STAMMP - A statistical model and processing pipeline for PAR-CLIP data reveals transcriptome maps of mRNP biogenesis factors

Phillipp Torkler

München 2015

Dissertation zur Erlangung des Doktorgrades
der Fakultät für Chemie und Pharmazie
der Ludwig–Maximilians–Universität München

STAMMP - A statistical model and processing pipeline for PAR-CLIP data reveals transcriptome maps of mRNP biogenesis factors

Phillipp Torkler
aus
Melle, Deutschland

2015

Erklärung:

Diese Dissertation wurde im Sinne von §7 der Promotionsordnung vom 28. November 2011 von Herrn Dr. Johannes Söding betreut.

Eidesstattliche Versicherung:

Diese Dissertation wurde eigenständig und ohne unerlaubte Hilfe erarbeitet.

München, den 26. März 2015

Phillipp Torkler

Dissertation eingereicht am: 26.03.2015

1. Gutachter: Dr. Johannes Söding
2. Gutachter: Prof. Dr. Patrick Cramer

Mündliche Prüfung am: 27.04.2015

Acknowledgements

During my years working at the Gene Center many different people supported me and my work which I would like to thank here.

First, I would like to thank my mentor Dr. Johannes Söding for all his trust, support and guidance throughout my entire time in his research group. Your way of thinking and tackling scientific problems shaped my being as a scientist.

I am also thankful to my second supervisor Prof. Dr. Patrick Cramer for the opportunity to participate in projects in close collaboration of bioinformatics and biochemistry.

I am grateful to Dr. Dietmar Martin, Prof. Dr. Klaus Förstemann, Prof. Dr. Roland Beckmann and Prof. Dr. Karl-Peter Hopfner for offering their time as being members of my examination committee.

Carlo, my dear PAR-CLIP-buddy, I especially would like to thank you for the nice and successful collaboration, the ristretto time and the countless discussions about what to do next. I am sure there will be more to come.

Many thanks are going to all current and former groups and colleagues I worked with. Thank you Bene, Björn, Henrik, Sebastian, Carina, Theresa, Verena, Katha, Steffi F and of course Achim for giving me a pleasant start at the Gene Center. Thanks to the former 'Södings' Holger and Eckhart for their early advices about thesis planning. I like to thank Anja, Armin, Jessica, Markus, Stefan, Susi and Vincent for the great working atmosphere in the group, endless cakes, all the help, the discussions and activities also besides daily work. Matthias, Mark, Chris, Daniel, Juri, Ziga, Julien and the table soccer thank you for being real sportsmen. I also like to thank my colleagues Christoph, Sarah and Tobias for the cozy atmosphere during my excursion to the 5th floor. I enjoyed working with all of you.

Rainer, Ebi, Buchse and Pälle thanks to you for always remembering me that there are so many other important things despite work and science.

I am deeply grateful to my parents Annette and Andreas who always supported me in everything I did and encouraged me to follow my own thoughts.

Finally and foremost it is your love and believe in me Steffi, that constantly drives me forward. I am thankful that I can spend my life with you.

Summary

Today's understanding of function and relevance of RNA molecules has been shaped in 1958 by Francis Crick: the so called 'Central Dogma'. It defines RNA as being a messenger between the information storage of the cell the DNA and its executive form, proteins. This central dogma, however, has been extended over the last decades. Besides their central role, several additional biological functions have been attributed to RNA molecules (1.1). In eukaryotes, co-transcriptional processes such as $5'$ capping, splicing and $3'$-processing, are an important part of mRNA maturation. Those processes are dominantly regulated via the direct interaction between RNA and RNA-binding proteins (1.2). Specific interactions between proteins and RNA are established by RNA binding domains (RBD, 1.2). In addition, cellular functions like post-transcriptional generegulation or RNA half-life and decay are mediated and controlled by protein-RNA interactions as well. Hence, a precise analysis of protein-RNA interaction as well as the complex interconnection of their networks is necessary to get a detailed understanding of central biological functions.

In order to detect protein-RNA interactions the recently developed photoactivatable ribonucleoside-enhanced CLIP (PAR-CLIP) approach was adapted to the model system *Saccharomyces cerevisiae* by Carlo Bäjen (AG Cramer). PAR-CLIP offers the precise determination of protein-RNA interaction via UV crosslink-specific induced point mutations based on previously incorporated nucleotide analogues into growing transcripts (1.3.1).

Analysis of current high-throughput next generation sequencing (NGS) data as generated by PAR-CLIP experiments is dependent on sound computational frameworks to formulate and answer biological questions and to gain biological insights out of raw experimental data. Up to now, no software is available covering all necessary steps for PAR-CLIP data analysis. Furthermore, available software solutions for protein-RNA interaction determination or PAR-CLIP measurements (1.3.2) depend on misleading assumptions resulting in high false discovery rates.

This work aimed at producing a pipeline for PAR-CLIP data dubbed as `STAMMP` (2.1) that covers all necessary analysis steps of PAR-CLIP data. `STAMMP` includes pre-process procedures adopted to the needs of PAR-CLIP data analysis (2.2) and a statistical mixture model for reliable detection of PAR-CLIP binding sites based on crosslink induced point mutations (2.3). Additionally, `STAMMP` introduces a normalization procedure of PAR-

CLIP binding sites with RNA-seq data to correct for transcript abundance effects and to obtain binding strength measurements per binding site (2.4) followed by exhaustive post-process analysis steps to directly infer biological results (2.6).

Compared to available protein-RNA interaction site detection methods (1.3.2) `STAMMP` shows a clear improvement in both the number of found sites and the achieved false discovery rate (3.1.3).

Despite the fact, that technical biases are known for related experimental procedures like ChIP-seq researches have begun to investigate possible technical biases affecting PAR-CLIP data only recently. In concordance to current biochemical results the analysis of PAR-CLIP data here indicates, that PAR-CLIP experiments are indeed affected by offtarget effects (3.1.1), sequence biases (3.1.5) and technical background biases (3.1.6), that need to be taken into account during analysis.

The development of `STAMMP` went along with the analysis of 25 PAR-CLIP data sets measured in *Saccharomyces cerevisiae* and revealed insights into the biogenesis of mRNAs (3.2) as well as the transcriptome surveillance machinery in yeast (3.3.2). In brief, the necessity for the correction of transcription abundance (3.2.4), the conserved recognition of pre-mRNA introns (3.2.5), a unified recognition of pre-mRNA polyadenylation sites conserved between yeast and human (3.2.6) and a connection of splicing and $3'$-processing events (3.2.9) could be shown with `STAMMP` among other observations. Additionally, the PAR-CLIP analysis supported the hypothesis for selective termination processes and degradation of ubiquitous ncRNA in order to maintain transcriptome surveillance (3.3.2).

Thus, the newly developed analysis pipeline `STAMMP` was able to overcome previous pitfalls and bottlenecks in PAR-CLIP data analysis and proved a valuable tool for the in-depth evaluation of high-throughput data.

Publications

Parts of this work have been published or are in the process of publication:

2014 **Transcriptome maps of mRNP biogenesis factors define pre-mRNA recognition**

Baejen C*, **Torkler P***, Gressel S, Essig K, Söding J, Cramer P
Mol Cell. 2014 Sep 4;55(5):745-57

Author contributions: C.B. and P.C. designed the study. C.B. established protocols and planned experiments. C.B., S.G., and K.E. performed experiments. P.T., J.S., C.B., and P.C. designed data analysis methods. P.T. and J.S. developed the analysis package. P.T. carried out data analysis. P.C. and C.B. wrote the manuscript with input from all authors. P.C. and J.S. supervised the work.

2013 **Transcriptome surveillance by selective termination of noncoding RNA synthesis**

Schulz D*, Schwalb B*, Kiesel A, Baejen C, **Torkler P**, Gagneur J, Söding J, Cramer P
Cell. 2013 Nov 21;155(5):1075-87

Author contributions: D.S. and P.C. conceived and designed the study. D.S. performed ChIP-seq and 4tU-seq. C.B. performed PAR-CLIP. B.S., D.S., J.S., and J.G. designed data analysis. B.S., A.K., P.T., D.S., and C.B. carried out data analysis. D.S. and P.C. wrote the manuscript with input from all authors. P.C. supervised the project.

* contributed equally.

Contents

List of Figures

List of Tables

1 Introduction

Ribonucleic acid (RNA) is one of the three components of the central dogma of molecular biology proposed by Francis Crick in 1958 [1]. The dogma firstly described the general flow of biological information in life and states the well known cascade of steps to transform the sequentially stored biological information of the deoxyribonucleic acid (DNA) into RNA known as transcription, which can be followed by translation, a process, where messenger-RNA (mRNA) is compiled into protein. Additional to the central dogma, a variety of different functions are nowadays known for RNA besides being mRNA alone. Proteins are interacting with all kinds of RNAs, e.g. to ensure proper RNA production, guiding them to specific cellular components, regulating their decay and protecting them from unwanted digestion [2, 3]. During these processes proteins and RNAs forming ribonucleo-protein (RNPs) and messenger ribonucleo-protein complexes (mRNPs). Therefore, a precise understanding of protein-RNA interactions is needed to get details about central cellular processes. However, only recent developed biochemical technologies provide methods to pinpoint the precise *in vivo* interactions of proteins and RNAs transcriptome wide [4].

In the majority of cases today's research of molecular biology is done in collaborations of biochemists, focusing onto inventions and improvements of biological measurement techniques and computational biologists, focusing onto inventions and improvements in analysis systems and theoretical models [5]. This work introduces methods for analyzing measurements of protein-RNA interactions derived from photoactivatable ribonucleoside-enhanced CLIP (PAR-CLIP) experiments. The remainder of this chapter gives a brief overview about the biological, technical and mathematical background.

1.1 A brief overview about RNA

All genetic information of an organism is located in every nucleus of a cell and is stored in the double-stranded DNA helix composed of the nucleotides adenine (A), guanine (G), cytosine (C) and thymine (T) [6]. Similar to the DNA the polymeric RNA molecule is a chain of the four nucleotides adenine (A), guanine (G), cytosine (C) and uracil (U). As stated by the central dogma, the DNA serves as a template for the generation of messenger RNA (mRNA) containing a copy of the coding sequence of proteins a process known as transcription [1]. According to the dogma, transcription is followed by translation, the

generation of proteins by the ribosome guided by the information encoded in the mRNA.

Nowadays, the general statement that DNA is transcribed into mRNA has been extended by numerous details over the past decades. In brief, transcription starts by unwinding the DNA double-strand by helicases and the genetic information is read in $3' - 5'$ direction by the RNA polymerase (Pol) to generate RNA transcripts in $5' - 3'$ orientation.

During mRNA transcription modifications are made to the transcript by RNA-binding proteins and several quality checkpoints have to be passed which are connected with each other which is known as the biogenesis of mRNAs. Several proteins associated with mRNA maturation are recruited via phosphorylation and dephosphorylation of the C-terminal-domain (CTD) of PolII while other proteins can interact directly with RNA molecules via different RNA binding domains (RBD, see 1.2) [7].

After the transcription machinery is guided to the transcription start site (TSS) via transcription factors (TF) the pre-initiation complex (PIC) assembles and starts transcription. The growing transcript gets protected against rapid degradation starting at the $5'$ end of the transcript by capping enzymes which chemically modify the first nucleotide of the RNA [8, 9]. On the DNA level most eukaryotic genes consists of exons and introns whereas exons contain protein coding sequences (CDS) and introns do not contain CDS. As a consequence, introns have to be removed from transcripts which is known as splicing to ensure proper protein translation [10, 11]. In addition, different exons can be combined to generate different proteins originating from the same TSS. A process known as alternative splicing [12, 13].

Finally, mRNA maturation is completed after the $3'$-end-processing of transcripts which contains the cleavage and release of the mRNA followed by the poly-adenylation of the $3'$-end to protect the transcript from rapid degradation. Successfully matured mRNAs are packaged and transported out of the nucleus into the cytoplasm [14].

Besides classical mRNAs which contain the coding sequence of proteins other classes of RNAs are known which do not carry protein coding sequences and thus are not subsequently transcribed. Four different polymerases exist in eukaryotic organisms whereas each polymerase is responsible for the generation of specific classes of RNAs. RNA polymerase I (PolI) transcribes ribosomal RNA (rRNA), RNA polymerase II (PolII) is responsible for the transcription of mRNA, small nuclear RNA (snRNA) and micro RNA, RNA polymerase III (PolIII) synthesizes transfer RNA (tRNA) and rRNA, and the mitochondrial polymerase (mtPol) transcribes RNA molecules from mitochondrial DNA [15, 16, 17].

The view on the function of RNAs has been expanded dramatically compared to the dominating central dogma of the understanding of the function of RNAs over the last decades. Besides carrying protein coding information RNA molecules can fold into structures, interact with other proteins and can be functional active [18, 19]. In addition,

comparisons of the transcriptome and the genome of organisms revealed that transcription is pervasive throughout the genome which gave rise to non-coding RNA (ncRNA) or long non-coding RNA (lncRNA) while only a minority of the transcribed regions are translated into proteins [20]. The complexity of organisms correlates more with the amount of transcribed ncRNA than with the number of genes [21]. Examples could be found that some ncRNAs are indeed functional and facilitate e.g. gene regulation and protein-protein interactions [22]. Expression of ncRNA is cell type specific and thus, suggesting their importance for specific cellular functions and fate [23].

However, pervasive transcription and ncRNAs leads to a more complex transcriptome which has to be regulated in order to maintain cellular function [24]. Transcription of ncRNAs can overlap with genes leading to interference of gene transcription, and e.g. ncRNAs can inhibit other RNA molecules from translation. Thus, the cell needs mechanisms to distinguish between functional and non-functional RNAs and further systems for the regulation of RNA half-life of different RNA classes. As a consequence different transcription termination pathways for mRNA coding transcripts and ncRNA could be determined recently [25, 26]. In yeast the Nrd1-Nab3-Sen1 pathway is responsible for the rapid degradation of ncRNA while the precise termination of mRNA proteins is still not fully understood [3].

Numerous RNA binding proteins are associated to these highly dynamic regulatory processes and recent discoveries strongly suggest, that regulation of transcripts and the determination of cellular function and fate is regulated post-transcriptional by a network of RNA binding proteins and the crosstalk of these processes [27]. However, most investigations of transcription regulation have been done on the level of DNA analyzing the interaction of transcription factors and gene expression. Most analysis of the interactions of proteins and RNAs are done *in vitro* and only recent developed biochemical methods offer precise whole transcriptome *in vivo* data to study protein-RNA interactions precisely (see 1.3.1).

1.2 Target recognition of RNA-binding proteins

As stated in 1.1 proteins interact in many different ways with all classes of RNA and at all steps of a lifetime of a RNA molecule. In order to elucidate biological functions of RNA-binding proteins (RBPs) and determine their binding specificities a knowledge of how protein-RNA interactions are facilitated is necessary. Contrary to the first idea, that many different RNA-binding domains (RBDs) are responsible for the diverse binding behaviours of RBPs only a few RNA-binding modules facilitate many different functions by integrating multiple RBDs in various structural arrangements into a RBP [28]. Here, a brief introduction about the arrangements, combinations and specificities of RBDs is

given.

Often, a protein consists of multiple RBDs (Figure 1.1A) where each domain for itself shows only a weak sequence specificity. However, by integrating multiple copies of weak domains a protein gains high specificities and affinities through the combination of domains while maintaining high flexibility and regulation by assembling or dissembling them. The modularity of RBPs lead to several effects listed in the following:

- a protein can interact with longer stretches of nucleotides of a RNA (Figure 1.1B)
- distant sequence of a molecule can be bound (Figure 1.1C)
- different RNA molecules can be bound by the same protein (Figure 1.1D)
- a protein can facilitate RNA structure changes (Figure 1.1E)
- a RBP functions as a spacer to facilitate the binding of an additional protein (Figure 1.1G)

The modularity of RBPs is mirrored in evolution. For example, long linkers between two RBDs tend to be as conserved as the domain itself if the precise positioning of the domains is mandatory [29]. Summarized information about the RNA-recognition motif (RRM), the K-homology domain (KH domain) and the zinc fingers (ZF), that are parts of the proteins analyzed in chapter 3 are given here.

From these three classes, the RRM is the best described one which is typically 80-90 amino acids long and forms four stranded anti-parallel β-sheets with two helices packed against it whereas the RNA recognition is usually mediated via the surface of the β-sheet. Usually, a single RRM recognizes RNA sequences between four and eight nucleotides length [30] via three conserved residues consisting of an Arg or Lys that form a salt bridge to the backbone and two aromatic residues that make interactions to nucleobases [28]. The binding of an RRM is mainly, but not limited to RNA sequence recognition only. A few examples show an interaction with other proteins as well.

The KH-domain is typically around 70 amino acids long and can bind to ssRNA and ssDNA. In contrast to the RRM no aromatic residues are present and the binding of typically four nt long stretches is facilitated by hydrogen bonding, shape, and electrostatic interactions.

At last, zinc fingers (ZF) which are well known domains of protein-DNA interactions can also mediate binding of proteins to RNAs and characterized by the arrangement of secondary structures around one or more zinc ions to stabilize the fold. Based on the basis of the residues interacting with zinc the ZF domains are distinguished. Normally a protein consists of multiple zinc finger repeats in order to facilitate sequence specificity.

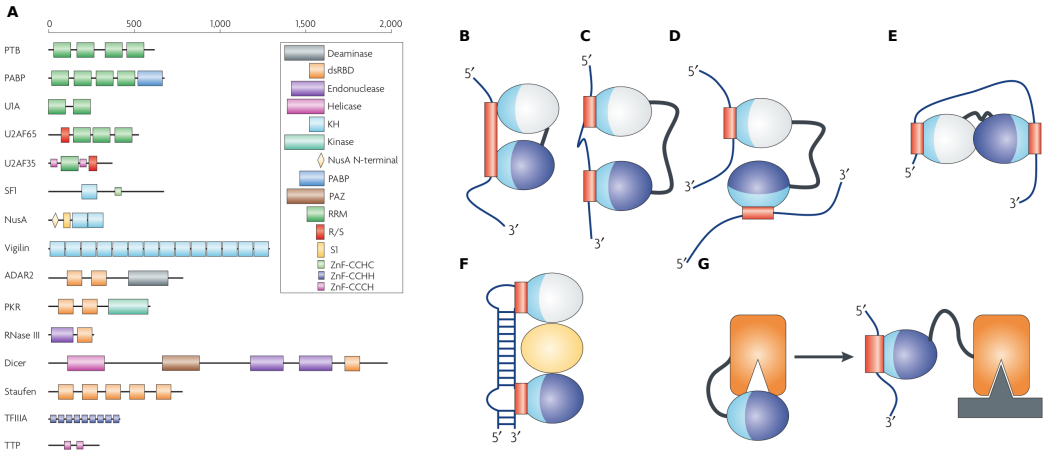

Figure 1.1: **An overview of RBPs modular structure and multiple functionality taken from Lunde et al. [28]**
(A) Demonstration for the variability in the number of copies of RBDs per RBP
(B) Multiple domains are used to enhance sequence specific protein-RNA interactions
(C) Disordered long linkers facilitate the recognition of distant motifs
(D) Disordered long linkers facilitate the recognition of distinct RNA molecules
(E) Rearrangement of the topology of a RNA molecule
(F) RBP are used as spacers to position other proteins in dependence of the RNA molecule
(G) RNA binding coupled to environmental conditions

1.3 Technical background

Current research in system biology and biochemistry often consists of a wet lab part subsequently followed by specific bioinformatic analyses. This study focuses on the bioinformatic analysis part of protein-RNA interaction data obtained by PAR-CLIP experiments. As a consequence only a brief introduction into the biochemical methods and the history of development is given in 1.3.1. A detailed introduction of available computational PAR-CLIP analysis methods is given in 1.3.2.

1.3.1 Cross-linking and immunoprecipitation (CLIP) and its improvements

First whole-genome *in vivo* protein-RNA interaction measurements were published in 2000 using a technique known as RNA immunoprecipitation (RIP) chip [31]. In principle a RNA binding protein is immunoprecipitated and isolated together with the bound RNA. Bound RNA is reverse transcribed into cDNA followed by microarray analysis. Although first insights into protein-RNA interactions were achieved by RIP-chip this technique has several drawbacks. Precise binding sites cannot be detected due to low resolutions, microarray measurements are error-prone and only stable mRNPs can be detected. To circumvent these shortcomings and to get more precise protein-RNA measurements a method called ultraviolet (UV) crosslinking and immunoprecipitation (CLIP) was developed [32, 33].

CLIP utilizes the fact, that *in vivo* crosslinks between protein-RNA interactions are induced only at sites of direct contact between proteins and RNAs using UV light. After crosslinking and subsequent cell lysis, the protein of interest gets immunoprecipitated and bound RNA is reverse transcribed into cDNA. In contrast to RIP-chip the cDNA fragments were subjected to Sanger sequencing and mapped to the genome to infer protein-RNA interactions.

The Sanger method used in the original publication was replaced by modern high-throughput sequencing technologies to get more precise data due to higher sequencing depth and is known as high-throughput sequencing of CLIP cDNA library (HITS-CLIP or CLIP-seq, Figure 1.2 middle) [34]. The resolution of CLIP and HITS-CLIP is dominated by the length of cDNA fragments subjected for sequencing.

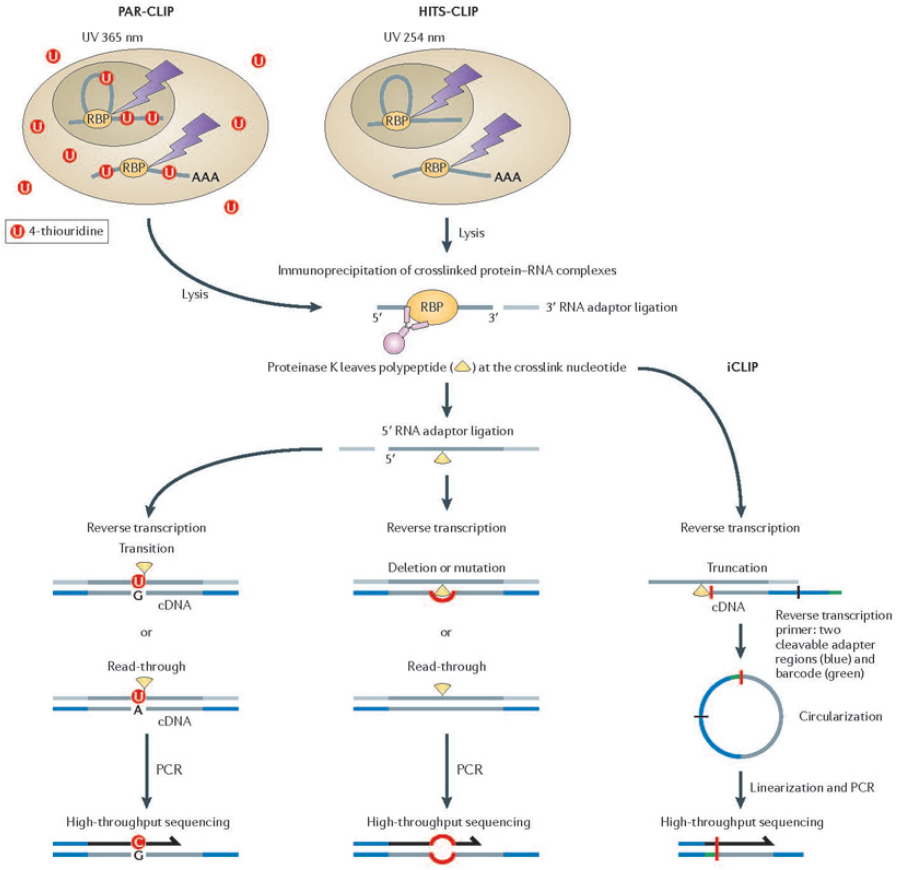

Figure 1.2: **Overview of the protein-RNA detection methods PAR-CLIP, (HITS)-CLIP and iCLIP taken from [4]**
For PAR-CLIP cells are labeled with 4-thiouridine and protein-RNA interactions are crosslinked with 365nm UV light. In contrast, all other CLIP protocols avoid the usage of nucleotide analogues, but use 254nm UV light. Cell lysis, immunoprecipitation of the crosslinked protein-RNA complexes, adapter ligation and proteinase K digestion are performed in every protocol. After reverse transcription and sequencing PAR-CLIP utilizes crosslink-induced specific point mutations for crosslink site detection while HITS-CLIP searches for deletions or mutations and iCLIP leverages cDNA truncation sites.

Along with HITS-CLIP another method for the study of *in vivo* protein-RNA interactions was developed by Granneman et al. known as CRAC [35]. Granneman could show, that mutation and deletion events in reads are introduced by crosslinks and can be used to pin point protein-RNA interaction sites precisely.

To further improve the resolution of protein-RNA interactions two recent approaches were developed. The photoactivatable ribonucleoside-enhanced CLIP (PAR-CLIP) approach uses nucleotide analogues like 4-thiouridine (4-SU) or 6-thioguanosine (6-SG) to further enhance the CLIP procedure (Figure 1.2 left) [36]. Compared to the original CLIP protocol the usage of 4-SU or 6-SG allows to use UV-light with a wavelength of 365nm instead of 256nm leading to lower levels of noise due to the fact, that crosslinks only occur between nucleotide analogues and proteins and not between other nucleotides and proteins. In addition, nucleotide analogues cause specific point mutations during the reverse transcription(Figure 1.2 left). 4-thiouridine leads to specific T→C mutations and 6-thioguanosine leads to G→A mutations. These artificially introduced mutations can be utilized for a precise protein-RNA binding site detection as described in 1.3.2 and 2.3. PAR-CLIP induces more and specific mutations compared to CRAC and HITS-CLIP which leads to a clear distinction between true and false protein-RNA interaction sites.

The second recently published approach for nucleotide resolution measurements of protein-RNA interactions is known as individual nucleotide resolution CLIP (iCLIP) [32, 33]. In comparison to PAR-CLIP no nucleotide analogues are used. Reverse transcription can be aborted at crosslink sites due to a drop off of the polymerase caused by peptide rests that remain at the crosslink site after proteinase K digestion. The resulting truncated fragments would get lost in the standard CLIP protocol. By using an adapter which consists of a $3'$-primer, cleavage-site and $5'$-primer these fragments can be captured, circularized, linearized (Figure 1.2 right) and subjected to sequencing. The sequencing results are searched for truncation sites indicating the position of the crosslink.

1.3.2 Methods for binding site detection in PAR-CLIP data

Reliable binding site detection is one of the key components of an analysis pipeline for PAR-CLIP data in order to separate signal from noise and to get reliable information for subsequent analysis in post-processing steps. This section gives a detailed overview about published methods for PAR-CLIP binding site detection and focuses on the mathematical background of each presented method. More general information about CLIP data analysis are covered in 2.1.

The question how to identify reliable binding sites in PAR-CLIP data was first addressed in the original PAR-CLIP publication by Hafner et al. [36]. Their preliminary analysis revealed that 4-SU induces T→C mutations which

- are the most observed mutations in the data sets

- are most of the times inside or in close vicinity of binding motifs (Figure 1.3A,B)

- lead to the strongest motif enrichments, if binding sites are sorted according to the number of T→C mutations (Figure 1.3C).

Hafner et al. selected all read clusters with at least two crosslink positions and, second, ranked all clusters by the total number of T→C mutations (see supplementary information in [36]) to separate true binding sites from noise. All subsequently published binding site detection methods are based on the T→C characteristics first described and utilized by Hafner.

Most published methods for binding site detection consist of a two step process where one step utilizes T→C mutations and a second step exploits the read coverage. In the following this chapter gives a specific introduction into the available PAR-CLIP analysis methods `naive`, `PARalyzer`[37], `wavClusteR`[38] and `PIPE-CLIP`[39].

1.3.2.1 Naive approach - Hafner like

The simplest approach to define a decision criterion for PAR-CLIP binding site detection is to introduce the thresholds δ_r for the number of reads r_i and δ_m for the number of T→C mutations m_i at genomic position i. In addition to these criteria only reads that mapped to annotated mRNAs were analyzed in the original publication [36]. In this work the naive approach is defined that a genomic position is considered as a true binding site

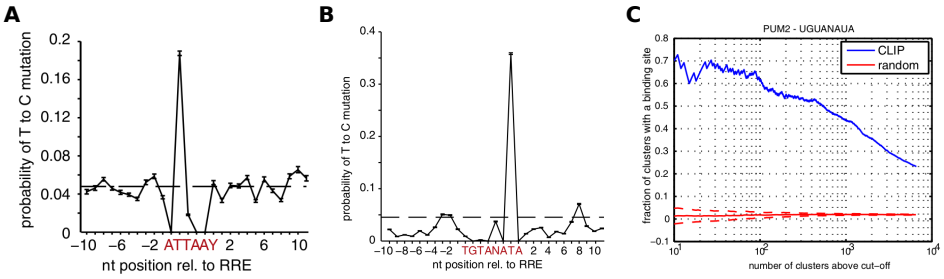

Figure 1.3: **T→C characteristics taken from [36]**
The figures were taken from [36]
(A) T→C positional mutation frequency for PAR-CLIP clusters anchored at the ATTAAY RNA-recognition element (RRE) show the highest probability of T→C mutations at the first T in the motif. The T→C shows clear enrichment compared to the background (dashed line)
(B) T→C positional mutation frequency for PAR-CLIP clusters anchored at the 8-nt recognition motif from all motif-containing clusters. Although, three possible T could be used for crosslink facilitation only the last T is preferred for crosslinking.
(C) The fraction of clusters containing the binding motif of the analyzed factor PUM2 is shown on the y-axis. All PAR-CLIP sites are sorted according to the number of observed T→C mutations (x-axis). A dependency of a decreasing fraction of clusters containing the binding motif and the amount of observed T→C mutations can be seen.

if

$$r_i > \delta_r \quad \text{and} \tag{1.1}$$

$$m_i > \delta_m \tag{1.2}$$

regardless at which genomic position a read was mapped.

