universität wien

DISSERTATION / DOCTORAL THESIS

Titel der Dissertation /Title of the Doctoral Thesis

„Bioinformatical analysis of RNA - protein interactions in AU-rich element mediated decay“

verfasst von / submitted by

Mag. rer. nat. Jörg Fallmann

angestrebter akademischer Grad / in partial fulfilment of the requirements for the degree of

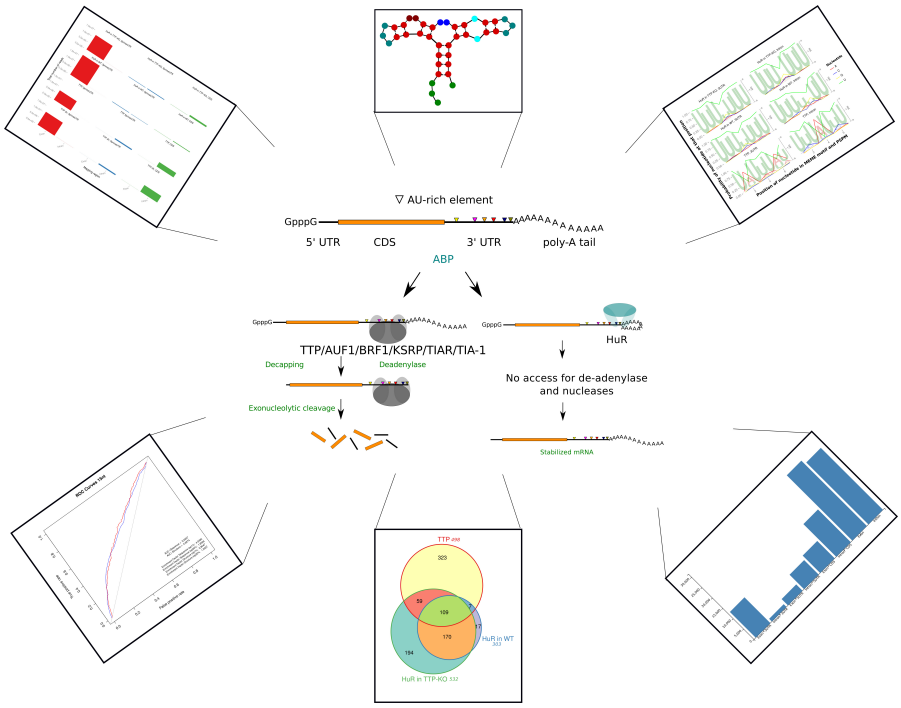
Doctor of Philosophy (PhD)

Wien, 2016 / Vienna 2016

Studienkennzahl lt. Studienblatt / degree programme code as it appears on the student record sheet:	A 794 685 490
Dissertationsgebiet lt. Studienblatt / field of study as it appears on the student record sheet:	Molekulare Biologie
Betreut von / Supervisor:	Univ.- Prof. Dipl.- Phys Dr. Ivo L. Hofacker

BIOINFORMATICAL ANALYSIS OF RNA - PROTEIN INTERACTIONS IN AU-RICH ELEMENT MEDIATED DECAY

JÖRG FALLMANN



TTP and HuR in AMD

Jörg Fallmann: *Bioinformatical analysis of RNA - protein interactions in AU-rich element mediated decay*

A thesis submitted in partial fulfillment of the requirements for the degree of

Ph.D.

at the

Institut für theoretische Chemie
Fakultät für Chemie
Universität Wien

ABSTRACT

This work is concerned with the interaction of RNA binding proteins (RPBs) and their RNA target sites. The main focus lies on proteins interacting with AU-rich elements (AREs), so called AU-rich binding proteins (ABPs). Their targets, AREs, are cis-acting RNA motifs found throughout genes in many higher organisms. Their function in AU-rich element mediated decay and the factors discriminating between active (bound) and inactive (unbound) status of such RNA elements is subject to this study.

We analyzed PAR-iCLIP data, identifying main targets of tristetraprolin (TTP) and HuR (ELAVL1) in LPS induced, primary bonemarrow derived marcophages (BMDMs) in mouse. The influence of RNA secondary structure on binding, cooperative vs. competitive behavior of RBPs, correlation with mRNA decay rates, over-represented binding motifs and differences between early and late immune-response binding of TTP are part of this thesis.

Furthermore, we compare our dataset to an over-expression study and investigate the potential of our predictions in mouse for their portability to human. Our previously published database for AREs in human and mouse (AREsite) was updated during this thesis and replaced by a new version (AREsite2), which contains annotations of AU-/GU- and U-rich elements in all genic regions (exons, introns, UTRs) of coding and non-coding genes in human, mouse, fruit fly, zebrafish and bandworm.

Together with an example analysis of AREsite2 derived data, this thesis presents a comprehensive analysis of RNA elements and their sequence and structure features crucial for functional RNA-RBP interaction.

ZUSAMMENFASSUNG

In dieser Arbeit wird die Interaktion von RNA bindenden Proteinen (RBPS) mit ihren Ziel-RNAs untersucht. Das Hauptaugenmerk liegt dabei auf Proteinen, die mit so genannten AU-reichen Elementen (ARE) interagieren, so genannte AU-reich bindende Proteine (ABPs). Der Interaktionspartner, AREs, sind cis-regulatorische Elemente welche entlang vieler Gene in höheren Organismen zu finden sind. Ihre Funktion im so genannten AU-rich element mediated decay (AMD), also dem gezielten Abbau von mRNA, und Faktoren mit denen aktive

(gebundene) von inaktiven (ungebundenen) Elementen voneinander unterscheidbar werden, sind Teil dieser Studie.

Zu den behandelten Themen gehört die Analyse von PAR-iCLIP Daten, bei der primäre Ziele von Tristetraprolin (TTP) und HuR (ELAVL1) in LPS induzierten, primären "bonemarrow derived" Makrophagen (BMDMs) aus Maus identifiziert werden. Des weiteren befasst sich diese Arbeit mit dem Einfluss von RNA Sekundärstrukturen auf Bindung, kooperatives bzw. kompetitives Bindungsverhalten von RBPs, Korrelation mit mRNA Abbauraten, überrepräsentierte Bindemotife und Unterschiede in der frühen und späten Phase der Immunantwort.

Ein Vergleich unseres Datensets mit einem Überexpressions Datenset und die Untersuchung des Potentials Ergebnisse von Maus auf Mensch zu übertragen sind weitere Punkte dieser Arbeit. AREsite, eine von uns publizierte Datenbank wurde im Verlauf dieser Arbeit überarbeitet und durch eine neue Version, AREsite2 ersetzt. Diese neue Datenbank enthält Annotationen von AU-/GU-/U- reichen Elementen in allen genischen Regionen (exons, introns, UTRs) von kodierenden und nicht-kodierenden Genen in Mensch, Maus, Fruchtfliege, Zebrafisch und Bandwurm.

Zusammen mit der Integration experimenteller Ergebnisse in AREsite2 präsentiert diese Arbeit eine umfassende Analyse von RNA Elementen und deren Sequenz- und Struktureigenschaften die für funktionelle RNA-RBP Interaktion eine Rolle spielen.

ACKNOWLEDGMENTS

First of all I want to thank my supervisor *Ivo Hofacker* for providing me with the opportunity of starting this thesis in his group. His dedication to all fields of RNA research and his guidance over the years were a constant source of information and inspiration. He made it possible for me to follow my own research interests with only few constraints. I also want to thank *Christoph Flamm* for a lot of interesting discussions, motivating words and catching enthusiasm for science in general. They were both great guides along my journey expanding my horizon beyond wetlab molecular biology, and still are.

Thanks also to *Peter Stadler*, who was an important guide during this thesis and an awesome new boss. My thanks also go to *Andrea Tanzer*, who spent a lot of time listening to my complaints and provided me with both, scientific and real life related guidance. My thanks also go to *Stephan Bernhart*, who is not only a scientific adviser and found time to proofread this thesis, but whom I also consider a dear colleague and friend, always ready to jump in when help is needed or beer to be drunk.

At this point I want to expand my thanks to the whole team at TBI *Florian Eggenhofer, Stefan Badelt, Peter Kerpedjiev, Michael Wolfinger, Fabian Amman, Ronny Lorenz, Roman Ochsenreiter, Stefan Hammer, Sven Findeiss and all the others* for the outstanding atmosphere they created, both at work and in private.

Andrea. The comfort you grant me and your support over the years have made my life richer in so many ways. Thank you for being at my side, sharing time, experiences and all the good and the not so good. You are the love of my life.

Family. Without you, none of this would have been possible. The support of my father *Klaus Fallmann* and my mother *Christa Fallmann* was always unconditionally and freely and they always made me feel independent enough to explore while providing the safety to come back. You do a great job as parents. My sister and best women *Anja Fallmann* makes me a proud brother, thank you all.

Last but not least I want to thank *Judith Ivansits, Gerlinde Aschauer and Nicola Wiskocil*. Juli for her endless support in all organizational matters, many shared meals and having a great time. Nicola for her commitment in all RNA-DK related stuff. Gerlinde for always helping out, even on short notice or long distance, many years of navigating diverse projects through organizational waters and participating in funny events.

FUNDING

This work was funded, in part, by the Austrian DK RNA program FG748004, and by the University of Vienna Research platform 323500 and by the Austrian Science Fund (FWF) grant SFB 43.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACRONYMS

ARE AU-rich element

ABP AU-rich binding protein

AMD AU-rich element mediated mRNA decay

TTP Tristetraprolin

HuR/ELAVL1 Hu-antigen R/Embryonic lethal abnormal vision like 1

RNA/DNA Ribonucleic acid/Deoxyribonucleic acid

mRNA messenger-RNA

sRNA small-RNA

ncRNA non-coding RNA

CLIP Crosslink and Immunoprecipitation

PAR-CLIP PhotoActivatable Ribonucleoside-enhanced CLIP

iCLIP Individual nucleotide resolution CLIP

PAR-iCLIP PhotoActivatable Ribonucleoside-enhanced Individual nucleotide resolution CLIP

REST REpresentational State Transfer

1 INTRODUCTION

1.1 RNA-PROTEIN INTERACTIONS

A key mechanism for the survival of each cell and as a consequence whole organisms, is tight and correct regulation of gene expression. This includes localization and turnover of all forms of nucleic acids and proteins in a cell. With growing complexity of higher organisms, so grows the complexity of regulatory mechanisms.

Vital part of this multi-facetted regulatory machinery is the interplay between ribo-nucleic acids (RNA), either coding (mRNA) or non-coding (ncRNA), and regulatory factors like proteins. Modulation of the spatial-temporal expression of RNA molecules is crucial for keeping the balance between synthesis (transcription), translation, transport and decay of mRNAs, ncRNAs and proteins. The extreme versatility of single RNA molecules in terms of sequence and structural features is reflected by an equal complexity of RNA binding domains and binding preferences of proteins.

A crucial layer in gene expression regulation is tight control of (post-) transcriptional fate by proteins that interact directly with cis-acting RNA motifs. Such mechanisms allow a fast response to environmental stress, infection or developmental necessities. With comprehensive understanding of key players and their interactome, it should become possible to develop strategies *e. g.* for medical applications.

Synthetic biology approaches that exploit such information, have a broad bandwidth of potential use cases. From synthetic proteins that regulate specific RNAs, over the modification of natural RNA binding sequences to the repair of non-functional natural RNAs by the introduction of synthetic RNA sequences that are designed for specific half-life effects *e. g.* combined with the CRISPR-Cas [54] system.

This work is focused on the interplay between RNA, tristetraprolin (TTP) and Hu antigen R (HuR) two proteins acting as main players in AU-rich element mediated decay (AMD), a key RNA half-life control mechanism in metazoa. TTP is crucial for the correct resolving of immunological response, mainly due to its RNA degrading function. HuR has been described as ubiquitous RNA interaction partner, with mainly stabilizing function. Utilizing a relatively new approach to identify RNA-protein interactions in a high-throughput manner, known as CLIP-Seq , we investigated key features for successful interaction and compared them for their predictive power for *in silico* determination of active RNA-protein interaction sites.

The next sections will summarize RNA and protein features that have been identified as crucial parts in the interplay between these two components of life.

1.1.1 *RNA binding proteins*

Hundreds of RNA binding proteins (RBPs) have been shown to be involved in virtually all aspects of (post-transcriptional) gene expression regulation (see *e.g.* [9, 21, 111]). Gerstberger et al. [38] present a manually curated collection of more than 1.500 RBPs in human, highlighting their vast number and thus potential for interaction and regulation. Regulation is usually initiated by direct interaction between RBP and target RNA, requiring more or less specific sequence motifs [30] and accessible binding sites. Many of the known RBPs seem to prefer single stranded binding regions, although some have been shown to interact with structured RNA sites [7].

The versatility of RBPs makes it hard to predict interaction partners from amino-acid sequence alone. However, it is usually not the whole protein that interacts with a target, but specific parts, known as domains. Such domains and their function in RNA-RBP interaction are topic of the next section.

1.1.2 *RNA binding domains*

Most forms of protein interactions, both with other proteins and nucleic acids, require specific conserved (tertiary) structures with certain amino-acid content, known as domains. RNA binding proteins (RBPs) contain RNA-binding domains (RBDs). Although these domains are very specific and employ different interaction mechanisms, they can share some features that enable RNA-protein interactions.

Protein domains can in general fold independently of the rest of the protein, and this fold plays a crucial role in the specificity of RNA recognition. Hydrogen bonds with the backbone, as well as specific interactions between nucleotides and amino-acids are common for sequence specific binding. Electrostatic interactions and stacking thereof contribute to the affinity of a protein to RNA. In general, protein domains are 35-90 amino acids in size and interact with a small stretch of nucleotides (3-5nt). To increase affinity and specificity they often work in combination with other RBDs in the same protein, thus highlighting the modularity of most proteins. However, some RBPs do not contain such a canonical RBD and remain to be investigated in more detail.

One distinguishes sequence-specific from sequence-unspecific binding. Sequence-specificity can be achieved via two strategies, i) hy-

drogen bonds between the protein backbone and RNA bases which are highly dependent of the protein fold (hydrophobic sidechains are looking towards the RNA, almost no intramolecular stacking of RNA bases instead intermolecular with sidechains, RNA bases not exposed to solvent, very rigid and specific scaffold) and ii) hydrogen bonds between amino-acid sidechains and RNA bases (intramolecular RNA base stacking, RNA bases exposed to solvent).

While the structural dependencies of i) make it nearly impossible to derive preferred target sequences without structure information (*e. g.* crystal structures), in case of ii) it should be possible to predict target preferences from amino-acid sequence information alone [7]. However, RBPs often employ a mix of strategies to bind their targets, which makes prediction of target sequences challenging in either case. For some RBPs specific sequence preferences are known, for others they can be guessed from the available RBDs.

However, due to the versatility of RBDs in combination, exact sequence preferences are sometimes hard to predict although general binding preferences are known. Some of the most common and best studied RNA-binding domains, are described in the following. Detailed reviews on this topic are presented in *e. g.* Auweter et al. [7], Cook et al. [29], Lunde et al. [88], McHugh et al. [96], which build the basis for the next sections.

1.1.2.1 *RNA recognition motif (RRM)*

The RNA recognition motif (RRM), also known as RNA binding domain (RBD) is one of the most common RBDs in eucaryotes, and found throughout many forms of life. It has been shown not only to be important for RNA/DNA-protein interactions, but also for protein-protein interactions [24].

Its two conserved motifs, RNP-1 and RNP2, consist of 8 and 6 mostly positively charged or aromatic amino acids. With a span of ~90 amino acids the RRM consists of a four-strand antiparallel β-sheet, the primary RNA-binding surface, packed against two α-helices, crucial for RRM-RRM interaction, in a β1α1β2β3α2β4 topology [24, 29], see fig. 1a. The mode of RRM-RNA recognition is highly versatile.

Canonical interactions require contacts between RNP-1 and RNP-2 of the β-sheet, while non-canonical interactions can involve loop-regions and N- or C- terminal RRM flanking amino acids. Contacts have been shown to involve 4 to 6 nucleotides, depending on the RRM interaction site. RRMs are often found in tandems or triplets and can be separated by a flexible linker, arranged as a continuous RNA-binding platform oriented in the same direction, or forming an RNA-binding cleft or can interact back to back, forcing the RNA to loop around the protein [29].

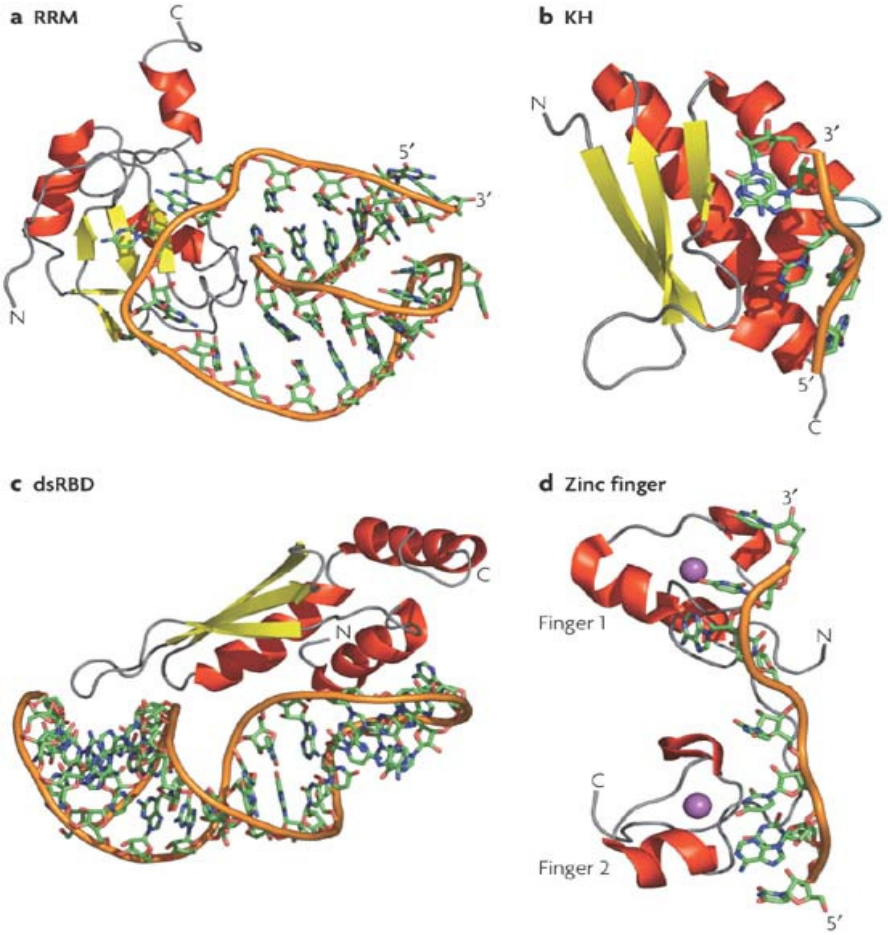


Figure 1: **RNA binding domains at a glance** , adopted from Lunde et al. [88]. a) RRM of human U1A in contact with single stranded RNA via the protein β-sheet and two loops b) KH domain of Nova-2 bound to short stretch of single stranded RNA via conserved GXXG in protein loop c) dsRBD of Rnt1 bound to RNA helix via conserved protein loop d) Zink fingers of TIS11d bound to AU-rich element via hydrogen bonds between protein backbone and RNA bases, zink finger α-helices and β-sheets are coordinated by a Zink atom

RRMs have been shown to interact with AU-rich elements (AREs) [145]. Proteins that bind RNA in the context of secondary structures may involve other forms of RRM-RNA interactions.

1.1.2.2 *Pumilio homology domain*

The Pumilio Homology Domain (PUM-HD) consists of 8 PUF (Pumilio and FBF) repeats, three α-helices each, of a 36 amino acid motif. The repeats pack together to form a right-handed super helix that binds RNA in the inner face (concave side) of the domain, while the outer face (convex side) mediates protein-protein interactions [146].

Each RNA nucleotide contacts two consecutive repeats, with the bases interacting with protein side-chains. Recognition of usually 8

target nucleotides by a repeat involves only a few well-conserved amino acids. Due to its well characterized recognition mechanism, a recognition code for the PUF repeat has been developed, allowing the design of custom PUM-HD domains which bind new motifs [23].

Increasing the number of repeats even allows to bind to nucleotide stretches longer than the usual 8 bases [29].

1.1.2.3 *K homology domain*

hnRNP K homology (KH) domains, are present in different folds in all domains of life. Two α-helices, a variable loop sequence containing a conserved GXXG (G glycine, X represents glycine, arginine or lysin) motif, and a β-strand build the RNA binding cleft of KH domains, see fig. 1b.

Often combined in multiples, or "augmented" each binding cleft can interact with four or more RNA bases to enhance binding affinity and specificity [29, 136].

1.1.2.4 *Double-stranded RNA binding domain*

Double-stranded RNA binding domains (dsRBDs) are 65-70 amino acids in size. They consists of two a-helices packed against a three-strand antiparallel b-sheet in a αβββα fold. RNA binding capabilities are derived from both α-helices and a loop region between two of the β-strands, see fig. 1c.

They play a role in post-transcriptional regulation, RNA editing, RNA processing and RNA localization. The deep and narrow major groove of the A-form RNA double helix leads to the assumption that dsRBDs are not sequence specific, but recognize the double-stranded RNA (dsRNA) shape only. Mismatches or bulges in RNA duplexes may effect target specificity by dsRBD containing proteins.

However, sequence-specific contacts between the protein and the minor groove have been shown, *e. g.* in the case of ADAR2 [29, 92].

1.1.2.5 *Zinc fingers*

Zinc fingers are a large and diverse class of protein domains. They act as DNA-, RNA-, and protein-binding domains, coordinating zinc as common property, see fig. 1d. Their three-dimensional structures vary and their evolutionary origins may be independent.

Mechanisms behind recognition of and interaction with individual targets of zinc finger proteins remain to be understood completely. However some classes tend to follow a trend, C2H2 zinc fingers are usually DNA binding, CCCH zinc fingers are primarily single-stranded RNA binding. CCHC zinc knuckles bind RNA in viral and metazoan proteins. Metazoan CCHC zinc knuckles show RNA binding in the context of proteins that also contain another RBD.

Stacking interactions and hydrogen bonds are crucial for RNA recognition by CCCH proteins. Such stacks are formed intermolecularly, between the RNA and backbone atoms, highlighting the importance of the domain fold for RNA recognition [7, 21, 29].

1.1.3 *Protein binding elements*

Interaction between RNA and proteins depends on both, a protein domain that recognizes the target RNA, as well as RNA elements that are recognizable by the protein domain. While a number of protein domains have been studied and characterized in detail, RNA elements crucial for successful interaction are in general less well studied. There are a number of RNA characteristics than can promote protein interaction, both on sequence and structure level.

This section focuses on elements for single-stranded RNA binding proteins, which in general require the target RNA sequence to be unpaired or in a defined structural context like the loop section of a hairpin loop. The most prominent metazoan RNA sequence elements in this category are AU-rich elements (AREs) and GU- or U- rich elements (GU/UREs), which are in the focus of this thesis. These sequence motifs are found in many coding and non-coding RNAs, throughout genic regions including UTRs as well as CDS and intronic and exonic regions. Although other RNA elements, like *e. g.* the PUF repeats, important for RNA recognition by PUM-HD proteins exist, this work is only concerned about AU/GU/U-rich elements.

1.1.3.1 *ARE*

AU-rich elements (AREs) are cis-acting sequence elements that have been categorized as A/U flanked versions of the AUUUA core motif, with a total of 5-13nts. These motifs are bound by AU-rich binding proteins (ABPs). Three classes of AREs have been defined. Class I AREs consist of several dispersed copies of the AUUUA motif within U-rich regions, class II AREs consist of at least 2 overlapping UUAU-UUA(U/A)(U/A) nonamers and class III AREs are U-rich regions that do NOT contain the AUUUA pentamer.

The strict motif definition can be explained by early experiments that focused on single, well characterized targets of ABPs, which often contain repeated versions of such elements. Recent high-throughput experiments show that this strict definition is not feasible for all targets and interacting sites, where often variations of these motifs are found that also contain guanine or cytosin.

The best studied ABPs tristetraprolin (TTP), HuR and Auf1 contain zinc fingers domains and/or RRMs interacting directly with AREs. The latter have been shown to play an important role in RNA half-life control [13, 39].

1.1.3.2 *URE*

U-rich elements (UREs) are defined as RNA stretches of 7-9 nt that consist mostly of uridine. In contrast to AREs, these motif definition allows for some variance in composition, often cytosin or guanin can be found in such elements.

HuR is one of the most prominent URE binders, and like AREs, UREs have been shown to influence RNA turnover [22, 63].

1.1.3.3 *GRE*

GU-rich elements (GREs), are similar to AREs, but have guanin flanking U-rich stretches of RNA. GRE-binding proteins like CUG-binding protein 1 (CUGBP1), have been reported to influence RNA half-life, similar to ABPs and UBPs [74, 140].

1.2 RNA CYCLE OF LIFE

Complex organisms require complex regulatory mechanisms. In the last years, RNA has gained more and more attention as crucial part of gene expression control. From the synthesis of RNA by RNA-polymerase transcription from its DNA template, processing via various complexes from de-/capping, de-/adenylation, splicing, modification to the final un-/stable molecule, RNA undergoes a vast number of processing steps.

All sorts of RNA, be it transfer RNA or messenger RNA, micro RNA or long non-coding RNA, are affected by these or other processes. A major part of RNA half-life control is performed directly by proteins or protein-complexes, which are subject of this thesis.

The amount of available RNA is always depending on the ratio between synthesis and decay. The former has been target of investigations for a long time, and a lot is known about mechanisms and regulation of RNA synthesis. Decay on the other hand is not as well

investigated, although there is no reason why decay should be less regulated than synthesis.

1.2.1 *RNA synthesis*

RNA is synthesized from a DNA via RNA-polymerase, a class of enzymes found in procaryotes as well as eucaryotes. While bacteria only have one kind of RNA-polymerase, eucaryotic organisms express three types, each one required for the synthesis of specific forms of RNA.

RNA-polymerase binds DNA at certain positions, unwinds the DNA double-helix and initiates transcription by joining the first RNA nucleotides complementary to the DNA template. The freshly synthesized RNA strand is elongated until a stop signal is reached, the nascent RNA molecule is released and RNA polymerase detaches from the DNA template.

This very brief and simple description of transcription (see figure 2) already shows multiple stages where RNA synthesis can be regulated, from DNA accessibility for RNA-polymerase, to initiation factors, proof-reading mechanisms, and many more. However, at the end of transcription, a nascent RNA molecule is available for further processing.

1.2.2 *RNA maturation and processing*

In procaryotes, a freshly synthesized RNA is already available for protein translation by ribosomes or other processes. In eucaryotes, the nascent RNA is still in the nucleus and has to either be exported, or undergo a series of processing steps before being translated or functioning as ncRNAs.

For eucaryotic mRNAs, several processing steps, from 5′-end capping over splicing to 3′-end poly-adenylation (see fig. 2) ensure that the correct messenger RNA is being synthesized and released into the cytosol. At each of the many processing steps, tight regulation of RNA fate is ensured.

1.2.3 *RNA half-life control*

Once a mature RNA molecule has been synthesized, a series of RNA half-life control mechanisms ensure correct turnover. While nonsense-mediated decay (NMD) ensures that only correctly processed RNAs (not containing premature termination codons PTC) are retained in the cytosol, various control points ensure correct translation of mR-

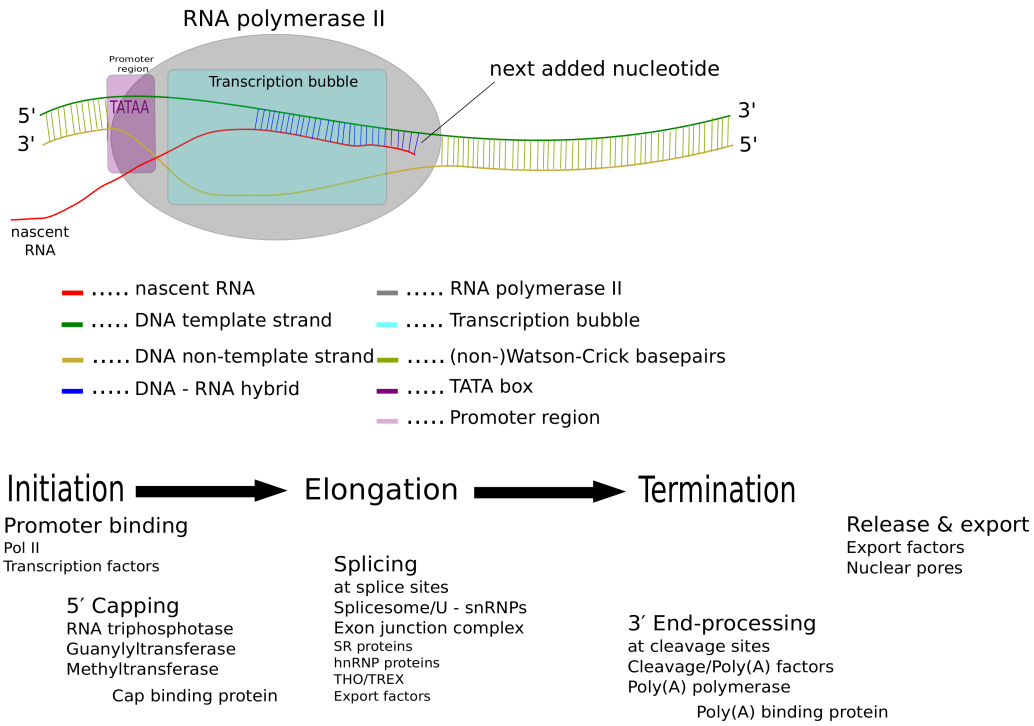


Figure 2: **Transcription of messenger RNA (mRNA) and timeline** This schematic covers the minimum of required factors for trancription. The timeline below indicates crucial steps in transcription and involved enzymes, which highlights available stages for regulation.

NAs and further processing of other RNAs (see Garneau et al. [36] for a review), known as post-transcriptional regulation.

Among the most abundant mechanisms in metazoa is RNA half-life-control by ABPs, known as ARE mediated decay (AMD), figure 3 shows mRNA decay pathways including AMD, which is shown in more detail in figure 4.

Such half-life control mechanisms are crucial for cell fate, as the correct amount of available RNA is key to processes like differentiation, response to environmental stress, proliferation and many more. Understanding these mechanisms will help to identify novel ways of disease treatment.

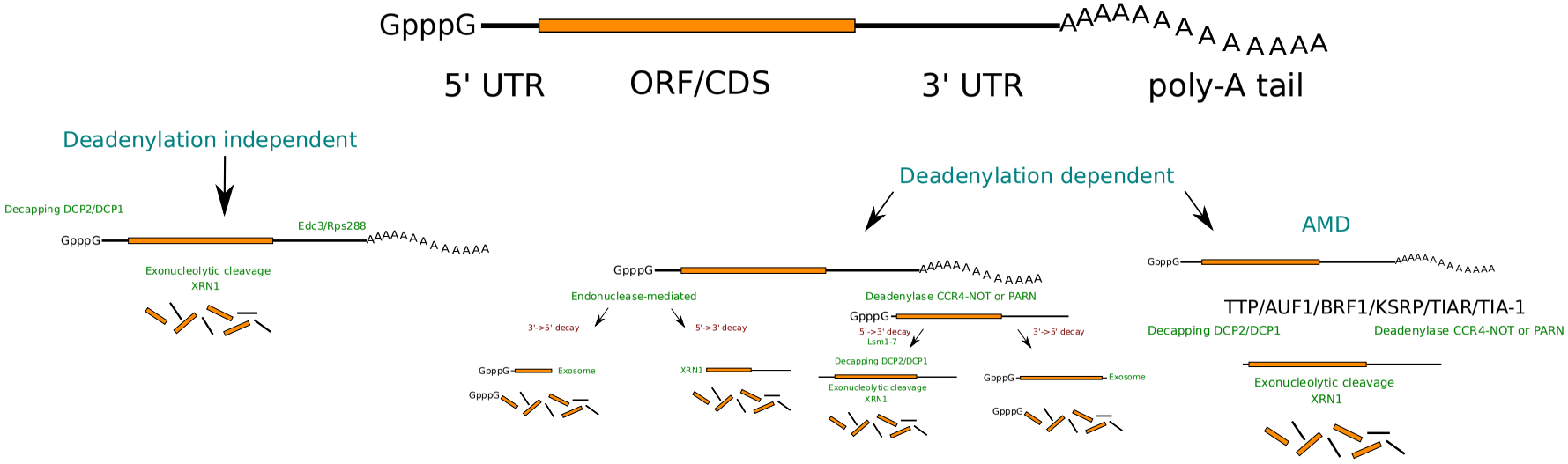


Figure 3: **Overview of mRNA decay mechanisms in eucaryotes** Decay mechanisms are divided into deadenylation dependend and independend. Independent decay relies on 5′ decapping and exonucleolytic cleavage, while independent decay can start on both ends and be either endo- or exonucleolytic.

1.2.4 *AMD*

AU-rich element mediated decay is a mechanism of RNA half-life-control that requires direct interaction of trans-acting ABPs with their cis-acting RNA target site (AREs) and the subsequent recruiting of mRNA processing factors.

Although AMD can be seen as one of the key mechanisms controlling gene expression, our understanding of its details is still limited. Upon interaction, a RNA-protein complex is formed, that initializes mRNA decapping, deadenylation and subsequently RNA decay (see fig. 4).

However, for certain ABPs like HuR (ELAV1) it has been shown, that their interaction with RNA can prevent decay, thus stabilizing the transcript, although the exact mechanisms remain unknown. Stabilizing effects of ABPs do not necessarily have to be active, they can also come in form of antagonistic binding effects, blocking other RBPs, miRNAs or yet unidentified destabilizing factors [13, 19, 126, 141].

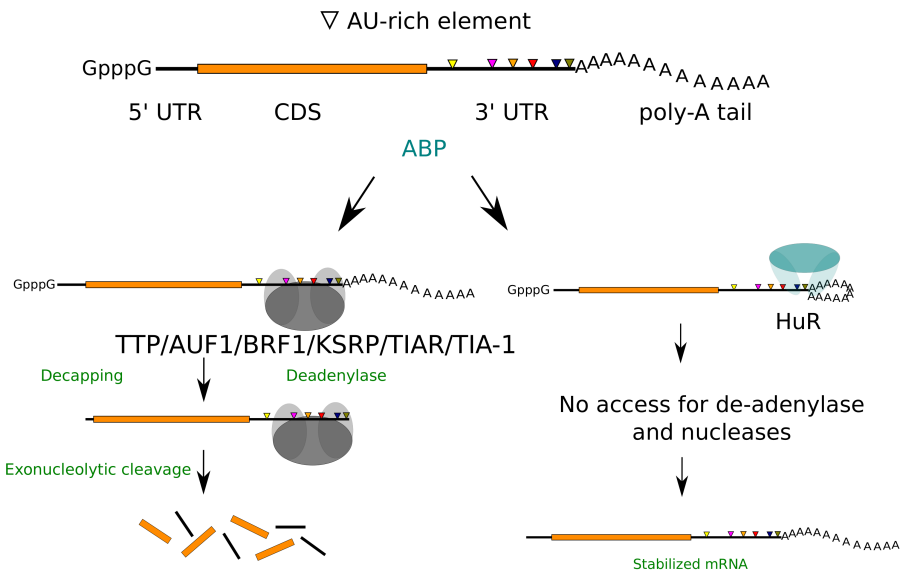


Figure 4: **AU-rich element mediated mRNA decay** AMD is a special type of deadenylase dependent mRNA decay. Upon interaction of i) TTP with mRNA, the latter becomes deadenylated, decapped and is subject to exonucleolytic decay or ii) HuR with mRNA leads to protection of the poly-A tail from deadenylase, which stabilizes the mRNA

1.3 AU-RICH BINDING PROTEINS

This class of proteins was named after their preference for AREs. This classification however is rather old, and more recent experiments often show that proteins classified as ABPs, often have other preferences as well, *e. g.* HuR which would be better classified as U-rich binding protein, due to its preference for UREs. The current explosion of CLIP-Seq and other experiments and the new insights into binding preferences generated have the potential to change this outworn type of classification and lead to some less strict and more flexible categories.

1.3.1 *TTP*

Tristetraprolin (TTP) is a CCCH tandem zinc-finger protein known to interact with single-stranded RNA molecules. It has a destabilizing effect on its RNA targets. Predominantly found in the cytoplasm, it has been shown to be able to shuttle into the nucleus. TTP preferentially binds the core UUAUUUAUU of class II AREs, promotes deadenylation and degradation of RNAs.

Its tandem zinc fingers can bind to adjacent 5′ -UAUU- 3′ subsites on the single-stranded target RNA, potentially interacting with two RNA copies at once [12, 52]. A crystal structure of TTP zinc fingers bound to a synthetic strand of mRNA can be seen in figure 5A.

TTPs binding preferences have been investigated in detail for some of its targets, but not in a systematic, transcriptome wide way until recently. Although partial crystal structures of TTP zinc fingers exist, so far no structure of the whole protein is available.

As one of the major regulators of mRNA stability, especially during immune-stress response, TTP was one of the two AMD related proteins studied in this thesis. While it's expression levels under normal conditions are low, induced immunological stress *e. g.* via lipopolysaccharid (LPS) induction has a strong effect on TTP expression.

TTP is itself regulated by phosphatases and kinases, which are believed to modify the carboxyterminal domain (CTD) of TTP [118], thereby regulating its activity. However, exact mechanisms for this regulation are not known, and not topic of this study.

1.3.2 *HuR*

HuR (human antigen R) preferentially binds the nonamer NNUUNNUUU. It can shuttle between the cytoplasm and the nucleus and contains three RRMs. Two N-terminal binding to ARE motifs and the C-terminal motif binding to poly-A tails, which potentially prevents deadenylase from interaction and stabilizing the RNA-protein complex [12].

A sketch of the RRM1 of HuR is shown in figure 5B. In contrast to TTP, HuR is a well studied protein, although its binding preferences and mode of action are still not fully understood. As potential counterpart to TTP, HuR-CLIP-Seq data in $TTP^{+/+}$ and TTP^{-} cells were analyzed during this thesis, with focus on direct cooperative or antagonistic effects.

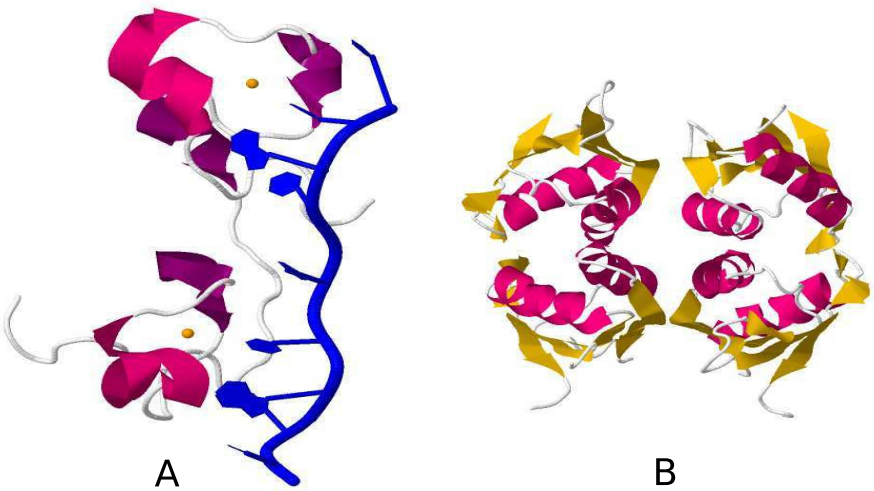


Figure 5: **AU-rich binding proteins TTP and HuR** A) Sketch of a tandem CCCH-zinc finger domain of TTP in contact with a class II AU-rich element (PDB: 1RGO). The nucleic acid is shown in blue and the two zinc ligands are shown in orange.
B) Sketch of the first two tandem RRMs (RRM1) of HuR that is known to bind ARE motifs(PDB-ID: 3HI9)

1.3.3 *Auf1*

Auf1 (ARE/Poly(U)-binding/degradation factor), also known as heterogeneous nuclear ribonucleoprotein D (hnRNPD) is a RNA destabilizing factor. It preferentially binds to U-rich elements, but has also been shown to interact with poly-A and AU-rich elements [42]. Four isoforms of Auf1 have been identified, all containing two RRMs for RNA-interaction.

Its diverse binding preferences and the number of isoforms available, make Auf1 a complex interaction partner for RNA, which is

known to be less sequence sensitive than other RBPs. It is well possible that different isoforms have different effect on RNA targets, maybe even stabilizing their target.

All three ABPs discussed here, have shown the ability to shuttle between nucleus and cytosol, potentially interacting with RNAs from synthesis to decay, underlining their important role as regulators of gene expression and cell fate. Their exact binding mechanics and parameters that influence their de-/stabilizing effects remain yet to be investigated.

1.3.4 *TTP and HuR in AMD*

Both TTP and HuR, can be found to interact or compete with each other for single stranded target sites, either having an agonistic or antagonistic effect on the stability of their target RNAs providing the cell with a fast response mechanism to environmental or developmental conditions. Studies comparing both ABPs have reported a wide range of interaction between the two, from only marginal overlap to vast amounts of shared binding sites [86, 102, 120].

Overlap of target sites in immunostimulated primary mouse macrophages and influence on RNA half-life under physiological conditions is one of the major points addressed in this thesis. As both ABPs are known to be able to shuttle between nucleus and cytosol, enabling them to act in an auto-regulatory manner on their own mRNA or pre-mRNA respectively, their cooperativity/competition is an even more intriguing target for further investigation.

1.4 IDENTIFYING RNA-PROTEIN INTERACTIONS

RNA-protein interactions are a central part of the complex interactome of organisms and as such their interplay and underlying mechanisms are not simple to investigate. It is not trivial to distinguish true binding sites from sites sharing sequence and/or structure features, especially as interaction is not necessarily functional. Often proteins interact with their target not only at specific sites, but in a probing manner known as diffusional search [97], further complicating interaction analysis.

In principle, investigating interactions requires some knowledge of the interaction partners, sometimes in form of specific probes, antibodies, cell-types or substrates. Early methods to investigate protein-nucleic acid interactions were footprinting techniques, where enzymes or chemicals are used to digest or modify nucleic acid unprotected from the protein body, resulting in a “footprint” of the protein on its target.

Electrophoretic Mobility Shift Assay (EMSA), utilizes band-shift during gel electrophoresis between bound and unbound nucleic acid to identify if interaction happens. These methods, are useful to predict interaction on nucleotide level and footprints of specific proteins. However, due to the specifics of the experiments they are unsuitable for large scale experiments without detailed knowledge of interaction partners.

Experimental methods for the characterization of RNA-RBP interactions can generally be broken down into *in vitro* assays, which means free from other interacting factors and under experimental conditions and *in vivo* approaches which capture a snapshot of RBP binding to RNAs at natural expression levels or after induction.

RNA-centric methods use mass spectrometry to potentially identify all RBPs bound to an RNA of interest. Protein-centric methods focus on a specific protein of interest which is crosslinked via UV-light or formaldehyde to its target, which is then co-immunoprecipitated with the protein.

While RNA-centric methods allow the identification of novel RBP interactions, protein-centric methods require knowledge of the protein of interest and specific antibodies for the IP. However, protein-centric methods can easily be applied in a high-throughput manner and require lower amounts of starting material.

This section contains a general description of the two former mentioned approaches based on the reviews from Cook et al. [29] and McHugh et al. [96], while the high-throughput part will be described in more detail later on.

1.4.1 *RNA-centric methods*

RNA-centric methods purify an RNA of interest and identify interacting proteins or protein complexes via methods like mass spectrometry (MS). This allows the detection of novel RBPs, as well as RBPs for which antibodies are hard to come by. However, detection of RNA interacting RBPs via RNA-centric methods requires the purification of enough protein mass, which requires an extraordinary amount of starting material [10]. In contrast to nucleic acids purified protein cannot be amplified, which makes RNA-centric methods challenging for low abundancy RNAs and proteins.

in vitro approaches (see fig. 6), use a synthetic RNA bait to capture RBPs from cellular extracts, while the technically more challenging *in vivo* approaches (see fig. 7), preserve the context of competing or assisting RNA-protein interactions.

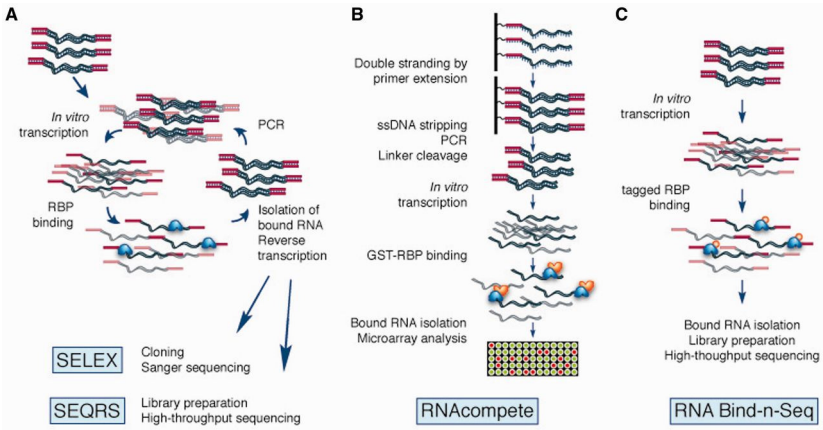


Figure 6: **RBP in vitro assays**, adopted from Cook et al. [29] with permision. A) SELEX and SEQRS where RNAs undergo several rounds of binding and amplification and resulting pools are analyzed via sequencing at the end (SELEX) or after each round (SEQRS) B) RNAcompete assays binding affinity of proteins with designed RNAs on microarray C) RNA Bind-n-Seq sequences protein concentration dependent amounts of bound RNAs

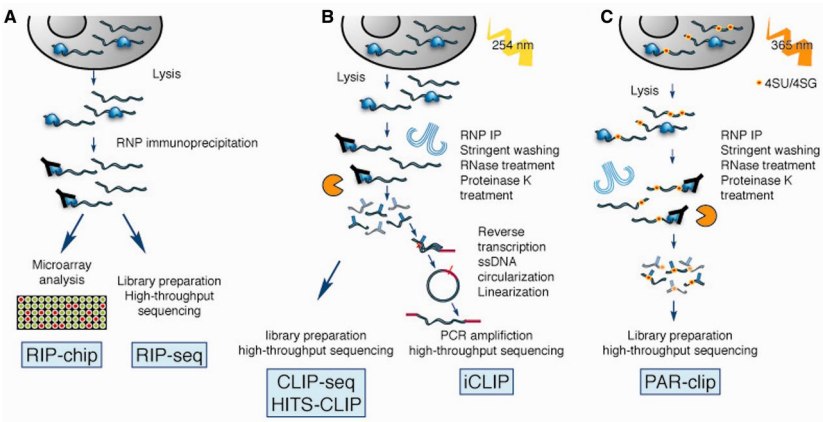


Figure 7: **RBP in vivo assays**, adopted from Cook et al. [29] with permision. A) RIP-chip and RIP-seq assay bound RNAs after IP B) CLIP-Seq methods, co-IP of bound RNAs after UV-crosslinking and identification of targets via NGS C) PAR-CLIP first treats cells with modified U or G nucleoside analogs for higher crosslinking efficiency

Purification of RNA-RBP complexes from extract or lysate is a challenging procedure. Re-association or formation of non-specific interactions can occur under native conditions, which can be prevented using stringent, high-salt wash conditions. Denaturing condition on the other hand require crosslinked complexes, with crosslinking strategies varying from low efficiency to technically challenging for later identification of purified peptides.

Tag based purification of the RNA of interest is an alternative where the RNA is tagged with naturally occurring interaction sites, *e. g.* the MS2 viral coat protein binding RNA stem loop structure [20] or artificially designed RNA aptamers. Such tags are then used for purification via resin- or solid support- coupled MS2 protein, streptavidin or histidin. Depending on the tag, elution of complexes from resin or solid support can either be conducted by boiling in SDS, which will dissolve all specific and unspecific complexes, or via specific elusion, *e. g.* via biotin in excess in case of streptavidin. The more specific the elution, the higher the detection sensitivity.

MS is used for the identification of interacting complexes. Non-quantitative methods compare purified proteins from the RNA of interest and a control. Total protein stained samples are separated by gel electrophoresis and bands present only in the sample but not the control are extracted and proteins identified by MS. Whole proteome methods require quantitative MS, where all proteins in the sample can be identified, including those not visible on a gel. Non-specific proteins can be excluded by analyzing a control. Metabolic labeling, chemical labeling, or spike-ins can be used to tag proteins for MS analysis. Isotopes of the proteins are compared to provide direct quantification of peptide ratios from sample and control to discriminate true binding from non-specific interactions [29, 96].

1.4.2 *Protein-centric methods*

Protein-centric methods require access to specific purification methods for the protein of interest *in vivo* , or a way to express a tagged version *in vitro* . Most common are antibodies which allow immunoprecipitation (IP) of the protein. Consequentially, the quality and specificity of the antibody has a huge impact on the reliability of the results. Co-immunoprecipitated RNA is then reverse transcribed into cDNA, PCR amplified and sequenced to identify interaction partners. PCR amplification of protein bound RNA allows to detect interaction partners even when less starting material is available, in contrast to RNA-centric methods.

1.4.2.1 *In vitro*

A common method for the identification of binding motifs for RBPs is Systematic Evolution of Ligands by EXponential enrichment (SELEX) [65, 156]. Randomized RNA oligos are incubated with an RBP of interest, followed by reverse transcription (RT) of bound RNAs. Resulting cDNA is then PCR amplified and in vitro transcribed. This process is repeated, each time increasing the amount of high-affinity binding sites in the pool. Sanger sequencing is then applied to the enriched sequences to finally identify the binding motifs.

SELEX enriches high-affinity motifs, which may exclude some functional binding sites with lower affinity, and it is not possible to deduce quantitative affinity information for sub-optimal motifs. SELEX in combination with high-throughput sequencing is known as SEQRS, where resulting pools are sequenced after each round of selection, which gives some information on sub-optimal motifs as well.

Binding specificities of RBPs can be probed by RNAcompete [110], where a purified Glutathione S-transferase(GST)-tagged RBP of interest is incubated with a pool of ~40 nt long RNAs which are designed to represent all 9-mers in a compact way. RNA is incubated in excess, so that molecules compete for a limited amount of protein binding sites, which allows to deduct relative affinity from abundance after a single-step selection. Eluted RNAs are then hybridized to a microarray for detection.

A comparable approach is RNA-bind'n-seq [72], where the protein of interest is expressed in vitro. Different concentrations of protein are then incubated with random RNAs of length 40nt. After IP and sequencing, the ratio of protein concentration and bound RNA can be used to determine real dissociation constants (K_d) from such experiments, while simultaneously allowing to infer simple secondary structure preferences, as 40nt is long enough to preserve basic structures.

However, neither SELEX, nor SEQRS, nor RNAcompete are able to detect complex secondary structure constraints of interactions, as the RNA oligos used are too small for structures more complex than simple hairpins. RNAcompete oligos are even designed to prevent complex structures, to represent all single-stranded 9-mers in the most compact way. Only RNA Bind-n-Seq has the potential to be used for RNA secondary structure probing and allows the deduction of off-rates in context of single nucleotide mutations, which enables binding affinity decomposition into sequence and structure features.

1.4.2.2 *In vivo*

For *in vivo* methods, native and denaturating purification methods have to be distinguished. Native purification methods, known as RNA immunoprecipitation (RIP), preserve physiological conditions and thus also native RNA-protein and protein-protein complexes during purification . However, during purification, the protein can interact with RNAs which are not present in the same cell compartment and could not interact *in vivo* . Furthermore, unspecific interactions with RNAs that are highly abundant in cells, *e. g.* rRNAs, can interfere and mask specific interactions with low-abundancy targets.

Denaturing methods for RNA-protein interactions crosslink the protein of interest to the target RNA. Crosslinking takes a snapshot of current interactions, thus preventing the interaction of protein with RNA in a non-*in vivo* manner in later steps of purification. Crosslinking with short wavelength UV light creates covalent bonds between aromatic amino acids of the protein and RNA nucleotides in close proximity without crosslinking proteins with other proteins.

Followed by antibody-purification, this methods are known as CLIP (crosslink and immunoprecipitation) [132]. RNA-protein complexes are denatured in sodiumdodecylsulfate (SDS) and retrieved from SDS gel after purification.

Several types of CLIP procedures have been proposed [68], *e. g.* HITS-CLIP (HIgh-Throughput Sequencing of RNA isolated by CrossLinking ImmunoPrecipitation) [153], iCLIP (Individual-nucleotide resolution CLIP) [67], PAR-CLIP (PhotoActivatable-Ribonucleoside-enhanced CrossLinking and ImmunoPrecipitation) [45] (see fig. 8) to name the most common ones. Together with eCLIP (enhanced CLIP) [138], ir-CLIP (infrared CLIP) [154], hiCLIP (RNA hybrid and individual-nucleotide resolution ultraviolet crosslinking and immunoprecipitation) [127], CLASH (crosslinking, ligation, and sequencing of hybrids) [70] and CRAC (cross-linking and analysis of cDNAs) [41] a bandwidth of experimental designs are available, each with certain advantages and limitations.

They all rely on the same principle, crosslinking protein residues and adjacent nucleotides with UV light, varying details to achieve different outcomes. As an example, in PAR-CLIP nucleotide analogs like thio-uridine or thio-guanine are introduced into the cell as crosslinking agents. This circumvents the otherwise low efficiency of UV-crosslinking at 254nm, as the nucleotide analogs can be crosslinked with long-wave UV light (365nm), but it works only with cultured cells which readily utilize the nucleotide analogs. The biochemistry behind UV-crosslinking is still not completely understood, so that it remains unclear which interactions might be missed completely or to what extent.

What is known, is that reverse transcriptase (RT) misreads crosslinked nucleotides with a higher as usual rate, or drops off completely. PAR-CLIP exploits this behaviour, as the introduced nucleotide analogs in case of thio-uridine are misinterpreted as guanines by the RT, which introduces T2C transitions in the resulting sequencing reads. These transitions can be used to pinpoint interaction sites. iCLIP, as another example, takes advantage of the fact that the amino acid tag left at the crosslink site after proteinase digestion often causes termination of reverse transcription to pinpoint the exact interaction site.

However, CLIP-Seq variants are not bias free. Certain nucleotides and amino acids are preferentially crosslinked by UV-light, and crosslink efficiency varies between proteins, just as the incorporation rate of nucleotide analogs, which varies between cell types and is considered low. PAR-CLIP only creates bonds at the nucleotide analog, so tags will be enriched at locations with several repeats of that base. Furthermore, crosslinking only occurs at sites where nucleotides and aromatic side chains are in close proximity, so even if a nucleotide analog is incorporated a crosslink only happens if the analog is close to the actual binding site. A conceptual problem can arise if the interacting amino-acid side chains are not aromatic, thus can not be crosslinked, and therefore simply not be seen in a CLIP-Seq experiment.

