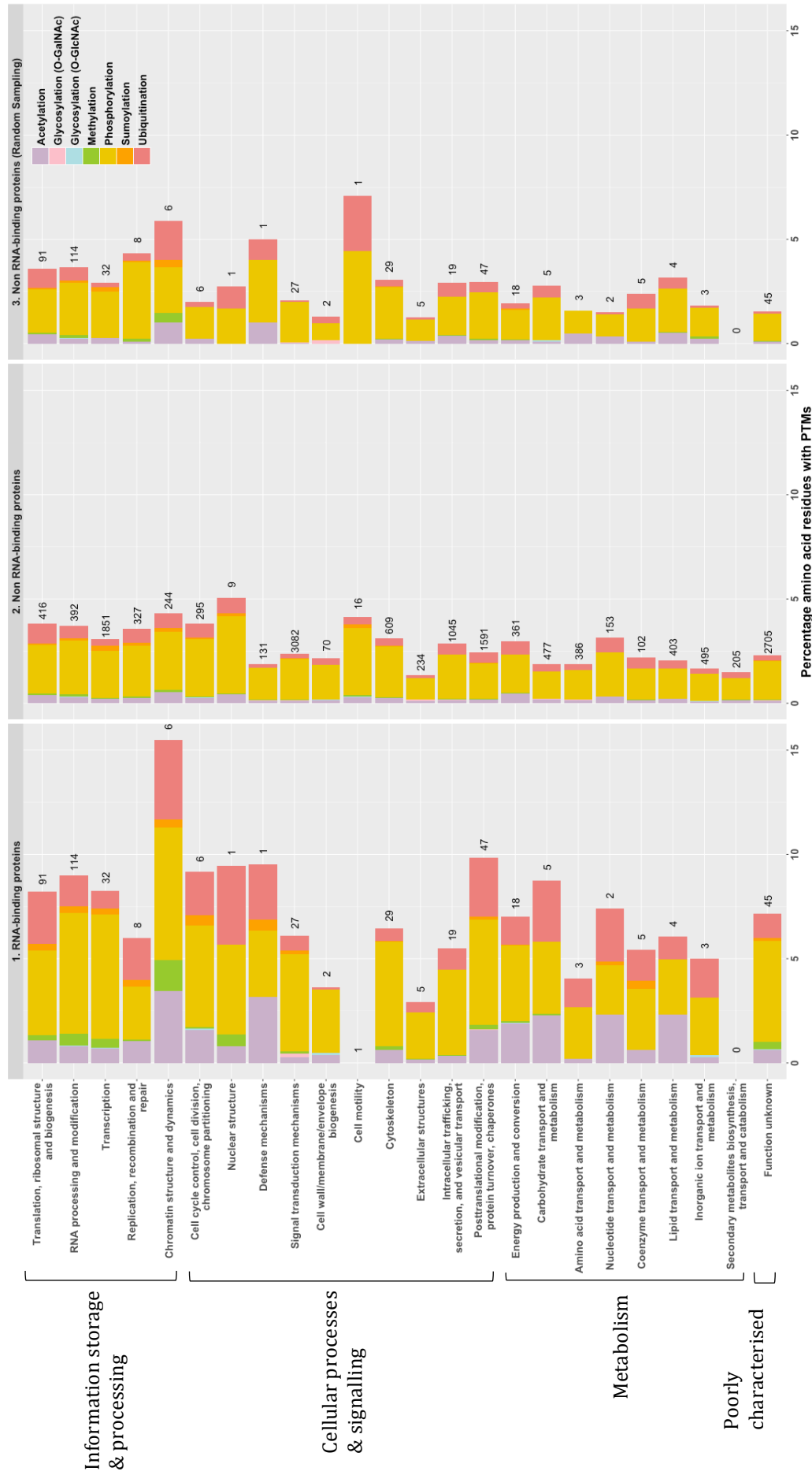
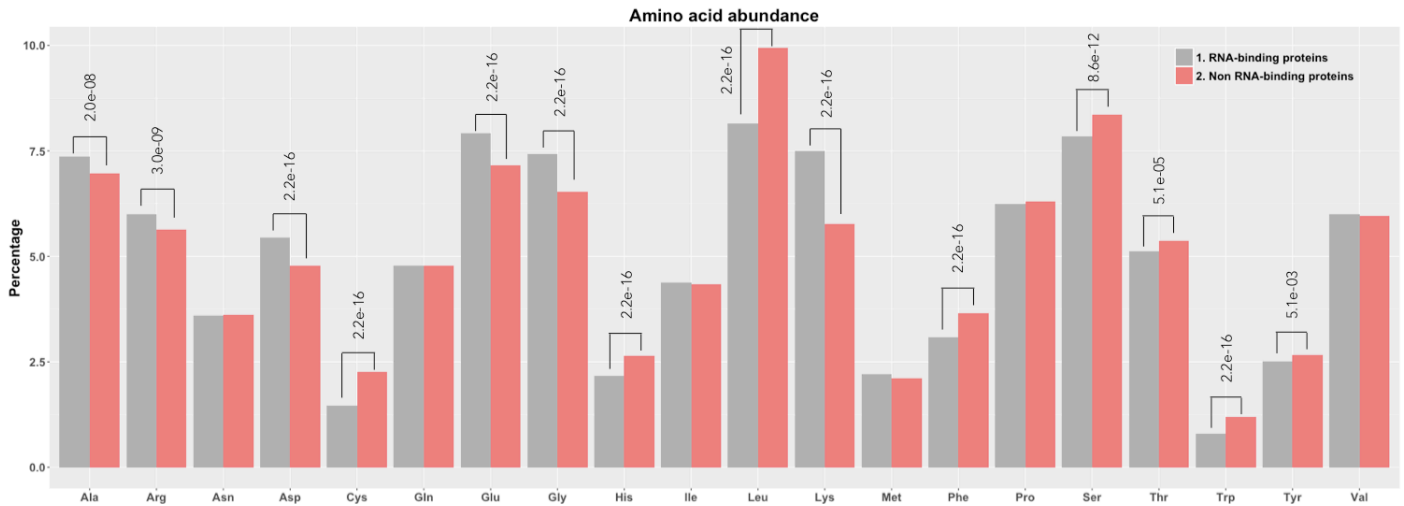


Figure 3.8 Functional classifications of RBPs and non-RBPs. RBPs are also involved in metabolic functions, most PTMs show an equal distribution across different functional classes. The numbers on top of each bar represents number of proteins in that functional class. Cell motility class outlier in RBPs is not shown. Random down-sampled non-RBPs show similar distribution of PTMs across various functional classes.

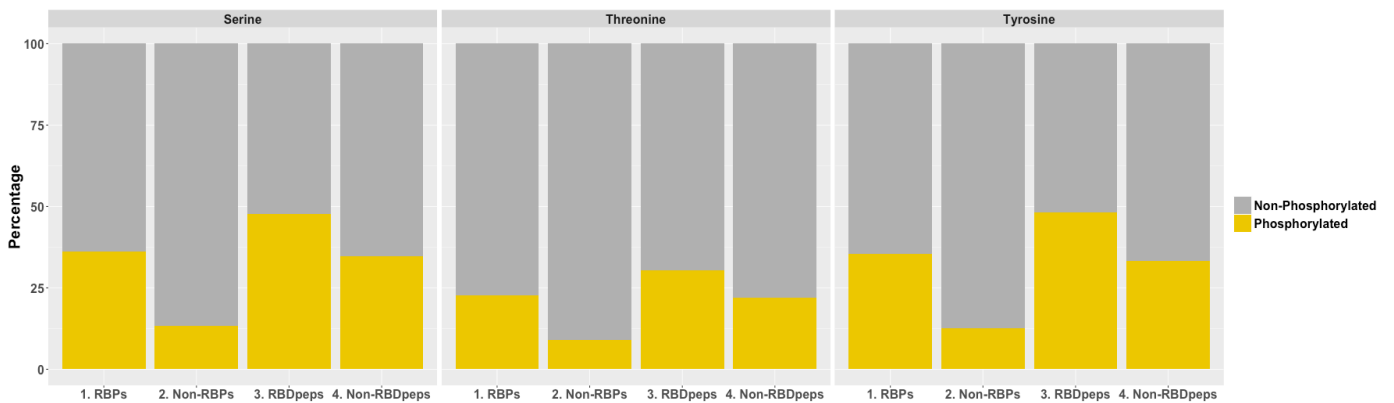
chromatin structure and dynamics. To investigate any sampling bias between RBPs and non-RBPs, I have randomly down-sampled non-RBPs from each respective functional class to match the number of RBPs. The percentage of modified amino acids and the distribution of PTMs within the random down-sampled non-RBPs indicate that the distribution is not due to any sampling bias (Figure 3.8).

3.3.6 Amino acid abundance

From the above analyses it is observed that phosphorylation, ubiquitination and acetylation are overrepresented in RBPs. These modifications are predominantly targeted at particular amino acids such as serine, threonine, tyrosine, arginine and lysine and it is therefore possible that the observed enrichment in PTMs is due to a bias in the sequence composition of RBPs towards these amino acid residues. I compared amino acid abundances in RBPs and non-RBPs to inspect enrichment of these amino acids. I observe a significant enrichment of alanine, arginine, aspartate, glutamate, glycine and lysine amino acid residues in RBPs, but depletion of cysteine, histidine, leucine, phenylalanine, serine, threonine, tryptophan and tyrosine. Presence of amino acids such as asparagine, glutamine, isoleucine, methionine, proline and valine remain unchanged in RBPs and non-RBPs (Figure 3.9A). Charged amino acid residues such as arginine, lysine, aspartate and glutamate are preferred at the protein surface where the surface electrostatic potential is conducive for electrostatic interactions with RNA. A large majority of protein-RNA interactions involve the highly negatively charged RNA sugar-phosphate backbone (Ellis et al., 2007). Charged amino acids such as lysine, arginine, aspartate and glutamate frequently interact with RNA through backbone interactions, interactions at the major and minor grooves or stacking with nucleotide bases (Ananth et al., 2013; Morozova et al., 2006) and are observed in RBPs more than expected by chance. Aromatic amino acids such as phenylalanine, tyrosine and tryptophan form stacking interactions with nucleosides (Morozova et al., 2006), but are less common (Figure 3.9A). Although phosphorylation is the most prevalent PTM in RBPs, it is



A



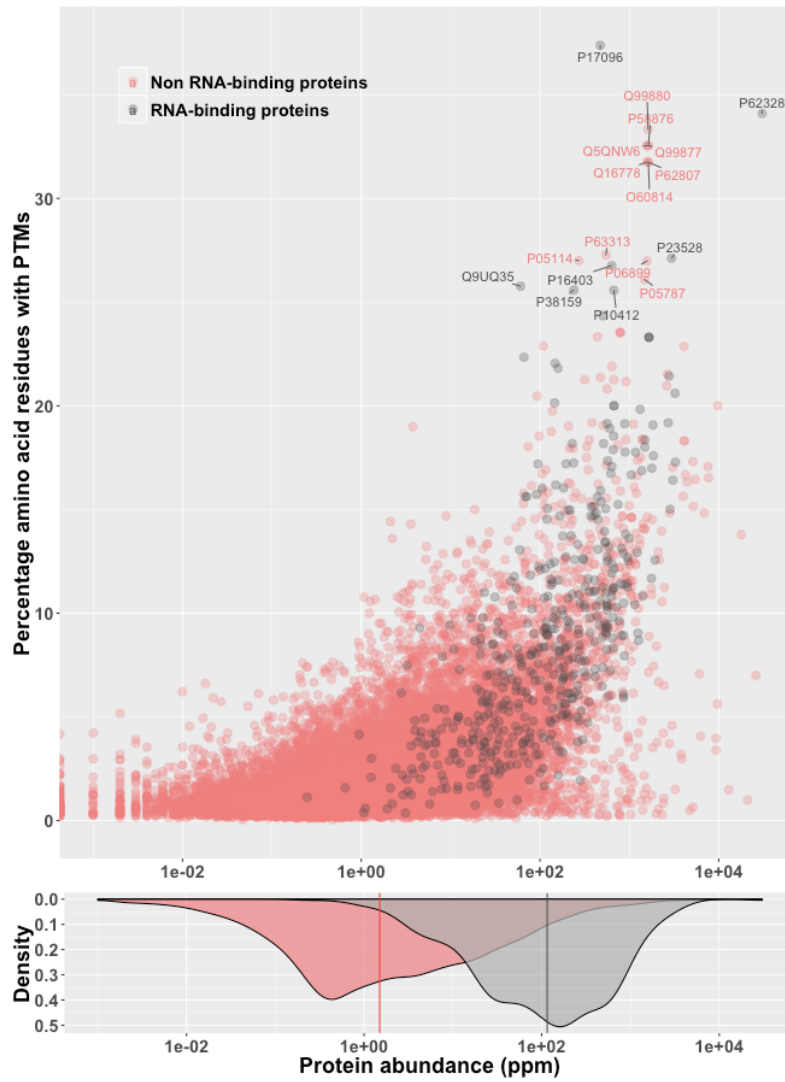
B

Figure 3.9 Amino acid abundance in RBPs and non-RBPs. (A) Amino acids occurrence within RBPs as a percentage of their total sequence lengths. Glycine and charged amino acids such as lysine, arginine, glycine, glutamate are observed significantly more within RBPs than expected by chance, similarly cysteine and aromatic amino acids such as phenylalanine, tyrosine and tryptophan are less frequent in RBPs. (B) The fraction of serine, threonine and tyrosine residues phosphorylated in RBPs is significantly more compared to non-RBPs and similarly they are targeted more in RBDpeps compared to non-RBDpeps.

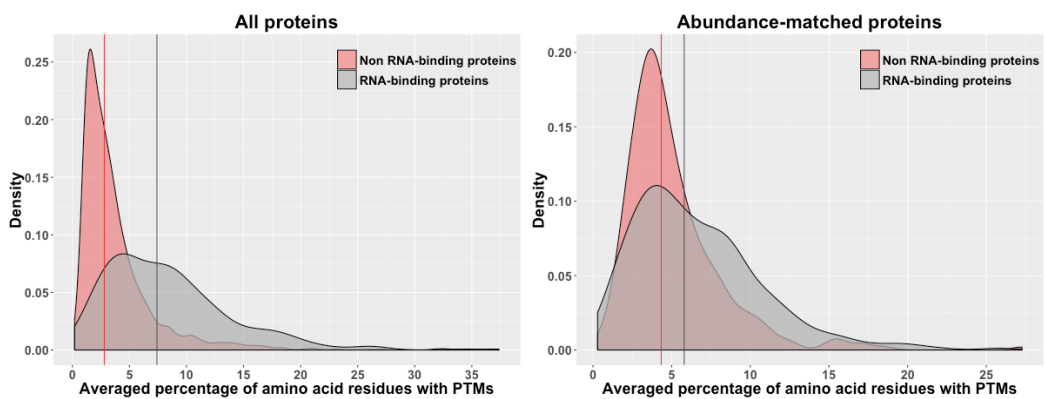
interesting to note that the fraction of residues that are mainly targeted for phosphorylation, such as serine, threonine and tyrosine, is less in RBPs compared to non-RBPs. Comparison of phosphorylated and non-phosphorylated residues shows that a large fraction of serine, threonine and tyrosine residues in non-RBPs are not modified (Figure 3.9B); 36.20% of all serine residues in RBPs are phosphorylated compared to only 13.40% in non-RBPs, similarly 22.81% of threonine and 35.38% of tyrosine residues in RBPs are phosphorylated compared to 8.95% and 12.49% of threonine and tyrosine residues respectively in non-RBPs. This fraction of phosphorylated residues is significantly higher in RBDpeps, for example 47.64% of serine residues within RBDpeps are targets for phosphorylation and similarly for threonine and tyrosine, wherein 30.33% and 48.11% residues within RBDpeps respectively are modified.

3.3.7 Protein abundance

PTMs have been shown to influence stability and cellular abundances of proteins (Elia et al., 2008; Vazquez et al., 2000). The PTM levels are in turn influenced by the abundances of their targeted as well as their modifying proteins such as the 'writer' and 'eraser' enzymes (Beltrao et al., 2013). Since protein abundances and PTMs are linked, I investigated whether the abundances of RBPs and non-RBPs play a role in the observed differences in their PTM levels. It can be seen from Figure 3.10A, that the PTM levels positively correlate with their protein abundances: PTMs occur more in highly abundant proteins compared to the PTM levels found in low abundance proteins in both RBPs and non-RBPs. Comparison of their relative protein abundances also show that RBPs are significantly more abundant in the cell than non-RBPs (Figure 3.10A, bottom panel). Human RBPs are found to be nearly a hundred-fold more abundant (median abundance of RBPs = 115.5 ppm) than non-RBPs (median abundance of non-RBPs = 1.5 ppm). In agreement to the findings, an earlier study by (Gerstberger et al., 2014) observed an abundance of RBP transcripts compared to the non-RBP transcripts across various human tissues. Despite having different cellular abundances both RBPs and non-RBPs show similar trends of PTM levels with respect to their abundances, therefore, in order to check the influence of protein abundances on



A



B

Figure 3.10 Protein abundance and PTMs. (A) PTMs occur more in highly abundant proteins compared to low abundance proteins in both RBPs and non-RBPs. RBPs are also more abundant in the cell than non-RBPs. Protein abundance is measured in ppm (parts per million). (B) Distributions of averaged percentage of amino acid residues with PTMs among all and abundance-matched proteins. Vertical lines indicate the median values.

PTM levels, I compared PTM levels of RBPs and non-RBPs that have the same cellular abundances. Figure 3.10B shows the average percentage of PTM levels of abundance-matched proteins. It is seen that RBPs and non-RBPs with the same abundances have similar distributions of PTM levels. The median of the distribution of averaged percentage PTM levels in RBPs is 5.79, which is similar to the median of the distribution of averaged percentage PTM levels of non-RBPs, which is 4.33. Since the abundance-matched pairs have similar PTM levels, this therefore indicates that much of the difference in the observed PTM levels is due to the differential protein abundances. The comparatively high levels of PTMs observed in RNA-binding proteins is largely due to the high abundances of RBPs within the cell. The shoulder observed in the distribution of RBPs (Figure 3.10B) perhaps indicates a sub-population of RBPs with higher PTM levels.

3.3.8 Structural validation

RNA-binding peptides identified by (Castello et al., 2016) determine RNA-binding sites at a peptide-wide resolution. In order to infer RNA-protein interactions at the amino acid level high-resolution and in the presence of PTMs, I compare RNA-binding regions in proteins identified by (Castello et al., 2016) with their corresponding experimental structures. Using interactions from the available structures of RNA-protein complexes in PDB, I find that nearly a third (32.41%) of the amino acids involved in RNA-protein interactions in structural complexes are also identified by RBDpeps (true-positives), however a majority of the amino acid residues that interact with RNA in structural complexes (67.59%) were not part of RBDpep (false-negatives) (Figure 3.11). Interactions that are not present in the structure but identified by (Castello et al., 2016) (false-positives) were not computed because of unavailability of structural data for those regions. There are two reasons for the low percentage of true-positives: first, shorter RNA molecules in the structure may not cover all the interaction sites on the protein, second, interatomic interactions within the structure were used to infer interacting amino acid residues, which will not include the neighbouring stretch of amino acids that are part of the peptide sequence in RBDpep (Castello et al., 2016).

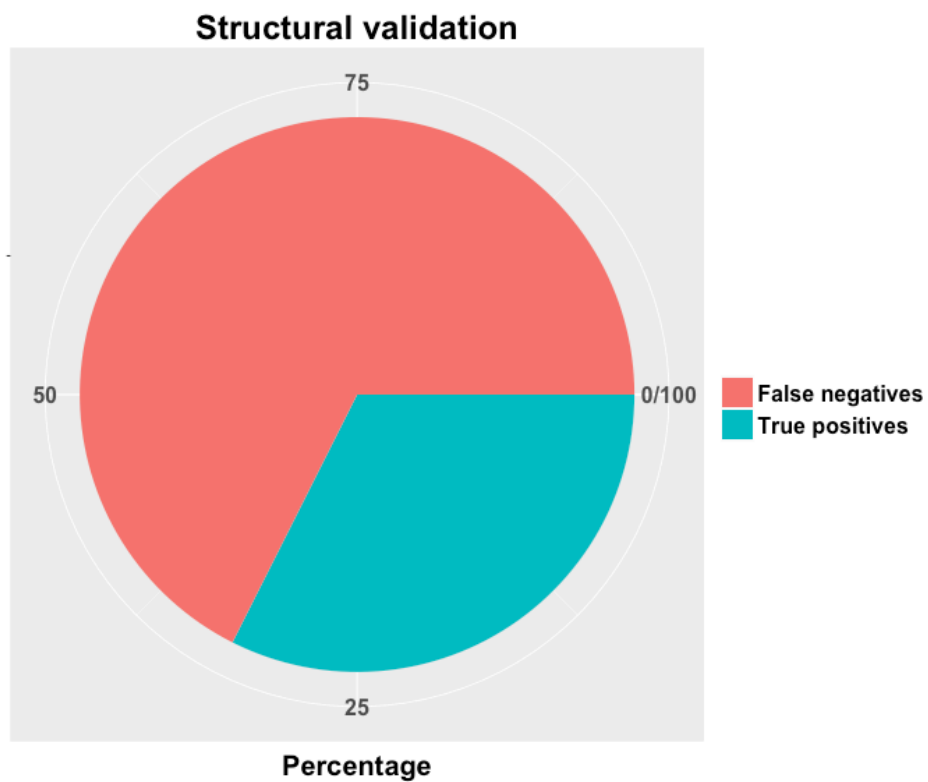


Figure 3.11 Structural validation of RNA-binding peptides. (Castello et al., 2016) identifies 32.41% amino acid residues that interact with RNA in RNA-protein complexes (true-positives), but does not identify 67.59% residues that show interactions in RNA-protein complexes (false-negatives).

3.3.9 Regulation of RNA-protein interactions mediated by post-translational modifications

Results from the previous sections have clearly shown a significant presence of PTMs in RBPs, and they allude to a prominent role of PTMs in the functions of these RBPs. A large amount of literature evidence documents the influence of PTMs in regulating RNA-protein interactions and ultimately the functions of RBPs. PTMs affect RNA-protein interactions in diverse ways; one of the common modes of regulation is through directly influencing interactions through their presence at the RNA-binding sites. The RNA-protein interactions are largely mediated by surface electrostatics, which involve hydrogen bond formation and van der Waals interactions. Charged moieties such as phosphate and acetyl groups at the binding sites bring about local changes in the electrostatic potential, which can disrupt the hydrogen bond donor-acceptor complementarity.

Phosphorylation of a single amino acid residue within the RNA-binding site has been shown to affect affinity for RNA. Phosphorylation of Ser46 in the rubella virus capsid negatively regulates RNA-binding by decreasing its affinity (Law et al., 2003). This phosphorylation mediated decrease in affinity for RNA binding has been suggested to prevent nonspecific binding of cellular RNA and/or premature assembly of nuclear capsids, while dephosphorylation at the latter stages of virus assembly mediates increase affinity for RNA for efficient nucleocapsid assembly (Law et al., 2003). Insights into the influence of PTMs on interactions are provided by studies employing targeted modifications such as phosphomimetic substitutions, which use amino acid substitutions that mimic phosphorylated proteins. A phosphomimetic replacement of Ser318 by Asp318 within the RNA Recognition Motif 3 (RRM3) was used to infer the mechanism of RNA recognition by the Human antigen R (HuR) protein (Scheiba et al., 2014). The mutated residue does not influence the protein structure, but the affinity to bind RNA was slightly lower, which was suggested to be due to the electrostatic repulsion effect between Asp318 and the RNA backbone (Scheiba et al., 2014). Interestingly a similar phosphomimetic substitution of Ser138 by Asp138 in the

RRM3 of HuR protein clearly showed an enhanced affinity to type III AU-rich element (ARE) mRNAs but reduced affinity to type I and type II AREs (Schulz et al., 2013), suggesting influence by other factors. Changes in local surface electrostatics by PTMs are more influential for those interactions that are highly specific and involve conserved interaction sites. An example of such disruption in RNA-protein interaction is observed in the iron regulatory protein 1 (IRP1). The phosphomimetic substitution of an evolutionarily conserved amino acid residue Ser711 by Glu711 completely abolishes binding of IRP1 to mRNA iron-response elements (Fillebeen et al., 2005). This disruption in RNA-protein interaction is attributed directly to changes in local surface electrostatics leading to loss of interaction as the substitution did not alter protein stability nor induce misfolding (Fillebeen et al., 2005). Interestingly, the phosphomimetic substitution of Ser711 by Glu711 also had a detrimental effect on the catalytic activity of IRP1, wherein its aconitase activity was severely impaired. The phosphomimetic mutant of IRP1 displayed minimal capacity to generate the intermediate *cis*-aconitate from citrate and was only partially able to convert *cis*-aconitate to isocitrate (Fillebeen et al., 2005), probably because of its proximity to one of the active site residues Arg713 (Walden et al., 2006) (Figure 3.12A,B), indicating a broader influence of PTMs on protein functions. However, not all PTMs at the binding sites are detrimental to RNA-protein interactions. Acetylation of Sam68, a member of the STAR family of KH domain containing RNA-binding proteins, is shown to enhance RNA-binding (Babic et al., 2004). The acetylation of lysine residues within the RNA-binding region of highly conserved GSH domain by acetyltransferase CBP positively regulates association of Sam68 with the poly(U) RNA substrate and this enhanced association is suggested to play a role in tumor cell proliferation (Babic et al., 2004).

Apart from directly regulating interactions at the RNA-binding sites, PTMs can also indirectly regulate RNA-binding through changes in protein conformation, or affect oligomerisation by switching affinity towards proteins. For example methylation of the nuclear poly(A) binding protein PABPN1, favours RNA-binding but weakens affinity of PABPN1 towards the nuclear import receptor protein transportin (Fronz et al., 2011). RNA and transportin compete for

binding to PABPN1 and is regulated by its methylation. However on the contrary phosphorylation of the RNA-binding protein p54(nrb) does not affect its interaction with proteins but selectively diminishes its binding to 5' splice sites, poly(A), poly(U) and poly(C) homopolymers but not to poly(G), non coding RNA Neat1 and PIR-1 RNA (Bruelle et al., 2011). Similarly acetylation of TDP-43, a highly conserved RNA and DNA-binding protein, results in impairing its RNA-binding but in turn promotes protein accumulation that resemble pathological inclusions in amyotrophic lateral sclerosis (Cohen et al., 2015).

A few structural studies have provided insights into the competitive regulation of RNA and protein binding by PTMs by bringing about changes in entropy. The component of the yeast U5 small nuclear ribonucleoprotein (snRNP) Aar2 competes with protein, U5-specific helicase Brr2, and di-snRNA U4/U6 for binding to U5-specific protein Prp8 (Weber et al., 2013). Phosphomimetic mutation of Ser253 to Glu253 in Aar2 results in a ten-fold reduced affinity towards Prp8 compared to wild-type Aar2 and is due to the enthalpic gain/entropic loss resulting from the folding and immobilisation of unstructured region. This change in conformation in Aar2 no longer results in competitive binding to Prp8 but allows both Brr2 and U4/U6 di-snRNA to cooperatively bind to Prp8 (Weber et al., 2013). Similarly entropy changes brought about by N-acetylation of Lys50 in the Tat peptide decreases binding affinity towards HIV-1 TAR RNA (Kumar and Maiti, 2013). Acetylation of Lys50 is essential for Tat to associate with p300/CBP-associated factor (PCAF). Mutation of Lys50 to Arg50, which prevents acetylation of Tat, results in disassociation of Tat from TAR RNA but promotes its association with PCAF (Mujtaba et al., 2002).

Phosphorylation of the histone chaperone B23/nucleophosmin, which binds to ribosomal RNA (rRNA) chromatin and stimulates rRNA transcription, too results in its decreased RNA-binding (Okuwaki et al., 2002). Conserved threonine residues within the intrinsically disordered regions of B23 play a role in RNA-binding and phosphomimetic mutations of these residues result in significantly lower RNA-binding activity, but do not influence disorderedness indicating a

major role of phosphorylation, but not structure, influenced loss of RNA-binding (Hisaoka et al., 2014; Hisaoka et al., 2010).

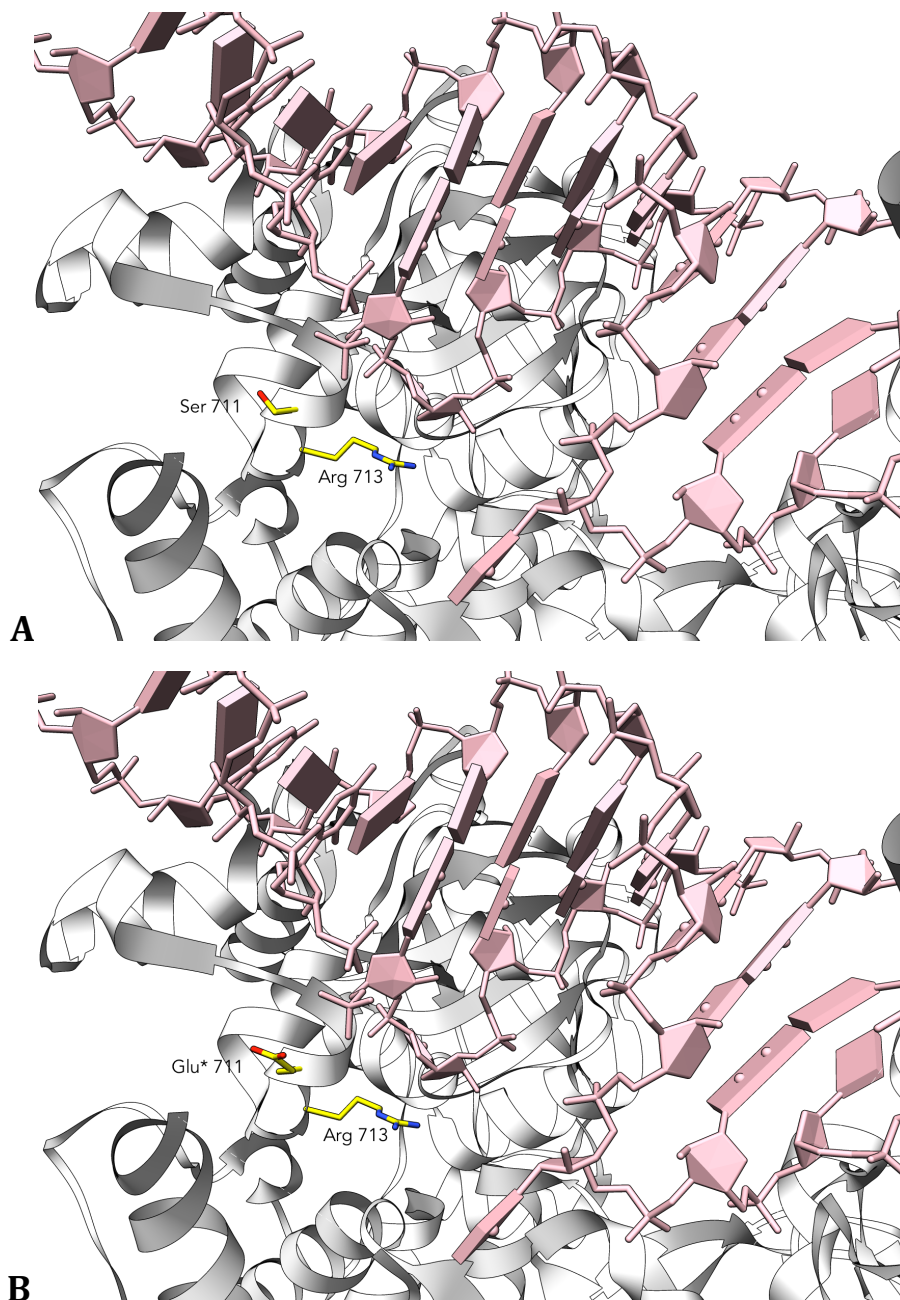


Figure 3.12 Comparison of effects of phosphorylation on the RNA-binding of IRP1. (A) Structure of rabbit IRP1 wild-type protein in complex with frog ferritin H IRE-RNA (pink) (PDB: 3SNP). The *cis*-aconitase active site residue Arg 713 and the conserved Ser 711 are shown as sticks (yellow). (B) The structure is mutated *in-silico* to replace Ser 711 with a phosphomimetic residue Glu 711, which may reduce affinity with the RNA-backbone due to the presence of negative electrostatic potential. *In-silico* mutation was carried out in UCSF Chimera package with Dunbrack rotamer libraries (Dunbrack, 2002).