1.3.2.2 PARalyzer

`PARalyzer` [37] was the first bioinformatic software specifically introduced for analyzing PAR-CLIP data and was used in different PAR-CLIP studies indicating that the RNA-binding protein HuR couples pre-mRNA processing and mRNA stability in human [40] and analyzing the binding behavior of LIN28 in human [41].

In brief, the model proposed by `PARalyzer` utilizes experimentally introduced T→C mutations with two kernel-density estimates. One estimator for T→C mutation events and one estimator for T→T non-mutation events. Read groups exceeding a previously specified number of minimum reads and where the likelihood of T→C events as defined by the kernel-density estimator is higher than T→T events are considered as binding sites.

In detail, for a read group of length L `PARalyzer` defines $x^{(i)}_{\mathrm{T}\to\mathrm{C}}$ and $x^{(i)}_{\mathrm{T}\to\mathrm{T}}$ with $i \in \{1,\ldots,L\}$ as the number of observed T→C and T→T events at position i. The values of $n_{\mathrm{T}\to\mathrm{C}}$ and $n_{\mathrm{T}\to\mathrm{T}}$ are defined as the total number of T→C and T→T events within the read group. Thus, the likelihoods for T→C and T→T events within a read group are defined as:

$$f_{\mathrm{T}\to\mathrm{C}}(j) = \sum_{i=1}^{L} \frac{x^{(i)}_{\mathrm{T}\to\mathrm{C}}}{n_{\mathrm{T}\to\mathrm{C}}} \times \frac{1}{\sqrt{2\lambda^2\pi}} e^{-\frac{||i-j||^2}{2\lambda^2}} \tag{1.3}$$

$$f_{\mathrm{T}\to\mathrm{T}}(j) = \sum_{i=1}^{L} \frac{x^{(i)}_{\mathrm{T}\to\mathrm{T}}}{n_{\mathrm{T}\to\mathrm{T}}} \times \frac{1}{\sqrt{2\lambda^2\pi}} e^{-\frac{||i-j||^2}{2\lambda^2}} \tag{1.4}$$

with $j \in \{1,\ldots,L\}$. A Gaussian kernel with globally fixed parameter $\lambda = 3$ was used for class-specific density estimation (see Material and Methods in [37] for more details).

Equations 1.3 and 1.4 are then used to produce a non-parametric estimate for T→C conversion and T→T non-conversion events, respectively:

$$k_{\mathrm{T}\to\mathrm{C}}(j) = \frac{f_{\mathrm{T}\to\mathrm{C}}(j)}{\sum_{i=j}^{L} f_{\mathrm{T}\to\mathrm{C}}(j)} \tag{1.5}$$

$$k_{\mathrm{T}\to\mathrm{T}}(j) = \frac{f_{\mathrm{T}\to\mathrm{T}}(j)}{\sum_{i=j}^{L} f_{\mathrm{T}\to\mathrm{T}}(j)} \tag{1.6}$$

Thus, genomic position j for which $k_{\mathrm{T}\to\mathrm{C}}(j) > k_{\mathrm{T}\to\mathrm{T}}(j)$ are defined as binding sites. An

example of the class-specific density estimations for T→C (red) and T→T (blue) events is given in Figure 1.4.

1.3.2.3 wavClusteR

In 2012 Sievers et al. introduced the Bioconductor [42] package `wavClusteR` for binding site detection in PAR-CLIP data based on the analysis of T→C transitions [38]. `wavClusteR` was the first method that suggested to distinguish between experimentally and non experimentally induced mutations in PAR-CLIP data for reliable binding site detection. In general, the method uses a two step process. First, a non-parametric two-component mixture model describing experimentally and non experimentally induced mutations is used to find reliable binding sites. Second, a wavelet-based peak calling is used for the determination of peak boundaries. The original publication describes a continuous wavelet transform (CWT) for peak boundary definition which was replaced by the Mini-Rank Norm (MRN) algorithm in later versions of the Bioconductor package (see actual `wavClusteR` vignette for details).

According to the original publication let $\mathcal{A} = \{A,C,G,T\}$ be the nucleotide alphabet and $\mathcal{S} = \{(g,r)|g,r \in \mathcal{A} \wedge g \neq r\}$ be the set of substitutions of any base g in the reference genome to any other base r in the read. Next, the relative substitution frequency (RSF) for each genomic position i and observed substitution s is introduced as:

$$\hat{x}_{s,i} = \frac{y_{s,i}}{z_i}, s \in \mathcal{S}, \tag{1.7}$$

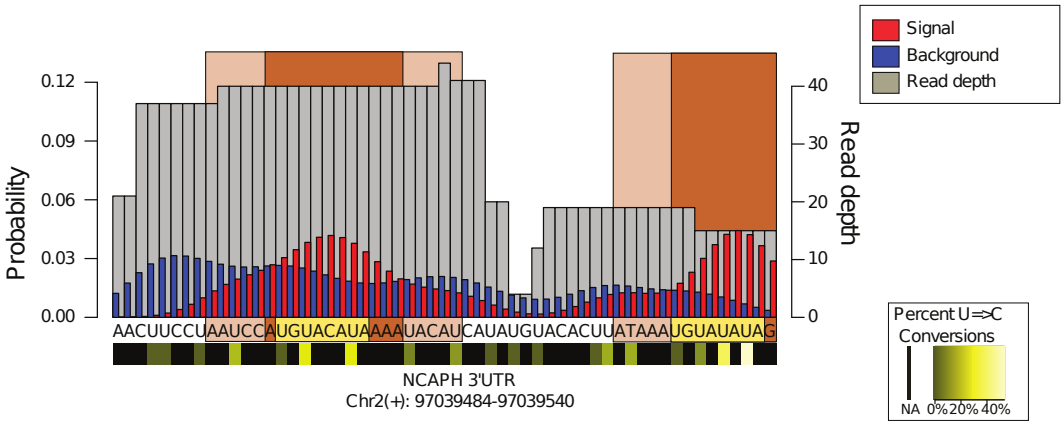

Figure 1.4: **Example of `PARalyzer` interaction site identification**
The figure was taken from [37] to illustrate the binding site detection proposed by `PARalyzer`. The entire genomic region corresponds to a single read-group from the Pumilio2 library. The orange region represents the nucleotides where the T→C signal kernel density estimate is above T→T background. The light pink locations are the full interaction sites extended by up to 5 nucleotides. A light gold box highlights the sequences that match the known Pumilio2 binding motif.

where z_i is the total coverage at position i and $y_{s,i}$ the number of observed substitution s at position i. The number of substitutions $y_{s,i}$ is binomially distributed and parameterized by z_i and $x_{s,i}$, whereas x_s is distributed according to a probability density function (PDF) $p_s, x_s \sim p_s$ consisting of a two component mixture to model experientally and non experimentally induced mutations.

The overall goal of the mixture model is to define a lower bound α and an upper bound β for RSF values in such a way, that a RSF at position i is considered as experimentally induced and thus likely to be true if $\alpha > x_{TC,i} < \beta$.

The mixture model p_s is described as:

$$p_s(x) = \underbrace{\lambda_{s,1} p_{s,1}(x)}_{\text{non experientally}} + \underbrace{\lambda_{s,2} p_{s,2}(x)}_{\text{experientally}}, \tag{1.8}$$

where $\lambda_{x,k}$ are the mixing coefficients with $\lambda_{s,k} \geq 0$, $\sum_k \lambda_{s,k} = 1$ and k being the component index so that $k = 1$ is referred to the non experimentally induced part and $k = 2$ is referred to the experimentally induced part. In order to find reliable binding sites the following expression has to be solved:

$$p_{TC}(x) = \underbrace{\lambda_1 p_1(x)}_{\text{non experientally}} + \underbrace{\lambda_2 p_2(x)}_{\text{experientally}}, \tag{1.9}$$

According to [38] equation 1.9 can be treated as a binary classification problem answering the question when p_{TC} is dominated by $\lambda_2 p_2(x)$. Thus, the authors derive the posterior class probability:

$$P(K = 2|X = x) = \frac{\lambda_2 p_2(x)}{\lambda_1 p_1(x) + \lambda_2 p_2(x)} \tag{1.10}$$

which can be rewritten to:

$$P(K = 2|X = x) = \frac{p_{TC}(x) - \lambda_1 p_1(x)}{p_{TC}(x)} \tag{1.11}$$

by using equation 1.9 or as the log-odds ratio:

$$\log \frac{P(K = 2|X = x)}{P(K = 1|X = x)} = \log \frac{\lambda_2 p_2(x)}{\lambda_1 p_1(x)} \tag{1.12}$$

with parameters $\lambda_1, \lambda_2, p_1(x), p_2(x)$.

After parameter estimation `wavClusteR` uses equation 1.10 to calculate the posterior probability of an observation being generated by the experimentally component. Next,

`wavClusteR` reports a lower and an upper bound defined as 'support' so that a genomic position i is considered as a binding site if the RSF at position i is within the calculated boundaries (see Figure 1.5B).

See Sievers et al. [38] for more details about the estimation of their proposed PDFs. Here, only the proposed estimation for the mixing coefficients λ_k is recapped for later discussions. In order to estimate the mixing coefficients Siever et al. introduce the count function f defined as:

$$f : \mathcal{S} \rightarrow \mathbb{N}_0, f(s) = \sum_{j=1}^{G} \mathrm{I}(s,j) \tag{1.13}$$

with

$$\mathrm{I}(s,j) = \begin{cases} 1, & \text{if } s \text{ is observed at least once at position } j \\ 0, & \text{otherwise} \end{cases} \tag{1.14}$$

where G equals the size of the genome. The function $f(s)$ indicates the number of genomic positions with at least one substitution s. The mixing coefficient λ_2 for the weight of experimentally induced mutations is then defined as:

$$\hat{\lambda}_2 = \frac{f(TC) - \tilde{f}}{f(TC)}, \tilde{f} = arg\,max f(n) \tag{1.15}$$

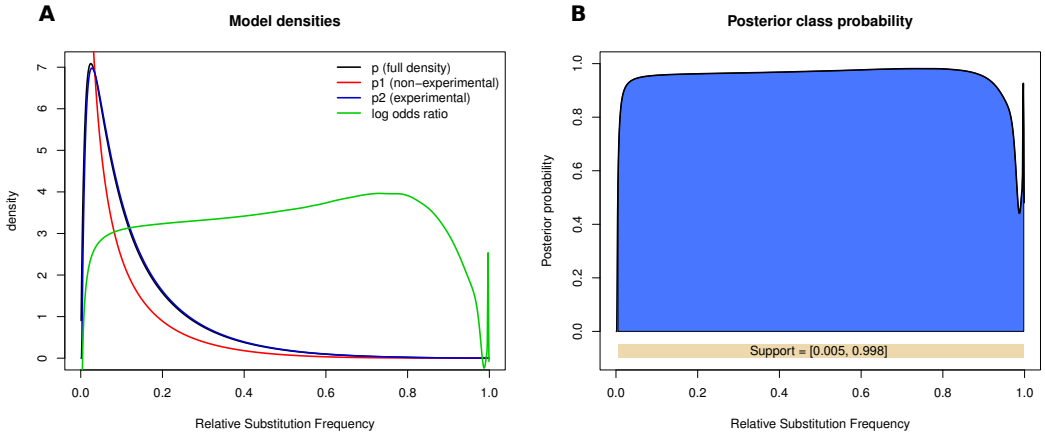

Figure 1.5: **Model Estimations and Posterior Class Probabilities Estimated by `wavClusteR`**
The plots were generated with the original `wavClusteR` package and the Nrd1 data of chromosome 4 from [24].
(A) Shown are the estimated densities for p,p_1,p_2 of equation 1.10 and the density of equation 1.12 calculated by the `fitMixtureModel` function of `wavClusteR`
(B) The resulting posterior class probability of equation 1.10 is shown, i.e. the probability that a given relative substitution frequency (RSF, horizontal axis) has been experimentally induced. The area under the curve corresponding to the returned RSF interval is colored, and the RSF interval indicated.

1.3.2.4 PIPE-CLIP

Likewise to `wavClusteR` the binding site detection proposed in `PIPE-CLIP` uses a two step process [39]. In brief, `PIPE-CLIP` finds enriched clusters of reads via a zero truncated negative binomial model and selects binding sites using a model for crosslink induced mutations. Only positions which pass both steps successfully are reported as reliable binding site. In contrast to the other presented methods `PIPE-CLIP` can be used to analyze PAR-CLIP, HITS-CLIP and iCLIP data.

For the selection of binding sites in PAR-CLIP data `PIPE-CLIP` focuses on experimentally induced T→C mutations. In detail, each genomic position i is characterized by the total number of mapped reads k_i and the number of observed T→C mutations m_i. In `PIPE-CLIP` m_i is modeled by a binomial distribution parameterized by k_i and the success rate τ and thus defined as:

$$p(m_i|\tau,k_i) = \binom{k_i}{m_i} \tau^{m_i}(1-\tau)^{(k_i-m_i)}, \tag{1.16}$$

where the success rate τ is defined as k_i/`genomsize`. The probabilities obtained by the binomial distribution are then used to calculate p-values to access the statistical significance of the mutation. In a final step, the p-values are used to determine false discovery rates via the Benjamin-Hochberg method and genomic positions with FDRs less than a user-specified threshold are reported [43, 39].

2 Methods

The details of the computational pipeline for PAR-CLIP data analysis are presented in this chapter. After a general introduction about the fundamental analysis steps in 2.1 each step is explained in detail in the remainder of this chapter.

The PAR-CLIP experiments which served as input for the development of the analysis pipeline presented here where performed by Carlo Bäjen in the model organism *Saccharomyces cerevisiae*. All data presented in this work is meanwhile available under the GEO accession number GSE59676 and published in Bäjen and Torkler et al. [44], as well as in Schulz and Schwalb et al. [24]. Note, that the photoactivalable nucleotide analogue 4-SU was replaced in the yeast protocol by 4-thiouracil (4-tU). This chapter only contains information regarding data analysis of PAR-CLIP data. For a detailed description of the biochemically methods used in the yeast protocol please take a look in the supplementary material of the publications [44, 24].

To avoid massive over-fitting of the pipeline development to the in-house generated data, I also took a close look at the Nrd1/Nab3 yeast data previously published by Creamer et al. [45] especially during the beginning of the development.

2.1 General workflow of PAR-CLIP data analysis

Basically the analysis of high-throughput next generation sequencing (NGS) data as generated by CLIP experiments consists of the four basic building blocks

1. Pre-Processing
2. Binding Site Detection
3. Data Normalization
4. Post-Processing

and are typically presented as a pipeline which covers all necessary steps to get from raw sequencing data to interpretable data representations as well as data quality measurements (Figure 2.1). This work presents software solutions for each building block whereas each block is implemented separately to facilitate high flexibility to users. The pipeline is mainly written in Python 3 with dependencies on SciPy and R [46, 47].

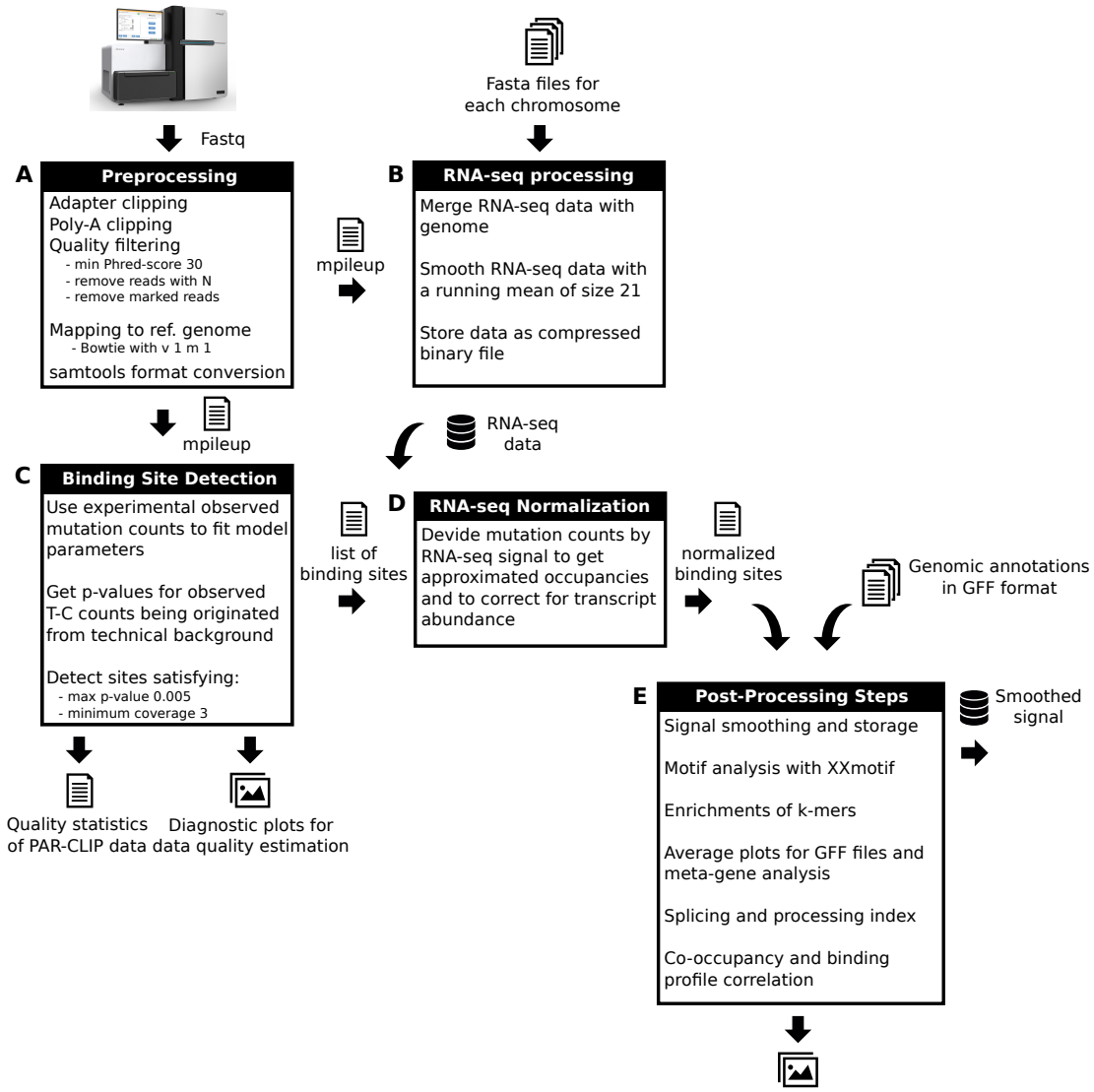

Figure 2.1: **General pipeline overview**
(A) The pre-processing step takes the raw sequence information in fastq format as input and covers all raw data processing steps like adapter clipping, filtering of low quality reads and reference mapping. The pre-processing is used for both PAR-CLIP sequencing data and RNA-seq data and stores the processed data into the pileup format offered by SAMtools.
(B) RNA-seq processing takes the pileup from (A) as well as fasta files for each genomic chromosome as input. Next, the RNA-seq signal is smoothed and stored in binary format for the normalization step (D)
(C) The statistical model from 2.3 detects true binding sites in the pileup generated by (A). Binding sites are stored in a simple tab delimited text format and diagnostic plots for mutation events as well as quality information are stored.
(D) Detected binding sites data from (C) are normalized with the smoothed RNA-seq data from (B).
(E) Normalized binding sites and genomic annotations in GFF format serve as input for various post-processing scripts which return interpretable plots listed in the box.

In general the pre-processing block is responsible for basic raw data handling of the sequencing data to remove adapter sequences, filter out low quality reads and map the reads back to the reference genome. Normally, this block consists of a sequence of widely used third-party software. In 2.2 a detailed list of the used programs is given as well as explanations for parameters settings especially adapted for PAR-CLIP experiments.

The binding site detection method is the key component of an analysis pipeline to separate signal from noise. For this purpose a statistical model is introduced which gives the probability that an observed number of mutations (see chapter 1.3.1 and Figure 1.2) is caused by technical errors rather then by crosslink. A detailed description of the model is given in 2.3.

Normalization of CLIP data is necessary to correct for transcript abundance and is highlighted in 2.4.

At last the normalized binding sites can be subjected to further analysis to get biological insights of the analyzed protein. The plots and analysis currently supported by the pipeline are explained in 2.6.

2.2 Pre-processing

The pre-processing steps in `STAMMP` consist of a sequence of third-party software adapted to the needs of PAR-CLIP data analysis. The first three pre-processing steps are performed by in-house developed tools from Alexander Graf.

First, adapter sequences are removed from the sequencing reads. Next, $3'$-ends of sequencing reads are scanned for long poly-A stretches which are also clipped. Thus, reads from already poly-adenylated mRNAs can also be captured and further processed. Third, reads with poor quality are removed. A rigorous quality filtering is necessary for PAR-CLIP data, because PAR-CLIP is based on specifically introduced mutations. All reads are removed which have

- at least one nucleotide with a Phred score < 30
- at least one unknown nucleotide 'N'
- a mark set by the illumina internal quality check.

These strict filter parameters enhance the likelihood that observed T$\rightarrow$C mutations in reads can be trusted and are more likely to be true mutations instead of bad sequencing signals. Weaker filter parameters could lead to higher amounts of noise.

Clipped and filtered reads are then subjected to an alignment procedure to map the reads to the reference genome of the underlying organism. Besides the filtering step the alignment procedure is crucial for PAR-CLIP data. The precise handling of mutations during the alignment step must be considered carefully for PAR-CLIP data analysis to

avoid mistakes. Due to a large number of parameters for the alignment procedure the mutations from the PAR-CLIP signal can be lost during this step easily. Currently, `Bowtie` is used for the alignments of the reads [48]. In default settings `Bowtie` uses the `−n alignment mode` which is coupled to the `−e` parameter. The following criteria are used for the alignment of reads in default settings:

- alignments may not have more than `−n` mismatches
- the sum of the quality scores of all mismatches may not exceed `−e`

Due to the fact that PAR-CLIP data benefits from artificially introduced mutations with high Phred-scores `Bowtie`'s parameters are set to the `−v alignment mode` which is mutually exclusive from the `−n` mode to avoid problems. Here, no Phred-scores are considered during the alignment procedure and only alignments with up to `−v 1` mismatches which map uniquely to the genome `−m 1` are reported. The score independent alignment procedure coupled with the rigorous quality filtering at the beginning is well suited for the alignment of PAR-CLIP reads.

Finally, the uniquely mapped reads are converted into the `pileup` format using `SAMtools` [49]. After file format conversion the `pileup` is passed to the binding site detection explained in the following chapter.

2.3 A novel statistical mixture model for binding site detection in PAR-CLIP data

As stated in 1.3.2 the identification of reliable binding sites is a crucial part of PAR-CLIP data analysis as it forms the basis for all subsequent analysis steps. Here, I introduce the statistical model for binding site selection implemented in `STAMMP` - a statistical mixture model for PAR-CLIP data - which distinguishs between experimentally induced and non experimentally induced mutations. Compared to the published methods summarized in 1.3.2 the general assumption of `STAMMP` is related to the general assumption of `wavClusteR`. However, the model used in `STAMMP` to describe non experimentally induced mutations as well as the way to finally find reliable binding sites is different. The model proposed by `STAMMP` is used to calculate the probability that an observed number of mutations at a specific genomic position is caused by technical errors rather than from crosslink induced mutations.

Let $\mathcal{A} = \{A,C,G,T\}$ be the nucleotide alphabet and $\mathcal{S} = \{(g,r)|g,r \in \mathcal{A} \wedge g \neq r\}$ be the set of all possible mutations. The number of an observed mutation $s \in \mathcal{S}$ at genomic position i is defined as m_i^s and the number of mapped reads at position i is r_i. As a consequence of this, PAR-CLIP measurements consists of data (m_i^s, r_i) for both T→C transitions as well as every other mutation defined by $\mathcal{S}$. However, the

assumption in STAMMP is, that (m_i^{TC}, r_i) data is a mixture of experimentally induced and non experimentally induced mutations whereas $(m_i^{s'}, r_i)$ data with $s' \in \mathcal{S} \setminus (T,C)$ is exclusively generated by error sources. In general, there are two rough sources of errors. First, mutations can be induced by technical errors like sequencing-, PCR- or alignment-errors and second, errors can be caused by sequence polymorphisms in the analyzed cells resulting in SNPs. Due to the nature of the two origins of errors it is reasonable to assume that mutations s' can be utilized to fit the negative distribution $p(m_i^{s'}|r_i, \theta)$ with parameters θ to define the probability to get m_i^s mutations due to everything except the photoactivated crosslink.

The source of mismatches (MM) to the reference genome caused by technical errors can be described by a beta binomial distribution

$$\mathrm{Bb}(m_i^{s'}|r_i, \alpha_0, \alpha_1) = \binom{r_i}{m_i^{s'}} \frac{\mathrm{B}(m_i^{s'} + \alpha_0, r_i - m_i^{s'} + \alpha_1)}{\mathrm{B}(\alpha_0, \alpha_1)} \tag{2.1}$$

where B is the beta function with the form

$$\mathrm{B}(\alpha_0, \alpha_1) = \frac{\Gamma(\alpha_0)\Gamma(\alpha_1)}{\Gamma(\alpha_0 + \alpha_1)} \tag{2.2}$$

The mismatches due to polymorphisms in the population can be described by a binomial distribution

$$\mathrm{Bin}(m_i^{s'}|r_i, \rho) = \binom{r_i}{m_i^{s'}} \rho^{m_i^{s'}} (1-\rho)^{r_i - m_i^{s'}} \tag{2.3}$$

with ρ being the rate of mismatches due to SNPs.

Assuming a relative weight ω of sequencing errors the negative distribution is given by

$$p(m_i^{s'}|r_i, \alpha_0, \alpha_1, \rho, \omega) = \omega\, \mathrm{Bb}(m_i^{s'}|r_i, \alpha_0, \alpha_1) + (1-\omega)\, \mathrm{Bin}(m_i^{s'}|r_i, \rho) \tag{2.4}$$

with parameters $\theta = \{\alpha_0, \alpha_1, \rho, \omega\}$.

In order to calculate the negative distribution given in equation 2.4 the unknown parameters θ have to be specified. A solution to this problem is by maximizing the log-likelihood of $p(m_i^{s'}|r_i, \theta)$. This corresponds to choosing the values of θ for which the probability of the observed data is maximized [50]. The log likelihood of equation 2.4 is given by

$$\mathrm{LL}(m_i^{s'}|r_i,\theta) = \log \prod_{i\in\mathrm{Genome}} p(m_i^{s'}|r_i,\theta) \rightarrow \max \tag{2.5}$$

$$= \sum_{r=1}^{\infty} \sum_{m^{s'}=0}^{r} f(m^{s'},r) \log p(m^{s'}|r,\theta) \rightarrow \max \tag{2.6}$$

where $f(m,r)$ is a count function which returns the number of observed $(m^{s'}, r)$ mismatches in a data set. Thus, the count function is defined as

$$f(m^{s'}, r) = \sum_{i\in\mathrm{Genome}} \mathrm{I}(m_i^{s'} = m^{s'} \wedge r_i = r) \tag{2.7}$$

with I being the indicator function.

To optimize the log-likelihood function the quasi-Newton method of Broyden, Fletcher, Goldfarb, and Shanno known as BFGS as offered in the SciPy package is used [51, 52, 53, 54, 46]. A re-parametrization according to

$$\begin{aligned} \alpha_0 &= \exp^{\alpha_0} \\ \alpha_1 &= \exp^{\alpha_1} \\ \rho &= \frac{1}{1+\exp^{\rho}} \\ \omega &= \frac{1}{1+\exp^{\omega}} \end{aligned}$$

assures, that the parameters are estimated according to their domains. Equation 2.5 is optimized separately for the mutation data for $r \in \{5,\ldots,20\}$ if enough data is available. Start parameters are randomly chosen for each initialization and the optimization is performed at least 100 times per r to prevent sticking into local optima. Lastly, the parameters that lead to the best log-likelihoods per r are averaged and used as global parameters for the p-value calculation.