Formaldehyde crosslinking can be an alternative, but requires more elaborate purification methods. Affinity tag coupled proteins can be used to purify in denaturing conditions (guanidine or urea) and work with UV as well as formaldehyde crosslink protocols. However, a tagged version of the protein of interest has to be available for expression in the studied cell line.

Protein occupancy profiling is a technique similar to CLIP, except that RNPs are not immunoprecipitated but purified via oligo (dT) beads or biotinylation. Cross-linking is also possible with formaldehyde, followed by RNA digestion, cross-link reversion and sequencing of the purified RNA [96].

An ongoing challenge is the extraction of target RNAs and specific protein binding sites by *in silico* methods, which follows such experiments and is discussed in the next section.

1.5 NGS

Next generation sequencing (NGS), is a high-throughput method following most of the experimental RNA-RBP interaction detection approaches discussed so far. The combination with high-throughput methods allows to identify a huge number of interactions at once, as well as comparisons between different experimental setups, *e. g.* knockout-wildtype, timelines, concentration dependencies, and more. Initially,

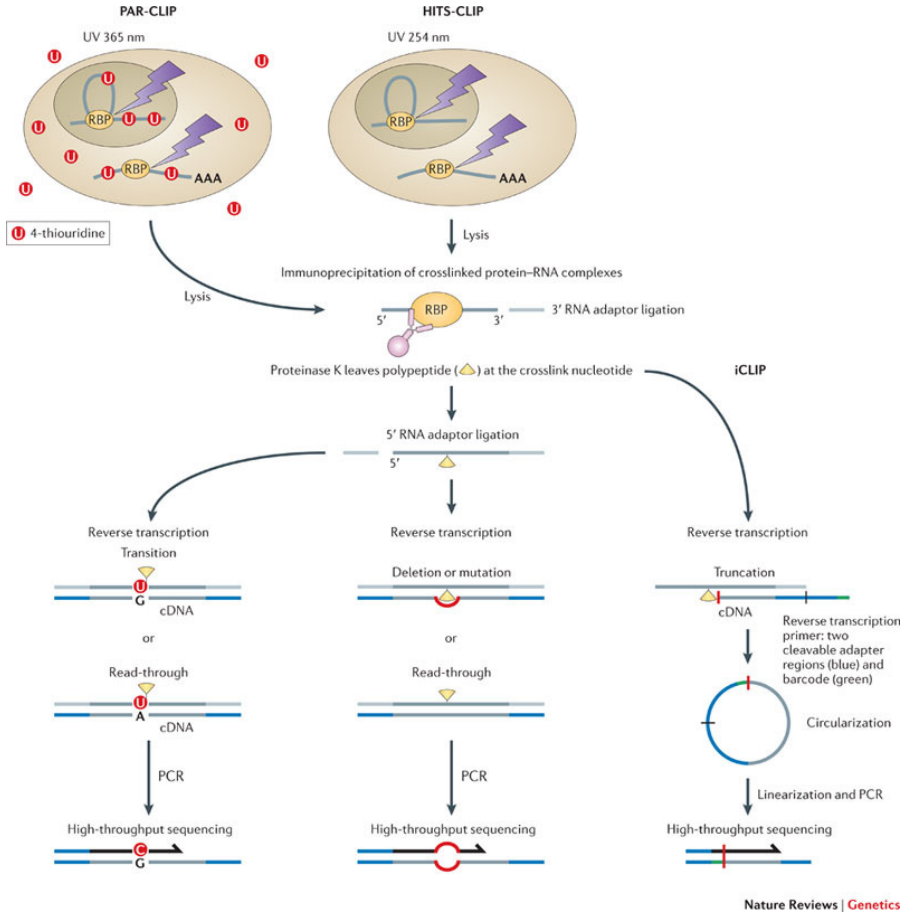


Figure 8: **Most used CLIP-Seq methods and their differences**, adopted from König et al. [68]. High-throughput sequencing of RNA isolated by ultraviolet (UV) crosslinking and immunoprecipitation (HITS-CLIP), photoactivatable ribonucleoside-enhanced CLIP (PAR-CLIP) and individual nucleotide resolution CLIP (iCLIP). CLIP methods differ by UV-light wavelength used for crosslinking, introduction of nucleoside analogs (PAR-CLIP), introduction of transitions (PAR-CLIP) or deletions (HITS-CLIP) during reverse transcription at crosslink sites and addition of 5′ and 3′ adapter separately (PAR- and HITS-CLIP) or in one step (iCLIP via circularization and linearization)

sequencing of short DNA stretches was a costly process. It became affordable with the introduction of first parallel sequencing strategies. The ever decreasing cost for sequencing experiments and the increase in precision and throughput make NGS a great and feasible resource for all kinds of experiments, even large consortia projects like ENCODE [27] or the 1000 genomes project [28].

Several strategies of read amplification and sequence identification have successfully been applied in high-throughput life sciences. Amplification strategies can be grouped into Emulsion PCR and Bridge amplification, or are not needed at all (single molecule sequencing

strategy). The actual sequence identification step is either sequencing by synthesis, or sequencing by ligation. Both forms can use fluorescent dyed nucleotide-triphosphates where the incorporation of bases triggers pyrophosphate release, emitting flashes of light unique for each base, which are recorded to infer the sequence of the newly synthesized DNA generated from the cDNA sample.

Alternatively, the nucleotides are added sequentially, so that signal can only be generated by those currently in solution (*e. g.* ION-torrent). Sequencing by ligation uses DNA ligase instead of DNA polymerase, which is used as key enzyme during sequencing by synthesis. Di-base primers are ligated to the nascent DNA strand in multiple rounds for the former, while single nucleotide tri-phosphates are incorporated for the latter.

The main differences between today's most popular sequencing platforms are the maximum length of reads that can be sequenced, whether single- or paired-end read sequencing is possible, the latter allowing long sequences to be read from both ends, or if mate-pair sequencing is feasible, where two reads with a given linker size can be retrieved from each sequence in the sample. Further differences are the number of reads that can be analyzed in parallel and the sequencing speed. It is safe to assume that none of the platforms will be best for all types of analysis, so that the right combination of sequencing platform and experimental setup varies from case to case (see *e. g.* Solonenko et al. [124]).

For a recent overview on sequencing platforms and perspectives see *e. g.* Mardis [89], Pareek et al. [106] and table 1.

Besides the obvious use for experiments like RNA-Seq or RNA-RBP studies, NGS methods can be applied to any experiment which can be measured in terms of sequencing reads. Thus, NGS data processing and analysis is a rather young field with a great potential.

1.5.1 *The general workflow of NGS experiments*

Fragmentation of target DNA/RNA is performed, either by sonication or digestion by restriction enzymes. Fragmentation is also the first critical step in an NGS experiment, as enzymatic digestion can bias the outcome due to enzyme cut site preferences, while fragmentation by sonication is a more random process, which renders it bias free but also less reproducible than enzymatic digestion [66]. The main advantage of enzymatic digestion for CLIP-Seq is the smaller fragment size which allows higher resolution of binding sites compared to lager sonication fragments [123]. Best would be to mix different enzymes for digestion, thus preventing potential biases.

Table 1: **NGS technologies compared**, adopted from Mardis [90], Metzker [99], van Dijk et al. [137] ∗ Average read-lengths.‡ Fragment run. §Mate-pair run. Frag, fragment; GA, Genome Analyzer; GS, Genome Sequencer; MP, mate-pair; N/A, not available; NGS, next-generation sequencing; PS, pyrosequencing; RT, reversible terminator; SBL, sequencing by ligation; SOLiD, support oligonucleotide ligation detection.

Platform	Library/template preparation	NGS chemistry	Read length (bases)	Run time (days)
Roche/454's GS FLX Titanium	Frag, MP/emPCR	PS	330∗	<0.5
Illumina/ Solexa's GAII	Frag, MP/solid-phase	RTs	75 or 100	4‡, 9§
Life/APG's SOLiD 3	Frag, MP/emPCR	Cleavable probe SBL	50	7‡, 14§
Pacific Biosciences	Frag only/single molecule	Real-time	964∗	N/A
Ion Torrent	Frag, MP/emPCR,single molecule	Semiconductor, pH change	200	<0.5

Gb per run	Pros	Cons	Biological applications
0.45	Longer reads improve mapping in repetitive regions; low run times	Low throughput; high reagent cost; high error rates in homopolymer repeats	Bacterial and insect genome de novo assemblies; medium scale (<3 Mb) exome capture; 16S in metagenomics
18‡, 35§	Currently the most widely used platform in the field; highest throughput; many compatible protocols	Sequence complexity needed; low multiplexing capability of samples	Variant discovery by whole-genome resequencing or whole-exome capture; gene discovery in metagenomics
30‡, 50§	Two-base encoding provides inherent error correction	Long run times and short reads	Variant discovery by whole-genome resequencing or whole-exome capture; gene discovery in metagenomics
N/A	Reads up to 20kb and more; low run time; single molecule runs	High cost; high error rates; low throughput; limited range of applications	Full-length transcriptome sequencing; complements other resequencing efforts in discovering large structural variants and haplotype blocks
1	No optical scanning, no fluorescent nucleotides; low run time; many applications	High error rates in homopolymer repeats	Transcriptome/Exome sequencing; bacterial and insect sequencing; targeted sequencing of genes

cDNA is then reverse-transcribed if necessary, as the sequencing reaction works on DNA not RNA. Adapter sequences for polymerase chain reaction (PCR) and sequencing are ligated to the cDNA fragments. PCR enrichment ensures sequencing depth which means that enough copies of each fragment in the mix are available for sequencing. The parallel processing of multiple experiments/conditions/replicates on one sequencing lane can be realized by multiplexing, where samples are barcoded and then mixed to decrease sequencing costs, which on the other hand also decreases sequencing depth.

The sequencing reaction is performed in a highly parallel fashion, precise steps depend on the applied protocol. Resulting reads are then converted from signal to sequence if necessary. Samples are demultiplexed, quality controlled and post-processed according to experiment and sequencing protocol.

1.5.2 *RNA-Seq*

High-throughput quantification of transcript levels is possible with RNA-Seq [147]. More and more replacing microarrays, RNA-Seq has become the method of choice for the evaluation of gene expression on transcript level. The manifold variations of this technique allow direct assessment of RNA expression, half-life, modifications, genotyping, genome assembly and many more. As expression levels of RNAs have a direct influence on the occurrence and effect of RNA-RBP interactions, RNA-Seq results were incorporated into this study. In brief, during RNA-Seq cellular RNAs are extracted, fragmented, converted to cDNA, adapters are aligned and cDNA is sequenced.

Resulting reads are aligned to the genome or transcriptome to generate exonic, intronic, junction or other reads, depending on the protocol followed. RNA-Seq allows to calculate expression values for genes from read counts, predict (novel) transcript-isoforms, investigate single-nucleotide-polymorphisms (SNPs) or other modifications, differential expression profiles and much more. The analysis of RNA-Seq reads differs from *e. g.* CLIP-Seq generated reads in many ways, for a recent publication on RNA-Seq analysis see *e. g.* Conesa et al. [26].

The general workflow for read analysis is the same than for other NGS experiments. However, due to the vast amount of reads necessary for reliable analysis of such experiments, the amount of data produced and time consumed is a factor to be considered during the planing phase, as are the demands for adequate computational hardware.

1.5.3 *CLIP-Seq*

As mentioned above (see 1.4.2.2), immunoprecipitation techniques require a specific antibody against the protein of choice, which is used to extract the latter from cell-lysates. These protein-centric methods can be quantified via microarrays (RIP-ChIP) or NGS techniques (RIP-Seq). Crosslink-IP (CLIP) techniques further require a crosslinking agent, *e. g.* UV-light, to create a covalent bond between the protein of interest and its RNA target. Coupled with NGS methods, CLIP-Seq techniques have gained growing attention as method for RNA-RBP interaction studies. Depending on the kind of CLIP technique used (iCLIP, HITS-CLIP, Par-CLIP etc.), downstream analysis requires specific algorithms to filter signal from noise.

The general workflow is as follows: If required and possible, cells are cultured in medium with nucleotide analogs like thio-uridine. They are then exposed to UV-light, which creates covalent bonds between nucleotides respectively their analogs, and juxtaposed amino-acids of

interacting proteins. The crosslinked RNA is then co-immuno-precipitated with the protein of interest via specific antibodies.

Cells are then treated with DNAse and RNAse for fragmentation of RNA that is not protected by the interacting protein footprint. Proteinase is added to digest protein residues up to a small portion of amino-acids that stay covalently bound to their target RNA. Reverse transcriptase synthesizes cDNA from the RNA templates, readily incorporating transitions or deletions when protein remnants are encountered, or simply dropping off the template. After sequencing, these mutations can be used to identify interaction sites, and if available, transitions can be used to distinguish signal from noise.

However, CLIP-Seq signal is a qualitative measure for RBP targets, and a quantitative measure only for the relative amount of protein titert by it. It indicates which RNAs are targets and which are not, but gives no quantitative measure of binding strength or affinity, as the number of crosslinks depends on a series of factors. For one there is the number of protein molecules available for binding. Ideally, MS studies accompany such an experiment, so that the amount of available RBP is known, but this is very expensive and not at all standard. However, as most CLIP-Seq studies focus on one protein, it should be save to assume a comparable amount of available protein for binding throughout replicates and conditions.

So although not known, the real amount of protein in the cell is not as important as the amount of RNA available for binding, which is the second factor. Highly abundant RNAs will likely produce more CLIP signal than spurious RNAs, independent of the binding affinities. As mentioned before, RNA-Seq experiments can be used to quantify the relative amount of a specific RNA in comparison to the rest, which can be used to normalize CLIP signal.

However, one has to be aware of very low expressed transcripts which can introduce a bias into such a normalization, as well as the fact that very abundant RNAs will be down-ranked, even if they are strong targets, but their number is higher then the amount of available protein, such that not every copy of RNA can be bound by a protein. For this thesis, we integrated expression values derived from RNA-Seq experiments into our findings, to rank targets by normalized CLIP signal and for downstream analysis like motif finding.

1.5.4 *Processing of NGS data*

A general analysis workflow for NGS and CLIP-Seq experiments is shown in figure 9.

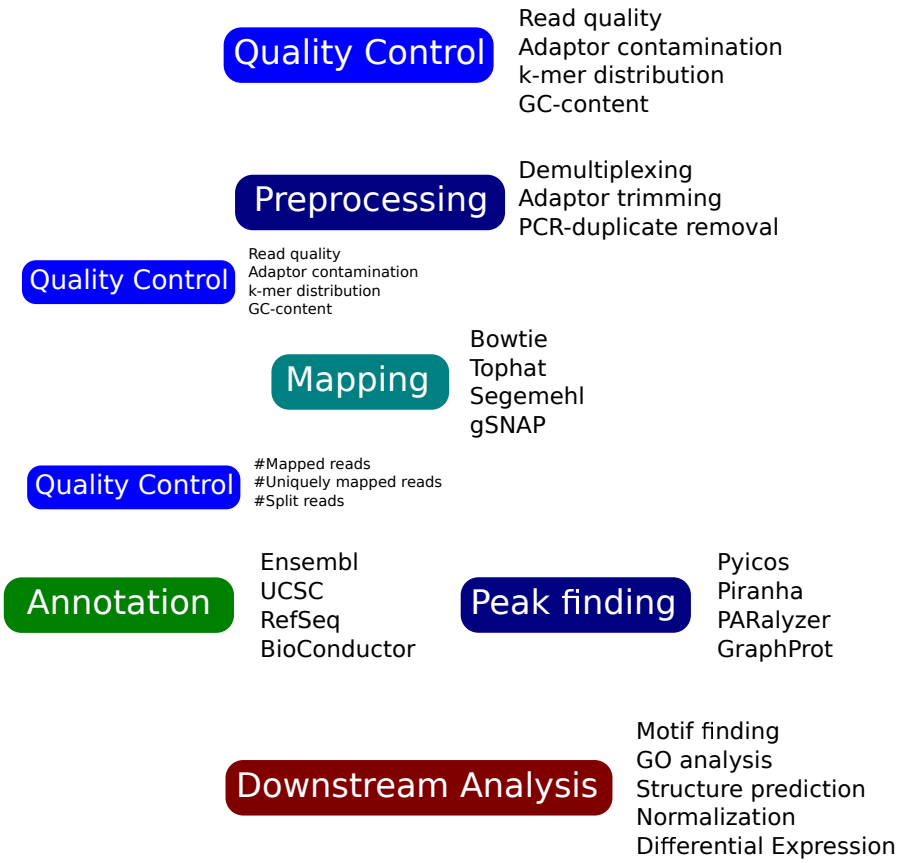


Figure 9: **CLIP-Seq processing pipeline** Quality control steps accompany a CLIP-Seq experiment after sequencing and every step of analysis. After preprocessing and mapping of reads, peak regions are defined and annotated. These peak regions are subject to downstream analysis regarding sequence and structure preferences, normalization and more.

1.5.4.1 *File formats*

Since it is very common for bioinformatical analysis to work on sequences, or strings to use a more informatical term, there are a lot of file formats, often specific for the task at hand. This is also the case for NGS data. This section will give a brief overview on file formats used during analysis of the underlying experiments. For a more complete picture, please refer to one of the many up-to-date www sources on this topic
(*e. g.* https://en.wikibooks.org/wiki/Next_Generation_Sequencing_(NGS)/Introduction#File_format_and_terminology and https://genome.ucsc.edu/FAQ/FAQformat.html).

While the FASTA format is a very common and widely used sequence format, consisting only of a header with some information followed by the sequence, this format is ancient and had to be adopted to deal with additional information. The so called FASTQ format is able to

store sequence quality and additional information of sequences and is often the starting point of NGS analysis. As mapping adds additional information, like number of matched sites, mutation events (INDELs, mis/matches) FASTQ was not enough and the most widely used SAM format [79] was invented.

Together with its binary counterpart BAM and the related tools for reading and processing (SAMtools [79] or PicardTools http://broadinstitute.github.io/picard/) the SAM format contains all the information required for downstream NGS analysis.

Information on annotation, gene features and the likes are stored in Browser Extensible Data (BED) and General Feature Format (GFF) or Gene Transfer Format (GTF) format or similar, mostly tab separated file formats, which are usually human readable and read/editable by hand or tools like the BEDtools suite [108], several peak finders/Differential Expression (DE) analysis tools and more.

Furthermore, (indexed, binary) formats for the easy upload of files to web-services like the UCSC genome browser [59] exist *e. g.* Wig or BigWig. Variant calling, *e. g.* for SNP detection require information often stored in the Variant Call Format (VCF). There is a multitude of further formats, mostly for specialized tasks of downstream analysis available, but the above mentioned formats are those most common and of importance for this thesis.

1.5.4.2 *Pre-processing*

As many RNA-Seq and CLIP-Seq experiments are run at different conditions for comparison, it is very common to multiplex samples. The addition of random barcodes together with fixed barcodes for each sample allows the parallel sequencing of multiple samples on one lane and the later split into the single samples.

Tools are available that allow manual splitting by barcode, *e. g.* the FastX toolkit (http://hannonlab.cshl.edu/fastx_toolkit/), however this is often done as a service by sequencing facilities. In rare cases, as for the CLIP-Seq data used here, custom built code has to be used to split samples by complex barcodes.

The next (optional) step is adapter trimming, as reads are often shorter than the maximum sequencing length of the sequencer, leading to readthrough into adapter sequences. Even under optimal conditions, at least barcodes and some PCR-primer adapters have to be cleaved from the reads, which can be done again with the FastX-toolkit, alternative programs like Cutadapt [91] or of course using custom built code.

1.5.4.3 *Quality control*

A very essential step during NGS analysis is quality control (QC). In principal it is recommended to perform QC after every step during pre- and post- processing, to ensure a correct basis for downstream analysis. Remaining adapter-sequences, calls with low quality and reads of wrong length can be identified and removed from the dataset.

A very handy tool for this task is FastQC (`http://www.bioinformatics.babraham.ac.uk/projects/fastqc/`). It creates statistics on overall sequence quality, as called by the sequencing machine, compares sequences to known adapters, analyses k-mer enrichment for bias estimation and more. Some experiments require specialized quality control, in this case *e. g.* analysis of T2C transition rates, which are not yet covered by established tools.

However, quality control is something that should in any case be adapted to the data at hand, as *e. g.* over-sequencing (high read-duplication levels) is a common problem in ChIP-Seq data, while it is simply not possible in RNA-Seq experiments. Also the influence of remaining adapter sequences, the number of uniquely mappable reads and so on are quality criteria that depend on the experiment and can not be generalized. So far no quality control pipeline specifically for CLIP-Seq experiments is available.

1.5.4.4 *Read mapping or assembly*

The high versatility of NGS experiments leads to a very diverse set of tools for read mapping or assembly of *de novo* genomes/transcriptomes, reviewed in *e. g.* Reinert et al. [112] and Simpson and Pop [122], which are the basis for this section. The most prominent software for read alignment is the Basic Local Alignment Search Tool (BLAST) [4], which uses a database of indexed k-mers that are compared to the query and extended until a threshold is reached to find the highest-scoring segment pairs.

Although this local alignment heuristic works very good for small datasets, it is simply not efficient enough to deal with NGS data, where millions of reads have to be aligned. NGS read alignment is non-trivial. Usually reads are mapped to a reference genome, which is not identical with the genome of the organisms the reads are derived from. This means that mapping algorithms have to deal with differences coming from sequencing errors, as well as naturally occurring single nucleotide polymorphisms (SNPs), insertions/deletions (InDels) of small regions and even large-scale complex variations of thousands of nucleotides, *e. g.* transposable elements.

Furthermore, reads derived from mature mRNA do not contain intronic regions, which are usually larger than exonic ones, so reads mapping to two or more exons have to be split to span the exon-intron-exon structure. Sometimes mismatches arise from the experimental method used, *e. g.* from bisulfite treatment during epigenomic NGS or T2C transitions from thio-uracil crosslink in PAR-iCLIP . However, independent of the source of reads and whether they come from single or paired-end runs, the challenge is to map them back to the position on the reference genome where they were derived from. This is done by solving an approximate matching problem, approximate because of above mentioned challenges (SNPs, *etc.*).

Two main approaches to deal with the large number of input and the large size of reference genomes exist, filtering and indexing. During filtering, one excludes regions on the reference where no approximate match is possible, thereby shrinking the search space. Indexing is based on pre-processing of the reference sequence, the reads or both to allow a quick lookup of potential mapping locations without scanning of the whole reference. Schbath et al. [119] divide algorithms for read mapping in three categories, those that use hash tables, Burrows-Wheeler Transform or suffix trees/arrays as underlying data structure.

Reinert et al. [112] also mention the enhanced suffix array and the FM-index, which is based on the Burrows-Wheeler transform. The FM-index is less memory demanding, but not as fast as the memory demanding suffix array. Irrespective of the underlying data structure, in the end the user will get a list of reads that could be mapped to the reference genome, the genomic location of the best match or matches and some information on the alignment and its quality. This assumes, that such a reference genome or at least a reference genome of some closely related species exists. If this is not the case, one has to generate a reference by genome/transcriptome assembly.

Even today, most sequencing technologies are limited to read lengths of not more than 150 bases, 10-20kb in case of PacBio (see table 1). The human genome on the other hand has a size of 3Gb. Reconstructing these huge genomes is possible by assembling read fragments at overlapping positions, generated *e. g.* via shotgun sequencing as proposed by Staden [125]. A simple approach for assembly is to iteratively join reads in decreasing overlap quality, starting with best matching reads and ending with reads with only small overlap.

Such nascent assembled sequences are known as contigs (contiguous sequence of bases). Greedy assemblers use this strategy of locally optimal joining, but are limited in their usefulness when it comes to repeat regions. Graph based assemblers represent reads and their relationships as vertices and edges in a graph and try to find a walk

through that graph that best reconstructs the underlying genome. In the simplest modes, each read is a vertex and linked with an edge to overlapping graphs. OLC assemblers [58] (Overlay, Layout, Consensus) first find overlapping vertices, then build an ordered layout for the graph and return a consensus sequence computed from the graph. Finding overlaps is a problem similar to the alignment of reads and best solved via indexing.

OLC assemblers became unfeasible when confronted with the large number of reads derived from NGS experiments and led to the development of De Bruijn graph based assemblers [107]. There, each read is broken into a sequence of overlapping k-mers, distinct k-mers are added as vertices to graph and those derived from adjacent positions are linked by an edge. De Bruijn graphs represent all copies of repeats as single segment in the graph with multiple entry and exit points, thereby collapsing repeat regions. Solving the assembly problem for such a graph can be formulated as an Eularian path problem, visiting each edge in the graph once. Assemblers usually construct contigs from unambiguous, unbranched regions of the graph and not the whole sequence at once. However, generating such a De Bruijn graph of k-mers for higher eucaryotes, given mismatches and repeats is extremely memory consuming.

In recent years Bloom filters, which use bit arrays, indexed by multiple hash functions or FM-index structures are used to deal with the memory consumption. String graphs were recently [103] discovered as an elegant way to represent overlap-based assemblies. Reads that are substrings of other reads, or contained by other reads are removed and transitive edges are removed from the graph. The resulting string graph shares properties with the De Bruijn graph without the need to generate k-mers.

As it is far beyond the scope of this thesis to give a detailed overview of all available implementations of the described algorithms, or a comparison of the latter, please refer to available literature for more information, *e. g.* [11, 112, 116, 119, 122].

1.5.4.5 *Downstream analysis*

After successful QC and mapping, reads are available for downstream analysis. In case of CLIP-Seq this means definition of peak regions, for the distinction of real binding sites from noisy data, sequence/structure motif search, annotation of bound regions and more. RNA-Seq reads allow the identification of gene/transcript expression levels, differential expression analysis, transcript isoform detection and more, mostly depending on read counts per defined region (*e. g.* gene).

A general protocol for downstream analysis is not available, as it strongly depends on the type of investigation conducted. Furthermore, one has to keep in mind that what works for one experiment may prevent conclusive findings in an other, or even worse lead to false positives, which are often hard to identify later on. At this point it should be mentioned that detailed description of data sources and experimental and bioinformatical processing steps is crucial for reproducibility of experiments.

Several projects aim at helping the user to create reproducible analysis pipelines, among them larger projects like `Snakemake` https://bitbucket.org/snakemake/snakemake/wiki/Home, `Bpipe` [117], and several commercial and smaller projects. During the course of this thesis we developed `ViennaNGS` [150], a lightweight `Perl6` based toolbox for the generation of reusable pipelines, some of which were used for downstream analysis steps within this thesis.

1.6 BINDING SITE IDENTIFICATION, NORMALIZATION AND MOTIF PREDICTION

No matter whether RNA- or protein-centric methods were used to identify interaction partners, the resulting collection of reads in case of (high-throughput) sequencing or fluorescence intensities in case of microarrays, or MS spectra, require advanced bioinformatical methods to identify binding partners and/or interaction sites. This section will focus on computational methods for binding site prediction from next generation sequencing related experiments and motif prediction from protein-centric experimental data.

In general, the first step is to identify true binding sites by filtering spurious and unspecific binding. Such sites are then used to identify binding motifs (see figure 10 for a workflow). The latter can then be used for binding site predictions, given that their quality is good enough and that the protein of interest has binding preferences.

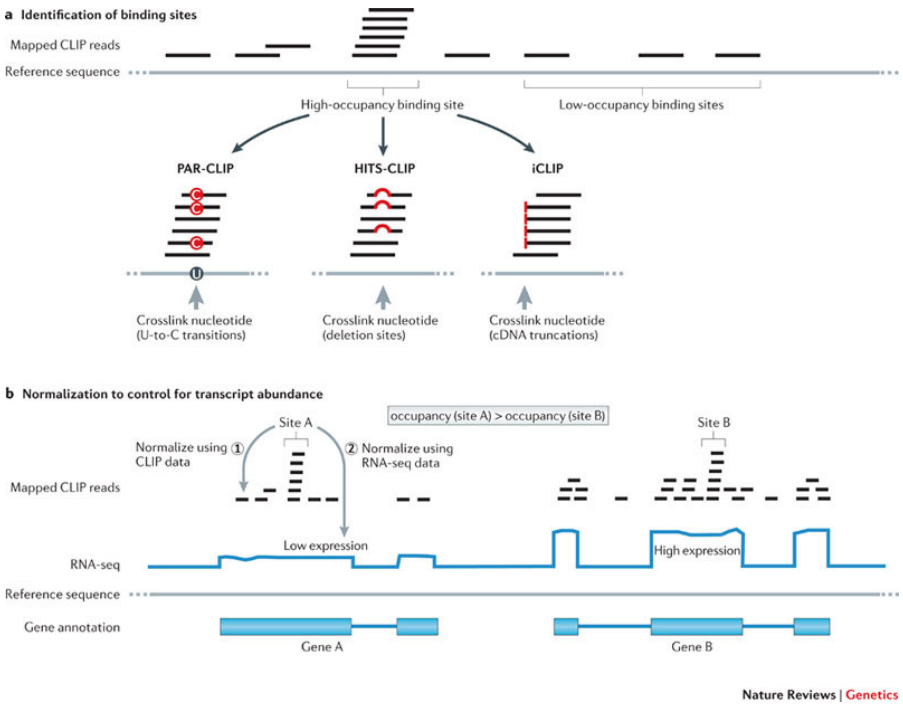


Figure 10: **CLIP-Seq peak finding and normalization** from König et al. [68]. A) Regions with enriched signal (crosslink events) are filtered from background with peak finder algorithms. B) CLIP-Seq signal of such regions depends on the amount of available transcript and total signal over transcript as well as transcript abundance can be used for normalization.

1.6.1 *Defining binding sites from CLIP-Seq experiments*

The major challenge in CLIP-Seq binding site prediction is missing negative control. Without negative control, one has to come up with a measure, that allows to distinguish true binding from background binding. Algorithms for binding site identification work on read counts in defined genomic regions or sequence stretches derived from data directly. A straight forward way to distinguish real binding from noise is the random distribution of reads over such a defined region (*e. g.* the gene body) and calculating the probabilities for finding the read density observed in the experiment. Such algorithms allow the computation of p-values for peak regions and enrichment values between theoretical and experimental signals. Pyicos [3], is one implementation of such an algorithm, where false discovery rates (FDRs) are calculated from CLIP-Seq experiments without control experiments.

Paralyzer [32] utilizes T2C transitions, introduced in PAR-CLIP experiments by reverse transcriptase when crosslink sites are encountered. Comparing the smoothened kernel density estimates (used to infer the probability density function) for transitions and non-transitioned nucleotides in binding regions, allows enrichment analysis for sites with more transition events then expected from background.

Piranha [135] compares read counts in bins with one or more reads to a negative binomial distributed read model. It theory, it allows to call peaks in all sorts of CLIP experiments, although correct estimation of bin size by the user is necessary.

The number of tools for peak detection and CLIP analysis is growing steadily (see table 2 for an overview), so this work only lists the so far most widely used implementations. A source for discussion is the elimination of background from CLIP-Seq experiments, as high signal does not automatically indicate strong binding, and the reverse is true as well.

The common approach of selecting only signal rich binding sites into the final set of peaks can lead to false positives, as some regions tend to show high signal across conditions and protein of interest, which suggests background binding. On the other hand one might miss important binding sites with low signal due to low expression of target sites. Challenges like these remain to be solved on the computational side, however, adequate experiment quality will always be of the essence for successful CLIP-Seq analysis.

Table 2: **CLIP-seq data specific processing tools**, adopted from Reyes and Ficarra [113]

TOOL	YEAR	EXPERIMENT	FOCUS	MAIN ADVANTAGE	RECOMMENDED CASE	AVAILABILITY
Paralyzer [32]	2011	PAR-CLIP	Peak detection	Exploits T to C mutations to Improve Signal to noise ratio	PAR-CLIP data	http://www.genome.duke.edu/labs/ohler/research/PARalyzer/
wavClusteR [25]	2012	PAR-CLIP (BAM format)	Noise and false positives reduction Peak detection	Distinguishes between non-experimentally and experimentally induced transitions	PAR-CLIP data	https://github.com/FedericoComocglio/wavClusteR
Piranha [135]	2012	CLIP-seq and RIP-seq (BED or BAM)	Noise and false positives reduction Peak detection CLIP-seq data comparison [correction for transcript abundance]	Corrects the reads dependence on transcript abundance	CLIP-seq and Transcript abundance data	http://smithlab.use.edu
mCarts [155]	2013	CLIP-seq	Sites prediction on different samples	Considers accessibility in local RNA secondary structures and cross-species conservation	RBP motif	http://zhanglab.c2b2.columbia.edu/index.php/MCarts
dCLIP [144]	2014	CLIP-seq	Peak detection CLIP-seq data comparison [correction for transcript abundance]	Detects differential binding regions in comparing two CLIP-seq experiments	several CLIP-seq datasets and Transcript abundance data	http://qbrc.swmed.edu/software/
PIPE-CLIP [155]	2014	CLIP-seq (SAM or BAM)	Noise and false positives reduction Statistical assessment Peak detection	Provides a significance level for each identified candidate binding site	HITS-CLIP, iCLIP	http://pipeclip.qbrc.org/
GraphProt [94]	2014	CLIP-seq and RNAcompete	Peak detection Sites prediction on different samples	Detects RBP motif secondary structure common characteristics. It estimates binding affinities	RBP motifs that are NOT located within single-stranded regions	http://www.bioinf_uni-freiburg.de/Software/GraphProt/

1.6.2 *Binding motif prediction*

Once binding sites are identified, the next logical step is to search for binding preferences of the protein of interest. Search for the preferred binding motif is a routine task with CLIP-Seq data, identification of such a motif is, however, non-trivial.

Motif finding is described as the problem of discovering of motifs without any prior knowledge of how the motifs look [55]. Given a set of sequences, the task is to find subsequences that occur more often than expected, meaning that they are over-represented. This means that the motif of interest will occur in many input sequences and can in principle be found by aligning the input sequences and searching for conserved regions. However, motifs do not have to be fully conserved, and they can even consists of sub-motifs themselves, or at least show some variability in their nucleotide content. Alignments can be used to generate Position Weight Matrices (PWM), which assign each position in a sequence a probability for containing a certain nucleotide. From such a PWM, the frequency of a given motif in the input can be computed and compared to the background frequency (*e.g.* number of motifs in genes), such that a score for over-representation is derived. Many implementations of algorithms that utilize this or equal strategies exist (see table 3 for an exemplary overview), among which MEME [8] is the most widely used. Applying an expectation maximization (EM) algorithm to find the most over-represented motifs in a set of sequences, MEME successfully predicted binding motifs for a set of RBPs from data.

RBP binding motifs can in general be predicted by DNA motif finders, which either only consider RNA sequence, or include RNA secondary structure. Accessibility of motifs is not a factor when considering DNA motifs, as DNA is in general in a double stranded B-form $\alpha-$helical structure, which allows (sequence specific) DNA binding proteins to interact with its major groove. RNA on the other hand is less accessible when double-stranded, due to its A-form $\alpha-$helical geometry, which results in a very deep and narrow major groove and a shallow and wide minor groove, both not accessible for proteins. Thus, most RBPs are thought to prefer single stranded RNA (ssRNA) regions for interaction. To correctly predict binding motifs for RBPs it is therefore interesting to include accessibility of binding sites. MEMERIS [47], predicts the probability of being unpaired for analyzed regions and incorporates this single-strandedness into MEME motif prediction, making it more appropriate for ssRNA binding protein motif prediction.

However, it is not only interesting to consider the accessibility of the preferred motif, but also to get an idea of the structural con-

text, as motif embedding regions can of course influence the binding behaviour of RBPs. RNAcontext [57] interprets the probability for each nucleotide in a binding site to be found in a particular RNA secondary structure (hairpin, multi loop, interior loop, etc.) derived from *e. g.* RNAplfold. It combines this information to extract the preferred structural context of a motif, accepting loss of nucleotide resolution. Furthermore it can deal with affinity data and use the affinity of a protein to binding sites to refine the motif search and predict affinities for identified motifs. This on the other hand, requires such a dataset for optimal performance, which is not standard, and is optimized for short sequences, ignoring broader context which may be important for successful interaction.

GraphProt [94] is a graph kernel-based machine learning algorithm, extracting motifs that were highly predictive for binding from a set of bound and unbound sequences. These motifs can be used to predict binding affinities and de novo binding sites, not present in the experimental output. A main advantage over RNAcontext is that the full secondary structure information is conserved and not just a structure profile per motif, which decreases the error-rate and can be used to identify structural preferences of RBPs with higher resolution.

Motif finding algorithms incorporating gaped positions have not yet been extensively applied to RNA-protein interaction data, although many RBPs contain more than one RNA interaction site and thus have the potential to bind gaped motifs.

However, MEME works well for many RBPs, presumably because of their preference for ssRNA regions, and has successfully been used for binding motif prediction with our dataset (see 2.2.8.1).

Table 3: **Motif finding algorithms used for analyzing RBP-RNA interaction data, adopted from Cook et al. [29]**

Algorithm	Input	Type of motif generated	Considers secondary structure?
MEME [8]	Positive (and optionally, negative) sequences	PWM	No
PhyloGibbs [121]	Positive (and optionally, negative) sequences	PWM	No
REFINE [114]	Positive sequences	N/A, Filtering procedure to only consider sequences containing three enriched hexamers; filtered sequences are then submitted to another motif finding algorithm	No
cERMIT [37]	Rank ordered sequences	PWM	No
DRIMUST [35]	Rank ordered sequences	IUPAC motif, possibly gapped	No
StructuRED [40]	Positive and negative sequences	PWM in a hairpin loop	Yes, considers possible hairpin loops up to 7 bases with at least 3 paired bases
TEISER [75]	Sequences and scores (e.g., stability scores)	PWM in a hairpin loop	Yes, considers possible hairpin loops with stems 4-7 bases long and loop sizes of 4-9 bases
RNAcontext [57]	Sequences and affinity scores	PWM with structural context scores	Yes, learns the preferred structural context of each base in a motif
GraphProt [94]	Positive and negative sequences	graph-based sequence and structure motifs, can be visualized with logos	Yes, models RNA structure using a graph-based encoding
CMfinder [152]	Positive sequences	structured sequence	Yes, SCFG-based, examines the most stable structures in the input
RNApromo [109]	Positive sequences	structured sequence	Yes, SCFG-based, optimizes a motif from an initial set of substructures generated from the input
#ATS [81]	Positive and negative sequences	IUPAC	Yes, scores candidate binding sites by accessibility
MEMERIS [47]	Positive and negative sequences	PWM	Yes, uses accessibility as prior knowledge to guide motif finding toward single-stranded regions

1.6.3 *RNA-RBP databases*

With a growing number of experiments and RNA-RBP interaction predictions, online databases collecting this kind of data emerged. Such databases make it possible to compare RBP targets for shared/unique sequence and/or structure features, shared motifs and more and build the basis for many downstream analysis tasks. Table 4 shows a number of currently available databases for RNA-RBP interaction studies, some of them even offering ready to use analysis pipelines and tools.

Table 4: **Databases for RNA-RBP interaction data**, adopted from Cook et al. [29]

Database	URL	Features
RBPDB [30]	http://rbpdb.ccbr.utoronto.ca/	Direct observations of protein-RNA interactions in metazoans, both low- and high-throughput
CISBP-RNA [111]	http://cisbp-rna.ccbr.utoronto.ca/	Directly observed and predicted (by homology with known proteins) motifs. Tools for scanning sequences and comparing motifs
starBase [80]	http://starbase.sysu.edu.cn/	RBP-RNA and miRNA-RNA interactions from CLIP data
doRiNA [16]	http://dorina.mdc-berlin.de/	mRNA-centric or RBP-centric search of CLIP data including combinatorial search
CLIPz [62]	http://www.clipz.unibas.ch/	Storage and analysis (mapping reads, extracting clusters, mapping T2C conversions) of CLIP data
CLIPdb [133]	http://lulab.life.tsinghua.edu.cn/clipdb/	CLIPdb aims to characterize the regulatory networks between RNA binding proteins (RBPs) and various RNA transcript classes by integrating large amounts of CLIP-Seq (including HITS-CLIP, PAR-CLIP and iCLIP as variantions) data sets
AREsite2 [34]	http://rna.tbi.univie.ac.at/AREsite	Database of AU-/GU-/U-rich elements in human, mouse, zebrafish, fruit fly and worm with information to overlap with CLIP-Seq identified RBP binding sites

1.6.4 *Expression level estimation from RNA-Seq data*

Transcript expression levels contain a lot of information that is important for the correct interpretation of biological consequences of *e. g.* experimental conditions, cell state or in our case inflammatory response. The type of expressed transcripts and their expression rate can give insight into gene expression control mechanisms and regulatory networks and show relevant changes under different cellular conditions. It is quite obvious, that the expression rate of a target RNA has influence on the amount of protein that can interact with this target. This makes RNA-Seq an important part of the thesis at hand, both, to normalize CLIP-Seq signal and to analyze changes between the different states of LPS induction investigated here.

In general, existing algorithms for the estimation of transcript/gene expression can be divided into count based and transcript isoform abundance based methods. While the former assign read counts to defined regions and are mostly used for differential-expression (DE) analysis, the latter assign fragments or reads to regions which can either be pre-defined or inferred from read coverage. This allows the prediction of expression levels of de-novo transcripts without prior annotation.

Kanitz et al. [56] state that gene level expression estimates obtained by cumulating transcript isoform abundance are more accurate than those from "count-based" methods. Among the most widely used implementations for transcript isoform abundance estimations is Cufflinks [129], which was also used for this thesis, while DESeq [5] respectively the newer version DESeq2 [85] is most commonly used for count based analysis of DE.

1.7 RNA STRUCTURE

Interactions between RNAs and proteins are influenced by the structural context of binding sites. As many RBPs either bind single-stranded regions, or have certain structural preferences, like hairpin loops, RNA secondary structure is a critical aspect to be considered for successful binding site prediction.

Although ABPs, as their name implies, show a preference for certain RNA motifs, this sequences alone may not be sufficient for effective binding. The influence of "structuredness", which means the general probability of a region to form secondary structures, on RNA-protein interaction is one of the main motivators for this thesis.

In biology, it is generally known that structure defines function. What is true for proteins with their modular buildup dictating their function, holds also true for RNAs. For proteins, where tertiary structure is crucial for function, the fold of a protein into the correct tertiary structure is the main step from peptide chain to functional protein, driven by hydrophobic forces. RNAs however, have a hierarchical folding, where basepairs and helices (known as secondary structure) are formed first and then complex tertiary structures can be formed. In contrast to proteins, where secondary structure is mainly the aggregation of polypeptides into α—helices and β—sheets, RNA secondary structure already contains a lot of information, including the potential of an RNA for intra- and intermolecular interactions.

RNA secondary structure elements (see fig. 11 for an overview) are formed via intramolecular interactions of nucleotides. Such interactions form base-pairs via hydrogen bonds between corresponding nucleotides. The standard set of RNA base-pairs (AU,GC) is known as Watson-Crick-base-pairs , named after the famous discoverers of DNAs double-helical structure [148]. GC-base-pairs can form three hydrogen bonds between their Watson-Crick edges, while AU-base-pairs can only form two. This is important considering their energy contributions, which is higher for GC- than for AU-base-pairs . The most important stabilizer of RNA secondary structure however, are stacking interactions, where base-pairs in close proximity generate an energy bonus from electrostatic forces of the stacking nucleotides. This energy bonus has a huge impact on the thermodynamics of RNA secondary structure , as adjacent base-pairs (stems) become more favorable than separated ones. The same holds true for the energy contributions of loop regions, which depend on the type and amount of bases in the loop (a minimum of four is required for a loop to form).

Besides canonical base-pairs other interactions between nucleotides are occurring in nature like *e. g.* the Non-Watson-Crick (or non-canonical)

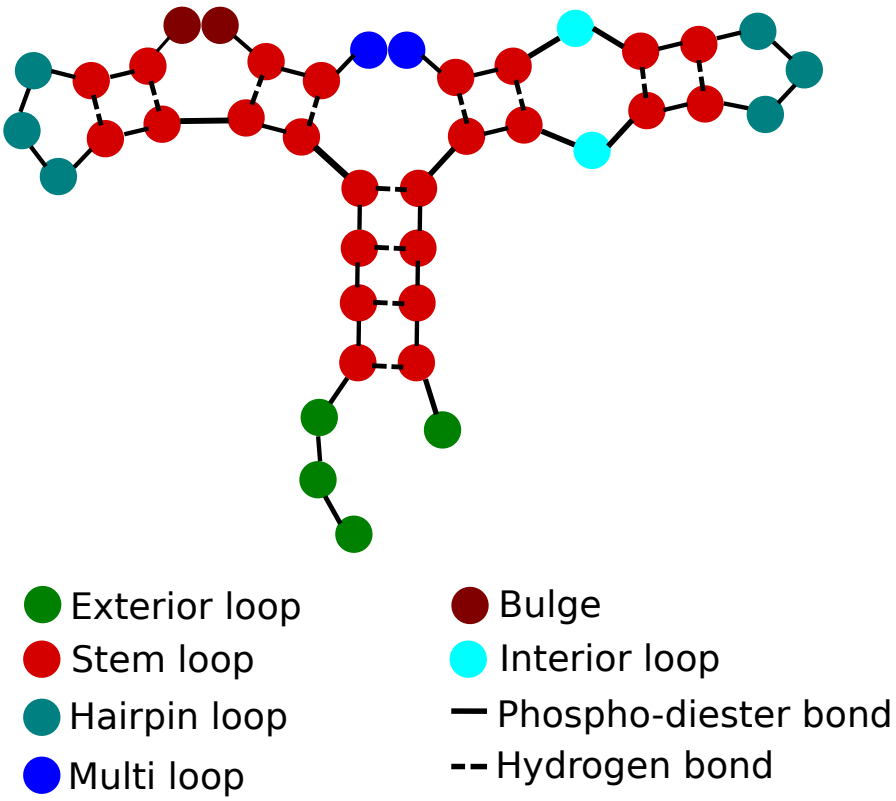


Figure 11: **Overview of RNA secondary structure elements** Loop types that occur in RNA molecules and are distinguished by *in silico* structure prediction algorithms due to their differing thermodynamic effects. One distinguishes stem loops, hairpin loops, multi loops, bulges, interior loops and exterior loops.

wobble-basepair GU. RNA bases can not only interact via the "standard" Watson-Crick-edge, they can also form bonds between their Hoogsteen- or CH-edge and their Sugar-edge. These edges even allow the formation of base-pairs between three bases at once, known as base triplets, influencing the stability of helices and tertiary as well as quaternary structures.

So far not mentioned are long range interactions like pseudo-knots or kissing hairpins, which also contribute to RNA secondary structure formation. They are a form of intramolecular base-pairing where a stem or loop region interacts with another non-adjacent stem or loop regions. Such interactions are usually not very frequent *in vivo* and hard to compute *in silico* , as they explode the search-space for potential RNA secondary structure , thus they are neglected from most prediction algorithms. In general such structures are treated as tertiary interactions.

The Leontis-Westhof annotation [76–78] (see fig. 12) introduces a set of motifs that use the three edges of nucleotides in different conformations, to categorize 3D-interactions in a 2D fashion. Among these

motifs are sarcin/ricin loops, kink-turns, C-loops and A-minor motifs, which can all be seen as building blocks for RNA tertiary structure.

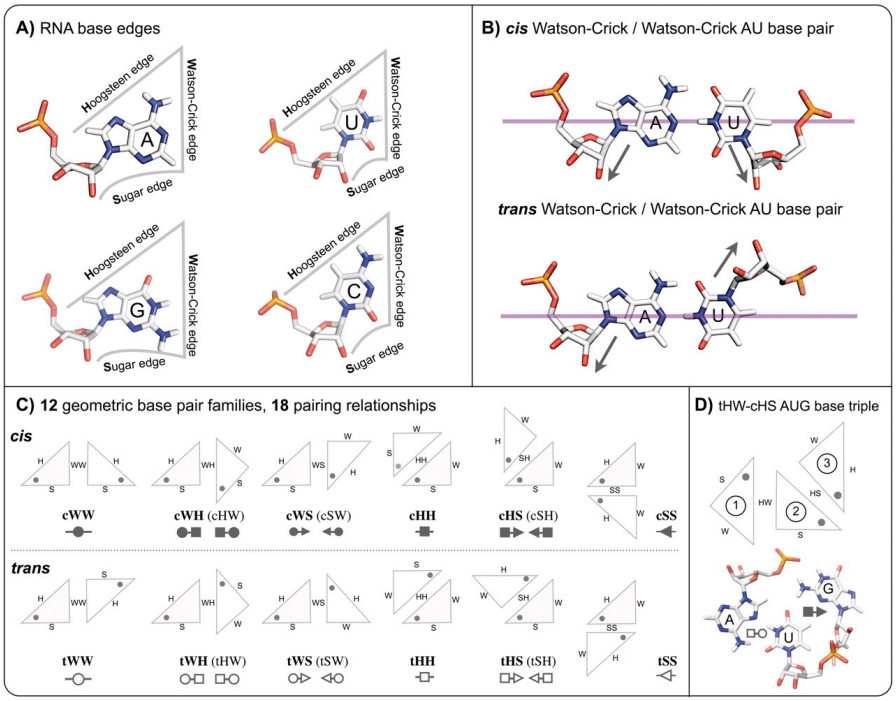


Figure 12: **Summary of Leontis/Westhof base pairing classification**, from Abu Almakarem et al. [1]. (A) Three edges for base pairing interactions, the Hoogsteen (H), Watson-Crick (W) and Sugar (S) edges which include the 2'-OH group of the riboses. (B)Nucleotides can pair in cis and trans conformation, with the glycosidic bonds of the nucleotides on the same side in the cis configuration, and on opposite sides in the trans configuration. (C) Schematic representations of each of the 12 basic base pair families. Triangles represent bases, circles represent W edges, squares represent H edges and triangles represent S edges. Filled in symbols represent cis and open symbols, trans base pairs. (D) Schematic showing a representative regular base triple, where the central base pairs with each of the other two bases using a distinct base edge.

Secondary structure influences binding potential as well as binding influences the ensemble of potential RNA secondary structures. An inaccessible binding site requires some kind of energy contribution to unfold and become accessible, while a bound protein can prevent secondary structures from forming or provide the energy needed to form it. *in vivo* RNA can be expected to be in constant contact with several RBPs and binding factors like other RNAs (*e. g.* miRNAs, metal ions, etc.), which all influence and are influenced by the ensemble of potential secondary structures a RNA molecule can form.

in silico methods allow the prediction of potential secondary structures for RNA molecules at given temperatures and since recently

also under the constraint of other interaction partners. Also recently developed were experimental high-throughput methods for measurement of RNA structuredness which allow the assessment of secondary structures *in vitro* or even *in vivo* . Together with experiments for the identification of binding sites, these methods are the basis for further investigations of structural influence on RNA-RBP interaction.

1.7.1 *Experimental determination of RNA secondary structure*

Footprinting techniques determine the secondary structure of RNA molecules by cutting the RNA using RNAses specific to single or double stranded RNA, or utilizing small molecular reagents cleaving or modifying nucleotides in proportion to their accessibility.

Selective 2'-hydroxyl acylation and primer extension (SHAPE) and its derivative SHAPE-Seq [87] and SHAPE-Seq 2.0 [84], as well as PARS [61] are techniques, that can be applied to experimentally validate RNA secondary structures in a high-throughput manner in an *in vitro* setting.

DMS-Seq [115] even allows this *in vivo* . However, such experiments, similar to computational predictions, do not return a single structure as they work on the whole set of available RNA molecules. As a consequence, one gets a snapshot of all structures formed by the specific RNA molecule at time of the experiment. While *in vitro* experiments lack the "real life" environment of cellular compartments, they allow to investigate RNA secondary structure without interference of other molecules. *in vivo* experiments on the other hand return a more realistic look on RNA secondary structure , as they are probed in an natural environment.

However, such experiments alone might provide insights into structure, but to really understand RNA secondary structure dynamics, such experiments would best be combined with knowledge of interaction partners that might influence structure formation.

Thermodynamic measurements of energy contributions of single base-pairs and loop-types are the basis for free-energy based algorithms for RNA secondary structure prediction. However, only some datasets have been published, most widely used are optical melting measurements (see *e. g.* Turner and Mathews [130]). Parameters derived from such or chemical modification experiments [93, 142] are readily incorporated into RNA secondary structure prediction algorithms.

1.7.2 *in silico prediction of RNA secondary structure*

Free-energy based algorithms build on the assumption, that thermodynamically stable structures are more likely to exist *in vivo* than un-

stable ones. The Zuker algorithm [157] is the basis for programs that predict secondary structures using their thermodynamic probabilities. More complex approaches consider all possible structures via partition functions, based on an algorithm first proposed by McCaskill [95]. The fact that functional RNA secondary structures are more likely to be conserved through evolution than non-functional ones is the basis for covariation algorithms. As simultaneous folding and alignment of RNA sequences is computationally costly, most implementations use heuristics for their predictions.

In principle algorithms for the computation of RNA secondary structure rely on Dynamic Programming (DP). Nussinov and Jacobson [104] *et al.* presented the first efficient algorithm for the prediction of RNA secondary structure with a maximum number of base pairs. However, as described earlier, stacking interactions play a crucial role for correct RNA secondary structure formation and are not modeled by this type of algorithm.

First steps towards Minimum-Free-Energy (MFE) structure predictions were done by decomposing RNA secondary structure into their respective loop regions, which are enclosed by stems that contribute most of the stacking energy. The idea is to estimate the energy of a structure by decomposing it into its loops and summing up their energy contributions, the most famous algorithm that solved this problem efficiently was presented by Zuker [157]. However, this allowed only the prediction of simplified structures and no suboptimal solutions. Suboptimal structures are important, as RNA is not as a static molecule, always folding into its MFE structure, but as a dynamic entity, which folds and unfolds upon interaction or changes in cellular environment.

Modeling all suboptimal structures of a RNA molecule is a non-trivial task, as an exponential number of structures is possible, first efficiently solved by Wuchty et al. [151]. This approach works only for sequences of small length and is thus not applicable for exhaustive predictions of most naturally occurring RNAs. However, a key point for secondary structure prediction in RNA-RBP interaction studies is not to determine a specific structure, like the minimum-free-energy (MFE) structure, but more the gain of accessibility information.

As a stretch of RNA must be accessible for most RBPs to interact, the most likely secondary structure is of less importance than its accessibility derived from the ensemble of structures an RNA can form. McCaskill [95] presented an algorithm that allows the exhaustive calculation of base-pairing probabilities from the Bolzmann distributed ensemble of structures in thermal equilibrium. Derivatives of this approach are used to compute local sequence accessibility (see *e. g.* RNAplfold [15]).