3.4 Conclusion

In this study I have shown that RNA-binding proteins are enriched in sites for post-translational modifications such as phosphorylation, acetylation, methylation and ubiquitination. The experimentally known RBPs constitute only about 5% of the total human proteome (isoforms not included), nevertheless they show a significant enrichment in PTM sites over other non-RBPs, which suggest that RBPs are a highly regulated class of proteins. Given the dynamic nature of interaction of RNA with proteins of various kinds at different stages of its life cycle (Lunde et al., 2007), I hypothesize that PTMs, as regulatory switches, represent a fine-tuning mechanism that can dynamically influence association or disassociation of proteins with RNA.

PTM sites were found enriched in RNA-binding regions, as also observed by (Castello et al., 2016). The presence of more PTMs sites on regions that interact with RNA could provide a direct means to regulate RNA-protein interactions by changing the local surface electrostatics. It is well known that RBPs comprise disordered regions and are sites for PTMs. Comparison of PTM sites on disordered and globular regions of RBPs and non-RBPs indicates that despite a significant presence of disordered regions within RBPs, there is no enrichment of PTM sites in disordered regions. Amino acid composition has shown that RBPs are enriched with charged RNA-binding residues such as lysine, arginine, aspartate and glutamate, but are also depleted in other amino acids, most importantly, serine, threonine and tyrosine, which are the major targets for phosphorylation. Phosphorylation is significantly enriched in RBPs, but a relatively lower fraction of phosphorylation-targeted amino acids indicate that these amino acids are more frequently phosphorylated than those in non-RBPs. For example, I find that nearly a third of all the serine residues in RBPs are targets for phosphorylation and the chances of them being phosphorylated are even higher if these residues are present within RNA-binding sites. About half of all the serine and tyrosine residues within RBDpeps are sites for phosphorylation.

Although PTMs are enriched in RBPs, the above analyses have indicated that this enrichment is neither due to the presence of more disordered regions in RBPs nor due to the bias in amino acid composition of RBPs. The enrichment of PTMs in RBPs is associated with increased cellular abundance. Although there is a positive correlation between PTM levels and abundances, correlation does not necessarily imply causation; it is not clear whether PTMs lead to abundance or vice versa or both.

RBPs identified by the RBDmap technique (Castello et al., 2016) overlap nearly a third of proteins annotated as RNA-binding in Swiss-Prot (release 2016_10). This partial overlap between the two datasets can be attributed to the experimental design. RBDmap will miss some of the proteins if they are i) bound to non-polyadenylated RNAs, this includes proteins that bind to small RNAs and ribosomal RNAs ii) weakly bound or have low cross-linking efficiency, iii) interact with sugar-phosphate backbone and not the nucleotides or iv) do not have cleavage sites for LysC or ArgC peptidases (Castello et al., 2016). Moreover, the RBPs identified by (Castello et al., 2016) are from a single experiment - UV crosslinking followed by oligodT capture and mass spectroscopy, while the RBPs curated in Swiss-Prot are identified using various experimental techniques and therefore covers a wide range of protein types. Despite the above experimental limitations 72% of RBPs identified by RBDmap are novel and are not annotated as such in Swiss-Prot. As mentioned earlier in the methods section 3.2.2, the number of RBPs not detected by RBDmap and therefore assigned as non-RBPs represents a small fraction (7.92%) of the entire human non-RBP set and may not significantly influence observations. Protein abundances also play an important role in their identification, wherein abundant proteins are readily detected by techniques such as mass spectrometry (Millioni et al., 2011). The authors report that the experimental methods of RBDmap are not selective for identifying highly abundant RBPs within the cell (Castello et al., 2016). By using an independent control datasets of RBPs from Swiss-Prot, any biases in RBP coverage and detection influenced by protein abundances were minimised.

PTMs in RBPs are indicative of its significance in regulation of RNA-protein interactions, however it should also be noted that PTMs do not fully represent the regulatory mechanisms of RNA-protein interactions, post-transcriptional modifications of RNA substrates and other cellular factors that govern the relative concentration of RNA and protein concentrations *in vivo* could also influence RNA-protein interactions. Availability of large-scale post-transcriptional data and high-resolution experimental structures of full-length RNA-protein complexes will be useful in fully understanding the regulatory interactions between RNA and proteins.

3.5 References

Ananth, P., Goldsmith, G., and Yathindra, N. (2013). An innate twist between Crick's wobble and Watson-Crick base pairs. *RNA* 19, 1038-1053.

Anantharaman, V., Koonin, E.V., and Aravind, L. (2002). Comparative genomics and evolution of proteins involved in RNA metabolism. *Nucleic acids research* 30, 1427-1464.

Antson, A.A. (2000). Single-stranded-RNA binding proteins. *Current opinion in structural biology* 10, 87-94.

Arnesen, T. (2011). Towards a functional understanding of protein N-terminal acetylation. *PLoS biology* 9, e1001074.

Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., *et al.* (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature genetics* 25, 25-29.

Babic, I., Jakymiw, A., and Fujita, D.J. (2004). The RNA binding protein Sam68 is acetylated in tumor cell lines, and its acetylation correlates with enhanced RNA binding activity. *Oncogene* 23, 3781-3789.

Baltz, A.G., Munschauer, M., Schwanhauser, B., Vasile, A., Murakawa, Y., Schueler, M., Youngs, N., Penfold-Brown, D., Drew, K., Milek, M., *et al.* (2012). The mRNA-bound proteome and its global occupancy profile on protein-coding transcripts. *Molecular cell* 46, 674-690.

Bedford, M.T., and Clarke, S.G. (2009). Protein arginine methylation in mammals: who, what, and why. *Molecular cell* 33, 1-13.

Beltrao, P., Bork, P., Krogan, N.J., and van Noort, V. (2013). Evolution and functional cross-talk of protein post-translational modifications. *Molecular systems biology* 9, 714.

Bhandari, D., Guha, K., Bhaduri, N., and Saha, P. (2011). Ubiquitination of mRNA cycling sequence binding protein from *Leishmania donovani* (LdCSBP) modulates the RNA endonuclease activity of its Smr domain. *FEBS letters* 585, 809-813.

Blackwell, E., and Ceman, S. (2012). Arginine methylation of RNA-binding proteins regulates cell function and differentiation. *Molecular reproduction and development* 79, 163-175.

Brauer, U., Zaharieva, E., and Soller, M. (2014). Regulation of ELAV/Hu RNA-binding proteins by phosphorylation. *Biochemical Society transactions* 42, 1147-1151.

Brown, A.S., Mohanty, B.K., and Howe, P.H. (2015). Computational Identification of Post Translational Modification Regulated RNA Binding Protein Motifs. *PloS one* 10, e0137696.

Bruelle, C., Bedard, M., Blier, S., Gauthier, M., Traish, A.M., and Vincent, M. (2011). The mitotic phosphorylation of p54(nrb) modulates its RNA binding activity. *Biochemistry and cell biology = Biochimie et biologie cellulaire* 89, 423-433.

- Butt, A.M., Khan, I.B., Hussain, M., Idress, M., Lu, J., and Tong, Y. (2012). Role of post translational modifications and novel crosstalk between phosphorylation and O-beta-GlcNAc modifications in human claudin-1, -3 and -4. *Molecular biology reports* 39, 1359-1369.
- Calabretta, S., and Richard, S. (2015). Emerging Roles of Disordered Sequences in RNA-Binding Proteins. *Trends in biochemical sciences* 40, 662-672.
- Castello, A., Fischer, B., Eichelbaum, K., Horos, R., Beckmann, B.M., Strein, C., Davey, N.E., Humphreys, D.T., Preiss, T., Steinmetz, L.M., *et al.* (2012). Insights into RNA biology from an atlas of mammalian mRNA-binding proteins. *Cell* 149, 1393-1406.
- Castello, A., Fischer, B., Frese, C.K., Horos, R., Alleaume, A.M., Foehr, S., Curk, T., Krijgsveld, J., and Hentze, M.W. (2016). Comprehensive Identification of RNA-Binding Domains in Human Cells. *Molecular cell* 63, 696-710.
- Castello, A., Hentze, M.W., and Preiss, T. (2015). Metabolic Enzymes Enjoying New Partnerships as RNA-Binding Proteins. *Trends in endocrinology and metabolism: TEM* 26, 746-757.
- Castello, A., Horos, R., Strein, C., Fischer, B., Eichelbaum, K., Steinmetz, L.M., Krijgsveld, J., and Hentze, M.W. (2013). System-wide identification of RNA-binding proteins by interactome capture. *Nature protocols* 8, 491-500.
- Chavez, J.D., Weisbrod, C.R., Zheng, C., Eng, J.K., and Bruce, J.E. (2013). Protein interactions, post-translational modifications and topologies in human cells. *Molecular & cellular proteomics : MCP* 12, 1451-1467.
- Chen, Y.C., Sargsyan, K., Wright, J.D., Huang, Y.S., and Lim, C. (2014). Identifying RNA-binding residues based on evolutionary conserved structural and energetic features. *Nucleic acids research* 42, e15.
- Choudhary, C., Kumar, C., Gnad, F., Nielsen, M.L., Rehman, M., Walther, T.C., Olsen, J.V., and Mann, M. (2009). Lysine acetylation targets protein complexes and co-regulates major cellular functions. *Science* 325, 834-840.
- Cohen, T.J., Hwang, A.W., Restrepo, C.R., Yuan, C.X., Trojanowski, J.Q., and Lee, V.M. (2015). An acetylation switch controls TDP-43 function and aggregation propensity. *Nature communications* 6, 5845.
- Cook, K.B., Kazan, H., Zuberi, K., Morris, Q., and Hughes, T.R. (2011). RBPDB: a database of RNA-binding specificities. *Nucleic acids research* 39, D301-308.
- Darnell, R.B. (2010). HITS-CLIP: panoramic views of protein-RNA regulation in living cells. *Wiley interdisciplinary reviews. RNA* 1, 266-286.
- De Guzman, R.N., Turner, R.B., and Summers, M.F. (1998). Protein-RNA recognition. *Biopolymers* 48, 181-195.
- Dosztanyi, Z., Csizmok, V., Tompa, P., and Simon, I. (2005). IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content. *Bioinformatics* 21, 3433-3434.
- Drazic, A., Myklebust, L.M., Ree, R., and Arnesen, T. (2016). The world of protein acetylation. *Biochimica et biophysica acta* 1864, 1372-1401.

- Duan, G., and Walther, D. (2015). The roles of post-translational modifications in the context of protein interaction networks. *PLoS computational biology* *11*, e1004049.
- Dunbrack, R.L., Jr. (2002). Rotamer libraries in the 21st century. *Current opinion in structural biology* *12*, 431-440.
- Dunker, A.K., Brown, C.J., Lawson, J.D., Iakoucheva, L.M., and Obradovic, Z. (2002). Intrinsic disorder and protein function. *Biochemistry* *41*, 6573-6582.
- Elia, A., Constantinou, C., and Clemens, M.J. (2008). Effects of protein phosphorylation on ubiquitination and stability of the translational inhibitor protein 4E-BP1. *Oncogene* *27*, 811-822.
- Ellis, J.J., Broom, M., and Jones, S. (2007). Protein-RNA interactions: structural analysis and functional classes. *Proteins* *66*, 903-911.
- Fillebeen, C., Caltagirone, A., Martelli, A., Moulis, J.M., and Pantopoulos, K. (2005). IRP1 Ser-711 is a phosphorylation site, critical for regulation of RNA-binding and aconitase activities. *The Biochemical journal* *388*, 143-150.
- Fronz, K., Guttinger, S., Burkert, K., Kuhn, U., Stohr, N., Schierhorn, A., and Wahle, E. (2011). Arginine methylation of the nuclear poly(a) binding protein weakens the interaction with its nuclear import receptor, transportin. *The Journal of biological chemistry* *286*, 32986-32994.
- Fukuda, H., Sano, N., Muto, S., and Horikoshi, M. (2006). Simple histone acetylation plays a complex role in the regulation of gene expression. *Briefings in functional genomics & proteomics* *5*, 190-208.
- Gerstberger, S., Hafner, M., and Tuschl, T. (2014). A census of human RNA-binding proteins. *Nature reviews. Genetics* *15*, 829-845.
- Gnad, F., Gunawardena, J., and Mann, M. (2011). PHOSIDA 2011: the posttranslational modification database. *Nucleic acids research* *39*, D253-260.
- Hafner, M., Landthaler, M., Burger, L., Khorshid, M., Hausser, J., Berninger, P., Rothballer, A., Ascano, M., Jr., Jungkamp, A.C., Munschauer, M., *et al.* (2010). Transcriptome-wide identification of RNA-binding protein and microRNA target sites by PAR-CLIP. *Cell* *141*, 129-141.
- Hisaoka, M., Nagata, K., and Okuwaki, M. (2014). Intrinsically disordered regions of nucleophosmin/B23 regulate its RNA binding activity through their inter- and intra-molecular association. *Nucleic acids research* *42*, 1180-1195.
- Hisaoka, M., Ueshima, S., Murano, K., Nagata, K., and Okuwaki, M. (2010). Regulation of nucleolar chromatin by B23/nucleophosmin jointly depends upon its RNA binding activity and transcription factor UBF. *Molecular and cellular biology* *30*, 4952-4964.
- Hornbeck, P.V., Zhang, B., Murray, B., Kornhauser, J.M., Latham, V., and Skrzypek, E. (2015). PhosphoSitePlus, 2014: mutations, PTMs and recalibrations. *Nucleic acids research* *43*, D512-520.
- Huang, K.Y., Su, M.G., Kao, H.J., Hsieh, Y.C., Jhong, J.H., Cheng, K.H., Huang, H.D., and Lee, T.Y. (2016). dbPTM 2016: 10-year anniversary of a resource for post-translational modification of proteins. *Nucleic acids research* *44*, D435-446.

Huerta-Cepas, J., Szklarczyk, D., Forslund, K., Cook, H., Heller, D., Walter, M.C., Rattei, T., Mende, D.R., Sunagawa, S., Kuhn, M., *et al.* (2016). eggNOG 4.5: a hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences. *Nucleic acids research* *44*, D286-293.

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

Jurica, M.S., and Moore, M.J. (2003). Pre-mRNA splicing: awash in a sea of proteins. *Molecular cell* *12*, 5-14.

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

Konig, J., Zarnack, K., Luscombe, N.M., and Ule, J. (2012). Protein-RNA interactions: new genomic technologies and perspectives. *Nature reviews. Genetics* *13*, 77-83.

Kota, V., Sommer, G., Durette, C., Thibault, P., van Niekerk, E.A., Twiss, J.L., and Heise, T. (2016). SUMO-Modification of the La Protein Facilitates Binding to mRNA In Vitro and in Cells. *PloS one* *11*, e0156365.

Kumar, S., and Maiti, S. (2013). The effect of N-acetylation and N-methylation of lysine residue of Tat peptide on its interaction with HIV-1 TAR RNA. *PloS one* *8*, e77595.

Kurotani, A., Tokmakov, A.A., Kuroda, Y., Fukami, Y., Shinozaki, K., and Sakurai, T. (2014). Correlations between predicted protein disorder and post-translational modifications in plants. *Bioinformatics* *30*, 1095-1103.

Law, L.M., Everitt, J.C., Beatch, M.D., Holmes, C.F., and Hobman, T.C. (2003). Phosphorylation of rubella virus capsid regulates its RNA binding activity and virus replication. *Journal of virology* *77*, 1764-1771.

Lee, D.Y., Teyssier, C., Strahl, B.D., and Stallcup, M.R. (2005). Role of protein methylation in regulation of transcription. *Endocrine reviews* *26*, 147-170.

Lunde, B.M., Moore, C., and Varani, G. (2007). RNA-binding proteins: modular design for efficient function. *Nature reviews. Molecular cell biology* *8*, 479-490.

Manning, G., Plowman, G.D., Hunter, T., and Sudarsanam, S. (2002). Evolution of protein kinase signaling from yeast to man. *Trends in biochemical sciences* *27*, 514-520.

Martin, C., and Zhang, Y. (2005). The diverse functions of histone lysine methylation. *Nature reviews. Molecular cell biology* *6*, 838-849.

McKee, A.E., Minet, E., Stern, C., Riahi, S., Stiles, C.D., and Silver, P.A. (2005). A genome-wide in situ hybridization map of RNA-binding proteins reveals anatomically restricted expression in the developing mouse brain. *BMC developmental biology* *5*, 14.

Meister, G. (2013). Argonaute proteins: functional insights and emerging roles. *Nature reviews. Genetics* *14*, 447-459.

Millioni, R., Tolin, S., Puricelli, L., Sbrignadello, S., Fadini, G.P., Tessari, P., and Arrigoni, G. (2011). High abundance proteins depletion vs low abundance proteins enrichment: comparison of methods to reduce the plasma proteome complexity. *PloS one* *6*, e19603.

Minguez, P., Letunic, I., Parca, L., and Bork, P. (2013). PTMcode: a database of known and predicted functional associations between post-translational modifications in proteins. *Nucleic acids research* *41*, D306-311.

Mitchell, S.F., and Parker, R. (2014). Principles and properties of eukaryotic mRNPs. *Molecular cell* *54*, 547-558.

Mittal, N., Roy, N., Babu, M.M., and Janga, S.C. (2009). Dissecting the expression dynamics of RNA-binding proteins in posttranscriptional regulatory networks. *Proceedings of the National Academy of Sciences of the United States of America* *106*, 20300-20305.

Morozova, N., Allers, J., Myers, J., and Shamoo, Y. (2006). Protein-RNA interactions: exploring binding patterns with a three-dimensional superposition analysis of high resolution structures. *Bioinformatics* *22*, 2746-2752.

Mujtaba, S., He, Y., Zeng, L., Farooq, A., Carlson, J.E., Ott, M., Verdin, E., and Zhou, M.M. (2002). Structural basis of lysine-acetylated HIV-1 Tat recognition by PCAF bromodomain. *Molecular cell* *9*, 575-586.

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

O'Connor, S.E., and Imperiali, B. (1996). Modulation of protein structure and function by asparagine-linked glycosylation. *Chemistry & biology* *3*, 803-812.

Okuwaki, M., Tsujimoto, M., and Nagata, K. (2002). The RNA binding activity of a ribosome biogenesis factor, nucleophosmin/B23, is modulated by phosphorylation with a cell cycle-dependent kinase and by association with its subtype. *Molecular biology of the cell* *13*, 2016-2030.

Panigrahi, A.K., Ernst, N.L., Domingo, G.J., Fleck, M., Salavati, R., and Stuart, K.D. (2006). Compositionally and functionally distinct editosomes in *Trypanosoma brucei*. *RNA* *12*, 1038-1049.

Pejaver, V., Hsu, W.L., Xin, F., Dunker, A.K., Uversky, V.N., and Radivojac, P. (2014). The structural and functional signatures of proteins that undergo multiple events of post-translational modification. *Protein science : a publication of the Protein Society* *23*, 1077-1093.

Ranum, L.P., and Day, J.W. (2004). Pathogenic RNA repeats: an expanding role in genetic disease. *Trends in genetics : TIG* *20*, 506-512.

Rothbart, S.B., and Strahl, B.D. (2014). Interpreting the language of histone and DNA modifications. *Biochimica et biophysica acta* *1839*, 627-643.

Scheiba, R.M., de Opakua, A.I., Diaz-Quintana, A., Cruz-Gallardo, I., Martinez-Cruz, L.A., Martinez-Chantar, M.L., Blanco, F.J., and Diaz-Moreno, I. (2014). The C-terminal RNA binding motif of HuR is a multi-functional domain leading to HuR oligomerization and binding to U-rich RNA targets. *RNA biology* *11*, 1250-1261.

Schulz, S., Doller, A., Pardini, N.R., Wilce, J.A., Pfeilschifter, J., and Eberhardt, W. (2013). Domain-specific phosphomimetic mutation allows dissection of different protein kinase C (PKC) isotype-triggered activities of the RNA binding protein HuR. *Cellular signalling* *25*, 2485-2495.

Steeg, P.S., Palmieri, D., Ouatas, T., and Salerno, M. (2003). Histidine kinases and histidine phosphorylated proteins in mammalian cell biology, signal transduction and cancer. *Cancer letters* *190*, 1-12.

Szymanski, C.M., and Wren, B.W. (2005). Protein glycosylation in bacterial mucosal pathogens. *Nature reviews. Microbiology* *3*, 225-237.

Tian, B., Bevilacqua, P.C., Diegelman-Parente, A., and Mathews, M.B. (2004). The double-stranded-RNA-binding motif: interference and much more. *Nature reviews. Molecular cell biology* *5*, 1013-1023.

Timchenko, L.T., Miller, J.W., Timchenko, N.A., DeVore, D.R., Datar, K.V., Lin, L., Roberts, R., Caskey, C.T., and Swanson, M.S. (1996). Identification of a (CUG)_n triplet repeat RNA-binding protein and its expression in myotonic dystrophy. *Nucleic acids research* *24*, 4407-4414.

Turner, M., Galloway, A., and Vigorito, E. (2014). Noncoding RNA and its associated proteins as regulatory elements of the immune system. *Nature immunology* *15*, 484-491.

Ule, J., Jensen, K., Mele, A., and Darnell, R.B. (2005). CLIP: a method for identifying protein-RNA interaction sites in living cells. *Methods* *37*, 376-386.

Varadi, M., Zsolyomi, F., Guharoy, M., and Tompa, P. (2015). Functional Advantages of Conserved Intrinsic Disorder in RNA-Binding Proteins. *PloS one* *10*, e0139731.

Vazquez, F., Ramaswamy, S., Nakamura, N., and Sellers, W.R. (2000). Phosphorylation of the PTEN tail regulates protein stability and function. *Molecular and cellular biology* *20*, 5010-5018.

Velankar, S., Dana, J.M., Jacobsen, J., van Ginkel, G., Gane, P.J., Luo, J., Oldfield, T.J., O'Donovan, C., Martin, M.J., and Kleywegt, G.J. (2013). SIFTS: Structure Integration with Function, Taxonomy and Sequences resource. *Nucleic acids research* *41*, D483-489.

Verdin, E., and Ott, M. (2015). 50 years of protein acetylation: from gene regulation to epigenetics, metabolism and beyond. *Nature reviews. Molecular cell biology* *16*, 258-264.

Vodermaier, H.C. (2004). APC/C and SCF: controlling each other and the cell cycle. *Current biology : CB* *14*, R787-796.

Vriend, G. (1990). WHAT IF: a molecular modeling and drug design program. *Journal of molecular graphics* *8*, 52-56, 29.

Walden, W.E., Selezneva, A.I., Dupuy, J., Volbeda, A., Fontecilla-Camps, J.C., Theil, E.C., and Volz, K. (2006). Structure of dual function iron regulatory protein 1 complexed with ferritin IRE-RNA. *Science* *314*, 1903-1908.

Wang, M., Herrmann, C.J., Simonovic, M., Szklarczyk, D., and von Mering, C. (2015). Version 4.0 of PaxDb: Protein abundance data, integrated across model organisms, tissues, and cell-lines. *Proteomics* *15*, 3163-3168.

Wang, Y.C., Peterson, S.E., and Loring, J.F. (2014). Protein post-translational modifications and regulation of pluripotency in human stem cells. *Cell research* *24*, 143-160.

Ward, J.J., Sodhi, J.S., McGuffin, L.J., Buxton, B.F., and Jones, D.T. (2004). Prediction and functional analysis of native disorder in proteins from the three kingdoms of life. *Journal of molecular biology* 337, 635-645.

Weber, G., Cristao, V.F., Santos, K.F., Jovin, S.M., Heroven, A.C., Holton, N., Luhrmann, R., Beggs, J.D., and Wahl, M.C. (2013). Structural basis for dual roles of Aar2p in U5 snRNP assembly. *Genes & development* 27, 525-540.

Will, C.L., and Luhrmann, R. (2011). Spliceosome structure and function. *Cold Spring Harbor perspectives in biology* 3.

Woodsmith, J., Kamburov, A., and Stelzl, U. (2013). Dual coordination of post translational modifications in human protein networks. *PLoS computational biology* 9, e1002933.

Yang, Y.C., Di, C., Hu, B., Zhou, M., Liu, Y., Song, N., Li, Y., Umetsu, J., and Lu, Z.J. (2015). CLIPdb: a CLIP-seq database for protein-RNA interactions. *BMC genomics* 16, 51.

Yao, C., Weng, L., and Shi, Y. (2014). Global protein-RNA interaction mapping at single nucleotide resolution by iCLIP-seq. *Methods in molecular biology* 1126, 399-410.

Zeidan, Q., and Hart, G.W. (2010). The intersections between O-GlcNAcylation and phosphorylation: implications for multiple signaling pathways. *Journal of cell science* 123, 13-22.

Chapter 4

Long non-coding RNA mediated regulation of gene expression in hereditary hemochromatosis

4.1 Introduction

Non-coding transcripts form a major part of the human transcriptome comprising nearly 48% (Pertea, 2012) to 68% (Iyer et al., 2015) of transcriptional output. Although by definition non-coding RNAs do not encode proteins, growing evidences suggest their involvement in crucial biological functions; they serve as key regulatory molecules in gene expression at epigenetic, transcriptional and post-transcriptional levels, protein localisation and serve as organisational frameworks for subcellular structures (Santosh et al., 2015; Wilusz et al., 2009). Aberrant activity of non-coding RNAs have been linked to various conditions such as cancer and other metabolic diseases. Hereditary hemochromatosis is a genetic metabolic disorder that results in excessive iron concentrations in the body. The aetiology of hemochromatosis is well known but the regulatory mechanism of non-coding RNAs is only beginning to be understood. In order to understand the role of non-coding RNAs in hemochromatosis I have investigated various aspects of lncRNA interaction with coding and non-coding transcripts. Firstly, I study homology of lncRNA transcripts between mouse and human. Next, by predicting miRNA-binding sites shared between mRNAs and lncRNAs, I study the competing regulatory interactions between them mediated by miRNAs. I then investigate correlation of

gene expression between sense-antisense pairs of mRNAs and long non-coding RNAs and their association with genomic regulatory elements to understand the *cis*- and *trans*-regulatory influence of lncRNAs on gene expression.

The genetic message encoded in the genome is conveyed through intermediary messenger RNAs (mRNA), which are translated into functional proteins, a process termed gene expression. The expression of a gene is controlled by different mechanisms which include regulating the number of transcribed copies of mRNA, the number of mRNAs available for translation into proteins, regulating the translational machinery or by regulating the proper folding and function of the translated protein product itself.

A significant fraction of the genome encodes non-protein-coding genes (Encode et al., 2007), which include ribosomal RNAs (rRNAs), transfer RNAs (tRNAs), long non-coding RNAs (lncRNAs; > 200 nucleotides) and short non-coding RNAs such as micro-RNAs (miRNA; ~22 nucleotides), small interfering RNAs (siRNAs; 20 to 25 nucleotides), PIWI-associated RNAs (piRNAs; 26 to 30 nucleotides), among others. Short non-coding RNAs regulate gene expression by a process called RNA interference (RNAi), wherein miRNA, siRNA and piRNA target complementary sequences on the 3' untranslated regions (UTR) of mRNA. The argonaute protein, which constitutes the RNA induced silencing complex (RISC) together with miRNA or siRNA, cleaves target mRNA upon complementary binding of miRNA seed sequences (nucleotides 2 to 7 from 5'-end) to mRNA resulting in loss of gene expression (Wilson and Doudna, 2013).

The modes of action of lncRNAs on regulating gene expression, however, are quite different. LncRNAs have been shown to influence gene expression at both their proximal loci (*cis*-acting) as well as at distal loci on a different chromosome (*trans*-acting). LncRNAs activate or repress gene expression through several mechanisms, which include affecting accessibility through chromatin modulation, binding to transcription factors, binding to protein-transport factors, forming lncRNA-DNA triplex structures or mimic DNA-binding sites (Geisler and Coller, 2013). One of the best-studied examples of a *cis*-acting

lncRNAs is the 'X inactive specific transcript (Xist)' (Ng et al., 2007). Xist regulates dosage compensation in female cells by associating with one of the X-chromosome from which it was transcribed and causes its inactivation (Ng et al., 2007). Xist mediates this inactivation/transcriptional silencing by acting as a scaffold for multiple regulatory proteins. The A-repeat region of Xist directly binds to the scaffold attachment factor A and histone deacetylase 1 (HDAC1) and promotes deacetylation by HDAC3 and demethylation of histone 3 on lysine 4 (H3K4) (Engreitz et al., 2016). Xist also indirectly recruits polycomb repressive complexes 1 and 2 (PRC1 and PRC2) through binding of heterogeneous nuclear ribonucleoproteins K (hnRNPK) and SMRT/HDAC1-associated repressor protein (SHARP) to the B-F repeat (Engreitz et al., 2016; McHugh et al., 2015). PRC1 and PRC2 mediate trimethylation of histone H3 on lysine 27 (H3K27me3), a repressive chromatin mark, across the inactive X chromosome. The histone methyltransferase SETDB1, one of the other proteins recruited by Xist deposits repressive H3K9me2 and H3K9me3 marks. It has been noted that lncRNAs also take part in activating gene expression in *cis*. An example of such activating lncRNA is the 'HOXA transcript at the distal tip' (HOTTIP) (Wang et al., 2011). HOTTIP is transcribed from the 5' tip of the HOXA locus, which includes a cluster of genes expressed during embryonic development (Wang et al., 2011). HOTTIP binds WD repeat-containing protein 5 (WDR5) and recruits mixed-lineage leukaemia (MLL) proteins, which are SET-domain-containing lysine methyltransferases and deposit H3K4me3 marks near the transcription start sites of multiple 5' HOXA genes (Wang et al., 2011). Knockdown on WDR5 inhibits expression of 5' HOXA genes and HOTTIP RNA, while overexpression of HOTTIP is implicated in lung, pancreatic, colorectal, prostate and gastric cancers (Lian et al., 2016). Some of the other *cis*-acting lncRNAs include AIR (Nagano et al., 2008) and ANRIL (Yap et al., 2010).

