Bio-computational identification and characterization of RNA-binding proteins in bacteria

German Title: *Bioinformatische Identifikation und Charakterisierung von RNA-bindenden Proteinen in Bakterien*

Doctoral thesis for a doctoral degree
at the Graduate School of Life Sciences,
Julius-Maximilians-Universität Würzburg,
Section: *Infection and immunity*

submitted by

Malvika Sharan

from

Ranchi, India

Würzburg, 2017

Submitted on:

Members of the Doctoral Thesis Committee:

Chairperson: Prof. Jörg Schultz

Primary Supervisor: Prof. Jörg Vogel

Supervisor (Second): Prof. Thomas Dandekar

Supervisor (Third): Dr. Ana Eulalio

Supervisor (Fourth): Dr. Cynthia Sharma

Date of Public Defence:

Date of Receipt of Certificates:

Summary

RNA-binding proteins (RBPs) have been extensively studied in eukaryotes, where they post-transcriptionally regulate many cellular events including RNA transport, translation, and stability. Experimental techniques, such as cross-linking and co-purification followed by either mass spectrometry or RNA sequencing has enabled the identification and characterization of RBPs, their conserved RNA-binding domains (RBDs), and the regulatory roles of these proteins on a genome-wide scale. These developments in quantitative, high-resolution, and high-throughput screening techniques have greatly expanded our understanding of RBPs in human and yeast cells. In contrast, our knowledge of number and potential diversity of RBPs in bacteria is comparatively poor, in part due to the technical challenges associated with existing global screening approaches developed in eukaryotes.

Genome- and proteome-wide screening approaches performed *in silico* may circumvent these technical issues to obtain a broad picture of the RNA interactome of bacteria and identify strong RBP candidates for more detailed experimental study. Here, I report APRICOT ("Analyzing Protein RNA Interaction by Combined Output Technique"), a computational pipeline for the sequence-based identification and characterization of candidate RNA-binding proteins encoded in the genomes of all domains of life using RBDs known from experimental studies. The pipeline identifies functional motifs in protein sequences of an input proteome using position-specific scoring matrices and hidden Markov models of all conserved domains available in the databases and then statistically score them based on a series of sequence-based features. Subsequently, APRICOT identifies putative RBPs and characterizes them according to functionally relevant structural properties. APRICOT performed better than other existing tools for the sequence-based prediction on the known RBP data sets. The applications and adaptability of the software was demonstrated on several large bacterial RBP data sets including the complete proteome of *Salmonella* Typhimurium strain SL1344. APRICOT reported 1068 *Salmonella* proteins as RBP candidates, which were subsequently categorized using the RBDs that have been reported in both eukaryotic and bacterial proteins. A set of 131 strong RBP candidates was selected for experimental confirmation and characterization of RNA-binding activity using RNA co-immunoprecipitation followed by high-throughput sequencing (RIP-Seq) experiments. Based on the relative abundance of transcripts across the RIP-Seq libraries, a catalogue of enriched genes was established for each candidate, which shows the RNA-binding potential of 90% of these proteins. Furthermore, the direct targets of few of these putative RBPs were validated by means of cross-linking and co-immunoprecipitation (CLIP) experiments.

This thesis presents the computational pipeline APRICOT for the global screening of protein primary sequences for potential RBPs in bacteria using RBD information from all kingdoms of life. Furthermore, it provides the first bio-computational resource of putative RBPs in *Salmonella*, which could now be further studied for their biological and regulatory roles. The command line tool and its documentation are available at https://malvikasharan.github.io/APRICOT/.

Zusammenfassung

RNA-bindende Proteine (RBPs) wurden umfangreich in Eukaryoten erforscht, in denen sie viele Prozesse wie RNA-Transport, -Translation und -Stabilität post-transkriptionell regulieren. Experimentelle Methoden wie C*ross-linking and Koimmunpräzipitation mit nachfolgedener Massenspektromentrie / RNA-Sequenzierung* ermöglichten eine weitreichende Charakterisierung von RBPs, RNA-bindenden Domänen (RBDs) und deren regulatorischen Rollen in eukaryotischen Spezies wie Mensch und Hefe. Weitere Entwicklungen im Bereich der hochdurchsatzbasierten Screeningverfahren konnten das Verständnis von RBPs in Eukaryoten enorm erweitern. Im Gegensatz dazu ist das Wissen über die Anzahl und die potenzielle Vielfalt von RBPs in Bakterien dürftig.

In der vorliegenden Arbeit präsentiere ich APRICOT, eine bioinformatische Pipeline zur sequenzbasierten Identifikation und Charakterisierung von Proteinen aller Domänen des Lebens, die auf RBD-Informationen aus experimentellen Studien aufbaut. Die Pipeline nutzt *Position Specific Scoring Matrices* und *Hidden-MarkovModelle* konservierter Domänen, um funktionelle Motive in Proteinsequenzen zu identifizieren und diese anhand von sequenzbasierter Eigenschaften statistisch zu bewerten. Anschließend identifiziert APRICOT mögliche RBPs und charakterisiert auf Basis ihrer biologischeren Eigenschaften. In Vergleichen mit ähnlichen Werkzeugen übertraf APRICOT andere Programme zur sequenzbasierten Vorhersage von RBPs. Die Anwendungsöglichkeiten und die Flexibilität der Software wird am Beispiel einiger großer RBP-Kollektionen, die auch das komplette Proteom von *Salmonella* Typhimurium SL1344 beinhalten, dargelegt. APRICOT identifiziert 1068 Proteine von *Salmonella* als RBP-Kandidaten, die anschließend unter Nutzung der bereits bekannten bakteriellen und eukaryotischen RBDs klassifiziert wurden. 131 der RBP-Kandidaten wurden zur Charakterisierung durch *RNA co-immunoprecipitation followed by high-throughput sequencing* (RIP-seq) ausgewählt. Basierend auf der relativen Menge an Transkripten in den RIP-seq-Bibliotheken wurde ein Katalog von angereicherten Genen erstellt, der auf eine potentielle RNA-bindende Funktion in 90% dieser Proteine hindeutet. Weiterhin wurden die Bindungstellen einiger dieser möglichen RBPs mit C*ross-linking and Co-immunoprecipitation* (CLIP) bestimmt.

Diese Doktorarbeit beschreibt die bioinformatische Pipeline APRICOT, die ein globales Screening von RBPs in Bakterien anhand von Informationen bekannter RBDs ermöglicht. Zudem enthält sie eine Zusammenstellung aller potentieller RPS in *Salmonella*, die nun auf ihre biologsche Funktion hin untersucht werden können. Das Kommondozeilen-Programm und seine Dokumentation sind auf https://malvikasharan.github.io/APRICOT/ verfügbar.

Table of Contents

List of figures

List of tables

Abbreviation index

AAC	Amino acid composition
ADAR	Adenosine deaminases acting on RNA
AGO	Protein Argonaute
ALS	Amyotropic lateral sclerosis
ANKHD1	Ankyrin repeat and KH domain-containing protein 1
APRICOT	Analyzing Protein RNA Interaction by Combined Output Technique
ATP	Adenosine triphosphate
AU-rich elements	Adenylate-uridylate-rich elements
AUC	Area Under the Curve
AUF1	AU-binding factor-1
BCL-2	B-cell leukemia
BLAST	Basic Local Alignment Search Tool
CAFA	Critical Assessment of Function Annotation
CASP	Critical Assessment of Structure Prediction
CDD	Conserved Domain Database in NCBI
CLIP-Seq	Cross-linked RNA co-immunoprecipitation followed by RNA-Seq
COGs	Clusters of Orthologous Groups of proteins
CSD	Cold shock domain
CSPs	Cold shock proteins
CspX	Cold shock protein X (X stands for different proteins such as CspA, B, C, E, F, H
CsrA	Carbon storage regulator
CTE	Constitutive transport element
DDX	DEAD- and DEAH-box RNA helicase families
DE	Differential expression
DEAD/DEAH box	Domain with conserved amino acid 'Asp-Glu-Ala-Asp/His' motif
DHX9	ATP-dependent RNA helicase A
DNA	Deoxyribonucleic acid
Dsrm	Double-stranded RNA-binding motif
EF-Tu	Elongation factor thermo unstable
EIF	Eukaryotic initiation factor
eIF2D	Eukaryotic initiation factor 2D
ELAV	Embryonic Lethal Abnormal Vision (RNA-binding protein)
EXOSC	Exosome complex component
FMR/FXR	Fragile X mental retardation
FN	False negative
FP	False positive
FPR	False positive rate
GO	Gene Ontology
GTP	Guanosine 5'-triphosphate
HDAC1	Histone deacetylase 1
Hfq	Host factor of bacteriophage Q
HIV-1	Human Immunodeficiency Virus
HMM	Hidden Markov Model
hnRNP	heterogeneous nuclear ribonucleoproteins
hpi	hours post infection
HuR	Hu-antigen R (HuR)

IDR	Intrinsically disordered regions
IPP	Intrinsically disordered proteins
IQR	Interquartile region
KEGG	Kyoto Encyclopedia of Genes and Genomes
KH	K Homology domain (K refers to human protein K)
KOW domain	Kyrpides, Ouzounis, and Woese (acronym for the authors)
KSRP	The KH-type splicing regulatory protein
KSRP	K homology splicing regulatory protein
La	LA motif RNA-binding domain
LARP	LA-related protein
LSM	Sm-like proteins
MCC	Matthews correlation coefficient
MCT1	Monocarboxylate-Transporter 1
mRNA	Messenger ribonucleic acid
mRNP	Messenger ribonucleic acid binding proteins
MSA	Multiple sequence alignment
NBench	Nucleic Acid Binding Prediction Benchmark
NCBI	National Center for Biotechnology Information
NIP7	Nucleolar Pre-rRNA Processing Protein 7
NOVA	Neuro-oncological ventral antigen 1
nr	Non-redundant database in NCBI
NT	CoIP sample of empty plasmid as non-target control
PABP	Poly(A)-binding protein
PABPN1	Polyadenylate-binding protein 1
PDB	Protein Data Bank
PIWI gene	Named after P-element Induced WImpy testis in *Drosophila*
PRK	Protein Clusters
PRKR	Interferon-inducible double-stranded RNA-dependent protein kinase activator
PRMT5	Protein arginine *N*-methyltransferase 5
PSI-BLAST	Position-Specific Iterative BLAST
PSSM	Position Specific Score Matrix
PTB	Polypyrimidine tract-binding proteins
PUA	Pseudouridine synthase and Archaeosine transglycosylase domain
PUF	Poly(U)-binding-splicing factor
PWM	Position Weight Matrix
PyPI	Python Package Index
Q in Q1, Q2, Q3	Quartile
RBD	RNA-binding domains
RBP	RNA-binding proteins
RIP-Seq	RNA co-immunoprecipitation followed by RNA-Seq
RNA	Ribonucleic acid
RNA-Seq	High-throughput RNA-Sequencing
RNP	Ribonucleoproteins
ROC curve	Receiver operating characteristic curve
RPS-BLAST	Reverse Position-Specific BLAST
RRM	Ribonucleic acid recognition motifs
RRP	Ribosomal RNA-processing protein
rRNA	Ribosomal ribonucleic acid
SalCom	*Salmonella* Compendium

SAM	Sterile alpha motif
SKI	A nuclear proto-onco gene, discovered at Sloan-Kettering Institute
SL1344	*Salmonella* Typhimurium strain SL1344
SM protein	Sm site binding proteins
SmpB	SsrA-binding protein
snRNP	Small nuclear ribonucleoproteins
SPT6H	Transcription elongation factor SPT6
SR	RNA binding domains with long repeats of Serine (S) and Arginine (R)
SRBD1	S1 RNA-binding domain-containing protein 1
sRNA	Small RNA
SVM	Support Vector Machine
Sxl	Protein sex-lethal
TGTs	Archaeal archaeosine synthases
THUMP	Thiouridine synthases, RNA methylases, and pseudouridine synthases
TIA-1	T-cell intracellular antigen 1
TIAR	T-cell intracellular antigen 1-related
TMM	Trimmed Mean of M-value
TN	True negative
TP	True positive
TPR	True positive rate
tRNA	transfer ribonucleic acid
TTP	Tristetraprolin
TTR	Turnover and translation regulatory
U2AF	U2 Small Nuclear RNA Auxiliary Factor
UNKL	Putative E3 ubiquitin-protein RING finger protein unkempt-like
UTR	Untranslated region
WD40	40 amino acids terminating in a tryptophan-aspartic acid dipeptide
YTH	YT521-homology domain
ZBP1	Zipcode-binding protein
ZnF/ZF	Zinc finger

Chapter 1

Introduction

1.1 Overview

The classic central dogma of molecular biology presents the one-directional flow of genetic information from DNA to RNA via transcription, and from coding RNA to proteins via translation. However, it is apparent that many processes in this highly-regulated pathway also involve so-called non-coding RNAs, where RNA molecules play the role of regulators rather than simply information carriers in diverse biological systems. Advances in RNA biology have demonstrated that RNA molecules participate in many layers of gene regulation, associated with both transcription and translation (Moore & Proudfoot, 2009; Wang *et al.*, 2015; de Klerk *et al.*, 2015). These layers of riboregulation often involve two main players: RNAs and their protein accomplices. RNA-binding proteins (RBPs) and ribonucleoprotein complexes are important post-transcriptional regulators in several processes such as RNA splicing, transport, localization, translation, stabilization, degradation, and quality control (Chothia *et al.*, 1986; Lund *et al.*, 2007; Castello & Lesk, 2012; Burd & Dreyfuss, 1994; Baltz *et al.*, 2012; Kwon *et al.*, 2013, Gerstberger *et al.*, 2014, Strein *et al.*, 2014, de Klerk & 't Hoon 2015; Merchese *et al.*, 2016; Anji *et al.*, 2016). Such regulatory mechanisms can involve either brief, transient interactions or stable binding of regulatory RNAs with RBPs. A major focus of RNA research in past decades has been towards the characterization of a handful of well-known RBPs and their structural and functional importance in cellular systems. However, the recent development of high-throughput system-wide approaches by means of interactome capture and mass spectrometry has enabled the identification of a large number of unexplored RBPs in a proteome-wide manner (Castello *et al.*, 2012; Baltz *et al.*, 2012; Michell *et al.*, 2013). So far, such studies have been conducted in eukaryotes by capturing protein-RNA complexes by means of *in vivo* experiments. Unfortunately, these approaches cannot be applied to non-eukaryotic systems due to technical limitations, such as the lack of poly-A tails in bacterial mRNAs. Hence, an even more global approach should be designed to identify RBP candidates in non-eukaryotes that could be subjected to experimental validation. This issue can be addressed by computational techniques (Puton *et al.*, 2012; Si *et al.*, 2015; Gerstberger *et al.*, 2014; Miao

& Westhof, 2015; Freeberg & Kim, 2016), which can be specialized for RBP identifications in different organisms, including bacteria, by using the existing sequence and domain knowledge from a wide range of organisms.

In this thesis, I present the first high-throughput identification of novel RBPs in the bacterial species *Salmonella* Typhimurium by means of a bio-computational approach. In the first part, I report a bioinformatic software, APRICOT ("Analyzing Protein RNA Interaction by Combined Output Technique"), which was developed for the identification and characterization of RBPs in a complete proteome using information derived from known RBPs across all kingdoms of life. In the second part, I provide insights gained from subsequent experimental studies of the computationally-identified RBP candidates in bacterial species, *Salmonella* Typhimurium, which were further analysed for their regulatory roles by integrating knowledge from of high-throughput sequencing approaches.

1.2 RNA-binding proteins and RNA-binding domains

Proteins that can bind to one or several RNA molecules in cells to form ribonucleoproteins (RNPs) are known as RNA-binding proteins (RBPs). The term RBPs is often used to indicate any RNA-interacting protein that plays important roles in cellular processes but does not necessarily require stable binding to RNAs. These proteins play central roles in both the nucleus and cytoplasm of eukaryotes, and have been characterized with cellular functions and structural roles across all the kingdoms of life (Peal *et al.*, 2011; Nishtala *et al.*, 2016).

Several RBPs are comprised of low complexity, intrinsically disordered regions that recognize and interact with their RNA targets (Castello *et al.*, 2012; 2016). However, the majority of RBPs contain a single or multiple characteristic RNA-binding domains (RBDs), which are conserved sequence motifs that facilitate the specificity and affinity towards RNA targets (Cléry & Allain, 2011). The RBDs can be categorized into two main classes: classical and non-classical. The classical RBDs, which include a small number of domain families, are known for their high abundance in eukaryotic RBPs (**Table 1.1**). A few examples are discussed later, along with their biological context.

Table 1.1 ***List of all the classical and few non-classical RBDs with their examples compiled from the literature.***

Classical RBDs	Function	Examples from Human proteome	References
RRM	RNA recognition motif that binds to ssRNA.	TIA-1, TIAR, PTBP1, ELAV, hnRNPs, snRNPs, SR, U2AF, Sxl, La, PABP, Hu/HuR	Maris *et al.*, 2005, Cléry *et al.*, 2008; Cassola *et al.*, 2010; Daubner *et al.*, 2013, Colombrita *et al.*, 2013
KH	K-homology domains occur in multiple copies and recognise different RNAs in cooperative manner.	KSRP, FMR1, ANKHD1, EXOSC, FXR	Siomi *et al.*, 1993; Siomi *et al.*, 1994; Grishin *et al.,* 2001; Velverde *et al.*, 2008
DEAD	Contain DEAD-box helicase and are involve d in RNA metabolism.	DDX3X, HDAC1	Schmid & Linder, 1992; de la Cruz, 1999; Rocak & Linder, 2004; Matsui *et al.*, 2006; Linder & Jankowsky, 2011; Banroques *et al.*, 2008
CSD	Supposed to help the cell to survive in higher then optimum temperatures.	CSDE1, CSDC2, CspA, ATP-dependent RNA helicase DeaD	Wistow, 1990; Landsman, 1992; Jones & Inouye 1994
La	Acts as RNA-polymerase III transcription factoes in the nucleus and as translation factor in cytoplasm.	LARP1, Lupus La protein	Jacks *et al.*, 2003; Alfano *et al.*, 2004; Kotok-Kogan *et al.*, 2008
zf-CCCH	Recognizes different RNAs and lead to functions such as alternative splicing, degradation etc.	ZC3H (3, 8, 10, 13, 15), UNKL, Roquin-1	Klug, 1999; Laity *et al.*, 2001; Matthews & Sunde, 2002; Brown, 2005; Hall, 2005; Gamsjaeger *et al.*, 2007; Hamad *et al.*, 2014
PIWI	Bind to small interfering RNAs and micro RNAs and play crucial roles in their biogenesis and function.	AGO (2,4), PRMT5	Yuan *et al.*, 2005; Faehnle *et al.*, 2013

PUF	Play several roles in eukaryotes such as cytoplasmic de-adenylation, translational repression.	PUF (3, 60), Pumilio homolog 3	Zamore *et al.*, 1997; Wang *et al.*, 2001; Wang *et al.*, 2002; Spassov & Jurecic, 2003
S1	The S1 domain is an essential in protein translation as it interacts with the ribosome and messenger RNA.	SRBD1, SPT6H, RRP5 homolog	Boni *et al.*, 1991; Ringquist *et al.*, 1995; Bycroft *et al.,* 1997
dsrm	Binds to dsRNAs.	ADAR, PRKRA, DHX9, Staufen homolog 1	Manche *et al.*, 1992; St Johnston *et al.*, 1992; Kim *et al.*, 1994; Bycroft *et al.*, 1995
PUA	Binds to dsRNAs and leads to ribosome biogenesis and translation regulation.	dyskerin, MCT1, eIF2D, and NIP7	Ramamurthy *et al.*, 1999; Mizutani *et al.*, 2004; Sivaraman *et al.*, 2004; Pérez-Arellano *et al.*, 2007

RNA-Recognition Motif (RRM) superfamily

The RRM fold, also known as the ribonucleoprotein (RNP) domain, is abundant in both bacterial and eukaryotic proteins and is present in an exceptionally high number of predicted RBP candidates in metazoans (references in **Table 1.1**). The RRM domain is a 90-amino acid (aa) long sequence and comprises of two conserved sub-domains of length eight and six aa, referred to as RNP1 and RNP2 respectively. Though also known for its DNA- and protein-binding capability, this domain has been intensively studied for its RNA-binding behaviours. RRM domains have been shown to mainly be involved in interactions with single-stranded RNA and are found in about 2% of human genes (Maris *et al.*, 2005). Examples of proteins encoding this domain are the human nucleolysin TIA-1 and TIAR proteins (Tessier *et al.*, 2014) that bind to AU-rich elements and lead to alternative pre-RNA splicing and regulation of mRNA translation. Another protein is neuronal ELAV protein that binds to the 3' UTR of mRNAs to confer stability or direct nuclear export, depending upon its target (Colombrita *et al.*, 2013; Darnell, 2013). Additional examples of RRM-containing RBPs include heterogeneous nuclear ribonucleoproteins (hnRNP) (Brunetti *et al.*, 2015) and small nuclear ribonucleoproteins (U1 and U2 snRNP) (Fischer *et al.*, 2011), proteins involved in alternative splicing (SR, U2AF, Sxl) (Zhu & Krainer, 2000; Guth *et al.*, 2001; Penalva & Sánchez, 2003),

and proteins that regulate RNA stability and translation (La, PABP, Hu) (Yang *et al.*, 2011; Casper *et al.*, 2013; Lu *et al.*, 2014). To date, there are more than 300 tertiary structures are available for the RRM domain in various protein contexts in the protein structure databases. These structures along with its complexes with RNAs display four characteristic β strands and two conserved α helices arranged in α-β sandwich. The 4-stranded β sheet acts as its primary RNA-binding surface. In bacteria, nearly 100 proteins have been found with RRM domains, which show certain differences from the eukaryotic RRM proteins (Maris *et al.*, 2005). For example, bacterial RRM proteins are short sequence of 100 aa and have single copy of RRM domain whereas eukaryotic proteins are generally longer and contain multiple copies of RRMs (Maris *et al.*, 2005).

K homology (KH) domains

The hnRNP KH domain is present in wide variety of RBPs and like RRM, functions in RNA recognition (references in the **Table 1.1**). This 70-aa long domain is often found in proteins in multiple copies and recognizes one or several different AU-rich regions in target RNAs. Proteins containing this domain are involved in diverse processes, including splicing, post-transcriptional regulation, and translational control (Glisovic *et al.,* 2008). This domain found in both eukaryotic and bacterial proteins, and are categorized as Type I and Type II, based on the differences in their KH folds (Grishin, 2001). The nucleic acid targets of KH domains are typically single-stranded RNA or DNA. They recognise 4-nucleotide regions of diverse patterns such as UCAC, UAAC, TCCC, CCCT, or TTTT (Cléry & Allain, 2013). Only a few PDB structures are available for KH-RNA or KH-single stranded DNA complexes. Examples of KH domain-containing proteins include the KSRP proteins, which bind to G-rich targets and regulate Let-7 microRNA biogenesis (Lee *et al.*, 2016). The fourteen KH repeats containg protein vigilin, can bind to tRNAs, mRNAs, and rRNAs. Another example of KH-containing proteins is fragile X mental retardation protein 1 (FMR1). This protein can strongly bind to GU-rich RNA targets and is involved in transporting mRNA from the nucleus to the cytoplasm (Darnell & Richter, 2012).

Cold shock domains (CSDs)

Cold shock proteins (CSPs) are found in eukaryotes, bacteria, and archaea and are comprised mostly of an evolutionarily conserved domain called the S1-like CSD (references in the **Table 1.1**). The eukaryotic gene regulatory factor Y-box protein contains a CSD and regulates transcription and translation of genes that contain the pyrimidine-rich Y-box

sequence in their promoters (Lee *et al.*, 1994). In bacteria, the CSP family includes cold-inducible proteins such as CspA and CspB of *Salmonella* and *E. coli* (Bae *et al.,* 1997; Phadtare & Inouye, 2001), which are highly expressed at lower growth temperatures, and non-cold inducible CSPs such as CspC and CspE (Phadtare & Inouye, 1999). These CSPs preferentially bind poly-pyrimidine regions of single-stranded RNA and DNA and lead to increased translation by the ribosome, mRNA degradation, or transcription termination (Manival *et al.,* 2001; Glisovic *et al.,* 2008).

DEAD-box helicases

The largest family of RNA helicases is the DEAD-box proteins, which includes proteins found in most eukaryotes and prokaryotes (references in the **Table 1.1**). DEAD-box helicases are involved in many aspects of RNA metabolism (Rocak & Linder, 2004). Nine common conserved sequences (the so-called Q-motif, motif 1, motif 1a, motif 1b, motif II, motif III, motif IV, motif V, and motif VI) of this domain are involved in ATP binding, hydrolysis, intramolecular rearrangements, and RNA interaction. These helicases modulate cellular processes like pre-mRNA processing and rearrangement of RNP complexes (Banroques *et al.,* 2008). The DEAH and SKI families are two related groups of DEAD-box-domain proteins, which together with DEAD-box are referred to as DExD/H proteins (Tanner & Linder, 2001).

Zinc finger (ZF) domains

ZF domains were originally recognized as DNA-binding domains, but are now more intensively studied for their RNA-binding capacity. ZF domains are approximately thirty aa long, and interact with their targets by means of conserved aa motifs such as CCHH, CCCH, or CCCC (references in **Table 1.1**). Many structures of ZF-containing RBPs are available in PDB, which together illustrate that this domain is comprised of a ββα protein fold where a Zn^{2+} ion holds a β-hairpin and an α-helix together. ZF domain-containing proteins can have one or multiple of these domains. The CCHH/C2H2-type ZF domains occur most frequently, in which two cysteines and histidines co-ordinate a zinc ion (Brayer *et al.,* 2008). C2H2-ZF proteins contain ~10 repeats of this domain. These are functionally diverse transcription factors, which account for about 700 human proteins (Schmitges *et al.,* 2016). For example, ZN268 acts as a transcriptional repressor, whereas ZF-containing TFs like KLF4, SP1, and ZNF423 activate or repress transcription in response to physiological stimuli. The CCCH-ZF protein Tis11d is a member of the tristetrapolin (TTP) protein family that is involved in the control of the inflammatory response. This protein recognises AU-rich elements in the 3'-end of its

target mRNA and leads to degradation of the transcript. Another CCCH ZF domain-containing protein is Muscleblind-like 1 (MBNL1), which interacts with its target RNAs by means of four CCCH ZF domains, thereby regulating alternative splicing to promote muscle differentiation (Cho & Tapscott, 2007). Inactivation of MBNL1 is associated with myotonic dystrophy. An example of CCHC ZF-containing RNP is the HIV-1 nucleocapsid (NC) protein, which recognise motifs other than an AU-rich element (Zargarian *et al.,* 2014). These examples show that ZF-containing proteins can specifically recognize different RNA sequences. The ZF proteins have not yet been reported in bacteria.

Pseudouridine synthase and archaeosine transglycosylase (PUA)-domain

The PUA domain family is highly conserved across all kingdoms of life and catalyses the isomerization of uridine to pseudouridine in its RNA targets (references in **Table 1.1**). Domains of this family are 64-96 aa long and contain a common RNA recognition surface with specific α/β architecture consisting of two α helices and six β strands, and folds into a β–sandwich structure. This domain has been identified in archaeal and eukaryotic pseudouridine synthases, archaeal archaeosine synthases, and a family of predicted archaeal and bacterial rRNA methylases (Aravind & Koonin, 1999; Coltri *et al.,* 2007).

Examples of PUA-containing proteins in eukaryotes include dyskerin, MCT1, eIF2D, and NIP7. Dyskerin is a small nucleolar ribonucleoprotein, which binds to an ACA motif of H/ACA RNA, thereby allowing its stable anchoring to tRNAs. MCT1 (multiple copies T-cell malignancies 1) is involved in ribosome biogenesis and translational regulation, and facilitates m^7G cap complex binding. It requires eIF4E for the interaction with the m^7G cap through its PUA domain (Cerrudo *et al.*, 2014). EIF2D (eukaryotic translation initiation factor 2D) is required for ribosome biogenesis and translation regulation, and acts as trafficking receptor for phosphoglycoproteins. This RBP binds to dsRNA in tRNA or rRNA (Dever & Green, 2012; Cerrudo *et al.*, 2014). NIP7 interacts with pre-rRNA and leads to 60S ribosome subunit biogenesis and translation regulation (Coltri *et al.,* 2007). Except for dyskerin protein, these proteins are absent in *E. coli, Salmonella enterica* and *Shigella dysenteriae.*

Non-classical RBDs

Unlike the above-described classical RBDs (**Table 1.1**), non-classical RBDs are much higher in number in sequence databases. Only a few RBP examples exist for different classes of non-classical RBDs. A few examples of non-classical RBDs (**Table 1.2**) are GTP-EFTU (Miller &

Weissbach, 1997), LSM (He & Parker, 2000; Kufel *et al.*, 2003; Yong *et al.*, 2004), YTH (Kang *et al.*; 2014; Harigaya *et al.*, 2006), SAM (Aviv *et al.*, 2003; Kim & Bowie, 2003), and Helicase C (Gorbalenya *et al.*, 1989; Caruthers *et al.*, 2000).

Table 1.2 ***Examples of classical and non-classical RBDs with their examples compiled from the literatures.***

Classical RBDs	Examples from the Human proteome
RRM	TIA-1, TIAR, PTBP1, ELAV, hnRNPs, snRNPs, SR, U2AF, Sxl, La, PABP, Hu/HuR,
KH	KSRP, FMR1, ANKHD1, EXOSC, FXR
DEAD	DDX3X, HDAC1
CSD	CSDE1, CSDC2, CspA, ATP-dependent RNA helicase DeaD
La	LARP1, Lupus La protein
zf-CCCH	ZC3H (3, 8, 10, 13, 15), UNKL, Roquin-1
PIWI	AGO (2,4), PRMT5
PUF	PUF (3, 60), Pumilio homolog 3
S1	SRBD1, SPT6H, RRP5 homolog
dsrm	ADAR, PRKRA, DHX9, Staufen homolog 1
PUA	dyskerin, MCT1, eIF2D, and NIP7

Using RBDs for the prediction of RBPs

RBPs are often classified based on their RBDs, as the specificity of their targeting is achieved by these motifs and can therefore give insights into their functional implications and regulatory mechanisms. Several studies have identified RNA-binding motifs in RBPs and further high-resolution information on the RNA-binding residues has been reported, which are derived from the RNA-protein structures. Owing to the amount of sequence information available for RBDs, this serves as an important resource for the development of computational techniques to identify new RBPs, to predict their RNA targets, to model RNA-protein complexes, and to identify residues important for RNA binding. Such knowledge is particularly helpful for carrying out prediction and exploratory studies in bacteria, where technical issues currently render global screening approaches such as interactome capture challenging.

1.3 Regulatory roles of RNA-protein interactions

RBPs affect regulation of gene expression in response to environmental changes by coordinating each step of the lifecycle of an RNA molecule from its transcription to its degradation. Upon transcription, specific RBPs interact with RNAs to form RNP complexes. These RNPs are further processed, localized, translated, stabilized, or degraded, often based on cellular conditions (Pullman *et al.*, 2014; Janga *et al.*, 2011; Spitale *et al.*, 2015). Malfunction in RBPs or their RNA targets can lead to dysregulations that underlie various human diseases, including cancers (Ling *et al.*, 2013; Mohan *et al.*, 2014; Lu *et al.*, 2014; Brinegar & Cooper, 2016). Several studies have been conducted in both human and bacteria to identify the regulatory roles of RBPs under routine and altered conditions in order to identify specific RNAs and biological pathways they regulate (Michaux *et al.*, 2012; Wilf *et al.*, 2013; Ariyachet *et al.*, 2013; Liu *et al.*, 2014; Wang *et al.*, 2015; Figueroa-Angulo *et al.*, 2015; Vembar *et al.*, 2015; Burke & Portnoy, 2016).

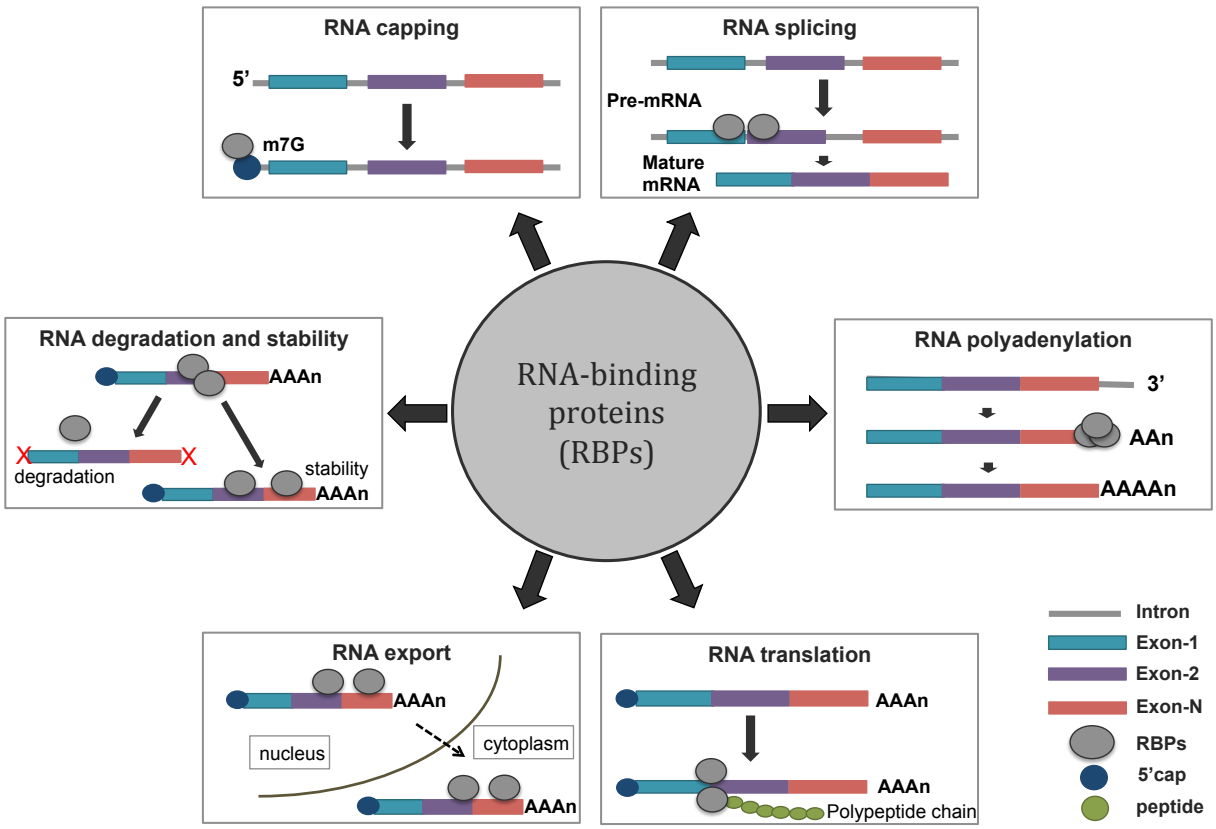


Figure 1.1 ***Illustration of several important regulatory roles of RNA-binding proteins.*** *(Adapted from Sutherland* et al.*, 2015)*

The major part of this thesis will focus on exploring and cataloguing bacterial RBPs. Since the development and design of the studies are greatly based on the landmark studies

conducted in human, it is important to give an overview of few of the important regulatory roles of RNAs, which is illustrated in **Figure 1.1** and are briefly discussed below.

Polyadenylation

All eukaryotic mRNAs, except those encoding replication-dependent histones, are processed and a 3' modification of about 200 adenylate residues referred to as poly(A) tails are added post-transcriptionally (reviewed in Jacobson & Peltz, 1996; Wickens *et al.*, 1997; Garneau *et al.*, 2007). The process of polyadenylation involves the cleavage of transcripts between a specific upstream (AAUAAA) and downstream (U/GU rich) sequences, followed by the addition of a poly(A) tail of varying lengths by poly(A) polymerase. Polyadenylation therefore leads to 3' isoforms of mRNAs, and has specific regulatory effects on various processes that affect the metabolism and expression of mRNAs, such as nuclear transport, translation efficiency, and stability. The polyadenylation of non-coding isoforms (those lacking the entire ORF) is decayed 38% faster compared to the coding isoforms (Gupta *et al.,* 2014). For example, CPSF proteins binds to AAUAAA together with nuclear poly(A) binding protein (PABPN1) to initiate poly(A) polymerase activity (Mangus *et al.,* 2003; Kuhn & Wahle 2004). PABPN1 requires an RBD and C-terminal arginine-rich domain to bind to the poly(A) tail (Liebold *et al.*, 2015). Any alterations to its coding region, such as GCG expansion in human leads to an autosomal dominant late onset neuromuscular disease called oculopharyngeal muscular dystrophy (Garibaldi *et al.,* 2015).

RNA stability, translation, and degradation

RBPs are core regulators of messenger RNA stability and translation in mammalian cells. Translational regulation provides a rapid mechanism to control gene expression by immediately modulating the production of proteins (reviewed by Di Liegro *et al.*, 2014). Usually RBPs in eukaryotes control expression by interacting with common features of mRNAs such as the 3' poly(A) tail, 5'cap structure, or 5' and 3' untranslated regions (UTRs) (Pullmann *et al.*, 2007; Abdelmohsen, 2012).

A group of classical RBPs, Hu proteins binds to AU-rich sequences in 3' UTRs of mRNA targets and influence diverse post-transcriptional aspects from splicing, to translation, to degradation of their RNA metabolism (reviewed by Hinman & Lou, 2008). Besides the Hu family proteins, additional factors that promote target degradation are AU-binding factor-1 (AUF1) (Yoon *et al.*, 2015), hnRNP (Dreyfuss *et al.* 1993), K homology splicing regulatory

protein (KSRP) (Valverde *et al.*, 2008), and tristetraprolin (TTP) (Brooksa & Blackshear, 2013). All of these proteins are termed "turnover and translation regulatory" (TTR) RBPs and they are involved in shuttling proteins between the nucleus and cytoplasm. Other examples of such RBPs are T-cell intracellular antigen 1 (TIA-1) and TIA-1-related (TIAR) proteins (Zheng *et al.*, 2005), polypyrimidine tract-binding proteins (PTB) (Kamath *et al.*, 2001), nucleolin (Borer *et al.*, 1989), and heterogeneous nuclear ribonucleoproteins (hnRNPs) (Samarina *et al.*, 1968). The targets of these proteins are involved in diverse cellular processes such as cell growth, cell cycle, stress response, proliferation, senescence, and carcinogenesis (Abdelmohsen, 2012).

This family of proteins highlight a very important feature of RBPs, which is their capability to bind to a diverse range of targets and confer target-specific regulation. Furthermore, multiple RBPs of different regulatory roles can act on a same target to infer cooperative or competitive effects.

Alternative splicing

Alternative splicing is a process of differential inclusion of exons from precursor mRNAs to generate mature isoforms and the production of different, but related, proteins with specific functions (Glisovic *et al.*, 2008). This mechanism is extensively regulated by RBPs, and involves the exclusion of introns by recognition and joining of a 5' and 3' splice site pair using a large complex of small nuclear RNAs and proteins called the spliceosome (Ideler & Jan, 2012). These complexes show high specificity for their target mRNAs, however some precursor mRNAs are recognized by different complexes in different cell types. This behaviour of spliceosomes gives rise to an altered binding specificity and splicing effect, depending on the specific biological condition or experimental set-up. The complex regulatory roles of alternative splicing are widespread in the eukaryotic cells and often linked to developmental and disease processes.

Bioinformatic studies have revealed an enrichment of regulatory RNA motifs near alternative exons with different splice site strengths. For example, serine/arginine-rich (SR) proteins regulate alternative splicing through the recruitment of the splicesome-forming small nuclear ribonucleoproteins (snRNP) U1 U2AF (Lee *et al.*, 2015). The NOVA proteins recognize YCAY (where Y represents a pyrimidine base) motifs in hnRNP mRNAs and lead to alternative splicing (Dredge *et al.*, 2015). Similarly, Fox protein shows specificity for the (U)GCAUC binding motifs near exons, leading to brain- or muscle-specific splicing patterns

(Kuroyanagi, 2009). Enrichment of this motif near exons with breast and ovarian tumour-specific patterns has been linked to the role of Fox proteins in these cancers (Kuroyanagi, 2009).

The genome-wide splicing maps of several RBPs are comparable in terms of their binding positions. For example, proteins like Nova, Fox, hnRNP L and PTB can silence exon inclusion by binding to the splice sites or exons, however binding to regions downstream of exons enhances exon inclusion (Witten *et al.*, 2011).

RNA editing and modification

Base modifications such as methylations in mRNAs and noncoding RNAs, referred as the "epitranscriptome", are involved in post-transcriptional events such as insertion, deletion, and substitution of nucleotides. About 60 RNA modifications such as base isomerization, base alteration, and ribose 2'-hydroxyl group methylation are known in the mRNAs and tRNAs (Cantera *et al*., 2010).

The widely-studied RNA modification is the conversion of the nucleobase adenosine (A) to inosine (I) in RNA molecules (Glisovic *et al.*, 2008, Sakurai *et al*., 2010). Such modification activity has been shown for ADAR (adenosine deaminases acting on RNA) proteins, which catalyse the deaminiation of adenosine to inosine in RNAs (Sakurai *et al*., 2010). Other examples of RNA-editing are conversion of adenosine to 6-methyladenosine (m^6A) (Dominissini *et al.,* 2012; Meyer *et al.,* 2012) and modification of cytosine to5-methylcytosine (m^5C) (Squires *et al.,* 2012). Though these modifications are widely studies in eukaryotes, they are reported in bacterial small non-coding RNAs (Cantera *et al*., 2010; reviewed by Marbaniang & Vogel, 2016).

Such modifications expand the diversity of gene products by producing RNA sequences that are different from those originally encoded by the genome. The majority of examples of RNA editing have been identified in the non-coding regions of RNAs. Several examples of editing have also been described in protein-coding regions that give rise to changes in the amino acid sequences of their encoded proteins. For example, editing of an A to I in the glutamate receptor GluR-B mRNA results in incorporation of a glutamine rather than arginine, which leads to altered protein function (Glisovic *et al.*, 2010). Mutations in the *ADAR* gene of *Drosophila* lead to neuronal dysfunction, whereas in in mice it has been linked to embryonic lethality (Keegan *et al.*, 2005).

mRNA export and localization

In eukaryotes, mature RNAs are exported from the nucleus to the cytoplasm after the events of transcription, splicing, and 3'-polyadenylation. This export involves the assembly of a cargo-carrier complex of RNAs and RBPs in the nucleus, which is then translocated through the nuclear pore complex and released into cytoplasm (reviewed by Steward, 2007). The RNA annealing protein Yra1 and the mRNA export factor Mex67 are examples of RBPs that associate with mRNA to facilitate its export to the cytoplasm (Cole, 2000; Oeffinger & Zenklusen, 2012). Other examples of RNA-binding export factors include recyclable hnRNP shuttling proteins such as nuclear polyadenylated RBPs and Nab4 (HRP1), as well as the nucleolar protein Npl3, which are involved in transcription, intermediate metabolism, and ribosomal biogenesis, respectively (Guisbert *et al.*, 2005). Further localization of mRNAs in the cytoplasm following nuclear export is important for gene expression as it allows the spatial regulation of protein production to a specific target site of the cell.

The different splice variants of mRNAs due to rapid cleavages in different cells can determine their localization in the cells facilitated by RBPs. One example is the multiple KH domain-containing zipcode-binding protein ZBP1, which binds to β-actin mRNA at a 54-nucleotide localization element ("zipcode") located in its 3' UTR (Ross *et al.,* 1997). ZBP1 further localizes mRNAs within the cytoplasm in several asymmetric cell types so that they can then be translated at the site where they are required. Fragile X mental retardation protein (FMRP) is another RBP involved in several processes of RNA metabolism including the stimulus-induced localization of several dendritic mRNAs, such as that encoding β-actin in neuronal dendrites (Dictenberg *et al.,* 2008).

1.4 Biological features of RBPs in eukaryotes and bacteria

Given the important regulatory roles of RBPs, several techniques have been developed for the high-throughput analysis of their targets and their functions, which often utilize co-immunoprecipitation/co-purification, protein mass spectrometry, and RNA deep sequencing. These experimental approaches have contributed to our understanding of RBPs and their roles in different organisms. Notably, due to developments in high-throughput mass-spectrometry and sequencing approaches, it is, in principle, possible to perform global analyses to comprehensively catalogue all RBPs in an organism.

1.4.1 Overview of eukaryotic RBPs

Several studies have been conducted to identify and characterize RBPs as post-transcriptional regulators in human, mouse, and yeast (Castello *et al.*, 2012; Baltz *et al.*, 2012; Mitchell *et al.*, 2012; Kwon *et al.*, 2013; Gerstberger *et al.*, 2014; Conrad *et al.*, 2015). Landmark studies for genome-wide screening of RBPs in human used RNA interactome capture techniques, which involve UV-induced covalent crosslinking of bound proteins and RNAs for the purification of RNP complexes. Such interactome capture primarily depend on the poly(A) tails of coding transcripts, which is important for the control of their expression, export, stability, and translation in eukaryotic and several archaeal species (Shi, 2012; Derti *et al.*, 2012; Régnier & Marujo, 2013). In this approach, proteins that covalently bind to polyadenylated RNAs following UV crosslinking are characterized as RBPs (Castello et al. 2012; Baltz et al., 2012). These studies facilitated the *in vivo* identification of over 1500 human RBPs including hundreds of previously unexplored candidates. The various features of eukaryotic RBPs and their classification give us an insight into the fundamental requirement of RBPs in varieties of biological processes and provides a consistent resource to guide future research in this regard. A few major features of RBPs listed below.

RNA-Binding domains and motifs

Due to an increasing amount of experimental and structural data, it has been possible to carry out multiple sequence alignments and computational predictions of the structures of these candidates to identify their functional components. These components include highly conserved protein domains and short motifs, which are the functional and structural units of proteins of diverse functionalities. Castello *et al.* (2012) classified the RBPs based on the two classes of domains called classical RBDs that are characterized in a large number of RBPs and non-classical RBDs that are characterized in several RBPs as well. A few classical RBPs are listed in the **Table 1.1** and the non-classical RBDs are listed in the **Table 1.3** along with their distribution across archaeal, bacterial and eukaryotic protein sequences including RBPs as computationally mapped by the InterPro domain database. Ray *et al.* (2013) classified RNA binding motifs of RBPs that are highly conserved across different species. These components are essential for the recognition of specific RNA targets and hence, confer binding specificity to RBPs.

A resource of 1,542 manually-curated human RBPs have been compiled that interact with specific RNA targets (Castello *et al.*, 2012; Baltz *et al.*, 2012; Gerstberger *et al.*, 2014). In

these studies, proteins have been catalogued along with their structural and functional features according to their RBDs and interacting RNA partners. RBP classification was carried out based on their RBDs, as well as their co-occurrences with other RBDs or different functional domains. For example, ssRBDs like RRMs, KH, ZF, and cold-shock domains that recognize small 4-6 nucleotide segments frequently co-occur in multiple repeats or in combinations, which increases the RNA binding affinity of RBPs and also provide an evolutionary basis for protein functions.

Table 1.3 ***Examples of non-classical RBDs that have been identified in experimental studies of specific RBPs.***

RBDs are indicated, along with the number of archaeal (A), bacterial (B) and eukaryotic (E) protein sequences computationally mapped by the InterPro domain database.

Non-classical RBDs	**Total proteins (A: Archaea, B: Bacteria, E: Eukaryotes) (Source: InterProScan)**
Alba	A: 447
APOBEC_N	B: 16, E: 1042
Brix	A: 255, B: 148, E: 6345
Btz	E: 1160
eRF1_3	A: 1051, B: 50, E: 2881
Fibrillarin	A: 529, B: 5, E: 1578
FtsJ	A: 457, B: 13992, E: 5323
Gar1	B: 10, E: 1638
GTP_EFTU	A: 2, B: 12226, E: 1076
Helicase_C	A: 5401, B: 167636, E: 104874
LSM	A: 2822, B: 45059, 23194
PseudoU_synth_1	A: 366, B: 13100, E: 3533
RNase_PH	A: 18, B: 7655, E: 2
SAM	B: 342, E: 24701
SAP	A: 96, B: 2874, E: 10958
SpoU	A: 383, B: 5669, E: 7
THUMP	A: 1507, B: 12316, E: 1566
TRM	A: 516, B: 98, E: 1767
TROVE	A: 7, B: 760, E: 571
TrpBP	B: 569
YTH	A: 4, B: 40, E: 3528
zf-C2H2	A: 1097, B: 1571, E: 197826
zf-CCHC	A: 2, B: 95, E: 43951

Disordered regions in RBPs

In interactome capture studies (Castello *et al.,* 2012), several candidate RBPs were identified without any canonical RBD. These proteins presumably rely on low complexity disordered regions for interaction with their targets and have roles in both RNA metabolism and regulation. RBPs with disordered sequences, referred to as intrinsically disordered proteins (IDPs), have common structural characteristics such as low hydrophobicity, high net charge, and low amounts of ordered secondary structure (Calabretta *et al.,* 2015). A subclass of IDPs possesses low complexity regions of 1-10 amino acids, which serve to identify and bind their targets. This relatively disordered secondary structure confers flexible structures to IDPs that are stabilised to adapt a rigid structure upon binding with their ligands. For example, RGG/RG and the related RGG/YGG motifs are widely present in IDPs such as nucleolin and the FMR protein. FMR binds with G4-quadruplex secondary structure to specific target RNAs, which induces a stable structure of an RGG/RG motif in the protein that facilitates strong interactions between the arginine and the G4 sequences. This suggests that RGG/RG disordered regions are essential for the RNA recognition by providing a more accessible conformation to allow interaction with their target RNAs. The IDRs in RBPs are also linked with formation of RNP ultrastructures, termed as 'assemblages' (Calabretta *et al.,* 2015). These dynamic RNP granules regulate RNA-processing, bioavailability, degradation, and transport and carry out RNA metabolism. For example, stress granule and p-body assembly depends on the presence of untranslated mRNAs (Ayache *et al.*, 2015).

Although there are many proteins that rely on disordered regions for their RNA binding ability (Castello *et al.*, 2016), based on the available data, RBDs serve as a primary resource for carrying out computational prediction of RBPs.

Tissue specific expression of RBPs

98% of paralogous RBPs did not have any tissue specificity for their expression as they are ubiquitously transcribed in tissues such as germline, brain, muscle, liver, and bone marrow. For example, many other RBPs such as ribosomal proteins, components of the spliceosome, and those involved in RNA transport and turnover, are expressed ubiquitously. Approximately 5% of RBP families and their isoforms have one or more members that showed tissue specificities. For example, the ELAV-like family is neuronal (Colombrita *et al.*, 2013), whereas the PIWI and DDX4 helicase proteins such as DDX3X and DDX3Y are highly germline-specific (Girard *et al.*, 2006).

The RBP classes that are ubiquitously expressed, show differential levels of expression across tissues, hence, their loss affects only the tissues where they have the highest expression. For example, some members of the fragile X mental retardation syndrome related protein family (FMR1, FXR1 and FXR2) (Agulhon *et al.*, 1999) do not have target specificity; however, FMR1 shows the highest expression levels in brain, thyroid, and gonads. Loss of these proteins leads to mental retardation, pre-mature ovarian insufficiency in fragile X syndromes like ataxia, as well as skeletal defects. Hence, these RBPs can be linked to important human genetic diseases (Wang *et al.*, 2016).

Target specificities of RBPs

The classification of RBPs has been carried out based on their predominant target RNA classes like mRNA-binding, non-coding (nc) RNA-binding, tRNA-binding, pre-rRNA-binding, and small nuclear (sn) RNA-binding, and small nucleolar (sno) RNA-binding. Several proteins have been reported for their involvements in mRNA binding (692 proteins) and ribosome binding (169 proteins). Whereas few other RBPs are involved in the biogenesis of rRNAs (122), snoRNAs (41), and of charged tRNAs (130). A total of 122 RBPs are characterized for their roles in RNA degradation, transcriptional silencing, or activation or repression of indirect targets through interaction with other ncRNAs (microRNAs, PIWI-interacting RNAs, and long ncRNAs). Such binding specificities of RBPs toward their RNA targets could facilitate the identification of members of specific pathways associated with post-transcriptional gene regulation, independent of their functions. The relative sizes of several classes of RBPs are constant across phylogenies, for instance, 38%, 12%, and 14% mRNA-, tRNA- and rRNA-binding proteins, respectively but they show varying levels of evolutionary conservation. The primary sequences of RBPs that bind to rRNA are most stable across species, with an average similarity of 51%, whereas other classes show 30% or less similarity between human and yeast, showing the divergence in their biogenesis. Across different tissue samples, transcript abundance of RBP encoding genes was recorded to be ~6 times higher than that of transcription factor encoding genes, even though they both are important regulatory elements and account for almost an equal number of genes (Gerstberger *et al.*, 2014).