The latter and other programs from the famous ViennaRNA package [83] have been used in this thesis to predict the effects of accessibility on RNA-protein interactions and cooperativity/competition. An interesting feature of such predictions is that not only accessibility can be predicted, but also the probability of being in a certain structure (*e. g.* hairpin, stem, bulge, ..).

However, one has to be aware that such predictions, as they are made on a local rather than global scale, are very context-sensitive. This means that when analyzing *e. g.* CLIP-Seq target sites, the length of the surrounding region one selects for folding has a strong impact on the results.

1.8 GENE ONTOLOGY

Gene Ontology (GO) uses defined terms to describe gene properties. It covers three domains, Cellular Component (*e. g.* cell parts or extracellular environment), Molecular Function (*e. g.* binding, catalysis) and Biological Process (*e. g.* signal transduction, metabolic process). The Gene Ontology Consortium http://geneontology.org/ and its GO project are concerned with the development of a consistent computational representation of how genes encode biological functions at the molecular, cellular and tissue system levels in form of GO terms.

Such terms exist for most genes in many organisms and can be used to analyze functions of a set of gene by GO-term enrichment, to find over- or under- represented GO terms. Such terms can be seen as indicators for *e. g.* molecular functions specific to the analyzed set of genes. However, GO-enrichment analysis results depend strongly on the set of available GO-terms for the organism and genes of interest, as well as the selected background, and have to be interpreted with care. They can, however, help to identify differences between experiments/conditions in a broader context than just a list of target genes as GO term comparison allows some conclusion on cell/tissue/organisms wide changes in broad context.

2 RESULTS

2.1 ARESITE 2.0

As described in section 1.1.3.1, AU-rich elements are cis-acting sequence motifs, preferentially bound by ARE binding proteins (ABPs). Together with U- and GU-rich elements, these sequence patterns represent a set of potential binding sites for proteins that have a crucial influence on RNA function and half-life.

With the publication of AREsite [43], a first attempt was made at annotating potential ABP target motifs in human and mouse protein coding 3'UTRs and evaluating their functional properties in terms of accessibility and conservation.

Our recently published update AREsite2 [34] contains annotations of motifs in all genic regions (exons, introns, UTRs) of coding and non-coding genes in human, mouse, fruit fly, zebrafish and bandworm. This vastly increased amount of information is accessible either via a web interface, or a new REST API for semi-automated retrieval of information.

2.1.1 *Improvements*

A comparison of features between `AREsite` versions 1 and 2 is provided in table 5, adopted from Fallmann et al. [34].

The updated database provides additional information on the level of genomes/transcriptomes/motifs analyzed, genic regions annotated, accessibility of data and integration of experimental data.

This section will provide some information on the annotated motifs in included organisms, the intersection with published CLIP datasets for the RBPs HuR, TTP and Auf1 and an outlook on how such curation can help predicting functional binding sites from the huge set of annotated motifs.

2.1.2 *AU-/GU-/U-rich elements in AREsite2*

AREsite2 contains information on 378,019,727 motifs alone in human and more than 1.5×10^9 motifs in total. This is a manifold of what was covered in the first version of the database. Figure 13 shows a comparison of numbers of human and mouse coding and non-coding genes, containing at least one copy of the core ARE (ATTTA), URE

Table 5: **Summary of features of `AREsite` and `AREsite2`, respectively**

Genic feature	AREsite	AREsite2
3'UTRs	✓	✓
5'UTRs		✓
CDS		✓
introns		✓
exons	✓	✓
mRNAs	✓	✓
non-coding RNAs		✓
Species	**AREsite**	**AREsite2**
human	✓	✓
mouse	✓	✓
zebrafish		✓
fruitfly		✓
C.elegans		✓
Motif feature	**AREsite**	**AREsite2**
AREs	✓	✓
UREs/GREs		✓
Motif accessibility	✓	✓
Secondary structures in overlap		✓
Conservation information	✓	✓
Result download	✓	✓
Database dump		✓
Related literature	✓	✓
REST interface		✓
CLIP-Seq integration		✓

(TTTTT), GRE (GTTTG), or the poly-A signal (AWTAAA) in exon/intron/UTR/CDS.

Mouse is a well established model organism, not least for the portability of findings to human. Especially the portability of findings concerning gene expression regulation make mouse a valid model system for investigations in this area. Comparing motif numbers and location in human and mouse, it becomes obvious that mouse also has a high potential for investigations concerning AU/GU/U-rich elements and their biological function. However, there are differences, as is to be expected even between closely related organisms. AREsite2 provides the means to compare binding element related findings between these two and other model organisms.

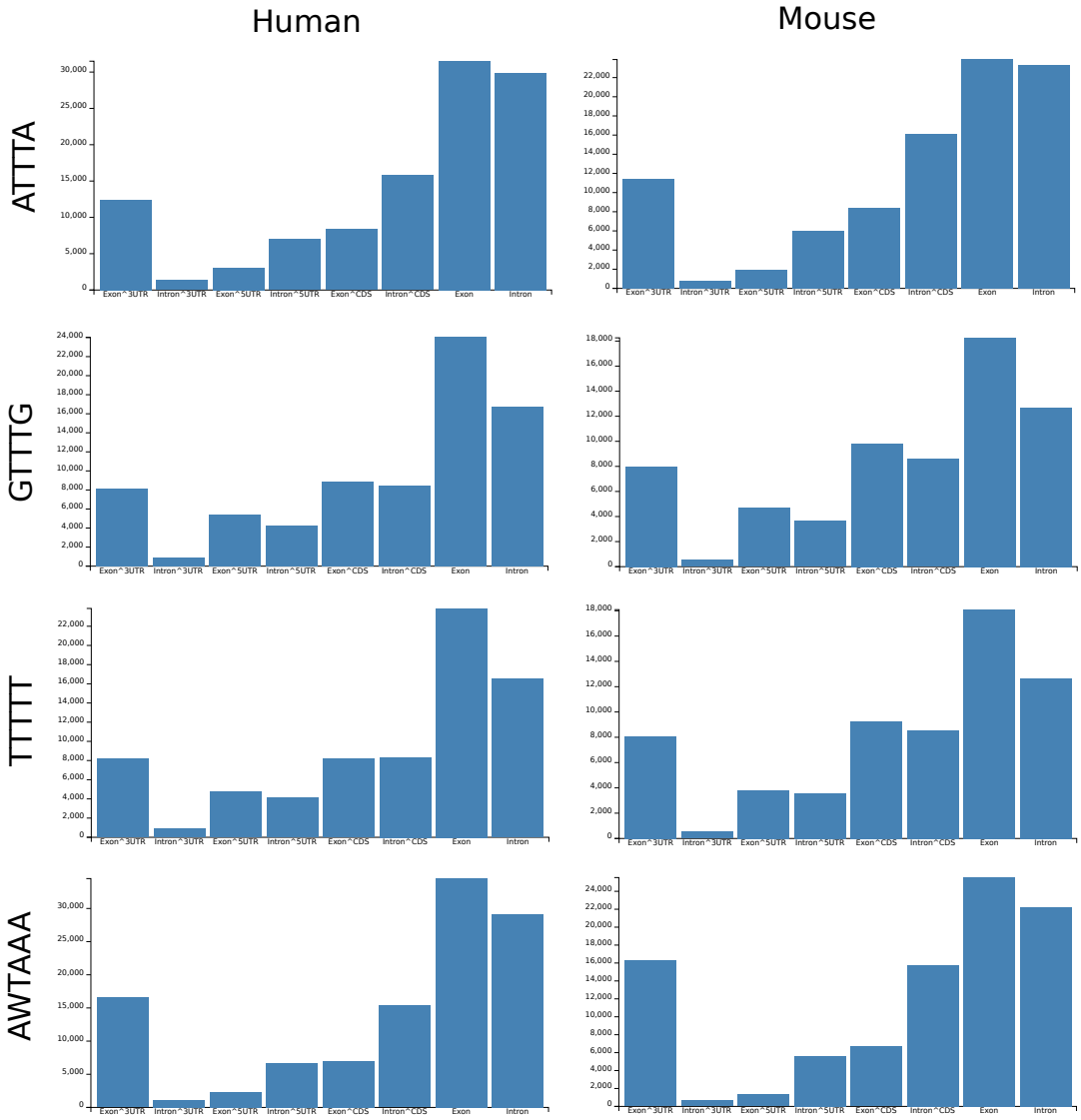


Figure 13: **Genes with motifs annotated in AREsite2** A comparison of genes with at least one copy of the core ARE ATTTA, URE TTTTT, GRE GTTTG, and the poly-A signal AWTAAA in human and mouse. Motifs are binned by the genic region they are located in.

2.1.2.1 *Accessibility of AU-/GU-/U-rich elements in AREsite2*

Presence or absence of a motif in a given gene is an important information. However, presence alone is not evidence enough for interaction with a protein. A motif that is hidden in secondary structures which prevent proteins from interacting will most likely have no biological function.

Similar to its predecessor, AREsite2 includes the local structuredness of motif sites in terms of opening energies and accessibility probabilities. RNAplfold [14] was used to calculate these terms for each gene

in the database, considering short range interactions (W = 80, L = 40) as well as mid range interactions (W = 240, L = 160).

Depending on the protein of interest, binding preferences may include some sort of secondary structure. To make information on this level available, AREsite2 incorporates stable secondary structures in overlap with annotated motifs from genome wide scans with RNALfoldZ [44, 49]. Locally stable RNA secondary structures were predicted for all included genomes and filtered by Z-score. Motifs of interest in overlap with such a stable structure are visualized with Forna [60], which allows the user to inspect and interact with these structures.

Summing up, AREsite2 contains a lot of information on motif location, annotation, accessibility, conservation and more for a vast amount of potential RBP targets. Some of this motifs were extracted from the database for downstream analysis, which is the topic of the next sections.

2.1.3 *Integration of CLIP-Seq datasets*

As mentioned before, more than 1.5×10^9 motifs were annotated for AREsite2. This, however, does not mean that all of those motifs are actively bound by RBPs, or at least not that they are all functional at once in RNA half-life regulation. To produce a more comprehensive picture of functional target sites, CLIP-Seq datasets, retrieved from the CLIP database CLIPdb [133] or directly from source (*e. g.* Mukherjee et al. [102], Sedlyarov et al. [120]) were integrated. Preprocessed CLIP-Seq datasets were intersected with annotated motifs, to extract motifs with experimental evidence for interaction in terms of CLIP signal. Those motifs are considered active and part of the positive set for further investigation. Motifs without overlap are considered inactive and part of the negative set.

With those datasets several downstream analysis steps were conducted, as described in the next sections. However, one has to keep in mind, that this type of analysis is prone to error by various sources. For one, the set of positive (bound) motifs is depending on the quality of the CLIP-Seq experiment. It is commonly accepted, that CLIP-Seq does not guarantee full saturation of binding sites, and is of course depending on the cell type and conditions used for the experiment. This makes it particularly hard to generate an adequate negative set, as one wants to prevent false negatives, or negatives that have no biological meaning in the context of the experiment conducted. RNA-Seq derived expression data can be integrated, to filter for expressed transcripts and get rid of motifs without possible function due to

their location on unexpressed transcripts. Further filters can be applied, *e. g.* one can filter motifs in regions without a certain regulatory role, which in case of mRNA half-life control would mean to exclude motifs not within 3'UTRs.

As for most of the experimental datasets in AREsite2 no accompanying RNA-Seq experiment is available and the aim is to investigate principle differences between all bound and unbound sites, no further filtering steps were conducted for the results presented in the next section. This has a strong influence on the results derived with both datasets, even when the principal analysis is similar, which is discussed in section 3.1.

2.1.4 *AU/GU/U-richness vs accessibility of motifs*

Motif presence and accessibility are the basis for successful interaction with RBPs. However, just those two criteria alone are no guarantee for interaction and much less for activity in terms of biological function. The question at hand is which features allow to distinguish active from inactive motifs. In a first attempt to define such a feature or set of features, the previously described positive and negative sets were analyzed for their mean A+U-richness and accessibility.

This analysis was conducted for hg38 and mm10, with CLIP-Seq sites of TTP (3 h and 6 h after LPS induction for mouse, extracted from [120]), HuR and Auf1 (human only), after lift-over (mm9 to mm10, hg18 to hg38) where necessary.

Figures 14 to fig 17 show mean mono- and di-nucleotide content of AU/GU/U-rich elements in positive and negative sets with 15nt flanking region. Although all possible dinucleotides were considered, only those of interest for this analysis (AU,UU,GU, each the sum over their permutations) are shown. It is important to mention that AU and GU classes also contain AA/UU and GG/UU respectively.

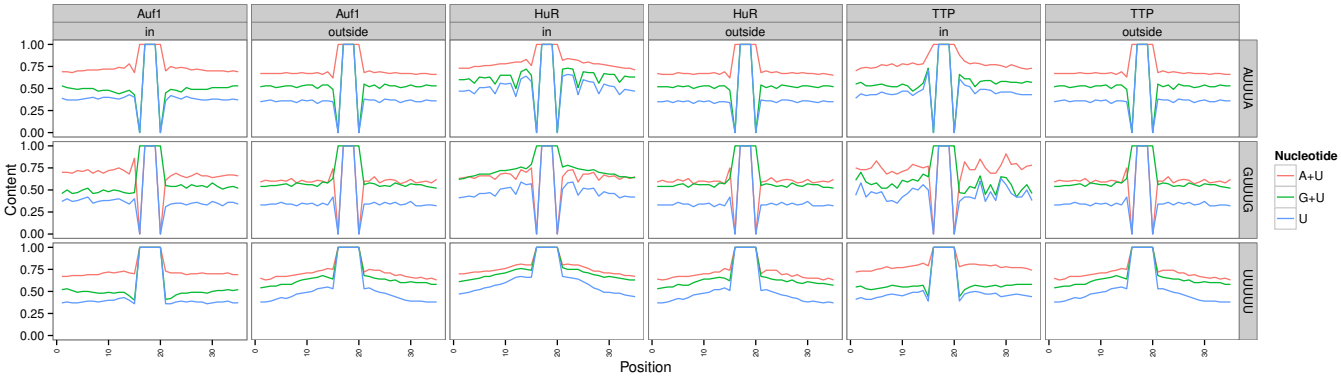


Figure 14: **Mean mono-nucleotide composition of AU/GU/U-rich motifs in human** A comparison of the mono-nucleotide composition of AU/GU/U-rich elements overlapping CLIP-Seq signal of Auf1, HuR and TTP and without overlap. For each protein and motif combination nucleotide content of motifs in and outside of CLIP-Seq defined binding regions is compared.

The A+U mono-nucleotide content of flanking regions is for all proteins and all motif classes higher for motifs overlapping CLIP-Seq signal than for the negative set. This effect is particularly strong for TTP and HuR. The same is true when only comparing U content. On the contrary, the G+U nucleotide content for AUUUA and UUUUU motif flanking regions of Auf1 and TTP is higher in the negative set, which shows that motifs in GU-rich context are less likely to be bound and therefore active.

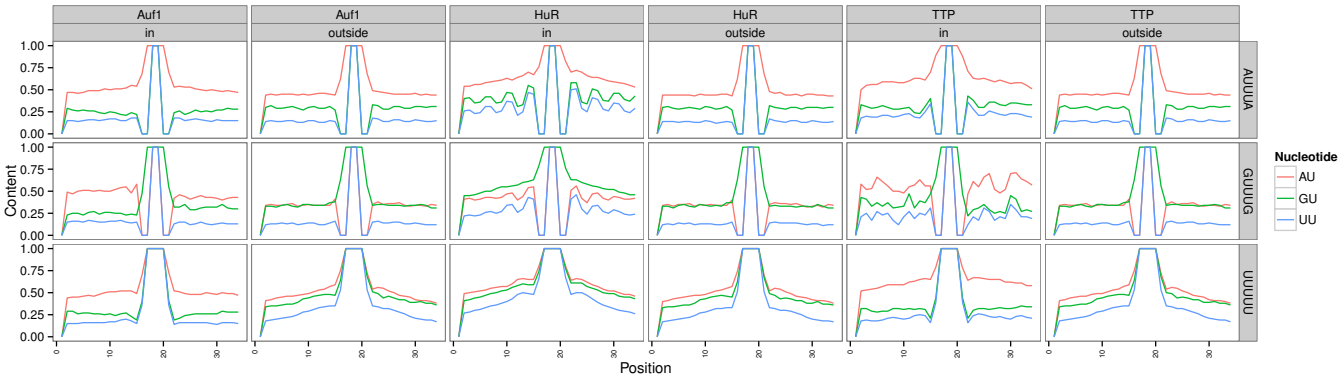


Figure 15: **Mean di-nucleotide composition of AU/GU/U-rich motifs in human** A comparison of the di-nucleotide composition of AU/GU/U-rich elements overlapping CLIP signal of Auf1, HuR and TTP and without overlap, similar to figure 14

The content of AU and UU dinucleotides is in most cases higher in the positive set for all proteins and motif classes, similar to the mono-nucleotide content. Interestingly, UU and GU di nucleotide content for Auf1 in general and TTP in the UUUUU motif class is higher for the negative set. In summary this indicates that TTP and Auf1 prefer AU-rich flanking regions around their interaction sites, while HuR seems to care only for U richness, independent of co-occurring nucleotides.

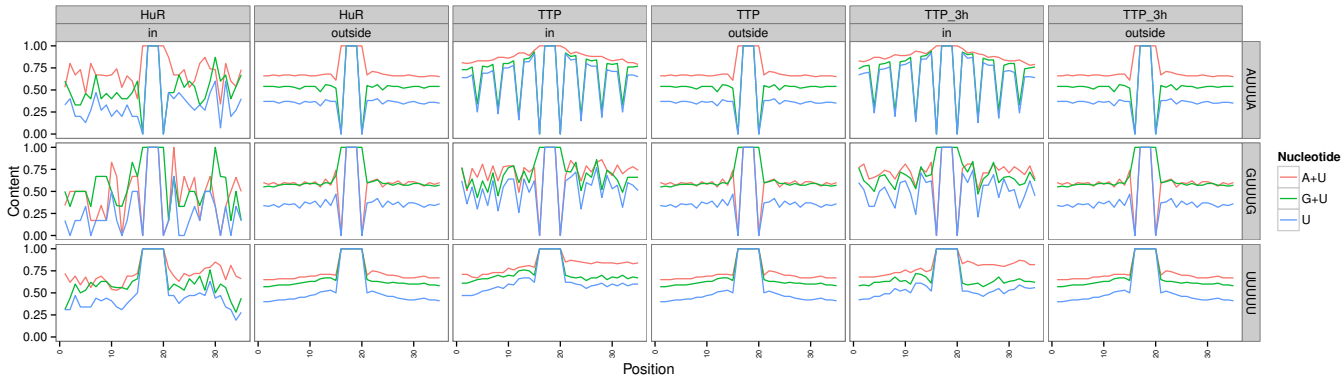


Figure 16: **Mean mono-nucleotide composition of AU/GU/U-rich motifs in mouse** A comparison of the mono-nucleotide composition of AU/GU/U-rich elements overlapping CLIP signal of HuR and TTP (3 h and 6 h after LPS induction) and without overlap, similar to figure 14

In mouse, the relative small amount of publicly available binding sites for HuR render this analysis step hard to compare. The content of all mono-nucleotides in the HuR negative set however, resembles that seen for human, thus it seems safe to assume similar preferences. The TTP sets show that for motifs of the AUUUA class, the A+U content stays high in the positive set, while the G+U and U content varies. The variation can be interpreted in a way, that AUUUA motifs in TTP bound sites are preferentially embedded in A+U rich regions, with recurring As, many Us and only a small portion of Gs. In general, the distributions over all motifs indicate, that TTP bound motifs are embedded in regions rich in U and also A, while G+U content compared to U content indicates that Gs are more often found in regions flanking unbound motifs, similar to the human motif sets.

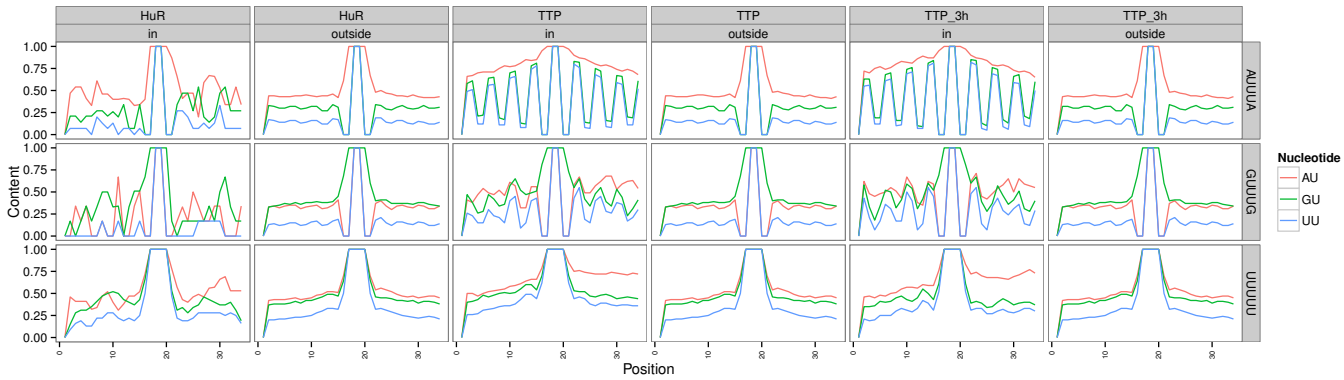


Figure 17: **Mean di-nucleotide composition of AU/GU/U-rich motifs in mouse** A comparison of the mono-nucleotide composition of AU/GU/U-rich elements overlapping CLIP signal of HuR and TTP (3 h and 6 h after LPS induction) and without overlap, similar to figure 14

The di nucleotide content reveals no new results compared to the mono-nucleotide content. Again AU dinucleotides are very common flanking motifs bound by TTP in both datasets, while G containing dinucleotides are rather rare.

The three RBPs investigated here are known to prefer single stranded regions for binding. This leads to the hypothesis, that active motifs are found in a more single stranded surrounding than others. As AREsite2 contains RNAplfold derived accessibility predictions for all regions overlapping annotated motifs, the positive and negative sets could easily be examined for their accessibility.

Figures 18 and 19 show probabilities of being unpaired over a stretch of 5nt along a 35nt long region, embedding AU/GU/U-rich elements of positive and negative sets in their center, for all proteins of interest.

Comparable to the A+U content of motifs in the positive set, flanking regions around bound motifs are in general more accessible then those in the negative set. However, for Auf1 and TTP sites in the UUUUU motif class, the opposite is the case, here motifs in the negative set have a higher probability of being unpaired than those from the positive set. AUUUA class motifs in the TTP positive set show the highest difference in the regions shortly before and overlapping the motif. GUUUG class motifs in the Auf1 set show a similar picture, where the region upstream of the motif in the positive set is more accessible in comparison to the negative set, while the opposite is true downstream of the motif. For TTP and GUUUG class motifs the downstream region in the positive set is also more accessible than the upstream region.

In mouse, the HuR dataset is hard to interpret due the low number of available bindingsites, however, at least for UUUUU class motifs,

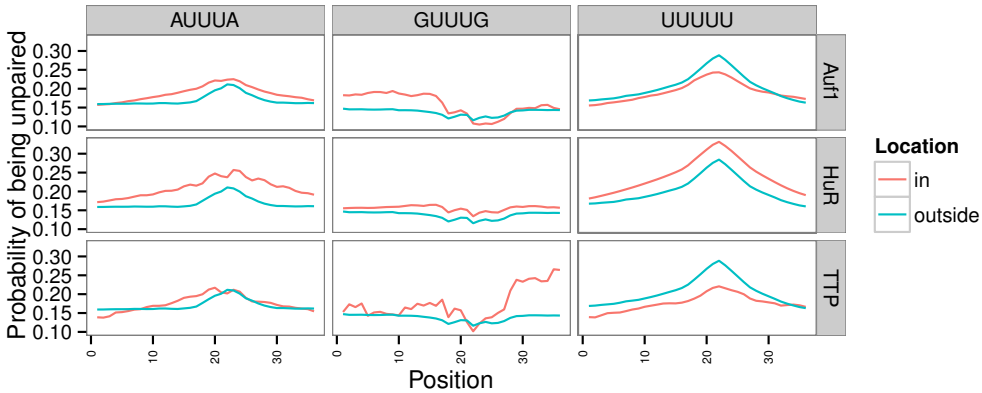


Figure 18: **Mean probability of being unpaired of AU/GU/U-rich motifs in human** A comparison of the accessibility of AU/GU/U-rich elements overlapping CLIP signal of Auf1, HuR and TTP and without overlap. Accessibility is measured in terms of probability of being unpaired over a stretch of 5 nucleotides, corresponding to the length of the investigated core motifs.

accessibility of motif embedding regions is higher for the positive set than for the negative set. TTP shows comparable accessibility of motif sets in both conditions, especially the negative set is very similar. Due to the high number of overlapping binding sites or binding sites in close proximity (see section 2.2.7.1), also the positive sets are comparable.

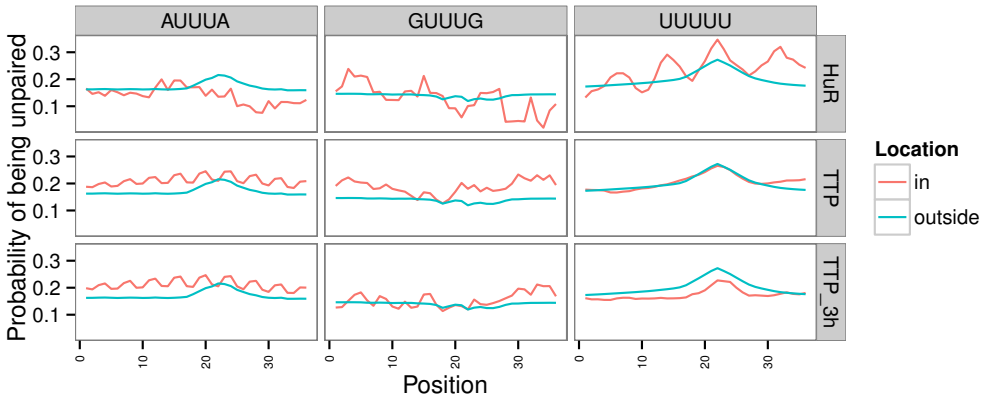


Figure 19: **Mean probability of being unpaired of AU/GU/U-rich motifs in mouse** A comparison of the accessibility of AU/GU/U-rich elements overlapping CLIP signal of Auf1, HuR and TTP and without overlap, similar to figure 18.

2.1.5 *The search for a discriminator*

After what we learned so far, higher AU/U content and accessibility of regions embedding active elements, at least for their core binding motifs, a point to discuss remains the predictive power of these findings. Predicting novel binding sites for ABPs is a challenging task. For this section data from AREsite2 was utilized to investigate the power of AU/GU/U-richness vs. accessibility as discriminator between bound and unbound elements.

In a first step the descriptive power of former described features is analyzed and visualized with receiver-operating-characteristic (ROC) curves. A ROC-curve is generated when the number of true positives is plotted against the number of false positives as a function of a threshold of a certain feature. In this case AU/GU/U-content and/or accessibility are used as thresholds to show how well one of these features describes if a certain sample is from the positive or negative set. The area under the ROC curve (AUROC) helps to compare how well a descriptor performs, the higher the AUROC, the higher its descriptive power. A ROC close to the diagonal (corresponding to an AUROC of 0.5) resembles a random selection, which means the descriptor is uninformative. A curve that goes below the diagonal is not a good descriptor, but its negative can still be useful as predictor.

Figures 20 to 22 show the results of this analysis for all three investigated ABPs in human and the discriminative power in terms of area under the ROC curve (AUC).

Figures 25 to 24 show the respective results for mouse. For each protein the motif with best results was selected, appendix A.1 contains plots for all motif-protein combinations.

2.1.5.1 *Human Auf1 targets*

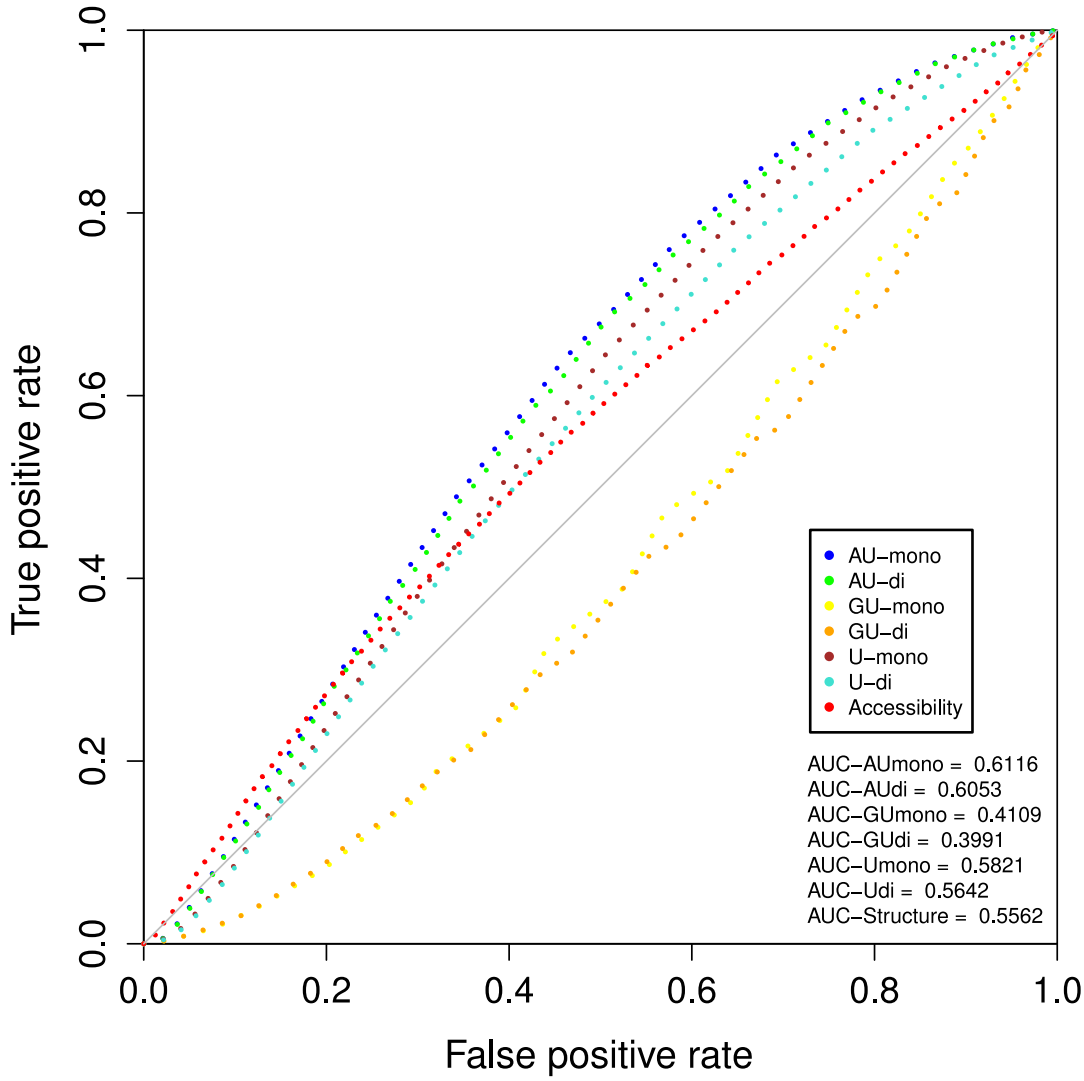


Figure 20: **Descriptor analysis of nucleotide content vs accessibility of Auf1 bound/unbound AU-rich elements (AUUUA) in human** ROC curve to visualize the descriptive power of accessibility and nucleotide content in terms of Area under the ROC curve (AUC).

For Auf1, which is known to bind AU- as well as GU- and U-rich elements, the highest predictive power lies in AU-content of sequences (see figures 20 and 46). Mono-nucleotide as well as di-nucleotide AU-content have the highest AUROC for all investigated motif classes. U-content and accessibility are weaker descriptors, GU-content is in all cases no valid descriptor. U-rich motifs targeted by Auf1 (see fig 46) can not be distinguished from unbound ones with any of the descriptors essayed here.

2.1.5.2 *Human HuR targets*

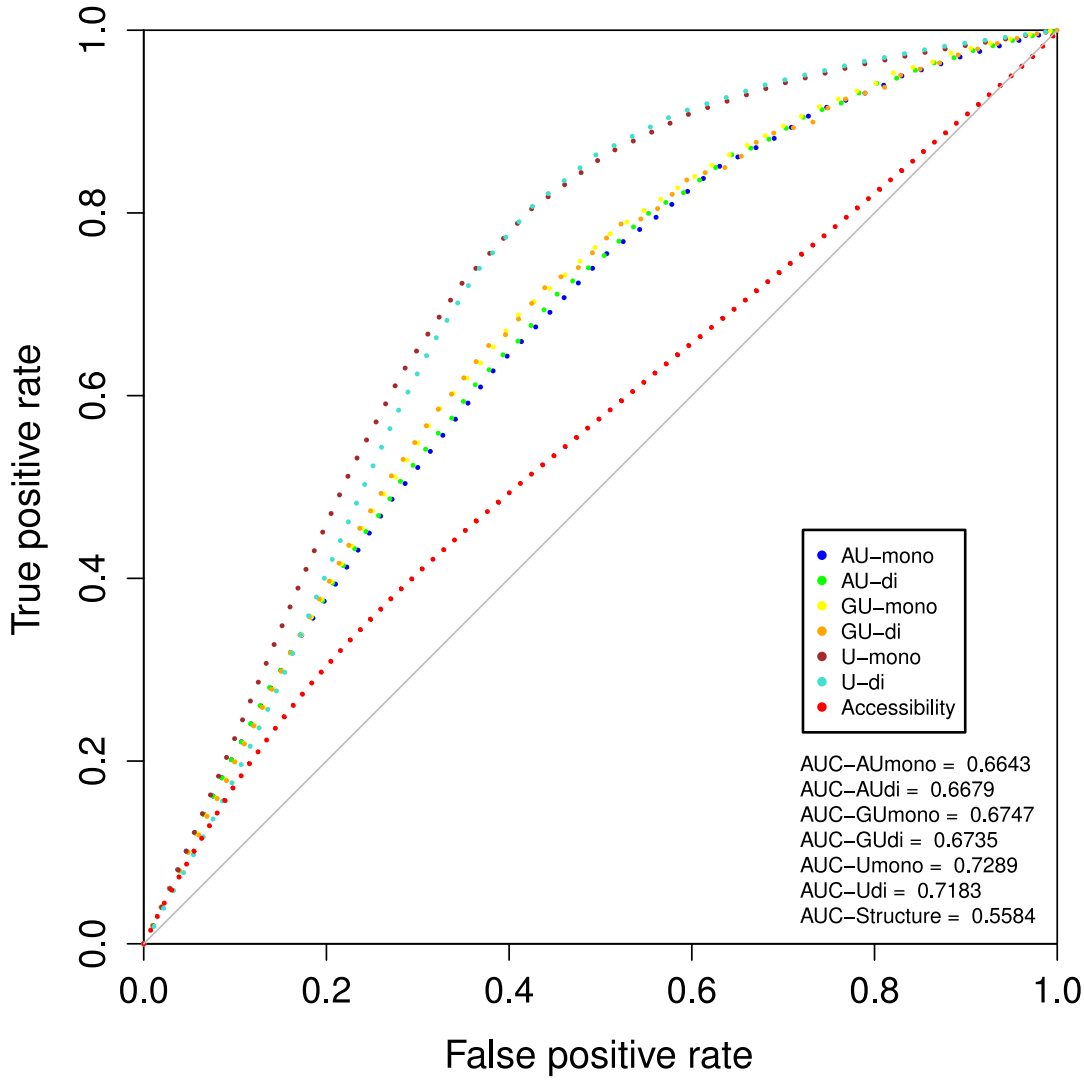


Figure 21: **Descriptor analysis of nucleotide content vs accessibility of HuR bound/unbound U-rich elements (UUUUU) in human** ROC curve to visualize the descriptive power of accessibility and nucleotide content in terms of Area under the ROC curve (AUC).

For all subsets of HuR targets, mono- and di-nucleotide content have a higher potential as descriptors than motif accessibility (see figures 21 and 47). Depending on the investigated motif family, either AU- or GU- di-nucleotide or U- mono-nucleotide content can be considered reasonable descriptors for bound and unbound motifs. Accessibility of motifs is a rather weak descriptor for HuR targets. This is in direct contrast to the findings from our datasets in section 2.2.11.4.

2.1.5.3 *Human TTP targets*

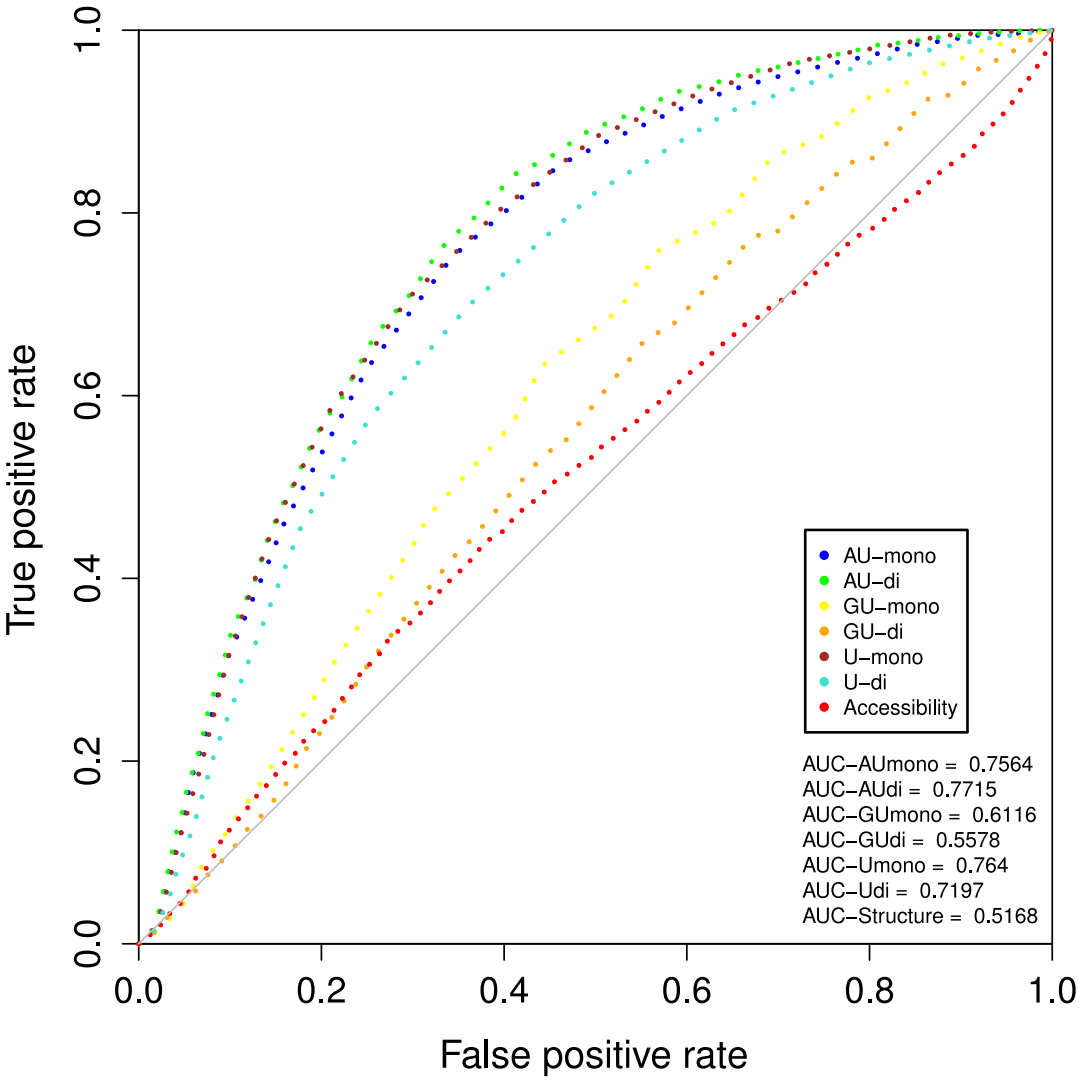


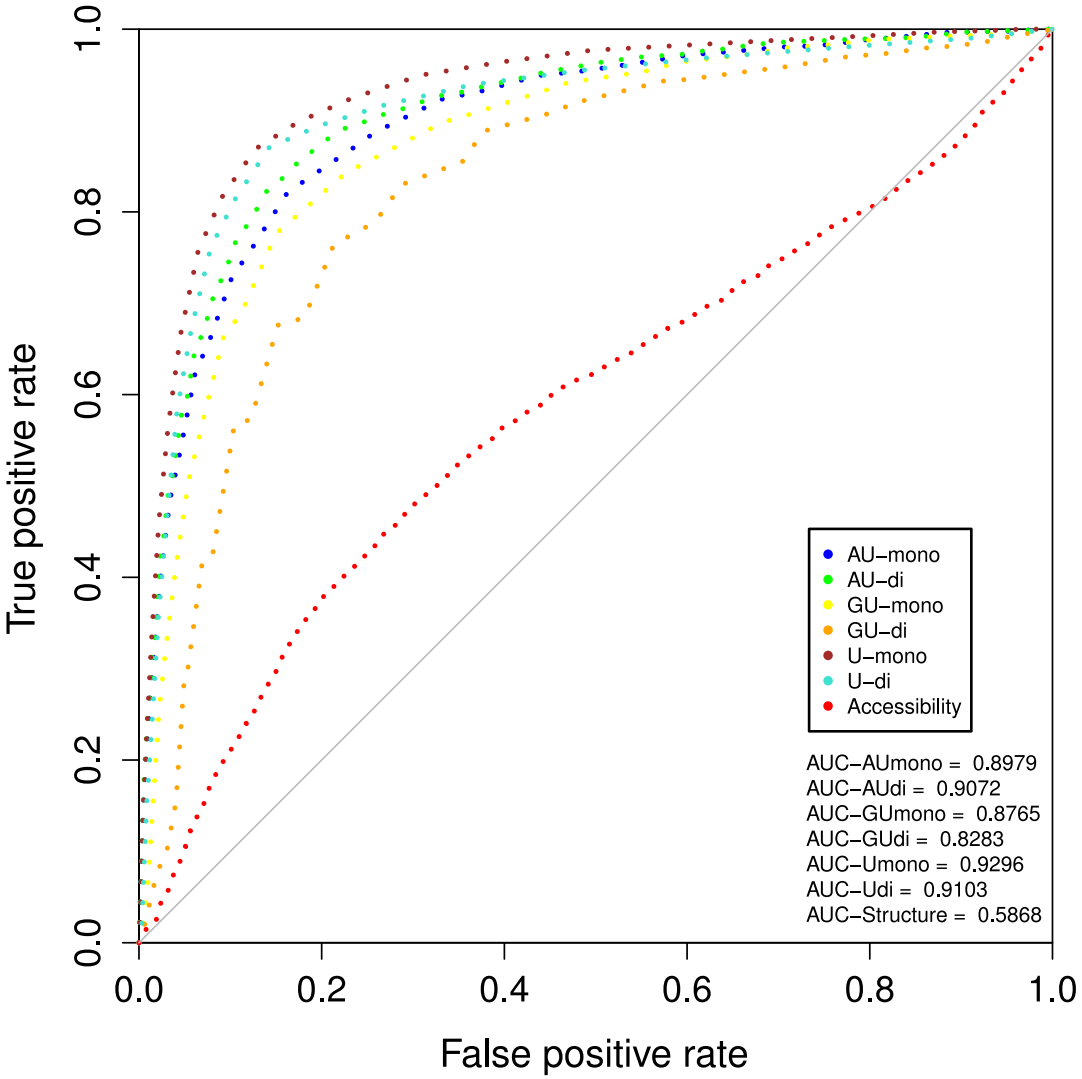
Figure 22: **Descriptor analysis of nucleotide content vs accessibility of TTP bound/unbound AU-rich elements (AUUUA) in human** ROC curve to visualize the descriptive power of accessibility and nucleotide content in terms of Area under the ROC curve (AUC).

The picture for TTP targets resembles that of HuR targets in human, with the difference, that AU-mono and di-nucleotide content are in every case the best (in the case of U-rich motifs even the only valid) descriptors, followed by U-mono-content (see figures 22 and 48). Accessibility of motifs is again either a weaker, or in case of U-rich motifs, no useful descriptor.

Summing up, AU-content seems to be a valid descriptor for all three ABPs, followed by U-mono-nucleotide content. GU-content is a weak or no valid descriptor in most cases. Accessibility of motifs could not

be shown to be a useful, or better descriptor than nucleotide content in any dataset.

2.1.5.4 *Mouse TTP targets 3 h after LPS induction*



AUUUA

Figure 23: **Descriptor analysis of nucleotide content vs accessibility of TTP bound/unbound AU-rich elements in mouse, 3 h after LPS induction** ROC curve to visualize the descriptive power of accessibility and nucleotide content in terms of Area under the ROC curve (AUC).

TTP targets of the AU-rich type can be well distinguished from unbound motifs by their nucleotide content, best by U-content, followed by AU-content (see figures 23 and 50). Embedding regions of GU- and U- rich motifs are better described by AU-content. For the first time, even GU-content shows some descriptor potential, although not comparable to AU- and U- content. Accessibility of motifs has descriptive

power only in case of AU-rich motifs, again weaker than nucleotide content, but AU-rich motifs have to be considered the preferred binding motifs of TTP.

2.1.5.5 *Mouse TTP targets 6 h after LPS induction*

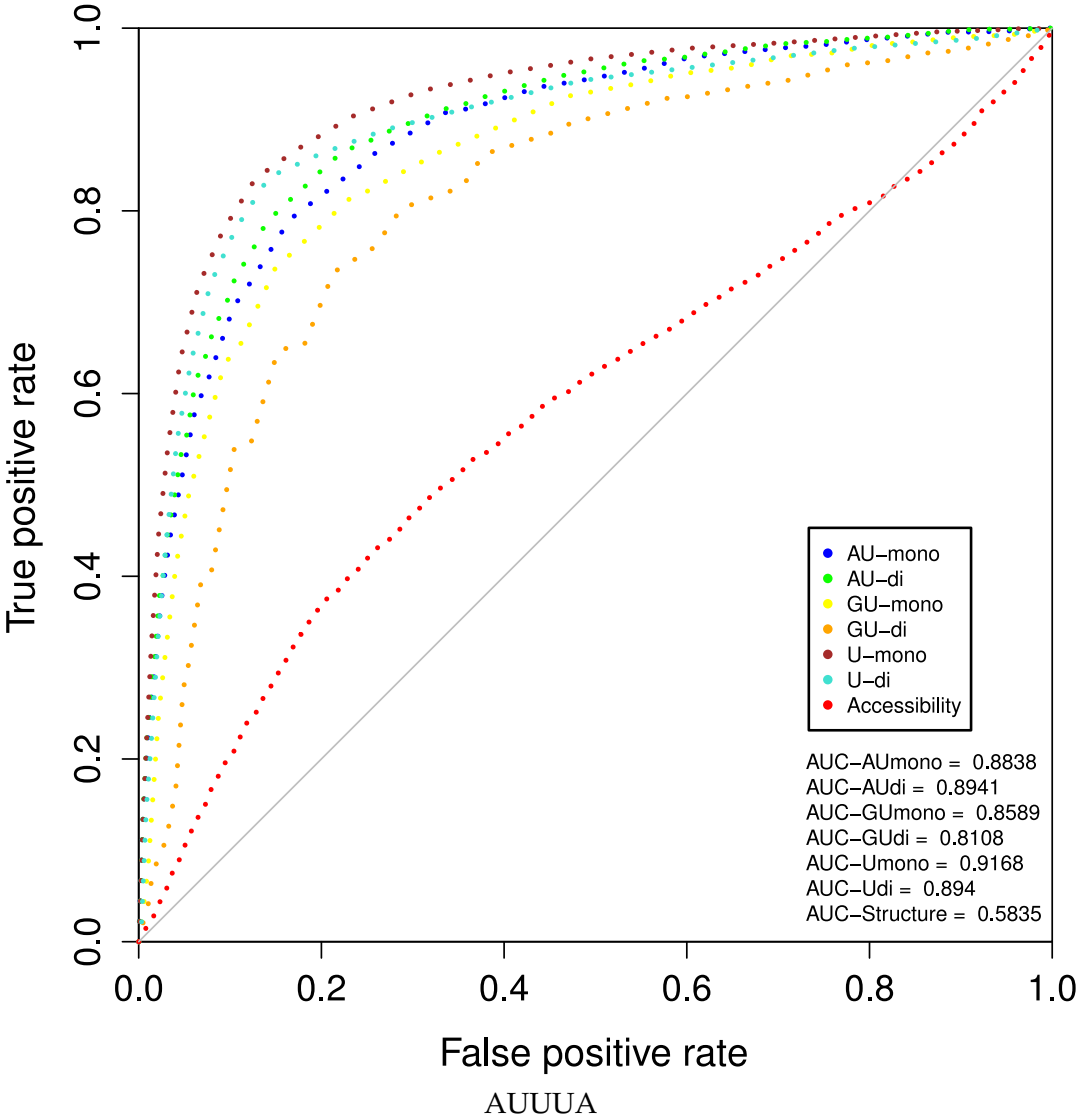
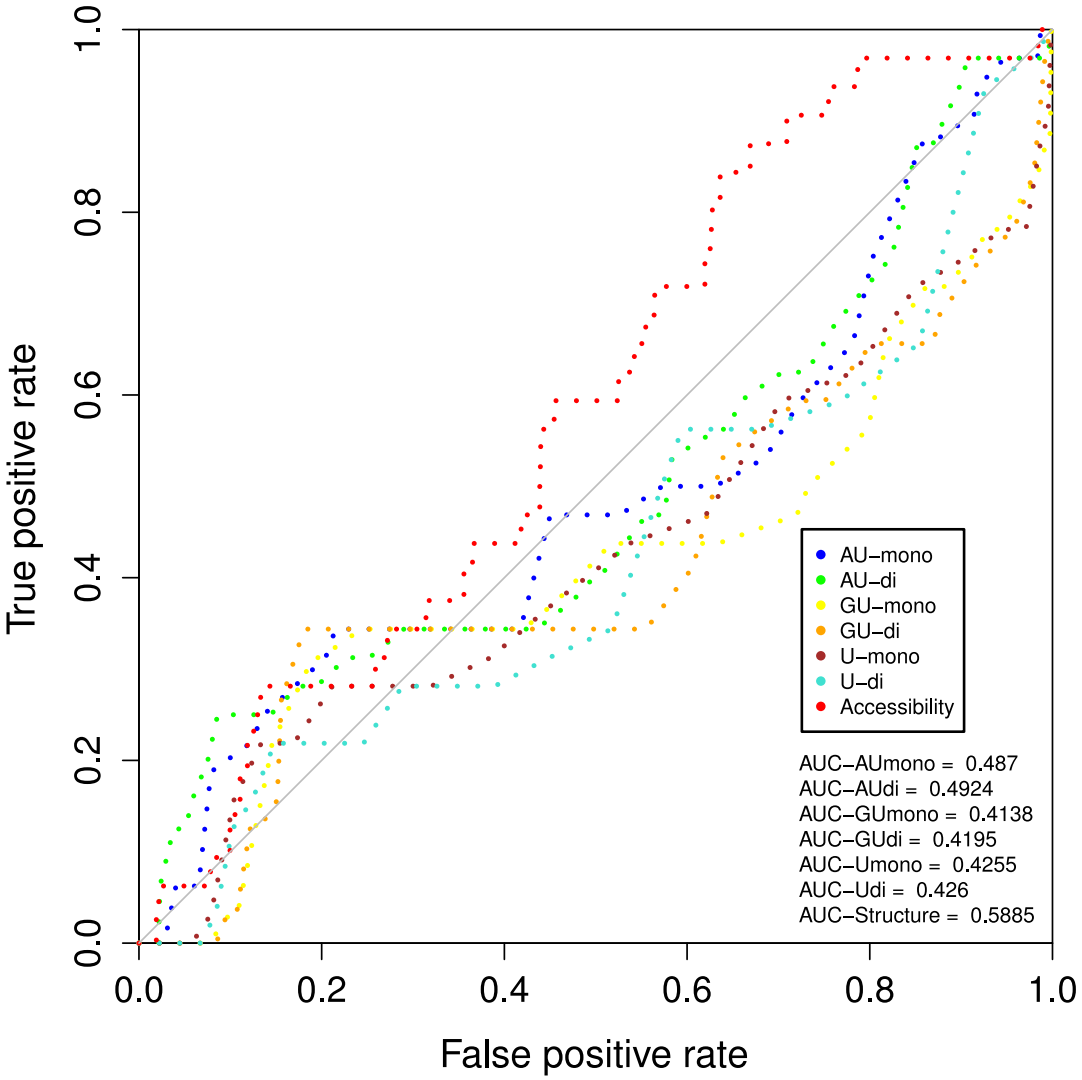


Figure 24: **Descriptor analysis of nucleotide content vs accessibility of TTP bound/unbound AU-rich elements in mouse, 6 h after LPS induction** ROC curve to visualize the descriptive power of accessibility and nucleotide content in terms of Area under the ROC curve (AUC).

6 h after LPS induction, TTP binding sites can still best be distinguished from unbound sites by their AU- and U- content, presenting a similar picture than 3 h after LPS induction (see figures 24 and 51). In case of the 6 h dataset, accessibility is a slightly better descriptor than for the 3 h dataset.

2.1.5.6 *Mouse HuR targets*



UUUUU

Figure 25: **Descriptor analysis of nucleotide content vs accessibility of HuR bound/unbound U-rich elements in mouse** ROC curve to visualize the descriptive power of accessibility and nucleotide content in terms of Area under the ROC curve (AUC).

In contrast to human target sites, HuR targets of AU- and GU-rich families, show no promising descriptor in our set of descriptors in mouse (see figures 25 and 49). Only the U-rich class of motifs has a potential, although not very good, descriptor in form of accessibility of motif embedding regions. This can partially be explained by the small set of binding sites available and the majority of those not in overlap with AU- and GU-rich motifs, while U-rich motifs are clearly favored. However, this effect was already visible in section 2.1.4. If none of the investigated features show potential as descriptors, two possible explanations come to mind. Either the true discriminating

feature was not part of the analysis, or presence of a motif alone is indeed the only necessity for binding.

Summing up, AU- and U- content of motifs and embedding regions are promising descriptors for TTP targets and unbound sites in human and mouse. Auf1 and HuR targets in human can also be described by AU content, while HuR targets could not be distinguished by any of the presented descriptors in mouse. Accessibility of motifs shows only low descriptive potential for the ABPs and motif families investigated here. When compared to section 2.2.11.4, where PAR-iCLIP derived datasets were normalized to transcript expression levels and filtered for biological relevance, it becomes obvious, that an analysis as conducted here is strongly influenced by the quality of the available data and downstream analysis.