A number of lncRNAs are shown to function in *trans*. One of the well-known examples includes the 'HOX transcript antisense RNA' (HOTAIR) (Tsai et al., 2010). The antisense HOTAIR is transcribed from the HOXC locus between genes HOXC11 and HOXC12 on human chromosome 12 and carries out silencing of genes on the HOXD locus on chromosome 2 (Rinn et al., 2007). The 5' domain

and the 3' domains of HOTAIR were shown to bind PRC2 and LSD1 respectively (Rinn et al., 2007; Tsai et al., 2010). PRC2 comprises H3K27 methylase EZH2, SUZ12 and EED, while LSD1 is a demethylase that mediates enzymatic demethylation of H3K4me2 (Tsai et al., 2010). Both PRC2 and LSD1 can bind to multiple proteins providing it with DNA target specificity (Tsai et al., 2010). The HOTAIR complex, by a yet unknown mechanism, guides PRC2 and LSD1 to various genomic locations resulting in their silencing via H3K27me3 by PRC2 and H3K4 demethylation by LSD1 (Rinn et al., 2007; Tsai et al., 2010). It is shown that the microRNA miR-141 regulates expression of HOTAIR (Chiyomaru et al., 2014) and that HOTAIR expression is altered in many cancers including breast (Gupta et al., 2010), gastric (Hajjari et al., 2013) and pancreatic cancer (Kim et al., 2013). Interestingly recent studies have cast doubts on the function of HOTAIR in silencing HOXD cluster (Schorderet and Duboule, 2011; Selleri et al., 2016). Unlike in humans, HOTAIR was observed not to significantly influence HoxD cluster of genes in mouse (Schorderet and Duboule, 2011). HOTAIR is poorly conserved in sequence between human and mouse. The complete deletion of HoxC cluster in mouse does not influence expression pattern or chromatin marks on target HoxD genes (Schorderet and Duboule, 2011), but on the contrary a 4-kb deletion within the HoxC cluster in mouse was shown to derepress expression of HoxD gene cluster leading to severe phenotypes (Li et al., 2013). Further recent follow up studies on HoxC cluster deletion in mice have reconfirmed previous observations that HOTAIR has no major role in *trans*-regulation of gene expression of HoxD locus, but allude to a *cis*-regulation of expression of neighbouring HoxC11 and HoxC12 genes (Amandio et al., 2016). The differences in observations between the two studies on the role of HOTAIR on gene expression regulation of HoxD locus is attributed to different genetic backgrounds of mice used (inbred C57BL/6 versus mixed background C57BL/6 and CBA) and transcriptome profiling of cells from different tissues (tail tip fibroblasts versus forelimb, hindlimb, genital tubercle, and lumbosacral, sacrocaudal, and caudal trunk) and at different developmental stages (newborn mice versus E12.5 embryos) (Li et al., 2016a; Selleri et al., 2016). Additional studies on HOTAIR are proposed to clearly understand the role of HOTAIR in regulation of expression of HoxD locus (Li et al., 2016a).

Recently a novel mechanism of gene expression regulation has been proposed involving both lncRNAs and miRNAs termed competing endogenous RNA (ceRNA) (Salmena et al., 2011). Protein non-coding transcripts, such as lncRNAs and pseudogenes, that share common miRNA binding sites or miRNA response elements (MREs) with mRNAs, compete with mRNAs to bind the same pool of miRNAs (Salmena et al., 2011). This crosstalk between non-coding RNAs and mRNAs is proposed to regulate their respective gene expression levels (Salmena et al., 2011). Based on their relative abundances and the number of MREs, non-coding RNAs act as molecular sponges in sequestering miRNAs and thereby influence protein expression levels (Figure 4.1). This mode of gene expression regulation has been observed in a few cases: the tumour suppressor gene PTEN and its pseudogene PTENP1 are both targets of miRNAs miR-19b and miR-20a (Poliseno et al., 2010). Overexpression of PTENP1 3' un-translated region (UTR) transcripts derepresses expression of PTEN transcript and protein, indicating that PTENP1 3' UTR functions as a decoy to bind miR-19b and miR-20a and promotes PTEN mRNA expression (Poliseno et al., 2010). Other complex processes have been observed where competitive associations with miRNAs are used as a means to auto-regulate protein activity. The lncRNA HULC acts as a decoy to bind miR-372, which also targets the 3' UTR of PRKACB. The transcription factor CREB, which is phosphorylated by PRKACB, auto regulates its activity by overexpressing HULC which in turn sustains PRKACB expression (Wang et al., 2010).

In this chapter I investigate the functions and regulatory role of lncRNAs, using various approaches, which includes studying the ceRNA hypothesis on regulating gene expression in hereditary hemochromatosis. This work is in collaboration with Dr. Martina Muckenthaler and Dr. Kamesh Rajendra Babu from the University Hospital Heidelberg. Hereditary hemochromatosis (HH) is an autosomal condition that causes systemic iron overload due to mutations in one or more iron response genes (Bomford, 2002). Characteristics of hemochromatosis include increased absorption of iron, hyperferremia and tissue iron overload. Other associated complications include liver cirrhosis, cancer, diabetes, heart failure and arthritis (Muckenthaler, 2014). HH is

observed in people of northern European descent with an estimated prevalence of 0.4% to 9.2% in the population (Bomford, 2002). To date five types of hemochromatosis have been described: Type 1, which is the most prevalent subtype, is autosomal recessive caused by the mutation C282Y in the HH gene *Hfe* (Feder et al., 1996); Type 2 or juvenile hemochromatosis is a rare autosomal recessive condition caused by an unidentified locus (Roetto et al., 1999); Type 3 is an autosomal recessive condition caused by mutation in the transferrin receptor 2 protein *Tfr2* (Camaschella et al., 2000); Type 4 hemochromatosis is autosomal dominant and is due to mutation in the intestinal iron transporter ferroportin *Fpn* (*Slc40a1*) (Njajou et al., 2001) and Type 5, which is an autosomal dominant condition caused by mutation in the H-subunit of iron storage protein ferritin (Kato et al., 2001).

The hepatic proteins affected in type 1, type 3 and type 5 hemochromatosis, namely *Hfe*, *Tfr2* and ferritin respectively, are involved in upstream iron sensing, regulating iron uptake and storage (Pantopoulos, 2008; Worthen and Enns, 2014). In comparison, the protein affected in type 4 hemochromatosis is the downstream major cellular iron exporter ferroportin (Njajou et al., 2001). The regulatory mechanisms by which, intestinal iron absorption and export into the bloodstream by ferroportin, offer unique insights into homeostasis of iron in the body. Figure 4.2 illustrates the systemic iron metabolism pathway. The liver secretes a peptide hormone hepcidin (*Hamp*) in response to high systemic iron levels and inflammation (Nemeth et al., 2004). Hepcidin binds to its only known receptor ferroportin, present on duodenal enterocytes and macrophages (Donovan et al., 2005). Upon hepcidin binding, ferroportin undergoes ubiquitination, which causes it to be internalised and degraded (Nemeth et al., 2004; Qiao et al., 2012). The gain-of-function mutation C326S makes ferroportin resistant to hepcidin (Altamura et al., 2014) which prevents its internalisation and degradation (Fernandes et al., 2009) resulting in unchecked iron export into the bloodstream causing systemic iron overload.

Two mouse models, *Slc40a1*^{C326S/C326S} (*Fpn*-C326S), which has iron overload and resembles pathology of human HH type 4, and ferroportin trap *Slc40a1*^{trap} (*Fpn*-

Trp), which resembles iron deficiency, are used in this study (see methods 4.2.1 for description of mouse models). Small non-coding RNA has been shown to play a regulatory role in iron homeostasis (Castoldi et al., 2011). The liver specific miRNA miR-122 controls systemic iron homeostasis in mouse by targeting 3' UTR of mRNAs that encode activators of hepcidin (*Hamp*) transcription, such as hemochromatosis (*Hfe*), hemojuvelin (*Hjv*) and bone morphogenetic protein receptor type 1 A (*Bmpr1a*) (Castoldi et al., 2011). Other miRNAs associated with iron metabolism include miR-Let-7d, miR-196, miR-320 and miR-485-3p (Yujing Li, 2013). Although miRNA mediated regulation of iron homeostasis is well documented, very little is known about the role of lncRNAs. I investigate the potential regulatory role of lncRNAs in iron homeostasis by exploring possible interactions between lncRNAs, miRNAs and mRNAs and the correlation in expression of mRNA and lncRNA pairs.

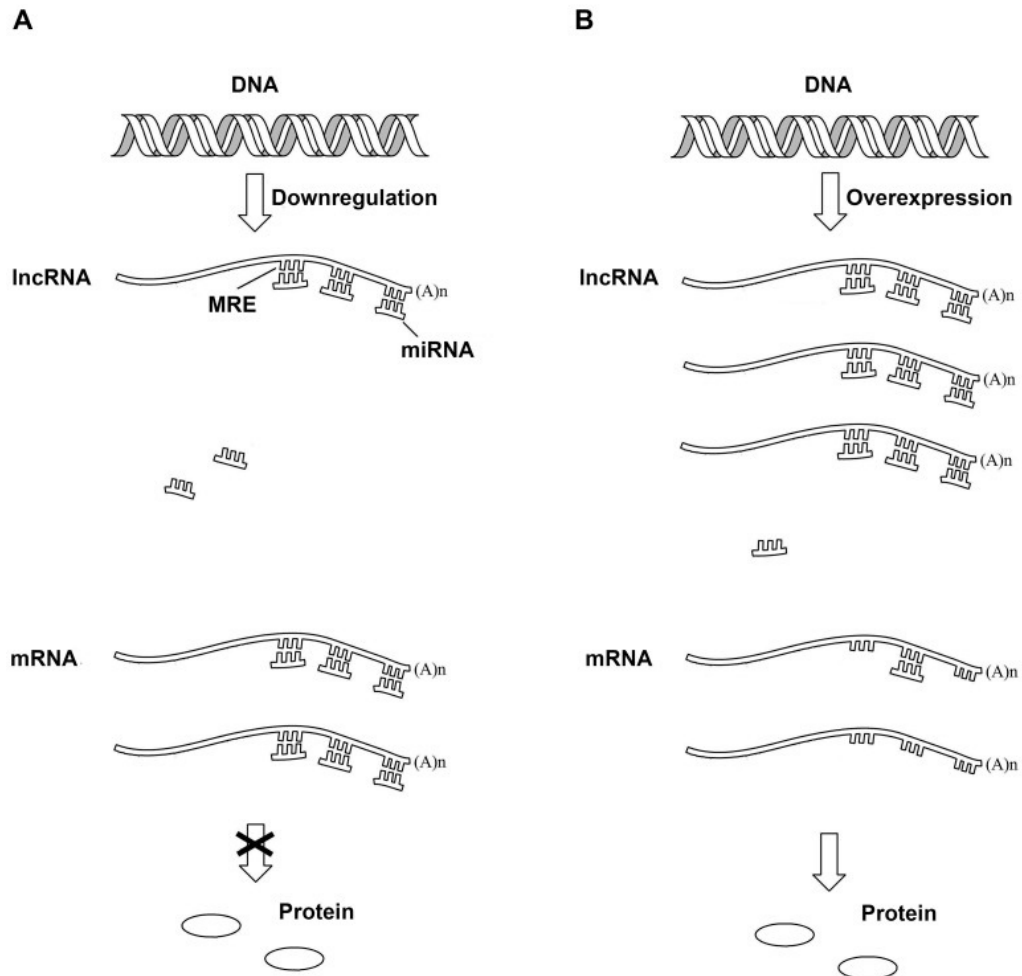


Figure 4.1 Schematic representation of competing endogenous RNA (ceRNA) hypothesis. When long non-coding RNA and mRNA share the same miRNA response elements (A) downregulation of lncRNAs causes free miRNAs to target mRNA resulting in decreased protein expression, while (B) overexpression of lncRNAs sequesters (sponges) away miRNAs which results in availability of mRNA translation. Figure adapted from (Xia et al., 2014) DOI: 10.1038/srep06088.

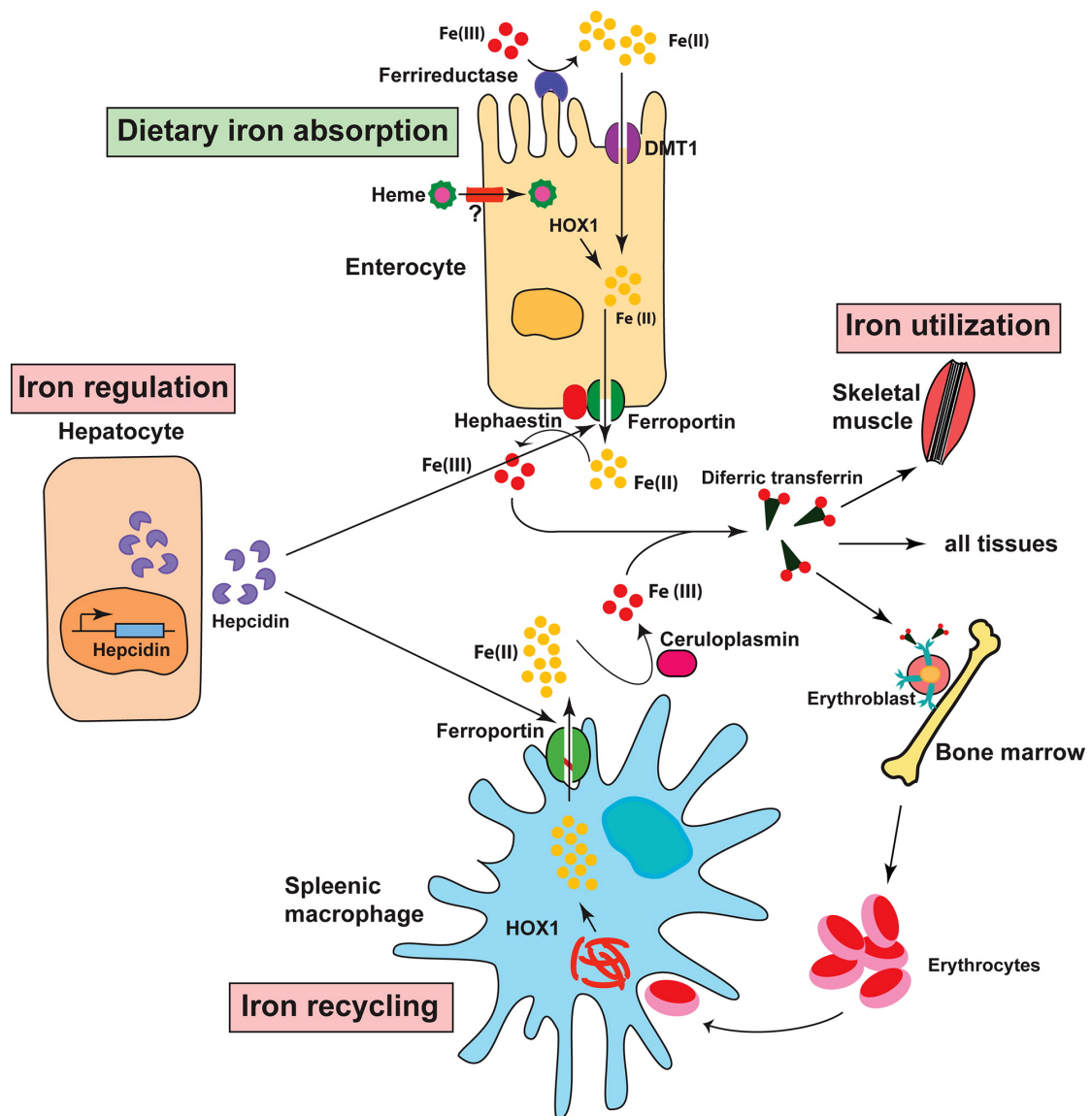


Figure 4.2 Schematic representation of systemic iron metabolism pathway. The dietary iron is absorbed through the divalent metal transporter 1 (Dmt1) present on the enterocytes and is exported and released into systemic circulation by ferroportin (Slc40a1). The peptide hormone hepcidin (Hamp) binds to ferroportin in response to high systemic iron levels and inflammation and regulates its activity. Figure adapted from (Pantopoulos et al., 2012) DOI: 10.1021/bi300752r

4.2 Methods

4.2.1 Mouse models of hereditary hemochromatosis

Two mouse models were used to study the effects of iron metabolism in hereditary hemochromatosis. The *Slc40a1* locus was targeted to introduce a C326S point mutation (Altamura et al., 2014). First the *Slc40a1*^{trap} (Fpn-Trp) trapped allele was generated by introducing the β -Geo cassette into the sixth intron of *Slc40a1* locus alongside the C326S mutation in the seventh exon (Altamura et al., 2014). The *Slc40a1*^{trap} mouse line was obtained by injecting the targeted embryonic stem cells into mouse blastocysts followed by germline transmission of the targeted allele (Altamura et al., 2014). The *Slc40a1*^{trap} line was crossed with CRE-deletor strain to remove β -Geo cassette resulting in *Slc40a1*^{wt/C326S} mice. The *Slc40a1*^{C326S/C326S} (Fpn-C326S) homozygous mutant mouse line was then obtained by intercrossing heterozygous *Slc40a1*^{wt/C326S} mice (Altamura et al., 2014). *Slc40a1*^{C326S/C326S} mice mimic hereditary hemochromatosis and exhibit high transferrin saturation and increased serum ferritin levels (Altamura et al., 2014). The above experiments were performed by Dr. Martina Muckenthaler and colleagues at the University Hospital Heidelberg and EMBL, Heidelberg.

4.2.2 RNA sequencing and differential expression analysis

To identify differentially expressed transcripts, total RNA isolates were obtained from liver of *Slc40a1*^{C326S/C326S}, *Slc40a1*^{trap}, and wild-type mice at the age of 10 weeks (3 mice per group). Ribosomal RNAs were depleted and isolated RNA was subjected to strand-specific RNA sequencing using the HiSeq 2500 system (Illumina). Sequence reads were aligned to the mouse reference genome assembly GRCm38 (Ensembl) using TopHat2. RNA transcripts were annotated by aligning against Rfam (Nawrocki et al., 2015). RNA transcripts that do not contain ORFs (open reading frame) were predicted as non-coding RNAs. Differentially expressed transcripts were analysed using DESeq2 package in R and Bioconductor. Differentially expressed RNA transcripts were selected based

on False Discovery Rate (FDR) cut-off < 0.1 and P-value < 0.05 . The above experiments were performed by Dr. Martina Muckenthaler and colleagues at the University Hospital Heidelberg and EMBL, Heidelberg.

In Fpn-C326S mouse model 193 mRNA transcripts show expression ≥ 2 fold and 225 mRNA transcripts are expressed ≤ 0.5 fold compared to wild-type mice. Among lncRNAs, 11 transcripts are expressed ≥ 2 fold and 19 transcripts have an expression value ≤ 0.5 fold. In order to include more transcripts for analysis, the expression fold change values were relaxed to ≥ 1.5 fold and < 1 fold. 22 lncRNA transcripts expressed ≥ 1.5 fold and 30 lncRNA transcripts are expressed < 1 fold compared to wild-type mice were considered.

In Fpn-Trp mouse model 268 mRNA transcripts show expression greater than or equal to two fold and 128 mRNA transcripts are expressed lower than 0.5 fold compared to wild-type mice respectively. Among lncRNAs, 129 lncRNA transcripts are expressed greater than or equal to two fold and 33 lncRNAs are expressed less than 0.5 fold compared to wild-type mice respectively.

4.2.3 Homology of lncRNAs

Non-coding homologues of differentially expressed mouse lncRNAs were searched against non-coding RNA human RefSeq database (NR_) using nhmmer (Wheeler and Eddy, 2013) at default parameters (gap open probability, popen: 0.03; gap extension probability, pextend: 0.75; significant E-value threshold, incE: 0.01). Syntenic relations between mouse lncRNAs and human transcripts were inferred from Ensembl 84 (Aken et al., 2016) by comparing protein-coding genes within the vicinity. The mouse lncRNA was considered syntenic if the lncRNA gene shared similar context with the human lncRNA including homologous protein-coding genes in the neighbourhood. Coding potential of differentially expressed lncRNAs were investigated using Coding Potential Assessment Tool (CPAT) (Wang et al., 2013) and Coding Potential Calculator (CPC) (Kong et al., 2007).

4.2.5 miRNA target site prediction

A software package was developed in object oriented Perl to predict conserved miRNA-binding sites on lncRNAs and 3' UTRs of mRNAs. Although numerous software packages are available for predicting miRNA targets, I developed local miRNA target prediction software to develop object oriented programming skills. Perfect reverse complementary of miRNA seed sequence (6mer or 7-mer) on lncRNA or the 3' UTR of mRNAs was assumed sufficient to call those sites as miRNA recognition elements (MREs) or miRNA-binding sites. Evolutionary conservation of miRNA-binding sites between mouse and human transcripts were inferred using genomic alignments from UCSC MAF (multiple alignment format) files. The multiz60way genomic alignments of mammalian genomes were queried from UCSC Table Browser using genomic coordinates of mouse lncRNAs (GRCm38/mm10) and the alignments used as input. A list containing 192 miRNAs expressed in mouse liver was obtained from Dr. Anton Enright's lab (EMBL-EBI, Hinxton). This list was compared with the list of miRNAs that are expressed in the adult mouse liver downloaded from the microRNA expression and sequence analysis database (mESAdb) (Kaya et al., 2011) corresponding to Beuvink et al., dataset (Beuvink et al., 2007). After comparison 67 miRNA families were found to be common between both sets and were used in the analysis. miRNA seed sequences were downloaded from miRBase (Kozomara and Griffiths-Jones, 2014). Experimentally validated miRNA-target sites in mouse 3' UTR of mRNAs were downloaded from TarBase version 7.0 (Paraskevopoulou et al., 2016).

4.2.6 Competing endogenous RNA network

The competing endogenous RNA interaction network was modelled using common miRNA interactions shared between lncRNAs and mRNAs. Differentially expressed lncRNAs and mRNAs that are annotated to be involved in iron homeostasis and miRNAs expressed in mouse liver, which target both lncRNAs and mRNAs, form nodes of the network. The predicted miRNA-binding sites in lncRNAs and experimentally validated miRNA-binding sites on 3' UTR of mRNAs

form the edges. SwissProt reviewed protein-coding genes that are involved in iron homeostasis were obtained by querying UniProt using the terms 'iron ion homeostasis [55072]' and 'Mus musculus (Mouse) [10090]' in the advanced 'Gene Ontology [GO]' and Organism [OS]' search options respectively. The competing endogenous RNA network is represented as Sankey diagram and was generated in R using the rCharts package.

4.2.7 Sense-antisense mRNA-lncRNA pairs

Differentially expressed long non-coding RNAs were grouped into two classes: antisense RNA and long intervening ncRNA (lincRNA). Antisense RNAs were defined as long non-coding RNAs that overlap intronic and/or exonic regions of a sense mRNA and are transcribed in the opposite direction relative to the sense mRNA. The antisense non-coding RNAs were further classified into 9 sub-groups based on the Ensembl 84 regulatory features they are associated with. Association with a regulatory feature was defined as the significant overlap of antisense RNA exon with a genomic regulatory feature. If the antisense RNA exons overlapped more than one regulatory feature, the longest overlap with the regulatory feature was considered. These sub-groups are: (1) CTCF-binding site associated, (2) Enhancer associated, (3) No regulatory feature, (4) Open chromatin associated, (5) Promoter associated (bi-directional), (6) Promoter associated (non bi-directional), (7) Promoter flanking region associated (bi-directional), (8) Promoter flanking region associated (non bi-directional) and (9) Transcription factor (TF) binding site associated.

The terminologies associated with promoter and promoter-flanking regions used in this study are described as follows- Promoter associated (bi-directional): the antisense ncRNA and sense mRNA share (overlap) the same promoter region and their 5' ends orient towards each other (bi-directional gene pair). Promoter associated (non bi-directional): the antisense ncRNA and sense mRNA have individual promoters. Promoter flanking region associated (bi-directional): the antisense ncRNA and sense mRNA share (overlap) the same promoter flanking region and their 5' ends orient towards each other. Promoter flanking region

associated (non bi-directional): the antisense ncRNA has its individual promoter-flanking region and does not overlap with sense mRNA.

4.2.8 LincRNA-adjacent mRNA pairs

LincRNAs were defined as long non-coding RNAs that are present within intergenic regions and do not overlap a protein-coding gene. Upstream and downstream protein-coding genes of both strands from transcription start site of lincRNAs were identified and classified into 4 groups: (1) Upstream gene on antisense strand, (2) Upstream gene on sense strand, (3) Downstream gene on antisense strand and (4) Downstream gene on sense strand.

4.2.9 Gene Ontology

Protein coding genes on both strands present in either direction within 500KB and 1MB vicinity of differentially expressed lincRNAs were tested for enrichment of the following terms; in biological process: cellular iron ion homeostasis, cellular response to iron ion, ferrous iron import into cell, iron ion homeostasis, iron ion import, iron-sulphur cluster assembly, multicellular organismal iron ion homeostasis, negative regulation of iron ion transmembrane transport, response to iron ion; and in molecular function: ferric iron binding, ferrous iron binding, iron channel inhibitor activity, iron ion binding, iron-responsive element binding and iron-sulphur cluster binding. The enrichment analysis was limited to these terms to reduce background signal and increase the sensitivity of statistical tests.

4.3 Results

4.3.1 Overview of Fpn-C326S and Fpn-Trp datasets

Martina Muckenthaler, Kamesh R. Babu and colleagues at the University Hospital Heidelberg and EMBL, Heidelberg have experimentally identified protein coding and non-coding transcripts expressed in the liver tissues of mouse hereditary hemochromatosis model systems Fpn-C326S and Fpn-Trp. Gene expression fold changes, compared to wild-type mice, were inferred for 4,125 mRNAs and 255 non-coding RNAs in the Fpn-C326S mouse model and for 7,562 mRNA and 2,266 non-coding RNAs in the Fpn-Trp mouse model. This data from their study is used to further investigation. Figure 4.3A,B shows the distribution of fold changes of differentially expressed transcripts from the two data sets. The ncRNAs are marginally overexpressed in both the data sets compared to the wild type (the mean expression value of all ncRNAs from Fpn-C326S mouse model is 1.35 and the mean expression value of all ncRNAs from Fpn-Trp is 1.26). On an average the ncRNAs are slightly overexpressed compared to their protein coding genes (the mean expression value of mRNAs from Fpn-C326S mouse model is 1.12 and the mean expression value of mRNAs from Fpn-Trp is 1.07). Only lncRNAs were selected from these datasets for further analysis. Investigation of protein-coding potential of differentially expressed lncRNAs from the two datasets shows that a majority of the transcripts have low protein coding potential (Figure 4.3C,D). A few lncRNAs were annotated with high protein coding potential but a closer inspection indicated that these transcripts as antisense or processed pseudogenes.

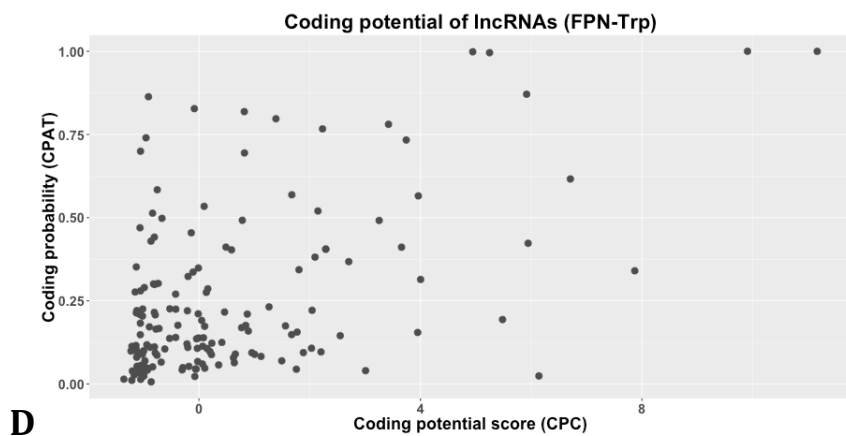
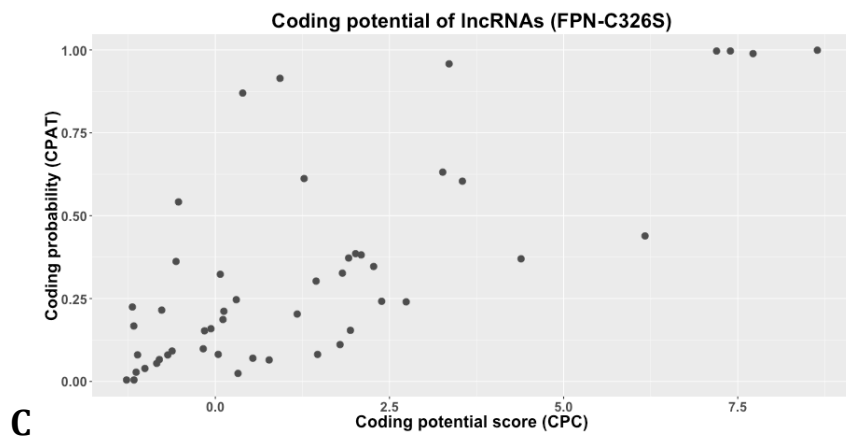
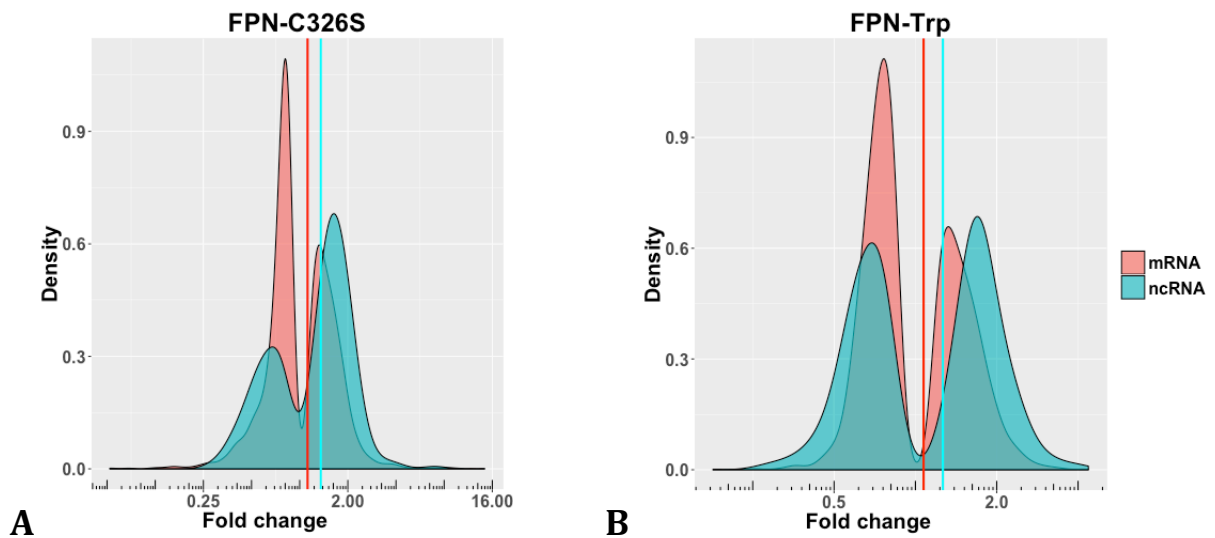


Figure 4.3 Overview of Fpn-C326S and Fpn-Trp datasets. The distribution of coding and non-coding transcripts in the two datasets (A) Fpn-C326S iron overload and (B) Fpn-Trp iron deficiency mouse models. The protein coding potential of non-coding transcripts (C, D) are poor. Transcripts with greater than 75% CPAT score are annotated pseudogenes.

