Computational studies of RNA modification-dependent RNA-protein networks

Charlotte Capitanchik

A dissertation submitted in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

of

University College London.

Student Number: 17130416

March, 2022

I, Charlotte Capitanchik, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the work.

Abstract

The covalent modification of RNA nucleotides is a powerful layer of post-transcriptional control of gene expression across the tree of life. Historically, only abundant modifications on abundant RNAs such as tRNA and rRNA could be studied, due to methodological limitations. In the past decade, leaps forward in biochemistry and high throughput sequencing methods have enabled mapping of RNA modifications across all RNA species. In particular this thesis focuses on the most abundant internal modification of mRNA, N6-methyladenosine (m6A), and how RNA binding proteins (RBPs) interact with RNA modifications to impact RNA life cycle. Alongside these experimental developments have come new computational challenges. Integration of many datasets must be approached carefully, with a view to extract as much biological information as possible. Throughout this work I describe the development of open source computational tools for the analysis and visualisation of CLIP data. A computational pipeline based on hierarchical pre-mapping steps enables accurate quantification of non-coding RNAs from individual nucleotide resolution crosslinking and immunoprecipitation (iCLIP) datasets. Using the pipeline I describe novel tRNA binding for the DEAH-box helicase DDX3X and identify widespread binding of NSun2 and Trmt2A to pre-tRNAs. In collaboration with the lab of Prof. Folkert van Werven, I integrate m6A miCLIP with m6A-reader protein iCLIP data, alongside functional datasets in WT and methyltransferase deletion conditions in order to uncover the role of m6A in early budding yeast meiosis. Surprisingly, we find that the sole yeast m6A-binding protein, Pho92p, binds in both an m6A-dependent and an m6A-independent manner. m6A-dependent Pho92p binding partners are implicated in mRNA decay coupled to translation. Taken together I present powerful computational tools that will be of use to the wider community, alongside the interesting biological insights they have already enabled.

Impact Statement

Broadly, the study of RNA binding proteins (RBPs) and their RNA partners, alongside the understanding of how RNA nucleotide modifications contribute to these binding networks, is a rich field, producing insights which will certainly have applications to many medical fields in the long-term. In the course of my doctoral work I have developed and contributed to several pieces of software that will help researchers more effectively analyse data produced to identify RNAs bound by RNA binding proteins, and RNAs containing modified nucleotides, and to present the results of such analyses.

Understanding how m6A reader proteins function is of particular importance to the field of cancer research. m6A is dysregulated in a staggering number of cancers (X.-Y. Chen et al., 2019). A small molecule inhibitor of the methyltransferase has been pursued as a treatment for Acute myeloid leukaemia (AML) (Yankova et al., 2021). Furthermore, YTH reader proteins are themselves also being explored as targets to modify the cross-presentation of tumour antigens and activation of CD8+ T cells and to selectively kill AML cells (Han et al., 2019; Paris et al., 2019). Work in model systems, such as budding yeast, allows us to come to a more detailed understanding of complex molecular mechanisms, which is crucial for interpreting the effects of interventions targeting complex human pathways.

From an academic perspective, I describe my contributions to four manuscripts. One is a review describing the advances in technologies to map m6A transcriptome-wide, which I weave into my thesis introduction (Capitanchik et al., 2020). Another describes my work in collaboration with the group of Prof. Folkert van Werven to uncover the mechanisms of m6A-dependent RBP networks in budding yeast meiosis (Varier et al., 2022). In my final results chapters I touch upon my work with Dr. Lisa Strittmatter to develop the analysis methods for a variant CLIP protocol focused on identifying helicase binding sites in spliceosomes of defined conformation (Strittmatter et al., 2020). Further, I briefly present how this led to a collaboration with Dr. Anob Chakrabarti to produce a computational tool that enables researchers to effectively visualise and normalise many CLIP samples in one plot, with additional annotations (Chakrabarti et al., 2021). Manuscripts describing the tRNA binding of DDX3X, my ncRNA-aware CLIP pipeline and the interactive CLIP peak caller Clippy are all in preparation.

The further impact of the computational tools described in this work will be felt when they are integrated into the web analysis platform iMaps (imaps.goodwright.com), where they will be made available as part of a wider suite of tools. The user-friendly platform will allow integrative analysis across many datasets, with the aim of increasing the pace of biological insight in the face of potentially overwhelming amounts of data.

Acknowledgements

Most PhDs are turbulent, but I'd say anyone graduating around now has had a little bit more turbulence than usual. I am beyond grateful for the mentorship of my two supervisors, Prof. Jernej Ule and Prof. Nick Luscombe, which has enabled me to succeed despite the odds. Both have supported me to grow as a scientist and as a person. I feel very lucky to have spent the past four years working between the two groups in such a kind and intellectually stimulating atmosphere, with so many incredible scientists. The only problem is that makes it a little hard to move on.

When I started my PhD, Nobby Chakrabarti, Fede Agostini and Igor Ruiz de Los Mozos took me under their wing and I learnt so much about how to do bioinformatics from them. Thanks also to Sebastian Steinhauser and Aylin Cakiroglu for the drinks, the misery and the skateboarding. Also thanks to Aylin for involving me in the Crick Data Challenge, which was a highlight of my PhD and continues to open up new opportunities for me.

I would be remiss not to thank Miha Modic, Flora Lee and Rupert Faraway (and all the Ule lab members I bothered) for the precious time spent teaching me how to work in the wet lab (and all the drinks).

My PhD is built on the foundation of lots of great collaborations, thank you especially to Lisa Strittmatter, Patrick Toolan-Kerr, Oscar Wilkins, Lorea Blazquez, Radhika Varier, Dora Sideri, Aleksej Drino, Nobby all the time and most recently a massive thank you to Marc Jones who has taught me loads about programming and is also a very nice person. Thank you also to Giulia Manferrari for sharing your work with me and being lots of fun.

Time spent with friends has kept me vaguely sane, thank you to the ride or

dies: Molly, Ruairi, Claire, Andrew, Emma, Nick, Larry and Sam. Thank you to the incredible women on my PhD programme: Simran, Lisa, Windie, who make me feel optimistic about the future. Thank you Holly for being the opposite of a fair-weather friend - a storm buddy?

Shout out to my Mum for all the support and raising me and all that! You taught me to always give my best. Also to my three brothers: Louis, Manny and Zane - you inspire me to be better. I am also very grateful to my grandparents, especially my Grandma, whose unwavering belief in me has pushed me forward.

Finally thank you to Alex, my partner in the truest sense of the word.

Dedication

I would like to dedicate this thesis to the memory of my Grandfather, Professor David Capitanchik; my Father, Danny Capitanchik and to my friend and fellow scientist, Diva Elizabeth Gibson.

I am heartbroken that I can never again celebrate in your company, but I am grateful in the knowledge that any achievement I make is only possible because I was loved by you.

Preface

This thesis describes my PhD work carried out at The Francis Crick Institute, London, UK, between September 2017 and March 2022 under the supervision of Prof. Nicholas Luscombe and Prof. Jernej Ule. The work has been partially published in multiple manuscripts and has often been the result of fruitful collaborations. Below I detail these manuscripts and the work carried out by others in contribution to each results chapter.

Introduction The survey of existing m6A methodologies was published in *Frontiers in Genetics*:
Capitanchik, C., Toolan-Kerr, P., Luscombe, N. M., & Ule, J. (2020). How do you identify m6A methylation in transcriptomes at high resolution? a comparison of recent datasets. *Frontiers in genetics*, 11, 398.

A computational pipeline for ncRNA-aware miCLIP and iCLIP analysis All lab work regarding DDX3X was performed by Dr. Aleksej Drino.

N6-methyladenosine-dependent RBP networks in Yeast Meiosis All of the lab work in this chapter - iCLIP, miCLIP, RNA-Seq, m6A translation assays were performed by Dr. Radhika Varier and Dr. Dora Sideri. RNA-Seq was pre-processed by Dr. Harshil Patel. The work is available as a pre-print:
Varier, R. A.*, Sideri, T.*, Capitanchik, C.*, Manova, Z., Calvani, E., Rossi, A., ... & van Werven, F. (2022). m6A reader Pho92 is recruited co-transcriptionally and couples translation efficacy to mRNA decay to promote meiotic fitness in yeast. *bioRxiv*.

CLIP Data Visualisation All of the lab work involved in the psiCLIP project was performed by Dr. Lisa Strittmatter and all data analysis was performed by me.

The project is published in *Nature Communications*:

Strittmatter, L. M.*, Capitanchik, C.*, Newman, A. J., Hallegger, M., Norman, C. M., Fica, S. M., ... & Nagai, K. (2021). psiCLIP reveals dynamic RNA binding by DEAH-box helicases before and after exon ligation. *Nature communications*, 12(1), 1-15.

CLIPplotR was developed with Dr. Anob Chakrabarti and is available as a pre-print:

Chakrabarti, A. M., Capitanchik, C., Ule, J., & Luscombe, N. M. (2021). clipplotr-a comparative visualisation and analysis tool for CLIP data. *bioRxiv*.

The Clippy project was initiated by me, further development of the source code and interactive visualisation was performed by Dr. Marc Jones.

* co-first authorship

Contents

List of Figures

List of Tables

List of Abbreviations

CDS	Coding DNA Sequence
Co-IP	Co-Immunoprecipitation
DNA	Deoxyribose Nucleic Acid
eCLIP	Enhanced Cross-Linking and Immunoprecipitation
GUI	Graphical User Interface
iCLIP	Indivudal nucleotide resolution Cross-Linking and Immunoprecipitation
IDR	Intrinsically Disordered Region
m6A	N6-methyladenosine
miCLIP	Methylation Indivudal nucleotide resolution Cross-Linking and Immunoprecipitation
mRNA	Messenger RNA
ncRNA	Non-coding RNA
PAR-CLIP	Photoactivatable Ribonucleoside-enhanced Cross-Linking and Immunoprecipitation
PCR	Polymerase Chain Reaction
PCA	Principal Component Analysis

RBP	RNA-binding protein
RNA	Ribose Nucleic Acid
rRNA	Ribosomal RNA
RT	Reverse Transcription/Transcriptase
SAM	S-Adenosyl methionine
snRNA	Small Nuclear RNA
SVM	Support Vector Machine
TLC	Thin Layer Chromatography
tRNA	Transfer RNA
UTR	Untranslated Region
YTH	YT521-B homology

Chapter 1

Introduction

> You are not obligated to complete the work, but neither are you free to desist from it.
>
> *Rabbi Tarfon, Pirke Avot 2:21*

The life of an RNA molecule begins at transcription and from that very moment it's existence is negotiated through complex and dynamic networks at the co- and post-transcriptional level. RNA processing mechanisms act to change the information content of mRNA, for example by alternative splicing or RNA editing, or to modulate the volume of this information by controlling rates of RNA synthesis, processing, degradation and translation. The life of an RNA is hectic as it passes through the hands of many RNA-binding proteins (RBPs) all acting together to achieve the miraculous: delivering said RNA to the right place at the right time.

A growing body of evidence suggests that an important player in fine-tuning gene expression is the language of covalent modifications of RNA nucleotides (Roignant & Soller, 2017). Both the RNA nucleotide base and sugar-phosphate backbone can harbour modifications. Such modifications are evolutionarily ancient, with many being present from the very beginnings of life in the RNA world. In fact, many tRNA methyltransferases can be traced back to the last universal common ancestor (LUCA), signifying their importance for life on this planet (Becker et al., 2018; Rana & Ankri, 2016; Weiss et al., 2018). It's compelling to imagine an early primordial soup, abuzz with chemical reactions, atoms traded to form all possible nucleotide variations—the useful ones staying with us. The same modifications can be found in Eukarya, Eubacteria and Archaea, although the amounts, locations and functions of specific modifications differ from organism to organism, suggesting that RNA modification is an attractive substrate for evolution.

1.1 Introduction to eukaryotic RNA modifications

1.1.1 RNA modifications in the process of gene expression

The process of protein-coding gene expression requires recruitment of RNA Polymerase II (RNAPII) to genes to catalyse the transcription of DNA into messenger RNA (mRNA). pre-mRNA must be processed into mature mRNA through capping, splicing and polyadenylation to enable correct translation of the coding RNA sequence into protein in the cytoplasm (Figure 1.1).

In general, RNA modifications participate in eukaryotic gene expression from start to finish. The moment the first pre-mRNA nucleotide exits the RNA Polymerase II (RNAPII) exit tunnel several enzymes engage to catalyse the addition of a methylated guanosine nucleotide (m7G) (Martinez-Rucobo et al., 2015). This process, called capping, protects the pre-mRNA from 5'-exonucleases and later facilitates translation initiation. The first and second transcribed nucleotides are additionally methylated at the 2'O position of the ribose sugar (2'-O*me*), these methylations are important in mammals for designating mRNAs as “self” in comparison to invading viral RNAs (Hyde & Diamond, 2015). Where this first nucleotide is an adenosine, 2'-O*me* can be combined with N6-methyladenosine (m6A) modification by PCIF1/CAPAM, to form m6Am, which is proposed to enhance translation of such transcripts (Akichika et al., 2019).

As transcription progresses, U1 snRNP bound to RNAPII waits in the wings for signs of the first intron, so that it can bind to the 5' splice site (5'SS), an event which triggers the process of pre-mRNA splicing to excise intronic sequences (Z. Zhang et al., 2021). This splicing process is catalysed by a dynamic molecular

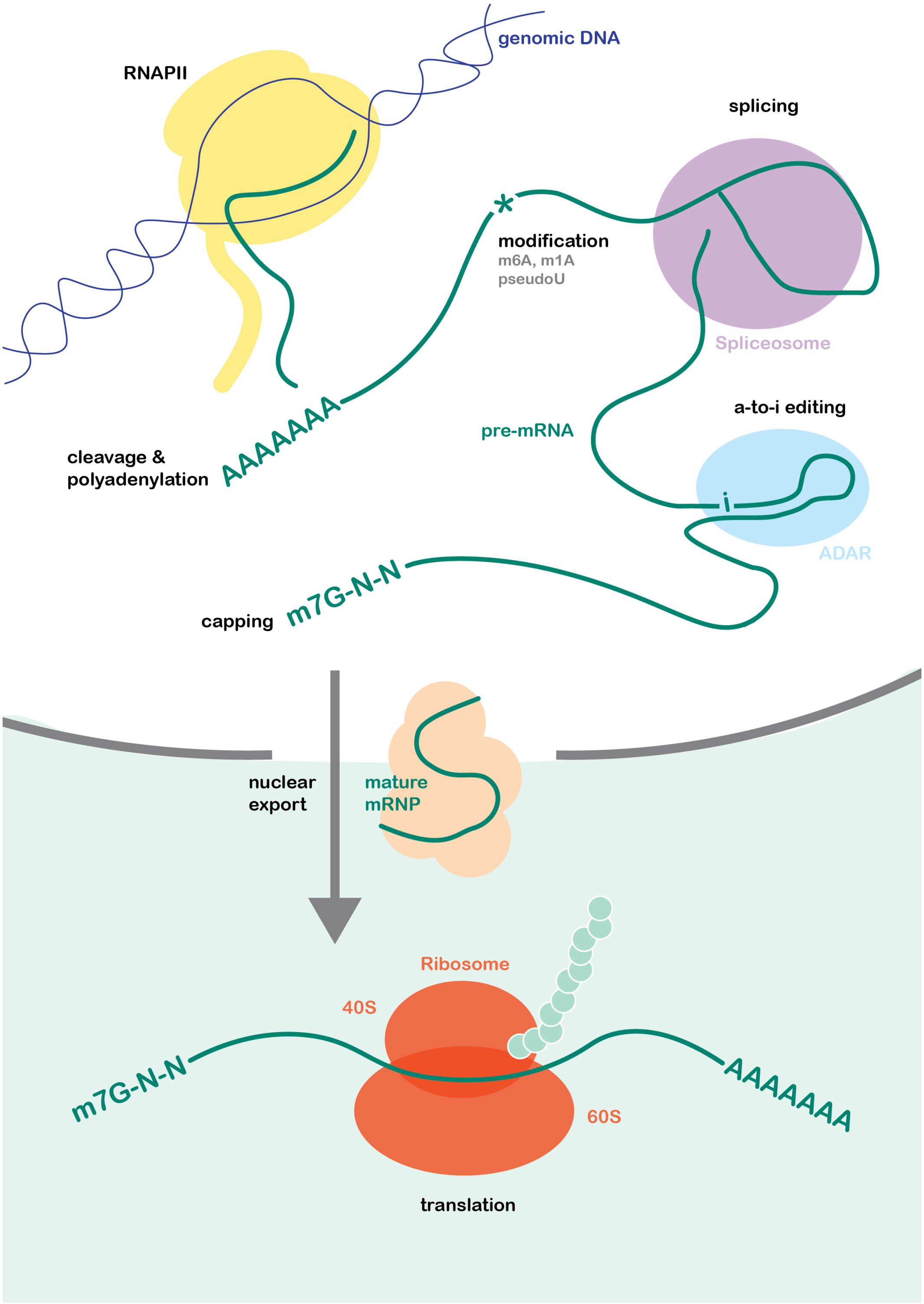


Figure 1.1: The process of protein-coding gene expression. pre-mRNA is transcribed in the nucleus by RNAPII. The pre-mRNA is capped, spliced and polyadenylated. During this process the RNA may also be modified in several different ways, and possibly edited. Mature mRNA is packaged and exported to the cytoplasm where it is translated by the ribosome.

machine called the spliceosome, which contains structural RNAs which themselves are modified.

On some transcripts, as RNAPII reaches somewhere around about the stop codon, m6A is catalysed and how and why this happens is the subject of much discussion in this thesis. As RNAPII reaches the end of the gene, the mRNA will be cleaved, somewhere in between a highly conserved AAUAAA hexamer and a downstream U/GU-rich sequence element. Poly(A) polymerase (PAP) catalyses the addition of the poly(A) tail, which is rapidly bound by poly(A) binding protein (PABP). In the cytoplasm, this poly(A)-PABP complex binds to eIF4G scaffold protein to stimulate translation initiation (Gallie, 1991).

1.1.2 Modification of structural RNAs

Two critical processes in gene expression are presided over by complex RNA-protein machines. Splicing is performed by the spliceosome, which contains structural RNAs called small nuclear RNAs (snRNAs). Translation of mRNA to protein is catalysed by the ribosome, which contains highly structured ribosomal RNAs and is aided by transfer RNAs which are loaded with the required amino acids.

Spliceosomal snRNAs contain many modifications: their caps are trimethylated, several uridines are converted to a 5-ribosyl isomer (pseudouridine, Ψ) and they, too, harbour 2'-O*me*. Mammalian U6 snRNA contains a single m6A in the sequence that recognizes the 5'SS (ACm6AGA); this seems especially important for splicing of introns where this snRNA adenosine meets the intron also at an adenosine, with work suggesting that methylation of the snRNA improves the thermostability of the intron-snRNA duplex in this context (Ishigami et al., 2021).

Translation of proteins needs to happen at the right time, in the right place and with fidelity. The most well understood function of RNA modifications is in the safeguarding of accurate translation, which follows as tRNAs are the most highly modified RNAs in the cell (Figure 1.2). RNA modifications which sterically block Watson-Crick base pairing are important for maintaining the open loops in tRNA structure, for example m1G in the anticodon loop and m1A in the T loop. m1G37 adjacent to the anticodon wobble position is also important for preventing frameshifts, in fact other bulky modifications at the Watson-Crick interface are also used for this purpose at this position, including cyclic N6-threonylcarbamoyladenosine (Björk et al., 1989; Klassen et al., 2017). Certain modifications are crucial for binding of ions, m5C in tRNAs promotes distant binding of Mg2+ which stabilises stem structure, and is required for ribosome binding (Y. Chen et al., 1993; Dao et al., 1994).

The ribosome structure itself also requires stabilising RNA modifications which are deposited during ribosome biogenesis, for example pseudouridines enhance base stacking interactions leading to more stable structure (Davis, 1995). Not only this, but rRNA modifications are enriched in functional areas of the ribosome that could impact translation, for example 60% of yeast rRNA modifications occur in functional regions such as the large subunit A, P and E sites, the peptidyl transferase active site and polypeptide exit tunnel (Decatur & Fournier, 2002). Over the past decade a growing body of research suggests the existence of "specialised ribosomes" differing in sequence, expression, localisation and interestingly modification content, opening up the possibility of previously unrealised layers of translational control (Xue & Barna, 2012).

1.1.3 Modification of mRNAs

mRNA itself can be modified in ways other than capping. m6A is often described as "the most abundant internal modification of mRNA". This statement contains several caveats: "of mRNA" is included because pseudouridine, the first discovered RNA modification, is the most abundant modification in the cell (Machnicka et al., 2013) and "internal" is included because of the previously discussed 5' capping modifications invariably present on mRNAs. Behind m6A in terms of abundance is adenosine to inosine deamination, catalysed by ADAR enzymes, which is more often termed "RNA editing", it is most abundant in brain mRNAs where it is estimated to be present at a rate of 1 in every 17,000 nucleotides (Paul & Bass, 1998). The distinction between modification and editing is generally between mod-

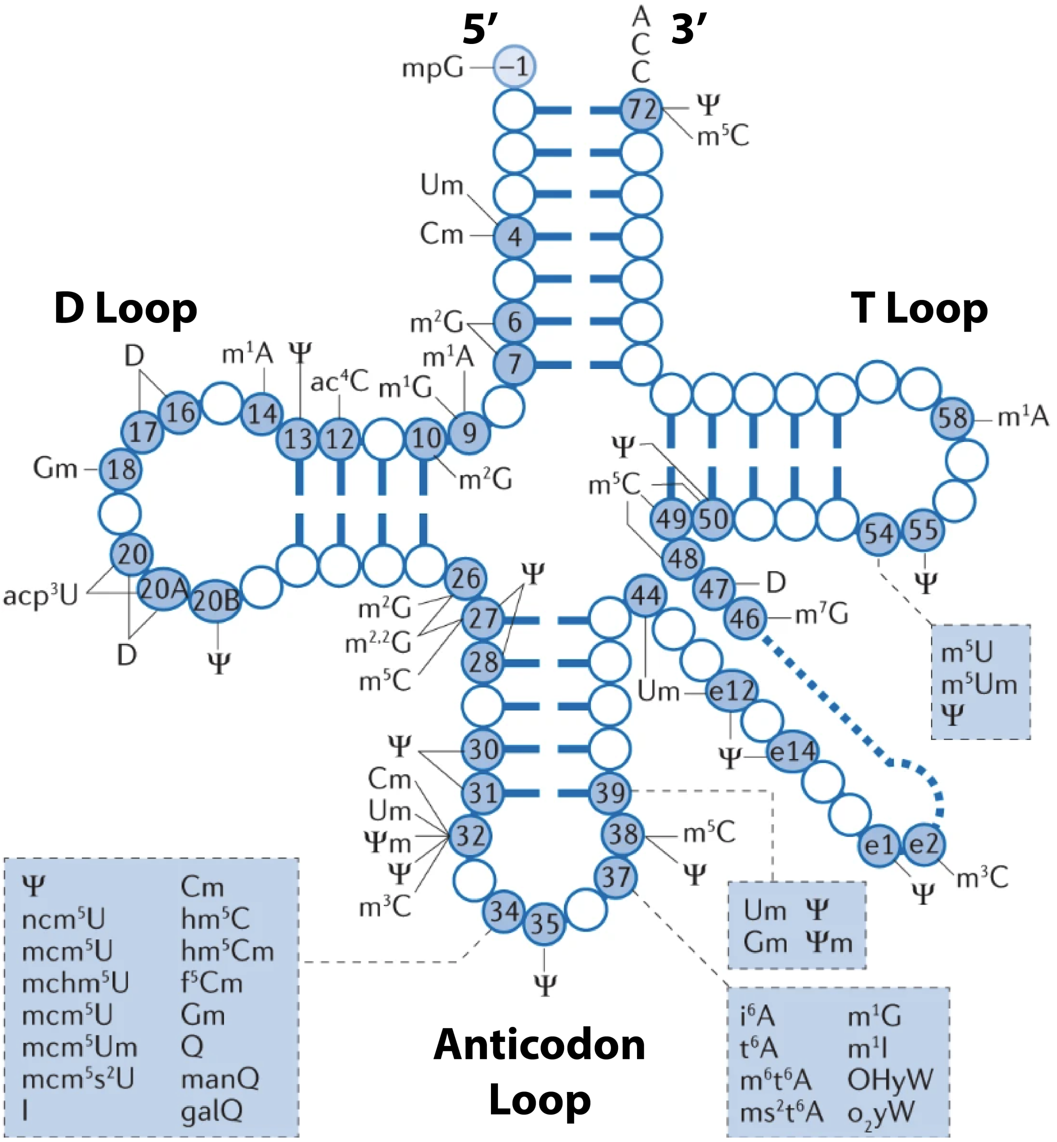


Figure 1.2: tRNAs are the most highly modified RNA in the cell. Shown are human cytoplasmic tRNA modifications with abbreviations as set in the MODOMICS database (Machnicka et al., 2013). Figure modified from (Suzuki, 2021).

ifications which cannot change the resulting protein sequences, and so-called edits which can. In the case of inosine, if present in a coding sequence it will be "read" by tRNAs as a cytosine rather than adenosine. When this changes the amino acid it can have functional consequences for the protein. Other internal modifications of mRNA are present at much lower frequencies and therefore have proved troublesome to accurately detect and map. Mass spectrometry approaches to detecting rare RNA modifications in poly(A)+ purified RNA are often hampered by persistent rRNA contamination and many sequencing approaches face high false positive rates (Anreiter et al., 2021). Pseudouridine is estimated to occur on 89-1929 transcripts and there is one validated m1A site in *ND5* mRNA (Carlile et al., 2014; Safra et al., 2017; Schwartz et al., 2014). Also suggested to occur in mRNA are 2'-O*me* of ribose (particularly of uridine nucleotides), m7G, m5C, hydroxy-m5C, m3C and N4-acetylcytidine (Anreiter et al., 2021).

1.1.4 History of m6A research

m6A was first described in the 1970's: in HeLa cell poly(A)+ RNA m6A was found to be present at an average rate of 1 in every 1000 adenosine nucleotides (Wei et al., 1975). More recent measurements by quantitative mass spectrometry of total RNA from post-mortem human tissues agree suggesting that 0.11–0.23% of adenosines are modified with m6A (J. Liu, Li, et al., 2020). There is some evidence that m6A levels change substantially in cancer, whilst this is mostly inferred from expression of methyltransferases, mass spectrometry measurements in circulating tumour cells vs. whole red blood cells were found to be consistently doubled, although it's notable no measurement rose above ∼0.3% (W. Huang et al., 2016). Interestingly, m6A levels in budding yeast meiosis echo these measurements, being ∼0.15-0.2% of adenosines (Varier et al., 2022). In *Drosophila* m6A peaks in early developmental stages, but never reaches above 0.1% in mRNAs (Kan et al., 2017; Lence et al., 2016). Such a narrow range possibly hints at metabolic constraints within eukaryotic cells, or perhaps a very deleterious impact of levels much higher than this.

Long before any m6A sites were mapped to specific residues, or the enzymes involved were known, the consensus sequence motif was deduced by chromatography methods to be N-(G/A)-m6A-C-N, in mouse RNA, with the 5' nucleotide being G ~75% of the time and A ~25%. The 5' N was determined to be G or A 90% of the time and the 3' N was rarely a G (Schibler et al., 1977). This motif is commonly referred to as DRACH (D=G/A/U,R=G/A, H=A/C/U) or occasionally RRACH. Fascinatingly, it was already proposed in this early work that m6A was installed in mRNA in the nucleus, due to the same sequences being identified in heterogenous nuclear RNA and mRNA fractions (Schibler et al., 1977).

For decades the only known precise positions of m6A on mRNA were on the highly abundant bovine prolactin mRNA and in the Rous sarcoma virus RNA genome. m6A was localised to 108nt in the prolactin mRNA 3'UTR (Horowitz et al., 1984) and seven sites of m6A were identified at single nucleotide resolution in the 3' half of the Rous sarcoma virus genome (Kane & Beemon, 1985). It was determined that modification of Rous sarcoma virus genome was not constitutive, which seems to be a general feature of m6A- that individual mRNA molecules in a population will be heterogenous at modification sites. Second, authors noted the clustering of m6A in specific regions, where multiple sites would be in close proximity (Kane & Beemon, 1985). To this day it's unclear what this says about the kinetics of m6A or what consequences this might have for function.

1.1.5 Components of the m6A-methylating complex and their assembly on RNA

In mammals, m6A is installed by a Mettl3/Mettl14 heterodimer, with Mettl3 being the catalytically active enzyme (Schöller et al., 2018). The binding interface between Mettl3 and Mettl14 creates a groove for RNA contacts, and N-terminal Mettl3 zinc finger motifs also contribute to RNA binding with some apparent recognition of RNA secondary structure (Meiser et al., 2020; Śledź & Jinek, 2016). Wilms' tumor 1-associating protein (WTAP) is also required for *in vivo* mRNA modification and is thought to recruit the Mettl3/Mettl14 complex to nuclear speckles

(Ping et al., 2014). In mammalian cells, Mettl3, Mettl14 and WTAP all Co-IP with RNAPII and H3K36me3 modified chromatin (Huang et al. 2019). Additionally, the same m6A marks are seen in chromatin-associated and nuclear mRNA fractions as in whole cell purified mRNAs, suggesting m6A is installed in nascent RNAs, potentially co-transcriptionally (Ke et al. 2017). Further proteins are also required for *in vivo* methylation, making the mRNA m6A-methylating complex a large seven-component molecular machine, including Zc3h13, Virilizer (also known as KIAA1429 or VIRMA), Hakai and Rbm15 (Patil et al., 2016; J. Wen et al., 2018; Yue et al., 2018).

In mammals, there are at least three other m6A methlytransferases. METTL5-TRMT112 methyltransferase complex installs the m6A modification at position 1832 of human 18S rRNA (van Tran et al., 2019), whilst ZCCHC4 installs m6A at position 4220 in 28S rRNA (Ma et al., 2019). Further, Mettl16 methylates U6 snRNA, MAT2A and various other ncRNAs in highly conserved hairpin loops of specific sequence (Pendleton et al., 2017; Shima et al., 2017; Warda et al., 2017).

1.2 Methods for mapping m6A across transcriptomes

Note that the following section is adapted from my review article (Capitanchik et al., 2020).

In order to understand the general principles of m6A-mediated mRNA regulation, first we need to discern which adenosines are methylated on which transcripts in a given cell or tissue at a given time. Over the past decade there have been many leaps forward in technologies measuring m6A, starting with antibody-based approaches, and more recently branching out into enzymatic, metabolic and direct RNA sequencing methodologies; all of which are discussed in the following section (Figure 1.3).

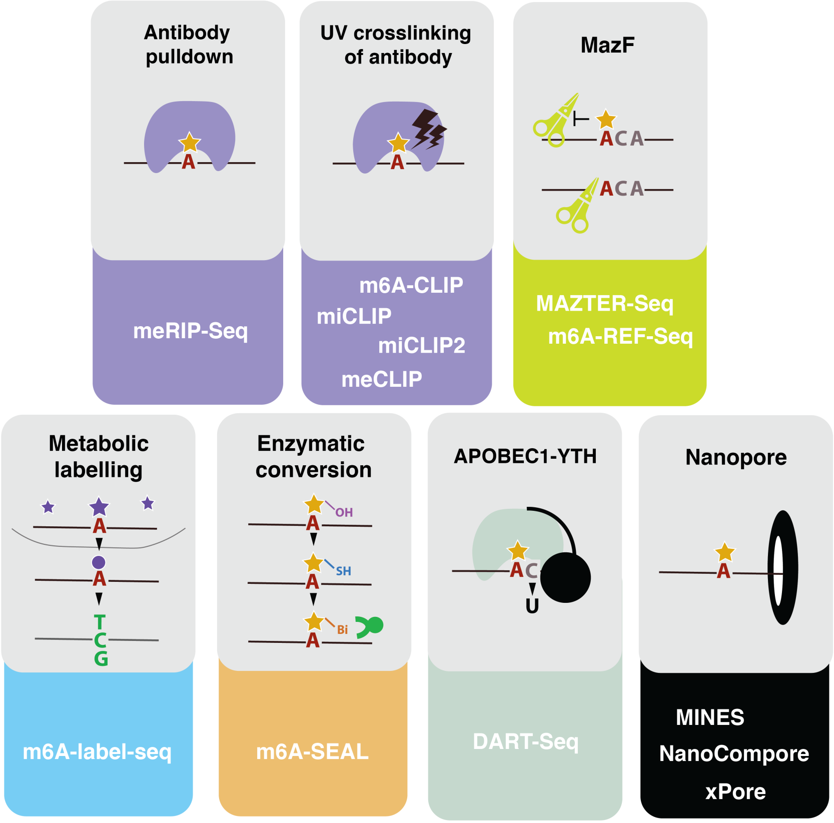


Figure 1.3: A summary of methods for mapping m6A across transcriptomes. Figure updated from (Capitanchik et al., 2020).

1.2.1 Antibody-based methods

In 2012, two groups simultaneously published the first methods to map m6A transcriptome-wide, m6A-seq and MeRIP-seq, utilising antibodies that bind m6A to perform RNA immunoprecipitation, followed by high-throughput sequencing (Dominissini et al., 2012; K. D. Meyer et al., 2012). The resolution of m6A-seq is limited to the length of RNA fragments, with no objective way of determining where in the fragment the modification occurred.

Greater resolution was achieved by UV-crosslinking the antibody to RNA, following the principles of the CLIP protocol (J.-H. Lee et al., 2021). Such approaches were simultaneously developed in the laboratories of Samie Jaffrey and Robert Darnell, named miCLIP and m6A-CLIP respectively (Ke et al., 2015; Linder et al., 2015). Here, purified RNA is incubated *in vitro* with an m6A antibody. The polyclonal Abcam and Synaptic System antibodies were determined to be most efficient at enriching for m6A and gave the most predictable mapping signatures; as a result, they remain the most commonly used antibodies in subsequent miCLIP publications. Following immunoprecipitation, the antibody is digested with proteinase K, leaving an amino acid adduct attached to the RNA base. During preparation of the cDNA library, the reverse transcriptase either reads through this crosslinked adduct, causing a substitution or deletion mutation, or is stopped, resulting in cDNA truncation. These signals can be analysed computationally to give near single nucleotide resolution of the modification site (Haberman et al., 2017).

m6A-CLIP is conceptually similar, but slightly different: the protocol exclusively uses the Synaptic Systems antibody which optimises for cDNA truncations rather than mutations, RNA fragments are size-selected prior to immunoprecipitation and cDNAs are purified using BrdU. There are also differences in the starting RNA-to-antibody ratios - miCLIP uses an excess of RNA, whereas m6A-CLIP uses an excess of antibody.

Since then, an adaptation of the eCLIP protocol was introduced, called m6A-eCLIP (meCLIP) which uses the Abcam m6A antibody. The analysis focuses entirely on C-to-T conversions and filters for DRACH motifs in the resulting data (Roberts et al., 2021; Van Nostrand et al., 2016). More recently, miCLIP2 improved on previous protocols with several experimental optimisations, choosing to optimise cDNA truncations using the Synaptic Systems antibody (Körtel et al., 2021). Namely, bead-based cDNA size selection is performed as opposed to gel based in miCLIP, saving time. Prior to this, a first PCR of six cycles means that cDNA molecules lost in the size selection are still part of the final library. Further, longer unique molecular identifiers (UMIs) ensures that they do not become saturated at high coverage positions. Finally, as opposed to the original iCLIP-style adapter ligation which required circularisation of the molecule before relinearisation, miCLIP2 directly ligates the 3’ sequencing adaptor to the 3’ end, which minimises loss of the sample. Taken together, these improvements result in higher complexity libraries that can be produced using less input material. Additionally, an adaptive boosting machine learning model was trained on WT vs. Mettl3 KO data, which can be used to call sites of m6A modification from miCLIP2 data to remove non-specific background signal (Körtel et al., 2021).

Despite these advances the promiscuity of m6A antibodies is still a significant drawback, they are known to have additional preference for unmodified poly(A) stretches and m6Am (Schwartz et al., 2013). Studies generally tackle this issue by only reporting sites found within the consensus DRACH motif or by perturbing methyltransferase activity. Neither is optimal: DRACH-only reporting prevents discovery of m6A in non-canonical motifs, whereas knockout or knockdown controls exclude sites that can be modified by another methyltransferase. Furthermore, disrupting the m6A machinery may introduce global changes in RNA abundance that are difficult to account for, except with the careful use of input libraries and spike-ins (J. Liu, Dou, et al., 2020).

Methods that rely on crosslink-induced mutations as the readout require higher read coverage to call sites than those based on truncation detection. Additionally, for all strategies integrating multiple control datasets (methyltransferase depletion, RNA input, etc) the increased variance may reduce the statistical power to call sites. In summary, though antibody-based methods have been fundamental to paving the way for transcriptomic analysis of m6A, and remain the most common way to survey the modification, issues with antibody specificity make the development and use of orthogonal approaches desirable.

1.2.2 Enzymatic methods

The known biases in antibody-based mapping have led to the development of alternative methods. DART-seq employs the *in vivo* expression of an m6A-recognising YTH protein domain fused to the APOBEC1 enzyme which deaminates cytosine to uracil (K. D. Meyer, 2019). Thus, this construct deaminates cytosine residues nearby to m6A sites bound by YTH, which can be subsequently quantified as mutations in sequencing. Chromatography studies suggest that m6A is invariably followed by cytosine (Wei et al., 1976) meaning individual m6As could theoretically be identified at single nucleotide resolution, although in practise $\sim$80% of identified sites are >20 nucleotides away from the nearest miCLIP site. On the otherhand, as the YTH-APOBEC1 construct can be transiently expressed in cells, library preparation is more straightforward than either the antibody or enzyme-based methods, because no post-extraction treatment of the RNA is required. Likely for this reason, libraries can be produced with as little as 10ng of total RNA. One concern is that using APOBEC1 alone as a control might introduce false negatives, as it also has 3'UTR preference and 70% of APOBEC1-only deaminated cytosines are preceded by an adenosine.

The discovery of MazF endoribonuclease, an enzyme that cuts preferentially at ACA over m6A-CA sites, led to the development of two new methods: MAZTER-seq and m6A-REF-seq (Garcia-Campos et al., 2019; Imanishi et al., 2017; Z. Zhang et al., 2019). mRNA treated with Mazf produces fragments flanked by ACA sequences, therefore any internal ACA within a fragment is reasoned to be m6A modified. By carefully characterising the biases of MazF the authors were able to calculate stoichiomtetric information, which is missing from many of the described methods. Clearly, the limitation of the MazF enzyme to ACA sites and the extensive filtering requirements on the resulting datasets do mean that MazF-based methods alone cannot provide a full transcriptome-wide map of m6A. Nevertheless, these methods offer valuable complementary information.

More recently, the m6A-SEAL approach exploited unstable reaction intermediates produced by the human m6A demethylase FTO to enable labelling of m6A-containing RNA fragments (Y. Wang et al., 2020). RNA was purified from cells then incubated with recombinant human FTO protein in conditions conducive to oxidation at the m6A methyl group, creating N6-hydroxymethyladenosine (hm6A). This hydroxyl group could then be thiolated using dithiothreitol (DTT) to become N6-dithiolsitolmethyladenosine (dm6A). At this point the RNA is then biotin labelled and enriched using streptavidin pulldown and RNA fragments are prepared into sequencing libraries. Due to the many enzymatic/chemical processing steps, 5ug of poly(A)+ RNA is required as starting material, also the reproducibility between replicates is quite poor. Additionally, the resolution is lower than other nonantibody methods as fragments are $\sim$100nt long.

1.2.3 Metabolic approaches

m6A-label-seq is an innovative approach relying on feeding cells with Se-allyl-L-selenohomocysteine instead of methionine, which means the methyl group on the SAM cofactor is replaced with an allyl group, and therefore that allyl-SAM will be used instead of SAM in the methylation reaction leading to N6-allyladenosine (a6A) instead of N6-methyladenosine (Shu et al., 2020). RNA is purified and fragmented, and a6A-containing fragments are enriched using a commercial antibody. The a6A is then cyclysed using iodination and in reverse transcription this nucleotide is converted to a T,C or G which can be detected as a mutation after high throughput sequencing. Due to the many manipulations the method requires 5ug of poly(A)+ RNA, however it does offer true single nucleotide resolution. Another drawback is that treatment with allyl-SAM does cause many gene expression changes in the cell, including in the m6A machinery components themselves.

1.2.4 Direct RNA sequencing methods

The ideal scenario would be to read m6A from transcripts with no enzymatic or metabolic manipulation, or antibody enrichment at all. Direct RNA Nanopore sequencing has proved a popular way to try and achieve this aim. As RNA five-mers move through the immobilised protein pore, changes in current and pore dwell time are used to identify bases. Theoretically, modified bases should look different to their unmodified counterparts. However, in practise deconvolving the raw signal to measure m6A sites is not so straightforward. Investing in optimising Nanopore methods is worthwhile, because with little RNA preparation single molecule information is available. This makes it possible to study the combinatorics of m6A sites on the same transcripts—something impossible by any other means. It should also be possible to study multiple different kinds of modification simultaneously in one experiment.

The first Nanopore m6A transcriptome was produced for yeast mRNA (H. Liu et al., 2019). A Support Vector Machine (SVM) called EpiNano was trained on Nanopore sequencing data of synthetic transcripts containing m6A residues in every possible five-mer combination to identify the most informative signals that distinguish m6A from other nucleotides. The raw current intensities alone were found to be poor predictors of methylation status; instead, the selected training features included mean per-base quality, mismatch frequency and deletion frequency. The model achieved 90% prediction accuracy for the training dataset. It was then used to recover 363 previously identified, high-confidence m6A sites, previously identified using m6A-seq, which it was able to do with 87% accuracy.

An alternative approach, MINES (m6A Identification using Nanopore Sequencing), was applied to produce the first human m6A Nanopore transcriptome (Lorenz et al., 2019). This method applied Tombo, a program that was previously developed to detect *de novo* modifications in Nanopore DNA sequencing data using the base-calling error rate (Oxford Nanopore Technologies, 2018). The authors trained random forest models using the Tombo modification values to classify the m6A status of four RGACH motifs. Those RGACH sites overlapping with HEK293 miCLIP and HeLa m6A-CLIP sites (Ke et al., 2017; Linder et al., 2015) were labelled as true positives in the training data, and the models achieved an average accuracy of 79%, identifying 35% of m6A sites identified with CLIP-based methods (in part due to the motif restriction).

A further approach is NanoCompore (Leger, Amaral, Pandolfini, Capitanchik, et al., 2019a), which compares Nanopore signals between two datasets and therefore does not require a training dataset. Specifically, this is achieved by contrasting the median current intensities and dwell times of kmers between the experiment and a control with perturbed modifications (e.g. wildtype vs. knockdown, or *in vitro* modified vs. unmodified controls). For this reason, NanoCompore isn't restricted to m6A and can be readily extended to other modifications which have a reliable control. Another approach, xPore, is conceptually similar (Pratanwanich et al., 2021). A major advantage is that it avoids being biased by the accuracy of previous mapping methods (ie. CLIP) to assign modification sites, as site identification is instead determined by the sensitivity to a specific modification enzyme. Of course, the dependence on a comparison between samples is also a limitation, as reliable controls are currently unavailable for many modifications and biological systems, and specific sites or RNA species are often modified by distinct enzymes. As a result, there is probably a reduced risk of false positive site assignment at the cost of sensitivity.

Interestingly, a much simpler comparative approach was used to map the *Arabidopsis thaliana* m6A transcriptome (Parker et al., 2020), in which the base-calling error rate was used as the sole input. Signals were compared between a vir-1 (*Arabidopsis* m6A methyltransferase) mutant and a vir-1 restored line, m6A was assigned where the error rate was two-fold greater in the control line compared to mutant. Taking this approach 66% of identified m6A sites fell within five nucleotides of a miCLIP peak.

The above methods demonstrate that direct RNA sequencing can be used to detect m6A. A common limitation pertains to the resolution and accuracy of modification assignment for transcripts with low sequencing depth. As five nucleotides move through the pore at a time it can be hard to determine the exact site of modification. However, with third-generation sequencing technologies developing rapidly, the benefits of using direct sequencing to map RNA modifications are likely to push the boundaries of the field.

1.2.5 The best way to map m6A across transcriptomes

Currently, antibody-based CLIP approaches, alongside powerful predictive algorithms such as m6aboost, are still likely to provide the most "bang for your buck" in an experiment aimed at mapping m6A generally in a single condition, or in making comparisons across conditions. Whilst the false positive rate is high, the coverage of m6A sites is high across a range of RNA input amounts, and a broad range of sequence contexts can be detected. Furthermore, the method doesn't require any genetic editing which can be difficult in many systems. That being said, the future of m6A mapping is likely to exist in a combination of chemical and Nanopore approaches. Efforts to chemically modify m6A residues to a bulkier modification that can create a reliable reverse transcription signature currently have high specificity, but relatively low efficiency. Optimising the efficiency of such chemical reactions could allow us to achieve similar or better m6A coverage than miCLIP, with a much lower false positive rate. I think that the ideal approach would combine these efficient chemical reactions with high depth Nanopore sequencing. Currently Nanopore-based methods still also face issues with specificity, but again in this context converting m6A to a bulkier modification could aid detection by creating a more distinctive signature as the nucleotide moves through the pore. Nanopore-based methods also still have a problem with low coverage transcriptome-wide that will hopefully be alleviated in the years to come.

1.3 m6A readers and function

Knowing the sites of m6A modification is only the first step to understanding function. It is becoming increasingly apparent that in order to understand the functions of m6A modification, we need to understand the principles guiding RBPs that bind to or interact with m6A sites. Only through these proteins are m6A impacts on mRNA fate made possible.

1.3.1 YTH proteins: canonical m6A readers

In mammals, there are five proteins containing YT521-B homology (YTH) domains—these domains confer m6A-binding capacity. Some organisms, such as *S. cerevisiae*, have only one YTH protein, whilst *A. thaliana* has eleven (Arribas-Hernández, Rennie, Köster, et al., 2021; Schwartz et al., 2013). m6A is bound in a highly conserved hydrophobic aromatic cage, as determined by crystal structures of m6A-modified RNA oligonucleotides and the YTH domains of human YTHDC1 (Xu et al., 2014), human YTHDF2 (T. Zhu et al., 2014), human YTHDF1 (Xu et al., 2015), *Zygosaccharomyces rouxii* MRB1 (ZrMRB1) and *S. cerevisiae* Pho92p (Xu et al., 2015). However, reported *in vitro* affinities of this interaction suggest it might not be very stable, and may require additional interactions to stabilise the RNA-protein contact (Patil et al., 2018). iCLIP of *A. thaliana* YTH protein ECT2 identified crosslinking to both m6A motifs and upstream U-rich motifs, which were determined to represent contacts of low complexity regions of the protein stablising the YTH domain interaction (Arribas-Hernández, Rennie, Schon, et al., 2021). In mammals, three of the YTH proteins are paralogs, resulting from two gene duplication events - these form the YTH domain-containing family (YTHDF). Whilst all three share a high degree of sequence similarity, a C-terminal YTH domain and mostly cytoplasmic cellular localisation, YTHDF1 and YTHDF3 are more closely related than YTHDF2.

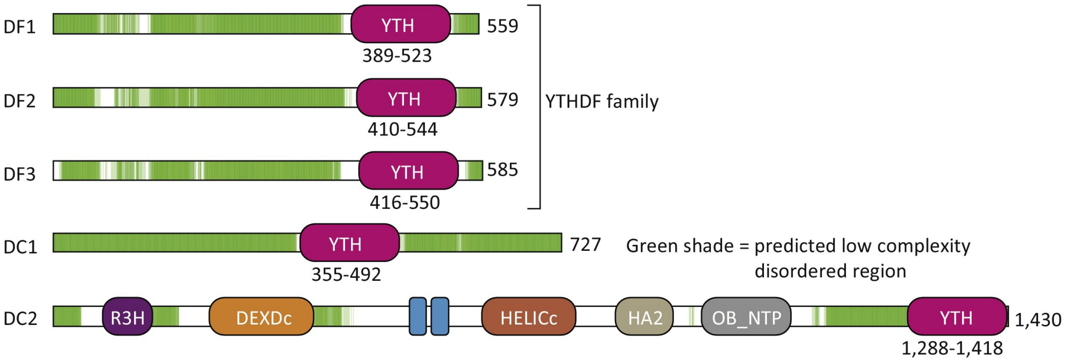


Figure 1.4: Domain structures of YTH domain-containing proteins in mammals. Figure adjusted from (Patil et al., 2018), with permission from Elsevier.

The remaining YTH proteins, YTHDC1 and YTHDC2 are very different—both from each other and the YTHDF proteins (Figure 1.4). YTHDC1 is larger than the YTHDF proteins, has an internal YTH domain and is localised to the nucleus. YTHDC2 is a swiss-army knife of a protein with a C-terminal YTH domain, but also domains consistent with its designation as a member of the DEAH/RNA helicase A (RHA) helicase family (DEXDc, HELICc, OB-fold) (Jain et al., 2018). Additionally, it contains an R3H domain involved in nucleotide binding (Grishin, 1998) and ankyrin repeats (ANK) which are usually involved in protein-protein interactions (J. Li et al., 2006). In some organisms, such as *D. melanogaster* and *C. elegans* the YTH domain has been lost from YTHDC2 - note that *C. elegans* has no m6A in mRNA, but *Drosophila* does (Jain et al., 2018). Structural and *in vitro* work suggests that the YTHDC1 and YTHDC2 YTH domains have less affinity for m6A than the YTHDF family domains, where Tyr260 is replaced by a Leu in the YTHDC proteins (Luo & Tong, 2014). All five proteins have predicted low complexity disordered regions which may contribute to RNA and protein interactions.

One might suppose that specificity of m6A function could be guided by different sequence preferences of different YTH proteins directing mRNAs to different fates. However, this doesn't appear to be the case. *In vitro* work subjecting purified YTH domains to isothermal titration calorimetry (ITC) suggests that different YTH domains recognise different DRACH motifs with similar affinity, with the exception of the YTHDC1 YTH domain which shows some preference for GG-m6A-CU over GA-m6A-CU (Xu et al., 2015). Interestingly, *Drosophila* YTHDC1 and YTHDF proteins prefer A-rich DRACH motifs, which correlates with *Drosophila* bias in m6A sequence contexts (Kan et al., 2021).

1.3.2 Non-canonical m6A readers/interactors

Alongside YTH proteins, several other proteins have been proposed to read or interact with m6A, although the structural basis of these interactions mostly remains to be determined. The most well characterised "non-canonical" group of reader proteins are insulin-like growth factor-2 mRNA-binding proteins 1, 2, and 3 (IGF2BP1–3). These proteins use their third and fourth KH domains (KH3–4) to interact with m6A, along with some flanking sequences; however, the exact nature of this interaction is unclear, as many RBPs contain KH domains but do not specifically recognise m6A (H. Huang et al., 2018). Despite this, IGF2BP1–3 seem to bind m6A specifically in UGG-m6A-C context to stabilise m6A-containing transcripts (H. Huang et al., 2018).

A common technique to identify novel m6A readers is to incubate cell lysate with an m6A-modified bait oligo and compare pull down of proteins using the bait vs. an unmodified oligo of the same sequence. By using this method with neuronal cell lysate (HT-22 cells), Proline rich coiled-coil 2A (Prrc2a) and Prrc2c were identified. Interestingly, PAR-CLIP suggested Prrc2a has a similar binding preference to IGF2BP (UGGAC) and might also perform a similar function in stablising transcripts as Prrc2a m6A-binding appears to stabilise the *Olig2* transcript (R. Wu et al., 2019). This is despite Prrc2a not having any annotated KH domains.

A further validated non-canonical binder of m6A is the translation initiation factor complex eIF3, which binds to m6A in DRACH sequence context. *In vivo*, this promotes cap-independent translation of certain transcripts containing m6A in their 5'UTR. The m6A binding capacity could not be assigned to a specific subunit alone, leading the authors to conclude that the m6A recognition occurs at a multi-subunit interface, representing yet another mechanism of m6A binding (K. D. Meyer et al., 2015). Determining the structural basis of these non-canonical m6A interactions will be critical in the design of future studies, as being able to target mutations that abolish m6A-reading ability will enable us to fully untangle the contribution of m6A to the function of these RBPs.

In 2012, an m6A-bait mass spectrometry screen identified HuR/ELAVL1 as an m6A reader too (Dominissini et al., 2012). Through further research it was established that HuR is actually binding to U-rich regions adjacent to m6A sites (K. Chen, Lu, et al., 2015; Y. Wang et al., 2014). In the paper the authors conclude that HuR is increasing RNA stability due to blocking microRNA targeting (Y. Wang et al., 2014), however I think it's possible that HuR binding inhibits YTHDF binding, because the U-rich sequences are required for stabilisation - meaning that these transcripts are no longer targeted for decay.

m6A can also change RBP binding in more indirect ways: m6A modification at hairpin loops can melt RNA secondary structure to reveal single stranded HNRNPC binding sites (N. Liu et al., 2015)

1.3.3 Role of YTH proteins in translation and decay

In human HeLa cells, siRNA knockdown of YTHDF1 resulted in reduced translational efficiency of shared YTHDF1/2 target transcripts (identified by PAR-CLIP), but no change in mRNA half-life (X. Wang et al., 2015). Conversely siRNA knockdown of YTHDF2 resulted in a longer half-life but no change in translational efficiency, suggesting a differing role for the two proteins (X. Wang et al., 2014; X. Wang et al., 2015). siRNA knockdown of YTHDF3 in HeLa cells reduced translation efficiency of target transcripts, but interestingly tethering of YTHDF3 alone to a luciferase reporter didn't increase its translation, but tethering both YTHDF1 and YTHDF3 to the reporter increased its translation more than just YTHDF1 alone (Shi et al., 2017). Flag-tagged Co-IP of the YTH proteins suggest they interact with each other in an RNA-independent manner. Depletion of YTHDF3 led to more association of YTHDF1/2 with non-specific targets, whereas depletion of YTHDF1/2 led YTHDF3 to bind less to RNA, suggesting YTHDF3 recruitment to RNA is somewhat dependent on YTHDF1/2. The authors also suggest YTHDF3 works with YTHDF2 to promote mRNA decay as triple YTHDF1/2/3 depletion led to more m6A accumulation than YTHDF1/2 depletion (Shi et al., 2017).

Mouse knockout of YTHDF1 leads to mice with learning and memory defects, with experiments suggesting that YTHDF1 promotes translation in hippocampal neuron cell culture, especially in response to potassium chloride depolarization (Shi et al., 2018). Further evidence demonstrated that YTHDF1 and m6A sites in the Robo3.1 mRNA were required for correct neuronal development in mice (Zhuang et al., 2019). In mouse embryonic fibroblasts, reprogramming to iPSCs is hindered by shRNA knockdown of YTHDF2/3, but not YTHDF1 (J. Liu, Gao, et al., 2020). Results of Co-IP experiments and combinatorial knockdowns suggested that YTHDF2 mediates RNA decay through the CCR4-NOT complex, whilst YTHDF3 recruits the PAN2-PAN3 hetero-multimeric complex (Du et al., 2016; J. Liu, Gao, et al., 2020).

Despite these experiments other work in HEK293, mouse embryonic stem cell development, arabadopsis and zebrafish development suggests that the binding sites, and function in mRNA decay, of the YTHDF proteins is almost entirely overlapping (Arribas-Hernández, Rennie, Schon, et al., 2021; Kontur et al., 2020; Lasman et al., 2020; Zaccara & Jaffrey, 2020). I would suggest that in light of the previously detailed work, and differing expression of YTHDF1/2/3 across cell types and developmental stages that the reality is more nuanced, and that whilst perhaps there may be cellular contexts in which the three proteins are interchangeable, there are probably details that mean this is somehow sub-optimal for cells. For example in mouse gametogenesis YTHDF2 knockout is lethal compared to YTHDF1/3 knockout which has no impact, due to differences in spatial localisation and abundance (Lasman et al., 2020). Its unclear if YTHDF1/3 could rescue the defect if their spatial localisation and abundance were matched, or if these factors are intrinsically linked to the YTHDF2 sequence in some way.

Certainly, context is key and an interesting area of future work will be to study the impact of post-translational modifications of YTH proteins. For example, its been shown that YTHDF2 can be SUMOylated in response to hypoxia in human cell lines, leading to increased mRNA affinity (Hou et al., 2021).

1.4 The Mettl3 network in cell fate choices across the tree of life

Across species m6A is utilised in different ways, but one unifying theme seems to be use of m6A in dynamic cellular transitions, especially during development (Frye et al., 2018). It's possible that m6A is an attractive evolutionary substrate for these processes where rapid mRNA turnover is required.

Mettl3 knockout in mouse and human embryonic stem cells prevents exit from pluripotency, which correlates with increased abundance of pluripotency transcripts such as *Nanog* (Batista et al., 2014; Geula et al., 2015). In zebrafish, YTHDF2 knockout results in delayed decay of maternal transcripts, leading to delayed embryonic development overall (Zhao et al., 2017). Further, endothelial cells in zebrafish embryos treated with Mettl3 morpholino failed to differentiate to hematopoietic lineages, again seemingly due to loss of YTHDF2 mediated decay of endothelial markers (C. Zhang et al., 2017). Disruption of MTA (METTL3) expression in *Arabidopsis thaliana* has recessive embryo-lethality (Zhong et al., 2008).

In flies, methyltransferase (Ime4) knockout leads to adult flies that cannot fold their wings correctly or fly (Haussmann et al., 2016; Kan et al., 2017). Additionally, Ime4 knockout and knockout of YTH protein, YT521-B, leads to a reduced female population due to increased inclusion of a male-specific exon in the sex-determination factor Sex lethal (*Sxl*). This causes erroneous upregulation of X chromosome transcripts, leading to incorrect gene dosage in affected females (Haussmann et al., 2016; Kan et al., 2017).

In honey bees, hypermethylation of caste-specific transcripts encourages worker bee development, and inhibition of m6A leads to larvae with queen-like features (M. Wang et al., 2021). In the malaria parasite *Plasmodium falciparum*, m6A is dynamic and peaks at 30 hours post-infection with m6A levels as high as 0.7% of all adenosines, however its unclear what function this has in the parasite life cycle (Baumgarten et al., 2019). The parasite *Toxoplasma gondii* uses m6A to regulate polyadenylation of transcripts, where methyltransferase knockout leads to run-on transcription into developmental stage-specific repressed genes (Farhat et al., 2021).