In section 2.2.11.5 features extracted from PAR-iCLIP defined binding sites were also used to train a linear discriminator and assess its predictive power with the here presented datasets as testsets.

2.2 PAR-iCLIP of TTP and HuR in primary mouse macrophages

To directly address the role of TTP and HuR in the inflammatory macrophage transcriptome, PAR-iCLIP experiments in LPS induced primary mouse macrophages were analyzed. The raw dataset consists of PAR-iCLIP experiments in TTP 3 h and 6 h after LPS induction, and HuR 6 h after LPS induction in $TTP^{+/+}$ and $TTP^{-/-}$ primary macrophages.

As we have shown in Sedlyarov et al. [120] direct influence of TTP binding on RNA half-life is observable and only a handful of binding sites seem to be targeted by TTP and HuR in a directly antagonistic manner. Furthermore, we focus on the quantification of our CLIP analysis with RNA-Seq data, to account for transcript expression rates, and show the influence of secondary structure vs. AU-richness on functionality of ARE motifs.

Recently published Ago-CLIP-Seq sites in mouse macrophages were integrated and analyzed for overlaps with our dataset. GO analysis of identified target genes in all examined conditions conclude this chapter.

2.2.1 *Processing of PAR-iCLIP reads*

Preprocessing of reads retrieved from PAR-iCLIP and RNA-Seq protocols includes demultiplexing, discarding of PCR artifacts, barcode trimming and removing adapters with `Cutadapt` [91]. Statistical analysis with FASTQC (S. Andrews: `http://www.bioinformatics.babraham.ac.uk/projects/fastqc/` was conducted to validate read quality and clipping efficiency (results not shown).

Reads were mapped to the mouse genome (*Mus musculus*, assembly NCBI m37 (April 2007, strain C57BL/6J)) with `Segemehl` [50]. Only uniquely mapped reads were used for further analysis to avoid ambiguous binding signal.

Figure 26 shows the number of PAR-iCLIP reads remaining after each processing step.

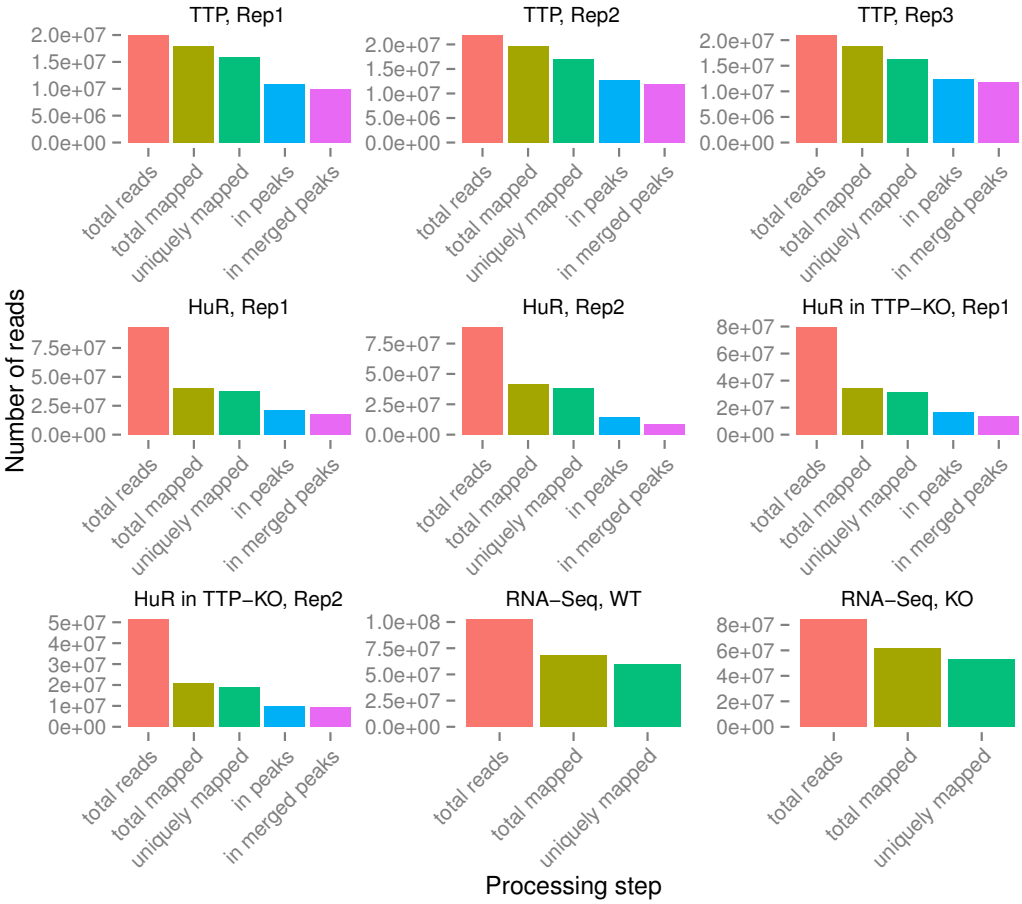


Figure 26: **Amount of remaining Par-iCLIP-Seq reads after each processing step for all samples** Total reads in sample, total mapped reads, uniquely mapped reads, reads in peak regions and reads in peak regions we see in all replicates of a sample.

With more than 10×10^6 uniquely mapped reads per replicate in the final peak set, our CLIP-Seq dataset has higher sequencing depth than comparable ones [143], allowing us to apply stringent filtering without loosing too many true positives.

2.2.2 *Crosslink site extraction and analysis*

CLIP [131] (Crosslinking and Immunoprecipitation) is a method to study interactions between nucleic acids and proteins. A key feature of CLIP techniques is to establish a covalent crosslink between RNA bases and aromatic protein residues via UV light. Par-CLIP [45] (Photoactivatable-Ribonucleoside-Enhanced Crosslinking and Immunoprecipitation) was developed to increase the amount of crosslinked protein-RNA residues.

Cells are incubated with thio-uridine, a photoactive uridine analog, which is incorporated into newly transcribed RNA. UV treatment at 365nm ensures site-specific crosslinking between aromatic amino acid residues and thio-uridine. Analysis of crosslink sites are based on T2C transitions, which occur when reverse transcriptase (RT) reads through a crosslink site and interprets thio-uridine as guanine, thus introducing a cytidine into the cDNA strand. iCLIP [67] (Individual-nucleotide resolution UV Cross-Linking and Immunoprecipitation) on the other hand uses UV-light at a standard wavelength for unspecific RNA-protein crosslinking at 254nm, with the difference that only the 3′ RT-primer is annealed before the cDNA synthesis step.

Reverse transcriptase tends to drop-off of the RNA template when encountering a crosslink-site, so that the cDNA strand ends one nucleotide before the crosslink site, thus allowing identification of crosslink sites with nucleotide resolution. For more details please refer to section 1.4.2.2.

CROSSLINK SITE EXTRACTION The here used PAR-iCLIP method combines advantages of both techniques, high yield and nucleotide resolution of crosslink sites. To take full advantage of this high resolution, only the theoretical crosslink site, i.e. the position one nucleotide upstream of the read start was considered, rather than the entire read. Using the whole read would lead to a signal shift away from the actual crosslink site, resulting in artificial peak patterns and binding site analysis.

2.2.3 *Peak finding and filtering*

A peak is defined as a region with a significantly higher number of read pileup at a given genomic position than would be expected by chance. The Pyicos [3] ModFDR method was applied for peak finding, together with a modified filtering algorithm for the use with PAR-iCLIP crosslink sites, which can be seen as reads of length one. Due to the nucleotide resolution of PAR-iCLIP , peak width can range from one nucleotide for very sharp signals, to several hundred nucleotides for regions with e.g. multiple consecutive binding sites. Our custom filtering method splits peak regions surrounding the highest peak

signal, henceforth named summit in accordance with Pyicos, once certain height-thresholds are reached.

Cutoffs were defined based on signals detected in known TTP targets. Peaks with a summit signal below 100 pileups are considered background and discarded. With a sliding window approach, starting from the summit, a peak is first split when its height falls below 30% of the summit signal. Emerging subpeaks with a summit above this cutoff and 100 pileups are then recursively split when their signal falls below 10% of their summit.

Final split-peaks contain a high amount of crosslink signal and allow to analyze protein binding sites with high resolution. Replicates of each experimental setup were analyzed separately. Width and position of peaks vary slightly between experiments. For the ranked lists of TTP and HuR target genes, peaks from all replicates were collected and peaks that do not have an overlap with peaks in all other replicates were filtered out (see 'in_merged_peaks' in fig. 26).

Resulting filtered peaks were then subject to downstream analysis, e.g. annotation and motif analysis.

2.2.4 *Transition analysis*

In order to verify the specificity of PAR-iCLIP for thymidines as cross-link sites, the nucleotide composition around the 5′ ends of all reads was analyzed.

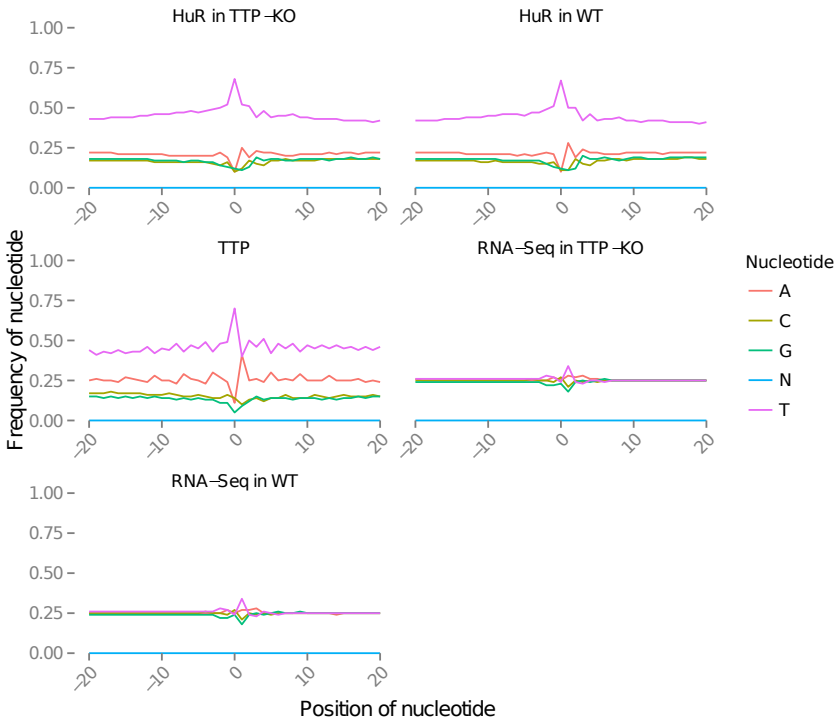


Figure 27: **Nucleotide distribution around 5′ends of reads.** To verify the specificity of PAR-iCLIP for Ts as cross-link sites, the nt composition around the 5′ ends of all reads was analyzed. Position 0, which is the putative crosslink site, is in ~ 66% of the cases thymidine. The same analysis has been conducted for RNA-Seq experiments, showing a more or less equal distribution of nucleotides along the reads.

The first nucleotide upstream of the 5′-end of each read was extracted, here called position 0, which represents the potential crosslink site. As expected, the majority of all reads (~66%, fig. 27) show a thymidine at this position of the reference genome.

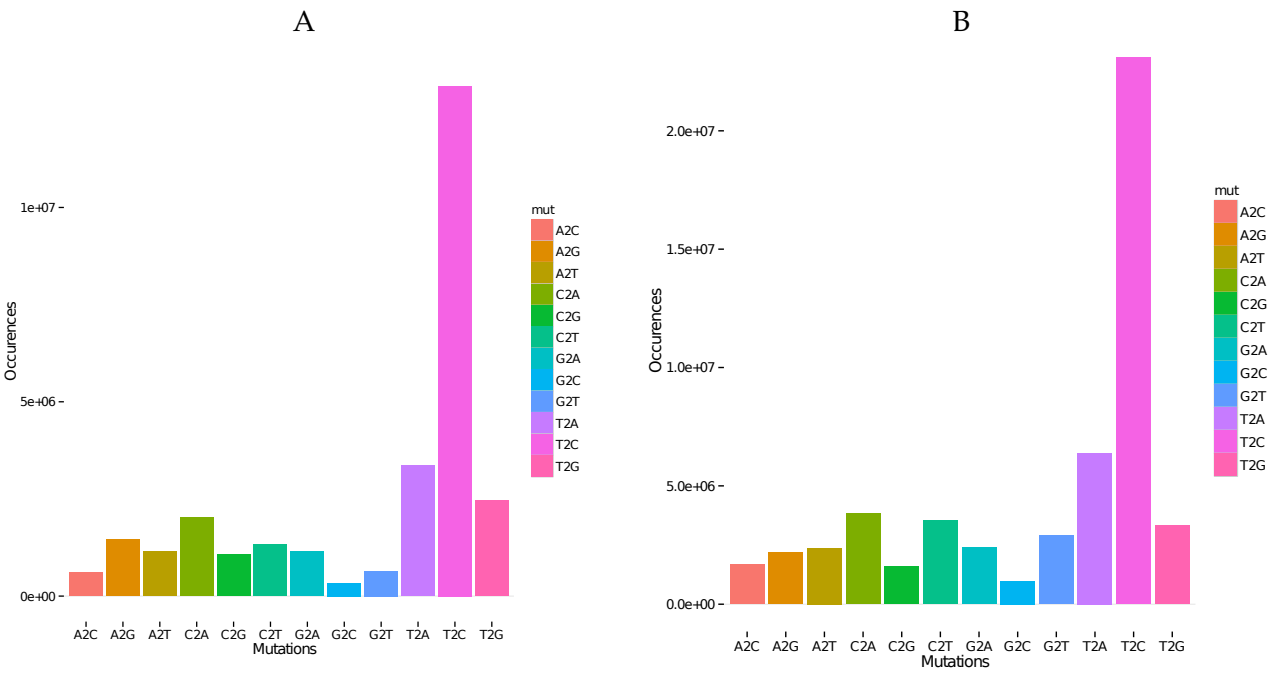


Figure 28: **Transitions in TTP PAR-iCLIP** When reverse transcriptase encounters a thiouridine in the RNA template, it is recognized as guanine, resulting in the incorporation of a cytidine in the newly synthesized cDNA strand. Later on this T2C transitions can be used to detect read-through events. (A) T2C transitions represent ~ 46% of all occurring transitions and transversions in TTP samples. (B)show a similar ratio of T2C to other mutations in HuR in WT, HuR in KO is equivalent to the WT dataset and not shown

To check whether these T were indeed involved in a cross-link, a 'transition map' of all T2C transitions observed within reads was constructed. 8.9 million reads (55%) contain at least one T2C transition, which makes up 2.4 million unique positions on the genome. T2C transitions represent ~ 46% of all occurring transitions and transversions in TTP samples (fig. 28). Furthermore, 9.65 million. reads (59%) show a T on position 0 that is contained in our transition map and thus has been used for cross-linking at least once. 78% of all reads have a T2C transition within 3 nucleotides of position 0 and make up 94% of final peak regions which are subject to further analysis.

2.2.5 *Genomic distribution of binding sites*

2.2.5.1 *Gene Annotation for human and mouse*

Mouse genome assembly mm9/GRCm37 (source: ENSEMBL [33] and genome annotation ENSEMBL v67 `http://may2012.archive.ensembl.org/Mus_musculus/Info/Index` were used for annotation. Genomic coordinates of all protein coding genes were retrieved via the ENSEMBL Perl API `http://www.ensembl.org/info/docs/api/core/index.`

html#api on all 3 levels (gene, transcript, exon). Mouse-human orthologs were retrieved via the ENSEMBL Biomart tool [64].

2.2.5.2 *Annotation of binding sites in ENSEMBL mouse genes*

Crosslinks derived from uniquely mapped reads in peak regions were annotated with ENSEMBL derived information.

For gene statistics an exon first approach was applied, where all transcript isoforms of a target gene are taken into account: A peak region is classified as exonic if it occurs in an exon of at least one transcript isoform; it is intronic if it occurs in an intron of at least one isoform and never in an exon.

2.2.5.3 *Binding site distribution*

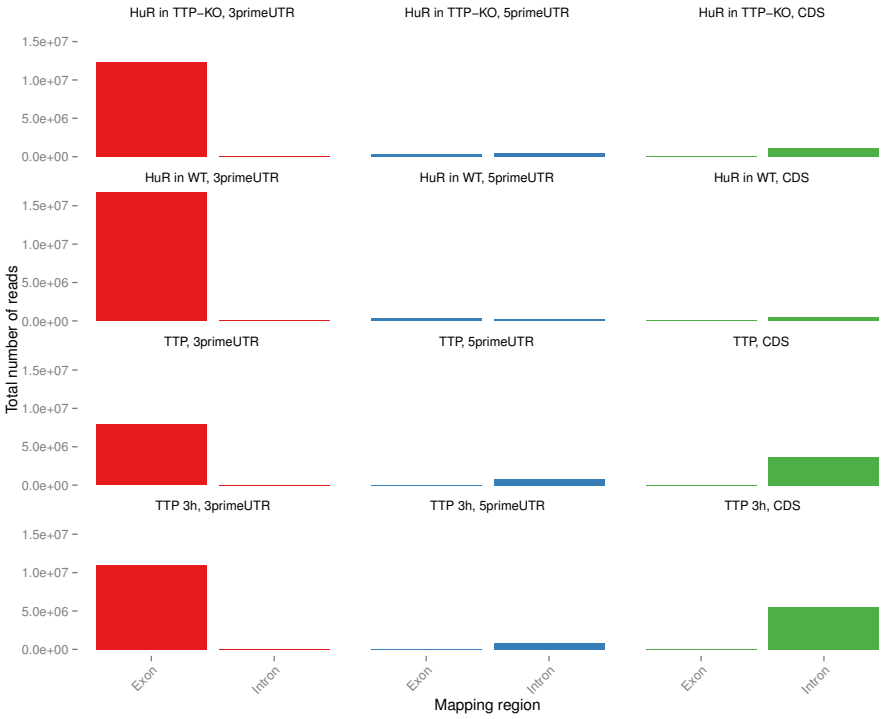


Figure 29: **Genic distribution of crosslinks in peaks** for one representative replicate of TTP and HuR in WT and TTP-KO BMDMs after filtering.

Both TTP and HuR favor binding to 3'UTRs. However, TTP also shows a prominent amount of signal in intronic regions, especially introns in the CDS of target genes in both conditions. Figure 29 shows the intronic and exonic localization of PAR-iCLIP reads in 5'UTRs, CDS and 3'UTRs. 63% (63% 3 h LPS) of total TTP PAR-iCLIP signal map to 3'UTR exons, 30% (32% 3 h LPS) to CDS introns and 6% (5% 3 h LPS) to 5'UTR introns. The remaining reads are found in exons of 5'UTRs and CDS. 10% of TTP derived PAR-iCLIP signal in introns

originates from a single intron (intron 4) of Immune-responsive gene 1 (Irg1) (fig. 29, tab. 26).

For HuR, 94% (WT) and 87% (TTP-KO) of total PAR-iCLIP signal maps to 3'UTR exons, 2% (WT) and 8% (TTP-KO) to CDS introns and 1% (WT and TTP-KO) to CDS exons. 3% of the signal in WT cells originates from 5'UTRs, (2% exonic, 1% to intronic), in TTP-KO this number is slightly higher with 5%, we detect 3% intronic signal and 2% exonic.

Table 6 shows the number of peaks and PAR-iCLIP signal in introns and exons for TTP 6 h and HuR in WT and TTP-KO. 5' UTRs were not considered here given the low amount of signal derived from this genomic element (fig. 29). Only 6 h experiments are included for better comparability with HuR.

A total of 498 genes are bound by TTP (6 h LPS). We find more peaks in CDS introns (855) than in 3'UTRs (566) and also more genes with intronic peaks (337 intronic, 196 3'UTR). However, the highest amount (~66%) of PAR-iCLIP signal is derived from peaks in 3'UTRs. In both HuR conditions clearly more and stronger peaks can be found in 3'UTRs (1,935 peaks/16,639,802 reads in WT and 1,465 peaks/14,033,808 reads in TTP-KO) than in introns (179/598,838 reads in WT and 434 peaks/1,501,148 reads in TTP-KO) and more genes show 3'UTR binding, 234 (365 in TTP-KO) genes with 3'UTR peaks, 77 (212 in TTP-KO) genes with intronic peaks.

SUMMARY Most binding signal of TTP and HuR in primary mouse BMDMs, independent of the analyzed condition is located in exonic regions of 3'UTRs of protein coding genes. In case of HuR this preference holds true for both, peak numbers and CLIP-Seq signal, for TTP this is only true for signal.

At this point it is to be considered, that peak numbers depend directly on the cutoffs set during peak filtering, thus we consider the amount of CLIP-Seq signal, which is directly derived from the number of reads containing crosslinks, a more stable indicator for strong/weak targets.

However, the high amount of intronic binding indicates some function for these regions. A possible explanation could be titration of TTP via circular intronic RNA sponge molecules as another layer of regulation.

Although RNA-Seq data of analyzed BMDMs is available, no circular RNA fragments could be identified. This, however, should be investigated in separate experiments, designed for the detection of circular RNAs.

2.2.6 *Quantification and Normalization of RNA-seq and PAR-iCLIP data*

PAR-iCLIP can not distinguish between poor binding sites in highly expressed targets and good binding sites in targets with low expression. In order to introduce a measure for binding site strength, RNA-Seq experiments of primary BMDMs (WT and TTP-KO) upon 3 h and 6 h of LPS induction [120] were performed. Expression rates were calculated using `Cufflinks` v.2.0.2 [128] with ENSEMBL exon annotation as regions of interest.

The resulting FPKM values for each expressed transcript isoform were then used to normalize PAR-iCLIP derived signals on two levels. Peak areas were normalized by the sum of transcript FPKMs that overlap the peak region to define $\mathrm{PeakScore}_{\mathrm{normalized}}$ (eq 1).

$$\mathrm{PeakScore}_{\mathrm{normalized}} = \frac{\mathrm{PeakArea}}{\mathrm{FPKM}_{\mathrm{transcript}}} \tag{1}$$

Gene-wide PAR-iCLIP signal was normalized by gene expression rates ($\mathrm{FPKM}_{\mathrm{gene}}$) to define $\mathrm{GeneScore}_{\mathrm{normalized}}$ (eq 2). $\mathrm{FPKM}_{\mathrm{gene}}$ were calculated as the sum of FPKMs of all transcript isoforms ($\mathrm{FPKM}_{\mathrm{transcript}}$) of each gene containing the peak region within the mature mRNA.

Only transcripts of FPKM $\geqslant$ 10 were considered, as we expect TTP targets under inflammatory stress to be overexpressed.

$$\begin{aligned} \mathrm{FPKM}_{\mathrm{gene}} &= \sum_{\mathrm{transcript\ contains\ peak}} \mathrm{FPKM}_{\mathrm{transcript}} \\ \mathrm{GeneScore}_{\mathrm{normalized}} &= \frac{\sum \mathrm{PeakArea}}{\mathrm{FPKM}_{\mathrm{gene}} + (\mathrm{median}_{\mathrm{FPKM}} * \alpha)} \end{aligned} \tag{2}$$

Sparse data correction (median $\mathrm{FPKM}_{\mathrm{transcript}}$ is added to $\mathrm{FPKM}_{\mathrm{gene}}$ before normalization) was applied, to avoid spurious high GeneScores of low expressed genes. Using this GeneScores (α was set to 1 for all tables shown) we were able to generate a ranked list of potential TTP (and HuR) target genes in mouse (tab. 22,25,23,24).

2.2.6.1 *Normalization of PAR-iCLIP signal*

In order to rank target genes independent of their expression and therefore see which ones play an important role in the model system, normalized GeneScores (equ. 2) were calculated.

To pass this filter, at least one transcript isoform has to (i) have an FPKM$\geqslant$10 and (ii) this isoform must contain at least one peak. This filter works well for exonic and 3'UTR regions, it does ,however, not allow to normalize intronic regions without the assumption that intron levels are comparable to exon levels.

The problem that arises is that tools for the calculation of transcript expression work on exon level, which makes sense in a biological way, as mature mRNAs are not thought to contain introns. As our dataset does not allow to distinguish cytosolic from nuclear fraction derived transcripts, it is thus impossible to quantify intron expression levels. Although for nascent RNA, introns will be available in comparable amounts than exons, co-transcriptional splicing and unknown intron stability make it impossible to infer these numbers from our dataset.

As regulatory function of ABP binding sites was so far only correlated with 3'UTR binding and as of now, no functional role of intronic binding could be established and to circumvent analysis based on too many assumptions, it was decided to exclude intron normalization from downstream analysis.

About a quarter of TTP 6 h target genes (133 of 498) are expressed above threshold, almost all (130) have TTP binding sites in 3'UTRs (tab. 6). For HuR, 43% and 32% (WT and TTP-KO, respectively) of targets are highly expressed, again, almost all binding sites reside in 3'UTRs.

Normalization allows to remove all unspecific or weak targets and downrank genes that show high TTP signal just because of mRNA abundance. High GeneScores indicate, that the gene is 'strongly' bound by TTP. Two possible scenarios can lead to identical gene scores: (i) high PAR-iCLIP signal exists because TTP binds every copy of mRNA and therefore has at least one high-affinity binding site or (ii) a gene has multiple binding sites, but of medium to low affinity.

Therefore the normalized PeakScore (equ. 1) is calculated, which reflects the strength of the interaction between RNA and protein in dependence of the expression rate of the targeted mRNA. With this information it is possible to identify binding sites that are preferably bound by TTP or HuR, i.e. sites that could directly influence mRNA regulation, presenting a list (see 27, 28, 29, 30) of candidate binding sites for further experimental analysis (each list represents one replicate of the corresponding experiment, after filtering for peaks that occur in all replicates).

Ranked lists of TTP (and HuR) target genes normalized by expression are presented in tables 22, 23, 24, 25). In contrast to the full list of targets, these genes can be seen as important actors in inflammatory response (highly expressed) and quantitative information about TTP/HuR affinity is available. This curated list of target genes was used for several downstream analysis steps (discriminator analysis, correlation analysis with RNA decay, etc.).

2.2.7 *TTP and HuR target genes revealed by PAR-iCLIP*

Tables 18, 19 and 20 show the top 10 target genes of TTP and HuR in WT and TTP-KO cells identified 6 h after LPS induction. Genes are ranked based on PAR-iCLIP signal.

Table 21 shows the top ten target genes of TTP after 3 h of LPS induction. Top targets after 3 h correlate strongly with those 6 h after induction. Most of the following analysis steps were conducted for the 6 h sample only to compare with HuR samples.

Among the top 10 targets of TTP (tab. 18) are many well studied target genes in macrophages[17, 18, 53, 69] and other cell types[71], as for instance Tnf-alpha, Cxcl2, Zfp36 (TTP), and Ccl3 . In addition, we detected targets where TTP binds not only to 3'UTRs, but to intronic regions, for instance Immuno-responsive gene Irg1.

Top targets for HuR in WT and TTP-KO contain more or less the same set of genes, but differ slightly in rank (tab. 19, 20). Again, among the top targets we find many genes known for their interaction with HuR [73], *e. g.* . ActB, Cd44, Marcks.

The full lists of target genes, including gene expression rates from accompanying RNA-Seq experiments, PAR-iCLIP signal and gene annotation is provided as supplemental material of Sedlyarov et al. [120].

2.2.7.1 *Gene counts*

116 (WT) and 168 (TTP-KO) of the 499 TTP target genes identified in this study are also bound by HuR (fig. 30A). 279 genes are bound by HuR in both WT and TTP-KO, where 170 genes are not targeted by TTP at all, and 109 genes are also bound by TTP. Only 7 genes are targeted by TTP and HuR in WT, but not in TTP-KO, while 59 TTP targets are only bound by HuR under the absence of TTP. 323 genes show only TTP binding sites, and 381 are exclusively bound by HuR (17 only in WT, 194 only in TTP-KO).

Figure 30B shows the number of binding sites in our samples, with and without direct overlap, where the majority of TTP binding sites do not directly overlap HuR binding sites in both conditions.

To address a possible direct antagonistic behaviour of TTP and HuR when targeting the same mRNA, overlaps in binding regions from TTP and HuR PAR-iCLIP in WT and TTP-KO (tab. 7) were analyzed . `BEDtools v2.17` [108] was used to compare binding sites within and between all experimental settings. Only overlaps on the same strand (-s) with minimum overlap of 1nt were considered.

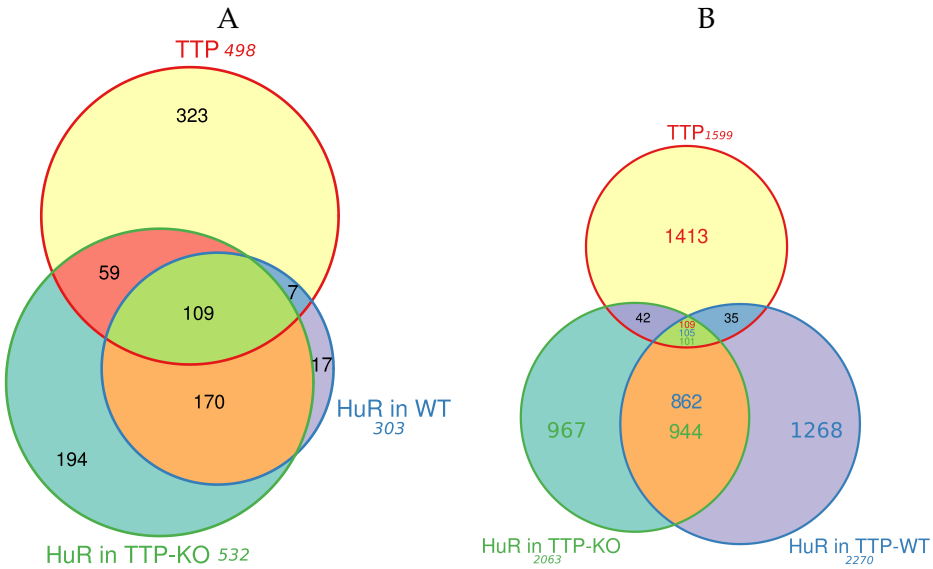


Figure 30: **Venn diagram (A)** shows the number of genes in our samples that either contain no or at least one peak of one or two of the other ABPs. In contrast to table 7, these peaks do not have to overlap. **(B)** This Venn diagram shows the number of binding sites with and without overlap in our samples color-coded by ABP of numbers differ

Table 6 summarizes numbers of peaks, genes with peaks, transcripts with peaks, and PAR-iCLIP signal in peaks for all samples 6 h after LPS induction and compares it to RNA-Seq normalized numbers.

Table 7 summarizes numbers of peaks and genes and PAR-iCLIP signal for TTP and HuR derived binding sites in overlap for all samples 6 h after LPS induction and compares it to RNA-Seq normalized numbers.

Table 6: TTP/HuR Peak Analysis. Shown are the number of genes, peaks and transcripts containing peak regions in total sample, 3'UTR and introns for one representative replicate of each sample after peak filtering and the amount of Par-iCLIP signal derived from this peaks. Also shown are these numbers for the RNA-Seq normalized samples, together with PeakScore and GeneScore.

	TTP	TTP 3'UTR	TTP CDS Introns	HuR	HuR 3'UTR	HuR CDS Introns	HuR in TTP-KO	HuR in TTP-KO 3'UTR	HuR in TTP-KO CDS Introns
Raw Mapping Data									
Number of peaks	1,598	566	855	2,286	1,935	179	2,074	1,465	434
Number of genes with peaks	498	199	337	303	238	77	532	361	211
Number of transcripts with peaks	1,125	347	712	654	484	154	1,187	710	448
Par-iCLIP signal in peaks	11,967,358	7,895,107	4,021,532	17,648,904	16,639,802	598,838	14,033,808	12,185,054	1,501,148
Raw Mapping Data normalized to RNA-seq									
Number of peaks	461	454	—	1,390	1,321	—	940	897	—
Number of genes with peaks	132	129	—	130	127	—	168	163	—
Number of transcripts with peaks	152	146	—	149	145	—	188	182	—
Mean PeakScore	153	154	—	169	172	—	219	221	—
Mean GeneScore	157	169	—	156	168	—	159	173	—

Table 7: TTP/HuR Overlap Analysis. Shown are the number of peaks and genes with peaks in direct overlap between TTP and HuR samples together with the signal in these peaks and signal at directly overlapping positions. Also shown are these numbers for the RNA-Seq normalized samples, together with the PeakScore of overlapping peaks and at directly overlapping positions.

	TTP overlap with HuR	TTP overlap with HuR in TTP-KO	TTP overlap with HuR in WT and TTP-KO	HuR with TTP	HuR in TTP-KO with TTP	HuR in WT with HuR in TTP-KO	HuR in TTP-KO with HuR in WT
Overlap between Samples							
Genes in direct overlap	51	67	42	51	67	256	256
Peaks in direct overlap	144	151	109	140	143	967	1045
Par-iCLIP signal in overlapping peaks	6,978,339	6,937,437	6,573,603	7,179,554	5,556,129	15,167,525	11,894,417
Par-iCLIP signal in overlapping peaks with nucleotide resolution	6,460,972	5,822,110	5,434,502	5,358,712	4,171,787	14,610,473	11,622,224
Overlap between Normalized Samples							
Genes in direct Overlaps	34	34	27	34	34	112	112
Peaks in direct Overlaps	110	100	81	105	93	557	607
PeakScore in overlapping peaks	46,586	41,346	40,701	16,133	12,260	190,320	144,113
PeakScore in overlapping peaks with nucleotide resolution	37,875	36,517	30,092	12,652	9,772	182,786	141,893

Directly overlapping TTP and HuR peaks were detected for 51 genes in WT and 67 genes in TTP-KO. 42 genes show overlapping peaks in both conditions. PAR-iCLIP signal from nucleotides at direct peak overlaps with HuR in WT cells is higher than in overlaps with HuR in TTP-KO. For binding sites used by TTP and HuR in both WT and TTP-KO, we observe average peak lengths of 21.71nt (TTP), 23.16nt (HuR in WT) and 24.13nt (HuR in KO) (not significantly different in two-sample t-test).

This insignificant difference indicates that TTP does not displace HuR from binding sites when both proteins are present, as would be the case when direct competition were the standard mode of interaction.

467 TTP target genes were identified 3 h after LPS induction, 142 targets can be found in the 6 h dataset but not in the 3 h dataset, and 166 in the 6 h and not the 3 h dataset. In most cases these genes do have TTP bound in each of the other datasets, but either very weakly or not in all replicates and are thus excluded during the filtering procedure.

On peak level, a total of 837 peaks are in direct overlap and 762 peaks not directly overlapping, with 212 peaks within 50nt distance to each other. As the stringent filtering and experimental noise have to be considered, it is likely that TTP targets the same genes and sites under both conditions, however, with differing affinity and quantity (see section 2.2.14).

SUMMARY TTP 3 h and 6 h after LPS induction preferentially target the same genes, only interaction strength and reproducibility between replicates differ. HuR in WT and TTP-KO do bind different sets of genes, although top targets remain mostly the same.

Antagonistic behaviour between TTP and HuR could not be observed as the default modus operandi for mRNA stability regulation in our experiments. Only a small subset of genes shows directly overlapping binding sites, indicating a more indirect regulatory mechanism. We could also not detect a significant influence on peak length in overlapping binding sites between WT and KO HuR-CLIP-Seq, which could have indicated active displacement of HuR in presence of TTP.

2.2.8 *Motif Analysis*

To find sequence motifs that act as biologically functional entities within our PAR-iCLIP datasets, k-mer analysis and motif enrichment analysis were conducted. MEME [8] finds overrepresented sequence motifs based on expectation derived from a background model. The commandline version of MEME was used to detect over-represented sequence motifs in the peak regions. MEME generates a background model based on nucleotide frequencies of input sequences.

Since the aim is to identify motifs that are enriched in specific genomic elements (introns and 3'UTRs) rather than all regions of PAR-iCLIP signal, individual background models for those genomic elements annotated in ENSEMBL protein coding genes were generated manually. Peak regions shorter than MEME's minimum sequence length (8nt) were extended on both ends to a minimal length of 9nt. Using the custom background models and the "any number of repeats" mode of MEME with motif-length of 7 yielded the best results regarding both motif information and gene coverage (see tab. 9).

2.2.8.1 *Analysis of sequence motifs in TTP and HuR binding sites*

Before over-represented binding motifs were analyzed with MEME, a simple k-mer count was conducted with KAnalyze [6] in TTP and HuR samples before and after RNA-Seq normalization. Top 5 7-mers overall, in 3' UTRs and intronic regions are summarized in table 8. Although k-mer analysis does simply count all k-mers in a given sequence set without enrichment or any other measure of significance, the motifs found this way resemble the MEME identified binding motifs of TTP and HuR already very well. The top 5 k-mers can be seen as rotations of the consensus UAUUUAU motif for TTP and UUUUUUU motif for HuR, including point mutations.

MEME analysis is similar to k-mer analysis focused on CDS introns and 3'UTR exons since these genomic partitions contain most of the PAR-iCLIP signal. In contrast to k-mer counts, the motifs derived from this over-representation assay compact all variation in a single motif, which allows to compare the information content of variation on sequence level.

Table 8: Top 5 7-mers in Par-iCLIP peaks for TTP and HuR samples in all, 3′UTR and intronic peak regions. Also shown are 7-mers derived from these regions in normalized datasets. Non-U nucleotides are colored for visualization purposes.

7mers		TTP				HuR-WT				HuR-KO		
not normalized	All	UUUAUUU	1206	6.72%	All	UUUUUUU	1842	7.92%	All	UUUUUUU	3506	13.47%
		UAUUUAU	1160	6.47%		UUUGUUU	374	1.60%		UUUGUUU	597	2.29%
		AUUUAUU	1115	6.22%		UUUUGUU	316	1.35%		UUGUUUU	499	1.91%
		UUAUUUA	1110	6.19%		UUGUUUU	316	1.35%		UUUUGUU	490	1.88%
		UUUUUUU	272	1.51%		UGUUUUU	209	0.89%		UUUCUUU	351	1.34%
	3′UTR	UAUUUAU	143	2.79%	3′UTR	UUUUUUU	1565	7.90%	3′UTR	UUUUUUU	2326	12.94%
		UUUAUUU	122	2.38%		UUUGUUU	310	1.56%		UUUGUUU	362	2.01%
		UUUUUUU	118	2.30%		UUGUUUU	254	1.28%		UUGUUUU	301	1.67%
		AUUUAUU	118	2.30%		UUUUGUU	253	1.27%		UUUUGUU	298	1.65%
		UUAUUUA	111	2.16%		UGUUUUU	180	0.90%		UUUCUUU	232	1.29%
	Intron	UUUAUUU	1083	8.54%	Intron	UUUUUUU	159	6.29%	Intron	UUUUUUU	1032	14.03%
		UAUUUAU	1013	7.99%		UUUGUUU	61	2.41%		UUUGUUU	233	3.16%
		UUAUUUA	997	7.87%		UUUUGUU	60	2.37%		UUGUUUU	194	2.63%
		AUUUAUU	996	7.86%		UUGUUUU	58	2.29%		UUUUGUU	190	2.58%
		UAUCUAU	163	1.28%		UGUUUUG	31	1.22%		UUUAUUU	128	1.74%
normalized	All	UUUUUUU	88	2.25%	All	UUUUUUU	834	6.44%	All	UUUUUUU	1053	9.57%
		UAUUUAU	87	2.23%		UUUGUUU	174	1.34%		UUUGUUU	180	1.63%
		UUUAUUU	59	1.51%		UUUUGUU	133	1.02%		UUUUGUU	139	1.26%
		AUUUAUU	59	1.51%		UUGUUUU	132	1.02%		UUGUUUU	137	1.24%
		UUAUUUA	58	1.48%		UUUUUGU	90	0.69%		UUUCUUU	106	0.96%
	3′UTR	UUUUUUU	88	2.29%	3′UTR	UUUUUUU	814	6.74%	3′UTR	UUUUUUU	997	9.33%
		UAUUUAU	85	2.21%		UUUGUUU	172	1.36%		UUUGUUU	178	1.66%
		AUUUAUU	59	1.53%		UUUUGUU	131	1.04%		UUUUGUU	137	1.28%
		UUUAUUU	58	1.51%		UUGUUUU	130	1.03%		UUGUUUU	134	1.25%
		UUAUUUA	57	1.48%		UGUUUUU	89	0.07%		UUUCUUU	105	0.98%

MEME analysis reveals that over-represented motifs derived from peak regions of TTP PAR-iCLIP differ slightly for 3'UTRs and intronic regions. While in 3'UTRs the already well described ARE heptamer *U[AU]UUU[AU]U* is detected, the U-rich *UUUAUUU* motif tends to be over-represented in intronic regions. Furthermore, intronic motifs may contain Cs in 2nt distance to the central A, which itself is more pronounced than in 3'UTRs. However, the motif found in intronic regions can be seen as shifted version of the 3'UTR motifs, thus there seems to be no significant difference in binding motif choice by TTP. This is in strong correlation to k-mer analysis results.

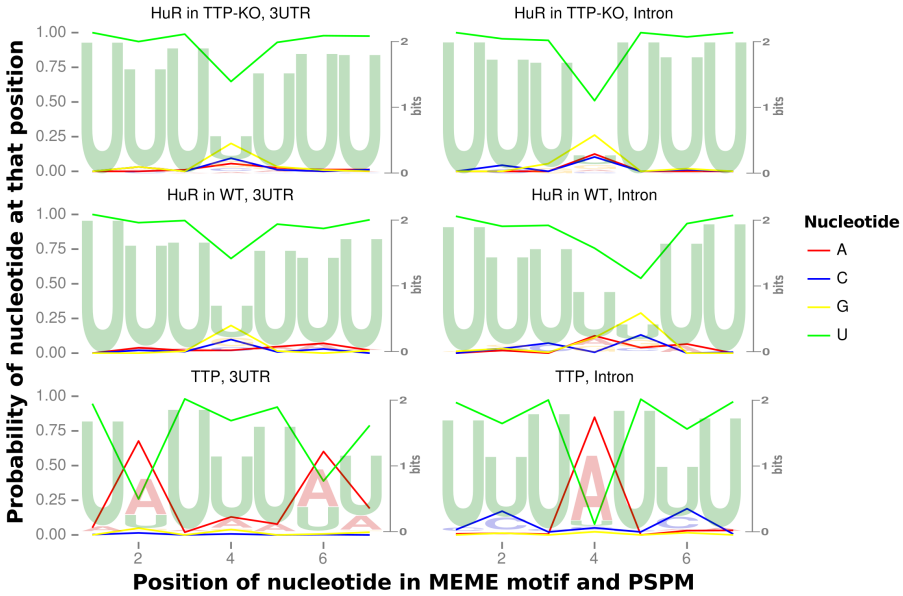


Figure 31: **MEME motif** Probability of a nucleotide to be found in the most over-represented MEME motif for peak regions of all samples divided in 3' UTR and Intron located peaks for comparison. The second Y-Axis and corresponding lines show the information bit-score of the motif.

The HuR datasets reveal only small sequence differences in peak regions of different elements. 3'UTR as well as intronic regions show a U-rich heptamer, which can contain guanine (G) or cytosine (C) around position 4. However, all extracted motifs show a high similarity to previously published ARE, or U-rich motifs. Figure 31 shows the probability for each nucleotide to be present at any position in the top over-represented MEME motifs for TTP, HuR and HuR in TTP-KO.

Figure 32 shows the 7mer MEME motif for TTP 3 h after LPS induction for bindingsites in 3'UTRs and Introns. Motifs have a high correlation with those in the 6 h dataset, indicating that TTP has a high affinity towards presented motifs throughout immune response.

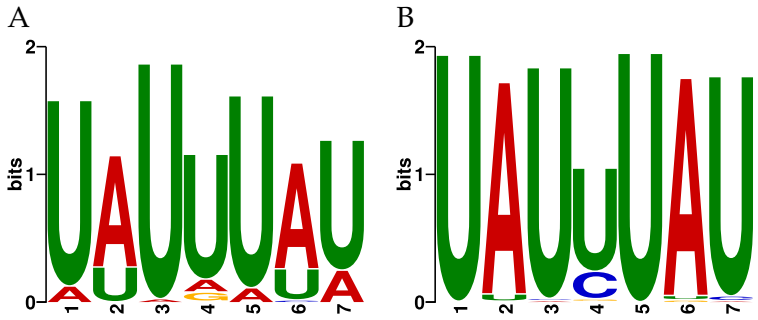


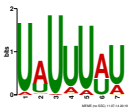
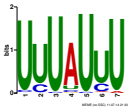
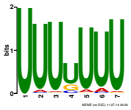
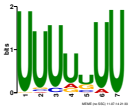
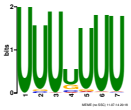
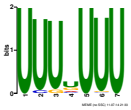
Figure 32: **TTP 3h motif in 3'UTRs and introns.** MEME 7-mer logos of **(A)** 3'UTR and **(B)** intronic binding sites

To check how much of our observations can be explained by the MEME consensus motifs, the fraction of CLIPed genes that contain the latter was computed, as well as the fraction of PAR-iCLIP signals that fall into this motifs. To do so, the sequences of all binding sites detected were scanned using the regular expression of the motif provided by MEME.

This regular expression represents the most likely form of the motif, i.e. not all possible isoforms of the motif are taken into account, only those nucleotides are considered that best fit the multilevel consensus.

Table 9 summarizes the motifs found and their occurrences. For this analysis and direct comparison to the HuR dataset, we focused on 6 h datasets only.

Table 9: **Par-iCLIP signal and target gene coverage of MEME motifs** derived from the regular expression describing the most probable motif per sample and genomic partition.

Sample	MEME regular expression and motif	% of total peak signal with overlap	% of total peak signal in overlap	% of total peak count	% of genes with peaks and motif	% of genes where peak and motif overlap	% of ENSEMBL mm9 protein coding genes with motif
TTP 3′UTR	U[AU]UUU[AU]U 	89%	36%	31%	96%	66%	59%
TTP Intron	UUUAUUU 	68%	34%	31%	96%	65%	84%
HuR 3′UTR	UUUUUUU 	64%	38%	14%	92%	83%	39%
HuR Intron	UUUU[UG]UU 	56%	30%	32%	95%	38%	90%
HuR in TTP-KO 3′UTR	UUU[UG]UUU 	74%	49%	29%	95%	89%	51%
HuR in TTP-KO Intron	UUU[UG]UUU 	66%	38%	50%	96%	62%	91%

89% of TTP PAR-iCLIP signal in 3'UTRs originates from the motif *U[AU]UUU[AU]U* and 68% of intronic signal maps to the motif *UUUAUUU*. 64% of HuR PAR-iCLIP signal in 3'UTRs is derived from the motif *UUUUUUU* and 56% of intronic signal from the motif *UUUU[UG]UU*. The most over-represented motif in both genomic elements for HuR in TTP-KO cells is *UUU[UG]UUU*. 74% of 3'UTR signal and 66% of intronic signal comes from this MEME motif.

To evaluate if and by how much TTP motifs in 3'UTRs and introns differ in usage by TTP, the same analysis as described, was conducted with swapped motifs. The amount of PAR-iCLIP signal detected within 3'UTR peaks when searched for the intronic motif decreased by 14% compared to the original one. In the other case, 10% more signal coverage is observed when using the 3'UTR motif in introns, highlighting TTPs preference of the known core motif, and that the intronic MEME motif is just a shifted version.

SUMMARY TTP does not discriminate intronic from exonic binding sites by sequence motif. Most of the PAR-iCLIP signal in our experiments results from TTP/HuR interacting with already well described motifs. Taken together, the strongest, most frequent and over represented TTP and HuR motifs identified here confirm the commonly known and published ones.

2.2.8.2 *Analysis of sequence motifs in TTP and HuR overlapping binding sites*

To further analyze TTP and HuR overlapping binding regions for differences with distinct binding sites, MEME motifs of the latter were computed and compared to motifs derived from non-overlapping sites.

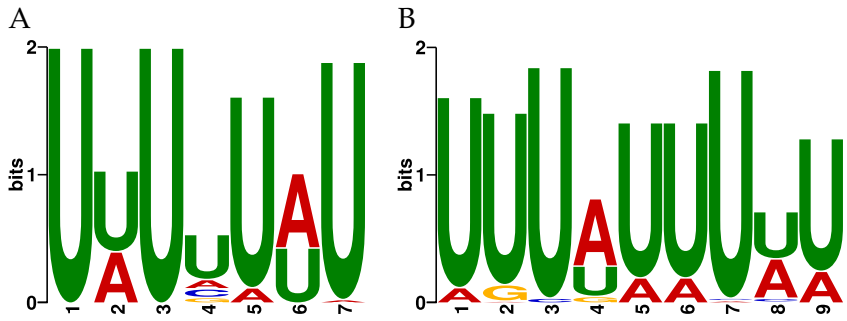


Figure 33: **Overlapping binding sites between TTP and HuR in 3'UTRs.** MEME logos of **(A)** 7-mer and **(A)** 9-mer over-represented motifs

Figures 33A and 33B show the MEME sequence logo of 79 hand curated binding sites where TTP and HuR overlap. In contrast to `[UA]UAU[UC]UAU[AU]` TTP and `[AU]UUU[UG]UUU[AU]` HuR only binding sites, one can see a merged motif, which can be described as `[AU]U[AU]U[UAGC]U[AU]U[AU]`.

So for most of the overlapping binding sites (66/79) a sort of consensus motif can be found, which is in general U-rich, as required for HuR binding, but does also contain As as required for TTP binding. The 9mer motif also fulfills these requirements, however with more variability.

SUMMARY Overlapping sites of TTP and HuR binding are neither in the class of typical TTP nor HuR motifs. They rather represent a merged version of both, rich in Us as required by both RBPs with some A content as required for TTP binding.

This indicates that overlapping sites represent a third class of motif, not favored over canonical ones by either RBP. The potential of the resulting motifs for their usefulness in prediction of co-regulatory binding sites remains to be investigated in detail.

2.2.8.3 *Analysis of sequence motifs in TTP and HuR non-overlapping binding sites*

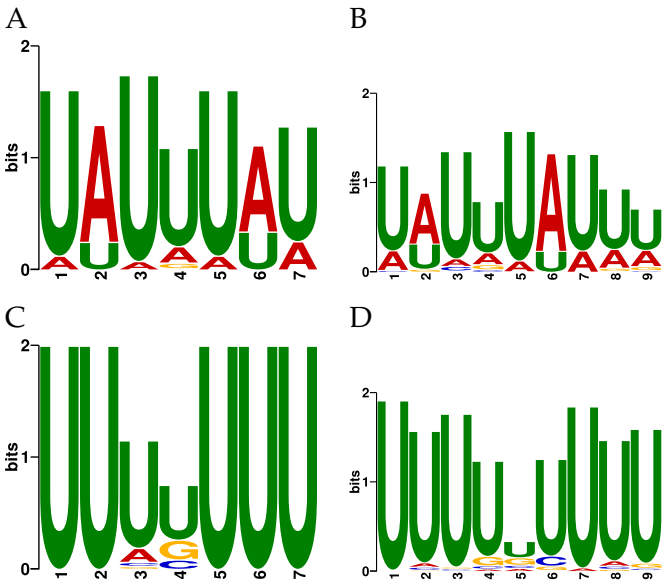


Figure 34: **Non HuR overlapping TTP binding sites in 3'UTRs.** MEME **(A)** TTP 7-mer, **(B)** TTP 9-mer,**(C)** HuR 7-mer and **(D)** HuR 9-mer non-overlapping 3'UTR binding sites.

Figures 34 **A and B** show the MEME sequence logo of TTP binding sites without HuR overlap. Figures 34 **C and D** show the MEME sequence logo of HuR binding sites without TTP overlap. The motifs are similar to the motifs derived from the total set of binding sites, which is not surprising as most binding sites do not overlap. However, compared to sequence motifs in overlapping sites the most obvious difference is the U-content, which is higher in overlapping sites, a probable prerequisite for HuR binding.

Furthermore, HuR PAR-iCLIP signals between WT and KO conditions was compared. If direct competition were the case, one would expect to find higher HuR signal in TTP-KO. However, when using non-normalized datasets, which is not ideal as changes in expression levels are expected, HuR signal in KO is lower than HuR signal in WT. Even after normalization to mRNA levels, no significant increase in HuR signal at these exact sites was detected, further strengthening a non-direct competition model between TTP and HuR.

SUMMARY Sequence motifs extracted from non-overlapping sites have high similarity to canonical binding motifs identified here. This can be explained by the fact that only few overlapping sites could be identified, but it still indicates that there are separate classes of motifs for individual and competitive binding.

2.2.9 *Human/Mouse conserved binding sites*

A strong indicator of function for some genomic region is conservation. Evolution depends on the survival of the fittest. So if a binding site for a protein that is conserved between organisms has a function in all of them, it is expected to be conserved. This section describes the conservation of 6 h PAR-iCLIP derived binding sites between mouse (mm9) and human (hg19). Binding site coordinates were lifted from mouse to human as follows.

2.2.9.1 *Coordinate lift-over mouse-human*

In order to identify homologous TTP binding sites in human and mouse, we extracted syntenic regions using the `liftover` tool of the `Kent source tree` [59].

Conservation of identified TTP and HuR binding sites of one representative replicate between mm9 and hg19 was investigated in a similar manner to the comparison with Mukherjee et al. [102] (sec. 2.2.10). Lift-over parameters were set to 95% and 10% sequence similarity for highly conserved and conserved subsets respectively.

Table 10 provides an overview of binding site conservation for TTP and HuR, where highly conserved means 95% and conserved means 10% sequence similarity between mouse and human.

Table 10: **Human/Mouse conserved binding sites**

	TTP	HuR
Highly conserved total	580	1445
Conserved total	759	1817
Highly conserved 3'UTR	424	1,246
Conserved 3'UTR	503	1,568
Highly conserved Introns	143	96
Conserved Introns	242	123

Although multiple intronic binding sites are conserved, 3'UTR sites are conserved more often. The TTP/HuR binding site in Irg1 intron 4 is not conserved among mouse and human, thus gives no indication for a possible sponge function in human.