RBPs with binding affinity to similar RNA classes affect the same tissues and lead to similar pathologies. For example, defects in ribosomal proteins and rRNA biogenesis factors cause bone marrow and skin related diseases, while defects in genes that encode mRNA-

binding RBPs are liked to neurodegenerative and neuromuscular diseases like amyotropic lateral sclerosis (ALS) (Vanderweyde *et al.*, 2013).

Expression dynamics of RBPs

In addition to classifications based on corresponding domains and target specificities, expression dynamics such as co-expressions can also be used to define roles for RBPs in important biological events such as ovarian and brain developments (Gerstberger *et al.,* 2014). RBPs function as direct binders and/or transient interacting partners like chaperones in RNP complexes to perform maturation, processing, regulation and transportation of RNAs. The relative abundance of RBPs and other proteins that act co-operatively or competitively, affect the overall regulatory response under different conditions. Examples of RBPs with cooperative behaviours are splicing factors, which can lead to different splicing patterns of the same precursor mRNA, and U1 snRNPs, which control alternative polyadenylation sites. Examples of RBPs with competitive behaviours are ELAVL1, which is involved in mRNA target regulation by miRNAs, and Pumilio homologs, which act synergistically with miR-221 and miR-222 to destabilise cyclin dependent kinase inhibitor 1B (Gerstberger *et al.*, 2014).

1.4.2 Overview of bacterial RBPs

Most bacterial mRNAs lack poly(A) tails; therefore, experimental techniques that rely on mRNA polyadenylation cannot be directly applied to bacterial systems. Hence, no global studies have so far been conducted on bacterial systems (Barquist & Vogel, 2015). Most of the knowledge of bacterial RBPs comes from independent characterization of single candidates in model bacteria such as *E. coli*. Only a few RBPs (besides ribosomal proteins) have been characterized in detail, such as Hfq, CsrA, ProQ, CspA, CspB, and SmpB (**Table 1.4**). Some of these RBPs have also been well characterized in other model organisms, such as the model pathogen *Salmonella* Typhimurium, which like *E. coli*, is also a Gram-negative enterobacterium.

Bacterial pathogens use several strategies to survive in challenging environmental conditions. Traditionally, the expression of virulence factors has been viewed as being controlled by transcription factors. However, in recent years it has become clear that bacterial pathogens also express many non-coding RNAs and RBPs, contributing to the post-transcriptional regulation by the modulation of RNA decay, translation initiation efficiency, or transcript elongation.

Two well characterized RBPs, Hfq and CsrA are widely conserved in different bacterial species, including many pathogens, which together with their targets comprise large post-transcriptional regulons (Romeo, 1998; Chao & Vogel, 2010; Storz *et al.*, 2011; Westermann *et al.*, 2016; Holmqvist *et al.*, 2016). Hfq protein acts as a chaperone that stabilizes bound sRNAs and helps them regulate their mRNA targets in enteric model bacteria such as *Escherichia coli* (Tree *et al.,* 2014) and *Salmonella enterica* (Chao *et al.*, 2010). Like Hfq, CsrA/RsmA family act as a global regulator by binding with their targets via GGA motifs in *E.coli, Salmonella* (Romeo *et al.* 1993; Holmqvist *et al.,* 2016).

Table 1.4 ***Different RBPs reported in enterobacterial species.***

RBPs	RBDs	References	Functions as RBPs
AmiR	ANTAR	Galperin, 2006	Transcription antitermination
Bgl/Sac family	CAT (Co-AntiTerminator RNA-binding domain)	Declerck *et al.*, 1999	Protein-mediated antitermination, control the expression of carbohydrates untilizing genes
Csp (A-B)	Cold shock domains	Phadtare *et al*., 1999	transcription antiterminators, prevent RNA degradation
CsrA	CsrA	Romeo *et al*., 1993	Global regulator, compete with the ribosome for binding to the mRNAs and regulate their expression post-transcriptionally
Hfq	LSM	Chao *et al.*, 2010	Global regulator, post-transcriptional expression regulation of interacting sRNAs and mRNAs
ProQ	ProQ, ProQ/FinO domain	Chault *et al.*, 2011; Smirnov *et al.*, 2016	RNA chaperone, regulation of ProP activity
RNaseE	DSRM	Mian *et al.*, 1997	ssRNA-specific endoribonucleases, RNA stabilization and decay
YhbY	RapZ-like family	Ostheimer *et al.*, 2002	Found in bacteria and archaea, involve in ribosome assembly
SmpB	Small protein B	Wower *et al*., 2002; Giudice *et al.*, 2014	Degrades proteins synthesized from damaged RNAs in hybrid with transfer-messenger RNA

In contrast to the limited number of RBPs, several hundreds of non-coding RNAs (ncRNAs) have been discovered in bacteria that are linked to various regulatory processes. These ncRNAs interact with either mRNAs or proteins and lead to the expression of transcription factors or specific regulons such as riboswitches that are considered global regulators of

virulence (Walters & Storz, 2009; Storz *et al.*, 2011). Functional studies of a small number of these RNA targets have revealed that, in association with their corresponding RBPs, they regulate several diverse physiological functions including virulence, translation, and stress responses. Despite the identification of hundreds of non-coding RNA targets, only a small number of them have been validated to have important roles in RNA-mediated gene regulations (**Table 1.4**).

These bacterial RBPs have been thoroughly studies in only a few bacteria and are not present in all bacterial species. For example, only approximately 50% of sequenced bacterial genomes encode an Hfq homologue (Chao & Vogel, 2012). Furthermore, several pathogens with well-established repertoire of small RNA (sRNA) transcripts that likely participate in post-transcriptional gene regulation do not always depend on these regulators for their functionalities. Therefore, the identification of novel RNA-binding proteins in pathogenic bacteria is an important next step, which will facilitate the understanding of the regulatory networks that regulate their lifestyle. In order to understand the mechanisms involved in such RNA-regulated events in bacteria, it is crucial to identify and characterize the proteins that interact with these regulatory RNAs.

1.5 Bioinformatic approaches for RBP prediction

Due to the central roles of RBPs in diverse biological processes, it is crucial to identify them and understand their regulatory mechanisms. Several RBPs have been experimentally characterized and studied for their biological properties in numerous organisms. Additionally, several structures of protein-RNA complexes have been solved experimentally, providing biophysical information on the nature of the interaction between nucleic acids and domains, as well as specific amino acid residues, of RBPs.

As discussed earlier, experimental techniques for the high-throughput identification of RBPs are not only labour-intensive and costly, but also not convenient for all biological systems such as bacteria. Therefore, computational methods are being established for the identification of candidate RBPs in the genomes of diverse organisms. Specifically, information obtained from global screening studies of several human RBPs has contributed to the development and refinement of computational tools for identification of additional eukaryotic RBPs and their corresponding RNA-binding domains. A considerable number of

protein-RNA complex crystal structures also constitute an important resource for the development of bio-computational tools.

1.5.1 Prediction of RNA-binding residues in proteins

More than fifteen computational methods have been developed to characterize RBPs by predicting RNA-binding residues derived from the known protein-RNA structures. Most of these methods rely upon a curated set of structures of RNA-protein complexes to generate classifiers or features for the development of machine learning-based models like Support Vector Machine (SVM: see Materials and Methods section). Other non-SVM computational techniques are also built upon structural and physico-chemical models derived from subsets of RBP sequences.

In a comprehensive assessment analysis called Nucleic Acid Binding Prediction Benchmark (*NBench*), Miao and Westhof (2015 & 2016) demonstrated the strengths and weaknesses of such computational methods on data sets of different compositions to avoid the bias introduced by curated data sets used for their development. All of these SVM and non-SVM tools for the prediction of RNA-binding residues are listed with their important features in the **Table 1.5**.

Table 1.5 ***The important features of the tools for the identification of RNA-binding residues.*** *(Reference: Miao & Westhof* et al.*, 2015)*

Name of the software	Sequence/structure features	Main predictive feature	Reference
DR_bind1	Structure	ASA, EC, Q	Chen *et al.*, 2007
KYG	Structure	PSSM, RP	Kim *et al.*, 2006
RBRDetector	Structure	PSSM	Yang *et al.*, 2014
BindN	Sequence	HP, Q	Wang *et al.*, 2006
BindN+	Sequence	PSSM, HP, Q	Wang *et al.*, 2010
PPRInt	Sequence	PSSM, RP	Kumar *et al.*, 2008
PRBR	Sequence	PSSM, RP	Ma *et al.*, 2011
Predict_RBP	Sequence	PSSM, RP, ASA	Wang *et al.*, 2011
RBRIdent	Sequence	PSSM, SS, Q	Xiong *et al.*, 2015
RNABindR	Sequence	PSSM	Terribilini *et al.*, 2011
RNABindRPlus	Sequence	PSSM	Walia *et al.*, 2014
aaRNA	Sequence and Structure	PSSM, ASA, SS, EC	Li *et al.*, 2014
RNAProSite	Sequence and Structure	Other parameters	Sun *et al.*, 2016
PRNA	Sequence and Structure	PSSM, RP, ASA, HP	Liu *et al.*, 2010
RBscore	Sequence and Structure	RP, ASA, Q, SA	Miao & Westhof, 2015

(Abbreviations used in the 3rd column - PSSM: position specific scoring matrix derived from sequence alignment, RP: residue propensity, ASA: accessible surface area, HP: hydrophobicity, SS: secondary structure, EC: conservation entropy, Q: electrostatic/pKa, SA: structural alignment)

A successful model for RBP RNA-binding residue prediction should perform equally well across different data sets with stable predictive ability reflected in terms of performance measures such as true positive rate, false positive rate, and accuracy. Based on these criteria, *BindN+* (Wang *et al.*, 2010), *RNABindRPlus* (Walia *et al.*, 2014), *aaRNA* (Li *et al.*, 2014), *RNAProSite* (Sun *et al.*, 2016) and *RBscore* (Miao & Westhof, 2015) show a good overall performance, however they still demonstrated data set bias. Some RBP-related methods such as *RNABindR* (Terribilini *et al.*, 2007), *KYG* (Oanh *et al.*, 2006), *aaRNA* (Li *et al.*, 2014), *RNAProSite* (Sun *et al.*, 2016), and *RBscore* (Miao & Westhof, 2015), though not trained on DNA-binding proteins, could predict DNA-binding residues as well. It was observed in the *NBench* study that machine-learning methods have better discriminative ability for RNA-binding residues compared to other methods. Several examples of these machine-learning tools are *PRNA, Predict_RBP, RNABindRPlus* (Walia *et al.*, 2014), and *RBscore_SVM* (Miao & Westhof, 2015). However, this does not qualify them as overall good predictive approaches, as they did not perform equally well on all the data sets. This poses the challenge of dealing with the non-discriminative performance of such predictive tools, which may lead to false-positive predictions of a non-RBP as an RBP. The structure-based tools performed better than sequence-based tools, indicating that the structural information captures information on RNA-binding residues with better specificity.

1.5.2 Prediction of RNA-binding proteins

In principle, tools for RNA-binding residue prediction should identify RBPs based on the assumption that RNA-binding residues will be recognized only in RBPs. However, the above tools were not designed for this purpose. In the absence of a pre-defined set of experimentally confirmed RBPs, such tools may lead to the false characterization of non-RBPs as having RNA-binding residues. Only a few computational tools, such as *RNApred* (Kumar *et al.*, 2008), *SPOT-Seq-RNA* (Yang *et al.*, 2014), and *catRAPID signature* (Livi *et al.*, 2016) have been specifically developed for the identification of RBPs. These programs characterize RBPs using sequence-based features, such as biochemical properties, structural properties, and their evolutionary relationship (Zhao *et al.*, 2011; Puton *et al.*, 2012; Si *et al.*, 2015). *RNApred* is a web-based application for the prediction of RBPs from protein sequences based on amino acid composition and position-specific scoring matrix (PSSM) based evolutionary information. A large number of proteins can be analyzed using the

composition-based method; however, PSSM-based approaches can process only one protein at a time. The underlying software is trained on positive and negative sets of proteins to build a SVM-based method for the classification of RBPs and non-RBPs. *SPOT-Seq-RNA* is a template-based method, which uses the information derived from RBP structures. The software is implemented as a web-server and command-line tool and can process only one query protein at a time. The software has been coupled with SPARKS X (Zhou & Zhou, 2005) and DRNA (Zhao *et al.*, 2011) for the prediction of template-based structures and binding affinity, respectively.

The tool *catRAPID signature* is relatively new software that uses sequence-based physico-chemical features for the identification of RBPs. The cut-off for the identification of RBPs has been derived from an SVM, which has been trained on human proteins. In addition to RBP recognition, this software also predicts RNA-binding regions in the query proteins. Like other tools, it requires an amino acid sequence as a query, but can process 100 proteins at a time. In past decade, a significantly higher number of tools for RNA-binding residue predictions within RBPs have been developed, compared to the tools developed for identification of RBPs themselves. While an improvement in accuracy can be observed for both types of tools, an obvious data set bias in the underlying training and test sets of lower diversity and smaller size exists, as they are curated specifically for the method in question. Other limitations include the number of query proteins that can be processed by the tools, which is often only one query at a time; hence, these tools are not conveniently applicable on large data sets.

Most of these computational methods, specifically for the identification of RNA-binding residues, are computationally expensive and are limited in their functionalities due to the restricted nature of their training data sets. Moreover, due to the limitation of number of query proteins the tools are designed to process, makes them unsuitable for a large-scale data analysis (Miao & Westhof, 2015).

1.6 Aim of the study

The main aim of this study is to establish a bio-computational approach for the identification and characterization of novel RBPs in bacterial proteomes using the pathogen *Salmonella* Typhimurium as a model organism. The first part of the thesis describes a

computational pipeline that I developed for the large-scale analysis of protein sequences to identify putative RBPs based on RBD classes. The second section of the thesis gives an overview of high-throughput sequencing-based experimental characterization of the resulting computationally identified RBPs.

Salmonella species (Gram-negative enterobacteria) are intracellular pathogens of eukaryotic hosts such as humans, and are transmitted via contaminated water and food (Santos *et al.*, 2001). *Salmonella* is well characterized and has been studied intensively for its regulation of virulence and survival factors, including post-transcriptional regulation by riboregulators. Therefore, it was used as a model to expand our search for novel RBPs in a systematic manner. A proteome-wide computational-based identification of RBPs, followed by RNA co-immunoprecipitation and high-throughput sequencing (RIP-Seq) (Selth *et al.*, 2009; Zhao *et al.*, 2010) of candidate proteins was carried out in *Salmonella* Typhimurium SL1344. These proteins were further characterized using publicly available infection-relevant transcriptomes and genetic data sets, which collectively present a resource of RBPs in *Salmonella* that will aid in designing further exploratory and validation studies in bacteria.

Chapter 2

Computational identification and characterization of RBPs using APRICOT

The important discoveries on the regulatory roles of RBPs in eukaryotes have stirred interest in the scientific community to move on from the characterization of a handful of RBPs to the identification of RBPs on a large-scale. Many RBPs within a complex with other proteins function in a cooperative or competitive manner, which control the levels or behaviour of RNAs depending on biological conditions (Ciafre & Galardi, 2013). Although these experimental studies revealed considerable details about the mechanisms and biological roles of eukaryotic RBPs, it has been speculated that there are many more unidentified and uncharacterized RBPs that might play central roles in different biological pathways. It is technically possible to capture RBPs *in vivo* that are still prone to biases. Furthermore, the existing approaches require a specific biochemical characteristic such as mRNA poly(A) tails in the case of eukaryotic mRNAs and relatively sophisticated experimental set-ups for other organisms restricting our understanding of RBPs in other domains of life. Nonetheless, the catalogue of information now available from studies of RBPs in eukaryotes may now be applied to identify and characterize the important players of RNA-protein interactions in bacteria via the development of computational approaches to identify RBPs and their RNA-binding residues.

In this endeavour, I intended to develop a bioinformatic pipeline called APRICOT (Analyzing Protein RNA Interaction using Combined Output Technique). This software carries out sequence-based analysis of entire proteomes to identify conserved functional domains associated with experimentally validated RBPs, such as RBDs, and provide an overview of their physico-chemical properties. One of the main emphases of the pipeline is to computationally identify RBP candidates in large sets of query proteins. The challenging aspect of processing large-scale data is taken into consideration in order to ultimately apply this method to the bacterial proteome sets. In this chapter, an integrated computational pipeline has been designed with state-of-the-art sequence-based algorithms and tools to infer the statistical significance of computational predictions of protein functions.

2.1 Overview of APRICOT pipeline for RBP identification

APRICOT incorporates several distinct modules for the identification and characterization of proteins, which are assembled into a command-line tool. A human RBP PTBP1 (Sawicka *et al.,* 2008) is used as an example to describe representative modules of the pipeline. PTBP1 is an mRNA regulator that contains several repeated RBDs, specifically the highly abundant eukaryotic RRM domain (Dye & Patton, 2001).

The pipeline can be explained in three parts: program input, intermediate analysis and program output (**Figure 2.1A**). The input requires two set of information, which are the query proteins and functional classes of interest. The query proteins are subjected to the intermediate analysis, which included processes such as domain prediction, selection of domains of interest, and scoring and ranking of the predicted domains. Upon analysis by the pipeline, the program output is generated that include the selected proteins with the functional domains of interest and their annotation by different biological functions. The pipeline and its components have been described below in detail and shown in **Figure 2.1B**.

2.1.1 Program input

APRICOT requires two inputs for its execution: query proteins and the functional class of interest (**Figure 2.1B**).

1. Query proteins:

The query proteins can be provided either as a list of gene IDs, protein IDs, or amino acid sequences (subcommand query). When protein IDs (for example, P26599 for PTBP1) are supplied to APRICOT, queries are directly searched against the UniProt database to retrieve their amino acid sequences and available annotations. The retrieved sequences are then used as input for domain prediction. If the users are interested in a particular species, the search for query-associated information can be limited to that specific organism by providing a corresponding taxonomic identifier (for example, 9606 for human). Since APRICOT has been designed to process multiple queries, the motif prediction can be dynamically carried out for the functional characterization of an entire proteome set corresponding to a taxonomy ID (subcommand taxid). When amino acid sequences are supplied as queries, APRICOT skips the sequence and their annotation retrieval step and directly proceeds to the domain predictions.

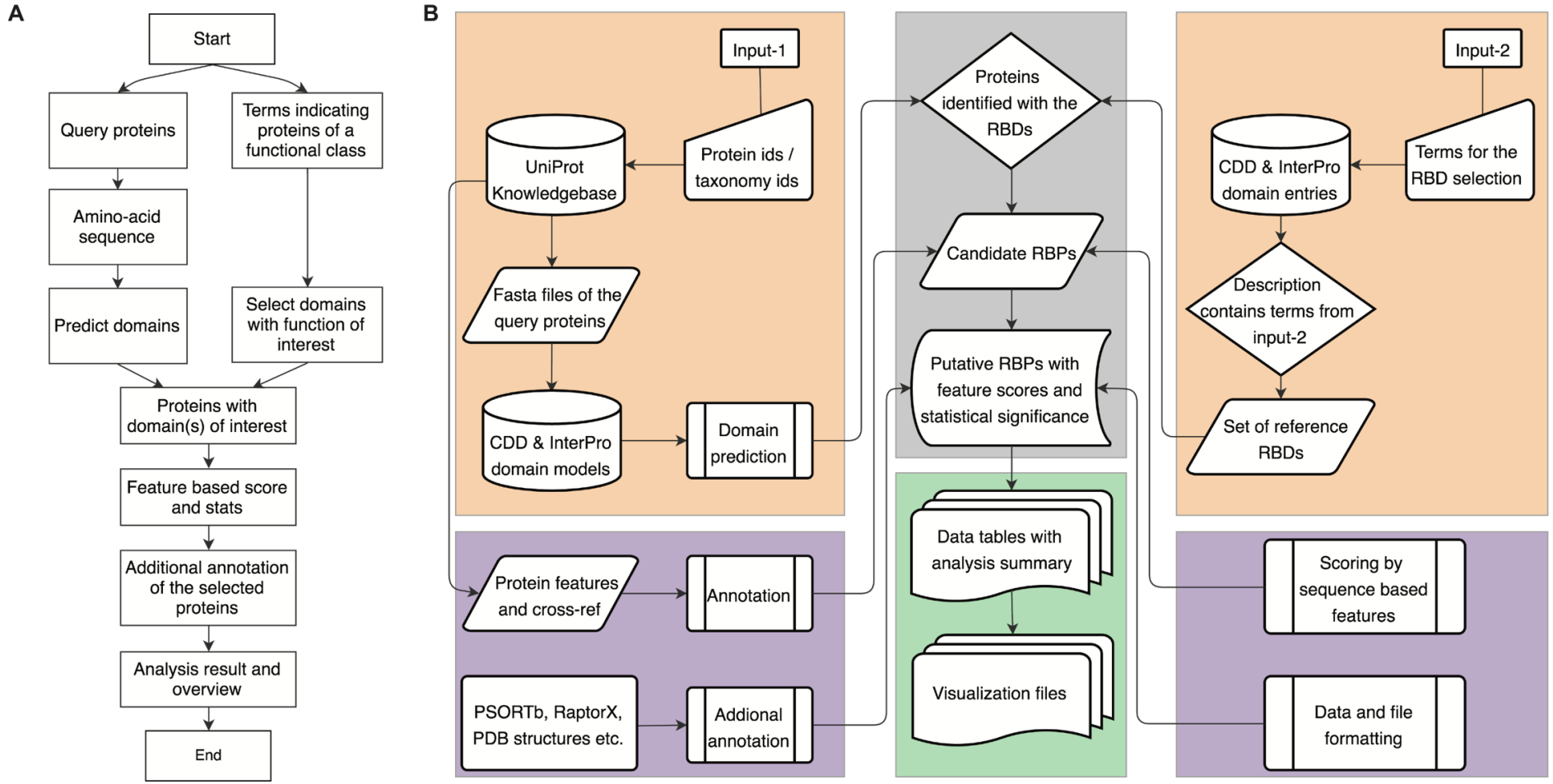


Figure 2.1 *Architecture of APRICOT. (**2.1A**) A simplified overview of the processes involved in APRICOT analysis. (**2.1B**) Flow-chart showing different components of APRICOT pipeline for the characterization of RNA-binding proteins. Modules for the primary analysis involving the processing of user-provided inputs (orange boxes) and the downstream analysis, which includes modules for the identification of RBPs candidates (grey box) and the modules for the annotation and feature-based scoring of putative RBPs (purple boxes). APRICOT generates comprehensive results for each analysis, which are represented by means of tables and visualization files (green box).*

2. Functional class of interest:

Depending of the functional class(s) of interest, users must provide an appropriate list of terms or keywords (subcommand keywords) such as names of domain families, Pfam IDs, GO or MeSH terms, which are referred herein as domain selection keywords. APRICOT employs a string-based search algorithm using these terms to select relevant domains or reference domain sets from the relevant resources, which are then subsequently used to identify proteins of functional relevance that contain these domains. Multi-word terms like 'RNA binding' can also be provided as 'RNA-binding' by using a hyphen as a connector. In such cases, APRICOT looks for co-occurrences of these terms in the same context in an annotation. There is no strict convention on the nature of keywords that can be provided as domain selection keywords. However, due to the domain selection method, any general or ambiguous terms should be avoided to exclude irrelevant domains from the analysis. For example, use of the term 'RNA' alone does not unambiguously refer to RBDs. Therefore, users should use definite terms such as 'RNA-binding' or domain families like the classical RBDs RRM (RNA recognition motif) or KH (K Homology) to indicate groups of RBDs. Alternatively, use of MeSH terms (Liu *et al.*, 2009) can also provide more specific information for the domain selection. For instance, a total of 198 descriptors have been listed for the MeSH term 'RNA binding' in the NCBI database, which can be directly used as the domain selection keywords for the collection of specific RNA binding protein-related domains.

In order to maintain stringent selection of truly functional domains, APRICOT by default does not allow the selection of a domain entry if the annotated domain selection term has either a prefix or a suffix. This opens the possibility of omitting some relevant entries from the domain selection keywords, but it also ensures exclusion of several non-relevant domains that might be included by chance. Nonetheless, users can allow a prefix or suffix by including a hash symbol (#) at the beginning or end of a term, respectively. For example, by inputting '#RNA-binding' one can allow the inclusion of 'tRNA-binding', 'mtRNA-binding', etc., while use of 'RNA-bind#' would allow varying verb forms for bind such as binder, binding etc. Furthermore, one can allow both prefixes and suffixes (for example, #RNA-bind#).

Optionally, for the classification of predicted domains, a second set of keywords can be provided (*result classification keywords*). This list can be comprised of terms associated with biological functions, enzymatic activities, or specific biochemical or structural features. For example, the predicted RNA-related domain data could be divided into the classification tags

of RRM, ribosome, synthetase, helicases, etc. Such classifications can help users tremendously in navigating through large data sets or for the selection of a representative protein for a certain function conferred by the domains. When users do not provide result classification terms, APRICOT uses the domain selection terms for this purpose as well.

2.1.2 Modules for domain prediction and annotations

The core functionality of APRICOT involves a multi-step process for the selection of proteins by identifying functional sites or domains of interest in their sequences, followed by their annotation with various biological features. These steps are described below and illustrated in **Figure 2.2**.

1. Reference databases and associated tools

APRICOT requires a set of query proteins as input for which the presence of RBDs is to be determined. Basic information, e.g. amino acid sequences and taxonomy data, are retrieved from the UniProt Knowledgebase (Magrane & UniProt Consortium, 2011). In addition, a reference domain set is collected from domain databases based on functional classes specified by the user.

The domain resources used in this study are Conserved Domain Database (CDD) (Marcher-Bauer *et al.*, 2015) and InterPro (Mitchell *et al.*, 2015), which consist of predictive models and signatures representing protein domains, families, and functional sites from multiple publicly-available databases. CDD includes domain entries as Position-Specific Score Matrices (PSSMs) that are generated from multiple sequence alignments (MSA) of representative amino acid sequences obtained from several domain databases, including Pfam (Finn *et al.*, 2016), TIGRFAM (Haft *et al.*, 2003), SMART (Schultz *et al.*, 1998; Letunic *et al.*, 2014), COGs (Tatusov *et al.*, 1997; Galperin *et al.*, 2015), several NCBI curated domains like PRK or Protein Clusters (ONeill *et al.*, 2010), and multi-model superfamilies of proteins (Gough *et al.*, 2001). For the identification of domains in a given protein sequence, the PSSM entries in CDD are queried via Reverse Position-Specific BLAST (RPS-BLAST), a variant of the popular Position-Specific Iterative BLAST (PSI-BLAST) (Altschul *et al.*, 1990; Altschul & Koonin, 1997). CDD (v3.14) contains annotations for 50,648 domains. Entries from every domain resource are assigned an individual PSSM identifier that allows redundant entries of domains.

InterPro is a similar consortium that consists of domain entries as predictive models and signatures obtained from different databases, such as Pfam (28, 29), TIGRFAMs (Haft *et al.*, 2003), SMART (Schultz *et al.*, 1998; Letunic *et al.*, 2014), PROSITE patterns and profiles (Sigrist *et al.*, 2013), HAMAP (Pedruzzi *et al.*, 2015), PRINTS (Attwodd *et al.*, 2012), PIRSF (Wu *et al.*, 2004), ProDom (Bru *et al.*, 2005), PANTHER (Mi *et al.*, 2010), GENE3D (Lam *et al.*, 2016), and SUPERFAMILY (Gough *et al.*, 2001). Most of these databases contain domain entries as Hidden Markov Models (HMM) (Krogh *et al.*, 1994), probabilistic models derived from sequence alignments, which capture information on both substitution and indel frequencies. A query protein can be queries against these domain entries using HMMER based tools (Mistry *et al.*, 2013). Several member databases contain PSSM domain models built from the multiple sequence alignments of representative amino acid sequences from the UniProt protein database, which can be queried by BLAST-based methods or a single model search algorithm, which have been integrated into InterProScan 5 (Jones *et al.*, 2014). As of May 2016, InterPro (v.57) contained 29,175 domain models of which several are annotated with Gene Ontology (GO) terms (Ashburner *et al.*, 2000).

InterPro and CDD consortia have only three databases in common (Pfam, TIGRFAM, and SMART) that account for approximately 20,000 domains.

2. Selection of reference domain set

A string-based selection of domain families and functional motifs is carried out using the *domain selection keywords* to create a reference domain set. Domains are selected from the collections of domain entries from the domain databases collected by the CDD and InterPro consortiums when they match at least one of the provided terms in their annotations. In this analysis, I considered recent human interactome studies (Castello *et al.*, 2012; Baltz *et al.*, 2012; Kwon *et al.*, 2013; Gerstberger *et al.*, 2014) as comprehensive resources for building a reference RBD set. To report only high-confidence RBPs by avoiding the selection of ambiguous and functionally irrelevant domains, all the classical RBDs were included in *domain selection keywords* (**Figure 2.2A**). In order to account for ribosomal proteins, 109 terms related to RNA-binding ribosomal domains (Gerstberger *et al.*, 2014) were also included in *domain selection keywords*. An additional term '#RNA-bind#' was introduced to include any additional RBDs in the reference set that are well described as RBDs in databases but are not classified as classical RBDs (**Figure 2.2B**). Using these domain selection keywords, a total of 4,951 RBD entries were curated from CDD (1,951 entries) and InterPro (3,000

entries), referred as *reference domain set*, which was used for filtering domain predictions in the downstream analysis.

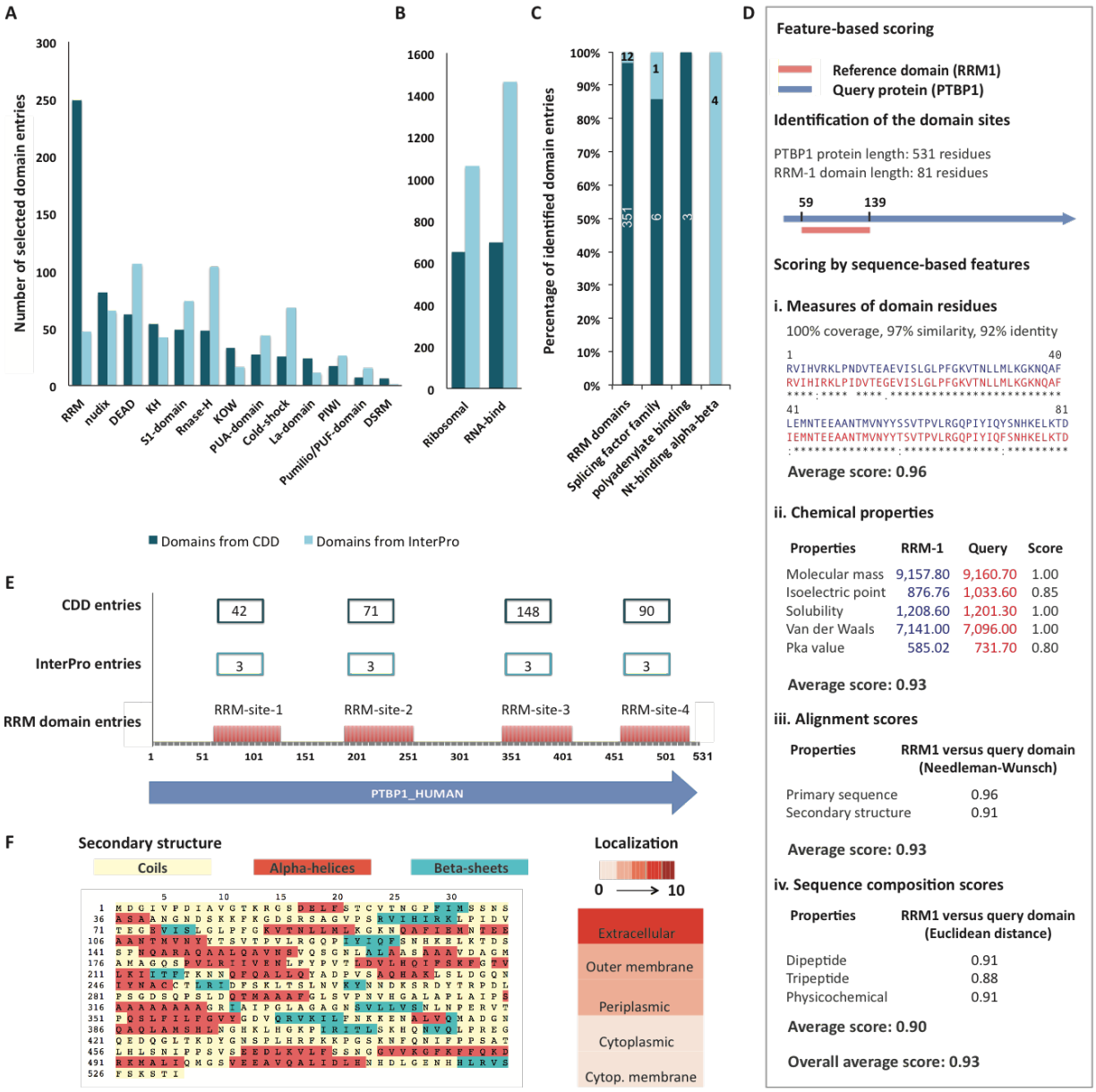


Figure 2.2 ***Different components of APRICOT for the characterization of RBPs illustrated using the human RBP PTBP1.***

*(**2.2A**) Bar chart showing the distribution of the known RNA binding domains collected from the CDD and the InterPro consortium. Several of these domains were selected by more than one domain selection term. (**2.2B**) Additional domains selected by RNA-binding ribosomal domains and the term 'RNA-bind'. (**2.2C**) Domain entries from CDD and InterPro database, which were identified by APRICOT in PTBP1. (**2.2D**) A schematic workflow illustrating different processes involved in feature-based scoring resulted from a comparative analysis of RRM-1 domain (RRM1_PTBP1) and the corresponding domain identified in PTBP1. As shown in the scheme, the features involved in this analysis have been classified into four categories, each comprising of specific set of sequence-based features. The*

*features are scored by Bayesian probabilities in a range of 0 to 1, where 1 signifies a complete match between the reference and the domain identified in the query. (**2.2E**) The four RRM sites in PTBP1 corresponding to different RRM entries from CDD and InterPro. (**2.2F**) Visualization of additional annotations of PTBP1 protein by secondary structure and probability of subcellular localizations generated by APRICOT.*

3. Domain prediction

In this step, query amino acid sequences are screened for all the possible domains from the databases without filtering a certain functional class. The sequences are then subjected to domain prediction using RPS-BLAST and InterProScan to query their corresponding databases CDD and InterPro, respectively (**Figure 2.2C**). By default, APRICOT uses both CDD and InterPro for the domain predictions. However, users can choose one of the databases to reduce the run-time. Since the primary requirement of this module is the amino acid sequences of the query proteins in FASTA format, users can analyze novel or partial sequences even when the gene/protein IDs are unknown or absent.

4. Selection of proteins by functional domains of interest

This module allows the selection of relevant proteins from the query sets based on the predicted domains obtained in the previous step. The proteins are considered as candidates if they contain one of the domains of interest. Cut-offs for various statistical parameters (discussed below) can be defined for the selection of the predicted domains to identify such candidates.

5. Feature-based scoring

Divergent domains can be predicted in a protein, which might correspond to partial sequence of a functional domain. This module ranks the domain predictions by their functional relevance by accounting for such partial conservation of domains. A comparative analysis is carried out between the protein sequences that are predicted in the candidate proteins as a specific region of a domain of interest and the corresponding fragments of their reference consensus sequence. This comparison is done for a number of sequence-based features, which are discussed below in detail.

5.1. Global alignment scores (primary sequence and secondary structure)

A global alignment refers to an alignment between two sequences from the beginning till the end based on its similarity in order to find out the best possible alignment. To calculate the extent of similarity between the predicted domain region in the query and its corresponding reference sequence, APRICOT carries out their global alignments using the Needleman-Wunsch algorithm (Needleman & Wunsch, 1970) implemented in Biopython (Chapman *et al.*, 2000; Cock *et al.*, 2009). This algorithm uses dynamic programming to compare biological sequences and uses match scores and gap penalties, although in APRICOT no gap penalty was introduced. The similarity scores are calculated for the global alignments of two sequence features: primary amino acid sequence and secondary structure. The similarity scores between the query and reference sequences range from 0 to 1, where 1 is a complete match.

5.2. Chemical properties (average mass, pKa, and pI)

The chemical properties of a protein can help in determining their biological activities in living cells, their involvement in cellular processes, their three-dimensional folding, and their structural stability. Therefore, APRICOT analysis also computes the similarity between the region of predicted domains in the query proteins and its corresponding fragments in the references for several chemical properties (average mass, pKa, and isoelectric point (pI)) (Zamyatin, 1972; Chothia, 1976; Tanford, 1968; P.J. Linstrom and W.G. Mallard, Nahway, N.J., 11(1989)). The values for each feature in the predicted domains are divided by the values of the corresponding feature in the reference domains, and a score in the range of 0 to 1 is obtained which suggests the extent of functional similarity in the predicted domains. This analysis is based on the assumption that a relative domain conservation in query and reference sequence will result to similar chemical properties.

5.3. Euclidean distances of protein compositions (di-peptides, tri-peptides, and physico-chemical properties)

The composition of a protein refers to the fraction of each amino acid group (di-peptides and tri-peptides, which refers to the occurrences of 2 and 3 amino acids in an ordered sequence) or properties (physico-chemical or secondary structure) in the amino acid sequence. It has been shown that function-specific information (for example, subcellular localization, secondary structure, enzyme families, and membrane protein types) can be specified by such compositions (Eisenhaber *et al.*, 1996; Reczko & Hatzigerrorgiou, 2004; Cai & Chou, 2006; Shen & Chou *et al.*, 2008; Habib *et al.*, 2008; Otaki *et al.*, 2010; Yu *et al.*, 2010;

Källberg *et al.*, 2012). Therefore, the similarity in composition between the predicted domains and their corresponding references can reflect the functional significance of the predictions and therefore inform the user of the putative biological function conferred by the identified domain. These similarities are calculated by Euclidean distance of di- and tri-peptide composition and composition of physico-chemical properties between the reference and the predicted domains in the query protein. The similarity score (1-Euclidean distance) is represented in a range of 0 to 1, where 1 stands for an absolute match.

5.4. Similarity between predicted sites and reference domains (domain similarity, identity, gaps and coverage)

The last set of properties considered for feature-based scoring is the sequence similarity by means of identity, physico-chemical similarity and gaps, as well as coverage obtained for the predicted domain sites in the query, with respect to the sites in their reference domains. The domain coverage is calculated by dividing the residue counts of the predicted domain site in the query protein by the original length of the reference domain. The similarity, identity, and gaps are calculated by dividing the corresponding residue counts in the predicted domain by the calculated domain coverage (rather than the full length of the domain). Each of these parameters is reported in a value range of 0 to 1. The coverage value of 1 indicates an identification of a complete domain in the query. The similarity and identity value of 1 indicates an absolute match in the fraction of domain identified in the query. The gap value of 0 means no gap in the sequence, which is represented as the measure of 1- gap so that a score closer to 1 represents a favourable scenario.

5.5. Scoring and ranking of predicted domains in a protein set

The relative similarity between the predicted functional site and the reference domain consensus for these sets of features are calculated. The feature-based scoring is carried out for each predicted domain in each query protein in a range of 0 to 1 as described in the previous sections (**5.1-5.4**). These scorings represent the Bayesian probabilistic score functional potential of the predicted motifs, where 1 indicates the highest probability (**Figure 2.2D**). Ultimately, these features are combined and ranked with other putative proteins of interest to determine high confidence RBP predictions. This combined output of relative similarity score denotes the 'combined output' (the CO in the acronym of APRICOT).

6. Additional annotations of the selected proteins

Upon selection of proteins of functional relevance, users can choose to further annotate these proteins by information like sub-cellular localization by PSORTb (Yu *et al.*, 2010), 8-state secondary structure by RaptorX (Källberg *et al.*, 2012), additional GO annotation, and structural homologs, which are discussed below in detail.

6.1. Identification sub-cellular localization of the proteins

Information regarding the subcellular localization of a protein provides further insight into the potential function of a protein by giving an idea of their roles in protein complexes or biological pathways. Several tools have been trained on large data sets to capture such information in bacterial proteins, such as the SVM-based CELLO2GO (Yu *et al.*, 2014) and PSLDoc (Chang *et al.*, 2008), amino acid composition-based SLP-Local (Matsuda *et al.*, 2005), composition and structural features-based PSL101 (Su *et al.*, 2007), composition and GO clustering-based Gneg-PLoc (Chou & Shen, 2006), Gpos-PLoc (Shen & Chou, 2007), Cell-PLoc (Chou & Shen, 2008), Gneg-mPLoc 2010), combined method-based SubcellPredict (Bulashevska & Eils, 2006), and HensBC (Niu *et al.*, 2008). Software has also been developed to predict such information for eukaryotic proteins, such as neural-network and HMM-based PredSL (Petsalaki *et al.*, 2006), domain-based PSCL (Wang *et al.*, 2011), those based on PseAA (pseudo amino acid composition) such as PseAAC (Shen & Chou, 2008), as well as SVM-based SecretomeP (Bendtsen *et al.*, 2004). A more complete list of such localization tools is available online at http://www.psort.org/. One of the most recently developed tools that incorporates knowledge obtained from the data sets associated with all kingdoms of life is PSORTb. PSORTb provides a list of five localization sites (cytoplasmic, cytoplasmic membrane, cell wall, extracellular, and secondary localization) and an associated probability score (0-10 indicating low to high probability). A standalone version of PSORTb (v.3.315) is integrated into the APRICOT pipeline for computational prediction of the subcellular localization of selected proteins.

6.2. Secondary structure calculation by RaptorX

In principle, an amino acid sequence that aligns well with annotated proteins could be considered as a functional homolog. However, amino acid conservation at the sequence level is not always obvious when dealing with the sequences where only functional domains are conserved whereas rest of the sequence shares only secondary structure homology. In such cases, the selection of true homologs based on primary sequences is difficult. To address this problem, the candidate proteins can also be compared to known proteins at the

structural level in APRICOT. The structure prediction tool RaptorX (Källberg *et al.*, 2012) has been integrated in the pipeline for the prediction of protein secondary structures.

6.3. Tertiary structure homologs from Protein Data Bank (PDB)

Tertiary structures are critical to identify the ligand partners of proteins and achieve a high-resolution annotation. Out of several millions of proteins available in the non-redundant (nr) database of NCBI (Pruitt *et al.*, 2009), only 105,417 proteins and 5,198 protein/nucleic-acid complexes (November 2015) have been crystalized. There are numerous computational tools available for the estimation of tertiary structures of proteins, such as Phyre2 (Kelley *et al.*, 2015), CPHModels (Nielsen *et al.*, 2010), and I-TASSER online (Roy *et al.*, 2010). These methods are computationally demanding and are available only as web-servers, making their integration difficult into automated workflows. Hence, in order to provide a quick insight into the potential binding mechanisms of selected proteins, APRICOT extracts the structure homologs and the available annotation from the PDB database.

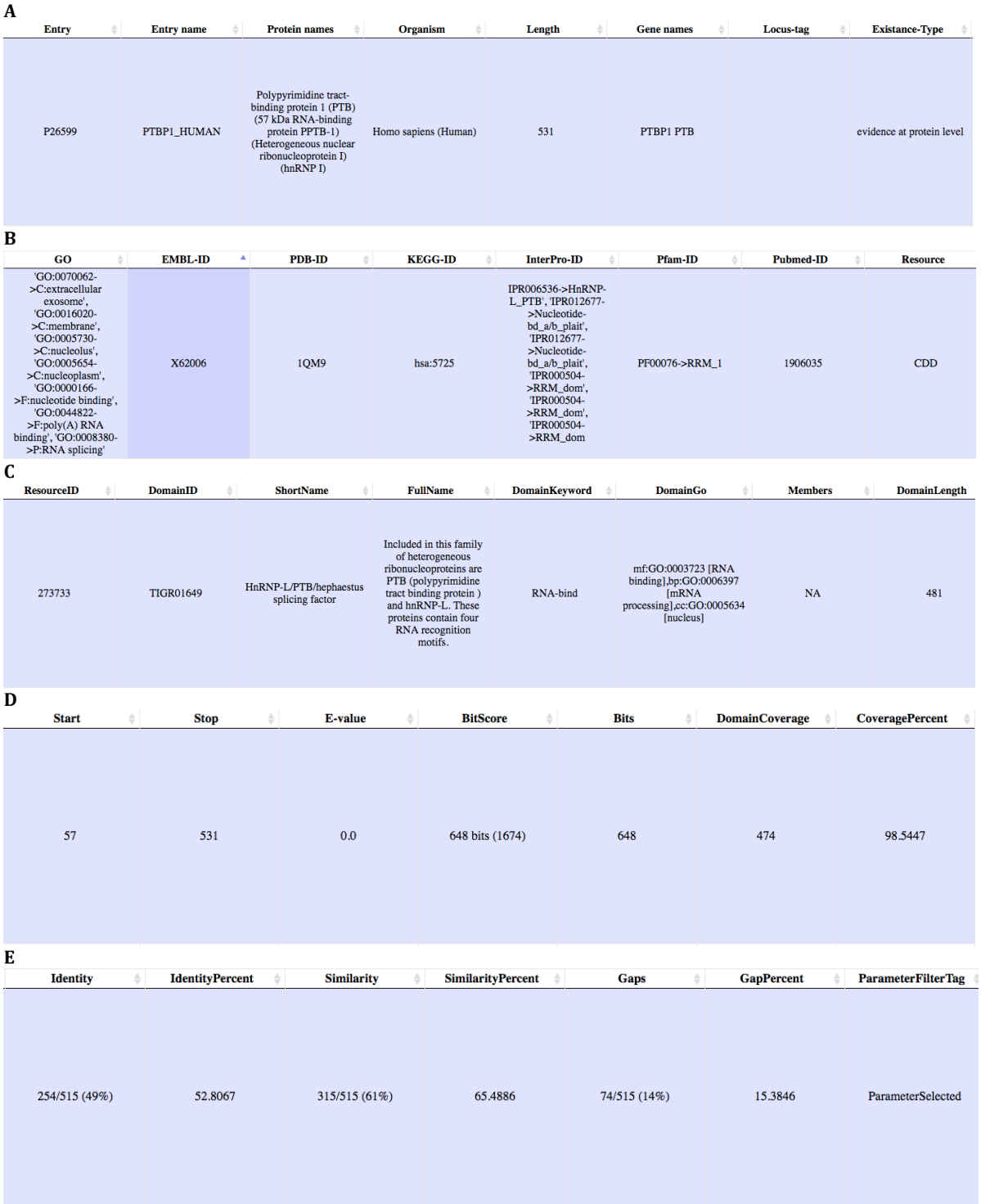
6.4. Gene Ontology

The Gene Ontology or GO Consortium (Reference Genome Group of the Gene Ontology Consortium, 2009) is a bioinformatics initiative for unifying annotation by means of a controlled vocabulary. GO terms are widely used for standard annotation of a gene with various information, including cellular localization, biological processes, and molecular function. GO is determined by extracting all GO terms available for a protein in the UniProt database (Magrane M & UniProt Consortium, 2011) and for its domains from the InterPro and CDD databases. In order to achieve a broader GO-catalogue for each candidate protein, Blast2GO (Conesa & Götz, 2008) can also be executed from APRICOT (subcommand blast2go) when already installed by users.

2.1.3 Program output

A comprehensive result is returned by APRICOT at each step of the analysis and is stored with relevant information that serves as the input for subsequent steps. For example, the data for predicted domains can be repeatedly used for extracting proteins of different functional classes. The selected proteins are provided in a tabular format with the statistics on domain prediction and corresponding annotations obtained from UniProt and the comparative analysis (Figure 2.3). To allow easy navigation through the large-scale analysis data, the results can be classified using result classification keywords into smaller subsets of

proteins with similar enzymatic activities or other functional aspects. Additionally, graphs and charts are provided to aid the visualization of the resulting data.



A

Entry	Entry name	Protein names	Organism	Length	Gene names	Locus-tag	Existance-Type
P26599	PTBP1_HUMAN	Polypyrimidine tract-binding protein 1 (PTB) (57 kDa RNA-binding protein PPTB-1) (Heterogeneous nuclear ribonucleoprotein I) (hnRNP I)	Homo sapiens (Human)	531	PTBP1 PTB		evidence at protein level

B

GO	EMBL-ID	PDB-ID	KEGG-ID	InterPro-ID	Pfam-ID	Pubmed-ID	Resource
'GO:0070062->C:extracellular exosome', 'GO:0016020->C:membrane', 'GO:0005730->C:nucleolus', 'GO:0005654->C:nucleoplasm', 'GO:0000166->F:nucleotide binding', 'GO:0044822->F:poly(A) RNA binding', 'GO:0008380->P:RNA splicing'	X62006	1QM9	hsa:5725	IPR006536->HnRNP-L_PTB', 'IPR012677->Nucleotide-bd_a/b_plait', 'IPR012677->Nucleotide-bd_a/b_plait', 'IPR000504->RRM_dom', 'IPR000504->RRM_dom', 'IPR000504->RRM_dom	PF00076->RRM_1	1906035	CDD

C

ResourceID	DomainID	ShortName	FullName	DomainKeyword	DomainGo	Members	DomainLength
273733	TIGR01649	HnRNP-L/PTB/hephaestus splicing factor	Included in this family of heterogeneous ribonucleoproteins are PTB (polypyrimidine tract binding protein) and hnRNP-L. These proteins contain four RNA recognition motifs.	RNA-bind	mf:GO:0003723 [RNA binding],bp:GO:0006397 [mRNA processing],cc:GO:0005634 [nucleus]	NA	481

D

Start	Stop	E-value	BitScore	Bits	DomainCoverage	CoveragePercent
57	531	0.0	648 bits (1674)	648	474	98.5447

E

Identity	IdentityPercent	Similarity	SimilarityPercent	Gaps	GapPercent	ParameterFilterTag
254/515 (49%)	52.8067	315/515 (61%)	65.4886	74/515 (14%)	15.3846	ParameterSelected

Figure 2.3 ***Screenshot of an HTML output of APRICOT analysis comprising of information on domain prediction and corresponding annotations.***

*The columns in the table have been explained in three parts: (**2.3A**) annotation of the protein retrieved from the UniProt knowledgebase, (**2.3B**) annotation of the domains in the protein selected by APRICOT, and (**2.3C**) statistical output for each parameter associated with the domain predictions.*

2.2 Data sets used in this study

2.2.1 Training sets

For the identification of the most suitable parameters and their corresponding cut-offs for domain selection, training sets were collected from the manually-curated and reviewed subset of the UniProt Consortium – SwissProt (UniProt Consortium, 2015). A positive set of proteins was selected by using the keyword 'RNA-binding'. A second set of proteins was selected by using all the terms indicating functional association of proteins with nucleic acid. A third set comprising all the uncharacterized and hypothetical proteins from the database also was selected. These three sets of proteins were subtracted from the SwissProt data, and the remaining data (consisting of 271,219 proteins) were considered as the resource for a negative set. All redundant protein sequences from both positive and negative sets were removed by clustering the sequences using BLASTclust (Altschul *et al*, 1999) with a 90% sequence identity cut-off. Following these steps, a total of 4,779 non-redundant proteins were compiled into the positive set and a set of 5,834 proteins was selected for the negative set, referred to henceforth as 'SwissProt-positive' and 'SwissProt-negative', respectively.

2.2.2 Test sets

To consistently evaluate the sensitivity, specificity, and accuracy of APRICOT, a pair of positive and negative set was obtained from NCBI Reference Sequence (RefSeq), a non-redundant (nr) database (Pruitt *et al.*, 2012), using the terms 'RNA-bind' and 'periplasmic', respectively. The former term retrieved 4,470 RBPs from various organisms in all kingdoms. The term 'periplasmic', which retrieved 5,836 bacterial periplasmic proteins, was considered as a resource for non-RNA-binding proteins based on the assumption that the majority of periplasmic proteins lack RBDs. Using BLASTclust from the NCBI-BLAST package (Altschul *et al*, 1999), the proteins in each set were clustered by 75% sequence similarity, which resulted into 687 proteins in the positive set and leaving 1,199 proteins in negative set, henceforth referred as 'nr-positive' and 'nr-negative', respectively. An additional pair of positive and negative sets was obtained from the RNApred webserver (Si *et al.*, 2015), which will be referred as 'RNApred-positive' (377 proteins) and 'RNApred-negative' (355 proteins).