After parameter estimation equation 2.4 is used to calculate the probability that the observed number of T→C mutations m_i^{TC} is caused by technical errors. To speed up the binding site detection the probability distributions are pre-calculated and stored. Next, p-values are calculated for each genomic position with T→C mutations based on equation 2.4 where the p-value is given by

$$p(M \geq m_i^{TC}|r_i,\theta) = \sum_{j=m_i^{TC}}^{r_i} p(j|r_i,\theta) \tag{2.8}$$

Genomic positions with a p-value less or equal than a defined cutoff (default: 0.002) are reported as protein-RNA binding site.

2.4 Normalization of PAR-CLIP binding sites with RNA-seq data to correct for transcript abundance

PAR-CLIP data or in general CLIP data depends on the expression and the transcript lifecycle of the underlying transcripts [55, 4]. The higher the expression of a transcript the more signal is measured in CLIP and PAR-CLIP experiments. Therefore, a normalization procedure to correct for transcript abundance is mandatory to compare binding sites in a quantitative manner (Figure 2.2).

The suggestion of Kishore et al. to correct CLIP and PAR-CLIP counts by mRNA expression levels measured by RNA-seq experiments is used in `STAMMP` [55]. However, the RNA-seq data for PAR-CLIP data normalization should be generated under the same conditions to avoid expression biases introduced by the PAR-CLIP protocol.

To correct for transcript abundance the T→C mutation counts m_i^{TC} of binding sites reported by the statistical model (2.3) are divided by an averaged number of RNA-seq reads a_i. In default settings, `STAMMP` uses a moving average with a nucleotide width of $w = 10$ to slightly smooth the RNA-seq read counts r_i^{RNA} so that

$$a_i = \frac{1}{2w+1} \sum_{j=i-w}^{i+w} r_j^{RNA} \tag{2.9}$$

.

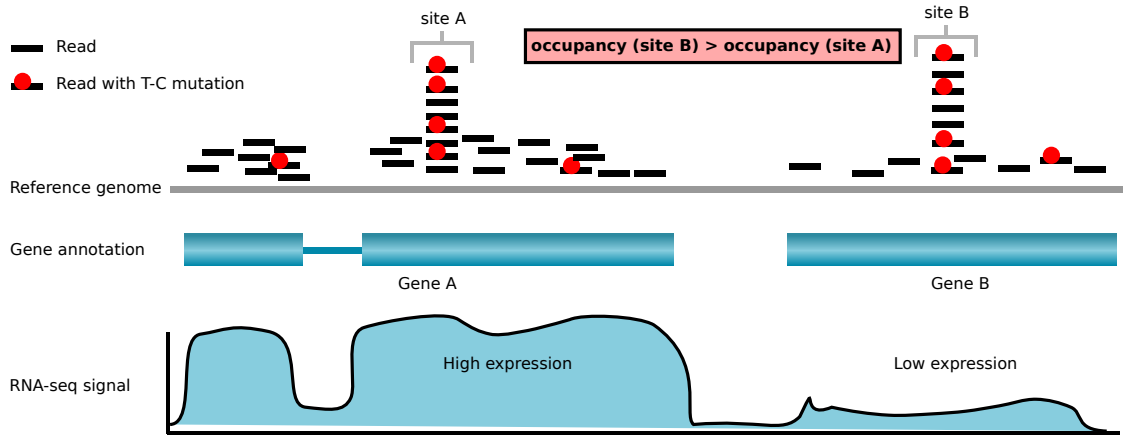

Figure 2.2: **Normalization scheme for PAR-CLIP data in dependence on [4]**
A schematic overview for the normalization of PAR-CLIP data by RNA-seq data to correct for transcript abundance is shown. `STAMMP` divides the observed T→C mutation counts by the number of RNA-seq transcripts mapped at that position. Thus, the binding at site B is stronger compared to site A despite the fact, that the number of total counts for site A and B is identical.

The normalized m_i^{TC} counts are expressed as the occupancy of a factor to a site given by

$$occ(m_i^{TC}) = \frac{m_i^{TC}}{a_i} \tag{2.10}$$

Thus, the affinity of a protein to a detected binding sites is determined in natural way. RNA-seq data used for normalization in this work has been generated under the same conditions like the PAR-CLIP experiments in order to derive read count data as close as possibile to the PAR-CLIP induced conditions.

2.5 Generation of precise annotation data

Post-processing analysis of sequencing data strongly depends on precise annotation data to reveal binding patterns of proteins. The PAR-CLIP data analyzed in this work was generated in *Saccharomyces cerevisiae* and the most complete and accurate *S.cer* annotation accessible so far is based on ChIP-chip experiments and published by Xu et al. in 2009 [56]. However, the same laboratory published precise transcript-isoform (TIF) data of yeast based on sequencing data in 2013 which showed a lot of transcript variants per annotated transcript [57]. In order to get more precise annotations compared to the 2009 ChIP-chip data the TSS and pA positions of the most dominant isoform per annotation were taken instead of the older ChIP-chip based TSS and pA positions (Figure 2.3).

2.6 Post-Processing steps in `STAMMP`

Previous to post-processing the normalized protein occupancies are aligned to the their genomic locations and are slightly smoothed with a moving average according to equation 2.9 with the differences, that $w = 5$ and r_j^{RNA} is replaced by $occ(m_i^{TC})$. Next, occupancies above the 97% quantile are set to 100% to fix the scale and make the measurements comparable between different proteins. After smoothing and rescaling the final genome-wide occupancies are stored in compressed binary format and are ready for further analysis.

2.6.1 Mutation rate analysis for data quality assurance

`STAMMP` offers exploratory plots in order to check

- if the incorporation of the 4tU during the PAR-CLIP experiment was successful
- if the crosslink induced T→C mutations are more likely than other mutations
- and therefore to validate if the experiment was successful at all

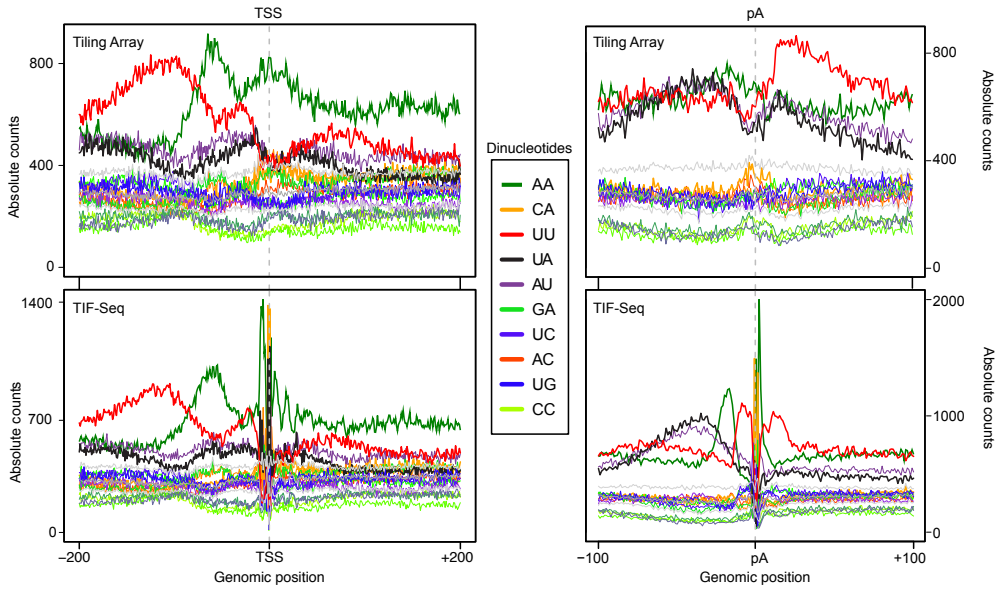

Figure 2.3: **Comparison of TSS and pA annotations from tiling array and TIF-Seq data, respectively**
To annotate TSS and pA sites, recent TIF-Seq data from [57] was used which yielded much sharper sequence features around TSS and pA sites than the previous annotation from [56].

First, genome wide average mutation rates $p(x \to y)$ are reported by STAMMP for every possible mutation of nucleotide x to another nucleotide y with x,$y \in \{A,C,G,T\}$ and $x \neq y$ so that

$$p(x \to y) = \frac{\text{number of observed } x \to y \text{ mutations}}{\text{number of sequenced } x} \tag{2.11}$$

Second, $q(x \to y)_i$ is defined as the local mutation rate for every possible mutation at genomic position i so that

$$q(x \to y)_i = \frac{\text{number of observed } x \to y \text{ mutations at } i}{\text{number of reads at } i} \tag{2.12}$$

The $q(x \to y)_i$ are collected for each position with at least one observed mutation. Next, the $q(x \to y)_i$ are reported as boxplots for each mutation $x \to y$. Thus, it can be easily checked if the T→C mutations differ from other mutations.

Third, the local mutation rates are plotted as a scatter-plot where $q(x \to y)_i$ is shown on the y-axis and the according coverage on the x-axis. T→C mutations are shown in blue and all other mutations are shown in orange. Thus, variations in the mutation rate for different amounts of coverage can be checked.

Examples for mutation rate plots are given in 3.1.1 (Figure 3.1).

2.6.2 Mutation rate analysis around binding sites

The strongest n non-overlapping PAR-CLIP sites detected by `STAMMP`'s statistical model are centered at the crosslinked T nucleotide. Next, the averaged T→C mutation rates around these central crosslink sites are plotted to check if more T positions are enriched for crosslinks or if the mutation rates are similar to the background.

2.6.3 Whole annotation occupancy profiles

Any list of genomic annotations given in a General Feature Format (GFF) can be plotted as a 2-dimensional heat-map as well as an average of all annotations of the GFF-file for the sense and the anti-sense strand. Each line of a GFF corresponds to one genomic annotation/feature and contains the information about the chromosomal start and stop position of each feature among other things.

The heat-map shows the occupancy values around user defined distances of the start sites given in the GFF file. Additionally, the stop positions are plotted as well. Thus, binding preferences within selected annotations can be displayed. In cases where a large amount of annotations is present the occupancy data can be further smoothed as specified by the user.

Annotation wide averages are plotted around both start and stop positions with a user defined width. Again the sense and anti-sense data is plotted as well as the the averages for different length classes of the given transcripts.

Examples for whole annotation heatmaps can be found in 3.2.3 (Figure 3.8C, D) showing the distribution of the PAR-CLIP signals of the Cbc2 capping enzyme accros mRNA coding transcripts in yeast and e.g. in 3.2.5 (Figure 3.13A) that displays the binding of Nam8, Mud2 and Ist3 in yeast introns. Examples for annotation wide averages are given in 3.2.3 (Figure 3.8H) showing the Cbc2 binding for different length classes of yeast transcripts.

2.6.4 Motif analysis and general enrichments of k-mers

Detected binding sites are sorted according to their estimated occupancies and a user defined number (default: 1 000) of the strongest binding sites are subjected as positive set for the motif finding tool `XXmotif`. Overlapping binding sites are filtered out of the positive set in order to consider a genomic location only ones as a binding site. The filter process always keeps the binding site with the highest occupancy if multiple binding sites are located within one region.

In default settings sequences ±12 nt around the crosslinked T nucleotide are forwarded to `XXmotif` with the parameters

`– – zoops – –merge – motif – threshold LOW – –max – match – positions 10`.

In principle a motif finder searches for short sequences which are enriched in a positive sequence set compared to the occurence in a background or negative set. `XXmotif` normally calculates the background model based on the mono- and di-mer frequencies of the positive input set and ranks candidate motifs according to their significance. The 25 nt short PAR-CLIP sequences often contain strong sequence biases and can therefore lead to inapplicable background models. Thus, `STAMMP` randomly selects 10 000 T nucleotides out of a transcriptome GFF file and subjects sequences ± 12 nt around the randomly chosen T positions as negative set to `XXmotif` in order to get reliable transcriptome background frequencies. The motifs found by `XXmotif` are then returned to the user.

Additional to the binding-motif search a k-mer approach is implemented in `STAMMP` to detect weak k-mer preferences in dependence of the change of the occupancy. Therefore, binding sites are sorted according to their occupancy and are binned. The bin sizes start with sizes of 128, 256, 512, 1 024, 2 048 and 4 096 binding sites per bin and the the size of 4 096 sites per bin is continuously used until no binding site is left over. A log-odd score $S(x)$ is calculated for each k-mer x within a bin according to

$$S(x) = \log_2 \left(\frac{p(x)}{n(x)} \right) \tag{2.13}$$

where $p(x)$ is the probability to observe the k-mer x within the bin and $n(x)$ is the probability to observe k-mer x in the negative set. Probabilities for the negative set are estimated by randomly chosen sequence fragments of the transcriptome. The resulting data is plotted as a heatmap where the sorted bins are drawn at the x-axis, all possible k-mers are drawn at the y-axis and $S(x)$ is color-coded from red to white to green for log-odd scores ranging from $[-3.5,3.5]$.

PWMs found by XXmotif analysis are e.g. shown in 3.2.6 (Figure 3.15B) and enrichments of k-mers can be found in 3.2.6 (Figure 3.15C).

2.6.5 Position dependent k-mer counts

Despite looking for enriched motifs or k-mers as in 2.6.4 `STAMMP` also searches for possible k-mer patterns around detected binding sites. Therefore, PAR-CLIP sites detected by `STAMMP`'s statistical model are centered at the crosslinked T nucleotide. For a given number of k and distance d the occurrence of each k-mer at each genomic position within $\pm d$ nt around all selected crosslink sites is plotted on the y-axis, whereas the genomic position is depicted on the x-axis. Thus, positional sequence biases around binding sites can be discovered easily.

In 3.1.5 (Figure 3.6) position dependent k-mer counts can be found for Nrd1, Pub1, Rna15 and Sub2.

2.6.6 Enrichment of RNA secondary structure features

RNA molecules can form secondary structures like hairpins and stems due to the pairing of bases of the same molecule. Instead of direct binding of proteins to consensus sequences as analyzed in 2.6.4 and 2.6.5 proteins can also bind to secondary structure features. Sequences 100 nt ± the crosslinked T nucleotide serve as input for the program `CapR` which calculates probabilities for each RNA base position belonging to a specific secondary structural category. Currently, the six RNA structural categories

- stem part
- hairpin loop
- bulge loop
- internal loop
- multibranch loop
- exterior loop

are considered by `CapR` to investigate the structural features of RNAs [58]. According to the position dependent analysis of k-mers in 2.6.5 the structural probabilities are plotted on the y-axis and the genomic position is plotted at the x-axis. The probabilities of the structural features are averaged at each genomic position over all analyzed sequences.

2.6.7 Splicing index calculation

Given the annotation $\mathcal{A}$ of splice sites a sequence file containing exon-intron (EI), intron-exon (IE) and exon-exon (EE) junctions ±50nt is build by `STAMMP` and used for the mapping step of the pre-processing procedure described in 2.6 instead of a complete reference genome sequence. Thus, read counts for

$$\begin{aligned} N_i^{EI} &= \text{number of reads covering exon-intron junctions} \\ N_i^{IE} &= \text{number of reads covering intron-exon junctions} \\ N_i^{EE} &= \text{number of reads covering exon-exon junctions} \end{aligned}$$

with $i \in \mathcal{A}$ can be obtained from the artificially mapped PAR-CLIP data. The splicing index $S(\mathcal{A})$ is defined as

$$S(\mathcal{A}) = \log_2 \left(\frac{2 \sum_{i \in \mathcal{A}} N_i^{EE}}{\sum_{i \in \mathcal{A}} N_i^{EI} + \sum_{i \in \mathcal{A}} N_i^{IE}} \right) \tag{2.14}$$

which gives an estimate for the preference for the binding of a protein to spliced or unspliced mRNA.

Splicing indices for 23 proteins can be found in 3.2.5 (Figure 3.13C, Figure 3.14C,D) and in 3.2.9 (Figure 3.18A).

2.6.8 Processing index calculation

Similar to the idea of the splicing index in 2.6.7 the processing index $P(\mathcal{A})$ for an annotation $\mathcal{A}$ of mRNAs of an organism is an estimate for the preference of a protein to bind to uncleaved, premature mRNAs rather than already cleaved, poly-adenylated, mature mRNAs. To calculate the processing index, STAMMP computes annotation wide averages N_i of the read counts r_i of every genomic position i located ±50nt upstream and downstream of pA-sites according to

$$N_i = \frac{1}{|\mathcal{A}|} \sum_{j \in \mathcal{A}}^{\mathcal{A}} r_i^j$$

The assumption is, that averaged read counts N_i^{down} downstream of a pA site can only occur from premature mRNAs, so that $N_i^{prem} = N_i^{down}$, whereas read counts N_i^{up} upstream of a pA site are a mixture of mature mRNA counts N_i^{mat} and pre-mRNA counts N_i^{prem}. Therefore, $N_i^{up} = N_i^{mat} + N_i^{prem}$. For increased robustness with regard to different transcript isoforms and uncertainties in the exact location of pA sites, STAMMP defines the mean of the N_i^{up} values 50nt upstream of the pA-site as M^{up} and the mean of the N_i^{down} values 50nt downstream of the pA-site as M^{down}. The splicing index $P(\mathcal{A})$ is then defined as

$$P(\mathcal{A}) = \log_2 \left(\frac{M^{down}}{\max(1, M^{up} - M^{down})} \right) \tag{2.15}$$

Processing indices are calculated for 23 proteins in 3.2.9 and are displayed in (Figure 3.18A)

2.6.9 Calculation of binding profiles and correlation matrices

STAMMP offers to analyze the change in the binding profile dependent on the occupancy as well as to detect similarities in binding profiles between different factors. More precisely, for each factor f and all transcripts of a given GFF-annotation file, the occupancies in the region between the start and the stop position are rescaled to an equal length of 300 bins. In this way, each transcript t has a resized profile $p^{f,t}$, where $p_i^{f,t}$ is the occupancy of factor f at transcript t at location bin $i \in \{1, \ldots, 300\}$. Next, the mean occupancy per transcript is calculated and each $p^{f,t}$ is assigned to one of 10 equal-sized quantiles

(deciles). For each of these 10 deciles d `STAMMP` sums up the resized profiles $p^{f,t}$ to obtain the whole decile average occupancies which results in averaged binding shapes $p^{f,d}$ for each factor f for each decile d so that

$$p_i^{f,d} = \frac{1}{|d|} \sum_{j}^{|d|} p_i^{f,j} \tag{2.16}$$

For each pair of factors f and f' and each decile d, `STAMMP` computes the Pearson correlation between their binding profile shapes $p^{f,d}$ and $p^{f',d}$ as a measure of the similarity of their binding profiles

$$\mathrm{cor}(f,f') = \frac{\sum_i (p_i^{f,d} - \overline{p}_i^{f,d})(p_i^{f',d} - \overline{p}_i^{f',d})}{\sqrt{\sum_i (p_i^{f,d} - \overline{p}_i^{f,d})^2 \sum_i (p_i^{f',d} - \overline{p}_i^{f',d})^2}} \tag{2.17}$$

Thus, it can be investigated if the binding of a factor is different for different levels of occupancy and if different factors tend to bind at similar regions. The correlations ranging from [−1,1] are then plotted color-coded from red(−1) to white(0) to green(1) as a matrix where all analyzed factors and occupancy deciles are drawn at x- and y-axis as shown in 3.2.9 (Figure 3.19).

2.6.10 Calculation of *total* co-occupancies

To calculate the tendency of pairs of factors f to co-occupy similar subsets of transcripts, `STAMMP` computes the pairwise Pearson correlations of their *total* occupancies $z^{f,t}$ over all transcripts t

$$\mathrm{cor}(f,f') = \frac{\sum_t (z^{f,t} - \overline{z}^{f,t})(z^{f',t} - \overline{z}^{f',t})}{\sqrt{\sum_t (z^{f,t} - \overline{z}^{f,t})^2 \sum_t (z^{f',t} - \overline{z}^{f',t})^2}} \tag{2.18}$$

Noise can be reduced by weighting up/down the contribution of occupancy values of binding sites to the total occupancy of a transcript that is typical/atypical of the binding location of the factor within a transcript. Based on the assumption that the strongest quantile of the averaged binding shapes $p^{f,d}$ (see 2.6.8) represents the typical binding behavior of a protein, the average profile is rescaled to the [0,1] interval to get a weight w_i^f defined as

$$w_i^f = \frac{p_i^{f,d}}{\max(p^{f,d})} \tag{2.19}$$

which is then used to get the weighted total occupancy of a factor f to a transcript t given by

$$z^{f,t} = \sum_{i}^{|p^{f,t}|} w_i^f p_i^{f,t} \tag{2.20}$$

Again the correlations are plotted as a matrix where the analyzed factors are listed on the x- and y-axis as shown in 3.2.9 (Figure 3.18C).

2.6.11 Calculation of *local* co-occupancies

In contrast to the total occupancy STAMMP also calculates the tendency of pairs of factors A and B to bind locations in the transcriptome near to each other defined as the *local* co-occupancy. The average occupancy of factor B within ± 12 nt of occupancy peaks of factor A (unsmoothed occupancy data) is calculated. To suppress statistical noise, only peaks of A above the 75% quantile of all peaks of A are selected to focus on sites which are strongly bound by factor A. Next, the average occupancy of B is divided by the background occupancy of B, which is estimated by averaging the occupancy of B within 25 nt windows out of 2 000 randomly selected positions in the transcriptome. The enrichments of the signal of B around binding sites of A are then plotted as heatmap whereas the factors B are located on the y-axis and the factors A are on the x-axis as shown in 3.2.9 (Figure 3.18D). Note, that this matrix does not have to be symmetric, because factor B might co-localized with factor A but not vice versa.

3 Results & Discussion

The results presented here are separated into a technical and biological part. Technical results presented in 3.1 are primarily focusing on a comparison of the performance of the binding site detection of `STAMMP` to other available methods as well as general characteristics of PAR-CLIP data. All results presented in this chapter are based on the PAR-CLIP experiments generated by Carlo Bäjen (AG Cramer). The achieved biological results using `STAMMP` are described in 3.2 and are published in [44] and [24].

3.1 `STAMMP` - a statistical mixture model for PAR-CLIP data

3.1.1 Offtarget proteins affect PAR-CLIP mutation probabilities

Mutation events in PAR-CLIP experiments show a high amount of variability between different experiments (Figure 3.1). Both the probability to observe possibly crosslink induced T→C mutations and the amount of noise show strong variations.

Besides the mentioned sources of error like sequencing-, PCR- and alignment-errors which can lead to different levels of data quality other proteins can induce T→C mutations as well and thus affect the number of observed T→C mutations referred to as offtarget effects. In theory, observed T→C mutations can also be caused by proteins which bind

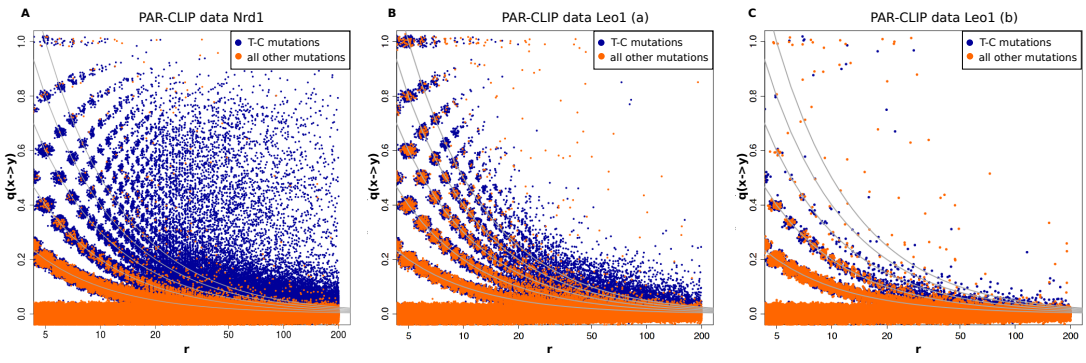

Figure 3.1: **Scatterplots of local mutations rates vs. coverage**
Three different PAR-CLIP experiments are shown to demonstrate the variation of crosslink induced mutations in PAR-CLIP data. The PAR-CLIP data falls of in quality from left (A) to right (C) as the separation between T→C (blue) and other mutations (orange) vanishes from left to right. Local mutation rates $q(x \to y)_i$ are plotted on the y-axis and the corresponding coverage r_i is plotted on the x-axis. As described in 2.3 the negative distribution $p(m_i|r_i,\theta)$ is fitted to the data points represented in orange.

in close vicinity to the protein of interest. If another protein binds to a 4-tU in the neighborhood of an immunoprecipitated protein a T→C mutation can be induced which is not caused by the protein of interest. As a consequence of this, additional T→C counts can be introduced by offtarget effects.

To investigate the contribution of offtarget proteins to the number of T→C mutations the probability of observing T→C mutations in the neighborhood of footprints is compared to the average probability of mutation events. The pre-processing described in 2.2 is repeated with altered parameters for `Bowtie`. The number of allowed mutations `v` is raised from 1 to 3. After the mapping procedure the newly mapped reads R that overlap with the 15 000 strongest binding sites of an analyzed protein are used to approximate the rate λ_{off} of offtarget induced mutations as given by

$$\lambda_{\text{off}} = \frac{\sum_{r \in R} \sum_{i \in |r|} I(n_i^r = C \wedge n_i^g = T)}{\sum_{r \in R} \sum_{i \in |r|} I(n_i^g = T)} \tag{3.1}$$

where I is the indicator function, n_i^r is the nucleotide in read r at position i and n_i^g is the corresponding nucleotide in the reference genome. Only count data of reads which are not located ± 4 nt around the selected binding sites are considered for λ_{off}. The assumption is, that the analyzed protein gets crosslinked to the measured binding site and that additional observed T→C mutations in close proximity on the same read, but not directly at the detected crosslinked sites are likely to be generated by other proteins which were also bound to the fragment.

Offtarget probabilities are distributed around an average offtarget probability of 0.004 (Figure 3.2A, black line) and can be estimated with low amount of data as the rate does not change with an increasing number of analyzed reads. The 25% and 75% quantile rates are 0.003 and 0.005 as indicated in dark gray (Figure 3.2A).

In comparison the offtarget probability largely exceeds the expected average mutation probability (Figure 3.2B). These differences are a strong indicator, that observed T→C events are indeed affected by offtarget effects. From this one can follow, that weak parameters for binding site selections tend to find a high number of binding sites which are likely to be offtarget errors only.

3.1.2 Offtarget rates facilitate realistic test data sets for benchmarking

Comparing existing methods of binding-site detection is necessary to find the optimal method for PAR-CLIP data analysis. Here, a quantitative comparison based on real biological data has been done for the methods introduced in 1.3.2 and `STAMMP` for the first time.

Real biological data should be preferred for method comparison instead of simulated data sets in order to evaluate the methods under real conditions. Thus, the benchmark

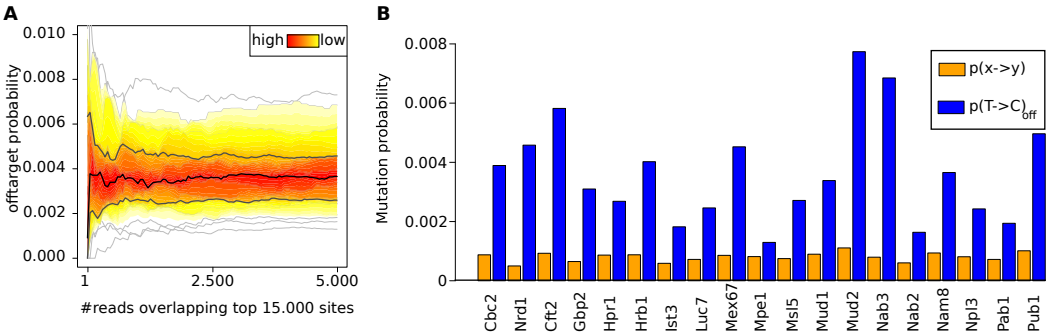

Figure 3.2: **Comparison of estimated offtarget mutation rates and genomic mutation rates**
(A) The change of the estimation of offtarget probabilities in dependence of the number of analyzed reads is shown as clusterplot for 26 PAR-CLIP data sets. The estimated offtarget probability is shown on the y-axis and the number of analyzed reads that overlap with one of the top 15k binding sites is shown on the x-axis. Offtarget probabilities are estimated for 26 PAR-CLIP data sets from different sequencing machines and 2 different labs. The distribution of the offtarget estimations is shown as colorcode from yellow (few data points) to red (many data points). The median of the offtarget probabilities is shown as black line and the 25% quantile and 75% quantile as dark gray lines.
(B) Estimated offtarget probabilities (blue) are compared against the average observed mutation rates. The offtarget rate is higher for all data sets compared to the expected probability based on transcriptome average mutation rates.

proposed here relies on data of 25 *S.cer* data sets published in [44, 24]. All data sets were pre-processed equally according to 2.2. Next, the pre-processed data was forwarded to each introduced binding-site detection method. Only data from the largest yeast chromosome (chrIV) was used for evaluation, due to technical limitations of some methods.