However, all this is not to say that there is a universal requirement for mRNA m6A in eukaryotic development - *C. elegans* for example, have only rRNA and snRNA m6A methlytransferases and no detectable m6A in mRNA (Sendinc et al., 2020).

1.5 CLIP methodologies help us to study protein-RNA interactions

Many of the existing insights into RBP readers of m6A have relied at least in part on crosslinking and immunoprecipitation technologies, which have allowed comprehensive mapping of m6A sites through miCLIP and reader protein binding through iCLIP or related variants. In this section I will give a more detailed description of the iCLIP experimental method and analysis of the resulting data, highlighting where this contrasts with miCLIP.

1.5.1 (m)iCLIP experimental Principles

In iCLIP and related methods, cells are first exposed to UV-C (254 nm) irradiation, resulting in *in vivo* covalent protein-RNA crosslinking (Figure 1.5, (F. C. Y. Lee & Ule, 2018)). In the case of PAR-CLIP, or 4sU-iCLIP, cells are first incubated with uridine analog 4-thiouridine (4sU), or in yeast, 4-thiouracil (4tU) and treated instead with UV-A (365 nm). This can increase crosslinking efficiency for some RBPs, interestingly YTH proteins seem to be among them.

After crosslinking, cells are lysed and RNA is fragmented in the cellular lysate using RNase I (König et al., 2010). In PAR-CLIP a second round of RNase digestion is performed on beads. In irCLIP all RNase digestion is performed on beads using nuclease S1 which leaves a 3' OH group on RNA fragments, such that an additional phosphatase step can be skipped - the downside is that this digestion is less efficient (Zarnegar et al., 2016). The RNase digestion step must be carefully optimised to prevent overdigestion which leads to biases that impact binding site identification (Haberman et al., 2017) and underdigestion which could lead to poor resolution and retention of larger complexes.

In the case of miCLIP, there is no *in vivo* crosslinking. RNA is first purified from cells, it can be poly(A)+ selected or depleted of rRNA at this stage, before being chemically fragmented by zinc(III)-mediated RNA cleavage.

Following RNA fragmentation, immunoprecipitation (IP) is performed on beads. In the miCLIP protocol at this point the m6A antibody of choice is crosslinked to RNA fragments (Figure 1.5 (4)). Due to the crosslinking, protein-RNA complexes can be washed more stringently than in a traditional RNA-IP approach, using ionic detergents and high salt buffers. The SeqRv adapter is ligated to fragmented RNA, which later provides sequence complimentarity to the RT primer.

Subsequently, the protein-RNA complexes are eluted from the beads and run on an SDS-PAGE gel and transferred to a nitrocellulose membrane, where the transfer helps to remove unbound RNAs due to the poor RNA-binding capacity of nitrocellulose. Typically the SeqRv adapter is intercalated with an infared dye so that protein-RNA complexes can be visualised on the membrane (Zarnegar et al., 2016).

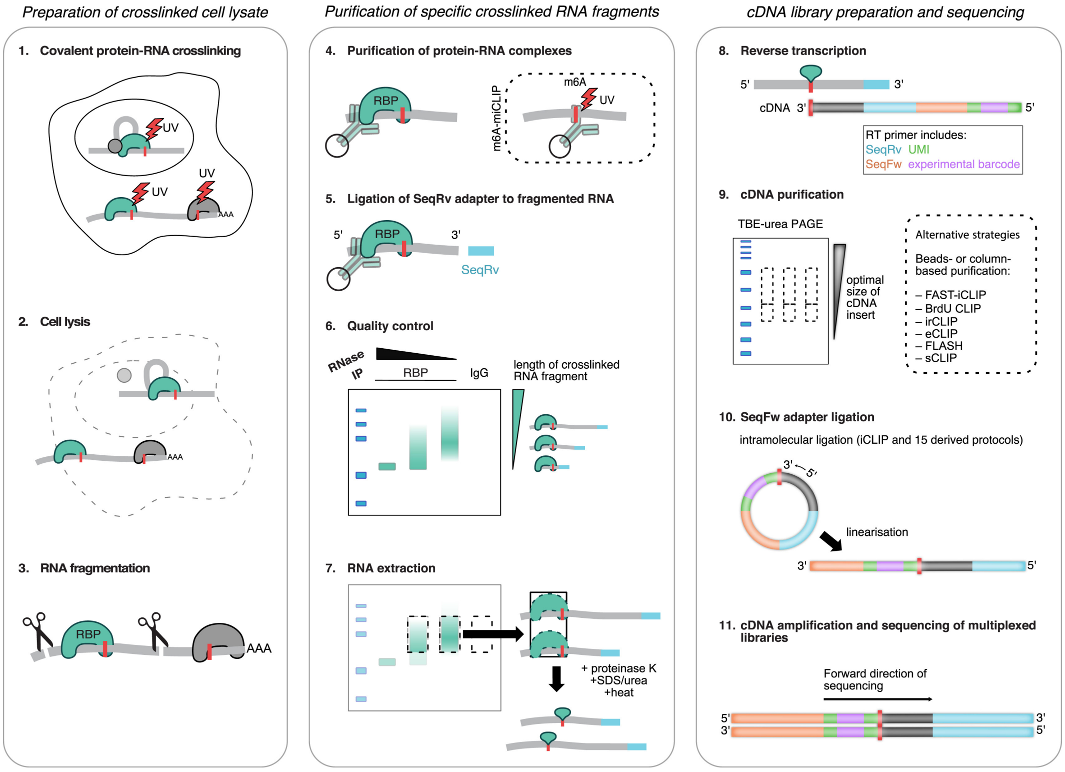

Preparation of crosslinked cell lysate
Purification of specific crosslinked RNA fragments
cDNA library preparation and sequencing
1. Covalent protein-RNA crosslinking
UV
UV
UV
AAA
2. Cell lysis
AAA
3. RNA fragmentation
RBP
AAA
4. Purification of protein-RNA complexes
RBP
m6A-miCLIP
m6A
UV
5. Ligation of SeqRv adapter to fragmented RNA
5'
RBP
3'
SeqRv
6. Quality control
RNase
IP
RBP
IgG
length of crosslinked RNA fragment
7. RNA extraction
+ proteinase K
+SDS/urea
+heat
8. Reverse transcription
5'
3'
cDNA 3'
5'
RT primer includes:
SeqRv UMI
SeqFw experimental barcode
9. cDNA purification
TBE-urea PAGE
optimal size of cDNA insert
Alternative strategies
Beads- or column-based purification:
– FAST-iCLIP
– BrdU CLIP
– irCLIP
– eCLIP
– FLASH
– sCLIP
10. SeqFw adapter ligation
intramolecular ligation (iCLIP and 15 derived protocols)
3'
5'
linearisation
3'
5'
11. cDNA amplification and sequencing of multiplexed libraries
Forward direction of sequencing
5'
3'
3'
5'

Figure 1.5: A schematic of the iCLIP protocol. Figure reproduced from (F. C. Y. Lee & Ule, 2018).

Some protocols, such as eCLIP, omit this visualisation and cut blindly from the membrane based on an estimated size range. Visualisation should show a distribution of fragment sizes above the expected molecular weight of the RBP. The membrane is then cut and the crosslinked RBP is digested with proteinase K to leave a small amino acid adduct at the crosslink position (Figure 1.5 (7)).

Following this the reverse transcription reaction is performed. The RT primer anneals to the SeqRv sequence. The RT primer contains an experimental barcode so many experiments can be combined into the same library for sequencing. Another important feature of the RT primer is the unique molecular identifier (UMI) which contains a sequence of random nucleotides of defined length that can be used later in the analysis to remove PCR duplicates. Use of different RT enzymes and conditions can result in truncation, mutation or small deletions at the crosslink site, every approach will also result in differing proportions of readthrough events (Figure 1.6). For the iCLIP and miCLIP methods used in this thesis, truncation events are optimised (Haberman et al., 2017).

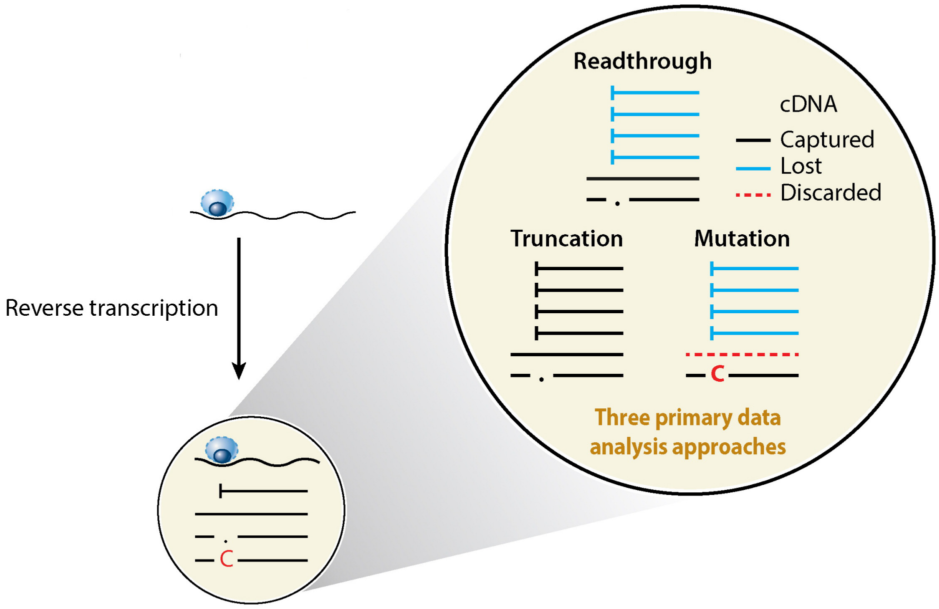


Figure 1.6: Possible reads produced from different CLIP approaches. Reverse transcription events resulting from a nucleotide-amino acid adduct are represented. iCLIP protocols maximise truncation events, whereas PAR-CLIP protocols encourage mutations at crosslink sites. Other protocols such as HITS-CLIP rely on analysis of readthrough cDNAs, where a short deletion may occur at the crosslink site. Adapted from (Chakrabarti et al., 2018).

cDNAs must then be purified, either by gel or beads-based approach to remove free RT primers or adapters, that could end up dominating sequencing libraries. Following the final adapter ligation step, a test PCR is performed to determine how many cycles are required for the final library amplification. A lower number of PCR cycles generally suggests a higher quality, more complex, library. Once individual samples are multiplexed the libraries are submitted for short read sequencing. Single-end sequencing is sufficient for most CLIP protocols, but paired-end can be required in protocols where the UMI/experimental barcode is split between 5' and 3' cDNA ends.

1.5.2 Analysis of (m)iCLIP data

Once the libraries are returned from sequencing the process of computational analysis begins. Reads are first demultiplexed into separate samples based on the experimental barcode, the 5' barcode is removed and the UMI is moved into the fastq header, using a tool such as Ultraplex (Wilkins, Capitanchik, et al., 2021). Individual fastq files are then run through FastQC to assess sequencing quality and adapter contamination (Andrews et al., 2010). Some analysis pipelines remove PCR duplicates at this stage based on read identity and UMIs (Flynn et al., 2015), however removing them later after mapping allows for more stringent removal.

The 3' Illumina sequencing adapter can be trimmed using cutadapt, or a wrapper such as Trim Galore! (Krueger, 2017; Martin, 2011). It's prudent to subsequently run the trimmed reads through FastQC again, to be certain that adapters have been completely removed.

Trimmed reads are then mapped to the reference genome or transcriptome depending on the protein to be analysed (Figure 1.7(1)). Where one is especially interested in RBPs binding to mature transcripts at positions that may cross exon-exon junctions, then transcriptome mapping can be more useful, but this also excludes the vast majority of peak callers from use. An optional step here is pre-mapping to an index comprised of tRNA and rRNA, to prevent them from contaminating the genomic mapping - this is discussed in more detail in Chapter 2. In the case of genome mapping, spliced alignment is important and so a splicing-aware aligner should be used such as STAR (Dobin et al., 2013). Also important is consideration in the mapping step of the type of crosslinking events that should be most abundant in the dataset at hand. Where iCLIP preserves information about the precise crosslink position at 5' cDNA start position for example, then it is important that soft-clipping of the 5' end of the read is disabled. Where there is an expectation of a high mutation rate, the inherent penalties of the aligner may need to be adjusted.

At this point, PCR duplicates can be removed by collapsing of reads that have identical UMIs and cDNA start positions, either through custom scripting or a software such as UMI-tools (T. Smith et al., 2017)(Figure 1.7(2)). Crosslinking positions from iCLIP are defined as the cDNA start position minus one nucleotide. Mutation-based protocols such as PAR-CLIP require mismatch calling, with careful SNP filtering, to identify crosslink sites (Corcoran et al., 2011).

Once crosslink sites have been extracted, peaks can be called depending on the desired downstream analysis, this is covered in great detail in Chapter 4 (Figure 1.7(3)). Following peak calling a large number of downstream analyses are possible depending on the research question. Commonly, researchers will want to know the binding motif of the RBP, whether for quality control of the dataset or because it is unknown. This is typically performed by an expectation maximisation algorithm such as STREME/DREME from the MEME software suite (Bailey, 2021). Alternatively, a tool such as PEKA can be used where every kmer present in the binding sites is given an enrichment score against kmer frequency in a background set derived from crosslinks not in peaks (Kuret et al., 2021). This is an especially useful approach to control for sequences that represent crosslinking biases, which should be present more frequently in the background set, it does however require strictly defined genomic annotations due to analysis in distinct genomic regions.

Integration of orthogonal functional data such as quantification of splicing or alternative polyadenylation events from RNA-Seq of RBP knockdown can be a powerful way to explore the position-dependent principles of RNA regulation by the RBP or RNA modification of interest (Rot et al., 2017).

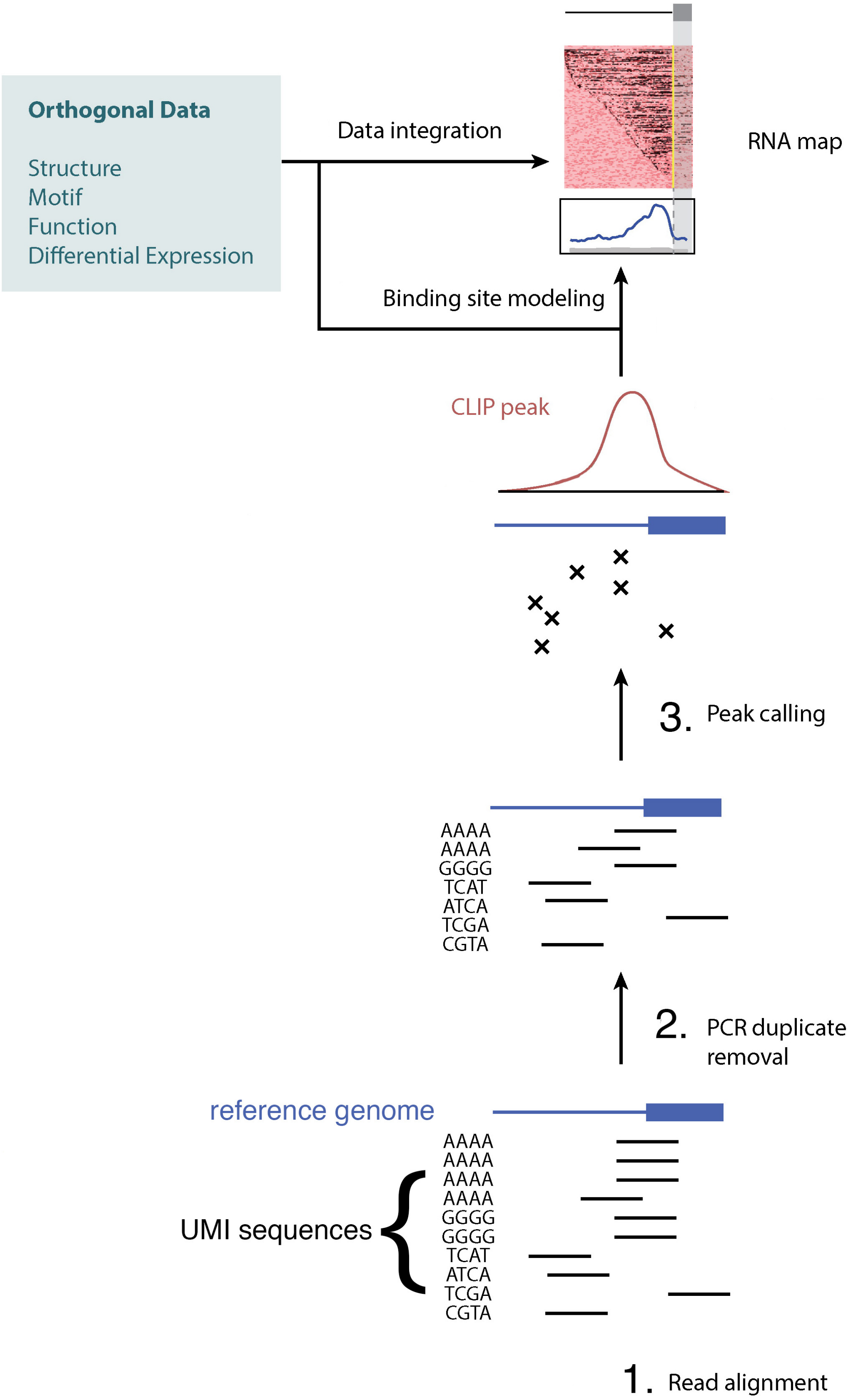


Figure 1.7: Overview of CLIP data processing. First reads are aligned to a reference genome or transcriptome, PCR duplicates are removed and crosslink sites are extracted (represented by x). Peaks can be called and taken for further downstream analysis including motif finding, or RNA maps where orthogonal functional data is integrated with CLIP peaks to identify positional principles of regulation. Adapted from (Chakrabarti et al., 2018).

1.6 Outline and Aims

The cellular functions of m6A are predominantly facilitated by RNA binding proteins which recognise the modification and recruit other machineries. miCLIP and iCLIP allow us to identify the sites of modification and RBP binding respectively, thus offering valuable insights into potential molecular mechanisms governing interactions between m6A and RBPs. When integrated with additional functional data, we are able to explore the functional relevance of these RNA modification dependent RBP-networks. Whilst crucial for our understanding of these processes, such analyses present complex computational challenges which are the focus of the current work.

In chapter two, I describe a novel, reproducible computational pipeline for analysing miCLIP and iCLIP data with a focus on the accurate quantification of repetitive non-coding RNA species. tRNAs, rRNA and snRNA are highly modified, but are commonly inappropriately quantified or ignored in CLIP analyses. This is due to their repetitive nature in the genome which makes assignment of short reads to specific genes difficult. I examine multiple options for resolving such reads and settle on a comprehensive pre-mapping approach guided by the logic of RNA abundance. Different categories of ncRNA are quantified and summarised in different ways informed by our prior biological knowledge. I focus especially on tRNAs, and describe an approach that uses the reads themselves to define clusters of tRNA isotypes that can be accurately resolved. Using this pipeline I am able to discern that RNA methylases NSun2 and Trmt2A bind extensively to pre-tRNAs. Furthermore, in collaboration with Dr. Aleksej Drino in the lab of Prof. Matthias Schaefer, I use my pipeline to characterise the previously unrecognised tRNA-binding of DEAH-box helicase DDX3X from publically available iCLIP datasets.

In chapter three, I collaborate with Dr. Radhika Varier and Dr. Dora Sideri in the lab of Prof. Folkert van Werven to investigate m6A-dependent RBP interactions in budding yeast meiosis. Yeast meiosis is a powerful model through which to understand the role of m6A and YTH proteins in cellular differentiation, due to the tight time window in which m6A is present in budding yeast and also the relative simplicity of the system, having one m6A writer and one YTH protein. We set out to use iCLIP and miCLIP methods to understand the function of m6A and potential readers in yeast meiosis, and therefore to elucidate general rules of m6A networks that could inform studies in other organisms. By careful analysis of RBP iCLIP in WT vs. methyltransferase deletion genetic backgrounds and integration of m6A miCLIP data alongside public m6A datasets, I confirm the m6A-dependence of budding yeast's only YTH protein—Pho92p. By means of such analysis, I am also able to reject the hypothesis that a high scoring hit in a mass spectrometry screen, Gis2p, binds RNA in an m6A-dependent manner. I discover a surprising amount of m6A-independent binding in our Pho92p datasets, which is located at the 3' end of transcripts in a similar manner to m6A-dependent binding, but is not dependent on the consensus m6A sequence motif. Further, by integration of publicly available ribosome profiling data and RNA-Seq datasets produced in the van Werven lab, I explore the potential function of Pho92p binding of m6A. I also make a surprising finding of Pho92p binding sites which increase in binding upon methyltransferase deletion, with a specificity for mRNAs encoding proteasomal components.

Finally, in chapter four I describe two valuable software tools that bring data visualisation to the forefront of CLIP analyses, with a focus on ease of use and interpretation for the user. The second of which, Clippy, I used extensively in the yeast meiosis project. Such data visualisation will be essential to the field at large moving forwards, as computational biologists and bench scientists alike aim to glean biological insights from complex bioinformatics datasets. I describe how the first tool, clipplotr, developed with Dr. Anob Chakrabarti, was born out of a need to visualise many CLIP datasets in one plot and the considerations one must make when condensing data into such a format. The resulting tool simplifies the process of producing such visualisations and has been widely used by bench scientists and bioinformaticians in our lab alike. The second tool, Clippy, developed with Dr. Marc Jones, is an interactive peak caller for CLIP data addressing several issues with peak calling of bioinformatics datasets; namely, the 'black box' nature of peak calling, that can make it hard for users to understand why some regions are called as peaks and others are not. Clippy shines light on the process of choosing parameters for your dataset by means of an interactive visualisation tool that is launched in the user's browser. We demonstrate that in addition to this novel feature, Clippy performs comparably or better than existing state of the art CLIP peak callers.

Taken together, I present a suite of computational tools that will improve the performance and accessibility of CLIP analysis, as well as describing the valuable biological insights that have already resulted from the application of such tools in a variety of scenarios pertaining to the understanding of RNA modification-dependent RBP networks.

Chapter 2

Software development for iCLIP and miCLIP analysis of structural and repetitive ncRNAs with multiple genomic copies

2.1 Introduction

2.1.1 The Problem

RNA modification is highly prevalent among rRNA and tRNA, where tRNA has the highest known proportion of modified nucleotides of any RNA species (Machnicka et al., 2013). snRNAs are also modified, and study of RBP-snRNA interactions are of interest to the research community, because of their importance to the process of pre-mRNA splicing. However, rRNA, tRNA and snRNA are highly repetitive in the genome, which makes the reads from these experiments difficult to accurately map. tRNAs for example, are infamously hard to resolve using short-read sequencing data for this reason: of 433 high confidence tRNA genes identified in GtRNAdb for humans, 172 are exact duplicate copies at another genomic location and the rest can be highly similar (Chan & Lowe, 2009). Consequently, in sequencing libraries, reads derived from these non-coding RNAs will often map to multiple places in the genome. This is a challenge for RNA-Seq experiments where the goal is quantification at the level of gene or transcript, but even more challenging for miCLIP where the goal is to identify exact sites of RNA modification and iCLIP where the goal is to identify protein binding sites. The high modification content of tRNAs also makes them difficult for reverse transcriptase enzymes to navigate. For example, modifications that cause steric hindrance to Watson-Crick base pairing can result in cDNA truncation, mutations or short deletions.

Further to wanting to quantify repetitive non-coding RNAs, we also do not want repeat-derived reads to contaminate results on mRNAs. For example, tRNAs and snRNAs can be encoded within introns of protein coding genes - without proper classification one can mistakenly attribute tRNA and snRNA reads to the protein coding genes themselves. This famously occurred in the case of m1A, where researchers originally misattributed hundreds of mRNA m1A sites. Many of these sites were called due to bioinformatics errors, like assigning a tRNA site to an mRNA, because it resided in the pre-mRNA intron (Schwartz, 2018). This is mostly a problem for tRNA and rRNA which are not (fully) included in the main genomic annotation files that are used.

Common approaches to dealing with multimapping in other bioinformatics domains are to randomly assign multimappers or to split the score of multimappers between all locations. However, considering the example of tRNAs, it becomes apparent that such techniques will be inaccurate. Where distinct groups of tRNAs are more similar than others, these specific groups will be penalised by such strategies, likewise regions of tRNAs that are more similar will have fewer counts by these approaches. In the realms of RNA-Seq and CLIP several computational tools provide more sophisticated solutions to help resolve and quantify such repetitive non-coding RNAs, however fall short of providing all the functionality one would like. In this chapter I pay specific attention to resolving as much tRNA information as possible, given that tRNAs are the most heavily modified RNAs in the cell. In the present section I will describe existing solutions, followed by a description and characterisation of a reproducible ncRNA-aware CLIP pipeline I developed and finally, the application of this pipeline to several relevant datasets.

2.1.2 Technical features of a good bioinformatics pipeline

Before embarking on the description of various pipelines I will address: what makes a good bioinformatics pipeline or software in general? The issue has been the subject of much discussion in the past decade, as attitudes and standards in bioinformatics software have shifted rapidly towards more reproducible, sustainable choices. Famously in 2009 a team of analysts aimed to reproduce the analysis of microarray datasets from 18 papers published in *Nature Genetics* (Ioannidis et al., 2009). At this time none of the processing code from any of the 18 papers had been published to a public code repository such as GitHub, meaning the analysts wrote code based on the written method descriptions. For over half the articles the analysis was not reproducible, half of these were due to unavailable data, but the other half was due to unavailable software, insufficient method descriptions or incorrect results reached when using the given method descriptions. It is notable that although the authors conclude that code should be made available as a standard, they highlight that:

> ”We should caution that even a detailed code would not make a novel and complicated analysis trivial to reproduce, nor would it totally eliminate the possibility for bias...It is not necessary for the code trail to be given in each minute detail, but important decision nodes should be described.”

This is a nuance that in practice is hard for journals to police - most journals in 2022 require software/code to be deposited in a public repository such as GitHub, but the extra detail in this assessment is that as well as being available, code needs to be interpretable and relatively easy to run.

One way this has been addressed, especially in the pre-processing of bioinformatics data, is with the renaissance of workflow languages such as common workflow language (CWL), workflow description language (WDL), Snakemake and

Nextflow (Wratten et al., 2021). For reproducible downstream analyses popular choices are code notebooks, such as Jupyter notebooks for Python and R notebooks for R. With these tools, code and resulting tables, graphs and commentary text are weaved together such that code is easily connected to key results. Code notebooks are such a useful format that journal eLife is developing Executable Research Articles inspired by them (Tsang & Maciocci, 2020). Workflow languages facilitate the handling of several important aspects of reproducibility:

- *Readability.* By laying out analysis as a series of steps with separated inputs, outputs and parameters, all workflow languages simplify the process of interpreting other people's code.

- *(Some) error checking.* Workflow languages will check that expected inputs and outputs are available or generated in the pipeline. For example, Snakemake does this by compiling a graph of processes before running a workflow, whereas Nextflow does this at run time of the individual processes.

- *System portability and scalability.* Many workflow languages include facilities to make a workflow portable to any computational environment, for example different operating systems, or to transfer from local execution, to high-performance cluster or cloud execution. This means the same code can be used to analyse one dataset, or millions of datasets.

- *Code portability.* Modularisation of the processes in workflows makes writing workflows even faster and easier. In this case, a module is written for one process, for example Bowtie read alignment, and then it can be re-used at multiple points in the same pipeline or in different pipelines. This prevents errors introduced by code duplication, which frequently occur when a change needs to be made in many places rather than just one place. Further, community repositories of modules, such as Nextflow's nf-core project, mean that the modules are tested by many more users, so they are more likely to be more robust to error.

- *Containerisation.* A painful part of reproducing any analysis involving a large number of software packages is installing the correct versions of the packages and also resolving any clashing dependencies between them. Containers are units of software that contain an entire runtime environment and any software packages installed by the author. In practice, Docker and Singularity are commonly used to create and manage containers (Yuen et al., 2021). The benefit of having a container for a pipeline is that the user no longer has to install any of the required software or resolve any software conflicts. Moreover, if the pipeline is run in a container then the author doesn't have to worry about accommodating differences in how operating systems work, for example MacOS vs. Linux, because code run within the container is running within the container's operating system. Containers become even more powerful when combined with modules, as having individual containers for individual modules means the author doesn't have to manage many clashing software packages.

- *Version control.* A common situation for a bioinformatician is going to reproduce an analysis from a paper and realising that the code in the public repository is now quite different from the time of publication. This is where version control is very important. GitHub is the main way to maintain version control as versions of your repositories can be published at any point in time and labelled - allowing, for example, for a version to be associated with a publication. Knowing the versions of software used in an analysis is also quite critical as newer releases of software frequently introduce breaking changes. While workflow languages do not enforce any version control as such, the communities around them are pushing for these standards. Nextflow nf-core workflows are all strictly version controlled. Most workflow languages offer ways to integrate with containers, which themselves are version controlled, or with conda - a package manager. The simplest way to specify software versions is with a conda environment specification written in YAML, this presents a step up from simply listing required software versions in documentation because an environment can be generated from the YAML specification directly.

2.1.3 Existing CLIP solutions

At the present time of writing there are three CLIP-specific solutions available for the problem of repetitive ncRNA mapping. Two, the eCLIP Family Aware pipeline (referred to as eFAP from now on) and FAST-iCLIP, represent end-to-end data processing pipelines for eCLIP and iCLIP data respectively (Van Nostrand, Pratt, Yee, et al., 2020; Zarnegar et al., 2016). On the other hand, CLAM (CLIP-seq Analysis of Multi-mapped reads) specifically takes multimapped BAM files as input and returns the user with a new BAM file with reassigned multimappers (Z. Zhang & Xing, 2017). CLAM uses an expectation maximisation approach to assign multimapping reads to the loci which have most unique read mappings in the surrounding area, which is a useful logic to bias towards RNAs that are most likely expressed. In practice the approach is very computationally demanding, and therefore only appropriate in the case of small datasets where multimapping represents only a small portion of the library, and therefore is not useful in the case of RNA modification data containing a high percentage of tRNA and rRNA. A useful characteristic however, is that crosslinks can be subsequently derived from the processed BAM, therefore CLAM is agnostic to the library preparation protocol.

eFAP is written in common workflow language, a workflow management language based on JSON and YAML. The pipeline consists of two steps: first, Bowtie2 is used to map paired-end CLIP reads to a manually curated set of repeat elements from RepBase, Ensembl annotations, GtRNAdb and rRNA from NCBI. The reads mapped to this custom database are resolved to the level of repeat families - any read mapping between two repeat families is not taken further for analysis. Second, the repeat family resolved reads and uniquely mapping genomic reads are merged such that the best mapping for each read is chosen. Such a strategy is sensible and clearly retrieves a lot of information lost by standard mapping approaches (Van Nostrand, Pratt, Yee, et al., 2020). The main issues for my analysis are: production of the ncRNA database is not fully described, the treatment of tRNAs is not ideal and the code-base is not very accessible to an outsider. tRNA sequences from both RepBase and GtRNADB are mapped to together; as the RepBase sequences are usually fragments of tRNA pseudogenes it is likely that reads could multimap between the genuine tRNA and the RepBase fragments, which would not be resolved by the pipeline. Whilst the pipeline is technically written in CWL, it is mostly calling custom perl scripts that can run up to 1795 lines long with few comments (in the case of "duplicate_removal_inline_paired.count_region_other_reads_masksnRNAs_andreparse SE-andPE_20201210_simple.pl"). Therefore, to edit or adjust the analysis to suit a specific purpose presents a major challenge to the outsider. Further, currently only the human genome is supported.

FAST-iCLIP is a pipeline originally written to accompany the Fully Automated and Standardized iCLIP (FAST-iCLIP) protocol (Flynn et al., 2015), with a focus on including viral sequences in mapping to enable study of host-pathogen RBP-RNA interactions. The code was subsequently updated on the release of the infrared-CLIP (irCLIP protocol), mostly to improve the speed (Zarnegar et al., 2016). It is predominantly written in python, but does not use a workflow language, so the steps in the pipeline are not modular, and thus require considerable work to edit or modify parameters. Similarly, debugging the pipeline is frustrating because the entire pipeline must be run from scratch each time to debug, whereas popular workflow managers auto-detect steps that have already been successful.

Mouse and human prepared annotation databases are provided, but there are no instructions or facilities for creating such databases for other organisms. This is a problem because the sequences included are undocumented so it is near impossible to know their origin. The GitHub repository states "We will release details of generating annotation files for other genomes shortly in future.", but the last update was in 2017. At the present time of writing Bowtie2 is used for all mapping steps in default mode - where a random alignment is chosen in the case of multiple best alignments (Langmead & Salzberg, 2012). This can create randomness in some quantifications

- if a read maps between different repeat families it becomes random whether its count is attributed to one or the other, which is undesirable. Additionally Bowtie2 has been superseded by more modern aligners when it comes to mapping spliced reads, which in the case of RBPs binding mature mRNA transcripts, we might expect quite a few. The TopHat, HISAT and HISAT2 aligners makes use of Bowtie2 in their underlying algorithms, but STAR has been found to outperform them (Dobin et al., 2013; Kim et al., 2015). Therefore, for the genome alignment step STAR should be used - notes in the GitHub repository suggest this was planned but not implemented.

Further, PCR duplicates are removed before mapping, which is a valid strategy, but in this instance only accounts for duplicates which are exact sequence matches - in other CLIP pipelines PCR duplicates are removed after mapping based on start position and UMI - which means reads that contains mismatches to each other can still be removed as PCR duplicates. This allows for the capture of those duplicates where, for example, one read may have acquired a point mutation during Illumina sequencing.

To summarise, in its current iteration FAST-iCLIP is now dated, and not optimal to use for the purpose of general ncRNA analysis in CLIP. The mapping order of reads to indexes: exogenous virus sequences → repeat sequences → endovirus sequences → tRNA → genome, is inflexible and does not make sense for the majority of available CLIP data, but will be useful for those specifically studying RBPs in the context of host-virus interactions.

2.1.4 Existing RNA-Seq solutions to quantifying tRNAs

Arguably more efforts have been made towards improving quantification of tRNAs from an abundance (RNA-Seq) standpoint, than from an RBP binding (CLIP) standpoint. Innovations of both the experimental and computational variety have improved tRNA abundance estimates dramatically. From an experimental standpoint, efforts to remove RNA modifications from tRNAs can improve the processivity of first-strand synthesis, and therefore resulting sequencing libraries. In 2015,

DM-tRNA-seq introduced the use of a combination of wildtype *E. coli* AlkB and a D135S mutant, to remove m1A, m3C and m1G respectively (Zheng et al., 2015). The second innovation was to use thermostable group II intron reverse transcriptase (TGIRT; InGex) with high processivity to accommodate passage through both remaining modifications and stable tRNA structures. Subsequently, **Hydro-tRNA-seq** took a different approach, using alkaline hydrolysis to produce shorter RNA fragments (Gogakos et al., 2017). This results in the disruption of structures and opening up the more 5' regions of tRNAs for sequencing, that may have previously been "blocked" by modifications in the body of tRNAs which could interrupt first-strand synthesis primed from only the 3' end of the tRNA molecule. The downside is that shorter reads are harder to resolve at the computational level. A further invention in **YAMAT-Seq** involves the use of a hybridised Y-shape adapter which is ligated to tRNA molecules by bacteriophage T4 RNA Ligase 2 (Rnl2) - this takes advantage of the base pairing between 5' and 3' tRNA ends in full length mature tRNAs (Shigematsu et al., 2017). In 2020, **QuantM-tRNAseq** introduced a further variation, using a splint adapter ligation strategy to the 3' terminal CCA of mature tRNAs and processive SuperScript IV reverse transcriptase. Importantly the authors compared the tRNA isotype abundances quantified with QuantM-tRNAseq against all previous methods and found that it most faithfully replicated abundances found via tRNA microarray (Pinkard et al., 2020). Most recently, in 2021, **mim-tRNAseq** presented a further improved approach, from both an experimental and bioinformatics perspective (Behrens et al., 2021). On the experimental side, the authors returned to TGIRT, which was previously abandoned due to low priming efficiency and bias resulting from its template switching mechanism. To address these issues, DNA adapters were added to tRNA 5' and 3' ends and reverse transcription was performed at a lower temperature in a low salt buffer for a longer time period. Under these conditions TGIRT had increased processivity and could read-through modifications with low fidelity, resulting in consistent mutations at these sites which could be later analysed.

From a computational angle, mim-tRNAseq maps reads to mature tRNA transcript sequences using GSNAP aligner (Behrens et al., 2021). GSNAP is provided with data of tRNA modifications downloaded from the MODOMICS database, such that positions of tRNA modification are not counted by the alignment algorithm as mismatches. A first round of mismatch-lenient mapping is made to add any frequently mismatched bases to the modification index, ready for a stringent second round of mapping. Mature tRNA sequences were clustered using usearch at the isotype level, although sequences within an isotype group could have multiple clusters if their identity was less than 95%. The final mapping is performed to the representative centroid cluster sequences. Reads are then further deconvolved if they contain a base that is only present in one sequence of the cluster, otherwise they remain assigned to the centroid sequence.

Previously, MINTmap was the computational standard for tRNA-seq mapping, especially suited to the mapping of tRNA fragments (Loher et al., 2017). With this approach, every possible tRNA fragment is annotated with all possible originating tRNAs; reads are then assigned to these fragments. The downside to the approach is that it does not perform more higher level summaries (e.g. at isotype level), so this is left to the user to resolve if desired.

More recently, another mapping method involving clustering of tRNA sequences was proposed. Consensus sequences were produced by collapsing non-redundant tRNA sequences at an arbitrary Levenshtein distance threshold and uniquely mapping reads were quantified to the consensus reference sequences (Pichot et al., 2021). The difficulty with this method is setting the distance threshold, although the authors describe ways to go about this it requires some optimisation.

2.2 Structural and repetitive ncRNA-aware CLIP pipeline development

2.2.1 iCLIP testing datasets

In order to develop and characterise my pipeline, I tested it with iCLIP (or iCLIP-like) data for 8 different proteins. I chose two datasets where I expected high tRNA enrichment: miCLIP of NSun2, an m5C methyltransferase targeting the majority of tRNAs at positions C48, C49 in the variable loop junction (Blanco et al., 2014; Hussain et al., 2013) and miCLIP of Trmt2A, which catalyses methylation of tRNA U54 (Carter et al., 2019). As described in the introduction, the miCLIP in this context refers to iCLIP of NSun2/Trmt2A with a mutation that stabilises a covalent bond between the protein and RNA during the methylation reaction. For this reason, UV-C irradiation is not required in these library preparations.

I chose two datasets specifically for their enrichment of snRNAs, namely spliceosome iCLIP (SmB iCLIP) under mild washing conditions (Briese et al., 2019) and Prpf8 iCLIP (as Prpf8 is a protein at the spliceosomal core, it makes many contacts with snRNAs (Blazquez et al., 2018)). I then chose hnRNPC iCLIP, because I expected to see enrichment of Alu elements (Zarnack et al., 2013; Zarnegar et al., 2016). Finally, I chose three proteins which have good iCLIP data and that to my knowledge have not been exhaustively analysed in this way before: Ptbp1, Tia1 and Tdp-43 (Haberman et al., 2017; Rot et al., 2017; Z. Wang et al., 2010). I expected that these three proteins would have little enrichment for repetitive non-coding RNAs, serving predominantly as a negative control set. This is because described functions of these proteins are mostly related to pre-mRNA processing, all three are involved in pre-mRNA splicing, and Tia1 and Tdp-43 are involved in alternative polyadenylation decisions (Linares et al., 2015; Rot et al., 2017; Tollervey et al., 2011; Z. Wang et al., 2010). That being said, I expected Ptbp1 and Tia1 iCLIP data to have enrichment of some transposable elements, as they have previously been reported to bind LINEs (Attig et al., 2018).

2.2.2 Pre-mapping or post-hoc assignment of genomic multimappers?

To assign reads mapping to repetitive non-coding RNAs, all reads could be first pre-mapped to different categories of ncRNAs before being mapped to the genome. Alternatively, all reads could be mapped to the genome allowing many multimappers and then assigned post-hoc to repetitive ncRNA families based on available annotations (Figure 2.1A). To help decide between the two strategies I checked whether pre-mapping or genomic mapping captured the most tRNA and snRNA-derived reads in the chosen CLIP datasets (Figure 2.1B,C). For this test I used Bowtie for pre-mapping and STAR for genomic mapping (Dobin et al., 2013; Langmead, 2010).

I assume for known tRNA and snRNA binders that if a read can be assigned to tRNA or snRNA respectively then it should be, as this likely reflects the biological reality. In general there are three possible sources of reads being assigned as tRNA/snRNA: 1) reads that represent genuine RBP binding to tRNA/snRNA, 2) reads that represent contamination in the CLIP library of unbound tRNA/snRNA sequences or 3) reads that have sequences similar to tRNA/snRNA, but are actually derived from another RNA type, that have been bioinformatically mis-assigned as tRNA/snRNA - these could be genuinely bound by the RBP or not.

I found that the pre-mapping strategy recovered more snRNA and tRNA reads for the known snRNA and tRNA binders than the genome strategy. This difference could be quite substantial, with pre-mapping increasing the percentage of reads mapping to tRNA/snRNA by as much as 20% in the case of NSun2 tRNA and 15% in the case of Prpf8 snRNA (Figure 2.1B,C). This likely has consequences for downstream quantification of different tRNA/snRNA types. The reads lost by the genome strategy will be those that map equally well to other positions in the genome, which are likely to be unannotated copies or fragments of tRNA/snRNA. These unannotated regions will likely be biased towards certain groups of tRNA/snRNA.

I also checked the assignment of tRNA/snRNA for a group of proteins unknown to be tRNA/snRNA binders as a form of negative control. Generally, the pre-mapping strategy did assign more reads as tRNA/snRNA. I propose that this is mostly representing the reality of where these reads came from for several reasons: 1) Generally, the unknown snRNA binders with the most pre-mapped snRNA signal, Tdp-43 and Tia1, also have higher signal from the genomic strategy, similar for the unknown tRNA binders. 2) The NSun2 replicate 1 library is of poorer quality to NSun2 replicate 2 - it contains only 250,954 reads compared to the second replicate 11,527,392, so where the first replicate contains many more assigned snRNA reads than the second, I propose this is from experimental contamination of the data.

In the FAST-iCLIP pipeline the pre-mapping strategy is used. In the eFAP pipeline, the best mapping is chosen - whether it is genomic or pre-mapped, and where there is a tie the pre-mapped is preferred. In the end it partially becomes a theoretical argument - if a read maps to tRNA with two mismatches, and equally well or better to somewhere else in the genome - which mapping is more reliable? I would argue that tRNA is highly abundant within cells and that it is more likely such a read is genuinely tRNA-derived. Therein lies the logic of hierarchical pre-mapping, where the order of mapping steps is based on decreasing RNA abundance.

2.2.3 Choice of alignment algorithm

The choice of alignment algorithm and settings when aligning to non-coding RNAs is important, because they are more modified than mRNAs one should allow for more mutations - however there is a balance between allowing mutations and being too lenient. In the previous test I used Bowtie with an absolute maximum of two mismatches allowed. In the mim-tRNAseq manuscript Bowtie is tested against Bowtie2 and GSNAP for tRNA alignment (Behrens et al., 2021). There are some differences between the use cases of CLIP and mim-tRNAseq - with mim-tRNAseq tRNA modifications are also inputted to GSNAP such that mutations at these positions are not considered to be mismatches; this makes sense because of their optimised reverse transcription that promotes mutation at these positions. Reads in mim-tRNAseq are often longer than CLIP reads because of this, and frequently span

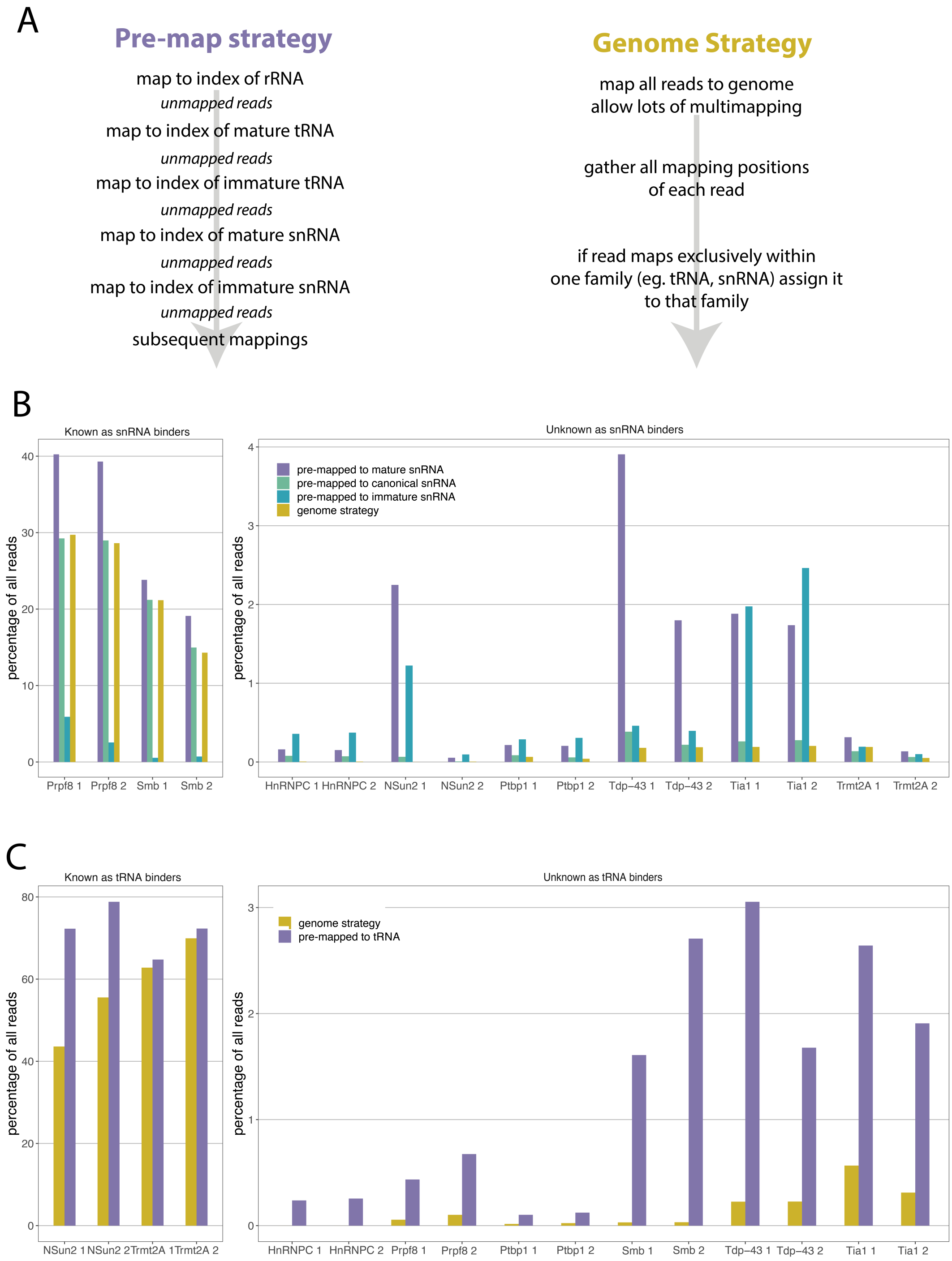


Figure 2.1: Comparison of pre-mapping and genomic mapping strategies. A) Schematic describing pre-mapping vs. genomic mapping strategies. B) Percentage of all reads mapping to snRNAs by pre-mapping or genomic mapping strategies for known snRNA binders and unknown snRNA binders. C) Percentage of all reads mapping to tRNAs by pre-mapping or genomic mapping strategies for known tRNA binders and unknown tRNA binders.

entire tRNAs (∼75nt). In iCLIP data I expect the mutation profile at RNA modifications to be less consistent so I opted for comparison of Bowtie vs. Bowtie2. The advantage of Bowtie is that you can set a maximum number of mismatches, so you have relatively fine grained control over what is considered a suitable mapping, it may also perform better for reads <50nt in length. In Bowtie2 you have much less control, but the advantage of short-gapped alignment and improved indel detection.

In the mim-tRNAseq manuscript the authors found a big improvement on tRNA mapping for Bowtie2 vs. Bowtie, however I find little difference between the two in the case of iCLIP data when I compared the tools for mapping to a mature tRNA index over multiple iCLIP datasets (Figure 2.2). For iCLIP in fact, Bowtie aligns more tRNA reads. Interestingly, this effect is more apparent for certain libraries over others, and it would be good in future to assess what qualities of the libraries lead to this effect. It could potentially be due to more sensitive mapping of shorter reads with Bowtie.

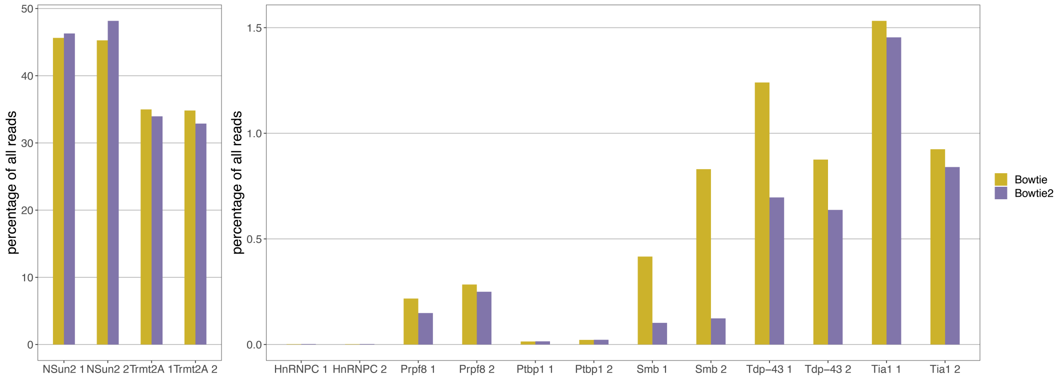


Figure 2.2: Comparison of Bowtie vs. Bowtie2 for iCLIP tRNA read assignment. iCLIP reads were mapped to a mature tRNA index using either Bowtie (yellow) or Bowtie2 (purple).

2.2.4 snRNA

Like tRNA, there are many annotated snRNA genes in the genome representing different sub-families (Table 2.1). U1, U2, U4, U5, U6, U4atac, U6atac, U11 and U12 are all spliceosomal RNAs which bind to protein partners to form small nuclear ribonucleoproteins (snRNPs) that participate in the removal of pre-mRNA introns as part of the major and/or minor spliceosomes. U7 snRNA is a short distant relative to the spliceosomal snRNAs, which participates in replication-dependent histone mRNA 3' end processing (Dominski & Marzluff, 1999). What about U3, U8 and U13? The"U" RNAs were so named because these RNAs were found to be uridine-rich when compared to rRNA or mRNAs (Hodnett & Busch, 1968; Reddy & Busch, 1988), later U3, U8 and U13 were distinguished as snoRNAs due to their localisation in the nucleolus, which is due to their essential role in rRNA processing (Reddy et al., 1981; Reddy & Busch, 1988). Finally, 7SK snRNA forms a snRNP which facilitates RNAPII-mediated pre-mRNA elongation, and also acts as a transcription factor to promotes expression of many snRNA genes (Egloff et al., 2017).

snRNA	Number of genes	Number proposed to be functional
U1	145	8
U2	90	3
U4	97	2
U5	31	5
U6	1310	7
U7	156	NA
U4atac	19	3
U6atac	43	1
U11	6	1
U12	12	1
7SK	1	1

Table 2.1: Human snRNA genes annotated in Gencode v39. Number of proposed functional genes taken from (Marz et al., 2008).

snRNA annotations represent many variants and copies, a minority are likely to be functioning snRNA genes, whilst many more represent pseudogenes resulting from gene duplication followed by mutation, or by some form of reverse transcription and integration. An analysis of spliceosomal snRNA promoter structures suggested the number of human spliceosomal snRNA genes which might be functional (Table 2.1, (Marz et al., 2008)). A portion of the U6 pseudogenes are chimeric with L1 long interspersed nuclear element (LINE) sequences which reflects their origin from L1-mediated retrotransposition events (Buzdin et al., 2002). There is some evidence that variants of U1 snRNA are expressed, especially in stem cells, where they might promote pluripotency (O'Reilly et al., 2013; Vazquez-Arango et al., 2016). U5 snRNA variants have been shown to form variant spliceosomes, where five U5 snRNA variants (A,B,D,E,F) make up the majority of U5 snRNPs (Mabin et al., 2021; Sontheimer & Steitz, 1992).

Alongside pre-mapping to all snRNA sequences, and post-genomic mapping assignment of multimappers, another option would be pre-mapping to canonical snRNA sequences. In the case of spliceosomal snRNAs for example, this might help to prioritise functional sequences and also more precisely map the positions of crosslinks on the RNA, because multimapping at different positions on slightly different snRNA copies does not have to be resolved. In my testing I found that mapping to canonical snRNA sequences reduced the amount of reads assigned to snRNAs in comparison to pre-mapping to all snRNA sequences, in general leading to a similar number of assigned reads as the genome strategy (Figure 2.1B). This suggests that the extra reads assigned by mapping to all mature snRNAs come from divergent sequences that are more likely to be represented elsewhere in the genome, unannotated as snRNA sequences. For this reason I chose to use mapping to all snRNA sequences to assign their abundance, but complemented this with mapping to canonical snRNA sequences separately to give positional information.

Furthermore, I also investigated the assignment of snRNA reads to their gene families. In both Prpf8 and SmB iCLIP libraries the majority of mapped cDNAs could be uniquely attributed to one snRNA gene family, despite the fact that these cDNAs might map to many individual genes within these families (Figure 2.3A). These designations make sense, because Prpf8 is closely associated with U5 snRNA, while U1 snRNA is the most abundantly expressed in the cell so is expected to dominate the SmB data. I checked to see if there was a bias towards multimapping between certain snRNA gene familes, and found that the U6 category was most likely to be mapped to by cDNAs which also map to other families (Figure 2.3B). U6 has the largest number of genes/pseudogenes and I propose that its possible that within this sequence space there are genes that contains sequences resembling other snRNAs, although this would require further closer investigation. I also noticed that the amount of multimapping cDNAs between families was more for the SmB libraries (as high as 10% of snRNA mapped cDNAs). I hypothesised this could be due to shorter mapped cDNA lengths, and indeed I did find that the SmB data had a peak of mapped cDNA length around 12 nt which was absent in the Prpf8 data, which could be contributing to this increased multimapping (Figure 2.3C). Other factors are likely important, such as where within the snRNAs the reads are derived from, as some areas will be more unique than others and the Sm binding site is likely more degenerate than Prpf8 snRNA contacts.

2.2.5 Resolving tRNAs at higher resolution

tRNAs can be divided into categories based on the amino acid that is loaded, the anticodon that is decoded (isotype) and by the rest of the sequence (isodecoder). Even at the isodecoder level, tRNAs which have exactly the same sequence might be transcribed from many genomic loci (Figure 2.4).

The effect on short read mapping is that a given tRNA-derived read will often map equally well to different tRNA genes, isodecoders, isoacceptors and less frequently even to tRNAs in different amino acid categories. This makes quantification troublesome as we would like to assign each read as a single count to a single gene. My solution is to group tRNAs at some level of hierarchy and summarise counts within those groups, however the question is then, how should these groups be defined?

In all available CLIP pipelines (eFAP and FAST-iCLIP) groupings are made at the level of isoacceptors and reads that are ambiguous at the isoacceptor level, i.e. that map between two or more tRNA isoacceptors, are discarded. However I observed that this could be a large proportion of reads. For example, in the case of NSun2 miCLIP, whilst 99% reads mapped either singly or multiply within amino acid tRNA families, around 20% multimapped between isoacceptor groups (Table 2.2). I hypothesised that multimapping between isoacceptor groups will be biased to certain tRNA groups that are more similar in sequence, therefore by discarding

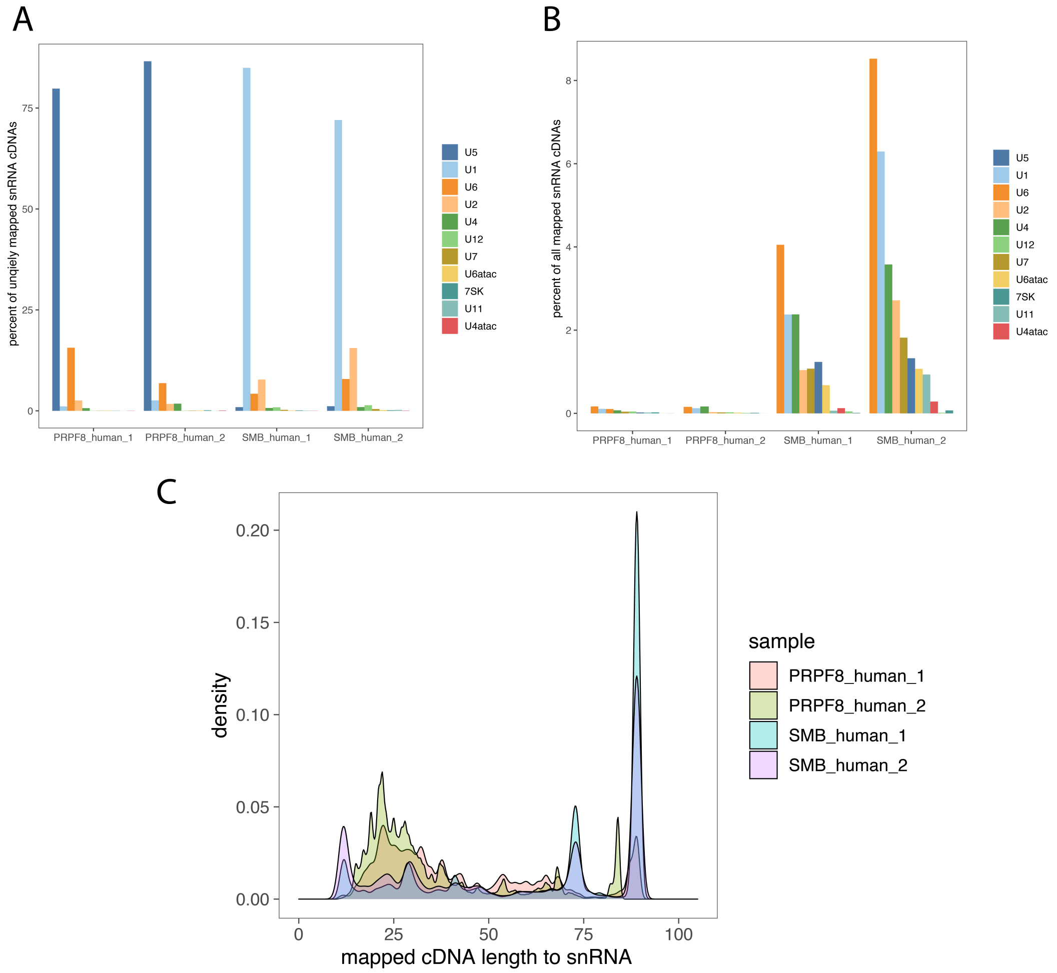


Figure 2.3: snRNA cDNA assignment to snRNA gene families. A) cDNA assignments to different snRNA gene families using only those cDNAs that uniquely mapped to a gene family. Presented as percentage of all snRNA mapped cDNAs. B) Gene families that were most frequently the source of multimapping between families, here the *y* axis represents number of multimapped alignments made to the gene family divided by the total number of mapped snRNA cDNAs for the indicated sample. C) Density plot of mapped cDNA alignments to snRNA for the indicated samples.