The sensitivity of the pipeline was also tested on other positive data sets collected from various resources, such as RBPDB (Cook *et al.*, 2011), RNAcompete (Ray *et al.*, 2009), rbp86, rbp109, rbp107 (Cheng *et al.*, 2008), and RBRIdent (Xiong *et al.*, 2015), consisting of 1,101,

205, 86, 109, 107 and 281 proteins, respectively. These negative and positive RBP data sets are listed in the **Table 2.1**.

Table 2.1 ***List of all the positive and negative RBP data sets used in the development and benchmarking of APRICOT.***

Data set	Number of proteins	Comments	Reference	Link
SwissProt-positive	4557	Non-redundant protein sets selected from SwissProt database using the keyword 'RNA-bind'.	Magrane *et al.*, 2010	http://www.uniprot.org/
SwissProt-negative	5864	Non-redundant protein sets selected from SwissProt database from the set of potential non-nucleic acid binding proteins.	Magrane *et al.*, 2010	http://www.uniprot.org/
RBPDB	1101	RBPs obtained from RBPDB, a database of eukaryotic RNA-binding protein specificities.	Cook *et al.*, 2010	http://rbpdb.ccbr.utoronto.ca/
RNAcompete	205	RBPs obtained from the RNAcompete study conducted for the systematic analysis of RNA binding specificities in eukaryotes.	Ray *et al.*, 2009	http://hugheslab.ccbr.utoronto.ca/supplementary-data/RNAcompete
rbp86	86	RBPs compiled from the Protein Data Bank (PDB) using a maximum resolution of 3 Å and sequence identity less than 70%.	Cheng *et al.*, *2008*	http://doi.org/10.1186/1471-2105-9-S12-S6
rbp109	109	RBPs compiled from the Protein Data Bank using a maximum resolution of 3.5 Å and sequence identity less than 30%.	Cheng *et al.*, *2008*	http://doi.org/10.1186/1471-2105-9-S12-S6
rbp107	107	RBPs compiled from the Protein Data Bank using a maximum resolution of 3.5 Å and sequence identity less than 25%.	Cheng *et al.*, *2008*	http://doi.org/10.1186/1471-2105-9-S12-S6
RBRIdent	281	RBPs used for the development of an improved classifier named RBRIdent to identify the RNA-binding residues.	Xiong *et al.*, 2015	http://166.111.152.91/RBRIdent
nr-positive	687	RBPs selected from NCBI RefSeq, non-redundant databases using the keyword 'RNA-bind'.	Pruitt *et al.*, 2005	http://www.ncbi.nlm.nih.gov/
nr-negative	1199	Potential non-RNA-binding proteins selected from NCBI RefSeq, non-redundant databases using the keyword 'periplasmic'.	Pruitt *et al.*, 2005	http://www.ncbi.nlm.nih.gov/

In order to show the practical applications of APRICOT as a tool for large-scale analysis of data like complete proteome sets, two model organisms were evaluated. The *E. coli* K12 genome (taxonomy ID: 83333) was used as an example for bacterial species, while *Homo sapiens* (taxonomy ID: 9606) was used as an example for eukaryotic species. These genomes consist of 4,479 and 70,076 protein entries in the UniProt database, respectively. APRICOT was used to select positive RBP sets from both proteomes in order to quantify the accuracy with which the pipeline identifies RBPs in these relatively well-characterized genomes. I considered 1,535 non-redundant human proteins as a positive set, which have either been proposed as RBPs in the global experiment-based studies or have been reported in independent publications (Castello *et al.*, 2012, Baltz *et al.*, 2012; Kwon *et al.*, 2013; Gerstberger *et al.*, 2014).

In contrast, due to lack of such global studies in bacteria, beside ribosomal proteins, only a few proteins such as Hfq (Storz *et al.*, 2011), CsrA (Chao & Vogel, 2010), YhbY (Ostheimer *et al.*, 2002), SmpB (Wower *et al*., 2002), ProQ (Chaulk *et al.*, 2011), CspA (Phadtare & Inouye, 1999), and CspB (Phadtare & Inouye, 1999) have been reported as RBPs in *E. coli*. Hence, a larger RBP reference of *E. coli* K12 was retrieved from UniProt database using the GO term GO:0003723 for RNA-Binding, and was comprised of 160 proteins including the above-mentioned known RBPs.

2.3 Parameter optimization for domain predictions

2.3.1 Assessment criteria

The statistical parameters for domain predictions in the training set, as well as the performance of the tool on the test sets, were evaluated by using standard binary criteria of sensitivity (SN), specificity (SP), accuracy (ACC), Matthews Correlation Coefficient (MCC), and *F*-measure, using the following equations (where TP, FN, TN and FP are true positive, false negative, true negative and false positive respectively):

$$SN = \frac{TP}{TP + FN}$$

$$SP = \frac{TN}{TN + FP}$$

$$ACC = \frac{TP + TN}{TP + TN + FP + FN}$$

$$MCC = \frac{TP \times TN - FN \times FP}{\sqrt{(TP + FN) \times (TP + FP) \times (TN + FP) \times (TN + FN)}}$$

$$Fmeasure = \frac{2 \times SN \times SP}{SN + SP}$$

Receiver operating characteristic (ROC) curve and their area under the curve (AUC) was used as a criterion for accuracy, which was plotted using false positive rate (FPR or 1-SP) and true positive rate (TPR or SN).

2.3.2 Parameter optimization for the selection of predicted domains

The training sets SwissProt-positive (4,779 proteins) and SwissProt-negative (5,834 proteins) were analyzed with APRICOT in order to evaluate the ability of the pipeline to accurately differentiate RBPs from non-RBPs. For this evaluation, statistical parameters of sequence similarity, residue identity, residue gap, and E-value of the domain prediction was used to describe the similarity between a query and its corresponding reference. Unlike residue identity, sequence similarity accounts for edit operations such as positive substitutions, thereby capturing the secondary structure information at a better resolution. An E-value for searches of homologs against a database represents the likelihood that a given match in a sequence is purely by chance, meaning that a lower E-value reflects a more significant match. I describe an additional parameter namely the domain coverage, which is the percentage of the length predicted as domain in the query compared to the original length of reference domain. Generally, lower domain coverage suggests a random similarity of the predicted domain, whereas higher domain coverage reflects a higher potential of a domain to be functionally relevant.

Initially, the analysis of the training sets with a naïve approach was investigated, which involved InterProScan and CDD based batch-search methods with their default settings. Analysis by InterProScan achieved a TPR of 0.77 and CDD achieved a TPR of 0.79. Several proteins in the CDD-based method were annotated as RBDs with coverage lower than 10% and sequence similarity lower than 5%, which indicated poor conservation of the functional domains. Similarly, InterProScan failed to characterize several RBPs due to its stringent filtering criteria. Interestingly, several RBPs were reported by only one of the methods. Hence, when the results from both the analyses were combined, an increased TPR of 0.82 was achieved. This clearly shows the potential to achieve higher sensitivity by the combined approach that is implemented in APRICOT. The training data sets were further analyzed by

APRICOT, which predicted thousands of RBD entries in both the positive and the negative sets that were evaluated using systematically varying cut-offs of each parameter to optimize the identification of RBPs. The corresponding ROC curves were generated and optimal cut-off ranges were defined by identifying the values of the parameters that show an optimal TPR (closer to 1) and FPR (closer to 0) with high ACC (closer to 1), resulting into statistically significant AUC, MCC, and *F*-measures.

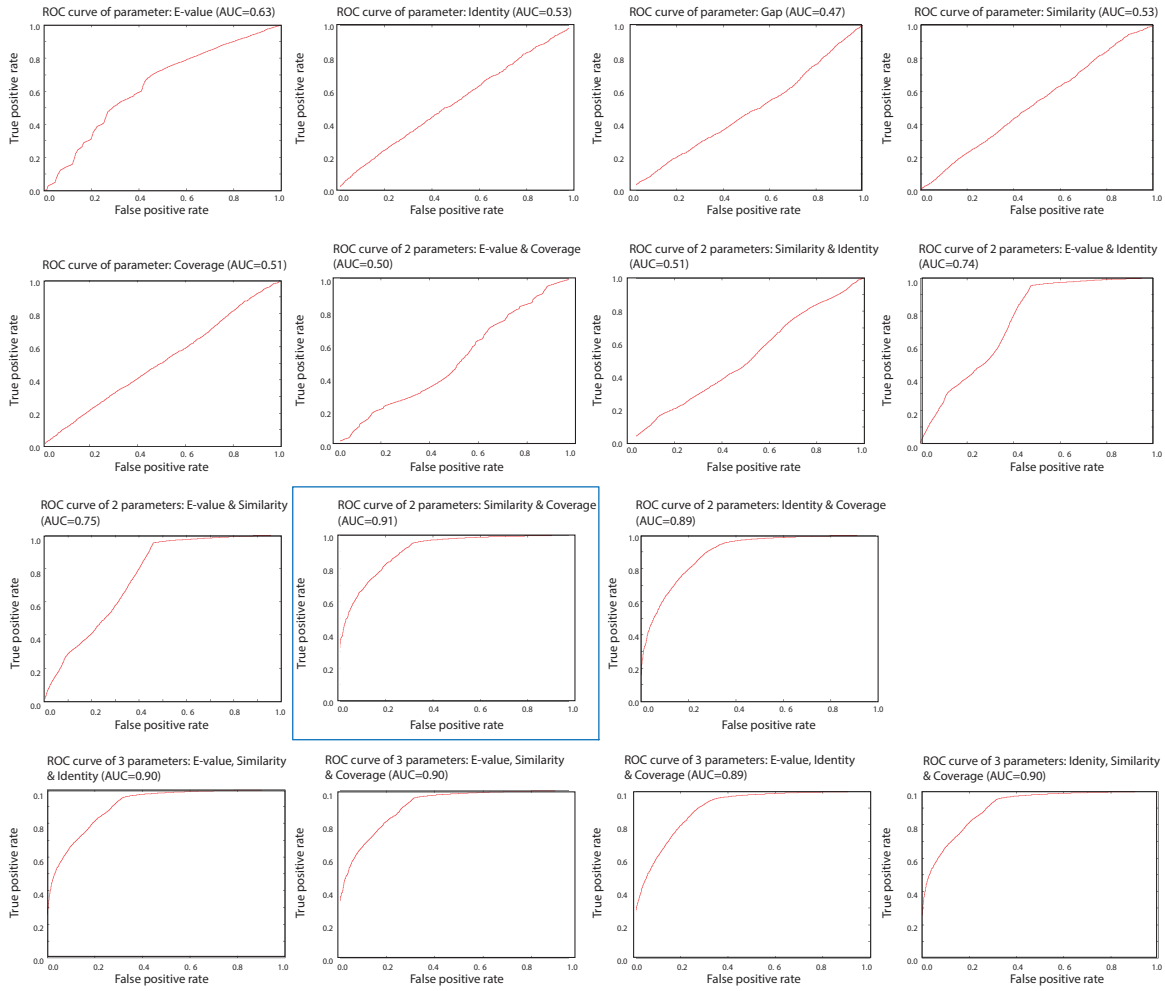


Figure 2.4 ***Assessment of the marginal contributions of all the domain prediction parameters to the overall accuracy of APRICOT with which it identifies RBPs.***

The Receiver Operating Characteristic (ROC) curves and their Area Under the Curves (AUC) for each parameter and combinations of parameters to assess their marginal contributions to the overall accuracy of APRICOT with which it identifies RBPs. The highest AUC value of 0.91 was achieved by the combination of three parameters: sequence similarity and domain coverage (highlighted in the third row). Using the combination of any three parameters (last row), the AUC value of 0.90 was achieved.

For the coverage of predicted domains, the minimum cut-off was recorded to be 39% that attained an accuracy, TPR, FPR, MCC value, and *F*-measure of 0.81, 0.87, 0.24, 0.63 and 0.81, respectively. Using a higher cut-off of 60%, a lower TPR 0.81, but a better FPR of 0.16, was obtained. This consequently shows a better ACC and *F*-measure. Similarly, the optimal threshold for the minimum cut-off of sequence similarity was recorded to be 24%, which attains accuracy, TPR, FPR, MCC value of and *F*-measure of 0.81, 0.83, 0.20, 0.63 and 0.81, respectively. Similarly, as shown in the ROC curve, by using a minimum cut-off of 15% for the residue identity and at a maximum E-value cut-off of 0.01, the high accuracies of 0.81 and 0.82 were achieved. The decision values of the parameters were further ranked, individually and in combinations, for all the predicted RBD entries in the training sets, and ROC curves and AUCs were generated to identify their marginal contributions on overall accuracy in detecting RBDs (**Figure 2.4**).

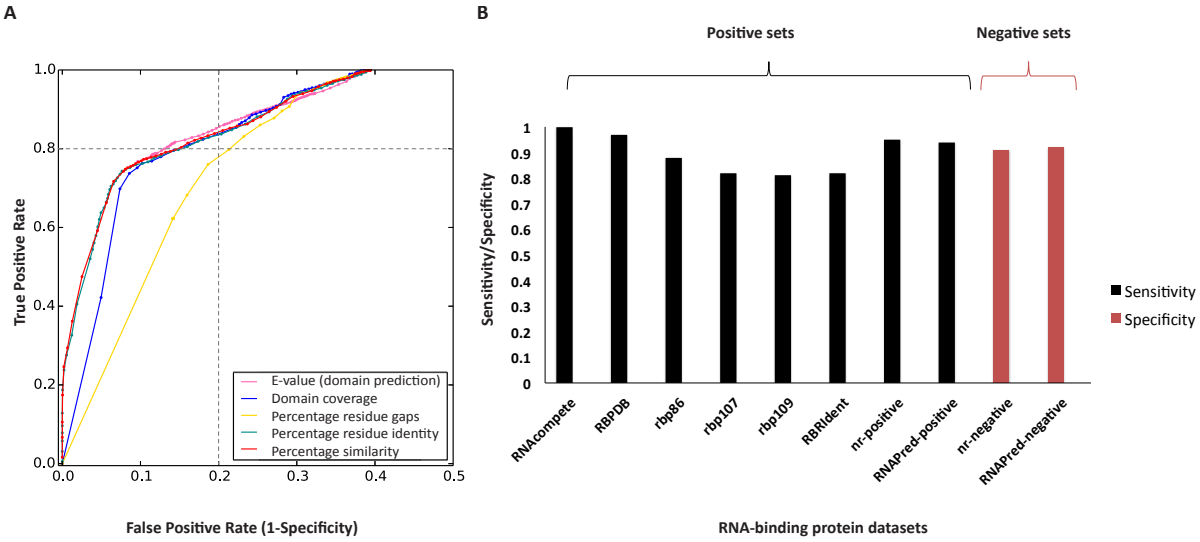


Figure 2.5 ***Selection of parameter cut-offs for RBP selection and the performance assessment of APRICOT on different data sets.***

*(**2.5A**) The ROC curves were generated for the domain prediction parameters of domain E-value (magenta), coverage (blue), residue gap (yellow), residue identity (green) and similarity (red). The optimal ranges for the parameters were defined for the selection of predicted domains at a considerably high accuracy (> 0.8 as indicated by the dashed lines) on the training sets (SwissProt-positive and SwissProt-negative). The minimum cut-off for most contributing parameters, percentage domain coverage and percentage similarity were recorded to be 39% and 24% respectively, which together attained an accuracy of 0.82. (**2.5B**) The bar chart illustrates the performance of APRICOT on different data sets by means of sensitivity (shown in black) and specificity (shown in red). APRICOT*

was evaluated on 8 positive data sets and 2 negative data sets, which showed an average sensitivity of 0.90 and an average specificity of 0.91.

This evaluation led to the selection of domain coverage and sequence similarity as the default parameters for the APRICOT analysis with their minimum cut-offs of 39% and 24%, respectively. The analysis by APRICOT using the selected parameters with their defined cut-offs achieves a TPR of 0.85, which is higher than the naïve approach. The MCC and *F*-measure achieved for the APRICOT analysis of the training sets are 0.64 and 0.82, respectively. This demonstrates the efficiency of the selected parameters and their cut-offs in identifying RBPs with a high accuracy of 0.82.

2.4 Assessment of pipeline performance for the identification of RBPs

A variety of positive data sets were analyzed by APRICOT, on which the pipeline achieved sensitivity in a range of 0.81 to 1 (**Figure 2.5B**), demonstrating its high efficiency in domain-based characterization of RBPs. A more detailed evaluation of the pipeline performance was carried out on the paired data set of nr-positive and nr-negative, and RNApred-positive and RNApred-negative (**Table 2.2**).

Table 2.2 ***Performance of APRICOT on positive and negative pair of data sets obtained from NCBI database and RNApred method.***

Data sets	**RNApred**	**NCBI (nr)**
Data set types		
Positive set	376 proteins	687 proteins
Negative set	355 proteins	1,199 proteins
Measures of performance assessment		
TP proteins	344 proteins	657 proteins
FP proteins	47 proteins	119 proteins
TPR (SN)	0.96	0.97
FPR (1-SP)	0.1	0.13
Accuracy	0.93	0.92
MCC	0.86	0.85
F measure	0.93	0.92

The complete proteomes of *H. sapiens* containing 70,076 UniProt protein entries was subjected to domain prediction. A known set of 1,540 non-redundant RBPs was used as a

positive reference set, of which 25 RBPs have not been defined with any globular domains. The reference domain set was considered for the initial identification of RBPs using pre-defined cut-offs for the aforementioned default parameters. Upon filtering of proteins by predicted domains, 1,091 from the reference RBP set were reported with at least one RBD from the reference domain set, showing a sensitivity of 0.71. By including the non-classical RBDs in the reference domain set, 68 more proteins could be recognized by APRICOT as RBPs. Moreover, 201 additional RBPs could be recognized by further including domains listed as *RBDs unknown*. The remaining 180 proteins that are not identified as RBPs by APRICOT do not contain RBDs and are listed as RNA-related proteins (Gerstberger *et al.*, 2014).

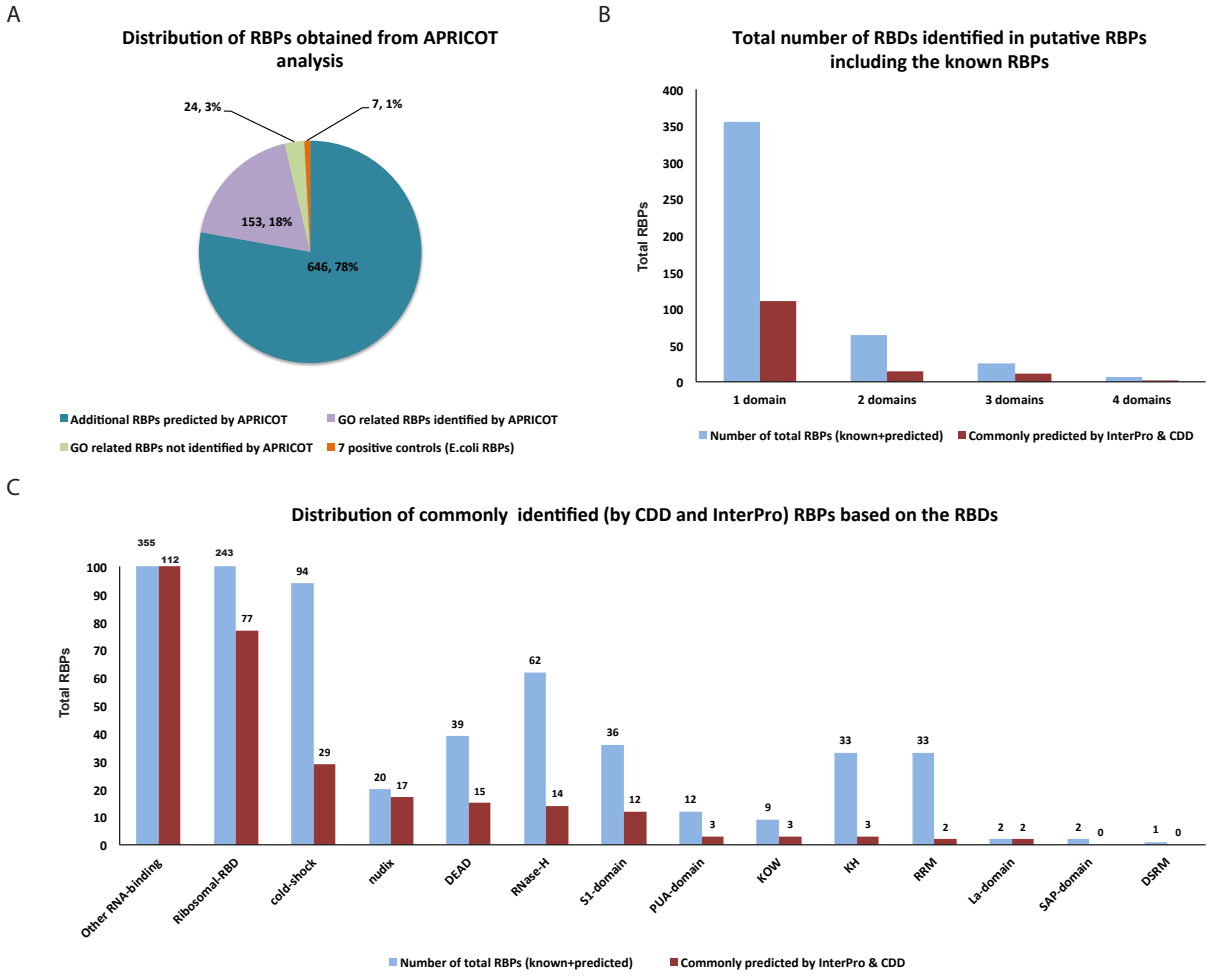


Figure 2.6 ***Analysis of the complete proteome of E. coli K-12 by APRICOT using the default parameters for the identification of RBPs.***

*(**2.6A**) The pie chart illustrating the distribution of proteins into positive RBP controls, and GO defined RBPs identified or not identified as such by APRICOT. (**2.6B**) Distribution of APRICOT identified RBPs based on the number of RBDs identified in its sequences by any one of the CDD or InterPro databases or commonly by both the databases. (**2.6C**) Distribution of RBPs that are identified as such by both the databases into different classes of RBPs.*

A similar analysis of the complete proteome of *E. coli* K-12 was carried out using APRICOT with default parameters and the reference domain set (**Figure 2.6**). In the initial characterization of RBPs, 673 sequences were selected as RBP candidates by RPS-BLAST and 502 sequences by InterProScan analysis. These proteins account for 806 RBP candidates, of which 369 proteins were identified as putative RBPs by both the methods. From the full proteome set, APRICOT could successfully identify the known RBPs: Hfq, CsrA, YhbY, SmpB, ProQ, CspA and CspB, likely due to highly conserved RBDs in their sequences that have been previously characterized for their regulatory roles (**Table 2.3**).

Table 2.3 ***List of known RBPs in Escherichia coli with their conserved RBDs and their corresponding regulatory roles.***

Protein name	Domains	Domain Coverage (%)	Residue Similarity (%)	Residue Identity (%)
Hfq	RRM_RBM7	41.33	25.33	17.33
CsrA	CsrA	84.06	72.46	49.28
ProQ	ProQ/FINO family	100	66.67	55.26
YhbY	RNA_bind_YhbY	97.89	89.47	80
SmpB	SsrA-binding domain	99.31	67.36	46.53
CspA	Cold shock domain	98.51	77.61	67.16
CspB	Cold shock domain	97.01	73.13	67.16

Furthermore, from the GO term-derived 160 RBPs from *E. coli* K-12, 129 were identified correctly by APRICOT, which demonstrated a sensitivity of 0.80. APRICOT failed to identify the remaining 24 proteins as RBPs because either the predicted RBDs did not pass the parameter filters or the reference domain set lack specific domains associated with these proteins. These unidentified RBPs included CRISPR system Cascade subunits, toxic proteins, and several enzymatic proteins like ribonucleases, tRNA-dihydrouridylases, and mRNA interferases.

The feature-based scores were calculated for each domain selected from the predicted data, which facilitates the differentiation of highly reliable RBD predictions from the low confidence RBD predictions. Query proteins that consist of high confidence RBDs were further annotated with additional information, namely subcellular localization, secondary structures, GO terms, and tertiary structures.

These proteome-wide analyzes clearly demonstrate a high sensitivity of the pipeline in identifying RBPs based on functional domains. However, it also shows a limitation: the characterization of the queries depends on the functional domains and motifs selected from the databases based on the user-provided terms.

2.5 Comparative assessment of RBP prediction tools

Several computational approaches are developed for the prediction of nucleic acid binding sites. Only four tools - namely SVMprot (Cai & Chou, 2005), RNApred (Kumar *et al.*, 2011), SPOT-Seq-RNA (Yang *et al.*, 2014) and catRAPID signature (Livi *et al.*, 2015) – have been originally described as RBP predictors. SVMprot was designed to predict RBPs by the SVM-based classification of protein primary sequences into functional families (54 Pfam families) and it was made available as a webserver. Since the tool is no longer available, it was not included in this analysis. RNApred uses SVM models that are developed with amino acid compositions and PSSMs and is available as a webserver. SPOT-Seq-RNA, which uses structural similarity based predictions of the RBPs. It also allows the identification of the binding residues and binding affinities using SPARKSX (Zhou & Zhou, 2005) and DRNA tools (Yang *et al.*, 2014), respectively, and is available as both the webserver and command-line tool. The fourth program, catRAPID signature, is an SVM-based method to identify RNA-binding proteins and their binding regions based on physico-chemical properties. An assessment of the APRICOT in comparison with the aforementioned tools was conducted to evaluate their methods and potential to predict RBPs correctly (**Table 2.4A**).

Unlike other tools that have been trained or constructed on a certain reference set, APRICOT is not established on any fixed set of references as it selects reference domains for each analysis based on the user provided keywords. Therefore, it is capable of using any new RNA-binding domains that might be added in the integrated domain sources in future. APRICOT selects proteins that are predicted with statistically significant RBDs and scores them in comparison with their reference consensus sequence for various features using Needleman-Wunsch alignment scores, Euclidean distance, and similarity-based scores. At the end of the analysis, the scores for each property are combined to obtain a Bayesian probabilistic score in a range of 0 to 1, where 1 indicates the best hits. The results from all the intermediate steps are provided to allow users to evaluate different statistical aspects of their study.

For an unbiased evaluation of the relative performances of APRICOT with *RNApred, SPOT-Seq-RNA*, and *catRAPID signature*, the two data sets *RBscore_R130* (130 RBPs) and *RBscore_R116* (116 RBPs) were used. These are the training and test sets created for the *RBscore_SVM* approach in *NBench*. On *RBscore_R130*, APRICOT achieved a TPR of 0.88, whereas *RNApred, SPOT-Seq-RNA*, and *catRAPID signature* attained much lower TPRs of 0.79, 0.82, and 0.55, respectively. On the *RBscore_R116*, which is indicated as a challenging set in *NBench*, APRICOT achieved a comparatively low TPR of 0.67. However, this was still higher than the TPRs achieved by *RNApred* (0.66), *SPOT-Seq-RNA* (0.51), and *catRAPID signature* (0.47). The performances of naïve RPS-BLAST were also checked, which is used for the batch-search of domain in CDD, and InterProScan, which is used for motif prediction in InterPro consortium. On both the data sets the naïve approaches for domain identification showed lower performances compared to their combined performance. Both the methods in their default setting achieved a TPR of 0.82 on the RBscore_R130 by identifying 107 RBPs. On the *RBscore_R116*, RPS-BLAST and InterProScan showed performances higher than SPOT-Seq-RNA but lower than APRICOT and *RNApred* by achieving TPR of 0.55 and 0.57, respectively.

APRICOT performed better than other tools in all the assessment metrics used for the evaluation of *RBscore_R246* (RBPs from both the data sets) as positive set and RNApred-negative (355 proteins) by achieving highest accuracy, MCC and *F*-measure of 0.88, 0.75 and 0.86 respectively (**Table 2.4B**).

Table 2.4 ***A comparative assessment of APRICOT and existing tools for RBP prediction.***

(2.4A) Important features have been evaluated as a measure of their predictive ability. (2.4B) A set of 246 RBP (RBscore_R246) and a negative set of 355 non-RBPs (RNApred-negative) were tested using different assessment metrics, where APRICOT achieved highest accuracy, MCC and F*-measure.*

RBP prediction tools	**APRICOT**	**catRAPID signature**	**SPOT-Seq-RNA**	**RNApred**
(2.4A) Main features of the tools				
Main criteria for RBP characterization	RNA-binding motifs and domain families	Physico-chemical properties	RBP structure homologs	SVM classification by composition features of proteins
Additional analysis	Sequence-based scoring of domain (includes physico-chemical properties)	Prediction of RNA-binding regions	RNA-binding residue prediction and binding affinity	PSSM-based evolutionary information
Availability	Command-line and	Webserver	Webserver and	Webserver

	Docker image		command-line	
Query types	Amino acid sequences / gene names / UniProt protein / taxonomy ids	Amino acid sequences	Amino acid sequence	Amino acid sequences
Allowed number of query proteins	Unlimited	100 proteins or total number of submitted characters = 100000	One query at a time	Unlimited for composition or one query at a time for the PSSM based analysis
Probability scores for RBPs	Bayesian score (0-1), 1 = best score	SVM score (Threshold -0.2)	Z-score	SVM score (Threshold -0.2)
Main criteria for RBP characterization	RNA-binding motifs and domain families	Physico-chemical properties	RBP structure homologs	SVM classification by composition features of proteins
(2.4B) Performance assessment				
TP (proteins)	193	125	166	180
FP (proteins)	44	150	6	102
TPR (SN)	0.79	0.51	0.67	0.73
FPR (1-SP)	0.12	0.42	0.02	0.29
ACC	0.83	0.54	0.83	0.72
MCC	0.66	0.1	0.69	0.44
F-measure	0.83	0.54	0.8	0.72

2.6 Prediction of RNA-binding sites

A comparative assessment of the programs developed for the prediction of nucleic acid binding sites was next carried out in NA Binding Prediction Benchmark (Miao & Westhof, 2015). A total of 16 tools for the prediction of RNA-binding residues, 5 tools for the prediction of DNA-binding residues, along with several data sets obtained from the structures of protein-nucleic acid complexes were included in this study (available at http://ahsoka.u-strasbg.fr/nbench/index.html). APRICOT identifies RBPs among large-scale query sets and further characterizes them by biological functions, whereas the 16 tools in *NBench* predict RNA-binding residues in the pre-defined RBPs. Hence, the motivation behind developing APRICOT is fundamentally different from the tools involved in *NBench*. However, APRICOT and these tools can complement each other by first identifying RBPs and their corresponding RBDs with APRICOT and further obtaining a high-resolution annotation by

identifying RNA-binding residues using the best performing tools from *NBench.* To evaluate the potential of this idea, 3,657 PDB entries were acquired, consisting of 24 different RNA-related data sets in *NBench*, selected at a resolution cut-off of 3.5 Å. This data set was subjected to analysis by APRICOT and a comparative assessment was carried out between the identified RBD sites and the nucleic acid binding residues at the distance cut-off of 3.5 Å in each PDB entry (**Figure 2.7**).

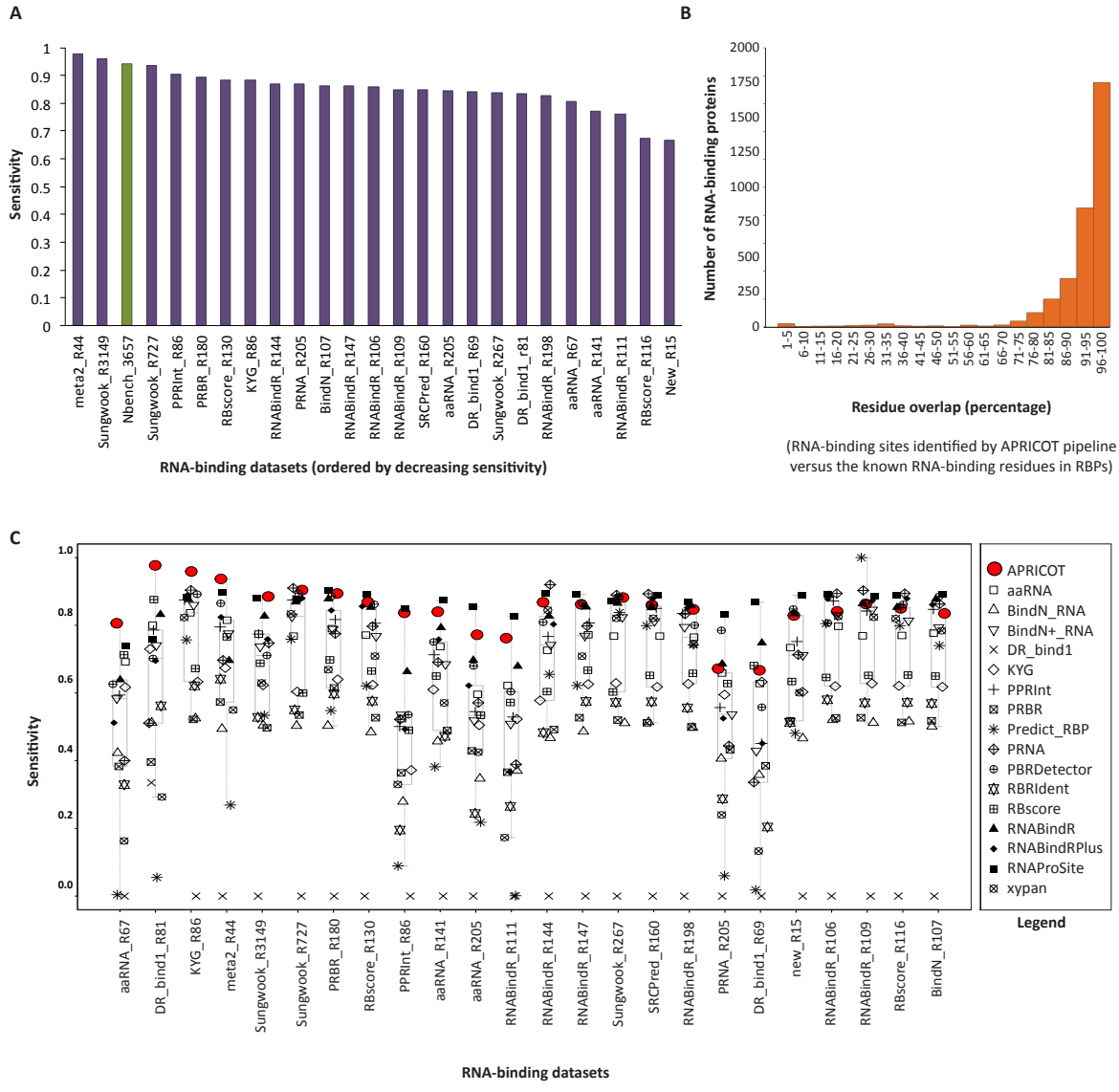


Figure 2.7 ***A comparative assessment between the identified RBD sites by APRICOT and the nucleic acid binding residues identified the tools discussed in NBench.***

*(**2.7A**) The bar chart showing the specificities achieved by APRICOT on different data sets, including the entire set of 3,657 RBPs (NBench_3657 shown in green). (**2.7B**) Distribution of RNA-binding proteins based on the percentage of overlapping RNA-binding residues defined in NBench with the RNA-binding sites identified by APRICOT. The RNA-binding sites were identified in 3,445 of NBench_3657, of which 3,304 proteins have more than 70% of their RNA-binding residues*

*overlapping with the RNA-binding sites. (**2.7C**) Boxplots showing the sensitivities achieved by APRICOT in identifying RNA-binding sites (in red) and other RNA-binding residue prediction tools in identifying RNA-binding residues (in black) on NBench data sets. On all the data sets, APRICOT achieved sensitivities higher than or as good as high performing tools.*

The RNA-binding residues of 3,340 (91%) PDB entries overlap with the APRICOT predicted RBD sites was observed that showing an overall sensitivity of 0.91 (**Figure 2.7A, Figure 2.7B**). The *NBench* tools were ranked by their sensitivities to identify RNA-binding residues together with APRICOT for its ability to identify RNA-binding sites on 24 data sets. As shown in **Figure 2.7C**, APRICOT was among the best performing tools compared to the other tools in *NBench* across the 21 diverse data sets. In agreement with the observations made for the tools in *NBench*, APRICOT showed the lower sensitivity on the New_R15 set (15 new structures) and RBscore_R116 (116 proteins, mentioned as a difficult set). Furthermore, unlike most of the tools that do not show discriminative potential for RNA and DNA binding residues, APRICOT showed a high specificity (0.7) when 1,374 DNA binding proteins were included in this analysis. This evaluation demonstrates that APRICOT's domain prediction-based analysis is an extremely efficient approach to identify RBPs and their corresponding potential RNA-binding region in the query sequences. Furthermore, it also implies that the resolution of the RBP studies could be enhanced significantly by first identifying the RBPs using APRICOT, followed by the analysis with the tools for the identification of RNA-binding residues in the predicted RBD sites.

2.7 Identification of other functional classes by APRICOT

APRICOT modules are applicable for the functional identification of not only RBPs, but they can be easily adapted for one or multiple other functional classes. As a part of the Critical Assessment of Function Annotation (CAFA), a project to assess the methods for computational annotation of protein functions (Radivojac *et al.*, 2013), APRICOT was successfully used to annotate a bacterial data set comprising of more than 1 million proteins by a wide number of biological functions (Jiang *et al.*, 2016). In order to emphasize the aspect of APRICOT as a tool for the characterization of other functional classes of proteins, kinase proteins from *E. coli* strain K-12 as the reference set were chosen. Kinases catalyse the transfer of phosphate groups to a substrate molecule using ATP as a phosphate donor. In the UniProt database, 110 proteins from *E. coli* K-12 are annotated with various kinase

activities (for example, Serine/threonine-protein kinase, Signal histidine kinase, Shikimate kinase etc.) and are tagged by the GO term (GO:0016301) for kinase activity.

The APRICOT pipeline was supplied with the term 'kinase' for the selection of reference domain set and the pipeline was applied to the kinase proteins. Out of 110, 106 kinase proteins were identified correctly by APRICOT, achieving a sensitivity of 0.96. The set of proteins that was discarded by APRICOT contain kinase-associated domains that were not present in the reference domain set due to the domain selection constraints of APRICOT. This analysis suggests that APRICOT is also efficient in the characterization of proteins based on pre-defined sets of domains associated with functional classes other than RBPs. However, it should be noted that the accuracy of the results depends on the choice of terms for the domain selection.

2.8 Concluding remarks

APRICOT is an integrated pipeline for the sequence-based identification and annotation of the query proteins based on the functional motifs and domains of interest known from the experimental data. Notably, here I report APRICOT primarily as a tool for the sequence-based identification of RBPs, which uses a consistent set of reference RBDs from CDD and InterPro domain databases. Technically, the PSSM-based approach of CDD is built upon ungapped motifs, whereas the HMM probabilistic models of InterPro can handle motifs with insertions and deletions. By combining the predictive abilities of the CDD and InterPro consortia, APRICOT provides a broader scope for domain characterization.

By applying the pipeline to a variety of test sets, it was also ensured that the efficiency of APRICOT does not depend on the underlying data sets, unlike the tools that are trained and tested on a small and often curated data set. In addition, APRICOT has been extensively trained and optimized for the identification of diverse RBP sets from all domains of life. Hence, it can be used to extend our knowledge of RBPs in systems other than eukaryotes. One of such example is presented in the next chapter, where the tool is applied for the identification and characterization of RBPs in bacterial pathogen *Salmonella* Typhimurium.

Other existing tools for RBP identification can process very few queries at a time (most often only one) via their webserver. Therefore, the capacity of APRICOT to deal with a large-scale data set is one of its important features, which allows it to process data sets as large as

the complete human proteome. For instance, APRICOT could successfully identify RBPs like CsrA, ProQ, YhbY, and SmpB in *E.coli* with their respective RBD motifs with domain coverage of higher than 80% and residue similarity close to 70%. In addition, APRICOT suggested a number of proteins in *E.coli* that can potentially interact with RNAs via RBDs and hence, could be experimentally studied.

Due to the automated framework and accessibility of different modules of the pipeline, APRICOT can be conveniently adapted for the characterization of other functional classes. Additionally, by applying the tool for the identification of the kinases in *E. coli*, I demonstrated that the tool is not built on a fixed set of domain information; instead it allows users to characterize proteins based on the functional classes of their interest.

Chapter 3

High-throughput screening of putative RBPs in *Salmonella* Typhimurium

Bacterial RBPs have been shown to bind to specific sRNAs or mRNAs for the post-transcriptional regulation of gene expression, and have been studied intensively in enterobacteria such as *Salmonella.* The gram-negative flagellated bacterial species *Salmonella enterica* subsp. *enterica* is of an immense interest of scientific community due to its intracellular eukaryotic pathogenicity. A relative small number of proteins have been reported as RBP in *Salmonella* and other bacterial proteomes (**Chapter 1**). However, the total number of known human RBPs (>1500) corresponding to >5% of the entire proteome, indicates that a much higher number of proteins in other proteomes, including those of bacteria, could potentially have RNA-binding abilities.

Using the APRICOT software (described in the previous chapter), a primary screening of RBPs was carried out in the proteome, which was followed by an RNA-Sequencing and Immunoprecipitation-based experimental validation of several of the potential RBP candidates carried out in the lab of Prof. Jörg Vogel. Various technical methods were implemented for the validation of these RBP candidates and for the identification of the genes enriched. Furthermore, other publicly available data sets of dual RNA-Seq (Westermann *et al.*, 2016), *Salmonella* Compendium (Kröger *et al.*, 2013), and transposon-directed insertion site sequencing (TraDIS) (Langridge et al., 2009; Chaudhuri *et al*., 2013; van Opijnen & Camilli, 2014), were included in this study in order to derive the biological functionality of the proteins in terms of their relevance as regulatory components.

Due to the two main functional aspects (computational and wet lab experiments) of this study, this chapter can be divided into two sections. The first part includes the methods for RBP identification and their selection for the CoIP-based experimental study and sequencing in the Chapters 3.1-3.3. The second part discusses the high-throughput sequencing and its downstream analysis in the Chapters 3.4-3.8.

3.1 Identification of RBPs in *Salmonella* Typhimurium SL1344

Using the APRICOT software, a comprehensive set of reference-RBDs was selected from the domain databases and the proteome of *Salmonella* Typhimurium SL1344 was subjected to the analysis in order to identify proteins with the reference domains. RBP candidates that were identified with one of the functional domains of interest were scored by means of sequence-based features, which include information like chemical properties, amino acid composition, structural properties, alignment scores, and measures of similarity with respect to the reference domains. The feature-based scoring facilitated the ranking of the candidate proteins and allowed the selection of high-confidence candidates that possess highly conserved RBP motifs. Additionally, APRICOT annotated the selected putative RBPs by information such as secondary structure, subcellular localization, and Gene Ontology. The results obtained from these analyses in SL1344 are discussed below in detail.

3.1.1 Selection of RNA-binding domains

In addition to the various bacterial RBDs reported in literature (**Table 3.2**), a reference for RNA-binding domains was retrieved using a range of terms comprising classical and non-classical domains (Castello *et al.*, 2012), which are characterized as such due to their presence in the well-characterized mRNA binding proteins (mRNPs). In addition, to account for other well-defined non-mRNPs and ribosome-related RBDs, the term 'RNA-bind' and terms associated with the RNA-binding ribosomal domains (Gerstberger *et al.*, 2014) were used.

Table 3.1 ***The classification of prokaryotic RBDs based on the literature.***

RBDs in bacteria	References
PIWI	Kumar *et al.*, 2012
KH	Matus-Ortega *et al.*, 2007
S1 (studied with KH domain)	Wong *et al.*, 2013
SAM (homolog *rimO* in bacteria)	Fontecava *et al.*, 2004
CRM	Ostheimer *et al.,* 2002 & 2003
ANTAR	Shu & Zhulin, 2002
PNPase: Polyribonucleotide nucleotidyltransferase	Symmons *et al.*, 2002
CAT (Co-AntiTerminator RNA-binding domain)	Declerck *et al.*, 1999
TRAP	Yakhnin *et al.*, 2006
Cold shock domains	Sachs *et al.*, 2012
DSRM (found in *E. coli* RNases III, H1)	Mian *et al.*, 1997
PUA (bacterial RsmF)	Demirici *et al.*, 2010

S4 (EF-Tu, EF-G) (e.g. RpsD)	Björkman *et al.*, 1999
RRM	Maryuyama *et al.*, 1999
ZnF-CCCH	Deng *et al.*, 2012
TROVE (can co-occur with WD40)	Bateman & Kickhoefer, 2003
THUMP (e.g. thiI-like 4-thiouridine synthases)	Waterman *et al.*, 2006
SRP54 (P48 and FtsY)	Montoya *et al.*, 1997
CheY (bacterial two-component signalling systems) (e.g. AmiR)	Galperin, 2006
Pseudouridine synthetase (TruB, RsuA, TruD, Pus4)	Koonin, 1996
LSM (Hfq)	Lease & Woodson, 2004
CRISPR-associated protein Cse3	Makarova *et al.*, 2006
Aconitase B- Eukaryotic mAcn (IRE)	Walden *et al.*, 2006
OST-HTH (DUF88)	Anantharaman *et al.*, 2010

APRICOT retrieved 655 domain entries from CDD and 593 diverse entries from the InterPro database using 16 domain identifiers that are well-defined in the databases or listed as classical RBDs, which are: RRM, Nudix, DEAD, KH, RNase H, S1-domain, cold-shock domain, La-domain, PIWI, pumilio, DSRM, zf-CCCH, SAP-domain, and PUF-domain. In addition, 5,816 additional domains were included in the reference sets, which account for 1,952 non-classical RBDs, 1,715 RNA-binding ribosomal domains, and 2,149 other non-mRNP related RBDs. In addition to the important classical and non-classical RBDs (**Figure 2.2A-2.2B**), the detailed statistics of domain entries selected by the specific terms denoting non-classical RBDs and RBD-unknown have been shown in the **Figure 3.1A** and **Figure 3.1B,** respectively.

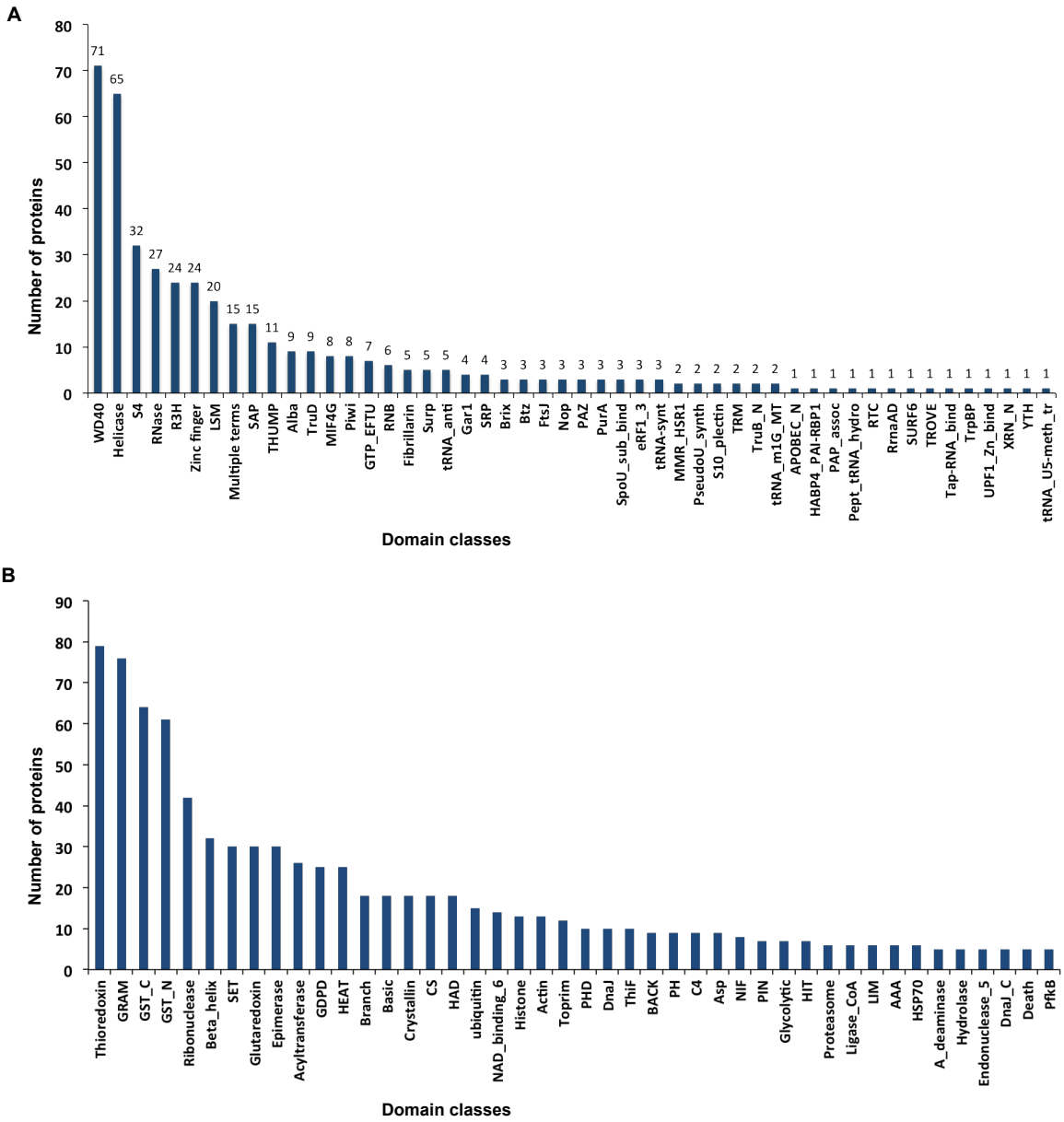


Figure 3.1 ***The numbers of domain entries selected by specific terms indicating different classes of domains.***

*Domain entries from CDD and InterPro database that were retrieved using the terms associated with the selected classes of non-classical RBDs (**3.1A**) and RBD-unknown (**3.1B**).*

3.1.2 Identification of RNA-binding proteins

APRICOT has been trained on a large set of positive and negative RBPs to rank the parameters and their respective cut-offs for the identification of RBPs that contain classical RBDs with an optimal accuracy (~0.8). The training sets indicated that two parameters contribute the most in the identification of RBPs, which are domain coverage and sequence similarity. The domain coverage denotes the fraction of reference consensus identified in the query proteins and the sequence similarity indicates the extent of similarity between the reference and identified domain region for which the minimum cut-offs were set as 39% and

24% respectively. It should be noted that the similarity threshold used by APRICOT does not necessarily takes the structural conformation that plays an important role in RNA binding into account.

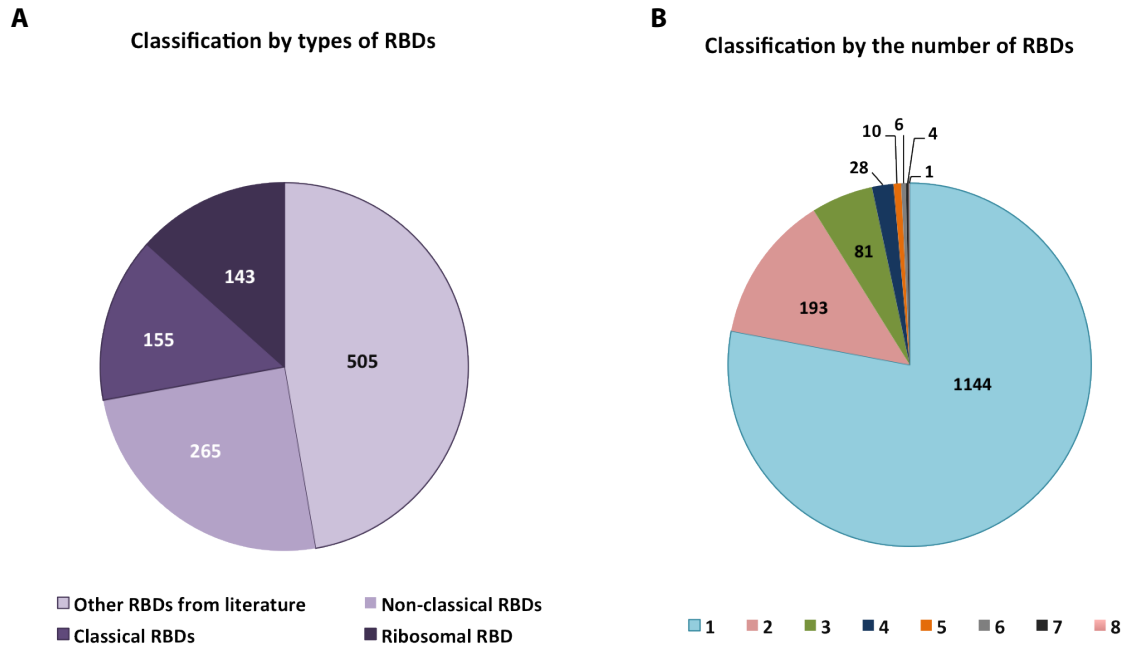


Figure 3.2 ***The classification of 1068 proteins that were computationally identified as RBPs by APRICOT in Salmonella Typhimurium SL1344.***

(3.2A) *Pie chart illustrating the distribution of putative RBPs across different RBD types. 155 proteins were identified as RBPs based on the occurrences of classical RBDs in their sequences.* ***(3.2B)*** *Pie chart showing the number of RBDs predicted in the putative RBPs. The majority of them were predicted to harbour only 1 RBD, whereas 323 proteins were identified with 2 or more RBDs.*

APRICOT identified 1,068 proteins as putative RBPs, which constitutes about 20% of the entire proteome of SL1344, using the pre-defined parameters and their cut-offs (**Figure 3.2**). Among these putative RBPs, 155 proteins were predicted to have classical RBDs, 265 proteins were predicted to have non-classical RBDs, 143 proteins were predicted to have RNA-binding ribosomal domains, 96 proteins are annotated as transcription factors, and the remaining 500 proteins were predicted to have potentially non-mRNP-related RBDs. From the set of known bacterial RBPs identified in the literature (Van Assche *et al.*, 2015), in *Salmonella* Typhimurium the few proteins that have been characterized as such are Hfq, CsrA, ProQ, SmpB, CspA, CspB, and YhbJ (RapZ). These proteins were successfully identified with high confidence as RBPs by APRICOT with their respective RBDs (**Table 3.3**).