The task of binding-site detection is a binary classification problem. Genomic sites with observed T→C mutations can be classified as binding-sites (positive) or non-binding-sites (negative). There are four possible different outcomes in classification problems:

- True positive (TP): a site classified as positive is a true interaction site
- False positive (FP): a site classified as positive is no interaction site
- True negative (TN): a site classified as negative is no interaction site
- False negative (FN): a site classified as negative is a true interaction site

Various measurements for the evaluation of classifiers were defined based on these counts. Here, only the true-positive-rate (TPR), false-positive-rate (FPR), precision and false-

discovery-rate (FDR) are recapitulated:

$$\text{TPR} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{3.2}$$

$$\text{FPR} = \frac{\text{FP}}{\text{FP} + \text{TN}} \tag{3.3}$$

$$\text{precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{3.4}$$

$$\text{FDR} = \frac{\text{FP}}{\text{FP} + \text{TP}} \tag{3.5}$$

Receiver operator characteristics (ROC) or ROC-curves are a standard procedure for the comparison of different classifiers by plotting the TPR versus the FPR. Unfortunately, the mandatory true-positive counts, false-negative counts and false-positive counts for ROC analyses cannot be determined directly from biological data. It cannot be distinguished if found binding sites are true interactions sites or not. Therefore, the number of found binding sites is a mixture of true positives and false positives. In addition, it cannot be distinguished how many sites which are not classified as true binding site are actually no interaction site. From these observations it follows, that the standard ROC-curve cannot be used for classifier evaluation.

To circumvent these shortcomings of real data an approximation for the number of false-positives is used which allows a comparison of the methods in a ROC-like fashion. Each introduced method is used to find binding sites in PAR-CLIP data with observed G→A mutations. Sites with G→A mutations are no protein-RNA interaction sites. Their only source of origin are technical errors. As stated in 3.1.1 PAR-CLIP induced T→C counts can be additionally affected by offtarget proteins. Non-T→C mutations have a lack of offtarget induced effects. Thus, G→A mutations are artificially introduced into the PAR-CLIP data with the estimated rate $\lambda_{\text{off}} = 0.004$ to correct for this missing effect. As a consequence, the final G→A mutations serve as an estimator for the number of false-positively classified sites for each method after correction for unequal nucleotide occurrences.

The different methods are evaluated by comparing the number of found binding sites with the estimated FDR based on the approximated false-positive counts. The methods proposed by `wavClusteR`, `naive` and `STAMMP` only use one user defined parameter for binding site detection. Hence, these methods are compared by plotting the number of called binding sites on the y-axis and the corresponding FDR on the x-axis for various thresholds.

Binding-site-detections proposed by `PARalyzer` and `PIPE − CLIP` are based on various user defined parameters. A comparison of all methods is done by using the default

parameter settings for each method. PARalyzer was used with the parameters:

$$\begin{aligned}
\text{bandwidth} &= 3\\
\text{conversion} &= T > C\\
\text{minimum read count per group} &= 5\\
\text{minimum read count per cluster} &= 5\\
\text{minimum read count for KDE} &= 5\\
\text{minimum cluster size} &= 10\\
\text{minimum conversion location for cluster} &= 1\\
\text{minimum conversion count for cluster} &= 1\\
\text{minimum read count for cluster inclusion} &= 5\\
\text{minimum read length} &= 13\\
\text{maximum number of non conversion mismatches} &= 0\\
\text{additional nucleotides beyond signal} &= 5
\end{aligned}$$

PIPE − CLIP was used with the parameters:

$$\begin{aligned}
\text{Remove PCR duplicate} &= \text{NO}\\
\text{FDR for determining Enriched CLuster} &= 0.01\\
\text{FDR for determining Reliable Mutations} &= 0.01
\end{aligned}$$

wavClusteR was used with the interval of [0.2,0.9], the threshold for the naive approach was $\delta_m \geq 2$ and the p-value cut-off for STAMMP was 0.002.

3.1.3 STAMMP finds more binding sites at low FDRs

A comparison of the binding site performances of the single threshold based methods STAMMP, wavClusteR and naive using the benchmark data sets as described in 3.1.2 shows in general a better performance of STAMMP (Figure 3.3). Starting with the most weak parameters for each method a site is considered as a true binding site with at least one observed T→C mutation which leads to equal FDRs for each method between $7\% - 20\%$ dependent of the tested data sets. Depending on the analyzed factor the maximal number of called sites varies between a minimum number of ∼11 000 found sites for the splicing factor Ist3 and a maximum number of ∼165 000 found sites for the elongation factor Npl3. Using more strict thresholds for all methods the statistical model of STAMMP quickly starts to find more binding sites at lower FDRs compared to the other methods in 14/24 data sets (Figure 3.3). At a fixed FDR of 5% STAMMP finds on average around 2.03 more binding sites compared to the naive approach and around 3.2 more binding sites compared to wavClusteR.

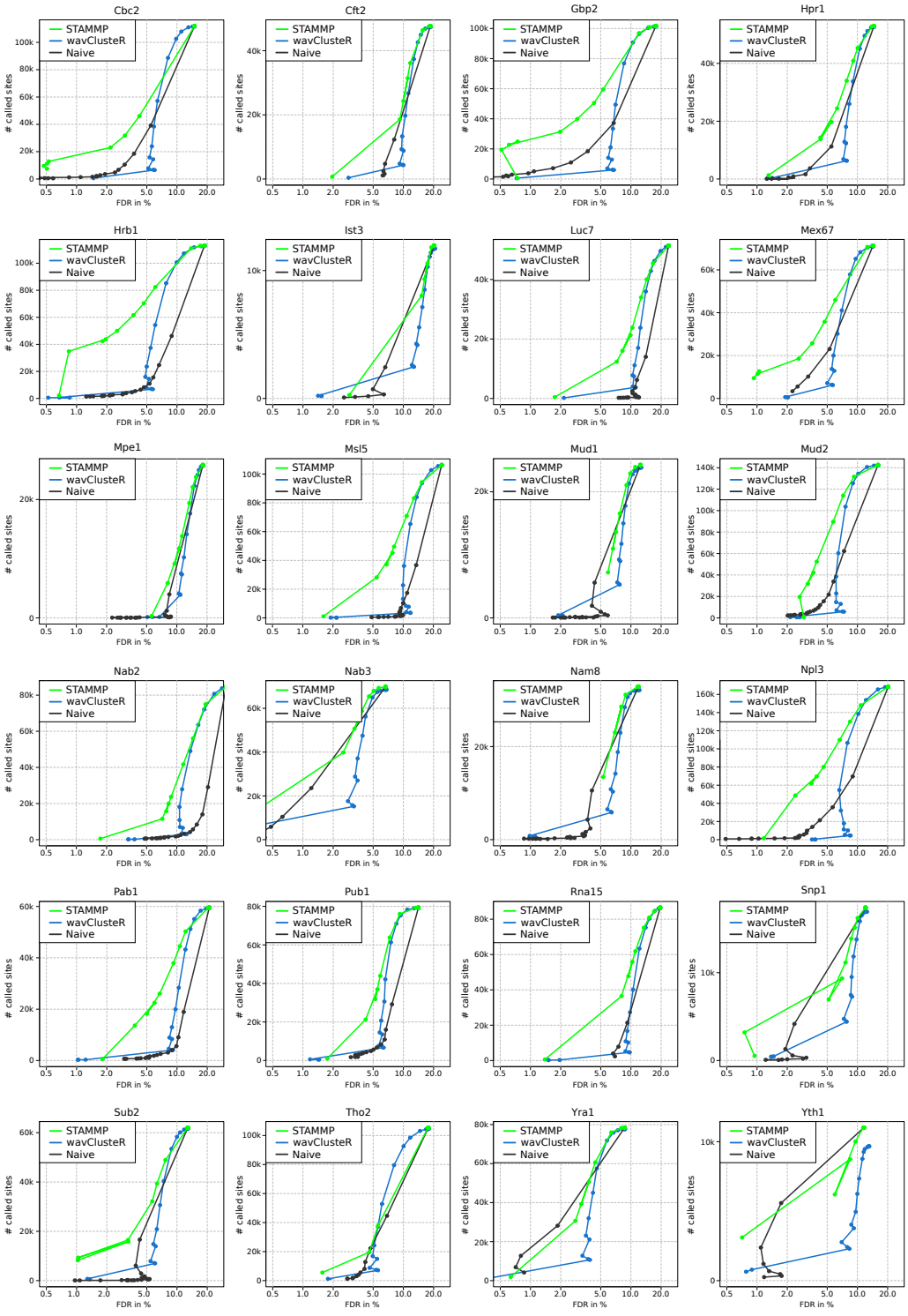

Figure 3.3: **Detailed comparison of binding site performances of `STAMMP`, `wavClusteR` and `naive` on 24 PAR-CLIP data sets**
Performances are evaluated on PAR-CLIP data of chromosome IV of *S.cer* experiments. Offtarget based mutations were additionally introduced as described in 3.1.2. The number of detected binding sites is depicted on the y-axis of each plot and the corresponding achieved FDR based on called G→A mutations is shown on the x-axis. `STAMMP` shows a clear improvement in 14/24 data sets and has a similar performance in 10/24 data sets.

The comparison of all tested methods using their suggested parameter settings (see 3.1.2) shows that the statistical model used in `STAMMP` achieves overall better performances compared to already published methods. FDRs achieved by `STAMMP` are pairwise compared to the FDRs achieved by `wavClusteR`, `naive`, `PARalyzer` and $\texttt{PIPE}-\texttt{CLIP}$ (Figure 3.4A). The majority (85) of data points is located above the diagonal supporting the statement that FDRs achieved by `STAMMP` are lower. Only a minority (19) of data sets fall under the diagonal. However, the FDRs achieved by `STAMMP` for this data remain comparable to the FDRs achieved by other methods in 12/19 cases as the worst FDR of `STAMMP` is only 16%. The colorcode from red (-3.5) to white (0) to green (+3.5) represents the enrichment of called sites which is given by

$$\log_2\left(\frac{\#\text{found sites by } \texttt{STAMMP}}{\#\text{found sites by other method}}\right) \tag{3.6}$$

Only three comparisons show a strong reduction in the number of called binding sites by `STAMMP` while lowering the FDR and four examples show a minor reduction in the number of called sites. The remaining majority of comparisons indicates that `STAMMP` achieves both a reduction in the FDR while finding more reliable binding sites (Figure 3.4A). All cases where `STAMMP` shows a slightly worse FDR more binding sites are called compared to other methods.

Overall `STAMMP` achieves an average FDR of 4.63% and 50% of all tested data sets get FDRs between 2.03% (25% quantile) and 6.92% (75% quantile, Figure 3.4B). In comparison the `naive` method shows the second best performance with an average FDR of

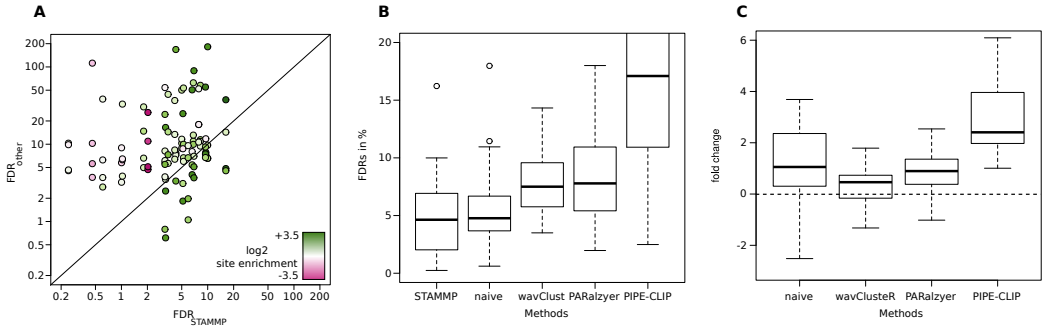

Figure 3.4: **Comparison of FDRs and number of found binding sites with default parameters**
(A) Pairwise comparison of the FDR achieved by `STAMMP` (x-axis) to the FDR achieved by `naive`, `wavClusteR`, `PARalyzer` or $\texttt{PIPE}-\texttt{CLIP}$ (y-axis). The $\log_2$ enrichment of the number of sites found by `STAMMP` compared to another method is colorcoded from $[-3.5,3.5]$
(B) Distribution of the achieved FDRs (y-axis) for each tested method (x-axis) based on 26 PAR-CLIP data sets.
(C) Distribution of the $\log_2$ enrichments (y-axis) as given in equation 3.6. Each introduced method is compared against the number of found sites by `STAMMP` (x-axis).

4.77% and FDRs of 3.67% and 6.68% for the 25% quantile and 75% quantile. The particularly introduced PAR-CLIP analysis methods `wavClusteR`, `PARalyzer` and `PIPE − CLIP` achieve average FDRs of 7.5%, 7.79% and 17.1% and values for the 25% quantile of 5.76% 5.41% and 10.92%. Likewise the 75% quantile boundaries for these methods are 9.58%, 10.94% and 26%.

The pairwise comparison of the $\log_2$ enrichment of the number of found sites as given in equation 3.6 shows a clear enrichment in the number of found sites by `STAMMP` in most of the tested data sets (Figure 3.4C).

Note, that the RSF decision criteria for binding site detection as proposed by `wavClusteR` can in principle be used to achieve better performances than the `naive` approach. However, `wavClusteR`'s model proposes RSF values within [0.003,0.009] for the lower boundary and between [0.995,0.999] for the upper boundary for most of the data sets which is similar to accept every position as a true binding site with at least one observed T→C mutation leading to the worst FDRs. The mixing coefficient estimation of `wavClusteR` as proposed in equation 1.15 is likely to be the reason for bad RSF value determination. The analysis of the data used here leads to mixing coefficients for λ_2 close to 1. As a consequence, the term $\lambda_1 p_1(x)$ that models the non-experimentally induced mutations gets irrelevant.

The bad performance of `PIPE − CLIP` is likely to be caused by the choice of the parameterization for the success rate τ in equation 1.16. No reasonable explanation can be found to estimate the number of observed mutation by approximating the success rate by dividing the number of reads by the genomesize.

Although no detailed comparisons about running time and memory-consumption are done here please note, that the `naive` approach and `STAMMP` are the fastest tested methods with the lowest memory-consumption which only take minutes to get binding sites from the mpileup file. Only `PARalyzer` shows economically justifiable results. Unfortunately, the version of `PIPE − CLIP` used here was not able to handle a single experiment (data file around 1GB) at once without running out of memory on a standard desktop computer. The binding site detection of `wavClusteR` is as fast as `STAMMP`. However, the proposed subsequent step to find crosslink boundaries takes several days even for data of one yeast chromosome only.

3.1.4 Ordering of binding sites affects motif performance

A small fraction of seven proteins out of 25 different RBPs namely Nrd1, Nab3, Gbp2, Pub1, Pab1, Yth1, and Hrb1 show sequence specificities as detected by `XXmotif`. From these proteins only Nrd1 shows a strong sequence specificity whereas the remaining proteins only show an enrichment of binding sequences within the strongest bound sequences.

To further evaluate the method performances reported binding sequences of a sequence

specific protein are scanned for the occurrence of the corresponding binding motif. Therefore, the binding-sites found by each method on the yeast chromosome IV are sorted according to their score proposed by each method. If no score is available, the binding sites are sorted according to the number of observed T→C mutations m_i^{TC} as proposed in [36]. Next, the specific binding sequence of a protein is searched ± 12 nt around a crosslinked site. If the binding motif is present the binding site is considered as a TP and as a FP otherwise. Then, the cumulated precision is plotted on the y-axis and the numbers of already searched sites is plotted on the x-axis.

Motif performances are obtained for the Nrd1 binding sites with the consensus motif 'UGUA' [45] (Figure 3.5A) and for the Gbp2 binding sites with the consensus motif 'GGUG' [59] (Figure 3.5B). The p-value ordering proposed by `STAMMP` shows an improved motif find performance for the Nrd1 data compared to the other methods. `PIPE` − `CLIP`'s ordering shows the second best performance, while there is no difference in motif finding between `wavClusteR`, `naive` and `PARalyzer`.

In contrast, the p-value ordering of `STAMMP` shows the worst performance for the Gbp2 data compared to the other methods. In general, the enrichment of the Gbp2 consensus motif is not as strong as in the case of Nrd1.

These contrary results in the motif performance and the lack of more PAR-CLIP data of sequence specific RBPs do not allow a final conclusion about the best ordering of binding of PAR-CLIP binding sites. Additional technical biases might affect the ordering of

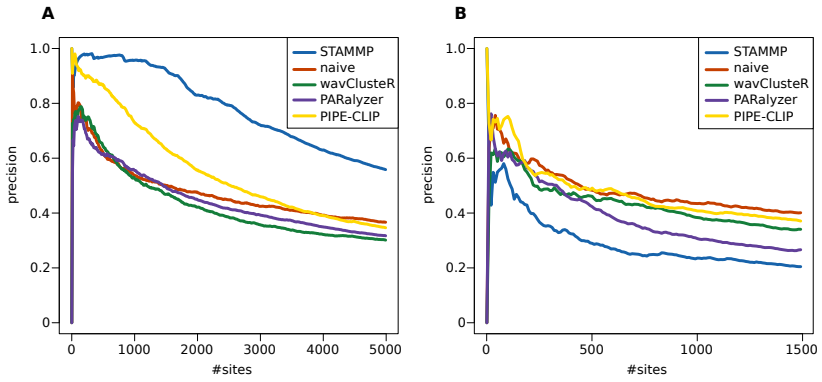

Figure 3.5: **Influece of binding site ordering to motif find performances on Nrd1 and Gbp2 data sets**
(A) Nrd1 binding sites are sorted according to the score value proposed by each method and the sites are scanned for the presence of the Nrd1-consensus sequence UGUA. If the motif was found the sites is considered as true positive and as false positive otherwise. For the Nrd1 data the sorting according to `STAMMP`s p-value shows the best performance, followed by `PIPE` − `CLIP`.
(B) Gbp2 binding sites are sorted according to the score value proposed by each method and the sites are scanned for the presence of the binding the Gbp2 binding motif GGUG. `PIPE` − `CLIP` and the `naive` ordering show the best results for Gbp2. Compared to Nrd1 the overall presence of the consensus motif of Gbp2 is reduced.

binding sites which can lead to artificially enrichments or depletions of sequences biasing the motif analysis.

3.1.5 PAR-CLIP data shows position dependent *k*-mer biases.

RNA-binding proteins show differences in the positions of their binding sequences in dependence of the crosslinked site. The 4-mer counts analyzed here are based on the first 2000 highest occupied binding sites per protein.

Nrd1 shows the strongest sequence specificity of all analyzed proteins. The Nrd1 binding motif UGUA shows no positional variance as it is specifically peaked direct at the crosslinked site, although there are in principle two possible U nucleotides to facilitate a crosslink (Figure 3.6A).

In contrast, the poly-U binding motif of Pub1 is not as positioned at the central 4-tU as the Nrd1 motif. Here, most of the Us of the binding sequence can facilitate crosslinks to the protein (Figure 3.6B). Additionally, poly-Us are in general enriched in Pub1 data.

Rna15 tends to bind downstream of poly-U enriched sequences. Poly-Us are slightly enriched upstream, but not downstream of Rna15 sites and show the strongest enrichment at the crosslinked U (Figure 3.6C).

Besides the analysis of preferred binding motifs PAR-CLIP measurements based on 4-tU crosslinking tend to predominantly bind to AAAU or poly-A stretches followed by a 4-tU. As an example the 4-mer data of Sub2 is shown in Figure 3.6D. The AAAU peak upstream of the crosslinked site is present in 21/25 analyzed proteins. More strikingly, the AAAU 4-mer is also present in binding sites with lower occupancies. Thus, it is likely that PAR-CLIP measurements based on 4-tU tend to have a sequence bias favoring AAAU stretches for crosslinking.

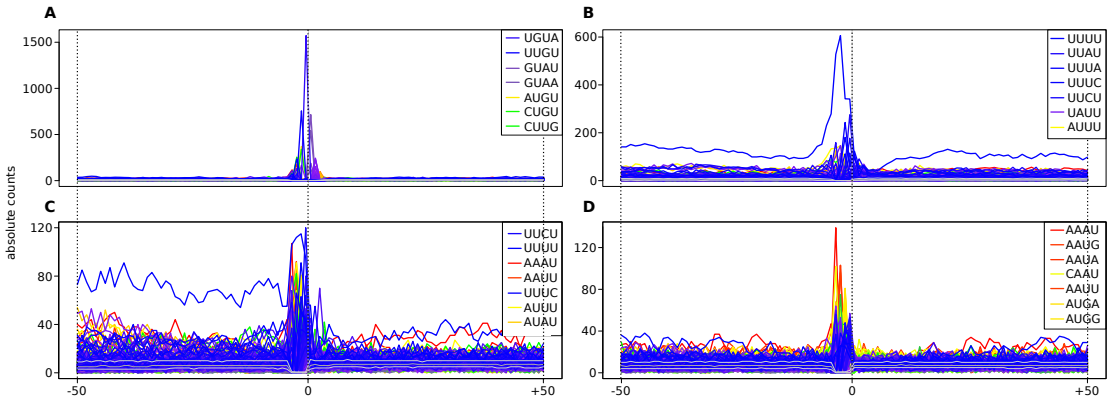

Figure 3.6: **Position dependent *k*-mer counts for Nrd1(A), Pub1(B), Rna15(C) and Sub2(D) data**
The absolute 4-mer counts around the 2000 strongest bound sites of Nrd1(A), Pub1(B), Rna15(C) and Sub2(D) are shown. 4-mers in the legend are sorted according to the observed peak height for each protein.

3.1.6 Overlaps of binding sites in PAR-CLIP data indicate technical background bias

Biochemical experiments often tend to show systematic biases which have to be taken into account and corrected for in order to receive reliable analyses. For ChIP-chip data, analyses have been performed to check and correct for biases prior to post-process analyses [60, 7]. Likewise, it could be shown, that highly expressed loci were always enriched in ChIP-seq analyses regardless of the analyzed protein [61]. To check if the immunoprecipitation or the PAR-CLIP protocol in general show similar biases the 25 published data sets are analyzed for the degree of overlap of binding sites of factors A and B as defined by the Jaccard similarity coefficient $\mathcal{J}_{A,B}$ which is defined as

$$\mathcal{J}_{A,B} = \frac{|A \cap B|}{|A \cup B|} \tag{3.7}$$

In fact, PAR-CLIP data sets show a degree of similarity with an average of 0.07 and with a maximum $\mathcal{J}_{A,B}$ of 0.22 (Figure 3.7A) indicating strong immunoprecipitation biases. Although no intentional correction is given in STAMMP, the proposed expression normalization (see 2.4) corrects for this bias implicitly. An binding site ordering according to their occupancies coupled with a binning of binding sites of 300 sites per bin followed by the calculation $\mathcal{J}_{A,B}$ per bin shows, that bins with strong binding sites of factor A do not show high values for $\mathcal{J}_{A,B}$ with bins of binding sites of another factor B (Figure 3.7B).

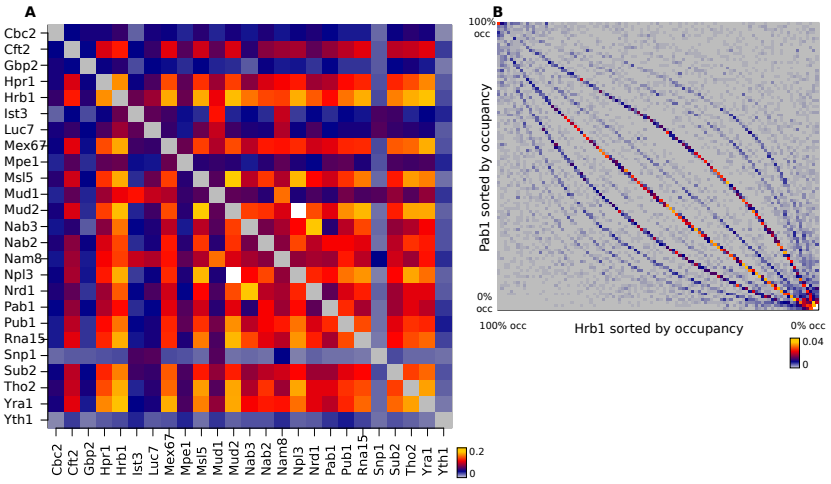

Figure 3.7: **PAR-CLIP data sets show a high degree of overlap**
(A) The Jaccard-coefficient of binding sites $\mathcal{J}_{A,B}$ detected in two PAR-CLIP data sets A and B is shown in heat color code. Sites of factor A are drawn on the y-axis and sites of factor B are drawn on the x-axis. Entries of the diagonal indicating a 100% overlap of a factor with itself were set to 0.
(B) Bins of Pab1 and Hrb1 sites are sorted according to their occupancies and are shown exemplaray for all analyzed data sets. The overlap of bins is shown in heat color code. The lower the occupancy of a factor gets the higher is the overlap between two different factors.

The lower the occupancy the higher the degree of overlap indicating, that a general background binding is present in PAR-CLIP data. Recently, the general background binding was confirmed biochemically in another study [62].

3.2 Transcriptome maps of mRNP biogenesis factors define pre-mRNA recognition

The analysis pipeline represented by `STAMMP` was used to analyze 25 RNA-binding proteins in *S.cer* which were published in 2 publications and presented here and in 3.3.

All results presented in this section were obtained in collaboration with Carlo Bäjen and are published in [44].

3.2.1 Summary

Biogenesis of eukaryotic messenger ribonucleoprotein complexes (mRNPs) involves the synthesis, splicing, and 3′ processing of pre-mRNA, and the assembly of mature mRNPs for nuclear export. We mapped 23 mRNP biogenesis factors onto the yeast transcriptome, providing $10^4 - 10^6$ high-confidence RNA interaction sites per factor. The data reveal how mRNP biogenesis factors recognize pre-mRNA elements *in vivo*. They define conserved interactions between splicing factors and pre-mRNA introns, including the recognition of intron-exon junctions and the branchpoint. They also identify a unified arrangement of 3′ processing factors at pre-mRNA polyadenylation (pA) sites in yeast and human, which results from an A-U sequence bias at pA sites. Global data analysis indicates that 3′ processing factors have roles in splicing and RNA surveillance, and that they couple mRNP biogenesis events to restrict nuclear export to mature mRNPs.

3.2.2 Introduction

Biogenesis of eukaryotic mRNAs involves pre-mRNA synthesis by RNA polymerase (Pol) II and cotranscriptional RNA processing, which encompasses 5′ capping, intron splicing, and 3′ RNA cleavage and polyadenylation (3′ processing). The mature mRNA is packaged with RNA-binding proteins into messenger ribonucleoprotein particles (mRNPs) and exported to the cytoplasm where it directs protein synthesis. Factors for mRNP biogenesis are recruited cotranscriptionally by interactions with the C-terminal domain (CTD) of Pol II [63, 64, 65, 66], and by interactions with the emerging pre-mRNA transcript [67, 68, 69, 70, 71].

Mapping of mRNP biogenesis factors onto pre-mRNA and mature mRNA promises insights into RNA determinants for splicing, 3′ processing, and RNA export, and into the coupling between these processes. Biogenesis factors can in principle be mapped onto the transcriptome by *in vivo* protein-RNA crosslinking and immunoprecipitation (CLIP) [33]. CLIP is based on UV light-induced crosslinking and identifies direct protein-RNA interaction sites after sequencing of the crosslinked RNA regions [72]. CLIP-based methods could indeed provide transcriptome maps for several human 3′ processing factors [73] and mRNA-binding proteins in the yeast *Saccharomyces cerevisiae* [74].

However, mRNP biogenesis factors have not been systematically mapped onto pre-mRNA, likely due to difficulties in trapping short-lived RNAs in cells, and due to the complexity caused by different pre-mRNA species.

Here we present high-confidence transcriptome maps for 23 mRNP biogenesis factors in yeast, where pre-mRNA complexity is low because spliced protein-coding genes contain only single introns. These maps were obtained by photoactivatable-ribonucleoside-enhanced (PAR)-CLIP, which was developed in human cells [36] and recently adopted to yeast [45, 24]. Compared to other CLIP methods, PAR-CLIP uses less invasive, low-energy UV light, which results in a specific U-to-C base transition at the crosslinked sites that facilitates their precise localization at low false-positive rates.

Our analysis includes factors implicated in 5′ cap binding, splicing, 3′ processing, and mRNA export. Six of these 23 factors, namely Gbp2, Hrp1/Nab4, Mex67, Nab2, Pab1, and Tho2, were recently mapped using a technique called CRAC [74]. This published study focused on the distribution of RBPs between mRNAs and noncoding (nc) transcripts, whereas we focus here on pre-mRNA recognition during mRNP biogenesis. We show that PAR-CLIP captures short-lived pre-mRNA intermediates, and provide insights into the *in vivo* RNA-binding preferences of mRNP biogenesis factors, the recognition of introns and 3′ processing sites in pre-mRNA, and the interdependence of different steps in mRNP biogenesis.

3.2.3 Transcriptome maps of mRNP biogenesis factors

To map mRNP biogenesis factors over cellular RNA at high resolution, we optimized the PAR-CLIP protocol and obtained high RNA labeling efficiencies with 4-thiouracile (4tU) in exponentially growing yeast cells (Experimental Procedures). We found conditions that led to very high reproducibility between biological replicates (Figure 3.9) and enabled high 4tU incorporation levels of 2% [75] without significant changes in cellular mRNA abundance (Figures 3.8A and 3.9B). We also developed a computational pipeline for PAR-CLIP data analysis (P.T., C.B., A. Graf, S. Krebs, P.C, and J.S., un-published data). This pipeline includes a statistical model for crosslink site determination and an analysis of the sequence neighborhood of crosslinked sites with the motif discovery tool XXmotif [76].