4.3.2 Homologues of lncRNAs

Similarity between sequences is often associated with similarity in function (Joshi and Xu, 2007). Identifying functionally annotated transcripts with significant sequence similarity to the differentially expressed mouse lncRNAs might aid in inferring their putative functions. In order to infer function through homology I have carried out sequence similarity searches to identify lncRNA homologues in humans. Searches for lncRNA homologues in the previous studies have suggested that, unlike protein-coding genes, lncRNAs do not share significant sequence similarities between other species (Nam and Bartel, 2012; Ulitsky et al., 2011). To find sequence homologues of the differentially expressed lncRNAs in Fpn-C326S iron overload mouse model, I first scanned these sequences across the database of known RNA families in Rfam (Nawrocki et al., 2015). The Rfam database comprises a collection of covariance models, multiple sequence alignments and consensus secondary structures of non-coding RNA families (Griffiths-Jones et al., 2003). Two lncRNA sequences showed significant homology to RNA families. The lncRNA Trp53cor1 is homologous to the lincRNA-p21 families RF01889 (E-value: $4.5e-34$), RF01890 (E-value: $2.9e-13$) and the lncRNA Rab26os is significantly similar to the Snord60 family RF00271 (E-value: $7.7e-20$). Rfam (version 12.0; September 2014) contains 2,450 non-coding RNA families and represent the largest collection of non-coding RNA families, but they are by no means exhaustive. Therefore to obtain a better coverage these sequences were searched against the human reference non-coding RNA sequence database (NR_) using nhmmer (Wheeler and Eddy, 2013). The nhmmer sequence search resulted in alignments of low query coverage against target sequences (Figure 4.4), indicating poor sequence similarity with other lncRNAs. Only 6 mouse lncRNAs showed significant sequence similarities (query coverage greater than 50% and E-value $\leq 10^{-25}$) with 15 human non-coding RNAs. Table 4.1 lists some of the lncRNA homologues that show significant sequence similarities with human non-coding transcripts. A complete list of homologues of lncRNA expressed in the Fpn-Trp iron deficient mouse model is listed in the appendix (Table A2).

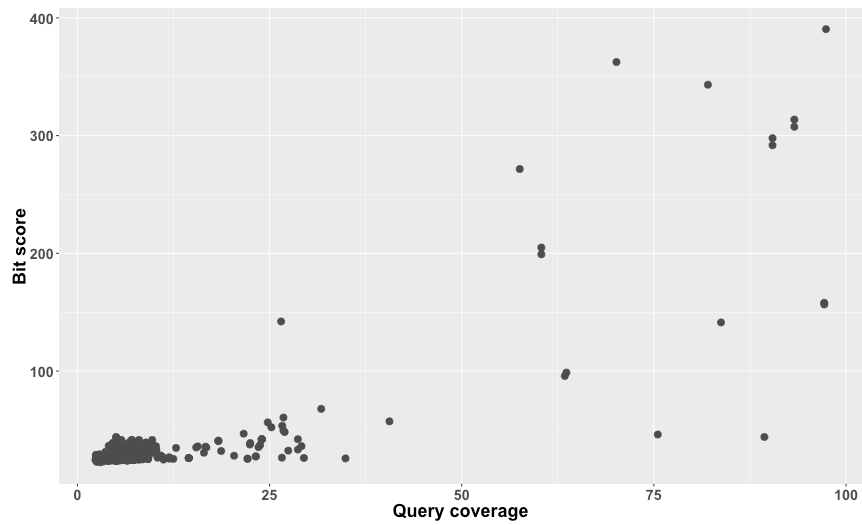


Figure 4.4 Sequence similarity of differentially expressed lncRNAs in Fpn-C326S iron overload mouse model. Most mouse lncRNAs show poor sequence conservation with other non-coding transcripts. Sequence homologues with greater than 50% sequence coverage are annotated pseudogenes and lncRNAs that belong to the serine protease inhibitor and aldo-keto reductase family. All targets have an E-value $\leq 10^{-5}$.

Evolutionary sequence conservation is widely used as an indicator of homology among protein-coding genes. Sequence similarity search works better in identifying homologues of protein-coding genes, which are under selective pressures to maintain nucleotide or amino acid sequence conservation, but it is often less sensitive in identifying lncRNA homologues which have very little sequence conservation. In addition to sequence similarity other approaches have been suggested, such as comparison of RNA secondary structure, function and analysis of expression from syntenic loci, to identify lncRNA homologues (Diederichs, 2014). Secondary and tertiary structures are more robust to changes in sequence. Co-variations of paired nucleotides can still conserve secondary structures by maintaining base-pairing properties without having to conserve sequence (Eddy and Durbin, 1994), thereby allowing detection of homologues between lncRNAs with low sequence similarity. But while predicting RNA secondary structures is feasible for short RNA segments, the process becomes computationally expensive with increase in sequence length. Multiple RNA sequence alignments are essential in predicting statistically significant evolutionary conserved RNA secondary structures for comparison (Rivas et al., 2017), but the lack of homology of lncRNAs in these datasets does not merit the use of secondary structure comparison approach. Syntenic analysis offers another approach to identify homologous lncRNAs. It has been shown that protein-coding genes adjacent to a lncRNA gene are likely to have orthologs adjacent to lncRNA, indicating that genomic positions of lncRNAs can be conserved despite low sequence homology (Ulitsky et al., 2011). I have therefore compared the genomic loci of these lncRNAs by comparing their neighbouring protein-coding genes as references. I observe that a few lncRNAs show syntenic relation with the human non-coding RNAs based on the conserved genomic organisation (Table 4.2).

4.3.3 miRNA target site prediction and conservation

It is well known that a contiguous and perfect base pairing between miRNA seed sequences (nucleotides 2 to 8) and 3' UTR of mRNAs promotes mRNA degradation and/or prevent translation (Filipowicz et al., 2008). Using a miRNA

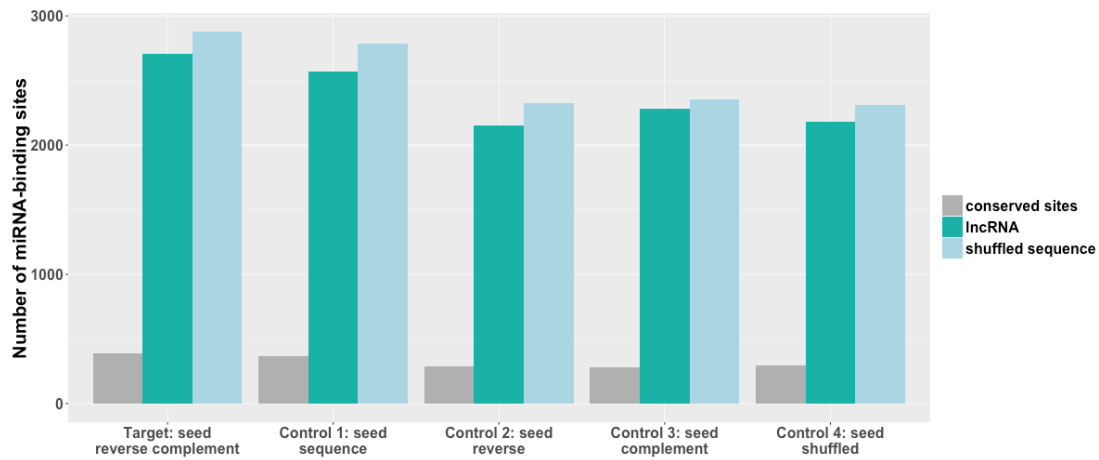
target prediction pipeline I predicted miRNA-binding sites in mouse lncRNAs and 3' UTRs of mRNAs and compared their conservation across human transcripts. The number of predicted miRNA-binding sites in mouse lncRNAs is higher compared to the conserved target sites. Similarly the number of predicted miRNA-binding sites in 3' UTRs of mRNAs is higher than the conserved binding sites between mouse and human mRNAs (Figure 4.5). LncRNA and mRNA sequences were shuffled and used as control sequences. Comparison of predicted miRNA-binding sites between target sequences and control shuffled sequences did not show any significant difference, indicating no enrichment of miRNA-binding sites. Further, four control experiments were carried out (Control 1 to 4), which included (i) matching occurrences of identical miRNA seed sequence and (ii) reversed seed sequence, (iii) complement of seed sequence and (iv) shuffled seed sequence on target transcripts. When compared with shuffled transcript sequences, the mRNA and lncRNA transcripts do not show a significant difference in the number of control miRNA sites (Figure 4.5), which indicates that the liver miRNAs do not preferentially target mRNA and lncRNA transcripts expressed in iron overload mice, than any transcript just by chance.

Mouse lncRNA ID	Mouse lncRNA gene	Human ncRNA accession	Human ncRNA gene	Query coverage (%age)	E-value	Bit score
ENSMUSG00000071414	Gm6736	NR_026743	AKR1C6P	97.161	2.00E-44	158
ENSMUSG00000071414	Gm6736	NR_073125	AKR1E2	93.270	1.50E-91	313.6
ENSMUSG00000071414	Gm6736	NR_073126	AKR1E2	90.431	9.30E-87	297.8
ENSMUSG00000071414	Gm6736	NR_027916	AKR1C8P	63.407	1.40E-25	95.8
ENSMUSG00000071414	Gm6736	NR_073127	AKR1E2	60.358	1.30E-58	204.9
ENSMUSG00000076576	Igkv6-32	NR_027293	BMS1P20	75.504	6.30E-10	46.1
ENSMUSG00000082087	Gm12138	NR_026743	AKR1C6P	97.161	4.30E-44	156.6
ENSMUSG00000082087	Gm12138	NR_073125	AKR1E2	93.270	8.30E-90	307.5
ENSMUSG00000082087	Gm12138	NR_073126	AKR1E2	90.431	4.90E-85	291.8
ENSMUSG00000082087	Gm12138	NR_027916	AKR1C8P	63.617	1.50E-26	98.7
ENSMUSG00000082087	Gm12138	NR_073127	AKR1E2	60.358	5.30E-57	199.2
ENSMUSG00000083534	H2-M6-ps	NR_001434	HLA-H	97.389	6.40E-115	390.4
ENSMUSG00000083534	H2-M6-ps	NR_027822	HLA-L	57.544	5.80E-79	271.6
ENSMUSG00000085355	3010003L21Rik	NR_026806	FLJ13224	82.026	1.10E-100	343.2
ENSMUSG00000090555	Gm8893	NR_073112	SERPINB1	89.372	3.10E-10	44
ENSMUSG00000090555	Gm8893	NR_015340	SERPINA13P	83.736	1.10E-39	141.3
ENSMUSG00000090555	Gm8893	NR_110563	SERPINA2	70.129	1.40E-106	362.5

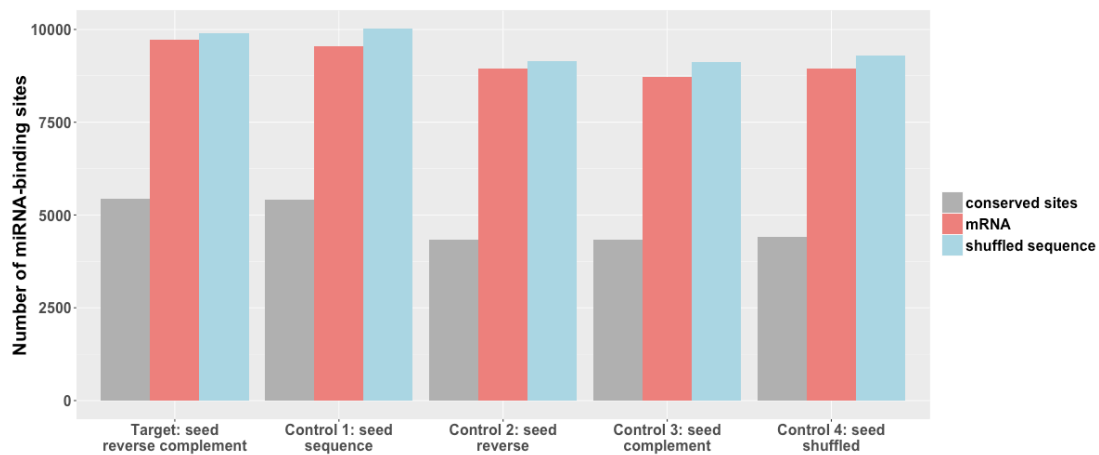
Table 4.1 Sequence homologues of mouse lncRNAs in human identified using nhmmer.

Mouse lncRNA ID	Mouse lncRNA gene	Human ncRNA ID	Human ncRNA gene
ENSMUSG00000085355	3010003L21Rik	ENSG00000177340	FLJ13224
ENSMUSG00000053889	Kirrel3os	ENSG00000257271	KIRREL3-AS1
ENSMUSG00000052188	Gm14964	ENSG00000237363	AP006288.1
ENSMUSG00000085132	Gm12265	ENSG00000197815	RP1-253P7.4
ENSMUSG00000074918	Inafm2	ENSG00000259330	INAFM2
ENSMUSG00000087404	Gm11752	ENSG00000261978	CTD-2529O21.2
ENSMUSG00000085439	Rapgef4os1	ENSG00000228016	RAPGEF4-AS1

Table 4.2 Differentially expressed mouse lncRNAs and their syntenic human homologues.



A



B

Figure 4.5 Predicted miRNA-binding sites in mouse (A) differentially expressed lncRNAs and (B) mRNAs associated with iron ion homeostasis. Target experiment denotes identifying miRNA seed sequence-binding sites (6-mers) on lncRNA and mRNA transcripts and their shuffled sequences. Control experiments (Controls 1 to 4) denote identifying occurrences of miRNA control seed sequences on transcripts and their shuffled sequences.

4.3.4 Competing endogenous RNA network

It is estimated that nearly 74% to 92% of all protein-coding genes in the four model genomes- worm, fruit fly, mouse and human- are regulated by miRNAs (Miranda et al., 2006). Both lncRNAs and protein-coding transcripts exhibit binding sites for multiple miRNAs and in most cases can be bound by more than one miRNA at the same time (Peter, 2010; Wu et al., 2010) resulting in a tightly controlled mechanism of regulating gene expression. It has been proposed that lncRNAs and miRNAs compete for miRNA-binding and regulate gene expression (Salmena et al., 2011). In order to understand the influence of shared miRNA-binding sites in lncRNAs and mRNAs on gene expression I have modelled a competing endogenous RNA (ceRNA) network, using predicted and experimentally validated miRNA-binding sites shared between lncRNAs that are differentially expressed in mouse iron overload models and mRNAs involved in iron homeostasis.

Figure 4.6 shows a total of 630 interactions that are predicted between miRNA-mRNA and miRNA-lncRNA for 7-mer miRNA seed sequences. The interaction matrix consists of 67 miRNAs expressed in mouse liver, 57 mRNAs and 42 differentially expressed lncRNAs (genomic alignments of 10 lncRNAs could not be generated from Ensembl and therefore were not used for the analysis). For the interaction matrix with 6-mer miRNA seed, refer to the appendix (Figure A2). The interaction matrix shows predicted miRNA interactions, but it is not straightforward to interpret the competing interactions between mRNA and lncRNA using this layout. Therefore I have depicted the competing interactions using a network interaction layout comprising nodes and edges (Figure 4.7).

57 mouse protein-coding genes that have been previously associated with iron homeostasis, were selected to model the ceRNA network in order to focus on genes that play an essential role in iron metabolism. It is possible to model the ceRNA network using differentially expressed mRNAs, but the scale of the network, involving 418 mRNA nodes, would be too dense to navigate and identify regulatory mechanism of genes that are relevant to iron homeostasis.

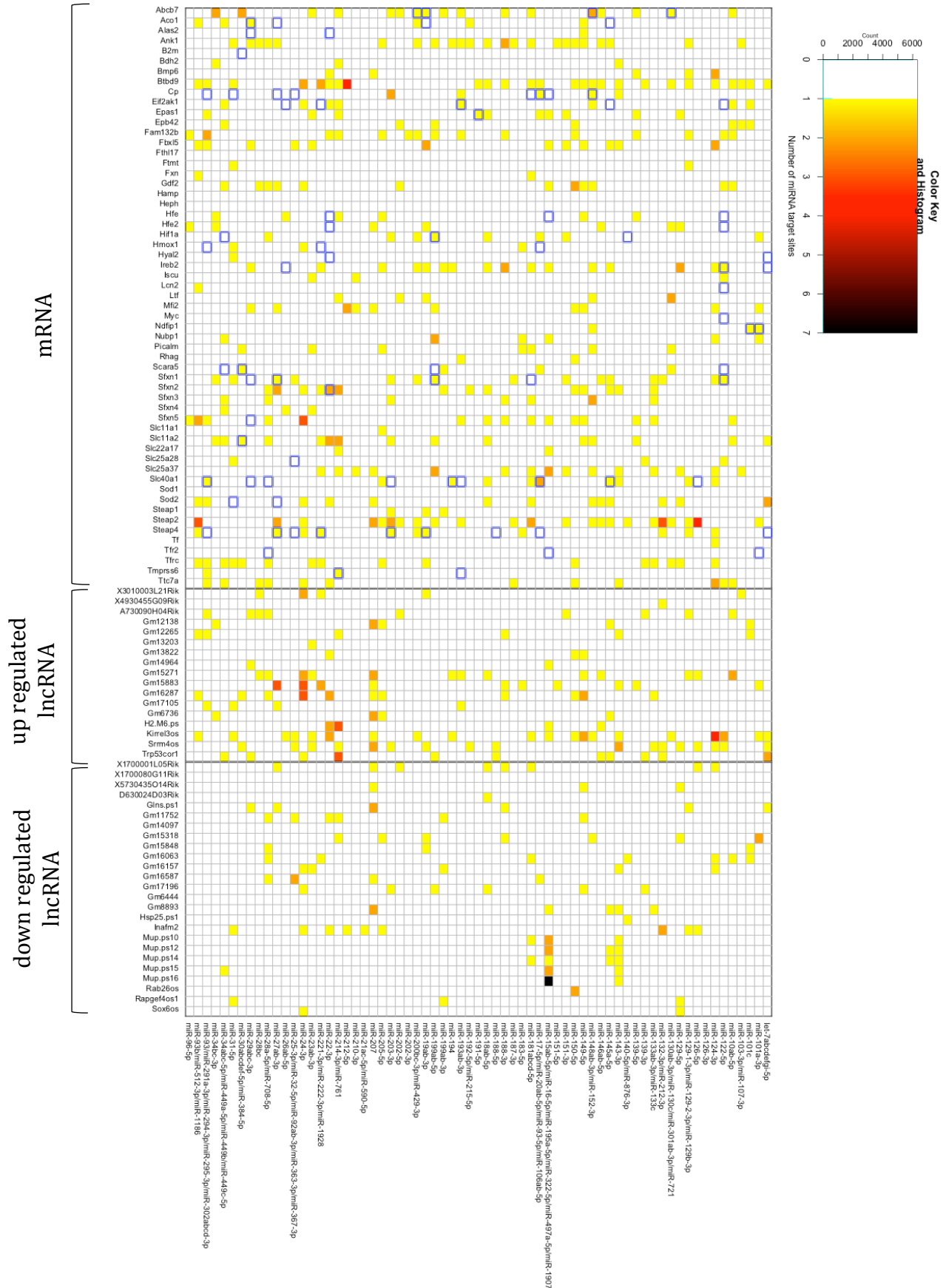


Figure 4.6 All predicted miRNA binding sites (7-mer seed sequence) on mRNAs and lncRNAs. Boxes with blue borders are experimentally validated miRNA targets from TarBase v7.0.

For example a ceRNA network developed using 630 predicted miRNA targets for a 7-mer seed sequence length is shown in Figure 4.7 to illustrate the dense interactions within the network. This network has 407 edges that connect 60 miRNAs with 50 mRNAs. Including all 418 differentially expressed mRNAs would not only make this network very dense and difficult to visualise but also dilute away regulatory interactions that could be important for hemochromatosis. Therefore, I have used an approach to reduce the background noise by filtering down interactions that include only experimentally validated interactions between miRNAs and iron homeostasis regulating mRNAs and predicted conserved binding sites between miRNAs and differentially expressed lncRNAs. Figure 4.8 shows a filtered-down ceRNA model with shared miRNA interactions between mRNA and lncRNAs; most miRNAs target more than one mRNA or lncRNA, and a single mRNA or lncRNA is targeted by more than one miRNA, forming a tightly regulated many-to-many interaction network. Only those miRNAs that target both mRNA and lncRNA are shown here, other miRNAs that only targeted either mRNA or lncRNA have been omitted in the network, since they do not form a part of the ceRNA hypothesis. The expression of mRNAs in the ceRNA network is regulated by the relative abundance of lncRNAs (Salmena et al., 2011) and the strength of regulation depends on the number of miRNA binding sites on these targets. Some of the strongly targeted protein-coding genes include *Abcb7*, *Steap4*, *Slc40a1* (Ferroportin-1) and *Sfxn1*.

The mitochondrial ATP-binding cassette sub-family B member 7 (*Abcb7*) is an exporter of mitochondrial Fe-S cluster proteins and is essential for the biogenesis of cytosolic Fe-S proteins (Pondarre et al., 2006). Mutations in *Abcb7* have been linked to cause sideroblastic anaemia, which causes mitochondrial iron deposition (Bekri et al., 2000) and most importantly knockdown of *Abcb7* results in mitochondrial iron overload and cellular iron deficiency (Cavadini et al., 2007). Another important protein targeted is the metalloreductase *Steap4*. The *Steap* family of metalloreductases stimulate cellular uptake of iron and copper by reducing iron from ferric (Fe^{3+}) to ferrous (Fe^{2+}) and copper from cupric (Cu^{2+}) to cuprous (Cu^{1+}) (Ohgami et al., 2006). One of the main pathological symptoms of hemochromatosis is the excessive deposition of iron in

mRNA

miRNA

lncRNA

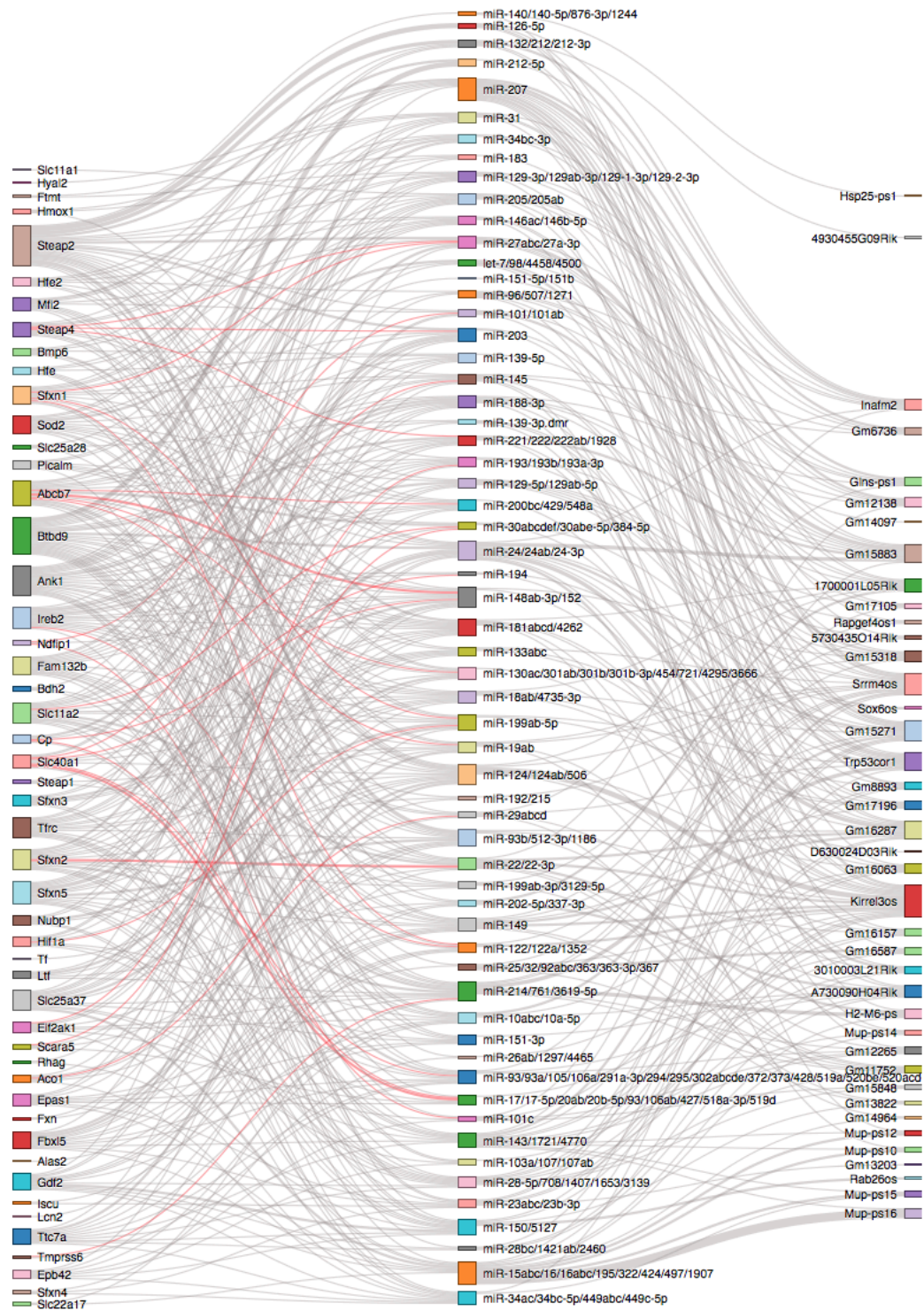


Figure 4.7 The ceRNA network for iron overload mouse model Fpn-C326S. Protein-coding transcripts associated with iron homeostasis are on the left, miRNAs expressed in mouse liver are in the middle and differentially expressed lncRNAs are on the right. The edges represent 7-mer seed sequence interactions with mRNA and lncRNAs and the edge thickness is proportional to the number of miRNA-binding sites. Red lines are predicted interactions that have also been validated in mouse liver tissue as annotated by TarBase v7.0, gray lines are predicted interactions.

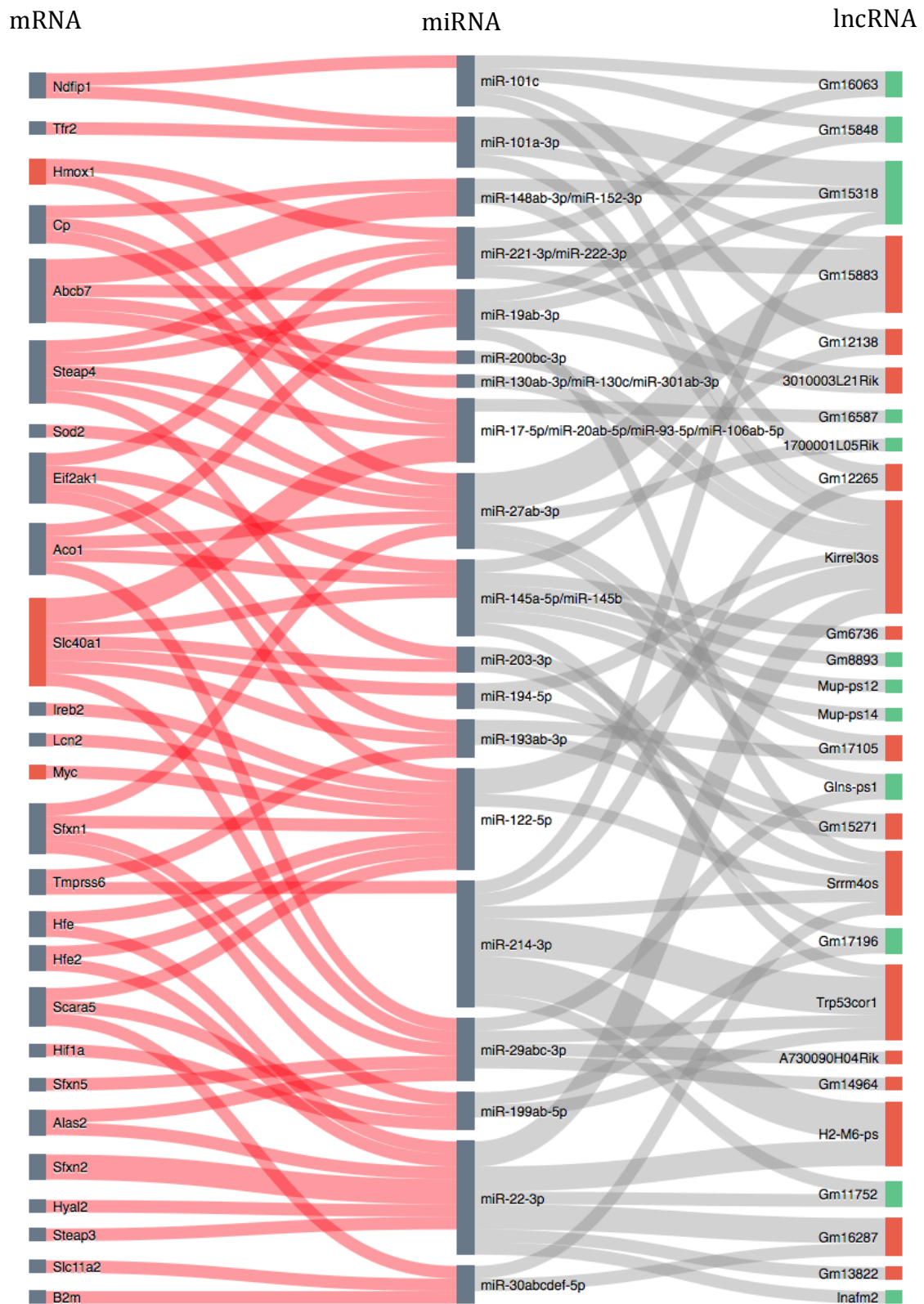


Figure 4.8 Filtered down ceRNA network. Edges in red are experimentally validated interactions from TarBase v7.0, whereas edges in gray are predicted. Experimentally validated interactions that were not predicted are also included. Transcripts with red nodes are over expressed, green are under expressed and gray nodes are not differentially expressed in Fpn-C326S iron overload mouse model.

tissues such as liver and heart (Andrews, 1999). The reduction of iron by Steap2, Steap3, and Steap4 and the stimulation of increased uptake of free non-transferrin-bound iron are thought to influence hemochromatosis (Ohgami et al., 2006). Slc40a1 (ferroportin) is an important intestinal iron exporter and mutations in ferroportin leading to resistance for hepcidin is one of the main causes of type 4 hemochromatosis (Njajou et al., 2001). Post-transcriptionally ferroportin is regulated by iron regulatory proteins (Muckenthaler et al., 2008) and post-translationally by hepcidin (De Domenico et al., 2007). Other protein-coding genes in the ceRNA network include those that code for mitochondrial iron transporter Slc11a2, hereditary hemochromatosis proteins Hfe, homojuvelin Hfe2, translation initiation factor Eif2ak1 and the iron sensor aconitate hydratase Aco1.