Table 3.2 ***Positive RBPs used as control libraries in the RIP-Seq based screening of RBPs in*** **Salmonella** ***Typhimurium.***

These proteins were successfully identified as RBPs by APRICOT, and the identified RBDs and their corresponding statistics are indicated.

Gene name	Synonym/ Locus_tag	Domain id	Short name	Domain length	E value	Coverage percent	Identity percent	Similarity percent
yhbJ	SL1344_3295	PF03668	RapZ-like family	284	1.00E-175	98.94	64.43	75
smpB	SL1344_2660	PF01668, TIGR00086	SsrA-binding protein	68	7.00E-36	95.58	58.82	73.52
proQ	SL1344_1775	PF04352, SM00945	ProQ, ProQ/FinO domain	114	2.00E-46	100.00	55.26	67.54
csrA	SL1344_2806	PF02599	CsrA	54	3.00E-27	98.14	70.37	90.74
cspB	*cspJ cspG* SL1344_1924	PRK09890	Cold-shock protein	70	4.00E-34	98.57	80.00	91.42
cspA	SL1344_3615	PRK09890, PF00313	Cold-shock protein, DNA-binding	70	1.00E-34	98.48	74.24	84
hfq	SL1344_4295	PRK00395, PF01423	Hfq, LSM	79, 66	4.00E-07	94.93, 81.81	77.21, 25.75	87.34, 48.48

To infer functionalities and possible regulatory roles, the putative RBPs were further classified based on the type of their identified RBDs, their co-occurrences with other RBDs or non-RBDs, and known functions (**Figure 3.3**). Among the classical RBDs, cold-shock domains were predicted in 87 proteins, which was the domain class with the maximum number of putative RBPs. Other domain classes that accounted for an average of 36 proteins are KH, DEAD, and S1-domains. Meanwhile, a KOW and La-domain were identified in 9 and 8 proteins, respectively, and DSRM and PIWI domains were identified in only one protein each. RNA-binding ribosomal protein domains were identified in 255 proteins and non-mRNP-related RBDs were identified in 479 proteins. Among the non-classical RBDs, various SAM domains, S4-domain, Helicase_C, and tRNA-anti were identified in a several proteins. However, most of the non-classical domains were identified in one putative RBP only, which is similar to the distribution of these RBDs reported in the human candidate RBP complement (Castello *et al.*, 2012; Gerstberger *et al.*, 2014). Among the proteins identified with RNA-binding ribosomal domains, 68 were originally annotated as ribosomal proteins in the databases, 53 of them were annotated as hypothetical or uncharacterized proteins, whereas the rest accounted for proteins with other functions.

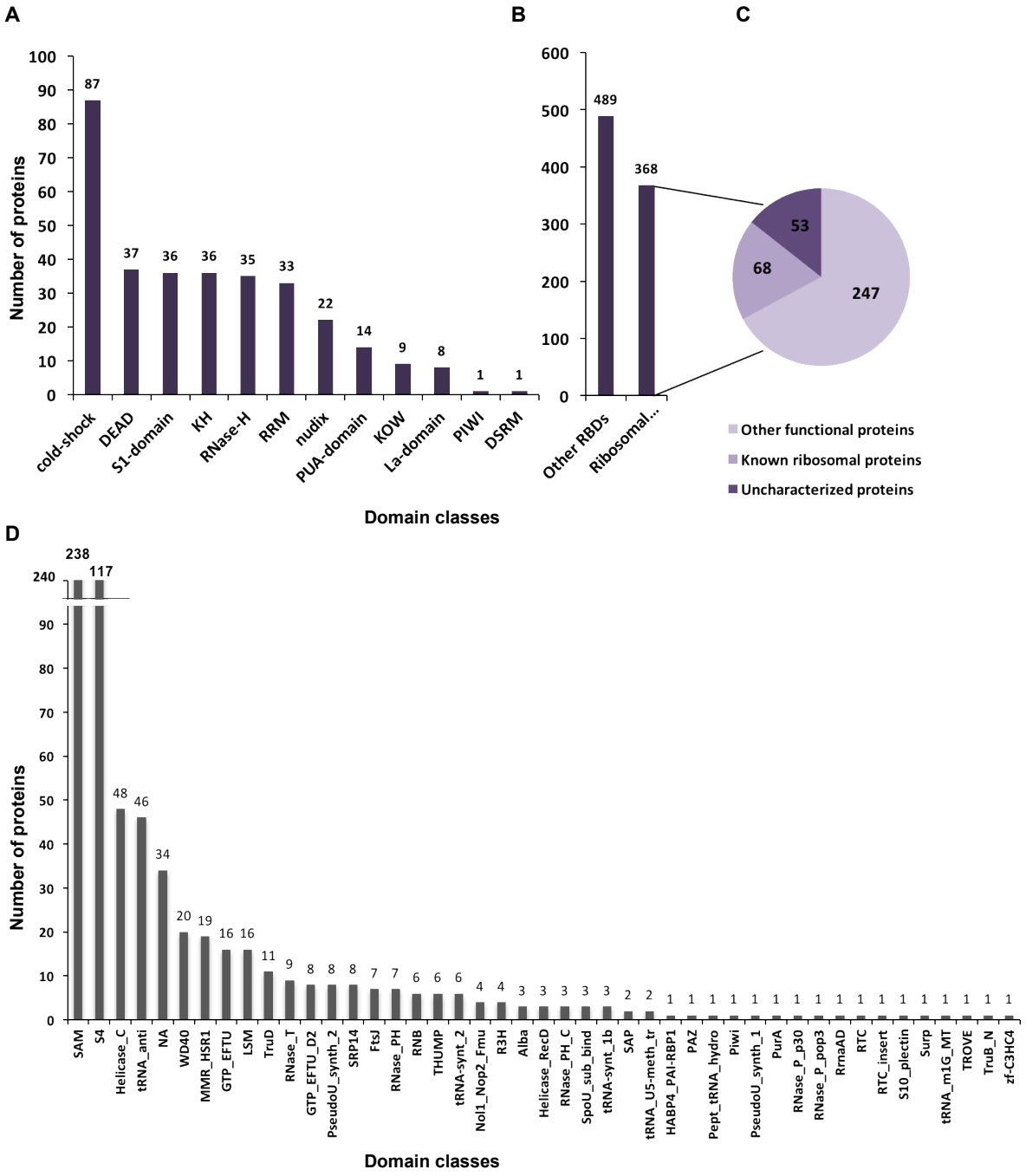


Figure 3.3 ***The classification of predicted RBPs by the domain types.***

*Bar charts showing the classification of the proteins into classical domains (**3.3A**), non-classical domains (**3.3B**) and other RBDs: domains that are annotated as RNA-binding based on literature (further separated in **3.3D**). (**3.3C**) Pie chart illustrating the classification ribosomal domains containing proteins into the groups based on their known functions in the UniProt database.*

3.2 Transcript abundance of computationally-identified RBP-encoding genes in transcriptomic data

The human RBPs, in general, are expressed at a higher level compared to the rest of the proteome and have been reported to contribute significantly to the total expressed protein-coding transcriptome (Gerstberger *et al.*, 2014). This pattern of expression has been linked to important roles of RBPs in RNA processing and post-translational gene regulation. In order to determine if such expression patterns are also present for the putative RBPs identified by APRICOT in SL1344, the expression levels of putative RBPs in two transcriptomic data sets was measured. The first data set was taken from SalCom, which comprises RNA-Seq data generated under infection-relevant conditions (Kröger *et al.*, 2013). The second data set was obtained from the dual RNA-Seq based study of *Salmonella*-infected HeLa cells (Westermann *et al.*, 2016), which quantifies the expression of transcripts from both host and pathogen.

In addition to the RBPs, the expression patterns of transcription factors (TFs) were also retrieved. The total number of genes encoding RBPs and TFs are almost similar in the human genome (~1,500). In contrast, in SL1344, only 216 proteins (25% of the putative RBP-encoding genes) are annotated as TFs.

In human, TFs represent 3% of the transcript abundance in data sets obtained from different tissues and cancers, whereas RBP-encoding genes contribute up to 20% of the transcript abundance (Gerstberger *et al.*, 2014). To determine the existence of such expression pattern in SL1344, the relative abundance of TFs and RBPs in the aforementioned SL1344 transcriptomic data sets was calculated.

3.2.1 SalCom data sets

SalCom is a transcriptomic compendium for *Salmonella enterica* serovar Typhimurium comprised of RNA-Seq samples generated under 20 infection-related conditions (Kröger *et al.*, 2013). A total 3,790 genes (85%) out of 4,456 coding genes were recorded as expressed (read count > 5) in at least one environmental condition. SalCom included expression profiles for 1,040 out of 1,068 APRICOT-selected candidate RBP-encoding genes, and 213 out of 216 transcription factors. A total of 96 TF encoding genes which were annotated as such by Go terms, were identified also as genes encoding RBPs based on APRICOT analysis. These genes were considered only as TFs for the purpose of quantification; hence, I used 944 putative RBPs in the subsequent analysis. In the SalCom data set, 841 out of 944 RBPs (89%) and 204

out of 213 TFs (96%) were sufficiently expressed in at least one of the conditions (**Figure 3.4**).

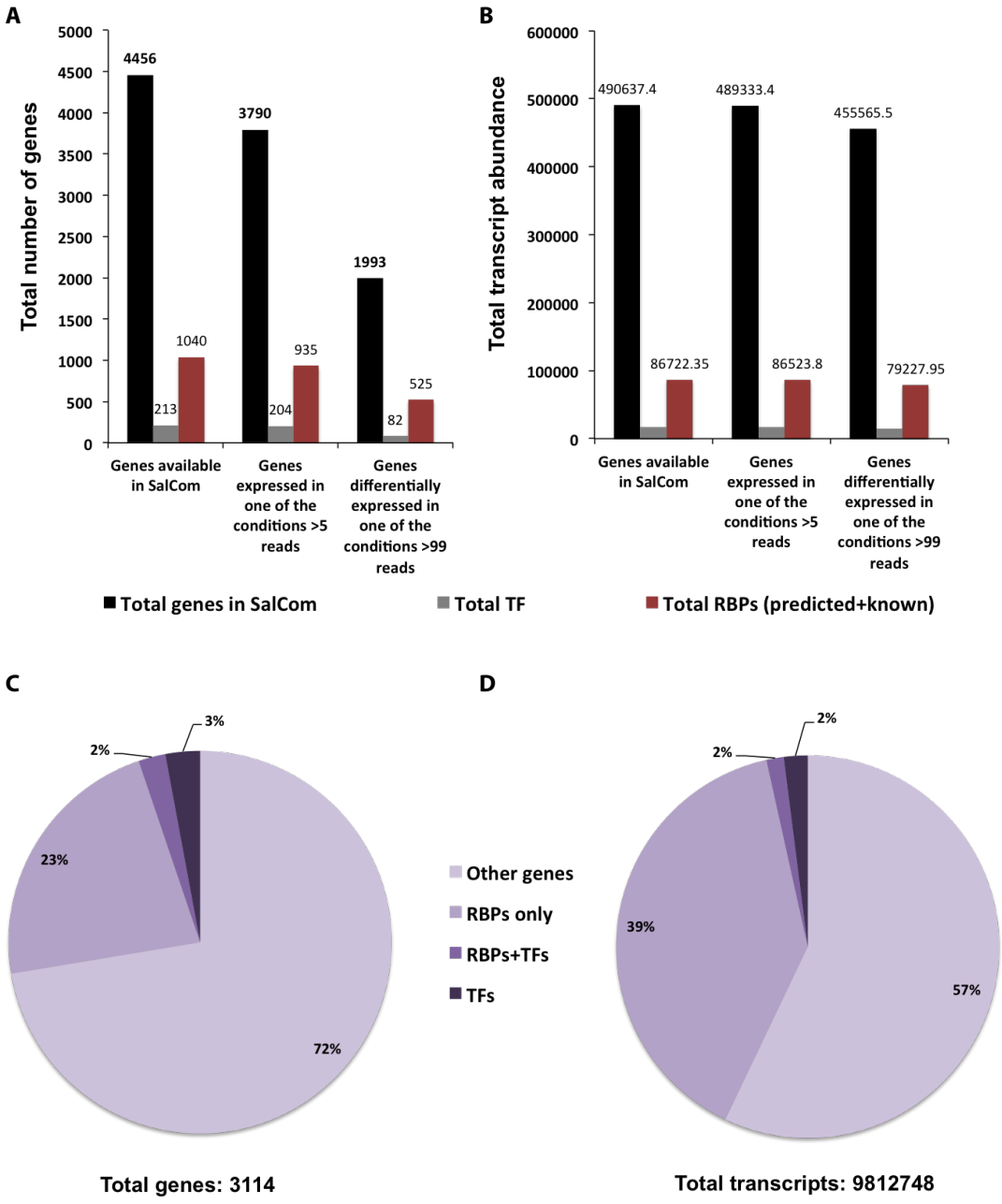


Figure 3.4 ***Annotation of the putative RBP encoding genes and their comparison with the transcription factors (TFs) using the transcriptomic data sets in SalCom.***

(3.4A) *Numbers RBP encoding genes and the TFs in the complete gene set in SalCom.* ***(3.4B)*** *The relative transcript abundance of the RBP encoding genes and the TFs in the complete SalCom transcriptomic data sets.* ***(3.4C)*** *Distribution of the 3114 differentially expressed genes (expression cut off = >5 reads) into different classes of putative RBPs, TFs, RBPs+TFs (when the TFs are also predicted as RBPs) and other genes.* ***(3.4D)*** *Relative transcript abundance of the differentially expressed genes in the different gene classes (as described in the* ***Figure 3.4C****).*

Similar to the transcript abundance of TFs in human transcriptome data sets, the TF-encoding genes of SL1344 contribute an average of 3% to the transcriptome in the SalCom data set, while a high transcript abundance of 41% was recorded for all the putative RBPs of which ~50% of transcripts correspond to the RNA-binding ribosomal proteins. The highly-expressed set of genes that are mapped with read counts of more than 100 in at least one RNA-Seq data set in SalCom consisted of 1,993 genes (45% of the total) that include 488 candidate RBP-encoding genes (51% of total RBPs) and 82 TFs (38% of the total TFs), representing 41% and 3% of the total transcriptome, respectively (**Figure 3.5)**.

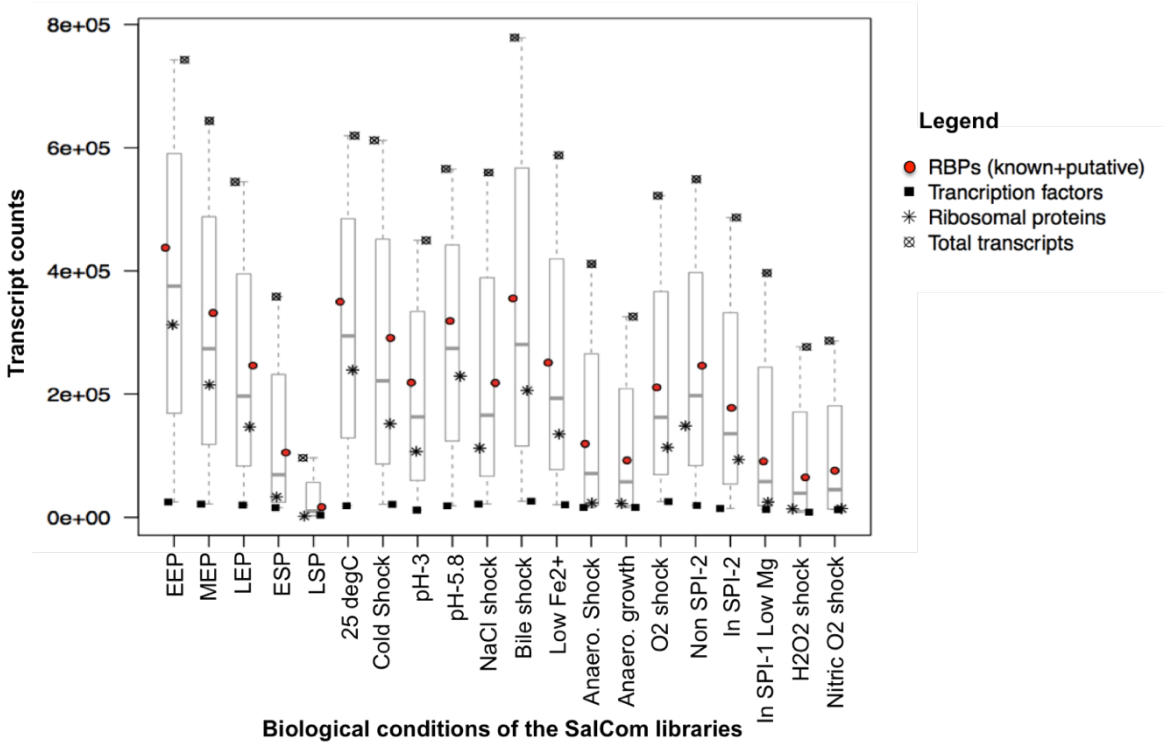


Figure 3.5 ***Total transcript contributions of the highly expressed candidate RBP encoding genes and TFs to the complete transcriptome data set of Salcom.***

A threshold of the mapped read count >100 was used for the selection of the highly-expressed genes. As shown in the boxplots, the RBP encoding genes account for an average of 40% of the entire transcriptome, half of which correspond to the ribosomal RBP encoding genes. In contrast <3% transcripts correspond to the TFs in different conditions.

In the set of putative RBP-encoding genes, 68 genes are annotated as ribosomal proteins and represent 24% of the total transcriptome. These RNA-binding ribosomal proteins account for only 6% of the total putative RBP genes but they represent 55% of the transcripts corresponding to RBP-encoding genes. In summary, this evaluation indicates that in the complete transcriptome, 18.5% of transcripts represent non-ribosomal RBPs, 22.5%

transcripts represent RNA-binding ribosomal proteins, and 3% of transcripts account for TFs (**Table 3.4**). These proteins were further inspected in a different dataset, which is discussed in detail later in this chapter.

Figure 3.6 ***Gene expression profiles of genes in SalCom data set.***

(Images below)

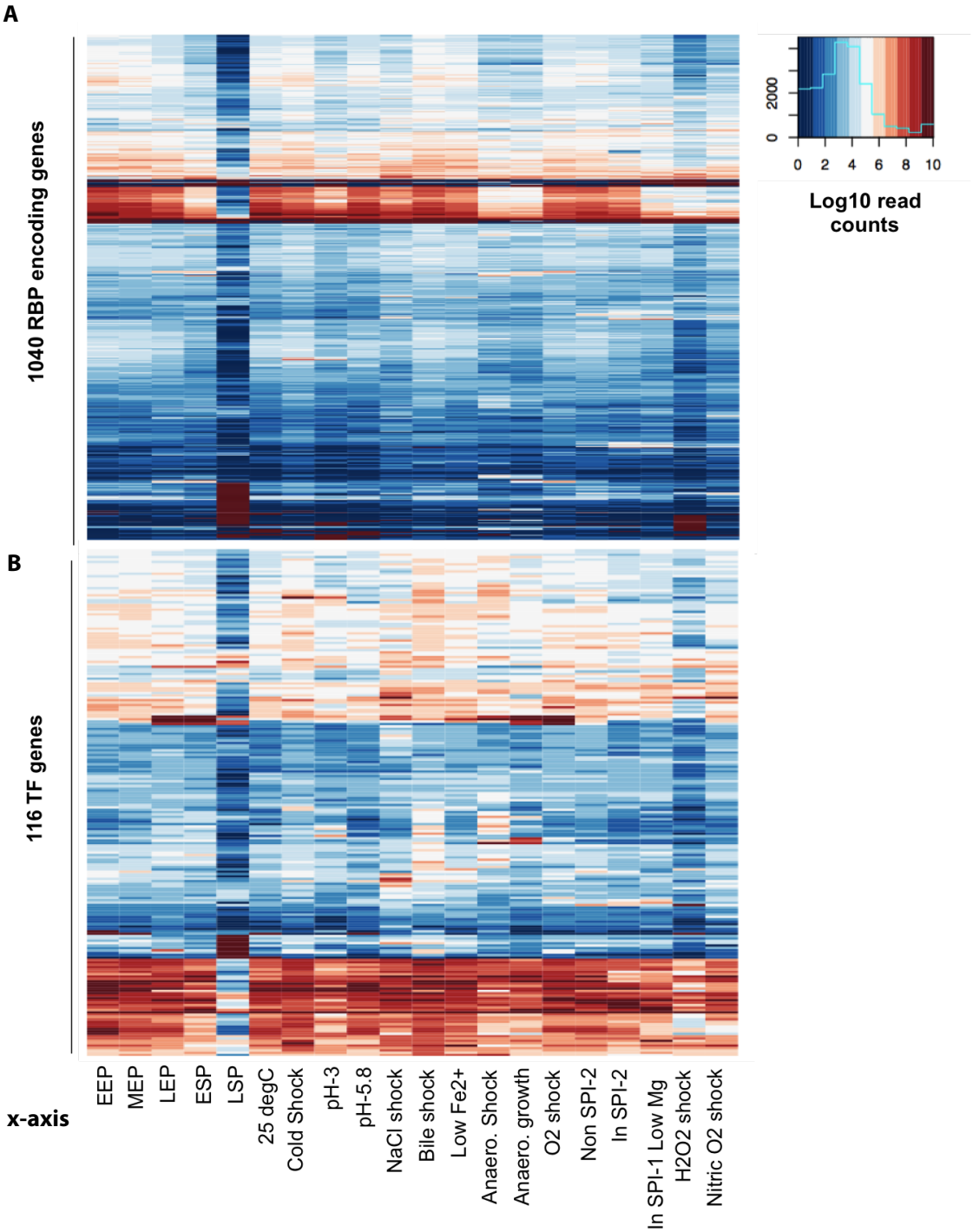


Figure 3.6A-3.6B *Heatmaps in 3.6A and 3.6B illustrate the expression profiles in terms of the transcript abundance of the candidate RBP encoding genes (3.6A) and TFs (3.6A) in the y-axis, where the x-axis corresponds to the biological conditions of the samples.*

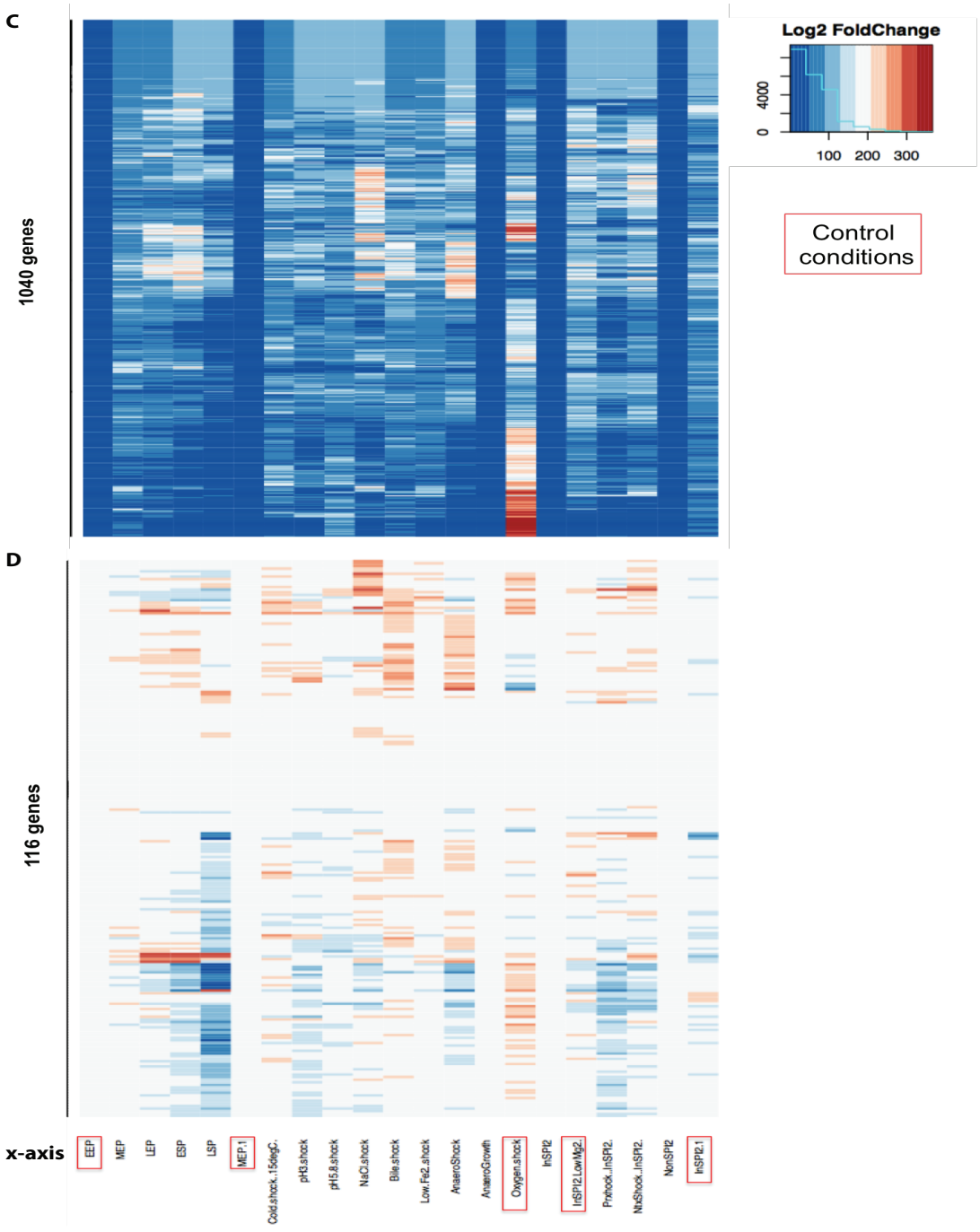


Figure 3.6C-3.6D *Heatmaps in 3.6C and 3.6D illustrate the expression profiles in terms of the differential expressions of the candidate RBP encoding genes (3.6C) and TFs (3.6D) in the y-axis, where the x-axis corresponds to the biological conditions of the samples.*

Table 3.3 ***The abundance of candidate RBPs (ribosomal and non-ribosomal) in contrast to the GO derived transcription factors across the transcriptomic data sets of SalCom.***

Gene type	Total genes	Genes in SalCom
Complete	4657	4456
Transcription factor	216	213
RBPs (putative+known)	1068	1040

Furthermore, using a minimum fold-change cut-off of 1.5, 3,114 expressed genes showed differential expression in at least one of the infection relevant-conditions in the SalCom data set. A total of 860 genes encoding putative RBPs (75% of the total RBPs) and 163 TFs (76% of the total TFs) were recorded as differentially expressed in at least one of the conditions, which represent 28% and 5% of the total number differentially expressed genes, respectively. In the set of 1,993 highly expressed genes, 1,809 genes were differentially expressed, which included 439 candidate RBP-encoding genes (46% of the total RBPs) and 77 TFs (4% of the total TFs), accounting for 26% and 4% of the total highly expressed gene sets, respectively. The detail of the expression patterns in terms of the transcript abundance is shown in **Figure 3.6.** The transcript levels of the candidate RBP-encoding genes in the expression profiles of SL1344 suggest that RBP- and ribosome-related transcripts may have important roles in the translation-related processes due to their high abundance.

3.2.2 Dual RNA-Seq data sets

To further understand the expression of the putative RBPs identified in this study in the context of infection, the expression patterns of the candidates in a dual RNA-Seq data set (Westermann *et al.*, 2016) were also retrieved. Dual RNA-Seq is a technique that has been developed to understand the gene expression patterns of both the host and bacterial pathogen simultaneously during infection. High-resolution RNA-Seq libraries were generated for HeLa cells both before (0 hour) and after infection by SL1344 at different time points of 2, 4, 8, 16 and 24 hours post-infection (hpi) (Westermann *et al.*, 2016). These data sets were analyzed to capture the cumulative abundance of transcripts corresponding to the RBPs and TFs of both human and *Salmonella*. A large proportion of the transcripts in the RNA-Seq data sets represent the human transcriptome, whereas transcripts corresponding to *Salmonella* range from 1% in the libraries generated for the early time-points for infection to 16% in the libraries generated from samples isolated 24 hours post-infection.

The relative expression levels of RBPs and TFs from HeLa and SL1344 are reported here with respect to their reference organism. In agreement with the previous observations by Gerstberger *et al.* (2014), the transcripts related to the human TFs represent an average of 3% of the transcriptomes in both infected and non-infected libraries. The fraction of transcriptome corresponding to SL1344 in the infected libraries represented a similar level of expression (~3%) for the 216 bacterial TFs. The expression level of human RBP encoding genes, which were reported to represent 20% of the transcriptome (Gerstberger *et al.*, 2015), represented ~9% of the total transcriptome in non-infected data sets, increasing only slightly in the infected data set. Interestingly, higher expression of the putative RBP-encoding genes of SL1344 was observed, which contributed more than 10% to the fraction of SL1344 transcriptome of infected libraries in the early stages of infection (2, 4, and 8 hpi), reaching up to 16% and 23% in the libraries generated for the later stages of infection (16 and 24 hours, respectively) (**Figure 3.7**). This steady increase in the transcript contributions between 2 – 24 hpi suggests that the putative *Salmonella* RBP-encoding genes are more highly expressed in the later stages of infection. In the samples infected at 2, 4, 8, 16 and 14 hpi, the number of expressed RBP encoding genes are 313 (30%), 553 (51%), 805 (75%), 861 (80%) and 1062 (99%), respectively. These observations suggested that there are important regulatory roles of the putative RBPs during *Salmonella* infection of host-cells, as only a 30% of RBPs are expressed in the early stage of infection but all are expressed at the late infection stage, and at higher levels.

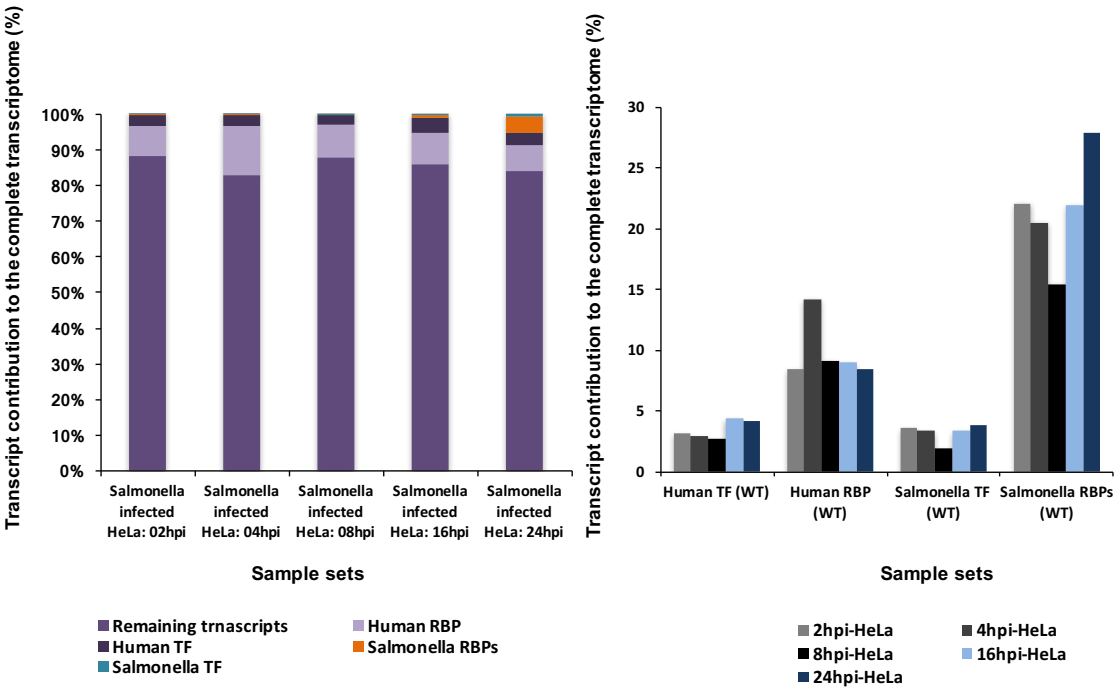


Figure 3.7 ***Quantification of transcript abundance of RBP and TF encoding genes in Salmonella infected libraries of the dual RNA-Seq data set.***

*(**3.7A**) Human RBP and TF encoding genes represent an average of 9% and 3% of the total dual RNA-Seq transcriptome respectively in all* Salmonella*-infected libraries. The contribution of* Salmonella *transcripts to the total transcriptome of the dual RNA-Seq data set (human+*Salmonella*) is considerably smaller due to its smaller genome size. **(3.7B)** Bar chart represents the relative transcript abundance of human and* Salmonella *RBP and TF encoding genes to their own transcriptome. The contribution of genes that encode human RBPs, human TFs and* Salmonella *TFs are consistent in all the samples. However, the* Salmonella *candidate RBP encoding genes of contributed >10% to the fraction of SL1344 transcriptome of infected libraries in the early stages of infections (2, 4 and 8 hpi) reaching up to 16% and 23% in the libraries generated for the later stage of infection at 16 and 24 hpi, respectively.*

To further distinguish genes that play important roles in late stages of infection compared to that important in the early stage of infection, a differential expression analysis (for example, in DESeq2 the adjusted *P* value (Padj) < 0.1 and differential expression log2fold change < -1/> +1 was taken as threshold) was carried out using libraries created at 2 hpi as control samples. At 4 hpi, only 31 human genes and 2 *Salmonella* genes were reported as differentially expressed, indicating only slight changes in gene expression during the early hours of infection. However, a consistent increase in the number of differentially expressed genes was observed in the dual RNA-Seq samples generated at the later time points for both the organisms. At 8 hpi, 112 human genes (3 RBPs and 11 TFs encoding), and 86 *Salmonella*

genes (8 predicted RBPs encoding, for example, *sigD*, *invC*,*prgK,* and *siiC*) and 3 TFs were differentially expressed. At 16 hpi, 747 human genes, which included 14 RBP-encoding genes and 50 TFs, as well as 120 *Salmonella* genes, which included 21 RBP-encoding genes (for example, genes observed at 8 hpi, *nikA*, *fimZ*, *tlpA,* and 2 plasmid-encoded genes) and 3 TFs were differentially expressed. At 24 hpi, the differentially expressed genes in the samples from all the discussed functional groups differed only slightly from the samples associated with 16 hpi.

3.3 Selection of RBP candidates for the experimental validation

From the set of putative RBPs identified by APRICOT in the SL1344 proteome in this study, we selected a subset of 131 candidate proteins, as well as 6 positive controls (Hfq, CsrA, CspA, CspB, SmpB, and YhbJ) that are characterized with RBDs known in both eukaryotes and prokaryotes from the set of classical RBDs (RRM, DEAD, KH, S1, S4, cold-shock, PUA, and LSM). Several less studied domains like SAM, KOW, THUMP (Aravind *et al.*, 2001), WD40, and Nudix (Yang *et al.*, 2010) were also selected. Additionally, it was taken under consideration whether the genes encoding these putative RBPs are expressed in at least one of the conditions in SalCom data set, or are expressed in the dual RNA-Seq infection data set.

In the dual RNA-Seq data set, all positive controls were differentially expressed (DESeq analysis using non-infected sample as control) at both early and later stages of infections with an exception of *smpB*, which was not differentially expressed at 2 hpi. From the set 131 selected candidate genes (excluding the positive controls), 43, 78, 104, 110 and 131 are differentially expressed at 2, 4, 8, 16 and 24 hpi, respectively, which is representative of the proportion of all differentially-expressed APRICOT-selected RBPs that encode genes at these time points (**Figure 3.8**). Of the 43 genes that were differentially expressed 2 hours post-infection, 39 were also found to have altered expression at later infection time points (**Table 3.5**).

Table 3.4 ***A set of 39 genes is differentially expressed across the dual RNA-Seq data sets corresponding to the different time points of infection.***

These genes are selected using the log2 fold-change (log2FC) > 2 compared to the non-infected control and a p-adjusted value < 0.1. All these genes were later subjected to RIP-Seq based experimental studies.

Gene name (locus tag)	2 hpi - Log2FC	4 hpi - Log2FC	8 hpi - Log2FC	16 hpi - Log2FC	24 hpi - Log2FC
aceF (SL1344_0153)	2.25	6.98	7.74	6.75	10.62
acnA (SL1344_1644)	3.91	6.15	8.09	7.27	10.14
acnB (SL1344_0159)	2.71	6.64	6.40	6.93	10.50
cysN (SL1344_2913)	4.58	5.08	7.13	7.53	10.07
deoB (SL1344_4496)	2.61	6.18	6.28	7.85	11.69
dnaG (SL1344_3184)	4.64	6.64	5.48	6.04	12.68
dnaK (SL1344_0012)	3.59	6.78	6.34	5.73	10.46
engA (SL1344_2481)	3.78	6.37	6.67	6.57	9.22
engB (SL1344_3948)	3.42	4.84	4.62	6.47	9.93
ffh (SL1344_2650)	3.54	6.38	7.30	6.02	11.73
glk (SL1344_2371)	4.32	5.18	6.52	5.73	8.21
gltD (SL1344_3303)	2.82	6.29	6.73	6.37	3.96
grxB (SL1344_1102)	3.86	5.52	6.50	6.38	8.47
hflX (SL1344_4296)	3.18	7.31	5.54	6.03	11.04
infB (SL1344_3259)	1.97	4.77	3.31	5.22	9.55
lepA (SL1344_2545)	4.03	5.84	5.85	6.88	9.95
ligA (SL1344_2390)	3.96	5.48	7.27	6.94	10.14
lysC (SL1344_4156)	4.40	5.80	6.67	5.24	8.33
mreB (SL1344_3346)	3.78	6.49	6.12	5.65	10.96
nusA (SL1344_3260)	4.08	7.11	5.15	6.85	11.01
nusB (SL1344_0412)	4.45	6.80	7.12	5.92	11.00
obgE (SL1344_3273)	2.93	6.23	4.56	6.06	12.38
pheT (SL1344_1272)	5.13	9.25	9.77	7.52	8.61
pnp (SL1344_3255)	3.42	8.16	6.64	6.43	10.11
ppiB (SL1344_0529)	5.11	5.80	7.71	7.04	9.47
prfA (SL1344_1704)	4.26	5.45	6.83	6.25	9.25
rho (SL1344_3876)	4.05	7.55	8.34	6.89	9.56
rluD (SL1344_2622)	4.31	8.47	8.00	7.06	10.56
rplW (SL1344_3405)	2.81	7.91	4.16	5.45	10.90
rpmA (SL1344_3275)	2.82	7.02	5.04	7.34	11.65
rpsB (SL1344_0217)	3.79	7.09	5.40	5.97	10.59
rpsC (SL1344_3401)	3.35	7.68	4.64	5.73	11.21
rpsD (SL1344_3383)	4.01	7.47	6.40	6.70	11.00
sdhB (SL1344_0717)	3.71	6.07	8.36	7.32	8.75
secA (SL1344_0136)	2.68	6.14	6.17	5.98	8.84
selB (SL1344_3647)	4.11	4.48	6.57	6.64	9.44
tolB (SL1344_0730)	3.99	6.79	6.63	6.01	11.22
tyrS (SL1344_1381)	3.89	5.76	6.12	5.96	10.45
uvrB (SL1344_0775)	3.08	5.81	6.33	7.00	11.23

Classification of RBP candidates by their RBDs

The selected 131 putative RBPs were next classified based on their RBD architecture identified by APRICOT (**Appendix Table 1**). The categorization was carried out based on the

class of domains identified in the selected proteins. The sub-categories were defined to indicate if the proteins contained RBDs from single class or several classes and if the RBD occurred in one location or multiple locations.

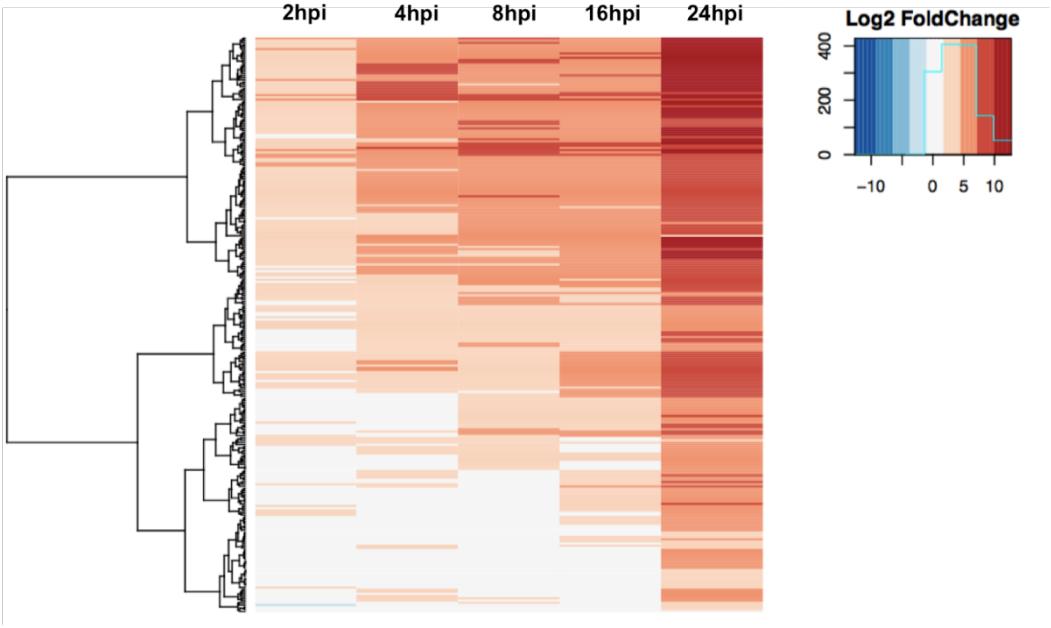


Figure 3.8 ***Differential expression of 131 selected RBP encoding genes in the dual RNA-Seq data of*** **Salmonella-*infected HeLa samples.***

*The differential expression (shown in log2 fold-change values) are obtained from DESeq2 analysis, where the triplicated mock treated HeLa samples were used as controls and salmonella infected HeLa samples were used as treated libraries**.** A set of 39 genes was consistently observed at all the time point of* Salmonella *infections as differentially expressed (**Table 3.5**).*

Several proteins were classified to have only a single RBD class, such as ribosomal domains, DEAD, LSM, cold-shock RRM, SAM, or several non-mRNP domains. Other RBD classes contained less than 3 proteins each. Some proteins were identified with multiple RBDs that occurred single or multiple times in their sequences. For example, many ribosomal domains, cold-shock domains, and KH domains were observed together with other RBDs in the same proteins. In contrast, LSM, WD40, KOW, and Nudix domains did not occur with other RBD classes. A set of 38 proteins were classified into a separate group as they contained repeated RNA-binding sites, most of which were annotated with domain entries from several RBD types from different databases (**Figure 3.9A**). Several proteins were identified that had several non-RNA-binding domains in addition to a single or multiple RBDs (**Figure 3.9B)**.

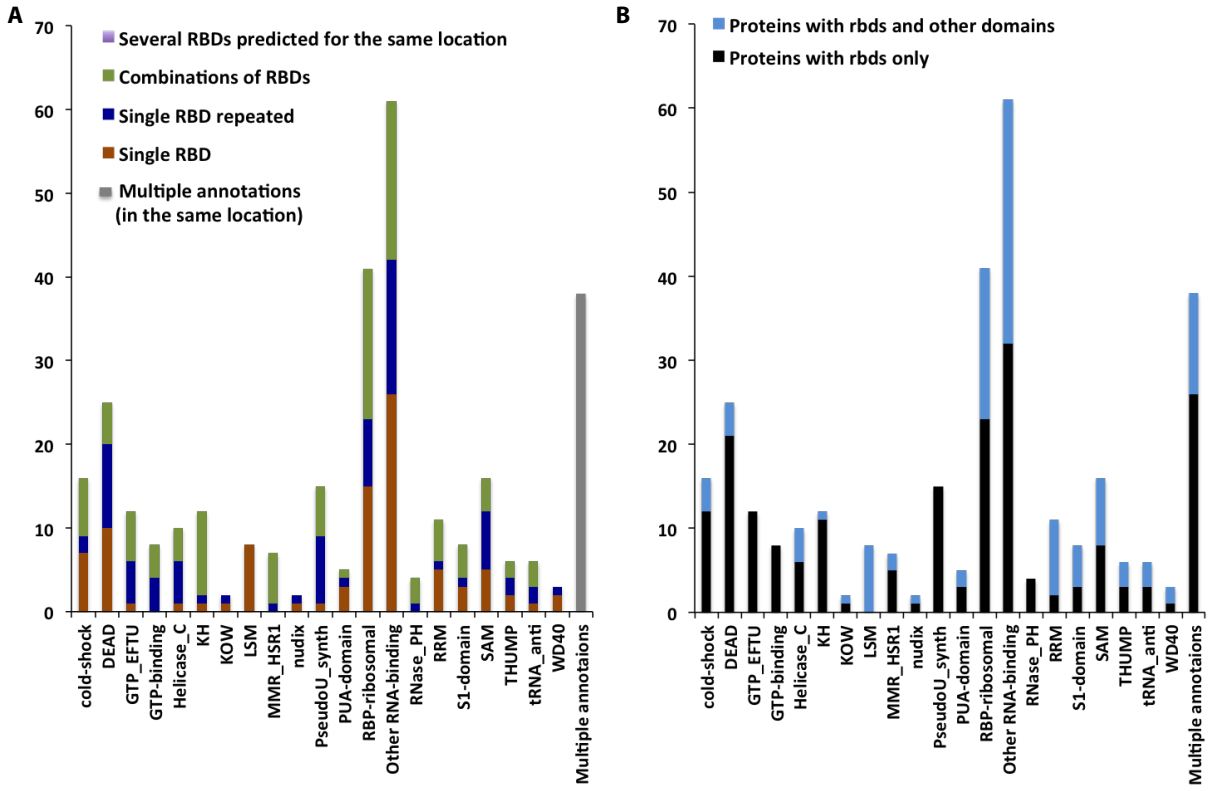


Figure 3.9 ***Domain architecture of RBP candidates.***

Total number of proteins (Y axis) identified with (3.9A) repeated RNA-binding sites that correspond to the domain entries from different databases and (3.9B) non-RNA-binding domains in addition to one or multiple RBDs.

3.4 RNA co-immunoprecipitation combined with sequencing (RIP-Seq) of candidate BPs analysis

The method of co-immunoprecipitation (CoIP) followed by high-throughput RNA sequencing is called RIP-Seq. The coding region of the protein of interest is cloned into a plasmid with a C-terminal 3xFLAG tag, introduced into the organism of interest. In this study, *Salmonella* Typhimurium SL1344 was subjected to CoIP with an anti-FLAG antibody. RNAs that are co-purified with the RBP of interest are then identified by RNA-sequencing. In this study, each of the selected candidate FLAG-tagged RBPs, as well as the tagged positive controls Hfq, CsrA, YhbJ, SmpB, CspA, and CspB and strains carrying empty plasmids as the non-target controls (NT), was subjected to RIP-Seq (see Materials and Methods).

3.4.1 RIP-Seq-based experimental validation of RBP candidates

CoIP combined with sequencing is an extremely useful approach to identify the RNA targets that either directly bind or transiently interact with the RBPs of interest. Hence, RIP-Seq serves as a suitable technique to generate a genome-wide quantitative snapshot of RNA-binding capacity of a protein. The 131 candidate proteins, together with the known RBPs positive controls (Hfq, CsrA, SmpB, CspA, CspB, and YhbJ), were subjected to RIP-Seq analysis (see Materials and Methods). An additional set of samples called NT are generated using empty vectors (containing only a 3xFLAG tag but no candidate RBP), which in principle do not have any RNA binding partner. Hence, NT samples were used as neutral controls to capture any background noise such as highly abundant RNAs that might be non-specifically purified during immunoprecipitation and appear as targets in all the RIP-Seq libraries. In total, 10 replicates of two control samples, Hfq and NT, were also included to validate the reproducibility of the CoIP experiments. Details of sequencing outcomes from each of the positive controls are listed in **Table 3.6** and the data associated to the RIP-Seq libraries (**Appendix Table 1**) will be deposited at GEO.

Table 3.5 ***The total number of enriched targets from the different classes of RNAs in the RIP-Seq samples corresponding to the known RBP controls.***

Samples	sRNAs	mRNAs	tRNAs	rRNAs	Plasmid genes	Total targets
CspA	22	321	0	0	4	347
CspB	29	277	1	0	2	309
CsrA	19	135	2	7	4	167
Hfq	83	501	0	0	15	599
SmpB	16	160	15	3	9	203
YhbJ	4	77	6	0	0	87

The data for each RIP-Seq library was processed by removing adapter sequences from the reads using cutadapt (Version 1.8) (Martin, 2011), and subjected to quality trimming using the fastq_quality_trimmer tool from FastX suite (Version 0.0.13) (http://hannonlab.cshl.edu/fastx_toolkit/) with a phred-score cut-off of 20. READemption version 0.3.7 (Förstner *et al.*, 2014), which uses Segemehl-version 0.1.3 (Hoffmann *et al.*, 2009) for the read alignment, was used for mapping of the reads (minimum length cut-off 12 nt) to the reference genome of *Salmonella* Typhimurium SL1344 (NC_016810, NC_017718, NC_017719 and NC_017720) obtained from NCBI and annotated for the different RNA classes (mRNAs, rRNAs, tRNAs and sRNAs).

3.4.2 Quantification and normalization of quantified reads in RIP-Seq

In general, the gene expression in RNA-Seq samples is captured in terms of read counts. Gene-wise quantification of read counts was carried out for each library to determine the expression profiles of each RNA class in their respective samples. The global changes in the expression of a gene between two samples can be determined by calculating the difference in the fold change of its read counts. In order to accurately estimate such a difference or the relative expression levels of the genes between the samples, various gene expression normalization strategies have been developed. Most of these strategies use the assumption that the majority of the genes are not differentially expressed. One such method is Trimmed Mean of M-values or TMM (Robinson *et al.*, 2010), which is a widely-used method for the calculation of size factors of RNA-Seq data for the further statistical identification of differentially expressed genes.

For comparison of transcript levels in the RIP-Seq libraries in this study, a modified TMM normalization method was introduced that uses a reference-free approach for the normalization of the quantified RIP-Seq data (personal discussion with Dr. Lars Barquist). Unlike the standard TMM approach that requires a reference sample, here a reference gene *set* was derived that comprised the genes with a minimum of 10 transcripts in each sample. The normalization factor was further determined for each sample using the 'calcNormFactors' function of edgeR (empirical analysis of digital gene expression data in R), an R-package for differential expression analysis of RNA-Seq expression profiles (Robinson *et al.*, 2010). The 'calcNormFactors function' finds scaling factors to normalize the libraries by relative RNA composition. The normalized value for each gene in a sample is determined by dividing the raw transcript counts by the size factor of the respective sample derived by scaling the corresponding normalization factor value by the maximum value of the normalization factor obtained in the data set.

3.4.3 Selection of the enriched genes

The enrichment of particular transcripts in the CoIP sample of a candidate RBP may indicate a particular class of binding partners of the protein in study (Faoro & Ataide, 2014). In our study, RIP-Seq libraries of empty plasmid (NT) samples were used as a negative control, which theoretically should not show enrichment of transcripts corresponding to specific genes. However, transcripts from some genes were in fact enriched in NT samples as well, which may represent transcripts that are non-specifically co-purified, independently of any specific RNA-binding activity.