For each factor, we obtained between 25 000 and 800 000 high-confidence protein-RNA binding sites at a p value below $5 \times 10^{-}3$, which corresponds to false discovery rates between only 0.18% and 3.5% (Table 3.1).

Biogenesis Event	Factor/ Subunit	Complex	RNA-Binding Domain	PAR-CLIP Crosslink Sites	False Discovery Rate %
Capping	Cbc2	CBC	RRM	98 034	0.178
Splicing	Luc7	U1 snRNP	ZF	93 261	1.035
	Mud1	U1 snRNP	RRM	99 384	1.918
	Nam8	U1 snRNP	RRM	151 813	1.675
	Snp1	U1 snRNP	RRM	25 493	0.447
	Ist3	U2 snRNP	RRM	66 003	3.184
	Mud1	BBP-U2AF65	RRM	801 430	1.769
	Msl5	BBP-U2AF65	ZN	476 370	1.961
3′ processing	Rna15	CFIA	RRM	582 756	3.463
	Mpe1	CPF	ZF	122 500	2.262
	Yth1	CPF(PFI)	ZF	59 049	3.432
	Cft2	CPF(CFII)	-	189 866	1.723
	Pab1	-	RRM	233 513	2.052
	Pub1	-	RRM	371 902	1.332
Export	Hpr1	THO/TREX	-	249 887	1.913
	Tho2	THO/TREX	-	400 965	1.064
	Sub2	TREX	-	228 620	1.085
	Mex67	TREX	-	288 579	1.010
	Yra1	Export adaptor	RRM	400 156	1.064
	Nab2	Export adaptor	ZF	283 606	2.413
	Npl3	Export adaptor	RRM	770 240	1.282
	Hrb1	SR-like	RRN	395 402	0.976
	Gbp2	SR-like	RRM	65 692	0.182

Table 3.1: **mRNP Biogenesis Factors Analyzed here by PAR-CLIP**

We applied the optimized protocol to 23 mRNP biogenesis factors that showed reproducible PAR-CLIP signals (Table 3.1). These included the Cbc2 subunit of the cap-binding complex (CBC) and components of the splicing machinery, namely the yeast homologs of the branchpoint (BP)-binding protein BBP (Msl5) and U2AF65 (Mud2), and subunits of the snRNPs U1 (Luc7, Mud1, Nam8/Mud15, Snp1) and U2 (Ist3/Snu17). Factors in the 3′ processing machinery included the Rna15 subunit of cleavage factor (CF) IA, and three subunits of the cleavage and polyadenylation factor (CPF), namely Mpe1, Yth1 (in the CPF subcomplex PFI), and Cft2/Ydh1 (CPF subcomplex CFII). We also included nine proteins implicated in mRNP export, in particular subunits of the THO/TREX complex (Hpr1, Tho2, Sub2), the export factor Mex67, and its putative mRNA adaptors Nab2, Npl3 (also known as Nop3 or Nab1), and Yra1/She11, and

the SR-like factors Gbp2 and Hrb1. We also studied the poly(A)- and poly(U)-binding proteins Pab1 and Pub1 that regulate mRNP export and stability [77].

Biological replicates for a random selection of factors revealed a high reproducibility (Figures 3.9C-E). The obtained data map the protein-RNA interaction landscape underlying mRNP biogenesis (Figures 3.10, 3.11A, and 3.11B).

3.2.4 RNA abundance normalization reveals capped transcripts

PAR-CLIP crosslinks for the CBC subunit Cbc2 clustered at the 5′ ends of mRNAs as expected, but often extended for several hundred nucleotides (nt) downstream (Figure 3.8C). We found that Cbc2 binding appeared more focused at mRNA 5′ ends after the data were corrected for RNA abundance (Figure 3.8D), as measured by RNA-Seq under the same experimental conditions (Figure 3.8B). We estimated relative occupancies of the crosslinked factors along mRNAs by dividing the frequency of U-to-C transitions by the RNA-Seq signal at this site. The normalization reduced the transcript-to-transcript signal fluctuation, led to an even distribution of estimated occupancy signals over RNAs with different abundance (Figure 3.8E), and abolished a weak artificial correlation of PAR-CLIP signals with RNA levels (Figure 3.8F). As a result, the distribution of crosslinking sites over transcripts was independent of the number of observed crosslinks (Figure 3.8G, and 3.8F), confirming the high data quality. The normalization procedure thus prevents misinterpretation due to systematic overrepresentation of abundant transcripts. In the normalized data, strongest binding of Cbc2 was observed within the first ∼90 nt downstream of the transcription start site (TSS) within the 5′ untranslated region (5′ UTR) of mRNAs (Figure 3.8G, 3.8H, and 3.10).

The normalization also enhanced Cbc2 signals on ncRNA transcripts (Figure 3.8C and 3.8D, green panels), facilitating the detection of capped ncRNAs (Figure 3.8H and 3.11). Widespread Cbc2 binding was observed at the 5′ end of divergent ncRNA transcripts that emerged from bidirectional promoters antisense to mRNAs. Cbc2 sites were found from ∼120 nt upstream of the TSS of the sense transcript, with the peak of Cbc2 crosslinking at ∼250 nt (Figure 3.8H, left panel). This is consistent with the presence of two distinct Pol II initiation complexes for sense and divergent transcription from bidirectional promoters [78], and indicates that divergent transcripts are capped before they associate with the Nrd1 complex that triggers their degradation [79, 25, 24]. Cbc2 also crosslinked to antisense RNA 100–300 nt upstream of the polyadenylation (pA) site, identifying capped antisense ncRNAs at the 3′ ends of many genes (Figure 3.8H, right panel). We also identified Cbc2-binding sites in cryptic unstable transcripts (CUTs) and stable untranslated transcripts (SUTs) [80], with stronger signals for CUTs (Figure 3.12A and 3.8B).

The Cbc2 data enabled comparison with the recent CRAC-based mapping of Cbc1, the other subunit of CBC [74]. Both Cbc1 and Cbc2 showed RNA interactions at the

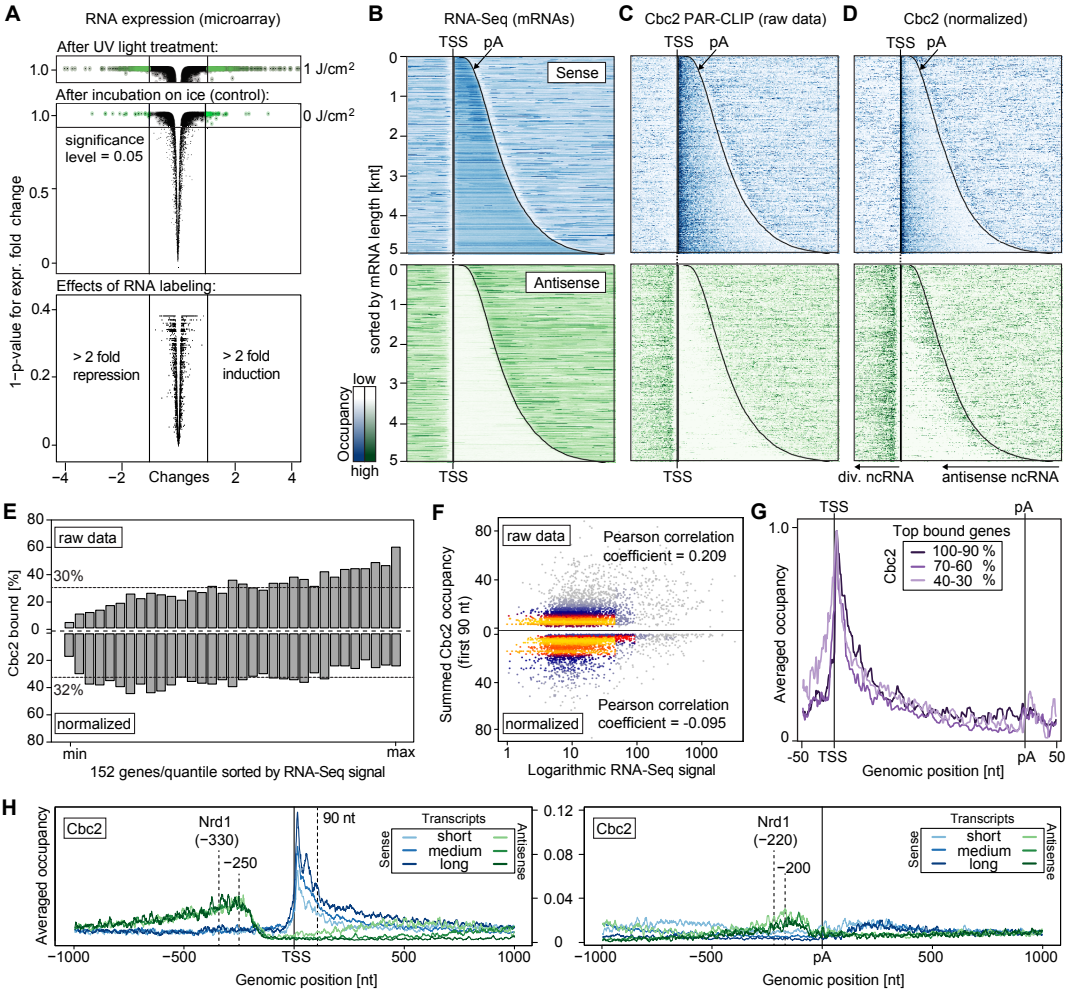

Figure 3.8: **RNA Abundance-Normalized PAR-CLIP Estimates Factor Occupancies over the Yeast Transcriptome**

(A) 4-thiouracil (4tU) labeling has only a very minor effect on cellular mRNA levels. Vulcano plots of expression fold changes for mRNAs measured by Affymetrix microarrays show that only few mRNAs significantly change their abundance due to RNA labeling, incubation on ice, and UV light exposure.

(B) Smoothed Cbc2 RNA-Seq data in sense (blue) and antisense (green) direction for all open reading frame-containing transcribed regions (ORF-Ts). ORF-Ts are sorted by length and aligned at their transcription start site (TSS).

(C) Smoothed, raw Cbc2 RNA-binding strength as measured by the number of PAR-CLIP U-to-C transitions per U site in sense (blue) and antisense (green) direction for all ORF-Ts sorted by length and aligned at their TSS.

(D) Normalization of PAR-CLIP signals reduces noise. Relative Cbc2 occupancy estimated by dividing the number of U-to-C transitions for each U site by the RNA-Seq signal at the corresponding genomic position in sense (blue) and antisense (green) direction for all ORF-Ts.

(E) Normalization of PAR-CLIP signals facilitiates interpretation as occupancy profiles. Whereas raw PAR-CLIP binding strength (shown in C) strongly depends on mRNA level, normalized occupancies (shown in D) are independent of mRNA levels. The y axis shows the percentage of transcripts bound by Cbc2, where 'bound' is defined as the sum of the first 90 nt of a transcript $\geq$ the mean of the sums of the first 90 nt of all ORF-Ts.

(F) Normalization abolishes the dependence of estimated occupancy on mRNA level. Pearson correlation between mRNA level and the PAR-CLIP binding strength in the first 90 nt of each ORF-T before (top) and after (bottom) RNA abundance normalization.

(G) Cbc2-binding profiles are independent of factor occupancy. Transcript-averaged Cbc2 occupancy for three mRNA level classes (100% − 90%, 70% − 60%, and 40% − 30% expression quantile).

(H) Transcript-averaged Cbc2 occupancies in sense (blue) and antisense (green) directions, centered at the TSS (left) and the polyadenylation site (pA) (right), for short (0-1 kb), medium (1-2 kb), and long (2-5 kb) transcripts.

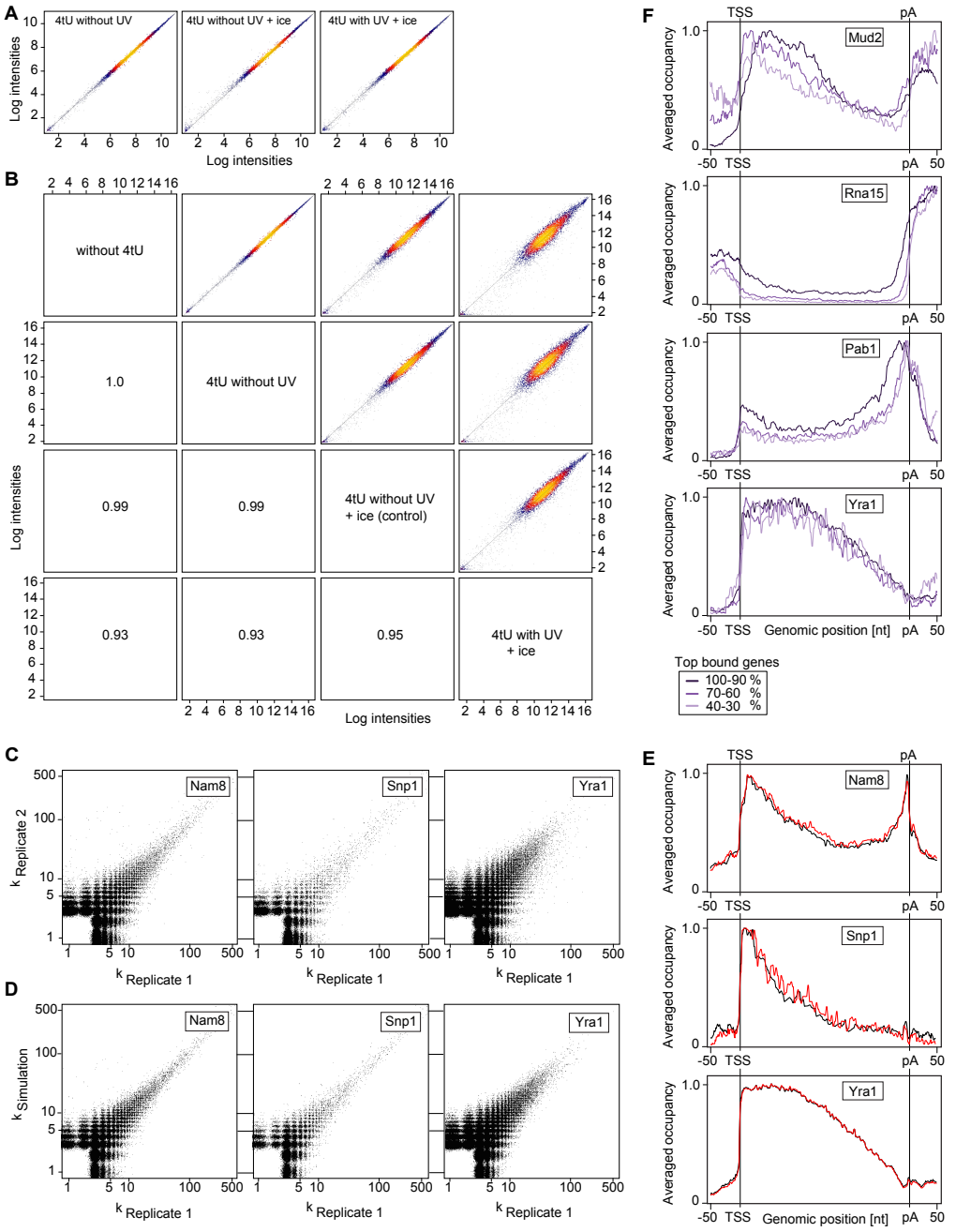

Figure 3.9: **4tU labeling and UV-treatment leave gene expression levels nearly unchanged**

(A) Correlation of expression levels of between pairs of biological replicates with same treatment: without 4tU labeling after 4tU-labeling, and after subsequent UV-light treatment with an energy dose of 1 J/cm^2 (related to Figure 3.8A)

(B) Correlation of expression levels between cells after the various treatment steps during the PAR-CLIP procedure.

(C) Correlation of replicates for Nam8, Snp1 and Yra1.

(D) For comparison with panel C, the simulated correlation by Poisson distribution for Nam8, Snp1 and Yra1 is provided.

(E) Averaged occupancy profiles of replicates for Nam8, Snp1 and Yra1.

(F) Occupancy profiles are independent of factor occupancy. Transcript-averaged occupancy for three expression level classes [100% − 90%, 70% − 60%, and 400% − 30% expression quantile] of Mud2, Rna15, Pab1, and Yra1. This demonstrates that occupancy profiles are reliable even at lowly occupied genes (related to Figure 3.8G).

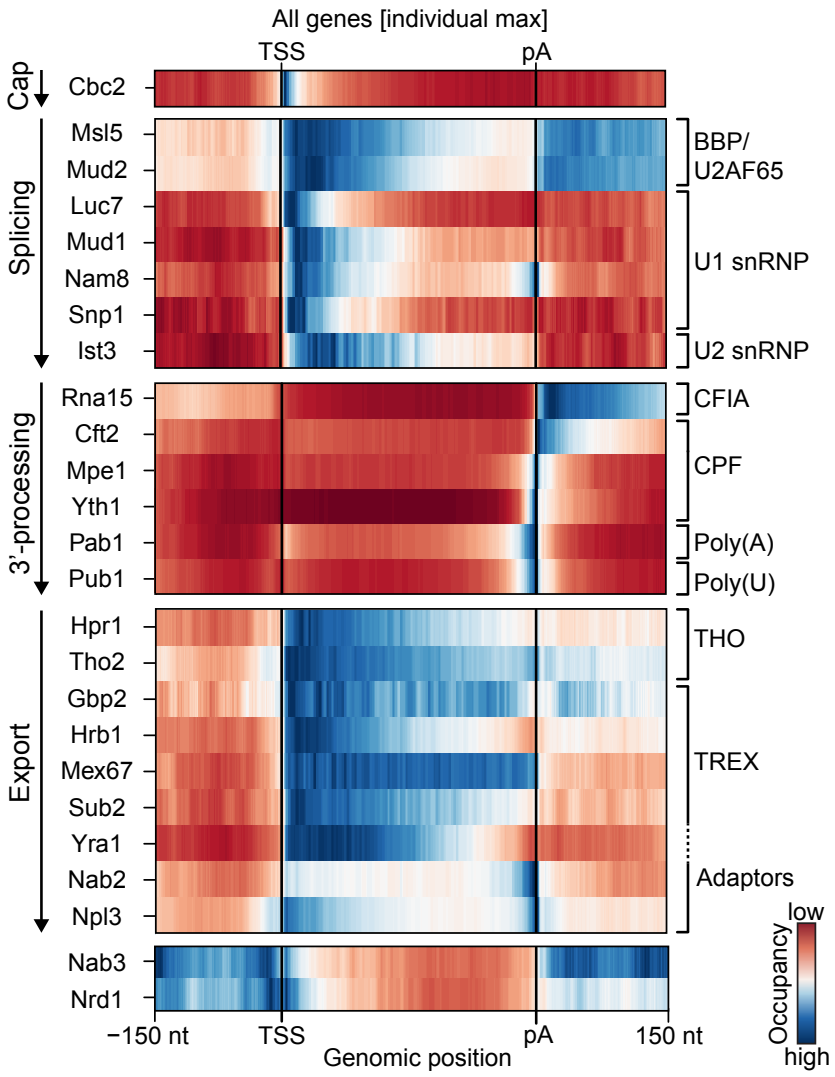

Figure 3.10: **RNA-Binding Profiles for 23 mRNP Biogenesis Factors**
Transcript-averaged occupancy profiles of mRNP biogenesis factors, scaled such that their TSSs and pA sites coincide. The color code shows the occupancy relative to the maximum occupancy per profile (dark blue).

5′ ends of transcripts, cross-validating the studies (a detailed comparison also for other factors is found in 3.11C). However, the PAR-CLIP protocol and normalization procedure used here apparently led to more focused signals at RNA 5′ ends (Figure 3.8G and 3.8H; see Figures 3.11C for comparison of other factors) and enhanced signals for short-lived RNAs and RNAs with low abundance (Figures 3.10 and 3.12A-3.12D), prompting us to use it for an investigation of factors involved in the recognition of short-lived pre-mRNA elements.

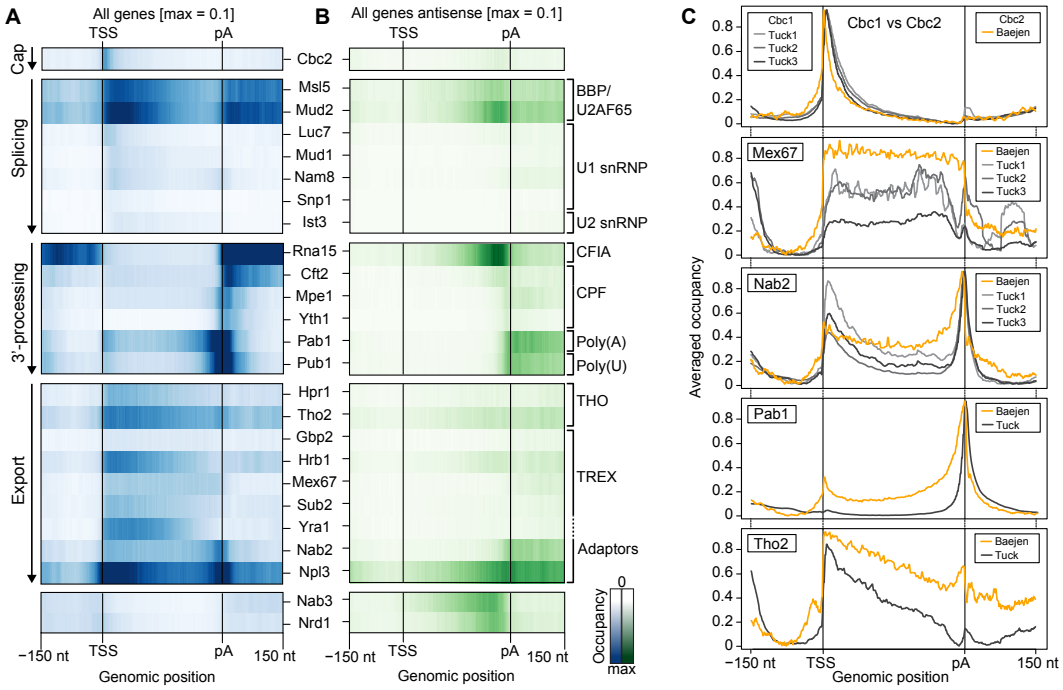

Figure 3.11: **Overview of occupancy profiles of all investigated proteins on ORF-Ts**
Smoothed occupancy profiles around all ORF-Ts were aligned at their TSS, length-scaled such that their pA sites coincided, and the occupancies averaged over all transcripts.
(A) Occupancy profiles on sense strand, i.e., on the proper mRNA.
(B) Occupany on the transcripts antisense to the annotated mRNA direction. Note the high occupancy of early termination factors Nab3 and Nrd1, termination factor Rna15, splicing factor Mud2, and export adaptor Npl3 on antisense transcripts.
(C) Comparison of transcript-averaged occupancy profiles. as measured by CRAC (black, [74]) and by PAR-CLIP (orange, this work).

3.2.5 Conserved recognition of pre-mRNA introns

Intron recognition is the initial step in pre-mRNA splicing and was extensively studied *in vitro* [81]. It begins with binding of BBP to the BP and binding of U2AF65 to a pyrimidine-rich region between the BP and 3′ splice site (3′ SS),and continues with binding of the U1 snRNP to the 5′ SS. The resulting complex E is then remodeled, and U2 snRNA displaces BBP by base pairing with the BP region, positioning U2 snRNP near the 3′ SS and giving rise to complex A (Figure 3.13G). The protein-RNA interactions underlying intron recognition may be largely conserved between yeast and human but have not been systematically analyzed *in vivo*.

Although introns are rapidly degraded *in vivo*, our protocol could capture intron sequences bound by splicing factors involved in intron recognition (Figures 3.13A, 3.13B, and 3.14A). Crosslinking signals for the BBP homolog Msl5 and the U2AF65 homolog Mud2 spanned entire introns and showed peaks near the 5′ SS and the 3′ SS, respectively (Figure 3.13D). The BP motif UACUAAC was detected around Mud2- and Msl5-bound

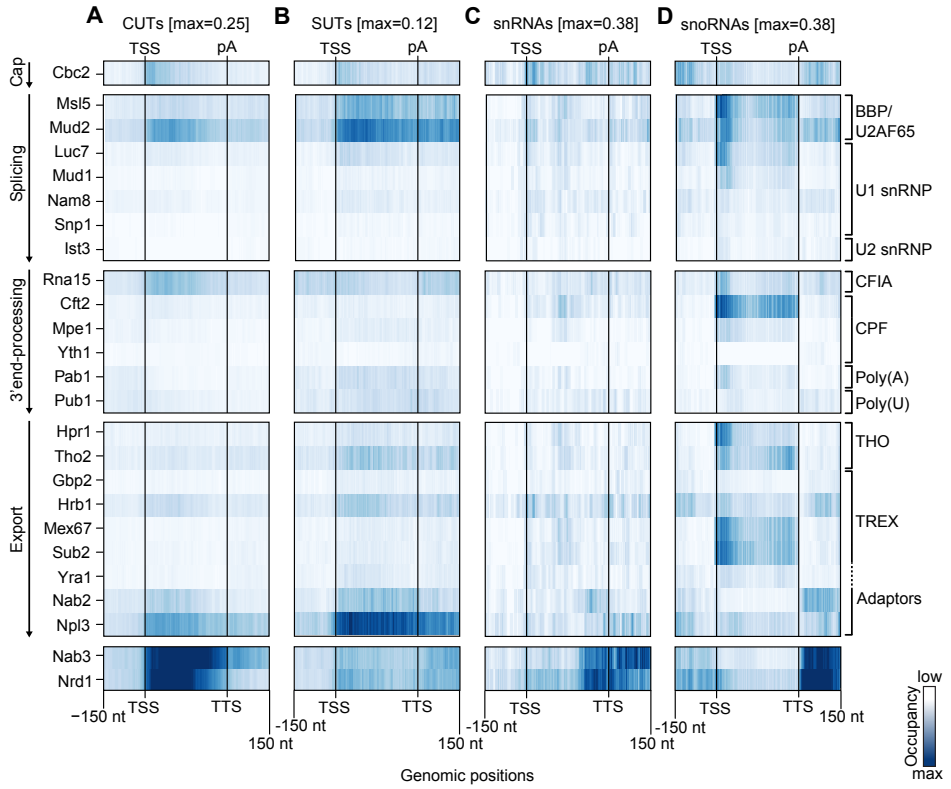

Figure 3.12: **Overview of occupancy profiles of all investigated proteins on non-coding RNAs**
For each factor, transcript class-averaged occupancies aligned at TSS and scaled to coincide at their transcription termination sites (TTS). Occupancies for each factor were divided by the maximum attained over any transcript, including all ORF-Ts.
(A) Occupancy profiles on cryptic unstable transcripts (CUTs),
(B) on stable unnotated transcripts (SUTs),
(C) small nuclear RNAs (snRNAs), and
(D) small nucleolar RNAs (snoRNAs).

sites in intron-containing genes (Figures 3.13A and 3.14A) and is generally located within ∼50 nt upstream of the 3′ SS (Figure 3.13E). When we averaged crosslink density after aligning introns at the BP, Msl5 displayed a peak on the BP (Figure 3.13F), consistent with binding of yeast Msl5 to the BP *in vivo*. Mud2 and Ist3 peaked 15 nt and 27 nt downstream of the BP, respectively (Figure 3.13F). Thus we could resolve binding of the U2AF65 homolog Mud2 to a pyrimidine/U-rich region that was defined in the human system [82, 83]. These results agree with *in vitro*-derived functions of the Msl5-Mud2 complex in BP recognition [84], and in bridging between the BP and U1 snRNP at the 5′ SS [85]. Msl5 and Mud2 also crosslinked to intron-less RNAs (Figures 3.10, and 3.11A), consistent with scanning of RNAs for U-rich regions by the U2AF65-BBP complex.

Crosslinks of U1 snRNP subunits peaked ∼17 nt downstream of the 5′ SS (Figure 3.13D). Motif searches around crosslinking peaks (±12 nt) detected the consensus 5′ SS sequence GUAUGU in Luc7, Mud1, Nam8, and Snp1 data (Figures 3.13A and 3.14A).