Figure 35A provides an overview of target genes for TTP and HuR in our dataset that have orthologs in human and their overlap (*i. e.* genes targeted by TTP and HuR). Almost all TTP targets have orthologs (444

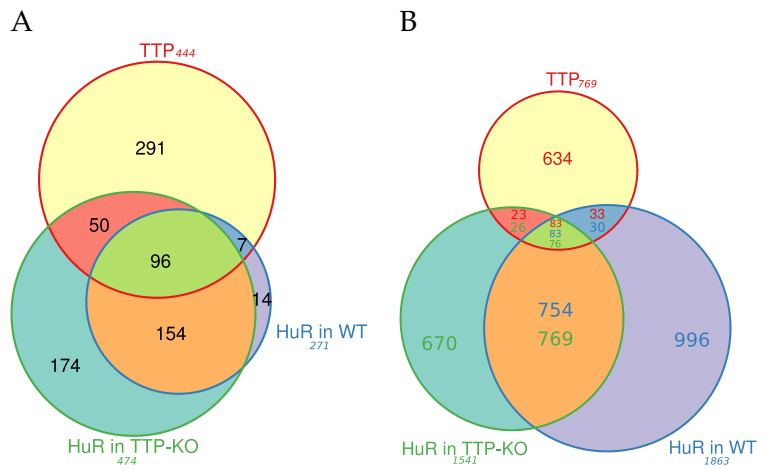


Figure 35: **TTP and HuR targets with human-mouse orthologs. (A)** This figure shows the TTP-HuR targeted conserved genes and overlaps **(B)** This figure shows conserved TTP and HuR bindingsites and overlaps. For convenience, color-codes show numbers for each ABP and experimental setting, since some binding sites overlap more than one BS in another set.

of 498), the same holds true for HuR in WT (271 of 303) and HuR in $TTP^{-/-}$ (474 of 532).

In figure 35B TTP binding sites which are conserved among mouse and human in any replicate and overlap conserved binding sites of HuR in TTP-WT and TTP-KO are shown. When comparing to fig. 30 one sees that relatively more HuR binding sites are conserved than TTP binding sites. Overlapping binding sites between all three or pairs of ABPs are well conserved, in general better than non-overlapping ones.

SUMMARY Binding sites in 3'UTRs are more often conserved between mouse and human than those in intronic regions. The ratio of overlapping binding sites between TTP and HuR is similar for binding sites lifted to human and those in mouse. The large number of TTP and HuR target genes conserved between mouse and human highlights the potential for portability of findings in this dataset to human.

2.2.10 *Comparison of our findings with Mukherjee et al. [102]*

2.2.10.1 *Comparison of experimental setup*

Targets from Sedlyarov et al. [120] show some overlap to a previous comprehensive study of TTP binding sites in a human HEK cell line by Mukherjee et al. [102], however, the majority of targets, and especially most of the top targets, were not identified there (for more details see section 2.2.10.3).

The two studies compared here differ greatly in their experimental setup.

(i) TTP is not expressed in human embryonic kidney cells (HEK) and had to be introduced via transfection with an expression vector. In contrast, primary mouse BMDMs are the natural defense system against bacterial infections and TTP is known to be expressed during the early inflammatory response, as mimicked by LPS stimulation.

(ii) Overexpression of TTP in HEK cells most likely alters the stocheometry of TTP-target interactions leading to artifacts. LPS induction of TTP expression in BMDMs ensures more natural conditions and thus allows us to detect and analyze native targets of TTP.

(iii) Previous studies (see e.g. Copeland et al. [31] or Osuchowski et al. [105] or Webb et al. [149]) underline similarities in inflammatory/endotoxin response in mice and humans, which emphasizes the impact of our findings on TTP function in BMDM inflammatory response and the high number of conserved target genes. 444 out of 498 TTP target genes identified in our study have annotated orthologs in human, indicating that these genes might be true targets of TTP in human as well.

(iv) TNF-α is one of the key players in inflammatory response, if not controlled it can cause systemic inflammatory response syndrome (cytokine storm) which is lethal to human as well as murine cell lines. While TNF-α is among, if not the top TTP targets, it can not be found among the target list of Mukherjee et al. [102].

Taken together, Sedlyarov et al. [120] identified TTP targets with human orthologs in a native, non-over-expressed system, allowing the findings of this study to be transferred from the model system mouse to human.

2.2.10.2 *Comparing Target Genes Sets*

Mukherjee et al. [102] identified 2,143 human genes to be targeted by TTP in HEK cells. 1,925 of these genes have orthologs in mouse (genome assembly NCBI m37 [mm9], ENSEMBL annotation 67), 942

Table 11: **Comparison of experimental conditions of TTP binding studies**

Condition	Sedlyarov et al. [120]	Mukherjee et al. [102]
cell system	mouse primary bone marrow derived macrophages (BMDB)	human embryonic kidney cells (HEK)
TTP induction	LPS stimulation	transfection of expression vector
TTP levels	native	overexpression
CLIP method	PAR-iCLIP	PAR-Clip

of which are 'high confidence' orthologs. Only 107 (48 high confidence) of these human genes with known mouse orthologs are represented in our set of TTP targets in BMDMs.

444 out of 498 TTP target genes identified in our study have orthologs in human. As for HuR, 271 of 303 target genes in wild type BMDMs and 474 of 532 HuR targets in TTP-KO have human orthologs according to ENSEMBL.

Sedlyarov et al. [120] provides a list of 500 genes which represent main TTP targets in inflammatory response. The majority of these genes have orthologs in human (see Figure 35 in section 2.2.9), underlining the importance and portability of mouse models in order to study ABP related human disease mechanisms.

2.2.10.3 *Comparing TTP Signals*

Coordinates of TTP peaks in human (hg19) identified by Mukherjee et al. [102] were 'lifted' to mouse coordinates (mm9) with -minMatch=0.05.

3,316 binding sites out of 4,625 total binding sites of the Mukherjee et al. [102] dataset could be assigned to homologous loci in mouse. 2,731 "lifted" binding sites are within annotated mouse genes.

1,925 human target genes have orthologs in mouse, for 1,896 of those the binding sites could be lifted as well. This is because in some cases the actual TTP binding site is not sufficiently conserved between mouse and human.

Only 32 binding sites (from 29 human genes) overlap directly with binding sites from the Sedlyarov et al. [120] dataset, *e.g.* Zfp36, but not TNF. Zpf36 and PFKFB3 are two of the top 10 targets (12 of top50) that have been identified in both studies.

248 out of 1,598 binding sites in the Sedlyarov et al. [120] dataset are in ±50nt distance to a total of 3,316 sites Mukherjee et al. [102] (2,731 in annotated genes).

Table 12: **TTP target genes with binding sites identified in both studies.** After liftover of coordinates, we found TTP binding sites to be conserved between human and mouse in the following 35 genes

Gene name						
6330409N04Rik	Mdm2	Actb	Med13	Adrbk1	Mxd1	Aff1
Nfkbia	Anxa5	Papd7	App	Pfkfb3	Arl8a	Plaur
Atf3	Ppp1r15a	B2m	Ppp3r1	Brd4	Rabep1	Cebpb
Sdc4	Cep170	Tet2	Cxcl1	Tnfaip3	Dennd4b	Zeb2
Ets2	Zfp36	Etv3	H3f3b	Hivep2	Ier3	Mcl1

These 248 binding sites come from 35 genes out of 498 from our dataset (tab. 12).

SUMMARY For the comparison of Sedlyarov et al. [120] PAR-iCLIP data to the dataset of Mukherjee et al. [102], several aspects have to be kept in mind. Although lift over works well in general, it lacks precision for small intervals (i.e. binding sites), therefore a range of ±50nt around binding sites was considered a reasonable range to compare these intervals between organisms.

Furthermore, the investigated cell types differ vastly. HEK cells without native TTP expression are expected to show different binding behaviour then the native LPS induced BMDM approach. However, a small set of genes was identified in both studies, containing *e. g.* TTP itself (Zfp36), highlighting auto-regulation as important mechanism in TTP controlled mRNA stability regulation in human and mouse.

2.2.11 *Structure vs. Sequence analysis*

2.2.11.1 *ARE analysis*

Potential ABP binding sites as defined by AREsite [43], are investigated for their "activity" in TTP/HuR regulation of mRNA stability. Therefore positions of annotated consensus motifs are compared with binding sites identified by PAR-iCLIP experiments, to generate "positive" and "negative" sets for further analysis. Conservation of these motifs between human and mouse is shown in tab. 13.

Genomic coordinates of all sites listed in the ARE database that correspond to the consensus motifs for HuR (*TTTKTTT*) and TTP (*WATTTAW*) identified in this PAR-iCLIP analysis were extracted. The following subsets were created: (i) sites conserved between mouse and human (ii) sites residing in target genes expressed in BMDMs (iii) sites that overlap with binding sites identified in this study.

To determine overlaps between features, `BEDtools v2.17` [108] was used and only overlaps on the same strand (-s) with minimum overlap of 1nt were considered.

13,862 of 17,411 TTP motifs in the AREsite1 database reside in transcripts that are expressed in BMDMs. Almost all of these motifs (12,290) are conserved between human and mouse, but only a small fraction (249) is indeed used by TTP. HuR follows a similar trend (see tab. 13).

Table 13 summarizes numbers for TTP and HuR ARE core motifs bound and unbound as well as conserved and not conserved between human and mouse.

Furthermore, AREs are divide into motifs in transcripts that are (i) expressed in BMDMs used in this PAR-iCLIP experiments and in those that (ii) do and do not overlap with binding sites identified in this study.

While the vast majority of ARE motifs (*WATTTAW* as well as *TTTKTTT*) can be found in transcripts expressed in our cell-lines also have conserved sites in human, only a small amount of those show overlaps with identified binding sites. The ratio between conserved and unconserved motifs is highest for those not in overlap with PAR-iCLIP signal of any ABP.

Table 13: ARE motifs used and unused by TTP/HuR and (non-) conserved between human and mouse

AREs	TTTKTTT (HuR)			WATTTAW (TTP)		
total	47,887			17,411		
conserved	33,187			15,034		
unconserved	14,700			2,377		
expressed	40,419			13,862		
expressed & conserved	28,443			12,290		
expressed & unconserved	11,976			1,572		
AREs in overlap with Par-iCLIP binding site	HuR KO	HuR WT	TTP	HuR KO	HuR WT	TTP
total	2,232	1,574	112	76	84	249
conserved	1,319	904	74	70	77	217
unconserved	913	670	38	6	7	32
expressed	2,232	1,574	112	76	84	249
expressed & conserved	1,319	904	74	70	77	217
expressed & unconserved	913	670	38	6	7	32
AREs NOT in overlap with Par-iCLIP binding site	HuR KO	HuR WT	TTP	HuR KO	HuR WT	TTP
total	45,658	46,313	47,779	17,335	17,327	17,164
conserved	31,869	32,283	33,113	14,964	14,957	14,819
unconserved	13,789	14,030	14,666	2,371	2,370	2,345
expressed	38,190	38,845	40,311	13,786	13,778	13,615
expressed & conserved	27,125	27,539	28,369	12,220	12,213	12,075
expressed & unconserved	11,065	11,306	11,942	1,566	1,565	1,540

However, the ratio of bound motifs is small compared to unbound ones, and conserved bound motifs are always more than unconserved ones.

The "positive" and "negative" sets derived from this ARE analysis where used to compare sequence and structure features that lead to binding by TTP and/or HuR.

2.2.11.2 *Structure analysis*

RNAplfold [14] can calculate the energy needed to open potential secondary structures on a stretch of RNA, which allowed us to compare differences in the structuredness of binding site embedding regions and non-bound regions. For this analysis flanking regions of 28nt were added to the positive and negative sets described in section 2.2.11.1.

Then the opening energies for said sequences were calculated and the latter binned in 7nt steps, which corresponds to the length of the extracted consensus motifs and allows to compare AU-content of both sets with their structuredness.

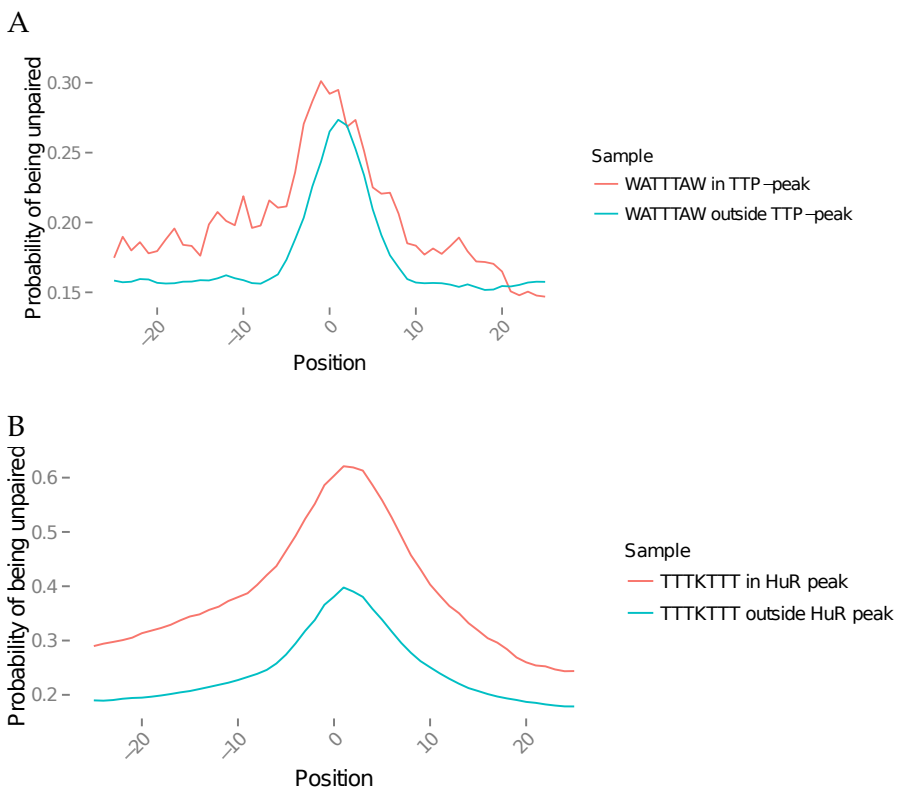


Figure 36: **Structural context of TTP and HuR binding sites. (A)** Accessibility of regions embedding ARE motifs with and without overlap with TTP binding sites, **(B)** Accessibility of regions embedding ARE motifs with and without overlap with HuR binding sites

Figures 36A and 36B show the mean probability of being unpaired for a ±28nt context around ARE core motifs (WAUUUAW for TTP, where W can either be A or U and UUUKUUU for HuR where K can either be U or G) in and outside of peak regions of TTP and HuR. In both cases ARE motifs within binding sites show a higher probability of being unpaired than motifs without peak signal.

Figures 37A and 37B show the RNAplfold derived opening energies for potential secondary structures ±28 nt around peak regions in bins

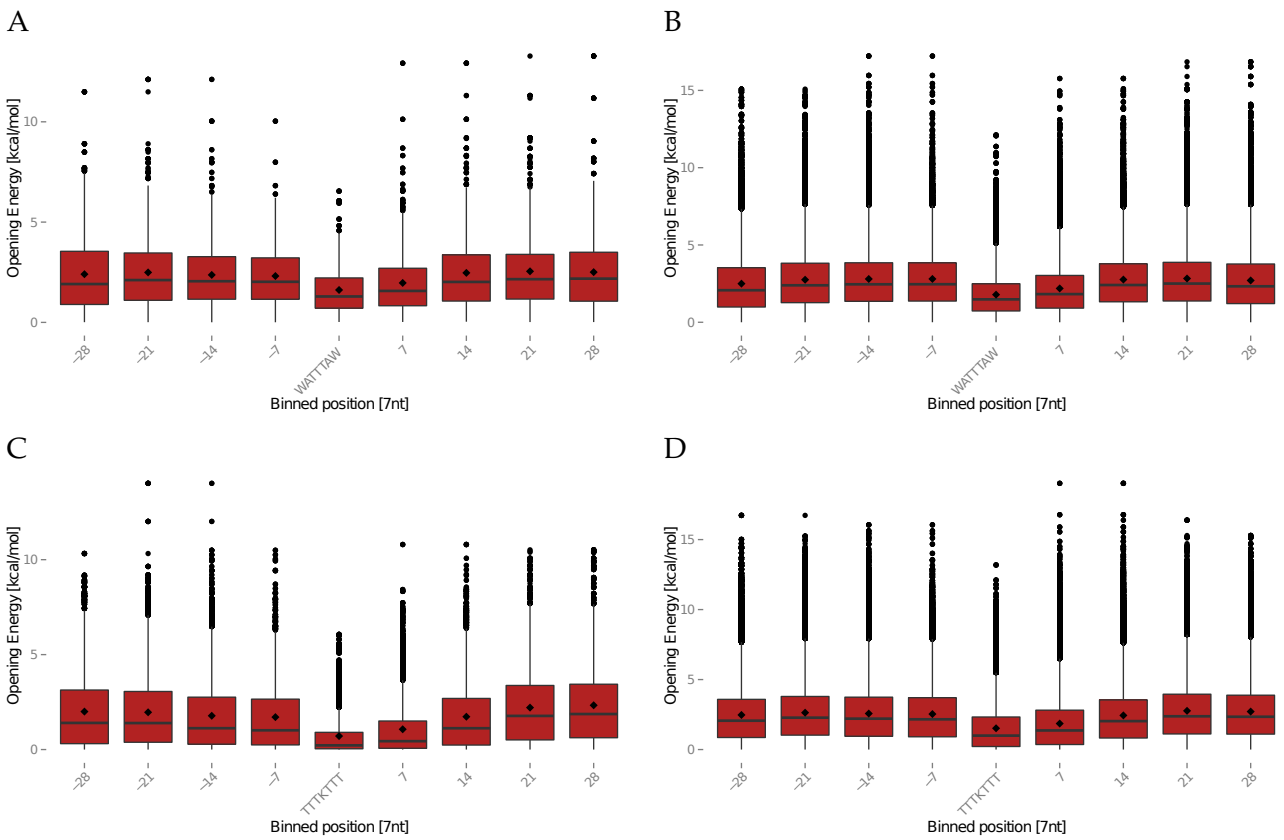


Figure 37: **Opening energy for secondary structures in ARE motif embedding regions in- and outside of TTP and HuR binding sites. (A)** ARE overlapping TTP binding sites, **(B)** ARE outside TTP binding sites, **(C)** ARE overlapping HuR binding sites, **(D)** ARE outside HuR binding sites; - ... median • ... mean

of 7nts. Opening energies for motifs in TTP/HuR peaks are lower than for motifs without peak overlap, which means that the former are less likely to be found in stable secondary structures than the latter. As opening energy and probability of being unpaired are not independent terms, this finding is not unexpected, but visualization as box plot allows to get a feeling for mean and median opening energy, which are both lower for bound motifs (Note the differing scales on y-axis).

2.2.11.3 *Sequence analysis*

A+U-content analysis of flanking regions around ARE motifs in both sets is shown in figure 38. The region close to the actual WATTTAW motif has a higher A+U-content in TTP/HuR binding sites (median > 80%) compared to motifs outside, however the A+U content remains in general very high (median ~ 70%).

Regions more distant to the central motif show comparable A+U content in bound and unbound regions for TTP as well as HuR.

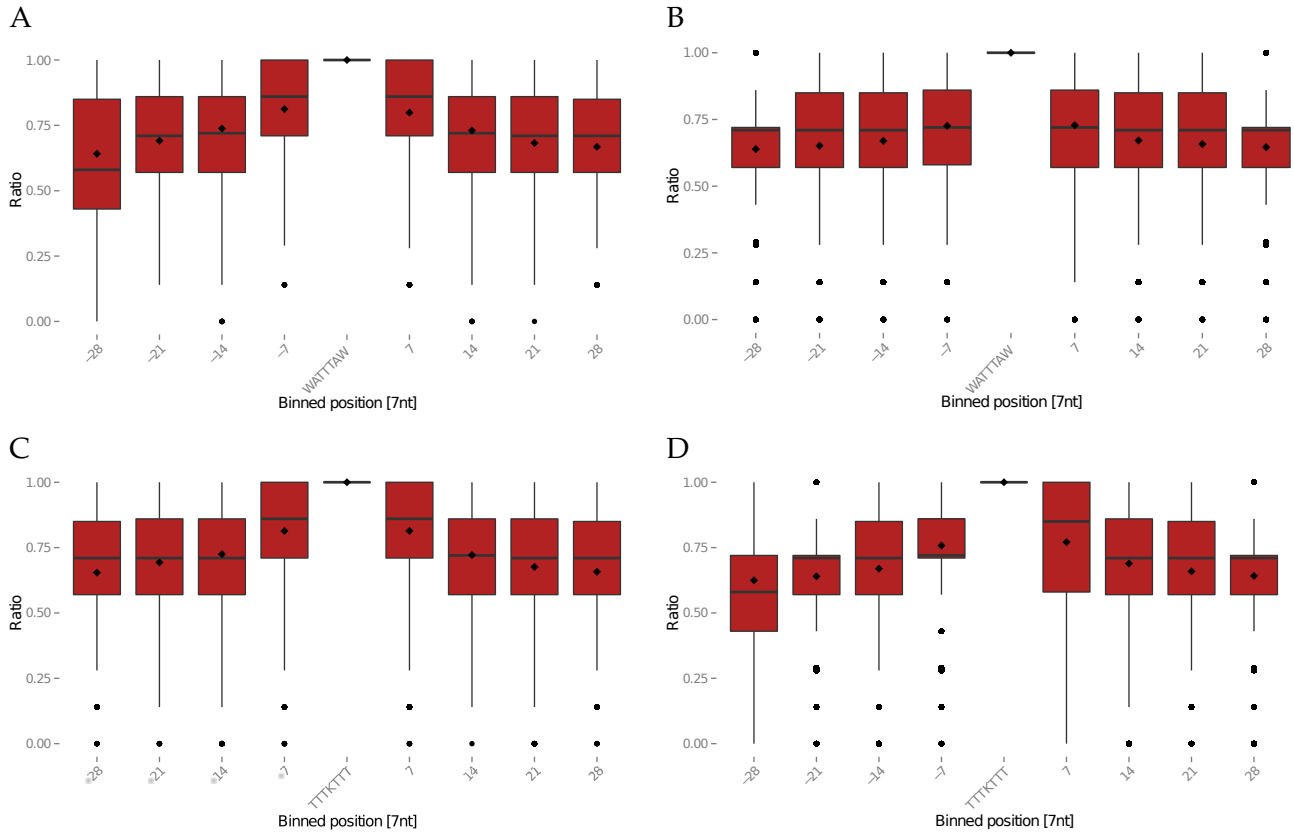


Figure 38: **AU content of regions embedding ARE motifs in- and outside of TTP and HuR binding sites. (A)** A+U content overlapping TTP binding sites, **(B)** A+U content outside TTP binding sites, **(C)** A+U content overlapping HuR binding sites, **(D)** A+U content outside HuR binding sites; - ... median • ... mean

2.2.11.4 *Comparison of sequence and structure as descriptors for active binding sites*

Table 14 compares Wilcoxon ranked sum test derived p-values describing the significance in differences on sequence (A+U-content) and structure (opening energies) level between flanking regions around ARE motifs with and without overlapping TTP/HuR binding sites.

Differences are in all cases significant, however, for TTP differences in A+U-content and opening energy are comparable, while for HuR differences in opening energy are more significant than in A+U-content.

Table 14: **Wilcoxon rank sum test**, of A+U content and opening energy of binding site flanking regions

WATTTAW in TTP

	A+U content	**Opening energy**
Flanking region [nt]	**p-value**	**p-value**
15	7.1603e-11	6.1496e-12
20	2.2114e-12	8.8905e-11
25	1.9143e-12	2.4849e-11
30	3.2151e-11	4.5697e-10
35	4.9798e-08	6.4108e-09

TTTKTTT in HuR

	A+U content	**Opening Energy**
Flanking region [nt]	**p-value**	**p-value**
15	9.1934e-41	9.6389e-151
20	4.3190e-50	1.5994e-145
25	6.1551e-56	2.7504e-141
30	3.1049e-57	8.8364e-135
35	1.8690e-56	2.2810e-117

To further evaluate if structuredness can be used as a descriptor for bound and unbound ARE motifs, we performed a Receiver-Operator-Characteristic (ROC) analysis comparing the A+U-content of regions embedding ARE motifs with and without overlap of binding sites of TTP and HuR with the energy required to open potential RNA secondary structures (fig. 39), similar to section 2.1.5, in a region ± 15 nt and ± 25 nt around motifs.

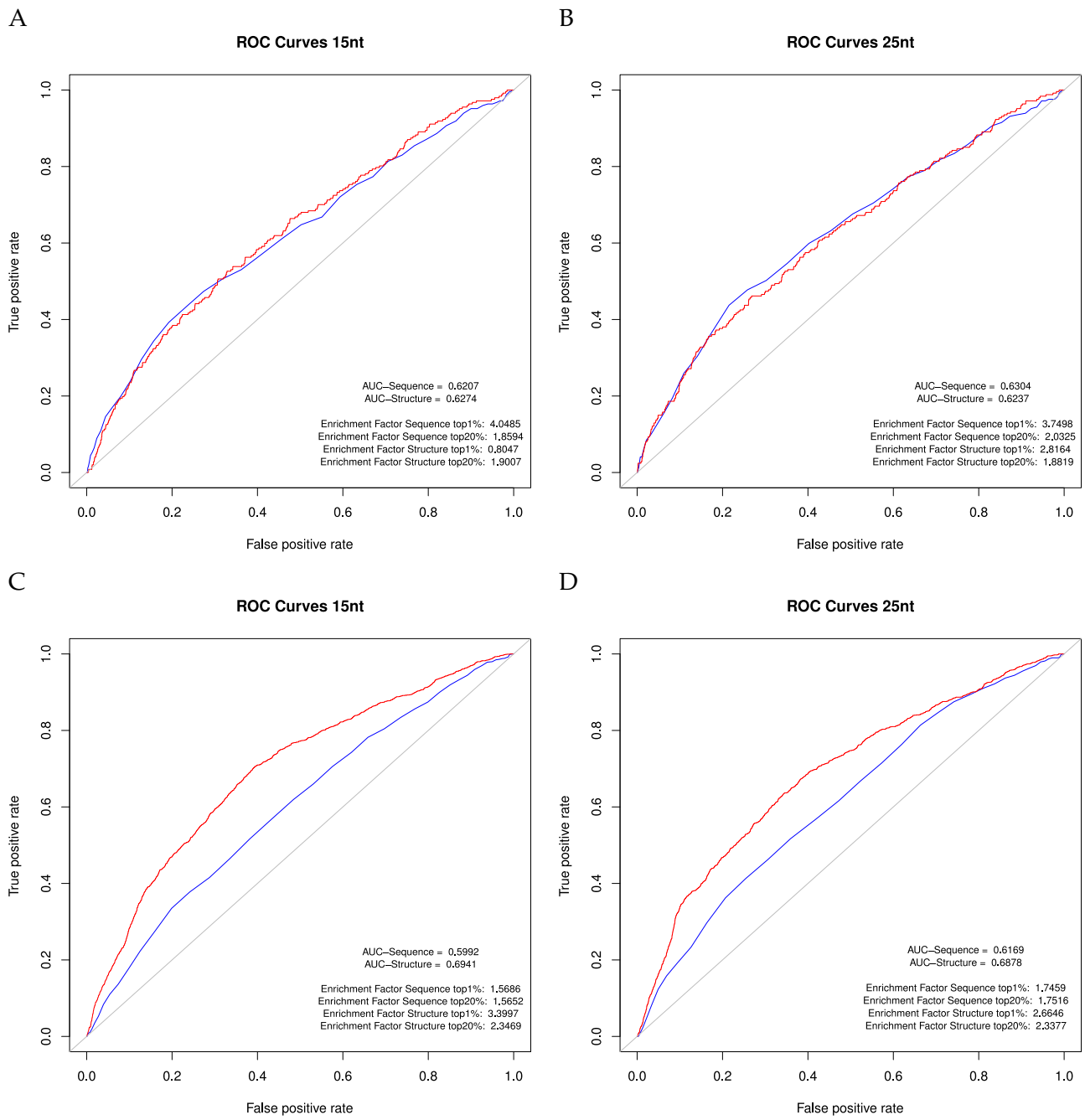


Figure 39: **Descriptor analysis of ARE flanking regions with and without overlap of TTP/HuR binding sites** **(A)** ROC-curve of ±15nt around TTP binding sites, **(B)** , ROC-curve of ±25nt around TTP binding sites, **(C)** ROC-curve of ±15nt around HuR binding sites, **(D)** ROC-curve of ±25nt around HuR binding sites

For TTP Sequence and Structure derived AUCs (Area under the ROC-Curve) are almost similar, while for HuR structure derived AUCs are higher. For TTP AU-richness of bound motifs is already higher than for unbound ones, so that AU-content as well as opening energy are equally useful for distinguishing bound from unbound ARE mo-

tifs.

For HuR however, opening energy seems to be a better descriptor of bound and unbound sites than AU-content alone, which is in general very high, also for unbound motifs, see section 2.2.11.3. For both proteins, the descriptive power of structural context ±15nt around binding sites is higher than for the ±25nt context. This indicates that TTP as well as HuR binding depends on context in close proximity to actual binding sites.

SUMMARY Analysis of accessibility of TTP/HuR un-/bound sites reveals that bound sites are in general embedded in a more accessible environment that unbound sites. Also the AU-content is higher in the surroundings of bound sites.

This indicates that active target sites for both RBPs need to be accessible and AU-rich in a broader context than just the actual binding site, rendering the existence of secondary structure prerequisites besides single-strandedness unlikely.

2.2.11.5 *Linear discriminator analysis*

We now know that accessibility and AU-content as descriptors can be used to distinguish bound from unbound motifs in our normalized and filtered PAR-iCLIP dataset. What remains to be investigated is if these descriptors can be used to train a discriminator that can distinguish bound from unbound sites in a larger context.

To that purpose a linear discriminator was trained with the R MASS library [139] using the PAR-iCLIP dataset descriptors AU-content and opening energy for training. The dataset was split 9:1, where 90% of the positive and negative set were used for training and the remaining 10% for testing of the linear discriminators. Once trained, this discriminators were also tested against the AREsite2 derived positive and negative sets described in section 2.1.3.

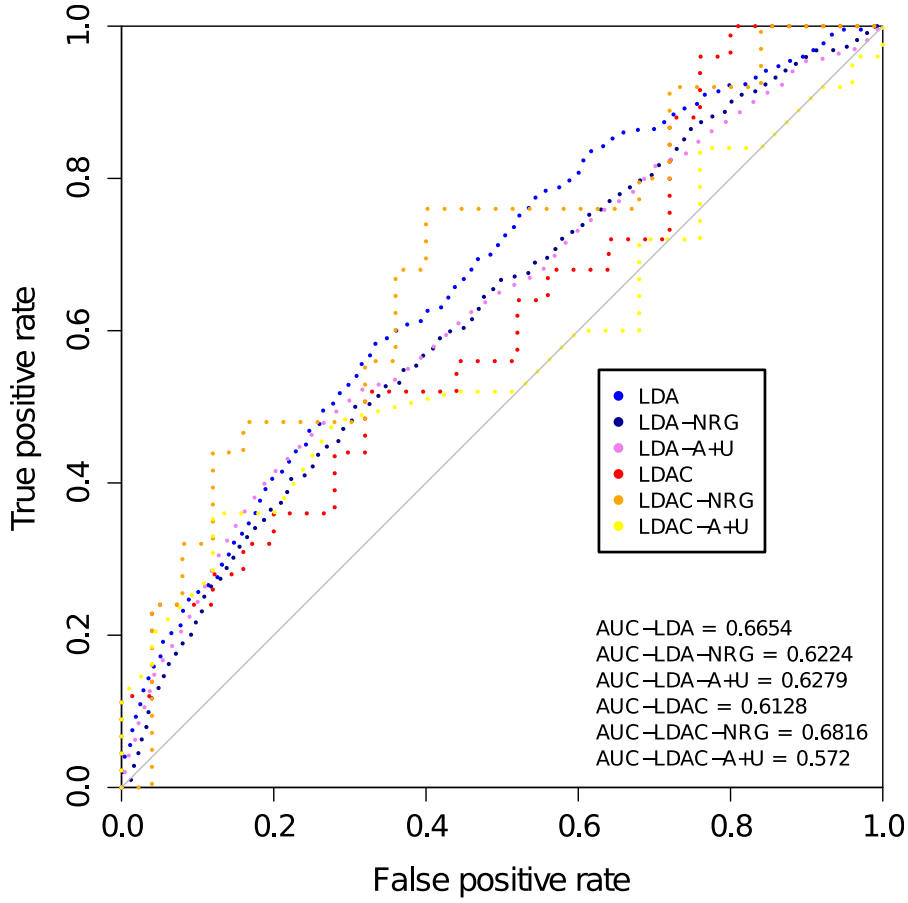


Figure 40: **LDA analysis for TTP** Linear discriminator analysis of sequence and/or structure for TTP binding sites. LDA was trained with the PAR-iCLIP derived dataset and tested on AREsite2 derived motifs with CLIP-Seq signal overlap. The plot shows a comparison of predictive power. LDA, LDA-NRG and LDA-A+U and corresponding AUCs show descriptive power for training with 90% of the training set and testing on the remaining 10%, while LDAC, LDAC-NRG and LDAC-A+U show predictive power when tested on the AREsite2 derived dataset. NRG stands for accessibility in terms of opening energy, A+U for A and U sequence content respectively, no addition means a combination of both descriptors was tested.

Figures 40 and 41 show ROC curves for both tests for TTP and HuR respectively. The blue to violet curves stand for the 9:1 test and the red to yellow curves for the test with the AREsite2 dataset. For each test the discriminators are either opening energy and A+U content or one of the two. For TTP, AUCs for the 9:1 test indicate medium predictive power for all three discriminators, with the highest AUC for a combination of both, opening energy and A+U-content. When testing on the AREsite2 dataset, opening energy outperforms both, A+U-content and the combination of both descriptors.

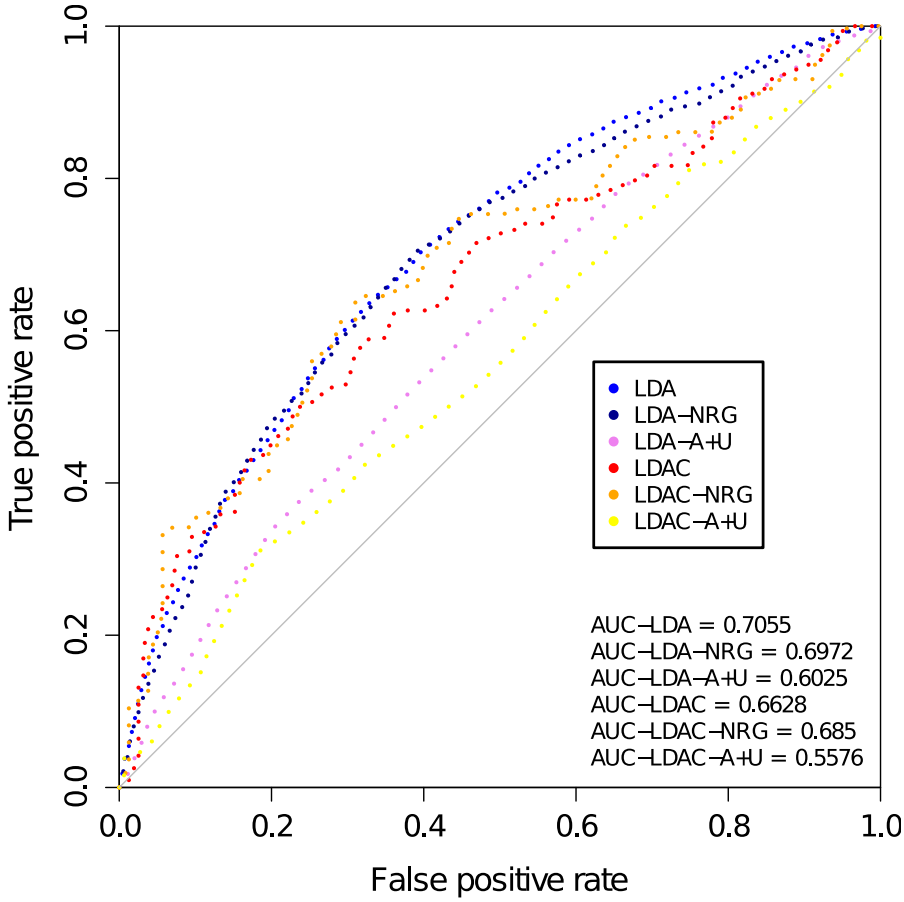


Figure 41: **LDA analysis for HuR** Linear discriminator analysis of sequence and structure for HuR binding sites. LDA was trained with the PAR-iCLIP derived dataset and tested on AREsite2 derived motifs with CLIP-Seq signal overlap. The plot shows a comparison of predictive power. LDA, LDA-NRG and LDA-A+U and corresponding AUCs show descriptive power for training with 90% of the training set and testing on the remaining 10%, while LDAC, LDAC-NRG and LDAC-A+U show predictive power when tested on the AREsite2 derived dataset. NRG stands for accessibility in terms of opening energy, A+U for A and U sequence content respectively, no addition means a combination of both descriptors was tested.

For HuR opening energy has the highest AUC for both test sets, and outperforms A+U-content in predictive power. Similar to the descriptor analysis in section 2.2.11.4, the LDA analysis shows that accessibility of motifs is a good discriminator between HuR bound and unbound motifs.

SUMMARY Linear discriminator analysis (LDA) for the ABPs TTP and HuR, shows that accessibility and AU-content can be used to successfully discriminate between bound and unbound motifs. In case of HuR, accessibility of a motif is even a better discriminator than A+U-content.

Although LDA is a rather simple way of training for a discriminator, it already shows promising results and highlights the value of secondary structure predictions for machine learning approaches for protein-RNA interaction studies.

2.2.12 *miRNAs and TTP/HuR*

Lu et al. [86] recently published an Ago-CLIP-Seq dataset in mouse BMDMs. This experiment aims at identifying miRNA interaction sites in the same biological context than our PAR-iCLIP experiment. As cross-regulation of ABPs and miRNAs could be shown in this study, we extracted miRNA binding sites from the Lu et al. [86] data and intersected them with our PAR-iCLIP peak regions.

Figure 42 shows a Venn diagram of binding sites in $\pm 50nt$ distance between these datasets. Only a minority of binding sites overlap, interestingly most of them with HuR in the TTP-KO sample.

All overlapping sites are excellent candidates for further experiments, focusing on the extend of miRNA RBP cross-regulation in detail.

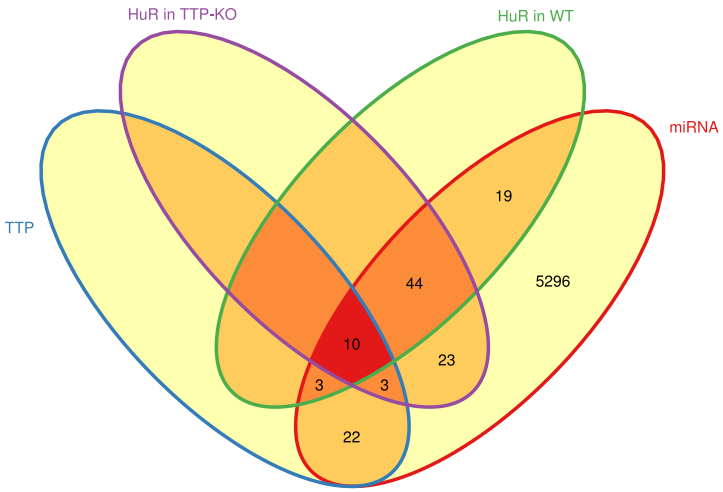


Figure 42: **Venn diagram** presenting the number of binding sites in $\pm 50nt$ distance between TTP/HuR from PAR-iCLIP and miRNAs extracted from Lu et al. [86].

SUMMARY This preliminary analysis shows already that miRNA binding sites have overlaps with both RBPs, thus co-regulatory function as described *e. g.* by Lu et al. [86] is a factor to be considered for further analysis.

Such analysis could shed light on the mechanics behind RNA half-life regulation, especially in the context of the auto-regulatory function of both RBPs.

2.2.13 *Cooperative vs. competitive binding*

So far we have shown that direct overlaps between TTP and HuR are rarely found in our dataset. This indicates that direct competition of these two ABPs is only relevant for a minority of targeted genes. To address indirect competition, we focused on genes with binding site of both proteins, but without direct overlap.

Competition does not necessarily require both proteins to be located at the exact same or overlapping stretch of nucleotides, but can potentially occur via structural constraints that are the result of binding of one competitor. Such interaction introduces or titrates energy to/from the system, which can lead to refolding, potentially un/blocking previously un/paired regions.

Lin and Bundschuh [82] present a model for the calculation of cooperative binding free energy, where the free energy of a RNA molecule bound by two interaction partners is derived from the sum of of the energy of both partners interacting separately minus the end state and ground state.

Equation 3 describes the four states which are used to calculate $\Delta\Delta G$, the cooperative binding free energy. A negative $\Delta\Delta G$ indicates antagonistic binding effects, a positive $\Delta\Delta G$ indicates cooperative effects. The new constraint folding option in the ViennaRNA package 2.0 [83], using a pair of binding sites as constraints, allows to calculate all terms required for such an investigation.

$$\begin{aligned}
\Delta G_{0\to1} &= \Delta G^1 - \Delta G^0 - RT \times \ln\left(\frac{c_1}{K_{D,1}}\right) \\
\Delta G_{0\to2} &= \Delta G^2 - \Delta G^0 - RT \times \ln\left(\frac{c_2}{K_{D,2}}\right) \\
\Delta G_{1\to12} &= \Delta G^{12} - \Delta G^1 - RT \times \ln\left(\frac{c_2}{K_{D,2}}\right) \\
\Delta G_{2\to12} &= \Delta G^{12} - \Delta G^2 - RT \times \ln\left(\frac{c_1}{K_{D,1}}\right)
\end{aligned} \tag{3}$$

$$\Delta\Delta G = \Delta G_{0\to1} - \Delta G_{2\to12} = G^1 + G^2 - G^{12} - G^0$$

Non-overlapping pairs of binding sites in 3'UTRs with minimal distance of 10nt between and within experiments were extracted from our dataset. Minimum free energy of 3'UTR sequences with/without these binding sites as constraints were computed with RNAfold [48]. Figure 43 shows histograms of $\Delta\Delta G$ computed from all binding site pairs from our PAR-iCLIP dataset and the Ago-CLIP-Seq data from Lu et al. [86], between and within samples on the same 3'UTR.

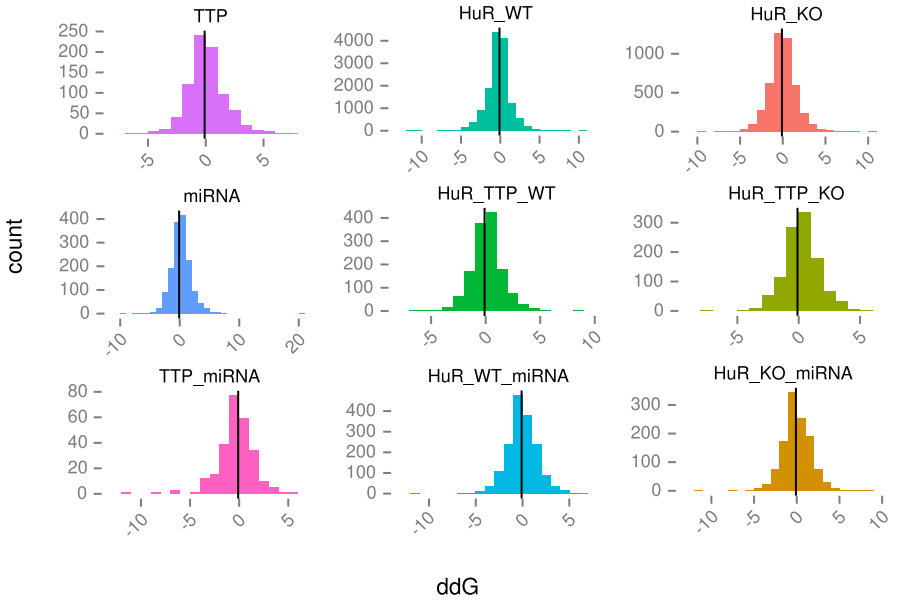


Figure 43: ΔΔG **for all pairs of 3'UTR binding sites** between and within samples of TTP and HuR from PAR-iCLIP and miRNA from Lu et al. [86].

In general only small to no ΔΔG for pairwise binding could be seen. However, the here investigated ABPs bind single-stranded RNA and effects on already unpaired regions were expected to be small. Furthermore, we find some of the top TTP targets to be effected in a cooperative manner by other molecules of TTP as well as HuR or miRNA binding the same target. The TTP transcript itself is one of the RNAs where miRNA-27 binding has effect on HuR and TTP binding site accessibility.

Lu et al. [86] propose a regulatory feedback loop where HuR binding stabilizes the TTP transcript, while miRNA-27 binding has destabilizing effects. Furthermore, they show that in the absence of HuR, more miRNA-27 is bound to the TTP transcript, inferring that miRNA-27 and HuR compete for binding. This was shown for binding sites with direct overlap. Our data suggest an even more complex picture.

While we compute an overall negative (antagonistic) effect of miRNA-27 binding on the accessibility of TTP binding sites on its own mRNA, we also see a positive effect on HuR binding sites on the same target.

This suggests, that miRNA-27 contributes positively to the expression of TTP, once by indirectly displacing TTP from its binding site, which would otherwise have a negative effect on TTP expression, while in parallel rendering a HuR binding site more accessible. However, one has to be aware that this preliminary analysis is not capable of solving the complex interactome between TTP/HuR and miRNAs, but note that this kind of investigation should be subject to future studies.

SUMMARY While this analysis was conducted with all non-overlapping pairs of binding sites sharing the same 3'UTR, pre-selection of high potential sites (*e. g.* by high PeakScore) should be included. Larger peak regions, which could potentially contain more then one molecule of TTP/HuR should also be split into smaller regions, as especially these sites have a high potential of co-regulation.

Summing up, this preliminary investigation shows some promising results, which have a potential for further studies with more complex models of interaction and constraints.

2.2.14 *TTP directly influences mRNA half-life*

A key function of TTP is initiation of degradation of target mRNAs. To test whether direct correlation between TTP binding and mRNA decay can be found, Pearson correlation of normalized gene score for TTP 3 h and 6 h targets with mRNA decay experiments published in Sedlyarov et al. [120] was investigated. Normalized gene scores were used to cope with the influence of RNA expression on CLIP-Seq signal.

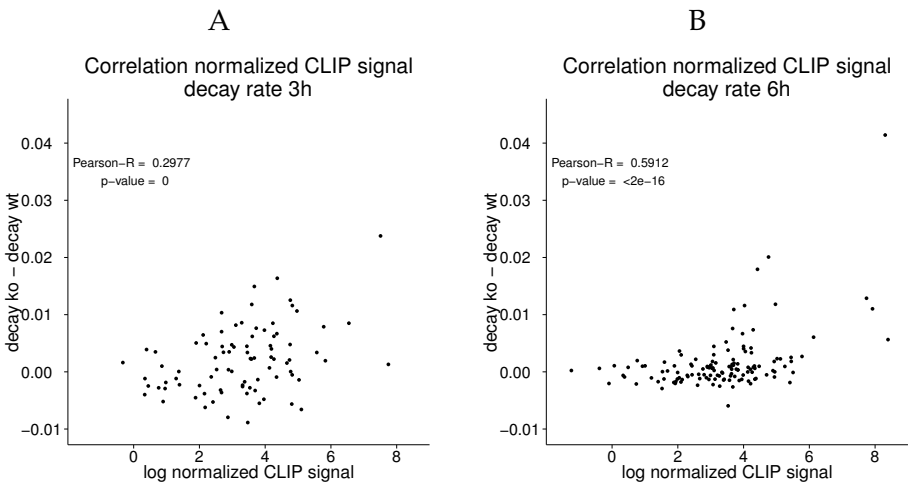


Figure 44: **Correlation analysis of mRNA decay and PAR-iCLIP signal** A) compares mRNA decay rate 3 h after LPS induction with RNA-Seq normalized PAR-iCLIP signal for TTP, B) compares mRNA decay rate 6 h after LPS induction with RNA-Seq normalized PAR-iCLIP signal for TTP.

Figure 44 shows dot-plots comparing the normalized CLIP-Seq signal to the mRNA decay rate in WT and TTP-KO cells. As the presence of TTP is thought to decrease mRNA stability upon interaction, the difference between decay rate in KO cells and decay rate in WT cells was calculated. Most of the genes show only marginal difference between

both conditions, resulting in the majority of datapoints between -.01 and .01 on the y-axis.

However, some genes with higher CLIP-Seq signal show also increased decay rate. At 3 h after LPS induction, where TTP already binds target genes, correlation with decay rate changes is only weak (Pearson-R = 0.2977; 95%CI: lower = 0.0941, upper = 0.4774). 6 h after LPS induction we see a significantly (p-val: .004, z=2.65) higher correlation (Pearson-R = 0.5912; 95%CI: lower = 0.4660, upper = 0.6932) between CLIP-Seq signal and decay.

This indicates a direct influence of TTP bound to a target mRNA and decay of the latter for the 6 h dataset which is in perfect agreement with the biological model of TTP resolving the inflammatory response.

2.2.15 *GO Analysis of TTP and HuR target genes*

An indicator for the molecular function of genes are associated GO-terms. We used the generated lists of target genes of TTP and HuR for GO enrichment analysis to investigate gene function differences between TTP 3 h and 6 h and HuR in WT and TTP-KO with the tools `DAVID` [51] and `PantherDB` [100] and the `R` package `TopGO` [2].

2.2.15.1 *GO-enrichment for TTP binding sites in UTR, intron and overall*

Target genes were divided in three sets, those containing all genes, those with binding sites in 3'UTR and those with binding sites in intronic regions only and investigated `TopGO` GO-term enrichment for each subclass. As expected, we see a clear bias towards inflammatory related GO-terms in over-all and 3'UTR bound TTP target genes. Those with exclusive intronic binding sites lack these GO-terms.

The top enriched GO-term annotation clusters of `DAVID` derived GO-terms for the 3 h and 6 h TTP dataset comparing all genes and those with only 3'UTR binding sites can be found in the supplements (see tables 33, 34, 35 and 36). In both conditions (3 h and 6 h), GO-term related to inflammation and cytokine activity are ranked higher in the 3'UTR datasets, which indicates that only 3'UTR binding of TTP plays a direct role for inflammatory response.

2.2.15.2 *GO term enrichment comparison between TTP datasets*

To further investigate the role of TTP binding during inflammatory response, we compared the number of genes with specific GO-terms and log fraction (observed vs expected) of GO terms for TTP targets 3 h and 6 h after LPS induction with `PantherDB`.

Table 15: **GO-term enrichment for TTP target genes, and genes containing exclusively 3′UTR or intronic peak regions.** Analysis was conducted with TopGO and the set of all expressed transcripts as background.

GO-ID	GO-Term	Annotated	Significant	Expected	Rank
		TTP			
GO:0005125	cytokine activity	26	15	6.78	1
GO:0008009	chemokine activity	10	8	2.61	2
GO:0042379	chemokine receptor binding	11	8	2.87	3
GO:0005126	cytokine receptor binding	32	16	8.34	4
GO:0001664	G-protein coupled receptor binding	17	10	4.43	5
		TTP 3′UTR			
GO:0008009	chemokine activity	10	8	2.22	1
GO:0005125	cytokine activity	26	14	5.77	2
GO:0005126	cytokine receptor binding	27	14	5.99	3
GO:0042379	chemokine receptor binding	11	8	2.44	4
GO:0042802	identical protein binding	25	13	5.55	5
		TTP Introns			
GO:0003676	nucleic acid binding	74	58	49.52	1
GO:0005524	ATP binding	59	47	39.48	2
GO:0030554	adenyl nucleotide binding	59	47	39.48	3
GO:0032559	adenyl ribonucleotide binding	59	47	39.48	4
GO:0001883	purine nucleoside binding	74	57	49.52	5

Figure 45 shows results as retrieved from PantherDB for both gene sets. In the 3 h category, more genes with corresponding GO-terms related to immune response, response to stress or cell communication are annotated, while in the 6 h category terms like cell proliferation, immune system processes and cellular defense response are more present (fig. 45A). When directly comparing the log fraction of observed vs. expected GO terms between both conditions (see fig. 45B), this trend of the 3 h towards early and 6 h towards late immune response related processes stays the same.

This supports the model of TTPs role during inflammatory response, which in the first 3 h of infection starts to control stress and stimulus induced genes, while it later on primarily targets genes relevant for inflammatory response and proliferation, both necessary for successful ceasing of inflammatory response and preventing the immune system from over-reaction.

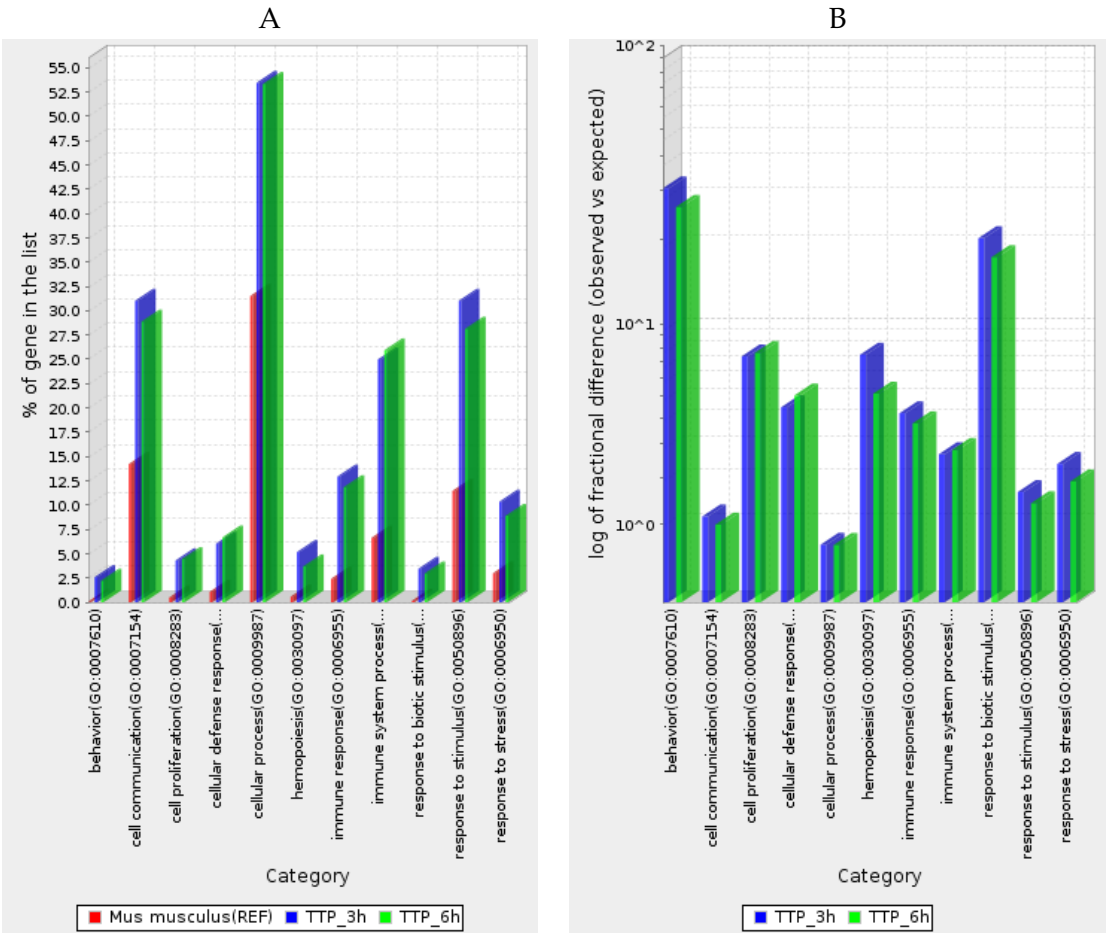


Figure 45: **PantherDB GO term comparison** A) Shows the %-age of PAR-iCLIP derived target genes that are enriched for a certain GO-term in background (mus musculus) and TTP 3 h and 6 ,h after LPS induction, B) shows the difference between observed and expected GO-term enrichment for PAR-iCLIP derived target genes for TTP 3 h and 6 h after LPS induction.