Such unspecific enrichment of transcripts could result from heir abundance in the cells, which might lead to non-specific co-purification with the experimental reagents (König *et al*., 2012). This poses one of the crucial challenges of RIP-Seq data analysis, which is to separate transcripts that are enriched in specific CoIP samples from the transcripts that are enriched non-specifically in several CoIP samples. In order to identify the set of genes that are truly enriched in each RIP-Seq library, the statistical dispersions of normalized read counts were first measured by determining quartiles of the read distribution of each gene across our RIP-Seq samples. The NT replicates were used to measure the abundance of each transcript that resulted from unspecific binding of RNAs to the antibody-Sepharose, and Hfq libraries were used to represent the enrichment profile of a representative global RBP.

Quartiles (Q) of a ranked set of values are points that divide the data set into four equal quarters (Hyndman & Fan, 1996). For example, the first quartile (Q1) divides the lowest 25% of the data set from the rest, second quartile (Q2) divides the data set in half, and third quartile (Q3) divides upper 25% of the data set from the rest. The interquartile range (IQR) is the difference between the Q3 and Q1. A term 'fence' is used to indicate upper and lower limit of the data. Any data lying outside the fence can be considered to be extreme value or outlier. Based on this principle, the upper fence for each gene was recorded as minimum threshold for the enrichment of a gene by calculating Q3+1.5*(IQR) of the normalized read count distribution across the RIP-Seq libraries (**Figure 3.10**). In order to avoid the genes that are found in all the samples and potentially include unspecific binding transcripts, an additional cut-off of a minimum of 50 reads in was applied during the selection of the enriched genes.

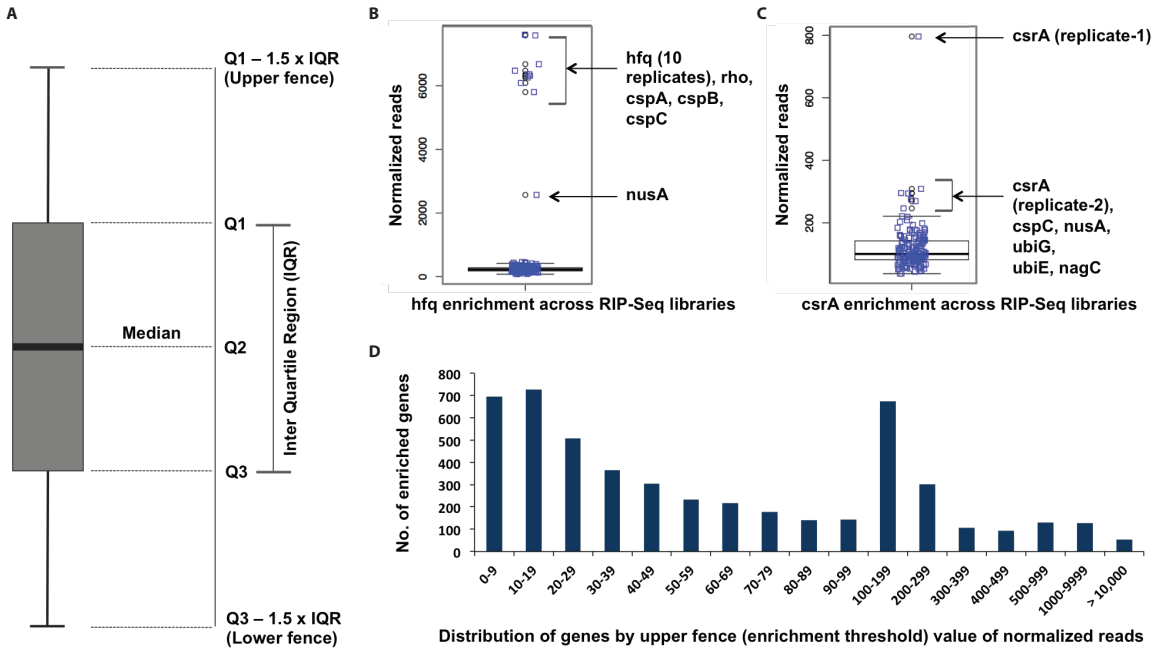


Figure 3.10 ***Overview of variance based threshold used for the selection of enriched genes in the RIP-Seq data sets.***

(3.10A) The quartile based variance analysis can be explained by means of a boxplot. In this study, the variance analysis was carried out on the modified TMM normalized read counts for every gene across each library. Subsequently, the upper fences (Q3+1.5(IQR)) of the quartile distributions were recorded as minimum thresholds for the enrichment of the corresponding genes. (3.10B and 3.10C) Boxplots showing the variance analysis on the normalized reads across the RIP-Seq libraries for the positive controls Hfq (3.10B) and CsrA (3.10C). hfq and csrA genes are enriched in their own RIP-Seq libraries, which positively validates the threshold selection criteria for the gene enrichment. Furthermore, additional RIP-Seq libraries where these genes are enriched suggest more binding partners of these RNAs. (3.10D) The bar chart shows the distribution of upper fence values (in x-axis) versus the number of genes (y-axis). This shows the expression patterns of the genes, which could be classified into lowly expressed genes (read count <10 reads), highly expressed genes (read count >100 reads) and moderately expressed genes (read count between >10 and <100).*

The transcripts from the RIP-Seq libraries were assigned to the different RNA classes (mRNAs, tRNAs, rRNAs, and sRNAs) of the SL1344 reference genome and the gene-wise quantification and its normalization was carried out. A modified TMM-normalization approach was applied to further identify enriched genes in the RIP-Seq libraries (see Materials and Methods).

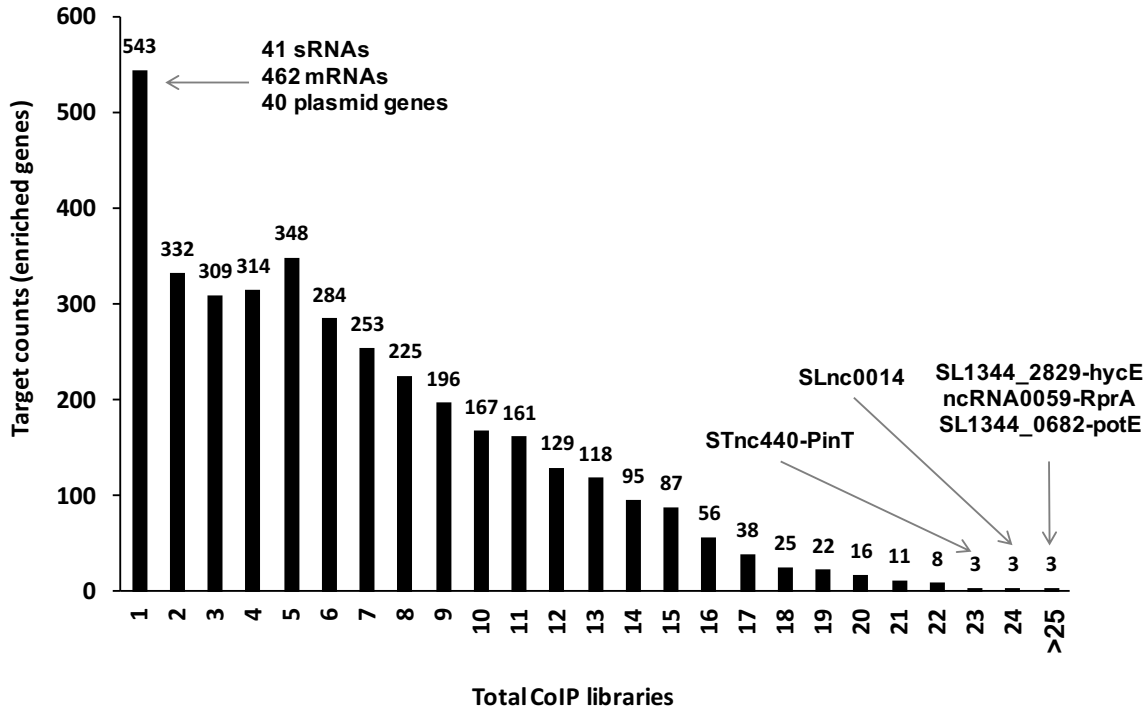


Figure 3.11 ***Distribution of 3746 genes expressed across the RIP-Seq libraries.***

Only 28 genes were enriched in more than 20 libraries, whereas 543 genes were recorded as enriched in only one of the RIP-Seq libraries. This enrichment patterns indicate differential target affinities of the RBPs.

For the computation of normalization factors independent of any reference or control library, genes that were found to have a minimum of 10 reads in each RNA-Seq library were selected, which consisted of a total of 799 genes. Subsequently, the size-factors were estimated, which were used for the normalization of the gene-wise quantification data. Further, the statistical dispersions of normalized read counts of each gene across RIP-Seq samples were calculated by the quartile approach and the upper fences (Q3+1.5*IQR) of the gene expressions were derived. The selection of enriched genes in our study uses a two-fold selection strategy: 1) the expression level of the genes must be higher than their corresponding upper-fence cut-off calculated from the quartile approach, and 2) there must be a minimum of 50 transcripts corresponding to the genes that satisfy the upper-fence cut-off criteria to avoid weakly-expressed genes. Upon selection of enriched RNAs, or candidate targets of each protein, a set 1,246 coding or sRNAs genes was identified that did not show enrichment in any RIP-Seq library, which were removed from further analysis. Hence, 3,746 genes that were enriched in at least one RIP-Seq sample were used for further analysis.

Only 28 genes were enriched in more than 20 samples, whereas 543 genes were detected as enriched in only one of the samples (**Figure 3.11**). These observations indicated that

according to the aforementioned gene selection criteria, most of the potential non-specific binders were excluded from appearing as enriched genes. Our selection strategy might exclude a few potentially enriched genes that are weakly expressed, but it avoids the inclusion of false positives. This was verified by recording the enrichment levels of the known RBP binders in our positive controls (discussed later).

3.5 Classification of RIP-Seq libraries by interacting RNA classes

The RBP candidates were next classified based on the genes enriched in their corresponding libraries, as this can explain the similar or contrasting functionalities of these proteins based on their interacting partners. As reported in the literature, Hfq is a global RNA-binding protein that has several mRNA and sRNA binding partners (Chao & Vogel, 2010; Holmqvist *et al.*, 2016). Consistent with such studies, Hfq was determined in our study to have 80 sRNA targets (the mean of 10 replicates) analysis, which was the highest number of sRNA targets for any of the RBPs we tested. These enriched sRNAs included many known sRNA targets, such as ArcZ, ChiX, CyaR, DapZ, DsrA, GcvB, GlmZ, InvR, MicF, OmrA, OmrB, PinT, RprA, RybB, RydC, and SgrS (Chao *et al.*, 2012).

In addition, ~500 mRNA targets were also found to be enriched in the Hfq tagged libraries compared to the NT control, which was a higher number than for most of the RBPs (**Figure 3.12A**). A few transcripts encoded on the SL1344 plasmids were also enriched in Hfq libraries. However, no enrichment of tRNAs or rRNAs was recorded. The list of enriched genes in the CoIP sample of second positive control, CsrA (**Figure 3.12B**), was comprised of 19 sRNAs, including the known CsrA-binding antagonizing sRNAs CsrB and CsrC, as well as PinT (Holmqvist *et al.*, 2016). Additionally, 135 mRNAs, 2 tRNAs and 7 rRNAs were also enriched in the CsrA tagged library, which demonstrated a relatively global binding pattern like that of Hfq, but to more diverse classes of RNA targets. More than 300 genes were enriched in the FLAG-tagged CoIP samples of CspA and CspB (**Figure 3.12C**), including 140 shared targets and indicating similarity in their specificities as RBPs. YhbJ, also known as RapZ (Göpel, 2014), was included in this study as a control for RBPs that have specific binding partners. It has been reported that RapZ binds two specific sRNAs, GlmY and GlmZ (Göpel, 2014), which were also enriched in our study. Together, these outcomes of the RIP-Seq experiments with previously characterized RBPs shows the effectiveness of our CoIP

approach for identifying targets of RBPs analysis further classify them based on their preference for specific RNA classes.

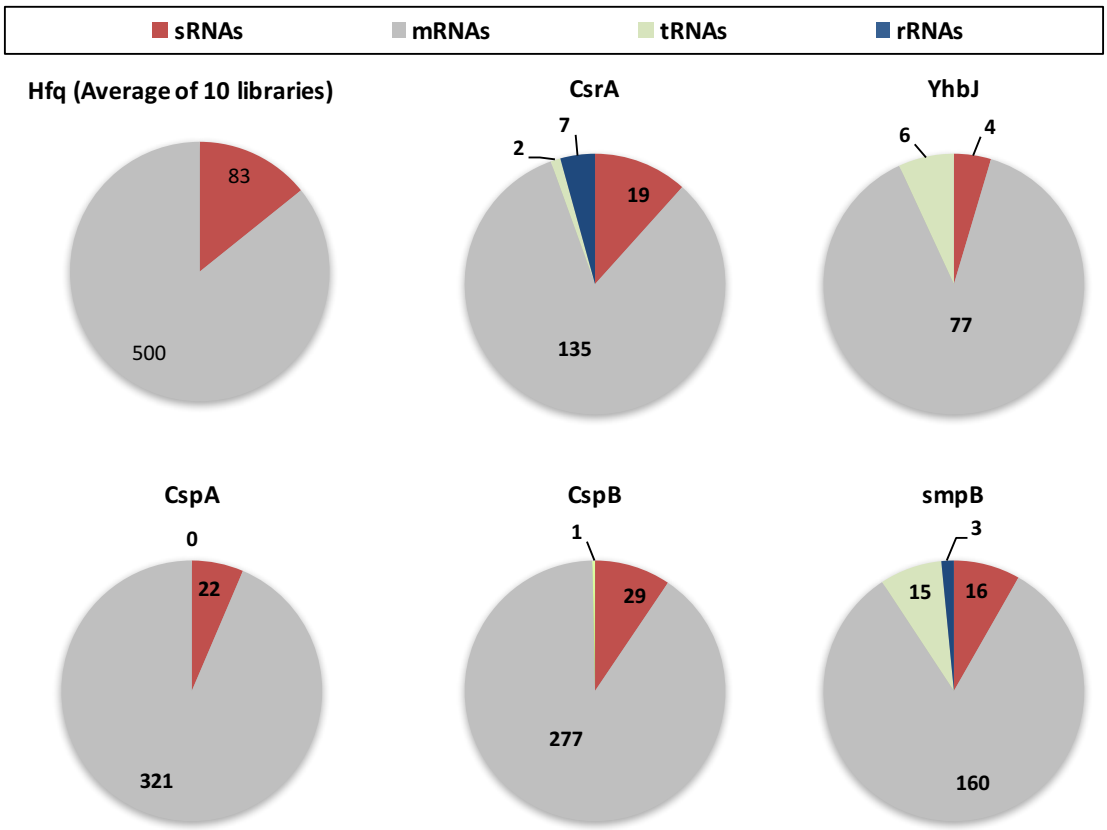


Figure 3.12 ***Gene enrichment profiles of the positive control libraries of Hfq, CsrA, YhbJ, CspA, CspB and SmpB.***

As shown in the pie charts, an average of 500 mRNA targets and 83 sRNA targets were recorded in the Hfq libraries. Hfq has a higher number sRNA targets compared to other libraries, which include known targets namely, ArcZ, ChiX, CyaR, DapZ, DsrA, GcvB, GlmZ, InvR, MicF, OmrA, OmrB, PinT, RprA, RybB, RydC, and SgrS. Similarly, among the enriched targets for CsrA, 19 sRNAs including the known targets CsrB, CsrC, and PinT were recorded. In the library of YhbJ, its 2 known sRNA targets/glmY and glmZ) were enriched. More than 300 genes were enriched in CspA and CspB libraries that shared 140 common targets indicating similar and/or cooperative activities. SmpB library enriched targets from all the 4 classes of RNAs (mRNA, tRNA, rRNA, and sRNA), which was also the case for CsrA libraries.

In our empty plasmid sample, as non-target (NT) controls, only a few genes were enriched (ranging from 4 to 12 targets across the 10 NT libraries). There are an additional 15 proteins in our set of candidates that showed less than 12 targets from different RNAs classes, which can be considered as low confidence RBPs or false positives based on their enrichment patterns similar to the NT libraries (**Table 3.7**).

Table 3.6 ***The RIP-Seq libraries of the candidate RBPs that showed an enrichment profiles similar to the background controls (NT samples).***

These 15 out of 131 proteins enriched less than 12 targets from different classes of RNA.

Samples	sRNAs	mRNAs	tRNAs	rRNAs	Total targets
NT-1	2	2	0	0	4
NT-2	1	3	0	0	4
NT-3	1	3	0	0	5
NT-4	1	4	0	0	5
NT-5	2	5	0	0	7
NT-6	1	5	3	0	9
NT-7	1	9	0	0	10
NT-8	2	6	0	0	11
NT-9	0	7	5	0	12
NT-10	0	7	5	0	12
citC (SL1344_0612)	0	7	0	0	7
deoB (SL1344_4496)	6	5	0	0	12
dnaK (SL1344_0012)	0	13	0	0	13
glk (SL1344_2371)	0	8	0	0	8
glnK (SL1344_0456)	0	5	0	0	5
lysC (SL1344_4156)	1	6	0	0	7
rluA (SL1344_0096)	0	8	1	0	9
rph (SL1344_3700)	1	9	0	0	10
rpsC (SL1344_3401)	0	5	0	0	5
rrmA (SL1344_1764)	2	5	0	0	7
sgaB (SL1344_4317)	6	6	0	0	13
SL1344_1712	0	13	0	0	13
SL1344_2929	0	2	1	0	3
SL1344_3369	2	8	1	0	11
SL1344_3587	0	7	0	0	7
SL1344_3646	0	9	0	0	9
yffB (SL1344_2445)	0	4	0	0	4
yqcB (SL1344_2945)	0	9	3	0	12

Besides the positive controls, the CoIP libraries of 43 proteins showed enrichment of more than 10 sRNAs, including that of RpsD, TolA, Res, Rho, TufB, TolB, GatB, AcnA, YdiL. and AefA, which al have more than 30 enriched candidate sRNAs targets. Besides Hfq, the CoIP libraries that showed about 500 enriched mRNAs were SL1344_3412 (TufB), SL1344_3876 (Rho), SL1344_3260 (NusA), SL1344_2200 (YejH), SL1344_1196, SL1344_1766 (CspC), and SL1344_0617 (CspE). Several proteins, including Hfq, were found to have several plasmid genes as their targets. More than 20 plasmid genes were enriched in the CoIP libraries of the RBP candidates Rho, NusA, RluC, Pnp, and TufB. A few proteins, namely TolB, Res, NusB, GatB, YbdG, and SL1344_3432 had >30 tRNA targets enriched in their CoIP libraries. These

proteins show an exceptionally high preference for tRNA targets compared to 100 other libraries in our data set, most of which did not have any tRNA targets.

Similarly, more than 100 libraries in our data set did not show any preference for rRNA targets. However, several proteins, including AcnA, AceA, RluC, Res, RpsD, CaiD and TraR, had more than 14 rRNA targets. Interestingly, several proteins showed specificity for certain RNA classes, whereas a few other enriched large proportions of genes from all the RNA classes. Two examples of the RBPs enriching different proportions of RNA targets from all the RNA classes are Res and TolB. The putative RNA targets of the Res protein included 44 sRNAs, 73 mRNAs, 81 tRNAs, 15 rRNAs, and 7 plasmid-encoded genes. The targets of TolB showed its higher binding preference for sRNAs (35 targets), mRNAs (83 targets), tRNAs (82 targets) and a lower preference for rRNAs (4 targets) and plasmid genes (3 targets).

Table 3.7 ***The classification of the candidate RBPs based on the total number of RNA targets enriched in their RIP-Seq data.***

(Please refer to Appendix-1 for details).

Target number	Total samples	Controls
> 200	35	Hfq, CspA, CspB, SmpB
> 100, < 200	23	CsrA
> 50, < 100	32	YhbJ
< 50, > NT targets	32	None
<= NT targets	15	NT (control not counted as sample)

For the purpose of classification, the candidate RBPs were ordered by the number of targets and their enrichment of the different RNA classes (**Table 3.8**). The positive control, Hfq was established as the top-ranking protein in our data set due to its overall high number of targets, as well as having the highest number of sRNA and mRNA targets. The other three proteins with large number of targets, Rho, TufB and NusA, were ranked high in our list demonstrating a similar enrichment pattern as Hfq. These proteins also showed a high preference for plasmid genes. By only considering the overall number of enriched genes, 35 proteins were categorized together on the basis of having a total number of enriched genes higher than 200 (global).

Besides the positive controls Hfq (500 targets on average), CspA (346 targets), CspB (309 targets), and SmpB (202 targets), this group of proteins included other cold-shock proteins such as CspC (461 targets), CspE (447 targets), and CspD (213 targets), which were recorded

to have unique enriched targets from sRNA and mRNA classes. A total of 23 proteins were grouped together that had a number of enriched targets in a range of 100 to 200 (similar to the global binders). This group of proteins included the positive control CsrA (165 targets) and another cold shock protein, CspH (100 targets). In addition, 64 proteins that had 13-99 enriched genes (relatively specific binding partners), were grouped together. This group included YhbJ (87 targets), which showed enrichment of its previously described specific sRNA binding partners. The remaining proteins that enriched less than 12 targets were placed in the bottom of our ranked list, together with the NT controls (potentially non-binders).

We next explored the RNA-binding activity of one class of proteins in more detail: the pseudouridine synthases (TruD, RluB, RluC, RluD, YciL, YjbC, YqcB, and RluA), which were subjected to RIP-Seq based analysis in our study. Posttranscriptional modifications of cellular RNAs are carried out by pseudouridine synthases, hence it is not surprising that these proteins were computationally identified to bind to RNAs. Upon RIP-Seq analysis, four members of this family (TruD, RluB, RluC, and RluD) were identified to enrich more than 100 targets. YciL and YjbC libraries enriched 96 and 30 targets respectively. The two remaining proteins: YqcB and RluA, were placed with the NT controls as non-binders. The pseudouridine synthases were highlighted as an important group of RNA-binding proteins.

3.6 Hierarchical clustering of RIP-Seq samples

Next, an unsupervised hierarchical clustering (Eisen *et al.*, 1998) and principle component analysis (PCA) (Hotelling, 1933; Alter *et al.*, 2000; Jolliffe, 2002; Ringner, 2008) of the RIP-Seq libraries were carried out to statistically derive their relatedness based on their enriched, potentially target genes. This analysis was carried out on the normalized gene expression profiles corresponding to each library. The clustering approach was used in order to further group the RBPs based on the similarity of their RIP-Seq enrichment profiles based on the calculation of the distances between their expression profiles (See Material and Method). This led to the clustering of genes with similar expression profiles and samples with similar enrichment profiles.

The expression profile for each gene per library was recorded as a table in such a manner that the rows correspond to genes and columns correspond to the samples. Genes that were

not enriched in any RIP-Seq samples were removed from the table. The expression table was then subjected to unsupervised hierarchical clustering (see Materials and Methods) in order to cluster the samples in the x-axis and genes in y-axis based on their expression patterns. In x- and y-axis, the samples with similar enrichment profile and genes with similar expression levels should occur together in the same cluster, which has been shown in the heat-maps generated for normalized read counts (**Figure 3.13**).

As observed in the heat-map, several clusters comprising of only RBPs with a high specificity for only tRNAs, only rRNAs, or only several highly expressed sRNAs were observed on the y-axis. On the x-axis, this approach efficiently clustered the 10 Hfq RIP-Seq libraries together. The NT libraries were grouped separately in different clusters, which were distantly located from the Hfq showing an apparent difference in their expression profiles. Similarly, other positive controls (CsrA, CspA, and CspB) that bind to several targets appeared together in another cluster that was closely located to the Hfqs. YhbJ occurred in a cluster located away from the Hfq and CsrA libraries highlighting a markedly different expression profile from the global RBPs. Upon observing the expected clustering patterns of the positive and neutral background controls, I used this approach of hierarchical clustering for the classification of the RBP-candidates in our RIP-Seq libraries based on the expression and enrichment profiles.

The cluster analysis was next repeated on the expression profiles while retaining only one out of 10 Hfq libraries in the expression table in order to identify candidate RBPs that have RIP-Seq profiles similar to that of Hfq (**Figure 3.14**). As discussed previously, proteins like Rho, CspC, CspE, TufB, and NusA had more than 100 candidate RNA targets, like Hfq, suggesting they could be more global RNA-binding proteins. These 6 libraries, along with 15 other libraries, clustered together, underscoring their similar in expression patterns. The next cluster was comprised of 13 libraries, including three positive controls (CspA, CspB and SmpB). CsrA clustered with 13 other libraries that included 2 cold shock proteins (CspD and CspH). These three clusters together accounted for 43 of the candidate RBP RIP-Seq libraries (excluding the controls for global RBPs), which appear to exhibit a more global RNA binding profile.

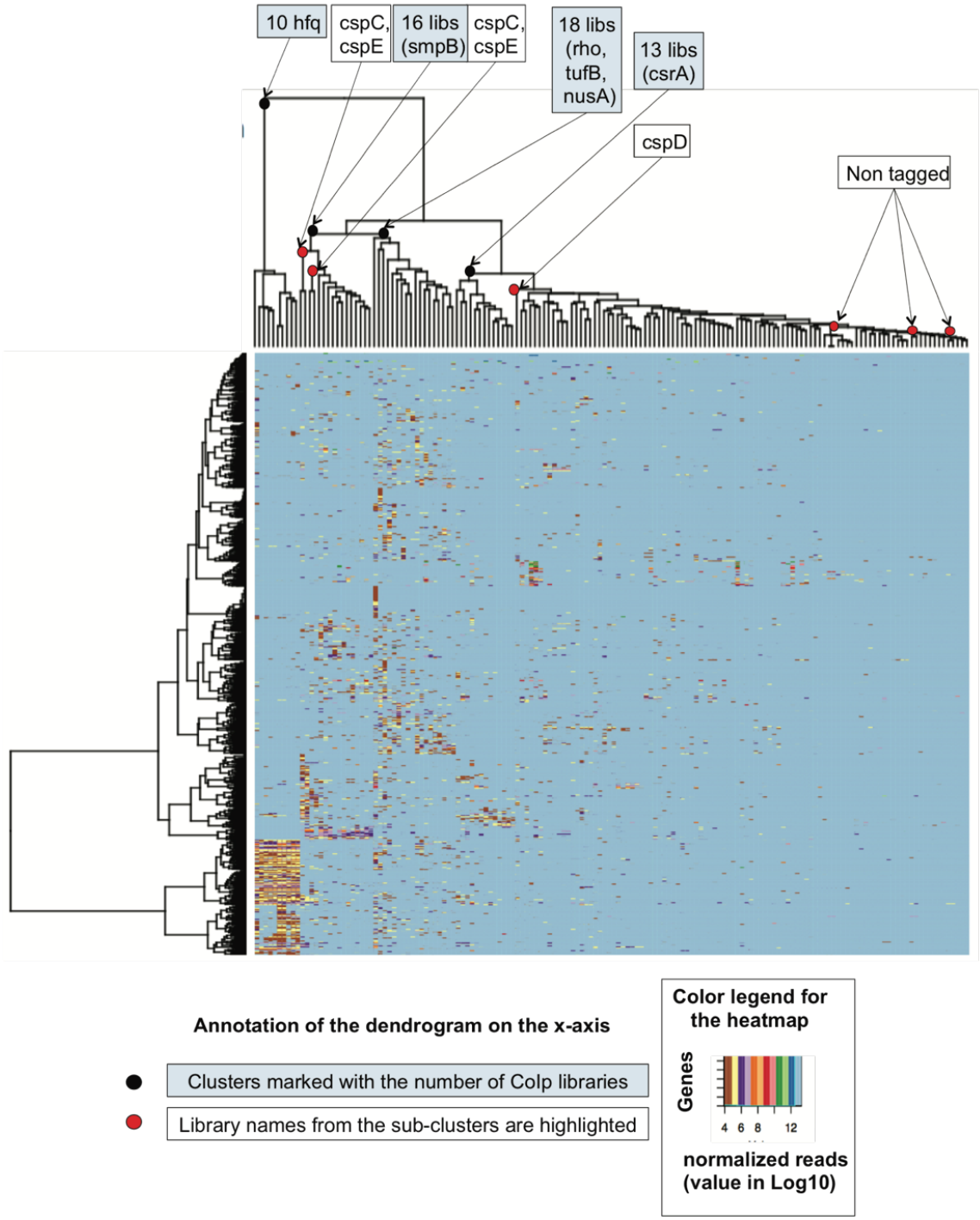


Figure 3.13 ***Hierarchical clustering of the RIP-Seq libraries based on their normalized read counts.***

As shown in the heatmap, samples with similar enrichment profiles were grouped together in the x-axis and genes with similar expression levels were grouped together in the y-axis. Three positive controls (Hfq, CsrA and SmpB) and the not-targets as neutral non-target samples are annotated on the heatmap. Additionally, few clusters have been pointed for the interesting enrichment profiles and clustering patterns of the RBP candidates.

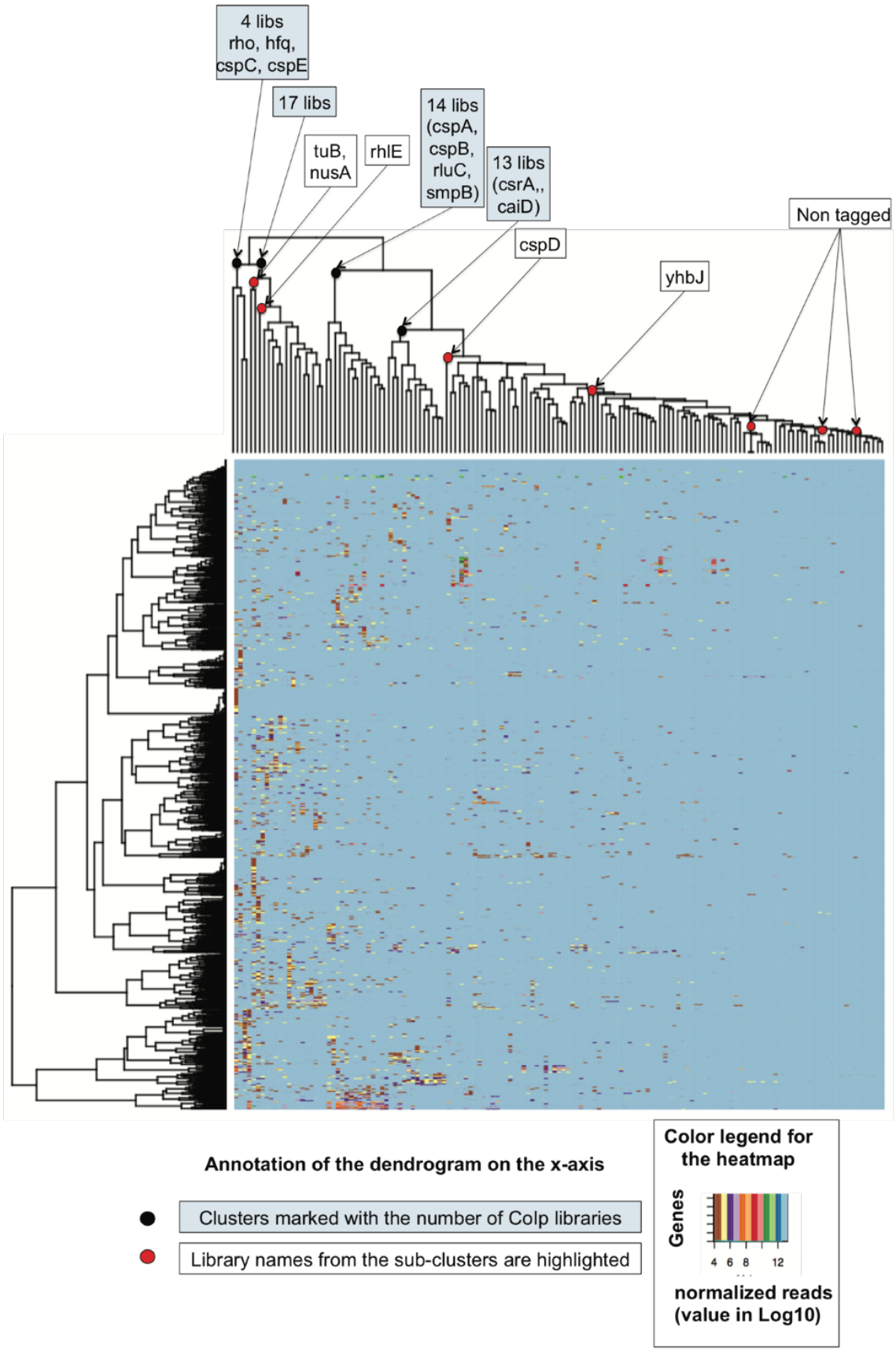


Figure 3.14 ***Hierarchical clustering of the RIP-Seq libraries based on their enrichment profiles by using one representative Hfq library out of 10 as a positive control.***

This analysis was performed to avoid any clustering bias that could have been introduced by the significant enrichment profiles of Hfq replicates.

The NT-containing clusters included an additional 22 libraries with a low number of RNA-targets (3-23 targets) and may therefore represent false positives from the APRICOT analysis. The clusters located in between the more global RBPs and the NT-containing clusters accounted for the remaining 67 RBP candidates, as well as the YhbJ control for a specific RBP. Based on these observations, the RIP-Seq data could be classified into three classes, global RBPs (43 proteins), specific RBPs (66 proteins), and low-target proteins (22 proteins).

The above clustering analysis provides an effective method to separate libraries with different number of targets. However, the wide range of expression of transcripts (ranging from 0 to several thousand across the genes) does not allow their clustering based on gene enrichment. Hence, in order to conduct enrichment-based clustering, the expression profiles were transformed into binary values. The set of genes that are not enriched were given a value of 0 each, while the enriched genes were given a value of 1 each. The positive enrichment was determined when the total number of transcripts was higher than 50 reads and also above the quartile based cut-off for each gene.

When the clustering was performed on the resulting binary enrichment table **(Figure 3.15)**, all the Hfq libraries again clustered together, while similar pattern was seen for 8 of 10 NT libraries, which clustered separately. CsrA clustered with 19 other libraries and YhbJ occurred in separate cluster with 6 other libraries. The positive controls CspA and CspB occurred together with the cluster of 14 libraries that included SmpB and the two cold-shock proteins CspC and CspE. When the RIP-Seq libraries of the candidate RBPs cluster with positive controls (such as Hfq and CsrA), it signified a certain level of similarities in their enrichment patterns, for example their binding specificity to common targets. In contrast, the clustering of RIP-Seq libraries with NT libraries showed the low RNA-binding potential of those proteins, for example, the potential non-RBP SL1344_3369 clustered with 8 NT libraries. The remaining 2 NT libraries were present in a separate cluster highlighting the low binding potential of these candidate RBPs. This cluster comprises of 14 non-NT libraries: YffB, Glk, RpsC, GlnK, RrmA, CitC, SL1344_2929, YqcB, YhjS, RluA, SL1344_3646, Rph, SL1344_3369, DeoB, and LysC. Based on the enrichment based clustering, the RIP-Seq data can be divided in the four groups: global RBPs (51 proteins), specific RBPs (52 proteins) and low-target proteins (28 proteins). The maximum number of targets for the NT library was recorded as 12, which allowed the distribution of the low-target proteins into 2 groups. The first group comprised of 15 proteins that have total targets <12, therefore they could be

classified as NT-like non-RNA-binding proteins. The second group that comprised of 13 proteins could be labelled as possible RBPs due to target number > 12.

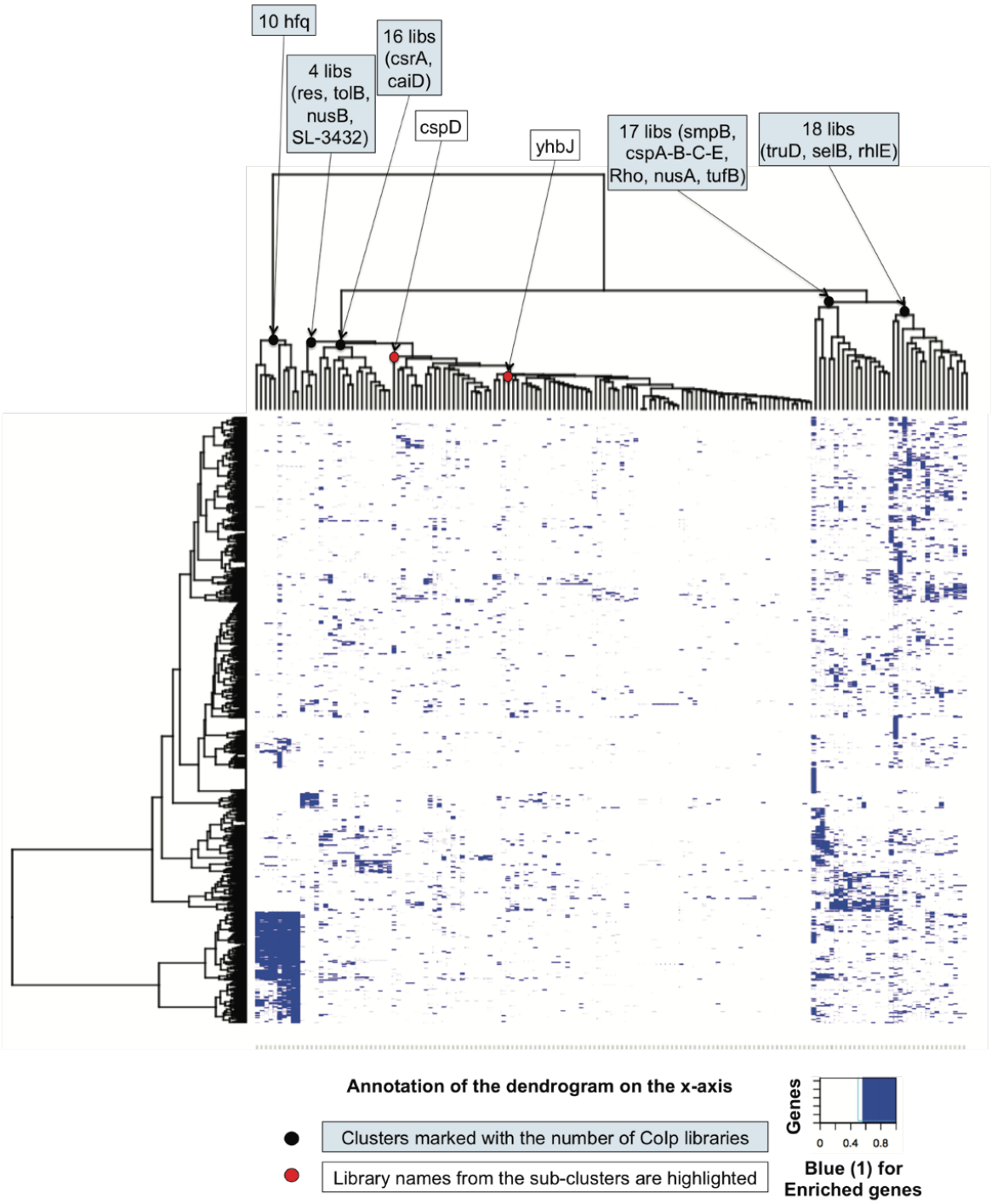


Figure 3.15 ***Cluster analysis of the binary transformed values of enrichment table.***

For each RIP-Seq library, the enriched and non-enriched genes were denoted with 1 and 0 respectively. Upon unsupervised clustering, the Hfq libraries were grouped together and the similar pattern was seen for 8 of 10 non-target (empty plasmid) libraries, which indicated a successful clustering of the libraries with similar enrichment.

3.7 Functional characterization of putative RBPs

3.7.1 TraDIS-based characterization of the RBPs

Genetic elements called transposons can randomly integrate in genomes, in a process that is mediated by transposase enzymes. The enzyme can recognize inverted repeats in the flanking region of a transposon, which is inserted into its target sequence by introducing double strand breaks. Such transposons have been adapted for random insertional mutagenesis of bacterial strains, allowing forward genetic screens, which directly link a phenotype with genotype. Recently, several deep-sequencing-based techniques have been developed to allow genome-wide identification of transposon insertion sites in high-density mutant pools before and after selection. One such study was carried out in *Salmonella* Typhi by introducing insertion every 13 bp in the genome, followed by transposon-directed insertion site sequencing (TraDIS) (Langridge *et al.*, 2009; Chaudhuri *et al*., 2013; van Opijnen & Camilli, 2014), which determined the essentialities of genes under different growth and infection-related conditions.

In one of these studies (Chaudhuri *et al.*, 2013), a library of around 10,000 transposon mutants was generated using Tn5 and Mu transposons, and this mutant pool was then subjected to selection in different hosts. This was followed by TraDIS to identify insertion sites in the input and output pools and estimate the contribution of each non-essential *Salmonella* gene to fitness. TraDIS assigned insertion sites and fitness scores (log2 fold-change) values obtained from the DESeq analysis of the input versus output reads, where a negative fitness score signifies an attenuated mutant and a positive score signifies a mutant that are abundant in the output pool than in the input. Genes were categorized as important in colonization if they were disrupted in one or multiple mutant in any of the host species. I used this data set in order to characterize genes that encode computationally identified RBPs in *Salmonella* Typhimurium and are important under infection-relevant conditions annotated by TraDIS.

From the set of 1,068 putative RBP-encoding genes identified by APRICOT (Chapter 2), 750 genes were available in the TraDIS data set, of which 501 genes were reported to have a statistically significant score (Chaudhuri *et al.*, 2013). By further including a maximum *P* value threshold of 0.05, a total of 110 genes were identified that were suggested to have a potential role in colonization. This included three of the RBP positive control genes (*hfq*,

yhbJ, and *smpB*). The high-significance set of 110 genes includes several genes that are annotated as enzymes such as helicases (*recG*, *recQ*, *srmB*, *uvrD,* and SL1344_3253), nucleases (*ams*, *recD*, *recJ*, *rnG*, *rnhB*), synthases (*carB*, *ssaN*, *truB*, *atpD,* and SL1344_1651), and synthetases (*purA*, *purD,* and *yjeA)*.

It also included the *sseC* and *ssaN* genes that are related to the type III secretion systems. Other examples include, genes encoding a global response regulator (*arcA*), two-component response regulators (*dcuR*, *ssrB*) and transcriptional regulatory proteins (*hilA, phoP*, *slyA*, *yihW*, SL1344_4350, SL1344_3095, SL1344_2314). The RIP-Seq libraries involved in this study included 19 proteins including Hfq, YhbJ and SmpB that are encoded by the genes that are reported in the TraDIS investigation as relevant to colonization of animal reservoirs and hence zoonosis (**Table 3.7**). These genes can be further investigated and characterized for their exact roles during infection.

Table 3.8 ***The functional characterization of the RIP-Seq libraries of the RBP candidates using TraDIS data sets.***

The table contains a set of 19 candidate RBP encoding genes that includes the positive controls hfq*,* yhbJ *and* smpB*, were highlighted in this analysis that are reported to be relevant for zoonosis in TraDIS investigation. (reference: Choudhury et al., 2013)*

Chromosome/ plasmid	Location	Transposon	Gene name	Product	Fitness Score	*P* value
Chromosome	187898	Mu	*acnB*	aconitate hydratase 2 (citrate hydro-lyase 2)	-15.00	0.018
Chromosome	161170	Mu	*secA*	preprotein translocase SecA subunit	-15.00	1.02E-40
Chromosome	161180	Mu	*secA*	preprotein translocase SecA subunit	-15.00	1.03E-42
Chromosome	3468322	Mu	*deaD*	ATP-dependent RNA helicase (dead-box)	-15.00	0.015
Chromosome	815168	Mu	*tolA*	tolA protein	-15.00	2.53E-39
Chromosome	815202	Mu	*tolA*	tolA protein	-15.00	1.62E-38
Chromosome	815461	Mu	*tolB*	tolB protein precursor	-15.00	1.53E-34
Chromosome	814537	Mu	*tolA*	tolA protein	-7.31	1.16E-06
Chromosome	4069900	Mu	*thdF*	thiophene and furan oxidation protein	-6.20	1.44E-19
Chromosome	4070152	Mu	*thdF*	thiophene and furan oxidation protein	-5.54	7.87E-17
Chromosome	3488855	Tn5	SL1344_3270	Intergenic: ftsJ-yhbY, overlap SL1344_3270	-5.45	1.88E-19

Chromosome	3959574	Mu	*recG*	ATP-dependent DNA helicase	-5.29	9.08E-20
Chromosome	800596	Mu	*sdhB*	succinate dehydrogenase iron-sulfur protein	-5.06	4.68E-19
Chromosome	3959603	Mu	*recG*	ATP-dependent DNA helicase	-5.01	6.79E-16
Chromosome	3958654	Mu	*recG*	ATP-dependent DNA helicase	-4.89	2.17E-13
Chromosome	3468339	Mu	*deaD*	ATP-dependent RNA helicase (dead-box protein)	-3.96	5.82E-12
Chromosome	3468785	Mu	*yrbN*	conserved sORF	-3.71	2.05E-12
Chromosome	2837910	Tn5	*smpB*	SsrA (tmRNA)-binding protein	-3.56	2.41E-11
Chromosome	4204949	Mu	*rfaH*	transcriptional activator	-3.50	6.84E-10
Chromosome	188099	Mu	*acnB*	aconitate hydratase 2 (citrate hydro-lyase 2)	-3.41	2.60E-08
Chromosome	2725260	Mu	*lepA*	GTP-binding protein LepA	-3.25	1.87E-08
Chromosome	1838770	Mu	*ychF*	hypothetical ATP/GTP-binding protein	-3.18	2.90E-08
Chromosome	1839467	Tn5	*ychF*	hypothetical ATP/GTP-binding protein	-3.10	1.07E-07
Chromosome	1839351	Mu	*ychF*	hypothetical ATP/GTP-binding protein	-2.99	6.69E-07
Chromosome	1838804	Mu	*ychF*	hypothetical ATP/GTP-binding protein	-2.94	9.91E-07
Chromosome	187272	Mu	*acnB*	aconitate hydratase 2 (citrate hydro-lyase 2)	-2.82	1.23E-06
Chromosome	187038	Mu	*acnB*	aconitate hydratase 2 (citrate hydro-lyase 2)	-2.78	6.38E-06
Chromosome	187828	Tn5	*acnB*	aconitate hydratase 2 (citrate hydro-lyase 2)	-2.76	5.01E-06
Chromosome	3468515	Mu	*deaD*	ATP-dependent RNA helicase (dead-box protein)	-2.69	3.74E-05
Chromosome	1773061	Mu	*yciL*	hypothetical pseudouridine synthase	-2.53	6.47E-05
Chromosome	186784	Mu	*acnB*	aconitate hydratase 2 (citrate hydro-lyase 2)	-2.46	2.49E-05
Chromosome	3892359	Tn5	*selB*	selenocysteine-specific elongation factor	-2.26	0
Chromosome	2806154	Mu	*rluD*	FtsH suppressor protein SfhB	-2.19	0
Chromosome	3948421	Mu	*rph*	RNase PH	-1.76	0.006

3.7.2 KEGG pathway and Gene Ontology enrichment analysis

To further characterize the potential biological functions of each candidate RBP, the enriched genes set of each RIP-Seq sample were subjected to Gene Ontology (GO) (Ashburner *et al.*, 2000) and KEGG pathway (Kanehisa *et al.*, 2000 & 2012) enrichment analysis. The GO enrichment analysis was carried out for each of the ontology categories (biological processes, cellular components and molecular functions) and Fisher's exact test (Fisher *et al.*, 1922) *P* values were calculated using a background of the complete GO profile of *Salmonella* Typhimurium obtained from the UniProt database (Margrane & UniProt Consortium, 2011). Furthermore, enrichment scores were calculated by dividing the ratio of genes for each GO term in the enriched set by the ratio of the genes for the corresponding term in the background gene set. A maximum *P* value cut-off of 0.05 and minimum enrichment score cut-off of 2 was used for selecting the enriched GO terms. A similar enrichment analysis for KEGG pathways was carried out. Vornoi diagrams using Voronto (Santamaria & Pierre, 2012) were generated where the genes are mapped to the KEGG pathways according to their RIP-Seq enrichment level. The functional annotations obtained from these two approaches were used for hypothesizing the functional importance of RBP candidates and their potential regulatory roles in biological system.

The functional analysis of the target enrichment revealed several important regulatory pathways functionally associated with *Salmonella* infection and virulence for several of the candidate RBPs. A heatmap was generated to visualize the range of pathways that were enriched in the RIP-Seq samples (**Figure 16**). As shown, genes encoding the two well-known RBPs: Hfq and CsrA, enrich for genes such as SL1344_1030 and SL1344_1784 that are functionally associated with the *Salmonella* infection-related KEGG pathways. Two more genes from the same pathway, SL1344_2845 and SL1344_4300, showed high enrichment exclusively in Hfq libraries. These two genes were enriched in several RBP candidate CoIPs namely Rho, SL1344_1651, SL1344_2639, and several cold shock proteins (CspC, CspE, and CspH). These candidate RBPs, as well as CsrA, had enrichment of several other genes related to the infection related KEGG pathways, which are SL1344_1030, SL1344_1784, SL1344_2858, SL1344_2861, SL1344_2862, SL1344_2674, SL1344_2863 and SL1344_2864. Two pseudouridine synthetases, RluC and YjeQ also enriched the mRNAs encoded by these genes, whereas the CoIP of the pseudouridine synthase RluB showed instead enrichment of SL1344_1030, SL1344_2861, and SL1344_2674. Another RBP candidate, CaiD, exhibited high enrichment of SL1344_2863, SL1344_2864 and SL1344_1888 genes, which are also

functionally associated with *Salmonella* infection. One plasmid-encoded candidate RBP, SL1344_P1_0024 (TraR), shows an enrichment of the mRNA targets (SL1344_4213, SL1344_1784, SL1344_2863, SL1344_2864 and SL1344_2756 mRNAs) that are functionally associated with the infection related KEGG pathways.

A set of RBP candidates (CsrA, Rho, CspC, CspE, CspH, YjeQ, SL1344_1651, and SL1344_2639) enriched several genes (SL1344_2674, SL1344_1030, SL1344_1784, SL1344_2858, SL1344_2861, SL1344_2862, SL1344_2863, and SL1344_2864) are functionally associated to the bacterial invasion of epithelial cells. Several of those genes were also enriched in the Hfq libraries (SL1344_2845, SL1344_1030, SL1344_1784), the RluC library (SL1344_2674, SL1344_1030, SL1344_1784, SL1344_2858), the RluB library (SL1344_1030, SL1344_2861, SL1344_2674), the YjbC library (SL1344_2674, SL1344_1030), the CaiD library (SL1344_2863, SL1344_2864), and the TraR library (SL1344_2863, SL1344_2864).