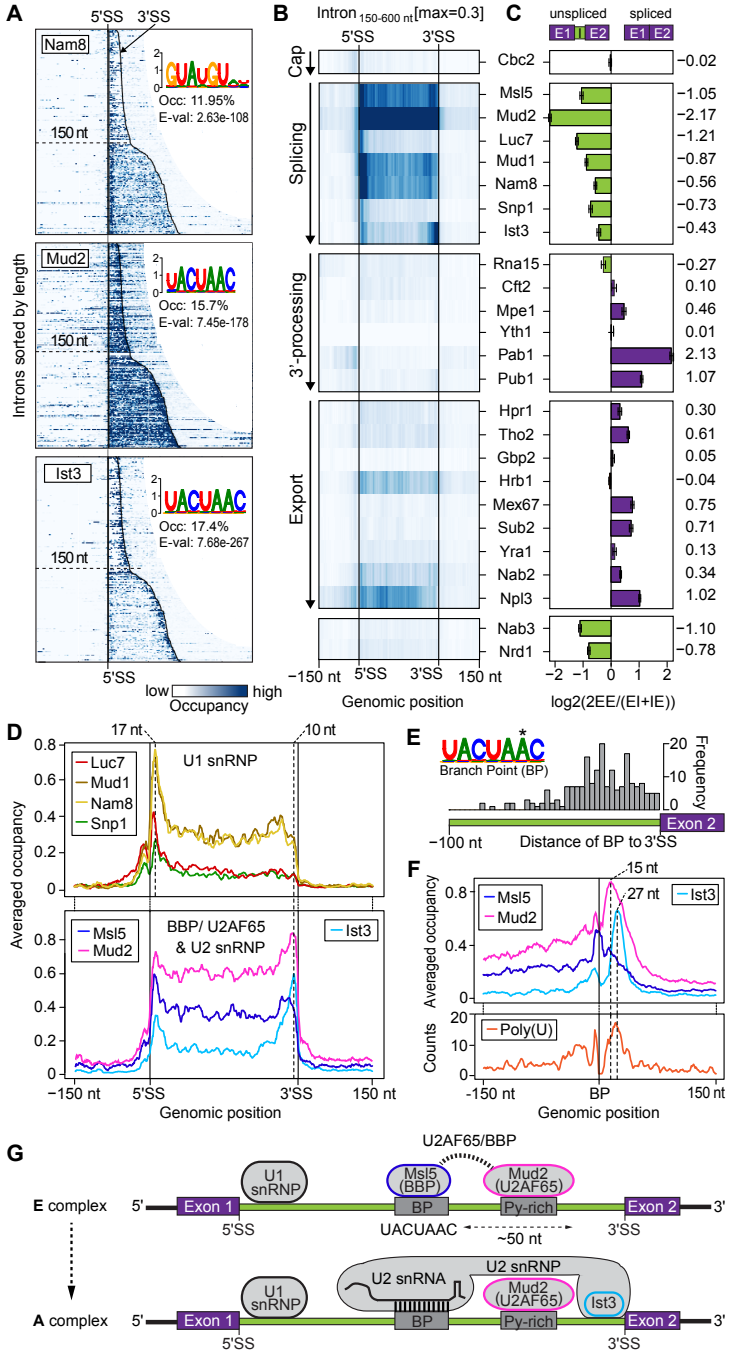

Figure 3.13: **Conserved Recognition of Pre-mRNA Introns *In Vivo***

(A) Normalized and smoothed occupancy profiles of U1 subunit Nam8, Mud2 (human U2AF65), and U2 subunit Ist3 around introns of up to 600 nt length. Introns were sorted by length and aligned at their 5′ splice site (5′ SS).

(B) Transcript-averaged occupancy profiles of all factors around introns between 150 and 600 nt length.

(C) Splicing factors show high affinity for unspliced RNAs. Splicing indices (2.6.7) indicate the binding preference for spliced versus unspliced RNAs for all factors.

(D) Intron-averaged factor occupancy profiles show binding of U1 snRNP near the 5′ SS and binding of the U2 snRNP and the commitment complex (BBP/U2AF65) over the entire intron with a peak at the 3′ splice site (3′ SS).

(E) The branchpoint (BP) lies within 50 nt upstream of the 3′ SS. Distance distribution of the branch point (BP) motif from the 3′ SS.

(F) Yeast Msl5 (human BBP) binds the BP *in vivo*, whereas Mud2 (U2AF65) and U2 snRNP (Ist3) bind downstream of the BP. Transcript-averaged occupancy profiles of Msl5, Mud2, and Ist3, centered at the BP (top), compared to the poly(U) distribution over the same region (bottom).

(G) Model of factors recognizing an intron during formation of E and A complexes.

As expected, crosslink sites of U1 snRNP subunits were not significantly enriched around the BP (Figure 3.14B). The U2 subunit Ist3 crosslinked mainly ∼10 nt upstream of the 3′ SS (Figure 3.13D). These results agree with the *in vitro*-derived binding of U1 and U2 snRNPs near the 5′ SS and the 3′ SS, respectively [81]. The splice site RNA motifs were apparently responsible for recruitment of U1 and U2 snRNPs, because their subunits generally did not crosslink to intron-less RNAs (Figure 3.11A). To investigate the order of factor binding to introns, we calculated a 'splicing index' (Figures 3.13C, 3.14C, and 3.14D, 2.6.7) [86]. All splicing factors obtained negative splicing indices, demonstrating preferential binding to unspliced RNA. The strongest preference for unspliced over spliced RNA was obtained for Mud2, the weakest for Ist3. Thus our *in vivo* data support the two-state model of intron recognition derived from *in vitro* studies (Figure 3.13G).

3.2.6 Unified recognition of pre-mRNA polyadenylation sites

In human cells, recognition of the pA site involves several RNA sequence elements that are bound by the cleavage and polyadenylation specificity factor (CPSF) complex [67, 69, 70]. The CPSF subunit CPSF-160 recognizes the pA signal (PAS) sequence AAUAAA upstream of the pA site. Subunits CPSF-100 and CPSF-30 bind neighboring U-rich sequences, and subunit CPSF-73 cleaves the RNA [67, 69]. Homologous subunits are found in the yeast CPSF counterpart CPF, which also contains additional proteins, such as Mpe1 [87].

After extensive trials we could map CPF subunits Cft2/Ydh1 (CPSF-100), Yth1 (CPSF-30), and Mpe1 onto pre-mRNA (Figure 3.15A). Cft2 crosslinked to regions flanking the pA site, consistent with binding near the cleavage site *in vitro* [88]. Yth1 showed a peak ∼17 nt upstream of the pA site, consistent with *in vitro* results [89], and with localization of its human counterpart CPSF-30 *in vivo* [73]. Mpe1 gave rise to a peak ∼6 nt upstream the pA site, explaining why it is an essential factor required for 3′ processing [87]. Although Cft1/Yhh1 (CPSF-160) and Ysh1 (CPSF-73) did not show PAR-CLIP signals, these data locate the yeast CPSF counterpart CPF at the pA site *in vivo* and define many of its subunit-RNA interactions.

Human CPSF is assisted by the CstF complex, which binds to pre-mRNA downstream of CPSF [67, 69]. However, the yeast CstF counterpart CFIA is believed to bind upstream of the CPSF counterpart CPF [67, 69], and this model is based on *in vitro* evidence that the CFIA subunit Rna15 binds upstream of the pA site [90]. In contrast, we observed very strong crosslinking of Rna15 downstream of the pA site *in vivo*, with a peak at ∼16 nt (Figure 3.15A). These results agree with an alternative *in vitro* study [88], and with binding of the human Rna15 homolog CstF64 downstream of the pA site *in vivo* [73]. The PAR-CLIP peak for Rna15 downstream of the pA site is also consistent with an occupancy peak observed in the same region by chromatin immunoprecipitation [91]. Thus CFIA is

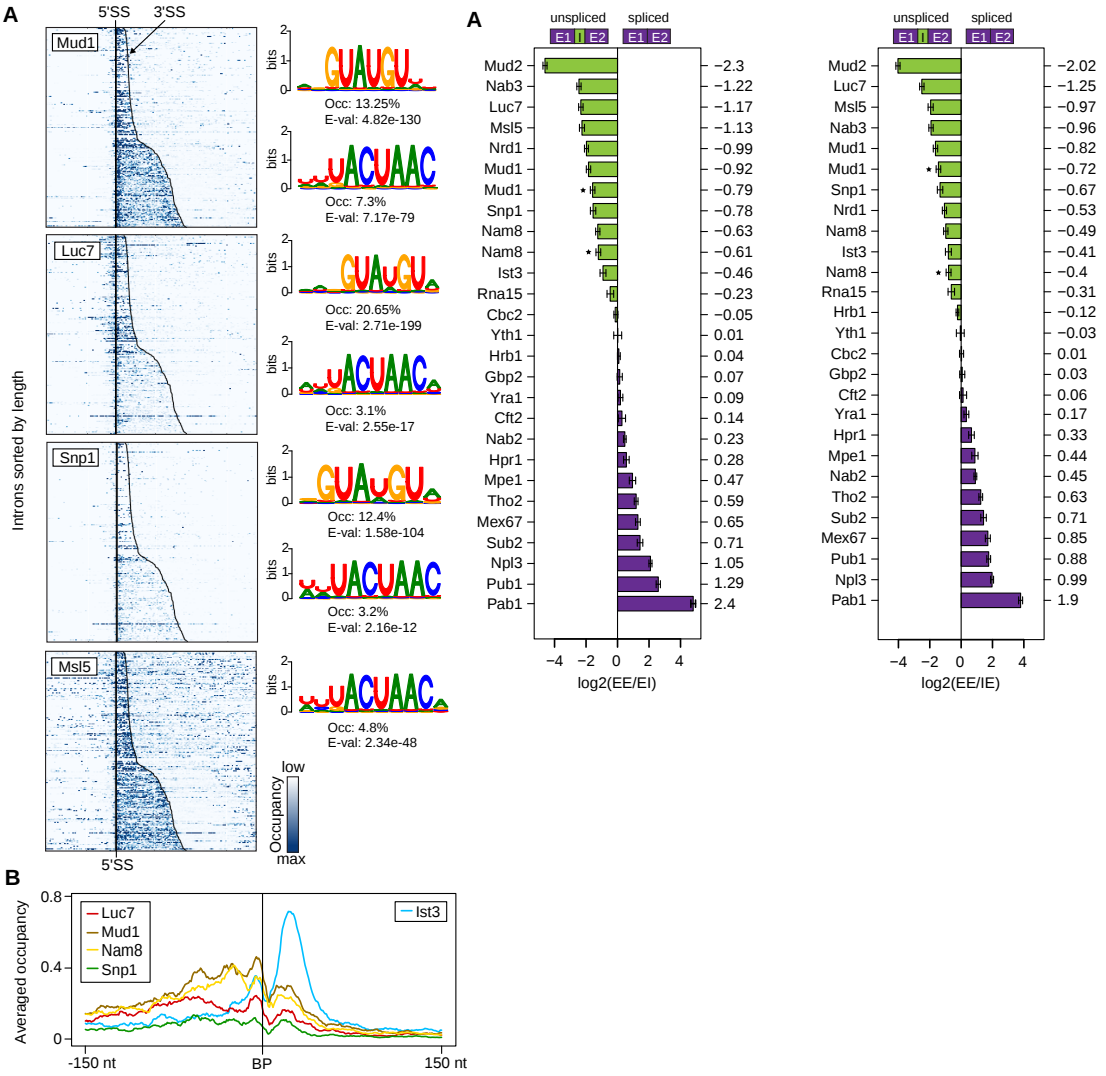

Figure 3.14: **Occupancy of splicing factors around introns and the branch point (BP)**
The splicing index is robust with respect to using the coverage of coverage exon-intron or intron-exon junctions.
(A) Occupancy profiles of the U1 snRNP splicing factors Mud1, Luc7, and Snp1, and the BBP/Msl5 derived from PAR-CLIP experiments for all introns. Each line represents an intron, and introns are sorted by length and aligned at their 5′-SS. Motifs found by XXmotif to be enriched ±20 nt around the cross-linking sites are shown next to the factors around which they are enriched.
(B) Average occupancy profiles for the U1 snRNP Luc7, Mud1, Nam8 and Snp1, and the U2 snRNP Ist3 around the branch point (BP).
(C) Splicing index calculated using coverage of exon-intron (EI) junctions.
(D) Splicing index calculated using coverage of inton-exon (IE) junctions.

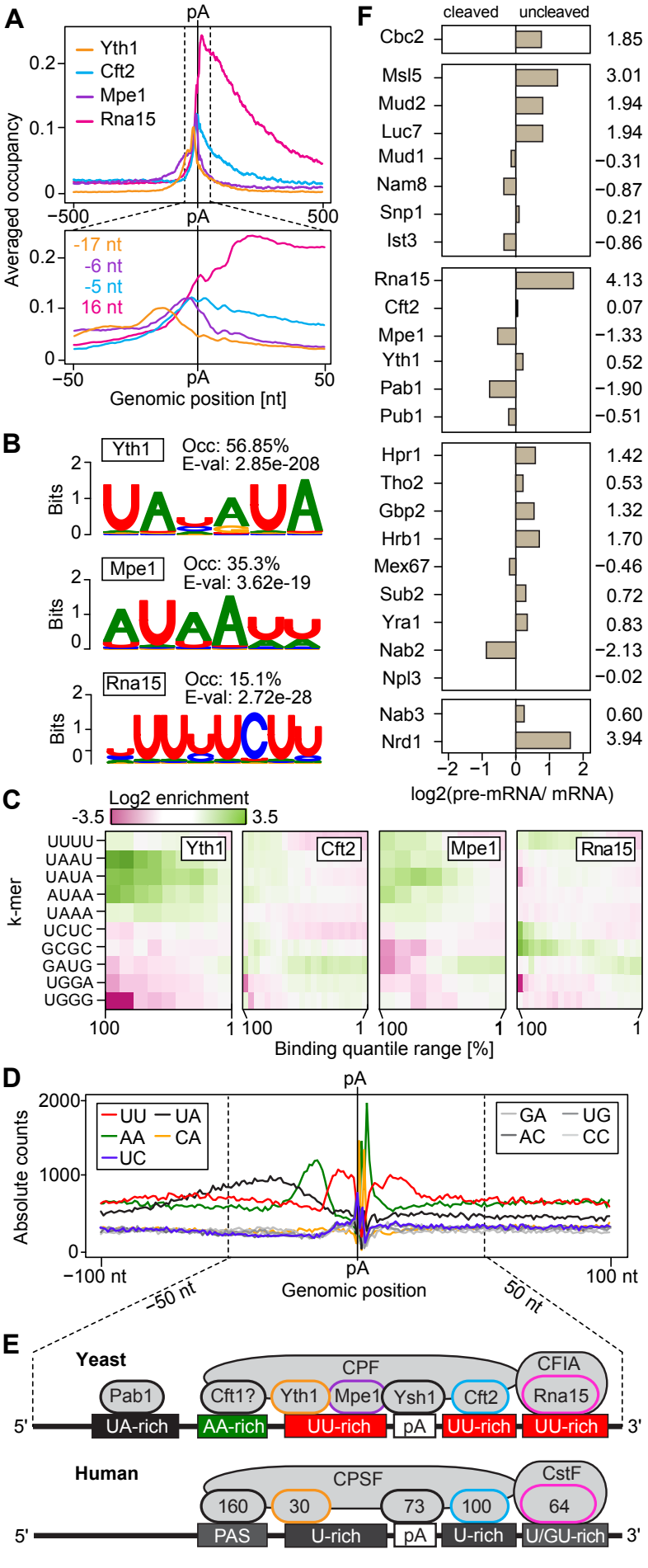

Figure 3.15: **Unified Model for Polyadenylation Site Recognition *In Vivo***

(A) CFIA subunit Rna15 binds downstream of CPF complex and the pA site. Averaged occupancy profiles of Rna15 and CPF subunits Cft2, Mpe1, and Yth1, aligned at the pA site show that CPF binds at the pA site, whereas CFIA binds downstream *in vivo*.

(B) RNA motifs enriched in a window of ±25 nt around the crosslinked sites with fraction of occurrence and XXmotif E-value.

(C) RNA 3′ processing factors have distinct tetramer-binding preferences. Shown are log-odd scores for enrichments of selected tetramers (y axis) for bins of binding sites ranging from 100% to 1% occupancy (x axis).

(D) Sequences around the pA site exhibit an 'A-U bias'. Shown is the distribution of nucleotide composition around the pA site.

(E) Unified model for pA site recognition in *S. cerevisiae* and human by the two major, conserved 3′ processing complexes CPF (CPSF) and CFIA (CstF) bound on pre-mRNA.

(F) Processing indices measuring the tendency of the factors to bind to uncleaved pre-mRNA rather than cleaved RNA, computed as log2 odds ratios uncleaved versus cleaved RNA bound by the factor.

located downstream, rather than upstream, of the pA site and CPF, consistent with the position of the human CstF complex downstream of the pA site and downstream of the CPF counterpart CPSF. These results lead to a unified model for pA site recognition by the two conserved 3′ processing complexes bound to pre-mRNA (Figure 3.15E).

3.2.7 Definition and decoration of mRNA 3′ ends

To investigate how RNA sequence-binding preferences of processing factors define the pA site, we searched for sequence motifs around crosslinking peaks. Peaks for Yth1 and Mpe1 often contained the motifs UAUAUA ('efficiency element'; [92, 93]) and AUAAUU, respectively, and Cft2, Mpe1, and Yth1 generally preferred RNA sites containing U/A-rich tetramer sequences (Figure 3.15C). Rna15 did not show enrichment for the 'positioning element' AAUAAA (Figure 3.15C) that had been reported to bind *in vitro* [90]. Instead, it preferentially bound regions *in vivo* that were enriched with the A-less tetrameric motifs UUUU and UCUC (Figure 3.15C), and a motif search yielded the sequence UUUUCUU (Figure 3.15B) that is very similar to the downstream U-rich element [93]. More generally, pA site regions exhibit a bias in nucleotide composition. Whereas the immediate vicinity of the pA site showed an enrichment of UU and AA dinucleotides up- and downstream, respectively, flanking regions showed the inverse, a phenomenon we refer to as A-U bias (Figures 3.15D).

The A-U bias apparently directs binding of CPF subunits upstream and around pA sites and binding of Rna15 downstream of pA sites, due to corresponding sequence preferences of these factors (Figure 3.15C). In some yeast mRNAs, the A-rich upstream region contains a positioning element [90] that may bind Cft1 and may correspond to the human polyadenylation signal [94], and an UA-rich efficiency element [92] that may bind CFIB/Hrp1 [95]. These two elements are, however, dispensable for RNA cleavage *in vitro* [88], consistent with our view that the A-U bias, rather than specific sequence elements, underlies pA site recognition. A similar A-U bias was observed around human pA sites [73] and befits the conserved arrangement of 3′ processing factors revealed here.

Additional data showed that the AU/UU-rich regions upstream of pA sites bind Pab1 and Pub1 (Figure 3.16). Both factors gave rise to crosslinking near the 3′ end of mRNAs (Figures 3.16A and 3.16C). Pab1 bound upstream of the pA site to the UAUAUA 'efficiency element' motif (Figures 3.16A and 3.16B) as described [59, 74], and showed some depletion at the Yth1 site (Figures 3.15E and 3.16C). Pub1 occupied both UA-rich regions in the 3′ UTR [96, 97] and poly(U) tracts (Figure 3.16B) but also bound upstream of the open reading frame (ORF) in the 5′ UTR (Figure 3.16D) as described [98, 99, 100]. Pub1 and Pab1 were generally depleted from the translated ORF (Figures 3.16C and 3.16D), indicating that these factors are displaced during translation in the cytoplasm. Taken together, these data may be explained as follows. The two major 3′ processing

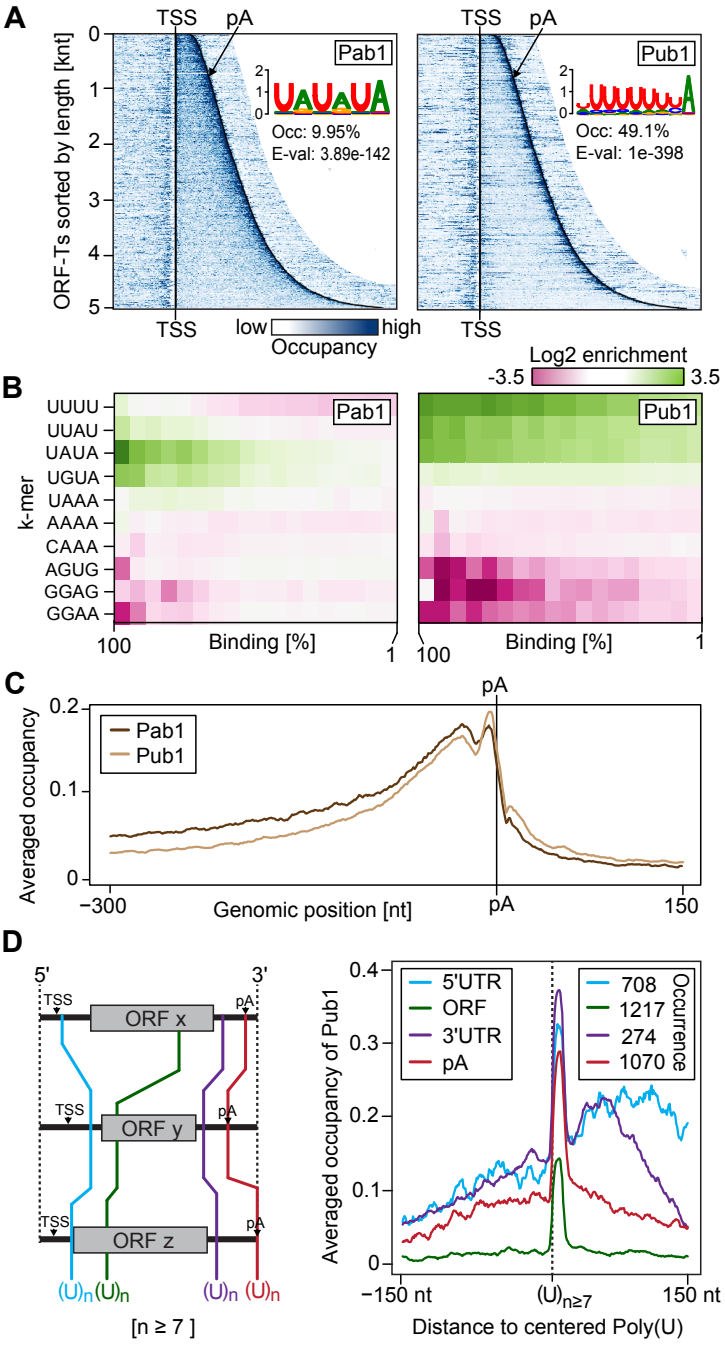

Figure 3.16: **Pab1 and Pub1 Bind UA-and U-Rich Sequences at mRNA 3′ Ends**
(A) Normalized and smoothed occupancy profiles of the poly(A)-binding protein Pab1 and the poly(U)-binding protein Pub1 derived from PAR-CLIP data in sense direction for all ORF-Ts. ORF-Ts were sorted by length and aligned at their transcription start site (TSS). The motifs were enriched around binding sites (±25 bp).
(B) Pab1 and Pub1 bind to U/A-rich sequences. Log2 enrichment of selected tetramer motifs around Pab1 (left) and Pub1 (right) binding sites compared to unbound sequence regions, analyzed within 18 equal-sized bins of occupancy quantiles between 100% and 1% site occupancy (x axis).
(C) Averaged occupancy profiles of Pab1 and Pub1 derived from PAR-CLIP data in sense direction for all ORF-Ts, centered at the pA site of all ORF-Ts.
(D) Pub1 preferentially binds poly(U)n ≥ tracts near the pA site. Average occupancy profiles of Pub1 around Poly(U)n ≥ tracts within the 5′ UTR, ORF, or 3′ UTR, or near pA sites.

complexes CPF and CFIA bind to pre-mRNA regions with an A-U bias and pA site, causing RNA cleavage and polyadenylation, and subsequent release of 3′ processing factors, which enables complete decoration of the mRNA 3′ end with Pab1 and Pub1.

3.2.8 Transcription-coupled mRNP export

In our current view, mRNA export begins with the recruitment of the THO/TREX complex during Pol II elongation [101, 102]. Mature mRNA is then exported from the nucleus by the heterodimeric export factor Mex67-Mtr2 [103, 104]. Mex67 uses mRNA adaptor proteins such as Nab2, Npl3, and Yra1 [105, 106, 107, 108]. PAR-CLIP analysis revealed similar distributions of the THO subunits Tho2 and Hpr1 over mRNAs (Figure 3.17A) and no mRNA preferences, indicating that the THO complex is a general factor associated with Pol II transcripts. Tho2 gave stronger signals, consistent with its role in THO complex recruitment [109, 110]. Mex67 bound RNA *in vivo* (Figure 3.17B), explaining how it remains bound to mRNA after release of adaptor proteins. Mex67 did not show preferences for RNA motifs, consistent with its function as a general export factor, and consistent with data obtained by CRAC ([74], Figure 3.11C). The export adaptors Nab2, Npl3, and Yra1 showed different crosslinking patterns, indicating specific, nonredundant functions (Figure 3.17C). The number of mRNAs bound by two or three export adaptors was limited (Figure 3.17C), showing that these factors exhibit mRNA preferences, as suggested by purification of mRNAs associated with Yra1 [111]. Yra1 occupancy decreased before the pA site, whereas Npl3 also showed crosslinking at 3′ ends, consistent with its influence on pA site choice [112, 113]. Whereas Nab2 preferentially bound short mRNAs (Figure 3.17C), Yra1 and Sub2 preferred long mRNAs (Figure 3.17D). Nab2 crosslinking density was also stronger at the 3′ ends of ORF-Ts as described (Figure 3.17A) [74], consistent with its known influence on 3′ processing [114, 115, 116, 74]. Nab2 sites were enriched for the motif GUAG (Figure 3.17C) as described [59]. Thus components of the mRNA export machinery show preferences for RNAs with specific sequences and lengths.

3.2.9 Global analysis links splicing to 3′ processing

We now subjected all PAR-CLIP data to a global analysis (Figure 3.18). In addition to the splicing index (Figures 3.13C, 3.14C), we introduced a 'processing index' (2.6.8) that estimates whether factors preferentially bind uncleaved or cleaved RNA (Figures 3.15F). A plot of splicing versus processing indices (Figure 3.18A) indicates how the composition of protein-RNA complexes is remodeled during mRNP biogenesis (Figure 3.18B). We further calculated for each pair of factors the Pearson correlation coefficient of the total weighted occupancies over transcripts (Figure 3.18C, 2.6.10). This estimates the extent to which factors co-occupy the same transcripts. We further determined the extent to which two factors colocalize in a window of 25 nt around binding sites (Figure 3.18D, 2.6.11). Finally, we computed for each pair of factors the Pearson correlations between their averaged occupancy profiles, to measure the shape similarity of binding profiles (Figure 3.19, 2.6.9).

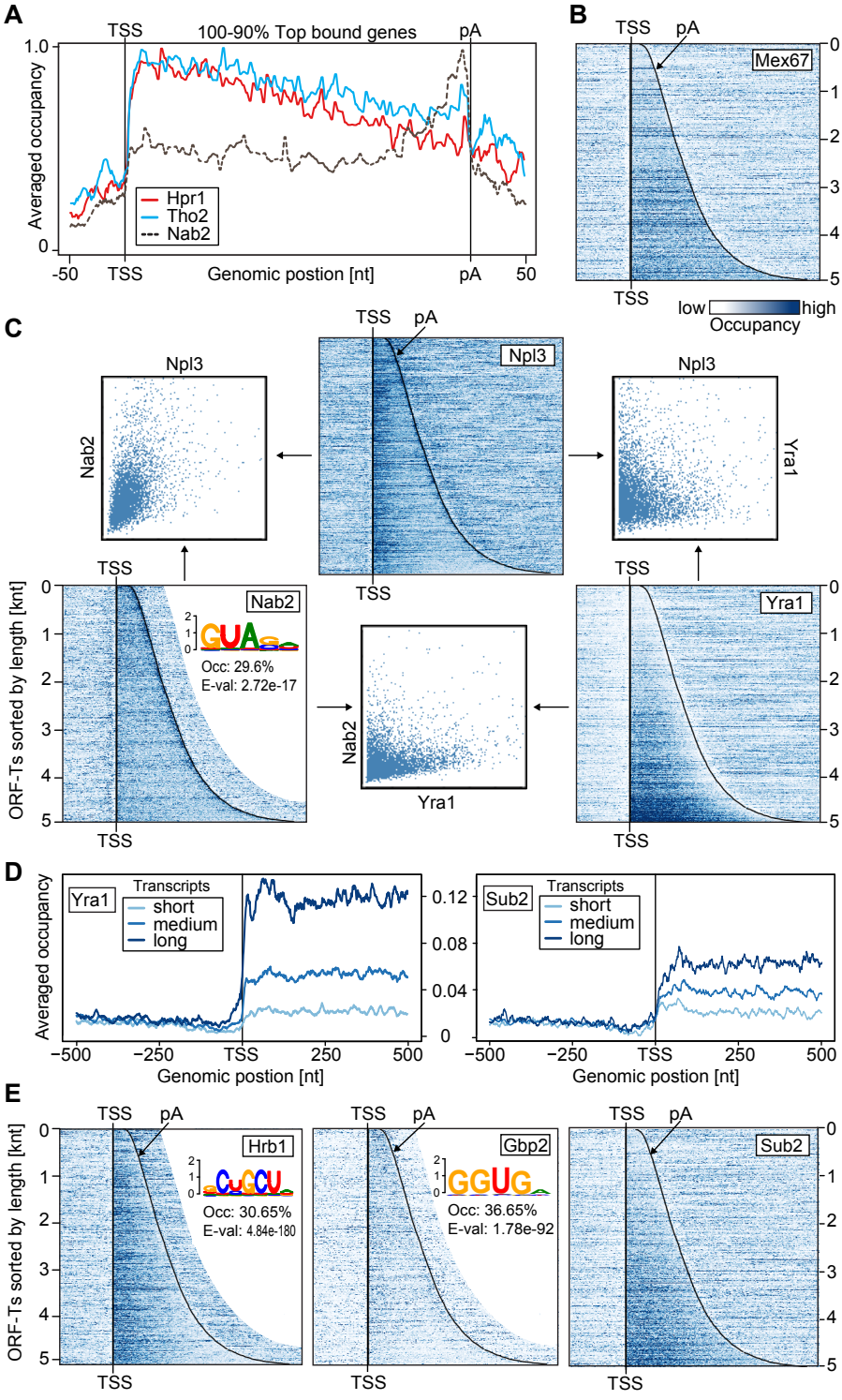

Figure 3.17: **Export Adaptors Differ in Their mRNA-Binding Preference**
(A) Normalized and smoothed average occupancy profiles for Nab2 and the THO complex subunits Tho2 and Hpr1, derived from PAR-CLIP experiments for top-bound ORF-Ts (100%˘90%).
(B) Normalized and smoothed average occupancy profiles for the general export factor Mex67, derived from PAR-CLIP experiments for all ORF-Ts, sorted by length and aligned at their TSS.
(C) The export adaptors Yra1, Npl3, and Nab2 have distinct mRNA-binding preferences. Pairwise correlation scatterplots for occupancies of Yra1, Npl3, and Nab2 on ORF-Ts.
(D) Transcript-averaged Yra1 and Sub2 occupancies in sense directions, centered at the TSS, for short (0–1 kb), medium (1–2 kb), and long (2–5 kb) transcripts.
(E) Occupancy profiles for SR-like proteins Hrb1 and Gbp2 and for Sub2, derived from PAR-CLIP experiments for all ORF-Ts, sorted by length and aligned at their TSS.