2.2.15.3 *GO-analysis for orthologous TTP targets in human and mouse*

To further investigate if the findings in our dataset can be transferred to human, GO-term related differences and commonalities between TTP target genes identified in this study and their orthologs annotated in human were investigated.

Mouse/Human conserved genes show no obvious GO-term related difference in comparison to the total target list for TTP (444 of 498 genes are conserved). Table 16 shows a brief summary of cell-fate related GO terms enriched in human and mouse with `DAVID`, the full list of the top three clusters can be found in the supplements (tab. 37).

Table 16: **Summary of GO-terms of TTP-target genes with orthologs in human and mouse** This table lists only those GO-terms (molecular function) that were enriched for both human and mouse. For the whole set see table 37

(Negative-) Regulation of	Human	Mouse
apoptosis	✓	✓
programmed cell death	✓	✓
cell death	✓	✓
(acute) inflammatory response	✓	✓
immune system development	✓	✓
endocytosis	✓	✓

2.2.15.4 *GO-comparison for our target genes and Mukherjee et al. [102] target genes*

TTP target genes in our dataset to those derived from [102] were already compared for their overlap. As vast differences between those datasets were detected, GO-term enrichment of both datasets was conducted, to check the function of genes targeted by TTP under the hard to compare conditions during both experiments.

The top 3 cluster for TTP targeted genes in our dataset contain endocytosis, defense response, response to wounding, inflammatory response (see tables 33, 34, 35 and 36).

When annotating GO-terms for the Mukherjee et al. [102] target gene list with DAVID (see tab. 17),we do find immune response in cluster 244 (not shown) for the first time, top enriched terms are nuclear lumen and intracellular organelle lumen. Terms like nucleic acid binding and transcription factor binding are shown as most enriched by rank in Mukherjee et al. [102]. These results further highlight the difference between our study of TTP function in a native system compared to over-expression studies, as we are able to show the importance of TTP for inflammatory/wound response.

2.2.15.5 *HuR target genes*

Tables for `TopGO` enriched GO-terms in HuR targets can be found in the supplements (chapter A, table 31 and 32). GO-term enrichment for HuR target genes in WT and TTP-KO shows cytoskeleton associated GO-terms, which fits top HuR targets like ActB and highlights that HuR has no specific function during inflammatory response compared to TTP. Top GO-terms of these datasets are similar and differ mainly by rank.

Table 17: **Mukherjee et al. [102] TTP target genes GO-terms** annotated with DAVID

Category	Term	Count	%
Annotation Cluster 1	Enrichment Score: 35.1501		
GOTERM_CC_FAT	GO:0031981 nuclear lumen	342	14.7034
GOTERM_CC_FAT	GO:0070013 intracellular organelle lumen	381	16.3801
GOTERM_CC_FAT	GO:0043233 organelle lumen	384	16.5090
GOTERM_CC_FAT	GO:0031974 membrane-enclosed lumen	389	16.7240
GOTERM_CC_FAT	GO:0005654 nucleoplasm	220	9.4583
GOTERM_CC_FAT	GO:0044451 nucleoplasm part	142	6.1049
GOTERM_CC_FAT	GO:0005730 nucleolus	160	6.8788
Annotation Cluster 2	Enrichment Score: 27.9342		
SP_PIR_KEYWORDS	nucleus	913	39.2519
SP_PIR_KEYWORDS	Transcription	433	18.6156
SP_PIR_KEYWORDS	transcription regulation	425	18.2717
GOTERM_MF_FAT	GO:0003677 DNA binding	473	20.3353
GOTERM_BP_FAT	GO:0045449 regulation of transcription	513	22.0550
GOTERM_BP_FAT	GO:0006350 transcription	435	18.7016
SP_PIR_KEYWORDS	dna-binding	380	16.3371
GOTERM_BP_FAT	GO:0051252 regulation of RNA metabolic process	357	15.3482
GOTERM_BP_FAT	GO:0006355 regulation of transcription, DNA-dependent	345	14.8323
GOTERM_MF_FAT	GO:0030528 transcription regulator activity	284	12.2098
GOTERM_MF_FAT	GO:0003700 transcription factor activity	186	7.9966
GOTERM_MF_FAT	GO:0043565 sequence-specific DNA binding	120	5.1591
Annotation Cluster 3	Enrichment Score: 14.9707		
GOTERM_CC_FAT	GO:0043228 non-membrane-bounded organelle	431	18.5297
GOTERM_CC_FAT	GO:0043232 intracellular non-membrane-bounded organelle	431	18.5297
GOTERM_CC_FAT	GO:0005856 cytoskeleton	179	7.6956

3 DISCUSSION AND OUTLOOK

3.1 ARESITE 2.0

Already more than 1,000 visitors and 13,000 served requests of AREsite2 in the first half year show, that the recently published update is readily accepted as resource for RNA-protein interaction investigations. This resource is not only interesting for the broad community, it was also used in this thesis to investigate main differences between TTP, HuR and Auf1 bound and unbound motifs. AU/GU/U-content as well as the accessibility of motif embedding regions was analyzed for all motifs contained in AREsite2 . CLIP-Seq experiments, processed by CLIPdb or directly from source are also part of the database, allowing to investigate motifs with and without experimental evidence.

Positive and negative datasets were curated from the database. The positive set consists of motifs with overlapping CLIP-Seq signal and the negative set consists of motifs without overlapping signal. These datasets were not filtered any further due to the lack of RNA-Seq or equivalent data.

Section 2.1.4 shows, that for all combinations of motifs and proteins, AU- and U-content is higher in the positive set than in the negative set. This indicates, that unbound motifs are more often found isolated than bound ones, which are embedded in AU/U-rich context. Analyzing accessibility in terms of probability of being unpaired for both sets shows that bound motifs are in general more accessible then unbound ones, and that their peak accessibility is found around the center of the motif.

Since AU-content and structuredness are correlated, the question arises, which feature is best used for target site prediction. To investigate the descriptive power of these findings, Receiver Operating Characteristic (ROC) curves were computed in section 2.1.5.

ROC analysis shows that accessibility is in most cases not a very good descriptor for both sets, while AU- and U-content of embedding regions has potential. Together, these findings show that there are differences between bound and unbound motifs, even if they are not strong enough to be good predictors. Comparing this results to the ROC analysis with our Sedlyarov et al. [120] dataset in section 2.2.11.4, the impact of thorough preprocessing becomes evident. While the unfiltered, non RNA-Seq normalized AREsite2 data suggest only weak descriptive power for secondary structure, analysis of our filtered and RNA-Seq normalized PAR-iCLIP data shows

the opposite. Binding site quality can only be judged by comparing CLIP-Seq signal to transcript abundance. This highlights the need for accompanying experiments when investigating complex mechanisms like RNA-protein interactions with CLIP-Seq experiments.

One has to keep in mind that due to the nature of CLIP-Seq experiments, which will under normal circumstances not lead to a fully saturated target list, not all motifs that are potentially targeted by one of the three RBPs is in the positive set. While this leads to false negatives, which impact downstream analysis, there are strategies available to deal with this problem.

Additional experiments like RNA-Seq allow to filter for motifs that are located on expressed transcripts, hence available for binding and help to curate more accurate positive and negative sets. Furthermore, expression rates derived from such experiments allow to normalize CLIP-Seq signal to the amount of available target, establishing the means to compare different proteins for their binding behavior and required binding features.

The lack of such experimental data can to some degree be circumvented with *in silico* approaches like *e. g.* GraphProt , which estimates sequence and structure features from a set of validated binding sites and predicts new target sites that share common features. It is, however, obvious that adequately preprocessed, high quality datasets will profit more from such methods than raw datasets, as features used for predictions are extracted directly from the initial set of binding sites.

Comparing descriptor analysis of AREsite2 data and RNA-Seq normalized PAR-iCLIP data emphasizes that the power of RNA secondary structure for binding site discrimination is tightly coupled to adequate processing of CLIP-Seq derived bindingsites, last but not least normalization. The LDA analysis in section 2.2.11.5 validates, that it is indeed possible to discriminate active binding sites from inactive ones, given a good enough training set.

The example analysis (see section 2.1.4) discussed here shows that data from AREsite2 can readily be used for detailed investigation of RNA-RBP interactions. Combining AREsite2 data, PAR-iCLIP derived training sets, a more advanced machine learning algorithm and ideally also additional expression and RNA stability data can potentially be used to predict effects of RBP binding under certain conditions and/or in different cell types.

3.2 PAR-ICLIP

The introduction of CLIP-Seq and its derivatives rendered high-resolution mapping of protein binding sites on RNA molecules in a high-throughput fashion a feasible tool for molecular biology. As for all next generation sequencing based protocols, generation of large datasets is faster than the actual data analysis, highlighting the need for case specific,

thorough and fast bioinformatical processing. While many tools and services offer general and fast analysis of NGS data, this is often not enough to deal with the specifics of certain experimental protocols. A great part of this thesis is concerned with the analysis of PAR-iCLIP derived data, a hybrid method of iCLIP and PAR-CLIP without any ready to use analysis pipeline available. The following sections will discuss the PAR-iCLIP related findings presented in chapter 2, section 2.2.

3.2.1 *Verification of method*

This thesis shows that PAR-iCLIP can be used to identify binding sites of RNA binding proteins with nucleotide resolution (fig. 27) and higher yield than comparable methods (fig. 26).

Under the present experimental conditions, PAR-iCLIP crosslinking should only be possible between thio-uridine and aromatic amino-acid residues. A high rate of thymidine at position 0 of the mapped reads, i.e. the hypothetical crosslink site could be shown for our Sedlyarov et al. [120] dataset. This suggests, that crosslinking at incorporated thio-uridines efficiently causes reverse transcriptase to fall off during reverse transcription directly at the cross-link site (~66% of reads). For the remaining 34% of reads the hypothetical cross-linked site is at an A,G or C. However, the nearest T2C transition is usually within 1ont from the observed position zero. Thus, reverse transcriptase occasionally reads through crosslinks, but seems to fall off in the immediate surrounding in most cases. Experiments that would establish a benchmark for thio-uridine incorporation, crosslink efficiency and reverse transcriptase read-through and drop-off rates are not available yet. Thus, we accept the high rate of thymidine in position 0 and the fact that T2C transitions ($\sim$ 46% of all analyzed transitions, fig. 28) are observed far more often than any other mutation ($\sim 4-5$ times more often than the next most frequent transition) as indicators for high-quality CLIP-Seq .

Although not corrected for biases from mapping errors, sequencing quality or SNP-events, the high amount of T2C transitions can only be explained by read-through events. However, this also shows the main advantage of this method compared to regular PAR-CLIP . For the latter, read-through is mandatory for library preparation, leading to PCR duplicate rates of 80% and more, simply as the sequencing depth is limited.

In case of this study, focusing on transition events only would have led to a loss of many uniquely mappable reads, *i. e.* sequencing depth. So far no standard method for quantification of read-through and crosslink events and incorporation of findings into CLIP-Seq processing and normalization was established. The RNA-Seq based normalization of PAR-iCLIP reads presented in this thesis (see section 2.2.6)

made it possible to compare results for TTP and HuR, including RNA stability correlation analysis, which would otherwise have been problematic. RNA-Seq expression rates were also used to filter the set of unbound motifs, which increased the quality and significance of all downstream analysis steps.

3.2.2 *HuR binds preferentially to 3'UTRs of mouse BMDM mRNAs*

A proposed mechanism for HuRs RNA stabilizing function is that HuR binds to the 3'UTR of its target and the poly-A tail, preventing deadenylases from degrading the latter. PAR-iCLIP derived HuR signal stems almost exclusively (~90%) from binding to 3'UTRs exons. This is true for HuR in $TTP^{+/+}$ (WT) and as well as in $TTP^{-/-}$ (KO) cells. Although we observed intronic binding of HuR (3% in WT and 7% in KO), the amount of crosslinks in this regions is so small, that most peaks were discarded during peak filtering. Similar results have been reported in previous studies of HuR CLIP-Seq, *e. g.* Uren et al. [134]. The fact that signal stems almost exclusive from 3'UTRs contributes to the proposed mechanism for HuRs stabilizing function.

3.2.3 *TTP also binds to intronic regions of mouse BMDM mRNAs*

Although the exact mechanism behind TTPs RNA destabilizing function is still not fully understood (see Brooks and Blackshear [17] for a review), so far only 3'UTR binding could be shown to influence RNA half-life. Also in this work, the majority (53%) of TTP PAR-iCLIP crosslinks could be mapped to 3'UTRs. However, to some surprise, we could annotate a third (32%) of them in intronic regions of coding sequences. 10% of this signal originates from intron 4 of Immune-responsive gene 1 (Irg1). Intronic binding of TTP has been observed before [102], however, it remains elusive whether TTP in this case also causes mRNA destabilization or performs other, yet uncharacterized functions.

Introns might act as sponges that titer TTP away from its regular target sites in 3'UTRs and by this increase target mRNA stability in a cis- or trans- acting manner. Such intronic sponges were described as circular RNAs [98], that for instance control abundance of free/reactive miRNAs [46].

In Sedlyarov et al. [120] we could show that TTP is available in the nucleus, the same compartment were introns are spliced. Thus, our results suggest that introns play a role in regulating or at least tuning concentrations of free TTP. Whether or not this observation applies to other introns remains to be seen, given that introns often contain AU-repeats which are potential TTP binding sites.

The same is true for the idea of circularized intronic sponges, which seems an intriguing explanation for TTPs unexpected binding behav-

ior. The huge amount of signal stemming from a single intron (Irg1 intron 4) and the fact that TTP binds to the spliced out version of the latter, suggest sponge function. As such a sponge has a strong influence on the amount of TTP available for binding, tight control of its own half-life could be established by de-/circularization. Additional, targeted RNA-Seq experiments would allow to proof this concept if circularization point spanning reads can be found.

Anyway, even if Irg1 intron 4 can be shown as TTP sponge in mouse, it is not conserved in human (in contrast to many other intronic ARE-elements, see 2.2.11.1), but the existence of such an interesting regulatory mode should nonetheless be investigated beyond the scope of this one intron.

3.2.4 *Identified target genes and implications*

Many of the known TTP and HuR target genes are present in our top target lists (see section 2.2.7) which further supports the usefulness of the applied PAR-iCLIP method. In section 2.2.6.1 we show that normalization by RNA-Seq estimated expression rates could successfully be applied, leading to normalized target lists that were used for many downstream analysis steps.

This normalization is of importance for the investigated system, immune response in primary mouse BMDMs. It allows to re-rank targets according to TTPs/HuRs binding preference in relation to expression changes due to LPS induction and/or TTP knockout. Although this does not allow direct inference of binding affinities, it is a crucial first step towards such an analysis. Quantified binding signal gives a direct measure to compare targets and protein binding preferences over a range of experimental conditions.

Without previous normalization, analysis steps like correlation with mRNA decay would be rather meaningless. A remaining challenge is the normalization of intronic signal, however, due to the unknown effect of such binding, without consequences for this thesis.

3.2.5 *Different binding region equals different binding motif?*

MEME analysis (see section 2.2.8.1) confirmed published binding motifs for both TTP and HuR. TTP shows permutations of ARE and U-rich motifs throughout peaks in 3'UTRs and intronic regions. The motifs vary slightly in sequence, however, TTP seems to recognize the 3'UTR and intronic motifs comparably well.

To validate this, introns were searched with the 3'UTR motif and vice versa and signal coverage on found motifs was calculated. While signal in UTRs slightly decreased when searching with the intronic motif, the 3'UTR motif applied on intronic binding sites showed more signal coverage. The motif derived from 3'UTRs allows for more vari-

ation due to the variable flanking positions allowing either A or U and thus covers, as expected, more signal. However, this is a strong indicator that TTP does not discriminate 3'UTR from intronic binding sites by motif.

96% of TTP target genes contain the top over-represented MEME motifs somewhere in their 3'UTRs or intronic regions (see tab. 9 in section 2.2.8.1). In about 65% of those genes the potential binding sites are indeed used for binding, as indicated by PAR-iCLIP signals.

Visual inspection of the remaining 35% of genes revealed slight variations of the TTP core motif in TTP peaks, for instance, an additional U in the center (*UAUUUUUAU*) or motifs were the flanking Us are missing (*AUUUA*) or even motifs entirely consisting of Us (*UUUUUUUU*) were found. In the latter cases, overlapping PAR-iCLIP signal of TTP and HuR was observed, explaining the U-richness of these motifs.

We propose that overlapping motifs present a third class of binding motif (see section 2.2.8.2), which can be seen as a hybrid motif for both proteins. However, due to the low number of overlaps, the potential of the derived motif for prediction of other overlapping sites remains to be validated.

3.2.6 *Binding sites are often conserved between mouse and human*

Investigating the conservation of binding sites for TTP and HuR between mouse and human (see section 2.2.9) shows that sites located in 3'UTRs are frequently conserved. Also intronic sites are conserved, but to a lesser extend.

As so far only 3'UTR binding could be shown to influence mRNA stability, this indicates that conclusions drawn here for mouse can be ported to human. The same is true for the identified target genes, where most have orthologs in human.

It is important to emphasize such a finding, as for many interesting studies, including most knockout or knockdown experiments, this provides a way to investigate protein-RNA interactions in a native setting in a model system without having to rely on over-expression in "artificial" cell lines.

However, there is still no guarantee that findings drawn from such studies can be ported 1:1, as cells from model organisms like mouse can and do behave different from human cells, especially under stress.

3.2.7 *Overlap analysis reveals not only competitive binding*

TTP and HuR are known to have antagonistic effects on mRNA half-life. While TTP is a known RNA destabilizer, HuR can prolong mRNA

half-life. It has been shown that both act in complexes with other proteins and/or regulatory factors like miRNAs [17, 86, 101], which makes it hard to identify direct cooperative and antagonistic behavior.

So far, it remained elusive whether they act on the same targets and binding sites under native conditions, i.e. if they compete for the same binding sites thereby directly antagonizing each other. The Sedlyarov et al. [120] PAR-iCLIP experiments analyzed in this thesis provide the basis for competition analysis under native conditions.

To see if TTP and HuR indeed compete for target genes, we analyzed overlapping peak regions for TTP with HuR in WT and in KO cells (see section 2.2.7.1). While 23% of genes containing TTP peaks also contain peaks of HuR in WT (34% in KO), only 10% of those show directly overlapping peak regions with HuR in WT (13% in KO) and 8% under both conditions at the same genomic position.

Thus, TTP and HuR indeed target the same genes to some extent, but they do only rarely share the same binding regions. While HuR in KO binds 75% (229) more genes than in WT, we only detect 50 more genes that are also targeted by TTP. This does not support direct competition as default regulatory mechanism. However, HuR might very well be just one among many protein or (nc)RNA agents that are able to interfere with TTP binding and vice versa, especially under inflammatory stress conditions.

We conclude, that TTP and HuR compete directly for certain targets, but our data suggest that this is not the default. This is in contrast to the study of Mukherjee et al. [102], who found over 80% overlap. However, we investigate the role of both proteins in a native setting, without over-expression or "artificial" cell lines like HEK cells, both potentially resulting in many false positives and non-functional targets. We focus on the role of TTP and HuR in the specific process of immune response in a native setting without overexpression. Thus, we may miss potential targets, either because these targets are simply not expressed in BMDMs, or they vanish in comparison to targets important during inflammatory stress. However, for any CLIP-Seq experiment, there is always a tradeoff between finding as many potential binding sites as possible and finding binding sites that have a real biological meaning for a system. The ideal case would be to combine both kinds of study, to first draw conclusions on general binding behavior, and then investigate a more specialized case to see if conclusions drawn from the general investigation still hold true. This was successfully done in this thesis, highlighting both, differences as well as commonalities between two TTP focused CLIP-Seq studies.

3.2.8 *Cooperative vs. competitive binding in broader context*

For the PAR-iCLIP dataset of TTP and HuR binding sites, no preference for direct competition could be found. In general the binding sites of both proteins do not overlap directly, obvious when comparing their preferred binding motifs. This, however, does not exclude competitive behavior, as competitive effects do not necessarily need direct overlaps. To investigate this in more detail, all pairs of binding sites for TTP/HuR and miRNAs extracted from an Ago-CLIP-Seq experiment [86] were compared, see section 2.2.13.

With a simple model (see section 2.2.13, equation 3), changes in binding free energy in presence of binding partners on the same 3'UTR were analyzed. Although this study is only preliminary, a proposed regulatory feedback-loop between miRNA-27, HuR and TTP could already be shown to be even more complex.

However, although the simple model used here is by far not complete and it requires more effort and data to come up with a more convincing model, first results show potential for future investigations in this direction.

To shed more light onto this complex topic, additional data has to be included. Efforts like eCLIP experiments applied by the GENCODE consortium make many new CLIP-Seq datasets available in a comparable manner. However, the full complexity mechanism of RNA-protein interactions will only be deciphered if the full spectrum of interactions is taken into account, including protein-RNA as well as RNA-RNA and protein-protein interactions on sequence as well as on structure level.

3.2.9 *Is sequence or structure the better predictor for functional binding sites*

Taking together the results from this thesis, it seems that both, sequence and structure are reasonable descriptors for bound and unbound motifs in human and mouse. For motifs derived from ARE-site2, one has to keep in mind that this dataset is unfiltered and for sure contains a lot of false negatives, simply due to the fact that only a limited amount of CLIP-Seq experiments is available and that these experiments are not saturated and not accompanied by matching RNA-Seq experiments or other adequate measures of transcript abundance.

For PAR-iCLIP binding sites in mouse, which were investigated in more detail, it could be shown that accessibility, or opening energy is a solid, if not even better descriptor than AU-content. The main conclusion that can be drawn from this investigation is that bound sites are usually found in a context which is both, more AU-rich and accessible than unbound motifs.

Linear discriminators, trained with AU-content and opening energy as descriptors for the PAR-iCLIP dataset and tested there, as well as with the AREsite2 dataset, prove that accessibility of binding sites is a solid discriminator for both, TTP and HuR binding. Although AU-content is already a good discriminator, accessibility of motif embedding regions outperforms it in both test sets, especially for HuR.

One has to keep in mind that the proteins investigated here are all known to have strong preferences for specific binding site sequence, as was confirmed here too. These motifs are highly enriched in Uridine and Adenine, which increases the chance of finding unpaired stretches embedding binding sites. This suggests that the high AU-content of the surrounding region serves to make the binding site more accessible. For a high quality binding site set and given careful RNA secondary structure prediction, accessibility of motif embedding region provides a layer of information that should definitely be integrated into target site prediction.

3.3 CONCLUDING REMARKS AND OUTLOOK

Many studies that focus on interactions between proteins and RNA have been presented over the last years, with numbers growing from day to day. Their experimental design has changed over the years, from single target, single interaction, gel electrophoresis driven to whole cell, transcriptome wide interactome analysis driven by high throughput sequencing advancements.

Sparked by the invention of CLIP-Seq techniques, studies nowadays present target lists containing hundreds or thousands of interaction sites for single proteins. This evolution from single target to the whole interactome allows researchers to draw conclusions about binding preferences and interaction networks on a whole new scale. For the first time ever, it is possible to get detailed knowledge about the role and behavior of an RNA binding protein in an *in vivo* setting and on a cell wide scale, probably even in single cell resolution in the not so far future.

This leads to new insights and allows to investigate correlation between interaction and cellular events that were not possibly drawn before.

However, with all this new data and information it is still important to keep in mind that CLIP-Seq also has its limits.

One limitation of CLIP-Seq is that it can only resolve interactions that are exposed to the crosslinking agent, which is usually UV-light. The saturation rate of UV-crosslinking seems to be highly variable, between cell-types as well as between proteins, and so far no control experiment which could be used to deduct saturation rates was shown.

Another drawback is the lack of a solid negative control, preventing experimental validation of false positives. IP-bases techniques can potentially introduce a range of errors, highly depending on the quality of the used antibody, although quality control is possible to ensure specificity. CLIP-Seq alone can not determine the affinity of a protein for certain targets and is only a quantitative measure in terms of which RNAs are targeted by the protein of interest.

Still, CLIP-Seq is a solid technique for the investigation of proteins that directly interact with RNAs, as long as certain quality standards are fulfilled.

Any CLIP-Seq experiment profits from matching RNA-Seq -experiments, which allow to draw conclusions about transcript abundance and can be used to normalize CLIP-Seq signal to available copies of target RNAs. Furthermore, accompanying experiments allow to extend the information derived from CLIP-Seq experiments from binding site location to biological function, *e. g.* RNA half-life control.

Only very few experiments are concerned with the cooperative or antagonistic effects of RBPs. Such effects can, however, have a huge impact on the interactome of a protein in a certain cell type under certain conditions. Another point to keep in mind is that many RBPs interact not alone but in a complex with other proteins, RNPs or other molecules, which potentially have a strong effect on the choice of target.

This could be circumvented by combining *in vivo* approaches like CLIP-Seq with *in vitro* - experiments like RNA-bind'n-seq, which exclude naturally occurring partners/competitors and allow to focus on a single player. Such experiments, however, can never cope with the complexity of *in vivo* experiments, and results have to be compared carefully.

The combination of CLIP-Seq with other experiments also allows to draw conclusions about the consequences of successful binding, as could be shown for TTP and mRNA decay [120]. Although the correlation between decay and TTP binding signal is strongest for a specific TTP target (Tnf-α), a general trend towards TTP function as regulator of late immune response could be shown.

The case of TTP and mRNA decay is special and it may be harder to find consequences of other RNA-RBP interactions, however, this part remains the most interesting, as interaction without consequences is only half of the story.

Taken together, CLIP-Seq has a huge potential for the investigation of RNA-protein interactions. Careful planing and selection of adequate cellular system, as well as accompanying experiments like RNA-Seq are prerequisites for a comprehensive investigation.

The next years will bring further advances in experimental and *in silico* approaches which will shed more light on the complex interactome of higher cells, creating a basis for synthetic approaches that allow us to take control and directly influence the balance towards our needs.

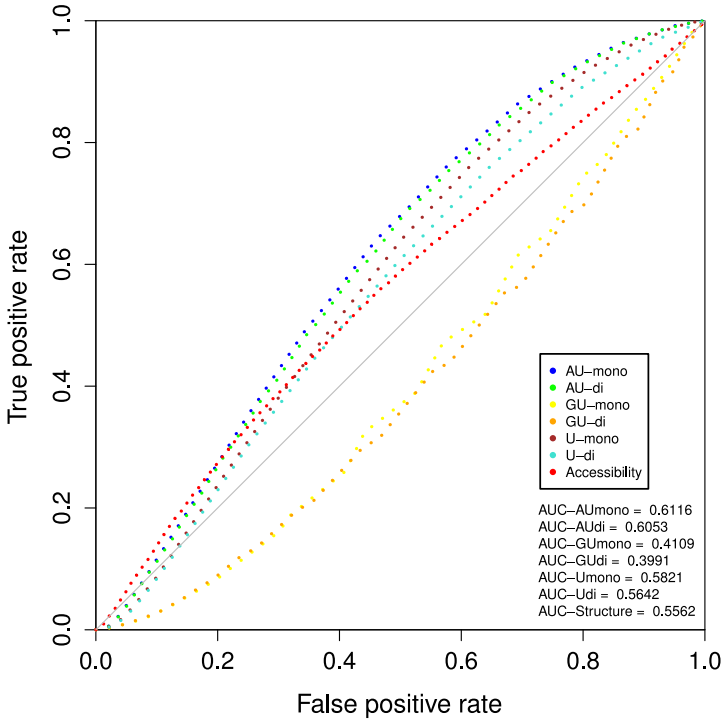
The End

A

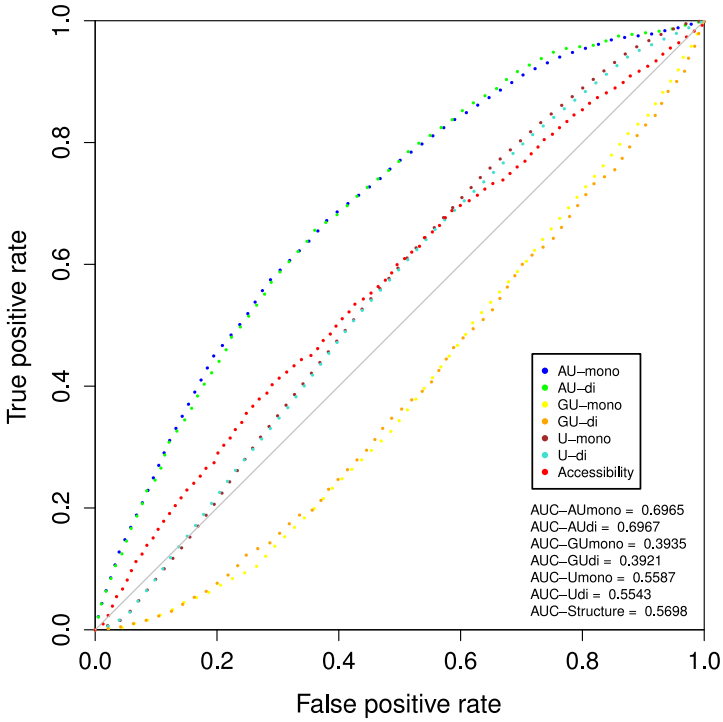
APPENDIX

A.1 ARESITE2_SUPPLEMENTS

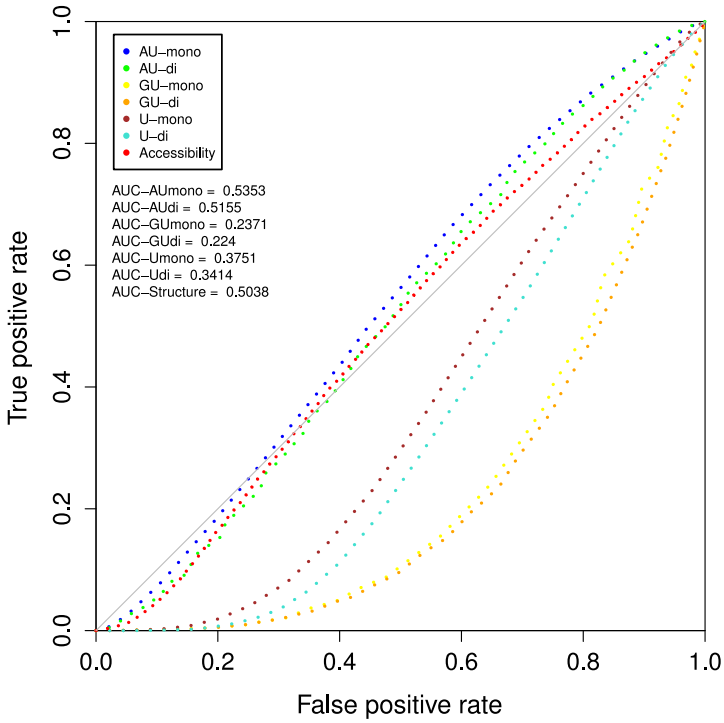
This section contains supplementary figures for section 2.1.5. Figures 46 to 48 show ROC curves from the descriptor analysis of AREsite2 derived HuR, TTP and Auf1 bound and unbound AU/GU/U-rich elements in human. Figures 49 to 51 show ROC curves from the descriptor analysis of AREsite2 derived HuR and TTP bound and unbound AU/GU/U-rich elements in mouse. In contrast to the figures in section 2.1.5, the here presented figures contain ROC curves for all combinations of motifs and proteins, even if the motifs are not main targets for the protein of interest.



AUUUA



GUUUG



UUUUU

Figure 46: **Descriptor analysis of nucleotide content vs accessibility of Auf1 bound/unbound AU/GU/U-rich elements in human** ROC curve to visualize the descriptive power of accessibility and nucleotide content in terms of Area under the ROC (AUC).

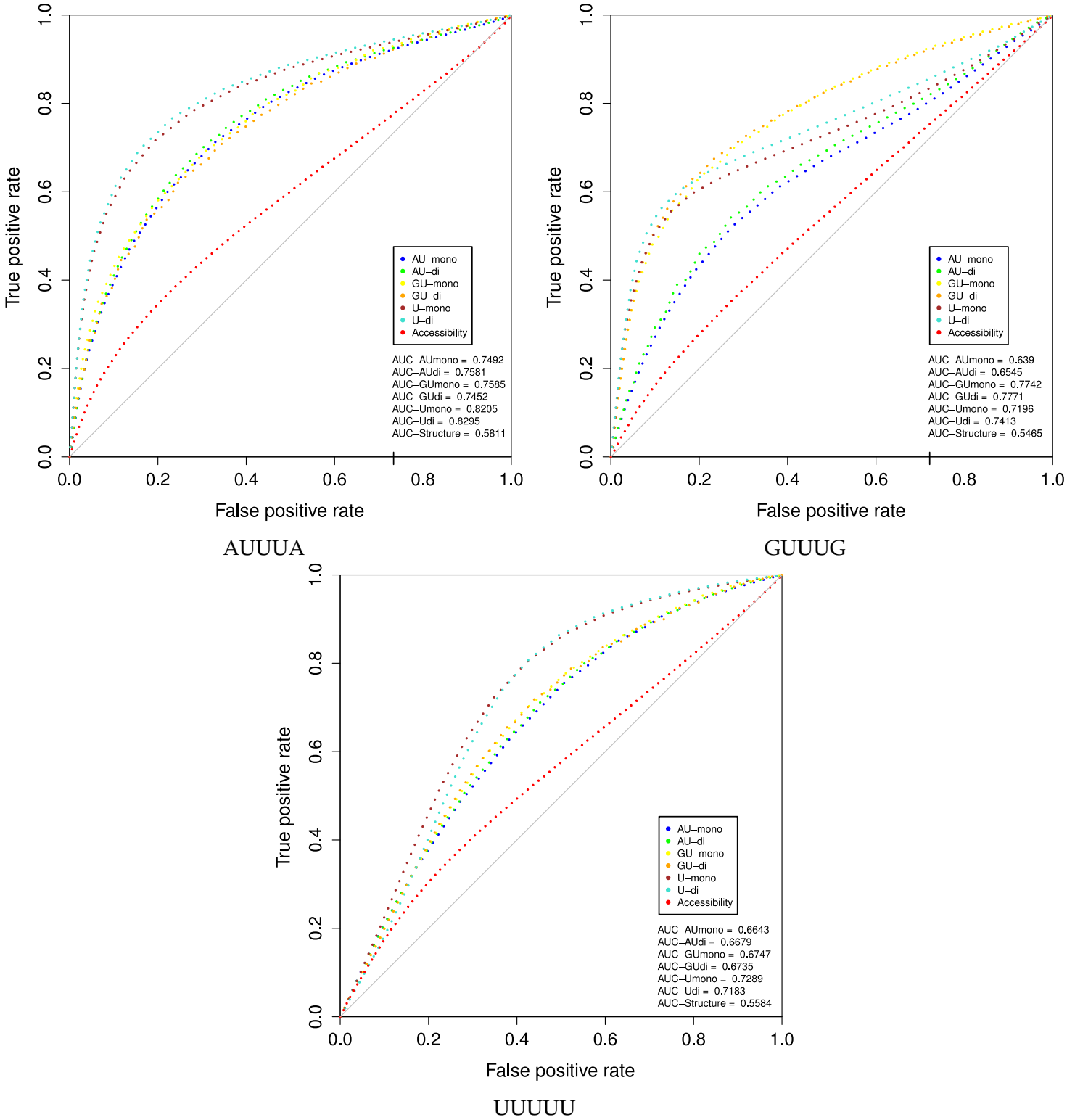


Figure 47: **Descriptor analysis of nucleotide content vs accessibility of HuR bound/unbound AU/GU/U-rich elements in human** ROC curve to visualize the descriptive power of accessibility and nucleotide content in terms of Area under the ROC (AUC).

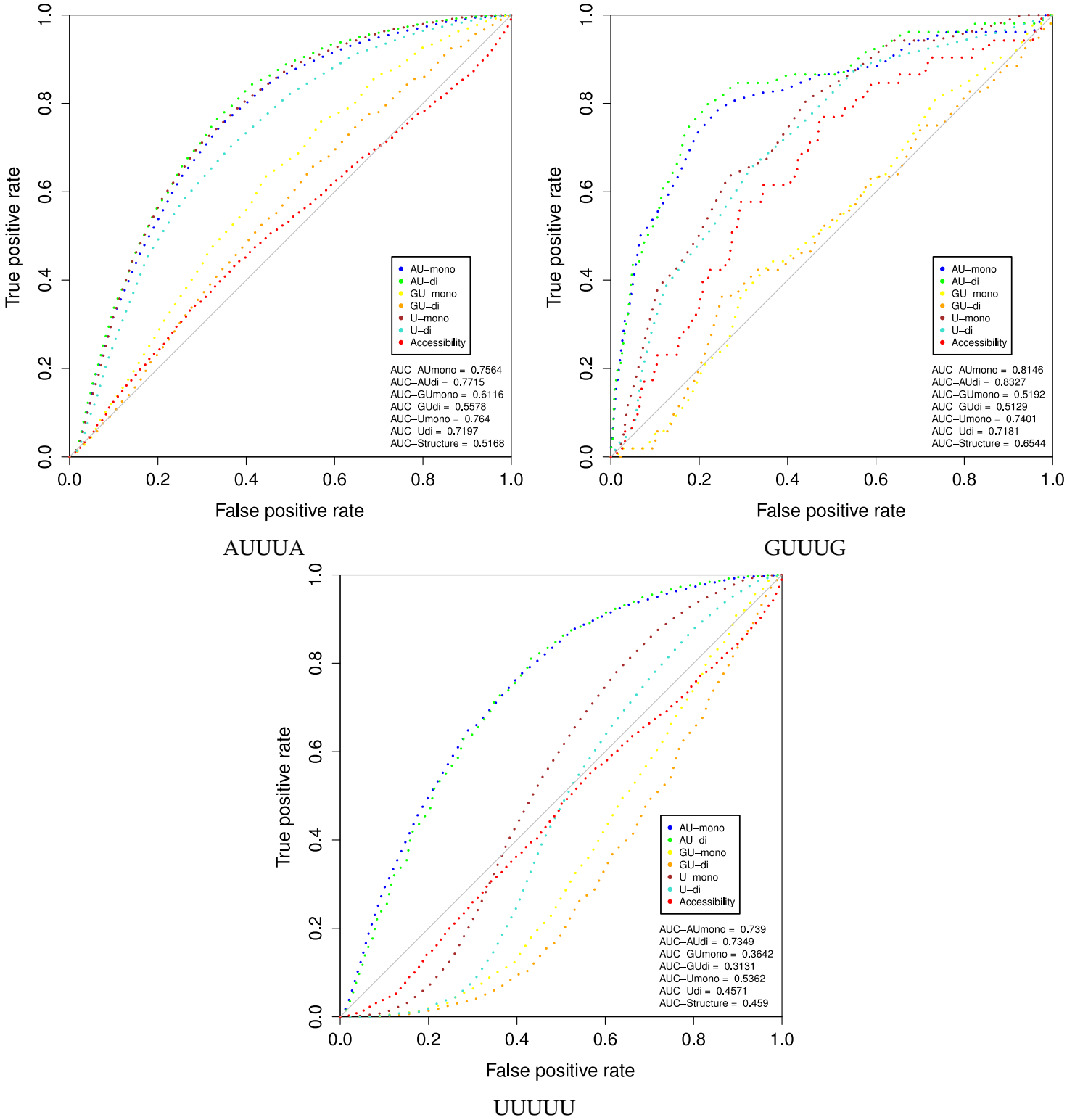


Figure 48: **Descriptor analysis of nucleotide content vs accessibility of TTP bound/unbound AU/GU/U-rich elements in human** ROC curve to visualize the descriptive power of accessibility and nucleotide content in terms of Area under the ROC (AUC).

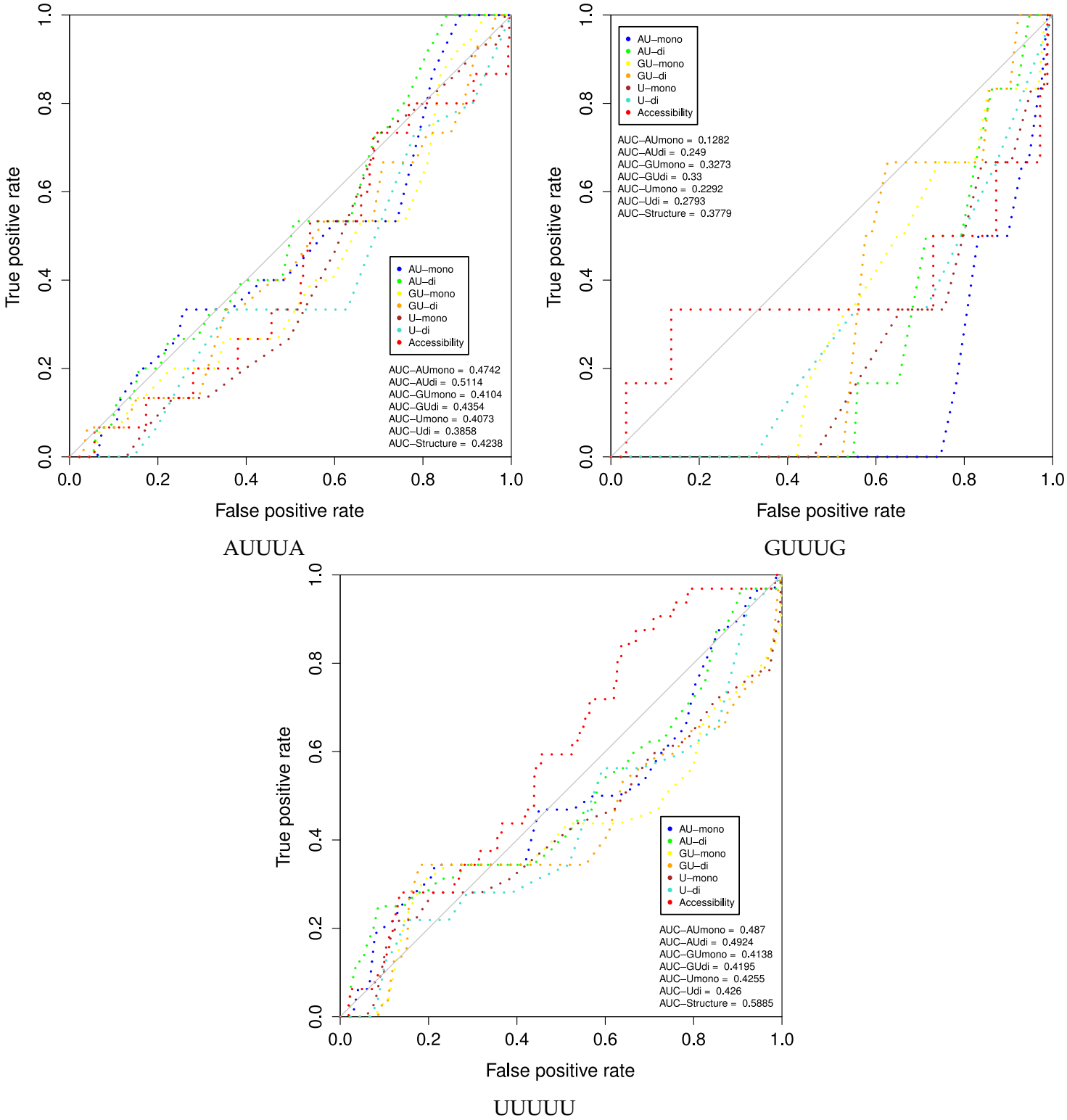


Figure 49: **Descriptor analysis of nucleotide content vs accessibility of HuR bound/unbound AU/GU/U-rich elements in human** ROC curve to visualize the descriptive power of accessibility and nucleotide content in terms of Area under the ROC (AUC).

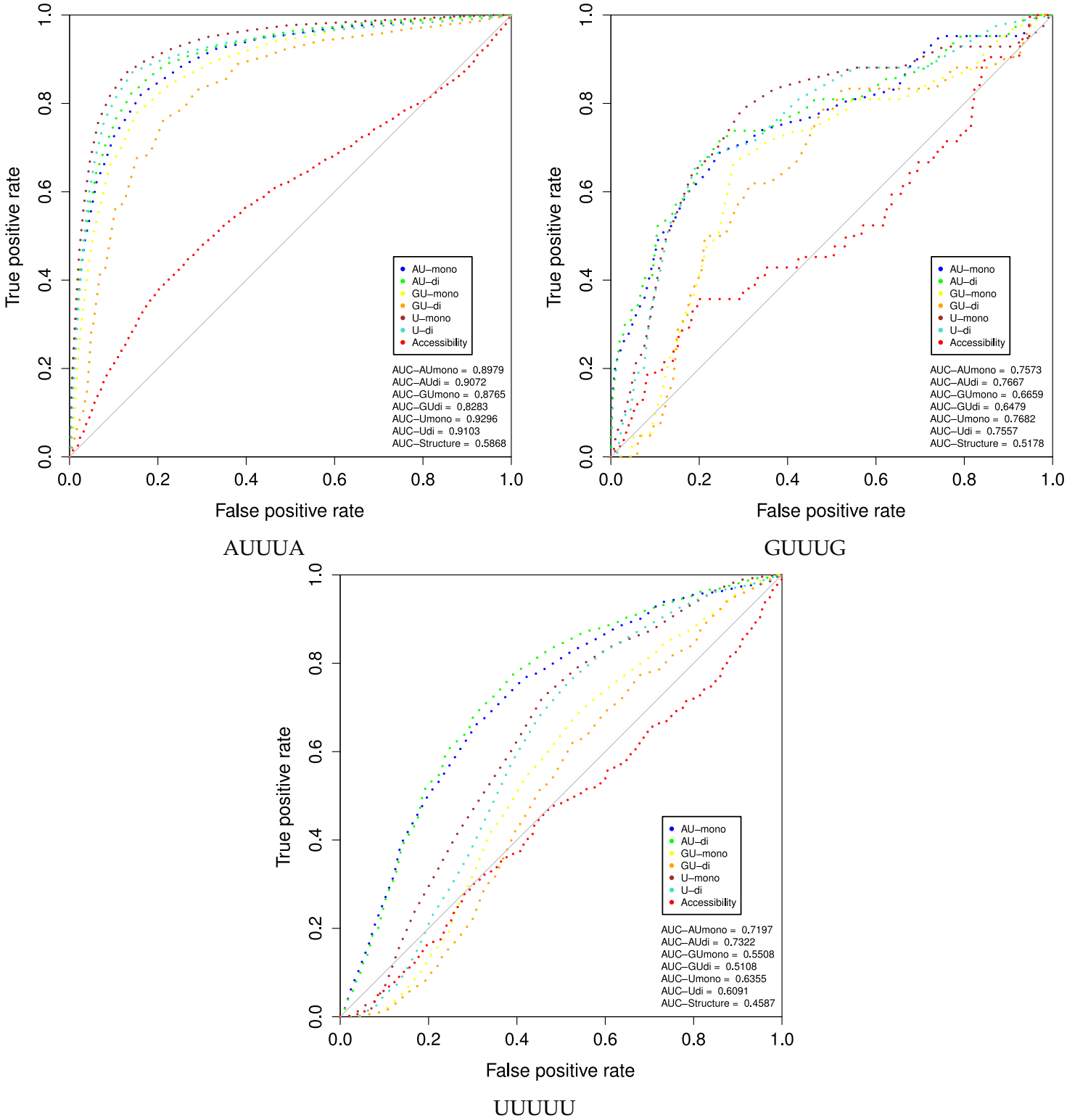


Figure 50: **Descriptor analysis of nucleotide content vs accessibility of TTP 3h bound/unbound AU/GU/U-rich elements in human** ROC curve to visualize the descriptive power of accessibility and nucleotide content in terms of Area under the ROC (AUC).

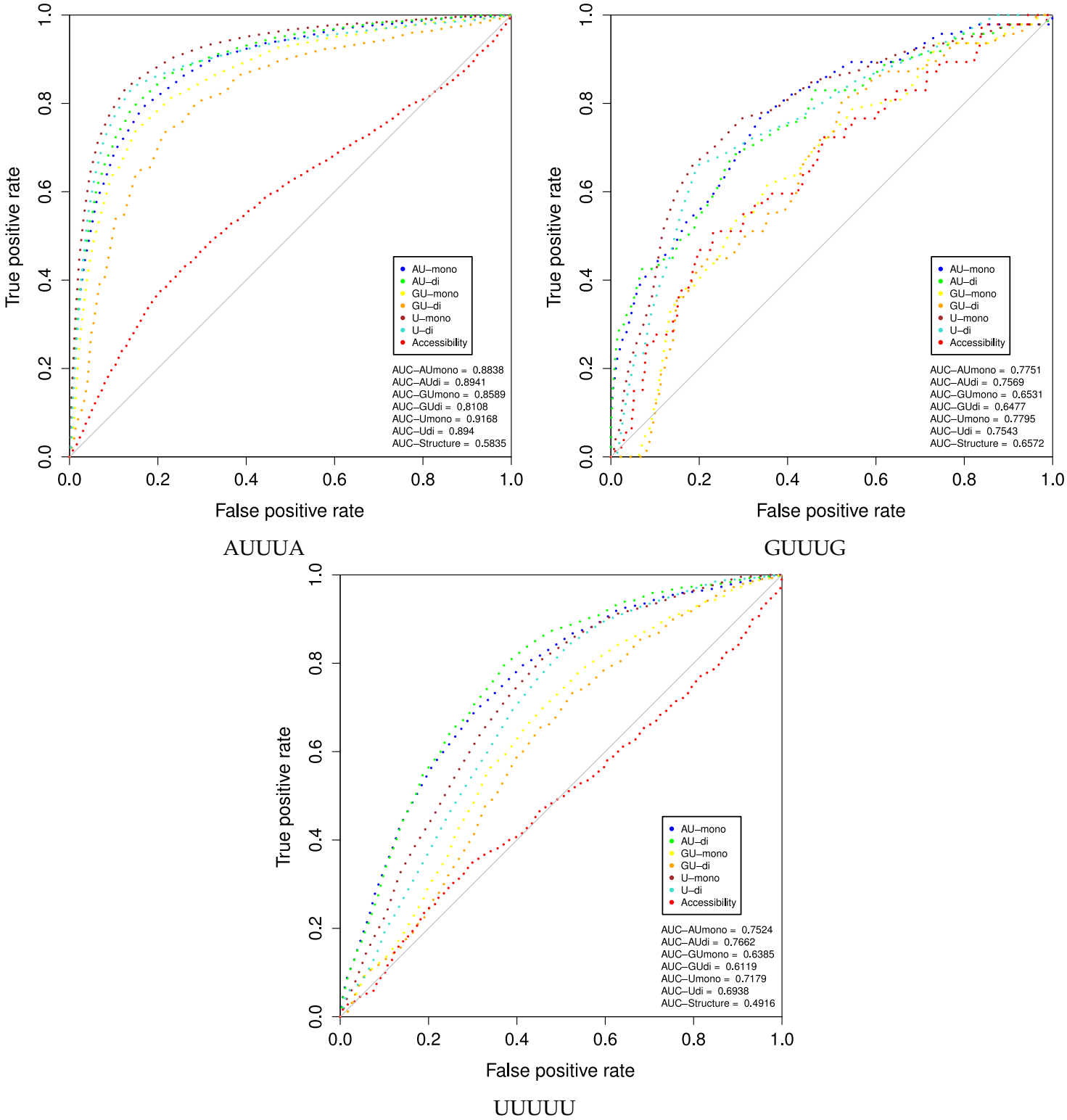


Figure 51: **Descriptor analysis of nucleotide content vs accessibility of TTP 3h bound/unbound AU/GU/U-rich elements in human** ROC curve to visualize the descriptive power of accessibility and nucleotide content in terms of Area under the ROC (AUC).

A.2 PAR-ICLIP SUPPLEMENTS

This section provides supplementary information to the analysis of PAR-iCLIP data for TTP and HuR in mouse macrophages in section 2.2.

A.2.1 *Top 10 targets*

Tables of top target genes and introns of TTP and HuR, both normalized with RNA-Seq and without normalization, are presented here.

Table 18: **Top 10 TTP target genes** with the highest Par-iCLIP signal in their union of peaks, ranked by mean among replicates. Shown is the region of the strongest peak, amount of signal in this peak region, the genomic region the peak maps to, the mean Par-iCLIP signal across replicates and the gene expression rate in FPKMs.

Gene	ENSEMBL ID	Strongest peak genomic region	Strongest peak signal	Partition of strongest peak	Mean Par-iCLIP signal	Gene FPKM
Tnf	ENSMUSG00000024401	chr17:35336601 35336674: 1	1609463	Exon 3UTR	1,495,449	161.068
Ptgs2	ENSMUSG00000032487	chr1:151952917 151952975:1	1569150	Exon 3UTR	1,495,412	112.354
Ccl3	ENSMUSG00000000982	chr11:83461513 83461602: 1	1368647	Exon 3UTR	1,271,499	922.847
Irg1	ENSMUSG00000022126	chr14:103451288 103451374:1	605044	Intron CDS	644,498	491.032
Cxcl2	ENSMUSG00000058427	chr5:91334336 91334391:1	688666	Exon 3UTR	561,015	87.8696
Pid1	ENSMUSG00000045658	chr1:84243404 84243539: 1	250560	Intron 5UTR Intron CDS	464,130	134.445
Zfp36	ENSMUSG00000044786	chr7:29161949 29161990: 1	339483	Exon 3UTR	388,062	65.2556
Pfkfb3	ENSMUSG00000026773	chr2:11411408 11411482: 1	131423	Intron CDS	176,135	145.673
Emr1	ENSMUSG00000004730	chr17:57542299 57542404:1	96764	Intron CDS	174,432	379.894
Ccl4	ENSMUSG00000018930	chr11:83478043 83478073:1	141977	Exon 3UTR	173,624	703.314

Table 19: **Top 10 HuR target genes in WT** cells with the highest Par-iCLIP signal in their union of peaks, ranked by mean among replicates. Shown is the region of the strongest peak, amount of signal in this peak region, the genomic region the peak maps to, the mean Par-iCLIP signal across replicates and the gene expression rate in FPKMs.