Figure 3.16 ***Visualization of two pathways: bacterial infections and bacterial secretion system that are enriched in the RIP-Seq samples.***

*The **Figures 3.16A and 3.16B** (shown below) are heatmaps generated by Voronto tool to highlight the high (red spectrum) and low (blue spectrum) enrichment of the pathways in different RIP-Seq libraries (x-axis) based on the genes involved in these pathways (y-axis).*

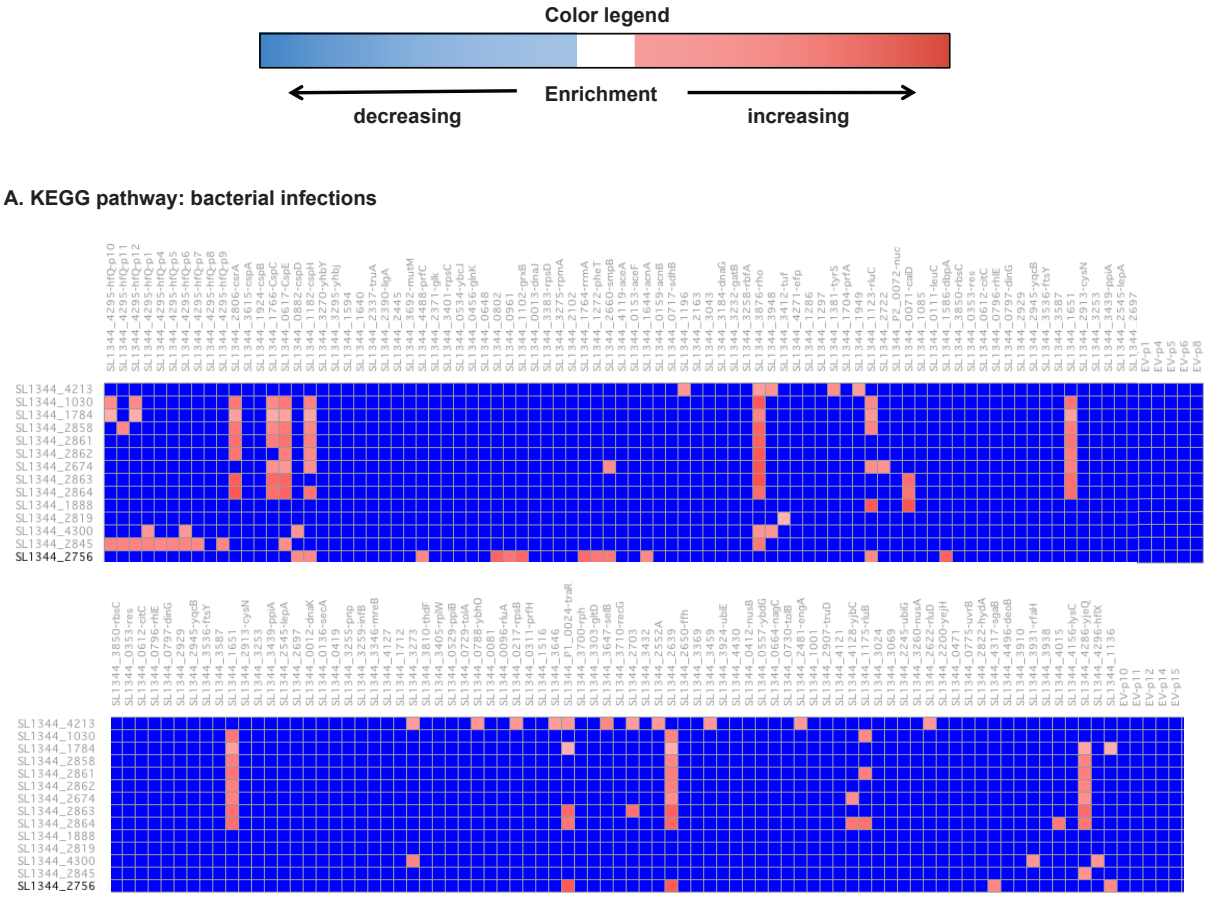


Figure 3.16A Heatmaps showing enrichment of KEGG pathways associated with bacterial infections corresponding to 14 genes (x-axis). *The y-axis corresponds to the various RIP-Seq libraries, where the last columns correspond to NT samples (denoted by EV) that does not enrich these pathways.*

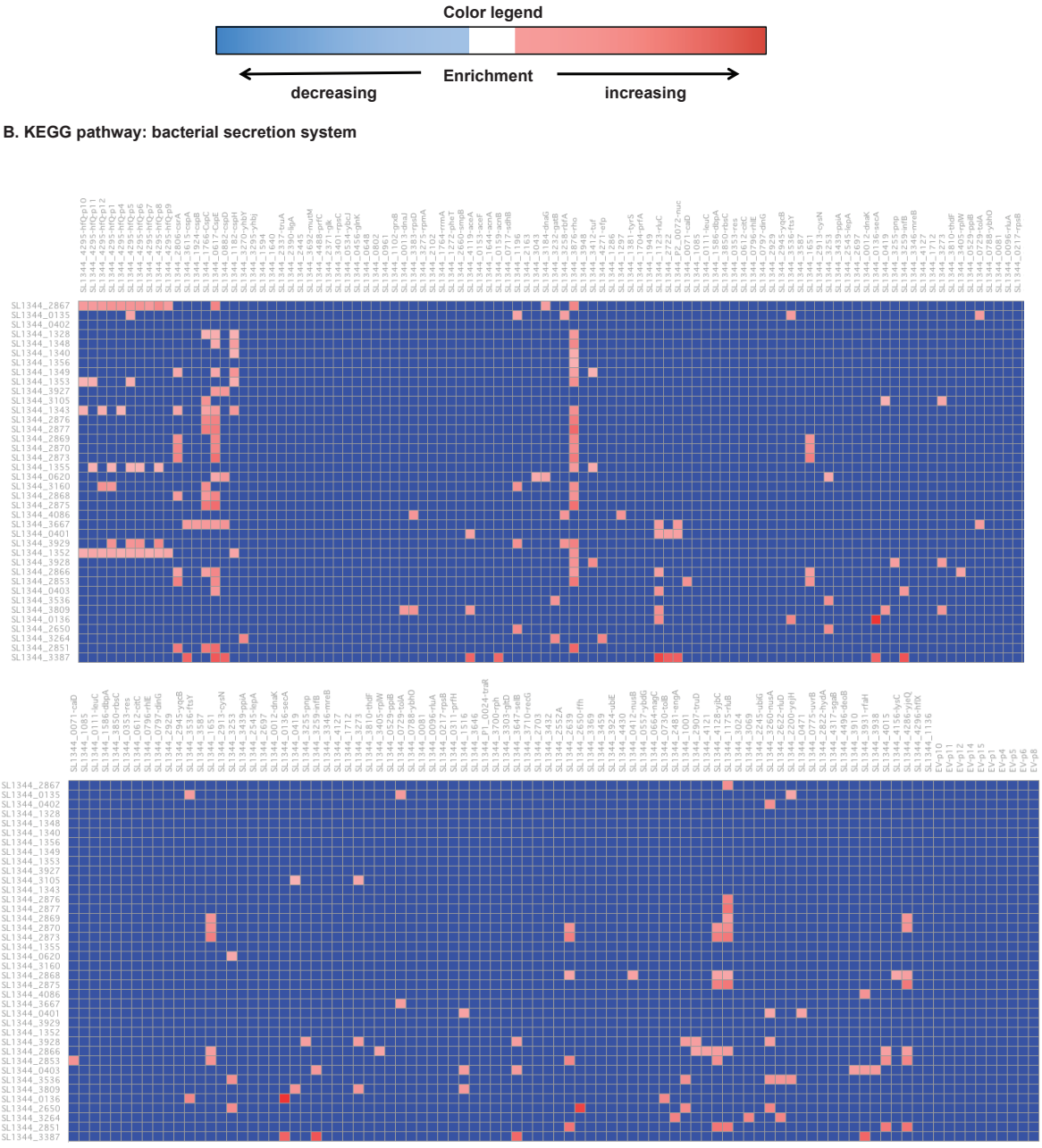


Figure 3.16B Heatmaps showing enrichment of KEGG pathways associated with bacterial secretion systems corresponding to 38 genes (x-axis). *The y-axis corresponds to the various RIP-Seq libraries, where the last columns in the lower heatmap correspond to NT.*

Several genes related to flagellar assembly were also enriched in the RIP-Seq data of several candidate RBP candidates (**Appendix Table 2**). RBPs (controls and candidates) that enriched 3-10 of these genes are Hfq, YhbJ, CspB, Rpsd, Sdhb, Nuc, DinG, Pnp, SL1344_3273, SelB, SL1344_4430, TolB, UvrB, and SL1344_0471. Another group of proteins comprised of

CspA, AceA, AceF, AcnB, Tuf, SL1344_2722, RhlE, TruD, and RluB, enriched more than 10 genes from this pathway.

Chemotaxis is also closely related with *Salmonella* interactions with the host, and some candidate RBPs enriched mRNAs related to this pathway (**Appendix Table 3**). More than 2 of those genes were enriched as targets in the RIP-Seq libraries of CspA, CspB, CspC, AcnB, SdhB, RluB, YejH, SL1344_1136, Tuf, EngA, and SL1344_3273. RBP candidates that enriched several genes from this pathway are SL1344_1196, SL1344_3948, SL1344_2722, RhlE, SelB, SL1344_4430, and SL1344_1001.

A more descriptive functional analysis of RBPs by means of GO and KEGG enrichment analysis will be discussed in the next section for the two cold shock proteins.

3.7.2 Functional characterization of CSPs as RBP candidates

Two cold-inducible proteins: CspA and CspB are characterized as RBPs in gram-negative bacteria including *Salmonella* (Phadtare *et al*, 1999). These CSPs bind to RNA by means of conserved RNP1 and RNP2 motifs, which are prevalent in eukaryotic Y-box proteins. Due to these conserved domains, CspA and CspB were used as positive controls in this RBP screening study whereas the remaining CSPs: CspC, CspD, CspE and CspH, were subjected to the RIP-Seq analysis. An analysis of these CSP RIP-Seq libraries highlighted the sample specific and common enrichment patterns among them. For example, PCA analysis of these libraries successfully clustered CspA and CspB together in one group, and CspC and CspE in another group (**Figure 3.17A**). In contrast, CspH clustered away from either of these clusters showing an enrichment pattern, which is specific for this particular CSP. CspC and CspE were were grouped with the positive RBP controls based on their expression profiles with ~400 putative RNA targets, strongly suggesting that they have *bona fide* RNA-binding activity.

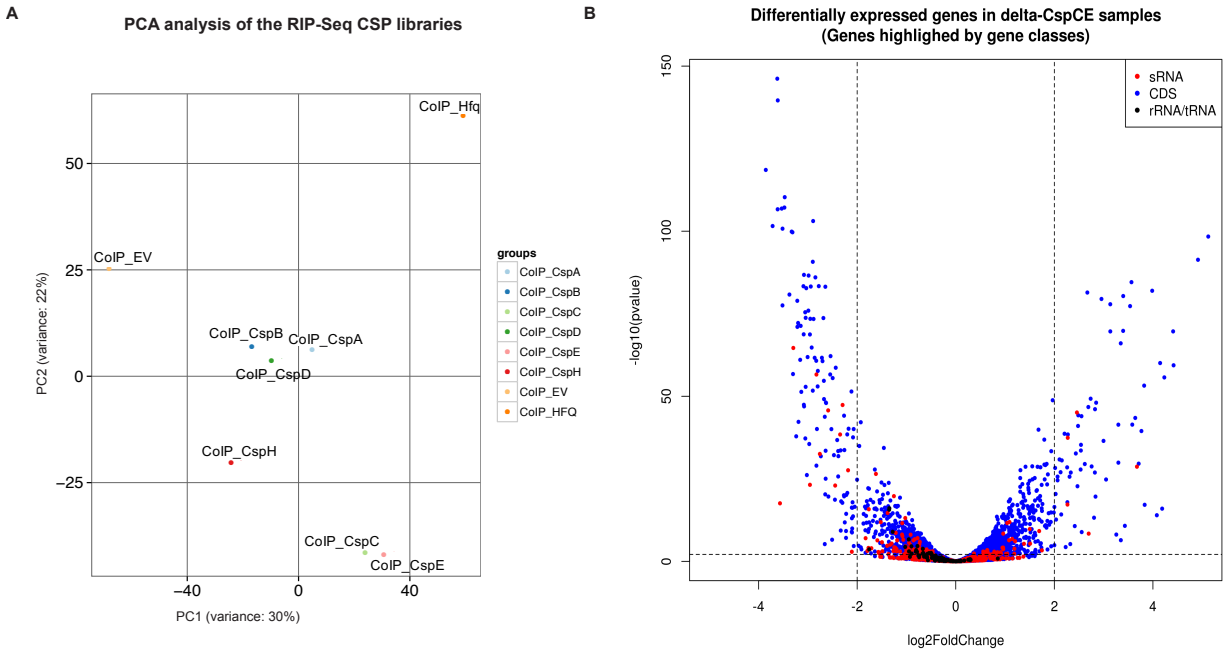


Figure 3.17 ***Overview of samples involved in RIP-Seq analysis of Cold Shock Proteins and full transcriptome analysis of ΔcspCE.***

(3.17 A) PCA analysis of the RIP-Seq libraries demonstrates the similar and dissimilar enrichment profiles of different CSPs. As shown, CspA and CspB cluster with CspH. Similarly, CspC and CspE cluster together. CspH shows different enrichment profile compared to other CSPs. In contrast, Hfq and the non-target samples do not cluster with any CSPs. (3.17B) Scatter plot showing various classes of differentially expressed (DE) genes in the ΔcspCE transcriptomics samples.

The next approach to functionally annotate these proteins included the analysis of the global transcriptome changes by sequencing a *Salmonella* strain deleted of both *cspC* and *cspE* (Δ*cspCE*) compared with the wild-type parental strain. This study was conducted on three biological replicates for each strain and the sequence data was computationally analyzed using the READemption RNA-Seq analysis pipeline (Förstner *et al.*, 2014). The differential expression (DE) analysis was carried out by *DESeq2* (Anders & Huber, 2010; Love *et al.*, 2014), which revealed 590 downregulated (log2 fold-change <-1) and 536 upregulated (log2 fold-change >1) genes in the Δ*cspCE* strain compared to WT. The set of regulated genes included primarily mRNAs and sRNAs (**Figure 3.17B**).

Figure 3.18 Functional analysis of RIP-Seq libraries of CSPs and ΔcspCE (knockdown samples) transcriptomics data sets.

(Figures 3.18 A-D are shown below)

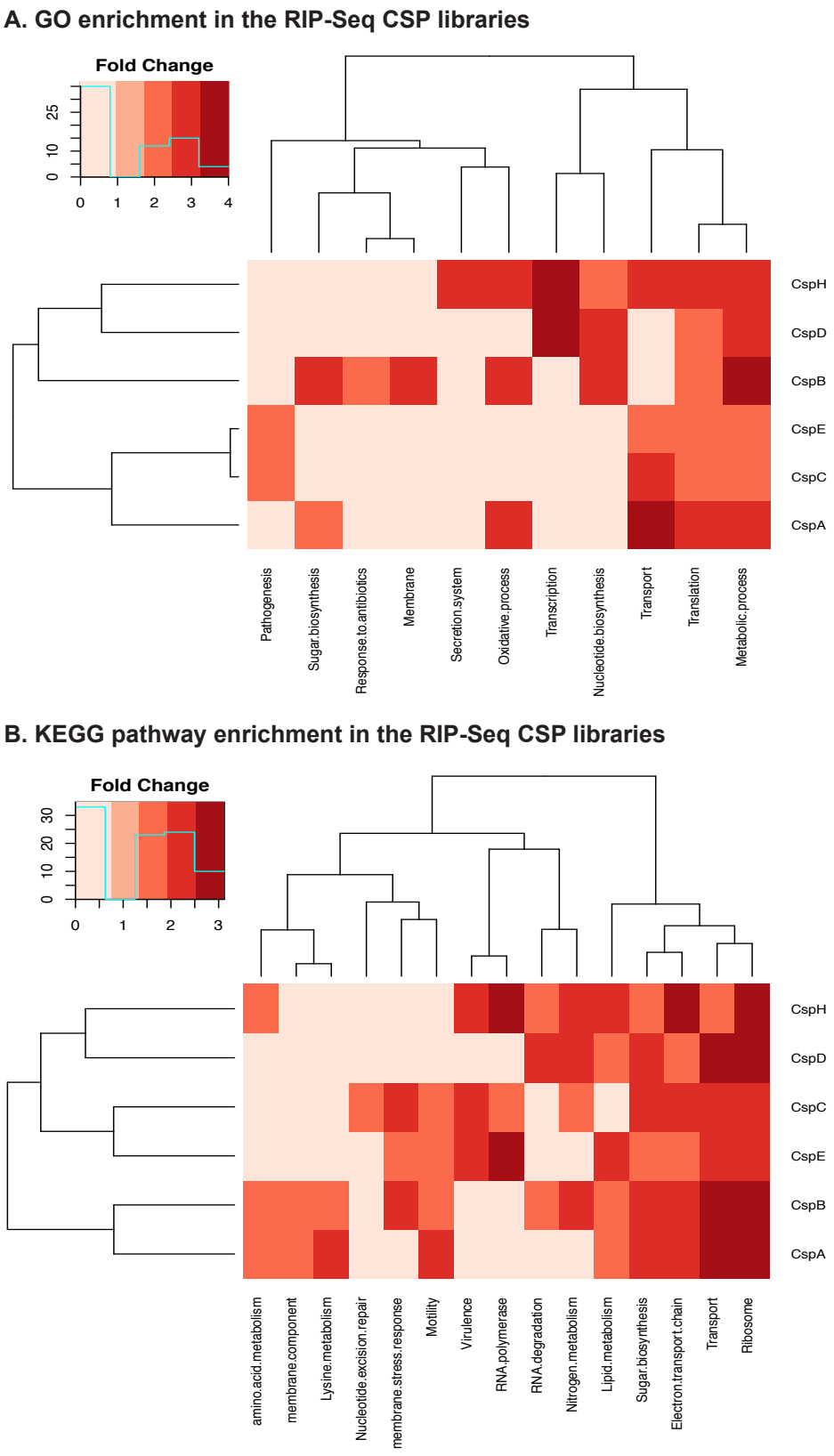


Figure 3.18A-B *Functional enrichment analysis carried out on the enriched genes in the different CSP libraries, which highlights the enrichment of GO terms (3.18A) and KEGG pathways (3.1BB) associated with infections in the CspC and CspE RIP-Seq libraries.*

C. GO enrichment in the ΔcspCE CSP libraries

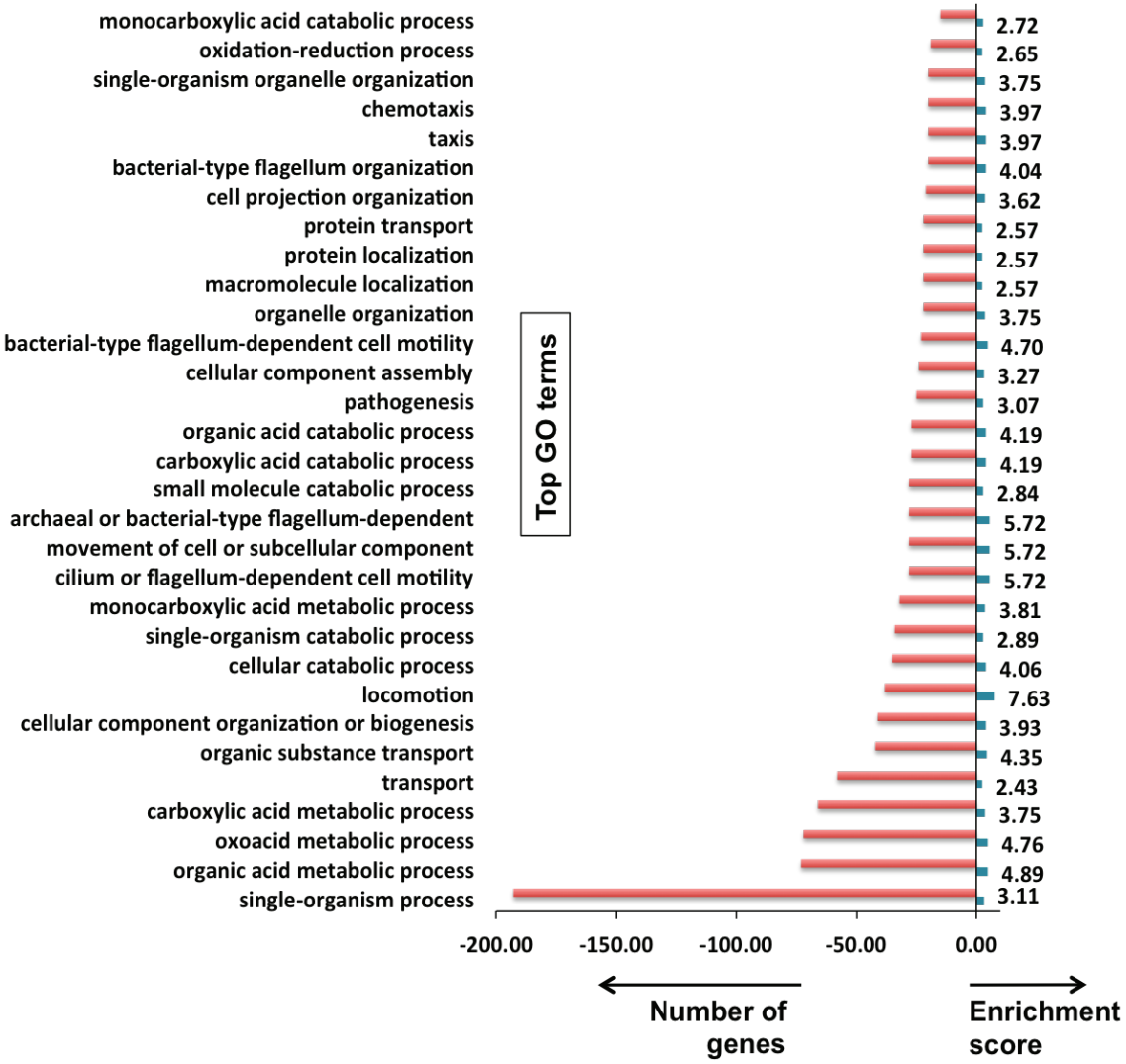


Figure 3.18C *Top GO terms obtained from the GO enrichment analysis of the genes that are differentially expressed (DE) in the ΔcspCE libraries. The red bar represents the number of genes and the blue bars represent their corresponding enrichment score (see materials and methods).*

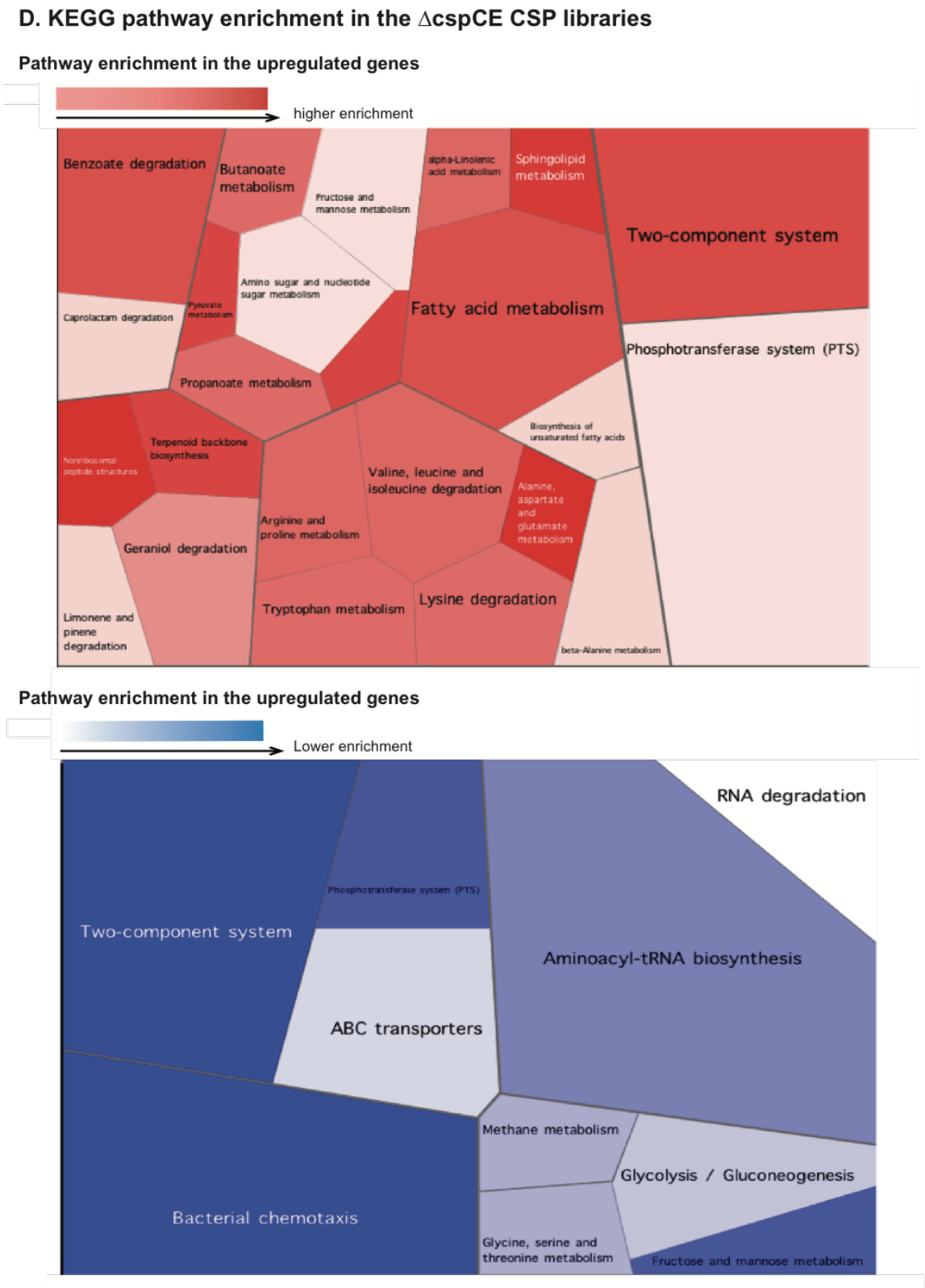


Figure 3.18D *Representation of enriched KEGG pathways among the DE genes.*

3.8 Concluding remarks

This study was designed to obtain a preliminary indication of the number proteins that can potentially bind to RNAs in *Salmonella* Typhimurium SL1344*,* and further characterize them for their biological roles. The primary screening conducted by APRICOT protein characterization pipeline indicated that about 10% proteins in the complete proteome contain functional domains that could potentially have binding affinity towards RNAs. The experimental set-up for the validation of several of these candidates was carried out by means of RIP-Seq-based high-throughput sequencing approach to account for the RNA binding partners of RBPs. The analysis of these sequencing data showed that many of these candidates could interact with one or multiple RNA classes.

RIP-Seq is labour intensive and costly procedure when performed on a large number of proteins. The problem of dealing with a large sample set was addressed by limiting the number of RBP candidates for the validation study by only selecting predicted RBPs that have important regulatory or infection relevant roles as per the existing literature based annotations. Hence, for the RIP-Seq based validation study, a set of 131 RBPs was selected by exploring their biological characteristics available in publicly-available experimental data sets for *Salmonella*, including SalCom, a *Salmonella*-HeLa cell dual RNA-Seq data set, and TraDIS infection data set (**Appendix Table 1**).

Another technical issue with this approach is the differentiation between primary targets and unspecific secondary binders. This could be addressed partially by using sample replicates to account for only those RNAs that consistently bind with RBPs. However, a replicated study on 131 proteins could still be cost-intensive, and abundantly expressed secondary binders might still be falsely counted as RBP targets. In order to address this, we carried out a pooled exploratory analysis of the entire RIP-Seq data set generated in this study by introducing a suitable normalization approach. The normalization in this study uses a modified TMM, which was very effective at excluding transcripts that are abundant in all the samples and appear as false positive RBP binders. Furthermore, a stringent target selection criterion was established by using a cut-off derived from quartile approach applied on the entire RIP-Seq samples. Based on an intensive analysis of the 131 test samples, only 13 RBP candidates were indicated as non-RBPs as they exhibit enrichment patterns similar to the NT controls that do not bind to RNAs. Furthermore, by integrating knowledge from different resources such as GO, KEGG, and TraDIS data set, initial biological characterization

of some candidate RBPs was carried out, which could be further classified based on their functional relevance.

In the related studies for the target validation and functional characterization of CspC and CspE CLIP-Seq experiment was carried out in our lab (Dr. Charlotte Michaux) that showed a high overlap with their RIP-Seq expression profiles. These enriched RNAs, from both RIP-Seq and CLIP-Seq, were subjected to GO and KEGG enrichment analysis, which revealed their potential roles in pleiotropic stress response-related biological processes (data not shown). Furthermore, by means of complete transcriptome analysis of a double deletion strain (Δ*cspCE)* compared to WT, it was suggested that these proteins might have complementary roles in these phenotypes. Additional phenotypic experimental studies further characterized these proteins for their regulatory roles in virulence of *Salmonella* (data not shown). This provided a proof-of-principle that RIP-Seq based screening of candidate RBPs in *Salmonella* could participate in important regulatory roles.

As one of the first exploratory bio-computational analysis for the identification of RBPs in model bacterial pathogen, this study can serve as a valuable resource for the scientific community. In addition to providing a resource of RBP candidates in *Salmonella*, but also specifies a step-wise framework for the characterization of their biological function.

Chapter 4

Discussion

4.1 The APRICOT computational pipeline: potential applications and scope

It has been well established that RBPs play several important roles in global regulatory networks (Moore & Proudfoor, 2009; Wang *et al.*, 2015; de Klerk *et al.*, 2015). Several RBPs have been documented and characterized in eukaryotes, using cross-linking and mass-spectrometry based methods (Castello *et al.,* 2012; Baltz *et al.*, 2012; Kwon *et al.*, 2014). Furthermore, by using co-immunoprecipitation experiments, the binding partners of several of these RBPs have been identified, which provide insights into the functional roles of these proteins (König *et al.*, 2012). Such global studies by means of wet-lab experiments have been developed and reproduced in eukaryotic model organisms like human, mouse and yeast (Castello *et al.,* 2012; Baltz *et al.*, 2012; Mitchell *et al.*, 2013). Though these approaches are efficient, they are extremely tedious, time consuming and costly. Moreover, they are not readily applicable to other systems without extensive modification of the techniques. For example, the cross-linking based interactome capture depends on the pull-down of eukaryotic mRNAs via the poly(A) tails, which are absent from most bacterial mRNAs. Other deep-sequencing methods like CLIP and CoIP followed by high-throughput sequencing (RIP-seq) can be conducted on a single protein at a time, in order to capture their genome-wide binding partners and characterize their binding specificity. It has been proposed from studies of the human genome that about 10% of the entire proteome has the potential to bind RNAs (Gerstberger *et al.*, 2014); hence carrying out such experiments on the complete proteome of other organisms might be unfeasible and/or inefficient. As listed already (Chapter 1), bioinformatic tools provide an important alternative to reduce the list of candidate RBPs and decrease the number of false positive results. Most bio-computational approaches for RBP identification are established on the information obtained from experimentally determined structural information of RNA-protein complexes. Several tools have been developed independently of such RNA-protein structures by instead involving bio-computationally derived information from all eukaryotic RBP sequences and identifying proteins comprising similar physico-chemical properties to those of known RBPs. However, these methods are limited in their application to specific data types due to manually curated training sets.

Hence, these tools are prone to miss out RBPs that do not qualify as such due to the underlying reference data used in their development (Miao & Wetshoff, 2015).

I set out to develop an approach that is established independent of a specific set of proteins to avoid any bias caused by such underlying data set. Furthermore, I applied this approach to present an exploratory analysis of RBPs in bacterial proteomes in a large-scale manner. A generic software pipeline (APRICOT) was designed that could be customised to analyse the proteome of an organism to identify proteins with biological roles of interest, based on a given set of functional domains. To specifically identify RBPs in bacterial proteome, a set of RBDs was used, which are annotated in domain databases for their RNA relevant structural and regulatory roles.

Since bacterial proteomes differ from eukaryotic proteomes due to phylogenetic divergence, there are obvious challenges when applying knowledge obtained from eukaryotic data sets for the RBP predictions in bacteria. To account for such differences, a wide variety of reference features were selected to capture both the structural information and physico-chemical properties. One such feature is the information of subcellular localization of proteins that can help tremendously in annotating newly sequenced genomes, designing experimental studies, and identifying drug targets and biomarkers. The eukaryotic experimental set-up for RBP characterization by interactome capture lacks subcellular spatial information, which has been partially addressed by the SerIC experimental approach (Conrad *et al.*, 2015). However, it is not feasible to derive localization information for every protein in a cellular system in a single experiment. Computational approaches provide a faster option to predict localization signals coded in the protein sequences to give insight into their subcellular location. Several tools have been trained on large data sets to capture such information in bacterial proteins. To allow users to classify their set of query proteins or putative RBPs, a highly effective tool - PSORTb (Yu *et al.*, 2010) has been integrated in the APRICOT pipeline. PSORTb software have been trained on bacterial and eukaryotic data sets to predict subcellular localization signals in the protein sequence. This added feature of APRICOT provides users not only the possibility to identify proteins with domains of interest, but further identify their possible roles in biological network by placing them into the cellular context. The feature of secondary structure by searching homology for tertiary structures provide further insight into the possible substrate specificity of proteins. Such information, in the context of putative RBPs, further allows the modelling of a protein-RNA complex computationally, which can then be experimentally

verified. In order to proceed further with the biological understanding of predicted RBPs, APRICOT has also been provided with a tool for the prediction of secondary structures by means of RaptorX (Källberg *et al.*, 2012), and tertiary structure homology search against the PDB database (Kouranov *et al.*, 2006). Furthermore, using Gene Ontology terms, which associate gene-products to specific biological processes, cellular components, and molecular functions, proteins of interest can be clustered to derive their relatedness in biological systems (GO consortium 2009). To enable this feature for putative RBPs, APRICOT provides comprehensive Gene Ontology information, which is compiled from the UniProt knowledgebase (Magrane M & UniProt Consortium, 2011), InterPro domain resources (Mitchell *et al.*, 2015), and additional calculation by blast2go (Conesa & Götz, 2008). An analysis step was included in the pipeline to calculate the distance between the domains identified in the query proteins and their corresponding reference domain in terms of these functional properties.

The number and functional understanding of proteins in the databases is increasing exponentially due to the availability of genome sequences, as well as the technical advances in experimental studies brought about by high-throughput techniques like mass-spectrometry and RNA-sequencing. One of the major concerns resulted by this rapid increase in newly identified proteins is how these new ORF candidates will be accurately and descriptively annotated. Community-driven projects like Critical Assessment of Structure Prediction (CASP) (Moult *et al.*, 2014) and Critical Assessment of Function Annotation (CAFA) (Radivojac *et al.*, 2013) are intended to annotate these proteins with structural and functional information and further improve the quality of annotation with the help of the experimental scientific community. In this endeavour, our pipeline was included in the CAFA project (Jiang *et al.*, 2016), where it performed a complete annotation of more than 1 million bacterial proteins with gene ontology terms. This inspired the modularity of the pipeline, which makes it adaptable for the annotation of different data sets with practically any functional class. This modularity allows users to introduce modifications required for their analysis, like the type of query input (protein name, synonymous IDs, FASTA sequence, or taxonomy IDs), choice of reference domains (terms for domain selection), and reusability of intermediate data. Several options are available for defining filtering criteria such as selection of reference domains (one or several domain families), extent of similarity between the reference and predicted domains, similarity of their physico-chemical

properties, and additional structural similarities. The tool has been extensively tested on different sets, and improved for its ability to annotate proteins.

The applicability of this approach for RBP identification was demonstrated in several data sets composed of proteomes of varying sizes and phylogenetic source. The pipeline successfully processed these data sets, showing high specificity and sensitivity for known RBPs. A high accuracy was further observed for APRICOT predicted functional sites, which coincide with the binding residues known from the PDB structures. These observations could be reproduced when the complete *E. coli* and human proteomes were analysed. The pipeline could identify all known RBPs and suggested many additional putative RBPs. This capability of the pipeline successfully addresses the issue of dealing with large data sets, especially for the computational processing of complete proteome sets from different organisms.

Since most of the bioinformatics tools depend on a number of other software packages, one of the common concerns is to maintain their consistency and reproducibility. I developed APRICOT as a command-line tool, which is publicly available on PyPI (Python Package Index) and GitHub. To make it feasible for the users to use the command-line tool efficiently, I have generated and intensively tested a Docker image, which is available on the Docker hub. Users can carry out a frictionless installation and usage of the program along with the associated software and data dependencies, allowing users to conveniently use APRICOT on different platforms. A comprehensive documentation has been provided that includes installation requirements, instructions, and example data sets to test the software effortlessly. Extensive resources, including video tutorials, have been developed in order to present the concepts behind APRICOT, as well as guide the beginner users through hands-on sessions. In near future, the software will be extended with a functional and sustainable web server for its public accessibility.

As apparent from the rich collection of data sets and integration of various computational approaches in the pipeline, APRICOT provides an ideal scenario for the analysis of large sets of query proteins in an automated manner. Using statistically derived high importance parameters and their respective cut-offs, a clear distinction is provided between high confidence domain prediction and the domains that are predicted by chance. The Bayesian scoring offers an easy solution to the users to rank their sets of proteins predicted with domains of interest by their biological relevance.

4.2 Resource for bio-computationally characterized RBPs in bacteria

One of the main focuses of biological research has been on the development of methods for the identification and characterization of human proteins, including RBPs. Among bacteria, many model organisms have been completely sequenced but a much lower amount of proteins have been intensively characterized, which includes ~10 RBPs in the model enterobacteria *Salmonella* and *E. coli*.

One major highlight of this thesis is the exploratory analysis of the complete proteome set of *Salmonella* Typhimurium SL1344 in order to identify RBPs. Using the APRICOT computational pipeline, 1068 RBPs were predicted, of which 372 proteins were annotated as such by both the main components of the tools, CDD (Marcher-Bauer *et al.*, 2015) and InterPro (Mitchell *et al.*, 2015). By further introducing the criteria of filtering by domains that are annotated in both eukaryotes and bacteria, a set of 131 proteins was selected for analysis by RIP-Seq experiments (**Appendix Table 1**). The experimental set-up chosen for the validation study, RIP-Seq, is a widely-used technique in RNA biology research for the co-purification followed by high-throughput sequencing identification of RNA targets of proteins of interest (Ule *et al.*, 2003; Selth *et al.*, 2009; Zhao *et al.*, 2010; König *et al.*, 2012).

One major challenge of this approach is to differentiate primary binding partners of RBPs from the secondary or indirect targets (Selth *et al.*, 2009). In an initial setup, the analysis of the RIP-Seq sample of each protein was carried out against a control of RIP-Seq data set generated for bacterial samples transformed with an empty plasmid, named non-targeting (NT) samples. Interestingly, the 10 different NT samples had different expression profiles. It was intuitively expected that if no protein is overexpressed there should not be an enrichment of targets. However, the NT samples showed low enrichment of a few (<15 targets) RNA targets when compared with other NT control samples. We used samples with the overexpression of Hfq as the positive controls, which as expected showed enrichment of several RNA targets. The 10 Hfq samples, which were sequenced in different sequencing pools, had several common targets including various known mRNA and sRNA genes. However, they did not show a completely identical enrichment profile, possibly introduced by the library preparation or the sequencing process. Therefore, it was crucial to introduce a normalization method that could exclude targets that appear in each sample non-specifically or low confidence targets that could occur due to low number of transcripts. Based on the modified-TMM normalization analysis, the targets that are enriched in NT samples and

consistently appear in all the samples were excluded. In addition, a minimum transcript count of 50 reads to select a target was applied. Upon analysis, 118 of the 131 candidate proteins were identified as putative RBPs as they enriched sample specific targets, whereas 15 samples had low number of targets and exhibited an enrichment profile similar to the NT samples.

The more target specific technique CLIP-Seq helps in identifying the direct binders and the binding motifs that are specific to the majority of RNA targets, hence other targets that do not bind to such specific sites on proteins can be regarded as secondary or low confidence binders. In the related studies in our lab, Dr. Charlotte Michaux carried out more specific experimental studies for the target validation and functional characterization of positive RBPs: CspC and CspE using CLIP-Seq. RNA targets of the individual proteins were enriched in CLIP-Seq data sets, which showed a high overlap with their RIP-Seq expression profiles. These enriched RNAs, from both RIP-Seq and CLIP-Seq were subjected to GO and KEGG enrichment analysis that revealed their potential roles in pleiotropic stress response-related biological processes. Furthermore, by means of complete transcriptome analysis of a double deletion strain (Δ*cspCE)* compared to WT, it was suggested that these proteins might have complimentary roles in these phenotypes. Additional phenotypic experimental studies further characterized these proteins for their regulatory roles in virulence of *Salmonella*. This provided a proof-of-principle that RIP-Seq based screening of candidate RBPs in *Salmonella* could predict important regulatory roles.

The large amount of sequencing data generated in this work serves as a substantial resource to begin to understand the biological roles of these candidate proteins. In addition to provide a computational pipeline, this study also explores the success rate of its prediction by means of experimental validation studies, as well as an experimental data set of potential RNA binding partners for multiple candidate RBPs. However, this exploratory study of RBPs also highlighted technical challenges associated with large scale bio-computational screening of protein sets, which is discussed in detail in the next section. In summary, this study is the first step in the direction of comprehensive analysis of regulatory RBPs in bacteria and constitutes an important resource to guide high-resolution characterization studies.

4.3 Limitations

The analyses described above clearly demonstrate the efficiency of APRICOT for prediction of RBPs in bacteria, nevertheless, there are limitations to this pipeline. One obvious shortcoming is that since the pipeline has not been designed to identify new domains in the protein sequences, the software largely depends on the availability of domains in the databases. A second limitation relates to the correct selection of reference domain sets, which is influenced by the user-provided terms for the compilation of domains from databases. The MeSH terms, Pfam domain identifiers, protein family names, and generic biological terms like gene ontology can be used to denote a functional class and select domains accordingly. Since the domain selection takes place by searching a string in the annotation of domain entries, any use of ambiguous terms may lead to the selection of functionally irrelevant domains. Similarly, by using an extremely specific term, one can influence the accuracy of the reference set by limiting the domain search space to only few domains that are overtly annotated with the specified terms. For example, a large number of ribosomal proteins and domains are defined in human proteomes. However, not all of them are RNA-binding. Hence, instead of using 'ribosomal' as the term to indicate these domains, the specific RNA-binding ribosomal domains (Gerstberger *et al.*, 2015) were defined to exclude undesirable non-RNA-binding domains. Similarly, only 71% of the human RBPs could be identified using well-annotated domains like classical RBDs and RNA-binding ribosomal domains. However, the remaining RBPs could be identified by further including non-classical RBDs.

The current version of APRICOT tool is not designed to find new domains, hence the limitation of the software in dealing with only the known domains is unavoidable. However, by establishing a verified set of keywords for the important functional classes, the second limitation of term selection can be resolved. Scientific community involved in different fields of protein research can contribute to the development and improvement of this kind of consistent sets of terms or domain families related to the specific protein classes.

Other limitation relates to the validation study by RIP-Seq experiments. The main challenge is to identify and deal with the biases, such as batch effect introduced by technical handling of the data sets. The RIP-Seq samples used in this study were generated by the overexpression of a protein of interest by means of plasmids (see materials and methods). This experimental setup could have non-desirable biological effect on the sample conditions

(Selth *et al.*, 2009). For example, it was observed that NT samples, which was sequenced in 10 different sequencing pools, showed enrichment of a few genes (<15 targets with low transcript abundance) even though it was expected to show no target enrichment in the absence of protein overexpression. Additionally, Hfq samples were also sequenced in 10 sequencing pools as positive controls, which showed enrichment of much higher number of targets than previously reported (Chao *et al.*, 2010; Holmqvist *et al.*, 2016). These observations highlighted the enrichment of non-specific targets in the samples, which could not be avoided even by using replicated sample sets. Beside these two sets of samples (Hfq and NT), other samples were not produced in replicates due the associated cost- and labour-intensive experiments. Hence, for these samples it could not be directly verified if the targets are enriched as a result of expressed proteins or due to unspecific secondary bindings to the primary targets. Furthermore, based on their transcript abundance, a few genes were found enriched consistently across each sample.

Such unspecific secondary binders can introduce inconsistency in the quantification of target enrichment across different samples including replicated samples of Hfq and NT samples. An experimental setup where a background sample is generated for each RIP-Seq sample could have solved this issue by allowing a direct comparison of both the samples and identify the primary targets even in the absence of replicates. However, for the available set of samples that lack such background controls, the quality of target selection was optimized by integrating a normalization process by TMM method (Robinson *et al.*, 2010). The standard TMM requires a reference or control sample for the selection of reference gene set to calculate size factors for each sample, which is further used for normalizing them. However, due to a lack of an appropriate reference, a modification was introduced in this method (suggested by Dr. Lars Barqist) that globally compared the enrichment profiles of the samples and selected reference gene set for the size factor calculation that have more than 10 reads in each sample. Upon normalization of the samples, the target genes were recorded for each sample. This modified TMM not only in excluded the genes that initially appeared as targets due to their enrichment in each sample non-specifically, but also highlight the sample specific targets efficiently, which could be verified for the positive controls.

4.4 Future applications

4.4.1 Characterization of other functional classes

The APRICOT pipeline has been designed in such a way that it can be easily adapted for the characterization of functional protein classes other than RNA-binding. For each query protein, all possible domains are predicted. By further defining a comprehensive non-ambiguous set of terms that indicates certain functional properties of a protein class, domain entries are collected from the data sets. This reference set of domains is used for the selection of proteins in which these domains are predicted. Users can choose different properties to identify proteins of interest without re-running the domain prediction analysis. The default parameters have been defined based on the data sets related to RBPs and non-RBPs. These parameters worked equally well for the characterization of kinases. However, for different functional classes SVM based parameter ranking followed by optimal cut-off range can be calculated in a similar manner as done for RBPs. In this regard, it is also important to establish the domain sets or the search term that selects only relevant domains of interest.

4.4.2 Identification of dual-specificity DNA-RNA binders

Most of the proteins are multifunctional, as they may participate in different biological processes. Using serial interactome capture (SerIC) technique, Conrad *et al.* (2015) identified a group of proteins with dual binding specificity to RNA and DNA in their study of nuclear interactome. A standard IC approach involves UV induced cross-linking of proteins with their RNA targets, which are isolated by poly(A)-RNA to oligo d(T) magnetic beads and stringently washed before subjecting them to LC-MS/MS detection. A considerably high amount of DNA was observed in the IC purification of nuclear fraction, hence they included another round of chemical and enzymatic treatment, oligo d(T) capture, stringent washes followed by LC-MS/MS detection. Even with these extreme measures to avoid DNA-binding protein recovery, 80 proteins were recovered as DNA-RNA-binding proteins (DRBPs) in the nuclear transcriptome that were previously annotated as DNA-binding proteins. Sixty of these proteins have RBDs or experimentally verified RNA targets. The remaining proteins were recognized as novel DRBPs, a few of which are transcription factors or with roles in RNA-splicing, processing, stability and DNA damage response. The examples of these novel DRBPs are, BCLAF and THRAP3, components of SNARP complex, which regulates the stability of cyclinD1 and also couples DDR and alternative splicing (Conrad *et al.*, 2015). Several DRBPs

couple transcription and RNA splicing, for example PHF5A, which activates CX43 and as a part of splicing factor 3b protein complex, it couples splicing factor and DNA helicase. These examples including a few other DRBPs such as kinases, ZF domain containing proteins and replication-dependent linker histones, which give a new insight into the multifunctional proteins.

These observations indicate that this dual specificity can allow proteins to have multifunctional ability to act as regulators in related or unrelated cellular events. This less understood set of interesting proteins, can be explored by means of computational characterization techniques in more systematic manner. APRICOT could be extended to allow users to identify proteins with dual-functionality as abovementioned DRBPs.

4.4.3 Identification of domain co-occurrence

Systems biology aims to capture the interaction of individual components like DNA, RNA, proteins and small molecules, in a biological environment like cellular system. This approach is different from a reductionist way of studying one target at a time, where the main focus is to build a biological network to understand the structural and functional dynamics of the whole system. As mentioned earlier, most of the proteins are multifunctional in nature and respond differently according to the environmental conditions. Since most of the functional aspects of a protein can be captured by its domain architecture, it is possible to computationally identify the co-occurrence of two or more domains and derive their possible functions (Wang *et al.*, 2011). APRICOT in its primary analysis identifies all the possible domains in a query; hence, it could be programmed to capture the co-occurrences of domains defined by users. This feature can allow users to go away from reductionist model of identifying one-to-one interaction and identify protein interaction networks, which can give insights into the possible functions and interactions of proteins of interest in proteome level.

4.4.4 Subcellular spatial resolution of RBPs

The nuclear and cytoplasmic interactome study by SerIC (Conrad *et al.*, 2015) highlighted the different repertoire of RBPs, which could be annotated by their subcellular spatial resolution. In APRICOT, we use PSORTb (Yu *et al.*, 2010) to allow users to classify putative RBPs based on their putative subcellular localization in the downstream analysis. This feature of the pipeline could be further developed by analysing a large protein set to derive

a possible functional correlation of known RBPs with the computationally predicted subcellular localization. Such correlation can be used as reference to characterize the localization information of novel or non-annotated RBPs in future.

4.4.5 Using intrinsically disordered regions of RBPs

Interactome capture of eukaryotic RBPs has revealed that while most RBPs have RBDs, a large number of proteins with RNA-binding activity do not possess a canonical RBD but instead acquire their ability to bind to RNAs by evolution of low complexity and disordered amino acid sequences. As discussed earlier, these proteins, known as intrinsically disordered proteins (IDPs), have equally important regulatory functions as RBPs and are also likely underrepresented in our catalogue of putative RBPs (Castello *et al.*, 2012; 2016). A subclass of IDPs comprise a low complexity region of 1-10 amino acids long, which provides structural profiles to the IDPs. In order to identify such IDPs computationally, it is important to use compositional information of amino acid sequences that could be connected to the low complexity region, structural properties and physico-chemical features. So far, APRICOT uses canonical domains as a basis for the identification of RBPs, however sequence composition of the known IDPs could be generated and used as reference for the characterization query proteins that lack RBDs.

4.4.6 High-throughput sequencing-analysis based characterization of RBPs

The processing and analysis of RIP-Seq libraries was carried out by a standardized computational approach. Users can take the analysis framework discussed in this study for *Salmonella* Typhimurium as a reference and use the computational method established for this study. This high-throughput sequencing-based characterization would involve the initial mapping of the transcripts to the reference genome, gene-wise quantification of the reads for each library followed by a reference-free normalization of quantified reads by abovementioned modified TMM normalization approach. This normalization further allows identifying sets of libraries that show relatively higher enrichment of certain targets and avoids the inclusion ambiguous transcripts as targets that appear in several libraries. The efficiency of this approach is evident in the data sets used in this study; hence, by extending APRICOT pipeline, a streamlined analysis of RIP-Seq based characterization of candidate RBPs can be allowed. This will give a standard platform for the identification and experimental characterization of proteins of interest.

Furthermore, as described for the two positive RBPs CspC and CspE, more specific studies such as CLIP-Seq-based target validation and experimentally based phenotypic characterization can be further used to understand the regulatory roles.

4.5 Conclusions and perspective

Current research in biological science is inseparable from bioinformatic approaches and computational analysis. Developments in the techniques for high-throughput sequencing have accelerated the growth of bio-computational methods allowing both the fields to grow in a symbiotic manner. Applying this principle to the discovery of RBPs is not a new concept. Several tools have been developed using the existing knowledge of RBP sequences and structures. Advances in the proteome-wide discovery of RBPs in the human genome have encouraged the expansion of such research to bacterial organisms, in order to understand the underlying regulatory roles that drive their adaptation and survival under adverse conditions. With the increasing number of newly sequenced genomes, it is feasible to multiply-align genomic locations followed by identification of functionally conserved motif sequences. Integration of such motif information of proteins with structural and sequence-based properties allows a better discovery of protein functional classes.

The computational pipeline of APRICOT has been developed for high-throughput screening of RBPs. This software pipeline provides a convenient approach to process large number of queries and annotate them by highly conserved RNA-binding motifs and their physico-chemical properties. Though optimized largely on RBP data sets, the tool has been tested on other functional groups of proteins to verify the multifunctional annotation capacity of the pipeline.

As shown for human, *E.coli* and *Salmonella* proteomes, the framework of this study can be used as a template to carry out similar studies in other proteomes. An informed selection of candidates was carried out for their experiments validation as it was more practical to narrow down the long list of computationally identified RBPs to a smaller subset by using different biological factors such as conservation of sequence and physico-chemical properties. The regulatory elements of RBPs taken in this study are functionally conserved domains; however further new classes of regulatory elements could be included in future. For example, the intrinsically disordered regions, which are much smaller in size compared

to RBDs, could form another layer of information. Studies are being carried out to understand and characterize these linear motifs in RNA binding context and further discovery of these underlying patterns across larger data sets can be carried out using advanced machine learning approaches. Such information will boost RBP discovery by revealing new unexplored classes of protein-RNA interactions. Furthermore, developments in new experimental technology by cross-linking and more efficient sequencing methods can improve the accuracy of the pipeline by providing a complete picture of RBP catalogue in a proteome of interest.

Chapter 5

Materials and Methods

Programming languages

APRICOT software has been mainly developed using the object-oriented Python programming language. I have used Linux to automate the software installation and analysis. To execute specific statistical anaysis and visualization of RNA-Seq data, different packages of R programming language were also used.

Git was used for version control, storage, and sharing of the codes via GitHub. Dockerfile has been developed to generate image for the containerization of the software. The Docker images (malvikasharan/apricot and malvikasharan/apricot_with_dependencies) are hosted on the Docker hub (https://hub.docker.com/malvikasharan).

Availability of APRICOT software

The APRICOT software, its documentations, and links to various data sets and tutorials are available on the home page of the software: https://malvikasharan.github.io/APRICOT/. Additionally, the software package has been submitted to Python Package Index (PyPI), which can be downloaded from https://pypi.python.org/pypi/bio-apricot.