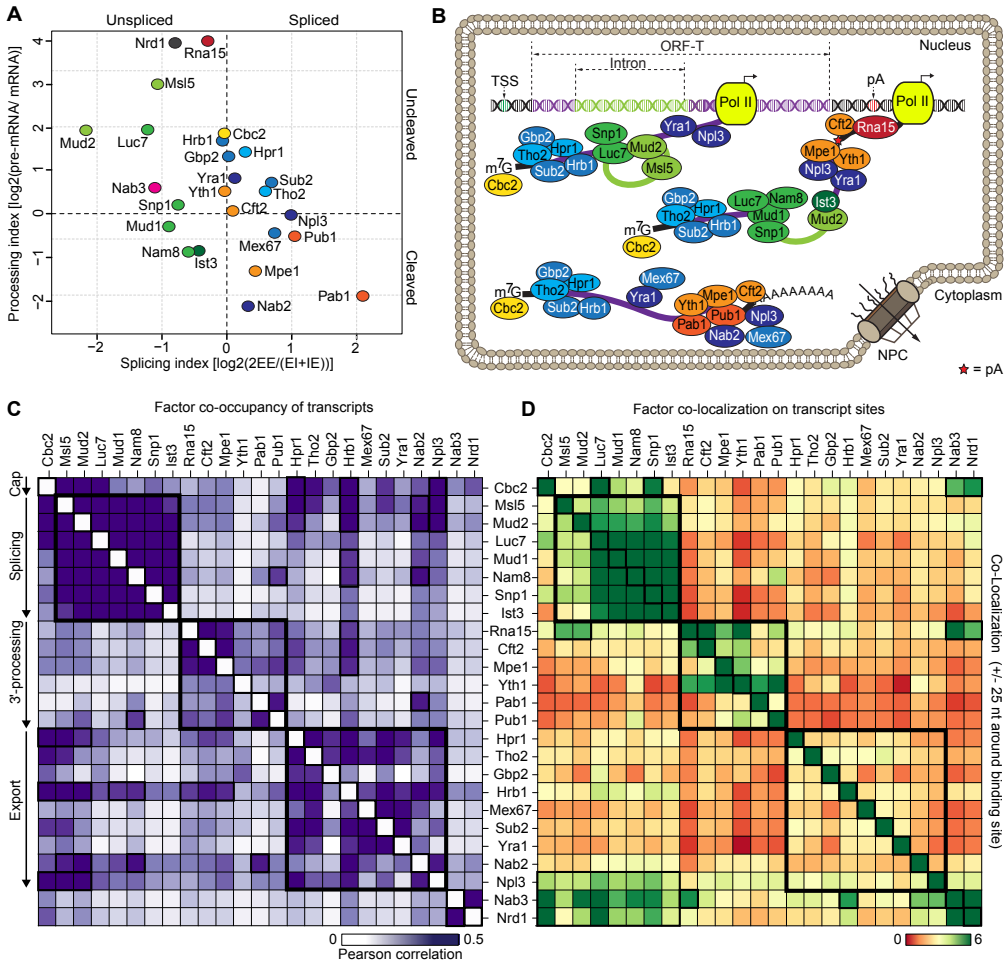

Figure 3.18: **Global Analysis Reveals Links between Splicing, 3′ Processing, and Export**
(A) 'Splicing index' and 'processing index' for all analyzed factors (yellow, capping; orange/red, 3′ processing; green, splicing; blue, export; black/pink, RNA surveillance). A splicing index of 0 (1) indicates binding only to unspliced (spliced) mRNA. Similarly, a processing index of 0 (1) signifies binding only to uncleaved pre-mRNA (cleaved mRNA). The splicing index is averaged over all intron-containing ORF-Ts; the processing index is averaged over all ORF-Ts.
(B) Model for mRNP biogenesis resulting from PAR-CLIP-based occupancy measurements.
(C) Factor co-occupancy of transcripts. Pairwise Pearson correlation coefficient of the total weighted occupancies over entire transcripts for all factors.
(D) Factor colocalization on transcript sites. Colocalization for each pair of factors in a window of 25 nt around the centered binding sites (column).

This global analysis provided evidence for an ancient link between splicing and 3′ processing. Splicing factors fell into two groups when sorted by their processing index (Figure 3.15F and 3.18A). The splicing factors Mud2, Msl5, Snp1, and Luc7 preferentially bound uncleaved RNA, whereas other splicing factors preferred cleaved RNA (Figure 3.18B). Mud2 and Msl5 profiles were correlated with those of 3′ processing factors Rna15 and Cft2, and Nam8 correlated with Mpe1, Pab1, and Pub1 (Figure 3.19). Also, Mud2, Msl5, and Nam8 crosslinked near the pA site (Figure 3.10), although this could also reflect

splicing-independent functions of these factors at pA sites. Nam8 tended to colocalize with Pub1, whereas Mud2 and Msl5 tended to co-occupy transcripts with Hpr1, Hrb1, Nab2, and Npl3, and they colocalized with Rna15 (Figures 3.18C and 3.18D). Indeed, Rna15 preferentially bound unspliced mRNAs (Figures 3.13C and 3.18A), but also showed the lowest processing index (Figures 3.15F and 3.18A), confirming its early binding to pre-mRNA [94, 117].

These results indicate that the machineries for splicing and 3′ processing interact in yeast, as inferred by genetics [118], although it is currently believed that such an interac-

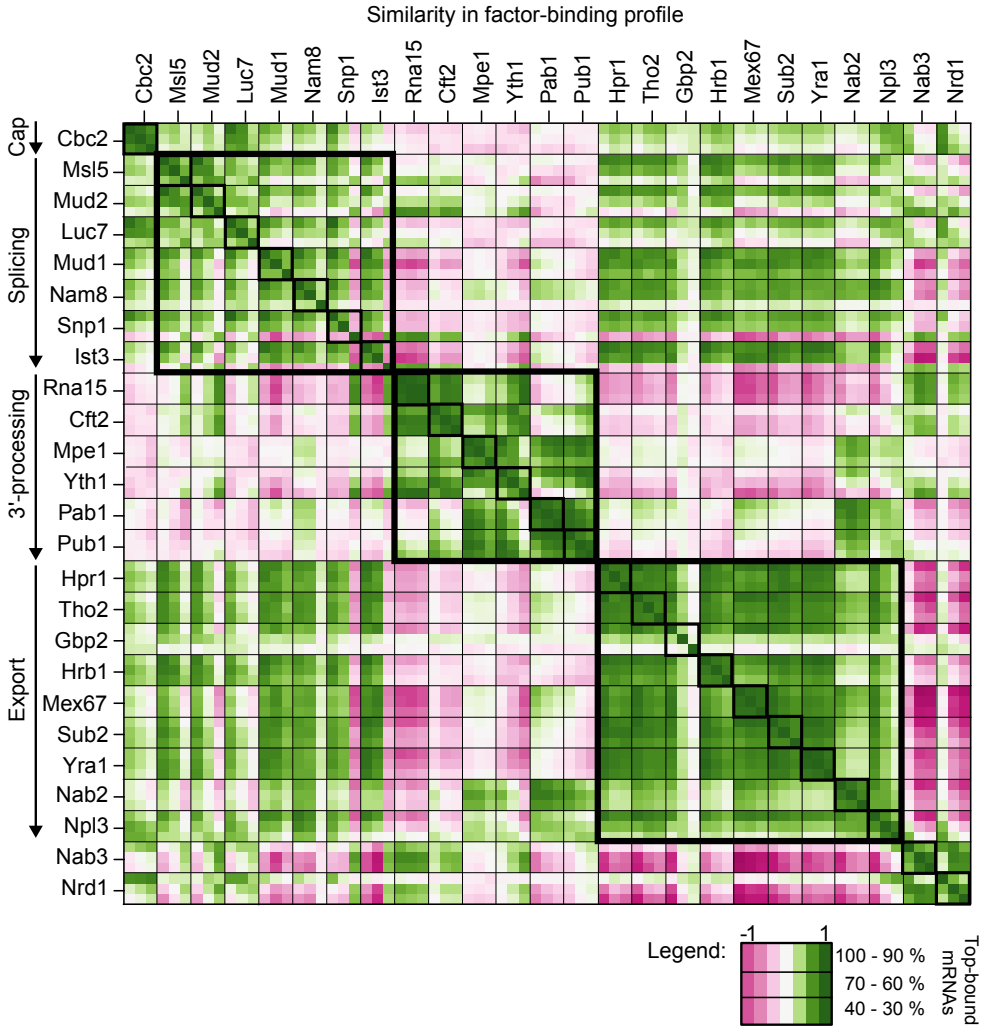

Figure 3.19: **Similarity matrix of factor-binding profiles**
Similarity matrix of factor-binding profiles shows blocks of functionally linked factors with similar occupancy profiles along transcripts For each pair of factors, the matrix shows the color-coded Pearson correlations between the occupancy profiles averaged over ORF-Ts in occupancy quantile ranges $100\% - 90\%$, $70\% - 60\%$ and $40\% - 30\%$. Each cell corresponding to a pair of factors thus shows 3×3 color coded Pearson correlations. In cells that are entirely green in all 3×3 subcells, profiles are similar to each other across all three quantile ranges of occupancy. Note that export factors profiles show similarity to splicing factor profiles.

tion is restricted to mammalian cells [119, 70, 120]. Indeed, 3′ processing may assist in splicing, but not the other way around, because unspliced and spliced transcripts recruit 3′ processing factors to a similar extent (data not shown). This model is consistent with the known stimulation of splicing by 3′ processing in human cells [121].

3.2.10 Transcript surveillance and fate

The global analysis also elucidated how nuclear export is restricted to mature mRNPs. First, export factors preferred spliced over unspliced mRNA, and generally did not bind uncleaved RNA (Figures 3.13C and 3.18A). The highest splicing index was found for Pab1, which binds mature mRNA [122], whereas the lowest splicing index was found for Mud2, which is expected to initiate intron recognition [81]. Second, binding profiles for export factors except Nab2 differed from those of 3′ processing factors (Figure 3.19), reflecting that export factors select 3′ processed mRNAs. Third, the SR proteins Gbp2 and Hrb1 [123] overlapped with THO/TREX subunits, and Hrb1 tended to bind the same transcripts as the Mud2-Msl5 complex (Figures 3.18C). This is consistent with a role of Gbp2 and Hrb1 in restricting mRNA export to spliced transcripts [124]. Gbp2 and Hrb1 showed distinct RNA-binding motifs (Figure 3.17E), and Hrb1 colocalized with splicing factors Luc7 and Snp1 (Figure 3.18D), consistent with a role in splicing [125, 126, 81].

A subset of 3′ processing factors also showed occupancy profiles that were similar to those of RNA surveillance factors Nrd1 and Nab3 (Figure 3.19). Rna15 colocalized with Nrd1 and Nab3 on transcripts (Figure 3.18D) and crosslinked to aberrant divergent ncRNAs (Figure 3.11). This indicates that some 3′ processing factors are part of the RNA surveillance machinery that terminates and degrades aberrant RNAs, as predicted by genetics [25]. Nrd1 and Nab3 colocalized with Cbc2 (Figure 3.18D) and preferentially bound uncleaved pre-mRNA, in accordance with their role in triggering early termination of transcription. Nrd1 and Nab3 colocalized with splicing factors on introns (Figures 3.13C and 3.18D), likely because of their preferential binding to noncoding RNA regions [24]. These observations are consistent with a general nuclear RNA surveillance pathway and suggest that a transient surveillance/3′ processing complex takes a decision during RNA synthesis of whether a transcript is subjected to degradation or to polyadenylation and nuclear export.

3.2.11 Conclusion

Here we report high-confidence transcriptome maps for 23 protein factors involved in mRNP biogenesis in the eukaryotic model system *S. cerevisiae*. We demonstrate that PAR-CLIP efficiently captures short-lived unspliced and uncleaved pre-mRNAs. This allowed mapping of splicing factors onto introns and of 3′ processing factors within regions downstream of the pA site, which are rapidly removed and degraded in cells. The distri-

bution of factors over various pre-mRNA species that result from events during mRNP biogenesis enabled integration of the data into a model for mRNP biogenesis based on factor occupancy.

The three most notable insights from our data include (1) the observation of intron recognition by the Mud2-Msl5 (human U2AF65-BBP) and the snRNPs U1 and U2 *in vivo*, (2) a unified, conserved arrangement of the two major 3′ processing complexes CPF and CFIA (human CPSF and CstF) around the pA site, and (3) links of the 3′ processing machinery to RNA splicing and nuclear RNA surveillance. An analysis of the RNA sequences underlying the crosslinked sites recovered known splicing motifs and revealed a conserved 'A-U bias' at the pA site. It also defined eight specific RNA motifs bound by biogenesis factors, of which three were new, and showed that most factors exhibited binding preferences for certain RNA tetrameric motifs.

Our results support the emerging concept that RNA-binding factors, in contrast to DNA-binding factors, generally show binding preferences, rather than specificities, and exhibit site promiscuity. To achieve high target specificity, multiple interactions of RNA-binding subunits within a functional complex are required and/or additional protein interactions of factors must occur, such as binding to the Pol II CTD. Synergistic factor binding is evident within the machineries for splicing and 3′ processing. It explains how sites in pre-mRNA can be located with confidence despite little sequence conservation and a scarcity of motifs in RNA. It also explains how mRNA, which is restricted in its sequence due to its coding nature, can evolve to specifically bind multifactor complexes.

Finally, global analysis of our data revealed that processes involved in mRNP biogenesis are more tightly coupled than generally thought. An ancient link between 3′ processing and splicing apparently coordinates both processes and generates mature mRNPs that are selected for nuclear export. In particular, we observed direct RNA interactions of splicing factors at the pA site and a differential distribution of splicing factors on pre-mRNAs before and after their 3′ cleavage. How 3′ processing may influence spliceosome dynamics and how the composition of protein-RNA complexes is remodeled several times during mRNA biogenesis may be analyzed in the future.

3.3 Nrd1, Nab3, Sen1 binding site analysis

Parts of the analysis of the PAR-CLIP data presented in this chapter were published in Schulz and Schwalb et al. [24] and served as input for further analysis in Schulz and Schwalb [24]. Here, I summarize the Nrd1/Nab3 binding site analysis provided by `STAMMP` and add the binding site information of Sen1.

3.3.1 Summary

The transcriptome of eukaryotic genomes consists of more transcripts as expected by annotations. A characteristic that is described as pervasive transcription. In *Saccharomyces cerevisiae* it could be shown that a major fraction of ncRNA originates from bi-directional promoters [56] with two individual pre-initiation complexes (PICs) [78]. However, in order to maintain proper cellular functions the generation of ncRNA should be avoided. With a conditional depletion of Nrd1 from the nucleus of *Saccharomyces cerevisiae* using the anchor-away technique [127] and an additional genome wide monitoring of transcriptional changes using 4tU-seq [128] Schulz and Schwalb et al. could show, that Nrd1 globally restricts ncRNA transcription by early termination [24]. The coupling of anchor-away and 4tU-seq lead to a new annotation of 1 526 Nrd1-dependent unterminated transcripts (NUTs) and showed that 59% of all NUTs originate antisense from 5′ and 3′ nucleosome depleted regions (NDRs) flanking known genes [24]. Nrd1 is known to recognize a primary sequence element in RNAs [45] and interacts with Nab3 and Sen1 [26]. In concordance, the analysis of PAR-CLIP data shows that most binding sites of Nrd1, Nab3 and Sen1 are located antisense of the TSS and pA signals of known transcripts.

3.3.2 Nrd1,Nab3 and Sen1 primarily target antisense ncRNA originating from NDRs at transcription start sites and pA sites of transcripts

In contrast to the published data in [24] the PAR-CLIP signals shown here are normalized with the RNA-seq signal from Bäjen and Torkler et al. [44] which leads to similar results (Figure 3.20A, compare to Figures 3A and 3B from [24]). The normalization in Schulz and Schwalb was done with 4-tU-seq data after nuclear depletion of Nrd1 to avoid potential normalization biases, due to sequencing problems of short lived ncRNAs [24]. However, the antisense ncRNA data is analog between 4-tU-seq and RNA-seq normalized data (Figure 3.20A and Figure 1B from [24]) indicating that RNA-seq data can be used for reliable normalization even of ncRNA. Interestingly, the normalization with RNA-seq data leads to weaker Nrd1 signals in sense transcripts compared to the 4-tU-seq normalization (Figure 3.20A and Figure 1A from [24]).

A comparison of transcript averaged occupancies of Nrd1 and its interaction partners Nab3 and Sen1 shows similar binding patterns (Figure 3.20B). Sen1 and Nrd1 show strikingly similar patterns for both the binding to antisense transcripts originating from NDRs at TSSs and pAs. Although the binding pattern of Nab3 is almost identical to Nrd1 and Sen1 the binding of Nab3 is weaker at antisense transcripts originated from NDRs at TSSs compared to the Nrd1 and Sen1 occupancies. Conversely, the occupancy of Nab3 is stronger for antisense transcripts originated from NDRs at pAs compared to the occupancies of Nrd1 and Sen1 (Figure 3.20B).

Nrd1/Nab3/Sen1 show their strongest occupancy at the same genomic locations at

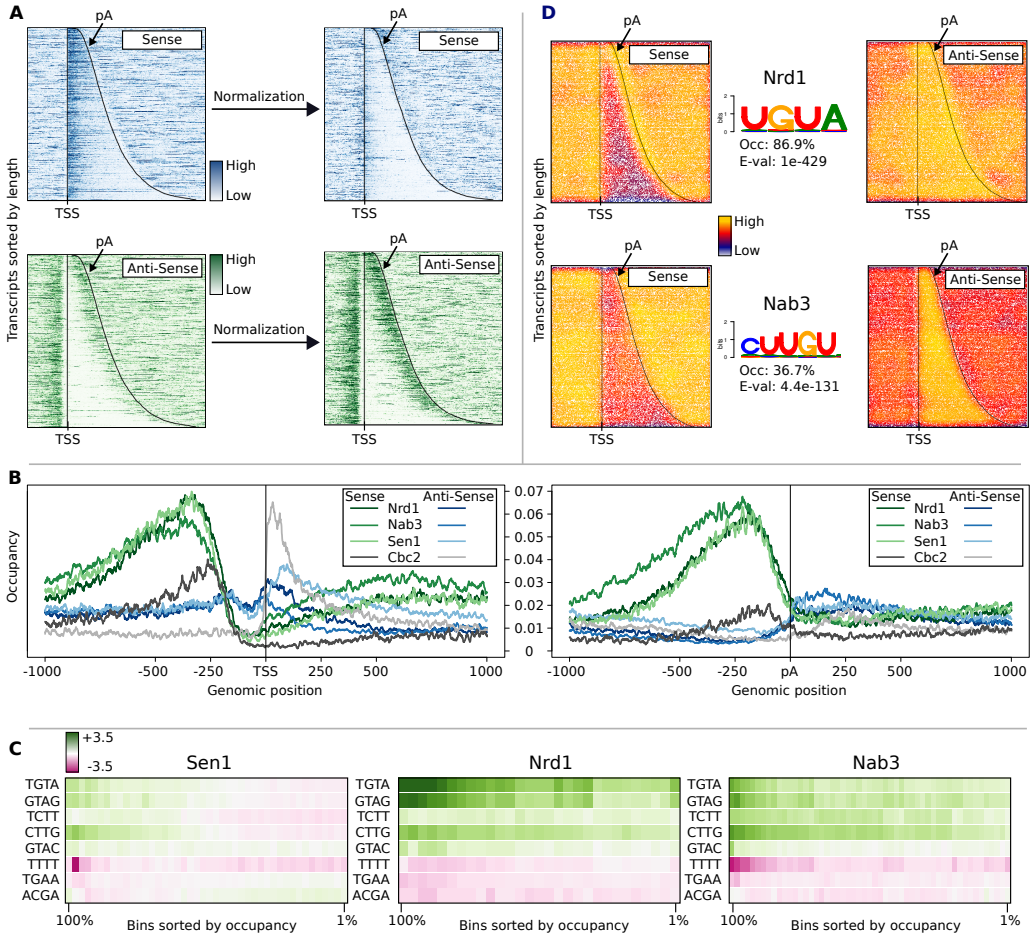

Figure 3.20: **Nrd1,Nab3 and Sen1 binding targets in *Saccharomyces cerevisiae***
(A) Smoothed Nrd1 unnormalized (left) and normalized (right) occupancies in sense (blue) and antisense (green) direction for all open reading frame-containing transcribed regions (ORF-Ts). ORF-Ts are sorted by length and aligned at their transcription start site (TSS).
(B) Transcriped-averaged occupancies for Nrd1/Nab3/Sen1 and Cbc2 in sense and antisense direction centered at the transcriptional start site (TSS) (left) and the polyadenylation site (pA) (right).
(C) Log2 enrichment of selected tetramer motifs (y axis) around Sen1 (left), Nrd1 (middle) and Nab3 (right) binding sites compared to unbound sequence regions, analyzed within sequence bins sorted by occupancy from 100% to 1% occupancy (x axis).
(D) Scatter-plots of the occurences of the tetramer motifs UGUA (top row) and CUUG (bottom row) in sense (left) and antisense ORF-Ts (right). Position weight matrices (PWMs) obtained by XXmotif for the top bound 1000 Nrd1 (top) and Nab3 (bottom) sequences are shown in the middle.

∼300nt upstream of TSSs and at ∼240nt upstream of pAs which indicates a strong relationship between these proteins. The comparison of the Nrd1/Nab3/Sen1 occupancies to the peak of the capping factor Cbc2 shows that the occupancies for Nrd1/Sen1 are almost identical for transcripts from NDRs at TSSs and pAs (∼0.65 at TSS and ∼0.06 for pA) while the Cbc2 occupancy shows a difference by a factor of 2 (∼0.04 for transcripts of NDRs at TSSs and ∼0.02 at pAs). The difference of the occupancy of Nab3 compared to its interaction partners and the disimilar occupancy of Cbc2 shows a small difference

between ncRNA originated from NDRs at TSSs and at pAs.

Motif and k-mer analysis of the PAR-CLIP data reveal strong binding preferences to the 4-mers UGUA and GUAG for Nrd1 and CUUG and UCUU for Nab3 (Figure 3.20C) which is consistent with published binding preferences for Nrd1 and Nab3 [45, 59]. However, Nrd1 shows a stronger preference to primary sequence elements compared to Nab3 which shows also weak preferences to Nrd1 related 4-mers. In addition, the motif enrichment is positively correlated with the described RNA-seq normalization (Figure 3.20C). Thus, the stronger the PAR-CLIP occupancy the stronger the enrichment of the 4-mer. Despite the strong agreement of the Nrd1 and Sen1 binding patterns no convincing binding preference is observed for Sen1 (Figure 3.20C) indicating that the RNA recognition for the Nrd1/Nab3/Sen1 complex is mediated via the binding preferences of Nrd1 and Nab3.

A closer look to the distribution of the Nrd1-preferred 4-mer UGUA across the *S.cer* transcriptome (Figure 3.20D top row) shows a depletion in sense transcripts and a weak enrichment in antisense transcripts and thus indicating that ncRNAs are detected via sequence specificities. The 4-mer analysis of the Nab3-preferred 4-mer CUUG revealed even stronger observations (Figure 3.20D bottom row) especially for the occurrence of CUUG in antisense transcripts.

4 Conclusion & Outlook

A complete computational pipeline for PAR-CLIP data analysis referred to as `STAMMP` is presented in this work. `STAMMP` covers all necessary steps to get from raw data to interpretable plots and to shed light on the functions of *in vivo* protein-RNA interactions. The general analysis of PAR-CLIP data shows, that mutation counts which are widely used for both binding site detection and binding strength determination have to be corrected for their transcript expression rate dependency.

Additionally, PAR-CLIP data is impaired by severe technical biases likely to be caused by immunoprecipitation and offtarget effects, that falsifies biological results. The RNA-seq normalization shown here addresses both effects. Dividing the observed mutation counts by RNA-seq counts corrects for transcript abundance and lead to a natural determination of binding strength. After normalization technical immunoprecipitation background binding shows only strong influence on weakly bound binding sites. Strongly bound sites are not affected heavily. As a consequence, the normalization of PAR-CLIP data is absolutely mandatory to get reliable biological results.

PAR-CLIP data analyzed in this study is affected by a general sequence bias where binding sites tend to bind to 4-tU after previous poly-A stretches. If this effect is a general PAR-CLIP bias or if it is specific to the protocol used here may be analyzed in the future.

To sum up, data pre-processing has a huge impact on PAR-CLIP data quality and the following biological conclusion. However, proposed analysis protocols neglect these pre-process steps to a large extend and are only focusing onto the detection of binding sites based on mutation rates. Due to these facts, PAR-CLIP studies based on unnormalized data and their biological conclusions should be considered with caution.

For the purpose of the comparison of PAR-CLIP binding site detection methods a benchmark based on real biological data of 25 data sets is introduced here to compare PAR-CLIP analysis methods under real conditions. `STAMMP`'s statistical model for binding site detection shows a superior performance compared to the analysis methods proposed by `wavClusteR`, `PARalyzer` and $\texttt{PIPE}-\texttt{CLIP}$. More binding sites are found by `STAMMP` while lowering the FDR in less time.

However, the treatment of sequence biases and offtarget effects are not considered so far and have to be approached in upcoming improvements of `STAMMP`. In addition, the parameter estimations can be done separately for various sequencing depths as long as

enough count data is available rather then averaging parameters like it is done in the current version. Finally, it should be the aim to extend the p-value based detection method to a full bayesian model.

Besides technical results `STAMMP` revealed insights into the biogenesis of mRNAs as well as the transcriptome surveillance machinery in yeast. The protocol successfully captured interactions of proteins to pre-mRNAs as well as mature mRNAs allowing the determination of protein-RNA interactions in the matter of time. A conserved recognition of pre-mRNA introns, a unified recognition of pre-mRNA polyadenylation sites between yeast and human as well as a connection of splicing and $3'$-processing events are the most notable results.

As no specific binding motifs could be inferred for most of the analyzed proteins the *in vivo* folding of RNA molecules is one major remaining obstacle for both precise binding motif finding as well as the inference of factor co-occupancies and co-localizations. More sophisticated models are needed to precisely address these questions.

Bibliography

[1] Francis Crick et al. Central dogma of molecular biology. *Nature*, 227(5258):561–563, 1970.

[2] Gideon Dreyfuss, V Narry Kim, and Naoyuki Kataoka. Messenger-RNA-binding proteins and the messages they carry. *Nature Reviews Molecular Cell Biology*, 3(3):195–205, 2002.

[3] Odil Porrua and Domenico Libri. Transcription termination and the control of the transcriptome: why, where and how to stop. *Nature Reviews Molecular Cell Biology*, 2015.

[4] Julian König, Kathi Zarnack, Nicholas M Luscombe, and Jernej Ule. Protein–RNA interactions: new genomic technologies and perspectives. *Nature Reviews Genetics*, 13(2):77–83, 2012.

[5] Bonnie Berger, Jian Peng, and Mona Singh. Computational solutions for omics data. *Nature Reviews Genetics*, 14(5):333–346, 2013.