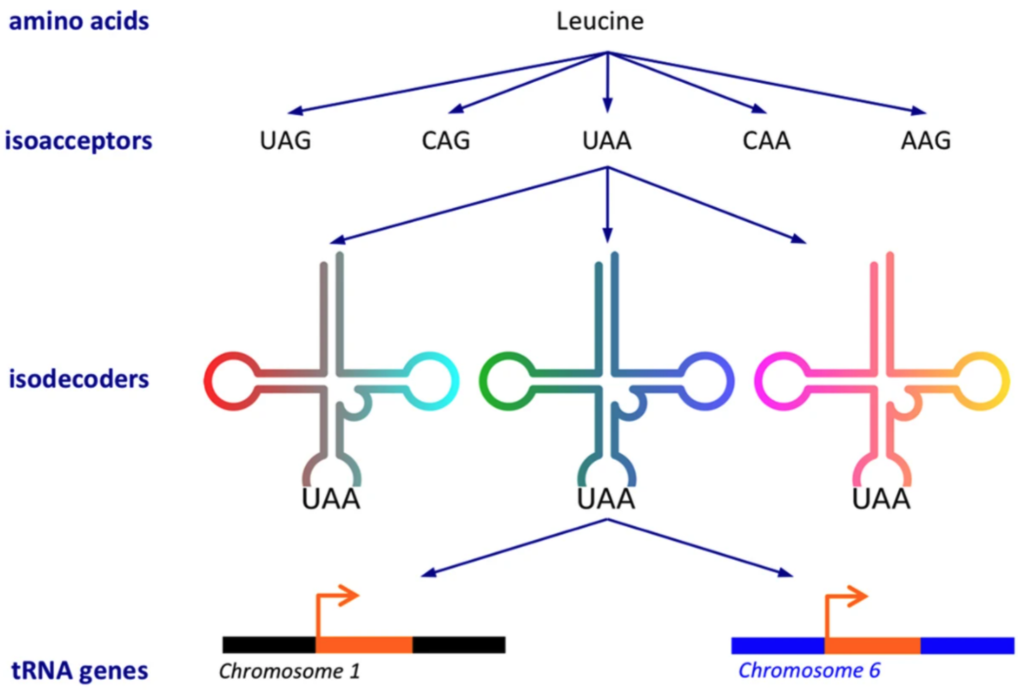


Figure 2.4: Hierarchy of tRNA groupings adapted from (Loher et al., 2017).

reads that multimap between isotypes one could be biasing the count data against certain tRNA isoacceptor groups. By calculating the Levenshtein distances between consensus sequences from all human tRNA isoacceptor groups it is clear to see that some have very few nucleotide differences, whilst others have many (Figure 2.5). For example, one might expect reads to commonly multimap between Leu$^{\mathrm{TAG}}$ and Leu$^{\mathrm{AAG}}$ because on average genes in these groups have only one nucleotide difference. One could create groupings based simply on a distance cutoff from this heatmap, however this will cause a loss of resolution in certain data unnecessarily. Consider the isoacceptors Gly$^{\mathrm{GCC}}$ and Gly$^{\mathrm{CCC}}$: on average genes from these groups will have four nucleotides different - if these nucleotides are close together such that they are all contained within a 40nt CLIP read, then the CLIP data is able to discriminate between the two groups. On the other hand, if these differentiating nucleotides are not in a region where the protein is bound, then the CLIP data would not be able to discriminate between the two groups. Therefore, I have opted to let the data itself dictate the resolution at which the quantification is made.

To address the issue I wrote python code that would merge isotype groups

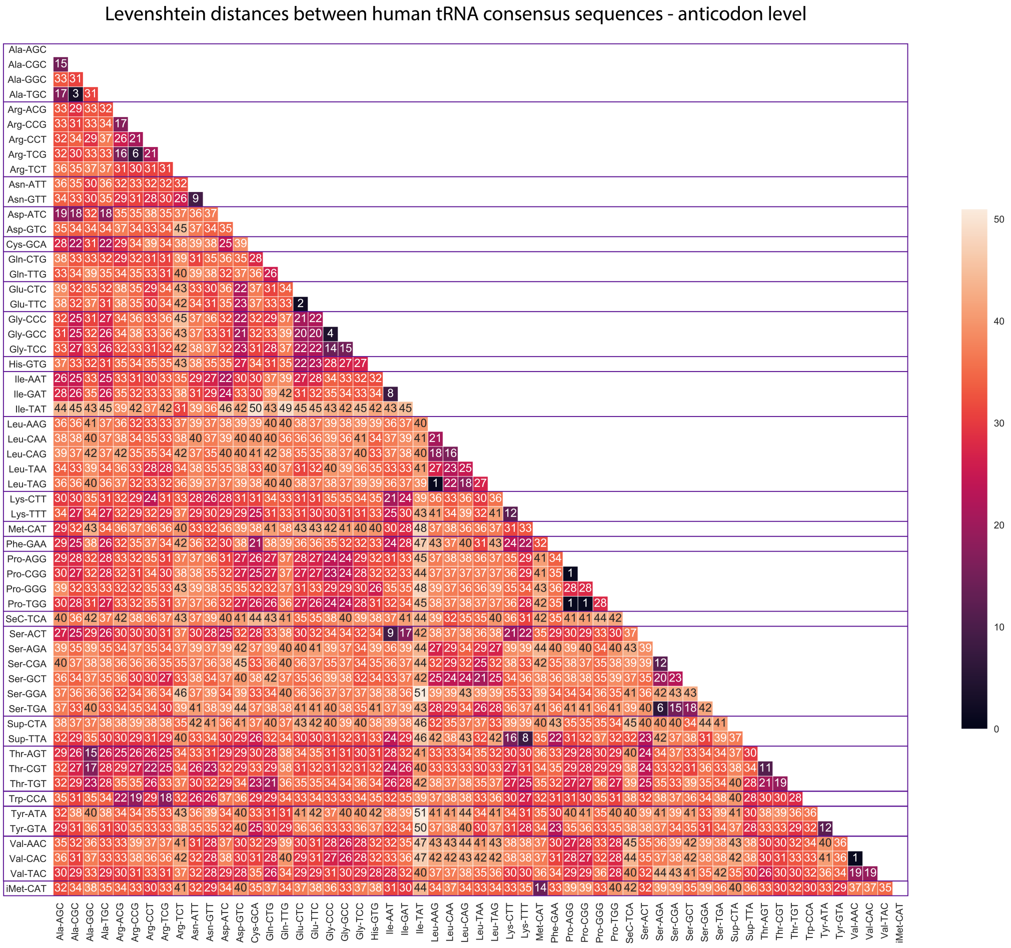


Figure 2.5: Levenshtein distances between human tRNA isotype consensus sequences.

where reads were found to be frequently multimapping. To decide how many groups should be made, I took a data-focused approach. The user sets a "fraction merged" value such that isotype groups are merged until the fraction of reads that would be lost to multimapping, are in fact recovered. In the present analysis I used a value of 0.9 so that I required 90% of multimapped isotype reads to be recovered. In my analysis of test data, I found that this threshold made sensible groupings, whilst also minimising the diminishing returns of making larger tRNA groupings for the sake of recovering fewer and fewer reads. The details of the tRNA mapping are automatically generated and output in a summary statistics table, so that the user of the pipeline can assess the assignment of reads easily, as shown for analysis of NSun2 data (Table 2.2).

By examining test m1A miCLIP produced in our lab by Dr. Paulo Gameiro, I found that discarding reads that multimap between tRNA isotypes (as in FAST-iCLIP and eFAP pipelines) can be misleading. For example for this particular miCLIP data you can see that by discarding reads that multimap between Ser-TGA and Ser-AGA you might be misled into thinking that position 67 is not modified in Ser-AGA when in fact this is ambiguous because more reads multimap to that locus between Ser-TGA and Ser-AGA than uniquely map to either isotype (Figure 2.6A). The situation is similar between Leu-CAG and Leu-CAA (Figure 2.6B).

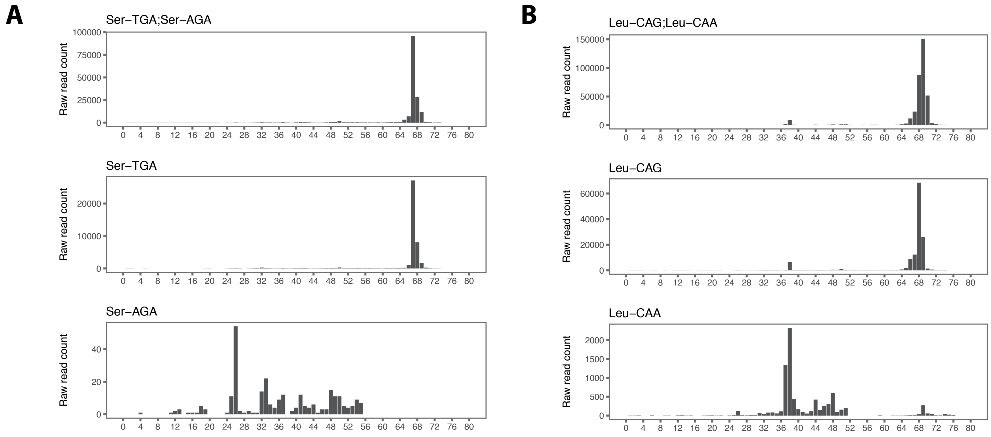


Figure 2.6: Misleading tRNA counts when discarding reads that multimap between isotypes. A) m1A miCLIP read count data summarised per position along tRNA transcript. Top panel shows the profile when isotypes Ser-TGA and Ser-AGA are combined, bottom two panels show the results of treating the isotypes as separate. B) Same as A, but for Leu-CAG and Leu-CAA isotypes.

In contrast to RNA-Seq, CLIP data also contains positional information, therefore it is important to resolve the cDNA start positions. In the eFAP and FAST-iCLIP pipelines the position of a multimapping tRNA read is set based on the mapping to the longest tRNA sequence. In my pipeline I have opted to take the mode of all mapped positions, and where this is not possible a position is randomly chosen based on a seed value. I also report the median and maximum distance discrepancies between these ambiguous positions in the summary report, so that the user is alerted if this difference is very large (Table 2.2).

Metric	**Value**	**Sample name**
Total mapped reads to trna	28852	NSUN2_HEK293T
Total single mapping reads to trna	12980	NSUN2_HEK293T
Number of reads that are ambiguous at amino acid level	292	NSUN2_HEK293T
Percentage of reads used for amino acid summary	0.99	NSUN2_HEK293T
Number of reads that are ambiguous at anticodon level	6192	NSUN2_HEK293T
Percentage of reads used for anticodon summary	0.79	NSUN2_HEK293T
The actual fraction of ambiguous reads that are recovered by merging anticodons	0.92	
The anticodons that were merged are	Leu-CAA;Leu-CAG—Ser-AGA;Ser-TGA—Glu-CTC;Glu-TTC—Pro-AGG;Pro-CGG;Pro-TGG—Thr-AGT;Thr-CGT—Val-AAC;Val-CAC—Leu-AAG;Leu-TAG—Gly-CCC;Gly-GCC	NSUN2_HEK293T
Number of reads that have an ambiguous position at anticodon level	17	NSUN2_HEK293T
Median distance between ambiguous positions at anticodon level	2	NSUN2_HEK293T
Biggest distance between ambiguous positions at anticodon level	2	NSUN2_HEK293T

Table 2.2: Contents of ”tRNA_summary_stats.tsv” output.

2.2.6 Immature sequences

To capture RBP binding to nascent snRNA and tRNA transcripts I also included an "immature" category which is mapped to after the "mature" category. For snRNA genes I included a 50nt flank upstream and downstream. In the case of tRNA genes some have introns, which are included in the "immature" sequence set, in addition to the 50nt flanking regions. Mature tRNA sequences also have an additional 5' G for histidine tRNAs, and a 3' CCA added, where the "A" becomes the site of amino acid attachment (Loher et al., 2017). To my knowledge, both FAST-iCLIP and eFAP pipelines include immature and mature sequences in their mapping indexes, but do not separate the two in their summaries.

2.2.7 Mitochondrial chromosome

I chose to map to the mitochondrial chromosome before the nuclear genome because much mitochondrial sequence is present in the human nuclear genome, with at least 612 independent integrations (Woischnik & Moraes, 2002). It was found that the human genome contains exact copies of several mitochondrial tRNAs (Telonis et al., 2014). I reasoned that if an RBP binds to mitochondrial RNAs, then these reads can be uniquely mapped if they are first mapped to the mitochondrial chromosome before the nuclear genome, otherwise they could end up multimapped to the nuclear genome in a way that is hard to resolve. Unfortunately it was not possible to test this very well: because of the limitations of the datasets there was proportionately little mitochondrial RNA mapping. I expected there might be some in the NSun2 data, as NSun2 methylates both nuclear and mitochondrial-encoded tRNAs(Van Haute et al., 2019). In the second replicate there were some mitochondrial mapped reads, mostly to tRNA, which did increase when I mapped to the mitochondrial genome before the nuclear genome (Figure 2.7).

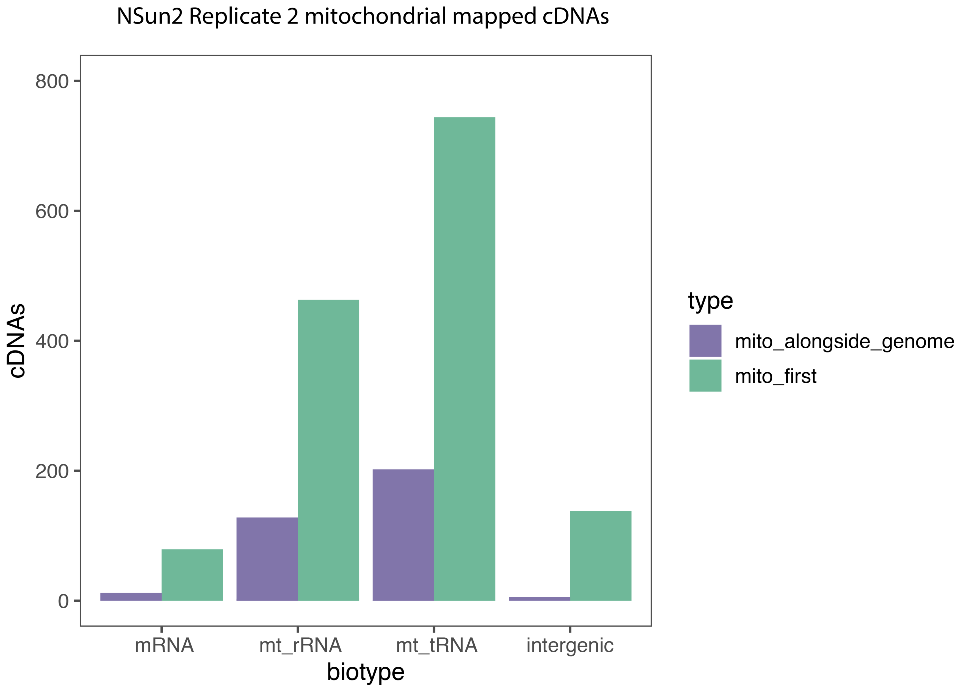


Figure 2.7: NSun2 mitochondrial mapped cDNAs. Green represents mapping to the mitochondrial chromosome first, before mapping unmapped reads to the nuclear genome, whereas purple represents mapping to both sequences at the same time.

2.2.8 Repeats

Repeat elements, composed mostly of transposable elements, present an important binding partner for some RBPs. Short interspersed nuclear elements (SINEs) represent ∼13% of the human genome, in primates these are mostly comprised of Alu elements which are sequences derived from the 7SL signal recognition RNA (Lander et al., 2001). HnRNPC binds to Alu elements within some coding gene introns to prevent the deleterious exonisation of these Alu elements (Zarnack et al., 2013). Long interspersed nuclear elements (LINEs), representing ∼21% of the human genome (Lander et al., 2001), can also be bound by RBPs. For example, MATR3 and PTBP1/2, alongside other repressive RBPs, prevent inclusion of LINE-derived exons that are located within long introns (Attig et al., 2018). Such repression of these elements allows them to persist in genomes without negative impact on the host, giving the sequences an opportunity to evolve into functional exons (Attig & Ule, 2019).

Repeat elements are curated and annotated in Repbase, a database running since the early 1990's (Bao et al., 2015; Kojima, 2018). Transposable elements are classified at the highest level as either DNA transposons, LTR retrotransposons (such as endogenous retroviruses ERV1, etc.) or Non-LTR retrotransposons (LINEs and SINEs). The classification becomes more detailed at the level of superfamilys/-clades. In practice the classes present in the latest Repbase annotation (at time of writing) are: DNA, LINE, Low_complexity,LTR, RC, Retroposon, RNA, rRNA, Satellite, scRNA, Simple_repeat, SINE, snRNA, srpRNA, tRNA and a final Unknown category, which consists of ultraconserved elements that are hard to classify due to their age. Each class is then split into further categories I will refer to as families. For example, the SINE class has 57 families including many Alus (eg. AluJb, AluJo, AluJr, etc.). Within each family are further sequences, for example in total the SINE class contains 72769 sequences.

I mapped reads from the miCLIP/iCLIP libraries to all the Repbase repeat sequences, removed PCR duplicates and then assigned reads to repeat classes. I was surprised to find that even at this highest level of categorisation, consistently only 50% of repeat mapped cDNAs could be uniquely assigned to a class (Figure 2.8). For now, the information of how many cDNAs can be uniquely assigned, and how many are multimapped at the class level is printed to a log file and all repeat cDNAs are reported in the abundance summary. In future iterations of the pipeline it would be useful to address this in a better way - however this is an improvement on eFAP and FAST-iCLIP pipelines where reads multimapping between classes are discarded without letting the user know.

2.3 Results

2.3.1 Pipeline Overview

In the end I split my pipeline into two separate Snakemake workflows: `prepare-annotation` and `analyse-samples`, available at

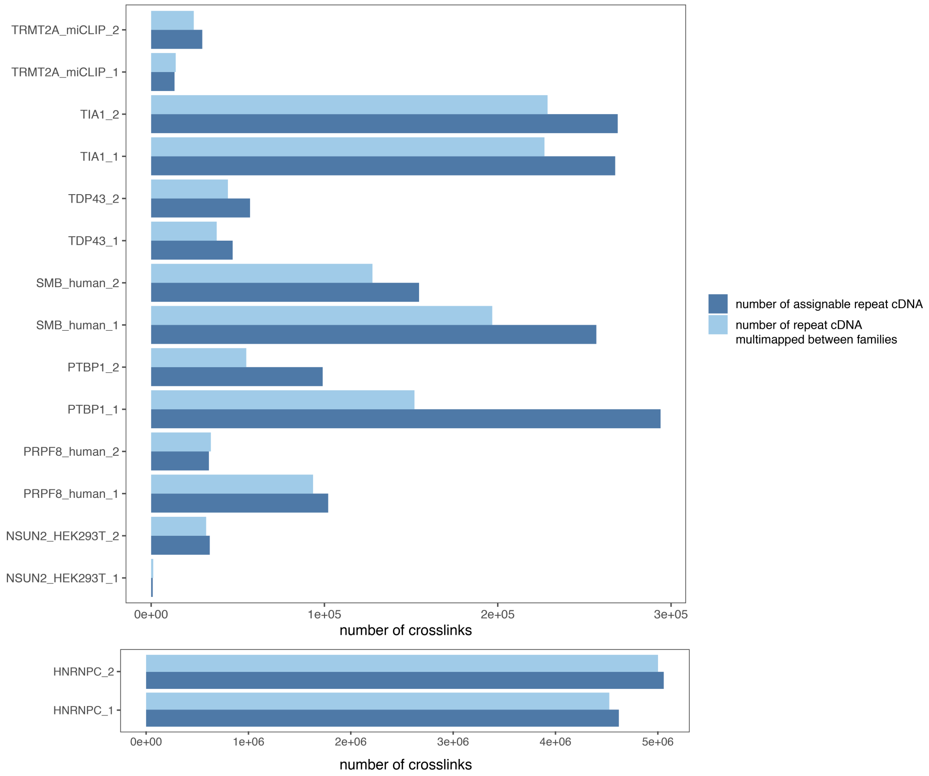


Figure 2.8: Assigning repeat-derived cDNAs to repeat classes. Column graph showing the total number of cDNAs that could be uniquely assigned to a single repeat class (dark blue) and those that multimapped between repeat classes (light blue).

https://github.com/ulelab/ncawareclip. The `prepare-annotation` workflow downloads and processes all of the annotation and sequence files from their original web sources and creates necessary indexes and processed annotation files. This process in eFAP and FAST-iCLIP is completely undocumented and users download ”pre-prepared” annotation databases prepared by the authors. By maintaining this connection to original sequences and annotation sources, every manipulation of the files is recorded and therefore made accessible to those who might want to tweak or edit the process for their own purpose. This also facilitates extending the pipeline to any genome. A species-specific configuration file is passed to the `prepare-annotation` workflow, for example the Human configuration file looks like this:

```
# Files for index generation and mapping #
```

```
species: "Human_hg38"

# Sequences and annotation files #

rDNA_sequence: "--taxon␣human␣RNA45SN1,RNA5S1"
canonical_snRNA_sequences: "--taxon␣human␣RNU1-1,RNU2-1,RNU4-1,RNU5A-1,RNU6-1"
mature_tRNA_sequence: "http://gtrnadb.ucsc.edu/GtRNAdb2/genomes/eukaryota/Hsapi38/
    hg38-mature-tRNAs.fa"
mature_snRNA_sequence: "ftp://ftp.ebi.ac.uk/pub/databases/gencode/Gencode_human/
    release_39/gencode.v39.transcripts.fa.gz"
immature_tRNA_bed: "http://gtrnadb2.ucsc.edu/genomes/eukaryota/Hsapi38/hg38-tRNAs.
    tar.gz"
genome_annotation: "ftp://ftp.ebi.ac.uk/pub/databases/gencode/Gencode_human/
    release_39/gencode.v39.primary_assembly.annotation.gtf.gz"
genome_sequence: "ftp://ftp.ebi.ac.uk/pub/databases/gencode/Gencode_human/
    release_39/GRCh38.primary_assembly.genome.fa.gz"
repeats: "http://hgdownload.soe.ucsc.edu/goldenPath/hg38/database/nestedRepeats.txt
    .gz"
mito_chromosome: "chrM"
```

Once the `prepare-annotation` pipeline has been run, the user can then proceed to the `analyse-samples` workflow. Note that `prepare-annotation` only needs to be run once per species. This is also set up using a configuration file, here the user just needs to provide the paths to their demultiplexed fastq files - I provide paths to the SmB and Prpf8 iCLIPs in the following example. There are also two options that can be set as needed. `remove_duplicates_ncRNA` allows users to decide whether to remove PCR duplicates over rRNA, tRNA, snRNA and repeats. This option is included to allow for the analysis of historical data that may have very short random barcodes - in this case the UMIs could become depleted at highly crosslinked positions in these highly abundant RNAs. This option is also useful if the user wants to compare iCLIP data mapped using the pipeline with another data type that doesn't have UMI information.

```
# Define samples #
samples:
    SMB_human_1: /camp/home/capitac/capitac/ncawareclip/Data/
        Smb_mild_HEK293_rep1_ERR3450336.fq.gz
    SMB_human_2: /camp/home/capitac/capitac/ncawareclip/Data/
        Smb_mild_HEK293_rep2_ERR3450337.fq.gz
    PRPF8_human_1: /camp/home/capitac/capitac/ncawareclip/Data/
        PRP8_Hela_rep1_ERR2822467.fq.gz
    PRPF8_human_2: /camp/home/capitac/capitac/ncawareclip/Data/PRP8_Hela_rep2_ERR2822468.fq.gz

# Indexes generated by prepare-annotation workflow #
rRNA_index: ../prepare-annotation/Human_hg38/rRNA/rRNA_index/rRNA_bowtie
rRNA_index_bowtie2: ../prepare-annotation/Human_hg38/rRNA/rRNA_index/rRNA_bowtie2
rRNA_index_star: ../prepare-annotation/Human_hg38/rRNA/rRNA_STAR_index
mature_tRNA_index: ../prepare-annotation/Human_hg38/tRNA/all_tRNA_index/all_tRNA_bowtie
mature_tRNA_index_bowtie2: ../prepare-annotation/Human_hg38/tRNA/all_tRNA_index/all_tRNA_bowtie2
immature_tRNA_index: ../prepare-annotation/Human_hg38/tRNA/all_immature_tRNA_index/all_immature_tRNA_bowtie
mature_snRNA_index: ../prepare-annotation/Human_hg38/snRNA/all_snRNA_index/all_snRNA_bowtie
immature_snRNA_index: ../prepare-annotation/Human_hg38/snRNA/all_immature_snRNA_index/all_immature_snRNA_bowtie
canonical_snRNA_index: ../prepare-annotation/Human_hg38/snRNA/canonical_snRNA_index/canonical_snRNA_bowtie
mito_index: ../prepare-annotation/Human_hg38/mito_separated/Mito_star_index
repeats_index: ../prepare-annotation/Human_hg38/repeats/repeats_index/repeats_bowtie
genome_minus_mito_index: ../prepare-annotation/Human_hg38/mito_separated/genomeMinusMito_star_index
genome_regions: ../prepare-annotation/Human_hg38/full_genome/regions.gtf.gz

# Optional parameters #
remove_duplicates_ncRNA: True
fraction_trnaisotype_merge: "0.9"
```

In `analyse-samples` input fastq files are first run through Trim Galore! to trim reads for sequencing adaptors, sequencing quality (bases of less than Q=20 are trimmed from the 3' end) and length, allowing a minimum length of 11nt (Krueger, 2017). Fastq files are also run through FastQC to give an initial indication of sequencing quality (Andrews et al., 2010). Following this, reads are sequentially mapped to rRNA → mature tRNA → immature tRNA → mature snRNA → immature snRNA → mitochondrial chromosome → repeats → nuclear genome (Figure 2.9). Each mapping step is performed using Bowtie, allowing two mismatches, aside from the mitochondrial and nuclear genome mapping which is performed using STAR, allowing for spliced read mapping and soft clipping of the 3' end of reads (Dobin et al., 2013; Langmead, 2010). PCR duplicates are then

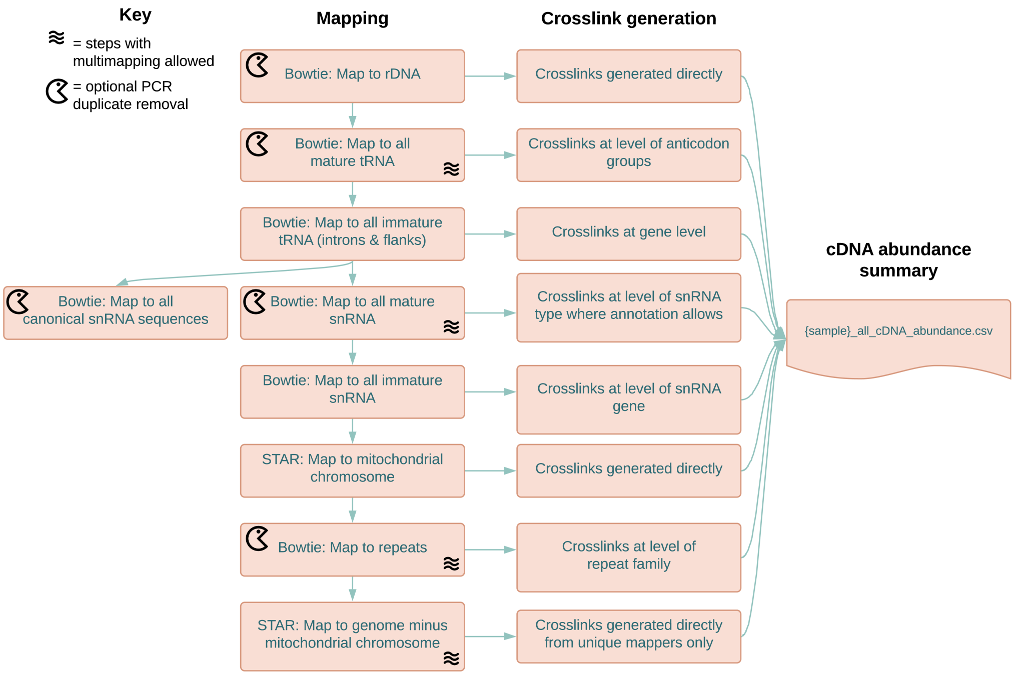

Key
≋ = steps with multimapping allowed
= optional PCR duplicate removal
Mapping
Crosslink generation
Bowtie: Map to rDNA
Bowtie: Map to all mature tRNA
Bowtie: Map to all immature tRNA (introns & flanks)
Bowtie: Map to all canonical snRNA sequences
Bowtie: Map to all mature snRNA
Bowtie: Map to all immature snRNA
STAR: Map to mitochondrial chromosome
Bowtie: Map to repeats
STAR: Map to genome minus mitochondrial chromosome
Crosslinks generated directly
Crosslinks at level of anticodon groups
Crosslinks at gene level
Crosslinks at level of snRNA type where annotation allows
Crosslinks at level of snRNA gene
Crosslinks generated directly
Crosslinks at level of repeat family
Crosslinks generated directly from unique mappers only
cDNA abundance summary
{sample}_all_cDNA_abundance.csv

Figure 2.9: Flow diagram showing steps of the ncRNA-aware CLIP pipeline: specifically the `analyse-samples` workflow.

removed using the UMI and mapped start position of reads. For steps without multimapping enabled crosslinks are generated directly by taking the start position of uniquely mapped cDNAs. Mitochondrial and nuclear genome mapped crosslinks are annotated using the regions file outputted by iCount segment (König et al., 2010)[https://github.com/ulelab/iCount-Mini]. Multimapped mature tRNA are processed to give crosslink information at the level of roughly tRNA isotype, as previously described. Mature snRNA is processed to give crosslink abundance at the level of snRNA family, canonically mapped snRNA crosslinks can then be used to investigate positional information. Multimapped repeats are processed to give crosslinks at the level of repeat class and separately repeat family, to give more detailed information.

2.3.2 Pipeline run on test data

The pipeline accurately assigns cDNA abundances to the appropriate RNA types (Figure 2.10). Increased proportions of snRNA are captured in the SmB and Prpf8 iCLIP libraries and the high repeat content of HnRNPC, Tia1 and Ptbp1 libraries is also captured. Tia1 has extensive 3'UTR binding which is well known and corresponds with its role in regulating alternative polyadenylation (Rot et al., 2017). Especially interesting is the abundance of NSun2 and Trmt2A mapping to immature tRNA sequences. This is not without precedent - it was previously described that a yeast orthologue of NSun2 requires an intron *in vitro* to methylate m5C34 in tRNA-Leu-CAA (Brzezicha et al., 2006). In the case of Trmt2A, the *E. coli* enzyme catalysing tRNA m5U54, TrmA, is required as a tRNA chaperone during its maturation, hinting that possibly Trmt2A is catalysing m5U54 in pre-tRNA (Keffer-Wilkes et al., 2020).

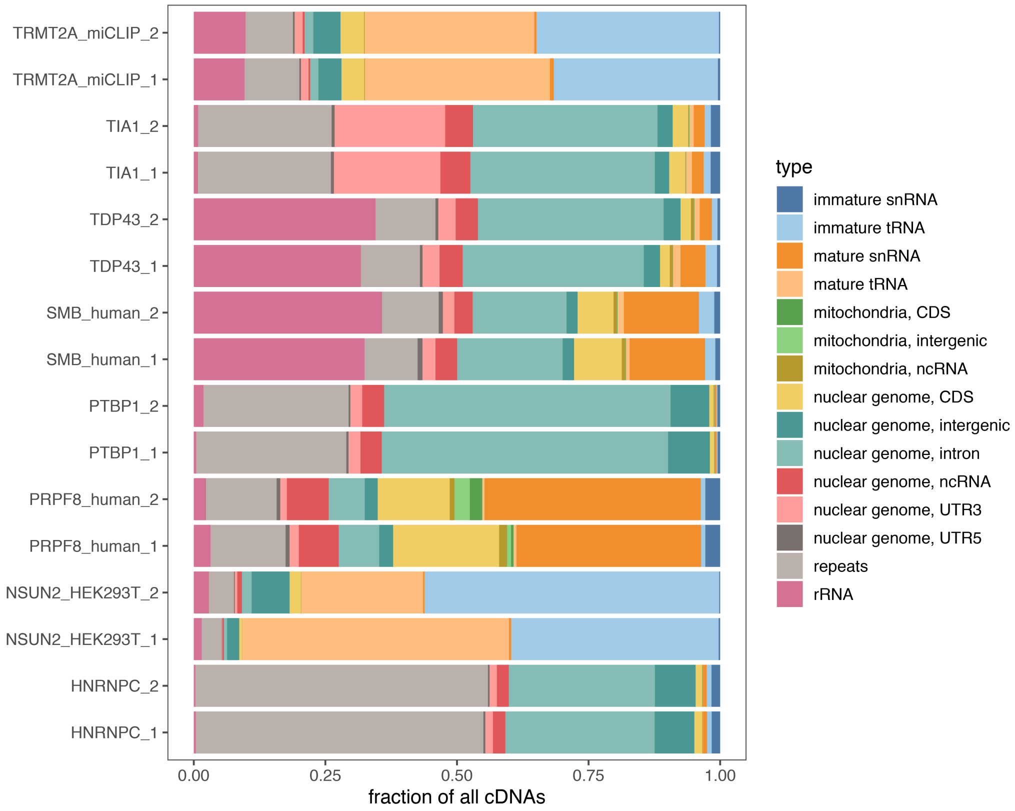


Figure 2.10: Final results of pipeline: RNA type abundance summary. The contents of [sample]_all_cDNA_abundance.csv is plotted for each sample.

TDP-43 and SmB iCLIP have the highest proportion of ribosomal RNA mapping, with over 25% for both proteins (Figure 2.10). To my knowledge there are no reported interactions between SmB and ribosomes, furthermore it has been shown that SmB doesn't co-sediment with polysomes (Aviner et al., 2017). This SmB data was produced with mild washing conditions, so the most likely source of rRNA is from contamination during sample preparation. Indeed, when harsher washing conditions are used, the percentage of rRNA in the library decreases (Faraway, 2021).

The rRNA in the TDP-43 data is less likely to be a contaminant due to the stringent washing conditions used and also some previous literature which suggests some contact between TDP-43 and ribosomal proteins. TDP-43 has been shown to co-IP with ribosomal protein subunits, and has been specifically found to associate with stalled ribosomes in stress granules (Freibaum et al., 2010; Higashi et al., 2013). This interaction may be mediated by the protein RACK1 (Russo et al., 2017). More recently it's been suggested that TDP-43 is an enhancer of translation in specific mRNAs in a neuronal context (Neelagandan et al., 2019). However, it's difficult to determine if this is via direct interaction with ribosomes and potentially rRNA, or an indirect effect related to TDP-43's role in mRNA transport, for example. On close inspection TDP-43 crosslinks are abundant at positions across all ribosomal RNAs, which suggests the interaction identified by iCLIP is likely to be non-specific (data not shown). To exclude cross-reactivity between TDP-43 antibodies and ribsomal proteins it would be useful to perform mass spectrometry analysis of crosslinked protein-RNA samples prepared for iCLIP. Taken together, more biochemical work is required to untangle the exact interactions enabling TDP-43's ability to impact translation.

2.3.3 Case Study: DDX3X is a novel tRNA binding protein

2.3.3.1 Background

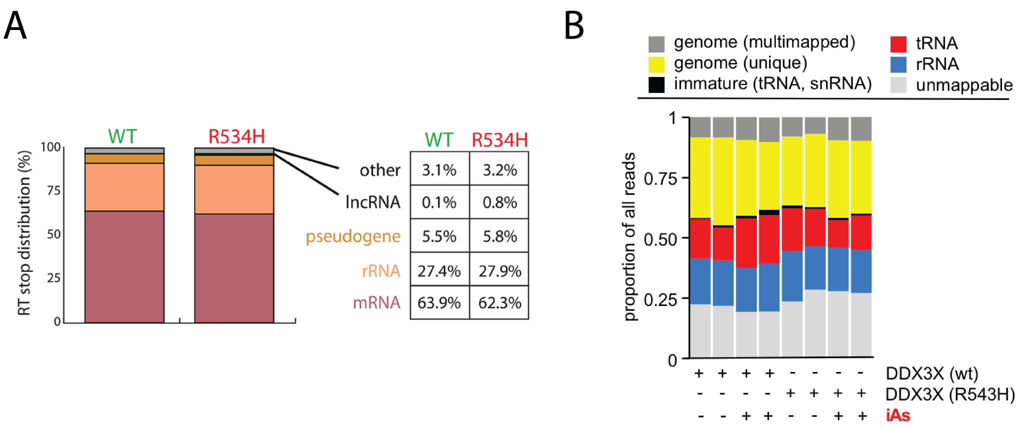
tRNAs are well known for their role in mRNA translation, where they carry amino acids to ribosome-mRNA complexes to facilitate synthesis of nascent polypeptide chains. Alongside this essential role, it has become increasingly understood that tRNAs can also be subject to tightly controlled fragmentation to produce tRNA-derived small RNAs (tsRNAs). These tsRNAs participate in control of apoptosis, inhibition of endogenous retrotransposon activity and intercellular communication including transgenerational epigenetic inheritance (J.-T. Wen et al., 2021). In mammals, the most well understood mechanism of tsRNA biogenesis is mediated by the endonuclease RNase A-family member Angiogenin (ANG). Under stress conditions, ANG is post-translationally modified and becomes active, nicking tRNAs in their anticodon loops, resulting in the production of 5' and 3' tRNA halves. The 5' tsRNAs are seemingly more abundant and therefore hypothesised to be more functionally relevant (Drino et al., 2020).

Dr. Aleksej Drino from the group of Prof. Matthias Schaefer was interested in discovering novel regulators of the tsRNA biogenesis process and so designed a screen to identify RBPs bound to 5' tRNA halves. HEK293 cells were stressed by exposure to inorganic sodium arsenite (iAs) followed by UV-crosslinking, protein extraction and chromatographic fractionation followed by mass spectrometry of tsRNA-containing RNPs. 5' tsRNA-containing RNPs were pooled and further separated by ion exchange chromatography and fractions containing a positive signal ratio of 5' tsRNA/tRNA were subjected to mass spectrometry to determine the identity of co-migrating proteins. Since RNPs co-migrating with 5' tsRNAs contained full-length tRNAs, fractionation was also performed on protein extracts that were not exposed to iAs as a control. Triplicate mass spectrometry analyses of control RNPs as well as 5' tsRNA-containing RNPs after enrichment from stressed cells identified 114 proteins which were exclusively detected in 5' tsRNA-containing RNPs and not in the corresponding fractions originating from control cells. Various proteins that were detected in 5' tsRNA/tRNA-containing fractions are well known known for their roles in tRNA biology (e.g. TRMT10C, RTCA, RNAseT2, EIF2, EIF5A, EEF1G, TUFM, EEF2), which supported the notion that fractions contained tRNA-interacting proteins. Notably, while no RNA exonuclease was among the detected proteins, several highly conserved RNA helicases such as eIF4A1/2, DDX5/17, DDX3X and DDX39A/B co-migrated in 5' tsRNA/tRNA-containing RNPs. After further biochemical characterisation, Dr. Aleksej Drino found evidence of DDX5 and DDX3X binding to tRNAs. He hypothesised that these helicases could have the potential to act on nicked tRNAs during the oxidative stress response and via their RNA unwinding activity, liberate the tsRNAs from the full length tRNA. This would represent a critical missing step in our current view of tsRNA biogenesis.

Fortunately, high quality FAST-iCLIP data was publicly available for FLAG tagged DDX3X from a previous study, and so I collaborated with Aleksej to reanalyse this data through my pipeline, to see if we could characterise the tRNA binding further (Oh et al., 2016).

2.3.3.2 DDX3X iCLIP analysis

Published DDX3X iCLIP was available in sodium arsenite vs. wild type conditions and also with expression of a helicase-dead mutant (DDX3R534H) (Oh et al., 2016). By reanalysing this data through my pipeline I found 15-20% of reads mapped to tRNA - something which was surprisingly missed in the original manuscript, despite the fact that the FAST-iCLIP pipeline was used for analysis (Figure 2.11).



	WT	R534H
other	3.1%	3.2%
lncRNA	0.1%	0.8%
pseudogene	5.5%	5.8%
rRNA	27.4%	27.9%
mRNA	63.9%	62.3%

Figure 2.11: DDX3X iCLIP biotype distribution. A) Analysis of RT stops as presented in (Oh et al., 2016). B) Read mapping distributions for all DDX3X iCLIP samples processed through my ncRNA-aware mapping pipeline.

Furthermore, I found that whilst the overall proportion of tRNA reads in stress decreased in matched RNA-Seq from the original manuscript, the amount of tRNA reads went up in DDX3X iCLIP under stress and this effect was abrogated in the helicase-dead mutant iCLIP data (Figure 2.12A). This supports the idea that

DDX3X is involved in processing tRNAs in response to stress, in a manner that is dependent on its helicase activity. I was interested in the positioning of DDX3X on tRNAs, so I produced a metaprofile of iCLIP and RNA-Seq signal over all tRNA genes. I found that the iCLIP data was highly enriched at the anticodon loop (Figure 2.12B). This was not seen in the RNA-Seq, which gave me confidence that this was not only due to RT stops at an RNA modification, e.g. m1A or m7G.

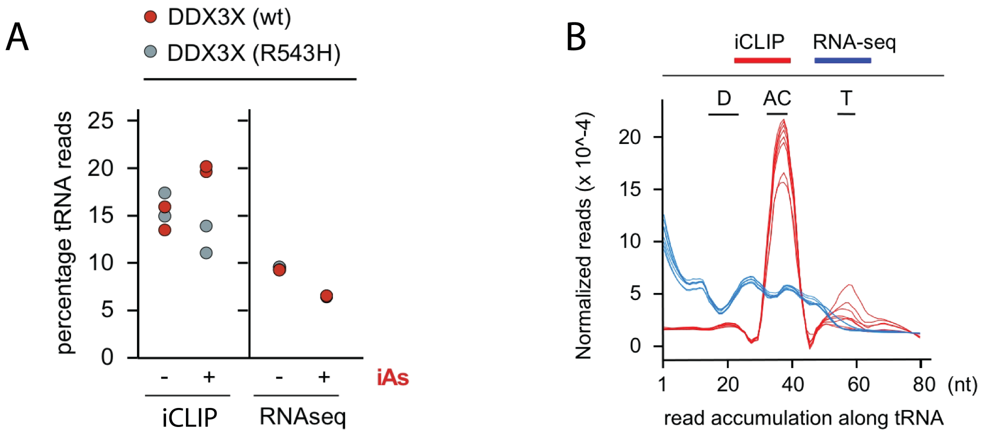


Figure 2.12: DDX3X binding under steady state and stress conditions. A) Changes in DDX3X-mediated iCLIP-derived tRNA reads and in tRNA expression data obtained from duplicate experiments using DDX3X or DDX3X (R534H) during steady-state conditions and after iAs exposure. Normalized abundance of tRNA-derived reads from iCLIP (left panel) or total RNA-sequencing experiments (right panel) are shown. B) Metaplot depicting positional information of DDX3X-mediated iCLIP signatures on tRNA sequences obtained from (Oh et al., 2016). 5' most nucleotide positions from individual tRNA-derived reads were determined and normalised against library size for individual samples (red). As a comparison, tRNA-derived reads from total RNA-seq data treated in the same way are shown (blue). Individual lines represent replicate experiments. Letters indicate the positions of D-, anticodon- (AC) and T- loops within tRNA sequences.

Finally I was interested in whether DDX3X exhibited any specificity to binding certain tRNA isotypes. Using my pipeline I was able to resolve the reads to near isotype resolution (Figure 2.13). Interestingly, I found that the most abundant tRNA isotype in the iCLIP data, Lys^{UUU}, was not the most abundant in the RNA-Seq data, which suggested a specificity that was not simply guided by tRNA abundance (Figure 2.13 A,B). To validate that Lys^{UUU} was indeed a target of DDX3X, Dr. Aleksej

Drino affinity purified tRNA-Lys^{UUU} from HEK293 cells and subjected it to an electrophoretic mobility shift assay (EMSA) using MBP-DDX3X. The results showed that increasing concentrations of MBP-DDX3X shifted tRNA-Lys^{UUU} to completion (Figure 2.13 C). From analysis of triplicate EMSA he was able to calculate the dissociation constant (Kd) between MBP-DDX3X and tRNA-Lys^{UUU} as 25 ± 3.2 nM. This is similar to the affinity of *S. cerevisiae* Aspartyl-tRNA synthetase to tRNA (Kd= 30nM) and much stronger than *Methanocaldococcus jannaschii* Trm5 binding to tRNA-Cys (Kd= 700nM) (Yang et al., 2013).

2.3.3.3 Conclusions

Further to demonstrating that DDX3X is a novel tRNA binding protein, this case study highlights the utility of my pipeline in undertaking rigorous analysis of iCLIP data, as the tRNA binding was missed in the initial DDX3X iCLIP manuscript.

An open question is the meaning of finding DDX3X crosslinks predominantly in the anticodon loop of tRNAs. This means that reads extend into the 3' half of tRNAs, whilst DDX3X was enriched in mass spectrometry of 5' tRNA halves. Additionally, 15% of DDX3X iCLIP reads map to tRNA under steady state conditions where little fragmentation is occurring. Whilst this increases to 20% upon stress induction, the 5% increase in binding is to the same tRNA isotypes. One explanation could be that DDX3X binds to full length tRNAs either 3' or 5' to where ANG would nick once activated. Crosslinking is often preferential to single stranded regions of RNA, so in this case crosslinks to the anticodon loop could represent binding on either side. Perhaps once the nicking has occurred, the unwinding activity of DDX3X is fast, and therefore iCLIP mostly captures the steady state scenario where DDX3X is poised at certain tRNAs. We must remember, also, that various hybridisation studies have suggested that only 0.1–5% of a given tRNA isoacceptor yield tsRNAs, even under stress conditions (Oberbauer & Schaefer, 2018; Yamasaki et al., 2009). Therefore, the overall ratio of tsRNA to whole tRNAs is likely to be very low even in the stress condition. It has been proposed that whilst the total amount is low, the production is likely localised to certain places in the cell where a

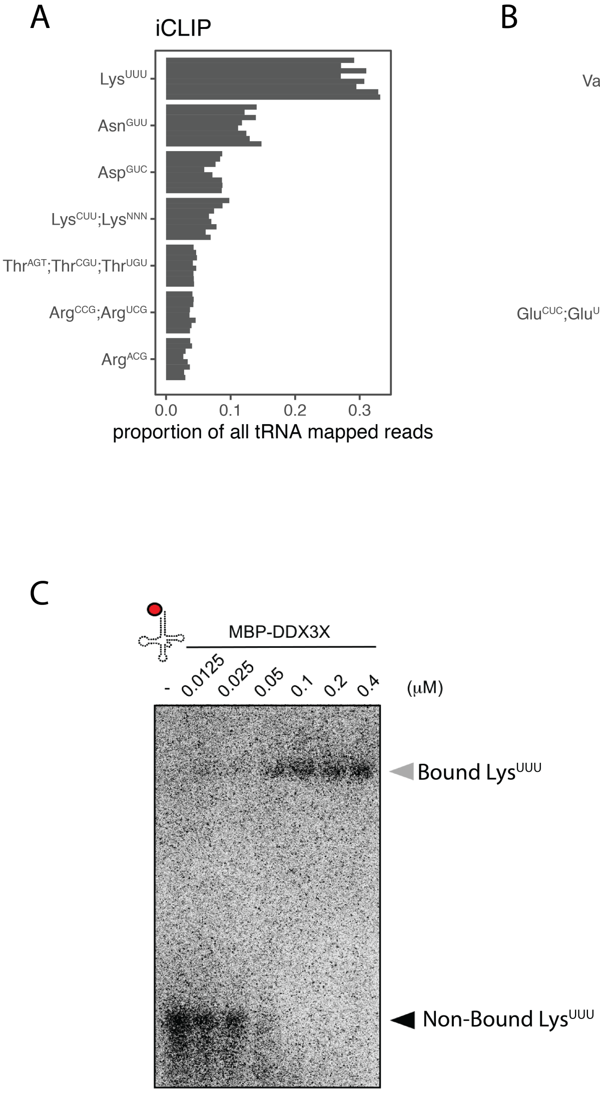


Figure 2.13: DDX3X isotype specificity. A) Ranking of DDX3X-derived iCLIP reads obtained from (Oh et al., 2016) mapping to tRNAs according to their normalized abundance. Individual rows at each tRNA isodecoder box represent the proportion of tRNA reads among all mapped tRNA sequences. Individual grey bars represent individual iCLIP experiments. B) As in A, but for matched total RNA-Seq data processed in the same way. C) Representative EMSA performed by Dr. Aleksej Drino after combining increasing molarities of MBP-DDX3X and 5' end-labeled tRNA-Lys^{UUU} (30 nM final). UV-crosslinked RNPs were separated using nPAGE. Black arrowhead, non-bound tRNAs; grey arrowhead, DDX3X-tRNA-Lys^{UUU} complexes.

high local tsRNA concentration could exert function (Oberbauer & Schaefer, 2018).

This leads to a related question: what does binding enrichment of certain tRNA isotypes in iCLIP actually mean? If the kinetics of unwinding is generally fast such that we do not capture it with iCLIP, then binding enrichment in iCLIP could represent substrates that DDX3X struggles with. Similarly, the binding we see in iCLIP could represent contacts that have nothing to do with tsRNA biogenesis. DDX3X is involved in a staggering number of cellular processes, including some related to translation (Mo et al., 2021). To uncover the true meaning of the binding we detected will require further careful biochemical investigations.

2.4 Discussion

2.4.1 On tRNA resolution with iCLIP

In this chapter I have developed a ncRNA-aware CLIP analysis pipeline that accurately reports cDNA abundance over a range of repetitive ncRNAs. One of my main focuses setting out was to quantify tRNAs at the highest possible resolution while discarding as few cDNAs as possible, which I achieved by grouping at the level of tRNA isotype and then further merging groups on a case-by-case basis as determined by the data itself.

One point to consider is theoretically how much tRNA information we can extract from iCLIP data. Because iCLIP data is produced using a reverse transcription step that prioritises truncations at crosslink sites, we might expect that this process would also favour truncation at RNA modifications at the Watson-Crick interface. This could be problematic in two ways: 1) in the case where an RNA modification "interrupts" reverse transcription towards a more 5' crosslink event, resulting in a cDNA product that is too short to map and 2) similarly to (1), but where the cDNA is long enough to map, a site of RNA modification could be misinterpreted as the crosslink site, when in reality the crosslink site would be somewhere 5' to the modification. One possible way to mitigate this experimentally could be to use the approach introduced by DM-tRNA-seq and treat crosslinked, purified RNA *in vitro* with AlkB and a D135S mutant to remove m1A, m3C and m1G, before proceeding to library preparation (Zheng et al., 2015).

Additionally, there could be an impact of structure on the *in vivo* crosslinking. Crosslinking is more likely in single stranded regions so we might expect that crosslinks would be more likely in the D, anticodon or T loops. As the T loop is typically 15-20 nucleotides away from the 3' end of the tRNA it might be harder to get reads derived from this region - if they are much shorter they would be much harder to map.

2.4.2 Further development

Currently the pipeline is written in Snakemake, and the `analyse-samples` and `prepare-annotation` workflows have become quite unwieldy, with substantial code duplication for all the Bowtie mapping steps for example. I would like to port it to Nextflow to take advantage of the nf-core modules repository and general modular DSL2 syntax, which would make the pipeline more readable and easier to edit. Eventually it would be ideal that the ncRNA-aware steps become available as a part of the official nf-core/clip-seq pipeline, which currently just removes tRNA/rRNA mapped reads. The second benefit to a switch to Nextflow DSL2 is that currently my Snakemake pipeline is not containerised, meaning that people could run into problems in other compute environments that would be impossible for me to anticipate. Using Nextflow DSL2 each module is run in its own docker container, this: a) makes the pipeline as a whole very portable, it should be able to be run in any compute environment, b) makes the containerisation trivial to implement - for nf-core modules it is already implemented, and for additional modules if the program is available on BioConda, a Biocontainer is already automatically generated. Porting into Nextflow will also allow the pipeline to be integrated into the iMaps platform (imaps.goodwright.com), providing the pipeline with a graphical user interface and therefore making it accessible to biologists without coding experience.

Further to these technical improvements, I found that consistently 50% of repeat element derived reads multimap between repeat families. Using these reads for quantification could drastically change how individual elements are quantified. It would be useful to first determine what the distribution of read lengths is for these remaining multimapped reads as it might be that they are too short to reliably map. Further, it would be interesting to compare different alignment algorithms to determine if any are better at aligning these reads uniquely. A simulation of reads derived from repeat elements would help in this instance, to determine the accuracy of different algorithms. One possible solution could be to come back to the CLAM approach, but only inputting the aligned regions that are relevant to these multimapped repeat reads in order to reduce the memory requirements.

Chapter 3

N6-methyladenosine-dependent RBP networks in Yeast Meiosis

3.1 Introduction

3.1.1 Budding yeast life cycle and entry into meiosis

Under nutrient-rich conditions the budding yeast *Saccharomyces cerevisiae* exists in a diploid state. If the cell is nitrogen starved, but still has access to glucose, the yeast will begin growing in a pseudohyphal state, where cells elongate and line up end-on-end after mitotic divisions in an attempt to find a more hospitable location. However, with a complete lack of nitrogen, sugar starvation, and additionally the presence of a non-fermentable carbon source such as acetate, the yeast will undergo meiosis (sporulation) to produce haploid gametes (Figure 3.1A) (Neiman, 2005; Strudwick et al., 2010). These gametes, also called spores, are contained within a sac called an ascus, which protects them from environmental insults. The process of returning to the mitotic cell cycle is called germination. The first step of this process, breakdown of the ascus, occurs once the yeast is exposed to glucose again, although a full return to mitotic growth requires additional nutrients (Joseph-Strauss et al., 2007).

At the molecular level, the transcription factor inducer of meiosis 1 (Ime1p) is the master regulator of the meiotic program (Figure 3.1B). Once induced Ime1p binds to the DNA-binding protein Ume6p which triggers degradation of Ume6p, a repressor which binds many early meiotic gene promoters where it recruits the ISW2 chromatin remodelling complex and the Rpd3-Sin3 histone deacetylase (Strudwick et al., 2010). An important Ume6p target is IME2, which is a serine-threonine protein kinase, which itself phosphorylates Ime1p causing its degradation (H. E. Smith & Mitchell, 1989). *ime2*-Δ cells can undergo another round of DNA replication, suggesting the activity of Ime2p is critical for ensuring a transition to the next stage of meiosis (Guttmann-Raviv et al., 2002).

Meiosis is tightly regulated to prevent cells making costly mistakes, for example inducing a lethal meiosis in a haploid cell. In haploid cells, the induction of upstream lncRNA *IRT1* by transcription factor Rme1p, prevents transcription of the downstream *IME1* (van Werven et al., 2012). Similarly, expression of the yeast m6A methyltransferase Ime4p, is inhibited by an antisense lncRNA at the *IME4* (*IME4-AS*) locus in haploid cells (Figure 3.1B) (Hongay et al., 2006). When the cell is ready to sporulate, the a1/α2 repressor complex silences *IME4-AS* and also binds to sites in the *RME1* promoter, to prevent its expression, therefore releasing *IME1* from transcriptional repression by *IRT1*.