Among the lncRNAs Gm15318, Gm15883, Kirrel3os, Srrm4os, Trp53cor1 and H2-M6-pseudogene have more than one binding sites for the same miRNAs. Unlike protein-coding genes, little functional information is available for these lncRNAs to directly implicate their role in regulating gene expression. One of the aims of this project is to functionally annotate these differentially expressed lncRNAs by studying their regulatory interactions using the ceRNA network. LncRNA Gm15318 (Ensembl: ENSMUSG00000086010) is down regulated in iron overload condition (fold change of 0.64) compared to the wild type mouse. It is antisense to and overlaps the second and third exons of protein coding gene trefoil factor 1 (Tff1) on chromosome 17 and is upstream of Tff2 and Tff3. Gm15318 shares synteny with human chromosome 21, but does not have any homologues. Tff1 expression is induced in the gut of iron-deprived rats and has been suggested to take part in increased iron absorption (Collins, 2006). The opposite strand (antisense) of Kin of IRRE like protein-3 (Kirrel3-os, Ensembl: ENSMUSG00000053889) lies within the intronic region of Kirrel3. It is one of the strongly targeted lncRNAs and has more than one conserved binding sites for many miRNAs, including miR-122. The role of Kirrel3 in iron homeostasis is not known. The intergenic lncRNA tumor protein p53 pathway corepressor 1 (Trp53cor1 or lincRNA-p21, Ensembl: ENSMUSG00000085912) (fold change of 1.93) is a well-studied lncRNA in mouse liver, which takes part in regulating

expression of nearby protein-coding genes (Recio et al., 2013; Yoon et al., 2012). Trp53cor1 expression is induced by p53; it acts as a downstream repressor in the p53 transcriptional response and plays a role in triggering apoptosis (Huarte et al., 2010). Trp53cor1 associates with DNA-binding protein hnRNP-K resulting in transcriptional repression at specific genomic loci (Huarte et al., 2010). Trp53cor1 also represses translation of its target genes by associating with translational repressor RCK (Yoon et al., 2012). Inhibition of Trp53cor1 results in affecting the expression of hundreds of genes that are normally repressed by p53 in mouse embryonic fibroblasts (Huarte et al., 2010).

MiRNA families that target the largest number of transcripts within the ceRNA network include miR-27, miR-145, miR-122, miR-214, miR-29 and miR-22. MiR-122 targets hemochromatosis protein Hfe and Hfe2, miR-145 and miR-22 regulate expression of transferrin receptor 1 (TfR1) and miR-214 targets lactoferrin (Yujing Li, 2013). Considering mutual interaction sites in Figure 4.8, it can be seen that a few miRNAs target more mRNAs than non-coding RNAs and vice-versa. For example miR-17 family and miR-122 target more mRNAs than lncRNAs, but miR-214-3p targets more lncRNAs than mRNAs. MiR-17-5p targets protein-coding genes heme oxygenase 1 (Hmox1), ceruloplasmin (Cp), Steap4 and ferroportin (Slc40a1), but has a competing binding site only on one of the down regulated lncRNAs Gm16587. Likewise, targets of miR-122-5p include eight protein-coding genes and two upregulated lncRNA targets. MiR-17-5p has less competition from lncRNAs and is free to regulate the expression of its target protein-coding genes, but the two over expressed lncRNAs, Kirrel3os and Srrm4os, could sponge away miR-122-5p leading to an unregulated expression of its targeted mRNA transcripts. The targets of miR-214-3p includes more lncRNA targets than mRNAs and most of these lncRNA targets are over expressed, which indicates that miR-214-3p has minimal impact on the regulation of expression of its targeted protein-coding transcripts including transmembrane serine protease Tmprss6.

It is evident from the ceRNA network that the regulation of gene expression is not straightforward; the regulation of mRNA expression depends on the relative

concentration of all three classes of transcripts involved – mRNA, miRNA and lncRNA. Since the expression of mRNAs are redundantly regulated, sponging away one or a family of miRNAs by one or a few lncRNAs could still result in a regulatory effect by other miRNAs.

4.3.5 Co-expression of sense-antisense mRNA-lncRNA pairs

To gain further insights into the role of lncRNAs in regulating gene expression, I investigate the relation between sense-antisense mRNA-lncRNA pairs expressed in Fpn-Trp iron deficient mouse model. LncRNAs that are in the antisense orientation to the protein-coding genes could regulate gene expression by forming complementary base pairs with mRNAs. Antisense lncRNAs, also called natural antisense transcripts, have been shown to regulate their sense mRNAs (Katayama et al., 2005; Pelechano and Steinmetz, 2013), which include inhibition of splicing of neuroblastoma MYC, ErbA and ZEB2 mRNAs (Beltran et al., 2008; Krystal et al., 1990; Munroe and Lazar, 1991), forming imperfect complementary base pairs and inhibiting translation of TP53 (Abdelmohsen et al., 2014), silencing DHRS4 gene cluster both in *cis* and *trans* by physically interacting with epigenetic modifiers (Li et al., 2012) and also promoting translation of protein PHO1;2 (Jabnoute et al., 2013).

In this dataset 469 lncRNAs were found to overlap protein-coding loci and were defined as antisense RNAs and considered for analysis. Further, I grouped these sense-antisense pairs based on the genomic regulatory features the antisense lncRNAs are associated with such as enhancer associated, bi-directional promoter associated or transcription factor associated, and others (see method section 4.2.7, Figure 4.9). Among the ncRNA-mRNA sense-antisense pairs, nearly three quarters of ncRNAs (359, 76.54%) associate with (overlap) various genomic regulatory features such as promoters, enhancers, transcription factor binding sites, CTCF binding sites, etc., and the rest (110, 23.45%) do not overlap any regulatory regions. A majority of the antisense lncRNAs that overlap regulatory regions are associated with bi-directional promoters (134, 28.57%) followed by no regulatory feature (110, 23.45%) and non bi-directional

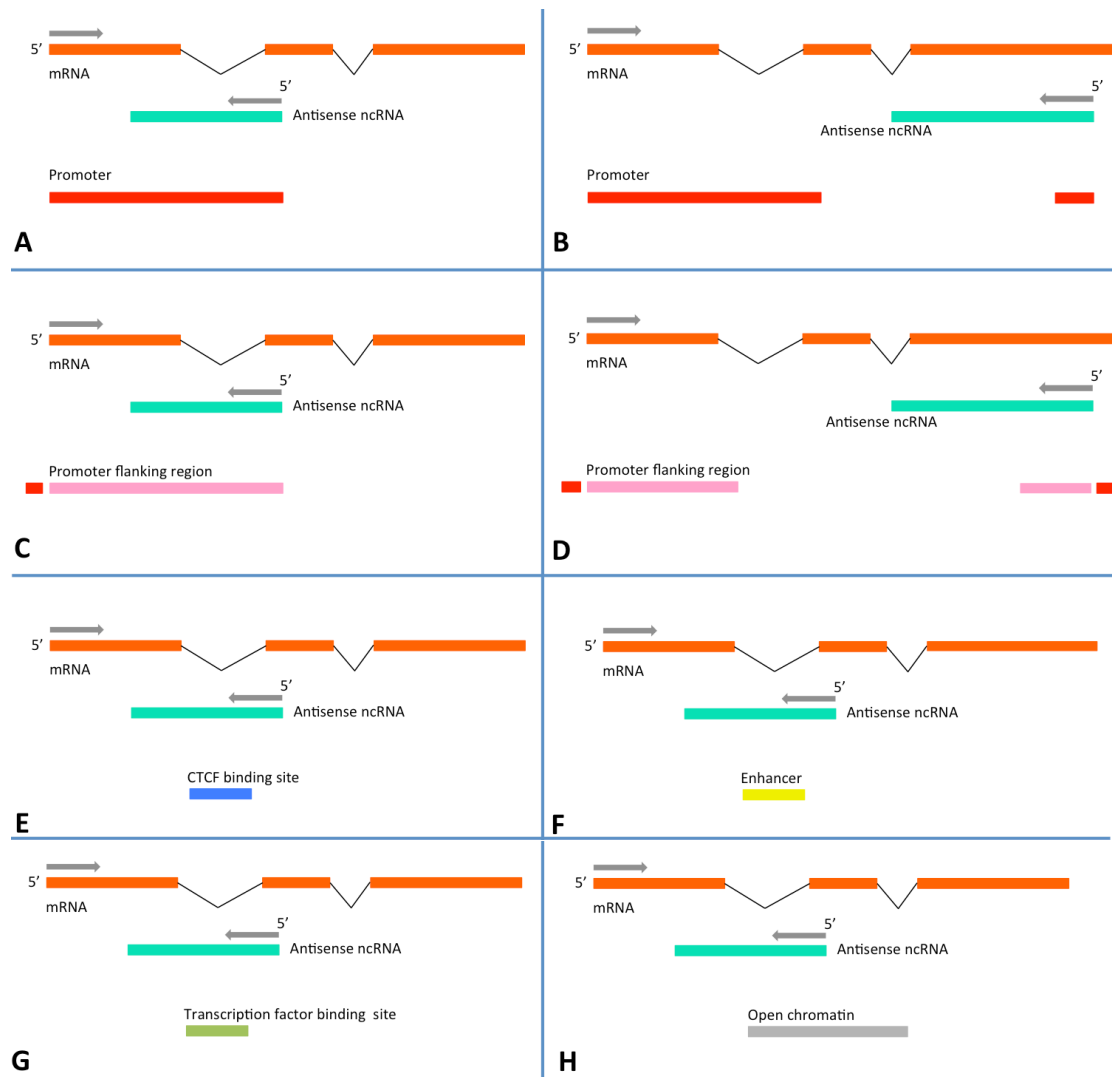


Figure 4.9 Schematic representations of sense-antisense mRNA-lncRNA pairs associated with various genomic regulatory features. (A) Bi-directional promoter associated, (B) Non-bidirectional promoter associated, (C) Bidirectional promoter flanking region associated, (D) Non-bidirectional promoter flanking region associated, (E) CTCF binding site associated, (F) Enhancer associated, (G) Transcription factor (TF) binding site associated and (H) Open chromatin associated.

Expression correlation of FPN-Trp antisense ncRNA v/s sense mRNA

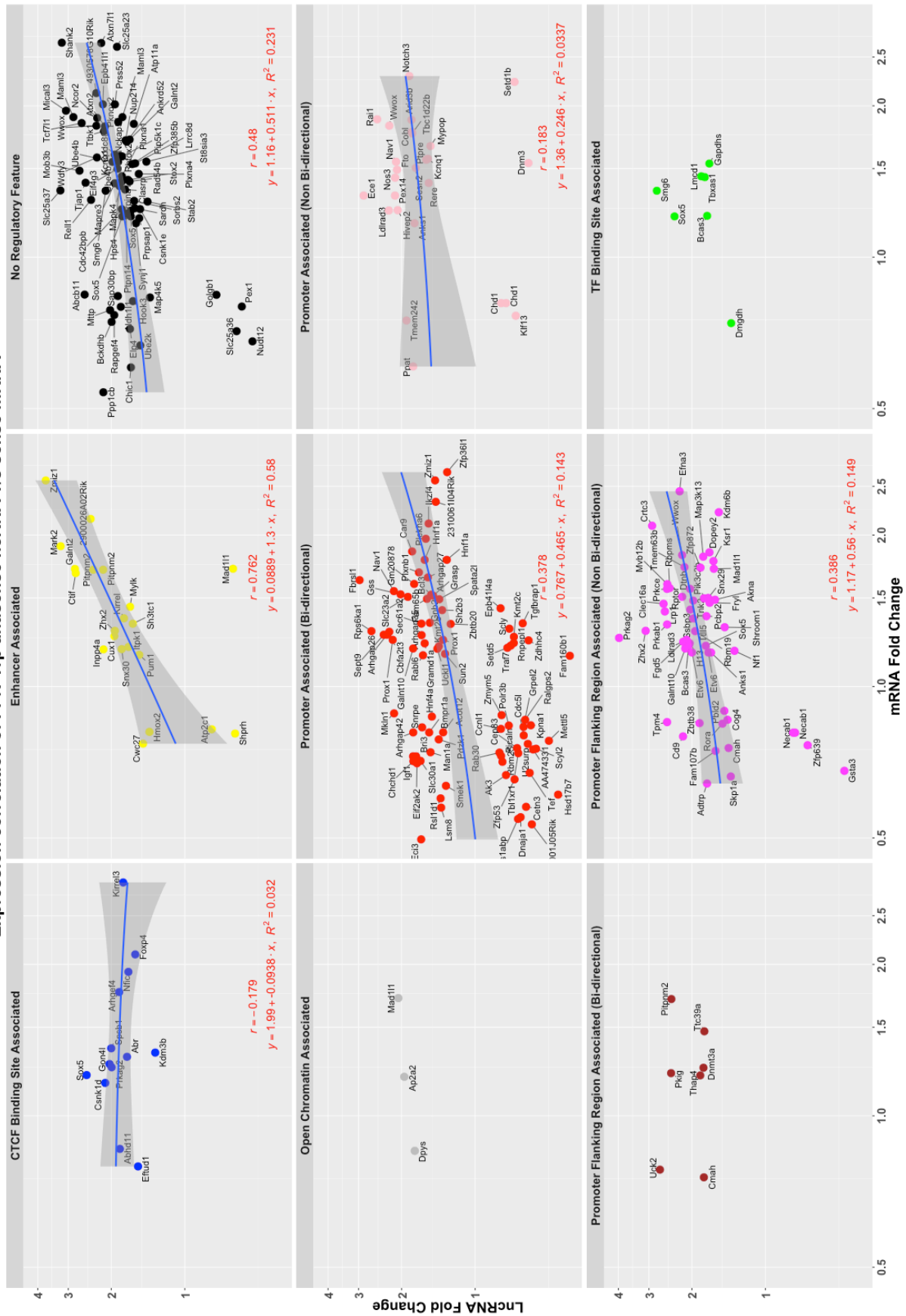


Figure 4.10 Correlation of gene expression values of sense-antisense mRNA-lncRNA transcripts. Sample sizes associated with Open chromatin, Promoter flanking region (Bi-directional) and TF binding site are small to compute statistically significant correlations.

promoter flanking regions (92, 19.61%). Among these 469 sense-antisense mRNA-lncRNA pairs, expression values of 288 mRNAs were experimentally captured in this dataset and this set of 288 sense-antisense mRNA-lncRNA pairs were used for further analysis. Figure 4.10 shows the correlation in the gene expression values of these groups of sense-antisense pairs. The enhancer associated sense-antisense pairs show a strong positive correlation (Pearson correlation coefficient, $r = 0.76$) and followed by non-regulatory feature associated pairs ($r = 0.48$), bidirectional promoter associated pairs ($r = 0.37$), non-bidirectional promoter flanking region associated pairs ($r = 0.38$) and transcription factor associated pairs ($r = 0.31$). Enhancer associated antisense RNAs are transcribed from canonical RNA genes whose exonic regions overlap an enhancer element. These enhancer associated antisense RNAs are different from the recently described class of ncRNAs termed enhancer-derived RNAs or eRNAs (Li et al., 2016), which are transcripts transcribed from intergenic or intragenic enhancers and are largely non-polyadenylated (De Santa et al., 2010; Kim et al., 2010; Ren, 2010). Enhancer RNAs are pervasively transcribed and are involved in regulating expression of their cognate mRNAs and have been studied in detail (Li et al., 2016) but very little is known about enhancer associated antisense RNAs. Since both enhancer associated antisense RNAs and eRNAs comprise enhancer regulatory sequences, they could perform similar functional roles. Several mechanisms have been put forth through which eRNAs are thought to regulate transcription of protein-coding genes in the vicinity – eRNAs positively influence enhancer-promoter looping and gene transcription, they act as scaffolds to bind transcription factors at enhancers, to bind and inhibit transcriptional repressors or act in *trans* by translocation to distal sites (Li et al., 2016). eRNAs largely promote gene activation and their knock down have been shown to result in downregulation of cognate coding genes (Li et al., 2016). Similar to eRNAs, enhancer-associated antisense RNAs have also indicated a positive co-regulated expression with their sense transcripts (Onodera et al., 2012).

Among 288 sense-antisense pairs, I discuss a few mRNA-lncRNA pairs that may play a regulatory role in iron homeostasis (Table 4.3). Mitoferrin-1 (Slc25a37) is

an important mitochondrial iron importer expressed in foetal and adult hematopoietic tissues and is important for erythroid iron assimilation and heme biosynthesis (Shaw et al., 2006). Defects in mitoferrin-1 impair iron import, synthesis of heme and Fe-S cluster or storage of mitochondrial ferritin (Chen and Paw, 2012). The spliced antisense lncRNA Gm27222 (Ensembl: ENSMUSG00000098248) does not overlap any genomic regulatory features. It lies within the intronic region of mitoferrin-1 and does not form an RNA duplex with the mature mitoferrin-1 mRNA. Like other intronic antisense lncRNAs (Louro et al., 2009), I predict that Gm27222 may possibly regulate gene expression through transcriptional interference by interacting with the promoter region of mitoferrin-1 through imperfect base pairing.

Transcription of non-coding antisense transcripts by bi-directional promoters is relatively widespread (Seila et al., 2008; Wei et al., 2011). Transcription of antisense ncRNA pairs were thought to result in transcriptional gene silencing by forming complementary base pairs with sense mRNA or DNA or by competing for the same pool of general transcription factors (Villegas and Zaphiropoulos, 2015; Wei et al., 2011). However many studies have shown a positive regulatory influence of antisense lncRNA on mRNA expression (Beltran et al., 2008; Katayama et al., 2005), including those that are transcribed from bidirectional promoters (Uesaka et al., 2014). Promoter derived antisense lncRNAs have been shown to act in *cis* promoting sense mRNA expression through sequence specific DNA demethylation (Imamura et al., 2004; Tomikawa et al., 2011) or by chromatin remodelling through displacing positioned nucleosomes (Wei et al., 2011).

The sense-antisense pair lncRNA Gm17110 (Ensembl: ENSMUSG00000090779) and the bone morphogenetic protein receptor type-1A (Bmpr1a) mRNA, are bidirectionally transcribed by a common promoter. Gm17110 is antisense to Bmpr1a and completely overlaps its first exon. Bmpr1a plays a central role in signal transduction and expression of hepcidin; binding of Bmp to the serine/threonine kinase receptor Bmpr1a results in downstream signalling and expression of hepcidin (Babitt et al., 2007; Mayeur et al., 2014). There are no

Antisense ncRNA	ncRNA fold change	Sense mRNA	mRNA description	mRNA fold change	Genomic regulatory feature	Reference
Gm27222	3.231	Slc25a37	Mitoferrin-1	1.357	No regulatory feature	(Chen and Paw, 2012)
Gm17110	1.542	Bmpr1a	Bone morphogenetic protein receptor, type 1A	0.81	Promoter associated (bi-directional)	(Babitt et al., 2007)
Hnf4aos	1.503	Hnf4a	Hepatic nuclear factor 4, alpha	0.8713	Promoter associated (bi-directional)	(Matsuo et al., 2015)
Igf1os	1.795	Igf1	Insulin-like growth factor 1	0.726	Promoter associated (bi-directional)	(Ackerman and Gems, 2012)
2310010J17 Rik	0.782	Picalm	phosphatidylinositol binding clathrin assembly protein	0.727	Promoter associated (bi-directional)	(Scotland et al., 2012)

Table 4.3 Gene expression correlations between sense-antisense mRNA-lncRNA pairs associated with iron homeostasis

reported evidence of regulatory interactions between Gm17110 and Bmpr1a. However, like other promoter-associated antisense lncRNAs (Imamura et al., 2004; Tomikawa et al., 2011), due to its positional overlap with the promoter and sequence complementarity with the coding region, I predict that Gm17110 regulates mRNA expression by either influencing transcription through interacting with the promoter, or translation by forming an RNA-duplex with the mRNA. Other bidirectional promoter associated sense-antisense mRNA-lncRNA pairs such as Hnf4a-Hnf4aos, Igf1-Igf1os and Picalm-2310010J17Rik might also regulate gene expression through a similar mechanism.

4.3.6 Correlation of expression of lincRNAs and adjacent mRNAs

The long intervening noncoding RNAs (lincRNAs) do not overlap any protein-coding loci. These transcripts are thought to act on either protein-coding genes that are found in proximity to the lincRNA or on distal protein-coding genes. The functions of distal or *trans* acting lincRNAs is independent of the site of transcription and their targets could be present elsewhere in the cell, on the other hand the functions of proximal targeting lincRNAs or *cis* acting lincRNAs, is transcription site dependent and their direct targets are found in the vicinity (Ulitsky and Bartel, 2013). Most lincRNAs are adjacent to and are found within 10 Kb of protein-coding genes and are thought to play a role in regulating gene expression of their nearby protein-coding genes (Ponjavic et al., 2009; Ulitsky and Bartel, 2013). In order to investigate any such regulatory interactions of lincRNAs I have analysed the gene expression correlation between the differentially expressed lincRNAs in Fpn-Trp mouse model and their neighbouring protein-coding genes.

Upstream and downstream protein-coding genes from both strands of lincRNA transcription start sites were studied. Among 118 lincRNAs, proximal protein-coding genes were identified for 98 lincRNAs and with a gene expression value in the dataset. The gene expression correlations of lincRNA-adjacent mRNA pairs are shown in Figure 4.11. I observe a positive correlation in gene expression between lincRNAs and the proximal protein-coding genes (Pearson correlation

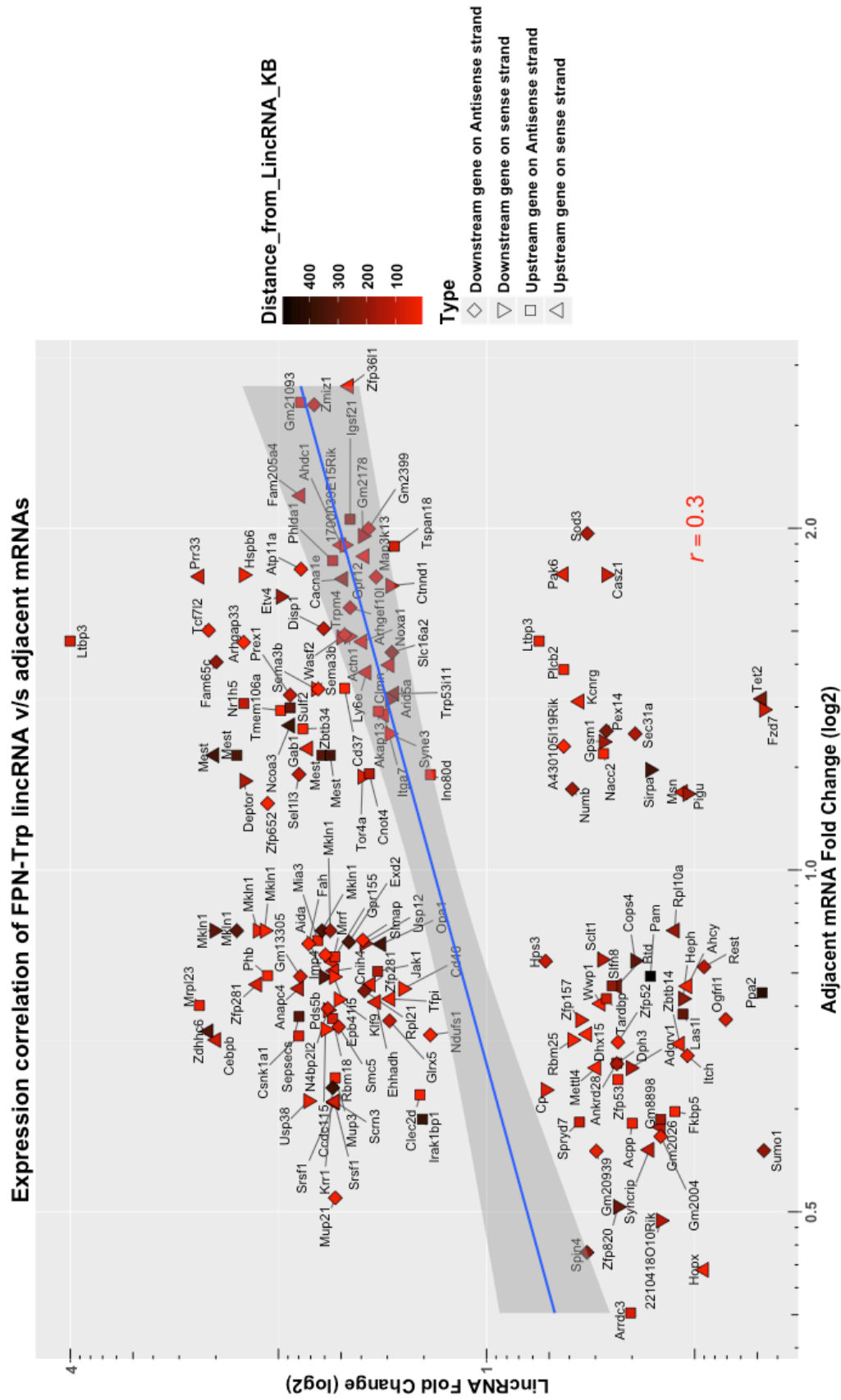


Figure 4.11 Correlation of gene expression values of lincRNA-adjacent mRNA transcripts.

coefficient, $r = 0.3$) (Figure 4.11), which is similar, but weak in comparison, to that observed for most groups of sense-antisense mRNA-lincRNA pairs (Figure 4.10). Investigation of the local protein-coding neighbourhood of lincRNAs identified a few mRNAs involved in maintaining iron homeostasis, which include hephaestin (Heph), ceruloplasmin (Cp) and Cebpb within 0.2 Mb vicinity of lincRNAs (Table 4.4). Not much is known about the differentially expressed lincRNAs found nearby these protein-coding genes. Heph is found 211 Kb downstream of lincRNA F630028010Rik (Ensembl: ENSMUSG00000078122) on the sense strand. Heph is a copper-dependant transmembrane ferroxidase, responsible for the uptake of dietary iron from the intestinal enterocytes and oxidises ferrous to ferric before releasing it into circulation (Vashchenko and Macgillivray, 2012; Vulpe et al., 1999). Heph associates or interacts with ferroportin-1 and function together in export of iron from the intestinal cells (Han and Kim, 2007). Decreased Heph activity and low Heph expression level have been linked to systemic iron deficiency in mice (Chen et al., 2006).

Ceruloplasmin (Cp), a ferroxidase, is present 58 Kb downstream of lincRNA 4632415L05Rik (Ensembl: ENSMUSG00000048106). The plasma Cp is homologous to Heph and has a distinct mechanism to Heph, wherein it regulates iron export from tissues stores (Jiang et al., 2015) and intestinal cells (Cherukuri et al., 2005). Defects in Cp causes iron overload in the liver, brain and kidney (Jiang et al., 2016; Jiang et al., 2015; Kono et al., 2006). Interestingly the gene expression levels of both Heph (fold change: 0.77) and Cp (fold change: 0.64) are down regulated in the Fpn-Trp mouse model along with their upstream lincRNAs when compared to wild-type mice, which might suggest a possible role of these lincRNAs in regulatory the expression of Heph and Cp in iron deficient mice.

The gene encoding transcription factor Cebpb is 92 Kb upstream of lincRNA 9230111E07Rik (Ensembl: ENSMUSG00000087624). The Cebpb expression is down regulated (fold change: 0.71), however the expression of its downstream lincRNA 9230111E07Rik is over two-fold upregulated. Cebpb regulates transcription of the peptide hormone hepcidin (Hamp) (Sow et al., 2009), which

LincRNA	LincRNA fold change	Adjacent mRNA	mRNA fold change	mRNA description	Gene neighbourhood	Reference
F630028010Rik	0.519	Heph	0.770	Hephaestin	Sense strand; Downstream	(Vulpe et al., 1999)
4632415L05Rik	0.080	Cp	0.640	Ceruloplasmin	Sense strand; Downstream	(Patel et al., 2002)
9230111E07Rik	2.460	Cebpb	0.708	CCAAT/enhancer-binding protein beta	Sense strand; Upstream	(Sow et al., 2009)

Table 4.4 Gene expression correlations between lincRNA and adjacent mRNA pairs associated with iron homeostasis.

in turn regulates activity of ferroportin. Decreased expression of *Cebpb* contributes to low levels of hepcidin gene expression in mouse liver (Shpyleva et al., 2011).

4.3.7 Gene ontology enrichment

Apart from the protein-coding genes immediately upstream and downstream to lincRNAs, I also investigated protein-coding neighbourhood within 1 MB of both antisense and lincRNAs and carried out enrichment analysis for functions in iron metabolism, to identify local gene clusters that regulate iron homeostasis. Among 1608 protein-coding genes in the neighbourhood of lincRNAs expressed in iron overload model *Fpn-C326S*, 29 protein-coding genes were identified that are involved in iron metabolism. The enrichment of neighbouring genes for functions directly associated in maintaining iron ion homeostasis identified 4 protein-coding genes within 0.5 Mb and 6 protein-coding genes within 1 Mb of ncRNAs. The enrichment of iron ion homeostasis related genes within the neighbourhood of lincRNAs are not significant (Chi-squared test, p-value: 0.947). Similar investigation of protein-coding neighbourhood of lincRNAs in iron deficient mouse model *Fpn-Trp*, did not show significant enrichment of iron ion homeostasis genes near lincRNAs compared to other protein-coding genes (Chi-squared test, p-value: 0.553). Among 6883 protein coding genes, 89 iron metabolism related genes were found in the neighbourhood, out of which 15 genes were found with 0.5 Mb and 25 genes were found within 1 Mb, which directly associate with maintaining iron ion homeostasis. These genes include *Hamp*, *Hamp2*, *Heph*, *Cp*, *Smad4*, *Tfr2*, *Tfrc* and others. For a complete list of genes, involved in iron metabolism, nearby lincRNAs expressed in *Fpn-C326S* and *Fpn-Trp* mice models see appendix (Tables A3, A4).

4.4 Conclusion

lincRNAs form a major fraction of the transcriptome; they are dynamically expressed, alternatively spliced, and associate with chromatin of actively transcribed genes (Geisler and Collier, 2013; Mattick, 2009). Although lincRNAs

are ubiquitously found and differentially expressed in various non-homeostatic conditions, very little is known about a large number of them regarding their evolution and functions. One view is that most annotated lncRNAs are non-functional and are mere products of non-specific transcription and do not offer any functional advantage (Struhl, 2007), but growing evidences indicate that lncRNAs are bona fide transcripts and involved in important biological functions (Santosh et al., 2015; Wilusz et al., 2009).