Gene	ENSEMBL ID	Strongest peak genomic region	Strongest peak signal	Partition of strongest peak	Mean Par-iCLIP signal	Gene FPKM
Actb	ENSMUSG00000029580	chr5:143665308 143665354: 1	2722347	Exon 3UTR	4,328,409	2125.05
Sdc4	ENSMUSG00000017009	chr2:164250606 164250657: 1	750536	Exon 3UTR	1,210,840	1035.69
Cd44	ENSMUSG00000005087	chr2:102653598 102653677: 1	201012	Exon 3UTR	484,712	118.355
Marcks	ENSMUSG00000069662	chr10:36855743 36855811: 1	176464	Exon 3UTR	338,082	34.8283
Cd47	ENSMUSG00000055447	chr16:49912930 49912997:1	81075	Exon 3UTR	314,126	207.251
Rsad2	ENSMUSG00000020641	chr12:27129919 27130033: 1	127846	Exon 3UTR	250,643	731.273
Dpysl2	ENSMUSG00000022048	chr14:67423346 67423376: 1	124015	Exon 3UTR	233,583	35.6168
Sod2	ENSMUSG00000006818	chr17:13209961 13210016:1	47171	Exon 3UTR	205,347	147.627
Lrrc58	ENSMUSG00000034158	chr16:37883337 37883363:1	80680	Exon 3UTR	197,313	31.4827
Gas7	ENSMUSG00000033066	chr11:67500616 67500639:1	75013	Exon 3UTR	174,711	269.692

Table 20: **Top 10 HuR target genes in TTP-KO** cells with the highest Par-iCLIP signal in their union of peaks, ranked by mean among replicates. Shown is the region of the strongest peak, amount of signal in this peak region, the genomic region the peak maps to, the mean Par-iCLIP signal across replicates and the gene expression rate in FPKMs.

Gene	ENSEMBL ID	Strongest peak genomic region	Strongest peak signal	Partition of strongest peak	Mean Par-iCLIP signal	Gene FPKM
Actb	ENSMUSG00000029580	chr5:143665308 143665354: 1	2021844	Exon 3UTR	2,694,442	1951.29
Sdc4	ENSMUSG00000017009	chr2:164250606 164250658: 1	539468	Exon 3UTR	1,185,836	1176.58
Cd44	ENSMUSG00000005087	chr2:102653589 102653677: 1	140158	Exon 3UTR	317,935	125.522
Marcks	ENSMUSG00000069662	chr10:36855743 36855811: 1	171505	Exon 3UTR	298,742	33.627
Dpysl2	ENSMUSG00000022048	chr14:67423346 67423377: 1	124258	Exon 3UTR	215,899	37.6543
Ccl4	ENSMUSG00000018930	chr11:83477995 83478073:1	180070	Exon 3UTR	171,895	804.768
Maf	ENSMUSG00000055435	chr8:118229043 118229131: 1	62558	Exon 3UTR Intron CDS	169,473	22.3215
Cebpb	ENSMUSG00000056501	chr2:167515726 167515803:1	119220	Exon 3UTR	167,776	77.609
Cd47	ENSMUSG00000055447	chr16:49912930 49913019:1	80184	Exon 3UTR	166,291	202.494
Ahnak	ENSMUSG00000069833	chr19:9093056 9093174:1	176816	Exon 3UTR Intron CDS	165,812	62.3047

Table 21: **Top 10 TTP target genes 3 h after LPS induction** with the highest Par-iCLIP signal in their union of peaks, ranked by mean among replicates. Shown is the region of the strongest peak, amount of signal in this peak region, the genomic region the peak maps to and the gene expression rate in FPKMs.

Gene	ENSEMBL ID	Strongest peak genomic region	Strongest peak signal	Partition of strongest peak	Mean Par-iCLIP signal	Gene FPKM
Tnf	ENSMUSG00000024401	chr17:35336601-35336670:-1	4184700	Exon 3UTR	3.377.951	430.724
Ccl3	ENSMUSG00000000982	chr11:83461513-83461583:-1	1508497	Exon 3UTR	1.271.768	1180.6
Irg1	ENSMUSG00000022126	chr14:103451319-103451374:1	924921	Intron CDS	899.394	362.916
Cxcl2	ENSMUSG00000058427	chr5:91334347-91334391:1	1366627	Exon 3UTR	851.323	283.305
Ptgs2	ENSMUSG00000032487	chr1:151952948-151952975:1	743934	Exon 3UTR	792.441	90.5835
Ccl4	ENSMUSG00000018930	chr11:83478043-83478073:1	288978	Exon 3UTR	254.379	1251.1
Zfp36	ENSMUSG00000044786	chr7:29161950-29161972:-1	225159	Exon 3UTR	249.373	53.6747
Adap2	ENSMUSG00000020709	chr11:79987127-79987183:1	147799	Intron CDS	189.247	76.5687
Slc28a2	ENSMUSG00000027219	chr2:122269203-122269247:1	108576	Intron CDS	186.819	36.0989
Pid1	ENSMUSG00000045658	chr1:84243404-84243475:-1	88863	Intron 5UTR | Intron CDS	160.950	75.1633

A.2.2 *Top 10 RNA-Seq normalized targets*

Table 22: **Top 10 TTP target genes normalized** with the highest Par-iCLIP signal in their union of peaks normalized by gene expression and ranked by mean among replicates. Shown is the region of the strongest peak, the normalized amount of signal in this peak region (PeakScore), the genomic region the peak maps to, the mean Par-iCLIP signal across replicates (GeneScore) and the gene expression rate in FPKMs.

Gene	ENSEMBL ID	Strongest peak genomic region	PeakScore	Partition of strongest peak	Mean GeneScore	Gene FPKM
Ptgs2	ENSMUSG00000032487	chr1:151952917-151952975:1	13966	Exon 3UTR	4416	112.354
Tnf	ENSMUSG00000024401	chr17:35336601-35336674:-1	9992	Exon 3UTR	4053	161.068
Cxcl2	ENSMUSG00000058427	chr5:91334336-91334391:1	8490	Exon 3UTR	2755	87.8696
Zfp36	ENSMUSG00000044786	chr7:29161975-29161990:-1	5202	Exon 3UTR	2296	65.2556
Ccl3	ENSMUSG00000000982	chr11:83461513-83461612:-1	1483	Exon 3UTR	459	922.847
Pfkfb3	ENSMUSG00000026773	chr2:11393704-11393750:-1	4382	Exon 3UTR	323	145.673
Il12b	ENSMUSG00000004296	chr11:44226968-44227004:1	814	Exon 3UTR	305	13.825
Dck	ENSMUSG00000029366	chr5:89210411-89210460:1	361	Exon 3UTR	246	23.4804
Nlrp3	ENSMUSG00000032691	chr11:59380344-59380370:1	776	Exon 3UTR	233	38.507

Table 23: **Top 10 HuR target genes normalized** in WT cells with the highest Par-iCLIP signal in their union of peaks normalized by gene expression and ranked by mean among replicates. Shown is the region of the strongest peak, the normalized amount of signal in this peak region (PeakScore), the genomic region the peak maps to, the mean Par-iCLIP signal across replicates (GeneScore) and the gene expression rate in FPKMs.

Gene	ENSEMBL ID	Strongest peak genomic region	PeakScore	Partition of strongest peak	Mean GeneScore	Gene FPKM
Cd44	ENSMUSG00000005087	chr2:102653598-102653677:-1	15304	Exon 3UTR	678	118.355
Sept2	ENSMUSG00000026276	chr1:95404442-95404464:1	5297	Exon 3UTR	641	67.0354
Tmem2	ENSMUSG00000024754	chr19:21932508-21932570:1	1689	Exon 3UTR	492	19.3048
Cd81	ENSMUSG00000037706	chr7:150253345-150253462:1	5221	Exon 3UTR	470	56.7185
Actb	ENSMUSG00000029580	chr5:143665116-143665156:-1	3272	Exon 3UTR	453	2125.05
Lass6	ENSMUSG00000027035	chr2:68951708-68951723:1	10179	Exon 3UTR	437	266.681
Otub1	ENSMUSG00000024767	chr19:7273421-7273439:-1	749	Exon 3UTR	408	45.9404
Stat5a	ENSMUSG00000004043	chr11:100745933-100745949:1	746	Exon 3UTR	401	29.0541
Nrp2	ENSMUSG00000025969	chr1:62842753-62842842:1	2617	Exon 3UTR	386	56.9863
Irak2	ENSMUSG00000060477	chr6:113644740-113644795:1	681	Exon 3UTR	381	29.624

Table 24: **Top 10 HuR target genes normalized** in TTP-KO cells with the highest Par-iCLIP signal in their union of peaks normalized by gene expression and ranked by mean among replicates. Shown is the region of the strongest peak, the normalized amount of signal in this peak region (PeakScore), the genomic region the peak maps to, the mean Par-iCLIP signal across replicates (GeneScore) and the gene expression rate in FPKMs.

Gene	ENSEMBL ID	Strongest peak genomic region	PeakScore	Partition of strongest peak	Mean GeneScore	Gene FPKM
Ahnak	ENSMUSG00000069833	chr19:9093056-9093174:1	13924	Exon 3UTR	2253	62.3047
Nrp2	ENSMUSG00000025969	chr1:62842753-62842842:1	3088	Exon 3UTR	752	58.9408
Lass6	ENSMUSG00000027035	chr2:68951767-68951876:1	10638	Exon 3UTR	659	213.957
Actb	ENSMUSG00000029580	chr5:143665329-143665355:-1	2408	Exon 3UTR	599	1951.29
Pmp22	ENSMUSG00000018217	chr11:62972196-62972289:1	3452	Exon 3UTR	595	42.804
Calcrl	ENSMUSG00000059588	chr2:84171724-84171734:-1	5395	Exon 3UTR	532	65.1811
Otub1	ENSMUSG00000024767	chr19:7273399-7273446:-1	616	Exon 3UTR	446	45.6824
Cd44	ENSMUSG00000005087	chr2:102653560-102653584:-1	8150	Exon 3UTR	443	125.522
Upf1	ENSMUSG00000058301	chr8:72855832-72855890:-1	1895	Exon 3UTR	429	10.1765
Marcks	ENSMUSG00000069662	chr10:36855743-36855811:-1	5100	Exon 3UTR	427	33.627

Table 25: **Top 10 TTP target genes 3 h after LPS induction normalized**, with the highest Par-iCLIP signal in their union of peaks normalized by gene expression and ranked by mean among replicates. Shown is the region of the strongest peak, the normalized amount of signal in this peak region (PeakScore), the genomic region the peak maps to, the mean Par-iCLIP signal across replicates (GeneScore) and the gene expression rate in FPKMs.

Gene	ENSEMBL ID	Strongest peak genomic region	PeakScore	Partition of strongest peak	Mean GeneScore	Gene FPKM
Ptgs2	ENSMUSG00000032487	chr1:151952948-151952975:1	8212.69	Exon 3UTR	2316.37	90.5835
Tnf	ENSMUSG00000024401	chr17:35336471-35336485:-1	9715.51	Exon 3UTR	1824.02	430.724
Zfp36	ENSMUSG00000044786	chr7:29161950-29161972:-1	4194.88	Exon 3UTR	696.72	53.6747
Ccl3	ENSMUSG00000000982	chr11:83461513-83461583:-1	1277.74	Exon 3UTR	340.79	1180.6
Cxcl2	ENSMUSG00000058427	chr5:91334347-91334391:1	4823.87	Exon 3UTR	323.89	283.305
Il12b	ENSMUSG00000004296	chr11:44226982-44226996:1	634.7	Exon 3UTR	306.56	15.3565
49334-26M1-1Rik	ENSMUSG00000021133	chr12:81981728-81981780:1	993.48	Exon 3UTR	262.52	22.1811
Trim30a	ENSMUSG00000030921	chr7:111559516-111559582:-1	641.22	Exon 3UTR	164.21	47.6107
Dck	ENSMUSG00000029366	chr5:89210411-89210422:1	358.71	Exon 3UTR	152.84	12.0545
Il10ra	ENSMUSG00000032089	chr9:45062196-45062230:-1	1115.03	Exon 3UTR	143.53	31.5313

A.2.3 *Top 10 TTP intronic targets*

Table 26: **Top10 mouse introns** and overlap of uniquely mapped TTP reads before peak filtering. Shown is the gene, the genomic position of the intron, gene and transcript IDs containing the intron, total Par-iCLIP signal overlapping the intron, nucleotides with signal, length of the intron, signal normalized to length, and expression rate of the intron containing gene.

Gene	Chromosome	Intron start	Intron end	Strand	ENSEMBL gene ID	ENSEMBL transcript ID with intron number	Gene type	Par-iCLIP signal	nt with signal	nt length	Signal / length	Gene FPKM
Irg1	chr14	103450742	103453728	+	ENSMUSG00000022126	ENSMUST00000022722_4	protein coding	394932	2163	2986	0.72	491.032
Pid1	chr1	84035082	84155806	-	ENSMUSG00000045658	ENSMUST00000168574_2 ENSMUST00000167490_3 ENSMUST00000051845_2	protein coding	98139	11367	120724	0.09	134.445
Slc28a2	chr2	122267587	122272621	+	ENSMUSG00000027219	ENSMUST00000110525_4 ENSMUST00000110524_3 ENSMUST00000028652_4	protein coding	80216	1824	5034	0.36	270.778
Pfkfb3	chr2	11411372	11411916	-	ENSMUSG00000026773	ENSMUST00000171188_4 ENSMUST00000170196_4 ENSMUST00000114846_4 ENSMUST00000114845_4 ENSMUST00000114844_4 ENSMUST00000100411_4 ENSMUST00000049849_4 ENSMUST00000028114_4	protein coding	73128	289	544	0.53	145.673
Emr1	chr17	57542284	57546256	+	ENSMUSG00000004730	ENSMUST00000086763_5 ENSMUST00000004850_5	protein coding	63554	1768	3972	0.45	379.894
Adap2	chr11	79984263	79987589	+	ENSMUSG00000020709	ENSMUST00000164168_7 ENSMUST00000021050_7	protein coding	62995	1668	3326	0.50	145.461
Lass6	chr2	68699862	68772571	+	ENSMUSG00000027035	ENSMUST00000028426_1	protein coding	45743	6474	72709	0.09	266.681
Lyn	chr4	3605511	3665871	+	ENSMUSG00000042228	ENSMUST00000103010_1 ENSMUST00000041377_1	protein coding	42052	5195	60360	0.09	156.841
Csf3r	chr4	125709543	125711413	+	ENSMUSG00000028859	ENSMUST00000106162_6 ENSMUST00000030673_5	protein coding	41811	1071	1870	0.57	17.0415
Dennd1a	chr2	38015370	38098921	-	ENSMUSG00000035392	ENSMUST00000130472_2 ENSMUST00000102787_2	protein coding	36949	4334	83551	0.05	104.527

A.2.4 *RNA-Seq normalized PAR-iCLIP peaks*

This section contains the highest ranked peaks by RNA-Seq normalized PAR-iCLIP signal of TTP and HuR.

Table 27: **Top 10 TTP peaks after normalization** by expression. Shown is the region of the peak, the normalized amount of signal in this peak region (PeakScore), the genomic partition and sequence of the peak.

Gene	ENSEMBL ID	Peak genomic region	PeakScore	Partition	Sequence
Ptgs2	ENSMUSG00000032487	chr1:151952910-151952975:1	13856.44	Exon CDS 3UTR	TAAAAGTCTACTGACCATATTTATTTATTTATGTGAAGAATTTAATTTAATTATT-TAATATTTATA
Tnf	ENSMUSG00000024401	chr17:35336601-35336674:-1	9992.42	Exon 3UTR	CCTCTATTTATATTTGCACTTATTATTTATTATTTATTTATTATTTATTTATTTGCTTAT-GAATGTATTTATTT
Cxcl2	ENSMUSG00000058427	chr5:91334336-91334391:1	8489.92	Exon 3UTR	CTGTCTGAGAGTTCACTTATTTATTTATCTATGTATTTATTTATTTATTAATTCCA
Zfp36	ENSMUSG00000044786	chr7:29161949-29161990:-1	5202.36	Exon 3UTR	CTTTATTTATTGTATTAAGATTTTATAGTATTTATATATATT
Pfkfb3	ENSMUSG00000026773	chr2:11393671-11393751:-1	4305.38	Exon 3UTR	GATATTTTCATTTGTAATACTTGAAGTTTATTTTTTTATTATTTTGATAGCAGATGTGC-TATTTATTTATTTAATATGTAT
Ccl3	ENSMUSG00000000982	chr11:83461513-83461611:-1	1248.72	Exon 3UTR	TTCACTTGAAATTTTATTTAATTTAATCCTATTGGTTTAATACTATTTAATTTTG-TAATTTATTTTATTGTCATACTTG-TATTTGTGACTATTTATTCT
Cd274	ENSMUSG00000016496	chr19:29461644-29461708:1	1065.68	Exon 3UTR	TAAATGGTTGCTCACTATGCATTTTCTGTGCTCTTCGCCCTTTTTATTTAATGTATG-GATATTTA
Cflar	ENSMUSG00000026031	chr1:58814688-58814730:1	1030.35	Exon 3UTR	ATTGTATAATGTATATCATATTGTATATATTGTAATATATATA
Zfp36	ENSMUSG00000044786	chr7:29161998-29162007:-1	845.11	Exon 3UTR	CCCTTTATTT
Il12b	ENSMUSG00000004296	chr11:44226968-44227004:1	813.99	Exon 3UTR	TTGAAATATTTAAGTAATTTATGTATTTATTAATTTA

Table 28: **Top 10 TTP peaks 3 h after LPS induction and after normalization** by expression. Shown is the region of the peak, the normalized amount of signal in this peak region (PeakScore), the genomic partition and sequence of the peak.

Gene	ENSEMBL ID	Peak genomic region	PeakScore	Partition	Sequence
Tnf	ENSMUSG00000024401	chr17:35336601-35336670:-1	9715.51	Exon 3UTR	TATTTATATTTGCACTTATTATT TATTATTTATTTATTATTTAT-TTATTTGCTTATGAATGTATTTATTT
Ptgs2	ENSMUSG00000032487	chr1:151952948-151952975:1	8212.69	Exon 3UTR	AATTTAATTTAATTATTTAATAT TTATA
Cxcl2	ENSMUSG00000058427	chr5:91334347-91334391:1	4823.87	Exon 3UTR	TTCACTTATTTATTTATCTATGT ATTTATTTATTTATTAATTCCA
Zfp36	ENSMUSG00000044786	chr7:29161950-29161972:-1	4194.88	Exon 3UTR	GATTTTATAGTATTTATATATAT
Ptgs2	ENSMUSG00000032487	chr1:151952925-151952942:1	3167.09	Exon 3UTR	CATATTTATTTATTTATG
Zfp36	ENSMUSG00000044786	chr7:29161975-29161989:-1	1294.89	Exon 3UTR	TTTATTTATTGTATT
Ccl3	ENSMUSG00000000982	chr11:83461513-83461583:-1	1277.74	Exon 3UTR	CTATTGGTTTAATACTATTTAATT TTGTAATTTATTTTATTGT-CATACTTGTATTTGTGACTATTTATTCT
Il1ora	ENSMUSG00000032089	chr9:45062196-45062230:-1	1115.03	Exon 3UTR	GAACCTTTATTTATTTATTTGCTC ACTTATTTATT
49334-26M1-1Rik	ENSMUSG00000021133	chr12:81981749-81981773:1	993.48	Exon 3UTR	TTTAATTTAATTTTAATTGTTCTAT
Cflar	ENSMUSG00000026031	chr1:58814689-58814731:1	955.47	Exon 3UTR	TTGTATAATGTATATCATATTGTA TATATTGTAATATATATAA

Table 29: **Top 10 HuR peaks in WT after normalization** by expression. Shown is the region of the peak, the normalized amount of signal in this peak region (PeakScore), the genomic partition and sequence of the peak.

Gene	ENSEMBL ID	Peak genomic region	PeakScore	Partition	Sequence
Cd44	ENSMUSG00000005087	chr2:102653598-102653670:-1	4763.7	Exon 3UTR	TGTTCAGTATGACTTTTTTTTTTATTTTGTTTTATTTTGTTTTGTTTTGTTTTGTTTTTTTAGGTTACTTTAT
Lass6	ENSMUSG00000027035	chr2:68951785-68951970:1	4227.07	Exon 3UTR	ACTTTTTTTGGTTTTTTGTTTTGTTTGTTTGGGTTTTTTTTGTTTTTTGTTTTTTGTTTTTTGTTTTTTGTTTTTTTTTAGTGTTAGCTGCAAAGCTACAAAGCTCTGGAATGGTTACATTATGATTCTGGAACGTTCGCTCAAAGCTTAATTAGCATAAGTGTGGACCACACTCCAGCTTAATTC
Cd44	ENSMUSG00000005087	chr2:102654134-102654187:-1	2936.96	Exon 3UTR	GATTATGTTTAGCATAAAATTTTCTATTCCTTTTTTATTTTATGTCATTTTTTT
Lrrc58	ENSMUSG00000034158	chr16:37883337-37883363:1	2562.68	Exon 3UTR	GTCATCACTATTGGCAATGAGCGGTTC
Lrrc58	ENSMUSG00000034158	chr16:37883542-37883573:1	2021.78	Exon 3UTR	GATTACTGCTCTGGCTCCTAGCACCATGAAGA
Cd44	ENSMUSG00000005087	chr2:102654091-102654130:-1	1970.18	Exon 3UTR	GTCTGTTCCAATTTATGAAAATAGCATTGCTTTCTGAAAT
Marcks	ENSMUSG00000069662	chr10:36855542-36855658:-1	1786.11	Exon 3UTR	CCTTCTTTCTTTACTTTTACTTTTTTTTTTTTTTTTTTGGCATCAGTATTAATGTTTTTTGCATACTTTGCATCTTTATTAAAAAAGTGTAAACTTTCTTTGTCAGATCTATAGACAT
Cd44	ENSMUSG00000005087	chr2:102653452-102653500:-1	1614.01	Exon 3UTR	AAGTCTGAGTCTTTGTAGCACATCAGTGTGGCCTTAGTATGGTCCTCCT
Arf5	ENSMUSG00000020440	chr6:28376440-28376505:1	1362.49	Exon 3UTR	GGTTTTGGTTTTGGTTTTTTGATTTTTTTTTCTTTTTTTTCTTTCTTTTTTTTTTGTTTTGGGGGTA
Cd44	ENSMUSG00000005087	chr2:102653560-102653583:-1	1333.22	Exon 3UTR	GGTATAAATTGATTCATAATAAGT

Table 30: **Top 10 HuR peaks in TTP-KO after normalization** by expression. Shown is the region of the peak and the normalized amount of signal in this peak region (PeakScore).

Gene	ENSEMBL ID	Peak genomic region	PeakScore	Partition	Sequence
Ahnak	ENSMUSG00000069833	chr19:9093056-9093174:1	13924.17	Exon 3UTR	TTTTTTTTTTTTTTTTTGGTTTCTGTTTTCTTTTCTTTTTCCTTTTTCTT TCTTTCTTTTTTTTTTTGCATTGAACGCTGAGAGCTCGTGTTTACT AAGCAAGCTTTTGTGTTTATCTT
Lass6	ENSMUSG00000027035	chr2:68951767-68951876:1	10638.4	Exon 3UTR	AACTCTCCTTAATGTGAAACTTTTTTTGGTTTTTTGTTTTGTTTGTT TGGGTTTTTTTTGTTTTTTGTTTTTTGTTTTTTGTTTTTTGTTTTTTTT TAGTGTTAGCTGCA
Cd44	ENSMUSG00000005087	chr2:102653589-102653677:-1	8149.72	Exon 3UTR	GGTCCTTTGTTCAGTATGACTTTTTTTTTTATTTTGTTTTATTTTGTT TTGTTTTGTTTTGTTTTTTTAGGTTACTTTATCAGAAGCAT
Marcks	ENSMUSG00000069662	chr10:36855743-36855811:-1	5100.22	Exon 3UTR	ATTTTAAAGATTTTTTTTTAAATTTCACATTTTTTTT AAGCAGCAAATTTTTTGTTTTTTGTTTTTTTT
Calcrl	ENSMUSG00000059588	chr2:84171652-84171714:-1	4920.96	Exon 3UTR	ATTTGTAATTTTCTTTTTTTTTTTTTCTTTCTTTTTTTTTTTATATTTGT AATTTTCTGAACTTT
Maf	ENSMUSG00000055435	chr8:118229043-118229131:-1	4396.53	Exon 3UTR	GTGTGTTTGTGTGTTTTATTTTGTTTGGATTTTCTTTTTTCTTCTTCT TCTTCTTTGTTTTTGTTTATTATTACTATTACTAATTTTTT
Cflar	ENSMUSG00000026031	chr1:58813287-58813338:1	4315.1	Exon 3UTR	CTTTCTTTCTTTCTTTCTTTCTTTCTTTCTTTCTTTCTTTCTTTCCTT CCTT
Cd44	ENSMUSG00000005087	chr2:102654134-102654187:-1	3679.29	Exon 3UTR	GATTATGTTTAGCATAAAATTTTCTATTCCTTTTTTATTTTATGT CATTTTTTT
Pmp22	ENSMUSG00000018217	chr11:62972195-62972289:1	3452.05	Exon 3UTR	TCTTAAACAACTTTACATCCTAACACTATAACCAAGCTCAGT ATCTTTGTTTTGTTTCGTTTTTTTTTTTAATCTTTTTGTTTTGCT CAGACATA
Dpysl2	ENSMUSG00000022048	chr14:67423346-67423377:-1	3299.97	Exon 3UTR	TTTTTTTTTTTTTTTTTTTTTTTTTTCCTCTT

A.2.5 *GO-term analysis*

This section presents results from GO-term enrichment of TTP and HuR target genes analyzed with TopGO and DAVID.

Table 31: **GO-term enrichment for HuR target genes** , and genes containing exclusively 3'UTR or intronic peak regions, analyzed with TopGO and expressed mouse BMDM genes as background

GO-ID	GO-Term	Annotated	Significant	Expected	Rank	weight
		HuR				
GO:0003779	actin binding	27	14	6.52	1	0.0014
GO:0005102	receptor binding	63	25	15.2	2	0.0030
GO:0005080	protein kinase C binding	4	4	0.97	3	0.0033
GO:0051015	actin filament binding	4	4	0.97	4	0.0033
GO:0008092	cytoskeletal protein binding	38	17	9.17	5	0.0034
GO:0005178	integrin binding	9	6	2.17	6	0.0079
GO:0001653	peptide receptor activity	3	3	0.72	7	0.0139
GO:0008528	G-protein coupled peptide receptor activ...	3	3	0.72	8	0.0139
GO:0016597	amino acid binding	3	3	0.72	9	0.0139
GO:0016887	ATPase activity	18	9	4.34	10	0.0142
		HuR 3UTR				
GO:0003779	actin binding	24	14	6.88	1	0.0018
GO:0038023	signaling receptor activity	15	10	4.3	2	0.0022
GO:0008092	cytoskeletal protein binding	32	17	9.18	3	0.0023
GO:0005080	protein kinase C binding	4	4	1.15	4	0.0066
GO:0051015	actin filament binding	4	4	1.15	5	0.0066
GO:0005102	receptor binding	57	25	16.35	6	0.0069
GO:0005178	integrin binding	8	6	2.29	7	0.0084
GO:0004871	signal transducer activity	33	16	9.47	8	0.0102
GO:0060089	molecular transducer activity	33	16	9.47	9	0.0102
GO:0001653	peptide receptor activity	3	3	0.86	10	0.0233
		HuR Introns				
GO:0036094	small molecule binding	58	9	3	1	0.00036
GO:0097367	carbohydrate derivative binding	54	8	2.8	2	0.00152
GO:0000166	nucleotide binding	55	8	2.85	3	0.00174
GO:1901265	nucleoside phosphate binding	55	8	2.85	4	0.00174
GO:0005516	calmodulin binding	2	2	0.1	5	0.00249
GO:0097159	organic cyclic compound binding	91	10	4.71	6	0.00270
GO:1901363	heterocyclic compound binding	91	10	4.71	7	0.00270
GO:0032553	ribonucleotide binding	46	6	2.38	8	0.01712
GO:0001882	nucleoside binding	47	6	2.43	9	0.01908
GO:0008026	ATP-dependent helicase activity	6	2	0.31	10	0.03309

Table 32: **GO-term enrichment for HuR target genes in TTP-KO**, and genes containing exclusively 3'UTR or intronic peak regions, analyzed with TopGO and expressed mouse BMDM genes as background

GO-ID	GO-Term	Annotated	Significant	Expected	Rank	weight
			HuR in TTP-KO			
GO:0005102	receptor binding	66	32	20.72	1	0.0018
GO:0042623	ATPase activity, coupled	20	12	6.28	2	0.0070
GO:0016887	ATPase activity	23	13	7.22	3	0.0099
GO:0016209	antioxidant activity	6	5	1.88	4	0.0132
GO:0008026	ATP-dependent helicase activity	13	8	4.08	5	0.0231
GO:0070035	purine NTP-dependent helicase activity	13	8	4.08	6	0.0231
GO:0005539	glycosaminoglycan binding	11	7	3.45	7	0.0269
GO:0001653	peptide receptor activity	3	3	0.94	8	0.0307
GO:0008528	G-protein coupled peptide receptor activ...	3	3	0.94	9	0.0307
GO:0016597	amino acid binding	3	3	0.94	10	0.0307
			HuR in TTP-KO 3UTR			
GO:0038023	signaling receptor activity	12	9	4.14	1	0.0046
GO:0005080	protein kinase C binding	4	4	1.38	2	0.0139
GO:0005102	receptor binding	56	27	19.33	3	0.0173
GO:0016209	antioxidant activity	6	5	2.07	4	0.0203
GO:0008092	cytoskeletal protein binding	31	16	10.7	5	0.0326
GO:0003779	actin binding	24	13	8.29	6	0.0341
GO:0001653	peptide receptor activity	3	3	1.04	7	0.0407
GO:0004930	G-protein coupled receptor activity	3	3	1.04	8	0.0407
GO:0008528	G-protein coupled peptide receptor activ...	3	3	1.04	9	0.0407
GO:0005515	protein binding	337	125	116.35	10	0.0407
			HuR in TTP-KO Introns			
GO:0003950	NAD+ ADP-ribosyltransferase activity	5	3	0.67	1	0.018
GO:0008026	ATP-dependent helicase activity	9	4	1.2	2	0.021
GO:0070035	purine NTP-dependent helicase activity	9	4	1.2	3	0.021
GO:0004386	helicase activity	11	4	1.47	4	0.045
GO:0042623	ATPase activity, coupled	11	4	1.47	5	0.045
GO:0004693	cyclin-dependent protein serine/threonin...	3	2	0.4	6	0.048
GO:0005516	calmodulin binding	3	2	0.4	7	0.048
GO:0030332	cyclin binding	3	2	0.4	8	0.048
GO:0097472	cyclin-dependent protein kinase activity	3	2	0.4	9	0.048
GO:0016763	transferase activity, transferring pento...	7	3	0.93	10	0.053

Table 33: **TTP 6h target genes GO-terms** analyzed with DAVID and full mouse geneset as background

Category	Term	Count	%	PValue	Bonferroni	FDR
Cluster 1	Enrichment Score: 35.1501					
GOTERM_CC_FAT	GO:0031981 nuclear lumen	342	14.7034	6.0514E-48	3.8850E-45	9.0203E-45
GOTERM_CC_FAT	GO:0070013 intracellular organelle lumen	381	16.3801	3.5805E-43	2.2987E-40	5.3371E-40
GOTERM_CC_FAT	GO:0043233 organelle lumen	384	16.5090	6.2427E-42	4.0078E-39	9.3054E-39
GOTERM_CC_FAT	GO:0031974 membrane-enclosed lumen	389	16.7240	7.7691E-42	4.9878E-39	1.1581E-38
GOTERM_CC_FAT	GO:0005654 nucleoplasm	220	9.4583	1.7533E-33	1.1256E-30	2.6135E-30
GOTERM_CC_FAT	GO:0044451 nucleoplasm part	142	6.1049	1.4309E-22	9.1861E-20	2.1328E-19
GOTERM_CC_FAT	GO:0005730 nucleolus	160	6.8788	3.3727E-20	2.1653E-17	5.0273E-17
Annotation Cluster 2	Enrichment Score: 27.9342					
SP_PIR_KEYWORDS	nucleus	913	39.2519	8.2109E-91	4.8198E-88	1.2090E-87
SP_PIR_KEYWORDS	Transcription	433	18.6156	3.6657E-35	2.1518E-32	5.3977E-32
SP_PIR_KEYWORDS	transcription regulation	425	18.2717	8.3034E-35	4.8741E-32	1.2227E-31
GOTERM_MF_FAT	GO:0003677 DNA binding	473	20.3353	8.5738E-31	1.0271E-27	1.3858E-27
GOTERM_BP_FAT	GO:0045449 regulation of transcription	513	22.0550	2.7688E-29	1.0411E-25	5.1110E-26
GOTERM_BP_FAT	GO:0006350 transcription	435	18.7016	6.7254E-29	2.5287E-25	1.2415E-25
SP_PIR_KEYWORDS	dna-binding	380	16.3371	4.3548E-28	2.5563E-25	6.4124E-25
GOTERM_BP_FAT	GO:0051252 regulation of RNA metabolic process	357	15.3482	2.3148E-19	8.7038E-16	4.2730E-16
GOTERM_BP_FAT	GO:0006355 regulation of transcription, DNA-dependent	345	14.8323	7.6448E-18	2.8744E-14	1.4112E-14
GOTERM_MF_FAT	GO:0030528 transcription regulator activity	284	12.2098	7.9259E-13	9.4952E-10	1.2811E-9
GOTERM_MF_FAT	GO:0003700 transcription factor activity	186	7.9966	4.5876E-9	5.4959E-6	7.4151E-6
GOTERM_MF_FAT	GO:0043565 sequence-specific DNA binding	120	5.1591	5.5021E-7	6.5894E-4	8.8932E-4
Annotation Cluster 3	Enrichment Score: 14.970696834067018					
GOTERM_CC_FAT	GO:0043228 non-membrane-bounded organelle	431	18.5297	3.0612E-22	1.9653E-19	4.5631E-19
GOTERM_CC_FAT	GO:0043232 intracellular non-membrane-bounded organelle	431	18.5297	3.0612E-22	1.9653E-19	4.5631E-19
GOTERM_CC_FAT	GO:0005856 cytoskeleton	179	7.6956	0.0131	0.9998	17.8015

Table 34: **TTP 6h 3'UTR target genes GO-terms** analyzed with DAVID and full mouse geneset as background

Category	Term	Count	%	PValue	Bonferroni	FDR
Annotation Cluster 1	Enrichment Score: 8.7706					
SP_PIR_KEYWORDS	cytokine	23	11.5578	1.6761E-17	4.2740E-15	2.1842E-14
GOTERM_MF_FAT	GO:0005125 cytokine activity	23	11.5578	4.4607E-17	1.4185E-14	6.0138E-14
GOTERM_BP_FAT	GO:0006955 immune response	32	16.0804	8.4421E-16	1.3438E-12	1.4766E-12
GOTERM_BP_FAT	GO:0009611 response to wounding	26	13.0653	9.4676E-14	1.4328E-10	1.5752E-10
GOTERM_CC_FAT	GO:0005615 extracellular space	29	14.5729	6.3711E-13	1.2042E-10	7.9121E-10
GOTERM_BP_FAT	GO:0006954 inflammatory response	20	10.0503	7.0282E-12	1.0634E-8	1.1690E-8
KEGG_PATHWAY	mmu04060:Cytokine-cytokine receptor interaction	22	11.0553	1.3056E-10	1.2665E-8	1.4409E-7
GOTERM_CC_FAT	GO:0044421 extracellular region part	31	15.5779	4.9639E-10	9.3818E-8	6.1641E-7
GOTERM_BP_FAT	GO:0006952 defense response	24	12.0603	9.4373E-10	1.4279E-6	1.5697E-6
GOTERM_BP_FAT	GO:0006935 chemotaxis	13	6.5327	3.0606E-9	4.6306E-6	5.0906E-6
GOTERM_BP_FAT	GO:0042330 taxis	13	6.5327	3.0606E-9	4.6306E-6	5.0906E-6
INTERPRO	IPR001811:Small chemokine, interleukin-8-like	9	4.5226	3.7019E-9	1.6770E-6	5.2568E-6
GOTERM_MF_FAT	GO:0008009 chemokine activity	9	4.5226	7.2196E-9	2.2958E-6	9.7332E-6
GOTERM_MF_FAT	GO:0042379 chemokine receptor binding	9	4.5226	8.9977E-9	2.8613E-6	1.2130E-5
SP_PIR_KEYWORDS	inflammatory response	11	5.5276	9.8773E-9	2.5187E-6	1.2872E-5
SP_PIR_KEYWORDS	chemotaxis	10	5.0251	1.8066E-8	4.6069E-6	2.3543E-5
SMART	SM00199:SCY	9	4.5226	3.0890E-8	3.5833E-6	3.5242E-5
KEGG_PATHWAY	mmu04621:NOD-like receptor signaling pathway	11	5.5276	3.7055E-8	3.5943E-6	4.0892E-5
GOTERM_BP_FAT	GO:0007626 locomotory behavior	15	7.5377	4.6897E-7	7.0930E-4	7.8002E-4
KEGG_PATHWAY	mmu04062:Chemokine signaling pathway	15	7.5377	9.1421E-7	8.8674E-5	0.0010
INTERPRO	IPR000827:Small chemokine, C-C group, conserved site	6	3.0151	2.1186E-6	9.5926E-4	0.0030
PIR_SUPERFAMILY	PIRSF001950:small inducible chemokine, C/CC types	6	3.0151	6.4525E-6	8.1914E-4	0.0075
GOTERM_CC_FAT	GO:0005576 extracellular region	36	18.0905	6.1890E-5	0.0116	0.0768
GOTERM_BP_FAT	GO:0007610 behavior	15	7.5377	1.8074E-4	0.2393	0.3002
SP_PIR_KEYWORDS	Secreted	29	14.5729	0.0011	0.2454	1.4286

Table 35: **T** TP 3h target genes GO-terms analyzed with DAVID and full mouse geneset as background

Category	Term	Count	%	PValue	Bonferroni	FDR
Annotation Cluster 1	Enrichment Score: 5.2476					
SP_PIR_KEYWORDS	SH2 domain	14	3.0635	1.1175E-6	3.3631E-4	0.0015
INTERPRO	IPR000980:SH2 motif	14	3.0635	2.7853E-6	0.0023	0.0043
UP_SEQ_FEATURE	domain:SH2	13	2.8446	3.8892E-6	0.0053	0.0064
SMART	SM00252:SH2	14	3.0635	8.4482E-5	0.0173	0.1064
Annotation Cluster 2	Enrichment Score: 4.6457					
GOTERM_BP_FAT	GO:0006952 defense response	37	8.0963	4.3853E-10	8.8847E-7	7.5511E-7
SP_PIR_KEYWORDS	cytokine	19	4.1575	2.7947E-7	8.4117E-5	3.7364E-4
GOTERM_BP_FAT	GO:0009611 response to wounding	27	5.9081	5.0732E-7	0.0010	8.7356E-4
GOTERM_MF_FAT	GO:0005125 cytokine activity	19	4.1575	5.6524E-7	2.7467E-4	8.1071E-4
KEGG_PATHWAY	mmu04060:Cytokine-cytokine receptor interaction	24	5.2516	1.0562E-6	1.3519E-4	0.0012
SP_PIR_KEYWORDS	inflammatory response	12	2.6258	2.4818E-6	7.4673E-4	0.0033
GOTERM_BP_FAT	GO:0006954 inflammatory response	20	4.3764	3.2306E-6	0.0065	0.0056
GOTERM_CC_FAT	GO:0005615 extracellular space	24	5.2516	7.0445E-4	0.1662	0.9158
GOTERM_CC_FAT	GO:0044421 extracellular region part	28	6.1269	0.0094	0.9117	11.5593
GOTERM_CC_FAT	GO:0005576 extracellular region	38	8.3151	0.4755	1.0000	99.9781
SP_PIR_KEYWORDS	Secreted	30	6.5646	0.8454	1.0000	100.0000
Annotation Cluster 3	Enrichment Score: 4.5479					
GOTERM_BP_FAT	GO:0010324 membrane invagination	18	3.9387	4.2362E-6	0.0085	0.0073
GOTERM_BP_FAT	GO:0006897 endocytosis	18	3.9387	4.2362E-6	0.0085	0.0073
GOTERM_BP_FAT	GO:0016044 membrane organization	19	4.1575	1.4898E-4	0.2606	0.2562
GOTERM_BP_FAT	GO:0016192 vesicle-mediated transport	26	5.6893	2.4057E-4	0.3858	0.4134

Category	Term	Count	%	PValue	Bonferroni	FDR
Annotation Cluster 1	Enrichment Score: 6.1927					
GOTERM_MF_FAT	GO:0005125 cytokine activity	19	11.6564	9.3743E-15	2.7232E-12	1.2412E-11
SP_PIR_KEYWORDS	cytokine	19	11.6564	1.6477E-14	3.7628E-12	2.1050E-11
GOTERM_BP_FAT	GO:0006955 immune response	27	16.5644	9.3658E-14	1.1994E-10	1.5270E-10
GOTERM_BP_FAT	GO:0009611 response to wounding	20	12.2699	3.3334E-10	4.2668E-7	5.4323E-7
GOTERM_BP_FAT	GO:0006952 defense response	22	13.4969	6.1634E-10	7.8891E-7	1.0044E-6
KEGG_PATHWAY	mmu04060:Cytokine-cytokine receptor interaction	19	11.6564	7.5101E-10	5.7828E-8	7.9236E-7
GOTERM_BP_FAT	GO:0006954 inflammatory response	16	9.8160	1.9565E-9	2.5043E-6	3.1884E-6
GOTERM_CC_FAT	GO:0005615 extracellular space	21	12.8834	2.9808E-9	4.2030E-7	3.5215E-6
SP_PIR_KEYWORDS	inflammatory response	10	6.1350	2.4274E-8	5.5588E-6	3.1099E-5
GOTERM_MF_FAT	GO:0008009 chemokine activity	8	4.9080	2.5446E-8	7.4301E-6	3.3863E-5
INTERPRO	IPR001811:Small chemokine, interleukin-8-like	8	4.9080	2.6576E-8	9.7533E-6	3.6603E-5
GOTERM_MF_FAT	GO:0042379 chemokine receptor binding	8	4.9080	3.0798E-8	8.9929E-6	4.0986E-5
GOTERM_CC_FAT	GO:0044421 extracellular region part	23	14.1104	1.4239E-7	2.0077E-5	1.6822E-4
SMART	SM00199:SCY	8	4.9080	1.9198E-7	1.7086E-5	2.0840E-4
SP_PIR_KEYWORDS	chemotaxis	7	4.2945	1.5994E-5	0.0037	0.0205
INTERPRO	IPR000827:Small chemokine, C-C group, conserved site	5	3.0675	2.8988E-5	0.0106	0.0399
PIR_SUPERFAMILY	PIRSF001950:small inducible chemokine, C/CC types	5	3.0675	6.0720E-5	0.0061	0.0675
GOTERM_BP_FAT	GO:0042330 taxis	8	4.9080	6.1333E-5	0.0755	0.0999
GOTERM_BP_FAT	GO:0006935 chemotaxis	8	4.9080	6.1333E-5	0.0755	0.0999
GOTERM_CC_FAT	GO:0005576 extracellular region	27	16.5644	5.6237E-4	0.0763	0.6624
GOTERM_BP_FAT	GO:0007626 locomotory behavior	9	5.5215	0.0016	0.8639	2.5067
KEGG_PATHWAY	mmu04062:Chemokine signaling pathway	9	5.5215	0.0020	0.1446	2.1171
INTERPRO	IPR002473:Small chemokine, C-X-C/Interleukin 8	3	1.8405	0.0039	0.7618	5.2421
INTERPRO	IPR001089:Small chemokine, C-X-C	3	1.8405	0.0039	0.7618	5.2421
SP_PIR_KEYWORDS	Secreted	23	14.1104	0.0056	0.7256	6.9798
INTERPRO	IPR018048:Small chemokine, C-X-C, conserved site	3	1.8405	0.0063	0.9033	8.3956
PIR_SUPERFAMILY	PIRSF002522:CXC chemokine	3	1.8405	0.0068	0.4982	7.3115
GOTERM_BP_FAT	GO:0007610 behavior	9	5.5215	0.0321	1.0000	41.2178

Table 37: **TTP mouse-human orthologous target genes GO-terms** analyzed with DAVID and full mouse geneset as background

Category	Term	Count	%	PValue	Bonferroni	FDR
Annotation Cluster 1	Enrichment Score: 9.6639					
GOTERM_BP_FAT	GO:0042981 regulation of apoptosis	54	12.3288	1.6257E-10	3.9780E-7	2.8608E-7
GOTERM_BP_FAT	GO:0043067 regulation of programmed cell death	54	12.3288	2.3440E-10	5.7358E-7	4.1250E-7
GOTERM_BP_FAT	GO:0010941 regulation of cell death	54	12.3288	2.6753E-10	6.5465E-7	4.7080E-7
Annotation Cluster 2	Enrichment Score: 7.4952					
GOTERM_BP_FAT	GO:0006916 anti-apoptosis	23	5.2511	1.4956E-8	3.6597E-5	2.6320E-5
GOTERM_BP_FAT	GO:0043066 negative regulation of apoptosis	30	6.8493	3.2844E-8	8.0365E-5	5.7798E-5
GOTERM_BP_FAT	GO:0043069 negative regulation of programmed cell death	30	6.8493	4.4837E-8	1.0971E-4	7.8904E-5
GOTERM_BP_FAT	GO:0060548 negative regulation of cell death	30	6.8493	4.7444E-8	1.1609E-4	8.3492E-5
Annotation Cluster 3	Enrichment Score: 5.6560					
GOTERM_BP_FAT	GO:0043065 positive regulation of apoptosis	30	6.8493	1.9526E-6	0.0048	0.0034
GOTERM_BP_FAT	GO:0043068 positive regulation of programmed cell death	30	6.8493	2.2443E-6	0.0055	0.0039
GOTERM_BP_FAT	GO:0010942 positive regulation of cell death	30	6.8493	2.4571E-6	0.0060	0.0043

BIBLIOGRAPHY

[1] A. S. Abu Almakarem, A. I. Petrov, J. Stombaugh, C. L. Zirbel, and N. B. Leontis. Comprehensive survey and geometric classification of base triples in RNA structures. *Nucleic Acids Research*, 40(4):1407–1423, February 2012. ISSN 0305-1048, 1362-4962. doi: 10.1093/nar/gkr810. URL http://nar.oxfordjournals.org/lookup/doi/10.1093/nar/gkr810.

[2] Adrian Alexa and Jorg Rahnenfuhrer. *topGO: topGO: Enrichment analysis for Gene Ontology*, 2010. R package version 2.16.0.

[3] Sonja Althammer, Juan González-vallinas, Cecilia Ballaré, Miguel Beato, and Eduardo Eyras. Pyicos: a versatile toolkit for the analysis of high-throughput sequencing data. *Bioinformatics (Oxford, England)*, 27(24):3333–40, December 2011. ISSN 1367-4811. doi: 10.1093/bioinformatics/btr570. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3232367&tool=pmcentrez&rendertype=abstract.

[4] Stephen F. Altschul, Warren Gish, Webb Miller, Eugene W. Myers, and David J. Lipman. Basic local alignment search tool. *Journal of molecular biology*, 215(3):403–410, 1990. URL http://www.sciencedirect.com/science/article/pii/S0022283605803602.

[5] Simon Anders and Wolfgang Huber. Differential expression analysis for sequence count data. *Genome biology*, 11(10):R106, January 2010. ISSN 1465-6914. doi: 10.1186/gb-2010-11-10-r106. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3218662&tool=pmcentrez&rendertype=abstracthttp://genomebiology.com/2010/11/10/R106.

[6] Peter Audano and Fredrik Vannberg. KAnalyze: a fast versatile pipelined K-mer toolkit. *Bioinformatics (Oxford, England)*, 30(14):2070–2, July 2014. ISSN 1367-4811. doi: 10.1093/bioinformatics/btu152. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4080738&tool=pmcentrez&rendertype=abstract.

[7] S. D. Auweter, F. C. Oberstrass, and F. H.-T. Allain. Sequence-specific binding of single-stranded RNA: is there a code for recognition? *Nucleic Acids Research*, 34(17):4943–4959, September 2006. ISSN 0305-1048, 1362-4962. doi: 10.1093/nar/gkl620. URL http://nar.oxfordjournals.org/lookup/doi/10.1093/nar/gkl620.

[8] T L Bailey and C Elkan. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proceedings / ... International Conference on Intelligent Systems for Molecular Biology ; ISMB. International Conference on Intelligent Systems for Molecular Biology*, 2:28–36, January 1994. ISSN 1553-0833. URL http://www.ncbi.nlm.nih.gov/pubmed/7584402.

[9] Richard H. Baltz. Combinatorial biosynthesis of cyclic lipopeptide antibiotics: a model for synthetic biology to accelerate the evolution of secondary metabolite biosynthetic pathways. *ACS Synthetic Biology*, page 120809123853000, August 2012. ISSN 2161-5063. doi: 10.1021/sb3000673. URL http://dx.doi.org/10.1021/sb3000673http://pubs.acs.org/doi/abs/10.1021/sb3000673.

[10] Marcus Bantscheff, Markus Schirle, Gavain Sweetman, Jens Rick, and Bernhard Kuster. Quantitative mass spectrometry in proteomics: a critical review. *Analytical and Bioanalytical Chemistry*, 389(4):1017–1031, September 2007. ISSN 1618-2642, 1618-2650. doi: 10.1007/s00216-007-1486-6. URL http://link.springer.com/10.1007/s00216-007-1486-6.

[11] Suying Bao, Rui Jiang, WingKeung Kwan, BinBin Wang, Xu Ma, and You-Qiang Song. Evaluation of next-generation sequencing software in mapping and assembly. *Journal of human genetics*, 56 (6):406–414, 2011. URL http://www.nature.com/jhg/journal/v56/n6/abs/jhg201143a.html.

[12] Andrew Barker, Michael R Epis, Corrine J Porter, Benjamin R Hopkins, Matthew C J Wilce, Jackie a Wilce, Keith M Giles, and Peter J Leedman. Sequence requirements for RNA binding by HuR and AUF1. *Journal of biochemistry*, 151(4):423–37, April 2012. ISSN 1756-2651. doi: 10.1093/jb/mvs010. URL http://www.ncbi.nlm.nih.gov/pubmed/22368252.

[13] Carine Barreau, Luc Paillard, and H.B. Beverley Osborne. AU-rich elements and associated factors: are there unifying principles? *Nucleic acids research*, 33(22):7138–50, January 2005. ISSN 1362-4962. doi: 10.1093/nar/gki1012. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1325018&tool=pmcentrez&rendertype=abstracthttp://nar.oxfordjournals.org/content/33/22/7138.short.

[14] Stephan Bernhart. *Variations on RNA folding - Locally stable structures and RNA hybridization*. PhD thesis, 2007.

[15] Stephan H Bernhart, Ivo L Hofacker, and Peter F Stadler. Local RNA base pairing probabilities in large sequences. *Bioinformatics (Oxford, England)*, 22(5):614–5, March 2006. ISSN 1367-

4803. doi: 10.1093/bioinformatics/btk014. URL http://www.ncbi.nlm.nih.gov/pubmed/16368769.

[16] K. Blin, C. Dieterich, R. Wurmus, N. Rajewsky, M. Landthaler, and A. Akalin. DoRiNA 2.0–upgrading the doRiNA database of RNA interactions in post-transcriptional regulation. *Nucleic Acids Research*, 43(D1):D160–D167, January 2015. ISSN 0305-1048, 1362-4962. doi: 10.1093/nar/gku1180. URL http://nar.oxfordjournals.org/lookup/doi/10.1093/nar/gku1180.

[17] Seth a Brooks and Perry J Blackshear. Tristetraprolin (TTP): interactions with mRNA and proteins, and current thoughts on mechanisms of action. *Biochimica et biophysica acta*, 1829(6-7): 666–79, 2013. ISSN 0006-3002. doi: 10.1016/j.bbagrm.2013.02.003. URL http://www.ncbi.nlm.nih.gov/pubmed/23428348.

[18] Matthias Bros, Nadine Wiechmann, Verena Besche, Julia Art, Andrea Pautz, Stephan Grabbe, Hartmut Kleinert, and Angelika B Reske-Kunz. The RNA binding protein tristetraprolin influences the activation state of murine dendritic cells. *Molecular immunology*, 47(5):1161–70, February 2010. ISSN 1872-9142. doi: 10.1016/j.molimm.2009.11.002. URL http://www.ncbi.nlm.nih.gov/pubmed/19945750.

[19] Fatima Cairrao, Anason S Halees, Khalid S a Khabar, Dominique Morello, and Nathalie Vanzo. AU-rich elements regulate Drosophila gene expression. *Molecular and cellular biology*, 29(10):2636–43, May 2009. ISSN 1098-5549. doi: 10.1128/MCB.01506-08. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2682044&tool=pmcentrez&rendertype=abstract.

[20] Jannette Carey, Vicki Cameron, Pieter L. De Haseth, and Olke C. Uhlenbeck. Sequence-specific interaction of R17 coat protein with its ribonucleic acid binding site. *Biochemistry*, 22 (11):2601–2610, 1983. URL http://pubs.acs.org/doi/abs/10.1021/bi00280a002.

[21] Alfredo Castello, Bernd Fischer, Katrin Eichelbaum, Rastislav Horos, Benedikt M. Beckmann, Claudia Strein, Norman E. Davey, David T. Humphreys, Thomas Preiss, Lars M. Steinmetz, Jeroen Krijgsveld, and Matthias W. Hentze. Insights into RNA Biology from an Atlas of Mammalian mRNA-Binding Proteins. *Cell*, 149(6):1393–1406, June 2012. ISSN 00928674. doi: 10.1016/j.cell.2012.04.031. URL http://linkinghub.elsevier.com/retrieve/pii/S0092867412005764.

[22] Sung-Hee Chang, Yi-Chien Lu, Xi Li, Wan-Ying Hsieh, Yuquan Xiong, Mallika Ghosh, Todd Evans, Olivier Elemento, and

Timothy Hla. Antagonistic function of the RNA-binding protein HuR and miR-200b in post-transcriptional regulation of vascular endothelial growth factor-A expression and angiogenesis. *The Journal of biological chemistry*, 288(7):4908–21, February 2013. ISSN 1083-351X. doi: 10.1074/jbc.M112.423871. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3576095&tool=pmcentrez&rendertype=abstract.