3.1.2 m6A-networks in yeast meiosis

In yeast, m6A is catalysed by Ime4p, a Mettl3 ortholog. Ime4p is bound by Slz1p and Mum2p, an ortholog to mammalian WTAP (Figure 3.2A). Together, these three proteins are known as the MIS (<u>M</u>um2p, <u>I</u>me4p, <u>S</u>lz1p) complex (Agarwala et al., 2012). Homozygous deletion of *IME4* results in severe sporulation defects (Hongay et al. 2006). DNA replication proceeds with wild type kinetics, but during Anaphase I chromosome segregation is defective (Agarwala et al., 2012). Notably, m6A has only been observed in yeast during early meiosis, peaking at roughly three hours, coinciding with DNA replication (Figure 3.2B) (Agarwala et al., 2012). m6A can also be induced by treating cells with rapamycin, which mimics nitrogen starvation (Bodi et al., 2015). Ime4p expression is detectable from the onset of meiosis,

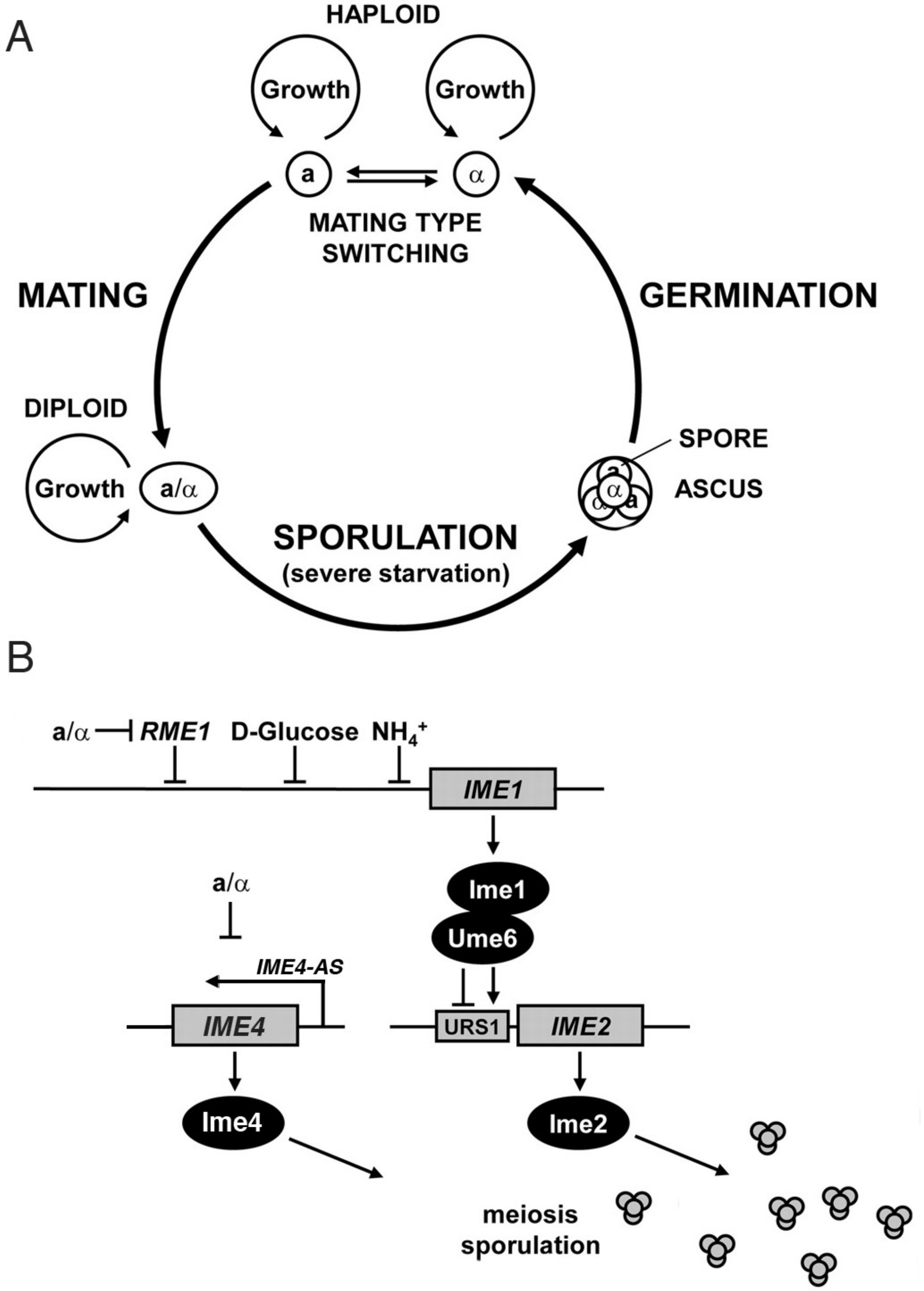


Figure 3.1: Transcriptional control of the *Saccharomyces cerevisiae* life cycle. A) When exposed to nutrient rich conditions budding yeast will grow as a diploid, upon severe starvation sporulation will occur. Upon return to nutrients, the haploid spores will germinate and mate to create diploid cells again. B) Glucose, nitrogen and *RME1* all repress the expression of *IME1* in diploid yeast cells. Activation of yeast mating factor, removal of glucose and nitrogen releases this repression. Ime1p binds to transcriptional repressor Ume6p, which causes degradation of Ume6p and release of repression at early meiotic genes including *IME2*. Similarly, activation of yeast mating factor represses the antisense lncRNA repressor of *IME4*, causing Ime4p expression. Figure adapted from (Strudwick et al., 2010) to include information from (Hongay et al., 2006).

with mRNA levels highest between hours 3—6, and transcriptional repression beginning from hour 5 onwards as *IME4-AS* becomes detectable (Agarwala et al., 2012). Interestingly, a catalytic dead mutant of Ime4p (*ime4-D349A, W351A*) has an intermediate meiotic phenotype between *ime4*Δ and WT, suggesting that Ime4p has both catalytic and non-catalytic functions in yeast meiosis. (Agarwala et al., 2012).

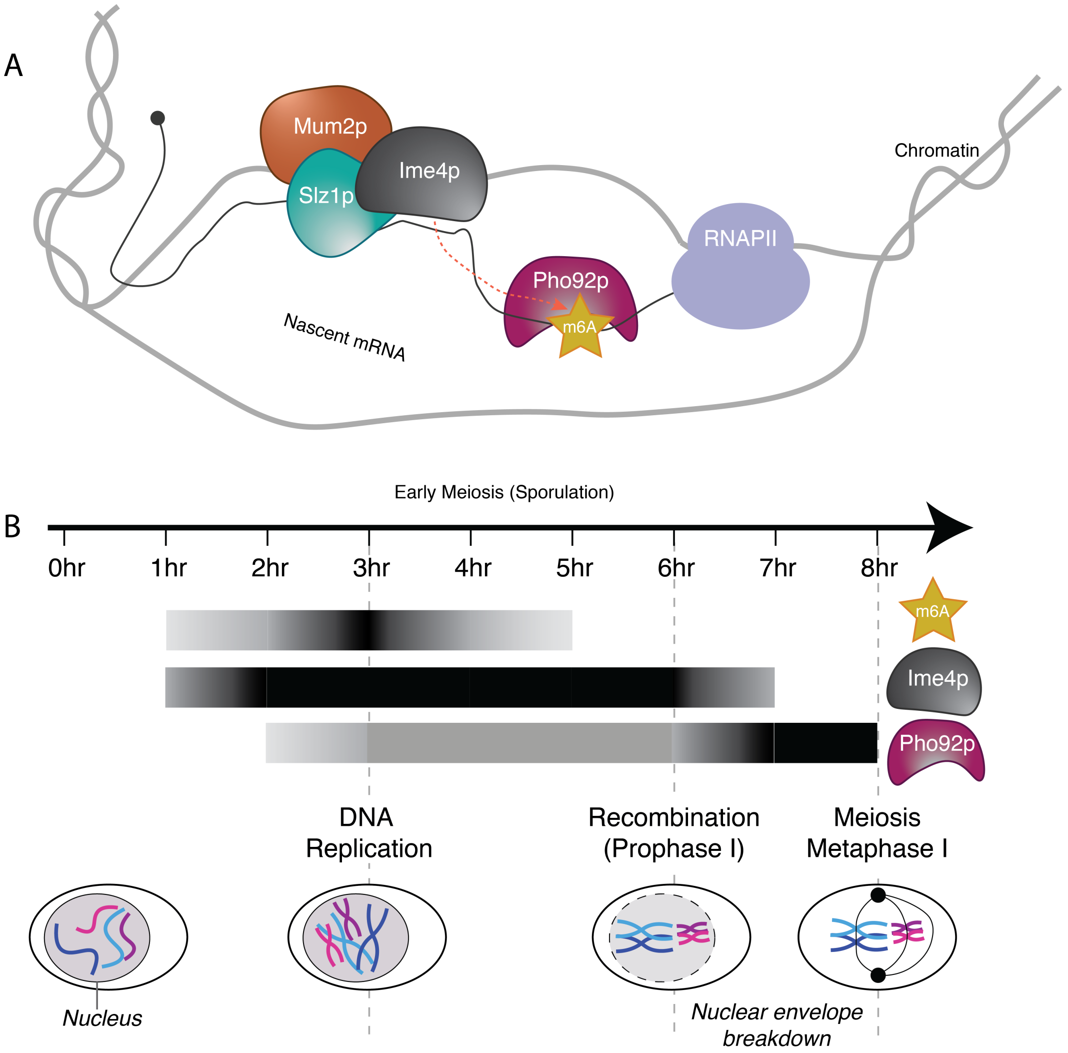


Figure 3.2: m6A-dependent RBP networks in yeast meiosis. A) The MIS complex methylates RNA in the nucleus, here suggested to occur at chromatin. Pho92p binds to this m6A. B) Timeline of early meiosis indicating the expression levels of Ime4p and the level of m6A as measured in (Agarwala et al., 2012) and Pho92p as measured by Dr. Radhika Varier (data not shown). Note that Pho92p is expressed under normal nutrient-rich yeast conditions, but to a much lower level (Kang et al., 2014).

It is likely that most m6A methylation of yeast mRNA occurs in the nucleus, as in mammalian cells. Ime4p has been shown to co-localise with Slz1p and Mum2p in the nucleus in a manner dependent on Slz1p and its N-terminal nuclear localisation signal (Schwartz et al., 2013). However, Slz1p is not essential for m6A methylation itself, as ectopic expression of Ime4p and Mum2p alone is sufficient for $\sim 75\%$ of methylation (Agarwala et al., 2012). This perhaps explains why *MUM2* deletion has a similar phenotype to *IME4* deletion, whilst *SLZ1* deletion has a less severe phenotype—there is still a delay to chromosome segregation, but the final proportion of asci containing the correct number of spores is similar to WT (Agarwala et al., 2012).

Previously, several methods have been used to map the locations of m6A modifications in yeast meiosis. Results from thin layer chromatography (TLC) analysis showed $\sim 1\%$ of adenosines in poly(A)+ RNA are m6A modified in sporulating yeast (Bodi et al., 2010). *IME2* mRNA was the first yeast transcript to be discovered as m6A-modified, as detected by m6A antibody pull down followed by RT-qPCR. By protecting either end of the mRNA before performing TLC, the m6A was localised to the 3' end of the transcript. *IME1* and *IME4* transcripts were also found to be modified by the m6A IP followed by RT-qPCR approach (Bodi et al., 2010).

Following in the footsteps of m6A-Seq studies in mammalian cells, m6a-Seq was subsequently performed on yeast mRNAs using the anti-m6A Synaptic Systems polyclonal antibody (Dominissini et al., 2012; Schwartz et al., 2013). m6A in *IME2* mRNA was validated by m6A-Seq and also by loss of m6A-Seq signal upon mutation of the methylated adenosine, however sites in *IME1* and *IME4* were not validated by m6A-Seq and so they weren't checked by mutational analysis. Overall, by comparing m6A-Seq performed in WT vs. *ime4*-Δ lines, 1308 m6A sites were identified within 1183 genes. 711 of these sites contained a peak that was < 5nt away from a consensus RGAC, and of these most conformed to an extended consensus of ANRG(m6A)CNNU, suggesting a more constrained m6A motif in yeast compared to mammals. m6A sites were enriched at the 3' end of transcripts, as in mammalian studies (Schwartz et al., 2013).

With the advent of antibody-independent approaches to mapping m6A, MAZTER-Seq was developed, utilising the m6A-sensitive endoribonuclease MazF. MazF cuts RNA in an ACA sequence context with preference for unmethylated ACA vs. (m6A)CA, which can be exploited in library preparation (Garcia-Campos et al., 2019). Through application in SK1 sporulating yeast, a logistic regression model was built capable of predicting stoichiometries of RRACA m6A sites from cis regulatory information alone, with the top three most important predictors being the identity of the +4, -2 and -4 nucleotides with respect to the m6A site. This led the authors to conclude that m6A in yeast is genetically "hard-coded" and likely only regulated by altering expression levels of the MIS complex. However, RRACA sites represent an estimated 16%–25% of yeast m6A so the problem of yeast m6A is still far from solved.

3.1.3 Pho92p is an m6A-reader

Pho92p, also known as MRB1p, is the sole YTH protein in *S. cerevisiae*, sharing most sequence similarity with YTHDF2 (Kang et al., 2014). The YTH domain shares extensive sequence similarity to all mammalian YTH domains and a crystal structure shows that it binds m6A similarly, by encasing it in a hydrophobic aromatic cage (Xu et al., 2015). As well as the C-terminal YTH domain, Pho92p has an N-terminal low complexity domain which hypothetically could also interact with RNA. It has been suggested that the plant YTH ortholog ECT2, binds to poly(U) and UNUNU via its IDR (Arribas-Hernández, Rennie, Schon, et al., 2021).

Ultimately the function of m6A methylation in yeast meiosis is poorly understood, however whatever function it has is likely to be facilitated by RBP effectors. Critically, the binding sites of Pho92p are as yet unknown, hindering our understanding of its function. Based on the known functions of other YTH orthologs it is possible Pho92p could contribute to mRNA decay or translation. It has been shown that m6A-modified yeast mRNAs are enriched in polysome fractions, suggesting a role for m6A related to active translation (Bodi et al., 2010). A lower abundance of m6A-modified target RNAs upon induction of m6A has also been reported, suggesting a role of m6A in mRNA decay (Schwartz et al., 2013).

Pho92p itself was originally characterised as part of the phosphate signal transduction (PHO) pathway. It has been shown to co-immunoprecipiate with Pop2p of the Pop2–Ccr4–Not deadenylase complex, to reduce the the stability of the *PHO4* mRNA in phosphate rich conditions where its not needed (Kang et al., 2014). Expression of the Pho92p YTH domain alone did not regulate the stability of *PHO4* mRNA, supporting that YTH proteins act via recruitment of other protein complexes via N-terminal, non-YTH, regions.

3.1.4 Gis2 as a novel m6A-reader

To identify further m6A readers in yeast meiosis, Dr. Radhika Varier collaborated with the group of Michiel Vermeulen to perform a pull down of yeast cell proteins bound to either an m6A-modified bait RNA oligo (comprising four GG(m6A)CU repeats) or a matched unmodified control bait (Figure 3.3) (Varier et al., 2022). The experiment identified Pho92p as the most enriched protein bound to m6A-modified oligos in early meiotic cell extracts, the second most enriched protein was Gis2p.

Gis2p is the *S.cerevisiae* ortholog of human CCHC-type zinc finger nucleic acid-binding protein, CNBP, sharing 59% sequence similarity (Sammons et al., 2011). Interestingly, under zinc-deficiency, expression of *GIS2* is shut down by run on transcription of an upstream RNA, potentially as a zinc-saving mechanism by yeast cells. Gis2p has 7 CCHC-type zinc fingers, and therefore contains 7 zinc molecules per protein - estimated to constitute 2–8% of the total zinc requirement of the cell (Taggart et al., 2018). Neither Gis2p or CNBP protein has been implicated in m6A-dependent networks previously.

Gis2p associates with polysomes, translation factors such as translation initiation factor eIF4G and poly(A) binding protein in a manner that is RNAse-dependent, suggesting a role in translation that depends on binding to mRNAs rather than protein-protein interactions (Rojas et al., 2012; Sammons et al., 2011). Furthermore, ectopic expression of Gis2p in HEK293 cells with a reporter based on human ornithine decarboxylase mRNA, which harbours an internal ribosome en-

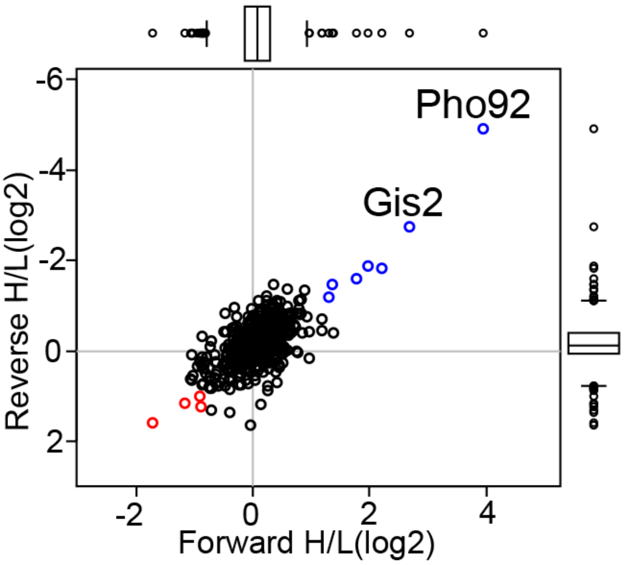


Figure 3.3: Mass spectrometry screen to identify m6A readers in early meiosis, performed by Dr. Radhika Varier in collaboration with the group of Michiel Vermeulen (Varier et al., 2022). Scatter plot displaying proteins identified in m6A consensus oligo pull down versus control. In short, cells were grown in nutrient rich medium till saturation, shifted to nutrient depleted media, and then induced to enter meiosis using *CUP1* promoter fused to *IME1*. Protein extracts were incubated using m6A and control RNA baits bound to streptavidin beads. Eluted proteins were differentially labelled with light and heavy dimethyl isotopes, mixed, and proteins from forward and reverse label swap reactions were identified by MS.

try site (IRES), demonstrated that Gis2p could facilitate cap-independent translation, similar to CNBP. Interestingly, spot dilution vegetative growth assays showed no growth defect in *gis2*-Δ, and no difference in growth to WT strains when cells were treated with different translation inhibitors, suggesting Gis2p is not essential in nutrient-rich conditions, even under translational stress (Sammons et al., 2011). Despite this, Gis2p has been found to localise to p bodies and stress granules following arsenite stress (Rojas et al. 2012). Additionally, vegetative *gis2*Δ cells do have a larger size, and *GIS2*-overexpressing cells have a slight growth defect (Jorgensen et al., 2002; Scherrer et al., 2011).

Identification of Gis2p RNA binding partners was previously performed from yeast grown in nutrient rich media using an RNA affinity purification approach, followed by incubation of Gis2p affinity purified RNAs and input RNAs with a DNA microarray (Scherrer et al., 2011). This methodology is likely to introduce some false negatives and false positives as Gis2p could associate with different mRNAs during the affinity purification procedure than it might *in vivo*. This is prevented in the iCLIP protocol, where UV-crosslinking is performed *in vivo* prior to cell lysis. Nevertheless, RNA-IP approach identified repeating GAN motifs as the main sequence feature of Gis2p binding sites. The analysis also suggested most of these GAN motif stretches occur in-frame and start from the first codon position in a transcript. Different categories of Gis2p-bound transcripts changed differently in response to *gis2*-Δ or *GIS2*-overexpression. Gis2p-bound nucleolar/rRNA biogenesis related transcripts increased in abundance in *gis2*-Δ and correspondingly decreased in abundance upon *GIS2*-overexpression. Bound transcripts related to chromatin architecture functions were upregulated upon *GIS2*-overexpression, whilst *GIS2* deletion had little effect. The authors additionally find a lot of overlap between Gis2p targets and the targets of other yeast RBPs, particularly ribosome-associated Gbp2p and nucleolar protein Nsr1p, so the relative contributions of Gis2p and other RBPs to yeast mRNA fate remains to be determined (Scherrer et al., 2011).

3.1.5 Aims of this chapter

In this chapter I analyse m6A miCLIP data produced at 4 hours into sporulation in WT vs. *ime4*-Δ genetic backgrounds to establish a high confidence set of m6A sites in early yeast meiosis. I then further analyse and integrate Pho92p 4TU-iCLIP and Gis2 iCLIP in WT vs. *ime4*-Δ yeast to determine the binding sites of these RBPs and whether their binding is dependent on Ime4p. I integrate the m6A miCLIP data to corroborate whether Ime4p-dependence means m6A-dependence. Finally I explore the RNA abundance and translational efficiency of these RBP targets to address fundamental questions about the function of m6A in yeast meiosis.

3.2 Materials and Methods

3.2.1 iCLIP/miCLIP pre-processing

iCLIP and miCLIP samples were prepared by Dr. Radhika Varier and Dr. Dora Sideri using the improved iCLIP protocol (F. C. Y. Lee et al., 2021). Mock miCLIP data was prepared by excluding the crosslinking step and subsequent IP with anti-m6A antibody. This results in a library similar to standard RNA-seq, however because it was generated using miCLIP-style library preparation any biases from this process should be represented. Note that this could be referred to as input, but I am hesitant to use this term as mock miCLIP is very different to input used in ChIP-seq, a technique CLIP is conceptually related to, where signal is usually weak due to IP with IgG and is used to determine binding regions that are false positives. In contrast, mock miCLIP datasets have coverage similar to RNA-seq and are used only for normalisation of CLIP signal to transcript abundance in this particular context. Note that both miCLIP and mock miCLIP were performed on poly(A)+ selected RNA.

Reads were demultiplexed using iCount demultiplex and subsequently trimmed for adapter sequences and also for PHRED score >30 using Trim Galore! (Krueger, 2017). Due to the high ncRNA content in miCLIP libraries, a sequential mapping strategy was used for all libraries, which is available as a Snakemake (Köster & Rahmann, 2012) pipeline from www.github.com/ulelab/ncawareclip. Mapping was first to representative *Saccharomyces cerevisiae* snRNA and rRNA sequences downloaded from NCBI (Pruitt et al., 2005), followed by mature tRNA sequences (3' CCA and 5' G added) and the SacCer3 mitochondrial chromosome before being mapped to the SK1 MvO genome (available from http://cbio.mskcc.org/public/SK1_MvO/).

PCR duplicates were removed using the unique molecular identifiers (UMIs). The start positions of uniquely mapping reads were taken as crosslinks. Detailed annotations were taken from (Chia et al., 2021). tRNA annotations were downloaded from UCSC table browser, which sources the annotations from GtRNAdb (Chan &

Lowe, 2016; Karolchik et al., 2004).

All genome browser screenshots display stranded crosslinks per million (CPM) normalised bigWig files. Crosslinks at each position were divided by the total number of genomic crosslinks in the sample multiplied by one million.

3.2.2 iCLIP/miCLIP Differential Analysis

Peaks were called on iCLIP and miCLIP data with Clippy v1.2.0 (https://github.com/ulelab/clippy/releases/tag/v1.2.0). Reproducible, high quality replicate samples were combined for peak calling and then individual sample coverage over these peaks was calculated using `bedtools map` (Quinlan, 2014). Peaks were filtered to have at least 5 cDNAs in 3 WT or 3 *ime4*-Δ replicates, to come to a preliminary list of binding sites.

To determine Ime4p-dependence, WT vs. *ime4*-Δ samples for Pho92p iCLIP, Gis2p iCLIP and m6A miCLIP were compared using DESeq2 (Love et al., 2014) whilst controlling for gene expression changes by including measurements from mock miCLIP samples as a contrast in the linear model. Genes with less than 20 cDNA counts across 3 replicates were discarded from the analysis. P values were calculated using a likelihood-ratio test. A stringent threshold of log2FoldChange $\leq$ -2 and adjusted p value of $<$ 0.001 was used to determine differentially bound sites for Pho92p, Gis2p iCLIP. For m6A miCLIP a threshold of log2FoldChange $\leq$ -1 and adjusted p value of $<$ 0.001 were used. Due to the high depth of the iCLIP datasets, sites were further filtered based on a DESeq2 base mean $>$ 200, which is a measure of coverage in both iCLIP and mock iCLIP samples. The value is calculated as the average of the normalized cDNA counts per peak from all samples, divided by their size factors.

Peak assignment to transcriptomic regions was performed using the following heirarchy to resolve any overlapping annotations: snoRNA > ncRNA > STOP codon > 3' UTR > 5'UTR > last 100nt CDS > first 100nt CDS > CDS > intergenic.

3.2.3 Published m6A sites definition

To come to a group of published m6A sites to compare our data against, I took all Mazter-Seq sites with confidence group > 1 and all m6A-Seq sites (Garcia-Campos et al., 2019; Schwartz et al., 2013). In order to robustly map these sites to the MvO SK1 genome assembly, I expanded all intervals to 150nt, retrieved the sequences and used BLAST to get mappings (Ye et al., 2006). I filtered mappings to those that were unique and in the case of Mazter-Seq, perfectly aligned to an “ACA” sequence. To create a consensus list, any m6A-seq region that overlapped with Mazter-Seq site(s) was removed, under the assumption that the signal was representing the sites detected at higher resolution by Mazter-Seq - although it is possible that adjacent non-ACA sequence context m6A sites would not be represented in the list. This procedure resulted in a list of 1297 m6A regions.

3.2.4 Distance to nearest m6A

Distances between peak sets were calculated using `bedtools closest` with parameters `-s -t first -d` (Quinlan, 2014).

3.2.5 GO term enrichment

Gene list enrichment analysis was performed using YeastEnrichr, specifically using KEGG 2019 pathways (https://maayanlab.cloud/YeastEnrichr/) (H. C. Chen et al., 2013; Kuleshov et al., 2016).

3.2.6 Motif analysis

Motifs were discovered in peak regions by resizing all peaks to 100nt and obtaining their fasta sequences to submit to STREME (Bailey, 2021). Either shuffled sequences or another peak set were used as background, as indicated in the main text. Motifs were plotted around the centre of peaks using a custom script available upon request.

3.2.7 Metaprofiles

Bigwigs were generated from bedgraphs using UCSC BedGraphToBigwig (Kuhn et al., 2013) and metaprofiles were generated around regions of interest using DeepTools ComputeMatrix and PlotHeatmap (Ramırez et al., 2014). Further integration and plotting was performed using R with ggplot2, dplyr, data.table, stringr and cowplot packages (Dowle et al., 2019; Wickham, 2010, 2016; Wickham et al., 2015; Wilke et al., 2019).

3.3 Results

3.3.1 iCLIP/miCLIP quality control

Replicates generally had good correlation as measured by Pearson's correlation, scatter plots and Principal Component Analysis (PCA) performed at the peak level (Figure 3.4). Additionally, (m)iCLIP library sizes generally exceeded 1 million cDNAs (Figure 3.5). Poly(A) selection in the miCLIP and mock miCLIP libraries appears to have worked in enriching for non-rRNA RNAs (Figure 3.5 A,D). Notably, smaller Pho92p iCLIP libraries, which we presume to be of lower quality, have a higher proportion of non-coding and mitochondrial RNAs present, suggesting that this signal is likely to be noise. The decision to exclude some replicates (indicated in red) was based on a combination of the samples having fewer crosslinks compared to other replicates (any sample with fewer than 1 million cDNAs was removed), indicating variation in library preparation, and also clustering further away from replicates in PCA and/or poorer correlation to other replicates.

It is interesting that the proportion of variance accounted for by genetic background was much less for Gis2p (34%) than for the miCLIP (80%) or Pho92p (58%).

3.3.2 Differential CLIP analysis

Having chosen reproducible replicates, I wanted to determine the Ime4p-dependence in all experiments. This required first defining binding sites, and then calculating differential enrichment between WT and *ime4*-Δ conditions. It was important to account for changes in gene expression between the two conditions, otherwise it is impossible to distinguish loss of binding from reduced RNA abundance and *vice versa*. Previously, differential iCLIP analysis has been performed using DESeq2 and including expression data as a contrast in the linear model (Zarnack et al., 2013). This method has also been used in differential analysis of meRIP-Seq (McIntyre et al., 2020). Several packages have emerged to specifically address such differential analysis in meRIP-Seq, which is conceptually similar to

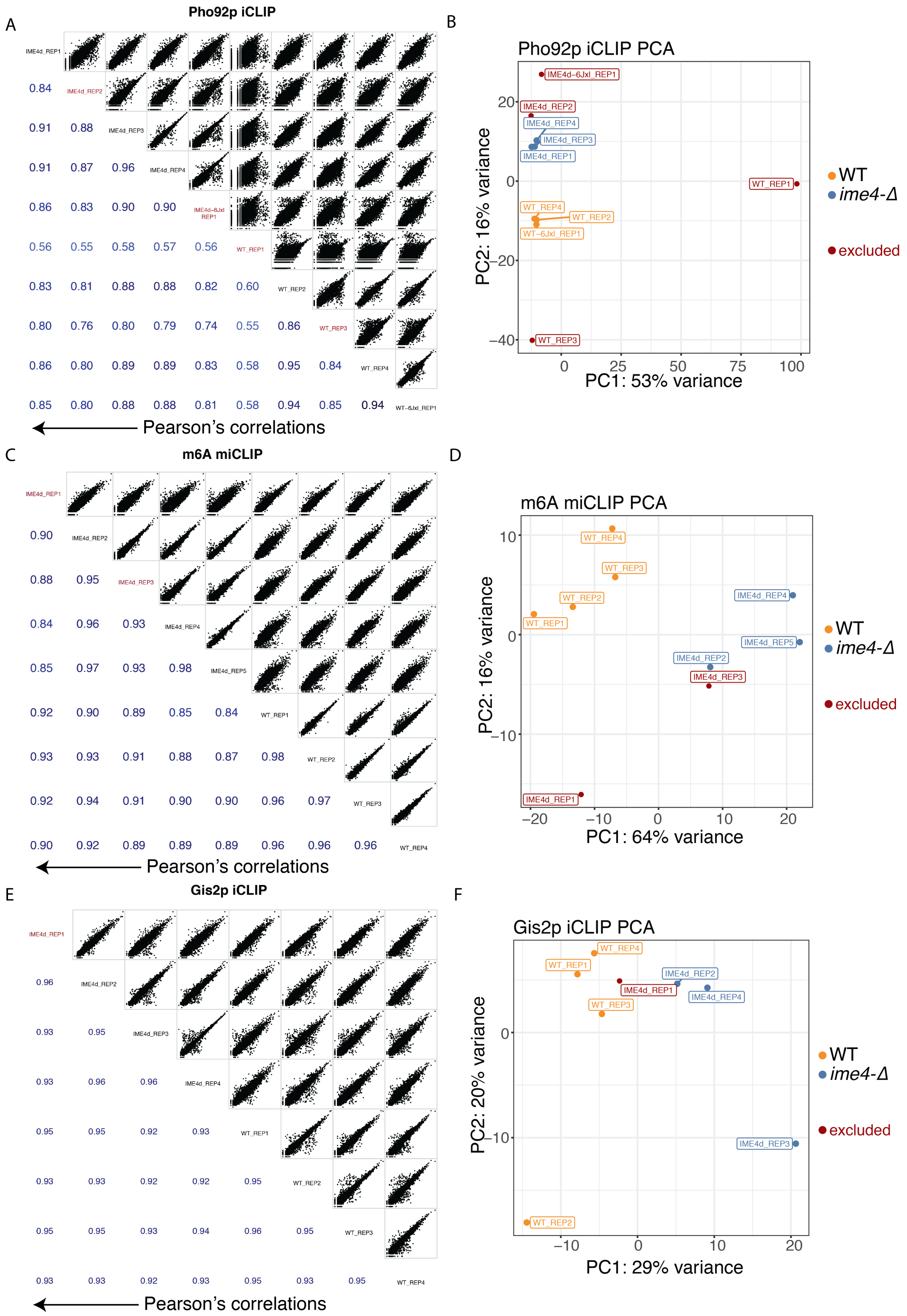


Figure 3.4: Correlation between iCLIP and miCLIP replicate samples, quantified at the peak level. A) On the left, a correlation matrix showing all Pho92p iCLIP experiments plotted against each other, in the top right quadrant as scatter graphs and showing Pearson's correlations in the bottom left. On the right, a PCA plot with points coloured by yeast genotype and whether the replicate was eventually excluded from further analysis. B) As in A but for m6A miCLIP. C) As in A but for Gis2p iCLIP.

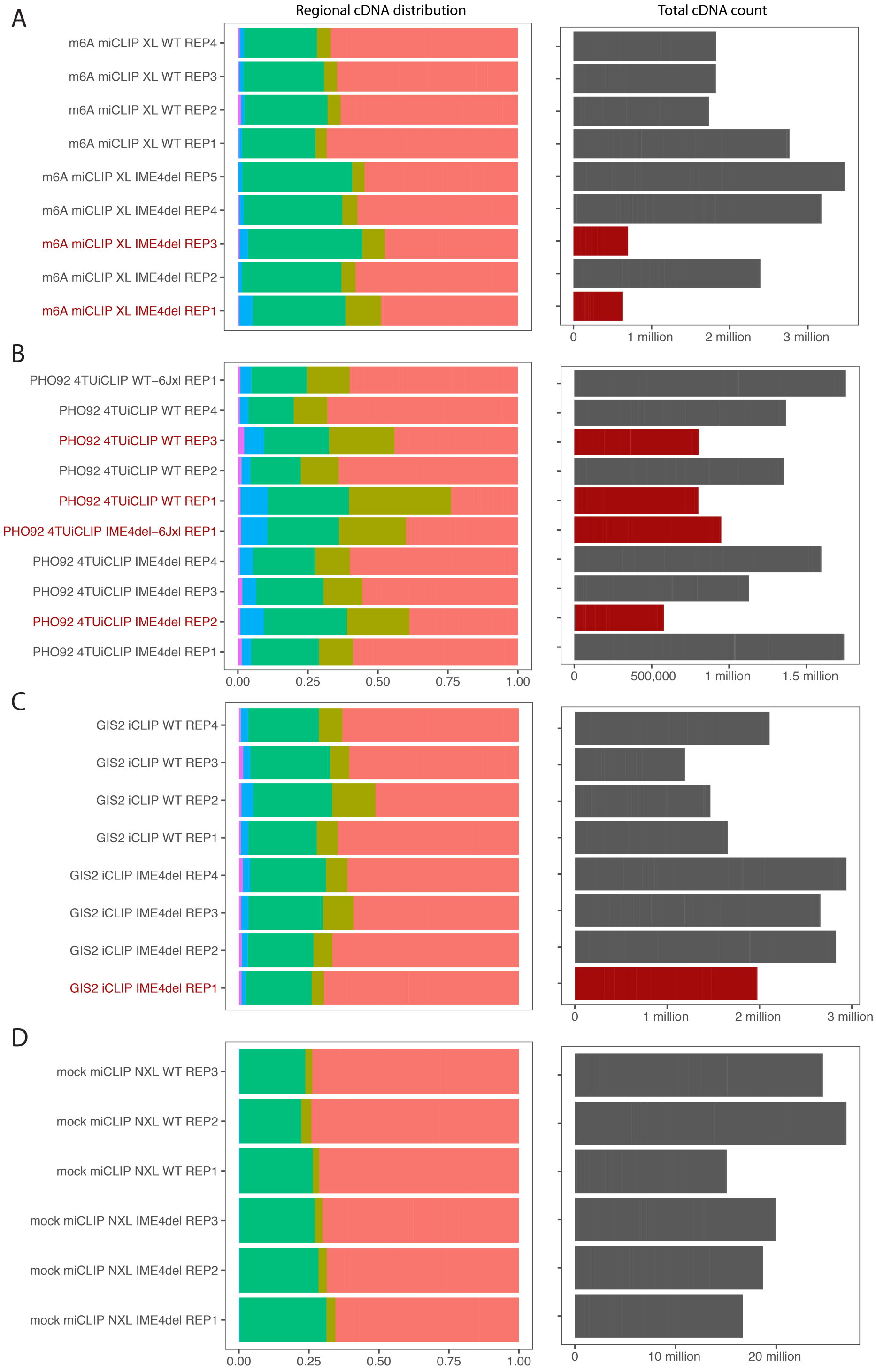


Figure 3.5: cDNA count and library distributions for iCLIP and miCLIP. A) On the left, regional distributions are shown for mapped cDNAs from miCLIP libraries for different transcript types; on the right, the total cDNA count for each library is displayed. Red colour indicates those replicates excluded from further analysis. B) As in A but for Pho92p iCLIP libraries. C) As in A but for Gis2p iCLIP libraries. D) As in A but for mock miCLIP libraries.

the current analysis, the most up to date being the Quad Negative Binomial (QNB) package (N. Liu et al., 2017). Even though DESeq2 does use the negative binomial model to estimate over-dispersion, this estimate is based only on the (m)iCLIP samples in this use case and not the control expression data, in this case the mock miCLIP. In QNB, the over-dispersion in (m)iCLIP samples and mock control samples for each condition is separately accounted for. Therefore there were several decisions to make regarding how to approach the differential binding/methylation analysis.

I examined QNB vs. DESeq2 approaches and whether to use mock counts summarised at the gene or peak level. I hypothesised that more differential sites may be called when using gene level mock counts due to reduced variation compared to peak level summarised counts. I found that the DESeq2 approach consistently called more Ime4p-dependent peaks than QNB, and that using gene level summarised mock counts vs. peak level counts doesn't make much of a difference for this specific dataset (in this analysis defined by log2 fold change $<$ -1 and p value $<$ 0.05) (Figure 3.6A).

An increase in called peaks will be a mix of true and false positives. I wanted to evaluate the likely ratios between the two. Due to a lack of ground truth I decided to examine the proportion of RGAC motif-containing peaks in both Ime4p-dependent and independent categories, using RGAC motif occurrence as a proxy for a ”true” Ime4p-dependent Pho92p peak (Figure 3.6B).

Whilst QNB did produce a higher proportion of RGAC-containing Ime4p-dependent peaks (40% vs. 36% for the gene level analysis), it also produced nearly 3X fewer Ime4p-dependent peaks (349 vs. 956), therefore on balance I chose to procede with DESeq2. I chose to use gene level mock counts because it seemed to make more sense conceptually, we are interested after all in the level of gene expression, not really mock signal at a given position in the gene - however based on these results I could have chosen peak level counts and I would have obtained similar results.

Analysis using this strategy revealed that 642 Pho92p peaks in 507 genes

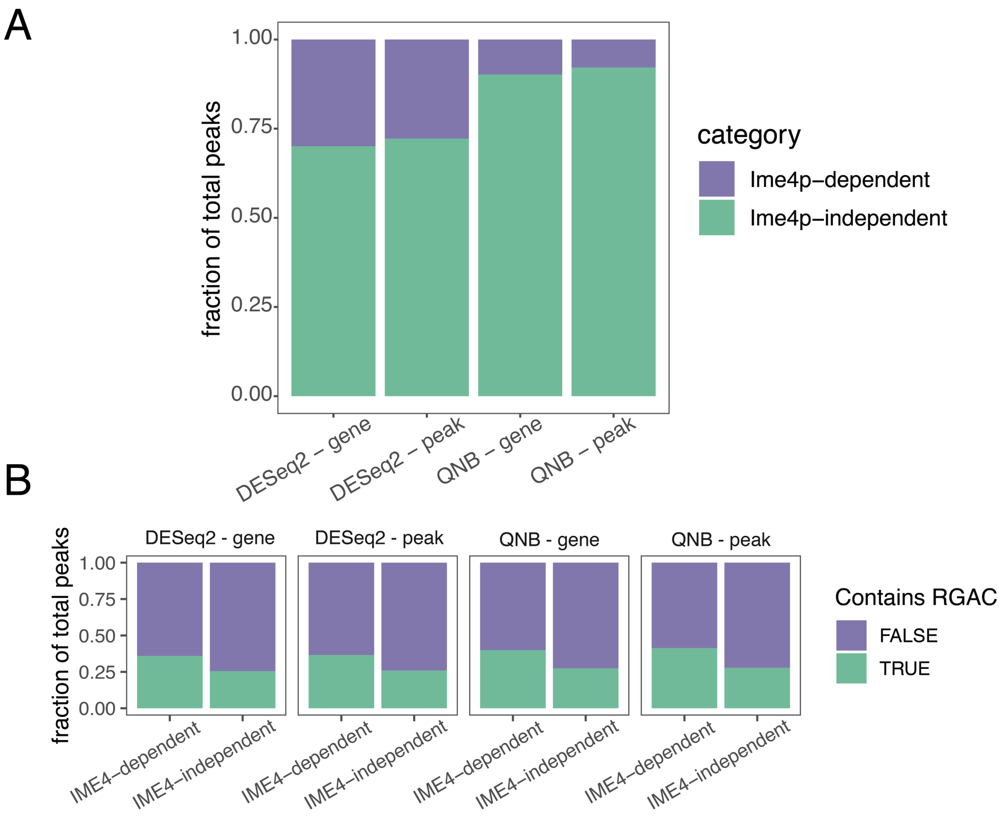


Figure 3.6: Comparison of QNB vs. DESeq2 for differential CLIP analysis. A) Fraction of total peaks called for each differential analysis strategy that were defined as Ime4p-dependent with criteria log2 fold change < -1 and p value < 0.05. Gene and peak refer to the level at which mock miCLIP data was quantified as an expression control. B) Proportion of total peaks retrieved in each differential analysis strategy which contained an RGAC motif, with peaks split up into Ime4p-dependent and independent as defined in A.

(16.7% of all detected peaks) were found to be reliably reduced in *ime4*-Δ (log2FoldChange <= -2, adjusted p value < 0.001) (Figure 3.7A). Using a less stringent criteria I could designate up to 30% (1130/3823 peaks at log2FoldChange <= 0, adjusted p value < 0.05) of detected Pho92p peaks as reduced in *ime4*-Δ; however, for subsequent analysis, I will refer to the 642 stringently defined peaks as "Ime4p-dependent" Pho92p binding sites. This means that surprisingly, a large subset of Pho92p binding sites do not decrease in *ime4*-Δ cells, indicating that Pho92p can also bind transcripts in an Ime4p independent manner. In contrast, analysis of Gis2p iCLIP revealed 3563 peaks, of which only 43 peaks in 24 genes were decreased in *ime4*-Δ representing only 1.2% of all Gis2p peaks (log2FoldChange <= -2, adjusted p value < 0.001) (Figure 3.7B).

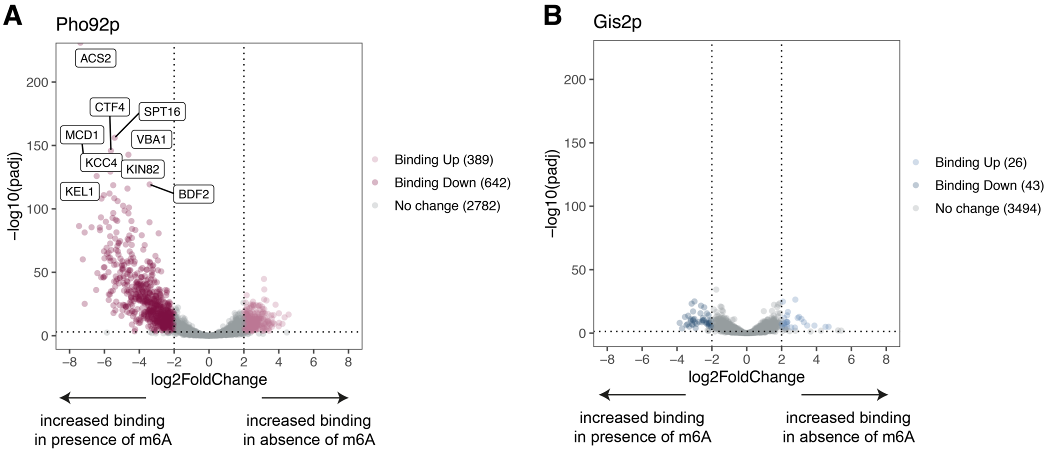


Figure 3.7: Volcano plots describing differential iCLIP analysis. A) Pho92p iCLIP peak volcano plot, showing log2 fold change in WT vs. *ime4*-Δ conditions on the x axis and -log10(adjusted p value) on the y axis. Dotted lines denote the log2FoldChange <= -2, adjusted p value < 0.001, threshold. Peaks with increased signal in *ime4*-Δ are coloured in light pink, whereas those that decrease are coloured in dark magenta. Identities of genes harbouring several of the top downregulated binding sites are indicated with labels. B) As in A but for Gis2p. The few up and downregulated binding sites are coloured in light and dark blue respectively.

The most differentially bound transcript by Pho92p was *ACS2* mRNA, which encodes acetyl-coenzyme A synthetase 2 (Figure 3.7A). This is an 'anaerobic' isozyme of acetyl-coenzyme A synthetase, which is required for glucose metabolism, which is limited during yeast meiosis, and so potentially m6A-bound Pho92p plays a role in regulating the expression of this enzyme in early meiosis (van den Berg et al., 1996). Other Ime4p-dependent Pho92p binding partners included *CTF4* and *MCD1* mRNA, which both encode proteins important for sister chromatid cohesion. Mcd1p is an essential subunit of the cohesin complex and Ctf4p is a protein suggested to link DNA replication to sister chromatid cohesion as it is also responsible for recruiting proteins to replication forks (Guacci et al., 1997; Hanna et al., 2001). The timing of m6A deposition, coinciding with DNA replication, suggests these mRNAs could be very relevant meiotic targets of Pho92p regulation.

I wanted to be confident that the peak categories defined by the DESeq2 analysis reflected the raw data well, and so I plotted crosslink per million (CPM) normalised representative samples of all (m)iCLIPs against their designated categories (Figure 3.8). Reassuringly, for Pho92p and miCLIP Ime4p-dependent sites, there is enriched WT iCLIP signal at Ime4p-dependent peaks and very little signal in *ime4*-Δ (m)iCLIP samples (Figure 3.8 A,B). Gis2p signal also shows the same pattern, however the scale of reduction in signal between WT and *ime4*-Δ is much less dramatic (Figure 3.8 C). I also plotted mock miCLIP signal for the same peak categories (Figure 3.8 D-F). Interestingly, the genes containing Ime4p-dependent peaks seem to be generally upregulated in *ime4*-Δ. It is difficult to discern if this is a statistical issue - ie. if the fold changes are more extremely opposite in CLIP vs. mock is an Ime4p-dependent peak more likely to be called? Or whether this in fact represents a biological truth - that transcripts containing m6A, Ime4p-dependent Pho92 or Gis2 binding are generally upregulated in *ime4*-Δ. It is likely to be a mixture of both.

3.3.3 Analysis of m6A miCLIP data

In order to determine if Ime4p-dependence is a suitable proxy for m6A-dependence, I needed to process our miCLIP data to come to a list of m6A sites.

After peak calling, I filtered miCLIP peaks to have $>$ 5 counts across all 3 WT miCLIP replicates, this left me with 9519 miCLIP peaks which I took forward for DESeq2 analysis. I expected true m6A sites to decrease in signal in *ime4*-Δ. I found 1286 miCLIP peaks in 870 genes that were reduced in *ime4*-Δ cells (log2FoldChange $<=$ -1, adjusted p value $<$ 0.001). To help me to understand what thresholds to use to find true m6A sites, I examined the enrichment for RGAC motifs within the peaks, reasoning that true m6A sites would be contained within the canonical motif. Visualising as a volcano plot, it is reassuring that downregulated miCLIP peaks also appear to contain more RGAC motifs (Figure 3.9A).

I followed this by testing for RGAC enrichment given different stringencies of thresholding the miCLIP (Figure 3.9B). I calculated the background probability of detecting an RGAC as follows:

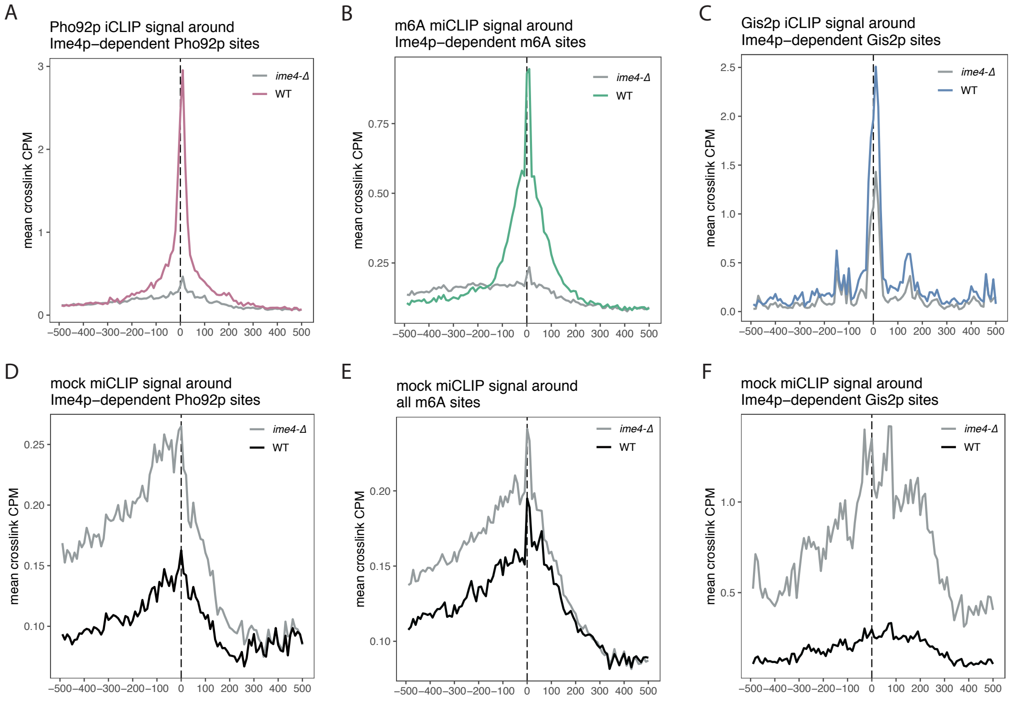


Figure 3.8: Metaprofiles of raw iCLIP signal over categories defined by DESeq2 analysis. A) Crosslinks per million normalised representative Pho92p samples are plotted around Ime4p-dependent Pho92p binding sites as characterised by DESeq2 analysis, pink denotes the WT sample and grey denotes the *ime4*-Δ sample. B) As in A but for representative m6A miCLIP samples over Ime4p-dependent m6A sites, green denotes the WT sample and grey denotes the *ime4*-Δ sample. C) As in A but for representative Gis2p iCLIP samples over Ime4p-dependent Gis2p binding sites, blue denotes the WT sample and grey denotes the *ime4*-Δ sample. D-F) Crosslinks per million normalised representative WT and *ime4*-Δ mock miCLIP samples effectively showing gene expression over the corresponding binding site categories: D) Ime4p-dependent Pho92p binding sites, E) Ime4p-dependent m6A sites and F) Ime4p-dependent Gis2p sites.

$$P(RGAC) = F * (m - (k - 1))$$

where F = the proportion of RGAC kmers found in all SK1 gene sequences, m = the median peak length of m6A miCLIP peaks and k = the length of the kmer, in this case 4. To calculate F I used the EMBOSS compseq tool (http://emboss .bioinformatics.nl/cgi-bin/emboss/compseq)(Rice et al., 2000). I was interested to compare this to the enrichment of RGAC in Pho92p Ime4p-dependent and independent peaks and also the Gis2p data.

Several interesting things emerge from this analysis. First, whilst increasing stringency of thresholding improves the proportion of m6A miCLIP peaks containing RGAC, this hits a plateau, as there is no increase between ”high” and ”highest” categories. Second, Pho92p Ime4p-dependent peaks have a higher RGAC enrichment than the highest stringency m6A miCLIP thresholding. I suspect this is because the Pho92p data is of higher quality - we suspect more noise from the m6A antibody. Third, Gis2p shows a very small enrichment for RGAC, which while significant by chi-square test (p <0.001), the magnitude of enrichment suggests it could be within the margin of error of my background estimate, which is a simplification. Fourth, the m6A miCLIP data shows RGAC enrichment even with no filtering for Ime4p-dependency which I interpret to mean that there are probably true m6A sites in even the unfiltered category, which become false negatives in our analysis. Finally, it appears that the thresholding for Pho92p Ime4p-dependent sites is likely set in a good place, because Ime4p-independent sites have a similar RGAC enrichment to the expected background level. Note the chi-square test (p <0.001) even suggests mild depletion of RGAC in the Ime4p-independent sites, but again this is likely within the margin of error of my background estimate.

I continued to study the positional enrichment of m6A miCLIP peaks. I plotted the profile of high confidence Ime4p-dependent miCLIP peaks vs. Ime4p-independent over genes (Figure 3.10A). The Ime4p-dependent peaks were more 3’ biased, which is expected based on the literature. The m6A antibodies are notoriously noisy, and so extensive Ime4p-independent signal is likely to represent bind-

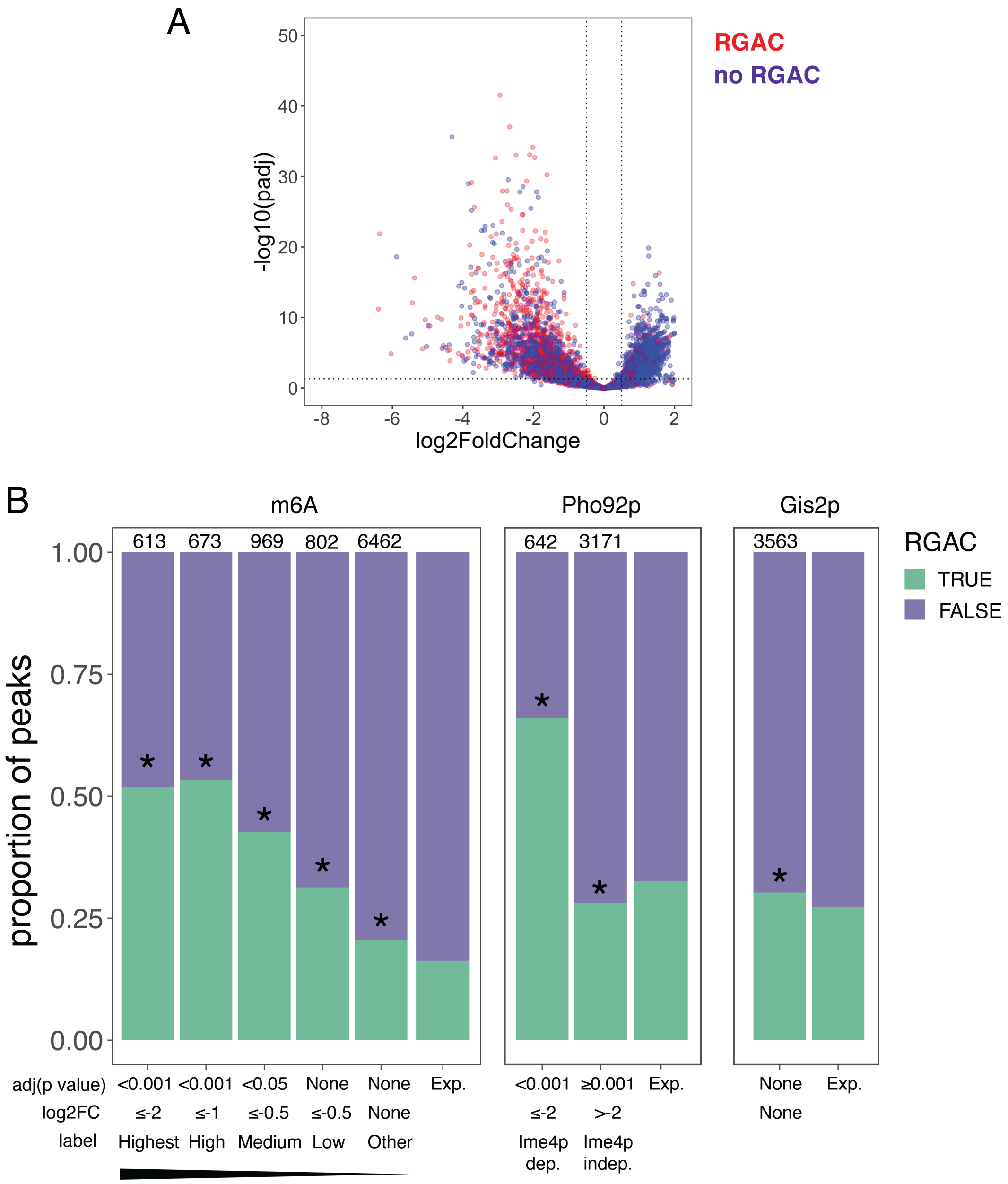


Figure 3.9: Testing thresholding of miCLIP data. A) A volcano plot of m6A miCLIP data analysed with DESeq2. The x axis displays log2 fold change WT vs. *ime4*-Δ samples and the y axis shows -log10(adjusted p value). Each dot represents a miCLIP peak, where red means the peak contains at least one RGAC motif and purple means there is not RGAC motif. Dotted lines represent "medium" miCLIP thresholding with log2 fold change $\leq$ 0.5 and adjusted p value $<$ 0.05. B) Stacked bar graphs showing the proportion of peaks containing RGAC at different levels of thresholding for m6A miCLIP, Pho92p and Gis2p iCLIP. The numbers at the top of the bars denote the number of peaks in the given category. "Exp." bars denote calculated expected RGAC enrichment based on frequency of RGAC in SK1 transcripts.

ing at non-m6A sites, such as sites with poly(A) stretches. Next, I assigned all peaks to a specific transcriptomic region, finding nearly 50% of Ime4p-independent binding at the 3' end of transcripts, but no bias towards overlapping the CDS or 3'UTR (Figure 3.10B). As previously reported in mammals, there was an enrichment for peaks overlapping the STOP codon. The STOP codon itself could hypothetically be m6A-modified if it were UGA followed by a cytosine (UAG would not be modified, UAA is possible, but not canonical). To investigate this, I took the sequences of STOP codons +1 nucleotide that overlapped either an Ime4p-dependent (130 STOP codons) or independent miCLIP peaks (342 STOP codons), I found only 6 instances of GAC motif in Ime4p-dependent peaks and 5 instances in the independent peaks (Figure 3.10C). This suggests that whilst a UGA STOP followed by a cytosine is enriched in Ime4p-dependent miCLIP peaks, the number is so few that it's unlikely that the STOP codon itself being methylated is of importance, but more the general 3' end of the transcript region.