In this chapter I have explored various methods to understand functions and regulation of lncRNAs expressed in two different mouse models of iron homeostasis related to hereditary hemochromatosis. Identification of lncRNA homologues through sequence conservation indicated little similarity with human lncRNA transcripts, but through syntenic analysis a few homologues of mouse lncRNA were identified in humans. A pipeline was developed to predict conserved miRNA binding sites in lncRNA and mRNAs, which showed that the target sites of miRNAs were less conserved among lncRNAs compared to the sites within 3'UTRs of mRNAs. The predicted and experimentally validated miRNA binding sites were used to develop a competing endogenous RNA network, which identified shared interactions between mRNA, miRNA and lncRNAs. A few of these interactions, viz. miR-193a-3p/Slc40a1 (Fpn)/Tmprss6, miR-19a-3p/Aco1/Steap4 and miR-122-5p/Sfxn1/Eif2ak1, among others were experimentally tested in the Fpn-C326S mouse model by our collaborators, however they did not observe any significant regulation in their gene expression (unpublished data), suggesting that the predicted regulatory interactions within the ceRNA network may not fully represent physiological regulatory interactions. One of the challenges in interpreting the ceRNA network is the many-to-many interaction between mRNAs, miRNAs and lncRNAs, which makes the regulation redundant, wherein expression of an mRNA is controlled by more than one miRNAs and lncRNAs and therefore elimination of any one miRNA or lncRNA might not fully impact the steady-state expression levels of the mRNA. Another factor that influences ceRNA regulation is the abundance of low-affinity or background MREs (6 nucleotide binding sites and non-canonical sites). It has been shown that the higher numbers of background MREs significantly

contribute to competition and greatly reduce the effect from ceRNA regulation (Denzler et al., 2016).

Further, the antisense lncRNAs and lincRNAs were investigated by studying their expression correlation with the mRNAs that are in sense orientation and in the nearby vicinity respectively. Classification of lncRNAs into sense-antisense pairs and lincRNAs has given some interesting insights into their possible mechanisms for regulating gene expression. A positive correlation in the expression of antisense lncRNAs and mRNAs was observed, which was strong among antisense lncRNAs that overlapped an enhancer element or bidirectional promoter regions. Similar co-expression was also seen between lincRNAs and their immediate upstream and downstream protein-coding genes. The positive co-expression of lincRNAs and their neighbouring mRNAs have been observed in other studies and has been attributed to be a general phenomenon (Ponjavic et al., 2009)

Recent study of lncRNAs have provided insights into their functions with respect to their genomic location and association with genomic regulatory features (Amaral et al., 2016). A number of lncRNA promoters in mouse and human are found conserved in their genomic position relative to orthologous protein coding genes, these syntenic promoter-associated lncRNAs are termed positionally conserved RNAs (pcRNAs) (Amaral et al., 2016). A majority of pcRNAs are transcribed bi-directionally and are shown to be highly tissue specific. These pcRNAs are co-induced and due to the shared transcriptional regulatory elements show co-regulated gene expression with their corresponding protein coding genes. A knockdown of either one member of the pair results in the down-regulation of the other forming a positive feedback-loop and interdependence (Amaral et al., 2016).

Most pcRNAs are also enriched with CTCF binding regions close to their transcription start sites, which take part in genome looping contacts. These pcRNAs are preferentially located at boundaries of such genome looping contact points or loop anchoring points. The pcRNAs whose promoters overlap a loop anchor point are called topological anchor point RNAs (tapRNAs) and share high

local sequence similarity in mouse and human (Amaral et al., 2016). Enhancers are often found at the other end of the loop containing tapRNAs (Amaral et al., 2016). It is suggested that tapRNAs are involved in formation of loop structures and by bringing the contact points of loop together the enhancers are brought in close proximity to the tapRNAs and may result in gene expression regulation (Amaral et al., 2016). Comparison of lincRNAs and sense-antisense mRNA-lincRNA pairs against Amaral et al.'s (Amaral et al., 2016) set of pcRNAs and tapRNAs, identified 19 ncRNAs as pcRNAs and 21 ncRNAs identified as tapRNAs, suggesting a similar functional mechanism of these lincRNAs.

Finally, gene ontology enrichment identified a few iron metabolism associated genes present in close neighbourhood of lincRNAs, but not enriched compared to other protein-coding genes, which suggested that the iron ion homeostasis associated genes are not part of a regulatory cluster controlled by these differentially expressed lincRNAs. Through various analyses I have identified 37 lincRNAs that seem to play a regulatory role in iron metabolism. Figure 4.12 summarises these differentially expressed lincRNAs from both mouse models of iron homeostasis. Six out of seven lincRNAs that were found to be syntenic with human transcripts take part in ceRNA interactions. Interestingly the set of lincRNAs that were predicted to share sequence similarity with human transcripts were not found to be syntenic. Among lincRNAs, Trp53cor1 is involved in ceRNA interactions and some of the antisense lincRNAs that overlap protein-coding genes associated with iron homeostasis were neither syntenic nor observed in ceRNA interactions.

In conclusion, in this chapter I have attempted to infer homology, functions and regulatory mechanisms of lincRNAs expressed in two mouse models of iron homeostasis using an array of different methods which has provided a broad understanding of their functional properties and the complexities in working with them. These lincRNAs provide a rich substrate for understanding the subtleties of regulation of iron homeostasis and could be promising targets that can be further investigated experimentally.

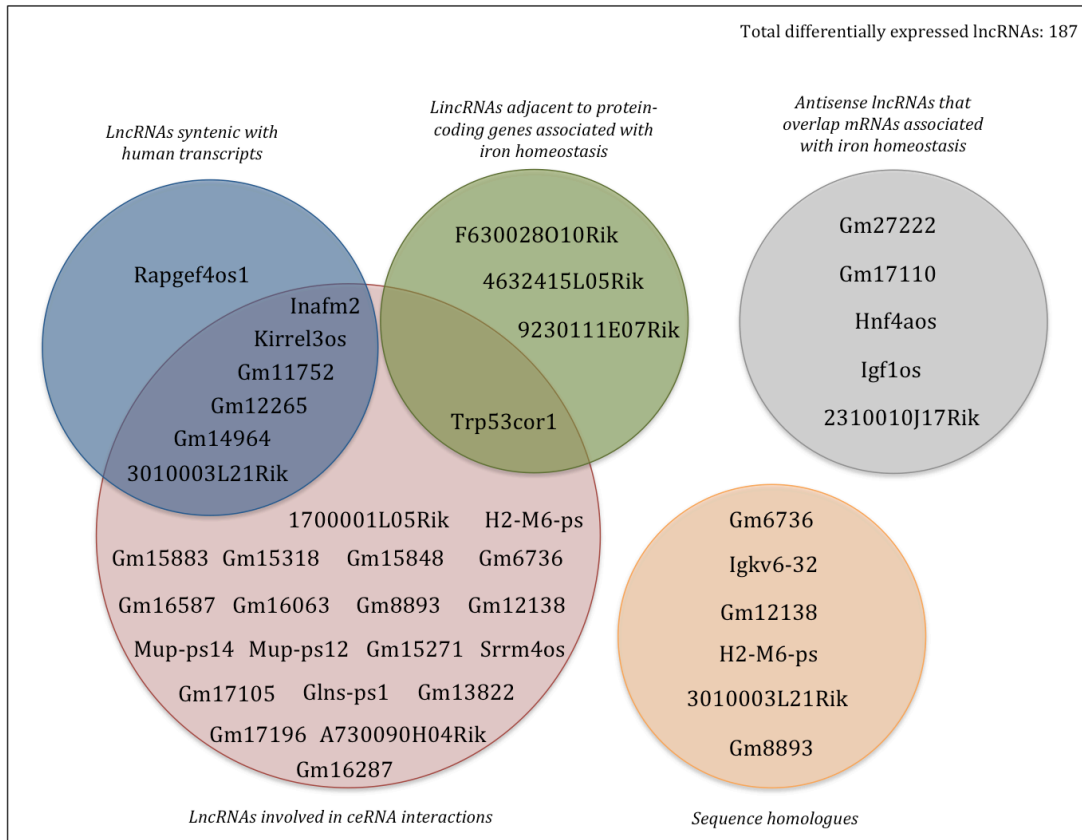


Figure 4.12 Venn diagram showing differentially expressed lncRNAs from mouse models Fpn-C326S and Fpn-Trp that are predicted to regulate iron homeostasis.

4.5 References

Abdelmohsen, K., Panda, A.C., Kang, M.J., Guo, R., Kim, J., Grammatikakis, I., Yoon, J.H., Dudekula, D.B., Noh, J.H., Yang, X., *et al.* (2014). 7SL RNA represses p53 translation by competing with HuR. *Nucleic acids research* 42, 10099-10111.

Ackerman, D., and Gems, D. (2012). Insulin/IGF-1 and hypoxia signaling act in concert to regulate iron homeostasis in *Caenorhabditis elegans*. *PLoS genetics* 8, e1002498.

Aken, B.L., Ayling, S., Barrell, D., Clarke, L., Curwen, V., Fairley, S., Fernandez Banet, J., Billis, K., Garcia Giron, C., Hourlier, T., *et al.* (2016). The Ensembl gene annotation system. *Database : the journal of biological databases and curation* 2016.

Altamura, S., Kessler, R., Grone, H.J., Gretz, N., Hentze, M.W., Galy, B., and Muckenthaler, M.U. (2014). Resistance of ferroportin to hepcidin binding causes exocrine pancreatic failure and fatal iron overload. *Cell metabolism* 20, 359-367.

Amandio, A.R., Necsulea, A., Joye, E., Mascrez, B., and Duboule, D. (2016). Hotair Is Dispensable for Mouse Development. *PLoS genetics* 12, e1006232.

Amaral, P.P., Leonardi, T., Han, N., Vire, E., Gascoigne, D.K., Arias-Carrasco, R., Buscher, M., Zhang, A., Pluchino, S., Maracaja-Coutinho, V., *et al.* (2016). Genomic positional conservation identifies topological anchor point (tap)RNAs linked to developmental loci. *bioRxiv* doi:10.1101/051052.

Andrews, N.C. (1999). Disorders of iron metabolism. *The New England journal of medicine* 341, 1986-1995.

Babitt, J.L., Huang, F.W., Xia, Y., Sidis, Y., Andrews, N.C., and Lin, H.Y. (2007). Modulation of bone morphogenetic protein signaling in vivo regulates systemic iron balance. *The Journal of clinical investigation* 117, 1933-1939.

Bekri, S., Kispal, G., Lange, H., Fitzsimons, E., Tolmie, J., Lill, R., and Bishop, D.F. (2000). Human ABC7 transporter: gene structure and mutation causing X-linked sideroblastic anemia with ataxia with disruption of cytosolic iron-sulfur protein maturation. *Blood* 96, 3256-3264.

Beltran, M., Puig, I., Pena, C., Garcia, J.M., Alvarez, A.B., Pena, R., Bonilla, F., and de Herreros, A.G. (2008). A natural antisense transcript regulates Zeb2/Sip1 gene expression during Snail1-induced epithelial-mesenchymal transition. *Genes & development* 22, 756-769.

Beuvink, I., Kolb, F.A., Budach, W., Garnier, A., Lange, J., Natt, F., Dengler, U., Hall, J., Filipowicz, W., and Weiler, J. (2007). A novel microarray approach reveals new tissue-specific signatures of known and predicted mammalian microRNAs. *Nucleic acids research* 35, e52.

Bomford, A. (2002). Genetics of haemochromatosis. *Lancet* 360, 1673-1681.

Camaschella, C., Roetto, A., Cali, A., De Gobbi, M., Garozzo, G., Carella, M., Majorano, N., Totaro, A., and Gasparini, P. (2000). The gene TFR2 is mutated in a new type of haemochromatosis mapping to 7q22. *Nature genetics* 25, 14-15.

Castoldi, M., Vujic Spasic, M., Altamura, S., Elmen, J., Lindow, M., Kiss, J., Stolte, J., Sparla, R., D'Alessandro, L.A., Klingmuller, U., *et al.* (2011). The liver-specific microRNA miR-122 controls systemic iron homeostasis in mice. *The Journal of clinical investigation* *121*, 1386-1396.

Cavadini, P., Biasiotto, G., Poli, M., Levi, S., Verardi, R., Zanella, I., Derosas, M., Ingrassia, R., Corrado, M., and Arosio, P. (2007). RNA silencing of the mitochondrial ABCB7 transporter in HeLa cells causes an iron-deficient phenotype with mitochondrial iron overload. *Blood* *109*, 3552-3559.

Chen, C., and Paw, B.H. (2012). Cellular and mitochondrial iron homeostasis in vertebrates. *Biochimica et biophysica acta* *1823*, 1459-1467.

Chen, H., Huang, G., Su, T., Gao, H., Attieh, Z.K., McKie, A.T., Anderson, G.J., and Vulpe, C.D. (2006). Decreased hephaestin activity in the intestine of copper-deficient mice causes systemic iron deficiency. *The Journal of nutrition* *136*, 1236-1241.

Cherukuri, S., Potla, R., Sarkar, J., Nurko, S., Harris, Z.L., and Fox, P.L. (2005). Unexpected role of ceruloplasmin in intestinal iron absorption. *Cell metabolism* *2*, 309-319.

Chiyomaru, T., Fukuhara, S., Saini, S., Majid, S., Deng, G., Shahryari, V., Chang, I., Tanaka, Y., Enokida, H., Nakagawa, M., *et al.* (2014). Long non-coding RNA HOTAIR is targeted and regulated by miR-141 in human cancer cells. *The Journal of biological chemistry* *289*, 12550-12565.

Collins, J.F. (2006). Gene chip analyses reveal differential genetic responses to iron deficiency in rat duodenum and jejunum. *Biological research* *39*, 25-37.

De Domenico, I., Ward, D.M., Langelier, C., Vaughn, M.B., Nemeth, E., Sundquist, W.I., Ganz, T., Musci, G., and Kaplan, J. (2007). The molecular mechanism of hepcidin-mediated ferroportin down-regulation. *Molecular biology of the cell* *18*, 2569-2578.

De Santa, F., Barozzi, I., Mietton, F., Ghisletti, S., Polletti, S., Tusi, B.K., Muller, H., Ragoussis, J., Wei, C.L., and Natoli, G. (2010). A large fraction of extragenic RNA pol II transcription sites overlap enhancers. *PLoS biology* *8*, e1000384.

Denzler, R., McGeary, S.E., Title, A.C., Agarwal, V., Bartel, D.P., and Stoffel, M. (2016). Impact of MicroRNA Levels, Target-Site Complementarity, and Cooperativity on Competing Endogenous RNA-Regulated Gene Expression. *Molecular cell* *64*, 565-579.

Diederichs, S. (2014). The four dimensions of noncoding RNA conservation. *Trends in genetics* : *TIG* *30*, 121-123.

Donovan, A., Lima, C.A., Pinkus, J.L., Pinkus, G.S., Zon, L.I., Robine, S., and Andrews, N.C. (2005). The iron exporter ferroportin/Slc40a1 is essential for iron homeostasis. *Cell metabolism* *1*, 191-200.

Eddy, S.R., and Durbin, R. (1994). RNA sequence analysis using covariance models. *Nucleic acids research* *22*, 2079-2088.

Encode, P.C., Birney, E., Stamatoyannopoulos, J.A., Dutta, A., Guigo, R., Gingeras, T.R., Margulies, E.H., Weng, Z., Snyder, M., Dermitzakis, E.T., *et al.* (2007). Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* *447*, 799-816.

Engreitz, J.M., Ollikainen, N., and Guttman, M. (2016). Long non-coding RNAs: spatial amplifiers that control nuclear structure and gene expression. *Nature reviews. Molecular cell biology* *17*, 756-770.

Feder, J.N., Gnirke, A., Thomas, W., Tsuchihashi, Z., Ruddy, D.A., Basava, A., Dormishian, F., Domingo, R., Jr., Ellis, M.C., Fullan, A., *et al.* (1996). A novel MHC class I-like gene is mutated in patients with hereditary haemochromatosis. *Nature genetics* *13*, 399-408.

Fernandes, A., Preza, G.C., Phung, Y., De Domenico, I., Kaplan, J., Ganz, T., and Nemeth, E. (2009). The molecular basis of hepcidin-resistant hereditary hemochromatosis. *Blood* *114*, 437-443.

Filipowicz, W., Bhattacharyya, S.N., and Sonenberg, N. (2008). Mechanisms of post-transcriptional regulation by microRNAs: are the answers in sight? *Nature reviews. Genetics* *9*, 102-114.

Geisler, S., and Collier, J. (2013). RNA in unexpected places: long non-coding RNA functions in diverse cellular contexts. *Nature reviews. Molecular cell biology* *14*, 699-712.

Griffiths-Jones, S., Bateman, A., Marshall, M., Khanna, A., and Eddy, S.R. (2003). Rfam: an RNA family database. *Nucleic acids research* *31*, 439-441.

Gupta, R.A., Shah, N., Wang, K.C., Kim, J., Horlings, H.M., Wong, D.J., Tsai, M.C., Hung, T., Argani, P., Rinn, J.L., *et al.* (2010). Long non-coding RNA HOTAIR reprograms chromatin state to promote cancer metastasis. *Nature* *464*, 1071-1076.

Hajjari, M., Behmanesh, M., Sadeghizadeh, M., and Zeinoddini, M. (2013). Up-regulation of HOTAIR long non-coding RNA in human gastric adenocarcinoma tissues. *Medical oncology* *30*, 670.

Han, O., and Kim, E.Y. (2007). Colocalization of ferroportin-1 with hephaestin on the basolateral membrane of human intestinal absorptive cells. *Journal of cellular biochemistry* *101*, 1000-1010.

Huarte, M., Guttman, M., Feldser, D., Garber, M., Koziol, M.J., Kenzelmann-Broz, D., Khalil, A.M., Zuk, O., Amit, I., Rabani, M., *et al.* (2010). A large intergenic noncoding RNA induced by p53 mediates global gene repression in the p53 response. *Cell* *142*, 409-419.

Imamura, T., Yamamoto, S., Ohgane, J., Hattori, N., Tanaka, S., and Shiota, K. (2004). Non-coding RNA directed DNA demethylation of Sphk1 CpG island. *Biochemical and biophysical research communications* *322*, 593-600.

Iyer, M.K., Niknafs, Y.S., Malik, R., Singhal, U., Sahu, A., Hosono, Y., Barrette, T.R., Prensner, J.R., Evans, J.R., Zhao, S., *et al.* (2015). The landscape of long noncoding RNAs in the human transcriptome. *Nature genetics* *47*, 199-208.

Jabnourne, M., Secco, D., Lecampion, C., Robaglia, C., Shu, Q., and Poirier, Y. (2013). A rice cis-natural antisense RNA acts as a translational enhancer for its cognate mRNA and contributes to phosphate homeostasis and plant fitness. *The Plant cell* *25*, 4166-4182.

Jiang, B., Liu, G., Zheng, J., Chen, M., Maimaitiming, Z., Chen, M., Liu, S., Jiang, R., Fuqua, B.K., Dunaief, J.L., *et al.* (2016). Hephaestin and ceruloplasmin facilitate iron metabolism in the mouse kidney. *Scientific reports* *6*, 39470.

Jiang, R., Hua, C., Wan, Y., Jiang, B., Hu, H., Zheng, J., Fuqua, B.K., Dunaief, J.L., Anderson, G.J., David, S., *et al.* (2015). Hephaestin and ceruloplasmin play distinct but interrelated roles in iron homeostasis in mouse brain. *The Journal of nutrition* *145*, 1003-1009.

Joshi, T., and Xu, D. (2007). Quantitative assessment of relationship between sequence similarity and function similarity. *BMC genomics* *8*, 222.

Katayama, S., Tomaru, Y., Kasukawa, T., Waki, K., Nakanishi, M., Nakamura, M., Nishida, H., Yap, C.C., Suzuki, M., Kawai, J., *et al.* (2005). Antisense transcription in the mammalian transcriptome. *Science* *309*, 1564-1566.

Kato, J., Fujikawa, K., Kanda, M., Fukuda, N., Sasaki, K., Takayama, T., Kobune, M., Takada, K., Takimoto, R., Hamada, H., *et al.* (2001). A mutation, in the iron-responsive element of H ferritin mRNA, causing autosomal dominant iron overload. *American journal of human genetics* *69*, 191-197.

Kaya, K.D., Karakulah, G., Yalciner, C.M., Acar, A.C., and Konu, O. (2011). mESAdb: microRNA expression and sequence analysis database. *Nucleic acids research* *39*, D170-180.

Kim, K., Jutooru, I., Chadalapaka, G., Johnson, G., Frank, J., Burghardt, R., Kim, S., and Safe, S. (2013). HOTAIR is a negative prognostic factor and exhibits pro-oncogenic activity in pancreatic cancer. *Oncogene* *32*, 1616-1625.

Kim, T.K., Hemberg, M., Gray, J.M., Costa, A.M., Bear, D.M., Wu, J., Harmin, D.A., Laptewicz, M., Barbara-Haley, K., Kuersten, S., *et al.* (2010). Widespread transcription at neuronal activity-regulated enhancers. *Nature* *465*, 182-187.

Kong, L., Zhang, Y., Ye, Z.Q., Liu, X.Q., Zhao, S.Q., Wei, L., and Gao, G. (2007). CPC: assess the protein-coding potential of transcripts using sequence features and support vector machine. *Nucleic acids research* *35*, W345-349.

Kono, S., Suzuki, H., Takahashi, K., Takahashi, Y., Shirakawa, K., Murakawa, Y., Yamaguchi, S., and Miyajima, H. (2006). Hepatic iron overload associated with a decreased serum ceruloplasmin level in a novel clinical type of aceruloplasminemia. *Gastroenterology* *131*, 240-245.

Kozomara, A., and Griffiths-Jones, S. (2014). miRBase: annotating high confidence microRNAs using deep sequencing data. *Nucleic acids research* *42*, D68-73.

Krystal, G.W., Armstrong, B.C., and Battey, J.F. (1990). N-myc mRNA forms an RNA-RNA duplex with endogenous antisense transcripts. *Molecular and cellular biology* *10*, 4180-4191.

Li, L., Helms, J.A., and Chang, H.Y. (2016a). Comment on "Hotair Is Dispensable for Mouse Development". *PLoS genetics* *12*, e1006406.

Li, L., Liu, B., Wapinski, O.L., Tsai, M.C., Qu, K., Zhang, J., Carlson, J.C., Lin, M., Fang, F., Gupta, R.A., *et al.* (2013). Targeted disruption of Hotair leads to homeotic transformation and gene derepression. *Cell reports* *5*, 3-12.

Li, Q., Su, Z., Xu, X., Liu, G., Song, X., Wang, R., Sui, X., Liu, T., Chang, X., and Huang, D. (2012). AS1DHRS4, a head-to-head natural antisense transcript, silences the DHRS4 gene cluster in cis and trans. *Proceedings of the National Academy of Sciences of the United States of America* *109*, 14110-14115.

- Li, W., Notani, D., and Rosenfeld, M.G. (2016b). Enhancers as non-coding RNA transcription units: recent insights and future perspectives. *Nature reviews. Genetics* 17, 207-223.
- Lian, Y., Cai, Z., Gong, H., Xue, S., Wu, D., and Wang, K. (2016). HOTTIP: a critical oncogenic long non-coding RNA in human cancers. *Molecular bioSystems* 12, 3247-3253.
- Louro, R., Smirnova, A.S., and Verjovski-Almeida, S. (2009). Long intronic noncoding RNA transcription: expression noise or expression choice? *Genomics* 93, 291-298.
- Matsuo, S., Ogawa, M., Muckenthaler, M.U., Mizui, Y., Sasaki, S., Fujimura, T., Takizawa, M., Ariga, N., Ozaki, H., Sakaguchi, M., *et al.* (2015). Hepatocyte Nuclear Factor 4alpha Controls Iron Metabolism and Regulates Transferrin Receptor 2 in Mouse Liver. *The Journal of biological chemistry* 290, 30855-30865.
- Mattick, J.S. (2009). The genetic signatures of noncoding RNAs. *PLoS genetics* 5, e1000459.
- Mayeur, C., Lohmeyer, L.K., Leyton, P., Kao, S.M., Pappas, A.E., Kolodziej, S.A., Spagnolli, E., Yu, B., Galdos, R.L., Yu, P.B., *et al.* (2014). The type I BMP receptor Alk3 is required for the induction of hepatic hepcidin gene expression by interleukin-6. *Blood* 123, 2261-2268.
- McHugh, C.A., Chen, C.K., Chow, A., Surka, C.F., Tran, C., McDonel, P., Pandya-Jones, A., Blanco, M., Burghard, C., Moradian, A., *et al.* (2015). The Xist lncRNA interacts directly with SHARP to silence transcription through HDAC3. *Nature* 521, 232-236.
- Miranda, K.C., Huynh, T., Tay, Y., Ang, Y.S., Tam, W.L., Thomson, A.M., Lim, B., and Rigoutsos, I. (2006). A pattern-based method for the identification of MicroRNA binding sites and their corresponding heteroduplexes. *Cell* 126, 1203-1217.
- Muckenthaler, M.U. (2014). How mutant HFE causes hereditary hemochromatosis. *Blood* 124, 1212-1213.
- Muckenthaler, M.U., Galy, B., and Hentze, M.W. (2008). Systemic iron homeostasis and the iron-responsive element/iron-regulatory protein (IRE/IRP) regulatory network. *Annual review of nutrition* 28, 197-213.
- Munroe, S.H., and Lazar, M.A. (1991). Inhibition of c-erbA mRNA splicing by a naturally occurring antisense RNA. *The Journal of biological chemistry* 266, 22083-22086.
- Nagano, T., Mitchell, J.A., Sanz, L.A., Pauler, F.M., Ferguson-Smith, A.C., Feil, R., and Fraser, P. (2008). The Air noncoding RNA epigenetically silences transcription by targeting G9a to chromatin. *Science* 322, 1717-1720.
- Nam, J.W., and Bartel, D.P. (2012). Long noncoding RNAs in *C. elegans*. *Genome research* 22, 2529-2540.
- Nawrocki, E.P., Burge, S.W., Bateman, A., Daub, J., Eberhardt, R.Y., Eddy, S.R., Floden, E.W., Gardner, P.P., Jones, T.A., Tate, J., *et al.* (2015). Rfam 12.0: updates to the RNA families database. *Nucleic acids research* 43, D130-137.

Nemeth, E., Tuttle, M.S., Powelson, J., Vaughn, M.B., Donovan, A., Ward, D.M., Ganz, T., and Kaplan, J. (2004). Heparin regulates cellular iron efflux by binding to ferroportin and inducing its internalization. *Science* 306, 2090-2093.

Ng, K., Pullirsch, D., Leeb, M., and Wutz, A. (2007). Xist and the order of silencing. *EMBO reports* 8, 34-39.

Njajou, O.T., Vaessen, N., Joosse, M., Berghuis, B., van Dongen, J.W., Breuning, M.H., Snijders, P.J., Rutten, W.P., Sandkuijl, L.A., Oostra, B.A., *et al.* (2001). A mutation in SLC11A3 is associated with autosomal dominant hemochromatosis. *Nature genetics* 28, 213-214.

Ohgami, R.S., Campagna, D.R., McDonald, A., and Fleming, M.D. (2006). The Steap proteins are metalloreductases. *Blood* 108, 1388-1394.

Onodera, C.S., Underwood, J.G., Katzman, S., Jacobs, F., Greenberg, D., Salama, S.R., and Haussler, D. (2012). Gene isoform specificity through enhancer-associated antisense transcription. *PloS one* 7, e43511.

Pantopoulos, K. (2008). Function of the hemochromatosis protein HFE: Lessons from animal models. *World journal of gastroenterology* 14, 6893-6901.

Pantopoulos, K., Porwal, S.K., Tartakoff, A., and Devireddy, L. (2012). Mechanisms of mammalian iron homeostasis. *Biochemistry* 51, 5705-5724.

Paraskevopoulou, M.D., Vlachos, I.S., and Hatzigeorgiou, A.G. (2016). DIANA-TarBase and DIANA Suite Tools: Studying Experimentally Supported microRNA Targets. *Current protocols in bioinformatics* 55, 12 14 11-12 14 18.

Patel, B.N., Dunn, R.J., Jeong, S.Y., Zhu, Q., Julien, J.P., and David, S. (2002). Ceruloplasmin regulates iron levels in the CNS and prevents free radical injury. *The Journal of neuroscience : the official journal of the Society for Neuroscience* 22, 6578-6586.

Pelechano, V., and Steinmetz, L.M. (2013). Gene regulation by antisense transcription. *Nature reviews. Genetics* 14, 880-893.

Pertea, M. (2012). The human transcriptome: an unfinished story. *Genes* 3, 344-360.

Peter, M.E. (2010). Targeting of mRNAs by multiple miRNAs: the next step. *Oncogene* 29, 2161-2164.

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

Pondarre, C., Antiochos, B.B., Campagna, D.R., Clarke, S.L., Greer, E.L., Deck, K.M., McDonald, A., Han, A.P., Medlock, A., Kutok, J.L., *et al.* (2006). The mitochondrial ATP-binding cassette transporter Abcb7 is essential in mice and participates in cytosolic iron-sulfur cluster biogenesis. *Human molecular genetics* 15, 953-964.

Ponjavic, J., Oliver, P.L., Lunter, G., and Ponting, C.P. (2009). Genomic and transcriptional co-localization of protein-coding and long non-coding RNA pairs in the developing brain. *PLoS genetics* 5, e1000617.

Qiao, B., Sugianto, P., Fung, E., Del-Castillo-Rueda, A., Moran-Jimenez, M.J., Ganz, T., and Nemeth, E. (2012). Hepcidin-induced endocytosis of ferroportin is dependent on ferroportin ubiquitination. *Cell metabolism* 15, 918-924.

Recio, L., Phillips, S.L., Maynor, T., Waters, M., Jackson, A.F., and Yauk, C.L. (2013). Differential expression of long noncoding RNAs in the livers of female B6C3F1 mice exposed to the carcinogen furan. *Toxicological sciences : an official journal of the Society of Toxicology* 135, 369-379.

Ren, B. (2010). Transcription: Enhancers make non-coding RNA. *Nature* 465, 173-174.

Rinn, J.L., Kertesz, M., Wang, J.K., Squazzo, S.L., Xu, X., Bruggmann, S.A., Goodnough, L.H., Helms, J.A., Farnham, P.J., Segal, E., *et al.* (2007). Functional demarcation of active and silent chromatin domains in human HOX loci by noncoding RNAs. *Cell* 129, 1311-1323.

Rivas, E., Clements, J., and Eddy, S.R. (2017). A statistical test for conserved RNA structure shows lack of evidence for structure in lncRNAs. *Nature methods* 14, 45-48.

Roetto, A., Totaro, A., Cazzola, M., Cicilano, M., Bosio, S., D'Ascola, G., Carella, M., Zelante, L., Kelly, A.L., Cox, T.M., *et al.* (1999). Juvenile hemochromatosis locus maps to chromosome 1q. *American journal of human genetics* 64, 1388-1393.

Salmena, L., Poliseno, L., Tay, Y., Kats, L., and Pandolfi, P.P. (2011). A ceRNA hypothesis: the Rosetta Stone of a hidden RNA language? *Cell* 146, 353-358.

Santosh, B., Varshney, A., and Yadava, P.K. (2015). Non-coding RNAs: biological functions and applications. *Cell biochemistry and function* 33, 14-22.