To make the analysis using the software easy for both computer experts and non-experts and improve reproducibility of results, the Docker images have been created and intensively tested, and are hosted on the Docker hub (https://hub.docker.com/r/malvikasharan/apricot/). Docker is a software containerization platform, which allows packaging of complete file system of software: codes, package dependencies, system tools, system libraries, applications etc., which runs independently of environment.

I have created video tutorials that give an overview of the software and demonstrate various ways to install software and execute analysis. The videos are available online for which the links are provided in the home page of the software.

Installation of APRICOT software

The software can be installed in a local system via pip, git, or Docker, which is discussed below in detail.

1) Installation via pip: The software package can be installed from the Python Package Index using a requirement specified using the command `pip install bio-apricot`.

This will not only install the software but also other Python packages, matplotlib, numpy, scipy, openpyxl and requests, which are required to use APRICOT. This installation allows user to test the functionality of the software. However, to carry out a functional analysis other dependencies such as tools and databases are required, which are listed online in the following document: https://github.com/malvikasharan/APRICOT/blob/master/software_dependencies.md.

2) Installation via git: The current version of software is submitted to the GitHub repository that can be downloaded using the command `git clone https://github.com/malvikasharan/APRICOT`. Additionally, all the aforementioned Python libraries and dependencies should be installed in order execute the software.

3) Docker image: Docker image of the APRICOT software allows the packaging of the software along with all its dependencies, which ensures the execution of the tool on any platform reproducibly and without any error. The Docker should be installed locally whereupon the docker image of APRICOT can be fetched from the Docker hub using the command `docker pull malvikasharan/apricot`.

The tutorials (documents and video) are available, which can guide novice users through the installation, example analysis and real case analysis. The links to all the materials are available of the GitHub website: http://malvikasharan.github.io/APRICOT/.

Domain databases

APRICOT uses an extensive set of protein domain databases to ensure a larger search space for reference signature motifs. The two main resources are Conserved Domain Database (CDD) and InterPro, which provide a comprehensive collection of conserved motifs curated from a large set of protein sequences and their respective search interfaces. Each of these domain resources or consortium comprises of a large number of domain entries

derived from classic multiple sequence alignment (MSA) of representative sequences of proteins. These MSAs of domains are available in the form of position weight matrix (PWM) or profile hidden Markov models (HMM), which correspond to an average of 10 databases each. Each database has been discussed separately with a total number of entries that were available at the time of writing this thesis (July 2016).

CDD

CDD v3.15 contains 52,411 domain entries of which large numbers of domain entries correspond to NCBI CDD curation effort (11,474 domains), Pfam (16,230 domains) and TIGRFAMs (4,488 domains). Other major resources are SMART (1,013 domains), Clusters of Orthologous Groups of proteins (COGs) (4,873 domains) and Protein Clusters (PRK) (10,885 domains). The NCBI CDD curation project curates a sequence/structure/function relationship of domain families, taking the correlation of residue conservation patterns and functional properties into account. These domains are present in the database as Position Specific Scoring Matrices, which has been discussed in detail in the section *algorithms and tools*. The important members of CDD have been briefly described below.

Pfam is one of the biggest protein family databases that contain manually curated domain entries defined by probabilistic profile hidden Markov models (HMM). Pfam classifies these entries into protein domain families, created from highly representative sequences from UniProt knowledgebase. TIGRFAM is a domain family database that comprises of domain entries as HMM, built from sequence alignments and are annotated accordingly. For each domain family in both Pfam and TIGRFAMs a curated cut-off is assigned for the corresponding HMM that serves as a criterion for the selection and annotation of a query protein. Like the previously described databases, the SMART database comprises of more than 1,200 domains as a profile HMM built from multiple sequence alignments of representative sequences. These domains constitute about 500 domain families linked with signalling, extracellular and chromatin associated proteins. COG is a phylogenetic classification of proteins or group of paralogs from all domains of life. COG uses complete microbial genomes for the orthology based functional characterization of 26 functional categories in 5 COGs including 10,000 proteins each. PRK is an NCBI database that contains protein sequences derived from complete genomes of archaea, bacteria, plants, fungi, protozoans and viruses.

InterPro

InterPro database contains protein signatures such as domains, families and functional sites from various databases. These signatures are present as predictive models in the form of PWM or HMM. The curators assign matching signatures from different databases to a single InterPro identifier in order to maintain consistency by providing non-redundant annotations. At the time of writing this thesis, InterPro database comprises of 29,415 predictive models from its associated database, which are: Pfam (16,295 entries), TIGRFAM (4,488 entries), SMART (1,312 entries), Gene3D (2,626 entries), PANTHER (95,118 entries), PIRSF (3,285 entries), PRINTS (2,106 entries), ProDom (1,894 entries), PROSITE (2,445 entries), and SUPERFAMILY (2,019 entries). The first three databases, which are also the members of CDD, have already been discussed earlier. The remaining databases are briefly described below.

The Gene3D database provides sequence annotations for the protein databases Ensembl, UniProt, and RefSeq. It uses HMM, graph-theory based method, and CATH domain families for domain identification. As a part of the Gene Ontology Reference Genome Project, the PANTHER database includes HMM of domain families, categorized into subfamilies that are used for classification and identification of protein function. PIRSF stands for Protein Informatics Resources SuperFamily, and represents a protein classification and annotation resource generated from the evolutionary relationships between protein sequences. The PRINTS database is a collection of protein family fingerprints. These families of groups of motifs provide biological context compiled from matching motif neighbors. The ProDom database catalogues protein domain families generated from the comparison of all protein sequences. In integrates structural information from SCOP database. At PROSITE, a set of documentation entries of protein domains, domain families and profiles that are used for the annotation of UniprotKB entries is curated. Finally, SUPERFAMILY is a resource for structural classification of proteins (SCOP) at superfamily levels for complete genomes. It also comprises of domain specific Gene Ontology for functional assignments of proteins.

Domain prediction tools

APRICOT uses domain prediction tools for the primary analysis of query proteins for the collection of their functional units. As discussed earlier, APRICOT uses the two major domain consortia CDD and InterPro, which are two comprehensive collections of domains from

diverse domain databases. Both these resources use specialized tools for querying protein sequences of interest against its domain entries.

CDD uses Reversed PSI-BLAST (RPS-BLAST) to search a query protein against its PSSM entries. Basic Local Alignment Search Tool (BLAST) is a method to search for similar sequences in a database by identifying regions of local alignment. PSI-BLAST stands for Position-Specific Iterative BLAST, which uses protein BLAST (BLASTp) to query protein sequences against a database and derives PSSM profiles from multiple sequence alignments (MSA) of matching sequences. The consensus created from the PSSM is used for querying more matching sequences, which are merged with previously discovered sequences to create a new corresponding consensus. This search for the matching sequences can be carried out in several iterations, which can be set by the users. This iterative way to look for matching profiles is very efficient for the identification of remote conservation; however, its run time depends on the number of iterations. RPS-BLAST provides a faster alternative by reducing the run time by allowing users to search their query against a pre-compiled PSSM in CDD by directly searching for the matching profiles in the query sequences in one pass.

InterPro uses InterProScan software for the annotation of protein sequences with domains from different members of the consortium. Most of the applications linked to InterPro, such as Pfam, TIGRFAM, SMART, PIRSF, and PRINTS, use a variety of methods based on HMMER and BLAST algorithms to query PWM and HMM entries from the databases. The Java-based architecture of InterProScan provides a platform to execute database specific search applications in a parallelized manner, which provides a combined output.

Algorithms used in this study (short descriptions with important references)

1. Position-Specific Scoring Matrices (PSSM)

One of the methods to capture conservation patterns in matching sequences is PSSM. PSSM searches for profiles by multiply aligning the matching sequences and producing a matrix of scores for each position in the alignment where the highly-conserved positions are scored higher than the weakly conserved regions (zero for no conservation). This score matrix is used to generate a consensus, which is subsequently used for the identification of additional matching sequences. The high-scoring matches from next rounds are added to the multiple alignments, the position-based scoring is carried out, and profiles are refined. The

new profile can be further used as query for subsequent searches or until convergence. This iterative method for the search of similar sequences allows the identification of divergent sequences. Therefore, PSSM is an efficient probabilistic method of scoring a residue in an alignment at the particular column (a position in the aligned sequence) as a part of a meaningful alignment. The probability of a residue type *a* occurring in a column *u* of the PSSM $q_{u,a}$ and the probability of this residue to occur at any other sequence including the backgrounds as p_a that are not related to the alignment (Zvelebil & Marketa, 2008). The log odd form for a PSSM element can be estimated by:

$$m_{u,a} = log \frac{q_{u,a}}{p_a}$$

In this thesis, I have implemented PSSM by the integration of the RPS-BLAST approach, which queries PSSM containing CDD database for the identification of pre-computed matching profiles in one pass without requiring the iterative searches.

2. Hidden Markov model (HMM)

A Markov model is an approach for probabilistic modelling of sequence conservation by predicting the sequence of state changes based on sequence of observation. In other words, it predicts the likelihood of a sequence to have descended from a particular sequence and hence helps in building a model of the most probable consensus from a set of related sequences in the form of a set of rules or scores. The hidden Markov model generates two states of information: an underlying state path that occurs while transitioning from one state to another, and observed sequences where residues are emitted from one state in the state path. In a state path or hidden Markov chain, only the observed sequence is given as the next position in the sequence or state depends only on the current state. Since only the observed sequence is given, the underlying state path (which state to go next) is hidden, which is modelled by HMM using emission transition (emitting a residue when a state is visited) and transition (moving from one state to other) probabilities (Krogh *et al.*, 1994). The probability of an amino acid *a* occurring in a sequence of length $L\ (x_1,\ldots,x_L)$ is the sum over all possible paths (sequence of states q_1 to q_N, where *q* is state in HMM and *N* is the number of states in a path) that could produce that sequence, which is written as follows:

$$Prob(x_1 \ldots x_n|model) = \sum_{paths\ q_0 \ldots q_{N+1}} Prob(x_1 \ldots x_n, q_1 \ldots q_{N+1}|model)$$

In several domain databases like Pfam, TIGRFAM, and PRINTS comprise of domain entries as HMM which can be queried using database specific software such as HMMER (Eddy, 2009 & 2011) assembled in the previously mentioned InterProScan software.

3. Euclidean distance

Euclidean distance is a distance between two points in Euclidean space, which is used in the context of protein bioinformatics for the distance calculation between two amino acid sequences for certain features. For two sequences *x* and *y,* the distance between a feature *a* in sequence *x* and its corresponding feature *b* in sequence *y* can be calculated as follows (Alaxander Bogomoly, The distance formula):

$$dist\big((x,y),(a,b)\big) = \sqrt{(x-a)^2 + (y-b)^2}$$

In this thesis, Euclidean distance has been used in the context of annotation scoring by calculating the distance of protein compositions (di-peptides, tri-peptides, and physico-chemical properties) between reference and predicted domains. These distances are used in scoring the annotation of predicted domains by 1-Euclidean distance (0 to 1, 1 = absolute match). The second usage of this algorithm is in the clustering of CoIP libraries based on their enrichment profiles, which places the CoIP sample with similar enrichment profiles together in a cluster compared to the CoIP samples with different enrichment profiles.

4. Needleman-Wunsch Algorithm

The Needleman-Wunsch algorithm (Needleman & Wunsch, 1970) is a dynamic programming algorithm for the alignment of sequences where higher scores are assigned for a match and lower scores for a mismatch. Dynamic programming solves smaller independent problems and creates a scoring matrix for different possible alignments. Thereafter, a high scoring alignment in the matrix is identified that indicates an optimal alignment of two sequences. In APRICOT, the Needleman-Wunsch algorithm calculates the extent of similarity between the predicted domain region in the query and its corresponding reference sequence for the annotation based scoring. In their original publication, the matches or mismatches were used for the scoring of the alignment of two proteins without using any penalty for the gaps. The scoring of two sequences (S(A) and S(B)) requires a two dimensional array (matrix $F_{i,j}$ for the entry in row *i* and column *j*) which is denoted as follows:

$$F_{i,j} = \max(F_{i,j-1} + S(A_i, B_i), F_{i-1,j} + S(A_i, B_i))$$

5. Support Vector Machine (SVM)

SVMs are supervised learning models used analysis for classification and regression analysis. The SVM training algorithm builds a model from a set of objects by assigning them to different categories, making it a non-probabilistic binary linear classifier. When a new object is introduced, the SVM model maps it to a category based on which side of the gap it falls on, predicted by the classifiers. In addition to performing linear classification, SVMs can efficiently perform a non-linear classification using what is called the kernel trick, implicitly mapping their inputs into high-dimensional feature spaces.

6. Calculation of accuracy

The Receiver Operating Characteristic (ROC) curve is a plotting system used to show the performance by a binary classifier system in various thresholds for its discriminatory parameters. The true positive rates, sensitivity versus the false positive rates, or 1-specificity are plotted at different parameter thresholds. Area Under Curve (AUC) stands for area under the ROC curve, and tests if a positive instance ranks higher than a negative instance. This analysis allows the selection of suboptimal parameter threshold in terms of class distribution. In this thesis, ROC and AUC analyses were used for each parameter and combinations of parameters used for domain prediction to assess their marginal contributions to the overall accuracy of APRICOT with which it identifies RNA-binding proteins. The decision values of the parameters were evaluated for all the predicted RNA-binding domain entries in the training sets. Subsequently, the ROC and AUC were generated for the ranking of the parameters using the module “ROC Curve for Binary SVM” (authors: Tingfan Wu, Chien-Chih Wang, and Hsiang-Fu Yu) from the LIBSVM Version 3.21, a package for support vector machines (Chang & Lin, 2015).

7. Trimmed Mean of M-values (TMM) for normalization

TMM normalization is used for estimating relative RNA production levels in different RNA-Seq libraries (Robinson *et al.*, 2010). TMM allows the calculation of scaling factors between samples that can be used for the analysis of differentially expressed genes that uses a weighted trimmed mean of the log expression ratios. By defining one sample as the reference, a reference gene set is selected that has non-zero read counts. Based on this selected set of genes, the effective library sizes for the non-reference samples are estimated, which are further used for calculating the TMM normalization factors followed by the

normalization of read counts in each sample. In this study, a modified approach was used, where rather than using gene set selected from one reference sample, genes are selected that are expressed across all the samples. The size factors are calculated for each sample by taking only these genes into account, which are further used for the normalization of exact gene expression values of the samples.

A weighted TMM of the log expression ratio (fold-change in terms of relative production) of a gene in two samples can be determined by an empirical strategy. TMM equates the overall expression levels of genes between the samples assuming that the majority of them are not differentially expressed (Robinson *et al.*, 2010).

8. Principle Component Analysis (PCA)

PCA is a statistical procedure often used for exploratory data analysis that uses linear combinations of the data to define a new set of unrelated variables or principal components (Alter *et al.*, 2000; Jolliffe, 2002; Ringner, 2008). PCA is used for identifying the reduced dimensionalities of a data set to account for the variation in the data set. The dimensions are identified in terms of principal components, along which the variation in the data is maximal. The data sets can be represented by plotting the few principle components, which allows a visual assessment of the similarities and differences between different data sets. In this study PCA has been used to represent RIP-Seq samples with a smaller number of variables (shown for the CSPs in the **Figure 3.16A**), and detect dominant patterns of gene expression. Since similarities between data sets are correlated to the distances in the projection of the space defined by the principal components, PCA was used to highlight the similarity and difference between the gene expression data with respect to the principal components.

9. Cluster analysis of expression data

Cluster analysis is an exploratory approach for the classification of objects in a manner such that the similar objects are placed in the same group/cluster, and objects of a different nature are clustered with the objects of its kind. Analysis by allowing unsupervised learning allows the identification of the subset of objects of a similar pattern (Eisen *et al.*, 2000). In this study, I used this approach to cluster RIP-Seq samples by their enrichment patterns. The clustering was computed by using the heatmap.2 command within the ggplots package of the R programming language (Gregory *et al.*, 2016) by using Ward's clustering method and the Euclidean distance measure. The Ward's minimum variance method aims to find

compact spherical clusters and the Euclidean distance was used to compute the distance between the clusters. The clustering analysis of RIP-Seq data set led to the grouping of samples with similar enrichment profiles and gene sets with similar enrichment profiles across the samples.

10. GO and KEGG Pathway enrichment analysis

An enrichment analysis is a statistical approach to identify a set of genes expressed in a certain condition if they belong to same GO term (biological processes, cellular components and molecular functions) (Ashburner *et al.*, 2000) and KEGG pathway (Kanehisa *et al.*, 2000 & 2012). The enrichment analysis is carried out on a set of genes with respect to a background set of genes for the calculation of the statistical significance of the analysis. In this study, for the identification of enriched GO terms and pathways, this information of enriched genes is used against entire genome as background by including the level of enrichment of these genes into account. Fisher's exact test *P* values (Fisher *et al.*, 1922) were calculated as a measure of statistical significance, which, unlike approximate values, calculates the deviation from a null hypothesis closer to the exact values.

CoIP library preparation

(conducted by Drs. Charlotte Michaux, Nora C. Marbaniang, and Erik Holmqvist)

The coding regions of all the selected candidate RBPs and positive controls were cloned into a pBAD24-derived plasmid with and additional C-terminal 3xFLAG tag and introduced into *Salmonella* Typhimurium SL1344. The tagged strains, together with a empty plasmid as non-target (NT) controls were grown in LB supplemented with ampicillin overnight (220 rpm, 37 °C). Five hundred microliters of overnight culture was then diluted 1into 50ml fresh medium. At an OD_{600} of 0.2, L-arabinose at the final concentration of 0.2% was added in order to induce overexpression of the proteins of interest. At an OD_{600} of 2, 100 OD of culture was collected by centrifugation (4700 rpm, 40 min, 4 °C) and subjected to CoIP according to the protocol of (Chao *et al.*, 2012). Briefly, bacteria were re-suspended in 800 µl of ice-cold lysis buffer (20 mM Tris, pH 8.0, 150 mM KCl, 1 mM $MgCl_2$, 1 mM dithiothreitol), and disrupted with 1 ml glass beads (BioSpec Products, 0.1 mm diameter) by shaking at 30 Hz for 10 min. The cleared lysate obtained after centrifugation (16000 rcf, 15 min, 4 °C) was incubated with 0.5µl/OD anti-FLAG antibody (Sigma, F1804) at 4 °C for 30 min and incubated

with 64 µl pre-washed Protein-A Sepharose (Sigma, P6649) for an additional 30 min. After 5 washes in ice-cold lysis buffer, the sepharose was subjected to RNA extraction using phenol:chloroform:isoamyl alcohol (PCI 25:24:1, pH 4.5; Roth, X985.3). After DNase I digestion (Life Technologies), the RNA was used to construct cDNA libraries by Vertis Biotechnologie AG (Freising, Germany), and sequenced on the in-house MiSeq apparatus (Illumina) with 100 cycles in a strand-specific manner. All the RIP-Seq libraries involved in this study will be deposited at GEO.

Software and packages for data visualization

APRICOT uses matplotlib and ggplot packages from Python programming language to visualize the analysis output of the software.

I used R packages for creating scatter plot, box-plot, clustering, and heatmap images for Chapter 2 and Chapter 3 of this thesis. Other software used to generate images in the projects described in this thesis are as follows: circos for circular visualization of RNA-Seq genome expression, Adobe Illustrator, graphical features of Microsoft PowerPoint/Excel, and Voronto mapper for expression of ontology.

References

Abdelmohsen, K. (2012). Modulation of Gene Expression by RNA Binding Proteins: mRNA Stability and Translation.

Agulhon, C., Blanchet, P., Kobetz, A., Marchant, D., Faucon, N., Sarda, P., Moraine, C., Sittler, A., Biancalana, V., Malafosse, A., et al. (1999). Expression of FMR1, FXR1, and FXR2 genes in human prenatal tissues. J. Neuropathol. Exp. Neurol. *58*, 867–880.

Alfano, C., Sanfelice, D., Babon, J., Kelly, G., Jacks, A., Curry, S., and Conte, M.R. (2004). Structural analysis of cooperative RNA binding by the La motif and central RRM domain of human La protein. Nat. Struct. Mol. Biol. 11, 323–329.

Alter, O., Brown, P.O., and Botstein, D. (2000). Singular value decomposition for genome-wide expression data processing and modeling. Proc. Natl. Acad. Sci. U.S.A. *97*, 10101–10106.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990). Basic local alignment search tool. J. Mol. Biol. 215, 403–410.

Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25, 3389–3402.

Anantharaman, V., Zhang, D., and Aravind, L. (2010). OST-HTH: a novel predicted RNA-binding domain. Biol. Direct *5*, 13.

Anders, S., and Huber, W. (2010). Differential expression analysis for sequence count data. Genome Biology 11, R106.

Anji, A., and Kumari, M. (2016). Guardian of Genetic Messenger-RNA-Binding Proteins. Biomolecules 6, 4.

Aravind, L., and Koonin, E.V. (1999). Novel predicted RNA-binding domains associated with the translation machinery. J. Mol. Evol. 48, 291–302.

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. Nat. Genet. 25, 25–29.

Attwood, T.K., Coletta, A., Muirhead, G., Pavlopoulou, A., Philippou, P.B., Popov, I., Romá-Mateo, C., Theodosiou, A., and Mitchell, A.L. (2012). The PRINTS database: a fine-grained protein sequence annotation and analysis resource--its status in 2012. Database (Oxford) 2012, bas019.

Aubourg, S., Kreis, M., and Lecharny, A. (1999). The DEAD box RNA helicase family in Arabidopsis thaliana. Nucleic Acids Res. 27, 628–636.

Ayache, J., Bénard, M., Ernoult-Lange, M., Minshall, N., Standart, N., Kress, M., and Weil, D. (2015). P-body assembly requires DDX6 repression complexes rather than decay or Ataxin2/2L complexes. Mol Biol Cell *26*, 2579–2595.

Baltz, A.G., Munschauer, M., Schwanhäusser, 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.

Barquist, L., and Vogel, J. (2015). Accelerating Discovery and Functional Analysis of Small RNAs with New Technologies. Annu. Rev. Genet. 49, 367–394.

Bateman, A., and Kickhoefer, V. (2003). The TROVE module: a common element in Telomerase, Ro and Vault ribonucleoproteins. BMC Bioinformatics *4*, 49.

Bendtsen, J.D., Jensen, L.J., Blom, N., Von Heijne, G., and Brunak, S. (2004). Feature-based prediction of non-classical and leaderless protein secretion. Protein Eng. Des. Sel. 17, 349–356.

Björkman, J., Samuelsson, P., Andersson, D.I., and Hughes, D. (1999). Novel ribosomal mutations affecting translational accuracy, antibiotic resistance and virulence of Salmonella typhimurium. Mol. Microbiol. *31*, 53–58.

Bogomolny A., Copyright 2011-2016. The Distance Formula from Interactive Mathematics Miscellany and Puzzles. http://www.cut-the-knot.org/pythagoras/DistanceFormula.shtml

Boni, I.V., Isaeva, D.M., Musychenko, M.L., and Tzareva, N.V. (1991). Ribosome-messenger recognition: mRNA target sites for ribosomal protein S1. Nucleic Acids Res. *19*, 155–162.

Brown, R.S. (2005). Zinc finger proteins: getting a grip on RNA. Curr Opin Struct Biol 15, 94–98.

Bru, C., Courcelle, E., Carrère, S., Beausse, Y., Dalmar, S., and Kahn, D. (2005). The ProDom database of protein domain families: more emphasis on 3D. Nucleic Acids Res. 33, D212–215.

Brunetti, J.E., Scolaro, L.A., and Castilla, V. (2015). The heterogeneous nuclear ribonucleoprotein K (hnRNP K) is a host factor required for dengue virus and Junín virus multiplication. Virus Res. 203, 84–91.

Bulashevska, A., and Eils, R. (2006). Predicting protein subcellular locations using hierarchical ensemble of Bayesian classifiers based on Markov chains. BMC Bioinformatics 7, 298.

Burd, C.G., and Dreyfuss, G. (1994). Conserved structures and diversity of functions of RNA-binding proteins. Science 265, 615–621.

Bycroft, M., S, G., Ag, M., M, P., and D, S.J. (1995). NMR solution structure of a dsRNA binding domain from Drosophila staufen protein reveals homology to the N-terminal domain of ribosomal protein S5. EMBO J 14, 3563–3571.

Bycroft, M., Hubbard, T.J., Proctor, M., Freund, S.M., and Murzin, A.G. (1997). The solution structure of the S1 RNA binding domain: a member of an ancient nucleic acid-binding fold. Cell 88, 235–242.

Cai, Y.-D., and Chou, K.-C. (2006). Predicting membrane protein type by functional domain composition and pseudo-amino acid composition. J. Theor. Biol. 238, 395–400.

Calabretta, S., and Richard, S. (2015). Emerging Roles of Disordered Sequences in RNA-Binding Proteins. Trends Biochem. Sci. 40, 662–672.

Cantara, W.A., Crain, P.F., Rozenski, J., McCloskey, J.A., Harris, K.A., Zhang, X., Vendeix, F.A.P., Fabris, D., and Agris, P.F. (2011). The RNA Modification Database, RNAMDB: 2011 update. Nucleic Acids Res. *39*, D195–201.

Cassola, A., Noé, G., and Frasch, A.C. (2010). RNA recognition motifs involved in nuclear import of RNA-binding proteins. RNA Biol 7, 339–344.

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. Mol Cell *63*, 696–710.

Chang, J.-M., Su, E.C.-Y., Lo, A., Chiu, H.-S., Sung, T.-Y., and Hsu, W.-L. (2008). PSLDoc: Protein subcellular localization prediction based on gapped-dipeptides and probabilistic latent semantic analysis. Proteins *72*, 693–710.

Chao, Y., and Vogel, J. (2010). The role of Hfq in bacterial pathogens. Curr. Opin. Microbiol. 13, 24–33.

Cho, D.H., and Tapscott, S.J. (2007). Myotonic dystrophy: emerging mechanisms for DM1 and DM2. Biochim. Biophys. Acta *1772*, 195–204.

Chapman, B., and Chang, J. (2000). Biopython: Python Tools for Computational Biology. SIGBIO Newsl. 20, 15–19.

Chaulk, S.G., Smith Frieday, M.N., Arthur, D.C., Culham, D.E., Edwards, R.A., Soo, P., Frost, L.S., Keates, R.A.B., Glover, J.N.M., and Wood, J.M. (2011). ProQ is an RNA chaperone that controls ProP levels in Escherichia coli. Biochemistry 50, 3095–3106.

Chen, Y.C., Wu, C.Y., and Lim, C. (2007). Predicting DNA-binding amino acid residues from electrostatic stabilization upon mutation to Asp/Glu and evolutionary conservation. Proteins *67*, 671–680.

Cheng, C.-W., Su, E.C.-Y., Hwang, J.-K., Sung, T.-Y., and Hsu, W.-L. (2008). Predicting RNA-binding sites of proteins using support vector machines and evolutionary information. BMC Bioinformatics 9, S6.

Chang, C.-C., and Lin, C.-J. (2011). LIBSVM : a library for support vector machines. ACM Transactions on Intelligent Systems and Technology, 2:27:1--27:27. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

Chaudhuri, R.R., Morgan, E., Peters, S.E., Pleasance, S.J., Hudson, D.L., Davies, H.M., Wang, J., van Diemen, P.M., Buckley, A.M., Bowen, A.J., et al. (2013). Comprehensive Assignment of Roles for Salmonella Typhimurium Genes in Intestinal Colonization of Food-Producing Animals. PLoS Genet *9*.

Chothia, C. (1976). The nature of the accessible and buried surfaces in proteins. J. Mol. Biol. 105, 1–12.

Chothia, C., and Lesk, A.M. (1986). The relation between the divergence of sequence and structure in proteins. EMBO J 5, 823–826.

Chou, K.-C., and Shen, H.-B. (2006). Large-scale predictions of gram-negative bacterial protein subcellular locations. J. Proteome Res. 5, 3420–3428.

Chou, K.-C., and Shen, H.-B. (2008). Cell-PLoc: a package of Web servers for predicting subcellular localization of proteins in various organisms. Nat Protoc 3, 153–162.

Cléry, A., and Allain, F.H.-T. (2013). From structure to function of RNA binding domains (Landes Bioscience).

Cléry, A., Blatter, M., and Allain, F.H.-T. (2008). RNA recognition motifs: boring? Not quite. Curr. Opin. Struct. Biol. 18, 290–298.

Cock, P.J.A., Antao, T., Chang, J.T., Chapman, B.A., Cox, C.J., Dalke, A., Friedberg, I., Hamelryck, T., Kauff, F., Wilczynski, B., *et al.* (2009). Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics 25, 1422–1423.

Colombrita, C., Silani, V., and Ratti, A. (2013). ELAV proteins along evolution: back to the nucleus? Mol. Cell. Neurosci. 56, 447–455.

Conesa, A., and Götz, S. (2008). Blast2GO: A comprehensive suite for functional analysis in plant genomics. Int J Plant Genomics 2008, 619832.

Conrad, T., Albrecht, A.-S., Costa, V.R. de M., Sauer, S., Meierhofer, D., and Ørom, U.A. (2016). Serial interactome capture of the human cell nucleus. Nature Communications 7, 11212.

Cook, K.B., Kazan, H., Zuberi, K., Morris, Q., and Hughes, T.R. (2011). RBPDB: a database of RNA-binding specificities. Nucleic Acids Res. 39, D301–308.

Darnell, R. (2012). CLIP (cross-linking and immunoprecipitation) identification of RNAs bound by a specific protein. Cold Spring Harb Protoc *2012*, 1146–1160.

Daubner, G.M., Cléry, A., and Allain, F.H.-T. (2013). RRM-RNA recognition: NMR or crystallography…and new findings. Curr. Opin. Struct. Biol. 23, 100–108.

De Klerk, E., and 't Hoen, P.A.C. (2015). Alternative mRNA transcription, processing, and translation: insights from RNA sequencing. Trends Genet. 31, 128–139.

De la Cruz, J., Kressler, D., and Linder, P. (1999). Unwinding RNA in Saccharomyces cerevisiae: DEAD-box proteins and related families. Trends Biochem. Sci. *24*, 192–198.

De Lima Morais, D.A., Fang, H., Rackham, O.J.L., Wilson, D., Pethica, R., Chothia, C., and Gough, J. (2011). SUPERFAMILY 1.75 including a domain-centric gene ontology method. Nucleic Acids Res. 39, D427–434.

Declerck, N., Vincent, F., Hoh, F., Aymerich, S., and van Tilbeurgh, H. (1999). RNA recognition by transcriptional antiterminators of the BglG/SacY family: functional and structural comparison of the CAT domain from SacY and LicT. J. Mol. Biol. *294*, 389–402.

Demirci, H., Larsen, L.H.G., Hansen, T., Rasmussen, A., Cadambi, A., Gregory, S.T., Kirpekar, F., and Jogl, G. (2010). Multi-site-specific 16S rRNA methyltransferase RsmF from Thermus thermophilus. RNA *16*, 1584–1596.

Deng, H., Liu, H., Li, X., Xiao, J., and Wang, S. (2012). A CCCH-type zinc finger nucleic acid-binding protein quantitatively confers resistance against rice bacterial blight disease. Plant Physiol. *158*, 876–889.

Derti, A., Garrett-Engele, P., Macisaac, K.D., Stevens, R.C., Sriram, S., Chen, R., Rohl, C.A., Johnson, J.M., and Babak, T. (2012). A quantitative atlas of polyadenylation in five mammals. Genome Res. 22, 1173–1183.

Dictenberg, J.B., Swanger, S.A., Antar, L.N., Singer, R.H., and Bassell, G.J. (2008). A direct role for FMRP in activity-dependent dendritic mRNA transport links filopodial-spine morphogenesis to fragile X syndrome. Dev. Cell *14*, 926–939.

Dominissini, D., Moshitch-Moshkovitz, S., Schwartz, S., Salmon-Divon, M., Ungar, L., Osenberg, S., Cesarkas, K., Jacob-Hirsch, J., Amariglio, N., Kupiec, M., et al. (2012). Topology of the human and mouse m6A RNA methylomes revealed by m6A-seq. Nature *485*, 201–206.

Dye, B.T., and Patton, J.G. (2001). An RNA recognition motif (RRM) is required for the localization of PTB-associated splicing factor (PSF) to subnuclear speckles. Exp. Cell Res. 263, 131–144.

Eddy, S.R. (2009). A new generation of homology search tools based on probabilistic inference. Genome Inform *23*, 205–211.

Eddy, S.R. (2011). Accelerated Profile HMM Searches. PLoS Comput. Biol. *7*, e1002195.

Eisenhaber, F., Frömmel, C., and Argos, P. (1996). Prediction of secondary structural content of proteins from their amino acid composition alone. II. The paradox with secondary structural class. Proteins 25, 169–179.

Eisen, M.B., Spellman, P.T., Brown, P.O., and Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns. Proc. Natl. Acad. Sci. U.S.A. *95*, 14863–14868.

Faehnle, C.R., Elkayam, E., Haase, A.D., Hannon, G.J., and Joshua-Tor, L. (2013). The making of a slicer: activation of human Argonaute-1. Cell Rep *3*, 1901–1909.

Faoro, C., and Ataide, S.F. (2014). Ribonomic approaches to study the RNA-binding proteome. FEBS Letters *588*, 3649–3664.

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 Res. 42, D222–230.

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 Res. 44, D279–285.

Fischer, U., Englbrecht, C., and Chari, A. (2011). Biogenesis of spliceosomal small nuclear ribonucleoproteins. Wiley Interdiscip Rev RNA 2, 718–731.

Fisher, R.A. (1922). On the Interpretation of χ2 from Contingency Tables, and the Calculation of P. Journal of the Royal Statistical Society 85, 87–94.

Fontecave, M., Atta, M., and Mulliez, E. (2004). S-adenosylmethionine: nothing goes to waste. Trends Biochem. Sci. *29*, 243–249.

Förstner, K.U., Vogel, J., and Sharma, C.M. (2014). READemption—a tool for the computational analysis of deep-sequencing–based transcriptome data. Bioinformatics 30, 3421–3423.

Freeberg, M.A., and Kim, J.K. (2016). Mapping the Transcriptome-Wide Landscape of RBP Binding Sites Using gPAR-CLIP-seq: Bioinformatic Analysis. Methods Mol. Biol. 1361, 91–104.

Galperin, M.Y. (2006). Structural classification of bacterial response regulators: diversity of output domains and domain combinations. J. Bacteriol. *188*, 4169–4182.

Gamsjaeger, R., Liew, C.K., Loughlin, F.E., Crossley, M., and Mackay, J.P. (2007). Sticky fingers: zinc-fingers as protein-recognition motifs. Trends Biochem. Sci. *32*, 63–70.

Gerstberger, S., Hafner, M., and Tuschl, T. (2014). A census of human RNA-binding proteins. Nat Rev Genet 15, 829–845.

Girard, A., Sachidanandam, R., Hannon, G.J., and Carmell, M.A. (2006). A germline-specific class of small RNAs binds mammalian Piwi proteins. Nature *442*, 199–202.

Giudice, E., Macé, K., and Gillet, R. (2014). Trans-translation exposed: understanding the structures and functions of tmRNA-SmpB. Front Microbiol *5*.

Göpel, Y. (2014). GlmY and GlmZ: a hierarchically acting cascade composed of two small RNAs.

Grishin, N.V. (2001). KH domain: one motif, two folds. Nucleic Acids Res 29, 638–643.

Guth, S., Tange, T.Ø., Kellenberger, E., and Valcárcel, J. (2001). Dual function for U2AF(35) in AG-dependent pre-mRNA splicing. Mol. Cell. Biol. 21, 7673–7681.

Habib, T., Zhang, C., Yang, J.Y., Yang, M.Q., and Deng, Y. (2008). Supervised learning method for the prediction of subcellular localization of proteins using amino acid and amino acid pair composition. BMC Genomics 9 Suppl 1, S16.

Haft, D.H., Selengut, J.D., Richter, R.A., Harkins, D., Basu, M.K., and Beck, E. (2013). TIGRFAMs and Genome Properties in 2013. Nucleic Acids Res. 41, D387–395.

Hall, T.M. (2005). Multiple modes of RNA recognition by zinc finger proteins. Curr Opin Struct Biol 15, 367–373.

Helm, M., and Alfonzo, J.D. (2014). Posttranscriptional RNA Modifications: playing metabolic games in a cell's chemical Legoland. Chem. Biol. *21*, 174–185.

Hinman, M.N., and Lou, H. (2008). Diverse molecular functions of Hu proteins. Cell Mol Life Sci *65*, 3168–3181.

Holmqvist, E., Wright, P.R., Li, L., Bischler, T., Barquist, L., Reinhardt, R., Backofen, R., and Vogel, J. (2016). Global RNA recognition patterns of post-transcriptional regulators Hfq and CsrA revealed by UV crosslinking in vivo. EMBO J. 35, 991–1011.

Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. Journal of Educational Psychology 24, 417–441.

Huaiyu, M., Dong, Q., Muruganujan, A., Gaudet, P., Lewis, S., and Thomas, P.D. (2010). PANTHER version 7: improved phylogenetic trees, orthologs and collaboration with the Gene Ontology Consortium. Nucleic Acids Res. 38, D204–210.

Hyndman, R.J., and Fan, Y. (1996). Sample Quantiles in Statistical Packages. The American Statistician *50*, 361–365.

Jacks, A., Babon, J., Kelly, G., Manolaridis, I., Cary, P.D., Curry, S., and Conte, M.R. (2003). Structure of the C-terminal domain of human La protein reveals a novel RNA recognition motif coupled to a helical nuclear retention element. Structure *11*, 833–843.

Jiang, Y., Oron, T.R., Clark, W.T., Bankapur, A.R., D'Andrea, D., Lepore, R., Funk, C.S., Kahanda, I., Verspoor, K.M., Ben-Hur, A., *et al.* (2016). An expanded evaluation of protein function prediction methods shows an improvement in accuracy. Genome Biol. 17, 184.

Matthews, J.M., and Sunde, M. (2002). Zinc fingers--folds for many occasions. IUBMB Life 54, 351–355.

Jolliffe, I.T. (2002). Principal Component Analysis, Springer.

Jones, P.G., and Inouye, M. (1994). The cold-shock response--a hot topic. Mol Microbiol 11, 811–818.

Jones, P., Binns, D., Chang, H.-Y., Fraser, M., Li, W., McAnulla, C., McWilliam, H., Maslen, J., Mitchell, A., Nuka, G., *et al.* (2014). InterProScan 5: genome-scale protein function classification. Bioinformatics 30, 1236–1240.

Källberg, M., Wang, H., Wang, S., Peng, J., Wang, Z., Lu, H., and Xu, J. (2012). Template-based protein structure modeling using the RaptorX web server. Nat Protoc 7, 1511–1522.

Kanehisa, M., and Goto, S. (2000). KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res. 28, 27–30.

Kanehisa, M., Goto, S., Sato, Y., Furumichi, M., and Tanabe, M. (2012). KEGG for integration and interpretation of large-scale molecular data sets. Nucleic Acids Res. 40, D109–114.

Kelley, L.A., Mezulis, S., Yates, C.M., Wass, M.N., and Sternberg, M.J.E. (2015). The Phyre2 web portal for protein modeling, prediction and analysis. Nat Protoc 10, 845–858.

Kim, U., Wang, Y., Sanford, T., Zeng, Y., and Nishikura, K. (1994). Molecular cloning of cDNA for double-stranded RNA adenosine deaminase, a candidate enzyme for nuclear RNA editing. Proc. Natl. Acad. Sci. U.S.A. *91*, 11457–11461.

Kim, C.A., and Bowie, J.U. (2003). SAM domains: uniform structure, diversity of function. Trends Biochem. Sci. *28*, 625–628.

Klimke, W., Agarwala, R., Badretdin, A., Chetvernin, S., Ciufo, S., Fedorov, B., Kiryutin, B., O'Neill, K., Resch, W., Resenchuk, S., *et al.* (2009). The National Center for Biotechnology Information's Protein Clusters Database. Nucleic Acids Res. 37, D216–223.

Klug, A. (1999). Zinc finger peptides for the regulation of gene expression. J Mol Biol 293, 215–218.

König, J., Zarnack, K., Luscombe, N.M., and Ule, J. (2012). Protein-RNA interactions: new genomic technologies and perspectives. Nat. Rev. Genet. *13*, 77–83.

Koonin, E.V. (1996). Pseudouridine synthases: four families of enzymes containing a putative uridine-binding motif also conserved in dUTPases and dCTP deaminases. Nucleic Acids Res. *24*, 2411–2415.

Kotik-Kogan, O., Valentine, E.R., Sanfelice, D., Conte, M.R., and Curry, S. (2008). Structural analysis reveals conformational plasticity in the recognition of RNA 3' ends by the human La protein. Structure 16, 852–862.

Kouranov, A., Xie, L., de la Cruz, J., Chen, L., Westbrook, J., Bourne, P.E., and Berman, H.M. (2006). The RCSB PDB information portal for structural genomics. Nucleic Acids Res. 34, D302–305.

Kröger, C., Colgan, A., Srikumar, S., Händler, K., Sivasankaran, S.K., Hammarlöf, D.L., Canals, R., Grissom, J.E., Conway, T., Hokamp, K., *et al.* (2013). An infection-relevant transcriptomic compendium for Salmonella enterica Serovar Typhimurium. Cell Host Microbe 14, 683–695.

Krogh, A., Brown, M., Mian, I.S., Sjölander, K., and Haussler, D. (1994). Hidden Markov models in computational biology. Applications to protein modeling. J. Mol. Biol. 235, 1501–1531.

Kumar, M., Gromiha, M.M., and Raghava, G.P.S. (2011). SVM based prediction of RNA-binding proteins using binding residues and evolutionary information. J. Mol. Recognit. 24, 303–313.

Kumar, M.S., and Chen, K.C. (2012). Evolution of animal Piwi-interacting RNAs and prokaryotic CRISPRs. Brief Funct Genomics *11*, 277–288.

Kuroyanagi, H. (2009). Fox-1 family of RNA-binding proteins. Cell Mol Life Sci 66, 3895–3907.

Kwon, S.C., Yi, H., Eichelbaum, K., Föhr, S., Fischer, B., You, K.T., Castello, A., Krijgsveld, J., Hentze, M.W., and Kim, V.N. (2013). The RNA-binding protein repertoire of embryonic stem cells. Nat Struct Mol Biol 20, 1122–1130.

Laity, J.H., Bm, L., and Pe, W. (2001). Zinc finger proteins: new insights into structural and functional diversity. Curr Opin Struct Biol 11, 39–46.

Lam, S.D., Dawson, N.L., Das, S., Sillitoe, I., Ashford, P., Lee, D., Lehtinen, S., Orengo, C.A., and Lees, J.G. (2016). Gene3D: expanding the utility of domain assignments. Nucleic Acids Res. 44, D404–409.

Landsman, D. (1992). RNP-1, an RNA-binding motif is conserved in the DNA-binding cold shock domain. Nucleic Acids Res 20, 2861–2864.

Langridge, G.C., Phan, M.-D., Turner, D.J., Perkins, T.T., Parts, L., Haase, J., Charles, I., Maskell, D.J., Peters, S.E., Dougan, G., et al. (2009). Simultaneous assay of every Salmonella Typhi gene using one million transposon mutants. Genome Res *19*, 2308–2316.

Lease, R.A., and Woodson, S.A. (2004). Cycling of the Sm-like protein Hfq on the DsrA small regulatory RNA. J. Mol. Biol. *344*, 1211–1223.

Letunic, I., Doerks, T., and Bork, P. (2015). SMART: recent updates, new developments and status in 2015. Nucleic Acids Res. 43, D257–260.

Linder, P., and Jankowsky, E. (2011). From unwinding to clamping - the DEAD box RNA helicase family. Nat. Rev. Mol. Cell Biol. 12, 505–516.

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

Livi, C.M., Klus, P., Ponti, R.D., and Tartaglia, G.G. (2015). catRAPID signature: Identification of Ribonucleoproteins and RNA-Binding Regions. Bioinformatics btv629.

Love, M.I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-Seq data with DESeq2. Genome Biology 15, 550.

Lunde, B.M., Moore, C., and Varani, G. (2007). RNA-binding proteins: modular design for efficient function. Nat Rev Mol Cell Biol 8, 479–490.

Ma, X., Guo, J., Wu, J., Liu, H., Yu, J., Xie, J., and Sun, X. (2011). Prediction of RNA-binding residues in proteins from primary sequence using an enriched random forest model with a novel hybrid feature. Proteins *79*, 1230–1239.

Magrane, M., and UniProt Consortium (2011). UniProt Knowledgebase: a hub of integrated protein data. Database (Oxford) 2011, bar009.

Manche, L., Green, S.R., Schmedt, C., and Mathews, M.B. (1992). Interactions between double-stranded RNA regulators and the protein kinase DAI. Mol. Cell. Biol. 12, 5238–5248.

Makarova, K.S., Grishin, N.V., Shabalina, S.A., Wolf, Y.I., and Koonin, E.V. (2006). A putative RNA-interference-based immune system in prokaryotes: computational analysis of the predicted enzymatic machinery, functional analogies with eukaryotic RNAi, and hypothetical mechanisms of action. Biol. Direct *1*, 7.

Marbaniang, C.N., and Vogel, J. (2016). Emerging roles of RNA modifications in bacteria. Curr. Opin. Microbiol. *30*, 50–57.

Marchese, D., de Groot, N.S., Lorenzo Gotor, N., Livi, C.M., and Tartaglia, G.G. (2016). Advances in the characterization of RNA-binding proteins. Wiley Interdiscip Rev RNA 7, 793–810.

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 Res. 43, D222–226.

Maris, C., Dominguez, C., and Allain, F.H.-T. (2005). The RNA recognition motif, a plastic RNA-binding platform to regulate post-transcriptional gene expression. FEBS J. *272*, 2118–2131.

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal 17, pp. 10–12.

Maruyama, K., Sato, N., and Ohta, N. (1999). Conservation of structure and cold-regulation of RNA-binding proteins in cyanobacteria: probable convergent evolution with eukaryotic glycine-rich RNA-binding proteins. Nucleic Acids Res. *27*, 2029–2036.

Matsuda, S., Vert, J.-P., Saigo, H., Ueda, N., Toh, H., and Akutsu, T. (2005). A novel representation of protein sequences for prediction of subcellular location using support vector machines. Protein Sci. 14, 2804–2813.

Matsui, T., Hogetsu, K., Usukura, J., Sato, T., Kumasaka, T., Akao, Y., and Tanaka, N. (2006). Structural insight of human DEAD-box protein rck/p54 into its substrate recognition with conformational changes. Genes Cells *11*, 439–452.

Matus-Ortega, M.E., Regonesi, M.E., Piña-Escobedo, A., Tortora, P., Dehò, G., and García-Mena, J. (2007). The KH and S1 domains of Escherichia coli polynucleotide phosphorylase are necessary for autoregulation and growth at low temperature. Biochim. Biophys. Acta *1769*, 194–203.

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

Mian, I.S. (1997). Comparative sequence analysis of ribonucleases HII, III, II PH and D. Nucleic Acids Res. *25*, 3187–3195.

Miao, Z., and Westhof, E. (2015). A Large-Scale Assessment of Nucleic Acids Binding Site Prediction Programs. PLOS Comput Biol 11, e1004639.

Miao, Z., and Westhof, E. (2016). RBscore&NBench: a high-level web server for nucleic acid binding residues prediction with a large-scale benchmarking database. Nucl. Acids Res. gkw251.

Mistry, J., Finn, R.D., Eddy, S.R., Bateman, A., and Punta, M. (2013). Challenges in homology search: HMMER3 and convergent evolution of coiled-coil regions. Nucleic Acids Res. 41, e121.

Mitchell, A., Chang, H.-Y., Daugherty, L., Fraser, M., Hunter, S., Lopez, R., McAnulla, C., McMenamin, C., Nuka, G., Pesseat, S., *et al.* (2015). The InterPro protein families database: the classification resource after 15 years. Nucleic Acids Res. 43, D213–221.

Mitchell, S.F., Jain, S., She, M., and Parker, R. (2013). Global analysis of yeast mRNPs. Nat Struct Mol Biol 20, 127–133.

Mizutani, K., Machida, Y., Unzai, S., Park, S.-Y., and Tame, J.R.H. (2004). Crystal structures of the catalytic domains of pseudouridine synthases RluC and RluD from Escherichia coli. Biochemistry *43*, 4454–4463.

Montoya, G., Svensson, C., Luirink, J., and Sinning, I. (1997). Crystal structure of the NG domain from the signal-recognition particle receptor FtsY. Nature *385*, 365–368.

Moore, M.J., and Proudfoot, N.J. (2009). Pre-mRNA Processing Reaches Back toTranscription and Ahead to Translation. Cell *136*, 688–700.

Moult, J., Fidelis, K., Kryshtafovych, A., Schwede, T., and Tramontano, A. (2014). Critical assessment of methods of protein structure prediction (CASP) — round x. Proteins *82*, 1–6.

Najafabadi, H.S., Mnaimneh, S., Schmitges, F.W., Garton, M., Lam, K.N., Yang, A., Albu, M., Weirauch, M.T., Radovani, E., Kim, P.M., *et al.* (2015). C2H2 zinc finger proteins greatly expand the human regulatory lexicon. Nat Biotech 33, 555–562.

Needleman, S.B., and Wunsch, C.D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. J. Mol. Biol. 48, 443–453.

Nielsen, M., Lundegaard, C., Lund, O., and Petersen, T.N. (2010). CPHmodels-3.0--remote homology modeling using structure-guided sequence profiles. Nucleic Acids Res. 38, W576–581.

Nishtala, S., Neelamraju, Y., and Janga, S.C. (2016). Dissecting the expression relationships between RNA-binding proteins and their cognate targets in eukaryotic post-transcriptional regulatory networks. Sci Rep 6, 25711.

Niu, B., Jin, Y.-H., Feng, K.-Y., Lu, W.-C., Cai, Y.-D., and Li, G.-Z. (2008). Using AdaBoost for the prediction of subcellular location of prokaryotic and eukaryotic proteins. Mol. Divers. 12, 41–45.

Van Opijnen, T., and Camilli, A. (2013). Transposon insertion sequencing: a new tool for systems-level analysis of microorganisms. Nat Rev Microbiol 11.

O'Connell, M.A., Mannion, N.M., and Keegan, L.P. (2015). The Epitranscriptome and Innate Immunity. PLoS Genet. *11*, e1005687.

Ostheimer, G.J., Barkan, A., and Matthews, B.W. (2002). Crystal structure of E. coli YhbY: a representative of a novel class of RNA binding proteins. Structure 10, 1593–1601.

Ostheimer, G.J., Williams-Carrier, R., Belcher, S., Osborne, E., Gierke, J., and Barkan, A. (2003). Group II intron splicing factors derived by diversification of an ancient RNA-binding domain. EMBO J. *22*, 3919–3929.

Otaki, J.M., Tsutsumi, M., Gotoh, T., and Yamamoto, H. (2010). Secondary structure characterization based on amino acid composition and availability in proteins. J Chem Inf Model 50, 690–700.

Peal, L., Jambunathan, N., and Mahalingam, R. (2011). Phylogenetic and expression analysis of RNA-binding proteins with triple RNA recognition motifs in plants. Mol. Cells 31, 55–64.

Pedruzzi, I., Rivoire, C., Auchincloss, A.H., Coudert, E., Keller, G., de Castro, E., Baratin, D., Cuche, B.A., Bougueleret, L., Poux, S., *et al.* (2015). HAMAP in 2015: updates to the protein family classification and annotation system. Nucleic Acids Res. 43, D1064–1070.

Penalva, L.O.F., and Sánchez, L. (2003). RNA binding protein sex-lethal (Sxl) and control of Drosophila sex determination and dosage compensation. Microbiol. Mol. Biol. Rev. 67, 343–359, table of contents.

Pérez-Arellano, I., Gallego, J., and Cervera, J. (2007). The PUA domain - a structural and functional overview. FEBS J. *274*, 4972–4984.

Petsalaki, E.I., Bagos, P.G., Litou, Z.I., and Hamodrakas, S.J. (2006). PredSL: a tool for the N-terminal sequence-based prediction of protein subcellular localization. Genomics Proteomics Bioinformatics 4, 48–55.

Phadtare, S., and Inouye, M. (1999). Sequence-selective interactions with RNA by CspB, CspC and CspE, members of the CspA family of Escherichia coli. Mol. Microbiol. 33, 1004–1014.

Pruitt, K.D., Tatusova, T., Klimke, W., and Maglott, D.R. (2009). NCBI Reference Sequences: current status, policy and new initiatives. Nucleic Acids Res. 37, D32–36.