I further sought to determine if Ime4p-dependent miCLIP peaks would contain the RGAC motif if I performed a search *de novo*. I used STREME to search for enriched motifs in Ime4p-dependent miCLIP peak sequences resized to 100nt intervals vs. shuffled dinucleotide-content-matched background sequences (Figure 3.10D). The top enriched motif was the canonical RGAC motif, as previously identified. Further, the RGAC was contained within an extended sequence context of -4 A and +4 U which validates previous findings (Schwartz et al., 2013). Of the Ime4p-dependent miCLIP peak sequences containing such a motif, ∼80% contained one, suggesting they contain one m6A, whereas ∼20% contained two (Figure 3.10E). The Ime4p-dependent miCLIP peaks range from 11—167 nucleotides in width, with 77% being under 50nt (Figure 3.10F). I was curious to see how the RGAC motif was positioned around peaks and so I plotted RGAC motif enrichment around peak centers (Figure 3.10G). Whilst there was enrichment at peak centers, the RGAC motif could also be further away, indicating it could be adjacent to the miCLIP peak also, which I observed while studying genome browser tracks. I hypothesised this is due to the m6A antibody crosslinking adjacent to m6A sites where

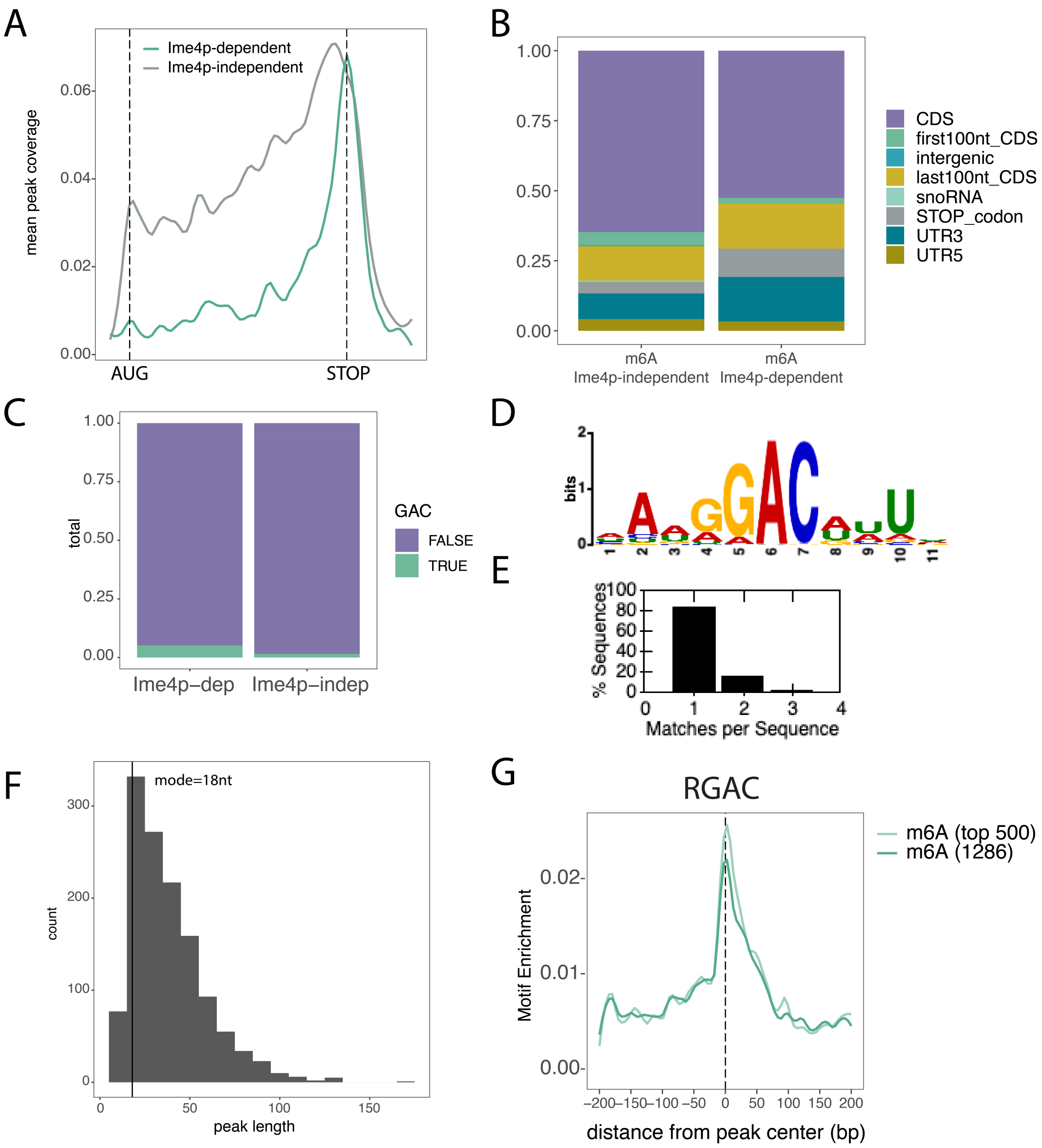


Figure 3.10: Positional and motif enrichment of miCLIP data. A) Metagene profile of miCLIP peaks over yeast genes, green indicates Ime4p-dependent peaks and grey indicates Ime4p-independent. B) Stacked bar graph showing peak annotations to different transcriptomic regions. C) Stacked bar graph showing proportion of STOP codons overlapped by a miCLIP peak that are encoded by UGA followed by a C (in green). D) Sequence logo for top enriched Ime4p-dependent m6A motif. E) Proportion of Ime4p-dependent m6A sites containing the motif in D, which have 1 or more matches per peak sequence. F) Histogram distribution of miCLIP peak lengths. G) Enrichment of RGAC motif around Ime4p-dependent m6A sites (dark green) and the top 500 of these (light green).

there is a more favourable sequence for crosslinking, for example a poly(U) stretch.

Next I wanted to check if I identified m6A sites previously reported in the literature. *NAM8* and *RAD54* transcripts were both validated to have m6A sites by m6A-Seq (Schwartz et al., 2013). The three sites across the two genes are also found by our miCLIP (Figure 3.11,3.12). Interestingly *NAM8* mRNA is also bound by Gis2p at the 5' end, and upon further inspection there appears to be Ime4p-dependent Pho92p binding at the *NAM8* mRNA m6A site, although this wasn't captured in the full transcriptomic analysis due to stringent thresholding (Figure 3.11C).

Ideally, I would like to resolve the m6A sites to single nucleotide resolution. However, the raw crosslinking data is not easy to parse in this regard. For example, in *NAM8* mRNA m6A peak region the actual RGAC motif has much fewer crosslinks than other sequences within the peak (Figure 3.11B). The 5' most dashed box region of crosslinking may be explained by the multiple trinucleotide U's, the most 3' dashed region also contains a U trinucleotide. The middle dashed box is less obvious, but I suspect this might be due to the m6A antibody binding at poly(A) stretches with higher affinity than m6A itself. Therefore, I decided to proceed with peak regions at lower resolution to integrate with the iCLIP data; feeling confident that due to all the metrics previously described, they were likely to represent true m6A sites in the near vicinity.

After having validated the previously reported *NAM8* and *RAD54* transcript m6As, I wanted to investigate the sites in *IME1*, *IME2* and *IME4*. As previously described, the *IME2* site was identified in (Bodi et al., 2010) and validated by (Schwartz et al., 2013), whereas *IME1* and *IME4* sites were identified by (Bodi et al., 2010) but not found by (Schwartz et al., 2013).

Firstly, we did identify the m6A site in *IME2* mRNA, falling into the medium confidence category (log2FC=-0.6, padj= 0.007). *IME1* mRNA was very interesting, there was a clear Ime4p-dependent Pho92p site, however we did not detect an m6A. Looking at the genome browser tracks it does seem possible that there is an m6A in that position. The peak was quantified in my analysis, but after correcting

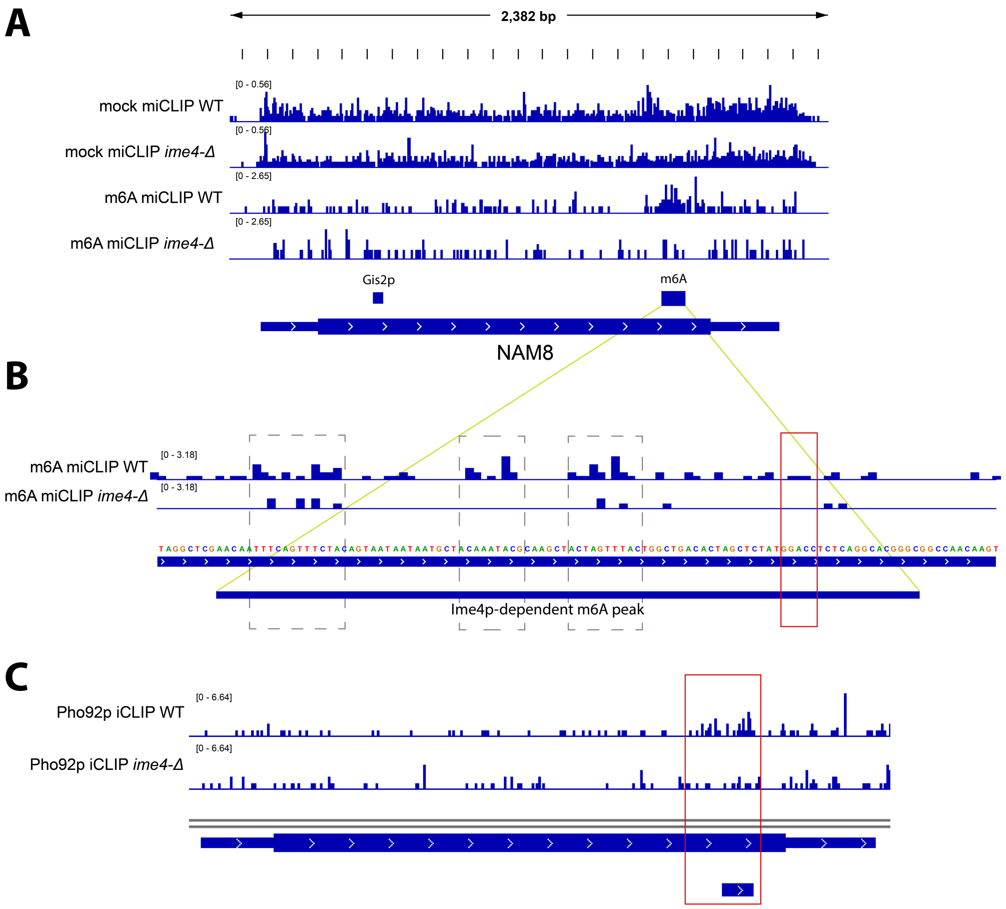


Figure 3.11: Detailed study of the *NAM8* gene. A) Genome browser screenshot showing miCLIP and mock miCLIP representative sample crosslinks over the *NAM8* gene. Scale refers to crosslinks per million. Gis2p binding site is indicated and high confidence Ime4p-dependent m6A site. B) Zoom-in on the Ime4p-dependent m6A site showing miCLIP crosslinks and the sequence at that region. Highlighted in orange is the m6A consensus motif previously reported to be modified. Other areas of high miCLIP crosslink signal in the WT sample are indicated with dashed grey boxes. C) Representative Pho92p iCLIP samples are shown over the *NAM8* gene, Pho92p crosslinks that look to indicate Ime4p-dependent binding are highlighted with an orange box and the m6A site is shown below with a blue box.

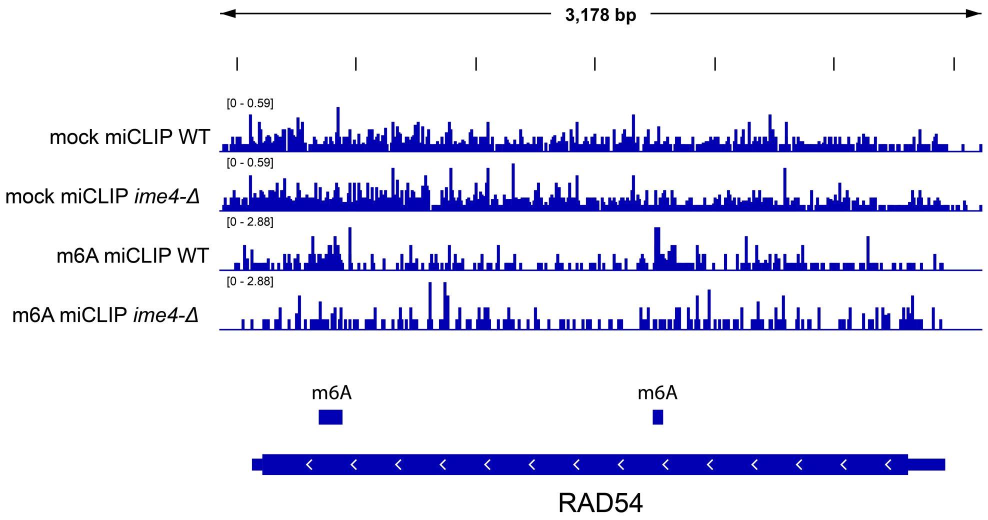


Figure 3.12: m6A sites in *RAD54* mRNA. Genome browser screenshot showing miCLIP and mock miCLIP representative sample crosslinks over the *RAD54* gene. Scale refers to crosslinks per million. High confidence Ime4p-dependent m6A sites are shown below with blue boxes.

for the change in RNA abundance of *IME1* the actual quantified log2fold change was 0.15—a small increase in *ime4*-Δ. Also interesting, is that by examining the Gis2p crosslinks it looks like there could be a Gis2p binding site that only occurs when Pho92p is absent - suggestive of competitive binding of the two proteins. Again, this Gis2p binding site was not quantified by my analysis because overall the Gis2p signal over the gene was very low. This case study highlights the limits of detection in such an analysis; when both the RNA abundance and (m)iCLIP signal decreases, it is hard to detect responsiveness to the condition. Likely, this is exactly the same reason why (Schwartz et al., 2013) failed to identify the *IME1* m6A site in their bioinformatics analysis, whilst it was detected by targeted m6A IP followed by RT-qPCR (Bodi et al., 2010). To further validate that there was in fact an m6A site in this region, I used Nanocompore data (Leger, Amaral, Pandolfini, Capitanchik, et al., 2019b). I found that there was a high scoring kmer in an AGACU sequence context within this region, suggesting that it is indeed a bonafide m6A site. Note, that due to MAZTER-Seq only detecting m6A in an ACA sequence context the *IME1* site would also not be detected by MAZTER-Seq (Figure 3.13).

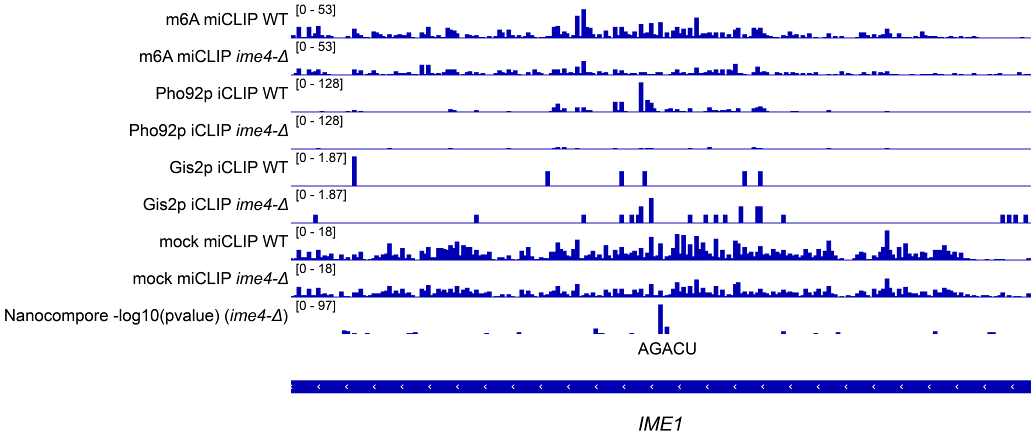


Figure 3.13: A detailed study of *IME1*. Genome browser screenshot displaying m6A miCLIP, Pho92p and Gis2p iCLIP crosslinks per million over *IME1*. An Ime4p-dependent Pho92p binding site is present, mock miCLIP data shows expression of *IME1* in WT vs. *ime4*-Δ conditions. The bottom track shows Nanocompore signal as -log10(p value), the consensus m6A motif AGACU is indicated below.

I proceeded to overlap all our miCLIP-defined m6A sites with a curated list of published m6A sites collated from m6A-Seq and MAZTER-Seq (See Methods section 3.2.3, (Garcia-Campos et al., 2019; Schwartz et al., 2013)). Overall, the majority of m6A sites which did overlap, did so either exactly, or within 50nt (Figure 3.14A). 40% of published sites are within 50nt of a medium confidence m6A miCLIP site. (Figure 3.14B). Taking all miCLIP peaks submitted to DESeq2 analysis, we could capture a maximum 55% of published sites (Figure 3.14C).

Whilst the overlap was good, it was lower than I expected. I wondered if one reason behind this could be that the gene expression between the different experiments: our present work, m6A-Seq and MAZTER-Seq, might be different. The protocol used in the present work to synchronise meiosis - induction of *IME1* from the *CUP1* promoter, produces a much more synchronised cell population than the method used in the m6A-Seq and MAZTER-Seq papers.

To investigate this possibility I investigated the expression levels of the three categories of genes in our WT mock miCLIP data and reciprocally in WT input sequencing data from (Schwartz et al., 2013), that is: genes identified as m6A-modified in our miCLIP data alone, genes identified as m6A-modified in both our

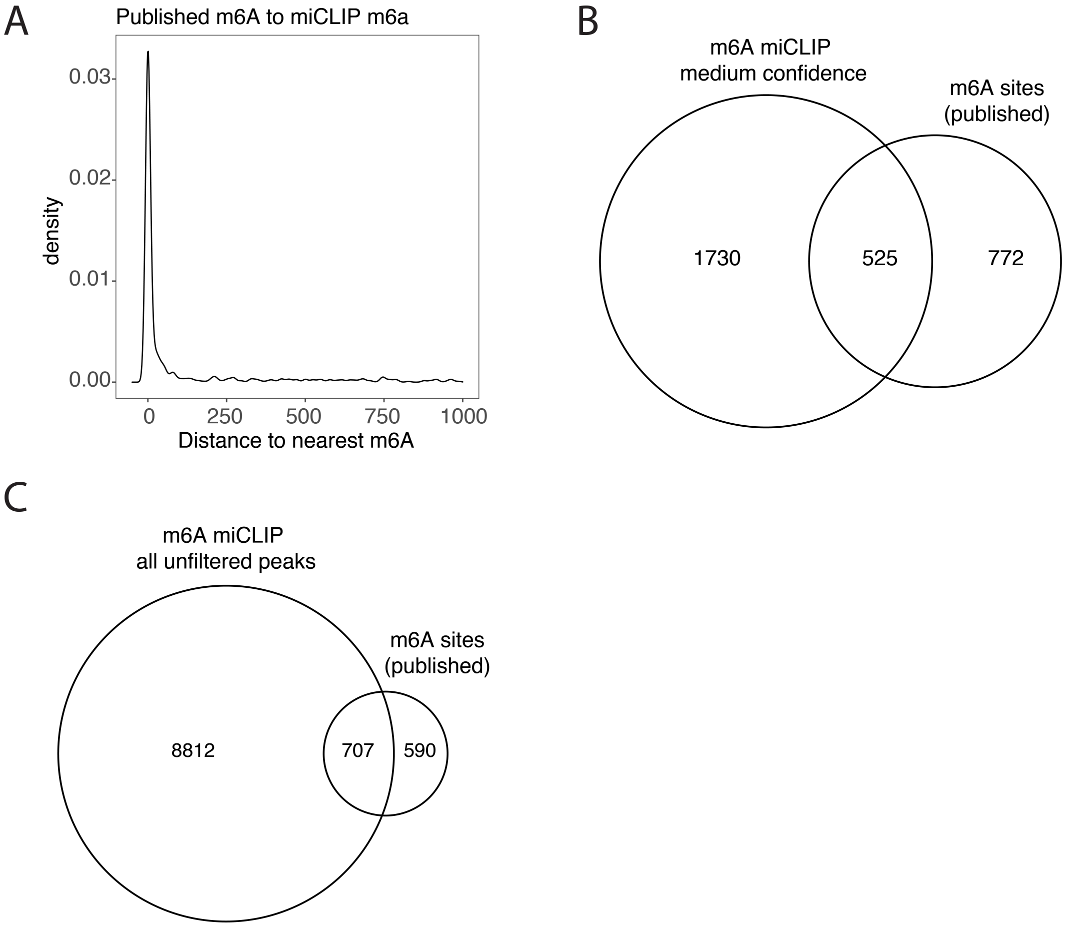


Figure 3.14: Overall agreement between miCLIP m6A sites and published sites. A) Density plot of distance from published m6A site to nearest miCLIP-defined m6A site. B) Venn diagram showing overlap between medium confidence m6A miCLIP sites and published m6A sites. c) Venn diagram showing overlap between all m6A miCLIP peaks and published m6A sites.

miCLIP and published data, and genes that are identified only in the published data. I was surprised to find that in fact, genes that were identified in published data alone were more lowly expressed in both expression data from the present project, and the matched published expression data (Figure 3.15).

This could be caused by discrepancies in data processing, for example, being more stringent in my thresholding from the beginning. Alternatively, this could be experimental - perhaps the stringency of the miCLIP experimental method, with UV crosslinking and harsh washes, biases data towards detecting more highly expressed genes compared to m6A-Seq or MAZTER-Seq. However, this wouldn't explain

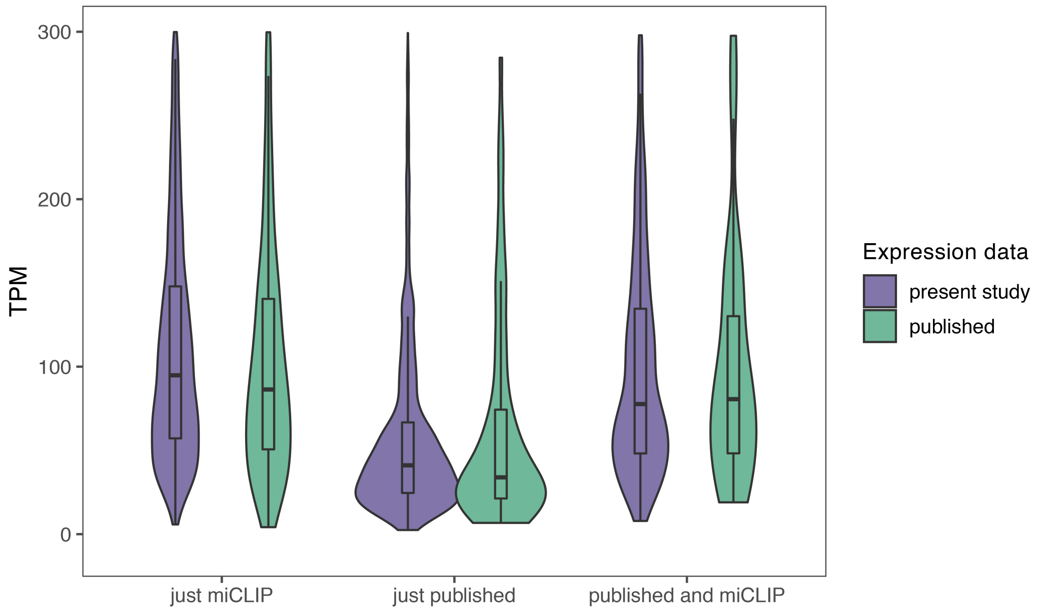


Figure 3.15: Expression of genes containing published and/or miCLIP-defined m6A sites. Violin plot with box plot overlay showing the normalised abundance (TPM) of transcripts containing m6A defined by just miCLIP, just the published data or by both datasets. The expression data is divided into WT data from the present study (purple) and from the m6A-Seq paper (green).

why the sites found only in miCLIP data, residing on more highly expressed genes, were not identified in the published work.

Another possiblity is that the higher expression sites only found in miCLIP are more likely to be noise—to investigate, I looked at the RGAC enrichment in peaks found across expression deciles in the miCLIP data. In the raw m6A miCLIP peaks, as the expression decile increases so do the proportion of peaks without RGAC; however, this mostly seems to be corrected by the thresholding for Ime4p-dependence (Figure 3.16).

Our miCLIP data might also differ from the published m6A-Seq because the Abcam polyclonal m6A antibody (ab151230) was used as opposed to the Synaptic Systems antibody used in the published work.

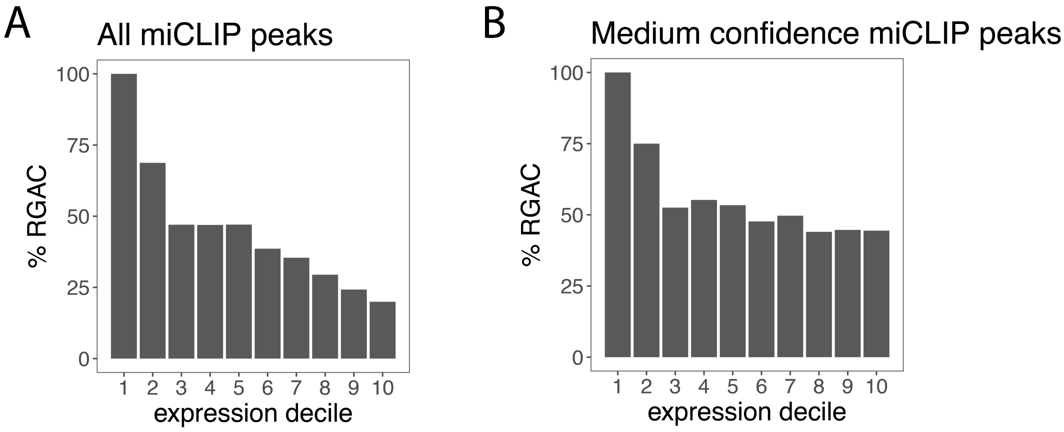


Figure 3.16: RGAC content of miCLIP peaks stratified by transcript expression decile. A) Transcripts containing at least one m6A miCLIP peak were stratified by expression decile, then the proportion of RGAC-containing peaks were calculated within the deciles. B) The same as A but for medium confidence m6A miCLIP peaks.

3.3.4 Pho92p is an m6A-reader, whilst Gis2p is not

Having established our miCLIP data to be of sufficient quality, I returned to the Pho92p and Gis2p data, to scrutinise the m6A-dependency of their binding to RNA. I began by calculating the distance between Pho92p or Gis2p binding sites to the nearest m6A. It is clear that most Pho92p sites that reduce in binding in *ime4*-Δ, either directly overlap with an m6A site, or are very near to one. In contrast those Pho92p sites that increase binding, are widely distributed with respect to m6A, regardless of whether miCLIP-defined m6A sites or published m6A sites are used (Figure 3.17A,B). Gis2p binding sites have a less clear relationship to m6A positions - important in the interpretation is that the categories of differential binding are much smaller for Gis2p.

Having concluded both that: a) there are few Gis2p binding sites which decrease in signal upon Ime4p deletion and b) the few that do decrease do not overlap with m6A sites, I felt confident that Gis2p is not an m6A-binding protein.

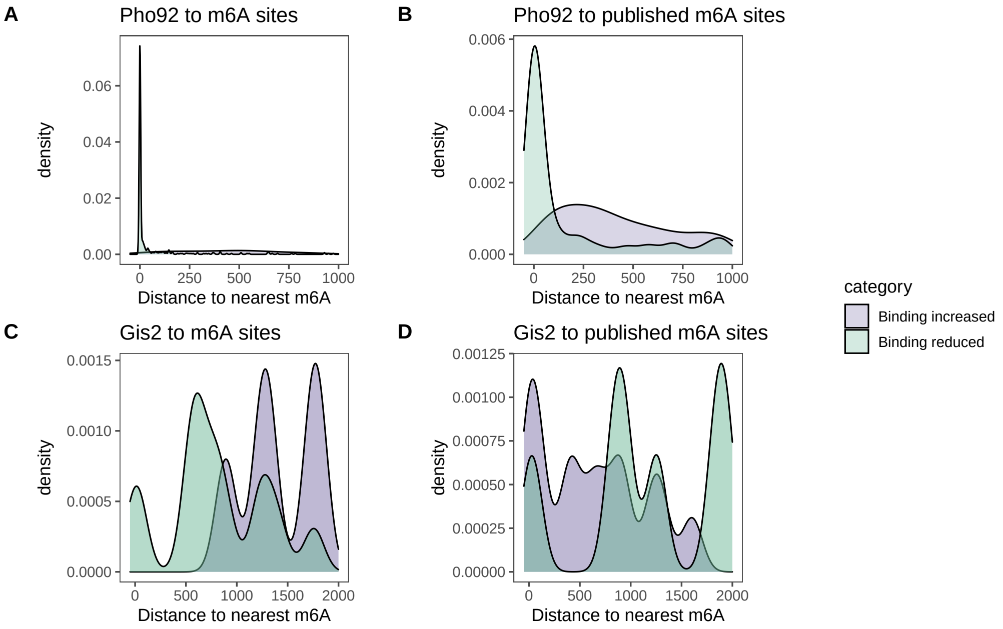


Figure 3.17: Distance to nearest m6A from Pho92p and Gis2p binding sites. A) Pho92p peaks are split into those where binding increases in *ime4*-Δ (purple) and those where binding is reduced (green). A density plot is shown of the distance from each binding site to the nearest m6A as defined by miCLIP. B) The same as in (A) but defining m6A sites by published data. C, D) The same as in (A, B), but for Gis2p binding sites.

3.3.5 Characterising Pho92p binding to RNA

To further characterise Pho92p RNA binding, I looked at its positional enrichment over transcripts. I found that Pho92p binds at the 3' end of RNAs, regardless of whether binding is Ime4p-dependent or not (Figure 3.18 A,B). I saw this effect in my categorised peaks, but I also verified it in the raw crosslinking data (Figure 3.18 D). Consistent with m6A pattern at mRNAs, Pho92p binding was detected predominantly at the 3' end of transcripts, with 23% (145/642) of Ime4-dependent binding sites directly overlapping a STOP codon (Figure 3.18 C). Whilst Ime4-independent Pho92 sites also occurred predominantly at 3' transcript ends, only 13% of these sites directly overlapped a STOP codon, suggesting it is possible m6A is involved in this positioning.

Next I used STREME to search for enriched motifs in Ime4p-dependent Pho92p peak sequences resized to 100nt intervals vs. shuffled dinucleotide-content-

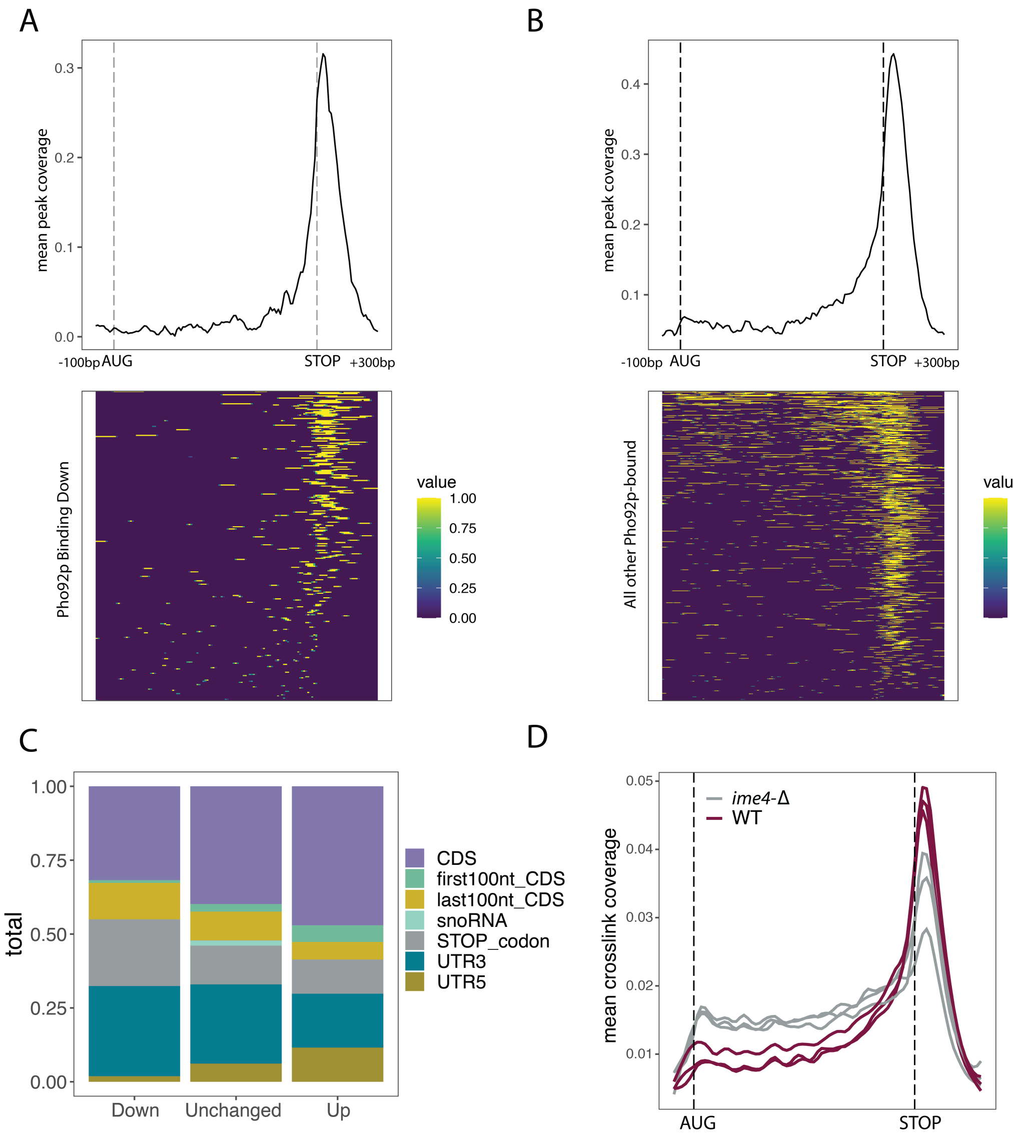


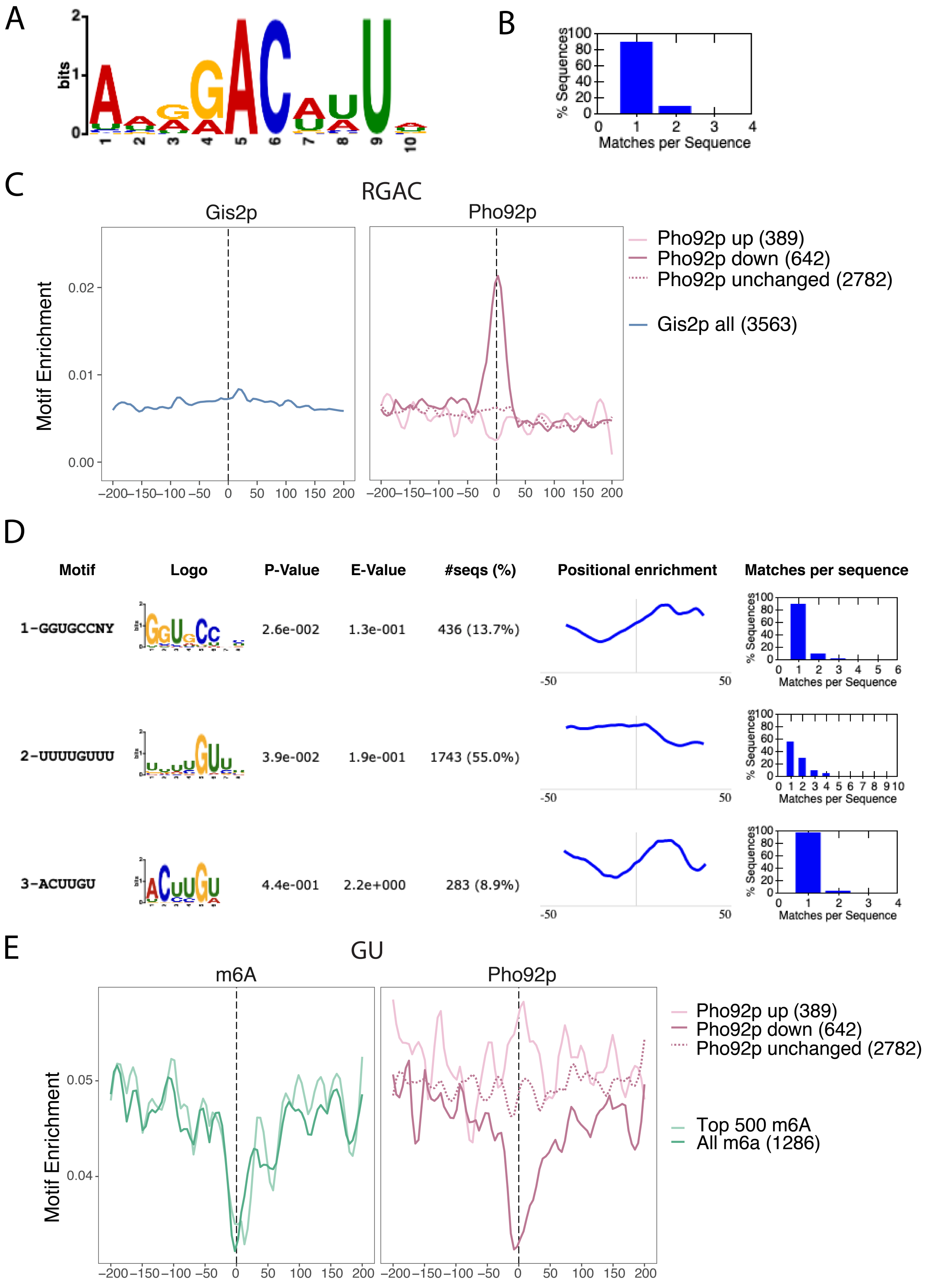
Figure 3.18: Pho92p binding site distribution over transcripts. A) Metagene profile of Ime4p-dependent Pho92p binding site distribution over yeast genes extended by 100bp upstream and 300bp downstream. A binding site is denoted as 1, values intermediate between 1 and 0 are due to the smoothing used to generate the heatmap image. B) As in A but for Ime4p-independent Pho92p binding sites (those that are unchanged or upregulated). C) Regional transcript distribution of Pho92p binding sites separated into those which are downregulated, unchanged or upregulated in *ime4*-Δ. D) Metagene profile of CPM-normalised iCLIP crosslink signal over yeast genes, showing WT and *ime4*-Δ *samples.*

matched background sequences. I found that the most enriched motif was almost indistinguishable from the extended m6A consensus motif found in our miCLIP data (Figure 3.19 A), with a similar distribution among peak sequences (Figure 3.19 B). Furthermore, by plotting the enriched RGAC motif around Pho92p peak centres of different categories it was apparent RGAC enrichment was a feature of binding sites that were downregulated in *ime4-*Δ and not in those that were upregulated or unchanged (Figure 3.19 C). I also verified RGAC was not enriched in Gis2p binding, again supporting the idea that it is not an m6A binding protein. This directional dependency on Ime4p is therefore present for Pho92p RGAC motif enrichment, but not for the 3' positional binding of Pho92p.

I was curious if there were any sequence features of Ime4p-independent Pho92p binding, so I searched using STREME with the Ime4p-independent peak sequences as foreground and the dependent sequences as background. I found several low complexity motifs, which all shared GU dinucleotides 3.19 D). To explore this further I plotted the enrichment of these dinucleotides around different categories of Pho92p binding site and also m6A sites. Interestingly there did appear to be some enrichment of GU dinucleotides in the category of upregulated Pho92p binding sites, whereas such sequences were depleted around m6A, or m6A-dependent binding 3.19 E).

3.3.6 Functional impact of Pho92p and m6A on RNA networks

Having identified and characterised sites of m6A and Pho92p binding, I next sought to address what functional impact this has for yeast cells undergoing meiosis. To understand the functional networks occupied by bound/modified genes, I first performed a gene ontology enrichment analysis using YeastEnrichr, specifically using KEGG 2019 pathways (https://maayanlab.cloud/YeastEnrichr/) (H. C. Chen et al., 2013; Kuleshov et al., 2016). (Figure 3.20). The top enriched term for both high confidence m6A sites and Ime4p-downregulated Pho92p binding was the MAPK signalling pathway. Interestingly for Pho92p, meiosis was the second most enriched term, but this did not appear in the top five terms for m6A sites, suggesting m6A



Motif	P-Value	E-Value	#seqs (%)
1-GGUGCCNY	2.6e-002	1.3e-001	436 (13.7%)
2-UUUUGUUU	3.9e-002	1.9e-001	1743 (55.0%)
3-ACUUGU	4.4e-001	2.2e+000	283 (8.9%)

Figure 3.19: Pho92p motif analysis. A) Sequence logo for most enriched motif in Pho92p Ime4p-dependent binding sites, y axis displays information content. B) Of Pho92p Ime4p-dependent binding sites containing the motif in A, a bar graph displaying those binding sites with 1 or more occurences of the motif. C) RGAC motif enrichment plotted around Gis2p binding sites (left, blue) and different categories of Pho92p binding site (right, pink). D) Top three enriched motifs in Pho92p Ime4p-independent binding sites, showing various attributes as calculated by STREME. E) GU dinucleotide enrichment plotted around m6A sites (left, green) and different categories of Pho92p binding site (right, pink).

function at meiotic genes is Pho92p-dependent. The MAPK pathway is highly important for yeast meiosis; in early meioisis, MAP kinases are important for nuclear reorganisation (Stone et al., 2000) and in late meiosis MAP kinases are critical for the formation of the spore walls (R. E. Chen & Thorner, 2007).

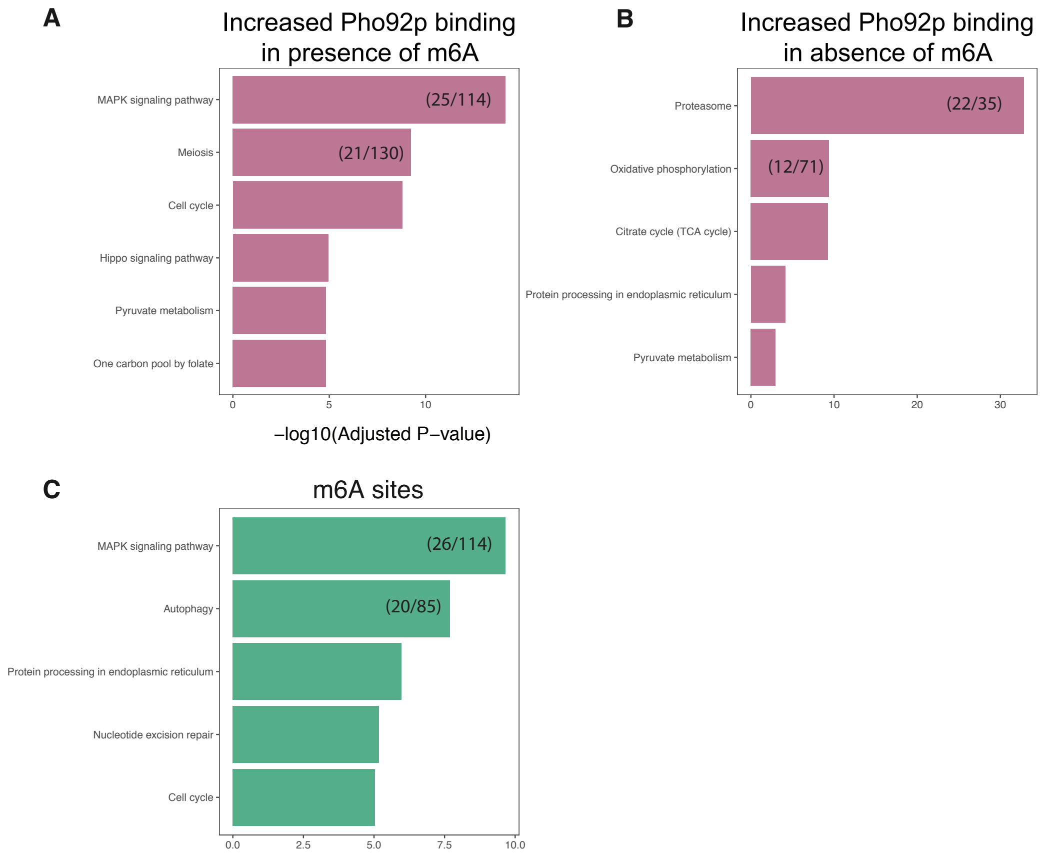


Figure 3.20: Pho92p and m6A gene GO term enrichments. A) GO term enrichments for transcripts containing Pho92p binding sites which decrease in *ime4*-Δ. The top six most enriched terms are shown, the x axis shows -log10(adjusted p value). The black numbers on the bars denote the number of genes identified out of the total number of genes in that GO term category. B) As in A but for transcripts containing Pho92p binding sites which increase in *ime4*-Δ. C) As in A but for transcripts containing high confidence m6A sites.

Interestingly, the majority of proteasomal genes (22/35) had Pho92p binding that was upregulated in *ime4*-Δ. Recycling proteins is critical during yeast meiosis, where there is no environmental nitrogen, so all amino acids for making new proteins must be scavenged from existing proteins.

Next, I investigated the RNA abundance of Pho92p bound genes when *PHO92* or *IME4* are deleted. I identified that Pho92p binding sites, specifically downregulated sites, are most enriched on transcripts that are upregulated in *ime4*-Δ and/or *pho92*-Δ, (Figure 3.21). Like the majority of Pho92p binding, such regulated binding is also at the 3' end of transcripts. This result suggests that Pho92p might be involved in the degradation of its bound transcripts. Given that it is Pho92p binding sites that are downregulated in *ime4*-Δ (Figure 3.21 A,B) that are the most enriched in transcripts that are upregulated upon loss of Ime4p and Pho92p, which suggests that this is very likely to be m6A-dependent binding. Also interesting is that a subset of genes which are downregulated in *ime4*-Δ have Pho92p binding sites which increase in signal in the same condition (Figure 3.21 D), however this doesn't appear to be the case for downregulated genes in *pho92*-Δ (Figure 3.21 C), which suggests this isn't caused by a direct effect of Pho92p binding. Additionally, Pho92p binds in an Ime4p-independent manner to genes that change in all directions in *ime4*-Δ RNA-Seq, suggesting again that *ime4*-Δ causes gene expression changes that have an independent mechanism to Pho92p-bound m6A targeting.

I was also interested to examine whether Pho92p binding of transcripts had any impact on their translation. To answer this question I integrated our data with published measurements of translational efficiency as measured by ribosome profiling during a meiotic time course (Brar et al., 2012). I found that m6A-modified transcripts were generally more highly translated than other genes expressed at the same time point (Figure 3.22 A). Furthermore, any form of Pho92p binding regardless of Ime4p-dependence seemed to increase translation efficiency even further (Figure 3.22 B).

Spurred on by these preliminary insights into Pho92p and m6A function, Dr. Radhika Varier performed an experiment to identify the fates of m6A-modified transcripts under WT and *pho92*-Δ conditions at 4hrs into sporulation. By inhibiting RNAPII with thiolutin, then measuring the relative amount of m6A in poly(A) purified RNA, she could monitor the decay of m6A-modified transcripts. Fascinatingly, she found that the decay of m6A-modified transcripts was dependent on Pho92p, as indicated by the RNA-Seq (Figure 3.22 C). Furthermore, by additionally treating

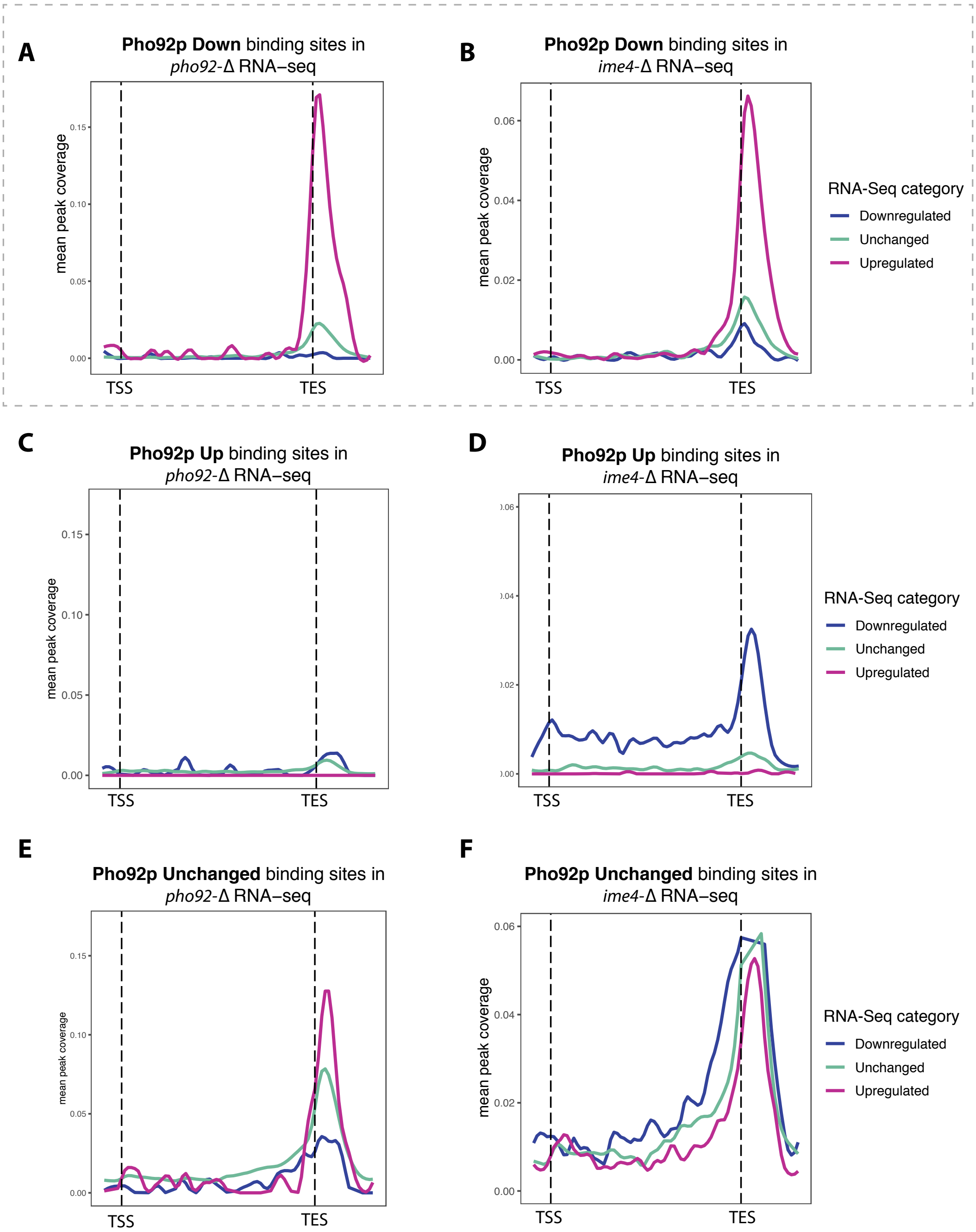


Figure 3.21: Behaviour of Pho92p-bound genes in *ime4*-Δ and *pho92*-Δ. Pho92p metagene profiles split into genes that are downregulated, unchanged or upregulated in *ime4*-Δ (B,D,F) and *pho92*-Δ (A,C,E) RNA-Seq. Metagenes are also stratified by Pho92p binding sites which decrease in *ime4*-Δ (A,B), those which increase (C,D) and those which don't change (E,F).

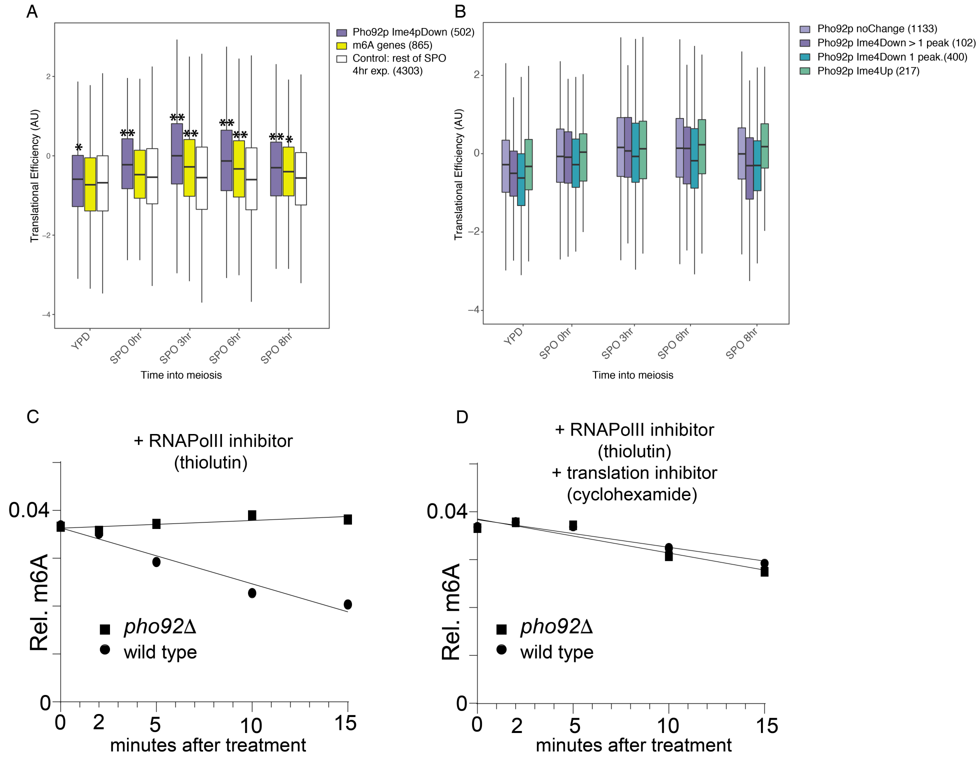


Figure 3.22: The role of m6A and Pho92p in meiotic translation. A) Boxplot showing translational efficiency as calculated in (Brar et al., 2012) at different time points into meiosis. YPD refers to nutrient-rich Yeast Extract–Peptone–Dextrose media, whereas SPO indicates sporulation media with potassium acetate and raffinose. Transcripts are grouped by Ime4p-dependent Pho92p binding and whether they are m6A modified. As a control, genes not bound by Pho92p or m6A modified, but deemed to be expressed at the 4hr time point, were taken. Note that genes could be present in multiple groups. *p< 0.005,**p< 0.001, groups compared to control using two-sided Welch's t-test, p values were Bonferroni corrected for multiple testing. B) As in A, but Pho92p binding is split into categories based on Ime4p-dependence. C) WT and *pho92*-Δ cells were treated with thiolutin at SPO 4hr to inhibit RNAPolII, and harvested at 5 time points. Poly(A)+ RNA was extracted and m6A mass spectrometry was performed to measure the relative m6A/A ratio, shown on the y axis. D) As in C, but cyclohexamide was added with thiolutin at time 0 to additionally inhibit translation. The experiments in C and D were conceptualised and performed by Dr. Radhika Varier.

cells with cyclohexamide, and therefore blocking translational elongation, the effect of Pho92p in enhancing RNA decay was lost (Figure 3.22 D). This suggests that the decay of Pho92p-bound transcripts is dependent on active translation.

3.3.7 Characterisation of Gis2p RNA binding.

Having analysed Pho92p and m6A on transcripts, I returned to the Gis2p data, to see if any conclusions could be made about the protein's function in meiotic cells. Generally, I found Gis2p binding to occur at any position along transcripts (Figure 3.23 A), with no specific preference, as noted by previous studies, for the beginning of the CDS (Figure 3.23 B). Around half of transcripts bound by Gis2p had more than one binding site (Figure 3.23 C), and these binding sites had a repeating GWW motif context as previously reported, but with perhaps the most salient feature being repeated spaced out G's (Figure 3.23 D, E). Such GWW motifs were clearly enriched in Gis2p binding sites and absent, or even depleted, at m6A and Pho92p binding sites (Figure 3.23 F).

I again, used YeastEnrichr, to find enriched KEGG 2019 pathways for Gis2p-bound genes (Figure 3.24). All proteasomal genes are bound by Gis2p, as well as approximately half of meiotic annotated genes and those involved in oxidative phosphorylation. I would venture that probably most expressed genes are bound by Gis2p - given that we identify over 2000 bound transcripts and at an estimate, $\sim$4000 are expressed at any given time (assuming some limits to our detection at more lowly expressed transcripts). Despite this, the identification of every proteasome annotated gene is noteworthy. I was interested especially because in my Pho92p analysis I identified that 22/35 annotated proteasome genes had increased Pho92p binding in *ime4*-Δ.

This observation led me to examine the overlap between Pho92p and Gis2p binding. I found roughly a quarter of Pho92p and Gis2p binding sites overlapped (Figure 3.25 A). Consistent with my previous observation the most enriched GO term for Gis2p and Pho92p bound genes was the proteasome, with 32/35 genes bound by both proteins (Figure 3.25 B). I wanted to establish what the co-binding

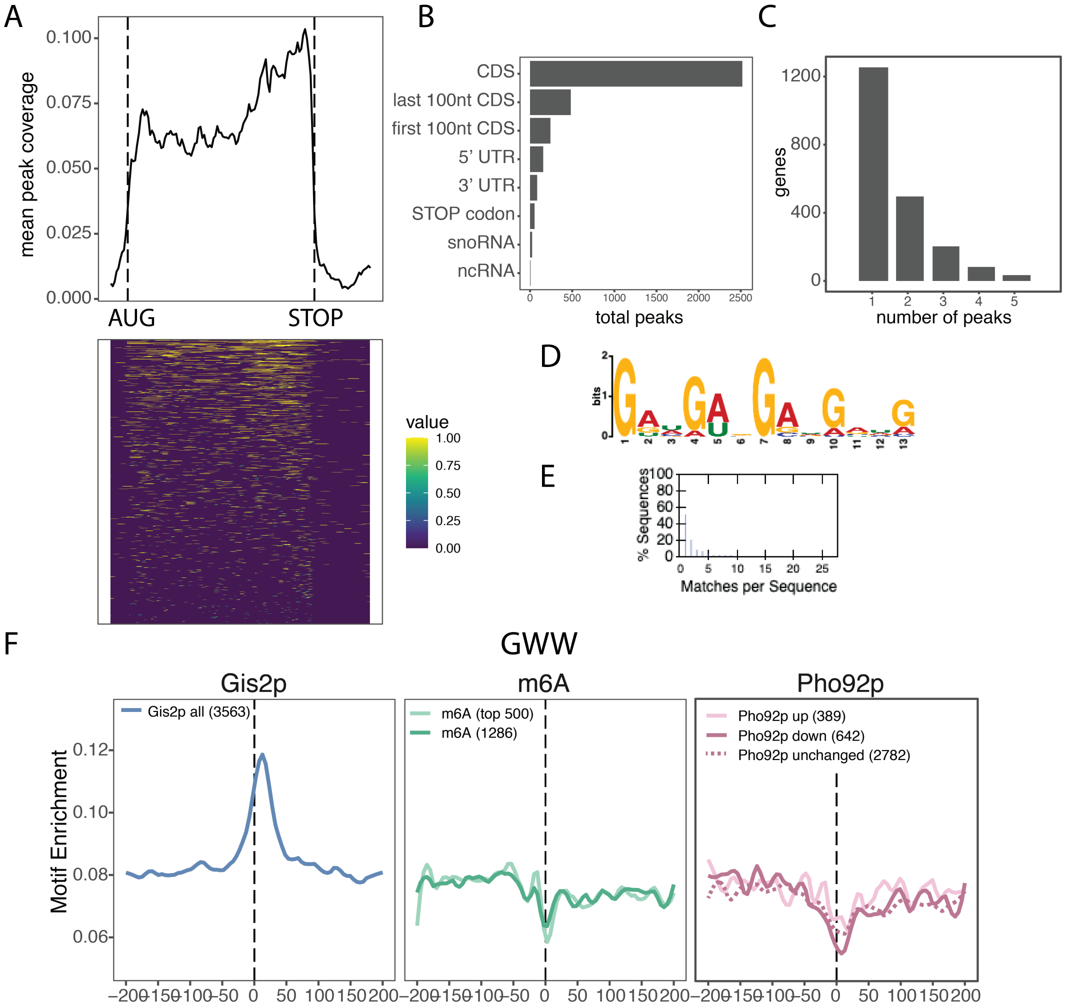


Figure 3.23: Gis2p binding characteristics. A) Gis2p metagene profile shown as both a heatmap and line graph. Each nucleotide is designated 1 or 0 depending on whether a Gis2p binding site is present. Intermediate values are a result of vector smoothing that occurs to generate the image. B) Gis2p binding site annotations to transcript regions. C) The number of Gis2p peaks per gene that contains at least one binding site. D) Sequence logo for the most enriched motif by STREME analysis. E) The number of occurrences of the motif in (D) per 100nt rescaled Gis2 binding site sequences. F) Occurrence of GWW (W=A/U) motif across Gis2p binding sites (blue), m6A sites (green) and Pho92p sites (purple).