Schorderet, P., and Duboule, D. (2011). Structural and functional differences in the long non-coding RNA hotair in mouse and human. *PLoS genetics* 7, e1002071.

Scotland, P.B., Heath, J.L., Conway, A.E., Porter, N.B., Armstrong, M.B., Walker, J.A., Klebig, M.L., Lavau, C.P., and Wechsler, D.S. (2012). The PICALM protein plays a key role in iron homeostasis and cell proliferation. *PloS one* 7, e44252.

Seila, A.C., Calabrese, J.M., Levine, S.S., Yeo, G.W., Rahl, P.B., Flynn, R.A., Young, R.A., and Sharp, P.A. (2008). Divergent transcription from active promoters. *Science* 322, 1849-1851.

Selleri, L., Bartolomei, M.S., Bickmore, W.A., He, L., Stubbs, L., Reik, W., and Barsh, G.S. (2016). A Hox-Embedded Long Noncoding RNA: Is It All Hot Air? *PLoS genetics* 12, e1006485.

Shaw, G.C., Cope, J.J., Li, L., Corson, K., Hersey, C., Ackermann, G.E., Gwynn, B., Lambert, A.J., Wingert, R.A., Traver, D., *et al.* (2006). Mitoferrin is essential for erythroid iron assimilation. *Nature* 440, 96-100.

Shpyleva, S.I., Muskhelishvili, L., Tryndyak, V.P., Koturbash, I., Tokar, E.J., Waalkes, M.P., Beland, F.A., and Pogribny, I.P. (2011). Chronic administration of 2-acetylaminofluorene alters the cellular iron metabolism in rat liver. *Toxicological sciences : an official journal of the Society of Toxicology* 123, 433-440.

Sow, F.B., Alvarez, G.R., Gross, R.P., Satoskar, A.R., Schlesinger, L.S., Zwilling, B.S., and Lafuse, W.P. (2009). Role of STAT1, NF-kappaB, and C/EBPbeta in the macrophage

transcriptional regulation of hepcidin by mycobacterial infection and IFN-gamma. *Journal of leukocyte biology* 86, 1247-1258.

Struhl, K. (2007). Transcriptional noise and the fidelity of initiation by RNA polymerase II. *Nature structural & molecular biology* 14, 103-105.

Tomikawa, J., Shimokawa, H., Uesaka, M., Yamamoto, N., Mori, Y., Tsukamura, H., Maeda, K., and Imamura, T. (2011). Single-stranded noncoding RNAs mediate local epigenetic alterations at gene promoters in rat cell lines. *The Journal of biological chemistry* 286, 34788-34799.

Tsai, M.C., Manor, O., Wan, Y., Mosammamaparast, N., Wang, J.K., Lan, F., Shi, Y., Segal, E., and Chang, H.Y. (2010). Long noncoding RNA as modular scaffold of histone modification complexes. *Science* 329, 689-693.

Uesaka, M., Nishimura, O., Go, Y., Nakashima, K., Agata, K., and Imamura, T. (2014). Bidirectional promoters are the major source of gene activation-associated non-coding RNAs in mammals. *BMC genomics* 15, 35.

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

Ulitsky, I., Shkumatava, A., Jan, C.H., Sive, H., and Bartel, D.P. (2011). Conserved function of lincRNAs in vertebrate embryonic development despite rapid sequence evolution. *Cell* 147, 1537-1550.

Vashchenko, G., and Macgillivray, R.T. (2012). Functional role of the putative iron ligands in the ferroxidase activity of recombinant human hephaestin. *Journal of biological inorganic chemistry : JBIC : a publication of the Society of Biological Inorganic Chemistry* 17, 1187-1195.

Villegas, V.E., and Zaphiropoulos, P.G. (2015). Neighboring gene regulation by antisense long non-coding RNAs. *International journal of molecular sciences* 16, 3251-3266.

Vulpe, C.D., Kuo, Y.M., Murphy, T.L., Cowley, L., Askwith, C., Libina, N., Gitschier, J., and Anderson, G.J. (1999). Hephaestin, a ceruloplasmin homologue implicated in intestinal iron transport, is defective in the sla mouse. *Nature genetics* 21, 195-199.

Wang, J., Liu, X., Wu, H., Ni, P., Gu, Z., Qiao, Y., Chen, N., Sun, F., and Fan, Q. (2010). CREB up-regulates long non-coding RNA, HULC expression through interaction with microRNA-372 in liver cancer. *Nucleic acids research* 38, 5366-5383.

Wang, K.C., Yang, Y.W., Liu, B., Sanyal, A., Corces-Zimmerman, R., Chen, Y., Lajoie, B.R., Protacio, A., Flynn, R.A., Gupta, R.A., *et al.* (2011). A long noncoding RNA maintains active chromatin to coordinate homeotic gene expression. *Nature* 472, 120-124.

Wang, L., Park, H.J., Dasari, S., Wang, S., Kocher, J.P., and Li, W. (2013). CPAT: Coding-Potential Assessment Tool using an alignment-free logistic regression model. *Nucleic acids research* 41, e74.

Wei, W., Pelechano, V., Jarvelin, A.I., and Steinmetz, L.M. (2011). Functional consequences of bidirectional promoters. *Trends in genetics : TIG* 27, 267-276.

Wheeler, T.J., and Eddy, S.R. (2013). nhmmer: DNA homology search with profile HMMs. *Bioinformatics* 29, 2487-2489.

- Wilson, R.C., and Doudna, J.A. (2013). Molecular mechanisms of RNA interference. *Annual review of biophysics* 42, 217-239.
- Wilusz, J.E., Sunwoo, H., and Spector, D.L. (2009). Long noncoding RNAs: functional surprises from the RNA world. *Genes & development* 23, 1494-1504.
- Worthen, C.A., and Enns, C.A. (2014). The role of hepatic transferrin receptor 2 in the regulation of iron homeostasis in the body. *Frontiers in pharmacology* 5, 34.
- Wu, S., Huang, S., Ding, J., Zhao, Y., Liang, L., Liu, T., Zhan, R., and He, X. (2010). Multiple microRNAs modulate p21Cip1/Waf1 expression by directly targeting its 3' untranslated region. *Oncogene* 29, 2302-2308.
- Xia, T., Liao, Q., Jiang, X., Shao, Y., Xiao, B., Xi, Y., and Guo, J. (2014). Long noncoding RNA associated-competing endogenous RNAs in gastric cancer. *Scientific reports* 4, 6088.
- Yap, K.L., Li, S., Munoz-Cabello, A.M., Raguz, S., Zeng, L., Mujtaba, S., Gil, J., Walsh, M.J., and Zhou, M.M. (2010). Molecular interplay of the noncoding RNA ANRIL and methylated histone H3 lysine 27 by polycomb CBX7 in transcriptional silencing of INK4a. *Molecular cell* 38, 662-674.
- Yoon, J.H., Abdelmohsen, K., Srikantan, S., Yang, X., Martindale, J.L., De, S., Huarte, M., Zhan, M., Becker, K.G., and Gorospe, M. (2012). LincRNA-p21 suppresses target mRNA translation. *Molecular cell* 47, 648-655.
- Yujing Li, Z.S. (2013). The Crosstalk between Micro RNA and Iron Homeostasis. *International Journal of Genomic Medicine* 01.

Appendix

Table A1. List of true and putative domain atrophy cases with atrophy scores between 0.15 and 1. 197

Table A2. Homologues of lncRNAs differentially expressed in iron deficient mouse model FPN-Trp. 200

Figure A1. Comparison of PTMs between candidate RBPs and other proteins.202

Figure A2. All predicted miRNA binding sites (6-mer) on mRNAs and lncRNAs. 203

Table A3. Protein-coding genes involved in iron metabolism present within 1 MB vicinity of lncRNAs that are expressed in iron overload mouse model FPN-C326S. 204

Table A4. Protein-coding genes involved in iron metabolism present within 1 MB vicinity of lncRNAs that are expressed in iron deficiency mouse model FPN-Trp. 207

Table A1

UniProt ID	PfamA accession	Domain name	Domain atrophy type	Type	Atrophy score	Domain interval (d)	Unmatched HMM (D)	HMM model length (L)
LUXF_PHOLE	PF00296	Bac_lucifera se	N-terminal end-bounded atrophy	True	0.319	71	169	307
LUXF_PHOPO	PF00296	Bac_lucifera se	N-terminal end-bounded atrophy	True	0.319	71	169	307
B4E7B5_BURCJ	PF00501	AMP-binding	N-terminal end-bounded atrophy	True	0.192	71	151	417
B4EL89_BURCJ	PF00501	AMP-binding	N-terminal end-bounded atrophy	True	0.182	75	151	417
Q8AAN6_BACTN	PF00501	AMP-binding	N-terminal end-bounded atrophy	True	0.168	82	152	417
Q8GPH0_ENTAG	PF00501	AMP-binding	N-terminal end-bounded atrophy	True	0.161	95	162	417
A6QFS4_STAAE	PF02826	2-Hacid_dh_C	Downstream domain-bounded atrophy	True	0.236	39	81	178
C3PBM5_BACAA	PF02826	2-Hacid_dh_C	Downstream domain-bounded atrophy	True	0.236	40	82	178
Q1HVC6_EBVA8	PF00716	Peptidase_S21	N-terminal end-bounded atrophy	Putative	0.788	9	265	325
Q7M1H4_SOLPE	PF03767	Acid_phosphat_B	N-terminal end-bounded atrophy	Putative	0.748	21	193	230
RPOC2_SINAL	PF04983	RNA_pol_Rpb1_3	N-terminal end-bounded atrophy	Putative	0.582	12	104	158
IRS1A_XENLA	PF02174	IRS	N-terminal end-bounded atrophy	Putative	0.49	1	50	100
Q8LFU8_ARATH	PF00795	CN_hydrolase	N-terminal end-bounded atrophy	Putative	0.473	0	88	186
YR307_MIMIV	PF00481	PP2C	N-terminal end-bounded atrophy	Putative	0.412	6	111	255
MCF2_HUMAN	PF13716	CRAL_TRIO_2	N-terminal end-bounded atrophy	Putative	0.409	18	79	149
Q99N72_MOUSE	PF13716	CRAL_TRIO_2	N-terminal end-bounded atrophy	Putative	0.409	18	79	149
Q7M3I1_SHEEP	PF00244	14-3-3	N-terminal end-bounded atrophy	Putative	0.326	33	110	236
KADL_ENCCU	PF00406	ADK	N-terminal end-bounded atrophy	Putative	0.311	0	47	151
D5MNX5_ZOBGA	PF00722	Glyco_hydro_16	N-terminal end-bounded atrophy	Putative	0.297	6	61	185
YEZB_BACSU	PF01740	STAS	N-terminal end-bounded atrophy	Putative	0.291	11	45	117
MTM2_METJA	PF01555	N6_N4_Mtase	N-terminal end-bounded atrophy	Putative	0.29	53	120	231
RBR1_MAIZE	PF11934	DUF3452	N-terminal end-bounded atrophy	Putative	0.281	7	46	139
PUR_ARATH	PF04845	PurA	N-terminal end-bounded atrophy	Putative	0.239	29	81	218
KCD11_HUMAN	PF02214	BTB_2	N-terminal end-bounded atrophy	Putative	0.213	20	40	94
KCD11_MOUSE	PF02214	BTB_2	N-terminal end-bounded atrophy	Putative	0.213	18	38	94
Q9S9E5_BRANA	PF00234	Tryp_alpha_amyl	N-terminal end-bounded atrophy	Putative	0.211	4	23	90
Q9S9E6_BRANA	PF00234	Tryp_alpha_amyl	N-terminal end-bounded atrophy	Putative	0.211	4	23	90

Q9S9E7_BRANA	PF00234	Tryp_alpha_amyl	N-terminal end-bounded atrophy	Putative	0.211	4	23	90
Q9S9F0_BRANA	PF00234	Tryp_alpha_amyl	N-terminal end-bounded atrophy	Putative	0.211	4	23	90
PAAK_THET2	PF00501	AMP-binding	N-terminal end-bounded atrophy	Putative	0.199	68	151	417
COX3_CORGL	PF00510	COX3	N-terminal end-bounded atrophy	Putative	0.198	24	75	258
PAAK_ECOLI	PF00501	AMP-binding	N-terminal end-bounded atrophy	Putative	0.192	71	151	417
Q9S9E8_BRANA	PF00234	Tryp_alpha_amyl	N-terminal end-bounded atrophy	Putative	0.189	4	21	90
Q9S9E9_BRANA	PF00234	Tryp_alpha_amyl	N-terminal end-bounded atrophy	Putative	0.189	4	21	90
Y497_MYCPN	PF02126	PTE	N-terminal end-bounded atrophy	Putative	0.185	4	61	308
PAAK_AZOEV	PF00501	AMP-binding	N-terminal end-bounded atrophy	Putative	0.182	75	151	417
CGD2L_LUPAN	PF00234	Tryp_alpha_amyl	N-terminal end-bounded atrophy	Putative	0.178	3	19	90
A6L0Y5_BACV8	PF00501	AMP-binding	N-terminal end-bounded atrophy	Putative	0.177	80	154	417
LUTR_BACSU	PF00392	GntR	N-terminal end-bounded atrophy	Putative	0.172	0	11	64
PAAK_PSEPU	PF00501	AMP-binding	N-terminal end-bounded atrophy	Putative	0.17	80	151	417
Q7LZT8_9VIRU	PF00729	Viral_coat	N-terminal end-bounded atrophy	Putative	0.161	45	76	192
Q7M3I2_SHEEP	PF00244	14-3-3	N-terminal end-bounded atrophy	Putative	0.161	12	50	236
PDXL4_ARATH	PF01680	SOR_SNZ	C-terminal end-bounded atrophy	Putative	0.718	1	151	209
Q14DL6_HUMAN	PF07686	V-set	C-terminal end-bounded atrophy	Putative	0.596	1	69	114
YFF1_YEAST	PF07691	PA14	C-terminal end-bounded atrophy	Putative	0.568	8	91	146
RGPA1_RAT	PF02145	Rap_GAP	C-terminal end-bounded atrophy	Putative	0.468	25	113	188
ABCAB_HUMAN	PF00005	ABC_tran	C-terminal end-bounded atrophy	Putative	0.453	19	81	137
Q8AB22_BACTN	PF00754	F5_F8_type_C	C-terminal end-bounded atrophy	Putative	0.38	5	54	129
ISAA_STAAU	PF01464	SLT	C-terminal end-bounded atrophy	Putative	0.355	29	72	121
ISAA_STAAM	PF01464	SLT	C-terminal end-bounded atrophy	Putative	0.355	29	72	121
ISAA_STAAN	PF01464	SLT	C-terminal end-bounded atrophy	Putative	0.355	29	72	121
ISAA_STAA8	PF01464	SLT	C-terminal end-bounded atrophy	Putative	0.355	29	72	121
ISAA_STAAC	PF01464	SLT	C-terminal end-bounded atrophy	Putative	0.355	29	72	121
ZCCHV_MOUSE	PF00644	PARP	C-terminal end-bounded atrophy	Putative	0.345	37	108	206
SON_HUMAN	PF14709	DND1_DSR_M	C-terminal end-bounded atrophy	Putative	0.287	11	34	80
SON_MOUSE	PF14709	DND1_DSR_M	C-terminal end-bounded atrophy	Putative	0.287	10	33	80
Q16K62_AEDAE	PF00102	Y_phosphatase	C-terminal end-bounded atrophy	Putative	0.217	7	58	235
ECT1_YEAST	PF01467	CTP_transf_2	C-terminal end-bounded atrophy	Putative	0.21	53	86	157
TPT1L_HUMAN	PF00838	TCTP	C-terminal end-bounded atrophy	Putative	0.206	2	36	165

NAAA_RAT	PF02275	CBAH	C-terminal end-bounded atrophy	Putative	0.196	65	127	316
IELK1_BAURF	PF00197	Kunitz_legume	C-terminal end-bounded atrophy	Putative	0.193	6	40	176
NAAA_HUMAN	PF02275	CBAH	C-terminal end-bounded atrophy	Putative	0.19	67	127	316
XYLJ_PSEPU	PF01557	FAA_hydrolase	C-terminal end-bounded atrophy	Putative	0.188	6	47	218
ID5A_PROJU	PF00197	Kunitz_legume	C-terminal end-bounded atrophy	Putative	0.182	3	35	176
VDHAP_CHICK	PF01425	Amidase	C-terminal end-bounded atrophy	Putative	0.181	56	136	441
ASAH1_CAEEL	PF02275	CBAH	C-terminal end-bounded atrophy	Putative	0.18	70	127	316
ID5A_ADEPA	PF00197	Kunitz_legume	C-terminal end-bounded atrophy	Putative	0.176	4	35	176
METL8_MOUSE	PF08241	Methyltransf_11	C-terminal end-bounded atrophy	Putative	0.158	4	19	95
PURK_STAAN	PF02826	2-Hacid_dh_C	Downstream domain-bounded atrophy	Putative	0.236	39	81	178
PURK_STAAM	PF02826	2-Hacid_dh_C	Downstream domain-bounded atrophy	Putative	0.236	39	81	178
AUBA_PYRFU	PF10150	RNase_E_G	Downstream domain-bounded atrophy	Putative	0.199	105	159	271
TLP_ORYSJ	PF00314	Thaumatococin	Within-domain atrophy	Putative	0.286	0	61	213
XYNA_CRYAL	PF00331	Glyco_hydro_10	Within-domain atrophy	Putative	0.241	0	77	320
CBR_DUNBA	PF00504	Chloroal_b-bind	Within-domain atrophy	Putative	0.167	22	48	156
Q7M3I2_SHEEP	PF00244	14-3-3	Within-domain atrophy	Putative	0.153	0	36	236

Table A1. List of true and putative domain atrophy cases with atrophy scores between 0.15 and 1.

Table A2

Mouse gene ID	Ensembl Biotype	Mouse gene symbol	Human gene symbol	Human gene ID	Rfam Accession
ENSMUSG00000029447	processed transcript	Cct6a	CCT6A	ENSG00000146731	
ENSMUSG00000053332	processed transcript	Gas5	GAS5	ENSG00000234741	
ENSMUSG00000056579	processed transcript	Tug1	TUG1	ENSG00000253352	
ENSMUSG00000060183	polymorphic pseudogene	Cxcl11	CXCL11	ENSG00000169248	
ENSMUSG00000064043	processed transcript	Trerf1	TRERF1	ENSG00000124496	
ENSMUSG00000064380	snoRNA	Gm26448	SNORA73A	ENSG00000274266	RF00045
ENSMUSG00000064422	snRNA	Gm22502	RNU6-750P	ENSG00000212248	RF00026
ENSMUSG00000064493	snoRNA	Snora28	SNORA28	ENSG00000272533	RF00400
ENSMUSG00000064595	snoRNA	Gm22300	SNORA44	ENSG00000252840	RF00405
ENSMUSG00000064602	snoRNA	Snora41	SNORA41	ENSG00000207406	RF00403
ENSMUSG00000064634	snoRNA	Gm22620	SNORA1	ENSG00000206834	RF00408
ENSMUSG00000064637	snoRNA	Snora20	SNORA20	ENSG00000207392	RF00401
ENSMUSG00000064721	snoRNA	Gm25855	SNORD25	ENSG00000275043	RF00054
ENSMUSG00000064796	misc RNA	Terc	Telomerase-vert.1	ENSG00000277925	RF00024
ENSMUSG00000064797	snoRNA	Gm24357	SNORD6	ENSG00000202314	RF00342
ENSMUSG00000064925	snoRNA	Snora62	SNORA62	ENSG00000272015	RF00091
ENSMUSG00000065037	misc RNA	Rn7sk	RN7SKP178	ENSG00000201875	RF00100
ENSMUSG00000065281	snoRNA	Gm24452	SNORD27	ENSG00000252128	RF00086
ENSMUSG00000065634	snoRNA	Gm24252	SNORA24	ENSG00000275994	RF00399
ENSMUSG00000065663	snoRNA	Gm22579	SNORA25	ENSG00000252550	RF00402
ENSMUSG00000065734	snoRNA	Snord49a	SNORD49A	ENSG00000277370	RF00277
ENSMUSG00000074918	antisense	Inafm2	INAFM2	ENSG00000259330	
ENSMUSG00000076609	IG C gene	Igkc	IGKC	ENSG00000211592	
ENSMUSG00000077549	snoRNA	Snord71	SNORD71	ENSG00000223224	RF00576
ENSMUSG00000077677	snRNA	Gm24468	RNU6-679P	ENSG00000212305	RF00026
ENSMUSG00000084453	snRNA	Gm24596	RNU6-98P	ENSG00000206900	RF00026
ENSMUSG00000087819	snoRNA	Gm25117	SNORA48	ENSG00000212383	RF00554
ENSMUSG00000087881	scaRNA	Gm22442	SCARNA21	ENSG00000252835	RF00602
ENSMUSG00000087968	scaRNA	Gm25395	SCARNA6	ENSG00000252798	RF00478
ENSMUSG00000088176	misc RNA	Gm23094	7SK	ENSG00000271394	RF00100
ENSMUSG00000088573	misc RNA	Gm24530	RN7SKP141	ENSG00000251976	RF00100
ENSMUSG00000088929	snoRNA	Gm24299	SNORD5	ENSG00000239195	RF01161
ENSMUSG00000089011	snoRNA	Gm24879	SNORA48	ENSG00000212383	RF00554
ENSMUSG00000089015	snRNA	Gm24996	RNU6-871P	ENSG00000251931	RF00026
ENSMUSG00000089296	snRNA	Gm23205	RNU6-387P	ENSG00000223263	RF00026
ENSMUSG00000089607	snRNA	Gm22500	RNU6-412P	ENSG00000252243	RF00026
ENSMUSG00000089634	Processed pseudogene	Nat8b	NAT8		

ENSMUSG00000092341	lincRNA	Malat1	MALAT1	ENSG00000278217	
ENSMUSG00000092713	snoRNA	Gm22858	SNORD53 SNORD92	ENSG00000265706	RF00325
ENSMUSG00000092837	ribozyme	Rpph1	RPPH1	ENSG00000277209	
ENSMUSG00000093183	misc RNA	Gm25687	RN7SL277P	ENSG00000240490	RF00017
ENSMUSG00000094152	lincRNA	Slc6a16	SLC6A16	ENSG00000063127	
ENSMUSG00000097059	lincRNA	Fam120aos	FAM120AOS	ENSG00000188938	
ENSMUSG00000097571	lincRNA	Jpx	JPX	ENSG00000225470	
ENSMUSG00000097589	processed transcript	Dleu2	DLEU2	ENSG00000231607	
ENSMUSG00000098234	lincRNA	Snhg6	SNHG6	ENSG00000245910	
ENSMUSG00000100826	processed transcript	Snhg14	SNHG14	ENSG00000224078	
ENSMUSG00000101609	antisense	Kcnq1ot1	KCNQ1OT1	ENSG00000269821	
ENSMUSG00000103081	Polymorphic pseudogene	Pcdhgb8	PCDHGB3		
ENSMUSG00000104213	IG C gene	Ighd	IGHD		
ENSMUSG00000104960	processed transcript	Snhg8	SNHG8	ENSG00000269893	
ENSMUSG00000064451	snoRNA	Snora23	SNORA23	ENSG00000201998	RF00319
ENSMUSG00000064853	snoRNA	Gm23442	SNORA38	ENSG00000200816	RF00428
ENSMUSG00000064943	snRNA	Gm23240	RNU1-125P	ENSG00000252561	RF00003
ENSMUSG00000064994	snoRNA	Gm22422	SNORA70	ENSG00000206886	RF00156
ENSMUSG00000065145	misc RNA	Vaultrc5	VTRNA3-1P	ENSG00000199422	RF00006
ENSMUSG00000065402	miRNA	Mir122	MIR122	ENSG00000207778	
ENSMUSG00000077563	snoRNA	Snora68	SNORA68	ENSG00000207166	RF00263
ENSMUSG00000084638	snRNA	Gm23889	RNU6-777P	ENSG00000201135	RF00026
ENSMUSG00000088273	snoRNA	Gm23123	SNORA48	ENSG00000212383	RF00554
ENSMUSG00000088428	rRNA	Gm22556	RNA5SP111	ENSG00000223318	RF00001
ENSMUSG00000088705	snRNA	Gm25549	RNU6-694P	ENSG00000200941	RF00026
ENSMUSG00000089542	snoRNA	Gm25835	SNORD10	ENSG00000238917	RF01290
ENSMUSG00000092274	lincRNA	Neat1	NEAT1	ENSG00000245532	
ENSMUSG00000093007	miRNA	Mir15a	MIR15A	ENSG00000275952	
ENSMUSG00000094405	snRNA	Gm23143	RNU5E-1	ENSG00000199347	RF00020
ENSMUSG00000094411	snoRNA	Snord16a	SNORD16	ENSG00000199673	RF00138
ENSMUSG00000096037	rRNA	n-R5s136	RNA5S12	ENSG00000199270	RF00001

Table A2. Homologues of lincRNAs differentially expressed in iron deficient mouse model FPN-Trp.

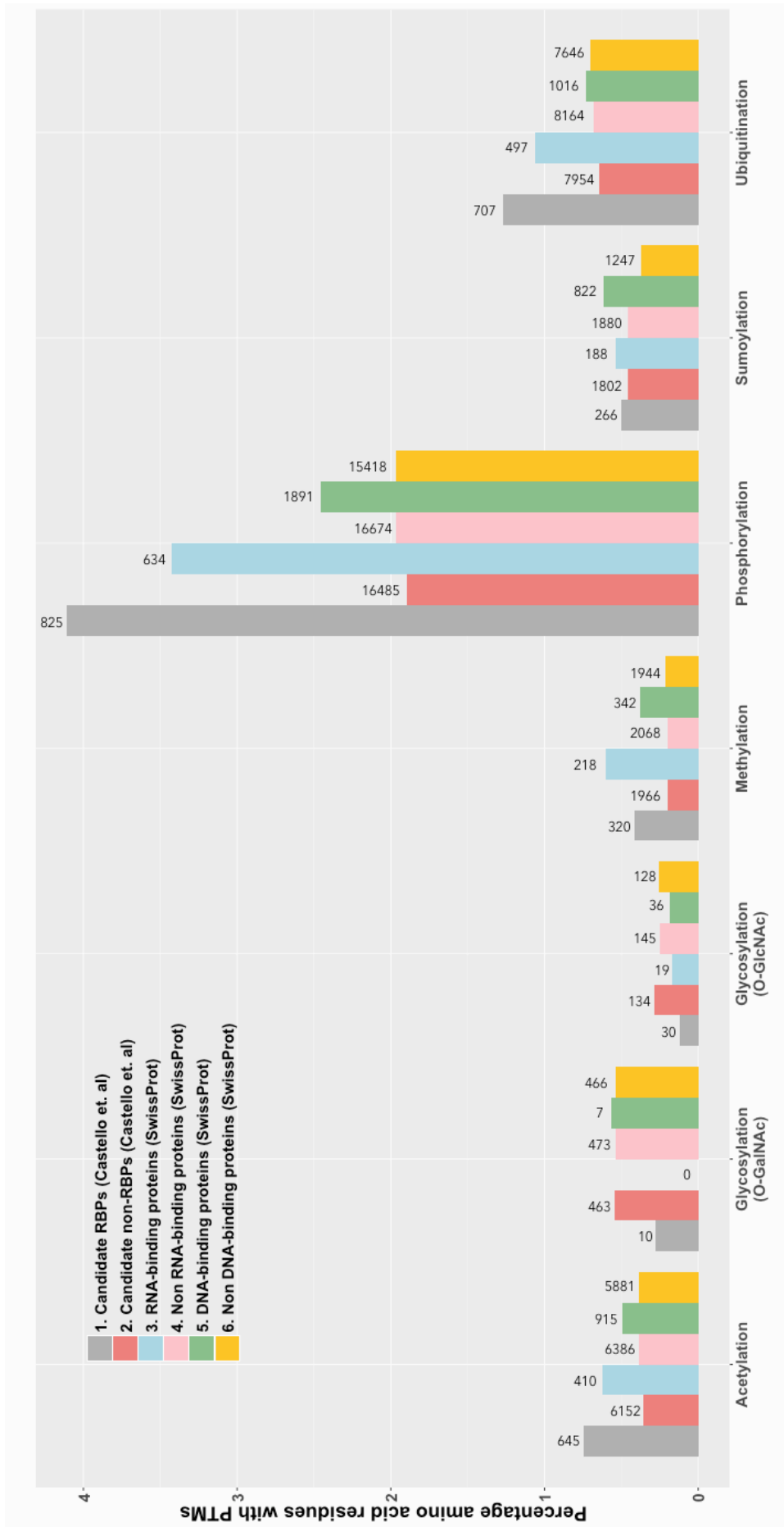


Figure A1 Comparison of PTMs between candidate RBPs and other proteins. Numbers on top of each bar denotes the number of proteins with the particular PTM.

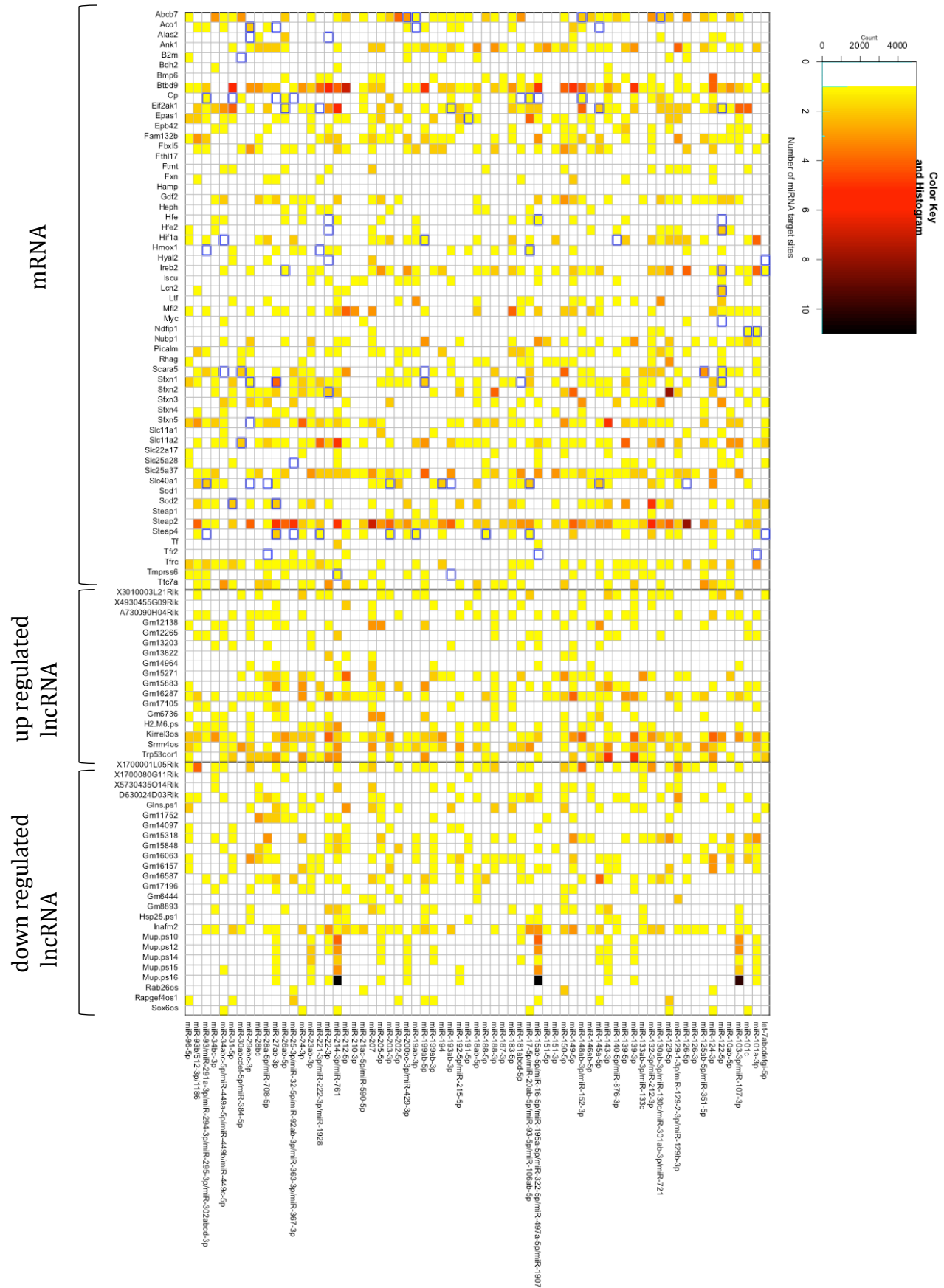


Figure A2. All predicted miRNA binding sites (6-mer) on mRNAs and lncRNAs. Boxes with blue borders are experimentally validated miRNA targets.