Pruitt, K.D., Tatusova, T., Brown, G.R., and Maglott, D.R. (2012). NCBI Reference Sequences (RefSeq): current status, new features and genome annotation policy. Nucleic Acids Res. 40, D130–135.

Puton, T., Kozlowski, L., Tuszynska, I., Rother, K., and Bujnicki, J.M. (2012). Computational methods for prediction of protein–RNA interactions. Journal of Structural Biology 179, 261–268.

Radivojac, P., Clark, W.T., Oron, T.R., Schnoes, A.M., Wittkop, T., Sokolov, A., Graim, K., Funk, C., Verspoor, K., Ben-Hur, A., *et al.* (2013). A large-scale evaluation of computational protein function prediction. Nat. Methods 10, 221–227.

Ramamurthy, V., Swann, S.L., Spedaliere, C.J., and Mueller, E.G. (1999). Role of cysteine residues in pseudouridine synthases of different families. Biochemistry 38, 13106–13111.

Ray, D., Kazan, H., Chan, E.T., Peña Castillo, L., Chaudhry, S., Talukder, S., Blencowe, B.J., Morris, Q., and Hughes, T.R. (2009). Rapid and systematic analysis of the RNA recognition specificities of RNA-binding proteins. Nat. Biotechnol. 27, 667–670.

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

Reczko, M., and Hatzigerrorgiou, A. (2004). Prediction of the subcellular localization of eukaryotic proteins using sequence signals and composition. Proteomics 4, 1591–1596.

Reference Genome Group of the Gene Ontology Consortium (2009). The Gene Ontology's Reference Genome Project: a unified framework for functional annotation across species. PLoS Comput. Biol. 5, e1000431.

Régnier, P., and Marujo, P.E. (2013). Polyadenylation and Degradation of RNA in Prokaryotes (Landes Bioscience).

Robinson, M.D., and Oshlack, A. (2010). A scaling normalization method for differential expression analysis of RNA-Seq data. Genome Biology 11, R25.

Robinson, M.D., McCarthy, D.J., and Smyth, G.K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics 26, 139–140.

Roden, J.C., King, B.W., Trout, D., Mortazavi, A., Wold, B.J., and Hart, C.E. (2006). Mining gene expression data by interpreting principal components. BMC Bioinformatics *7*, 194.

Romeo, T. (1998). Global regulation by the small RNA-binding protein CsrA and the non-coding RNA molecule CsrB. Mol. Microbiol. 29, 1321–1330.

Roy, A., Kucukural, A., and Zhang, Y. (2010). I-TASSER: a unified platform for automated protein structure and function prediction. Nat Protoc 5, 725–738.

Sachs, R., Max, K.E.A., Heinemann, U., and Balbach, J. (2012). RNA single strands bind to a conserved surface of the major cold shock protein in crystals and solution. RNA *18*, 65–76.

Sakurai, M., Yano, T., Kawabata, H., Ueda, H., and Suzuki, T. (2010). Inosine cyanoethylation identifies A-to-I RNA editing sites in the human transcriptome. Nat. Chem. Biol. *6*, 733–740.

Santamaría, R., and Pierre, P. (2012). Voronto: mapper for expression data to ontologies. Bioinformatics 28, 2281–2282.

Santos, R.L., Zhang, S., Tsolis, R.M., Kingsley, R.A., Adams, L.G., and Bäumler, A.J. (2001). Animal models of Salmonella infections: enteritis versus typhoid fever. Microbes Infect. *3*, 1335–1344.

Sawicka, K., Bushell, M., Spriggs, K.A., and Willis, A.E. (2008). Polypyrimidine-tract-binding protein: a multifunctional RNA-binding protein. Biochem. Soc. Trans. 36, 641–647.

Schmid, S.R., and Linder, P. (1992). D-E-A-D protein family of putative RNA helicases. Mol Microbiol 6, 283–291.

Schwartz, S. (2016). Cracking the epitranscriptome. RNA *22*, 169–174.

Selth, L.A., Gilbert, C., and Svejstrup, J.Q. (2009). RNA immunoprecipitation to determine RNA-protein associations in vivo. Cold Spring Harb Protoc *2009*, pdb.prot5234.

Shen, H.-B., and Chou, K.-C. (2007). Gpos-PLoc: an ensemble classifier for predicting subcellular localization of Gram-positive bacterial proteins. Protein Eng. Des. Sel. 20, 39–46.

Shen, H.-B., and Chou, K.-C. (2008). PseAAC: a flexible web server for generating various kinds of protein pseudo amino acid composition. Anal. Biochem. 373, 386–388.

Shen, H.-B., and Chou, K.-C. (2010). Gneg-mPLoc: a top-down strategy to enhance the quality of predicting subcellular localization of Gram-negative bacterial proteins. J. Theor. Biol. 264, 326–333.

Shi, Y. (2012). Alternative polyadenylation: new insights from global analyzes. RNA 18, 2105–2117.

Si, J., Cui, J., Cheng, J., and Wu, R. (2015). Computational Prediction of RNA-Binding Proteins and Binding Sites. Int J Mol Sci 16, 26303–26317.

Sigrist, C.J.A., de Castro, E., Cerutti, L., Cuche, B.A., Hulo, N., Bridge, A., Bougueleret, L., and Xenarios, I. (2013). New and continuing developments at PROSITE. Nucleic Acids Res. 41, D344–347.

Siomi, H., Matunis, M.J., Michael, W.M., and Dreyfuss, G. (1993). The pre-mRNA binding K protein contains a novel evolutionarily conserved motif. Nucleic Acids Res. *21*, 1193–1198.

Siomi, H., Choi, M., Siomi, M.C., Nussbaum, R.L., and Dreyfuss, G. (1994). Essential role for KH domains in RNA binding: impaired RNA binding by a mutation in the KH domain of FMR1 that causes fragile X syndrome. Cell *77*, 33–39.

Sivaraman, J., Iannuzzi, P., Cygler, M., and Matte, A. (2004). Crystal structure of the RluD pseudouridine synthase catalytic module, an enzyme that modifies 23S rRNA and is essential for normal cell growth of Escherichia coli. J. Mol. Biol. *335*, 87–101.

Smirnov, A., Förstner, K.U., Holmqvist, E., Otto, A., Günster, R., Becher, D., Reinhardt, R., and Vogel, J. (2016). Grad-seq guides the discovery of ProQ as a major small RNA-binding protein. Proc Natl Acad Sci U S A *113*, 11591–11596.

Spassov, D.S., and Jurecic, R. (2003). The PUF family of RNA-binding proteins: does evolutionarily conserved structure equal conserved function? IUBMB Life 55, 359–366.

Spitale, R.C., Flynn, R.A., Zhang, Q.C., Crisalli, P., Lee, B., Jung, J.-W., Kuchelmeister, H.Y., Batista, P.J., Torre, E.A., Kool, E.T., *et al.* (2015). Structural imprints in vivo decode RNA regulatory mechanisms. Nature *519*, 486–490.

Storz, G., Vogel, J., and Wassarman, K.M. (2011). Regulation by small RNAs in bacteria: expanding frontiers. Mol. Cell 43, 880–891.

Strein, C., Alleaume, A.-M., Rothbauer, U., Hentze, M.W., and Castello, A. (2014). A versatile assay for RNA-binding proteins in living cells. RNA 20, 721–731.

Shu, C.J., and Zhulin, I.B. (2002). ANTAR: an RNA-binding domain in transcription antitermination regulatory proteins. Trends Biochem. Sci. *27*, 3–5.

Squires, J.E., Patel, H.R., Nousch, M., Sibbritt, T., Humphreys, D.T., Parker, B.J., Suter, C.M., and Preiss, T. (2012). Widespread occurrence of 5-methylcytosine in human coding and non-coding RNA. Nucleic Acids Res. *40*, 5023–5033.

Su, E.C.-Y., Chiu, H.-S., Lo, A., Hwang, J.-K., Sung, T.-Y., and Hsu, W.-L. (2007). Protein subcellular localization prediction based on compartment-specific features and structure conservation. BMC Bioinformatics 8, 330.

Sutherland, J.M., Siddall, N.A., Hime, G.R., and McLaughlin, E.A. (2015). RNA binding proteins in spermatogenesis: an in depth focus on the Musashi family. Asian J. Androl. *17*, 529–536.

Symmons, M.F., Williams, M.G., Luisi, B.F., Jones, G.H., and Carpousis, A.J. (2002). Running rings around RNA: a superfamily of phosphate-dependent RNases. Trends Biochem. Sci. *27*, 11–18.

Tanford, C. (1968). Protein denaturation. Adv. Protein Chem. 23, 121–282.

Tatusov, R.L., Fedorova, N.D., Jackson, J.D., Jacobs, A.R., Kiryutin, B., Koonin, E.V., Krylov, D.M., Mazumder, R., Mekhedov, S.L., Nikolskaya, A.N., *et al.* (2003). The COG database: an updated version includes eukaryotes. BMC Bioinformatics 4, 41.

Tessier, S.N., Audas, T.E., Wu, C.-W., Lee, S., and Storey, K.B. (2014). The involvement of mRNA processing factors TIA-1, TIAR, and PABP-1 during mammalian hibernation. Cell Stress Chaperones 19, 813–825.

Tree, J.J., Granneman, S., McAteer, S.P., Tollervey, D., and Gally, D.L. (2014). Identification of Bacteriophage-Encoded Anti-sRNAs in Pathogenic Escherichia coli. Mol Cell *55*, 199–213.

UniProt Consortium (2015). UniProt: a hub for protein information. Nucleic Acids Res. 43, D204–212.

Ule, J., Jensen, K.B., Ruggiu, M., Mele, A., Ule, A., and Darnell, R.B. (2003). CLIP identifies Nova-regulated RNA networks in the brain. Science *302*, 1212–1215.

Upton, G., and Cook, I. (1996). Understanding Statistics (OUP Oxford).

Valverde, R., Edwards, L., and Regan, L. (2008). Structure and function of KH domains. FEBS J 275, 2712–2726.

Van Assche, E., Van Puyvelde, S., Vanderleyden, J., and Steenackers, H.P. (2015). RNA-binding proteins involved in post-transcriptional regulation in bacteria. Front Microbiol 6, 141.

Vanderweyde, T., Youmans, K., Liu-Yesucevitz, L., and Wolozin, B. (2013). The Role Stress Granules and RNA Binding Proteins in Neurodegeneration. Gerontology *59*.

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.

Walia, R.R., Xue, L.C., Wilkins, K., El-Manzalawy, Y., Dobbs, D., and Honavar, V. (2014). RNABindRPlus: A Predictor that Combines Machine Learning and Sequence Homology-Based Methods to Improve the Reliability of Predicted RNA-Binding Residues in Proteins. PLOS ONE 9, e97725.

Wang, C.-C., Fang, Y., Xiao, J., and Li, M. (2011). Identification of RNA-binding sites in proteins by integrating various sequence information. Amino Acids *40*, 239–248.

Wang, E.T., Taliaferro, J.M., Lee, J.-A., Sudhakaran, I.P., Rossoll, W., Gross, C., Moss, K.R., and Bassell, G.J. (2016). Dysregulation of mRNA Localization and Translation in Genetic Disease. J. Neurosci. *36*, 11418–11426.

Wang, K., Hu, L.-L., Shi, X.-H., Dong, Y.-S., Li, H.-P., and Wen, T.-Q. (2012). PSCL: predicting protein subcellular localization based on optimal functional domains. Protein Pept. Lett. 19, 15–22.

Wang, L., and Brown, S.J. (2006). BindN: a web-based tool for efficient prediction of DNA and RNA binding sites in amino acid sequences. Nucleic Acids Res. *34*, W243–248.

Wang, L., Huang, C., Yang, M.Q., and Yang, J.Y. (2010). BindN+ for accurate prediction of DNA and RNA-binding residues from protein sequence features. BMC Syst Biol *4 Suppl 1*, S3.

Wang, X., Zamore, P.D., and Hall, T.M. (2001). Crystal structure of a Pumilio homology domain. Mol Cell 7, 855–865.

Wang, X., McLachlan, J., Zamore, P.D., and Hall, T.M. (2002). Modular recognition of RNA by a human pumilio-homology domain. Cell 110, 501–512.

Wang, Z., Zhang, X.-C., Le, M.H., Xu, D., Stacey, G., and Cheng, J. (2011). A Protein Domain Co-Occurrence Network Approach for Predicting Protein Function and Inferring Species Phylogeny. PLoS One 6.

Wang, Z., Sun, X., Zhao, Y., Guo, X., Jiang, H., Li, H., and Gu, Z. (2015). Evolution of gene regulation during transcription and translation. Genome Biol Evol 7, 1155–1167.

Waterman, D.G., Ortiz-Lombardía, M., Fogg, M.J., Koonin, E.V., and Antson, A.A. (2006). Crystal structure of Bacillus anthracis ThiI, a tRNA-modifying enzyme containing the predicted RNA-binding THUMP domain. J. Mol. Biol. *356*, 97–110.

Waters, L.S., and Storz, G. (2009). Regulatory RNAs in bacteria. Cell *136*, 615–628.

Westermann, A.J., Förstner, K.U., Amman, F., Barquist, L., Chao, Y., Schulte, L.N., Müller, L., Reinhardt, R., Stadler, P.F., and Vogel, J. (2016). Dual RNA-Seq unveils noncoding RNA functions in host-pathogen interactions. Nature 529, 496–501.

Wistow, G. (1990). Cold shock and DNA binding. Nature 344, 823–824.

Wong, A.G., McBurney, K.L., Thompson, K.J., Stickney, L.M., and Mackie, G.A. (2013). S1 and KH domains of polynucleotide phosphorylase determine the efficiency of RNA binding and autoregulation. J. Bacteriol. *195*, 2021–2031.

Wower, J., Zwieb, C.W., Hoffman, D.W., and Wower, I.K. (2002). SmpB: a protein that binds to double-stranded segments in tmRNA and tRNA. Biochemistry 41, 8826–8836.

Wu, C.H., Nikolskaya, A., Huang, H., Yeh, L.-S.L., Natale, D.A., Vinayaka, C.R., Hu, Z.-Z., Mazumder, R., Kumar, S., Kourtesis, P., *et al.* (2004). PIRSF: family classification system at the Protein Information Resource. Nucleic Acids Res. 32, D112–114.

Xiong, D., Zeng, J., and Gong, H. (2015). RBRIdent: An algorithm for improved identification of RNA-binding residues in proteins from primary sequences. Proteins 83, 1068–1077.

Yakhnin, H., Yakhnin, A.V., and Babitzke, P. (2006). The trp RNA-binding attenuation protein (TRAP) of Bacillus subtilis regulates translation initiation of ycbK, a gene encoding a putative efflux protein, by blocking ribosome binding. Mol. Microbiol. *61*, 1252–1266.

Yang, Q., Gilmartin, G.M., and Doublié, S. (2010). Structural basis of UGUA recognition by the Nudix protein CFIm25 and implications for a regulatory role in mRNA 3' processing. Proc Natl Acad Sci U S A 107, 10062–10067.

Yang, X.-X., Deng, Z.-L., and Liu, R. (2014). RBRDetector: improved prediction of binding residues on RNA-binding protein structures using complementary feature- and template-based strategies. Proteins *82*, 2455–2471.

Yang, Y., Zhao, H., Wang, J., and Zhou, Y. (2014). SPOT-Seq-RNA: predicting protein-RNA complex structure and RNA-binding function by fold recognition and binding affinity prediction. Methods Mol. Biol. 1137, 119–130.

Yeung, K.Y., and Ruzzo, W.L. (2001). Principal component analysis for clustering gene expression data. Bioinformatics *17*, 763–774.

Yu, C.-S., Lin, C.-J., and Hwang, J.-K. (2004). Predicting subcellular localization of proteins for Gram-negative bacteria by support vector machines based on n-peptide compositions. Protein Sci. 13, 1402–1406.

Yu, C.-S., Cheng, C.-W., Su, W.-C., Chang, K.-C., Huang, S.-W., Hwang, J.-K., and Lu, C.-H. (2014). CELLO2GO: a web server for protein subCELlular LOcalization prediction with functional gene ontology annotation. PLoS ONE 9, e99368.

Yu, N.Y., Wagner, J.R., Laird, M.R., Melli, G., Rey, S., Lo, R., Dao, P., Sahinalp, S.C., Ester, M., Foster, L.J., *et al.* (2010). PSORTb 3.0: improved protein subcellular localization prediction with refined localization subcategories and predictive capabilities for all prokaryotes. Bioinformatics 26, 1608–1615.

Yuan, Y.-R., Pei, Y., Ma, J.-B., Kuryavyi, V., Zhadina, M., Meister, G., Chen, H.-Y., Dauter, Z., Tuschl, T., and Patel, D.J. (2005). Crystal structure of A. aeolicus argonaute, a site-specific DNA-guided endoribonuclease, provides insights into RISC-mediated mRNA cleavage. Mol. Cell 19, 405–419.

Zamore, P.D., Williamson, J.R., and Lehmann, R. (1997). The Pumilio protein binds RNA through a conserved domain that defines a new class of RNA-binding proteins. RNA *3*, 1421–1433.

Zamyatnin, A.A. (1972). Protein volume in solution. Prog. Biophys. Mol. Biol. 24, 107–123.

Zhao, J., Ohsumi, T.K., Kung, J.T., Ogawa, Y., Grau, D.J., Sarma, K., Song, J.J., Kingston, R.E., Borowsky, M., and Lee, J.T. (2010). Genome-wide identification of Polycomb-associated RNAs by RIP-seq. Mol Cell *40*, 939–953.

Zhao, H., Yang, Y., and Zhou, Y. (2010). Structure-based prediction of RNA-binding domains and RNA-binding sites and application to structural genomics targets. Nucl. Acids Res. gkq1266.

Zhao, H., Yang, Y., and Zhou, Y. (2011). Highly accurate and high-resolution function prediction of RNA binding proteins by fold recognition and binding affinity prediction. RNA Biol 8, 988–996.

Zhou, H., and Zhou, Y. (2005). SPARKS 2 and SP3 servers in CASP6. Proteins 61 Suppl 7, 152–156.

Zhu, J., and Krainer, A.R. (2000). Pre-mRNA splicing in the absence of an SR protein RS domain. Genes Dev. 14, 3166–3178.

Zvelebil, M.J., and Baum, J.O. (2008). Understanding Bioinformatics (Garland Science).

Appendix

Appendix Table 1

Proteins analyzed for the CoIP based screening of RBPs in *Salmonella* (discussed in the Chapter 3)

Gene names	Locus tag	UniProt ID	Protein names
cspA (control)	SL1344_3615	A0A0H3NMU5	Cold shock protein
cspB (control)	SL1344_1924	E1WGN1	Cold shock-like protein CspJ
csrA (control)	SL1344_2806	A0A0H3NF42	Carbon storage regulator
Hfq (control)	SL1344_4295	A0A0H3NIR6	RNA-binding protein Hfq
smpB (control)	SL1344_2660	A0A0H3NER3	SsrA-binding protein (Small protein B)
yhbJ (control)	SL1344_3295	A0A0H3NGG1	RNase adapter protein RapZ
aceA	SL1344_4119	A0A0H3NIN9	Isocitrate lyase
aceF	SL1344_0153	A0A0H3N7M4	Acetyltransferase component of pyruvate dehydrogenase complex (EC 2.3.1.12)
acnA	SL1344_1644	A0A0H3NC48	Aconitate hydratase (Aconitase) (EC 4.2.1.3)
acnB	SL1344_0159	A0A0H3N962	Aconitate hydratase B (EC 4.2.1.3) (EC 4.2.1.99) (2-methylisocitrate dehydratase)
aefA	SL1344_0471	A0A0H3N8H3	Integral membrane protein AefA
caiD	SL1344_0071	A0A0H3NH35	Carnitinyl-CoA dehydratase (EC 4.2.1.149) (Crotonobetainyl-CoA hydratase)
citC	SL1344_0612	A0A0H3N9D7	[Citrate [pro-3S]-lyase] ligase (EC 6.2.1.22)
comA	SL1344_3459	A0A0H3NMJ9	Competence gene-DNA binding and transport
cspC	SL1344_1766	A0A0H3NHM6	Cold shock-like protein CspC
cspD	SL1344_0882	A0A0H3NA23	Cold shock-like protein CspD
cspE	SL1344_0617	A0A0H3N9E2	Cold shock-like protein cspE
cspH	SL1344_1182	A0A0H3NFR3	Cold shock protein (CspH)
cysN	SL1344_2913	A0A0H3NR47	Sulfate adenylyltransferase subunit 1 (EC 2.7.7.4)
dbpA	SL1344_1586	A0A0H3NBN6	ATP-dependent RNA helicase DbpA (EC 3.6.4.13)
deaD	SL1344_3253	A0A0H3NI69	ATP-dependent RNA helicase DeaD (EC 3.6.4.13) (Cold-shock DEAD box protein A)
deoB	SL1344_4496	A0A0H3NQ37	Phosphopentomutase (EC 5.4.2.7) (Phosphodeoxyribomutase)
dinG	SL1344_0797	A0A0H3NEL5	Probable ATP-dependent helicase DinG
dnaG	SL1344_3184	A0A0H3NG55	DNA primase (EC 2.7.7.-)

dnaJ	SL1344_0013	A0A0H3N7I0	Chaperone protein DnaJ
dnaK	SL1344_0012	A0A0H3NCG3	Chaperone protein DnaK (HSP70) (Heat shock 70 kDa protein) (Heat shock protein 70)
efp	SL1344_4271	A0A0H3NPH5	Elongation factor P (EF-P)
engA	SL1344_2481	A0A0H3NEC5	GTPase Der (GTP-binding protein EngA)
engB	SL1344_3948	A0A0H3NNK6	Probable GTP-binding protein EngB
ffh	SL1344_2650	A0A0H3NEQ4	Signal recognition particle protein (Fifty-four homolog)
ftsY	SL1344_3536	A0A0H3NH43	Signal recognition particle receptor FtsY (SRP receptor)
gatB	SL1344_3232	A0A0H3NS78	PTS system, galactitol-specific IIB component
glk	SL1344_2371	A0A0H3NDY0	Glucokinase (EC 2.7.1.2) (Glucose kinase)
glnK	SL1344_0456	A0A0H3N8F4	Nitrogen regulatory protein P-II
gltD	SL1344_3303	A0A0H3NSF0	Glutamate synthase (NADPH) small chain
grxB	SL1344_1102	A0A0H3NAK6	Glutaredoxin 2
hflX	SL1344_4296	A0A0H3NPK0	GTPase HflX (GTP-binding protein HflX)
hscC	SL1344_0648	A0A0H3NE72	Chaperone heat shock protein
hydA	SL1344_2822	A0A0H3NQW9	Hydrogenase maturation protein
infB	SL1344_3259	A0A0H3NGC8	Translation initiation factor IF-2
lepA	SL1344_2545	A0A0H3NEE6	Elongation factor 4 (EF-4) (EC 3.6.5.n1) (Ribosomal back-translocase LepA)
leuC	SL1344_0111	A0A0H3NCP8	3-isopropylmalate dehydratase large subunit (EC 4.2.1.33)
lig	SL1344_2390	A0A0H3NFL2	DNA ligase (EC 6.5.1.2) (Polydeoxyribonucleotide synthase [NAD(+)])
lysC	SL1344_4156	A0A0H3NVC6	Aspartokinase (EC 2.7.2.4)
mreB	SL1344_3346	A0A0H3NMB9	Rod shape-determining protein
mutM	SL1344_3692	A0A0H3NTS0	Formamidopyrimidine-DNA glycosylase (Fapy-DNA glycosylase) (EC 3.2.2.23)
nagC	SL1344_0664	A0A0H3N9I6	N-acetylglucosamine repressor
nuc	SL1344_P2_0072	A0A0H3NXU0	Nuclease
nusA	SL1344_3260	A0A0H3NM77	Transcription termination/antitermination protein NusA
nusB	SL1344_0412	A0A0H3NDL2	N utilization substance protein B homolog (Protein NusB)
obgE	SL1344_3273	A0A0H3NSC0	GTPase Obg (EC 3.6.5.-) (GTP-binding protein Obg)
pheT	SL1344_1272	A0A0H3NB12	Phenylalanine--tRNA ligase beta subunit (EC 6.1.1.20)
pnp	SL1344_3255	A0A0H3NM73	Polyribonucleotide nucleotidyltransferase (EC 2.7.7.8) (Polynucleotide phosphorylase)
ppiA	SL1344_3439	A0A0H3NGU9	Peptidyl-prolyl cis-trans isomerase (PPIase) (EC 5.2.1.8)

ppiB	SL1344_0529	A0A0H3NDV8	Peptidyl-prolyl cis-trans isomerase (PPIase) (EC 5.2.1.8)
prfA	SL1344_1704	A0A0H3NCA7	Peptide chain release factor 1 (RF-1)
prfC	SL1344_4488	A0A0H3NWG0	Peptide chain release factor 3 (RF-3)
prfH	SL1344_0311	A0A0H3NHU7	Hypothetical peptide chain release factor
rbfA	SL1344_3258	A0A0H3NI71	Ribosome-binding factor A
rbsC	SL1344_3850	A0A0H3NHS2	High affinity ribose transport protein
recG	SL1344_3710	A0A0H3NN06	ATP-dependent DNA helicase RecG (EC 3.6.4.12)
res	SL1344_0353	A0A0H3NHZ1	Type III restriction-modification system enzyme (StyLTI) (Ec 3.1.21.5)
rfaH	SL1344_3931	A0A0H3NJR3	Transcription antitermination protein RfaH
rhlE	SL1344_0796	A0A0H3N9C8	ATP-dependent RNA helicase RhlE (EC 3.6.4.13)
rho	SL1344_3876	A0A0H3NJK0	Transcription termination factor Rho (EC 3.6.4.-) (ATP-dependent helicase Rho)
rlmL	SL1344_1001	A0A0H3NFA4	Ribosomal RNA large subunit methyltransferase K/L
rluA	SL1344_0096	A0A0H3NCN6	Pseudouridine synthase (EC 5.4.99.-)
rluB	SL1344_1175	A0A0H3NAE6	Pseudouridine synthase (EC 5.4.99.-)
rluC	SL1344_1123	A0A0H3NK80	Pseudouridine synthase (EC 5.4.99.-)
rluD	SL1344_2622	A0A0H3NEN7	Pseudouridine synthase (EC 5.4.99.-)
rph	SL1344_3700	A0A0H3NMZ6	Ribonuclease PH (RNase PH) (EC 2.7.7.56) (tRNA nucleotidyltransferase)
rplW	SL1344_3405	A0A0H3NIE5	50S ribosomal protein L23
rpmA	SL1344_3275	A0A0H3NGE3	50S ribosomal protein L27
rpsB	SL1344_0217	A0A0H3NCZ1	30S ribosomal protein S2
rpsC	SL1344_3401	A0A0H3NGS0	30S ribosomal protein S3
rpsD	SL1344_3383	A0A0H3NGP6	30S ribosomal protein S4
rrmA	SL1344_1764	A0A0H3NDQ7	rRNA guanine-N1-methyltransferase
sdhB	SL1344_0717	A0A0H3N9N1	Succinate dehydrogenase iron-sulfur subunit (EC 1.3.5.1)
secA	SL1344_0136	A0A0H3NCR8	Protein translocase subunit SecA
selB	SL1344_3647	A0A0H3NTL9	Selenocysteine-specific elongation factor
sgaB	SL1344_4317	A0A0H3NPM4	Hypothetical PTS system IIB protein
thdF	SL1344_3810	A0A0H3NU44	tRNA modification GTPase MnmE (EC 3.6.-.-)
thiI	SL1344_0419	A0A0H3NI88	tRNA sulfurtransferase (EC 2.8.1.4)
tolA	SL1344_0729	A0A0H3NAV1	TolA protein
tolB	SL1344_0730	A0A0H3N961	Protein TolB
traR	SL1344_P1_0024	A0A0H3NYC6	Conjugative transfer protein

truA	SL1344_2337	A0A0H3NJT6	tRNA pseudouridine synthase A (EC 5.4.99.12)
truD	SL1344_2907	A0A0H3NFD6	tRNA pseudouridine synthase D (EC 5.4.99.27)
tufB	SL1344_3412	A0A0G2PMS1	Elongation factor Tu (EF-Tu)
tyrS	SL1344_1381	A0A0H3NCU4	Tyrosine--tRNA ligase (EC 6.1.1.1) (Tyrosyl-tRNA synthetase)
ubiE	SL1344_3924	A0A0H3NHZ8	Ubiquinone/menaquinone biosynthesis C-methyltransferase UbiE (EC 2.1.1.163)
ubiG	SL1344_2245	A0A0H3NF69	Ubiquinone biosynthesis O-methyltransferase
uvrB	SL1344_0775	A0A0H3NB00	UvrABC system protein B (Protein UvrB) (Excinuclease ABC subunit B)
ybcJ	SL1344_0534	A0A0H3NDW2	Uncharacterized protein
ybdG	SL1344_0557	A0A0H3N990	Hypothetical membrane protein
ybhO	SL1344_0788	A0A0H3N9U5	Cardiolipin synthase B (CL synthase) (EC 2.7.8.-)
ybiO	SL1344_0802	A0A0H3NEM1	Hypothetical membrane protein
yceG	SL1344_1136	A0A0H3NFM2	Hypothetical secreted protein
ychF	SL1344_1712	A0A0H3NC19	Ribosome-binding ATPase YchF
yciL	SL1344_1651	A0A0H3NDH0	Pseudouridine synthase (EC 5.4.99.-)
yciM	SL1344_1640	A0A0H3NLQ9	Lipopolysaccharide assembly protein B
ydiL	SL1344_1297	A0A0H3NAV0	Uncharacterized protein
ydiS	SL1344_1286	A0A0H3NG36	Hypothetical electron transfer flavoprotein-quinone oxidoreductase
yegD	SL1344_2102	A0A0H3ND59	Uncharacterized protein
yejH	SL1344_2200	A0A0H3NDL4	Hypothetical helicase
yffB	SL1344_2445	A0A0H3NPH2	Uncharacterized protein
ygfZ	SL1344_3024	A0A0H3NRH2	tRNA-modifying protein YgfZ
yggB	SL1344_3043	A0A0H3NFT9	Hypothetical membrane protein
yggJ	SL1344_3069	A0A0H3NRQ8	Ribosomal RNA small subunit methyltransferase E (EC 2.1.1.193)
yhbY	SL1344_3270	A0A0H3NGD7	Uncharacterized protein
yhfA	SL1344_3432	A0A0H3NGV2	Uncharacterized protein
yhjS	SL1344_3587	A0A0H3NMT2	Uncharacterized protein
yigI	SL1344_3910	A0A0H3NUH7	Uncharacterized protein
yigZ	SL1344_3938	A0A0H3NNJ6	Uncharacterized protein
yjbC	SL1344_4128	A0A0H3NID0	Pseudouridine synthase (EC 5.4.99.-)
yjeQ	SL1344_4286	A0A0H3NPJ0	Putative ribosome biogenesis GTPase RsgA (EC 3.6.1.-)
yjhP	SL1344_4430	A0A0H3NJ47	Uncharacterized protein

ymdC	SL1344_1085	A0A0H3NA63	Cardiolipin synthase C (CL synthase) (EC 2.7.8.-)
ynaI	SL1344_1594	A0A0H3NLL2	Hypothetical membrane protein
yqcB	SL1344_2945	A0A0H3NFH6	Hypothetical RNA pseudouridylate synthase
yrdC	SL1344_3369	A0A0H3NSP6	Threonylcarbamoyl-AMP synthase (TC-AMP synthase) (EC 2.7.7.87)
SL1344_0081	SL1344_0081	A0A0H3N7S7	Hypothetical lipoprotein
SL1344_0961	SL1344_0961	A0A0H3NBP0	Antitermination Protein q
SL1344_1196	SL1344_1196	A0A0H3NAH7	Hypothetical membrane protein
SL1344_1516	SL1344_1516	A0A0H3NBH3	Hypothetical lipoprotein
SL1344_1949	SL1344_1949	A0A0H3NEF1	Hypothetical portal protein
SL1344_2163	SL1344_2163	A0A0H3NDB3	Hypothetical oxidoreductase
SL1344_2552A	SL1344_2552A	A0A0H3NPX7	Predicted bacteriophage protein
SL1344_2639	SL1344_2639	A0A0H3NGF0	Predicted bacteriophage protein
SL1344_2697	SL1344_2697	A0A0H3NEV8	Predicted bacteriophage protein
SL1344_2703	SL1344_2703	A0A0H3NQC8	Hypothetical bacteriophage protein
SL1344_2722	SL1344_2722	A0A0H3NKV8	Bacteriophage P4 DNA primase
SL1344_2929	SL1344_2929	A0A0H3NH71	Uncharacterized protein
SL1344_3646	SL1344_3646	A0A0H3NHD3	Hypothetical sugar kinase
SL1344_4015	SL1344_4015	A0A0H3NNS8	Hypothetical carbohydrate kinase
SL1344_4121	SL1344_4121	A0A0H3NKA5	Uncharacterized protein
SL1344_4127	SL1344_4127	A0A0H3NKA8	Uncharacterized protein

Appendix Table 2

Salmonella genes involved in flagellar assembly that are enriched in one of the CoIP samples in this study (discussed in the Chapter 3):

Gene names	Locus tag	UniProt id	Protein names
flgA	SL1344_1110	A0A0H3NA79	Flagella basal body P-ring formation protein
flgB	SL1344_1111	A0A0H3NFK4	Flagellar basal body rod protein FlgB
flgC	SL1344_1112	A0A0H3NAL6	Flagellar basal-body rod protein FlgC
flgD	SL1344_1113	A0A0H3NK70	Basal-body rod modification protein FlgD
flgE	SL1344_1114	A0A0H3NC32	Flagellar hook protein FlgE
flgF	SL1344_1115	A0A0H3NA83	Flagellar basal body protein
flgG	SL1344_1116	A0A0H3NFK8	Flagellar basal-body rod protein FlgG (Distal)
flgH	SL1344_1117	A0A0H3NAM2	Flagellar L-ring protein (Basal body L-ring)
flgI	SL1344_1118	A0A0H3NK77	Flagellar P-ring protein (Basal body P-ring)
flgK	SL1344_1120	A0A0H3NA87	Flagellar hook-associated protein 1 (HAP1)
flgL	SL1344_1121	A0A0H3NFL1	Flagellar hook-associated protein 3
flgM	SL1344_1109	A0A0H3NC27	Negative regulator of flagellin synthesis
flgN	SL1344_1108	A0A0H3NK64	Flagella synthesis protein FlgN
flhAa	SL1344_1848	A0A0H3NHV9	Flagellar biosynthesis protein FlhA
flhB	SL1344_1849	A0A0H3NCQ1	Flagellar biosynthetic protein FlhB
flhC	SL1344_1859	A0A0H3NCR1	Flagellar transcriptional regulator FlhC
flhD	SL1344_1860	A0A0H3NMG3	Flagellar transcriptional regulator FlhD
fliC	SL1344_1888	A0A0H3NMJ6	Flagellin
fliD	SL1344_1889	A0A0H3NE27	Flagellar hook-associated protein 2 (HAP2)
fliE	SL1344_1897	A0A0H3NCU5	Flagellar hook-basal body complex protein FliE
fliF	SL1344_1898	A0A0H3NML0	Flagellar M-ring protein
fliG	SL1344_1899	A0A0H3NE38	Flagellar motor switch protein FliG
fliH	SL1344_1900	A0A0H3NCJ5	Flagellar assembly protein FliH
fliI	SL1344_1901	A0A0H3NI18	Flagellum-specific ATP synthase
fliJ	SL1344_1902	A0A0H3NCU9	Flagellar FliJ protein
fliK	SL1344_1903	A0A0H3NML6	Flagellar hook-length control protein
fliM	SL1344_1905	A0A0H3NCK0	Flagellar motor switch protein FliM
fliN	SL1344_1906	A0A0H3NI23	Flagellar motor switch protein FliN

fliO	SL1344_1907	A0A0H3NCV8	Flagellar protein
fliP	SL1344_1908	A0A0H3NMM0	Flagellar biosynthetic protein FliP
fliQ	SL1344_1909	A0A0H3NE46	Flagellar biosynthetic protein FliQ
fliR	SL1344_1910	A0A0H3NCK7	Flagellar biosynthetic protein FliR
fliS	SL1344_1890	A0A0H3NCI3	Flagellar protein FliS
fliT	SL1344_1891	A0A0H3NI05	Flagellar protein FliT
fljB	SL1344_2756	A0A0H3NEZ8	Flagellin
motA	SL1344_1858	A0A0H3NHX2	Motility protein A
motB	SL1344_1857	A0A0H3NCF1	Motility protein B

Appendix Table 3

Salmonella genes involved in bacterial chemotaxis that are enriched in one of the CoIP samples in this study (discussed in the chapter-3):

Gene names	Locus tag	UniProt id	Protein names
cheA	SL1344_1856	A0A0H3NDZ5	Chemotaxis protein CheA
cheB	SL1344_1852	A0A0H3NCE6	Chemotaxis response regulator protein-glutamate methylesterase (EC 3.1.1.61)
cheM	SL1344_1854	A0A0H3NCQ6	Methyl-accepting chemotaxis protein II
cheR	SL1344_1853	A0A0H3NHW6	Chemotaxis protein methyltransferase (EC 2.1.1.80)
cheW	SL1344_1855	A0A0H3NMF8	Purine binding chemotaxis protein
cheY	SL1344_1851	A0A0H3NDY8	Chemotaxis protein CheY
cheZ	SL1344_1850	A0A0H3NMF2	Protein phosphatase CheZ (EC 3.1.3.-) (Chemotaxis)
dppA	SL1344_3596	A0A0H3NIX1	Periplasmic dipeptide transport protein
fliG	SL1344_1899	A0A0H3NE38	Flagellar motor switch protein FliG
fliM	SL1344_1905	A0A0H3NCK0	Flagellar motor switch protein FliM
fliN	SL1344_1906	A0A0H3NI23	Flagellar motor switch protein FliN
malE	SL1344_4166	A0A0H3NIS1	Periplasmic maltose-binding protein
mglB	SL1344_2167	A0A0H3NEZ1	D-galactose-binding periplasmic protein
motA	SL1344_1858	A0A0H3NHX2	Motility protein A
motB	SL1344_1857	A0A0H3NCF1	Motility protein B
rbsB	SL1344_3851	A0A0H3NU95	D-ribose-binding periplasmic protein
SL1344_2283	SL1344_2283	A0A0H3NDU3	Hypothetical receptor/regulator protein
SL1344_3126	SL1344_3126	A0A0H3NFZ7	Methyl-accepting chemotaxis protein
tcp	SL1344_3542	A0A0H3NMQ9	Methyl-accepting chemotaxis citrate transducer
trg	SL1344_1556	A0A0H3NGY0	Methyl-accepting chemotaxis protein III (Mcp-iii) (Ribose and galactose chemoreceptor protein)
tsr	SL1344_4464	A0A0H3NL96	Methyl-accepting chemotaxis protein

Curriculum vitae

Personal Details

Born: January 6th, 1987 in Ranchi, India
Nationality: India

Education

May 2012 – June 2016

PhD student in the groups of Prof. Dr. Jörg Vogel (RNA Biology Group), Dr. Ana Eulalio (Host RNA Metabolism) and Dr. Konrad U. Förstner (Core Unit Systems Medicine)
Institute for Molecular Infection Biology (IMIB)
University of Würzburg, Würzburg, Germany.
Co-supervisors: Prof. Thomas Dandekar (University of Würzburg) and Dr. Cynthia Sharma (IMIB)

October 2009 – February 2012

Masters of Science in Life Science Informatics
Bonn-Aachen International Center for Information Technology (B-IT)
University of Bonn. Bonn, Germany.
Co-supervisors: Prof. Martin Hoffman-Apitius (The Fraunhofer Institute – SCAI, Sankt Augustin), Prof. Heiko Schoof (Max Planck Institute for Plant Breeding Research, Köln) and Dr. Bianca H. Habermann (Max Planck Institute for Biology of Ageing, Köln)
Thesis title: "morFeus: a computational tool to detect remotely conserved orthologs using symmetrical best hits and orthology network scoring". (Grade: 2 ECTS)

July 2005 – June 2008

Bachelors of Science in Biotechnology
Peoples' Education Society Institute for Applied Sciences
Bangalore University. Bangalore, India.
Grades – 86 percent (in the scale of 100 percent)

Scientific internships

October 2010 – February 2012: Student Researcher in the group of Dr. Bianca H. Habermann**,** Max Planck Institute for Biology of Ageing (Bioinformatics Group)**.**
Köln, Germany

June 2010 – August 2010: Student assistant, Bayer AG Research Centre (Bioinformatics Department) under the supervision of Dr. Florian Sohler.
Wuppertal, Germany.

February 2010 – May 2010: Student assistant, B-IT, University of Bonn (Department of Life Science Informatics) under the supervision of Dr. Alexandra Reitelmann.
Bonn, Germany.

Würzburg, 2016

_____

Malvika Sharan

List of publications

Publications associated to the present work and beyond (during the PhD)

The listed publications are associated to my PhD work and co-operations therein:

- Sharan, M., Förstner, K.U., Eulalio, A., Vogel, J. (2017). APRICOT: an integrated computational pipeline for the sequence-based identification and characterization of RNA-binding proteins. *Nucleic Acids Research* (in press)
- Sunkavalli, U., Aguilar, C., Silva, R.J., Sharan, M., Cruz, A.R., Tawk, C., Maudet, C., Mano, M., Eulalio, A. (2017). Analysis of Host MicroRNA Function Uncovers a Role for miR-29b-2-5p in *Shigella* Capture by Filopodia. *PLOS Pathogens* (in press)
- Chowdhury, S.R., Reimer A., Sharan M., Kozjak-Pavlovic, V., Eulalio, A., Prusty, B.K., Fraunholz, M., Karunakaran, K., Rudel, T. (2017). *Chlamydia trachomatis* Prevents Mitochondrial Fragmentation Via The miR-30c-P53-DRP1 Axis. *The Journal of Cell Biology* (in press)
- Contributing author to the CAFA Consortium paper: Jiang, Y., *et al.* (2016). An expanded evaluation of protein function prediction methods shows an improvement in accuracy. *Genome Biol*ogy. *17*, 184.
- Contributing author to the crowdsourced paper: Budd, A., *et al.* (2015). Ten simple rules for organizing an unconference. *PLoS Computational Biology*. 11, e1003905.
- Sharan M. (2015). Computational identification and characterization of RNA-binding proteins in Salmonella (conference publication), *F1000 Research*. *ISMB-ECCB Conference*.
- Maudet, C., Mano, M., Sunkavalli, U., Sharan, M., Giacca, M., Förstner, K.U., and Eulalio, A. (2014). Functional high-throughput screening identifies the miR-15 microRNA family as cellular restriction factors for Salmonella infection. *Nature Communications 5*, 4718.

The manuscripts related to the PhD work that are submitted or in preparation

- Michaux, C., Holmqvist, E., Vasicek, E., Sharan, M., Barqist, L., Gunn, J.S., Vogel, J. CspC and CspE: two RNA-binding proteins involved in stress response and virulence in *Salmonella* Typhimurium

- Tawk, C., Sharan, M., Eulalio, A., Vogel, J. A paucity of RNA targeting potential amongst bacterial effector proteins.
- Maudet, C., Aguilar, C., Lisowski, C., Tawk, C., Sharan, M., Förstner, K.U., Mano, M., Eulalio, A. *Salmonella* decreases let-7i-3p expression to promote multiple steps of bacterial infection.
- Betancur, J.C., Sharan, M., Lopez, D. *et al.*, Cell-fate decision defines acute and chronic infection cell types in *Staphylococcus aureus.*

Previous publication

- Wagner, I., Volkmer, M., Sharan, M., Villaveces, J.M., Oswald, F., Surendranath, V., and Habermann, B.H. (2014). morFeus: a web-based program to detect remotely conserved orthologs using symmetrical best hits and orthology network scoring. BMC Bioinformatics *15*, 263.

Attended conferences and courses

Conferences/workshops and courses

1. Software Carpentry workshop for novices - Python, Unix & version control
   October 26-27, 2016
   Würzburg, Germany
   Co-organizer and trainer

2. EMBO Practical Course - Non-coding RNA in infection
   September 23, 2016
   Würzburg, Germany
   Trainer

3. Invited at the Bhabha Atomic Research Centre (BARC) Organized by Dr. R. Mukhopadhyay from Homi Bhabha Centre for Science Education, TIFR
   December 27, 2015
   Mumbai, India
   Invited talk

4. 10th International Symposiums organized by Graduate School of Life Science October 14-15, 2015
   Würzburg, Germany
   Poster presentation

5. Software Writing Skills for Your Research - Workshop for Novices by FOSTER, Helmholtz Centre Potsdam - GFZ
   September 23-25, 2015
   Potsdam, Germany
   Trainer

6. 23rd Intelligent Systems for Molecular Biology (ISMB) - 14th European Conference on Computational Biology (ECCB)
   July 9-11, 2015
   Dublin, Ireland.
   Selected Talk (won the best scientific presentation award)

7. RNA 2015, The Twentieth Annual Meeting of the RNA Society
   May 26-31, 2015
   Madison, WI, USA.
   Poster Presentation

8. FOSTER-UNESCO open science for doctoral schools
   April 23-24, 2015
   Paris, France

Representative of the Graduate School of Life Sciences, University of Würzburg at the meeting for designing strategy to promote open science in research.

9. 9th International Symposiums organized by Graduate School of Life Science October 14-15, 2014
   Würzburg, Germany
   Selected Talk

10. EMBO Practical Course - Non-coding RNA in infection
    October 12-18, 2014
    Würzburg, Germany
    Trainer

11. Mol-Micro Meeting Würzburg, Institute for Molecular Infection Biology
    May 7-9, 2014
    Würzbug, Germany.
    Poster presentation

12. EMBO Conference Series Visualizing biological data (VIZBI)
    March 5-7, 2014
    Heidelberg, Germany
    Virtual participation

13. 8th International Symposiums organized by Graduate School of Life Science October 9-10, 2013
    Würzburg, Germany
    Poster presentation (won the best poster award)

14. 3rd conference on Regulating with RNA in Bacteria
    June 4-8, 2013
    Würzburg, Germany
    Poster presentation

15. The 21st annual international conference on Intelligent Systems for Molecular Biology (ISMB) & 12th European Conference on Computational Biology (ECCB)
    July 21-23, 2013
    Berlin, Germany
    Poster presentation

16. 7th International Symposiums organized by Graduate School of Life Science October 16-17, 2012
    Würzburg, Germany
    Poster presentation

17. The 11th European Conference on Computational Biology
    September 9-12, 2012
    Basel, Switzerland
    Poster presentation

GSLS workshops on the transferable-skills

- Poster design
- Presentation skills
- Good scientific practice
- Scientific writing and publishing
- Developing your marketing strategy: Cover-letters and CVs
- Introduction to biotech industry, from idea to product
- Quality management and audit in biotech industry
- Project management in biotech industry
- Open access and copyright in science

Independent courses/workshops

- SciFund Challenge outreach training for scientists
  October 2015 - November 2015
  Certification course
    - Co-founded WUBSyB (Würzburg Unseminar in Bioinformatics and Systems Biology) to provide a networking platform for the bioinformatics enthusiasts in the University of Würzburg.
- Software Carpentry training course for trainers
  January 2015 - June 2015
  Certification course
    - Co-organized and trained at the Software carpentry workshops at the University of Würzburg, Helmholtz Centre - Potsdam, and European Molecular Biology Laboratory - Heidelberg.

Contributions by others

The work described in this doctoral thesis was carried out in the groups of Prof. Dr. Jörg Vogel and Dr. Ana Eulalio at the Institute for Molecular Infection Biology (IMIB) at the University of Würzburg, Würzburg, Germany. Several parts of the work discussed in this dissertation have been contributed by others and are indicated below.

Dr. Jörg Vogel conceived the idea of the projects concerning RBP screening in bacterial system and provided important guidance throughout the development of the work. Dr. Ana Eulalio provided important supervision and much needed scientific discussions throughout my PhD projects that helped tremendously in the development and completion of my thesis.

APRICOT software was developed under the supervision of Dr. Konrad Förstner who contributed considerably to the project by code development and other intellectual inputs.

APRICOT was developed as an independent pipeline as a result of specific requirements in different projects, which were shared with Caroline Taouk and Dr. Charlotte Michaux in the lab of Prof. Jörg Vogel. Their criticism and scientific support is indispensable in shaping of this project from the users' point of view.

Drs. Charlotte Michaux, Nora C. Marbaniang and Erik Holmqvist carried out the wet-lab experiments for the CoIP based screening of RBPs and their characterization in *Salmonella*, who also contributed to the finalization of RNA-Seq analysis methods. Dr. Victoria McParland sequenced cDNA libraries on the MiSeq platform.

The normalization approach for the analysis of the RIP-Seq samples was intensively discussed with Dr. Lars Barquist who suggested the modified TMM method used in this thesis.

The GO and KEGG pathway enrichment analysis method was discussed and developed with Dr. Lei Li. The idea of enrichment scores was developed with Dr. Charlotte Michaux.

The preliminary drafts of this thesis were kindly read and commented on by the members of my thesis committee, Dr. Konrad Förstner, Dr. Sarah Svensson, Dr. Charlotte Michaux, Dr. Sandy Pernitzsch and Caroline Taouk. Jens Hör and Konrad translated the abstract from English to German.

Acknowledgements

This thesis would not have been possible without the help, guidance and generosity of many people. I thank everyone who has in any way contributed to making my professional and personal life, truly a learning experience.

I would like to thank my supervisors Prof. Jörg Vogel and Dr. Ana Eulalio, for giving me a wonderful opportunity to be a part of their incredible scientific groups at the Institute of Molecular Infection Biology, University of Würzburg and perform this work under their expert supervision, which have been instrumental to my scientific development. I am thankful to Prof. Thomas Dandekar for his guidance by means of valuable feedback on my work and numerous motivating discussions. Thanks to Dr. Cynthia Sharma for providing her scientific inputs. My appreciation extends to Prof. Jörg Schultz for his willingness to chair the doctoral committee.

My incessant gratitude to Dr. Konrad Förstner for being an excellent teacher and mentor who not only patiently guided me in my PhD projects but also shared his enthusiasm for bioinformatics and open knowledge, and inspired me to find my niche within science, for which I will always be grateful.

I express my gratitude to all present and former colleagues in the Vogel group, Eulalio group and bioinformatics group – especially Dr. Charlotte Michaux, Caroline Taouk and Thorsten Bischler – for providing me continuous support and possibilities to learn in a competent scientific environment. I thank the Graduate School of Life Sciences of the University of Würzburg for giving me an opportunity to become a part of their exceptional interdisciplinary program. Thanks to Dr. Gabriel Blum-Oehler and Dr. Carolin Kisker for promoting women in science and supporting me in contributing to the program through Doctoral Researchers' Council.

I would like to acknowledge my colleagues and friends from the European Molecular Biology Laboratory, who inspired me to get through the last bits of my thesis. I extend my gratitude to Dr. Aidan Budd, for being a great mentor, motivator and a fantastic human being.

My heartfelt thanks to Ana Y., Anni, Antonio, Barbara, Disha, Ester, Gaurav, Gudrun, Hilde, Jorge, Jürgen, Kamini-KD, Nobi, Norman, Pixie, Rasha, Shiron, Sonika, Sophie, Tamara, and many friends from Ranchi, Bonn, Bangalore, and Kham-dinner-tables for their invaluable fellowship, solidarity and all they've given me through their kindness.

I am infinitely thankful to my wonderful confidants Niharika, Abhishek, Amrita and Phenol, for always having my back.

I am forever indebted to Maa and Papa, who have provided me their love and unwavering support beyond measures and whose value to me grows with age.

Finally, I wish for Zindagi and all the beautiful little ones to be nurtured in a world fostered by science and humanity.

Affidavit/declaration

I hereby confirm that my thesis titled “Bio-computational identification and characterization of RNA-binding proteins in bacteria” is the result of my own work. I did not receive any help or support from commercial consultants. All sources and/or materials applied are listed and specified in the thesis.

I confirm that this thesis has not yet been submitted as part of another examination process neither in identical nor in similar form.

Würzburg, 2016 _____

Malvika Sharan

Hiermit erkläre ich an Eides statt, die Dissertation „Bioinformatische Identifikation und Charakterisierung von RNA-bindenden Proteinen in Bakterien.“ eigenständig, d.h. insbesondere selbstständig und ohne Hilfe eines kommerziellen Promotionsberaters, angefertigt und keine anderen als die von mir angegebenen Quellen und Hilfsmittel verwendet zu haben.

Ich erkläre außerdem, dass die Dissertation weder in gleicher noch in ähnlicher Form bereits in einem anderen Prüfungsverfahren vorgelegen hat.

Würzburg, 2016 _____

Malvika Sharan