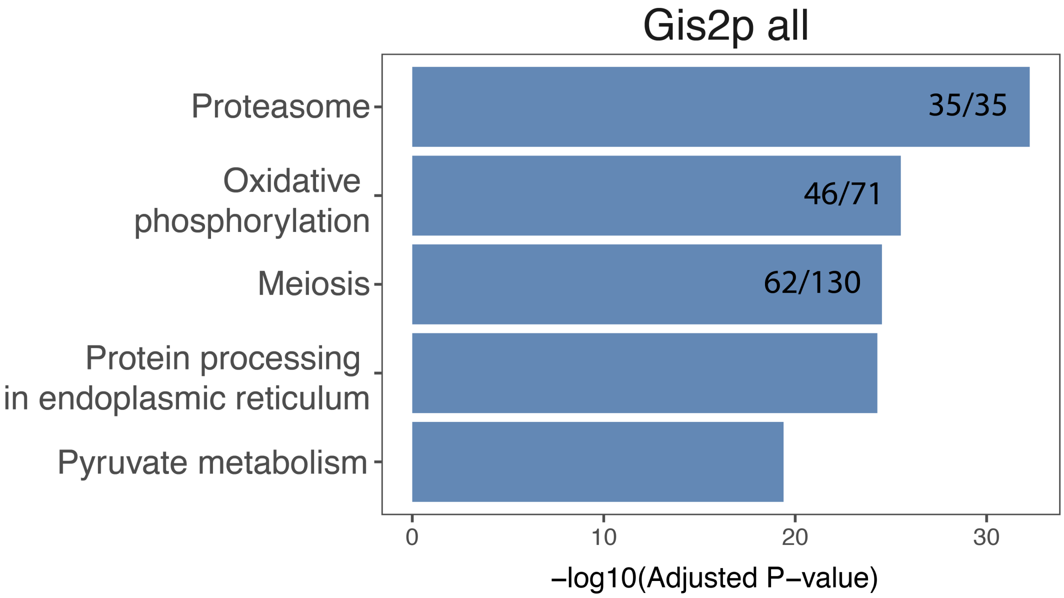


Figure 3.24: Gis2p gene GO term enrichments taken from KEGG 2019 pathways. The number on the bar indicates the number of genes bound by Gis2p out of the total number of genes in the given category.

of these RBPs at proteosomal transcripts might mean for their fate. I compared the abundance of proteasomal transcripts vs. those involved in oxidative phosphorylation, the second most enriched GO term, in a variety of deletion conditions involving *PHO92*, *GIS2* and *IME4* (Figure 3.25 C, D). The RNA-Seq experiments were performed in two batches with a degree of technical variance between experiment 1 and experiment 2 as can be seen by comparing *ime4*-Δ from experiment 1 and 2, and also the triple deletion RNA-Seq experiment which was performed twice (Figure 3.25 C, D). Despite this variance, we can make some conclusions. Deletion of *GIS2* alone appears to have little effect on proteasomal transcripts, whereas deletion of *PHO92* either alone, or in combination with *GIS2* causes a small amount of downregulation. The greatest impact is seen when deleting *IME4*, causing the transcripts to be more substantially downregulated. Because we know that in *ime4*-Δ these transcripts are more bound by Pho92p, this might suggest that the increased binding of Pho92p leads to some degradation. However, this could be some indirect impact of *ime4*-Δ. The small downward trend in such transcripts following single *PHO92* deletion would suggest a more complex relationship. Moreover, these results suggest that Gis2p is not impacting the RNA abundance of these transcripts.

The relevance of these changes in proteasomal transcripts does seem significant when compared to the transcripts encoding components of the oxidative phosphorylation pathway, which comparatively don't change very much in any of the deletion conditions (Figure 3.25 D).

I also returned to the translational efficiency dataset to ask if there was any combinatorial impact of Gis2p and Pho92p binding on translation (Brar et al., 2012). I found that in general, Gis2p binding alone on transcripts was correlated with a slightly higher translational efficiency, which makes sense because it has been previously found to interact with translation factors (Figure 3.25 E). Pho92p binding alone correlated in a bigger boost to translational efficiency than Gis2p alone. Transcripts that were bound by both did seem to have a higher translational efficiency than transcripts with either protein bound alone, suggesting there may be some combinatorial impact on translation.

To investigate whether there was any positional relationship between Gis2p and Pho92p binding I studied the proteasomal bound genes in the genome browser (Figure 3.26). Gis2p and Pho92p binding sites could be at either end of the transcript as in *RPN2* mRNA (Figure 3.26 A). Binding sites could also be overlapping as in *PRE2* mRNA (Figure 3.26 B). There didn't appear to be a clear relationship. One point to note is that because these genes become downregulated in *ime4*-Δ, it would be hard for us to designate them as m6A modified, using the *ime4*-Δ control. There is however, unfiltered m6A miCLIP signal in these genes, suggesting that they could be m6A-modified (data not shown).

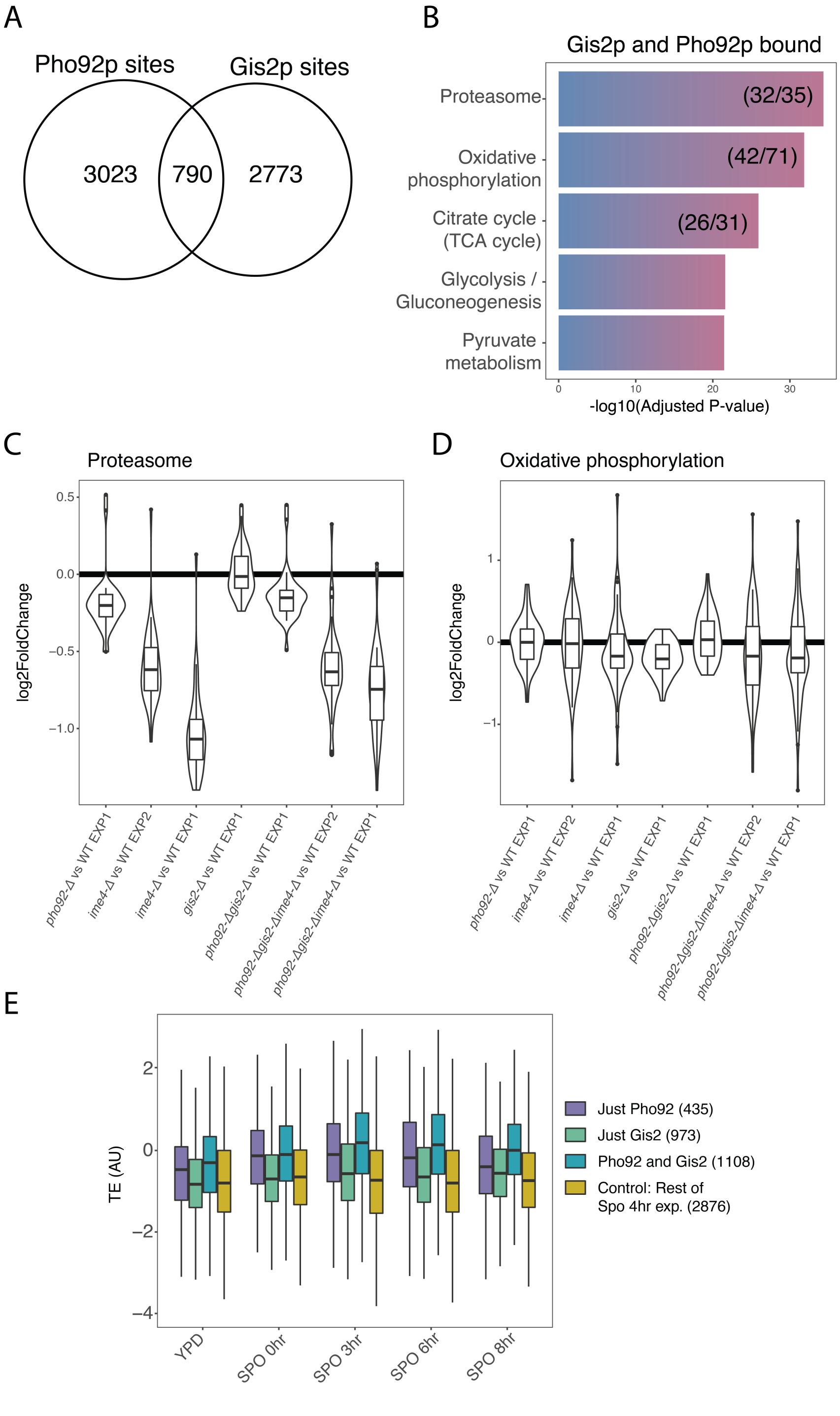


Figure 3.25: Fate of transcripts bound by both Gis2p and Pho92p. A) Venn diagram showing number of Gis2p and Pho92p sites within 100nt of each other. B) GO term enrichments of genes bound by both Gis2p and Pho92p. C) Violin plot with overlayed boxplot showing log2 fold change of proteasomal genes in variety of deletion conditions, shown on the x axis. D) Same as C but instead for genes involved in oxidative phosphorylation. E) Boxplot showing translational efficiency as calculated in (Brar et al., 2012) at different time points into meiosis. Transcripts are grouped by whether they are bound by Pho92p alone, Gis2p alone or by both RBPs. As a control, genes not bound by Pho92p or Gis2p, but deemed to be expressed at the 4hr time point, were taken.

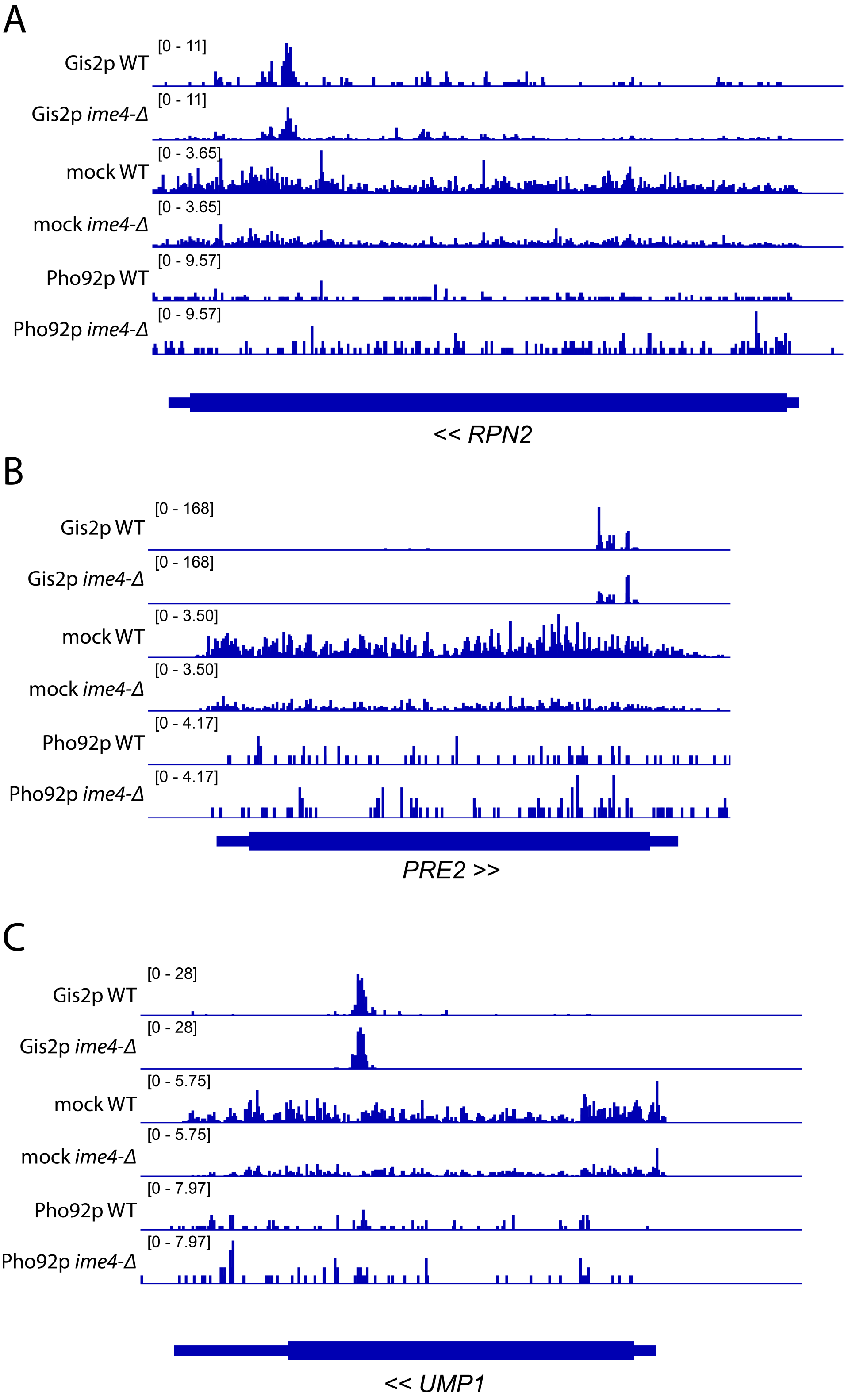


Figure 3.26: Pho92p and Gis2p iCLIP at proteasomal genes. A) Representative Pho92p and Gis2p iCLIP samples at the *RPN2* gene in WT and *ime4*-Δ conditions. A Gis2p binding site can be seen at the 3' end of the gene and a Pho92p binding site is seen at the 5' end in *ime4*-Δ condition only. Mock miCLIP shows the gene expression. Signal is shown as crosslinks per million. B) Genome browser track as in A but at the *PRE2* gene. At this gene both Gis2p and Pho92p binding occurs more at the 3' end of the transcript. C) As in A but for the *UMP1* gene, Gis2p and Pho92p binding occurs at the 3' end of the gene. Again, Pho92p binding is increased in *ime4*-Δ

3.4 Conclusions

In this chapter, I aimed to understand the relationship between m6A, Pho92p and Gis2p in yeast meiosis by analysing and integrating a combination of miCLIP, iCLIP and functional genomics data. Whilst some conclusions are clear, a lot of functional mechanistic details remain obscure and will require further investigation.

3.4.1 Pho92p couples mRNA decay to active translation

My analysis confirms previous work demonstrating Pho92p is a bonifide m6A reader: roughly 30% of binding sites could designated as downregulated in *ime4*-Δ; these binding sites were mostly overlapping or near to m6A sites. When searching for motifs *de novo* the extended m6A consensus motif was found and Pho92p is positioned at the 3' end of transcripts as m6A is. Given that the next most enriched protein in the original mass spectrometry assay was Gis2p, and there is no other YTH-domain containing protein in the yeast genome, this suggests Pho92p is likely the only m6A reader in yeast. However, it is possible that other proteins could interact with m6A flanked by a different sequence context, or indirectly. By integrating iCLIP-defined binding sites with RNA-Seq data I found that transcripts with Pho92p binding sites which were downregulated in *ime4*-Δ, were upregulated in both *ime4*-Δ and *pho92*-Δ, suggesting Pho92p is targeting bound transcripts for degradation in an m6A-dependent manner.

Interestingly I found that Pho92p binding sites show striking 3' end enrichment regardless of whether binding is Ime4p-independent or at m6A sites, which would suggest that the positioning is not entirely dependent on m6A. Dr. Radhika Varier found by Co-IP and ChIP-qPCR that Pho92p accumulates at chromatin over the course of gene transcription, in a manner that is dependent on polymerase-associated factor 1 complex (Paf1C) (data not shown). This would suggest transcription or chromatin factors contribute towards Pho92p positioning on transcripts, which may explain why Ime4p-independent Pho92p binding does not need a strong RNA sequence context. The question then, is why Pho92p binding that is Ime4p-independent does not have the same impact on RNA abundance as Ime4p-dependent binding, as I found in my integration of RNA-Seq data. A simple mechanism could be that m6A locks Pho92p binding to RNA, which in turn allows the Pho92p-mRNAs to make it intact to translating ribosomes, where they will be degraded. It is likely that without m6A locking Pho92p in place, the protein-RNA interaction is less stable, and Pho92p dissociates from the mRNA before the complex makes it to the ribosome. Longer occupancy on transcripts could also give Pho92p more of an opportunity to be post-translationally modified which could facilitate its interactions with mRNA decay factors. This could be investigated by performing comparative iCLIP of Pho92p from nuclear and cytoplasmic fractions - if this hypothesis is true one would expect that Ime4p-independent binding sites will have more signal in the nuclear fraction vs. the cytoplasmic fraction, whereas Ime4p-dependent binding will remain much the same between nucleus and cytoplasm.

An alternative hypothesis could be that transcripts with Pho92p-bound m6A also have additional RBPs bound that help to facilitate the role in decay. This might be supported by the fact that any Pho92p binding seemed to indicate a higher translation efficiency of the transcript, however this result is very correlatory and requires validation. To explore the preliminary bioinformatic results from the translational efficiency datasets, it would be useful to perform ribosome profiling with deletion or depletion of Pho92p. Higher "translation efficiency" as measured by ribosome profiling could also indicate ribosome stalling, so it might be worth performing monosome vs. disome profiling to distinguish these possibilities (C. C.-C. Wu et al., 2020). Infact, one mechanism of mRNA decay coupling to translation that has previously been described involves ribosome collosions on mRNA, leading to ubiquitination of the 40S ribosome subunit by E3 ubiquitin ligase Hel2p (ZNF598 in mammals) . This chain of events instigates recruitment of the CCR4-NOT deadenylase complex.

3.4.2 Regulation of proteasomal transcripts

Given that the proteasome may be involved in the coupling of mRNA translation and decay, I was surprised to find that 22/35 proteasomal transcripts had Pho92p binding sites which increased in *ime4*-Δ, these transcripts were also bound by Gis2p. Whilst I couldn't find any Ime4p-dependent m6A sites on these transcripts, this is likely because they are downregulated in *ime4*-Δ. I did find WT m6A miCLIP peaks across these transcripts suggesting they could be m6A-modified, but this would require more validation. One possible reason why there might be m6A at proteasomal transcripts, but that Pho92p binding goes up in *ime4*-Δ could be that the MIS complex remains bound at these transcripts, blocking Pho92p binding when Ime4p is present. Another RBP could also display this same behaviour. It would be interesting to tag some of these transcripts, such as *RPN2*, *PRE2* and *UMP1* with MS2 loops, UV crosslink and pull them down, followed by protein mass spectrometry to determine which other RBPs are bound at these transcripts in WT vs. *ime4*-Δ conditions. Assembly of some elements of the proteasome is co-translational, for example Rpt1p and Rpt2p are thought to interact as nascent polypeptides in an ”assemblysome” also containing Not1p, the scaffold of the Ccr4–Not complex (Panasenko et al., 2019). Pho92p has been shown to interact with the Ccr4-Not complex. Taken together, the role of m6A, Pho92p and Gis2p in assembly of the proteasome warrants further investigation. As a starting point, it would be interesting to check proteasome assembly in *ime4*-Δ, *pho92*-Δ and Ime4p catalytic mutant.

3.4.3 Gis2p is not an m6A reader

Prior to the present study Gis2p had never been investigated as an m6A reader. I conclude that its discovery in the mass spectrometry data was likely an artefact for the following reasons: a) very few Gis2p binding sites are downregulated in *ime4*-Δ, and unlike Pho92p or the miCLIP data, just as many sites are upregulated as downregulated, b) of the sites that are downregulated few are close to m6A and c) Gis2p binding sites show no enrichment for RGAC sequence context.

So, the question remains, why was Gis2p enriched in the original mass spectrometry screen? I would posit that Gis2p is a highly abundant RBP in yeast cells, which has a rather degenerate binding motif. In my motif analysis, one could conclude that the main feature is repeating spaced out G nucleotides which were present on the RNA oligo probes. So, why enriched binding when m6A was present? I speculate that the non-modified probe could form some secondary structure between G and C nucleotides, which melted with the addition of m6A, making the oligo more accessible. It is also possible Gis2p was able to piggyback on Pho92p binding to m6A oligos.

Previous studies have shown that whilst Gis2p is an abundant RBP, the only observed impact in *gis2*-Δ is larger cells. I found that Gis2p binds the majority of mRNAs, sometimes with multiple binding sites, suggesting it is a fundamental component of mRNPs. Therefore, the lack of more extreme phenotype is surprising. I found that transcripts bound by Gis2p seemed to have slightly higher translation efficiency, and those occupied by both Pho92p and Gis2p have even higher translational efficiency (TE) than transcripts bound by either protein alone, suggesting Gis2p is involved somehow in modulating translation. This perceived increase in TE could again be indicative of ribosome collisions. GWW codons encode negatively charged amino acids such as aspartate and glutamate, which have been shown to impact ejection times from the ribosome exit tunnel (Sabi & Tuller, 2015), perhaps Gis2p is involved in somehow mitigating this impact, or engaging with nascent peptides that get stuck.

3.4.4 Limitations to the present analysis

Whilst we uncover useful information, there are limitations to the present analysis. For all of the CLIP data there is a bias towards detection of binding/m6A on more highly expressed genes, especially where differential analysis is performed that requires some degree of peak coverage across several replicates. I also identified that peaks on genes that increased in expression in *ime4*-Δ were more likely to be called as downregulated. It is unclear how much of this effect is biological, because m6A causes decay of transcripts, or somewhat technical in that peaks are more confidently called as downregulated when the gene expression changes in the opposing direction. To assess the impact of technical bias it would be useful to simulate data corresponding to different scenarios and run the same analysis.

In the case of m6A detection, antibody approaches remain imperfect due to off-target antibody binding and crosslinking biases. Whilst we have mitigated this somewhat by including the *ime4*-Δ control, it would be better to apply a more specific method in the first place. As described in the Introduction, methods that involve chemically modifying the m6A residue, such as m6A-SAC-seq, are showing promise and in the future may offer a more reliable way to map m6A in the transcriptome (Hu et al., 2022).

Another limit to the detection of differential binding or m6A sites, is that *ime4*-Δ causes widespread changes in gene expression. This increased variability reduces our power to call differential peaks. It would be good to repeat some of the experiments with the Ime4p catalytic mutant strain, which should have less dramatic gene expression changes, which should improve the detection of differential peaks. Furthermore, by doing RNA-Seq in this background we could determine which effects are due to m6A, and which are due to other impacts of Ime4p on gene expression/-translation.

3.4.5 Further investigations of the N6-methyladenosine dependent regulatory network in yeast meiosis

To further investigate function, it would be ideal to produce Pho92p and Gis2p degron-tagged yeast strains, such that effects could be monitored after rapid depletion, as a lot of the present results rely on deletion strains which could have compensated for intial defects in unpredictable ways. A common method to achieve this is to use an auxin-inducible degron system (Nishimura et al., 2020). In the example of Gis2p, where the function is unclear in deletion strains, perhaps surviving cells have adapted to life without the RBP, but monitoring the phenotype after rapid depletion might be more fruitful.

The results from the RNA-Seq data and thiolutin experiment suggest that it would be interesting to study transcriptome-wide RNA stability changes after Pho92p depletion, either in *pho92-*Δ or a *pho92-AID* strain. This would be possible using a technique that involves metabolic labelling, for example SLAM-Seq, although this might be difficult to achieve across the meiotic time course where gene expression changes rapidly (Herzog et al., 2017). In the thiolutin experiment it was only possible to monitor m6A-modified RNA; therefore, it would be good to integrate m6A miCLIP data with the SLAM-Seq to deduce if the effect is really specific to m6A-modified transcripts.

Chapter 4

CLIP Data Visualisation

As the cathedral is to its foundation
so is an effective presentation of
facts to the data.

*Willard C. Brinton, Graphic
Methods For Presenting Facts.
(1914)*

4.1 Introduction

With complex bioinformatics datasets it can be easy to hide behind the command line and focus purely on tables of numbers. However visualisation is a critical part of both the analysis process and also of conveying the findings of the analysis to a broader audience.

In this chapter I describe two projects that highlight the importance of thoughtful data visualisation. The first highlights different normalisation strategies for CLIP data and ways to effectively condense many experiments into few graphs, as well as how design choices can impact the interpretation of the reader. The second focuses on visualisation in the analysis process for a step where visualisation is commonly missing - describing a new peak caller with an interactive user interface for testing parameters.

4.2 psiCLIP and clipplotr

4.2.1 Background

Purified spliceosome iCLIP (psiCLIP) is a variant of the standard iCLIP protocol, optimised in order to study the RNA binding sites of spliceosomal proteins in defined stages of splicing (Strittmatter et al., 2020). psiCLIP is particularly useful for capturing the binding of spliceosomal helicases, which are notoriously difficult to map in electron cryo-microscopy (cryo-EM) structures, presumably due to the dynamic nature of their binding. Eight spliceosomal helicases safeguard the fidelity of pre-mRNA splicing, and understanding their activity is critical to the understanding of the splicing process as a whole (Cordin & Beggs, 2013). psiCLIP is performed on *in vitro*-assembled spliceosomes which are stalled in the same manner as those prepared for cryo-EM structures, offering the possibility to integrate the resulting information.

This project produced an unprecedented volume of CLIP data - we published 93 individual samples, but in the course of developing and refining the methodology probably analysed around double this number. The *in vitro* spliceosomes are assembled on stereotyped splicing substrates, variants of the *ACT1* and *UBC4* pre-mRNA transcripts, therefore we required a visualisation approach that would enable us to make quick comparisons of multiple CLIP samples at once across these substrates, and also a normalisation strategy that would make these comparisons as meaningful as possible.

After developing such a visualisation in the course of the psiCLIP project, I then worked alongside Dr. Anob Chakrabarti to produce a generalised tool that could enable others, including those with limited bioinformatics expertise, to quickly produce such visualisations for their own projects.

4.2.2 Smoothing CLIP data

iCLIP data across transcripts is most often presented as a column graph, with positions along the transcript on the *x* axis and cDNA counts at each position plotted on the *y* axis (Figure 4.1A). While nucleotide-resolution column graphs convey precise positional information, they cannot be used to present data across multiple experimental conditions on a single plot since overlapping bars would not be visible. Furthermore, it can be difficult to interpret the meaning of individual highly crosslinked nucleotides - by presenting individual nucleotide crosslinking in such plots we may be drawing attention to specific nucleotides in a way that an untrained eye might construe to be especially biologically meaningful. For example, in this particular plot, the eye is drawn to the individual crosslink around position +30 (Figure 4.1A). There are many reasons why a certain nucleotide in particular may be highly crosslinked, for example this nucleotide might represent a position where a "crosslinkable" amino acid comes into contact with a preferentially "crosslinkable" RNA nucleotide - regardless of whether this nucleotide is used by the protein in its recognition of the RNA. In an RNA with strong secondary or tertiary structure, crosslinking may also be constrained. As I found in Chapter 3 for miCLIP data, crosslinking can sometimes occur in a U-rich tract adjacent to the suspected binding/modification motif. For these reasons, a broader area of generally high crosslinking is more reliably interpreted as corresponding to RBP binding, versus focus on single highly crosslinked positions.

Therefore, I explored smoothing the data of individual psiCLIP helicase experiments using Gaussian smoothing (Figure 4.1B). Curves with a window size of ten nucleotides reflected the crosslinking profile well, and also agreed with the nine nucleotides of RNA bound to the two RecA domains in structures of DEAH-box helicases (He et al., 2017). The window of ten nucleotides will not always be appropriate for every RBP - some may have more punctate, or others more spread out, binding profiles.

4.2.3 Different normalisation for different purposes

Further to smoothing of the data, when presenting multiple replicate samples together one must consider how to appropriately normalise these datasets to make comparisons between them meaningful. In the case of psiCLIP, different normali-

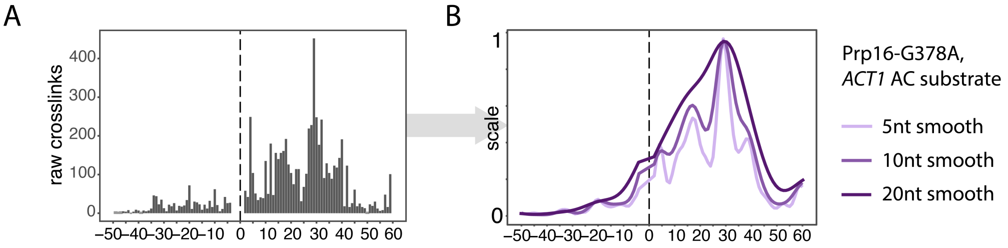


Figure 4.1: Smoothing psiCLIP data. A) An example of Prp16 data mapped to a spliceosomal substrate shown as a crosslink column graph. Position along the spliceosomal substrate is represented as relative to the intron branch point A nucleotide, here 0. B) The same data after Gaussian smoothing using different window sizes.

sation strategies were used depending on the research question. Most common in iCLIP analyses is to take a "crosslinks per million" approach where crosslinks at a given position are divided by the total number of crosslinks in the library and then multiplied by 1 million (See Figure 6 (Hallegger et al., 2021) for an example). This is analogous to the reads per million (RPM) approach used in RNA-Seq, but for CLIP approaches normalising for gene length is not appropriate as we are interpreting single nucleotide positions (Mortazavi et al., 2008). For proteins that do not crosslink very efficiently, variations in library complexity between replicate samples can render the crosslinks per million approach still insufficient. In the case where we are simply interested in changes in position of binding on a given RNA sequence, rather than binding frequency or strength, we can normalise crosslinks by dividing by the maximum number of cDNAs mapped to a single position, an approach often referred to as "max peak".

In the psiCLIP project we were interested in the binding position of helicase Prp16 on the stereotyped splicing substrates *ACT1* and *UBC4*. Prior to our work, Prp16 was thought to bind at an invariant distance from the intronic branch point adenosine, but we noticed that with longer distances between branch point and 3' splice site the peak of Prp16 binding appeared to be further away from the branch point. Simply normalising replicates by library size made it difficult to make comparisons (Figure 4.2A). Instead, to specifically answer the question of whether there was a difference in the position of binding peak between *UBC4* and *ACT1* transcripts, I subtracted the untagged protein control signal from the UV-crosslinked signal, smoothed the data further using a 20nt window and used max peak normalisation. It was now much clearer that there was indeed a difference in binding position, which I also confirmed to be statistically significant ($p < 0.001$, $\alpha = 0.05$, Student's unpaired two sided *t* test) (Figure 4.2B).

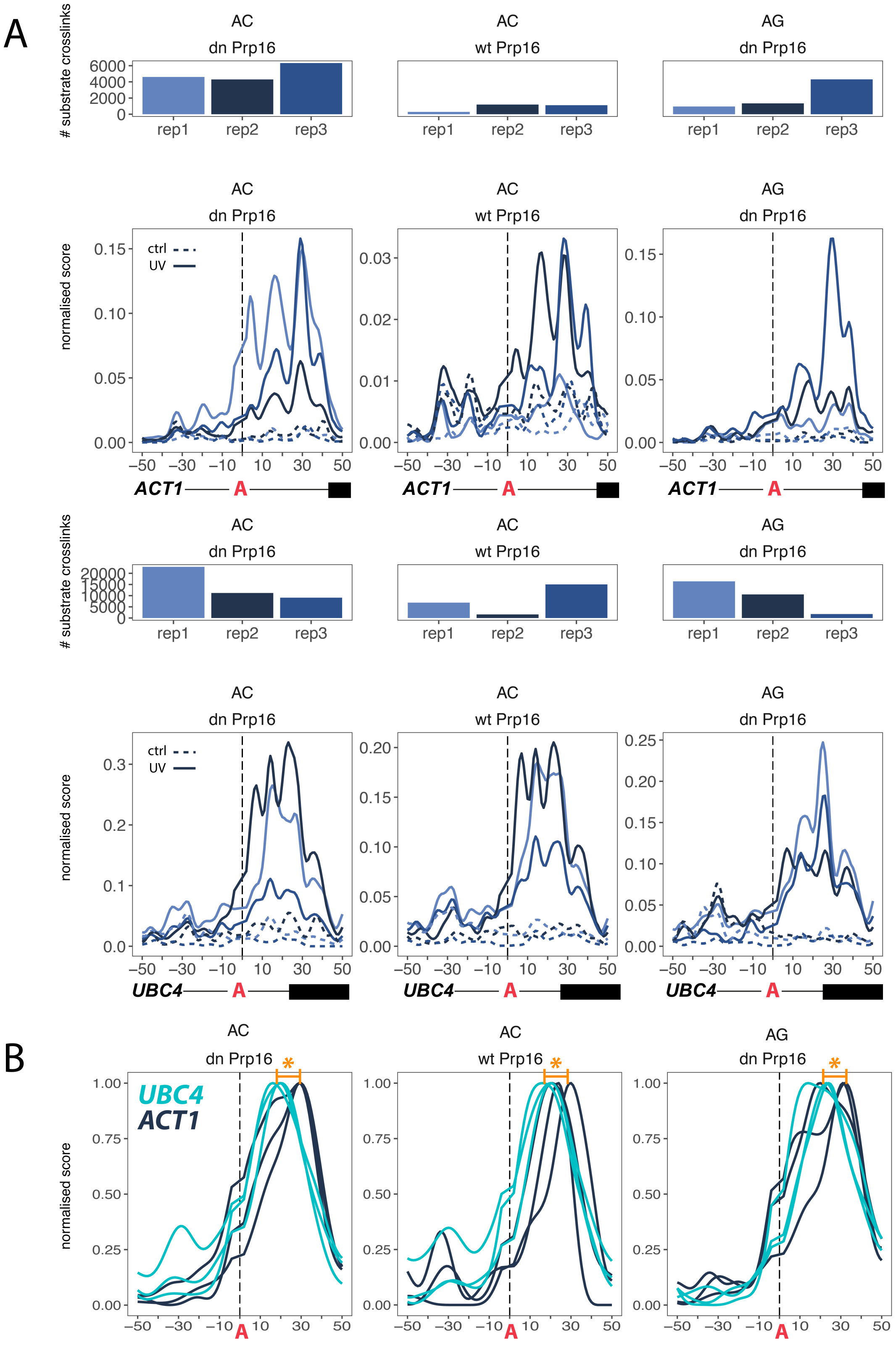


Figure 4.2: Normalising psiCLIP data. A) Summary of 36 Prp16 psiCLIP samples across different substrates (*UBC4* and *ACT1*, with both AG 3' splice site and AC 3' splice site) and also WT vs. mutant (dn for dominant negative) Prp16 protein. Smoothed signal is plotted relative to the branch point adenosine. Above each smoothed line graph the number of crosslinks per UV-crosslinked sample is shown. Note that generally larger libraries have more enriched binding. B) Here the 36 samples are further condensed, the untagged control signal is subtracted from the UV signal, a 20nt Gaussian smooth is used and *ACT1* and *UBC4* samples are plotted in the same graphs, but in different colours.

4.2.4 Creating a user-friendly tool for plotting many CLIP replicates

Following my experiences with psiCLIP and the experiences of Dr. Anob Chakrabarti with an iCLIP project involving visualisations of multiple iCLIP experiments of different TDP43 mutants, we worked together to produce a general tool that would enable researchers to visualise CLIP data across transcripts in a useful way (Chakrabarti et al., 2021; Hallegger et al., 2021). The tool is written in R using the R packages optparse, data.table, ggplot2, ggthemes, cowplot, patchwork, zoo and smoother, and the Bioconductor packages rtracklayer and GenomicFeatures (Dowle et al., 2019; Gentleman et al., 2004; Hamilton, 2015; Huber et al., 2015; Lawrence et al., 2009; Wickham, 2011; Wilke et al., 2019). clipplotr is available at www.github.com/ulelab/clipplotr where there is extensive documentation. The code is primarily split into two halves: one half focuses on retrieving and visualising the relevant gene annotations for the user-specified region, whilst the other half focuses on loading, normalising and smoothing the input datasets as required. Four types of "tracks" can be plotted:

- the *crosslink track* represents CLIP datasets provided as bed or bedgraph files, these can be grouped into different categories e.g. for replicate datasets of different proteins. The samples can be normalised by library size (crosslink per million), max peak or by size factors provided by the user. Smoothing can be performed with either a rolling mean or Gaussian smooth, where the smoothing window or span is given by the user.

- the *auxiliary track* is a way of providing any additional annotations, for example repeat elements, SNPs or called peak regions.

- the *coverage track* enables plotting of orthogonal coverage-based data such as ribosome profiling, or RNA-Seq, which again can be grouped in any way.

- the *annotation track* is provided by the user as a GTF file and represents transcript annotations for the plotted region. The user can choose to plot the annotation as 'transcript' or 'meta-transcript'. With the transcript option, all transcripts in the region are plotted and coloured by gene. With the meta-transcript option all transcripts for a given gene are condensed such that all annotated exons are combined. This option is especially useful for when large numbers of redundant transcripts clutter the annotation track.

The plot in Figure 4.3 is a reproduction of Figure 1C from (Zarnack et al., 2013), produced with the following single line of code:

```
./clipplotr \
--xlinks 'hnRNPC_iCLIP_rep1_LUjh03_all_xlink_events.bedgraph.gz,
    hnRNPC_iCLIP_rep2_LUjh25_all_xlink_events.bedgraph.gz,
    U2AF65_iCLIP_ctrl_rep1_all_xlink_events.bedgraph.gz,
    U2AF65_iCLIP_ctrl_rep2_all_xlink_events.bedgraph.gz,
    U2AF65_iCLIP_KD1_rep2_all_xlink_events.bedgraph.gz,
    U2AF65_iCLIP_KD2_rep1_all_xlink_events.bedgraph.gz' \
--labels 'hnRNPC␣1,hnRNPC␣2,U2AF2␣WT␣1,U2AF2␣WT␣2,U2AF2␣KD␣1,U2AF2␣KD␣2' \
--colours '#586BA4,#324376,#0AA398,#067E79,#A54D69,#771434' \
--groups 'hnRNPC,hnRNPC,U2AF2␣WT,U2AF2␣WT,U2AF2␣KD,U2AF2␣KD' \
--normalisation libsize \
--smoothing rollmean \
--smoothing_window 50 \
--auxiliary 'Alu_rev.bed.gz' \
--auxiliary_labels 'reverse␣Alu' \
--coverage 'ERR127306_plus.bigwig,ERR127307_plus.bigwig,ERR127308_plus.bigwig,
    ERR127309_plus.bigwig,ERR127302_plus.bigwig,ERR127303_plus.bigwig,
    ERR127304_plus.bigwig,ERR127305_plus.bigwig' \
--coverage_labels 'CTRL1␣1,CTRL1␣2,CTRL2␣1,CTRL2␣2,KD1␣1,KD1␣2,KD2␣1,KD2␣2' \
--coverage_colours '#A1D99B,#74C476,#31A354,#006D2C,#FDAE6B,#E6550D,#FC9272,#
    DE2D26' \
--coverage_groups 'CTRL,CTRL,CTRL,CTRL,KD,KD,KD,KD' \
--gtf gencode.v34lift37.annotation.gtf.gz \
--region 'chr1:207513000:207515000:+' \
--highlight '207513650:207513800' \
--annotation transcript \
--output 'CD55.pdf'
```

Whilst the tool has been tested widely within multiple labs and used in several publications (Hallegger et al., 2021; J.-H. Lee et al., 2021), the command to run clipplotr still needs to be executed on the terminal, and the relevant software dependencies must be installed. This requires a degree of technical expertise to operate. In the near future we hope to integrate the tool with the online web server iMaps (imaps.goodwright.com), so that clipplotr can be run from a graphical user interface

(GUI) with no requirement for any programming from the end user.

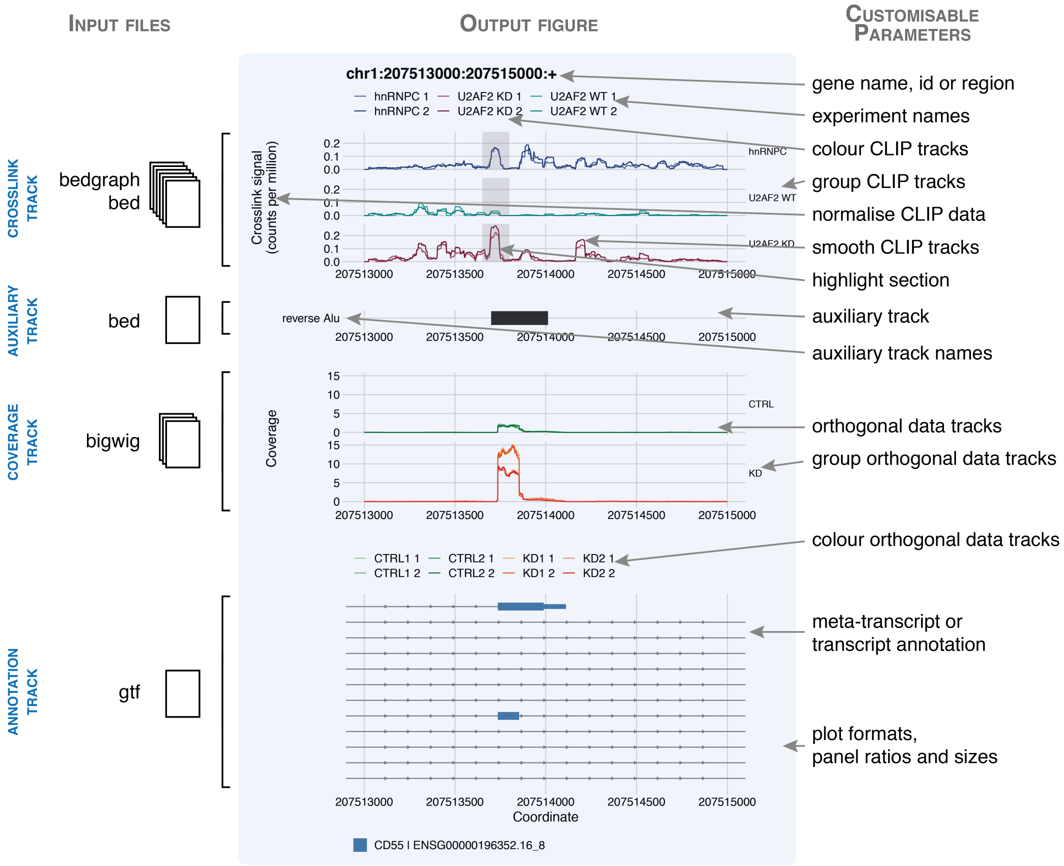


Figure 4.3: Annotated view of clipplotr output. A figure generated by a clipplotr command using data from (Zarnack et al., 2013) is inset in blue, demonstrating all four types of data track. The input file formats required for each track type is indicated to the left. On the right the customisable parameters are annotated that can be specified in the single clipplotr command. Figure is reproduced from (Chakrabarti et al., 2021).

4.3 A new interactive peak caller: Clippy

4.3.1 Background

Peak calling is a crucial step in analysis of genomic and transcriptomic datasets. For iCLIP data, peak calling of crosslinks serves multiple purposes: allowing researchers to home in on interesting binding sites to further manipulate and validate at the bench, whilst also enabling downstream bioinformatics analyses such as differential binding and integration with other data types. Theoretically the perfect peak caller would be able to distinguish true crosslink positions generated by target RBP-RNA contacts from background consisting of non-target protein crosslinks and events that are not actually due to crosslinking, eg. single nucleotide polymorphisms (SNPs) in PAR-CLIP data, or truncations in iCLIP that are due to RNA modifications, non-crosslinked reads, contamination of libraries, sequencing errors etc. In practice, other considerations are important: ease of installation and use (including parameter adjustment and interpretability), speed of the program and requirements for auxiliary information, such as transcript annotations and input controls.

Different peak callers may in fact be useful in different situations based on these practical and technical requirements. Peak-callers may perform differently when faced with an RBP that binds sharply to a clear motif (eg. RBFOX2 (Van Nostrand et al., 2016)) vs. a more promiscuous binder without a clear motif (eg. FMRP, FUS (Rogelj et al., 2012)). Moreover, it is important to define our purpose in peak calling. One might explain that we call peaks because clusters of single nucleotide crosslinks represent the footprint of an RNA-binding protein on RNA, which we consider to be the binding site, and so we hope that by peak-calling we capture this binding site. However, in practice the reality is more complex, crosslinking sites may not always correspond to binding sites and peak-calling is most often a way to summarise our data to submit for further downstream analysis, which might influence the parameters we choose. For example, researchers wanting to mutate a binding site for later functional studies will be willing to sacrifice true positive sites for the sake of minimising false positives, as a few very strong candidates may suffice. However, bioinformaticians looking at overlapping two datasets might prefer less stringent peak calling to maximise the possibility of overlap. Similarly, when undertaking a differential analysis it can be better to assign broader peaks to reduce the variability introduced by shorter peaks with lower counts.

In this section I describe Clippy - a Python-based peak-caller for CLIP data. Clippy is unique in that it provides an interactive peak calling app that launches in the user's browser. This app allows the user to explore how their parameter choice affects peak calling on several genes of interest. For example, a user might specifically choose a low, medium and high expression gene to ensure that their chosen parameters perform well over these ranges or a user might have prior biological knowledge that enables them to check parameters on certain key target transcripts. This solves several issues with current peak calling: 1) to optimise parameters one must run code multiple times blindly, load results into a genome browser to compare, and perhaps go back to run the code again; in the worst case a user might not explore a peak caller's settings at all, and use sub-optimal default settings because of this time cost. By utilising Clippy's interactive app this iterative process is made quick and easy. 2) The algorithms behind peak callers can feel like a "black box" to biological researchers, who are often left wondering why one region is called as a peak, whilst another is not. Allowing the user to interactively explore thresholding provides transparency and clarity to the process.

The full benefits of Clippy are as follows:

- Interactive peak calling mode enables exploration of parameters on test genes.
- Fast to run - 30 minutes to 1 hour for a large human dataset.
- Easy to install through the Bioconda package repository, with a linked Docker container that auto-updates with each new release (Grüning et al., 2018).
- Minimal transcript annotation requirements. A GTF must be supplied but the only required feature is "gene", this makes the peak-caller very permissive to non-standard model organisms or custom annotations.

- Can be run on transcriptomic or genomic-mapped crosslinks.
- Can run on any type of CLIP data, as it runs on a bed file of crosslinks defined by the user.
- Parameters are simple and straightforward to intuit.
- Users can choose to calculate separate minimum thresholds for exons vs. introns to account for coverage differences.

Many of the latest developments in CLIP peak calling have been to incorporate the data from input controls, such as in programs like PureCLIP (Krakau et al., 2017). However, it has recently been shown that the broad application of input controls is not as straightforward as one might assume, and domination of certain RBPs in these controls could actually further bias peak calling rather than improving it (Kuret et al., 2021). Therefore, we focused on finding a tool that is able to perform well without input data, and recommend that such enrichment analyses should be performed downstream to the initial peak calling itself.

To test Clippy we ran it alongside a selection of the most commonly-used CLIP peak callers: Piranha, Clipper, iCount and PureCLIP (König et al., 2010; Krakau et al., 2017; Lovci et al., 2013; Uren et al., 2012) (Table 4.1). We also tested Paraclu: a CAGE-seq peak caller that has previously been repurposed for CLIP data (Hallegger et al., 2021; Kuret et al., 2021). We performed a comparison of CLIP peak callers across different CLIP types and proteins, using: Rbfox2 eCLIP data and Ptbp1 and Tia1 iCLIP data (Table 4.2).

4.3.2 Clippy algorithm and interface

Clippy is a wrapper around the scipy `find_peaks` function tailored for CLIP data, that I have developed together with Dr. Marc Jones (Virtanen et al., 2020). This algorithm is used across data science to find peaks in a wide range of datasets, from identifying peaks of volcanic plume density in the field of geophysics (Haley et al., 2021) to finding movement artefacts in bioradar signal produced by wearables

Peak caller	Concept	Installation	Annotation requirement	Reference
Piranha	Uses a zero-truncated negative binomial distribution to assign statistical significance to bins of crosslink values.	Available from Bioconda	None	(Uren et al., 2012)
Clipper	Fits smoothed splines to read depth to define clusters of crosslink sites.	Requires compilation.	A number of annotations are built into the software.	(Yeo et al., 2009)
iCount	Permutation of crosslink locations to determine significant crosslink sites, which subsequent merging.	Available from Bioconda	Gencode or Ensembl formatted gtf	(König et al., 2010)
PureCLIP	Hidden Markov Model to segment the genome into regions enriched and non-enriched for crosslink sites.	Available from Bioconda	FASTA file for the reference genome.	(Krakau et al., 2017)
Paraclu	Progressive peak splitting to determine “stable” peaks.	Available from Bioconda	None	(Frith et al., 2008)
Clippy	Wraps the find_peaks function from the scipy.signal Python library to identify peaks.	Available from Bioconda	Any gtf/gff with “gene” lines in the third column. For exon thresholding, “exon” lines are also required.	This chapter

Table 4.1: Characterisation of CLIP peak callers.

during human sleep (Anishchenko et al., 2019). The algorithm is based around two key parameters used to identify peaks:

1. *Prominence:* The prominence of a peak is a topographic term calculated as the distance between the height of the peak and its nearest contour line. In mountaineering terms it can be thought of as the shortest height drop that can be taken from a given peak to reach a higher one.

2. *Relative height:* The relative height parameter can be used to define the width of a given peak. In this sense, a line is drawn at a relative fraction of the prominence of a peak and the width of that line is determined to be the peak width. In practical terms lower relative width values will result in shorter peaks.

In order to apply the algorithm to CLIP data I introduced a few extra components. Firstly, I decided to call peaks on a gene-by-gene basis so that certain thresholds could be calculated on a gene-by-gene basis. My idea was that this could help to ensure better peak calling across a range of gene expression values. Secondly, I introduced smoothing of the crosslink data before peak calling, the rationale behind this being that users can make broader peaks with larger smoothing windows and also to combat some of the biases observed when focusing on single crosslinks (as discussed for psiCLIP). I also introduced a minimum threshold that must be exceeded by the crosslink signal to be called as a peak, which is set as the mean crosslink signal across the gene. The threshold for prominence is set as the mean crosslink signal across the gene + (standard deviation in crosslink signal * a user-defined scaling factor). By introducing this scaling factor users can increase the prominence threshold if they have especially noisy or high coverage datasets (Figure 4.4).

An interactive parameter search app was built using the Dash framework (https://dash.plotly.com/). Launching the app is triggered by execution of a single Clippy command in the terminal; the app can then be accessed in any browser - here I show Google Chrome (Figure 4.5). A downsampled version of the Pho92p

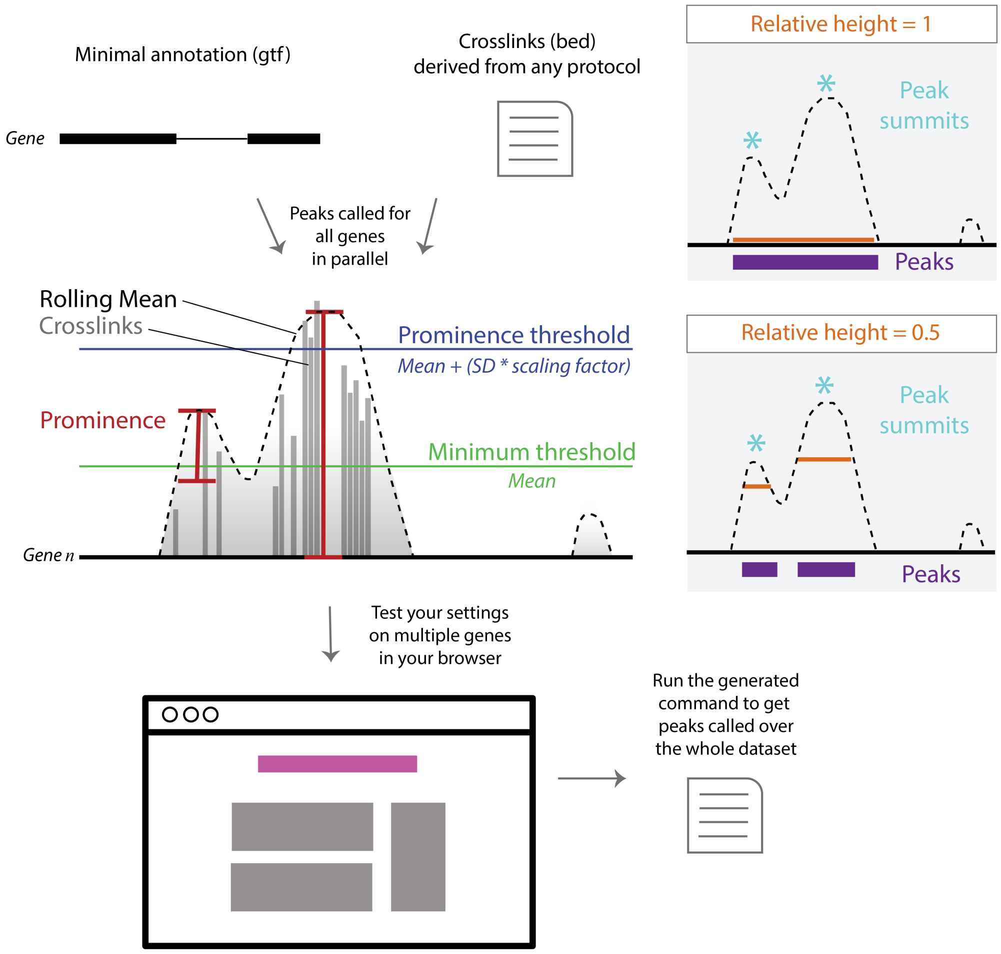
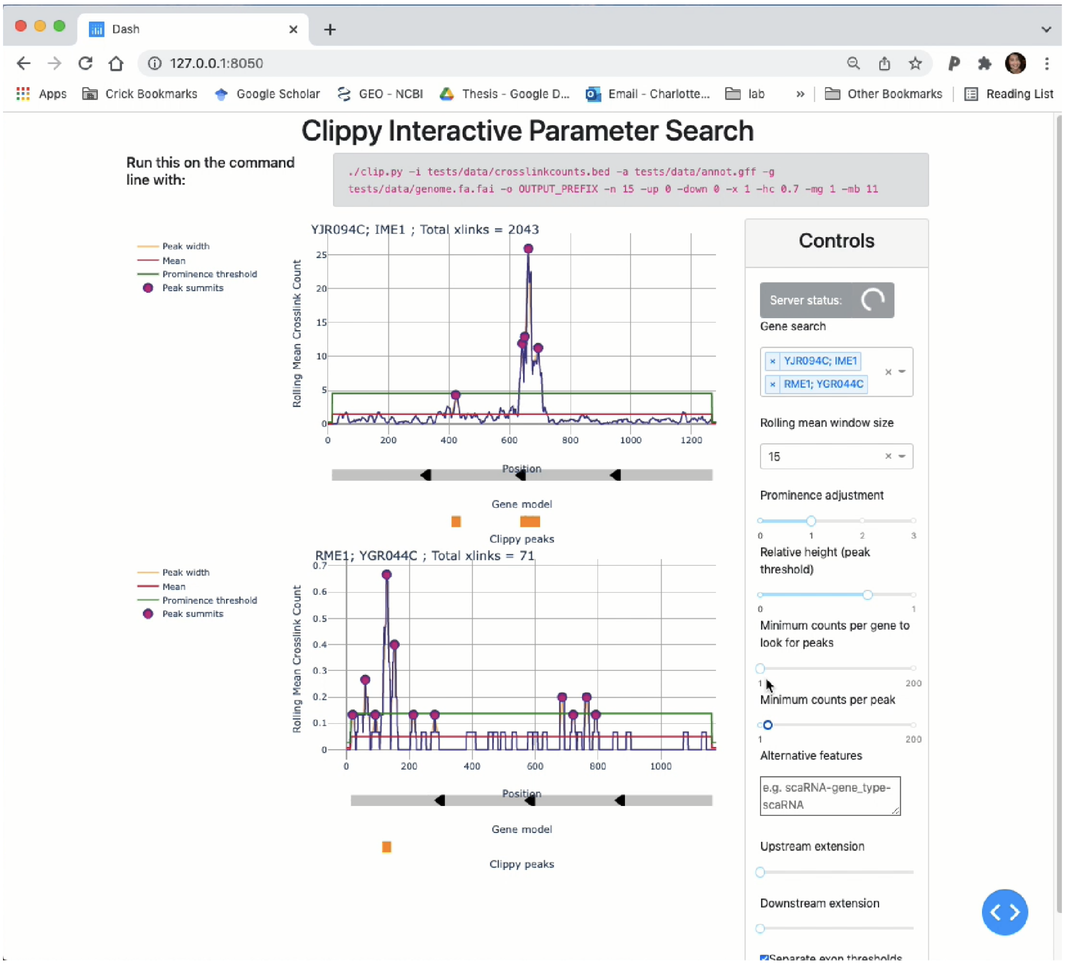


Figure 4.4: Schematic of Clippy algorithm and operation.

iCLIP data is packaged with Clippy as an example dataset. The app has three main sections: at the top in red text the command to run Clippy with the currently trialed settings is dynamically displayed, on the right a control panel allows users to adjust every parameter and on the left CLIP signal is plotted across genes chosen to be visualised. In this specific example I chose *IME1* and *RME1*. The total number of crosslinks across the gene is printed in the plot titles - it is clear that *IME1* has a lot more coverage (2043 crosslinks) than *RME1* (71 crosslinks), but in this example I have found parameters that work well for both genes. The final peaks are displayed in orange beneath the crosslink signal graphs.



Clippy Interactive Parameter Search

Run this on the command line with:

```
./clip.py -i tests/data/crosslinkcounts.bed -a tests/data/annot.gff -g
tests/data/genome.fa.fai -o OUTPUT_PREFIX -n 15 -up 0 -down 0 -x 1 -hc 0.7 -mg 1 -mb 11
```

Controls

Server status:

Gene search: YJR094C; IME1, RME1; YGR044C

Rolling mean window size: 15

Prominence adjustment: 1

Relative height (peak threshold)

Minimum counts per gene to look for peaks

Minimum counts per peak

Alternative features: e.g. scaRNA-gene_type-scaRNA

Upstream extension

Downstream extension

☑ Separate exon thresholds

Figure 4.5: Screenshot of Clippy interactive mode.

4.3.3 Performance

In order to test Clippy's performance against state of the art CLIP peak callers, Dr. Marc Jones pre-processed public Rbfox2 eCLIP data and Ptbp1 and Tia1 iCLIP data using nf-core/clipseq (https://nf-co.re/clipseq), merged replicate samples and ran the peaks callers using a custom Nextflow workflow (https://github.com/luslab/peak-benchmarking, (Table 4.2)). Clippy was run using parameters: window size 15, adjustment factor 1, min gene count 5 and min peak count 5, the current default, and all other peak callers were run with default settings. RNA-Seq corresponding to TIA knockout and PTBP1/2 knockdown (Table 4.2) was mapped using nf-core RNA-Seq pipeline (DOI: 10.5281/zenodo.5550247, (Ewels et al., 2020)), and splicing analysis was performed with rMATS (Shen et al., 2014).