Table A3

ncRNA EnsEMBL gene id	mRNA gene in vicinity (within 1MB)	Gene Ontology ID	Category	Description
ENSMUSG00000000031	Th	GO:0008199	Molecular Function	ferric iron binding
ENSMUSG00000000031	Th	GO:0008198	Molecular Function	ferrous iron binding
ENSMUSG00000000031	Th	GO:0005506	Molecular Function	iron ion binding
ENSMUSG00000090357	Hamp2	GO:0006879	Biological Process	cellular iron ion homeostasis
ENSMUSG00000090357	Hamp2	GO:0097690	Molecular Function	iron channel inhibitor activity
ENSMUSG00000090357	Hamp2	GO:0034760	Biological Process	negative regulation of iron ion transmembrane transport
ENSMUSG00000090357	Hamp	GO:0006879	Biological Process	cellular iron ion homeostasis
ENSMUSG00000090357	Hamp	GO:0055072	Biological Process	iron ion homeostasis
ENSMUSG00000090357	Hamp	GO:0097690	Molecular Function	iron channel inhibitor activity
ENSMUSG00000090357	Hamp	GO:0034760	Biological Process	negative regulation of iron ion transmembrane transport
ENSMUSG00000062132	Hamp2	GO:0006879	Biological Process	cellular iron ion homeostasis
ENSMUSG00000062132	Hamp2	GO:0097690	Molecular Function	iron channel inhibitor activity
ENSMUSG00000062132	Hamp2	GO:0034760	Biological Process	negative regulation of iron ion transmembrane transport
ENSMUSG00000062132	Hamp	GO:0006879	Biological Process	cellular iron ion homeostasis
ENSMUSG00000062132	Hamp	GO:0055072	Biological Process	iron ion homeostasis
ENSMUSG00000062132	Hamp	GO:0097690	Molecular Function	iron channel inhibitor activity
ENSMUSG00000062132	Hamp	GO:0034760	Biological Process	negative regulation of iron ion transmembrane transport
ENSMUSG00000079011	Hmox1	GO:0006879	Biological Process	cellular iron ion homeostasis
ENSMUSG00000079011	Hmox1	GO:0055072	Biological Process	iron ion homeostasis
ENSMUSG00000079011	Hmox1	GO:0034395	Biological Process	regulation of transcription from RNA polymerase II promoter in response to iron
ENSMUSG00000086754	Nfs1	GO:0018283	Biological Process	iron incorporation into metallo-sulfur cluster
ENSMUSG00000086754	Nfs1	GO:0051536	Molecular Function	iron-sulfur cluster binding
ENSMUSG00000089842	Ogfod2	GO:0005506	Molecular Function	iron ion binding
ENSMUSG00000086166	Rtel1	GO:0051539	Molecular Function	4 iron, 4 sulfur cluster binding
ENSMUSG00000086166	Rtel1	GO:0051536	Molecular Function	iron-sulfur cluster binding
ENSMUSG00000085546	Nfs1	GO:0018283	Biological Process	iron incorporation into metallo-sulfur cluster
ENSMUSG00000085546	Nfs1	GO:0051536	Molecular Function	iron-sulfur cluster binding
ENSMUSG00000093594	Plod1	GO:0008198	Molecular Function	ferrous iron binding

ENSMUSG00000093594	Plod1	GO:0005506	Molecular Function	iron ion binding
ENSMUSG00000086868	Cyp2b19	GO:0005506	Molecular Function	iron ion binding
ENSMUSG00000086868	Cyp2a12	GO:0005506	Molecular Function	iron ion binding
ENSMUSG00000086868	Cyp2f2	GO:0005506	Molecular Function	iron ion binding
ENSMUSG00000086868	Egln2	GO:0005506	Molecular Function	iron ion binding
ENSMUSG00000073144	Tbxas1	GO:0005506	Molecular Function	iron ion binding
ENSMUSG00000073144	Kdm7a	GO:0005506	Molecular Function	iron ion binding
ENSMUSG00000093629	Isca2	GO:0051537	Molecular Function	2 iron, 2 sulfur cluster binding
ENSMUSG00000093629	Isca2	GO:0051539	Molecular Function	4 iron, 4 sulfur cluster binding
ENSMUSG00000093629	Isca2	GO:0016226	Biological Process	iron-sulfur cluster assembly
ENSMUSG00000093629	Isca2	GO:0008198	Molecular Function	ferrous iron binding
ENSMUSG00000093629	Isca2	GO:0097428	Biological Process	protein maturation by iron-sulfur cluster transfer
ENSMUSG00000093629	Isca2	GO:0051536	Molecular Function	iron-sulfur cluster binding
ENSMUSG00000085295	Cyp7a1	GO:0005506	Molecular Function	iron ion binding
ENSMUSG00000089746	Etfdh	GO:0051539	Molecular Function	4 iron, 4 sulfur cluster binding
ENSMUSG00000089746	Etfdh	GO:0051536	Molecular Function	iron-sulfur cluster binding
ENSMUSG00000086128	Fech	GO:0051537	Molecular Function	2 iron, 2 sulfur cluster binding
ENSMUSG00000086128	Fech	GO:0055072	Biological Process	iron ion homeostasis
ENSMUSG00000086128	Fech	GO:0030350	Molecular Function	iron-responsive element binding
ENSMUSG00000086128	Fech	GO:0005506	Molecular Function	iron ion binding
ENSMUSG00000086128	Fech	GO:0051536	Molecular Function	iron-sulfur cluster binding
ENSMUSG00000085444	Bbox1	GO:0005506	Molecular Function	iron ion binding
ENSMUSG00000078122	Heph	GO:0006879	Biological Process	cellular iron ion homeostasis
ENSMUSG00000078122	Heph	GO:0055072	Biological Process	iron ion homeostasis
ENSMUSG00000078122	Heph	GO:0006826	Biological Process	iron ion transport
ENSMUSG00000078122	Heph	GO:0008198	Molecular Function	ferrous iron binding
ENSMUSG00000086130	Plod1	GO:0008198	Molecular Function	ferrous iron binding
ENSMUSG00000086130	Plod1	GO:0005506	Molecular Function	iron ion binding
ENSMUSG00000079505	Tnf	GO:0045994	Biological Process	positive regulation of translational initiation by iron
ENSMUSG00000086605	Ptgis	GO:0005506	Molecular Function	iron ion binding
ENSMUSG00000074876	B2m	GO:1903991	Biological Process	positive regulation of ferrous iron import into cell

ENSMUSG00000074876	B2m	GO:0055072	Biological Process	iron ion homeostasis
ENSMUSG00000074876	B2m	GO:0071281	Biological Process	cellular response to iron ion
ENSMUSG00000074876	B2m	GO:1904434	Biological Process	positive regulation of ferrous iron binding
ENSMUSG00000086914	Cyp19a1	GO:0005506	Molecular Function	iron ion binding
ENSMUSG00000089712	Ogfod1	GO:0005506	Molecular Function	iron ion binding
ENSMUSG00000089712	Ciapin1	GO:0051537	Molecular Function	2 iron, 2 sulfur cluster binding
ENSMUSG00000089712	Ciapin1	GO:0016226	Biological Process	iron-sulfur cluster assembly
ENSMUSG00000089712	Ciapin1	GO:0051536	Molecular Function	iron-sulfur cluster binding
ENSMUSG00000090220	Ogfod2	GO:0005506	Molecular Function	iron ion binding
ENSMUSG00000093565	Nthl1	GO:0051539	Molecular Function	4 iron, 4 sulfur cluster binding
ENSMUSG00000093565	Nthl1	GO:0051536	Molecular Function	iron-sulfur cluster binding
ENSMUSG00000093565	Nubp2	GO:0051539	Molecular Function	4 iron, 4 sulfur cluster binding
ENSMUSG00000093565	Nubp2	GO:0016226	Biological Process	iron-sulfur cluster assembly
ENSMUSG00000093565	Nubp2	GO:0051536	Molecular Function	iron-sulfur cluster binding
ENSMUSG00000085772	Hba-x	GO:0005506	Molecular Function	iron ion binding

Table A3. Protein-coding genes involved in iron metabolism present within 1 MB vicinity of lncRNAs that are expressed in iron overload mouse model FPN-C326S.

Table A4

ncRNA EnsEMBL gene id	mRNA gene in vicinity (within 1MB)	Gene Ontology ID	Category	Description
ENSMUSG00000097904	Pole	GO:0051539	Molecular Function	4 iron, 4 sulfur cluster binding
ENSMUSG00000097904	Pole	GO:0051536	Molecular Function	iron-sulfur cluster binding
ENSMUSG00000097904	Hscb	GO:0016226	Biological Process	iron-sulfur cluster assembly
ENSMUSG00000087384	Nubp1	GO:0006879	Biological Process	cellular iron ion homeostasis
ENSMUSG00000087384	Nubp1	GO:0051539	Molecular Function	4 iron, 4 sulfur cluster binding
ENSMUSG00000087384	Nubp1	GO:0016226	Biological Process	iron-sulfur cluster assembly
ENSMUSG00000087384	Nubp1	GO:0051536	Molecular Function	iron-sulfur cluster binding
ENSMUSG00000000031	Th	GO:0008199	Molecular Function	ferric iron binding
ENSMUSG00000000031	Th	GO:0008198	Molecular Function	ferrous iron binding
ENSMUSG00000000031	Th	GO:0005506	Molecular Function	iron ion binding
ENSMUSG00000100738	Epas1	GO:0055072	Biological Process	iron ion homeostasis
ENSMUSG00000097472	Cyp4f18	GO:0005506	Molecular Function	iron ion binding
ENSMUSG00000104945	Ltf	GO:0055072	Biological Process	iron ion homeostasis
ENSMUSG00000104945	Ltf	GO:0005506	Molecular Function	iron ion binding
ENSMUSG00000078247	Sod2	GO:0055072	Biological Process	iron ion homeostasis
ENSMUSG00000089842	Ogfod2	GO:0005506	Molecular Function	iron ion binding
ENSMUSG00000090220	Ogfod2	GO:0005506	Molecular Function	iron ion binding
ENSMUSG00000086050	Jmjd6	GO:0005506	Molecular Function	iron ion binding
ENSMUSG00000089755	Ttc7	GO:0006879	Biological Process	cellular iron ion homeostasis
ENSMUSG00000062132	Hamp2	GO:0006879	Biological Process	cellular iron ion homeostasis
ENSMUSG00000062132	Hamp2	GO:0097690	Molecular Function	iron channel inhibitor activity
ENSMUSG00000062132	Hamp2	GO:0034760	Biological Process	negative regulation of iron ion transmembrane transport
ENSMUSG00000062132	Hamp	GO:0006879	Biological Process	cellular iron ion homeostasis
ENSMUSG00000062132	Hamp	GO:0055072	Biological Process	iron ion homeostasis
ENSMUSG00000062132	Hamp	GO:0097690	Molecular Function	iron channel inhibitor activity
ENSMUSG00000062132	Hamp	GO:0034760	Biological Process	negative regulation of iron ion transmembrane transport

ENSMUSG00000085524	Nfs1	GO:0018283	Biological Process	iron incorporation into metallo-sulfur cluster
ENSMUSG00000085524	Nfs1	GO:0051536	Molecular Function	iron-sulfur cluster binding
ENSMUSG00000105759	Ogfod2	GO:0005506	Molecular Function	iron ion binding
ENSMUSG00000086769	Nos3	GO:0005506	Molecular Function	iron ion binding
ENSMUSG00000100629	Ogfod3	GO:0005506	Molecular Function	iron ion binding
ENSMUSG00000090778	Fto	GO:0008198	Molecular Function	ferrous iron binding
ENSMUSG00000085514	Brip1	GO:0051539	Molecular Function	4 iron, 4 sulfur cluster binding
ENSMUSG00000085514	Brip1	GO:0051536	Molecular Function	iron-sulfur cluster binding
ENSMUSG00000097620	Sdhb	GO:0051537	Molecular Function	2 iron, 2 sulfur cluster binding
ENSMUSG00000097620	Sdhb	GO:0051539	Molecular Function	4 iron, 4 sulfur cluster binding
ENSMUSG00000097620	Sdhb	GO:0051538	Molecular Function	3 iron, 4 sulfur cluster binding
ENSMUSG00000097620	Sdhb	GO:0051536	Molecular Function	iron-sulfur cluster binding
ENSMUSG00000075265	Tfrc	GO:0006879	Biological Process	cellular iron ion homeostasis
ENSMUSG00000075265	Tfrc	GO:0097286	Biological Process	iron ion import
ENSMUSG00000075265	Tfrc	GO:0005381	Molecular Function	iron ion transmembrane transporter activity
ENSMUSG00000075265	Tfrc	GO:0071281	Biological Process	cellular response to iron ion
ENSMUSG00000085772	Hba-x	GO:0005506	Molecular Function	iron ion binding
ENSMUSG00000097000	Cyp4f18	GO:0005506	Molecular Function	iron ion binding
ENSMUSG00000101360	Th	GO:0008199	Molecular Function	ferric iron binding
ENSMUSG00000101360	Th	GO:0008198	Molecular Function	ferrous iron binding
ENSMUSG00000101360	Th	GO:0005506	Molecular Function	iron ion binding
ENSMUSG00000086868	Cyp2b19	GO:0005506	Molecular Function	iron ion binding
ENSMUSG00000086868	Cyp2a12	GO:0005506	Molecular Function	iron ion binding
ENSMUSG00000086868	Cyp2f2	GO:0005506	Molecular Function	iron ion binding
ENSMUSG00000086868	Egln2	GO:0005506	Molecular Function	iron ion binding
ENSMUSG00000091908	Cyp1a2	GO:0005506	Molecular Function	iron ion binding
ENSMUSG00000091908	Cyp1a1	GO:0005506	Molecular Function	iron ion binding
ENSMUSG00000091908	Cyp11a1	GO:0005506	Molecular Function	iron ion binding
ENSMUSG00000090263	Smad4	GO:0006879	Biological Process	cellular iron ion homeostasis
ENSMUSG00000058934	Pah	GO:0005506	Molecular Function	iron ion binding
ENSMUSG00000087014	Rev3l	GO:0051539	Molecular Function	4 iron, 4 sulfur cluster binding

ENSMUSG00000087014	Rev3l	GO:0051536	Molecular Function	iron-sulfur cluster binding
ENSMUSG00000086344	Cmah	GO:0051537	Molecular Function	2 iron, 2 sulfur cluster binding
ENSMUSG00000086344	Cmah	GO:0051536	Molecular Function	iron-sulfur cluster binding
ENSMUSG00000087492	Tph1	GO:0005506	Molecular Function	iron ion binding
ENSMUSG00000100121	Tbxas1	GO:0005506	Molecular Function	iron ion binding
ENSMUSG00000100121	Kdm7a	GO:0005506	Molecular Function	iron ion binding
ENSMUSG00000103409	Hyal2	GO:0060586	Biological Process	multicellular organismal iron ion homeostasis
ENSMUSG00000103409	Mon1a	GO:0006879	Biological Process	cellular iron ion homeostasis
ENSMUSG00000097246	Ercc2	GO:0051539	Molecular Function	4 iron, 4 sulfur cluster binding
ENSMUSG00000097246	Ercc2	GO:0051536	Molecular Function	iron-sulfur cluster binding
ENSMUSG00000103164	Hyal2	GO:0060586	Biological Process	multicellular organismal iron ion homeostasis
ENSMUSG00000103164	Mon1a	GO:0006879	Biological Process	cellular iron ion homeostasis
ENSMUSG00000086192	Lcn2	GO:0055072	Biological Process	iron ion homeostasis
ENSMUSG00000086192	Lcn2	GO:0005506	Molecular Function	iron ion binding
ENSMUSG00000087484	Eif2ak1	GO:0055072	Biological Process	iron ion homeostasis
ENSMUSG00000086199	Brip1	GO:0051539	Molecular Function	4 iron, 4 sulfur cluster binding
ENSMUSG00000086199	Brip1	GO:0051536	Molecular Function	iron-sulfur cluster binding
ENSMUSG00000091184	P4ha1	GO:0005506	Molecular Function	iron ion binding
ENSMUSG00000097231	Ercc2	GO:0051539	Molecular Function	4 iron, 4 sulfur cluster binding
ENSMUSG00000097231	Ercc2	GO:0051536	Molecular Function	iron-sulfur cluster binding
ENSMUSG00000097275	Nos3	GO:0005506	Molecular Function	iron ion binding
ENSMUSG00000025644	P4htm	GO:0005506	Molecular Function	iron ion binding
ENSMUSG00000098146	Hamp2	GO:0097690	Molecular Function	iron channel inhibitor activity
ENSMUSG00000098146	Hamp2	GO:0034760	Biological Process	negative regulation of iron ion transmembrane transport
ENSMUSG00000098146	Hamp	GO:0006879	Biological Process	cellular iron ion homeostasis
ENSMUSG00000098146	Hamp	GO:0055072	Biological Process	iron ion homeostasis
ENSMUSG00000098146	Hamp	GO:0097690	Molecular Function	iron channel inhibitor activity
ENSMUSG00000098146	Hamp	GO:0034760	Biological Process	negative regulation of iron ion transmembrane transport
ENSMUSG00000097665	Fdx1l	GO:0051537	Molecular Function	2 iron, 2 sulfur cluster binding
ENSMUSG00000097665	Fdx1l	GO:0051536	Molecular Function	iron-sulfur cluster binding

ENSMUSG00000084870	Dnajc24	GO:0008198	Molecular Function	ferrous iron binding
ENSMUSG00000085145	Cisd3	GO:0051537	Molecular Function	2 iron, 2 sulfur cluster binding
ENSMUSG00000085145	Cisd3	GO:0051536	Molecular Function	iron-sulfur cluster binding
ENSMUSG00000052403	Ndor1	GO:0005506	Molecular Function	iron ion binding
ENSMUSG00000085941	Nos2	GO:0005506	Molecular Function	iron ion binding
ENSMUSG00000086893	Cdkal1	GO:0051539	Molecular Function	4 iron, 4 sulfur cluster binding
ENSMUSG00000086893	Cdkal1	GO:0051536	Molecular Function	iron-sulfur cluster binding
ENSMUSG00000086645	Mfi2	GO:0055072	Biological Process	iron ion homeostasis
ENSMUSG00000086645	Mfi2	GO:0005506	Molecular Function	iron ion binding
ENSMUSG00000086645	Mfi2	GO:0097286	Biological Process	iron ion import
ENSMUSG00000085440	Cyp4v3	GO:0005506	Molecular Function	iron ion binding
ENSMUSG00000097149	Hamp2	GO:0006879	Biological Process	cellular iron ion homeostasis
ENSMUSG00000097149	Hamp2	GO:0097690	Molecular Function	iron channel inhibitor activity
ENSMUSG00000097149	Hamp2	GO:0034760	Biological Process	negative regulation of iron ion transmembrane transport
ENSMUSG00000097149	Hamp	GO:0006879	Biological Process	cellular iron ion homeostasis
ENSMUSG00000097149	Hamp	GO:0055072	Biological Process	iron ion homeostasis
ENSMUSG00000097149	Hamp	GO:0097690	Molecular Function	iron channel inhibitor activity
ENSMUSG00000097149	Hamp	GO:0034760	Biological Process	negative regulation of iron ion transmembrane transport
ENSMUSG00000094152	Ftl1	GO:0005506	Molecular Function	iron ion binding
ENSMUSG00000091192	Rxra	GO:0045994	Biological Process	positive regulation of translational initiation by iron
ENSMUSG00000097838	Cmah	GO:0051537	Molecular Function	2 iron, 2 sulfur cluster binding
ENSMUSG00000097838	Cmah	GO:0051536	Molecular Function	iron-sulfur cluster binding
ENSMUSG00000097838	Gpld1	GO:0071282	Biological Process	cellular response to iron(II) ion
ENSMUSG00000101609	Th	GO:0008199	Molecular Function	ferric iron binding
ENSMUSG00000101609	Th	GO:0008198	Molecular Function	ferrous iron binding
ENSMUSG00000101609	Th	GO:0005506	Molecular Function	iron ion binding
ENSMUSG00000085178	Aloxe3	GO:0005506	Molecular Function	iron ion binding
ENSMUSG00000085178	Alox12b	GO:0005506	Molecular Function	iron ion binding
ENSMUSG00000085178	Alox8	GO:0005506	Molecular Function	iron ion binding
ENSMUSG00000086826	Cygb	GO:0005506	Molecular Function	iron ion binding
ENSMUSG00000086826	Jmjd6	GO:0005506	Molecular Function	iron ion binding

ENSMUSG00000090873	Tfr2	GO:0006879	Biological Process	cellular iron ion homeostasis
ENSMUSG00000090873	Tfr2	GO:0097460	Biological Process	ferrous iron import into cell
ENSMUSG00000090873	Tfr2	GO:0055072	Biological Process	iron ion homeostasis
ENSMUSG00000090873	Tfr2	GO:0010039	Biological Process	response to iron ion
ENSMUSG00000090873	Tfr2	GO:0071281	Biological Process	cellular response to iron ion
ENSMUSG00000090873	Cyp3a13	GO:0005506	Molecular Function	iron ion binding
ENSMUSG00000104585	Etfdh	GO:0051539	Molecular Function	4 iron, 4 sulfur cluster binding
ENSMUSG00000104585	Etfdh	GO:0051536	Molecular Function	iron-sulfur cluster binding
ENSMUSG00000086533	Ercc2	GO:0051539	Molecular Function	4 iron, 4 sulfur cluster binding
ENSMUSG00000086533	Ercc2	GO:0051536	Molecular Function	iron-sulfur cluster binding
ENSMUSG00000107102	Lias	GO:0051539	Molecular Function	4 iron, 4 sulfur cluster binding
ENSMUSG00000107102	Lias	GO:0051536	Molecular Function	iron-sulfur cluster binding
ENSMUSG00000087593	Ercc2	GO:0051539	Molecular Function	4 iron, 4 sulfur cluster binding
ENSMUSG00000087593	Ercc2	GO:0051536	Molecular Function	iron-sulfur cluster binding
ENSMUSG00000087028	Ndor1	GO:0005506	Molecular Function	iron ion binding
ENSMUSG00000100287	Cyp11b2	GO:0005506	Molecular Function	iron ion binding
ENSMUSG00000087030	Tet1	GO:0005506	Molecular Function	iron ion binding
ENSMUSG00000087030	Dna2	GO:0051539	Molecular Function	4 iron, 4 sulfur cluster binding
ENSMUSG00000087030	Dna2	GO:0051536	Molecular Function	iron-sulfur cluster binding
ENSMUSG00000100199	Cyp2d10	GO:0005506	Molecular Function	iron ion binding
ENSMUSG00000100199	Cyp2d9	GO:0005506	Molecular Function	iron ion binding
ENSMUSG00000100199	Cyp2d26	GO:0005506	Molecular Function	iron ion binding
ENSMUSG00000097057	Tmprss6	GO:0006879	Biological Process	cellular iron ion homeostasis
ENSMUSG00000097057	Tmprss6	GO:0055072	Biological Process	iron ion homeostasis
ENSMUSG00000097320	Hamp2	GO:0006879	Biological Process	cellular iron ion homeostasis
ENSMUSG00000097320	Hamp2	GO:0097690	Molecular Function	iron channel inhibitor activity
ENSMUSG00000097320	Hamp2	GO:0034760	Biological Process	negative regulation of iron ion transmembrane transport
ENSMUSG00000097320	Hamp	GO:0006879	Biological Process	cellular iron ion homeostasis
ENSMUSG00000097320	Hamp	GO:0055072	Biological Process	iron ion homeostasis
ENSMUSG00000097320	Hamp	GO:0097690	Molecular Function	iron channel inhibitor activity
ENSMUSG00000097320	Hamp	GO:0034760	Biological Process	negative regulation of iron ion transmembrane transport

ENSMUSG00000106237	Tfr2	GO:0006879	Biological Process	cellular iron ion homeostasis
ENSMUSG00000106237	Tfr2	GO:0097460	Biological Process	ferrous iron import into cell
ENSMUSG00000106237	Tfr2	GO:0055072	Biological Process	iron ion homeostasis
ENSMUSG00000106237	Tfr2	GO:0010039	Biological Process	response to iron ion
ENSMUSG00000106237	Tfr2	GO:0071281	Biological Process	cellular response to iron ion
ENSMUSG00000106237	Cyp3a13	GO:0005506	Molecular Function	iron ion binding
ENSMUSG00000099137	P4ha3	GO:0005506	Molecular Function	iron ion binding
ENSMUSG00000086128	Fech	GO:0051537	Molecular Function	2 iron, 2 sulfur cluster binding
ENSMUSG00000086128	Fech	GO:0030350	Molecular Function	iron-responsive element binding
ENSMUSG00000086128	Fech	GO:0005506	Molecular Function	iron ion binding
ENSMUSG00000086128	Fech	GO:0051536	Molecular Function	iron-sulfur cluster binding
ENSMUSG00000097380	Fa2h	GO:0005506	Molecular Function	iron ion binding
ENSMUSG00000010492	Rtel1	GO:0051539	Molecular Function	4 iron, 4 sulfur cluster binding
ENSMUSG00000010492	Rtel1	GO:0051536	Molecular Function	iron-sulfur cluster binding
ENSMUSG00000084788	Cmah	GO:0051537	Molecular Function	2 iron, 2 sulfur cluster binding
ENSMUSG00000084788	Cmah	GO:0051536	Molecular Function	iron-sulfur cluster binding
ENSMUSG00000087223	Hfe2	GO:0006879	Biological Process	cellular iron ion homeostasis
ENSMUSG00000087223	Hfe2	GO:0055072	Biological Process	iron ion homeostasis
ENSMUSG000000100975	Glrx5	GO:0051537	Molecular Function	2 iron, 2 sulfur cluster binding
ENSMUSG000000100975	Glrx5	GO:0051536	Molecular Function	iron-sulfur cluster binding
ENSMUSG00000075585	Cyp2w1	GO:0005506	Molecular Function	iron ion binding
ENSMUSG000000101599	Ndufs1	GO:0051537	Molecular Function	2 iron, 2 sulfur cluster binding
ENSMUSG000000101599	Ndufs1	GO:0051539	Molecular Function	4 iron, 4 sulfur cluster binding
ENSMUSG000000101599	Ndufs1	GO:0051536	Molecular Function	iron-sulfur cluster binding
ENSMUSG00000048106	Cp	GO:0006879	Biological Process	cellular iron ion homeostasis
ENSMUSG00000097162	Picalm	GO:0055072	Biological Process	iron ion homeostasis
ENSMUSG00000097162	Picalm	GO:0097459	Biological Process	iron ion import into cell
ENSMUSG00000099708	2410016006 Rik	GO:0005506	Molecular Function	iron ion binding
ENSMUSG00000097703	2410016006 Rik	GO:0005506	Molecular Function	iron ion binding
ENSMUSG00000093577	Hfe2	GO:0006879	Biological Process	cellular iron ion homeostasis
ENSMUSG00000093577	Hfe2	GO:0055072	Biological Process	iron ion homeostasis
ENSMUSG00000075591	Cyp3a13	GO:0005506	Molecular Function	iron ion binding

ENSMUSG00000097285	Fam132b	GO:0006879	Biological Process	cellular iron ion homeostasis
ENSMUSG00000093565	Nthl1	GO:0051539	Molecular Function	4 iron, 4 sulfur cluster binding
ENSMUSG00000093565	Nthl1	GO:0051536	Molecular Function	iron-sulfur cluster binding
ENSMUSG00000093565	Nubp2	GO:0051539	Molecular Function	4 iron, 4 sulfur cluster binding
ENSMUSG00000093565	Nubp2	GO:0016226	Biological Process	iron-sulfur cluster assembly
ENSMUSG00000093565	Nubp2	GO:0051536	Molecular Function	iron-sulfur cluster binding
ENSMUSG00000097059	Phf2	GO:0005506	Molecular Function	iron ion binding
ENSMUSG00000089889	Eif2ak1	GO:0055072	Biological Process	iron ion homeostasis
ENSMUSG00000097613	Aco2	GO:0051539	Molecular Function	4 iron, 4 sulfur cluster binding
ENSMUSG00000097613	Aco2	GO:0005506	Molecular Function	iron ion binding
ENSMUSG00000097613	Aco2	GO:0051538	Molecular Function	3 iron, 4 sulfur cluster binding
ENSMUSG00000097613	Aco2	GO:0051536	Molecular Function	iron-sulfur cluster binding
ENSMUSG00000105130	Cyp51	GO:0005506	Molecular Function	iron ion binding
ENSMUSG00000097312	Alkbh8	GO:0005506	Molecular Function	iron ion binding
ENSMUSG00000097073	Nt5e	GO:0008198	Molecular Function	ferrous iron binding
ENSMUSG00000097890	Rev3l	GO:0051539	Molecular Function	4 iron, 4 sulfur cluster binding
ENSMUSG00000097890	Rev3l	GO:0051536	Molecular Function	iron-sulfur cluster binding
ENSMUSG00000097375	Phf2	GO:0005506	Molecular Function	iron ion binding
ENSMUSG00000078122	Heph	GO:0006879	Biological Process	cellular iron ion homeostasis
ENSMUSG00000078122	Heph	GO:0055072	Biological Process	iron ion homeostasis
ENSMUSG00000078122	Heph	GO:0006826	Biological Process	iron ion transport
ENSMUSG00000078122	Heph	GO:0008198	Molecular Function	ferrous iron binding
ENSMUSG00000106069	Tet2	GO:0008198	Molecular Function	ferrous iron binding

Table A4. Protein-coding genes involved in iron metabolism present within 1 MB vicinity of lncRNAs that are expressed in iron deficiency mouse model FPN-Trp.