[6] James D Watson, Francis HC Crick, et al. Molecular structure of nucleic acids. *Nature*, 171(4356):737–738, 1953.

[7] Andreas Mayer, Michael Lidschreiber, Matthias Siebert, Kristin Leike, Johannes Söding, and Patrick Cramer. Uniform transitions of the general RNA polymerase II transcription complex. *Nature Structural & Molecular Biology*, 17(10):1272–1278, 2010.

[8] Eun-Jung Cho, Toshimitsu Takagi, Christine R Moore, and Stephen Buratowski. mRNA capping enzyme is recruited to the transcription complex by phosphorylation of the RNA polymerase II carboxy-terminal domain. *Genes & Development*, 11(24):3319–3326, 1997.

[9] Eric B Rasmussen and JOHN T Lis. In vivo transcriptional pausing and cap formation on three Drosophila heat shock genes. *Proceedings of the National Academy of Sciences*, 90(17):7923–7927, 1993.

[10] Susan M Berget, Claire Moore, and Phillip A Sharp. Spliced segments at the 5'terminus of adenovirus 2 late mRNA. *Proceedings of the National Academy of Sciences*, 74(8):3171–3175, 1977.

[11] John Rogers and Randolph Wall. A mechanism for RNA splicing. *Proceedings of the National Academy of Sciences*, 77(4):1877–1879, 1980.

[12] Brenton R Graveley. Alternative splicing: increasing diversity in the proteomic world. *Trends in Genetics*, 17(2):100–107, 2001.

[13] Benjamin J Blencowe. Alternative splicing: new insights from global analyses. *Cell*, 126(1):37–47, 2006.

[14] Robin Reed and Ed Hurt. A conserved mRNA export machinery coupled to pre-mRNA splicing. *Cell*, 108(4):523–531, 2002.

[15] Christoph Engel, Sarah Sainsbury, Alan C Cheung, Dirk Kostrewa, and Patrick Cramer. RNA polymerase I structure and transcription regulation. *Nature*, 502(7473):650–655, 2013.

[16] Patrick Cramer, David A Bushnell, and Roger D Kornberg. Structural basis of transcription: RNA polymerase II at 2.8 ångstrom resolution. *Science*, 292(5523):1863–1876, 2001.

[17] Rieke Ringel, Marina Sologub, Yaroslav I Morozov, Dmitry Litonin, Patrick Cramer, and Dmitry Temiakov. Structure of human mitochondrial RNA polymerase. *Nature*, 478(7368):269–273, 2011.

[18] Thomas R Cech and BL Bass. Biological catalysis by RNA. *Annual Review of Biochemistry*, 55(1):599–629, 1986.

[19] Ivo L Hofacker, Walter Fontana, Peter F Stadler, L Sebastian Bonhoeffer, Manfred Tacker, and Peter Schuster. Fast folding and comparison of RNA secondary structures. *Monatshefte für Chemie/Chemical Monthly*, 125(2):167–188, 1994.

[20] Sarah Djebali, Carrie A Davis, Angelika Merkel, Alex Dobin, Timo Lassmann, Ali Mortazavi, Andrea Tanzer, Julien Lagarde, Wei Lin, Felix Schlesinger, et al. Landscape of transcription in human cells. *Nature*, 489(7414):101–108, 2012.

[21] Ryan J Taft, Michael Pheasant, and John S Mattick. The relationship between non-protein-coding DNA and eukaryotic complexity. *Bioessays*, 29(3):288–299, 2007.

[22] John L Rinn and Howard Y Chang. Genome regulation by long noncoding RNAs. *Annual Review of Biochemistry*, 81, 2012.

[23] Alessandro Fatica and Irene Bozzoni. Long non-coding RNAs: new players in cell differentiation and development. *Nature Reviews Genetics*, 15(1):7–21, 2014.

[24] Daniel Schulz, Bjoern Schwalb, Anja Kiesel, Carlo Baejen, Phillipp Torkler, Julien Gagneur, Johannes Soeding, and Patrick Cramer. Transcriptome surveillance by selective termination of noncoding RNA synthesis. *Cell*, 155(5):1075–1087, 2013.

[25] Hannah E Mischo and Nick J Proudfoot. Disengaging polymerase: terminating RNA polymerase II transcription in budding yeast. *Biochimica et Biophysica Acta (BBA)-Gene Regulatory Mechanisms*, 1829(1):174–185, 2013.

[26] Eric J Steinmetz, Nicholas K Conrad, David A Brow, and Jeffry L Corden. RNA-binding protein Nrd1 directs poly (A)-independent 3'-end formation of RNA polymerase II transcripts. *Nature*, 413(6853):327–331, 2001.

[27] Daniel J Hogan, Daniel P Riordan, André P Gerber, Daniel Herschlag, and Patrick O Brown. Diverse RNA-binding proteins interact with functionally related sets of RNAs, suggesting an extensive regulatory system. *PLoS Biology*, 6(10):e255, 2008.

[28] Bradley M Lunde, Claire Moore, and Gabriele Varani. RNA-binding proteins: modular design for efficient function. *Nature Reviews Molecular Cell Biology*, 8(6):479–490, 2007.

[29] Ewan Birney, Sanjay Kumar, and Adrian R Krainer. Analysis of the RNA-recognition motif and RS and RGG domains: conservation in metazoan pre-mRNA splicing factors. *Nucleic Acids Research*, 21(25):5803–5816, 1993.

[30] Sigrid D Auweter, Florian C Oberstrass, and Frédéric H-T Allain. Sequence-specific binding of single-stranded RNA: is there a code for recognition? *Nucleic Acids Research*, 34(17):4943–4959, 2006.

[31] Scott A Tenenbaum, Craig C Carson, Patrick J Lager, and Jack D Keene. Identifying mRNA subsets in messenger ribonucleoprotein complexes by using cDNA arrays. *Proceedings of the National Academy of Sciences*, 97(26):14085–14090, 2000.

[32] Jernej Ule, Kirk B Jensen, Matteo Ruggiu, Aldo Mele, Aljaž Ule, and Robert B Darnell. CLIP identifies Nova-regulated RNA networks in the brain. *Science*, 302(5648):1212–1215, 2003.

[33] 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, 2005.

[34] Donny D Licatalosi, Aldo Mele, John J Fak, Jernej Ule, Melis Kayikci, Sung Wook Chi, Tyson A Clark, Anthony C Schweitzer, John E Blume, Xuning Wang, et al. HITS-CLIP yields genome-wide insights into brain alternative RNA processing. *Nature*, 456(7221):464–469, 2008.

[35] 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.

[36] Markus Hafner, Markus Landthaler, Lukas Burger, Mohsen Khorshid, Jean Hausser, Philipp Berninger, Andrea Rothballer, Manuel Ascano Jr, Anna-Carina Jungkamp, Mathias Munschauer, et al. Transcriptome-wide identification of RNA-binding protein and microRNA target sites by PAR-CLIP. *Cell*, 141(1):129–141, 2010.

[37] David L Corcoran, Stoyan Georgiev, Neelanjan Mukherjee, Eva Gottwein, Rebecca L Skalsky, Jack D Keene, Uwe Ohler, et al. PARalyzer: definition of RNA binding sites from PAR-CLIP short-read sequence data. *Genome Biol*, 12(8):R79, 2011.

[38] Cem Sievers, Tommy Schlumpf, Ritwick Sawarkar, Federico Comoglio, and Renato Paro. Mixture models and wavelet transforms reveal high confidence RNA-protein interaction sites in MOV10 PAR-CLIPses data. *Nucleic Acids Research*, page gks697, 2012.

[39] Beibei Chen, Jonghyun Yun, Min Soo Kim, Joshua T Mendell, and Yang Xie. PIPE-CLIP: a comprehensive online tool for CLIP-seq data analysis. *Genome Biol*, 15:R18, 2014.

[40] Neelanjan Mukherjee, David L Corcoran, Jeffrey D Nusbaum, David W Reid, Stoyan Georgiev, Markus Hafner, Manuel Ascano Jr, Thomas Tuschl, Uwe Ohler, and Jack D Keene. Integrative regulatory mapping indicates that the RNA-binding protein HuR couples pre-mRNA processing and mRNA stability. *Molecular Cell*, 43(3):327–339, 2011.

[41] Markus Hafner, Klaas EA Max, Pradeep Bandaru, Pavel Morozov, Stefanie Gerstberger, Miguel Brown, Henrik Molina, and Thomas Tuschl. Identification of mRNAs bound and regulated by human LIN28 proteins and molecular requirements for RNA recognition. *RNA*, 19(5):613–626, 2013.

[42] Robert C Gentleman, Vincent J. Carey, Douglas M. Bates, and others. Bioconductor: Open software development for computational biology and bioinformatics. *Genome Biology*, 5:R80, 2004.

[43] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 289–300, 1995.

[44] Carlo Baejen, Phillipp Torkler, Saskia Gressel, Katharina Essig, Johannes Söding, and Patrick Cramer. Transcriptome maps of mRNP biogenesis factors define pre-mRNA recognition. *Molecular Cell*, 55(5):745–757, 2014.

[45] Tyler J Creamer, Miranda M Darby, Nuttara Jamonnak, Paul Schaughency, Haiping Hao, Sarah J Wheelan, and Jeffry L Corden. Transcriptome-wide binding sites for components of the Saccharomyces cerevisiae non-poly (A) termination pathway: Nrd1, Nab3, and Sen1. *PLoS Genetics*, 7(10):e1002329, 2011.

[46] Eric Jones, Travis Oliphant, Pearu Peterson, et al. SciPy: Open source scientific tools for Python, 2001–. [Online; accessed 2015-03-02].

[47] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2014.

[48] Ben Langmead, Cole Trapnell, Mihai Pop, Steven L Salzberg, et al. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol*, 10(3):R25, 2009.

[49] Heng Li, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, Richard Durbin, et al. The sequence alignment/map format and SAMtools. *Bioinformatics*, 25(16):2078–2079, 2009.

[50] Christopher M Bishop et al. *Pattern recognition and machine learning*, volume 1. springer New York, 2006.

[51] Charles George Broyden. The convergence of a class of double-rank minimization algorithms 1. general considerations. *IMA Journal of Applied Mathematics*, 6(1):76–90, 1970.

[52] Roger Fletcher. A new approach to variable metric algorithms. *The Computer Journal*, 13(3):317–322, 1970.

[53] Donald Goldfarb. A family of variable-metric methods derived by variational means. *Mathematics of Computation*, 24(109):23–26, 1970.

[54] David F Shanno. Conditioning of quasi-Newton methods for function minimization. *Mathematics of Computation*, 24(111):647–656, 1970.

[55] Shivendra Kishore, Lukasz Jaskiewicz, Lukas Burger, Jean Hausser, Mohsen Khorshid, and Mihaela Zavolan. A quantitative analysis of CLIP methods for identifying binding sites of RNA-binding proteins. *Nature Methods*, 8(7):559–564, 2011.

[56] Zhenyu Xu, Wu Wei, Julien Gagneur, Fabiana Perocchi, Sandra Clauder-Münster, Jurgi Camblong, Elisa Guffanti, Françoise Stutz, Wolfgang Huber, and Lars M Steinmetz. Bidirectional promoters generate pervasive transcription in yeast. *Nature*, 457(7232):1033–1037, 2009.

[57] Vicent Pelechano, Wu Wei, and Lars M Steinmetz. Extensive transcriptional heterogeneity revealed by isoform profiling. *Nature*, 497(7447):127–131, 2013.

[58] T Fukunaka, Haruka Ozaki, Goro Terai, Kiyoshi Asai, Wataru Iwasaki, and Hisanori Kiryu. CapR: revealing structural specificities of RNA-binding protein target recognition using CLIP-seq data. *Genome Biol*, 15(1):R16, 2014.

[59] Daniel P Riordan, Daniel Herschlag, and Patrick O Brown. Identification of RNA recognition elements in the Saccharomyces cerevisiae transcriptome. *Nucleic Acids Research*, 39(4):1501–1509, 2011.

[60] Matthias Siebert, Michael Lidschreiber, Holger Hartmann, and Johannes Soding. A guideline for ChIP-chip data quality control and normalization (prot 47). *Tech. rep., Gene Center Munich, Ludwig-Maximilians-Universität*, 2009.

[61] Leonid Teytelman, Deborah M Thurtle, Jasper Rine, and Alexander van Oudenaarden. Highly expressed loci are vulnerable to misleading ChIP localization of multiple unrelated proteins. *Proceedings of the National Academy of Sciences*, 110(46):18602–18607, 2013.

[62] Matthew B Friedersdorf and Jack D Keene. Advancing the functional utility of PAR-CLIP by quantifying background binding to mRNAs and lncRNAs. *Genome Biol*, 15(1):R2, 2014.

[63] Stephen Buratowski. Progression through the RNA polymerase II CTD cycle. *Molecular Cell*, 36(4):541–546, 2009.

[64] Martin Heidemann and Dirk Eick. Tyrosine-1 and threonine-4 phosphorylation marks complete the RNA polymerase II CTD phospho-code. *RNA Biology*, 9(9):1144–1146, 2012.

[65] Jing-Ping Hsin and James L Manley. The RNA polymerase II CTD coordinates transcription and RNA processing. *Genes & Development*, 26(19):2119–2137, 2012.

[66] Roberto Perales and David Bentley. "cotranscriptionality": the transcription elongation complex as a nexus for nuclear transactions. *Molecular Cell*, 36(2):178–191, 2009.

[67] Serena Chan, Eun-A Choi, and Yongsheng Shi. Pre-mRNA 3'-end processing complex assembly and function. *Wiley Interdisciplinary Reviews: RNA*, 2(3):321–335, 2011.

[68] Michaela Müller-McNicoll and Karla M Neugebauer. How cells get the message: dynamic assembly and function of mRNA-protein complexes. *Nature Reviews Genetics*, 14(4):275–287, 2013.

[69] Corey R Mandel, Yun Bai, and Liang Tong. Protein factors in pre-mRNA 3'-end processing. *Cellular and Molecular Life Sciences*, 65(7-8):1099–1122, 2008.

[70] Nick J Proudfoot. Ending the message: poly (A) signals then and now. *Genes & Development*, 25(17):1770–1782, 2011.

[71] Markus C Wahl, Cindy L Will, and Reinhard Lührmann. The spliceosome: design principles of a dynamic RNP machine. *Cell*, 136(4):701–718, 2009.

[72] Miha Milek, Emanuel Wyler, and Markus Landthaler. Transcriptome-wide analysis of protein–RNA interactions using high-throughput sequencing. In *Seminars in Cell & Developmental Biology*, volume 23, pages 206–212. Elsevier, 2012.

[73] Georges Martin, Andreas R Gruber, Walter Keller, and Mihaela Zavolan. Genome-wide analysis of pre-mRNA 3' end processing reveals a decisive role of human cleavage factor I in the regulation of 3' UTR length. *Cell Reports*, 1(6):753–763, 2012.

[74] Alex Charles Tuck and David Tollervey. A transcriptome-wide atlas of RNP composition reveals diverse classes of mRNAs and lncRNAs. *Cell*, 154(5):996–1009, 2013.

[75] Alex Andrus and Robert G Kuimelis. Base composition analysis of nucleosides using HPLC. *Current Protocols in Nucleic Acid Chemistry*, pages 10–6, 2001.

[76] Holger Hartmann, Eckhart W Guthöhrlein, Matthias Siebert, Sebastian Luehr, and Johannes Söding. P-value-based regulatory motif discovery using positional weight matrices. *Genome Research*, 23(1):181–194, 2013.

[77] David A Mangus, Matthew C Evans, and Allan Jacobson. Poly (A)-binding proteins: multifunctional scaffolds for the post-transcriptional control of gene expression. *Genome Biol*, 4(7):223, 2003.

[78] Ho Sung Rhee and B Franklin Pugh. Genome-wide structure and organization of eukaryotic pre-initiation complexes. *Nature*, 483(7389):295–301, 2012.

[79] Torben Heick Jensen, Alain Jacquier, and Domenico Libri. Dealing with pervasive transcription. *Molecular Cell*, 52(4):473–484, 2013.

[80] Maxime Wery, Marta Kwapisz, and Antonin Morillon. Noncoding RNAs in gene regulation. *Wiley Interdisciplinary Reviews: Systems Biology and Medicine*, 3(6):728–738, 2011.

[81] Cindy L Will and Reinhard Lührmann. Spliceosome structure and function. *Cold Spring Harbor Perspectives in Biology*, 3(7):a003707, 2011.

[82] Cameron D Mackereth, Tobias Madl, Sophie Bonnal, Bernd Simon, Katia Zanier, Alexander Gasch, Vladimir Rybin, Juan Valcárcel, and Michael Sattler. Multi-domain conformational selection underlies pre-mRNA splicing regulation by U2AF. *Nature*, 475(7356):408–411, 2011.

[83] Kathi Zarnack, Julian König, Mojca Tajnik, Inigo Martincorena, Sebastian Eustermann, Isabelle Stévant, Alejandro Reyes, Simon Anders, Nicholas M Luscombe, and Jernej Ule. Direct competition between hnRNP C and U2AF65 protects the transcriptome from the exonization of Alu elements. *Cell*, 152(3):453–466, 2013.

[84] J Andrew Berglund, Katrin Chua, Nadja Abovich, Robin Reed, and Michael Rosbash. The splicing factor BBP interacts specifically with the pre-mRNA branchpoint sequence UACUAAC. *Cell*, 89(5):781–787, 1997.

[85] Nadja Abovich and Michael Rosbash. Cross-intron bridging interactions in the yeast commitment complex are conserved in mammals. *Cell*, 89(3):403–412, 1997.

[86] Claudia Schneider, Grzegorz Kudla, Wiebke Wlotzka, Alex Tuck, and David Tollervey. Transcriptome-wide analysis of exosome targets. *Molecular Cell*, 48(3):422–433, 2012.

[87] L. T. Vo, M. Minet, J. M. Schmitter, F. Lacroute, and F. Wyers. Mpe1, a zinc knuckle protein, is an essential component of yeast cleavage and polyadenylation factor required for the cleavage and polyadenylation of mRNA. *Molecular and Cellular Biology*, 21(24):8346–8356, Dec 2001.

[88] Bernhard Dichtl and Walter Keller. Recognition of polyadenylation sites in yeast pre-mRNAs by cleavage and polyadenylation factor. *The EMBO Journal*, 20(12):3197–3209, 2001.

[89] Silvia ML Barabino, Martin Ohnacker, and Walter Keller. Distinct roles of two Yth1p domains in 3'-end cleavage and polyadenylation of yeast pre-mRNAs. *The EMBO Journal*, 19(14):3778–3787, 2000.

[90] Stefan Gross and Claire L Moore. Rna15 interaction with the A-rich yeast polyadenylation signal is an essential step in mRNA 3'-end formation. *Molecular and Cellular Biology*, 21(23):8045–8055, 2001.

[91] Andreas Mayer, Amelie Schreieck, Michael Lidschreiber, Kristin Leike, Dietmar E Martin, and Patrick Cramer. The spt5 C-terminal region recruits yeast 3' RNA cleavage factor I. *Molecular and Cellular Biology*, 32(7):1321–1331, 2012.

[92] Zijian Guo, Patrick Russo, Ding-Fang Yun, JS Butler, and Fred Sherman. Redundant 3'end-forming signals for the yeast CYC1 mRNA. *Proceedings of the National Academy of Sciences*, 92(10):4211–4214, 1995.

[93] Joel H Graber, Gregory D McAllister, and Temple F Smith. Probabilistic prediction of Saccharomyces cerevisiae mRNA 3'-processing sites. *Nucleic Acids Research*, 30(8):1851–1858, 2002.

[94] Zijian Guo and Fred Sherman. 3'-end-forming signals of yeast mRNA. *Trends in Biochemical Sciences*, 21(12):477–481, 1996.

[95] Marco M Kessler, Michael F Henry, Elisa Shen, Jing Zhao, Stefan Gross, Pamela A Silver, and Claire L Moore. Hrp1, a sequence-specific RNA-binding protein that shuttles between the nucleus and the cytoplasm, is required for mRNA 3'-end formation in yeast. *Genes & Development*, 11(19):2545–2556, 1997.

[96] Radharani Duttagupta, Bin Tian, Carol J Wilusz, Danny T Khounh, Patricia Soteropoulos, Ming Ouyang, Joseph P Dougherty, and Stuart W Peltz. Global analysis of Pub1p targets reveals a coordinate control of gene expression through modulation of binding and stability. *Molecular and Cellular Biology*, 25(13):5499–5513, 2005.

[97] Shobha Vasudevan and Stuart W Peltz. Regulated ARE-mediated mRNA decay in Saccharomyces cerevisiae. *Molecular Cell*, 7(6):1191–1200, 2001.

[98] Ying Cui, Kevin W Hagan, Shuang Zhang, and Stuart W Peltz. Identification and characterization of genes that are required for the accelerated degradation of mRNAs containing a premature translational termination codon. *Genes & Development*, 9(4):423–436, 1995.

[99] Maria J Ruiz-Echevarría, Carlos I González, and Stuart W Peltz. Identifying the right stop: determining how the surveillance complex recognizes and degrades an aberrant mRNA. *The EMBO Journal*, 17(2):575–589, 1998.

[100] Maria J Ruiz-Echevarría and Stuart W Peltz. The RNA binding protein Pub1 modulates the stability of transcripts containing upstream open reading frames. *Cell*, 101(7):741–751, 2000.

[101] Rosa Luna, Ana G Rondón, and Andrés Aguilera. New clues to understand the role of THO and other functionally related factors in mRNP biogenesis. *Biochimica et Biophysica Acta (BBA)-Gene Regulatory Mechanisms*, 1819(6):514–520, 2012.

[102] Katja Sträßer, Seiji Masuda, Paul Mason, Jens Pfannstiel, Marisa Oppizzi, Susana Rodriguez-Navarro, Ana G Rondón, Andres Aguilera, Kevin Struhl, Robin Reed, et al. TREX is a conserved complex coupling transcription with messenger RNA export. *Nature*, 417(6886):304–308, 2002.

[103] Patric Grüter, Carlos Tabernero, Cayetano von Kobbe, Christel Schmitt, Claudio Saavedra, Angela Bachi, Matthias Wilm, Barbara K Felber, and Elisa Izaurralde. TAP, the human homolog of Mex67p, mediates CTE-dependent RNA export from the nucleus. *Molecular Cell*, 1(5):649–659, 1998.

[104] Alexandra Segref, Kishore Sharma, Valérie Doye, Andrea Hellwig, Jochen Huber, Reinhard Lührmann, and Ed Hurt. Mex67p, a novel factor for nuclear mRNA export, binds to both poly (A)+ RNA and nuclear pores. *The EMBO Journal*, 16(11):3256–3271, 1997.

[105] Alexandra Hackmann, Thomas Gross, Claudia Baierlein, and Heike Krebber. The mRNA export factor Npl3 mediates the nuclear export of large ribosomal subunits. *EMBO Reports*, 12(10):1024–1031, 2011.

[106] Nahid Iglesias, Evelina Tutucci, Carole Gwizdek, Patrizia Vinciguerra, Elodie Von Dach, Anita H Corbett, Catherine Dargemont, and Françoise Stutz. Ubiquitin-mediated mRNP dynamics and surveillance prior to budding yeast mRNA export. *Genes & Development*, 24(17):1927–1938, 2010.

[107] Susana Rodríguez-Navarro and Ed Hurt. Linking gene regulation to mRNA production and export. *Current Opinion in Cell Biology*, 23(3):302–309, 2011.

[108] Murray Stewart. Nuclear export of mRNA. *Trends in Biochemical Sciences*, 35(11):609–617, 2010.

[109] Sebastián Chávez, Traude Beilharz, Ana G Rondón, Hediye Erdjument-Bromage, Paul Tempst, Jesper Q Svejstrup, Trevor Lithgow, and Andrés Aguilera. A protein complex containing Tho2, Hpr1, Mft1 and a novel protein, Thp2, connects transcription elongation with mitotic recombination in Saccharomyces cerevisiae. *The EMBO journal*, 19(21):5824–5834, 2000.

[110] Kamil Gewartowski, Jorge Cuéllar, Andrzej Dziembowski, and José María Valpuesta. The yeast THO complex forms a 5-subunit assembly that directly interacts with active chromatin. *Bioarchitecture*, 2(4):134–137, 2012.

[111] Haley Hieronymus and Pamela A Silver. Genome-wide analysis of RNA–protein interactions illustrates specificity of the mRNA export machinery. *Nature Genetics*, 33(2):155–161, 2003.

[112] Miriam E Bucheli, Xiaoyuan He, Craig D Kaplan, Claire L Moore, and Stephen Buratowski. Polyadenylation site choice in yeast is affected by competition between Npl3 and polyadenylation factor CFI. *RNA*, 13(10):1756–1764, 2007.

[113] Pritilekha Deka, Miriam E Bucheli, Claire Moore, Stephen Buratowski, and Gabriele Varani. Structure of the yeast SR protein Npl3 and interaction with mRNA 3'-end processing signals. *Journal of Molecular Biology*, 375(1):136–150, 2008.

[114] JAMES T Anderson, SCOTT M Wilson, KSHAMA V Datar, and MAURICE S Swanson. NAB2: a yeast nuclear polyadenylated RNA-binding protein essential for cell viability. *Molecular and Cellular Biology*, 13(5):2730–2741, 1993.

[115] Deanna M Green, Kavita A Marfatia, Emily B Crafton, Xing Zhang, Xiaodong Cheng, and Anita H Corbett. Nab2p is required for poly (A) RNA export in Saccharomyces cerevisiae and is regulated by arginine methylation via Hmt1p. *Journal of Biological Chemistry*, 277(10):7752–7760, 2002.

[116] Ronald E Hector, Keith R Nykamp, Sonia Dheur, James T Anderson, Priscilla J Non, Carl R Urbinati, Scott M Wilson, Lionel Minvielle-Sebastia, and Maurice S Swanson. Dual requirement for yeast hnRNP Nab2p in mRNA poly (A) tail length control and nuclear export. *The EMBO Journal*, 21(7):1800–1810, 2002.

[117] Thomas C Leeper, Xiangping Qu, Connie Lu, Claire Moore, and Gabriele Varani. Novel protein–protein contacts facilitate mRNA 3'-processing signal recognition by Rna15 and Hrp1. *Journal of Molecular Biology*, 401(3):334–349, 2010.

[118] Guillaume Chanfreau, Suzanne M Noble, and Christine Guthrie. Essential yeast protein with unexpected similarity to subunits of mammalian cleavage and polyadenylation specificity factor (CPSF). *Science*, 274(5292):1511–1514, 1996.

[119] Harold G Martinson. An active role for splicing in 3'-end formation. *Wiley Interdisciplinary Reviews: RNA*, 2(4):459–470, 2011.

[120] Yongsheng Shi, Di Giammartino, Dafne Campigli, Derek Taylor, Ali Sarkeshik, William J Rice, John R Yates III, Joachim Frank, and James L Manley. Molecular architecture of the human pre-mRNA 3' processing complex. *Molecular Cell*, 33(3):365–376, 2009.

[121] Andrea Kyburz, Arno Friedlein, Hanno Langen, and Walter Keller. Direct interactions between subunits of CPSF and the U2 snRNP contribute to the coupling of pre-mRNA 3' end processing and splicing. *Molecular Cell*, 23(2):195–205, 2006.

[122] CHRISTIANE BRUNE, SARAH E MUNCHEL, NICOLE FISCHER, ALEXANDRE V PODTELEJNIKOV, and KARSTEN WEIS. Yeast poly (A)-binding protein Pab1 shuttles between the nucleus and the cytoplasm and functions in mRNA export. *RNA*, 11(4):517–531, 2005.

[123] Merle Windgassen and Heike Krebber. Identification of Gbp2 as a novel poly (A)+ RNA-binding protein involved in the cytoplasmic delivery of messenger RNAs in yeast. *EMBO Reports*, 4(3):278–283, 2003.

[124] Alexandra Hackmann, Haijia Wu, Ulla-Maria Schneider, Katja Meyer, Klaus Jung, and Heike Krebber. Quality control of spliced mRNAs requires the shuttling SR proteins Gbp2 and Hrb1. *Nature Communications*, 5, 2014.

[125] Tracy L Kress, Nevan J Krogan, and Christine Guthrie. A single SR-like protein, Npl3, promotes pre-mRNA splicing in budding yeast. *Molecular Cell*, 32(5):727–734, 2008.

[126] Haihong Shen and Michael R Green. RS domains contact splicing signalsand promote splicing by a commonmechanism in yeast through humans. *Genes & Development*, 20(13):1755–1765, 2006.

[127] Hirohito Haruki, Junichi Nishikawa, and Ulrich K Laemmli. The anchor-away technique: rapid, conditional establishment of yeast mutant phenotypes. *Molecular Cell*, 31(6):925–932, 2008.

[128] Mai Sun, Björn Schwalb, Daniel Schulz, Nicole Pirkl, Stefanie Etzold, Laurent Larivière, Kerstin C Maier, Martin Seizl, Achim Tresch, and Patrick Cramer. Comparative dynamic transcriptome analysis (cDTA) reveals mutual feedback between mRNA synthesis and degradation. *Genome Research*, 22(7):1350–1359, 2012.