Dataset	Data Type	Cell line	Source	Reference
Tia1 and Tial1	iCLIP	HeLa	ArrayExpress: E-MTAB-432	(Z. Wang et al., 2010)
Ptbp1	iCLIP	HEK293	ArrayExpress: E-MTAB-5027	(Haberman et al., 2017)
Ptbp1-Ptbp2 knockdown	RNA-Seq siRNA PTBP1 and PTBP2	HEK293	GEO: GSE69656	(Gueroussov et al., 2015)
DKO of TIA1 and TIAL1, rescued with DOX overexpression of TIA1-FH or TIAL1-FH	RNA-Seq	Flp-In T-REx HEK293	NCBI BioProject: PRJNA400256	(C. Meyer et al., 2018)
Rbfox2 eCLIP	eCLIP	HepG2	ENCODE: ENCSR456FVU	(Van Nostrand, Freese, et al., 2020)

Table 4.2: Datasets used for testing peak callers.

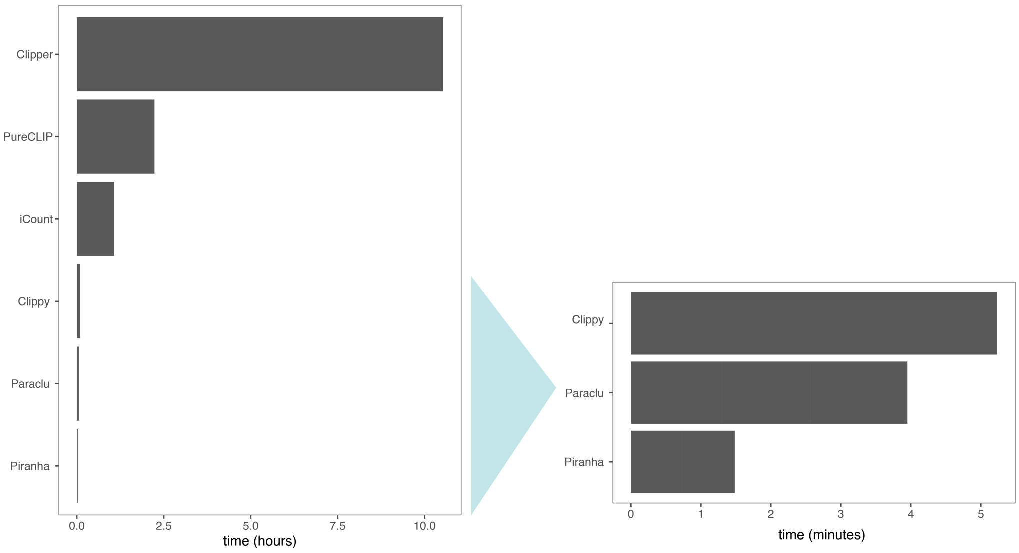


Figure 4.6: Runtime of peak callers for Ptbp1 dataset.

Clippy ran in ~5 minutes for the largest dataset (Ptbp1), much faster than iCount, PureCLIP or Clipper which took between an hour and ten hours to complete (Figure 4.6). Only Paraclu and Piranha were faster, taking 4 and 1.5 minutes respectively.

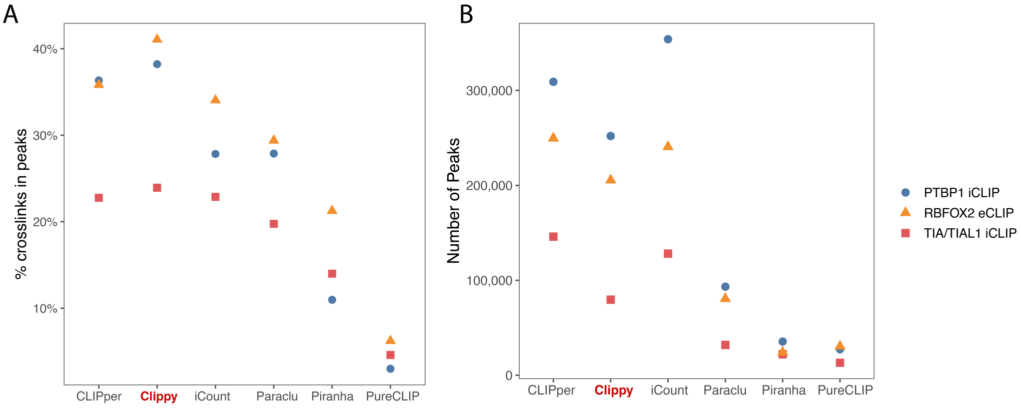


Figure 4.7: Number of peaks and percentage of crosslinks contained within peaks. A) Percentage of all crosslinks contained within peaks for Ptbp1 iCLIP, Rbfox2 eCLIP and Tia1/Tial1 iCLIP data analysed with six different peak callers. B) The number of peaks called for each dataset by each peak caller.

Clippy consistently produced peaks containing the most crosslinks, despite not calling the most peaks (Figure 4.7). PureCLIP consistently called few peaks and discarded the most crosslinks. Called peaks could be true positives or they might be false positives. In the absence of ground truth, there are a few ways to assess the precision of Clippy. In the context of CLIP data, it is easier to identify surrogate true positive outcomes, than false positives. For example, Ptbp1, Rbfox2 and Tia1/Tial1 all regulate splicing events in the transcriptome, therefore we can make an assumption that a binding site near to a regulated exon is a true positive. However, it is harder to make the statement that a binding site not near to a regulated exon is a false positive. This is because in a technical sense, our RNA-Seq might not be sensitive enough to detect every regulated exon and in a biological sense the binding site might be near to a different transcriptomic landmark that is involved in a separate function of the RBP. It's also possible that RBPs might bind lots of regions in the transcriptome without being functional at all.

A common way to examine peaks around regulated exons is to plot an RNA map (Rot et al., 2017). In a splicing RNA map exons are split into categories: silenced (increased inclusion in RBP knockdown RNA-Seq), enhanced (decreased inclusion in RBP knockdown), constitutive (percent spliced in values near to 1) and control (percent spliced in values are consistent between WT and RBP knockdown, but are not so high to suggest the exon is constitutively spliced in). CLIP peaks are then plotted as a metaprofile around exons, split into the different categories. This strategy has enabled the discovery of RBP position-dependent principles of RNA regulation. It has also been previously used to test how well peak callers are able to identify binding sites near regulated exons (Chakrabarti et al., 2018). To produce the RNA maps in this section I used code developed by Aram Amalietti available from https://github.com/ulelab/rna_maps.

Another way to address the performance is to look at sequence motif enrichment from peak sequences. PEKA is a motif finding algorithm for CLIP data, that generates enrichment scores for all kmers of a given length found within peak sequences (Kuret et al., 2021). These scores are validated against scores from *in vitro* methods such as RNA Bind-n-Seq (RNBS) and RNAcompete (RNAC) (Lambert et al., 2014; Ray et al., 2017). I reasoned that examining these scores would give a sense of whether Clippy, and the other peak callers, are truly capturing the important characteristics of RBP binding sites. Further, usually the authors of peak callers run a tool like DREME and report the most enriched motif, which is less useful as you do not have a sense of what other sequences might be enriched (Bailey, 2011).

Applying RNA maps to Ptbp1 and Tia1/Tial1 peaks, I found that Clippy produced the most enriched peaks across regulated exons, followed by Paraclu (Figure 4.8A,C). I removed Clipper from the RNA maps because it had dramatically higher signal than any other peak caller, obscuring evaluation of the other peak callers, but this signal was also in "off-target" regions. Combined with the long run time and poor motif performance it seems sub-optimal compared to the other options. In terms of recovering motifs, Clippy was able to capture both Ptbp1 and Tia1/Tial1 motifs (Figure 4.8B,D). Running Clippy with a longer rolling mean window enabled more variant motifs to be detected, which deserves further investigation. It is of note that the very stringent peak callers seem to perform better at motif detection when binding is more punctate and restricted to one motif (ie. Tia1/Tial1) than when binding is more spread, and seemingly multivalent (Ptbp1). It is concerning that even though Clipper called nearly the most peaks, it was unable to detect Ptbp1 motifs.

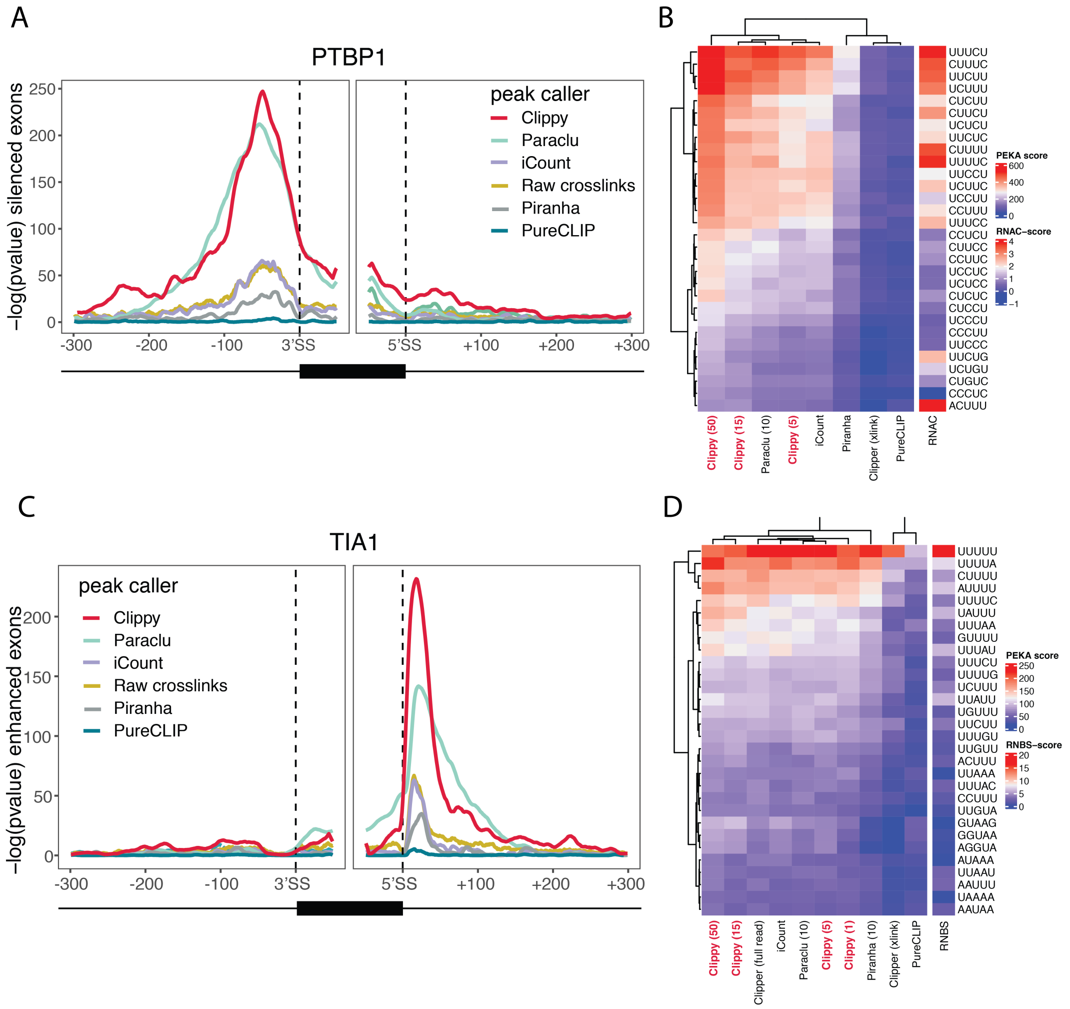


Figure 4.8: RNA maps and PEKA enrichment scores for Ptbp1 and Tia1/Tiall1. A) RNA map showing Ptbp1 iCLIP peaks over exons that are silenced by Ptbp1, as determined by Ptbp1/2 siRNA knockdown. Different colours of line indicate peaks called by different peak callers. The *y* axis is -log(pvalue) determined by chi-squared test on the frequency of iCLIP peaks in silenced exons vs. control exons at a given position. The *x* axis shows distance in nucleotides from the 3' splice site or 5' splice site of the cassette exon. B) Heatmap showing PEKA score of 5-mers enriched in peaks called by different peak callers, note Clippy is shown with varying window size. The right-most column displays *in vitro* binding scores for the 5-mers as determined by RNAcompete. C) As in A, but for Tia1/Tiall1 iCLIP data. D) As in B, but for Tia1/Tiall1 iCLIP data with *in vitro* RNA Bind-n-Seq data.

4.4 Conclusions

In this chapter I have described the development of two software tools for the visualisation and analysis of CLIP data. The first, clipplotr, enables users to normalise, smooth and visualise their data over a gene or genomic region of interest, alongside complementary annotations or data. The strength of this tool lies in its ability to sensibly condense large amounts of data into a digestable plot from which one can make biological observations. The second tool, Clippy, is a peak caller with a unique interactive graphical interface, allowing users to explore the impact of different parameter choices on called peaks. I have demonstrated that Clippy outperforms, or matches performance, of existing CLIP peak callers in various scenarios, demonstrating it to be a useful addition to CLIP analysis pipelines.

Two important analysis issues are addressed in this chapter, that of data normalisation and parameter choice. Some degree of normalisation of individual samples is required for their comparison. The question of how to normalise is an interesting one - currently the possibilities within clipplotr are by: library size, where the crosslinks at each position are divided by the total crosslink count and multiplied by one million; the maximum peak, where the crosslinks within a window are divided by the cDNA count of the highest crosslink; or by size factors, these are integers derived by other means, for example via DeSeq2, and crosslinks are divided by the given number. Whilst these options are useful, there are several situations in which they may not be sufficient. For example, in the case where a user would like to compare two CLIP samples where the RBP of interest is dramatically differentially expressed, or has it's RNA-binding capacity inhibited in one condition, simple library size normalisation won't capture the true nature of the binding dynamics. Due to a large number of simply missing binding sites in this example, the remaining binding sites will have their normalised counts inflated, even if they are bound to the same extent as the other condition. To avoid such misinterpretation, a spike-in mixture of known concentration at the experimental stage is useful to normalise against (K. Chen, Hu, et al., 2015). In future, capacity to normalise against spike-ins will be added to clipplotr.

Another aspect of normalisation that is important in the context of CLIP data, is normalisation of binding signal to gene expression. When comparing RBP binding across two genes, a researcher may want to know if the increased binding in one gene can be explained by higher gene expression. In order to add such normalisation clipplotr would require gene expression information from the given cell type or tissue, which could be generated from matched RNA-Seq data for example. Currently it is possible for users to plot RNA-Seq data as a coverage track below the crosslink track, but in a figure with many tracks the user might want to condense this information further by normalising the crosslink track by the coverage track, which could be added as an option in the future.

In terms of parameter choice, an issue for both clipplotr and Clippy is the selection of an appropriate smoothing window for the crosslink data, especially when analysing a novel RBP with limited or no prior expectations. Through analysis of multiple RBP datasets we find that a window of 15 nucleotides is usually a good starting point, which can be iteratively adjusted by eye. An ideal situation would be if this parameter could be inferred from the crosslinking pattern itself, such that there was a way to automatically detect broad or punctate binders. However, this problem is not trivial, due to the difficulty of designing a universal criteria that should be optimised against. One possible way would be for the tool to run across varying window sizes and for one to be selected based on enrichment of kmer sequences. The issue with relying on sequence is that some RBPs don't have a strong sequence context and so this approach would fail for them. A recent tool "Stoaty-Dive" aims to classify different categories of peaks within a single CLIP dataset, however PureCLIP peaks extended by 20nt in either direction are used as input, which is already introducing bias into the number and length of peaks called (Heyl & Backofen, 2021).

In the future, more work will be required to establish criteria which describe accurate and useful CLIP peaks. This will be of great help in the automation of parameter selection for peak detection.

Chapter 5

Discussion

In the course of my thesis I have explored computational approaches for analysing crosslinking and immunoprecipitation data in the context of RNA modifications, and in the wider study of RNA binding proteins. I describe three new openly available software packages: a ncRNA-aware end-to-end CLIP analysis pipeline written in the Snakemake workflow language, which I've tested with an array of miCLIP and iCLIP datasets; clipplotr, a tool for smoothing, normalising and presenting multiple CLIP datasets across transcripts of interest alongside supplementary data and finally Clippy, a peak caller with a unique interactive visualisation that enables users to explore parameters on transcripts of interest before applying to whole datasets.

These computational approaches have already facilitated interesting biological insights, in the case of the ncRNA-aware CLIP pipeline I have been able to identify that NSun2 and Trmt2a extensively modify pre-tRNA transcripts and also explore the novel tRNA binding of DEAH-box helicase DDX3X. I used Clippy extensively in the exploration of m6A in yeast meiosis to call peaks across miCLIP and iCLIP datasets. Here the ability to moderate the length of peaks was especially useful in choosing broader peak regions suitable for differential CLIP analysis. clipplotr was born out of the need for a tool capable of visualising many CLIP datasets at once, a problem I encountered in the course of working on the psiCLIP project, which involved $>$90 final iCLIP samples.

In this discussion I will explore the impact of the various findings made in this thesis and the opportunities that present themselves as a result.

5.1 Specificity of m6A and its readers

In chapter 3 I explored iCLIP data for Pho92p produced in both WT and methyltransferase deletion backgrounds by Dr. Radhika Varier. I found that approximately 30% of binding sites could be designated as downregulated in *ime4*-Δ. It's possible that a higher percentage of sites are m6A-dependent, but were beyond our detection limit. Specifically I found that calling sites as m6A-dependent might be harder when the transcript is also downregulated. Despite this, the m6A consensus sequence motif is enriched in m6A-dependent Pho92p binding, but not in the sites I designated as m6A-independent binding, suggesting that these categories are genuinely meaningful.

This observation raises the question of what specifies Pho92p binding: in the case of m6A-dependent binding, why some m6A sites and not others? And in the case of the independent binding, how is this specified in the absence of m6A, or the m6A sequence motif? Despite searching, I couldn't find a convincing preference of Pho92p for certain m6A consensus motifs over others, and alongside previous *in vitro* work it seems unlikely that this is the reason for binding of certain m6As. Rather I suspect the answer lies in U-rich sequences that are bound by Pho92p's intrinsically disordered regions. I found GU repeats enriched in analysis of m6a-independent Pho92p binding sites and perhaps with enough multivalent interactions over degenerate sequence Pho92p is still able to interact with RNA. This binding might be weaker than that specified by m6A, although its worth noting that the m6A binding isn't particularly strong in the first place compared to other RBP-RNA affinities and in itself has been suggested to be stablised by U-rich interactions of intrinsically disordered regions (Arribas-Hernández, Rennie, Köster, et al., 2021). Interestingly the effect of Pho92p on mRNA downregulation/decay was most extreme for Pho92p bound to m6A. We propose this is due to stronger interactions maintaining Pho92p-mRNA binding from the nucleus into the cytoplasm, but this isn't certain. It remains to be investigated what the function of this m6A-independent binding is, if any. As more CLIP data in this model system becomes available it will be interesting to see if Pho92p binding is impacted by nearby binding of other RBPs in U-rich sequence contexts, which could also go part of the way to explaining why certain m6A sites are bound over others.

There could also be a contribution of chromatin factors in positioning Pho92p if it is indeed loaded co-transcriptionally, as the 3' end binding is consistent regardless of sequence at that given position.

The dependency of YTH protein binding on m6A across transcriptomes remains to be explored in human or mice, although extensive redundancy in steady state cell lines might make this hard to investigate. Expression of the different human YTH proteins differs along developmental timelines and in different cellular contexts, therefore perhaps it will prove more fruitful to choose conditions where a certain YTH protein is more dominantly expressed, in order to determine their individual functions and dependency on m6A.

5.2 The promise of single molecule resolution

As described in this thesis, advances in high throughput sequencing methods have vastly improved the detection of m6A on mRNAs. In the case of the yeast m6A project, alongside miCLIP we were also keen to explore Nanopore sequencing as a method that is unaffected by m6A antibody biases and also offers single molecule resolution. Dr. Dora Sideri prepared RNA which was sequenced by Dr. Tommasso Leonardi at the Italian Institute of Technology and analysed using the Nanocompore approach where changes in voltage and retention time of 5mers in the pore are assessed between WT and methyltransferase knockout conditions to identify m6A sites. I contributed to the benchmarking of the method with some comparison to miCLIP data (Leger, Amaral, Pandolfini, Capitanchik, et al., 2019b).

However, I believe that exciting future developments will revolve around analyses that are able to extract information about the combinatorics of RNA modification sites on the same molecule. The dynamics of m6A deposition are currently unclear, its unknown if modification is cooperative, ie. does methylation at one site make methylation at adjacent sites more likely? Preliminary analysis of the Actin transcripts suggests not (Leger, Amaral, Pandolfini, Capitanchik, et al., 2019a), but ideally we would develop a unified computational approach that could summarise such information across all quantifiable sites.

Methods such as MAZTER-seq are able to quantify the proportion of mRNA transcripts of a given gene that are methylated at a given point in time. It would be ideal to validate such quantifications using Nanopore, which should be a more unbiased approach, given that MAZTER-seq requires enzymatic digestion of RNA. While we still know little about how m6A results in mRNA fate changes, we know even less about how this operates across a population of transcripts for a given gene. Are the number of modified transcripts regulated at some level, can this be changed and how? Further, does having more sites of modification on a transcript make a fate change more extreme? This is partially addressed by analyses of bulk modification levels, however in these cases its unclear if transcripts with two modification sites are any more modified than transcripts with one, i.e. modification could be split between the two sites rather than the two sites being constitutively modified.

5.3 The future of bioinformatics in RNA biology and beyond

The software tools presented in this thesis all address burgeoning needs that have emerged in the CLIP field: that of accurate ncRNA quantification and how to visualise ever expanding volumes of CLIP and orthogonal data to reach meaningful biological conclusions. The issue of analysing and visualising large volumes of bioinformatics data is not isolated to the CLIP field alone, but is in fact a growing problem in bioinformatics as a whole. There is also increasing interest in making the latest advances in bioinformatics technologies accessible to the biologists who are producing the data, who are most commonly not bioinformatics experts themselves. Making analyses as straightforward to implement as possible not only benefits those without coding experience, but also lightens the load of experienced bioinformaticians to explore exciting new frontiers.

The rapid adoption of workflow languages such as Nextflow and Snakemake, has greatly simplified bioinformatics processes over a short period of time. As an example, at the beginning of my PhD in 2017, bioinformaticians I knew all had their own version of RNA-Seq workflow, commonly a series of bash scripts. This meant that many bioinformaticians spent valuable research time repeating a task that had been performed thousands of times before, and collaboration between bioinformaticians wasn't straightforward, with different researchers in the same group coming to slightly different results from analysing the same data. Now, it is possible to run a very good RNA-Seq workflow from Nextflow's nf-core project with a single line of code, or even without coding or any access to high performance computing by launching the workflow on the cloud using Nextflow Tower's graphical user interface (GUI) (https://tower.nf). The potential of these advances for accelerating biomedical discovery is awe-inspiring. During the global COVID-19 pandemic, the COVID-19 Genomics UK (COG-UK) Consortium created a platform for integrating and analysing rapidly produced SARS-CoV-2 viral genomes. The platform incorporates several Nextflow pipelines, some of which are run daily - highlighting the robustness of the language. One analyses raw sequencing data (https://github.com/connor-lab/ncov2019-artic-nf), another performs quality control, moves files and write reports (https://github.com/SamStudio8/elan-nextflow/). Datapipe performs multiple sequence analysis of all sequences that pass a quality threshold in order to call variants (https://github.com/COG-UK/datapipe). Further, a Snakemake pipeline called Grapevine runs phylogentic analyses across all database sequences to produce updated phylogenetic trees (https://github.com/COG-UK/grapevine). By July 2021, the platform contained over 550,000 sequences. Before the widespread adoption of workflow languages, such rapid analysis at such a massive scale would have required considerably more work to architect.

From the perspective of CLIP analysis, within Prof. Luscombe's group we developed nf-core/clipseq (https://nf-co.re/clipseq): a gold standard analysis pipeline for CLIP data. As the Nextflow language moves from its first iteration - DSL1, to its second - DSL2, I have worked extensively, alongside others in the Luscombe group, on porting the existing CLIP pipeline from DSL1 to DSL2, whilst also integrating the latest software improvements - including Ultraplex, an ultra-fast demultiplexer (Wilkins, Capitanchik, et al., 2021) and Clippy (https://github.com/goodwright/imaps-nf). The DSL2 language updates support the creation of software modules, community-curated gold standard software modules are made available by the nf-core project, so many common steps have already been written and contributed including those for Bowtie and STAR mapping for example (https://github.com/nf-core/modules) (Ewels et al., 2020).

Separately to these developments, Prof. Jernej Ule has been developing iCLIP analysis server platforms for the past ten years. The iCount server was developed originally in partnership between Prof. Jernej Ule and Prof. Tomaž Curk, representing a free GUI for iCLIP analysis which doubled as a database repository for iCLIP data. As I joined the group in 2017, an updated platform developed in partnership with the company Genialis was released (https://imaps.genialis.com/iclip). The improvements included improved options for data security and sharing alongside the concept of 'Collections'. A collection contains all the data supporting a single project or paper, enabling meaningful organisation of data that can then be shared with other single users or groups of users at varying levels: for example read-only or full access. Further, the iCLIP pipeline underlying the platform was updated, and the look and feel of the website was improved.

With the rapid developments in Nextflow, we saw the opportunity to overhaul the iMaps platform entirely to wrap around Nextflow DSL2 pipelines on the backend and in November 2020 partnered with the company Goodwright, comprised of Dr. Sam Ireland and Alex Harston, to build iMaps 2.0 (https://imaps.goodwright.com/). The power of integrating with Nextflow is multilayered: 1) with appropriate abstraction of iMaps software layers **any** Nextflow

DSL2 pipeline can be easily integrated into the platform: this means we will be able to expand the platform beyond iCLIP analysis to include functional genomics data such as RNA-Seq, QuantSeq, ribosome profiling and also expand beyond RNA to integrate analysis of chromatin layers, for example cut and run (https://nf-co.re/cutandrun). 2) As bioinformaticians are able to use exactly the same version-controlled pipelines accessed through iMaps or through the command line this heightens reproducibility across projects and reduces the inconvenience of achieving different results via the web server than via the command line. Further to the use of Nextflow in the backend, we also made changes to improve the front end, namely to improve user account management - now it is easy to see which groups you are a part of and which datasets you own. Additionally, we introduced more signifiers of a data objects origin, making it easier to trace the flow of data through a pipeline. Further to all of this we created a public iMaps Slack workspace to gather feedback from users and answer questions with the aim of creating a supportive community around the platform (imapsgroup.slack.com).

More than being a platform for analysis, iMaps also functions as a user-populated database. Once users are happy with their collections it is simply one click to make that collection public and searchable. We plan to introduce quality thresholding into the search for data objects, to facilitate integrative analysis across all datasets present in the platform. Other platforms in this field are more limited in scope, as they exclusively perform only one of these functions. The most established provider of bioinformatics analysis workflows by GUI is currently Galaxy, which in late 2020 released the CLIP-explorer pipeline (https://clipseq.usegalaxy.eu/) (Heyl et al., 2020). Unlike nf-core/clip-seq, there is no clear route for users to contribute to the development of CLIP-explorer or to report issues, and the results of user's analyses are not collated and shared publicly, making meta-analyses intractable. Conversely, ENCORI (previously starBase http://starbase.sysu.edu.cn/index.php) and POSTAR2 (http://lulab.life.tsinghua.edu.cn/postar/) are the main CLIP databases, and both suffer from a lack of consistent curation for CLIP data or interactive visualisation (J.-H.

Li et al., 2014; Y. Zhu et al., 2019). Moreover, these databases lack quality control analyses and don't report some of the key analysis parameters, and are therefore in danger of rapidly falling out-of-date in the face of rapidly developing technologies and standards. iMaps overcomes these limitations by being openly developed on GitHub and establishing coherent annotation standards.

Integration of the visualisation tools described in this thesis will enhance iMaps and broaden the user base of the tools, as currently both still require use of the command line to initially run. Further, by porting my ncRNA-aware CLIP pipeline into Nextflow, I will make it more widely accessible to the community and I am certain that the routine accurate quantification of ncRNA will yield useful, unexpected discoveries in the years to come. By streamlining the process of data pre-processing, computational biologists will be able to shift focus on to more interesting downstream analyses. Moreover, by facilitating high quality data curation, exciting meta-analyses will become more viable. In the field of CLIP meta-analyses are often limited to ENCODE eCLIP data due to convenience; however, by populating iMaps with a wide range of data, more expansive meta-analyses will become possible. In the near future, expansion of the iMaps platform to include orthogonal datasets such as RNA-Seq, quantSeq, ribosome profiling and chromatin profiling methodologies will open up even more exciting possibilities for data visualisations and integration. Prioritising datasets across different species, alongside integration with cross-species resources, will enable analysis from an evolutionary vantage point. Powerful modelling approaches, such as deep learning of binding site specificities, require carefully prepared input datasets which are currently incredibly time consuming to curate. Through inclusion of thoughtful quality control filters and thresholds, iMaps should open the doors to novel analyses and insights.

To summarise, I believe these rapid developments in bioinformatics will accelerate insights into RBP biology and beyond. By focusing our attention on how to improve and facilitate collaboration, how to present data in a biologically meaningful way and by making usability a priority, every researcher benefits.

Bibliography

Agarwala, S. D., Blitzblau, H. G., Hochwagen, A., & Fink, G. R. (2012). RNA methylation by the MIS complex regulates a cell fate decision in yeast. *PLoS genetics*, *8*(6), e1002732. https://doi.org/10.1371/journal.pgen.1002732

Akichika, S., Hirano, S., Shichino, Y., Suzuki, T., Nishimasu, H., Ishitani, R., Sugita, A., Hirose, Y., Iwasaki, S., Nureki, O., & Suzuki, T. (2019). Cap-specific terminal N 6-methylation of RNA by an RNA polymerase II-associated methyltransferase. *Science*, *363*(6423). https://doi.org/10.1126/science.aav0080

Andrews, S. et al. (2010). FastQC: A quality control tool for high throughput sequence data.

Anishchenko, L., Evteeva, K., Korostovtseva, L., Bochkarev, M., & Sviryaev, Y. (2019). Respiratory rate determination during sleep by CW doppler radar. *2019 International Conference on Biomedical Innovations and Applications (BIA)*, 1–4. https://doi.org/10.1109/BIA48344.2019.8967462

Anreiter, I., Mir, Q., Simpson, J. T., Janga, S. C., & Soller, M. (2021). New twists in detecting mRNA modification dynamics. *Trends in biotechnology*, *39*(1), 72–89. https://doi.org/10.1016/j.tibtech.2020.06.002

Arribas-Hernández, L., Rennie, S., Köster, T., Porcelli, C., Lewinski, M., Dr Dorothee Staiger, P., Andersson, R., & Brodersen, P. (2021). Principles of mRNA targeting via the arabidopsis m6a-binding protein ECT2. *eLife*, *10*. https://doi.org/10.7554/eLife.72375

Arribas-Hernández, L., Rennie, S., Schon, M., Porcelli, C., Enugutti, B., Andersson, R., Nodine, M. D., & Brodersen, P. (2021). The YTHDF proteins ECT2 and ECT3 bind largely overlapping target sets and influence target mRNA abundance, not alternative polyadenylation. *eLife*, *10*. https://doi.org/10.7554/eLife.72377

Attig, J., Agostini, F., Gooding, C., Chakrabarti, A. M., Singh, A., Haberman, N., Zagalak, J. A., Emmett, W., Smith, C. W. J., Luscombe, N. M., & Ule, J. (2018). Heteromeric RNP assembly at LINEs controls Lineage-Specific RNA processing. *Cell*, *174*(5), 1067–1081.e17. https://doi.org/10.1016/j.cell.2018.07.001

Attig, J., & Ule, J. (2019). Genomic accumulation of retrotransposons was facilitated by repressive RNA-Binding proteins: A hypothesis. *BioEssays: news and reviews in molecular, cellular and developmental biology*, e1800132. https://doi.org/10.1002/bies.201800132

Aviner, R., Hofmann, S., Elman, T., Shenoy, A., Geiger, T., Elkon, R., Ehrlich, M., & Elroy-Stein, O. (2017). Proteomic analysis of polyribosomes identifies splicing factors as potential regulators of translation during mitosis. *Nucleic acids research*, *45*(10), 5945–5957. https://doi.org/10.1093/nar/gkx326

Bailey, T. L. (2011). DREME: Motif discovery in transcription factor ChIP-seq data. *Bioinformatics*, *27*(12), 1653–1659. https://doi.org/10.1093/bioinformatics/btr261

Bailey, T. L. (2021). STREME: Accurate and versatile sequence motif discovery. *Bioinformatics*. https://doi.org/10.1093/bioinformatics/btab203

Bao, W., Kojima, K. K., & Kohany, O. (2015). Repbase update, a database of repetitive elements in eukaryotic genomes. *Mobile DNA*, *6*, 11. https://doi.org/10.1186/s13100-015-0041-9

Batista, P. J., Molinie, B., Wang, J., Qu, K., Zhang, J., Li, L., Bouley, D. M., Lujan, E., Haddad, B., Daneshvar, K., Carter, A. C., Flynn, R. A., Zhou, C., Lim, K.-S., Dedon, P., Wernig, M., Mullen, A. C., Xing, Y., Giallourakis, C. C., & Chang, H. Y. (2014). M(6)a RNA modification controls cell fate transition in mammalian embryonic stem cells. *Cell stem cell*, *15*(6), 707–719. https://doi.org/10.1016/j.stem.2014.09.019

Baumgarten, S., Bryant, J. M., Sinha, A., Reyser, T., Preiser, P. R., Dedon, P. C., & Scherf, A. (2019). Transcriptome-wide dynamics of extensive m6a mRNA methylation during plasmodium falciparum blood-stage development. *Nature microbiology*, *4*(12), 2246–2259. https://doi.org/10.1038/s41564-019-0521-7

Becker, S., Schneider, C., Crisp, A., & Carell, T. (2018). Non-canonical nucleosides and chemistry of the emergence of life. *Nature communications*, *9*(1), 5174. https://doi.org/10.1038/s41467-018-07222-w

Behrens, A., Rodschinka, G., & Nedialkova, D. D. (2021). High-resolution quantitative profiling of tRNA abundance and modification status in eukaryotes by mim-tRNAseq. *Molecular cell*. https://doi.org/10.1016/j.molcel.2021.01.028

Björk, G. R., Wikström, P. M., & Byström, A. S. (1989). Prevention of translational frameshifting by the modified nucleoside 1-methylguanosine. *Science*, *244*(4907), 986–989. https://doi.org/10.1126/science.2471265

Blanco, S., Dietmann, S., Flores, J. V., Hussain, S., Kutter, C., Humphreys, P., Lukk, M., Lombard, P., Treps, L., Popis, M., Kellner, S., Hölter, S. M., Garrett, L., Wurst, W., Becker, L., Klopstock, T., Fuchs, H., Gailus-Durner, V., Hrabě de Angelis, M., ... Frye, M. (2014). Aberrant methylation of tRNAs links cellular stress to neuro-developmental disorders. *The EMBO journal*, *33*(18), 2020–2039. https://doi.org/10.15252/embj.201489282

Blazquez, L., Emmett, W., Faraway, R., Pineda, J. M. B., Bajew, S., Gohr, A., Haberman, N., Sibley, C. R., Bradley, R. K., Irimia, M., & Ule, J. (2018). Exon junction complex shapes the transcriptome by repressing recursive splicing. *Molecular cell*, *72*(3), 496–509.e9. https://doi.org/10.1016/j.molcel.2018.09.033

Bodi, Z., Bottley, A., Archer, N., May, S. T., & Fray, R. G. (2015). Yeast m6a methylated mRNAs are enriched on translating ribosomes during meiosis, and under rapamycin treatment. *PloS one*, *10*(7), e0132090. https://doi.org/10.1371/journal.pone.0132090

Bodi, Z., Button, J. D., Grierson, D., & Fray, R. G. (2010). Yeast targets for mRNA methylation. *Nucleic acids research*, *38*(16), 5327–5335. https://doi.org/10.1093/nar/gkq266

Brar, G. A., Yassour, M., Friedman, N., Regev, A., Ingolia, N. T., & Weissman, J. S. (2012). High-resolution view of the yeast meiotic program revealed by ribosome profiling. *Science*, *335*(6068), 552–557. https://doi.org/10.1126/science.1215110

Briese, M., Haberman, N., Sibley, C. R., Faraway, R., Elser, A. S., Chakrabarti, A. M., Wang, Z., König, J., Perera, D., Wickramasinghe, V. O., Venkitaraman, A. R., Luscombe, N. M., Saieva, L., Pellizzoni, L., Smith, C. W. J., Curk, T., & Ule, J. (2019). A systems view of spliceosomal assembly and branchpoints with iCLIP. *Nature structural & molecular biology*, *26*(10), 930–940. https://doi.org/10.1038/s41594-019-0300-4

Brzezicha, B., Schmidt, M., Makalowska, I., Jarmolowski, A., Pienkowska, J., & Szweykowska-Kulinska, Z. (2006). Identification of human tRNA:m5C methyltransferase catalysing intron-dependent m5c formation in the first position of the anticodon of the pre-tRNA leu (CAA). *Nucleic acids research*, *34*(20), 6034–6043. https://doi.org/10.1093/nar/gkl765

Buzdin, A., Ustyugova, S., Gogvadze, E., Vinogradova, T., Lebedev, Y., & Sverdlov, E. (2002). A new family of chimeric retrotranscripts formed by a full copy of U6 small nuclear RNA fused to the 3' terminus of l1. *Genomics*, *80*(4), 402–406. https://doi.org/10.1006/geno.2002.6843

Capitanchik, C., Toolan-Kerr, P., Luscombe, N. M., & Ule, J. (2020). How do you identify m6 a methylation in transcriptomes at high resolution? a comparison of recent datasets. *Frontiers in genetics*, *11*, 398. https://doi.org/10.3389/fgene.2020.00398

Carlile, T. M., Rojas-Duran, M. F., Zinshteyn, B., Shin, H., Bartoli, K. M., & Gilbert, W. V. (2014). Pseudouridine profiling reveals regulated mRNA pseudouridylation in yeast and human cells. *Nature*, *515*(7525), 143–146. https://doi.org/10.1038/nature13802

Carter, J.-M., Emmett, W., Mozos, I. R., Kotter, A., Helm, M., Ule, J., & Hussain, S. (2019). FICC-Seq: A method for enzyme-specified profiling of methyl-5-uridine in cellular RNA. *Nucleic acids research*. https://doi.org/10.1093/nar/gkz658

Chakrabarti, A. M., Capitanchik, C., Ule, J., & Luscombe, N. M. (2021). *Clipplotr - a comparative visualisation and analysis tool for CLIP data*. https://doi.org/10.1101/2021.09.10.459763

Chakrabarti, A. M., Haberman, N., Praznik, A., Luscombe, N. M., & Ule, J. (2018). Data science issues in studying Protein–RNA interactions with CLIP technologies. *Annual Review of Biomedical Data Science*, *1*(1), 235–261. https://doi.org/10.1146/annurev-biodatasci-080917-013525

Chan, P. P., & Lowe, T. M. (2009). GtRNAdb: A database of transfer RNA genes detected in genomic sequence. *Nucleic acids research*, *37*(Database issue), D93–7. https://doi.org/10.1093/nar/gkn787

Chan, P. P., & Lowe, T. M. (2016). GtRNAdb 2.0: An expanded database of transfer RNA genes identified in complete and draft genomes. *Nucleic acids research*, *44*(D1), D184–9. https://doi.org/10.1093/nar/gkv1309

Chen, H. C., Tseng, C. K., Tsai, R. T., Chung, C. S., & Cheng, S. C. (2013). Link of NTR-mediated spliceosome disassembly with DEAH-box ATPases prp2, prp16, and prp22. *Molecular and cellular biology*, *33*(3), 514–525. https://doi.org/10.1128/mcb.01093-12

Chen, K., Lu, Z., Wang, X., Fu, Y., Luo, G.-Z., Liu, N., Han, D., Dominissini, D., Dai, Q., Pan, T., & He, C. (2015). High-ResolutionN6-Methyladenosine (m6a) map using Photo-Crosslinking-Assisted m6a sequencing. *Angewandte Chemie*, *127*(5), 1607–1610. https://doi.org/10.1002/ange.201410647

Chen, K., Hu, Z., Xia, Z., Zhao, D., Li, W., & Tyler, J. K. (2015). The overlooked fact: Fundamental need for Spike-In control for virtually all Genome-Wide analyses. *Molecular and cellular biology*, *36*(5), 662–667. https://doi.org/10.1128/MCB.00970-14

Chen, R. E., & Thorner, J. (2007). Function and regulation in MAPK signaling pathways: Lessons learned from the yeast saccharomyces cerevisiae. *Biochimica et biophysica acta*, *1773*(8), 1311–1340. https://doi.org/10.1016/j.bbamcr.2007.05.003

Chen, X.-Y., Zhang, J., & Zhu, J.-S. (2019). The role of m6a RNA methylation in human cancer. *Molecular cancer*, *18*(1), 103. https://doi.org/10.1186/s12943-019-1033-z

Chen, Y., Sierzputowska-Gracz, H., Guenther, R., Everett, K., & Agris, P. F. (1993). 5-methylcytidine is required for cooperative binding of mg2+ and a conformational transition at the anticodon stem-loop of yeast phenylalanine tRNA. *Biochemistry*, *32*(38), 10249–10253. https://doi.org/10.1021/bi00089a047

Corcoran, D. L., Georgiev, S., Mukherjee, N., Gottwein, E., Skalsky, R. L., Keene, J. D., & Ohler, U. (2011). PARalyzer: Definition of RNA binding sites from PAR-CLIP short-read sequence data. *Genome biology*, *12*(8), R79. https://doi.org/10.1186/gb-2011-12-8-r79

Cordin, O., & Beggs, J. D. (2013). RNA helicases in splicing. *RNA biology*, *10*(1), 83–95. https://doi.org/10.4161/rna.22547

Dao, V., Guenther, R., Malkiewicz, A., Nawrot, B., Sochacka, E., Kraszewski, A., Jankowska, J., Everett, K., & Agris, P. F. (1994). Ribosome binding of DNA analogs of tRNA requires base modifications and supports the "extended anticodon". *Proceedings of the National Academy of Sciences of the United States of America*, *91*(6), 2125–2129. https://doi.org/10.1073/pnas.91.6.2125

Davis, D. R. (1995). Stabilization of RNA stacking by pseudouridine. *Nucleic acids research*, *23*(24), 5020–5026. https://doi.org/10.1093/nar/23.24.5020

Decatur, W. A., & Fournier, M. J. (2002). rRNA modifications and ribosome function. *Trends in biochemical sciences*, *27*(7), 344–351. https://doi.org/10.1016/s0968-0004(02)02109-6

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., & Gingeras, T. R. (2013). STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics*, *29*(1), 15–21. https://doi.org/10.1093/bioinformatics/bts635

Dominissini, D., Moshitch-Moshkovitz, S., Schwartz, S., Salmon-Divon, M., Ungar, L., Osenberg, S., Cesarkas, K., Jacob-Hirsch, J., Amariglio, N., Kupiec, M., Sorek, R., & Rechavi, G. (2012). Topology of the human and mouse m6a RNA methylomes revealed by m6a-seq. *Nature*, *485*(7397), 201–206. https://doi.org/10.1038/nature11112

Dominski, Z., & Marzluff, W. F. (1999). Formation of the 3' end of histone mRNA. *Gene*, *239*(1), 1–14. https://doi.org/10.1016/s0378-1119(99)00367-4

Dowle, M., Srinivasan, A., Gorecki, J., Chirico, M., Stetsenko, P., Short, T., Lianoglou, S., Antonyan, E., Bonsch, M., Parsonage, H., et al. (2019). Package 'data. table'. *Extension of 'data. frame.*

Drino, A., Oberbauer, V., Troger, C., Janisiw, E., Anrather, D., Hartl, M., Kaiser, S., Kellner, S., & Schaefer, M. R. (2020). Production and purification of endogenously modified tRNA-derived small RNAs. *RNA biology*, *17*(8), 1104–1115. https://doi.org/10.1080/15476286.2020.1733798

Du, H., Zhao, Y., He, J., Zhang, Y., Xi, H., Liu, M., Ma, J., & Wu, L. (2016). YTHDF2 destabilizes m(6)a-containing RNA through direct recruitment of the CCR4-NOT deadenylase complex. *Nature communications*, *7*, 12626. https://doi.org/10.1038/ncomms12626

Egloff, S., Vitali, P., Tellier, M., Raffel, R., Murphy, S., & Kiss, T. (2017). The 7SK snRNP associates with the little elongation complex to promote snRNA gene expression. *The EMBO journal*, *36*(7), 934–948. https://doi.org/10.15252/embj.201695740

Ewels, P. A., Peltzer, A., Fillinger, S., Patel, H., Alneberg, J., Wilm, A., Garcia, M. U., Di Tommaso, P., & Nahnsen, S. (2020). The nf-core framework for community-curated bioinformatics pipelines. *Nature biotechnology*, *38*(3), 276–278. https://doi.org/10.1038/s41587-020-0439-x

Faraway, R. (2021). *Non-canonical spliceosomal assembly and function* (U. Jernej, Ed.; Doctoral dissertation). UCL (Univeristy College London). UCL (Univeristy College London).

Farhat, D. C., Bowler, M. W., Communie, G., Pontier, D., Belmudes, L., Mas, C., Corrao, C., Couté, Y., Bougdour, A., Lagrange, T., Hakimi, M.-A., & Swale, C. (2021). A plant-like mechanism coupling m6a reading to polyadenylation safeguards transcriptome integrity and developmental gene partitioning in toxoplasma. *eLife*, *10*. https://doi.org/10.7554/eLife.68312

Flynn, R. A., Martin, L., Spitale, R. C., Do, B. T., Sagan, S. M., Zarnegar, B., Qu, K., Khavari, P. A., Quake, S. R., Sarnow, P., & Chang, H. Y. (2015). Dissecting noncoding and pathogen RNA–protein interactomes. *RNA*, *21*(1), 135–143. https://doi.org/10.1261/rna.047803.114

Freibaum, B. D., Chitta, R. K., High, A. A., & Taylor, J. P. (2010). Global analysis of TDP-43 interacting proteins reveals strong association with RNA splicing and translation machinery. *Journal of proteome research*, *9*(2), 1104–1120. https://doi.org/10.1021/pr901076y

Frith, M. C., Valen, E., Krogh, A., Hayashizaki, Y., Carninci, P., & Sandelin, A. (2008). A code for transcription initiation in mammalian genomes. *Genome research*, *18*(1), 1–12. https://doi.org/10.1101/gr.6831208

Frye, M., Harada, B. T., Behm, M., & He, C. (2018). RNA modifications modulate gene expression during development. *Science*, *361*(6409), 1346–1349. https://doi.org/10.1126/science.aau1646

Gallie, D. R. (1991). The cap and poly(a) tail function synergistically to regulate mRNA translational efficiency. *Genes & development*, *5*(11), 2108–2116. https://doi.org/10.1101/gad.5.11.2108

Garcia-Campos, M. A., Edelheit, S., Toth, U., Safra, M., Shachar, R., Viukov, S., Winkler, R., Nir, R., Lasman, L., Brandis, A., Hanna, J. H., Rossmanith, W., & Schwartz, S. (2019). Deciphering the "m6a code" via Antibody-Independent quantitative profiling. *Cell*, *178*(3), 731–747.e16. https://doi.org/10.1016/j.cell.2019.06.013

Gentleman, R. C., Carey, V. J., Bates, D. M., Bolstad, B., Dettling, M., Dudoit, S., Ellis, B., Gautier, L., Ge, Y., Gentry, J., Hornik, K., Hothorn, T., Huber, W., Iacus, S., Irizarry, R., Leisch, F., Li, C., Maechler, M., Rossini, A. J., . . . Zhang, J. (2004). Bioconductor: Open software development for computational biology and bioinformatics. *Genome biology*, *5*(10), R80. https://doi.org/10.1186/gb-2004-5-10-r80

Geula, S., Moshitch-Moshkovitz, S., Dominissini, D., Mansour, A. A., Kol, N., Salmon-Divon, M., Hershkovitz, V., Peer, E., Mor, N., Manor, Y. S., Ben-Haim, M. S., Eyal, E., Yunger, S., Pinto, Y., Jaitin, D. A., Viukov, S., Rais, Y., Krupalnik, V., Chomsky, E., . . . Hanna, J. H. (2015). Stem cells. m6a mRNA methylation facilitates resolution of naıve pluripotency toward differentiation. *Science*, *347*(6225), 1002–1006. https://doi.org/10.1126/science.1261417

Gogakos, T., Brown, M., Garzia, A., Meyer, C., Hafner, M., & Tuschl, T. (2017). Characterizing expression and processing of precursor and mature human tRNAs by Hydro-tRNAseq and PAR-CLIP. *Cell reports*, *20*(6), 1463–1475. https://doi.org/10.1016/j.celrep.2017.07.029

Grishin, N. V. (1998). The R3H motif: A domain that binds single-stranded nucleic acids. *Trends in biochemical sciences*, *23*(9), 329–330. https://doi.org/10.1016/s0968-0004(98)01258-4

Grüning, B., Dale, R., Sjödin, A., Chapman, B. A., Rowe, J., Tomkins-Tinch, C. H., Valieris, R., Köster, J., & Bioconda Team. (2018). Bioconda: Sustainable and comprehensive software distribution for the life sciences. *Nature methods*, *15*(7), 475–476. https://doi.org/10.1038/s41592-018-0046-7

Guacci, V., Koshland, D., & Strunnikov, A. (1997). A direct link between sister chromatid cohesion and chromosome condensation revealed through the analysis of MCD1 in s. cerevisiae. *Cell*, *91*(1), 47–57. https://doi.org/10.1016/s0092-8674(01)80008-8

Gueroussov, S., Gonatopoulos-Pournatzis, T., Irimia, M., Raj, B., Lin, Z.-Y., Gingras, A.-C., & Blencowe, B. J. (2015). An alternative splicing event amplifies evolutionary differences between vertebrates. *Science*, *349*(6250), 868–873. https://doi.org/10.1126/science.aaa8381

Guttmann-Raviv, N., Martin, S., & Kassir, Y. (2002). Ime2, a meiosis-specific kinase in yeast, is required for destabilization of its transcriptional activator, ime1. *Molecular and cellular biology*, *22*(7), 2047–2056. https://doi.org/10.1128/MCB.22.7.2047-2056.2002

Haberman, N., Huppertz, I., Attig, J., König, J., Wang, Z., Hauer, C., Hentze, M. W., Kulozik, A. E., Le Hir, H., Curk, T., Sibley, C. R., Zarnack, K., & Ule, J. (2017). Insights into the design and interpretation of iCLIP experiments. *Genome biology*, *18*(1), 7. https://doi.org/10.1186/s13059-016-1130-x

Haley, S., Behnke, S., Edens, H., et al. (2021). Observations show charge density of volcanic plumes is higher than thunderstorms. *Journal of geophysical research*.

Hallegger, M., Chakrabarti, A. M., Lee, F. C. Y., Lee, B. L., Amalietti, A. G., Odeh, H. M., Copley, K. E., Rubien, J. D., Portz, B., Kuret, K., Huppertz, I., Rau, F., Patani, R., Fawzi, N. L., Shorter, J., Luscombe, N. M., & Ule, J. (2021). TDP-43 condensation properties specify its RNA-binding and regulatory repertoire. *Cell*, *184*(18), 4680–4696.e22. https://doi.org/10.1016/j.cell.2021.07.018

Hamilton, N. (2015). Smoother: Functions relating to the smoothing of numerical data [Accessed: 2020-1-4].

Han, Y., Feng, J., Xia, L., Dong, X., Zhang, X., Zhang, S., Miao, Y., Xu, Q., Xiao, S., Zuo, Z., Xia, L., & He, C. (2019). CVm6A: A visualization and exploration database for mas in cell lines. *Cells*, *8*(2). https://doi.org/10.3390/cells8020168

Hanna, J. S., Kroll, E. S., Lundblad, V., & Spencer, F. A. (2001). Saccharomyces cerevisiae CTF18 and CTF4 are required for sister chromatid cohesion. *Molecular and cellular biology*, *21*(9), 3144–3158. https://doi.org/10.1128/MCB.21.9.3144-3158.2001

Haussmann, I. U., Bodi, Z., Sanchez-Moran, E., Mongan, N. P., Archer, N., Fray, R. G., & Soller, M. (2016). M(6)a potentiates sxl alternative pre-mRNA splicing for robust drosophila sex determination. *Nature*, *540*(7632), 301–304. https://doi.org/10.1038/nature20577

He, Y., Staley, J. P., Andersen, G. R., & Nielsen, K. H. (2017). Structure of the DEAH/RHA ATPase prp43p bound to RNA implicates a pair of hairpins and motif va in translocation along RNA. *RNA*, *23*(7), 1110–1124. https://doi.org/10.1261/rna.060954.117

Herzog, V. A., Reichholf, B., Neumann, T., Rescheneder, P., Bhat, P., Burkard, T. R., Wlotzka, W., von Haeseler, A., Zuber, J., & Ameres, S. L. (2017). Thiol-linked alkylation of RNA to assess expression dynamics. *Nature methods*, *14*(12), 1198–1204. https://doi.org/10.1038/nmeth.4435

Heyl, F., & Backofen, R. (2021). StoatyDive: Evaluation and classification of peak profiles for sequencing data. *GigaScience*, *10*(6). https://doi.org/10.1093/gigascience/giab045

Heyl, F., Maticzka, D., Uhl, M., & Backofen, R. (2020). Galaxy CLIP-Explorer: A web server for CLIP-Seq data analysis. *GigaScience*, *9*(11). https://doi.org/10.1093/gigascience/giaa108

Higashi, S., Kabuta, T., Nagai, Y., Tsuchiya, Y., Akiyama, H., & Wada, K. (2013). TDP-43 associates with stalled ribosomes and contributes to cell survival during cellular stress. *Journal of neurochemistry*, *126*(2), 288–300. https://doi.org/10.1111/jnc.12194

Hodnett, J. L., & Busch, H. (1968). Isolation and characterization of uridylic acid-rich 7 S ribonucleic acid of rat liver nuclei. *The Journal of biological chemistry*, *243*(24), 6334–6342.