[23] Cheom-Gil Cheong and Traci M. Tanaka Hall. Engineering RNA sequence specificity of Pumilio repeats. *Proceedings of the National Academy of Sciences*, 103(37):13635–13639, 2006. URL http://www.pnas.org/content/103/37/13635.short.

[24] Antoine Cléry, Markus Blatter, and Frédéric H-T Allain. RNA recognition motifs: boring? Not quite. *Current Opinion in Structural Biology*, 18(3):290–298, June 2008. ISSN 0959440X. doi: 10.1016/j.sbi.2008.04.002. URL http://linkinghub.elsevier.com/retrieve/pii/S0959440X08000584.

[25] Federico Comoglio, Cem Sievers, and Renato Paro. Sensitive and highly resolved identification of RNA-protein interaction sites in PAR-CLIP data. *BMC Bioinformatics*, 16(1), December 2015. ISSN 1471-2105. doi: 10.1186/s12859-015-0470-y. URL http://www.biomedcentral.com/1471-2105/16/32.

[26] Ana Conesa, Pedro Madrigal, Sonia Tarazona, David Gomez-Cabrero, Alejandra Cervera, Andrew McPherson, Michał Wojciech Szcześniak, Daniel J. Gaffney, Laura L. Elo, Xuegong Zhang, and Ali Mortazavi. A survey of best practices for RNA-seq data analysis. *Genome Biology*, 17(1), December 2016. ISSN 1474-760X. doi: 10.1186/s13059-016-0881-8. URL http://genomebiology.com/2016/17/1/13.

[27] ENCODE Project Consortium and others. The ENCODE (ENCyclopedia of DNA elements) project. *Science*, 306(5696):636–640, 2004. URL http://science.sciencemag.org/content/306/5696/636.short.

[28] The 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature*, 526(7571):68–74, September 2015. ISSN 0028-0836, 1476-4687. doi: 10.1038/nature15393. URL http://www.nature.com/doifinder/10.1038/nature15393.

[29] K. B. Cook, T. R. Hughes, and Q. D. Morris. High-throughput characterization of protein-RNA interactions. *Briefings in Functional Genomics*, 14(1):74–89, January 2015. ISSN 2041-2649, 2041-2657. doi: 10.1093/bfgp/elu047. URL http://bfgp.oxfordjournals.org/cgi/doi/10.1093/bfgp/elu047.

[30] Kate B Cook, Hilal Kazan, Khalid Zuberi, Quaid Morris, and Timothy R Hughes. RBPDB: a database of RNA-binding specificities. *Nucleic acids research*, 39(Database issue):D301–8, January 2011. ISSN 1362-4962. doi: 10.1093/nar/gkq1069. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3013675&tool=pmcentrez&rendertype=abstract.

[31] Shannon Copeland, H Shaw Warren, Stephen F Lowry, Steve E Calvano, and Daniel Remick. Acute Inflammatory Response to Endotoxin in Mice and Humans Acute Inflammatory Response to Endotoxin in Mice and Humans. *Clinical and Diagnostic Laboratory Immunology*, 12(1):60–67, 2005. doi: 10.1128/CDLI.12.1.60.

[32] David L Corcoran, Stoyan Georgiev, Neelanjan Mukherjee, Eva Gottwein, Rebecca L Skalsky, Jack D Keene, and Uwe Ohler. PARalyzer: Definition of RNA binding sites from PAR-CLIP short-read sequence data. *Genome Biology*, 12(8):R79, 2011. ISSN 1465-6906. doi: 10.1186/gb-2011-12-8-r79. URL http://genomebiology.com/2011/12/8/R79.

[33] Fiona Cunningham, M. Ridwan Amode, Daniel Barrell, Kathryn Beal, Konstantinos Billis, Simon Brent, Denise Carvalho-Silva, Peter Clapham, Guy Coates, Stephen Fitzgerald, Laurent Gil, Carlos García Girón, Leo Gordon, Thibaut Hourlier, Sarah E. Hunt, Sophie H. Janacek, Nathan Johnson, Thomas Juettemann, Andreas K Kähäri, Stephen Keenan, Fergal J. Martin, Thomas Maurel, William McLaren, Daniel N. Murphy, Rishi Nag, Bert Overduin, Anne Parker, Mateus Patricio, Emily Perry, Miguel Pignatelli, Harpreet Singh Riat, Daniel Sheppard, Kieron Taylor, Anja Thormann, Alessandro Vullo, Steven P. Wilder, Amonida Zadissa, Bronwen L. Aken, Ewan Birney, Jennifer Harrow, Rhoda Kinsella, Matthieu Muffato, Magali Ruffier, Stephen M. J. Searle, Giulietta Spudich, Stephen J. Trevanion, Andy Yates, Daniel R. Zerbino, and Paul Flicek. Ensembl 2015. *Nucleic Acids Research*, 43(D1):D662–D669, January 2014. ISSN 0305-1048, 1362-4962. doi: 10.1093/nar/gku1010. URL http://nar.oxfordjournals.org/lookup/doi/10.1093/nar/gku1010http://nar.oxfordjournals.org/content/43/D1/D662.full.pdf.

[34] Jörg Fallmann, Vitaly Sedlyarov, Andrea Tanzer, Pavel Kovarik, and Ivo L. Hofacker. AREsite2: an enhanced database for the comprehensive investigation of AU/GU/U-rich elements. *Nucleic Acids Research*, 44(D1):D90–D95, January 2016. ISSN 0305-1048, 1362-4962. doi: 10.1093/nar/gkv1238. URL http://nar.oxfordjournals.org/lookup/doi/10.1093/nar/gkv1238.

[35] Barrett C Foat and Gary D Stormo. Discovering structural cis-regulatory elements by modeling the behaviors of mRNAs.

Molecular Systems Biology, 5, April 2009. ISSN 1744-4292. doi: 10.1038/msb.2009.24. URL http://msb.embopress.org/cgi/doi/10.1038/msb.2009.24.

[36] Nicole L. Garneau, Jeffrey Wilusz, and Carol J. Wilusz. The highways and byways of mRNA decay. *Nature Reviews Molecular Cell Biology*, 8(2):113–126, February 2007. ISSN 1471-0072, 1471-0080. doi: 10.1038/nrm2104. URL http://www.nature.com/doifinder/10.1038/nrm2104.

[37] Stoyan Georgiev, Alan P. Boyle, Karthik Jayasurya, Xuan Ding, Sayan Mukherjee, Uwe Ohler, and others. Evidence-ranked motif identification. *Genome Biol*, 11(2):R19, 2010. URL http://www.biomedcentral.com/content/pdf/gb-2010-11-2-r19.pdf.

[38] Stefanie Gerstberger, Markus Hafner, and Thomas Tuschl. A census of human RNA-binding proteins. *Nature Reviews Genetics*, 15(12):829–845, November 2014. ISSN 1471-0056, 1471-0064. doi: 10.1038/nrg3813. URL http://www.nature.com/doifinder/10.1038/nrg3813.

[39] Timothy J T.J. Gingerich, Jean-Jacques J.J. Feige, and Jonathan LaMarre. AU-rich elements and the control of gene expression through regulated mRNA stability. *Animal health research reviews Conference of Research Workers in Animal Diseases*, 5(1):49–63, 2004. doi: 10.1079/AHRR200460. URL http://journals.cambridge.org/production/action/cjoGetFulltext?fulltextid=766260.

[40] Hani Goodarzi, Hamed S. Najafabadi, Panos Oikonomou, Todd M. Greco, Lisa Fish, Reza Salavati, Ileana M. Cristea, and Saeed Tavazoie. Systematic discovery of structural elements governing stability of mammalian messenger RNAs. *Nature*, April 2012. ISSN 0028-0836. doi: 10.1038/nature11013. URL http://dx.doi.org/10.1038/nature11013http://www.nature.com/doifinder/10.1038/nature11013.

[41] Sander Granneman, Grzegorz Kudla, Elisabeth Petfalski, and David Tollervey. Identification of protein binding sites on U3 snoRNA and pre-rRNA by UV cross-linking and high-throughput analysis of cDNAs. *Proceedings of the National Academy of Sciences*, 106(24):9613–9618, 2009. URL http://www.pnas.org/content/106/24/9613.short.

[42] Frances M. Gratacós and Gary Brewer. The role of AUF1 in regulated mRNA decay. *Wiley Interdisciplinary Reviews: RNA*, 1 (3):457–473, November 2010. ISSN 17577004. doi: 10.1002/wrna.26. URL http://doi.wiley.com/10.1002/wrna.26.

[43] Andreas R AR Gruber, Jörg Fallmann, Franz Kratochvill, Pavel Kovarik, and Ivo L Hofacker. AREsite: a database for the comprehensive investigation of AU-rich elements. *Nucleic acids research*, 39(Database issue):1–4, November 2011. ISSN 1362-4962. doi: 10.1093/nar/gkq990. URL http://nar.oxfordjournals.org/content/39/suppl_1/D66.shorthttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3013810&tool=pmcentrez&rendertype=abstracthttp://www.ncbi.nlm.nih.gov/pubmed/21071424.

[44] A.R. Gruber, S.H.F. Bernhart, Y. Zhou, and Hofacker I.L. Rnalfoldz: efficient prediction of thermodynamically stable, local secondary structures. *German Conference on Bioinformatics*, 173: 12–21, 2010.

[45] Markus Hafner, Markus Landthaler, Lukas Burger, Mohsen Khorshid, Jean Hausser, Philipp Berninger, Andrea Rothballer, Manuel Ascano, Anna-carina Jungkamp, Mathias Munschauer, Alexander Ulrich, Greg S Wardle, Scott Dewell, Mihaela Zavolan, and Thomas Tuschl. Transcriptome-wide identification of RNA-binding protein and microRNA target sites by PAR-CLIP. *Cell*, 141(1):129–41, April 2010. ISSN 1097-4172. doi: 10.1016/j.cell.2010.03.009. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2861495&tool=pmcentrez&rendertype=abstracthttp://dx.doi.org/10.1016/j.cell.2010.03.009.

[46] Thomas B Hansen, Trine I Jensen, Bettina H Clausen, Jesper B Bramsen, Bente Finsen, Christian K Damgaard, and Jø rgen Kjems. Natural RNA circles function as efficient microRNA sponges. *Nature*, 495(7441):384–8, March 2013. ISSN 1476-4687. doi: 10.1038/nature11993. URL http://www.ncbi.nlm.nih.gov/pubmed/23446346.

[47] Michael Hiller, Rainer Pudimat, Anke Busch, and Rolf Backofen. Using RNA secondary structures to guide sequence motif finding towards single-stranded regions. *Nucleic acids research*, 34(17):e117, January 2006. ISSN 1362-4962. doi: 10.1093/nar/gkl544. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1903381&tool=pmcentrez&rendertype=abstract.

[48] I. L. Hofacker, W. Fontana, P. F. Stadler, L. S. Bonhoeffer, M. Tacker, and P. Schuster. Fast folding and comparison of RNA secondary structures. *Monatshefte fi¿œr Chemie Chemical Monthly*, 125(2):167–188, February 1994. ISSN 0026-9247. doi: 10.1007/BF00818163. URL http://www.springerlink.com/index/10.1007/BF00818163.

[49] I.L. Hofacker, B. Priwitzer, and P.F. Stadler. Prediction of locally stable rna secondary structures for genome-wide surveys. *Bioinformatics*, 20:186–190, 2004.

[50] Steve Hoffmann, Christian Otto, Stefan Kurtz, Cynthia M Sharma, Philipp Khaitovich, Jörg Vogel, Peter F Stadler, and Jörg Hackermüller. Fast mapping of short sequences with mismatches, insertions and deletions using index structures. *PLoS computational biology*, 5(9):e1000502, September 2009. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1000502. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2730575&tool=pmcentrez&rendertype=abstract.

[51] Da Wei Huang, Brad T Sherman, and Richard A Lempicki. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nature Protocols*, 4(1):44–57, December 2008. ISSN 1754-2189, 1750-2799. doi: 10.1038/nprot.2008.211. URL http://www.nature.com/doifinder/10.1038/nprot.2008.211.

[52] Brian P Hudson, Maria a Martinez-Yamout, H Jane Dyson, and Peter E Wright. Recognition of the mRNA AU-rich element by the zinc finger domain of TIS11d. *Nature structural & molecular biology*, 11(3):257–64, March 2004. ISSN 1545-9993. doi: 10.1038/nsmb738. URL http://www.ncbi.nlm.nih.gov/pubmed/14981510.

[53] Faoud T Ishmael, Xi Fang, Maria Rosaria Galdiero, Ulus Atasoy, William F C Rigby, Myriam Gorospe, Chris Cheadle, Cristiana Stellato, and Glucocorticoids Gcs. Role of the RNA-Binding Protein Tristetraprolin in Glucocoticoid-Mediated Gene Regulation. *The journal of Immunology*, pages 17–19, 2008.

[54] Martin Jinek, Krzysztof Chylinski, Ines Fonfara, Michael Hauer, Jennifer A Doudna, and Emmanuelle Charpentier. A Programmable Dual-RNA-Guided DNA Endonuclease in Adaptive Bacterial Immunity. *Science (New York, N.Y.)*, 337(6096):816–821, June 2012. ISSN 1095-9203. doi: 10.1126/science.1225829. URL http://www.sciencemag.org/content/337/6096/816.abstracthttp://www.sciencemag.org/content/early/2012/06/27/science.1225829.fullhttp://www.ncbi.nlm.nih.gov/pubmed/22745249.

[55] Neil C. Jones and Pavel Pevzner. *An introduction to bioinformatics algorithms*. Computational molecular biology. MIT Press, Cambridge, MA, 2004. ISBN 978-0-262-10106-6.

[56] Alexander Kanitz, Foivos Gypas, Andreas J Gruber, Andreas R. Gruber, Georges Martin, and Mihaela Zavolan.

Comparative assessment of methods for the computational inference of transcript isoform abundance from RNA-seq data. *Genome biology*, 16(1):150, December 2015. ISSN 1465-6914. doi: 10.1186/s13059-015-0702-5. URL http://genomebiology.com/2015/16/1/150http://www.genomebiology.com/content/pdf/s13059-015-0702-5.pdfhttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4511015&tool=pmcentrez&rendertype=abstract.

[57] Hilal Kazan, Debashish Ray, Esther T Chan, Timothy R Hughes, and Quaid Morris. RNAcontext: a new method for learning the sequence and structure binding preferences of RNA-binding proteins. *PLoS computational biology*, 6(7):e1000832, January 2010. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1000832. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2895634&tool=pmcentrez&rendertype=abstract.

[58] John D. Kececioglu and Eugene W. Myers. Combinatorial algorithms for DNA sequence assembly. *Algorithmica*, 13(1-2):7–51, 1995. URL http://link.springer.com/article/10.1007/BF01188580.

[59] W. J. Kent, C. W. Sugnet, T. S. Furey, K. M. Roskin, T. H. Pringle, a. M. Zahler, and a. D. Haussler. The Human Genome Browser at UCSC. *Genome Research*, 12(6):996–1006, May 2002. ISSN 1088-9051. doi: 10.1101/gr.229102. URL http://www.genome.org/cgi/doi/10.1101/gr.229102.

[60] Peter Kerpedjiev, Stefan Hammer, and Ivo L. Hofacker. Forna (force-directed RNA): Simple and effective online RNA secondary structure diagrams. *Bioinformatics (Oxford, England)*, 2015. URL http://www.tbi.univie.ac.at/newpapers/pdfs/TBI-p-2015-1.pdfhttp://bioinformatics.oxfordjournals.org/content/early/2015/06/21/bioinformatics.btv372.full.pdf.

[61] Michael Kertesz, Yue Wan, Elad Mazor, John L Rinn, Robert C Nutter, Howard Y Chang, and Eran Segal. Genome-wide measurement of RNA secondary structure in yeast. *Nature*, 467(7311):103–7, September 2010. ISSN 1476-4687. doi: 10.1038/nature09322. URL http://www.ncbi.nlm.nih.gov/pubmed/20811459.

[62] M. Khorshid, C. Rodak, and M. Zavolan. CLIPZ: a database and analysis environment for experimentally determined binding sites of RNA-binding proteins. *Nucleic Acids Research*, 39(Database):D245–D252, January 2011. ISSN 0305-1048, 1362-4962. doi: 10.1093/nar/gkq940. URL http://nar.oxfordjournals.org/lookup/doi/10.1093/nar/gkq940.

[63] Jongmin Kim and Erik Winfree. Synthetic in vitro transcriptional oscillators. *Molecular systems biology*, 7(465):465, February 2011. ISSN 1744-4292. doi: 10.1038/msb.2010.119. URL http://www.nature.com/doifinder/10.1038/msb.2010.119http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3063688&tool=pmcentrez&rendertype=abstract.

[64] R. J. Kinsella, A. Kahari, S. Haider, J. Zamora, G. Proctor, G. Spudich, J. Almeida-King, D. Staines, P. Derwent, A. Kerhornou, P. Kersey, and P. Flicek. Ensembl BioMarts: a hub for data retrieval across taxonomic space. *Database*, 2011(0):bar030–bar030, July 2011. ISSN 1758-0463. doi: 10.1093/database/bar030. URL http://database.oxfordjournals.org/cgi/doi/10.1093/database/bar030.

[65] S J Klug and M Famulok. All you wanted to know about SELEX. *Molecular biology reports*, 20(2):97–107, January 1994. ISSN 0301-4851. URL http://www.ncbi.nlm.nih.gov/pubmed/7536299.

[66] Ellen Knierim, Barbara Lucke, Jana Marie Schwarz, Markus Schuelke, and Dominik Seelow. Systematic Comparison of Three Methods for Fragmentation of Long-Range PCR Products for Next Generation Sequencing. *PLoS ONE*, 6(11):e28240, November 2011. ISSN 1932-6203. doi: 10.1371/journal.pone.0028240. URL http://dx.plos.org/10.1371/journal.pone.0028240.

[67] Julian König, Kathi Zarnack, Gregor Rot, Tomaz Curk, Melis Kayikci, Blaz Zupan, Daniel J Turner, Nicholas M Luscombe, and Jernej Ule. iCLIP reveals the function of hnRNP particles in splicing at individual nucleotide resolution. *Nature structural & molecular biology*, 17(7):909–15, July 2010. ISSN 1545-9985. doi: 10.1038/nsmb.1838. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3000544&tool=pmcentrez&rendertype=abstract.

[68] Julian König, Kathi Zarnack, NM Luscombe, and Jernej Ule. Protein-RNA interactions: new genomic technologies and perspectives. *Nature Reviews Genetics*, 13(February):77–83, 2011. ISSN 1471-0056. doi: 10.1038/nrg3141. URL http://www.nature.com/nrg/journal/v13/n2/abs/nrg3141.htmlhttp://discovery.ucl.ac.uk/1344586/.

[69] Franz Kratochvill, Christian Machacek, Claus Vogl, Florian Ebner, Vitaly Sedlyarov, Andreas R Gruber, Harald Hartweger, Raimund Vielnascher, Marina Karaghiosoff, Thomas Rülicke, Mathias Müller, Ivo Hofacker, Roland Lang, and Pavel Kovarik. Tristetraprolin-driven regulatory circuit controls quality and timing of mRNA decay in inflammation. *Molec-*

ular systems biology, 7(560):560, January 2011. ISSN 1744-4292. doi: 10.1038/msb.2011.93. URL http://www.ncbi.nlm.nih.gov/pubmed/22186734.

[70] Grzegorz Kudla, Sander Granneman, Daniela Hahn, Jean D Beggs, and David Tollervey. Cross-linking, ligation, and sequencing of hybrids reveals RNA-RNA interactions in yeast. *Proceedings of the National Academy of Sciences of the United States of America*, 108(24):10010–5, June 2011. ISSN 1091-6490. doi: 10.1073/pnas.1017386108. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3116431&tool=pmcentrez&rendertype=abstract.

[71] Wi S Lai, Joel S Parker, Sherry F Grissom, Deborah J Stumpo, and Perry J Blackshear. Novel mRNA targets for tristetraprolin (TTP) identified by global analysis of stabilized transcripts in TTP-deficient fibroblasts. *Molecular and cellular biology*, 26(24):9196–208, December 2006. ISSN 0270-7306. doi: 10.1128/MCB.00945-06. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1698545&tool=pmcentrez&rendertype=abstract.

[72] Nicole Lambert, Alex Robertson, Mohini Jangi, Sean McGeary, Phillip A. Sharp, and Christopher B. Burge. RNA Bind-n-Seq: Quantitative Assessment of the Sequence and Structural Binding Specificity of RNA Binding Proteins. *Molecular Cell*, 54(5): 887–900, June 2014. ISSN 10972765. doi: 10.1016/j.molcel.2014.04.016. URL http://linkinghub.elsevier.com/retrieve/pii/S109727651400327X.

[73] Svetlana Lebedeva, Marvin Jens, Kathrin Theil, Björn Schwanhäusser, Matthias Selbach, Markus Landthaler, and Nikolaus Rajewsky. Transcriptome-wide Analysis of Regulatory Interactions of the RNA-Binding Protein HuR. *Molecular cell*, June 2011. ISSN 1097-4164. doi: 10.1016/j.molcel.2011.06.008. URL http://www.ncbi.nlm.nih.gov/pubmed/21723171.

[74] Ju Youn J.E. J.Y. Jerome E Lee, Jeffrey Wilusz, Bin Tian, and Carol J C.J. Wilusz. Systematic analysis of cis-elements in unstable mRNAs demonstrates that CUGBP1 is a key regulator of mRNA decay in muscle cells. *PloS one*, 5(6):e11201, January 2010. ISSN 1932-6203. doi: 10.1371/journal.pone.0011201. URL http://www.ncbi.nlm.nih.gov/pubmed/20574513http://dx.plos.org/10.1371/journal.pone.0011201.

[75] L. Leibovich and Z. Yakhini. Efficient motif search in ranked lists and applications to variable gap motifs. *Nucleic Acids Research*, 40(13):5832–5847, July 2012. ISSN 0305-1048, 1362-4962.

doi: 10.1093/nar/gks206. URL http://nar.oxfordjournals.org/lookup/doi/10.1093/nar/gks206.

[76] N B Leontis and E Westhof. Geometric nomenclature and classification of RNA base pairs. *RNA (New York, N.Y.)*, 7(4):499–512, April 2001. ISSN 1355-8382. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1370104&tool=pmcentrez&rendertype=abstract.

[77] N.B. Leontis and Eric Westhof. Analysis of RNA motifs. *Current Opinion in Structural Biology*, 13(3):300–308, June 2003. ISSN 0959440X. doi: 10.1016/S0959-440X(03)00076-9. URL http://www.sciencedirect.com/science/article/pii/S0959440X03000769http://linkinghub.elsevier.com/retrieve/pii/S0959440X03000769.

[78] N.B. Neocles B Leontis, Jesse Stombaugh, and Eric Westhof. The non-Watson-Crick base pairs and their associated isostericity matrices. *Nucleic acids research*, 30 (16):3497–531, August 2002. ISSN 1362-4962. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=134247&tool=pmcentrez&rendertype=abstracthttp://nar.oxfordjournals.org/content/30/16/3497.short.

[79] Hai Li, Wei Chen, Yue Zhou, Parveen Abidi, Orr Sharpe, William H W.H. William H Robinson, Fredric B F.B. Kraemer, and Jingwen Liu. Identification of mRNA binding proteins that regulate the stability of LDL receptor mRNA through AU-rich elements. *Journal of lipid research*, 50(5):820–31, May 2009. ISSN 0022-2275. doi: 10.1194/jlr.M800375-JLR200. URL http://www.jlr.org/content/50/5/820.shorthttp://www.ncbi.nlm.nih.gov/pubmed/19141871http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2666168&tool=pmcentrez&rendertype=abstract.

[80] Jun-Hao Li, Shun Liu, Hui Zhou, Liang-Hu Qu, and Jian-Hua Yang. starBase v2.0: decoding miRNA-ceRNA, miRNA-ncRNA and protein-RNA interaction networks from large-scale CLIP-Seq data. *Nucleic Acids Research*, 42(D1):D92–D97, January 2014. ISSN 0305-1048, 1362-4962. doi: 10.1093/nar/gkt1248. URL http://nar.oxfordjournals.org/lookup/doi/10.1093/nar/gkt1248.

[81] X. Li, G. Quon, H. D. Lipshitz, and Q. Morris. Predicting in vivo binding sites of RNA-binding proteins using mRNA secondary structure. *RNA*, 16(6):1096–1107, June 2010. ISSN 1355-8382. doi: 10.1261/rna.2017210. URL http://rnajournal.cshlp.org/cgi/doi/10.1261/rna.2017210.

[82] Yi-Hsuan Lin and Ralf Bundschuh. RNA structure generates natural cooperativity between single-stranded RNA binding proteins targeting 5′ and 3′UTRs. *Nucleic Acids Research*, 43(2): 1160–1169, 2015. doi: 10.1093/nar/gku1320.

[83] Ronny Lorenz, Stephan H Bernhart, Christian Höner Zu Siederdissen, Hakim Tafer, Christoph Flamm, Peter F Stadler, and Ivo L Hofacker. ViennaRNA Package 2.0. *Algorithms for molecular biology : AMB*, 6:26, January 2011. ISSN 1748-7188. doi: 10.1186/1748-7188-6-26. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3319429&tool=pmcentrez&rendertype=abstract.

[84] D. Loughrey, K. E. Watters, A. H. Settle, and J. B. Lucks. SHAPE-Seq 2.0: systematic optimization and extension of high-throughput chemical probing of RNA secondary structure with next generation sequencing. *Nucleic Acids Research*, 42(21): e165–e165, December 2014. ISSN 0305-1048, 1362-4962. doi: 10.1093/nar/gku909. URL http://nar.oxfordjournals.org/lookup/doi/10.1093/nar/gku909.

[85] Michael I Love, Wolfgang Huber, and Simon Anders. This Provisional PDF corresponds to the article as it appeared upon acceptance . Fully formatted Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, 15(12):550, 2014. ISSN 1465-6906. doi: 10.1186/s13059-014-0550-8. URL http://genomebiology.com/2014/15/12/550http://www.genomebiology.com/content/pdf/s13059-014-0550-8.pdf.

[86] Yi-Chien Lu, Sung-Hee Chang, Markus Hafner, Xi Li, Thomas Tuschl, Olivier Elemento, and Timothy Hla. ELAVL1 Modulates Transcriptome-wide miRNA Binding in Murine Macrophages. *Cell reports*, 9(6):2330–43, December 2014. ISSN 2211-1247. doi: 10.1016/j.celrep.2014.11.030. URL http://www.ncbi.nlm.nih.gov/pubmed/25533351.

[87] J.B. Julius B Lucks, Lei Qi, Vivek K V.K. Mutalik, Denise Wang, and Adam P A.P. Arkin. Versatile RNA-sensing transcriptional regulators for engineering genetic networks. *Proceedings of the National Academy of Sciences of the United States of America*, 108(21):8617–8622, May 2011. ISSN 1091-6490. doi: 10.1073/pnas.1015741108. URL http://www.pnas.org/cgi/content/abstract/108/21/8617http://www.ncbi.nlm.nih.gov/pubmed/21555549http://www.pnas.org/content/108/21/8617.short.

[88] Bradley M Lunde, Claire Moore, and Gabriele Varani. RNA-binding proteins: modular design for efficient function. *Nature*

reviews. Molecular cell biology, 8(6):479–90, June 2007. ISSN 1471-0072. doi: 10.1038/nrm2178. URL http://www.ncbi.nlm.nih.gov/pubmed/17473849.

[89] Elaine R. Mardis. A decade's perspective on DNA sequencing technology. *Nature*, 470(7333):198–203, February 2011. ISSN 0028-0836, 1476-4687. doi: 10.1038/nature09796. URL http://www.nature.com/doifinder/10.1038/nature09796.

[90] Elaine R. Mardis. Next-Generation Sequencing Platforms. *Annual Review of Analytical Chemistry*, 6(1):287–303, June 2013. ISSN 1936-1327, 1936-1335. doi: 10.1146/annurev-anchem-062012-092628. URL http://www.annualreviews.org/doi/abs/10.1146/annurev-anchem-062012-092628.

[91] Marcel Martin. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet. journal*, pages 10–12, 2011. URL http://journaldev.embnet.org/index.php/embnetjournal/article/view/200.

[92] Grégoire Masliah, Pierre Barraud, and Frédéric H. T. Allain. RNA recognition by double-stranded RNA binding domains: a matter of shape and sequence. *Cellular and Molecular Life Sciences*, August 2012. ISSN 1420-682X, 1420-9071. doi: 10.1007/s00018-012-1119-x. URL http://link.springer.com/10.1007/s00018-012-1119-x.

[93] David H Mathews, Matthew D Disney, Jessica L Childs, Susan J Schroeder, Michael Zuker, and Douglas H Turner. Incorporating chemical modification constraints into a dynamic programming algorithm for prediction of RNA secondary structure. *Proceedings of the National Academy of Sciences of the United States of America*, 101(19):7287–7292, 2004.

[94] Daniel Maticzka, Sita J. Lange, Fabrizio Costa, and Rolf Backofen. GraphProt: modeling binding preferences of RNA-binding proteins. *Genome Biol*, 15(1):R17, 2014. URL http://www.biomedcentral.com/content/pdf/gb-2014-15-1-r17.pdf.

[95] J S McCaskill. The equilibrium partition function and base pair binding probabilities for RNA secondary structure. *Biopolymers*, 29(6-7):1105–19, 1990. ISSN 0006-3525. doi: 10.1002/bip.360290621. URL http://www.ncbi.nlm.nih.gov/pubmed/1695107http://www.hubmed.org/display.cgi?uids=1695107.

[96] Colleen A. McHugh, Pamela Russell, and Mitchell Guttman. Methods for comprehensive experimental identification of rna-protein interactions. *Genome Biol*, 15:203, 2014. URL http://www.biomedcentral.com/content/pdf/gb4152.pdf.

[97] G. V. Mechetin and D. O. Zharkov. Mechanisms of diffusional search for specific targets by DNA-dependent proteins. *Biochemistry (Moscow)*, 79(6):496–505, June 2014. ISSN 0006-2979, 1608-3040. doi: 10.1134/S0006297914060029. URL http://link.springer.com/10.1134/S0006297914060029.

[98] Sebastian Memczak, Marvin Jens, Antigoni Elefsinioti, Francesca Torti, Janna Krueger, Agnieszka Rybak, Luisa Maier, Sebastian D Mackowiak, Lea H Gregersen, Mathias Munschauer, Alexander Loewer, Ulrike Ziebold, Markus Landthaler, Christine Kocks, Ferdinand le Noble, and Nikolaus Rajewsky. Circular RNAs are a large class of animal RNAs with regulatory potency. *Nature*, 495(7441):333–8, March 2013. ISSN 1476-4687. doi: 10.1038/nature11928. URL http://www.ncbi.nlm.nih.gov/pubmed/23446348.

[99] Michael L. Metzker. Sequencing technologies - the next generation. *Nature Reviews Genetics*, 11(1):31–46, January 2010. ISSN 1471-0056, 1471-0064. doi: 10.1038/nrg2626. URL http://www.nature.com/doifinder/10.1038/nrg2626.

[100] Huaiyu Mi, Sagar Poudel, Anushya Muruganujan, John T. Casagrande, and Paul D. Thomas. PANTHER version 10: expanded protein families and functions, and analysis tools. *Nucleic Acids Research*, 44(D1):D336–D342, January 2016. ISSN 0305-1048, 1362-4962. doi: 10.1093/nar/gkv1194. URL http://nar.oxfordjournals.org/lookup/doi/10.1093/nar/gkv1194.

[101] Neelanjan Mukherjee, Patrick J Lager, Matthew B Friedersdorf, Marshall A Thompson, and Jack D Keene. Coordinated post-transcriptional mRNA population dynamics during T-cell activation. *Molecular Systems Biology*, 5(288):288, 2009. URL http://www.ncbi.nlm.nih.gov/pubmed/19638969.

[102] Neelanjan Mukherjee, Nicholas C Jacobs, Markus Hafner, Elizabeth a Kennington, Jeffrey D Nusbaum, Thomas Tuschl, Perry J Blackshear, and Uwe Ohler. Global target mRNA specification and regulation by the RNA-binding protein ZFP36. *Genome biology*, 15(1):R12, January 2014. ISSN 1465-6914. doi: 10.1186/gb-2014-15-1-r12. URL http://www.ncbi.nlm.nih.gov/pubmed/24401661.

[103] Eugene W. Myers. The fragment assembly string graph. *Bioinformatics*, 21(suppl 2):ii79–ii85, 2005. doi: 10.1093/bioinformatics/bti1114. URL http://bioinformatics.oxfordjournals.org/content/21/suppl_2/ii79.abstract.

[104] R Nussinov and a B Jacobson. Fast algorithm for predicting the secondary structure of single-stranded RNA. *Proceedings*

of the National Academy of Sciences of the United States of America, 77(11):6309–13, November 1980. ISSN 0027-8424. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=350273&tool=pmcentrez&rendertype=abstract.

[105] Marcin F Osuchowski, Daniel G Remick, James a Lederer, Charles H Lang, Ansgar O Aasen, Mayuki Aibiki, Luciano C Azevedo, Soheyl Bahrami, Mihaly Boros, Robert Cooney, Salvatore Cuzzocrea, Yong Jiang, Wolfgang G Junger, Hiroyuki Hirasawa, Richard S Hotchkiss, Xiang-An Li, Peter Radermacher, Heinz Redl, Reinaldo Salomao, Amin Soebandrio, Christoph Thiemermann, Jean-Louis Vincent, Peter Ward, Yong-Ming Yao, Huang-Ping Yu, Basilia Zingarelli, and Irshad H Chaudry. Abandon the mouse research ship? Not just yet! *Shock (Augusta, Ga.)*, 41(6):463–75, 2014. ISSN 1540-0514. doi: 10.1097/SHK.0000000000000153. URL http://www.ncbi.nlm.nih.gov/pubmed/24569509.

[106] Chandra Shekhar Pareek, Rafal Smoczynski, and Andrzej Tretyn. Sequencing technologies and genome sequencing. *Journal of Applied Genetics*, 52(4):413–435, November 2011. ISSN 1234-1983, 2190-3883. doi: 10.1007/s13353-011-0057-x. URL http://link.springer.com/10.1007/s13353-011-0057-x.

[107] Pavel A. Pevzner, Haixu Tang, and Michael S. Waterman. An Eulerian path approach to DNA fragment assembly. *Proceedings of the National Academy of Sciences*, 98(17):9748–9753, 2001. URL http://www.pnas.org/content/98/17/9748.short.

[108] Aaron R Quinlan and Ira M Hall. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics (Oxford, England)*, 26(6):841–2, March 2010. ISSN 1367-4811. doi: 10.1093/bioinformatics/btq033. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2832824&tool=pmcentrez&rendertype=abstract.

[109] Michal Rabani, Michael Kertesz, and Eran Segal. Computational prediction of RNA structural motifs involved in posttranscriptional regulatory processes. *Proceedings of the National Academy of Sciences of the United States of America*, 105(39):14885–90, September 2008. ISSN 1091-6490. doi: 10.1073/pnas.0803169105. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2567462&tool=pmcentrez&rendertype=abstract.

[110] Debashish Ray, Hilal Kazan, Esther T Chan, Lourdes Peña Castillo, Sidharth Chaudhry, Shaheynoor Talukder, Benjamin J Blencowe, Quaid Morris, and Timothy R Hughes. Rapid and

systematic analysis of the RNA recognition specificities of RNA-binding proteins. *Nature biotechnology*, 27(7):667–70, July 2009. ISSN 1546-1696. doi: 10.1038/nbt.1550. URL http://www.ncbi.nlm.nih.gov/pubmed/19561594.

[111] Debashish Ray, Hilal Kazan, Kate B. Cook, Matthew T. Weirauch, Hamed S. Najafabadi, Xiao Li, Serge Gueroussov, Mihai Albu, Hong Zheng, Ally Yang, Hong Na, Manuel Irimia, Leah H. Matzat, Ryan K. Dale, Sarah A. Smith, Christopher A. Yarosh, Seth M. Kelly, Behnam Nabet, Desirea Mecenas, Weimin Li, Rakesh S. Laishram, Mei Qiao, Howard D. Lipshitz, Fabio Piano, Anita H. Corbett, Russ P. Carstens, Brendan J. Frey, Richard A. Anderson, Kristen W. Lynch, Luiz O. F. Penalva, Elissa P. Lei, Andrew G. Fraser, Benjamin J. Blencowe, Quaid D. Morris, and Timothy R. Hughes. A compendium of RNA-binding motifs for decoding gene regulation. *Nature*, 499(7457):172–177, July 2013. ISSN 0028-0836, 1476-4687. doi: 10.1038/nature12311. URL http://www.nature.com/doifinder/10.1038/nature12311.

[112] Knut Reinert, Ben Langmead, David Weese, and Dirk J. Evers. Alignment of Next-Generation Sequencing Reads. *Annual Review of Genomics and Human Genetics*, 16(1):133–151, August 2015. ISSN 1527-8204, 1545-293X. doi: 10.1146/annurev-genom-090413-025358. URL http://www.annualreviews.org/doi/abs/10.1146/annurev-genom-090413-025358.

[113] Reyes and Elisa Ficarra. Computational Methods for CLIP-seq Data Processing. *Bioinformatics and Biology Insights*, page 199, October 2014. ISSN 1177-9322. doi: 10.4137/BBI.S16803. URL http://www.la-press.com/computational-methods-for-clip-seq-data-processing-article-a4405.

[114] D. P. Riordan, D. Herschlag, and P. O. Brown. Identification of RNA recognition elements in the Saccharomyces cerevisiae transcriptome. *Nucleic Acids Research*, 39(4):1501–1509, March 2011. ISSN 0305-1048, 1362-4962. doi: 10.1093/nar/gkq920. URL http://nar.oxfordjournals.org/lookup/doi/10.1093/nar/gkq920.

[115] Silvi Rouskin, Meghan Zubradt, Stefan Washietl, Manolis Kellis, and Jonathan S. Weissman. Genome-wide probing of RNA structure reveals active unfolding of mRNA structures in vivo. *Nature*, 505(7485):701–705, December 2013. ISSN 0028-0836, 1476-4687. doi: 10.1038/nature12894. URL http://www.nature.com/doifinder/10.1038/nature12894.

[116] M. Ruffalo, T. LaFramboise, and M. Koyuturk. Comparative analysis of algorithms for next-generation sequencing

read alignment. *Bioinformatics*, 27(20):2790–2796, October 2011. ISSN 1367-4803, 1460-2059. doi: 10.1093/bioinformatics/btr477. URL http://bioinformatics.oxfordjournals.org/cgi/doi/10.1093/bioinformatics/btr477.

[117] S. P. Sadedin, B. Pope, and A. Oshlack. Bpipe: a tool for running and managing bioinformatics pipelines. *Bioinformatics*, 28(11):1525–1526, June 2012. ISSN 1367-4803, 1460-2059. doi: 10.1093/bioinformatics/bts167. URL http://bioinformatics.oxfordjournals.org/cgi/doi/10.1093/bioinformatics/bts167.

[118] Heike Sandler and Georg Stoecklin. Control of mRNA decay by phosphorylation of tristetraprolin. *Biochemical Society transactions*, 36(Pt 3):491–6, June 2008. ISSN 0300-5127. doi: 10.1042/BST0360491. URL http://www.ncbi.nlm.nih.gov/pubmed/18481987.

[119] Sophie Schbath, Véronique Martin, Matthias Zytnicki, Julien Fayolle, Valentin Loux, and Jean-François Gibrat. Mapping Reads on a Genomic Sequence: An Algorithmic Overview and a Practical Comparative Analysis. *Journal of Computational Biology*, 19(6):796–813, June 2012. ISSN 1066-5277, 1557-8666. doi: 10.1089/cmb.2012.0022. URL http://online.liebertpub.com/doi/abs/10.1089/cmb.2012.0022.

[120] Vitaly Sedlyarov, Jörg Fallmann, Florian Ebner, Jakob Huemer, Lucy Sneezum, Masa Ivin, Kristina Kreiner, Andrea Tanzer, Claus Vogl, Ivo Hofacker, and Pavel Kovarik. Tristetraprolin binding site atlas in the macrophage transcriptome reveals a switch for inflammationÂ resolution. *Molecular Systems Biology*, 12(5):n/a–n/a, 2016. ISSN 1744-4292. doi: 10.15252/msb.20156628. URL http://dx.doi.org/10.15252/msb.20156628.

[121] Rahul Siddharthan, Eric D. Siggia, and Erik van Nimwegen. PhyloGibbs: A Gibbs Sampling Motif Finder That Incorporates Phylogeny. *PLoS Computational Biology*, 1(7):e67, 2005. ISSN 1553-734X, 1553-7358. doi: 10.1371/journal.pcbi.0010067. URL http://dx.plos.org/10.1371/journal.pcbi.0010067.

[122] Jared T. Simpson and Mihai Pop. The Theory and Practice of Genome Sequence Assembly. *Annual Review of Genomics and Human Genetics*, 16(1):153–172, August 2015. ISSN 1527-8204, 1545-293X. doi: 10.1146/annurev-genom-090314-050032. URL http://www.annualreviews.org/doi/abs/10.1146/annurev-genom-090314-050032.

[123] Anil Kumar Singh and Swaranjit Singh Cameotra. Rhamnolipids production by multi-metal-resistant and plant-growth-

promoting rhizobacteria. *Applied biochemistry and biotechnology*, 170(5):1038–56, July 2013. ISSN 1559-0291. doi: 10.1007/s12010-013-0244-9. URL http://www.ncbi.nlm.nih.gov/pubmed/23640260.

[124] Sergei a Solonenko, J César Ignacio-Espinoza, Adriana Alberti, Corinne Cruaud, Steven Hallam, Kostas Konstantinidis, Gene Tyson, Patrick Wincker, and Matthew B Sullivan. Sequencing platform and library preparation choices impact viral metagenomes. *BMC genomics*, 14:320, January 2013. ISSN 1471-2164. doi: 10.1186/1471-2164-14-320. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3655917&tool=pmcentrez&rendertype=abstract.

[125] R. Staden. A strategy of DNA sequencing employing computer programs. *Nucleic acids research*, 6(7):2601–2610, 1979. URL http://nar.oxfordjournals.org/content/6/7/2601.short.

[126] Georg Stoecklin, Thomas Mayo, and Paul Anderson. ARE-mRNA degradation requires the 5′-3′ decay pathway. *EMBO reports*, 7(1):72–7, January 2006. ISSN 1469-221X. doi: 10.1038/sj.embor.7400572. URL http://www.ncbi.nlm.nih.gov/pubmed/16299471.

[127] Yoichiro Sugimoto, Alessandra Vigilante, Elodie Darbo, Alexandra Zirra, Cristina Militti, Andrea DÁmbrogio, Nicholas M. Luscombe, and Jernej Ule. hiCLIP reveals the in vivo atlas of mRNA secondary structures recognized by Staufen 1. *Nature*, 519(7544):491–494, March 2015. ISSN 0028-0836, 1476-4687. doi: 10.1038/nature14280. URL http://www.nature.com/doifinder/10.1038/nature14280.

[128] Cole Trapnell, Brian a Williams, Geo Pertea, Ali Mortazavi, Gordon Kwan, Marijke J van Baren, Steven L Salzberg, Barbara J Wold, and Lior Pachter. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature biotechnology*, 28(5):511–5, May 2010. ISSN 1546-1696. doi: 10.1038/nbt.1621. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3146043&tool=pmcentrez&rendertype=abstract.

[129] Cole Trapnell, Adam Roberts, Loyal Goff, Geo Pertea, Daehwan Kim, David R Kelley, Harold Pimentel, Steven L Salzberg, John L Rinn, and Lior Pachter. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nature Protocols*, 7(3):562–578, March 2012. ISSN 1754-2189. doi: 10.1038/nprot.2012.016. URL http://www.ncbi.nlm.nih.gov/pubmed/22383036http://www.nature.com/doifinder/10.1038/nprot.2012.016.

[130] Douglas H Turner and David H Mathews. NNDB: the nearest neighbor parameter database for predicting stability of nucleic acid secondary structure. *Nucleic acids research*, page gkp892, 2009.

[131] Jernej Ule, Kirk B Jensen, Matteo Ruggiu, Aldo Mele, Aljaz Ule, and Robert B Darnell. CLIP identifies Nova-regulated RNA networks in the brain. *Science (New York, N.Y.)*, 302(5648):1212–5, November 2003. ISSN 1095-9203. doi: 10.1126/science.1090095. URL http://www.ncbi.nlm.nih.gov/pubmed/14615540.

[132] Jernej Ule, Kirk Jensen, Aldo Mele, and Robert B. Darnell. CLIP: A method for identifying protein-RNA interaction sites in living cells. *Methods*, 37(4):376–386, December 2005. ISSN 10462023. doi: 10.1016/j.ymeth.2005.07.018. URL http://linkinghub.elsevier.com/retrieve/pii/S1046202305001787.

[133] Jumpei Umetsu. CLIPdb : a CLIP-seq database for protein-RNA interactions. *BMC genomics*, 16(1):51, 2015. ISSN 1471-2164. doi: 10.1186/s12864-015-1273-2. URL http://www.biomedcentral.com/1471-2164/16/51http://www.biomedcentral.com/content/pdf/s12864-015-1273-2.pdf.

[134] Philip J Uren, Suzanne C Burns, Jianhua Ruan, Kusum K Singh, Andrew D Smith, and Luiz O F Penalva. Genomic analyses of the RNA-binding protein Hu antigen R (HuR) identify a complex network of target genes and novel characteristics of its binding sites. *The Journal of biological chemistry*, 286(43):37063–6, October 2011. ISSN 1083-351X. doi: 10.1074/jbc.C111.266882. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3199453&tool=pmcentrez&rendertype=abstract.

[135] Philip J Uren, Emad Bahrami-Samani, Suzanne C Burns, Mei Qiao, Fedor V Karginov, Emily Hodges, Gregory J Hannon, Jeremy R Sanford, Luiz O F Penalva, and Andrew D Smith. Site identification in high-throughput RNA-protein interaction data. *Bioinformatics (Oxford, England)*, 28(23):3013–3020, September 2012. ISSN 1367-4811. doi: 10.1093/bioinformatics/bts569. URL http://www.ncbi.nlm.nih.gov/pubmed/23024010.

[136] Roberto Valverde, Laura Edwards, and Lynne Regan. Structure and function of KH domains: Structure and function of KH domains. *FEBS Journal*, 275(11):2712–2726, June 2008. ISSN 1742464X. doi: 10.1111/j.1742-4658.2008.06411.x. URL http://doi.wiley.com/10.1111/j.1742-4658.2008.06411.x.

[137] Erwin L. van Dijk, HÃ©lÃšne Auger, Yan Jaszczyszyn, and Claude Thermes. Ten years of next-generation sequencing technology. *Trends in Genetics*, 30(9):418–426, September 2014.

ISSN 01689525. doi: 10.1016/j.tig.2014.07.001. URL http://linkinghub.elsevier.com/retrieve/pii/S0168952514001127.

[138] Eric L Van Nostrand, Gabriel A Pratt, Alexander A Shishkin, Chelsea Gelboin-Burkhart, Mark Y Fang, Balaji Sundararaman, Steven M Blue, Thai B Nguyen, Christine Surka, Keri Elkins, Rebecca Stanton, Frank Rigo, Mitchell Guttman, and Gene W Yeo. Robust transcriptome-wide discovery of RNA-binding protein binding sites with enhanced CLIP (eCLIP). *Nature Methods*, 13(6):508–514, March 2016. ISSN 1548-7091, 1548-7105. doi: 10.1038/nmeth.3810. URL http://www.nature.com/doifinder/10.1038/nmeth.3810.

[139] W. N. Venables and B. D. Ripley. *Modern Applied Statistics with S*. Springer, New York, fourth edition, 2002. URL http://www.stats.ox.ac.uk/pub/MASS4. ISBN 0-387-95457-0.

[140] I.A. Irina A Vlasova, Nuzha M N.M. Tahoe, Danhua Fan, Ola Larsson, Bernd Rattenbacher, J.R. Julius R. J.R. Stern-John, Jayprakash Vasdewani, George Karypis, Cavan S C.S. Reilly, Peter B. P.B. Bitterman, Others, and Paul R Bohjanen. Conserved GU-rich elements mediate mRNA decay by binding to CUG-binding protein 1. *Molecular cell*, 29(2): 263–270, 2008. ISSN 10972765. doi: 10.1016/j.molcel.2007.11.024. URL http://www.sciencedirect.com/science/article/pii/S1097276507008192.

[141] Christopher von Roretz and Imed-Eddine Gallouzi. Decoding ARE-mediated decay: is microRNA part of the equation? *The Journal of cell biology*, 181(2):189–94, April 2008. ISSN 1540-8140. doi: 10.1083/jcb.200712054. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2315667&tool=pmcentrez&rendertype=abstract.

[142] A E Walter, D H Turner, J Kim, M H Lyttle, P Müller, D H Mathews, and M Zuker. Coaxial stacking of helixes enhances binding of oligoribonucleotides and improves predictions of RNA folding. *Proc Natl Acad Sci U S A*, 91(20):9218–9222, Sep 1994.

[143] T. Wang, G. Xiao, Y. Chu, M. Q. Zhang, D. R. Corey, and Y. Xie. Design and bioinformatics analysis of genome-wide CLIP experiments. *Nucleic Acids Research*, May 2015. ISSN 0305-1048. doi: 10.1093/nar/gkv439. URL http://nar.oxfordjournals.org/lookup/doi/10.1093/nar/gkv439http://nar.oxfordjournals.org/content/early/2015/05/09/nar.gkv439.full.pdf.

[144] Tao Wang, Yang Xie, and Guanghua Xiao. dCLIP: a computational approach for comparative CLIP-seq

analyses. *Genome biology*, 15(1):R11, January 2014. ISSN 1465-6914. doi: 10.1186/gb-2014-15-1-r11. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4054096&tool=pmcentrez&rendertype=abstract.

[145] Xiaoqiang Wang and T M Tanaka Hall. Structural basis for recognition of AU-rich element RNA by the HuD protein. *Nature Structural Biology*, 8(2):141–145, 2001. URL http://www.nature.com/nsmb/journal/v8/n2/abs/nsb0201_141.html.

[146] Xiaoqiang Wang, Phillip D. Zamore, and Traci M. Tanaka Hall. Crystal structure of a Pumilio homology domain. *Molecular cell*, 7(4):855–865, 2001. URL http://www.sciencedirect.com/science/article/pii/S1097276501002295.

[147] Zhong Wang, Mark Gerstein, and Michael Snyder. RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics*, 10(1):57–63, January 2009. ISSN 1471-0064. doi: 10.1038/nrg2484. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2949280&tool=pmcentrez&rendertype=abstracthttp://www.nature.com/nrg/journal/v10/n1/abs/nrg2484.html.

[148] J.D. Watson and F.H.C. Crick. Molecular structure of nucleic acids. *Nature*, 171(4356):737–738, 1953. URL http://www.nature.com/physics/looking-back/crick/.

[149] Benjamin a Webb, Sherry Hildreth, Richard F Helm, and Birgit E Scharf. Sinorhizobium meliloti chemoreceptor McpU mediates chemotaxis toward host plant exudates through direct proline sensing. *Applied and environmental microbiology*, 80(11):3404–15, June 2014. ISSN 1098-5336. doi: 10.1128/AEM.00115-14. URL http://www.ncbi.nlm.nih.gov/pubmed/24657863.

[150] Michael T. Wolfinger, Jörg Fallmann, Florian Eggenhofer, and Fabian Amman. ViennaNGS: A toolbox for building efficient next-generation sequencing analysis pipelines. *F1000Research*, 4, 2015. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4513691/.

[151] S Wuchty, W Fontana, I L Hofacker, and P Schuster. Complete suboptimal folding of RNA and the stability of secondary structures. *Biopolymers*, 49(2):145–65, February 1999. ISSN 0006-3525. doi: 10.1002/(SICI)1097-0282(199902)49:2<145::AID-BIP4>3.0.CO;2-G. URL http://www.ncbi.nlm.nih.gov/pubmed/10070264.

[152] Z. Yao, Z. Weinberg, and W. L. Ruzzo. CMfinder–a covariance model based RNA motif finding algorithm. *Bioinformatics*, 22(4):445–452, February 2006. ISSN 1367-4803,

1460-2059. doi: 10.1093/bioinformatics/btk008. URL http://bioinformatics.oxfordjournals.org/cgi/doi/10.1093/bioinformatics/btk008.

[153] Gene W Yeo, Nicole G Coufal, Tiffany Y Liang, Grace E Peng, Xiang-Dong Fu, and Fred H Gage. An RNA code for the FOX2 splicing regulator revealed by mapping RNA-protein interactions in stem cells. *Nature structural & molecular biology*, 16(2): 130–7, February 2009. ISSN 1545-9985. doi: 10.1038/nsmb.1545. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2735254&tool=pmcentrez&rendertype=abstract.

[154] Brian J Zarnegar, Ryan A Flynn, Ying Shen, Brian T Do, Howard Y Chang, and Paul A Khavari. irCLIP platform for efficient characterization of protein-RNA interactions. *Nature Methods*, 13(6):489–492, April 2016. ISSN 1548-7091, 1548-7105. doi: 10.1038/nmeth.3840. URL http://www.nature.com/doifinder/10.1038/nmeth.3840.

[155] Chaolin Zhang, Kuang-Yung Lee, Maurice S Swanson, and Robert B Darnell. Prediction of clustered RNA-binding protein motif sites in the mammalian genome. *Nucleic acids research*, 41(14):6793–807, August 2013. ISSN 1362-4962. doi: 10.1093/nar/gkt421. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3737533&tool=pmcentrez&rendertype=abstract.

[156] Bob Zimmermann, Ivana Bilusic, Christina Lorenz, and Renée Schroeder. Genomic SELEX: A discovery tool for genomic aptamers. *Methods*, 52(2):125–132, October 2010. ISSN 10462023. doi: 10.1016/j.ymeth.2010.06.004. URL http://linkinghub.elsevier.com/retrieve/pii/S1046202310001581.

[157] M Zuker. On finding all suboptimal foldings of an RNA molecule. *Science*, 244(4900):48–52, 1989. URL http://www.ncbi.nlm.nih.gov/pubmed/2468181.

COLOPHON

This document was typeset using the typographical look-and-feel `classicthesis` developed by André Miede. The style was inspired by Robert Bringhurst's seminal book on typography "*The Elements of Typographic Style*". `classicthesis` is available for both LaTeX and LyX:

http://code.google.com/p/classicthesis/

Happy users of `classicthesis` usually send a real postcard to the author, a collection of postcards received so far is featured here:

http://postcards.miede.de/

Final Version as of August 24, 2016 (`classicthesis` final version).