Hongay, C. F., Grisafi, P. L., Galitski, T., & Fink, G. R. (2006). Antisense transcription controls cell fate in saccharomyces cerevisiae. *Cell*, *127*(4), 735–745. https://doi.org/10.1016/j.cell.2006.09.038

Horowitz, S., Horowitz, A., Nilsen, T. W., Munns, T. W., & Rottman, F. M. (1984). Mapping of n6-methyladenosine residues in bovine prolactin mRNA. *Proceedings of the National Academy of Sciences of the United States of America*, *81*(18), 5667–5671. https://doi.org/10.1073/pnas.81.18.5667

Hou, G., Zhao, X., Li, L., Yang, Q., Liu, X., Huang, C., Lu, R., Chen, R., Wang, Y., Jiang, B., & Yu, J. (2021). SUMOylation of YTHDF2 promotes mRNA degradation and cancer progression by increasing its binding affinity with m6a-modified mRNAs. *Nucleic acids research*. https://doi.org/10.1093/nar/gkab065

Hu, L., Liu, S., Peng, Y., Ge, R., Su, R., Senevirathne, C., Harada, B. T., Dai, Q., Wei, J., Zhang, L., Hao, Z., Luo, L., Wang, H., Wang, Y., Luo, M., Chen, M., Chen, J., & He, C. (2022). M6a RNA modifications are measured at single-base resolution across the mammalian transcriptome. *Nature biotechnology*, 1–10. https://doi.org/10.1038/s41587-022-01243-z

Huang, H., Weng, H., Sun, W., Qin, X., Shi, H., Wu, H., Zhao, B. S., Mesquita, A., Liu, C., Yuan, C. L., Hu, Y.-C., Hüttelmaier, S., Skibbe, J. R., Su, R., Deng, X., Dong, L., Sun, M., Li, C., Nachtergaele, S., . . . Chen, J. (2018). Recognition of RNA n6-methyladenosine by IGF2BP proteins enhances mRNA stability and translation. *Nature cell biology*, *20*(3), 285–295. https://doi.org/10.1038/s41556-018-0045-z

Huang, W., Qi, C.-B., Lv, S.-W., Xie, M., Feng, Y.-Q., Huang, W.-H., & Yuan, B.-F. (2016). Determination of DNA and RNA methylation in circulating tumor cells by mass spectrometry. *Analytical chemistry*, *88*(2), 1378–1384. https://doi.org/10.1021/acs.analchem.5b03962

Huber, W., Carey, V. J., Gentleman, R., Anders, S., Carlson, M., Carvalho, B. S., Bravo, H. C., Davis, S., Gatto, L., Girke, T., Gottardo, R., Hahne, F., Hansen, K. D., Irizarry, R. A., Lawrence, M., Love, M. I., MacDonald, J., Obenchain, V., Oleś, A. K., . . . Morgan, M. (2015). Orchestrating high-throughput genomic analysis with bioconductor. *Nature methods*, *12*(2), 115–121. https://doi.org/10.1038/nmeth.3252

Hussain, S., Sajini, A. A., Blanco, S., Dietmann, S., Lombard, P., Sugimoto, Y., Paramor, M., Gleeson, J. G., Odom, D. T., Ule, J., & Frye, M. (2013). NSun2-mediated cytosine-5 methylation of vault noncoding RNA determines its processing into regulatory small RNAs. *Cell reports*, *4*(2), 255–261. https://doi.org/10.1016/j.celrep.2013.06.029

Hyde, J. L., & Diamond, M. S. (2015). Innate immune restriction and antagonism of viral RNA lacking 2-o methylation. *Virology*, *479-480*, 66–74. https://doi.org/10.1016/j.virol.2015.01.019

Imanishi, M., Tsuji, S., Suda, A., & Futaki, S. (2017). Detection of n6-methyladenosine based on the methyl-sensitivity of MazF RNA endonuclease. *Chemical communications*, *53*(96), 12930–12933. https://doi.org/10.1039/c7cc07699a

Ioannidis, J. P. A., Allison, D. B., Ball, C. A., Coulibaly, I., Cui, X., Culhane, A. C., Falchi, M., Furlanello, C., Game, L., Jurman, G., Mangion, J., Mehta, T., Nitzberg, M., Page, G. P., Petretto, E., & van Noort, V. (2009). Repeatability of published microarray gene expression analyses. *Nature genetics*, *41*(2), 149–155. https://doi.org/10.1038/ng.295

Ishigami, Y., Ohira, T., Isokawa, Y., Suzuki, Y., & Suzuki, T. (2021). A single m6a modification in U6 snRNA diversifies exon sequence at the 5' splice site. *Nature communications*, *12*(1), 3244. https://doi.org/10.1038/s41467-021-23457-6

Jain, D., Puno, M. R., Meydan, C., Lailler, N., Mason, C. E., Lima, C. D., Anderson, K. V., & Keeney, S. (2018). Ketu mutant mice uncover an essential meiotic function for the ancient RNA helicase YTHDC2. *eLife*, *7*. https://doi.org/10.7554/eLife.30919

Jorgensen, P., Nishikawa, J. L., Breitkreutz, B.-J., & Tyers, M. (2002). Systematic identification of pathways that couple cell growth and division in yeast. *Science*, *297*(5580), 395–400. https://doi.org/10.1126/science.1070850

Joseph-Strauss, D., Zenvirth, D., Simchen, G., & Barkai, N. (2007). Spore germination in saccharomyces cerevisiae: Global gene expression patterns and cell cycle landmarks. *Genome biology*, *8*(11), R241. https://doi.org/10.1186/gb-2007-8-11-r241

Kan, L., Grozhik, A. V., Vedanayagam, J., Patil, D. P., Pang, N., Lim, K.-S., Huang, Y.-C., Joseph, B., Lin, C.-J., Despic, V., Guo, J., Yan, D., Kondo, S., Deng, W.-M., Dedon, P. C., Jaffrey, S. R., & Lai, E. C. (2017). The m(6)a pathway facilitates sex determination in drosophila. *Nature communications*, *8*, 15737. https://doi.org/10.1038/ncomms15737

Kan, L., Ott, S., Joseph, B., Park, E. S., Dai, W., Kleiner, R. E., Claridge-Chang, A., & Lai, E. C. (2021). A neural m6A/Ythdf pathway is required for learning and memory in drosophila. *Nature communications*, *12*(1), 1458. https://doi.org/10.1038/s41467-021-21537-1

Kane, S. E., & Beemon, K. (1985). Precise localization of m6a in rous sarcoma virus RNA reveals clustering of methylation sites: Implications for RNA processing. *Molecular and cellular biology*, *5*(9), 2298–2306. https://doi.org/10.1128/mcb.5.9.2298-2306.1985

Kang, H.-J., Jeong, S.-J., Kim, K.-N., Baek, I.-J., Chang, M., Kang, C.-M., Park, Y.-S., & Yun, C.-W. (2014). A novel protein, pho92, has a conserved YTH domain and regulates phosphate metabolism by decreasing the mRNA stability of PHO4 in saccharomyces cerevisiae. *Biochemical Journal*, *457*(3), 391–400. https://doi.org/10.1042/BJ20130862

Karolchik, D., Hinrichs, A. S., Furey, T. S., Roskin, K. M., Sugnet, C. W., Haussler, D., & Kent, W. J. (2004). The UCSC table browser data retrieval tool. *Nucleic acids research*, *32*(Database issue), D493–6. https://doi.org/10.1093/nar/gkh103

Ke, S., Alemu, E. A., Mertens, C., Gantman, E. C., Fak, J. J., Mele, A., Haripal, B., Zucker-Scharff, I., Moore, M. J., Park, C. Y., Vågbø, C. B., Kusśnierczyk, A., Klungland, A., Darnell, J. E., Jr, & Darnell, R. B. (2015). A majority of m6a residues are in the last exons, allowing the potential for 3' UTR regulation. *Genes & development*, *29*(19), 2037–2053. https://doi.org/10.1101/gad.269415.115

Ke, S., Pandya-Jones, A., Saito, Y., Fak, J. J., Vågbø, C. B., Geula, S., Hanna, J. H., Black, D. L., Darnell, J. E., Jr, & Darnell, R. B. (2017). M6a mRNA modifications are deposited in nascent pre-mRNA and are not required for splicing but do specify cytoplasmic turnover. *Genes & development*, *31*(10), 990–1006. https://doi.org/10.1101/gad.301036.117

Keffer-Wilkes, L. C., Soon, E. F., & Kothe, U. (2020). The methyltransferase TrmA facilitates tRNA folding through interaction with its RNA-binding domain. *Nucleic acids research*, *48*(14), 7981–7990. https://doi.org/10.1093/nar/gkaa548

Kim, D., Langmead, B., & Salzberg, S. L. (2015). HISAT: A fast spliced aligner with low memory requirements. *Nature methods*, *12*(4), 357–360. https://doi.org/10.1038/nmeth.3317

Klassen, R., Bruch, A., & Schaffrath, R. (2017). Independent suppression of ribosomal +1 frameshifts by different tRNA anticodon loop modifications. *RNA biology*, *14*(9), 1252–1259. https://doi.org/10.1080/15476286.2016.1267098

Kojima, K. K. (2018). Human transposable elements in repbase: Genomic footprints from fish to humans. *Mobile DNA*, *9*, 2. https://doi.org/10.1186/s13100-017-0107-y

König, J., Zarnack, K., Rot, G., Curk, T., Kayikci, M., Zupan, B., Turner, D. J., Luscombe, N. M., & Ule, J. (2010). iCLIP reveals the function of hnRNP particles in splicing at individual nucleotide resolution. *Nature structural & molecular biology*, *17*(7), 909–915. https://doi.org/10.1038/nsmb.1838

Kontur, C., Jeong, M., Cifuentes, D., & Giraldez, A. J. (2020). Ythdf m6a readers function redundantly during zebrafish development. *Cell reports*, *33*(13), 108598. https://doi.org/10.1016/j.celrep.2020.108598

Körtel, N., Rücklé, C., Zhou, Y., Busch, A., Hoch-Kraft, P., Sutandy, F. X. R., Haase, J., Pradhan, M., Musheev, M., Ostareck, D., Ostareck-Lederer, A., Dieterich, C., Hüttelmaier, S., Niehrs, C., Rausch, O., Dominissini, D., König, J., & Zarnack, K. (2021). Deep and accurate detection of m6a RNA modifications using miCLIP2 and m6aboost machine learning. *Nucleic acids research*. https://doi.org/10.1093/nar/gkab485

Köster, J., & Rahmann, S. (2012). Snakemake—a scalable bioinformatics workflow engine. *Bioinformatics*, *28*(19), 2520–2522. https://doi.org/10.1093/bioinformatics/bts480

Krakau, S., Richard, H., & Marsico, A. (2017). PureCLIP: Capturing target-specific protein–RNA interaction footprints from single-nucleotide CLIP-seq data. *Genome biology*, *18*(1), 240.

Krueger, F. (2017). Trim galore!

Kuhn, R. M., Haussler, D., & Kent, W. J. (2013). The UCSC genome browser and associated tools. *Briefings in bioinformatics*, *14*(2), 144–161. https://doi.org/10.1093/bib/bbs038

Kuleshov, M. V., Jones, M. R., Rouillard, A. D., Fernandez, N. F., Duan, Q., Wang, Z., Koplev, S., Jenkins, S. L., Jagodnik, K. M., Lachmann, A., McDermott, M. G., Monteiro, C. D., Gundersen, G. W., & Ma'ayan, A. (2016). Enrichr: A comprehensive gene set enrichment analysis web server 2016 update. *Nucleic acids research*, *44*(W1), W90–7. https://doi.org/10.1093/nar/gkw377

Kuret, K., Amalietti, A. G., & Ule, J. (2021). *Positional motif analysis reveals the extent of specificity of protein-RNA interactions observed by CLIP*. https://doi.org/10.1101/2021.12.07.471544

Lambert, N., Robertson, A., Jangi, M., McGeary, S., Sharp, P. A., & Burge, C. B. (2014). RNA Bind-n-Seq: Quantitative assessment of the sequence and structural binding specificity of RNA binding proteins. *Molecular cell*, *54*(5), 887–900. https://doi.org/10.1016/j.molcel.2014.04.016

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., Funke, R., Gage, D., Harris, K., Heaford, A., Howland, J., Kann, L., Lehoczky, J., LeVine, R., McEwan, P., . . . International Human Genome Sequencing Consortium. (2001). Initial sequencing and analysis of the human genome. *Nature*, *409*(6822), 860–921. https://doi.org/10.1038/35057062

Langmead, B. (2010). Aligning short sequencing reads with bowtie. *Current protocols in bioinformatics / editoral board, Andreas D. Baxevanis ... [et al.]*, *32*(1), 11–17.

Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with bowtie 2. *Nature methods*, *9*(4), 357–359. https://doi.org/10.1038/nmeth.1923

Lasman, L., Krupalnik, V., Viukov, S., Mor, N., Aguilera-Castrejon, A., Schneir, D., Bayerl, J., Mizrahi, O., Peles, S., Tawil, S., Sathe, S., Nachshon, A., Shani, T., Zerbib, M., Kilimnik, I., Aigner, S., Shankar, A., Mueller, J. R., Schwartz, S., ... Hanna, J. H. (2020). Context-dependent functional compensation between ythdf m6a reader proteins. *Genes & development*, *34*(19-20), 1373–1391. https://doi.org/10.1101/gad.340695.120

Lawrence, M., Gentleman, R., & Carey, V. (2009). Rtracklayer: An R package for interfacing with genome browsers. *Bioinformatics*, *25*(14), 1841–1842. https://doi.org/10.1093/bioinformatics/btp328

Lee, F. C. Y., Chakrabarti, A. M., Hänel, H., Monzón-Casanova, E., Hallegger, M., Militti, C., Capraro, F., Sadée, C., Toolan-Kerr, P., Wilkins, O., Turner, M., König, J., Sibley, C. R., & Ule, J. (2021). *An improved iCLIP protocol*. https://doi.org/10.1101/2021.08.27.457890

Lee, F. C. Y., & Ule, J. (2018). Advances in CLIP technologies for studies of Protein-RNA interactions. *Molecular cell*, *69*(3), 354–369. https://doi.org/10.1016/j.molcel.2018.01.005

Lee, J.-H., Wang, R., Xiong, F., Krakowiak, J., Liao, Z., Nguyen, P. T., Moroz-Omori, E. V., Shao, J., Zhu, X., Bolt, M. J., Wu, H., Singh, P. K., Bi, M., Shi, C. J., Jamal, N., Li, G., Mistry, R., Jung, S. Y., Tsai, K.-L., ... Li, W. (2021). Enhancer RNA m6a methylation facilitates transcriptional condensate formation and gene activation. *Molecular cell*, *81*(16), 3368–3385.e9. https://doi.org/10.1016/j.molcel.2021.07.024

Leger, A., Amaral, P. P., Pandolfini, L., Capitanchik, C., et al. (2019a). RNA modifications detection by comparative nanopore direct RNA sequencing. *BioRxiv*.

Leger, A., Amaral, P. P., Pandolfini, L., Capitanchik, C., Capraro, F., Barbieri, I., Migliori, V., Luscombe, N. M., Enright, A. J., Tzelepis, K., Ule, J., Fitzgerald, T., Birney, E., Leonardi, T., & Kouzarides, T. (2019b).

RNA modifications detection by comparative nanopore direct RNA sequencing. https://doi.org/10.1101/843136

Lence, T., Akhtar, J., Bayer, M., Schmid, K., Spindler, L., Ho, C. H., Kreim, N., Andrade-Navarro, M. A., Poeck, B., Helm, M., & Roignant, J.-Y. (2016). M(6)a modulates neuronal functions and sex determination in drosophila. *Nature*, *540*(7632), 242–247. https://doi.org/10.1038/nature20568

Li, J., Mahajan, A., & Tsai, M.-D. (2006). Ankyrin repeat: A unique motif mediating protein-protein interactions. *Biochemistry*, *45*(51), 15168–15178. https://doi.org/10.1021/bi062188q

Li, J.-H., Liu, S., Zhou, H., Qu, L.-H., & Yang, J.-H. (2014). Starbase v2.0: Decoding miRNA-ceRNA, miRNA-ncRNA and protein-RNA interaction networks from large-scale CLIP-Seq data. *Nucleic acids research*, *42*(Database issue), D92–7. https://doi.org/10.1093/nar/gkt1248

Linares, A. J., Lin, C.-H., Damianov, A., Adams, K. L., Novitch, B. G., & Black, D. L. (2015). The splicing regulator PTBP1 controls the activity of the transcription factor pbx1 during neuronal differentiation. *eLife*, *4*, e09268. https://doi.org/10.7554/eLife.09268

Linder, B., Grozhik, A. V., Olarerin-George, A. O., Meydan, C., Mason, C. E., & Jaffrey, S. R. (2015). Single-nucleotide-resolution mapping of m6a and m6am throughout the transcriptome. *Nature methods*, *12*(8), 767–772. https://doi.org/10.1038/nmeth.3453

Liu, H., Begik, O., Lucas, M. C., Ramirez, J. M., Mason, C. E., Wiener, D., Schwartz, S., Mattick, J. S., Smith, M. A., & Novoa, E. M. (2019). Accurate detection of m6a RNA modifications in native RNA sequences. *Nature communications*, *10*(1), 4079. https://doi.org/10.1038/s41467-019-11713-9

Liu, J., Gao, M., Xu, S., Chen, Y., Wu, K., Liu, H., Wang, J., Yang, X., Wang, J., Liu, W., Bao, X., & Chen, J. (2020). YTHDF2/3 are required for somatic reprogramming through different RNA deadenylation pathways.

Cell reports, *32*(10), 108120. https://doi.org/10.1016/j.celrep.2020.108120

Liu, J., Dou, X., Chen, C., Chen, C., Liu, C., Xu, M. M., Zhao, S., Shen, B., Gao, Y., Han, D., & He, C. (2020). N6-methyladenosine of chromosome-associated regulatory RNA regulates chromatin state and transcription. *Science*, *367*(6477), 580–586. https://doi.org/10.1126/science.aay6018

Liu, J., Li, K., Cai, J., Zhang, M., Zhang, X., Xiong, X., Meng, H., Xu, X., Huang, Z., Peng, J., Fan, J., & Yi, C. (2020). Landscape and regulation of m6a and m6am methylome across human and mouse tissues. *Molecular cell*, *77*(2), 426–440.e6. https://doi.org/10.1016/j.molcel.2019.09.032

Liu, N., Dai, Q., Zheng, G., He, C., Parisien, M., & Pan, T. (2015). N(6)-methyladenosine-dependent RNA structural switches regulate RNA-protein interactions. *Nature*, *518*(7540), 560–564. https://doi.org/10.1038/nature14234

Liu, N., Zhou, K. I., Parisien, M., Dai, Q., Diatchenko, L., & Pan, T. (2017). N6-methyladenosine alters RNA structure to regulate binding of a low-complexity protein. *Nucleic acids research*, *45*(10), 6051–6063. https://doi.org/10.1093/nar/gkx141

Loher, P., Telonis, A. G., & Rigoutsos, I. (2017). MINTmap: Fast and exhaustive profiling of nuclear and mitochondrial tRNA fragments from short RNA-seq data. *Scientific reports*, *7*, 41184. https://doi.org/10.1038/srep41184

Lorenz, D. A., Sathe, S., Einstein, J. M., & Yeo, G. W. (2019). Direct RNA sequencing enables m6a detection in endogenous transcript isoforms at base specific resolution. *RNA*. https://doi.org/10.1261/rna.072785.119

Lovci, M. T., Ghanem, D., Marr, H., Arnold, J., Gee, S., Parra, M., Liang, T. Y., Stark, T. J., Gehman, L. T., Hoon, S., Massirer, K. B., Pratt, G. A., Black, D. L., Gray, J. W., Conboy, J. G., & Yeo, G. W. (2013). Rbfox proteins regulate alternative mRNA splicing through evolutionarily conserved RNA bridges. *Nature structural & molecular biology*, *20*(12), 1434–1442. https://doi.org/10.1038/nsmb.2699

Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome biology*, *15*(12), 550. https://doi.org/10.1186/s13059-014-0550-8

Luo, S., & Tong, L. (2014). Molecular basis for the recognition of methylated adenines in RNA by the eukaryotic YTH domain. *Proceedings of the National Academy of Sciences of the United States of America*, *111*(38), 13834–13839. https://doi.org/10.1073/pnas.1412742111

Ma, H., Wang, X., Cai, J., Dai, Q., Natchiar, S. K., Lv, R., Chen, K., Lu, Z., Chen, H., Shi, Y. G., Lan, F., Fan, J., Klaholz, B. P., Pan, T., Shi, Y., & He, C. (2019). N6-Methyladenosine methyltransferase ZCCHC4 mediates ribosomal RNA methylation. *Nature chemical biology*, *15*(1), 88–94. https://doi.org/10.1038/s41589-018-0184-3

Mabin, J. W., Lewis, P. W., Brow, D. A., & Dvinge, H. (2021). Human spliceosomal snRNA sequence variants generate variant spliceosomes. *RNA*, *27*(10), 1186–1203. https://doi.org/10.1261/rna.078768.121

Machnicka, M. A., Milanowska, K., Osman Oglou, O., Purta, E., Kurkowska, M., Olchowik, A., Januszewski, W., Kalinowski, S., Dunin-Horkawicz, S., Rother, K. M., Helm, M., Bujnicki, J. M., & Grosjean, H. (2013). MODOMICS: A database of RNA modification pathways–2013 update. *Nucleic acids research*, *41*(Database issue), D262–7. https://doi.org/10.1093/nar/gks1007

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal*, *17*(1), 10–12. https://doi.org/10.14806/ej.17.1.200

Martinez-Rucobo, F. W., Kohler, R., van de Waterbeemd, M., Heck, A. J. R., Hemann, M., Herzog, F., Stark, H., & Cramer, P. (2015). Molecular basis of Transcription-Coupled Pre-mRNA capping. *Molecular cell*, *58*(6), 1079–1089. https://doi.org/10.1016/j.molcel.2015.04.004

Marz, M., Kirsten, T., & Stadler, P. F. (2008). Evolution of spliceosomal snRNA genes in metazoan animals. *Journal of molecular evolution*, *67*(6), 594–607. https://doi.org/10.1007/s00239-008-9149-6

McIntyre, A. B. R., Gokhale, N. S., Cerchietti, L., Jaffrey, S. R., Horner, S. M., & Mason, C. E. (2020). Limits in the detection of m6a changes using MeRIP/m6A-seq. *Scientific reports*, *10*(1), 6590. https://doi.org/10.1038/s41598-020-63355-3

Meiser, N., Mench, N., & Hengesbach, M. (2020). RNA secondary structure dependence in METTL3-METTL14 mRNA methylation is modulated by the n-terminal domain of METTL3. *Biological chemistry*, *402*(1), 89–98. https://doi.org/10.1515/hsz-2020-0265

Meyer, C., Garzia, A., Mazzola, M., Gerstberger, S., Molina, H., & Tuschl, T. (2018). The TIA1 RNA-Binding protein family regulates EIF2AK2-Mediated stress response and cell cycle progression. *Molecular cell*, *69*(4), 622–635.e6. https://doi.org/10.1016/j.molcel.2018.01.011

Meyer, K. D. (2019). DART-seq: An antibody-free method for global m6a detection. *Nature methods*, *16*(12), 1275–1280. https://doi.org/10.1038/s41592-019-0570-0

Meyer, K. D., Patil, D. P., Zhou, J., Zinoviev, A., Skabkin, M. A., Elemento, O., Pestova, T. V., Qian, S.-B., & Jaffrey, S. R. (2015). 5' UTR m(6)a promotes Cap-Independent translation. *Cell*, *163*(4), 999–1010. https://doi.org/10.1016/j.cell.2015.10.012

Meyer, K. D., Saletore, Y., Zumbo, P., Elemento, O., Mason, C. E., & Jaffrey, S. R. (2012). Comprehensive analysis of mRNA methylation reveals enrichment in 3' UTRs and near stop codons. *Cell*, *149*(7), 1635–1646. https://doi.org/10.1016/j.cell.2012.05.003

Mo, J., Liang, H., Su, C., Li, P., Chen, J., & Zhang, B. (2021). DDX3X: Structure, physiologic functions and cancer. *Molecular cancer*, *20*(1), 38. https://doi.org/10.1186/s12943-021-01325-7

Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L., & Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature methods*, *5*(7), 621–628. https://doi.org/10.1038/nmeth.1226

Neelagandan, N., Gonnella, G., Dang, S., Janiesch, P. C., Miller, K. K., Küchler, K., Marques, R. F., Indenbirken, D., Alawi, M., Grundhoff, A., Kurtz, S., & Duncan, K. E. (2019). TDP-43 enhances translation of specific mRNAs linked to neurodegenerative disease. *Nucleic acids research*, *47*(1), 341–361. https://doi.org/10.1093/nar/gky972

Neiman, A. M. (2005). Ascospore formation in the yeast saccharomyces cerevisiae. *Microbiology and molecular biology reviews: MMBR*, *69*(4), 565–584. https://doi.org/10.1128/MMBR.69.4.565-584.2005

Nishimura, K., Yamada, R., Hagihara, S., Iwasaki, R., Uchida, N., Kamura, T., Takahashi, K., Torii, K. U., & Fukagawa, T. (2020). A super-sensitive auxin-inducible degron system with an engineered auxin-TIR1 pair. *Nucleic acids research*, *48*(18), e108. https://doi.org/10.1093/nar/gkaa748

Oberbauer, V., & Schaefer, M. R. (2018). tRNA-Derived small RNAs: Biogenesis, modification, function and potential impact on human disease development. *Genes*, *9*(12). https://doi.org/10.3390/genes9120607

Oh, S., Flynn, R. A., Floor, S. N., Purzner, J., Martin, L., Do, B. T., Schubert, S., Vaka, D., Morrissy, S., Li, Y., Kool, M., Hovestadt, V., Jones, D. T. W., Northcott, P. A., Risch, T., Warnatz, H.-J., Yaspo, M.-L., Adams, C. M., Leib, R. D., . . . Cho, Y.-J. (2016). Medulloblastoma-associated DDX3 variant selectively alters the translational response to stress. *Oncotarget*, *7*(19), 28169–28182. https://doi.org/10.18632/oncotarget.8612

O'Reilly, D., Dienstbier, M., Cowley, S. A., Vazquez, P., Drozdz, M., Taylor, S., James, W. S., & Murphy, S. (2013). Differentially expressed, variant U1 snRNAs regulate gene expression in human cells. *Genome research*, *23*(2), 281–291. https://doi.org/10.1101/gr.142968.112

Oxford Nanopore Technologies. (2018). Tombo: Detection of non-standard nucleotides using the genome-resolved raw nanopore signal. *Oxford Nanopore Technologies*.

Panasenko, O. O., Somasekharan, S. P., Villanyi, Z., Zagatti, M., Bezrukov, F., Rashpa, R., Cornut, J., Iqbal, J., Longis, M., Carl, S. H., Peña, C., Panse, V. G., & Collart, M. A. (2019). Co-translational assembly of proteasome subunits in NOT1-containing assemblysomes. *Nature structural & molecular biology*, *26*(2), 110–120. https://doi.org/10.1038/s41594-018-0179-5

Paris, J., Morgan, M., Campos, J., Spencer, G. J., Shmakova, A., Ivanova, I., Mapperley, C., Lawson, H., Wotherspoon, D. A., Sepulveda, C., Vukovic, M., Allen, L., Sarapuu, A., Tavosanis, A., Guitart, A. V., Villacreces, A., Much, C., Choe, J., Azar, A., . . . Kranc, K. R. (2019). Targeting the RNA m6a reader YTHDF2 selectively compromises cancer stem cells in acute myeloid leukemia. *Cell stem cell*. https://doi.org/10.1016/j.stem.2019.03.021

Parker, M. T., Knop, K., Sherwood, A. V., Schurch, N. J., Mackinnon, K., Gould, P. D., Hall, A. J., Barton, G. J., & Simpson, G. G. (2020). Nanopore direct RNA sequencing maps the complexity of arabidopsis mRNA processing and m6a modification. *eLife*, *9*. https://doi.org/10.7554/eLife.49658

Patil, D. P., Chen, C.-K., Pickering, B. F., Chow, A., Jackson, C., Guttman, M., & Jaffrey, S. R. (2016). M(6)a RNA methylation promotes XIST-mediated transcriptional repression. *Nature*, *537*(7620), 369–373. https://doi.org/10.1038/nature19342

Patil, D. P., Pickering, B. F., & Jaffrey, S. R. (2018). Reading m6a in the transcriptome: m6A-Binding proteins. *Trends in cell biology*, *28*(2), 113–127. https://doi.org/10.1016/j.tcb.2017.10.001

Paul, M. S., & Bass, B. L. (1998). Inosine exists in mRNA at tissue-specific levels and is most abundant in brain mRNA. *The EMBO journal*, *17*(4), 1120–1127. https://doi.org/10.1093/emboj/17.4.1120

Pendleton, K. E., Chen, B., Liu, K., Hunter, O. V., Xie, Y., Tu, B. P., & Conrad, N. K. (2017). The U6 snRNA m6a methyltransferase METTL16 regulates SAM synthetase intron retention. *Cell*, *169*(5), 824–835.e14. https://doi.org/10.1016/j.cell.2017.05.003

Pichot, F., Marchand, V., Helm, M., & Motorin, Y. (2021). Non-Redundant tRNA reference sequences for deep sequencing analysis of tRNA abundance and epitranscriptomic RNA modifications. *Genes*, *12*(1). https://doi.org/10.3390/genes12010081

Ping, X.-L., Sun, B.-F., Wang, L., Xiao, W., Yang, X., Wang, W.-J., Adhikari, S., Shi, Y., Lv, Y., Chen, Y.-S., Zhao, X., Li, A., Yang, Y., Dahal, U., Lou, X.-M., Liu, X., Huang, J., Yuan, W.-P., Zhu, X.-F., . . . Yang, Y.-G. (2014). Mammalian WTAP is a regulatory subunit of the RNA n6-methyladenosine methyltransferase. *Cell research*, *24*(2), 177–189. https://doi.org/10.1038/cr.2014.3

Pinkard, O., McFarland, S., Sweet, T., & Coller, J. (2020). Quantitative tRNA-sequencing uncovers metazoan tissue-specific tRNA regulation. *Nature communications*, *11*(1), 4104. https://doi.org/10.1038/s41467-020-17879-x

Pratanwanich, P. N., Yao, F., Chen, Y., Koh, C. W. Q., Wan, Y. K., Hendra, C., Poon, P., Goh, Y. T., Yap, P. M. L., Chooi, J. Y., Chng, W. J., Ng, S. B., Thiery, A., Goh, W. S. S., & Göke, J. (2021). Identification of differential RNA modifications from nanopore direct RNA sequencing with xpore. *Nature biotechnology*, *39*(11), 1394–1402. https://doi.org/10.1038/s41587-021-00949-w

Pruitt, K. D., Tatusova, T., & Maglott, D. R. (2005). NCBI reference sequence (RefSeq): A curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic acids research*, *33*(Database issue), D501–4. https://doi.org/10.1093/nar/gki025

Quinlan, A. R. (2014). BEDTools: The Swiss-Army tool for genome feature analysis. *Current protocols in bioinformatics / editoral board, Andreas D. Baxevanis ... [et al.]*, *47*, 11.12.1–34. https://doi.org/10.1002/0471250953.bi1112s47

Ramırez, F., Dündar, F., Diehl, S., Grüning, B. A., & Manke, T. (2014). Deeptools: A flexible platform for exploring deep-sequencing data. *Nucleic acids research*, *42*(Web Server issue), W187–91. https://doi.org/10.1093/nar/gku365

Rana, A. K., & Ankri, S. (2016). Reviving the RNA world: An insight into the appearance of RNA methyltransferases. *Frontiers in genetics*, *7*, 99. https://doi.org/10.3389/fgene.2016.00099

Ray, D., Ha, K. C. H., Nie, K., Zheng, H., Hughes, T. R., & Morris, Q. D. (2017). RNAcompete methodology and application to determine sequence preferences of unconventional RNA-binding proteins. *Methods*, *118-119*, 3–15. https://doi.org/10.1016/j.ymeth.2016.12.003

Reddy, R., Li, W. Y., Henning, D., Choi, Y. C., Nohga, K., & Busch, H. (1981). Characterization and subcellular localization of 7-8 S RNAs of novikoff hepatoma. *The Journal of biological chemistry*, *256*(16), 8452–8457.

Reddy, R., & Busch, H. (1988). Small nuclear RNAs: RNA sequences, structure, and modifications. In M. L. Birnstiel (Ed.), *Structure and function of major and minor small nuclear ribonucleoprotein particles* (pp. 1–37). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-73020-7_1

Rice, P., Longden, I., & Bleasby, A. (2000). EMBOSS: The european molecular biology open software suite. *Trends in genetics: TIG*, *16*(6), 276–277. https://doi.org/10.1016/s0168-9525(00)02024-2

Roberts, J. T., Porman, A. M., & Johnson, A. M. (2021). Identification of m6a residues at single-nucleotide resolution using eCLIP and an accessible custom analysis pipeline. *RNA*, *27*(4), 527–541. https://doi.org/10.1261/rna.078543.120

Rogelj, B., Easton, L. E., Bogu, G. K., Stanton, L. W., Rot, G., Curk, T., Zupan, B., Sugimoto, Y., Modic, M., Haberman, N., Tollervey, J., Fujii, R., Takumi, T., Shaw, C. E., & Ule, J. (2012). Widespread binding of FUS along nascent RNA regulates alternative splicing in the brain. *Scientific reports*, *2*, 603. https://doi.org/10.1038/srep00603

Roignant, J.-Y., & Soller, M. (2017). M6a in mRNA: An ancient mechanism for Fine-Tuning gene expression. *Trends in genetics: TIG*, *33*(6), 380–390. https://doi.org/10.1016/j.tig.2017.04.003

Rojas, M., Farr, G. W., Fernandez, C. F., Lauden, L., McCormack, J. C., & Wolin, S. L. (2012). Yeast gis2 and its human ortholog CNBP are novel components of stress-induced RNP granules. *PloS one*, *7*(12), e52824. https://doi.org/10.1371/journal.pone.0052824

Rot, G., Wang, Z., Huppertz, I., Modic, M., Lenče, T., Hallegger, M., Haberman, N., Curk, T., von Mering, C., & Ule, J. (2017). High-Resolution RNA maps suggest common principles of splicing and polyadenylation regulation by TDP-43. *Cell reports*, *19*(5), 1056–1067. https://doi.org/10.1016/j.celrep.2017.04.028

Russo, A., Scardigli, R., La Regina, F., Murray, M. E., Romano, N., Dickson, D. W., Wolozin, B., Cattaneo, A., & Ceci, M. (2017). Increased cytoplasmic TDP-43 reduces global protein synthesis by interacting with RACK1 on polyribosomes. *Human molecular genetics*, *26*(8), 1407–1418. https://doi.org/10.1093/hmg/ddx035

Sabi, R., & Tuller, T. (2015). A comparative genomics study on the effect of individual amino acids on ribosome stalling. *BMC genomics*, *16 Suppl 10*, S5. https://doi.org/10.1186/1471-2164-16-S10-S5

Safra, M., Sas-Chen, A., Nir, R., Winkler, R., Nachshon, A., Bar-Yaacov, D., Erlacher, M., Rossmanith, W., Stern-Ginossar, N., & Schwartz, S.

(2017). The m(1)a landscape on cytosolic and mitochondrial mRNA at single-base resolution. *Nature*. https://doi.org/10.1038/nature24456

Sammons, M. A., Samir, P., & Link, A. J. (2011). Saccharomyces cerevisiae gis2 interacts with the translation machinery and is orthogonal to myotonic dystrophy type 2 protein ZNF9. *Biochemical and biophysical research communications*, *406*(1), 13–19. https://doi.org/10.1016/j.bbrc.2011.01.086

Scherrer, T., Femmer, C., Schiess, R., Aebersold, R., & Gerber, A. P. (2011). Defining potentially conserved RNA regulons of homologous zinc-finger RNA-binding proteins. *Genome biology*, *12*(1), R3. https://doi.org/10.1186/gb-2011-12-1-r3

Schibler, U., Kelley, D. E., & Perry, R. P. (1977). Comparison of methylated sequences in messenger RNA and heterogeneous nuclear RNA from mouse L cells. *Journal of molecular biology*, *115*(4), 695–714. https://doi.org/10.1016/0022-2836(77)90110-3

Schöller, E., Weichmann, F., Treiber, T., Ringle, S., Treiber, N., Flatley, A., Feederle, R., Bruckmann, A., & Meister, G. (2018). Interactions, localization, and phosphorylation of the m6a generating METTL3-METTL14-WTAP complex. *RNA*, *24*(4), 499–512. https://doi.org/10.1261/rna.064063.117

Schwartz, S. (2018). M1a within cytoplasmic mRNAs at single nucleotide resolution: A reconciled transcriptome-wide map. *RNA*, *24*(11), 1427–1436. https://doi.org/10.1261/rna.067348.118

Schwartz, S., Agarwala, S. D., Mumbach, M. R., Jovanovic, M., Mertins, P., Shishkin, A., Tabach, Y., Mikkelsen, T. S., Satija, R., Ruvkun, G., Carr, S. A., Lander, E. S., Fink, G. R., & Regev, A. (2013). High-resolution mapping reveals a conserved, widespread, dynamic mRNA methylation program in yeast meiosis. *Cell*, *155*(6), 1409–1421. https://doi.org/10.1016/j.cell.2013.10.047

Schwartz, S., Bernstein, D. A., Mumbach, M. R., Jovanovic, M., Herbst, R. H., León-Ricardo, B. X., Engreitz, J. M., Guttman, M., Satija, R., Lander, E. S., Fink, G., & Regev, A. (2014). Transcriptome-wide mapping reveals widespread dynamic-regulated pseudouridylation of ncRNA and mRNA. *Cell*, *159*(1), 148–162. https://doi.org/10.1016/j.cell.2014.08.028

Sendinc, E., Valle-Garcia, D., Jiao, A., & Shi, Y. (2020). Analysis of m6a RNA methylation in caenorhabditis elegans. *Cell discovery*, *6*(1), 47. https://doi.org/10.1038/s41421-020-00186-6

Shen, S., Park, J. W., Lu, Z.-X., Lin, L., Henry, M. D., Wu, Y. N., Zhou, Q., & Xing, Y. (2014). rMATS: Robust and flexible detection of differential alternative splicing from replicate RNA-Seq data. *Proceedings of the National Academy of Sciences of the United States of America*, *111*(51), E5593–601. https://doi.org/10.1073/pnas.1419161111

Shi, H., Wang, X., Lu, Z., Zhao, B. S., Ma, H., Hsu, P. J., Liu, C., & He, C. (2017). YTHDF3 facilitates translation and decay of n6-methyladenosine-modified RNA. *Cell research*, *27*(3), 315–328. https://doi.org/10.1038/cr.2017.15

Shi, H., Zhang, X., Weng, Y.-L., Lu, Z., Liu, Y., Lu, Z., Li, J., Hao, P., Zhang, Y., Zhang, F., Wu, Y., Delgado, J. Y., Su, Y., Patel, M. J., Cao, X., Shen, B., Huang, X., Ming, G.-L., Zhuang, X., … Zhou, T. (2018). M6a facilitates hippocampus-dependent learning and memory through YTHDF1. *Nature*, *563*(7730), 249–253. https://doi.org/10.1038/s41586-018-0666-1

Shigematsu, M., Honda, S., Loher, P., Telonis, A. G., Rigoutsos, I., & Kirino, Y. (2017). YAMAT-seq: An efficient method for high-throughput sequencing of mature transfer RNAs. *Nucleic acids research*, *45*(9), e70. https://doi.org/10.1093/nar/gkx005

Shima, H., Matsumoto, M., Ishigami, Y., Ebina, M., Muto, A., Sato, Y., Kumagai, S., Ochiai, K., Suzuki, T., & Igarashi, K. (2017). S-

Adenosylmethionine synthesis is regulated by selective N6-Adenosine methylation and mRNA degradation involving METTL16 and YTHDC1. *Cell reports*, *21*(12), 3354–3363. https://doi.org/10.1016/j.celrep.2017.11.092

Shu, X., Cao, J., Cheng, M., Xiang, S., Gao, M., Li, T., Ying, X., Wang, F., Yue, Y., Lu, Z., Dai, Q., Cui, X., Ma, L., Wang, Y., He, C., Feng, X., & Liu, J. (2020). A metabolic labeling method detects m6a transcriptome-wide at single base resolution. *Nature chemical biology*. https://doi.org/10.1038/s41589-020-0526-9

Śledź, P., & Jinek, M. (2016). Structural insights into the molecular mechanism of the m(6)a writer complex. *eLife*, *5*. https://doi.org/10.7554/eLife.18434

Smith, H. E., & Mitchell, A. P. (1989). A transcriptional cascade governs entry into meiosis in saccharomyces cerevisiae. *Molecular and cellular biology*, *9*(5), 2142–2152. https://doi.org/10.1128/mcb.9.5.2142-2152.1989

Smith, T., Heger, A., & Sudbery, I. (2017). UMI-tools: Modeling sequencing errors in unique molecular identifiers to improve quantification accuracy. *Genome research*, *27*(3), 491–499. https://doi.org/10.1101/gr.209601.116

Sontheimer, E. J., & Steitz, J. A. (1992). Three novel functional variants of human U5 small nuclear RNA. *Molecular and cellular biology*, *12*(2), 734–746. https://doi.org/10.1128/mcb.12.2.734-746.1992

Stone, E. M., Heun, P., Laroche, T., Pillus, L., & Gasser, S. M. (2000). MAP kinase signaling induces nuclear reorganization in budding yeast. *Current biology: CB*, *10*(7), 373–382. https://doi.org/10.1016/s0960-9822(00)00413-9

Strittmatter, L. M., Capitanchik, C., Newman, A. J., Hallegger, M., Norman, C. M., Fica, S. M., Oubridge, C., Luscombe, N. M., Ule, J., & Nagai, K. (2020). *PsiCLIP reveals dynamic RNA binding by DEAH-box helicases* *before and after exon ligation*. https://doi.org/10.1101/2020.03.15.992701

Strudwick, N., Brown, M., Parmar, V. M., & Schröder, M. (2010). Ime1 and ime2 are required for pseudohyphal growth of saccharomyces cerevisiae on nonfermentable carbon sources. *Molecular and cellular biology*, *30*(23), 5514–5530. https://doi.org/10.1128/mcb.00390-10

Suzuki, T. (2021). The expanding world of tRNA modifications and their disease relevance. *Nature reviews. Molecular cell biology*, *22*(6), 375–392. https://doi.org/10.1038/s41580-021-00342-0

Taggart, J., Wang, Y., Weisenhorn, E., MacDiarmid, C. W., Russell, J., Coon, J. J., & Eide, D. J. (2018). The GIS2 gene is repressed by a Zinc-Regulated bicistronic RNA in saccharomyces cerevisiae. *Genes*, *9*(9). https://doi.org/10.3390/genes9090462

Telonis, A. G., Loher, P., Kirino, Y., & Rigoutsos, I. (2014). Nuclear and mitochondrial tRNA-lookalikes in the human genome. *Frontiers in genetics*, *5*, 344. https://doi.org/10.3389/fgene.2014.00344

Tollervey, J. R., Curk, T., Rogelj, B., Briese, M., Cereda, M., Kayikci, M., König, J., Hortobágyi, T., Nishimura, A. L., Zupunski, V., Patani, R., Chandran, S., Rot, G., Zupan, B., Shaw, C. E., & Ule, J. (2011). Characterizing the RNA targets and position-dependent splicing regulation by TDP-43. *Nature neuroscience*, *14*(4), 452–458. https://doi.org/10.1038/nn.2778

Tsang, E., & Maciocci, G. (2020). Welcome to a new ERA of reproducible publishing [Accessed: 2022-2-2].

Uren, P. J., Bahrami-Samani, E., Burns, S. C., Qiao, M., Karginov, F. V., Hodges, E., Hannon, G. J., Sanford, J. R., Penalva, L. O. F., & Smith, A. D. (2012). Site identification in high-throughput RNA-protein interaction data. *Bioinformatics*, *28*(23), 3013–3020. https://doi.org/10.1093/bioinformatics/bts569 van den Berg, M. A., de Jong-Gubbels, P., Kortland, C. J., van Dijken, J. P., Pronk, J. T., & Steensma, H. Y. (1996). The two acetyl-coenzyme a synthetases of saccharomyces cerevisiae differ with respect to kinetic properties and transcriptional regulation. *The Journal of biological chemistry*, *271*(46), 28953–28959. https://doi.org/10.1074/jbc.271.46.28953

Van Haute, L., Hendrick, A. G., D'Souza, A. R., Powell, C. A., Rebelo-Guiomar, P., Harbour, M. E., Ding, S., Fearnley, I. M., Andrews, B., & Minczuk, M. (2019). METTL15 introduces n4-methylcytidine into human mitochondrial 12S rRNA and is required for mitoribosome biogenesis. *Nucleic acids research*, *47*(19), 10267–10281. https://doi.org/10.1093/nar/gkz735

Van Nostrand, E. L., Freese, P., Pratt, G. A., Wang, X., Wei, X., Xiao, R., Blue, S. M., Chen, J.-Y., Cody, N. A. L., Dominguez, D., Olson, S., Sundararaman, B., Zhan, L., Bazile, C., Bouvrette, L. P. B., Bergalet, J., Duff, M. O., Garcia, K. E., Gelboin-Burkhart, C., . . . Yeo, G. W. (2020). A large-scale binding and functional map of human RNA-binding proteins. *Nature*, *583*(7818), 711–719. https://doi.org/10.1038/s41586-020-2077-3

Van Nostrand, E. L., Pratt, G. A., Shishkin, A. A., Gelboin-Burkhart, C., Fang, M. Y., Sundararaman, B., Blue, S. M., Nguyen, T. B., Surka, C., Elkins, K., Stanton, R., Rigo, F., Guttman, M., & Yeo, G. W. (2016). Robust transcriptome-wide discovery of RNA-binding protein binding sites with enhanced CLIP (eCLIP). *Nature methods*, *13*, 508. https://doi.org/10.1038/nmeth.3810

Van Nostrand, E. L., Pratt, G. A., Yee, B. A., Wheeler, E. C., Blue, S. M., Mueller, J., Park, S. S., Garcia, K. E., Gelboin-Burkhart, C., Nguyen, T. B., Rabano, I., Stanton, R., Sundararaman, B., Wang, R., Fu, X.-D., Graveley, B. R., & Yeo, G. W. (2020). Principles of RNA processing from analysis of enhanced CLIP maps for 150 RNA binding proteins.

Genome biology, *21*(1), 90. https://doi.org/10.1186/s13059-020-01982-9

van Tran, N., Ernst, F. G. M., Hawley, B. R., Zorbas, C., Ulryck, N., Hackert, P., Bohnsack, K. E., Bohnsack, M. T., Jaffrey, S. R., Graille, M., & Lafontaine, D. L. J. (2019). The human 18S rRNA m6a methyltransferase METTL5 is stabilized by TRMT112. *Nucleic acids research*, *47*(15), 7719–7733. https://doi.org/10.1093/nar/gkz619

van Werven, F. J., Neuert, G., Hendrick, N., Lardenois, A., Buratowski, S., van Oudenaarden, A., Primig, M., & Amon, A. (2012). Transcription of two long noncoding RNAs mediates mating-type control of gametogenesis in budding yeast. *Cell*, *150*(6), 1170–1181. https://doi.org/10.1016/j.cell.2012.06.049

Varier, R. A., Sideri, T., Capitanchik, C., Manova, Z., Calvani, E., Rossi, A., Edupuganti, R. R., Ensinck, I., Chan, V. W. C., Patel, H., Kirkpatrick, J., Faull, P., Snijders, A. P., Vermeulen, M., Ralser, M., Ule, J., Luscombe, N. M., & van Werven, F. J. (2022). *M6a reader pho92 is recruited co-transcriptionally and couples translation efficacy to mRNA decay to promote meiotic fitness in yeast*. https://doi.org/10.1101/2022.01.20.477035

Vazquez-Arango, P., Vowles, J., Browne, C., Hartfield, E., Fernandes, H. J. R., Mandefro, B., Sareen, D., James, W., Wade-Martins, R., Cowley, S. A., Murphy, S., & O'Reilly, D. (2016). Variant U1 snRNAs are implicated in human pluripotent stem cell maintenance and neuromuscular disease. *Nucleic acids research*, *44*(22), 10960–10973. https://doi.org/10.1093/nar/gkw711

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., . . . SciPy 1.0 Contributors. (2020). SciPy 1.0: Fundamental algorithms for scientific computing in python. *Nature methods*, *17*(3), 261–272. https://doi.org/10.1038/s41592-019-0686-2

Wang, M., Xiao, Y., Li, Y., Wang, X., Qi, S., Wang, Y., Zhao, L., Wang, K., Peng, W., Luo, G.-Z., Xue, X., Jia, G., & Wu, L. (2021). RNA m6a modification functions in larval development and caste differentiation in honeybee (apis mellifera). *Cell reports*, *34*(1), 108580. https://doi.org/10.1016/j.celrep.2020.108580

Wang, X., Lu, Z., Gomez, A., Hon, G. C., Yue, Y., Han, D., Fu, Y., Parisien, M., Dai, Q., Jia, G., Ren, B., Pan, T., & He, C. (2014). N6-methyladenosine-dependent regulation of messenger RNA stability. *Nature*, *505*(7481), 117–120. https://doi.org/10.1038/nature12730

Wang, X., Zhao, B. S., Roundtree, I. A., Lu, Z., Han, D., Ma, H., Weng, X., Chen, K., Shi, H., & He, C. (2015). N(6)-methyladenosine modulates messenger RNA translation efficiency. *Cell*, *161*(6), 1388–1399. https://doi.org/10.1016/j.cell.2015.05.014

Wang, Y., Li, Y., Toth, J. I., Petroski, M. D., Zhang, Z., & Zhao, J. C. (2014). N6-methyladenosine modification destabilizes developmental regulators in embryonic stem cells. *Nature cell biology*, *16*(2), 191–198. https://doi.org/10.1038/ncb2902

Wang, Y., Xiao, Y., Dong, S., Yu, Q., & Jia, G. (2020). Antibody-free enzyme-assisted chemical approach for detection of n6-methyladenosine. *Nature chemical biology*. https://doi.org/10.1038/s41589-020-0525-x

Wang, Z., Kayikci, M., Briese, M., Zarnack, K., Luscombe, N. M., Rot, G., Zupan, B., Curk, T., & Ule, J. (2010). iCLIP predicts the dual splicing effects of TIA-RNA interactions. *PLoS biology*, *8*(10), e1000530. https://doi.org/10.1371/journal.pbio.1000530

Warda, A. S., Kretschmer, J., Hackert, P., Lenz, C., Urlaub, H., Höbartner, C., Sloan, K. E., & Bohnsack, M. T. (2017). Human METTL16 is a n6-methyladenosine (m6a) methyltransferase that targets pre-mRNAs and various non-coding RNAs. *EMBO reports*, *18*(11), 2004–2014. https://doi.org/10.15252/embr.201744940

Wei, C. M., Gershowitz, A., & Moss, B. (1975). Methylated nucleotides block 5' terminus of HeLa cell messenger RNA. *Cell*, *4*(4), 379–386. https://doi.org/10.1016/0092-8674(75)90158-0

Wei, C. M., Gershowitz, A., & Moss, B. (1976). 5'-terminal and internal methylated nucleotide sequences in HeLa cell mRNA. *Biochemistry*, *15*(2), 397–401. https://doi.org/10.1021/bi00647a024

Weiss, M. C., Preiner, M., Xavier, J. C., Zimorski, V., & Martin, W. F. (2018). The last universal common ancestor between ancient earth chemistry and the onset of genetics. *PLoS genetics*, *14*(8), e1007518. https://doi.org/10.1371/journal.pgen.1007518

Wen, J., Lv, R., Ma, H., Shen, H., He, C., Wang, J., Jiao, F., Liu, H., Yang, P., Tan, L., Lan, F., Shi, Y. G., He, C., Shi, Y., & Diao, J. (2018). Zc3h13 regulates nuclear RNA m6a methylation and mouse embryonic stem cell Self-Renewal. *Molecular cell*, *69*(6), 1028–1038.e6. https://doi.org/10.1016/j.molcel.2018.02.015

Wen, J.-T., Huang, Z.-H., Li, Q.-H., Chen, X., Qin, H.-L., & Zhao, Y. (2021). Research progress on the tsRNA classification, function, and application in gynecological malignant tumors. *Cell death discovery*, *7*(1), 388. https://doi.org/10.1038/s41420-021-00789-2

Wickham, H. (2010). Stringr: Modern, consistent string processing. *The R journal*, *2*(2), 38. https://doi.org/10.32614/rj-2010-012

Wickham, H. (2011). Ggplot2. *Wiley Interdisciplinary Reviews: Computational Statistics*, *3*(2), 180–185. https://doi.org/10.1002/wics.147

Wickham, H. (2016). *Ggplot2: Elegant graphics for data analysis*. Springer.

Wickham, H., Francois, R., Henry, L., Müller, K., et al. (2015). Dplyr: A grammar of data manipulation. *R package version 0. 4*, *3*.

Wilke, C. O., Wickham, H., & Wilke, M. C. O. (2019). Package 'cowplot'. *Streamlined Plot Theme and Plot Annotations for 'ggplot2*.

Wilkins, O. G., Capitanchik, C. et al. (2021). Ultraplex: A rapid, flexible, all-in-one fastq demultiplexer. *Wellcome Open.*

Woischnik, M., & Moraes, C. T. (2002). Pattern of organization of human mitochondrial pseudogenes in the nuclear genome. *Genome research*, *12*(6), 885–893. https://doi.org/10.1101/gr.227202

Wratten, L., Wilm, A., & Göke, J. (2021). Reproducible, scalable, and shareable analysis pipelines with bioinformatics workflow managers. *Nature methods*, *18*(10), 1161–1168. https://doi.org/10.1038/s41592-021-01254-9

Wu, C. C.-C., Peterson, A., Zinshteyn, B., Regot, S., & Green, R. (2020). Ribosome collisions trigger general stress responses to regulate cell fate. *Cell*, *182*(2), 404–416.e14. https://doi.org/10.1016/j.cell.2020.06.006

Wu, R., Li, A., Sun, B., Sun, J.-G., Zhang, J., Zhang, T., Chen, Y., Xiao, Y., Gao, Y., Zhang, Q., Ma, J., Yang, X., Liao, Y., Lai, W.-Y., Qi, X., Wang, S., Shu, Y., Wang, H.-L., Wang, F., … Yuan, Z. (2019). A novel m6a reader prrc2a controls oligodendroglial specification and myelination. *Cell research*, *29*(1), 23–41. https://doi.org/10.1038/s41422-018-0113-8

Xu, C., Liu, K., Ahmed, H., Loppnau, P., Schapira, M., & Min, J. (2015). Structural basis for the discriminative recognition of N6-Methyladenosine RNA by the human YT521-B homology domain family of proteins. *The Journal of biological chemistry*, *290*(41), 24902–24913. https://doi.org/10.1074/jbc.M115.680389

Xu, C., Wang, X., Liu, K., Roundtree, I. A., Tempel, W., Li, Y., Lu, Z., He, C., & Min, J. (2014). Structural basis for selective binding of m6a RNA by the YTHDC1 YTH domain. *Nature chemical biology*, *10*(11), 927–929. https://doi.org/10.1038/nchembio.1654

Xue, S., & Barna, M. (2012). Specialized ribosomes: A new frontier in gene regulation and organismal biology. *Nature reviews. Molecular cell biology*, *13*(6), 355–369. https://doi.org/10.1038/nrm3359

Yamasaki, S., Ivanov, P., Hu, G.-F., & Anderson, P. (2009). Angiogenin cleaves tRNA and promotes stress-induced translational repression. *The Journal of cell biology*, *185*(1), 35–42. https://doi.org/10.1083/jcb.200811106

Yang, F., Wang, X. Y., Zhang, Z. M., Pu, J., Fan, Y. J., Zhou, J., Query, C. C., & Xu, Y. Z. (2013). Splicing proofreading at 5' splice sites by ATPase prp28p. *Nucleic acids research*, *41*(8), 4660–4670. https://doi.org/10.1093/nar/gkt149

Yankova, E., Blackaby, W., Albertella, M., Rak, J., De Braekeleer, E., Tsagkogeorga, G., Pilka, E. S., Aspris, D., Leggate, D., Hendrick, A. G., Webster, N. A., Andrews, B., Fosbeary, R., Guest, P., Irigoyen, N., Eleftheriou, M., Gozdecka, M., Dias, J. M. L., Bannister, A. J., . . . Kouzarides, T. (2021). Small-molecule inhibition of METTL3 as a strategy against myeloid leukaemia. *Nature*, *593*(7860), 597–601. https://doi.org/10.1038/s41586-021-03536-w

Ye, J., McGinnis, S., & Madden, T. L. (2006). BLAST: Improvements for better sequence analysis. *Nucleic acids research*, *34*(Web Server issue), W6–9. https://doi.org/10.1093/nar/gkl164

Yeo, G. W., Coufal, N. G., Liang, T. Y., Peng, G. E., Fu, X.-D., & Gage, F. H. (2009). An RNA code for the FOX2 splicing regulator revealed by mapping RNA-protein interactions in stem cells. *Nature structural & molecular biology*, *16*(2), 130–137. https://doi.org/10.1038/nsmb.1545

Yue, Y., Liu, J., Cui, X., Cao, J., Luo, G., Zhang, Z., Cheng, T., Gao, M., Shu, X., Ma, H., Wang, F., Wang, X., Shen, B., Wang, Y., Feng, X., He, C., & Liu, J. (2018). VIRMA mediates preferential m6a mRNA methylation in 3'UTR and near stop codon and associates with alternative polyadenylation. *Cell discovery*, *4*, 10. https://doi.org/10.1038/s41421-018-0019-0

Yuen, D., Cabansay, L., Duncan, A., Luu, G., Hogue, G., Overbeck, C., Perez, N., Shands, W., Steinberg, D., Reid, C., Olunwa, N., Hansen, R., Sheets, E., O'Farrell, A., Cullion, K., O'Connor, B. D., Paten, B., & Stein, L. (2021). The dockstore: Enhancing a community platform for sharing reproducible and accessible computational protocols. *Nucleic acids research*, *49*(W1), W624–W632. https://doi.org/10.1093/nar/gkab346

Zaccara, S., & Jaffrey, S. R. (2020). A unified model for the function of YTHDF proteins in regulating m6A-Modified mRNA. *Cell*. https://doi.org/10.1016/j.cell.2020.05.012

Zarnack, K., König, J., Tajnik, M., Martincorena, I., Eustermann, S., Stévant, I., Reyes, A., Anders, S., Luscombe, N. M., & Ule, J. (2013). Direct competition between hnRNP C and U2AF65 protects the transcriptome from the exonization of alu elements. *Cell*, *152*(3), 453–466. https://doi.org/10.1016/j.cell.2012.12.023

Zarnegar, B. J., Flynn, R. A., Shen, Y., Do, B. T., Chang, H. Y., & Khavari, P. A. (2016). irCLIP platform for efficient characterization of protein–RNA interactions. *Nature methods*, *13*(6), 489–492. https://doi.org/10.1038/nmeth.3840

Zhang, C., Chen, Y., Sun, B., Wang, L., Yang, Y., Ma, D., Lv, J., Heng, J., Ding, Y., Xue, Y., Lu, X., Xiao, W., Yang, Y.-G., & Liu, F. (2017). M6a modulates haematopoietic stem and progenitor cell specification. *Nature*, *549*(7671), 273–276. https://doi.org/10.1038/nature23883

Zhang, Z., Chen, L.-Q., Zhao, Y.-L., Yang, C.-G., Roundtree, I. A., Zhang, Z., Ren, J., Xie, W., He, C., & Luo, G.-Z. (2019). Single-base mapping of m6a by an antibody-independent method. *Science Advances*, *5*(7), eaax0250. https://doi.org/10.1126/sciadv.aax0250

Zhang, Z., Chen, T., Chen, H.-X., Xie, Y.-Y., Chen, L.-Q., Zhao, Y.-L., Liu, B.-D., Jin, L., Zhang, W., Liu, C., Ma, D.-Z., Chai, G.-S., Zhang, Y., Zhao, W.-S., Ng, W. H., Chen, J., Jia, G., Yang, J., & Luo, G.-Z. (2021). Systematic calibration of epitranscriptomic maps using a synthetic modification-free RNA library. *Nature methods*. https://doi.org/10.1038/s41592-021-01280-7

Zhang, Z., & Xing, Y. (2017). CLIP-seq analysis of multi-mapped reads discovers novel functional RNA regulatory sites in the human transcriptome. *Nucleic acids research*, *45*(16), 9260–9271. https://doi.org/10.1093/nar/gkx646

Zhao, B., Nachtergaele, S., Roundtree, I. A., & He, C. (2017). Our views of dynamic n6-methyladenosine RNA methylation. *RNA*. https://doi.org/10.1261/rna.064295.117

Zheng, G., Qin, Y., Clark, W. C., Dai, Q., Yi, C., He, C., Lambowitz, A. M., & Pan, T. (2015). Efficient and quantitative high-throughput tRNA sequencing. *Nature methods*, *12*(9), 835–837. https://doi.org/10.1038/nmeth.3478

Zhong, S., Li, H., Bodi, Z., Button, J., Vespa, L., Herzog, M., & Fray, R. G. (2008). MTA is an arabidopsis messenger RNA adenosine methylase and interacts with a homolog of a sex-specific splicing factor. *The Plant cell*, *20*(5), 1278–1288. https://doi.org/10.1105/tpc.108.058883

Zhu, T., Roundtree, I. A., Wang, P., Wang, X., Wang, L., Sun, C., Tian, Y., Li, J., He, C., & Xu, Y. (2014). Crystal structure of the YTH domain of YTHDF2 reveals mechanism for recognition of n6-methyladenosine. *Cell research*, *24*(12), 1493–1496. https://doi.org/10.1038/cr.2014.152

Zhu, Y., Xu, G., Yang, Y. T., Xu, Z., Chen, X., Shi, B., Xie, D., Lu, Z. J., & Wang, P. (2019). POSTAR2: Deciphering the post-transcriptional regulatory logics. *Nucleic acids research*, *47*(D1), D203–D211. https://doi.org/10.1093/nar/gky830

Zhuang, M., Li, X., Zhu, J., Zhang, J., Niu, F., Liang, F., Chen, M., Li, D., Han, P., & Ji, S.-J. (2019). The m6a reader YTHDF1 regulates axon guidance through translational control of robo3.1 expression. *Nucleic acids research*. https://doi.org/10.1093/nar/gkz157